Índice general

Estadística Descriptiva

1.1. Preliminares

En muchos problemas reales nos enfrentamos al estudio de fenómenos que, bajo unas condiciones determinadas, pueden presentar valores/estados diferentes, lo que puede dar lugar a grandes volúmenes de información —en muchos casos numérica— y estructuras asociativas complejas. En la actualidad, la facilidad para la obtención, almacenamiento y transmisión de dicha información es enorme —no se puede decir máxima— y la tarea de generar una base de datos que contenga los resultados de la observación directa o de la experimentación se ha simplificado hasta tal punto que el protagonismo en el análisis de datos ha pasado a los métodos. En este momento, en el que se puede observar y registrar "casi todo", la Estadística —tanto Descriptiva como Inferencial— se ha convertido en una herramienta indispensable para la extracción de conocimiento a partir de los datos.

Las situaciones más sencillas se presentan cuando nuestro interés se centra en uno o dos aspectos —variables— del fenómeno objeto del estudio. Si, además, dichas variables tienen carácter cuantitativo, la "oferta" de procedimientos para el análisis se amplia notablemente.

En el caso univariante una sencilla tabla de frecuencias, o cualquiera de sus versiones gráficas: diagramas de tallos y hojas, histograma, gráficos de cajas, etc., es suficiente para poner en evidencia los atributos fundamentales de la distribución —en la actualidad, los métodos gráficos están adquiriendo un protagonismo notable gracias al extraordinario avance de las tecnologías de la información y de la comunicación—. Normalmente no manejaremos la distribución de frecuencias "completa" a la hora de tomar decisiones que involucren a nuestro fenómeno. Como mucho, utilizaremos una serie de medidas que valoren aspectos relevantes de la misma tales como la localización (centralización) y la dispersión respecto al centro. En uno y otro caso, se pueden emplear distintos procedimientos que dependen de la interpretación que se dé al concepto. Lo usual es situar el centro de la distribución en el valor proporcionado por la media, la mediana o la moda y medir la dispersión mediante la desviación típica —raíz cuadrada positiva de la varianza —. La descripción resulta especialmente simple cuando la distribución es unimodal —sólo una moda— y simétrica.

Cuando el problema es bivariante el diagrama de dispersión y la recta de regresión, junto con la varianza residual, constituyen la base de cualquier estudio descriptivo. El coeficiente de correlación lineal también juega un papel relevante en la fase inicial de un análisis estadístico bidimensional.

1.2. Problemas resueltos

La siguiente colección de problemas muestra la manera en que se resuelven algunos casos prácticos de estadística descriptiva.

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

$$\frac{1}{404}\left[(0^2 \times 53) + (1^2 \times 110) + (2^2 \times 82) + \cdots + (12 \times 1)\right] = \frac{4561}{404} = 11.29$$

y

$$\bar{x}^2 = 2.532^2 = 6.412$$

resulta que $v_x = 11.29 - 6.411 = 4.878$.

Como la desviación típica es la raíz cuadrada positiva de la varianza, nuestra medida de la dispersion es $\sqrt{4.878} = 2.209$.

Problema 3. 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?

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:

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

o, 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 4. Se dispone del siguiente conjunto de datos:

$$25, 32, 37, 45, 48, 32, 40, 45, 26, 33, 15, 47, 35, 52, 55, 46, 35, 34, 60, 20, 18$$

(a) Representar, mediante un diagrama de tallos y hojas, el conjunto de datos. Tome para el tallo el dígito de las decenas y para la hoja el de las unidades.

(b) Calcule la media, la varianza y la desviación estándar de las observaciones.

Solución

(a) Ordenamos las observaciones de menor a mayor. El conjunto de datos queda así:

$$15, 18, 20, 25, 26, 32, 32, 33, 34, 35, 35, 37, 40, 45, 45, 46, 47, 48, 52, 55, 60$$

Con esta ordenación se obtiene de forma inmediata el siguiente diagrama de tallos y hojas

1	58
2	056
3	2234557
4	055678
5	25
6	0

(b) El conjunto de datos está compuesto por 21 observaciones. La media de las observaciones es

$$\bar{x} = \frac{\sum_{i=1}^{21} x_i}{21} = \frac{780}{21} \approx 37.14$$

Para calcular la varianza, se utilizará la versión de la varianza insesgada:

$$s^2 = \frac{S_{xx}}{n-1} \quad \text{donde} \quad S_{xx} = \sum_{i=1}^{21}(x_i - \bar{x})^2 = \sum_{i=1}^{21} x_i^2 - \frac{(\sum_{i=1}^{21} x_i)^2}{n}.$$

De los datos, se obtiene que $\sum_{i=1}^{21} x_i^2 = 31970$. Por tanto,

$$s^2 = \frac{S_{xx}}{20} = \frac{1}{20}\left(31970 - \frac{(780)^2}{21}\right) \approx 149.93.$$

Finalmente, la desviación estándar es $s = \sqrt{149.93} \approx 12.24$.

Problema 5. Considere un conjunto de datos con las siguientes observaciones:

$$x_1 = 10, x_2 = 15, x_3 = 11, x_4 = 17, x_5 = 20, x_6 = 18, x_7 = 14, x_8 = 13, x_9 = 17$$

(a) Calcule la media y la mediana de las observaciones.

(b) Si se reemplaza el dato $x_5 = 20$ por $x_5 = 65$, ¿cuánto valdría ahora la media y la mediana? ¿Cómo interpreta el resultado obtenido?

Solución

(a) La media de las observaciones es

$$\bar{x} = \frac{\sum_{i=1}^{9} x_i}{9} = \frac{135}{9} = 15.$$

Ahora ordenamos las observaciones de menor a mayor y resulta el siguiente conjunto de datos: