Estadística

Soluciones ejercicios: Descriptiva

Ver 8

Emilio Letón

1. Nivel 1

1. ¿Puede haber conjuntos de datos distintos entre sí con igual media, mediana, mínimo, máximo y amplitud?

   SOLUCIÓN:

   Sí, por ejemplo, los datos

   $$\{1,4,5,5,5,6,6,6,7,10\}, \{1,2,3,4,5,6,7,8,9,10\} \text{ y } \{1,1,1,1,1,10,10,10,10,10\}.$$

2. Explicar razonadamente por qué $\sum_{i=1}^{n}\left(x_i-\bar{x}\right)$ no es una buena medida de dispersión.

   SOLUCIÓN:

   Porque $\sum_{i=1}^{n}\left(x_i-\bar{x}\right)=0$ siempre, ya que

   $$\sum_{i=1}^{n}\left(x_i-\bar{x}\right)=\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\bar{x}=n\bar{x}-\bar{x}\sum_{i=1}^{n}1=n\bar{x}-\bar{x}n=0.$$

3. Deducir la fórmula que expresa la varianza muestral de un conjunto de datos como la media de los cuadrados del conjunto de datos menos el cuadrado de la media del conjunto de datos.

   SOLUCIÓN:

   Están pidiendo demostrar que $s^2=\overline{x^2}-\overline{x}^2$. Y esto es cierto porque

   $$s^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2=\frac{1}{n}\left(\sum_{i=1}^{n}\left(x_i^2+\overline{x}^2-2x_i\overline{x}\right)\right)=\frac{1}{n}\sum_{i=1}^{n}x_i^2+\frac{1}{n}n\overline{x}^2-\frac{2}{n}\overline{x}\sum_{i=1}^{n}x_i$$

   $$=\frac{1}{n}\sum_{i=1}^{n}x_i^2+\overline{x}^2-2\overline{x}^2=\overline{x^2}-\overline{x}^2$$

4. Demostrar que la desviación tipica muestral es cero si y solo sí el conjunto de datos es constante.

   SOLUCIÓN:

   En primer lugar se prueba que si la desviación típica es cero entonces la variable es constante. Esto es cierto porque $s=0$ implica que $\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2=0$, es decir que todos los sumandos son cero (al ser todas cantidades mayores o iguales que cero), por lo que $x_i=\bar{x}$ para todo $i=1,...,n$, con lo que el conjunto de datos es constante.

En segundo lugar se trata de probar que si el conjunto de datos es constante, es decir, $x_i = a$ para todo $i = 1, ..., n$, entonces $s = 0$. Esto es cierto porque $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} a = \frac{1}{n}na = a$, y por tanto,

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(a-a)^2 = 0$$

con lo que $s = 0$.

5. ¿Puede ocurrir que haya dos conjuntos de datos $x : \{x_1, ..., x_n\}$ e $y : \{y_1, ..., y_n\}$ distintos entre sí y que sin embargo tengan igual media y varianza?

SOLUCIÓN:

Sí. Al ser $\overline{x} = \overline{y}$, se tiene que cumplir que $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$. Por otra parte al ser $s_x^2 = s_y^2$, y verificarse que $s_x^2 = \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right]$ y que $s_y^2 = \frac{1}{n}\left[\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right]$, se tiene que verificar que $\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i^2$. Por tanto basta tomar dos conjuntos de datos x e y distintos entre sí y que verifiquen las condiciones anteriores. Se puede tomar de ejemplo el conjunto de datos $x : \{4, 5, 9\}$ e $y : \{3, 8, 7\}$, ya que $\sum_{i=1}^{n} x_i = 18 = \sum_{i=1}^{n} y_i$ y $\sum_{i=1}^{n} x_i^2 = 122 = \sum_{i=1}^{n} y_i^2$.

6. Decir si son verdaderas o falsas las siguientes afirmaciones. En caso de que sean verdaderas demostrarlo y en caso de que sean falsas dar un contraejemplo.

   a) En un conjunto de datos $x : \{x_1, x_2, ..., x_n\}$ donde hay k valores distintos $x_1, x_2, ..., x_k$ y donde cada valor distinto x_j $(j = 1, .., k)$ se repite con una frecuencia relativa de $fr(x_j)$, se tiene que la varianza $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ verifica que

$$s^2 = \left[\sum_{j=1}^{k} x_j^2 \cdot fr(x_j)\right] - \left[\sum_{j=1}^{k} x_j \cdot fr(x_j)\right]^2$$

   b) Para cualquier conjunto de datos $x : \{x_1, x_2, ..., x_n\}$ se tiene que la varianza $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ siempre verifica que $s^2 > 0$.

