Estadística I
Grado en Matemáticas, UAM, 2014-2015

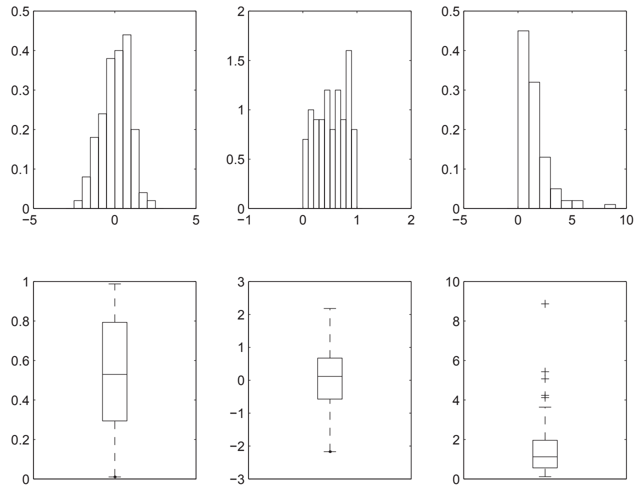
Hoja 1 (Estadística descriptiva)

Resúmenes y representaciones de muestras

1. 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.

2. Para tres muestras de datos se han representado los correspondientes histogramas y diagramas de cajas. Relaciona cada histograma con el diagrama de cajas que le corresponde:



3. Dada una muestra $(x_1, x_2, \ldots, x_n)$, comprueba que la media $\overline{x}$ es aquel valor respecto del cual la dispersión cuadrática media es mínima, es decir, considera la función d dada por

$$a \in \mathbb{R} \mapsto d(a) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2\,,$$

y comprueba que el mínimo de d se alcanza en $a = \overline{x}$ (por tanto, el valor mínimo de d es justamente la varianza de los x_i).

4. Comprueba que

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

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?).

5. Prueba que si para un cierto $\varepsilon \geq 0$ se tiene que $|x_i| \leq \varepsilon$ para todo $i = 1, \dots, n$, entonces se cumple que $V \leq \varepsilon^2$.

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

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

Obtén una expresión recursiva análoga para las varianzas V_n y V_{n+1}.

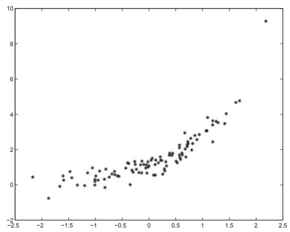
7. La media de las variaciones mensuales del PIB de la Comunidad de Murcia de los nueve primeros meses del año ha sido del 0.1 %. ¿Cuál debe ser la media de los últimos tres meses para que la media anual cumpla el objetivo del 0.2 %?

Correlaciones, covarianzas y recta de regresión

8. Responde razonadamente a las siguientes cuestiones:

a) Si para todo i se cumple que $y_i < x_i$, ¿el coeficiente de correlación entre x e y es negativo?

b) Se muestra a continuación la nube de puntos de una muestra de dos variables (X, Y):



De los tres valores siguientes: 1 %, 83 % y −73 %, ¿cuál podría corresponder al coeficiente de correlación entre x e y?

9. 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: $(1,1),\ (2,3),\ (3,4),\ (4,z)$.

10. Prueba que, si para un cierto $\varepsilon > 0$ (y reales a, b) se cumple que

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

entonces

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

11. 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?

12. 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 lineal en el problema transformado y comenta la bondad del ajuste.

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

Ejercicios adicionales

13. Dada una muestra $(x_1, y_1), \dots, (x_n, y_n)$ 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 $y = b\,x$. Da la fórmula de $\hat{b}$ y la expresión del menor error cuadrático en términos de la muestra.

14. Disponemos de una muestra $(x_1, y_1), \dots, (x_n, y_n)$ con $n \geq 2$. Para el ajuste, vamos a considerar que los distintos puntos tienen importancia relativas distintas dadas por unos pesos $\pi_1, \dots, \pi_n$, 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 $(x_1, y_1), \dots, (x_n, y_n)$ 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 $(x_1, y_1), \dots, (x_n, y_n)$ 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:

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

Aquí $\mathrm{dist}\big((x_0, y_0); y = a + bx\big)$ 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 $\rho(x, y)$.

Ejercicios de ordenador

17. En la hoja de cálculo `datos-hoja-ej1.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.