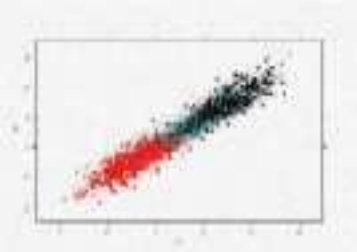
ESTADÍSTICA (SISTEMAS)

Profesores: Hilario Navarro. Jorge Martín



DEPARTAMENTO DE ESTADÍSTICA,
INVESTIGACIÓN OPERATIVA Y CÁLCULO
NUMÉRICO

Primera unidad didáctica.
Soluciones a los problemas propuestos de Estadística Descriptiva

Curso 2004-2005

Problema 1. En 1978, H. Cavendish realizó una serie de 29 experimentos con objeto de medir la densidad de la tierra. Sus resultados, tomando como unidad de densidad la del agua, fueron:

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.68	5.85	

Analizar descriptivamente estos datos.

Solución

Si optamos por un diagrama de tallos y hojas para tener una descripción gráfica de la distribución de frecuencias resulta

48	8
49	
50	7
51	0
52	6799
53	04469
54	2467
55	03578
56	12358
57	59
58	5

Respecto a la localización del centro de la distribución obtenemos:

- *Media:*
$$\bar{x} = \frac{1}{29}\sum_{i=1}^{29} x_i = 5.448$$
- *Mediana:* La posición central de las observaciones ordenadas está ocupada por la que toma el valor 5.46.

Para valorar la dispersión de la distribución elegimos la desviación típica. El cálculo de la varianza lo efectuamos mediante la fórmula

$$v_x = \frac{1}{29}\sum_{i=1}^{29} x_i^2 - \bar{x}^2$$

resultando $v_x = 0.049$.

Como consecuencia, el valor de la desviación típica es

$$\sqrt{0.049} = 0.221$$

□

Problema 2. La siguiente es la distribución del número de artículos defectuosos encontrados en 404 lotes de un producto manufacturado. Calcule la *media, mediana, varianza y desviación* típica para iniciar la descripción de los datos referidos.

Nº de items defectuosos	*Nº de lotes*
0	53
1	110
2	82
3	58
4	35
5	20
6	18
7	12
8	9
9	3
10	1
11	2
12	1

Solución

Para el cálculo de las medidas de centralización y dispersión pedidas, tenemos que tener en cuenta que partimos de la distribución de frecuencias. Concretamente, según la tabla del enunciado, el valor 0 se ha presentado en 53 ocasiones (frecuencia absoluta n_i), el 1 en 110, el 2 en 82 y así sucesivamente. Entonces, la media será el resultado del siguiente cálculo

$$\bar{x} = \frac{1}{404}\left[(0 \times 53) + (1 \times 110) + (2 \times 82) + (3 \times 58) + \cdots + (12 \times 1)\right]$$

es decir,

$$\bar{x} = \frac{1023}{404} = 2.532.$$

La mitad del número de observaciones es 202. Observando la tabla vemos que el primer valor que acumula una frecuencia igual o superior a este número es el 2 y, por tanto, es la mediana de esta distribución.

El hecho de que la media supere a la mediana nos indica que la cola derecha tiene una "extensión" mayor que la izquierda.

¿Cuánto se dispersan los datos en torno al centro de la distribución? Para responder a esta pregunta calcularemos la desviación típica. La media de las desviaciones cuadráticas (varianza) la obtendremos mediante la expresión

$$v_x = \frac{1}{404}\sum_i n_i x_i^2 - \bar{x}^2.$$

Dado que

$$\frac{1}{404}\left[(0^2 \times 53) + (1^2 \times 110) + (2^2 \times 82) + \cdots + (12 \times 1)\right] = \frac{4671}{404} = 11.562$$

y

$$\bar{x}^2 = 2.532^2 = 6.411$$

resulta que $v_x = 11.562 - 6.411 = 5.241$.

Como la desviación típica es la raíz cuadrada positiva de la varianza, nuestra medida de la dispersion es $\sqrt{5.241} = 2.289$.

□

Problema 3. Los datos siguientes muestran la asociación existente entre las medidas de dos aspectos interesantes de un fenómeno determinado. Calcule la recta de mínimos cuadrados y la varianza residual.

