ESTADÍSTICA Y DECISIONES
Vol 1: La normal univariada, tablas de contingencia
Versión 18, junio 2016

José Rodríguez

13 de junio de 2016

TALLER SOBRE MANEJO DE CALCULADORA

Para el curso se necesita una calculadora que tenga las operaciones fundamentales, raíz cuadrada y paréntesis. La que hay en casa puede servir o también una Casio fx-350MS. Por favor marcarla con el nombre o con algún logo. El siguiente taller se hace en la calculadora y se reporta: "yo ya sé manejar mi calculadora" (no hay que presentar nada más). **Los errores de calculadora no tienen disculpa: es necesario intuir el rango de la respuesta para detectar errores de tecleo o de manejo**.

Si $a = 5$, $b = 2{,}1$, $c = -3{,}4$, $d = 3{,}2 \times 10^4$, $e = 2{,}7 \times 10^{-5}$, $f = -4{,}9 \times 10^{-4}$ ejecute las siguientes operaciones:

1. $b - 7c =$? En caso de duda, usamos paréntesis:
   $b - 7c = 2{,}1 - 7 \times (-3{,}4) = 25{,}9$
2. $a - 5b = -5{,}5$
3. $\sqrt{a + 5b} = \sqrt{(a + 5b)} = 3{,}93$. *Paréntesis obligatorio.*
4. $\sqrt{a - 5b}$. La raíz par de un negativo no existe.
5. $10^0 = 1$.
6. $10^3 = 1000$,
7. $10^{-2} = 0{,}01$,
8. $10^{-5} = 0{,}00001$.
9. $3{,}5 \times 10^{-5} = 0{,}000035$ Ayuda: calcule $0{,}000035 \times 1$.
10. $a^2 = 25$,
11. $d = 32000$
12. $a^b = 29{,}36$
13. $a^{-b} = 0{,}034 = 1/(a^b)$
14. $ad = 160000$
15. $ad/(cb) = -22408{,}96$
16. $\frac{\sqrt{a+5b}}{(a^b-(\frac{ad}{cb}))}$ Las operaciones muy largas las hacemos por pedazos desde adentro hacia afuera: $\frac{\sqrt{(a+5b)}}{(a^b-(\frac{ad}{cb}))} = \frac{3{,}93}{(29{,}36-(-22408{,}96))} = 0{,}00017$
17. $\frac{3}{7} + \frac{2}{7} = \frac{5}{7}$
18. $\frac{3}{7} - \frac{2}{5} = \frac{15-14}{35} = \frac{1}{35}$
19. $\frac{3}{7} \times \frac{2}{5} = \frac{6}{35}$
20. $\frac{3/7}{2/5} = \frac{15}{14}$
21. $50/100 = numerador/denominador = 0{,}5$. Si el numerador aumenta, el resultado aumenta. Si el numerador disminuye, el resultado disminuye. Si el denominador aumenta, el resultado disminuye. Si el denominador disminuye el resultado aumenta.
22. Las binomiales del capítulo II se aproximan con 4 cifras, lo demás con 2.

PREFACIO

La estadística es la herencia de los príncipes que por su vanagloria encuestaron sus imperios para enumerar la gente, los soldados, los navíos, los recursos y las rentas públicas. Más luego y en la medida en que el Estado era entendido como una responsabilidad para ser administrada, la estadística se convirtió en la herramienta que conectaba las necesidades del pueblo con los proyectos del Estado.

No es de admirar que la estadística sea hoy en día de uso diario tanto en lo personal como en lo empresarial pues, en cierta medida, cada ser humano es un príncipe de su propia vida y cada empresa un pequeño imperio. Y como si fuera poco, la estadística se confunde hoy en día con el método científico, que consiste básicamente en contrastar lo que uno ve con lo que uno cree. Más extensamente tenemos que el método científico tiene como objetivo responder preguntas muy básicas:

1) ¿Cómo es el mundo? Tenemos la estadística descriptiva, presentada en el primer capítulo.

2) ¿Por qué el mundo es como es? Formulamos creencias que al ser revestidas de matemáticas se denominan modelos, como en el capítulo dos, en el cual introducimos una abstracción del juego de lanzar una moneda y ver qué cae.

3) ¿Cómo puedo decidir si lo que creo es verdad o no? Hacemos experimentos y los estudiamos para contrastar lo que vemos con lo que creemos. A esto dedicamos todo el resto del libro.

4) ¿Quién es el responsable de lo que creemos? Antiguamente la ciencia se revestía de tanta gloria y de tanto renombre que se producía la impresión de que la ciencia simplemente retraba el mundo tal cual es y por ello la ciencia era la verdad absoluta. Pero ya hace bastante tiempo que los filósofos se dieron cuenta de que la ciencia no era la fuente de la verdad, sino que la ciencia es una subcultura humana que antes que nada busca descripciones efectivas. Su efectividad puede hacerse parecer a la del macho aquel que sabe que su mujer le funciona dándole con la culata del fusil. Y eso es algo que, según él, es una verdad experimental contundente. Pero por otro lado, si uno renuncia a la efectividad, se queda con la incertidumbre y el desasosiego que ella causa. Pero eso no es tan malo: ¿qué mejor comienzo para la creatividad que no saber qué hacer ni para dónde coger? Pues bien, la estadística crea opciones, pero es el ser humano quien toma decisiones y quien debe responder.

Podemos ahora enunciar el **objetivo global del curso:** entender el método científico de acuerdo a los estándares de la modernidad, incluyendo manejo de datos, formulación y estudio de hipótesis concernientes a experimentos de productividad y de homogeneidad, y tablas de contingencia.

Metodología: los estudiantes se organizan por grupos, preparan en casa la tarea asignada y la exponen en el tablero, lo cual da puntos para la nota de quices y tablero. Obsérvese que el profesor no explica, son los estudiantes los que le explican a él y a sus compañeros. El que más aprende es el que pasa al tablero pues es casi seguro que entre todos encontramos hasta el más mínimo error. Por eso, pasar al tablero da punto extra aunque la solución haya tenido errores. Para poder pasar al tablero se necesita comprar un marcador pero que sea borrable (dry erase). El fuerte de la clase consiste en resistir la lluvia de preguntas que el profesor le hace a los estudiantes para que todos nos aseguremos de que se entiende el material. Sabemos por nuestra mucha experiencia, es decir, por los muchísimos rajados, que el trabajo diario para entender y para agilizar es necesario para pasar la materia.

El fruto de tanto esfuerzo es tremendamente formativo pues produce un o una estudiante bien preparado o preparada para:

- Utilizar el método científico en su pensar diario así sea de forma cualitativa.
- Argumentar y defender las ideas propias.
- Emprender grandes obras gracias a la disciplina del trabajo quizá lento pero continuo.
- Estudiar por su cuenta el material que sea.
- Trabajar eficientemente tanto sólos como en equipo.
- Enfrentar noblemente cualquier proceso de fiscalización que es la base de la democracia transparente.

Aunque se ha hecho un esfuerzo para que los ejemplos parezcan realistas, en realidad son todos inventados.

En nuestro curso hay varios tipos de estudiantes: los supervisores, los que luchan por un 3, los que pelean por un 5, los admirables, los colombianos, los héroes y los quemados. Distinguirlos es muy sencillo:

1. Los supervisores se acomodan en un grupo y miran cómo trabajan los demás.
2. Los que luchan por un tres trabajan en grupo y se reparten los trabajos.
3. Los que pelean por un 5 hacen todo solos todos los días, comparan procedimientos y respuestas en su grupo de estudio y alegan entre ellos hasta que todo quede claro.
4. Hay otras personas que empiezan desconectados porque vienen con experiencias muy negativas en matemáticas, luego pasan a luchar por un 3, pero después ya no se satisfacen con un 4. ¡Son los admirables!
5. Tenemos también alumnos tremendamente inteligentes que dejan todo para lo último, un estilo muy colombiano, y por supuesto, se rajan.
6. ¿Y qué de los héroes? Son personas que trabajan y trabajan y siguen trabajando pero sin fruto. Pero no se rinden y contra todo pronóstico repuntan hacia el tercero o incluso cuarto parcial y lo hacen con honores.
7. Finalmente hay personas que tratan de hacer el héroe pero se queman pues su objetivo es obtener las respuestas sin entender los por qués. Es aconsejable retirarse si para el parcial III uno aún no ha logrado conectarse al curso.

José

PROGRAMA DE ESTADISTICA PARA SOCIALES, I -2016		
La clase empieza a las en punto y termina a las y 50		
Dia	Tema	Problemas
Enero 18L	Descriptiva: 1.1 Datos 1.2 Histogramas 1.3 Mediana	Taller I, (pág. 47): 1,2,3,4
19Ma	1.4-1.5 La media, media ponderada, Varianza	5,6,7
21J	1.6-1.7 Datos agrupados. Rango intercuartilar. probabilidad, Taller Excel,pág. 267	8,9,10
22V	Cap 2 Probabilidad. Distribuciones, 2.1-2.2 La binomial $Bi(2, 1/2)$,	11,12,13
25L	2.3 El triángulo de Pascal, la $Bi(n, p)$; 1.7 Marco global	14,15,16.
26Ma	Repaso. Quiz sorpresa, quizá hoy o tal vez después.	Parcial 1. ver 1 (pág. 51)
28J	Repaso	P1. ver 2
29V	Repaso	P1. ver 3
Febrero 1L	Quiz con Hoja personal de fórmulas y calculadora	
2Ma	Repaso	P1. ver 4
4J	Parcial I con Hoja personal de fórmulas y calculadora	
5V	4.1 La normal como envolvente	Taller II, pág. 117:1,2,3
8L	4.2 a 4.4 El teorema-z; la tabla z	4,5,6
9Ma	5.2 Aplicación de la normal	7,8,9
11J	5.2 La noción de interesante	14, 15, 16
12V	5.4 La Hipótesis nula. IC (Intervalos de confianza) de la media con un dato.	23, 31a y b
15L	6.1 El TLC (Teorema del Límite Central): teoría y aplicaciones	34, 35,36
16Ma	6.1 El TLC, 6.2 IC de la media con varios datos	37,38,40a y b
18J	Repaso. Quiz sorpresa, quizá hoy o tal vez después.	P2. Ver 1 (pág. 127)
19V	Repaso	P2. Ver 2
22L	Quiz con libro y calculadora	
23Ma	Repaso	P2. Ver 3
25J	Repaso	P2. Ver 4
26V	Parcial II con libro y calculadora	
29L	LA CIENCIA SE VISTE DE GALA	
	9.2 Experimentos prototipo: la media + tlc, una proporción	Taller III, pág. 191: 1,2,3
Marzo 1Ma	9.2 Dos proporciones, datos apareados, dos varianzas	4, 5,6
3J	9.2 Dos medias con varianzas iguales	7, 8
4V	9.2 Dos medias con varianzas diferentes	9
7L	Quiz: ¿Con cuál caso, con cuántas y cuáles colas se resuelve cada problema del 21 al 51?	pág 194
8Ma	Repaso. Quiz sorpresa, quizá hoy o tal vez después.	P3 Ver 1 (pág. 202)
10J	Repaso	P3 Ver 2
11V	Quiz con folleto de fórmulas	
14L	Repaso	P3 Ver 3
15Ma	Repaso.	P3 Ver 4
17J	Parcial III con folleto de fórmulas.	
18V	Cap 11: Los 10 experimentos básicos. Una media	Taller IV, pág. 253: 1,2,3
Marzo 21/25	SEMANA DE TRABAJO INDIVIDUAL,	
28L	Una media, una varianza, una proporción	4, 6, 8
29M	Dos proporciones, una y dos varianzas	9,10,11
31J	Dos varianzas. Dos medias. Árboles	12,13,15,
Abril 1V	12.1 Probabilidad condicional.	18, 19
4L	12.1 Árboles	20, 21
5Ma	12.2 Tablas de contingencia	22, 23
7J	12.2 Homogeneidad de proporciones	24,25, 26
8V	Repaso.Quiz sorpresa, quizá hoy o tal vez después.	P4. ver 1 (pág. 262)
11L	Repaso	P4. ver 2
12Ma	Quiz con folleto de fómulas	
14J	Repaso	P4. ver 3
15V	Quiz con folleto de fómulas	
18L	Repaso	P4. ver 4
19Ma	Quiz con folleto de fómulas	
21J	Parcial IV con folleto de fórmulas.	
22V	ENTREGAR PROYECTO pág 297 . Repaso final	FINAL ver. 1, pág 299
25L	Repaso final	versión 1, cont.
26Ma	Repaso final	FINAL ver. 2
28J	Quiz: simulacro del final	
29V	Repaso final	versión 2, 3 y 4
Mayo 2L-6V	Repaso final	
Mayo 11-25	EXAMEN FINAL CON FOLLETO DE FÓRMULAS	

Índice general

1. Descriptiva **1**
- 1.1. Histogramas y tortas . . . 1
- 1.2. Datos ordinales . . . 2
- 1.3. Media y varianza de datos numéricos o cuantitativos . . . 4
- 1.4. Lista de datos cuantitativos, exactitud total . . . 9
- 1.5. Datos cuantitativos agrupados: exactitud total . . . 10
- 1.6. Datos cuantitativos agrupados: exactitud aproximada . . . 12
- 1.7. Marco global . . . 19
- 1.8. Conclusión y retrospectiva . . . 23

2. Probabilidad **25**
- 2.1. Introducción . . . 25
- 2.2. Distribuciones . . . 28
- 2.3. La binomial . . . 30
- 2.4. Binomiales a toda velocidad (Triángulo de Pascal) . . . 40
- 2.5. Conclusión . . . 45

3. Taller y parcial I **47**
- 3.1. Taller I . . . 47
- 3.2. Parcial I: con hoja de fórmulas hecha personalmente . . . 51

4. La normal **59**
- 4.1. Idealización poblacional de una tendencia central . . . 59
- 4.2. La Z = la normal estándar . . . 61
- 4.3. Por qué se necesita una tabla? . . . 65
- 4.4. La Z como prototipo . . . 66
- 4.5. El protocolo-z . . . 67
- 4.6. Conclusión . . . 73

5. Hipótesis **75**
- 5.1. Ciencias duras y blandas . . . 75
- 5.2. La noción de interesante . . . 76
- 5.3. Eventos interesantes sobre la normal . . . 78
- 5.4. Eventos interesantes en la binomial . . . 82
- 5.5. Aproximación de la binomial por la normal . . . 83
- 5.6. La significancia de un evento . . . 84
- 5.7. La Hipótesis nula . . . 85
- 5.8. Intervalos de confianza . . . 94
- 5.9. Fabricando decisiones . . . 99
- 5.10. Protocolos y experimentos . . . 101
- 5.11. Repetitividad y ruido . . . 101
- 5.12. Conclusión . . . 104

6. La media y el experimento básico **105**
6.1. El teorema del límite central: TLC . . . 107
6.2. El IC para la media . . . 111
6.3. Conclusión . . . 116

7. Taller y parcial II **117**
7.1. Taller II . . . 117
7.2. Parcial II: con libro abierto . . . 126

8. Experimentos y estadígrafos de contraste **149**
8.1. Estudio de la media . . . 150
8.2. Estudio de una proporción . . . 155
8.3. Otros estadígrafos de contraste . . . 155
8.4. Tablas estadísticas . . . 157
8.5. Conclusión . . . 162

9. La ciencia se viste de gala **163**
9.1. Instrucciones para resolver problemas . . . 163
9.2. Ejemplos prototipo . . . 165
9.3. Conclusión . . . 189
9.4. la tabla de la z . . . 190

10.Taller y parcial III **191**
10.1. Ejercicios básicos . . . 191
10.2. III Parcial: con folleto de fórmulas . . . 199

11. Los 10 experimentos básicos **207**
11.1. Introducción . . . 207
11.2. La lista de los 10 . . . 207
11.3. ¿Cómo se decide un experimento? . . . 208
11.4. Notación . . . 209
11.5. El diccionario de experimentos . . . 209

12.Árboles y tablas de contingencia **239**
12.1. Árboles y probabilidades condicionales . . . 239
12.2. Tablas de contingencia: Independencia de dos factores . . . 243

13.Parcial IV **253**
13.1. Taller IV . . . 253
13.2. Parcial IV: con folleto de fórmulas . . . 258

14.Talleres complementarios **267**
14.1. Taller 1: Excel pa'ya . . . 267
14.2. Taller 2: Estadística no paramétrica y el teorema de Chevichev . . . 273
14.3. Taller 3: Test de ajuste . . . 283
14.4. Taller 4: Gnumeric y OpenOffice. . . . 290
14.4.1. Un diamante por un dolar . . . 290
14.4.2. ¡Linux a la orden! . . . 291
14.4.3. Instalación de Gnumeric . . . 291
14.4.4. Las tablas de OpenOffice y Gnumeric . . . 292
14.4.5. Un ejemplo con Gnumeric . . . 293
14.5. Taller 5: Investigación. . . . 295

15.Proyecto **297**

16.Examen final: con folleto de fórmulas. **299**
16.1. Anexo: la media muestral de distribuciones NO normales . . . 308

Capítulo 1

Descriptiva

1 ***Objetivo*** *¿Cómo es el mundo? Veremos una clasificación de los datos en diversos tipos, sus representaciones gráficas, la moda, la mediana, la media, cuartiles, datos normales vs. datos extremos.*

1.1. Histogramas y tortas

Todo el mundo hace estadística a toda hora como lo vemos en el siguiente ejemplo.

2 ***Ejemplo*** *A Juana, una hermosa mujer de 1.42m de estatura, le gustan los hombres no muy altos porque, según ella, los hombres altos tienden a ser más seguros de sí mismos por lo que pueden aceptar los reclamos con facilidad pero el problema es que 'cuando les da la chiripiorca pueden matarla a una de un solo estrujón'. De sus compañeros, los más interesantes para Juana con sus alturas respectivas son:*

-Pedro: 1.70; Juan: 1.60; Alberto: 1.65; Abel: 1.60; Roberto: 1.60

Es inútil convencer a Juana de que hay estadística pues cada dato es algo único y especial para ella, y su lista es lo único que ella acepta. Sin embargo, las terceras personas raramente se interesan en la lista de todos los datos. Las presentaciones que agrupan los datos de manera natural son las preferidas, a saber **tablas y gráficas de frecuencias absolutas y/o relativas**. La lista anterior puede resumirse en la siguiente tabla de frecuencias:

Tabla de Frecuencias		
Estatura	Absoluta	Relativa
1.60	3	3/5 = 0.6
1.65	1	1/5= 0.2
1.70	1	1/5=0.2
Total	5	1

Esta tabla puede representarse por medio de gráficas. La representación gráfica de una tabla de frecuencias absolutas de datos cuantitativos se denomina **histograma** o **diagrama de barras**. La representación gráfica circular de una tabla de frecuencias relativas, de datos categóricos o cuantitativos, se denomina **torta**, la cual se acostumbra a dar en porcentajes.

El histograma correspondiente a la tabla de frecuencias absolutas inmediatamente anterior es el siguiente:

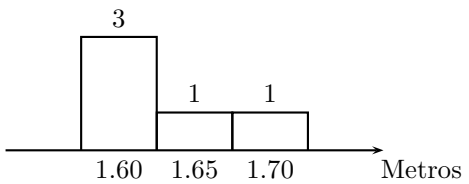


Figura 1.0. Histograma de la altura de la muestra R54.

La torta asociada a la tabla de frecuencias relativas es la siguiente:

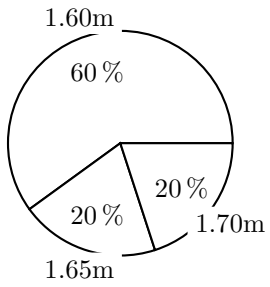


Figura 1.1. Torta que representa la distribución de alturas de la muestra R54.

3 ***Ejemplo*** *Cómo se calcula una torta.*

Para calcular una torta, se hallan las frecuencias relativas y se traducen a grados. Digamos, si la frecuencia relativa de la categoría C1 es 0,23 o 23 %, entonces la categoría C1 expande en la torta un ángulo de

$$\text{ángulo} = 23 \times \tfrac{360}{100}$$

De esa forma, si hay sólo una categoría con el 100 %, le corresponde toda la circunferencia que es de 360 grados.

4 ♣ ***Definición.*** *Los datos son* ***numéricos o cuantitativos*** *cuando se representan naturalmente por números y su promedio tiene sentido. Por ejemplo, la edad de un grupo de jóvenes. Los datos cuantitativos se representan naturalmente con un histograma sobre un eje de números reales, pues estos números son la esencia de lo que es cuantitativo. Los datos son* ***ordinales*** *cuando se pueden poner en línea de mayor a menor: ejemplo: la edad aproximada medida según: viejo, maduro, joven, niño. También se pueden graficar sobre un eje, los menores a la izquierda, los mayores a la derecha. En general, a los datos ordinales se les asocia números y se tratan como tales. Ej: a niño se le asocia 10 (años), a joven 25, a maduro 45, a viejo 70. Incluso uno puede atraverse a sacar medias pero sólo para tener una idea de su verdadero valor pues los resultados no son muy confiables. Los datos son* ***categóricos*** *cuando no tiene sentido promediarlos ni tampoco ordenarlos de mayor a menor. Ejemplo: los colores del cabello: negro, amarillo, castaño, rojo. Los datos estrictamente categóricos quedan bien representados por una torta y no quedan muy bien sobre un eje, pues uno diría que están ordenados.*

1.2. Datos ordinales

Los exámenes clasificatorios son muy comunes y a todos nos interesa saber cuál puesto obtuvimos y cosas por el estilo.

5 ***Ejemplo*** *Las notas numéricas en un examen generan datos ordinales pues uno siempre puede comparar dos notas para decir si son iguales o cuál es la mayor. Consideremos la siguiente tabla:*

Muchachos y sus notas	
Nombre	**Puntaje en Examen**
Camila Fernández	45
Pablo Rodríguez	63
Juana Maldonado	39
Mónica Bernal	44
Andrés Umaña	53
Juan Pérez	57
Diana Suárez	60
Pedro Hernández	37
María Puentes	47
Felipe Rueda	53

Vemos que la *moda*, el dato que más se repite, es 53.

Para clasificar unos datos ordinales, los ordenamos de menor a mayor:

Puntajes, rango y puesto			
Nombre	Nota	Rango	Puesto
Pedro	37	0	10
Juana	39	1	9
Mónica	44	2	8
Camila	45	3	7
María	47	4	6
Felipe	53	5	4
Andrés	53	5	4
Juan	57	7	3
Diana	60	8	2
Pablo Rodríguez	63	9	1

Cuando son tan pocos datos, como aquí, uno simplemente dice: Andrés es tan bueno como Felipe, pero Juan es mejor que ambos y todos ellos son mejores que Mónica. Aunque generalmente con eso es suficiente, de todas maneras a la gente competitiva (que somos casi todos) nos gusta declarar nuestro estatus muy oficialmente: Entre 10 Juan está por encima de 7. Formalizando, tenemos los conceptos de rango (ranking en una jerarquía) y puesto.

6 ♣ ***Definiciones.*** *El* ***rango*** *de un dato ordinal es el número de datos que son estrictamente menores que él. El* ***puesto*** *de un dato es el número de datos mejores que él más uno. Supongamos que los datos en una carrera son: llegó primero Juan, luego Luis que llegó empatado con Alfonso, pero Diego llegó de último. Tenemos: Juan llegó en el primer puesto (0 personas que llegaron antes que él + 1 = 1), Luis y Alfonso llegaron en el segundo puesto (1 persona antes que ellos +1 = 2), Diego llegó en el cuarto puesto (3 personas que llegaron antes que él +1 = 4). Análogamente, el rango de Diego es cero (quedó de último), el de Luis y Alfonso es uno, pues le ganaron a Diego, y el de Juan es 3, pues llegó antes que 3 personas. El puesto hace una comparación con los que van por encima, el rango con los que van por debajo.*

Una definición muy importante es la siguiente:

7 ***La mediana: es el dato de la mitad.*** *Para hallarla, se ordenan en fila todos los datos de menor a mayor. Si el número de datos es impar, la mediana queda en el dato de la mitad. Por ejemplo, si los datos ya ordenados son 1, 2, 3, 4 y 5, la mediana es 3. Si el número de datos es par, se toma el promedio de los dos centrales: si los datos son 1, 2, 3, y 4, la mediana es 2.5. La mediana es el referente natural de comparación para datos ordinales: ¿Cómo le fue a 2? En ambos casos le fue mal pues quedó por debajo de la mediana.*

8 ***Problema del reemplazo de toda la población por un sólo dato****: al tomar decisiones, a veces uno opta por la estrategia de tomar un dato como el representante de toda la población. Puede hacerse por capricho o puede hacerse usando algún criterio, como buscar la mejor representatividad. En este caso, una receta para tener en cuenta es: si los datos son categóricos y no pueden compararse, como los colores del pelo, tome la* ***moda*** *o sea el evento que más se repite: en un país nordico, píntese el cabello de rubio, pero en uno del trópico, de negro. Si los datos son ordinales y pueden compararse de menor a mayor, tome la* ***mediana****, que es el dato de la mitad después de ordenar. Si los datos son cuantitativos y se pueden promediar, tome la* ***media*** *que es el promedio. Para hallar la media usamos los procedimientos señalados más abajo. No se considera muy inteligente reemplazar toda la población por un único dato y parar ahí: lo menos que hay que hacer es estimar el descontento que eso puede causar. Para datos cuantitativos, el descontento se mide por la varianza y la desviación. Para datos ordinales el descontento se mide por el* ***rango intercuartilar*** *que lo veremos al final del capítulo.*

1.3. Media y varianza de datos numéricos o cuantitativos

9 ♣ ***Definiciones.*** *La* ***media*** *de un conjunto de n datos numéricos, para los que tiene sentido sumar y dividir, se halla sumando todo y dividiendo por el número de datos, n. Oficialmente: si tenemos una* ***lista*** *de datos numéricos X y a un dato lo notamos x_i, pues aparece en la lista en el lugar i, la* ***media o promedio*** *es*

$$\bar{X} = \frac{\Sigma x_i}{n} = \frac{\Sigma x}{n}$$

lo cual se lee: sume todos los datos y divida por el número de datos, n. Ejemplo: Sea X la lista de datos 1,2 3. La media es

$$\bar{X} = \frac{1+2+3}{3} = \frac{6}{3} = 2$$

La media es el número que mejor representa los datos si éstos valen por igual. La media debe dar entre el mínimo y el máximo de lo datos o sino hay un error de cálculo.

Varianza muestral s^2*: mide la variabilidad de los datos, entre más variados, más grande es la varianza. Desde el punto de vista de la toma de decisiones, la varianza mide el descontento de los datos de una muestra al ser representados por la media (se divide por n en estadística descriptiva y por $n-1$ en estadística inferencial, como es nuestro caso)*

$$s^2 = \frac{\Sigma(x_i-\bar{X})^2}{n-1} = \frac{\Sigma(x-\bar{X})^2}{n-1}$$

Lo cual se lee: sume los descontentos cuadráticos $(x-\bar{X})^2$, y divida por $n-1$. Ejemplo para X 1,2,3 y media 2:

$$s^2 = \frac{(1-2)^2+(2-2)^2+(3-2)^2}{2} = \frac{1+0+1}{2} = 1$$

La ***desviación estándar muestral*** *o simplemente la* ***desviación*** *s es la raíz de la varianza:*

$$s = \sqrt{\frac{\Sigma(x_i-\bar{X})^2}{n-1}} = \sqrt{\frac{\Sigma(x-\bar{X})^2}{n-1}}$$

Para los datos 1,2 3, la desviación vale 1.

Tanto la desviación como la varianza son medidas del descontento de los datos al representárseles por la media. Son índices de variabilidad de la población: entre más grandes, la población es más heterogénea, más inhomogénea con respecto a poblaciones que tienen la misma media. La desviación viene en las mismas unidades de los datos por lo que para gráficas es preferible a la varianza. Ejemplo: si los datos se refieren a la altura de unos muchachos y vienen en centímetros, la varianza viene en centímetros al cuadrado pero la desviación viene en centímetros y se puede dibujar en la gráfica indicando una suerte de distancia promedio entre los datos y la media.

Coeficiente de variación*: mide la homogeneidad de los datos con respecto a la magnitud promedio que sirve de unidad de medida.*

$$C = \frac{s}{\overline{X}}$$

Para los datos X 1,2,3, con desviación 1 y media 2, $C = \frac{1}{2} = 0{,}5$.

(En los textos, el coeficiente de variación se da generalmente en porcentaje).

10 ***Homogeneidad y coeficiente de variación para datos positivos.*** *Si el* ***coeficiente de variación*** *es del orden de 0.01, la población puede tomarse como homogénea y ser representada por la media. Si es de 0.1 como más o menos homogénea y sabremos que la media causará protestas. Pero si el coeficiente de variación supera 0.5, la población no tiene nada de homogénea y no debe pretenderse que la media represente la población. Esta regla puede ser anulada por una mejor información sobre el contexto del problema.*

11 ***La media no es suficiente en la toma de decisiones:***

La media es el representante democrático de toda la lista de datos cuantitativos, y por eso podemos preguntarnos: ¿desde cuando para acá un representante democrático representa sin error, sin insatisfacción, a todos sus representados? Pues claro que eso nunca pasa a no ser que se trate de clones. Por eso, una vez que tenemos la media, debemos inventar un índice, un **estadígrafo o estadístico**, la varianza o la desviación, que nos diga qué tan bien o qué tan mal es la representación de la población por la media. Distingamos tres casos:

1) Si la población es muy homogénea, la media será un representante fiel de toda la población.

2) Si la población es más o menos homogénea, la media estará sesgada a favor de los que se le parecen y en contra de los que no.

3) Si la población es muy inhomogénea, lo mejor es que no sea representada por la media.

Todo esto se ve reflejado en la toma de decisiones. Pensemos en la tarea de vestir a un grupo de personas: en cuanto a altura del cuerpo, los indígenas nativos son muy homogéneos. Por lo tanto, un vendedor de camisas puede comprar a muy bajo precio grandes cantidades de una misma talla S, que para ellos representa la media, e ir a la selva y cambiarlos por canastos, flechas y plumas, haciendo un gran negocio. Por otro lado, los soldados campesinos no son tan homogéneos, pero en caso de emergencia, una misma talla les sirve, aunque evidentemente, unos se verán mejor que otros. Pero si se trata de vestir a la población en general, decidir vestirlos con la misma talla sería lo mismo que quebrar el negocio.

En el ejemplo de vestir una población, la media es el representante básico de los datos, la cual se complementa con la varianza, un medidor del descontento de dicha representación. Pero hay otros casos en los cuales la varianza es lo que más importa. Proponemos dos ejemplos ilustrativos. El primero es el de los grandes aguaceros: en pluviometría se hace un registro de la cantidad de lluvia que cae por cada hora por cada decímetro cuadrado. Uno estima la variabilidad de los datos colectados calculando la varianza. En general, el valor medio en una región semiseca es muy moderado. Y, si la varianza es pequeña, el clima es semiseco en todo tiempo. Pero los grandes desastres ocurren cuando cae un chaparón que dura 35 minutos, que crea grandes ríos que no duran más de hora y media y que se llevan medio mundo por delante. Los grandes chaparrones no pesan mucho sobre la media pero sí pueden hacer crecer la varianza de manera significativa: la varianza es la que se asocia al peligro.

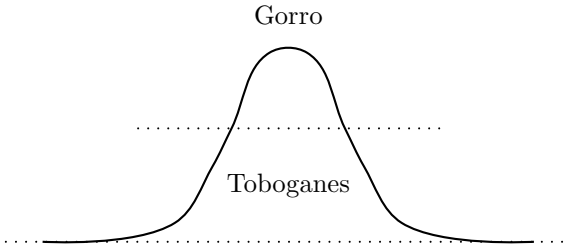
Otro ejemplo. Según dice Karla, su novio, Alfonso, es muy tierno pero el problema es que por largos tiempos se desaparece y la deja viuda y para tratar de remediar el error cuando se aparece se comporta como un chupete. Karla ha pensado terminarle, pues ya varias veces le ha dicho que no le interesa tener novio por e-mail, sino que lo quiere presencial. ¿Qué es lo que alegan estos novios en términos estadísticos? Según Alfonso, lo que importa es la media del tiempo que él le dedica a ella: las largas esperas las compensa con muy largas visitas. Pero según Karla, lo que causa el problema es la varianza tan alta del tiempo entre citas, el tiempo que él la deja esperando, el cual es muy inhomogéneo, pues toma valores desde minutos hasta de millones de años. Karla está desesperada pero Alfonso no puede entender y cree que está haciendo las cosas lo mejor que puede. Ellos no se gritan, pero Karla llora del desespero al ver que no puede terminarle. Y no sabe por qué.

12 ***Ejemplo.*** *Estimación gráfica de la desviación estándar.*

Ya sabemos cómo calcular con toda exactitud tanto la media como la desviación de un conjunto de datos numéricos. Sin embargo, en este mundo no todo es exactitud y por diversas razones uno necesita aprender a estimar dichos valores a vuelo de pájaro. Necesitamos valores que siendo aproximados sean de todas manera bastante buenos. La manera inteligente de hacerlo es usar el histograma de los datos para trazar una campana, la campana que mejor se ajuste a los datos, y usar dicha campana para hacer un estimativo tanto de la media como de la desvicación.

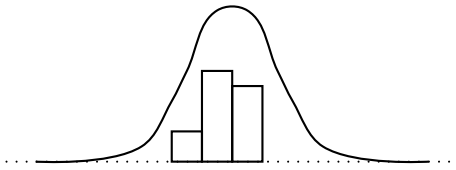
Aprendamos a trazar la **campana que mejor se ajuste a los datos**. A las campanas les decimos oficalmente **campanas de Gauss**.

Para dibujar una campana primero se dibuja un gorro y después se le añaden dos toboganes:

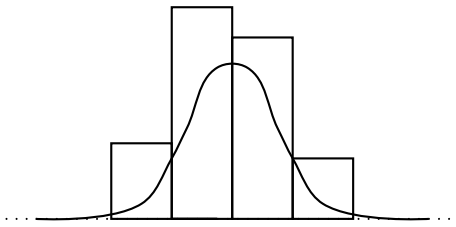


Una campana de Gauss es una idealización de datos que se agrupan como una montañita. Encontrar la campana que mejor idealice unos datos puede ser difícil. Algunos errores comunes son los siguientes:

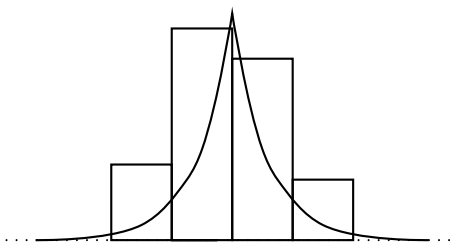
Mal: Campana muy por encima, como una cobija

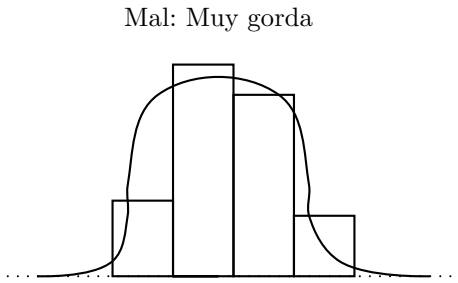


Mal: Muy por debajo, muy tacaña



Mal: Muy delgada





Una buena camapana es algo así:

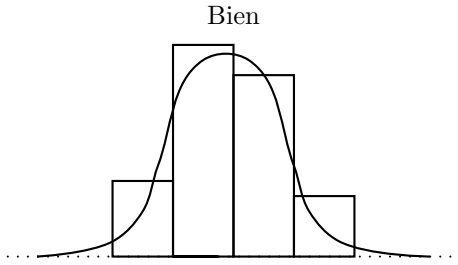


Figura 1.2. Una buena campana se ajusta bien a los datos, está bien centrada, no es ni alta, ni baja, ni gorda, ni flaca, ni achatada, ni empinada. Es como un hilo de cirugía, procurando compensar lo que queda por fuera con lo que queda por dentro.

Estudiémos la gráfica siguiente para aprender a **estimar la media y la desviación de unos datos cuantitativos** usando la campana de mejor ajuste:

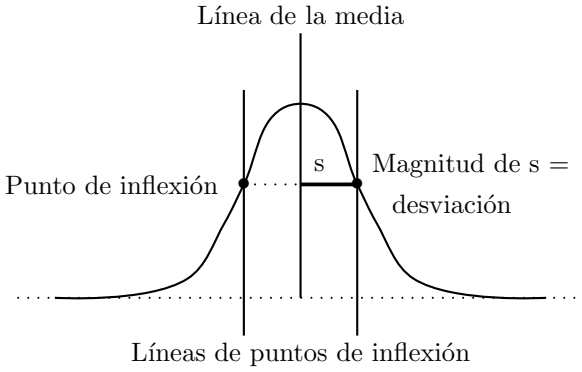


Figura 1.3. El eje de simetría de la campana determina la línea de la media. Hay dos puntos en donde la parábola se choca con los toboganes: son los puntos de inflexión. Ellos determinan las líneas de puntos de inflexión. La distancia de la línea de la media a cualquiera de las dos líneas de inflexión es la desviación. La media da el promedio y la desviación mide la dispersión de los datos: entre más grande, más heterogéneos. Para hallar la varianza se eleva la desviación al cuadrado.

13 ***Ejemplo*** *La campana de mejor ajuste a unos datos tiene la silueta siguiente: ¿Cuánto vale la media y cuánto la desviación?*

La media es el eje de simetría vertical de la campana.

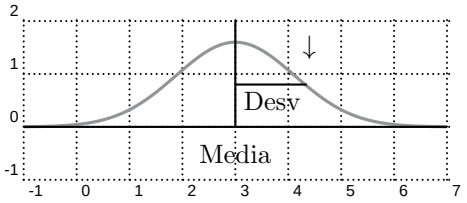


Figura 1.4. La media de una campana es el eje de simetría: la media es 3. Un punto de inflexión es donde el gorro se choca con un tobogán: el superior está en 4.4. La desviación es la distancia entre el punto de inflexión y la media: 4.4-3= 1.4. Su varianza es entonces $1{,}4^2 = 1{,}96$.

Para determinar la desviación de una campana, primero se halla el punto de inflexión. Para ello, se descompone la campana en un gorro hacia abajo y dos toboganes. El punto de inflexión es aquel donde el gorro se choca con uno de los toboganes. En la gráfica el punto de inflexión superior está indicado por una flechita: ahí la campana no abre ni hacia arriba ni hacia abajo. Por ejemplo, tenemos que la curva abre hacia abajo en 3 y en sus cercanías. Pero en 5 y en sus cercanías, la curva abre hacia arriba. Lo mismo pasa en 1 y sus cercanías: la curva abre hacia arriba. Pero en 4.4 uno no sabe hacia a dónde abre la curva: tenemos ahí un punto de inflexión. Lo mismo pasa en 1.6: es otro punto de inflexión. Hay dos puntos de inflexión, pero ponemos la atención en el que queda al lado de arriba. 4.4.

Luego que uno tiene el punto de inflexión, la desviación s es igual a la distancia horizontal entre la media y el punto de inflexión $s = 4{,}4 - 3 = 1{,}4$.

14 ***Ejemplo*** *Estimemos visualmente la* ***varianza*** *de los datos X 1,2,3,4,5, con media 3.*

El histograma es el siguiente:

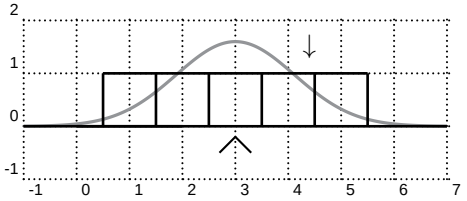


Figura 1.5. Los datos 1,2,3,4,5 tienen media 3. El punto de inflexión es 4.4. Por lo tanto, la desviación es 4.4-3= 1.4. Su varianza es entonces $1{,}4^2 = 1{,}96$*, i.e. 2.*

15 ***Ejemplo*** *La cadena de datos aleatorios X dada por 3, 4, 3,6,4,5,7,6,4,5,8,3 tiene media 4.83. Calculemos visualmente su desviación y varianza. Hallemos también los* ***datos extremos*** *que son por definición aquellos que quedan alejados de la media más de 2 desviaciones. Para el histograma contamos cuántas veces aparece cada número. Por ejemplo, hay 3 treces:*

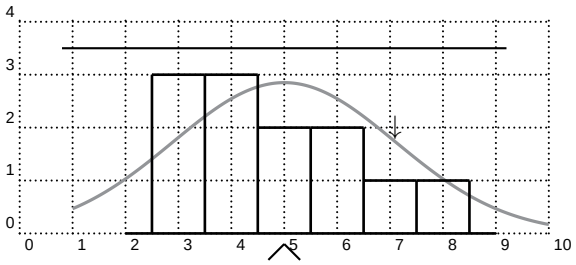


Figura 1.6. Media = 5. Punto de inflexión = 7.1. Desviación = 7.1-5 = 2.1 y la varianza 4.41. Según el criterio de las dos desviaciones, los datos extremos son los menores que 0.8 ó mayores que 9.2.

Imponemos la campana que mejor se ajuste a nuestro histograma: que no esté corrida ni a la derecha ni a la izquierda, que no quede ni alta ni baja y que la parte del histograma que queda como exceso sobre la campana se compense con lo que queda como defecto, lo que le falta al histograma para llegar a la campana. Usemos la capana para estimar gráficamente la media y la varianza. La media es 5 que es el eje de simetría de la campana. El punto de inflexión es donde la parábola que abre hacia abajo se

choca con los toboganes: es 7.1. La desviación es la distancia entre el punto de inflexión y la media. La desviación nos da 7.1 -5 = 2.1 y por ende la varianza es 4.41.

Habiendo hallado visualmente la desviación, podemos hallar los datos extremos según el criterio de las dos desviaciones. Para obtener las barreras que determinan los datos extremos sumamos y restamos a la media dos veces la desviación. Tenemos $5 \pm 4{,}2$ que da 0.8 como el extremo inferior y 9.2 como superior. Lo que queda entre 0.8 y 9.2 se considera normal y se ha señalado con una barra horizontal encima de la campana. Pero, todo lo que quede abajo de 0.8 es extremo y todo lo que quede arriba de 9.2 es extremo. Revisando los datos, vemos que en la muestra tomada no hay datos extremos. Un dato extremo llama la atención y puede exigir decisiones especiales. Por ejemplo a los profesores que clasifican como extremos inferiores en la encuesta a los estudiantes se les podría poner en período de prueba y se les hace un seguimiento asesorado. Si repiten, podrían no tener renovación de contrato.

1.4. Lista de datos cuantitativos, exactitud total

Presentamos 5 métodos para calcular la varianza, dos de ellos enseguida y los otros tres en la sección siguiente. Con excepción del quinto todos son exactos.

16 ***Ejemplo: primer método.*** *Hallemos la media y la varianza de la notas 5, 3, 4.5 y 2 usando la definición.*

Elaboramos la tabla siguiente:

Cálculo de la varianza de unas notas		
Materia	$x = Nota$	$(x - X)^2$
Estadística	5	$(5 - 3{,}6)^2 = 1{,}96$
Economía	3	$(3 - 3{,}6)^2 = 0{,}36$
Teatro	4.5	$(4{,}5 - 3{,}6)^2 = 0{,}81$
Contabilidad	2	$(2 - 3{,}6)^2 = 2{,}56$
Total	$\sum x = 14{,}5$	$\sum (x - X)^2 = 5{,}69$

Para hacer esta tabla, primero se computa la suma de la $x's$, en la segunda columna: da 14,5. Por tanto, la media es la suma de todo entre el número de datos que se denota n. Da $14{,}5/4 = 3{,}6$. Sabiendo la media, 3,6, podemos calcular la segunda columna. Y después la varianza:
$s^2 = \frac{\sum (x - \bar{X})^2}{n-1} = \frac{5{,}69}{3} = 1{,}896$
Por tanto, $s = \sqrt{1{,}896} = 1{,}37$.

La varianza se calcula más fácilmente con ayuda del siguiente teorema:

17 ◇ ***Teorema para el segundo método:*** *La varianza muestral de una lista de n datos es igual a*

$$s^2 = \frac{n \sum x^2 - (\sum x)^2}{n(n-1)}$$

El símbolo $\sum x^2$ significa: eleve cada dato al cuadrado y sume todo. El símbolo $(\sum x)^2$ significa: sume los datos y el resultado elévelo al cuadrado. La mejor manera de usar este teorema es por medio de una tabla, como sigue:

Segundo método para los datos 5, 3, 4.5 y 2: Nos ayudamos de la siguiente tabla

Cálculo de la varianza de unas notas		
Materia	$x = Nota$	$x^2 = Nota^2$
Estadística	5	25
Economía	3	9
Teatro	4.5	20.25
Contabilidad	2	4
Total	$\sum x = 14{,}5$	$\sum x^2 = 58{,}25$

Por lo tanto, la media es
$\bar{X} = \frac{14,5}{4} = 3,6$
La varianza es
$s^2 = \frac{n\sum x^2 - (\sum x)^2}{n(n-1)}$
$s^2 = \frac{4(58,25) - (14,5)^2}{4(3)}$
$s^2 = \frac{233 - 210,25}{12} = \frac{22,75}{12} = 1,896$
La desviación estándar es
$s = \sqrt{1},896 = 1,37.$

El **coeficiente de variación**: es

$C = \frac{1,37}{3,6} = 0,38$

Como C es cercano a 0.5 concluimos que los datos son muy inhomogéneos y que de ninguna manera conviene representarlos por la media. Para entenderlo, pensemos en el estudiante de la tabla que lleva su promedio en 3.6. Basándonos en dicho promedio, podemos decir que el estudiante es regular. Pero si no añadimos nada más, estaríamos juzgándolo apresuradamente, pues estaríamos pasando por alto que es buenísimo en estadística en tanto que va muy mal en contabilidad. Teniendo todo en cuenta, especialmente su cinco en estadística, podemos desafiarlo a ser excelente en todo, pues tiene con qué. Esta consejería no se basa en la media únicamente, sino que usa al tiempo media y varianza,

El coeficiente de variación nos podría servir para detectar **outliers**, elementos que están alejados de la media de una manera extrema. Cuando los hay, la toma de decisiones puede ser un poco rebuscada:

18 ***Ejemplo: la estabilidad de la mediana.*** *Consideremos los datos -1, 0, 1 que corresponden a las ganancias en tres operaciones de negocios. Su media es cero y por experiencia sabemos que no hay razón para designar a alguno de los datos como outlier, como extremo. Pero si añadimos el dato 40, que vino de un negocio extra con el Gobierno y que sería un dato extremo, un outlier, los datos quedan -1, 0, 1, 40. La nueva media es 10 y nos damos cuenta que un outlier causó un corrimiento tremendo en la estimación de la media. Y lo dicho de la media también es válido para la varianza. Ahora bien, de las dos estimaciones, cero y diez, ¿cuál es más confiable para armar planes a futuro? La más confiable es la primera porque se basa en el cuerpo de la distribución, en lo usual, en aquella parte donde los datos se agrupan, mientras que la segunda se deja arrastrar por un oultier que salió por azar y que seguramente no saldrá en próximas ocasiones. Una manera más oficial de considerar el problema es tomando en cuenta la mediana. La mediana de los datos -1, 0, 1 es 0. Y la de -1, 0, 1, 40 es 0.5 que es el promedio de los datos 0 y 1. Inferimos que la mediana es bastante estable ante la presencia de un único outlier o cuando su proporción es mínima con respecto a los datos. Es por dicha estabilidad que la mediana es muy útil en la toma de decisiones.*

1.5. Datos cuantitativos agrupados: exactitud total

Cuando hay una larga lista de datos cuantitativos que se repiten, uno puede reemplazar la lista por una **tabla de frecuencias absolutas** y usar dicha tabla para los cálculos de la media y varianza.

19 ***Ejemplo, tercero y cuarto métodos.*** *Un grupo de 40 estudiantes toma diverso número de créditos, pero se da la casualidad de que el número de créditos puede ser 4, 7, 10, 13, 16, 19. La tabla de frecuencias absolutas es como sigue:*

Distribución recarga académica	
Créditos	# estudiantes
4	3
7	6
10	10
13	15
16	4
19	2
Total	40

Si tenemos una tabla con m renglones y marcas de clase x_i y frecuencias absolutas F_i, para calcular la media se procede así:

Paso 1: Se calcula el número de datos

$$n = \sum F = F_1 + F_2 + ... + F_m.$$

Paso 2: Se calcula la suma ponderada de todos los términos:

$$\Sigma xF = x_1F_1 + x_2F_2 + ... + x_nF_m$$

Paso 3: La media es:

$$\bar{X} = \frac{\Sigma(xF)}{n} = \frac{\Sigma(xF)}{\Sigma F} = \frac{\Sigma(x_iF_i)}{\Sigma F_i} = \frac{(x_1F_1)+....+(x_nF_m)}{F_1+F_2+...+F_n}$$

El **tercer método** para calcular la varianza es apropiado para cálculos a mano:

Paso 1: Se calcula la suma ponderada de todos los cuadrados:

$$\Sigma x^2F = x_1^2F_1 + x_2^2F_2 + ... + x_n^2F_m$$

Paso 2: Se calcula el término *Suma de los cuadrados x x*:

$$S_{xx} = \Sigma x^2F - \frac{(\Sigma xF)^2}{n}$$

Paso 3: La varianza es

$$s^2 = \frac{S_{xx}}{n-1}$$

Precaución: la varianza es una suma de cuadrados, por tanto debe ser positiva o hay error numérico.

Paso 4: La desviación s es la raíz cuadrada de la varianza:

$$s = \sqrt{\frac{S_{xx}}{n-1}}$$

Para calcular la media y la varianza de los créditos tomados por los estudiantes, podemos hacer uso de una tabla adecuada que usa el tercer método:

Cálculo de la varianza			
$x =$ *créditos*	F = frecuencia	xF	x^2F
4	3	12	48
7	6	42	294
10	10	100	1000
13	15	195	2535
16	4	64	1024
19	2	38	722
Total	$n = \sum F = 40$	$\Sigma xF = 451$	$\Sigma x^2F = 5623$

Ahora podemos calcular la media:

$\bar{X} = \frac{\Sigma(xF)}{n} = \frac{451}{40} = 11{,}27$

El termino S_{xx}:

$$S_{xx} = \Sigma x^2 F - \frac{(\Sigma xF)^2}{n} = 5623 - \frac{451^2}{40} = 5623 - \frac{203401}{40} = 5623 - 5085{,}025 = 537{,}975$$

$$S_{xx} = 537{,}975$$

Por lo tanto, la varianza muestral es

$$s^2 = \frac{S_{xx}}{n-1} = \frac{537{,}975}{39} = 13{,}79$$

y la desviación estándar es

$$s = \sqrt{13{,}79} = 3{,}71$$

El **cuarto método** para calcular la varianza mezcla los métodos uno y tres y es el siguiente:

$s^2 = \frac{\Sigma(F_i(x_i - \bar{X})^2)}{n-1} = \frac{\Sigma(F(x-\bar{X})^2)}{n-1}$, donde se usan las frecuencias absolutas F y n es la suma de las frecuencias, es decir el número de datos. Interpretar este método y aplicarlo queda como ejercicio.

El **quinto método** para calcular la varianza es inmediato si partimos del método 4:

$s^2 = \frac{\Sigma(F(x-\bar{X})^2)}{n-1} = \Sigma(\frac{F}{n-1}(x-\bar{X})^2)$

Lo cual quiere decir que una suma dividida por un número es igual a la suma de los números divididos por el mismo número. Ahora, si n es grande, podemos reemplazar $n-1$ por n:

$s^2 = \Sigma(\frac{F}{n}(x-\bar{X})^2)$

Pero como la frecuencia absoluta dividida por el número de elementos es igual a la frecuencia relativa, $\frac{F}{n} = f$, nos queda

$s^2 = \Sigma(f(x-\bar{X})^2) = \Sigma((x-\bar{X})^2)f = \Sigma((x_i-\bar{X})^2)f_i$

Aunque no haremos ejercicios en este capítulo para este quinto método, en el capítulo siguiente será el único método a usar pero en la forma de probabilidades: cuando n es grande, la frecuencia relativa se puede usar con confianza para tomar decisiones pero se dice **probabilidad** en vez de frecuencia relativa. Ejemplo: de 200 estudiantes de 8 secciones, 40 perdieron el primer parcial, lo cual da una frecuencia relativa de $\frac{40}{200} = \frac{1}{5} = 0{,}2$. Para planificar el siguiente semestre podemos decir: la probabilidad de perder es de 0.2 ó de 20 %. Al hablar de probabilidades estamos enfatizando toma de decisiones que afectan a toda la población, estamos imponiendo una verdad, estamos infiriendo de una muestra a toda la población. Se formaliza cambiando la notación: $\bar{X}$, s^2 y s denotan respectivamente la media, varianza y desviación de una muestra. Si la muesta es grande, podemos utilizar los valores como estimación de la media, varianza y desviación de la población, las cuales se denotan con letras griegas: μ (mu), σ^2 (sigma cuadrado), σ (sigma).

1.6. Datos cuantitativos agrupados: exactitud aproximada

Cuando hay muchos datos que son diversos, como las alturas de un grupo de personas, o pocos datos que están muy dispersos, como el ingreso de 20 familias de estrato alto, uno tiene la opción de redondearlos o agruparlos por rangos, intervalos o clases. Redondear o agrupar disminuye la exactitud

pero permite ver la tendencia, por ejemplo si es central (cuando los datos se agrupan alrededor de la media), o si es hacia el borde inferior, o si es bimodal (que tiene dos jorobas, es decir, con dos modas bien definidas). Una vez que uno tiene los intervalos, uno decide cuál va a ser el representante o **marca de clase** de cada intervalo.

20 ***¿Cuántas clases debo escoger?*** *Para decidir en cuántas clases, rangos o intervalos se van a agrupar los datos tenemos varios criterios. El primero es el de la comodidad, por ejemplo: las notas de un curso van de 0 a 5 y queda bien dividirla en 5 clases, de 0 a 1, de 1 a 2 y así sucesivamente. El criterio del 7: se saca el valor mínimo y el máximo. La diferencia entre los dos se divide en 7 partes iguales que van a representar las clases del muestreo, o quizá uno prefiera dividir no por 7 sino por un número entre 5 y 15. El criterio de la abundancia dice: entre más datos haya, uno puede hacer más subdivisones pero no tantas que varias clases queden pobremente representadas. El criterio académico: el número de clases es proporcional al logaritmo del número de datos. Una vez que uno tiene los intervalos, como representante o* ***marca de clase*** *puede tomarse el punto medio de cada clase o intervalo. Estas recetas básicas se modifican según el objetivo personal.*

21 ***¿Cómo se eligen las marcas de clase?*** *Tenemos varios criterios. El de la comodidad dice: busque algo cómodo, digamos el extremo derecho del intervalo, o el izquierdo, o el centro. Tenga en cuenta que estamos en un método aproximado, por tanto no se fatigue buscando exactitud inoportuna. El criterio del contexto dice: lea bien el contexto y los objetivos administrativos y entonces tome la marca de clase por sentido común. Por ejemplo: una profesora quiere motivar a sus pequeños estudiantes a estudiar matemáticas, entonces durante las primeras lecciones les pone la nota máxima a todo lo que sea prometedor y una nota intermedia a todo lo demás.*

22 ***Ejemplo*** *En general, los créditos que 40 estudiantes toman cada semestre pueden tomar valores muy diversos. Una forma de redondear contiene los siguientes intervalos de créditos: si alguien toma entre 3 y 5 créditos, redondeamos a 4. O dicho en términos oficiales: al rango de 3 a 5 créditos le damos la marca de clase 4. Similarmente al intervalo* $6-8$ *le corresponde 7, al* $9-11$ *le corresponde 10, al* $12-14$ *le corresponde 13, al* $15-17$ *le corresponde 16, al* $18-20$ *le corresponde 19. A continuación se hace la tabla de frecuencia absoluta de muchachos que toman determinados créditos por semestre.*

Una vez que los datos están agrupados, se designan las marcas de clase que son números que reemplazan toda la clase, como por ejemplo el punto medio de la clase. Pero según el criterio personal, uno puede tomar el borde inferior, o el superior o incluso alguien por fuera de la clase. Se hace un tabla de frecuencias absolutas, con la cual uno puede dibujar el histograma. Para hallar la media y la varianza aproximadas de los datos, a la tabla de frecuencias se le aplica el mismo procedimiento que en el caso anterior.

23 ***Ejemplo*** *Tenemos los datos siguientes que corresponden al peso en gramos de unas muestras al azar de orfebrería.*

48.5, 49.2, 51.0, 50.9, 48.7, 50.5, 49.5, 50.8, 50.0, 51.1,
50.7, 51.1, 49.3, 49.1, 50.0, 48.7, 51.2, 49.2, 49.0, 49.3,
49.1, 48.7, 48.6, 49.2, 49.9, 50.1, 50.1, 50.9, 52.4, 50.2,
49.5, 50.9, 51.4, 49.7, 49.8, 50.8, 50.3, 51.8, 50.0, 51.1,
51.3, 50.1, 50.4, 51.1, 49.8, 49.8, 50.1, 50.7, 50.1, 49.9,
49.7, 51.1, 49.7, 49.9, 49.8, 50.5, 49.7, 50.8, 50.7, 50.7,
50.8, 51.5, 51.0

La tarea es primero agruparlos en clases definidas por los intervalos con bordes siguientes: 48.5, 49.5, 50.5, 51.5, 52.5. Después hay que hacer la tabla de frecuencias absolutas y calcularle su media y su varianza. Los intervalos son cerrados por abajo y abiertos por arriba. Así se evita doble contabilidad. Para indicarlo, se pone un signo menos como exponente al final de la notación del intervalo, como en $48{,}5-49{,}5^{-}$*, eso quiere decir que si el dato 49.5 llega a aparecer, y para no contabilizarlo dos veces, una vez en este intervalo y otra en el siguiente, no se cuenta en el intervalo* $48{,}5-49{,}5^{-}$ *(pues de ahí se quitó), sino se cuenta dentro del intervalo siguiente,* $49{,}5-50{,}5^{-}$*. La marca de clase se denota* x*.*

El agrupamiento designado produce la siguiente tabla de frecuencias absolutas que se complementa para poder sacar la media y la varianza:

Peso en gramos				
Intervalos	x	F	xF	x^2F
$48{,}5 - 49{,}5^-$	49	13	637	31213
$49{,}5 - 50{,}5^-$	50	24	1200	60000
$50{,}5 - 51{,}5^-$	51	23	1173	59823
$51{,}5 - 52{,}5^-$	52	3	156	8112
Total		$n = \sum F = 63$	$\sum xF = 3166$	$\sum x^2F = 159148$

$\bar{X} = \frac{\Sigma xF}{\sum F}$

$\bar{X} = \frac{3166}{63} = 50{,}25$

Calculamos S_{xx}:

$$S_{xx} = \Sigma x^2F - \frac{(\Sigma xF)^2}{n} = 159148 - \frac{3166^2}{63} = 159148 - \frac{10023556}{63} = 159148 - 159104 = 43{,}9$$

Con S_{xx} podemos calcular la varianza:

$$s^2 = \frac{S_{xx}}{n-1} = \frac{43{,}9}{62} = 0{,}70$$

y la desviación estándar es la raíz de la varianza:

$$s = \sqrt{0{,}70} = 0{,}8366$$

El coeficiente de variación C es

$C = \frac{s}{\bar{X}} = \frac{0{,}8366}{50{,}25} = 0{,}016$

Como el valor de C es aproximadamente 0,01, la población se considera homogénea y uno puede tratar a los miembros de la población como si fuesen iguales a la media. Por lo tanto, uno puede hacer un reporte que diga: las piezas de orfebrería tenían un peso promedio de 50,25 con una ligera variación. Si alguien preguntara qué tan grande es esa variación, uno respondería que la desviación era menor a uno mientras que la media era 50 y es por eso que la variación se considera poca o ligera.

Los datos cuantitativos por ser numéricos también pueden ordenarse y al hacerlo permiten que se les aplique técnicas ordinales. Veamos:

24 ***Regla de tres para percentiles y cuartiles.***

Todos nosotros hemos usado expresiones como "le gané al 97 %". En Estadística se dice informalmente: "Soy el percentil 97" (percentil = dividir en 100 partes iguales y calcular el porcentaje de los que están por debajo). Formalmente, la definición de percentil la hacemos mediante el siguiente procedimiento:

Para **hallar el dato que corresponde al percentil k sobre n datos ya ordenados de menor a mayor** primero se calcula la localidad L_k que es un simple dato intermedio y está dado por

$L_k = n \times \frac{k}{100}$.

Si la localidad L_k está dada por un fraccionario o decimal, se aproxima por arriba a la posición más cercana y el dato correspondiente es el percentil k. **Si L_k está dada por un entero**, el dato correspondiente al percentil k es el promedio entre las posiciones L_k y la siguiente. Ejemplo:

Los datos son los primeros 8 números pares 2, 4, 6, 8,... 16 y queremos hallar el dato que corresponde al percentil 47. Tenemos que $n = 8$. La localidad está dada por

$L_{47} = 8 \times \frac{47}{100} = 3{,}76$.

Como obtuvimos un fraccionario o decimal, aproximamos por arriba a 4. Miramos quién está en la posición 4: está el 8 y concluímos: el percentil 47 es 8. (Intriga: el 8 también es el percentil 48 y 49. Preguntamos: ¿Cómo deberá definirse el percentil asignado a 4?)

Calculemos ahora al percentil 25. La localidad está dada por

$L_{25} = 8 \times \frac{25}{100} = 2.$

Como obtuvimos un entero, promediamos los datos que están en esa posición y la siguiente: en la posición 2 está un cuatro y en la 3 un seis. El promedio de 4 y 6 es 5. El percentil 25 es 5.

De los percentiles los más famosos son los cuartiles (dividir en 4 partes iguales y calcular el porcentaje de las que están por debajo):

El **cuartil 1**, Q_1, es el percentil 25 que, como ya sabemos, es 5. La **mediana** es el **cuartil 2**, Q_2, o el percentil 50 cuya localidad está dada por:

$L_{50} = 8 \times \frac{50}{100} = 4.$

Como obtuvimos un entero, promediamos los datos que están en esa posición y la siguiente. Es decir, la mediana, cuartil 2, Q_2, o percentil 50 es el promedio de 8 y 10 que da 9, conforme esperamos por sentido común, pues abajo del 9 hay tantos datos como arriba de él.

El **cuartil 3**, Q_3, sería el percentil 75 cuya localidad está dada por:

$L_{75} = 8 \times \frac{75}{100} = 6.$

Como obtuvimos un entero, promediamos los datos que están en esa posición y la siguiente. En la posición 6 hay un 12 y el siguiente dato es 14. El promedio es 13, así que $Q_3 = 13$. Para terminar, el cuartil 4, Q_4, sería el percentil 100 que siempre se interpreta como el valor máximo, en la última posición y que corresponde a un 16.

Veamos un ejemplo en el cual se aproxima siempre hacia arriba. Si los datos son 11 y **ya ordendos de menor mayor** son:

1,2,2,2,3,3,3,3,3.5,4,5.

donde 3.5 representa un único dato. Tenemos que el percentil 17 se calcula así:

$L_{17} = 11 \times \frac{17}{100} = 1{,}87$ que por ser fraccionario se aproxima a 2. En dicha posición está un 2. Ese sería el percentil 17.

El **cuartil 1**, Q_1, es el percentil 25 cuya posición está dada por:

$L_{25} = 11 \times \frac{25}{100} = 2{,}75 \to 3.$

En la posición 3 está un 2: ese es el cuartil 1: $Q_1 = 2$. La **mediana** es el **cuartil 2**, Q_2, o el percentil 50 cuya posición está dada por:

$L_{50} = 11 \times \frac{50}{100} = 5{,}5 \to 6.$

En la posición 6 está un 3, ese sería la mediana: $Q_2 = 3$. El **cuartil 3**, Q_3, sería el percentil 75 cuya posición está dada por:

$L_{75} = 11 \times \frac{75}{100} = 8{,}25 \rightarrow 9$.

En la posición 9 está un 3.5, ese sería el cuartil 3: $Q_3 = 3{,}5$. Para terminar, el cuartil 4, Q_4, sería el percentil 100 que sería la última posición, la 11, y que corresponde a un 5.

Veamos otro ejemplo en el cual la localidad simpre viene de número entero. Para ello, añadimos un 5 a los datos del ejemplo anterior obteniendo $n = 12$ y la muestra ya ordenada quedaría:

1,2,2,2,3,3,3,3,3.5,4,5,5.

Los cuartiles serían

$Q_1 \rightarrow L_{25} = 12 * 0{,}25 = 3 \rightarrow$ Promedio de datos 3 y 4 = (2+2)/2 = 2 = Q_1;
$Q_2 \rightarrow L_{50} = 12 * 0{,}50 = 6 \rightarrow$ Promedio de datos 6 y 7 = (3+3)/2 = 3 = Q_2;
$Q_3 \rightarrow L_{75} = 12 * 0{,}75 = 9 \rightarrow$ Promedio de datos 9 y 10 = (3.5+4)/2 =3.75 = Q_3;

25 ***Ejemplo: diagrama de tallo y hojas*** *Presentamos un método para reportar y analizar de forma aproximada datos cuantitativos que muestra cómo uno puede añadirle creatividad a sus presentaciones. La trampa es tener en cuenta que los datos cuantitativos también se pueden ordenar y por tanto podemos hablar de percentiles y cuartiles. Analicemos los siguiente datos que corresponden a las notas en un examen sobre cuidado del bebé presentado por unos futuros papás, cuyas esposas esperan el primer hijo. Se comparan dos grupos, uno de ingenieros y otro de antropólogos.*

Las notas del grupo de papás ingenieros son:

1.2 2.3 3.2 3.3 1.7
2.1 2.2 3.2 3.3 4.5
1.9 2.1 2.4 2.5 3.4
2.2 3.2 1.5 3.2 2.3
Las notas del grupo de papás antropólogos son:

2.3 3.8 3.3 4.5 4.8
3.8 4.1 3.9 4.6 3.4
2.1 3.4 4.5 3.2 4.5

Para hacer el diagrama de tallo y hojas, primero se dividen los datos en intervalos apropiados y se ordenan los datos. Como son datos entre 0 y 5, dividimos los datos en 5 clases, de 0 a 1, de 1 a 2,..., de 4 a 5.

Las notas del grupo de papás ingenieros son:
4.5
3.2 3.2 3.2 3.2 3.3 3.3 3.4
2.1 2.1 2.2 2.2 2.3 2.3 2.4 2.5
1.2 1.5 1.7 1.9

Las notas del grupo de papás antropólogos son:

4.1 4.5 4.5 4.5 4.6 4.8
3.2 3.3 3.4 3.4 3.8 3.8 3.9
2.1 2.3

Después se ponen todos los datos en un árbol en donde se observa un ordenamiento de menor a mayor desde abajo hacia arriba y desde el tronco o centro del árbol hacia afuera de él. Para hacerlo con elegancia, los datos no se copian sino que se presentan mediante la convención natural siguiente: si sobre el tallo se indica el rango 3, el dato 2 sobre dicho nivel se interpreta como 3.2 y así con todos.

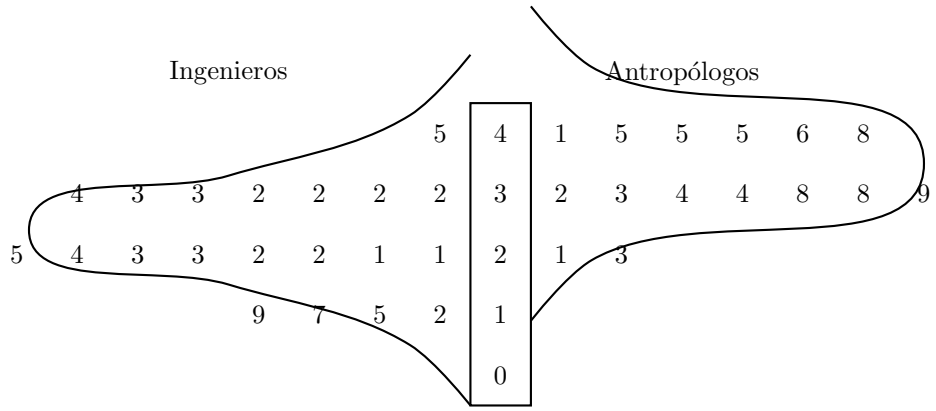


Figura 1.7. Diagrama de tallo (con los intervalos) y hojas(con los datos) que representa las notas en un examen sobre cuidado del bebé presentado por dos grupos de padres cuyas esposas esperan el primer hijo.

Los siguientes puntos pueden enriquecer la discusión:

1. Hay 15 datos del grupo de antropólogos. Su mediana es el octavo dato: es 3.8
2. Hay 20 notas del grupo de ingenieros. Su mediana está entre 2.4 y 2.3, digamos en 2.35.
3. Trazamos las campanas, las cuales no deben quedar corridas, ni a un lado ni al otro. Las campanas no se pintan como cobijas, por encima de todo, sino como hilos de cirugía, unos tramos por dentro y otros por fuera, de forma que lo que quede por fuera equilibre lo queda por dentro. Vemos que la media (la coordenada x del eje de simetría y distinta de la mediana) de la nota de los ingenieros queda como en 2.4, mientras que la de los antropólogos en aproximadamente 3.3.
4. La desviación para ambos grupos de notas podría ser 0.8. La desviación es la distancia entre el eje de simetría, la media, y el punto de inflexión, donde la campana cambia de curvatura.
5. De los antropólogos hay 6 de 15 que obtuvieron una nota mayor que 4, mientras que 19 de 20 ingenieros obtuvieron una nota inferior a 4. Es decir, El 40 % de los antropólogos tiene mejor puntaje que el 95 % de los ingenieros.
6. Hay 4 ingenieros de 20 (20 %) que tienen menor nota que el antropólogo que sacó la menor nota.

26 ***Ejemplo: Utilizando nuestra definición de percentil y cuartil discutamos la relación entre ingenieros y antropólogos del ejemplo anterior.*** *Sea k un número entre 0 y 100. Para hallar el percentil k calculamos*

$L_k = n \times \frac{k}{100}$.

Si L_k es un entero, la posición del percentil k es el promedio entre las posiciones L_k y la siguiente. Si L_k es fraccionario, se aproxima por arriba a la posición más cercana. Apliquemos esto:

Para los ingenieros se tiene:

$Q_1 \to L_{25} = 20 * 0{,}25 = 5 \to$ *Promedio entre datos 5 y 6 = (2.1+2.1)/2 = 2.1;*
$Q_2 \to L_{50} = 20 * 0{,}50 = 10 \to$ *Promedio entre datos 10 y 11= (2.3+2.4)/2 = 2.35;*
$Q_3 \to L_{75} = 20 * 0{,}75 = 15 \to$ *Promedio entre datos 15 y 16 = (3.2+3.2)/2 =3.2;*

Para los antropólogos:

$Q_1 \to L_{25} = 15 * 0{,}25 = 3{,}75 \to$ *Dato 4 = 3.3;*
$Q_2 \to L_{50} = 15 * 0{,}50 = 7{,}5 \to$ *Dato 8 = 3.8;*
$Q_3 \to L_{75} = 15 * 0{,}75 = 11{,}25 \to$ *Dato 12 = 4.5;*

Vemos que el tercer cuartil de los ingenieros es menor que el primer cuartil de los antropólogos. Esto significa que el 75 % de los ingenieros está por debajo de Q_1, es decir, por debajo del 75 % de los antropológos.

Para sacar de manera aproximada los cuartiles en un diagrama de tallo y hojas se procede así: uno va recorriendo los datos del menor hasta el mayor. Uno divide la lista de datos en dos partes iguales. La división es la mediana. La primera mitad se divide en dos y así se localiza el primer cuartil. El tercer cuartil se halla dividiendo la segunda mitad en dos (esta metodología aproximada puede mejorarse a gusto personal).

27 ***El boxplot, diagrama de caja o gráfica de cuadro o diagrama de bigotes***

¿Cómo se sintetizan los datos cuantitativos de manera óptima? Si se trata de una **distribución de tendencia central** (con datos que se agrupan alrededor de la media), usamos la media y la desviación. Pero si se trata de resolver la pregunta para distribuciones de cualquier tipo, centrales o no, una salida que es muy aceptada es tener en cuenta que los datos cuantitavos también son ordinales y podemos pensar en términos de mediana y cuartiles, los cuales son mucho más generales. Veamos.

28 ***EL resumen de los 5 valores y el boxplot*** *es el conjunto conformado por el valor mínimo, el primer cuartil, la mediana, el tercer cuartil y el valor máximo. Es una síntesis ideal de los datos. Genéricamente, llamamos* ***estadígrafo*** *a los valores que se calculan a partir de una muestra de datos. Como ejemplos tenemos a la media, desviación, coeficiente de variación, mediana, cuartiles, mínimo, máximo.*

Existe una forma gráfica de representar los 5 valores y se denomina **boxplot** o **diagrama de caja** o **gráfica de cuadro** o **diagrama de bigotes**. Este diagrama consiste en una línea recta que se dibuja, por ejemplo, encima del histograma, y que empieza en el mínimo, termina en el máximo y tiene una caja no muy gruesa entre el primer cuartil y el tercero y con una línea vertical que denota la mediana. Si la distribución es simétrica, la caja será simétrica alrededor de la mediana. Pero cuando la distribución es asimétrica, la caja también lo es: si se explaya hacia arriba ó hacia abajo, la caja hace lo mismo.

29 ***Ejemplo.*** *Usando* ***regla de tres****, hallemos el resumen de 5 valores y dibujemos el boxplot para las calificaciones en un examen: 1, 3, 2, 4, 3, 2, 3, 2, 3.5, 5, 3.*

Solución: Ordenamos los datos de menor a mayor: 1,2,2,2,3,3,3,3,3.5,4,5. Hay 11 datos. Por el ejemplo 24, en la página 14, ya sabemos que los cuartiles son : 2, 3 y 3.5. Como el mínimo valor es 1 y el máximo 5, el boxplot nos quedaría así:

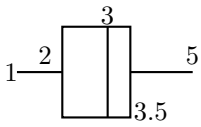


Figura 1.9. Boxplot que muestra el mínimo, el primer cuartil, la mediana, el tercer cuartil y el máximo de unas calificaciones en un examen. Obsérvese que las posiciones respetan la escala.

La caja dice que la distribución está explayada hacia abajo y muy recogida hacia arriba. El nivel de dificultad del examen fue diseñado para que pueda pasar quien haya hecho un esfuerzo realizable pero que quien quiera sacar una nota alta demuestre perfecto dominio de la temática.

En **Gnumeric,** una alternativa profesional y gratis a Excel y que se busca por su nombre en Internet, se pueden hallar los 5 valores con *insert, functions, statistics, quartile.*. En la casilla *quartnumber* uno pone uno de los números 0, 1, 2, 3, 4. Para indicar los datos, se hace click en *array Area* y se negrea la región que los contiene. Con cero da el mínimo valor. Con 1 da el primer cuartil, con 2 da la mediana, con 3 el tercer cuartil y con 4 el valor máximo.

Solución usando la función *quartil* de **Gnumeric** :

Min = 1,
Primer cuartil: 2.5,
Segundo cuartil = mediana = 3,
Tercer cuartil = 3.25,
Máximo = 5.

Atención: cuando uno usa un paquete es necesario decir cuál es porque es casi seguro que las definiciones y procedimientros que éste utiliza son diferentes de los de los demás. En este caso, Gnumeric da respuestas distintas a las nuestras para los cuartiles 1 y 3.

30 ***El rango intercuartilar y el diagrama de bigotes que muestra los valores extremos***

Podemos utilizar el resumen de 5 datos para hallar los **eventos extremos** según el **criterio del rango intercuartilar**. Para fijar ideas, consideremos los siguientes datos que indican el número de cuadernos + libros que cargan unos estudiantes tomados al azar de grado 8:

4,5,6,6,6,8,10,10,10,10,10,11,11,12,18.

Su resumen de 5 datos según el **paquete R** (http://www.evoljava.com/dl/Refacil.pdf), es: 4 (mínimo), 6 (cuartil 1), 10 (mediana = cuartil 2), 10.5 (cuartil 3), 18 (Máximo).

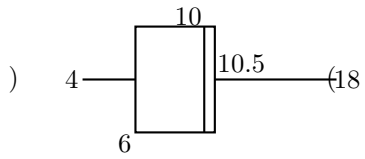


Figura 1.10. Diagrama de bigotes en el cual se indican con paréntesis las fronteras que separan los valores normales de los extremos. En este diagrama no hay valores extremos por abajo pero 18 lo es por arriba pues es mayor que 17.25.

El **rango intercuartilar** RI es por definición igual al cuartil tres menos el uno: $RI = Q_3 - Q_1 = 10{,}5 - 6 = 4{,}5$. A dicho valor lo multiplicamos por 1.5 (siempre se multiplica el rango intercuartilar por 1.5) para hallar la **aleta cuartilar**, la cual da 6.75. Para delimitar lo que es extremo por arriba sumamos la *aleta cuartilar* al tercer cuartil: 10.5+6.75 = 17.25. Lo que esté por encima de dicho valor es extremo del lado superior. En nuestro caso 18 es extremo por arriba. Para delimitar lo que es extremo por abajo restamos la **aleta cuartilar** del primer cuartil: 6-6.75 = -0.75. Lo que esté por debajo de dicho valor es extremo del lado inferior. No tenemos valores extremos por abajo.

Estos datos fueron utilizados por la Rectora para diagnosticar un exceso de peso sobre las espaldas de los niños y proponer a los padres de familia que mejor podrían pasarse a la tecnología de libros electrónicos y tablets.

1.7. Marco global

Un primer objetivo de la estadística es describir al mundo, lo cual implica observarlo, registrarlo y reportarlo. Pero el mundo es tan supremamente complicado que es preciso delimitar el rango y tipo de las observaciones. Eso sucede naturalmente si uno se fija un **proyecto**, que puede nacer de una simple pregunta, como por ejemplo: ¿Por qué soy más bajito que la mayoría? En general, todo proyecto implica la demarcación de una población que es el conjunto universal o conjunto de referencia, digamos, las personas de mi Universidad o de mi empresa. Las observaciones deben ser registradas. Debido a que hacer y registrar observaciones cuesta dinero y esfuerzo, nos restringimos a una muestra que debe ser seleccionada al azar. Formalicemos:

31 ♣ ***Definición.*** *El* ***investigador*** *es la persona que define el proyecto y quien toma las decisiones. La* ***población*** *o* ***universo*** *es el conjunto de referencia y está conformado por los objetos que se van a ver afectados por las decisiones tomadas. La* ***muestra*** *es el subconjunto sometido a estudio directo y sobre el cual se toman las decisiones. Una* ***muestra es al azar*** *cuando no existe ninguna forma de saber cuál será el próximo elemento a registrar. Una* ***variable*** *es una característica que se mide u observa en cada elemento de la muestra. Un* ***dato o evento*** *es un registro de una caracterísitca en una observación (preferiblemente simple), es el valor de la variable en un elemento de la muestra. Las* `decisiones` *afectan a toda la población pero se toman estudiando la muestra. Si la muestra queda mal hecha, las decisiones pueden resultar muy equivocadas. El primer requisito para que una muestra sea buena es que represente maximal y fielmente a la población, para lo cual es necesario que sea al azar. Una* ***v.a. (variable aleatoria)*** *es una característica que se mide en una muestra al azar. Una* ***v.a. es discreta*** *cuando los datos originados son todos enteros, 0,1,2,3,... y sus negativos. El número de hijos es una variable discreta. Una* ***v.a. es continua*** *cuando se necesitan cifras decimales de arbitraria precisión para reportar los datos. La altura de las personas es una variable continua. A veces se añaden* ***protocolos de muestreo y/o de medición*** *que son especificaciones para tomar la muestra y para hacer las mediciones u observaciones de tal forma que la representatividad y aleatoriedad de los datos quede a la vista. Con frecuencia se ignora toda esta sofisticación y se habla de población y muestra en referencia directa a los eventos o datos.*

Recalcamos que la elección del proyecto, del universo, de la muestra, de las variables a medir, y las mismas observaciones se hacen por seres humanos y por lo tanto tienen un valor subjetivo intrínseco: dos seres humanos nunca ven, desean o añoran lo mismo. No importa cuán automatizado sea el método de registrar las observaciones, el que lo automatizó lo hizo de acuerdo a valores y compromisos personales, es decir subjetivos. Por esta sola razón, la ciencia es una subcultura humana. Por lo tanto, su conexión con la **verdad** (la **realidad** tal cual es) es siempre discutible. Y precisamente eso la hace fuerte: la ciencia que tenemos hoy en día ha sido discutida hasta la saciedad por las generaciones pasadas y no se derrumba porque a alguien se le ocurra criticarla. Dicho claramente: si Usted tiene algo que criticarle a la ciencia y desea ser oído, más le vale que se vuelva un reconocido, respetado y apreciado experto.

Una vez que uno tiene un proyecto y reúne datos, hay que analizarlos. Pero eso se hace secuencialmente, pues en general es algo complicado. El primer paso es reportar los datos para que las tendencias básicas se hagan visibles. El paso más crucial de todos es decidir cuáles datos son aceptables y cuáles son rechazados, cuáles datos son muy importantes y cuáles casi despreciables. A veces eso se hace por criterios simples: si uno al medir la temperatura de un niño encuentra que es 12 grados centígrados, uno inmediatamente sabe que ese dato es falso. Pero a veces eso involucra toda una inercia filosófica. Por ejemplo, la televisión reporta de cuando en cuando el registro de cámaras de seguridad que muestran almas en pena y esos reportes vienen lo mismo que de un castillo inglés, que de un viejo hotel en los Estados Unidos o que de una estación eléctrica en Pasto. Y sin embargo, los físicos no se dan por aludidos. ¿Por qué? Porque para ellos y para la ciencia moderna sólo existe lo que se puede registrar y manipular a voluntad. Oficialmente se dice que la realidad no se ve directamente sino sólo a través de cosmovisiones. ¿Qué es entonces una **cosmovisión**? Es una ideología auto sostenida o inercia de procedimiento que organiza el registro de datos decidiendo su relevancia y pertinencia para la construcción de la verdad en el círculo social dado.

32 ***Ejemplo*** *¿Por qué trabajamos? Trabajamos, en general, para ganar el pan de cada día, el cual comemos con afán y ansiedad para poder tener fuerzas para poder trabajar más. Eso implica que las cosas que tienen directa pertinencia con mi trabajo son los generadores de datos de máxima importancia para mí. En cambio, los datos relacionados con su trabajo, perdone Usted, casi no me interesan. Esta ideología es auto sostenida por razones biológicas, por supervivencia y puede polarizar más del 60 % de la vida de un ser humano.*

33 ***Estadística y decisiones*** *Pocas veces tiene uno tanto tiempo y tantos recursos como para hacer un censo de una población grande. La estadística nos ayuda a tomar decisiones basándonos en los datos de una muestra al azar, algo que se obtiene a mucho menor costo.*

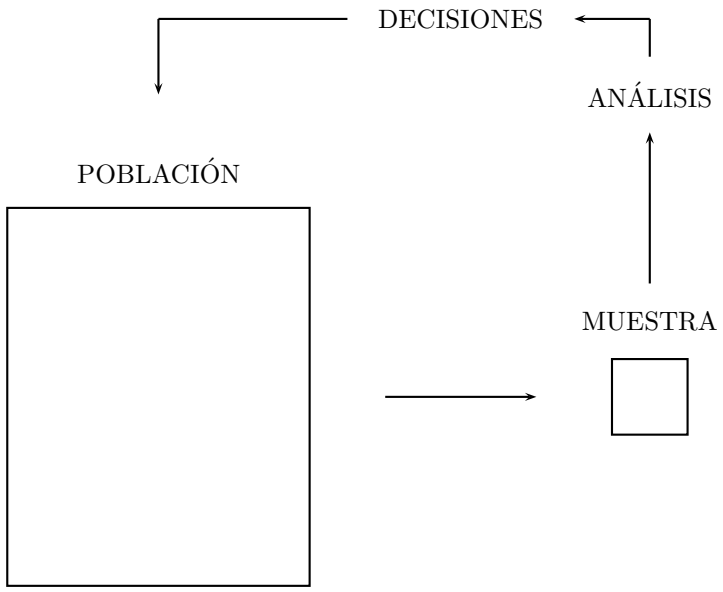


Figura 1.10. El objetivo central de la estadística es disminuir el costo de la toma de decisiones gracias a que se estudian los datos de una muestra al azar y no de un censo.

Donde hay muestreo, hay incertidumbre, donde hay censo no hay incertidumbre. En estadística tomamos decisiones bajo incertidumbre pero pagamos un precio en equivocaciones, un precio que eventualemente puede ser tolerable.

Veamos cómo se implementa este programa. Estaremos dedicados en buena parte de nuestro curso a estudiar decisiones basándonos en el **promedio**, el cual es el número que él solo es la mejor representación de todos los datos, y en la **varianza**, una medida de su dispersión o variablidad y que mide el descontento global de ser representado por la media. Por tanto, seguimos el siguiente algoritmo o procedimiento:

1. Hacemos un muestreo al azar, sacamos la media y la varianza de la muestra. A esta parte se la denomina **estadística descriptiva**.

2. Puesto que las decisiones afectan a toda la población y no sólo a la muestra, lo que debemos hacer es estimar cuáles podrían ser los valores de la media y varianza de la población. En definitiva, los valores poblacionales son iguales a los de la muestra pero a los que se le concede un margen de error. A este paso se le llama **estadística inferencial**.

3. Sobre los valores poblacionales y teniendo en cuenta la incertidumbre debida al muestreo, se toman las decisiones apropiadas.

34 ***Ejemplo*** *Dora le pone una cita a un pretendiente a las 3 en punto. Ella llega faltando 10 para las 3, pues a ella le interesa saber qué tan puntual es el joven que tanto dice quererla. a) ¿Cuál es la decisión a tomar, el investigador (el que toma la decisión), la población, la muestra, la variable a medir, los eventos? b) ¿Qué tan racional es la actitud de Dora? c) La variable a medir: ¿Es aleatoria? ¿Es discreta? ¿Es continua? d) ¿Cuál es la decisión deseable? e) ¿Cómo se toma la decisión de aceptarle una nueva cita?*

a) La decisión a tomar por Dora es si el pretendiente es o no confiable. La investigadora es Dora pues ella es quien toma la decisión. La muestra es el pretendiente. La variable a medir es la hora de llegada del pretendiente a su primera cita. El evento o dato a registrar es la hora de llegada del pretendiente en el día de hoy. La población incluye a todos los pretendientes en todas sus primeras citas.

b) La hora de llegada del pretendiente a la primera cita denota su puntualidad que da un índice de confiabilidad. Tener una relación con una persona confiable es algo maravilloso. En caso contrario hay que estar dispuesto a aguantarse continuamente toda serie de desplantes. Por lo tanto, Dora está muy bien orientada y sabe lo que quiere.

c) La variable a medir es continua, pues el tiempo puede medirse con una precisión muy grande que puede requerir muchos decimales. Aunque la muestra no es al azar, pues es el pretendiente, su hora de llegada sí lo es pues hay demasiadas perturbaciones incontroladas que afectan la hora de llegada a una cita.

d) Es aconsejable descartar sin derecho a opinar a los pretendientes que llegen tarde a su primera cita pues es muy probable que sean no confiables en el resto de compromisos. Pero ¿qué es llegar tarde? Como regla estadística, todos los pretendientes (el conjunto universal o población) llegan muy puntuales a su primera cita, así que Dora ha decidido que llegar despúes de la 3 y 10 es llegar tarde.

e) En general, el procedimiento para tomar decisiones es el siguiente: el valor de la muestra (llegar tarde, llegar a tiempo) se contrasta con el valor esperado según los objetivos fijados (debe llegar a tiempo) teniendo en cuenta el margen de error aceptable sacado de la población. Si hay una gran discrepancia entre la muestra y lo que se cree, se considera que los datos rechazan lo que se cree, pero si la discrepancia es pequeña en relación a la población, se tolera y se decide que los objetivos se han cumplido. En la vida diaria, medir la discrepancia entre los valores de la muestra y lo supuesto depende de cada quien. En nuestro ejemplo, Dora declara retardado a quien la haga esperar más de 10 minutos. Nosotros veremos formalismos estándares que involucran al azar que resulta de factores no controlados como que un accidente produjo un enorme trancón.

35 ***Ejemplo: retirarse o no retirarse.*** *En una universidad se permite retirarse de una materia si N, la nota acumulada hasta la fecha límite de retiros, es muy baja y no se desea correr el riesgo de perderla. Veamos de qué manera nos ayuda la estadística en nuestra toma de decisiones.*

Nota	Resultado final		
	X: Pasó	Y: No pasó	Totales
Alta: $N \geq 3{,}4$	21	0	21
Mediana: $3{,}0 \leq N \leq 3{,}3$	7	3	10
Baja: $2{,}1 \leq N \leq 2{,}9$	2	3	5
Totales	30	6	36

Analicemos esta tabla:

1. La muestra corresponde a estudiantes de una clase determinada (caso real de una clase en el 2012 en la que vieron estadística con este libro. Hubo además 3 personas con menos de 2 que se retiraron).

2. Lo que queremos es saber qué consejo darle a un estudiante cuando se enfrente a la decisión de retirarse o no. Por tanto, la población se refiere a los estudiantes potenciales que puedan tomar el mismo curso en un futuro.

3. Hay dos variables a tener en cuenta. La primera es la nota equivalente acumulada hasta la fecha límite de retiros. Dicha nota es una variable continua y va de 0 a 5 y es la nota que cada estudiante sacaría si se terminara ahí mismo el semestre. Se pasa con 3. La segunda variable es si pasó o no. Pasar puede codificarse como uno y no pasar como cero, por lo tanto es una variable discreta.

4. Como se trata de una clase específica puede alegarse que la muestra no es al azar. Con todo, podemos conceder que los estudiantes que se inscribieron en dicha clase (una de 6) son al azar y que por tanto representan fielmente la población total. Por consiguiente, el análisis de la tabla sí es de ayuda en consejería.

5. Tenemos que decidir si la tabla indica o no que la nota equivalente tiene algo que ver con el destino definitivo, pasar o no pasar. Miramos la tabla y nos inclinamos en favor de la tesis que dice que

la tabla evidencia una relación entre la nota equivalente y la condición final. Defendemos nuestro punto de vista como sigue. Primero: Una nota equivalente alta es garantía absoluta de pasar: entre 21 personas con nota alta, el 100 % pasaron. De los que tenían nota mediana, pasaron el 70 % entre 10 personas. Y de 5 personas con nota baja, el 40 % pasaron. Teniendo en cuenta que cada categoría está bien representada exceptuando quizá la última, concluimos que entre mejor sea la nota equivalente, más alta es la probabilidad de pasar.

6. Pero en definitiva, ¿Cuál es el consejo práctico que uno debe dar a los estudiantes? Proponemos lo siguiente. Si no se puede correr el riesgo de perder la materia, lo mejor es que se retire si la nota equivalente no es alta. Si se tiene una nota alta y se decide quedarse hay que seguir en la misma tónica de disciplina diaria y de esforzarse por entender y relacionar. Si se puede correr un riesgo moderado, uno puede quedarse si la nota no es baja. Si se puede correr un riesgo alto, uno puede quedarse a pesar de que su nota sea baja pero si ya se entendió en que se está fallando y si se está seguro de poder mejorarlo.

Veamos un ejemplo de una tabla hipotética que no nos sirve para consejería:

Nota	Resultado final		
	X: Pasó	Y: No pasó	Totales
Alta: $N \geq 3{,}4$	15	6	21
Mediana: $3{,}0 \leq N \leq 3{,}3$	7	3	10
Baja: $2{,}1 \leq N \leq 2{,}9$	3	2	5
Totales	25	11	36

Esta tabla no sirve para el propósito para la cual fue creada pues lo que ella dice es que 2/3 de los estudiantes de la clase pasan sin importar la nota equivalente. Eso es posible en un curso cortacabezas, en el cual todo el mundo lleva nota equivalente en la fecha de retiros cerca de 3 (la nota mínima para pasar) y por tanto, dependen crucial y definitivamente en lo que pase en la parte final del semestre y sobre todo en el examen final que es acumulativo. Con todo, tenemos un consejo: prepárese bien para cada instante como si fuese el único de todo el curso.

Nuestro objetivo global para el curso es aprender a formular problemas como éste, resolverlos y tomar decisiones según los estándares de la modernidad, en los cuales el azar se tiene en cuenta formal y explícitamente.

1.8. Conclusión y retrospectiva

El método científico propone una forma de conocer y de analizar tanto lo que vemos como lo que pensamos del mundo que observamos. El método comienza invitando a elegir un tema de estudio, a formular un proyecto, a definir un universo que nos servirá de referente. En estadística, al conjunto universal se le llama **población**. En seguida, la primera pregunta que este método le hace al investigador es: ¿Qué ve Usted? El investigador responde: sobre los objetos de una **muestra**, subconjunto del universal, tomada al azar veo los siguientes datos para la variable X presentados en esta tabla...y representados en la siguientes gráficas (tortas e histogramas) que denuncian la tendencia siguiente (ej: central). Pero para poder hacer una gráfica adecuada y seguir con el análisis de datos, lo primero que uno puede hacer es decidir con qué tipo de datos se trabaja:

Si son datos **cualitativos o categóricos**, como el color de una camiseta de un solo fondo (blanca, negra, amarilla,..), uno puede hallar la tabla de **frecuencias tanto absolutas como relativas** y hacer la representación gráfica por medio de una **torta**. Se puede sacar la **moda**, que es el dato que más se repite.

Si se trata de datos **ordinales**, que dan una clasificación, como el orden de llegada a la meta, uno los ordena, sacando **el rango, los percentiles, la mediana**. El primer **decil** corresponde al percentil 10, el primer **cuartil** corresponde al percentil 25, el tercer **cuartil** corresponde al percentil 75: decimos

que Juan está por encima del tercer cuartil cuando está más arriba que el 75 % de los participantes. La **mediana** es el segundo cuartil o el percentil 50.

Si se trata de datos **cuantitativos**, como el peso en gramos de una muestra de joyas precolombinas en oro, a los cuales se les puede sacar la **media, la varianza, la desviación y el coeficiente de variación**, lo primero que hay que decidir es si uno desea resultados exactos o resultados aproximados y a continuación situarse en el procedimiento correspondiente. ¿Cómo justificamos la aproximación de datos? Sucede que en muchas ocasiones la toma de decisiones no cambia aunque los datos cambien un poquito: si el peso promedio durante el último semestre de dos jóvenes de la misma altura y contextura física es 68 y 69Kgs, lo más seguro es que compren pantalones de la misma talla. Esa tolerancia permite simplificar el trabajo usando métodos aproximados. Al respecto, vimos dos métodos: uno gráfico y el otro por agrupamiento de datos, los cuales se complementan.

Si uno desea hallar la media y desviación de una lista de datos **cuantitativos**, tal y como vienen, uno sigue el ejemplo 17 de la pág. 9. Si los datos vienen en una tabla de frecuencias absolutas pero sin aproximar, uno utiliza la metodología del numeral 19 de la pág. 10. Si uno desea aproximar por intervalos, se utiliza lo visto en el ejemplo 23 de la pág. 13.

Usamos histogramas preferiblemente para reportar tablas de frecuencias absolutas de datos cuantitativos o numéricos y tortas para reportar frecuencias relativas de cualquier tipo de datos, pero especialmente de datos categóricos.

La moda, la mediana y la media son **estadígrafos** que reemplazan un conjunto de datos por un único número. Desde el punto de vista del método científico, dichos estadígrafos responden a la pregunta siguiente: ¿Podría Usted sintetizar de manera óptima sus datos en un único número? Y después, ¿Cómo complementaría Usted de manera óptima la información dada por dicho estadígrafo? Para datos cuantitativos, la respuesta es: como la media es el representante democrático de los datos, lo que hay que añadir es una medida del descontento que dicha representación causa entre los datos. Eso se mide con la varianza o la desviación estándar o el coeficiente de variación.

Tenemos dos criterios para determinar los eventos extremos de datos cuantitativos. Si una distribución es de tendencia central -cuyo histograma se puede aproximar bien por una montaña con un único gran pico- podemos aproximar el histograma por una campana, estimar su desviación y declarar como extremo todo lo que esté a más de dos desviaciones alejado de la media. Pero si tenemos dificultad en aceptar que los datos se pueden aproximar por una campana, podemos usar el criterio del rango intercuartilar que es $Q_3 - Q_1$. Si multiplicamos dicho valor por 1.5 tenemos la aleta intercuartilar. Si sumamos a Q_3 dicha aleta obtenemos la frontera de los datos extremos por arriba. Y, si restamos de Q_1 dicha aleta obtenemos la frontera correspondiente por abajo.

La estadística es en cierto sentido el brazo armado de la ciencia. La ética en relación con el manejo de las armas comienza con el siguiente consejo: no desenfunde sus armas a no ser que esté seguro de necesitar usarlas. Parece que este consejo es muy beneficioso para todos los amantes de la estadística y puede ser usado muy sabiamente. Por ejemplo: los análisis estadísticos indican que las facilidades del manejo computacional han aumentado con los años y que por ende la edad en la que los niños comienzan a jugar con los computadores disminuye y ellos pasan cada vez más tiempo entretenidos con esos aparatos. En un reporte al gran público, toda la serie de diagramas estadísticos fue reemplazada por una única gráfica en la cual se veía en primer plano la espalda de un portatil abierto detrás de la cual se alcanzaba a ver la punta de la cabeza de un ser muy pequeño que sostenía en sus piernas al portátil y del cual apenas sobresalían los pies en sus tenis de niño. Por supuesto que la inclusión de una gráfica estadística hubiese quedado como una redundancia.

Capítulo 2

Probabilidad

36 ***GRAN OBJETIVO PARA EL RESTO DEL SEMESTRE: Teniendo una descripción básica del mundo, uno no puede evitar preguntarse por qué es como es. Acá aparece la ciencia diciendo que es vano responder a esa pregunta a no ser que uno pueda corroborarla, haciendo experimentos o registrando más observaciones. Y aún así, uno no puede obtener una total seguridad de que sus conclusiones sean ciertas. Eso se debe a que para un número finito de datos, uno no puede presumir conocer la verdad, sino tan sólo proponer un cierto esquema de probabilidades. Por eso, nuestro propósito no puede ser otro que armar al estudiante de ciencias sociales con herramientas suficientes para que pueda ver, entender y sentir experimentalmente que la ciencia, toda la ciencia, es una artesanía, una depurada y hermosa artesanía, cuya conexión con la verdad es apenas tentativa.***

37 ***Objetivo del capítulo****: Introducir conceptos de probabilidad, variable aleatoria, muestreo al azar, distribución binomial (nuestro primer modelo o creencia matematizada). Cálculo de una Binomial, triángulo de Pascal.*

2.1. Introducción

El método científico, corroborar lo que uno cree, es innato en todo ser humano. Nada más miremos lo que hace una niña que anda por su época del realismo mágico: ella con su hermanito pequeño juegan con aviones de papel. Además de mantener a su hermanito felizmente ocupado, la niña se ha trazado un objetivo: hacer que su avión vuele lo más lejos posible. Y ella tiene una teoría: soplar al avión antes de lanzarlo lo capacita para que vuele lejos. Pero ella no considera que su teoría sea simplemente una creencia, sino que es algo que da poder para controlar el universo. Por eso, ella hace varios experimentos: si soplar sirve de algo, soplar un poco más debe hacer avanzar al avión un poco más. Y efectivamente, eso es lo que hace. Lamentablemente son tantas las variables que entran en juego que la niña pronto se cansará de hacer ciencia y no podrá este día darse cuenta de que su teoría es falsa.

Vemos pues que el método científico, corroborar lo que uno cree, es innato en todo ser humano. Pero se atribuye a Galileo el haberlo formalizado y convertido en la punta de lanza de una ideología que llegó a convertirse en lo que hoy se llama ciencia. Galileo trabajaba con objetos muy estables, como rodillos y planetas, y con leyes físicas muy poderosas, es decir, muy sencillas, como la gravedad y la óptica clásica. Su método le pareció tan obviamente necesario y tan extremadamente liberador que llegó a decir que cualquiera que tuviese ojos en la frente podría llegar con su método a la verdad. Una verdad contante y sonante muy en contraste con los agüeros y creencias por fe.

¿Cómo es posible que algo tan de todos, como lo es el método científico, sea considerado en muchos círculos como patrimonio exclusivo de la ciencia que por su majestuosidad nada parece tener que ver con el vulgo?

En realidad, debemos reconocer que la física, la química y aún la biología han logrado adelantos asombrosos. Y que, por contraste, la gente común y corriente nos seguimos muriendo de cualquier mal. ¿Qué es lo que pasa? Lo que pasa es que cualquiera puede esquivar una piedra si se la lanzan a más de 7 pasos, pero ¿quién podrá esquivar una lluvia de asteroides? Lo que queremos decir es que la ciencia exitosa lo es porque puede trabajar con mundos simplificados, pero en cambio, la vida real no para de retarlo a uno de millones de formas. Y sin embargo, uno logra liberarse de muchos mitos: bastantes niños prueban a sus 5 años la teoría del soplo sobre sus aviones pero para los 15 años ya casi nadie lo intenta.

Si uno por instinto puede lograr grandes cosas en esta vida tan complicada, ¿podrá lograr algo mejor si arma al instinto de una metodología altamente depurada? Nuestra creencia es que sí, pero no es obvio, ni fácil, ni barato, ni libre de errores. Con todo, hemos enfrentado en este trabajo el reto de presentar esta temática de forma verdaderamente entendible a los estudiantes de ciencias sociales.

Lo primero que haremos es desmentir la falsa impresión de que la complejidad de un sistema implica incapacidad total para conocerlo y controlarlo. Porque si somos un poco más humildes que Galileo podremos ganar mucho: una descripción adecuada de un sistema complejo nos puede permitir conocer y controlar un sistema no en su totalidad pero sí de forma que puede ser relativamente satisfactoria. Desistir de un control absoluto es lo mismo que desistir de la verdad absoluta. Por lo tanto, de entrada debemos conformarnos con verdades parciales, relativas. ¿Cómo se implementa eso en ciencia? Pues en vez de decir que mañana saldrá el sol, una verdad absoluta para los que no saben de colisiones con estrellas de neutrones, decimos que la probabilidad de cara al lanzar una moneda es de 1/2. Y, ¿qué ganamos con eso? Podemos predecir que si lanzamos muchas monedas, aproximadamente la mitad saldrán cara. Pero, ¿qué tiene que ver eso con mi vida práctica como científico social? Podemos decir que muchos tipos de eventos en la vida son semejantes a lanzar una moneda.

Por ejemplo, una persona que llega a pedir empleo quizá sea un futuro ladrón o quizá no. Pero sabemos que las mamás cabeza de familia prefieren no robar para así gozar de una estabilidad que garantice la supervivencia de sus bebés, pero con todo hay excepciones: decimos que más del 90 % de las mamás cabeza de familia cuidan la empresa donde laboran (dando por descontado un bueno y justo trato) o bien que la probabilidad de que una madre cabeza de familia sea ladrona es inferior a 0.1. Entonces, ¿cuál es una decisión sabia? Muchos opinan que además de dar buen trato a la empleada, hay que ayudarla a sentirse vigilada, y por eso ordenan requisas por rutina, juntamente con observación continuada por circuito cerrado. Otros consideran que encontrar una empleada en la que se pueda confiar vale más que el 20 % del activo de la empresa y por eso no invierten mucho en esquemas de seguridad anti-fraude.

Definición: Una **probabilidad** es una frecuencia relativa en una población. Como se trata de una población, necesitamos un censo para averiguar una probabilidad. En general, los censos son muy costos. Por ello, tenemos otra definición operativa: una **probabilidad** es la idealización de una frecuencia relativa a partir de una muestra grande tomada al azar. Decimos que **un muestreo es al azar** cuando es imposible predecir cuál será el próximo dato. Por ejemplo, si de 1000 personas que pasan por una calle principal, 245 llevan bluejean, podemos decir que la probabilidad de que una persona vista esa prenda es 0.245. Algunos problemas permiten definir las probabilidades sin necesidad de censos ni de muestreos. Por ejemplo, si estamos haciendo un experimento que consiste en lanzar un dado, la probabilidad de cualquier cara es 1/6 pues idealizamos el dado a una figura simétrica por todo lado. Dicho de otra manera, si las frecuencias relativas de cada cara no tienden a aproxmarse a 1/6, el dado está mal hecho.

38 *Axiomas fundamentales*

Nos queda ahora fácil entender por qué la probabilidad debe tener tres propiedades básicas. Usamos el termino **evento** para indicar algo que es posible. Por ejemplo, si se trata de lanzar una moneda, caer cara es un evento.

1. La probabilidad de cualquier evento está entre cero y uno, pues la probabilidad es la idealización de una frecuencia relativa.

2. La probabilidad de la unión de dos eventos que nada tienen en común es la suma de las probabilidades. Por ejemplo, al lanzar un dado la probabilidad de 4 es 1/6 y la de 5 es 1/6, por la tanto, la probabilidad de que caiga 4 o 5 es 2/6. Pero en general, *la probabilidad de la unión de dos eventos es igual a la suma de las probabilidades de los dos eventos por separado menos la probabilidad de la intersección.* Esto es para evitar doble contabilidad. Por ejemplo, la probabilidad de que al lanzar un dado caiga un número menor que 5 es 4/6. Por otra parte, la probabilidad de que caiga un número par es 1/2. ¿Cuál es la probabilidad de que caiga un número par o menor que 5? Esos números son 1,2,3,4,6. Por lo tanto, la probabilidad es 5/6. Pero 1/2+4/6= 3/6+4/6 = 7/6 que es mayor que uno. Para recobrar el 5/6 debemos restar de 7/6 la probabilidad de los eventos comunes, a saber, los que son pares y son menores que 5, que son 2 y 4, cuya probabilidad es 2/6. Por lo que 7/6-2/6=5/6, que era lo que buscábamos.

3. La probabilidad del universo es uno, pues la suma de todas las frecuencias relativas da uno.

39 ***Ejemplo*** *Establecer las diversas probabilidades en el siguiente caso: Se pregunta a cada una de 250 personas si son simpatizantes del candidato A o del B. Se encontró que 140 son simpatizantes de A, 130 lo son de B. Además, entre todos ellos, hay 25 que son simpatizantes de ambos candidatos.*

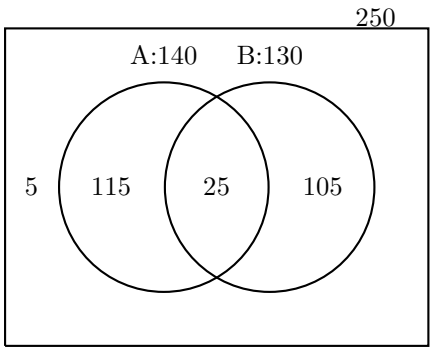


Figura 2.0. Diagrama de disyunciones de Venn: la suma de las partes da la suma del todo.

a) Si seleccionamos una de esas personas al azar, la probabilidad de que sea simpatizante de A es $140/250 = p(A)$.

b) Si seleccionamos una de esas personas al azar, la probabilidad de que sea simpatizante sólo de A es $115/250 = p(A - B)$.

c) Si seleccionamos una de esas personas al azar, la probabilidad de que sea simpatizante de B es $130/250 = p(B)$.

d) Si seleccionamos una de esas personas al azar, la probabilidad de que sea simpatizante sólo de B es de $105/250 = p(B - A)$.

e)Si seleccionamos una de esas personas al azar, la probabilidad de que sea simpatizante de ambos candidatos es de $25/250 = p(A \cap B)$.

f) Si seleccionamos una de esas personas al azar, la probabilidad de que sea simpatizante de al menos uno de los candidatos es de $245/250 = p(A \cup B)$.

g) Si seleccionamos una de esas personas al azar, la probabilidad de que su simpatía no tenga que ver con ninguno de los dos candidatos es $5/250 = p((A \cup B)')$.

h) Si seleccionamos una de esas personas al azar, la probabilidad de que no sea simpatizante de A es $110/250 = p(A')$.

i) Si seleccionamos una de esas personas al azar, la probabilidad de que no sea simpatizante exclusivo de A es $135/250 = p((A - B)')$.

j) Si seleccionamos una de esas personas al azar, la probabilidad de que no sea simpatizante de B es $120/250 = p(B')$.

k) Si seleccionamos una de esas personas al azar, la probabilidad de que no sea simpatizante exclusivo de B es $145/250 = p((B - A)')$.

l) La probabilidad condicional de ser simpatizante de B dado que ya se es de A: $25/140 = p(B/A)$.
m) La probabilidad condicional de ser simpatizante de A dado que ya se es de B: $25/130 = p(A/B)$.

40 ♣ ***Definición.*** *Variable aleatoria = v.a.*

Hay una jerga oficial que se usa y que al principio es muy engorrosa pero que después de entenderla y de familiarizarse un poco se convierte en algo muy útil e imprescindible para comunicarse. Empecemos entendiendo qué se quiere decir cuando se dice X **es una v.a. o variable aleatoria**. Supongamos que uno está haciendo una encuesta, digamos, preguntando el ingreso de las personas que tienen un empleo. Cada persona responde cuanto gana. Ese es el valor de $X=$ sueldo de un empleado. Pero además, si el muestreo es aleatorio y no podemos predecir qué responderá el próximo entrevistado, decimos que X es una variable aleatoria. Cuando uno va reuniendo datos, uno tiene un histograma, y cuando el número de datos tiende al infinito, al histograma de frecuencias relativas se le llama la distribución de la variable aleatoria dada. La distribución tiene una media que se denota μ y una varianza que se denota σ^2 .

41 ***Comentario.*** *La media μ de una variable aleatoria X también es conocida como* ***valor esperado*** *y se puede denotar $E(X)$. De esta forma queremos hacer énfasis en el contraste entre media de datos y media de una variable aleatoria. La media de unos datos es la media de algo que ya pasó, mientras que la media de una variable aleatoria es algo que ni pasó ni pasará (pues se necesitaría reunir un número infinito de datos), pero de tal manera que la media de los resultados experimentales estará por ahí cerca de la media de la variable aleatoria. Por la misma razón, utilizamos una notación especial para la varianza de una variable aleatoria: σ_X^2 ó, cuando se sobreentienda, σ^2.*

Uno tiene la tendencia a pensar que ser aleatorio significa no saber absolutamente nada. Eso es falso: lo que no se sabe es cuál será el próximo dato. Pero uno puede predecir con cierta certeza, por ejemplo, cuál será la moda del sueldo de las personas: será el mínimo. Algunas variables aleatorias vienen tipificadas, lo cual puede suceder por dos motivos: si uno hace un histograma de las frecuencias relativas, los histogramas van adoptando una cierta silueta, por ejemplo, una fuerte tendencia a agruparse alrededor de la media. Decimos en ese caso que se trata de una **distribución de tendencia central**. Cada vez que decimos sobre el valor de una variable que es aproximadamente igual a tanto, estamos en un caso de una distribución de tendencia central. Por ejemplo, ¿cuánto ha estado valiendo el dolar? Aproximadamente 0,70 euros. Eso puede significar que la media es 0,70 y que el valor tiene una cierta fluctuación alrededor de ésta, la cual se tipifica diciendo que la desviación es 0,05.

Pero también puede pasar que uno esté haciendo un experimento con dos opciones, como el de lanzar una moneda, que puede caer cara o sello. Decimos que tenemos **una distribución binomial**, pues binomial significa dos nombres, opciones o posibilidades: cara vs. sello; bueno vs. malo; chico vs. grande.

El trabajo que sigue es predecir la silueta del histograma de las frecuencias relativas que aproximadamente habrán de salir en un arreglo experimental específico o en un muestreo sistematizado. Para entender en qué consiste el experimento más natural con una distribución binomial pensemos en una moneda. La lanzamos n veces y contamos el número de caras que cayeron. El número de sellos no hay que contarlos, pues es n menos el número de caras. Además, la probabilidad de cara es igual a la probabilidad de sello, igual a $1/2$. Podemos pensar también en las familias con n hijos: aproximadamente, en unos casos mejor que en otros, la probabilidad de varón es igual a la probabilidad de niña $(1/2)$. Podemos contar entonces el número de varones en familias con n hijos.

2.2. Distribuciones

La estadística se edifica sobre la teoría de probabilidades, la cual es una historia sin fin. Es usual dedicarle un semestre a estudiarla, pero nosotros vamos a dedicarle solo una mínima atención con la certeza de que es mucho, muchísimo, lo que ganaremos en visión y entendimiento. Vamos, pues, a estudiar la distribución binomial, que se refiere a situaciones en las que aparecen dos opciones, como cara y sello, sí o no, blanco o negro.

42 ***Ejemplo*** *Alturas de unos jóvenes (muestreo real).*

Las alturas de un grupo de muchachos se redondeó a su valor más cercano en decímetros. La siguiente es la tabla de frecuencias absolutas de dichas marcas de clase:

Alturas en dms de un grupo de jóvenes	
Altura	Frecuencia
15	2
16	11
17	15
18	4
19	2
	Total= 34

El histograma asociado es

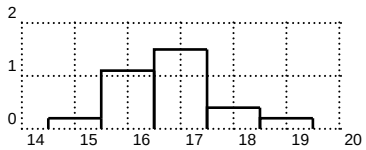


Figura 2.1. Distribución de la altura en decímetros de 34 personas. La unidad vertical es de 10. La distribución es de tendencia central: los datos se agrupan hacia el centro de la distribución. ¿Por qué?

Nos preguntamos: ¿De dónde salen las tendencias? Las alturas de las personas siguen una distribución de tendencia central, pero bien pudiera ser que los datos tratarán de alejarse de la media, como sucede con algunos descriptores de aquellos adolescentes que les incomoda parecerse a la mayoría.

Para explicar de dónde salen las tendencias, tenemos una idea para poner a prueba: una **tendencia central** sale de sumar efectos de varios factores aleatorios, caprichosos, impredecibles, independientes, cada uno de los cuales puede aportar o no aportar a la tendencia. La idea es la siguiente:

Imaginemos que un factor aleatorio es semejante a lanzar una moneda. Si cae cara, hay aporte a la tendencia, pero si cae sello no hay aporte. Si lanzamos, digamos, 4 monedas, habrá en promedio 2 caras y lo que pase en la realidad debe andar por ahí cerca de dos: cerca de la media, generando una distribución de tendencia central.

En retrospectiva podría decirse: un evento cualquiera genera un error con respecto a la media que puede ser por exceso o por defecto. Pero cuando muchos eventos entran a interactuar, los excesos y los defectos con respecto a la media tienden a aniquilarse mutuamente y todo tiende a congregarse alrededor de la media, alrededor del centro de la distribución.

43 ***Ejemplo*** *Juego con monedas.*

Lanzamos una moneda 4 veces y contamos el número de caras y ese es el número que se registra como dato. Digamos, si salió cara- cara- sello- cara, registramos un 3 por ser 3 el número de caras.

Se repite lo mismo muchas veces. Obtenemos una secuencia de números que van desde 0 hasta 4. Se hace el histograma correspondiente, con 5 clases. El resultado será una distribución de tendencia central, agrupada alrededor de 2. Veamos los resultados de un caso real:

Número de caras al lanzar una moneda 4 veces	
Número de caras	Frecuencia
0	3
1	8
2	16
3	6
4	1
	Total de experimentos = 34

El histograma asociado es

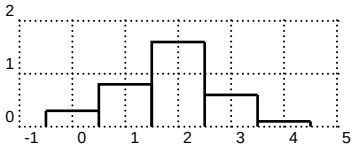


Figura 2.2. Se lanza una moneda 4 veces y se cuenta el número de caras, X. Se repite eso 34 veces y se grafica el histograma de la distribución de X. La unidad vertical es de 10. La distribución es de tendencia central: esperamos en promedio 2 caras y los datos se agrupan hacia la media de la distribución, pues los errores aleatorios por exceso o por defecto con respecto a la media tienden a compensarse.

44 ***Pregunta: ¿Podría mostrarme una tendencia que no sea central?***

Respuesta: Si lanzamos un dado muchas veces y registramos como dato el número de puntos que queda hacia arriba, las frecuencias relativas de los diferentes números del 1 al 6 tienden a ser muy similares y al hacer el histograma uno ya no ve una montaña sino una mesa de altura 1/6. A esa tendencia se le llama la **distribución uniforme**. Y a la tendencia que sale del juego con monedas, cara y sello, se llama **distribución binomial**.

Otro ejemplo: Todos tenemos problemas en la vida. A toda hora. Pero en la generalidad de los casos, uno puede supervivir con relativa alegría y paz. Sin embargo, a veces aparecen problemas que representan una crisis que nos quita el sueño. Y a veces, muy raras veces, uno tiene momentos en los cuales tiene dos problemas críticos. Y por lo menos una vez en la vida, uno tiene un momento en el que todo le va mal, para lo cual es suficiente que uno tenga 3 problemas gordos. Vemos que la distribución de problemas críticos tiene su gran moda en cero, sigue una frecuencia relativa muy moderada para uno, luego algo pequeño para dos, mientras que tres es apenas visible. Tendencias de este tipo se modelan por la **distribución exponencial**.

A una distribución de tendencia central la modelamos con monedas, es decir, con la binomial. Similarmente, a una distribución uniforme la modelamos con un dado, y ahora nos preguntamos. ¿cómo modelamos una distribución exponencial?

Una distribución exponencial resulta en el siguiente experimento: se va lanzando una moneda con probabilidad de cara cercano a uno y se anota como dato el número de veces que fue necesario lanzarla para que cayera cara por primera vez. Lo más probable es que la primera cara caiga bien pronto y que rara vez uno tenga que lanzar la moneda muchas veces para lograr la primera cara.

2.3. La binomial

En un juego con monendas, uno tira la monenda determinado número de veces y cuenta el número de caras. Desde el punto de vista estadístico, eso es exactamente lo mismo que formar un grupo personas y contar el número de mujeres. O contar en un grupo de personas a los que van por Ordúz, candidato a la alcaldía. A todas las situaciones que presentan repeticiones de eventos con dos opciones las referimos con el mismo nombre: **distribución binomial**. Tenemos que especificar: 1) las dos opciones en mente: cara o sello, hombre o mujer, a favor de Ordúz o no a favor. Sin embargo, es usual abstraer y decir que todas las binomiales son juegos de monedas en los cuales se cuenta el número de caras X (o de hombres o de los que están a favor de Ordúz). 2) El número de lanzamientos n (o de grupos de personas entrevistadas) 3) La probabilidad de cara p. Es claro que la probabilidad de cara, p, puede ser cualquier número entre cero y uno. Por ejemplo, la probabilidad de mujer en grupos de rock pesado es 0.01.

45 ***Ejemplo: Binomiales y democracia***

En general, en una distribución binomial (con dos opciones por evento), no necesariamente la probabilidad de un evento es igual a la probabilidad del otro.

Por ejemplo, en Sudáfrica antes de 1990 cada persona era clasificada por el gobierno como blanca o como negra. Los turistas e inmigrantes también recibían dicha clasificación. Ser clasificado como blanco era la dicha pues el gobierno le abría a uno las puertas en todas partes. Pero el gobierno no era tan

cruel: si se presentaba un negro pero había algunos rastros físicos que delataran un ancestro blanco, el paciente era clasificado como blanco. Con la revolución social liderada por Mandela y su gente, los negros recibieron prebendas. Pero aquellos negros que eran clasificados como blancos no recibieron nada. Sucedía además que en tiempo del Apartheid ellos eran bienvenidos por el gobierno mas no por la sociedad. Así que en el mejor de los casos tenían empleos muy bajos que no mejoraron con la ruptura del Apartheid pero estas personas tampoco recibieron nada después del cambio de Constitución pues ellos pertenecían al grupo que lo tenía todo.

Es claro que los negros eran muchos más que los blancos, pero el gobierno adoptó a muchos negros, con algún rastro blancoide, para hacerle creer al mundo que los blancos eran muchos. Digamos $p(\{blanco\}) = 0{,}2$ mientras que $p(\{negro\}) = 0{,}8$.

Ahora conectemos esas probabilidades con la democracia, de la cual hay varias interpretaciones. La más elemental es que todo ciudadano tiene derecho al voto, que es el instrumento que la democracia usa para decidir el futuro, eligiendo el mandatario a gusto personal. En el Apartheid, sólo el 20 % de la población tenía derecho al voto. Después del revolcón de Mandela, todo el mundo tiene derecho al voto. Pero todos sabemos que si uno se queda ahí, no va ganando mucho. Un poco más complicado es lograr que en la democracia los diversos frentes o grupos sociales tengan igualdad de oportunidades. Ahora bien, ¿qué significa tener igualdad de oportunidades?

Algunos alegan que eso significa que en cualquier estamento de gobierno los diversos grupos estén representados proporcionalmente a su frecuencia relativa. Por ejemplo, si se trata de los blancos y negros de Sudáfrica, que el 80 % del parlamento sea negro. Pero otros pensadores más soñadores se imaginan que la democracia no consiste en instalar un gobierno democrático sino en convertir la democracia en una vivencia, lo cual significaría que en cualquier estamento sea de gobierno o de trabajo o de estudio, los diversos grupos estén representados proporcionalmente a su frecuencia relativa. Eso quiere decir que está bien que se piense que el 80 por ciento de los obreros de las muchas minas de oro de ese país sean negros, pero que también pase lo mismo con los puestos directivos.

Pero hay que ver que una democracia no requiere que todos tengan acceso a la propiedad. Con que no haya extrema pobreza, ya uno puede darse por muy bien servido en cualquier democracia. ¿Querrá eso decir que la democracia no es más que el abrigo de oveja con el cual se viste el lobo del capitalismo para poder gozar de las suyas?

Por otro lado, también hay que reconocer que la democracia que convierte frecuencias relativas en ley puede resultar aburrida y fuera de contexto: los diversos grupos pueden tener ideas distintas sobre lo que significa tener éxito social o económico. Por ejemplo: está el citadino que se preparó estudiando duramente y logró conseguir un buen empleo y que tarde la noche llega a su casa a ver por señal satelital programas de aves, ríos y mares. Y este Fulano parece ser feliz. En contraste, está el que es dichoso viendo por la tardecita a los pájaros y garzas y no por televisor sino que los ve subiendo rió arriba, desde su propia casa. Así que lo importante no es que todo el mundo sea todo, sino que cada cual tenga acceso al libre desarrollo de su personalidad, dentro de su ideología, cultura y necesidades particulares. Esa línea de pensamiento lleva a la conclusión de que la forma más desarrollada de gobierno tiende a favorecer al anarquismo.

Regresando de nuevo al estudio de las distribuciones, notemos que en una discusión real no se habla de dos grupos solamente sino de varios. Cuando uno tiene dos opciones uno tiene todo lo que se necesita con la binomial. Similarmente, existe una multinomial para el estudio de experimentos o muestreos con varias opciones. No las veremos, ni a las multinomiales ni a las otras miles de distribuciones importantes, pues con la binomial aprenderemos el modus operandi, que es el que necesitamos saber para aplicar la estadística a la toma de decisiones.

De otra parte, podemos generalizar diciendo que en una distribución binomial, en la cual contamos el número de eventos distinguidos en cierto experimento o muestreo sistematizado, la probabilidad de dicho evento es p, que puede tomar cualquier valor entre cero y uno, y no solamente $1/2$.

Ahora pasamos a temas más formales.

46 ***Representación de una binomial***

Un juego de cara o sello con monedas puede representarse mediante un árbol. Consideremos una moneda real, con probabilidad de cara igual a la de sello igual a 1/2. Si la tiramos una vez, tenemos dos opciones:

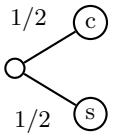


Figura 2.3. Árbol que representa las dos opciones, cara o sello, cuando uno tira una moneda con probabilidad de cara igual a la de sello igual a 1/2.

Si habiendo tirado la moneda una vez, una la tira otra vez, el experimento conjunto de tirar la moneda dos veces se representa así:

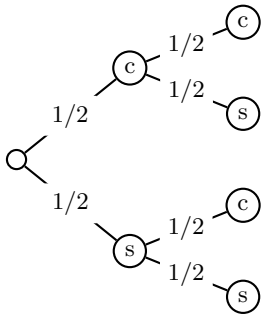


Figura 2.4. Este árbol muestra las cuatro opciones que se tienen cuando uno tira una moneda dos veces: cc, cs, sc, ss. También se indican las probabilidades particulares, cada una igual a 1/2. Éste es el árbol de la $Bi(2, 1/2)$*, que se lee: árbol de la binomial dos un medio que significa que se tira una moneda dos veces cuando la probabilidad de cara es* 1/2.

Lo que nosotros llamamos binomial, cara o sello, es una abstracción genérica de un experimento con dos opciones. Por ejemplo, cara puede representar perder un examen y sello ganarlo. La probabilidad de perder el primer parcial generalmente es baja, digamos 1/8, es decir de cada 8 muchachos hay uno que lo pierde. Por consiguiente, la probabilidad de ganarlo es 7/8. Para interpretar los resultados generalmente asumimos que tirar una moneda, diagamos, 3 veces es lo mismo que tirar 3 monedas. Así, con respecto a los exámenes, aplicamos los resultados no a un muchacho que presenta 35 exámenes sino a 35 muchachos.

Atención: lo que salió cuando uno tiró una moneda no interfiere con lo que pasará cuando la moneda se tira otra vez. Decimos que los eventos son y deben ser independientes. Al hacer aplicaciones es preciso tenerlo en cuenta. Por ejemplo, si estamos hablando de muchachos que presentan un examen, debemos garantizar que no haya copia.

En los eventos o posibles resultados de una binomial uno cuenta sólo el número de caras, que puede representar el número de los que perdieron un examen, o el número de personas afectadas negativamente por el tratado de libre comercio con Irán. La ocpión complementaria no se cuenta pues se deduce de lo que pasa con la primera.

Un experimento genérico con una binomial se denota así:

$$X \sim Bi(n, p)$$

que se lee: X se distribuye como una binomial ene pe. Y que significa: X es una variable aleatoria que cuenta el número de caras en un experimento en el que se lanzan n monedas y tal que la probabilidad de cara sea p y tal que tirar una moneda no interfiere con lo que pasa con lo que sale al lanzarla las otras veces.

Representemos con un árbol la binomial $X \sim Bi(3, 1/8)$, es decir, un experimento en el que se tiran 3 monedas y se cuenta el número de caras, cuando la probabilidd de cara es 1/8:

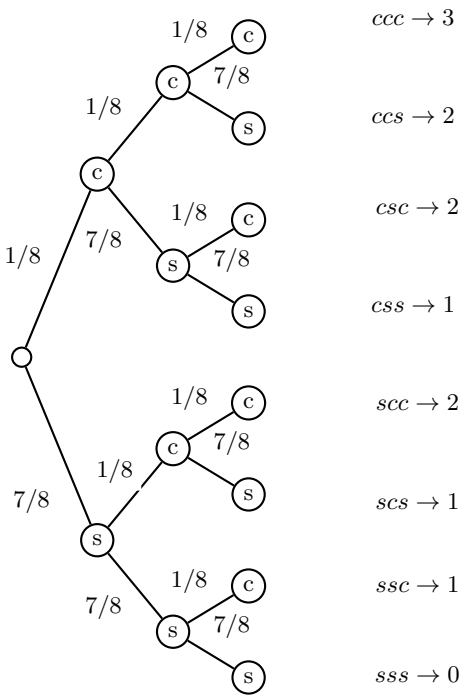


Figura 2.5. Árbol de eventos de la $X \sim Bi(3, 1/8)$: se tira una moneda 3 veces, X cuenta el número de caras y la probabilidad de cara es 1/8. Cada rama da un evento, por ejemplo, la quinta rama de arriba a abajo es scc, el número de caras es $X = 2$.

47 ***Predicciones de la binomial.***

Como asumimos que el azar está en el fondo de todo, no podemos hacer predicciones absolutas pero sí estadísticas. Por ejemplo, si jugamos con monedas verdaderas con probabilidad de cara $p = 1/2$ igual a la de sello y si tiramos 30 monedas no podemos asegurar cuantas caerán cara, pero esperamos que el número de caras se agrupe alrededor de 15 y que no siempre caigan 15 caras es decir que debe haber variabilidad con varianzas y desviaciones no nulas. Para hacer predicciones estadísticas necesitamos que lo que pase en un experimento no repercuta sobre su repetición. Veamos esto al detalle.

48 ***Regla del producto de probabilidades.***

Para poder calcular predicciones respecto de una binomial, necesitamos una regla operacional que nos diga qué significa que los datos vengan al azar. Pues eso quiere decir que si acabó de salir cara, entonces no sabemos qué caerá ahora. Podemos decir eso en un lenguaje operativo, que permita hacer cálculos fácilmente, como sigue:

Cuando dos eventos A y B son **independientes**, es decir, cuando lo que le pase al uno nada tiene que ver con lo que le pase al otro, la probabilidad de que sucedan los dos, sea al tiempo, o sea uno después del otro, es igual al producto de las probabilidades. A esto lo llamamos la **regla del producto de las probabilidades** para eventos independientes. Eso lo notamos como:

$p(AB) = p(A)p(B)$, si A y B son independientes.

Aplicamos la regla del producto para hallar la probabilidad de que, por ejemplo, al lanzar dos monedas, ambas caigan cara. Como la probabilidad de cara es 1/2, la probabilidad de que dos monedas caigan cara es $1/2 \times 1/2 = 1/4$.

49 ***Resumen: El experimento binomial.*** *Muchas situaciones de la vida real quedan bien modeladas por un experimento en el cual se tira una moneda varias veces y se cuenta el número de caras. Con monedas reales, la probabilidad de cara es cercana a un medio y lo mismo la de sello. Pero tratándose de aplicaciones, la probabilidad de cara puede representar la probabilidad de pasar un examen y puede ser cualquier valor entre cero y uno. Oficialmente decimos, sea X la v.a., (variable aleatoria, que no se sabe que saldrá después) que cuenta el número de caras cuando una moneda se tira n veces, entoces X tiene una distribución binomial con probailidad de cara igual a p. Se denota $X \sim Bi(n,p)$, por ejemplo, $X \sim Bi(3,1/8)$, lo cual se lee: X se distribuye como una binomial con $n = 3$ y $p = 1/8$. Extensivamente significa: X es una variable aleatoria que cuenta el número de caras en un experimento en el que se lanza una moneda 3 veces y tal que la probabilidad de cara es* $1/8$.

Al tirar una moneda 3 veces, un evento posible puede ser cara-cara-sello que se denota ccs. En ese caso X toma el valor $x = 2$, pues cayeron dos caras. Para calcular la probabilidad de ccs no podemos decir que todos los eventos son igualmente probables, pues claramente es más probable que caigan sellos y no caras, pues la probabilidad de sello es 7/8. Debemos entonces calcular las probabilidades por la regla de la multiplicación de probabilidades para eventos independientes. En el caso ccs, el número de caras es $x = 2$ y la probabilidad es entonces $(1/8)(1/8)(7/8) = (1/8)^2(7/8)$. Lo notamos

$$ccs \to 2 \to (1/8)^2(7/8)$$

Todas las opciones o eventos posibles aparecen automáticamente en el siguiente árbol:

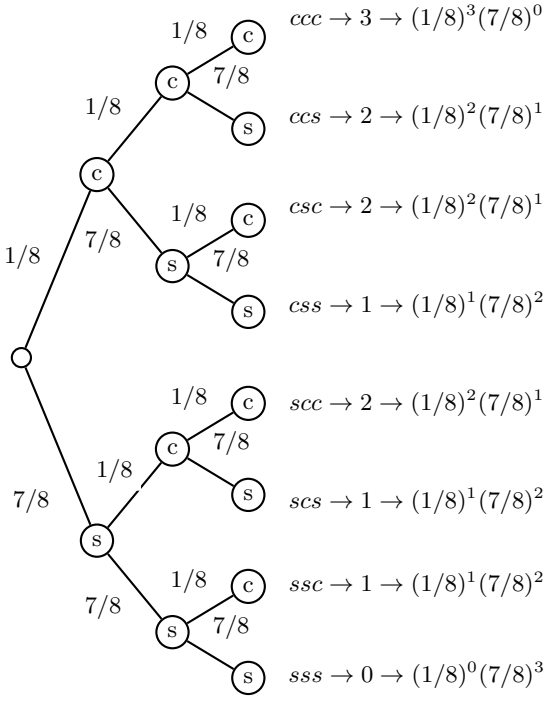


Figura 2.6. Árbol de eventos de la $X \sim Bi(3,1/8)$: se tira una moneda 3 veces, X cuenta el número de caras y la probabilidad de cara es 1/8. Cada rama da un evento, por ejemplo, la quinta rama de arriba a abajo es scc, su número de caras es $X = 2$ y su probabilidad es $(7/8)(1/8)(1/8) = (1/8)^2(7/8)^1$. Pero, atención, la probabilidad de dos caras es $p(X = 2) = p(2) = 3(1/8)^2(7/8)^1$ pues hay 3 eventos con dos caras: ccs,csc, scc.

Este árbol se lee horizontalmente así: se comienza el experimento tirando la moneda la primera vez y cae cara o sello. Es el primer nivel en columna vertical con c, s. Se tira la moneda por segunda vez

y si la primera cayó cara, la segunda puede caer cara o sello. Si la primera vez cayó sello, la segunda puede caer cara o sello. Es el segundo nivel con 4 nodos en columna vertical. El tercer nivel relata lo que puede pasar con la tercera vez que tiramos la moneda, por todo 8 opciones.

Con la información gráfica pasamos a la tabla de probabilidades: X cuenta el número de caras en un experimento en el que se lanza una moneda 3 veces tal que la probabilidad de cara es 1/8. La siguiente tabla contiene las probabilidades de los valores de X, su media y su varianza. Observemos que hay tres eventos que tienen $x = 2$, a saber ccs, csc, scc. Por eso hay que poner un tres en el renglón de $x = 2$ y en la segunda columna.

La media, μ, y la varianza, σ^2, se calculan por las fórmulas siguientes:

$$Media = \mu = \Sigma x p(x)$$

$$Varianza = \sigma^2 = \Sigma(x_i - \mu)^2 p_i = \Sigma(x - \mu)^2 p(x)$$

La siguiente tabla muestra la forma de calcular todo:

$X \sim Bi(3, 1/8)$			
x	$p(x) =$ prob de x	$xp(x)$	$(x-\mu)^2 p(x)$
0	$(1/8)^0(7/8)^3 = 0{,}6699$	$0\ (0.6699) = 0$	$(0 - 0{,}375)^2(0{,}6699) = 0{,}0942$
1	$3(1/8)^1(7/8)^2 = 0{,}2871$	$1\ (0.2871) = 0.2871$	$(1 - 0{,}375)^2(0{,}2871) = 0{,}1122$
2	$3(1/8)^2(7/8)^1 = 0{,}0410$	$2(0.0410) = 0.0820$	$(2 - 0{,}375)^2(0{,}0410) = 0{,}1083$
3	$(1/8)^3(7/8)^0 = 0{,}0020$	$3\ (0.0020) = 0.0059$	$(3 - 0{,}375)^2(0{,}0020) = 0{,}0145$
Σ	$\Sigma p(x) = 1{,}000$	$\mu = \Sigma x p(x) = 0{,}3750$	$\sigma^2 = \Sigma(x-\mu)^2 p(x) = 0{,}3282$

Las probabilidades, en la segunda columna, se usan para calcular la media y la varianza en la tercera y cuarta. El valor de la media, calculado en la tercera columna, se usa para calcular la varianza en la cuarta columna.

La media vale 0.375. La varianza vale $\sigma^2 = 0{,}3282$, la desviación es su raíz cuadrada $\sigma = 0{,}5728$.

Interpretemos los resultados en términos de un estudiante cuya probabilidad de pasar un examen de una determinada materia sea 1/8, la cual tiene 3 exámenes en el semestre: vemos que de 3 exámenes pasa en promedio 0.375. Hablando claramente, perderá todos los exámenes.

Usamos $p(x)$ para hacer el histograma o **diagrama de barras de probabilidad**:

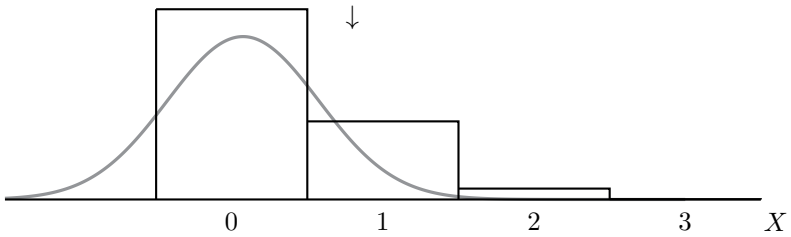


Figura 2.7. Histograma o diagrama de barras de probabilidad de $X \sim Bi(3, 1/8)$. La media está arribita de cero, la desviación estándar vale aproximadamente 0.5. La campana se ha dibujado procurando compensar los excesos con los defectos, es decir, lo que sobre un lado debe compensarse con lo que le falte en otro. Estas gráficas se llaman oficialmente: diagramas de barras de probabilidad.

Para dibujar el histograma se mira cual es el mayor valor entre las probabilidades. En nuestro ejemplo es $0{,}669 \sim 0{,}7$. Por lo tanto, se abre campo para que 0.7 quepa en la gráfica y ocupe ojalá todo el espacio vertical. Los demás valores se grafican en proporción. Se calcula por la gráfica la media y la desviación y se compara con los valores hallados de la tabla. Una gran discrepancia indica que hay algo por corregir.

Todo eso quiere decir que hay variabilidad en los datos y que al aproximar los datos por una campana, los puntos de inflexión quedan a una distancia de 0.57 de la media. En definitiva: al lanzar una moneda tres veces tal que la probabilidad de cara sea 1/8, como la media es 0.37, veremos caer cero caras y una

cara con relativa gran frecuencia. Fuera de ellos, el 2 (dos caras) caerá rara vez y el 3 (tres caras) casi nunca.

Para lucir exactitud total, uno puede usar quebrados:

La suma en la segunda columna de la tabla debería dar uno. La discrepancia se debe a que usamos una aproximación de sólo 3 cifras decimales. Pero que la suma es uno exacto, podemos demostrarlo reconociendo que la suma en quebrados es la expansión de un binomio al cubo:

$(1/8)^0(7/8)^3 + 3(1/8)^1(7/8)^2 + 3(1/8)^2(7/8)^1 + (1/8)^3(7/8)^0$
$= (1/8 + 7/8)^3 = 1^3 = 1$

Lo cual nos dice, por lo menos, que vamos bien pues la probabilidad de todo el mundo da uno. Calculemos la media y la varianza de la $Bi(3, 1/8)$ usando quebrados:

La fórmula para la media

$$media = \mu = \Sigma xp = \Sigma x_i p_i = \Sigma x(i)p(i) = x_0p_0 + ... + x_np_n$$

que se convierte en

$media = \mu = 0p(0) + 1p(1) + 2p(2) + 3p(3)$
$= 0 + (1) * 3(1/8)^1(7/8)^2 + 2 * 3(1/8)^2(7/8)^1 + 3(1/8)^3(7/8)^0$
$= 0 + 3 * (7^2)/8^3 + 6 * 7/8^3 + 3/8^3 = (147 + 42 + 3)/512 = 192/512 = (3 * 64)/(8 * 64) = 3/8 = 0{,}375$

Eso quiere decir, que si se repite el experimento 8 veces, nos caerán en promedio tres caras. Observemos que la media se puede escribir como $3(1/8)$ que corresponde a np: ¿tendremos aquí una coincidencia o una ley?

La varianza es:

$$Varianza = \sigma^2 = \Sigma(x_i - \mu)^2 p_i = \Sigma(x - \mu)^2 p(x)$$

Como la media vale $3/8$, tenemos:
$\sigma^2 = (0 - 3/8)^2 p(0) + (1 - 3/8)^2 p(1) + (2 - 3/8)^2 p(2) + (3 - 3/8)^2 p(3)$
$= (3/8)^2(7/8)^3 + (5/8)^2 * 3(1/8)(7/8)^2 + (13/8)^2(3 * (1/8)^2(7/8) + (21/8)^2(1/8)^3$
$= (1/8^5)(3087 + 25 * 3 * 49 + 169 * 3 * 7 + 441)$
$= 10752/32768 = (21 * 512)/(64 * 512) = 21/64 = 3(1/8)(7/8).$
Para hallar la desviación σ, se saca la raíz cuadrada de la varianza.
$\sigma = \sqrt{3(1/8)(7/8)} = \sqrt{0{,}328} = 0{,}57$

50 ***Ejemplo*** *Estudiémos $X \sim Bi(4, 1/2)$, es decir, hagamos la lista de todas los eventos, con el valor de X, el número de caras, y su probabilidad. Hagamos el histograma para X, demostremos que vamos bien demostrando que la suma de todas las probabilidades da uno, calculemos la media y la varianza. Interpretemos el problema en términos del regaño que Luis le pegó a Maricela.*

Respuesta:
Todos los casos son:

ssss, sssc, sscs, scss, csss, sscc, scsc, sccs, cssc, cscs, ccss, cccs, ccsc, cscc, sccc, cccc.

Tenemos 16 casos, todos equiprobables, pues $p(c) = p(s)$. Hay un único caso con $X = 0$, ssss, y hay 4 con una cara: csss, scss, sscs, sssc. La tabla de probabilidades para X, el número de caras, es

$X \sim Bi(4, 1/2)$				
Valores en la primera fila, probabilidades en la segunda				
0	1	2	3	4
1/16	4(1/16)	6(1/16)	4(1/16)	1/16

la suma da uno, pues 1+4+6+4+1 = 16 en el numerador que es lo mismo que en el denominador. Otra forma de verificar que la suma de todas las probabilidades da uno es:

$1/16 + 4(1/16) + 6(1/16) + 4(1/16) + 1/16$
$= (1/2)^4 + 4(1/2)^3(1/2) + 6(1/2)^2(1/2)^2 + 4(1/2)^3(1/2) + (1/2)^0(1/2)^4$
$= (1/2 + 1/2)^4 = 1.$

Cálculo de la media:
$Media = \mu = \sum x_i p_i = \sum_{i=1}^{n} x_i p_i$
$= 0(1/16) + 1(4/16) + 2(6/16) + 3(4/16) + 4(1/16)$
$= 0 + 4/16 + 12/16 + 12/16 + 4/16 = 32/16 = 2 = 4(1/2).$

Cálculo de la varianza:
$\sigma^2 = \Sigma(x_i - \mu)^2 p(x_i)$
$= (0-2)^2 * 1/16 + (1-2)^2 * 4/16 + (2-2)^2 * 6/16 + (3-2)^2 * 4/16 + (4-2)^2 * 1/16$
$= 4/16 + 4/16 + 4/16 + 4/16$
$= 1$

Como la desviación es la raíz cuadrada de la varianza, la desviación es:
$\sigma = 1.$

Podemos resumir todo en el siguiente histograma que también tiene el nombre de **diagrama de barras de probabilidad**:

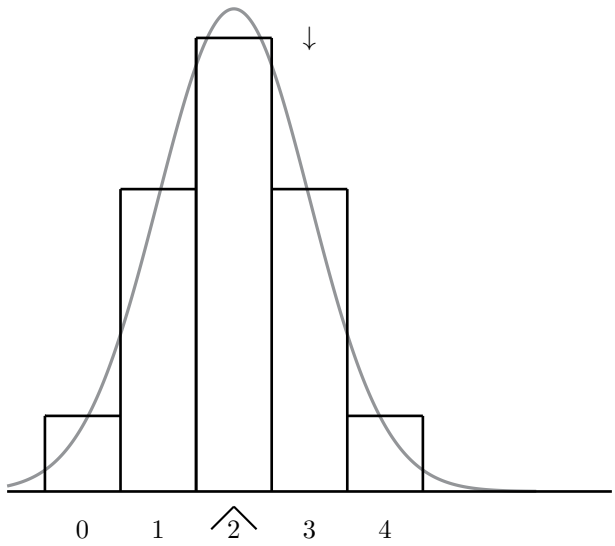


Figura 2.8. Histograma de $X \sim Bi(4, 1/2)$. La media está en 2, la desviación estándar vale 1. Obsérvese la forma de campana de la envolvente, la cual se ha dibujado de tal manera que compensa los excesos con los defectos.

Ahora tratemos de comprender el regaño que Luis le pegó a Maricela. Todo comenzó porque mientras Luis esperaba a Maricela, se dio cuenta de que empezaba a sufrir desde antes de empezar a esperarla pues a veces ella no llegaba. Y tuvo tiempo para contabilizar lo que había pasado en las últimas 4 citas: dos llegó y otras dos no llegó y si a una llega a la otra yano. Y habiendo llegado Maricela, Luis le dijo: *¿Me podrías dar la moneda que controla nuestras citas para echarla a la basura?*

51 ***Ejemplo: Sara contra Laura*** Sara estaba muy enojada con Laura porque ella le quitó el novio. Y por eso Sara le dijo algo que quería decirle desde tiempo atrás pero que nunca se había atrevido: *Laura, en materia de hombres tu eres muy aguerrida, pero cuando se trata de tu profesión como juez, eres una blandengue. Digo esto porque de los últimos 5 casos, todos los has fallado a favor del sindicado. Los hombres hacen lo mismo pero es porque reciben billete a lo loco y por eso son muy ricos, pero a ti no se te compra con dinero sino con una lágrima de cocodrilo. Por eso es que este país está como está, porque el mal no se castiga sea por corrupción o sea por falta de visión y de autoridad.*

Laura bajó la cabeza y dijo: *Sara, estás en lo cierto: a veces me gustaría ser más dura.*

Sara sintió que iba ganando la pelea y dijo en tono reconciliador: *No creas que yo estoy en el otro extremo. Porque si bien mis últimos 4 casos han sido fallados en contra del sindicado, una mirada más amplia, a los últimos 80 casos, permite ver que la probabilidad de que yo condene a alguien es 3/8, muy cercana a 3/9 que es la probabilidad de condenar al sindicado por parte de todo el colegio de jueces de nuestra ciudad.*

Laura levantó la cabeza y la miró. No dijo ni una sola palabra, pero con su actitud le estaba diciendo: *por eso estoy orgullosa de que seas mi amiga, porque tengo mucho que aprender de ti.*

Sara captó el mensaje, se sintió satisfecha y se fue a seguir trabajando. Pero por la noche, cuando pudo reflexionar, se dio cuenta de que se había dejado envolver por la dulzura de Laura, la misma dulzura que los hombres, a quienes ella llama miserables, no pueden resistir. Y sintió verguenza de sí misma y desde ese día nunca más volvió a hablarle. Y todos creen que es por causa del novio. Pero eso, por supuesto, no es cierto.

52 ***Ejemplo*** $X \sim Bi(1,p)$.

Lista de todos los eventos posibles: no hay más que dos posibilidades, o la moneda lanzada cae cara o cae sello. Si X cuenta el número de caras, X puede tomar los valores 0 o bien 1. Toma el valor 1 con probabilidad p, pues cae una cara, y el valor cero con probabilidad $1-p$, pues no cae cara.

$X \sim Bi(1,p)$		
Eventos	X	P
s	0	$1-p$
c	1	p
Suma		1

Aplicando las fórmulas de media y varianza a la $Bi(n,p)$ tenemos:

$\mu = 0p(0) + 1p(1) = 0 + p(1) = p$

pues $p(1)$ es la probabilidad de que caiga una cara cuando se lanza una sola moneda y eso vale p.

Este resultado dice que si la probabilidad de caer cara es p entonces al lanzar una moneda en promedio nos caerán p caras y si lanzamos n monedas nos caerán np caras en promedio. Por ejemplo, si lanzamos dos monedas y $p = 1/2$ entonces en promedio nos caerá 1 cara.

La varianza es:

$\sigma^2 = (0-p)^2p(0)+(1-p)^2p(1) = p^2(1-p)+(1-p)^2p = (1-p)(p^2+(1-p)p) = (1-p)(p^2+p-p^2) = (1-p)p$

Definiendo q como $1-p$, podemos notar a la varianza

$$\sigma^2 = pq$$

y a la desviación como

$$\sigma = \sqrt{pq}$$

53 ***Ejemplo*** $X \sim Bi(2,p)$

Si $X \sim Bi(2,p)$ estamos contando el número de caras cuando lanzamos dos monedas y la probabilidad de cara es p. Si uno lanza dos monedas, pueden caer los siguientes eventos: *cc, cs, sc, ss*. Si la probabilidad de cara es p y la de caer sello es q con $p+q=1$, entonces en el caso *cc* hay 2 caras y la probabilidad es p^2, en el caso *cs* hay una cara y la probabilidad es pq. En el caso *sc* hay una cara y la probabilidad es qp y en el caso *ss* hay cero caras y la probabilidad es q^2.

$X \sim Bi(2,p)$		
Evento	X	probabilidad
ss	0	$(1-p)^2$
sc	1	$(1-p)p$
cs	1	$p(1-p)$
cc	2	p^2

Por tanto, la tabla de probabilidades de la distribución de X es:

$X \sim Bi(2,p)$	
x_i	p_i
0	$(1-p)^2$
1	$2(1-p)p$
2	p^2
Suma	$\Sigma p_i = 1?$

Verifiquemos que la suma de todo el mundo da uno:

$(1-p)^2 + 2(1-p)p + p^2 = [(1-p)+p]^2 = 1^2 = 1$

Cálculo de la media:

$\mu = 0p(0) + 1p(1) + 2p(2) = 0 + (1)2(1-p)p + (2)p^2 = 2p(1-p+p) = 2p$
Cálculo de la varianza:

$\sigma_X = (0-2p)^2p(0) + (1-2p)^2p(1) + (2-2p)^2p(2)$
$= 4p^2(1-p)^2 + (1-p-p)^2 2(1-p)p + 4(1-p)^2p^2$
$= 4p^2(1-p)^2 + [(1-p)^2 - 2(1-p)p + p^2]2(1-p)p + 4(1-p)^2p^2$
$= 4p^2(1-p)^2 + 2(1-p)^3p - 4(1-p)^2p^2 + 2(1-p)p^3 + 4(1-p)^2p^2$
$= 2(1-p)^3p + 2(1-p)p^3 + 4(1-p)^2p^2$
$= 2p(1-p)[(1-p)^2 + p^2 + 2(1-p)p]$
$= 2p(1-p)[(1-p)+p]^2 = 2p(1-p)(1)$
$= 2p(1-p)$

54 ***Ejemplo*** *Estudiemos la distribución $X \sim Bi(3,p)$. En este caso, se trata de lanzar una moneda 3 veces y de contar el número de caras, sabiendo que la probabilidad de cara es p.*

La siguiente es la lista de todos los eventos, del valor de X y de la probabilidad correspondiente:

$sss \to 0 \to p^0(1-p)^3$
$ssc \to 1 \to p^1(1-p)^2$
$scs \to 1 \to p^1(1-p)^2$
$css \to 1 \to p^1(1-p)^2$
$scc \to 2 \to p^2(1-p)^1$
$csc \to 2 \to p^2(1-p)^1$
$ccs \to 2 \to p^2(1-p)^1$
$ccc \to 3 \to p^3(1-p)^3$

Con esta lista podemos hacer el histograma para las probabilidades de X.

