27. Respuestas y soluciones

□ Capítulo 3

3.1. Si tomamos como referencia el PBI por habitante, la relación se reduce a que el PBI por habitante de Brasil es el 84% del de Argentina.

3.2. La cantidad de conductores es generalmente mayor entre las 18 h y 20 h (hora pico) que entre las 14 h y las 16 h. Es esperable que la cantidad de accidentes sea mayor. La proporción de accidentes es una medida más adecuada que la cantidad en este caso.

3.3. Se trata de noticias como:

- Los argentinos comen muy mal: exceso de carne y poca verdura. Se come un 75% más de carnes rojas de lo recomendado y un 46% menos de verduras.
- La cena navideña es 10 veces más calórica de lo sugerido.
- Referido al fútbol. Entre los cinco grandes suman 87 campeonatos, el 76% de los títulos.

3.4. Se trata de avisos como los de los ejercicios 5 y 7.

3.5. ¿Qué significa que la piel naranja sea menos visible? ¿Cómo se mide? ¿El producto fue efectivo en todas las mujeres? ¿Cómo se mide que una piel esté lisa? ¿Cómo se obtuvo el -1,9cm? ¿Es un promedio? ¿Es la reducción mínima? ¿Es la reducción máxima? ¿Las 44 que realizaron la autoevaluación están entre las 50 que participaron del test clínico?

3.6. Explique las siguientes frases:

- Le puedo pagar a lo sumo $500 por ese trabajo: Le puede pagar $500 pero no más de ese valor
- Le voy a pagar como mínimo $500 por ese trabajo. Le puede pagar $500 y tal vez más.
- Quiero que vuelvas como máximo a las 11 de la noche. A las 11 o antes de esa hora.
- Se presentaron por lo menos 10 personas para el puesto de encargado de control de calidad. Se presentaron 10, 11, 12, etc. pero no se sabe cuántos exactamente.
- No más de 10 personas se presentaron para el puesto de chofer. Pueden haberse presentado: ninguna, una, dos, ..., ó 10 personas pero no más.

3.7. "Un chico de 8 a 12 años puede perder hasta un litro de transpiración durante dos horas de actividad un día caluroso", afirma una publicidad.

- Un procedimiento para estimar cuanto líquido puede perder un chico por transpiración durante dos horas:

Pese varios chicos antes y después de realizar una actividad física durante 2 horas.

- "Hasta dos litros" significa que puede
- no perder nada
- perder 1 litro
- perder 2 litros
- perder 1,5 litros

Capítulo 6

6.1.

- Una encuesta de opinión contacta a 1.243 adultos y les pregunta, ¿ha comprado un billete de lotería en los últimos 12 meses? **Población:** todos los adultos. No aclara ninguna característica. **Muestra:** 1.243 adultos.
- Durante la reunión anual del colegio de abogados, todos los presentes (2.500), llenaron una encuesta referida al tipo de seguro que prefería para su automóvil. **Población:** 2.500 miembros del colegio de abogados.
- En 1968 se realizó en Holanda un test de inteligencia a todos los varones de 18 años que estaban realizando el Servicio Militar Obligatorio. **Población:** todos los varones de 18 años de Holanda en 1968.
- El INDEC lleva a cabo la Encuesta Permanente de Hogares (EPH) en la que se encuestan 25.000 hogares para captar información sobre la realidad económico-social de la República Argentina. **Población:** todos los hogares. **Muestra:** 25.000 hogares.

6.2.

- Sesgos por subcubrimiento. Podrían, por ejemplo, quedar sub-representados sectores sociales con mayores dificultades para movilizarse o aquellos que piensan que van a perder ya que tendrán menor entusiasmo para realizan el esfuerzo de ir a votar.
- Sesgo para tratar de agradar o para no quedar mal con los demás miembros del bloque.

6.3.

Indique cuál es el tipo de muestreo realizado en cada caso.

- Cada alumno escribe su nombre en un papel, lo pone en una bolsa y el director elige 100 papeles. Se trataría de un muestreo aleatorio simple, pero con este procedimiento surge la pregunta de cuán bien fueron mezclados los nombres en la bolsa. Podrían salir elegidos más alumnos de un curso porque los papeles estarían juntos.
- A cada alumno se le asigna un número entre 1 y 2.500 y se seleccionan generando 100 números al azar de cuatro dígitos utilizando algún programa de computación: muestreo aleatorio simple.
- Para cada año, se asigna a cada alumno un número entre 1 y 500, y se elige 1 de cada 25 alumnos: muestreo sistemático.
- Se eligen al azar una división de cada uno de los años y se seleccionan 20 alumnos de cada división: muestreo en dos etapas.
- Se eligen al azar 60 alumnos de los primeros 3 años y 40 alumnos de los últimos dos años: muestreo estratificado.

- Se eligen al azar 60 alumnos de los primeros 3 años y 40 alumnos de los últimos dos años. Se seleccionan en forma separada los varones y las mujeres de acuerdo con la proporción de mujeres y varones que tiene la escuela: muestreo proporcional al género.

6.4.

a) Que la encuesta no dice nada porque arrastra el sesgo por respuesta voluntaria.

6.5.

a) VÁLIDA
b) NO VÁLIDA
c) VÁLIDA
d) NO VÁLIDA
e) VÁLIDA
f) NO VÁLIDA

□ Capítulo 7

7.1.

a. Ambas dimensiones de la figura, alto y ancho, fueron reducidas a la mitad, el área es 0,5 x 0,5 = 0,25 veces más pequeña. Nuestros ojos responden al área que se redujo a la cuarta parte. Para que el área se reduzca a la mitad la altura y el ancho deben corregirse por el factor $\sqrt{0,5} = 0,707$. Es importante que la relación de los valores se muestre mediante áreas; son éstas las que realmente producen la impresión visual.

b. La relación entre los valores de los bancos está magnificada, produciendo una sensación de mayor caída que la real. El artista representó los valores de los bancos por los diámetros de los círculos. Por ejemplo, para el BNP Paribas la relación entre los diámetros es

$$\frac{D_2}{D_1} = \frac{108}{32,5} = 3,32$$

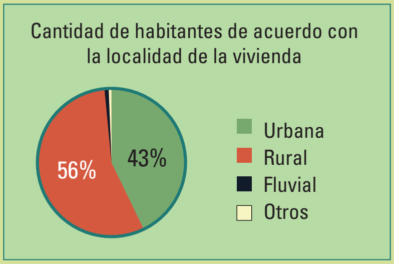
(D_1 = diámetro del círculo menor y D_2 = diámetro del círculo mayor). Los radios mantienen la misma relación,

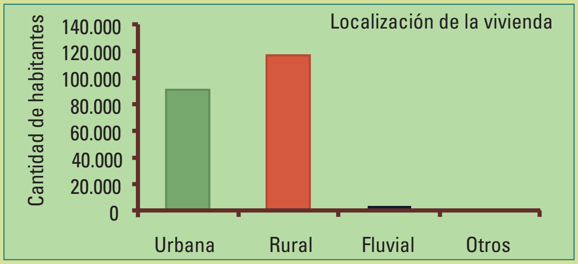
$$\frac{R_2}{R_1} = 3,32$$

El cociente entre las respectivas áreas es $$\frac{\text{Área1}}{\text{Área2}} = \left(\frac{R_2}{R_1}\right)^2 = 3,32^2 = 11,02$$

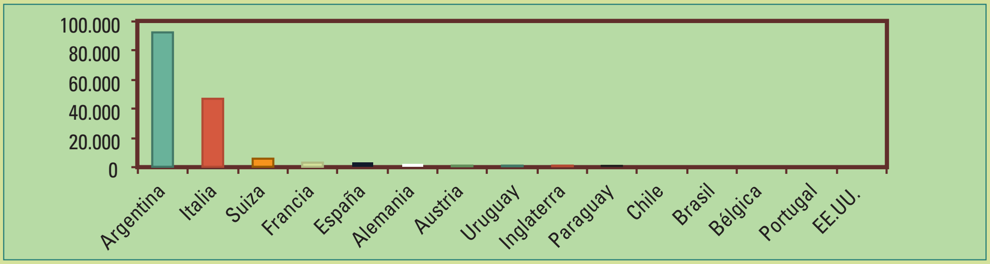
Visualmente el año 2007 aparece como 11 veces mayor, cuando en realidad lo es un poco más de 3 veces.