SOLUCIÓN:

   a) Es verdadera.

   En primer lugar, se tiene que $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{j=1}^{k} x_j fr(x_j)$ y que $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \sum_{j=1}^{k} x_j^2 fr(x_j)$ y dado que

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = \overline{x^2} - \overline{x}^2$$

   se tiene que $s^2 = \overline{x^2} - \overline{x}^2 = \left[\sum_{j=1}^{k} x_j^2 \cdot fr(x_j)\right] - \left[\sum_{j=1}^{k} x_j \cdot fr(x_j)\right]^2$.

b) Es falsa.

Si el conjunto de datos es constante $x : \{a, a, ..., a\}$, se tiene que $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} a = \frac{1}{n}na = a$, con lo que $s^2 = \frac{1}{n}\sum_{i=1}^{n}(a-a)^2 = 0$. Lo que sí se verifica siempre es que $s^2 \geq 0$ para cualquier conjunto de datos.

7. La media de las edades (en años) de un grupo de personas vale m_0. Dentro de 10 años, la media valdrá m_1. ¿Cuál o cuáles de las siguientes respuesta es correcta?

a) $m_0 = m_1 + 10$.

b) $m_0 + 10 = m_1$.

c) $m_0 > m_1$.

d) $m_0 < m_1$.

e) $m_0 = m_1$.

f) Depende de las edades.

SOLUCIÓN:

Se deja para el alumno.

8. La media de las edades (en años) de un grupo de personas vale m_a y en meses m_m. ¿Cuál o cuáles de las siguientes respuesta es correcta?

a) $12m_a = m_m$.

b) $m_a = 12m_m$.

c) $m_a > m_m$.

d) $m_a < m_m$.

e) $m_a = m_m$.

f) Ninguna de las anteriores.

SOLUCIÓN:

Se deja para el alumno.

9. La desviación típica de las edades (en años) de un grupo de personas vale d_0. Dentro de 10 años, la media valdrá d_1. ¿Cuál o cuáles de las siguientes respuesta es correcta?

a) $10d_0 = d_1$.

b) $d_0 > d_1$.

c) $d_0 < d_1$.

d) $d_0 = d_1$.

e) Depende de las edades.

f) Ninguna de las anteriores.

SOLUCIÓN:

Se deja para el alumno.

10. La desviación típica de las edades (en años) de un grupo de personas vale d_a y en meses d_m. ¿Cuál o cuáles de las siguientes respuesta es correcta?

    a) $12d_a = d_m$.

    b) $d_a = 12d_m$.

    c) $d_a = d_m$.

    d) $d_a > d_m$.

    e) $d_a < d_m$

    f) Depende de las edades.

    SOLUCIÓN:

    Se deja para el alumno.

2. Nivel 2

1. Demostrar que si construimos un conjunto de datos z mezclando $n_1$valores de x y n_2 valores de y, la media de z es $\overline{z} = \frac{n_1}{n_1+n_2}\overline{x} + \frac{n_2}{n_1+n_2}\overline{y}$.

   SOLUCIÓN:

   El conjunto de datos z que se considera es $z : \{x_1, x_2, ..., x_{n_1}, y_1, y_2, ..., y_{n_2}\}$. Por tanto, su media es

   $$\overline{z} = \frac{x_1 + x_2 + ... + x_{n_1} + y_1 + y_2 + ... + y_{n_2}}{n_1 + n_2} = \frac{n_1\overline{x} + n_2\overline{y}}{n_1 + n_2} = \frac{n_1}{n_1 + n_2}\overline{x} + \frac{n_2}{n_1 + n_2}\overline{y}.$$

   Es decir, si nos dan dos medias, para poder combinarlas no es correcto hacer la media de las medias, hay que hacer la media ponderada.