$X \sim Bi(3,p)$	
x_i	p_i
0	$p^0(1-p)^3$
1	$3p^1(1-p)^2$
2	$3p^2(1-p)^1$
3	$p^3(1-p)^0$
Suma	$\Sigma p_i = 1?$

Demostremos que la suma es uno:
$(1-p)^3 + 3p^1(1-p)^2 + 3p^2(1-p)^1 + p^3(1-p)^0 = [(1-p)+p]^3 = 1^3 = 1$

Lo cual nos dice, por lo menos, que vamos bien pues la probabilidad de todo el mundo da uno. Calculemos la media y la varianza de la $Bi(3,p)$. La fórmula para la media es

$$\mu = \Sigma xp = \Sigma x_i p_i = x_1 p_1 + ... + x_n p_n$$

que se convierte en

$\mu = 0p(0) + 1p(1) + 2p(2) + 3p(3)$
$= 0 + (1)3p^1(1-p)^2 + 2 \times 3p^2(1-p)^1 + 3p^3$
$= 3p(1-p)^2 + 6p^2(1-p) + 3p^3$
$= 3p[(1-p)^2 + 2p(1-p) + p^2] = 3p[(1-p)+p]^2$
$= 3p[1]^2$
$= 3p$

La varianza es:
$\sigma^2 = \Sigma(x_i - \mu)^2 p_i$

Como la media vale $3p$ y utilizando la notación $q = 1-p$ obtenemos:
$= 3p[1]^2$
$\sigma_X = (0-3p)^2 p(0) + (1-3p)^2 p(1) + (2-3p)^2 p(2) + (3-3p)^2 p(3) =$
$= (3p)^2 q^3 + (1-3p)^2(3pq^2) + (2-3p)^2(3p^2 q) + (3-3p)^2 p^3 =$
$= 9p^2q^3 + (1-6p+9p^2)(3pq^2) + (4-12p+9p^2)(3p^2q) + (9-18p+9p^2)p^3 =$
$= 9p^2q^3 + 3pq^2 - 18p^2q^2 + 27p^3q^2 + 12p^2q - 36p^3q + 27p^4q + 9p^3 - 18p^4 + 9p^5$

Para entender que es ese enredo, clasifiquemos los términos por el exponente total, según el grado:
5: $9p^2q^3 + 27p^3q^2 + 27p^4q + 9p^5 = 9p^2(q^3 + 3pq^2 + 3p^2q + 3p^3) = 9p^2(q+p)^3 = 9p^2$
4: $-18p^2q^2 - 36p^3q - 18p^4 = -18p^2(q^2 + 2pq + p^2) = -18p^2(q+p)^2 = -18p^2$
3: $3pq^2 + 12p^2q + 9p^3 = 3pq^2 + 3p^2q + 9p^2q + 9p^3 = 3pq(q+p) + 9p^2(q+p) = 3pq + 9p^2$

Sumando, obtenemos: $9p^2 - 18p^2 + 3pq + 9p^2 = 3pq$.

2.4. Binomiales a toda velocidad (Triángulo de Pascal)

Hemos visto cómo se calculan binomiales de diverso tipo y ya nos dimos cuenta de que la tarea puede llegar a ser muy engorrosa a medida que se aumenta el número de monedas por experimento.

Nos complace anunciar que tenemos una buena noticia: hay un método simple y extrarápido para sacar la tabla de probabilidades y lo vamos a explicar ya. El primer paso es saber calcular los coeficientes binomiales para lo cual tenemos la siguiente ayuda:

55 ***El triángulo de Pascal.*** *Tiene la estructura siguiente:*

(1)				1		1				
(2)			1		2		1			
(3)		1		3		3		1		
(4)	1		4		6		4		1	
(5)	1	5		10		10		5		1

El primer renglón del triángulo sale de
$(a+b)^1 = 1a + 1b$

Los coeficientes de $(a+b)^2$ se leen sobre el renglón número dos del triángulo y son 1, 2, 1:
$(a+b)^2 = 1a^2 + 2ab + 1b^2$

Los coeficientes de $(a+b)^3$ se leen sobre el renglón número tres del triángulo y son 1, 3, 3, 1:
$(a+b)^3 = 1a^3 + 3a^2b + 3ab^2 + b^3$.

El triángulo nos permite predecir cuáles serán los coeficienes para $n = 4$. Son 1, 4, 6, 4, 1:
$(a+b)^4 = 1a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$.

Además podemos reproducir el triángulo notando lo siguiente: el triángulo de Pascal tiene unos en los bordes y empieza su primer renglón con dos unos. De ahí en adelante cada nuevo renglón se construye a partir del anterior. La suma de dos números seguidos da el valor intermedio del renglón siguiente. Ejemplo: Tomamos $1 + 2$ del segundo renglón y obtenemos el primer tres del tercero. Hacemos lo mismo con 6 y 4 del cuarto renglón y obtenemos el 10 del quinto. Y así hasta llegar a la potencia que necesitemos. Por ejemplo, para $n = 5$ tenemos que los coeficientes son 1, 5, 10, 10, 5 y 1:

$(a+b)^5 = a^5 + 5a^4b^1 + 10a^3b^2 + 10a^2b^3 + 5ab^4 + b^5$

Notemos que para la potencia n hay $n+1$ términos y que mientras que las potencias de a bajan, las de b suben y ambas van desde 0 hasta n.

56 ***Ejemplo*** *Calculemos la $Bi(3, 1/8)$ a toda velocidad.*

Paso 1: Verbalizamos la pregunta: se trata de estudiar un experimento en el cual se lanzan 3 monedas, tal que la probabilidad de cara es 1/8 y X cuenta el número de caras. X puede tomar valores 0,1,2,3.

Paso 2: Hacemos una tabla de probabilidades pero sin coeficientes, teniendo en cuenta tan sólo los exponentes. Por ejemplo, para $x = 0$, tenemos cero caras y 3 sellos, por tanto el término tiene la forma $(1/8)^0(7/8)^3$, pues la probabilidad de cara es 1/8 y la de sello 7/8. Pero si $x = 2$, hay dos cara y un sello: tenemos el término $(1/8)^2(7/8)^1$. Todo eso lo ponemos en la tabla:

$X \sim Bi(3, 1/8)$	
x_i	p_i
0	$(1/8)^0(7/8)^3$
1	$(1/8)^1(7/8)^2$
2	$(1/8)^2(7/8)^1$
3	$(1/8)^3(7/8)^0$
Suma	No se suma

Paso 3: Se completa la tabla con los coeficientes tomados del triángulo de Pascal 1,3,3,1. Añadimos dichos coeficientes a la tabla, uno por renglón. No es necesario preocuparse si se empieza desde arriba o desde abajo pues los coeficentes son simétricos con respecto a la mediana. Nos queda:

$X \sim Bi(3,1/8)$	
x_i	p_i
0	$1(1/8)^0(7/8)^3 = 0{,}699$
1	$3(1/8)^1(7/8)^2 = 0{,}287$
2	$3(1/8)^2(7/8)^1 = 0{,}040$
3	$1(1/8)^3(7/8)^0 = 0{,}001$
Suma	$\Sigma p_i = 1{,}03$

Entendamos lo que significan los coeficientes del triángulo de Pascal. Si x_i vale 0, eso quiere decir que estamos estudiando el caso de que de 3 monedas, cero caigan cara. Y eso se da de una única forma, que todas caigan sello, sss. Si x_i vale 1, eso quiere decir que estamos estudiando el caso de que de 3 monedas, una caiga cara. Y eso se da de 3 formas distintas: css,scs,ssc. Si x_i vale 2, estamos estudiando el caso de que de 3 monedas, dos caigan cara. Y eso se da de 3 formas distintas: ccs,csc,scc. Y si x_i vale 3, se trata del caso en el que de 3 monedas, todas caen cara. Y eso se da de una única forma: ccc.

Paso 4: Test: la suma total debe dar uno excepto por aproximaciones en el manejo de la calculadora. En efecto, estamos expandiendo $(p+q)^n$ donde p es la probabilidad de cara y q la de sello por lo que $q = 1-p$. Por tanto, $(p+q)^n = (p+(1-p))^n = 1^n = 1$. Es por eso que si la suma total es diferente de UNO, más allá de una cifra decimal, hay un error, el cual debe corregirse.

Paso 5: Histograma.

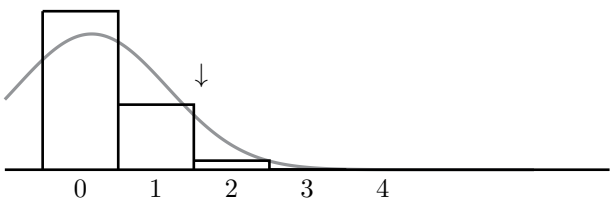


Figura 2.9. Histograma de $X \sim Bi(3,1/8)$. La media está arribita de cero, la desviación estándar vale aproximadamente 1. La campana se ha dibujado de tal manera que compense los excesos con los defectos.

Para dibujar el histograma se mira cual es el mayor valor: $0{,}699 \sim 0{,}7$. Por lo tanto, se abre campo para que 0.7 quepa en la gráfica y ocupe toda el espacio vertical. Los demás valores se grafican en proporción. Se calcula por la gráfica la media y la desviación.

Paso 6: Ahora puede calcularse la media y la varianza por fórmulas supercortas:
La media de una binomial $Bi(n,p)$ es $\mu = np$.
La varianza de una binomial $Bi(n,p)$ es $\sigma^2 = np(1-p)$.
En nuestro caso, la binomial $Bi(3,1/8)$, tenemos:
La media de la binomial $Bi(3,1/8)$ es $\mu = 3(1/8) = 0{,}375$.
La varianza es $\sigma^2 = np(1-p) = 3(1/8)(7/8) = 0{,}328$.
La desviación es $\sigma = \sqrt{0{,}328} = 0{,}572$

57 ***Ejemplo*** *Calculemos la tabla de probabilidades de la binomial $X \sim Bi(6,1/4)$.*

Al proceder igual que antes encontramos un problema: el triángulo de Pascal no tiene el renglón con $n = 6$. Solución: pues como no lo tiene, nos lo inventamos. Para ello recordamos que al sumar dos números consecutivos sobre el mismo renglón se obtiene el número inmediatamente debajo sobre el renglón inferior. Además hay que emparedar con unos. Nuestro triángulo queda como:

$$
\begin{array}{ccccccccccccc}
 & & & & & 1 & & 1 & & & & & \\
 & & & & 1 & & 2 & & 1 & & & & \\
 & & & 1 & & 3 & & 3 & & 1 & & & \\
 & & 1 & & 4 & & 6 & & 4 & & 1 & & \\
 & 1 & & 5 & & 10 & & 10 & & 5 & & 1 & \\
1 & & 6 & & 15 & & 20 & & 15 & & 6 & & 1
\end{array}
$$

Por lo tanto, la tabla de probabilidades de la binomial $X \sim Bi(6, 1/4)$ es

$X \sim Bi(6, 1/4)$	
x_i	p_i
0	$1(1/4)^0(3/4)^6 = 0{,}178$
1	$6(1/4)^1(3/4)^5 = 0{,}356$
2	$15(1/4)^2(3/4)^4 = 0{,}297$
3	$20(1/4)^3(3/4)^3 = 0{,}132$
4	$15(1/4)^4(3/4)^2 = 0{,}033$
5	$6(1/4)^5(3/4)^1 = 0{,}004$
6	$1(1/4)^6(3/4)^0 = 0{,}000$
Suma	$\Sigma p_i = 1$

Explicitemos el significado de algunos coeficentes del triángulo de Pascal.

Si x_i vale 0, eso quiere decir que estamos estudiando el caso de que de 6 monedas, cero caigan cara. Y eso se da de una única forma, que todas caigan sello, ssssss. Si x_i vale 1, eso quiere decir que estamos estudiando el caso de que de 6 monedas, una caiga cara. Y eso se da de 6 formas distintas: cssssss,scssss,sscsss, ssscss, sssscs,ssssssc. Si x_i vale 2, estamos estudiando el caso de que de 3 monedas, dos caigan cara. Y eso se da de muchas formas distintas, 15 por todas: ccssss,cscsss,sccsss... Si x_i vale 5, se trata del caso en el que de 6 monedas, 5 caen cara. Y eso se da de 6 formas: cccccs,ccccsc,cccscc,ccsccc,cscccc, sccccc. Y si x_i vale 6, se trata del caso en el que de 6 monedas, todas caen cara. Y eso se da de una única forma: cccccc.

58 ***Ejemplo*** *Por tradición se sabe que la probabilidad de resolver un punto cualquiera de un examen dado es* $1/4$*, es decir, de* 4 *puntos al azar, un muchacho cualquiera resuelve* 1*. a) ¿Cómo se distribuyen las notas en un examen de 6 puntos completamente independientes si lo único que se tiene en cuenta es si se resolvió el punto o no? b) ¿Cuál es la probabilidad de que un estudiante cualquiera resuelva por lo menos 4 puntos?*

Solución. a) Estamos en un problema de dos opciones, resolver un punto o no hacerlo. Por lo tanto, se trata de una binomial, lo cual queda evidente si llamamos cara al evento de resolver el punto y sello al de no resolverlo. Como son 6 puntos, estamos en el caso de 6 monedas con probabilidad de cara igual a $1/4$. En conclusión, la nota X se distribuye como $X \sim Bi(6, 1/4)$, la cual fue estudiada en el ejemplo anterior.

b) La probabilidad de resolver por lo menos cuatro puntos es igual a la probabilidad de resolver 4 puntos más la de resolver 5 más la de resolver 6. Se denota

$$p(X \geq 4) = p(4) + p(5) + p(6) = 0{,}03 + 0{,}004 + 0 = 0{,}034$$

Concluímos que es inusual que alguien pueda resolver 4 puntos o más.

59 ***Ejemplo*** *Calculemos la tabla de probabilidades de la binomial* $X \sim Bi(8, 1/2)$*. Para ello ampliamos el triángulo de Pascal hasta 8 y tenemos en cuenta que* $(1/2)^8 = 1/256$*.*

$X \sim Bi(8,1/2)$	
x_i	p_i
0	$1(1/2)^0(1/2)^8 = 1(1/2)^8 = 1/256$
1	$8(1/2)^1(1/2)^7 = 8(1/2)^8 = 8/256$
2	$28(1/2)^2(1/2)^6 = 28(1/2)^8 = 28/256$
3	$56(1/2)^3(1/2)^5 = 56(1/2)^8 = 56/256$
4	$70(1/2)^4(1/2)^4 = 70(1/2)^8 = 70/256$
5	$56(1/2)^5(1/2)^3 = 56(1/2)^8 = 56/256$
6	$28(1/2)^6(1/2)^2 = 28(1/2)^8 = 28/256$
7	$8(1/2)^7(1/2)^1 = 8(1/2)^8 = 8/256$
8	$1(1/2)^8(1/2)^0 = 1(1/2)^8 = 1/256$
Suma	$\Sigma p_i = 1$

Verificamos que la suma da uno:
$(1/2)^8 + 8(1/2)^7(1/2) + 28(1/2)^6(1/2)^2 + 56(1/2)^5(1/2)^3$
$+70(1/2)^4(1/2)^4 + 56(1/2)^3(1/2)^5 + 28(1/2)^2(1/2)^6 + 8(1/2)(1/2)^7 + (1/2)^8 = (1/2+1/2)^8 = 1^8 = 1$

60 ***El histograma o diagrama de barras de la*** $Bi(8,1/2)$***:***

Para graficar el **histograma o diagrama de barras** de la distribución binomial graficamos las probabilidades para cada valor:

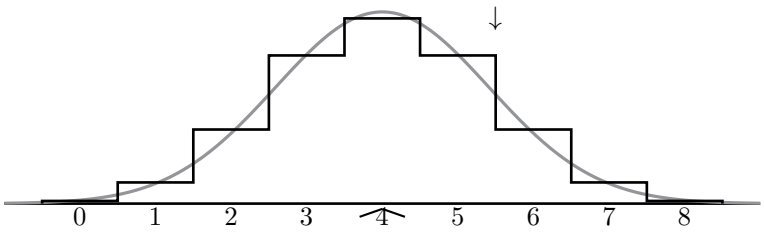


Figura 2.10. La $Bi(8,1/2)$ *queda muy bien aproximada por una campana. Las alturas salen de los numeradores de la tabla de probabilidades.*

Esta distribución tiene media $\mu = np = 8(1/2) = 4$, varianza $\sigma^2 = npq = 8(1/2)(1/2) = 2$ donde $q = 1 - p$ y desviación igual a $\sigma = 1{,}4$.

61 ***El coeficiente binomial***

Cuando hay que calcular coeficientes binomiales con un n grande, hay una regla que es más rápida que el triángulo de Pascal y se llama la regla del **coeficiente binomial**. Para entenderla, empecemos considerando la siguiente expansión, donde $q = 1 - p$:

$$(p+q)^5 = 1p^0q^5 + 5p^1q^4 + 10p^2q^3 + 10p^3q^2 + 5p^4q^1 + 1p^5q^0$$

Teniendo en cuenta que n es el número de monedas y que r es la variable que cuenta el número de caras, $0, 1, ..., n$, la expresión anterior se generaliza en

$$(p+q)^n = \binom{n}{0}p^0q^n + \binom{n}{1}p^1q^{n-1} + \binom{n}{2}p^2q^{n-2} + \binom{n}{3}p^3q^{n-3} + ... + \binom{n}{n-1}p^{n-1}q^1 + \binom{n}{n}p^nq^0$$

donde usamos el símbolo $\binom{n}{r}$ para indicar el **coeficiente binomial para** n **monedas y** r **caras** en la expansión de $(p+q)^n$ correspondiente al término cuyas potencias son p^rq^{n-r}.

Para calcular el coeficiente $\binom{n}{r}$ usamos la siguiente fórmula:

$\binom{n}{r} = \frac{n!}{r!(n-r)!}$

donde el símbolo ! es el **factorial** y representa el producto del número por todos los demás menores que él. Ejemplo: el coeficiente de a^3b^2 en $(a+b)^5$ es

$\binom{5}{3} = \frac{5!}{3!2!} = \frac{5\times4\times3\times2\times1}{(3\times2\times1)\times(2\times1)} = \frac{120}{12} = 10$

Regla especial: 0! es uno.

Para calcular $\binom{5}{2}$ con la **Calculadora Casio**, se escribe 5 luego se oprime la tecla nCr y luego se escribe 2. Debe dar 10. Sin calculadora, se simplifica inteligentemente, por ejemplo:

$C_8^{15} = \frac{15!}{8!7!} = \frac{9\times10\times11\times12\times13\times14\times15}{1\times2\times3\times4\times5\times6\times7} = \frac{9\times10\times11\times12\times13\times14}{1\times2\times4\times6\times7} = \frac{9\times10\times11\times12\times13}{1\times4\times6} = \frac{9\times5\times11\times12\times13}{1\times2\times6}$
$= \frac{9\times5\times11\times13}{1} = 6435$

Los coeficientes hallados se conectan a la expansión binomial multiplicándolos por p y q a sus potencias respectivas.

Con la nueva notación, la fórmula general para $p(r)$, la probabilidad de r caras con $p(cara) = p$ y $n-r$ sellos con $p(sello) = q = 1-p$ en una distribución binomial $Bi(n,p)$ cuando se tiran n monedas, se escribe oficialmente como

$p(r) = \binom{n}{r} p^r q^{n-r}$

Por ejemplo, para $Bi(5,0.23)$ tenemos que la probabilidad de 2 caras es

$p(2) = \binom{5}{2} (0.23)^2(0.77)^{5-2} = 0.2415$

2.5. Conclusión

El estudio de sistemas complejos es posible si uno cambia el concepto de verdad absoluta por el de probabilidad, la cual puede entenderse como una frecuencia relativa en una gran muestra tomada al azar. Una moneda, con cara y sello, nos ha servido de modelo para ilustrar cómo se trabaja. En consecuencia, hemos introducido la distribución binomial y hemos aprendido a calcularla y a comprenderla.

Capítulo 3

Taller y parcial I

62 ***Objetivo:*** *El siguiente taller contiene el tipo de preguntas que pueden aparecer en el parcial. A continuación proponemos unos ejemplos de parcial que el estudiante deberá poder resolver antes de 45 minutos.* ***HAY RESPUESTAS PARA CASI TODOS LOS EJERCICIOS en la pág. 131.***

3.1. Taller I

1 . Estudiar los siguientes datos (ayuda: ejemplo 2, pág. 1. Respuestas en la pág 131):

a) Los colores dominantes de banderas de partidos políticos: rojo, rojo, azul, amarillo, amarillo, verde, amarillo.

b) Sentimiento de personas hacia el líder de la organización: cariño, indiferencia, desprecio, desprecio, admiración, admiración, rabia, desprecio.

c) La duración de los aplausos a los candidatos a Edil en una localidad: muy poca, mucha, muchísima, más bien poca, casi nada, muchísima. ¿Podemos considerar que en esta localidad ya hay una democracia madura en el sentido de que la comunidad, heterogénea como es, se siente bien representada?

2 . Diseñe dos cascos que protejan la cabeza de lesiones en dos deportes extremos, el uno de patinaje sobre obstáculos y el otro de combate con palo. Haga los dos dibujos de los dos cascos. La tabla de frecuencias relativas de las diferentes lesiones es la siguiente (la región temporal está sobre la oreja, la parietal va por encima de la cabeza, la occipital por detrás):

Lesiones en dos deportes extremos		
Región	Patinaje	Combate
Mentón	0.4	0
Nariz	0.1	0
Pómulos	0.1	0.4
Frente	0.1	0
Temporal	0.1	0.4
Parietal	0.1	0.2
Occipital	0.1	0

3 . a) Calcular el rango y el puesto de cada partido si se clasifican de acuerdo al número de votos en cientos en la localidad F. b) Halle la mediana. c) ¿Cómo le fue a Trabajo y paz? (Ayuda: ejemplo 5-7, pág. 2). Los datos son:

Partidos y sus votos en la localidad F	
Nombre	**Número de votos, cientos**
Naturaleza en paz	7
Seguridad en la calle	6
Trabajo para jóvenes	11
Comida para indigentes	1
Muerte a violadores	7
Estudio gratis	4
Trabajo y paz	4
Más policía	7
No a la corrupción	8

4 . Ejercicio para practicar los 4 métodos exactos para calcular la media, la varianza, la desviación. Tenemos 5 listas de datos que corresponden a las notas de unos estudiantes en 5 cursos. Para aprender, tome una lista cualquiera y sobre ella aplique los 4 métodos. Todos deben dar lo mismo. Para mecanizar y agilizar tome otra lista y haga lo mismo. (Ayuda: ejemplos 16 y 17 pág. 9, y ejemplo 19 pág. 10). Para el método 4 haga una tabla con x, F, $(x-\bar{x})^2$, $F(x-\bar{x})^2$ con su suma y reemplace en la fórmula.

a) 3, 4, 5, 3, 2, 2, 3, 4.

b) 3, 4, 3, 4, 5, 3, 2, 1.

c) 1, 4, 2, 3, 4, 2, 3, 3, 2.

d) 4, 5, 4, 5, 3, 4, 2, 1, 3.

e) 2, 3, 2, 3, 1, 2, 3.

5 . Considere la silletería de un teatro: como regla general y sin excepciones todas las sillas son iguales. ¿Cómo es eso posible, dado que los asistentes a los teatros tienen tan enorme variabilidad morfológica? ¿Es apropiado decir que el diseño de las sillas se basa en un único descriptor o estadígrafo morfológico y que de éste sólo se tomó el de una sola persona? ¿Cuál podría ser dicho descriptor? Ahora piense en los zapatos que usamos: casi son sobre medidas. ¿Cómo se explica la infinita tolerancia de la variabilidad en el caso del teatro y la total ausencia de ella en el caso de los zapatos?

6 . En una universidad las notas definitivas van de 0 a 5 en unidades, décimas y centécimas. En cierto curso, las diversas componentes de nota se manejan así: 4 parciales y una nota de tareas, cada uno 13 %. Total 65 %. Un proyecto 10 %, examen final: 25 %.

a) ¿Cuál es la nota que corresponde al 100 % si las notas de un estudiante son, en el orden enunciado antes, 3.4, 4.5, 4.4, 2.3, 4.6, 3.5, 1.8? A la nota definitiva se le llama *promedio ponderado*.

Ayuda: imaginemos una clase en la cual la nota definitiva sale de 4 notas todas de igual valor. Las notas son a, b, c, d. La nota definitiva es el promedio:

$(a+b+c+d)/4 = a/4 + b/4 + c/4 + d/4 = a \times 0{,}25 + b \times 0{,}25 + c \times 0{,}25 + d \times 0{,}25$.

Vemos que todas las notas están ponderadas por el mismo valor y que esto se puede generalizar a cualquier situación.

b) Las notas de un estudiante en sus 4 parciales, tareas, y proyecto son respectivamente 4.5, 1.2, 2.3, 3.4, 4.5, 3.8. ¿Cuánto debe sacar en su final para que su 100 % le quede en 3.76 para compensar lo perdido en otras materias?

c) Un estudiante termina su 100 % en 4.24. Necesita una centésima para llegar a lo que necesita que es 4.25. ¿Por cuánto debe hacer un reclamo en el examen final para lograr su objetivo?

7 . El porcentaje de software ilegal en 25 empresas

20 24 51 47 57
56 74 52 58 59
47 58 74 43 71
71 96 22 39 75
47 68 64 55 61
Para los datos dados:

a) Haga un diagrama de tallo y hojas (Ejemplo 25 pág. 16). Dibuje la campana que mejor se ajuste a dicho diagrama y estime gráficamente la media y la desviación. Con la media y desviación halladas y usando el criterio de las dos desviaciones, determine gráficamente los datos extremos. (Ayuda: ejemplos 12 hasta 15 pág. 6).

b) Calcule los cuartiles. Recuerde que el cuartil 1 es el percentil 25 con $k = 25$, el cuartil 2 es el percentil 50 con $k = 50$ y el 3 tiene $k = 75$. (Ejemplo 24, pág. 14).

c) Usar el diagrama de tallo para agrupar por intervalos adecuados y definir las marcas de clase. Hallar las tablas de frecuencias absolutas y relativas. Usar la tabla de frecuencias absolutas para calcular por fórmulas la media, la varianza, la desviación y el coeficiente de variación (ejemplo 23 pág. 13). Compare los resultados exactos con los aproximados sacados de la campana: deben ser muy parecidos.

d) Aconséjele una medida al Gobierno, el cual debe fortalecer la legalidad del software: si no aumentamos la legalidad, nuestros ingenieros no tendrán motivación para trabajar, ni habrá competencia por la exelencia y siempre seremos esclavos de las potencias.

8 . Los datos siguientes se refieren al número de preguntas en una semana que unos alumnos hacen al profesor: 3, 4, 3, 4, 0, 0, 2, 1, 0, 4, 3,0, 4, 0, 5, 3, 0, 0, 4, 5, 12, 0.

a) Hallar numéricamente los cuartiles (Ejemplo 24, pág. 14), el resumen de 5 valores y dibujar el boxplot o gráfica de cuadro (Ejemplo 29, pág. 18).

b) Halle los datos extremos según el criterio del rango intercuartilar. Sobre el boxplot, gráfica de cuadro o de cuadro y bigotes señale la frontera de los datos normales (Ejemplo 30, pág. 19).

c) Dibuje un histograma de los datos. Trace la campana que mejor se ajuste, dibuje y mida la desviación. Utilícela para determinar los valores extremos según el criterio de las dos desviaciones (Ayuda: ejemplos 12, 13 y 15 pág. 6). ¿Son congruentes las dos metodologías, la de las dos desviaciones y la del rango intercuarticular, para determinar los valores extremos? Recomendación: cuando el histograma se ajuste bastante bien a una campana, use el criterio de las dos desviaciones para tomar decisiones en relación a los valores extremos. De lo contrario, use el criterio del rango intercuartilar.

d) Quite los valores extremos según el criterio del rango intercuartilar y calcule la media, desviación y coeficiente de variación de los datos restantes. ¿Hay una variación significativa?

9 . Las notas de unos muchachos en un examen fueron: 1, 2.1, 1.8, 1.5, 5, 1.8. Ellos descubrieron un outlier, lo quitaron, recalcularon la media que les dió 1.64 y con ella en mano se fueron a la pelea con el profesor. Pero éste les respondió: “en estos datos no hay ningún dato extremo pues los 6 datos tienen media 2.2 y desviación 1.42. Usando el criterio de las dos desviaciones, lo normal es lo que está entre -0.64 y 5.04. Como el posible candidato para outlier es 5 y éste está en la región de lo normal, no hay outliers”. Al oírlo, los muchachos primero se rieron, luego se carcajiaron y después comenzaron a llorar. Cuando ya se cansaron, uno de ellos dijo: ¡pensemos un poquito! Y en efecto, así lo hicieron y ganaron una estupenda batalla. ¿Qué fue lo que pensaron? (Ej 18, pág. 10).

10 . De un grupo de 120 personas entrevistadas, 80 piensan que la política sobre el manejo de la crisis de los secuestrados por los violentos debe incluir una acción militar. Hay 20 que opinan que debe incluirse una negociación. Y hay 8 que creen que uno puede jugar con las dos opciones: negociación y fuerza. Plantear y calcular las diversas probabilidades que surgen de esta situación, siguiendo el ejemplo 39, pág. 27.

11 . De 150 países que asistirán a la próxima ronda sobre energía limpia (no contaminante), 70 tienen yacimientos de petróleo y 130 tienen fuentes hidroeléctricas. Y hay 4 que no tienen ninguna de las dos. Plantear y calcular las diversas probabilidades que surgen de esta situación, siguiendo el ejemplo 39, pág. 27. Ayuda: haga una disyunción de Venn como en el siguiente dibujo en el cual aparecen 3 incógnitas por lo que necesitamos 3 ecuaciones independientes.

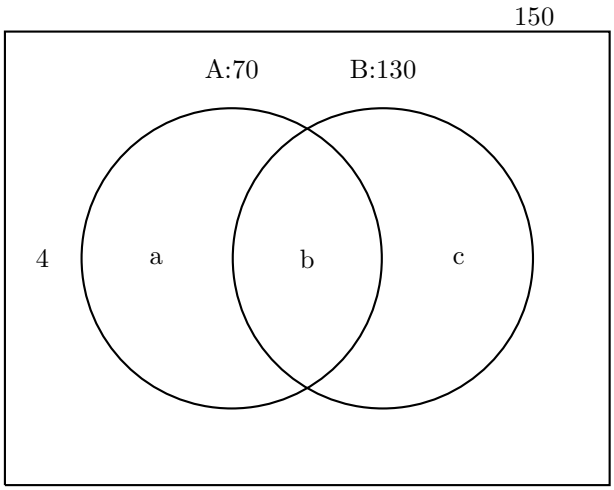


Figura 3.1. Diagrama de disyunciones de Venn: la suma de las partes da el todo.

12 . La probabilidad de que Adelita llegue tarde a clase es 2/9 y tiene 3 clases por semana. Desarrolle la $Bi(3, 2/9)$: árbol generador, lista de eventos, tabla de probabilidades, histograma, media, varianza, desviación. Interprete los resultados en términos de las veces que Adelita llega tarde a clase. (Ayuda: ejemplos 46 -49, pág. 32.)

13 . Utilice un método corto para hallar la tabla de probabilidades de la $Bi(6, 1/3)$. Haga el histograma correspondiente. Luego calcule la media, la varianza y la desviación. Interprete los resultados en términos de las veces que Paquito sufrirá para tomarse la sopita teniendo en cuenta 6 días de la semana, pues el domingo comen a la carta. (Ayuda: ejemplo 55 y siguientes, pág. 41).

14 . Dos amigas se hacen juntas en los parciales. Ellas siempre están muy bien preparadas, sin embargo se miran la una a la otra. ¿Por qué? No es para copiarse sino para saber quién va más rápido. La probabilidad de que se miren en un parcial es de 0.2. a) Si en el semestre responden 4 parciales, ¿cuál es la probabilidad de que se miren al menos 1 vez y se ganen la enemistad del profesor, quien nada sabe sobre la sicología de las amigas? Ayuda: Reformule el enunciado en términos de una binomial y observe que la probabilidad solicitada es $p(1) + p(2) + p(3) + p(4)$. b) ¿cuál es la probabilidad de que se miren al menos 1 vez si se tienen en cuenta los 24 parciales de todas las carreras del semestre? ¿Las echarán de la U? Ayuda: Hay una forma extracorta de resolver la parte a). Encuéntrela y use la misma metodología para resolver la parte b).

15 . a) Arme una historia para cada una de las siguientes frecuencias relativas o demuestre que no es posible. Ejemplos: 2/5 (historia = de cada 5 personas en la cafetería, 2 eran hombres), 1/9 (de cada 9 mujeres, una es partidaria de que el aborto sea estricta materia de conciencia personal). Ejercicios: 3/7, -4/9, 0.89, -1.9, 2.3.

b) Arme una telenovela para cada una de las sumas de las siguientes frecuencias relativas o demuestre que no es posible: 1/2 + 1/2 (la mitad eran hombres, la otra mitad mujeres), 1/2 + 1/3 + 1/6, 0.5 + 0.6, -0.3 + 0.8, 0.2+0.5+0.3, 0.4 -0.4.

c) ¿Cuáles son las propiedades básicas que cumplen las frecuencias relativas y sus sumas? d) ¿Cuáles son las propiedades básicas de una probabilidad sabiendo que puede entenderse como la frecuencia relativa medida sobre una muestra muy grande y al azar?

16 . Los estudiantes fueron llegando a clase, unos temprano, otros a tiempo y otros tarde, pero como el profesor no llegó, se fueron a las y cuarto. a) ¿Cuál es la decisión a tomar, el investigador, la muestra, la variable a medir, los eventos, la población? b) ¿Qué tan racional es la actitud de los estudiantes? c) La variable a medir: ¿Es aleatoria? ¿Es discreta? ¿Es continua? d) ¿Cuál es la decisión deseable? e) ¿Cómo se contrasta la muestra con la población y con lo que se cree al momento de tomar la decisión de irse o no de clase? Ayuda: Ej. 31 -34, pág 20.

17 . Nosotros conocemos cuatro métodos para calcular la varianza de una lista de datos. Todos los métodos dan la misma respuesta. Pero nunca lo hemos probado. Y sin embargo, a algunos estudiantes les intriga saber cómo se demuestra tal equivalencia. Para ayudarlos, vamos a demostrar que

$s^2 = \frac{\Sigma(x-\bar{X})^2}{n-1} = \frac{n\sum x^2-(\sum x)^2}{n(n-1)}$.

El ejercicio consiste en explicar todos los pasos.

Para ello, comenzamos desarrollando una resta al cuadrado: cuadrado del primero menos dos veces el primero por el segundo más el cuadrado del segundo.

$s^2 = \frac{\Sigma(x-\bar{X})^2}{n-1} = \frac{\Sigma(x^2-2x\bar{X}+\bar{X}^2)}{n-1}$

Ahora aplicamos la ley de sumas que dice $\Sigma(a+b) = \Sigma a + \sum b$. Tenemos:

$\frac{\Sigma(x^2-2x\bar{X}+\bar{X}^2)}{n-1} = \frac{\Sigma x^2-\Sigma 2x\bar{X}+\Sigma\bar{X}^2}{n-1}$.

Ahora aplicamos la ley que dice que una constante puede factorizarse:

$\Sigma ka = k\Sigma a$. Obtenemos:

$\frac{\Sigma x^2-\Sigma 2x\bar{X}+\Sigma\bar{X}^2}{n-1} = \frac{\Sigma x^2-2\bar{X}\Sigma x+\bar{X}^2\Sigma 1}{n-1}$

El último término sale de la siguiente igualdad: $\Sigma\bar{X}^2 = \bar{X}^2\Sigma 1$, puesto que al sacar la constante $\bar{X}^2$ queda uno. ¿Cuánto vale la suma de uno? Es la suma de n unos porque la suma corre sobre todos los términos de la lista que son n. Por tanto:

$\frac{\Sigma x^2-2\bar{X}\Sigma x+\bar{X}^2\Sigma 1}{n-1} = \frac{\Sigma x^2-2\bar{X}\Sigma x+n\bar{X}^2}{n-1}$

Ahora bien, Sabemos que la media es $\bar{X} = \Sigma x/n$ por tanto:

$\frac{\Sigma x^2-2\bar{X}\Sigma x+n\bar{X}^2}{n-1} = \frac{\Sigma x^2-2(\Sigma x/n)\Sigma x+n(\Sigma x/n)^2}{n-1}$

$= \frac{\Sigma x^2-2(\Sigma x)^2/n)+(\Sigma x)^2/n}{n-1} = \frac{n\Sigma x^2-(\Sigma x)^2}{n(n-1)}$.

3.2. Parcial I: con hoja de fórmulas hecha personalmente

63 ***Objetivo:*** *Proponemos enseguida varios parciales diferentes para que así pueda el estudiante aprender a calibrar el tiempo que siempre es muy corto: 45 minutos. Para el parcial no se puede sacar libro ni apuntes, se puede usar la calculadora y una hoja de fórmulas que cada quién ha hecho previamente a gusto personal.*

Consejos: *cuando haya que calcular una media y/o una desviación de unos datos cuantitativos, lo primero que hay que hacer es clasificar el problema en uno de los casos siguientes y aplicar la metodología correspondiente:*

1. *Se trata de una lista de datos: ejemplo 17, pág. 9.*

2. *Se trata de muchos datos que deben ser agrupados en una tabla de frecuencias absolutas sin aproximar: ejemplo 19, pág. 10.*

3. *Se trata de muchos datos que se agrupan por intervalos en una tabla de frecuencia absolutas: ejemplo 23, pág. 13.*

4. *Se trata de una tabla de frecuencias acumuladas, como en el problema 1 de la versión 2 del primer parcial, numeral 65, pág. 53. En ese caso, lo que hay que hacer es calcular la tabla de frecuencias absolutas y seguir como de costumbre.*

5. *Se trata de una distribución binomial: ejemplo 49, pág. 34. Estas mismas fórmulas sirven para cualquier distribución discreta y para tablas de datos con frecuencias relativas.*

Para el bien de todos, adjuntamos una lista de errores que uno deberá evitar:

1. *No indicó a qué evento se referían las respuestas numéricas.*

2. *Una probabilidad le dio mayor que uno.*

3. *Interpretó mal el enunciado de un problema relacionado con los diagramas de Venn y en consecuencia todo le quedó mal.*

4. *No hizo los diagramas de Venn.*

5. *No indicó los nombres de los conjuntos en los diagramas de Venn.*

6. *El árbol de la binomial le quedó con unas ramas largas y otras cortas, cuando todas deberían tener el mismo tamaño. Como consecuencia, algunas opciones quedaron sin listar.*

7. *Hizo el árbol bien pero contó mal el número de caras en una rama. Por lo tanto, hubo un error en los coeficientes binomiales pues le dio 1 2 4 1 cuando le debió haber dado 1 3 3 1. No descubrió el error pues no tuvo en cuenta que los coeficientes dan simétricos con respecto a su mediana.*

8. *No cumplió con la definición de X de una binomial: Si no se dice nada sobre la X de una binomial, se asume que cuenta el número de caras y en la tabla de probabilidades el valor de X y el exponente de la probabilidad de cara coinciden. Si el valor de X coincide con el exponente de la probabildiad de sello, está mal, a menos que se espeficique que X representa el número de sellos.*

9. *No verificó que la suma de las probabilidades en una tabla de la binomial debe dar uno y por tanto no descubrió un error en los cálculos.*

10. *No insertó los coeficientes binomiales en la tabla de las probabilidades de la binomial.*

11. *Hizo el triángulo de Pascal pero no lo utilizó para la tabla de probabilidades.*

12. *La tabla de la binomial le resultó bimodal (con dos montañitas) y siguió como si nada pues no tuvo en cuenta que una binomial siempre es unimodal (una sola montañita). Además, la suma de las probabilidades no fue uno, error que no identificó.*

13. *Utilizó coeficientes binomiales equivocados y no dijo de dónde los sacó, pues no sabe reproducir el triángulo de Pascal.*

14. *Se equivocó en una multiplicación con la calculadora y le dio una respuesta un poquito diferente a la verdadera. El error se hubiese detectado si mentalmente hubiese juzgado la respuesta. Por ejemplo, 23 por 30 no puede dar 1700 sino un número algo mayor que 600.*

15. *Se equivocó en los cálculos de la media y la desviación, le dio una respuesta muy diferente de la verdadera pero no lo detectó por no haber usado el histograma como referencia.*

16. *No dio el valor de la media sino que dijo que estaba en un intervalo.*

17. *Utilizó una n equivocada en el cálculo de la media y varianza.*

18. *Para calcular la media y la varianza utilizó sobre las marcas de clase el método para una lista de datos cuando debió haber utilizado el que corresponde a una tabla de frecuencias absolutas.*

19. *Calculó la media a partir de frecuencias acumuladas sin enterarse de que primero debía sacar la tabla de frecuencias absolutas.*

20. *Hizo el histograma de frecuencias acumuladas cuando debío haber hecho el de absolutas.*

21. *Confunde p(A/B) con p(B/A): el denominador indica el universo, el conjunto de referencia.*

22. *No terminó los problemas por cuanto no tiene la costumbre de hacer tareas todos los días e inmediatamente que las asignan. Ni mucho menos le interesa ir adelantado pues no se ha enterado de que esta materia la puede dejar en cinco.*

64 ***Versión 1***

1. Calcular el rango y el puesto de cada persona si se clasifican de acuerdo a la nota obtenida en un examen. Halle la mediana. Halle la nota y el grupo de alumnos que define el percentil 45. Halle la nota y el grupo de alumnos que definen el cuartil 3. ¿Cómo le fue a Pedro?

Muchachos y sus notas	
Nombre	**Puntaje en Examen**
Camila Fernández	15
Pablo Rodríguez	26
Juana Maldonado	39
Mónica Bernal	21
Andrés Umaña	43
Juan Pérez	37
Diana Suárez	42
Pedro Hernández	46
María Puentes	37
Felipe Rueda	48

2. El peso en kilos de un grupo de 25 muchachas fue:
60 64 50 57 72
56 64 52 48 57
47 48 64 63 81
71 46 72 59 65
67 58 63 45 51
Para los datos dados:
a) Agrupar por intervalos adecuados y definir las marcas de clase.
b) Hallar las tablas de frecuencias absolutas y relativas.
c) Hacer el histograma y usarlo para estimar la media y la desviación.
d) Calcular por fórmula para datos agrupados la media, la varianza, la desviación y el coeficiente de variación.
e) Calcule la moda y la mediana.

3. $X \sim Bi(3, 2/7)$ a) Desarrolle: árbol generador, lista de eventos, tabla de probabilidades, histograma, media, varianza, desviación. b) ¿Cuál es la probabilidad de que salgan al menos 2 eventos distinguidos (dos caras)?

65 ***Versión 2***

1. A partir de una tabla de frecuencias absolutas se puede calcular la **tabla de frecuencias acumuladas** sumando todo lo que está por debajo. Ejemplo: supongamos que tenemos la tabla siguiente de frecuencias absolutas que clasifica unos jugadores de tenis de mesa por rangos. La notación que usamos para los intervalos es así: el extremo se incluye en el intervalo si se cierra con paréntesis rectangular y no se incluye si se cierra con paréntesis redondo. Así, el dato 10 se incluye en el segundo intervalo y no en el primero.

Puntajes	
Intervalos	**Frecuencia absoluta**
[0, 10)	1
[10, 15)	5
[15, 20)	7
[20, 25)	15
[25, 30)	8
[30, 35)	2

La tabla asociada de frecuencia acumulada es la siguiente:

Puntajes	
Intervalos	**Frecuencia acumulada**
[0, 10)	1
[0, 15)	6
[0, 20)	13
[0, 25)	28
[0, 30)	36
[0, 35)	38

Cuado Juan llegue a leer la tabla dirá: yo tengo 27 puntos, por tanto clasifico en el intervalo [0,30) y soy mejor que 28 jugadores como mínimo y como máximo 35.

La siguiente es una tabla de frecuencia acumulada de las edades de unos padres de familia:

Edades de padres	
Intervalos	**No. acumulado de padres**
[35, 40)	7
[35, 45)	12
[35, 50)	17
[35, 55)	45
[35, 60)	52
[35, 65)	58

a) Calcule la tabla de frecuencias absolutas. Ejemplo: El segundo intervalo es [40, 45) y su frecuencia absoluta es 12 -7 = 5. Tome como marca de clase el borde inferior de cada intervalo. Ej: la marca del segundo intervalo es 40.

b) Dibuje el histograma.

c) Calcule la media y la varianza por gráfica y por fórmulas.

d) Calcule el coeficiente de variación.

e) Haga una interpretación.

2. Use un método corto para hallar la tabla de probabilidades de la $Bi(5, 1/3)$ junto con su media, varianza y desviación. Haga el histograma correspondiente. Extienda la tabla para calcular la media, la varianza y la desviación. Los dos métodos deben coincidir. ¿Cuál es la probabilidad de que salgan **a lo sumo** (como máximo) dos caras?

3. De un grupo de 150 estudiantes, 110 piensan seguir hasta el posgrado en esta misma universidad inmediatamente después de terminar pregrado, 70 de los cuales quieren trabajar al mismo tiempo. Hay otros 30 que quieren adquirir una buena experiencia de trabajo antes de meterse a estudiar más. Y hay otros que desean ir al exterior, a viajar o a estudiar. Plantear y calcular las diversas probabilidades que surgen de esta situación.

66 ***Versión 3***

1. Use diagramas de tallo y hojas para representar los siguientes datos de puntajes de rendimiento global de dos grupos de jueces, uno de Medellín y otro de Bogotá. Sobre los diagramas, estime las medianas, las medias, las desviaciones y compárelas. Para cada grupo calcule los 5 valores y dibuje el boxplot o diagrama de caja correspondiente. Calcule el rango intercuartilar, determine los valores extremos y añada los bigotes correspondientes al diagrama de caja.

De Bogotá:	De Medellín:
50 64 60 57 62	35 78 49 45 37
66 65 43 58 57	70 68 87 91 23
45 49 64 73 81	13 82 91 33 23
71 56 62 49 65	45 67 90 22 31
67 48 63 55 51	

2. En una Revista de Ciencias Sociales ya saben que los escritores que salen de la Universidad Colombia se destacan por la buena calidad de sus artículos y que la probabilidad de que escriban algo que valga la pena publicar es 18 % (otras universidades no sobrepasan el 5 %). Esta vez los de la U Colombia mandaron 7 artículos. Calcular las siguientes probabilidades:

   a) de que al menos (mínimo) 3 artículos salgan publicados.

   b) de que a lo sumo (máximo) 4 salgan publicados.

   c) de que todos salgan rechazados.

3. De un grupo de 34 filósofos, 12 son positivistas (que tratan de enmarcar a la filosofía dentro del método científico) y 18 son existencialistas (que le dan valor primordial a la conciencia y al sufrimiento) y hay otros 10 que pertenecen a otras corrientes. Hacer los diagramas de Venn y calcular las probabilidades correspondientes.

67 ***Versión 4: todo sobre la binomial***

1. La probabilidad de que en el transcurso de un examen un estudiante se declare enfermo y abandone la clase es 1/2000. En una clase de 50 estudiantes dos de ellos se declararon enfermos, abandonaron la clase y después solicitaron un examen repositorio. Pero el profesor considera que cuando dos estudiantes se declaran enfermos, al menos uno es por engaño. Para tratar de entender el punto de vista del profesor, calcule la probabilidad de que a) nadie se enferme de los 50, b) uno se enferme, c) dos se enfermen, d) al menos un estudiante de los 50 se enferme. e) Respaldan sus resultados al profesor? Atención: este problema por ser de binomiales grandes no se resuelve ni por árboles ni por triángulo de Pascal: se usa la calculadora o la fórmula para factoriales.

2. Cuando un estudiante abandona un examen porque se declara enfermo, inmediatamente va a la enfermería a que lo atiendan. El médico le da un recibo que consta que vino a la enfermería. Dicha constancia no vale de nada para el profesor. Lo único que vale es un veredicto que declare incapacidad para continuar con el examen. La probabilidad de que el médico de tal concepto es 1/100 pues el médico es partidario de enfrentar la vida con valor. Por otro lado y como regla general, los que se declaran enfermos no tienen un respaldo de trabajo en tareas ni en tablero. En una clase magistral hay 7 de tales estudiantes. a) Usando el triángulo de Pascal, halle la distribución de la cantidad de dichos muchachos a los que se les daría permiso para un nuevo examen suponiendo que todos ellos se declarsen enfermos. Usando fórmulas calcule la media y la desviación. Haga el diagrama de barras correspondiente.b) ¿Cuál es la probabilidad de que se le de permiso al menos a un estudiante de los 7?

3. De los estudiantes de una gran universidad que abandonaron sus exámenes, tres de ellos recibieron permiso para presentar el repositorio por razones médicas. La probabilidad de pasar el examen es 0.01. Es muy baja pues en general no alcanzan a prepararse a última hora. a) Haga el árbol de eventos, úselo para calcular la tabla de probabilidades de pasar, extiéndala para calcular la media y la desviación. Haga el diagrama de barras correspondiente. b) ¿Cuál es la probabilidad de que al menos un estudiante de los 3 pase el examen?

4. Tome una decisión: ¿Es o no inteligente declarse enfermo? Redacte un instructivo que explique concisa y claramente a un estudiante inexperto la razón de su veredicto y el remedio a tanto problema. Especifique qué hacer en un caso de una situación de deterioro extremo de la salud.

68 ***Versión 5***

1. Calcule la media y la desviación de las siguientes notas en un curso usando tres métodos exactos diferentes: 3, 4, 3, 5, 2, 3, 4, 5, 4, 3.

2. a) Use algún método rápido para calcular las probabilidades de la $Bi(10, 1/7)$. b) Use fórmulas para calcular la media, la varianza y desviación y el coeficiente de variación. Use sus resultados para darle un consejo a don José que piensa irse a su ciudad natal a visitar a sus padres para lo cual debe recorrer unos 300Km. Sucede que a don José le gusta adelantar en curva y a lo largo del trayecto hay 10 curvas muy apropiadas para adelantar... y para ganarse una buena multa. La probabilidad de multa es 1/7 en cada curva. Los policías funcionan de manera independiente. Pero don José es muy precavido y ya alistó lo de 3 multas. c) ¿Cuál es la probabilidad de que a don José le alcance para pagar las multas que se gane? d) ¿Cuál es la probabilidad de que se le vaya en multas lo que tenía apartado para divertirse, que es lo equivalente a otras 3 multas? e) ¿Cuál es la probabilidad de que quede hipotecado si se gana las 10 multas?

3. La siguiente es una teoría para explicar por qué la altura de las personas tiene una distribución de tendencia central. La teoría dice: uno tiene genes de crecimiento, los cuales vienen de dos tipos y todo funciona como una $Bi(4, 3/7)$, de tal manera que si cae cara, uno tiene un gen que lo hace crecer 40 cms y si cae sello, el gen lo hace crecer 50 cms. Si un individuo tiene cccc, mide 160; en tanto que el individuo que viene con cscs, mide 180. Calcule la tabla de alturas de las personas con sus probabilidades, la media, la varianza y haga el correspondiente histograma.

69 ***Versión 6***

1. Sea $X \sim Bi(6, 4/7)$. Indicando de dónde sale todo:

   a) Use un método rápido para calcular la tabla de probabilidades

   b) Halle $p(X > 2)$.

   c) Calcule por fórmulas la media y la desviación.

2. El entrenador le dijo al estudiante de clavados: impúlsese y para empezar a girar enrósquese sobre sí mismo como un erizo y para terminar el giro, estire las piernas y métase en el agua como una flecha. El muchacho ya sabe girar pero no sabe entrar al agua como una flecha, no puede hacerlo porque no ha podido saber cuándo deben estirarse las piernas. El papá tomó 30 videos al azar de clavados de expertos y midió el tiempo en milésimas de segundo entre el momento que estuvo con la cabeza hacia abajo y el instante en que estiró las piernas:

   ```
   22 19 40 32 21 34
   68 29 31 45 36 54
   19 45 46 23 76 20
   12 28 33 66 27 55
   31 55 68 55 70 16
   ```

   a) Haga un diagrama de tallo y hojas. Haga la tabla de frecuencias absolutas correspondiente.
   b) Use el diagrama de tallo y hojas para estimar la media y la desviación por medio de la campana que mejor se ajusta a los datos.

c) Utilice la tabla de frecuencias absolutas para calcular la media, la desviación y el coeficiente de variación por fórmulas.
d) Use el criterio de las dos desviaciones para determinar si hay valores extremos. Decirlos.
e) Utilice el diagrama para estimar la mediana, el cuartil uno y el 3.
f) Utilice el criterio del rango intercuartilar para determinar si hay valores extremos.
g) Dibuje el boxplot con sus bigotes.
h) De un consejo en cuanto al momento de estirar las piernas. Atención: es imposible saber cuándo han pasado, digamos, 20 milisegundos. Pero se puede saber cuándo los ojos apuntan sobre un referente externo.

El tiempo resultará traumáticamente corto en éste y en todos los parciales para el que sabe hacer las cosas pero es lento por no haber hecho muchos ejercicios para agilizar, ni por haber exigido muchos quices para quitarse el intenso frío que congela en los exámenes.

Capítulo 4

La normal

70 ***Objetivo: Aprenderemos a manejar la distribución normal, la cual es un modelo poblacional de una distribución experimental o muestral de tendencia central.***

4.1. Idealización poblacional de una tendencia central

Un histograma viene de una muestra, la campana que mejor lo aproxima es una idealización que se aplica a toda la población.

71 ***Variable aleatoria***

En estadística se deben tomar decisiones que afectan a toda la población y no sólo a la muestra. Por tanto, como mínimo, los datos deben representar fielmente la población. Para lograrlo, uno debe procurar que la muestra venga al azar para no crear sesgos o tendencias en alguna dirección. Una **v.a. (variable aleatoria)** es una lista de datos impredecibles en el sentido que saber una sección inicial de nada sirve para predecir los que vienen después. Por ejemplo, si uno se para en una calle transitada y registra la edad, uno obtendrá una lista cuyos datos son aleatorios pues en general uno no podrá predecir quien vendrá después de una señora cuarentona.

72 ***La normal***

Una v.a. tiene una **tendencia central** cuando su datos se agrupan formando una montaña de un único pico y cuando practicamente no hay datos muy alejados de ella. Uno puede caricaturizar un histograma de tendencia central que es muestral por una campana que es poblacional.

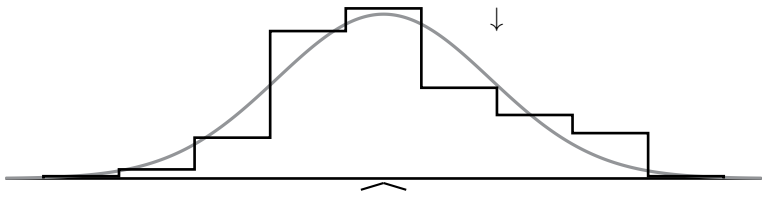


Figura 4.0. Una muestra al azar cuyo histograma sea de tendencia central, como la altura de las personas, da origen a un campana que es una idealización poblacional.

Pero entre todas las campanas, ¿cuál es la mejor? Proponemos el siguiente criterio: la campana que mejor se ajusta a un histograma de tendencia central es aquella que tiene como media la media del histograma y como desviación la desviación de la distribución.

Cuando la campana denota la distribución de una variable aleatoria X (que viene de un muestreo al azar) se denomina la **distribución normal**. Se determina por dos **parámetros** (valores ideales, poblacionales), la media μ (mu) y la desviación estándar σ (sigma). Se denota

$$X \sim N(\mu, \sigma)$$

y se lee: X **es una variable aleatoria con distribución normal con media** μ **y desviación** σ. Tengamos presente que la media de unos datos se denota $\bar{X}$ y su desviación s que son letras latinas. Pero la media de una normal, que es una idealización de un posible infinito número de datos, se denota μ y su desviación σ que son letras griegas.

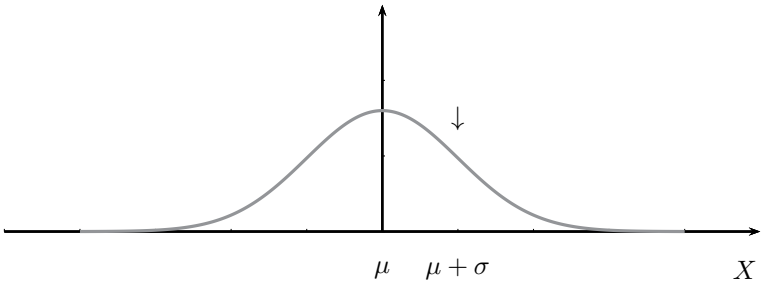


Figura 4.1. La Campana de Gauss. El centro de simetría representa la media. El punto de inflexión superior es donde la parábola se choca con el tobogán superior, el cual se indica con una flechita vertical. Una unidad horizontal representa una desviación estándar, es decir, la distancia entre la media y el punto de inflexión.

La **media** es el lugar sobre el eje horizontal que divide la campana en dos y la **desviación** es la distancia que hay sobre el eje horizontal entre la media y cualquiera de los dos punto de inflexión, donde la parábola se choca con los toboganes.

La campana es la idealización de una distribución de tendencia central y del diagrama de barras de la binomial. Por lo tanto el área bajo la campana se interpreta como una probabilidad y por consiguiente **toda la campana tiene área uno y toda porción de la campana debe tener un área entre cero y uno.**

73 ***Cómo estimar la media y la desviación.*** *Ya sabemos calcular gráficamente tanto la media como la desviación de la campana que mejor aproxima unos datos, lo cual sirve pero requiere mucho trabajo. Cuando uno necesita apreciaciones a la velocidad del rayo, el* ***criterio de la doble mitad*** *es muy sencillo y práctico: se determina el mínimo y el máximo. La media está en la mitad entre el mínimo y el máximo y lo que mida la mitad de la mitad del rango de valores observados es la desviación.*

74 ***Ejemplo. El número de amigos de las primíparas.***

Sea X la v.a. (variable aleatoria) que cuenta el número de amigos de las alumnas de primer semestre. La que menos tiene amigos, tiene 3 y la que más tiene, tiene 11. La media vale el promedio de 11 y 3 que es 7. Entre el mínimo y el máximo hay 11-3 = 8 que es igual al rango de valores observados. Su mitad es 4 y su doble mitad es 2. Por tanto, la desviación vale 2. Sucintamente, $X \sim N(\mu = 7, \sigma = 2)$. Demos una respuesta aproximada a la siguiente pregunta: ¿qué proporción de niñas tienen un número de amigos entre 9 y 11? El área de la región que queda bajo la curva y entre 9 y 11 se denota $p(9 < X < 11)$ y es la probabilidad de que el número de amigos de una alumna de primer semestre esté entre 9 y 11.

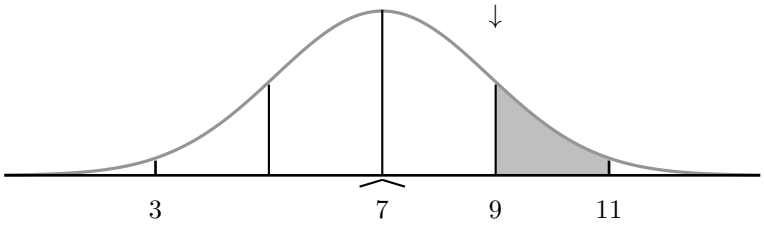


Figura 4.2. El área bajo toda la campana es UNO pues es la probabilidad de todo el mundo. La región que queda entre 9 y 11 tiene un área relativa aproximada que es menor que 1/4 = 0.25, quizá algo parecido a 0.15.

El área señalada cubre aproximadamente un 15 % de la campana. Eso significa que de 100 niñas de primer semestre tomadas al azar, aproximadamente unas 15 tienen entre 9 y 11 amigos de confianza.

4.2. La Z = la normal estándar

El área de una región se interpreta como una probabilidad y matemáticamente corresponde a una integral. La Campana de Gauss es particular en el sentido de que no se puede integrar por fórmulas. Toca hacerlo por trucos de aproximación especiales, lo cual ha resultado un complique. Como todas estas cuentas son tan tediosas, se ha optado por hacer los cálculos una sola vez en la vida y ponerlos en una tabla. Pero hay un problema: uno puede variar la media y la desviación de la campana y obtener una distribución diferente. Por tanto, necesitaríamos un número infinito de tablas. La buena noticia es que sólo se necesita una tabla y precisamente la que se refiere a la normal más sencilla posible:

75 ♣ ***Definición.*** *La normal que tiene media cero y desviación 1 se llama* ***la normal estándar o la Z*** *que cumple:*

$$Z \sim N(\mu = 0, \sigma = 1)$$

y cuya gráfica es la siguiente:

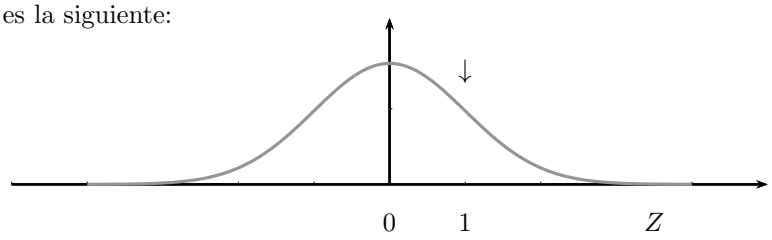


Figura 4.3. La Z es una normal con media cero y desviación estándar uno.

Para hallar el área de una región bajo la campana utilizamos una tabla, la tabla de la Z o de la distribución de Z. En realidad, hay muchas versiones de la tabla de la Z. Por eso, cuando uno vaya a interpretar una tabla, lo primero que hay que hacer es mirar el código de interpretación de la tabla.

76 ***Nuestro código.***

Para cada número z sobre la campana de la normal estándar, nuestra tabla da la probabilidad de obtener un evento menor que z, lo cual es igual al área bajo la curva y a la izquierda de z. Dicha área se denota $\Psi(z)$ (tridente de zeta) y es el área o probabilidad antes de z. A Ψ se le llama **la función de distribución de** Z. `Es el equivalente del percentil`.

77 ***Ejemplo*** *Ilustremos el significado de la igualdad* $\Psi(0) = 0{,}5$*:*

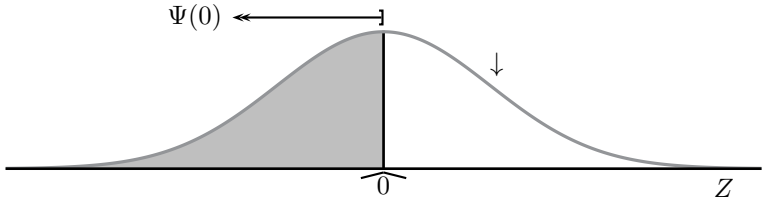


Figura 4.4. $\Psi(0)$ *es el área que está debajo de la campana de la Z y a la izquierda del punto* 0*. Como toda la campana tiene área UNO,* $\Psi(0) = 0{,}5$ *pues es el área de la mitad de la campana, la cual es simétrica. La flechita vertical indica el punto de inflexión positivo que para la Z es 1.*

Supongamos que Z representa el tiempo de llegada a clase que está descrito por una normal con media cero (el tiempo de comienzo de clase, digamos a las 10 en punto) y desviación un minuto. Podemos decir que todos se esfuerzan por llegar a tiempo y que antes de las diez llegan tantas personas como después. Análogamente y teniendo en cuenta que la desviación vale 1, donde está la flechita vertical, $\Psi(1) = 0.8$, que quiere decir que el 80 % de las personas llegan antes del minuto después de haber comenzado clase. Muy pocas personas, menos del 5 %, se retardan más de 2 minutos: $1 - \psi(2) < 0{,}05$.

78 *LA TABLA DE LA Ψ PARA LA Z:*

La tabla tiene 4 columnas, dos para la parte negativa de la Z y dos para la positiva. Esto se debe a que la media de la Z es cero, y por tanto hay valores de z tanto positivos como negativos (la Z representa el eje de la normal estándar, pero z se refiere a un valor específico). Nuestra tabla presenta para cada valor de z un número que es el área o probabilidad antes de z y que se denota $\Psi(z) = p(Z < z)$ (psi de z) y se le llama **función de distribución** de la normal estándar. La gráfica es la siguiente:

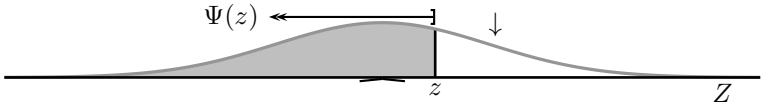


Figura 4.5. Sobre la campana normal estándar, media cero y desviación uno, el área antes de z se denota Ψ(z) (tridente o psi de z) y es el valor reportado por nuestra tabla:

LA Ψ PARA $Z \sim N(0,1)$			
$\Psi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds$			
1 cola: $\Psi = \alpha$, 2 colas $\Psi = \alpha/2$			
z-	$\Psi(z)$	z+	$\Psi(z)$
0	0.5000	0	0.5000
-0.10	0.4602	0.10	0.5398
-0.20	0.4207	0.20	0.5793
-0.30	0.3821	0.30	0.6179
-0.40	0.3446	0.40	0.6554
-0.50	0.3085	0.50	0.6915
-0.60	0.2743	0.60	0.7257
-0.70	0.2420	0.70	0.7580
-0.80	0.2119	0.80	0.7881
-0.90	0.1841	0.90	0.8159
-1.00	0.1587	1.00	0.8413
-1.10	0.1357	1.10	0.8643
-1.20	0.1151	1.20	0.8849
-1.30	0.0968	1.30	0.9032
-1.40	0.0808	1.40	0.9192
-1.50	0.0668	1.50	0.9332
-1.60	0.0548	1.60	0.9452
-1.64	0.05	1.64	0.95
-1.70	0.0446	1.70	0.9554
-1.80	0.0359	1.80	0.9641
-1.90	0.0287	1.90	0.9713
-1.96	0.025	1.96	0.975
-2.00	0.0228	2.00	0.9772
-2.10	0.0179	2.10	0.9821
-2.20	0.0139	2.20	0.9861
-2.30	0.0107	2.30	0.9893
-2.40	0.0082	2.40	0.9918
-2.50	0.0062	2.50	0.9938
-2.60	0.0047	2.60	0.9953
-2.70	0.0035	2.70	0.9965
-2.80	0.0026	2.80	0.9974
-2.90	0.0019	2.90	0.9981
-3.00	0.0013	3.00	0.9987
-3.10	0.0010	3.10	0.9990
-3.80	0.0001	3.80	0.9999
$-\infty$	0	∞	1

79 ***Ejemplo: Hallar Ψ para un z dado.*** *De acuerdo con la tabla de la Ψ se tiene que:*

1. $\Psi(-1.4) = 0.0808$*;*
2. $\Psi(0.3) = 0.6179$*;*
3. $\Psi(2.3) = 0.9893$*.*
4. $\Psi(-30) = 0^+$*, un poquitico más de cero, pues está entre* 0 *y* 0.0001*.*
5. $\Psi(30) = 1^-$*, un poquitico menos que uno, pues está entre* 0.9999 *y* 1*.*

80 ***Ejemplo: Hallar z para un Ψ dado.*** *Hallemos z tal que $\Psi(z) = 0.2420$. En la tabla, columna del Ψ localizamos 0.2420. Vemos que $z = -0.70$. Pero no hay nadie cuyo Ψ sea 12 ni tampoco -4.*

81 ***Aclaraciones para quienes gustan del rigor.*** *Para cada número z sobre la campana de la normal estándar, nuestra tabla da la probabilidad de obtener un evento menor que z, lo cual es igual al área bajo la curva y a la izquierda de z.*

Observemos que el código de interpretación de nuestra tabla está en el primer renglón y dice:

$\Psi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds$

¿Qué significa este jeroglífico?

Para que se nos grabe, repitamos una vez más: la tabla reporta la función de distribución psi de z que es igual al área o integral bajo la Campana de Gauss de la Z y que queda antes de z, desde menos infinito hasta z.

La función de distribución de la Z da para cada z la probabilidad de que sobre la Z un evento sea menor que z.

Y todo esto también se escribe formalmente como sigue. Si

$$f(s) = \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}}$$

denota la campana de la distribución Z, la normal estándar, tenemos:

$$\Psi(z) = p(-\infty < Z < z) = \int_{-\infty}^{c} f(z)d(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds$$

Tengamos en cuenta:

- la Z con mayúsculas denota una variable aleatoria cuyos valores siguen una distribución normal estándar, por ejemplo, la altura de unas personas.
- La z con minúsculas indica un evento de la normal estándar, por ejemplo, la altura de Juan.
- La f que da la fórmula de la campana se llama la **densidad de probabilidad** de la Z.
- La Ψ se llama la **función de distribución** de la normal estándar Z y $\Psi(z)$ da la probabilidad de que un evento sea menor que z.

Decimos que nuestra tabla de la Z es una tabla para la **cola inferior** pues da al área bajo la campana desde menos infinito hasta z sobre el eje horizontal.

82 $\diamond$ ***Teorema del área grande menos el área chica:*** *El área o probabilidad $p(a < z < b)$ es igual al área grande $\Psi(b)$ menos el área chica $\Psi(a)$. Formalmente:*

Si $\Psi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds$*, entonces*

$$p(a < z < b) = \Psi(b) - \Psi(a).$$

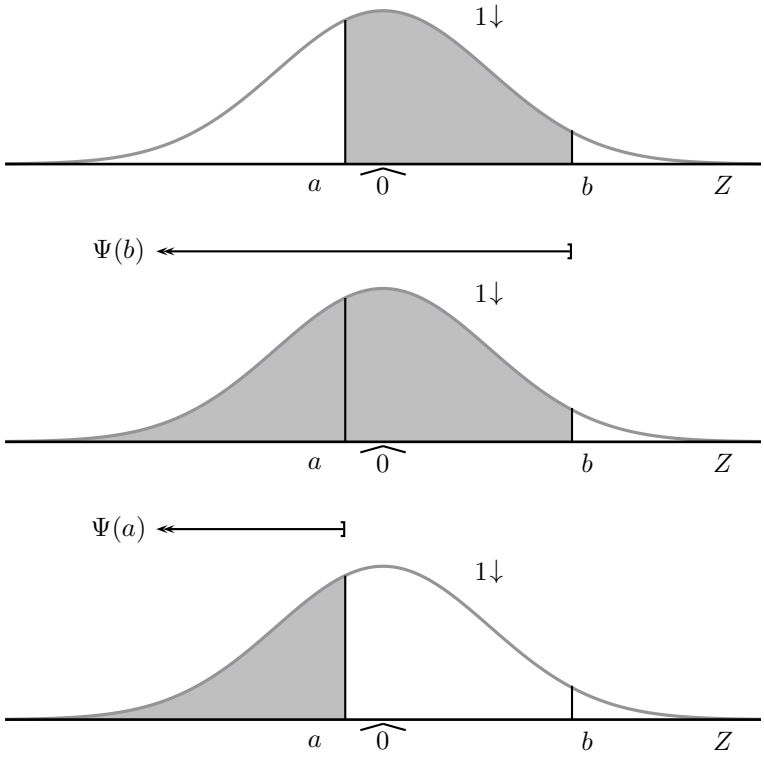


Figura 4.6. El área sombreada en la parte superior es $p(a < Z < b)$, la probabilidad de que Z esté entre a y b. Eso es igual a todo lo que queda antes de b (figura intermedia) menos todo lo que queda antes de a (figura inferior). El área grande, la sombreada en la figura intermedia, es $P(-\infty < Z < b) = \psi(b)$. El área chica, la sombreada en la figura inferior, es $P(-\infty < Z < a) = \Psi(a)$. El área de arriba es igual al área grande menos el área chica. Es decir $p(a < Z < b) = \Psi(b) - \Psi(a)$.

Este teorema nos dice cómo usar la tabla: para saber la probabilidad de que Z tome un valor entre a y b hallamos el área bajo la campana de la Z desde a hasta b. Para ello miramos la tabla para $z = b$ y le restamos el valor de la tabla para $z = a$.

Prueba por integrales. Sin escribir el integrando, que se sobreentiende, podemos decir que:

$$P(a < Z < b) = \int_a^b = 0 + \int_a^b = (\int_{-\infty}^a - \int_{-\infty}^a) + \int_a^b = \int_{-\infty}^a + \int_a^b - \int_{-\infty}^a = \int_{-\infty}^b - \int_{-\infty}^a = \Psi(b) - \Psi(a).$$

83 ***Ejemplo*** *Usemos la tabla de la Ψ y el teorema anterior para calcular $p(-1{,}7 < z < 2{,}9)$.*

Solución: Para hallar la probabilidad o área de $p(-1{,}7 < z < 2{,}9)$ restamos el área grande menos el área chica:

$$p(-1{,}7 < z < 2{,}9) = \Psi(2{,}9) - \Psi(-1{,}7) = 0{,}9981 - 0{,}0446 = 0{,}9535$$

84 ***Investigación*** *Aprenda a usar Excel para calcular $\Psi(z)$ para cualquiera z y con mucha precisión. Excel tiene casi todo lo que veremos en el curso. Sin embargo,* ***Gnumeric*** *tiene fama de ser muy exacto.*

85 ***La tabla de la Z y su Ψ puede reemplazarse por la siguiente gráfica que uno puede reproducir a pulso para reemplazar la tabla.***

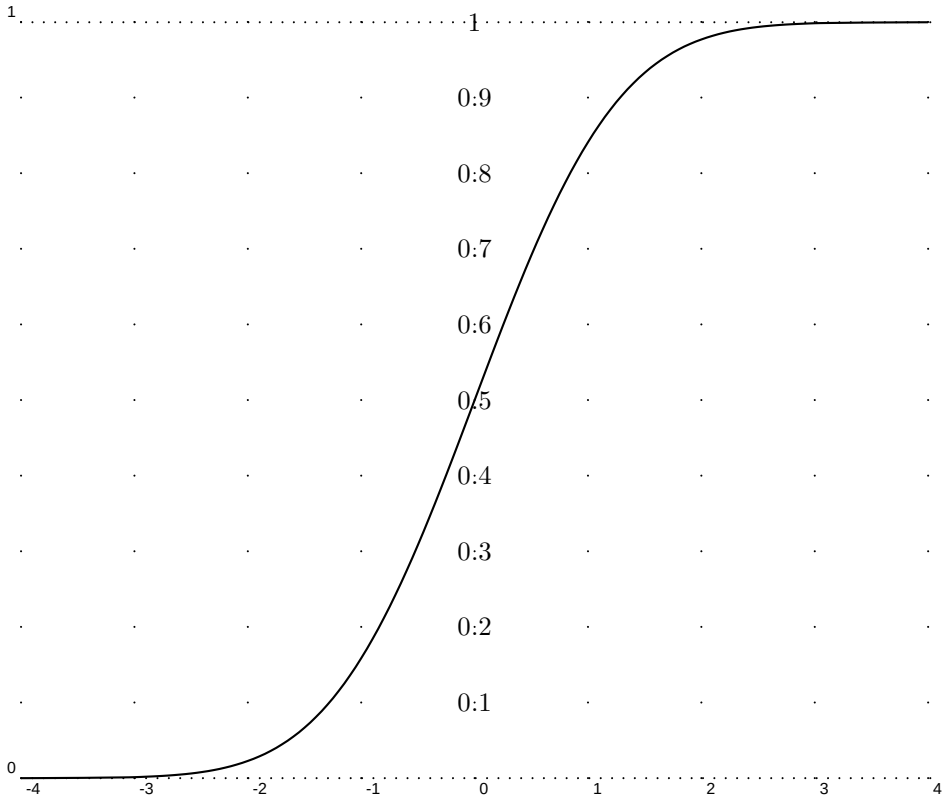


Figura 4.7. La gráfica de $\Psi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds$.

4.3. Por qué se necesita una tabla?

La tabla de la normal es una ayuda para calcular áreas. Pero eso puede hacerse por integración. ¿Por qué no usamos dicha opción en vez de mortificarnos con una tabla? Para entenderlo, nos conviene añadirle un poco de rigor a la discusión:

86 ♣ ***Definición.*** *La Campana de Gauss es la curva dada por la función:*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

donde e representa el número de Euler, cuyo valor es aproximadamente 2,71. Si se grafica esta función, se obtiene una campana, como todas las que hemos visto. Esta función así de enredada es la mejor aproximación continua al diagrama de barras de una binomial y su ajuste es casi perfecto cuando n es grande (mayor que 10) y p se acerca a $1/2$.

En la jerga oficial se dice: la campana es el mejor ajuste continuo al diagrama de barras de la binomial que es una distribución discreta. Decimos que una **v.a. (variable aleatoria) es discreta** cuando cuenta sobre los números enteros, 0, 1, 2, 3, 4 ... y sus negativos: la binomial, que cuenta el número de caras, es discreta. En un juego de dados, si X es lo que salga en la cara hacia arriba, X es discreta. Una **v.a. es continua** cuando sus eventos están asociados al eje real, por ejemplo, la altura de las personas, cuyo reporte exige decimales en varias cifras. Toda normal con su campana es continua.

Sucede que mientras la integral de coseno es seno y la de x^2 es $x^3/3$, no existen fórmulas sencillas para hallar la integral bajo la campana. Lo que existe es una técnica llamada análisis numérico, cuya

aplicación es, como mínimo, engorrosa. Por eso, el uso de una tabla ha resultado muy práctico y ha sido adoptado por todos. Últimamente, los programas de computador que han incorcopado el análisis numérico en sus subrutinas han llegado a ser de mucha utilidad y reemplazan la tabla y dan precisión de 9 decimales.

4.4. La Z como prototipo

Hicimos una tabla para ahorrarnos los cálculos de las áreas bajo la campana de la Z. Pero sucede que variando la media y la desviación se obtienen infinidad de campanas: ¿Necesitaremos una tabla para cada campana? Ahora vamos a ver algo sorprendente: con una sola tabla, la que ya tenemos, es suficiente para saber cualquier área bajo cualquier campana de cualquier distribución normal con cualquier media y con cualquier desviación. Y todo se logra con un ligero cambio de perspectiva. Veámos en pasos simples lo que forma una gran solución.

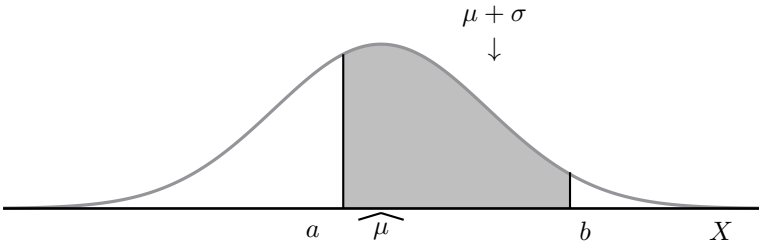


Figura 4.8. El área de la región sombreada es aproximadamente igual a 0.6 o sea algo más de media campana. Pero, ¿cuánto vale la media y cuánto la desviación? Sin importar el valor de la media μ o el de la desviación σ, el área sigue siendo aproximadamente 0.6.

El primer paso es darnos cuenta de que el área de una porción sobre una campana cualquiera depende de la proporción con respecto a toda la campana y no de los valores específicos de la media y la desviación, tal como se ve en la gŕafica acompañante.

El segundo paso es calcar el problema dado en X sobre la Z como sigue:

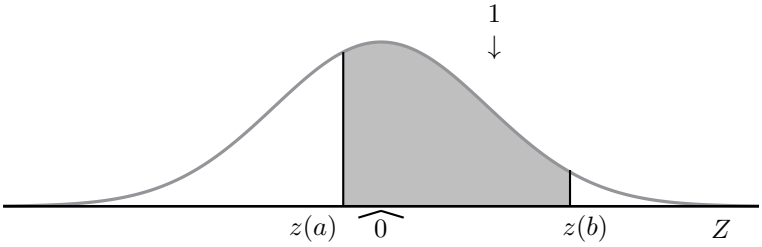


Figura 4.9. El área de la región sombreada sobre la figura anterior para una v.a. $X \sim N(\mu, \sigma)$ es exactamente igual al área de la región sombreada sobre la Z con media cero y desviación 1.

Por tanto, todo lo que nos falta es encontrar la manera de pasar de la X a la Z. Para ello usamos una transformación adecuada:

87 ♣ ***Definición. La transformada-z*** *para calcar la X sobre la Z está dada por*

$z(x) = \frac{x-\mu}{\sigma}$

Observemos que
$z(\mu) = \frac{\mu-\mu}{\sigma} = 0$
$z(\mu+\sigma) = \frac{\mu+\sigma-\mu}{\sigma} = 1.$

Eso quiere decir que la transformada-z cambia la media de la X por la media de la Z que es cero y la desviación de la X por la desviación de la Z que es uno. Eso es lo que significa calcar la X sobre la Z. En general: la $z(x)$ da el número de desviaciones estándar a partir de la media a la que queda x sobre X, signo positivo arriba, signo negativo abajo. Por ejemplo, si $x = \mu + 2\sigma$
$z(\mu+2\sigma) = \frac{\mu+2\sigma-\mu}{\sigma} = +2.$

88 ***Ejemplo*** *La siguiente es una cadena de datos que corresponden a la longitud del cabello de unas señoritas, el cual se ha medido en la nuca: 11, 12.4, 9.5, 8.1, 10.4, 11.1, 9.2, 7.9. La media de estos datos es 9.95 = 10 y su desviación es 1.1. Ejecutemos la* ***transformada*** *z:* $z = (x-\mu)/\sigma = (x-10)/1{,}1$ *y hallemos el percentil asociado.*

La longitud del cabello de las señoritas tiene una tendencia central -dada por el largo del cabello que esté de moda- pero no sigue exactamente una distribución normal, pues la cola superior, la de cabello muy largo, es mucho más extendida y poblada que la cola inferior, donde figuran las que tienen cabello extracorto. Sin embargo, a cualquier distribución se le puede aplicar cualquier cambio de escala y en particular la transformada-z, a la cual se le puede asociar el valor Ψ de la tabla:

La transformada-z produce sobre los datos del largo del cabello:

$z(11) = (11-10)/1{,}1 = 0{,}90$; $\Psi(0{,}90) = 0{,}8159$
$z(12{,}4) = (12{,}4-10)/1{,}1 = 2{,}18$; $\Psi(2{,}18) = 0{,}98$
$z(9{,}5) = (9{,}5-10)/1{,}1 = -0{,}45$; $\Psi(-0{,}45) = 0{,}32$
$z(8{,}1) = (8{,}1-10)/1{,}1 = -1{,}73$; $\Psi(-1{,}73) = 0{,}04$
$z(10{,}4) = (10{,}4-10)/1{,}1 = 0{,}36$; $\Psi(0{,}36) = 0{,}64$
$z(11{,}1) = (11{,}1-10)/1{,}1 = 1$; $\Psi(1) = 0{,}8413$
$z(9{,}2) = (9{,}2-10)/1{,}1 = -0{,}73$; $\Psi(-0{,}73) = 0{,}23$
$z(7{,}9) = (7{,}9-10)/1{,}1 = -1{,}91$; $\Psi(-1{,}91) = 0{,}02$

La transformada-z cuenta a partir de la media en unidades de desviaciones estándar. El signo es el usual: positivo a la derecha, negativo a la izquierda. En el caso del cabello, signo positivo indica cabello más largo que la media, signo negativo = cabello más corto que la media. Vemos que la transformada-z funciona como un **medidor de la discrepancia o diferencia** entre un valor dado y la media, pero reportada en unidades de desviaciones estándar.

¿Cuál es el significado de Ψ? Si los datos se ajustan bien a una campana, el valor de Ψ de cada dato es una buena aproximación al **percentil del dato**, es decir, a la proporción de datos menores que él. Para decisiones administrativas, esta aproximación podría ser más que suficiente. Por ejemplo, el 30 % de las señoritas tienen el cabello más corto que 9.5cm, pero sólo el 4 % lo tienen más corto que 8.1cm. Por tanto, hay un gran mercado para productos de auto-peluquería.

89 ♣ ***Definición.*** *Mientras que la transformada z toma un evento x y le asocia su equivalente z, tenemos la* ***transformada-****x, que toma un valor de z y le asocia el valor de x correspondiente:*
Si $z = (x-\mu)/\sigma$ *entonces:*
$z\sigma = x - \mu$
$x = \mu + z\sigma$

90 ***Ejemplo*** *Sea* $X \sim N(5,\ 0.7)$*. Hallemos algunas transformadas x:*
a) Si $z = -1$*, que significa que hay un evento localizado una desviación abajo de la media, entonces el valor de x es* $x = 5 + (-1)(0{,}7) = 4{,}3$*.*
b) Si z =1.2, que significa que hay un evento localizado 1.2 desviaciones arriba de la media, entonces el valor de x es $x = 5 + [1{,}2(0{,}7)] = 5 + 0{,}84 = 5{,}84$*.*

4.5. El protocolo-z

Aprendamos ahora a usar la trasformada-z para calcular diversas probabilidades tales como la probabilidad de ser mayor que algo o de estar entre dos valores dados.

91 ***Ejemplo sobre novias y adolescentes.*** *El tiempo en horas por día que unos adolescentes dedican a la novia* $X \sim N(2,\ 0.7)$*. Hallar la proporción de muchachos a quienes se acusa de vivir en la casa de la novia por dedicarle más de 2 horas y media al día.*

Como 2.5 está arriba de la media, la proporción debe ser menor que 0.5, y por estar antes de una desviación arriba de la media, ésta debe ser algo así como 0.3. Para saberlo exactamente, calculamos la transformada-z de 2.5:

$$z(2{,}5) = \tfrac{x-\mu}{\sigma} = \tfrac{2{,}5-2}{0{,}7} = 0{,}71.$$

El Ψ nos da 0.75. Esta es la proporción de los que están abajo de $z = 0{,}71$. Como nos preguntan los que están por arriba, restamos este valor de uno, la probabilidad de todo el universo, y nos da 0.29. Es decir, por tanto, el 29 % de muchachos viven en la casa de la novia y no ayudan para pagar los servicios.

92 ***Ejemplo*** *Los ingresos de una oficina de abogados.*

Una oficina de 2 abogados recibe mensualmente por concepto de servicios de asesoría y representación una suma X que es una v.a. normalmente distribuida con media 8 y desviación 3. Para que los abogados se sientan a gusto con sus ingresos deben ganar más de 7 (para sus gastos) pero no más de 13, pues por ahora no les resulta rentable volverse secuestrables. Nuestro problema es calcular la probabilidad de que el ingreso X esté entre 7 y 13, lo cual se escribe así:

$$\text{Hallar } p(7 < X < 13) \text{ dado que } X \sim N(\mu = 8, \sigma = 3)$$

En general, si $X \sim \mathrm{N}(\mu, \sigma)$, $p(a < X < b)$ es igual al área bajo la campana entre a y b.

Con lo que veremos enseguida podremos calcular $p(7 < X < 13)$ para el ingreso de la oficina de abogados. Si dicha probabilidad resultase ser, digamos, 0.20, eso querría decir que el 20 % de los meses el dúo estaría conforme y el 80 % inconforme con las entradas, bien sea porque son muy bajas, y no les alcanzaría para sus gastos, o porque son demasiado altas y afectaría su sentimiento de seguridad. Pero si dicho valor fuese mayor a 0.90, entonces diríamos que por lo menos el 90 % de los meses el dúo estaría conforme con sus ingresos, y que probablemente estarían muy contentos con su profesión.

93 ***Protocolo-z para calcular*** $p(a < X < b)$

Supongamos que una v.a. X tiene media μ y desviación σ. Para calcular $p(a < X < b)$ se aplica el siguiente protocolo que ilustraremos para hallar $p(7 < X < 13)$ dado que $X \sim N(\mu = 8, \sigma = 3)$.

Paso 1. Se dibuja la campana de X, con media μ y desviación σ, con a y b.

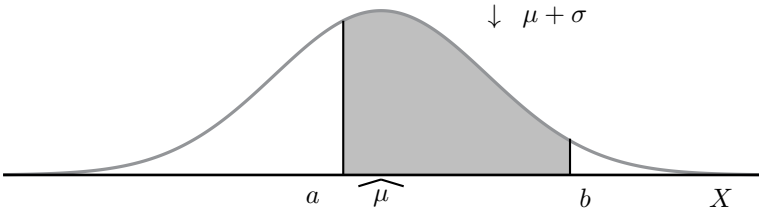


Figura 4.9. El área entre a y b se interpreta como la probabilidad de que un evento esté entre a y b, la cual se denota $p(a < X < b)$.

Paso 2. Se dibuja otra campana igualita, pero con escala horizontal diferente: la media es cero y la desviación es UNO, la cual es la **distribución normal estándar,** denotada Z. La campana Z tiene la misma forma que la campana de X, pero no las mismas unidades. Por eso, lo que antes eran a y b ahora son nuevas coordenadas denotadas $z(a)$ y $z(b)$.

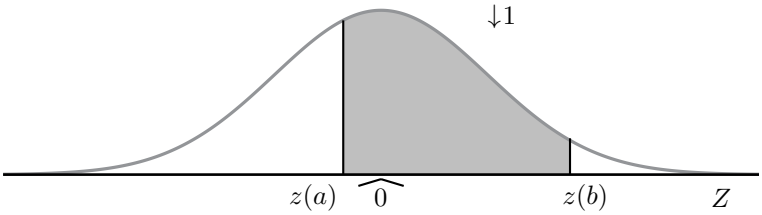


Figura 4.10. El área entre a y b en la campana de la X es exactamente igual al area entre $z(a)$ y $z(b)$ en la campana de la Z, es decir: $p(a < X < b) = p(z(a) < Z < z(b))$.

Paso 3. Para calcular la probabilidad de que X esté entre a y b calculamos la probabilidad de que Z esté entre $z(a)$ y $z(b)$ puesto que las áreas son iguales, lo cual se escribe como:

$$p(a < X < b) = p(z(a) < Z < z(b)).$$

Paso 4. ¿Cómo hallamos las coordenadas Z de a y de b, $z(a)$ y $z(b)$? Para ello utilizamos la **La transformada-z** , como sigue:

$z(x) = \frac{x-\mu}{\sigma}$.

En particular, para $a = 7$, $b = 13$, $\mu = 8$ y $\sigma = 3$ tenemos

$z(7) = \frac{7-\mu}{\sigma} = \frac{7-8}{3} = \frac{-1}{3} = -0{,}333.$
$z(13) = \frac{13-\mu}{\sigma} = \frac{13-8}{3} = \frac{5}{3} = 1{,}666.$

Ahora bien:
$p(a < X < b) = p(z(a) < Z < z(b))$ que se transforma en
$p(7 < X < 13) = p(-0{,}333 < Z < 1{,}666)$.

Paso 5. El área dentro de la campana de la Z que va desde $z(a)$ hasta $z(b)$ es igual al área grande $\Psi(z(b))$ menos el área chica $\Psi(z(a))$. Concretamente:

$$p(a < X < b) = p(z(a) < Z < z(b)) = \Psi(z(b)) - \Psi(z(a)).$$

Para nuestro caso, tenemos:
$p(7 < X < 13) = p(-0{,}333 < Z < 1{,}666) = \Psi(1{,}666) - \Psi(-0{,}333)$
En la tabla tomamos los valores más cercanos. Tenemos que $\Psi(1{,}7) = 0{,}9554$ y que $\Psi(-0{,}30) = 0{,}3821$. Podemos responder entonces:
$p(7 < X < 13) = \Psi(1{,}666) - \Psi(-0{,}333) = 0{,}9554 - 0{,}3821 = 0{,}5733$
Conclusión, en esta oficina viven quejándose bien porque creen ganar muy poco o también porque creen ganar mucho (seguramente tienen que pagar muchos impuestos).

Podemos resumir todo muy oficialmente:

94 ◇ ***Teorema-z:*** *Todas las campanas son lo mismo. Es decir:*
Si X sigue una distribución normal con media μ y desviación σ, entonces, usando la transformada-z definida por $z(x) = \frac{x-\mu}{\sigma}$, se tiene que X queda transformada en una Z o normal estándar (media cero y desviación 1).

Y además:

$$P(a < X < b) = P(z(a) < Z < z(b)) = \Psi(z(b)) - \Psi(z(a))$$

lo cual se lee: el área bajo la campana de una v.a. X entre dos puntos a y b es exactamente lo mismo que la correspondiente área bajo la campana de la Z entre la transformada-z de a y la transformada-z de b, la cual es igual al área grande menos el área chica.

95 ***Ejemplo*** *Sea X una variable aleatoria distribuida normalmente con media 8 y desviación estándar 2. Si nosotros reescalamos esta variable por*

$$z = z(x) = \frac{x-8}{2}$$

estamos cambiando la media antigua, 8, por la media nueva cero, y dividiendo las antiguas unidades por 2. Así, la campana que antes tenía $\mu = 8$ y $\sigma = 2$ tiene ahora, después del reescalamiento, media cero y desviación estándar UNO.

96 ***Ejemplo: Teorema del área grande menos área chica.*** *En referencia al sueldo de un profesional de servicios, hallamos $p(1 < X < 7)$ dado que $X \sim N(\mu = 4, \sigma = 2)$.*

Solución: Primero calcamos el problema original de la X sobre la Z con ayuda de la transformada z:
$z = (x - \mu)/\sigma$
$z(1) = \frac{1-4}{2} = -1{,}5$
$z(7) = \frac{7-4}{2} = 1{,}5$
$p(1 < X < 7) = p(z(1) < Z < z(7)) = p(-1{,}5 < Z < 1{,}5)$
Una vez sobre la Z, restamos el área grande menos el área chica:
$p(1 < X < 7) = \Psi(1{,}5) - \Psi(-1{,}5) = 0{,}9332 - 0{,}0668 = 0{,}8664.$
Interpretación: nuestro profesional de servicios estará satisfecho con sus ingresos el 86 % de los meses.

97 ***Ejemplo*** *Sea $X \sim N(\mu = 2{,}5, \sigma = 0{,}8)$, calculemos $P(1{,}1 < X < 7{,}3)$*

Solución: $z(1{,}1) = (1{,}1 - 2{,}5)/0{,}8 = -1,75,$
$z(7{,}3) = (7{,}3 - 2{,}5)/0{,}8 = 6$
$P(1{,}1 < X < 7{,}3) = P(-1{,}75 < Z < 6) = \Psi(6) - \Psi(-1{,}75) = 1 - 0{,}0402$
$= 0{,}9598$

98 ***Ejemplo*** *Un joven, cuyo papá tiene una finca sin oficio, notó que a las muchachas de las ciudades grandes les gusta tener perros de hermoso pelaje. Para una línea dada de perros, la belleza del pelo depende de varias variables, una de los cuales es su longitud X. Supongamos que X se distribuye normalmente con media 2.8 y desviación estándar 0.7, donde las unidades son centímetros. La longitud de un pelo hermoso está entre 3.2 y 4. ¿Cuántos perros debe criar el joven para conseguir 40 perros con la longitud apropiada del pelo?*

En primer lugar, nosotros calculamos la probabilidad de estar entre 3.2 y 4. Para ello, notamos que cuando la distribución es continua, **la probabilidad de un cierto intervalo es el área bajo la curva y encima del intervalo**. Cuando la curva es integrable, el área se representa por una integral. Para matar la curiosidad, escribámosla explícitamente:

$$P(3{,}2 < X < 4) = \int_{3{,}2}^{4} \frac{1}{0{,}7\sqrt{2\pi}} e^{-\frac{(x-2{,}8)^2}{2(0{,}7)^2}} dx$$

Ahora, nosotros usamos la transformada-z

$$z(x) = \frac{x-2{,}8}{0{,}7}$$

Ahora bien:

$z(4) = \frac{4-2{,}8}{0{,}7} = 1{,}7$
$z(3{,}2) = \frac{3{,}2-2{,}8}{0{,}7} = 0{,}57$

Puesto que la transformada-z simplemente calca áreas, el área bajo la campana de la X entre 3.2 y 4 es lo mismo que el área bajo la campana Z entre 0.57 y 1.7. Por una vez en la vida, vemos cómo se escribe eso con integrales:

$$P(3{,}2 < X < 4) = \int_{3{,}2}^{4} \frac{1}{0{,}7\sqrt{2\pi}} e^{-\frac{(x-2{,}8)^2}{2(0{,}7)^2}} dx = = \int_{0{,}57}^{1{,}7} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = P(0{,}57 < Z < 1{,}7)$$

En futuras ocasiones basta con escribir:

$$P(3{,}2 < X < 4) = P(0{,}57 < Z < 1{,}7) = \Psi(1{,}7) - \Psi(0{,}57) = 0{,}9554 - 0{,}70 = 0{,}25$$

Esto significa que el 25 % de los perros, o la cuarta parte, tienen pelo con la longitud conveniente para que luzcan hermosos. Como necesitamos 40 perros de pelo hermoso para poder sellar un negocio con una tienda, debemos multiplicar ese porcentaje por una cantidad Q desconocida, tal que $0{,}25Q = 40$. Por consiguiente, $Q = 40/0{,}25 = 160$.

99 ***Ejemplo: Problema modelo de las camisas.*** *Discutamos un problema real que queda en la interfase de la biología, la cultura y la industria de la moda: decidir el número de unidades por talla para una prenda dada para una población humana. Observemos que los de talla S son mucho menos que los de talla M y por eso es inteligente hacer de cada talla un número de prendas proporcional a la gente que usa dicha talla.*

El cuerpo humano tiene una variabilidad muy amplia en las medidas de sus muchos diversos componentes, tanto que cada cuerpo es único y diferente de los otros. Pero una compañía no puede producir a bajo costo, digamos, una camisa especial para cada persona. Es más, los hombres son, en general, muy tolerantes con las medidas de sus camisas y todo lo que requieren es que les quede holgada. Como consecuencia, existe una única variable o **descriptor** que sirve como indicador de la talla de la camisa de un hombre: el ancho de la parte de la espalda que queda de hombro a hombro. A este descriptor lo llamamos *espalda* y es el que usaremos.

Una compañía que produce vestidos decide confeccionar camisas para una población de 2.500 hombres, lo que se mide es la variable aleatoria X dada por lo que mida en centímetros la espalda del cliente. X tiene una distribución normal con media 40 centímetros y desviación estándar 5 cm. Hay cuatro tallas: S,M,L,XL. Calcular el número de camisas por talla si el rango de cada talla es
S: $X < 32$
M: $32 \leq X < 45$
L: $45 \leq X < 53$
XL: $X \geq 53$.

Solución. **Paso 1.** Dibujamos el problema (para mejor visibilidad, el borde de la talla XL se ha corrido a la izquierda):

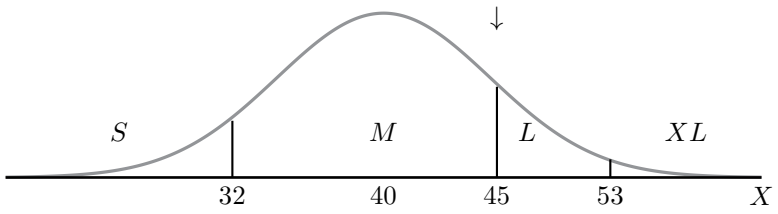


Figura 4.11. Camisas para 2500 hombres que visten 4 tallas.

Paso 2: Calcamos el problema sobre la Z. Para ello calculamos la transformada-z correspondiente a los bordes de los intervalos que definen las tallas:

$z(32) = \frac{32-40}{5} = -1{,}6$
$z(45) = \frac{45-40}{5} = 1$
$z(53) = \frac{53-40}{5} = 2{,}6$

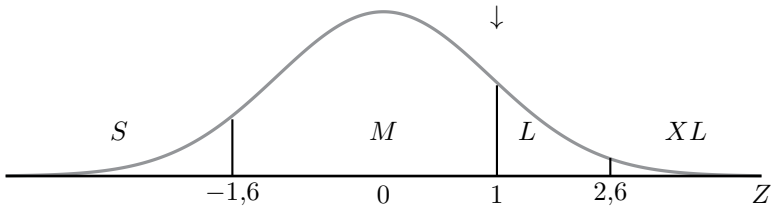


Figura 4.12. El problema sobre la X es calcado sobre la Z.
Paso 3: Calculamos la Ψ de cada z:
$\Psi(-1{,}6) = 0{,}0548$
$\Psi(1) = 0{,}8413$
$\Psi(2{,}6) = 0{,}9953$

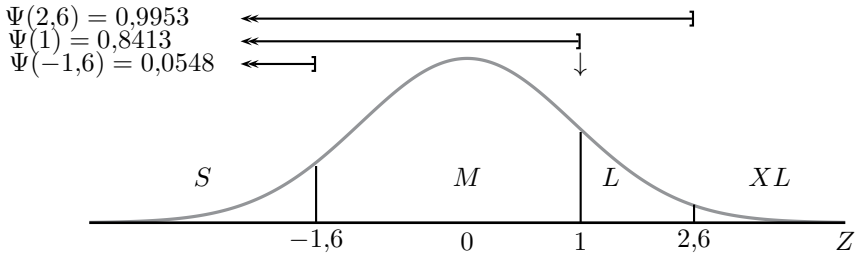


Figura 4.13. Hallamos Ψ para cada límite.

Paso 4: Calculamos la probabilidad de cada talla utilizando el teorema del área grande menos el área chica:

S: $\Psi(z(32)) - \Psi(z(-\infty)) = \Psi(-1{,}6) - \Psi(-\infty) = 0{,}0548 - 0 = 0{,}0548$

M: $\Psi(z(45)) - \Psi(z(32)) = 0{,}8413 - 0{,}0548 = 0{,}7865$

L: $\Psi(z(53)) - \Psi(z(45)) = 0{,}9953 - 0{,}8413 = 0{,}154$

XL: $\Psi(z(\infty)) - \Psi(z(53)) = 1 - 0{,}9953 = 0{,}0047$

Paso 5: Pasamos de probabilidades a número de camisas:

S: $0.0548 \times 2500 = 137$

M: $0.7865 \times 2500 = 1966$

L: $0.154 \times 2500 = 385$

XL: $0.0047 \times 2500 = 12$

Paso 6: Nos aseguramos de que todo va bien. Hay dos cosas que debemos hacer. Primera: sumamos el número de camisas para constatar que tenemos 2500: 137 + 1966 + 385 + 12 = 2500. Salió bien. Segunda: verificamos que la respuesta esté de acuerdo con la gráfica de la campana: esperamos muy pocos individuos XL, pocos de S, un poco más de L y muchos de M. El orden por frecuencias debe ser: $XL < S < L << M$ y eso es precisamente lo que dice nuestra respuesta.

Vamos bien.

Paso 7: Decidimos la marca de clase, es decir, decidimos cuántos centímetros debe tener la espalda da las camisas dependiendo de la talla:

Para la talla S podemos tomar 33, para que no le quede estrecha a nadie.

Para la talla M, tomemos 46.

Para la talla L, que sea 54.

Para la talla XL tengamos en cuenta que son muy pocas las personas de esa talla, es decir que en su mayoría pueden estar por ahí cerca de 53. Por eso ponemos 57 como la marca de clase de la talla XL.

100 ***Ejemplo*** *Los outliers fueron puestos fuera del objetivo de la compañía.*

Una compañía que produce vestidos decide confeccionar camisas para una población de 10.000 hombres, cuyas espaldas tienen una distribución normal con media 40 centímetros y desviación estándar 5 cm. Hay cuatro tamaños: S,M,L,XL, distribuidos simétricamente con respecto a la media. El rango de cada tamaño es una desviación estándar y la media de la distribución queda entre M y L.

La compañía considera que no es rentable confeccionar prendas para gente muy pequeña o muy grande, pues ellos son una minoría. Se ha decidido que esa fracción cubra el 5 % de la población, es decir que sean los que estén alejados de la media más de dos desviaciones estándar. A ellos se les conoce con el nombre genérico de **outliers** (que quedan fuera).

Definir quién es outlier y quién no es una decisión personal. Se aconseja que el grupo de outliers sea pequeño, para que no todo el mundo esté en esa clase, pero no tan pequeño que no se encuentre nadie. Si no se tiene un criterio personal, la proporción de outliers recomendada, por defecto, es del 5 %: son pocos, pero no demasiado pocos, y están en los confines del universo, bien lejos de la media.

La mitad de estos outliers está en la cola izquierda o inferior de la campana y la otra mitad en la cola superior. Así que la compañía cubre el 95 % de la población en la región localizada alrededor

de la media. Nosotros sabemos de memoria que esto implica que la región correspondiente cubre dos desviaciones estándar alrededor de la media. Dicha región obedece la desigualdad $-2 < Z < 2$. De esto nosotros podemos saber las medidas de cada talla:

Como la media está en 40 cm y la desviación estándar vale 5, se deduce que la espalda mínima está a dos desviaciones abajo de la media: $X = 40 - 2(5) = 30$.

La espalda máxima, queda a dos desviaciones arriba de la media $X = 40 + 2(5) = 50$.

Por consiguiente, el rango de la medida de la espalda va de 30cm a 50cm y comprenderá 4 tallas. Los rangos de las tallas, en centímetros, son:

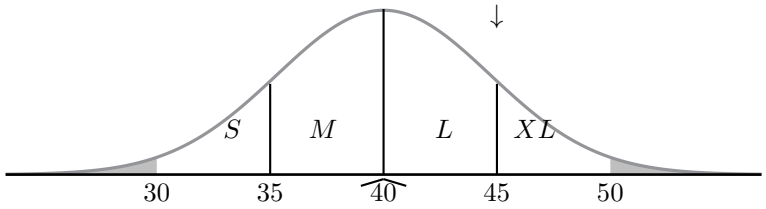


Figura 4.14. Camisas para 9500 hombres que visten 4 tallas. Eran 10000 hombres pero los muy altos y los muy bajos (en sombra) suman el 5 %, o sea 500 hombres, y no entran en el diseño.

Talla S entre una y dos desviaciones abajo de la media, comprende el intervalo: (30,35) y la marca de clase podría medir 37cm de espalda.

Talla M entre la media y una desviación estándar abajo de esta: (35, 40), y la marca de clase podría medir 42cm de espalda.

Similarmente:

Talla L: (40,45) con marca de clase 47 cm de espalda.

Talla XL: (45,50) con marca de clase 52 cm de espalda.

Aplicamos la tecnología aprendida en el ejemplo anterior y llegamos a la respuesta: la compañía confeccionará 1350 camisas con una medida de espalda igual a 35cm, 3400 camisas con una medida de espalda de 40cm, 3400 camisas con una medida de espalda de 45cm, y 1350 camisas con una medida de espalda de 50cm. Nosotros tenemos un total de 1350 +1350 + 3400 +3400 = 2700 + 6800 = 9500 camisas. Nosotros omitimos a 500 hombres que formularán órdenes especiales a sus sastres locales o que de algún modo tratarán de acomodarse al mercado.

101 ***Proyecto*** *Lo que hemos hecho con las camisas, rehacerlo con otra prenda, digamos, un sombrero, una camisa para dama, un sostén, un bikini, una sudadera. Decida primero cuál es el descriptor fundamental de las medidas humanas con relación a su prenda. Suponga una distribución normal para el descriptor elegido y proceda a definir las tallas y a estimar el mercado y la producción requerida por cada talla. Los especialistas proyectan la impresión de que diseñar la producción de bikinis es de lo más difícil del mundo porque al contrario de las camisas para hombre, un descriptor no es suficiente para describirlo y al menos se necesitan 3 o mucho mejor 4 datos independientes. El resultado es que no haya dos mujeres que sea vean iguales y la producción en masa se vuelve una tarea muy complicada a no ser que se usen telas muy elásticas, por lo cual ésa es la tendencia que se impone.*

4.6. Conclusión

Las distribuciones de tendencia central son aquellas cuyos datos son cuantitativos y se agrupan alrededor de la media. Hemos propuesto idealizarlas por la normal, la cual es la envolvente continua de la binomial: podemos sugerir que todo fenómeno que produce una distribución de tendencia central está determinado por varios factores independientes que pueden tomar una de dos opciones. Una normal está completamente definida por su media y su desviación. Gracias a la transformada-z, hemos podido calcular probabilidades asociadas a una normal cualquiera usando una única tabla, la de la normal estándar.

Capítulo 5

Hipótesis

102 ***Objetivo: Teniendo datos que analizar y modelos sintetizadores, lo que viene enseguida en el método científico es el estudio de creencias o sospechas. En realidad, existe un formalismo para describir dentro de la estadística lo que es una creencia o hipótesis y la forma como se la decide aceptar o rechazar. Por misterioso que todo eso pueda parecer, sale de problemas muy naturales y de conceptos muy sencillos. Comenzamos con una disertación sobre ciencias duras vs. ciencias blandas. Luego pasamos a decidir cuándo un evento es estadísticamente interesante, lo cual nos permitirá introducir el formalismo de las hipótesis nula vs. hipótesis alterna.***

5.1. Ciencias duras y blandas

Nuestra época tiene mucho que mostrar a las generaciones venideras. Por una parte, los muertos por guerras y asesinatos de Estado durante el siglo XX pueden sobrepasar los 200 millones de personas. ¿Qué tanto serán 200 millones de asesinatos a escala de la historia del mundo, será mucho o será poco? ¿Se nos recordará por eso? Seguramente sí. Pero también se nos recordará por la ciencia, la cual ha sido el gran aporte de los últimos 4 siglos al desarrollo de la historia del mundo. No es seguro que las generaciones futuras se enorgullezcan de la ciencia, pues eso del desarrollo a veces tiene el sabor de un amargo suicidio planetario. Pero lo que sí es obvio es que la ciencia es de lo mejor que tenemos para mostrar además de ser la calle sin salida en la que nos hemos metido. Es por eso que aprender el método científico es algo obligatorio para toda persona que escoja el ámbito académico como medio para adquirir una profesión para la vida.

El nacimiento de la ciencia se asocia con frecuencia a la física clásica, la cual se distingue porque sus objetos son claramente definidos, claramente distinguibles, muchos de sus descriptores se pueden medir con alta precisión, se les puede poner en condiciones iniciales caprichosas y finamente definidas, se les puede conservar en un ambiente delicadamente controlado, y se puede seguir su conducta por largo tiempo.

Con tantos requisitos fácilmente cumplidos, la física clásica llegó a ser el prototipo de las **ciencias duras**. Sin embargo, tantas facilidades no se cumplen por todas partes. Por ejemplo, la Astronomía que es la base de la cosmología, la que explica el origen y fin del universo, no tiene la opción de poner dos planetas en el lugar que se le antoje ni con la velocidad inicial que desee. Eso la priva de hacer experimentos, y tiene que conformarse con observar y medir. Y sin embargo, nadie se atrevería a decir que la astronomía no es ciencia.

Por tanto, en cualquier disciplina que se pueda observar y medir, también deberá poderse hacer ciencia.

Por otro lado, el sobre-desarrollo de la física clásica generó el advenimiento de la mecánica cuántica. Con eso, la ciencia se vio obligada a reformularse a sí misma en términos estadísticos, pues uno ya no predice la conducta de un objeto específico sino la conducta de poblaciones de objetos.

La introducción de la probabilidad a la ciencia y la capacidad de conformarse con observar y medir son dos características que han permitido que otras disciplinas no físicas también quieran ser llamadas

ciencias. Es por eso que hoy tenemos las Ciencias Sociales, como la sicología: antes se le ponía atención a lo que el paciente sentía, pero ahora se registra y se conduce, se manipula su conducta. Ese encasillamiento de la sicología dentro de la ciencia se atribuye a Skinner.

Pero hay una diferencia entre la astronomía y la Sicología: en la primera podemos hacer una contabilidad de la influencia del mundo alrededor del subsistema que estamos observando, mientras que en la segunda eso no se puede hacer: los hilos que sostienen la personalidad son demasiado sutiles e invisibles para que puedan estirarse, pintarse o cortarse. Por eso, se habla de **ciencias duras y de ciencias blandas**.

Para nosotros, una **ciencia dura** es aquella en la cual la responsabilidad en la toma de decisiones y en la fabricación de la verdad es mínima por parte del investigador. En cambio, en **una ciencia blanda**, la responsabilidad en la toma de decisiones y en la fabricación de la verdad por parte del investigador es máxima: sin investigador no habría verdad, ni decisión, ni ciencia.

Las ciencias más duras son las que tienen más blandito el corazón: no podemos explicar por medio de las ciencias el origen del universo o de la vida sin hacer un enorme esfuerzo de fe.

5.2. La noción de interesante

Lo primero que haremos para entrar en nuestro estudio del aporte de la estadística al método científico es precisar un detalle aparentemente insignificante, pero que se convertirá a lo largo de todo el curso en el pivote alrededor del cual girará la toma de decisiones. Ese detalle se refiere a la clasificación de un evento como **'normal'** o común y corriente, en contraposición a ser **'estadísticamente interesante'** o **'anormal' o 'outlier'**. En los siguientes ejemplos veremos algo que convertiremos en ley: **lo interesante en una distribución central es lo que es muy extremo y tiene muy poca probabilidad de ocurrencia.**

103 ***Ejemplo*** *Los recién nacidos*

El objetivo de un hospital en relación con los partos es aumentar la probabilidad de supervivencia y buena salud tanto del bebé como de la madre. Una variable de alta importancia en la toma de decisiones es el peso del bebé, el cual puede estimarse antes del parto gracias a una ecografía. El peso tiene una distribución de tendencia central que se puede caricaturizar por una normal. Los bebés que se sitúan alrededor de la media nacen naturalmente en un parto de bajo riesgo, generalmente son saludables y salen rápidamente para la casa. Pero los bebés de bajo peso, especialmente los que nacen antes de tiempo, aunque lo hacen naturalmente tienen alta mortalidad. De forma análoga, los bebés muy pesados son preocupantes por el daño que la mamá pueda sufrir en el parto y generalmente nacen por cesárea. En conclusión, los niños o eventos interesantes para un hospital son aquellos que son extremos quedando lejos de la media y que ocurren con baja probabilidad.

Para fijar ideas, pongamos que sí o que tal vez no que el peso de los bebés tiene una distribución normal con media 3 y desviación 0.5 (kilogramos). Es decir, $X \sim N(3, 0.5)$. Si aplicamos el criterio de las 2 desviaciones, los eventos extremos son aquellos que quedan alejados de la media más allá de dos desviaciones. Son los que quedan más abajo de 2 y más arriba de 4. A todo bebé que clasifique en uno de dichos rangos se le prestará atención especial.

104 ***Una y dos colas:*** *Muchas veces uno está interesado en los eventos extremos pero no de ambos lados de la distribución sino de uno solo. Por ejemplo, un director de un equipo masculino de básquet está interesado en los hombres muy altos y los demás es como que ni existen. En este caso se trata de* ***la cola superior****. En cambio algunos varones prefieren las mujeres de baja estatura, por lo que ellos se fijan en* ***la cola inferior*** *de la distribución de altura. Similarmente, existe la creencia de que un jinete de caballos de carrera debe ser preferiblemente de poco peso para no cansar el caballo, es decir, debe estar en* ***la cola inferior*** *de la distribución de peso. Pero para los carros de carrera, una mezcla de inteligencia y valor prima sobre el peso, y hay que aceptar a quien sea inteligente y aguerrido sin importar que sea pesado o liviano. A no ser que sea demasiado extremo.*

Hemos visto de qué manera una clasificación estadística se convierte naturalmente en una luz que ilumina la toma de decisiones administrativas. Pero sucede que implementar las decisiones cuesta dinero, tiempo, esfuerzo. Eso implica que uno se vea obligado a redefinir las barreras de lo que separa lo normal de lo anormal. Por ejemplo, si el jefe del hospital se viera escaso de presupuesto, él podría declarar que lo anormal será lo que quede a tres desviaciones alejados de la media. Y como estos giros son tan normales, tan rutinarios, se ha introducido un lenguaje sencillo y práctico al respecto.

105 ♣ ***Definición: la significancia.*** *Cuando nosotros tenemos una distribución de tendencia central, la primera cosa que nosotros hacemos para clasificar un evento como interesante o no interesante es definir un número llamado* ***alfa****, α, positivo y cercano a cero. Este α se denomina* ***nivel de significancia*** *y es por definición la proporción (está entre cero y uno) de lo que se define por nosotros como* ***'outlier', marginal, extremo, o estadísticamente interesante****.*

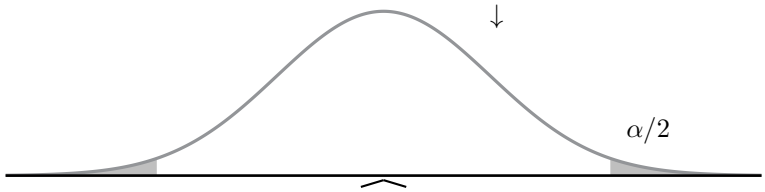


Figura 5.0. Todos los eventos extremos forman un conjunto cuya probabilidad es α que se llama la ***significancia****. Cuando se trata de una cola, toda la significancia va hacia la cola prescrita, pero cuando se trabaja con dos colas, la significancia se reparte, la mitad arriba, la mitad abajo. Con dos colas, como en esta gráfica, se trabaja en cada cola con $\alpha/2$. Con una sola cola, se trabaja con α en la cola prescrita.*

En lugar de dar α, **el nivel de significancia**, nosotros podríamos decir que hemos escogido una **confianza** de $1-\alpha$. Por ejemplo, la confianza asociada a $\alpha = 0{,}05$ es 0.95 o 95 %.

La elección de α es un asunto personal. Uno desearía considerar los casos muy extremos, es decir, tomar α tan pequeño como fuese posible. Pero podría pasar que para encontrar un objeto verdaderamente extremo resulte ser muy difícil y muy caro. Es por eso que **la significancia α se toma por defecto como 0.05**, i.e., **la confianza es del 95 %**. Este valor da un buen compromiso entre factibilidad y eficacia.

106 ***La zeta crítica z_c.*** *La significancia representa sobre la campana un área que indica la proporción de los eventos definidos como extremos. Sobre la campana normal estándar, la zeta correspondiente se saca de la tabla y se denomina* ***z-crítico*** *y se denota z_c. Por ejemplo, si α = 0.0107, este alfa es un área y por tanto se localiza en la columna del tridente. La z-crítica con cola inferior asociada a dicho alfa es z_c = -2.3. Lo que quede más abajo es extremo. Lo que quede más arriba es normal, no interesante.*

107 ***Ejemplo*** *Hallemos la significancia asociada al criterio de las dos desviaciones para eventos extremos. Debido a que la zeta mide el número de desviaciones que un evento queda arriba o abajo de la media, decir que los eventos interesantes son aquellos que quedan alejados de la media más allá de dos desviaciones es lo mismo que decir que la z_c tiene un valor absoluto mayor que 2. Hallar la significancia asociada es lo mismo que hallar la probabilidad de ser más extremo en valor absoluto que 2. La probabilidad de ser menor que -2 nos la da la tabla y es $\Psi(-2) = 0{,}0228$. Como la campana de la Z es simétrica, esa misma proporción queda por encima de $z = 2$. Por tanto, definir los valores extremos con dos desviaciones es lo mismo que trabajar con una significancia de 0.0456. En vez de ser tan exactos,* ***es usual usar la siguiente aproximación: el criterio de las dos desviaciones es equivalente a trabajar con una significancia de 0.05 o con una confianza del 95 %.***

108 ***¿Interesante para qué?***

Un niño enano es muy interesante, es decir, es muy rara la mamá que no lo lleve al médico. ¿Por qué? Pues porque tal vez los médicos descubran qué le pasa al niño, si es que tiene una pésima asimilación de los alimentos o qué. Análogamente, toda mamá se alegra de que hijo sea alto, pero se preocupa si se estira más que un resorte pasando los dos metros cuando sus compañeros todavía no llegan a 1.50: ¿será que mi niño tiene cáncer?

Vemos que los eventos interesantes tienen dos clases de explicaciones: suceden por puro azar dentro de los mecanismos que explican la distribución que se estudie, o bien suceden por otras razones. Y si la ciencia está interesada en progresar, ¿a quién estudiará, a los normales o a los anormales? Tenemos la impresión de que es más fácil descubrir algo nuevo si nos dedicamos al estudio de los anormales.

Por supuesto que los normales pueden serlo no porque el mecanismo que genera la distribución los sitúe cerca de la media, sino por otras razones. Digamos: alguien que tiene relativamente muchos genes de altura pero una mala asimilación bien puede tener una altura normal pero sufrir de algunos quebrantos de salud.

Queremos enfatizar que **la noción de interesante** es algo que no existe *in abstracto*, sino que es una clasificación ligada a personas reales, y como tal es **algo personal y a lo mejor circunstancial**. Por ejemplo, a algunas personas les puede parecer que medir 180 es algo interesante, mientras que para otras eso no tiene ninguna importancia, pues su deporte favorito es el básquet y ahí se es interesante cuando se llega a los dos metros.

El hecho de que la noción de interesante sea algo personal lo convierte en un tema de responsabilidad individual. Como la noción de interesante está en la raíz de la toma de decisiones, por ello enfatizamos: **las decisiones estadísticas contienen un elemento de libre albedrío y por lo tanto cada cual debe declararse responsable de las decisiones que tome sobre todo si se basa en la estadística**. **No hay decisiones científicas sin responsabilidad**.

109 ***Ejemplo*** *David vs. Goliat*

Como regla, los gigantes tienen problemas de salud. Una persona de 2.50m de altura puede necesitar usar un palo para sostenerse y caminar o no se puede mover tan rápido como las personas normales debido a los picos muy altos en la tensión arterial. Así, la jirafa tiene válvulas apropiadas en las arterias del cuello para evitar el daño del cerebro cuando ella baja la cabeza. Ahora, según la **Biblia**, la altura de Goliat era seis codos y un palmo, aproximadamente 2.90m pues un codo mide 45 cm y un palmo 22.

Está claro que Goliat era muy anormal y que quedaba como a 10 desviaciones estándares arriba de la media, lo cual se puede decir de otra manera: Goliat era anormal por lo extremadamente alto con una significancia cercana a cero, es decir, de forma absoluta. Vemos claramente que acá tenemos un caso interesante en la cola superior.

Podemos interpretar la altura extraordinaria de Goliat como una exageración compuesta con fines educativos y también para glorificar la memoria del gran Rey David. Muy a propósito: los escolásticos que escribieron la **Biblia de Jerusalén** dicen que, según los manuscritos tempranos, el joven David no mató a Goliat, sino que éste fue vencido por un hombre valiente cuyo nombre no ha sido registrado.

Pero por otro lado, por exagerados que hayan sido los reporteros bíblicos, es razonable pensar que Goliat era grande comparado con sus amigos y enemigos pero tal vez no al ser comparado con la potencialidad biológica. Eso quiere decir que no corría ningún riesgo biológico por ser algo más grande que la gente alta de su medio. Por eso, pudo ser perfectamente cierto que él era un guerrero desde su juventud. Así, él era ágil y seguro.

Hay otro detalle: David recibió la espada de Goliat con agrado cuando él huía desarmado. Esto significa que una persona de altura normal podía manejar la espada de Goliat. Eso no se debe a que Goliat era casi normal sino a que en ese tiempo no era posible construir espadas largas, pues estas eran hechas de hierro o a lo mejor de bronce (hierro y cobre). El acero no se conocía. Por otro lado, los filisteos, de donde venía Goliat, eran inmigrantes de Creta, donde se conocía mucho de metales, por consecuencia del vulcanismo de la isla. Por eso, Goliat tenía espada y lanza, pero los israelitas tenían espadas de relativamente mala calidad. Por supuesto que para ellos, usar piedra y onda era de lo mejor y un buen palo valía por dos espadas. Apoyamos por lo tanto la tradición de que Goliat fue puesto fuera de combate por una pedrada.

5.3. Eventos interesantes sobre la normal

Primero estudiaremos cómo hallar los eventos interesantes sobre la Z dada una significancia y según el número y modo de colas. Luego pasaremos a una normal cualquiera.

110 ***Ejemplo*** *Señalemos sobre la Z los eventos interesantes según diversas circunstancias.*

- Con $\alpha = 0{,}05$ hallemos los eventos interesantes en la cola inferior. Como es con la cola inferior, toda la significancia va al lado de abajo. La significancia es una probabilidad, por tanto es un área, por lo cual corresponde en la tabla a $\Psi = \alpha = 0{,}05$ y encontramos que el zeta correspondiente, el **zeta crítico**, es $z_c = -1{,}64$. Todos los valores de z inferiores a $-1{,}64$ son interesantes.

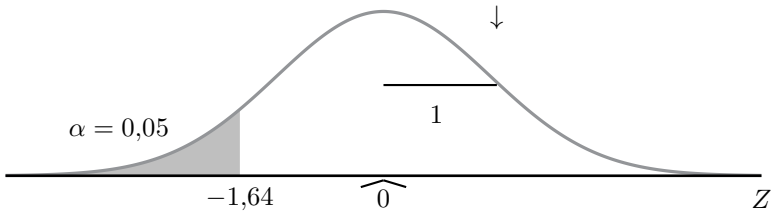


Figura 5.1. Todos los valores menores que $-1{,}6$ *son interesantes si la significancia es* $0{,}05$ *y con la cola inferior.*

- Con $\alpha = 0{,}01$ hallemos los eventos interesantes en la cola inferior. Como es con la cola inferior, toda la significancia va al lado de abajo. Usamos la tabla de la Ψ para hallar el z_c correspondiente a $\Psi = \alpha = 0{,}01$. Es $z_c = -2{,}3$. Todos los valores de z inferiores a -2.3 son interesantes.

- Con $\alpha = 0{,}05$ hallemos los eventos interesantes en la cola superior, por lo que toda la significancia va al lado de arriba. Para hallar la z_c miramos en la tabla el z correspondiente a $\Psi = \alpha = 0{,}05$. Nos da $-1{,}6$, en la cola inferior. Como la campana es simétrica, el z que buscamos es el reflejado, o sea $+1{,}66$. Todos los valores de z superiores a $1{,}64$ son interesantes.

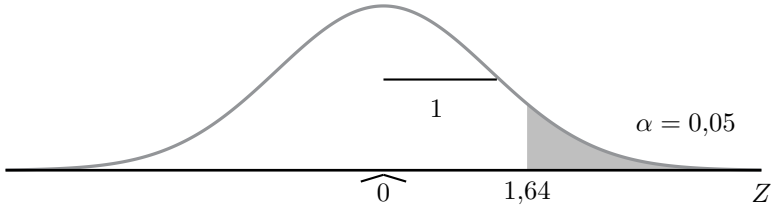


Figura 5.2. Todos los valores de z mayores que $1{,}64$ *son interesantes si la significancia es* $0{,}05$ *y con una cola.*

- Con $\alpha = 0{,}05$ hallemos los eventos interesantes con dos colas. Dividimos la significancia entre dos, por ser dos colas, y nos da $\alpha/2 = 0{,}025$. Vemos de la tabla de la Ψ que el z_c correspondiente a $\Psi = \alpha = 0{,}025$ es $-1{,}96$. Todos los valores de z menores que $-1{,}96$ son interesantes. Pero como se trata de dos colas, también son interesantes los eventos con z mayor que $1{,}96$.

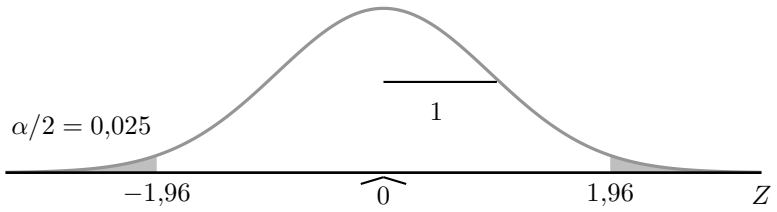


Figura 5.3. Con $\alpha = 0{,}05$ *y dos colas, los valores interesantes sobre Z son los menores que* $-1{,}96$ *y los mayores que* $1{,}96$.

- Con $\alpha = 0{,}04$ hallemos los eventos interesantes en las dos colas: repartimos la significancia entre dos: 0.02. A la cola inferior le toca 0.02 y a la cola superior 0.02. Miramos ahora en la tabla de la Ψ el z correspondiente a $\Psi = \alpha = 0{,}02$. Es $z_c = -2$. Un valor de z menor que -2 es muy interesante y por simetría lo es también cualquier valor mayor que 2.

  No pudimos discernir ningún cambio entre la significancia 0,05 y la de 0,04: eso se debe a que nuestra tabla es muy pobre. En caso de requerirse exactitud, lo mejor es usar un paquete apropiado. Uno que goza de muy buena fama por su precisión es **Gnumeric**, que es gratis, se maneja igual que Excel, y ahora también viene para Windows.

111 ***Ejemplo: Eventos interesantes para normales.*** *Con respecto a la altura de las personas señalemos los eventos interesantes en varias circunstancias.*

- Con $\alpha = 0{,}05$ hallemos los eventos interesantes en la cola inferior si la altura de la población se distribuye como $X \sim N(160, 10)$. Lo que buscamos es la barrera que divide los enanos de la gente normal. Como es con la cola inferior, toda la significancia va al lado de abajo. Usamos la tabla de la Ψ para hallar el z correspondiente a 0,05. Es $z = -1{,}6$. Todos los valores de z inferiores a $-1{,}6$ son interesantes. Lo que tenemos que hacer ahora es aplicar la transformada-x para pasar de la Z a la X: $x = \mu + z\sigma$. Como $z = -1{,}6$, $\sigma = 10$, tenemos $x = 160 + (-1{,}6)(10) = 160 - 16 = 144$.

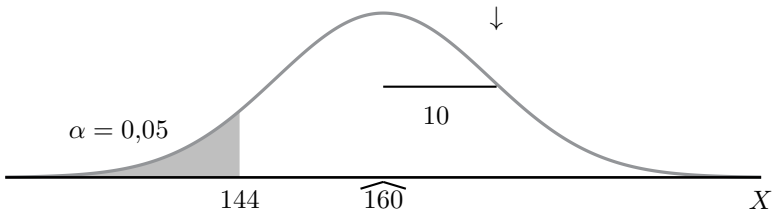


Figura 5.4. Todos los valores menores que 144 *son interesantes si la significancia es* 0,05 *y con la cola inferior.*

- Con $\alpha = 0{,}01$ hallemos los eventos interesantes en la cola inferior si $X \sim N(160, 12)$. Se trata del dueño de un circo que busca enanos. Como es con la cola inferior, toda la significancia va al lado de abajo. Usamos la tabla de la Ψ para hallar el z correspondiente a 0.01. Es $z = -2{,}3$. Por la transformada-x, eso corresponde a una altura $x = 160 - 2{,}3(12) = 160 - 27{,}6 = 132{,}4$. Es decir, que las personas interesantes en este caso son las que miden menos de 132.4.

- Con $\alpha = 0{,}05$ hallemos los eventos interesantes en la cola superior si $X \sim N(180, 5)$. Representamos la población de aspirantes a formar parte de la liga de básquet. De ellos el director debe fijarse en los más altos, por eso tomamos la cola superior y toda la significancia va al lado de arriba. Para hallar la z correspondiente miramos en la tabla el z correspondiente a $\Psi = 0{,}05$. Nos da $-1{,}6$, en la cola inferior. Como la campana es simétrica, el z que buscamos es el reflejado, o sea 1,6. Todos los valores de z superiores a 1,6 son interesantes. Por la transformada-x, la altura correspondiente es $x = \mu + z\sigma = 180 + 1{,}60(5) = 188$ que es la altura mínima para ser interesante (Precaución: ningún director se fija solamente en la altura, sino en una mezcla de altura y destrezas, por ejemplo, hay jugadores bajitos pero que son muy ágiles y que saltan más que uno grande).

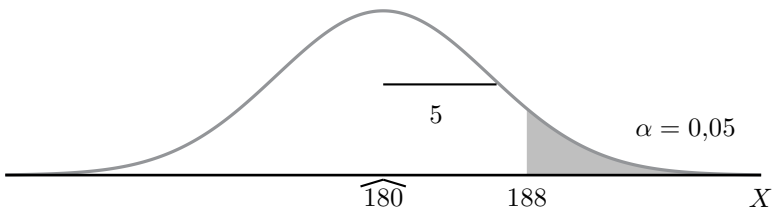


Figura 5.5. Un jugador con altura mayor que 188 es un buen candidato para la liga de básquet.

- Con $\alpha = 0{,}05$ hallemos los eventos interesantes con dos colas si $X \sim N(150, 10)$. Queremos diferenciar los enanos y los gigantes en una población indígena que consume alimentos sin hormonas. Dividimos la significancia entre dos, por ser dos colas, y nos da $\alpha/2 = 0{,}025$. Vemos de la tabla de la Ψ que el z correspondiente es $-1{,}96$. Todos los valores de z menores que $-1{,}96$ son interesantes. Pero como se trata de dos colas, también son interesantes los eventos con z mayor que 1,96. Ahora pasamos de la Z a la X usando la transformada-x: $x = \mu + z\sigma$. Para la cola inferior tenemos $x = 150 - (1{,}96)(10) = 130{,}4$. Para la cola superior, $x = 150 + (1{,}96)(10) = 169{,}6$.

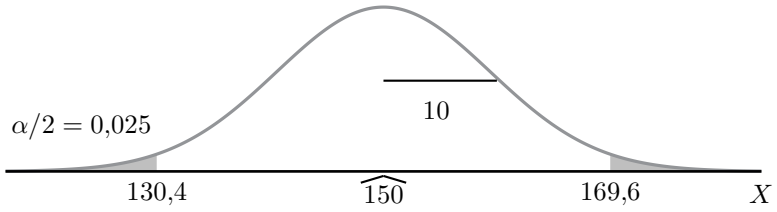


Figura 5.6. Con $\alpha = 0{,}05$ y dos colas, los valores interesantes sobre X son los menores que 130,4 *y los mayores que* 169,6.

- Con $\alpha = 0{,}04$ hallemos los eventos interesantes en las dos colas si $X \sim N(162, 12)$. Se trata de un ser humano desprevenido que mira a su alrededor captando los elementos extremos no importa que sean pequeños o grandes. Como se trata de las dos colas, repartimos la significancia entre dos: 0.02. A la cola inferior le toca 0.02 y a la cola superior 0.02. Miramos ahora en la tabla de la Ψ el z correspondiente a $\Psi = 0{,}02$. Es $z = -2$. Por lo tanto, una persona con altura $x = 162 + (-2)(12) = 162 - 24 = 138$ es declarado como muy interesante. Similarmente, una persona con altura $x = 162 + (2)(12) = 162 + 24 = 186$ es también muy interesante.

112 ***Ejemplo*** *Sea $X \sim N(\mu = 3, \sigma = 0{,}4)$ que puede ser el sueldo de un participante de un grupo de amigos. Hallemos algunos eventos interesantes:*

1. Si la significancia es $\alpha = 0{,}05$ y se toma la cola inferior, para burlarse de los que ganan poquito, se mira en la tabla de la Ψ la z correspondiente a 0.05. Dicho $z = -1{,}6$. Como $z = (x - \mu)/\sigma$ entonces $x = \mu + z\sigma = 3 + (-1{,}6)(0{,}4) = 3 - 0{,}64 = 2{,}36$, lo cual quiere decir que si alguien gana menos de 2.36 se hace blanco de las burlas.

2. Si la significancia es $\alpha = 0{,}01$ y se toma la cola superior, para saber quién va a jactarse, se mira en la tabla de la Ψ la z correspondiente a 1-0.01=.099. Esto se debe a que la tabla de la Ψ acumula lo del lado inferior. En este caso $z = 2{,}4$. Como $z = (x-\mu)/\sigma$ entonces $x = \mu + z\sigma = 3 + (2{,}4)(0{,}4) = 3 + 0{,}96 = 3{,}96$, lo cual quiere decir que si alguien gana más de 3.96 seguramente va a estar insoportable en las reuniones.

3. Si la significancia es $\alpha = 0{,}04$ y se toman la dos colas, para burlarse de los que ganan poquito y de los que ganan mucho, se mira en la tabla de la Ψ la z correspondiente a 0.02, pues hay 0.02 en cada cola. Dicho $z = -2$. Como $z = (x - \mu)/\sigma$ entonces $x = \mu + z\sigma = 3 + (-2)(0{,}4) = 3 - 0{,}8 = 2{,}2$, lo cual quiere decir que si alguien gana menos de 2.2 se hace blanco de las burlas. Pero como se trata de un problema con dos colas, la z también puede ser $+2$. Eso implica que la $x = 3 + (2)(0{,}4) = 3{,}8$, lo que marca el límite de los que ganan mucho en relación con su medio.

113 ***Ejemplo*** *Más allá de α.*

Un hombre es secuestrado pero la Autoridad establece un gran operativo militar para rescatarlo. Por eso la banda delictiva decide matarlo. Pero, debido a que les gusta también la acción y lo inesperado, el jefe decide lanzar una moneda. Si sale cara, lo soltarán; si sale sello, lo matarán. La moneda cae cara y el jefe, con el honor de todo un hombre, cumple la promesa y el secuestrado es liberado. A pesar de

que la probabilidad de ser soltado es 0.5, es decir, es muy grande, y es mucho mayor que $\alpha = 0.05$, el hombre le agradece a Dios y del mismo agradecimiento decide dar un regalo a los pobres equivalente a 10000 dólares. ¿Cómo puede explicarse eso? ¿No es eso una violación contra la teoría de probabilidades, pues qué hay de raro en que pase un evento de probabilidad 1/2? ¿Qué hay de raro?

Respuesta 1: el hombre valora su vida en 2 millones de dólares. Puesto que el salió favorecido con su propia vida, el ganó un millón que es 2 millones multiplicados por la probabilidad de ganar, 0.5. Pero él apenas donó 10000. Así que, su beneficio neto es de 990000. Respuesta 2: ¿Qué tan raro es que un jefe haya decidido dejar la suerte a una moneda y que además se arriesge a cumplir su palabra?

5.4. Eventos interesantes en la binomial

Un **evento es interesante** cuando su probabilidad acumulada en la cola correspondiente es menor que la significancia. Se acumula para que represente un área. Para binomiales pequeñas ($n < 11$), calculamos la tabla de probabilidades y sobre ella decidimos los eventos interesantes. Pero para binomiales grandes, aproximamos su diagrama de barras por una normal y usamos la tabla de la Z.

114 ***Ejemplo: Eventos interesantes para binomiales con*** $p = 1/2$***.***

Dos niños están apostando con monedas a la cara y sello. El juego consiste en lanzar una moneda 8 veces y en contar el número de caras que hayan caído. Y ahora bien: ¿quién gana? Es muy natural definir el juego de tal forma que el que gane sea el que saque más caras (o más sellos). Y, ¿por qué es esa tendencia tan natural? Porque sacar 6, o 7 caras es improbable, tal como lo vemos en la correspondiente tabla de probabilidades (Ej. 59, pág. 43):

$X \sim Bi(8, 1/2)$								
Valores en la primera fila, probabilidades en la segunda								
0	1	2	3	4	5	6	7	8
0.004	0.031	0.109	0.219	0.273	0.219	0.109	0.031	0.004

Observemos que la media es 4: si lanzamos 8 monedas, en promedio tendremos 4 caras, pero claro que pueden salir otras opciones. Sería extraordinario que salieran 8 caras, lo mismo que 8 sellos.

Hallemos ahora formalmente los eventos interesantes en varios casos.

- No hay eventos interesantes en la $Bi(8, 1/2)$ con significancia $\alpha = 0{,}001$ =0.1 % y con la cola superior. Por ejemplo, la significancia de 8, la probabilidad de ser tanto o más extremo que 8, es 0.004 y esto es más grande que 0.001. Con dos colas tampoco hay eventos interesantes pues cualquier valor es más probable que $\alpha/2 = 0{,}0005$.

- Los eventos interesantes en la $Bi(8, 1/2)$ con significancia $\alpha = 0{,}01$ =1 % y con dos colas son 0 y 8, pues estos dos valores tienen significancia $0.004 < 0.005$ ($\alpha/2$) y si añadimos a alguien más se nos pasa la significancia.

- Veamos qué pasa con la $Bi(8, 1/2)$, con significancia $\alpha = 0{,}05$ y con la cola superior. El evento 8 es interesante pues es extremo y muy raro. En efecto, su probabilidad es muy pequeña pues es 0.004 que es menor que 0.05. Para saber si 7 es interesante, sumamos su probabilidad a la de 8. Nos da $0.031 + 0.004 = 0.035$ que también es menor que 0.05. Por lo tanto, 7 es interesante. Para saber si 6 es interesante, sumamos su probabilidad a la de 7 y 8. Tenemos: $0.109 + 0.031 + 0.004 = 0.144$ que es mayor que 0.05. Por tanto, 6 no es interesante y nadie más lo es. Nuestra conclusión es que los datos interesantes con significancia 0.05 y la cola superior son 7 y 8.

- Los eventos interesantes en la $Bi(8, 1/2)$ con significancia $\alpha = 0{,}1$ y con dos colas son 0 y 1 del lado de abajo pues $p(0) = 0{,}004$ y $p(0) + p(1) = 0{,}035 < 0{,}05$ y del lado de arriba son 7 y 8 pues $p(8) = 0{,}004$ y $p(7) + p(8) = 0{,}035 < 0{,}05$.

115 ***Ejemplo:*** $p \neq 1/2$***.*** *Hallar los eventos interesantes con dos colas para* $\alpha = 0{,}05$ *sobre la* $Bi(4, 1/9)$.

Solución exacta: La tabla de probabilidades de $X \sim Bi(4, 1/9)$ es:

$X \sim Bi(4, 1/9)$				
X en la primera fila, probabilidades en la segunda y tercera, acumuladas en la cuarta				
0	1	2	3	4
$1(1/9)^0(8/9)^4$	$4(1/9)^1(8/9)^3$	$6(1/9)^2(8/9)^2$	$4(1/9)^3(8/9)^1$	$1(1/9)^4(8/9)^0$
0.6242	0.3121	0.0585	0.0049	0.0001
0.6242	-	0.0635 = 0.0585+0.0049	0.0050 = 0.0049+0.0001	0.0001

Acumulamos probabilidades desde ambas colas hasta cuando se sobrepase $\alpha/2 = 0{,}025$. En la cola inferior no hay eventos interesantes pues el candidato sería 0 con probabilidad $0{,}6242 > 0{,}025$. Los eventos 3 y 4 son interesantes en la cola superior. Y no hay más.

5.5. Aproximación de la binomial por la normal

La binomial da lugar a cálculos muy tediosos, pero cuando ella es simétrica puede ser aproximada de forma excelente por una campana, que de hecho es mucho más simple.

116 ◇ ***Teorema.*** *La binomial $Bi(n, p)$ se aproxima por la normal con media $\mu = np$ y desviación $\sigma = \sqrt{np(1-p)}$ y la aproximación es tanto mejor cuanto más grande sea n y cuanto más cerca esté p de 1/2. Para valores de p muy extremos, cerca de 0 ó de 1, la aproximación puede ser abusiva, pues dichas binomiales son muy asimétricas.*

117 ***Ejemplo*** *Aproximemos la $Bi(24, 1/2)$ por la Campana de Gauss.*

Sea $X \sim Bi(24, 1/2)$. Su valor esperado es $np = 24(1/2) = 12$ y su varianza es $\sigma^2 = np(1-p) = 24/4 = 6$. La aproximación de X por una Campana de Gauss debe hacerse teniendo en cuenta que la media sea 12 y la desviación $\sqrt{6} = 2{,}45$. Para ver como se aplica eso, calculemos $p(X = 14)$. De acuerdo al coeficiente binomial (ej. 61 pág. 44):

$$p(X = 14) = \tfrac{24!}{14!10!}(1/2)^{14}(1/2)^{10} = \tfrac{24!}{14!10!}(1/2)^{24} = 0{,}1168$$

Usemos ahora la Campana de Gauss:

$$p(X = x) = \tfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

$$p(X = 14) = \tfrac{1}{\sqrt{6}\sqrt{(2)(3{,}14)}}(2{,}7183)^{-\frac{1}{2}\frac{(14-12)^2}{6}} = \tfrac{1}{6{,}13}(2{,}7172)^{-0{,}3333} = 0{,}1169$$

Vemos que la aproximación normal es casi perfecta.

118 ***Ejemplo.*** *Hagamos un estudio gráfico de la aproximación de la $Bi(8, 1/2)$ por la Campana de Gauss.*

Esta distribución tiene media $\mu = np = 8(1/2) = 4$, varianza $\sigma^2 = npq = 8(1/2)(1/2) = 2$ y desviación igual a $\sigma = 1{,}4$. Para calcular $p(k)$, la probabilidad de k caras, tenemos en cuenta que hay un coeficiente C, que lo sacamos del triángulo de Pascal con $n = 8$, y además la probabilidad de cara a la k multiplicada por la probabilidad de sello a la $8 - k$. Por todo queda

$$p(k) = C(1/2)^k(1/2)^{8-k} = C(1/2)^8 = C(1/256) = C/256$$

La tabla y el histograma correspondiente son los siguientes:

$X \sim Bi(8, 1/2)$								
Valores en la primera fila, probabilidades en la segunda								
0	1	2	3	4	5	6	7	8
0.004	0.031	0.109	0.219	0.273	0.219	0.109	0.031	0.004

La gráfica siguiente nos muestra qué tan bien se aproxima el diagrama de barras de la binomial $Bi(8, 1/2)$ por la normal de media 4 y desviación 1.4.

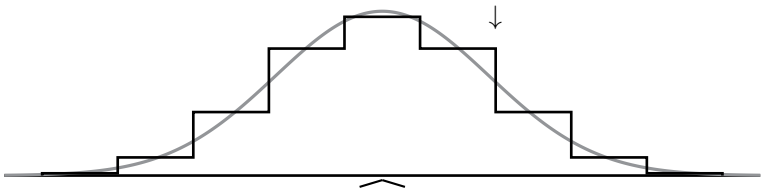


Figura 5.7. La $Bi(8, 1/2)$ queda muy bien aproximada por una campana.

Vemos que para $n = 8$, la curva continua en forma de campana se ajusta muy bien a la binomial y prácticamente se desvanece en los valores alejados de la media.

119 ***Ejemplo*** *Hallemos los eventos interesantes para la binomial $Bi(8, 1/2)$ con dos colas y significancia 0.05.*

La solución exacta la sacamos de la tabla de la binomial y nos da que los eventos interesantes con dos colas y significancia 0.05 son 0 y 8, pues su probabilidad total es $0.008 < 0.05$ y si añadimos los siguientes valores, 1 y 2, acumulamos 0.07 que es mayor que 0.05.

Ahora, aproximemos la binomial por la normal. La campana adecuada tiene media $\mu = np$ y desviación $\sigma = \sqrt{np(1-p)}$. En consecuencia, como $n = 8$ y $p = 1/2$, la media es $\mu = 8(1/2) = 4$ y la desviación es $\sigma = \sqrt{8(1/2)(1-1/2)} = 1{,}41$.

Por otro lado, como la significancia es 0,05 y es con dos colas, la dividimos entre dos y buscamos en la tabla la z asociada a $\Psi = 0{,}025$. Nos da un $z = -1{,}96$. Como es con dos colas, la otra cola está en $z = 1{,}96$. Para hallar la x correspondiente usamos la transformada-x que dice: $x = \mu + z\sigma$. Para la cola superior tenemos $x = 4 + (1{,}96)(1{,}41) = 6{,}76$. Para la z de la cola negativa, tenemos $x = 4 + (-1{,}96)(1{,}41) = 1{,}23$.

Podemos concluir que al aproximar la binomial por la normal los valores extremos para la significancia 0,05 y dos colas son los menores que 1.23 y los mayores que 6.76 y, por supuesto, nos restringimos a los valores posibles que la binomial puede tomar, los enteros entre 0 y 8. Así que en definitiva: los valores extremos para la significancia 0.05 y dos colas son $\{0, 1, 7, 8\}$. Comparando con la solución exacta, los valores 0 y 8 son correctos, pero la aproximación añade 1 y 7, que son incorrectos. Por esa razón, para valores menores que 10 usamos la aproximación con reservas.

5.6. La significancia de un evento

Los eventos compiten por significancia. Por ejemplo, alguien dice: "Fui de pesca con mi abuelo al Magdalena y cogí un pez de 5 libras", y entonces el otro responde: "Eso no es nada: nosotros con mi papá cogimos en el Orinoco uno de tres arrobas". En general, uno se jacta de algo cuando uno presume que ese algo es un evento que queda muy alejado de la media en una cierta distribución de tendencia central. También es claro que uno se jacta solamente cuando uno espera que nadie más pueda superarlo.

Para el caso del tamaño de los peces, si su peso sigue una distribución normal y si uno se jacta del gran tamaño del pez que uno pescó, la significancia puede medirse por la siguiente pregunta: ¿Quién podrá mostrar un pescado tan grande o más que el que nosotros cogimos? Esperamos que nadie sea capaz. Eso es equivalente a decir que la probabilidad de pescar un pez de mayor peso que el nuestro es casi cero.

