

Estadística Aplicada y Cálculo Numérico (Grado en Química)

Valeri Makarov

Dept. de Matemática Aplicada, U.C.M.

10/02/2015 – 29/05/2015

F.CC. Matemáticas, Desp. 420
<http://www.mat.ucm.es/~vmakarov>
e-mail: vmakarov@mat.ucm.es

Capítulo 2

Estadística descriptiva. Ajustes por mínimos cuadrados

Libro: Matemáticas B, 4-to de la ESO

Frecuencia relativa: Es la fracción

$$f_i = \frac{n_i}{n}$$

donde n es el número total de los datos. En nuestro caso $n = 30$.

También se puede expresar las **frecuencias relativas en porcentajes**.

Las ventajas de frecuencias relativas:

1. Unidades adimensionales
2. Proporcionan información que no "depende" del tamaño de la muestra

Frecuencia acumulada: Es el número de veces que ocurren los valores menores o iguales que un valor concreto.

$$a_i = \sum_{k=1}^i n_k$$

Es obvio que el último término es igual a n .

Podemos introducir la frecuencia acumulada relativa:

$$h_i = \frac{a_i}{n}$$

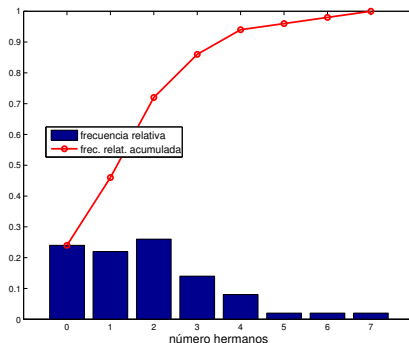
El último valor es igual a 1.

Problema 1a: Dibujar frecuencias relativas y acumuladas.

<i>Número de hermanos</i>	0	1	2	3	4	5	6	7
<i>Número de alumnos</i>	12	11	13	7	4	1	1	1

Resolución: $n = 50$

f_i	0,24	0,22	0,26	0,14	0,08	0,02	0,02	0,02
h_i	0,24	0,46	0,72	0,86	0,94	0,96	0,98	1,00



Medidas de tendencia central

- ▶ Frecuentemente al describir datos es conveniente **resumir la información con un solo número**.
- ▶ Este número que suele situarse hacia el centro de la distribución de datos se denomina **medida de tendencia central**.

Media aritmética de un conjunto $x_1, x_2, \dots, x_n$ se define como:

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

Mediana es el valor que ocupa el lugar central de los datos ordenados en orden creciente.

Si n es impar: $M = x_{(n+1)/2}, \quad x_1 \leq x_2 \leq \cdots \leq x_n$

Si n es par: $M = (x_{n/2} + x_{n/2+1})/2$

Ejemplo n impar:

1 2 3 3 4 **6** 7 7 8 8 9

$M = 6$. Para el mismo conjunto $\bar{x} = 5,27$.

Ejemplo n par:

1 2 3 3 4 **6 7** 7 8 8 9 10

$M = (6 + 7)/2 = 6,5$. Para el mismo conjunto $\bar{x} = 5,67$.

Mediana es más estable respecto a los valores atípicos

Medidas de dispersión

Dada una muestra: $x = \{x_1, x_2, \dots, x_n\}$, medidas de dispersión tienen como el objetivo dar **el campo de variabilidad** del conjunto de datos.

Varianza: se define como

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

Podemos desarrollar la formula de la varianza:

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

Covarianza

Para darnos una idea de la presencia de dependencias entre las variables x_i e y_i podemos usar la siguiente medida:

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Podemos desarrollar la fórmula:

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y}) = \left[\frac{1}{n} \sum_{i=1}^n x_i y_i \right] - \bar{x} \bar{y}$$

$s_{xy} > 0$: los valores grandes positivos de x_i suele corresponder a los valores grandes positivos de y_i

$s_{xy} < 0$: los valores grandes de x_i suelen corresponder a los valores grandes negativos de y_i .

$s_{xy} \approx 0$: no existe ninguna tendencia lineal entre las x e y .

Regresión lineal

Para una muestra

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

vamos a considerar la variable x como predictora y queremos encontrar una recta

$$y = ax + b$$

tal que mejor prediga la variable y .

Minimizamos el **error cuadrático medio**:

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

$E(a, b)$ es una función parabólica de dos variables.