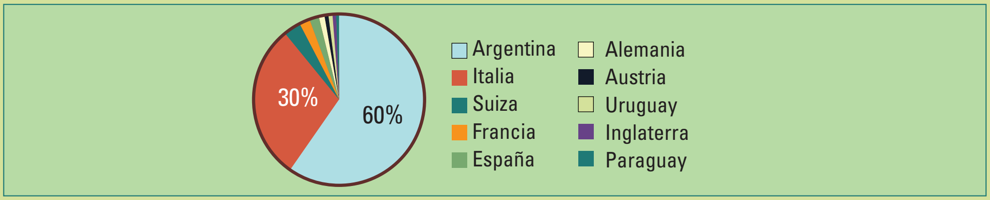
7.2. Los valores de las tres variables consideradas (Alfabetización, Nacionalidades y Localización) pueden representarse tanto en un gráfico de barras, como en un diagrama circular porque se trata de partes de un total.



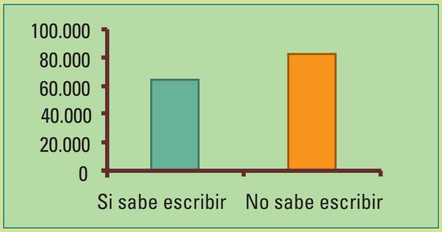


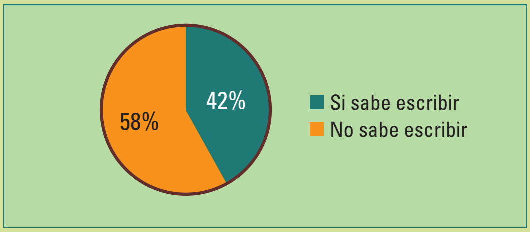
El 99% de la población es Urbana o Rural





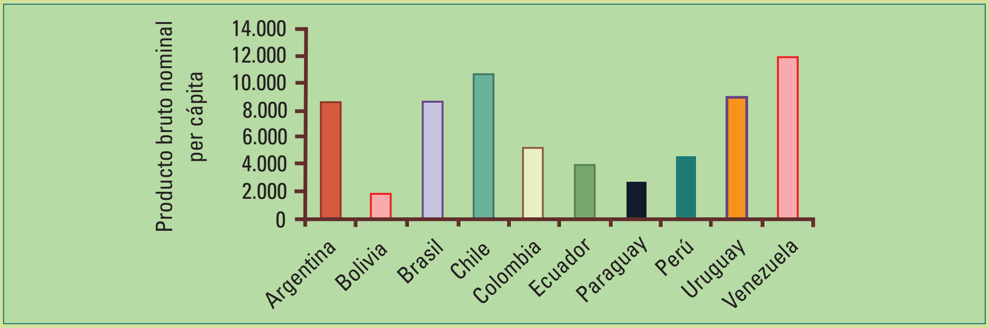
Dos nacionalidades abarcan al 90% de la población.





En 1887, más de la mitad de la población de la Provincia de Santa Fe no sabía escribir.

7.3. La altura de las barras representa el producto bruto total de cada país dividido por la cantidad de habitantes.



No se puede reemplazar la figura anterior por un gráfico circular porque los datos no representan una parte de un total.

Capítulo 8

8.1.

- ¿Están contentos los alumnos con el nuevo sistema de promoción? Una encuesta, porque interesa conocer la opinión de los alumnos sin modificarla.
- ¿El ausentismo de los alumnos es menor en verano que en invierno? Estudio observacional que no es una encuesta, se trata de describir un comportamiento.
- ¿El rendimiento de los alumnos en un examen es mejor si durante el mismo escuchan música de Vivaldi, en bajo volumen, en comparación con no escuchar nada? Estudio experimental, porque interesa comparar los resultados de aplicar dos "tratamientos": 1) con música de Vivaldi, 2) sin música de Vivaldi.

8.2. Similares a los del ejercicio 1.

8.3. Es un experimento porque está eligiendo a los niños y niñas al azar para dividirlos en los dos grupos a los que les enseñará utilizando canciones y a los que no. Ni los niños (ni sus padres) eligen en qué grupo participar.

Capítulo 9

9. 1.
Unidad muestral: una arandela
Variable: continua, diámetro de una arandela
Tamaño de la Muestra: 100

Población: todas las arandelas del lote
Parámetro: 1,908 cm
Estadístico: media muestral
Valor del estadístico: 1,915 cm

9. 2.
Unidad muestral: familia, el enunciado no especifica ni el año ni la región.
Variable: categórica con dos categorías: 1) madre sabe que un antibiótico no puede curar un resfrío, 2) madre no sabe que un antibiótico no puede curar un resfrío
Tamaño de la Muestra: 213
Población: todas las familias
Estadístico: porcentaje de madres que saben que un antibiótico no puede curar un resfrío
Valor del estadístico: 40%

9. 3.
Unidad muestral: hogar de la Argentina en el año 2001
Variable: categórica con dos categorías: 1) hogar heladera con freezer, 2) hogar sin heladera con freezer
Población: todos los hogares de la Argentina en el año 2001
Parámetro: porcentaje de hogares con heladera con freezer
Valor del parámetro: 50%

9. 4.
Unidad muestral: un auto
Variable: continua, el precio de un auto
Tamaño de la Muestra: 8
Estadístico: media muestral
Valor del estadístico: $ 21.880

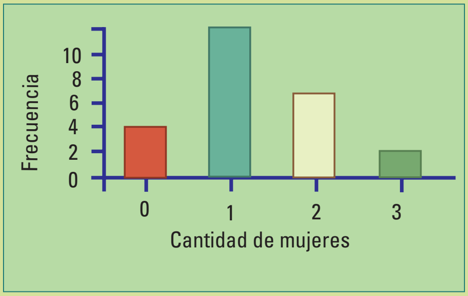
□ Capítulo 10

10.1.
a) 51 % no sorprendería pero 37% sí.
b) Ninguno de los dos porcentajes sorprendería si el tamaño de muestra fuera 35.

10.2.
i. (c) ii. (b) iii. (a) iv. (d)

10.3. (c) preciso y sin sesgo, a) preciso y con sesgo d) impreciso y sin sesgo, b) impreciso y con sesgo.

10.4. A continuación presentamos como ejemplo la cantidad de mujeres de cada muestra, al seleccionar 25 muestras de tamaño 3 de un curso que tiene 15 mujeres y 20 varones:



1 1 1 2 1 0 2 2 1 0 2 1 1 3 1 2 1 2 1 0 2 3 1 0 1 con un promedio de 1,28 mujeres por muestra de tamaño 3. Tenemos 4 muestras con 0 mujeres, 12 muestras con 1 mujer, 7 muestras con 2 mujeres y 2 muestras con 3 mujeres. Por azar se pueden obtener 4 entre 25 muestras con 0 mujeres. Por lo tanto, si la muestra seleccionada no tiene mujeres, no hay razones para sospechar discriminación.