Lo anterior nos motiva la siguiente definición:

120 ♣ ***Definición.*** *La* ***significancia de un evento*** *c en una distribución de tendencia central se denota $\alpha(c)$ y nos da la probabilidad de que alguien de la población sea más extremo que él. Es recomendable reportar la significancia de los eventos que uno logró obtener, entre más pequeña, más extremos. A la significancia de un evento se le llama también con el anglicismo* ***p-value****. Para la distribución normal tenemos:*

- *Si el dato c está en la cola inferior, su significancia es: $\alpha(c) = \Psi(z(c))$, pues la significancia de un evento que se distingue por ser pequeño es la proporción de eventos más pequeños que él.*
- *La significancia de un evento c que se distingue por ser grande es la proporción de eventos más grandes que él. Teniendo en cuenta que la campana de la Z es simétrica con respecto a 0:*
  $\alpha(c) = P(X > c) = P(X < -c) = \Psi(z(-c))$.
- *Para dos colas = $\alpha(c) = 2\Psi(z(-|c|))$, es decir, la significancia de un evento c es la proporción de eventos que son más extremos que él pero en valor absoluto y se mira en la tabla en la cola inferior.*
- *Los paquetes estadísticos generalmente dan la significancia o el p-value de los eventos para la cola superior.*

121 ***Ejemplo*** *La tabla de la Ψ que hemos reportado es la tabla de la significancia para la cola inferior del z indicado.*

122 ***Ejemplo: Significancia de un evento para normales.*** *Sea $X \sim N(\mu = 3, \sigma = 0{,}4)$ que puede ser el ingreso de los miembros de un grupo de amigos. Hallemos la significancia de algunos eventos:*

1. Pedro gana poquito, apenas 1.2. Su $z = (x - \mu)/\sigma = (1{,}2 - 3)/0{,}4 = -1{,}8/0{,}4 = -4{,}5$. Dicho valor ni siquiera está en la tabla, lo cual dice que $\Psi(4{,}5) \sim 0{,}001$, es decir, que de 1000 personas, sólo habría una que ganaría menos que Pedro. Pero como en un grupo de amigos no hay tanta gente, lo más probable es que Pedro sea el que menos gana. Su significancia es aproximadamente 0.001.
2. Juan gana 4.2. Su $z = (x - \mu)/\sigma = (4{,}2 - 3)/0{,}4 = 1{,}2/0{,}4 = 3$. La $\Psi(3) = 0{,}9987$. Lo cual dice que hay 99.87 % de personas que ganan menos que Juan, es decir, 0.13 % de personas que ganan más que Juan. Su significancia es 0.0013. Muy significante en la cola superior.
3. Alberto gana 3.2. Está demasiado cerca de la media, no se distingue ni por ganar poco ni por ganar mucho. Su $z = 0{,}5$ que da un a $\Psi(0{,}5) = 0{,}6915$. Eso quiere decir que lo que que está por abajo de él es el 0,6915. Y lo que está por encima es el resto, es decir, 0.3085. Como el está en la cola superior, su significancia con la cola superior es esa, 0.3085. Y con dos colas su significancia sería el doble: 0.6170.

5.7. La Hipótesis nula

Lo único que hemos hecho hasta ahora es revestir el sentido común con la terminología oficial para ver cómo se organiza un discurso. Sigamos este proceso de abstracción. **El paso del ejemplo siguiente es fundamental para todo el curso.**

123 ***Ejemplo*** *Nosotros usamos por instinto la estadística para tomar decisiones. Un caso que suena muy popular es el que sigue:*

Patricia sale de clase a las 2 y va llegando a su casa a eso de las tres y media. Rara vez antes de las tres o más tarde que las cuatro. La incertidumbre se debe mayormente a que ella se queda un poco en la U después de clase para poder hacer en grupo algunas tareas. Hoy es jueves. Patricia no ha llegado y ya son las cuatro y cuarto. La mamá ya anda nerviosa, pero se despreocupa después de una llamada de Patricia que le confirma que está haciendo tareas. Llegan las 6 de la tarde pero Patricia no aparece y sus llamadas no tranquilizan a la mamá, la cual ya ha comenzado a llamar a todas las amigas de Patricia, piensa llamar enseguida a la policia, luego al papá y al tío Paco. Ya no le podemos creer que se trata tan sólo de hacer tareas. Es evidente que a Patricia algo le pasó o algo está ocultando. Reformulemos esa historia en términos oficiales:

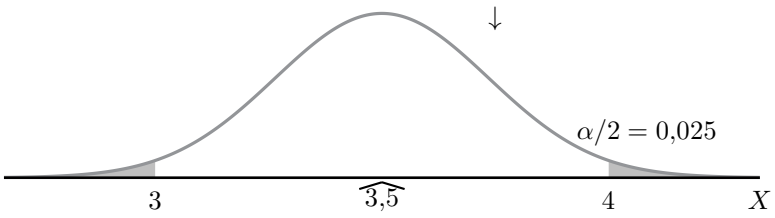


Figura 5.8. X es la hora de llegada de Patricia a su casa. Cuando $3 < x < 4$, la mamá toma una decisión: quedarse tranquila. Cuando $x < 3$ o $X > 4$ se toma otra: prender las alarmas.

Sea X la hora de llegada de Patricia a su casa. X tiene una distribución de tendencia central que idealizamos por una distribución normal. Para estimar la media y la desviación usamos el criterio de la doble mitad: 3 es el valor mínimo y 4 el máximo, 4-3=1, el cual dividimos en dos y da 0.5. La media está en 3.5 en tanto que la desviación vale 0.25. En resumen: $X \sim N(3.5,\ 1/4)$. Entre 3 y 4 hay dos desviaciones al lado y lado de la media. Como el criterio de las dos desviaciones es equivalente a una significancia del 0.05, el 5 % de las veces Patricia llega antes de las 3 o después de las 4. El otro 95 % de las veces Patricia llega entre las 3 y las 4. Eso quiere decir que el 2.5 % llega antes de las tres y el otro 2.5 % después de las cuatro. Trabajamos con dos colas, pues hay problema si Patricia llega tarde e igualemte si ella llega a casa demasiado temprano pues la mamá comienza a pensar si fue que tal vez ella no hizo las tareas. Así que mal si llega temprano, mal si llega tarde.

El contexto dice que hay barreras que dividen lo usual, lo tolerable, de lo inusual, lo intolerable. Lo usual es lo que está entre las 3 y las 4 y dentro de esas barreras está el 95 % de las veces de llegada de Patricia a su casa. La significancia es del 0.05. Si un evento es normal de acuerdo con dicha significancia, tomamos una decisión: la mamá se queda tranquila. Pero si un evento es anormal superior, la mamá toma otra decisión: comienza a llamar a las amigas de Patricia, a la Policía, al papá y al tío Paco.

Hemos usado un procedimiento para decidir una de dos explicaciones: Patricia se quedó haciendo tareas o a Patricia le pasó algo o algo está ocultando. Ahora aprendamos a formalizar todo eso:

Lo usual es que Patricia llegue alrededor de la media. Lo usual se denomina la H_o, la **hipótesis nula** y se escribe $H_o : \mu = 3{,}5$. Esto no es una ecuación, es una simbología que dice que lo usual es que Patricia se quede haciendo tareas, a veces se queda mucho, a veces se queda poco y qué tanto, es algo que depende de muchos factores no controlados y que por tanto no puede predecirse. Por eso, la hora de llegada X es aleatoria, pero su media es 3,5. La aleatoriedad produce variabilidad que se mide con la desviación y que es 1/4. El dato que queremos contrastar contra el transfondo de la H_o se llama **dato experimental**, $x = 6$. Tengamos en cuenta que pueden aparecer datos extremos que inviten a pensar que sucedió algo raro y que incentiven la búsqueda de explicaciones alternas que no forman parte del mundo habitual que se considera en la H_o. Esa opción se denomina la H_a, **hipótesis alterna**. En la H_a se especifica si la negación de la H_o se hace con la cola superior, con la cola inferior o con ambas colas.

Con respecto a la toma de decisiones tenemos:

$H_o : \mu = 3{,}5$ (Hay datos comunes y corrientes y son aquellos cuya diferencia con la media no es significativa teniendo en cuenta el número de colas, la significancia y la desviación. Lo que explica la media también explica el dato. Si aparece un dato así, tome la misma decisión que para la media.)

$H_a : \mu \neq 3{,}5$ (el dato es extremo con dos colas, la diferencia entre el dato y la media es significativa, invente otra explicación aparte de la H_o, la cual debe predecir una media distinta de 3.5).

Con una significancia $\alpha = 0{,}05$, **la RA, la región de aceptación de la** H_o, es el intervalo (3,4). **La RR, la región de rechazo de la** H_o, es lo que queda por fuera: $x < 3$ ó $x > 4$.

Como el dato experimental $x = 6$ queda en la zona de rechazo, rechazamos la H_o: somos partidarios de decir que hay algo anormal en lo que le está pasando a Patricia y que hay un factor extraño que ha causado un desajuste: Patricia está descubriendo que hablar de la vida con Juan es algo que la enamora.

Aclaración: Si la $H_o : \mu \neq 3{,}5$ se rechaza con la cola superior, lo cual dice que el dato experimental se considera extremo del lado superior, entonces la H_a es

$H_a : \mu > 3{,}5$ (dato extremo con la cola superior, invente otra explicación aparte de la H_o, la cual debe predecir que $\mu > 3{,}5$).

Si la H_o se rechaza con la cola inferior, lo cual dice que el dato experimental se considera extremo del lado inferior, entonces la H_a es

$H_a : \mu < 3{,}5$ (dato extremo con la cola inferior, invente otra explicación aparte de la H_o, la cual debe predecir que $\mu > 3{,}5$ y tome una decisión consecuente).

124 ♣ ***Definición.*** *La H_o**, hipótesis nula,** es una afirmación de igualdad de la media o de la varianza poblacionales que representa la conducta natural y espontánea del sistema que se esté estudiando y que incluye el azar que da variabilidad aleatoria. Decimos que una proposición es una **hipótesis alterna**, H_a, cuando esta niega a la H_o. Si hay una explicación para algo, entonces debe ser explicado o bien por la H_o o bien por la H_a, pero no por ambas. Por eso, hay que decidir cuál de las dos hipótesis explica el fenómeno propuesto. Si se rechaza la H_o, se está diciendo que debe haber un factor que crea un efecto sistemático que hace que los resultados se desvíen significativamente de lo predicho por la H_o. Es decir, estamos prediciendo que en circunstancias similares la misma gran discrepancia entre resultados y la H_o serán observados con una alta probabilidad que se hará más real entre más datos se consigan.*

125 ♣ ***Definición.*** *Al valor de z que separa la región de los eventos interesantes de la de los normales se llama z**-crítico**. Ejemplo: con dos colas y $\alpha = 0{,}05$, el z-crítico positivo vale 2 (con más exactitud vale 1.96).*

126 ***Ejemplo: estudiante engreído. Test sobre una media con un único dato y σ conocida.*** *Un estudiante exhibe con orgullo su 3.2 en un examen en el cual las notas se distribuyen normalmente con media 2.3 y desviación 0.3. ¿Se justifica su jactancia?*

Paso 1: Decidimos α que es la significancia, es decir, el porcentaje de eventos que, siendo extremos, serán tomados como anormales. Como no dijeron nada, ni tampoco tenemos criterios personales, se toma la significancia por defecto, 0.05.

Paso 2: Decidimos el número de colas y su posición. Como se trata de un estudiante que sacó una nota sobresaliente, estamos en la cola superior. Como es una única cola, toda la significancia va a ese lado y no se divide por dos.

Paso 3: Formulamos las hipótesis en juego. La hipótesis nula H_o es una igualdad que describe la conducta habitual y normal del mundo: $H_o : \mu = 2{,}3$. Es decir, todo el mundo debería sacar 2.3 de nota excepto por cosas del azar, el cual está contenido en la desviación poblacional que fue $\sigma = 0{,}3$.

La hipótesis alterna H_a niega la nula del lado de la cola superior: $H_a : \mu > 2{,}3$: además del azar, debe haber algún efecto sistemático que hace que al estudiante le vaya mejor que a la gran mayoría de la clase.

Paso 4: Anotamos el dato experimental: la nota en el examen fue de 3.2. Con lo dicho, ya hemos especificado la población y la muestra.

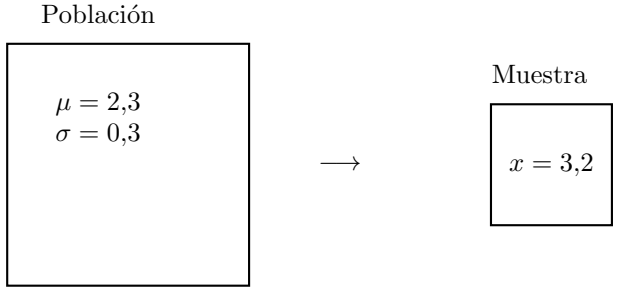


Figura 5.9. Hay una discrepancia entre la media de la población μ y el valor x: ¿se debe dicha discrepancia al azar, a la variación intrínseca de la población, o a algo sistemático?

Paso 5: Hallamos la z- crítica asociada a la significancia. Consultamos la tabla y buscamos la z cuya $\psi(z)$ sea 0.05. Es -1.6 que está en la cola inferior, pero como nosotros tenemos un caso relacionado con la cola superior, tomamos el valor reflejado que es $z_c = 1{,}6$. Grafiquemos:

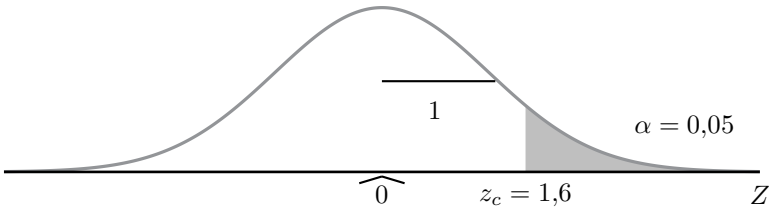


Figura 5.10. Si la significancia es 0,05*, todos los valores de z mayores que* 1,6 *son interesantes si se toma la cola superior.*

Paso 6: Usamos la transformada-x para hallar x_c, el x**-crítico**, asociado a $z_c = 1{,}6$:

$x_c = \mu + z\sigma = 2{,}3 + (1{,}6)(0{,}3) = 2{,}78$

Paso 7: Hacemos la gráfica correspondiente a X y localizamos RA y RR, las regiones de aceptación y de rechazo de la H_o, y también el dato experimental. Cuál región se toma abierta y cuál cerrada es una decisión totalmente arbitraria.

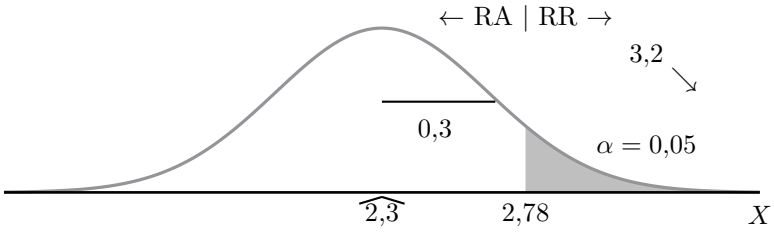


Figura 5.11. Todos los valores de x mayores que 2,78 *son interesantes cuando la significancia es* 0,05 *y se toma la cola superior. La región de aceptación de la H_o es RA = $(-\infty, 2.78]$, la cual contiene todos los eventos que consideramos comunes y corrientes. La región de rechazo de la H_o es RR = $(2.78, \infty)$. El dato experimental $x = 3{,}2$ queda en la RR, es extraordinario: la H_o se rechaza, la nota es anormal.*

Paso 8: Formalizamos el veredicto: la nota $x = 3{,}2$ es extraordinaria, no se ajusta a lo esperado según la H_o. Rechazamos la H_o.

Paso 9: Interpretamos el veredicto en términos del problema. Tenemos que elegir entre dos explicaciones:

1. H_o : de no haber sido por el azar, el estudiante habría sacado el promedio de la clase, 2.3.
2. H_a : hay un factor sistemático que hace que al estudiante le vaya mejor que a la gran mayoría de la clase.

Entre las dos explicaciones, elegimos la segunda: al muchacho le fue definitivamente mejor que a la gran mayoría. Él tiene una razón estadística para jactarse. ¡Ojalá pudiese entender que la estadística no lo es todo en la vida! Por nuestra parte, sería inteligente preguntarle qué es lo que hace, cuál es su método de estudio, cómo lo educó su papá, cómo lo trataron en el colegio, si espera o no que su novia le termine cuando le vaya mal.

127 ***Ejemplo*** *Si X se distribuye normalmente con desviación $\sigma = 1{,}7$ decidamos con $\alpha = 0{,}03$ las siguientes hipótesis nulas, cada una por separado:*

1. $H_o : \mu = 3$, $H_a : \mu < 3$ *sabiendo que hay un dato experimental* $x = -2$.
2. $H_o : \mu = 7$, $H_a : \mu > 7$ *sabiendo que hay un dato experimental* $x = 8{,}4$.
3. $H_o : \mu = 12$, $H_a : \mu \neq 12$ *sabiendo que hay un dato experimental* $x = 8$.

Solución:

1. La notación $H_o : \mu = 3$, $H_a : \mu < 3$ se interpreta así: La H_o es que $X \sim N(3,1.7)$ y se estudia con la cola inferior. Eso quiere decir que el intervalo de aceptación de la H_o es un intervalo de la forma (a, ∞). Para hallar el extremo inferior procedemos así: como se trata de la cola inferior, toda la significancia va hacia abajo. Miramos en la tabla de la Ψ el valor de z correspondiente a $\alpha = 0{,}03$ y da $z = -1{,}8$. Por tanto, por la transformada-x, el valor correspondiente de x es $x = \mu + \sigma z = 3 - 1{,}8(1{,}7) = 3 - 3{,}06 = -0{,}06$.

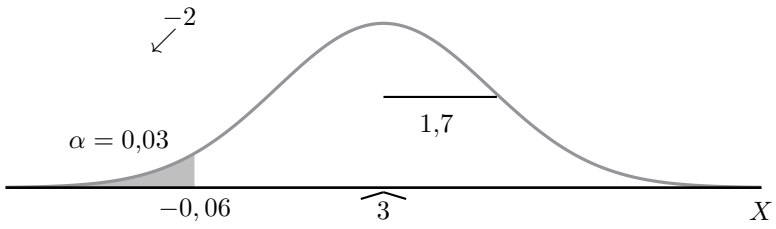


Figura 5.12. Todos los eventos con $x > -0{,}06$ quedan bien explicados por el azar, es decir, por las cosas que no se controlan en el experimento. Pero el evento $x = -2$ es demasiado extremo y no queda bien explicado por el azar: aconsejamos buscarse otra explicación.

El intervalo de aceptación de la H_o es (-0.06,∞). Como el dato experimental -2 no pertenece a dicho intervalo, la H_o se rechaza: el dato experimental es demasiado sospechoso, pues está demasiado abajo, y es mejor comenzar a buscar otra explicación fuera de la que se tiene, que la media vale 3.

2. La notación $H_o : \mu = 7$, $H_a : \mu > 7$ se interpreta así: La H_o es que $X \sim N(7,1.7)$ y se rechaza con la cola superior. Toda la significancia va hacia arriba.

   Vemos que el intervalo de aceptación de la H_o es un intervalo de la forma $(-\infty, b)$. Para hallar el extremo superior procedemos así: miramos en la tabla de la Ψ el valor de z correspondiente a 0,03 y da $z = -1{,}8$. Como se trata de la cola superior, usamos la simetría y tenemos $z = 1{,}8$.

   Por la transformada-x, el valor correspondiente de x es $x = \mu + \sigma z = 7 + 1{,}8(1{,}7) = 7 + 3{,}06 = 10{,}06$. El intervalo de aceptación de la H_o es (-∞, 10.06). Como el dato experimental 8.4 si pertenece a dicho intervalo, la H_o se acepta: aceptamos la explicación que se tiene, que la media vale 7 y que hay una variación debida al azar, a todos los factores no controlados.

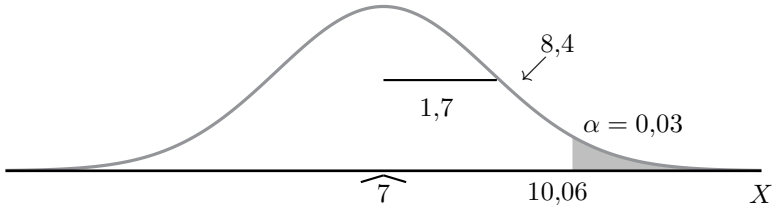


Figura 5.13. El evento 8,4 es común y corriente si X tiene una distribución normal con media 7, desviación 1,7 y si la significancia es 0,03 con la cola superior.

3. La notación $H_o : \mu = 12$, $H_a : \mu \neq 12$ se interpreta así: La H_o es que $X \sim N(12,1.7)$ y se rechaza con ambas colas. Eso quiere decir que el intervalo de aceptación de la H_o es un intervalo de la forma (a, b). Para hallar el extremo inferior procedemos así: como se trata de las dos colas, la significancia va la mitad hacia abajo y la otra mitad hacia arriba. Por lo tanto, lo que va a hacia abajo es $\Psi(a) = 0{,}015$. Miramos en la tabla de la Ψ el valor de z correspondiente a 0,015 y da $z = -2{,}2$.

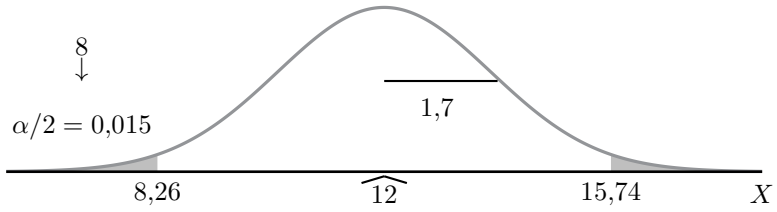


Figura 5.14. Con $\alpha = 0{,}03$ y dos colas, los valores interesantes sobre X son los menores que 8,26 y los mayores que 15,74. Como se ve, el dato $x = 8$ es extremo: su existencia queda mal explicada por azar, por las cosas comunes y corrientes que suceden como resultado de los factores no controlados. Aconsejamos pensar en que hay un factor sistemático en juego.

Por la transformada-x, el valor correspondiente de x es $x = \mu + \sigma z = 12 - 2{,}2(1{,}7) = 12 - 3{,}74 = 8{,}26 = a$. El extremo superior se obtiene por simetría: $b = 12 + 2{,}2(1{,}7) = 12 + 3{,}74 = 15{,}74$. El intervalo de aceptación de la H_o es (8.26,15.74). Como el dato experimental 8 no pertenece a dicho intervalo, la H_o se rechaza: la explicación que se tiene, que la media vale 12 y que hay una variación debida al azar, es insufiente para explicar el dato exprimental $x = 8$. Aconsejamos pensar en una causa sistemática que pueda producir desviaciones tan altas.

128 ***Ejemplo:Test sobre una binomial.*** *En la clase sobre* Desigualdad social en contra de la mujer *hay 21 mujeres y 10 hombres, pero en toda la universidad hay 4500 mujeres contra 5600 hombres. ¿Puede decirse que hay una sobre-representación de mujeres en esta clase?*

Solución en 15 pasos:

1. Distribución binomial, pues hay dos opciones: mujer y hombre. Decidimos que X cuenta el número de mujeres: $X \sim Bi(n, p)$.
2. Población = conjunto grande = Universidad. Muestra = conjunto pequeño = clase.
3. **La probabilidad** de mujer, p, **sale de la población** que en este caso es la Universidad con 4500 + 5600 = 10100 personas, de las cuales hay 4500 mujeres. Por tanto $p = 4500/10100 = 0{,}4456$.
4. n es el número de monedas lanzadas, es el número de elementos de la muestra: 21+10 = 31 monedas (personas), de las cuales cayeron 21 caras (mujeres). Por tanto, la frecuencia relativa observada es $f = 21/31 = 0{,}677$.

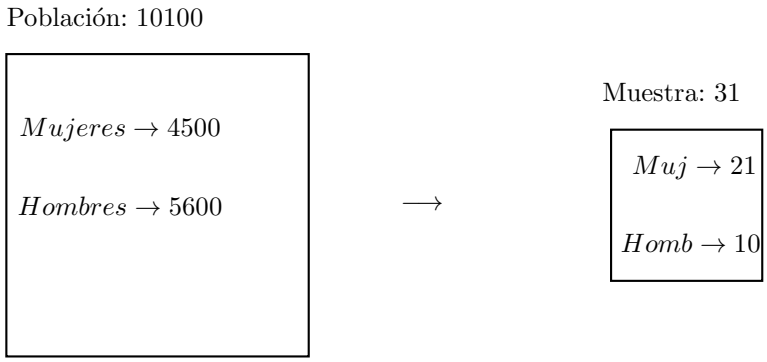


Figura 5.15. Hay una discrepancia entre la probabilidad de mujer en la población, $p = 4500/10100 = 0{,}4456$, y la frecuencia observada en la muestra, $f = 21/31 = 0{,}677$. ¿Se debe dicha discrepancia al azar, a que las reparticiones al azar crean variabilidad, ó a algo sistemático?

5. Problema: explique la discrepancia entre p y f. Si es pequeña se puede explicar por azar del muestreo, pero si es muy grande es mejor suponer que hay algo sistemático que la causa.

6. Número de colas: la palabra clave en la pregunta es sobre-representación, que quiere decir mayor de lo normal: se trata de una cola, la superior. Significancia = 0.05, pues no dijeron nada.

7. Distribución de X debida al azar: $X \sim Bi(n = 31, p = 0.4456)$.

8. Formalizamos la Hipótesis nula y la alterna:

   $H_o : p = 0{,}4456$. Para que la clase refleje la U no se requiere que la frecuencia observada de mujer sea 0.4456 pero las diferencias por azar han de ser pequeñas.

   $H_a : p > 0{,}4456$, la cual dice que hay una razón sistemática que hace que la frecuencia relativa de ser mujer en la muestra sea mayor que 0.4456.

   La H_o se rechaza con la cola superior y con significancia 0.05.

9. Calcular la binomial directamente es muy largo pues $n = 31$: aproximamos a la normal, la cual se define por su media y su desviación que deben ser las mismas de la Binomial:

   $\mu = np = 31 \times 0{,}4456 = 13{,}8$

   $\sigma^2 = np(1-p) = 13{,}8 \times 0{,}5544 = 7{,}65$, por lo que $\sigma = 2{,}76$

10. Reformulación: $X \sim Bi(n = 31, p = 0.4456) \sim N(\mu = 13.8, \sigma = 2.76)$. Esto implica una traducción de la H_o que ahora dice: $H_o : \mu = 13{,}8$. Se rechaza con la cola superior y con $\alpha = 0{,}05$.

11. La significancia es 0.05 y va a una sola cola, por tanto $z = -1{,}6$, la cual se convierte en 1.6 por ser la cola de arriba.

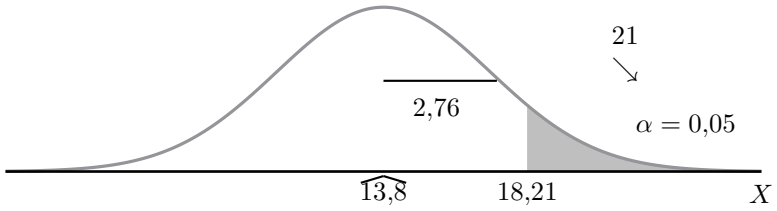


Figura 5.16. Si $H_o : p = 0{,}4456$, o lo que es lo mismo, $H_o : \mu = 13{,}8$, con la cola superior y $\alpha = 0{,}05$, la zona de aceptación de la H_o es lo que está abajo de 18.21. Como el dato experimental $x = 21$ está afuera de tal zona, rechazamos la H_o.

12. Hallamos el x crítico: $x_c = \mu + z\sigma = 13{,}8 + 1{,}6 \times 2{,}76 = 18{,}21$

13. Los eventos preocupantes son los mayores que 18.21.

14. Tomamos un veredicto: La Regíon de Aceptación de la H_o es $(-\infty, 18.21)$. La Región de Rechazo de la H_o es $(18.21, \infty)$. La discrepancia entre lo predicho por la H_o, que en promedio debería haber 13.8 mujeres, y lo observado, que hay 21, no se explica fácilmente por azar. Veredicto: se rechaza la H_o.

15. Expresamos el veredicto en términos del problema, lo cual evidencia un acto del libre albedrío, pues nadie está obligado a estar de acuerdo: Declaramos que existe una causa sistemática que hace que haya una sobre-representación de mujeres en la clase de sociales dada con respecto a la Universidad, por ejemplo, que la profesora se distingue por su lucha a favor de la mujer.

129 ***¿15 pasos?*** *Uno puede sentirse asustado por tantos pasos ejecutados en la solución del problema anterior. Nos parece que todo puede resumirse en: formular el problema sobre una binomial + aproximarlo sobre una normal + tomar un veredicto y expresar el veredicto en términos del problema.*

130 ***Formalismo.*** *Todo problema se replantea como el estudio de una hipótesis nula que contiene una igualdad. Por ejemplo, que la media vale 8, o que la varianza vale 3, o que la diferencia de medias es 0, o que la relación de las varianzas es 4. El estudio consiste en decir si la igualdad pregonada por la hipótesis nula se acepta o se rechaza. Una hipótesis nula jamás contiene una desigualdad, la cual queda bien en la Hipótesis alterna. La razón es que la metodología matemática que se usa para dirimir el problema, y que no será explicada, siempre considera la igualdad como la proposición que se calcula por naturalidad.*

131 ***Ejemplo*** *Las carreteras que unen las ciudades van llenándose de construcciones al lado y lado formando barrios extendidos. Cuando la carretera está densamente poblada, al conjunto de ciudad y carretera se denomina megalópolis. Hay partes del mundo, como en la región de Kyoto, Japón, donde las megalópolis encierran varias ciudades.*

Necesitamos decidir si un cierto territorio ya debe clasificarse como megalópolis e imponer al Estado todas las obligaciones que de eso se deriva: colegios, servicios, hospitales.

Para ello se hace un muestreo de 80 niños en un colegio que queda dentro de lo que oficialmente es la ciudad pero casi en el borde de su perímetro. De los 80 niños, 60 son de la ciudad y 20 son de lo que oficialmente es el campo. ¿Se ha poblado la región lo suficiente como para ser considerada como megalópolis? Por supuesto que asumimos que los niños de campo están tan ávidos de estudio como los de la ciudad.

Solución: Tenemos una hipótesis nula H_o que dice que ya hay megalópolis y que cualquier diferencia entre la proporción de niños de campo y de ciudad se debe a cosas del azar, cosas que no podemos controlar, por ejemplo, que las parejas con niños prefieran vivir en la ciudad o quizá al revés. Y nuestra hipótesis alterna H_a dice que las proporciones observadas son diferentes no por azar sino porque definitivamente la carretera está despoblada o quizá superpoblada.

Si para un niño, la probabilidad de ser de campo es p, podemos escribir nuestras hipótesis resumidamente: si ya hay megalópolis, dicha probabilidad deber ser igual a la de venir de ciudad. Por eso, la hipótesis nula nos dice:

H_o: $p = 1/2$.
$H_a : p \neq 1/2$.
Como no dijeron nada sobre la significancia, se rechaza la H_o con una significancia del 0.05.

A esto se le llama una **prueba o test de dos colas**. Decimos de dos colas, pues la H_a es la negación de la H_o que es una igualdad. Y como se puede ser desigual por arriba o por abajo, la H_a prevalece en el caso de que haya un evento raro, sea por exceso, en la cola superior, o por defecto, en la cola inferior. (Decimos que una **prueba o test es de una cola** cuando la H_a contiene una desigualdad, por ejemplo $H_a : p > 1/2$, que se cumple sólo en la cola superior.)

Decidamos cuál de las dos hipótesis es más opcionada por la estadística: elaboremos la H_o. Si es igualmente probable ser de campo que de ciudad, entonces la cantidad de niños de campo se distribuye como una $Bi(n = 80, p = 1/2)$.

Podemos aproximar la binomial por la normal: la campana que mejor se ajusta tiene media $\mu = np = 80 \times 0{,}5 = 40$ y la desviación es
$\sigma = \sqrt{npq} = \sqrt{80 \times 0{,}5 \times 0{,}5} = 4{,}5$
Resumiendo: si la H_o es cierta, deberemos contar en promedio con $\mu = 40$ niños de campo y una desviación $\sigma = 4{,}5$. Lo normal según la H_o, con significancia 0.05 y dos colas tiene $z = 1{,}96 = 2$. Por tanto, lo normal va entre 40 - 2(4.5) = 31 y 40 + 2(4.5) = 49 niños de campo entre los 80 enumerados.

Veredicto: De los 80 niños del colegio, apenas 20 son del campo. Podemos concluir que el campo y la ciudad aportan diferentes proporciones de niños al colegio estudiado. La ciudad y el campo aún no se han igualado en densidad de población.

132 ***Advertencia***

Nuestros eventos interesantes corresponden, aproximadamente, a aquellos que están lejos de la media. Esto está bien definido para las distribuciones con una tendencia central, pues el conjunto de eventos alejados de la media tiene una probabilidad muy baja. En ese caso, nuestra noción refleja lo que nosotros queremos precisamente: detectar los eventos que despiertan nuestra atención.

Pero ese no es siempre el caso.

Consideremos un dado. Sólo un dado. Tiene 6 caras, todas equiprobables. Esto significa que cuando el número de ensayos crezca, las frecuencias relativas de caras diferentes tenderán a asumir valores

similares. La distribución correspondiente se llama **la distribución uniforme**. Las probabilidades simplemente son

$p(k) = 1/6$ para $k = 1, 2, 3, 4, 5, 6$.

En este caso no hay eventos distinguidos pues todos los eventos son igualmente interesantes. Recalcamos: esta distribución no admite el concepto de eventos interesantes. Para que un evento sea interesante se necesita que llame la atención por algún motivo y que su probabilidad sea *comparativamente* muy baja.

Un dado no da eventos interesantes. Sin embargo, cuando se combinan varios dados, se suman distribuciones independientes y el resultado es una distribución de tendencia central, para la cual ya tiene sentido definir eventos interesantes y es por eso que los juegos de dados son muy populares.

133 *Los modelos compiten por credibilidad*

A una hipótesis nula también se le llama modelo o suposición o explicación, pues el objetivo de una hipótesis es explicar un evento o familia de eventos.

Las explicaciones o modelos compiten entre ellas por credibilidad. Y todo viene del sentido común. Para verlo, consideremos el caso de un hombre que para justificar su riqueza declara haberse ganado la lotería dos veces pero que al mismo tiempo trabaja en la Aduana en donde hay demasiada oportunidad para la corrupción. Es perfectamente posible que lo que él dice sea verdad. Pero la cruda realidad es que además de su familia nadie le cree. ¿Por qué? Porque asumir que él adquirió su enorme riqueza gracias a la corrupción ofrece una explicación que convierte su enorme riqueza en un evento altamente probable. En cambio, todo el mundo se rehúsa a creer que se ganó dos veces la lotería por una simple razón: las loterías sí explican las cosas pero lo hacen con una probabilidad demasiado baja.

Concluimos que la credibilidad de un modelo para explicar un evento aumenta entre más grande sea la probabilidad que el modelo le asigna al evento. Existe una escuela dentro de la estadística que toma estas apreciaciones como la razón y esencia del método científico: es la Estadística Bayesiana.

134 ***Ejemplo*** *Expliquemos por qué cayeron 15 caras de 16 lanzamientos.*

Considerando una moneda generalizada, basada en una ruleta, por ejemplo, se sabe que de 16 lanzamientos, 15 cayeron cara. Podemos explicar ese evento de muchas maneras: podemos decir que $p(cara) = 1/2$. Esa hipótesis explicaría el evento pero lo haría con una probabilidad muy baja. Veamos que el evento es un outlier. La media sería 8, la varianza 4, la desviación sería 2 y dos desviaciones serían 4. Por tanto los outliers serían 0,1,2,3,13,14,15,16. Como vemos 15 es un outlier.

Como en ese evento cayeron muchas más caras que sellos, podemos inferir que ese evento fue generado por una moneda generalizada cuya probabilidad de cara pudo ser, digamos, 3/4. En ese caso la media sería 12 y la desviación sería 3, por lo que dos desviaciones serían 6. Eso significaría que los outliers serían, aproximadamente, los números menores que 9 y nadie más. En ese caso 15 caras de 16 sería un evento corriente, normal.

Los dos modelos, $p = 1/2$, y $p = 3/4$ compiten pero gana el último. Y bien, ¿de todos los modelos, cuál sería el vencedor? Aunque tenemos medios matemáticos para saberlo, aventurémonos a proponer que $p = 15/16$ sería el modelo vencedor. Fijémonos que en ese caso la media sería 15, algo que nos ofrece mucha seguridad.

Llegamos a la conclusión de que un evento tiene infinidad de explicaciones plausibles. Para que uno se decida a favor de alguna en especial se requiere obtener más y más datos. Y nunca tendremos razones lógicas para gozar de certidumbre total. Por ello, la ciencia es un caminar, no un destino. O dicho más claramente: la conexión de la ciencia con la verdad es más una materia de fe y de irracionalidad que de certidumbre y lógica.

135 *Norma y normatividad*

Hemos designado como normal a aquel evento que queda en el centro de una distribución de tendencia central. Eso es perfecto para una variable numérica, cuantitativa, como el número de amigos que una persona tiene. Pero, ¿qué hay de los datos ordinales y categóricos? Veamos cuan importantes son los

datos categóricos y de qué forma se define en esos casos lo que es normal. Consideremos el caso del **Lenguaje** y luego el del **Derecho**.

La función del lenguaje es servir de medio de empaquetamiento de la información. Debido a la riqueza del lenguaje, dicha función puede ejecutarse de muchas maneras. La riqueza expresiva del lenguaje es imprescindible para que cada cual pueda expresar su personalidad. Por esa razón, siempre habrá variabilidad lingüística, la cual se describirá por datos categóricos y sus frecuencias: el 60 % de las personas dicen ¡Aló!.

Sin embargo, no es rara la ocasión en que la variabilidad se ve fuertemente menguada y aparece el concepto de **norma**, es decir, la **moda** adquiere un carácter obligante: podemos decir ¡Hola, qué tal! Pero son pocos los que dicen ¡Ajá, que tal!, expresión que todo el mundo entendería en el acto. De esa manera la moda se convierte en lo normal, en lo estético, en lo ético, en lo que debe ser.

Es decir, de datos categóricos hemos saltado a datos ordinales, a clasificaciones tan extremas que hasta decimos qué es bien y qué es mal. Esto es mucho más colorido en Derecho: las formas de relacionarse entre los miembros de una sociedad tienen una elevadísima variabilidad, pero tampoco tanta: nos estrechamos la mano para saludarnos, pero no las orejas. Evidentemente, la moda no es sólo lo normal, sino lo debido, lo moral, lo ético. Tanto en Lingüística como en Derecho, cuando la Academia de Lenguas y de Leyes entren a generar normatividad, la moda se convierte en la norma y llega a ser sinónimo de educado, académico.

Para que la normatividad o ley tenga el respaldo de la sociedad es absolutamente indispensable que lo que esté consignado en la ley sea la forma normal de conducta de la gran mayoría de personas. De esa forma, cuando el aparato coercitivo del Estado use su poder contra un contraventor, podrá percibirse entre la sociedad un sentimiento de descanso, de paz y hasta de alegría.

Hemos expuesto la tesis costumbrista de la norma, válida tanto en Linguística como en Derecho: la ley debe contener lo que es la costumbre, lo que es la **moda**. La estética, la ética y la moral no son primarias en esta tesis, sino que ellas figuran tras la costumbre. Una vez que la costumbre se establece, la estética, la moral y la ética forman la maquinaria invisible y todopoderosa cuyo único objetivo es hacer que la costumbre perdure.

La tesis costumbrista no es del agrado de todo el mundo. Por ejemplo: todo dictador se declara un instrumento de grandes ideales al momento de asumir el mando y está tan convencido de ello que no duda en sacrificar a la oposición.

¿Qué opción social tienen los nobles ideales? Muchísima. Veamos un ejemplo: todos hemos oído de alguno de los asesinatos ocurridos a la vista de todos porque el asesino se sintió dolido en su amor propio al oír una grosería en su contra. Y aunque los insultos verbales son muy frecuentes, los asesinatos que vengan las malas expresiones no. Con todo, cualquier asesinato de ese tipo es una catástrofe social. ¿Cómo trata de protegerse la sociedad de estas calamidades? Cultivado un noble ideal: ¡Cuán hermoso es ser decente! y cuando una persona decente oye dichas historias, ella siente que tiene algo que no cuesta nada pero que vale más que el oro y que la plata. Es así como nacen los 'valores' de la filosofía: pueden surgir de cualquier lado y pueden mantenerse en la sociedad por múltiples razones pero también pueden pasar de moda. Con todo, aquellos ideales que representan una ventaja para la supervivencia tienen más opciones de permanecer por largo tiempo y de ser transculturales.

Es tan fuerte la necesidad social de los ideales que hoy en día es impensable no contar con facultades de **Ciencia Política** donde se estudia, por ejemplo, la presión que los flujos económicos ejercen sobre la ley, pero además cómo se pueden regular o cambiar dichos flujos para que haya, por ejemplo, más estabilidad social.

En resumen: para datos categóricos el concepto de 'normal' también tiene sentido y se puede identificar con los eventos cuyas frecuencias relativas son dominantes, muy particularmente con la moda.

5.8. Intervalos de confianza

En esta sección nos ocuparemos de las siguientes preguntas:

1. Para una normal con una desviación dada, ¿cuánto vale la media μ si uno tiene un dato experimental dado?

2. Para una binomial, ¿cuánto vale p, la probabilidad de cara, si se lanzan n monedas y salieron k caras?

Ambas preguntas se responden de igual forma. Para verlo, consideremos el problema de la media en la normal: ¿cuánto vale la media sabiendo σ si tenemos un dato experimental x? Respuesta: No sabemos. Pero dado que tenemos un dato que pertenece a una distribución de tendencia central, lo mínimo que podemos decir es que μ debe estar por ahí cerca del dato. Decimos que un **número m es un valor plausible para la media** cuando dentro de la campana con media m y desviación σ el valor x es normal con la significancia dada y con dos colas. Eso implica que con los datos dados la $H_o : \mu = m$ se acepta. Al reunir todos los valores plausibles para la media, se tiene un intervalo que se denomina el **intervalo de confianza** o IC. Por tanto: no sabemos cuánto vale la media, pero es muy plausible que esté dentro del IC.

136 ***Ejemplo: Intervalo de confianza para una media con un único dato.*** *Hallemos constructivamente, para ver de dónde sale todo, el intervalo de confianza para μ con $\alpha = 0{,}008$ dado que hay un dato experimental $x = 9$ y sabiendo que $\sigma = 1{,}3$. Un número m pertenece a dicho intervalo si el dato $x = 9$ es normal con la significancia dada y con dos colas. Eso implica que con la $H_o : \mu = m$ se acepta pues $x = 9$ sería normal con dos colas y el α dada.*

Solución: **Paso 1**: *Intuyendo la respuesta.* Como el dato experimental es $x = 9$ y la desviación es 1.3, la media de la distribucíón debe estar por ahí cerca, digamos puede estar entre 5 y 15, pero sería increíble que la media fuese 1000 ó 10000. Lo que tenemos que hacer es reportar todas las medias que no generen protestas con la significancia dada. Para ello seguimos el procedimiento siguiente:

Paso 2: *Determinamos el z crítico asociado a la significancia*: sobre la Z, a la significancia $\alpha = 0{,}008$ le debe corresponder un valor de z crítico. Pero el dato puede ser anormal tanto por arriba como por abajo de la media poblacional y por eso el problema se resuelve con dos colas. Dividimos por tanto 0.008 entre dos y miramos en la tabla de la Z el z correspondiente a una área de 0.004. Da -2.65.

Paso 3: *Para una μ fija, determinemos lo normal, la región de aceptación de la H_o.* La normal asociada a la H_o tiene media desconocida μ y desviación $\sigma = 1{,}3$. Como el z crítico nos dio -2.65, eso significa que sobre X, los valores anormales empiezan en la media más o menos 2.65 veces la desviación estándar, es decir $2{,}65 \times 1{,}3 = 3{,}44$.

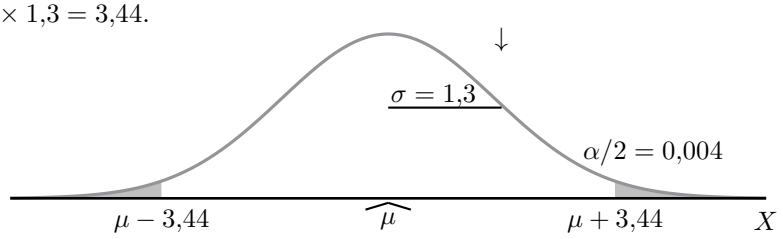


Figura 5.17. Con $\alpha = 0{,}008$ y dos colas, los eventos comunes y corrientes en una campana de media μ van desde $\mu - 3{,}44$ hasta $\mu + 3{,}44$.

Paso 4: *Fabricación de la campana recortada.* Recortamos de la campana las colas que contienen todos los eventos anormales. La campana recortada tiene centro en la media μ y tiene a cada lado una aleta que mide 3.44.

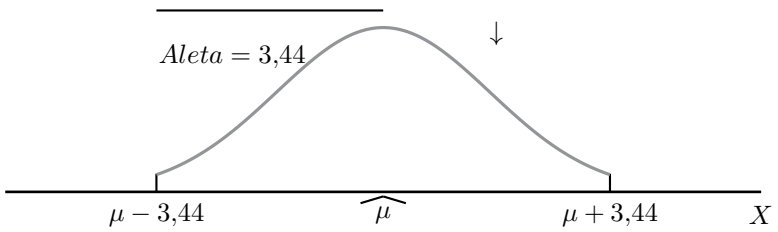


Figura 5.18. La campana recortada contiene todos los eventos normales, comunes y corrientes.

Paso 5: *Contraste con el dato experimental.* Tomamos la campana recortada y la ponemos sobre el eje real. Hay tres opciones: que la campana no contenga el dato por estar demasiado arriba, o que lo contenga, o que no lo contenga por estar demasiado abajo. Lo que tenemos que averiguar es el

subconjunto del eje real donde uno puede poner el centro de la campana, la media, para que ésta y el dato sean compatibles, es decir, que el dato quede dentro de la campana.

Imaginemos que empezamos haciendo un tanteo desde bien arriba y vamos bajando tratando de encontrar el primer punto donde la camapana contiene el dato experimental: en dicho caso el dato coincide con el corte inferior de la campana recortada:

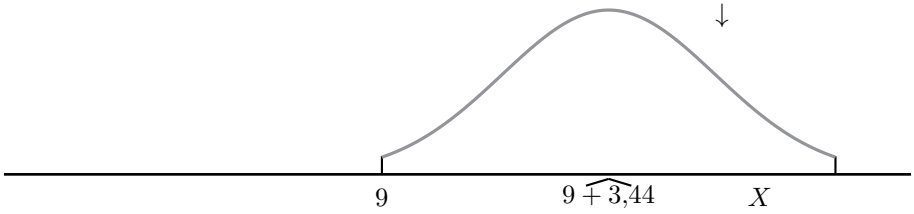


Figura 5.19. Campana recortada extrema hacia arriba que contiene el dato experimental como común y corriente. La media está una aleta arriba del dato experimental: $\mu = 9 + 3{,}44 = 12{,}44$.

Es decir, la campana extrema hacia arriba tiene su media en 9 + 3.44 = 12.44.

Si continuamos bajando, llegará un momento en que el dato coincida con el extremo superior de la campana recortada:

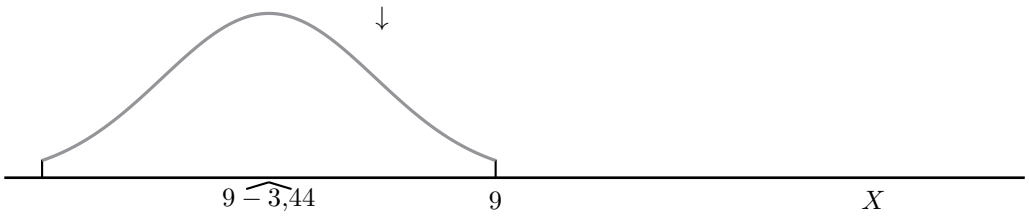


Figura 5.20. Campana recortada extrema hacia abajo que contiene el dato experimental como común y corriente. La media queda en 9-3.44 = 5.56

La media queda una aleta abajo del dato experimental μ = 9-3.44 = 5.56. Por tanto, el intervalo buscado empieza en el dato experimental menos 3.44 y termina en el dato experimental más 3.44. Eso da el intervalo

(9-3.44, 9+3.44) = (5.56, 12.44)

Paso 6: Cerramos el intervalo para que el dato experimental pueda estar en los bordes de lo que se considera normal [5.56, 12.44].

Paso 7: *Respuesta:* en el intervalo [5.56, 12.44] están todas las medias que son compatibles con el dato experimental $x_o = 9$ dado que $\sigma = 1{,}3$ y con la significancia 0.008 y dos colas. Se dice oficialmente: **el IC de la media con confianza del 99.2 % (1-0.008= 0.992) es [5.56, 12.44]**.

Paso 8: *Interpretación:* El IC nos permite responder a la pregunta: dado que el dato experimental es 9, ¿cuál es la media? Respuesta: no sabemos, pero con una confianza del 99.2 %, queda en el intervalo [5.56, 12.44]. La media poblacional puede estar o no dentro de este intervalo, pero si apostamos rutinariamente a que está en él, ganaremos el 99.2 % de las veces. Retrospectivamente, vemos que el IC resulta de poner la campana encima del dato experimental y moverla hacia arriba y hacia abajo hasta que ya no se pueda más pues el dato se saldría de la campana. Hacer ésto es es un riesgo calculado: el dato experimental salió al azar, por lo que puede estar en cualquier parte del eje real, dentro de la campana recortada o por fuera de ella. Pero la probabilidad de que esté en las colas es la significancia, la cual es 0.008 y, por tanto, la probabilidad de que esté dentro de la campana recortada es el resto, es decir, 0.992. Por eso decimos que trabajamos con una confianza del 99.2 %.

137 ♣ ***Definición.*** *El* ***IC, intervalo de confianza, para la media con confianza*** $1-\alpha$ ***y con dos colas dado el dato experimental*** x_o*, es el intervalo que contiene todos los números* m *tal que* $H_o : \mu = m$ *se acepta con significancia* α *y con dos colas.*

138 ***Ejemplo: Intervalo de confianza para una proporción.*** *Hallemos para una binomial, el IC intervalo de confianza del 88 % para* $p =$ ***probabilidad de cara****, si en 17 lanzamientos se obtuvieron 12 caras. Es decir, hallemos todos los valores posibles de* p *para los cuales* $f = 12/17$ *sería normal, no extremo.*

Solución. **Paso 1**: *Aterrizar:* Puesto que la frecuencia observada es $f = 12/17 = 0.70$, lo más confiable es decir que la probabilidad de cara debe estar por ahí cerca, 0. 65, 0.72. Pero no esperamos que $p = 0{,}1$ pues en ese caso, de 17 lanzamientos deberíamos esperar 1.7 caras, es decir 1 o 2, pero no 12.

Paso 2: *Determinamos el z crítico asociado a la significancia*: Sobre la Z, a la significancia $\alpha = 0{,}12$ le debe corresponder un valor de z crítico. Pero observemos que la p que sirve puede estar abajo o arriba del dato experimental, por eso, el problema se resuelve con dos colas. Dividimos la significancia 0.12 entre dos y vemos que el z correspondiente a una área de 0.06 es -1.55.

Paso 3: *Aproximación descarada de la binomial por la normal.* La aproximación de la binomial por la normal es decente cuando p es aproximadamente $1/2$ pues en ese caso la binomial es casi simétrica. Pero cuando p es muy pequeño o muy grande el histograma es muy asimétrico y la aproximación por la normal es abusiva. Procedamos con dicho atrevimiento: $X =$ número de caras se aproxima por la normal con una media anónima $np = 17p$ y desviación $\sqrt{17p(1-p)}$.

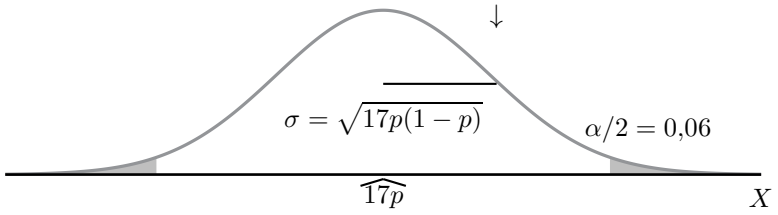


Figura 5.21. Aproximamos $X \sim Bi(n = 17, p)$ *por* $N(\mu = 17p, \sigma = \sqrt{17p(1-p)})$*. X cuenta el número de caras que va desde cero hasta n, el número de monedas, en este caso 17.*

Paso 4: *Reescalamiento* En el paso 3 encontramos una campana sobre el eje X que cuenta el número de caras que puede ir de cero a n, el número de monedas. Pero nos preguntan no sobre el número de caras sino sobre la probabilidad de cara. Una probabilidad es algo abstracto que se encarna en las frecuencias relativas. ¿Cómo pasamos del número de caras a la frecuencia relativa de caras? Dividiendo el número de caras X por el número de monedas, que en nuestro caso es 17. Así que toda la campana con media $np = 17p$ y desviación $\sqrt{17p(1-p)}$ la dividimos por 17 y nos queda otra campana con media p y desviación $\frac{\sqrt{17p(1-p)}}{17} = \frac{\sqrt{17}\sqrt{p(1-p)}}{\sqrt{17}\sqrt{17}} = \frac{\sqrt{p(1-p)}}{\sqrt{17}} = \sqrt{\frac{p(1-p)}{17}}$. Esta campana está sobre el eje $\frac{X}{17}$ que es una frecuencia relativa que encarna la probabilidad.

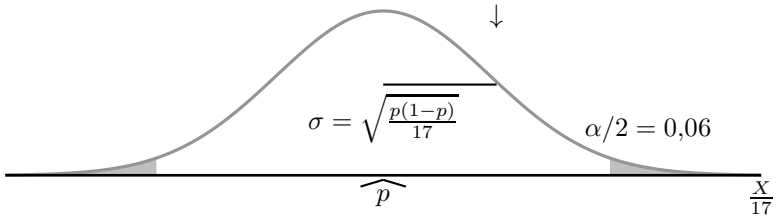


Figura 5.22. Reescalamos desde X *que cuenta el número de caras a la probabilidad que idealiza la frecuencia relativa de cara* $X/17$*. La gráfica dice:* $X/17 \sim N(p, \sqrt{\frac{p(1-p)}{17}})$.

Paso 5: *Fabricación de la campana recortada* Recortamos de la campana las colas que contienen todos los eventos anormales. Hallemos las barreras de corte: tenemos una normal con media desconocida p y desviación $\sigma = \sqrt{\frac{p(1-p)}{17}}$. Como la z crítica nos dio -1.55, entonces eso significa que sobre p, los valores anormales empiezan en la media p más o menos 1.55 veces la desviación estándar, es decir $1{,}55 \times \sqrt{\frac{p(1-p)}{17}}$. Dicho de otro modo: la campana recortada tiene centro en la media p y tiene a cada lado una aleta que mide $1{,}55 \times \sqrt{\frac{p(1-p)}{17}}$.

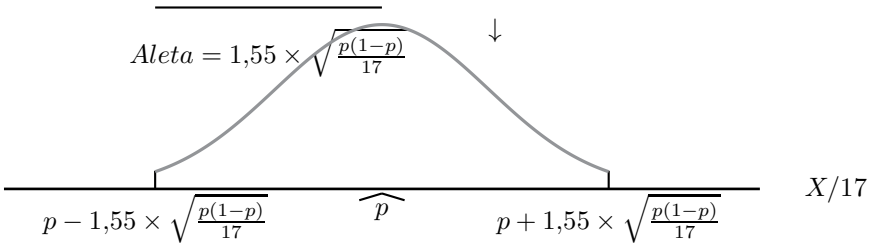


Figura 5.23. La campana recortada de $X/17$ contiene todos los eventos normales, comunes y corrientes. Tiene una aleta que mide $1{,}55 \times \sqrt{\frac{p(1-p)}{17}}$.

Paso 6: *Usamos la aproximación rígida* La campana recortada tiene una desviación igual a $\sqrt{\frac{p(1-p)}{17}}$, la cual depende de p. Eso quiere decir que si p cambia, la campana asociada o bien se ensancha o se angosta pero no permanece igual. Usamos la siguiente aproximación: en vez de considerar que la campana cambia cuando p cambia, nos imaginamos que permanece invariante, **rígida**. Y si ese es el caso, lo más verosímil según el experimento es que p sea igual a $f = \frac{12}{17}$ pues de 17 monedas cayeron 12 caras. Por tanto, la desviación de la campana rígida recortada sería entonces

$\sigma = \sqrt{\frac{p(1-p)}{17}} = \sqrt{\frac{\frac{12}{17}(1-\frac{12}{17})}{17}} = 0{,}11$
y la aleta
$1{,}55 \times \sqrt{\frac{p(1-p)}{17}} = 1{,}55\sqrt{\frac{\frac{12}{17}(1-\frac{12}{17})}{17}} = 0{,}17$

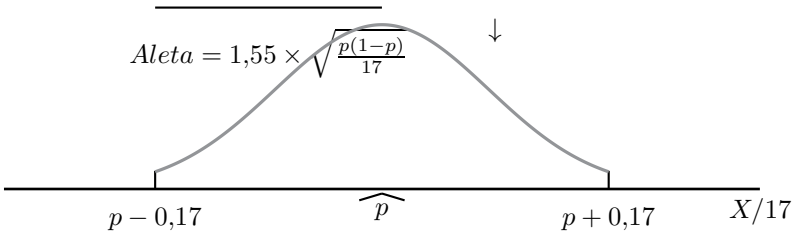


Figura 5.24. La campana recortada rígida contiene todos los eventos normales, comunes y corrientes. Tiene una aleta que mide $1{,}55 \times \sqrt{\frac{p(1-p)}{17}}$ pero con $p = 12/17 = 0{,}70$, lo que da 0.17.

Paso 7: *Contraste con el dato experimental* Puesto que el dato experimental es 12/17 que es algo fijo, acomodamos la campana recortada encima de él. Atención: el dato experimental salió al azar, por lo que puede estar en cualquier parte del eje real, dentro de la campana recortada o por fuera de ella. Pero la probabilidad de que esté en las colas es la significancia, la cual es 0.12 y, por tanto, la probabilidad de que esté dentro de la campana recortada es el resto, es decir, 0.88. Por eso decimos que trabajamos con una confianza del 88 %. Dejamos el dato experimental fijo y movemos la campana hacia la izquierda y hacia la derecha hasta donde ya no se pueda más, es decir, hasta que moverla un poquito más implica que el dato se convierta en anormal pues se sale de la campana. Claramente, la campana puede moverse dentro de un intervalo y lo que necesitamos es hallar los límites de dicho intervalo.

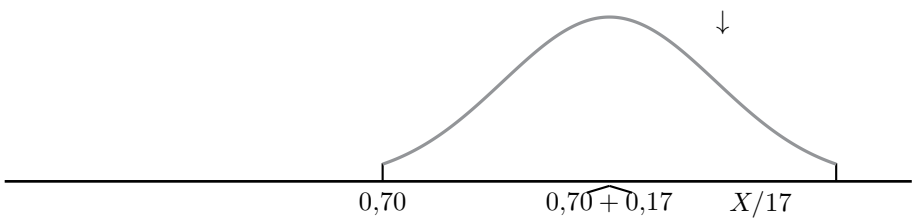


Figura 5.25. Campana recortada extrema hacia arriba que contiene el dato experimental.

Si movemos la campana hacia arriba del dato experimental, 12/17 = 0.70, la media de la campana puede moverse máximo 0.17. Es decir, la campana extrema hacia arriba tiene su media en 0.70 + 0.17.

Si movemos la campana hacia abajo y vamos hasta el extremo, vemos que el valor de la media correspondiente es 0.70-0.17.

.

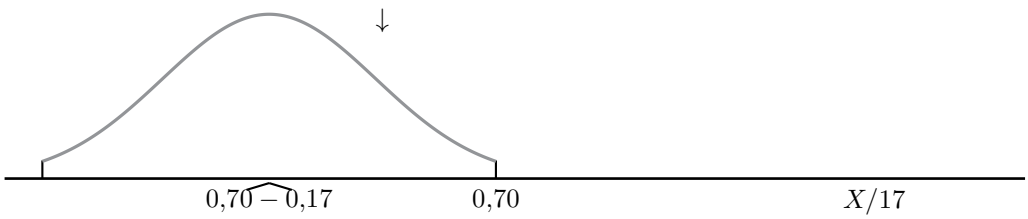


Figura 5.26. Campana recortada extrema hacia abajo que contiene el dato experimental.

Por tanto, el intervalo buscado empieza en el dato experimental, 0.70, menos 0.17 y termina en el dato experimental más 0.17. Eso da el intervalo

(0.70-0.17, 0.70+0.17) = (0.53, 0.87).

Paso 8: *Reporte de una respuesta atenta:* Cerramos el intervalo y tenemos en cuenta que las probabilidades van entre cero y uno, pues puede pasar que por las aproximaciones que hemos hecho nos den límites menores que cero o mayores que uno. Si es así, recortamos las puntas indeseables. En el caso presente no hubo puntas indeseables, por lo que la respuesta es: en el intervalo [0.53, 0.87] están todas las p que sustentan la $H_o : X \sim Bi(n = 17, p)$ con una significancia de 0.12 y con dos colas.

Paso 9: *Interpretación:* ¿Cuánto vale p, la probabilidad de cara, si de 17 lanzamientos se obtuvieron 12 caras? Respuesta: no sabemos. Pero con una confianza del 88 % p está en el intervalo [0.53, 0.87]. La p, poblacional, puede estar o no dentro de este intervalo, pero si apostamos rutinariamente a que p está en este intervalo, ganaremos el 88 % de las veces. Si quisiéramos aumentar la confianza, deberíamos disminuir la significancia y por tanto, el intervalo se expandería.

139 ♣ *Definición.* *El **IC**, intervalo de confianza, $1 - \alpha$ para p (la probabilidad de cara en una binomial) dada la frecuencia relativa observada $f = x/n$ es el intervalo que contiene todos los números q tales que $H_o : p = q$ se acepta con significancia α pues $f = x/n$ sería normal.*

5.9. Fabricando decisiones

Nosotros hemos desarrollado un criterio para escoger eventos que podrían servir como prometedores para una investigación extensa. Pero en algunas ocasiones de importancia biológica, ser especial según cierto descriptor podría implicar la diferencia entre vivir o morir. Por ejemplo, cuando hay falta de comida, un pájaro podría alimentar sólo al pollo más prometedor, digamos, al más fuerte, para que viva y el resto que aguante hambre y que se muera. Nosotros vemos que una decisión tomada por el pájaro padre, basado en cierta prueba cuantitativa, es una decisión que da vida a unos y condena a muerte a otros.

Un juego muy instructivo de decisiones que se relaciona con nuestra noción de interesante o extraño es el siguiente. El peso y el período de gestación de los bebés siguen una distribución de tendencia central. Los niños humanos que al nacer son demasiado livianos, están en la cola inferior, tienen más propensión hacia la muerte temprana en comparación con los niños normales o pesados. Hay también una variación por el tiempo de gestación y los bebés que nacen demasiado temprano también tienen una propensión hacia la muerte. Un doctor compasivo podría aconsejarle a la madre que aceptara la cruel realidad de la muerte inminente de su bebé que nació 4 meses antes de tiempo. No obstante, las madres normalmente no creen eso: ellas batallan desesperadamente para salvar la vida de sus bebés y, con la ayuda apropiada, ellas llegan a tener éxito.

Esas grandes victorias de vida son posibles porque el mecanismo de la generación de leche se dispara por el propio nacimiento y no cuando se complete el tiempo señalado de gestación que es de 9 meses. Es más, un bebé con cinco meses de gestación ya se ha desarrollado lo bastante como para digerir la

leche, pero él no puede todavía mamar y por consiguiente la madre se extraerá su leche y alimentará al bebé con un cuentagotas.

En conclusión, uno puede hacer una inferencia: los bebés muy prematuros se mueren. Pero las inferencias pueden ser erróneas: ¡Aunque los bebés prematuros se mueren, el mío vivió porque yo lo cuidé con todo mi instinto maternal! Cuando nos equivocamos en nuestras decisiones estadísticas o inferencias decimos que hemos cometido un **error de inferencia**. Hay dos tipos de tales errores, los mismos que uno puede cometer cuando uno juzga a alguien: uno puede declarar inocente a alguien culpable o uno puede declarar culpable a alguien inocente.

Cuando uno clasifica a alguien como anormal cuando en realidad no tiene nada de raro aparte de haber sido favorecido por el azar, uno comete un **error tipo** α. Y cuando uno clasifica como normal a alguien que es anormal pero con anormalidades que se compensan, uno comete un error **error tipo** β.

140 ***Ejemplo*** *Los submarinos.*

La altura de un grupo de muchachos, candidatos a trabajar en submarinos civiles, por ejemplo de oceanógrafos o de geofísicos, sigue una distribución normal $X \sim N(\mu = 170, \sigma = 10)$. Juan llegó a buscar trabajo en un submarino y el mide 1.98. ¿Le daremos trabajo a Juan?

Dilucidemos si se trata de dos colas o de una. En un submarino, las personas demasiado altas son los extraños, pues ocupan mucho espacio en los camarotes. A ellos se les rechazaría. Adicionalmente, los muchachos muy pequeños podrían ser hasta beneficiosos en muchas tareas que requieren un tamaño pequeño: ellos son bienvenidos. Así, los extraños son sólo las personas altas y nosotros estudiamos nuestra proposición que ha de ser rechazada con una sola cola, la superior.

Formulemos las hipótesis a discutir:
Ho : Juan es normal si $\alpha = 0{,}05$
Ha: Juan es un extraño. Lo rechazaremos por ser demasiado alto.

Puesto que la media es 1.70 y la desviación 0.10, y Juan mide 1.98, vemos que Juan está a casi 3 desviaciones estándar arriba de la media, por lo que es clasificado como alto, como extraño pero con dos colas. Ahora bien, todo lo que es extraño con dos colas también lo es con una: Juan es extraño con una sola cola y pertenece a los outliers de la cola superior.

El veredicto: nosotros rechazamos con una cola la hipótesis nula, Ho. Juan no es normal, él es un extraño, dado $\alpha = 0{,}05$.

Observemos nuestro idioma: nosotros operamos como jueces que toman decisiones que podrían afectar a otras personas. Por eso somos conscientes de nuestra responsabilidad. De hecho, la idea de trabajar en un submarino puede ser muy vivificante para muchos jóvenes. ¿Nosotros truncaríamos el sueño de Juan sólo porque él es alto? Sí, lo haríamos si así se requiriera. Nuestro deber es enseñarle a Juan que él podría volverse, digamos, un gran deportista o un guardabosques excelente o incluso entrar a la Infantería donde lo recibirían en el acto. Enfatizamos y nunca nos cansaremos de enfatizar que toda decisión conlleva una responsabilidad y que uno no puede ampararse en nada al momento de dar cuentas a sí mismo o a otros.

141 ***Principio de responsabilidad.*** *Cualquier ser humano es responsable de sus creencias y decisiones porque cualquier verdad es una creación de alguna mente por un acto injustificado de una voluntad libre.*

142 ***Ejemplo*** *Hay sol. Ésa es una verdad innegable. ¿En qué sentido nosotros somos responsables de esa creencia?*

Respuesta: Una catástrofe cósmica o ecológica podría cubrir la tierra con una espesa nube durante varias décadas y el sol dejaría de ser visible y la temperatura quizá quede bajo cero. Usted podría ocultar de los niños nacidos después de la hecatombe que allí arriba hay sol sobre la nube. O podría decirles: 'encima de esa nube, hay una enorme lámpara más poderosa que todas las lámparas juntas.' Pero cada uno de los niños respondería: ¿Me cree bobo? En ese momento usted experimentará el enorme peso que acarrea el creer que hay sol en el cielo.

143 ***En retrospectiva***

Debido a que lo que nosotros percibimos aparece como el resultado de muchos factores independientes, es común que los descriptores cuantitativos tengan una distribución de tendencia central, aproximable por una campana.

Las colas de esas distribuciones apuntan a los individuos interesantes para una investigación extensiva cuyo objetivo estaría en intentar descubrir el mecanismo que les hace ser extremos. Si es el mismo que explica a los individuos del centro de la distribución, no ha pasado nada. Pero quizá se descubra un mecanismo diferente y entonces hay avance en el conocimiento.

Pero las colas también son importantes en la naturaleza: ser o no ser un extremo podría volverse un asunto de vida o muerte. Esto significa que la estadística describe una parte fundamental de todo ambiente donde haya lucha por la vida, como lo es cualquier sociedad moderna. Y en esa lucha, se toman decisiones que no pocas veces se basan en clasificaciones de eventos como normales o anormales, según se pertenezca al centro de la distribución o a alguna cola. Hemos enfatizado que tales decisiones tienen carácter personal y que cada cual es responsable de ellas.

5.10. Protocolos y experimentos

144 ♣ ***Definición.*** *Una* ***condición inicial*** *es aquella que se establece al comenzar el registro de datos. Una* ***condición de frontera*** *es aquella que se conserva mientras el registro de datos se desarrolla. Un* ***protocolo o tratamiento*** *define un juego de condiciones iniciales o de frontera. Ejemplo: Póngase a un cultivo de brócoli a una temperatura de 13 grados centígrados a escuchar música de Bach.*

145 ♣ ***Definición.*** *Un* ***experimento*** *es el estudio de los cambios en alguna variable de salida, descriptor o cantidad que se mide, producidos por un cambio de tratamiento.*

Sucede que el número de descriptores puede ser infinito y por consiguiente también lo es el número de tipos de experimentos. Y todos son válidos. Sin embargo, hay experimentos típicos, fáciles de entender, fáciles de procesar. Los más sencillos se refieren a experimentos de productividad y a experimentos de homogeneidad. Asumimos que el descriptor es cuantitativo y que se registra como una variable aleatoria con valores x_i.

146 ♣ ***Definición.*** *Un* ***experimento de productividad media*** *es el que tiene como objetivo el estudio de los cambios en el promedio de la variable de salida. Ejemplo: ¿qué diferencia hay en el contenido promedio de vitamina C entre un cultivo de brócoli que escucha a Bach y otro que escucha a Debussy?*

147 ♣ ***Definición.*** *Un* ***experimento de homogeneidad*** *es el que tiene como objetivo el estudio de los cambios en la varianza de la variable de salida. Ejemplo: ¿por qué en sano juicio uno puede caminar en línea recta, homogéneamente, en cambio borracho uno va de un lado al otro y sin embargo cree que va derechito?*

5.11. Repetitividad y ruido

Una ciencia dura se precia de producir resultados repetibles. En mecánica clásica uno predecía el resultado exacto, pero en mecánica cuántica o en una ciencia social uno tiene la opción de predecir la forma de la distribución que se espera en la variable de salida.

No poder predecir resultados concretos sino tener que conformarse con resultados estadísticos es el precio que tenemos que pagar por hacer ciencia con lo cuántico y con las Ciencias Sociales. En lo cuántico, eso se debe a que las probabilidades son inherentes a la física fundamental. Pero en una ciencia social eso se debe a que cada subsistema estudiado es influenciado por un sinfín de variables que no podemos controlar, o quizá sí podamos pero nos sale muy caro en dinero, o tal vez no lo hagamos por un sentido ético.

148 ***Ruido*** *A todo lo que interfiere con un experimento y que no controlamos se le llama* ***ruido****.*

Acusación: Como no podemos controlar y ni siquiera determinar las condiciones iniciales, no tenemos protocolo, no tenemos experimento, no tenemos ciencia.

Defensa: la ciencia no consiste en saberlo todo ya, sino en un proceso que vaya develando la verdad y que quizá empiece con enunciados muy difusos. Pero la experiencia demuestra que la incertidumbre cede terreno al trabajo arduo y bien planeado. Pero, ¿por qué?

Porque para librarse del ruido, hay que permitírsele que se exprese, para que él se auto-aniquile, como en el silenciador de un carro. Eso significa algo muy concreto: entre más datos, será más fácil discernir lo sistemático de todo aquello que no podemos controlar.

149 *La binomial y el efecto del muestreo*

En muchas ocaciones una binomial sale naturalmente de una normal al poner una barrera: la llegada a clase es una v.a. que se distribuye normalmente, pero el que llegue después de las y 5 llega tarde. Y el que llegue antes no llega tarde. El tiempo de llegada depende de muchos factores no controlados, por ejemplo, del tráfico. Ahí está el azar. Pero una vez pasamos de la normal a la binomial, el azar debido a factores no controlados se desaparece y en su lugar aparece el efecto del muestreo: si uno toma 5 personas al azar, tal vez todos hayan llegado temprano, o quizá 3 ya llegaron y 2 estén por llegar. A este efecto se le llama **efecto del muestreo**. Pero en su raíz está el azar de todo lo que no se controla.

150 *Inferencia Estadística*

Consideremos un alegato entre Juan y Pedro. Juan lleva en Constitucional un promedio de 3.7 en los 3 parciales que han habido, y en cambio Pedro lleva un promedio de 3.5. Por eso Juan dijo: mis notas han sido mejores que las tuyas y eso demuestra que yo soy el mejor. A lo cual, Pedro respondió: tienes razón, tus notas han sido mejores que las mías, pues tú te preparaste para el examen pero yo me preparé para la vida, tratando de entender los porqués, lo cual es mucho más difícil. Pero ya veremos.

Nuestra apreciación de la situación Juan contra Pedro es la siguiente: es cierto que Juan fue el mejor hasta la fecha: en el pasado eso fue lo que pasó y no hay nada para remediarlo, es decir, no hay forma de añadir ni de quitar datos: el pasado está congelado.

Pero lo que Pedro tiene en mente es una inferencia estadística: en un futuro no muy lejano mis notas serán mejores que las tuyas pues mi método de estudio es muchísimo mejor que el tuyo.

Desde el punto de vista estadístico, toda la defensa de Pedro se resume a decir que él tiene la posibilidad de añadir datos y que por tanto las cosas pueden cambiar. Pero además, Pedro alega implícitamente que por más que el futuro traiga datos, todos ellos no harán más que respaldar su estrategia de estudio, entender los porqués, la cual es mejor que la de Juan, aprender procedimientos de memoria.

En general, el objetivo de la Estadística es utilizar unos pocos datos para inferir conclusiones robustas que sean válidas a largo plazo cuando se añada un gran número de datos. Si no hay la posibilidad de añadir una gran cantidad de datos, no hay problema estadístico que resolver. Eso se debe a que con pocos datos uno tendrá que aceptar la cruda realidad de lo que fue, pero no podrá decir nada de lo que pudo haber sido, pues por carecer de datos no se podrá discernir el efecto de lo que no puede controlarse.

En términos oficiales, a los pocos datos se les llama **muestra** y a todos los datos posibles juntos se les llama **población**, y uno debe esforzarse para que el muestreo sea al azar. El objetivo de la **estadística inferencial** es decidir si una conclusión que puede ser válida para la muestra se puede aplicar a toda la población.

Uno hace inferencias a toda hora. Por ejemplo, desde que acompañé a Paula el día que se sentía muy triste, ella se ha mostrado más cariñosa conmigo. Eso es un hecho, y como tal no puede cambiarse, ni nada. Pero de ese hecho uno tiende a inferir que Paula es agradecida, lo cual es una aseveración con poder predictivo: puedo cultivar la amistad de Paula acompañándola cuando se note triste. O uno puede inferir que a las mujeres les sienta bien que las acompañen cuando están aburridas, lo cual pienso aplicar con Marcela para conquistarla. El esquema es siempre el mismo: uno saca conclusiones de unos pocos datos que luego aplica a gran escala. Nuestro objetivo es ver cómo se hace eso de tal manera que uno pueda sustentar sus creencias, conclusiones o decisiones ante el mundo entero y sobre todo en la esfera académica.

Entremos ahora a considerar el efecto del ruido: si uno no puede controlar el universo entero, ¿cómo se hace para sacar conclusiones robustas que sean válidas en presencia de tantos factores que uno no puede controlar?

151 ◇ ***Primer teorema fundamental del método científico.*** *Entre más datos, más preciso será un experimento: el ruido será un problema en los experimentos pequeños, pero sus efectos acumulativos se cancelarán unos con otros en los experimentos grandes. Por eso, podemos esperar más exactitud, más certidumbre de un experimento grande con muchos datos, que de uno pequeño.*

Nuestro teorema está lejos de ser una entidad bien definida matemáticamente, pero podemos entender su espíritu con un razonamiento simple: el ruido es aleatorio, por lo tanto, a veces tirará para la derecha y a veces para la izquierda y terminará auto-aniquilándose. El trabajo que hemos hecho sobre la binomial nos ayuda a entender todo eso mucho mejor.

Recordemos que si $X \sim Bi(n,p)$, la media de Xes $\mu = np$, la varianza es $V = np(1-p)$ y la desviación estándar es $\sigma = \sqrt{np(1-p)}$. Con esto y aproximando la binomial por la normal podemos estimar los outliers para el nivel de significancia $\alpha = 0{,}05$: un individuo es un outlier si su distancia a la media es más grande que 2σ.

152 ***Ejemplo*** *Calculemos los outliers para $Bi(n, p = 1/2)$ en los casos $n = 4, 8, 64$.*

Para $n = 4$, $\mu = 4(1/2) = 2$, $\sigma = \sqrt{4(1/2)(1/2)} = 1$. Así que la frontera de la región que comprende a los individuos normales estará en $2 \pm 2(1) = 2 \pm 2$. Así, para $n = 4$, los outliers son menores que 0, o más grandes que 4. No hay ningún outlier por consiguiente. El porcentaje de eventos que clasifican como outliers e improbables es cero.

Para $n = 8$, $\mu = 8(1/2) = 4$, $\sigma = \sqrt{8(1/2)(1/2)} = \sqrt{2} = 1{,}4$. Así que la frontera de la región que comprende a los individuos normales estará en $4 \pm 2(\sqrt{2}) = 4 \pm 2{,}8$. Así, para $n = 8$, los outliers son menores que 1.2 o más grandes que 6.8. Por consiguiente, ellos son 0, 1, 7 y 8. El porcentaje de eventos que clasifican como outliers e improbables es 4 entre 9, o sea el 44.4 %.

Para $n = 64$, $\mu = 64(1/2) = 32$, $\sigma = \sqrt{64(1/2)(1/2)} = \sqrt{16} = 4$. Así que la frontera de la región que comprende a los individuos normales estará en 32 ± 8. Así, para $n = 64$, los outliers son menores que 24 o mayores que 40. Son: 0, 1, 2, 3.. ..23, 41, 42,.... 64. El porcentaje de eventos que clasifican como outliers e improbables es 48 entre 64 o 75 %.

En resumen, los eventos de una binomial tienden a acumularse alrededor de la media, pero entre más datos, entre más monedas se tiren, el grado de concentración es mayor de tal forma que es improbable que el número de caras k en un experimento difiera significativamente de la media esperada $\mu = np$. Dicho de otra manera: como $k \sim np$, entonce $k/n \sim p$. Es decir, uno puede estimar p la probabilidad de cara lanzando n monedas y contando el número de caras, k. En ese caso, p es aproximadamente k/n, pero entre más datos, entre más monedas se tiren, mejor es la aproximación.

153 ***Ruido***

Ahora veamos y captemos que la binomial es un modelo de ruido. Eso se hace así: la binomial viene de lanzar monedas. Antes contábamos el número de caras. Pero para interpretar a una binomial como ruido hay que seguir otro protocolo: supongamos que una moneda puede caer aleatoriamente ya sea sobre la alfombra o ya sea sobre el piso de baldosín y que ambos eventos tienen igual probabilidad, o sea 1/2. Cuando la moneda cae en la alfombra no hace ruido, pero cuando cae en el baldosín hace mucho ruido. Cuándo hace ruido y cuándo no es algo que no se puede predecir. Tenemos una fuente de ruido aleatorio. Si tenemos varias monedas, los diversos ruidos se suman y la intensidad del ruido total tiene una distribución binomial.

Démonos cuenta del juego de palabras utilizado en la argumentación anterior: la tarea era producir un modelo de ruido estadístico pero lo que hicimos fue encontrar un modelo de ruido acústico. ¿Acaso nos perdimos por el camino?

Lo que pasa es esto: el ruido acústico interfiere con la música de forma similar a como el ruido estadístico interfiere con los datos que uno quiere medir y registrar en el experimento. O si uno prefiere: si en nuestro experimento, todo lo que uno desea es registrar la música, el ruido acústico le mete

interferencia. Y análogamente a las técnicas de filtros de ruido que hay en tecnología de sonido, también tenemos en estadística una técnica para filtrar ruido: hacer que el ruido se autoexprese para que termine auto aniquilándose en su efecto total.

Al modelar el ruido, uno cuadra las cosas para que la media sea cero (se hace una translación). Eso implica, que cuando el número de eventos estudiados crece, el ruido tiende a auto-aniquilarse y el ruido total tiende entonces a concentrarse alrededor de cero, es decir, al rededor de nada, el ruido termina auto-aniquilándose.

Esta noción de ruido nos lleva a considerar un modelo muy popular de lo que es un dato experimental x_i, digamos lo que un estudiante saca en un parcial. El dato es la suma de un efecto determinista, μ_i, el cual es más o menos predecible por el estudio y la comprensión del estudiante, más un efecto del azar debido a que le preguntaron lo que estudió a última hora o tal vez porque no le preguntaron lo que estudió. Escribimos eso como

$$x_i = \mu_i + ruido_i$$

Si además, el $ruido_i$ sigue una distribución normal con media cero, entonces, μ_i es la media de x_i: un dato es una variable aleatoria distribuida normalmente con media μ_i y con desviación correspondiente σ_i. Como vemos, el valor medio de la distribución aparece naturalmente en la descripción de lo que es observado. Por esa razón, los experimentos que tienen que ver con la media, o sea con la productividad media, son los que primero se analizan. Sucede además que para hacerlo es necesario también analizar el comportamiento de la varianza, por lo que esos dos estadígrafos se analizan en compañía. Para entender todo eso, la clave está en ir paso a paso.

5.12. Conclusión

Toda la ciencia tiene un único fundamento biológico: la tendencia innata de todo ser humano de fijar su atención sobre los eventos que en una distribución central están muy alejados de la media. Clasificando los eventos de una distribución de tendencia central como normales, cerca de la media, o anormales, alejados de ella, podemos empezar a tomar decisiones. Puede que nuestras decisiones estén equivocadas y ese es el precio que tenemos que pagar por haber cambiado el concepto de la verdad por el de la probabilidad. Nuestro difícil desafío es minimizar a un precio razonable la proporción de decisiones equivocadas.

Capítulo 6

La media y el experimento básico

154 ***Introducción.***

En el capítulo pasado aprendimos el formalismo de la modernidad para expresiones del tipo: !Que alto que es Giga-Juan! En resumen, lo que hicimos fue lo siguiente: hicimos un muestreo aleatorio de las personas de la edad de Juan, les medimos la altura, agrupamos por rangos, hicimos la tabla de frecuencias absolutas, graficamos el histograma correspondiente, hallamos la campana envolvente que mejor se aproxima, le calculamos a dicha campana la media y la desviación. Este proceso de abstracción da una inferencia o extrapolación sobre la población y por lo tanto la media es poblacional y se denota μ y la desviación es poblacional y se denota σ. Todo este procedimiento o protocolo se resume escribiendo $X \sim N(\mu, \sigma)$ que se lee: las alturas de las personas definen una variable aleatoria con distribución normal de media μ y desviación σ.

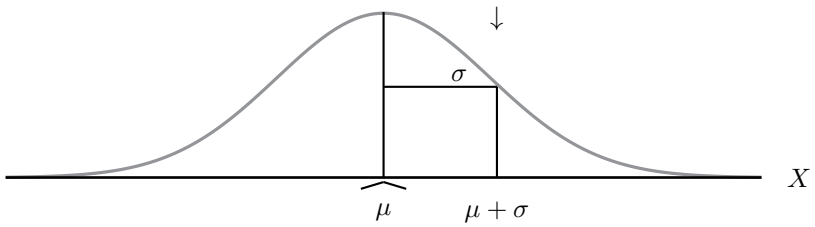


Figura 6.0. $X \sim N(\mu, \sigma)$: la variable aleatoria X se distribuye normalmente con media $\mu_X = \mu$ y desviación estándar $\sigma_X = \sigma$.

Luego, a la significancia por defecto de 0.05, que es un tridente por ser un área o proporción de lo que se considera extremo, le hallamos z_c, la z-crítica correspondiente de la cola superior, la cual nos permitió hallar x_c, el x-crítico correspondiente. Todo lo que quede arriba de x_c es extremo por arriba, todo lo que quede por abajo, es considerado normal. Giga-Juan obviamente queda muy por arriba y concluímos que es justo que se le llame Giga.

155 ***Objetivo*** *Aprenderemos el formalismo de la modernidad para expresiones del tipo: !Que alto que es el equipo de basquet de Giga-Juan!. Para ello tenemos que comparar el promedio de las alturas del equipo de Giga-Juan con todos los promedios posibles habidos y por haber de grupos de 9 elementos tomados al azar de la población de basquetbolistas. Delicado problema. La respuesta ya nos la calcularon los matemáticos y la ilustramos enseguida.*

156 ***Ejemplo*** *Se trata de armar equipos de básquet, deporte para el cual la altura es una variable importante. Una cierta población de básquetbolistas tiene miembros cuyas alturas se distribuyen normalmente $X \sim N(\mu_X = 175, \sigma_X = 10)$, unidades en centímetros. Si todos los equipos tienen 9 miembros, ¿cómo se distribuye la estatura promedio de los equipos si consideramos que al momento de formar equipos los lazos de amistad son más decisivos que la altura misma y que por lo tanto las alturas pueden considerarse al azar?*

Solución: La altura de las personas se describe por $X \sim N(\mu_X = 175, \sigma_X = 10)$. *Nos preguntan cómo se distribuye la altura promedio de muestras al azar con 9 elementos. Es decir, nos piden que hallemos la distribución de* $\bar{X}_9$ *que es la* ***media muestral de 9 elementos****. Sucede que cuando uno saca un dato, ese dato puede estar en casi cualquier parte con respecto a la media. Pero la media de varios datos no puede quedar en un extremo de la muestra y por tanto queda más cerca de la media poblacional que algunos datos. Por ejemplo, si los datos son 170, 180, 175, 190 con promedio 179, vemos que dicho promedio está más cerca de la media, 175, que 170, 180 y 190. Y con 9 datos, no sería raro que el promedio quedara más cerca de la media que unos 6 datos. Por tanto, esperamos que la desviación de la media muestral* $\bar{X}_9$ *sea menor que la de* X *y así la campana de* $\bar{X}_9$ *será más cerrada que la de* X.

El TLC, teorema del límite central (formalmente enunciado en el numeral anterior), trasforma estos comentarios cualitativos en un enunciado cuantitativo: la media de $\bar{X}_n$, *el promedio de muestras aleatorias de* n *elementos, es la misma media de* X, *en tanto que la desviación de* $\bar{X}_n$ *es la desviación de* X *dividida por la raíz cuadrada de* n, *el número de datos de cada muestra.*

En términos de los equipos de básquet tenemos: los equipos de básquet tendrán una altura promedio que varía de equipo a equipo pero de tal forma que la media de dichos promedios es la misma media de la población. Por otro lado, la desviación de la altura promedio de los equipos es la desviación de la altura de la población dividida por la raíz de 9. En notación matemática:

$\mu_{\bar{X}_9} = \mu_X = 175$

$\sigma_{\bar{X}_9} = \frac{\sigma_X}{\sqrt{9}} = \frac{10}{\sqrt{9}} = 3{,}33$

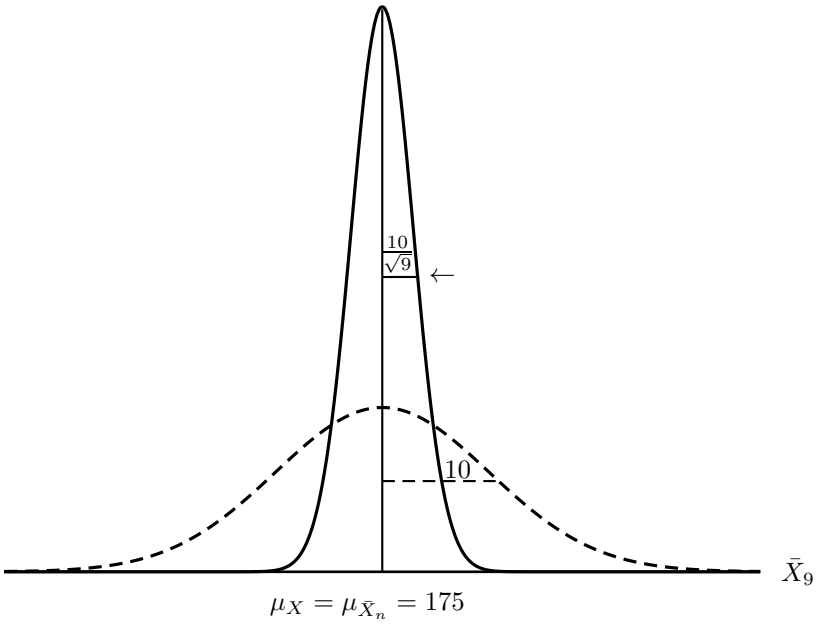


Figura 6.1. En línea punteada vemos la campana de X, la altura de los jugadores individuales. En la línea sólida vemos la campana de la altura promedio de los equipos de básquet de nueve miembros. La altura promedio de la altura de los jugadores es $\mu_X = 175$ y su desviación es $\sigma_X = 10$. La altura promedio de los equipos de 9 jugadores es igual a la altura media de los jugadores: $\mu_{\bar{X}_9} = \mu_X = 175$. Las medias de las dos campanas coinciden: la media de X es igual a la media de sus medias muestrales. Pero las desviaciones son diferentes: $\sigma_{\bar{X}_9} = \frac{\sigma_X}{\sqrt{9}} = \frac{10}{3} = 3{,}33$.

En general, la altura promedio por equipo está mucho más cerca de la media poblacional que los datos mismos, pues los valores extremos que salen por azar se balancean unos con otros. Vemos que los promedios de n *elementos también tienen una distribución normal con la misma media que los datos aislados pero con una desvación menor. Que tan menor nos lo dice el teorema del límite central formalizado enseguida. Sabiendo cómo es la distribución de los promedios de, en nuestro caso, 9 elementos, ya podemos juzgar si el equipo de Bsquet de Giga Juan es verdaderamente un outlier en su mundo de equipos.*

6.1. El teorema del límite central: TLC

De ahora en adelante supondremos que nuestras variables son aleatorias e independientes y que están normalmente distribuidas, y además que el proceso de muestreo es al azar y que el universo no cambia.

El **teorema del límite central** es la herramienta matemática para estudiar proposiciones con respeto a la media de varios datos. Por ejemplo, un significado del teorema del límite central es el siguiente: entre más grande sea la muestra aleatoria, más se acerca la media de la muestra a la media de la población. Formalicemos todo esto paso a paso:

157 ♣ ***Definición.*** *Cuando uno fija un programa científico, uno define el objeto de estudio o* ***población****. Por ejemplo: las señoras que llevan más de 10 años de casadas.*

158 ♣ ***Definición.*** *El protocolo que define una* ***variable aleatoria*** *X consta de a) una elección de lo que se le va a medir a la población, b) hacer un muestreo al azar c) hacer la medición y d) registrar dato por dato. Por ejemplo: a las señoras con más de 10 años de casadas se les mide el tiempo que de todo ese período dedicaron a estudiar. Asumimos que el proceso de registro no cambia las probabilidades de los demás eventos. Eso no pasa si la población tiene 4 datos equiprobables: al sacar un dato, las probabilidades de los otros datos cambian de 1/4 a 1/3. A la media de la variable aleatoria X se denota $\mu = \mu_X$ y a su varianza $\sigma^2 = \sigma_X^2$.*

159 ♣ ***Definición.*** *El protocolo que define una* ***muestra de*** *n elementos es el siguiente: se revuelve la población para que el muestreo sea al azar y se saca un puñado de n elementos. Es bueno garantizar que el muestreo no perturbe la población estudiada, que las probabilidades de los demás eventos no cambien y que no se creen correlaciones que antes no existían.*

160 ♣ ***Definición.*** *El protocolo que define la* ***media muestral de tamaño n*** *es el siguiente: se toma una muestra al azar de tamaño n, se mide cada objeto y se promedian los n datos. Eso da un único número, $\bar{x}_1$. Y se repite eso sin límite, $\bar{x}_2 ... \bar{x}_i ...$, para hacer el histograma y después aproximar por una campana. Esta nueva v.a. se llama la media muestral de n elementos y se denota $\bar{X}_n$. Para no correr el riesgo de que las probabilidades cambien al hacer el muestreo, los objetos se devuelven a la población una vez que se han medido y promediado. Se dice que el muestreo es* ***con reemplazo****.*

161 ◇ ***Observación:*** *Si X es una Campana de Gauss, entonces todo cuadra perfecto y la media muestral también sigue una distribución normal. Para distribuciones de tendencia central la media mustral es casi una campana y es usual aplicar la misma tecnología.*

162 ◇ ***TLC: teorema del límite central:*** *Si $X \sim N(\mu_X, \sigma_X)$ entonces $\bar{X}_n \sim N(\mu_X, \frac{\sigma_X}{\sqrt{n}})$. Atención: aunque X no tenga una distribución normal, la distribución de la media muestral $\bar{X}_n$ va aproximándose a una normal $N(\mu_X, \frac{\sigma_X}{\sqrt{n}})$ tanto más cuanto mayor sea el número de datos. Al término $\frac{\sigma_X}{\sqrt{n}}$ se le llama* ***error estándar de la estimación de la media*** *y es igual a la desviación estándar de la media muestral $\bar{X}_n$.*

Tratemos de entender el enunciado:

En la siguiente gráfica vemos una variable aleatoria X, digamos que se trate de la altura de una población muy grande de muchachos, la cual es normalmente distribuida:

Ahora bien, si de la población de muchachos sacamos al azar una puñada con n elementos y los medimos y sacamos el promedio, nos da un número: el promedio de la muestra. Si volvemos a repetir lo mismo nos dará otro número y al repetir y repetir nos dará $\bar{X}_n$, la **media muestral**, la cual también se distribuye normalmente.

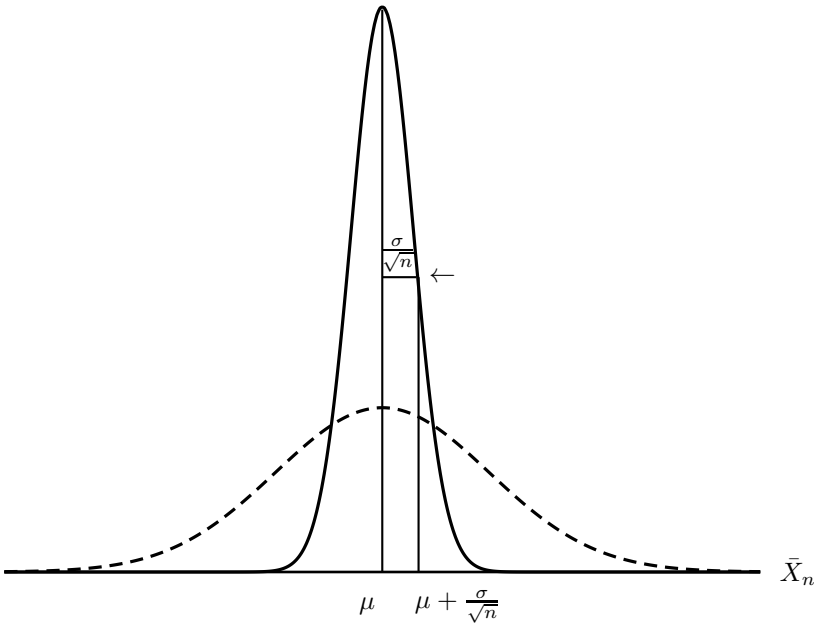


Figura 6.2. La variable aleatoria $\bar{X}_n$ se distribuye normalmente con media $\mu_{\bar{X}_n} = \mu_X = \mu$ y desviación estándar $\sigma_{\bar{X}_n} = \frac{\sigma_X}{\sqrt{n}}$. La flechita señala el punto de inflexión. En línea punteada vemos la campana de X, mucho más extendida. Las medias de las dos campanas coinciden: la media de X es igual a la media de sus medias muestrales. Pero las desviaciones son diferentes: mientras que los valores de X podrían estar casi en cualquier lado, los promedios de las muestras se parecen a la media poblacional y se agrupan cerca de ésta. Por tanto, la campana de las medias muestrales es más cerrada que la de X, tanto más cuanto mayor sea el número de datos. Y como es más cerrada, pues debe ser más alta.

Al promediar, se promedian los extremos y los diversos resultados salen más parecidos a la media poblacional de X que los datos mismos. Es decir, la campana de $\bar{X}_n$ se cierra, se comprime, sobre su media, la cual es la misma media de X pero su desviación y varianza disminuyen. **Mnemotecnia: al hacer una medición de la media hay que tener en cuenta dos factores. Primero: la variabilidad de lo que se mide debido al azar, que viene de todo lo que no se controla y que se representa por la desviación. Segundo: que entre más datos, menos incertidumbre. Todo junto se representa por la desviación de la media muestral que es la desviación original dividida por la raíz del número de datos.**

163 ***Ejemplo: Proporción de medias muestrales entre dos valores.*** *Si $X \sim N(\mu = 11, \sigma = 2)$ y si seleccionan muestras al azar de 12 eventos, hallemos la proporción de medias muestrales que estarán entre 10 y 14.*

La media muestral $\bar{X}_n$ tiene por el TLC una distribución normal con media 11 y desviación $2/\sqrt{12} = 0{,}57$. Por consiguiente, la transformada-z de 10 es:
$z_1 = (10 - 11)/0{,}57 = -1{,}75$.
Y la transformada-z de 14 es
$z_2 = (14 - 11)/0{,}57 = 5{,}26$
La probabilidad de que la media muestral esté entre 10 y 14 es: $\Psi(5{,}26) - \Psi(-1{,}75) = 1 - 0{,}04 = 0{,}96$.

164 ***Ejemplo: Test sobre una media con varios datos. Frío garantizado.*** *Un gran cuarto de refrigeración viene garantizado para $5°C$ en promedio y con una $\sigma = 4$. Una muestra al azar de 16 eventos dio una media de 8. ¿Es justa la garantía?*

Solución: **Paso 1: Determinamos el número de colas y la significancia.** La garantía es justa cuando los desajustes con respecto a la media no son significativos. El problema es con dos colas, pues la temperatura puede desequilibrarse por arriba o por abajo y ambos desajustes pueden ser indeseables. Tomamos $\alpha = 0{,}05$, lo cual con dos colas da una $z_c = 1{,}96$, o sea1,96 desviaciones estándar:

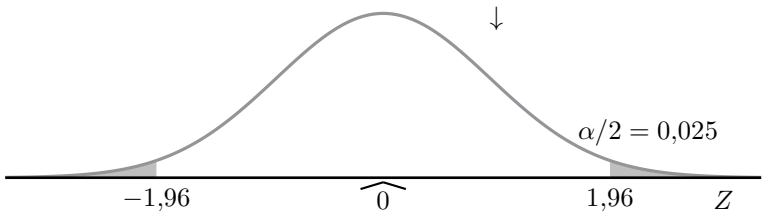


Figura 6.3. Todos los eventos extremos con dos colas forman un conjunto cuya probabilidad es 0.05.

Paso 2: Formulamos la hipótesis nula y la alterna.

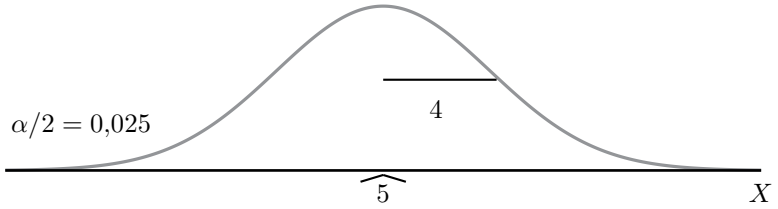


Figura 6.4. La temperatura X se distribuye normalmente con media 5 y desviación 4.

$H_o : \mu = 5$ (si no fuese por el azar, todos los datos serían 5).
$H_a : \mu \neq 5$ (el frío o el calor del exterior hace que los datos difieran sistemáticamente de 5.)
El desigual de la H_a dice que el problema se decide con dos colas, pues hay dos maneras de ser desigual, por arriba o por abajo. La significancia es 0.05, lo cual da $z = 1{,}96$, es decir, trabajamos con 1.96 desviaciones.

Paso 3: Fabricamos el mundo natural al problema. El resultado experimental reportado fue $\bar{X}_{16} = 8$. Este resultado no cabe en el histograma de la X, pues ahí van sólo datos aislados. Tampoco cabe en la Z, pues la Z es sólo una ayuda para el cálculo. Por tanto, hay que buscarle a ese resultado un mundo especial donde se halle a gusto. Ese mundo es el histograma de la v.a. $\bar{X}_{16}$. Ahora bien, la media de X fue 5 y la desviación de X fue 4: $\sigma_X = 4$, y como entre más datos menos incertidumbre, tenemos por el TLC que la media de la media muestral $\bar{X}_{16}$ es 5 y que la desviación de la media muestral $\bar{X}_{16}$ es

$\sigma_{\bar{X}_{16}} = \frac{4}{\sqrt{16}} = 1.$

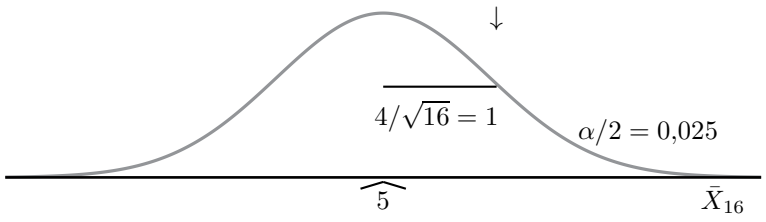


Figura 6.5. La variable $\bar{X}_{16}$ tiene media 5 y desviación 1. La desviación de $\bar{X}_{16}$ es más pequeña que la de X, pues 16 datos tomados al azar producirán una media que quedará mucho más cerca de la media poblacional que un dato aislado. Y en general, entre más datos, la campana de $\bar{X}_n$ más se cierra y por tanto la incertidumbre en la estimación de la media disminuye.

Paso 4: Calculamos las predicciones de la H_o, es decir, su zona de aceptación.
Como la z crítica fue 1.96, la región de aceptación de la H_o va desde 1.96 desviaciones abajo de la media hasta 1.96 desviaciones arriba de la media, pero en el espacio donde vive la v.a. $\bar{X}_{16}$, en el cual la desviación es 1, según el TLC. La región de aceptación es entonces el intervalo:
(5- 1.96(1), 5+1.96(1)) = (3.04, 6.96).

Es decir, si la H_o es cierta, las muestras de 16 datos tendrían un promedio que debería estar en el intervalo (3.04, 6.96).

Paso 5: Veredicto. Contrastamos la predicción de la H_o con el valor experimental que dice que la media de la muestra de 16 eventos fue $\bar{X}_{16} = 8$. Dicho valor experimental no pertenece a la región de aceptación: la H_o se rechaza.

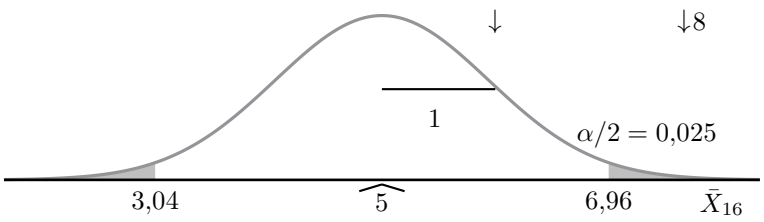


Figura 6.6. La variable $\bar{X}_{16}$ tiene media 5 y desviación 1. Lo normal va desde 3.04 hasta 6.94. El valor 8 es anormal.

Paso 6: Interpretación. Los datos experimentales nos dan pie para rechazar la tesis de que $\mu_X = 5$. Es decir, no es cómodo creer que la discrepancia entre el resultado experimental y la media predicha por la H_0 se deba tan sólo al azar. Es más cómodo creer que hay algo sistemático afectando el sistema: lo más probable es que el sistema regulador de la temperatura del cuarto de enfriamiento es de muy baja potencia y se deja ganar del aire circundante caliente. La garantía no es justa. Podemos reclamar.

165 ***Ejemplo*** *Un grupo de 30 mujeres que venían de un país lejano entró a un almacén a comprar zapatos, pero no había zapatos para ellas, pues su talla media era $\bar{X}_{30} = 40$ en tanto que la talla de las mujeres que forman la clientela, X, es una variable aleatoria con distribución normal tal que $\mu_X = 36$, $\sigma_X = 2$. ¿Es justo que regañen al Administrador por no haber previsto que llegarían mujeres de gran talla?*

Las hipótesis en estudio son:
$H_o : \mu = 36$
$H_a : \mu > 36$
El problema se resuelve con una cola, la superior, pues la pelea es por las mujeres de gran talla. Con $\alpha = 0{,}05$, la $z_c = 1{,}6$, es decir, 1.6 desviaciones estándar.

Usamos el TLC: la variable aleatoria media muestral tiene media 36, igual que la media de la población. Para calcular la desviación estándar de la media muestral usamos la fórmula:

$\sigma_{\bar{X}_n} = \frac{\sigma_X}{\sqrt{n}} = 2/\sqrt{30} = 0{,}37$. Por tanto
$(\bar{x}_{30})_c = \mu + z\sigma = 36 + 1{,}6 \times 0{,}37 = 36{,}592 \sim 37$.

Eso significa que el almacén debería estar en capacidad de atender a un grupo de 30 mujeres cuya talla promedio fuese menor que 37. Pero como el promedio de la talla de las mujeres turistas es de 40 deducimos que eso no pertenece a lo que la compañía debe considerar como plausible. Dicho formalmente, rechazamos la H_o con cola superior.

Veredicto: el administrador es inocente de toda culpa, pues el turismo de mujeres muy altas no es lo suficientemente fuerte como para generar un mercado específico. No sería ese el caso si las mujeres calzaran 37 y el administrador no tuviese zapatos para ellas, (suponiendo que no hay problemas de almacenamiento, ni por bodegas, ni por presupuesto).

166 ***Ejemplo*** *Un conectivo deductivo es una expresión del tipo: por lo tanto, se deduce que, podemos inferir, puesto que. En español, la palabra* `entonces` *a veces es conectivo deductivo, pero a veces no como en la siguiente expresión* `almorzamos y entonces pagamos y entonces nos fuimos.`

Para medir el estilo deductivo de un filósofo se cuenta el número de veces que utiliza un conectivo deductivo por cada 100 palabras. Para el filósofo ABC durante sus escritos primeros, la cantidad de conectivos deductivos por cada 100 palabras tiene una distribución normal con media 10 y desviación 3. En su último escrito, se seleccionó un texto al azar de 900 palabras y se encontró que había un

promedio de 13 conectivos deductivos por cada 100 palabras. Preguntamos: ¿Podemos deducir que ahora el joven filósofo es menos dogmático y más argumentativo? ¿Tendrá eso algo que ver con un incremento sustancial en las ventas de sus libros?

Solución:
Las hipótesis en estudio son:
$H_o : \mu = 10$
$H_a : \mu > 10$
Como no se dijo nada sobre la significancia de las decisiones, la tomamos igual al 5 %. El problema es con una cola, la superior. Eso significa que todo lo que esté arriba de 1.6 desviaciones estándares se toma como outlier.

Sea X la v.a. que denota el número de conectivos usados por cada 100 palabras.
Datos: $\mu_X = 10$; $\sigma_X = 3$
Tenemos que considerar también la media muestral de subíndice 9 (pues se analizarion 900 palabras): de acuerdo al teorema del límite central, su media es la misma que la de X, o sea 10, y su desviación estándar es $3/\sqrt{9} = 1$.

Puesto que la media del histograma de la media muestral es 10 y su desviación es 1, todo lo que quede arriba de $10 + 1{,}6 \times 1 = 11{,}6 = 12$ es extraño. Como nuestro dato experimental (de 9 eventos) dio media 13, eso resulta ser un dato extraño, fuera de serie. Por lo tanto, inferimos que ahora hay en promedio más conectivos deductivos que antes y no por el azar que corresponde a tomar muestras de 9 elementos.

Veredicto: se deduce que el escritor está esforzándose por mejorar su estilo y lo menos que se puede decir es que ahora es más fluido, conectado y posiblemente más coherente. Seguramente también mejoró otras cualidades, por lo que no es azar que las ventas hayan aumentado. En vista de que el joven filósofo ya logró entrar en el mercado, es buena inversión que se le compre la producción a futuro para que pueda dedicarse a escribir con más libertad y así todos ganemos.

6.2. El IC para la media

Nosotros ya sabemos cómo se halla el intervalo de confianza para la media cuando hay un dato experimental. Tal situación es en verdad rara, más usual es que uno tenga varios datos, lo cual permite hacer una estimación más precisa de la media.

167 ***Ejemplo: Intervalo de confianza para una media con varios datos.*** *Hallemos el intervalo de confianza del 99 % (con $\alpha = 0{,}01$) para la media de X, con distribución normal, la cual tiene $\sigma_X = 3{,}2$ si la media de 9 datos fue 14. Presentamos dos soluciones, una muy intuitiva, geométrica, y otra muy oficial, algebraica.*

Solución geométrica:
Paso 1: *Datos del problema.* Hallar el IC del 99 % de confianza (con $\alpha = 0{,}01$) para la media de $X \sim N(\mu = ?, \sigma = 3{,}2)$, si $\bar{X}_9 = 14$. EL IC contiene para la significancia dada todas las μ que son compatibles con el dato experimental $\bar{X}_9 = 14$, es decir, todas las μ que definen una campana para la cual el dato experimental $\bar{X}_9 = 14$ es común y corriente. EL IC responde a la pregunta: si se tiene que el promedio de 9 datos fue 14, ¿cuánto vale la media? Respuesta: No sabemos, pero con una confianza del 99 %, la media queda en el IC que ya mismo averiguamos.

Paso 2: *Buscamos el mundo adecuado para el experimento.* El resultado experimental $\bar{X}_9 = 14$ no cabe en el histograma de la X, pues ahí van sólo datos aislados. Tampoco cabe en la Z, pues la Z es sólo una ayuda para el cálculo. Por tanto, hay que buscarle a ese resultado un mundo especial donde se halle en casa. Ese mundo es el de la campana de la v.a. $\bar{X}_9$.

Paso 3: *Usamos el TLC para saber la distribución de la media muestral* $\bar{X}_9$. Tiene una distribución normal y como media a la misma media de X, es decir $\mu_{\bar{X}_9} = \mu_X$. Y también tenemos que $\sigma_{\bar{X}_n} = \frac{\sigma_X}{\sqrt{n}}$. En este caso, $\sigma_X = 3{,}2$ por lo que $\sigma_{\bar{X}_9} = 3{,}2/\sqrt{9} = 1{,}06$.

Paso 4: *Determinamos el z crítico asociado a la significancia.* Sobre la Z, a la significancia $\alpha = 0{,}01$ le debe corresponder un valor de z crítico. Pero observemos que la media puede estar abajo o arriba del dato experimental, por eso, el problema se resuelve con dos colas. Dividimos por tanto 0.01 entre

dos y miramos en la tabla de la Z el z correspondiente a una área de 0.005. Da -2.6, lo cual dice que trabajamos con 2.6 desviaciones estándar.

Paso 5: *Para una μ fija, determinemos lo normal.* La campana asociada a $\bar{X}_9$ tiene media desconocida $\mu_{\bar{X}_9}$ y desviación $\sigma_{\bar{X}_9} = 1{,}06$. Como el z crítico nos dio 2.6, eso significa que sobre $\bar{X}_9$, los valores anormales empiezan en la media más o menos 2.6 veces la desviación estándar, es decir $2{,}6 \times 1{,}06 = \pm 2{,}77$.

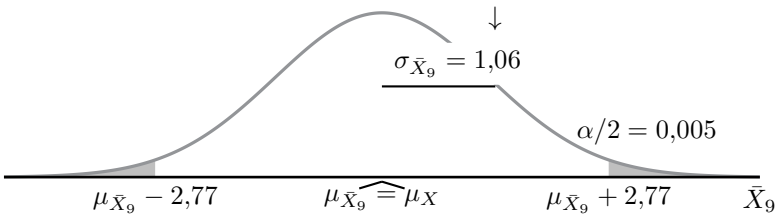


Figura 6.7. Con $\alpha = 0{,}01$ y dos colas, los eventos comunes y corrientes en la campana de la $\bar{X}_9$ van desde desde $\mu - 2{,}77$ hasta $\mu + 2{,}77$.

Paso 6: *Fabricación de la campana recortada.* Recortamos de la campana las colas que contienen todos los eventos anormales. La campana recortada tiene centro en la media $\mu_{\bar{X}_9}$ y tiene a cada lado una aleta que mide 2.77.

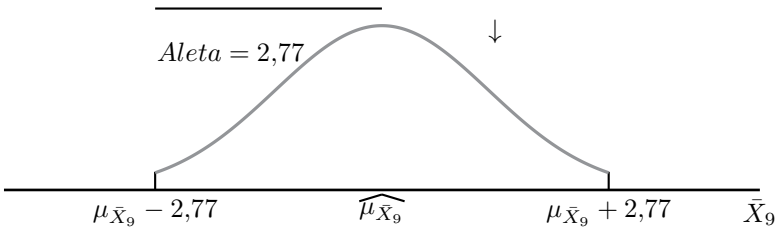


Figura 6.8. La campana recortada contiene todos los eventos normales, comunes y corrientes.

Paso 7: *Contraste con el dato experimental.* Tomamos la campana recortada y la ponemos encima del dato experimental, queriendo decir que la campana y el dato experimental son compatibles: un dato experimental es compatible con una campana cuando el dato clasifica como algo normal, cuya diferencia con la media se pueda explicar por azar y no por algún efecto sistemático que cree un evento extraordinario.

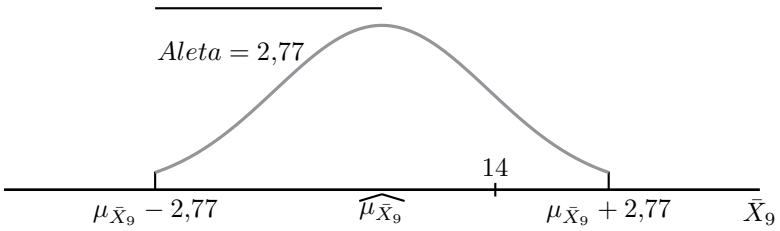


Figura 6.9. El dato experimental $\bar{X}_9 = 14$ es normal para la campana dada. La distancia entre la media y el dato experimental no debe ser mayor que la longitud de la aleta para que la media sea compatible con el dato experimental.

Paso 8: *Calculamos las medias extremas.* Movemos la campana recortada hacia la izquierda y hacia la derecha hasta donde ya no se pueda más, es decir, hasta que moverla un poquito más implique que el dato se convierta en anormal pues se sale de la campana.

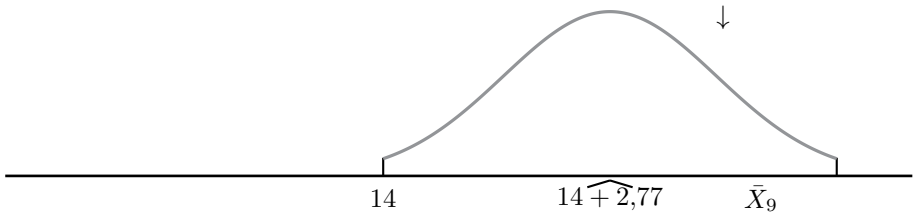


Figura 6.10. Campana recortada que contiene el dato experimental $\bar{X}_9 = 14$ en su extremo inferior. La campana ya no se puede correr más hacia arriba, pues el dato experimental dejaría de ser normal.

Claramente, la campana puede moverse dentro de un intervalo y lo que necesitamos es hallar los límites de dicho intervalo. Hallémoslo: si movemos la campana hacia arriba del dato experimental, $\bar{X}_9 = 14$, la media de la campana puede moverse máximo 2.77. Es decir, la campana extrema hacia arriba tiene su media en 14+2.77. Y si la movemos hacia abajo, la media puede moverse otro tanto.

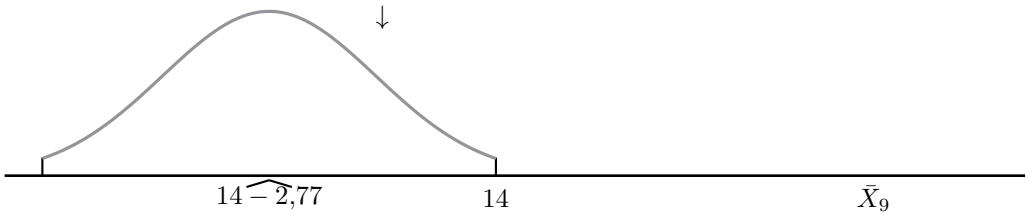


Figura 6.11. Campana recortada extrema hacia abajo que contiene el dato experimental como común y corriente.

Por tanto, el intervalo buscado empieza en el dato experimental menos 2.77 y termina en el dato experimental más 2.77. Eso da el intervalo

(14 - 2.77, 14+2.77) = (11.23, 16.77)

Paso 9: Cerramos el intervalo para que el dato experimental pueda estar en los bordes de lo que se considera normal: [11.23, 16.77].

Paso 10: *Respuesta.* Si hay un dato experimental, $\bar{X}_9 = 14$, y si $\sigma_X = 3{,}2$, con una confianza del 99 % la media está en el intervalo [11.23, 16.77].

Solución algebraica:

Paso 1: Datos del problema: hallar el IC del 99 % para la media de
$X \sim N(\mu = ?, \sigma = 3{,}2)$, si $\bar{X}_9 = 14$.

Paso 2: Averiguamos la z crítica. Como la confianza es 0,99, la significancia es $\alpha = 1 - 0{,}99 = 0{,}01$. Como se trata de un IC, la significancia se toma con dos colas. Por tanto, la dividimos entre dos, da 0,005 y miramos en la tabla para obtener $-2{,}6$. Como es con dos colas, hay dos zetas críticas: $\pm 2{,}6$.

Paso 3: Como el dato que nos dan es una media muestral, usamos el TLC.

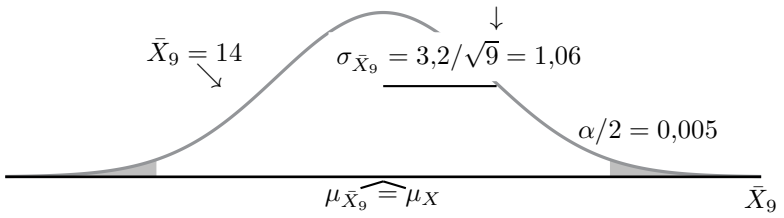


Figura 6.12. $\bar{X}_n$ tiene como media la misma media de X, $\mu_{\bar{X}_9} = \mu_X$, y como desviación, la desviación de X dividida entre la raiz cuadrada del número de datos: $\sigma_{\bar{X}_9} = 3{,}2/\sqrt{9} = 1{,}06$. Para calcular el IC asumimos que la media muestral observada es común y corriente.

El TLC nos da la distribución de la media muestral: $\bar{X}_n$ tiene una distribución normal y como media a la misma media de X, es decir $\mu_{\bar{X}_n} = \mu_X$. Y también tenemos que $\sigma_{\bar{X}_n} = \frac{\sigma_X}{\sqrt{n}}$. En este caso, $\sigma_X = 3{,}2$ por lo que $\sigma_{\bar{X}_n} = 3{,}2/\sqrt{9} = 1{,}06$.

Paso 4: Ahora calcamos el problema desde $\bar{X}_n$ sobre la Z.

Para ello, usamos la transformada-z cuya expresión general es

$z = \frac{x-\mu}{\sigma}$

y que en el caso de una media muestral $\bar{X}_n$ tiene la siguiente expresión:

$z(\bar{X}_n) = \frac{\bar{X}_n - \mu_{\bar{X}_n}}{\sigma_{\bar{X}_n}} = \frac{\bar{X}_9 - \mu_X}{\frac{\sigma_X}{\sqrt{9}}} = \frac{14-\mu_X}{\frac{3{,}2}{3}} = \frac{14-\mu_X}{1{,}06}$

Ahora bien, al intervalo de confianza del 99 % para la media de X es un intervalo que cumple con

$p(-2{,}6 < z(\bar{X}_n) < 2{,}6) > 0{,}99$

lo cual significa que, para la significancia dada, una media poblacional μ es compatible con la media de la muestra $\bar{X}_n$ si y sólo si $\bar{X}_n$ se clasifica como común y corriente en el mundo de las medias muestrales con media $\mu_{\bar{X}_9} = \mu_X$.

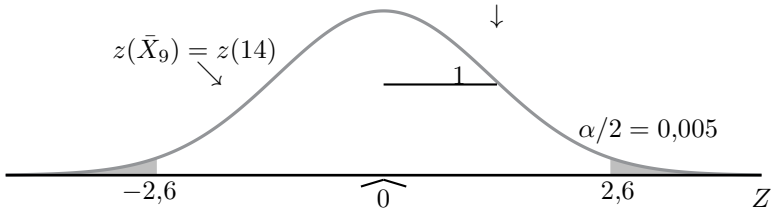


Figura 6.13. Como la media muestral observada es común y corriente en el mundo de las medias muestrales, su transformada-z es común y corriente en el mundo de la Z.

Paso 5: Elaboramos el intervalo. Reemplazando el valor para la transformada-z de la media tenemos:

$p(-2{,}6 < \frac{14-\mu_X}{1{,}06} < 2{,}6) > 0{,}99$.

Ahora multiplicamos en todo lado por 1,06:

$p((-2{,}6)(1{,}06) < 14 - \mu_X < (2{,}6)(1{,}06)) > 0{,}99$.

Restamos 14 de todas partes:

$p(-14 + (-2{,}6)(1{,}06) < -\mu_X < -14 + (2{,}6)(1{,}06)) > 0{,}99$.

Multiplicamos por menos uno, el cual reversa las desigualdades (por ejemplo: $1 < 2$ pero $-1 > -2$ pues -1 queda a la derecha de -2):

$p(14 + (2{,}6)(1{,}06) > \mu_X > 14 - (2{,}6)(1{,}06)) > 0{,}99$.

$p(14 + 2{,}77 > \mu_X > 14 - 2{,}77) > 0{,}99$.

$p(16{,}77 > \mu_X > 11{,}23) > 0{,}99$.

Orientemos las desigualdades a la manera usual, que apunten de menor a mayor

$p(11{,}23 < \mu_X < 16{,}77) > 0{,}99$.

Paso 6: Interpretación: la probabalidad condicional a los datos del problema de que la media quede entre 11.23 y 16.77 es mayor que 0.99. Lo cual quiere decir que si aseguramos que la media queda en ese intervalo, acertaremos por lo menos el 99 % de las veces y nos equivocaremos alrededor del 1 % si todo eso se hace muchas veces.

168 ◇ ***Teorema:*** *Sea $X \sim N(\mu, \sigma)$ y sea α la significancia. El intervalo de confianza del $(1-\alpha)\,\%$ para la media conociendo n datos con media $\bar{X}_n$ es*

$$(\bar{X}_n - z(\sigma/\sqrt{n}), \bar{X}_n + z(\sigma/\sqrt{n}))$$

donde z es el z-crítico en valor absoluto asociado a $\alpha/2$. Esa fórmula también se escribe rigurosamente como:

$$(\bar{X}_n + z_{\alpha/2}(\sigma/\sqrt{n}), \bar{X}_n + z_{(1-\alpha/2)}(\sigma/\sqrt{n}))$$

o abusando de la notación:

$$(\bar{X}_n - z_{\alpha/2}(\sigma/\sqrt{n}), \bar{X}_n + z_{\alpha/2}(\sigma/\sqrt{n}))$$

Demostración. Con todo lo que hemos aprendido en el cálculo del IC, podemos dar una demostración muy corta:

Si $X \sim N(\mu, \sigma)$, entonces, por el TLC, $\bar{X}_n \sim N(\mu_x, \sigma_x/\sqrt{n})$. A la significancia dada le corresponde un z crítico con dos colas que se mira en la tabla. Por lo tanto, los datos normales en $\bar{X}_n$ son aquellos que quedan dentro de la aleta de normalidad, cuya longitud es $z\sigma_x/\sqrt{n}$. Dicho de otra manera, una media es compatible con el dato experimental $\bar{X}_n$ cuando la distancia entre el dato y la media no es mayor que $z\sigma_x/\sqrt{n}$. Como sobre una línea una distancia se mide con el valor absoluto, tenemos:

$|\mu - \bar{X}_n| < z\sigma_x/\sqrt{n}$

Ahora recordamos que $|a| < b$ sí y sólo si a está entre $-b$ y b: $-b < a < b$. Aplicando esto, tenemos:

$-z\sigma_x/\sqrt{n} < \mu - \bar{X}_n < z\sigma_x/\sqrt{n}$

Pasando el término $-\bar{X}_n$ a sumar a todo lado:

$\bar{X}_n - z\sigma_x/\sqrt{n} < \mu < \bar{X}_n + z\sigma_x/\sqrt{n}$

lo cual dice que el IC para la media es:

$(\bar{X}_n - z\sigma_x/\sqrt{n}, \bar{X}_n + z\sigma_x/\sqrt{n})$

el cual puede cerrarse, según el gusto personal.

Gracias a este teorema existen calculadoras y paquetes que dan el IC de la media sin necesidad de entender de dónde salen las cosas. Usemos los datos del ejemplo anterior para ver cómo se aplica:

Si $X \sim N(\mu, \sigma)$, entonces, por el TLC, la v.a. media muestral con n elementos $\bar{X}_n \sim N(\mu_x, \sigma_x/\sqrt{n})$. El intervalo de confianza del 99 % de la media contiene todas las medias μ cuya distancia al dato experimental $\bar{X}_n|$ es menor que $z\sigma_x/\sqrt{n}$

$|\mu - \bar{X}_n| < z\sigma_x/\sqrt{n}$

donde
$\bar{X}_n = 14$
El z asociado a $\alpha/2 = 0{,}005$ es -2.6 y su valor absoluto es 2.6
$\sigma = 3{,}2$
$n = 9$;
reemplazando tenemos:

$$|\mu - 14| < (2{,}6)(3{,}2/\sqrt{9})$$

$$-(2{,}6)(3{,}2/\sqrt{9}) < \mu - 14 < (2{,}6)(3{,}2/\sqrt{9})$$

$$14 - 2{,}6(3{,}2/\sqrt{9}) < \mu < 14 + 2{,}6(3{,}2/\sqrt{9}))$$

$$14 - 2{,}6(3{,}2/3) < \mu < 14 + 2{,}6(3{,}2/3)$$

$$14 - 2{,}6(1{,}06) < \mu < 14 + 2{,}6(1{,}06)$$

$$14 - 2{,}77 < \mu < 14 + 2{,}77$$

$$11{,}23 < \mu < 16{,}77$$

Lo que nos dice que el IC del 99 % de confianza de la media es (11.23,16.77).

169 ***Para recordar siempre: no se puede estimar una media sin antes haber estimado la varianza o la desviación.***

La razón es que la desviación puede interpretarse como una incertidumbre.

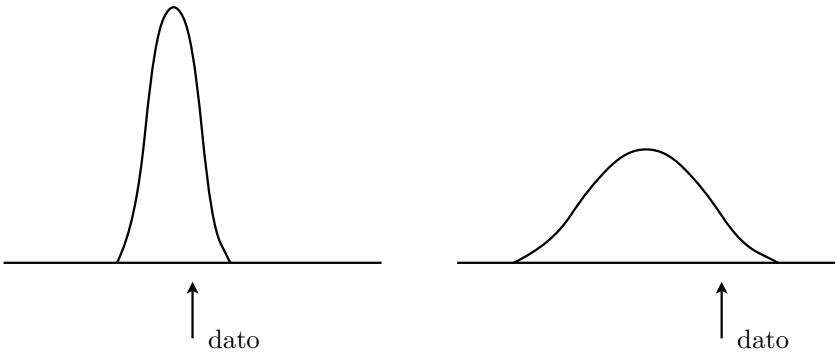


Figura 6.14. Cuando la campana es muy cerrada, cualquier dato estará bastante cerca de la media, pero cuando la campana es muy abierta, un dato cualquiera podría estar muy lejos de la media y se requerirían muchos datos para tener una estimación confiable de la media.

Si la incertidumbre descrita por la desviación es pequeña comparada con el orden de magnitud de la media, unos pocos datos son suficientes para estimar la media. Pero si la incertidumbre es grande, se requieren muchos datos para poder neutralizar dicha incertidumbre. Por eso, en la fórmula del intervalo de confianza aparece tanto el número de datos como la desviación.

6.3. Conclusión

Hemos argumentado la idea de que la ciencia es una artesanía, cuya belleza, poder y responsabilidad descansan sobre el ser humano que la ha creado siguiendo su instinto biológico. También hemos afrontado el difícil problema de extraer resultados experimentales que se puedan repetir con elevado índice de probabilidad sabiendo que no podemos controlar todo lo que influye sobre nuestro experimento. Hemos concluido que eso es perfectamente posible y que es suficiente reunir y reunir datos en el experimento para que el ruido termine auto-aniquilándose, lo cual ha quedado ilustrado con el teorema del límite central que habla sobre estimación de la media: entre más datos se promedie, el ruido tiende a auto-aniquilarse y el resultado se parecerá cada vez a la media de la población.

Enfaticemos la importancia de los intervalos de confianza en la toma de decisiones: es muy probable que los datos de una observación al azar o de un experimento nunca se repitan, por eso, son inútiles en la toma de decisiones. Pero adquirimos un tesoro si usamos los datos de la muestra para estimar los parámetros poblacionales, media o varianza, que nos dan el prospecto a futuro, lo cual se hace en los intervalos de confianza. Por ejemplo, si hoy me demoré almorzando 20 minutos (valor muestral), no es muy inteligente apartar 20 minutos cada día para el almuerzo. Es más prudente tener en cuenta que la gran mayoría de las veces, con un 95 % de confianza, el tiempo promedio de almuerzo en día laboral (valor poblacional) está entre 25 y 30 minutos.

Capítulo 7

Taller y parcial II

170 ***Objetivo y advertencia:*** *Hacer ejercicios de sobra y correctamente para quedar bien preparados. Por eso, las personas que vienen con experiencias muy negativas en matemáticas tienen que esforzarse en aprender a juzgar las respuestas con su sentido común. En particular sirve tener presente que la probabilidad es una idealización de frecuencia relativa y que por tanto está entre cero y uno y que la suma de todas las probabilidades da uno exacto. Por tanto, es prohibido cometer los siguientes errores garrafales: que se diga que una probabilidad es -0.7, o que sea 1.8, y que la suma de las probabilidades totales sea -0.9, o 0.77 o 7.2.*

7.1. Taller II

1 . a) Use el criterio de la doble mitad para proponer una normal, una campana, que represente la altura de sus compañeros de clase, la cual se puede tomar como una muestra al azar de los alumnos de la Universidad. b) Estime gráficamente la proporción de estudiantes de la U que miden entre 1.60m y 2 metros (ejemplos 70-74, pág. 59. Respuestas pág 140.)

2 . Use la tabla de la Ψ de la pág. 62 para calcular los siguientes valores de la función de distribución de la normal estándar y haga el dibujo correspondiente (ejemplos 75 - 79, pág. 61):

a) $\Psi(-2)$
b) Ψ(2,3)
c) $\Psi(-3)$
d) $\Psi(-5)$
e) $\Psi(-20)$
f) $\Psi(20)$

3 . Ilustrando con una gráfica hallar c tal que (ejemplo 80, pág. 63)

a) $\Psi(c) = 0,34.$

b) $\Psi(c) = 0,77.$

c) $\Psi(c) = 0,98.$

d) $\Psi(c) = 2,35.$

e) $\Psi(c) = -0,35.$

4 . Use la tabla de la Ψ para hallar las siguientes probabilidades (Ejs 82 y 83 pág. 64).

a) $p(-1{,}3 < z < 2{,}2)$
b) $p(-1{,}3 < z < -1{,}1)$
c) $p(-2{,}8 < z < 3{,}1)$
d) $p(3{,}5 < z < \infty)$

5 . a) Halle la media y la desviación de las notas siguientes: 3, 3, 3, 5, 4, 1, 2, 3, 4, 2, 3. b) Use la transformada z para hallar el valor z de cada una de las notas. Verbalice cada uno de dichos valores. Use la tabla de la Ψ para hallar una aproximación al percentil de cada dato. (Ej 88, pág. 67). Recuerde: la transformada-z de un evento mide la cantidad de desviaciones que el evento queda arriba o abajo de la media. Usamos la transformada-z para pasar de una campana cualquiera con media y desviación arbitrarias a una campana estandar con media cero y desviación 1. Así podemos usar la tabla de la Psi, del tridente, válida únicamente para la normal estándar, para hallar una aproximación al percentil.

6 . Halle los valores de x asociados a las z dadas según la normal dada (ej 90, pág. 67)

a) $z = 0{,}9$, $X \sim N(12, 6)$. *(Si un evento está arriba de la media 0.9 desviaciones, ¿cuál es dicho evento?)*

b) $z = 0{,}2$, $X \sim N$*(0.8, 7).*

c) $z = -2{,}3$, $X \sim N$*(0.001,0.00001) .*

d) $z = 1{,}6$, $X \sim N(1000, 120)$.

e) $z = 2{,}7$, $X \sim N(-200, 35)$.

7 . Calcule las probabilidades solicitadas, para una v.a. (variable aleatoria) X normalmente distribuida con la media y desviación estándar dadas e invente una telenovela apropiada para cada caso (Ejs 92, 93, 96 y 97, pág. 68). Ayuda: calque el problema sobre la Z usando la transformada-z, luego use el teorema del área grande menos el área chica. Escoja dos incisos y resuélvalos.

a) $\mu = 7$, $\sigma = 3$, $P(5 < X < 10) =?$
b) $\mu = 11$, $\sigma = 2$, $P(4 < X < 12) =?$
c) $\mu = 1120$, $\sigma = 85$, $P(1140 < X < 1235) =?$
d) $\mu = 13$, $\sigma = 0{,}45$, $P(10 < X < 50) =?$
e) $\mu = 320$, $\sigma = 40$, $P(-230 < X < 80) =?$
f) $\mu = 0{,}28$, $\sigma = 0{,}01$, $P(0{,}27 < X < 0{,}29) =?$

8 . Una compañía que produce camisas sabe que la espalda de sus clientes sigue una distribución normal con media 38cm y desviación 4cm. Se ha recibido un contrato para fabricar 5500 camisas. Se ha decidido que habrá 4 tallas definidas por su X correspondiente como sigue: $S : X < 33$; $M : 33 < X < 42$; $L : 42 < X < 46$; $XL : X > 46$. Decidir la cantidad de camisas de cada talla, y la marca de clase. (Ej 99, pág. 71)

9 . Una compañía que produce camisas sabe que la espalda de sus clientes sigue una distribución normal con media 40cm y desviación 5cm. Se ha recibido un contrato para fabricar 4500 camisas. Se ha decidido que habrá 4 tallas definidas por su transformada-z correspondiente como sigue: $S : z < -1{,}2$; $M : -1{,}2 < z < 1{,}2$; $L : 1{,}2 < z < 2{,}2$; $XL : z > 2{,}2$. Decidir la cantidad de camisas de cada talla, el intervalo para cada talla y la marca de clase.

Algunos errores que alguien cometió en estos dos problemas sobre camisas y que pueden evitarse son los siguientes:

1. *No verificó que las probabilidades no pueden ser negativas ni mayores que uno y que la suma de todas las probabilidades no puede ser distinta de uno.*
2. *No verificó que el número de camisas debió ser lo que se pedía que era 4500.*
3. *Le diéron 4366 camisas por todas cuando debería dar 4500 pero no se dio cuenta de que la discrepancia tan grande implica un error de cálculo. Una discrepancia mayor de 5 en 1000 ya delata un error.*
4. *El número de camisas de talla S y L le dieron mayor que el de talla M y no reaccionó: en general, los de talla M quedan en el centro de la distribución que por ser de tendencia central contiene a casi todo el mundo y por tanto deben ser los más numerosos.*
5. *No se dio cuenta de que el problema de las camisas tiene dos variantes que se resuelven por métodos distintos. El uno empieza con datos sobre X, el ancho de la espalda, el otro empieza con datos sobre la Z, datos estandarizados. En el primer caso hay que estandarizar, en el segundo no.*

10 . Nuestra tabla de la Ψ tiene 4 columnas, una para z negativo y otra para su Ψ, y otras dos para z positivo y su Ψ. Observe que en cada fila, la suma de los dos valores de Ψ es uno. ¿Es esto una coincidencia o un teorema? Demuestre su respuesta. En un primer esfuerzo, guíese por dibujos que apunten a la simetría de la campana. O también, use las propiedades de la integral para demostrar su resultado.

11 . Diferentes autores pueden producir tablas diferentes, y por consiguiente uno debe estar listo a entenderse con ellos. Ayudándose con gráficas adecuadas, diga ¿cómo puede uno calcular $\Psi(z)$ si una tabla da la información sobre:

a) $\int_0^z \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds$ para $z > o$
b) $\int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds$ para $z < o$
c) $\int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds$ para $z > o$
d) $\int_{-z}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds$ para $z > o$

12 . Demuestre o refute gráficamente que para cualquier significancia si un evento es anormal con dos colas, con mayor veras lo será con una cola.

13 . Consideremos la tesis de que la historia es lo que se recuerda instintivamente y se graba para que el futuro lo conozca. Si la fuerza de los recuerdos está dominada por el instinto, entonces es muy probable que los recuerdos consten en su mayoría de elementos críticos que causaron una descarga emocional gigante, hacia las afueras de la distribución. Por ejemplo, uno recuerda los grandes terremotos, pues aunque la tierra vive meciéndose a toda hora, los grandes terremotos son en verdad terroríficos por su potencia y destrucción.

a) Sobre un ejemplo de su vida real, asocie una campana a la tesis expuesta, en la cual el evento recordado corresponda a eventos lejos de la media en una cierta distribución.

b) Consideremos la siguiente objeción: No es tan obvio que todo sea por instinto. Por ejemplo: todos sabemos el día de nuestro nacimiento. Con todo, hay que reconocer que eso no ocurre en toda las culturas, digamos, algunos nativos de la selva no saben ni siquiera su edad. Pero precisamente ese detalle hace pensar que hay elementos culturales que definen la historia. Queda claro que la única forma de probar eso es dando ejemplos de casos que se recuerdan en unas culturas pero no en otras. Para calentar motores, elabore esta temática entre los miembros de su familia, en la cual se evidencie el aspecto personalizado de los recuerdos sobre los mismos hechos, circunstancias o períodos de tiempo. ¿Podrán todos los elementos culturales codificarse en forma de descriptores cuantitativos, a los cuales uno pueda asociarle una campana?

14 . Colas y significancia (Ejs 103 - 105 pág. 76). Tengamos presente: Si el interés está en lo máximo, se trabaja con la cola superior. Si en lo mínimo, con la cola inferior. Si en lo extremo, con ambas colas. Y también: la significancia es la proporción de los datos considerados como extremos.

a) Un profesor de un colegio está muy orgulloso de los cincos de su curso y no ve nada más. Si X es la nota del final, diga con cuáles colas trabaja. b) Como el profesor supo que los otros cursos no llegan ni a 4.5, también se puso orgulloso de los que sacaron más que eso: ¿Cambió la cola? ¿aumentó o disminuyó la significancia? c) La Rectora ha asignado un presupuesto para subsidiar a los estudiantes que viven muy lejos. Sea Y la distancia de la casa al colegio: ¿con cuáles colas ha de trabajar? Si hubo un recorte presupuestal, ¿aumentará o disminuirá la significancia? d) La Rectora contrató un tutor para preparar a sus alumnos para el Examen de Estado. Sea Z la nota en exámenes de preparación. Si el colegio es principiante, con cuáles colas ha de trabajar? A medida que el colegio madura, ¿qué le pasa a la significancia y al número de colas? e) La Oficina de Control de Colegios toma una muestra al azar y niega la acreditación a los colegios cuya nota promedio en un examen no sobrepase un valor mínimo dado. ¿A cuál cola debe prestarse atención prioritaria si W es la nota en una evaluación interna? ¿Qué le pasa a la significancia si la nota para pasar aumenta significativamente? f) A la Rectora le preocupa mucho su peso pues ha recibido reclamos de dar mal ejemplo. El problema es que los chocolates le encantan y después hace dieta por lo que su peso P vive subiendo y bajando. Ella se rige por el peso de las señoras de su estatura y edad que se distribuye como una normal con media 70 Kgs y desviación 3 Kgs. Cuando ella está sin novio y sobrepasa los 80 Kgs se prende una alarma que le dice: estoy requetegorda. ¿Con cuántas colas trabaja para declarar dieta? ¿Cuál es el valor crítico y cuál su significancia? Si tiene un novio gordito y cheverongo y que la hace sentir amada a pesar de lo gorda, ¿Qué le pasa al valor crítico? ¿Qué le pasa a la significancia?, ¿y a las colas?, ¿Y si se quiere conseguir un novio flaquito?

15 . Bosqueje una Z donde se señalen los eventos interesantes para la cola inferior cuando la significancia α sea 0.01, 0.05, 0.1, 0.2. Señale el z-crítico. Un evento es interesante cuando es más extremo que el valor crítico determinado por la significancia y la cola en cuestión. Haga lo mismo para la cola superior y para las dos colas (Ejs 110 pág. 78).

16 . Señale los eventos interesantes sobre las normales dadas (Ejs 111 y 112 pág. 80):

a) $X \sim N(15, 5)$, $\alpha = 0{,}01$, cola inferior.
b) $X \sim N(4, 1)$, $\alpha = 0{,}05$, cola superior.
c) $X \sim N(17, 8)$, $\alpha = 0{,}04$, dos colas.

17 . Señale los eventos interesantes sobre las binomiales dadas. Un error común que se comete al calcular los eventos interesantes de las binomiales es el siguiente: si la significancia es 0.01 con la cola superior, entonces uno dice que todo lo extremo por arriba que tenga probabilidad menor a 0.01 es interesante. Error. La solución correcta es:

1. El dato más extremo por arriba es interesante si su probabilidad es menor que 0.01.

2. El siguiente dato hacia abajo es interesante si su probabilidad más la del primero es menor que 0.01.

3. El tercer dato de arriba hacia abajo es interesante si la suma de su probabilidad más la acumulada hacia arriba es menor que 0.01.

Cuando se trata de una cola, comience por los valores más extremos y acumule probabilidad hasta que ya no pueda más pues se sobrepasa la significancia. Con dos colas: haga lo mismo pero con la mitad de la significancia. ¿Por qué se procede así? Porque la significancia es la probabilidad total de ser extremo. Eso implica un área o integral en el caso de las campanas o una suma en el caso de las binomiales. Los ejercicios son:

a) $X \sim Bi(4, 1/2)$, $\alpha = 0{,}07$, cola superior (Ej 114 pág. 82, fácil pues $p = 1/2$).
b) $X \sim Bi(7, 1/2)$, $\alpha = 0{,}01$, dos colas (Ej 114 pág. 82, fácil pues $p = 1/2$).
c) $X \sim Bi(4, 1/45)$, $\alpha = 0{,}01$, cola inferior, cola superior, dos colas (Ej 115 pág. 82 con $p \neq 1/2$).
d) $X \sim Bi(10, 1/5)$, $\alpha = 0{,}01$, dos colas (Ej 119 pág. 84, como n es grande, se puede aproximar por la normal).

18 . Dos niñas juegan con una moneda a la que saque más caras. Una de ellas sacó 18 caras de 20 monedas lanzadas. La otra niña se puso muy brava. Explique su reacción sicológica como algo compatible con la teoría de probabilidades, lo que nos hace pensar que la teoría de probabilidades tiene su base en el instinto. Aproxime la binomial por la normal. Se agradece un despliegue gráfico adecuado (Ej 119 pág. 84). Ayuda: 1) Especifique la binomial del juego, la n y la p. 2) Calcule la media y la desviación y dé la mejor aproximación normal a la binomial dada. 3) Tome la significancia por defecto, determine el número y posición de las colas y calcule los valores críticos. 4) Pruebe que 18 es interesante.

19 . Defina los eventos interesantes si se lanzan dos dados y la variable X cuenta el número de puntos que quedan en las caras que quedan hacia arriba.

20 . Halle la significancia del evento x con las condiciones dadas para la variable aleatoria X de distribución normal e invente una telenovela apropiada para cada caso (Ej 122 pág. 85):

a) $x = 54$ cuando $X \sim N(\mu = 80, \sigma = 12)$, cola inferior.
b) $x = 42$ cuando $X \sim N(\mu = 34, \sigma = 5)$, cola superior.
c) $x = -0{,}2$ cuando $X \sim N(\mu = 3{,}2, \sigma = 1{,}1)$, dos colas.
d) $x = -3$ cuando $X \sim N(\mu = 14, \sigma = 8)$, dos colas.
e) $x = 23$ cuando $X \sim N(\mu = -54, \sigma = 20)$, cola superior.
f) $x = -0{,}5$ cuando $X \sim N(\mu = -1, \sigma = 0{,}1)$, dos colas.

21 . Calcule el coeficiente de variación de las diversas binomiales y deduzca que se cierran cada vez más hacia la media, lo cual significa que los eventos probables se hacen cada vez más semejantes entre sí y además iguales a la media. Tal parecido aumenta con el n de la binomial. En términos de ruido y azar, todo se relee así: cuando uno lanza una moneda, uno tiene incertidumbre total y uno está a merced del azar: el ruido lo cubre todo. Pero cuando uno repite y repite, aparecen tendencias claras, definidas y determinísticas. Concretamente: Experimento básico. Lanzamos muchas monedas o una misma moneda muchas veces y contamos el número de caras. Experimento compuesto: Repetimos el experimento básico muchas veces y sacamos la media del número de caras. Observamos que la media y lo que salió en cada experimento básico son relativamente muy parecidos: la incertidumbre se esfumó, se autoaniquiló y la media resulta ser casi inconmovible.

22 . Multas y muertes.

La Secretaría de Tránsito impone cuotas obligatorias a todos los propietarios de vehículos y además impone multas a los violadores de sus normas.

a) Supongamos que la mayoría de ciudadanos procura ser muy puntual en sus pagos. Bosqueje una distribución apropiada para el caso de la frecuencia relativa de personas que deben una cantidad Q y explique por qué no es de tendencia central a pesar de que cada persona funciona libremente, es decir, decide de forma independiente. Defina qué eventos son los interesantes y explique por qué.

b) Dada la gran efectividad de los movimientos cívicos, en un pueblo deciden no pagar para después solicitar amnistía. Bosqueje la distribución de la deuda correspondiente. ¿Cuáles serían los eventos interesantes?

c) No es raro que muchos accidentes sucedan con participación de uno de los grandes deudores que viven poniendo en peligro la vida de todo el mundo y que nunca pagan. ¿Considera Usted que debe cancelársele la licencia a todo gran deudor en aras de la prevención de accidentes?

23 . (Prerequisitos: Ej. 123, pág. 85 hasta ej. 127, pág. 88.) El nivel de azúcar de la sangre tiene una distribución normal y gira en torno a su media, a veces sube y a veces baja, por lo que su desviación no es nula. Para que una sustancia sea clasificada como alimento, esta no debe hacer variar el azúcar de la sangre más allá de lo normal. Claramente aquí tenemos un caso de dos colas, pues si una sustancia baja mucho el azúcar es peligrosa y lo mismo pasa si la sube mucho.

Pero si uno ha detectado que una sustancia sube el azúcar desmesuradamente, entonces uno ha descubierto un remedio que sirve para subir el azúcar, bueno para los hipoglicémicos. Para probar que uno en realidad tiene uno de dichos remedios, uno hace una prueba de una cola, la superior. Similarmente, si uno ha descubierto una sustancia que baja el azúcar, uno ha descubierto un hipoglucemiante, y si además sirve en pequeñas dosis y no produce efectos colaterales, uno ha hecho un gran descubrimiento y corre el riesgo de volverse muy rico. La prueba estadística correspondiente es con una cola, la inferior. Miremos esto dentro de varios contextos plausibles.

Aclaraciones. Primera: PARA SABER EL NÚMERO DE COLAS EN UN PROBLEMA, uno lee la pregunta que hay que resolver y toma cuenta del objetivo: si simplemente se trata de establecer si algo es normal o no, es una prueba o test con dos colas. Si se trata de mejorar una marca, es con la cola superior. Si se trata de disminuir un problema, es con la cola inferior. Segunda: SOBRE EL USO DE LOS SIGNOS $>$ y $<$ para comparar dos números. Ambos signos se interpretan con la misma ley: el dato grande va sobre la apertura, que es lo grande, y el pequeño sobre el vértice, que es lo pequeño. Así, para denotar que 4 es mayor que 3 se puede escribir $4 > 3$ (4 es mayor que 3) ó bien $3 < 4$ (tres es menor que 4). Tercera: LOS PARES DE JEROGLÍFICOS CON la HIPÓTESIS NULA H_o y LA ALTERNA H_a se interpretan así:

- $H_o : \mu = 3$ y $H_a : \mu \neq 3$ se lee: lo normal gira alrededor de 3 y vamos a investigar si vale la pena pensar que hay un factor sistemático que haga que el mundo gire alrededor de una media diferente. La H_o se rechaza con dos colas pues hay dos maneras de ser diferente, por ser menor o por ser mayor.
- $H_o : \mu = 3$ y $H_a : \mu > 3$ se lee: lo normal gira alrededor de 3 y vamos a investigar si vale la pena pensar que hay un factor sistemático que haga que el mundo gire alrededor de una media mayor que 3. La H_o se rechaza con cola superior.
- $H_o : \mu = 3$ y $H_a : \mu < 3$ se lee: lo normal gira alrededor de 3 y vamos a investigar si vale la pena pensar que hay un factor sistemático que haga que el mundo gire alrededor de una media menor que 3. La H_o se rechaza con cola inferior.

a) Consideremos que el azúcar de la sangre de una persona saludable tiene media 100 y desviación 10. Cuando una persona consume el producto B en la cantidad que a ella le gusta, el azúcar se le baja a 78. Decida si el producto B sirve de alimento o si fuese mejor considerarlo como de remedio para bajar el azúcar.

b) Consideremos que el azúcar de la sangre de una persona saludable tiene media 100 y desviación 10. Cuando una persona consume el producto C en la cantidad que a ella le gusta, el azúcar se sube a 115. Queremos saber si el producto C sirve de alimento o si sería mejor considerarlo como de remedio para subir el azúcar.

c) Consideremos que el azúcar de la sangre de una persona saludable tiene media 100 y desviación 10. Cuando una persona consume el producto A en la cantidad que a ella le gusta, el azúcar se le baja a 93. Decida si el producto A sirve de alimento o si fuese mejor considerarlo como de remedio.

