Estadística I
Grado en Matemáticas, UAM, 2018-2019

Hoja 1 (Estadística descriptiva)

EJERCICIOS COMPUTACIONALES

1. En la hoja de cálculo `datos-hoja-ej1-18-19.xls` encontrarás unas cuantas series de datos y unas cuantas cuestiones de Estadística descriptiva para cada una de ellas. El objetivo es que utilices excel para darles respuesta.

RESÚMENES Y REPRESENTACIONES DE MUESTRAS

2. Determina razonadamente si las siguientes afirmaciones son verdaderas o falsas:

a) Si añadimos 7 a todos los datos de una muestra, el primer cuartil aumenta en 7 unidades y el rango intercuartílico no cambia.
b) Al restar 1 a cada dato de una muestra, la desviación típica siempre disminuye.
c) Si cambiamos el signo de todos los datos de una muestra, el coeficiente de asimetría también cambia de signo.
d) Al multiplicar por 3 todos los datos de una muestra, el coeficiente de asimetría no varía.
e) Si a una muestra de datos con media $\bar{x}$ se le añade un nuevo dato que coincide con $\bar{x}$, la media no cambia y la desviación típica disminuye.

3. a) Disponemos de una serie de datos $(x_1, \ldots, x_{100})$, ya ordenados de menor a mayor, cuya media muestral es $\bar{x}$. Ahora formamos una nueva serie añadiendo a la anterior los valores x_1 y x_{100}. ¿Qué condición se debe cumplir para que la media muestral de la nueva muestra coincida con $\bar{x}$, la media muestral de la muestra original?

b) Disponemos de una serie de datos emparejados $((x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n))$. Los datos x_i tienen media $\bar{x}$ y los datos y_i tienen media $\bar{y}$. Añadimos a la serie anterior un dato más, la pareja $(\bar{x}, \bar{y})$. Determina si la covarianza de la nueva serie es mayor, menor o igual que la covarianza de la serie original.

c) Tenemos una serie de datos $(x_1, x_2, \ldots, x_n)$, cuya media es a y cuya varianza muestral es b. Duplicamos ahora el tamaño de la serie añadiendo los opuestos (en signo) de los datos originales:

$$(x_1, x_2, \ldots, x_n, -x_1, -x_2, \ldots, -x_n) .$$

Llamemos b' a la varianza muestral de la nueva serie de datos. ¿Quién es mayor, b ó b'?

4. Tenemos una muestra $x_1, \ldots, x_n$. Denotemos su media por M_n. Añadimos un dato x_{n+1}, y la nueva media es M_{n+1}. Comprueba que

$$M_{n+1} = \frac{n}{n+1} M_n + \frac{1}{n+1} x_{n+1}.$$

Interpreta el resultado.

5. La media de las variaciones mensuales del PIB de la Comunidad de Murcia de los nueve primeros meses del año ha sido del 0.1 %. ¿Cuál debe ser la media de los últimos tres meses para que la media anual cumpla el objetivo del 0.2 %?

6. Dada una muestra $(x_1, \ldots, x_n)$, comprueba que

$$n^2 V_x = (n-1) \sum_i x_i^2 - \sum_{i \neq j} x_i x_j \,.$$

Deduce que la forma cuadrática dada por la matriz simétrica A con

$$a_{ij} = \begin{cases} n-1 & \text{si } i = j \\ -1 & \text{si } i \neq j \end{cases}$$

es (semi-)definida positiva (¿por qué no es definida positiva?).

7. Sea $(x_1, \ldots, x_n)$ una muestra. Prueba que si para un cierto $\varepsilon \geq 0$ se tiene que $|x_i| \leq \varepsilon$ para todo $i = 1, \ldots, n$, entonces se cumple que $V_x \leq \varepsilon^2$.

Correlaciones, covarianzas y recta de regresión

8. En cada una de las siguientes muestras, se ha sustituido un número por z. Si es posible, calcula z de forma que el coeficiente de correlación valga 1. Si no es posible, explica la razón.