2. Demostrar que si se tiene una proporción de sensores defectuosos de p_1 en una muestra de $n_1$sensores y una proporción de sensores defectuosos de p_2 en una muestra de n_2 sensores, la proporción global de sensores defectuosos p es $p = \frac{n_1}{n_1+n_2}p_1 + \frac{n_2}{n_1+n_2}p_2$.

   SOLUCIÓN:

   Si en la muestra primera hay una proporción de p_1 sensores defectuosos en la primera muestra, es porque el número de sensores defectuosos es r_1 verificando que

   $$p_1 = \frac{r_1}{n_1}.$$

   Análogamente para la muestra 2, el número de sensores defectuosos es r_2 verificando que

   $$p_2 = \frac{r_2}{n_2}.$$

   Por tanto el número de sensores defectuosos global es $r_1 + r_2$ de una muestra de $n_1 + n_2$ y la proporción global de sensores defectuosos p es

   $$p = \frac{r_1 + r_2}{n_1 + n_2} = \frac{p_1n_1 + p_2n_2}{n_1 + n_2} = \frac{n_1}{n_1 + n_2}p_1 + \frac{n_2}{n_1 + n_2}p_2.$$

   Es decir, si nos dan dos proporciones, para poder combinarlas no es correcto hacer la media de de las proporciones, hay que hacer una media ponderada de las proporciones.

3. Existen dos proveedores M y H que fabrican sensores. En una muestra de 300 sensores, se ha obtenido que el 10 % fallan y que el 40 % eran fabricados por el proveedor M. Además, en dicha muestra se ha observado que el número de sensores que fallan es igual para los dos proveedores. Se pide:

   a) Construir la tabla de frecuencias absolutas para las variables "Estado" y "Proveedor".

   b) ¿Distribución de frecuencias conjunta relativa? ¿Suman uno?

   c) ¿Distribución marginal relativa de la variable "Estado"? ¿Suman uno?

   d) ¿Distribución de frecuencias absolutas de "Estado" condicionada a que los sensores son del proveedor M? ¿Suman uno?

   e) ¿Distribución de frecuencias relativas de "Estado" condicionada a que los sensores son del proveedor M? ¿Y del H? ¿Suman uno?

   f) ¿Qué proveedor falla más: M o H?

SOLUCIÓN:

a)

Estado / Proveedor	M	H	
Falla	15	15	30
No Falla	105	165	270
	120	180	300

b)

Estado / Proveedor	M	H	
Falla	0.05	0.05	
No Falla	0.35	0.55	
			1

c)

Estado		
Falla	30	0.10
No Falla	270	0.90
	300	1

d)

Estado	M
Falla	15
No Falla	105
	120

e)

Estado	M
Falla	0.125
No Falla	0.875
	1

Estado	H
Falla	0.083
No Falla	0.926
	1

f) Se tiene que

$$fr\,(F|M) = \frac{fa\,(FyM)}{fa\,(M)} = \frac{fr\,(FyM)}{fr\,(M)} = 0{,}125 \neq fr\,(F) = 0{,}10$$

$$fr\,(F|H) = \frac{fa\,(FyH)}{fa\,(H)} = \frac{fr\,(FyH)}{fr\,(H)} = 0{,}083 \neq fr\,(F) = 0{,}10$$

Falla más el proveedor F.

3. Nivel 3

1. Decir si son verdaderas o falsas las siguientes afirmaciones. En caso de que sean verdaderas demostrarlo y en caso de que sean falsas dar un contraejemplo.

   a) Si en un conjunto de datos la media es igual a la mediana, entonces se concluye que necesariamente el conjunto de datos es simétrico.

   b) A partir de un conjunto de datos $\{x_1, ..., x_n\}$ se define otro conjunto de datos $\{y_1, ..., y_n\}$ de forma que $y_i = a + bx_i$ con a y b números reales cualesquiera (no necesariamente positivos). Si se calcula el coeficiente de correlación de Pearson r_{xy} entre los dos conjuntos de datos anteriores, entonces se tiene que $r_{xy} = +1$.

   c) Se tienen dos conjuntos de datos $x = \{x_1, ..., x_n\}$ e $y = \{y_1, ..., y_n\}$. A partir de ellos se definen dos nuevos conjuntos de datos x' e y' dados por

   $$x' = a + bx$$

   $$y' = c + dy$$

   con a, b, c y d números reales cualesquiera (no necesariamente positivos). Entonces se verifica que $r_{x'y'} = r_{xy}$.