24 . Una v.a. X se distribuye normalmente con media desconocida y $\sigma = 1{,}9$. Estudie la H_o dada a la luz del resultado experimental dado. Formule una telenovela apropiada a cada caso (Ej 123 - 127 pág. 85 y siguientes):

a) $H_o : \mu = 3$, $x = 10$, $\alpha = 0{,}02$, cola superior.

b) $H_o : \mu = 3$, $x = 10$, $\alpha = 0{,}02$, dos colas.

c) $H_o : \mu = -1$, $x = 4$, $\alpha = 0{,}01$, cola superior.

d) $H_o : \mu = -1$, $x = 4$, $\alpha = 0{,}01$, dos colas .

e) $H_o : \mu = 4$, $x = 1$, $\alpha = 0{,}05$, dos colas.

f) $H_o : \mu = -8$, $x = 3$, $\alpha = 0{,}022$, cola superior.

g) $H_o : \mu = -8$, $x = 3$, $\alpha = 0{,}022$, dos colas.

h) $H_o : \mu = 5$, $x = 2$, $\alpha = 0{,}00001$, cola inferior.

i) $H_o : \mu = 5$, $x = 2$, $\alpha = 0{,}00001$, dos colas .

25 . En una cierta comunidad grande hay 6 mujeres por cada 5 hombres. En una fiesta abierta a todos hay 50 personas de las cuales 35 son mujeres y 15 son hombres. ¿Podemos considerar que las proporciones dadas indican un desbalance de mujeres en lo que se refiere a la asistencia a esta fiesta? (Ej 128, pág. 90).

26 . En una clase de estadística hay 20 varones y la proporción histórica entre ellos de los que hacen tareas todos los días ha sido del 30 %. Pero un día una alumna muy hermosa dijo que los hombres son arrimados. Quince días después se verificó que los varones que hicieron tarea fueron 8. ¿Sirvió el tremendo regaño de la alumna o se puede explicar el cambio por azar? (Ej 128, pág. 90).

27 . Según la experiencia previa, de 10000 estudiantes que toman materias de matemáticas hay 4000 que consideran que les cuesta mucho, pero mucho trabajo. De una clase de 35 alumnos, hay 12 que sufren y sufren y no paran de sufrir. El profesor se jacta de su metodología alegando que en su clase se sufre menos que en el resto del mundo. ¿Es eso cierto?

28 . Si $X \sim Bi(n, p)$ se aproxima por una normal, ¿cómo se distribuye $\phi = X/n$? Explique claramente todo en términos de monedas.

29 . Invente:

a) Explique por qué cayeron 28 caras de 40 lanzamientos. Es decir, halle por tanteo un valor para $p(cara)$ tal que el resultado experimental sea común y corriente en ese caso si $\alpha = 0{,}05$ y dos colas (ej 133, Pág 93).

b) Explique por qué cayeron 38 caras de 40 lanzamientos si $\alpha = 0{,}01$, dos colas.

30 . Para explicar en una cierta situación por qué cayeron 18 caras de 25 lanzamientos, se tienen dos hipótesis con 2 colas. Primera que $p = 0{,}6$. Segunda que $p = 0{,}8$. ¿Cuál de las dos hipótesis goza de mayor credibilidad? (Ej 134, pág. 93)

31 . Halle constructivamente el intervalo de confianza para μ (el que contiene todas las medias poblacionales que son compatibles con los datos experimentales y siempre con 2 colas) conociendo un dato experimental x sacado al azar de una distribución normal y la desviación σ si (Ej 136, pág. 95)

a) $\alpha = 0.02$, $x = 12$, $\sigma = 2$
b) $\alpha = 0.04$, $x = 9$, $\sigma = 1{,}7$
c) $\alpha = 0.07$, $x = 41$, $\sigma = 13$
d) $\alpha = 0.06$, $x = -17$, $\sigma = 1{,}2$
e) $\alpha = 0.18$, $x = 0{,}001$, $\sigma = 0{,}0001$

32 . Halle constructivamente el intervalo de confianza para p (siempre con 2 colas) con el α dado conociendo que un experimento de n ensayos salieron c caras. Muestre todos los pasos. (Ej 138, Pág 97):

a) $\alpha = 0.02$, $n = 35$, $c = 12$
b) $\alpha = 0.03$, $n = 119$, $c = 58$
c) $\alpha = 0.04$, $n = 20$, $c = 17$
d) $\alpha = 0.05$, $n = 83$, $c = 49$
e) $\alpha = 0.18$, $n = 240$, $c = 135$

33 . Una familia independiente gana al mes $X \sim N(10, 2)$. Halle e ilustre gráficamente la forma como se distribuye el sueldo promedio de n meses tomados al azar cuando n vale a) 1, b) 4, c) 9, d) 16, e) 25, f) 36, g) 49. (Ej 154, pág. 105 a ej.156).

34 . Sea $\bar{X}_n$ la v.a. media muestral de n eventos tomados de una distribución normal (tomar una muestra de n eventos al azar de una población con distribución normal, sacarle la media y repetir y repetir). Halle la media y la desviación de $\bar{X}_n$ en cada uno de los siguientes casos. Dibuje las campanas tanto de X como de $\bar{X}_n$ (Ej 154, pág. 105 a ej.156):

a) $n = 8$, $X \sim N(\mu = 5, \sigma = 2)$
b) $n = 17$, $X \sim N(\mu = 3, \sigma = 1)$
c) $n = 4$, $X \sim N(\mu = 2, \sigma = 5)$
d) $n = 23$, $X \sim N(\mu = -8, \sigma = 4)$
e) $n = 35$, $X \sim N(\mu = -20, \sigma = 3)$

35 . Si $X \sim N(\mu = 8, \sigma = 3)$ y si seleccionan muestras al azar de 22 eventos, hallar la probabilidad o proporción poblacional de que la media muestral esté (Ej 163, pág. 108)

a) Entre 7 y 9.

b) Entre 4 y 10.

c) Entre 8 y 12.

d) Abajo de 7.

e) Arriba de 10.

f) A una distancia de la media mayor a 4.

36 . (Ej 164, pág. 108) Don Pedro López es taxista de profesión. Sus ganancias por día se distribuyen como $X \sim N(\mu = 5, \sigma = 1)$.

a) En los últimos 16 días hábiles, él ha ganado un promedio de 8: ¿Es eso normal?

b) El sábado por la noche, don Pedro le pegó a la esposa. Pero se arrepintió enseguida pues sabe de sobra que cuando le pega a la esposita en la siguiente semana le irá mal. Y preciso: en los siguientes 7 días don Pedro tuvo un promedio de 2. ¿Tendrá razón don Pedro en ligar la disminución de las ganancias con la conducta hacia la esposa?

c) Formule y resuelva una variante del problema anterior con TLC y demás en la cual don Pedro le regala una lavadora a la suegra, lo cual es hacer algo descomunalmente bueno por su esposa.

37 . (Ej 165, pág. 110) Un estudiante de séptimo semestre tiene notas distribuidas normalmente como N ($\mu = 3{,}5, \sigma = 0{,}3$). En este semestre, el papá le prometió un viaje si se ponía las pilas. Sus notas de las 6 materias dieron un promedio de 3.65, pero el papá no le dio ningún viaje. Explique.

38 . (Ej 166, pág. 110) De tiempo atrás, los empleados le han dado al Jefe un puntaje promedio por semestre de 4.1 con una desviación de 0.4. Pero en los tres últimos semestres, el jefe obtuvo un promedio de 4.2. El Jefe dice que él mejoró. Sus críticos le dicen que mejoró gracias al azar. ¿Qué dice Usted?

39 . Tome la significancia por defecto como 0.05. Para cada inciso, invente una telenovela adecuada y decida la $H_o : \mu = \mu_o$ vs. la H_a dada donde $\bar{X}_n$ es la media de una muestra de n eventos (Ejemplo: 165, pág. 110)

a) $\mu_o = 1$, $X \sim N(\mu =?, \sigma = 3)$, $\bar{X}_7 = 2$, $H_a : \mu > \mu_o$
b) $\mu_o = 12$, $X \sim N(\mu =?, \sigma = 1)$, $\bar{X}_9 = 3$, $H_a : \mu < \mu_o$
c) $\mu_o = 4$, $X \sim N(\mu =?, \sigma = 4)$, $\bar{X}_{16} = 2{,}5$, $H_a : \mu < \mu_o$
d) $\mu_o = 26$, $X \sim N(\mu =?, \sigma = 2)$, $\bar{X}_{25} = 4{,}7$, $H_a : \mu < \mu_o$
e) $\mu_o = 6{,}9$, $X \sim N(\mu =?, \sigma = 6)$, $\bar{X}_{36} = 5{,}8$, $H_a : \mu \neq \mu_o$
f) $\mu_o = 31$, $X \sim N(\mu =?, \sigma = 8)$, $\bar{X}_{27} = 23$, $H_a : \mu \neq \mu_o$

40 . Tome la significancia por defecto como 0.05. Halle por los dos métodos, gráfico y algebraico, el intervalo de confianza para la media con la media muestral dada. Muestre todos los pasos (ej 167, pág. 111). Calcule el error estándar en la estimación de la media (Teorema 162, pág. 107)

a) $X \sim N(\mu =?, \sigma = 3)$, $\bar{X}_7 = 2$
b) $X \sim N(\mu =?, \sigma = 1)$, $\bar{X}_9 = 3$
c) $X \sim N(\mu =?, \sigma = 4)$, $\bar{X}_{16} = 2{,}5$
d) $X \sim N(\mu =?, \sigma = 2)$, $\bar{X}_{25} = 4{,}7$
e) $X \sim N(\mu =?, \sigma = 6)$, $\bar{X}_{36} = 5{,}8$
f) $X \sim N(\mu =?, \sigma = 8)$, $\bar{X}_{27} = 23$

41 . Halle el IC para p si en 5 experimentos de 20 lanzamientos se obtuvieron en promedio por lanzamiento 7 caras. Tome $\alpha = 0{,}05$, 2 colas.

42 . Una cadena de diseños de blusas para dama tiene en su historial de ventas $\mu = 14$, $\sigma = 2,4$ y son 18 tiendas. Hay un premio a todos si el promedio global mejora significativamente en este mes. Si dicho promedio por tienda resultó ser de 0,8. ¿Dio resultado poner el premio?

43 . a) Infiera una fórmula general para el intervalo de confianza de la media de una normal $X \sim N(\mu, \sigma)$ dado un único dato x_o. Manipule su resultado para que le de

$x_o - z\sigma < \mu < x_o + \sigma$

donde z es el z asociado a la significancia tomada y con dos colas.

b) Haciendo manipulaciones algebraicas sobre el intervalo de confianza de la media de una normal, halle el IC para p de una binomial $X \sim Bi(n, p)$ si de n monedas salieron k caras. Manipule el resultado para llegar a

$f - z\sqrt{\frac{f(1-f)}{n}} < p < f + z\sqrt{\frac{f(1-f)}{n}}$

donde $f = k/n$ es la frecuencia relativa y z es el z asociado a la significancia tomada y con dos colas.

c) Comente sobre las virtudes y defectos de los dos métodos que tenemos para estimar p, el constructivo y el presente.

44 . Descifre el siguiente trabalenguas: Hay una contradicción entre el TLC y la metodología para hallar el IC de la media, pues el TLC dice que la media de la media muestral es la misma media, o sea que la media muestral es la media. Pero para hallar el IC de la media uno asume algo diferente, que la media está por ahí cerca de la media muestral sin ser necesariamente igual. Contradicción.

45 . Explique lo que significa y demuestre el teorema fundamental del método científico: entre más datos, el experimento para estimar la media es más preciso. Repita lo mismo pero para la probabilidad de éxito.

46 . En el TLC, la sigma nueva es la vieja dividida por la raíz del número de datos. Nosotros no hemos demostrado dicho teorema. Sin embargo, podemos hacerlo en un caso especial: demuestre el teorema para las binomiales y luego use la aproximación de una binomial por una normal, $Bi(n, 1/2) \sim N(n/2, \sqrt{n/4})$, para verificar o contradecir el TLC para estos casos.

7.2. Parcial II: con libro abierto

El **objetivo** del tema asignado para este parcial fue mostrar en la medida de lo posible de dónde salen las cosas. Una vez que sepamos eso, podremos automatizar, lo cual será hecho en el próximo tema. POR TAL RAZÓN NO SE PUEDE RESOLVER ESTE PARCIAL CON LOS MÉTODOS DEL PARCIAL III O IV a no ser que se explique de dónde sale todo y por qué funcionan los métodos usados.

Cómo estudiar: Este parcial es con libro abierto, apuntes y calculadora pues podrá preguntarse el desarrollo completo de los porqués. Lo primero que debe hacerse para resolver cada problema es entender bien su enunciado. Para ello se toma nota de las palabras claves -las que dan luz sobre cuál método usar- y a continuación se clasifica en la más afín de las categorías siguientes y se sigue el ejemplo:

1. Hallar Ψ para un z dado (ej 79, pág. 63).
2. Hallar z para un Ψ dado (ej 80, pág. 63).
3. Teorema del área grande menos área chica (ej 96, pág. 69).
4. Problema modelo de las camisas (ej 99, pág. 71).
5. Eventos interesantes para normales (ej 111, pág. 80).
6. Eventos interesantes para binomiales (si $p = 1/2$, ej 114, pág. 82. Si $p \neq 1/2$, ej 115, pág. 82). Si n grande, se aproxima por la normal, ej 119, pág. 84).
7. Significancia de un evento para normales (ej 122, pág. 85).
8. Test (decisión) sobre una media con un único dato (ej 126, pág. 87, estudiante engreído).
9. Test (decisión) sobre una binomial (ej 128, pág. 90, desbalance de mujeres).
10. Intervalo de confianza para una media con un único dato (ej 136, pág. 95).
11. Intervalo de confianza para una proporción (ej 138, pág. 97).
12. Proporción de medias muestrales entre dos valores (ej 163, pág. 108).
13. Test (decisión) sobre una media con varios datos (ej 164, pág. 108, cuarto de refrigeración).
14. Intervalo de confianza para una media con varios datos (ej 167, pág. 111).

171 ***Versión 1***

1. El puntaje en Inglés como segunda lengua en un examen de admisión, X, se distribuye como $X \sim N(35, 7)$. Hay 250 estudiantes, los cuales se van a distribuir en 4 cursos. Decir el número de estudiantes por curso sabiendo que se clasifican de acuerdo al puntaje (use gráficas apropiadas):
Principiantes: $X < 20$
Intermedios: $20 \leq X < 38$
Avanzados: $38 \leq X < 47$
De honores: $X \geq 47$.

2. Juanito mide 97cm y los niños de su edad tienen una altura $X \sim N(85, 6)$. Juan, el papá de Juanito, mide 186 y pertenece a una población cuya altura $Y \sim N(180, 10)$. Demuestre que Juanito es relativo a su población más alto que lo que es Juan relativo a la suya. Despliegue gráficas apropiadas. Ayuda: para medir la posición de un dato relativa a su población se usa la tranformada-z. Calcule la z de cada quien y compárelas para decidir quién es más extremo relativo a su población.

3. Uno de los indicadores de la salud es la tensión arterial. A una persona saludable le preocupa que se le baje y se desmaye o que se le suba arriba de 120 y pueda sufrir un derrame cerebral o un infarto de corazón debido a que se pueden reventar arterias importantes. En una persona hipertensa la tensión normalmente está alta y lo que quiere es bajarla. La tensión arterial del abuelito se distribuye como $X \sim N(154, 10)$, su tensión es muy alta y aunque ha pasado por muchos médicos no se mejora. a) ¿qué proporción de medias muestrales de 5 eventos estará arriba de 140?, es decir, ¿qué proporción del tiempo el abuelito tendrá la tensión superalta? b) Una nietecita logró saber que hay un médico que ha logrado milagros y con su propia plata ha llevado al abuelito a que lo formule. El médico lo somete a tratamiento y al hacer 5 mediciones distanciadas (para que sean independientes) y programadas al azar se ha producido un promedio de 128. Formule el informe del médico, con la H_o y demás, justificando sus honorarios teniendo en cuenta que trabaja con $\alpha = 0{,}06$. c) Los tíos no creen que el tratamiento haya servido sino que las mediciones resultaron favorables por mero azar. Formule su objeción en términos oficiales. d) Como los tíos creen que la nietecita no debe pagarle al médico, se ha formado un enorme lío jurídico. Decidieron llamarlo a usted como experto(a) en estadística para mediar: ¿Cuál ha de ser su concepto? Haga gráficas apropiadas.

172 ***Versión 2***

1. Una prenda para medicina ortopédica tiene un largo que se distribuye normalmente con media 15cm y desviación 3.5cm. Se requieren 200 prendas. Despliegue gráficas apropiadas y decir el número de prendas en cada talla sabiendo que hay 4 tallas cuyos rangos son:
S: $X < 10$
M: $10 \leq X < 17$
L: $17 \leq X < 20$
XL:$X \geq 20$.

2. Halle el IC para μ si $\sigma = 2{,}2$ conociendo que hay un dato experimental $x = -3$. Tome $\alpha = 0{,}06$. Muestre todos los pasos por alguno de los dos métodos.

3. Según el historial de un almacén, las ventas X durante un mes son tales que $X \sim N(4,\ 0.4)$. a) Si se toman 6 meses al azar, ¿qué proporción de medias muestrales estarán entre 2.2 y 5.3? b) Un nuevo administrador lleva 6 meses y ha vendido un promedio de 4.3 por mes. Como el contrato se le vence en 8 días, decida ya mismo con $\alpha = 0{,}06$ si Usted le va a renovar el contrato o no. Despliegue gráficas apropiadas. c) Después de que le renovaron el contrato al Administrador aparecieron rumores diciendo que tal decisión fue un **error tipo alfa**, es decir, fue una decisión que asumía que la mejoría observada se debía a alguna razón efectiva (una persona inteligente y bien preparada) que produjo un evento extraordinario, anormal, extremo cuando en realidad era fruto del azar. En una gráfica de ganancias contra tiempo muestre que ésto puede ser posible. Estime la probabilidad de dicho evento. d) ¿Cómo defendería Usted la decisión tomada ante dicha acusación? e) El promedio de ventas durante

los dos años siguientes fue de 4.00. Juzgue entre las dos perspectivas: ¿Fue la renovación del contrato una decisión errónea tipo alfa o no? f) Construya un ejemplo de error tipo beta. Un **error tipo beta** sucede cuando un evento se clasifica como normal cuando en realidad es extraordinario. Ejemplo: Luisa Fernanda entró con temor a la clase de estadística pues venía con la preconcepción de ser tarada para las matemáticas. Eso se debe a que en el colegio era del montón. Pero al trabajar el material y obtener la nota máxima llegó a la conclusión de que el colegio la había engañado: ella era muy buena para las matemáticas, pero una mala metodología le hizo creer que era del montón, sin esperanza.

173 *Versión 3*

1. Halle el intervalo de confianza del 93 % para la media de $X \sim N(\mu =?, \sigma = 2{,}3)$ sabiendo que el promedio de 5 datos experimentales es de 91. Muestre todos los pasos por alguno de los dos métodos.

2. Según la Dinastía (los altos oficales de la Fuerza Aérea), *lo más peligroso de una misión de alto riesgo es aterrizar la nave, pues uno siempre se queda corto de gasolina y no hay nadie a la mano para que ayude*. En su historial del año pasado para el Capitán Aguila 3, la cantidad de gasolina en galones que le quedaba después de aterrizar era $X \sim N(\mu = 2, \sigma = 0{,}4)$, por lo cual fue puesto como ejemplo. Pero en sus últimas 5 misiones de este año, él aterrizó con medio galón en promedio y por ello fue seriamente amonestado con memo a la hoja de vida y todo el cuento. ¿Es sensato tanto problema, tanto escándalo? ¿Es que acaso no se puede confiar en un as de combate? Despliegue gráficas apropiadas.

3. En el parcial I vimos los cuartiles. Revisemos ahora cómo se enmarcan dentro de la normal. Para hallar el cuartil uno en la Z tenemos que hallar la coordenada z tal que antes de ella quede la mejor aproximación al 25 % de la población. Por tanto, buscamos en la tabla el $\Psi = 0{,}25$ y miramos el z correspondiente: nos da $z = -0{,}70$. Para una v.a. $X \sim N(\mu = 3, \sigma = 2)$, el cuartil 1 es la transformada-x de $z = -0{,}70$ que es $x = \mu + z\sigma = 3 + (-0{,}7)(2) = 1{,}6$. Oficialmente escribimos: $P(X < 1{,}6) = \frac{1}{4} = 0{,}25$ para indicar que hemos hallado el cuartil 1.

Ejemplo 2, el cuartil 2 sobre Z es 0 pues antes de 0 queda el 50 % de la población. Sobre cualquier normal la transformada-x de $z = 0$ es la media: por tanto, el cuartil 2 siempre es la media:
$P(X < Q_2 = \mu) = \frac{2}{4} = \frac{1}{2}$.

Halle el cuartil indicado para la variable indicada y despliegue gráficas apropiadas:
a) Cuartil 1 para $X \sim N(7, 1)$.
b) Cuartil 2 para $X \sim N(8, 2)$.
c) Cuartil 3 para $X \sim N(9, 3)$.
d) Cuartil 7 para $X \sim N(10, 2)$.

174 *Versión 4*

1. En estadística la palabra decil significa múltiplo de la décima parte y se refiere a un porcentaje de la población por debajo del evento de referencia. Por ejemplo, para hallar el decil 7 sobre la Z buscamos en la tabla el z tal que antes de ella quede la mejor aproximación al 70 % de la población. Nos da $z = 0{,}50$. Esto implica que el decil 7 sobre cualquier campana es la transformada-x de 0.5. Ejemplo, si $X \sim N(3, 2)$ el decil 7 es la transformada-x de 0.5 que es $x = \mu + z\sigma = 3 + (0{,}5)(2) = 4$. Decimos: el decil k es la barrera d_k sobre la campana antes de la cual hay k décimos de toda la población. Oficialamente escribimos: el **decil** k sobre una distribución $X \sim N(\mu, \sigma)$ es d_k tal que $P(X < d_k) = \frac{k}{10}$.

Halle el decil indicado para la variable indicada y despliegue gráficas apropiadas.
a) Decil 5 para $X \sim N(7, 1)$.
b) Decil 3 para $X \sim N(8, 2)$.
c) Decil 5 para $X \sim N(9, 3)$.

d) Decil 10 para $X \sim N(10, 2)$.

2. Pedro el Grande ha entrado a un nuevo club, el cual tiene 30 jugadores. El ranking en dicho club se actualiza cada mes y se da por medio de una tabla de frecuencias acumuladas que muestra cuántos jugadores tienen un puntaje menor o igual que el borde superior del intervalo considerado.

Auto-ranking	
Intervalos	**No. acumulado de jugadores**
(10, 30]	2
(30, 35]	7
(35, 40]	15
(40, 45]	25
(45, 50]	28
(50, 70]	30

a) Tomando las marcas de clase como el extremo inferior de cada intervalo, halle la media, varianza y desviación del puntaje. Despliegue gráficas apropiadas. Ayuda: calcule la tabla de frecuencias absolutas y siga el ejemplo 23 de la pág. 13. b) Calcule el intervalo de confianza del 95 % para la media poblacional asumiendo que la desviación de la muestra es una excelente aproximación a la desviación de la población. Explique qué quiere decir esta terminología y qué es la muestra y qué la población. c) A Pedro le gusta, le encanta, que su novia lo admire. ¿Cuál sería el puntaje mínimo con que a Pedro le convendría aparecérsele?

3. Olivo hizo un estudio sociológico y llegó a la conclusión de que una buena parte de la violencia se explica no por maldad sino por conflictos en los cuales algún bando defiende sus derechos. Y ha llegado a la atrevida conclusión de que muchos conflictos se arreglarían si sus actores fuesen un poco más tolerantes, aguantando un tricito más a sus enemigos. Así que él mismo ha decidido tomarse un poco de su propia medicina. a) En el tiempo pasado, su número de agresiones por mes tenía una media de 8.1 con una desviación de 2.7. Pero el mes que acaba de terminar agredió a 5 personas (más que todo llamando estúpido a alguien). ¿Le sentó la medicina a Olivo? Despliegue gráficas apropiadas. b) Al cabo de 7 meses el promedio de agresiones de Olivo se mantuvo en 5. ¿Cambió la situación? c) Cuando Olivo supo que en el primer caso la Estadística decía que él no había cambiado mientras que en el segundo sí, preguntó: En el primer mes mi actitud fue tan seria como en los siguientes 6 meses, ¿cómo es posible que la Estadística me lo niegue? Respóndale a Olivo de manera sencilla y contundente.

175 *Versión 5*

1. En un experimento con monedas, de 16 lanzamientos cayeron 14 caras. Use la aproximación normal para hallar todas las p tales que el dicho experimento respalda con $\alpha = 0{,}08$ la $H_o : \ X \sim Bi(16, p)$. Muestre de dónde sale todo.

2. Halle el intervalo de confianza del 90 % para p, la probabilidad de cara en una binomial, si de 20 lanzamientos salieron 2 caras. Muestre todos los pasos.

3. En un país hay dos grupos, A y B. Por cada 2 personas de A hay 1 de B. En el presente Senado hay 170 personas todas elegidas por voto popular, 120 del grupo A y 50 del grupo B. Los del grupo B alegan que la democracia del país es ficticia, pues ellos, haciendo su mejor esfuerzo, tienen menos probabilidades de ser justamente representados en el Senado que los del grupo A. ¿Se justifica la acusación del grupo B? Despliegue gráficas apropiadas.

176 *Versión 6*

1. Desarrollando la binomial, halle los eventos interesantes de la cola superior con $\alpha = 0{,}04$ para $X \sim Bi(3,\ 0.08)$. Despliegue gráficas apropiadas.

2. Si $X \sim N(7, 2)$ y si se seleccionan muestras de 38 eventos, hallar la proporción de medias muestrales que estarán entre 6 y 9. Despliegue gráficas apropiadas.

3. En el presente semestre por cada 10 hombres en la U hay 8 mujeres, pero en la Facultad de Pintura y Escultura hay 120 personas, 70 mujeres y 50 hombres. ¿Se justifica decir que en dicha Facultad las mujeres están sobrerepresentadas? Despliegue gráficas apropiadas.

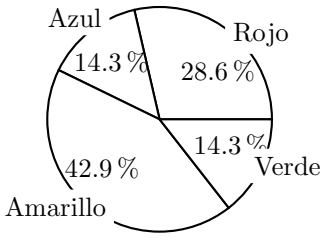
Respuestas Parte I

Problemas del Taller I

Pág 47; No 1.a. Los colores representan proyecciones ideológicas. El amarillo representa la riqueza prometida, el rojo corresponde al héroe que todos llevamos dentro. Los colores son datos que no se pueden sumar (no son datos cuantitativos), ni ordenar (no son datos ordinales), son datos cualitativos o categóricos a los que se les puede sacar una tabla de frecuencias relativas y una torta. Los grados asociados a cada categoría los sacamos por regla de tres: si a 1 le corresponde 360 grados, ¿cuántos le corrresponde a 0.286? Le corresponde 0.286 por 360 = 103.

Tabla de Frecuencias			
Colores	Absoluta	Relativa	Grados
Rojo	2	2/7 = 0.286	103
azul	1	1/7= 0.143	51
amarillo	3	3/7=0.429	154
verde	1	1/7=0.143	51
Total	7	1.001	359

La torta asociada a la tabla de frecuencias relativas es la siguiente:



El color más popular es el amarillo, lo cual significa que la gente se interesa sobre todo en la riqueza, lo cual puede reducirse a tener lo indispensable para supervivir.

b. La mayoría de personas podría atreverse a considerar que los datos son ordinales, aunque la forma de ordenarlos varía mucho de persona a persona. Una de ellas podría ser:

Indiferencia (1/ 8) $->$ desprecio (3/8) $->$ rabia (1/8) $->$ admiración (2/8) $->$ cariño (1/8).

Hay 3 personas de 8 que abrigan un sentimiento positivo, los demás negativo. Todos esperamos que el líder sea inteligente y se asesore para tratar de cambiar hacia lo positivo.

c. No tenemos motivos para analizar las letras iniciales. La duración de los aplausos son datos ordinales, pues el que recibe más aplausos queda mejor que el que recibe menos. La frecuencia relativa de que un aplauso dure muchísimo es 2/6, las demás son 1/6. Eso quiere decir que la comunidad ya tiene la dicha de contar con 2 ediles que la representen y con los cuales identificarse plenamente. Por ello consideramos que la democracia en la localidad está madura.

Pág 47; No 2 La idea es protejer los lugares donde se presentan lesiones y, si es posible, no proteger aquellos lugares que no peligran.

Pág 47; No 3 a) Rangos y puestos.

Puntajes, rango y percentiles			
Nombre	Votos	Rango	Puesto
Comida para indigentes	1	0	9
Estudio gratis	4	1	7
Trabajo y paz	4	1	7
Seguridad en la calle	6	3	6
Naturaleza en paz	7	4	3
Muerte a violadores	7	4	3
Más policia	7	4	3
No a la corrupción	8	7	2
Trabajo para jóvenes	11	8	1

b) Como hay nueve datos, la mediana sería el dato de la mitad, el quinto dato y esto corresponde a un desempate entre los que tienen 7 votos, que son tres partidos, Naturaleza en paz, Muerte a violadores y Más policía. c) A Trabajo y paz le fue muy mal pues quedó por debajo de la mediana.

Pág 48; No 4 Los 4 métodos deben dar los mismos resultados, o algo anda mal.

4.a $\bar{X} = 3{,}25$, $s = 1{,}035$

4.b $\bar{X} = 3{,}125$, $s = 1{,}246$

4.c $\bar{X} = 2{,}666$, $s = 1$

4.d $\bar{X} = 3{,}44$, $s = 1{,}33$

4.e $\bar{X} = 2{,}285$, $s = 0{,}756$

Pág 48; No 5 El diseño de la silletería de un teatro debe permitir que cada quien se siente donde quiera o quiźa donde le toque. Por lo tanto, todas las sillas deben ser iguales. Para escoger el patrón a seguir, se tiene en cuenta que una persona pequeña, como un niño, puede sentarse cómodamente en una silla grande, pero en general las personas robustas se sienten incómodas en sillas pequeñas. Por tanto, se escoge como patrón una persona sesgada hacia el lado de las robustas. El caso con los zapatos se soluciona si se tiene en cuenta que éstos le deben quedar a uno un poquito holgados pero no demasiado no sea que uno se tropiece. Por lo tanto, el ajuste debe ser casi perfecto.

Pág 48; No 6 a. Nota =3.296;

b. Nota necesaria imposible= 5.252;

c. Una décima es suficiente (es la aproximación en décimas a 0.04).

Pág 49; No 7 a). Primero que todo observamos que los datos van desde 20 hasta 96. Por tanto, dividimos los datos en intervalos usando el criterio de la flojera: 20 a 30, 30 a 40 y así hasta 90-100. Ahora ordenamos los datos en un diagrama de tallo y hojas (usamos la representación más sencilla):

```
20 22 24
39
43 47 47 47
51 52 55 56 57 58 58 59
61 64 68
71 71 74 74 75

96
```

Contamos los datos para estar seguros de que no falta nadie: son 25. Para estimar la media, trazamos mentalmente una campana: la media está alrededor de 50. La desviación alrededor de 10.

Determinemos los datos extremos: dos desviaciones miden 20. Por lo tanto, un dato es extremo por arriba si queda arriba de 70 y otro dato es extremo por abajo se queda abajo de 30. Siendo así las cosas, encontramos que hay tres valores extremos por abajo y en cambio por arriba habrían 6 datos lo cual no concuerda muy bien con el sentido común, pues casi todo el mundo sería extremo. Concluimos que la media no puede ser 50 sino que debe estar más arriba. Si ponemos la media en 55 y conservamos la misma desviación, los valores normales irían desde 35 hasta 75. Eso es más razonable así que declaramos que los valores extremos son 20, 22 y 24 por abajo y 96 por arriba.

b) Cuartiles por regla de tres. Aproximamos las posiciones hacia ariba pues dan decimales:

$Q_1 \rightarrow L_{25} = 25 * 0{,}25 = 6{,}25 \rightarrow$ Dato 7 = 47;
$Q_2 \rightarrow L_{50} = 25 * 0{,}50 = 12{,}5 \rightarrow$ Dato 13 = 57 = mediana;
$Q_3 \rightarrow L_{75} = 25 * 0{,}75 = 18{,}75 \rightarrow$ Dato 19 = 68;

Resumen de 5 valores: 20, 47, 57, 68, 96. La barrera de lo normal por abajo es 15.5. Como el mínimo es 20, no hay extremos por abajo. Barrera por arriba = 99.5. No hay valores extremos por arriba según el criterio del rango intercuartilar.

c) Cuando se agrupa por intervalos la respuesta depende del modo de agrupación y del número de decimales. Del diagrama de tallo y hojas sacamos las marcas de clase como el dato de la mitad y su frecuencia absoluta y con ello fabricamos la siguiente tabla:

Cálculo de la varianza			
x = mitad del intervalo	F = frecuencia	xF	x^2F
25	3	75	1875
35	1	35	1225
45	4	180	8100
55	8	440	24200
65	3	195	12675
75	5	375	28125
85	0	0	0
95	1	95	9025
Total	$n = \sum F = 25$	$\Sigma xF = 1395$	$\Sigma x^2F = 85225$

n = Número de objetos = $\sum F = 25$, $\sum xF = 1395$, $\sum x^2F = 85225$, Media = $\sum xF/n = 55.800$, $S_{xx} = \sum x^2F - (\sum xF)^2/n = 7384$, Varianza = $S_{xx}/(n-1) = 307.667$, Desvación = 17.540 , Coeficiente de variación = 0.314. Como este coeficiente está cerca de 0.5, consideramos que la población es altamente inhomogénea. Por lo tanto, deberán tenerse razones de peso para poder tomar decisiones al mismo tiempo que se asume que la media representa fielmente a toda la población.

Para hallar los datos extremos, sumamos y restamos a la media dos desviaciones: 55.8 $\pm$ 2(17.54) = 55.8 $\pm$ 35.08. La barrera inferior (que divide lo normal de lo extremo) es 20.72 y la superior 90.80. Datos extremos por abajo: 20, por arriba 96.

d) Funcionar con software ilegal es la manera natural de las compañías. Por tanto proponemos: 1) debe haber una transición suave e inteligente hacia la legalidad para no quebrar al país. 2) Hacer negocios como nación con las compañías extranjeras que producen programas de consumo masivo y no dejar que las compañías negocien solas. 3) Ofrecer capacitación sobre software gratuito que es muy abundante, de muy buena calidad y hoy en día ya es muy amigable. 4)Incentivar la formación de asociaciones de consumo.

Pág 49; No 8a) Los datos ordenados son:

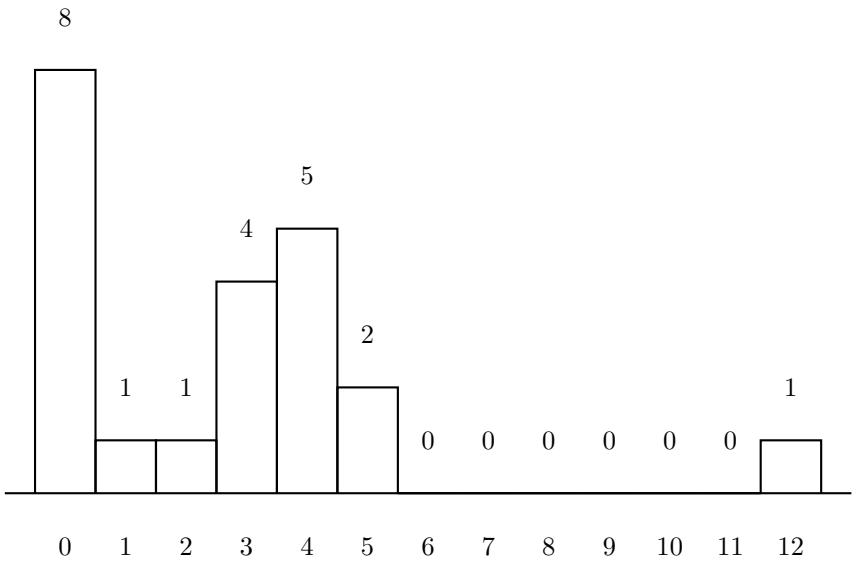
0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 3, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 12.

Datos del boxplot: min = 0; Q1 = 0; Q2 = 3; q3 = 4; max = 12

Se ve claramente que el 75 % de las personas hace por semana 4 preguntas o menos, aunque una buena proporción no hace ninguna pregunta.

b) El rango intercuartilar es q3-q1 = 4-0 = 4. La aleta intercuartilar es 1.5 veces el rango = 4 veces 1.5 = 6. Los datos extremos por abajo quedan debajo de q1 menos la aleta = 0 - 6 = -6 que se tiene que aproximar a cero. Los datos extremos por arriba quedan por encima de q3 + rango = 4 + 6 = 10. Por tanto, la persona que pregunta 12 veces es extrema de acuerdo al criterio del rango intercuartilar en esta población.

c) El histograma:



Los datos definitivamente quedan mal representados por una campana. Sin embargo, al tratar de dibujar la campana que mejor se ajusta al histograma, la media vale 2.6 y la desviación vale 2.8. El criterio de las dos desviaciones daría que 12 sería un valor extremo.

d) Sin el doce, los datos tienen media 2.14, varianza 3.7 y desviación 1.93. El coeficiente de variación es 0.90: tenemos una población altísimamente inhomogénea y de ninguna manera puede representarse por la media. El histograma tiene un pico en cero y una montañita con media 3. Esto nos dice que la población parece estar subdividida por lo menos en dos partes: los que están desconectados y nada tienen que preguntar y los que ya lograron conectarse y tienen dudas y preguntas. Los desconectados están a la moda. ¡Alerta morada!

Pág 49; No 9. Unos estudiantes calcularon el coeficiente de variación de todos los 6 datos que les dió 0.646, por lo que los datos se clasifican como altamente no homogéneos. Para probar que la causa era el 5, lo quitaron, calcularon la media de los 5 datos restantes que fue 1.64. La desviación les dió 0.41 y el coeficiente de variación 0.25, por lo que los nuevos datos ya era más o menos homogéneos. La región de lo normal, según el criterio de las 2 desviaciones, dió desde 0.82 hasta 2.46. Como el 5 queda muy por fuera de dicha región, es claramente un outlier. Otros estudiantes calcularon el rango de los normal de todos los 6 datos según el criterio del rango intercuartilar: les dió entre 0.6 y 3 y la mediana dió 1.8. Quitando el 5 lo normal va de 1.04 a 2.25 con la misma mediana. Se ve que el 5 es un outlier muy extremo. Ellos podrían alegar por eso que dicha nota no debe tenerse en cuenta en la negociación. Además, vemos que el criterio del rango intercuartilar es mucho más incisivo que el de las dos desviaciones.

Pág 50; No 10. Primero dilucidamos el enunciado: hay 80 personas partidarias de una acción militar. Esto es muy distinto de decir: hay 80 personas partidarias exclusivamente de una acción militar. Por consiguiente, las 80 personas ocupan un círculo completo y no una luna creciente.

a) $p(A) = 80/120$.
b) $p(A - B) = 72/120$.
c) $p(B) = 20/120$.
d) $p(B - A) = 12/120$.
e) $p(A \cap B) = 8/120$.
f) $p(AUB) = 92/120$.
g) $p[(AUB)'] = 28/120$.
h) $p(A') = 40/120$.
i) $p((A - B)') = 48/120$.
j) $p(B') = 100/120$.
k) $p((B - A)') = 108/120$.
l) $p(B/A) = 8/80$.
m) $p(A/B) = 8/20$.

Pág 50; No 11 Primero dilucidamos el enunciado: hay 70 países que tienen petroleo, de los cuales algunos tienen fuentes hidroeléctricas. Por tanto, se trata de un círculo completo. Las ecuaciones son: $a+b+c+4=150$; $a+b=70$; $b+c=130$. Las diyunciones son: $a=16$, $b=54$, $c=76$.

a) $p(A)=70/150$.
b) $p(A-B)=16/150$.
c) $p(B)=130/150$.
d) $p(B-A)=76/150$.
e) $p(A\cap B)=54/150$.
f) $p(A\cup B)=146/150$.
g) $p((A\cup B)')=4/150$.
h) $p(A')=80/150$.
i) $p((A-B)')=134/150$.
j) $p(B')=20/150$.
k) $p((B-A)')=74/150$.
l) $p(B/A)=54/70$.
m) $p(A/B)=54/130$.

Pág 50; No 12 La $Bi(3,2/9)$. Llegar tarde se nota a, llegar temprano como e. El árbol generador:

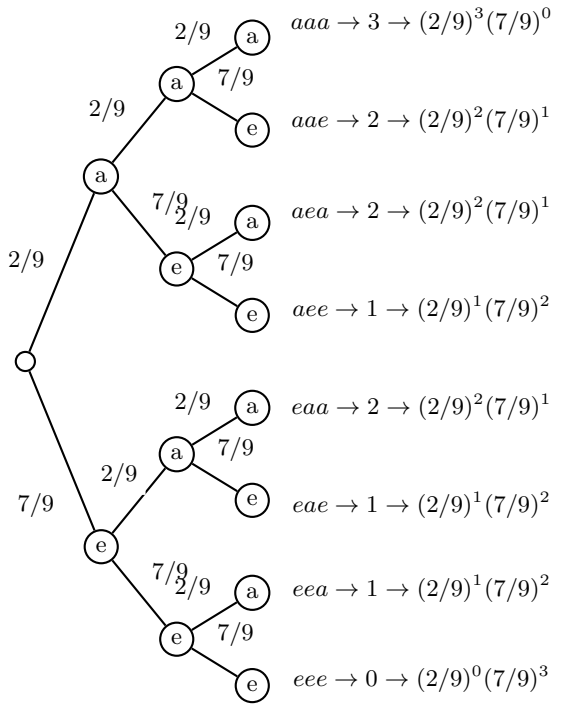


Figura 7.0. Cálculo gráfico de la $Bi(3,2/9)$.

La media, μ, y la varianza, σ^2, se calculan así:

$X\sim Bi(3,2/9)$			
x	$p(x)$	$xp(x)$	$(x-\mu)^2p(x)$
0	$(2/9)^0(7/9)^3=0{,}4705$	$0\ (0.4705)=0$	$(0-0{,}6668)^2(0{,}4705)=0{,}2091$
1	$3(2/9)^1(7/9)^2=0{,}4033$	$1\ (0.4033)=0.4033$	$(1-0{,}6668)^2(0{,}4033)=0{,}0448$
2	$3(2/9)^2(7/9)^1=0{,}1152$	$2(0.1152)=0.2305$	$(2-0{,}6668)^2(0{,}1152)=0{,}2048$
3	$(2/9)^3(7/9)^0=0{,}0109$	$3\ (0.0109)=0.0330$	$(3-0{,}6668)^2(0{,}0109)=0{,}0597$
Σ	$\Sigma p(x)=1{,}000$	$\mu=\Sigma xp(x)=0{,}6668$	$\sigma^2=\Sigma(x-\mu)^2p(x)=0{,}5184$

La media vale 0.6668. Como la varianza vale $\sigma^2 = 0{,}5184$, la desviación vale su raíz cuadrada:

$\sigma = \sqrt{0{,}5184} = 0{,}7201$

Como el promedio es 0.6668 por cada 3 clases, en dos semanas tendrá un promedio de llegada tarde de 1.3. Es decir, Adelita llega tarde aproximadamente una vez cada dos semanas.

Hagamos el histograma:

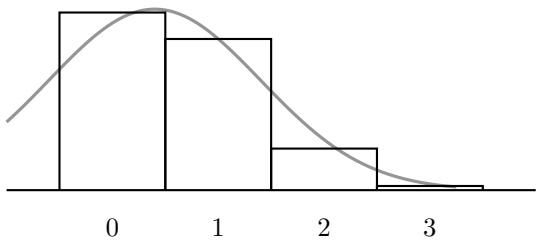


Figura 7.1. Histograma de $X \sim Bi(3, 2/9)$. La media está cerca de 0.5, la desviación estándar vale aproximadamente 1.

Pág 50; No 13 La tabla de probabilidades de $X \sim Bi(6, 1/3)$

$X \sim Bi(6, 1/3)$	
x_i	p_i
0	$1(1/3)^0(2/3)^6 = 0{,}0878$
1	$6(1/3)^1(2/3)^5 = 0{,}2634$
2	$15(1/3)^2(2/3)^4 = 0{,}3292$
3	$20(1/3)^3(2/3)^3 = 0{,}2195$
4	$15(1/3)^4(2/3)^2 = 0{,}0823$
5	$6(1/3)^5(2/3)^1 = 0{,}0165$
6	$1(1/3)^6(2/3)^0 = 0{,}0014$
Suma	$\Sigma p_i = 1{,}0001$

μ=2; σ^2=1.33; σ=1.155

Pág 50; No 14 a)$X \sim Bi(4{,}0.2)$ Probabilidad de que se miren al menos una vez $= p(1) + p(2) + p(3) + p(4) = 1 - p(0) = 0{,}5904$, es plausible que las echen. b) $1 - p(0) = 0{,}995$. Es muy probable que las echen: lo mejor es que se hagan separadas.

Pág 51; No 15 c) Toda frecuencia relativa está entre cero y uno, la suma de todas las frecuencias relativas da uno exacto. d) Una probabilidad tiene las mismas características básicas que una frecuencia relativa.

Pág 51; No 16. El tema del retardo del profesor se trata como sigue:

a) La decisión a tomar es irse o no de clase. Por lo tanto el investigador es la clase y la muestra es el profesor. La variable a medir es el tiempo de llegada a clase del profesor. El evento es la hora de llegada del día de hoy del profesor a su clase. La población está conformada por todos los profesores a los que se les registra su puntualidad en todas su clases.

b) Por múltiples razones un profesor puede demorarse un poquito para llegar a clase. La experiencia demuestra que en general el retardo de cualquier profesor no pasa de 5 minutos. Así que es muy sensato que los estudiantes se vayan al cuarto de hora de espera pues la probabilidad de que el profesor llegue después es prácticamente cero.

c) La variable a medir, el tiempo de llegada, es continua. Esta variable puede considerarse aleatoria pues hay múltiples razones que interfieren con la llegada a clase de un profesor, como encontrarse con

un reclamo inaplazable. Con todo, hay una queja contra la supuesta aleatoriedad: si a un profesor le toca llegar tarde un día, se esforzará sobremanera por llegar a tiempo el día siguiente, es decir, que a veces uno puede predecir el próximo dato.

d) Como regla estadística, todos los profesores procuran llegar antes de la hora prefijada para empezar clase pero en la práctica se retardan uno o dos minutos. Apoyamos la idea de que lo deseable es irse a las y cuarto, perdonar al profesor y armar la alaraca si vuelve a suceder.

e) En general, el procedimiento para irse o no de clase es el siguiente: el valor de la muestra (tiempo de retardo) se contrasta con el valor esperado según los objetivos fijados (el profesor debe llegar a tiempo). Si hay una gran discrepancia entre la muestra y lo que se cree (un retardo mayor que un cuarto de hora), se considera que los datos rechazan lo que se cree y se toma la decisión de irse de clase, pero si la discrepancia es pequeña (un retardo menor que un cuarto de hora), la discrepancia se tolera y se decide que los objetivos se han cumplido: el profesor llegó dentro del margen de tolerancia y hay clase.

Pág 53, Parcial 1, v.1: 1) Para calcular la mediana y los cuartiles, primero se ordena de menor a mayor, lo cual en la tabla siguiente es de arriba a abajo.

Muchachos y sus notas				
Nombre	**Puntaje en Examen**	Posición	Rango	Puesto
Camila Fernández	15	1	0	10
Mónica Bernal	21	2	1	9
Pablo Rodríguez	26	3	2	8
Juan Pérez	37	4	3	6
María Puentes	37	5	3	6
Juana Maldonado	39	6	5	5
Diana Suárez	42	7	6	4
Andrés Umaña	43	8	7	3
Pedro Hernández	46	9	8	2
Felipe Rueda	48	10	9	1

La mediana es el dato de la mitad, es decir, el promedio de 37 y 39 que es 38. Arriba de 38 hay 5 datos lo mismo que abajo. Percentil 45: $L_{45} = 10 \times 45/100 = 4{,}5$, como es una posición fraccionaria, se aproxima a la superior que es 5: el percentil 45 corresponde a Juan y María, ambos con 37 puntos. El cuartil 3 es el percentil 75 y lo tiene la posición 8, Andrés Umaña con 43 puntos. Pedro quedó arriba de la mediana y ocupó el puesto 2: le fue muy bien. ¿O acaso muy mal: por qué no fue el primero?

Pág 53, Parcial 1, v.1: 2) Las respuestas varían porque dependen de la partición, pero son aproximadamente: $\bar{X}$=59.2; s^2=103.33; s=10.45; Coeficiente de variación = 0.17

Pág 53, Parcial 1, v.1: 3) a) $\mu = 0{,}857$, $\sigma^2 = 0{,}6122$, $\sigma = 0{,}7824$ b) p(al menos dos caras) = p(2) + p(3) = 0.197.

Pág 53, Parcial 1, v.2: 1a) La tabla de frecuencias absolutas convenientemente aumentada para calcular la varianza es la siguiente:

Cálculo de la varianza, Edades de padres				
Intervalos	x	F = frecuencia	xF	x^2F
[35, 40)	35	7	245	8575
[40, 45)	40	5	200	8000
[45, 50)	45	5	225	10125
[50, 55)	50	28	1400	70000
[55, 60)	55	7	385	21175
[60, 65)	60	6	360	21600
	Total	$n = \sum F = 58$	$\Sigma xF = 2815$	$\Sigma x^2F = 139475$

$\bar{X} = 48.534$; $S_{xx} = 2850.431$; $s^2 = 50$; $s = 7.07$ Coeficiente de variación = 0.146 , población +/− homogénea.

Pág 53, Parcial 1, v.2: 2) Los coeficientes los sacamos del triángulo de Pascal:

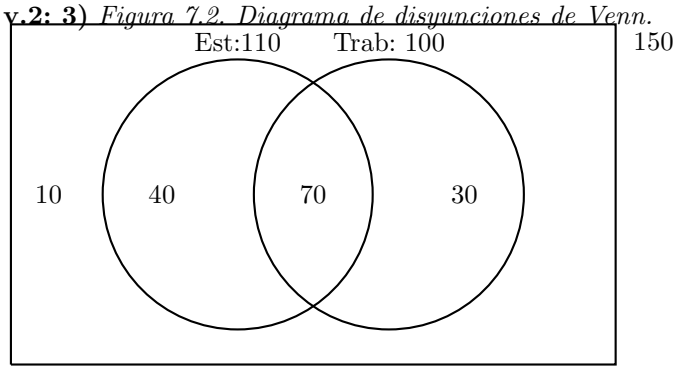
(1) 1 1
(2) 1 2 1
(3) 1 3 3 1
(4) 1 4 6 4 1
(5) 1 5 10 10 5 1

Insertamos los coeficientes del último renglón en la tabla de probabilidades, la cual es:

$X \sim Bi(5, 1/3)$			
x_i	p_i	$x_i p_i$	$(x_i - \mu)^2 p_i$
0	$1(1/3)^0(2/3)^5 = 0{,}1317$	0	0,3658
1	$5(1/3)^1(2/3)^4 = 0{,}3292$	0,3292	0,1463
2	$10(1/3)^2(2/3)^3 = 0{,}3292$	0,6584	0,0366
3	$10(1/3)^3(2/3)^2 = 0{,}1646$	0,4938	0,2926
4	$5(1/3)^4(2/3)^1 = 0{,}0412$	0,1646	0,2241
5	$1(1/3)^5(2/3)^0 = 0{,}0042$	0,0206	0,0457
Suma	$\Sigma p_i = 1{,}0001$	$\mu = 1{,}666$	var $= \sigma^2 = 1{,}111$

La desviación es 1,0541. El coeficiente de variación es 0.6, el cual es gigantesco: es muy mala idea representar esta distribución por la media. Se necesitan varias clases. La probabilidad de que salgan a lo sumo dos caras es $p(0) + p(1) + p(2) = 0{,}7901$.

Pág 53, Parcial 1, v.2: 3) *Figura 7.2. Diagrama de disyunciones de Venn.*



Probabilidad de hacer algo diferente a sólo estudiar = 110/150.

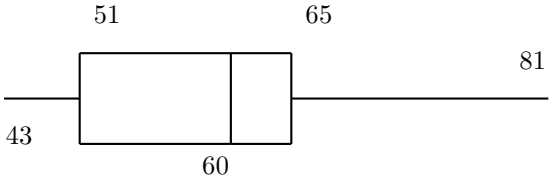
Pág 55, Parcial 1, v.3: 1) Para Bogotá: $\bar{X}$=59,6; s^2=117.6; s=10.84; Coeficiente de variación= 0,18. Diagrama de tallo y hojas para Bogotá:

```
43 45 48 49 49
50 51 55 56 57 57 58
60 62 62 63 64 64 65 65 66 67
71 73
81
```

A 'ojímetro', la media está aproximadamente en 58 y su desviación en 7. Los cuartiles:

$Q_1 \to L_{25} = 25 * 0{,}25 = 6{,}25 \to$ Dato 7 = 51;
$Q_2 \to L_{50} = 25 * 0{,}50 = 12{,}5 \to$ Dato 13 = 60 = mediana;
$Q_3 \to L_{75} = 25 * 0{,}75 = 18{,}75 \to$ Dato 19 = 65;

Resumen de 5 valores: 43, 51, 60, 65, 81.
El boxplot está determinado por los siguientes 5 valores: 43, 51, 60, 65 y 81:



Los que están por encima del tercer cuartil están desperdigados lo cual quiere decir que ser juez es una profesión muy difícil y con mucha dificultad se logra la excelencia. Los que están entre la mediana y el tercer cuartil están muy concentrados: es factible ser mejor que la mediana pero con dificultad se logra ser mejor que el 75 %. Las personas que cubren los cuartiles inferiores tienen diferentes vacíos difíciles de llenar y por eso cubren un rango amplio.

El rango intercuartilar es Q3 - Q1: 65-51 = 14. La aleta intercuartilar es 1.5 veces el rango intercuartilar: $1{,}5 \times 14 = 21$. Restamos este valor de Q1 para hallar los valores extremos del lado inferior: 51 - 21 = 30. Análogamente, los valores extremos por arriba son los mayores que el tercer cuartil más la aleta: 65 + 21 = 86. Vemos que no hay valores extremos ni por arriba ni por abajo: hay una preocupación efectiva entre los jueces de Bogotá por ir con el grupo. A un grupo así le funciona muy bien los programas de capacitación.

Pág 55, Parcial 1, v.3: 2) $X \sim Bi(7, 0.18)$. a) p(Mínimo 3) = p(3) + p(4) + p(5) + p(6) + p(7) = 0.115. b) p(máximo 4) = p(0)+ p(1) + p(2) + p(3) + p(4) = 0.997. c) p(todos rechazados) = p(0 aceptados) = p(0) = 0.249.

Pág 55, Parcial 1, v.3: 3) Tres ecuaciones: a + b = 12; b + c= 18; 10+a+b+c= 34. Solución:

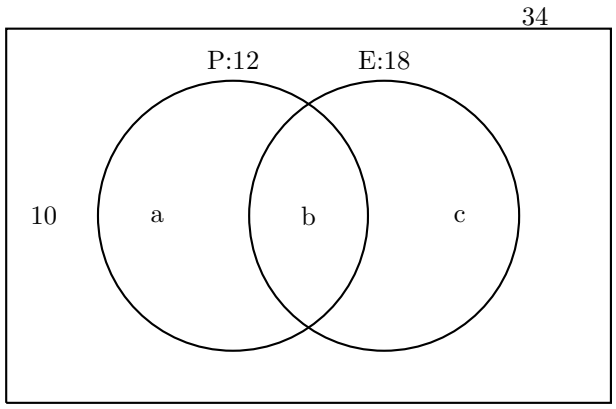


Figura 7.3. Diagrama de disyunciones de Venn: a= 6, b = 6, c = 12

Método alternativo: Todo = 34 = 10 + (P unido a E) = 10 + P + E - Intersección = 10 + 12 + 18 - Intersección. Por tanto: Intersección = 40 - 34 = 6. Probabilidad de ser positivista extremo = 6/34. Probabilidad de ser algo diferente de existencialista extremo = 22/34.

Pág 55, Parcial 1, v.4: 1a) $X \sim Bi(50, 1/2000)$, probabilidad de que nadie se enferme = p(0). **1d)** Probabilidad de que al menos uno se enferme = 1-p(0). **1e)** Como p(2) es casi cero en tanto que p(1) es pequeño pero visible, el profesor concluyó: por puro azar es posible que alguien resulte enfermo. Pero por puro azar es muy improbable que dos resulten enfermos. Por lo tanto, de dos enfermos al menos uno es mentiroso. **2b)** $X \sim Bi(7, 1/100)$, probabilidad de que se le de permiso al menos a uno = 1- p(0). **3)** $X \sim Bi(3, 1/100)$.

Pág 56, Parcial 1, v.5: 1) Media = 3.6, Varianza = 0.9333, Desviación = 0.9661, Coeficiente de variación = 0.2686.

Pág 56, Parcial 1, v.5: 2) los coeficientes de la $Bi(10, 1/7)$ son

1, 10, 45, 120, 210, 252, 210, 120, 45, 10, 1.

La tabla de probabililidades es:

a) Las probabilidades de la $X \sim Bi(10, 1/7)$	
x	$p(x)$
0	0.21405831774189094
1	0.35676386040580443
2	0.26757289343134305
3	0.11892128513703677
4	0.0346853745888381
5	0.006937074869208109
6	9.634826139789682E-4
7	9.176024830805634E-5
8	5.735015479108413E-6
9	2.124079792208631E-7
10	3.5401329622334536E-9
Sum	0.9999999999999998

b) La media es $np = 10/7 = 1{,}42$. La varianza es $np(1-p) = 1{,}22$, la desviación es 1,1. El coeficiente de variación es 0,78: la población es muy inhomogénea. c) Le alcanza para pagar si se gane 3 multas o menos con $p = 0{,}93$. d) Para que gaste todo lo que lleva debe ganarse 6 multsa o menos $p = 0{,}96$. d) Queda hipotecado si se gana las 10 multas con $p = p(10)$, casi cero. No es realista esperare que pase en un único viaje.

Pág 56, Parcial 1, v.5: 3) $X \sim Bi(4, 3/7)$. X cuenta el número de caras. Si X cae cara, uno crece 40cms, si cae sello, 50. La altura la denotamos h. Si caen 0 caras, entonces cayeron 4 sellos, por lo tanto la persona mide 200cms. La tabla básica es (por favor, terminar de rellenarla):

$X \sim Bi(4, 3/7)$				
x	h	$p(h)$	$hp(h)$	$(h-\mu)^2 p(h)$
0	200	$(3/7)^0(4/7)^4 = 0{,}107$		
1	190	$4(3/7)^1(4/7)^3 = 0{,}320$		
2	180	$6(3/7)^2(4/7)^2 = 0{,}360$		
3	170	$4(3/7)^3(4/7)^1 = 0{,}180$		
4	160	$(3/7)^4(4/7)^0 = 0{,}03$		
Σ		$\Sigma p(h) = 0{,}997$	$\mu = \Sigma hp(h) =$	$\sigma^2 = \Sigma(h-\mu)^2 p(h) =$

Según el histograma, la media vale alrededor de 185 y la desviación 10.

Pág 56, Parcial 1, v.6: 1) Probabilidades: (0) 0.0062; (1) 0.0496; (2) 0.1653; (3) 0.2938; (4) 0.2938; (5) 0.1567; (6) 0.0348; Media = 3.429; Var = 1.47; Desviación = 1.212.

Pág 56, Parcial 1, v.6: 2) Análisis cuantitativo: Media = 39.2, Var 336.3724, desviación = 18.3. Criterio de las dos desviaciones: Barrera superior de los datos normales = 75.8. Barrera inferior = 2.52. Análisis ordinal: Min = 12. Q1=23. Mediana = Q2 = 33.5. Q3 = 55. Max.=76. riterio del rango intercuartilar: Barrera superior de los datos normales = 103. Barrera inferior = -25. Como medir el tiempo resultó inútil, puede medirse al ángulo desde cuando se está com la cabeza hacia arriba en la última vuelta y la posición de la cabeza en el momento de estirar las piernas. Los ángulos se pueden medir usando referentes externos ojalá lejanos.

Problemas del Taller II

Pág 117; No 1.a Las respuestas son a gusto personal, por ejemplo: tomamos el mínimo como 1.50m y el máximo como 1.90m. La media queda en 1.70m y la desviación vale 10cm. b) Entre 1.60 y 2 metros queda más de la mitad, algo así como 60 %.

Pág 117; No 2.b. 0.9893

Pág 117; No 3.a) -0.40, **b)** 0.75, **d)** No existe pues el máximo valor de ψ es uno que lo toma en infinito. **e)** No existe pues el el mínimo valor de ψ es cero que lo toma en menos infinito.

Pág 117; No 4.a. 0.8893

Pág 118; No 5.) Media= 3; Desviación: 1,09; b) z(1)=-1.83; $\Psi(-1{,}83) = 0{,}03$, es decir, aproximadamente el 3 % de las personas tienen una nota menor a 1; z(2)=-0.91; $\Psi(-0{,}91) = 0{,}18$; z(3)=0; $\Psi(0) = 0{,}5$; z(4)= 0.91; $\Psi(0{,}91) = 0{,}81$; z(5)= 1.83; $\Psi(1{,}83) = 0{,}96$.

Pág 118; No 6. a) 17.4 b) 2.2 c) 0.0009977 d) 1192 e) -105.5

Pág 118; No 7.a. Probabilidad= 0.5993

b. Probabilidad= 0.6914

c. Probabilidad= 0.32

d. Probabilidad= 1

e. Probabilidad= 0

f. Probabilidad= 0.68

Pág 118; No 9 Camisas a producir=

S: 518

M: 3464

L: 455

XL: 63

Pág 120; No 14 a) El profesor se rige por la cola superior.

b) Al bajarse a 4.5 el profesor aumentó la significancia de trabajo pues la proporción de estudiantes que clasifican como extremos son ahora más pero la cola sigue siendo la superior pues sólo le importa lo mejor.

c) Cola superior. Un recorte disminuye la significancia pues ahora son menos los que son importantes, los favorecidos con subsidio.

d) Cuando el colegio es principiante debe mostrar por lo menos alguien bueno: cola superior, significancia muy baja. Cuando el colegio madura, puede mostrar un aumento masivo de la calidad, por tanto la significancia aumenta.

e) Para que el colegio sea una empresa rentable debe tener buen nombre. Por eso, debe atenderse a corto plazo a los que demuestren la mayor tasa de aprendizaje, normalmente en la cola superior de X. Si la nota para pasar aumenta sustancialmente, lo mejor es emprender un programa de mejoramiento a largo plazo: todos son importantes, la significancia, la proporción de los que son importantes, se acerca a uno.

Pág 120; No 15 Si $\alpha = 0{,}01$ el z_c de la cola inferior sería -2.3: todo lo que quede más abajo sería interesante. Con cola superior: lo que quede arriba de $z_c = 2{,}3$. Con dos colas, las $z_c = \pm 2{,}6$: lo que quede abajo de -2.6 junto a lo que quede arriba de 2.6 sería interesante.

Pág 120; No 16 a. (-∞, 3.5)

b. (5.6, ∞)

c. Eventos en donde $X < 1$ ó $X > 33$ pues la significancia se reparte entre las dos colas.

Pág 120; No 17.a. El único evento interesante es 4.

b. No hay.

c. Probabilidades halladas por triángulo de Pascal: $p(0) = 0{,}9140$, $p(1) = 0{,}0831$, $p(2) = 0{,}0028$, $p(3) = 0{,}00004$, $p(4) = 0{,}0000002$. Eventos interesantes para $\alpha = 0{,}01$ en la cola inferior: no hay. Cola superior: 2, 3 y 4 pues $p(2) + p(3) + p(4) = 0{,}00284 < 0{,}01$. Dos colas: 2, 3 y 4 pues $p(2) + p(3) + p(4) = 0{,}00284 < 0{,}005$

d. Probabilidades halladas por triángulo de Pascal: $p(0) = 0{,}11$, $p(1) = 0{,}26$, $p(2) = 0{,}30$, $p(3) = 0{,}20$, $p(4) = 0{,}08$, $p(5) = 0{,}0264$, $p(6) = 0{,}0055$, $p(7) = 0{,}0008$, $p(8) = 0{,}00007$, $p(9) = 0{,}0000004$, $p(10) = 0{,}0000001$. Los eventos interesantes para $\alpha = 0{,}01$ son 7,8,9,10 pues la suma de sus probabilidades es 0.00087 que es menor que 0.005.

Pág 121; No 18. Todo lo que hay que probar es que 18 entre 20 es un evento interesante. En efecto, usando la aproximación de la binomial por la normal con $\mu = 10$ y $\sigma = 2{,}23$ con significancia de 0.05 y en la cola superior, los eventos interesantes van desde 14 hasta 20, lo cual dice que sacar 18 caras entre 20 monedas lanzadas es realmente un evento muy extraño. Pero no fue esa la razón por la cual la niña se dolió en su alma, pues ella no es envidiosa, sino que ella se sintió demolida por su total incapacidad para hacer algo para mejorar la excelente marca de su amiguita. Es decir, ella sabe que el azar es incontrolable pero que en la naturaleza viene con personalidad propia: aquello que ha sido muy improbable en el pasado, será improbable en el futuro (cuando no hay razón para creer que la naturaleza de las cosas haya cambiado).

La coherencia del instinto con la teoría de probabilidades indica que el método científico tiene sus raíces en la biología. Por esa razón todo ser humano se las da de científico y es que realmente lo es y lo demuestra esgrimiendo fuertes críticas contra lo que no lo convence. Por ejemplo: nada ha podido hacer la Biología Moderna para convencer a todo el mundo de que el hombre surgió por evolución debido a la selección natural de una feliz sucesión de eventos al azar. Muy al revés, los grupos creacionistas, que promulgan la fe en que el hombre es una creación directa de Dios, han logrado hacerse muy odiosos por causa de sus acérrimos ataques contra la ciencia.

Pág 121; No 19. Se sabe por experiencia que al lanzar dos dados y contar el número de puntos que están en las caras que quedan hacia arriba, los eventos más significantes son aquellos en que las suma de los puntos de los dados sean 2 o 12, pues solo existe una combinación de números que dan este resultado, a saber el doble 1 y el doble seis. La probabilidad de que salga cada uno de estos eventos es 1/36. Por consiguiente la suma de ambos eventos es de 1/36+1/36=1/18=0.055. Esto indica que tanto las sumas de los puntos del doble 1 y el doble 6 se encuentran dentro del 6 % de la población ubicados después de 1.8 desviaciones de la media en una distribución normal, dándole gran significancia a los eventos.

Pág 121; No 20.a. 0.0179

b. 0.0548

c. 0.002

Pag 122; No 23 a). Este problema es con una cola, la inferior, pues un remedio sólo puede bajar el azúcar de una manera. Como no se dijo nada de la significancia, la tomamos como 0.05, cuya z asociada con una cola es -1.6. Esto dice que lo normal es lo que queda arriba de la media menos 1.6 veces la desviación estándar. La prueba que tenemos es:

$H_o : \mu = 100$ el producto B es un alimento. Se acepta cuando el dato experimental está arriba de 100 -1.6(10) = 100-16= 84.

$H_a : \mu < 100$ el producto B es un remedio para bajar el azúcar.

Como el dato experimental es 78, no respalda la H_o, la cual se rechaza: el producto B sirve de remedio para bajar el azúcar, siempre y cuando se consuma en las cantidades usuales.

Pag 122; No 23 b). Este problema es con una cola, la superior, pues un remedio sólo puede subir el azúcar de una manera. Como no se dijo nada de la significancia, la tomamos como 0.05, cuya z asociada con una cola es 1.6. Esto dice que lo normal es lo que queda abajo de la media más 1.6 veces la desviación estándar. La prueba que tenemos es:

$H_o : \mu = 100$ el producto C es un alimento. Se acepta cuando el dato experimental está abajo de 100 +1.6(10) = 100+16= 116.

$H_a : \mu > 100$ el producto C es un remedio para subir el azúcar.

Como el dato experimental es 115, respalda la H_o, la cual se acepta: el producto C sirve de alimento y no tiene méritos suficientes para ser considerado como remedio para subir el azúcar. Es decir, uno puede decir que es un remedio pero nadie lo compraría pues alimentos tan efectivos para subir el azucar como ese se encuentran por doquier.

Pag 122; No 23 c). Este problema es con dos colas, pues un remedio puede servir para bajar o para subir el azúcar. Como no se dijo nada de la significancia, la tomamos como 0.05 lo cual dice que lo normal es lo que está a dos desviaciones al lado y lado de la media. Por tanto, lo normal es lo que está entre 80.4 y 119.6. La prueba que tenemos es:

$H_o : \mu = 100$ el producto A es un alimento.

$H_a : \mu \neq 100$ el producto A no es un alimento.

Como el dato experimental es 93, es normal y la H_o se acepta: el producto A es un alimento.

Pág 122; No 24.a. Cola superior: $z(1 - 0{,}02) = 2$ Intervalo de aceptación $(-\infty, 6.4)$ como $x = 10$ decimos que la hipótesis nula se rechaza porque no entra dentro del intervalo.

Pág 122; No 24.b. Dos colas: $z(0{,}01) = \pm 2{,}3$ Región de aceptación = (-0.91,6.91), Ho se rechaza.

Pág 122; No 24.c. Cola superior: z(1-0.01)=2.3 RA = $(-\infty, 2.91)$, Ho: se rechaza.

Pág 122; No 24.d. Dos colas: $z(0{,}005) = \pm 2{,}6$, RA= (-5.42,3.42), Ho: se rechaza.

Pág 122; No 24.e. Dos colas: $z(0{,}05/2) = \pm 2$, RA = (0.6,7.4) Ho: se acepta.

Pág 122; No 24.f. Cola superior: $z(1 - 0{,}022) = 2$ $RA = (-\infty$, -4.6) Ho: se rechaza.

Pág 122; No 24.g. Dos colas: $z(0{,}022/2) = \pm 2{,}3$, RA = (-11.91, -4.09) Ho: se rechaza.

Pág 122; No 24.h. Cola inferior: $z(0{,}00001) = -4$ RA = $(-1.8, \infty)$. Ho: se acepta.

Pág 122; No 24.i. Dos colas: $z(0{,}00001/2) = -4$, RA(-2.31, 12.31) Ho:se acepta.

Pág 123; No 25. Problema sobre la fiesta de 50 personas. Solución:

1. En la comunidad grande, el universo, $p(\{Mujer\}) = 6/11$, pues hay 6 mujeres por cada 11 personas. Del universo tomamos una muestra al azar de 50 personas, de las cuales 35 son mujeres. Trabajamos con significancia igual a 0.05 y con dos colas, pues lo único importante es si la proporción de mujeres en la fiesta es igual o diferente de la proporción de mujeres en el universo. El desbalance puede ser de dos modos, por exceso o por defecto: tenemos dos colas.

2. Formulación del protocolo de decisión:

   H_o: $p = 6/11$.

   $H_a : p \neq 6/11$.

   Se rechaza la H_o con una significancia del 0.05.

   La H_o dice que la proporción de mujeres en la fiesta debe ser aproximadamente igual a la proporción de mujeres en el universo, excepto por cosas del azar. La H_a dice que no es cierto que la proporción de mujeres en la fiesta sea aproximadamente igual a la proporción de mujeres en el universo, sino que hay muchas razones sistemáticas que hacen que dicha proporción aumente o disminuya y que genéricamente se llama guerra de géneros.

3. Cálculo de de la H_o. Estamos con una binomial $X \sim Bi(50, 6/11)$ que cuenta mujeres y que puede aproximarse por una normal cuya media y desviación son:

   $\mu = np = 50(6/11) = 27{,}27$

   $\sigma = \sqrt{np(1-p)} = \sqrt{50(6/11)(5/11)} = 3{,}52$

   Tomamos la significancia igual a 0,05 con $z = \pm 1{,}96$. Lo normal sería todo lo que estuviese a 1.96 desviaciones al lado y lado de la media que es 27.27. Como $2\sigma = 6{,}89$, deducimos que el número de mujeres debería encontrarse entre 20.38 y 34.16 para que todo fuese normal. Decimos que la **región de aceptación de la** H_o es el intervalo (20.38, 34.16). Similarmente, la **región de rechazo de la** H_o es lo que queda antes de 20.38 unido a lo que queda después de 34.16.

4. Contraste de los datos experimentales o de muestreo con los permitidos por la H_o: hay 35 mujeres de 50, lo cual queda en la región de rechazo.

5. Veredicto: Rechazamos la H_o con $\alpha = 0{,}05$: la discrepancia entre lo predicho por la H_o, que las mujeres deberían ser un poco más que la mitad, y lo observado, que hay 35 mujeres de 50 personas, no se explica cómodamente por el azar, sino que se explica mejor diciendo que la H_o es falsa y que hay razones sistemáticas que hacen que las mujeres prefieran las fiestas más que los hombres, como que a las mujeres les gusta bailar en tanto que los hombres prefieren tomar trago con otros hombres en un bar para poder hablar de fútbol, carros y top-models.

Pág 123; No 26. $X \sim Bi(20, 0.3) \sim N(6, \sqrt{6(0{,}7)} = \sqrt{4{,}2} = 2{,}05)$.
$H_o : \mu = 6$, $H_a : \mu > 6$. $\alpha = 0{,}05$, cola superior, $z = 1{,}6$.
RA: $(-\infty, 6 + (1.6)(2.05)) = (-\infty, 6 + 3.27) = (-\infty, 9.27)$, $x = 8$ es normal. El regaño no sirvió.

Pág 123; No 27. X cuenta el número de estudiantes que sufren. $X \sim Bi(35, 0.4) \sim N(14, 2.898)$.
$H_o : p = 0{,}4$, $H_a : p < 0{,}4$, RA: $(9.36, \infty)$, $RR : (-\infty, 9.36)$. Como 12 pertenece a RA, aceptamos la H_o: la diferencia entre lo esperado según la H_o, que los estudiantes que sufren deben ser alrededor de 14, y lo observado, 12 estudiantes que sufren, es mínima y se explica cómodamente por azar. El profesor se jacta en vano, sin fundamento. Para defenderse, el puede reunir más datos o mejorar su metodología.

Pág 123; No 29. $p = 3/4$ sirve porque si
$X \sim Bi(40, 3/4) \sim N(30, \sqrt{30/4} = \sqrt{7{,}5} = 2{,}73)$ entonces $x = 28$ está a menos de una desvación de la media, 30, y por lo tanto es normal.

Pág 123; No 30. La teoría que goza de mayor credibilidad será aquella que asigne una mayor probabilidad al dato observado. De acuerdo a la primera teoría tenemos: $X \sim Bi(25, 0.6) \sim N(15, \sqrt{15(0{,}4)} = \sqrt{6} = 2{,}45)$ entonces $z(18) = \frac{18-15}{2{,}45} = \frac{3}{2{,}45} = 1{,}22$. De acuerdo a la segunda teoría, $X \sim Bi(25, 0.8) \sim N(20, \sqrt{20(0{,}2)} = \sqrt{4} = 2.)$ entonces $z(18) = \frac{18-20}{2.} = -1$. Por consiguiente, lo observado está a una

desviación de la media en la segunda teoría pero está más lejos en la primera. Por lo tanto, la segunda teoría tendrá más acogida.

Pág 123; No 31. Todos los intervalos de confianza son con dos colas, la significancia se divide entre dos.

a) $\alpha = 0{,}02$, $z = \pm 2{,}3$, IC = (12-(2.3)(2), 12 + (2.3)(2)) = (7.4, 16.6).

b) IC = (5.6,12.4).

c) IC = (17.6,64.6).

d) $\alpha = 0{,}06$, $z = \pm 1{,}9$, IC = (-17-(1.9)(1.2), -17+(1.9)(1.2))
= (-19.28,-14.72).

e) IC = (0.00087, 0.00113).

Pág 124; No 32. a) $\alpha = 0{,}02$, $n = 35$, $c = 12$, $z = \pm 2{,}3$, IC = (0.16, 0.52).

d) $\alpha = 0{,}05$, $z = \pm 2$, IC = (0.48,0.70).

e) $\alpha = 0{,}18$, $z = \pm 1{,}3$, IC = (0.52,0.60).

Pág 124; No 33. a) La media y la desviación de $\mu_{\bar{X}_1}$ son las mismas que las de X: Como el promedio de un único dato es el mismo dato, $\bar{X}_1 \sim X$, lo cual dice que la media muestral de 1 elemento se distribuye exactamente como la población misma. b) La media de $\mu_{\bar{X}_4}$ es la misma que la de X pero la desviación es 1. c) La media de $\mu_{\bar{X}_9}$ es la misma que la de X pero la desviación es 2/3. d) La media de $\mu_{\bar{X}_{16}}$ es la misma que la de X pero la desviación es 1/2. e) La media de $\mu_{\bar{X}_{25}}$ es la misma que la de X pero la desviación es 0.4. f) La media de $\mu_{\bar{X}_{36}}$ es la misma que la de X pero la desviación es 1/3. g) La media de $\mu_{\bar{X}_{49}}$ es la misma que la de X pero la desviación es 2/7.

Pág 124; No 34. a) $n = 8$, $X \sim N(5, 2)$; $\mu_{\bar{X}_8} = 5$; $\sigma_{\bar{X}_8} = 2/\sqrt{8} = 0{,}71$.

e) $\mu_{\bar{X}_{35}} = -20$; $\sigma_{\bar{X}_{35}} = 3/\sqrt{35} = 0{,}507$.

Pág 124; No 35. a) $\mu_{\bar{X}_{22}} = 8$; $\sigma_{\bar{X}_{22}} = 3/\sqrt{22} = 0{,}64$. $z(9) = \frac{9-8}{0{,}64} = 1{,}56$; $z(7) = \frac{7-8}{0{,}64} = -1{,}56$; Prob = $\Psi(1{,}56) - \Psi(-1{,}56) = 0{,}94 - 0{,}05 = 0{,}89$.

d) Abajo de 7 : $\Psi(z(7)) - \Psi(z(-\infty)) = 0{,}0668$.

f) El problema divide la recta real en tres segmentos: $(-\infty, 4)$, (4, 12) y $(12, \infty)$. Se toma un punto en cada uno, por ejemplo, 2, 5 y 15, y se chequea para ver cuáles están más cerca o más lejos de la media que 4 unidades. Hay dos regiones cuya distancia a la media es 4 o más: $(-\infty, 4)$, con probabilidad 0^+, y $(12, \infty)$, con probabilidad 0^+. Total 0^+.

Pág 124; No 36 a) Lo que pasa habitualmente es la H_o: que la media vale 5. Con significancia 0.05 la RA de la H_o con dos colas es (4.51,5.49). El dato experimental es que la media muestral de 16 datos es 8. Es muy anormal que le vaya tan extremadamente bien: la H_o se rechaza con dos colas, pues lo único que se pregunta es si eso es normal o anormal.

Pág 124; No 36 b) El análisis se hace con una cola: si le va igual que de costumbre, no hay nada de raro. Y si le llega a ir mejor que de costumbre, don Pedro pensará: eso no tiene nada de raro, a veces a uno le va bien y a veces le va mal. Pensará eso pues él está esperando el momento en que la vida vengará la injusticia que el cometió contra la esposa y eso no será haciéndolo ganar más que de costumbre. Eso quiere decir que las hipótesis son:

$H_o : \mu = 5$: me irá en promedio igual que de costumbre. Irle a don Pedro como de costumbre o mejor que de costumbre es algo que se atribuye al mismo mecanismo que genera toda la distribución. Más resumido: irle como de costumbre o mejor que de costumbre es algo que se puede explicar por puro azar, por las cosas incontroladas distintas a su relación con la esposa.

$H_a : \mu < 5$: me irá peor que de costumbre pues le pegué a mi esposa que siempre me ha sido fiel y yo no. Irle requetemal es algo que ya no se explica por azar sino que queda mejor explicado diciendo que es la justicia divina sobre su mala conducta, lo cual ya no es el mecanismo usual que genera la distribución.

Veredicto: La H_o se rechaza pues su RA con significancia 0.05 y cola inferior es (4.39, ∞) y 2 no pertenece a dicha región. La semana fue pésima, muy fuera de lo usual. Eso es muy simple de explicar pues pasan dos cosas: cuando Don Pedro le pega a su esposa, el se vuelve alérgico a los ruidos de su carro y cualquier ruidito le demuestra claramente que se necesita urgentemente un repuesto y, obviamente, se gasta todo haciendo un mantenimiento innecesario a su carro y entre tanto deja de trabajar. Por otra parte, le da miedo recoger jovencitas bonitas pues son las que drogan a los taxistas y tampoco recoge jóvenes sin corbata pues le da miedo que lo atraquen. Sigue de largo, perdiendo clientes al por mayor.

Pág 125; No 37. Veredicto con una cola, pues el estudiante tenía obligación de mejorar, que es diferente de demostrar que había sufrido algún cambio. Con una cola y $\alpha = 0{,}05$ el z crítico es 1,6. La desviación de X es 0.3 pero la de $\bar{X}_6$ es 0.12. Esto significa que la variabilidad de las notas es alta, que a veces la bien, mejor que 4 y a veces mal, pierde la materia. Pero el promedio por semestre de 6 materias es mucho más estable, pues lo malo se compensa con lo bueno: la desviación es pequeña y los promedios están pegados a 3.5. Esto implica que levantar la nota de una materia, en los parciales, puede ser realizable pero en cambio tratar de levantar un promedio puede generar un esfuerzo heróico y no siempre se realiza. Tenemos para $\bar{X}_6$ un valor crítico de 3.69. Concluimos que el papá fue justo al no darle el viaje a su hijo.

Pág 125; No 38. $X \sim N(4.1, 0.4)$, $\bar{X}_3 = 4{,}2$, $\mu_{\bar{X}_3} = 4{,}1$; $\sigma_{\bar{X}_3} = 0.4/\sqrt{3} = 0{,}23$. $H_o : \mu = 4{,}1$, $H_a : \mu > 4{,}1$. $\alpha = 0{,}05$, $z = 1{,}6$, cola superior. $RA : (-\infty,\ 4.1 + (1.6)(0.23)) = (-\infty,\ 4.468)$. $x = 4{,}2$ es normal. Veredicto: se acepta la H_o. El jefe mejoró gracias al azar.

Pág 125; No 39. a) $\mu_{\bar{X}_7} = 1$; $\sigma_{\bar{X}_7} = 3/\sqrt{7} = 1{,}13$. $H_o : \mu = 1$, $H_a : \mu > 1$. $\alpha = 0{,}05$, $z = 1{,}6$, cola superior. $RA : (-\infty,\ 1 + (1.6)(1.13)) = (-\infty,\ 2.8)$. $\bar{X}_7 = 2$ es normal. La H_o se acepta. Telenovela: un estudiante a veces anda mal de plata y a veces bien. Según él, su promedio genérico al final de la semana es de 1 con una desviación de 3, lo que indica que a veces se endeuda. Pero debido a que se decidió a dictar unas clases de estadística, sus ingresos durante las últimas 7 semanas han tenido un promedio de dos. El cree que ha mejorado por las clases, pero como su contabilidad es muy informal, no puede asegurarlo y por ahora la mejor explicación de su bonanza es el azar. Le convendría mejorar en la clase de estadística para que se pueda conseguir más alumnos.

e) $\mu_{\bar{X}_{36}} = 6{,}9$; $\sigma_{\bar{X}_{36}} = 6/\sqrt{36} = 1$. $H_o : \mu = 6{,}9$, $H_a : \mu \neq 6{,}9$. $\alpha = 0{,}05$, $z = \pm 2$, dos colas. RA: (6.9-(2)(1), 6.9+(2)(1)) = (4.9,8.9). $\bar{X}_{36} = 5{,}8$ es normal. La H_o se acepta. Telenovela: El sueldo de Juan, vendedor de lotería, ha sido en promedio 6.9. Pero desde que su esposa se dio cuenta de que le era infiel, hace 36 meses, el sueldo le bajó en promedio a 5.8. Le dijeron que el bajón se debe al azar y no a sus infidelidades y las peleas con su esposa, a lo cual el respondió: ¿Y por qué la conciencia me dice que es por mi culpa?

Pág 125; No 40. a) $\sigma_{\bar{X}_7} = 3/\sqrt{7} = 1{,}13$. $\alpha = 0{,}05$, $z = \pm 1{,}96$, dos colas.
IC= (2-(1.96)(1.13), 2 + (1.96)(1.13)) = (-0.21, 4.21).

e) $\sigma_{\bar{X}_{36}} = 6/\sqrt{36} = 1$. $\alpha = 0{,}05$, $z = \pm 1{,}96$, dos colas.
IC = (5.8-(1.96)(1), 5.8+(1.96)(1)) = (3.84,7.76).

Pág 125; No 42. Las ventas en toda la cadena fueron de 14,4. Con $\alpha = 0{,}05$, la aleta da 4.7. Pero el promedio global apenas se desvio 0.4 arriba de la media. Por tanto, poner el premio no sirvió de nada a la compañía. No se excluye sin embargo que a algunas tiendas les haya ido muy mal y a otras muy bien. Por eso, es posible que varias tiendas hayan ganado premio a costa de la empresa. La conclusión es que la implementación de la motivación estuvo mal diseñada. Alternativa: dar un premio si el total de ventas de toda la cadena mejora y dar premio a quienes superen un tope dado.

Pág 125; No 43. a) Si $X \sim N(\mu, \sigma)$ y si se tiene un único dato x_o, al repetir el procedimiento constructivo se concluye que

$x_o - z\sigma < \mu < x_o + z\sigma$

donde z es el z asociado a la significancia tomada y con dos colas.

b) Si $X \sim Bi(\mu, \sigma)$, X puede aproximarse por una normal con media $\mu = np$ y desviación $\sigma = \sqrt{np(1-p}$. Como el número observado de caras fue k entre n, podemos reemplazar todo para obtener:

$k - z\sqrt{np(1-p} < np < k + z\sqrt{np(1-p)}$

Dividiendo por n en todos lados, obtenemos:

$\frac{k}{n} - z\frac{\sqrt{np(1-p)}}{n} < p < \frac{k}{n} + z\frac{\sqrt{np(1-p)}}{n}$

Entremos a n dentro de la raíz:

$\frac{k}{n} - z\sqrt{\frac{np(1-p)}{n^2}} < p < \frac{k}{n} + z\sqrt{\frac{np(1-p)}{n^2}}$

Simplifiquemos

$\frac{k}{n} - z\sqrt{\frac{p(1-p)}{n}} < p < \frac{k}{n} + z\sqrt{\frac{p(1-p)}{n}}$

Ahora recordemos que una probabilidad es la idealización de una frecuencia relativa. En nuestro caso, la frecuencia relativa es $f = k/n$, lo cual nos permite inferir que el IC para p es

$f - z\sqrt{\frac{f(1-f)}{n}} < p < f + z\sqrt{\frac{f(1-f)}{n}}$

c) Comentarios. Primero: ambos métodos exigen aproximar una binomial por una normal, lo cual puede ser injusto si p no es 0.5. Segundo: ambos métodos asumen la aproximación rígida, es decir que

la desviación $\sqrt{np(1-p)}$ verdadera puede aproximarse por $\sqrt{f(1-f)}$. Tercero: ser consciente de las aproximaciones nos permite pensar que debe poderse descubrir estimaciones más precisas. Cuarto: el método construtivo es transparente y claro, pero en el presente método uno tiene que preguntarse: ¿Por qué en la última ecuación se reemplaza p por f sólo en los lados y no en el centro? Quinto: aunque el método constructivo es mucho más tedioso, nos permite descubrir que el IC contiene todos los valores del parámetro estimado que son compatibles con los datos y con la significancia dada α. Sabemos además que el IC es la respuesta a la pregunta: ¿Con los datos dados, cuánto vale el parámetro? A lo cual se responde: no sabemos con exactitud cuánto vale, pero si Usted apuesta a que el parámetro está en el IC, Usted ganará la apuesta el (1- α) por ciento de veces. Sexto: El objetivo del presente tema es el de ilustrar de qué manera la ciencia es una artesanía. Por eso, hemos preferido el método constructivo.

Pág 126; No 44. El problema que se ataca con el IC de la media es que la media de una muestra no es igual a la media de la población pero es información valiosa para estimar el rango donde la media poblacional pueda estar. Acá hay una única media muestral. Pero en el TLC, cuando se dice que la media de la media muestral es la misma media poblacional, lo que se dice es que el promedio de un número muy grande de medias muestrales de muestras tomadas al azar y con reemplazo, para que las probabilidades no cambien, es la media poblacional. Vemos que no se habla de lo mismo y no hay contradicción.

Pág 126; No 45. El IC de la media puede ser cortico o largo y su longitud denota una incertidumbre: entre más largo, mayor es la incertidumbre. Dicha longitud depende a su vez de la longitud de la aleta que es zeta veces la desviación sobre la raíz cuadrada del número de datos, n. Vemos que n aparece en el denominador, por eso, si n crece, la aleta se acorta y por tanto el IC también. El resultado final es que cuando el número de datos crece, la incertidumbre decrece. Por tanto, para uno mejorar la precisión, uno aumenta el tamaño de la muestra, pero eso cuesta tiempo y dinero y el costo no crece proporcionalmente a n sino al cuadrado de n.

Parcial II v.1 Pág 127. No 1) No. de estudiantes:
P: 4; I: 159; A: 75; H: 11.15;
No 2.) $z(Juan) = z(186) = 0{,}6$ pero $z(Juanito) = z(97) = 2$. Juanito es relativamente más alto.
No 3.a) $z(140) = -3{,}13$ y su tridente es 0.001. Eso significa que la tensión del abuelito es por costumbre demasiado alta cubriendo el 99.9 % del tiempo.

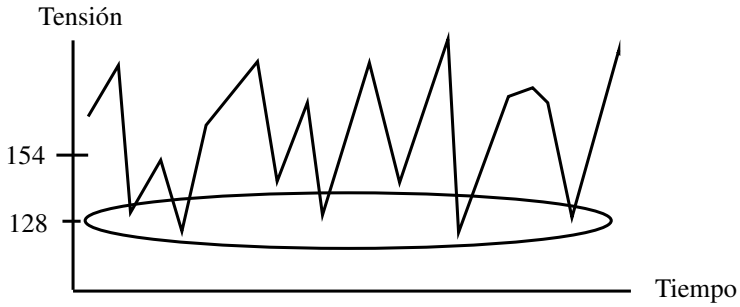


Figura 7.4. La tensión del abuelito vive subiendo y bajando alocadamente debido al azar, es decir, a factores no controlados y propios de la fisiología normal del abuelito. El óvalo enmarca un conjunto de 5 eventos que indican que uno puede tener por azar 5 eventos que sean bajones en la tensión del abuelito y sin contar para nada con el tratamiento. Vemos que la objeción de los tíos es en principio correcta pero su plausibilidad disminuye con el número de mediciones. En efecto, la probabilidad de que 5 eventos no observdos den un promedio inferior a 147.2 es 0.06 pero la probabilidad de tener un promedio de 128 en 5 mediciones es muchisísimo menor pues las mediciones deberían coincidir con los bajones. Ni siquiera un médico brujo se le mediría a lograr tal efecto conjunto.

b) Informe: $H_o : \mu = 154$, σ= 4.47, se rechaza con la cola inferior. RA de la $H_o = (147.2, \infty)$. Valor observado: $\bar{X}_5 = 128$. Veredicto: se rechaza la H_o. El tratamiento sí sirvió. Páguenme. c) Los tíos alegan: 128 fue clasificado como extremo con una significancia de 0.06. Pero la significancia es arbitraria y uno puede tomar la que uno quiera y la nuestra es realmente pequeña. En otras palabras: es perfectamente factible que el médico diga la verdad y que el promedio de 5 mediciones le dió efectivamente 128. Lo

que no creemos es que la causa haya sido el tratamiento sino que fue una casualidad a favor del médico. Eso se debe a que la tensión del abuelito no es fija sino que vive subiendo y bajando lo cual se especifica en que la desviación no es cero sino 10 y las subidas y bajadas se deben a factores no controlados y que son normales en la fisiología del abuelito, o como se dice, al azar. d) Concepto experto: El médico tiene mucha razón: un promedio de 128 en 5 mediciones es un valor extremadamente bajo. Pero por otra parte, no se excluye la posibilidad de que efectivamente el 128 haya sido una feliz coincidencia que nada tenga que ver con el tratamiento. Como ambas partes tienen razón, hay un conflicto bien fundamentado que hay que resolver.

Concepto experto: reunir más datos difícilmente resolvería el conflicto pues nunca habrían datos suficientes para convencer a un escéptico recalcitrante de que un tratamiento funciona. Sin embargo, la experiencia demuestra que la humanidad ha podido progresar científica y técnicamente tomando la directiva que los tratamientos funcionan cuando los resultados son clasificados como extremos con una significancia de alrededor del 5 %. En este caso, 128 da una $z = 2{,}6$ con una significancia de $0.0047 = 0.005$, lo cual quiere decir que mil ensayos sin tratamiento darían 5 resultados positivos. Eso implica que para repetir por azar el resultado del médico se requerirían hacer unos 200 ensayos. Naturalmente que es más barato pagarle. En definitiva: páguenle al médico.

Parcial II v.2 Pág 127. No 1) Prendas a producir:
S: 16.16; M: 128.98; L: 38.7; XL: 16.16;
No 2) α=0.06/2=0.03; 2 colas; IC=[-7.18,1.18]
No 3a 1^-.
No 3b) $H_o : \mu = 4$, α=0.06; cola superior; RR = $(4.25, \infty)$. Se rechaza la Ho: tenemos razones para creer que el aumento en las ventas no fue por azar, por factores no controlados, sino por la buena administración. Que se le renueve el contrato.
No 3c) La gráfica es casi la misma que la del problema 3 de la versión 1: el azar puede lograr cualquier cosa, pero el problema es que podría ser necesario un número demasaido grande de repeticiones para lograr eventos extremos y a favor. Como cualquier evento extremo, más alla del valor crítico, nos induce a una decisión de renovar el contrato y dicha decisión puede bien ser errónea tipo alfa, la probabilidad de un error alfa es la significancia tomada, en este caso 0.06.
No 3d) Apoyamos la decisión de renovarle el contrato porque es más confiable este administrador que quizá fue apoyado por el azar pero aquí mismo y con una probabilidad muchísimo menor que 0.06 y no otro del cual sólo se sepan buenas referencias.
No 3e) Asumiendo que las leyes del mercado se han mantenido constantes, la decisión fue claramente un error tipo alfa.
Parcial II v.3 Pág 128. No 1) IC = (89.15,92.85).
Parcial II v.3 Pág 128. No 2) La conducta ejemplar es la que sirve de referencia: $H_o : \mu = 2$. $H_a : \mu < 2$. Se rechaza con una sola cola, la inferior, pues si alguien llega con muy poca gasolina está poniendo en peligro su vida, su nave, el prestigio de la Fuerza Aérea y el Presupuesto Nacional. Pero si alguien llega con gasolina de sobra no comete falta alguna. Conducta particular: $\bar{X}_5 = 0{,}5$. Para juzgar dicha conducta tenemos: $\sigma_{\bar{X}_5} = 0{,}4/\sqrt{5} = 0{,}1788$. Como no dieron la significancia, se toma 0.05, lo cual da una z= 1.6, pues es con una sola cola. La H_o se acepta más arriba de 2 - (1.6)(0.1788) = 1.71 y se rechaza si está más abajo. Medio galón, está demasiado abajo. Es algo muy extremo por lo demasiado poquito. La amonestación es necesaria: no es que se dude de lo que un as puede hacer sino del tremendo mal ejemplo que es no reproducible por un piloto ordinario.
Parcial II v.4 Pág 128. No 2) a) Las frecuencias absolutas son: 2, 5, 8, 10, 3, 2. La media vale 36.167 y la desviación 8.972. b) Como son 30 datos se aplica el TLC. La nueva sigma es $8.972/\sqrt{30} = 1.638$. CI del 95 % (32.9718, 39.3622). Según lo que sabemos hasta ahora, para hallar el IC se necesita la desviación de la población, pero nosotros tenemos sólo la de una muestra. Por eso asumimos que las dos coinciden, lo cual no es trágico pues tenemos 30 datos que son bastantes. ¿Qué es la muestra y qué la población? Cada mes aparece una cartelera con la clasificación: eso es una muestra. La población reune los datos de las carteleras de todos los meses. c) La novia ve una población con media 36.167 y desviación 8.972. No se aplica el TLC porque a la novia sólo le importa Pedro que es una única persona. El x_c de la cola de arriba al 5 % de significancia es 36.167 + 1.643 (8.972) = 50.90= puntaje mínimo que inspira confianza al presentarse ante la novia.
Parcial II v.4 Pág 128. No. 3 a) Con una significancia del 5 % y cola inferior, la medicina se

declara inhabilitada: noy hay datos suficientes para respaldar la idea de que Olivo hay mejorado. b) La medicina sí sirvió. c) Aunque Olivo siempre se esforzó desde que decidió cambiar, un único mes no es garantía de que al próximo no aumente su agresividad, digamos a 14 agresiones, pero 7 meses de baja agresividad en promedio ya son suficientes para que uno confíe que Olivo sea menos agresivo en el futuro que sus acostumbradas 8.1 agresiones en promedio por mes.

Parcial II v.5 Pág 129. No 1) f= 0,875. σ=0,082 α=0.08/2=0.04; 2 colas; IC=[0.731,1.019]. Como las probabilidades están entre 0 y 1, la última respuesta es IC = [0.731, 1].

2) [0, 0.21] (se recorta por abajo pues las probabilidades no pueden ser negativas.)

3) Si se cuentan los de B, la RA de la H_o es $(-\infty, 46.52)$ y la de aceptación es $[46.52, \infty)$. Como hay 50 de B, eso es normal: no hay datos suficientes para alegar que la democracia es ficticia.

Parcial II v.6 Pág 129. No 1) $p(0) = 0{,}78$; $p(1) = 0{,}20$; $p(2) = 0{,}02$; $p(3) = 0{,}0005$; Eventos interesantes: 3 lo es pues $p(3) = 0{,}0005 < 0{,}04$; 2 lo es pues $p(3) + p(2) = 0{,}0005 + 0{,}02 = 0{,}02 < 0{,}04$; 1 no lo es pues $p(3) + p(2) + p(1) = 0{,}0005 + 0{,}02 + 0{,}20 = 0{,}22 > 0{,}04$;

No 3) Las mujeres sí están sobrerepresentadas. Se espera que como máximo haya 62.24 mujeres y hay 70.

Capítulo 8

Experimentos y estadígrafos de contraste

177 ***Objetivo: formular el esquema oficial del método científico: la ciencia es básicamente un contraste entre lo que se ve y lo que se cree.***

178 ***Introducción***

Restringiéndonos a distribuciones normales y binomiales, hemos aprendido a resolver problemas con nuestras propias manos para poder tener una experiencia vivencial de lo que significa que la ciencia moderna es una artesanía. Dicho más claramente: la construcción de la verdad en la ciencia moderna se hace por procedimientos en los cuales participan métodos matemáticos, al igual que el libre albedrío del ser humano. La responsabilidad del ser humano reside en la necesidad de tener que definir la significancia, mientras que los métodos matemáticos calculan la H_o propuesta. A partir de eso se construye la verdad como una proposición que no es ni verdadera ni falsa en sentido absoluto, sino sólo estadísticamente confiable tanto más cuanto más pequeña sea la significancia. Por eso, la ciencia es una estructura compuesta de verdades tentativas.

Detalle sociológico: la ciencia tiene muchas facetas y una de ellas es que genera empleo. Por eso nada más, la ciencia da origen a un flujo económico del cual dependen muchísimas personas: Inglaterra, por ejemplo, tiene una industria editorial de ciencia de alto nivel que se expresa en unas 5000 revistas y que se nutre del arduo trabajo de investigadores de todo el mundo que publican en Inglaterra para demostrarle a todos que han llegado a la cima. La estabilidad de los flujos económicos exige alta calidad en la ciencia y, como consecuencia, lo que en un momento se consideró alta calidad puede llegar a convertirse en una verdad absoluta, es decir en un lastre que inhibe la exploración del pensamiento alterno.

En fin, ya que hemos logrado saber, entender y vivir que la ciencia es una artesanía, tenemos de ahora en adelante la posibilidad de usar metodologías automáticas, tanto para el tipo de problemas que hemos visto como para otros de tipo totalmente diferente. En la base de dichas metodologías está lo que se denomina **estadígrafo de contraste**, que es un medidor de la discrepancia que hay entre lo predicho por la H_o y lo observado experimentalmente. Si la discrepancia es pequeña, la H_o se acepta, pero si la discrepancia es grande, la H_o se rechaza. Para saber si una discrepancia es grande o pequeña, se compara con los valores críticos dados por las tablas según la significancia dada. Asumiremos siempre que todas las variables aleatorias se distribuyen normalmente, sea por ley propia o por aproximación, como en el caso de las binomiales.

Llegamos así a nuestro esquema oficial : **el método científico es básicamente un contraste entre lo que se ve y lo que se cree.** Nuestras creencias son siempre muy sencillas, como que la media de una cierta distribución normal es 5 y que cualquier discrepancia entre lo que se cree y lo que se ve es un efecto debido puramente al azar. Esta es la H_o. Aceptamos la H_o cuando la discrepancia entre lo que se ve y lo que se cree no es mucha, pero la rechazamos si tal discrepancia se clasifica como grande. La clasificación de grande vs. pequeña se hace gracias al valor crítico del estadígrafo de contraste determinado por la significancia elegida y a veces por la representatividad del experimento, dada por el número de datos.

8.1. Estudio de la media

Con un ligero giro en la relectura de lo que ya sabemos, podemos entender claramente la idea de un estadígrafo de contraste:

179 ***Ejemplo*** *Decidamos con dos colas la H_o : $\mu = 5$ dado que un dato experimental dio 23, sabiendo que $\sigma = 4$.*

Solución a la antigua: La z asociada a la significancia por defecto 0.05 y con dos colas es 1.96. Por lo tanto, lo normal sobre Z es lo que está en el intervalo (-1-96,1.96).

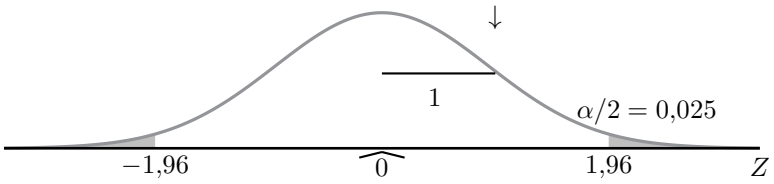


Figura 8.0. Si $Z \sim N(0,1)$, lo normal con dos colas y significancia 0.05 es lo que va de -1.96 a 1.96.

Ahora usamos la transformada-x para hallar los valores críticos sobre X:
$x = \mu + z\sigma$
Tenemos que el valor crítico izquierdo, X_-, y el derecho, X_+, son:
$X_- = 5 + (-1{,}96)(4) = -2{,}84$
$X_+ = 5 + (1{,}96)(4) = 12{,}84$.

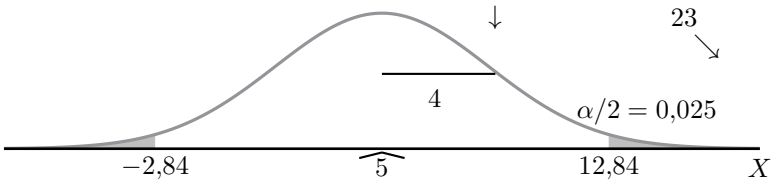


Figura 8.1. Si $X \sim N(5,4)$, lo normal es lo que va de -2.84 a 12.84. Como el dato experimental 23 está por fuera, rechazamos H_o: es incómodo creer a la luz de dicho resultado experimental que la media sea 5.

Vemos que la zona de aceptación de la H_o es el intervalo (-2.84,12.84). Como 23 no está ahí, lo clasificamos como extremo para la significancia dada y dos colas y, por tanto, rechazamos la H_o.

IMPORTANTE: Relectura por estadígrafo de contraste: Vamos a resolver el mismo problema pero de la forma como lo haremos de ahora en adelante. La idea fundamental se desprende del hecho de que todas las campanas son iguales y, por tanto, uno puede tomar las decisiones no sobre X sino directamente sobre Z, la campana estándar, la que tiene media cero y desviación uno. Para pasar de la X a la Z usamos la transformada y obtenemos z_{obs}, la **zeta-observada**:

$z_{obs} = \frac{x-\mu}{\sigma}$.

Reemplacemos:

$z_{obs} = \frac{23-5}{4} = \frac{18}{4} = 4{,}5$.

Para la significancia $\alpha = 0{,}05$ y dos colas, los valores críticos de zeta son $z_\pm = \pm\ 1.96$. lo normal en Z va desde -1.96 hasta 1.96. Como 4.5 está por fuera, decidimos que el z_{obs} es extremo, es anormal. Eso es suficiente para rechazar la H_o: la media no es cero, sobre Z, o lo que es lo mismo, la media no es 5, sobre X.

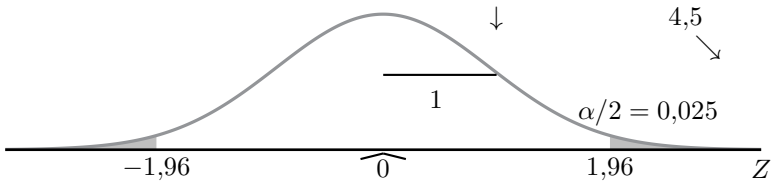


Figura 8.2. Si $Z \sim N(0,1)$, lo normal con dos colas y significancia 0.05 es lo que va de -1.96 a 1.96. Como la z experimental 4.5 está por fuera, rechazamos H_o: es incómodo creer a la luz de dicho resultado experimental que la media de la Z sea 0 o que la media de la X sea 5.

Releamos lo hecho de la forma con que lo haremos de ahora en adelante. Observemos que

$z_{obs} = \frac{23-5}{4} = \frac{18}{4} = 4{,}5$

puede interpretarse como sigue: se cree que el valor de la media es 5. Ponemos a prueba nuestra creencia haciendo un experimento para medir el valor real de la media: dio 23. Por tanto, si uno no cuestiona el valor de la desviación, podemos decir que z **mide en unidades apropiadas la discrepancia entre lo que se ve, 23, y lo que se cree, que la media vale 5**. La discrepancia medida entre lo que se ve y lo que se cree vale 4.5.

¿Es esta discrepancia grande o pequeña? Para saberlo, la comparamos con la z crítica asociada a la significancia por defecto 0.05, que es $1.96 = z_{0{,}025}$. Como 4.5 es más grande que 1.96, decidimos que la discrepancia es muy grande y rechazamos la H_o: no creemos que la media valga 5.

Lo fascinante es que toda la ciencia se puede plantear siguiendo el mismo protocolo: **el método científico es un contraste entre lo que se ve y lo que se cree. Ese contraste se mide con un estadígrafo de contraste. Si la *discrepancia observada* entre lo que se ve y lo que se cree es pequeña, ésta se explica por azar, por la variabilidad innata de los datos debida a factores no controlados, y se acepta lo que se cree, la H_o. Pero si la discrepancia es grande, es aconsejable rechazar lo que se creía y sugerir otra explicación a lo observado. La discrepancia observada es grande si es mayor que la discrepancia crítica sacada de la tabla correspondiente al estadígrafo usado. Es pequeña si es menor que el valor crítico. El primer ejemplo de estadígrafo de contraste es la z, pero cada esquema experimental genera sus propios estadígrafos.**

180 ***Ejemplo*** *Decidamos con dos colas la $H_o : \mu = 5$ dado que una muestra de 16 eventos dio una media de 8 sabiendo que $\sigma = 4$ (asumimos que estamos en una distribución normal).*

Solución: Sea X la v.a. de la que se habla. La media de la muestra de 16 eventos fue $\bar{X}_{16} = 8$. La desviación de X fue 4: $\sigma_X = 4$. Por el TLC, la v.a. media muestral también se distribuye normalmente, tiene media $\mu_{\bar{X}_{16}} = 5$ y desviación
$\sigma_{\bar{X}_{16}} = \frac{4}{\sqrt{16}} = 1$.

La transformada-z asociada a la media muestral observada $\bar{X}_{16} = 8$ es
$z_{obs} = \frac{\bar{X}_{16} - \mu_{\bar{X}_{16}}}{\sigma_{\bar{X}_{16}}} = \frac{8-5}{1} = 3$
Observemos que en el numerador aparece $\bar{X}_{16} = 8$, y por lo tanto en el denominador debe estar la desviación de la media muestral de $\bar{X}_{16}$, la cual es 1. No se puede mezclar un numerador en X con un denominador en otra variable Z o $\bar{X}_{16}$: todo debe ser homogéneo. z_{obs} mide, en unidades apropiadas, la discrepancia entre el dato experimental, 8, y lo presupuesto en la H_o, que la media vale 5.

Como no dijeron nada de la significancia, tomamos $\alpha = 0{,}05$ o sea dos desviaciones estándar. La región de aceptación sobre la Z es de -1.96 a 1.96. Como el z_{obs} no está en esa región, decidimos que la H_o se rechaza. Los datos observados nos dan pie para rechazar la hipótesis de que $\mu_X = 5$.

181 ***Notación*** *La z crítica con dos colas y con una significancia dada alfa, α, o una confianza de $1-\alpha$ se denota $z_{\alpha/2}$, en la cola inferior, y $z_{1-\alpha/2}$ en la cola superior. Si es con una cola, las notaciones correspondientes son z_α, en la cola inferior, y $z_{1-\alpha}$ en la cola superior. Esto mismo se utiliza para todos los demás estadígrafos. Es importante memorizar que esas notaciones pueden variar dependiendo de la forma como se hayan codificado las tablas, por lo que lo primero que hay que mirar en una tabla es el código de interpretación.*

182 ***Ejemplo*** *Rehagamos el problema de las turistas, ej. 165, pág. 110, que calzan 40 en promedio, mientras que la talla local, X, es una variable aleatoria con distribución normal tal que $\mu_X = 36$, $\sigma_X = 2$. ¿Es justo que regañen al Administrador por no tener zapatos para ellas?*

Solución: Medimos con una transformada-z la discrepancia entre lo hallado experimentalmente y lo esperado según la población local:

$$z_{obs} = \frac{\bar{X}_{30}-\mu_{\bar{X}_{30}}}{\sigma_x/\sqrt{n}} = \frac{\bar{X}_{30}-\mu_x}{\sigma_x/\sqrt{n}} = \frac{40-36}{0{,}37} = 11$$

El z-crítico es 1.96, pues tomamos la significancia como 0.05. Vemos que el z-observado es muchísimo más grande. De ello se deduce que las turistas generaron un evento extremadamente raro.

Veredicto: La discrepancia es demasiado grande: el administrador del almacén no tenía por qué haber previsto que harían falta zapatos extra-grandes. Con todo, la vendedora furtivamente cambió de lugar un zapato de hombre estilo italiano y eso fue suficiente para generar una venta estupenda.

183 ***Ejemplo*** *Se sabe que en el segundo parcial de Estadística, las notas se distribuyen normalmente y que $\mu = 3{,}2$ con $\sigma = 0{,}6$. En un curso de 28 estudiantes, el promedio fue 3.6.*

a) Formulemos H_o

b) Contrastemos H_o con el curso en observación.

c) Tomemos un veredicto con dos colas y digamos su justificación.

d) Tomemos un veredicto con una cola y digamos su justificación.

Solución: Realmente estamos afrontando la acusación generada por el hecho de que la notas del curso estuvieron muy altas con respecto a lo usual, ¿será que el nivel académico de la clase es mayor que el tradicional de la Universidad?

H_o : $\mu = 3{,}2$ con $\sigma = 0{,}6$. Es decir, cuando de una población con $\mu = 3{,}2$ y $\sigma = 0{,}6$ se saca una muestra de 28 individuos, la media muestral puede dar por mero azar cualquier valor, inclusive 3.6.

H_a : $\mu \neq 3{,}2$: la media experimental no proviene de una población con $\mu = 3{,}2$ y $\sigma = 0{,}6$ sino que viene de otra población, por ejemplo de una clase con un nivel académico menor.

Para dilucidar el interrogante, medimos la discrepancia normalizada según el TLC entre lo esperado según la H_o y lo observado:

$$z_{obs} = \frac{3{,}6-3{,}2}{0{,}6/\sqrt{28}} = 3.52$$

El z crítico con $\alpha = 0{,}05$ es 1.96, el cual es mucho menor que el z observado que es 3.52.

Veredicto: con $\alpha = 0{,}05$, el dato experimental es demasiado discrepante de lo esperado según H_o, la cual se rechaza con dos colas. Eso se justifica cuando lo único que nos interesa es comparar nuestro curso con un historial previo: las notas tradicionales y las de la muestra son discrepantes.

Cuando se trata de un profesor con mucha experiencia, a nadie le preocupa tal discrepancia, pues se sabe que a veces las notas son bajas y a veces altas aunque con frecuencia son normales. Por consiguiente, para el profesor con mucha experiencia se acepta la H_o.

Si se trata de un profesor primíparo que es desconocido en el medio, del cual no se sabe nada de antemano, se rechaza la H_o: las notas de su curso son diferentes a las notas del historial universitario. Es usual que lo llamen a dar explicaciones.

Pero si se trata de un profesor que además de tener mucha experiencia vino de un doctorado en educación matemática, él va a decir: mis notas son mejores que las correspondientes a los cursos pasados y no porque los exámenes sean más fáciles, sino porque he utilizado una excelente metodología para enseñar un magnífico material.

En ese caso, el profesor demuestra su pretensión mostrando que las notas no sólo son más altas en los parciales, que los hace cada profesor, sino también en el examen final, que lo hace la Universidad. O también puede pasar que quede mal, lo cual sucede sea porque sus notas son iguales a las usuales o bien porque son peores.

Como se ve, con este gran profesor estamos considerando
$H_o : \mu = 3{,}2$.
mientas que
$H_a : \mu > 3{,}2$

El z_{obs} es el mismo que antes:

$z_{obs} = \frac{3{,}6-3{,}2}{0{,}6/\sqrt{28}} = 3.52$

Se deduce que la región de aceptación de la H_o es toda la que queda a la izquierda del valor crítico de z determinado por $\alpha = 0{,}05$ y una cola. Eso equivale a encontrar el valor crítico de c tal que $\Psi(z_c) = 0{,}95$. Por la tabla vemos que $z_c = 1{,}64$. Y como $z_{obs} = 3{,}52$, se rechaza la H_o: el profesor defendió su honor con gran dignidad. Claro que nos falta ver qué pasa en el final que es el mismo para todos. Eso sí que va a ser interesante.

184 ***Ejemplo*** *Piscicultura por nativos ancestrales.*

Un estudiante de antropología trabaja en una excavación sobre un lago disecado. Datación con C^{14} de los restos de algas le dan una antigüedad cercana a los 1500 años. El objetivo de su proyecto es estudiar los hábitos de pesca de la población ancestral que vivía en su orilla. Según la literatura, la longitud X del pez PPP en el estrato E4 era normalmente distribuido: $X \sim N(20, 4)$, unidades en centímetros. El estudiante trabaja sobre un estrato más profundo y, por tanto, más antiguo E5. Podría encontrarse que la longitud promedio del pez creció con el tiempo, apuntando a la posibilidad de que los pobladores se inventaron la forma de nutrir los peces. Pero tal vez la población humana creció por lo que pescaban más y los peces no tendrían tiempo de crecer. Otra posibilidad sería que hubo una disminución de la población, por guerra, y de sus requerimientos alimenticios por lo que los peces podían crecer más. Por eso debe estudiarse la dinámica de la población humana. Se solicita la opinión de la estadística sabiendo que con 16 especímenes ha encontrado una media de 17 cm.

Solución. Tenemos que hacer una prueba de dos colas pues la incertidumbre actual es muy grande:

$H_o : \mu_x = 20$
$H_a : \mu_x \neq 20$

Nosotros tenemos $n = 16$, $\mu_X = 20 = \mu_{\bar{X_{16}}}$; $\sigma_X = 4$ y $\sigma_{\bar{X_{16}}} = \frac{4}{\sqrt{16}} = 1$. Por consiguiente, $\bar{X_{16}} \sim N(20, 1)$.

Los datos de la muestra produjeron $\bar{X_{16}} = 17$. El correspondiente z es

$z_{obs} = \frac{\bar{X}_n - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}$
$z_{obs} = \frac{17-20}{\frac{4}{\sqrt{16}}} = -3$

Como no dijeron nada sobre la significancia, la tomamos como 0.05. Como tomamos dos colas, todo lo que tenga una z más grande que 1.96 o menor que -1.96 será extraño. Eso es precisamente lo que tenemos: nuestro dato está a 3 desviaciones abajo de la media.

Veredicto. Nuestro análisis es compatible con la siguiente sospecha del estudiante: la longitud promedio del pez aumentó con el tiempo pues los peces fueron alimentados por la población indígena, lo cual corrobora la idea de que la piscicultura es en realidad muy antigua y fue reinventada en todos lados. Esta sugerencia podría corroborarse estudiando la calidad de los huesos de los nativos, cuya solidez también debió haber aumentado consecuentemente.

185 ***Ejemplo*** *Coltenis está buscando un nuevo representante de Colombia para enviarlo a fogueo por Europa. El rendimiento de un tenista se mide usando diversos descriptores y puede depender del gusto del coordinador. Usaremos el número de aces + passing shots por partido como nuestro descriptor. Uno de los tenistas opcionados es Juan Marulanda, cuyo desempeño de acuerdo a nuestro descriptor se distribuye como* $N(\mu = 9, \sigma = 2)$*. Pero en los últimos 5 partidos su promedio apenas fue de 6. Al notar su mal desempeño, el cual era evidente para todos, Juan fue donde el Coordinador y le dijo: 'yo para esto no sirvo'. Y cuando lo dijo, olía a borracho. ¿Es justificada su reacción desertora? Y ¿qué fue lo que el Coordinador le dijo que desde afuera se oían los sollozos? ¿Y por qué salió de su encuentro con el Coordinador derechito a entrenar?*

Solución: Como el mal desempeño de Juan se dio en un momento de indisciplina con borrachera y todo, la prueba estadística que le aplicamos es con una sola cola, pues se le acusa de haber disminuido intencionalmente su rendimiento. Tomamos $\alpha = 0{,}05$ para no hacer nada que pueda parecer sesgado.
$H_o : \mu = 9$
$H_A : \mu < 9$
$\bar{X}_5 = 6$
$\frac{\sigma}{\sqrt{5}} = \frac{2}{\sqrt{5}} = 0{,}89$
$z = \frac{6-9}{0{,}89} = -3{,}37$
El z crítico para la cola inferior es
$z(\alpha = 0{,}05) = -1{,}64$
Como el z experimental es -3.37 que es mucho menor que el z crítico, deducimos que el desempeño del deportista realmente fue muy inferior al usual. Puesto que eso se dio en una época de tomar trago, es justo que renuncie.

Y ¿qué fue lo que el Coordinador le dijo a Juan que desde afuera se oían los sollozos? Es supremamente claro para un Coordinador experimentado que para que un joven tenista de alto rendimiento se emborrache en tiempo de combate es porque debe haber causa mayor. En este caso, Juan fue traicionado por su novia, quien alega que él no le sirve pues no le dedica tiempo a ella. Y Juan confesó su dolor, se desahogó y por eso lloró, cosa que no había hecho ni siquiera borracho.

¿Y por qué salió Juan de su encuentro con el Coordinador derechito a entrenar? Porque Coltenis ha invertido en Juan mucho dinero en diversos gastos, en entrenamiento básico, en escuelas especiales, en fogueos y en viajes. Fuera de eso, la familia de Juan ha invertido millonadas en cientos de gastos que nadie nunca contabiliza. ¿Echaría un Coordinador toda esa plata al fuego simplemente por una novia descontenta? Por supuesto que no: La sagrada responsabilidad de un coordinador es la de poner vino y aceite sobre las heridas de sus estudiantes, para que sanen sin infectarse.

Pero entonces, ¿Para qué sirve la estadística si cualquier Coordinador puede anular sus directivas?

Los veredictos de la estadística se aplican mejor para casos estadísticos, es decir, para muchos eventos. Pero cuando se trata de eventos personalizados, cuando se trata de Juan Marulanda, las decisiones estadísticas no tienen mucha autoridad, ni siquiera porque los otros 23 descriptores con que midieron a Juan también mostraron que Juan para el tenis no sirve.

Casos como éste denotan que en cualquier campo existe una gran diferencia entre un técnico y un experto. Un técnico es el que cree que las técnicas llevan a la verdad y a la sabiduría. El experto utiliza las técnicas cuando y como su sabiduría se lo indican.

186 ***Ejemplo*** *Los gastos de una empresa de salud por persona durante un trimestre tienen una distribución de tendencia central que se aproxima por una normal* $N(\mu = 10, \sigma = 2)$*, donde una unidad*

representa 10.000 pesos. La empresa tiene la intención de cobrar una tarifa extra a los niños y a los ancianos pues se alega que ellos salen muy caros. Para ver si eso era verdad o no, se tomó una muestra al azar de 20 pacientes, entre niños y ancianos, y se obtuvo una media de 12. ¿Son en realidad más caros los niños y los ancianos?

Solución:
$X \sim N(\mu = 10, \sigma = 2)$
$H_o : \mu = 10$
$H_A : \mu > 10$
$\bar{X}_{20} = 12$
$\frac{2}{\sqrt{20}}$=0,4
$z_{obs} = \frac{12-10}{0,4}$=4,47

El z crítico con la cola derecha y $\alpha = 0,05$ es 1.64. Como el $z_{obs} = 4,47$ es mayor que el crítico, se rechaza la H_o: los niños y los ancianos sí representan comparativamente altos costos.

8.2. Estudio de una proporción

Debido a que las proporciones se estudian dentro de las binomiales, y que éstas se pueden aproximar por la normal, lo que hemos visto para la media de normales también se puede aplicar al estudio de proporciones. Veamos.

187 ***Ejemplo*** *Estamos en un juego con monedas generalizadas: se lanza una moneda 30 veces y se contó el número de caras que fue 19. El problema consiste en decidir con $\alpha = 0,05$ si en la moneda generalizada se tiene que $p(cara) = 1/2$, lo cual es la H_o. La H_a es la no igualdad, lo cual se estudia con dos colas, pues hay dos maneras de no ser igual, una por que se es menor y otra porque se es mayor.*

Solución: Calculemos la H_o:
Si $X \sim Bi(30, 1/2)$ entonces la media sería 15, la varianza sería 7.5 y la desviación 2.64. Aproximamos la binomial por la normal. Como estamos con una normal, podemos medir la discrepancia entre lo esperado y lo hallado con una transformada z. Como se lanzó la moneda 30 veces y salieron 19 caras, la discrepancia es:
$z = \frac{19-15}{2,64} = 1,515$
Criterio de aceptación de la H_o para $\alpha = 0,05$: se acepta la H_o si la $z(19)$ queda entre -1.96 y 1.96. Se rechaza la H_o en caso contrario.
Veredicto: como 1.515 queda entre -1.96 y 1.96, la discrepancia entre el experimento y la H_o queda bien explicada por el azar. Aparte del azar no se requieren mejores explicaciones. El experimento respalda la H_o, la cual se acepta.

8.3. Otros estadígrafos de contraste

- Se usa la z como estadígrafo de contraste para estudiar experimentos con medias cuando *se conoce la desviación poblacional* σ que es la desviación de todo el universo, bien porque se hizo un censo, una muestra muy grande o porque se leyó en la literatura. También se usa para decidir experimentos con proporciones.

- Se usa la t como estadígrafo de contraste para estudiar experimentos con medias cuando **NO** se conoce la desviación poblacional pero tenemos la desviación de una muestra, la cual notamos $s = s_X$. La t mide la discrepancia entre una media muestral y otra poblacional en unidades apropiadas:

  $t = \frac{\bar{X}_n - \mu_X}{\frac{s_X}{\sqrt{n}}}$

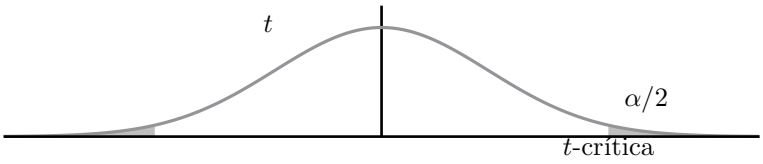


Figura 8.3. La t se parece a la z, pero es un poco más explayada. En una prueba de dos colas, hay dos t-críticas. Y a la derecha del valor crítico máximo queda la región de eventos extremos cuya probabilidad es $\alpha/2$.

Pero atención, entre más datos, más confiables serán las decisiones tomadas. Cuando eso pasa, se introduce un parámetro llamado **grados de libertad** y que se denota **gli**. La t es parecida a la z y por eso es simétrica, pero es más ancha, es decir, con más incertidumbre, pues la desviación no vino de un censo sino de una muestra. Cuando hay muchos datos, más de 30, la t se cierra y llega a ser igual a la z.

- Se usa la chi-cuadrado, χ^2 para comparar una varianza muestral VAR con una poblacional σ^2 que viene de una H_o. Los gli son $n-1$:

  $\chi^2 = \frac{(n-1)VAR}{\sigma^2}$ con $n-1$ gli

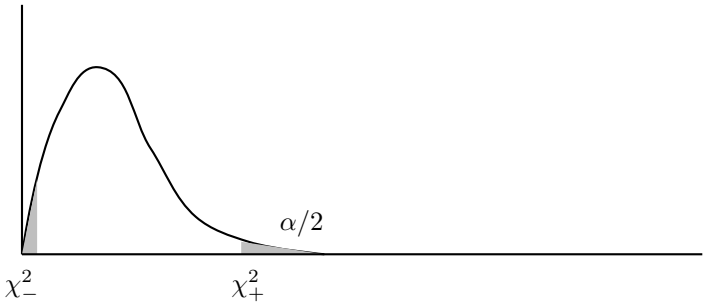


Figura 8.4. La tabla adjunta da los valores críticos χ^2_- *y* χ^2_+ *para un* α *dado, una y dos colas.*

Como la chi es un quebrado en el que hay divisiones de números positivos, la chi es siempre positiva, es decir, se restringe a la parte positiva del eje horizontal. La chi no es simétrica, por ello nuestra tabla tiene las chi-críticas para una significancia dada tanto la de arriba como la de abajo.

Cuando los gli son muy grandes y no están en la tabla, se usa la siguiente aproximación, donde $gli = n-1 = \nu$,

$$\chi^2 \sim N(\mu = \nu, \sigma = \sqrt{2\nu}).$$

- Se usa la F para comparar dos varianzas poblacionales conociendo las varianzas muestrales. **Atención:** la z y la t son estadígrafos que contrastan por substracción pero la chi y la F lo hacen por relaciones de quebrado. Por lo tanto, la H_o para comparar dos varianzas muestrales dice cuánto vale R, la relación esperada entre la varianza grande y la pequeña:

  $H_o : \frac{\sigma_G^2}{\sigma_P^2} = R$

  Decir en una H_o que las varianzas sean iguales es lo mismo que decir que $R = 1$. Cuando la prueba es con dos colas, la $H_a : R \neq 1$. Cuando la prueba es con una cola, debido a que la varianza mayor siempre se pone en el numerador, la $H_a : R > 1$.

  Estadígrafo de discrepancia: $F = \frac{VAR_G}{RVAR_P}$

  con $m-1$ gli en el numerador y $n-1$ gli en el denominador. Como en la F hay divisiones de números positivos, la F es siempre positiva, es decir, se restringe a la parte positiva del eje horizontal.

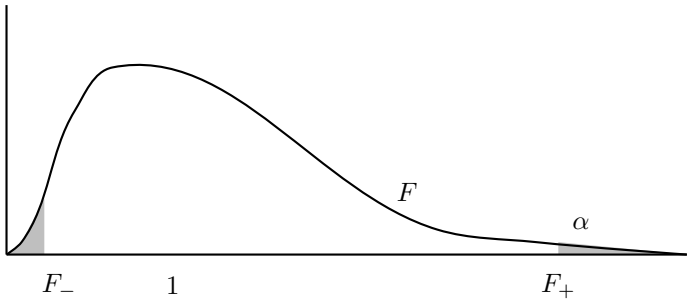


Figura 8.5. Esquema general de cualquier F. Para un α dado y dos colas, hay valores críticos denotados F_- y F_+. La moda está por ahí cerca de UNO.

La F es asimétrica, por eso se requiere dar las dos F-críticas para una significancia dada tanto la de arriba como la de abajo. Lamentamos informar que ninguna tabla presenta los dos valores críticos de la F, sino que da solamente el valor superior. Todo se remedia calculando el límite inferior usando la siguiente relación:

$F_{\alpha/2,m-1,n-1} = \frac{1}{F_{1-\alpha/2,n-1,m-1}}$

Para decidir si la discrepancia entre lo predicho por la H_o es grande o pequeña, se compara la F hallada con el valor crítico dado por la tabla de la F según la significancia. Pero atención, entre más datos hay más exactitud, por consiguiente debe haber algún parámetro que registre el número de datos tanto del numerador como del denominador: se llama 'grados de libertad del numerador', número de datos menos uno: $glinumerador = m - 1$ y 'grados de libertad del denominador $glidenominador = n - 1$.

188 ***¿De dónde salen los estadígrafos?***

Los **estadígrafos** que hemos visto son la z, la t, la F y la χ^2. Hemos visto también sus histogramas. En términos oficiales, lo que nosotros llamamos histograma de un estadígrafo se denomina **función de densidad de probabilidad**. A algunos estudiantes les parece intrigante saber de dónde sale todo eso. Nuestra respuesta es cualitativa y se basa en la z y es válida para los estadígrafos vistos y para cualquier otro que exista o pueda existir. Veámosla:

El estadígrafo z para medir la discrepancia entre x, el valor obtenido en un experimento que toma muestras al azar de una distribución normal, y μ, el valor predicho por una H_o, tiene la forma:

$z = \frac{x-\mu}{\sigma}$

y a esta z le hemos asociado una campana con media cero y desviación uno. Veamos el porqué: de una variable aleatoria normalmente distribuida con media μ y desviación σ tomamos un dato al azar x. Le calculamos la z, la discrepancia con respecto a la media. Nos da un resultado, al cual se le grafica como una rayita horizontal en un gráfica con ejes coordenados. Si repetimos lo mismo, nos da otra rayita y si repetimos lo mismo muchas veces, nos da un histograma. ¿Cómo se idealiza este histograma? La idealización más sencilla y más ajustada o la que mejor cuadra es una Campana de Gauss con media cero y desviación uno. Es por eso que la campana de la z representa lo que pasaría si la H_o fuese cierta, es decir si todos los datos provienen de una distribución normal con la media predicha.

Como generalización para cualquier estadígrafo tenemos: se saca un dato o un muestreo de la distribución elegida, se mide la discrepancia con lo predicho por la H_o que se estudie, se grafica el resultado del estadígrafo como una rayita horizontal sobre una gráfica. Se repite lo mismo muchas veces y se idealiza el histograma resultante. La curva que sale de la idealización es la función densidad del estadígrafo estudiado si la H_o es verdadera. Si dicha curva tiene una distribución de tendencia central, uno puede decidir, como de costumbre, qué es extraño y qué es normal y de esa forma obtener un criterio para aceptar o rechazar la H_o: cuando la discrepancia entre lo observado y lo predicho por la H_o es pequeña, no se considera relevante y la H_o se acepta, de lo contrario, la H_o se rechaza.

8.4. Tablas estadísticas

Presentamos a continuación las tablas de la t, la χ^2 y la F. Si uno llegase a necesitar mucha precisión, lo mejor es remitirse a Gnumeric o algún paquete estadístico.

La t

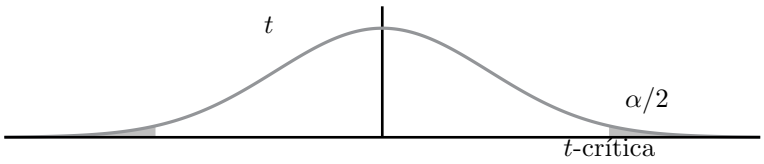


*Figura 8.6. La t se parece a la z, aunque es un poco más explayada. Cuando los **gli** (grados de libertad) crecen, la t tiende a la z pues la incertidumbre decrece. Para dos colas y para una significancia α hay dos valores simétricos de la t-crítica.*

La t-crítica. Para 2 colas: $\beta = \alpha$. Para 1 cola: $\beta = 2\alpha$						
gli	$\beta = 0{,}20$	$\beta = 0{,}10$	$\beta = 0{,}05$	$\beta = 0{,}04$	$\beta = 0{,}02$	$\beta = 0{,}01$
1	3.078	6.314	12.706	15.894	31.821	63.657
2	1.886	2.920	4.303	4.849	6.965	9.925
3	1.638	2.353	3.182	3.482	4.541	5.841
4	1.533	2.132	2.776	2.998	3.747	4.604
5	1.476	2.015	2.571	2.756	3.365	4.032
6	1.440	1.943	2.447	2.612	3.143	3.707
7	1.415	1.895	2.365	2.517	2.998	3.499
8	1.397	1.860	2.306	2.449	2.896	3.355
9	1.383	1.833	2.262	2.398	2.821	3.250
10	1.372	1.812	2.228	2.359	2.764	3.169
11	1.363	1.796	2.201	2.328	2.718	3.106
12	1.356	1.782	2.179	2.303	2.681	3.055
13	1.350	1.771	2.160	2.282	2.650	3.012
14	1.345	1.761	2.145	2.264	2.624	2.997
15	1.341	1.753	2.131	2.249	2.602	2.947
16	1.337	1.746	2.120	2.235	2.583	2.921
17	1.333	1.740	2.110	2.224	2.567	2.898
18	1.330	1.734	2.101	2.214	2.552	2.878
19	1.328	1.729	2.093	2.205	2.539	2.861
20	1.325	1.725	2.086	2.197	2.528	2.845
21	1.323	1.721	2.080	2.189	2.518	2.831
22	1.321	1.717	2.074	2.183	2.508	2.819
23	1.319	1.714	2.069	2.177	2.500	2.807
24	1.318	1.711	2.064	2.172	2.492	2.797
25	1.316	1.708	2.060	2.166	2.485	2.787
26	1.315	1.706	2.056	2.162	2.479	2.779
27	1.314	1.703	2.052	2.157	2.473	2.771
28	1.313	1.701	2.048	2.154	2.467	2.763
29	1.311	1.699	2.045	2.151	2.462	2.756
30	1.310	1.697	2.042	2.147	2.457	2.750
∞	1.285	1.645	1.960	2.054	2.326	2.576

Para $gli > 30$, la t y la z coinciden, por lo tanto, en ese caso puede usarse la z en vez de la t lo cual permite trabajar con cualquier significancia. **Se usa la t para muestras pequeñas, la z para muestras grandes.**

La χ^2 Chi-cuadrado Gráfica y Tablas

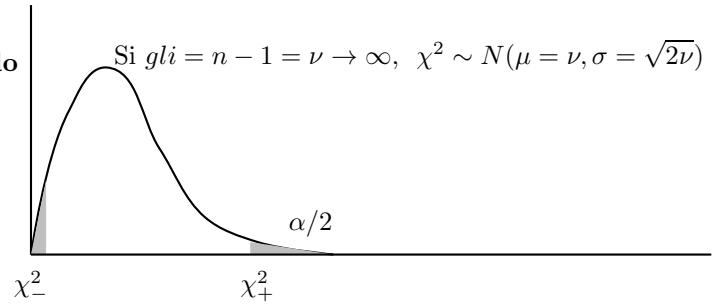


Figura 8.7. La tabla adjunta da los valores críticos χ^2_- y χ^2_+ para un α dado, una y dos colas.

χ^2-crítica. Con DOS colas $\beta = \alpha$; con UNA, $\beta = 2\alpha$. **gli** = grados de libertad										
gli	$\beta = 0{,}01$	$\beta = 0{,}01$	$\beta = 0{,}02$	$\beta = 0{,}02$	$\beta = 0{,}05$	$\beta = 0{,}05$	$\beta = 0{,}10$	$\beta = 0{,}10$	$\beta = 0{,}20$	$\beta = 0{,}20$
	χ^2_-	χ^2_+	χ^2_-	χ^2_+	χ^2_-	χ^2_+	χ^2_-	χ^2_+	χ^2_-	χ^2_+
1	0^+	7.879	0.00016	6.635	0.00098	5.024	0.0039	3.8415	0.016	2.706
2	0.010	10.596	0.020	9.210	0.051	7.378	0.103	5.991	0.211	4.605
3	0.072	12.838	0.115	11.345	0.216	9.348	0.352	7.815	0.584	6.251
4	0.207	14.860	0.297	13.276	0.484	11.143	0.711	9.488	1.064	7.779
5	0.412	16.750	0.554	15.086	0.831	12.832	1.145	11.070	1.610	9.236
6	0.676	18.547	0.872	16.812	1.237	14.449	1.635	12.592	2.204	10.645
7	0.989	20.277	1.239	18.475	1.690	16.013	2.167	14.067	2.833	12.017
8	1.344	21.955	1.646	20.089	2.180	17.534	2.733	15.507	3.490	13.362
9	1.735	23.589	2.088	21.665	2.700	19.023	3.325	16.918	4.168	14.684
10	2.156	25.188	2.558	23.208	3.247	20.483	3.940	18.307	4.865	15.987
11	2.603	26.756	3.053	24.724	3.816	21.919	4.575	19.675	5.578	17.275
12	3.074	28.298	3.571	26.217	4.404	23.336	5.226	21.025	6.304	18.549
13	3.565	29.818	4.107	27.687	5.009	24.735	5.892	22.361	7.041	19.812
14	4.075	31.318	4.660	29.141	5.629	26.119	6.571	23.685	7.789	21.064
15	4.601	32.799	5.229	30.578	6.262	27.487	7.261	24.995	8.547	22.306
16	5.142	34.267	5.812	32.000	6.908	28.845	7.962	26.295	9.312	23.541
17	5.697	35.717	6.408	33.407	7.564	30.189	8.672	27.586	10.085	24.769
18	6.265	37.154	7.015	34.804	8.231	31.526	9.390	28.868	10.865	25.989
19	6.844	38.580	7.633	36.189	8.906	32.852	10.117	30.143	11.651	27.202
20	7.434	39.996	8.260	37.564	9.591	34.169	10.851	31.410	12.442	28.411
21	8.034	41.398	8.897	38.930	10.283	35.478	11.591	32.669	13.239	29.615
22	8.643	42.793	9.542	40.287	10.982	36.780	12.338	33.922	14.041	30.812
23	9.260	44.181	10.196	41.635	11.688	38.074	13.090	35.172	14.848	32.006
24	9.886	45.557	10.856	42.979	12.401	39.362	13.848	36.413	15.658	33.195
25	10.519	46.927	11.524	44.312	13.120	40.645	14.611	37.650	16.473	34.381
26	11.160	48.288	12.198	45.641	13.844	41.921	15.379	38.884	17.291	35.562
27	11.807	49.644	12.878	46.959	14.573	43.193	16.151	40.110	18.113	36.739
28	12.461	50.989	13.564	48.274	15.308	44.459	16.928	41.335	18.939	37.913
29	13.121	52.331	14.256	49.584	16.047	45.721	17.708	42.557	19.768	39.084
30	13.786	53.669	14.953	50.889	16.791	46.976	18.492	43.770	20.599	40.254

A medida que el número de datos n crece, la gráfica de la χ^2 se corre hacia la derecha y se estira pareciéndose a una normal con media $\mu = n-1$ y desviación $\sigma = \sqrt{2(n-1)}$. Por lo que para $n > 30$ se puede aplicar la fórmula de los valores críticos de la Z: $\chi^2_- = \mu - z\sigma$, $\chi^2_+ = \mu + z\sigma$, donde la z se toma de acuerdo a la significancia y con dos colas (pero sin glis). En definitiva:

$\chi^2_- = n - 1 - z_{(\alpha,2)}\sqrt{2(n-1)}$ en tanto que $\chi^2_+ = n - 1 + z_{(\alpha,2)}\sqrt{2(n-1)}$. Ej. 238 pág. 224.

La F Gráfica y Tablas

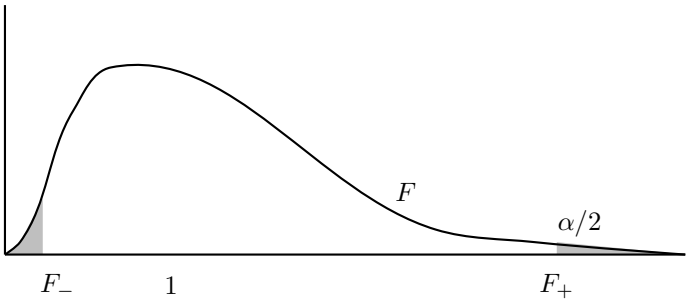


Figura 8.8. Esquema general de cualquier F. Para un α dado y dos colas, hay valores críticos denotados F_- y F_+. La moda está por ahí cerca de 1.

Tabla de F_+, valor crítico superior de la F, con DOS colas y $\alpha = 0{,}02$, o con UNA cola y $\alpha = 0{,}01$.

En el primer renglón están los gli = grados de libertad del numerador subrayados. En la primera columna los gli del denominador.

F_+ = Valor crítico superior de la F con DOS colas y $\alpha = 0{,}02$, o con UNA cola y $\alpha = 0{,}01$														
En el primer renglón están los gli del numerador. En la primera columna los gli del denominador														
gli	1	2	3	4	5	6	7	8	9	10	12	15	20	∞
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.00
l5	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	2.65
18	8.29	16.01	509	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	2.57
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.36
22	7.95	.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	1.60
120	6.85	4.79	395	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.00

Tabla de F_+, valor crítico superior de la F, con DOS colas y $\alpha = 0{,}05$, o con UNA cola y $\alpha = 0{,}025$

En el primer renglón están los gli = grados de libertad del numerador subrayados. En la primera columna los gli del denominador.

F_+ = Valor crítico superior de la F con DOS colas y $\alpha = 0{,}05$, o con UNA cola y $\alpha = 0{,}025$															
En el primer renglón están los gli del numerador. En la primera columna los gli del denominador															
gli	1	2	3	4	5	6	7	8	9	10	15	20	30	40	∞
2	38,5	39	39,1	39,2	39,2	39,3	39,3	39,3	39,3	39,3	39,4	39,4	39,4	39,4	39,4
3	17,4	16,0	15,4	15,1	14,8	14,7	14,6	14,5	14,4	14,4	14,2	14,1	14,0	14,0	13,9
4	12,2	10,6	9,9	9,6	9,3	9,1	9,0	8,9	8,9	8,84	8,65	8,55	8,46	8,41	8,26
5	10,0	8,43	7,76	7,38	7,14	6,97	6,85	6,75	6,68	6,61	6,42	6,32	6,22	6,17	6,02
6	8,81	7,25	6,59	6,22	5,98	5,81	5,69	5,59	5,52	5,46	5,26	5,16	5,06	5,01	4,85
7	8,07	6,54	5,88	5,52	5,28	5,11	4,99	4,89	4,82	4,76	4,56	4,46	4,36	4,30	4,14
8	7,57	6,05	5,41	5,05	4,81	4,65	4,52	4,43	4,35	4,29	4,10	3,99	3,89	3,83	3,67
9	7,20	5,71	5,07	4,71	4,48	4,31	4,19	4,10	4,02	3,96	3,76	3,66	3,56	3,50	3,34
10	6,93	5,45	4,82	4,46	4,23	4,07	3,94	3,85	3,77	3,71	3,52	3,41	3,31	3,25	3,08
11	6,72	5,25	4,63	4,27	4,04	3,88	3,75	3,66	3,58	3,52	3,32	3,22	3,11	3,06	2,89
12	6,55	5,09	4,47	4,12	3,89	3,72	3,60	3,51	3,43	3,37	3,17	3,07	2,96	2,90	2,7
13	6,41	4,96	4,34	3,99	3,76	3,60	3,48	3,38	3,31	3,24	3,05	2,94	2,83	2,77	2,60
14	6,29	4,85	4,24	3,89	3,66	3,50	3,37	3,28	3,20	3,14	2,94	2,84	2,73	2,67	2,49
15	6,19	4,76	4,15	3,80	3,57	3,41	3,29	3,19	3,12	3,06	2,86	2,75	2,64	2,58	2,40
16	6,11	4,68	4,07	3,72	3,50	3,34	3,21	3,12	3,04	2,98	2,78	2,68	2,56	2,50	2,32
17	6,04	4,61	4,01	3,66	3,43	3,27	3,15	3,06	2,98	2,92	2,72	2,61	2,50	2,44	2,25
18	5,97	4,55	3,95	3,60	3,38	3,22	3,09	3,00	2,92	2,86	2,66	2,55	2,44	2,38	2,19
19	5,92	4,50	3,90	3,55	3,33	3,17	3,05	2,95	2,88	2,81	2,61	2,50	2,39	2,33	2,14
20	5,87	4,46	3,85	3,51	3,28	3,12	3,00	2,91	2,83	2,77	2,57	2,46	2,34	2,28	2,09
21	5,82	4,41	3,81	3,47	3,25	3,08	2,96	2,87	2,79	2,73	2,53	2,42	2,30	2,24	2,05
22	5,78	4,38	3,78	3,4	3,21	3,05	2,93	2,83	2,76	2,69	2,49	2,38	2,27	2,20	2,01
23	5,74	4,34	3,75	3,40	3,18	3,02	2,90	2,80	2,73	2,66	2,46	2,3	2,23	2,17	1,97
24	5,71	4,31	3,72	3,37	3,15	2,99	2,87	2,77	2,70	2,63	2,43	2,32	2,20	2,14	1,94
25	5,68	4,29	3,69	3,35	3,12	2,96	2,84	2,75	2,67	2,61	2,41	2,30	2,18	2,1	1,91
30	5,56	4,18	3,58	3,24	3,02	2,86	2,74	2,65	2,57	2,51	2,30	2,19	2,07	2,00	1,79
40	5,42	4,05	3,46	3,12	2,90	2,74	2,62	2,52	2,45	2,38	2,1	2,06	1,94	1,87	1,64
50	5,34	3,97	3,39	3,05	2,83	2,67	2,55	2,45	2,38	2,31	2,10	1,99	1,86	1,79	1,55
60	5,28	3,92	3,34	3,00	2,78	2,62	2,50	2,41	2,33	2,27	2,06	1,94	1,81	1,74	1,49
120	5,15	3,80	3,22	2,89	2,67	2,5	2,39	2,29	2,22	2,15	1,94	1,82	1,68	1,61	1,32
Mil	5,03	3,70	3,12	2,79	2,57	2,42	2,30	2,20	2,12	2,06	1,84	1,72	1,58	1,49	1,13

Ejemplo 1: Significancia = 0.01, una cola. Gli de la varianza grande = Gli del numerador = 20, gli de la pequeña = gli del denominador = 13. La F vale 3.66.

Ejemplo 2: Significancia = 0.025, una cola. Gli de la varianza grande = Gli del numerador = 15, gli de la pequeña = gli del denominador = 7, La F = 4.56.

Ejemplo 3: Significancia = 0.10, dos colas. Gli de la varianza grande = Gli del numerador = 3, gli de la pequeña = gli del denominador = 17. La F vale 3.20.

Ejemplo 4: Significancia = 0.05, dos colas. Gli de la varianza grande = Gli del numerador = 15, gli de la pequeña = gli del denominador = 7. La F = F+ vale 4.56. La F con los gli intercambiados vale 3.29, la F- vale $1/3.29 = 0.304$.

Tabla de F_+, valor crítico superior de la F, con DOS colas y $\alpha = 0{,}10$, o con UNA cola y $\alpha = 0{,}05$

En el primer renglón están los gli = grados de libertad del numerador subrayados. En la primera columna los gli del denominador.

F_+ = Valor crítico superior de la F con DOS colas y $\alpha = 0{,}10$, o con una cola y $\alpha = 0{,}05$															
En el primer renglón están los gli del numerador. En la primera columna los gli del denominador															
gli	1	2	3	4	5	6	7	8	9	10	15	20	30	40	∞
2	18.5	19.0	19.1	19.2	19.3	19.3	19.3	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.62	8.59	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.75	5.72	5.61
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.50	4.46	4.37
6	599	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.81	3.77	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.41	3.38	3.34	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.08	3.04	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.86	2.83	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.70	2.66	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.57	2.53	2.41
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.47	2.43	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.38	2.34	2.21
14	460	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.31	2.27	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.25	2.20	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.19	2.15	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23	2.15	2.10	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.11	2.06	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.07	2.03	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.04	1.99	1.85
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.18	2.10	2.01	1.96	1.82
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.07	1.98	1.94	1.79
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.13	2.05	1.96	1.91	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11	2.03	1.94	1.89	1.74
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.09	2.01	1.92	1.87	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.84	1.79	1.63
40	408	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84	1.74	1.69	1.51
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87	1.78	1.69	1.63	1.44
60	4.00	3.15	2.76	253	2.37	2.25	2.17	2.10	2.04	1.99	1.84	1.75	1.65	1.59	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.75	1.66	1.55	1.50	1.26
∞	3.85	3.00	2.61	2.38	2.22	2.10	2.01	1.94	1.88	1.84	1.67	1.58	1.46	1.40	1.00

8.5. Conclusión

La ciencia es un contraste entre lo que se ve y lo que se cree. Lo que se cree se consigna en la H_o, la hipótesis nula, que siempre es una igualdad, por ejemplo, que la media vale 9. La alternativa a creer la H_o es la H_a, la hipótesis alterna, la cual dice si el estudio de la H_o es con una o con dos colas. La discrepancia entre lo que se ve y lo que se cree puede medirse, puede cuantificarse por medio de un estadígrafo de contraste y declararse fruto del azar, aceptando la H_o, o quizá resultado de un efecto sistemático no considerado en la H_o, rechazando la H_o. Eso depende de si la discrepancia es menor o mayor que el valor crítico del estadígrafo utilizado para medir la discrepancia. Los principales estadígrafos son la z, la t, la χ^2 y la F.

Capítulo 9

La ciencia se viste de gala

189 ***Objetivo*** *Implementar sobre una muestra seleccionada de experimentos el esquema oficial de la ciencia: el método científico es básicamente un contraste entre lo que se ve y lo que se cree. El material es fundamental pues constituye lo que será no sólo la rutina de la vida real sino la forma sencilla, clara y elegante de hacer ciencia.*

9.1. Instrucciones para resolver problemas

Se asume que los datos corresponden a una muestra aleatoria de una población grande. Cuando se tiene un solo dato, la varianza asociada es cero. Las variables cuantitativas se suponen con distribución normal. Decimos que los datos vienen apareados cuando vienen por pares, como ojo derecho vs. ojo izquierdo de la misma cabeza. Los datos son no apareados cuando vienen cada uno por su lado.