X	5	6	7	10	12	15	18	20
Y	7.4	9.3	10.6	15.4	18.1	22.2	24.1	24.8

Solución

La pendiente de la recta de mínimos cuadrados (recta de regresión) viene dada por la expresión

$$pend = \frac{cov_{x,y}}{v_x}$$

donde $cov_{x,y}$ es la covarianza entre las dos variables y v_x es la varianza de la variable X. Si para calcular la covarianza nos decidimos por la expresión

$$cov_{x,y} = \frac{1}{8}\sum_{i=1}^{8} x_i y_i - \bar{x}\bar{y},$$

necesitaremos los valores medios de ambas variables. Para los datos objeto de este estudio resulta:

- $\bar{x} = \dfrac{93}{8} = 11.625$
- $\bar{y} = \dfrac{131.9}{8} = 16.487$

Además,

- $\displaystyle\sum_{i=1}^{8} x_i y_i = [(5 \times 7.4) + (6 \times 9.3) + \cdots + (20 \times 24.8)] = 1801$

Como consecuencia,

$$cov_{x,y} = \frac{1801}{8} - (11.625 \times 16.487) = 33.464.$$

Ahora le toca el turno a la varianza de X. Para esta variable, la media de los cuadrados es

$$\frac{1}{8}(5^2 + 6^2 + 7^2 + \cdots + 20^2) = \frac{1303}{8} = 162.88.$$

Para obtener la varianza restamos el cuadrado de la media y resulta

$$v_x = 162.88 - 11.625^2 = 27.739.$$

Ya disponemos de todos los ingredientes para calcular la pendiente.

$$pend = \frac{cov_{x,y}}{v_x} = \frac{33.464}{27.739} = 1.206.$$

El valor positivo obtenido para la pendiente es compatible con la tendencia que se aprecia en los datos: "un aumento en el valor de X va acompañado por una variación del mismo tipo en el valor de Y".

La ordenada en el origen viene dada por

$$\bar{y} - (pend \times \bar{x}) = 16.487 - (1.206 \times 11.625) = 2.467.$$

Ya hemos calculado los elementos necesarios para expresar completamente la recta de regresión de Y sobre X:

$$Y = 2.467 + 1.206X.$$

¿Cómo describe esta línea la nube de puntos correspondiente? Una forma de valorar esta cuestión es medir la dispersión que presentan los puntos alrededor de la recta; y lo vamos a hacer a través de la varianza residual. Calcularemos en primer lugar el coeficiente de correlación lineal. Para ello necesitamos la varianza de la variable Y.

$$v_y = \frac{1}{8}(7.4^2 + 9.3^2 + 10.6^2 + \cdots + 24.8^2) - 16.487^2 = \frac{2507.07}{8} - 16.487^2 = 41.563.$$

Entonces,

$$r = \frac{cov_{x,y}}{\sqrt{v_x v_y}} = \frac{33.464}{\sqrt{27.739 \times 41.563}} = 0.986.$$

Comúnmente, un valor tan próximo a 1 se interpreta como que "la recta ajusta bastante bien a la nube de puntos". Una razón para esta afirmación se encuentra en la relación

$$\frac{varianza\ residual}{v_y} = 1 - r^2$$

que nos expresa que "a mayor coeficiente de correlación, menor dispersión relativa en torno a la recta de mínimos cuadrados".

En nuestro caso, el valor de dicha dispersión relativa es

$$\frac{varianza\ residual}{v_y} = 1 - 0.986^2 = 0.028$$

lo que supone una varianza residual de

$$varianza\ residual = 0.028 \times 41.563 = 1.164.$$

□

Problema 4. Se dispone de los siguientes datos referentes a 14 observaciones del par (X, Y):

$$\sum x_i = 517 \qquad \sum y_i = 346$$

$$\tfrac{1}{14}\sum x_i^2 = 2792.5 \qquad \tfrac{1}{14}\sum y_i^2 = 1246.7$$

$$\tfrac{1}{14}\sum x_i y_i = 1844.6$$

Se pide:

(a) Calcular la pendiente de la recta de regresión de Y sobre X.

(b) Obtener una medida del ajuste de dicha recta a la nube de puntos.

Datos auxiliares: Coeficiente de correlación $r = 0.98$.

Solución

(a) La pendiente de la recta de regresión de Y sobre X viene dada por el cociente

$$\frac{cov_{x,y}}{v_x} = \frac{\frac{1}{n}\sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2}$$

ó, alternativamente,

$$\frac{cov_{x,y}}{v_x} = r\sqrt{\frac{v_y}{v_x}}.$$

Sustituyendo en la primera expresión queda

$$\frac{1844.6 - \left(\frac{517}{14}\right)\left(\frac{346}{14}\right)}{2792.5 - \left(\frac{517}{14}\right)^2}.$$

(b) La varianza residual se define como el error cuadrático medio cometido con la recta de regresión de Y sobre X. Entonces, su valor nos dará una medida del ajuste de dicha recta a la nube de puntos. Con los datos del enunciado obtenemos

$$v_y = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 = 1246.7 - \left(\frac{346}{14}\right)^2 = 635.9$$

y $varianza\ residual = 635.9\left(1 - 0.98^2\right) = 25.18$.

