Análisis e Interpretación de Datos

MÁSTER UNIVERSITARIO EN ANÁLISIS Y VISUALIZACIÓN DE DATOS MASIVOS / VISUAL ANALYTICS AND BIG DATA

Miller Janny Ariza Garzón

Tema 3. Medidas que resumen la información

unir LA UNIVERSIDAD EN INTERNET

Universidad Internacional de La Rioja

Tabla de contenido

- Tema 3: Medidas que resumen la información.
  - Medidas de tendencia central.
  - Medidas de tendencia central robustas.
  - Medidas de dispersión.
  - Medidas de dispersión robustas.
  - Medidas de posición y forma.
  - Gráficos de caja. Datos atípicos y análisis exploratorio de datos.

Tabla de contenido

Medidas resumen de la información

Tendencia central

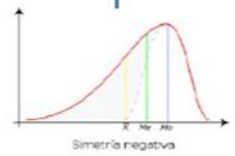
- Mediana = $x_{central}$
- Media
  - Media aritmética: $\frac{\sum x_i}{n}$
  - Media ponderada: $\frac{\sum w_i x_i}{\sum w_i}$
- Moda= n_{max}

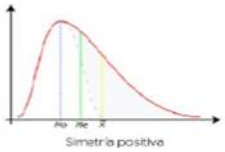
Dispersión

- Varianza: $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$
- Rango = $x_{max} - x_{min}$
- Coeficiente de variación: $cv = \frac{s}{\overline{x}} \times 100\%$
- Desviación típica: $s = \sqrt{s^2}$

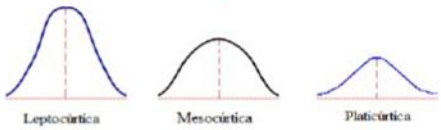
Posición y forma

- Asimetría



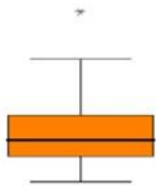


- Cuartiles y percentiles: Q_1, Q_2, Q_3 ; $P_1 \dots P_{99}$
- Tipificación: $z_i = \frac{x_i - \overline{x}}{s}$
- Apuntamiento



Valores atípicos

Caja y bigotes



Medidas que resumen la información

Se analiza la MUESTRA-POBLACIÓN y el CONTEXTO para decidir qué cantidades tienen sentido para caracterizar numéricamente el problema estadístico

Media

- Media aritmética: Es la media natural, la más usada y con más propiedades estadísticas. Le afectan mucho los valores atípicos. Describe situación de baja dispersión

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

- Media ponderada: Media teniendo en cuenta un ponderador w.

$$\frac{\sum_{i=1}^{n} x_i\, w_i}{\sum_{i=1}^{n} w_i}$$

- Media armónica: Usado para promediar tasas o cocientes.

$$\frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^{-1}\right)^{-1} \; ; \; x_i \neq 0 \; Para \; i = 1, \dots, n$$

- Media geométrica: Usado para promediar variaciones porcentuales (tasa de crecimiento porcentual promedio).

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \frac{1}{n}\sum_{i=1}^{n} \ln(x_i) \; ; \; x_i \geq 0 \; Para \; i = 1, \dots, n$$

Media armónica≤ Media geométrica ≤ Media aritmética

- Media cuadrática: es muy útil para calcular la **media** de variables **que** toman valores negativos y positivos. Ej, para calcular la **media** de errores de medida.

- Media truncada: Media aritmética sin valores extremos.
- Media winsorizada: Media aritmética imputando valores extremos.

Medias robustas

Ejemplos:

Ejemplo1: Promedio del costo de trabajo por hora para cada uno de los productos.

Nivel de mano de obra	Salario por hora en Euros	Horas de mano de obra por unidad	
		Producto 1	Producto 2
No calificado	5.00	1	4
Semicalificado	7.00	2	3
Calificado	9.00	5	3

Media ponderada.

Ejemplo2: Imaginemos una etapa ciclista cuyo trayecto tiene dos tramos muy distintos, de la misma longitud. En la primera mitad la carretera es estrecha y sinuosa, en continua pendiente y culmina en un puerto de primera categoría. Los organizadores estiman que en este primer trayecto la velocidad media será de **15 km por hora**. La segunda parte del trayecto es todo lo contrario: el firme es ancho y en ligero descenso, y tiene grandes rectas, lo que lo hace apto para que los corredores aceleren. Los organizadores calculan que la velocidad en ese tramo será de **30 km por hora**.

