Soluciones - Estadística descriptiva

Curso: Universidad 1

Solución del ejercicio 1

Planteamiento. Organizamos los datos en una tabla: la frecuencia absoluta cuenta repeticiones, la relativa la divide entre el total y las acumuladas van sumando. Las preguntas de «como mucho» se leen en las acumuladas y las de «al menos», con el complementario.

a. Es una variable cuantitativa discreta: toma valores enteros aislados.
b. Las relativas se obtienen dividiendo entre 20 y las acumuladas sumando hacia abajo.

x_i	f_i	h_i	F_i	H_i
0	4	0,20	4	0,20
1	8	0,40	12	0,60
2	5	0,25	17	0,85
3	2	0,10	19	0,95
4	1	0,05	20	1,00

c. $H_2 = \frac{12}{20} = 0{,}60$: el 60 % tiene como mucho un hermano.
d. Es el complementario: $1 - 0{,}60 = 0{,}40$, es decir, el 40 % (8 estudiantes).

Solución del ejercicio 2

Planteamiento. La media es el cociente entre la suma de productos valor por frecuencia y el número de datos. La mediana se localiza con las frecuencias acumuladas, buscando la posición central, y la moda es el valor más repetido.

a. $\bar{x} = \frac{0\cdot 4 + 1\cdot 8 + 2\cdot 5 + 3\cdot 2 + 4\cdot 1}{20} = \frac{28}{20} = 1{,}4$ hermanos.
b. Con 20 datos, la mediana es la media de los datos 10.º y 11.º, que valen 1 (la frecuencia acumulada pasa de 4 a 12 en el valor 1). Mediana $= 1$.
c. La moda es 1, el valor con mayor frecuencia (8).
d. La mediana y la moda (1 hermano) describen mejor el caso típico; la media, 1,4, no es un valor posible y la suben los pocos estudiantes con muchos hermanos.

Solución del ejercicio 3

Planteamiento. Usamos la fórmula abreviada de la varianza, que evita restar la media a cada dato: media de los cuadrados menos el cuadrado de la media. La desviación típica es su raíz cuadrada y el coeficiente de variación la compara con la media.

a. $0 \cdot 4 + 1 \cdot 8 + 4 \cdot 5 + 9 \cdot 2 + 16 \cdot 1 = 62$.

b. $\sigma^2 = \frac{62}{20} - 1{,}4^2 = 3{,}1 - 1{,}96 = 1{,}14$.
c. $\sigma = \sqrt{1{,}14} \approx 1{,}07$ hermanos y $CV = \frac{1{,}07}{1{,}4} \approx 0{,}76$, es decir, un 76 % de dispersión relativa.
d. El intervalo es $(0{,}33;\ 2{,}47)$, que contiene los valores 1 y 2: $8 + 5 = 13$ datos, el 65 %.

Solución del ejercicio 4

Planteamiento. Con datos agrupados cada intervalo se representa por su marca de clase. La media y la varianza se calculan como en datos discretos, y la mediana se estima por interpolación lineal dentro del intervalo donde la acumulada alcanza la mitad.

a. Marcas 155, 165, 175, 185. $\bar{x} = \frac{155 \cdot 6 + 165 \cdot 14 + 175 \cdot 12 + 185 \cdot 8}{40} = \frac{6820}{40} = 170{,}50$ cm.
b. $\sigma^2 = \frac{1}{40} \sum f_i (x_i - \bar{x})^2 = 94{,}75$, así que $\sigma \approx 9{,}73$ cm.
c. La mitad de los datos es 20; antes de $[160, 170)$ hay 6, así que la mediana está en ese intervalo: $Me = 160 + \frac{20 - 6}{14} \cdot 10 = 170{,}00$ cm.
d. $[160, 170)$, con la mayor frecuencia (14).

Solución del ejercicio 5

Planteamiento. Ordenamos los datos y calculamos la mediana y los cuartiles como medianas de cada mitad. El criterio de Tukey marca como atípicos los datos alejados más de 1,5 veces el rango intercuartílico de la caja.

a. La mediana es la media de los datos 7.º y 8.º: $\frac{21 + 22}{2} = 21{,}5$. Mitad inferior (7 datos): $Q_1 = 18$; mitad superior: $Q_3 = 26$.
b. $IQR = 26 - 18 = 8$ minutos.
c. Límites: $18 - 1{,}5 \cdot 8 = 6{,}0$ y $26 + 1{,}5 \cdot 8 = 38{,}0$. El valor 55 supera el límite superior: es atípico.
d. Caja de 18 a 26 con la mediana en 21,5, bigote inferior hasta 12, bigote superior hasta 30 y el 55 marcado aparte. La cola derecha es más larga: asimetría positiva.

