Actividad complementaria - Medidas para describir conjuntos de datos

Nayda Nataly Hernández Moreno

CORPORACIÓN UNIVERSITARIA IBEROAMERICANA

FACULTAD DE INGENIERIA (INGENIERIA EN CIENCIA DE DATOS VIR

(ID: 100193702)

Julian Ricardo López Hernández

Bogotá D.C. Colombia

02 de octubre de 2024

Medidas para describir conjuntos de datos

1. Media: Es la medida más conocida y útil de tendencia central, se usa para calcular un valor representativo de los números que se están promediando.

La media $\overline{x}$ de las observaciones $x_1 + x_2, \ \ldots, \ x_n$ está dada por:

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Ejemplo: la media de 10, 20, 20, 30, 30, 30, 40, 50, 50, 60 es

$$\overline{x} = \frac{10 + 20 + \cdots + 50 + 60}{10} = \frac{340}{10} = 34$$

2. Mediana: La palabra mediana es sinónimo de “medio” y la media muestral es en realidad el valor medio una vez se ordenan las observaciones de la más pequeña a la más grande Cuando las observaciones están denotadas por $x_1, \ldots., x_n$se utiliza el símbolo $\tilde{x}$ para representar la mediana muestral.

$$\tilde{x} = \begin{cases} el\ valor\ medio\ unico\ si\ n\ es\ impar; = \left(\frac{n+1}{2}\right)^{\acute{e}simo}\ valor\ ordenado. \\ promedio\ de\ los\ valores\ medios\ si\ \ n\ es\ par; \\ = promedio\ de \left(\frac{n+1}{2}\right)^{\acute{e}simo}\ y\ \left(\frac{n}{2}+1\right)^{\acute{e}simo}\ valores\ ordenados \end{cases}$$

Ejemplo: la mediana de 10, 20, 20, 30, 30, 30, 40, 50, 50, 60 es:

$$\frac{30+30}{2} = \frac{60}{2} = 30$$

1. ***Moda:*** La moda de una distribución se define como el valor de la variable que más se repite en un polígono de frecuencia la moda corresponde al valor de la variable que está bajo el punto más alto de la gráfica una muestra puede tener más de una moda. La moda $\hat{x}$, **para datos agrupados**, está definida por

$$\hat{x} = Li + \left(\frac{\triangle_1}{\triangle_1+\triangle_2}\right) I$$

Li = *frontera clase inferior*

$\triangle_1$ = *frecuencia absoluta mayor menos frecuencia absoluta anterior*

$\triangle_2$ = *frecuencia absoluta mayor menos frecuencia absoluta posterior*

Cuando los datos no están agrupados en intervalos, la moda es simplemente el valor que aparece con más frecuencia en el conjunto de datos. Para calcular la moda en datos no agrupados, se listan todos los valores del conjunto de datos, se cuenta la frecuencia de cada valor (cuántas veces aparece) y el valor con mayor frecuencia es la moda.

Ejemplo: La Moda de 10, 20, 20, 30, 30, 30, 40, 50, 50, 60 es 30 ya que se repite con mayor frecuencia

x_i	f_i
10	1
20	2
30	***3***
40	1
50	2
60	1

2. ***Varianza:*** Esta medida nos permite determinar el promedio aritmético de fluctuación de los datos respecto a su punto central o media. La desviación estándar nos da como resultado un valor numérico que representa el promedio de diferencia que hay entre los datos y la media.

   Para calcular la desviación estándar o típica se necesita calcular primero la varianza que es la sumatoria de diferencia de cada clase menos la media aritmética al cuadrado por su frecuencia por intervalo; esta sumatoria se divide entre el total de los datos.

   Suma de desviaciones = $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

   La varianza muestral, denotada por s^2 está dada por

$$s^2 = \frac{\sum (x_{i-\bar{x}})^2}{n-1} = \frac{S_{xx}}{n-1}$$

Ejemplo: Para calcular la varianza de 10, 20, 20, 30, 30, 30, 40, 50, 50, 60, primero se calcula su media:

$$10 + 20 + 20 + 30 + 30 + 30 + 40 + 50 + 50 + 60 = 340$$

$$n = 10 \; por \; lo \; que \; la \; media \; es:$$

$$\bar{x} = \frac{340}{10} = 34$$

Ahora, se aplica $s^2 = \frac{\sum (x_{i-\bar{x}})^2}{n-1} = \frac{S_{xx}}{n-1}$

x	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
10	10 - 34=-24	576
20	20 - 34= -14	196
20	20 – 34=-14	196
30	30 - 34= -4	16
30	30 - 34 = -4	16
30	30 - 34 =-4	16
40	40 - 34 = 6	36
50	50 - 34 =16	256
50	50 - 34 = 16	256
60	60 - 34 = 26	676
	Total	22405

Se suman todos los valores de las desviaciones cuadradas:

$$576 + 196 + 196 + 16 + 16 + 16 + 36 + 256 + 256 + 676 = 2240$$

Y se calcula la varianza:

$$s^2 = \frac{2240}{n-1} = \frac{2240}{10-1} = 248.89$$

Por lo tanto, la varianza muestral es:

$$s^2 = 248.89.$$

3. ***Desviación Estándar muestral,*** denotada por s , es la raíz cuadrada (positiva) de la varianza:

$$s = \sqrt{s^2}$$

$$s = \sqrt{248.89} = 15.78$$

4. ***Coeficiente de variación:*** El coeficiente de variación es una medida de la dispersión relativa de una variable estadística, calculada como la razón de la desviación estándar a la media, y se expresa como un porcentaje.

La fórmula para calcular el coeficiente de variación (CV) es:

$$CV = \frac{\sigma}{\mu}.100$$

$\sigma\ es\ la\ desviación\ estándar\ de\ los\ datos.$

$\mu\ es\ la\ media\ de\ los\ datos.$

Se calcula la media (μ) sumando todos los valores y dividiendo entre el número total de observaciones

$$\sigma = 15.78$$

$$\mu = 34$$

$$CV = \left(\frac{15.78}{34}\right) \cdot 100 = 46.4\%$$

5. ***Normalización Z:*** Es la distribución normal más importante en toda la probabilidad y estadística.

   En la cual a partir de un punto central de máxima frecuencia y es utilizada para cuando es imposible

   Si X tiene una distribución normal con media μ desviación estándar σ entonces:

$$Z = \frac{X - \mu}{\sigma}$$

Usamos los valores 10,20,20,30,30,30,40,50,50,60

x	z
10	$(10\text{-}34)/14.97 \approx -1.60$
20	$(20\text{-}34)/14.97 \approx -0.93$
20	≈ -0.93
30	$(30-34)/14.97 \approx -0.27$
30	≈ -0.27
30	≈ -0.27
40	$(40-34)/14.97 \approx 0.40$
50	$(50-34)/14.97 \approx 1.07$
50	≈ 1.07
60	$(60-34)/14.97 \approx 1.74$

Referencias:

Islas Salomón, C. A., Colín Uribe, M. P., & Morales Téllez, F. (2020). *Probabilidad y estadística*. Grupo Editorial Éxodo

Islas Salomón, C. A., Colín Uribe, M. P., & Morales Téllez, F. (2020). *Probabilidad y estadística*. Grupo Editorial Éxodo