Estadística descriptiva

PFG-JLF

UAM

Estadística I, 2018-2019

Datos/muestra

Punto de partida: disponemos

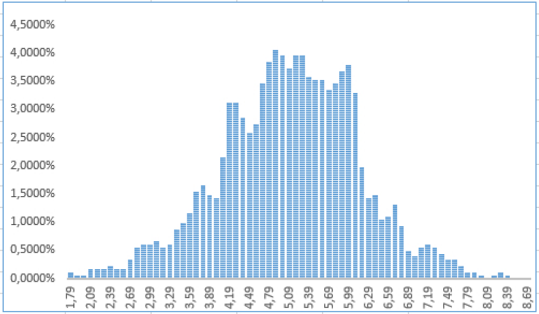
- de un conjunto de valores/datos (muestra)
- de una cierta característica/variable X
- en una población específica.

Notación para la muestra:

$$x_1, x_2, \ldots, x_n$$

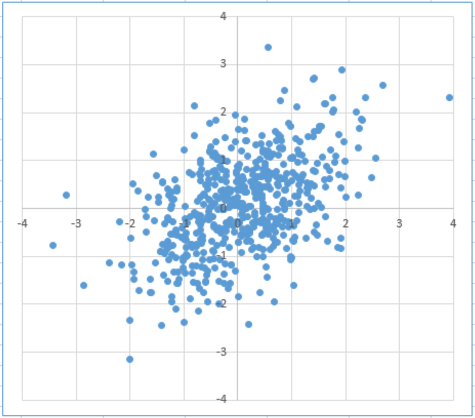
- El orden de los datos es irrelevante.
- El tamaño de la muestra es n.

dato1	1,793
dato2	4,910
dato3	4,181
dato4	3,444
dato5	3,932
dato6	5,573
dato7	7,242
dato8	5,654
dato9	6,089
dato10	4,450
dato11	5,490
dato12	5,957
dato13	5,828
dato14	3,992
dato15	4,816
dato16	3,592
dato17	5,480
dato18	5,939



01/08/2018	13,22
02/08/2018	13,23
03/08/2018	13,22
04/08/2018	13,22
05/08/2018	13,23
06/08/2018	13,21
07/08/2018	13,20
08/08/2018	13,20
09/08/2018	13,20
10/08/2018	13,18
11/08/2018	13,19
12/08/2018	13,19
13/08/2018	13,18
14/08/2018	13,16
15/08/2018	13,17
16/08/2018	13,17
17/08/2018	13,17
18/08/2018	13,17
19/08/2018	13,16

	X	Y
dato1	-0,17814212	0,39850918
dato2	1,28589318	0,53135167
dato3	0,72688509	-0,69749511
dato4	0,50832048	0,93855501
dato5	0,08383754	0,10192633
dato6	-0,28287457	-0,4339356
dato7	-1,22252332	0,07240645
dato8	-0,6164752	0,26955439
dato9	-1,423409	-2,44207263
dato10	1,70310211	0,78215113
dato11	-0,35485206	-1,52736313
dato12	-0,94234329	-0,98343951
dato13	0,94706423	1,78169846
dato14	0,40940862	0,66310579
dato15	-0,40834039	0,26060408
dato16	0,31072628	-0,14738146
dato17	-0,2289009	1,47048344
dato18	0,10688676	0,6006816
dato19	0,0787881	-0,41966805



Representación de los datos

- Se determinan una serie de clases $C_1, \ldots, C_k$ (generalmente intervalos),
- de manera que todos los datos caigan en alguna de las clases;
- se cuenta el número de datos en cada clase: $n_1, \ldots, n_k$ (números que suman n);
- o mejor, la frecuencia relativa en cada clase: $f_1, \ldots, f_k$, donde $f_j = n_j/n$ (estos números f_j suman 1);
- y se representan gráficamente, generalmente con un diagrama de barras.

Medidas descriptivas de la muestra

Media muestral:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Mediana: el "valor" que deja tantos datos a su izquierda como a su derecha. Se ordenan los datos de menor a mayor:

$$x_{i_1} \leq x_{i_2} \leq \cdots \leq x_{i_n} .$$

- Si n es impar, $n = 2r + 1$, $\mathrm{MED}_x = x_{i_{r+1}}$.
- Si n es par, $n = 2r$, se toma (habitualmente) $\mathrm{MED}_x = \frac{x_{i_r} + x_{i_{r+1}}}{2}$.