Datos A: $(1,1)$, $(2,3)$, $(2,3)$, $(4,z)$. Datos B: $(1,1)$, $(2,3)$, $(3,4)$, $(4,z)$.

9. Tenemos una muestra de datos emparejados $(x_1, y_1), \ldots, (x_n, y_n)$. Prueba que, si para un cierto $\varepsilon > 0$ (y para unos números reales a, b) se cumple que

$$|y_i - (ax_i + b)| \leq \varepsilon \quad \text{para } 1 \leq i \leq n\,,$$

entonces

$$|\mathrm{cov}_{x,y} - aV_x| \leq \sqrt{V_x}\, \varepsilon\,.$$

10. Los datos de mortalidad infantil (muertes por mil partos) en un país durante los años 2008-2012 fueron (tomando 2010 como año 0):

X: año	-2	-1	0	1	2
Y: mortandad	14.5	13.8	12.7	11.9	11.4

a) Ajusta a estos datos una ecuación de la forma $Y = ae^{bX}$, transformando primero a una regresión lineal.

b) Calcula el coeficiente de correlación de la regresión lineal y comenta la bondad del ajuste.

c) ¿Qué mortalidad infantil se espera para 2020 (año 10) si se da por bueno el ajuste anterior?

11. En la tabla se recogen medidas (bajo ciertas condiciones) del volumen de un determinado gas al someterlo a distintas presiones:

P presión	1	1.5	2	2.5	3
V volumen	1	0.76	0.62	0.52	0.46

a) Ajusta a estos datos una ecuación de la forma $V = aP^b$, transformando primero a una regresión lineal.

b) Calcula el coeficiente de correlación en el problema transformado y cuantifica la bondad del ajuste.

c) ¿Qué volumen corresponde a $P = 3.5$ si se da por bueno el ajuste anterior?

12. Como en el ejercicio 4, añadimos a una muestra $x_1, \ldots, x_n$ un dato x_{n+1}. Escribe una expresión para la varianza V_{n+1} de la muestra ampliada en términos de las características de la muestra original.

13. Dada una muestra $(x_1, y_1), \ldots, (x_n, y_n)$ con $n \geq 2$, se pide obtener la recta $y = \hat{b}x$, que pasa por el origen $(0,0)$, que da el menor error cuadrático medio de entre todas las rectas de ecuación $y = bx$. Da la fórmula de $\hat{b}$ y la expresión del error cuadrático mínimo en términos de la muestra.

14. Disponemos de una muestra $(x_1, y_1), \ldots, (x_n, y_n)$ con $n \geq 2$. Para el ajuste, vamos a considerar que los distintos puntos tienen importancia relativas distintas dadas por unos pesos $\pi_1, \ldots, \pi_n$, tales que $\pi_i > 0, 1 \leq i \leq n$ y $\sum_{i=1}^n \pi_i = 1$. Para una recta genérica de ecuación $y = a + bx$, se considera el error cuadrático ponderado:

$$\breve{E}(a, b) = \sum_{i=1}^{n} \pi_i \big(y_i - (a + bx_i)\big)^2 .$$

Halla la recta de ecuación $y = \breve{a} + \breve{b}x$ que da el menor error cuadrático ponderado.

15. Dada una muestra $(x_1, y_1), \ldots, (x_n, y_n)$ con $n \geq 3$, se pide obtener la parábola $y = \hat{a} + \hat{b}x + \hat{c}x^2$ que da el menor error cuadrático medio de entre todas las parábolas $y = a + bx + cx^2$. Esto es, obtener las fórmulas de $\hat{a}, \hat{b}$ y $\hat{c}$ y la expresión del menor error cuadrático en términos de la muestra. Generaliza al ajuste con polinomios de grado d, con $2 \leq d \leq n$.