Antes de tratar de resolver un problema, uno debe saber clasificarlo. Nuestras opciones son muy pocas: comparamos proporciones, medias o varianzas. Si comparamos proporciones sólo hay un caso, el cual se ilustra con un ejemplo modelo. Si comparamos medias puede ser que se trate de datos apareados o puede ser que se trate de datos no apareados. Si se trata de datos apareados, se busca el ejemplo correspondiente. Si se trata de comparar medias sobre datos no apareados, primero se averigua si vienen de poblaciones con varianzas iguales o con varianzas diferentes. Y cada caso se resuelve según su ejemplo.

Todos los problemas se resuelven por exactamente la misma metodología: se codifica el problema dentro de la camisa de fuerza de la ciencia que dice: **el método científico es un contraste medido en unidades apropiadas entre lo que uno observa (los datos) y lo que uno cree (la hipótesis nula).** Luego se resuelve el problema codificado y su solución se reinterpreta en términos de los enunciados del problema. Se siguen los pasos siguientes:

PASO 1: Se estudia el enunciado del problema para formular la hipótesis nula y la alterna. La hipótesis nula dice lo que se cree. La H_o siempre contiene una igualdad, por ejemplo que una diferencia de medias vale cero. La hipótesis alterna especifica con cuantas colas se trabaja. Para decidir si se trata de una cola o de dos colas se mira el contexto del problema:

1. Si se enfatiza una igualdad, la H_o es la igualdad y la H_a es la desigualdad, la cual es con dos colas, pues hay dos maneras de ser desigual.

2. Si el enunciado pregunta sobre una desigualdad (ser mejor, más alto, más grande, más rico), la hipótesis nula es la igualdad y la alterna contiene el símbolo de ser mayor que y es con una cola pues dado un número hay sólo una manera como otro número puede ser mayor.

3. Si se dice: yo tengo el doble de puntaje promedio que Usted, se trata con dos colas, la hipótesis nula es la igualdad: mi promedio es dos veces el tuyo. La hipótesis alterna es la desigualdad.

4. Si se dice: yo tengo más del doble de puntaje promedio que Usted, se trata con una cola, la hipótesis nula es la igualdad: mi promedio es dos veces el tuyo. La hipótesis alterna contiene el símbolo de ser mayor que.

PASO 2: Se especifica lo que se ve, lo observado. Por ejemplo, que la diferencia de las medias de dos muestras es 8. Se ruega encarecidamente homogenizar las dimensiones: si el problema habla de días, meses y años, reducir todo, por ejemplo, a meses.

PASO 3: Se mide la discrepancia entre lo que se ve y lo que se cree, entre lo observado y lo esperado. Para ello se usa un estadígrafo de contraste adecuado, eligiendo el caso adecuado entre los 10 planteados. Para el taller siguiente sólo se requieren los siguientes casos entre los 10 de la lista de los más importantes:

1. Comparar un dato con una media, o una media muestral con una media poblacional sabiendo la varianza poblacional.
2. Comparar dos proporciones.
3. Comparar dos varianzas.
4. Comparar dos medias. Para ello, primero hay que comparar las varianzas. Si las varianzas poblacionales resultan iguales, se aplica un procedimiento, son diferentes, otro.
5. Comparar las medias de datos apareados.

PASO 4: Se decide si la discrepancia medida es grande o pequeña. La discrepancia se define como grande si es más extrema que la discrepancia crítica. la discrepancia es pequeña cuando es menos extrema que el valor crítico de la discrepancia por puro azar, la cual se mira en las tablas.

La discrepancia crítica depende de la significancia α, la cual dice si un valor del estadígrafo considerado que mide diferencias o discrepancias es interesante o común y corriente. Si el problema no da la significancia, uno puede tomar por defecto $\alpha = 0{,}05$.

Para leer el valor crítico de un estadígrafo de una tabla se tiene en cuenta que cada tabla tiene su propio código de interpretación, el cual está al comienzo. Con tablas de otros libros conviene tener en cuenta:

- Si el problema es con dos colas y la tabla es con dos colas, se mira simplemente en la columna de la significancia dada.
- Si el problema es con una cola y la tabla es con una cola, se mira simplemente en la tabla de la significancia dada.
- Si el problema es con dos colas con significancia α y la tabla con una cola, se mira en la columna cuya significancia sea $\alpha/2$.
- Si el problema es con una cola con significancia α y la tabla con dos, se mira en la columna cuya significancia sea 2α.

PASO 5: Para terminar se toma un veredicto: Si la discrepancia dada por el estadígrafo de contraste es pequeña, la hipótesis nula se acepta. Pero si la tal discrepancia es grande, la hipótesis nula se rechaza. Luego se reescribe la decisión en los términos literarios del problema. **Un problema sin veredicto no vale nada** pues hay que recordar que a uno le pagan por tomar decisiones y no por hacer cálculos.

En resumen: la hipótesis nula, la de la igualdad dice que la población de datos no es homogénea, un proceso de muestreo aleatorio dará a veces un resultado y a veces otro. Por consiguiente, la diferencia entre lo esperado en la H_o y lo observado es un resultado del muestreo y no de un proceso sistemático. Decimos que la discrepancia resultante se debe al azar. La hipótesis alterna dice que aunque el azar en

el muestreo nunca deja de funcionar, si la discrepancia entre lo esperado según la H_o y lo observado es demasiado grande, entonces es más cómodo interpretar dicha enorme diferencia como el resultado de un proceso sistemático, es decir como el resultado de la actividad que se haya hecho. Ejemplo:

Uno a veces se siente fuerte y a veces débil. Si uno come y se siente fuerte, quizá no sea por la comida sino porque a veces uno se siente fuerte y a veces débil y esta vez uno se sintió fuerte por puro azar y la comida no tuvo nada que ver. Pero si uno definitivamente se siente muy fuerte, uno tiende a pensar que es por la comida y no hay nadie que lo haga cambiar de opinión.

9.2. Ejemplos prototipo

Aplicamos la caracterización siguiente: **el método científico es un contraste, la medida de una discrepancia, entre lo que se ve y lo que se cree en la H_o. Si la discrepancia entre lo que se ve y lo que se cree es pequeña, se acepta lo que se cree, la H_o. Si la discrepancia es grande, se repiensa lo que se cree y se busca otra explicación: se rechaza la H_o. El número de colas de rechazo se consigna en la H_a.**

190 ***Ejemplo: estudiante engreído. Test sobre una media con σ conocida*** *Un estudiante exhibe con orgullo su 3.2 en un examen en el cual las notas se distribuyen normalmente con media 2.3 y desviación 0.3. ¿Se justifica su jactancia?*

Paso 1: Formulamos lo que se cree (esperamos que el mundo se comporte como de costumbre, como la población). $H_o : \mu = 2{,}3$ y los datos experimentales deben estar por ahí cerca, sabiendo que la desviación vale 0.3. Es decir, todo el mundo debería sacar 2.3 de nota excepto por cosas del azar.

$H_a : \mu > 2{,}3$: además del azar, debe haber algún efecto sistemático que hace que al estudiante le vaya mejor que a la gran mayoría de la clase.

Paso 2: Lo que se ve: la nota en el examen fue de 3.2. La desviación poblacional: $\sigma = 0{,}3$.

Paso 3: Medimos la discrepancia observada entre lo que se ve y lo que se cree según la H_o. Para ello usamos una z:

$$z_{obs} = \frac{x-\mu}{\sigma} = \frac{3{,}2-2{,}3}{0{,}3} = \frac{0{,}9}{0{,}3} = 3.$$

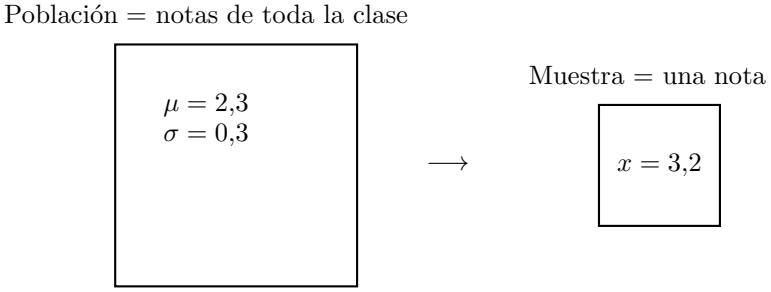


Figura 9.0. Hay una discrepancia entre la media de la población μ y el valor x: ¿se debe dicha discrepancia al azar, a la variación intrínseca de la población, o a algo sistemático?

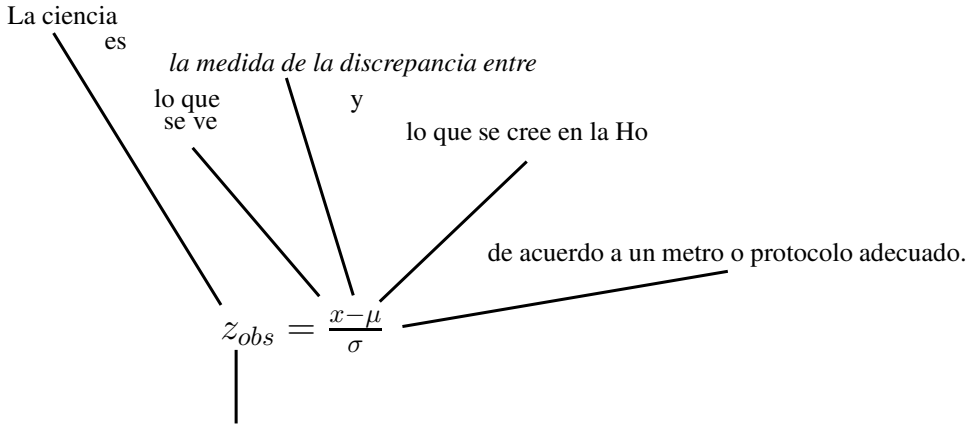


Figura 9.1. ¿Qué es la ciencia? Es básicamente la medida de la discrepancia entre lo que se ve y lo que se cree. Dicha discrepancia se mide con un estadígrafo de contraste que cuantifica la diferencia entre lo que se ve y lo que se cree de acuerdo a un metro o protocolo adecuado.

Paso 4: Decidimos si la discrepancia 3 es pequeña o grande.

La discrepancia crítica z_c con significancia 0.05 y cola superior es 1.64. Si la discrepancia observada es más extrema que el valor crítico, la discrepancia se considera grande y hay que repensar lo que se creía. Se rechaza la H_o. Es decir, no es cómodo explicar la discrepancia entre el dato observado y lo propuesto por la H_o diciendo que son cosas del azar. Pero la discrepancia es pequeña si es menos extrema que el valor crítico, y en dicho caso no hay necesidad de repensar lo que se creía en la H_o, la cual se acepta: la discrepancia entre lo que se ve y lo que se cree es pequeña y se considera fruto del azar, resultado de la interferencia de factores no controlados.

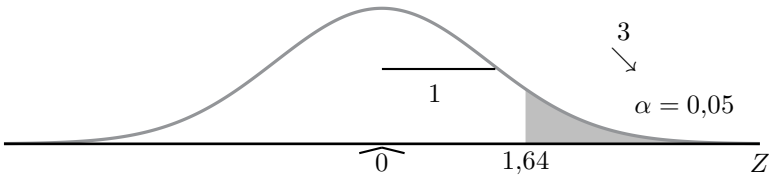


Figura 9.1. La discrepancia observada entre lo que se ve, que hay un dato con valor 3.2, y lo que se cree en la H_o, que la media vale 2.3, es $z_{obs} = 3$. Con la cola superior y $\alpha = 0{,}05$, el valor crítico de z es 1.64. Como la discrepancia $z_{obs} = 3$ es más extrema que el valor crítico decidimos que la discrepancia es grande y prometemos repensar lo que se creía: rechazamos la H_o.

Paso 5: Veredicto: Como la discrepancia $z_{obs} = 3$ es más extrema que 1.64, decidimos que la discrepancia es grande y la H_o se rechaza. Verbalicemos todo muy explícitamente. Tenemos que elegir entre dos plausibles explicaciones del dato experimental cuya discrepancia con respecto a la media es 3:

1. H_o : de no haber sido por el azar, el estudiante habría sacado el promedio de la clase, 2.3.

2. H_a : hay un factor sistemático que hace que al estudiante le vaya mejor que a la gran mayoría de la clase.

Entre las dos explicaciones elegimos gustosos la segunda: el muchacho sabe hacer que le vaya definitivamente mejor que a la gran mayoría. Tolerémosle su jactancia. Pero, si él pudo, ¿por qué nosotros no? ¡Claro que también podemos!

191 ***Ejemplo: el estudiante frustrado*** *Un estudiante no dijo absolutamente nada, pero con sólo mirarlo a la cara uno ya sabe que se siente muy frustrado al obtener una nota de 3.2 en un examen en el cual las notas se distribuyen normalmente con media 2.3 y desviación 0.3. Como que se siente muy bruto con tal nota. ¿Es sensato lo que el muchacho parece sentir?*

Solución: El estudiante parece ser solipsista, es decir que se fija metas como si fuese el único ser del universo. Y como que es de los que anhela la excelencia académica. Entonces su auto-calificación es algo personal sobre lo cual la estadística nada puede decir. Las metas solipsistas son importantes y definen la personalidad, las cosas etéreas que lo hacen a uno único, especial, irrepetible. Con todo, hay objetivos que conviene enmarcarlos dentro del contexto social. Si para alguien las notas en la U son del tipo social, entonces tengamos en cuenta que el examen al que se refiere el muchacho tuvo un promedio de 2.3 y que por tradición se sabe que la desviación es de 0.3, con distribución normal. Y ahora sí podemos estudiar estadísticamente el problema:

Paso 1: Formulamos lo que se cree (esperamos que el mundo se comporte como de costumbre, como la población). $H_o : \mu = 2{,}3$ y los datos experimentales deben estar por ahí cerca, sabiendo que la desviación vale 0.3. Es decir, todo el mundo debería sacar 2.3 de nota excepto por cosas del azar.

$H_a : \mu < 2{,}3$, lo cual se lee así: además del azar, debe haber algún efecto sistemático que hace que al estudiante le vaya peor que a la gran mayoría de la clase. Observemos que estamos asumiendo una cierta correlación entre la nota y los sentimientos: nota baja, sentimiento bajo; nota alta sentimiento alto.

Paso 2: Lo que se ve: la nota en el examen fue de 3.2. La desviación es 0.3 y se refiere a una enorme historia, por lo que podemos decir que se trata de una desviación poblacional: $\sigma = 0{,}3$.

Paso 3: Medimos la discrepancia obseravada entre lo que se ve y lo que se cree según la H_o. Para ello usamos una z:

$z_{obs} = \frac{x-\mu}{\sigma} = \frac{3{,}2-2{,}3}{0{,}3} = \frac{0{,}9}{0{,}3} = 3$.

Paso 4: Decidimos si la discrepancia 3 es pequeña o grande. Para ello tomamos la significancia 0.05 con la cola inferior que es -1.64.

Si la discrepancia está por debajo de ese valor, la discrepancia es más extrema y se rechaza la H_o: no es cómodo explicar el dato experimental diciendo que de no ser por el azar sería 2.3. Si la discrepancia es mayor que -1.64, se considera fruto del azar.

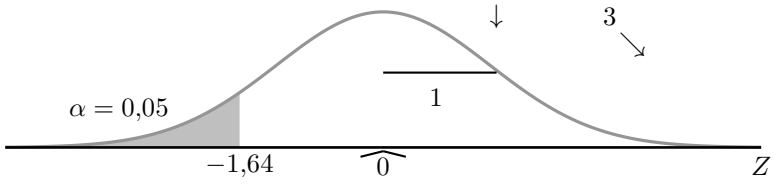


Figura 9.2. Si $H_o : \mu = 0$ en $Z \sim N(0,1)$, con la cola inferior y $\alpha = 0{,}05$, la zona de aceptación de la H_o es lo que está arriba de -1.6. Como el dato experimental con $z = 3$ está dentro de tal zona, aceptamos la H_o.

Paso 5: Veredicto: Como la discrepancia $z = 3$ está por encima de -1.6, decidimos que la discrepancia es normal y la H_o se acepta. Verbalicemos todo muy explícitamente. Tenemos que elegir entre dos plausibles explicaciones del dato experimental cuya discrepancia con respecto a la media es 3:

1. H_o : de no haber sido por el azar, el estudiante habría sacado el promedio de la clase, 2.3.

2. H_a : hay un factor sistemático que hace que al estudiante le vaya mal, más mal que a la gran mayoría de la clase.

Por supuesto que entre las dos explicaciones no nos queda más remedio que elegir la primera. Y es efectivamente lo que todo el mundo hace, pues si uno fuese amigo de ese muchacho, lo estaría recriminando por sus ideas tan excéntricas de sentirse mal cuando le fue mejor que a todo el mundo.

192 ***Ejemplo: bien preparado + TLC. Test sobre una media con $\bar{X}_n$ y σ conocidas.*** *El historial de notas de un estudiante se describe por $X \sim N(3.8,\ 0.3)$. Y de pronto se le ocurrió que quiere hacer un posgrado, y por ello ha decidido mejorar su método de estudio para subir su promedio. En las 6 materias de este semestre obtuvo un promedio de 4. Evidentemente, subió el promedio. Pero queda una duda: ¿la mejoría fue por azar o por haber mejorado su método de estudio y su disciplina?*

Lo primero que hacemos para resolver la duda es plantear la creencia fundamental que vamos a poner a prueba:

Paso 1: ¿Qué es lo que se cree? Que el mundo se comporte como de costumbre, o como la gran mayoría. $H_o : \mu = 3{,}8$, es decir, las notas deberían ser 3.8 excepto por cosas del azar, el cual se refleja en que la desviación no es cero sino 0,3.

$H_a : \mu > 3{,}8$, lo cual se lee así: además del azar, debe haber algún efecto sistemático que hace que al estudiante le vaya ahora mejor que antes. Puede ser su mejor método de estudio, haciendo los ejercicios a tiempo para poder hacer preguntas y resolver las dudas.

Paso 2: Lo que se ve: el promedio de 6 materias fue 4. La desviación poblacional es: $\sigma = 0{,}3$.

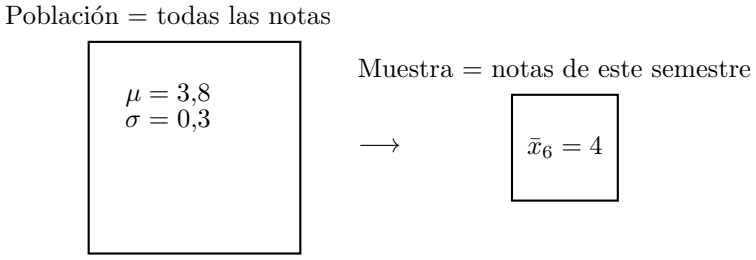


Figura 9.3. Hay una discrepancia entre la media de la población μ y la media de la muestra $\bar{x}_6$: ¿se debe dicha discrepancia al azar, a la variación intrínseca de la población, o a algo sistemático?

Paso 3: Medimos la discrepancia observada entre lo que se ve y lo que se cree según la H_o. Para ello usamos una z, pero poniendo atención a que la desviación que hay que tomar es la correspondiente al mundo donde se mueven los promedios de 6 materias, es decir, la desviación de $\bar{X}_6$, la cual por el TLC es $\frac{0{,}3}{\sqrt{6}} = 0{,}12$

$z_{obs} = \frac{\bar{X}_6 - \mu}{\sigma_{\bar{X}_6}} = \frac{4-3{,}8}{0{,}12} = \frac{0{,}2}{0{,}12} = 1{,}66.$

Paso 4: Decidimos si la discrepancia 1.66 es pequeña o grande. Para ello tomamos la significancia 0.05, que con la cola superior da un z = 1.64.

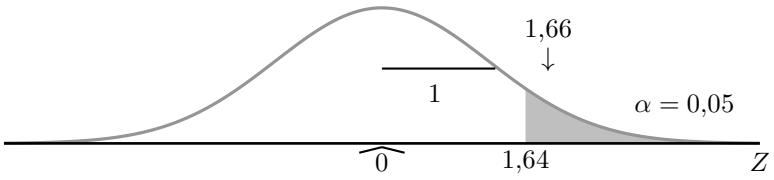


Figura 9.4. Si $H_o : \mu = 0$ en $Z \sim N(0,1)$, con la cola superior y $\alpha = 0{,}05$, la zona de aceptación de la H_o es lo que está abajo de 1.64. Como el dato experimental es $z_{obs} = 1{,}66$ está fuera de tal zona, rechazamos la H_o.

Si la discrepancia es más extrema que dicho valor, la discrepancia es grande y se rechaza la H_o: no es cómodo explicar el dato experimental diciendo que de no ser por el azar sería 3.8. Si la discrepancia es menor que 1.64, se considera fruto del azar.

Paso 5: Veredicto: Como la discrepancia $z_{obs} = 1.66$ está por encima de 1.64, decidimos que la discrepancia es grande, anormal y la H_o se rechaza. Verbalicemos todo muy explícitamente. Tenemos que elegir entre dos plausibles explicaciones del dato experimental cuya discrepancia con respecto a la media es 1.66:

1. H_o : de no haber sido por el azar, el estudiante habría sacado su promedio habitual, 3.8.
2. H_a : hay un factor sistemático que hace que al estudiante le vaya mejor que de costumbre. No en vano se hizo amigo de Lucía, que no es tan espigada que digamos pero es inteligente y muy disciplinada para estudiar. Es tan inteligente que ella no lo interrumpe cuando el habla.

Por supuesto que entre las dos explicaciones, elegimos gustosos la segunda.

193 ***Ejemplo: la mejor clase + la t. Test sobre una media con $\bar{X}_n$ y s conocidas.*** *Una clase anda pregonando que está sobrada pues su promedio entre 16 estudiantes es de 3.2 y desviación 0.3 en una materia en la cual las notas se distribuyen normalmente con media 2.3. ¿Se justifica su jactancia?*

Notamos que acá **no tenemos σ, la desviación poblacional, la de todo el universo, sino la desviación muestral,** s. Por lo tanto, no podemos usar el TLC. Usamos una pequeña variación, en la cual en vez de usar la z y σ usamos la t y s.

Paso 1: Esperamos que $H_o : \mu = 2{,}3$ y los datos experimentales deben estar por ahí cerca. Es decir, todas las clases deberían sacar en promedio 2.3 excepto por cosas del azar.
$H_a : \mu > 2{,}3$, lo cual se lee así: además del azar, debe haber algún efecto sistemático que hace que a esta clase le vaya mejor que a la gran mayoría de las clases. Tenemos un caso con una cola, la superior.

Paso 2: Lo que se ve: la nota promedio entre 16 estudiantes en el examen fue de 3.2. La desviación muestral fue : $s = 0{,}3$. Asumimos que la desviación hallada es típica de todas las clases.

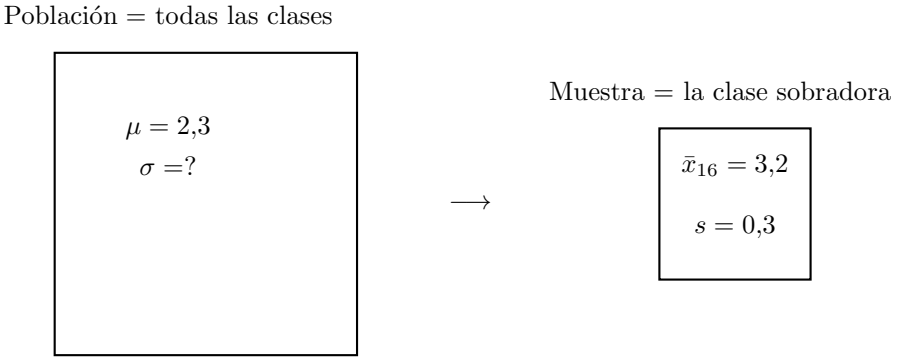


Figura 9.5. Hay una discrepancia entre la media de la población μ y la media de la muestra $\bar{x}_{16}$. Esta vez la muestra salió favorecida con una media muestral mayor que la media poblacional. ¿Qué pasará si volvemos a repetir el experimento? ¿Se conservará dicho favorecimiento?

Paso 3: Medimos la discrepancia observada entre lo que se ve y lo que se cree según la H_o. El problema es que no podemos usar la z de los ejemplos anteriores
$z_{obs} = \frac{\bar{X}_{16} - \mu}{\sigma_{\bar{X}_{16}}}$
porque no tenemos la desviación poblacional, la cual se obtiene por teoría o por el estudio de una muestra gigante. Pero nosotros sólo tenemos 16 datos. Por ello, nos vemos obligados a usar un estadígrafo de contraste diferente. Se aplica un **TLC modificado** que usa un estadígrafo denominado t:

$$t_{obs} = \frac{\bar{X}_n - \mu}{s/\sqrt{n}}$$

Observemos que en el denominador no va la desviación poblacional sino la muestral, la de la muestra, s. Ahora bien, en una distribución de tendencia central, es improbable que en una muestra pequeña salgan valores extremos, por lo tanto, la desviación de la muestra será menor que la desviación de la

población, la desviación real, la desviación de una muestra gigante. Como consecuencia, el quebrado de la t será mayor que el de la z. Por eso, la t tiene un histograma (función densidad) más explayado que el de la z. Sin embargo, a medida que el número de datos de la muestra se hace mayor, la desviación muestral tiende a aproximarse a la desviación poblacional y por tanto para un número grande de glis no hay mayor diferencia entre la t y la z. Por eso se dice en la literatura que **la t se usa como estadígrafo de contraste para el estudio de la media de muestras pequeñas**. Es usual considerar que no hay diferencia entre la z y la t a partir de 30 glis. Dicho con más claridad, cuando la muestra contiene más de 30 datos independientes, da lo mismo usar la z que la t por lo que se usa la z directamente, lo cual permite usar nuestras tablas para trabajar con cualquier significancia.

Para nuestro caso tenemos que la discrepancia entre lo que se ve y lo que se cree da:

$t_{obs} = \frac{\bar{X}_{16}-\mu}{s/\sqrt{16}} = \frac{3{,}2-2{,}3}{0{,}3/\sqrt{16}} = \frac{0{,}9}{0{,}3/4} = 12.$

Paso 4: Decidimos si la discrepancia 12 es pequeña o grande, lo cual se hace comparando la discrepancia hallada con el valor crítico dado por **la tabla de la t (pág. 158 y siguientes)**. Tomamos la significancia por defecto, $\alpha = 0{,}05$, con una cola (la superior). Miramos el código de interpretación de la tabla: calculamos $\beta = 0{,}10$, pues por tratarse de una cola es igual a 2α. Pero al ir a mirar en la tabla de la t, notamos que en dicha tabla hay una parámetro llamado **gli** (grados de libertad) que es igual al número de datos menos uno y que indica que entre más datos, más preciso es nuestro experimento y más contundente puede ser nuestra conclusión. Vamos entonces a la primera columna, la que tiene $\beta = 0{,}10$, y en gli = 16-1 = 15 obtenemos que la t crítica vale 1.753. Si la discrepancia es más extrema que ese valor, la discrepancia es grande y se rechaza la H_o: no es cómodo explicar el dato experimental diciendo que de no ser por el azar sería 2.3. Si la discrepancia experimental, dada por la t es menor que 1.753, se considera fruto del azar.

Paso 5: Veredicto: Como la discrepancia $t_{obs} = 12$ es más extrema que 1.753, decidimos que la discrepancia es muy grande y la H_o se rechaza.

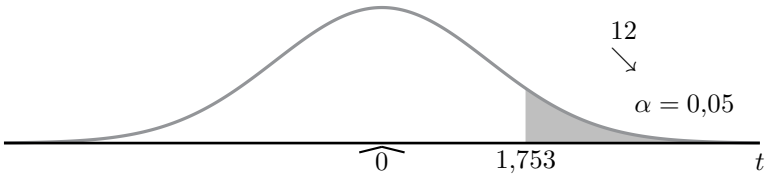


Figura 9.6. Si $H_o : \mu = 0$, con la cola superior y $\alpha = 0{,}05$, la zona de aceptación de la H_o es lo que está abajo de 1.753. Como el dato experimental con $t = 12$ está afuera de tal zona, rechazamos la H_o.

Verbalicemos todo muy explícitamente. Tenemos que elegir entre dos plausibles explicaciones del dato experimental cuya discrepancia con respecto a la media es 12:

1. H_o : de no haber sido por el azar, la clase habría sacado el promedio de 2.3.
2. H_a : hay un factor sistemático que hace que a esta clase le vaya mejor que a la gran mayoría de clases.

Entre las dos explicaciones elegimos, con algo de envidia, la segunda: a esta clase le fue definitivamente bien, mejor que a la gran mayoría, y no fue por azar. Eso significa que si volviésemos a repetir el experimento, a esta clase le volvería a ir definitivamente mejor que al promedio. Claro que existe un riesgo de que nos equivoquemos, el cual está dado por la significancia.

194 ***Ejemplo: proporción de mujeres afectadas. Test sobre una proporción.*** *La proporción de mujeres que consideran que un cierto fenómeno político en el Oriente Medio las afecta es 0.08 sobre una muestra de 3600 de estrato 3. Esto contrasta con lo que se esperaba después de una campaña publicitaria de los inmigrantes palestinos y libaneses, que debía ser 0.2 o más. ¿Se debe la diferencia entre lo que se ve y lo que se cree al azar o es mejor atribuirla al efecto de la propaganda?*

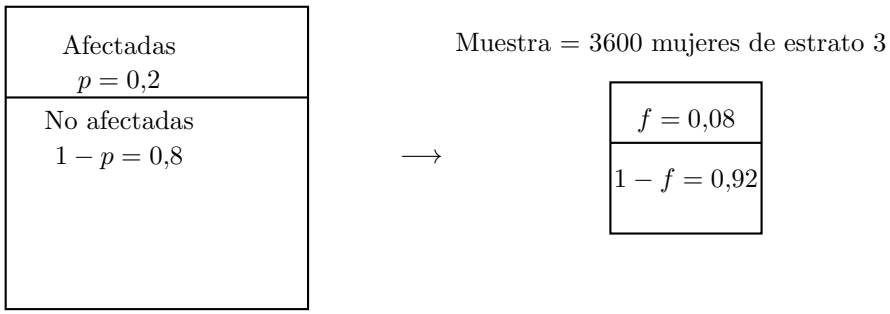


Figura 9.7. Hay una discrepancia entre la probabilidad esperada de ser afectada, $p = 0{,}2$, y la proporción $f = 0{,}08$ de mujeres afectadas en una muestra de 3600. ¿Qué pasará si volvemos a repetir el muestreo? ¿Se conservará la misma relación entre proporción observada y probabilidad?

Sea p la probabilidad de que una mujer sea afectada por los eventos en el Oriente Medio.

Paso 1: Formulamos lo que se cree. $H_o : p = 0{,}2$. De no ser por el azar en el muestreo, la proporción de afectadas después de la campaña debe ser 0.2.

$H_a : p < 0{,}2$, lo cual se lee así: además del azar, debe haber algún efecto sistemático (una enorme insensibilidad) que hace que la probabilidad de encontrar una mujer sensible a los eventos del Oriente Medio sea inferior a 0,2.

Paso 2: Lo que se ve: tamaño de la muestra $n = 3600$. La proporción observada para mujeres sensibles es $f = X/n = 0{,}08$ donde X es la cantidad de mujeres afectadas que es $X = 0{,}08 \times 3600 = 288$. Atención: **la proporción observada es una frecuencia relativa y por eso está entre 0 y 1.**

Paso 3: Medimos la discrepancia observada entre lo que se ve y lo que se cree según la H_o. Para ello podemos usar como estadígrafo de contraste a la z como sigue:

$$z_{obs} = \frac{f-p}{\sqrt{\frac{p(1-p)}{n}}}$$

$$z_{obs} = \frac{0{,}08-0{,}2}{\sqrt{\frac{0{,}2(1-0{,}2)}{3600}}} = \frac{-0{,}12}{\sqrt{\frac{0{,}2(0{,}8)}{3600}}} = \frac{-0{,}12}{0{,}00666} = -18$$

Usamos una z puesto que al aproximar la binomial por la normal, la desviación que ponemos viene de una espectativa, es teórica, es poblacional, y no de una muestra (si fuese de una muestra, pondríamos la t).

Paso 4: Decidimos si la discrepancia -18 es pequeña o grande. Para ello tomamos la significancia 0.05, que con la cola inferior da $z = -1{,}64$. Si la discrepancia es más extrema que ese valor, la discrepancia es grande y se rechaza la H_o. Si la discrepancia está arriba de -1.6, se considera fruto del azar.

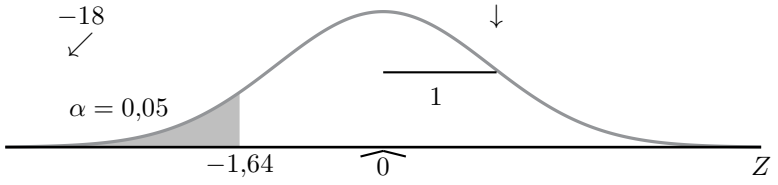


Figura 9.8. Si $H_o : \mu = 0$ en $Z \sim N(0,1)$, con la cola inferior y $\alpha = 0{,}05$, la zona de aceptación de la H_o es lo que está arriba de -1.64. Como el dato experimental con $z = -18$ está por fuera de tal zona, rechazamos la H_o.

Paso 5: Veredicto: Como la discrepancia $z = -18$ es más extrema que -1.64, decidimos que la discrepancia es anormal y la H_o se rechaza. Verbalicemos todo muy explícitamente. Tenemos que elegir entre dos plausibles explicaciones del dato experimental cuya discrepancia con respecto a la media es -18:

1. H_o : de no haber sido por el azar, la proporción de afectadas sería 0.2 o más.

2. H_a : hay un factor sistemático, una gran indiferencia, que hace que la campaña no haya despertado al nivel esperado la sensibilidad a los problemas en el oriente medio.

Nosotros optamos por la segunda explicación: la propaganda hecha no logró su objetivo entre las mujeres de estrato 3.

195 ***Ejemplo: comparación de proporciones de hombres y mujeres afectados.***

Test sobre comparación de dos proporciones. *Una muestra de 3600 mujeres de estrato 3 produjo 288 mujeres que consideran que un cierto fenómeno político en el Oriente Medio los afecta, en tanto que de otra muestra independiente de 3000 hombres del mismo estrato, 216 se consideran afectados. ¿Tienen las mujeres y los hombres diferente grado de sensibilidad a los problemas en el oriente medio?*

Sea p la probabilidad de que una mujer sea afectada por los eventos en el Oriente Medio. Sea r la probabilidad de que un hombre sea afectado por los eventos en el Oriente Medio. Sea $X = 288$ el número de mujeres afectadas y $Y = 216$ el número de hombres afectados. La muestra de mujeres es $m = 3600$, la de hombres es $n = 3000$.

Paso 1: Formulamos lo que se cree. $H_o : p - r = 0$. De no ser por el azar, la proporción de afectados no depende de ser hombre o mujer.

$H_a : p - r \neq 0$, lo cual se lee así: además del azar, debe haber algún efecto sistemático que hace que las mujeres y los hombres tengan diferente grado de sensibilidad a los fenómenos del Oriente Medio.

Paso 2: Lo que se ve: la proporción observada para mujeres $f = \frac{X}{m} = \frac{288}{3600} = 0{,}08$, la proporción observada para hombres $g = \frac{Y}{n} = \frac{216}{3000} = 0{,}07$. Cantidad de mujeres $m = 3600$. Cantidad de hombres $n = 3000$. La diferencia observada de proporciones es $f - g = 0{,}08 - 0{,}07 = 0{,}01$. La proporción global observada de afectados es

$$\hat{p} = \frac{X+Y}{m+n} = \frac{288+216}{6600} = 0,07$$

$1 - \hat{p} = 1 - 0{,}07 = 0,93 =$ proporción global observada de no afectados.

También se puede usar la fórmula alternativa

$$\hat{p} = \frac{fm+gn}{m+n} = \frac{0{,}08(3600)+0{,}07(3000)}{6600} = 0,07$$

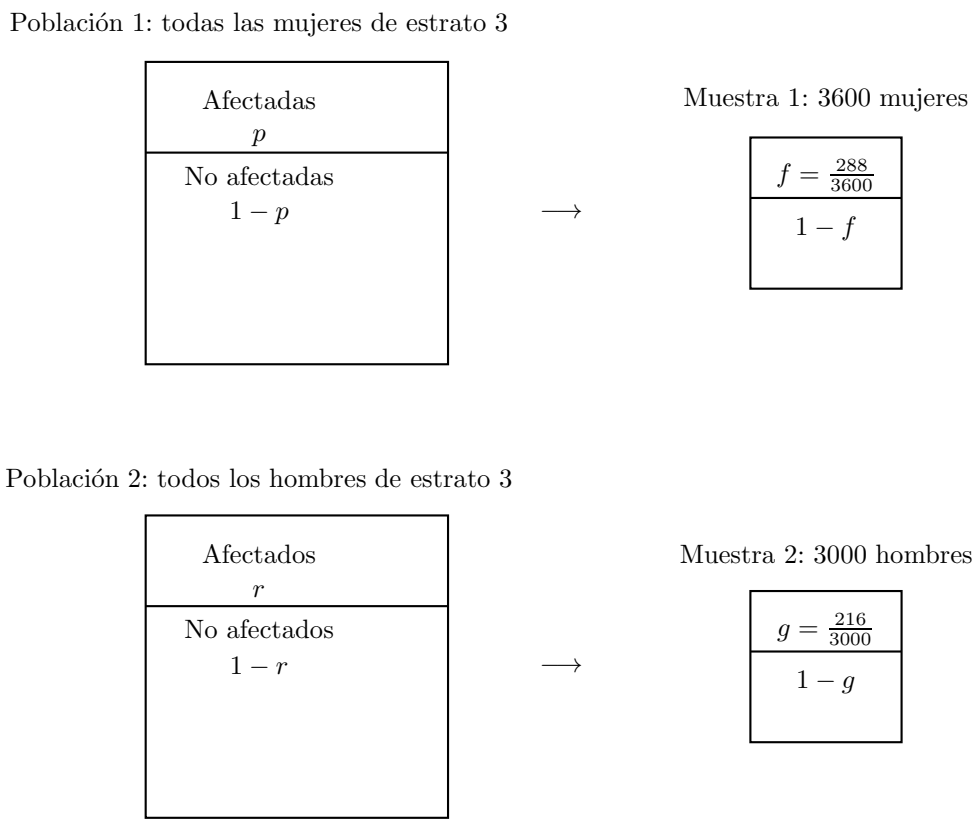


Figura 9.9. Hay una ligera discrepancia entre las proporciones observadas de afectados entre hombres y mujeres: $f - g \neq 0$. ¿Se debe dicha discrepancia al efecto azaroso del muestreo o a que las mujeres y los hombres tienen diferente probabilidad de ser afectados? Es decir, ¿podemos alegar que $p - r \neq 0$?

Paso 3: Medimos la discrepancia observada entre lo que se ve y lo que se cree según la H_o. Como la binomial se aproxima por una normal, podemos usar una z cuando la varianza se toma poblacional o una t cuando la varianza viene de las muestras:

$$t_{obs} = \frac{(f-g)-(p-r)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{m}+\frac{1}{n})}} = \frac{0{,}01-0}{\sqrt{0{,}07(0{,}93)(\frac{1}{3600}+\frac{1}{3000})}} = \frac{0{,}01}{0{,}0063074} = 1{,}58$$

Tomamos una t puesto que la varianza usada viene de muestreos. Los glis son $n+m-2$ que da 6598. Como eso es mucho mayor que 30, la t se aproxima por la z.

Paso 4: Decidimos si la discrepancia 1.58 es pequeña o grande. Para ello tomamos la significancia 0.05, que con dos colas da $z = 1{,}96$. Si la discrepancia está por encima de ese valor, la discrepancia es grande y se rechaza la H_o. Si la discrepancia está entre -1.96 y 1.96, se considera fruto del azar.

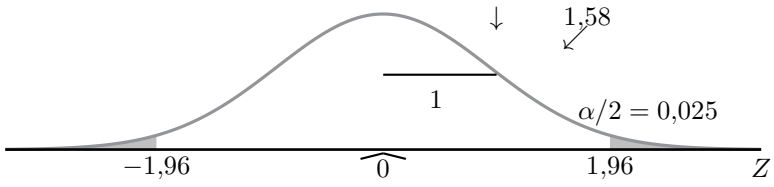


Figura 9.10. La discrepancia debida al azar se distribuye como $Z \sim N(0,1)$, lo normal es lo que va de -1.96 a 1.96. La discrepancia 1.58 entre lo predicho por la H_o y lo observado está dentro de lo normal. Aceptamos la H_o: el azar es suficiente para explicar la diferencia observada.

Paso 5: Veredicto: Como la discrepancia $z = 1{,}58$ es menos extrema que 1.96, decidimos que la discrepancia es normal y la H_o se acepta. Verbalicemos todo muy explícitamente. Tenemos que elegir entre dos plausibles explicaciones del dato experimental cuya discrepancia con respecto a lo esperado es 1.58:

1. H_o : de no haber sido por el azar, las proporciones de afectados serían las mismas tanto en hombres como en mujeres.
2. H_a : hay un factor sistemático que hace que la sensibilidad a los problemas en el oriente medio dependa del sexo. Ejemplo: ahora hay mucho reporte que muestra el dolor de la mujer palestina.

Nosotros optamos por la primera explicación: la diferencia observada en el grado de sensibilidad entre mujeres y hombres no se explica por un factor que crea efectos sistemáticos sino por el azar asociado al muestreo. No se excluye que estemos equivocados, pero se requerirían más datos para probarlo.

196 ***Ejemplo: comparación de medias para datos apareados, ojo derecho vs. ojo izquierdo.*** *Se dice que* ***los datos son apareados*** *cuando la manera natural de mirarlos y estudiarlos es por pares, como el ojo derecho y el ojo izquierdo, el esposo y la esposa, el presidente y el vicepresidente. Aunque sean apareados, deben tener algo que los haga independientes.*

La dioptría es la unidad de medida de la potencia de los lentes que usamos para ver mejor. Entre más dioptrías, más poderoso es el lente y más amplifica. Si uno usa gafas, estas pueden tener lentes con diversas dioptrías para cada ojo. Nos interesa saber si hay diferencia promedio entre el número de dioptrías del ojo de descanso vs. el ojo de trabajo de los relojeros, pues ellos usan un mono-ocular.

Sea X las dioptrías en el ojo de descanso, Y las dioptrías en el ojo de trabajo. La diferencia experimental con su signo entre las dioptrias del primer ojo y del segundo de LA MISMA PERSONA se denotan $D_i = X_i - Y_i$. Los D_i tienen una media $\bar{D}$ y una desviación s.

Dioptrías lente ojo de descanso, X, vs. ojo de trabajo, Y.

Paciente	X	Y	$D = X - Y$	D^2
Juan	0	1	-1	1
Pedro	1	0	1	1
José	1	2	-1	1
Luis	2	2	0	0
Alberto	1	0	1	1
Felipe	1	0	1	1
			$\sum D = 1$	$\sum D^2 = 5$

Solución: Este problema compara medias de datos apareados. Para este tipo de datos hay un tratamiento especial y distinto del que se da a datos no apareados. Sea $\bar{\Delta}$ = *delta barra* **la diferencia promedio poblacional para datos apareados** entre las dioptrías del ojo de descanso y el ojo de trabajo del mismo paciente.

Paso 1: Formulamos lo que se cree. $H_o : \bar{\Delta} = 0$, hacer trabajar al mismo ojo mientras que el otro descansa no causa, en promedio, ningún trauma adicional. La diferencia promedio predicha para una muestra muy grande de dioptrías de un ojo y del otro es cero.

$H_a : \bar{\Delta} \neq 0$, hacer trabajar al mismo ojo mientras que el otro descansa causa, en promedio, que un ojo quede diferente que el otro. Se decide con dos colas, pues preguntan sobre la diferencia entre los dos ojos. Como no dijeron nada de la significancia, tomamos $\alpha = 0{,}05$.

Paso 2: dato experimental: $\sum D_i = 1$ y hay $n = 6$ pares de datos, entonces $\bar{D} = 1/6$. Y como $\sum D_i^2 = 5$ entonces la desviación de D es:

$$s = \sqrt{\frac{n(\Sigma D^2) - (\Sigma D)^2}{n(n-1)}}$$
$$s = \sqrt{\frac{6(5)-1}{6(5)}} = \sqrt{\frac{29}{30}} = 0{,}983$$

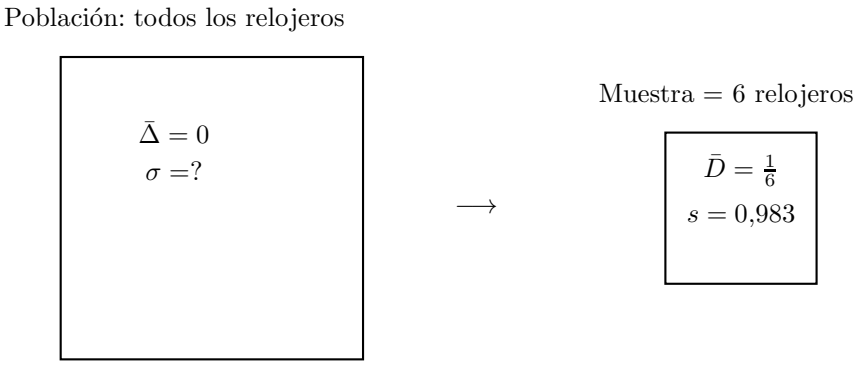


Figura 9.11. Hay una diferencia observada no nula $\bar{D} = \frac{1}{6}$ entre las dioptrías del ojo de descanso y del de trabajo en una muestra al azar de 6 relojeros: ¿Indica dicha diferencia que el muestreo fue muy tacaño y que el azar produjo fuerte intereferencia o que el ojo de trabajo se lesiona más que el de descanso?

Paso 3: Medimos la discrepancia entre lo que se ve y lo que se cree según la H_o. Lo esperado es $\bar{\Delta} = 0$. Lo observado es $\bar{D} = 1/6$.

Nuestro medidor de discrepancia entre lo hallado experimentalmente y lo esperado según la H_o es:

$$t_{obs} = \frac{\bar{D} - \bar{\Delta}}{\frac{s}{\sqrt{n}}}$$

$$t_{obs} = \frac{1/6-0}{0{,}983/\sqrt{6}} = \frac{1/6}{0{,}983/\sqrt{6}} \sim \frac{1/6}{1/\sqrt{6}} = \frac{\sqrt{6}}{6} = 0{,}415$$

Usamos una t y no una z porque se trata de una desviación muestral y no poblacional.

Paso 4: Decidimos si la discrepancia entre lo esperado y lo hallado es grande o chica. Para ello consultamos la tabla de la t (pág 158). Pero para ello necesitamos saber los grados de libertad, que para datos apareados son igual a número de pares de datos menos uno. Los grados de libertad miden la confiabilidad del experimento: entre más datos, más grados de libertad, mayor confiabilidad.

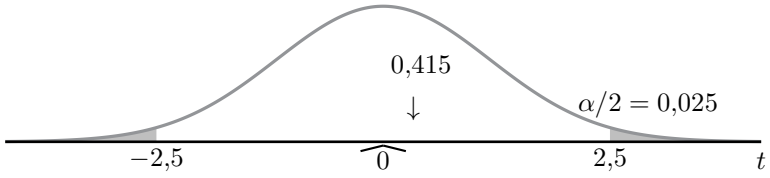


Figura 9.12. La discrepancia debida al azar se distribuye como una t con 5 g.l, lo normal es lo que va de -2.5 a 2.5. La discrepancia 0.415 entre lo predicho por la H_o y lo observado es pequeña. Aceptamos la H_o: no hay diferencia significativa en la población entre los traumas del ojo de trabajo y del de descanso.

En el presente caso, los gli son 6-1 = 5. Tenemos que el valor de la t crítica para dos colas con 5 gli y $\alpha = 0{,}05$ es $t_5 = 2{,}5$. Por tanto, la diferencia registrada, 0.415, es pequeña.

Paso 5: Tomamos un veredicto:

La H_o se acepta.

Paso 6: Interpretación: Nuestra muestra no nos ofrece razón para pensar que los promedios de dioptrías de ambos ojos sean diferentes. Sin embargo, nuestra conclusión parece basada más en la flojera que en otra cosa: se recomienda tomar más datos.

197 ***Ejemplo: los pollitos, comparación de varianzas no por resta sino por división.***
Las futuras mamás tienen la manía de sobarse la barriga e incluso de hacerlo con la barriga al aire. En parte por acariciarse ellas mismas, en parte por acariciar al bebé. Y en parte porque están programadas genéticamente para hacerlo, lo cual es necesario para que el bebé nazca sin defectos de visión. Eso lo sabemos gracias a experimentos con pollitos:

Un pollito intenta seguir el ejemplo de su madre y trata de comer semillas que se esparcen en la tierra. Parece simple. Pues no, no lo es: en primer término, el pollito debe discriminar entre lo que es una semilla y lo que no lo es. En segundo lugar, el pollito recogerá con su pico las semillas seleccionadas. Así que debe apuntar, pegarle, sincronizar la agarrudura del pico y tragar. Y si no lo hace, se muere. Es un reto a muerte.

De hecho, la historia empieza cuando el pollito todavía está en el huevo. El problema es que la información genética para construir el ojo no se especifica completamente. La razón es que se necesita un ajuste fino y armonioso entre los diversos componentes del ojo para que la luz quede enfocada donde toque. Sin dicha armonía, el pollito se vuelve miope o hipermiope.

Dicha regulación se logra como sigue: cuando el embrión aún está en la cáscara, la luz entra e ilumina el ojo en formación, y éste ajusta su desarrollo para que la luz quede perfectamente enfocada donde toca. Así se garantiza que al nacer, el pollito podrá ver con la nitidez necesaria para poder alimentarse.

¡Ajá! ¿Y cómo sabremos que esto no es simple carreta? Pues comparando los pollitos que salieron de un grupo de huevos que fueron iluminados con otros pollitos que salieron de otro grupo que no lo fueron.

En nuestro experimento, se le ofrece un blanco móvil a un pollito y medimos la precisión de sus golpes. Nos interesa demostrar que el pollito enfoca mejor su ataque cuando su embrión ha sido iluminado que cuando no. Para ello medimos la distancia del golpe al blanco y comparamos las varianzas de dicha distancia en dos circunstancias, desarrollo embrionario a oscuras y desarrollo permitiendo iluminación. Lo que esperamos es que un desarrollo embrionario bajo iluminación produzca pollitos que pueden enfocar mucho mejor sus ojos que cuando el desarrollo es a oscuras. La varianza de la distancia entre la posición del objetivo y el punto de golpe debe ser menor en el primer caso. En el segundo, los golpes deben estar desperdigados por el mundo entero y así la varianza debe ser mayor. Es decir, significativamente mayor.

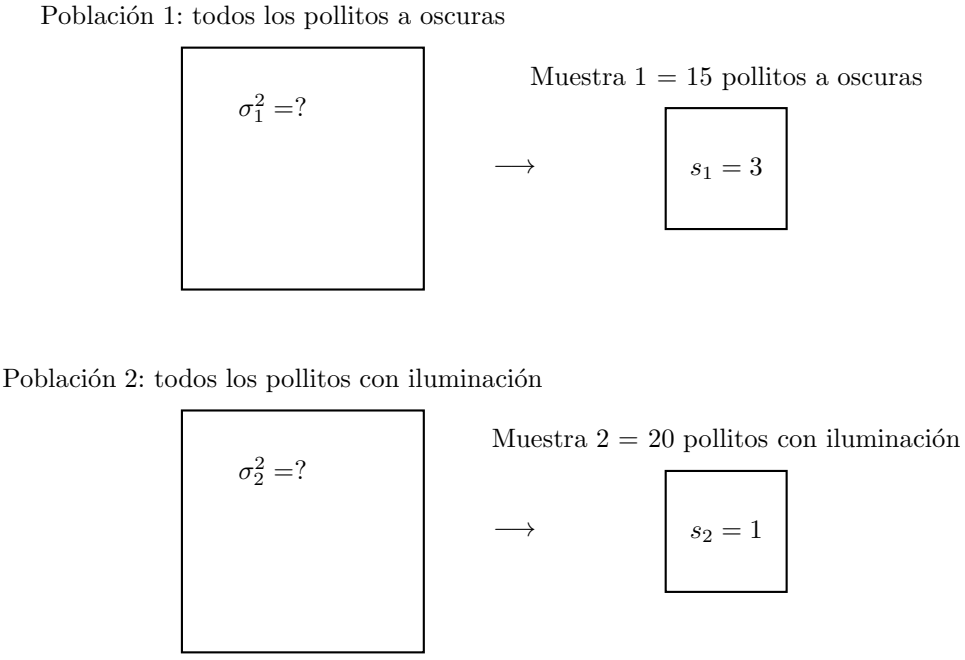


Figura 9.13. La distancia del golpe al objetivo tiene una varianza observada mayor a oscuras que con iluminación: ¿Es dicha diferencia lo suficientemente sólida como para permitirnos decir que los pollitos incubados con iluminación tienen mejor visión que los desarrollados completamente a oscuras?

En la primera circunstancia, una muestra de 15 golpes dio una desviación de 3mm y en la segunda, una muestra de 20 golpes dio una desviación de 1mm. ¿Logramos demostrar que la iluminación de los huevos influye sobre la capacidad de apuntar con el pico? Tomemos $\alpha = 0{,}01$.

Solución: estamos comparando homogeneidad, que compara varianzas (en vez de desviaciones). **Las varianzas se comparan no por resta sino por división, usando un quebrado, la varianza mayor arriba, la menor abajo.** Sea σ_1^2 la varianza poblacional, con un número infinito de datos, de la distancia del golpe al blanco bajo desarrollo a oscuras. Sea σ_2^2 la varianza poblacional de la distancia del golpe al blanco bajo desarrollo con iluminación. La división entre las varianzas poblacionales se denomina R, de relación:

$R = \frac{\sigma_1^2}{\sigma_2^2}$

Paso 1: Formulamos lo que se cree, la H_o. La hipótesis nula dice que el cambio propuesto no va a cambiar las cosas. Concretamente, que la iluminación de los huevos en nada va a cambiar la puntería del pollito, la cual se mide por la varianza de la distancia del punto de golpe al objetivo. Por tanto,

$H_o : R = \frac{\sigma_1^2}{\sigma_2^2} = 1,$

las varianzas poblacionales son iguales, la iluminación no causa efecto sistemático, repetible.

Paso 2: Lo que se ve: hay dos desviaciones muestrales, la primera tiene un valor de 3 y que viene de 15 datos. La segunda con un valor de 1 y que viene de 20 datos. Como la varianza es la desviación al cuadrado, tenemos:

$Var_1 = s^2{}_1 = 3^2 = 9$, la varianza grande con 15 datos o, lo que es lo mismo, 15-1 = 14 gli (grados de libertad).

$Var_2 = s_2^2 = 1^2 = 1$, la varianza pequeña con 20 datos o 20-1 = 19 gli

La relación entre las varianzas según el resultado observado, es

$R_{obs} = \frac{Var\ grande}{Var\ chica} = \frac{9}{1} = 9$ con 14 gli del numerador y 19 gli del denominador. Los gli dicen qué tan confiable es un experimento, pues entre más datos, más gli, mayor confiabilidad. La varianza grande siempre se pone arriba, los gli(grados de libertad) son el número de datos menos uno. Los gli de la varianza grande se llaman gli del numerador; los de la pequeña, gli del denominador.

Podemos resumir los datos en la tabla siguiente:

Varianza grande arriba						
	Descripción	s	$Var = s^2$	R_{obs}	n	gl
X	Oscuridad	3	9	9	15	14
Y	Luz	1	1		20	19

Paso 3: Medimos la discrepancia entre lo que se ve y lo que se cree según la H_o. Para ello, usamos una F de Fisher, que es una razón, un quebrado, entre lo que se ve, dado por R_{obs}, y lo que se cree en la H_o dado por R.

$F_{obs} = \frac{R_{obs}}{R} = \frac{9}{1} = 9$

Cuando la H_o es la igualdad de varianzas, la R_{obs} y la F_{obs} coinciden.

Paso 4: Determinamos la H_a. La hipótesis alterna es $H_a : R = \frac{\sigma_1^2}{\sigma_2^2} > 1$, lo cual se lee así: iluminar el embrión causa sistemáticamente un mejor desarrollo del ojo que hace que la varianza del error de puntería disminuya. A medida que el experimento funcione la varianza bajo iluminación tenderá a disminuir y como va en el denominador de R_{obs}, éste tenderá a aumentar al igual que la F_{obs}. Eso quiere decir que tenemos un test de una cola, la superior.

Paso 5: Decidimos si la discrepancia relacional F_{obs} entre lo que se ve y lo que se cree es pequeña o grande. Comparemos F_{obs} con el valor crítico F_+ de la tabla. Tomamos $\alpha = 0{,}01$ con una cola (ó $\alpha = 0{,}02$ y dos colas) y así podemos usar **la primera de las 3 tablas de la F, la que está en la pág. 160**. Hay otras 2 tablas para otras 2 combinaciones de α y número de colas.

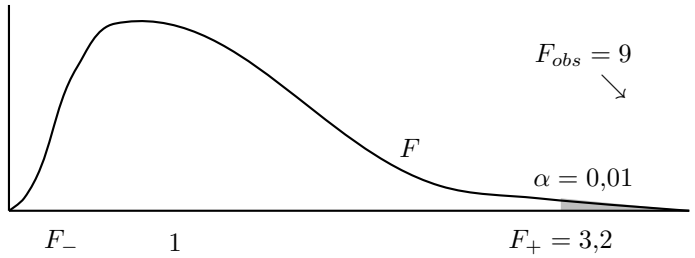


Figura 9.14. Para $\alpha = 0{,}01$ y una cola, 14 y 19 g.l, $F_+ = 3{,}2$. Una $F_{obs} = 9$ es demasiado grande, por lo que rechazamos la H_o, no creemos que la diferencia entre las varianzas muestrales se deba simplemente al azar. Postulamos que debe existir un efecto sistemático que hace que la varianza disminuya cuando los huevos son iluminados. La F no toma valores negativos porque viene de relaciones entre varianzas (que son positivas), y su moda está cerca de uno, pues está calibrada para varianzas iguales.

Los gli del numerador son 14. Buscamos un gli con indicador subrayado 14, pero no la hay. Tomamos 12. Los gli del denominador son 19. La F crítica correspondiente a 12 gli del numerador y 19 del denominador es 3.30, y sin muchos escrúpulos decidimos que para 14 y 19 la F sea 3.20. Se aproxima hacia abajo pues para 15 y 19 glis la F es 3.15.

Si la discrepancia relacional entre lo que se ve y lo que se cree en la H_o es más extrema que 3.2, la discrepancia entre lo que ve y lo que se cree es grande y se rechaza lo que se cree, la H_o. Si la discrepancia entre lo que ve y lo que se cree es menor que 3.2, se considera fruto del azar y se acepta lo que se cree, la H_o.

Paso 6: Veredicto. Como la discrepancia relacional observada $F_{obs} = 9$ está por encima de $F_+ = 3.2$, que es la F crítica, decidimos que es anormal y la H_o se rechaza.

Podemos resumir todo en la siguiente tabla:

Comparación de varianzas: varianza grande arriba					
	Descripción	s	$Var = s^2$	n	gl
X	Oscuridad	3	9	15	14
Y	Luz	1	1	20	19
R_{obs}	9				
H_o	$R = 1$				
H_a	$R > 1$				
F_{obs}	9				
α	0.01				
F_+	3.20				
Veredicto	H_o ×				

Paso 7: Interpretación. Tenemos que elegir entre dos plausibles explicaciones del dato experimental cuya discrepancia relacional con respecto a la H_o es 9:

1. H_o : de no haber sido por el azar, los factores no controlados, las dos varianzas muestrales serían iguales.

2. H_a : iluminar los huevos causa sistemáticamente que la varianza del error de puntería disminuya.

Elegimos la segunda explicación: iluminar los embriones permite la formación de un ojo altamente funcional, lo cual facilita el manejo del pico por el pollito, apuntando mejor, por lo que la varianza en el error de manejo del pico decrece sistemáticamente.

198 ***Ejemplo: germanismos.***

Comparación de medias cuando las varianzas poblacionales son iguales.

El estudio comparativo de dos escuelas de filósofos nacionales que abiertamente se declaran fuertemente influidas por el pensamiento germano tiene como objetivo contar el número de germanismos utilizados en sus escritos. Se compararon dos escuelas, la primera fue la escuela ***Nuevo Hegel*** *(filósofo alemán, maestro del idealismo:* Sein und Denke *[la existencia y el pensamiento] eran para él la misma cosa) y la segunda fue la escuela* ***Entendiendo a Wittgenstein*** *(filósofo austriaco neopositivista, que quería encuadernar la filosofía dentro de la ciencia. Este autor usa la palabra* Dinge *[cosa, objeto de la percepción, sujeto de la conversación] de una forma tan definitiva, que bien puede introducirse como germanismo). De la primera escuela se tomaron 4 libros y de cada libro se seleccionó un texto de 10000 palabras. En los 4 textos el promedio de germanismos fue* $\bar{X}_4 = 131$ *con* $Var_X = s_X^2 = 36,6$*. Por otra parte se estudiaron, con la misma técnica, 6 libros de la segunda escuela y tuvieron en promedio* $\bar{Y}_6 = 135$ *germanismos con* $Var_Y = s_Y^2 = 11,2$*. Decidamos si las dos escuelas tienen el mismo promedio de germanismos por cada 10000 palabras. Usemos* $\alpha = 0{,}02$*.*

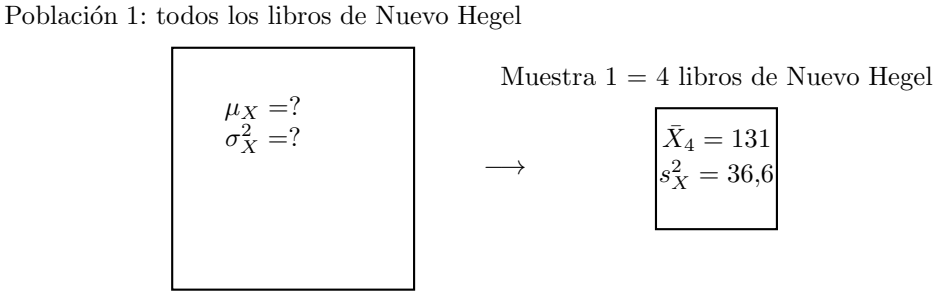


Figura 9.15. Hay una diferencia entre las medias muestrales, $\bar{X}_4 = 131$ *y* $\bar{Y}_6 = 135$*. Es lo suficientemente confiable dicha diferencia como para permitirnos decir que hay una diferencia sistemática entre el promedio de germanismos de las dos escuelas? ¿ O se deberá la diferencia observada al efecto azaroso del muestreo, predicho por las varianzas muestrales no nulas,* $s_X^2 = 36{,}6$ *y* $s_Y^2 = 11{,}2$*?*

Solución: Si la estadística no se practica, se olvida fácilmente. Hagamos por tanto un esfuerzo por poner en la mente algo que enriquece el sentido común y que aumenta la capacidad de retención y de análisis: **para poder comparar dos medias, es necesario tener en cuenta las varianzas**. Aclaremos esto. Primero que todo, uno toma dos muestras de datos cuantitativos, les saca las medias que son medias muestrales y a partir de ellas uno desea inferir si las medias de la población son iguales o diferentes. Pues bien, esto no puede hacerse a menos que se tome en consideración las varianzas. Veámoslo gráficamente:

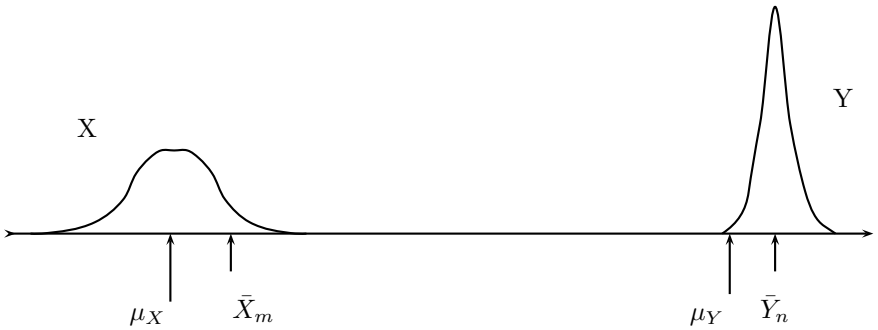


Figura 9.16. Cuando la diferencia de medias tanto poblacionales como muestrales es grande comparada con la dispersión de las campanas, la diferencia de medias es significativa: la misma relación de desigualdad se encontrará al repetir el muestreo.

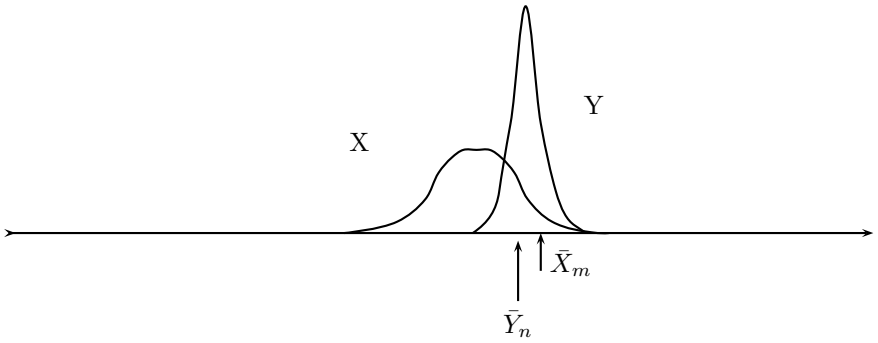


Figura 9.17. Cuando la diferencia de medias muestrales es pequeña comparada con la dispersión de las campanas, la diferencia de medias NO es significativa: podría encontrarse una relación de desigualdad que varía de muestreo a muestreo. Acá se ilustra un muestreo del cual podría inferirse resultados contrarios a los que son en realidad: uno podría pensar que Y es en promedio menor que X, pero la realidad es al revés. Para evitar esos errores, hacemos un test adecuado.

Por eso, para comparar medias se requiere saber de antemano la relación entre las varianzas poblacionales, si son iguales o si son diferentes. Ahora aprenderemos a comparar medias conociendo las varianzas muestrales. Cuando se conocen las varianzas poblaciones es un caso más sencillo que queda para el siguiente capítulo.

Comenzamos, pues, comparando las varianzas poblacionales. Reconocemos dos casos. Caso 1: las varianzas poblacionales son iguales. Caso 2: las varianzas poblacionales son diferentes. Estos casos se estudian por separado por la siguiente razón: si las varianzas (poblacionales) son iguales, puede considerarse que todos los datos, los de la muestra uno como los de la dos, provienen de una misma fuente y se pueden poner juntos, lo cual les dá más poder de resolución. Pero si las varianzas (poblacionales) son diferentes, los datos producen sus estimaciones por separado y no tienen tanto poder de resolución.

Para comparar varianzas seguimos el protocolo del ejemplo anterior que compara lo que se ve con lo que se cree usando un medidor de discrepancia relacional. Las notaciones para las variables son: X designa el número de germanismos por cada 10000 palabras en la escuela Nuevo Hegel, Y representa el número correspondiente a la escuela Entendiendo a Wittgenstein.

Paso 1: Formulamos lo que se cree para valores poblacionales. $H_o : R = \frac{\sigma_x^2}{\sigma_y^2} = 1$

$H_a : R = \frac{\sigma_x^2}{\sigma_y^2} \neq 1$ (en este tipo de ejercicios lo que se requiere es decidir si las varianzas son iguales o distintas, por lo tanto es con dos colas. Si las varianzas poblacionales son iguales se aplica el método

de este ejemplo pero si son diferentes, el del siguiente.)

Paso 2: Datos de lo que se ve, tomados de la muestra:
$Var_x = 36,6$ que viene de $m = n_x = 4$ datos y con 4-1=3 gli
$Var_y = 11,2$ que viene de $n = n_y = 6$ datos y con 6-1=5 gli
$R_{obs} = \frac{Varianza\ mayor}{Varianza\ menor} = \frac{Var_x}{Var_y} = \frac{36,6}{11,2} = 3,27$ con 3 gli en el numerador y 5 gli en el denominador.
Paso 3: Medimos la discrepancia entre lo que se ve y lo que se cree:
$F_{obs} = \frac{R_{obs}}{R} = \frac{3,27}{1} = 3,27$
Paso 4: Decidimos si la discrepancia entre lo que se ve y lo que se cree es grande o pequeña.

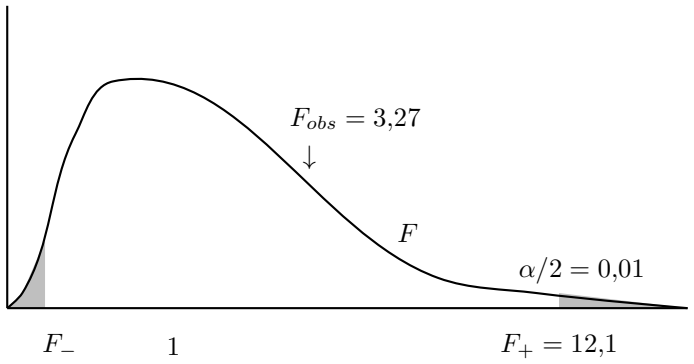


Figura 9.18. Para $\alpha = 0,02$ y dos colas, 3 y 5 g.l, $F_+ = 12,1$. Una $F_{obs} = 3,27$ es normal, por lo que aceptamos la H_o, la diferencia entre las varianzas muestrales puede deberse simplemente al azar.

Vemos que la $F_{obs} = 3,27$ es menos extrema que la F crítica, F_+=12.1 (**la primera de las 3 tablas de la F, la que está en la pág. 160**). Por lo tanto, la discrepancia relacional es pequeña.

Paso 5: Tomamos un veredicto: La F crítica con $\alpha = 0,02$, y $gl_x = 4 - 1 = 3$, $gl_y = 6 - 1 = 5$ es $12,1$. Como la F_{obs} es mucho menor que la F crítica, decidimos que las varianzas son iguales.

Podemos resumir todo en la siguiente tabla:

Comparación de varianzas: varianza grande arriba					
	Descripción	s	$Var = s^2$	n	gl
X	Nuevo Hegel		36.6	4	3
Y	Wittgenstein		11.2	6	5
R_{obs}	3.27				
H_o	$R = \frac{\sigma_x^2}{\sigma_y^2} = 1$				
H_a	$R \neq 1$				
R_{obs}	(var grande)/(var chica) = 3.27				
F_{obs}	$R_{obs}/R = 3.27$				
α	0.02				
$F_c = F_+$	12.1				
Veredicto	H_o OK				

Ahora podemos comparar las medias sabiendo que las varianzas son iguales:

Paso 1: Formulamos la hipótesis nula, lo que se cree que suceda excepto por cosas del azar:

$H_o : \mu_x = \mu_y$, es decir, $\mu_x - \mu_y = 0$
$H_a : \mu_x \neq \mu_y$, o sea $\mu_x - \mu_y \neq 0$.

La hipótesis nula dice que cualquier diferencia observada entre las medias de muestras se debe al azar y no a algún efecto sistemático. La hipótesis alterna dice que hay algo en el ambiente que crea efectos sistemáticos, repetibles, y que hace que las medias de las muestras sean diferentes.

Paso 2: Lo que se ve: las medias son $\bar{X}_4 = 131$ y $\bar{Y}_6 = 135$. La diferencia observada de medias es -4. Las varianzas muestrales son $Var_x = 36{,}6$ y $Var_y = 11{,}2$.

Paso 3: Medimos la discrepancia entre lo que se ve y lo que se cree. Para ello usamos una t pues se trata de comparar medias conociendo las varianzas de las muestras, pero primero tenemos que calcular la **desviación conjunta** dada por:

$$s_c = \sqrt{\frac{Var_x(n_x-1)+Var_y(n_y-1)}{n_x+n_y-2}} = \sqrt{\frac{36{,}66(3)+11{,}2(5)}{4+6-2}} = 4,55$$

donde $m = n_x$, $n = n_y$ representan el número de datos en la v.a. X y Y respectivamente.

La desviación conjunta debe dar un valor interpolado entre las dos desviaciones, o si no, está mal.

$$t_{obs} = \frac{(\bar{X_m}-\bar{Y_n})-(\mu_X-\mu_Y)}{s_c\sqrt{\frac{1}{m}+\frac{1}{n}}} = \frac{(\bar{X_m}-\bar{Y_n})-(0)}{s_c\sqrt{\frac{1}{m}+\frac{1}{n}}}$$

Observemos que cuando la H_o es la igualdad de medias, este quebrado compara la distancia entre las medias muestrales en relación con una medida conjunta de la dispersión de las campanas, tal como se espera del análisis gráfico anterior. Reemplazando valores tenemos:

$$t_{obs} = \frac{131-135}{4{,}55\sqrt{\frac{1}{4}+\frac{1}{6}}} = -1,36$$

Paso 4: Decidimos si la discrepancia entre lo que se ve y lo que se cree es grande o chica. Primero calculamos los grados de libertad de la t (pág 158): $m + n - 2 = 4 + 6 - 2 = 8$

La t-crítica vale 2.8 para $\alpha = 0{,}02$, 8 gli y dos colas.

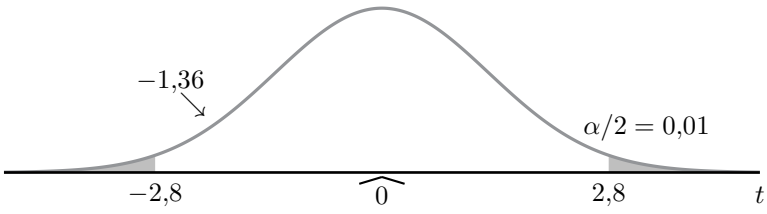


Figura 9.19. La discrepancia debida al azar se distribuye como t con 8 g.l, lo normal es lo que va de -2.8 a 2.8 . La discrepancia -1.36 entre lo predicho por la H_o y lo observado es normal.

Vemos que la t_{obs} es en valor absoluto mucho menor que el correspondiente valor crítico de la t con 8 grados de libertad.

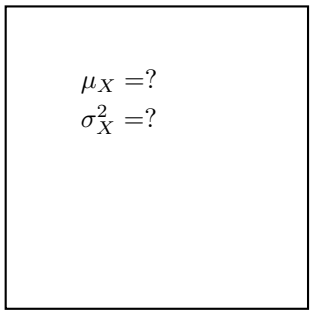
Paso 5: Veredicto: Se acepta la H_o pues la discrepancia es pequeña: las medias son iguales. Las dos escuelas parecen sentir la influencia germana de manera igual tanto en promedio como en varianza.

Podemos resumir todo en la siguiente tabla:

Comparación de medias					
	Descripción	*Media*	s	$Var = s^2$	n
X	Nuevo Hegel	131		36.6	4
Y	Wittgenstein	135		11.2	6
H_o	$\mu_X - \mu_Y = 0$				
H_a	$\mu_X - \mu_Y \neq 0$				
t_{obs}	-1.36				
α	0.02				
gl	8				
t_c	$\pm$2,8				
Veredicto	H_o OK				

199 ***Ejemplo: el Buda. Comparación de medias cuando las varianzas poblacionales son diferentes.*** *B.J., un filósofo de las religiones, está interesado en saber por qué las religiones orientales tienen cabida en cualquier cultura aún en el mundo actual tan moderno y tan materialista. Se ha encontrado con cosas que lo dejan pensativo, como la siguiente: un comerciante tiene un negocio que va a la bancarrota por falta de clientes. Pone un Buda al lado de la registradora y de cuando en cuando le soba la barriga y mentalmente le pide con fe y con reverencia que prospere su negocio. No pasan muchos días cuando el negocio se llena de clientela. El comerciante se asombra y se llena de pánico pues por primera vez en su vida entiende lo que significa amar a un dios ajeno, le da miedo de la ira de Dios y quita al Buda de su negocio. Y sin embargo, el negocio no se cae.*

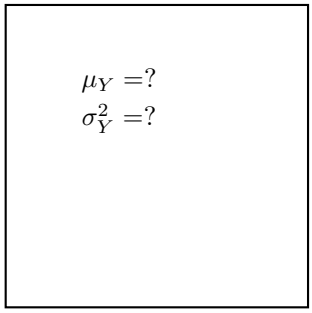
Población 1: todos los negocios con ceremonia budista



Muestra 1 = 37 negocios con ceremonia budista

$\longrightarrow$ $\bar{X}_{37} = 1{,}3$, $s^2_X = 0{,}75$

Población 2: todos los negocios sin ceremonia budista



Muestra 2= 41 negocios sin ceremonia budista

$\longrightarrow$ $\bar{Y}_{41} = 1{,}14$, $s^2_Y = 2$

Figura 9.20. La ganancia promedio con ceremonia budista, $\bar{X}_{37} = 1{,}3$, es mayor que sin ella, $\bar{Y}_{41} = 1{,}14$. ¿Es la diferencia lo suficientemente confiable como para permitirnos decir que se repetirá en el futuro y en otros muestreos? ¿ O se deberá la diferencia observada al efecto azaroso del muestreo, predicho por las varianzas muestrales no nulas, $s^2_X = 0{,}75$ y $s^2_Y = 2$?

Como estos fenómenos no son tan raros y a los comerciantes les intriga hablar del asunto con alguien entendido, B.J. ha logrado muchas entrevistas. Presentamos unos resultados preliminares sobre los promedios de ganancias netas de negocios que han practicado la ceremonia budista contrastados con las ganancias netas de negocios sin prácticas budistas. Los datos relevantes son:

De una muestra de 37 negocios del mismo corte que han practicado la ceremonia budista se tiene una ganancia promedio mensual de 1.3 con una varianza de 0.75. De una muestra de 41 negocios que no han practicado la ceremonia se tiene una ganancia promedio mensual de 1.14 y una varianza de 2. Con $\alpha = 0{,}02$ *decidamos si el incremento en las ventas es por azar o por algo sistemático.*

Solución: Se trata de una comparación de medias conociendo las varianzas de la muestra, lo cual exige que primero se comparen las varianzas poblacionales.

Comparación de varianzas poblacionales:
Como la varianza de la muestra de los negocios sin ceremonia budista es la más grande, estudiamos el quebrado de la varianza sin ceremonia budista entre la varianza con ceremonia budista:

Paso 1: Formulamos lo que se cree. El subíndice X significa 'con ceremonia budista' y el Y 'sin ceremonia budista'. La H_o se rechaza con dos colas pues decidimos si las varianzas poblacionales son o no iguales:

$H_o : \sigma_Y^2 = \sigma_X^2$, es decir $R = \sigma_Y^2/\sigma_X^2 = 1$
$H_A : \sigma_Y^2 \neq \sigma_X^2$, es decir $R = \sigma_Y^2/\sigma_X^2 \neq 1$

Paso 2: Lo que se ve: Anotamos la varianza grande y la varianza pequeña:

Sin ceremonia budista: $Var_Y = VAR_G = 2$, número de datos: $n_Y = 41$,
Con ceremonia budista: $Var_X = VAR_P = 0,75$, número de datos: $n_X = 37$,

$R_{obs} = Var_G/Var_P = 2/0{,}75 = 2{,}667$

Paso 3: Medimos la discrepancia entre lo que se ve y lo que se cree según la H_o:

$F_{obs} = \frac{R_{obs}}{R} = \frac{2{,}667}{1} = 2,667$

Paso 4: Decidimos si la discrepancia entre lo que se ve y lo que se cree es grande o pequeña: los gli son iguales a número de datos menos uno, tanto en el numerador como en el denominador.

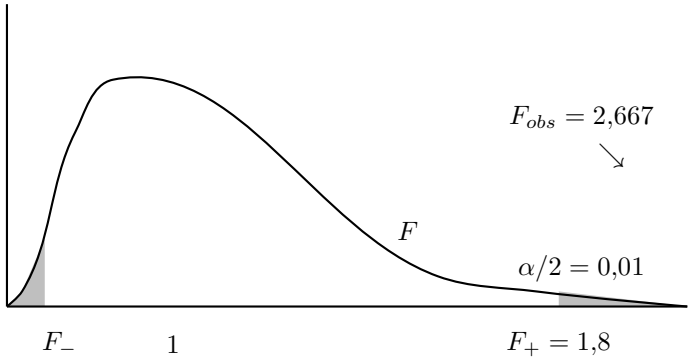


Figura 9.21. Una discrepancia de 2.66 es grande pues es mayor que la barrera que separa lo grande de lo pequeño, la cual es 1.8

La varianza del numerador es la más grande, por lo que los g.l del numerador son $41 - 1 = 40$, en tanto que los gli del denominador son $37 - 1 = 36$. Para $\alpha = 0{,}02$ y dos colas, 40 y 36 g.l, $F_+ = 1{,}8$ (

la primera de las 3 tablas de la F, la que está en la pág. 160.) Por lo tanto, una $F_{obs} = 2{,}66$ es más extrema, es grande.

Paso 5: Veredicto: La F_{obs} es mayor que la F crítica; por tanto decidimos que las varianzas poblacionales son diferentes, es decir, tenemos razones para creer que la diferencia entre las varianzas muestrales se dio no por azar sino por algo sistemático, que bien podría ser el Buda. Pero es difícil ser concluyente pues se requiere demostrar que el único cambio ocurrido y que pudo afectar el nogocio fue el Buda. Nuestro resumen en forma de tabla:

Comparación de varianzas: varianza grande arriba					
	Descripción	s	$Var = s^2$	n	gl
Y	Sin ceremonia budista		2	41	40
X	Con ceremonia budista		0.75	37	36
R_{obs}	2.667				
H_o	$R = 1$				
H_a	$R \neq 1$				
F_{obs}	2.667				
α	0.02, dos colas				
$F_c = F_+$	1.8				
Veredicto	H_o ×				

Ahora, sabiendo que las varianzas son distintas, investiguemos qué pasa con las medias:

Paso 1: Lo que se cree.
$H_o : \mu_X = \mu_Y$, o lo que es lo mismo: $\mu_X - \mu_Y = 0$ (el Buda ni fu ni fa)
$H_A : \mu_X > \mu_Y$ o lo que es lo mismo: $\mu_X - \mu_Y > 0$ (el Buda si ayuda)
Lo esperado bajo la H_o es $\mu_X - \mu_Y = 0$.

Paso 2: Lo que se ve.
Con ceremonia budista: $\bar{X}_{37} = 1,3$, $Var_X = 0,75$, número de datos: $n_X = 37$
Sin ceremonia budista: $\bar{Y}_{41} = 1{,}14$, $Var_Y = 2$, número de datos: $n_Y = 41$,

Lo observado es $\bar{X}_{37} - \bar{Y}_{41} = 1,3 - 1,14$.

Paso 3: Medimos la discrepancia entre lo que se ve y lo que se cree según la H_o. Usamos una t pero computada para el caso en el cual las varianzas se encontraron diferentes:

$$t_{obs} = \frac{(\bar{X}_{37} - \bar{Y}_{41}) - (\mu_X - \mu_Y)}{\sqrt{\frac{Var_X}{n_X} + \frac{Var_Y}{n_Y}}} = \frac{(1{,}3 - 1{,}14) - 0}{\sqrt{\frac{0{,}75}{37} + \frac{2}{41}}} = 0{,}609$$

Paso 4: Decidimos si la discrepancia hallada es grande o chica.

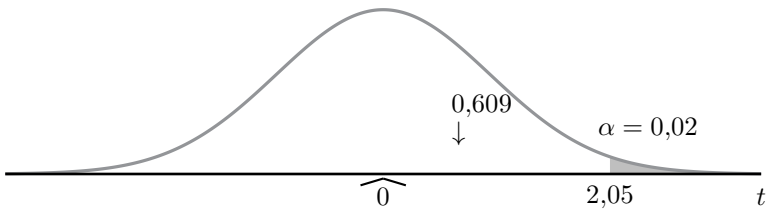


Figura 9.22. La discrepancia debida al azar se distribuye como t con 68 gli. La discrepancia 0.609 entre lo predicho por la H_o y lo observado es normal. Aceptamos la H_o: hasta la presente no hay diferencia significativa entre los altibajos en la ventas debidas al azar en comparación con lo que una ceremonia budista parece ofrecer.

Hallemos los gli

$$gli = \frac{(\frac{Var_X}{n_X}+\frac{Var_Y}{n_Y})^2}{\frac{(\frac{Var_X}{n_X})^2}{n_X+1}+\frac{(\frac{Var_Y}{n_Y})^2}{n_Y+1}} - 2 = \frac{(\frac{0,75}{37}+\frac{2}{41})^2}{\frac{(\frac{0,75}{37})^2}{38}+\frac{(\frac{2}{41})^2}{42}} - 2 = \frac{0,00477}{0,0000108+0,00005666} - 2 = 70,7 - 2 = 68,2$$

La t crítica con $\alpha = 0{,}02$, una cola y con gli dados por 68.2 es 2,05 (pág 158). Atención: para comparar medias hay dos métodos y para escoger el correcto se comparan las varianzas poblacionales para saber si son iguales o diferentes. Por tanto, trabajamos con dos colas. Pero al comparar las medias, el enunciado del problema es el que determina que la H_o se rechace con una o con dos colas. En este caso es con cola superior pues nos interesa saber si el aumento observado en las ventas puede ligarse a la ceremonia budista o si es mejor explicarlo por azar. ¿Qué debemos hacer con la significancia? ¿Dividirla por dos? ¿Multiplicarla por dos? ¿Dejarla igual? La respuesta correcta sale de la definición de significancia: es un número que indica la proporción de lo que se considera extremo. Por tanto, cuando la comparación de medias sea con dos colas hay que dividirla entre dos pero para una cola no: tomamos la significancia y la ponemos toda en la cola superior. Todo esto ya está involucrado en NUESTRAS tablas de la t y de la F y allí lo único que hay que mirar es el caso: una o dos colas y con la significancia intacta.

Paso 5: Veredicto. H_o se acepta. Discrepancia pequeña. Aún no hay datos suficientes para asegurar que una ceremonia budista ayude más que el azar para mejorar las ganancias.

Resumen en forma de tabla:

Comparación de medias					
	Descripción	*Media*	s	$Var = s^2$	n
Y	Sin ceremonia budista	1.14		2	41
X	Con ceremonia budista	1.3		0.75	37
H_o	$\mu_X - \mu_Y = 0$				
H_a	$\mu_X - \mu_Y > 0$				
t_{obs}	0.609				
α	0.02, cola sup.				
gl	68.2				
t_c	2,05				
Veredicto	H_o OK				

Nota: Los gli de la t cuando las varianzas poblacionales son diferentes tienen una fórmula complicada. Pero en el presente caso, es suficiente darse cuenta que el valor más pequeño de toda la tabla de la t es 1.6. Lo cual es más grande que 0.609. Por eso, no hay ninguna duda de que la H_o se acepta con cualquier significancia menor que 0.05.

200 ***Para recordar siempre: no se puede estimar una diferencia de medias, para decidir si son iguales o distintas, sin antes haber estimado las varianzas o las desviaciones.***

La razón es que la desviación puede interpretarse como una incertidumbre.

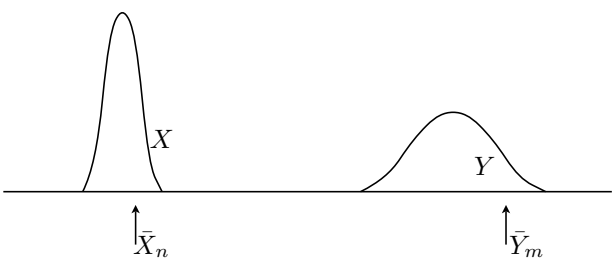


Figura 9.23. Cuando las campanas son muy cerradas en comparación con la distancia entre las medias, las medias muestrales estarán bastante cerca de las medias poblacionales, y no hay dificultad en decidir si las medias poblacionales son iguales o si son diferentes y cuál es la mayor.

Si la incertidumbre descrita por las desviaciones es pequeña comparada con el orden de magnitud de la diferencia de medias, unos pocos datos son suficientes para decidir la relación entre las medias.

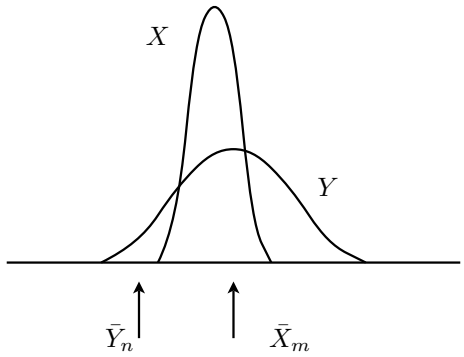


Figura 9.24. Cuando las campanas son muy abiertas en comparación con la distancia entre las medias, las medias muestrales pueden dar una noción equivocada de la relación entre las medias poblacionales. En esta gráfica, las medias muestrales salieron al revés de las medias poblacionales.

Si los datos se entrecruzan es porque la incertidumbre descrita por las desviaciones es grande comparada con la diferencia de medias. En ese caso se requieren muchos datos para poder rechazar la hipótesis nula de la igualdad de medias.

201 ***Ejemplo: el trasnocho. Comparación de medias para tablas de frecuencias absolutas.*** *Se les preguntó a unos estudiantes de Arquitectura, tomados al azar, cuántas horas trasnochaban por semana en época de proyectos. Lo mismo se hizo con otros estudiantes de Sociales. Los resultados se resumen en la tabla adjunta que da la distribución para cada facultad de las horas de trasnocho por semana. Por ejemplo, los estudiantes que trasnochan 17 horas por semana son 3 de arquitectura y 15 de sociales. ¿Hay diferencia promedio de trasnocho entre los dos grupos?*

Distribución detrasnocho		
Horas/semana	Arquitectura	Sociales
13	2	0
14	3	0
15	9	14
16	6	18
17	3	15
18	2	0
19	1	0
Total		

Solución resumida:

Paso 1: Calculamos medias y varianzas por los métodos del parcial I. Para sociales tenemos:

```
      x         F       xF      x2F = x cuadrado por F
      13        0        0        0
      14        0        0        0
      15       14      210     3150
      16       18      288     4608
      17       15      255     4335
      18        0        0        0
      19        0        0        0
Suma           47      753    12093
```

```
 n   = suma F = 47
 suma xF  =    753
 suma x2F  =  12093
 Media = suma xF / n =     16.021
 Sxx = suma x2F - (sumaxF)^2 /n =     28.979
 Varianza =  Sxx/(n-1) =      0.630
```

Para Arquitectura:

```
     x        F      xF     x2F
     13       2      26     338
     14       3      42     588
     15       9     135    2025
     16       6      96    1536
     17       3      51     867
     18       2      36     648
     19       1      19     361
Suma         26     405    6363

 n   = suma F = 26
 suma xF  =    405
 suma x2F  =   6363
 Media = suma xF / n =     15.577
 Sxx = suma x2F - (sumxF)^2 /n =     54.346
 Varianza =  Sxx/(n-1) =      2.174
```

Paso 2: Comparación de varianzas con alfa = 0.05:

```
varx = 0.63
vary = 2.174

Ho: R=1.0, dos colas.

RObs = vary/varx
RObs = 3.45

FObs = RObs/R
FObs = 3.45
glis F Num = 25
glis F Den = 46
alfa medios = 0.025
FCriticaDosColas = 1.9
FObsp > FCritico
discrepancia grande, varianzas diferentes.
```

Paso 3: comparación de medias sabiendo que las varianzas poblacionales son diferentes:

```
mediax=16.02
varx =0.63
nx = 47.0
mediay= 15.58
vary = 2.17
ny= 26
Ho: mux - muy = 0: dos colas.
```

```
Diferencia observada: mediax-mediay =  0.44

Discrepancia entre lo que se ve y lo que se cree en la Ho:
tobs =   1.43

Grados de libertad, fórmula larga: 33
alfa = 0.05
t-Crítico para dos colas = tc2 =  +/-2.0237
tobs pequeña: medias iguales.
Los datos permiten creer que ambos grupos trasnochan en promedio por igual.
```

9.3. Conclusión

Hemos pasado revista a varios y diversos ejemplos de problemas de la ciencia involucrada a la vida real y hemos verificado que el método científico es básicamente un contraste ente lo que se ve y lo que se cree. Nuestras creencias pueden ser referentes a una media de una distribución central, o de su varianza, o de la diferencia de dos medias o de la relación entre dos varianzas. También pueden ser de proporciones o de diferencia de proporciones. La discrepancia entre los valores registrados y los valores esperados se mide por un estadígrafo de contraste. Si de acuerdo con unos valores críticos, dados por tablas, la discrepancia es pequeña, dicha discrepancia se atribuye al azar. Pero si es grande, se considera prudente considerar que pueden haber efectos sistemáticos no tenidos en cuenta en la H_o que estén afectando los resultados.

9.4. la tabla de la z

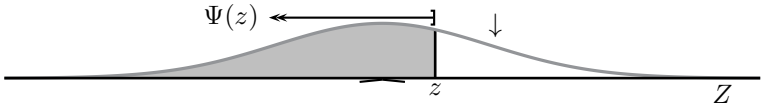


Figura 9.25. Sobre la campana normal estándar, media cero y desviación uno, el área antes de z se denota $\Psi(z)$ (tridente o psi de z) y es el valor reportado por nuestra tabla:

LA Ψ PARA $Z \sim N(0,1)$			
$\Psi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds$			
1 cola: $\Psi = \alpha$, 2 colas $\Psi = \alpha/2$			
z-	$\Psi(z)$	z+	$\Psi(z)$
0	0.5000	0	0.5000
-0.10	0.4602	0.10	0.5398
-0.20	0.4207	0.20	0.5793
-0.30	0.3821	0.30	0.6179
-0.40	0.3446	0.40	0.6554
-0.50	0.3085	0.50	0.6915
-0.60	0.2743	0.60	0.7257
-0.70	0.2420	0.70	0.7580
-0.80	0.2119	0.80	0.7881
-0.90	0.1841	0.90	0.8159
-1.00	0.1587	1.00	0.8413
-1.10	0.1357	1.10	0.8643
-1.20	0.1151	1.20	0.8849
-1.30	0.0968	1.30	0.9032
-1.40	0.0808	1.40	0.9192
-1.50	0.0668	1.50	0.9332
-1.60	0.0548	1.60	0.9452
-1.64	0.05	1.64	0.95
-1.70	0.0446	1.70	0.9554
-1.80	0.0359	1.80	0.9641
-1.90	0.0287	1.90	0.9713
-1.96	0.025	1.96	0.975
-2.00	0.0228	2.00	0.9772
-2.10	0.0179	2.10	0.9821
-2.20	0.0139	2.20	0.9861
-2.30	0.0107	2.30	0.9893
-2.40	0.0082	2.40	0.9918
-2.50	0.0062	2.50	0.9938
-2.60	0.0047	2.60	0.9953
-2.70	0.0035	2.70	0.9965
-2.80	0.0026	2.80	0.9974
-2.90	0.0019	2.90	0.9981
-3.00	0.0013	3.00	0.9987
-3.10	0.0010	3.10	0.9990
-3.80	0.0001	3.80	0.9999
$-\infty$	0	∞	1

Capítulo 10

Taller y parcial III

202 ***Objetivo:*** *muchos ejercicios y muchos parciales para quedar muy bien preparados.*

10.1. Ejercicios básicos

Hay que inventar una telenovela para cada problema.

1 . Estudiar la H_o dada a la luz del dato experimental dado x sabiendo que X se distribuye normalmente (Ej 179 , pág. 150 + Ej 190, pág. 165. Respuestas en la pág 311):

a) $H_o : \mu = 3$, dos colas, $\alpha = 0{,}05$, $\sigma = 1{,}1$, $x = 4$.

b) $H_o : \mu = 5$, cola superior, $\alpha = 0{,}02$, $\sigma = 0{,}2$, $x = 14$.

c) $H_o : \mu = -1$, cola inferior, $\alpha = 0{,}01$, $\sigma = 8$, $x = -9$.

2 . Estudiar la H_o dada a la luz del dato experimental dado $\bar{X}_n$ sabiendo que X se distribuye normalmente (Ej 192, pág. 168):

a) $H_o : \mu = 3$, dos colas, $\alpha = 0{,}05$, $\sigma = 1{,}1$, $\bar{X}_8 = 3{,}8$.

b) $H_o : \mu = 5$, cola superior, $\alpha = 0{,}02$, $\sigma = 0{,}2$, $\bar{X}_{14} = 6{,}2$.

c) $H_o : \mu = -1$, cola inferior, $\alpha = 0{,}01$, $\sigma = 8$, $\bar{X}_{23} = -9{,}8$.

3 . Estudiar la H_o dada a la luz del dato experimental dado $\bar{X}_n$ sabiendo que X se distribuye normalmente (Ej 193, pág. 169):

a) $H_o : \mu = 3$, dos colas, $\alpha = 0{,}05$. Calcule $\bar{X}_8$ y la desviación correspondiente a partir de los datos 3.9, 3, 3.7, 4.8, 2, 5, 3.5, 4.5. Ayuda: Ejemplo: 17, pág. 9. La media da $\bar{X}_8 = 3{,}8$, la desviación $s = 0{,}994$.

b) $H_o : \mu = 5$, cola superior, $\alpha = 0{,}01$, $s = 0{,}2$, $\bar{X}_{14} = 6{,}2$.

c) $H_o : \mu = -1$, cola inferior, $\alpha = 0{,}05$, $s = 8$, $\bar{X}_{23} = -9{,}8$.

d) $H_o : \mu = -1$, cola superior, $\alpha = 0{,}01$, $s = 1{,}3$, $\bar{X}_{67} = -11{,}2$.

4 . Decidir la H_o sobre la probabilidad de cara p con base en la proporción de la muestra f. Una proporción es una relación entre dos números. Si la proporción viene de una muestra es una frecuencia relativa. Pero si viene de una población es una probabilidad. En ambos casos la proporción es una número entre cero y uno. (Ej 194, pág. 170):

a) $f = 0{,}2$ con 15 datos, $p = 0{,}4$, dos colas, $\alpha = 0{,}07$.

b) Para f: 12 de 34 (es decir $f = 12/34$), $p = 0{,}5$, cola inferior, $\alpha = 0{,}03$.

c) Para f: 12 de 14, $p = 0{,}4$, cola superior, $\alpha = 0{,}09$.

5 . Decidir la H_o sobre la igualdad de proporciones (la H_o siempre es poblacional, se trata de probabilidades) con base en los datos (de frecuencias relativas) de muestras al azar e independientes (Ej 195, pág. 172):

a) $f = 0{,}2$ con 15 datos, para g: 5 de 17 datos, dos colas, $\alpha = 0{,}05$. (Ayuda: Si $f = X/n = X/15 = 0{,}2$ implica $X = 3$.)

b) Para f: 12 de 34; para g: 15 de 27. Cola inferior, $\alpha = 0{,}01$.

c) Para f: 12 de 14; para g: 15 de 27. Cola superior, $\alpha = 0{,}09$. Atención: no se puede usar nuestra tabla de la t pues no tiene tal significancia. Pero como los gli son mayores que 30, se puede usar la tabla de la z con cualquier significancia.

6 . Decidir la H_o sobre la igualdad de medias para datos apareados con base en la siguiente muestra al azar (Ej 196, pág. 174):

a) (1,2), (3,2), (4,2), (2,5), (1,3), (1,4), dos colas.

b) (1,4), (4,3), (6,2), (6,4), (3,1), (5,1), (5,0), (4,1), (5,1), dos colas.

c) (-1, -3), (-3,2), (-2,4), (3,-2), (-1,-2) (2,-1), cola superior.

7 . Estudiar la H_o sobre varianzas a la luz de $Var1$ y $Var2$, las varianzas de muestras independientes y al azar, sabiendo que X_1, X_2 se distribuyen normalmente (Ej 197, pág. 176):

a) $H_o : \sigma_1^2 = \sigma_2^2$, dos colas, $\alpha = 0{,}02$, $Var1 = 20$ con 9 datos, $Var2 = 12$ con 14 datos.

b) $H_o : \sigma_1^2 = \sigma_2^2$, cola superior, $\alpha = 0{,}01$, $Var1 = 12$ con 9 datos, $Var2 = 4$ con 14 datos.

c) Cola inferior: lo veremos en el próximo parcial.

8 . Demuestre que las varianzas poblacionales son iguales (H_a con dos colas) y estudie la H_o sobre medias a la luz de los resultados experimentales sabiendo que X, Y se distribuyen normalmente y que son independientes (Ej 198, pág. 179):

a) $H_o : \mu_X = \mu_Y$ con dos colas, $\alpha = 0{,}02$, $\bar{X}_{21} = 35$, $S_X = 8$, $\bar{Y}_{17} = 65$, $S_Y = 7$.

b) $H_o : \mu_X = \mu_Y$ con cola superior, $\alpha = 0{,}05$, $\bar{X}_6 = 123$, $S_X = 12$, $\bar{Y}_{15} = 85$, $S_Y = 13$.

c) $H_o : \mu_X = \mu_Y$ con cola inferior, $\alpha = 0{,}01$, $\bar{X}_{34} = 23$, $S_X = 4$, $\bar{Y}_{28} = 42$, $S_Y = 6$.

9 . Demuestre que las varianzas son diferentes (H_a con dos colas) y estudie la H_o sobre medias a la luz de los resultados experimentales sabiendo que X, Y se distribuyen normalmente y que son independientes (Ej 199, pág. 183):

a) $H_o : \mu_X = \mu_Y$, dos colas, $\alpha = 0{,}02$, $\bar{X}_{21} = 35$, $S_X = 8$, $\bar{Y}_{17} = 65$, $S_Y = 32$.

b) $H_o : \mu_X = \mu_Y$, cola superior, $\alpha = 0{,}01$, los datos para X son: 124, 126, 109.5, 136.5, 108, 134 (se calcula la media y la varianza por alguno de los métodos del parcial uno). Datos de Y: $\bar{Y}_{15} = 85$, $S_Y = 1{,}2$.

c) $H_o : \mu_X = \mu_Y$, cola inferior, $\alpha = 0{,}05$, $\bar{X}_{34} = 27$, $S_X = 4$, $\bar{Y}_{28} = 53$, $S_Y = 21$.

10 . Para cada uno de los ejemplos resueltos del capítulo 9 y en cuanto las tablas lo permitan, encuentre la significancia mínima a la cual se puede rechazar la H_o.

11 . Para cada test visto demuestre o refute que cuando n crece la probabilidad de rechazar la H_o aumenta. Aplicación: si uno tiene las sospecha de que cierta hipótesis nula es falsa pero no ha podido demostrarlo con sus datos, el diagnóstico es que éstos son muy pocos y lo que uno debe hacer es reunir más.

Los variados problemas que figuran a continuación son para que cada quien pueda practicar solución de problemas con folleto de fórmulas y nada más. Para cada problema, elija el caso con el cual ha de resolverse. Ejemplo: el problema de los excomulgados se resuelve con el caso 5 subcaso 2.

203 ***Problemas del corazón***

12 . Un paciente es aceptado de urgencia con diagnóstico de arritmia, es decir, que su corazón late sin ritmo, a veces muy despacio y a veces muy rápido cambiando de una lado al otro sin una razón a la vista. El paciente es sometido a un tratamiento. Diga cuál es el estadígrafo adecuado (media o varianza o qué) para decidir si el tratamiento funcionó o no.

13 . Otro paciente es aceptado de urgencia con diagnóstico de taquicardia, es decir, que su corazón late taqui-taqui-taqui, es decir, a toda prisa. El paciente es sometido a un tratamiento. Diga cuál es el estadígrafo adecuado (media o varianza o qué) para decidir si el tratamiento funcionó o no.

14 . Una pareja de esposos son internados con diagnóstico de taquicardia. Sus datos son registrados día por día. La esposa alega que ella se está mejorando más rápido que el esposo. Con las herramientas que tenemos, ¿cuál caso se utiliza para estudiar la creencia de la esposa? ¿ Qué pasa si el diagnóstico es por arritmia?

15 . Uno puede enfermarse de taquicardia sin enfermarse de arritmia y en cambio siempre que se enferma de arritmia también se enferma de taquicardia. ¿Es eso posible o es necesario?

16 . Cuando se presenta un paciente con arritmia y con taquicardia, ¿Qué debe curársele primero, la arritmia o la taquicardia?

17 . Entre la arritmia y la taquicardia, ¿cuál es 100 veces más peligrosa que la otra?

18 . ¿Le sería posible enumerar algunas prevenciones contra estos males?

19 . Un paciente es internado por arritmia, la varianza de 16 eventos de su frecuencia de latido antes del tratamiento es 1000. Al día tercero es 500 sobre una muestra al azar de 8 eventos. ¿Le darán de alta al paciente?

20 . Un paciente es internado por taquicardia, y se registran 16 eventos de su frecuencia de latido: la frecuencia media es 220 y la varianza es 1000. Se somete a tratamiento y al día tercero la frecuencia media es de 120 y su varianza es 500 sobre una muestra de 8 eventos. ¿Le darán de alta al paciente?

204 ***Problemas sobre la Edad Media (400-1500)***

21 . Alrededor del 1200 la ciudad europea más grande era París y después seguían Milán, Florencia, Génova y Nápoles, las cuales tenían alrededor de los 100000 habitantes. En realidad, una ciudad no tiene un número fijo de personas. Florencia por ejemplo, tenía una media histórica de 100000 y una desviación de 10000. Pero llegó la Peste Negra y la población de la ciudad cayó a 70.000. Demuestre que la Peste Negra fue un evento trascendente en la demografía de la ciudad y es por eso que es tan famosa.

22 . La verdad sea dicha, la Iglesia fue el puente entre la historia de los siglos y la modernidad. La Iglesia tenía poder, conocimiento y riqueza. Sucedía que si en la Edad Media alguien quería estudiar sólo tenía una opción: volverse monje. El monje Alberto, monje representativo de su comunidad y monje desde hacía 17 años, ocupaba su tiempo libre transcribiendo libros, los cuales adornaba página por página con elaborados motivos casi abstractos. El trascribía 12 libros al año, en promedio, con una desviación de 2. Pero cuando le llegó una colección de libros sobre Platón que había sido traducido del árabe por los españoles, su producción en los siguientes 3 años bajó a un promedio de 8 con una desviación de 4. ¿Afectó Platón al monje Alberto? ¿Afectó Platón a los monjes que son como él?

23 . Un monasterio de la Edad Media era realmente una beneficencia, un hospital, un hotel, un refugio, una universidad, una biblioteca religiosa y una biblioteca civil. No era difícil entender que un monasterio debía ser una hermosísima obra arquitectónica. En cada monasterio había un claustro que para todos los efectos era la plaza principal de la comunidad. Antes del 800, una muestra al azar de 35 claustros tenía de lado un promedio de 12m con una desviación de 2m. Pero después, una muestra de 13 plazoletas tenía una media de 14 con una desviación de 3. ¿Hay alguna diferencia entre el tamaño de los claustros entre las dos épocas? ¿Partió el año 800 en dos a la historia de la arquitectura religiosa?

24 . Debe ser que ya me estoy envejeciendo, pensó el gran Rey. Lo pensó porque ya había entrado la primavera y, honestamente, no quería levantarse de la cama para ir a la guerra. En los diez años antes de él, había por año un promedio de 0.4 batallas con una desviación de 0.1. Pero en una muestra de 7 años de su reinado, ha habido un promedio de 6 batallas por año con una desviación de 2. ¿Partió el Rey la historia de su mundo en dos partes en lo relativo al promedio de guerras por año?

25 . Cristo dijo 'Bendecid y no maldigáis'. Pero la Iglesia de la Edad Media tenía una fórmula para pronunciar la excomunión que consistía en una sarta de maldiciones que uno ni se imagina que puedan existir a no ser que uno lea las maldiciones del Deuteronomio. Había, sin embargo, una salvedad que condicionaba la aplicabilidad de las maldiciones: el arrepentimiento. Tales maldiciones eran pronunciadas por el Obispo en una pomposa y tenebrosa ceremonia. Pero eso aún distaba muy lejos de ser la inquisición. Es posible que en su teología, la Iglesia considerara que hay que castigar al hijo a quien se ama, pues es para su propio bien. De una muestra al azar de 14 excomulgados 8 se arrepintieron al ver que todas las puertas se le cerraban tanto para él como para su familia. Pero un hombre rico, poderoso e instruido fue excomulgado. Después de eso, de 16 excomulgados sólo 3 se arrepintieron. ¿Puede decirse en general que hay un cambio en la actitud hacia el arrepentimiento después de la excomunión de un hombre que sea rico, poderoso e instruido?

26 . La clase alta de la Edad Media, aparte del Señor Feudal y de sus nobles, fue conformada por los comerciantes que no dudaban en sobre-enriquecerse haciendo usura de sus productos en tiempo de crisis. Naturalmente que los más poderosos de ellos formaban una liga con su propio ejército pues las huelgas y revueltas no eran algo raro. En cierta región y durante el decenio del 1240 había un promedio de 7 huelgas por año con una desviación de 2. Pero en el siguiente decenio, el promedio llegó a 10 y una desviación de 3. ¿Podrá decirse que aumentaron en promedio las huelgas de forma estadísticamente significativa?

27 . La primera gran revolución Europea ocurrió en los siglos XII y XIII y su causa fue financiera: la tierra era en esos tiempos la única fuente de riqueza. Pero fue en ese entonces que el fenómeno banco-prestamista-dinero se impuso minorando el poder de la tierra y haciendo nacer una nueva clase: la burguesía. El Burgo-Maestre de la ciudad A alega que su localidad es menos usurera que la de B pues el número de soldados se conserva más homogéneo y como prueba ofrece los siguientes datos: 70 soldados en promedio durante los últimos 6 años con una desviación de 13 para su ciudad pero 83 soldados en promedio con una desviación de 16 para la ciudad B. ¿Tendrá razón el Burgo-maestre de A?

28 . Un efecto del poder del comercio fue el aumento de los pobres y mendigos. Durante los últimos 5 años antes de la muerte de Giorgio había un promedio de 18 mendigos en la ciudad con una desviación de 4, pero desde que el hijo quedó de dueño y señor, en éstos últimos 3 años ha habido un promedio de 23 mendigos con una desviación de 5. ¿Hay razón para creer que el nuevo señor es mucho más usurero?

29 . En la Edad Media nacieron los gremios con elementos de poder: cada gremio determinaba y controlaba el monopolio de su producción, elegían sus autoridades, escalafonaban a sus miembros, fijaban los precios y estudiaba el mercado para determinar las cantidades y calidades de cada producto. En una ciudad, la costumbre había sido vender por año 900 prendas para invierno. Pero en los últimos 6 años y debido al intenso frío se han requerido un promedio de 1200 prendas con una desviación de 50. Como se sabe que el frío viene de Rusia y no de Inglaterra ni de Escandinavia, los del gremio andan diciendo que la Madre Rusia es de muy buena suerte. ¿Se justifica dicha reacción?

30 . La demasiada maldad y opresión de la Edad Media incitó a muchos a apartarse del mundo, del siglo, como se le decía. Y se iban de ermitaños para consagrarse a la oración y a la penitencia. No era de extrañar que el número de ermitaños aumentara a la par con el número de soldados. Pero a veces, se vaticinaba que tanta maldad era una muestra inequívoca de la inminente venida del Señor para poner fin a tanto sufrimiento. Por ese tiempo, podía surgir el movimiento ermitaño con mucha fuerza. Si en una muestra al azar regional tomado el decenio pasado hubo 6 ermitaños en promedio por año, con una desviación de 2, y en los 3 años del nuevo ha habido un promedio de 12 con una desviación de 2, podemos decir que estamos en uno de esos momentos de efervescencia?

31 . En el último tiempo había habido en una muestra regional 3 ermitaños por cada 75 monjes. Pero ahora que se fundó la orden de los Cartujos, hay 4 ermitaños por cada 128 monjes. ¿Podemos decir que esa nueva orden ha ofrecido algo nuevo para la gente?

32 . Los humanos podemos reproducirnos como ratones. Pero en la Edad Media, el crecimiento poblacional fue moderado, excepto hacia su final. El crecimiento es moderado cuando la mortalidad infantil es alta. Una manera de estimar la tasa de mortalidad infantil es conociendo la diferencia de las edades de un hijo y del siguiente, pues si no hay mortalidad la diferencia debe ser aproximadamente de un año. Las siguientes parejas tomadas de una muestra al azar denotan dichas edades redondeadas a años: (12,8), (8,7), (9,6), (7,4), (7,2), (5,0), (3,0), (11,5). ¿Podemos decir que hay mortalidad infantil? Atención: la hipótesis nula es cero mortalidad lo cual significa que la diferencia esperada de medias es uno.

33 . El año antepasado hubo mucha escasez de trigo por lo que el año pasado hubo mucha siembra de ese grano. En consecuencia, en este año ha habido abundancia de trigo. Como hay mucho pan, venderlo se ha tornado algo difícil. Las ventas de pan en la semana han sido día por día y en unidades arbitrarias de peso, 23, 21, 23, 24, 25, 19, 23. Pero ahora hemos decidido hacer pan con alguna escultura sobre el lomo, como de un caballo o de una oveja. Las ventas han sido: 23, 25, 24, 23, 21, 23, 22. ¿Aconsejaría Ud a las demás panaderías que siguieran la misma nueva estrategia de fabricación?

34 . Nosotros, los artistas sagrados, fabricamos cuadros de santos, de santas, de apóstoles, de Jesús, de la Virgen María. La idea es que cada cual pueda tener un gran santo en su casa. En tiempo pasado vendíamos por mes aproximadamente 10 cuadros de la Virgen y 30 de los otros. En este mes que ha habido una aparición de la Virgen en Vézelay, hemos vendido 35 cuadros de la Virgen, pintados con las indicaciones de los videntes, y 2 de los otros. ¿Causó la aparición un cambio en las proporciones de ventas?

35 . En el arte Gótico hay una tendencia a acercarse a Dios. Eso se refleja en la altura de las columnas de las Catedrales. Una muestra al azar de 8 catedrales del arte Románico dio una altura promedio de 10 metros con una desviación de 2, pero en una muestra al azar de 10 en el arte Gótico la altura promedio era de 13 con una desviación de 3. ¿Fueron realmente las catedrales Góticas más altas que las Románicas?

36 . Una Catedral es una representación del cielo, majestuoso y eterno. Es difícil captar lo bien que los medioevos lograron entender eso, pero podemos intentarlo dando la duración en años de la construcción de algunas Catedrales: 70, 65, 67, 98, 120, 87, 77. Téngase en cuenta que la vida promedio de un ser humano era de unos 28 años. En contraposición a la Edad Media, la construcción de Catedrales de siglos anteriores demandaban años como los siguientes: 3, 6, 4, 3, 8, 6, 7. ¿Hubo un cambio en la duración de la construcción de una Catedral?

205 ***Problemas sobre Derecho tribal***

37 . Una cruda realidad de la vida es que no hay derecho sin coacción que haga justicia. Un instrumento de justicia en las sociedades primitivas fue el clan, el poder contante y sonante de la familia, lista a matar y a morir si fuese necesario. En materia de muertes, cada grupo genera su propia idiosincrasia que puede ser diferente de las de otros grupos. En una región A el número de muertes violentas, aparte de las causadas por los vengadores de sangre cuyo único objetivo es hacer justicia, es de 23 en promedio por año con una desviación de 3. En otra región de similar población, el promedio durante los últimos 3 años ha sido de 40. ¿Podrá decirse que hay idiosincrasias diferentes?

38 . Como regla infalible, todo clan que se hace poderoso es convertido por los más débiles en instrumento de derecho y de justicia. A él se recurre a poner quejas, a solicitar justicia, a pedir protección. Y también se viene a pagar los derechos correspondientes por tal subyugación. Hay dos clanes rivales que aún no son muy fuertes como para generar un monopolio de poder. Por lo tanto, sus allegados vienen y van y cambian de bando sin mucho inconveniente. En una muestra al azar de 100 personas se tuvo en el año pasado que el 34 % de la población estaba con el clan A, pero en este nuevo año, ahora que ha habido cambio de jefe, su proporción ha crecido al 74 % habiendo tomado una muestra al azar de 80 personas. ¿Se debe dicho cambio, como dice la gente, a que el nuevo jefe luce más cercano a sus aliados?

39 . No ha de ser raro que un clan protector se convierta en un clan opresor ante el cual sólo cabe invocar la protección de los dioses. Al mismo tiempo, en aras de los beneficios del comercio no es propicio ser injusto a toda hora. Por eso, todas las religiones involucran una porción de justicia social y familiar que en general es acatada por la gente buena. Las ceremonias religiosas involucran fiestas oficiales, seguramente ligadas con las estaciones, y ceremonias particulares con ofrendas para pedir servicios especiales. El Sacerdote de una tribu T ha quedado fuertemente conmovido por la muerte violenta de uno de sus allegados y ha decidido poner su granito de arena para que haya paz en la comarca. Por eso, ha decidido imponer fuertes sacrificios de perdón a todos los asesinos. En un tiempo antes de la nueva política, durante el cual se hicieron más de 200 sacrificios, el costo de cada sacrificio de perdón podía evaluarse en promedio en 1000 con una desviación de 100. Pero los últimos 5 sacrificios, después de la nueva política, tienen un costo promedio de 2000 con una desviación de 200. Cuando se supo la noticia, vino una comitiva a solicitar una rebaja de las tarifas. ¿Se justificada tal solicitud?

40 . El Sacerdote de la tribu J se corrompió. Aceptando prebendas personales ya no obliga a sacrificios costosos para limpiar el pecado. Sin embargo, la gente sigue viniendo por su bendición y así el descubre que tiene mucho más poder de lo que pensaba. Por tanto, ha decidido imponer una ley en la cual un día al año le traigan ofrendas especiales para él y para su familia. Antes de esta ley, digamos durante unos 2 años, recibía en promedio 45 visitas por mes con una desviación de 6, pero ahora, después de la ley, en los últimos 5 meses sólo ha recibido un promedio de 21 visitas con una desviación de 11. ¿Le convendrá al Sacerdote tomar el mismo una dosis de arrepentimiento?

41 . Debido a que el derecho se sustenta con la coacción, las leyes están en general bien adaptadas a su medio y en sociedades agrícolas pueden ser muy estables porque están ligadas a la tierra. Pero la tribu H ha venido fortaleciéndose en su papel de comerciante, lo cual la distancia de la tierra y de la estabilidad legal que ella garantiza. Ellos tenían una ley que era transmitida de generación en generación desde tiempos remotos. Pero ahora ha habido 5 reformas a la ley en los últimos 9 años. ¿Es justo pensar que tanta reforma se debe al nuevo oficio?

42 . Al ir cambiando de oficio, se cambia de leyes. La ley va tras la costumbre. Y eso es precisamente lo que enerva al Sacerdote de la tribu L. Así que ha llamado a todos para pegarles fuerte regaño y les dijo: "¿Acaso la belleza de los tulipanes cambia de año en año y con cada viento? Pues no. Como fue, así es, y como es, así será. De igual forma, hay sólo una ley, una ley eterna, una ley que está en cada estómago. Si esa ley les hace falta, vengan a mí, que yo la despertaré desde el fondo de sus sucias conciencias. Pero por lo demás, armen sus propios instrumentos de justicia y castigo." Sin darse cuenta, este Sacerdote ha creado una revolución jurídica: ahora hay leyes fundamentales, religiosas, y también leyes de costumbre, leyes civiles. Cada una con su estamento legal y de justicia. Ahora el líder del clan se ha tornado en alguien cuya autoridad va a la par con la del Sacerdote. Ha nacido el prototipo de Estado con separación de poderes religiosos y civiles. En una comarca, hay 1 tribu entre una muestra al azar de 10 donde hay separación de poderes y en otra hay 7 de una muestra al azar de 20. ¿Podemos decir que hay diferencia entre las dos comarcas en cuanto a su madurez jurídica?

43 . En una muestra de un grupo de tribus se contabilizó el número de leyes religiosas y civiles, como se indica por parejas (5,12), (5,13), (6,22), (5,17), (4,8), (4,6), (8,30), (10, 45). ¿Podemos decir que hay diferencia promedio entre el número de leyes religiosas y civiles?

44 . Cuando la tribu M fue creciendo, los problemas de convivencia aumentaron, por lo que decidieron escindirse en varias porciones y que cada se fuera por su lado. En el año antes de la división se presentaban 25 querellas en promedio por mes con una desviación de 3, pero en los últimos 7 meses y en todas las porciones juntas, hay 15 en promedio por mes con una desviación de 4. ¿Fue sabia la decisión de separarse?

45 . Entre más gente, más oportunidades de trueque. Y también más posibilidades de querellas y de violencia. Pero el poder de los negocios gana en esta lucha y, por eso, no hay forma de parar el crecimiento de los conglomerados humanos. El antiguo Jefe de una Tribu, quien reinó 12 años, decía que acostumbraba a tener menos de una pesadilla relacionada con violencia por año. Pero en los últimos dos años el nuevo jefe ha tenido 7. ¿Puede decirse que la violencia ha aumentado tomando esta tribu como una muestra al azar de toda la región?

46 . Hace frío y llueve. El Jefe de la tribu K está recostado mirando al vacío. En realidad, está intrigado por tratar de entender por qué el número de querellas por mes que tiene que resolver es más o menos constante y siendo que involucran a personas diferentes y a razones diferentes. En los últimos meses ha atendido el siguiente número de querellas: 7, 8, 9, 7, 8, 9, 7, 8, 7. ¿Se justifica la apreciación del jefe de que el número de querellas por mes es más o menos constante?

47 . El Jefe ha indagado qué pasa en la tribu vecina B. Allí, las querellas de los últimos meses han sido: 5, 8, 9, 8, 10, 4, 7, 5, 8, 9, 5. ¿Puede decirse que aquí si hay variabilidad con respecto a los datos del problema anterior? ¿En donde es mayor la variabilidad en la tribu K, del problema anterior, o en la B?

48 . En esa parte del mundo, las tribus saben contar y llevan memoria de la edad de las personas. a) Podemos pensar que en todas las tribus la edad de iniciación matrimonial de las niñas es aproximadamente la misma y que está condicionada por la biología. Para saber si eso es verdad o no, reportamos las edades de algunas niñas, muestra al azar, a las cuales se hace la ceremonia: 12, 13, 12, 14, 15, 12, 13, 14, 14, 15. ¿Respaldan estos datos la tesis discutida? b) Por otra parte, los muchachos no tienen rito especial pues se casan tan pronto pueden. Las edades de matrimonio de algunos jóvenes son: 16, 17, 15, 16, 17, 20, 18, 17, 16, 14. ¿Puede considerarse homogénea la edad de casarse de los varones? c) Las niñas se casan al año del rito de iniciación. ¿Son las edades promedio de las niñas y los jóvenes iguales al momento de casarse?

49 . El Jefe acaba de despertarse sobresaltado. Su esposa le pregunta angustiada sobre lo que se estaba soñando. El responde: "Vi una visión y aún la estoy viendo, en vívidos colores y muchos gritos, en que nosotros sacrificábamos a nuestros hermanos de la tribu vecina como ofrenda a los dioses, para que se acabara la sequía y que entre más hermanos sacrificábamos, mejores eran las cosechas. Aunque muchas tribus hacen eso, nosotros no lo hacemos y hemos podido supervivir y lo seguiremos haciendo. Pero de todas maneras ese sueño me mortifica, porque la estamos viendo negra. Lo que sí es necesario es estar prevenidos, no sea que nuestros hermanos sean los que nos lleven a nosotros a ser sacrificados. Además que han sido varios los robos del poco alimento que tenemos". En la tribu se sacrificaba cada 15 años a un adolescente nacido de su misma sangre, pero no a cualquier dios sino al Creador de todo, para que quedara claro que lo reconocían como el dador de la vida, de la lozanía, de la esperanza y de la fortaleza. En el vecindario, en cambio, había sacrificios humanos abundantes. En este año de sequía las cifras por tribu en una muestra al azar han sido las siguientes: 18, 20, 12, 9, 21, 13, 17, 8, 7. Pero en general, no sacrifican sino dos o tres por año. ¿Puede decirse que ha aumentado el número promedio de sacrificios humanos por la sequía?

50 . Un talismán es mucho más que un colmillo. Un buen talismán debe, sobre todo, haber sido bendecido por una alma santa. El Sacerdote que vive en las rocas goza de buen prestigio. Ultimamente ha esculpido los dientes que le llevan a santificar. Les ha puesto un venado, o una garza, o una danta, o lo que el cliente diga. Antes del cambio, y durante unos 9 años, le traían entre 25 y 35 colmillos por año. Pero ahora, y en menos de un año, ya ha esculpido más de 50. ¿Ha funcionado la estrategia?

51 . Una manifestación cultural de los pueblos indígenas americanos fueron sus casas más bien separadas pero unidas por caminos que estaban rodeados de matas florecidas. Eso no se había visto en Europa. En un pueblito, una muestra al azar de 8 casas tenía el siguiente número de matas: 20, 22, 18, 24, 21, 15, 17, 19. En otro pueblito había 12, 13, 15, 13, 14, 15, 13, 14, 12. a) ¿Puede decirse que el número de matas era algo más bien fijo para cada aldea? b) ¿Puede decirse que en un pueblito había más matas en promedio que en el otro?

10.2. III Parcial: con folleto de fórmulas

Este parcial es sin libro ni apuntes. Es con calculadora y el folleto de fórmulas que viene con el libro. **Lo primero que debe hacerse** es entender bien cada problema. A continuación se clasifica en una y sólo una de las categorías siguientes y se aplica la metodología indicada.

1. Comparación de una media muestral con una media poblacional conociendo la desviación poblacional. Se usa una z. Ej 192, pág. 168.
2. Comparación de la diferencia de dos medias experimentales con la diferencia esperada de dos medias poblacionales conociendo las varianzas poblacionales: se usa una z. Capítulo próximo.
3. Comparación de una media muestral con una media poblacional conociendo la desviación de la muestra: se usa una t. Ej 193, pág. 169.
4. Comparación de una proporción observada o frecuencia relativa con una probabilidad (poblacional): se usa una z. Ej 194, pág. 170.
5. Comparación de la diferencia de dos proporciones experimentales con la diferencia esperada de las probabilidades. Se usa a z y la t. Ej 195, pág. 172.
6. Comparación de la varianza de una muestra al azar con una varianza poblacional: se usa una chi-cuadrado χ^2. Capítulo próximo.

7. Comparación de la relación de dos varianzas experimentales con la relación esperada de varianzas poblacionales: se usa una F. Ej 197, pág. 176.

8. Comparación de la diferencia de dos medias experimentales con la diferencia esperada de dos medias poblacionales conociendo las dos varianzas muestrales de las cuales se infiere que las varianzas poblacionales son iguales. Primero se estudia si las varianzas poblacionales son o no iguales, por tanto es con dos colas y con una F, caso 7. Si las varianzas poblacionales resultan iguales, se comparan las medias (puede ser con una o con dos colas, depende del problema) con una t usando el método del ejemplo 198, pág. 179.

9. Comparación de la diferencia de dos medias experimentales con la diferencia esperada de dos medias poblacionales conociendo las dos varianzas muestrales de las cuales se infiere que las varianzas poblacionales son diferentes. Primero se estudia si las varianzas poblacionales son o no iguales, por tanto es con dos colas y con una F. Si las varianzas poblacionales son diferentes, se usa una t para comparar medias siguiendo el ejemplo 199, pág. 183.

10. Comparación de la diferencia de dos medias experimentales con la diferencia esperada de medias poblacionales para datos apareados (miopía ojo izquierdo vs. miopía ojo derecho): se usa una t. Ej 196, pág. 174.

Ayuda obligatoria: use su sentido común para tener una idea de la respuesta, si el factor enunciado causó un efecto o nó. Con dicho sesgo cuadre la matemática.

Ayuda para encontrar el caso que resuelve el problema. Se lee la pregunta del problema y se busca la palabra clave de la pregunta:

- Media: la nota promedio fue 4.2. Si no hay desviación o varianza, no se trata de la media, pues la media no vale sin la desviación: la media es el representante democrático de la población en tanto que la varianza mide el descontento de la población por dicha representación.

- Varianza (desviación): las notas fueron muy inhomogéneas, la mínima fue cero, la maxima 5 y todas estaban desparramadas por ahí. La varianza fue 1.5. La desviación fue de 1.2

- Porcentaje, frecuencia relativa o proporción observada o experimental: de 30 personas 18 pasaron que equivale al 60 %.

- Datos apareados: las novias sacaron mejor nota promedio que sus novios pues no dejan las tareas para el último momento.

Se tiene en cuenta si se trata de una muestra o de dos y por tanto de una media o de dos, o lo que sea. A continuación se busca el caso mirando los datos:

- Una media observada conociendo la varianza (desviación) poblacional $->$ caso1.

- Dos medias observadas conociendo las varianzas (desviaciones) poblacionales $->$ caso2.

- Una varianza muestral $->$ caso 6.

- Dos varianzas de dos muestras $->$ caso 7.

- Dos medias observadas conociendo las varianzas (desviaciones) muestrales$->$ caso 8 ó 9. Si las varianzas son iguales $->$ caso 8. Si son diferentes $->$ caso 9.

- Datos apareados con dos muestras $->$ caso 10.

- Una frecuencia relativa o proporción observada: caso 4.

- Dos frecuencias relativas o proporciones observadas: caso 5.

Cuando se da el tamaño de la muestra, la media, desviación y varianza son de la muestra. Cuando se dice que es lo usual, los datos son poblacionales. Las preguntas siempre se refieren a las poblaciones, pues son ellas las que sufren las decisiones que uno tome y es sobre ellas que uno formula la H_o. Para este parcial III la H_o es que el factor estudiado no produce ningún cambio en la v.a. observada: las medias poblacionales no cambian (la diferencia de medias da cero), las varianzas poblacionales son las mismas (su relación de quebrado da uno).

206 ***Ejemplo para mostrar que un problema de una cola puede resolverse tanto por la cola inferior como por la superior pero que hay que ser coherente en todos los pasos.***

En una muestra al azar de 25 jóvenes universitarios, 20 confesaron que le tenían temor al futuro. Pero el correspondiente número en otra muestra independiente de jóvenes que tenían trabajo, fue de 12 entre 29. ¿Hay datos suficientes para decir que tener trabajo es mejor para crear seguridad que estar estudiando?

SOLUCIÓN 1

Paso 1: Se encuentra el caso, se asignan las variables, se plantea la H_o, el número de colas y el estadígrafo de contraste:

Caso 5 subcaso 2.

$f = X/m = 20/25 = 0{,}8$, frecuencia relativa de los que tienen temor del grupo de los que estudian.

$g = Y/n = 12/29 = 0{,}414$, frecuencia relativa de los que tienen temor del grupo de los que trabajan.

$H_o : p - r = 0$ y una cola. Trabajar da lo mismo que estudiar respecto a la incertidumbre del futuro.

$\hat{p} = \frac{x+y}{m+n} = \frac{20+12}{25+29} = 0{,}5926$

$s_{Bi} = \sqrt{(\hat{p} * (1-\hat{p}) * (1/m + 1/n))} = 0{,}134$

$t_{obs} = \frac{(f-g)-(p-r)}{s_{Bi}} = \frac{0{,}386}{0{,}134} = 2{,}8800$

Paso 2: Se formula H_a, se halla t_c, se hace la gráfica, se toma un veredicto respecto a H_o, se interpreta el veredicto en términos de la pregunta del problema:

Formulemos la H_a:

para que tener trabajo sea mejor que estar estudiando, la frecuencia relativa de los que tienen pánico del grupo de los que trabajan g debe ser menor que aquella de los que estudian f. En dicho caso, $f - g$ sería positiva y la t_{obs} sería igualmente positiva:

$H_a : p - r > 0$, cola superior.

glis $= 25 + 18 = 43 > 30$:

$t_c = z_c = 1{,}64$ con significancia del 5 %.

Veredicto: la t_{obs} mide la discrepancia entre lo que se ve y lo que se cree en la H_o. Como t_{obs} es mayor, más extrema, que t_c, la discrepancia es grande. Por tanto rechazamos lo que creemos: se rechaza H_o.

Interpretación: los que trabajan la tienen más fácil que los que estudian respecto a la incertidumbre del futuro. Suficiente decir que los que estudian ni siquiera saben si obtendrán su diploma.

SOLUCIÓN 2

Paso 1: Se encuentra el caso, se asignan las variables, se plantea la H_o, el número de colas y el estadígrafo de contraste:

Caso 5 subcaso 2.

Asignamos las variables al revés:

$g = X/m = 20/25 = 0{,}8$, frecuencia relativa de los que tienen temor del grupo de los que estudian.

$f = Y/n = 12/29 = 0{,}414$, frecuencia relativa de los que tienen temor del grupo de los que trabajan.

$H_o : p - r = 0$ y una cola. Trabajar da lo mismo que estudiar respecto a la incertidumbre del futuro.

$\hat{p} = \frac{x+y}{m+n} = \frac{20+12}{25+29} = 0{,}5926$

$s_{Bi} = \sqrt{(\hat{p} * (1-\hat{p}) * (1/m + 1/n))} = 0{,}134$

$t_{obs} = \frac{(f-g)-(p-r)}{s_{Bi}} = \frac{-0{,}386}{0{,}13} = -2{,}8800$

Paso 2: Se formula H_a, se halla t_c, se hace la gráfica, se toma un veredicto respecto a H_o, se interpreta el veredicto en términos de la pregunta del problema:

Formulemos la H_a:

para que tener trabajo sea mejor que estar estudiando, la frecuencia relativa de los que tienen pánico del grupo de los que trabajan f debe ser menor que aquella de los que estudian g. En dicho caso, $f - g$ sería negativa y la t_{obs} sería igualmente negativa:

$H_a : p - r < 0$, cola inferior.

glis $= 25 + 18 = 43 > 30$:

$t_c = z_c = -1{,}64$ con significancia del 5 %.

Veredicto: la t_{obs} mide la discrepancia entre lo que se ve y lo que se cree en la H_o. Como t_{obs} es menor, más extrema, que t_c, la discrepancia es grande. Por tanto rechazamos lo que creemos: se rechaza H_o.

Interpretación: los que trabajan la tienen más fácil que los que estudian respecto a la incertidumbre del futuro. Suficiente decir que los que estudian ni siquiera saben si obtendrán su diploma.

207 ***Versión 1:***

1. La ciudad C no tiene metro. Se discute una decisión de sí o no acerca de construir uno. Habló el político A y se preguntó a una muestra al azar de 23 personas su decisión: 11 respondieron sí y 12 no. A continuación habló el político B. Se entrevistaron 17 personas al azar, 13 de las cuales apoyaron el sí y 4 el no.

a) ¿Podrá decirse que el político B obtuvo una mayor frecuencia relativa de positivos que A? b) ¿Cuál es la muestra en este problema? c) ¿Es 40 la población? y sino, ¿cuál es? d) ¿Es la pregunta a de estadística descriptiva o de inferencial y por qué? e) ¿Es acertado decir que como el político B obtuvo una mayor frecuencia relativa de positivos entonces el causó un cambio de opinión? f) ¿Es la pregunta e de estadística descriptiva o inferencial y por qué? g) ¿Qué es lo que hace la estadística inferencial? h) ¿Causó el político B un cambio de opinión (en la población)? i) El político B está seguro de que el sí causó un cambio de opinión (en toda la población). ¿Qué aconseja Usted para que el trate de probar lo que piensa?

2. Hay dos partidos dominantes. El número de curules en una muestra al azar de 7 años por cada partido ha sido

Número de curules							
Partido	Año 1	Año 2	Año 3	Año 4	Año 5	Año 6	Año 7
Transparencia	23	22	24	23	25	24	27
Honestidad	11	13	15	11	17	15	13

¿Hay razón para creer que la ventaja del momento seguirá del mismo lado en el futuro?

3. La Contraloría alega que tanto gasto de dinero en la reparación de la carretera A no puede ser. Para ello argumenta que en el año pasado con otro gobernador el gasto fue menor que ahora y sin embargo el régimen de lluvias ha disminuido en promedio con respecto al año pasado. Los datos que presenta la Contraloría son para el año pasado: precipitación de 140 en promedio con desviación de 0.1 sobre una muestra al azar de 15 mediciones. Para este año: promedio de 100 con una desviación de 50 sobre una muestra de 9 mediciones. a) Demuestre que la acusación de la Contraloría es correcta. b) Usted fue contratado como abogado para presentar una apelación y Usted la ganó: ¿Qué fue lo que hizo? Ayuda: haga una gráfica de las campanas y para que gane el pleito divida adecuadamente el eje horizontal en rangos de precipitación de lluvia: sequía, llovisna, aguacero, diluvio.

208 ***Versión 2***

1. En el Callejón de la Muerte, el número de muertes por trimestre en una muestra al azar de 8 trimestres del decenio pasado fue: 8, 6,4,6,8,7,6,4. En otra muestra al azar de 9 trimestres del presente decenio, el correspondiente número ha sido: 12,13,12,14,13,9,8,11, 10. ¿Puede decirse que las medias poblacionales son distintas? Tome $\alpha = 0{,}02$. Ayuda 1: no son datos apareados, pues tienen diferente

número de registros y no hay indicación de que los datos correspondan a meses homólogos. Ayuda 2: con los métodos del primer parcial saque las medias y las varianzas, aplique el método 7 para comparar varianzas poblacionales y luego el 8 ó 9 para comparar medias poblacionales.

2. Una jovencita que quiere hacer carrera como pop-star puso su sitio web y todo el cuento y ahora se preocupa por evaluar la fidelidad de sus fans (para ello se registra el IP de cada visita, lo cual no viola ninguna privacidad). La tabla de frecuencias absolutas siguiente da la distribución del número de visitas a su sitio de muestras al azar y no apareadas de fans tanto para el mes de mayo como para el mes de junio. Decida si el promedio de visitas aumentó, tal como ella lo esperaba. Ayuda para entender la tabla: pongamos como ejemplo el cuarto renglón: nos dice que de la muestra de mayo hubo 7 fans que visitaron el sitio 4 veces en tanto que para la muestra de junio fueron 8 los que visitaron su sitio 4 veces. Ayuda: Se calculan medias y varianzas por los métodos del parcial I. Para poder comparar medias poblacionales, primero debemos comparar varianzas (poblacionales) para decidir si éstas son iguales o diferentes (ejemplo de los pollitos). Como nos interesa saber si las varianzas poblaciones son iguales o no, el estudio es con dos colas. Pero el número de colas en la comparación de medias (ejemplos de los germanismos y del Buda) depende del contexto del problema. En este caso, lo clave es que la pop-star quiere saber si las visitas aumentaron en promedio. Ejemplo 201, pág 187.

Frecuencias absolutas del número de visitas		
x : # visitas	F1: Mes pasado	F2: Presente mes
1	4	3
2	8	5
3	5	8
4	7	8
5	2	4
6	1	4
Total		

3. Un candidato a la alcaldía tuvo un error garrafal en un debate televisado. De 30 personas entrevistadas que habían pensado votar por él, 22 consideran que el error no influirá en la conducta de los votantes. De 23 personas que habían pensado no votar por él, 20 consideran que ese error cambiará decididamente la conducta de los votantes. ¿Hay entre los dos grupos diferencia estadísticamente significativa de la percepción del efecto del error sobre la conducta de los votantes?

209 ***Versión 3***

1. Doce muchachos que se inscribieron en ingeniería electrónica tuvieron en su examen de admisión un puntaje promedio de 4.5 con una desviación de 0.2. Los de civil eran 24 y su promedio fue de 4.2 con una desviación de 0.4. a) ¿Puede decirse que los de electrónica son mejores en promedio que los de civil? b) Si la media tradicional para electrónica es de 4.6, ¿ tiene sentido decir que los muchachos recién inscritos manchan la tradición de su clase?

2. **Problema importante para entender por qué no se puede comparar tablas de frecuencias absolutas por el caso 10**. Consideremos la siguiente tabla que muestra la distribución de edades de dos grupos. El primer grupo es de niños de alrededor de 10 años y el segundo es de padres de alrededor de 40 años.

Distribución de las edades de dos grupos		
$Edad$ = en años	$Frec$ grupo niños	$Frec$ grupo padres
9	1	0
10	2	0
11	1	0
39	0	1
40	0	2
41	0	1

a) Halle las medias y varianzas, pruebe que las varianzas poblacionales son iguales. Pruebe que las edades promedio son diferentes. b) Notemos que la tabla tiene la estructura de datos apareados donde el criterio de apareamiento es la edad. Por tanto, el problema debe poder resolverse por el método de comparación de medias para datos apareados. Demuestre que con dicho método las medias son iguales. c) Debido a que los resultados se contradicen, ¿por cuál método opta Usted? ¿Por qué el método que da resultados falsos no puede usarse?

3. Juan y Pablo comparten un apartamento. Juan le dijo a Pablo: Usted es más vago que yo y eso es fácil de demostrar pues yo estudié la semana pasada 12 horas y en cambio usted no estudió sino 7. ¿Es verdad lo que dice Juan? Ayuda: determine la unidad de tiempo y relacione los datos con ella.

210 *Versión 4*

1. De una muestra al azar de 23 profesores y empleados de la U, 7 no parecen interesados en lo que pasa en el país. Pero después de un paro que bloqueó las calles de llegada a la U, de otra muestra al azar de 25, los no interesados fueron 3. ¿Son efectivos los paros para concientizarnos? Use $\alpha = 0{,}03$.

2. Una estudiante decidió hacer dieta para no engordar y se ha dado cuenta de que además de todo se le mejoró la visión por lo que ya no tiene que sentarse en la primera fila para ver el tablero. De una muestra de 17 niñas que no hacen ningún tipo de dieta, la distancia máxima a la cual pueden ver el tablero tiene una media de 5 metros y una desviación de 2. De otra muestra de 12 niñas que hacen dieta, la media correspondiente fue de 6 metros con una desviación de 1.5. ¿Es aconsejable hacer dieta para mejorar la visión? Tome $\alpha = 0{,}02$.

3. Una muestra al azar de 5 encuestadoras se demora en promedio 23 minutos por encuesta con una desviación de 8 minutos. Otro grupo de 7 se demora 16 minutos con una desviación de 2. La Señora Coordinadora ha comenzado a sospechar que, debido a que se les paga según el número de encuestas realizadas, el segundo grupo hace trampa, pues parecen conseguir muy rápidamente voluntarios que les respondan. Tome $\alpha = 0{,}02$.¿Se justifica de algún modo la sospecha de la Coordinadora?

211 *Versión 5*

1. De cuando en cuando aparecen muertas algunas personas que no parecen tener nada en común excepto que pueden ser mendigos, drogadictos, ladrones y prostitutos. Los muertos de esta clase en dos ciudades de igual población en una muestra al azar apareada de 7 meses han sido (3,2), (2,3), (2,5), (2,0), (3,0), (4,1), (2,1). ¿Se justifica decir que ambas ciudades son iguales en promedio en cuanto al número de muertos por mes del grupo mencionado?

2. María y Manuela se pelean por Juan. a) Manuela dice que Juan la llama en promedio por semana más a ella que a María. Pero María dice: demuéstramelo. Es muy simple, dice Manuela, aquí está la tabla de frecuencias absolutas de muestras al azar de las veces por semana que Juan nos ha llamado:

Distribución de las llamadas de Juan en # de semanas		
x = Veces por semana	*Frec* de María	*Frec* de Manuela
3	3	3
4	3	3
5	2	2
6	2	3
7	1	2
8	1	2
9	0	3
Total		

Por ejemplo, han sido 2 las semanas que Juan ha llamado a María 6 veces, en tanto que han sido 3 las semanas que Juan ha llamado a Manuela ese mismo número de veces. ¿Es cierto lo que alega Manuela? b) María ha notado que Juan también llama a Mónica y que la tabla de llamadas de muestras al azar es como sigue:

Distribución de las llamadas de Juan a Mónica	
Veces por semana	# de semanas
5	1
6	3
7	1
Total	

María se consuela diciendo que Juan llama más a Mónica que a Manuela. ¿Es eso verdad?

212 ***Versión 6***

Lea el material del Taller III, pág. 283 hasta donde dice *Ejemplo: distribución binomial*. Siga los ejemplos para resolver los siguientes problemas.

1. Todo el mundo resuelve los problemas fáciles pero entre más difíciles, menos son los que lo hacen. Una compañía que diseña exámenes propone uno de 5 puntos, que según ella tiene 4 problemas fáciles y uno difícil. Se toman grupos al azar de 30 personas y los que los resolvieron fueron: 25, 23, 22, 24, 6. ¿Es justa la garantía dada por la compañía?

2. La pelea por la colonización de Marte va a buen ritmo. De hecho, ya se corren simulaciones de los viajes de ida y vuelta, los cuales pueden durar aproximadamente 2 años. El historial en exploración espacial para antes del 2000 para 6 potencias se ajusta a las siguientes proporciones: 3, 2, 10, 1, 1, 1. Pero en el último decenio ha habido los siguientes misiones, listadas en el orden correspondiente: 20, 11, 73, 5, 6, 22. ¿Confirman los datos el decir de los analistas de que una nueva super-potencia está en acción y que el futuro dueño de Marte se torna misterioso?

Capítulo 11

Los 10 experimentos básicos

213 ***Objetivo:*** *Reunir en un solo diccionario todas las variantes de los 10 experimentos básicos. Resumiremos lo visto y haremos algunas extensiones inmediatas. Cuando se le haya olvidado todo y necesite repasar, mire todas las gráficas de este libro, cuando llegue a la sección 5.7, pág. 85, léala toda, lo mismo con el ejercicio 179, pág. 150. Después siga mirando las gráficas y concéntrese en este capítulo que tiene todo lo necesario en relación con las normales. Los árboles de probabilidades y las tablas de contingencia están en el siguiente capítulo. Hay talleres complementarios con temas adicionales.*

11.1. Introducción

214 *Una normal está totalmente descrita por la media y la desviación. Y por otro lado, la normal es muy común en la naturaleza y en la sociedad. Por eso, los experimentos básicos de la ciencia se han dirigido a dilucidar la media y la varianza o a comparar medias y varianzas de poblaciones normalmente distribuidas. Como vimos en el parcial II, el estudio de las proporciones y de la binomial también puede acomodarse dentro del de la normal.*

Ofrecemos diez casos, diez experimentos básicos. Cuando a uno le planteen un problema, uno trata de acomodarlo dentro de alguno de dichos casos y después se aplica la tecnología correspondiente. La media poblacional se denota μ y se considera que es la media de muchísimos datos al azar, o también la media de un censo. La media muestral es $\bar{X}_n$, que es la que viene de un muestreo, de un experimento. La desviación poblacional se denota σ y la muestral como s. La varianza muestral se denota $s^2 = VAR = s_X^2 = VAR_X = Var_X$.

Siempre asumimos que las muestras son al azar y que no son demasiado pequeñas, que las variables tienen una distribución normal y que si se trata de dos variables, ellas son independientes. Contraejemplo: suponiendo que tenemos cursos de 50 alumnos, no podemos aplicar un test de comparación de medias del número de los que pierden los quices contra el número de los que los ganan, pues la segunda información es redundante.

11.2. La lista de los 10

Los 10 experimentos básicos y más renombrados de la ciencia son los siguientes, algunos de los cuales ya los vimos en el capítulo pasado:

1. Estudio de una media sabiendo la desviación poblacional σ. Ejemplo: comparar una media experimental con una media teórica conociendo la desviación poblacional.

2. Comparar dos medias de X, Y sabiendo σ_X, σ_y. Ejemplo: comparar la diferencia de medias observadas con la diferencia esperada de dos medias teóricas o poblacionales conociendo las varianzas poblacionales.

3. Estudiar una media conociendo la desviación de una muestra al azar. Ejemplo: comparar una media experimental con una media teórica conociendo la desviación de la muestra.

4. Estudiar una proporción. Ejemplo: comparar una frecuencia relativa con una probabilidad (la frecuencia relativa de la población).

5. Experimentos con dos proporciones. Ejemplo: comparar la diferencia entre dos frecuencias relativas con la diferencia esperada de dos probabilidades.

6. Estudio de una varianza. Ejemplo: comparar una varianza de una muestra con una varianza poblacional.

7. Comparación de dos varianzas. Ejemplo: comparar la relación de quebrado entre dos varianzas muestrales con la relación de quebrado esperada entre las varianzas poblacionales.

8. Comparación de dos medias sabiendo, por el estudio de las varianzas de las dos muestras correspondientes, que las varianzas poblacionales son iguales. Hay dos casos:

   a) Si la H_o es que la diferencia de medias es cero.

   b) Si la diferencia de medias postulada en la H_o es cualquiera, cero o no. Ejemplo: comparar la diferencia de medias observadas con la diferencia esperada de dos medias teóricas o poblacionales conociendo las varianzas muestrales y habiendo inferido que las varianzas poblacionales son iguales.

9. Comparación de dos medias sabiendo, por el estudio de las varianzas de dos muestras, que las varianzas poblacionales son diferentes. Ejemplo: comparar la diferencia de medias observadas con la diferencia esperada de dos medias teóricas o poblacionales conociendo las varianzas muestrales y habiendo inferido que las varianzas poblacionales son diferentes.

10. Comparación de medias para datos apareados (miopía ojo izquierdo vs. miopía ojo derecho). Ejemplo: comparación de la diferencia de medias de datos apareados de dos muestras con la diferencia poblacional esperada.

Advertencia: en el estudio de todas las hipótesis nulas debe tenerse siempre en cuenta que *el método científico es un contraste entre lo que se ve y lo que se cree.*

11.3. ¿Cómo se decide un experimento?

Todos los experimentos de la ciencia se pueden acomodar dentro del siguiente esquema:

1. Se plantea una hipótesis básica que es la que hay que poner a prueba y que se denomina la H_o, hipótesis nula. Puede ser algo nacido de una teoría, o de la experiencia intuitiva de un investigador experimentado, o de los trabajos previos, o de trabajos relacionados. Por ejemplo, que una media vale 8, o que una varianza vale 10, o que una diferencia de medias vale 4, o que una varianza es 3 veces más grande que otra, o que una diferencia de proporciones vale 0.2, o que el ojo usado por los relojeros para mirar sus relojes (en el cual se ponen un monoscopio) es una dioptría más miope que el otro ojo. También se plantea una H_a que lo único que dice es si se trata de una prueba con una o con dos colas.

2. Se hace un experimento o un muestreo al cual se le mide lo que sea el objeto de estudio del experimento y lo exigido por la metodología. Por ejemplo, si nuestro objetivo es comparar la media bajo dos condiciones, para tomar un veredicto se miden las medias pero también se requiere cuantificar las varianzas.

3. Se mide la discrepancia entre lo predicho por la H_o y lo observado en el experimento o en el muestreo. Para ello se utiliza un estadígrafo de contraste.