Sin embargo, si queremos una medida del error con mayor capacidad de interpretación, deberemos calcular el valor relativo dado por el cociente

$$\frac{varianza\ residual}{v_y} = 1 - r^2$$

ó, equivalentemente, tomar el coeficiente r^2 como una medida del grado de ajuste: un valor próximo a 1 reflejará un buen ajuste y un valor cercano al 0 indicará la cualidad contraria. En definitiva, podríamos calcular directamente

$$r^2 = 0.96$$

concluyendo que, en este caso, el ajuste es bastante bueno.

□

Problema 5. A partir de una muestra de 26 observaciones de la variable X —que toma valores entre 320 y 430—, se obtuvo el siguiente diagrama de tallos y hojas:

32	55
33	49
34	
35	6699
36	34469
37	03345
38	9
39	2347
40	23
41	
42	4

(a) Reproduzca las 10 primeras observaciones (en la ordenación de menor a mayor).

(b) ¿Dónde está situada la mediana de la distribución? ¿Qué variación experimentaría dicha medida de centralización si el máximo de la distribución aumentara su valor en 10 unidades?

(c) Sabiendo que el valor medio es 370.7, ¿cómo mediría la dispersión de los datos respecto a este valor central? *(No se requiere realizar los cálculos)*

Solución

(a) Las observaciones pedidas son

$$325, 325, 334, 339, 356, 356, 359, 359, 363, 364$$

(b) La mediana de la distribución está situada en el punto

$$\frac{369 + 370}{2} = 369.5$$

Si el máximo de la distribución, que es 424, aumentara su valor en 10 unidades, la mediana estaría situada en el mismo punto —en 369.5—, ya que seguiríamos teniendo el mismo número de observaciones a cada lado.

(c) Mediante la desviación típica, que se define como la raíz cuadrada positiva de la varianza. Para el cálculo de esta última, se puede aplicar directamente la definición:

$$v_x = \frac{1}{n}\sum_i (x_i - \bar{x})^2$$

ó, equivalentemente,

$$\begin{aligned} v_x &= \frac{1}{n}\sum_i x_i^2 - \bar{x}^2 \\ &= \frac{1}{26}\left(325^2 + 325^2 + 334^2 + ... + 424^2\right) - 370.7^2 \end{aligned}$$

□

Problema 6. Para cada una de las condiciones que se indican a continuación, represente una nube de puntos (X, Y) que sea compatible con ella:

- Covarianza negativa.
- Pendiente de la recta de regresión de Y sobre X positiva.
- Correlación próxima a 1.
- Correlación nula.

(Nota: Justifique brevemente cada representación)

Solución

- *Breve justificación*: La relación entre **covarianza**, **coeficiente de correlación lineal** y **pendiente de la recta de regresión** de Y sobre X se pone de manifiesto en las siguientes igualdades:

$$Pendiente = \frac{cov_{x,y}}{v_x} = r\sqrt{\frac{v_y}{v_x}}$$

Por tanto, dichos elementos tendrán siempre el mismo signo (gráficas de la figura 1).

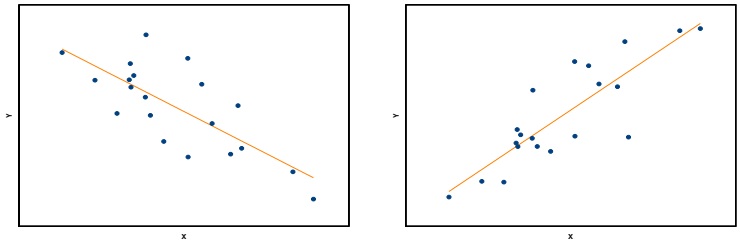


Figura 1: Covarianza negativa. Pendiente positiva

Además, el valor del **coeficiente de correlación** se refleja en la **varianza residual** según indica la siguiente expresión:

$$varianza\ residual = v_y\left(1 - r^2\right) \ .$$

Así, el caso de correlación próxima a 1 se corresponde con un valor pequeño para el cociente $\dfrac{varianza\ residual}{v_y}$ (gráfica izquierda figura 2), mientras que un coeficiente de correlación nulo supone una varianza residual cercana a su valor máximo, que es v_y (gráfica derecha figura 2).

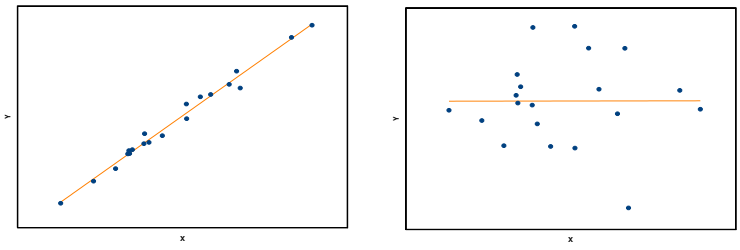


Figura 2: Correlación próxima a 1. Correlación nula

□