SOLUCIÓN:

a) Es falsa.

Por ejemplo el conjunto de datos $\{1, 1, 1, 1, 1, 5, 7, 8, 9, 10, 11\}$ verifica que la media al igual que la mediana es 5 y sin embargo los datos no son simétricos. Lo que sí se verifica es que si el conjunto de datos es simétrico entonces la media es igual a la mediana.

b) Es falsa.

Para los conjuntos de datos dados se tiene que

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\left(a + bx_i - \overline{a + bx_i}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\left(a + bx_i - \left(\frac{1}{n}\sum_{i=1}^{n}(a + bx_i)\right)\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\,(a + bx_i - a - b\overline{x}) = \frac{1}{n}b\sum_{i=1}^{n}(x_i - \overline{x})\,(x_i - \overline{x}) = \frac{b}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = bs_x^2$$

Por otra parte

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(a + bx_i - \overline{a + bx_i}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i - a - b\overline{x})^2 = b^2 s_x^2$$

con lo que $s_y = |b|\, s_x$.(se toma valor absoluto para asegurar que la desviación típica sea positiva: $s_y > 0$)

Por tanto, el coeficiente de correlación de Pearson r_{xy} es

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{b s_x^2}{s_x\, |b|\, s_x} = \frac{b}{|b|} = signo\,(b) \cdot 1$$

Con lo que si $b < 0$, se tiene que $r_{xy} = -1$

c) Es falsa.

El coeficiente de correlación de Pearson r_{xy} viene dado por

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

El coeficiente de correlación de Pearson $r_{x'y'}$ viene dado por

$$r_{x'y'} = \frac{s_{x'y'}}{s_{x'} s_{y'}} = \frac{\frac{1}{n}\sum_{i=1}^{n} \left(x'_i - \overline{x'}\right)\left(y'_i - \overline{y'}\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(x'_i - \overline{x'}\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(y'_i - \overline{y'}\right)^2}}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n} \left(a + bx_i - \overline{a + bx_i}\right)\left(c + dx_i - \overline{c + dy_i}\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(a + bx_i - \overline{a + bx_i}\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(c + dx_i - \overline{c + dy_i}\right)^2}}.$$

Dado que

$$\overline{a + bx_i} = \frac{1}{n}\sum_{i=1}^{n} (a + bx_i) = a + b\overline{x}$$

$$\overline{c + dy_i} = c + d\overline{y}$$

$$s_{x'y'} = \frac{1}{n}\sum_{i=1}^{n} \left(a + bx_i - \overline{a + bx_i}\right)\left(c + dx_i - \overline{c + dy_i}\right) = bd s_{xy}$$

se tiene que

$$r_{x'y'} = \frac{bd s_{xy}}{\sqrt{b^2} s_x \sqrt{d^2} s_y} = \frac{bd}{|b|\,|d|} r_{xy}$$

Si $signo(b) = signo(d)$, se tiene que $r_{x'y'} = r_{xy}$ y si $signo(b) \neq signo(d)$, $r_{x'y'} = -r_{xy}$. En el caso de que b y d sean positivos, se tiene que $r_{x'y'} = r_{xy}$, con lo que el coeficiente de correlación de Pearson no se modifica, es decir, no cambia la medida de asociación entre dos conjuntos de datos si se cambian las unidades, lo cual es lógico (sería catastrófico si cambiara: dos analistas de datos obtendrían conclusiones distintas trabajando con los mismos datos aunque en distintas unidades).

2. A partir de un conjunto de datos $\{x_1, ..., x_n\}$ se define otro conjunto de datos $\{y_1, ..., y_n\}$ de forma que $y_i = a + bx_i$. Expresar la relación que hay entre:

a) $\overline{x}$ y $\overline{y}$.

b) s_x^2 y s_y^2.

c) s_x y s_y.
(distinguir, si hiciera falta, los casos $b > 0, b = 0, b < 0$).