Medidas descriptivas de la muestra

Varianza muestral:

Definición:

$$V_x = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^2.$$

Regla habitual de cálculo:

$$V_x = \overline{x^2} - \overline{x}^2,$$

donde

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

Cuasivarianza:

$$s_x^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \overline{x})^2.$$

Desviación típica:

$$\sqrt{V_x} = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (x_j - \overline{x})^2} = \sqrt{\overline{x^2} - \overline{x}^2}$$

Cuasidesviación típica:

$$s_x = \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} (x_j - \overline{x})^2}.$$

Observaciones:

- $(n-1)\, s_x^2 = nV_x$.
- Las unidades de la desviación típica son las "correctas".
- Varianzas/desviaciones típicas muestrales grandes indican que los datos están bastante "dispersos" con respecto a la media muestral.
- $V_x = 0$ si y solo si datos constantes.

Cambios de escala/tipificación

Cambio de escala: para $a, b \in \mathbb{R}$, con $b \neq 0$,

$$x_1, \ldots, x_n \quad \longrightarrow \quad z_1, \ldots, z_n,$$

donde

$$z_i = a + b x_i \qquad \text{para cada } i = 1, \ldots, n.$$

Se tiene que

$$\overline{z} = a + b\,\overline{x}, \qquad V_z = b^2\, V_x \qquad \text{y} \qquad s_z^2 = b^2 s_x^2\,.$$

Tipificación. El cambio

$$x_i \mapsto z_i = \frac{x_i - \overline{x}}{\sqrt{V_x}}$$

transforma los datos $x_1, \ldots, x_n$ en datos $z_1, \ldots, z_n$ tales que

$$\overline{z} = 0 \qquad V_z = 1.$$

Otras medidas

Cuartiles. Para datos ya ordenados

$$x_1 \leq x_2 \leq \cdots \leq x_n$$

se definen los cuartiles

$$Q_0, Q_1, Q_2, Q_3, Q_4$$

como sigue:

- $Q_0 = \text{mín}(x_1, \ldots, x_n)$, $Q_4 = \text{máx}(x_1, \ldots, x_n)$.
- $Q_2 = \mathrm{MED}_x$.
- Q_3 es la mediana de los datos que están entre Q_2 y Q_4, y Q_1 es la mediana de los que están entre Q_0 y Q_2.

- rango de la muestra: $Q_4 - Q_0$;
- rango intercuartílico: $\mathrm{RIC} = Q_3 - Q_1$;
- valores atípicos:
  - mayores que $Q_3 + 1.5 \times \mathrm{RIC}$,
  - menores que $Q_1 - 1.5 \times \mathrm{RIC}$.

Coeficiente de asimetría muestral:

$$\mathrm{ASIM}_x = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{V_x^{3/2}}$$

Datos bidimensionales

Dos magnitudes, X e Y, medidas sobre los mismos individuos.

Muestra:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\,.$$

Se representan con un diagrama de dispersión.

Medidas de dependencia lineal

Covarianza muestral

$$\mathrm{cov}_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y}) = \overline{xy} - \overline{x}\cdot\overline{y},$$

donde

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i\, y_i.$$

- $\mathrm{cov}_{x,y} > 0$ indica dependencia positiva (relación directa entre las variables X e Y);
- $\mathrm{cov}_{x,y} < 0$ indica dependencia negativa (relación inversa entre las variables X e Y).

Se tiene (desigualdad de Cauchy–Schwarz) que

$$|\mathrm{cov}_{x,y}| \leq \sqrt{V_x}\sqrt{V_y}.$$

Con igualdad si y sólo si los datos (x_j, y_j) están *todos sobre una misma recta*.

Coeficiente de correlación

Si $V_x \neq 0$, $V_y \neq 0$,

$$\rho_{x,y} = \frac{\mathrm{cov}_{x,y}}{\sqrt{V_x}\sqrt{V_y}}.$$

Propiedades:

- su signo tiene el mismo significado que el de la covarianza;
- $-1 \leq \rho_{x,y} \leq 1$;
- invariante bajo cambios de escala;
- $\rho_{x,y} = \pm 1$ si y sólo si datos sobre una recta.

La recta de regresión

Tenemos una muestra del par de variables (X, Y):

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\,.$$

De entre todas las rectas $y = a + bx$, *¿cuál es la que "mejor" aproxima/explica la muestra?*

¿Qué significa "mejor ajuste"? Para $a, b \in \mathbb{R}$, definimos el error cuadrático medio

$$E(a, b) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (a + b\,x_i))^2$$

(errores "verticales"). Buscamos a, b que minimicen esta cantidad.