(SUGERENCIA: escribe matricialmente el sistema lineal resultante y expresa la solución en términos de las matrices involucradas.)

16. Dada una muestra $(x_1, y_1), \ldots, (x_n, y_n)$ con $n \geq 2$, donde las x_i y las y_i ya están tipificadas, se trata de obtener la recta $y = \hat{a} + \hat{b}x$ que da el menor error cuadrático medio medido en la forma:

$$\widetilde{E}(a, b) = \frac{1}{n} \sum_{i=1}^{n} \operatorname{dist}\big((x_i, y_i); y = a + bx\big)^2 .$$

Aquí $\operatorname{dist}\big((x_0, y_0); y = a + bx\big)$ denota la distancia euclídea del punto (x, y) a la recta de ecuación $y = a + bx$.

Verifica primero que $\hat{a} = 0$ y da una expresión de $\hat{b}$ en términos de $\rho(x, y)$.

HOJA 1

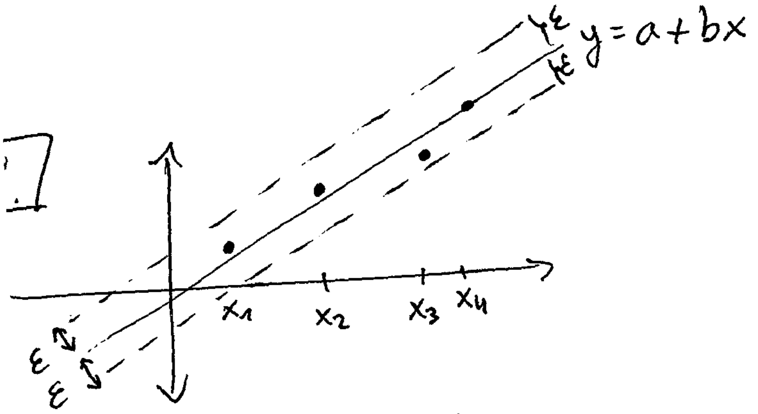
5. $\underbrace{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9}_{\text{media } = 0'1\%}, \underbrace{x_{10}, x_{11}, x_{12}}_{¿\text{media}=\mu?}$

$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{media} = 0'2\%}$

$$\frac{(9 \cdot 0'1) + (3 \cdot \mu)}{12} = 0'2 \Rightarrow 0'9 + 3\mu = 2'4 \Rightarrow \boxed{\mu = 0'5}$$

7. $$\bar{V} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2 \leq \overline{x^2} \leq \varepsilon^2$$

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq \frac{n\varepsilon^2}{n} = \varepsilon^2$$



Hipótesis: $|y_i - (bx_i + a)| \leq \varepsilon$

$$Cov_{x,y} - bV_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - \frac{b}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\left(y_1 - \bar{y} - b(x_i - \bar{x})\right) =$$

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

Entonces:

$$\left|\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\left(y_i - bx_i - a - (\bar{y} + b\bar{x} - a)\right)\right| \overset{\text{Cauchy-Schwartz}}{\leq} \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{1/2}}_{\sqrt{V_x}} \cdot$$

$$\cdot \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\left(y_i - bx_i - a - (\bar{y} - b\bar{x} - a)\right)^2\right)^{1/2}}_{\overset{\wedge|}{\varepsilon} \text{ por el ej. 7.}}$$

HOJA 2

2. $\mathbb{X} = (X_1, X_2, X_3)^T$ $\quad m = (1,1,0)$ $\quad V = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix}$

$\mathbb{Z} = (Z_1, Z_2, Z_3)^T$ $\quad \begin{cases} Z_1 = X_1 + X_2 \\ Z_2 = X_1 + X_2 + X_3 \\ Z_3 = 2X_1 + X_2 \end{cases}$

FORMA 1: $\mathbb{E}(Z_1) = \mathbb{E}(X_1 + X_2) = \mathbb{E}(X_1) + \mathbb{E}(X_2) = 1+1 = 2$