SOLUCIÓN:

a) Se verifica que $\overline{y} = a + b\overline{x}$ (siempre, no hace falta distinguir $b > 0, b = 0, b < 0$) ya que

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} (a + bx_i) = \frac{1}{n}(na + bn\overline{x}) = a + b\overline{x}.$$

b) Se verifica que $s_y^2 = b^2 s_x^2$ (siempre, no hace falta distinguir $b > 0, b = 0, b < 0$) ya que

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n} (a + bx_i - a - b\overline{x})^2 = \frac{1}{n}b^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 = b^2 s_x^2.$$

c) Se verifica que $s_y = |b|\, s_x$ ya que al extraer la raíz cuadrada en la expresión del apartado anterior, hay que tomar valor absoluto. Con la expresión $s_y = |b|\, s_x$ no hace falta distinguir los casos $b > 0, b = 0, b < 0$ y además se consigue que la desviación típica s_y sea siempre mayor o igual que cero.

3. A partir de un conjunto de datos $\{x_1, ..., x_n\}$ se construye otro $\{y_1, ..., y_n\}$ con $y_i = \frac{x_i - \overline{x}}{s_x}$ ("estandarización del conjunto de datos x"). Demostrar que el conjunto de datos $\{y_1, ..., y_n\}$ tiene media cero, varianza uno y desviación típica uno (utilizar un ejercicio anterior).

SOLUCIÓN:

En primer lugar reescribimos la forma en la que está dada el conjunto de datos y_i de forma que

$$y_i = \frac{-\overline{x}}{s_x} + \frac{1}{s_x} x_i.$$

Por tanto se tiene que

$$\overline{y} = \frac{-\overline{x}}{s_x} + \frac{1}{s_x}\overline{x} = 0; s_y^2 = \left(\frac{1}{s_x}\right)^2 s_x^2 = 1; s_y = \left|\frac{1}{s_x}\right| s_x = \frac{1}{s_x} s_x = 1$$

4. Decir si es verdadera o falsa la siguiente afirmación. En caso de que sea verdadera demostrarlo, y en caso de que sea falsa dar un contraejemplo:

"En un conjunto de datos dicotómicos (binarios) codificados con "0" y "1", la proporción de datos que toman el valor "1" es la media del conjunto dado."

SOLUCIÓN:

Es verdadera.

Si en el conjunto de datos de n observaciones hay n_0 valores con el código "0" y n_1 valores con el código "1", la proporción de datos que toman el valor "1" es $\frac{n_1}{n}$ y la media es

$$\frac{0 + ... + 0 + 1 + ..,1}{n} = 0\frac{n_0}{n} + 1\frac{n_1}{n} = \frac{n_1}{n} = \textit{proporción de unos.}$$

5. Decir si es verdadera o falsa la siguiente afirmación. En caso de que sea verdadera demostrarlo, y en caso de que sea falsa dar un contraejemplo:

   "En un histograma nunca puede haber huecos entre clases."

   SOLUCIÓN:

   Es falsa.

   Si hay valores atípicos puede haber huecos entre clases porque al discretizar la variable es posible que no haya valores en todos las clases que se determinen.

6. Decir si es verdadera o falsa la siguiente afirmación. En caso de que sea verdadera demostrarlo, y en caso de que sea falsa dar un contraejemplo:

   "A partir de un conjunto de datos $x : \{x_1, x_2, ..., x_n\}$ se construye un nuevo conjunto de datos z duplicando el conjunto de datos anterior, es decir $z : \{x_1, x_2, ..., x_n, x_1, x_2, ..., x_n\}$, teniendo z un tamaño muestral de $2n$. Entonces se verifica que $\overline{x} = \overline{z}, x_{med} = z_{med}$ aunque $s_x^2 \neq s_z^2$."

   SOLUCIÓN:

   Es falsa.

   Se verifica que $\overline{x} = \overline{z}, x_{med} = z_{med}$ y $s_x^2 = s_z^2$, ya que

$$\overline{z} = \frac{x_1 + x_2 + ... + x_n + x_1 + x_2 + ... + x_n}{2n} = \frac{2\left(x_1 + x_2 + ... + x_n\right)}{2n} = \overline{x}$$

$$s_z^2 = \overline{z^2} - \overline{z}^2 = \frac{x_1^2 + x_2^2 + ... + x_n^2 + x_1^2 + x_2^2 + ... + x_n^2}{2n} - \overline{z}^2$$

$$= \frac{2\left(x_1^2 + x_2^2 + ... + x_n^2 + x_1^2\right)}{2n} - \overline{x}^2 = \overline{x^2} - \overline{x}^2 = s_x^2$$