¿A qué velocidad media suponen los organizadores que rodará el pelotón?

Media armónica.

Ejemplos:

Ejemplo3. Si el crecimiento de las ventas en un negocio fue en los tres último años de 26%, 32% y 28%, hallar la media anual del crecimiento (discreto).

$$1+\mathrm{R_G}=\left[(1+\mathrm{R_1})(1+\mathrm{R_2})(1+\mathrm{R_3})...(1+\mathrm{R_T})\right]^{\frac{1}{\mathrm{T}}}$$

$$\mathrm{R_G}=\left[\prod_{t=1}^{T}(1+\mathrm{R_t})\right]^{\frac{1}{\mathrm{T}}}-1$$

Variable para la **media geométrica** es $(1+R_t)$.

$$R_G=\sqrt[3]{(1.26)*(1.32)*(1.28)}-1$$

$$R_G=0.286$$

Means in R:

```
data <- c(x1,x2,…,xn)
```

- Media aritmética: `m<-mean(data)`
- Media ponderada: `w<-c(w1,w2,…,wn)`
  `mp<- weighted.mean(data, w)`
- Media armónica: `ma<-1/mean(1/data)`
- Media geométrica: `mg<-exp(mean(log(data)))`
- Media cuadrática: `mc<-sqrt(sum(data^2)/length(data))`
- Media recortada: `mean(data, trim=porcentaje_recorte en cada extremo/100)`
- Media winsorizada: `library(psych)`
  `wm<-winsor.mean(data, trim = porcentaje/100)`

Medidas de dispersión:

La dispersión se refiere a la separación de los datos en una distribución, en comparación con una medida de tendencia central.

¿Por qué es importante medir la variabilidad?

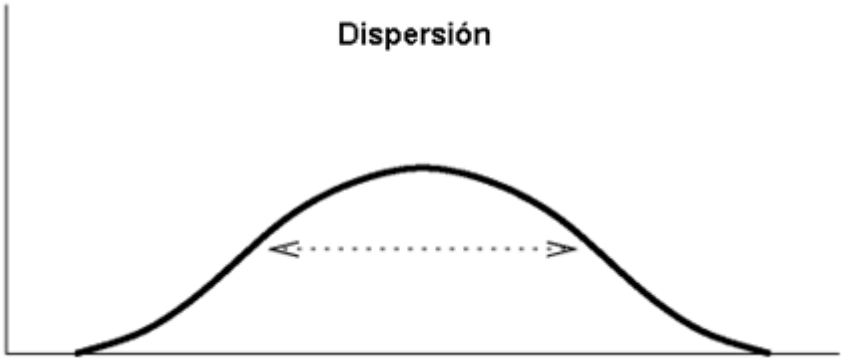
- Nos proporciona información para juzgar la calidad de la medida de tendencia central.
- Para medir la dispersión de los datos.
- Permite comparar la dispersión de diferentes conjuntos.
- Es una medida de riesgo

Medidas de dispersión:

Varianza	$s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1)$ ó $s^2 = \dfrac{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$
Desviación Estándar	$s = \sqrt{s^2} = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1)}$ ó $s = \sqrt{s^2} = \sqrt{\dfrac{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}}$

Relacionado con el segundo momento de una variable. K=2, momen

$$\mu_k = E[(X - \mu)^k]$$



Otra medida: **Rango**

Medidas de dispersión:

- **Coeficiente de variación:** medida relativa de variabilidad, medida en términos porcentuales.

$$CVE = \frac{S}{\bar{X}} x100\%, \qquad CV = \frac{\sigma}{\mu} x100\%$$

Expresa la desviación típica en porcentaje de la media. Proporciona una estimación de la magnitud de la desviación respecto a la magnitud de la media.

  - 0% - 10% Muy homogéneo
  - 10.1% - 20% Homogéneo
  - 20% - 50% Heterogéneo
  - Mayor que 50% Muy heterogéneo.

Medidas de dispersión:

In R: data <- c(x1,x2,…,xn)

- Desviación estándar: sd(data)
- Varianza: var(data)
- Coeficiente de variación: cve<-sd(data)/mean(data)

- Desviación estándar truncada:
  library(chemometrics)
  sd_trim(data,trim=p/100)
- Desviación estándar winsorizada:
  winsor.sd(data, trim = p/100)

Próxima sesión

- Tema 3: Medidas que resumen la información.

  - Medidas de posición y forma.
  - Gráficos de caja.
  - Datos atípicos y análisis exploratorio de datos

www.unir.net