10.5.

a) El margen de error para una confianza del 95% es aproximadamente $\frac{1}{\sqrt{n}}$

donde n es el tamaño de la muestra. La diferencia se debe a haber seleccionado distinta cantidad de varones y mujeres para realizar las entrevistas.

b) Es necesario incluir el margen de error debido a la variabilidad entre muestra y muestra.

□ Capítulo 13

13.1. Las tres respuestas son correctas. a) y b) son parte de la definición de estudios experimentales y observacionales. Una encuesta no impone ningún tratamiento, simplemente cuenta la frecuencia con la que aparece una respuesta, por lo tanto es observacional.

13.2. a) y b) son correctas y c) no. En un experimento tanto el grupo tratado como el grupo control son elegidos por el investigador. Es mejor cuando estos grupos son seleccionados al azar.

13.3. b) es el correcto. El primer estudio es observacional porque los sujetos no fueron asignados por el investigador al tratamiento (realización de actividad física).

13.4. a) es el correcto. El primer estudio es un estudio experimental con dos grupos tratamiento (1 litro; 2 litros) sin grupo control. El segundo es un estudio observacional porque el investigador no fijó cuanta leche se compraba en la casa para cada niño.

13.5 e) es el correcto. Se realizó un experimento en el cual los investigadores dividieron a los sujetos en grupo tratamiento y grupo control. Un censo debería estudiar a todos los que sufren de dolores de cabeza y no sólo a 20. Fueron comparadas las respuestas del grupo tratamiento (recibió chocolate) con la del grupo control (recibió un placebo). Las tabletas con gusto a menta sin chocolate eran el placebo.

13.6. a) Censo. Consulte en la dirección de la escuela donde podrán darle esa información.

Capítulo 15

15.1. Ambas preguntas se refieren a recodar nombres, esto no es una capacidad vinculada con la resolución de problemas, por lo tanto no son válidas para medir la inteligencia.

15.2. Comparar tasas de muerte, es decir la cantidad de muertes dividido la cantidad total de chicos que viaja en cada uno de los medios de transporte.

15.3.
a) La cantidad de muertes por cáncer aumenta a medida que la población se vuelve mayor. b) El porcentaje de muertes por cáncer aumenta a medida que la salud de la población mejora en general y no se muere por otras causas. c) El tiempo de supervivencia podría aumentar porque la enfermedad fue detectada antes; es decir porque el método diagnóstico (y no el tratamiento) es más efectivo.

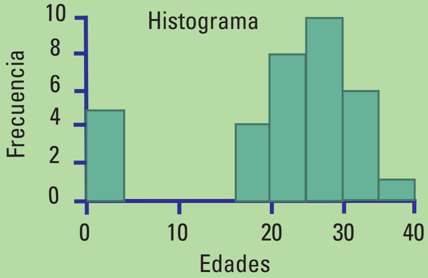
15.4. Una manera claramente inválida de medir "estado físico" sería preguntar si le interesa o no le interesa la política. Una forma válida consiste en comparar las pulsaciones por minuto de la persona en posición acostada y luego parada, (manteniendo 5 minutos de reposo en cada una de las posiciones antes de realizar las mediciones). Cuanto menor sea la diferencia mejor será su estado físico.

Capítulo 17

17.1. Tanto el diagrama tallo-hojas como el histograma muestran dos grupos de edades. El grupo de edades más pequeñas corresponde a la de los hijos, se trata de valores concentrados en el primer intervalo de clase. Las edades de los padres y madres se encuentran distribuidas en 5 intervalos de clase.

Diagrama tallo - hojas

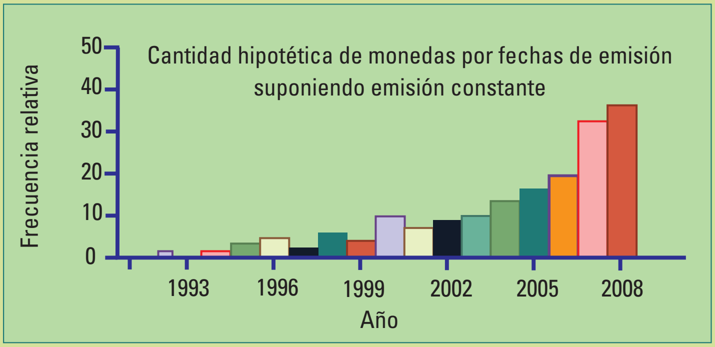
0 / 11122
0 /
1 /
1 / 899
2/ 0113334
2 / 5566678899
3 / 0012244
3 / 58



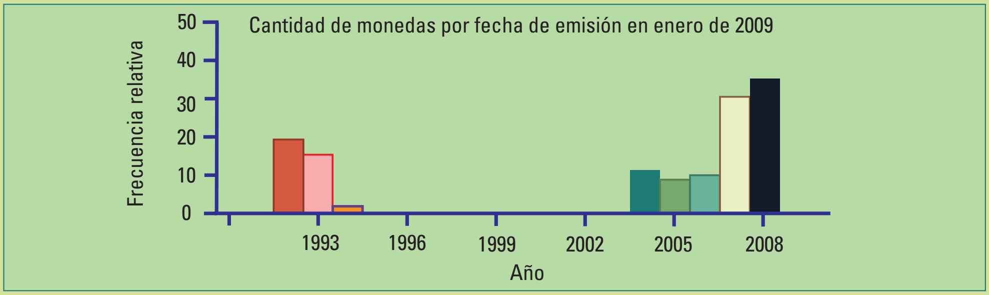
17.2. La mayoría de las familias no tendrán hijos y algunas tendrán hijos pequeños; aparecerá un pequeño promontorio en el lado más bajo de edades además del promontorio mayor correspondiente a las edades de los padres (c). Esta forma es similar a la observada en el diagrama tallo-hoja y el histograma del ejercicio 1.

17.3. Si la emisión de monedas fuera la misma todos los años y se perdieran una cierta proporción por año, se espera encontrar pocas monedas viejas y muchas nuevas. En este caso la distribución de las fechas de emisión tendría cola pesada a izquierda como se muestra en el siguiente histograma.

Año	Frecuencia	Frecuencia relativa
1992	20	0,15
1993	17	0,12
1994	1	0,01
2004	11	0,08
2005	9	0,07
2006	10	0,07
2007	33	0,24
2008	37	0,27



Utilizando 138 monedas de 10 centavos, seleccionadas en enero de 2008 obtenemos la siguiente distribución de frecuencias de la fecha de emisión:



El histograma resultante tiene una forma diferente a la esperada. Se observan dos grupos de fechas de emisión bien separados. El primero corresponde a las monedas emitidas entre los años 1992 y 1994, el segundo desde 2004 hasta 2008 estas últimas con una distribución más parecida a lo esperado. No encontramos monedas con año de emisión entre 1995 y 2003. Este resultado nos lleva a plantear nuevas preguntas, tal como suele ocurrir al realizar un análisis estadístico.