4. Se decide si la discrepancia entre la H_o y lo observado es pequeña o grande. Para ello se compara con el valor crítico del estadígrafo dado por la tabla para la significancia y con los grados de libertad correspondientes. Si el valor de la discrepancia, medida por el estadígrafo de contraste, es más extrema que el crítico, la discrepancia es grande. Pero si la discrepancia hallada no es más extrema que el valor crítico, la discrepancia es pequeña.

5. Se toma un veredicto: Si la discrepancia es pequeña, tal discrepancia se explica por azar, por efecto del muestreo, y se acepta la H_o. Si es grande, se rechaza la H_o y se dice que debe haber un efecto sistemático afectando el sistema. ¿De dónde sale este formalismo? Todo se debe a que la tabla tiene, para una significancia dada, el valor crítico del estadígrafo de contraste, el cual ha sido computado haciendo la distribución de lo que el estadígrafo de contraste daría si no hubiese nada más que el azar: la tabla calcula el valor crítico de la H_o.

11.4. Notación

El valor crítico de un estadígrafo depende de la significancia y del número de colas. El subíndice α **indica una sola cola** y por tanto toda la significancia va a la cola en referencia, en tanto que el subíndice $(\alpha, 2)$ **nos indica que se toman dos colas** y por tanto a cada cola va la mitad de la significancia. Por ejemplo: $z_{(\alpha,2)} = z$ crítico con significancia α y dos colas. Lo mismo con los demás estadígrafos.

La media muestral se denota $\bar{X_n}$, la poblacional μ. La varianza muestral se denota $VAR = VAR_X = s^2 = s_X^2$, la poblacional $\sigma^2 = \sigma_X^2$. La desviación muestral $s = s_X$, la poblacional $\sigma = \sigma_X$.

11.5. El diccionario de experimentos

Veamos ahora el detalle de los 10 experimentos básicos. Hay ejemplos para cada caso. Al final tenemos un resumen de todos los casos con sus procedimientos.

1. **Comparar la media muestral con la poblacional sabiendo σ, la desviación poblacional.** Puede contrastarse una creencia respecto a la media, es decir, podemos decidir una H_o, o podemos hallar el intervalo de confianza de la media dado una media muestral, la cual viene de un experimento concreto.

   Resumen para la media: si se sabe experimentalmente el valor de una media muestral $\bar{X_n}$, y si se conoce la desviación poblacional σ_X, y si se da una significancia α, entonces tenemos:

   a) Para comparar una media experimental con una media teórica conociendo la desviación poblacional, la $H_o : \mu = \mu_o$ dice: la discrepancia entre lo que se registre $\bar{X_n}$ y la media μ_o debería ser cero de no ser por el azar, representado por la desviación de la distribución.

   Para evaluar la hipótesis nula $H_o : \mu = \mu_o$ se puede usar el estadígrafo de contraste

   $$z = \frac{\bar{X_n} - \mu_o}{\frac{\sigma_X}{\sqrt{n}}}$$

   Este estadígrafo mide la discrepancia, en unidades apropiadas, entre lo encontrado experimentalmente $\bar{X_n}$ y lo predicho por la H_o que es μ: si la discrepancia es pequeña, se acepta la H_o, pero si es grande, se rechaza. El límite entre lo grande y lo pequeño está dado por la z crítica asociada a la significancia, $z_{(\alpha,2)}$ para dos colas, z_α para la cola inferior y $z_{1-\alpha}$ para la cola superior. Rechazar la H_o es equivalente a decir que es sensato creer que hay algún efecto sistemático, además del azar, que hace que el resultado experimental $\bar{X_n}$ se desvíe de la media.

   b) El intervalo de confianza $1 - \alpha$ para la media es el intervalo que reune a todas las medias para las cuales el dato experimental es un dato común y corriente y está dado por:

   $$\bar{X_n} - (z_{(\alpha,2)})\frac{\sigma_X}{\sqrt{n}} < \mu < \bar{X_n} + (z_{(\alpha,2)})\frac{\sigma_X}{\sqrt{n}}$$

   El intervalo de confianza para la media es tal que si uno apuesta a que la media poblacional está en dicho intervalo, uno acierta una proporción $1 - \alpha$ de las veces que apuesta. Eso se escribe así:

   $$Prob(\bar{X_n} - (z_{(\alpha,2)})\frac{\sigma_X}{\sqrt{n}} < \mu < \bar{X_n} + (z_{(\alpha,2)})\frac{\sigma_X}{\sqrt{n}}) = 1 - \alpha$$

c) Al término $\frac{\sigma_X}{\sqrt{n}}$ se le llama **error estándar de la estimación de la media** y es igual a la desviación estándar de la media muestral $\bar{X}_n$.

d) Para saber la media poblacional con total exactitud hay que hacer un censo, lo cual puede costar mucho. Ejemplo: el valor de un censo general para Colombia ronda por los 200000 millones de pesos. Lo inteligente es hacer una muestra pero hay que admitir un error en la precisión. Si se desea estimar una media poblacional de tal forma que la mitad de la longitud del intervalo de confianza sea no superior a E, lo cual da una medida de la incertidumbre de la estimación, hay que tomar **un número de datos no inferior a**

$$n = ((z_{(\alpha,2)})\tfrac{\sigma_X}{E})^2.$$

Una muestra nunca debe ser cero, y en caso de tener cifras decimales, se debe aproximar al número entero mayor.

215 ***Ejemplo*** *Si una v.a. tiene desviación 16 y la significancia es 0.05, por lo cual la z-crítica* $z_{(\alpha,2)}$ *es 2, y* $\bar{X_{36}} = 40$ *entonces, resolvamos las siguientes 4 preguntas:*

a*) Estudiémos la* $Ho : \mu = 20$ *que se rechaza con dos colas.*

b*) Hallemos el intervalo de confianza para la media poblacional.*

c*) Hallemos el error estándar en la estimación de la media poblacional.*

d*) Usemos* σ *para decidir el número de datos que se necesitaría para estimar la media poblacional con un error estándar no superior a* $E = 5$*. Tómese una significancia del 5 %.*

Solución:

a*) Podemos usar dos métodos para decidir la* $H_o : \mu = 20$ *con* $H_a : \mu \neq 20$*. El primero es decir que 20 no está en el intervalo de confianza de la media y por lo tanto se rechaza la* H_o*. El segundo es medir la discrepancia entre lo esperado y lo hallado:*

$z = \frac{\bar{X_n} - \mu_o}{\frac{\sigma_X}{\sqrt{n}}}$

$z = \frac{40-20}{\frac{16}{\sqrt{36}}} = \frac{20}{2{,}66} = 7{,}5$

Esta discrepancia es muy grande, pues la z crítica es 1.96, y la z experimental es 7.5. Por lo tanto, rechazamos la H_o*. Es decir, aconsejamos no creer que* $\bar{X_{36}} = 40$ *pueda ser explicado diciendo que la media vale 20 y que hay una efecto del azar dado por la desviación 16, sino que es sensato considerar que hay algún otro efecto sistemático afectando el resultado.*

b*) El intervalo de confianza para la media es:*

$40 - (1{,}96)(\frac{16}{\sqrt{36}}) < \mu < 40 + (1{,}96)(\frac{16}{\sqrt{36}})$

$40 - (1{,}96)(\frac{16}{6}) < \mu < 40 + (1{,}96)(\frac{16}{6})$

$40 - 5{,}22 < \mu < 40 + 5{,}22$

$34{,}78 < \mu < 45{,}22$

c*) El error estándar en la estimación de la media poblacional es la desviación de la media muestral* $= \frac{16}{\sqrt{36}} = 2{,}66$*.*

d*) Si se desea estimar una media poblacional con tolerancia o error estándar (igual a la mitad de la longitud del intervalo de confianza) no superior a* $E = 5$*, hay que tomar* ***un número de datos no inferior a*** *:*

$n = [(z_{(\alpha,2)})(\frac{\sigma_X}{E})]^2$ *donde* $z_{(\alpha,2)}$ *representa el z crítico con significancia* α *y con dos colas.*
$n = [(1{,}96)\frac{16}{5}]^2 = 39{,}3 = 40.$

216 ***Ejemplo*** *Hallemos el intervalo de confianza para la media si* $\sigma = 3$*,* $n = 25$*,* $\bar{X}_{25} = 18$*,* $\alpha = 0,05$

Respuesta: el intervalo de confianza para μ con $\alpha = 0,05$ es
$\bar{X}_n - z_{(\alpha,2)}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n + z_{(\alpha,2)}\frac{\sigma}{\sqrt{n}}$
es decir
$(18 - (1,96)\frac{3}{\sqrt{25}}, 18 + 1,96\frac{3}{\sqrt{25}}) = (16.8,\ 19.12)$

217 ***Ejemplo*** *La longitud media de un pez PPP en el estrato E4.*

Se estudia la longitud de un pez, la cual es una v.a. X con $\sigma_X = 4$. La literatura también dice que $\mu = 20$. Pero en el estrato E4 tenemos 16 datos con media $\bar{X_{16}} = 17$. Asumimos que la media muestral tiene una distribución normal. El intervalo de confianza para la media μ con $\alpha = 0{,}05$ es

$\bar{X}_n - (z_{(\alpha,2)})\frac{\sigma_X}{\sqrt{n}} < \mu < \bar{X}_n + (z_{(\alpha,2)})\frac{\sigma_X}{\sqrt{n}}$
$\bar{X_{16}} - 1{,}96\frac{4}{\sqrt{16}} < \mu < \bar{X_{16}} + 1{,}96\frac{4}{\sqrt{16}}$
$\bar{X_{16}} - 1{,}96(1) < \mu < \bar{X_{16}} + 1{,}96(1)$
$17 - 1{,}96(1) < \mu < 17 + 1{,}96(1)$
$15{,}04 < \mu < 18{,}96$

Observemos que 20 no está en este intervalo. Esto implica que el promedio de la literatura, basada en muchos datos, y la media obtenida por el estudiante indican cosas diferentes. Eso se debe a que el estudiante escarba en un estrato más bajo, más antiguo que aquel del cual se obtuvieron los datos de la literatura. La salida es alegar que la longitud del pez cambió con el tiempo.

218 ***Ejemplo*** *Hay un batallón que necesita mandar a hacer uniformes para sus soldados. Un grupo de 100 necesitan aproximadamente 300m de tela en promedio para sus uniformes. ¿Cuál sería el intervalo de confianza para la media por grupo, si $\sigma = 5$ y $\alpha = 0,05$?*

Solución: el intervalo de confianza está determinado por

$\bar{X}_n - z_{(\alpha,2)}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n + z_{(\alpha,2)}\frac{\sigma}{\sqrt{n}}$

donde $z_{(\alpha,2)}$ representa el z crítico para dos colas con significancia α. Según nuestra tabla, para $\alpha = 0{,}05$ el $z_{(\alpha,2)}$ está entre 1.9 y 2. Un valor casi exacto es 1.96. Reemplazando los valores obtenemos:
$300 - 1,96\frac{5}{\sqrt{100}} < \mu < 300 + 1,96\frac{5}{\sqrt{100}}$

$299{,}02 < \mu < 300{,}98$

219 ***Ejemplo*** *Si el error tolerable en la estimación de la media es $E = 2$, y si $\sigma_X = 4$, $\alpha = 0{,}05$ o sea $z = 2$, entonces*
$n = (2(4)/2)^2 = 4^2 = 16$.
Pero si $\epsilon = 0{,}1$ entonces $n = (2(4)/(0{,}1))^2 = 80^2 = 6400$.

Vemos que es muy costoso aumentarle precisión a una estimación. Para decidir de antemano el error tolerable en la precisión de la estimación de la media conviene tener en cuenta la desviación estándar: se vería muy mal proponer una tolerancia mayor a una desviación estándar. Supongamos que elegimos una tolerancia de un séptimo la desviación estándar:
$n = ((z_{(\alpha,2)})\frac{\sigma_X}{E})^2$.

$n = (2\frac{\sigma_X}{\sigma/7})^2$.

$n = (2 \times 7)^2 = 14^2 = 196 = 200$, pero también eso es mucho. Conformémonos con una precisión de la tercera parte de una desviación estándar:

$n = (2\frac{\sigma_X}{\sigma/3})^2 = 6^2 = 36$.

220 ***Ejemplo*** *Para fabricar una pintura que usaban los indígenas, pero dándole un toque moderno, se prepara una fécula de raíces que viene en polvo y sus partículas tienen un diámetro del orden de milímetro. Se desea saber si hay necesidad de cambiar de molino, pues ya está muy viejo, para lo cual se requiere una estimación exacta de la media con una tolerancia de una micra, o sea* $1 \times 10^{-6}m$ *. Se sabe que* $\sigma = 0{,}02mm = 2 \times 10^{-5}m$*. ¿Cuántas partículas deben estudiarse? Tómese* $\alpha = 0{,}05$*.*

Solución:

$n = \left[\frac{1{,}96(2*10^{-5})}{10^{-6}}\right]^2$

$n = 39, 2^2 = 1536, 64 \approx 1537$

Este trabajo microscópico es tan dispendioso que es bueno tratar de hacerlo utilizando medios fotográficos y densitometría. Pero en la práctica, la gente puede llegar a hacer todo eso simplemente restregando un poco de polvo entre sus dedos índice y pulgar.

2. **Comparar la diferencia de medias muestrales con la diferencia de medias poblacionales sabiendo las varianzas poblacionales:**

   Prueba de dos colas (ejemplo arbitrario) : $H_o : \mu_x - \mu_Y = a$ y $H_a : \mu_x - \mu_Y \neq a$

   Prueba de una cola: $H_o : \mu_x - \mu_Y = a$ y $H_a : \mu_x - \mu_Y < a$

   Estadígrafo de discrepancia entre lo observado $\bar{X_m} - \bar{Y_n}$ y lo esperado según H_o de la diferencia de dos medias: $\mu_X - \mu_Y$

$$z = \frac{(\bar{X_m} - \bar{Y_n}) - (\mu_X - \mu_Y)}{\sqrt{\frac{{\sigma_X}^2}{m} + \frac{{\sigma_Y}^2}{n}}}$$

   El intervalo de confianza para $\mu_X - \mu_Y$ se da por

$$(\bar{X_m} - \bar{Y_n}) - z_{(\alpha,2)}\sqrt{\frac{{\sigma_X}^2}{m} + \frac{{\sigma_Y}^2}{n}} < \mu_X - \mu_Y < (\bar{X_m} - \bar{Y_n}) + z_{(\alpha,2)}\sqrt{\frac{{\sigma_X}^2}{m} + \frac{{\sigma_Y}^2}{n}}$$

221 ***Ejemplo*** *¿Mejoró o no mejoró el atleta en esta semana en comparación con su desempeño de la semana pasada en el número de segundos para recorrer 70 metros? Asumimos que la desviación en ambos casos es la misma,* $\sigma = 1$*.*

Diez datos de la semana pasada: 12, 9, 8, 9, 7, 9, 8, 7, 8, 9 con media muestral 8,6*;* $m = 10$*. Nueve datos de esta semana: 12, 12, 11, 9, 10, 9, 10, 9, 11 con media muestral 9;* $n = 9$*.*

Debemos estudiar la hipótesis nula:

$H_o : \mu_X = \mu_Y$ (ahora está lo mismo que antes).

la cual se rechaza contra la hipótesis alterna:

$H_a : \mu_X \neq \mu_Y$ (Ahora está mejor o peor que antes).

Tomamos $\alpha = 0{,}05$ que da un z-crítico de 2.

Lo que nosotros sabemos es que el estadígrafo

$\frac{(\bar{X_m} - \bar{Y_n}) - (\mu_X - \mu_Y)}{\sqrt{\frac{{\sigma_X}^2}{m} + \frac{{\sigma_Y}^2}{n}}}$

tiene un distribución z. Nuestra notación es:

$\bar{X_m}$= la media muestral de la semana pasada = 8,6, $m = 10$;

$\bar{Y_n}$= la media muestral de esta semana = 9, $n = 9$.

μ_X y μ_Y son las medias poblaciones de la semana pasada y presente respectivamente. Recordamos que la media poblacional corresponde al caso en el cual se toma un número infinito de datos.

De acuerdo a la H_o:

$\mu_X - \mu_Y = 0$

Y sabemos que

$\sigma_X = \sigma_Y = 1$

Calculamos la z asociada a nuestro experimento:

$z = \frac{(8,6-9)-0}{\sqrt{\frac{1^2}{10}+\frac{1^2}{9}}} = \frac{-0,4}{\sqrt{\frac{1}{10}+\frac{1}{9}}} = \frac{-0,4}{0,46} = -0,87$

Veredicto: Dado que $|-0,87|$ es menos que el valor crítico 2, nosotros concluimos que la diferencia entre las dos muestras pudiera ser atribuible a las variables no controladas y que si bien uno ve una tendencia, aún no se ha logrado probar.

Cuando hay una tendencia que no se ha podido probar, lo que hay que hacer es reunir más datos, pues puede ser que la diferencia real entre el control y el experimental sea muy pequeña. Es lo mismo que pasa cuando uno negocea con empanadas, con cosas de poco valor: hay que vender muchas para poder hacerse a una riqueza. Y si uno produce unas empanadas un poquito mejores, hay que vender muchas para poder demostrar que en realidad la producción ha mejorado.

222 ***Ejemplo*** *Las notas de toda la carrera de Juan, un estudiante de Matemáticas que toma Historia para complementar la parte de sociales, son como sigue:*

$\mu_{HISTORIA} = 3,2$

$\sigma^2_{HISTORIA} = 1$

$\mu_{MATEMATICAS} = 4,5$

$\sigma^2_{MATEMATICAS} = 0,2$

Juan conoció al principio del semestre a una muchacha, Marta, que estudia Historia. Ahora tenemos que el promedio de sus dos materias en los 4 parciales son:

$n_1 = n_2 = 4$

$\bar{X}_{HISTORIA} = 4,2$

$\bar{X}_{MATEMATICAS} = 3,8$

Al ver el cambio en las notas, la muchacha en tono justiciero dice: ¿Para qué se estudia historia sino es para cambiar el curso del mundo?

Discusión: Antes de conocer a Marta, las notas de Juan en matemáticas eran mejores que sus notas en Historia. Pero después, sus notas promedio de Historia fueron mejores que las de matemáticas y además no le fue tan bien en Matemáticas como en el pasado. Esos cambios son un hecho en la vida de Juan y acá no hay estadística aplicable: lo que pasó en su vida, pasó y nada puede añadirse, nada puede quitarse, pues se trata de hechos perfectamente definidos. Pero no tenemos información suficiente para saber si la muchacha es la culpable del cambio de notas. Con todo, el contexto parece apoyar esa idea.

Ahora bien, la muchacha no dice que ella cambió a su amigo, cosa que todo el mundo parece saber, sino que dice que ella tiene la potencialidad de cambiar al mundo. Para apoyar su tesis podemos hacer un test que toma las notas de Juan como un muestreo aleatorio de las notas de todos los posibles amigos que ella pudiese tener en vez de Juan.

Comparamos con nuestro estadígrafo de discrepancia la diferencia de notas promedio entre Historia y Matemáticas antes y después de la aparición de Marta:

$$z = \frac{(x_1 - x_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$z = \frac{(3{,}8 - 4{,}2) - (4{,}5 - 3{,}2)}{\sqrt{\frac{0{,}2}{4} + \frac{1}{4}}} = 3{,}10$$

este valor es significativo tanto para una como para dos colas con una significancia de 0.05. Por lo tanto, la muchacha se jacta de algo que la estadística no puede desmentir.

3. **Comparar una media de una muestra con una media poblacional conociendo la desviación de la muestra.** Si se tiene la lista de datos, la desviación se calcula por:

$s = \sqrt{\frac{\Sigma(x_i - \bar{X})^2}{n-1}} = \sqrt{\frac{n(\Sigma x^2) - (\Sigma x)^2}{n(n-1)}}$

Prueba de dos colas: $H_o : \mu_x = a$ y $H_a : \mu_x \neq a$

Prueba de una cola: $H_o : \mu_x = a$ y $H_a : \mu_x < a$

Estadígrafo de contraste entre lo observado $\bar{X}_n$ y μ_X, lo esperado según H_o, :

$t_{obs} = \frac{\bar{X}_n - \mu_X}{\frac{s_X}{\sqrt{n}}}$

Para decidir si la discrepancia entre lo predicho por la H_o es grande o pequeña, se compara la t hallada con el valor crítico dado por la tabla de la t según la significancia. Pero atención, entre más datos hay más exactitud, por consiguiente debe haber algún parámetro que registre el número de datos: se llama 'grados de libertad', gli y es igual al número de datos menos uno: $gli = n - 1$.

223 ***Ejemplo***

Para reducir la tensión con los estudiantes de estadística, un profesor cuchilla hace el siguiente pacto de honor:

"Si la nota promedio del curso es menor que 3, subo lo mismo a la nota de todo el mundo para que el promedio llegue a 3".

El promedio de las notas de sus 20 estudiantes fue 2,8. Pero el profesor no subió nada. Después de mucho pelear, llegaron a la conclusión de que, siendo un curso de estadística, se sobreentendía que el enunciado explícito del pacto debía ser:

"El examen está calibrado para que la nota promedio sea 3. Por consiguiente, si la nota promedio del curso es estadísticamente menor que 3, subo lo mismo a la nota de todo el mundo para que el promedio llegue a 3".

Teniendo en cuenta que la desviación de las notas fue de 1. ¿Deberá subir el profesor 0,2 a todo el mundo?

Solución: Primero establezcamos la diferencia entre las dos lecturas del pacto. En el primer pacto sólo se trataba de una diferencia de números, sobre lo cual no hay y no puede haber alegato. En la segunda formulación sí hay mucho que decir, por ejemplo: la nota promedio del curso es *estadísticamente* menor que 3 si al repetir muchas veces el mismo curso y aplicar exactamente la misma metodología, el promedio de las notas de todos esos cursos es menor que tres. El problema es que tan sólo tenemos un curso, el presente. Procedemos así: Formulamos las hipótesis en juego:

$H_o : \mu = \mu_o = 3$ (si repetimos el curso muchas veces, el promedio es 3 y naturalmente que algunos cursos sacarán menos que 3, como es el caso presente).

$H_A : \mu < \mu_o$ (este curso está calibrado para que el promedio sea tres, pero ustedes lo perdieron por vagos y la prueba es que la diferencia entre 3 y 2.8 es enorme por tratarse de promedios y no se explica bien por cosas del azar).

Para dilucidar la pelea, calculamos la discrepancia entre lo predicho por la H_o y lo hallado experimentalmente:

$t_{obs} = \frac{2,8-3}{\frac{1}{\sqrt{20}}} = -0,91$

La discrepancia entre lo predicho por H_o, que la media es 3, y lo que pasó en la realidad del curso presente, promedio 2.8, es-0.91. ¿Será grande o pequeña dicha discrepancia? Para saberlo, miramos en la tabla la t-crítica.

Con 19 gli, $\alpha = 0,05$ y cola inferior, la t-crítica es $-1,729$. Por lo tanto, la discrepancia entre lo esperado y lo hallado es pequeña y no hay razón para creer que dicha discrepancia no se explique por los juegos normales del azar: no hay diferencia estadística entre 3 y 2.8.

Veredicto. El profesor no subirá nada: el profesor calibró bien su examen para que la nota promedio fuese 3. La divergencia actual entre el 3 y lo que sacaron en promedio se debe a cosas del azar. Pero queda una salvedad: la doble lectura del pacto pudo haber hecho perder la motivación de muchos. Además puede parecer injusto calibrar un examen para que el promedio sea 3. El alegato seguramente continuará.

224 ***Ejemplo*** *En un experimento, 9 datos tenían una media de 23 y una desviación de 2. Calcular el intervalo de confianza para la media para* $\alpha = 0,01$.

La solución: En este caso, nosotros tenemos: $n = 9$, $\bar{X}_n = 23$, $s = 2$. La t asociada a $\alpha = 0,01$, con dos colas y 9-1=8 grados de libertad, es $t = \pm 3,355$. El intervalo de confianza para la media se da por:

$\bar{X_n} - (t_{(\alpha,2)})\frac{s_X}{\sqrt{n}} < \mu < \bar{X_n} + (t_{(\alpha,2)})\frac{s_X}{\sqrt{n}}$

donde $t_{(\alpha,2)}$ representa la t crítica superior de significancia α y dos colas.

$23 - 3,355\frac{2}{\sqrt{9}} < \mu < 23 + 3,355\frac{2}{\sqrt{9}}$

$23 - 3,355\frac{2}{3} < \mu < 23 + 3,355\frac{2}{3}$

$23 - 3,355(0,666) < \mu < 23 + 3,355(0,666)$

$23 - 2,23 < \mu < 23 + 2,23$

$20,77 < \mu < 25,23$

225 ***Ejemplo*** *Comparemos una media experimental con una media de la literatura y especifiquemos el intervalo de confianza para nuestra media. En un muestreo de 20 datos se obtuvo* $\bar{X}_{20} = 7$ *cuando la media de la literatura es* $\mu = 5$. *Nuestra s, la desviación de la muestra, fue de 1.3.*

Hacemos una prueba de dos colas para la comparación de lo hallado con lo esperado según la literatura:

$H_o : \mu = 5$.

$H_a : \mu \neq 5$.

Para ello calculamos la discrepancia entre lo hallado y lo esperado:

$t_{obs} = \frac{\bar{X}_n - \mu}{s/\sqrt{n}}$

$t_{obs} = \frac{7-5}{1,3/\sqrt{20}}$

$t_{obs} = 6,88$

Ahora comparamos la $t_{obs} = 6,88$ con la t crítica según la tabla: con 19 gli y significancia 0.05 y con dos colas, la t crítica es ± 2.093 (pág. 158).

Veredicto: con una significancia del 5 %, concluimos que nuestros datos provienen de una población con media diferente a la considerada por la literatura.

Ahora el **Intervalo de confianza**:

$7-(2,093)\frac{1,3}{\sqrt{20}} < \mu < 7+(2,093)\frac{1,3}{\sqrt{20}}$
$7-0{,}61 < \mu < 7+0{,}61$
$6{,}39 < \mu < 7{,}61$

Vemos que la media de la literatura, que es 5, no pertenece a este intervalo: dicho valor es algo extraño para nuestros datos.

226 ***Ejemplo*** *El número de huevos puestos por las reinas de 20 colmenas ha tenido un promedio de 200 por día. Se desea incrementar la producción de huevos, pues se trata de una casta muy apacible. En un día en el que se ofreció alimento extra y apropiado a las abejas, las 20 colmenas produjeron* $\bar{X}_{20} = 500$ *huevos en promedio con una* $s_{20} = 100$. *Decida si el alimento funcionó y halle el intervalo de confianza para la media. Tómese* $\alpha = 0{,}01$.

Solución: prueba con una cola:
$H_o : \mu = 200$
$H_a : \mu > 200$

calculamos la discrepancia entre lo esperado y lo hallado:
$t_{obs} = \frac{\bar{X}_2 0-\mu}{s/\sqrt{n}}$
$_{obs}t = \frac{500-200}{100/\sqrt{20}} = 13{,}41$
la t crítica con una cola es $t(\alpha = 0{,}01) = 2{,}539$
Veredicto: se rechaza la H_o con g.l= 19 y $\alpha = 0{,}01$. La discrepancia entre la producción promedia antigua y la nueva queda explicada por el azar de forma muy incómoda y se explica mejor por el efecto positivo del alimento extra.

Calculemos ahora el **Intervalo de confianza** para la media de la nueva producción:

$500-2,861(\frac{100}{\sqrt{20}}) < \mu < 500+2,861(\frac{100}{\sqrt{20}})$
$436,06 < \mu < 563,97$

4. **Comparar una frecuencia relativa (de una muestra) con una probabilidad (proporción poblacional) p de una binomial.**

   - Para comparar una frecuencia relativa con una probabilidad (la frecuencia relativa poblacional) la $H_o : p = p_o$ dice: de no ser por el azar, la proporción encontrada de caras al hacer un experimento binomial debería ser p_o.
     Para evaluar la hipótesis nula $H_o : p = p_o$ se puede usar el estadígrafo de contraste

     $$z = \frac{f-p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$$

     Este estadígrafo mide la discrepancia, en unidades apropiadas, entre lo encontrado experimentalmente, f, y lo predicho por la H_o que es p_o: si la discrepancia es pequeña, se acepta la H_o, pero si es grande, se rechaza. El límite entre lo grande y lo pequeño está dado por la z crítica asociada a la significancia y que se denota $z_{(\alpha,2)}$ para dos colas, z_α para la cola inferior y $z_{1-\alpha}$ para la cola superior.

- El intervalo de confianza $1-\alpha$ para la p es el intervalo que reune a todas las p para las cuales el dato experimental es un dato común y corriente. Si $f=\frac{k}{n}$, el intervalo está dado por:

$$f-(t_{(\alpha,2)})\sqrt{\tfrac{f(1-f)}{n}}<p<f+(t_{(\alpha,2)})\sqrt{\tfrac{f(1-f)}{n}}$$

  Los $gli=n-1$. Usamos una t pues la desviación empleada viene de una muestra. Si $n>30$ podemos aproximar la t por la z.
  El intervalo de confianza para p es tal que si uno apuesta a que p está en dicho intervalo, uno acierta una proporción $1-\alpha$ de las veces que apuesta. Eso se escribe así:

$$Prob(f-(t_{(\alpha,2)})\sqrt{\tfrac{f(1-f)}{n}}<p<f+(t_{(\alpha,2)})\sqrt{\tfrac{f(1-f)}{n}})=1-\alpha$$

- Si se desea estimar una p sabiendo una estimación $\breve{p}$ y con un error estándar, la mitad de la longitud del intervalo de confianza, E, hay que tomar **un número de datos no inferior a**

$$n=((z_{(\alpha,2)})\tfrac{\sqrt{\breve{p}(1-\breve{p})}}{E})^2=((z_{(\alpha,2)})^2)\tfrac{\breve{p}(1-\breve{p})}{E^2}.$$

  Si no se conoce nada de p, no se tiene $\breve{p}$, se toma el peor caso, $\breve{p}=1/2$:

$$n=(z_{(\alpha,2)})^2\tfrac{1}{4E^2}=[z_{(\alpha,2)}/(2E)]^2.$$

227 ***Ejemplos*** *Nadie es profeta en su tierra. En efecto, en general los trabajos científicos de los profesores casi no son apreciados en su misma universidad. Según AH, el 30 % de los alumnos de posgrado de su facultad consideran que su mejor trabajo es bueno. a) Decidir el tamaño de la muestra para poder poner a prueba lo que el profesor dice si se desea estimar la probabilidad de que su mejor trabajo sea reconcocido como bueno con un error no mayor que* 0,1. *b) En un muestreo al azar de* 20 *estudiantes,* 5 *estuvieron de acuerdo. ¿Es justo lo que dice el profesor? c) Hallemos el IC correspondiente.*

- Deseamos estimar p sabiendo que $\breve{p}=0{,}3$ con $E=0{,}1$. Por tanto, hay que tomar **un número de datos no inferior a**

$$n=((z_{(\alpha,2)})^2)\tfrac{\breve{p}(1-\breve{p})}{E^2}.$$
$$n=(1{,}96)^2\tfrac{0{,}3(0{,}7)}{0{,}1^2}=(3{,}84)\tfrac{0{,}21}{4(0{,}01)}=80{,}67.$$

  Si se toma el peor caso, $p=1/2$:

$$n=[z_{(\alpha,2)}/(2E)]^2=[1{,}96/(2*0{,}1)]^2=96{,}04.$$

- Evaluemos la hipótesis nula $H_o: p=0{,}3$ que se rechaza con dos colas y significancia de 0.05 teniendo en cuenta que $f=5/20=1/4=0{,}25$. Para medir la discrepancia entre lo que se ve y lo que se cree se puede usar el estadígrafo de contraste

$$z=\frac{f-p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$$
$$z=\frac{0{,}25-0{,}3}{\sqrt{\frac{0{,}3(0{,}7)}{20}}}$$
$$z=\frac{-0{,}05}{\sqrt{\frac{0{,}21}{20}}}=-0{,}48$$

  Esta discrepancia es pequeña comparada con el valor de $z_c=1.96$. La H_o se acepta.

- El intervalo de confianza $1-\alpha$ para la p, teniendo en cuenta que $f=\frac{k}{n}=\frac{5}{20}=0{,}25$, $\alpha=0{,}05$, $gli=20-1=19$, $t=2{,}093$, está dado por:

$$f-(t_{(\alpha,2)})\sqrt{\tfrac{f(1-f)}{n}}<p<f+(t_{(\alpha,2)})\sqrt{\tfrac{f(1-f)}{n}}$$
$$0{,}25-(2{,}093)\sqrt{\tfrac{0{,}25(0{,}75)}{20}}<p<0{,}25+(2{,}093)\sqrt{\tfrac{0{,}25(0{,}75)}{20}}$$
$$0{,}25-0{,}19<p<0{,}25+0{,}19$$
$$0{,}06<p<0{,}44$$

228 ***Ejemplo*** *Se sabe desde antaño que por cada camiseta que se vende de Millonarios, se venden 2 de Santa Fe ($p_o = 0,33$). Millos goleó a Santa Fe. Se vendieron 300 de Millos y 150 de Santafé. a) ¿Tuvo la goleada efecto en las ventas? b) ¿Tiene una goleada efecto en las ventas?*

Solución: a) El efecto fue claro: la goleada disparó las ventas de las camisetas de goleador. b) Decidamos si podemos esperar que eso siempre pase. Aunque la pregunta literal es de dos colas, el contexto la forza en una sola:

$H_o : p = p_0$

$H_A : p > p_0$

La frecuencia experimental es $f = 300/450 = 0{,}667$. Como tenemos muchos datos, medimos con una z la discrepancia entre lo predicho por la H_o y lo hallado experimentalmente:

$$z = \frac{0{,}667-0{,}33}{\sqrt{\frac{0{,}33(1-0{,}33)}{450}}} = 15,034$$

Con una cola y $\alpha = 0,05$ el z-crítico vale ± 1.64. Como el z experimental nos dio 15, la discrepancia entre lo esperado y lo hallado es demasiado grande para ser explicada por el azar, por el vaivén de las ventas debida a factores no controlados.

Veredicto: H_o se rechaza con cola superor y $\alpha = 0,05$. Una goleada de Millos a Santafé tiene una influencia significativa y positiva en la venta de camisetas.

229 ***Ejemplo*** *En una muestra aleatoria de 85 candidatos a una carrera notarial, 10 resultaron sin conocimientos suficientes. Estime la proporción de rajados en el examen de admisión.*

Solución: el estimador insesgado de la frecuencia relativa de rajados es

$f = \hat{p} = \frac{x}{n} = \frac{10}{85}$=0,12.

El intervalo de confianza de la proporción de rajados p con $\alpha = 0{,}05$ es

$$f - (z_{(\alpha,2)})\sqrt{\tfrac{f(1-f)}{n}} < p < f + (z_{(\alpha,2)})\sqrt{\tfrac{f(1-f)}{n}}$$

donde aproximamos la t por la z por haber muchos datos. Reemplazando:

$$0,12 - 1,96\sqrt{\tfrac{0{,}12(1-0{,}12)}{85}} \leq p \leq 0,12 + 1,96\sqrt{\tfrac{0{,}12(1-0{,}12)}{85}}$$
$$0,051 \leq p \leq 0,189$$

230 ***Ejemplo*** *En una excavación hay dos tipos de cerámica A y B con 12 y 26 especímenes respectivamente. Hallemos el intervalo de confianza para la proporción de objetos tipo A. La literatura dice que en otras regiones la proporción de objetos tipo A es $p = 0{,}39$. ¿Tenemos alguna divergencia?*

Solución: Nosotros tenemos $n = 12 + 26 = 38$, $gli = 37 > 30$, tenemos muchos datos por lo que aproximamos la t por la z. La frecuencia relativa observada de tipo A es $f = 12/38$. Para $\alpha = 0{,}05$, el intervalo de confianza de la proporción p del espécimen de tipo A es

$$f - (z_{(\alpha,2)})\sqrt{\tfrac{f(1-f)}{n}} < p < f + (z_{(\alpha,2)})\sqrt{\tfrac{f(1-f)}{n}}$$

$\frac{12}{38} - 1{,}96\sqrt{\frac{\frac{12}{38}\frac{26}{38}}{38}} < p < \frac{12}{38} + 1{,}96\sqrt{\frac{\frac{12}{38}\frac{26}{38}}{38}}$

$0{,}32 - 2(0{,}075) < p < 0{,}32 + 2(0{,}075)$

$0{,}32 - 0{,}15 < p < 0{,}32 + 0{,}15$

$0{,}17 < p < 0{,}47$

Nosotros hemos obtenido $f = \frac{12}{38} = 0{,}317$ pero los datos publicados para otras regiones ofrecen una proporción $p = 0{,}39$. Para saber si hay divergencia razonamos así:

Nuestro intervalo de confianza para p es (0.17, 0.47) y dicho intervalo contiene la proporción publicada. Por lo tanto, no se ha evidenciado ninguna divergencia. Una solución alterna se encuentra como sigue:

calculamos el estadígrafo

$z = \frac{f-\mu_P}{\sigma_P} = \frac{f-p}{\sqrt{\frac{p(1-p)}{n}}}$

Para nuestros valores, $n = 38$, $f = 0{,}32$, $p = 0{,}39$ tenemos:
$z = \frac{0{,}32-0{,}39}{\sqrt{\frac{0{,}39(0{,}61)}{38}}} = \frac{-0{,}07}{sqrt\frac{0{,}39(0{,}61)}{38}} = \frac{-0{,}07}{0{,}039} = -1{,}79$

El valor crítico asociado a z con $\alpha = 0{,}05$ es $z = 1{,}96$. Esto significa que cada valor entre $(-1{,}96, 1{,}96)$ es considerado ordinario. Y como obtuvimos -1.79, deducimos que nuestro resultado es ordinario y que no tiene nada de raro. No tenemos algo nuevo.

231 ***¿Mejor una Binomial?*** *Para decidir nuestros tests sobre proporciones, es decir, sobre binomiales, hemos usado la Z y la t. Al usar la Z estamos aproximando la bonomial por la normal. Al usar la t estamos tomando en cuenta que cuando aumentamos el número de datos la incertidumbre disminuye y por tanto los IC se acortarían. Pero aún nos queda una protesta que debemos enfrentar: la t y su aproximación Z son perfectas para binomiales cuando p está cerca de 1/2 pues hay simetría en el diagrama de barras de la binomial y queda bien una aproximación por una campana. Pero si p está alejado de 0.5, la aproximación normal sería abusiva.*

¿Qué podemos hacer para tratar de resolver este problema? Lo que podría hacerse sería calcular la(s) binomial(es) exactamente y repetir con ellas todos los cálculos hechos para las campanas. Es muy posible que tengamos que hacer nuestras propias tablas y quizá necesitemos aprender simulación (podría aprenderla en el Vol V El método científico de www.evoljava.com). Como eso no es tan sencillo, optamos por usar la t y la z y cerrar los ojos a la protesta enunciada.

232 ***Ejemplo*** *Dos tipos de cerámica -continuación.*

Hay dos tipos de cerámica A y B y se espera que la proporción de elementos del tipo A sea 0.32. ¿Cuántos especímenes debemos recoger para lograr una exactitud de 0,05 en la estimación de p?

Solución: tomamos $\alpha = 0{,}05$, lo que da un $z = 2$.

$n = [(z_{(\alpha,2)})^2]\frac{p(1-p)}{\epsilon^2}$

$n = (1{,}96)^2\frac{0{,}22}{0{,}05^2} = 4(88) = 352.$

Nosotros hemos asumido que la frecuencia observada no cambiará a medida que excavamos. Pero eso es pedirle demasiado a la vida. Para curarnos en salud, tomamos el valor de p que de la máxima seguridad, el cual es 1/2. Obtenemos:

$n = (1{,}96)^2\frac{0{,}25}{0{,}05^2} = 4(100) = 400.$

5. **Comparar la diferencia entre dos frecuencias relativas $f-g = X/m - Y/n$ de la misma variable aleatoria pero de muestras de poblaciones distintas y al azar con la diferencia esperada de dos probabilidades $p-r$:**

   Prueba de dos colas: $H_o : p-r = a$ y $H_a : p-r \neq a$

   Prueba de una cola: $H_o : p-r = a$ y $H_a : p-r < a$

   Para medir la **Discrepancia entre lo observado $f-g$ y lo esperado $p-r$ para una H_o de diferencia de proporciones** hay que considerar varios casos:

   Caso 1: cuando se da p y r junto con las frecuencias relativas f y g, y los tamaños de las muestras m, n, primero calculamos la desviación conjunta binomial poblacional dada por

   $s_b = \sqrt{\frac{p(1-p)}{m} + \frac{r(1-r)}{n}}$

   y a continuación la discrepancia $= z = \frac{(f-g)-(p-r)}{s_b}$

   Se trata de una z pues s_b se calcula con valores poblacionales.

   Caso 2: Cuando la diferencia esperada es cero, $p-r = 0$, y se dan las frecuencias relativas f, g, los tamaños de las muestras m y n, podemos unir las dos muestras en una conjunta para estimar p:

   $\hat{p} = \frac{X+Y}{m+n} = \frac{fm+gn}{m+n}$

   $s_b = \sqrt{\hat{p}(1-\hat{p})(\frac{1}{m} + \frac{1}{n})}$

   discrepancia $= t = \frac{(f-g)-(p-r)}{s_b}$

   Se trata de una t pues s_b se calcula con valores muestrales. Los gli son $n+m-2$.

   Caso 3: Cuando no se da ni p ni r pero se da la diferencia esperada $p-r \neq 0$ junto con las frecuencias relativas f, g y los tamaños de las muestras m, n:

   $s_b = \sqrt{\frac{f(1-f)}{m} + \frac{g(1-g)}{n}}$

   discrepancia $= t = \frac{(f-g)-(p-r)}{s_b}$

   Puesto que s_b se calcula con valores muestrales de dos binomiales diferentes que aproximamos por normales, tenemos una t con gli igual al menor de $m-1$ y $n-1$. La diferencia de los gli del caso 2 al 3 se debe a que en el caso 2 podemos unir las dos muestras en una sola debido a la H_o que dice que las proporiones son iguales, pero en el caso 3no.

   Intervalo de confianza para la diferencia de proporciones $p-r$:

$$f-g-(t_{(\alpha,2)})\sqrt{\frac{f(1-f)}{m} + \frac{g(1-g)}{n}} < p-r < f-g+(t_{(\alpha,2)})\sqrt{\frac{f(1-f)}{m} + \frac{g(1-g)}{n}}$$

   gli = el mínimo entre $n-1$ y $m-1$. Cuando da mayor que 30 puede usarse una z.

233 ***Ejemplo*** *Se cree que la proporción de negros en la costa Caribe es de 0.3 mientras que la correspondiente proporción en la costa Pacífica es 0.7. Un muestreo de 200 personas en la costa Caribe da 70 negros y otro muestreo de 300 personas en la costa Pacífica da 235 negros. ¿Se respalda por nuestros muestreos la creencia anunciada en cuanto a la diferencia de proporciones? Como el tamaño de la menor muestra es mayor que 30, podemos usar una z en vez de la t.*

Solución:

$H_o : p - r = 0{,}7 - 0{,}3 = 0{,}4$

$H_a : p - r \neq 0{,}4.$

Como tenemos las probabilidades, que son poblacionales, podemos usar el caso 1 con la z: medimos la diferencia entre lo hallado por nuestro muestreo y lo predicho por la H_o por medio del estadígrafo:

$$z_{obs} = \frac{(f-g)-(p-r)}{\sqrt{\frac{p(1-p)}{m}+\frac{r(1-r)}{n}}}$$

donde $X = 235$, $m = 300$, $f = 235/300 = 0{,}78$, $Y = 70$, $n = 200$ y $g = 70/200 = 0{,}35$

$$z_{obs} = \frac{(0{,}78-0{,}35)-(0{,}4)}{\sqrt{\frac{0{,}7(0{,}3)}{300}+\frac{0{,}3(0{,}7)}{200}}} = \frac{0{,}0333}{0{,}0418} = 0{,}79$$

Como no dieron la significancia, la tomamos como 0.05, y como son dos colas, el $z_c = 1{,}96$.

Veredicto: la diferencia entre lo esperado y lo hallado no es lo suficientemente grande como para creer que ella no se deba a efectos del azar en el muestreo: la H_o se acepta.

234 ***Ejemplo*** *Rehagamos el ejemplo anterior usando el tercer caso que para estimar s_b usa muestras y una t. Tomemos la significancia como 0.03:*

Solución:

$H_o : p - r = 0{,}7 - 03 = 0{,}4$

$H_a : p - r \neq 0{,}4.$

Medimos la diferencia entre lo hallado por nuestro muestreo y lo predicho por la H_o por medio del estadígrafo:

$$t_{obs} = \frac{(f-g)-(p-r)}{\sqrt{\frac{f(1-f)}{m}+\frac{g(1-g)}{n}}}$$

donde $f = 235/300 = 0{,}78$ y $g = 70/200 = 0{,}35$

$$t_{obs} = \frac{(0{,}78-0{,}35)-(0{,}4)}{\sqrt{\frac{0{,}78(0{,}22)}{300}+\frac{0{,}35(0{,}65)}{200}}} = \frac{0{,}03}{\sqrt{0{,}00170}} = 0{,}72$$

Como los gli $= 300 + 200 - 2 > 30$ podemos aproximar la t por la z y trabajar con cualquier significancia. Para $\alpha = 0{,}03$ y dos colas, $z_c = \pm 2{,}20$. Vemos que $t_{obs} < t_c$. Podemos tomar nuestro

Veredicto: se acepta la H_o con una significancia del 0.03 y dos colas.

235 ***Ejemplo*** *Hallemos el intervalo de confianza para $p - r$ de acuerdo a nuestros datos del ejemplo anterior pero con $\alpha = 0{,}05$. Como los gli son muy grandes, aproximamos la t por la z:*

Solución:
$f-g-(z_{(\alpha,2)})\sqrt{\frac{f(1-f)}{m}+\frac{g(1-g)}{n}} < p-r < f-g+(z_{(\alpha,2)})\sqrt{\frac{f(1-f)}{m}+\frac{g(1-g)}{n}}$

donde $f = 235/300 = 0{,}78$ y $g = 70/200 = 0{,}35$, $m = 300$ y $n = 200$:

Tenemos que:

$\sqrt{\frac{0{,}78(1-0{,}78)}{300}+\frac{0{,}35(1-0{,}35)}{200}} = \sqrt{0{,}00170} = 0{,}041$

Reemplazando en el intervalo de confianza obtenemos

$0{,}78-0{,}35-1{,}96(0{,}041) < p-r < 0{,}78-0{,}35+1{,}96(0{,}041)$
$0{,}35 < p-r < 0{,}51.$

Veamos qué pasa si usamos abusivamente la fórmula del segundo caso:

$\hat{p} = \frac{X+Y}{m+n}$

$$f-g-(z_{(\alpha,2)})\sqrt{\hat{p}(1-\hat{p})(\tfrac{1}{m}+\tfrac{1}{n})} < p-r < f-g+(z_{(\alpha,2)})\sqrt{\hat{p}(1-\hat{p})(\tfrac{1}{m}+\tfrac{1}{n})}$$

Los datos son $X = 235$, $m = 300$, $f = 235/300 = 0{,}78$, $Y = 70$, $n = 200$, $g = 70/200 = 0{,}35$, $\hat{p} = (235+70)/(300+200) = 0{,}61$
El IC es entonces:

$$0{,}78-0{,}35-(1{,}96)\sqrt{0{,}61(1-0{,}61)(\tfrac{1}{300}+\tfrac{1}{200})} < p-r <$$
$$0{,}78-0{,}35+(1{,}96)\sqrt{0{,}61(1-0{,}61)(\tfrac{1}{300}+\tfrac{1}{200})}$$

Teniendo en cuenta que $\sqrt{0{,}61(1-0{,}61)(\frac{1}{300}+\frac{1}{200})} = 0{,}044$, el IC se simplifica en

$0{,}344 < p-r < 0{,}522$

Veredicto: se acepta la H_o con una significancia del 0.05 y dos colas.

236 ***Variante para usar la t.*** *Cuando los gli son menores que 30 usamos una t para ser muy rigurosos. Para ver cómo funciona, consideremos la variante del problema anterior: f = 8/10, g = 2/6, $H_o : p-r = 0{,}4$. Decidamos la H_o tanto por prueba de hipótesis como por IC.*

$t_{obs} = \frac{(f-g)-(p-r)}{\sqrt{\frac{f(1-f)}{m}+\frac{g(1-g)}{n}}}$

donde $f = 0{,}8$ y $g = 0{,}33$, por lo que $f-g = 0{,}47$

$t_{obs} = \frac{(0{,}8-0{,}33)-(0{,}4)}{\sqrt{\frac{0{,}8(0{,}2)}{10}+\frac{0{,}33(0{,}67)}{6}}} = \frac{0{,}07}{0{,}23} = 0{,}3$

Los gli son el menor entre 10-1 = 9 y 6-1 = 5 que da 5. Al tomar dos colas tenemos que $t_c = 2{,}5$, lo cual da que la H_0 se acepta: la diferencia de la proporción de personas de raza negra entre las dos costas sí puede decirse que sea 0.4.

Si uno halla el IC de la diferencia de medias, tenemos:

$$f - g - (t_{(\alpha,2)})\sqrt{\tfrac{f(1-f)}{m} + \tfrac{g(1-g)}{n}} < p - r < f - g + (t_{(\alpha,2)})\sqrt{\tfrac{f(1-f)}{m} + \tfrac{g(1-g)}{n}}$$

$$0{,}47 - (2{,}5)(0{,}23) < p - r < 0{,}47 + (2{,}5)(0{,}23)$$
$$-0{,}1 < p - r < 0{,}57$$

Como 0,4 pertenece a dicho intervalo, se acepta la H_o: es permitido creer que la diferencia de proporciones de personas de raza negra entre las dos costas es de 0.4.

6. **Comparar una varianza muestral con una poblacional:**

   Prueba de dos colas (ejemplo arbitrario): $H_o : \sigma^2 = a$ y $H_a : \sigma^2 \neq a$

   Prueba de una cola: $H_o : \sigma^2 = a$ y $H_a : \sigma^2 < a$

   Estadígrafo de contraste entre lo observado VAR y lo esperado σ^2 según H_o de una varianza: para las varianzas, lo esperado según la H_o y lo hallado experimentalmente no se compara restando sino **dividiendo**.

   $$\chi^2 = \tfrac{(n-1)VAR}{\sigma^2} \text{ con } n-1 \text{ gli}$$

   Fórmula general para el intervalo de confianza de la VARIANZA σ^2

   $$\tfrac{(n-1)VAR}{\chi^2_+} < \sigma^2 < \tfrac{(n-1)VAR}{\chi^2_-}$$

   donde χ^2_- y χ^2_+ son los valores críticos superior e inferior de la distribución chi-cuadrado. Para hallar el IC de la DESVIACIÓN se saca la raíz cuadrada en todo lado del IC de la varianza.

   Para hallar los valores críticos de la chi cuadrado hay varios casos:

   a) Cuando n es 31 o menos, los valores críticos de la chi-cuadrado se miran en la tabla, pág. 159.

   b) Si $n \geq 32$ los valores críticos se aproximan por:
   $\chi^2_- = n - 1 - z_{(\alpha,2)}\sqrt{2(n-1)}$
   $\chi^2_+ = n - 1 + z_{(\alpha,2)}\sqrt{2(n-1)}$.

   c) Si los gli son mayores que 100 se puede usar la siguiente aproximación para el IC de la desviación:

   $$s - \tfrac{z_{(\alpha,2)}}{\sqrt{2n}} < \sigma < s + \tfrac{z_{(\alpha,2)}}{\sqrt{2n}}$$

   Si preguntan el IC de la varianza, se eleva al cuadrado en todo lado del IC de la desviación.

   Tamaño de la muestra para la estimación de la DESVIACIÓN si por la literatura o por una muestra pequeña se sabe que vale aproximadamente D y si se desea que el error estandar (la mitad del intervalo de confianza) sea E:

   $n = \frac{1}{2}(\frac{z_{(\alpha,2)}D}{E})^2$

237 ***Ejemplo*** *Un locutor aprendiz está a punto de perder su trabajo pues aunque es muy inteligente y bien preparado su voz es muy monótona. Su novia le aconseja: piense en lo que dice, dése cuenta que hay unas palabras más importantes que otras y module la voz consecuentemente. Para saber si el joven está mejorando, le midieron la varianza de la entonación que dio un valor de 4 sobre 9 mediciones cuando el objetivo para esta semana era llegar a 5. ¿Cumplió el joven el objetivo? Tomar* $\alpha = 0{,}02$.

Solución: Tenemos $n-1=8$, $VAR=4$. Además:

$H_o : \sigma^2 = 5$

$H_a : \sigma^2 \neq 5$ (H_o se rechaza con dos colas).

El χ^2 experimental mide la discrepancia entre lo que se ve y lo que se cree:

$\chi^2 = \frac{(n-1)VAR}{\sigma^2} = \frac{8(4)}{5} = 6{,}4.$

Nosotros comparamos este valor con aquellos dados por la tabla de la chi-cuadrado, pág. 159. Como para este problema nosotros rechazamos la H_o con dos colas, entonces la región de aceptación de la H_o es un intervalo cuyos bordes se miran en la tabla. Hay que mirar siempre los DOS VALORES CRÍTICOS debido a que la χ^2 no es simétrica.

Así, para 8 gli con dos colas y $\alpha = 0{,}02$, tenemos que $\beta = \alpha = 0{,}02$ y por tanto $\chi^2_- = 1{,}646$ en tanto que $\chi^2_+ = 20{,}09$. Puesto que nuestro valor experimental de χ^2 pertenece a este intervalo, nosotros concluimos que la hipótesis nula es aceptable: no hay ninguna evidencia experimental para considerar que $\sigma^2 \neq 5$. El joven va cumpliendo el objetivo. ¡Que siga adelante con confianza!.

238 ***Ejemplo*** *La nitidez de la visión se mide evaluando la distancia máxima a la que se pueden distinguir dos puntos como diferentes. La nitidez varía por varias razones, por ejemplo, el tipo de dieta. La visión de un piloto debe ser buena y además buena a todas horas. La nitidez de la visión de un piloto dio en 49 mediciones distanciadas, para que fuesen independientes, una varianza muestral de* 17*. Pero debía ser 8 o menos. ¿Algún problema?*

Como el número de datos es grande, mayor que 30, aproximamos teniendo en cuenta que con $\alpha = 0{,}05$ y dos colas $z = \pm 1{,}96$,:

$\chi^2_- = n - 1 - z_{(\alpha,2)}\sqrt{2(n-1)} = 48 - (1{,}96)\sqrt{2(49-1)} = 48 - (1{,}96)(9{,}8) = 28{,}8$

$\chi^2_+ = n - 1 + z_{(\alpha,2)}\sqrt{2(n-1)} = 48 + (1{,}96)\sqrt{2(49-1)} = 48 + (1{,}96)(9{,}8) = 67{,}2.$

Por consiguiente, la región de aceptación de la $H_o : \sigma^2 = 8$, que se rechaza con dos colas, es (28.4, 67.6).

Calculemos ahora el χ^2 experimental

$\chi^2 = \frac{(n-1)VAR}{\sigma^2} = \frac{(48)17}{8} = 102$

Vemos que este valor es anormal. Podemos decir que la diferencia entre lo que debía ser y lo que resultó ser puede explicarse no por los efectos del azar asociados al tamaño finito del muestreo sino porque el piloto tiene algún problema. A revisión con el oftalmólogo y con el endocrinista.

239 ***Ejemplo*** *En un experimento, 9 datos tenían varianza 4. Calcular el intervalo de confianza para la varianza para* $\alpha = 0{,}02$.

Solución: En este caso, nosotros tenemos: $n-1=8$, $VAR=4$. Los valores de χ^2 asociados a dos colas, 8 grados de libertad y $\beta = \alpha = 0{,}02$, son $\chi^2_- = 1{,}646$ y $\chi^2_+ = 20{,}09$ (tabla de la chi-cuadrado, pág. 159). El intervalo de confianza se da por:

$\frac{(n-1)VAR}{\chi^2_+} < \sigma^2 < \frac{(n-1)VAR}{\chi^2_-}$

$\frac{8(4)}{20,09} < \sigma^2 < \frac{8(4)}{1,646}$

$1{,}59 < \sigma^2 < 19{,}5$

Observamos que la incertidumbre al estimar la varianza es mayor que para la media. Esto significa que con un esfuerzo moderado nosotros podemos estimar la media, el descriptor fundamental de una población. Pero para complementar la información dada por la media de la manera más eficaz, nosotros debemos agregar muchos datos, para estimar la varianza con una incertidumbre pequeña. Decimos complementar la información pues la varianza mide el descontento de la población al ser representada por la media.

240 ***Ejemplo*** *La luz juega un papel trascendente en la preparación de todo tipo de escenarios. A las lámparas de luz casi monocromática se les mide su calidad por la varianza de su espectro de radiación: entre menos varianza son más caras pues su color es más puro. Una de las mejores lámparas del mercado tiene una* $\sigma_X^2 = 8$*. Nuestras observaciones con 10 bombillos de una compañía primípara dieron* $s^2 = 6$*. Averigüemos con* $\alpha = 0{,}01$ *si estas nuevas lámparas son mejores que las de la compañía establecida en el mercado. Calculemos también el intervalo de confianza para la varianza de las frecuencias de radiación para las nuevas lámparas.*

Para comparar nuestros datos con la literatura:

$H_o : \sigma^2 = 8$

$H_a : \sigma^2 < 8$

$s^2 = VAR = 6$

$n = 10$

$g.l = n - 1 = 9$

Medidor de discrepancia relacional

$\chi^2 = \frac{(n-1)VAR}{\sigma^2} = \frac{9*6}{8} = 6,7$

Observemos que la nueva compañía pretende imponerse en el mercado con lámparas de mejor calidad, de menos varianza. Por tanto esperamos que el quebrado $\frac{S^2}{\sigma_0^2}$ sea menor que uno y que la correspondiente χ^2 sea menor que $n-1$. Por tanto, lo que tenemos que averiguar es la relación de nuestro χ^2 experimental con el χ^2 crítico pero del lado izquierdo.

La χ^2 crítico con la cola izquierda y para una significancia del 1 % es 2.088 (el cual se mira en la tabla con la significancia de dos colas tomando $\alpha = 0{,}01$, $\beta = 0{,}02$ y 9 gli). Vemos que el valor experimental de la chi es 6,7 mayor que el valor crítico. Es decir, aún no hay razón para decir que las nuevas lámparas son mejores que las otras.

Como eso de cola derecha y cola izquierda se presta a confusiones, podemos ayudarnos con el cálculo del **Intervalo de confianza**. La fórmula es:

$\frac{(n-1)VAR}{\chi_+^2} < \sigma^2 < \frac{(n-1)VAR}{\chi_-^2}$

Para $\alpha = 0{,}02$ y dos colas obtenemos $\chi_+^2 = 21{,}666$ y $\chi_-^2 = 2{,}088$. Por tanto

$\frac{9*6}{21,666} < \sigma^2 < \frac{9*6}{2,088}$

$2,49 < \sigma^2 < 25{,}86$

La varianza de las lámparas de la compañía antigua es 8 y dicho valor pertenece a nuestro intervalo de confianza, decidimos que no hay diferencia significativa en cuanto a la varianza de los dos tipos de lámpara.

241 *Por el conocimiento previo se sabe que la desviación de una cierta variable está alrededor de 20. Se desea evaluarla con una confianza del 95 % y con un error estandar de 5. ¿Cuántos elementos al azar debe tener la muestra?*

$n = \frac{1}{2}(\frac{z_{(\alpha,2)}D}{E})^2 = \frac{1}{2}(\frac{1{,}96\times 20}{5})^2 = 32$

7. **Comparar la relación de quebrado entre las varianzas de dos muestras independientes y aleatorias con la relación de quebrado entre dos varianazs poblacionales.**

Un experimento que compara varianzas también se llama de **homogeneidad**, mientras que uno que compara promedios también se llama de **productividad**. El primer objetivo es comparar la relación de quebrado entre dos varianzas muestrales con la relación de quebrado entre las varianzas poblacionales respectivas:

El estadígrafo de discrepancia para una H_o de proporción entre varianzas, la MAYOR EN EL NUMERADOR, se define en varios pasos:

primero: lo esperado se consigna en la hipótesis nula:

$H_o : \frac{\sigma_X^2}{\sigma_Y^2} = R$

Decir que las varianzas sean iguales es lo mismo que decir que $R = 1$. Cuando la prueba es con dos colas, la $H_a : R \neq 1$. Cuando la prueba es con una cola, debido a que la varianza mayor siempre se pone en el numerador, la $H_a : R > 1$.

Segundo: lo observado consiste en dos varianzas muestrales: VAR_G y VAR_P a lo cual se le saca la relación entre las dos, dividiendo la grande entre la pequeña:

$R_{obs} = \frac{VAR_G}{VAR_P}$

Tercero: estadígrafo de contraste entre lo observado R_{obs} y lo esperado R:

$F = \frac{R_{obs}}{R} = \frac{VAR_G}{RVAR_P}$

Cuarto: para decidir si la discrepancia entre lo predicho por la H_o es grande o pequeña, se compara la F_c dada por la tabla de la F según la significancia. Pero atención, entre más datos hay más exactitud, por consiguiente debe haber algún parámetro que registre el número de datos tanto del numerador como del denominador: se llama *grados de libertad del numerador*, gli.n. y es igual al número de datos menos uno: $glin. = m - 1$ y *grados de libertad del denominador*, $glid. = n - 1$.

Intervalo de confianza para la proporción de dos varianzas, $\frac{\sigma_X^2}{\sigma_Y^2}$:

$$\frac{VAR_G}{VAR_P F_+} < \frac{\sigma_X^2}{\sigma_Y^2} < \frac{VAR_G}{VAR_P F_-}$$

Teniendo las varianzas muestrales, nos falta mirar en la tabla los valores críticos de la F, al de la cola superior lo denotamos F_+ y al de la inferior F_-. El primero que se averigua es F_+: ese es el valor de la tabla para la significancia dada para dos colas con los grados de libertad en el numerador de la varianza grande y los grados de libertad del denominador para la varianza pequeña.

Después se calcula F_-

$F_- = \frac{1}{I}$ donde I denota la F_+ con los grados de libertad INTERCAMBIADOS.

Por ejemplo, tomemos $\alpha = 0{,}02$ con dos colas, 7 gli numerador y 12 gli denominador. La F de la tabla es 4.64. Ese valor es F_+, el valor crítico superior. Para encontrar F_- buscamos en la tabla de la misma significancia el valor correspondiente a 12 gli en el numerador y 7 gli en el denominador: da 6.47 que es I. Entonces $F_- = 1/6{,}47 = 0{,}155$.

242 ***Ejemplo*** *Unos estudiantes de antropología se unieron para montar un negocio de joyas precolombinas para exportar. Una de las líneas en consideración es la de joyas de plástico recubiertas en oro. Han tenido un problema y es que el recubrimiento puede caerse fácilmente por pedacitos. Después de mucho mirar y mirar, llegaron a la conclusión de que la causa fundamental*

es que es muy difícil hacer un recubrimiento homogéneo y que de hecho en algunos lugares queda muy débil.

Para medir la homogeneidad del recubrimiento se aprovecha la circunstancia de que al escarapelarse quedan gránulos, circulitos de recubrimiento. Se toma una muestra de gránulos y se les mide el diámetro y se les calcula la varianza y la desviación. Entre los métodos de recubrimiento que han probado, el mejor tiene una desviación de $s_2 = 0,061$ *milésimas de pulgada con* $n_2 = 10$*. Pero del SENA le mandaron un especialista que produjo una desviación de* $s_1 = 0,033$ *milésimas de pulgada con* $n_1 = 12$*. ¿Es mejor, más homogéneo, la nueva forma de recubrir?*

Respuesta:

Probemos la hipótesis nula $H_o : \sigma_1^2 = \sigma_2^2$, es decir, $R = 1$, contra la $H_a : \sigma_1^2 < \sigma_2^2$, es decir, $R > 1$, con $\alpha = 0,01$. Esta prueba es con una cola.

La H_o se rechaza si $F \geq 4{,}63$, valor crítico para $\alpha = 0{,}01$ para el cual

$g.l(numerador) = m - 1 = 9$

$g.l(denominador) = n - 1 = 11$

Como $R = 1$, $F_{obs} = R_{obs}/R = R_{obs}$ la calculamos como la varianza mayor sobre la menor:

$F_{obs} = \frac{0{,}061^2}{0{,}033^2} = 3,42$

Veredicto: Como $3,42 < 4,63$ se acepta H_o: no tenemos aún razón para creer que el método propuesto por el especialista sea mejor que lo que conocemos. Pero la solución no es echar al especialista del Sena sino reunir más datos pues de todas formas parece bien sospechar que su método es el mejor.

243 ***Ejemplo*** *Con 18 datos se halló en un primer tratamiento que* $s_1^2 = 482$ *y con 15 datos en un segundo tratamiento que* $s_2^2 = 48$*. Hallemos el intervalo de confianza con* $\alpha = 0{,}02$*, para la relación entre varianzas y decidamos si se justifica decir que las varianzas (poblacionales) son iguales.*

La varianza muestral mayor es la primera, por lo que dividimos la mayor entre la menor y queda una F con 17 gli en el numerador y 14 en el denominador:

Con $\alpha = 0{,}02$ y dos colas, $F_+ = 3{,}60$, el cual está en la tabla. Por otro lado, $F_- = \frac{1}{I}$ donde I denota la F_+ con los grados de libertad INTERCAMBIADOS.

$F_- = \frac{1}{I} = \frac{1}{3{,}39} = 0{,}29$

Reemplazando en el intervalo de confianza

$$\frac{VAR_G}{VAR_P F_+} < \frac{\sigma_G^2}{\sigma_P^2} < \frac{VAR_G}{VAR_P F_-}$$

$$\frac{482}{48(3{,}60)} < \frac{\sigma_G^2}{\sigma_P^2} < \frac{482}{48(0{,}29)}$$

$$2{,}79 < \frac{\sigma_G^2}{\sigma_P^2} < 34{,}6$$

Veredicto: Si las varianzas poblaciones fuesen iguales, la grande entre la chica daría UNO. Como UNO no pertenece a este intervalo, entonces estamos autorizados para decidir que los tratamientos producen varianzas diferentes.

244 ***Variantes*** *Con los mismos datos del ejemplo anterior, pongamos a prueba*

a) *que el primer tratamiento produce una varianza* 30 *veces mayor que el segundo.*

b) *que el primer tratamiento produce una varianza que es más grande que* 15 *veces la del segundo.*

Solución para la primera pregunta usando el IC. Si la varianza grande es 30 veces la menor entonces el quebrado de la varianza grande entre la pequeña da 30. Como 30 pertenece al IC de la relación entre varianzas, es permitido creer que la varianza grande es 30 veces la pequeña.

Solución para la primera pregunta por prueba de hipótesis. Tenemos que decidir entre

$H_o : R = 30$

$H_a : R \neq 30.$

La H_a es con dos colas pues en la formulación de la pregunta hay una afirmación de igualdad cuya negación puede ser de dos maneras, por mayor o por menor.

Con 18 datos se halló en un primer tratamiento que $s_1^2 = 482$ y con 15 datos en un segundo tratamiento que $s_2^2 = 48$.

Usamos la F como medidor de discrepancia relacional:

$F_{obs} = \frac{VAR_G}{VAR_PR} = \frac{482}{48(30)} = 0{,}33$

Como este valor es menor que uno, tenemos que hallar las dos F críticas. La significancia fue definida como $\alpha = 0{,}02$, por lo que con dos colas tenemos:

$F_+ = 3{,}60$

$F_- = \frac{1}{I} = \frac{1}{3{,}39} = 0{,}29$

Como 0,33 pertenece al intervalo de aceptación de la H_o que va de 0,29 a 3,60, aceptamos la H_o: está permitido creer que la primera varianza es 30 veces mayor que la segunda. Obsérvese que es imposible decir que la primera varianza es 40 veces más grande que la segunda.

Intriga: Podemos decir que las varianzas son iguales y también que la una es 30 veces mayor que la otra. Entonces, ¿al fín qué? ¿Son iguales o no son iguales las varianzas? La respuesta es que nosotros no podemos decidir sobre las varianzas pero sí aconsejar sobre lo que dá más confianza para ser creído por todos y que tenga una poca probabilidad de causar decisiones desatinadas.

Estudiémos la segunda pretensión: que el primer tratamiento produce una varianza que es más grande que 15 veces la del segundo. Tenemos que decidir entre

$H_o : R = 15$

$H_a : R > 15.$

EL IC hallado no nos sirve pues todos los IC son con dos colas pero la pregunta es con una sola.

La H_a es con la cola superior pues en la formulación de la pregunta se enfatiza *ser más grande que*. La significancia fue definida como $\alpha = 0{,}025$.

Con 18 datos se halló en un primer tratamiento que $s_1^2 = 482$ y con 15 datos en un segundo tratamiento que $s_2^2 = 48$.

Usamos la F como medidor de discrepancia relacional:

$F_{obs} = \frac{VAR_G}{VAR_PR} = \frac{482}{48(15)} = 0{,}66$

Este valor es menor que uno: ¿tendremos que hallar la F crítica por abajo, por arriba o ambas? Ya sabemos que se trata de una cola. Por lo tanto, es únicamente por un lado. ¿Pero por cuál? Para saberlo notemos que la relación

$F_{obs} = \frac{VAR_G}{VAR_PR}$

se hace más pequeña a medida que R crece, que es lo que se entiende al decir que la varianza uno es más grande que 15 veces la varianza dos. Por lo tanto, la cola de nuestro problema es la inferior. La significancia fue definida como $\alpha = 0{,}025$, por lo que con una cola tenemos que el valor por abajo de la F es:

$F_{=}\frac{1}{I} = \frac{1}{2,72} = 0{,}37$

Por tanto, la región de aceptación de la H_o va desde 0,37 hasta infinito. Como 0,66 pertenece a dicha región, aceptamos la H_o: la primera varianza es 15 veces mayor que la segunda. Además, como 0,66 no queda en la frontera del intervalo, también es permitido decir que la varianza del primer tratamiento es más grande que 15 veces la de la segunda.

245 ***Ejemplo*** *Se trata de Juan, un estudiante que dedica a la Universidad sólo 4 días a la semana. El resto va a trabajar. El joven tiene la sospecha de que es inconstante con su ritmo de estudio, tanto que la varianza de las horas de estudio puede variar enormemente de una semana a la otra. ¿Será eso verdad sabiendo que las horas de estudio de esta semana fueron por día 2, 3, 3, 6, en tanto que las horas de estudio de la semana pasada fueron por día 3, 3, 3 y 4? Resolvamos este problema calculando el intervalo de confianza de la relación entre las varianzas. Tomemos* $\alpha = 0{,}02$.

Solución:

La varianza se calcula por la fórmula (teorema 16 y ejemplo 17, pág. 9):

$s^2 = \frac{n\Sigma x^2 - (\Sigma x)^2}{n(n-1)}$

la cual es más fácil si nos ayudamos de una tabla que contenga tanto los datos como sus cuadrados:

Número de horas de estudio en los 4 días de esta semana		
	x	x^2
	2	4
	3	9
	3	9
	6	36
Total	$\sum x = 14$	$\sum x^2 = 58$

Varianza de las horas de estudio en esta semana

$s_1^2 = \frac{4(58)-196}{4(4-1)} = 3$

Número de horas de estudio en los 4 días de la semana pasada		
	x_i	x_i^2
	3	9
	3	9
	3	9
	4	16
Total	$\sum x = 13$	$\sum x^2 = 43$

Varianza de las horas de estudio la semana pasada

$s_2^2 = \frac{4(43)-169}{4*3} = 0,25.$

Vemos que la varianza grande es $3/0{,}25 = 12$ veces mayor que la pequeña. Es claro que el estudiante tiene razón para preocuparse y creer que es inconstante en su estudio. Sin embargo, si aceptamos que las dos semanas tomadas son muestras al azar de todas las semanas de estudio, nos queda elegir entre dos opciones. Primera: tan grande relación entre las varianzas se debe al azar, a que de un muestreo de dos semanas entre muchas a veces sale una relación y a veces otra. Segunda: dicha relación se debe a que el muchacho realmente es inconstante y la varianza de una semana no se compagina para nada con la varianza de otra semana cualquiera. Eso se decide con una prueba F y dos colas. Tomamos $\alpha = 0{,}02$.

$H_o : \sigma_1^2 = \sigma_2^2$ Entonces $R = 1$
$H_a : \sigma_1^2 \neq \sigma_2^2$

Estudiamos la discrepancia relacional con una F, pero ponemos la varianza mayor en el numerador:

$F_{obs} = F$ de Fisher $= \frac{s^2 grande}{s^2 pequeñaR}$
Como $R = 1$ cuando se estudia la igualdad entre varianzas, tenemos:
$F_{obs} = \frac{3}{0{,}25} = 12$

Nuestra $\alpha = 0,02$ con dos colas, y los grados de libertad son
Numerador: número de datos menos uno de la varianza grande $= 4 - 1 = 3$.
Denominador: número de datos menos uno de la varianza pequeña $= 4 - 1 = 3$.
Por lo que la $F_{crítica} = F_+ = 29,461$

Veredicto: H_0 se acepta: Juan es igual de inconstante de una semana a la otra.

Hallemos ahora el intervalo de confianza para R, la relación entre las varianzas poblacionales, donde usamos el subíndice G de grande y P de pequeña. La fórmula general es:

$$\frac{VAR_G}{VAR_P F_+} < \frac{\sigma_G^2}{\sigma_P^2} < \frac{VAR_G}{VAR_P F_-}$$

Sabemos que $F_+ = 29{,}461$, que la varianza (muestral) grande es 3 y que la pequeña es 0.25. Nos falta averiguar F_-:
$F_- = \frac{1}{I} = \frac{1}{29{,}461} = 0{,}0339$
Por consiguiente, el intervalo de confianza para R, la relación entre varianzas, es:

$$\frac{3}{0{,}25 \times 29{,}461} < \frac{\sigma_G^2}{\sigma_P^2} < \frac{3}{0{,}25 \times 0{,}0339}$$
$$0{,}407 < \frac{\sigma_G^2}{\sigma_P^2} < 353$$

Como $R = 1$ pertenece a este intervalo, entonces aceptamos la igualdad entre las dos varianzas, pues R es la relación entre ellas. Como consecuencia, no apoyamos la idea de Juan de que él es inconstante en su estudio. Sin embargo, lo más sabio es que reuna más datos.

Para la comparación de la diferencia entre dos medias de dos muestras independientes y aleatorias con la diferencia de dos medias poblacionales se requiere primero saber si las varianzas poblacionales son iguales o diferentes, lo cual se hace previamente con un test F, que está en el caso 7. Si las varianzas poblacionales son iguales, se usa el caso 8. Si son diferentes, el caso 9.

8. **Si las varianzas poblacionales se estiman a partir del experimento y resultan iguales, podemos usar como estadígrafo de contraste entre $\bar{X_m} - \bar{Y_n}$, lo observado, y $\mu_X - \mu_Y$, lo predicho por la H_o:**

$$t = \frac{(\bar{X_m} - \bar{Y_n}) - (\mu_X - \mu_Y)}{s_c \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

La desviación conjunta s_c puede expresarse como

$$s_c = \sqrt{\frac{VAR_X(m-1) + VAR_Y(n-1)}{n+m-2}}$$

Si el valor de la desviación conjunta no da entre las dos desviaciones, hay un error de aritmética.

El intervalo de confianza para $\mu_X - \mu_Y$ se da por

$$(\bar{X_m} - \bar{Y_n}) - t_{(\alpha,2)} s_c \sqrt{\frac{1}{m} + \frac{1}{n}} < \mu_X - \mu_Y < (\bar{X_m} - \bar{Y_n}) + t_{(\alpha,2)} s_c \sqrt{\frac{1}{m} + \frac{1}{n}}$$

Los grados de libertad son n+m-2.

246 ***Ejemplo*** *Un abogado obtuvo durante los 10 meses de trabajo del año pasado una ganancia promedio de $\bar{X}_{10} = 8,2$ con una varianza $s_X^2 = 4$. El abogado reconoce que sufre mucho pues es poco el cliente que no es muy conflictivo y, por eso, un día de enero de este año le dio por cambiar de estrategia y decidió llevarle la corriente a sus clientes. Claro que para él, esto es algo humillante y lo seguirá haciendo siempre y cuando su ganancia promedio aumente al menos 0.4. Después de 12 meses de trabajo sus ganancias tienen una media de*

$\bar{Y}_{12} = 9$ *y una varianza*

$s_Y^2 = 3$

Evidentemente, el promedio de los datos recolectados aumentó en más de 0.4. Pero ¿Tiene este abogado alguna base para decir que con clientes conflictivos es mejor llevarles la corriente y que su doctrina puede aplicarse a todo el mundo? Use una significancia igual a 0.02.

Análisis: Haciendo un gran esfuerzo por entender la justa pregunta, podemos imaginar que ese abogado y sus ganancias son una muestra aleatoria de la forma como les va a los abogados en general. Con esa concesión, podemos investigar si el cambio en el promedio de ganancias es estadísticamente significativo.

Primero hacemos un test para averiguar si las varianzas son las mismas:

Nuestra H_o dice que las varianzas muestrales son iguales por puro azar más no por alguna razón sistemática.

$R = 1 = \frac{VAR_G}{VAR_P}$

Nuestro medidor relacional de discrepancia es:

$F = \frac{VAR_G}{VAR_P} = \frac{4}{3} = 1,33$

La F crítica con $\alpha = 0,02; n_1 = 9; n_2 = 11$ es $4,63$.

Veredicto en cuanto a las varianzas: se acepta H_o: $\sigma_1^2 = \sigma_2^2$, las varianzas son iguales.

Para estudiar las medias, formulamos la H_o de que la diferencia teórica de medias es 0.4, o sea que la nueva estrategia afecta las ganancias lo mínimo necesario para que valga la pena la humillación que significa hacerse el sumiso.

$H_o : \mu_1 = \mu_2 + 0{,}4$ o bien $\mu_1 - \mu_2 = 0{,}4$

Procedemos. La desviación conjunta s_c puede expresarse como

$$s_c = \sqrt{\frac{VAR_X(m-1)+VAR_Y(n-1)}{n+m-2}} = \sqrt{\frac{3(11)+4(9)}{20}}$$

da 1.857. La t_{obs} viene de

$$t = \frac{(\bar{X_m}-\bar{Y_n})-(\mu_X-\mu_Y)}{s_c\sqrt{\frac{1}{m}+\frac{1}{n}}} = \frac{(9-8{,}2)-(0{,}4)}{1{,}857\sqrt{\frac{1}{12}+\frac{1}{10}}}$$

y da 0.503. Como este valor no tiene nada de extraño, pues todas las t son mayores que uno, podemos dar nuestro

Veredicto final: H_o se acepta con $\alpha = 0,02$: llevarle la corriente a los clientes aumenta las ganancias de un abogado típico lo necesario para soliviar la humillación intrínseca.

247 ***Ejemplo*** *Para saber si la Bienestarina (un complemento alimenticio para niños desarrollado y distribuido gratuitamente por el Estado) tiene algún efecto diferencial sobre la altura promedio de los niños de 5 años, se midió la estatura de una muestra aleatoria de niños que consumían regularmente Bienestarina y también se hizo lo mismo con una segunda muestra aleatoria que no lo consumía (grupo desplazado recién llegado a la ciudad). Ambos grupos venían de zonas socioeconómicamente muy similares. Los datos fueron:*

$\bar{X}_{150} = 62,7$
$\bar{Y}_{120} = 61,8$
$s_x = 2,62$, $s_x^2 = VAR_X = 6,864$,
$s_y = 2{,}50$, $s_y^2 = VAR_Y = 6,250$
Decidamos con $\alpha = 0,02$ *si las medias son iguales o diferentes.*

Solución: Averigüemos primero si se trata de varianzas iguales o distintas. Nuestro estadígrafo relacional de discrepancia es
$F_{obs} = 6{,}864/6{,}250 = 1{,}098$
La $F_c = 1{,}60$, por lo cual la F_{obs} claramente es pequeña: las varianzas son iguales.

Pasemos ahora al cálculo de la discrepancia entre las medias:
$s_c = \sqrt{\frac{6{,}25(119)+6{,}86(149)}{120+150-2}} = 2,56$
$t_{obs} = \frac{62{,}7-61{,}8}{2{,}56\sqrt{\frac{1}{120}+\frac{1}{150}}} = 2,863$

Con g.l $= 120 + 150 - 2 = 270 - 2 = 268$ la t es la misma z. Con dos colas y para $\alpha = 0{,}02$ la z crítica es ± 2.3 (pág. 158.

Veredicto: se rechaza H_o con dos colas, pues $t_{obs} > t_c$. La diferencia entre el promedio de las estaturas de los dos grupos es significativa. El Estado está cumpliendo con su deber.

248 ***Ejemplo*** *Se estudia el promedio de notas de estudiantes en 2 secciones de la clase de estadística.*

$\bar{X}_{20} = 3,5$
$n_X = 20$
$VAR_X = 4$
$\bar{Y}_{15} = 4$
$n_Y = 15$
$VAR_Y = 3,2$

Solución: acá no hay estadística probabilística, acá tenemos un censo y en un censo, lo que pasó, pasó y no hay nada que hacer. Evidentemente, el promedio de las notas de las dos secciones es diferente.

Consideremos la siguiente variante: el registro de datos se hizo porque los estudiantes de la sección con promedio más bajo protestaron pues alegan que están viendo un material que ellos consideran que es más difícil que el que ven en la otra sección. Ellos no piden que cambien el material sino que se les haga justicia. En atención a su protesta, tratemos de entenderlos: ellos creen que son una muestra aleatoria de la población estudiantil que eventualmente pueda tomar ese curso. Y que en realidad, el material es más difícil no sólo para ellos sino para la población potencial en general. La prueba de ello es que las notas en dicha sección son más bajas que la otra. Eso es lo que ellos parecen alegar.

Ahora sí estamos en un problema de inferencia estadística. Para comparar las medias, necesitamos saber primero la relación entre las varianzas.

$H_o : \sigma_1^2 = \sigma^2$
$F = \frac{4}{3,2} = 1,250$

La F crítica para $\alpha = 0{,}05$, con dos colas y con $gln = 19$ y $gld = 14$ es ± 2.84

Veredicto: H_o se acepta, las varianzas son iguales.

Pasemos a la consideración de las medias:
$H_o : \mu_1 = \mu_2$, es decir, $\mu_1 - \mu_2 = 0$
$H_A : \mu_1 \neq \mu_2$ es decir, $\mu_1 - \mu_2 \neq 0$
$s_c = \sqrt{\frac{4(19)+3,2(14)}{20+15-2}} = 1,913$
$t_{obs} = \frac{(\bar{X_m}-\bar{Y_n})-(\mu_X-\mu_Y)}{s_c\sqrt{\frac{1}{m}+\frac{1}{n}}} =$
$t_{obs} = \frac{(4-3,5)-0}{1,91\sqrt{\frac{1}{20}+\frac{1}{15}}} = 0{,}765$

Con $g.l = 20 + 15 - 2 = 33$, la t-crítica es 2,0345

Veredicto: H_o se acepta. Está permitido creer que el desbalance que pasó con estas dos secciones no se volverá a repetir en el futuro.

Variante: el experto opina. Como el problema se conocío fuera de clase, apareció un experto que dijo: "los datos originales permiten argumentar que la media poblacional grande es 0.6 mayor que la pequeña." Veámos si es cierto. Tenemos entonces la $H_o : \mu_1 - \mu_2 = 0{,}6$. Rechazamos con dos colas.

Las varianzas son iguales, $s_c = 1{,}913$

$$t_{obs} = \frac{(\bar{X_m}-\bar{Y_n})-(\mu_X-\mu_Y)}{s_c\sqrt{\frac{1}{m}+\frac{1}{n}}} = \frac{(4-3,5)-(0,6)}{1,913\sqrt{\frac{1}{20}+\frac{1}{15}}} = \frac{-0,1}{0,659} = 0{,}15$$

Esta t_{obs} es obviamente pequeña: la discrepancia entre lo que se ve y lo que se cree en la H_o es insignificante, por tanto aceptamos lo que se cree: es permitido creer que la media poblacional grande es 0.6 mayor que la pequeña.

Ahora tenemos un misterio: en la primera variante aceptamos que las medias son iguales y en la tercera que hay razones para creer que la media poblacional grande es 0.6 mayor que la pequeña. No hay ningún problema si sólo nos interesa saber qué es lo que está permitido creer. Pero si hay que tomar decisiones, ésto suena a una contradicción. Para ver cómo se resuleve, calculemos el IC de la diferencia de medias:

$$(\bar{X_m} - \bar{Y_n}) - t_{(\alpha,2)} s_c \sqrt{\tfrac{1}{m} + \tfrac{1}{n}} < \mu_X - \mu_Y < (\bar{X_m} - \bar{Y_n}) + t_{(\alpha,2)} s_c \sqrt{\tfrac{1}{m} + \tfrac{1}{n}}$$

donde la $t_{(\alpha,2)}$ es la t crítica con dos colas, 33 gli y significancia $\alpha = 0{,}05$:

$(4 - 3{,}5) - 2{,}02(1{,}913)\sqrt{\frac{1}{15} + \frac{1}{20}} < \mu_X - \mu_Y < (4 - 3{,}5) + 2{,}02(1{,}913)\sqrt{\frac{1}{15} + \frac{1}{20}}$

El IC de la diferencia de medias es por tanto: (-3.3719, 4.3719). Este intervalo contiene todas las diferencia de medias poblacionales que se justifican por los datos. Vemos que el cero está ahí, lo mismo que el 0.6 y lo mismo que muchos otros valores. También está el valor negativo 3.3719, el cual indica que en futuras ocasiones el curso al cual hoy le fue mal puede tener sistemáticamente notas mucho mejores que otros cursos. Ahora bien, ¿cuál es la más sensata de todas las posibles creencias? Dividimos el IC de la diferencia de medias en dos regiones, la que apoya al profesor (que el examen estuvo bien) y la que apoya a los muchachos (que el examen estuvo muy difícil). Consideramos que el profesor es apoyado por la región (-3.3719, 0) y que los estudiantes son apoyados por el intervalo (0, 4.3719). Los estudiantes tienen más apoyo: solicitamos que el profesor ceda.

Variante: varianzas más pequeñas: Para que el alegato de los muchachos tuviese una validez indiscutible sería suficiente que las varianzas disminuyeran. Tal sería el caso si las varianzas muestrales fuesen 0,2 y 0,3. Incluso, como los grados de libertad son mayores que 30 podemos calcular la discrepancia entre lo esperado según la H_o (que las medias poblacionales son iguales) y lo observado, por una z. Estaríamos diciendo que podemos tomar las varianzas muestrales como si fuesen poblacionales:

$z = \frac{(\bar{X_m} - \bar{Y_n}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{3{,}5 - 4}{\sqrt{\frac{0{,}2}{20} + \frac{0{,}3}{15}}} = -2{,}887$

la cual es un valor grande dado $\alpha = 0{,}05$.

Veredicto: H_o se rechaza. En esta variante los muchachos tienen razón: su propia experiencia es una causa suficiente para predecir que en el futuro su situación volverá a repetirse con otros cursos. Como la grandeza de la Universidad se edifica sobre la justicia y la disciplina, el profesor debe sentirse obligado a hacer algo para que la justicia brille y los muchachos puedan esforzarse con alegría.

Pero, ¡carambas!, ¿qué hacemos si el profesor considera que sus estudiantes deben nacer y crecer con la clara certeza de que ellos son mejores que los europeos, lo cual no es tan difícil?

9. **Comparación de la diferencia entre las medias de dos muestras aleatorias e independientes con la diferencia entre dos medias poblacionales sabiendo, por el estudio de las varianzas de las dos muestras, que las varianzas poblacionales son diferentes. Variables normalmente distribuídas:**

   La explicación de que toda diferencia entre medias es por azar se escribe en la hipótesis nula:

   $H_o : \mu_X - \mu_Y = a$

   Lo observado es $\bar{X_m} - \bar{Y_n}$

   Estadígrafo de contraste entre lo observado $\bar{X_m} - \bar{Y_n}$ y lo esperado según la H_o de diferencia de medias $\mu_X - \mu_Y$:

$$t = \frac{(\bar{X_m} - \bar{Y_n}) - (\mu_X - \mu_Y)}{\sqrt{\frac{VAR_X}{m} + \frac{VAR_Y}{n}}}$$

El intervalo de confianza para $\mu_X - \mu_Y$ se da por

$$(\bar{X_m} - \bar{Y_n}) - t_{(\alpha,2)}\sqrt{\frac{VAR_X}{m} + \frac{VAR_Y}{n}} < \mu_X - \mu_Y < (\bar{X_m} - \bar{Y_n}) + t_{(\alpha,2)}\sqrt{\frac{VAR_X}{m} + \frac{VAR_Y}{n}}$$

Los grados de libertad son calculados como sigue:

$$g.l = \frac{(\frac{Var_X}{n_X} + \frac{Var_Y}{n_Y})^2}{\frac{(\frac{Var_X}{n_X})^2}{n_X+1} + \frac{(\frac{Var_Y}{n_Y})^2}{n_Y+1}} - 2$$

249 ***Ejemplo*** *Una esposa que ama a su esposo está muy inconforme porque no se siente amada sino esclava. Y es una esclavitud de la que no puede salirse por más que trata. Pero al estudiar textos antropológicos, llega a la conclusión de que en algunas culturas ancestrales la mujer era oficialmente una esclava a la cual el marido tenía que comprar. Y se pregunta: ¿Cuál es la diferencia entre ser esclava oficialmente y ser reina de nombre pero esclava de hecho? Simplemente por burlarse un poco de sí misma y de su vida tan cruel decide hacer un experimento y comienza a tratar a su marido de "Sí señor", "No señor", y le trae su café caliente, y le prepara el baño, y cosas así que hacen sentir al marido como si fuese un Rey. Y precisamente cuando ya se iba cansando de sentirse ridícula su amiga le confirma lo que ella no se había atrevido a reconocer: que el marido ahora permanece mucho, muchísimo más tiempo en casa y juega más con los niños. Anotamos la duración del tiempo de juego con los niños por ocasión, en minutos, tanto antes como después del cambio de actitud de la esposa. Nos interesa saber si la duración promedio de los juegos ha sido triplicada.*

Después: $m = 12$, $\bar{X}_{12} = 34$, $VAR_X = 5$
Antes: $n = 7$, $\bar{Y}_7 = 12$, $VAR_Y = 90$.

Análisis: **Paso 1.** Comenzamos estudiando la varianza:

1. $H_o : \sigma_X^2 = \sigma_Y^2$
$H_A : \sigma_X^2 \neq \sigma_Y^2$

$F = \frac{90}{5} = 18$

La F_+, la F crítica superior con dos colas, significancia 0.10, $gln = 6$, $gld = 11$ es 3.09, y con significanica 0.01 es 5.07.

Veredicto: Como 18 es mucho mayor que 5.07, concluimos que las varianzas son diferentes, es decir, tenemos razones para creer que la diferencia entra las varianzas muestrales se originó no por azar sino por algo sistemático, que bien podría ser el cambio de actitud de la esposa.

Paso 2: Sabiendo que las varianzas son distintas, investiguemos si la media grande es el triple de la pequeña. Para ello vamos a suponer que la media poblacional de Y se puede tomar como su media muestral que fue 12. Lo que se cree es:

$H_o : \mu_X = 3\mu_Y$, es decir, $\mu_X - \mu_Y = 3\mu_Y - \mu_Y = 2\mu_Y = 24$

$H_A : \mu_X - \mu_Y \neq 24$, con dos colas.

Usamos una t como estadígrafo de discrepancia entre lo que se ve y lo que se cree. Lo que se ve es:

Después: $m = 12$, $\bar{X}_{12} = 34$, $VAR_X = 5$

Antes: $n = 7$, $\bar{Y}_7 = 12$, $VAR_Y = 90$.

$t = \frac{(\bar{X_m} - \bar{Y_n}) - (\mu_X - \mu_Y)}{\sqrt{\frac{VAR_X}{m} + \frac{VAR_Y}{n}}}$

$t = \frac{(34-12)-24}{\sqrt{\frac{5}{12} + \frac{90}{7}}} = \frac{-2}{\sqrt{13,27}} = -0,5$

Para sacar los gli utilizamos $Var_1 = 5$, $Var_2 = 90$ (acá el orden no importa):

$$g.l = \frac{(\frac{Var_X}{n_X} + \frac{Var_Y}{n_Y})^2}{\frac{(\frac{Var_X}{n_X})^2}{n_X+1} + \frac{(\frac{Var_Y}{n_Y})^2}{n_Y+1}} - 2$$

$g.l = \frac{(\frac{5}{12} + \frac{90}{7})^2}{\frac{(\frac{5}{12})^2}{12+1} + \frac{(\frac{90}{7})^2}{7+1}} - 2 = 6,02$

La t crítica con $\alpha = 0,01$, dos colas y con 6 gli es $\pm 3,14$. Por tanto,la $t_{obs} = -0,5$ es pequeña, lo cual dice que la discrepancia entre lo que se ve y lo que se cree no da lugar a dudas: aceptamos lo que se cree. El Rey ahora pasa el triple de tiempo jugando con los niños comparado con lo que pasaba antes cuando a duras penas era un simple marido cuyo único objetivo era trabajar y trabajar para mantener a la familia.

250 ***Ejemplo*** *La doctrina de los derechos humanos se basa en que el ser humano es un ser pensante y libre. La historia de esa doctrina se puede trazar hasta Descartes quien puso en el pensamiento la esencia del ser: 'pienso, luego existo', que es una forma muy personal de la filosofía idealista de Hegel quien creía que el ser y el pensamiento son lo mismo. Por lo tanto, al ser humano no se le puede pegar, sino educar; ni tampoco se le puede restringir de modo absoluto. Tal doctrina no es aceptada en todos lados, por ejemplo, los papás saben que una palmada es santo remedio para tanta molestadera de los niños, mucho mejor que la cantaleta de la mamá. Más seriamente y contrario a las religiones cristiana y musulmana, dicha doctrina alega que al hombre no se le puede restringir su derecho a definir cuáles relaciones sexuales son buenas y cuáles son malas. En el occidente moderno, que viene de un trasfondo cristiano que predica el temor a Dios y la hermosura de la santidad como norma de vida, hay ahora una cierta tendencia a definir las relaciones homosexuales como normales y permitidas, y la tendencia se acentúa. No así en los países de dominio musulmán, donde puede pasar que las leyes en contra del homosexualismo se endurezcan.*

El promedio del número de leyes por año en un cierto país de trasfondo cristiano P que tienen como objetivo mejorar las condiciones de los grupos de homosexuales, que son minoritarios, durante el decenio de los 90 fue de 1 y varianza 0.

Pero sucede que junto a esta pelea de simples ideologías hay respetables flujos económicos. Hay, por ejemplo, el comercio de niños de países en desarrollo, o los sofisticados laboratorios de biología transexual de los países desarrollados. Y las dos cosas se combinan. Por esa razón, un país en desarrollo que promueve el homosexualismo no hace más que humillar a su pueblo ante las potencias. Pero por otro lado, los que fabrican leyes quieren modernizarlas, lo cual significa ponerle el toque europeo. En los últimos años todo eso se ha ventilado bastante y, como consecuencia, en el país ya referido, durante los 7 primeros años del 2000 ha habido un promedio de 3 leyes por año con una varianza de 1.

¿Podemos decir que ha habido un cambio en el promedio de leyes al comparar el decenio del los 90 con los primeros años del 2000 para países como el P?

Solución: La telenovela narrada nos cuenta un hecho. La estadística nada puede quitarle, nada puede añadirle, y por supuesto que ha habido un incremento de leyes que tratan de mejorar la situación de los homosexuales. Pero si interpretamos la pregunta diciendo que el país mencionado

puede tomarse como una muestra al azar de los países de su clase, entonces ya podemos usar la estadística para decidir si el promedio de leyes ha cambiado en los países como P.

Comparamos las varianzas: La H_o que no puede ser otra que la igualdad de varianzas.

La F da infinito, pues se divide por cero, la varianza menor. Infinito es extremo para cualquier significancia. Concluimos que las varianzas de los dos períodos son diferentes.

Comparemos las medias, decidiendo si la diferencia de media poblacionales es cero:

$$t = \frac{(\bar{X_m} - \bar{Y_n}) - (\mu_X - \mu_Y)}{\sqrt{\frac{VAR_X}{m} + \frac{VAR_Y}{n}}}$$

$t = \frac{3-1}{\sqrt{\frac{1}{7} + \frac{0}{10}}} = 5{,}4$

Los gli están dados por

$$g.l = \frac{(\frac{Var_X}{n_X} + \frac{Var_Y}{n_Y})^2}{\frac{(\frac{Var_X}{n_X})^2}{n_X+1} + \frac{(\frac{Var_Y}{n_Y})^2}{n_Y+1}} - 2$$

$$g.l = \frac{(\frac{1}{7} + \frac{0}{10})^2}{\frac{(\frac{1}{7})^2}{7+1} + \frac{(\frac{0}{10})^2}{10+1}} - 2 = 6$$

La t crítica con $\alpha = 0{,}01$, dos colas y con 6 gli es 3,707.

Veredicto: H_o se rechaza. Discrepancia grande: podemos pensar que en los primeros años del 2000 se aprobaron más leyes a favor de los homosexuales que en el período de los 90 en los países de la clase del país de que trata la telenovela.

10. **Comparación de la diferencia de medias observadas con la diferencia de medias esperadas para datos apareados de variables independientes y normales (miopía ojo izquierdo vs. miopía ojo derecho).**

Los relojeros se ponen el monocular sobre el mismo ojo todas las veces. Nos interesa saber si el monocular causa miopía. Para decidirlo, registraremos los datos de miopía tanto del ojo de trabajo como del otro. **Pero de la misma persona,** y no el ojo de trabajo de una persona con el ojo de descanso del de otra. Tenemos datos apareados, o sea que vienen y se comparan por pares. Una hipótesis nula podría ser

$H_o : \bar{\Delta} = promedio\ poblacional\ esperado\ de\ las\ diferencias\ apareadas = 0.$

Ésto diría que simplemente nos interesa verificar si los dos ojos tienen igual grado de miopía (usamos $\bar{\Delta}$ para enfatizar que se trata de datos apareados). Otra H_o podría ser también

$H_o : \bar{\Delta} = 2$

lo cual quiere decir que definitivamente estamos convencidos de que el ojo de trabajo es mucho más miope y necesita lentes con más dioptrias que el ojo de descano.

Los datos vienen en n pares (x_i, y_i). Cada par de datos da origen a una diferencia $D_i = x_i - y_i$, la cual se reporta con su signo. Los D_i tienen una media $\bar{D}$, la diferencia observada de datos apareados, y una desviación s:

$\bar{D} = \frac{x_i - y_i}{n}$. (Uno puede cuadrar las cosas para que $\bar{D}$ siempre sea positiva.)

Estadígrafo de contraste para medir la discrepancia entre lo observado (dado por $\bar{D}$, el promedio observado de las diferencia apareadas), y lo esperado según la $H_o : \bar{\Delta} = promedio\ esperado\ de\ las\ diferencias\ apareadas = algo$

$$t = \frac{\bar{D} - (\mu_X - \mu_Y)}{\frac{s}{\sqrt{n}}} = \frac{\bar{D} - \bar{\Delta}}{\frac{s}{\sqrt{n}}}$$

Los gli de la t son $n-1$.

El intervalo de confianza para la diferencia apareada de medias es:

$$\bar{D} - t_{(\alpha,2)}\frac{s}{\sqrt{n}} < \bar{\Delta} = \mu_X - \mu_Y < \bar{D} + t_{(\alpha,2)}\frac{s}{\sqrt{n}}$$

251 ***Ejemplo*** *Deseamos saber cómo se relacionan las edades del varón y de la mujer en parejas de unión libre. En la tabla adjunta, X representa la edad del hombre en tanto que Y representa la edad de la mujer DE LA MISMA PAREJA. a) Puede decirse que la pareja tienen la misma edad? b) Puede decirse que la edad del hombre duplica a la de la mujer?*

Edades del hombre y de la mujer en parejas de unión libre

X	Y	$D_i = X - Y$	D_i^2
20	*19*	*1*	*1*
23	*17*	*6*	*36*
30	*20*	*10*	*100*
27	*23*	*4*	*16*
25	*15*	*10*	*100*
21	*19*	*2*	*4*
$\sum = 146$	$\sum = 113$	$\sum D = 33$	$\sum D^2 = 257$

Solución: a) Veamos si las edades son las mismas:

$H_o : \mu_1 = \mu_2$, con dos colas. Como son medias de datos apareados se usa una notación enfática:

$\bar{\Delta} = \mu_1 - \mu_2 = 0$ (no hay diferencia promedio de edades apareadas)

$\bar{D} = 33/6 = 5{,}5$ (promedio observado de diferencias de datos apareados)

$\alpha = 0{,}05$, gli= n-1 = 5, t-crítica: $t_5 = \pm 2{,}5$.

La fórmula para la desviación que aprendimos en el capítulo 1 era

$s = \sqrt{\frac{n(\Sigma x^2)-(\Sigma x)^2}{n(n-1)}}$

que para nuestro caso se convierte en

$s = \sqrt{\frac{n(\Sigma D^2)-(\Sigma D)^2}{n(n-1)}} = \sqrt{\frac{6(257)-(33)^2}{30}} = 3{,}88$

$t_{obs} = \frac{\bar{D}-\bar{\Delta}}{\frac{s}{\sqrt{n}}} = \frac{5{,}5-0}{3{,}88/\sqrt{6}} = \frac{5{,}5}{1{,}58} = 3{,}4$

Veredicto: H_o se rechaza: la edad promedio del hombre no es igual a la edad promedio de la mujer. Al ver la tabla, podemos concluir que la edad del hombre es mayor que la de la mujer.

b) Veamos si se puede decir que la edad del varón es el doble que la de la mujer.

Edad promedio del hombre = 146/6 = 24.3

Edad promedio de la mujer = 113/ 6 = 18.8

$H_o : \mu_1 = 2\mu_2$ (con dos colas)

$\bar{\Delta} = \mu_1 - \mu_2 = 2\mu_2 - \mu_2 = \mu_2 = 18{,}8$ (el hombre dobla a la mujer)

$\bar{D} = 33/6 = 5{,}5$

$\alpha = 0{,}05$, gli= 5, t-crítica: $t_5 = \pm 2{,}5$.

$s = \sqrt{\frac{n(\Sigma D^2)-(\Sigma D)^2}{n(n-1)}} = \sqrt{\frac{6(257)-(33)^2}{30}} = 3{,}88$

$t_{obs} = \frac{\bar{D}-(\mu_X-\mu_Y)}{\frac{s}{\sqrt{n}}} = \frac{5{,}5-18{,}8}{3{,}88/\sqrt{6}} = \frac{-13{,}47}{1{,}58} = -8{,}52$

Como la t_{obs} no está en el intervalo (-2.5, 2.5), rechazamos la H_o: no es aconsejable creer que el hombre doble a la mujer en edad.

Nota sobre la solución rigurosa: debido a que en este problema la H_o fue calculada a partir de los datos experimentales, se pierde uno de los gli de los datos. En vez de ser 5 gli son 4. La t_c vale 2.7 y la H_o se rechaza de todas maneras en ambas interpretaciones del problema.

Capítulo 12

Árboles y tablas de contingencia

252 ***Objetivo*** *Estudiamos probabilidades condicionales y la dependencia o independencia de dos factores.*

12.1. Árboles y probabilidades condicionales

En el capítulo 2 vimos la binomial y sus árboles de eventos cuando tirábamos varias monedas para ver si salía cara o sello. Sucede que las monedas no cambian cuando se tiran y, al volver a tirarlas, las probabilidades de cara o sello son las mismas que antes. En una frase, en la binomial las probabilidades no se gastan. Ahora veamos un ejemplo, en el cual las probabilidades cambian y, por lo tanto, no puede resolverse por binomial.

253 ***Ejemplo*** *Una baraja consta de 52 cartas, divididas en cuatro palos que son corazones, diamantes, tréboles y picas. Los corazones y diamantes son rojos, los tréboles y picas son negros. Cada uno de los palos consta de 13 cartas: As, 2, . . ., 10, J, Q y K. Se eligen al azar dos cartas* ***sin reemplazo = se saca la primera carta, se mira y se registra pero no se regresa a la mesa, y después se saca la segunda****. Hallemos la probabilidad de que las 2 cartas tengan alguna de las letras J,Q,K.*

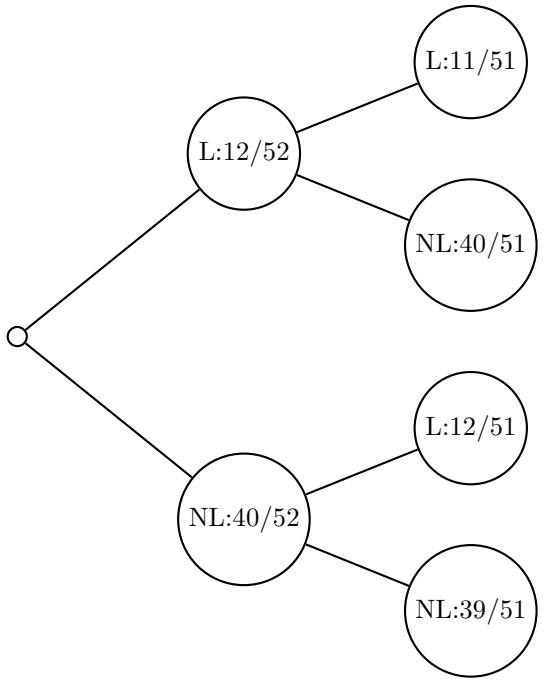


Figura 12.0. Sacamos dos cartas sin reemplazo. Puede salir L(letra) on NL (no letra).

Solución: Como por cada palo hay 3 letras, por todo hay 12. La probabilidad de sacar una letra es entonces 12/52. Pero si ya se sacó una carta sin reemplazo, y salió letra, puede salir letra con probabilidad 11/51, pues hay una letra menos y una carta menos. Pero si no salió letra en la primera carta, puede salir letra en la segunda con probabilidad 12/51 y no letra con probabilidad 39/51.

¿Cómo cambia el árbol si las cartas se sacaran **con reemplazo**, es decir cuando la carta sacada se regresa a la baraja, la cual vuelve a barajarse? Estaríamos en el caso de la binomial: probabilidad de letra 12/52 y de no letra 40/52 para cualquier carta.

Cuando las probabilidades cambian, lo hacen en dependencia de lo que pasó, por ejemplo, y se denominan **probabilidades condicionales**. Las probabilidades condicionales aparecen en la vida real cuando hay eventos encadenados, cada uno de los cuales puede tener varias opciones.

254 ***Ejemplo*** *Tania a veces está triste. La probabilidad de que esté triste es 0.2. Tania a veces hace la tarea y a veces no y todo depende de que esté o no triste. Si está triste, hace la tarea con probabilidad 0.1. Si no lo está, hace la tarea con probabilidad 0.7. Tania hoy hizo la tarea. ¿Respalda Ud la tesis de que Tania estuvo triste ayer por la tarde?*

Solución: la forma de actuar de Tania se describe por un árbol como sigue:

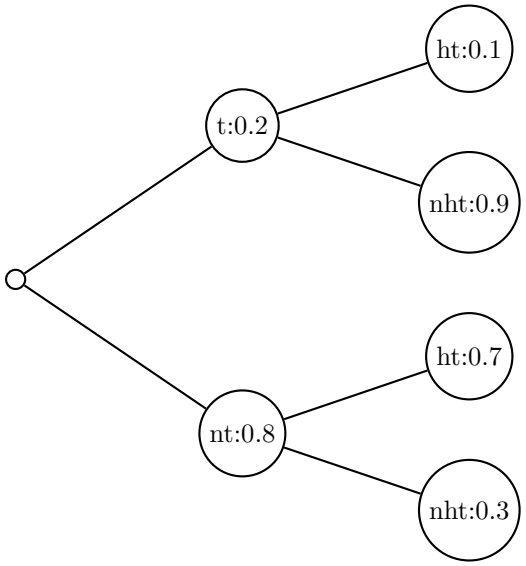


Figura 12.1. Forma de actuar de Tania: t (triste), nt (no triste); ht(hizo tarea), nht (no hizo la tarea).

Calculemos la probabalidad de que ayer estuviese triste sabiendo que vino con la tarea hecha. El universo de la probabilidad, lo que vale el 100 % , es que vino con la tarea hecha. Eso equivale a dos opciones, la primera vale (0.2)(0.1) que es la probabilidad de hacer la tarea dado que ella estuvo triste. La segunda opción es (0.8)(0.7), que corresponde a haber hecho la tarea dado que ella no estuvo triste. Como estar triste y no estarlo son eventos totalmente aparte que no tienen nada en común, la probabilidad de venir con la tarea es una suma sobre las dos opciones: (0.2)(0.1) + (0.8)(0.7). Ahora bien, de nuestro universo, haber hecho la tarea, tomamos la rama que corresponde a haber estado triste: (0.2)(0.1) . Por lo tanto, la probabilidad p de haber estado triste dado que hizo la tarea es:

$$p = \frac{(0{,}2)(0{,}1)}{(0{,}2)(0{,}1)+(0{,}8)(0{,}7)} = \frac{0{,}02}{0{,}58} = 0{,}034$$

Esta probabilidad es muy pequeña. La gran mayoría de personas se atrevería a asegurar que ella no estuvo triste cuando se enfrentó con la decisión de hacer la tarea. De una gran muestra de datos, al responder uno que Tania no estuvo triste, aproximadamente uno se equivocaría un 3 % de las veces.

Oficialmente, a la manera como hemos calculado la probabilidad condicional de que Tania haya estado triste dado que hoy ella hizo la tarea se llama **Regla de Bayes**.

255 ***Ejemplo*** *En un grupo hay 4 mujeres y 2 hombres. En otro hay 3 mujeres y 5 hombres.*

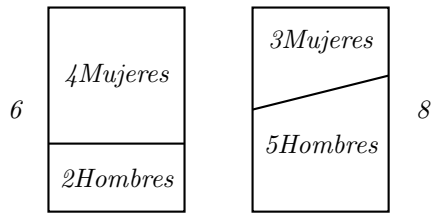


Figura 12.2. Los dos grupos al comienzo del experimento.

Una persona al azar sale del primer grupo y se pasa al segundo. Luego se elige al azar una persona del segundo grupo.

a. Hagamos un árbol que represente el experimento asignándoles probabilidades a las ramas.

b. Calculemos la probabilidad de que la persona elegida del segundo grupo sea mujer.

Solución: Este problema es una variante del problema de tirar una moneda dos veces y contar el número de caras. Para resolverlo, usemos la siguiente notación para los eventos:

- m= la persona que sale del primer grupo es mujer.
- h = la persona que sale del primer grupo es hombre.
- M= la persona que sale del segundo grupo es mujer.
- H = la persona que sale del segundo grupo es hombre.

La probabilidad de ser mujer de la primera persona del primer grupo es de 4/6 y de ser hombre 2/6, pues de 6 personas 4 son mujeres y 2 son hombres. Eso lo dibujamos así:

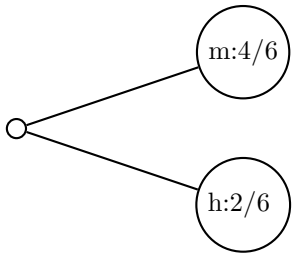


Figura 12.3. La primera persona puede ser m ó h con probabilidades 4/6 y 2/6 respectivamente.

Si la primera persona, salida del primer grupo, salió mujer, la situación inmediatamente antes de elegir la segunda persona es así:

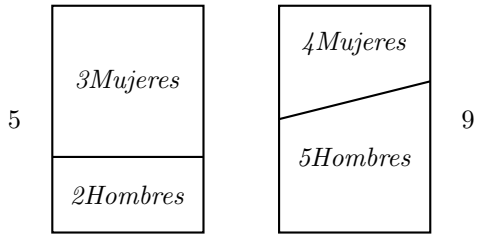


Figura 12.4. Si la primera persona sacada del primer grupo fue mujer, ésto es lo que queda.

En este caso, en el segundo grupo quedan 4 mujeres y 5 hombres. Por tanto, la probabilidad de salir mujer del segundo grupo sería 4/9 y la de hombre de 5/9.

Pero si la primera persona sacada del primer grupo fue hombre, quedan en el segundo grupo 3 mujeres y 6 hombres.

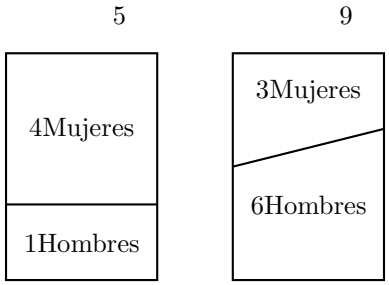


Figura 12.5. Si la primera persona sacada del primer grupo fue hombre, ésto es lo que queda.

Por tanto, la probabilidad de mujer es 3/9 y de hombre 6/9. Todo lo resumimos en el siguiente árbol:

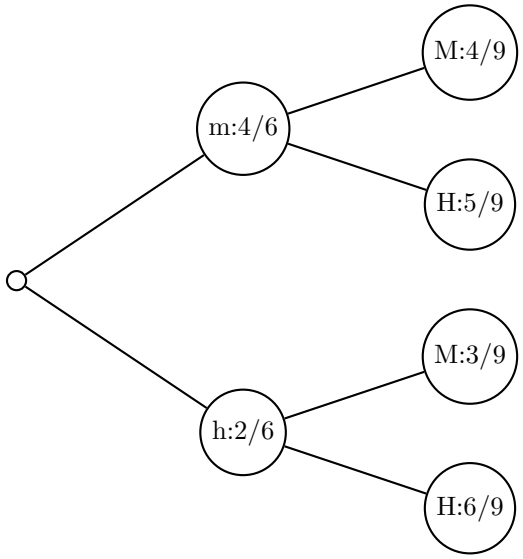


Figura 12.6. Las probabilidades de la segunda persona dependen de lo que pasó con la primera.

Ahora bien: las probabilidades a lo largo de un camino en el árbol se multiplican (más abajo explicamos por qué). Por lo tanto, la probabilidad de que la segunda persona sea mujer en la primera opción es: (4/6) (4/9). La probabilidad de que la segunda persona sea mujer en la segunda opción es (2/6)(3/9).

¿Cuál es la probabilidad de que la segunda persona sea mujer?

La segunda persona es mujer, sea porque sale de la primera opción o sea porque sale de la segunda. Como las dos opciones no tienen nada en común, las probabilidades se suman:

(4/6) (4/9) + (2/6)(3/9) = (16+6)/54 = 22/54.

Y esa es la respuesta.

256 ***Ley de la multiplicación***

Nos falta explicar por qué las probabilidades a lo largo de un camino se multiplican, como si se tratase de eventos independientes. La explicación intuitiva es que una vez que la primera persona salió del primer grupo, el género de la segunda persona depende únicamente de la proporción de mujeres que haya en el segundo grupo y realmente no importa de donde salieron las personas.

Veamos ahora una explicación rigurosa de por qué se multiplican las probabilidades a lo largo de una rama en este árbol. Tomemos las rama superior mM que para las otras ramas se procede igual.

Probabilidad de que la primera persona sea mujer, P(m) = 4/6.

Probabilidad de que la segunda persona sea mujer dado que la primera persona fue mujer = p(M/m) donde P(M/m) se lee: probabilidad de M dado que ocurrió m. Esta es la probabilidad condicional.

La probabilidad condicional se define como:

P(M/m) = p(mM)/p(m)

donde

P(mM) es la probabilidad de que la primera persona sea mujer y que la segunda sea mujer.

P(m) = probabilidad de sacar mujer del primer grupo = 4/6.

P(M/m) = la probabilidad de salir mujer en el segundo grupo dado que la primera persona fue mujer = 4/9.

Por tanto,

p(mM) = p(m) p(M/m).

p(mM) = p(m) p(M/m) = (4/6)(4/9)= 16/54.

Vemos que para llegar a mM hay que multiplicar la probabilidad de llegar a m por la de llegar a M dado que antes se había llegado a m. En palabras más claras, **la probabilidades a lo largo de un camino se multiplican, teniendo cuidado de que las probabilidades condicionales sean las naturales del problema.** Y así con todas las ramas de este árbol y con todos los árboles.

12.2. Tablas de contingencia: Independencia de dos factores

Cuando uno estudia la relación entre dos factores, es usual acomodar los datos en una **tabla de contingencia**, como la del siguiente ejemplo en el cual se contrasta la relación entre los **factores** género (con **eventos** hombre, mujer) y tamaño de los ojos (con **eventos** ojos grandes, ojos pequeños):

257 ***Consideremos la siguiente tabla de contingencia:***

Género	Tamaño de los ojos		
	Tiene ojos grandes	Tiene ojos pequeños	Totales
Hombre	20	50	70
Mujer	30	10	40
Totales	50	60	110

Los hombres en su mayoría tienen ojos pequeños, mientras que las mujeres en general tienen ojos grandes. Eso nos permite detectar el género de un ser humano simplemente mirándole los ojos. Este método no es perfecto, pero es bueno. Eso es lo que aprendemos por 'ojímetro' al mirar la tabla. La formalización estadística es algo sofisticada y es un formalismo que hay que aprender:

Tomamos como H_o la proposición de que los dos factores estudiados, tamaño de ojos vs. género, no tienen nada que ver el uno con el otro: decimos que son independientes. Más luego, mediremos la discrepancia entre lo que se observa, los datos, y lo que se cree según la H_o. Si esa discrepancia es pequeña, decimos que los factores son independientes. Pero si es grande, rechazamos la H_o y decimos que los factores son dependientes.

Supongamos que hemos demostrado que la H_o se rechaza. Los factores son dependientes, pero no sabemos la forma de la correlación: ¿cuáles son los hombres, los de ojos grandes o los de ojos pequeños? Para eso miramos los datos, y descubrimos que los hombres son los de ojos pequeños.

Veamos ahora el detalle de la formalización.

Para empezar, aprendamos a estimar las probabilidades a partir de los datos de la tabla.
Nosotros vemos que hay 20 hombres con ojos grandes. Otros 50 hombres no tienen los ojos grandes. Hay 40 mujeres, 30 de las cuales tienen ojos grandes y el resto tienen los ojos pequeños. De estos números nosotros formamos las proporciones: 20/70 es la proporción de hombres con los ojos grandes; 50/70 es la proporción de hombres con los ojos pequeños; 30/40 es la proporción de mujeres con los ojos grandes mientras 10/40 es la proporción de mujeres con los ojos pequeños. La proporción de mujer en la muestra es 40/110 mientras los de hombres es 70/110.

La proporción es una estimación de una probabilidad y siempre asumimos que la estimación se basa en un procedimiento de muestreo aleatorio. Si nosotros repetimos el procedimiento de la muestra, las proporciones observadas cambiarán. Pero cuando el tamaño de la muestra tiende al infinito, la proporción tiende a la probabilidad. Imaginemos que nuestra muestra es lo suficientemente grande como para permitirnos omitir la diferencia entre estimaciones y las probabilidades reales.
Usemos la siguiente nomenclatura para los atributos dados:

g= tener los ojos grandes.
p= tener los ojos pequeños.
h= ser un hombre.
m= ser una mujer.

Para el género: $p(h) = 70/110$; $p(m) = 40/110$ y $p(m) + p(h) = 1$
Para el tamaño de ojos: $p(g) = 50/110$; $p(p) = 60/110$ y $p(p) + p(g) = 1$

Para las observaciones registradas dentro de la tabla:
$p(h\ y\ g) = 20/110 = 2/11$.
$p(h\ y\ p) = 50/110 = 5/11$.
$p(m\ y\ g) = 30/110 = 3/11$.
$p(m\ y\ p) = 10/110 = 1/11$.
$p(h\ y\ g) + p(h\ y\ p) + p(m\ y\ g) + p(m\ y\ p) = 1$.

Hay otras proporciones que corresponden a las probabilidades condicionales. Entre las 50 personas con ojos grandes hay 20 que son varones. La proporción $20/50 = 2/5$ se denota $p(h/g)$ y se lee **la probabilidad condicional de ser hombre dado que se tienen los ojos grandes**. Concretamente:
$p(g/m)$ se lee: la probabilidad de tener los ojos grandes dado que se es mujer.
$p(g/h)$ se lee: la probabilidad de los ojos grandes ser dado que se es hombre.
$p(h/p)$ se lee: la probabilidad de ser hombre dado que se tienen los ojos pequeños.

El valor de estas probabilidades y de otras complementarias son:
$p(g/h) = 2/7$ y $p(p/h) = 5/7$ y los dos suman uno.
$p(g/m) = 3/4$ y $p(p/m) = 1/4$ que suman a uno.
$p(h/g) = 2/5$ y $p(m/g) = 3/5$ y los dos suman uno.
$p(h/p) = 5/6$ y $p(m/p) = 1/6$ que suman uno.

Hemos trabajado hasta aquí con frases como *tener ojos grandes* o *ser hombre*. A esas frases se les llama **proposiciones** pues pueden ser ciertas para algunas personas pero falsas para otras. A cada proposición se le asocia su **conjunto de verdad** que es el conjunto que contiene todos los elementos del universo que cumplen o hacen verdadera a la proposición. Así, el conjunto de verdad de ser hombre contiene a todos los hombres considerados en el conjunto universal y el conjunto de verdad de tener ojos grandes es aquel que contiene a todas las personas que tienen ojos grandes. En lo siguiente mezclaremos proposiciones y conjuntos de verdad pues uno los puede intercambiar. A los conjuntos de verdad se les nota con letras mayúsculas $A, B, ..., Z$ y a las proposiciones correspondientes con letras minúsculas $a, b, ..., z$.

La probabilidad de A condicionada a B es la división de la probabilidad de la intersección de A y B dividida por la probabilidad de B:

$p(A/B) = \frac{p(A \cap B)}{p(B)}$

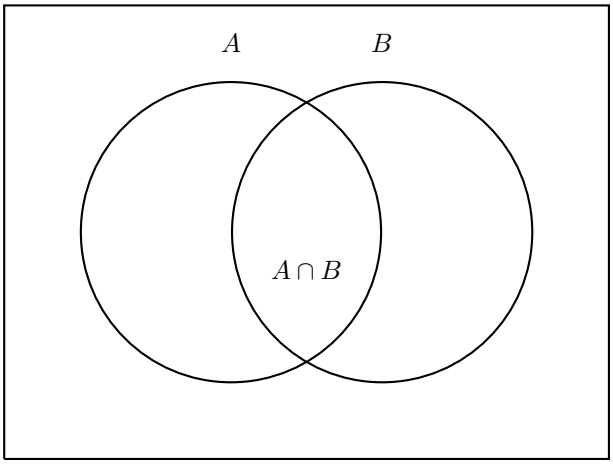


Figura 12.7. $p(A/B) = \frac{p(A \cap B)}{p(B)}$ *es la probabilidad de que A y B sucedan, tomando a B como universo.*

258 ***Ejemplo*** $p(h/g) = \frac{2}{5}$ *mientras que* $p(h \text{ y } g) = \frac{2}{11}$ *y* $p(g) = \frac{5}{11}$.
Con eso tenemos que $\frac{p(h \text{ y } g)}{p(g)} = \frac{\frac{2}{11}}{\frac{5}{11}} = \frac{2}{5}$.

259 ***Ejemplo*** *Calculemos* $p(h \text{ o } p)$*, la probabilidad de que una persona sea hombre o tenga los ojos pequeños. En el idioma Español, la o sin tilde es distinta de la ó tildada. La o sin tilde significa lo uno o lo otro o ambos. La ó tildada significa lo uno ó lo otro pero no ambos. Lo que nos preguntan se refiere a la o sin tilde, es decir a la unión de dos subconjuntos, lo uno o lo otro o ambos.*

Solución: $p(h \text{ o } p)$ corresponde a la probabilidad de la unión que es igual a la probabilidad del uno más la probabilidad del otro menos la probabilidad de la intersección. La fórmula general es:

$p(A \cup B) = p(A) + p(B) - p(A \cap B)$ (¿Por qué?)

$p(h \text{ o } p) = p(h) + p(p) - p(h \text{ y } p) = 70/110 + 60/110 - 50/110 = 80/110$

Los eventos h y p son independientes cuando $p(h/p) = p(h)$, lo cual se lee: la probabilidad condicional de h dado p es igual a la probabilidad de h, es decir, que p no influye sobre h, o lo que es lo mismo, que h es independiente de p. Ahora bien, tenemos que siempre es verdad que

$p(h/p) = \frac{p(h \text{ y } p)}{p(p)}$

pero cuando son independientes $p(h/p) = p(h)$. Se tiene por tanto

$p(h/p) = p(h) = \frac{p(h \text{ y } p)}{p(p)}$ es decir $p(h)p(p) = p(h \text{ y } p)$

Es usual traducir esta expresión en una definición no sobre las proposiciones sino sobre los eventos que estas representan:

260 ♣ ***Definición.*** *Dos* ***eventos*** A ***y*** B ***son independientes*** *si y sólo si* $p(A \cap B) = p(A)p(B)$.

261 ***Ejemplo*** *Dos eventos son independientes si lo que pase con el uno nada tiene que ver con lo que pase con el otro. El juego de dados ilustra muy bien la situación. Definamos X como la suma de los puntos de las caras que queden hacia arriba y calculemos su distribución.*

Consideremos dos dados iguales en todo menos en el color para poder distinguirlos. La probabilidad de cada cara en cada dado es 1/6. Ahora bien, si uno tira los dos dados, hay 36 opciones diferentes y equiprobables. Por eso, la probabilidad de que salga una pareja cualquiera de caras es 1/36 que también es igual a (1/6)(1/6). Todo gracias a que lo que salga en un dado nada tiene que ver con lo que salga en el otro.

La tabla de la suma de lo que dan las dos caras es:

La suma de dos dados						
+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Ahora bien, si X es la suma de lo que salga en las caras que quedan hacia arriba en los dos dados, entonces X toma valores entre 2 y 12. La probabilidad de $X = 2$ es 1/36 pues 2 sale en una única casilla (1,1). La probabilidad de $X = 7$ es 6/36 pues hay 6 opciones de 36 que salga una suma de 7: (6,1), (5,2), (4,3),(3,4), (2,5) y (1,6). La tabla que da las probabilidades de X es:

Distribución de X = la suma de dos dados										
2	3	4	5	6	7	8	9	10	11	12
1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Observamos que la suma de todas las probabilidades es $36/36 = 1$, lo cual nos dice que vamos bien. Los casos más improbables son los que tienen una suma de 2 ó de 12. Eso explica por qué en los juegos en los cuales se gana por avanzar sumando el puntaje de los dos dados, como en el parqués, se premia tanto los aces como los seises. Según la tabla y superponiendo mentalmente una campana, la media de X es 7 y su desviación es algo así como 2.5. Todo esto explica por qué el juego de parqués es realmente entretenido y siendo sencillo es amado por toda la familia.

262 ***Ejemplo*** *Demostremos, en referencia al ejemplo que relaciona el tamaño de los ojos con el género, que los eventos h y p no son independientes.*

Verifiquemos si se cumple o no que $p(h \text{ y } p) = p(h)p(p)$.

Por un lado, $p(h \text{ y } p) = 50/110$

Por el otro: $p(h) = 70/110$ en tanto que $p(p) = 60/110$. El producto de esas dos probabilidades es diferente de 50/110. Por lo tanto, los eventos no son independientes. Claro que nos queda la duda de saber si nuestra decisión se debe a un efecto del muestreo, que siendo al azar a veces da una cosa y a veces otra, o si en realidad, hay una ligazón entre los dos eventos. Esta intriga es genérica para todo par de eventos y se soluciona por el procedimiento siguiente:

263 ***Estudio de la H_o en una tabla de contingencia.***

El análisis estadístico de una tabla de contingencia no se dirige a entender si un par de eventos es o no independiente. Más bien se estudian todos los pares de eventos posibles quitando los ojos de los eventos y poniéndolos en los factores que éstos representan. Ejemplo: queremos saber si los dos factores, tamaño de ojos y género son independientes o dependientes.

Factor Uno: tamaño de ojos. Factor Dos: género.

Valores de la muestra			
	Tiene ojos grandes	Tiene ojos pequeños	Totales
Hombre	20	50	70
Mujer	30	10	40
Totales	50	60	110

La H_o dice, siempre, que los dos factores estudiados son independientes. Esto implica que todo par de eventos es independiente, i. e. que la probabilidad de una intersección de dos eventos es el producto de las probabilidades de los eventos: $p(e_1 \cap e_2) = p(e_1)p(e_2)$. Por ejemplo, la probabilidad de sacar dos caras lanzando una moneda dos veces es 1/4 y es igual al producto de sacar cara la primera vez por sacar cara la segunda vez. La segunda vez, la probabilidad no cambia pues lo que salió la primera vez para nada influye: los dos eventos son independientes.

Para los eventos univariantes de la tabla nosotros tenemos:
Para el género $p(h) = 70/110 = 7/11$; $p(m) = 40/110 = 4/11$.
Para el tamaño de ojos: $p(g) = 50/110 = 5/11$; $p(p) = 60/110 = 6/11$.

Para los eventos del bivariante y aplicando las hipótesis nulas:
$p(h \cap g) = p(h)p(g) = (7/11)(5/11) = 35/121$
$p(h \cap p) = p(h)p(p) = (7/11)(6/11) = 42/121$
$p(m \cap g) = p(m)p(g) = (4/11)(5/11) = 20/121$
$p(m \cap p) = p(m)p(p) = (4/11)(6/11) = 24/121$

Ahora a partir de la las probabilidades fabricaremos una tabla de contingencia de lo esperado según la H_o. Para ello, multiplicamos cada probabilidad por el total de personas consideradas en la tabla de contingencia que viene de la muestra. En nuestro caso, el total de personas 110. Nosotros obtenemos los números siguientes:
$N(h \cap g) = 110p(h)p(g) = 110(7/11)(5/11) = 350/11 = 31{,}82$
$N(h \cap p) = 110p(h)p(p) = 110(7/11)(6/11) = 420/11 = 38{,}18$
$N(m \cap g) = 110p(m)p(g) = 110(4/11)(5/11) = 200/11 = 18{,}18$
$N(m \cap p) = 110p(m)p(p) = 110(4/11)(6/11) = 240/11 = 21{,}82$
Todas estas cifras suman 110.
Metemos todos estos datos en una tabla, la tabla de lo esperado:

Valores esperados bajo H_o			
	Tiene ojos grandes	Tiene ojos pequeños	Totales
Varón	31.82	38.18	70
Mujer	18.18	21.82	40
Totales	50	60	110

¿Cómo podemos saber que entre tantos cálculos no hemos cometido algún error? Para estar seguros, comparamos las sumas totales, por columnas y renglones, entre lo esperado y lo observado: deben ser prácticamente iguales a no ser pequeños errores de aproximación.

Medimos la discrepancia entre lo observado, la primera tabla, y lo esperado según la H_o, la tabla segunda. Si la discrepancia es pequeña, los datos observados respaldan la H_o y se acepta, pero si la discrepancia es grande, los datos observados no respaldan la H_o y se rechaza.

$$Discrepancia = \sum \frac{(Observado_{ij} - Esperado_{ij})^2}{Esperado}$$

$$Discrepancia = \frac{(20-31{,}82)^2}{31{,}82} + \frac{(50-38{,}18)^2}{38{,}18} + \frac{(30-18{,}18)^2}{18{,}18} + \frac{(10-21{,}82)^2}{21{,}82} = 3{,}93 + 3{,}66 + 7{,}68 + 6{,}40 = 21{,}67$$

La discrepancia se distribuye APROXIMADAMENTE como una χ^2 con grados de libertad = (columnas -1)(Filas -1). El chi-cuadrado crítico superior para una significancia de 0.05 con UNA COLA y 1 grado de libertad es 3.841 pues por tratarse de una cola $\beta = 2\alpha = 0{,}10$. Así que, estamos autorizados para rechazar la hipótesis nula: nuestros dos factores son dependientes, por lo que tener o no tener

los ojos grandes puede usarse como un indicador relativamente confiable de ser o no ser mujer. Si la discrepancia hubiese dado infinito, el indicador sería totalmente confiable.

Para entender porqué se rechaza con una cola, la superior, tomemos la discrepancia

$$Discrepancia = \sum \frac{(Observado_{ij} - Esperado_{ij})^2}{Esperado}$$

y evaluémosla cuando lo esperado es exáctamente igual a lo observado, lo cual implica que lo observado respalda lo esperado y que la Ho se acepta. Pues bien, en dicho caso la discrepancia es cero, lo cual dice que cero pertenece a la región de aceptación de la Ho. Como la chi cuadrado (lo mismo que la F) no tiene valores negativos por tratarse de varianzas que nunca son negativos, al rechazar la Ho se hace con la cola superior. Con todo, hay excepciones en problemas muy especiales.

Nota: Aunque la discrepancia entre lo que se ve y lo que se cree bajo la H_o tiene una distribución que se aproxima a una χ^2, se sabe que en ocasiones la aproximación no es muy buena y por tanto se usan correcciones o se veta la metodología. Si las muestras no son al azar, es decir, si son dependientes, o si la frecuencia absoluta de alguna casilla en una tabla de cualquier tamaño es menor que 5, se veta la metodología. Una **corrección famosa es la de Yates** para tablas 2×2:

$$Discrepancia_{Yates} = \sum \frac{(Observado_{ij} - Esperado_{ij} - 0{,}5)^2}{Esperado}$$

264 ***Se decide con una cola***

Observemos que la discrepancia siempre da positiva y que su rango puede ser desde cero hasta valores muy grandes. En la generalidad de los casos de tablas de contingencia, la H_o se rechaza con una cola, la superior, pues la dependencia entre los dos factores estudiados se mide por la magnitud de la discrepancia, dada por el valor de la χ^2 experimental que cuando hay total independencia vale cero. Pero en algunos casos, un valor demasiado pequeño puede ser indicio de que los datos contienen algo interesante. Por ejemplo, que se hizo trampa en el análisis de los resultados, pues una discrepancia demasiado pequeña diría que los dos factores estudiados son absolutamente independientes, cosa que sería muy beneficiosa, por ejemplo, en un informe sobre la relación entre los ingresos de los militares y los casos de violencia por Paramilitarismo.

265 ***Cálculos simplificados***

Para tablas grandes, los cálculos de la discrepancia se hacen muy tediosos. Por ello, es conveniente usar el siguiente teorema que da un método mucho más fácil y cuya demostración queda de ejercicio. Ayuda: en la discrepancia original, se desarrolla el cuadrado, se simplifica donde se pueda y se tiene en cuenta que el número total de lo observado y el número total de lo esperado según la H_o coinciden y es igual a, digamos n.

$$Discrepancia = \sum \frac{(Observado_{ij} - Esperado_{ij})^2}{Esperado_{ij}} = (\sum \frac{(Observado_{ij})^2}{Esperado_{ij}}) - n$$

Observemos que la discrepancia es suma de cuadrados entre números positivos. Por tanto, la discrepancia en una tabla de contingencia entre lo que se ve y lo que se cree no puede ser negativa o sino hay un error de cálculo.

266 ***Ejemplo*** *Un campesino se burla del Gobierno a quien llama Gobierno asesino pues ha sido testigo de atropellos hechos por miembros del Ejército, quienes tildan de rebelde a todo campesino que para guardar su vida y la de su familia se ve obligado a ayudar a los violentos guardándoles cosas y diciendo si el Ejército está por ahí. Un Servicio Especial de Inteligencia ha reportado muchos de estos casos a las Altas Esferas. Allí la preocupación es grande, pues no parece haber un remedio contundente a este problema. Precisamente, hoy se escuchará un informe del Mayor R.E., quien alega que el Ejército jamás tendrá el apoyo popular por dos razones fundamentales. Primera, el castigo no satisface el dolor del Campesinado pues según éste estos dos factores son indpendientes, no tienen nada que ver el uno con el otro. Segunda: el Glorioso Ejército de Colombia no entiende lo que es el honor: se roba una*

cuchara y a quien ha ayudado a los violentos como mínimo lo amenaza e insulta grotescamente. El Mayor argumenta su primera pretensión con la tabla siguiente que muestra el número de atropellos discriminado por el castigo dado según el Código y por la magnitud del atropello según el Campesinado:

Tipo de castigo	*Lo que se ve: Magnitud del atropello*			
	X: Pequeña	*Y: intermedia*	*Z: Enorme*	*Totales*
A:Bajo	*4*	*3*	*2*	*9*
B:Mediano	*3*	*2*	*1*	*6*
C:Alto	*2*	*1*	*1*	*4*
Totales	*9*	*6*	*4*	$n = 19$

Según el Mayor, es necesario endurecer las sanciones y que se inculque entre el personal militar un mejor adoctrinamiento sobre la gloria y el honor. ¿Justifica la tabla la primera pretensión del Mayor R.E.? Tome la significancia como el 5 %. ¿Cuál sería su propuesta para ayudar a solucionar este terrible problema? Tenga en cuenta que la generalidad del Comando ve con desesperanza cualquier propuesta pues los campesinos avisan por celular a los violentos sobre la presencia del Ejército, la cual en el mejor de los casos se escapa pero en el peor arma mortales emboscadas.

Solución:

Paso 1: Estimamos las probabilidades univariadas:

$p(A) = 9/19$

$p(B) = 6/19$

$p(C) = 4/19$

$p(X) = 9/19$

$p(Y) = 6/19$

$p(Z) = 4/19$

Paso 2: Formulamos la H_o: los dos factores son independientes, el tipo de castigo según el Código Militar y la magnitud de los atropellos vistos desde la perspectiva del Campesinado.

Paso 3: Calculamos lo esperado según la H_o: La probabilidad de la intersección de dos eventos es el producto de las probabilidades de cada uno. Ejemplo:

$p(A \bigcap X) = p(A)p(X) = \frac{9}{19} \times \frac{9}{19}$,

Esta probabilidad se multiplica por el número total de atropellos que es 19 para hallar el número de atropellos esperado en la casilla (A,X):

$\frac{9}{19} \times \frac{9}{19} \times 19 = \frac{9}{19} \times 9 = 4{,}26$

Repitiendo el mismo procedimiento para las demás casillas, obtenemos la tabla de lo esperado según la H_o. **La forma rápida** de hacerlo es así: en la tabla de los esperado, en la casilla de la fila i con columna j se pone el resultado de multiplicar el total de la fila i por el total de la columna j y se divide por el gran total, que en este caso es $n = 19$, que es diferente de los glis.

Lo que se cree según la H_o				
	X: Pequeña	Y: intermedia	Z: Enorme	Totales
A:Bajo	4.26	2.84	1.89	8.99
B:Mediano	2.84	1.89	1.26	5.9
C:Alto	1.89	1.26	0.84	3.99
Totales	8.99	5.9	3.99	$n = 19$

Paso 4: Calculamos la discrepancia entre lo que se ve (tomado de la tabla de la página anterior) y lo que se cree según la H_o. Para ello, copiamos la tabla de lo que se ve, lo *observado*, y también la de lo que cree, lo *esperado*, y computamos el término $Observado^2/Esperado$ y sumamos:

Cálculo de la discrepancia			
Columna	*Observado*	*Esperado*	$Observado^2/Esperado$
1	4	4.26	3.75
	3	2.84	3.17
	2	1.89	2.12
2	3	2.84	3.17
	2	1.89	2.12
	1	1.26	0.79
3	2	1.89	2.12
	1	1.26	0.79
	1	0.84	1.19
			$\Sigma = 19.22$

Ahora, al total que dio 19.22 le restamos $n = 19$ lo cual da una discrepancia de 0.22. Esta discrepancia también se denomina la χ^2_{obs} chi-cuadrado observado. Alerta azul: todos estos cálculos pueden hacerse directamente en un solo paso con la calculadora sin necesidad de tabla, lo cual da una mejor aproximación pues la calculadora toma 6 decimales en todos los pasos.

Paso 5: Se decide si la discrepancia es grande o chica. Para ello la comparamos con el valor crítico de χ^2_+ con significancia del 5 % con una cola. **Los grados de libertad** se calculan multiplicando el número de filas menos uno por el número de columnas menos uno. En nuestro caso tenemos una tabla con 3 filas y 3 columnas . Por tanto, los $gli = (3-1)(3-1) = 4$. El valor crítico de la χ^2_+ con significancia 0.05 y la cola superior es 9.48, pues $\beta = 0{,}10$. La discrepancia entre lo que se cree en la H_o y lo que se ve es entonces catalogada como muy pequeña. Por tanto, no vale la pena cuestionar lo que se cree.

Paso 6: Tomamos un veredicto: La H_o de la independencia de los 2 factores se acepta.

Paso 7: Hacemos una interpretación en los términos del problema: El Mayor R.E. se lució: la Ley Penal Militar y los conceptos del Campesinado no tienen nada que ver. Como consecuencia, ni la aplicación de la más estricta justicia podrá lograr que el Campesinado tenga en buen concepto al Ejército y del Gobierno. En general, la aplicación de la justicia tiene 3 objetivos: primero, hacer justicia al agraviado. Segundo: desalentar cualquier intento de cometer una infracción. Tercero: que el Pueblo quede contento y sienta que tiene un Gobierno que lo respalda y no que lo oprime. Por tanto, ese tal Código hay que reformarlo y hay que hacerlo con la participación activa del Campesinado y, sin duda alguna, de los violentos: al fin y al cabo, una buena proporción de ellos fueron campesinos que se sintieron fuertemente heridos por los atropellos del Ejército.

267 *Comparación múltiple de proporciones*

Nosotros hemos visto cómo se usa el caso 5 para comparar dos proporciones usando una t. Para comparar tres, cuatro o cualquier número de proporciones no se usa una comparación de pares de proporciones las veces que sea necesario, sino que utilizamos una tabla de contingencia. En dicho caso, la H_o es la igualdad de todas las proporciones. Esto significa: todas las proporciones observadas vienen de muestras al azar de la misma población. La H_a es que hay al menos un par de proporciones diferentes, es decir, que no todas las muestras vienen de la misma población. Tratemos de entender por qué una comparación múltiple de proporciones puede hacerse con una tabla de contingencia. Para fijar ideas, discutamos un ejemplo concreto:

Se trata de una materia que tiene 3 secciones y ha llegado a la Jefatura la queja de que la sección 3 tiene un profesor que pasa a todo el mundo sin estar necesariamente bien preparado con el único objetivo de que a él le vaya bien en la encuesta que los estudiantes responden para evaluar a los profesores. Para abrir la investigación, el Jefe tomó las notas de los últimos exámenes y contabilizó los que pasaron y los que perdieron. Tenemos: en la sección 1 pasaron 15 de 22, en la 2 lo hicieron 18/24 y en la 3 pasaron 20/21.

Esto parece mucho más un censo que una muestra aleatoria, por lo que aquí no habría estadística inferencial sino descriptiva: la proporción de los que pasaron en la sección 3 es la mayor. Sin embargo, podemos pensar que estos exámenes registrados sin previo aviso son una muestra aleatoria de las posibilidades generales de las tres secciones. Si aceptamos dicho criterio, estamos en un caso de estadística

inferencial y ya podemos estudiar la extrapolación de lo que pasó a lo que puede pasar en general. Podemos razonar como sigue: si la acusación es falsa, la proporción de los que pasaron es constante entre las secciones a no ser por efectos del azar del muestreo. Por tanto, al cambiar de sección no cambian las proporciones, es decir, el factor *pasar* no influye sobre el factor *sección*, por tanto, los dos factores son independientes. Y viceversa, pretender que los factores son independientes es lo mismo que decir que la proporción de los que pasan no depende de la columna o sección.

Comentario: Este tipo de test tiene un nombre oficial: **test de homogeneidad de proporciones** y su objetivo es ver si dos o más muestras son homogéneas con respecto a un factor categórico. Se supone que las muestras son independientes. Lo que hicimos en el párrafo anterior fue demostrar que un test múltiple de homogeneidad se puede resolver con la misma metodología usada en las tablas de contingencia.

Solución: Acomodamos los datos en una tabla de contingencia y seguimos exactamente el mismo formalismo para independencia de factores. La tabla queda:

Observado				
Pasar	Sección			
	X= sección 1	Y=sección 2	Z= sección 3	Totales
A= perdió	7	6	1	14
B= pasó	15	18	20	53
Totales	22	24	21	$n = 67$

H_o : la proporción de los que pasan es la misma en todas las secciones, cuya mejor estimación es 53/67. Las diferencias de las proporciones con respeto a ésta es por puro azar.

H_a : hay al menos un par de proporciones que es diferente, no por azar sino por algo sistemático.

Número de colas: una, pues la discrepancia se mide utilizando un cuadrado que anula los signos correspondientes a estar por arriba o por abajo de lo esperado.

Lo esperado bajo la H_o es:

Esperado: Pasar vs sección				
	X= sección 1	Y=sección 2	Z= sección 3	Totales
A= perdió	4.60	5.01	4.39	14
B= pasó	17.40	18.99	16.61	53
Totales	22	24	21	$n = 67$

Por ejemplo, el valor de la columna 2 con fila 2 es $53 \times 24/67 = 18{,}99$. Averiguemos ahora la discrepancia entre lo que se ve y lo que se cree según la H_o:

Cálculo de la discrepancia		
Observado	*Esperado*	$Observado^2/Esperado$
7	4.597	10.659
6	5.015	7.1786
1	4.388	0.228
15	17.403	12.929
18	18.985	17.067
20	16.612	24.079
		$\Sigma = 72.139$

La disrepancia es

$\chi^2_{obs} = 72{,}139 - n = 72{,}139 - 67 = 5{,}139$ pues $n = 67$.

Discrepancia entre lo que se observa y lo que se cree según la H_o: 5.139. La discrepancia se distribuye bajo la H_o como una chi-cuadrado con 2 gli. Como trabajamos con una cola y con $\alpha = 0{,}05$, $\chi^2_c = 5{,}9915$.

La discrepancia observada es pequeña: los datos no permiten asegurar que haya al menos un par de proporciones diferentes. Nuestra interpretación es que no hay aún datos suficientes para descartar que las diferencias observadas entre las diversas secciones se deban al azar de un muestreo pequeño, un único examen, y por lo tanto aún no se puede aceptar que el profesor regala nota. Lo que debe hacerse es reunir datos de muchos examenes.

Capítulo 13

Parcial IV

Objetivo y advertencia: Mientras en los temas pasados aprendimos el por qué de los conceptos y procediminentos, en el presente los aprendemos a utilizar para resolver problemas. Las variantes posibles son muchísimas y no hay un ejemplo para cada una. Por tanto, una parte fundamental que demanda atención es aprender a sortear las diversas combinaciones de datos, métodos y preguntas para resolver el problema dado.

Número de colas: En el corte pasado aprendimos a manejar hipótesis nulas en las cuales una diferencia de medias o de proporciones era igual a cero. Ahora trataremos el caso cuando dicha diferencia no es cero. Eso implica que una media o proporción es mayor que la otra. Entonces uno se pregunta: ¿Será que ya no se necesita la hipótesis alterna? Para entender qué pasa, supongamos que tenemos la siguiente pregunta en el problema: ¿Puede decirse que la media mayor es 5 unidades más que la menor? Para captar el número de colas reescribimos la pregunta así: ¿Puede decirse que la media mayor es igual a la menor más 5 unidades? La H_o ahora es clara: $\mu_1 - \mu_2 = 5$ y la $H_a : \mu_1 - \mu_2 \neq 5$ sería con dos colas. Ahora bien, supongamos la siguiente variante: ¿Puede decirse que la media mayor es POR LO MENOS 5 unidades más que la menor? En este caso la H_o sería: $\mu_1 - \mu_2 = 5$ y la $H_a : \mu_1 - \mu_2 > 5$ de cola superior. De similar manera se trata la cola inferior y la comparación de varianzas.

13.1. Taller IV

Atención: al interpretar las fórmulas de los ejemplos resueltos, tengamos presente que el símbolo z_α **es el** z **crítico asociado a la significancia** α **con una cola**. Su signo se saca por contexto, positivo para la cola superior, negativo para la cola inferior. En contraste, $z_{(\alpha,2)}$ **es el** z **crítico asociado a la significancia** α **con dos colas**. Igual sucede con la t. Por otra parte, los intervalos de confianza siempre van con dos colas pues, por ejemplo, la media poblacional puede estar arriba o abajo de la media de la muestra.

1 . Hacer un muetreo cuesta. Por tanto es necesario minimizar el costo pero sin sobrepasar la tolerancia admitida. ¿Cuántos eventos debemos muestrear para estimar la media de X si queremos que el error de estimación de la media sea como máximo E sabiendo que la desviación poblacional es de σ_X y la significancia es α? (Ej 215-d, pág. 210. Respuestas en la pág 316)

a) $E = 2$; $\sigma_X = 0{,}3$; $\alpha = 0{,}05$

b) $E = 0{,}2$; $\sigma_X = 0{,}3$; $\alpha = 0{,}01$

c) $E = 2$; $\sigma_X = 3$; $\alpha = 0{,}02$

d) $E = 2$; $\sigma_X = 30$; $\alpha = 0{,}07$

2 . Hallar el intervalo de confianza para la media si una muestra al azar de 28 eventos dio una media muestral de 35 y si la desviación poblacional es de 1.7. Decidir si se justifica la apreciación de que la media poblacional es 42 (Ej 218, pág. 211).

3 . Hallar el intervalo de confianza para la media si una muestra al azar de 15 eventos dio una media muestral de 35 y una desviación muestral de 1.7. Decidir si se justifica la apreciación de que la media poblacional es 42 (Ej 224, pág. 215).

4 . Si una muestra al azar de 12 eventos dio una media de 3 y una desviación de 0.2, ¿Se justifica decir que la media poblacional vale 3.5? Resolver por prueba de hipótesis y por intervalo de confianza (Ej 225, pág. 215).

5 . Si una muestra al azar de 9 eventos dio una media de 14 y una desviación de 2, ¿Se justifica decir que la media poblacional vale 7? Resolver por prueba de hipótesis y por intervalo de confianza (Ej 225, pág. 215).

6 . Se hizo un censo y dio que la edad de matrimonio de las mujeres tenía media 22 años y una desviación de 4.3 años. No mucho tiempo después se hizo una muestra al azar de 23 mujeres y dio una media de 20.4 años con una desviación de 3.8. a) ¿Ha cambiado la media? b) ¿Ha cambiado la varianza? c) Un segundo muestreo de 19 mujeres dio una media de 22 años y una desviación de 2.5. Comparar con dos colas las medias usando las dos muestras. d) Hallar el intervalo de confianza para la relación entre las varianzas.