Solución del ejercicio 6

Planteamiento. La desviación típica tiene las unidades de los datos y depende de su tamaño. Para comparar grupos con medias muy diferentes usamos el coeficiente de variación, que mide la dispersión en relación con la media y no tiene unidades.

a. Los adultos: 11 kg frente a 0,45 kg.
b. $CV_1 = \frac{0{,}45}{3{,}3} \approx 0{,}136$ y $CV_2 = \frac{11}{72} \approx 0{,}153$.
c. Los adultos, por poco: $0{,}153 > 0{,}136$. En términos relativos la dispersión es parecida.
d. Porque las medias son muy distintas: 0,45 kg es mucho para un bebé y poco para

un adulto. El CV no tiene unidades y mide la dispersión en relación con la media.

Solución del ejercicio 7

Planteamiento. Calculamos para cada grupo la media y la desviación típica con la fórmula abreviada. Si las medias coinciden, la desviación típica decide la homogeneidad, y la mediana y los recuentos completan la comparación.

a. $\bar{x}_A = \frac{63}{10} = 6{,}3$ y $\bar{x}_B = \frac{63}{10} = 6{,}3$: las medias coinciden.
b. $\sigma_A^2 = \frac{\sum x^2}{10} - 6{,}3^2 = \frac{417}{10} - 39{,}69 = 2{,}01$, $\sigma_A \approx 1{,}42$. $\sigma_B^2 = \frac{465}{10} - 39{,}69 = 6{,}81$, $\sigma_B \approx 2{,}61$.
c. El grupo A, porque su desviación típica es mucho menor con la misma media.
d. $Me_A = 6$ y $Me_B = 6{,}5$. Con 8 o más: 2 alumnos en A y 4 en B. El grupo B tiene más notas altas, pero también más suspensos.

Solución del ejercicio 8

Planteamiento. Trabajamos con las marcas de clase para la media y la desviación típica. Para comprobar la promesa de la empresa estimamos el percentil 60 por interpolación, igual que la mediana pero buscando la posición del 60 % de los datos.

a. Marcas 2, 6, 10, 14: $\bar{x} = \frac{2\cdot 18 + 6\cdot 22 + 10\cdot 12 + 14\cdot 8}{60} = \frac{400}{60} \approx 6{,}67$ min.
b. $\sigma^2 = \frac{1}{60}\sum f_i (x_i - \bar{x})^2 \approx 16{,}09$, $\sigma \approx 4{,}01$ min.
c. $\frac{12+8}{60} = \frac{1}{3} \approx 33{,}3$ %.
d. El 60 % son 36 viajeros; antes de $[4, 8)$ hay 18, luego $P_{60} = 4 + \frac{36-18}{22}\cdot 4 \approx 7{,}27$ min. Como supera 6 minutos, la promesa no se cumple.

Solución del ejercicio 9

Planteamiento. Comparamos media y mediana con y sin el dato extremo. La media usa todos los valores y se desplaza mucho con un atípico; la mediana solo depende del orden y apenas cambia, por lo que es más representativa en datos asimétricos.

a. $\bar{x} = \frac{9270}{8} = 1158{,}8$ € y $Me = \frac{820+850}{2} = 835$ €.
b. Sin el ático: $\bar{x} = \frac{5770}{7} \approx 824{,}3$ € y $Me = 820$ €.
c. La mediana: apenas cambia (de 835 a 820), mientras que la media baja más de 300 €.
d. Sí: la media está inflada por un solo piso y ningún piso normal cuesta eso. Con datos asimétricos conviene informar de la mediana.

Solución del ejercicio 10

Planteamiento. Una transformación lineal $y = ax + b$ cambia la media del mismo modo y multiplica la desviación típica por $|a|$, mientras que sumar una constante no

altera la dispersión. Lo comprobamos transformando los datos.

a. $\bar{C} = \frac{38}{5} = 7{,}6$ °C; $\sigma_C^2 = \frac{362}{5} - 7{,}6^2 = 14{,}64$, $\sigma_C \approx 3{,}83$ °C.
b. $\bar{F} = 1{,}8 \cdot 7{,}6 + 32 = 45{,}68$ °F y $\sigma_F = 1{,}8\,\sigma_C \approx 6{,}89$ °F; sumar 32 no cambia la dispersión.
c. Los datos son 35,6, 41,0, 46,4, 50,0, 55,4 °F y su media es 45,68 °F.
d. Sí: $0{,}503$ en °C y $0{,}151$ en °F. No tiene sentido usar el CV con escalas cuyo cero es arbitrario, como las de temperatura.