17.4. Para estudiar las longitudes de las palabras, seleccione un artículo de una revista de deportes y otro de una de divulgación científica. Para cada uno de los artículos obtenga:

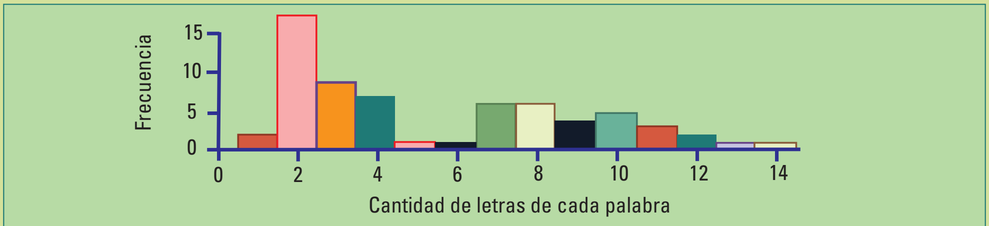
- la distribución de frecuencias
- la distribución de frecuencias relativas
- el histograma

de la variable "cantidad de letras" que tiene cada palabra. Compare las distribuciones obtenidas.

Lo haremos a continuación, como ejemplo, con el enunciado de este ejercicio:

Cantidad de letras por palabra	Palabras	Cantidad de palabras (Distribución de frecuencias)	Distribución de frecuencias relativas
1	y, a	2	0,03
2	de, un, de, de, de, de, de, la, la, de, de, el, de, la, de, lo, el, de	18	0,27
3	las, las, una, una, uno, los, que, las, con	9	0,14
4	Para, otro, Para, cada, cada, como, este	7	0,11
5	tiene	1	0,02
6	letras	1	0,02
7	revista, obtenga, palabra, Compare, haremos, ejemplo	6	0,09
8	estudiar, palabras, artículo, deportes, variable, cantidad	6	0,09
9	artículos, relativas, obtenidas, enunciado	4	0,06
10	longitudes, seleccione, científica, histograma, ejercicio	5	0,08
11	divulgación, frecuencias, frecuencias	3	0,05
12	distribución, distribución,	2	0,03
13	continuación	1	0,02
14	distribuciones	1	0,02
Total		66	1,03

Las frecuencias relativas no suman 1 por errores de redondeo.



Como la "cantidad de letras" es una variable numérica discreta los intervalos de clase del histograma están centrados en cada uno de los valores de la variable. El histograma muestra que el enunciado de este ejercicio tiene dos grupos de palabras, uno está formado por palabras cortas y el otro con palabras largas: el 55% son de a lo sumo 4 letras, el 40 % tiene entre 7 y 12 letras.

Un análisis similar puede realizarse con diferentes tipos de textos y en diferentes idiomas.

Capítulo 18

18.1.

e) Los diagramas tallo-hoja y los histogramas pueden mostrar los detalles que se esconden al calcular medidas resumen.

18.2.

d) La media, el desvío estándar, el máximo menos el mínimo, todos se ven afectados por la presencia de datos atípicos; la mediana y la distancia intercuartil, no.

18.3.

f) El desvío estándar sólo puede ser cero si todos los datos son iguales.

18.4.

b) Al dividir a todos los datos por 2, quedará el 20% entre 5 y 20. Luego al sumarles 10 quedará un 20% entre 15 y 30.

18.5) y 18.6)

a. Los **pesos** de varones y mujeres **por separado** tendrán medias y medianas similares si no hay alumnos o alumnas con pesos atípicos (es decir con pesos extremadamente bajos o extremadamente altos). Estos datos atípicos se detectan utilizando los gráficos - caja (box-plots). Si hay datos atípicos el desvío estándar puede ser demasiado grande y no representar a la mayoría de los pesos. Para los **pesos** de varones y mujere**s juntos,** las medidas resumen (media o mediana; distancia intercuartil o desvío estándar) pueden no ser una buena representación de los datos, porque se trata de dos grupos. En consecuencia el gráfico caja, que se obtiene a partir de medidas resumen, podría no dar una buena representación de los datos.

b. Registre la edad correspondiente a una fecha determinada. Por ejemplo el día que comienza el estudio. Utilice el formato de **edad decimal** expresada en años: Fecha del estudio: 29 de junio de 2008, Fecha de nacimiento: 15 de enero de 1992.

	Año	Mes	Día
Fecha del Estudio	2008	6	29
Fecha de nacimiento	1992	1	15
Edad	16	5	14

Edad: 16 años 5 meses y 14 días. Debemos convertir 5 meses y 14 días a años. Utilizaremos una forma sencilla que no tiene en cuenta la diferencia de días de los meses y los años bisiestos:

$$\text{Edad} = 16 + \frac{5}{12} + \frac{14}{365}$$

$$= 16{,}455 \text{ años.}$$

No deberían observarse demasiadas diferencias en las distribuciones de las edades de todos juntos y separados, varones y mujeres, como ocurre con peso. La media y el desvío estándar pueden resultar adecuadas.

18.7. Compare los resultados de los distintos años utilizando las medidas resumen que resultaron las más adecuadas para los pesos y para las edades y también histogramas.

18.8. La encuesta que realice en su división no será representativa de las opiniones de los demás años. Alumnos de años superiores suelen tener opiniones diferentes que los más chicos.

18.9. Para obtener **una muestra representativa de todos los años y de género**, utilice los registros de la dirección. Tome listados de varones y mujeres por separado. Puede utilizar un muestreo sistemático eligiendo, por ejemplo, uno de cada 10 alumnos/as.

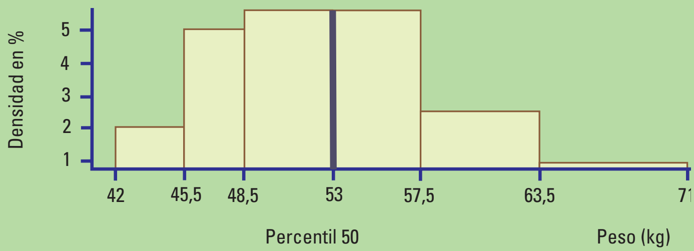
□ Capítulo 19

19.1. La figura 19.4 es una representación gráfica de los percentiles de los pesos. El percentil más pequeño, de 42 kg, es el del 3% indicando que hay un 3% de todas las mujeres de 16 años que tienen un peso menor. No sabemos cuánto menor, no sabemos cómo se distribuye ese 3% por debajo de 42 kg. Lo mismo ocurre con los pesos más altos que 71 kg, un 3% de los pesos de la población de mujeres de 16 años es mayor a 71 kg.

19.2. Complete la tabla de frecuencias siguiente utilizando la información de la figura 19.4

Intervalo de peso (kg)	Longitud del Intervalo	Frecuencia Relativa en %	Densidad Frecuencia Relativa en % / Longitud del Intervalo
[42 ; 45,5)	3,5	10 - 3 = 7	7 / 3,5 = 2,000
[45,5 ; 48,5)	3,0	25 - 10 = 15	15 / 3 = 5,000
[48,5 ; 53)	4,5	50 - 25 = 25	25 / 4,5 = 5,555
[53 ; 57,5)	4,5	75 - 50 = 25	25 / 4,5 = 5,555
[57,5 ; 63,5)	6,0	90 - 75 = 15	15 / 6 = 2,500
[63,5 ; 71)	7,5	97 - 90 = 7	7 / 7,5 = 0,933

19.3.