$\mathbb{E}(Z_2) = 2$

$\mathbb{E}(Z_3) = 3$

$$V(Z_3) = V(2X_1 + X_2) = \underbrace{V(2X_1)}_{\substack{= 4V(X_1) \\ = 4.1}} + \underbrace{V(X_2)}_{= 2} + \underbrace{2\mathrm{cov}(2X_1, X_2)}_{\substack{= 2.2\,\mathrm{cov}(X_1,X_2) \\ = 2.2.1 = 4}} = 1$$

También:

$$V(Z_3) = \mathbb{E}(Z_3^2) - \mathbb{E}(Z_3)^2 = 10$$

$$\mathrm{Cov}(Z_1, Z_2) = \mathrm{cov}(X_1 + X_2, Z_2) = \mathrm{cov}(X_1, Z_2) + \mathrm{cov}(X_2, Z_2) =$$
$$= \mathrm{cov}(X_1, X_1+X_2+X_3) + \mathrm{cov}(X_2, X_1+X_2+X_3) =$$
$$= \mathrm{cov}(X_1, X_1) + \mathrm{cov}(X_1, X_2) + \mathrm{cov}(X_1, X_3) + \mathrm{cov}(X_2, X_1) + \mathrm{cov}(X_2, X_2) + \mathrm{cov}(X_2, X$$
$$= 1 + 1 + 1 + 1 + 2 + 2 = 8$$

FORMA 2: $\mathbb{Z} = \underbrace{\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 2 & 1 & 0 \end{pmatrix}}_{= B} \mathbb{X}$ $\qquad \mathbb{E}(\mathbb{Z}) = B.\mathbb{E}(\mathbb{X}) = B\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$

Sabemos que $\mathbb{Z}$ es $N\left(\begin{pmatrix} 2 \\ 2 \\ 3 \end{pmatrix}, ?\right)$

BVB^t = matriz varianzas/covarianzas de $\mathbb{Z}$

$Y, Z_1, \ldots, Z_n$ v.a. indep. $N(0,1)$ $\rho \in (0,1)$

$$X_j = \sqrt{\rho}\, Y + \sqrt{1-\rho}\, Z_j$$

$\mathbb{X} = (X_1, \ldots, X_n)^T$ es un vector normal (normal multidimensional)

$$\mathbb{E}(X_j) = \sqrt{\rho}\, \mathbb{E}(Y) + \sqrt{1-\rho}\, \mathbb{E}(Z_j) = \sqrt{\rho} \cdot 0 + \sqrt{1-\rho} \cdot 0 = 0$$

$$V(X_j) = \mathbb{E}(X_j^2) - \mathbb{E}(X_j)^2 = \mathbb{E}(X_j^2) \overset{(1)}{=} \rho \underbrace{\mathbb{E}(Y^2)}_{1} + 2\sqrt{\rho}\sqrt{1-\rho}\, \underbrace{\mathbb{E}(\cancel{Y Z_j})}_{0 \text{ indep.}} +$$
$$+ (1-\rho) \underbrace{\mathbb{E}(Z_j^2)}_{1} = 1$$

$j = j$

$$(2) \quad \mathbb{E}(X_i X_j) = \rho \underbrace{\mathbb{E}(Y^2)}_{1} + \sqrt{\rho}\sqrt{1-\rho}\, \mathbb{E}(\cancel{Y Z_j})^{0} + \sqrt{\rho}\sqrt{1-\rho}\, \mathbb{E}(\cancel{Y Z_j})^{0} +$$
$$+ (1-\rho)\, \mathbb{E}(\cancel{Z_i Z_j})^{0} = \rho$$

$$\Rightarrow \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = N\left(\vec{0}, \begin{pmatrix} 1 & & \rho \\ & \ddots & \\ \rho & & 1 \end{pmatrix}\right)$$