Ayuda al detalle y respuestas: a) Los estadígrafos de un censo son de todo el mundo, por lo tanto son poblacionales. El censo da $\mu = 22$ y $\sigma = 4{,}3$. La primera muestra da $\bar{X}_{23} = 20{,}4$ y $s = 3{,}8$. Para comparar una media muestral con una media poblacional sabiendo la desviación poblacional se utiliza el caso 1 descrito en la parte a) del ejemplo 215 pág. 210: $z_{obs} = -1{,}78$; $z_c = 1{,}96$ No cambió la media. Atención: existe la posibilidad de usar s en vez de σ y aplicar una t pero preferimos una z pues σ es más exacta que s. b) Para comparar una varianza muestral con una poblacional se elevan las desviaciones al cuadrado y las varianzas se comparan siguiendo el caso 6 en el ejemplo 237 de la pág. 224. Rta: $\chi^2_{obs} = 17{,}18$. Como $\chi^2_- = 10{,}982$ y $\chi^2_+ = 36{,}781$ no cambió la varianza. Alternativa: las dos varianzas también pueden compararse por medio de una F, caso 7, como en el ejemplo 245 de la pág. 229, pero hay que poner los gli de la varianza de los datos del censo como infinito. c) Para comparar las medias sabiendo las medias y varianzas de dos muestreos se utiliza primero el caso 7 del folleto de fórmulas, para estudiar la igualdad entre varianzas, y después el caso 8 ó 9 del mismo folleto. d) El intervalo de confianza para la relación entre las varianzas se hace según el ejemplo 243 de la pág. 227.

7 . Estimar el intervalo de confianza para la proporción de positivos si de una muestra al azar de 350 salieron 178 positivos. Decidir si se justifica decir que la proporción de positivos es 0.7 (Ej 227, pág. 217).

8 . Se tienen dos definiciones de historia. Primera: la historia es lo que se recuerda y queda registrado para que el futuro lo conozca. Segunda: La historia es lo que se selecciona para ser recordado conforme a los intereses dominantes de cada cual, sea persona, grupo o país. A un grupo de 17 jóvenes se les pidió elegir una entre las dos definiciones de historia. De ellos, 7 escogieron la primera. a) Calcule el intervalo del 95 % de confianza para la proporción de personas que escogieron la segunda definición pues consideran que la historia es un reporte amañado de lo que pasó y que nunca habrá justicia sobre los engaños hechos por los grupos dominantes. b) ¿Justifican los datos la creencia de que las dos definiciones de historia se eligen por igual entre los jóvenes? Advertencia: esta pregunta es tramposa pues como habla de dos proporciones, uno pensaría que hace por el caso 5. Eso sería un grave error pues las proporciones no son independientes pues entre las dos suman UNO. Hay sólo una proporción independiente, que aporta información no redundante, y por tanto se hace con el caso 4. (Ej 227, pág. 217).

9 . Un estudiante se quejó diciendo que en otro curso el examen fue más fácil. Para sustentarlo mostró los datos siguientes. Curso propio: 44 estudiantes, pasaron 34. El otro curso: 43 estudiantes, pasaron 38. El profesor al ver los datos dijo: esto no contradice lo que siempre he pensado, a saber, que la proporción de los que pasan en el otro curso es 0.2 mayor que en el nuestro, lo cual se explica porque yo doy garantía de salir bien preparado de mi curso. ¿Tomando una confianza del 95 %, es aceptable lo que dice el profesor? Resuelva por prueba de hipótesis y por IC. (Ejs 234 y 235 , pág. 221)

10 . De una muestra al azar de 17 eventos, se obtuvo una varianza de 9 (esta varianza es la muestral). Hallar el intervalo de confianza para la varianza y para la desviación (poblacionales). a) Decidir si se justifica decir que la varianza vale 30 (Ej 240, pág. 225). b) ¿Qué pasa si hay 82 eventos? (Ayuda: cuando hay muchos eventos la tabla de la chi cuadrado no sirve y hay que usar fórmulas especiales: Ej 238, pág. 224.)

11 . Hallar el intervalo de confianza para la relación entre varianzas si $var1 = 0{,}8$ con 9 datos y $var2 = 2{,}3$ con 11 datos. ¿Se justifica decir que la proporción entre varianzas es de 1 a 7 (varianza poblacional grande entre varianza poblacional pequeña = 7)? (Ej 243, pág. 227).

12 . Tratamiento uno: 4, 5, 6, 5, 4, 5, 6, 4, 5, 5. Tratamiento dos: 4, 7, 8, 9, 6, 8, 4, 10, 11. Halle los intervalos de confianza para cada σ^2 (Ej 239, pág. 224) y para la relación de varianzas (Ej 245, pág. 229).

13 . Tratamiento uno: 4, 6, 8, 4, 5, 7, 8, 10, 9, 10. Tratamiento dos: 9, 5, 5, 6, 8, 9, 7, 11, 10, 11, 10. ¿Es la media mayor 0.3 unidades más grande que la menor ? Ayuda: primero se decide si las varianzas poblacionales son iguales o distintas (ej 197, pág. 176). Si las varianzas son iguales, las medias se estudian según el ej 246 pág. 231, si son diferentes, se sigue el ej 249 pág. 235.

14 . La Juez de la República Elsa Munevar tiene que decidir si el presunto lavador de dinero M.E. ha de quedar en libertad o si debe ser declarado culpable. La acusación presentada por el Fiscal dice lo siguiente: “Según los registros entregados a la Oficina de Impuestos Nacionales los 5 almacenes del acusado vendieron en el mes anterior a la intervención del Estado un promedio de 1.2 con una desviación de 0.2. En el mes posterior a la intervención y vigilancia estricta del Estado las ventas fueron 1.1 con una desviación de 0.3. Como se ve, las ventas cambiaron un 10 %. Eso respalda claramente la tesis de que antes había ingresos provenientes del lavado de dinero asociados al Narcotŕafico. Si unimos esto a las demás pruebas, no queda más remedio que declarar culpable a M.E.” A lo cual la Defensa arguyó: “Debido a que con una significancia del 5 %, se puede aceptar que las dos medias aludidas son iguales, y teniendo en cuenta que la presunción de inocencia domina sobre la de culpabilidad, exijo en nombre de la Constitución la libertad inmediata de mi cliente”. La Juez le (la) ha contratado como asesor(a) experto(a) en Estadística. Para ello a) Halle el intervalo de confianza del 95 % de la diferencia de medias. b) Decida si el Fiscal es respaldado por el IC que halló. c) Usando el mismo intervalo decida si la Defensa está en lo cierto. d)Tome su veredicto. e) Todo el mundo sabe que ME es lavador de dinero: ¿cuál es la estrategia usada por ME para tratar de evadir la estadística? f)¿Qué debería hacerse para contrarrestar su astucia? (Un renglón) (Ej 248, pág. 232).

15 . Una baraja consta de 52 cartas divididas en cuatro palos que son corazones, diamantes, tréboles y picas. Los corazones y diamantes son rojos, los tréboles y picas son negros. Cada uno de los palos consta de 13 cartas: As, 2, . . ., 10, J, Q y K. Se eligen al azar dos cartas sin reemplazo. a) Halle la probabilidad de que las 2 cartas tengan la letra J. b) Halle la probabilidad de que entre las dos cartas haya al menos una J(Ej 253, pág. 239).

16 . Bianca a veces se maquilla los ojos y a veces no. El 35 % de la veces se los maquilla . Cuando se los maquilla , el 55 % de las veces es invitada a salir. Cuando no, el 40 % de las veces la invitan a salir. Dado que hoy fue invitada, a) calcular la probabilidad de que ella se haya hoy maquilldo los ojos (Ej 254, pág. 240).b) ¿Se atrevería a decir que ella se los maquilló?

17 . El Señor Presidente ataca a sus opositores en el Congreso el 30 % de las veces que habla para la prensa. Cuando los ataca, el 5 % de las veces disminuye su popularidad, el 90 % permanece igual y el otro 5 % aumenta. Cuando no los ataca, el 15 % de las veces disminuye su popularidad, el 80 % permanece igual y el otro 5 % aumenta. Esta vez, la popularidad aumentó. ¿Atacó el Presidente a sus opositores ?

18 . De acuerdo con Melissa, los papás son terriblemente simples. Ellos pueden estar disgustados con la hija por una de 3 razones: la hija gasta mucha plata (70 %), la hija tiene amigos que no le convienen (20 %) y la hija ya no lo quiere como antes (10 %). Después de ponerse bravos, pueden reconciliarse así: por plata, el 40 %; por amigos, el 50 %, por falta de amor, el 70 %. Hoy ella llegó a casa y encontró al papá fríO y con cara de puño. Melissita puso a prueba su dulzura y falló. ¡Si tan sólo supiera la causa de la braveza!. Dele un diagnóstico probabilístico. Ayuda: se hace un árbol de cuya raíz salen 3 ramas: plata, amigos y amor. Cada rama se bifura en reconciliado y no-reconciliado. Para saber cuál es la razón más probable de la braveza del papá, se calcula la probabilidad de que el papá se haya puesto bravo por plata dado que no se ha reconciliado. Y así con las otras razones. La causa más probable es aquella que tenga una probabilidad condicionada mayor.

19 . Las artesanías hechas a mano tienen mayor valor económico que las hechas a máquina y además generan empleo. En una fábrica hay 3 operarios, P, J, L cuyo rendimiento corresponde al 25 %, 35 % y 40 % respectivamente. Debido a que hay una fuerte competencia por la producción y por la calidad, un operario se debate entre hacer menos pero mejor o hacer más pero peor. El 10 % de las piezas de P no pasan el control de calidad, el 7 % de las de J tampoco y lo mismo el 12 % de las piezas de L. Se decide dar un premio al mejor operario. ¿A quién debe premiarse?

20 . Una urna contiene 7 bolitas blancas y 12 negras. Una segunda urna contiene 4 bolitas blancas y 18 negras. Se extrae una bolita de la primera urna y se pasa a la segunda urna. Luego se extrae una bolita de la segunda urna. a) Hacer un árbol que represente el experimento asignándoles probabilidades a las ramas. b) Calcular la probabilidad de que la bolita extraída de la segunda urna sea negra. (Ej 255 pág. 241).

21 . Sea Y la variable aleatoria que resulta de sumar los puntos de las caras que quedan hacia arriba en un juego de tres dados bien balanceados. Halle la distribución de Y, es decir su tabla de probabilidades y dibuje su diagrama de barras. Por ejemplo, si los tres dados caen as, Y vale 3 y $p(Y = 3) = 1/216$ pues 3 sólo pasa con 3 aces entre $6^3 = 216$ opciones. Use dicho diagrama para estimar la media y la desviación. Ahora halle rigurosamente dichos valores. Ayuda: Siga el ejemplo 261, pág. 246 y haga la tabla de Y cuando el tercer dado cae 1 y los dos primeros dados caen cualquier valor. Repita el mismo procedimiento para los otros 5 valores posibles del tercer dado. Ahora para cada valor de Y entre 3 y 18 cuente las opciones y calcule la probabilidad. La suma total de las probabilidades debe dar 216/216.

22 . La tabla siguiente denota la magnitud de la rabia contra los violentos en una persona entrevistada y el candidato por el cual ella piensa votar en las próximas elecciones (sec 12.2, ej 257+258 +259 +262 pág. 243):

Rabia	Candidato escogido			
	X	Y	Z	Totales
A=Poca	13	8	4	
B=Mediana	12	24	8	
C=Mucha	2	22	15	
Totales				

Si se toma una persona al azar, halle la probabilidad de que dicha persona sea de

a) A

b) X

c) A y X que debe leerse de la tabla. No es p(A)p(X) pues nadie ha dicho que A y X sean independientes.

d) A o X (la unión).

e) A dado X: probabilidad de A dado que el universo es X que se nota $p(A/X)$.

f) Verifique que $p(A/X) = p(A \cap X)/p(X)$

g) ¿Puede decirse que los eventos A y X son independientes?

23 . Puesto que en clase se sigue un texto, no es claro para qué se requiere ir a clase. La siguiente tabla muestra la incidencia de ir o no ir a clase sobre la nota:

Asistencia a clase	Nota			
	X= Baja	Y = Mediana	Z= Alta	Totales
A=< 30 %	13	8	4	
B=30 %-50 %	12	24	8	
C=> 50 %	2	22	15	
Totales				

Si se toma una persona al azar, halle la probabilidad de que dicha persona sea de

a) B

b) Z

c) B y Z

d) B o Z (unión)

e) B dado Z. Probabilidad de B dado que el universo es Z.

f) Verifique que $p(B/Z) = p(B \cap Z)/p(Z)$

g) ¿Puede decirse que los eventos B y Z son independientes?

h) ¿Son los factores ir a clase y nota independientes? Ayuda: eventos y factores se tratan diferente (Ej 266, pág. 248).

i) ¿Vale la pena ir a clase?

j) Relacione la independencia de eventos con la de factores.

24 . Un grupo de Rock pesado da conciertos y sus enemigos dicen que son lunáticos, es decir, que la luna llena los excita y les da poder para ser tan expresivos. La siguiente tabla muestra el tipo de luna y la calidad del concierto dado. a) Analice la tabla por intuición y decida si los enemigos del grupo tienen razón. b) Use un test de independencia de factores para tomar una decisión (Ej 266, pág. 248). c) Use un test de homogeneidad de proporciones para tomar una decisión (Ej 267, pág. 250).:

Tipo de luna	Calidad del concierto			
	X= Baja	Y=Mediana	Z=Alta	Totales
A= Llena	6	7	19	
B=Otra	4	3	8	
Totales				

25 . El tiempo invertido en la preparación de una campaña, en semanas, y los resultados de la campaña se muestran en la tabla siguiente, la cual hay que analizar por intuicion + tecnología (Ej 266, pág. 248):

Observado: Tiempo invertido y tipo de resultado				
Tiempo invertido	Tipo de resultado			
	X= Pobre	Y=Mediano	Z=Bueno	Totales
A=< 10	23	43	14	
B=10 − 20	29	37	4	
C=> 20	2	8	9	
Totales				

26 . Leonardo se ha sentido presionado por sus amigas predilectas, quienes le exigen que elija novia entre ellas. Su mejor amigo le dice: tome una muestra aleatoria de las citas que ha tenido y a cada cita clasifíquela como buena o mala, según se haya sentido. Investigue para cada una de sus amigas la proporción de citas en que se sintió bien y elija como novia aquella niña cuya proporción de citas buenas es la mayor. Leonardo registra las frecuencias relativas observadas de buenas citas: con Patricia 6 de 8, Con Alba 10 de 13, con Tatiana 14 de 18. a) ¿Es el criterio elegido suficiente para que Leonardo elija novia? b) ¿Qué precauciones le aconsejaría Usted a Leonardo para que obtenga una muestra aleatoria, sin interferencia entre las amigas ni ellas mismas? Ayuda: Ej 267, pág. 250. Decida primero por intuición y luego por tecnología.

13.2. Parcial IV: con folleto de fórmulas

Este parcial es sin libro ni apuntes. Es con calculadora y con el folleto de fórmulas. Presentamos varios ejemplos de lo que podría ser un parcial. **Lo primero que hay que hacer para resolver cada pregunta de cada problema es relacionarla con alguno de los temas del diccionario de experimentos o con los de tablas de contingencia y después seguir los ejemplos, que pueden ser varios, pero sólo señalamos el primero.** ***Es necesario buscar el ejemplo que mejor se ajusta a la pregunta dada.*** Es buena idea trabajar con el resumen de fórmulas que se puede sacar en el examen final y no con el libro que no puede sacarse. Los casos son:

1. Comparación de una media muestral con una media poblacional conociendo la desviación poblacional σ: se usa una z. Ej 215, pág. 210.

2. Comparación de la diferencia de dos medias experimentales con la diferencia esperada de dos medias poblacionales conociendo las varianzas poblacionales: se usa una z. Ej 221, pág. 212.

3. Comparación de una media muestral con una media poblacional conociendo la desviación de la muestra: se usa una t. Ej 223, pág. 214.

4. Comparación de un porcentaje o proporción observada o frecuencia relativa con una probabilidad (poblacional): se usa una z. Ej 227, pág. 217.

5. Comparación de la diferencia de dos proporciones experimentales con la diferencia esperada de las probabilidades: se usa una z. Ej 233, pág. 220.

6. Comparación de la varianza de una muestra al azar con una varianza poblacional: se usa una chi-cuadrado χ^2. Ej 237, pág. 224 y siguientes: hay varios casos dependiendo del número de datos.

7. Comparación de la relación de dos varianzas experimentales con la relación esperada de varianzas poblacionales: se usa una F. Ej 242, pág. 226.

8. Comparación de la diferencia de dos medias experimentales con la diferencia esperada de dos medias poblacionales conociendo las dos varianzas muestrales de las cuales se infiere que las varianzas poblacionales son iguales. Para comparar varianzas se usa una F con dos colas y para medias una t. Ej 246, pág. 231. Hay dos casos:

    a) Si la H_o es que la diferencia de medias es cero.

    b) Si la diferencia de medias postulada en la H_o es cualquiera, cero o no.

9. Comparación de la diferencia de dos medias experimentales con la diferencia esperada de dos medias poblacionales conociendo las dos varianzas muestrales de las cuales se infiere que las varianzas poblacionales son diferentes. Para comparar varianzas se usa una F con dos colas y para medias una t. Ej 249, pág. 235.

10. Comparación de la diferencia de dos medias experimentales con la diferencia esperada de medias poblacionales para datos apareados (miopía ojo izquierdo vs. miopía ojo derecho): se usa una t. Ej 251, pág. 238.

11. Para calcular los intervalos de confianza de medias o sus diferencias, uno se enfoca en el caso que corresponda al estudio de medias y de allí toma la fórmula para dicho intervalo. Para comparar medias primero se comparan varianzas pues los IC para varianzas iguales son más angostos, menos incertidumbre, que los correspondientes cuando las varianzas son diferentes.

12. Árboles de probabilidades, Ej 255 pág. 241; Ej 254, pág. 240.

13. Probabilidades asociadas a tablas de contingencia: sec. 12.2, pág. 243; ej 257+258 +259 +262 pág. 243). Independencia de dos **eventos**: Ej. 261 pág. 246.

14. Independencia de dos **factores** dada una tabla de contingencia: se usa una χ^2 (Ej 266, pág. 248).

15. Homogeneidad de varias proporciones: se usa una χ^2 (Ej 267, pág. 250).

Existen variantes de problemas en los cuales uno tiene que hacer una elaboración previa. Por ejemplo: la media y la varianza de una lista de datos se calcula según el ejemplo 17 pág. 9. Pero si se trata de una tabla de frecuencias absolutas, por el ejemplo 19 pág. 10. Si lo que se da es una tabla de frecuencias acumuladas, lo primero que se hace es calcular la tabla de frecuencias absolutas asociadas y después se le calcula la media y varianza a dicha tabla. Si se da una tabla de frecuencias por rangos, hay que definir la marca de clase y calcular la tabla de frecuencias absolutas, como en el ejemplo 23 pág. 13.

Para decidir una H_o de dos colas tenemos dos métodos: el del estadígrafo de contraste y el del IC (intervalo de confianza). Uno puede elegir cualquier método pero el del IC permite discusiones más completas. Es imperativo no confundir las dos metodologías: si uno elige el método del estadígrafo de contraste, todas las comparaciones son entre estadígrafos, comparando el valor observado con los valores críticos. Pero si uno usa el IC, digamos de la varianza, todas las comparaciones son entre varianzas, la poblacional de la H_o con los valores extremos de la varianza dados por el IC respectivo.

268 ***Ejemplos cuando la H_o es una diferencia diferente de cero. Una y dos colas. Se enfatiza la coherencia en todos los pasos.***

En una muestra al azar de 25 jóvenes universitarios, 20 confesaron que le tenían temor al futuro. Pero el correspondiente número en otra muestra independiente de jóvenes que tenían trabajo, fue de 12 entre 29. ¿Puede decirse que la proporción de los que le tienen miedo al futuro entre los que estudian dobla la de los que trabajan?

SOLUCIÓN

Paso 0: advertencia. Cuando la Ho es una diferencia igual a cero para rechazar con dos colas no hay que preocuparse por quien restar de quien pues tanto la t como la z son simétricas. Pero cuando

la Ho es un diferencia diferente de cero para rechazar con dos colas es necesario preocuparse y ser coherente en los diversos pasos. Lo ilustraremos claramente.

Paso 1: Se encuentra el caso, se asignan las variables, se plantea la H_o, el número de colas y el estadígrafo de contraste:

Caso 5.

$f = X/m = 20/25 = 0{,}8$, frecuencia relativa de los que tienen temor del grupo de los que estudian. La probabilidad (poblacional) es p.

$g = Y/n = 12/29 = 0{,}414$, frecuencia relativa de los que tienen temor del grupo de los que trabajan. La probabilidad (poblacional) es r.

La H_o dice que $p = 2r$, es decir, $p - r = 2r - r = r$. Como no nos dan r, tomamos abusivamente $r = g = 0{,}414$ para poder hacer el problema. Tenemos $H_o : p - r = 0{,}414$. Se rechaza con dos colas pues la pregunta enfatiza una igualdad que se puede negar por arriba o por abajo. El caso es el 5 subacaso 3.

$$s_b = \sqrt{\frac{f(1-f)}{m} + \frac{g(1-g)}{n}} = \sqrt{\frac{0{,}8(0{,}2)}{25} + \frac{0{,}414(1-0{,}414)}{29}} = 0{,}1215$$

$$\text{discrepancia} = t_{obs} = \frac{(f-g)-(p-r)}{s_b} = \frac{(0{,}8-0{,}414)-(0{,}414)}{0{,}1215} = -0{,}227$$

Verifiquemos la coherencia en esta expresión: (0,414) es positiva y viene de la H_o en donde se calculó la probabilidad asociada a los que estudian menos la asociada a los que trabajan. Por lo tanto, en lo observado se debe computar la frecuencia relativa asociada a los que estudian menos la asociada a los que trabajan, tal como ha sido hecho.

Paso 2: Se formula H_a, se halla t_c, se hace la gráfica, se toma un veredicto respecto a H_o, se interpreta el veredicto en términos de la pregunta del problema:

Formulemos la H_a con dos colas:

$H_a : p - r \neq 0$.

glis = el más pequeño entre 25-1 y 29 -1. Da 24. La $t_c = 2{,}063$ con significancia del 5 %.

Veredicto: la t_{obs} mide la discrepancia entre lo que se ve y lo que se cree en la H_o. Como t_{obs} no tiene nada da extrema, la discrepancia es pequeña. Por tanto nos quedamos con lo que creemos: se acepta la H_o.

Interpretación: Los datos nos invitan a creer que la proporción de los que le tienen miedo al futuro entre los universitarios dobla a la de los que trabajan. Sería bueno hacer talleres sobre el tratamiento de las emociones negativas donde se plateen soluciones alternas al uso de medicamentos para clamar los nervios y de drogas sicoactivas.

VARIANTE. Con los mismos datos, resolver la pregunta: ¿Puede decirse que la proporción entre los universitarios de los que le tienen miedo al futuro es más del doble que entre los que trabajan ?

SOLUCIÓN 1 de la variante

Paso 1: Se encuentra el caso, se asignan las variables, se plantea la H_o, el número de colas y el estadígrafo de contraste:

Caso 5.

$f = X/m = 20/25 = 0{,}8$, frecuencia relativa de los que tienen temor del grupo de los que estudian. La probabilidad (poblacional) es p.

$g = Y/n = 12/29 = 0{,}414$, frecuencia relativa de los que tienen temor del grupo de los que trabajan.La probabilidad (poblacional) es r

La H_o dice que $p = 2r$, es decir, $p - r = 2r - r = r$. Como no nos dan r, tomamos $r = g = 0{,}414$ (esto es un abuso machetero para poder hacer el problema). Tenemos $H_o : p - r = 0{,}414$. Se rechaza con una cola pues la pregunta enfatiza una desigualdad que se puede negar de una sola manera. El caso es el 5 subacaso 3.

$$s_b = \sqrt{\frac{f(1-f)}{m} + \frac{g(1-g)}{n}} = \sqrt{\frac{0{,}8(0{,}2)}{25} + \frac{0{,}414(1-0{,}414)}{29}} = 0{,}1215$$

discrepancia $= t_{obs} = \frac{(f-g)-(p-r)}{s_b} = \frac{(0{,}8-0{,}414)-(0{,}414)}{0{,}1215} = -0{,}227$

Verifiquemos la coherencia en las dos restas: en ambas se resta la proporción asociada a los que estudian menos la asociada a los que trabajan.

Paso 2: Se formula H_a, se halla t_c, se hace la gráfica, se toma un veredicto respecto a H_o, se interpreta el veredicto en términos de la pregunta del problema:

Formulemos la H_a con una cola cola:

para que la frecuencia relativa de los que tienen pánico del grupo de los que estudian f sea más que el doble de la de los que estudian g se necesita que $(f-g) > 0{,}414$. En dicho caso, la t_{obs} sería igualmente positiva:

$H_a : p - r > 0$.

glis = el más pequeño entre 25-1 y 29 -1. Da 24. La $t_c = 1{,}7$ con sola superior y significancia del 5 %.

Veredicto: la t_{obs} mide la discrepancia entre lo que se ve y lo que se cree en la H_o. Como t_{obs} no tiene nada da extrema, la discrepancia es pequeña. Por tanto nos quedamos con lo que creemos: se acepta la H_o.

Interpretación: Los datos nos invitan a creer que la proporción de los que le tienen miedo al futuro entre los universitarios dobla a la de los que trabajan. Nuestro veredicto nos invita a no ser alarmista y a no andar pregonando que meterse a estudiar es condenarse al suicidio o a volverse drogadicto.

SOLUCIÓN 2 de la variante

Paso 1: Se encuentra el caso, se asignan las variables, se plantea la H_o, el número de colas y el estadígrafo de contraste:

Caso 5. Asignamos las variables al revés:

$g = X/m = 20/25 = 0{,}8$, frecuencia relativa de los que tienen temor del grupo de los que estudian. La probabilidad (poblacional) es r.

$f = Y/n = 12/29 = 0{,}414$, frecuencia relativa de los que tienen temor del grupo de los que trabajan. La probabilidad (poblacional) es p

La H_o dice que $r = 2p$, es decir, $p - r = p - 2p = -p$. Como no nos dan p, tomamos $p = f = 0{,}414$ para poder hacer el problema. Tenemos $H_o : p - r = -0{,}414$. Se rechaza con una cola pues la pregunta enfatiza una desigualdad que se puede negar de una sola manera. El caso es el 5 subacaso 3.

$s_b = \sqrt{\frac{f(1-f)}{m} + \frac{g(1-g)}{n}} = \sqrt{\frac{0{,}8(0{,}2)}{25} + \frac{0{,}414(1-0{,}414)}{29}} = 0{,}1215$

discrepancia $= t_{obs} = \frac{(f-g)-(p-r)}{s_b} = \frac{(0{,}414-08)-(-0{,}414)}{0{,}1215} = 0{,}227$

Verifiquemos la coherencia en las dos restas: en ambas se resta la proporción asociada a los que trabajan menos la asociada a los que estudian.

Paso 2: Se formula H_a, se halla t_c, se hace la gráfica, se toma un veredicto respecto a H_o, se interpreta el veredicto en términos de la pregunta del problema:

Formulemos la H_a con una cola cola:

La frecuencia relativa de los que tienen pánico del grupo de los que estudian g resulta ser más que el doble de la de los que trabajan f si $g = 0{,}9$ y $f = 0{,}414$. En dicho caso,

$t_{obs} = \frac{(f-g)-(p-r)}{s_b} = \frac{(0{,}414-0{,}9)-(-0{,}414)}{0{,}1215} = \frac{(0{,}414-0{,}9)+0{,}414}{0{,}1215} = -0{,}072$ que queda en la cola inferior. Por tanto, rechazamos con la cola inferior.

$H_a : p - r < 0$.

glis = el más pequeño entre 25-1 y 29 -1. Da 24. La $t_c = -1{,}7$ con cola inferior y significancia del 5 %.

Veredicto: la t_{obs} mide la discrepancia entre lo que se ve y lo que se cree en la H_o. Como $t_{obs} = -0{,}227$ no tiene nada da extrema en la cola inferior, la discrepancia es pequeña. Por tanto nos quedamos con lo que creemos: se acepta la H_o.

Interpretación: Los datos nos invitan a creer que la proporción de los que le tienen miedo al futuro entre los universitarios dobla a la de los que trabajan. Podría ser más del doble por azar (por causa de

la visita de un grupo internacional de música nadie estudió y se colgaron en las materias) pero no por algo sistemático. Que no cunda el pánico.

269 ***Versión 1***

1. Un examen puede salir mal diseñado por múltiples razones. Una de ellas es que sea imposible. Como consecuencia, todos los estudiantes tienen la misma nota, 1 para todos y varianza 0. Por lo tanto, se enfatiza que las notas de un buen examen deben separar los malos estudiantes de los regulares, de los buenos, de los muy buenos, de los excelentes. Eso implica que las notas deben estar desperdigadas y que por tanto su varianza debe ser alta. Se ha establecido en cierta universidad que un examen que se califica de 0 a 25 debe tener una varianza de por lo menos 18. Si la nota de un examen final, X, se distribuye normalmente, a) hallar el intervalo de confianza del 95 % para la varianza poblacional si la varianza de una muestra al azar de 16 datos tomados al azar fue de 32.8. b) ¿Diría Usted que el examen cumple con el requerimiento de la mínima varianza? c) Repetir los cálculos si el número de datos es 45. d) ¿Considera Usted que en este caso el examen clasifica como de buen diseño?

2. Un panadero se da cuenta de que el horno que tiene produce panes quemados en un lado y crudos en el otro, es decir, con alta variabilidad de cocimiento, de hecho, la varianza de cocimiento medida sobre 19 panes tomados al azar fue de 20.8. El panadero se dice a sí mismo: si logro bajar la varianza cinco veces, me voy de vacaciones. Y se pone a ajustar su horno. Después, sobre 16 panes tomados al azar la varianza fue de 10.2. a) Suponiendo que el cocimiento se distribuye normalmente, hallar el intervalo de confianza del 95 % para la relación entre varianzas poblacionales antes y después de arreglar el horno b)¿Tiene derecho el panadero a irse de vacaciones? Mejor dicho, ¿se justifica decir que dicha relación puede ser de 5 a 1, es decir, que la varianza grande entre la pequeña da 5 o que la relación entre varianzas poblacionales es 5? c) La esposa alegó: ese arreglo ni mejoró ni empeoró el horno: ¿es esto respaldado por el IC? d) ¿Cuál es el subintervalo del IC que favorece a la esposa (no irse de vacaciones) y cuál al esposo? e) ¿A qué acuerdo llegaron?

3. Un abogado tiene la impresión de que sus clientes más importantes muy sutilmente le exigen que haga un posgrado en derecho de patentes. Complacerlos es una decisión difícil, pues significa más esfuerzo y menos plata y en cambio no hay garantía de que su situación se mejore sustancialmente. Para ayudarse en su decisión, ha recolectado los siguientes datos que muestran la relación entre la universidad donde se estudia dicho posgrado y la mejoría de la situación financiera.

Universidad y mejora financiera				
Universidad	Mejora financiera			
	X:Poca	Y: Mediana	Z:Mucha	Totales
UniA	2	1	1	
UniB	3	2	1	
UniC	2	3	2	
Totales				

Por ejemplo, hubo 3 personas que estudiaron en UniC y que obtuvieron una mejoría mediana.

Por favor, analice la tabla y dele un consejo a ese abogado. Tome una significancia del 5 %.

270 ***Versión 2***

1. Es claro que abrir una autopista podría mejorar la calidad de vida pues el tiempo promedio de llegada al trabajo disminuye y por tanto permite dormir un poco más. El problema es la varianza. a) Pruebe que si la varianza del tiempo de llegada aumenta, la calidad de vida se ve comprometida pero que si dicha varianza disminuye, la calidad de vida definitivamente aumenta. Demuestre que la varianza debe disminuir. b) Antes de hacer la autopista, el tiempo de llegada de Juan a su trabajo tenía una varianza de 9.2 medida sobre 23 eventos tomados al azar. Después de que la abrieron, la varianza sobre 12 eventos fue de 3.1. Hallar el intervalo de confianza del 95 % para

la relación entre varianzas poblacionales antes y después de abrir la autopista c) ¿Es apropiado decir que la varianza grande es 5 veces la pequeña? d) ¿Es recomendable que Juan asuma, en el momento de tomar decisiones, que la varianza disminuyó a la quinta parte?

2. El tic-tac del corazón de Juan tiene una media por minuto de 74 y una varianza de 140, medida sobre 12 minutos. Por supuesto, eso es cuando no ve a Maritza. Porque cuando la ve, ¡mejor dicho!, cuando la ve el corazón se le acelera a una media de 112, y la varianza sobre los siguientes 5 minutos le llega a 180. a) ¿Puede decirse que la varianza poblacional se duplica? b) Puede decirse que la media poblacional se duplica? c) ¿Está enfermo Juan?

3. El Alcalde, siendo excepcionalmente intachable, cree que la única propaganda que debe hacer son sus buenas obras. Pero un asesor le pide que le eche una mirada a la siguiente tabla en la que se registra la cantidad de propaganda pagada y la aceptación del público para los proyectos del alcalde anterior.

Observado: Propaganda y aceptación			
Propaganda	Aceptación		
	X= Baja	Y=Alta	Totales
A=Poca	12	14	
B=Mucha	23	33	
Totales			

Si se toma una persona al azar, halle la probabilidad de que dicha persona sea de

a) A

b) X

c) A intersección X = $A \cap X$

d) A intersección X si fuesen independientes

e) ¿Puede decirse que los eventos A y X son independientes?

f) A o X (unión) = $A \cup X$

g) A dado X = A/X. Probabilidad de A dado que el universo es X

h) Verifique que $p(A/X) = p(A \cap X)/p(X)$

i) ¿Son los factores cantidad de propaganda y tipo de aceptación independientes?

j) Es la proproción de alta aceptación homogenea con respecto a la cantidad de propaganda?

k) ¿Le aconsejaría Ud. al Alcalde que invierta en propaganda?

271 ***Versión 3***

1. Hay dos partidos dominantes desde antaño. El número de curules en una muestra al azar de 7 años por cada partido ha sido (23, 11), (22,13),(24,15), (23,11), (25,17), (24,15), (27,13). ¿Hay razón para decir que el partido vencedor dobla en promedio a su rival?

2. Según el Gran Maestro Shogun, la mente es capaz de modificar el cuerpo para restaurar la salud. Patty no está enferma de nada pero le gusta esa teoría pues tiene la esperanza de que la misma técnica le ayude a controlar sus desordenados sentimientos que nunca hacen lo que se les manda. Ella hace sus ejercicios de control mental el 50 % de las veces. Uno de dichos ejercicios es: cierre los ojos, párese en un solo pie y resista. Cuando Patty hace sus ejercicios, puede controlar sus sentimientos el 30 % de las veces, pero cuando no los hace puede hacerlo el 12 % de las veces. Hoy se presentó en la vida de Patty un evento importante y salió exitosa controlando los sentimientos a la perfección. ¿Cuál es la probabilidad de que Patty haya hecho hoy sus ejercicios? ¿Los hizo o no los hizo?

3. De los estudiantes que aseguran que rinden mejor por la mañana se tomó una muestra al azar de 17 y 12 de ellos dijeron que se sentían mal en las clase después de las 2 de la tarde. Pero de los que dicen rendir mejor por la tarde se tomó una muestra al azar de 22 personas y de ellas 17 dijeron que les iba bien en clases después de las 2 de la tarde. a) ¿Puede decirse que la proporción de estudiantes a quienes les va mal por la tarde es 0.1 mayor en los del primer grupo que en el segundo? b) Halle el IC para la diferencia de proporciones. c) ¿Puede decirse que la diferencia de proporciones es 0.7?

272 ***Versión 4***

1. Mientras que la creencia popular dice que el período de las mujeres sucede como por reloj cada 28 días, una muestra al azar de 23 jovencitas dio un promedio de 27 días y una varianza de 14.2. a) ¿Se justifica la creencia popular en cuanto a la media para las jovencitas? b) ¿Se justifica la creencia popular en cuanto a la varianza para las jovencitas? (atención: la creencia popular dice dos cosas, que la media vale 28 y que la varianza vale cero pues la media es lo único que sucede y lo hace como si fuese guiada por un reloj bien calibrado. Habría varianza no cero si se observase retrazos y/o adelantos con respecto a los 28 días). Pero por otra parte, una muestra al azar de 15 señoras dio una media de 28.1 días y una varianza de 6 (días cuadrado). c) ¿Se justifica la creencia popular para las señoras)? (también incluye dos hipótesis nulas.) d) Hallar el intervalo de confianza para la relación entre varianzas. e) ¿Se justifica decir que las señoras disminuyen la irregularidad medida por la varianza a la mitad con respecto a las señoritas? f) ¿Se justifica decir que las señoras disminuyen la irregularidad a menos de la mitad con respecto a las jovencitas?

2. Una baraja consta de 52 cartas, divididas en cuatro palos que son corazones, diamantes, tréboles y picas. Los corazones y diamantes son rojos, los tréboles y picas son negros. Cada uno de los palos consta de 13 cartas: As, 2, . . ., 10, J, Q y K. Se eligen al azar dos cartas **sin reemplazo = se saca la primera carta, pero no se regresa a la mesa, y después se saca la segunda**. Halle:

(a) La probabilidad de que las 2 cartas tengan diferente color. Ayuda: haga el árbol cuyas opciones son color rojo vs. color negro.

(b) La probabilidad de que haya máximo una J. Ayuda: considere el árbol cuyas opciones son J vs. no-J

3. Para contrarrestar la acendrada opinión de que la carrera armamentista es perjudicial para el pueblo, hecho que parece verdad para los regímenes comunistas, un analista exhibe la siguiente tabla en la que se muestra el porcentaje del presupuesto del gobierno de países democráticos que se dedica a las fuerzas militares y la calidad de vida del ciudadano promedio. a) Demuestre que los factores son dependientes. b) Observando la tabla decida si la guerra está a favor o en contra de la calidad de vida. c) Proponga una explicación a sus resultados.

Observado: Presupuesto invertido y calidad de vida				
Presupuesto	Calidad de vida			
	X= Pobre	Y=Mediana	Z=Buena	Totales
A=< 3 %	23	43	4	
B=3-5 %	29	35	8	
C=> 5 %	2	3	5	
Totales				

273 ***Versión 5***

1. La distribución de notas en el parcial uno se describe en la siguiente tabla.

Notas en el parcial 1	
Intervalos	**No. acumulado de estudiantes**
(0, 1]	2
(0, 2]	7
(0, 3]	22
(0, 4]	32
(0, 5]	34

Para el segundo parcial hubo un quiz por semana y como consecuencia la media fue de 3.1 con una desviación de 0.4. a) Tomando el punto medio de cada intervalo como marca de clase, halle las frecuencias absolutas, calcule la media y la desviación. b) Halle el IC para la relación entre varianzas y decida si puede decirse o no que las varianzas son iguales. c) ¿Puede decirse que con la metodología del quiz por semana la media aumenta un 10 %? d) ¿Puede decirse que la varianza disminuye un 25 %? Obsérvese que los datos se refieren a censos, pues se considera que están todos los datos de toda la clase, sin embargo, la pregunta se refiere a la extrapolación de los resultados más allá de la clase, por lo que deben tomarse como muestras cuyo estudio se hace con las metodologías que estamos estudiando.

2. El teniente Rivas sale a hacer su ronda contra los violentos. Según las estadísticas, un teniente muere el 5 % de las veces que se combate. Pero además de combates hay escaramuzas, donde no se actúa a matar sino sólo para advertir al enemigo de que hay alguien en guardia. La proporción de acciones que son escaramuzas es del 85 %. Hoy hubo acción y el teniente no murió. Todos en el cuartel dicen que fue otra escaramuza más. ¿Será cierto? Ayuda: lea el texto para entenderlo, haga un árbol con dos pisos y decida qué va en el primero y qué en el segundo, si morir o no morir o bien combate o escaramuza; tenga presente que las probabilidades que salen de cada nodo deben sumar uno.

3. La tabla dada relaciona los datos de una muestra al azar encaminada a dilucidar si la proporción de personas físicamente fuertes depende o no del color de la piel.

Observado: Fortaleza y color de la piel				
Fortaleza	Color de la piel			
	X= Blanca	Y= Trigueña	Z=negra	Totales
A= Fuerte	12	15	18	
B= Débil	21	23	9	
Totales				

a) Decir si la proporción de personas fuertes es homogénea con respecto al color de la piel.

b) Hallar el IC para la proporción de A dado X.

c) ¿Puede decirse que la proporción de A dado Z es 10 % mayor que la de A dado X?

274 ***Versión 6***

1) Los hombres y las mujeres de una prestigiosa escuela de abogacía compiten por la supremacía. La tradición es que en el Examen Nacional de Calidad las mujeres tengan un promedio de 80 con una desviación de 11 y que los hombres tengan un puntaje promedio de 77 con una desviación de 6 (para no quedar mal, los hombres dicen que el examen no evalúa ni el pensamiento crítico ni la capacidad recursiva, sino sólo los conocimientos). En el 2001 una muestra al azar de 6 mujeres un promedio de 77 con una desviación de 8. Por otro lado, una muestra aleatoria de 7 hombres obtuvieron un puntaje promedio de 78 con una desviación de 4 a) ¿Hubo cambio en la media de las mujeres entre la tradición y el 2001? ¿Y de los hombres? b) Hubo cambio de la diferencia de medias entre géneros entre la tradición y el 2001? c) En el año que acaba de pasar, el promedio para 8 hombres fue de 81 con una desviación de 9 y el de 7 mujeres fue de 79 con una desviación de 8. ¿Hubo cambio en la media de las mujeres entre el 2001 y el año pasado? ¿Y de hombres? d) ¿Hubo cambio de la diferencia de medias entre el 2001 y el año pasado? e) Los hombres andan diciendo que de acuerdo a los datos del año que acaba de pasar, ellos son 10 puntos mejores que las mujeres. Demuestre que eso no se puede contradecir. f) Las mujeres respondieron: los mismo datos dicen que somos 5 puntos mejores que los hombres: ¿Es cierto? g) Reconcilie tanta pelea por medio del IC de la diferencia de medias.

275 ***Super 5***

1. Una baraja consta de 52 cartas, divididas en cuatro palos que son corazones, diamantes, tréboles y picas. Los corazones y diamantes son rojos, los tréboles y picas son negros. Cada uno de los palos consta de 13 cartas: As, 2, . . ., 10, J, Q y K. Se eligen al azar dos cartas **con reemplazo = carta sacada se devuelve y se baraja**. Halle la probabilidad de

a) Al menos un 3.

b) Máximo un 5.

c) Exactamente un 3 o exactamente un 5.

d) Con respecto a la pregunta en a): si la segunda carta no fue un 3, ¿cuál es la probabilidad de que la primera tampoco haya sido un 3?

2. De una muestra de 35 personas tomadas al azar antes de las elecciones, 17 se consideran a favor de un candidato, el cual perdió en las elecciones. Después de lo cual, de una muestra al azar de 40 personas, 10 se consideran fieles.

   a) Halle el IC para la diferencia de proporciones.

   b) ¿Puede decirse que la proporción de fieles cambió?

   c) ¿Puede decirse que la proporción de fieles disminuyó a la mitad?

3. Si en una muestra al azar de tamaño 43 se obtuvo una varianza de 12, hallar el IC de la varianza poblacional.

4. Sea $\bar{X}_{13} = 28$ con $s_X = 6$ y $\bar{Y}_{17} = 14$ con $s_Y = 2$.

   a) Halle el IC para la varianza de X.

   b) Halle el IC para la diferencia entre medias.

   c) Halle el IC para la relación entre varianzas.

   d) ¿Puede decirse que la media mayor es 5 veces la menor?

   e) ¿Puede decirse que la varianza mayor es 4 veces la menor?

   f) Puede decirse que la varianza de X vale 3?

5. Se trata de averiguar el efecto de la asistencia a fiestas como mecanismo para calmar los nervios y la calidad de las notas de los estudiantes:

Observado: Asistencia a fiestas y calidad de las notas				
Asistencia a fiestas	Calidad de las notas			
	X:Baja	Y: Mediana	Z:Alta	Totales
A: Baja	4	5	6	
B: Mediana	3	4	7	
C: Alta	3	2	1	
Totales				

   Si se toma una persona al azar, halle la probabilidad de que dicha persona sea de

   a) B

   b) Z

   c) B y Z

   d) B o Z (unión)

   e) B dado Z

   f) ¿Puede decirse que los eventos B y Z son independientes?

   g) ¿Cómo evalúa Ud el efecto de la asistencia a fiestas y la calidad de la nota (Ayuda: estudie primero la dependencia/independencia de los dos factores con el formalismo de las tablas de contingencia y luego guíese por la tabla).

Capítulo 14

Talleres complementarios

276 *Objetivo: aprender lo básico del manejo de una hoja de cálculo y lo mínimo sobre distribuciones que no se pueden aproximar por una normal. Hay 4 talleres, uno por parcial.*

14.1. Taller 1: Excel pa'ya

Aprenderemos lo básico de una hoja de cálculo. Este taller se puede hacer en Excel, Open Office o en Gnumeric. Es preciso que se inserten las funciones de Excel que resuelven los problemas automáticamente. Tiene dos partes, la primera se resuelve sólo lo que no se sepa. La segunda se resuelve toda en grupo sobre el archivo correspondiente y se manda al correo del profesor.

277 ***Parte 1: Excel básico.***

La aplicación más inmediata de Excel es hacer los trabajos de una calculadora y emplear gráficas.

Ejercicio 1: calculadora

1. Abra Excel y oblíguelo a desplegar una hoja rectangulada. Cada rectángulo es una casilla.

2. En cualquier casilla escriba 5 y enter. En la casilla del lado derecho escriba un 6 y enter.

3. En la siguiente casilla del lado derecho escriba un =. El igual es un símbolo que le indica a Excel que lo que sigue no es un número sino una instrucción para hacer cálculos. Después de haber escrito el igual, haga click en la casilla del 5. Aparecerá el nombre de la casilla, una letra de la columna y un número de la fila. Ahora escriba un + y a continuación haga click en la casilla del 6. Acabe con enter. Deberá aparecer 11, que es la suma de 5 y 6.

4. Juegue con diversos números y operaciones: la suma es +, la resta es - (guión corto), la multiplicación es *, la división es /.

5. Cualquier equivocación puede deshacerse oprimiendo al tiempo en el teclado las teclas Ctrl y Z.

Ejercicio 2: sumar pares de números de dos columnas

1. En una columna escriba los números 1, 2, 3. En la columna siguiente y al mismo nivel escriba los números 2, 4, 6. (el 3 y el 6 quedan en el último renglón).

2. Ahora sume los números del primer renglón, el 1 con el 2, debe dar 3. (Escriba el igual, haga click en el 1, escriba el +, haga click en el 2, termine con enter).

3. Para sumar los demás pares de números, no debe repetir el procedimiento anterior sino clonar la instrucción de la suma. Para hacerlo, haga click en la casilla que quiere clonar, aquella donde puso el igual y que ahora debe estar convertido en un 3. Arrastre el cursor al extremo inferior derecho de dicha casilla. El cursor se transformará en una cruz. Haga click allí mismo y sin soltar el ratón arrastre la cruz por toda la tercera columna hasta igualar todos los renglones que quiere sumar. Al soltar el ratón aparecerá el resultado de haber clonado la fórmula de la suma: aparecerá una columna con los números 3, 6, 9.

Ejercicio 3:Sume los primeros diez números.

1. Escriba los números de 1 a 10 en una columna.

2. Busque el ícono de autosuma

3. Para sumar automáticamente: se negrean los datos y se aplica una opción que se llama autosuma y que está en un ícono con una sumatoria. Para ejecutar la autosuma: Arrastrando el ratón con el botón izquierdo oprimido, se negrea o se señala o se pinta de azulito o de gris todas las casillas que quieren sumarse y se hace click en la opción suma del ícono de autosuma. La respuesta es 55.

4. Si no existe el ícono de autosuma, se procede así: se señala la casilla donde uno quiera la respuesta y se escribe Suma. Se señala la casilla del lado. Se busca el menu insert -¿function -¿se busca sum -¿se explora hasta que pregunte el rango de elementos a sumar -¿se hace click en el campo donde deben ir las casillas -¿se negrean las casillas de Excel que uno quiere sumar. -¿se acepta haciendo OK.

Ejercicio 4: Operaciones con los primeros diez números al cuadrado.

1. Escriba los diez primeros números, programe con el igual el cuadrado de 1: signo igual, click en la casilla del uno, estrella, click en la casilla del uno. Clone la fórmula. Use autosuma para hallar la suma. La suma es 385.

2. Para restar 5 de cada uno de los diez primeros números: Escribra los diez primeros números en una columna. En la siguiente columna, al lado del 1 escriba un 5. Clone el 5 para todos los renglones. Reste las dos columnas: en una tercera columna escriba y clone la fórmula de la resta. Ahora sume los elementos de la tercera columna. Da 5 como respuesta.

Ejercicio 5: Dibuje el histograma de frecuencias absolutas para

Tabla de Frecuencias absolutas	
Estatura	Frec. absoluta
Baja	3
Media	1
Alta	1

Para dibujar el histograma se hace click sobre un ícono que tenga la forma de histograma. Si no lo hay, sobre el que tenga la forma de torta y ahí se busca el ícono de histograma.

Ejercicio 6: Dibuje una torta o pie-chart para los siguientes datos:

Tabla de Frecuencias absolutas de votos	
Partido	Votos
A	2
B	6
C	3

Para dibujar la torta de frecuencias relativas se negrean los datos de la tabla de frecuencias absolutas y se llama un ícono que represente una torta. Es posible que de un histograma, en ese caso también aparecerá un menú gráfico con muchas opciones, una de ellas la torta. Se llama y listo.

Ejercicio 7: La parábola de x al cuadrado

Teclee los números del 0 al 10. Halle sus cuadrados. Busque un ícono que represente un histograma o una torta y juegue con las opciones hasta que vea la parábola de los cuadrados.

Ejercicio 8: Dibujar una función arbitraria

Dibuje la siguiente tabla que indica las horas de estudio de 4 estudiantes y su nota en un examen. Excel no acepta el punto decimal: debe cambiarse por una coma.

Horas de estudio y nota	
Horas de estudio	Nota
0	3
2	3.5
3	3.8
5	2.5

Use dos columnas para teclear los datos. Negree las dos columnas, llame el ícono de gráficas estadísticas, busque y aplique uno que se llame scatter o dispersión.

Ejercicio 9: Ordenar unos datos.

Escriba unos datos inventados y en desorden en una columna. Para ordenarlos: Se negrean los datos, y se aplica insert + función + sort (ordenar). Para recuperar los datos: ctrl + z.

Los datos se pueden ordenar por criterios alfanuméricos, como por ejemplo los nombres de una persona empezando con A y hasta la z.

Ejercicio 10: Contar unos datos.

Escriba unos datos inventados. Para contarlos: Se hace click en una casilla donde uno quiere la respuesta, se marca escribiendo conteo, se hace click en la del lado, se negrean los datos, y se aplica insert + función + count (contar). Para recuperar los datos: ctrl + z. Es posible que exista un ícono para ordenar: es una flechita hacia abajo o hacia arriba.

Ejercicio 11: Hallar la tabla de frecuencia absoluta de unos datos y dibujarla.

Se ordenan los datos. En una columna se escriben los rangos. Por ejemplo, 1 a 5, dice que uno considera los números entre 1 y 5. En la casilla del lado se hace click. Sobre la columna de datos se negrean los datos que corresponden al rango establecido y se cuentan. Luego con el rango 2 y así hasta terminar. Después de tener la tabla, uno puede dibujarla con scatter.

Ejercicio 12 (opcional) : Para una lista de datos, dibujar automáticamente el histograma por intervalos.

Excel ofrece la posibilidad de hacer todo en un sólo click. Ver:
http://support.microsoft.com/kb/214269/es

278 *Parte 2: Censos*

Un **censo** es una colección de datos sobre una variable que se mide sobre todos y cada uno de los elementos de la población. En contraposición, una muestra se refiere a una quizá mínima parte de la población. En un censo, lo que da es lo que es, sin incertidumbre. En una muestra, lo que da es una aproximación a lo que es, hay incertidumbre.

En este taller, estudiamos las notas reales de cursos de estadística pasados. Como están todos los alumnos, se trata de CENSOS. Se ha quitado el nombre y la dirección electrónica de cada quien. Las preguntas están al final.

Hay dos cursos A y B. El orden de la notas es el siguiente:

Notas de 4 parciales, quices y tablero, proyecto, 75 %, Examen Final, 25 %, 100 % y definitiva.

Las notas del curso A y las del B aparecen en sus tablas respectivas. Los porcentajes de B son algo diferentes. Trate de conseguir el archivo que contiene los datos o sino, teclee sólo los necesarios.

Notas del Curso A												
Nbre	Prog	1P	2P	3P	4P	QT	Prj	'75 %	EF	25 %	100 %	DEF
A1	CPOL	5.0	4.4	5.0	3.0	2.7	3.0	2.9	3.8	1.0	3.86	4.0
A2	DERE	3.5	4.4	5.0	4.4	3.7	4.0	3.1	3.6	0.9	4.03	4.5
A3	HIST	4.0	3.4	5.0	3.0	4.1	4.0	2.9	3.4	0.9	3.78	4.0
A4	ANTR	4.1	2.3	5.0	4.0	4.1	3.8	2.9	2.8	0.7	3.61	3.5
A5	DERE	4.7	4.5	2.6	3.5	3.2	3.0	2.7	1.9	0.5	3.18	3.0
A6	DERE	5.0	1.0	5.0	3.2	3.9	4.0	2.7	4.5	1.1	3.87	4.0
A7	DERE	3.5	1.2	2.6	3.0	4.3	3.0	2.2	2.3	0.6	2.77	3.0
A8	DERE	4.5	0.4			0.9	0.0	0.8		0.0	0.75	1.0
A9	ANTR	1.7	4.0	2.6	1.7	3.8	3.8	2.2	2.5	0.6	2.80	3.0
A10	ANTR	4.5	3.1	4.1	3.8	4.1	4.0	2.9	2.7	0.7	3.62	3.5
A11	DERE	4.0	4.6	3.2	3.5	3.9	4.0	2.9	2.8	0.7	3.59	3.5
A12	CPOL	2.8	4.0	3.7	4.3	3.9	3.0	2.7	1.2	0.3	3.03	3.0
A13	DERE	5.0	4.2	5.0	4.0	4.0	4.0	3.3	4.5	1.1	4.41	4.5
A14	EGIN	5.0	4.1	5.0	5.0	3.6	4.0	3.4	4.8	1.2	4.55	5.0
A15	CPOL	5.0	5.0	4.0	3.0	3.9	4.0	3.1	2.5	0.6	3.74	4.0
A16	DISE	4.5	3.5	3.0	2.8	3.2	3.0	2.5	3.3	0.8	3.34	3.5
A17	DERE	3.8		0.4	0.0	0.5	0.0	0.6	1.8	0.5	1.06	1.5
A18	ANTR	4.0	2.4	3.8	4.0	4.0	4.0	2.8	1.9	0.5	3.24	3.5
A19	CPOL	3.6	2.5	2.7	3.2	3.1	3.8	2.3	2.7	0.7	3.02	3.0
A20	ANTR	5.0	5.0	5.0	3.3	3.9	4.0	3.3	4.0	1.0	4.29	4.5
A21	DERE	5.0	3.5	2.1	4.2	3.9	3.0	2.7	3.2	0.8	3.52	3.5
A22	ANTR	3.2	4.5	3.0	3.4	4.2	4.0	2.8	2.3	0.6	3.36	3.5
A23	CPOL	5.0	2.6	3.6	2.3	2.4	3.8	2.4	1.5	0.4	2.82	3.0
A24	ANTR	3.0	3.1	2.3	3.2	4.0	3.8	2.4	2.9	0.7	3.13	3.0
A25	DERE		0.5			0.4	0.0	0.1		0.0	0.11	0.0
A26	FILO	5.0	5.0	5.0	4.0	3.8	4.0	3.4	4.0	1.0	4.36	4.5
A27	CPOL	5.0	3.0	2.5	3.7	3.0	3.0	2.5	2.4	0.6	3.13	3.0
A28	CPOL	3.7	4.1	2.1	3.6	2.0	3.0	2.3	2.9	0.7	3.04	3.0
A29	CPOL	4.8	5.0	4.7	4.3	5.0	4.0	3.5	4.4	1.1	4.59	5.0
A30	DERE	5.0	3.1	1.7	4.1	4.0	4.0	2.7	3.3	0.8	3.56	3.5
A31	CPOL	5.0	5.0	5.0	3.8	3.0	3.0	3.1	2.8	0.7	3.83	4.0
A32	CPOL	4.0	4.1	1.0	1.0	2.0	3.0	1.9	1.6	0.4	2.27	2.5
A33	DERE	4.7	5.0	5.0	4.0	3.5	3.8	3.3	3.1	0.8	4.04	4.0
A34	CPOL	5.0	3.8	2.0	2.0	2.3	3.8	2.3	3.4	0.9	3.20	3.0
A35	CPOL	2.2	2.2	2.2	1.5	0.3	0.0	1.1	1.5	0.4	1.47	1.5
A36	CPOL	3.8	2.7	1.0	0.5	1.9	3.0	1.6	1.0	0.3	1.83	2.0
A37	DERE	4.7	5.0	5.0	4.1	4.0	4.0	3.4	2.9	0.7	4.09	4.0

Notas del Curso B												
Nombre	Prog	1P	2P	3P	4P	QT	Proy	78 %	EFnl	22 %	100 %	DEF
B1	DERE	2.2	2.2			1.0		0.70		0.00	0.70	0.50
B2	DERE	4.7	3.8	4.7	4.8	4.2	4.0	3.41	1.00	0.22	3.63	3.50
B3	DERE	2.5	4.0	3.3	3.6	3.8	4.0	2.76	2.30	0.51	3.26	3.50
B4	DERE	4.0	4.3	5.0	3.0	3.5	4.0	3.09	1.10	0.24	3.34	3.50
B5	CPOL	2.7	2.5	4.2	2.0	2.0	1.0	1.87	2.10	0.46	2.33	2.50
B6	DERE	3.0	3.8			2.0		1.14		0.00	1.14	1.00
B7	LENG	0.0				0.3		0.03		0.00	0.03	0.00
B8	DERE	3.9	3.8	1.6	2.3	4.8	4.0	2.65	1.70	0.37	3.02	3.00
B9	DERE	2.5	3.1	2.6	2.2	3.8	4.1	2.38	4.00	0.88	3.26	3.50
B10	DERE	1.1	1.9	1.3		2.1		0.83		0.00	0.83	1.00
B11	CPOL	2.4	3.1	1.1		1.9		1.10		0.00	1.10	1.00
B12	DERE	3.0	4.7	3.0	2.5	2.8	4.0	2.60	2.80	0.62	3.22	3.00
B13	DERE	3.9	3.4	5.0	1.5	3.0	4.0	2.70	1.20	0.26	2.96	3.00
B14	ANTR	3.8	4.5	4.3	4.8	4.4	3.7	3.32	4.40	0.97	4.28	4.50
B15	DERE	4.7	4.6	4.2	4.0	4.2	5.0	3.46	3.50	0.77	4.23	4.00
B16	DERE	1.2	3.2	1.3		0.9		0.85		0.00	0.85	1.00
B17	DERE	2.6	3.9	2.7	2.5	3.2	3.7	2.42	0.70	0.15	2.57	2.50
B18	LENG	4.8	5.0	4.0	3.0	2.5	4.0	3.02	3.00	0.66	3.68	3.50
B19	DERE	3.3	3.5	3.0	2.3	2.1	5.0	2.49	4.70	1.03	3.53	3.50
B20	CPOL	0.0				0.1		0.01		0.00	0.01	0.00
B21	DERE	3.5	4.2	2.8	4.2	3.3	3.7	2.82	2.70	0.59	3.41	3.50
B22	DERE	3.0	4.0	3.9	4.7	4.8	4.0	3.17	2.00	0.44	3.61	3.50
B23	DERE	3.3				1.3		0.60		0.00	0.60	0.50
B24	DERE	4.6	4.1	4.6	4.2	4.1	4.0	3.33	2.70	0.59	3.93	4.00
B25	DERE	2.8	3.7	2.5	2.5	3.2	4.0	2.42	1.00	0.22	2.64	2.50
B26	DERE	1.9				0.3		0.29		0.00	0.29	0.50
B27	LENG	3.4	2.6	5.0	3.6	3.7	3.0	2.77	2.50	0.55	3.32	3.50
B28	DERE	2.2	2.3			1.2		0.74		0.00	0.74	0.50
B29	DERE	2.7	3.1	2.5	1.8	4.0	3.0	2.23	2.10	0.46	2.69	2.50
B30	DERE	4.2	5.0	5.0	5.0	5.0	4.6	3.74	4.60	1.01	4.76	5.00
B31	DERE	5.0	5.0	4.0	4.8	4.7	3.7	3.53	3.70	0.81	4.34	4.50
B32	LENG	5.0	5.0	5.0	5.0	3.9	3.0	3.49	3.50	0.77	4.26	4.50
B33	DERE	2.7	3.3	4.5	2.5	3.4	4.0	2.65	1.80	0.40	3.04	3.00
B34	DERE	4.4	3.9	2.0	4.5	4.1	5.0	3.10	3.00	0.66	3.76	4.00
B35	DERE	2.3						0.30		0.00	0.30	0.50
B36	DERE	5.0	4.6	5.0	3.5	3.9	5.0	3.50	3.50	0.77	4.27	4.50
B37	DERE	3.9	3.9	4.0	2.0	2.4	3.0	2.49	0.20	0.04	2.53	2.50
B38	DERE	2.6	3.8	1.0		1.9		1.20		0.00	1.20	1.00
B39	DERE	3.4	3.6	2.6	3.3	4.1	3.7	2.69	1.00	0.22	2.91	3.00
B40	DERE	3.3	4.3	4.4	4.3	4.2	3.7	3.15	3.20	0.70	3.85	4.00
B41	DERE	3.3	4.3	4.5	4.0	2.9	4.0	2.99	2.10	0.46	3.45	3.50
B42	DERE	2.6	3.5	1.0		1.9		1.17		0.00	1.17	1.00
B43	DERE	2.2	2.8	2.1		2.2		1.21		0.00	1.21	1.00
B44	LENG	0.0				0.0		0.00		0.00	0.00	0.00
B45	DERE	3.0	3.4	3.5	3.6	3.4	4.0	2.72	1.20	0.26	2.98	3.00
B46	DERE	3.8	3.6	4.0	1.1	2.7	4.0	2.50	2.60	0.57	3.07	3.00
B47	DERE	2.4	3.8	3.0	2.0	2.4	3.7	2.24	1.80	0.40	2.64	2.50

279 ***Ejercicios.*** *Respuests en la pág 324*

Las casillas vacías nunca se tienen en cuenta:

1. ¿A cuál curso le fue mejor en promedio en el segundo parcial?
2. ¿Cual curso tiene mayor varianza para el cuarto parcial?
3. ¿A quienes les fue mejor en promedio en el examen final en el curso A, a los abogados o a los de ciencia política?
4. Hacer una tabla de frecuencias absolutas de las notas del segundo parcial del curso A agrupando los datos por intervalos y dibujar el correspondiente histograma o diagrama de barras. ¿Es la distribución resultante de tendencia central?
5. Hacer una torta de la frecuencia relativa de los que pasaron el final del curso B.
6. Comparar las frecuencias relativas de los que pasaron el final en los dos cursos: ¿Cuál fue mayor?
7. Sustente o refute el siguiente análisis en referencia al curso B, el cual es el más reciente: de todos los estudiantes que comienzan el semestre hay algunos que se conectan al curso desde el puro principio y hay otros que no. La probabilidad de que se retiren o que pierdan el curso es mucho mayor para los que no pudieron conectarse desde un principio que para los conectados. Por otra parte, hay alumnos que sin ser de lo mejor no cesan de esforzarse y llegan a sacar notas excelentes, lo cual demuestra que ser el mejor es cuestión de ganas y no de ser un genio en matemáticas. Respecto del examen final, se percibe como muy difícil por muchos estudiantes, tanto que sacar 1.5 es difícil para algunos y por eso pierden la materia. Combinando todo junto, recomedamos : 1) hacer algo para que la probabilidad de conexión al curso aumente desde el primer día de clase, 2) dar incentivos para que los alumnos que no son fuertes en matemáticas se animen a luchar por un permanente excelente desempeño, 3) hacer un repaso fuerte al terminar el semestre para que el final deje de ser una amenaza y que el reto no sea sacar 3 sino 5.

280 ***Desafío*** *Siga el ejemplo 35 de la pág 22 y diseñe para el curso B una tabla como la allí mostrada e investigue de qué forma es útil para aconsejar si uno debe retirarse o no de la materia para no perderla y correr el riesgo de que lo echen de la Universidad. Ayudas:*

1. Calcule el 45 % correspondiente a los 3 primeros parciales.
2. Halle la nota equivalente sobre 5 como si el 45 % fuese todo el curso, es decir, el 100 %.
3. Al lado, ponga una columna que contenga la información de si pasó o no pasó.
4. Ordene los datos según el criterio descendente de la nota equivalente.
5. Particione a los estudiantes en dos o tres clases donde sea muy evidente la decisión a tomar, quedarse o retirarse.
6. Haga la tabla, como la del ejemplo, que muestre la incidencia de la nota equivalente con el resultado final, pasar o no pasar.
7. Formule explícitamente su consejo.

AYUDAS GENERALES

1. Conserve limpio el documento original:
Para ello, haga una copia del material en una nueva planilla.
Las planillas están numeradas y sus pestañas aparecen en la parte abajo izquierda de la ventana. Si uno desea una nueva planilla, se usa el menú insertar hoja o planilla o sheet. Uno puede copiar una columna aparte para poderla trabajar con más claridad:

Para copiar: se señala la parte que se quiere copiar, se teclea control + C, luego se hace click en el lugar donde uno quiere pegar lo copiado y después control + v.

Para señalar una región: Se hace click en la primera casilla de la región a señalar, se arrastra el cursor con el botón oprimido, se suelta el botón en el final de la región a copiar.

2. Para llamar una función como

promedio (average), o varianza (var), o count, para contar el número de elementos: Primero se pone en una celda un título o etiqueta que indique lo que se está calculando. Después se hace click en la casilla del lado,

para que allí se escriba el resultado de la función.

En el menú insertar se busca funciones o fórmulas

y después se localiza la función.

Cuando una ventana de diálogo pida datos,

se señala la región que contiene los datos y se hace click en aceptar.

Ejemplo: el promedio del primer parcial del curso A fue 4.27.

Su desviación fue 0.87.

Atención: asegúrese de tener en claro qué hacer con las casillas en blanco. Si desea tomarlas como cero, por favor, escriba el cero. Pero si desea tomarlas como un vacío que no debe ser tenido en cuenta, busque una opción de función que exprese explícitamente dicha deseo. O bien, copie aparte la columna, y elimine la casilla vacía: haga click con el botón derecho, elija eliminar y corra las casillas hacia arriba. Luego llame la función a calcular.

3. Para ordenar por carrera: se señala toda la región que contiene tanto todos los datos del curso como los títulos sobre las columnas. Se copia aparte. Se señala, y se llama el menú datos y luego ordenar.

Se le da como criterio de ordanamiento Programa.

Se acepta. Ej: en el curso A hay 7 antropólogos.

4. Un error puede deshacerse con control + z.

Atención: Las técnicas que hemos aprendido son quizá las más fáciles de ejecutar y de recordar. Pero si uno llegase a usar una hoja de cálculo por rutina, entonces es necesario tener en cuenta que existen formas profesionales para hacer tareas complejas. Tomemos como ejemplo hacer una tabla de frecuencias a bsolutas y su histograma o diagrama de barras. La forma experta de hacerlo en *Gnumeric* es como sigue: se teclean los datos uno por uno sobre una columna y sin dejar vacíos. Al lado se teclea una columna que contiene la lista de los cortes de los intervalos que definen los rangos para clasificar los datos. Por ejemplo, si los datos son números que representan las notas de un examen, la columna de datos contiene las calificaciones y la columna de rangos contiene los números 0,1,2,3,4,5, un número por celda. Luego se llama el menú *Tools* seguido de *Statistical analysis*, luego *Frequency tables* y de nuevo *frequency tables*. Aparece una ventana de diálogo, y sobre la casilla de diálogo *Inputrange*, uno señala con el cursor la región de la columna que contiene los datos. Después uno señala la tarjeta *Categories* y elije *Category ranges*, luego uno señala la segunda columna, aquella que contiene los bordes de los rangos. Pasa uno a la tarjeta *Graphs and options* y elije *Column chart*. Se termina con *OK*. *Gnumeric* despliega una hoja donde uno encuentra la tabla de frecuencias y el histograma.

14.2. Taller 2: Estadística no paramétrica y el teorema de Chevichev

Objetivo: Nosotros hemos visto cómo atacar los problemas estadísticos cuando se tiene una distribución de tendencia central que se caricaturiza como una normal. ¿Qué pasa con aquellas distribuciones que difieren mucho de una normal por tener varias jorobas o ser muy irregulares? Para responder dicha pregunta se ha creado la **estadística no paramétrica**, en contraposición a la que conocemos y que se llama **estadística paramétrica** y en la que todo depende del valor de **parámetros**, valores ideales, poblacionales, como la media y la varianza. Nosotros ya hemos trabajado con estadística no paramétrica en TLC pues es general y vale no sólo para distribuciones normales. Veremos también que el test de ajuste chi-cuadrado de una distribución es de hecho no paramétrico pues se aplica a cualquier tipo de datos categóricos, que también pueden ser ordinales (que se pueden ordenar) o cuantitativos (que se pueden promediar).

Recordemos en primer término una herramienta no paramétrica que es muy útil y muy popular: el boxplot o diagrama de caja que nos presenta de forma gráfica un resumen o síntesis de los datos y que nos permite definir los eventos extremos por el criterio del rango intercuartilar. Por otro lado, el TLC nos dice que la distribución de la media muestral tiene una distribución normal, lo cual es exacto para variables aleatorias con distribución normal y aproximadamente exacto para casi todo tipo de distribuciones, por lo cual también es no paramétrico.

En segundo lugar, presentamos el primer gran resultado teórico de la *estadística no paramétrica*, a saber, el *teorema de Chevichev*, el cual es válido para todas las distribuciones, sin importar si son normales o no. Probamos de forma entendible dicho teorema y mostramos cómo se aplica. Para ello, tenemos que ver cómo se usan las integrales para calcular la media y la varianza.

Ejercicio 1: Hallar el resumen de 5 valores para los datos del primer parcial del curso A. Calcular el rango intercuartilar y usarlo para determinar los valores extremos. Dibujar el *boxplot* con sus bigotes.

Ejercicio 2: Sobre la distribución z, la normal estándar:

1. Halle los cuartiles uno y tres.
2. Halle el rango intercuartilar y utilícelo para determinar los valores extremos.
3. Diga si está de acuerdo o no con la siguiente proposición: el criterio del rango intercuartilar para la determinación de eventos extremos es equivalente al criterio de las dos desviaciones que a su vez es equivalente a definir extremos con una significancia del 5 %.
4. Determine con cuantas desviaciones trabaja el criterio del rango intercuartilar y halle su significancia asociada.
5. Reformule el criterio de las dos desviaciones en términos del rango intercuartilar.

Ejercicio 3: Sobre **outliers** o **eventos extremos** :

1. Los profesores han solicitado en relación con la encuesta que la Universidad le hace a sus estudiantes que se detecten los outliers o datos extremos de la cola inferior y que se excluyan de los análisis ulteriores tales como el cálculo de la media. Ellos alegan que siempre hay alumnos resentidos que utilizan la encuesta para vengarse. ¿Cuál es su punto de vista al respecto?
2. Tanto el criterio de las dos desviaciones como el del rango intercuartilar tratan ambas colas de forma simétrica. Pero debido a que hay distribuciones muy asimétricas, como la $Bi(10000, 0{,}01)$, así mismo debe asimetrizarse la definición de outlier o dato extremo. ¿Qué opina Usted? ¿Cómo lo haría?
3. Considere una distribución en U que es todo lo contrario de una central. En esta última los datos tienden a congregarse hacia el centro de la distribución en tanto que en la primera los datos tienden a huír hacia ambos extremos. Un ejemplo de distribución central: el tiempo que se emplea para llegar al trabajo. Un ejemplo de distribución en U: en una sociedad donde el largo del cabello es un fuerte indicativo del género de una persona, los hombres tienden a llevarlo muy corto y las mujeres bastante largo. ¿Considera Usted que nuestras definiciones de datos extremos son aplicables a este caso? ¿Cuál sería una definición correcta de dato extremo para una distribuión en U?
4. La media es para distribuciones de tendencia central, el estadígrafo que mejor puede representar a toda la población. La varianza y la desviación son medidas del descontento de la población al pretender que sea representada por la media. Como indicador de homogeneidad definimos el coeficiente de variación. Proponga equivalentes en estadística no paramétrica.

El segundo resultado no paramétrico que veremos nos permitirá discriminar con dos colas lo común de lo extraordinario y es aplicable a todas las distribuciones de datos cuantitativos. Para prepararnos, veremos una regla que es válida únicamente para distribuciones normales y después la generalizamos.

Ejercicio 3: Hallar el resumen de 5 valores para los datos del primer parcial del curso A. Hacer el histograma, dibujar el *boxplot* mostrando el rango intercuartilar y los datos extremos.

281 ***Regla de las tres áreas*** *Usando la tabla de la Z se pueden establecer los valores señalados, los cuales conviene memorizar pues reemplazan la tabla cuando uno no la tiene a mano.*

a) El porcentaje de la población que queda entre una desviación al lado y lado de la media es del 68 %, lo cual se haya por el teorema del área grande menos el área chica:

$$p(-1 < Z < 1) = p(\mu - \sigma < X < \mu + \sigma) = 0{,}68$$

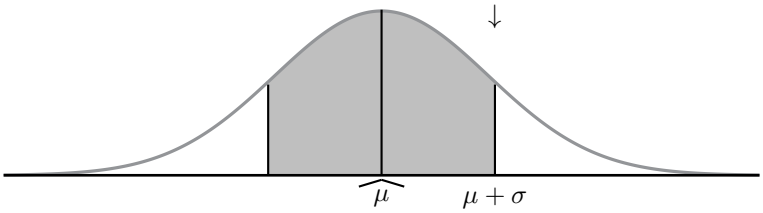


Figura 14.0. Lo que queda alrededor de la media y a menos de una desviación estándar cubre el 68 % de la población.

b) El porcentaje de la población que queda entre dos desviaciones a lado y lado de la media es del 95 %:

$$p(-2 < Z < 2) = p(\mu - 2\sigma < X < \mu + 2\sigma) = 0{,}95$$

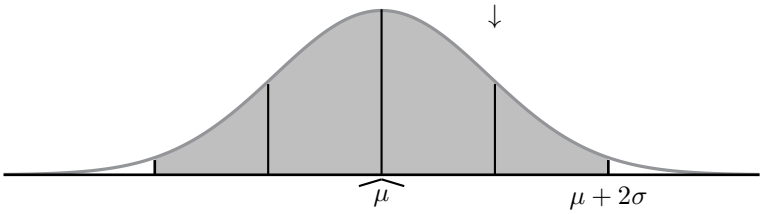


Figura 14.1. Lo que queda alrededor de la media y a menos de dos desviaciones estándares cubre el 95.5 % de la población.

c) El porcentaje de la población que queda entre tres desviaciones a lado y lado de la media es del 99.7 %.

$$p(-3 < Z < 3) = p(\mu - 3\sigma < X < \mu + 3\sigma) = 0{,}997$$

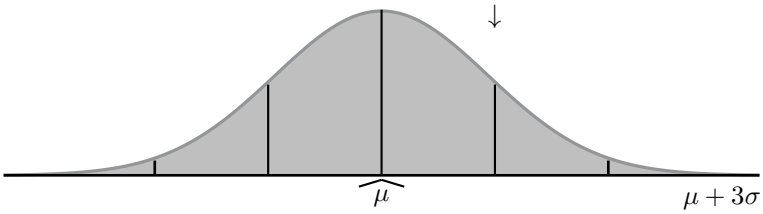


Figura 14.2. Lo que queda alrededor de la media y no más lejos que tres desviaciones estándares cubre el 99.7 % de la población.

d) Fuera de 4 desviaciones estándares al lado y lado de la media no hay prácticamente nada, a menos que se hagan muestreos muy grandes del orden de 10000 eventos:

$$p(Z < -4) + p(Z > 4) = p(X < \mu - 4\sigma) + p(X > \mu + 4\sigma) \sim 0$$

Ejercicio 4: Use la tabla de la z para demostrar los resultados de la regla de las 3 áreas.

Todos estos valores no dependen ni de la media ni de la varianza, son siempre los mismos para todas las campanas.

Aprendamos a escribir esta regla usando valores absolutos y desigualdades. Lo que hay que recordar es que si a y b son números sobre la recta, entonces $\mid a - b \mid$ representa la distancia entre ellos, pues es la resta pero siempre positiva. Teniedo en cuenta esto, la regla se escribe:

1. Lo que queda alrededor de la media y a menos de una desviación estándar cubre el 68 % de la población: $P(x :\mid x-\mu \mid < \sigma) = 0{,}68$. Esta expresión se lee: el área o probabilidad de la región que queda a una distancia menor (o igual) que una desviación es el 68 %.

2. Lo que queda alrededor de la media y a menos de dos desviaciones estándares cubre el 95.5 % de la población. $P(x :\mid x-\mu \mid < 2\sigma) = 0{,}955$. Esta expresión se lee: el área o probabilidad de la región que queda a una distancia menor (o igual) que dos desviaciones es el 95.5 %.

3. Lo que queda alrededor de la media y a menos de tres desviaciones estándares cubre el 99.7 % de la población. $P(x :\mid x-\mu \mid < 3\sigma) = 0{,}997$. Esta expresión se lee: el área o probabilidad de la región que queda a una distancia menor (o igual) que tres desviaciones es el 99.7 %.

Hemos formulado la regla de las tres áreas midiendo lo que está dentro de la región que queda cerca de la media y máximo a una distancia de tantas desviaciones. Podemos decir lo mismo en términos de lo que queda por fuera de dicha región:

1. Lo que queda fuera de más de una desviación estándar lejos de la media cubre el 100 % - 68 % = 32 % de la población: $P(x :\mid x-\mu \mid > \sigma) = 0{,}32$. Esta expresión se lee: el área o probabilidad de la región que queda a una distancia mayor (o igual) que una desviación es el 32 %.

2. Lo que queda alrededor de la media y a más de dos desviaciones estándares cubre el 4.5 % de la población. $P(x :\mid x-\mu \mid > 2\sigma) = 0{,}045$. Esta expresión se lee: el área o probabilidad de la región que queda a una distancia mayor (o igual) que dos desviaciones es el 4.5 %.

3. Lo que queda alrededor de la media y a más de tres desviaciones estándares cubre el 0.3 % de la población. $P(x :\mid x-\mu \mid > 3\sigma) = 0{,}003$. Esta expresión se lee: el área o probabilidad de la región que queda a una distancia mayor (o igual) que tres desviaciones es el 0.3 %.

Esta regla es válida para una v.a. con distribución normal, descrita por una Campana de Gauss. Pero sucede que nuestras campanas son idealizaciones de una distribución de tendencia central. Pero además de distribuciones de tendencia central hay muchas más, por ejemplo, una distribución que en vez de ser una montaña de un sólo pico tenga dos jorobas o 5 o una combinación complicada de valles y colinas. ¿Cómo haremos para calcular las barreras que dividen lo normal de lo extraordinario en dichos casos si sólamente tenemos la media y la desviación? Este tipo de preocupaciones tienen una respuesta conocida como el teorema de Chevichev. primero lo enunciaremos, luego lo probaremos y después daremos algunos ejemplos.

282 *Teorema de Chevichev*

Para una variable aleatoria X cualquiera de datos cuantitativos, con función densidad de probabilidad $p(x)$, media μ y desviación σ, sea k el número de desviaciones, el cual puede ser fraccionario o decimal. Entonces se tiene que

$$P(\mid x-\mu \mid \geq k\sigma) \leq \frac{1}{k^2}$$

lo cual dice que la probabilidad de los outliers que quedan más allá de k desviaciones alejados de la media es menor o igual que $\frac{1}{k^2}$. Esto es válido para todas las distribuciones, sean normales o no. Sin embargo, cuando se trata de distribuciones normales, es más exacto usar la Z que este teorema.

Para probar el Teorema de Chevichev necesitamos aprender a escribir la media y la varianza en términos de integrales o áreas.

283 *La media y varianza con integrales*

La media de una lista de datos es

$$\bar{X} = \frac{\sum x_i}{n}$$

donde n es el número de datos.
La varianza muestral es igual a

$$s^2 = \frac{\Sigma(x_i - \bar{X})^2}{n-1}$$

Cuando uno agrupa los datos en una tabla de frecuencias absolutas, sea porque los datos estén repetidos o porque uno hace rangos, con marca de clase x_i y frecuencia absoluta F_i, la media y la varianza se calculan por las siguientes fórmulas:

$$\bar{X} = \frac{\Sigma(xF)}{n} = \frac{\Sigma(x_i F_i)}{n} = \frac{(x_1 F_1) + ... + (x_n F_m)}{n}$$

$$s^2 = \frac{\Sigma(x_i - \bar{X})^2 F_i}{n-1}$$

Tengamos en cuenta que n es el número de datos en la lista de datos, pero m es el número de renglones en la tabla de frecuencias, es decir, el número de clases o rangos.

En vez de usar frecuencias absolutas, con F_i grande, uno puede usar frecuencias relativas, con f_i pequeña, donde la frecuencia relativa es la frecuencia absoluta dividida por el número de datos $f_i = F_i/n$. Asumiendo muchos datos, dividir por n o por $n-1$ es lo mismo. Con dicha concesión, la media y la varianza son:

$$\bar{X} = \Sigma(xf) = \Sigma(x_i f_i) = (x_1 f_1) + ... + (x_m f_m)$$

$$s^2 = \Sigma(x_i - \bar{X})^2 f_i = (x_1 - \bar{X})^2 f_1 + ... + (x_m - \bar{X})^2 f_m$$

Imaginémos que se trata de una variable cuantitativa X que puede tomar tanto valores enteros como fraccionarios y decimales. Decimos que toma valores en los reales. Ejemplo: la altura de las personas, que puede tomar el valor 1.62, por ejemplo. Con esa perspectiva, si uno dibuja una tabla como un diagrama de barras, uno tiende a imponer sobre su diagrama una curva, quizá una Campana de Gauss. Dicha curva representa la envolvente o, más rigurosamente, la función que uno obtendría si tuviese un enorme número de datos. Dicha función ideal se denomina **la función densidad** de X y se denota $p(x)$ puesto que es una probabilidad, una idealización de la frecuencia relativa. Cuando uno pasa de diagramas de barras a envolventes, porque se tienen muchos datos, uno pasa de frecuencias relativas a probabilidades y uno debe pasar de sumas a integrales y por eso la media y la varianza se escriben como:

$$\mu = \int x p(x) dx$$

$$\sigma^2 = \int (x - \mu)^2 p(x) dx$$

Cambiamos $\bar{X}$ por μ y s^2 por σ^2 para hacer énfasis en que trabajamos con cosas ideales, probabilidades a priori o con un número infinito de datos. No se trata de muestras, sino de idealizaciones o modelos teóricos. Tenemos también que

$$\frac{\sigma^2}{\sigma^2} = 1 = \frac{1}{\sigma^2} \int (x - \mu)^2 p(x) dx$$

donde se asume que la integral se extiende sobre todo el eje horizontal. Si es necesario, la función densidad $p(x)$ se define como cero donde no haya nadie. Más abajo utilizaremos esta igualdad.

284 ***Demostración del teorema de Chevichev.*** *Todo empieza así: si hablamos de números positivos, entonces*

$\frac{a}{b} \geq 1$ si $a \geq b$, por ejemplo, $\frac{7}{5} \geq 1$ pues $7 \geq 5$

En vez de números positivos, podemos tomar el valor absoluto de números cualesquiera, o bien el valor absoluto de la división. El valor aboluto se denota con barritas: $| -5 | = 5$. Tenemos:

$|\frac{a}{b}| \geq 1$ si $| a | \geq | b |$, por ejemplo, $|\frac{-7}{5}| \geq 1$ pues $| -7 | \geq | 5 |$.

Ahora aplicamos la desigualdad anterior a una distribución cualquiera, no sólo las normales sino de cualquier tipo, con media μ y desviación σ (siempre positiva) y para un x cualesquiera y un k positivo:

$|\frac{x-\mu}{k\sigma}| \geq 1$ si $\mid x-\mu \mid \geq k\sigma$

La desigualdad $\mid x-\mu \mid \geq k\sigma$ denota una región: cuando uno tiene el valor absoluto de una diferencia de números, uno lo puede interpretar como una distancia. Por ejemplo $\mid 5-9 \mid = 4$, lo cual nos dice que la distancia entre 5 y 9 es 4. Así mismo, decir que $\mid x-\mu \mid \geq k\sigma$ es lo mismo que decir que la distancia entre x y μ es mayor o igual que $k\sigma$. En una gráfica, eso se ve así:

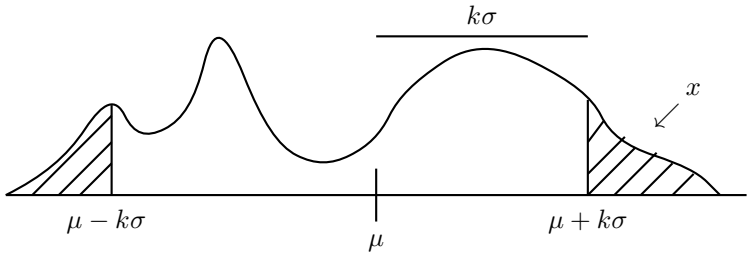


Figura 14.4. La condición $\mid x-\mu \mid \geq k\sigma$ dice que la distancia entre x y μ es mayor que $k\sigma$, y por eso denota la región achurada, con rayitas oblicuas.

Podemos elevar al cuadrado la desigualdad $|\frac{x-\mu}{k\sigma}| \geq 1$ para obtener:

$\left(\frac{x-\mu}{k\sigma}\right)^2 \geq 1$

la cual es válida si $\mid x-\mu \mid \geq k\sigma$. El sentido de la desigualdad no cambia al elevar al cuadrado pues si dos números son positivos y un número es mayor que otro, el cuadrado del primer número también será mayor que el cuadrado del segundo. Por ejemplo, $2 < 3$ y sus respectivos cuadrados cumplen una desigualdad en el mismo sentido: $4 < 9$. Observemos que la desigualdad que denota la región achurada no se eleva al cuadrado, pues es una condición. Lo único que se eleva al cuadrado es la desigualdad algebraica $|\frac{x-\mu}{k\sigma}| \geq 1$.

Podemos multiplicar una desigualdad por un número positivo y la desigualdad no cambia: $2 < 3$ y multiplicando por 4 obtenemos $8 < 24$. Si se multiplica una desigualdad por cero, queda cero igual a cero. Por tanto, podemos multiplicar la desigualdad anterior por $p(x)$ que nunca es negativa, y lo único que hay que tener en cuenta es que la desigualdad debe involucrar el igual, pues no se excluye que $p(x)$ sea cero en algún punto:

$\left(\frac{x-\mu}{k\sigma}\right)^2 p(x) \geq 1p(x) = p(x)$

Ahora integremos a ambos lados para obtener

$\int \left(\frac{x-\mu}{k\sigma}\right)^2 p(x)dx \geq \int p(x)dx$

No nos olvidemos que la integral se restringe a la región achurada, dada por $\mid x-\mu \mid \geq k\sigma$. El sentido de la desigualdad se conserva pues si una función queda debajo de otra, entonces sus áreas o integrales guardan la misma relación de orden:

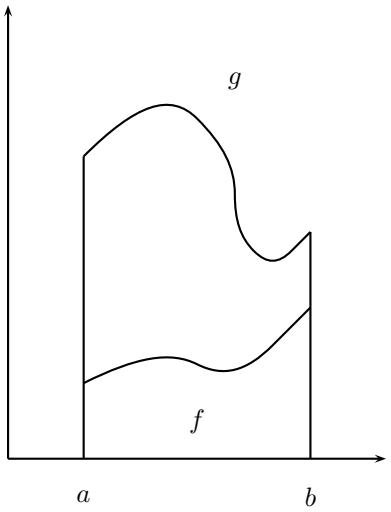


Figura 14.5. Sobre una región, que puede ser el intervalo $[a, b]$, se tiene que $f(x) < g(x)$ y por consiguiente las integrales o áreas sobre la misma región y entre las curvas y el eje horizontal guardan la misma relación de orden: $\int f(x)dx < \int g(x)dx$.

Ahora bien, la $\int p(x)dx$ sobre la región achurada denota el área bajo la curva y sobre dicha región y se interpreta como la probabilidad de dicha región y es igual a:

$P(\mid x - \mu \mid \geq k\sigma)$

que es la probabilidad de la región cuyos elementos x quedan lejos de la media a una distancia mayor o igual a $k\sigma$.

Por tanto, tenemos:

$\int \left(\frac{x-\mu}{k\sigma}\right)^2 p(x)dx \geq \int p(x)dx = P(\mid x - \mu \mid \geq k\sigma)$

que también se puede escribir como

$P(\mid x - \mu \mid \geq k\sigma) \leq \int \left(\frac{x-\mu}{k\sigma}\right)^2 p(x)dx$

donde la integral corre sobre la región achurada.

También podemos desarrollar el cuadrado en el quebrado para obtener:

$P(\mid x - \mu \mid \geq k\sigma) \leq \int \frac{(x-\mu)^2}{(k\sigma)^2} p(x)dx = \int \frac{(x-\mu)^2}{k^2\sigma^2} p(x)dx = \frac{1}{k^2\sigma^2} \int (x-\mu)^2 p(x)dx$

Acordémonos ahora que

$\frac{1}{\sigma^2} \int (x-\mu)^2 p(x)dx = 1$

por ende

$P(\mid x - \mu \mid \geq k\sigma) \leq \frac{1}{k^2\sigma^2} \int (x-\mu)^2 p(x)dx = \frac{1}{k^2}$

En conclusión tenemos:

$P(\mid x - \mu \mid \geq k\sigma) \leq \frac{1}{k^2}$

Fuerte objeción. El desarrollo que acabamos de ver tiene un fuerte inconveniente, a saber: la igualdad

$\frac{1}{\sigma^2}\int(x-\mu)^2 p(x)dx = 1$

es cierta si se asume que la región de integración es todo el eje horizontal, pero en nuestro desarrollo sólo estamos trabajando sobre la región achurada que nunca será igual a todo el eje. Para que el teorema quede rigurosamente demostrado hay que alegar lo siguiente: la integral sobre la región achurada, que representa un área, es menor que la integral de la misma función sobre todo el eje. Para involucrar esto en la argumentación, usamos la siguiente nomenclatura:

$\int_A f(x)dx$ denota la integral sobre la región achurada, en tanto que

$\int_M f(x)dx$ denota la integral sobre todo el mundo, es decir, sobre todo el eje horizontal.

Estamos diciendo que

$\int_A f(x)dx < \int_M f(x)dx.$

Con ese resultado, reescribimos nuestras desigualdades como sigue:

$P(\mid x-\mu \mid \geq k\sigma) \leq \int_A \frac{(x-\mu)^2}{(k\sigma)^2}p(x)dx = \int_A \frac{(x-\mu)^2}{k^2\sigma^2}p(x)dx$
$= \frac{1}{k^2\sigma^2}\int_A (x-\mu)^2 p(x)dx \leq \frac{1}{k^2\sigma^2}\int_M (x-\mu)^2 p(x)dx \leq \frac{1}{k^2}$

que en resumen dice:

Teorema de Chevichev: Para una variable aleatoria X cualquiera de datos cuantitativos, con función densidad de probabilidad $p(x)$, media μ y desviación σ, y sea k el número de desviaciones, el cual puede ser fraccionario o decimal. Entonces se tiene que

$$P(\mid x-\mu \mid \geq k\sigma) \leq \frac{1}{k^2}$$

lo cual dice que las regiones periféricas, la de los outliers, tienen una probabilidad total que decrece máximo como el inverso del cuadrado del número de desviaciones.

La notación usada puede ser algo engorrosa, por ello podemos tratar de ver cómo nos va con esta otra forma de enunciar el teorema:

$$P(D \geq k\sigma) \leq \frac{1}{k^2}$$

donde el símbolo $P(D \geq k\sigma)$ denota el área o probabilidad de la región achurada que queda alejada de la media a partir de k desviaciones. El teorema dice entonces: la región achurada que queda alejada de la media a partir de k desviaciones tiene una probabilidad o área que como máximo es $\frac{1}{k^2}$.

Ejemplo 1: Con $k = 2$ y para cualquier distribución sin importar que sea normal o no, se tiene que

$P(\mid x-\mu \mid \geq 2\sigma) \leq \frac{1}{2^2} = \frac{1}{4} = 0{,}25$

es decir, que en cualquier distribución se tiene que lo que queda más alejado que dos desviaciones tiene una probabilidad menor a 0.25, o lo que es lo mismo: la probabilidad de todo lo que queda alrededor de la media y a una distancia menor que dos desviaciones tiene una probabilidad que es mayor que 0.75.

Ejemplo 2. Reemplacemos $k = 1{,}3$ en el teorema de Chevichev para una v.a. X de la cual no se sabe que tenga una distribución normal pero con media 15 y desviación 1,7. Verbalizar los resultados.

Solución: Reemplazando $k = 1{,}3$ en el teorema de Chevichev tenemos

$P(\mid x-\mu \mid \geq 1{,}3\sigma) \leq \frac{1}{1{,}3^2} = \frac{1}{1{,}69} = 0{,}59$

Como la media vale 15 y la desviación vale 1.7, la condición $\mid x-\mu \mid \geq 1{,}3\sigma$ es igual a $\mid x-15 \mid \geq 1{,}3(1{,}7)$ que es lo mismo que $\mid x-15 \mid \geq 2{,}21$. Como el valor absoluto se puede interpretar como una distancia, la última desigualdad nos dice que la distancia entre la media y x debe ser mayor que 2,21, donde x es un valor que puede tomar X. Por tanto, el extremo superior de la región determinada por dicha desigualdad es $\mu + 2{,}21 = 15 + 2{,}21 = 17{,}21$ y el extremo inferior es $\mu - 2{,}21 = 15 - 2{,}21 = 12{,}79$.

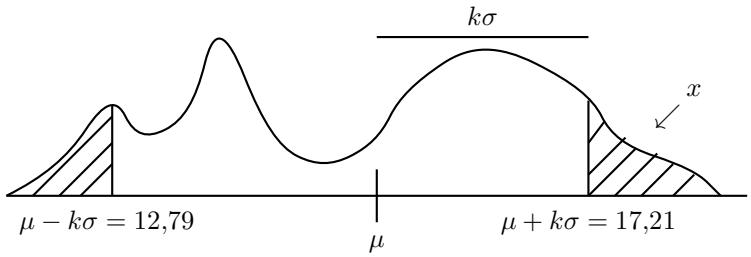


Figura 14.6. La condición $\mid x-\mu \mid \geq k\sigma$ dice que x debe estar más abajo que 12.79 o más arriba que 17.21, y por eso denota la región achurada, con rayitas oblicuas. Dicha región tiene un área o probabilidad menor que 0.59.

En conclusión, el teorema de Chevichev nos dice que la región que queda abajo de 12.79 unida a la que queda arriba de 17.21, la cual está achurada en la figura anterior, tiene un área o probabilidad que es menor que 0.59. Como vemos, no es inteligente considerar en este caso a la región achurada como una región de outliers o valores extremos.

Ejemplo 3. De una distribución sobre las ventas en un almacén se sabe únicamente que la media vale 14 y que la desviación vale 3. El dueño le ofrece a Ramón la posibilidad de quedarse como admistrador del negocio asumiendo que no hay valores abajo de 9 ni arriba de 19. ¿cuáles son los riesgos de Ramón?

Solución: $14 - 9 = 5$ y $19 - 14 = 5$ y 5 medido en unidades de desviación es $5 = 5/3 = 1{,}66$. Ahora, como no podemos asegurar que la distribución sea normal, nos toca usar el teorema de Chevichev con $k = 1{,}66$:

$$P(\mid x-\mu \mid \geq 1{,}66\sigma) \leq \frac{1}{1{,}66^2} = \frac{1}{2{,}7556} = 0{,}36$$

Este resultado nos dice que el trato propuesto incluye una incertidumbre en volumen de ventas que puede llegar al 36 %. Al ver dicho resultado, Ramón alega que dicha incertidumbre es muy grande y que por eso deben asumir que la media no es 14 sino 12. ¿Qué deberíamos responder? Lo que debemos pensar es que si a Ramón le va mal a todos nos va mal y por eso es mejor llevar la discusión a terreno entendible: es mejor conservar la media y ampliar el rango plausible de ventas a un valor de k alrededor de 3 o quizá 4.

Ejemplo 4. Para medir la brecha entre generaciones, un sociólogo estudia el largo del cabello de los varones, X, el cual no sigue una distribución normal, sino que su distribución se parece a la de una F: nadie calvo, muchos con cabello preferencialmente corto, algunos con cabello largo y unos pocos con cabello muy largo. Para edades entre 13 y 19 años, la media es 5 cm y la desviación 2. a)¿Qué porcentaje de muchachos tendrá un largo del cabello más largo que 8 cm o más corto que 2cm? b) Consideremos muestras al azar de 30 muchachos. ¿Qué porcentaje de dichas muestras tendrá una media que se desvíe más que un centímetro de la media?

Para resolver la primera pregunta tenemos que usar Chevichev pues X no se distribuye normalmente:

$P(\mid x-\mu \mid \geq k\sigma) \leq \frac{1}{k^2}$

$P(\mid x-5 \mid \geq k(2)) \leq \frac{1}{k^2}$

Para los extremos del intervalo considerado tenemos:

$\mid 8-5 \mid = \mid 2-5 \mid = 3 = 2k$

por lo que $k = 1{,}5$. Por tanto,

$P(\mid x-5 \mid \geq 1{,}5(2)) \leq \frac{1}{1{,}5^2} = 0{,}44$

lo cual se lee así: el porcentaje de muchachos tomados al azar que tengan un cabello más largo que 8 cm o más corto que 2 es menor que 44 %.

Para resolver la segunda pregunta tenemos: la media muestral $\bar{X}_{30}$ tiene una media igual a la media de X, es decir, 5cm, y una desviación igual a $2/\sqrt{30} = 0{,}365$. Esto es cierto no importa cómo sea la distribución de X, normal o no, lo único que se requiere es que el muestreo sea al azar de una población grande. Ahora tenemos dos opciones: ser demasiado rigurosos y aplicar de nuevo Chevichev para estos nuevos datos, o aplicar el TLC como una buena aproximación debido a que $n = 30$ es en este caso un número muy grande y que por tanto la media muestral tendrá una distribución que bien puede considerarse normal. Nosotros optamos por usar el TLC y todo lo que tenemos que hacer es calcular la transformada-z de los extremos del intervalo dado, los cuales son 4 y 6 (un cm. abajo/arriba de la media que es 5):

$z(6) = \frac{6-5}{0{,}365} = 2{,}74$
$z(4) = \frac{4-5}{0{,}365} = -2{,}74$

Ahora, lo que queda abajo de $z = -2{,}74$ es según la tabla de la Z igual a 0.0035 y por simetría es igual a lo que queda arriba de $z = 2{,}74$. Por tanto, lo que queda abajo de 4 o arriba de 6 tiene un porcentaje que es igual a 0.007, es decir el 0.7 %. Observemos que el TLC nos da una respuesta con igualdad, en cambio Chevichev nos da una respuesta con desigualdad: siempre que se hable de la media de una muestra no muy pequeña, mayor que 10, estamos invitados a usar el TLC por su sencillez.

Atención: el TLC también es no paramétrico. El TLC es una igualdad exacta para distribuciones normales pero se utiliza para distribuciones de prácticamente cualquier tipo como una aproximación muy buena, la cual es tanto mejor cuantos mayor sea el número de datos de la muestra.

285 ***Ejercicios sobre Chevichev.*** *Respuestas pág 325*

1. Reemplazar $k = 1$ y $k = 3$ en el teorema de Chevichev y verbalizar los resultados.

2. Compare para $k = 1$, $k = 2$ y $k = 3$ los resultados del teorema de Chevichev con los hallados para una distribución normal, lo cual lo vimos como la regla de las tres áreas. ¿Hay contradicción entre los resultados? Explique por qué sí o por qué no.

3. Reemplazar k =3.7 en el teorema de Chevichev para una distribución de la cual no se sabe que sea normal pero con media 20.5 y desviación 3.2. Verbalizar los resultados. b) Resuelva de nuevo el mismo ejercicio pero suponiendo que se sabe que la distribución es normal (ayuda: use la tabla de la z con $z = 3.7$).

4. De una distribución se sabe únicamente que la media vale 30 y que la desviación vale 5. El jefe alega que más abajo de 18 o más arriba de 42 no hay nadie. Use el teorema de Chevichev para ofrecerle una cortés enseñanza al jefe. Ayuda: reformule los límites propuestos por el jefe en términos de desviaciones y luego aplique el teorema de Chevichev.

5. De una distribución se sabe que es normal, que la media vale 30 y que la desviación vale 5. El jefe alega que más abajo de 18 o más arriba de 42 no hay nadie. Use herramientas apropiadas para ofrecerle una cortés enseñanza al jefe.

6. La edad de los hombres interesantes para una jovencita es una v.a. con media 25 y desviación 2. Hay un anciano que como mínimo tiene 35 años y que se siente atraído por ella. ¿Qué probabilidad tiene de que la jovencita lo encuentre interesante? ¿Qué pasa si la jovencita se entera de que el pretendiente tiene exactamente 35 años: aumenta o disminuye la probabilidad de aceptación? ¿El problema se resuelve con la cola inferior, con la superior o con dos colas o no importa? Lo malo de todo es que el teorema de Chevichev funcina con dos colas pues involucra un valor absoluto, por eso, su aplicación a problemas con una cola es abusiva pero necesaria pues es lo único que tenemos.

7. Con los datos del último ejemplo, calcular la probabilidad de que una media de una muestra al azar de 25 muchachos esté más alejada de la media poblacional que 2cm. Resolver el problema por los dos métodos: por Chevichev y asumiendo normalidad de la media muestral.

14.3. Taller 3: Test de ajuste

286 ***Objetivo*** *Nosotros hemos caricaturizado unos datos de tendencia central por una Campana de Gauss. Pero también sabemos por intuición que hay datos cuya distribución no se ajusta a una campana, por ejemplo, si presentan varias jorobas perfectamente definidas. Lo que veremos ahora será aprender a decidir rigurosamente si unos datos se ajustan o no a una distribución dada de antemano, la cual puede ser normal o cualquier otra. A ésto se le llama genéricamente un* ***test de ajuste****. Para clasificar si la discrepancia entre lo que se ve y lo que se cree es grande o chica, usamos una tabla de chi-cuadrado,* χ^2*, que está en la pág. 8.4, pág. 159.*

287 ***Ejemplo: distribución uniforme.***

La **distribución uniforme** es aquella que asigna igual probabilidad a todos los eventos posibles. Un juego con un dado se representa bien por una distribución discreta y cada número entero entre 1 y 6 tiene igual probabilidad 1/6. Se trata de una distribución discreta. Si hacemos una ruleta, su aguja puede caer en cualquier parte del círculo y todos los segmentos de arco de igual apertura tienen la misma probabilidad: se trata de una distribución uniforme continua.

En una clase hay 6 grupos de trabajo (tenemos una variable discreta) y ha surgido una pelea porque un grupo parece ser el preferido por el profesor. La acusación se basa en que las pasadas al tablero, que dan punto bueno para la nota, tiene la siguiente frecuencia absoluta: 12, 20, 12, 9, 8, 11. Estudiémos la acusación desde el punto de vista de un test de ajuste a la distribución uniforme.

Paso 1: Formulamos la H_o : todos los grupos tienen igual probabilidad de pasar al tablero. La probabilidad de que un grupo cualquiera pase al tablero es 1/6. La H_a es que alguna probabilidad no es 1/6. La significancia es 0.05. El número de colas lo decidiremos después.

Paso 2: Calculamos las predicciones de la H_o**.** Hay un total de 12+20+12+9+8+11 = 72 pasadas. Por lo tanto, el número esperado de pasadas de cada grupo es $(1/6) \times 72 = 12$.

Paso 3: Datos experimentales: 12, 20, 12, 9, 8, 11.

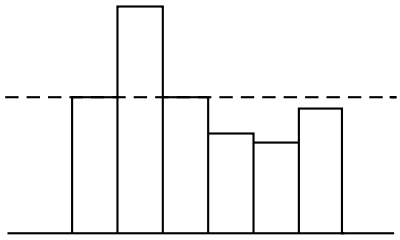


Figura 14.7. Histograma de datos observados. Lo esperado bajo la H_o *de una distribución uniforme se representa por la línea punteada.*

Paso 4: Medimos la discrepancia entre lo que se ve y lo que se cree. Hay infinidad de maneras de hacerlo, pero aquella que ha ganado en la ciencia utiliza la siguiente fórmula

$$Discrepancia = \chi^2_{obs} = \sum \frac{(Observado_i - Esperado_i)^2}{Esperado_i}$$

Observemos que la fórmula para medir la discrepancia involucra únicamente a las frecuencias absolutas tanto de lo observado como de lo esperado. Por eso, nuestro test de ajuste se puede aplicar a cualquier tipo de datos con tal que se dividan en clases y que la H_o quede bien epecificada.

La razón por la cual se ha adoptado esta forma de medir la discrepancia es que no importa con cuál distribución se trabaje, se produce un histograma que al ser idealizado concuerda bastante bien con la distribución chi-cuadrado (tabla de la chi-cuadrado, pág. 159). Con todo, es necesario tener presente

que para que cuadre la aproximación dicha, *el número mínimo de las frecuencias abolutas esperadas no debe ser menor que 5*. En nuestro ejemplo, el menor número esperado es 12, así que no hay problema.

Para nuestro caso, la discrepancia resulta ser:

$Discrepancia = \chi^2_{obs} = \frac{(12-12)^2}{12} + \frac{(20-12)^2}{12} + \frac{(12-12)^2}{12} + \frac{(9-12)^2}{12} + \frac{(8-12)^2}{12} + \frac{(11-12)^2}{12}$

$= 0 + 5{,}33 + 0 + 0{,}75 + 1{,}33 + 0{,}08 = 7{,}49$

Paso 5: Tomamos un veredicto. Para ello decidimos si la discrepancia hallada es grande o chica. Para hacerlo, tenemos que comparar la discrepancia hallada experimentalmente con el valor crítico dada por la tabla. Pero atención: nos falta decidir el número de colas.

Tengamos en cuenta que si la discrepancia es grande, inmediatamente diremos que la H_o se rechaza, que los grupos no tienen igual representación en el tablero. Pero puede pasar que la discrepancia sea muy pequeña, es decir, cercana a cero, pues siempre es positiva por tratarse de cuadrados. En ese caso lo que estamos diciendo es que lo que se observa caza perfectamente con lo que se espera. Por supuesto que no vamos a rechazar la H_o.

De lo dicho se desprende que un test de ajuste o de bondad de ajuste se rechaza sólamente con una cola, la superior. Con todo, uno bien puede tener criterios personales para hacer un test con dos colas o incluso con la cola inferior. Por ejemplo, una discrepancia muy pequeña puede insinuar que se están maquillando los datos para dar una imagen perfecta pero que siempre será sospechosa de ser falsa. Estos casos no serán analizados y siempre resolveremos los tests de ajuste con la cola superior.

También debemos definir los grados de libertad de la chi-cuadrado: son iguales a $k-1$, donde k es el número de clases, categorías o grupos.

Miramos en la tabla de la chi-cuadrado con 6-1= 5 gli y con una significancia de 0.05 en la cola superior. Obtenemos $\chi^2_c = 11{,}0705$ pues $\beta = 2\alpha = 0{,}10$. Vemos que la discrepancia observada es 7.49 y se clasifica como pequeña. Por tanto, aún no hay datos suficientes que desmientan la creencia de que todos los grupos tienen exactamente las mismas posibilidades de pasar al tablero.

288 *Ejemplo: distribución ad hoc.*

Una *distribución ad hoc* es aquella que se ajusta a propósito de un caso particular. En un país hay 5 partidos políticos. Según las oficinas de prensa, su número de miembros en millones es 2, 1.5, 0.8, 0.6, 4. Los analistas dicen que los partidos tienen un pésimo sistema de registro de membresía y que por lo tanto, no hay razón para creerle a las oficinas de prensa. Para demostrarlo, hicieron un muestreo al azar que dio la siguiente membresía en el mismo orden de numeración: 100, 40, 35, 38, 120. Hagamos un test de ajuste de lo observado a lo esperado de acuerdo a las oficinas de prensa.

Paso 1: Formulamos la H_o. El número total de personas en todos los partidos es 2+1.5+0.8+0.6+4 = 8.9 (millones). Por tanto, las frecuencias esperadas son, en el orden correspondiente, 2/8.9 = 0.225; 1.5/8.9 = 0.169; 0.8/8.9 = 0.090; 0.6/8.9 = 0.067; 4/8.9 = 0.449. La suma total da 1 cerrado.

Paso 2: Calculamos las predicciones de la H_o**.** La muestra tiene un total de 100+40+35+38+120=333 personas. Las frecuencias esperadas son proporcionales a las probabilidades. En el orden correspondiente, dichas frecuencia son: $333 \times 0{,}225 = 74{,}9$, $333 \times 0{,}169 = 56{,}3$, $333 \times 0{,}090 = 30$, $333 \times 0{,}067 = 22{,}3$, $333 \times 0{,}449 = 149{,}5$. Da un total de 333.

Paso 3: Datos experimentales: 100, 40, 35, 38, 120.

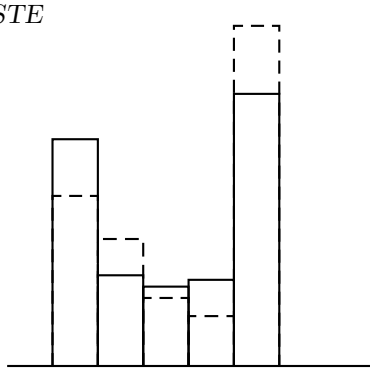


Figura 14.8. Histograma de datos observados en línea sólida. Lo esperado bajo la H_o declarada por las oficinas de prensa en línea punteada.

Paso 4: Medimos la discrepancia entre lo que se ve y lo que se cree.

$Discrepancia = \chi^2_{obs} = \sum \frac{(Observado_i - Esperado_i)^2}{Esperado_i}$

$Discrepancia = \chi^2_{obs} = \frac{(100-74,9)^2}{74,9} + \frac{(40-56,3)^2}{56,3} + \frac{(35-30)^2}{30} + \frac{(38-22,3)^2}{22,3} + \frac{(120-149,5)^2}{149,5}$
$= 8,41 + 4,71 + 0,83 + 11,05 + 5,82 = 30,82$

Paso 5: Tomamos un veredicto. Miramos en la tabla de la chi-cuadrado con 5-1= 4 gli y con una significancia de 0.05 en la cola superior. Obtenemos $\chi^2_c = 9,4877$ pues $\beta = 2\alpha = 0,10$. Como la discrepancia observada es 30,82, la cual es mucho mayor que el valor crítico de la chi-cuadrado, la discrepancia observada se clasifica como grande. Por tanto, concluímos que las frecuencias observadas en la muestra no se ajustan a las frecuencias esperadas. Los analistas tienen toda la razón. Nuestros resultados pueden complementarse con un estudio de comparación de una frecuencia observada con una probabilidad dada.

289 ***Ejemplo: distribución binomial.***

Vimos en el capítulo 2 cómo se calcula una distribución binomial, para lo cual era clave que los eventos fuesen al azar, independientes. Ahora veremos cuándo se ajustan unos datos a una distribución binomial.

Un sociólogo alega que la probabilidad de encontrar una persona que piensa que la injusticia es la causa más importante del terrorismo es 1/8. El asegura además que tales personas están al azar en cualquier familia o grupo. Como prueba, el dice que entró a una cafetería y fue preguntando en cada mesa que tenía un grupo de 3 personas sobre lo que cada quien creía. De 100 mesas estudiadas, la distribución de personas que creían la dicha tesis es como sigue: hubo 71 con cero , 26 con una, 2 con dos y 1 con tres. ¿Es cierto que los datos respaldan lo que el sociólogo dice?

Paso 1: Formulamos la H_o. El sociólogo alega dos cosas. La primera es sobre el valor de la probabilidad de que alguien piense que la causa fundamental del terrorismo es la injusticia. La segunda es que las personas que piensan así están al azar en cualquier parte. Esta última aseveración es asombrosa, pues dice que tener o no tener tal creencia es algo íntimimante personal, algo que escapa al control de la propaganda o a los sesgos conceptuales que dominan los círculos sociales a los que uno pertenece. Ambas pretensiones quedan bien representadas en la siguiente proposición: el número de personas que entre grupos de 3 tomados al azar creen que la injustica es la causa fundamental del terrorismo se distribuye como $X \sim Bi(3, 1/8)$.

Paso 2: Calculamos las predicciones de la H_o. La tabla de probabilidades de la binomial dada ya la calculamos en el ejemplo 49, pág. 34. Las probabilidades de cero, uno dos y tres son respectivamente: 0.6699, 0.2871, 0.0410 y 0.002. Como la muestra tiene 100 mesas, las frecuencias esperadas son 67, 29, 4 y 0. La suma total da 100.

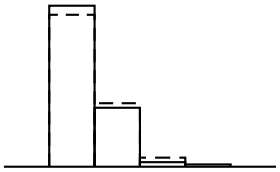


Figura 14.9. Histograma de datos observados en línea sólida. Lo esperado bajo la H_o declarada por el sociólogo en línea punteada.

Tenemos otro problema y es que hay frecuencias esperadas menores que cinco. El remedio es agrupar las clases pequeñas para que conjuntamente sumen más de cinco. Terminamos con 2 clases, la primera es la de cero y la segunda reune las clases 1, 2 o 3. La frecuencia esperada de la primera clase es 67 y la de la segunda es 33.

Paso 3: Datos experimentales: 71 y 29, pues como hicimos en lo esperado, lo hacemos también en lo observado: reunimos las últimas 3 clases en una sóla.

Paso 4: Medimos la discrepancia entre lo que se ve y lo que se cree.

$Discrepancia = \chi^2_{obs} = \sum \frac{(Observado_i - Esperado_i)^2}{Esperado_i}$

$Discrepancia = \chi^2_{obs} = \frac{(71-67)^2}{67} + \frac{(29-33)^2}{33}$

$= 0{,}72$

Paso 5: Tomamos un veredicto. Miramos en la tabla de la chi-cuadrado con 2-1= 1 gli y con una significancia de 0.05 en la cola superior. Obtenemos $\chi^2_c = 3{,}8415$ pues $\beta = 2\alpha = 0{,}10$. Como la discrepancia observada es 0,72, la cual es menor que el valor crítico de la chi-cuadrado, la discrepancia observada se clasifica como pequeña. Por tanto, concluímos que las frecuencias observadas en la muestra se ajustan a las frecuencias esperadas. No encontramos razón para contrariar al sociólogo.

Atención: la probabilidad de cara en este ejemplo se conocía de antemano. Pero en un muchos casos, uno no la conoce sino que la probabilidad de cara debe estimarse a partir de los mismos datos. Para ello, la probabilidad de cara se estima por el número total de caras entre el número total de monedas. Así, en el presente ejemplo, el número total de personas que creen que la causa fundamental del terrorismo es la injusticia es $1 \times 26 + 2 \times 2 + 1 \times 3 = 26 + 4 + 3 = 33$. Como son 100 mesas, cada una con 3 personas, hay 300 personas. Por lo tanto, la probabilidad de cara es $33/300 = 0.11$. Si en un caso como éste se hace un test de ajuste, uno tiene que restar dos grados de libertad y no uno. El primer grado se resta siempre porque el número total esperado de personas se tomó igual al de la muestra y el segundo porque uno de los parámetros de la hipótesis nula se calculó a partir de los mismos datos.

290 *Ejemplo: test de Kolmogorov-Smirnov para normalidad.*

El test chi-cuadrado también puede aplicarse para decidir si unos datos se ajustan o no a una normal. Sin embargo, específicamente para normalidad existen varios tests famosos y muy efectivos. Uno muy popular es el el **test de Kolmogorov Smirnov**. Para ver como se aplica, consideremos los siguientes datos que corresponden a las alturas en centímetros de 72 jóvenes:

145, 164, 171, 183, 177, 184, 168, 163, 157, 172,
179, 168, 172, 170, 172, 157, 162, 159, 164, 158,
173, 181, 162, 159, 167, 162, 166, 147, 163, 152,
156, 150, 170, 149, 180, 148, 163, 168, 169, 158,
150, 158, 174, 179, 168, 167, 182, 174, 168, 160,
180, 152, 160, 160, 148, 155, 173, 151, 149, 169,
161, 155, 160, 176, 177, 170, 169, 165, 146, 175,

156, 166

Agrupamos los datos comenzando desde 150 y en intervalos de longitud 10, cerrados por abajo y abiertos por arriba. La tabla de frecuencias absolutas correspondiente es la siguiente, la cual claramente denota una tendencia central pero necesitamos saber si los datos se ajustan o no a una distribucion normal. La tabla presenta las marcas de clase x y su frecuencia absoluta $F(x)$

Frecuencia absoluta de la altura en cms de unos jóvenes	
x	$F(x)$
150	12
160	24
170	24
180	12

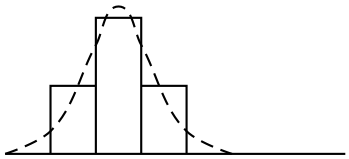


Figura 14.10. Histograma de datos observados en línea sólida. Aparentemente los datos se ajustan a una normal, la curva punteada.

Solución rigurosa:
Paso 1: Datos experimentales: Corresponden a las marcas de clase x y a su frecuencia absoluta observada $F(x)$. Añadimos a nuestra tabla otra columna con $A(x)$, la frecuencia absoluta acumulada (que acumula sobre $F(x)$). Añadimos una tercera columna con la frecuencia acumulada relativa $\Phi(x)$ que se halla dividiendo la frecuencia acumulada absoluta $A(x)$ por el número total de personas que es 72.

Paso 2: Formulamos la H_o. Para ello, calculamos la media y la desviación usando la metodología para tablas de frecuencias absolutas o histogramas, ejemplo 23 pág. 13. La media es 1.65 y la desviación es 9.5743. Por lo tanto, la H_o es que $X \sim N(1.65, 9.5743)$.

Paso 3: Calculamos las predicciones de la H_o. Para cada marca de clase x, calculamos la frecuencia acumulada predicha suponiendo que $X \sim N(165, 9.5743)$. Para ello, usamos la transformada-z para estandarizar los datos:

$z = \frac{x-165}{9,5743}$.

Al utilizar esta fórmula sobre x, la marca de clase, obtenemos los datos de la columna $z(x)$ de la tabla que está más abajo. Ahora miramos en la tabla de la z, numeral 78, pág. 63, el valor de $\Psi(z)$ para ver cuánto es el valor acumulado hasta z. El resultado está en la columna $\Psi(z)$. Eso sería lo que llevaríamos acumulado hasta $z = z(x)$ si nuestros datos tuviesen una distribución normal.

Paso 4: Medimos la discrepancia entre lo que se ve y lo que se cree. Lo observado está en la columna $\Phi(x)$ y lo esperado en la columna $\Psi(z(x))$. La discrepancia entre lo observado y lo esperado se mide en dos pasos. Primero: para cada clase se calcula el valor absoluto de la diferencia entre lo observado y lo esperado: $|\Phi(x) - \Psi(z(x))|$. Segundo: se toma el máximo valor de tales valores absolutos. La H_o dice que tales diferencias se deben al azar, la alterna que hay un efecto capaz de crear diferencias sistemáticamente. Tenemos una gran tabla con los siguientes datos:

La marca de clase de la altura es x, la frecuencia absoluta de x es $F(x)$, la frecuencia absoluta acumulada es $A(x)$, la frecuencia acumulada relativa de x es $\Phi(x)$, la frecuencia relativa acumulada según la hipótesis de normalidad es $\Psi(z(x))$. Tenemos también una última columna con $|\Phi(x) - \Psi(z(x))|$.

Test de Kolmogorov-Smirnov						
x	$F(x)$	$A(x)$	$\Phi(x)$	$z(x)$	$\Psi(z(x))$	$\lvert\Phi(x)-\Psi(z(x))\rvert$
150	12	12	0.167	-1.57	0.059	0.108
160	24	36	0.5	-0.522	0.301	0.199
170	24	60	0.833	0.522	0.699	0.134
180	12	72	1	1.57	0.941	0.058

El valor máximo de $|\Phi(x) - \Psi(z(x))|$ es $D = 0{,}199$. Esta es la discrepancia entre lo observado y lo esperado.

Paso 5: Tomamos un veredicto. Ahora debemos decidir si tal discrepancia es pequeña o grande. Para ello vamos a la tabla que está más abajo y que nos da el valor crítico de D y elegimos la significancia juntamente con el tamaño de la muestra, el cual en nuestro caso es 72. Vemos que 72 clasifica como mayor que 35 por lo que se le aplica la fórmula correspondiente a la significancia elegida:

Para $\alpha = 0{,}05$, el máximo valor correspondiente para la discrepancia permitida es $1{,}36/\sqrt{72} = 1{,}36/8{,}48 = 0{,}16$. Como la discrepancia observada es 0.199, la cual es más grande que 0.16, el valor permitido por azar, rechazamos la H_o y declaramos que los datos no se ajustan a una distribución normal aunque a todas luces lo sea de tendencia central. La fórmula $1{,}63/\sqrt{n} = 1{,}63/\sqrt{72} = 1{,}63/8{,}48 = 0{,}192$ da una aproximación que es válida cuando el número de eventos es mayor que 35 y $\alpha = 0{,}01$. Para dicha significancia, la H_o también se rechazaría: definitivamente es sensato alegar que los datos no tienen una distribución normal.

291 ***Ejercicios sobre ajuste.*** *Respuestas en la pág 325*

1. Un apostador empedernido falsifica un dado cambiando los puntos negros de unas caras por incrustaciones de plomo. Él hace unas pruebas y registra el número de veces que cae cada cara. Sus resultados son: la cara uno cayó 23 veces, la dos 25, la tres 19, la cuarta 30, la quinta 25, la sexta 22. Y el dice: obviamente tengo que apostar por el 4. ¿Usted qué opina?

2. Según los registros del año pasado, se vendieron 23 unidades tipo A, 45 tipo B, 70 tipo C, 41 tipo D y 10 tipo E. Con dichas cifras se hizo un pedido de 170 unidades. En el mes que pasó se vendieron 2 de A, 4 de B, 12 de C, 5 de D y 8 de E. ¿Hay razones para ordenar un pedido nuevo que se adapte a la nueva tendencia del mercado?

3. Por causa de un desastre ecológico de alta magnitud, la compañía LAJ ha sufrido un colpaso finaciero. Entre sus accionistas hay una cierta proporción que quieren vender sus acciones aún perdiendo. Es importante saber si dichas personas se encuentran al azar o si por el contrario andan en grupos que refuerzan sus convicciones mutuamente. Para ello, se encuestó el número de personas que entre grupos de 3, escogidos al azar en el pasillo, se declaran partidarios de vender a un menor precio. Se registraron 6 grupos con ninguna persona que quiera vender, 9 con una, 12 con dos y 20 con 3. Estimar la proporción de los que quieren vender y decidir si se agrupan al azar. Ayuda: la proporción de los que quieren vender es igual al número de personas que quieren vender dividido entre el número total de personas. Por ejemplo, como hay 12 grupos con dos personas que quieren vender, entre dichos grupos hay 24 personas que quieren vender. Atención: es este problema estamos calculando a partir de los datos uno de los parámetros de la H_o. Por eso hay que restar uno de los grados de libertad de la chi-cuadrado.

4. Los usuarios de un empresa muy importante se quejan de la mala atención pues ven que hay una caja desocupada. El Gerente responde: *tuvieron 6 meses para pagar pero todos se vinieron el último día.* La gente grita: *pues mande otro cajero.* El Gerente les proyecta el histograma del número de personas que pagaron por mes. El primer mes vinieron 230 personas, el segundo 2500, el tercero 7658, el cuarto 8752, el quinto 11375, el sexto 30400 y contando. Dice el Gerente: *claramente todo el mundo deja las diligencias para el último día. Por eso, aguanten.* Sale un jovencito y le dice: *Usted quiere decir que no hay una distribución normal, pero fíjese que si la hubiese, de todas formas se formaría un tumulto en los días cercanos al día medio de pago. Así que ponga un cajero.* El Gerente, que ha demostrado ser muy valiente al dar la cara, se queda mirando y dice: *lo del tumulto*

sería cierto pero su argumento es falso porque no es una distribución normal lo que buscamos. Y aún más, ni siquiera es uniforme.

a) Calcule la media y la desviación de la fecha de pago (se usa la tecnología para histogramas aprendida en el capítulo 1).

b) Demuestre usando K-S que la distribución no tiene una distribución normal con media y desviación las de la muestra.

c) ¿Qué quiere decir eso de que el argumento es falso? ¿Cuál sería la distribución óptima desde el punto de vista del Gerente? ¿Cómo podría lograrse?

TEST DE KOLMOGOROV-SMIRNOV SOBRE NORMALIDAD					
MUESTRA (N)	VALOR CRÍTICO DE D = MAX$\|\Psi(x) - \Psi(z)\|$				
	α =0.2	α =0.15	α =0.1	α =0.05	α =0.01
1	0.9	0.925	0.95	0.975	0.995
2	0.684	0.726	0.776	0.842	0.929
3	0.565	0.597	0.642	0.708	0.828
4	0.494	0.525	0.564	0.624	0.733
5	0.446	0.474	0.51	0.565	0.669
6	0.41	0.436	0.47	0.521	0.618
7	0.381	0.405	0.438	0.486	0.577
8	0.358	0.381	0.411	0.457	0.543
9	0.339	0.36	0.388	0.432	0.514
10	0.322	0.342	0.368	0.41	0.49
11	0.307	0.326	0.352	0.391	0.468
12	0.295	0.313	0.338	0.375	0.45
13	0.284	0.302	0.325	0.361	0.433
14	0.274	0.292	0.314	0.349	0.418
15	0.266	0.283	0.304	0.338	0.404
16	0.258	0.274	0.295	0.328	0.392
17	0.25	0.266	0.286	0.318	0.381
18	0.244	0.259	0.278	0.309	0.371
19	0.237	0.252	0.272	0.301	0.363
20	0.231	0.246	0.264	0.294	0.356
25	0.21	0.22	0.24	0.27	0.32
30	0.19	0.2	0.22	0.24	0.29
35	0.18	0.19	0.21	0.23	0.27
> 35	$1.07/\sqrt{n}$	$1.14/\sqrt{n}$	$1.22/\sqrt{n}$	$1.36/\sqrt{n}$	$1.63/\sqrt{n}$

Esta tabla da los valores críticos par $D = max|\Psi(x) - \Psi(z)|$ para muestras pequeñas y diversas significancias. Como hay un valor absoluto de por medio, el test queda automáticamente de dos colas. . Obsérvese que mientras en el test chi-cuadrado los grados de libertad están relacionados con el número de clases, en el test K-S lo que importa es el tamaño de la muestra. Por eso el K-S es más eficiente y con menos trabajo.

14.4. Taller 4: Gnumeric y OpenOffice.

Respuestas en la pág 326

Objetivo Aprender a usar OpenOffice y/o Gnumeric para calcular los valores de los estadígrafos de la normal y resolver problemas sobre hipótesis nulas e intervalos de confianza.

14.4.1. Un diamante por un dolar

¿Quién es aquel que encuentra un diamante que vale un millón de dólares y al querer venderlo dice: con que Usted me de 100 ya es suficiente? Expliquemos con ejemplos lo que esto significa.

292 ***Ejemplo*** *Si estamos comparando dos medias para saber si son iguales y se conoce la desviación poblacional y se rechaza con la cola inferior, usamos una z. Si la z_{obs} es 3.8, la igualdad de las medias se puede rechazar con una significancia igual a 0.05. Eso quiere decir que estamos asegurando que nuestros datos son más extremos que el 95 % de los casos. Pero, según la tabla de la z, la significancia mínima con la que eso puede hacerse es 0.0001. Lo que significa que sólo el 1 por diez mil de los casos tendrían valores más extremos que los nuestros. Ahora, bien, ¿qué es más contundente: decir que nuestros datos son más extremos que el 95 % o que el 99.99 %? Por supuesto que el segundo enunciado es más contundente.*

Vemos que establecer de antemano la significancia puede implicar menospreciar el valor de los datos. ¿Cuál es entonces el remedio? La solución es usar los computadores para que nos den el porcentaje de la población que es más extrema que nuestros datos. Por eso los computadores no aceptan ni rechazan hipótesis nulas sino sólamente dan el $p - value$ que en nuestros términos se llama la significancia del evento en cuestión: es la probabilidad de ser más extremo que nuestros datos. Si la significancia de un evento es 0.03 y se está trabajando con una significancia de 0.05, nuestro evento es interesante y se rechaza la H_o.

293 ***Receta para el rechazo o aceptación de hipótesis nulas usando p-values.*** *Si un evento es extremo con respecto a una significancia α, la H_o se rechaza pero de lo contrario se acepta. Ahora, el $p-value$ asociado a un evento representa la proporción de eventos que son más extremos que él. Por tanto, si $(p-value) < \alpha$, nuestro evento es más extremo que el valor crítico asociado a α y la H_o debe rechazarse. De otra forma, está permitido aceptarla.*

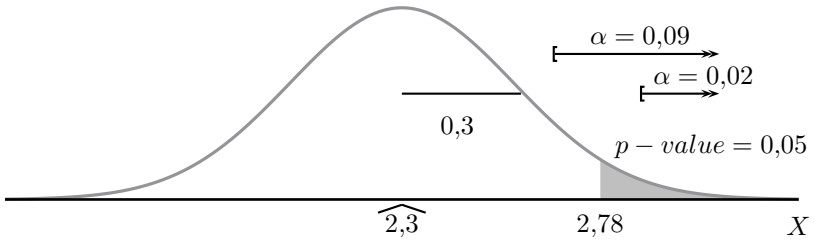


Figura 14.11. Tenemos la campana de probabilidad para una variable con distribución normal con media 2.3 y desviación 0.3. El p-value asociado al evento 2.78 con una cola es el área de la región sombreada correspondiente a la proporción de eventos que con la cola superior son más extremos que 2.78. Para dicho caso y la cola superior, el p-value es 0.05. Por tanto, para el evento 2.78 y significancia 0.02, la $H_o : \mu = 2{,}3$ se acepta porque en tal caso el evento es normal, pero se rechazaría si $\alpha = 0{,}09$ porque en tal caso el evento sería extremo. Vemos que los p-value pueden usarse pr decidir una H_o con cualquier significancia.

294 ***Ejemplo*** *Si usamos una significancia de 0.05 para comparar dos medias cuando se conoce la desviación de una muestra al azar con 7 datos, usamos una t. Como la t_{obs} es 3.4 y la t crítica con 6 gli y dos colas es ± 2.447, la igualdad de las medias se puede rechazar. Pero, según la tabla de la t, la significancia mínima con la que eso puede hacerse está entre 0.02 y 0.01. Excel nos dice que la significancia mínima es 0.014495928 = 0.015. Lo que significa que sólo el 15 por mil de los casos tendrían valores más extremos que los nuestros. Eso es mucho más contundente que decir que el 5 por ciento o 50 por mil de los casos son más extremos que nosotros.*

295 ***Ejemplo*** *Si usamos una significancia de 0.05 y la cola superior para comparar dos varianzas, usamos una F. Como la F_{obs} es 4.7 y la F crítica superior con 6 glinumerador y 8 denominador y dos colas es 3.58, la igualdad de las medias se puede rechazar. Pero, según las tablas de la F, la significancia mínima con la que eso puede hacerse está entre 0.01 y 0.05. Excel nos dice que la significancia mínima es $0.024292161 = 0.024$. Lo que significa que sólo el 24 por mil de los casos tendrían valores más extremos que los nuestros. Eso es más contundente que decir que el 5 por ciento o 50 por mil de los casos son más extremos que nosotros.*

Alerta: Cuando uno comienza a usar un paquete para hacer un estudio, lo primero que hay que hacer es resolver con ayuda del paquete unos problemas ya resueltos. De esa forma, uno puede conectarse al paquete con confianza. Sucede además que no es raro que uno encuentre un error de programación. Tal es el caso de Excel y de Gnumeric: ambos tienen errores de programación, pero hay una diferencia: los errores de Excel tienen que ver con funciones básicas, en cambio Gnumeric es famoso por la precisión de sus resultados, y sus errores se relacionan con la documentación y con la interfase con el usuario. Por eso, el presente taller se reducirá simplemente a saber cómo usar las funciones de Gnumeric y/o de OpenOffice, de la cuales sabemos que funcionan bien y pueden reemplazar a las tablas. También aprenderemos a usar algunos procedimientos, de los cuales podemos verificar que funcionan bien.

OpenOffice y Gnumeric son paquetes gratis, desarrollados para Linux pero ahora tambien se pueden correr bajo Windows. La versión para Windows de Gnumeric tiene en el momento presente algunos bugs de interacción con el usuario, pero con algo de tolerancia, uno puede usarlo.

14.4.2. ¡Linux a la orden!

Un programa que permite que uno use un computador se llama sistema operativo. Por la facilidad de manejo, Windows ha sido el sistema operativo preferido por la mayoría de personas, en tanto que Linux fue durante décadas patrimonio exclusivo de gurús muy diestros. Pero eso ha cambiado: si alguien ha deseado ponerle fin a la dulce esclavitud de Microsoft y sus imponentes productos, ahora tiene una opción para hacerlo. Personalmente recomiendo las versiones de Linux OpenSUSE y Ubuntu pues son gratis y en mi opinión ya son accesibles a la gente que desea un sistema operativo gratis, poderoso, fácil de usar para las tareas comunes y corrientes, y además en permanente evolución. En realidad, hay una fuerte competencia por simplicidad y efectividad en todas las versiones de Linux y no es raro oír aplausos a una cualquiera. Fedora, Debian, Fuduntu, Kubunto son otras versiones de Linux muy famosas.

La forma natural de cambiarse a Linux es la misma para todas las distribuciones, digamos que sea OpenSUSE. Es como sigue: uno no borra su sistema antiguo, al cual uno está acostumbrado. Más bien, uno consigue un disco duro y lo dedica para su nuevo sistema. OpenSUSE se baja de la red y se instala tan fácilmente como Windows. Es posible que OpenSUSE ponga problemas con los periféricos, por ejemplo, las impresoras, debido a que los drivers para Linux son más escasos que los que son para Windows. Con todo, la industria ha tomado en serio la siempre creciente importancia de Linux y ahora se esfuerza por producir drivers multiplataforma. Para cambiarse a Linux Ubunto, el cual es muy fuerte con periféricos, es todavía más simple: es buena idea echarle una mirada a su página web.

14.4.3. Instalación de Gnumeric

Para instalar Gnumeric:

1. En Windows, localice a Gnumeric con Google, bájelo e instálelo. Se inicia con un click sobre el ícono correspondiente.

2. En OpenSUSE: en la esquina izquierda-abajo, está el ícono de OpenSuse. Haga click sobre el, vaya a *computer*. Elija *Install Software*. De su clave de Administrador y espere a que la aplicación YAST2 se inicialice. En el buscador diga

   Gnumeric

   Termine con *Enter*. Siga las instrucciones (active algunas casillas donde se explica que ha de bajarse de la red) y espere a que el la aplicación YAST2, el instalador, se cierre por haber concluído su trabajo.

Para correr Gnumeric: abra una terminal o consola: haciendo click sobre el *KickOff Application launcher*, esquina inferior izquierda, y elija *terminal*). Sobre la terminal escriba

Gnumeric.

Espere algunos segundos y verá cómo se despliega la aplicación. Gnumeric tiene la imagen y semejanza de Excel y ha procurado ser compatible con Excel al máximo. Para verificarlo, haga algunas operaciones, una suma, una resta, pinte la gráfica de algún histograma.

Ejercicio 0: Hallemos la media y la varianza de los datos 1, 2, 3, 4, 2, 3, 3, 2, 3.

Para usar Gnumeric: teclee los datos sobre una columna, vaya al menu *Tools*, luego a *Statistical analysis*, y luego a *Descriptive statistics*. Una ventana de diálogo se abrirá con el cursor titilando sobre una casilla: con el cursor señale la región de la columna que contiene los datos y acepte con OK. Gnumeric despliega una hoja aparte con muchos datos, entre los cuales están la media, la varianza y la desviación. Ejemplo: sobre los datos propuestos, Gnumeric da una media de 2.55 y una varianza de 0.77. Es buena idea verificar el número de datos, que deben ser 9 y aparecen como *count*.

14.4.4. Las tablas de OpenOffice y Gnumeric

Las tablas de nuestro libro son muy limitadas, por ejemplo, las tablas de la F vienen con significancia 0.01, 0.02, 0.05 y los demás valores se quedan en el misterio. La ventaja inmediata de *Gnumeric* es que nos da valores críticos para cualquier significancia. Esa virtud se aprende a manejar muy fácilmente y hay tres métodos. Primero, se teclean los comandos que se listan a continuación. Segundo: se busca en el menú principal la sección *insert*, luego una va a *function* y allí busca el subconjunto de funciones *statistics* y luego señala la función deseada que lleva el mismo nombre que la función que proponemos teclear. Puede pasar que al teclear números, el programa no los acepte. El remdio es cambiar de modo al teclado de números, lo cual se hace oprimiendo una vez la tecla *Bloq Num*, en la parte superior derecha de dicho teclado. O también se puede ver qué sirve mejor, si una coma o un punto para señalar los decimales. Tercero: se puede usar copiar y pegar. Al hacerlo, *Gnumeric* puede insertar una comilla o un espacio, lo cual se investiga no en la celda donde se teclea sino en el editor de línea, inmediatamente abajo de la barra del menú. Dichas inserciones hay que borrarlas.

1. La z: la instrucción

   =normsinv(1- 0.4)

   da el z-crítico para la significancia $\alpha = 0{,}4$ con la cola superior. En éste caso el z crítico es 1.75068607125217.

   La instrucción

   =1-normsdist(zObs)

   da la significancia del z-experimental con la cola superior, es decir, la probabilidad de encontrar un z mayor que el z del experimento. Ejemplo: la significancia de la cola superior del valor $z = 1{,}75068607125217$ es 0.4.

2. La t depende de los grados de libertad. Para hallar el t-crítico con **dos colas** para la significancia 0.05 y 500 glise teclea

   =tinv(0.05,500)

   lo cual da 1.96471983746779. Vemos que con 500 gls, la t funciona igual que una z.

   La instrucción

   =tdist(1.96,50,2)

   da la significancia del t-experimental 1.96, con 50 grados de libertad y con 2 colas. Gnumeric da como significancia 0.05558087405522, lo cual dice que con 50 grados de libertad, la desviación de la muestra se puede considerar una magnífica estimación de la desvación poblacional y por tanto, la t da casi igual a la z. La instrucción

   =tdist(1.96,50,1)

da la significancia del t-experimental 1.96, con 50 grados de libertad y con la cola superior, que es 0.02779043702761.

3. La F:

   =finv(0.01,7,5)

   da el valor crítico con la cola superior para la significancia 0.01 con 7 gli en el numerador y 5 en el denominador. La respuesta es 10.4555108917609.

   La instrucción

   =fdist(10.4,7,5)

   da la significancia con la cola superior del valor $F_{obs} = 10{,}4$ con 7 gli en el numerador y 5 gli en el denominador. La respuesta es 0.01.

4. La Chi-cuadrado χ^2 Para hallar el valor crítico derecho de la χ^2 con significancia 0.01, 30 gli y la cola superior se teclea

   =chiinv(0.01,30)

   lo cual da 50.892.

   La instrucción

   =chidist(50.892,30)

   da la significancia de la cola superior del valor de χ^2-experimental igual a 50.892 y 30 grados de libertad, la cual es 0.01.

   Para saber el valor izquierdo de la distribución en OpenOffice se teclea la instrucción:

   =CHISQDIST(14.593,30)

   si uno desea saber la significancia de la cola inferior del chi cuadrado experimental 14.593 con 30 gl. La respuesta es 0.01.

   La instrucción

   =CHISQINV(0.01,30)

   permite saber en OpenOffice el χ^2 crítico de la cola inferior con significancia 0.01 y 30 gl. da 14.95.

   Trampa para saber en Gnumeric el chi-cuadrado crítico de la cola izquierda: si la significancia asignada a la cola izquierda es 0.025 y los gli son 8, uno teclea

   =chiinv(0.975,8)

   y da 2.17973074725265.

14.4.5. Un ejemplo con Gnumeric

Decidamos con significancia 0.04 y la cola superior que 2 es la media de los datos siguientes sabiendo que la desviación poblacional es 0.9. Los datos son: 1, 2, 3, 4, 2, 3, 3, 2, 3.

Solución: teclee los datos sobre una columna y active una casilla por ahí cerca en donde Gnumeric pondrá su respuesta. Es muy buena idea ponerle un título, nombre o label a las respuestas de Gnumeric. Para ello, en una casilla cualquiera, ponga el nombre zTestDC que significa test z con dos colas y haga click en la casilla del lado para que sea ahí donde Gnumeric ponga su respuesta. Sobre el menú *Insertar*, elija *Function*, luego elija la categoría *statistics* y luego busque *ztest*. Lea el mensaje que dice que se trata de una prueba de dos colas. Termine con OK. Se desplegará una ventana de diálogo que se rellena así: haga click sobre la casilla titulada *ref*, y a continucación señale la región que contiene los datos. Sobre la casilla x ponga la media de la hipótesis nula, en este caso, 2. Sobre la casilla marcada con *stddev* ponga el valor de la desviación poblacional, 0.9. Termine con OK.

Gnumeric pondrá su respuesta donde uno le indicó, la cual es 0.03. Este valor no es el z-crítico. Este 0.03 es la significancia con dos colas del z-observado Eso significa que el 3 % de la campana de la Z es más extremo con dos colas que el z-experimetal de nuestros datos. Ahora bien, nuestro problema ponía

una significancia de 0.04 en la cola superior. Eso es equivalente a decir que con dos colas la significnacia es 0.08. Como la significancia experimental con dos colas es 0.03, eso quiere decir que para la significancia dada en la H_o, que es 0.08, nuestros datos son extraordinarios y que la hipótesis nula se rechaza: no estamos autorizados a creer que la media poblacional puede ser 2 dado que la desviación poblacional es 0.9. Para poder aceptar la hipótesis nula, se requiere que la significancia experimental,la que arroja el paquete, sea mayor que la significancia dada en la H_o.

Ahora expliquemos en dónde está la falla de Gnumeric: resulta que el valor 0.03 no es la significancia de nuestro z-experimental con dos colas, como el paquete dice, sino con una, la superior. Aunque hay muchas salidas, la más confiable es hacer uno mismo sus cálculos y usar las tablas de Gnumeric.

Para nuestro caso, uno programa Gnumeric para que calcule el $z_{obs} = \frac{\bar{X}_9 - \mu}{\sigma/\sqrt{9}}$. El $\bar{X}_9$ uno lo toma directamente de Gnumeric, aproximadamente 2.55, el μ en la H_o vale 2, la desvación poblacional vale 0.9. El $z_{obs} = 1{,}851$. Para calcular el z-crítico con la cola superior y con significancia 0.04 se copia sobre una celda la instrucción

=normsinv(0.96)

se termina con enter y Gnumeric produce 1.75, lo cual significa que con una cola y significnacia 0.04, el z-crítico es 1.75. El 0.96 viene de ser la cola superior y significancia 0.04 y porque la función $= normsinv()$ da la probabilidad de ser menor que z_o, es decir $\psi(z_o)$. Como el z_{obs} es mayor que el z_c, entonces tenemos un conjunto de datos cuya media se clasifica como anormal, extraordinaria. Rechazamos la H_o.

Hemos dicho que podemos rechazar la H_o de que la media vale 2 con una significancia de 0.04 y la cola superior. Pero podemos preguntarnos: ¿Cuál es la significancia máxima con la cual se puede aceptar la H_o y con la cola superior? Para saberlo marcamos una casilla con SigniMax y al lado insertamos el comando

=1-normsdist(1.75)

terminamos con *Enter* y Gnumeric arrojará el valor 0.03 (el mismo que Gnumeric da para el test ztest pero a nosotros nos da con una cola y Gnumeric con dos). En vez de escribir 1.75, el valor del z_{obs}, uno puede encadenar directamente el resultado que Gnumeric calculó para nuestro z_{obs}. Lo que estamos diciendo es que para aceptar la H_o a la luz de los resultados experimentales es necesario trabajar con una significancia de 0.03 o menor en la cola superior.

Estas mismas tablas de la Z se pueden usar para comparar proporciones y pares de medias cuando uno conoce las desviaciones poblacionales.

Lo que hemos visto para la z, se adapta a los demás estadígrafos, así que sólo necesitamos saber cómo se les invoca. Eso lo podemos ver en la tabla siguiente:

Gnumeric ofrece en su menu Tools herramientas muy útiles para nuestro curso: compara varianzas y puede ejecutar les test de comparación de dos medias en todos los 3 casos, cuando se tienen varianzas iguales, varianzas diferentes y cuando los datos son apareados.

296 *Ejercicios*

Recordemos que para elevar un número al cuadrado, se multiplica el número por sí mismo y para sacar raíz cuadrada de 17 se usa la función

=sqrt(17)

que da 4.12310562561766.

Programe OpenOffice y/o Gnumeric y use sus tablas para resolver los siguientes problemas donde hay 3 conjuntos de datos:

A: 1, 2, 3, 4, 2, 3, 3, 2, 3.
B: 1, 14, 6, 7, 5, 16, 8, 7, 12, 39.
C: 3, 4, 5, 6, 2, 3, 4, 3, 2.

1. Estudio de una media sabiendo la desviación poblacional σ: La media poblacional de A es mayor que 2 sabiendo que la desviación poblacional es 0.9.

2. Comparar dos medias de X, Y sabiendo σ_X, σ_y. La media de C es por lo menos una unidad mayor que la da A, sabiendo que ambas desviaciones poblacionales valen 1.

3. Estudiar una media conociendo la desviación de una muestra al azar. Halle el intervalo de confianza para la media poblacional de A.

4. Estudiar una proporción: la proporción de los mayores que 3.5 en A es igual a 0.2.
5. Experimentos con dos proporciones. La proporción de los mayores que 3.5 en A es igual a la proporción de los mayores que 4.5 en C.
6. Estudio de una varianza. Halle el intervalo de confianza para la varianza de A.
7. Comparación de dos varianzas. La varianza poblacional de A es igual a la varianza poblacional de C.Tome $\alpha = 0{,}04$.
8. Comparación de dos medias sabiendo, por el estudio de las varianzas de las dos muestras correspondientes, que las varianzas poblacionales son iguales: La media de C es 1 unidad mayor que la de A.Tome $\alpha = 0{,}04$.
9. Comparación de dos medias sabiendo, por el estudio de las varianzas de dos muestras, que las varianzas poblacionales son diferentes. La media de C es 10 unidades mayor que la de A.Tome $\alpha = 0{,}04$.
10. Comparación de medias para datos apareados (miopía ojo izquierdo vs. miopía ojo derecho). La media de A es igual a la de C.
11. Resolver con Gnumeric el problema 266 página 248 sobre la independencia de dos factores. Se usa Tools de Gnumeric.

14.5. Taller 5: Investigación.

Consideremos un cambio de opinión creado por una noticia. Podemos tomar decisiones alrespecto usando compración de frecuencias o prueba de homogeneidad de proporciones siempre y cuando los datos sean independientes. Para que los datos sean independientes, las personas consultadas después de la noticia no deben ser las mismas que las consultadas antes de que se diera a conocer. Pero, es también muy natural hacer un estudio de cambio de opinión de la misma persona. En este caso no tenemos datos independientes sino apareados.

Para fijar ideas, consideremos que se discute hacer o no el metro. La noticia que se publicitó ampliamente fue que la ciudad está construida sobre una laguna que con los siglos se fue convirtiendo en un pantano y que después se disecó y que por lo tanto la hace altamente sensible a los terremotos y al deteriodo global por minifallas locales del terreno. Se entrevista a un grupo de personas y se les pregunta su opinión antes y después de que se publicara la noticia. La tabla de datos es la siguiente:

		Después	
		Si	No
	Si	20	30
Antes			
	No	34	24

Interpretación: hay 20 personas que querían el metro antes de la noticia y que aún lo quieren ahora. Análogamente, hay 30 personas que antes querían metro pero que desistieron una vez se pusieron a pensar en los costos astronómicos de mantenimiento.

La forma aceptada de manejo de estos datos se conoce como **prueba o test de McNemar para tablas de contingencia** 2×2 **para datos apareados**. Investigue y desarrolle el tema proponiendo unos 5 ejemplos, uno de cambio de opinión y otros 4 de otras cosas. Por ejemplo, ausencia o presencia de miopía en el ojo derecho vs. ojo izquierdo de la misma persona antes y después de una cirugía laser. Y, ¿qué hacer si quiero considerar diversos grados de miopía?

Capítulo 15

Proyecto

El proyecto tiene como objetivo relacionar el material mínimo de nuestro curso de estadística con la carrera de cada quien. El proyecto es para la mayoría de estudiantes algo excepcionalmente agradable y es común que sea muy bueno, incluso hay trabajos magníficos. Utilizaremos el proyecto para que sirva de repaso. Para ello, hay que inventarse y resolver 6 problemas que relacionen la carrera de cada quién con la temática mínima del curso.

297 ***El material mínimo. Si uno no sabe este material en su totalidad, ¿por qué habría de pasar el curso? Por tanto, se calificará con muchísimo rigor.***

1. De descriptiva: para una lista de datos calcular: media, varianza, desviación, cuartiles, mediana, rango intercuartilar, datos extremos según el criterio del rango intercuartilar y el de las dos desviaciones, boxplot y diagrama de bigotes. Sacar una tabla de frecuencia absoluta y la acumulada asociada, dibujar el histograma, y la tabla correspondiente de frecuencias relativas con su torta.