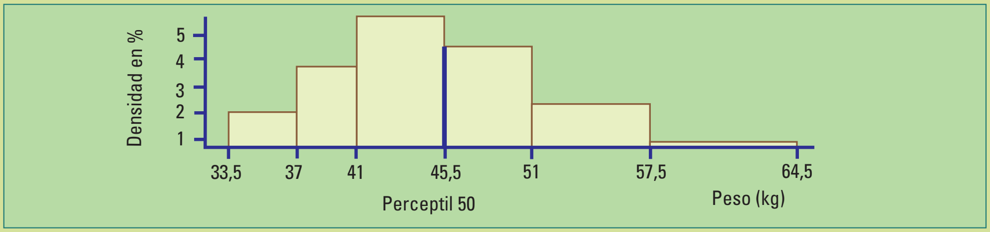
El área del histograma es la suma de las áreas de cada uno de los rectángulos de clase. Da como resultado 94 como se muestra a continuación:

Longitud de la base	Densidad (altura)	Área = Long. de la base x altura
3,5	7 / 3,5 = 2,000	3,5 x 7 / 3,5 = 7
3,0	15 / 3 = 5,000	3 x 15 / 3 = 15
4,5	25 / 4,5 = 5,555	4,5 x 25 / 4,5 = 25
4,5	25 / 4,5 = 5,555	4,5 x 25 / 4,5 = 25
6,0	15 / 6 = 2,500	6 x 15 / 6 = 15
7,5	7 / 7,5 = 0,933	7,5 x 7 / 7,5 = 7
Total		94

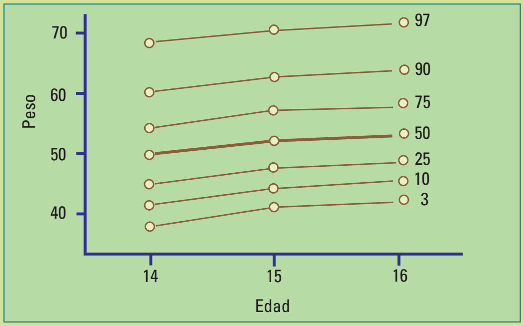
19.4. Percentiles del peso (kg)
13 años (mujeres)

Percentil	10	25	50	75	90	97
Peso	37	41	45,5	51	57,5	64,5

Intervalo de peso (kg)	Longitud del Intervalo	Frecuencia Relativa en %	Densidad Frecuencia Relativa en % / Longitud del Intervalo
[33,5 ; 37)	3,5	10-3 = 7	7 / 3,5 = 2,0000
[37 ; 41)	4,0	25-10 = 15	15 / 4 = 3,7500
[41 ; 45,5)	4,5	50 - 25 = 25	25 / 4,5 = 5,5550
[45,5 ; 51)	5,5	75-50 = 25	25 / 5,5 = 4,5454
[51 ; 57,5)	6,5	90-75 = 15	15 / 6,5 = 2,3100
[57,5 ; 64,5)	7 ,0	97-90 = 7	7 / 7 = 1,0000



19.5. y 19.6. La construcción de los diagramas tallo hoja permite ordenar los datos, esto facilita la obtención de los percentiles. Construya una tabla como la Tabla 19.2 y siga los pasos descriptos en la sección 19.1.2.



19.7. y 19.8. Debería obtener figuras similares a la siguiente:

En general, cada uno de los percentiles deberían aumentar con la edad; pero como se trata de diferentes alumnos en diferentes divisiones, pueden aparecer situaciones excepcionales y, entonces, no se cumpla ese aumento.

□ Capítulo 22

22.1.

a) Desde -1 hasta 1.

b) Cualquier número positivo.

22.2.

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{1}{n-1} \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1) \sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1)}}$$

$$= \frac{1}{n-1} \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1)}}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

22.3.

d) El coeficiente de correlación no cambia al sumar el mismo número a una de las variables o al multiplicarla por el mismo número.

22.4.

c) El coeficiente de correlación positivo muestra una tendencia de los valores mayores de una variable a estar acompañados de los valores mayores de la otra, pero para dos puntos en particular del diagrama todo es posible.

22.5.

c) Los tres diagramas muestran patrones no lineales fuertes. Sin embargo el coeficiente de correlación mide el grado de asociación lineal. Los dos primeros tienen r = 0 y el tercero un valor cercano a -1.

22.6.

b) Como el punto (3,11) pertenece a la recta y = 2 + b x, 11 = 2 + b x 3 entonces b= 3. Luego □ = 2 + 3.

22.7.

e) El coeficiente de correlación no puede ser mayor a 1.

22.8. El coeficiente de correlación se calcula como

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Vemos que las x's y las y's son intercambiables en la expresión anterior; no importa a los valores de cual de las variables llamemos x y a cual y.

Veamos ahora que cambiar de unidades producirá el mismo efecto en el numerador y el denominador y el resultado será el mismo:

Un cambio de unidades tiene la forma $z \rightarrow c\,z + d$. Si realizamos este cambio de unidades tendremos:

$$\frac{1}{n}\sum_{i=1}^{n}(cx_i + d) = c\,\frac{1}{n}\sum_{i=1}^{n}x_i + \frac{1}{n}\sum_{i=1}^{n}d = c\bar{x} + d \qquad \frac{1}{n}\sum_{i=1}^{n}(cy_i + d) = c\,\frac{1}{n}\sum_{i=1}^{n}y_i + \frac{1}{n}\sum_{i=1}^{n}d = c\bar{y} + d$$

Por lo tanto el coeficiente de correlación en las nuevas unidades es igual al de las unidades originales:

$$r = \frac{\sum_{i=1}^{n}((cx_i + d) - (c\bar{x} + d))((cy_i + d) - (c\bar{y} + d))}{\sqrt{\sum_{i=1}^{n}((cx_i + d) - (c\bar{x} + d))^2 \sum_{i=1}^{n}((cy_i + d) - (c\bar{y} + d))^2}} = \frac{c^2\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{c^2\sum_{i=1}^{n}(x_i - \bar{x})^2 c^2\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Finalmente como la media y el desvío son influidas fuertemente por un valor atípico, el coeficiente de correlación también.

22.9. Siga los lineamientos del ejemplo 1 de este capítulo. Elija al azar, por ejemplo 10 alumnos de cada año y solicite que registren durante un mes la cantidad de horas que dedican cada día a actividades sedentarias (mirar televisión, estudiar o utilizar la computadora) y las promedien. Obtenga medidas resumen como las de la tabla 22.2 y gráficos para comparar la distribución de las horas y la edad entre varones y mujeres. Construya diagramas de dispersión de las horas y la edad para varones y mujeres por separado y evalúe si hay diferencias. Ajuste una recta de cuadrados mínimos en cada uno de los diagramas y obtenga el coeficiente de correlación. Estime la diferencia en la cantidad de horas dedicadas a actividades sedentarias, a cada edad, entre varones y mujeres como se realizó en el ejemplo 9 de este capítulo.

□ Capítulo 23

23.1.
a) Cuanto mayor sea el tamaño de la muestra menor será el desvío estándar de la distribución de muestreo.
b. Es válido en general, para poblaciones grandes y muestras pequeñas en comparación con el tamaño de la población.
c. Vale siempre.
d. Es el resultado del TCL.

23.2.
a) Verdadera
b) Falsa. La distribución de muestreo de $\hat{\mathbf{p}}$ tiene un desvío estándar igual a $\sqrt{\frac{p(1-p)}{n}}$.
c. Falsa. La distribución de muestreo de $\hat{\mathbf{p}}$ es aproximadamente Normal cuando np o n (1-p) son suficientemente grandes (se suele tomar 5 ó 10).