Escribimos

$$
\begin{aligned}
E(a,b) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - (a + b\,x_i))^2 \\
&= a^2 + \overline{x^2}\,b^2 - 2\overline{y}\,a + 2\overline{x}\,ab - 2\overline{xy}\,b + \overline{y^2} \\
&= \underbrace{(a,b)\begin{pmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix}}_{\text{forma cuadrática}} \underbrace{-2\overline{y}\,a - 2\overline{xy}\,b}_{\text{términos lineales}} + \underbrace{\overline{y^2}}_{\text{constante}}
\end{aligned}
$$

La forma cuadrática es definida positiva (el determinante es justamente V_x), así que tiene al menos un mínimo.

Cálculo del mínimo:

$$\begin{cases} \dfrac{\partial E(a,b)}{\partial a} = 2a + 2\overline{x}b - 2\overline{y} = 0\,, \\ \dfrac{\partial E(a,b)}{\partial b} = 2\overline{x^2}b + 2\overline{x}a - 2\overline{xy} = 0\,. \end{cases}$$

Es decir,

$$\begin{cases} \overline{y} = a + b\,\overline{x}\,, \\ \overline{xy} = a\,\overline{x} + b\,\overline{x^2}\,. \end{cases}$$

Solución:

$$\widehat{b} = \frac{\mathrm{cov}_{x,y}}{V_x}$$

$$\widehat{a} = \overline{y} - \widehat{b}\,\overline{x} = \overline{y} - \left(\frac{\mathrm{cov}_{x,y}}{V_x}\right)\overline{x}$$

La recta de ecuación

$$y = \widehat{a} + \widehat{b}\,x$$

que da el mínimo error cuadrático medio, es la recta de regresión de Y sobre X.

Escrituras alternativas:

$$y - \overline{y} = \widehat{b}\,(x - \overline{x}) \qquad \text{o bien} \qquad \frac{y - \overline{y}}{\sqrt{V_y}} = \rho_{x,y}\,\frac{x - \overline{x}}{\sqrt{V_x}}.$$

Bondad de ajuste

El (mínimo) error cuadrático es $E(\widehat{a}, \widehat{b})$, que se puede escribir como

$$E(\widehat{a}, \widehat{b}) = \sqrt{V_y}\sqrt{1 - \rho_{x,y}^2}.$$

La cantidad $\rho_{x,y}^2$ es conocida como el coeficiente de determinación R^2.

Valores de R^2 próximos a 1 indican buen ajuste de la recta de regresión. Por ejemplo, R^2 del orden de 0.8 o 0.9 son "buenos" ajustes.

En el análisis de la recta de regresión se suelen dar

- los valores de $\widehat{a}$ y $\widehat{b}$,
- y el valor de R^2.

Transformación de datos

Ajuste logarítmico. Se quiere ajustar una curva del tipo

$$y = B \ln(x) + A$$

a unos datos (x_i, y_i), con $x_i > 0$.
Procedimiento:

- nueva variable $Z = \ln(X)$,
- transformamos los datos de la muestra: definimos $z_i = \ln(x_i)$,
- ajustamos recta de regresión a los (z_i, y_i),

$$y = \widehat{a} + \widehat{b} z.$$

- el ajuste a los datos originales será

$$y = \widehat{b} \ln(x) + \widehat{a},$$

es decir, $A = \widehat{a}$ y $B = \widehat{b} = \mathrm{cov}_{\ln(x),y} / V_{\ln(x)}$.

Ajuste exponencial. Se quiere ajustar una curva del tipo

$$y = C\, e^{Dx}$$

a unos datos (x_i, y_i), con $y_i > 0$.
Procedimiento:

- nueva variable $W = \ln(Y)$; equivalentemente, $Y = e^W$,
- transformamos los datos de la muestra: definimos $w_i = \ln(y_i)$,
- ajustamos recta de regresión a los (x_i, w_i),

$$w = \widehat{a} + \widehat{b}x,$$

- el ajuste a los datos originales será

$$y = e^w = e^{\widehat{a}} e^{\widehat{b}x},$$

es decir, $C = e^{\widehat{a}}$ y $D = \widehat{b} = \mathrm{cov}_{x,\ln(y)}/V_x$.