Mínimo de $E(a, b)$:

$$\frac{\partial E(a, b)}{\partial a} = -\frac{2}{n} \sum_{i=1}^n (y_i - ax_i - b)x_i = 0$$

$$\frac{\partial E(a, b)}{\partial b} = -\frac{2}{n} \sum_{i=1}^n (y_i - ax_i - b) = 0$$

Despejamos a y b usando definiciones de varianzas y covarianzas.

$$\frac{1}{n} \sum_{i=1}^n (y_i x_i - ax_i^2) - b\bar{x} = s_{xy} + \bar{x}\bar{y} - a(s_x^2 + \bar{x}^2) - b\bar{x} = 0$$

$$\bar{y} - a\bar{x} - b = 0$$

de donde:

$$a = \frac{s_{xy}}{s_x^2}$$

$$b = \bar{y} - \frac{s_{xy}}{s_x^2} \bar{x}$$

Finalmente **la recta de regresión** de y sobre x es

$$y - \bar{y} = \frac{s_{xy}}{s_x^2} (x - \bar{x})$$

Varianzas de regresión

La **varianza residual** es el error cuadrático mínimo:

$$V_r = \min(E(a, b))$$

Desarrollando la formula obtenemos:

$$\begin{aligned} V_r &= \frac{1}{n} \sum_{i=1}^n \left[(y_i - \bar{y}) - \frac{s_{xy}}{s_x^2} (x_i - \bar{x}) \right]^2 \\ &= \frac{1}{n} \sum_{i=1}^n \left[(y_i - \bar{y})^2 - 2 \frac{s_{xy}}{s_x^2} (x_i - \bar{x})(y_i - \bar{y}) + \frac{s_{xy}^2}{s_x^4} (x_i - \bar{x})^2 \right] \\ &= s_y^2 - 2 \frac{s_{xy}^2}{s_x^2} + \frac{s_{xy}^2}{s_x^2} \end{aligned}$$

Por lo tanto la varianza residual es:

$$V_r = s_y^2(1 - r^2), \quad r = \frac{s_{xy}}{s_x s_y}$$

El cociente

$$r = \frac{s_{xy}}{s_x s_y}$$

es el coeficiente de la correlación lineal.

Si $r = \pm 1$ entonces $V_r = 0$ y las variables están perfectamente correlacionadas (es decir sabiendo una podemos calcular con exactitud la otra).

La **varianza total** de la muestra podemos descomponer:

$$V_t \equiv s_y^2 = V_e + V_r = s_y^2 r^2 + s_y^2 (1 - r^2)$$

El término $V_e = s_y^2 r^2$ define la **varianza explicada** (por la regresión).

Regresión no lineal

A partir de ahora vamos a suponer que la función de la regresión

$$y = f(x)$$

no es lineal.

Regresión exponencial

Dada una nube de puntos: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, ajustar el modelo

$$y = ae^{bx}$$

Método de linealización

Tenemos una función exponencial para ajustar:

$$y = be^{ax}$$

Aplicamos logaritmo neperiano:

$$\ln(y) = \ln(b) + ax$$

Sustituyendo

$$Y = \ln(y) \quad B = \ln(b)$$

obtenemos la función lineal:

$$Y = ax + B$$

Para encontrar a y B usamos el método de la regresión lineal.

Ejemplo de linealización

Ajustar $y = be^{ax}$ a los datos:

x_k	0	1	2	3	4
y_k	1.5	2.5	3.5	5	7.5

1. Transformaciones:

$$\ln(y) = \ln(b) + ax, \quad Y = \ln(y), \quad B = \ln(b) \Rightarrow Y = ax + B$$

2. Datos linealizados:

x_k	0	1	2	3	4
$Y = \ln(y_k)$	0,406	0,916	1,253	1,609	2,015

3. Calculamos los promedios y la recta de regresión

$$\bar{x} = 2 \quad \bar{Y} = 1,240 \quad s_x^2 = 2 \quad s_{xY} = 0,782$$

$$Y - 1,240 = \frac{0,782}{2}(x - 2)$$

4. Calculamos a y B ($Y = ax + B$)

$$a = \frac{0,782}{2} = 0,391 \quad B = 1,240 - 0,782 = 0,458$$

5. Calculamos b

$$b = e^B = \exp(0,459) \approx 1,581$$

Por tanto:

$$y = 1,581e^{0,391x}$$