23.3. Éxito = salió 1 ó 2 al arrojar un dado.

a) Se arroja el dado n = 3 veces, se repite 60 veces y se obtienen los siguientes resultados:

Repetición nº	Cantidad de éxitos	$\hat{p}$
1	2	0,67
2	1	0,33
3	0	0,00
4	1	0,33
5	2	0,67
6	1	0,33
7	0	0,00
8	2	0,67
9	1	0,33
10	0	0,00
11	2	0,67
12	1	0,33
13	0	0,00
14	0	0,00
15	0	0,00
16	1	0,33
17	0	0,00
18	0	0,00
19	0	0,00
20	1	0,33

Repetición nº	Cantidad de éxitos	$\hat{p}$
21	0	0,00
22	1	0,33
23	0	0,00
24	0	0,00
25	1	0,33
26	0	0,00
27	2	0,67
28	2	0,67
29	2	0,67
30	1	0,33
31	1	0,33
32	1	0,33
33	2	0,67
34	3	1,00
35	2	0,67
36	1	0,33
37	1	0,33
38	0	0,00
39	0	0,00
40	0	0,00

Repetición nº	Cantidad de éxitos	$\hat{p}$
41	1	0,33
42	1	0,33
43	0	0,00
44	1	0,33
45	1	0,33
46	0	0,00
47	1	0,33
48	2	0,67
49	1	0,33
50	2	0,67
51	0	0,00
52	2	0,67
53	2	0,67
54	1	0,33
55	0	0,00
56	2	0,67
57	1	0,33
58	1	0,33
59	2	0,67
60	1	0,33

Al arrojar el dado 3 veces puede haber: ningún éxito, 1 éxito, 2 éxitos y 3 éxitos. Por lo tanto en la columna cantidad de éxitos solamente aparecen cuatro números (0; 1; 2 y 3) y también son cuatro los valores de la proporción de éxitos (0; 0,33; 0,77; 1)

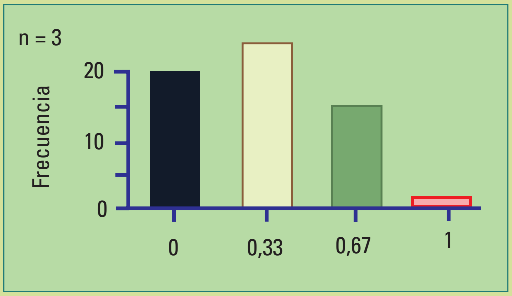
b) Se arroja el dado n = 30 veces, se repite 60 veces y se obtienen los siguientes resultados:

Repetición nº	Cantidad de éxitos	$\hat{p}$
1	10	0,33
2	11	0,37
3	12	0,40
4	10	0,33
5	10	0,33
6	10	0,33
7	13	0,43
8	14	0,47
9	4	0,13
10	10	0,33
11	16	0,53
12	3	0,10
13	9	0,30
14	9	0,30
15	13	0,43
16	10	0,33
17	11	0,37
18	9	0,30
19	8	0,27
20	14	0,47

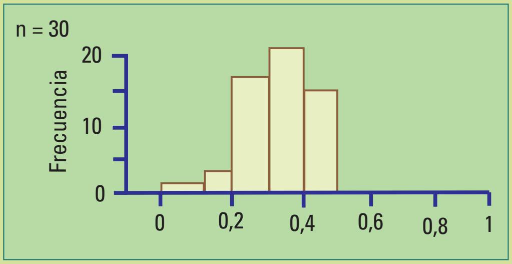
Repetición nº	Cantidad de éxitos	$\hat{p}$
21	10	0,33
22	13	0,43
23	13	0,43
24	9	0,30
25	12	0,40
26	8	0,27
27	5	0,17
28	13	0,43
29	10	0,33
30	9	0,30
31	8	0,27
32	14	0,47
33	7	0,23
34	10	0,33
35	8	0,27
36	4	0,13
37	13	0,43
38	14	0,47
39	14	0,47
40	8	0,27

Repetición nº	Cantidad de éxitos	$\hat{p}$
41	13	0,43
42	13	0,43
43	9	0,30
44	9	0,30
45	11	0,37
46	14	0,47
47	8	0,27
48	10	0,33
49	7	0,23
50	10	0,33
51	7	0,23
52	13	0,43
53	11	0,37
54	6	0,20
55	12	0,40
56	11	0,37
57	12	0,40
58	11	0,37
59	10	0,33
60	8	0,27

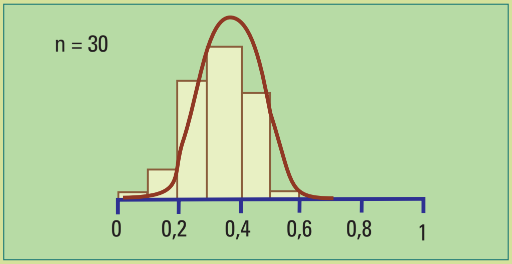
Al arrojar el dado 30 veces puede haber desde 0 hasta 30 éxitos. Hay también 31 valores diferentes en la columna encabezada por (0/30; 1/30; 2/30; ...;30/30)



El histograma anterior muestra la distribución de 60 proporciones muestrales de “éxito” en la repetición del experimento (arrojar un dado 3 veces, éxito= sale 1ó 2). El intervalo de clase que contiene al 0,33 es el más frecuente, pero la diferencia con las frecuencias del 0 y del 0,67 no es muy grande. Esta estimación está equivocada la mayoría de las veces.



El histograma anterior muestra la distribución de 60 proporciones muestrales de "éxito" en la repetición del experimento (arrojar un dado 30 veces, éxito= sale 1 ó 2). Las proporciones muestrales están más concentradas alrededor de la verdadera proporción (p =1/3) y su distribución puede aproximarse por la curva Normal:



En la práctica, la distribución de los valores de una variable en la población no cambia si se quita de ella una proporción muy pequeña. Los experimentos anteriores representan los resultados de muestreos aleatorios simples para estimar la proporción de éxitos (p) en una población (que solamente nosotros sabemos tiene p =1/3). Se trata de una simulación, repetimos muchas veces el cálculo de un estimador para conocer su distribución de muestreo, pero cuando lo usamos como estimador lo hacemos una única vez. Los histogramas anteriores muestran que es mejor usar muestras de tamaño 30 que de tamaño 3 para estimar una proporción. Cuanto más grande, mejor, siempre que el procedimiento de selección no tenga sesgo.

23.4.

a) Una calculadora da μ = 8,5 y σ = 5,346338. Se trata de parámetros poblacionales. N = 6 es el tamaño de la población. La fórmula de cálculo del desvío estándar poblacional difiere de s en que tiene $\sqrt{N}$ en vez de $\sqrt{(n-1)}$ en el denominador:

$$\sigma_X = \sqrt{\frac{1}{6}\sum_{i=1}^{6}(x_i - \mu_X)^2}$$

b) Enumere los valores de la variable de todas las 15 posibles muestras de tamaño n =2 y calcule $\bar{x}$ para cada una de ellas.

Intervalo de peso (kg)	Longitud del Intervalo	Frecuencia Relativa en %
Muestra 1	(2, 4)	$\bar{x}_1 = 3$
Muestra 2	(2,6)	$\bar{x}_2 = 4$
Muestra 3	(2, 9)	$\bar{x}_3 = 4{,}5$
Muestra 4	(2, 12)	$\bar{x}_4 = 7$
Muestra 5	(2, 18)	$\bar{x}_5 = 10$
Muestra 6	(4,6)	$\bar{x}_6 = 5$
Muestra 7	(4,9)	$\bar{x}_7 = 6{,}5$
Muestra 8	(4,12)	$\bar{x}_8 = 8$
Muestra 9	(4,18)	$\bar{x}_9 = 11$

⇒

Intervalo de peso (kg)	Longitud del Intervalo	Frecuencia Relativa en %
Muestra 10	(6,9)	$\bar{x}_{10} = 7{,}5$
Muestra 11	(6,12)	$\bar{x}_{11} = 9$
Muestra 12	(6,18)	$\bar{x}_{12} = 12$
Muestra 13	(9,12)	$\bar{x}_{13} = 10{,}5$
Muestra 14	(9,18)	$\bar{x}_{14} = 13{,}5$
Muestra 15	(12,18)	$\bar{x}_{15} = 15$

Observación. En este ejercicio se conoce completamente la población, y la distribución de muestreo de los valores de la media muestra.

c) Muestre que la media de las 15 medias muestrales ($\bar{x}$) es μ.