   El hecho de que $x_{med} = z_{med}$ se debe a que tanto sea n par como impar al duplicar los datos, el valor de la mediana sigue siendo el mismo, ya que no se altera el orden. Si n es impar, $x : \{x_1, ..., x_{med}, ..., x_n\}$ se tiene que el conjunto de datos

$$z : \{x_1, ..., x_{med}, ..., x_n, x_1, ...x_{med}, ..., x_n\} = \{x_1, x_1, ..., x_{med}, x_{med}, ...x_n, x_n\}$$

   tiene un número par de elementos, con lo que

$$z_{med} = \frac{x_{med} + x_{med}}{2} = x_{med}$$

   Si n es par, $x : \{x_1, ..., x_p, x_{p+1}..., x_n\}$ se tiene que $x_{med} = \frac{x_p + x_{p+1}}{2}$ y el conjunto de datos

$$z : \{x_1, ..., x_p, x_{p+1}, ..., x_n, x_1, ..., x_p, x_{p+1}, ..., x_n\} = \{x_1, x_1, ..., x_p, x_p, x_{p+1}, x_{p+1}...x_n, x_n\}$$

   tiene un número par de elementos, con lo que

$$z_{med} = \frac{x_p + x_{p+1}}{2} = x_{med}$$

   El falso el razonamiento de que $x : \{2, 5, 7, 8\}$ tiene $x_{med} = \frac{5+7}{2} = 8$ y que $z : \{2, 5, 7, 8, 2, 5, 7, 8\}$ tiene $z_{med} = \frac{8+2}{2} = 5$, ya que para calcular la mediana antes hay que ordenar los datos.

7. Decir si es verdadera o falsa la siguiente afirmación. En caso de que sea verdadera demostrarlo y en caso de que sea falsa dar un contraejemplo:

   "En un histograma, la clase que tiene más datos es donde se encuentra el valor que más se repite".

SOLUCIÓN:

Es falsa. La clase es un rango de valores, puede, incluso que en la clase donde más datos haya ningun valor se repita. Se puede usar como contraejemplo el dado por

$$x : \{10, 10, 10, 15, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30\}.$$

8. Decir si es verdadera o falsa la siguiente afirmación. En caso de que sea verdadera demostrarlo y en caso de que sea falsa dar un contraejemplo:

   "En distribuciones simétricas, la mediana es la diferencia entre el tercer cuartil y el primer cuartil".

   SOLUCIÓN:

   Es falsa.

   La diferencia entre el tercer cuartil y el primero es el rango intercuartílico y no la mediana. Se puede usar como contraejemplo el dado por

   $$x : \{1, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 6\},$$

   donde se observa que la distribución es simétrica, $Q_2 = \frac{3+4}{2} = 3{,}5$, $Q_1 = \frac{2+3}{2} = 2{,}5$ y $Q_3 = \frac{4+5}{2} = 4{,}5$.

   Se observa que $Q_3 - Q_1 = 1 \neq 3{,}5 = Q_2$.Lo que sí se verifica en distribuciones simétricas es que $Q_2 = \frac{Q_1+Q_3}{2}$.

9. Encontrar el valor a que minimiza $\sum_{i=1}^{n} (x_i - a)^2$ para un conjunto de n observaciones $x_1, ...x_n$.

   Aplicar el resultado deducido para demostrar que $\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \leq \frac{1}{n}\sum_{i=1}^{n}(x_i - x_{med})^2$.

   SOLUCIÓN:

   Se define la función $f(a) = (x_1 - a)^2 + ... + (x_n - a)^2$, con lo que para que

   $$\frac{\partial}{\partial a} f(a) = 2(x_1 - a)(-1) + ... + 2(x_n - a)(-1) = 0$$

   se tiene que $(x_1 - a) + ... + (x_n - a) = 0$, es decir que $x_1 + ... + x_n - na = 0$, o lo que es lo mismo $a = \frac{1}{n}\sum_{i=1}^{n} x_i$.