Solución del ejercicio 11

Planteamiento. Calculamos las medias, la covarianza y las desviaciones típicas. El coeficiente de correlación mide la fuerza de la relación lineal y la recta de regresión permite predecir, siempre dentro del rango de los datos observados.

a. $\bar{x} = 6{,}00$, $\bar{y} = 6{,}50$; $\sigma_{xy} = \frac{\sum x_i y_i}{6} - \bar{x}\bar{y} = \frac{261}{6} - 6{,}00 \cdot 6{,}50 \approx 4{,}50$.
b. $\sigma_x \approx 2{,}65$, $\sigma_y \approx 1{,}71$ y $r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \approx 0{,}996$: correlación positiva muy fuerte.
c. Pendiente $\frac{\sigma_{xy}}{\sigma_x^2} \approx 0{,}643$: $y = 6{,}50 + 0{,}643(x - 6{,}00)$, es decir $y \approx 2{,}643 + 0{,}643x$.
d. $y(6) \approx 6{,}50$. Con 20 horas saldría $15{,}50$, más de 10: no se puede extrapolar fuera del rango de los datos.

Solución del ejercicio 12

Planteamiento. Calculamos la recta de regresión y el coeficiente de determinación, que indica qué proporción de la variabilidad explica la recta. La pendiente se interpreta como el cambio medio de las ventas por unidad de publicidad.

a. $\bar{x} = 3$, $\bar{y} = 18{,}00$, $\sigma_{xy} = 5{,}80$, $\sigma_x^2 = 2$; pendiente $2{,}90$ y $y = 9{,}30 + 2{,}90x$.
b. $r \approx 0{,}989$ y $r^2 \approx 0{,}978$: el 97,8 % de la variabilidad de las ventas se explica por la recta.
c. Cada 1000 € adicionales de publicidad se asocian, de media, a unos 2900 € más de ventas.
d. No: la correlación no implica causalidad. Puede haber otras causas, como la temporada o una promoción, que aumenten a la vez ambas variables.

Solución del ejercicio 13

Planteamiento. La puntuación tipificada $z = \frac{x - \bar{x}}{\sigma}$ indica cuántas desviaciones típicas se aleja un dato de la media y permite comparar resultados de distribuciones distintas. Para el porcentaje usamos la tabla de la normal.

a. $z_L = \frac{7{,}5 - 6}{1{,}2} = 1{,}25$ y $z_P = \frac{8 - 7}{1{,}6} = 0{,}625$.
b. Lucía: está 1,25 desviaciones típicas por encima de la media, frente a 0,625 de Pablo.

c. $x = 7 + 1{,}25 \cdot 1{,}6 = 9$.
d. $P(Z > 1{,}25) \approx 0{,}106$, es decir, un 10,6 % de la clase.

Solución del ejercicio 14

Planteamiento. Construimos las frecuencias acumuladas y localizamos el intervalo que contiene cada posición buscada. La mediana y el percentil 90 se estiman por interpolación lineal dentro de su intervalo.

a. F: 10, 45, 85, 100.
b. 50 datos: la mediana está en $[200, 300)$: $Me = 200 + \frac{50-45}{40} \cdot 100 = 212{,}5$ L/día.
c. 90 datos: está en $[300, 400)$: $P_{90} = 300 + \frac{90-85}{15} \cdot 100 \approx 333{,}3$ L/día.
d. A partir de unos 333 litros diarios.

Solución del ejercicio 15

Planteamiento. Hacemos un estudio descriptivo completo: tabla, medias, dispersión, mediana y moda. Comparamos los resultados con los objetivos fijados y, como la variable es ordinal, valoramos qué medidas son más adecuadas.

a. $\bar{x} = \frac{3+14+45+180+150}{100} = 3{,}92$; $\sigma^2 = \frac{\sum f_i x_i^2}{100} - \bar{x}^2 = 0{,}99$, $\sigma \approx 1{,}00$.
b. Acumuladas: 3, 10, 25, 70, 100. La mediana (datos 50 y 51) es 4; la moda también es 4.
c. Media $3{,}92 \geq 3{,}8$: sí. Valoraciones 4 o 5: $45 + 30 = 75$ %, que supera el 70 %: sí.
d. Es una variable ordinal: la mediana, la moda y los porcentajes por categoría son más adecuados, porque la distancia entre «mala» y «regular» no tiene por qué ser la misma que entre «buena» y «excelente».
e. El protocolo cumple ambos objetivos: tres de cada cuatro pacientes lo valoran bien o muy bien, aunque un 10 % lo considera malo o muy malo y conviene revisar esos casos.