La media de las 15 medias muestrales ($\bar{x}$) es un parámetro poblacional ($\mu_{\bar{x}}$). La población (N = 15) está formada por los 15 valores posibles de $\bar{x}$.

$$\mu_{\bar{x}} = \frac{\sum_{i=1}^{n} \bar{x}_i}{N}$$

$$\mu_{\bar{x}} = \frac{\sum_{i=1}^{15} \bar{x}_i}{15}$$

$$= \frac{3 + 4 + 5{,}5 + 7 + 10 + 5 + 6{,}5 + 8 + 11 + 7{,}5 + 9 + 12 + 10{,}5 + 13{,}5 + 15}{15}$$

$$= \frac{127{,}5}{15}$$

$$= 8{,}5$$

$$\mu_{\bar{x}} = \mu$$

d) El desvío estándar de las 15 medias muestrales ($\bar{x}$) es un parámetro poblacional ($\sigma_{\bar{x}}$) de la misma población considerada en **c.** Usando una calculadora obtenemos el desvío estándar de los 15 valores:

$$\sigma_{\bar{x}} = \frac{1}{15} \sum_{i=1}^{15} (\bar{x}_i - \mu_{\bar{x}})^2$$

$$= 3{,}381321$$

El desvío estándar de las 15 medias muestrales es: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$

En efecto:

n = 2, N = 6, σ = 5,346338 $\quad \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{5{,}346338}{\sqrt{2}} \sqrt{\frac{6-2}{6-1}}$

$$= 3{,}381321$$

$$= \sigma_{\bar{x}}$$

e) En general, cuando el tamaño de la población N es muy grande en comparación con el tamaño muestral n, $\frac{N-n}{N-1} \approx 1$ y la expresión anterior se simplifica a $\frac{\sigma}{\sqrt{n}}$.

El error estándar no depende, en este caso, del tamaño de la población.

☐ Capítulo 24

24.1.

a) Como el tamaño de la muestra n = 40 es mayor a 30, podemos utilizar la aproximación por la Normal a la distribución de la media muestral. En este caso, el intervalo aproximadamente del 95% de confianza para la media muestral es de la forma $\bar{x} \pm 2 \frac{\sigma}{\sqrt{n}}$ por lo tanto resulta $3{,}09 \pm 2 \cdot \frac{0{,}25}{\sqrt{40}} \Rightarrow 3{,}09 \pm 0{,}08$

El intervalo es [3,01; 3,17] y tiene longitud 0,16.

b) $3{,}09 \pm 3 \cdot \frac{0{,}25}{\sqrt{40}} \Rightarrow 3{,}09 \pm 0{,}12$, es el resultando el intervalo [2,97; 3,21] que tiene longitud 0, 24.

c) La longitud del intervalo de confianza del 95% es $4 \frac{\sigma}{\sqrt{n}}$. Debemos hallar n tal que $4 \cdot \frac{0{,}25}{\sqrt{n}} = 0{,}1$. Por lo tanto $n = \left(4 \frac{0{,}25}{0{,}1}\right)^2$

$n = 100$

24.2.

c) es la correcta.

Las longitudes de los intervalos de confianza son para 95% y 99,7% respectivamente:

por el procedimiento 1 $4\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ y $6\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

por el procedimiento 2 $4\sqrt{\frac{1}{n}}$ y $6\sqrt{\frac{1}{n}}$

Para cualquiera de los dos procedimientos:

$$\frac{100 \text{ x (Longitud para 99,7\% - Longitud para 95\%)}}{\text{Longitud para 95\%}} = \frac{100 \text{ x } (6-4)}{4}$$
$$= \frac{100 \text{ x} 1}{2}$$
$$= 50$$

24.3.

c) es la correcta. El desvío estándar se estima por s = 1,4 y el intervalo de confianza es de la forma $\left[\bar{x}-2\left(\frac{s}{\sqrt{n}}\right) ; \bar{x}+2\left(\frac{s}{\sqrt{n}}\right)\right]$, quedando

$\left[6{,}8-2\ \frac{1{,}4}{\sqrt{49}} \quad ; 6{,}8+2\ \frac{1{,}4}{\sqrt{49}}\right]$. Esto es [6,8 -0,4; 6,8 +0,4]

24.4. No hay garantías que la verdadera proporción se encuentre dentro del intervalo. Ninguna de las respuestas es correcta.

24.5.

d) es correcta. Al aumentar el tamaño de la muestra de **n** a **kn**, la longitud del intervalo se divide por $\sqrt{\mathbf{k}}$**.**

24.6.

e) es correcta. Cuando se construye un intervalo de confianza se confía que el verdadero valor se encuentra dentro de él, pero no se puede estar seguro/a.

24.7.

b) es correcta. $n_1 = 400$ $n_2 = 300$ $\hat{p}_1 = 0{,}65$ $\hat{p}_2 = 0{,}48$

El desvío estándar estimado de la diferencia de proporciones es:

$$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{0{,}65(1-0{,}65)}{400}+\frac{0{,}48(1-0{,}48)}{300}}$$

$$= 0{,}0374$$

el IC para $\hat{p}$ - $\hat{p}$ es: (0,65-0,48) ± (2 x 0,0374) = 0,17 ± 0,0748

24.8. Primero verificamos que n $\hat{p}$ = 16 > 10 y n (1- $\hat{p}$) = 220 - 16 = 204 > 10. Sabemos que 220 es el 0,44% < 10 % de la población por lo que no es necesario utilizar una corrección por el tamaño de la población (el factor de corrección sería 0,998). La proporción muestral de piezas dañadas es $\hat{p}$ = 16 /220 = 0,072 y el desvío estándar estimado de $\hat{p}$ es:

$$\sigma_{\hat{p}} \text{ estimado } = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= \sqrt{\frac{0{,}072(1-0{,}072)}{220}}$$

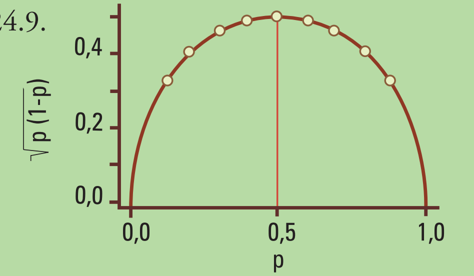
$$= 0{,}0174$$

El intervalo del 95% de confianza para la verdadera proporción de piezas dañadas es 0,072 ± 2 . 0,0174. Luego, tenemos un 95% de confianza que la proporción de piezas dañadas se encuentra entre 0,0372 y 1,1068. Como la cantidad de piezas del embarque es 50.000 este intervalo se traduce en el siguiente intervalo de piezas dañadas:

$$0{,}0372 \times 50.000 = 1.860 \quad y \quad 1{,}1068 \times 50.000 = 5.340$$

Tenemos un 95% de confianza que la cantidad de piezas dañadas se encuentra entre 0,0372 y 1,1068.

24.9.