   Por otra parte

   $$\frac{\partial^2}{\partial^2 a} f(a) = (-2)(-1) + ... + (-2)(-1) = 2n,$$

   con lo que siempre la segunda derivada es positiva, en particular para $a = \frac{1}{n}\sum_{i=1}^{n} x_i$, con lo que en dicho valor se alcanza un mínimo. Es decir la expresión que minimiza $\sum_{i=1}^{n}(x_i - a)^2$ es la media, por lo que para cualquier otro valor, por ejemplo, la mediana, dicha expresión será mayor, por tanto $\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \leq \frac{1}{n}\sum_{i=1}^{n}(x_i - x_{med})^2$.

4. Nivel 4

1. A partir de un conjunto de datos $x : \{x_1, ..., x_n\}$ se define otro conjunto de datos $y : \{y_1, ..., y_n\}$ de forma que $y_i = a + bx_i$. Expresar la relación que hay entre.

   a) x_{med} y y_{med}.

   b) $Q_1(x)$ y $Q_1(y)$.

   c) $Q_3(x)$ y $Q_3(y)$.

   d) $moda(x)$ y $moda(y)$.

   e) $amplitud(x)$ y $amplitud(y)$.

   f) $RI(x)$ y $RI(y)$.

   SOLUCIÓN:

   a) xxx.

   b) xxx.

   c) xxx.

   d) xxx.

   e) xxx.

   f) xxx

2. Decir si son verdaderas o falsas las siguientes afirmaciones:

   a) En un gráfico de caja ("box-plot"), los bigotes ("whiskers") corresponden a 1.5 veces el rango intercuartílico.

   SOLUCIÓN:

   a) Falso. El valor de 1.5 y 3 veces el rango intercuartílico se usan para definir las barreras internas y externas que determinan los valores atípicos. En el caso de que no haya datos atípicos, los bigotes corresponden al valor mínimo y máximo de los datos y en el caso de que haya datos atípicos al valor mínimo y máximo de los datos sin tener en cuenta lso datos atípicos.

3. Para el conjunto de datos $\{1, 10, 110, 120, 130, 140, 150, 160, 170, 180, 190, 379\}$ se pide:

   a) Calcular la media y la mediana.

   b) Dibujar el gráfico de caja ("box-plot")

   SOLUCIÓN:

   a) Para los 12 datos anteriores la media y la mediana son 145, ya que

   $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{12}(1 + 10 + ... + 379) = 145$$

   y la mediana, al ya estar ordenado los datos de menor a mayor y ser n par, es

   $$x_{med} = \frac{140 + 150}{2} = 145$$

b) Para dibujar el gráfico de caja, se necesita calcular Q_1,Q_3, el rango intercuartílico y los límites de las barreras internas y externas. Para calcular Q_1 se toma la parte del conjunto de datos ordenados de menor a mayor que son menores que el valor de la mediana, es decir $\{1, 10, 110, 120, 130, 140\}$, y sobre este conjunto de datos se calcula la mediana, con lo que al haber un número par de observaciones, $Q_1 = \frac{110+120}{2} = 115$. Para calcular Q_3 se toma la parte del conjunto de datos ordenados de menor a mayor que son mayores que el valor de la mediana, es decir $\{150, 160, 170, 180, 190, 379\}$, y sobre este conjunto de datos se calcula la mediana, con lo que al haber un número par de observaciones, $Q_3 = \frac{170+180}{2} = 175$. El rango intercuartílico es $RI = Q_3 - Q_1 = 175 - 115 = 60$, la barrera interna que viene dada por 1.5 veces el rango intercuartílico a partir de la caja, tiene por extremos $Q_1 - 1{,}5RI = 115 - 90 = 25$ y $Q_3 + 1{,}5RI = 175 - 90 = 265$ y la barrera externa que viene dada por 3 veces el rango intercuartílico a partir de la caja tiene por extremos $Q_1 - 3RI = 115 - 180 = -65$ y $Q_3 + 3RI = 175 + 180 = 355$. Con esta información, el gráfico de caja resultante se muestra en la Figura 1.