Solución del ejercicio 16

Planteamiento. Un crecimiento exponencial se convierte en lineal tomando logaritmos. Ajustamos la recta de regresión a los pares $(t, \ln y)$ y deshacemos el cambio para obtener los parámetros del modelo exponencial.

a. $\ln y = \ln a + \ln e^{bt} = \ln a + bt$: es una recta con pendiente b y ordenada $\ln a$.
b. $4{,}605;\ 5{,}011;\ 5{,}394;\ 5{,}799;\ 6{,}215$.
c. $\bar{t} = 2$, , pendiente $b \approx 0{,}4007$, ordenada $\approx 4{,}603$ y $r \approx 0{,}9999$: el ajuste lineal es casi perfecto.
d. $a = e^{4{,}603} \approx 99{,}8$ y $y \approx 99{,}8\, e^{0{,}4007t}$. Para $t = 5$: $y \approx 740$ mil bacterias.
e. $e^{bT} = 2 \Rightarrow T = \frac{\ln 2}{b} \approx 1{,}73$ h.

Solución del ejercicio 17

Planteamiento. La desigualdad de Chebyshev da una proporción mínima garantizada para cualquier distribución. Expresamos cada intervalo en número de desviaciones típicas k y aplicamos la fórmula $1-\frac{1}{k^2}$.

a. $k=\frac{100}{50}=2$: al menos $1-\frac{1}{4}=75\ \%$.
b. $k=3$: al menos $1-\frac{1}{9}\approx 88{,}9\ \%$.
c. $1-\frac{1}{k^2}=0{,}96\Rightarrow k^2=25$, $k=5$: el intervalo $[750,1250]$ h.
d. Chebyshev vale para cualquier distribución, incluidas las más desfavorables, así que solo da una cota mínima; conocer la forma de la distribución permite afinar.

Solución del ejercicio 18

Planteamiento. La media conjunta es la media ponderada por los tamaños. Para la varianza recuperamos la suma de cuadrados de cada grupo con la fórmula abreviada, las sumamos y aplicamos de nuevo la fórmula al conjunto.

a. $\bar{x}=\frac{30\cdot 6{,}2+20\cdot 7}{50}=\frac{186+140}{50}=6{,}52$.
b. Mañana: $30(1{,}5^2+6{,}2^2)=1220{,}7$. Tarde: $20(1^2+7^2)=1000$.
c. $\sigma^2=\frac{1220{,}7+1000}{50}-6{,}52^2\approx 1{,}9036$ y $\sigma\approx 1{,}380$.
d. Porque, además de la dispersión dentro de cada grupo, aparece la dispersión entre grupos debida a que sus medias son distintas.

Solución del ejercicio 19

Planteamiento. Calculamos el coeficiente de correlación con y sin el dato extremo. Así vemos que un solo punto alejado puede cambiar mucho el resultado, y que antes de calcular conviene dibujar el diagrama de dispersión.

a. $r\approx 0{,}741$.
b. $r\approx 0{,}832$.
c. Con el gran almacén, un solo punto muy alejado domina el cálculo de la covarianza y la desviación típica de y. Sin él, la relación entre los pequeños comercios se ve de forma más fiel.
d. Dibujar siempre el diagrama de dispersión, detectar datos atípicos y decidir, según el objetivo del estudio, si deben analizarse aparte.

Solución del ejercicio 20

Planteamiento. Demostramos las propiedades con manipulaciones algebraicas de sumatorios. Para los mínimos derivamos la función de a en el caso de la media y razonamos con distancias en el caso de la mediana.

a. $\sum x_i-n\bar{x}=n\bar{x}-n\bar{x}=0$.

b. Desarrollando el cuadrado: $\frac{1}{n}\sum x_i^2 - 2\bar{x}\frac{1}{n}\sum x_i + \bar{x}^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2$.
c. $S'(a) = -2\sum(x_i - a) = 0 \iff a = \bar{x}$, y $S'' = 2n > 0$: mínimo.
d. Con x_1: si a se mueve entre x_1 y x_3, la suma de las distancias a los extremos es constante, $x_3 - x_1$, y la distancia a x_2 es mínima cuando $a = x_2$, la mediana.