El gráfico de $\boxed{\sqrt{p(1-p)}}$ en función de p alcanza su máximo para p = 0,5.

Capítulo 25

25.1. Cuando se rechaza la hipótesis nula con un valor -p = 0,03 eso significa que

a. Falsa. Es posible rechazar la hipótesis nula en forma equivocada
b. Falsa. El valor-p nada dice respecto a la proporción de veces que la hipótesis nula es falsa.
c. Falsa.
d. Verdadera. El valor -p se calcula suponiendo que la hipótesis nula es verdadera.

25.2.

a. Verdadera. Es necesario tener razones vinculadas a la naturaleza del problema a estudiar y no con los datos, para elegir una hipótesis alternativa unilateral.
b. Falsa. La elección de la hipótesis alternativa no debe basarse en los datos.
c. Falsa. La elección de la hipótesis alternativa no depende de la magnitud del efecto que se interesa probar.

25.3.

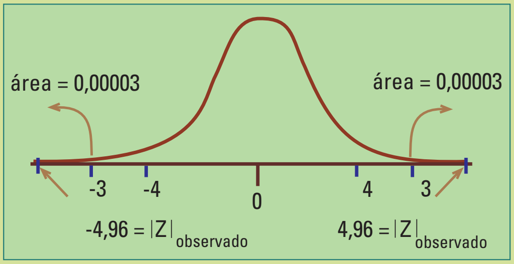
a. Falsa. La elección de la hipótesis alternativa no debe basarse en los datos.
b. Verdadera. Es necesario tener razones relacionadas con la naturaleza del problema a estudiar para elegir una hipótesis alternativa unilateral
c. Falsa. La elección de la hipótesis alternativa no está relacionada con el cálculo del valor - p.
d. Falsa. La elección de la hipótesis alternativa no depende de la cantidad de datos.
d. Falsa. La elección de la hipótesis alternativa no depende de la cantidad de datos.

25.4.

e. Verdadera. Aunque tenga la sospecha respecto a la dirección del resultado (las ambulancias tardan más que el tiempo especificado en H_0) debe plantear una alternativa bilateral, a menos que pueda probar que es imposible que el tiempo sea menor a 12 minutos.

25.5.

a. Verdadera. Tenemos. H_0: μ =12 y Ha: $\mu \neq$ 12.



$$n = 44, \frac{s}{\sqrt{n}} = \frac{8}{\sqrt{44}} = \frac{8}{6{,}63} = 1{,}21$$

$$z_{observado} = \frac{18-12}{1{,}21} = 4{,}96$$

El área hacia valores más alejados del cero desde -4,96 y 4,96 es menor al área 2 . 0,00003 =0,00006 de los valores más alejados desde -4 y 4.

Por lo tanto el valor -p (4,96; a dos colas) < 0,00006. Esto significa que si la hipótesis fuera verdadera, tendríamos un valor como el observado o más extremo menos de 1 de 1000 veces. Concluimos que la hipótesis nula no es verdadera y se rechaza la afirmación del intendente.

Un intervalo del 95% de confianza para el tiempo medio de tardanza de una ambulancia es más informativo:

$$\bar{x} \pm 1{,}96 \cdot \frac{s}{\sqrt{n}} = 18 \pm 1{,}96 \cdot 1{,}21$$
$$= 18 \pm 2{,}37$$
$$[\ 15{,}63\ ;\ 20{,}37\ ]$$

Con un 95% de confianza afirmamos que el tiempo medio que tarda una ambulancia en llegar al lugar del accidente se encuentra entre 15,63 y 20,37 minutos.

25.6.

a. y b. Falsas. Los datos no son parte de las hipótesis.

c. Verdadera. Hay que realizar un test para diferencia de proporciones bilateral por que no hay razones para suponer que la diferencia tendrá necesa ria mente una dirección.

d. Falsa .

25.7.

c. Verdadera. Aunque tenga la sospecha respecto a la dirección del resultado (las ambulancias tardan más que el tiempo especificado en H_0) debe plantear una alternativa bilateral, a menos que pueda probar que es imposible que el tiempo sea menor a 12 minutos.

Tenemos H_0: $p_1 - p_2 = 0$ y Ha: $p_1 - p_2 \neq 0$

$$\hat{p}_1 = \frac{68}{140} \qquad \hat{p}_2 = \frac{95}{180}$$
$$\hat{p}_1 = 0{,}49 \qquad \hat{p}_2 = 0{,}53$$

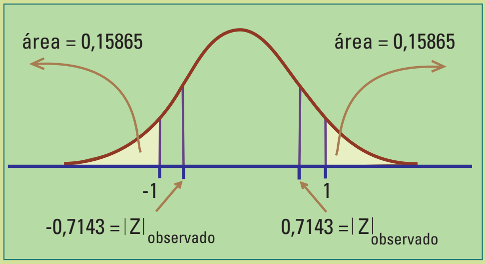
El estadístico del test es

$$\frac{\hat{p}_1 - \hat{p}_1 - 0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

y $$z_{observado} = \frac{\frac{68}{140} - \frac{95}{180}}{\sqrt{\frac{\frac{68}{140}\left(1 - \frac{68}{140}\right)}{140} + \frac{\frac{95}{180}\left(1 - \frac{95}{180}\right)}{180}}}$$

$$= \frac{0{,}49 - 0{,}53}{0{,}056}$$

$$= -0{,}7143$$



Como el test es a dos colas el p-valor es la suma de las áreas de las 2 colas, hacia valores alejados del cero bajo la curva N(0,1) a izquierda de -0,7143 y a derecha de 0,7143. Estas **áreas son mayores** que 2.0,15865 (las áreas que dejan hacia las colas el -1 y el 1, figura anterior).

27.8. ¿Cómo va a elegir las muestras? ¿Serán representativas de todos los profesores y alumnos del país? Si no lo es, restrinja el alcance de sus conclusiones a la población a que considere adecuada. Por ejemplo a todos los profesores y alumnos de su provincia, de su ciudad o de su escuela. Verifique que el tamaño de las muestras es suficientemente grande como para utilizar el TCL para la distribución de muestreo de las proporciones muestrales.

Llame “éxito”= una persona piensa que el cambio climático va afectar fuertemente sus vidas en los próximos 20 años.

Realice un test para diferencia de proporciones como en el del ejercicio 7 para comparar las opiniones entre:

- profesores y alumnos, p_1= proporción poblacional de éxitos entre los/as profesores/as, p_2= proporción poblacional de éxitos entre los alumnos/as
- mujeres y varones, p_1= proporción poblacional de éxitos entre los alumnos, p_2= proporción poblacional de éxitos entre las alumnas

Construya una estimación de la verdadera diferencia de proporciones utilizando un intervalo de confianza del 95%.

27.9.

c. Falsa. Los valores-p chicos (menores que 0,05) son evidencia en contra de la hipótesis nula.
b. Falsa. El tamaño de muestra no corrige el sesgo.
c. Falsa.
d. Verdadera.
e. Falsa.

Bibliografía recomendada

Estadística. David Friedman, Robert Pisani, Roger Purves y Ani Adhikari. **ISBN:** 848585568X 1993 Editor Antoni. Bosch Barcelona.

Interactive Statistics. Martha Aliaga, Brenda Gunderson. **ISBN:** 031497561 Prentice Hall 2006.

Introduction to the practice of Statistics. David S. Moore, George P McCabe. **ISBN:** 9780716764007. W.H. Freeman & Company 1989.

Statistics. Concepts and controversies. David S. Moore, William I. Notz **ISBN:** 9780716786368. W.H. Freeman & Company 2006