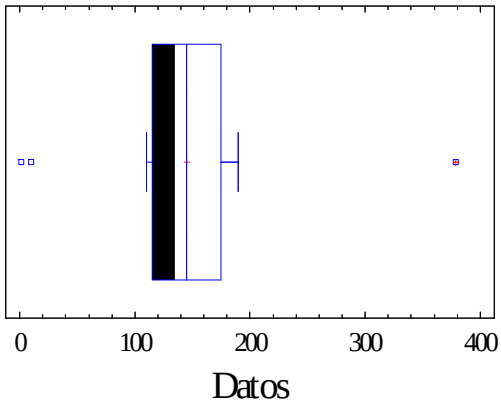

Figura1: box-plot

4. Se supone que se tienen 240 datos donde $\{1, 10, 110, 120, 130, 140, 150, 160, 170, 180, 190, 379\}$ son los 12 primeros y el resto de los datos consiste en la replicación del bloque anterior 19 veces más. Se pide para el conjunto de los 240 datos:

   a) Calcular la media y la mediana.

   b) Dibujar el gráfico de caja ("box-plot")

   c) Usando este ejemplo, contestar razonadamente si es verdadero o falso la afirmación: "Si la media y la mediana difieren poco, es porque los datos atípicos son escasos".

   d) ¿Es este conjunto de datos simétrico?

SOLUCIÓN:

a) La media de los 240 datos, utilizando la expresión para $k = 12$ datos agrupados es

$$\bar{x} = \sum_{j=1}^{12} x_j fr\left(x_j\right) = 1\frac{1}{20} + 10\frac{1}{20} + 110\frac{1}{20} + ... + 379\frac{1}{20} = 145.$$

La mediana del conjunto de 240 datos es la misma que la del conjunto de 12 datos que sirvió para su replicación, con lo que es

$$x_{med} = \frac{140 + 150}{2} = 145.$$

b) Para dibujar el gráfico de caja, se necesita calcular Q_1, Q_3, el rango intercuartílico y los límites de las barreras internas y externas. Estos valores, al igual que sucede con la mediana, coinciden con los valores respectivos para el conjunto de 12 datos que sirvió para su replicación, con lo que el gráfico de caja resultante es el de la Figura 1, aunque los valores atípicos se superponen, representando, por ejemplo, el valor atípico 379 no un único punto sino 20.

c) Falso. En este ejemplo no sólo difieren poco la media y la mediana, sino que son iguales, y se obtiene que hay 20 puntos atípicos que sobrepasan la barrera externa y 40 que sobrepasan la barrera interna.

d) No es simétrico. Este es un ejemplo de conjunto de datos en el que la media es igual a la mediana y el conjunto de datos no es simétrico, según se aprecia en el gráfico de caja de la figura 1 y en el histograma de la figura 2.

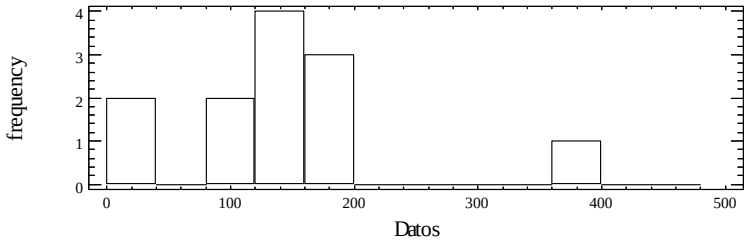

Figura 2: Histograma

5. Para los conjuntos de datos siguientes $\{1, 4, 5, 5, 5, 6, 6, 6, 7, 10\}$, $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ y $\{1, 1, 1, 1, 1, 10, 10, 10, 10, 10\}$ dibujar su box-plot.

SOLUCIÓN:

a) xxx

6. Decir si son verdaderas o falsas las siguientes afirmaciones:

a) En un diagrama de caja, la media siempre está comprendida en la caja del diagrama, es decir, entre el primer y tercer cuartil.

b) Los límites de los segmentos del diagrama de caja (bigotes), corresponden siempre al mínimo y al máximo de los datos.

SOLUCIÓN:

a) Falso. Es la mediana que la siempre estará dentro de la caja. La media puede, incluso, estar fuera de los bigotes. Se puede usar como contraejemplo el dado por

$$x : \{10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 30, 35, 40, 40, 50, 50, 50, 90, 1000, 1000\}.$$

b) Falso. En el caso de que haya valores atípicos, los bigotes llegan hasta el valor mínimo o máximo sin tener en cuenta los valores atípicos. En el caso de que no haya valores atípicos, sí sería cierta la afirmación. Se puede usar como contraejemplo el dado por

$$x : \{10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 30, 35, 40, 40, 50, 50, 90, 300, 300\}$$