2. De una binomial: sólo un ejemplo con n grande, digamos 20, el cual se resuelve con calculadora. Cálculo de probabilidades, de la media y varianza por fórmulas, diagrama de barras, envolvente normal. IC de una probabilidad. Estudio de una Ho de una proporción.

3. De la Campana de Gauss: una variante del problema de las camisas, y otra sobre proporción de medias muestrales entre dos valores.

4. `Dos problemas` del diccionario de experimentos para distribuciones normales. Uno de ellos ha de ser: comparación de dos medias cuando la diferencia de medias en la Ho no es cero.

5. Tablas de contingencia: probabilidades diversas. Test de independencia de factores. Reinterpretar como test múltiple de homogeneidad de proporciones.

Lo que uno hace para inventarse un problema es mirar cuál ejemplo o problema del libro le gusta para luego adaptar la telenovela a su carrera. Para que no sea plagio, es menester cambiar tanto el enunciado como todos los datos numéricos del ejemplo o problema.

`El proyecto se hace en grupo, por ejemplo, de 6.` **Se reparten el trabajo y cada uno revisa que los problemas de los demás estén bien**. Ésto es de sumo cuidado, pues el proyecto se califica como una tesis Y CON EXTREMO RIGOR, no por partes sino como un todo: puede obtenerse un 2 de nota definitiva por causa de un error garrafal, tal como:

1. Confundir muestra y población.

2. Decir que una probabilidad sea mayor que uno o negativa.

3. Decir que las probabilidades no sumen uno.

4. Resolver un problema de múltiple igualdad de proporciones con otro de igualdad de varianzas.

5. Confundir un problema de proporciones con otro de medias de datos numéricos.

6. Asignarle a una binomial una varianza a gusto personal.
7. Decir que un único dato tiene varianza no cero.
8. Presentar un proyecto incompleto o incluso un problema a medio hacer, sin la Ho o sin la interpretación.

Los datos pueden ser inventados, o pueden venir de una encuesta diseñada por los mismos estudiantes, o pueden sacarse de Internet o pueden ser algunos inventados y otros reales. Los estudiantes que han elegido hacer su propia encuesta siempre proyectan la impresión de haberse divertido e instruido. Hay derecho a tener máximo un problema mal. Fuera de tiempo no se aceptan proyectos.

Es obligatorio que todos los estudiantes del grupo revisen el material de todos: un buen proyecto es original, es creativo, sus problemas están claramente enunciados, están bien formulados, son profesionales y se atacan con el método adecuado, están bien resueltos y bien interpretados. Un buen proyecto se entrega completo. Cuando hay alguien que no haga su parte completa, se reporta su incumplimiento y le queda cero de nota o de lo contrario todos salen crucificados. Un mal proyecto tiene problemas repetidos, tiene plagios y no todos sus problemas son profesionales:

- **Un problema está repetido** cuando se resuelve por el mismo método que uno anterior del mismo estudiante, así que no le sirve de repaso, que es un objetivo básico del proyecto.
- **Un plagio** puede ser copiarse un problema de un libro, o de un proyecto ajeno de semestres anteriores o de internet. O puede ser cambiar la redacción de un problema de un libro o de una tarea pero conservando algunos números para ahorrarse los cálculos y el problema de inventarse algo coherente. Por eso, en el proyecto hay que mostrar los procedimientos pero está permitido usar Excel o Gnumeric o paquetes estadísticos de gusto personal. Si alguien comete plagio, todos los del grupo son culpables pues **es obligatorio que todos revisen el material de todos**.
- **Un problema es profesional** cuando no pudo haber sido inventado por el profesor de estadística por carecer de los conocimientos específicos de una profesión de ciencias sociales. Por eso, el proyecto se deja para final de semestre cuando ya los estudiantes han visto a profundidad temas propios de su carrera.

298 ***Ejemplo.*** *El tema elegido es la pareja presidente & vicepresidente. Sea X la popularidad del presidente y Y la del vicepresidente que se miden mes por mes.*

1. *Descriptiva: gráfica de la popularidad de una persona a lo largo de un año. Diagrama de tallo de X, Y en el tiempo. Gráfica de la distribución de X de todos los presidentes en todo un continente en el mismo mes.*
2. *Binomial. Universo: el continente en un año. Probabilidad de que haya una trifulca en la pareja en un mes. Probabilidad de que haya una catástofe continental de estabilidad en las parejas de gobierno.*
3. *Campana de Gauss: se dividen los rangos de popularidad: por el suelo, baja, mediana, alta.*
4. *Diccionario de experimentos: el tema se presta directamente para comparación de medias de datos apareados. Si hay una diferencia significativa de popularidad media dentro de las parejas a favor del presidente, uno podría argumentar que el vicepresidente es una figura diseñada (por intención o por corrupción) para amortiguar las caídas del gobierno y así enredar al pueblo. Las diversas varianzas miden estabilidad política que puede ser a lo largo de un período o a lo largo del continente.*
5. *Tablas de contingencia: estudio de dos factores. Primero: la importancia social de los programas de Estado (medida por el número de titulares de prensa positivos que despiertan). Segundo: la estabilidad ejecutiva de la pareja presidencial (medida por el grado de acuerdo en sus declaraciones). Test múltiple de homogeneidad de proporciones: La proporción de personas que consideran que el dúo presidencial forma un buen equipo es homogénea a lo largo de los diversos departamentos del país y además es del orden de 0.8.*

El proyecto puede ser escrito a mano, o puede ser en computador y mandarse por correo, o puede imprimirse.

Capítulo 16

Examen final: con folleto de fórmulas.

Respuestas en la pág 327.

El final es sin libro ni apuntes. Es con calculadora y con el folleto de fórmulas que no puede tener notas personales. Dura dos horas y puede tener 5 puntos con a,b,c. Este examen siempre incluye ligeras variantes de lo conocido. Siempre contiene problemas que se prestan a diversas y equivocadas interpretaciones si uno no lee bien. Hay una dificultad adicional: hay muchos temas y muchos métodos y equivocarse de método produce un cero (como de costumbre). Es pues natural que este examen sea percibido como muy difícil y que infunda miedo. En realidad, algunas personas no alcanza4n a sacar el uno que necesitan para pasar. Por tanto, es imprescindible preparar el final con responsabilidad, para lo cual se ha asignado tiempo suficiente. Lo peor que alguien puede hacer en el repaso es autoprotegerse detrás del grupo de estudio.

Destaquemos dos fuentes comunes de errores y una nota obligante sobre el rigor:

1. La palabra proporción se interpreta como porcentaje o frecuencia relativa, probabilidad y como relación entre varianzas. ¿Cómo saber a qué se refiere? Respuesta: Si la palabra proporción va sola y si es un dato puede referirse tanto a una frecuencia relativa muestral como a una poblacional o probabilidad. Si va sola y es en una pregunta, se refiere a una probabilidad porque casi nunca hacemos preguntas sobre las muestras sino sobre las poblaciones puesto que es a ellas a las que afectan nuestras decisiones y la proporción sobre una población es una probabilidad. Para que la palabra proporción pueda interpretarse como relación entre varianzas poblacionales, debe decir explícitamente eso.

2. Tratar de resolver una binomial por árboles cuando debió resolverse por fórmulas y calculadora. Ejemplo de problema que se resuelve por un árbol de 2 pisos: en una urna hay 3 bolitas azules y dos rojas. Se sacan 2 bolas sin reemplazo. ¿Cual es la probabilidad de que las dos bolas sean del mismo color? Este problema no puede hacerse por fórmulas de la binomial pues al sacar una bolita las probabilidades cambian: es necesario resolverlo por árboles. Ejemplo de un problema que se resuelve por las fórmulas de la binomial y calculadora: De veinte personas sacadas al azar, ¿cuál es la probabilidad de que al menos dos le tengan fobias a los musulmanes si la probabilidad de que una persona al azar les tenga fobia es 0.03? Este problema puede hacerse por la binomial pues como no dijeron el tamaño de la población, uno asume que es grande y que por tanto las probabilidades no cambian significativamente al sacar una persona y después otra. Tampoco dijeron por qué método hacerlo pues no puede hacerse por árboles pues tendría que tener 21 pisos y uno se demoraría 2 años. Así que debe hacerse por fórmulas y calculadora.

3. Para no sobrecargar el material no hemos exigido hasta hoy saber que nuestros procedimientos se basan en teoremas matemáticos que son válidos bajo condiciones dadas que hay que especificar. Pero para este momento, esto ya puede entenderse rápidamente y se exigirá en el final. Comencemos

por entender para qué se exige **muestreo al azar**: significa que obtenemos la máxima fidelidad en la representación de la población y así podremos extraer la máxima información útil para la toma de decisiones que afectará a toda la población. Cuando el muestreo no es al azar se denomina **sesgado** y se hace o por ignorancia o por querer favorecer a un grupo y oprimir al resto. Sigamos con **la binomial**, cara o sello, niño o niña. El desarrollo por árboles o por Pascal es exacto, pero la aproximación por campana requiere que p, la probabilidad de cara, no esté muy cerca ni de cero ni de uno. Para un caso así tenemos otra herramienta llamada la distribución de Poisson, que no hemos visto. Para poder aplicar la binomial y su aproximación a la normal también se necesita que la población no sea muy pequeña, por lo menos 10 veces más grande que la muestra, porque si no, al sacar un elemento pueden cambiar significativamente las probabilidades. **Los 10 experimentos básicos de la normal** son válidos si las distribuciones son normales. Pero, ¿qué pasa si las poblaciones son de tendencia central pero no normales, digamos, algo asimétricas? ¿Se pueden aplicar los teoremas? Respuesta: los experimentos con medias exigen que las medias muestrales del TLC sean normales, pero eso es cierto aún cuando la misma variable aleatoria en cuestión no sea perfectamente normal siempre y cuando la muestra no sea muy pequeña. Para distribuciones casi normales, una muestra de tamaño 5 puede ser suficiente. Pero para distribuciones muy asimétricas, puede necesitarse una muestra de tamaño 70 o más. Nuestros formalismos para **experimentos con varianzas** son celosos y es mejor no ambicionar su aplicabilidad a distribuciones no normales. Para casos anómalos tenemos otra herramienta llamada simulación (ilustrada en el anexo al final del capítulo). Cuando se comparen muestras, cada una debe ser al azar y deben ser independientes para que la información y fidelidad en la representación sean máxima. Como regla general, donde se usan aproximaciones, las muestras no deben ser muy pequñas. Por ejemplo, en las tablas de contingencia los datos por casilla ojalá sean mayor que 5 y todos de muestras independientes. Esto se debe a que la discripencia se aproxima a una chi-2 pero la aproximación se mejora a medida que los glis aumentan. Este tema tan delicado y tan complicado puede explorarse en Internet bajo el título *asunciones y restricciones para el test tal*.

299 ***Versión 1***

1. En las 5 secciones del curso Estadística 1 para Ciencias Sociales, la media de las notas finales de cada sección fueron 3.72, 4.11, 3.45, 3.08 y 4.53 respectivamente.

(a) Encuentre la media y la desviación de estas medias.

(b) Si en los cursos había 25, 22, 33, 19 y 38 estudiantes respectivamente, encuentre la media de las notas de los estudiantes de todos los cursos tomados como un todo.

2. Una baraja consta de 52 cartas, divididas en cuatro palos que son corazones, diamantes, tréboles y picas. Los corazones y diamantes son rojos, los tréboles y picas son negros. Cada uno de los palos consta de 13 cartas: As, 2, . . ., 10, J, Q y K. Se eligen al azar dos cartas **sin reemplazo**. Halle:

(a) La probabilidad de que ambas cartas sean de igual color.

(b) La probabilidad de que haya al menos una K.

(c) ¿Son independientes los eventos *las dos cartas son de igual color* y `haya al menos una K`?

3. Suponga que en una manifestación hay 20 banderas. De éstas, 6 son rojas y 14 son azules. Se eligen al azar dos banderas sin reemplazo.

a) Encuentre la probabilidad de que las dos banderas seleccionadas sean rojas.

(b) Encuentre la probabilidad de que al menos una bandera sea roja.

(c) Encuentre la probabilidad de que a lo sumo una bandera sea roja.

(d) Encuentre la probabilidad de que la primera bandera sea roja dado que la segunda fue azul.

Repita el ejercicio para el caso en el cual las banderas se sacan con reemplazo.

4. El problema en una fábrica de esferos es que el diámetro de las esferitas que van en las puntas de las minas no es homogéneo debido a imperfecciones de la maquiaria, lo cual significa que algunos esferos tienen la bolita muy grande y no dejan salir tinta mientras que a otros se les escurre y manchan todo por tener la bolita muy chiquita. a) Explique por qué este problema técnico tiene que ver tanto con la media como con la varianza. b)Supóngase que el diámetro sigue una distribución normal. Una muestra al azar de 15 esferitas dio una varianza de 0.04. Describa claramente la muestra y la población. c) Halle el IC

de la varianza del diámetro. d)Han ofrecido un nuevo procedimiento que promete disminuír la varianza a la tercera parte pero es muy costoso. El gerente dice: adoptaremos el nuevo procedimiento pagando todo lo que cuesta si eso es cierto. Una muestra al azar de 18 esferitas fabricada con la novedad dio que la varianza del diámetro fue de 0.02. Halle el IC para la relación entre varianzas, procedimiento nuevo vs. viejo. e) ¿La novedad cumple la demanda? Resuelva el problema por intervalo de confianza y por un test F. f) ¿Los IC hallados son muestrales o poblacionales? g) ¿Cuáles valores son los fundamentales a la hora de tomar decisiones: los muestrales o los poblacionales? h) Explique por qué el gerente no adoptó la novedad. i) El vendedor solicita que en vez de 18 esferitas, tomadas al azar, se tomen 50 y que se vea qué pasa con el procedimiento novedoso. La nueva varianza con 50 datos dio 0.015. Halle el IC de la nueva varianza j) Halle el IC de la relación entre varianzas en juego. k) Según el nuevo experimento, ¿cumple la novedad con la demanda? use el IC y también un test F. l) Si uno interpreta literalmente las preguntas anteriores, son con dos colas. Pero el sentido de la innovación técnica es con la cola superior pues se adoptará la novedad si la varianza se disminuye 3 ó 4 ó 9 veces y no se adoptará si la disminuye 3 veces o menos. Use un test F para resolver esta variante. m) Asesore al gerente con un consejo sobre si adoptar o no adpotar la novedad. n) Los espías de la Compañía acaban de reportar que la competencia adoptará el nuevo procedimiento. Reevalúe su decisión.

5. (a) Una variable aleatoria tiene distribución normal con media = 62.4. Encuentre su desviación estándar, si 31 % del área bajo la curva cae a la derecha de 79.2.

(b) El tiempo de uso del correo electrónico por sesión de Internet tiene media = 8 minutos y desviación = 2 minutos. Si se seleccionan muestras de 100 sesiones, ¿qué proporción de medias muestrales estarán entre 7.8 y 8.2 minutos?

300 ***Versión 2***

1.Un antropólogo observa que las mujeres de cierta población aborígen se pintan los labios o los pómulos. Tiene dos teorías para explicarlo. Teoría uno: se pintan para estar bellas y contentas consigo mismas o bien se pintan para atraer a los hombres. Los datos siguientes muestran el número de veces que diversas mujeres son miradas por los hombres dentro de la maloca en el lapso de una semana entre las 5 y 6 de la tarde.

Pómulos pintados de cafe: 87 68 92 79 83 67 71 92 112 75 77 102 79 78 85 75 72

Labios pintados de rojo: 101 100 87 95 98 81 117 107 103 90 100 99 94 94

(a) Represente gráficamente los datos mediante un diagrama de tallo y hojas. Hágalo en el mismo diagrama. Las hojas de la muestra de pómulos pintados de café hacia la izquierda y la de labios pintados de rojo hacia la derecha del tronco.

(b) Para cada grupo calcule la media, mediana y la moda.

(c) Para cada grupo calcule los cuartiles 1 y 3.

(d) Dibuje el boxplot o diagrama de caja mostrando el resumen de los 5 valores.

(e) Halle el rango intercuartilar y utilícelo para hallar los valores extremos de la primera serie de datos. Halle los outliers usando el método de la campana y las dos desviaciones. Compare los dos métodos.

(f) ¿Cuál teoría es apoyada por los datos?

2. Los estudiantes que toman la clase de estadística para Ciencias Sociales en general se demoran en promedio 6 minutos en llenar la encuesta de evaluación del curso. Suponga que el tiempo que toma un estudiante en llenar esta encuesta tiene una distribución normal con desviación estándar de 0.8 minutos.

(a) Calcule la probabilidad de que un estudiante demore entre 5 y 8 minutos en contestar la encuesta.

(b) Halle el tiempo que le tomó a un estudiante responder la encuesta si a la izquierda de este tiempo se encuentra el 13.14 % del área bajo la curva

3. De una muestra al azar de 75 estudiantes se obtuvo en la calificación en el Curso de Estadística una media de 3.55 con una desviación estándar de 0.85.

(a) Construya un intervalo de confianza del 94 % (significancia = 6 %) para el verdadero promedio (media poblacional).

(b) ¿Cuál debe ser el tamaño de la muestra si se desea estimar el verdadero promedio con una confianza del 99 % (significancia = 1 %) y con un error máximo de 0.1?

(c) En las preguntas anteriores, ¿era necesario suponer que la calificación queda bien aproximada por una distribución normal?

4. Considere los datos que aparecen en la tabla correspondiente a una encuesta realizada a 10 personas. Cada persona es clasificada como 0 si la persona tiene a lo sumo 25 años y como 1 en caso contrario. Se registra para cada persona el número de veces que escucha la BBC por semana:

Grupo de edad	0	0	1	1	0	1	1	0	1	1
No de veces que escucha la BBC	0	1	4	3	1	4	4	2	3	2

(a) Para cada grupo de edad halle media y desviación estándar de las veces que escucha la BBC por semana.

(b) Halle una frecuencia relativa que estime la probabilidad de que una persona tomada al azar del grupo de los viejitos escuche la BBC en un día cualquiera. Haga lo mismo con el grupo de los jóvenes. ¿Proporcionan los datos evidencia para afirmar que la probabilidad de que los viejitos escuchen la BBC es mayor que la de los jóvenes? Use $\alpha = 0{,}01$.

(c) Sea p la probabilidad de que los del grupo 0 escuchen la BBC en un día cualquiera. Estime p. Halle el error máximo de estimación de dicha probabilidad usando una confianza del 97 % (significancia = 3 %). Pruebe en el nivel de significancia del 5 % la hipótesis nula $p = 0{,}6$ versus la hipótesis alterna $p > 0{,}6$. Tenga en cuenta que el error máximo de estimación de un estadígrafo (media, varianza, proporción) para una significancia dada es la mitad de la longitud del intervalo de confianza correspondiente.

5. El estudiante a quien llaman el Halcón Dorado se las dio de muy vivo durante el semestre pasado y casi no estudió: la media de las notas en 6 materias del semestre pasado fue de apenas 3.26. Pero debido a su fuerte desprestigio, ha decidido recuperar su honor. En siete materias de este semestre obtuvo un promedio de 3.7. Tome $\alpha = 0{,}02$. a) ¿Podrá decirse que el Halcón mejoró su promedio en el estudio? b) ¿Es la pregunta anterior de estadística descriptiva o de estadística inferencial y por qué? c) ¿Es legal decir que como esta vez se lo propuso y mejoró, entonces puede mejorar cuando se lo propone? d) ¿Es la pregunta c de estadística descriptiva o de inferencial y por qué? e) ¿Qué es lo que hace la estadística inferencial? f) Las notas del semestre pasado tuvieron una desviación de 1.1 y las de éste semestre una desviación de 0.4. ¿Podrá decirse que el Halcón mejoró su regularidad en el estudio y por qué es bueno o malo mejorarla? g) ¿Podrá decirse que el Halcón Dorado puede mejorar su promedio cuando se lo propone? h) ¿Podrá decirse que el Halcón Dorado puede mejorar su regularidad cuando se lo propone?

301 ***Versión 3***

1. En una clase de 100 estudiantes graduados del pregrado, 54 estudiaron matemáticas, 69 estudiaron historia y 35 cursaron matemáticas e historia. Si se selecciona al azar uno de estos estudiantes, halle

(a) La probabilidad de que el estudiante tome matemáticas o historia;

(b) La probabilidad de que no tome ni matemáticas o historia;

(c) La probabilidad de que tome solo historia;

(d) Si se sabe que toma matemáticas, la probabilidad de que tome historia.

2. X es una v.a. de la cual lo único que se sabe es que la media vale 100 y la desviación 8.

(a) Halle la probabilidad de que un valor de X pueda estar por fuera del intervalo (85, 115).

(b) Hallar los valores de X que señalan la frontera de los outliers con una probabilidad total no mayor al 5 %.

3. La mesada (lo que un estudiante recibe de sus papás) se describe por una variable aleatoria con distribución normal con media = 800 y desviación estándar = 110. Hallar:

(a) La probabilidad de que un estudiante tenga una mesada inferior a 1000.

(b) La probabilidad de que el promedio de 30 estudiantes este entre 760 y 830.

(c) La mesada mínima de un estudiante para que él esté por encima del 99,29 %

4. Una muestra aleatoria de 26 abogadas declaran renta con una media de 3.2 y una desviación de 0.7. Otro grupo de 32 abogados arrojó un promedio de 3.5 y una desviación de 1.7. Tome $\alpha = 0{,}02$.

(a) ¿Las varianzas poblacionales son iguales?
(b) ¿La media de los abogados es menor que la de las abogadas?
(c) ¿La varianza de los abogados es 5 veces más grande que la de las abogadas?
(d) ¿La media de las abogadas es 0.6 menor que la de los abogados?

5. Si Usted es muy inteligente, usted debe aumentar su inteligencia a tal grado que parezca del montón. Habiendo dicho esto, el conferencista miró dos veces a todo el mundo y a continuación presentó la siguiente tabla:

Observado: Sueldo vs inteligencia				
Sueldo	Inteligencia			
	X = baja	Y = mediana	Z= alta	Totales
A=Alto	10	15	8	
B = Bajo	13	3	21	
Totales				

a) ¿Puede decirse que la proporción de A es homogénea con respecto a las columnas?
b) ¿Puede decirse que la diferencia entre la proporción de A en X y la de A en Z es de 0.4?
c) ¿Qué motiva el consejo del conferencista?
d) ¿Cómo explica Usted algo tan extraño?

302 ***Versión 4***

1. a)La probabilidad de que un Concejal se lance a la Cámara es de 0.3. De un grupo de 20 Concejales, ¿cuál es la probabilidad de que al menos 3 se lancen a la Cámara? b) Suponga que el 5 por mil de las personas de una gran ciudad están infectadas con VIH. Se eligen al azar 20 personas. ¿Cual es la probabilidad de que a lo sumo dos de ellas estan infectadas?

2. Dentro de un proyecto que busca mejorar la movilidad vial de la ciudad se construyó una gran avenida y se desea saber si los hábitos viales en vías ramificadas han variado. Se tienen 5 vías y el número de carros que ordinariamente transitaban por ellas entre 7 y 9 de la mañana en día laboral eran 120, 230, 55, 450, y 65. Recientemente se hizo una muestra al azar y se obtuvo en las vías respectivas: 160, 190, 83, 625 y 97. ¿Podemos decir que los hábitos viales se modificaron?

3. Se hizo una encuesta de sí o no a personas tomadas al azar para estimar la probabilidad de que una persona crea en el Congreso. De una muestra al azar de 35 personas, 6 respondieron sí creo. Para aumentar la credibilidad del Congreso se instauró un programa de TV en la franja de alta sintonía y por un gran canal. A los dos años se entrevistaron a 40 personas con la misma pregunta. De ellas, 12 respondieron afirmativamente. a) Los críticos dicen: es obvio que la credibilidad siguió siendo la misma. ¿Es verdad? b) La propaganda del Congreso dice: la credibilidad se duplicó. ¿es cierto? c) ¿Quién tiene razón según Usted?

4. La cantidad de muertos por año en ataques terroristas causados por dos grupos A y B tienen medias y varianzas históricas iguales a 22 y 5 contra 15 y 6 respectivamente, por lo que el primer grupo tiene fama de ser el más terrorífico. En el último tiempo, de una muestra al azar de 7 atentados del primer grupo hubo un promedio de 10 muertos, en tanto que en otra de 5 atentados del segundo hubo un promedio de 6 muertos. Aunque parece que los dos grupos se han ablandado, ¿considera Usted que hubo un cambio estadísticamente significativo en la diferencia promedio de muertes?

5. Las sesiones del Congreso ponen a los participantes muy tensos. Se les siguió la pauta a unos Congresistas para relacionar su consumo de cigarrillos con la asistencia a las sesiones:

Observado				
Consumo de cigarrillo	Asistencia			
	X= Nada	Y=Mas o menos	Z=Mucha	Totales
A=Fuman mucho	12	8	14	
B=Fuman poco o nada	13	21	9	
Totales				

Si se toma una persona al azar, halle la probabilidad de que dicha persona sea de

a) A

b) X

c) A y X

d) A o X (unión)

e) A dado X

f) Verifique que $p(A/X) = p(A \cap X)/p(X)$

g) ¿Puede decirse que los eventos A y X son independientes?

h) ¿Puede decirse que los dos factores, consumir cigarrillos y asistencia, son independientes?

i) Estudie la H_o de que la proporción de los que fuman mucho refleja los hábitos y costumbres de cada quién y que la asistencia a las sesiones no tiene nada que ver. Primero: demuestre que esta pregunta ya está resuelta en el inciso anterior, es decir, que se puede resolver con una tabla de contingencia. Segundo: resuelva la misma pregunta, la de las proporciones de los que fuman mucho, usando repetidas veces un test de comparación de proporciones: compare la proporción de los que fuman mucho en X con la de Y, luego la proporción en de X con la correspondiente en Z, y la de Y con la de Z. Tercero: resuelva la misma pregunta usando un test de ajuste a una distribución uniforme, contabilizando la proporción total de los que fuman mucho en relación con todo el mundo. Cuarto: ¿Dan las 3 formas de resolver el problema la misma solución? Explique por qué sí o por qué no.

j) Halle la probabilidad de que un senador tomado al azar fume mucho dado que su asistencia a las plenarias es mucha.

k) Halle la probabilidad de que un senador tomado al azar fume mucho dado que su asistencia a las plenarias es mediana o nada.

303 ***Versión 5***

1. Diga si es verdadero o falso. Si es verdadero, justificar. Si es falso, mostrar el contexto teórico que aclara todo.

   a) Una distribución de tendencia central es normal.

   b) La varianza es el cuadrado de la media.

   c) La desviación es una medida de asimetría.

   d) La dispersión se mide por la varianza y no por la desviación.

   e) Un estadístico o estadígrafo es un valor poblacional.

   f) Un estadígrafo estima un valor poblacional.

   g) Un estadígrafo no puede dar número decimal, pues se trata de algo preciso.

   h) La distribución normal es discreta pues es una aproximación de la binomial.

   i) La distribución normal es contínua pues se aproxima por una campana.

   j) Cuando hay muchos datos la normal se aproxima por una chi cuadrado.

   k) La F se usa para decidir si la normal es con una o con dos colas.

   l) La F se usa para decidir si dos varianzas muestrales son o no iguales.

m) La chi cuadrado sirve para decidir si una varianza es mayor que la otra.

n) Cuando uno tira un dado y después el otro, lo que salga en el segundo dado depende del primero pues por eso fue lanzado de segundo.

ñ) La F sirve para comparar dos proporciones pues la varianza de una proporción es np(1-p).

o) Un error tipo alfa es la probabilidad de que la hipótesis nula se rechace siendo verdadera.

p) En estadística no hay errores de decisiones pues en matemáticas no puede haber errores.

q) Un error tipo beta es uno menos alfa.

r) Cuando un juez libera a un culpable está cometiendo un error tipo alfa.

s) Cuando un juez condena a un inocente está cometiendo un error tipo alfa.

t) Dos proporciónes se comparan con una z pues ambas son femeninas.

u) Se pueden comparar dos proporciones pero no 3 ni mucho menos 4.

v) Cuando dos factores son independientes, $p(A \cap B) = p(A)p(B)$.

w) Para decidir si dos eventos son independientes, se usa una chi-cuadrado.

x) Las tablas de contingencia deciden si dos o más factores son independientes.

y) El criterio del rango intercuartilar para eventos extremos no se puede aplicar a campanas con desviación muy grande pues son muy anómalas.

2. En cada pregunta se dan dos proposiciones. Señalar VV si ambas son verdaderas, FF si ambas son falsas, VF si la primera es verdadera y la segunda falsa, FV si la primera es falsa y la segunda verdadera. Explique el contexto que clarifica la situación.

   a) Si la probabilidad de pasar el examen es 0.9, hoy paso. Si la probabilidad de pasar es 0.01, hoy pierdo.

   b) Si la probabilidad de pasar es 0.5, yo paso y mi compañero pierde. Si la probabilidad de pasar es 0.1, yo paso y mi compañero con toda seguridad pierde, pues yo ya pasé.

   c) Aunque la probabilidad de éxito sea grande, yo no estoy seguro de ganar y, analogamente, aunque la probabilidad de éxito sea muy pequeña yo no estoy seguro de perder. Por tanto, la teoría de probabilidades no sirve de nada.

   d) Toda normal está determinada por dos estadígrafos, la media y la desviación. La chi cuadrado sólo tiene un parámetro, a saber los grados de libertad.

   e) La Z es el mundo paralelo de toda normal y no se distinguen la una de la otra. La transformada-z se llama así pues transforma la Z en una normal.

   f) Una distribución uniforme siempre es discreta. Una normal siempre es continua.

   g) Cuando los gli crecen, la incertidumbre disminuye, por tanto la t debe cerrarse hacia cero (la desviación tiende a cero, la media también). Cuando los gli crecen, la chi-cuadrado se estira hacia arriba.

   h) Cuando los gli crecen, la t se aproxima a la normal. Cuando los gli crecen, la chi-cuadrado se aproxima a la normal.

   i) Cuando los gli crecen, la incertidumbre disminuye, por tanto la F tiene que aproximarse a uno. La significancia es la probabilidad de ser significativo.

   j) Cuando la significancia disminuye, el z crítico superior aumenta. Cuando la significancia disminuye la t crítica superior aumenta.

   k) Cuando la significancia aumenta, el chi-cuadrado crítico superior aumenta pues son proporcionales. Cuando la significancia aumenta, los valores extremos se hacen mayores.

   l) Cuando la significancia disminuye la probabilidad de hacer un error alfa aumenta. Cuando la significancia aumenta, la probabilidad de cometer un error beta disminuye.

   m) Cuando un test ha rechazado la Ho, uno reune más datos y con seguridad la Ho quedará aprobada. Cuando el número de datos aumenta, la incertidumbre disminuye, por tanto, no quedará duda de que la Ho es verdadera.

n) El azar no tiene leyes y por tanto la campana de una normal no tiene nada que ver con el azar. Ente más datos o grados de libertad, la probabilidad de rechazar la H_o en un test-t aumenta.

ñ) La ciencia es la verdad y si algo no está en la ciencia es mentira. La ciencia es la resta de lo que se ve menos lo que espera.

3. Discuta la ley de Pedro Demalas y algunas implicaciones en la toma de decisones: *cuando cae un rayo, me cae a mí,* lo cual significa que lo malo e improbable le sucede a uno, y lo bueno aunque poco probable, a los demás. El equivalente a esta ley en la cultura gringa moderna está muy bien explicado en las leyes de Murphy, de las cuales la más reconocida es: *Si algo puede salir mal, saldrá mal*. Y no importa que sea poco probable.

304 ***Gran repaso parte I***

1. La siguiente es una muestra al azar de las edades en meses de reclutamiento de niños por un grupo armado ilegal:

190	155	175	148	169
161	172	163	174	174
182	149	135	168	149
172	155	170	147	161
181	172	174	167	191

a) Dividir en intervalos adecuados, calcular y dibujar el diagrama de tallo y hojas correspondiente. Sobre el dibujo y sin ningún cálculo estimar la media y la desviación. Usando dichas estimaciones calcular el coeficiente de variación y hacer un juicio sobre el grado de homogeneidad de la muestra.

b) Calcular y dibujar el boxplot mostrando los valores extremos según el criterio del rango intercuartilar.

c) Las mamás que tienen forma de hacerlo sacan a sus niños de la región a los 11 años. ¿Según el boxplot qué tan razonable es ésto?

2. Las niñas altas tienen una mayor probabilidad de reclutamiento. Marcela y sus 3 amigas son todas altas y ya cumplieron los 11 años pero ninguna tiene forma de irse de la región. Si la probabilidad de reclutamiento correspondiente es de 0.3 y en referencia al grupo de las 4:

   a) ¿Cual es la probabilidad de que se lleven al menos a 2? b) Como el grupo de Marcela todo fue reclutado, ellas tienen temor de ser malditas: ¿Cuál es la probabilidad de que eso pase por azar? ¿Son malditas?

3. Los 200 soldados niños reciben un sueldo fijo y es el mismo para todos. Parte se lo gastan y parte lo envían a la casa. Por esta razón las mamás, sus vecinos y familiares no ven en el grupo armado una entidad necesariamente mala sino una fuente de trabajo. La cantidad que giran es una variable aleatoria X que se distribuye normalmente con media 100 y desviación 20. Hay malos hijos, buenos hijos e hijos modelo. Los buenos hijos son importantes porque se espera que también sean buenos hijos de la 'Organización' y como no son perfeccionistas será más fácil involucrarlos en 'misiones especiales'. Decir la cantidad de niños que son buenos hijos sabiendo que ellos giran entre 75 y 135.

4. Al llegar al cuartel todos los niños saben que vinieron a sufrir. Con todo, les llega el momento en que se rompen y les da tremenda lloradera. Eso ocurre de modo muy diverso, como una distribución de tendencia central con media 22 días y desviación 5 días. El grupo de Marcela y sus 3 amigas se rompieron en promedio a los 28 días por lo que ellas se creen muy valientes. La jefe dijo: ¡Igual de flojas a todo el mundo! Con una significancia del 5 % decidir si la Jefe está desprestigiando al grupo.

5. Los niños combatientes en su generalidad adquieren el síndrome de Ulises descrito en la Ilíada, el cual consiste en la urgencia de vivir bajo fuerte tensión que implique la necesidad de ser un héroe. Esa es la razón de reclutar niños. Debido al síndrome adquirido, muchos de estos combatientes cuando ya son versados en la guerra buscan ambientes más conflictivos y mejor pagos fuera del país, más que todo en África y en el Medio Oriente. Una muestra al azar de 7 combatientes que emigraron dió una edad promedio de 27 años. Si la desviación poblacional de dicha edad es de 3 años, hallar el intervalo de confianza del 95 % de su media poblacional. Usar cualquiera de los dos procedimientos (gráfico u algebraíco) pero mostrar de dónde sale todo.

305 ***Gran repaso parte II***

1. Al contrario de todos sus hermanos que tenían todos los ojos azules, ella nació con ojos verdes. La llamaron Esmeralda. Ahora tiene 11 años. Su trabajo en una mina de esmeraldas en Boyacá es sacar el barro de la tierra y revisarlo para ver si contiene una gota de aceite (una esmeralda de alto valor) pero que sea lo suficientemente grande para que su familia salga de pobre y ya no tenga que venir a embarrarse en la oscuridad. De hecho, sólo regresan al socavón cuando se les acaba la plata. A propósito, el sacavón donde van está tan gastado que ellos son los únicos que lo visitan pero al menos no viven con temor de que los maten. Esmeralda es la niña mimada de todos porque ella es la de la buena suerte. Es algo indiscutible según lo dicen todos pues se enguaca (encuentra una piedrita que les permite comprar mercado para dos semanas) rapidito. De una muestra al azar de 8 veces que fueron al socavón sin ella, 2 veces volvieron contentos. Pero de 23 veces que fueron con ella, 19 veces vinieron contentos. ¿Puede decirse que la proporción de regresar contentos con ella es el triple del caso de ir sin ella?

2. Agustín es un niño indígena en la Guanía que sirve de soporte en la minería del coltán, mezcla de columbio (modernamente llamado niobidio) y de tantalio. Son elementos claves para la tecnología de estado sólido que se utiliza en todo celular por lo cual la demanda es muy alta y además son materiales escasos, de hecho se les conoce como tierras raras. En una palabra, son minerales estratégicos, asociados hoy en día a la mucha plata, a los asesinatos y a la puja por el poder. En su región antes dominaba el grupo de los Guapos pero ahora son los Ayuelos los que mandan. A Agustín le parece que ambos grupos son iguales de peligrosos por lo impredecibles. Este término tan patético tiene una componente principal ligada a la mera presencia de los grupos, que infunden mucho terror, los cuales aparecen de cuando en cuando pero de manera muy aleatoria. El tiempo de espera en días entre una visita y la siguiente para una muestra aleatoria de 9 visitas de los Guapos tuvo una varianza de 120 pero los Ayuelos para una muestra al azar de 7 visitas la varianza fue de 140. ¿Puede decirse que las varianzas son iguales y que sus politicas de dominio por la sorpresa son similares para los dos grupos? (Atención: la varianza no es un descriptor que sea una medida trasparente y fiel del elemento sorpresa pero sí puede decirse que si no hay varianza en el tiempo de espera a la próxima visita, no hay sorpresa.)

3. Lolita aprendió a sacar platino de un afluente del río Atrato desde cuando era bebé. Ella se iba con su mamá, primero a acompañarla, después a jugar a la minería y a lo último para sostener el hogar. Hace 2 años, cuando cumplió los 15 resolvió casarse pues se dió cuenta de que su trabajo era valioso y que podía mantener un marido. Pues sucede que en su cultura cuando una mujer se casa el matrimonio le incluye mantener un marido borracho que se tira en su vicio incluso la plata del mercado. Un muestreo al azar de lo que el marido se gastaba en vicio por semana el año pasado fue: 3, 4, 3, 2, 5, 3, 2. Para este año ella sigue con su misma cantaleta pero además le reza la novena a Santa Lucía, la santa de los ojos, para que se los abra al marido y se de cuenta de lo mal que se está portando. Una muestra al azar de lo que se ha gastado este año dió: 3, 2, 1, 2, 3, 2, 5. a) Hallar las medias y varianzas de cada muestra. b) ¿Considera Usted que ha habido un cambio favorable en el promedio de lo que el marido se gasta en vicio?

4. El mono Núñez, de 12 años, está muy orgulloso de sus carteles en los cuales reclama cese a la minería y sí al agua en su región en los cerros entre Bogotá y Choachí. Lo que sucede es que en minería se utilizan explosivos, los cuales remueven las grietas de la tierra y por ahí se filtra el agua quizá en su totalidad. Es algo tormentoso que de un día al otro a uno se le seque la quebrada. El papá le ha dicho que tiene que ser más variado pues todos los carteles parecen lo mismo. Después

de una larga conversación llegaron a un acuerdo: los carteles deben tener diferentes elementos (árboles, casas, niños, vacas, ovejas,...). Antes, el número de elementos por cartel era en promedio 4 con desviación de 1. Puesto que había carteles con 2 elementos y otros con 3 y otros con 4 y hasta con 5 al niño le parecían muy variables. Ahora, después de la conversación, en una muestra al azar de 8 carteles el promedio del número de elementos es de 7 con una desviación de 3. a) ¿Considera Usted que ha habido un cambio en la varianza? b) ¿Puede decirse que el niño aprendió a subir la media?

5. Oro. Hay que encontrar la veta, seguirla dentro de la roca, extraer el mineral, macerarlo, lavarlo, atraparlo con azogue (mercurio), purificarlo del azogue a puro fuego, cambiarlo por armas, equipaje y comida. El Yonaz vive en cercanías del Nudo de los Pastos pero del lado del Putumayo. Él ya conoce todo el ciclo pues desde que fue reclutado a los 14 años por el grupo rebelde al que pertenece se declaró pacifista e inmediatamente lo mandaron a la mina. La minería es políticamente importante para el grupo pues disminuye la necesidad de secuestro y extorsión. Los niños mineros se dedican más que todo a extracción y maceración pues la refinación es cosa de gente grande, experimentada y de muy buen pulso. La proporción de tiempo que se dedica a cada actividad depende de factores externos como la situación militar en la localidad la cual decide si habrá o no gasolina para alguna maquinaria de ayuda. En la siguiente tabla se muestra la incidencia de la situación militar (favorable, neutra, desfavorable) sobre la distribución de personas trabajando bien sea en extracción o en maceración. Demuestre o refute que la proporción de maceradores es homogénea con respecto a la situación militar:

Observado: Tipo de actividad vs Situación militar				
	X (Favorable)	Y (neutra)	Z (Desfavorable)	Totales
A: Extracción	8	4	2	
B: Macerado	2	3	5	
Totales				

6. Gracias a las regalías del carbón de la Guajira construyeron una escuela a donde va Maya casi todos los días. Su nombre significa dulzura maternal y ya tiene 8 años. A veces no va porque como no siempre dan refrigerio, pues no tiene a qué ir. De los 30 días del mes pasado 18 dieron refrigerio. De los mismos 30 días ella fue 14 veces. Coincidió que cuando ella fue, hubo refrigerio 10 veces. a) Estime la probabilidad de refrigerio. b) Estime la probabilidad condicional de que Maya reciba refrigerio dado que va a la escuela. c) ¿Se justifica el desánimo de Maya por los refrigerios fallidos o es que ella tiene un problema distinto? Tenga en cuenta que ella se demora 2 horas desde su ranchería hasta la escuela y otras 2 ó 3 para volver.

16.1. Anexo: la media muestral de distribuciones NO normales

Para que la media muestral tenga una distribución normal no se requiere que la distribución inicial sea normal. Pero se requiere un muestra de tamaño suficiente, lo cual depende del grado de asimetría. Para estudiar el efecto de la simetría de la distribución original sobre la silueta de la media muestral ofrecemos el siguiente programa en Java. La tabla que es producida por el programa puede copiarse y pegarse en Excel, OpenOffice o Gnumeric para hacer la gráfica y juzgar si su silueta es normal o nó. También puede utilizarse la tabla para hacer un test de Kolmogorov-Smirnov de normalidad. Puede aprenderse Java aplicado a esta temática estudiando el Volumen V de www.evoljava.com.

```
/*
 *  Programa en Java para estudiar la influencia
    del tamaño de la muestra y
    de la simetría de una distribución X
    sobre la convergencia hacia la campana normal estandar
    de la media muestral de X con n elementos.
    Usamos un modelo:
    Se parte de una distribución
```

```
    que sólo toma dos valores 0 y 10,
    0 con probabilidad p
    y 10 con probabilidad 1-p.
    Cuando p está cerca de 0.5, la distribución es simétrica.
    Cuando p está cerca de 0 o de 1, es muy asimétrica.

    Cuando la distribución es ligeramente asimétrica,
    un tamaño de la muestra de 5 puede ser suficiente
    para que la media muestral luzca como una campana.
    Pero cuando es muy asimétrica, p = 0.1,
    con una muestra de 70 elementos uno ya podría
    conceder que la media muestral  sigue una
    distribución normal.

    La simetría se regula con p,
    el tamaño de la muestra con n.
    Cuando se corra el programa, se grafica en
    Excel, OpenOffice o Gnumeric
    una tabla de frecuencias absolutas
    que aparece al final del texto de resultados.
    Así uno puede saber de una sola mirada si
    la media muestral ya sigue o no
    una distribución normal.


 */
package tlcforstatistics;

import java.util.Random;


public class TLCforStatistics
{

    static Random  r = new Random();

    public static void main(String[] args)
    {
        //Tomamos un intervalo de 0 a 10.
        //Probabilidad de 0 es p,
        //la de 10 es 1-p,
        //la de los demás valores es cero.
        double p = 0.4;
        //Tamaño de la muestra
        double n = 5;

        int tablaFrecAbsoluta[] = new int[1001];
        for(int k = 0; k < 20; k++)
            tablaFrecAbsoluta[k] = 0;


        for(int repeticion = 0; repeticion < 3000; repeticion++)
        {
            double sum = 0;
            //La media muestral de n elementos es
```

```
            for(int i = 0; i < n; i++)
            {
                //Número al azar
                double m = r.nextDouble();
                //Si m < p cae cero y no suma nada
                if  ( m < p ) ;
                //Si m no es cero, cae 10 y se suma 10
                else sum = sum + 10;

            }
            //La media muestral es la suma recibida entre n
            double mediaMuestral = sum/n;

            //Contabilidad para llenar la tabla de
            //frecuencia absolutas
            System.out.println(repeticion + " Media muestral = "
                    + mediaMuestral);
            for(int k = 0; k < 1001; k++)
            {
                double limInf =  k * 0.1;
                double limSup = (k+1) * 0.1;

                if ( (mediaMuestral >= limInf)
                        && (mediaMuestral < limSup))
                    tablaFrecAbsoluta[k]++;

            }


        }
        //Impresión de la tabla.
        System.out.println("\nTabla de frecuencias absolutas,"
                + "\npegar en Excel, OpenOffice o Gnumeric "
                + "\ny graficar en XY o scatter mode");
        double x = 0;
        for(int k = 0; k < 1001; k++)
        {
            if (tablaFrecAbsoluta[k] > 0)
            {
                x = k;
                x = x/10;
                System.out.println(x + "\t" + tablaFrecAbsoluta[k]);
            }
        }
    }
}
```

Respuestas Parte II

Problemas del Taller III, pág. 191

Pág 191 ; No 1a. $z_{obs} = 0,90$; H_o se acepta.
b. $z_{obs} = 45$; H_o se rechaza.
Pág 191; No 2a. $z_{obs} = 2{,}06$; H_o se rechaza.
b. $z_{obs} = 22{,}45$; H_o se rechaza.
Pág 191; No 3a. $t_{obs} = 2{,}27$; gli = 7; $t_c = 2{,}365$ H_o se acepta.
b. $t_{obs} = 22{,}45$; gli = 13; $t_c = 2{,}650$ H_o se rechaza.
Pág 192 ; No 4a. $z_{obs} = -1{,}58$, $z_c = \pm 1{,}8$; H_o se acepta.
b. $z_{obs} = -1{,}71$, H_o se acepta.
Pág 192; No 5a $t_{obs} = -0{,}61$; $t_c = \pm 2{,}043$ con 30 gli, $\alpha = 0{,}05$; H_o se acepta.
b. $t_{obs} = -1{,}58$; $\alpha = 0{,}01$, cola inferior; $t_c = -2{,}4$ con gli = 59 (se puede usar la z), H_o se acepta.
Pág 192 ; No 6a. $\bar{D} = -1{,}0$ $s = 2{,}1$ $t_{obs} = -1{,}17$, gli = 5, $t_c = 2{,}571$; H_o se acepta.
b. $\bar{D} = 2{,}44$ $s = 2{,}4$ $t_{obs} = 3{,}05$, gli = 8; $t_c = 2{,}306$; H_o se rechaza.
Pág 192; No 7a. $F_{obs} = 1{,}66$; gln = 8; gld = 13; $F_c = 4{,}30$; H_o se acepta.
b. $F_{obs} = 3$; gln = 8; gld = 13; $F_c = 4{,}30$;H_o se acepta.
Pág 192; No 8a.
Ho: las medias poblaciones son iguales.
Ha: la media poblacional de X menos la media poblacional de Y es diferente de cero,
es decir, la media poblacional de X es diferente de la media poblacional de Y.
μ_1= 35; $s_1 = 8$; m = 21; μ_2= 65; $s_2 = 7$; n= 17; μ_1 - $\mu_2 = 0$;
Parte uno: varianzas:
Fobs = 1.3; Fc = 3.26 con $\alpha = 0.02$, dos colas, gln = 20; gld = 16
Varianzas iguales.
Parte 2 medias:
$s_c = 7.57$ t_{obs} = -12.144 gli = 36.0 $tc = \pm 2.3$ con alpha = 0.02 y dos colas y 36gl. La tc tiene signo más o menos, pues se trata de dos colas.
Región de aceptación de la Ho: desde -2.3 hasta 2.3.
Veredicto: Medias distintas.
No 8b.
Ho: las medias poblaciones son iguales.
Ha: la media poblacional de X menos la media poblacional de Y es mayor que cero,
es decir,
la media poblacional de X es mayor que la media poblacional de Y.
μ_1= 123; $s_1 = 12$; m = 6; $\mu_2 = 85$; $s_2 = 13$; n= 15; $\mu_1 - \mu_2 = 0$;
Varianzas:
Fobs = 1.17, La F crítica se mira con dos colas, porque se estudia la igualdad de varianzas.
$F_c = 4.62$, con dos colas, con $\alpha = 0.10$, gln= 14, gld = 5,
Varianzas iguales.
Medias:
$s_c = 12.74$ $t_{obs} = 6.17$ gli = 19.0
t_c= 1.729 por ser la cola superior, $\alpha = 0.05$, 19gl.

Veredicto: con una confianza del 0.05, es correcto creer que la media poblacional de X es mayor que la de Y.

No 8c.

Ho : las diferencia de medias es cero.

Ha: la diferencia de las medias poblacionales es menor que cero,

es decir,

la media poblacional X es menor que la media poblacional Y

μ_1= 23; s_1 = 4; m = 34; μ_2= 42; s_2 = 6; n= 28; μ_1 -μ_2 = 0;

Varianzas:

Fobs= 2.25;

La F crítica se mira con dos colas, por tratarse de la igualdad de varianzas.

Significancia = 0.02 (el doble de la de una cola)

Fcrítica = 2.5 aproximadamente, con dos colas, α = 0.02, gln = 27,

gld = 33

Las varianzas son iguales.

Medias:

sj = 5.0 t_{obs} = -14.89 gli = 60.0

Tcrítica = -2.326 (el menos es por ser la cola inferior); alpha = 0.01; gl=60;

Región de aceptación de la Ho: desde -2.326 hasta infinito.

Región de rechazo de la Ho: desde menos infinito hasta -2.326.

Veredicto: se rechaza la Ho: creemos que hay un efecto sistemático que hace que la media poblacional de X sea menor que la de Y.

Pág 193; No 9a. $F_{obs} = 16$, gls = 16 y 20, $\alpha = 0{,}02$, $F_c = 3{,}05$, Var diferentes. Comparación de medias: $t_{obs} = -3{,}771$; gli = 17; dos colas, $\alpha = 0{,}02$, $t_c = 2{,}567$; H_o se rechaza. Medias diferentes.

b. $\bar{X}_6 = 123$, $S_X = 12$, $F_{obs} = 100$, gls = 5 y 14, dos colas, $\alpha = 0{,}02$, $F_c = 4{,}7$, Var diferentes. Comparación de medias: $t_{obs} = 7{,}41$; gli = 5.0; cola superior, $\alpha = 0{,}01$, $t_c = 3{,}365$; H_o se rechaza.

c. Var diferentes, $t_{obs} = -7{,}21$; gli = 28.73; cola inferior, $\alpha = 0{,}05$, $t_c = 1{,}699$; H_o se rechaza.

Pág 194;No 20. $F_{obs} = 2$; $F_c = 4{,}56$ con $\alpha = 0{,}05$ y dos colas, 15gli numerador y 7 gli denominador. Varianzas iguales. Comparación de medias: $t_{obs} = 7{,}96$, $t_c = 2{,}8$ con 22 gli y $\alpha = 0{,}05$. Medias diferentes, el paciente va mejorando, pero eso no significa que esté bien. Como la arritmia = varianza aún es muy alta, que se quede otros dos días y despúes veremos.

Pág 194; No 21. Caso 1. Cola superior. Zobs= -3 ; Zcrít= -1,7 α = 0,05;

Pág 194; No 22. Caso 7 con dos colas + 8 ó 9. Fobs= 4; Fcrít= 6.23; α = 0.02, dos colas; gln=2; gld=16 ; Se acepta la Ho; Varianzas iguales. s_c= 2.309;

tobs= -4.42 tcrít= 2.552; α = 0,02, dos colas; gli = 18 ; Se rechaza la Ho; Hubo afectación.

Pág 194; No 23. Caso 7 con dos colas + 8 ó 9 con dos colas. Fobs= 2,25; Fcrít= 2.75; α =0.02; m-1=12 (f); n-1=34; Se acepta la Ho; s_c= 2,303;

tobs= 2.663; tcrít= 1.96; gli = 46 ; Se rechaza la Ho;

Pág 194; No 24. Caso 7 con dos colas + 8 ó 9 con dos colas. Fobs= 400; Fcrít= 5.8; α =0.02; m-1=6 ; n-1=9 ; Se rechaza la Ho;

tobs= 7.407; tcrít= 2.447; α =0,05; gli = 6 ; Se rechaza la Ho;

Pág 195; No 25. Caso 5 subcaso 2. La interpretación literal es con dos colas, pero la de la intriga es con una pues lo que lo uno quiere saber es si el mal ejemplo se propaga. σ_J = 0,176; tobs= 2.176; tcrít= 2; dos colas, α =0.05; Se rechaza la Ho;

Pág 195; No 26. Caso 7 con dos colas + 8 ó 9 con una cola. Fobs= 2.25;

Fcrít= 5.35; α =0.02; m-1=9 ; n-1=9 ; Se acepta la Ho;

s_c= 2.55;

tobs= 2.632; tcrít= 1.734; α =0.05× 2=0.1; gli = 18 ; Se rechaza la Ho;

Pág 195; No 27. Caso 7 con dos colas + 8 ó 9 con una cola. Fobs= 1,515;

Fcrít= 11; α =0.02; m-1=5 ; n-1=5 ; Se acepta la Ho;

s_c= 14.577;

tobs= -1.545 tcrít= -1.812 α =0.05×2=0.1; gli = 10 ; Se acepta la Ho;

Pág 195; No 28. Caso 7 con dos colas + 8 ó 9 con una cola. Fobs= 1.563;

Fcrít= 18; α =0.02; m-1=2 ; n-1=4 ; Se acepta la Ho;

s_c= 4.359;

tobs= 1.57; tcrít= 1.943; α =0.05×2=0.1; gli = 6 ; Se acepta la Ho;

Pág 195; No 29. Caso 3. $H_0 : \mu = 900$, se rechaza con la cola superior. $\bar{X}_6 = 1200$, $s = 50$, $t_{obs} = 14{,}6969$, gli = 5, $\alpha = 0{,}05$, $t_c = 2.015$. Se rechaza la H_o. La Madre Rusia se merece un te caliente con galletas y mermelada.

Pág 195; No 30. Caso 7 con dos colas + 8 ó 9 con una cola.

Pág 196; No 31. La expresión *último tiempo* puede interpretarse como algo poblacional, caso 4. Pero también como algo muestral pues puede ser un lapso corto, caso 5: Ho: p-r = 0, se rechaza con una cola pues se trata de ofrecer algo nuevo, es decir algo mejor, cola superior. f-g= 0.009, alfa = 0.05, pGorro = 0.034, sBi = 0.027, zObs = 0.3298, zc = 1.6. Se acepta la Ho: no hay datos suficientes para pretender que la nueva orden ofrezca algo mejor al pueblo.

Pág 196; No 32. Caso 10. $\bar{D} = 3{,}75$ s=1.58; $H_o : \Delta = \mu_1 - \mu_2 = 1$, se rechaza con la cola superior. $t_{obs} = 4{,}92$; $t_c = 1.8914$; gli = 7; Se rechaza la la Ho; es cómodo creer que hubo mortalidad infantil.

Pág 196; No 33. Caso 10. Datos apareados pues se tratan de los días de la semana día por día. $t_{obs} = 0{,}394$, $t_c = 1{,}943$ con gli = 6 y $\alpha = 0{,}05$ y una cola, la superior. Medias iguales, no hay necesidad de que los demás panaderos se fatiguen innecesariamente.

Pág 196; No 34. Caso 4 $p = 0{,}25$, $f = 0{,}95$, $z_{obs} = 9{,}83$ con $\alpha = 0{,}05$ y una cola, la superior, $z_c = 1{,}64$. La aparición de la Virgen cambia las proporciones a favor de sus cuadros. Sin embargo, eso no implica un incremento en las ventas.

Pág 196; No 35. Caso 7 con dos colas + 8 ó 9 con una cola.

Pág 196; No 36. Caso 7 con dos colas + 8 ó 9 con dos colas.

Pág 197; No 37. Caso 1 pues hay una desviación poblacional, dos colas.

Pág 197; No 38. Caso 5 subcaso 2: x= 34, m = 100; y = 59.2, n = 80; pGorro = 0.51; s = 0.07; tObs= 5.53; zc = 1.96 con dos colas y $\alpha = 0.05$. Hubo cambio, jefe más cercano a sus gente.

Pág 197; No 39. Caso 7 con dos colas + 8 ó con una cola.

Pág 197; No 40. Caso 7 con dos colas + 8 ó 9 con una cola.

Pág 197; No 41. Caso 4 con $p = 0$ y $f = 5/9$, una cola.

Pág 198; No 42. Caso 5 subcaso 2. El enunciado es ambiguo: una o dos colas, ambas quedan bien.

Pág 198; No 43. Caso 10. $H_o : \bar{\Delta} = 0$. $t_{obs} = -3{,}4262$. Dos colas, $t_c = \pm$ 2.3553 con 7 grados de libertad. Sí hay diferencia promedio entre el número de leyes civiles y religiosas y la gente es demasiado complicada.

Pág 198; No 44. Caso 7 con dos colas + 8 ó 9 con una cola.

Pág 198; No 45. Caso 5 subcaso 2, una cola.

Pág 198; No 46. Puede hacerse por coeficiente de variación. O bien por el caso 6 y con dos colas.

Pág 198; No 47. Caso 7, una cola.

Pág 198; No 48 a). Coeficiente de variación. Para las niñas: media 13.4, var= 1.38, s = 1.17, C = 0.08, cerca de 0.1, podemos considerar la población medianamente homogénea pero no homogénea. **b)** Para los varones: media = 16.6, var = 2.71, C = 0.09, edad medianamente homogenea.

c) Caso 7 con dos colas + 8 ó 9 con dos colas. Primero comparamos varianzas: $F_{obs} = 1.97$, $F_c =$ 4.02 con dos colas, y 9 gli en el numerador y 9 en el denominador. Varianzas iguales. Comparación de medias con dos colas: $H_o : \mu_1 = \mu_2$, $\alpha = 0{,}05$. sj = 1.43; $t_{obs} = 2.97$, $t_c = 2{,}1$ con dos colas y 18 gls. Medias diferentes.

Pág 199; No 49. Caso 3, una cola, la superior.

Pág 199; No 50. Caso 5. Pero como la primera muestra es muy grande (9 x 12 = 108 meses) puede tomarse como poblacional, caso 4, una cola, la superior.

Pág 199; No 51 a). Pueblito1: Media1 = 19.5, var1 = 8.28; s1 =2.87; coeficente de variación = 0.14, mediano, población medianamente homogénea. Pueblito2: Media2 = 13.44, var2 = 1.27; s2 =1.13; coeficente de variación = 0.08, mediano, población medianamente homogénea. b) Comparación de varianzas: FObs = 6.48; Fc = 4.54 con dos colas, $\alpha = 0{,}05$. Varianzas diferentes. Comparación de medias con una cola, la superior: tObs = 5.58; tc = 1.8 con cola superior y 9 gl. Se rechaza H_o: en el primer pueblito había más matas que en el otro.

Parcial 3, versión 1, pág. 202.

1a) $f = 0{,}478$, $g = 0{,}764$, Sí puede decirse que la frecuencia relativa de positivos del político B es mayor.

1b) Hay dos muestras en el problema, una de 23 personas y otra de 17.

1c) La población no es 40 = muestra total. La población es el conjunto de personas afectadas por la decisión que se tome. En este caso, todos los habitantes de la ciudad C saldrían afectados por la decisión. Esa sería la población.

1d) La pregunta a es de estadística descriptiva pues describe lo registrado, lo que se ve.

1e) La pregunta de si hubo o no un cambio en la opinión se refiere por derecho legítimo a la población, a todos los habitantes de la ciudad. Por tanto, no es acertado decir que como esta vez el político B obtuvo una mayor frecuencia relativa de positivos, entonces puede hacer lo mismo cuando quiera y que el causa un cambio de opinión en toda la población. La razón es que no se sabe si el aumento de la frecuencia relativa observada sea una casualidad del muestreo hecho o si es algo verdadero de toda la ciudad.

1f) La pregunta e es de estadística inferencial pues involucra una extrapolación de lo que se ve en la muestra a toda la población.

1g Lo que hace la estadística inferencial es decidir si lo que se ve en la muestra total amerita la misma conclusión en la población. Para ello se filtra el efecto del azar, en este caso dado por los muestreos que por ser finitos son inexactos y tergiversan los valores poblacionales. Para filtrar el efecto del azar se mide la discrepancia entre lo que se ve y lo que se cree en la H_o. Si la discrepancia es grande, se repiensa lo que se cree y se rechaza la H_o. Si es pequeña, se acepta lo que se cree en la H_o. Para decidir si es grade o pequeña se compara la discrepancia observada con los valores críticos dados por la tabla correspondiente.

1h) Comparación de la diferencia de dos proporciones experimentales con la diferencia de dos probabilidades.

$f = 0{,}478$, $g = 0{,}764$, $H_o : p - r = 0$. Dos colas, $\alpha = 0{,}05$. $\hat{p} = 0,6$;

$t_{obs} = -1,84$, con 38 glis por lo que se aproxima a una z con $z_c = 1{,}96$, proporciones iguales. Es sensato creer que la diferencia observada de frecuencias relativas es un efecto del muestreo, es decir, es algo que se dió ahora y que no se repetirá con elevada frecuencia en otros muestreos independientes y al azar y que por tanto no se trata de un efecto repetible y válido para toda la población.

1i) Para que el político B pueda probar que él si causó un cambio de opinión sólo le queda un camino: aumentar el tamaño del muestreo al azar pues entre más grande sea, más fiel es a la población.

2) Datos apareados, cola superior: $\bar{D} = 10.429$; s=2.225; tobs= 12.37; tcrít= 1.9 α =0,05; gli = 6.

Se rechaza la Ho, medias diferentes;

3a) Comparación de medias. Primero se comparan varianzas. Ho: varianzas iguales, se rechaza con dos colas. Fobs= 250000; α =0.05; m-1=8; n-1=14; dos colas, Fcrít= 3.29; Se rechaza la Ho: varianzas poblacionales diferentes. Comparación de medias: Ho: medias iguales, se rechaza con la cola superior. tobs= 2.4; tcrít= 1.85; una cola, α =0,05; gli = 8; Se rechaza la Ho: medias diferentes: en el último año llovió menos en promedio.

b) Defensa:

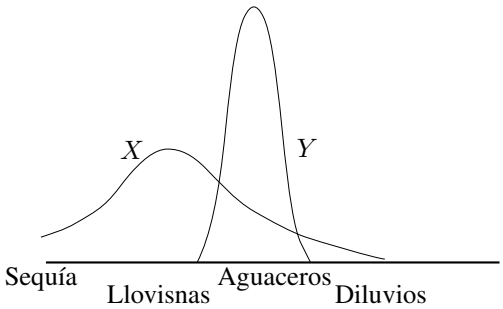


Figura 16.0. El gobernador anterior enfrentó un régimen de lluvias Y dominado por aguaceros de regular calibre (media grande), sin poder destructivo y muy homogéneos (varianza pequeña). El gobernador actual enfrenta un tiempo X dominado por llovisnas (media pequeña) pero con aguaceros diluviales muy esporádicos y de enorme poder destructivo (gran varianza pues los datos de precipitación están muy dispersos desde sequía hasta diluvio). Por eso se ha invertido mucha plata en reparaciones.

Parcial 3, versión 2, pág. 202;

1) Comparación de medias. Primero se comparan varianzas. Decenio pasado: $\bar{X} = 6{,}12$, s^2=2.41; Presente decenio: $\bar{Y} = 11{,}33$, s^2=4; Comparación de varianzas: Fobs= 1,65; Fcrít= 6,84; α =0.02 y dos colas; gln = 8 ; gld =7 ; Se acepta la Ho: vars iguales; Comparación de medias: s_c= 1.8; tobs= -5.9; tcrít= 2.6; α =0.02; dos colas, gli = 15. Se rechaza la Ho, medias diferentes.

2) Mes de mayo: $n = 27$, media = 2.92, var = 1.84. Mes de junio: $n = 32$, media = 3.53, var = 2.19. Comparación de varianzas: Fobs = 1.19, con $\alpha = 0{,}05$ y dos colas $F_c = 2.18$ para 31 y 26 gl. Varianzas iguales. Comparación de medias: $s_c = 1.425$, tobs = 1.62, gli = 57.0, con $\alpha = 0{,}05$ y una cola $t_c = -1{,}645$ medias iguales.

3) Comparación de dos proporciones. Contamos EN AMBAS MUESTRAS EL MISMO mismo MiSmO EVENTO, que el error sí influirá (no se pueden comparar caras con sellos sino que se comparan caras con caras ó sellos con sellos). $f = 8/30$, $g = 20/23$, $\widehat{p}$= 0,528; tobs $\sim$ Zobs= 4,37; Zcrít= 1.96; α =0.05 y dos colas; Se rechaza la Ho: proporciones diferentes.

Parcial 3 versión 3, Pág 203

1a) Comparación de medias con cola superior y significancia 0.05. Primero comparamos varianzas con dos colas. Tomamos la misma significancia pues asumimos que los dos tests, el de medias y el de varianzas, son independientes y no importa que el uno sea con una cola y el otro con dos. Fobs= 4; Fcrít= 3.2; con α =0.1 y dos colas; m-1=23; n-1=11; vars diferentes; Se comparan medias usando el método 9.

tobs= 3; tcrít= 2; α =0,05; gli = 35; cola superior, medias diferentes ;

2a) Comparación de medias por el método 8. Edad promedio de los niños: 10 años, varianza 0.66. Edad promedio de padres: 40 años, varianza 0.66. Varianzas iguales con cualquier significancia. t_{obs} = 51,9, t_c = 2,44 con 6 gli Medias diferenetes. **b)** Comparación de medias por datos apareados: $\bar{D} = 0$, $tobs = 0$, medias iguales con cualquier significancia. c) El método correcto es el primero pues toma todos los datos en cuenta. El método de datos apareados es incorrecto pues las edades no se toman en cuenta para nada. Por tanto, en general no se puede resolver un problema de comparación de medias de datos no apareados usando el método de datos apareados sobre su tabla de frecuencias absolutas.

3) Unidad de tiempo = una semana = 168 horas. Comparación de proporciones caso 5b, cola superior. $f = 12/168$, $g = 7/168$, $\widehat{p}$= 0,057; 1- $\widehat{p}$ = 0.943; Por haber muchos datos podemos aproximar la t por la z: zobs= 1.18; zcrít= 1,6; α =0,05; una cola, proporciones iguales.

Parcial 3 versión 4, Pág 204

1)Comparación de proporciones caso 5b con dos colas. $\widehat{p}$=0.208; glis = 23+25-2 = 46. Aproximamos la t por la z. La $z_{obs} = 1{,}57$; $z_c = 1{,}96$.

Discrepancia pequeña: proporciones iguales.

2) Comparación de medias. Primero se comparan varianzas. $F_{obs} = 1{,}7$; $F_c = 4{,}25$ con $\alpha = 0{,}02$, gli 16 y 11, 2 colas; varianzas iguales. Comparación de medias: s_c= 1.81; $t_{obs} = 1{,}47$; $t_c = 2{,}157$; gli = 27; $\alpha = 0{,}02$; una cola. Medias iguales: no hay datos suficientes para alegar que hacer dieta sirve para mejorar la visión.

3) Comparación de medias. Primero se comparan varianzas. $F_{obs} = 16$; $F_c = 9{,}15$ con 4 y 6 gl, $\alpha = 0{,}02$, 2 colas; varianzas diferentes. Comparación de medias: $t_{obs} = 1{,}91$; $t_c = 2{,}756$; gli = 4.54; $\alpha = 0{,}02$; una cola. Medias iguales: no hay datos suficientes para alegar que las encuestadoras hacen trampa.

Parcial 3 versión 5, Pág 204 1) Datos apareados. $t_{obs} = 1{,}03$, $t_c = \pm 2{,}44$ con gli = 6 y 2 colas. Ciudades iguales.

2) Comparación de medias. Primero se comparan varianzas. María: número de semanas = 12; Media = 4.83 por semana; Varianza = 2.697; desviación 1.64;

Manuela: número de semanas = 18; Media = 5.89 por semana; Varianza = 4.58; desviación = 2.19; María vs. Manuela: F_{obs} = 1.696, gliFNum = 17, gliFDen = 11, α = 0.05, dos colas, F_c = 3.2816, F_{obs} ¡F_c, Varianzas iguales. Comparación de medias: $s_c = 1.96$; $t_{obs} = 1.45$; gli = 28.0; Medias iguales.

Mónica: número de semanas = 5; Media = 6.0 por semana; Varianza = 0.5; Manuela vs. Mónica: F_{obs} = 9.1503, gliFNum = 17, gliFDen = 4, α = 0.05, dos colas, F_c = = 8.6113, Varianzas diferentes; Medias: tobs = 0.187; gli = 22; Medias iguales.

Problemas del Taller IV

Pág 253; No 1.a; $n = 0{,}08 \rightarrow 1$.

1.b $n = 15{,}21 \rightarrow 16$.

Pág 253; No 2; IC = (34.58, 35.41); $\mu \neq 42$.

Pág 254; No 3; $t_c = 2{,}145$, IC = (34.11, 35.89); $\mu \neq 42$.

Pág 253; No 4; $t_{obs} = -8{,}66$, $t_c = 2{,}201$; IC = (2.87, 3.12). No.

Pág 254; No 5; $t_{obs} = 10{,}5$, $t_c = 2{,}306$; IC = (12.46, 15.54). No.

Pág 254; No 7; f = 0.508 ; zObs= -7.838; IC = (0.45,0.56); No.

Pág 254; No 8; a) f= 0.588, IC = (0.3352,0.8413). b) Como las dos proporciones suman uno y son iguales, cada una vle 0.5. Como no son independientes, se toma sólo una. Ahora bien, 0.5 pertenece al IC. Por tanto, es justificado decir que son iguales.

Pág 255; No 9; f = 38/43, g = 34/44; f-g = 0.11; p - r = 0.2; s = 0.08; t_{obs} = -1.1. Como los gli son 42, pues es el mínimo entre 42 y 43, y es mayor que 30, aproximamos la t por la z: $t_c = 1{,}96$. Tomamos dos colas pues el profesor enfatiza una igualdad: es 0.2 mayor. Sería con una cola si el profesor hubiese dicho: la proporción de los que pasan en el otro curso es por lo menos 0.2 mayor que en el mío. Discrepancia no significativa: el profesor tiene razón. El IC para la diferencia de medias es (-0.045, 0.27). Como 0.2 pertenece a dicho intervalo, lo que dice el profesor es aceptable.

Pág 255; No 10; IC de la Var = (4.99,20.84); IC de la des = (2.23,4.56). No.

Pág 255; No 11; IC = (0.67,11.5). Si.

Pág 255; No 12; $s_1^2 = 0{,}54$, $s_2^2 = 6{,}027$; IC1 = (0.25,1.8); IC2 = (2.74, 22.117); IC relación = (2.72, 48.93);

Pág 255; No 13; $\bar{X}_{10} = 7{,}1$ $s_1^2 = 5{,}21$; $\bar{Y}_{11} = 8{,}27$, $s_2^2 = 5{,}018$; Para comparar medias, primero se comparan las varianzas: $F_{obs} = 1{,}03$, $F_c = 3{,}7$ con 9 y 10 gli Varianzas iguales; H_o la media mayor es 0.3 unidades más grande que la menor. Por tanto: H_o : μ grande - μ pequeña = 0.3. Como esto es positivo, debemos ser consecuentes en la resta de las medias muestrales: restamos la mayor menos la menor. Test de DOS COLAS (sería de una cola si dijera: la media mayor es POR LO MENOS 0.3 unidades más grande que la menor) $t_{obs} = 0.88$, $t_c = 2.09$, 2 colas, 19 gli Se acepta la H_o.

Pág 255; No 14 a) Primero comparamos varianzas. Fobs = 2.25. Fc = 9.6 con dos colas, significancia 0.05, gli 4 y 4. Varianzas iguales. Hallamos el IC de la diferencia de medias con el caso 8. sc = 0.255. tc = 2.3 con 8 gli. IC de la diferencia de medias = (-0.27 , 0.47). b) El Fiscal alega que la ventas disminuyeron el 10 % de 1.2, el cual es 0.12. Esta sería la Ho del Fiscal para la diferencia de medias. Como dicho valor está en el IC, es aceptable lo que él alega. c) El Abogado dice que la diferencia de medias es cero, el cual está en el IC. Se acepta su creencia. d) El Abogado es apoyado por todos los valores de diferencia de medias en (-0.27, 0) en tanto que el Fiscal es apoyado por (0, 0.47). El Fiscal tiene más apoyo y se decide a su favor: El Señor M.E. irá a la cárcel. e) La estrategia de M.E. para tratar de eludir la estadística es hacer que la cantidad de dinero lavado sea menor que la incertidumbre creada por tener una muestra con un número pequeño de datos. f) Para aumentar la seguridad reunimos más datos durante más meses.

Pág 255; No 15 Se hace un árbol de dos pisos y de dos opciones: J y NoJ a) P(ambas cartas tengan J) = p(J)p(J) = 0.0044 b) P(al menos una J) = p(J)p(J) + p(J)p(NJ) + p(NJ)p(J) = (4/52)(3/51) + (4/52)(48/51) + (48/52)(4/51) = 1 - p(NJ)p(NJ) = 1 - (48/52)(47/51) = 0.14909

Pág 256; No 16; a) P(se maquilló los ojos/fue invitada)= 42,5 %; b) La estadística no se aplica a casos aislados. Pero si nos permitiesen repetir la situación 100 veces y que sean 100 eventos independientes, sería sensato predecir que se maquilló los ojos 43 veces y que no se los maquilló 57 veces.

Pág 256; No 17; P(Atacó)= 30 %; P(atacar/popularidad aumentó) = 30 %.

Pág 256; No 18; La probabilidad de que el papá este bravo dado que la hija gasta mucha plata es 0.76. Como todas las probabilidades suman uno, las otras dos son menores que 0.24. Por lo tanto, el diagnóstico es: lo más probable es que el papá esté bravo por la demasiada gastadera de plata. El papá se lamenta: ¿Será que mi hija no sabe que la plata a mí me toca sudarla?

Pág 256; No 19; P(P y Pasa)= 22,5 %;

P(J y Pasa)= 32,6 %;

P(L y Pasa)= 35,2 %; Se debe premiar a L, pues es él quien produce más productos de calidad listos par vender.

Pág 256; No 20; a) El árbol que representa el experimento

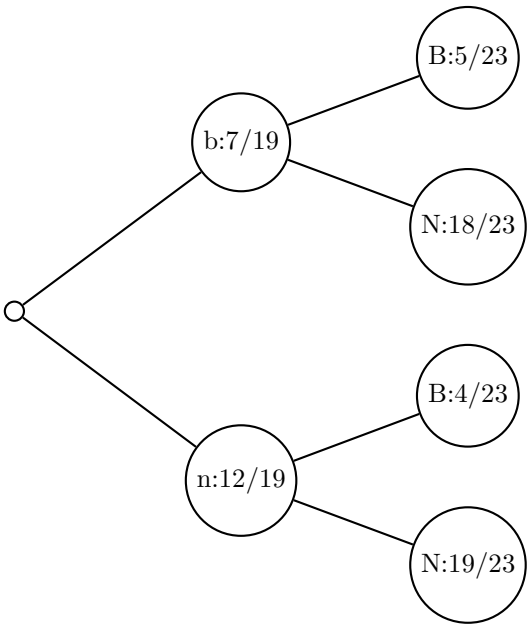


Figura 16.0. Las probabilidades de la segunda bolita dependen de lo que pasó con la primera bolita.

b) 0.81

Pág 256; No 21; Si el primer dado cae uno y los otros dos caen cualquier cosa, la tabla de la suma de los tres dados es:

La suma de 3 dados si el primero cae 1						
+	1	2	3	4	5	6
1	3	4	5	6	7	8
2	4	5	6	7	8	9
3	5	6	7	8	9	10
4	6	7	8	9	10	11
5	7	8	9	10	11	12
6	8	9	10	11	12	13

Como ésta, hay otras 5 tablas dando un total de 216 opciones. Por ejemplo, si el tercer dado cae 3 la tabla queda:

La suma de 3 dados si el primero cae 3						
+	1	2	3	4	5	6
1	5	6	7	8	9	10
2	6	7	8	9	10	11
3	7	8	9	10	11	12
4	8	9	10	11	12	13
5	9	10	11	12	13	14
6	10	11	12	13	14	15

Sobre las 6 tablas debemos contar las veces que en todas ellas caen los números entre 3 y 18 para hallar la tabla que da las probabilidades de Y:

Distribución de Y = la suma de tres dados. Cada valor de $p(y)$ debe dividirse por 216.																
y	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$p(y)\times$ 216	1	3	6	10	15	21	25	27	27	25	21	15	10	6	3	1

Por ojímetro, la media está en 10.5 y la desviación vale aproximadamente 3.5. Por fórmulas: media = 10.5, desviación = 3.24.

Pág 256; No 22; La tabla de lo observado con sus totales es:

Rabia	Candidato escogido			
	X	Y	Z	Totales
A=Poca	13	8	4	25
B=Mediana	12	24	8	44
C=Mucha	2	22	15	39
Totales	27	54	27	108

a) 25/108

b) 27/108

c) 13/108

d) (25+27-13)/108 = 39/108

e) 13/27

f) $p(A \cap X)/p(X) = (13/108)/(27/108) = 13/27 = p(A/X)$

g) $p(A \cap X) = 13/108 = 0{,}1203$ es diferente de $p(A)p(X) = 0{,}05$. Por tanto, los eventos A y X son dependientes.

Pág 257; No 23 h);

Esperado: Asistencia a clase y nota				
	Baja	Mediana	Alta	Totales
$< 30\,\%$	6.25	12.5	6.25	25
$30\,\% - 50\,\%$	11	22	11	44
$> 50\,\%$	9.75	19.5	9.75	39
Totales	27	54	27	108

$\chi^2_{obs} = 20{,}11; \chi^2_{crt} = 9{,}48$; gli = 4; $\alpha = 0{,}05$, una cola. Veredicto: los factores son dependientes. Eso significa ó bien que ir a clase mejora la nota ó que la empeora. Para dilucidarlo, miramos la tabla: ella apoya la idea de que la mejora.

Pág 257; No 24; a) Cuando hay luna llena la proporción de muy buenos conciertos es el triple que la de malos conciertos y también de regulares. Pero algo parecido pasa cuando no hay luna llena. Por lo tanto, se trata de un buen grupo por lo que para aceptar la acusación de los enemigos se necesitarían muchos más datos. b) Tabla de contingencia: Ho : los factores tipo de luna y calidad de concierto son independientes.

Esperado: Tipo de luna y calidad del concierto				
	Baja	Mediana	Alta	Totales
Llena	6.81	6.81	18.38	32
Otra	3.19	3.19	8.61	15
Totales	10	10	27	47

$\chi^2_{obs} = 0{,}38; \chi^2_{crt} = 5{,}99$; gli = 2; $\alpha = 0{,}05$, una cola. Veredicto: los factores son independientes. Es apresurado decir que el tipo de luna influye sobre la calidad del concierto. c) En el test de comparación de proporciones la Ho dice: las diversas proporciones de buenos conciertos es la misma no importa el tipo de luna. Para estudiar la Ho sirve el mismo procedimiento que en b): la Ho se acepta: el grupo es bueno con o sin luna llena.

Pág 258; No 25; Ho : los factores tiempo invertido y tipo de resultado son independientes.

Esperado: Tiempo invertido y tipo de resultado				
	X= Pobre	Y=Mediano	Z=Bueno	Totales
$A =< 10$	25.6	41.66	12.8	80
$B = 10 - 20$	22.4	36.44	11.2	70
$C = > 20$	6.07	9.9	3.03	19
Totales	54	88	27	169

$\chi^2_{obs} = 21{,}81; \chi^2_{crt} = 9{,}48$; gli $= 4$; $\alpha = 0{,}05$, una cola. Veredicto: los factores son dependientes. Menos de 20 semanas da resultados hacia la baja, más de 20 hacia la alta.

Pág 258; No 26; a) Test de homogeneidad de proporciones. H_o: la proporción de citas buenas es independiente de las columnas, es decir, del tipo de amiga. La tabla de lo observado es:

Observado: Tipo de cita vs diversas amigas				
Tipo de cita	Diversas amigas			
	X= Patricia	Y=Alba	Z=Tatiana	Totales
A= mala	2	3	4	
B= buena	6	10	14	
Totales				

La Ho que dice que las proporciones son homogéneas, es decir, que vienen de la misma población por muestreo al azar. Equivalente: las proporciones poblacionales son iguales. Tabla de lo esperado según la Ho:

Esperado: Tipo de cita vs. diversas amigas				
	X= Patricia	Y=Alba	Z=Tatiana	Totales
$A = mala$	1.846	3	4.154	9
$B = buena$	6.154	10	13.846	30
Totales	8	13	18	39

$\chi^2_{obs} = 0{,}024$, insignificante, se acepta la H_o. Que se busquen otro criterio, pues éste no sirvió. **b)** Para neutralizar el efecto de una cita con una amiga sobre la cita con la siguiente, se puede intercalar una o varias citas con niñas fuera del grupo considerado o esperar algún tiempo, digamos 3 días, entre citas.

Pág 262 Parcial IV v.1; **1)** Estimar una varianza sabiendo la varianza de una muestra. a) IC = (17.89, 78.57). b) Como 18 está en el IC, puede decirse que el examen cumple con el requirimieno de la mínima varianza. Sin embargo, los enemigos del profesor dirán solapadamente que no es buen examen pues es permitido decir que la varianza poblacional vale 17.9. No lo pueden decir abiertamente pues hay mucho campo para decir que sí es un buen examen. c) Como la tabla de la chi-cuadrado sólo va hasta 30 datos y tenemos 45, nos toca aproximarla por una normal con media $n - 1 = 44$ y desviación $= \sqrt{2(n-1)} = \sqrt{88} = 9{,}38$. Los valores críticos de dicha normal con dos colas y con el 95 % de confianza ($\alpha = 0{,}05$) son $\mu \pm z\sigma = 44 \pm 1{,}96(9{,}38)$. El valor superior es ${\chi^2}_+ = 62.39$ y el menor es ${\chi^2}_- = 25.61$. Al insertar dicha fórmula en el IC de la varianza que es $\frac{(n-1)VAR}{\chi^2_+} < \sigma^2 < \frac{(n-1)VAR}{\chi^2_-}$ obtenemos $\frac{(44)(32{,}8)}{62{,}39} < \sigma^2 < \frac{(44)(32{,}8)}{25{,}61}$ que se convierte en $23{,}13 < \sigma^2 < 56{,}34$. d) Aunque 18 no está en el nuevo IC, ahora queda claro (con el 95 % de confianza) que la varianza es mayor que 18, es decir que es muy alta, por lo que sí es muy buen examen.

Pág 262 Parcial IV v.1; **2)** Comparar dos varianzas muestrales. a) IC de R (relación entre varianzas)= (0.72, 5.51). b) Sí. c) La mujer del panadero alega que está pemitido pensar que R, la proporción verdadera entre las varianzas, podría valer 1, lo cual es verdadero pues 1 está en el IC de la relación entre varianzas. d) El R verdadero podría estar en cualquier lugar del IC e incluso afuera aunque con el 95 % de confianza está en el IC. Si R es 5 o mayor, el panadero gana. Pero si es menor que 5, el panadero pierde. d) El 5, que es la relación exigida, queda hacia el lado de arriba del IC, por lo tanto la proporción de valores mayores que 5 en el IC es muchísimo menor que la de ser menor. Por consiguiente, de irse de vacaciones estarían jugando a una lotería. Decisión: la situación de riesgo se

esclarecerá por sí sola si se reunen más datos: si el panadero arregló el horno, el IC de la relación de varianzas quedará muy cerrado en un valor muy alto, 5 o más. Pero si lo dañó o si no lo arregló lo suficiente, en un valor menor que 5. La decisión del viaje se pospone.

Pág 262 Parcial IV v.1; **3)** Tabla de contingencia. Esperado según la H_o de independencia de los dos factores:

Esperado: Universidad y mejora financiera			
	X: Poca	Y:Mediana	Z:Mucha
UniA	1.647	1.412	0.941
UniB	2.471	2.118	1.412
UniC	2.882	2.471	1.647

gli = 4, $\chi^2_{obs} = 0.9$, $\chi^2_c = 9.4877$. Factores independientes. Consejo: debido a que la U no influye en el futuro, puede matricularse en la más barata o quizá en la que le quede más cerca pues el tiempo es oro.

Pág 262 Parcial IV v.2; **1a)** Para entender cómo la disminución de la varianza afectaría la calidad de vida de Juan pensemos en un caso concreto: si no hay varianza en el tiempo de llegada (lo cual pasa en los metros), Juan tiene total control sobre su tiempo y puede planificar su vida a todo gusto. Pero si hay varianza, incertidumbre, en el tiempo de viaje, tendría que contentarse con planes plausibles que a lo mejor no le resulten. Ej. Juan se demoraba para llegar a su trabajo antes de la autopista en promedio 60 minutos y con una irregularidad de entre 30 y 90 minutos. Pero como él salía faltando 70 minutos para sus citas, muchas veces llegaba tarde. La autopista disminuyó tanto el promedio como la varianza del tiempo de llegada debido a que hay menos semáforos y menos trancones. Ahora Juan aprovecha para saludar a la novia y por eso su viaje se demora entre 50 y 70 minutos con media 60. Juan sigue saliendo faltando 70 minutos pero ahora llega a tiempo, empieza el día con muy buen genio, atiende mejor a sus clientes, vive más tranquilo y, como consecuencia, la amenaza de infarto se ha pospuesto.

1b) $\alpha = 0.05$, glNUm = 22, glDen = 11 $F_- = 0.371$, $F_+ = 3.22$, IC para la relación entre varianzas = (0.92 , 7.85)

c) Es permitido decir que la varianza disminuyó 5 veces pues 5 está en el IC.

d) No es recomendable que Juan asuma que la varianza disminuyó 5 veces pues debido a que 1 está en el IC también es permitido creer que nada cambió. Incluso, como 0.95 también está en el IC, también es posible que la varianza haya aumentado en vez de disminuir. Lo que Juan tiene que hacer obligatoriamente es reunir más datos para que la estimación de la relación de varianzas sea más precisa y la longitud del IC disminuya.

Pág 262 Parcial IV v.2; **2)**

Los datos son: $\bar{X}_{12} = 74$, $s_1^2 = 140$; $\bar{Y}_5 = 112$, $s_2^2 = 180$.

a)Para saber si es permitido decir que la varianza se duplica, se calcula el intervalo de confianza para la relación entre varianzas, la grande entre la pequeña, y se mira si 2 está en dicho intervalo. Para α = 0,05 y dos colas, F_+ con 4 y 11 gli es 4.27. Como F_+ con 11 y 4 gli es 8.84, entonces $F_- = 1/8{,}84 = 0{,}11$. El IC para la relación entre varianzas es (0.30, 11.68). Como 2 pertenece a dicho intervalo, se acepta la hipótesis de que la varianza se duplicó.

b) Para averiguar si las medias se duplican.

Paso 1: comparamos varianzas para saber si son iguales.

$H_o : R = \sigma_1^2/\sigma_2^2 = 1$ (las varianzas son iguales)

$H_a : R = \sigma_1^2/\sigma_2^2 \neq 1$ (las varianzas no son iguales). Es con dos colas.

Observemos que aunque pueda inferirse que las varianzas se duplican, aún no puede excluírse que sean iguales. La razón es que los datos permiten un gran rango de inferencias, el dado por el IC. Para decidir si es permitido decir que las varianzas poblacionales son iguales, todo lo que tenemos que hacer es mirar si el número 1 está en el rango del IC de la relación entre varianzas. Como sí está, es justificado creer que las varianzas son iguales. Ahora procedemos a comparar las medias:

Paso 2. Averiguamos si la media se duplicó sabiendo que las varianzas son iguales.

$H_o : \mu_2 = 2\mu_1$ es decir, $\mu_2 - \mu_1 = 2\mu_1 - \mu_1 = \mu_1 = \bar{X}_{12} = 74$

$H_a : \mu_2 - \mu_1 \neq 74$, es con dos colas pues preguntan si la media se duplicó o no se duplicó.

Todo lo demás se hace siguiendo el procedimiento de los germanismos: $t_{obs} = -5{,}5$; α=0,05; 2 colas; gli = 15; $t_c = 2{,}131$. Se rechaza la Ho; la segunda media no es el doble de la primera.

c) Un corazón saludable puede acelerar rápidamente pero se serena muy despacio, sin embargo, acelerado o en calma late muy regularmente, con una varianza muy baja. Por tanto, debido a que las varianzas tanto en calma como sobresaltado son muy grandes, Juan está demasiado enfermo y no es de amor. Además, cuando ve a Martiza la varianza le aumenta: nada de raro habrá si termina sufriendo un paro cardiaco o un infarto. Pero Martiza es inocente, que no la culpen.

Pág 262 Parcial IV v.2; **3)** a) p(A) = 26/82; b) p(X) = 35/82; c) P(A y X) = 12/82, se lee de la tabla; d) Si A y X fuesen independientes: P(A)P(X) = (26/82)(35/82) = 0.135. e) Como p(A y X) es diferente de P(A)P(X), estos eventos no son independientes. f) p(A o X) = 49/82; g) p(A dado X) = p(A/X) = 12/35; h) (12/82)/(35/82) = 12/35; i) La chi-cuadrado experimental es 0.18, los dos factores son independientes. j) Ho : la proporción de alta aceptación es independiente de la cantidad de propaganda. El procedimiento para el estudio de independencia de factores es el mismo que para el estudio de un test de homogeneidad de proporciones. Se acepta la Ho. k) El asesor no reunió datos suficientes al azar para inferir que la propaganda ayuda al buen nombre del Alcalde.

Pág 263 Parcial IV v.3; **1)** Comparación de medias de datos apareados. $\bar{D} = 10{,}42$; $s = 2{,}22$. $\bar{Y} = 13{,}57$ $H_o : \mu_2 - \mu_1 = 2\mu_1 - \mu_1 = \mu_1 = 13{,}57$. $t = (10{,}42 - 13{,}57)/(2{,}22/\sqrt{7}) = -3{,}74$ $t_c = \pm 2{,}44$ con 6 g.l, dos colas y significancia 0.05. Se rechaza la H_o. Más rigurosamente, los gli deben ser 5 y no 6 pues se usaron los datos para definir la H_o y por cada parámetro calculado se resta un gli. Con 5 gli, la H_o también se rechaza.

Pág 263 Parcial IV v.3; **2)** Árbol de probabilidad. La probabilidad de haber hecho los ejercicios dado que hoy pudo controlar sus sentimientos es de 0.15/0.21 = 0.71. En general las personas se atreverían a decir que ella los hizo pero sólo hasta que le preguntemos a ella sabremos la verdad. Sin embargo, si esta situación se repite de forma independiente muchas veces, al decir que ella hizo los ejercicios acertaríamos el 71 % de las veces y erraríamos el 29 %.

Pág 263 Parcial IV v.3; **3)** a) $f = 12/17 = 0.706$, $g = 5/22 = 0.227$. $H_o : p - r = 0{,}1$, se rechaza con dos colas pues se pregunta sobre una igualdad. Sería con una cola si se preguntara: ¿Puede decirse que la proporción de estudiantes a quines les va mal por la tarde es por lo menos 0.1 mayor en los del primer grupo que en el segundo? $t_{obs} = 2{,}6$, $t_c = \pm 2{,}12$ con gli = 16. Se rechaza la H_o. No es permitido decir que la proporción de estudiantes a quienes les va mal por la tarde es 0.1 mayor en el primer grupo que en el segundo. b) IC de la diferencia de proporciones (0.177, 0.779) c) Sí puede decirse que la diferencia de proporciones es 0.7 pues está en el IC.

Pág 264 Parcial IV v.4; **1)** Esperamos de un buen reloj que si da un tic cada segundo, no se atrace ni se adelante sino que se demore siempre lo mismo. Por tanto, decir que algo periodico funciona como un reloj, es decir que la varianza del período es cero.

a) $H_o : \mu = 28$ dado que $\sigma = 0$. Caso 1, ejemplo 215a, pág. 210. Se rechaza con dos colas, pues se trata de ser igual o diferente de 28. Se puede usar una $z_{obs} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{27-28}{0/\sqrt{23}} = -\infty$. La z crítica con significancia 0.05 y dos colas es 1.96. Por tanto, la $z_{obs} = -\infty$ es muy extrema. Se rechaza la H_o: los datos no apoyan la idea popular de que la media vale exactamente 28 dado que la varianza es cero. También existe otra forma de afrontar el problema y es tomando $s = \sqrt{14{,}2}$ y usando una t con 22gl, caso 3.

b) $H_o : \sigma^2 = 0$. Se rechaza con dos colas teniendo el dato experimental $s^2 = 14{,}2$ con 23 datos, o sea, 22gl. Se usa una χ^2, caso 6, Ej. 237, pág. 224. $\chi^2_{obs} = \infty$ y la χ^2 crítica con 22gli es 36.781 pues $\beta = \alpha$. Se rechaza la H_o. Hay datos de sobra para decir que la varianza no es cero. El período de las jovencitas no viene sincronizado como por reloj (este desajuste es real y es una causa de muchos embarazos no deseados, un calvario para la jovencita y para su familia).

c) Igual que con las jovencitas, hay datos suficientes para alegar que la varianza de las señoras no es cero y que la media no es 28 exacto.

d) El IC para la relación entre varianzas se resuelve siguiendo el ejemplo 243 de la pág. 227, caso 7. Varianza grande entre varianza pequeña = 14.2/6 = 2.36. F_+ con 22gli en el numerador y 14 gli en el denominador es 2.84. F_+ con 14gli en el numerador y 22 gli en el denominador es 2.49. $F_= 1/2{,}49 = 0{,}40$. El IC para R, la relación entre varianzas, va desde 2.36/2.84 = 0.83 hasta 2.36/0.40 = 5.9.

e) Como $R = 2$ pertenece al IC para R, eso quiere decir que la varianza grande entre la pequeña da 2 o que la varianza pequeña es la mitad de la grande. Por tanto, se justifica decir que las señoras disminuyen la varianza a la mitad.

f) También puede decirse que las señoras disminuyen la varianza a la tercera parte pues $R = 3$ está en el intervalo. Por tanto, es justificado decir que las señoras disminuyen la irregularidad en el

período a menos de la mitad. Por esa razón hay muchos menos embarazos no deseados entre las señoras que entre las jovencitas (cuando no se toman anticonceptivos y las relaciones se planifican conforme al reloj femenino que a veces pierde el ritmo).

Pág 264 Parcial IV v.4; **2)** Como se sacan dos cartas, el arbol tiene dos pisos y cada nodo tiene dos opciones: color rojo (R) y color negro(N). a) Diferente color significa RN y NR: p(RN) + p(NR) = 0.509. b) p(JN) + p(NJ) + p(NN) = 1-p(JJ) = 1 - (4/52)(3/51) = 0.995.

Pág 264; Parcial IV v.4 3) Ho: los factores son independientes. α=0,05; 4gl, cola superior; Discrepancia= 18.8; χ^2_{crt}= 9.48. Se rechaza la Ho = factores dependientes. b) Para saber cómo se relaciona la guerra con la calidad de vida miramos la tabla: entre mejor calidad de vida, más armas. c) Podría ser porque el que tiene plata quiere impresionar, o bien porque el que tiene plata quiere más y por eso fabrica armas, para vender y especular y para crear más trabajo. Esto es al contrario de lo que pasa en los regímenes comunistas en donde se dedica no un 5 % a las armas sino quizá un 50 % pues las armas nucleares, la guerra biológica y la compra masiva de traidores son todas demasiado caras.

Pág 264 Parcial IV v.5; **1)**.

a) La tabla de frecuencias absolutas asociada a la acumulada con su extensión correspondiente para calcular la varianza es la siguiente:

Cálculo de la varianza			
$x = créditos$	F = frecuencia	xF	x^2F
0.5	2	1	0.5
1.5	5	7.5	11.25
2.5	15	37.5	93.75
3.5	10	35	122.5
4.5	2	9	40.5
Total	$n = \sum F = 34$	$\Sigma xF = 90$	$\Sigma x^2F = 268{,}5$

La nota promedio en el primer parcial fue de $90/34 = 2.64$ con una desviación de 0.958 y una varianza de 0.917 con 33gl. La nota promedio para el segundo parcial fue 3.1 con una varianza de 0.16 y 33gl.

b) Hallamos el IC para R, la relación entre varianzas con 33 gli tanto en el numerador como en el denominador. $F_+ = 2{,}07$, $F_- = 0{,}48$: IC = (2.77, 11.93). Para decidir si las varianzas poblacionales son iguales o distintas miramos si $R = 1$ está en el IC de la relación entre varianzas. Como no está, aconsejamos creer que las varianzas poblacionales son distintas.

c) Sabiendo que las varianzas son distintas se aplica la metodología correspondiente para comparación de medias, caso 9, ej 249 pág. 235. En éste caso, la Ho dice que $\mu_2 = \mu_1 + 0{,}1\mu_1$ que es lo mismo que decir que $\mu_2 - \mu_1 = \mu_1 + 0{,}1\mu_1 - \mu_1 = 0{,}1\mu_1$ que se puede aproximar por $0{,}1(2{,}64) = 0{,}264$. La $t_{obs} = 1{,}1$, gli = 44.8, la t_c es mayor que 1.5, se acepta la H_o. Es permitido creer que la nota promedio mejoró un 10 %.

d) Decidamos si es creíble decir que la metodología del quiz disminuye la varianza un 25 %: La varianza original es 0.9171. Un 25 % menos corresponde a $0{,}25 \times 0{,}9171 = 0{,}2292$ menos. Ahora bien, $0{,}9171 - 0{,}2292 = 0{,}6879$. La nueva relación entre varianzas sería $0{,}6879/0{,}16 = 4{,}29$. Como éste número está en el IC de la relación entre varianzas, tenemos derecho a alegar que la metodología del quiz disminuye la varianza un 25 %.

Pág 264 Parcial IV v.5; **2)** Probabilidad de que haya sido escaramuza dado que no murió = 0.84. Es justificado creer que hubo escaramuza.

Pág 264 Parcial IV v.5; **3)**

a) Test de homogeneidad de proporciones. Si las proporciones de ser fuerte son homogéneas con respecto al color de piel entonces no cambian con el color de piel, o sea que los factores fortaleza y color de piel son independientes. Lo cual nos permite usar la misma metodología que para tabla de contingencia. Lo esperado según la H_o, que los factores son independientes, es:

Esperado: Fortaleza y color de la piel				
	X= Blanca	Y= Trigueña	Z=negra	Totales
A= Fuerte	15.15	17.449	12.398	45
B= Débil	17.847	20.551	14.602	53
Totales	33	38	27	98

$\chi^2_{Obs} = 6.5292$
$\chi^2_{+} = 5.9915$

Se rechaza la Ho, por tanto, el color de la piel es indicativo de la fortaleza física. Más rigurosamentre: la proporción de personas físicamente fuertes es dependiente del color de la piel. La tabla señala a la piel negra como indicativo de fortaleza.

b) Frecuencia observada para A: 12 entre 33. f = 0.3636, IC = (0.1962,0.5311)

c) Se trata de comparar la diferencia de dos proporciones muestrales contra la diferencia esperada de dos proporciones poblacionales. Se usa el caso 5 del folleto de fórmulas. La proporción muestral más grande es 18/27 =0.667 y la más pequeña es 12/33 = 0.3636. Lo esperado es que la diferencia de proporciones sea el 10 % de la menor: p-r = 0.03636, dos colas: tObs = 2.1599, glis = min(32, 26) = 26; tCritic = 2.05. No puede decirse que la proporción de A dado Z sea 10 % mayor que la de A dado X. Sería mejor decir que es el doble o algo así.

Pág 265 Parcial IV v.6; 1a) Caso 1 del diccionario de experimentos. Dentro de mujeres: n = 6, $\bar{X}$ = 77, σ = 11, μ = 80 z_{obs} = -0.668 Z_c = 2. No hubo cambio. Dentro de hombres: n = 7, $\bar{X}$ = 78, σ = 6, μ = 77, z_{obs} = 0.44, z_c = 2. No hubo cambio.

1b) Caso 2. μ_x= 77, μ_y= 80, σ_x = 6, σ_y = 11, $\bar{X}$ = 78, $\bar{Y}$ = 77, n_x = 7, n_y = 6, z_{obs} = 0.795, z_c = 2. No hubo cambio.

1c) Caso 8 o 9. Dentro de mujeres. Varianzas iguales. Comparación de medias, caso 8: t_{obs} = -0.4494, 11gl, α = 0,05, t_c = +/-2.21. No hubo cambio.

Dentro de hombres:

$\bar{X}$ =78, varx = 16, nx = 7, $\bar{Y}$= 81, vary = 81, ny= 8, $\mu_x - \mu_y = 0$. Comparamos varianzas: F_{obs} = 5.0625, con 7 y 6 gli F_{obs} = 5.0625, con 7 y 6gl, α = 0,05, F_c = 5.695. Varianzas iguales. Comparación de medias, caso 8: t_{obs}= -0.812, 13gl, $t_c = +/- 2{,}1603$. No hubo cambio.

1d) Debido a que las cuatro muestras son muy pequeñas, no podemos tomar a ninguna como poblacional, y por tanto, el problema no puede hacerse (con las herramientas que tenemos).

1e) $\bar{X}$=81, varx = 81, nx = 8, $\bar{Y}$= 79, vary = 64, ny= 7, μ_x = 89; mu_y = 79; $\mu_x - \mu_y$ = 10

Estudio de varianzas: R=1.0, F_{obs} = 1.2656, α = 0.05, dos colas, 7 y 6 gl, F_+ = 5.6955, varianzas iguales. Comparación de medias, caso 8:

t_{obs} = -1.8073

α = 0.05, dos colas (si dijeran: somos mejores que ellas más allá de 10 puntos, sería de una cola), gli = 13, t_c == +/ − 2,1604. Los hombres alegan algo que la estadística no puede contradecir, pero un mujer inteligente sabrá utilizar el IC de la diferencia de medias para defenderse.

1f) t_{obs} = 1.5814, t_c = +/-2.1604. Es correcto decir que las mujeres están 5 puntos por encima de los hombres.

1g) IC para la diferencia de medias: (-7.51, 11.51). El centro del intervalo está en 2, que sería la ventaja sobre las mujeres que un hombre podría alegar sin peligro de que lo rebatan.

Pág 265 Parcial IV Super 5 1a) Como es con reemplazo, las probabilidades no cambian. Se hace un árbol de dos pisos con dos opciones en cada nodo: 3 y No3. Probabilidad de al menos un 3 = $p(3\&3) + p(3\&No3) + p(No3\&3) = (4/52)(4/52) + (4/52)(48/52) + (48/52)(4/52)$.

1b) Se hace un árbol de dos pisos con dos opciones en cada nodo: 5 y No5. Probabilidad de máximo un 5 = $p(5\&No5) + p(No5\&5) + p(No5\&No5) = (4/52)(48/52) + (48/52)(4/52) + (48/52)(48/52)$.

1c) Haga un árbol de dos pisos con 3 opciones en cada nodo: un 3, un 5 y OtraCosa que no sea ni 3 ni 5. Para hallar la p de exactamente un 3 o exactamente un 5 se seleccionan los caminos en el arbol que clasifican con exactamente un tres, con exactamente un 5 o con ambos. Es igual a $p(3\&5) + p(3\&OtraCosa) + p(5\&3) + p(5\&OtraCosa) + p(OtraCosa\&3) + p(OtraCosa\&5) = (4/52)(4/52) + (4/52)(44/52) + (4/52)(4/52) + (4/52)(44/52) + (44/52)(4/52) + (44/52)(4/52)$

1d) El universo condicional es no ser 3 en la segunda carta y de ahí tomamos los caminos que tienen no 3 en la primera carta. Respuesta = $p(No3\&No3)/(p(3\&No3) + p(No3\&No3))$

= $(48/52)(4/52)/((4/52)(48/52) + (48/52)(48/52))$

2a) Caso 5, f = 0.48, g = 0.25, IC diferencia de proporciones = (0.03, 0.43)

2b) Si la proporción de fieles siguió igual, la diferencia de proporciones poblacionales debe ser cero. Como cero no pertenece al IC de la diferencia de proporciones, dicha H_o no se acepta: hubo cambio.

2c) La proporción original era 0.48. Su mitad es 0.24. La mayor menos la menor da 0.48 - 0.24 = 0.24. Este valor sí está en el IC de la diferencia de proporciones. Por lo tanto, es permitido decir que la

proporción de fieles disminuyó a la mitad.

3a) Caso 6. Como el número de grados de libertad es mayor que 30, se aproxima la curva de la chi-cuadrado por una campana de Gauss para así poder hallar los valores extremos. La media de la campana de Gauss que mejor aproxima la chi-cuadrado es 42 y su desviación es $\sqrt{2(42)} = 9{,}16$. Los valores críticos de la chi-cuadrado son los mismos de la campana, $\mu \pm z\sigma$. Con significancia del 0.05, la $z_c = 2$. El mayor es $\chi^2_+ = 42 + 2(9{,}16) = 60{,}33$. El valor crítico menor es $\chi^2_- = 42 - 2(9{,}16) = 18{,}33$. Ahora se reemplazan estos valores en el IC de la varianza.

4a) $s = 6$, $s^2 = 36$. IC de la varianza = (19.22, 98.96).

4b) Caso 7, $F_{obs} = 9$; $F_c = 2{,}88$. Varianzas diferentes. IC de la diferencia de medias: (10.29, 17.7).

4c) Caso 9, IC de la relación entre varianzas (3.1, 28.36).

4d) Decir que la media mayor es cinco veces la menor es decir que la mayor menos la menor es 4 veces la menor, es decir 4 veces 14= 56. Como 56 no está en el intervalo de la diferencia de medias, no se puede decir tal cosa.

4e) Como 4 está en el IC de la relación de varianzas, sí puede decirse que la varianza grande sea 4 veces la chica.

4f) Como 3 no está en el IC de la varianza, no puede decirse que la varianza valga 3.

5a) 14/35.

5b) 14/35.

5c) 7/35.

5d) 21/35.

5e) 7/14.

5f) $P(B \cap Z) \neq p(B)p(Z)$, son dependientes.

g) chi-cuadrado observado = 37.46 - 35 = 2.46. Valor crítico de la chi-cuadrado = 9.35. Discrepancia pequeña. Los datos insinúan que ir a fiestas es contraproducente, pero faltan datos para probarlo con un nivel de confianza del 95 %.

Problemas del taller complementario 1, Censos, Numeral 279 pág. 272

1. ¿A cuál curso le fue mejor en promedio para el segundo parcial?

Halle los promedios y compare:

Promedio de A: 3.51.

Promedio de B: 3.73.

2. ¿Cual curso tiene mayor varianza para el cuarto parcial?

Var de A: 1.3.

Var de B: 1.33.

3.¿A quienes les fue mejor en promedio en el examen final en el curso A, a los abogados o a los de ciencia política?

Tome el curso A, ordene por programa.

Calcule la media de los de derecho: 3.08.

calcule la media de los de ciencia política: 2.44.

4. ¿Es la distribución de las notas del segundo parcial del curso A de tendencia central? Se copia aparte la columna del parcial 2 del curso A, se ordenan los datos,

se cuentan los datos que están entre cero y uno, los que empiezan por 1, etc. Los que sacan 5 se cuentan con los que sacan 4. Da la siguiente tabla.

Notas del curso A	
Nota	**Frecuencia**
0	2
1	2
2	6
3	8
4	18

Para hacer el histograma, se señala la segunda columna y se hace click en el ícono de gráfica. La distribución no es de tendencia central: hubo un gran esfuerzo de todo el mundo por lograr la excelencia.

5. Hacer una torta que muestre la frecuencia relativa de los que pasaron el final del curso B. Se copia aparte la columna del final.

Se ordenan los datos.
Se cuenta el número total: 32.
Se cuentan los que pasaron: 11.
se hace una tabla:
Pasaron 11 Perdieron 21
Se señala toda la tabla con las etiquetas de subtítulos.
Se llama el menú de gráficos y se elije una torta (pie) o círculo.
6. Comparar las frecuencias relativas de los que pasaron el final en los dos cursos: ¿Cuál fue mayor?
Frecuencia relativa de los que pasaron en A: 14 de 35= 0.4 Frecuencia relativa de los que pasaron en B: 11 de 32= 0.34.

Problemas del taller complementario 2, Chevichev, numeral 285 pág. 282

Sobre estadística no paramétrica, ejercicio 1: hay 36 datos sin tener en cuenta una casilla vacía. Cada cuartil cubre 9 datos después de ordenados. El rersumen de los 5 valores: El mínimo es 1.7, el primer cuartil es 3.75, la mediana es 4.5, el tercer cuartil es 5 y el cuarto es 5.

Ejercicio 2. El primer cuartil de la z es -0.70, el tercero es 0.70. La distancia intercuartilar es 1.4. La aleta cuartilar es $1.5 \times 1.4 = 2.1$. Si le sumamos la aleta cuartilar al tercer cuartil, obtnemos: 07 + 2.1 = 2.8. Si le restamos la aleta del primer cuartil obtenemos -0.7-2.1 = -2.8. Eso significa que de acuerdo con el criterio intercuartilar, todo lo que esté por debajo de -2.8 es extremo por abajo y todo lo que esté por arriba de 2.8 es extremo por arriba. La significancia asociada a 2.8 es 0.0026. Así que el criterio intercuartilar no es igual al criterio de las dos desviaciones (con significancia 0.05 y dos colas) sino al criterio de 2.8 desvaciones y con significancia 0.0026 y dos colas. El criterio de las dos desviaciones para determinar los eventos extremos puede reformularse en términos del criterio intercuartilar así: si multiplicamos el rango intercuartilar,1.4, por x para dar una aleta que debe ser sumada al tercer cuartil y que de 2, obtenemos $0.7 + 1.4\text{k} = 2$, entonces $\text{k} = 1.3/1.4 = 0.9 \sim 1$. Es decir, el criterio de ls dos desviciones se traduce en sumar un rango intercuarticular, y no 1.5 rangos, al tercer cuartil.

Respuestas para los ejercicios de Chevichev:

1. Chevichev para distribuciones cualesquiera: $p(\mid x-\mu \mid \geq 1\sigma) \leq 1$; $p(\mid x-\mu \mid \geq 2\sigma) \leq 0{,}25$; $p(\mid x-\mu \mid \geq 3\sigma) = 0{,}11$;

2. Regla de las tres áreas para una distribución normal: $p(\mid x-\mu \mid \geq 1\sigma) = 0{,}32$; $p(\mid x-\mu \mid \geq 2\sigma) = 0{,}045$; $p(\mid x-\mu \mid \geq 3\sigma) \leq 0{,}003$;

3. Chevichev $p(\mid x-\mu \mid \geq 3{,}7\sigma) \leq 0{,}073$. Normal: 0.0002.

4. $z(42) = (42-30)/5 = 2{,}4$; Chevichev $p(\mid x-\mu \mid \geq 2{,}4\sigma) \leq 0{,}17$.

5. $z(42) = (42-30)/5 = 2{,}4 = z(18)$; Normalidad: $p(\mid x-\mu \mid \geq 2{,}4\sigma) = 0{,}0164$.

6. $z(35) = (35-25)/2 = 5$; Chevichev $p(\mid x-\mu \mid \geq 5\sigma) \leq 0{,}04$.

7. $\bar{X}_{25} = 25$ con $\sigma = 2/5 = 0{,}4$, $z(7) = (7-5)/0{,}4 = 5$; Chevichev $p(\mid x-\mu \mid \geq 5\sigma) \leq 0{,}04$. Asumiendo normalidad : $p(\mid x-\mu \mid \geq 5\sigma) = 0^{+}$.

Problemas del taller complementario 3, test de ajuste, numeral 291 pág. 288

1. Frecuencias observadas: 23, 25, 19, 30, 25 y 22. Frecuencias esperadas suponiendo una distribución uniforme: 24 para todas. Valor experimental de $\chi^2 = 2{,}83$, 5 gl, $\chi^2_c = 11{,}0705$ con significancia 0.05 y cola superior. Veredicto: No hay razón experimental para creer que el apostador haya logrado falsificar el dado.

2. Frecuencias observadas: 2,4,12,5,8, suma = 31. Frecuencias esperadas suponiendo una distribución ad hoc: 3.77, 7.38, 11.48, 6.72, 1.64. Suma = 30.99. Como la primera frecuencia esperada es menor que 5 la agrupamos con la segunda y hacemos lo mismo con lo observado. Queda en lo observado: 6, 12, 5, 8. En lo esperado: 11.15, 11.48, 6.72 y 1.64. Valor experimental de $\chi^2 = 27{,}5$, 3 gl, $\chi^2_c = 7{,}8147$ con significancia 0.05 y cola superior. Veredicto: El mercado ha cambiado. Pero hay una objeción: sería mejor comparar los registros del primer mes de este año con el promedio de los registros de los primeros meses de los años pasados y no con los de todo el año pasado.

3. Calculamos la probabilidad de los que quieren vender: $p = \frac{6\times0+9\times1+12\times2+20\times3}{(6+9+12+20)\times3} = \frac{9+24+60}{141} = 0{,}659$. Como hemos calculado la p a partir de los datos, llevamos un grado de libertad perdido. La expansión de la $Bi(3,\ 0.659)$ da las siguientes probabilidades: 0.0394, 0.2293, 0.443, 0.2870. La suma da 1. Al multiplicar por 47 obtenemos: 1.85, 10.78, 20.88, 13.49. Que suma 47. Como la primera clase está subrepresentada, agrupamos. Terminamos con los siguientes datos esperados: 12.63, 20.88 y 13.49.

Como lo observado es 6, 9, 12 y 20, se transforma en 15, 12 y 20. La $\chi^2_{obs} = 7{,}363$. La $\chi^2_c = 3{,}8$ con 3 -2 =1 gl, $\alpha = 0{,}05$ y la cola superior. Se resta un gli adicional pues la p se calculó a partir de lo observado. Las personas que quieren vender no parecen distribuidas al azar, hay una segregación, lo cual indica que la discusión acerca de vender o no vender se ha tornado irritante.

4. a) Media = 4.96: la gente llega a pagar en promedio casi al quinto mes. Desviación = 1.25. No todos llegan al tiempo.

b) Para decidir el test K-S lo hacemos sobre la tabla siguiente:

Test de Kolmogorov-Smirnov						
x	$F(x)$	$A(x)$	$\Psi(x)$	z	$\Psi(z)$	$\lvert\Psi(x) - \Psi(z)\rvert$
1	230	230	0.00377	-3.168	0.001	0.00278
2	2500	2730	0.0448	-2.37	0.0082	0.0034
3	7658	10388	0.17053	-1.568	0.0548	0.11574
4	8752	19140	0.3142	-0.768	0.2119	0.1024
5	11375	30515	0.5009	0.032	0.5	0.0009
6	30400	60915	1	0.832	0.7881	0.2119

Discrepancia entre lo predicho si se espera normalidad y lo observado = valor máximo de $|\Psi(x)-\Psi(z)|$ es 0,2119. El valor crítico con $\alpha = 0{,}05$ es $1{,}36/\sqrt{60915} = 0{,}005$. No hay normalidad.

c) La distribución oṕtima se logra así: a) A cada letra del alfabeto se le asigna un tiempo de presentación proporcional al número de clientes cuyo apellido empiece por esa letra b) Durante el tiempo señalado, a los usuarios correspondientes se les atiende en una cola de alta prioridad y a los demás una cola común y corriente. Variante: segregar por el número de la cuenta del servicio.

Problemas del taller complementario 4, Gnumeric, 296 pág. 294

Respuestas calculadas con Gnumeric:

1. $H_o : \mu = 3$, $H_a : \mu > 3$, una cola. $\bar{X}_9 = 2{,}55555555555556$, $s = 0{,}8819171036882$. $z_{obs} = 1{,}85185185185185$. La significancia es 0.05 y con la cola superior, el $z_c = 1{,}64485362695147$. Se rechaza la H_0. Es incómodo creer que la media sea 2 y apoyamos la idea de que sea mayor que 2.

2. $H_o : \mu_2 - \mu 1 = 1$, $H_a : \mu_2 - \mu 1 > 1$, cola superior. $z_{obs} = 0$ que se acepta con todas la significancias.

3. $\bar{X}_9 = 2{,}55555555555556$, $s = 0{,}8819171036882$, la significancia con dos colas es 0.05, el t-crítico con dos colas y 8 grados de libertad es 2.30600413520451. El intervalo va de 1.87765405955137 hasta 3.23345705155974.

4. $f = 1/9$, $H_o : p = 0{,}2$, dos colas, $z_{obs} = -0{,}66666666666667$. Se acepta la H_o ipsofacto.

5. $f = 1/9$, $g = 2/9$, $H_o : p - r = 0$, dos colas. $z_{obs} = 0{,}21320071635561$. La H_o se acepta ipso facto.

6. Hallemos el intervalo de confianza de una varianza usando las funciones de OpenOffice. Hay 9 datos, 8 gl. $s^2 = 0{,}77777777777778$. La chi-cuadrado con 8 gli y significancia 0.025 en la cola inferior se calcula por

   =CHISQINV(0.025,8)

   y da 2.18. La chi cuadrado con 8 gli y significancia 0.025 en la cola superior se calcula por

   =CHIINV(0.025,8)

   y da 17.53.

   (Estos valores coinciden con los que da Gnumeric, sólo que Gnumeric da 9 decimales.)

   Por tanto, el intervalo de confianza va desde 0.3549470748558 hasta 2.85423037716616.

7. $H_o : \sigma_2^2/\sigma_1^2 = 1$. Dos colas. Para comparar varianzas es Gnumeric, se va a *Tools*, después a *statistical analysis* y se elige *Two variances: F test*. La $F_{obs} = 2{,}28571428571429$, la F-crítica del lado superior vale 3.7410037631864. Se acepta la H_o con significancia 0.04. Las varianzas son iguales.

8. Sabiendo que las varianzas poblacionales son iguales, se usa una t que se puede hacer directamente con *Tools*. $H_o : \mu 2 - \mu 1 = 1$. la $t_{obs} = 0$, la t-crítica con dos colas vale 2.11990529922182. Se acepta que la media mayor es una unidad más que la menor.

9. Se prueba que las varianzas son diferentes y después de eso se estudian las medias. $H_o : \mu_2 - \mu_1 = 10$, dos colas. La $t_{obs} = 0{,}31258626265688$ en tanto que la t-crítica para dos colas es 2.39248504265803. Se acepta la H_o: está per,itido creer que la media mayor es 10 unidades más que la menor.

10. Se hace directo con Tools de Gnumeric. $H_o : \mu_1 - \mu_2 = 0$. Con dos colas. La $t_{obs} = 2{,}68328157299975$, la t-crítica con dos colas y 0.05 de significancia es 2.30600413520451. Se rechaza la H_o: tenemos suficientes datos para alegar que las medias poblacionales son diferentes.

11. En Gnumeric se busca *Tools*, después *Statistical analysis* después *Contingency tables* y se elige *Test de independencia*. Y listo.

Problemas del Final

Pág 300, versión 1, No 1.. a) $\bar{X} = 3{,}778$, $s = 0{,}5$

b) $\bar{X} = 3{,}85$ (es un promedio ponderado: se calcula el total de notas de cada curso, se suman todos estos totales para encontrar el gran total de notas y se divide por el número total de estudiantes).

Pág 300, versión 1, No 2a.. Con $R = rojo$ y $N = negro$ tenemos prob del mismo color = $p(RR) + p(NN) = 2(\frac{26}{52}\frac{25}{51}) = 0{,}4901$

b. Con $K = K$ y $N =$ no K, tenemos prob de al menos una K = $p(KK) + p(KN) + p(NK) = \frac{4}{52}\frac{3}{51} + \frac{4}{52}\frac{48}{51} + \frac{48}{52}\frac{4}{51} = \frac{396}{2652} = 0{,}14$

c. Sea $A =$ las dos cartas son del mismo color, $B =$ al menos una K. Hay que averiguar si $p(A)p(B) = p(A \text{ y } B)$. Las respuestas de lado izquierdo están en los puntos a) y b): $0{,}4901 \times 0{,}14 = 0{,}07$. Falta calcular $p(A \text{ y } B)$. Para ello hacemos una lista de todos los posibles eventos que representen ser del mismo color y tener al menos una K y sumamos las probabilidades. Pueden ser 1) p(Ser K roja y ser K roja) = (2/52)(1/51). 2) p(Ser K negra y ser K negra) = (2/52)(1/51). 3) p(Ser K roja y ser No K roja) = (2/52)(24/51). 4) p(Ser K negra y ser No K negra) = (2/52)(24/51). 5) p(Ser No K roja y K roja) = (24/52)(2/51). 6) p(Ser No K negra y K negra) = (24/52)(2/51). Sumando todo y escribiendo la respuesta con dos decimales da 0.07. Los eventos son independientes pues la igualdad se cumple. Con tres decimales ya no es cierto, por lo que decimos que los eventos tienen una ligera dependencia.

Pág 300, versión 1, No 3 Sin reemplazo a. $p(RR) = \frac{6}{20}\frac{5}{19}$

b. $p(RR) + p(RA) + p(AR) = \frac{6}{20}\frac{5}{19} + \frac{6}{20}\frac{14}{19} + \frac{14}{20}\frac{6}{19}$

c. $p(NN) + p(NR) + p(RN) = 1 - p(RR) = 1 - \frac{6}{20}\frac{5}{19}$

d. $p(RA)/[p(RA) + p(AA) = \frac{6}{20}\frac{14}{19}/[\frac{6}{20}\frac{14}{19} + \frac{14}{20}\frac{13}{19}]$

Con reemplazo a. $p(RR) = \frac{6}{20}\frac{6}{20}$

Pág 300, versión 1, No 4. a) La MEDIA del diametro de las esferitas debe estar calibrada para ser un tricito menor que el diámetro de la punta del lapicero, pero debido a que los procedimientos de fabricación no son perfectos, unas esferitas quedarán maś grandes de lo estipulado y otras más pequeñas y otras no tan esféricas. Este desajuste es la VARIANZA. b) La muestra es lo que se toma para hacer un experimento concreto, en este caso 15 esferas para cuantificar el procedimiento viejo. La población incluye todas las esferas fabricadas hasta la fecha y las que potencialmente puedan hacerse con el mismo procedimiento.

c) IC de la varianza del diámetro bajo el primer procedimiento = (0.0214,0.0995)

d) IC de la relación entre varianzas = (0.7266,5.8005)

e) Como 3 pertenece al IC de la relación entre varianzas, podemos creer que la novedad cumple la demanda. Resolvamos el mismo problema por un test F de comparación de dos varianzas. varx =0.04, nx = 15, vary = 0.02, ny= 18, α= 0.05, H_o : R=3.0, la varianza poblacional grande es tres veces la pequeña. $R_{Obs} = 2.0$, la varianza muestral grande es dos veces la pequeña. $F_{Obs} = 0.66667$,

$F_- = 0.3448$, $F_+ = 2.7526$, Se acepta la H_o: está permitido creer que la novedad cumple la demanda.

f) Todo lo que hacemos es para tomar decisiones, las cuales no pueden basarse en la muestra pues ésta es el espejismo del momento en tanto que las decisiones afectan a toda la población. Por tato, las decisiones deben tomarse sobre valores poblacionales que se infieren a partir de la muestra. Precisamente,

los IC son para estimar los valores poblaciones con el margen de error determinado por la significancia: entre más pequeña la significancia, más grande es el valor crítico del estadígrafo de contraste y más grande el error de estimación del valor poblacional.

g) Como posiblemente los resultados de un experimento nunca vuelvan a salir, los valores muestrales son inútiles para la toma de decisiones. Deben tenerse en cuenta las proyecciones a futuro que son dadas por los valores poblacionales.

h) Es ciertamente permitido creer que la varianza grande es 3 veces la pequeña, pero debido a que el IC de R contiene valores menores que uno, el método nuevo podría ser incluso peor que el viejo. Debido a que el nuevo procedimiento es muy caro, no se justifica la inversión.

i) Calculemos con el 95 % de confianza el IC de la varianza si con 50 datos $VAR = 0{,}015$. Como la tabla de la chi-cuadrado sólo va hasta 30 datos y tenemos 50, nos toca aproximarla por una normal con media $n-1 = 49$ y desviación $= \sqrt{2(n-1)} = \sqrt{98} = 9{,}899$. Los valores críticos de dicha normal con dos colas y con el 95 % de confianza ($\alpha = 0{,}05$) son $\mu \pm z\sigma = 49 \pm 2(9{,}899)$. El valor superior es ${\chi^2}_+ = 68.798$ y el menor es ${\chi^2}_- = 29.202$. Al insertar dicha fórmula en el IC de la varianza que es $\frac{(n-1)VAR}{\chi^2_+} < \sigma^2 < \frac{(n-1)VAR}{\chi^2_-}$ obtenemos $\frac{(49)(0{,}015)}{68{,}798} < \sigma^2 < \frac{(49)(0{,}015)}{29{,}202}$ que se convierte en $0{,}0117 < \sigma^2 < 0{,}0227$.

j) Por otro lado, los datos para estudiar la relación entre varianzas son: varx =0.04, nx = 15, vary = 0.015, ny= 50, α= 0.05. El IC de la relación entre varianzas es: $\frac{VAR_G}{VAR_P F_+} < \frac{\sigma_G^2}{\sigma_P^2} < \frac{VAR_G}{VAR_P F_-}$ que se convierte en $\frac{0{,}04}{0{,}015 \times 2{,}10} < \frac{\sigma_G^2}{\sigma_P^2} < \frac{0{,}04}{0{,}015 \times 0{,}4}$ lo cual da el IC para la relación entre varianzas poblacionales: $1.26 < \frac{\sigma_G^2}{\sigma_P^2} < 6.66$.

k) Como 3 está en el IC anterior, está permitido creer que la varianza se disminuye a la tercera parte. Veamos ahora qué nos dice un test F: H_o : R=3.0, la varianza se disminuye a la tercera parte. $R_{Obs} = 2.66$, $F_{Obs} = 0.888$, $F_- = 0.4$, $F_+ = 2.10$, Se acepta la H_o pues F_{Obs} está entre F_- y F_+.

l) La novedad se acepta si uno disminuye la varianza 3 veces o más y no la acepta si la disminuye menos de 3 veces. varx =0.04, nx = 15, vary = 0.015, ny= 50, H_o : R=3.0, la varianza se disminuye a la tercera parte. la H_o se rechaza si la varianza se disminuye más que a la tercera parte, digamos a la cuarta o quinta parte. ¿Es cola superior o inferior? Para no enredarse con división de quebrados entre quebrados averiguamos los valores críticos de la F de ambos lados pero con α= 0.05 y una cola. $R_{Obs} = 2.66$, $F_{Obs} = 2.66/3 = 0.888$, $F_+ = 1.87$. $F_- = 1/2.2 = 0.45$. Se acepta la H_o pues F_{Obs} está entre F_+ y F_-. Eso significa que está permitido creer que la varianza se disminuyó 3 veces o menos. Igualmente está permitido creer que la varianza diminuyó a la quinta parte pues $2.66/5 = 0.532$ está entre F_+ y F_-. Pero no está permitido creer que la varianza se disminuyó a la séptima parte pues $2.66/7 = 0.38$ está abajo de F_-.

m) Asesoría gerencial: el IC de la relación entre varianzas nos permite creer que el nuevo procedimiento disminuye la varianza a la tercera, cuarta o quinta parte. Pero por otro lado también podemos decir que la varianza sólo disminuyó 1.5 veces o incluso menos. Esto es demasiado poco para pagar tan alto precio. Recomendamos no adoptarla o hacerlo en período de prueba (es posible que una pequeña disminución en la varianza del diámetro de las esferas produzca una mejora significativa en la calidad de los esferos) o renegociar el precio en consecuencia de lo poco que con elevada confianza se espera mejorar.

Pág 300, versión 1, No 5.a). A la izquierda de 79.2 queda el 69 %. Se mira en la tabla de la Z en la cuarta columna y da un z de 0.5. Al despejar de la transformada-z, nos da $\sigma = 33{,}6$

b) 0.6826.

Pág 301, versión 2, No 1. Hay dos grupos P (pómulos pintados) y L (labios pintados).

Ordenamiento: para P: 67 68 71 72 75 75 77 78 79 79 83 85 87 92 92 102 112. (17 datos)

para L: 81 87 90 94 94 95 98 99 100 100 101 103 107 117.

b) Grupo P: media = 82, mediana 79, modas 75, 79 y 92. Grupo L: media 97.57, mediana 98.5, y modas 94 y 100.

e) El rango intercuartilar para la serie de los pómulos es la diferencia entre el cuartil 3 menos el uno = 87-75 = 12. La aleta intercuartilar es el rango intercuartilar por 1.5 = (12)(1.5) = 18. Los extremos por arriba según el criterio del rango intercuartilar son los que están arriba del tercer cuartil más la aleta: 87 + 18 = 105. Los extremos por abajo son los que están por debajo del primer cuartil menos la aleta = 75-18 = 57. Por tanto, no hay extremos por abajo pero por arriba está el 112. Para labios pintados: Barrera inferior = 83.5; barrera superior = 111.5.

f) Las niñas de la selva se comportan igual que nosotros: les gusta ir cambiando de pinta pero de todas maneras tienen una predilecta, que en el caso de ellas es pintarse los labios de rojo.

Pág 301, versión 2, 2a) p = 0.897

2b. Sea x el tiempo y z(x) su transformada z. Entonces el tridente de z(x) = 0.1314 pues es lo que queda a la izquierda, es decir, por debajo de él. Se halla la z y después la x = 5.12.

Pág 301, versión 2, No 3a. [3.364, 3.736]

3b. n= 383 **3c**. No, pues las preguntas que dependen del TLC no necesitan con urgencia que la v.a. en estudio siga una distribución normal.

Pág 301, versión 2, 4) a) Hay 4 personas en el grupo 0. Media y desviación por semana: $\bar{X} = 1$, $s^2 = 0{,}666$; $s = 0{,}82$ Hay 6 personas en el grupo 1: $\bar{X} = 3{,}33$, s= 0.816. b) Frecuencias relativas de escuchar la BBC por día: g = 1/7 ; f = 3.3/7 ; $t_{obs} = 1{,}33$; $t_c = 2{,}681$ con gli = 12, significancia 0.05 y con una cola. No hay datos suficientes para alegar que los jóvenes escuchan más la BBC que los viejos.

c) p = 0.14286, error= 0.4; Int = [0, 0.5];

Pág 301, versión 2, 5) a + b) Esta pregunta no es de estadística inferencial sino de descriptiva porque sólo se refiere a dos muestras sin ninguna implicación poblacional. Uno mira los datos y concluye: mejoró su promedio.

c) No es acertado decir que como esta vez se lo propuso y mejoró entonces puede mejorar cuando se lo propone. La razón es que falta considerar el efecto de todos los factores no controlados, el azar, sobre el promedio. Por ejemplo, pudo mejorar porque de pura casualidad los exámenes fueron un poquito más fáciles que en promedio.

d) La pregunta c es de estadística inferencial pues relaciona los datos de dos muestras con una extrapolación a futuro que involucra un valor promedio poblacional, el de las notas de aquí en adelante. Sin embargo la pregunta está mal formulada pues no está la descripción del azar que se da por el valor de la varianza.

e) El papel de la estadística inferencial es filtrar el efecto del azar (factores no controlados) de los datos de la muestra sobre la toma de decisiones.

f) Sí puede decirse como algo histórico que el Halcón mejoró su regularidad en el estudio pues la varianza de la notas disminuyó. Esto es bueno pues si uno tiene un promedio de 3.5 y una varianza pequeña, es posible que haya pasado todas las materias. Pero si la varianza es grande tuvo 5 en una materia fácil y un 1 en una difícil: eso no sirve.

g) Este problema es de estadística inferencial pues se trata de una extrapolación a futuro, es decir, hacia valores poblacionales. Primero comparamos varianzas para saber si son iguales o diferentes, $\alpha = 0{,}02$; $F_{obs} = 7{,}56$; $F_{crt} = 8{,}75$; Var iguales. Luego comparamos medias sabiendo que las varianzas son iguales: $t_{obs} = 0{,}99$, $t_{crt} = 2{,}32$ on una cola. Aunque mejoró su promedio , no hay datos suficientes para decir que puede mejorarlo cuando se lo propone. h) Problema aparte: ¿Mejoró la regularidad, es decir, disminuyó su irregularidad, su varianza? Este problema es con una cola: $F_{obs} = 7{,}56$; Como no hay tabla para $\alpha = 0{,}02$ con una cola, tomamos la tabla con $\alpha = 0{,}025$ y una cola. Con gln = 5, gld = 6, $F_{crt} = 5{,}98$. Vemos que, a pesar de que esta vez mejoró su regularidad, no hay datos suficientes para pretender que el estudiante pueda mejorar su regularidad cuando se lo proponga.

Pág 302, versión 3, 1a) 88/100;

b) 12/100;

c) 34/100;

d) 35/54

Pág 302, versión 3, 2a) No puede asumirse que la distribución sea normal. Debe hacerse por Chevichev.

$P(\mid x - \mu \mid \geq k\sigma) \leq \frac{1}{k^2}$

$P(\mid x - 100 \mid \geq k(8)) \leq \frac{1}{k^2}$

Para los extremos del intervalo considerado tenemos:

$\mid 115 - 100 \mid = \mid 100 - 85 \mid = 15 = 8k$

por lo que $k = 1{,}875$. Por tanto, $k^2 = 3{,}515625$ y

$P(\mid x - 100 \mid \geq 1{,}875(8)) \leq \frac{1}{1{,}875^2} = 0{,}285$

b) $P(\mid x - 100 \mid \geq k(8)) \leq \frac{1}{k^2} = 0{,}05$ $k = \sqrt{\frac{1}{0{,}05}} = 4{,}472$. Por tanto, $k\sigma = 4.472 \times 8 = 35.777$. Las fronteras de los outliers son: 100- 35.777 = 64.333 y 100+35.777 = 135.777.

Pág 302, versión 3, 3a) 0.9641

3b) 0.9104;

3c) Si el estudiante está por encima de 0.9929, lo que esa por debajo de él es 0.9929 y ése es su tridente. Hallar la z y después la x = 1075.

Pág 302, versión 3, 4) Utilizaremos los IC para resolver todas las preguntas. IC de la relación entre varianzas: (2.3335, 14.3245). a) Como 1 no pertenece al IC de la relación entre varianzas, puede decirse que éstas son diferentes. b) Sabiendo eso, el IC para la diferencia de medias: (-0.5, 1.1). Como 0 está en ese intervalo, puede decirse que la diferencia entre medias poblacionales es cero, es decir que son iguales. b) Atención: aunque se puede decir que la diferencia de medias es cero, pueden existir muchos otros posibles valores. Ejemplo: como dicho intervalo tiene valores negativos, también puede creerse que los abogados ganan en promedio menos que las abogadas. Si se resuelve este punto por prueba de hipótesis: $\alpha = 0{,}02$; $F_{obs} = 5{,}89$; $F_{crt} = 2{,}17$; Var diferentes. $t_{obs} = -0{,}908$; gli = 43.77. Medias iguales. Para demostrar por prueba de hipótesis que es posible creer que en la población los abogados ganan menos que las abogadas, se puede tomar como H_o que la diferencia de medias es -0.1 c) Como el IC es el de la relación de la varianza poblacional grande a la pequeña y como el 5 está en ese intervalo, puede decirse que la varianza de los abogados es 5 veces mayor que la de las abogadas. d) Como 0.6 está en el IC de la media mayor menos la menor, puede decirse que las abogadas ganan en promedio 0.6 menos que los abogados. Este fenómeno de discriminción ha sido denunciado muchas veces. Quizá se deba a que los hombres quieren ante todo maximizar las ganancias, en tanto que las mujeres buscan otras cosas como la alegría de tener un buen trabajo y de disfrutarlo, lo cual bien puede competir con el sueldo.

Pág 302, versión 3, 5) a)

Esperado: Sueldo vs inteligencia				
	X = baja	Y = mediana	Z= alta	Totales
A=Alto	10.843	8.486	13.672	33
B = Bajo	12.157	9.514	15.329	37
Totales	23	18	29	70

$\chi^2_{Exp} = 14.0362$, gli = 2.0, $\chi^2_c = 5.9915$. No puede decirse que la proporción de A sea homogénea con respecto a las columnas. Por lo tanto, a un grado de inteligencia le va mejor que a los otros. b) Frecuencia de A en X = 10/23. Frecuencia de A en Z = 8/29. $t_{obs} = 1{,}3$. Los gli son el mínimo entre 17 y 29, es decir, 17. $t_c = z_c = 2{,}11$ con dos colas y significancia 0.05. Sí. c) Que a los de inteligencia media les va mejor que a los de inteligencia alta. d) Los muy inteligentes quieren cambiarlo todo y son muy pendencieros. Pero al hacer los cambios, las cosas les salen mal pues la calidad no se improvisa. Por eso, hacer fuertes inversiones en investigación y desarrollo es el distintivo de las potencias.

Pág 303, versión 4, 1) a) Se estudia $(0{,}3+0{,}7)^{20}$. La probabilidad de que al menos tres se lancen es $p(3)+p(4)+...+p(20) = 1-(p(0)+p(1)+p(2)) = 1-(0{,}0008+0{,}0068+0{,}0278) = 1-0{,}0354 = 0{,}9646$. Para calcular $p(k)$ se utiliza la fórmula del **coeficiente binomial** pues se trata de expansiones de grado 20 que no se pueden hacer por el triángulo de Pascal. b) Se estudia $(0{,}005+0{,}995)^{20}$. La p de a los sumo (máximo) $2 = p(o)+p(1)+p(2) = 0{,}905+0{,}091+0{,}004 = 1^-$.

Pág 303, versión 4, 2) $\bar{D} = -47$, $s = 78{,}403$, $\bar{\Delta} = 0$, t_{obs} = -1.34, gli = 4, $\alpha = 0{,}05$, t_c = +/-2.776. No ha habido cambio de hábitos.

Pág 303, versión 4, 3) **a)** $H_o : p - r = 0$. f= 0.3, g= 0.17, $\hat{p} = 0.24$, $t_{obs} = 1.3$. Como gli > 30, $t_c = 1{,}96$ con $\alpha = 0{,}05$. Es justo decir que la credibilidad siguió siendo la misma.

b) $H_o : p - r = 2r - r = 0{,}17$

t_{obs} = -0.433. $t_c = 1{,}96$ pues gli > 30. Es justo decir que la credibilidad se duplicó.

c) Para sopesar quién tiene razón, hallamos el IC de la diferencia de proporciones: (-0.05 , 0.31). Vemos que el cero está casi a un extremo, mientras que 0.17 (la diferencia de medias en el caso de haberse duplicado) está casi en la mitad del IC, la cual es 0.12. Por lo tanto, le damos nuestro voto a la propaganda. Incluso, no podemos excluír que la credibilidad haya aumentado un 250 %. Pero decir tal cosa pondrá a pensar a los crítricos que dirán: tampoco se puede excluír que la credibilidad haya disminuído un 10 %.

Pág 303, versión 4, 4) $\mu_x = 22$, μ_y= 15. Por tanto, $H_o : \mu_x - \mu_y = 22 - 15 = 7$, $\sigma^2_x = 5$, $\sigma^2_y = 6$, $\bar{X} = 10$, $\bar{Y}$= 6, $n_x = 7$, $n_y = 5$, z_{obs} = -2.168 $z_c = \pm 1.96$, dos colas pues pregunta si hubo o no cambio significativo. Caso 2 del diccionario: comparar la diferencia de medias observadas con la diferencia de media poblacionales sabiendo las varianzas poblacionales. Se rechaza la H_o. La diferencia

original entre medias era 7, la muestral fue de 4. El cambio de medias es significativo y evidencia un fuerte ablandamiento del grupo más violento: es una magnífica oportunidad para citarlos a un dialogo de paz. EL IC de la diferencia de medias es (1.3, 6.69) que se halló con las diferencia observada de medias. La diferencia esperada en la H_o era 7, que no está en el IC, por lo tanto la H_o se rechaza.

Pág 303, versión 4, 5) Completamos la tabla:

Observado: Consumo de cigarrillo vs. asistencia				
	X= Nada	Y=Mas o menos	Z=Mucha	Totales
A=Fuman mucho	12	8	14	34
B=Fuman poco o nada	13	21	9	43
Totales	25	29	23	77

a) p(A) = 34/ 77.

e) p(A/X) = 12/25. El universo es X con un total de 25, de él se toma la parte que está en A, que son 12.

h) Lo esperado bajo la H_o de independencia de factores:

Esperado: Consumo de cigarrillo vs. asistencia				
	X= Nada	Y=Más o menos	Z=Mucha	Totales
A=Fuman mucho	11.04	12.81	10.16	34.01
B=Fuman poco o nada	13.96	16.19	12.84	42.99
Totales	25	29	23	77

Como $\chi^2_{obs} = 5{,}984$ y $\chi^2_c = 5{,}9915$, los factores son independientes.

i) Primero: este problema puede resolverse con una tabla de contingencia cuyo objetivo es dilucidar la relación entre dos factores, porque si las proporciones de los que fuman mucho no varían con el grado de asistencia es porque el factor asistencia no está ligado con el factor fumar. Y viceversa, si los dos factores están desligados, las proprociones deben ser las mismas. ii) Segundo: aplique 3 veces el caso 5b del folleto de fórmulas. iii) Tercero: Resolvamos el problema usando un test de ajuste a una distribución uniforme. Usamos dicha distribución porque es la única que refleja la pretensión de que las proporciones de los que fuman mucho no varían por el grado de asistencia a las plenarias. Pregunta delicada: ¿Cuál es el universo: los que fuman mucho o todo el mundo, los que fuman mucho y los que no? Respuesta: están preguntado sobre la proporción de los que fuman mucho con respecto a todo el mundo y no dentro del grado de asistencia. Por tanto, el universo es todo el mundo, que tiene 77 representantes. De éstos, tomamos los que fuman mucho que son 34. Por tanto la H_o dice $p = 34/77$ y es la misma para todas las columnas. Ahora transformemos esta H_o en otra en términos de una distribución uniforme: si la probabilidad de fumar mucho es p, la cantidad total de personas esperadas que fuman mucho es $p \times Total = (34/77) \times 77 = 34$.

Ahora tenemos el siguiente dilema respecto a cómo debe entenderse la H_o de una distribución uniforme. Tenemos dos opciones. Primera: el total 34 debe distribuirse uniformemente entre todas las columnas, es decir, a cada columna le tocan $34/3 = 11.3$. Segunda, la probabilidad 34/77 debe ser la misma para todas las columnas pero la cantidad esperada en cada columna debe ser proporcional al tamaño de la muestra dentro de cada columna. De estas dos opciones tomamos la segunda: la distribución uniforme se refiere a probabilidades iguales quedando libre el tamaño de la muestra para cada columna. Computemos ahora los esperados bajo la H_o para cada columna. Los que no asisten a las plenarias son 25, de los cuales esperamos $25 \times 34/77 = 11.039$ que fumen mucho. Los que asisten más o menos son 29, de los cuales esperamos $29 \times 34/77 = 12.805$ que fumen mucho. Hay 23 personas que asisten a todo, de ellos esperamos $23 \times 34/77 = 10.156$ que fumen mucho. El total de los esperados que fuman mucho son $11.039 + 12.805 + 10.156 = 34$.

Ahora computamos la discrepancia entre lo que se cree bajo la H_o y lo observado. Dicha discrepancia se distribuye como una chi cuadrado:

$$\chi^2_{obs} = \frac{(12-11{,}039)^2}{11{,}039} + \frac{(8-12{,}805)^2}{12{,}805} + \frac{(14-10{,}156)^2}{10{,}156} = 0{,}084 + 1{,}80 + 1{,}45 = 3{,}33.$$

Para decir si dicha discrepancia es grande o chica, tomamos una significancia de 0.05 en la cola superior y nos queda por averiguar los gl. Hay 3 columnas, por lo que empezamos con 3 gl. Restamos uno pues hay una ecuación que restringe todo, a saber que la suma de todo es 34. Por tanto, los gli son 2:

$\chi_c^2 = 5{,}995$

Como la chi cuadrado experimental es menor que el crítico, la discrepancia entre lo que se esperaba según la H_o y lo que se observa es pequeña y no amerita un cuestionamiento de la H_o: no tenemos datos suficientes para alegar que las proporciones de los que fuman mucho estén ligadas al grado de asistencia a las plenarias.

i) Cuarto: para este problema y con significancia 0.05 los 3 métodos dan lo mismo, pero no en general porque la eficiencia de los métodos no es igual, es decir, la cantidad de datos necesaria para rechazar la H_o es diferente. El método de la tabla de contingencia tiene cara de ser el más eficiente pues es el que usa todos los datos disponibles al tiempo en cambio los demás usan sólo parte de la información.

j)$p(A/Z) = p(A \cap Z)/p(Z) = 14/23$.

k) $p(A/(X \cup Y)) = p(A \cap (X \cup Y))/(X \cup Y)) = 20/54$.

Índice alfabético

Z, 68
$p(A \cup B)$, 245
$p - value$, 290
x-crítico, 88
z_{obs}, 150

boxplot, 49

a lo sumo, 54
aleta cuartilar, 19

Biblia, 78
    de Jerusalén, 78
Binomial
    Efecto del muestreo, 102
boxplot, 18

campana que mejor se ajuste a los datos, 6
campanas de gauss, 6
censo, 269
ciencia
    blanda, 76
    dura, 75, 76
Ciencia Política , 94
coeficiente binomial, 44, 330
coeficiente de variación, 5, 10
cola
    inferior, 76
    superior, 76
colas
    dos, 119
    una, 119
condición
    de frontera, 101
    inicial, 101
confianza, 77, 96
conjunto de verdad, 244
cosmovisión, 20
criterio
    del rango intercuartilar, 19
criterio de la doble mitad, 60
cuartil, 15
cuartiles, 14

dato, 20
    categórico, 2
    cuantitativo, 2
    numérico, 2
    ordinal, 2
dato experimental, 86
datos
    apareados, 174
    extremos, 8
decil k, 128
densidad de probabilidad, 63
Derecho, 94
descriptor, 71
desviación, 60
desviación conjunta, 182
desviación estándar, 4
diagrama
    de barras, 1
diagrama de barras, 44
diagrama de barras de probabilidad, 35, 37
diagrama de bigotes, 18
diagrama de caja, 18
distribución
    ad hoc, 284
    binomial, 28, 30, 300
    de tendencia central, 18, 28
    exponencial, 30
    normal, 59
    normal estándar, 68
    uniforme, 30, 93, 283

error
    alfa, 127
    beta, 128
    de inferencia, 100
    tipo α, 100
    tipo β, 100
error estándar, 107
    de la estimación de la media, 210
estadígrafo, 5, 18, 157
    de contraste, 149, 151, 155
estadígrafos, 24
estadística
    descriptiva, 21
    inferencial, 21
    no paramétrica, 273, 274
    paramétrica, 273
estadística inferencial, 102
estadísticamente interesante, 77

estadístico, 5
evento, 20, 26
    interesante, 82
eventos, 243
    extremos, 19, 274
    independientes, 33, 245
experimento
    homogeneidad, 101, 226
    productividad, 101, 226
experimentos
    básicos, 207
    con varianzas, 300
extremo, 77

factor, 243
factorial, 45
función
    de densidad de probabilidad, 157
    de distribución, 62, 63
    de distribución de Z, 61

gli = grados de libertad, 156, 170
gráfica
    de cuadro, 18
    de cuadro y bigotes, 49

Hegel, 179
hipótesis
    alterna, 86, 87
    nula, 86, 87
histograma, 1
historia, 119

IC, intervalo de confianza, 95
    media, 96
    proporción, 99
Inferencia Estadística, 102

Kolmogorov-Smironov
    tabla, 289

la función densidad , 277
la probabilidad condicional, 244
la región de aceptación de la H_o, 86
la región de rechazo de la H_o, 86
la regla de tres, 18
la t, 158
Lenguaje, 94
Los 10 experimentos básicos de la normal, 300

marca de clase, 13
media, 4, 60
    o promedio, 4
    intervalo de confianza, 96
    muestral, 106, 107
Mediana, 3
mediana, 4, 15
medidor de discrepancia , 67
moda, 3, 4, 94
muestra, 20, 23, 102, 107
muestra al azar, 20
muestreo
    al azar, 300
    con reemplazo, 107, 240
    sesgado, 300
    sin reemplazo, 239

número de colas en un problema, 122
nivel de significancia, 77
norma, 94
normal vs. anormal, 76

outlier, 10, 72, 77, 274

p-value, 85
paquete
    Gnumeric, 18
    R, 19
parámetro, 59, 273
percentil del dato, 67
población, 20, 23, 102, 107
principio de responsabilidad, 100
probabilidad, 12, 26
probabilidades condicionales, 240
promedio, 21
promedio ponderado, 48
proposición, 244
protocolo, 101
    de medición, 20
    de muestreo, 20
proyecto, 19
prueba, 122
    dos colas, 92
    una cola, 92
puesto, 3

RA, 86
rango, 3
rango intercuartilar, 4, 19
realidad, 20
región de aceptación de la H_o, 143
región de rechazo de la H_o, 143
Regla de Bayes, 241
responsabilidad, 78
ruido, 101

significancia
    nivel de, 77
significancia de un evento, 85

tabla
    de contingencia
        Datos apareados, 295
    corrección de Yates, 248

de contingencia, 243
de frecuencias absolutas, 1, 10
de frecuencias acumuladas, 53
de frecuencias relativas, 1
tendencia central, 28, 29, 59
teorema
Chevichev, 274
del límite central, 107
test, 122
de ajuste, 283
dos proprociones, 172
dos varianzas, 176
homogeneidad, 251
Kolmogorov Smirnov, 286
los 10 básicos, 207
medias con var diferentes., 183
medias con var iguales, 179
medias datos apareados, 174
una media, σ conocida, 87, 165
una media, σ conocida, 168
una media, s conocida, 169
una proporción, 170
torta, 1
transformada
x, 67
z, 66, 69
tratamiento, 101

v.a. (variable aleatoria), 20, 28, 59, 107
continua, 20, 65
discreta, 20, 65
valor esperado, 28
variable, 20
varianza, 8, 21
muestral, 4
verdad , 20

Wittgenstein , 179

z
crítico, 77, 79, 87
observada, z_{obs} , 150