Curso Análisis de datos

19 de julio de 2022

Análisis de datos multivariantes

Módulo 4
—
Análisis de datos multivariantes

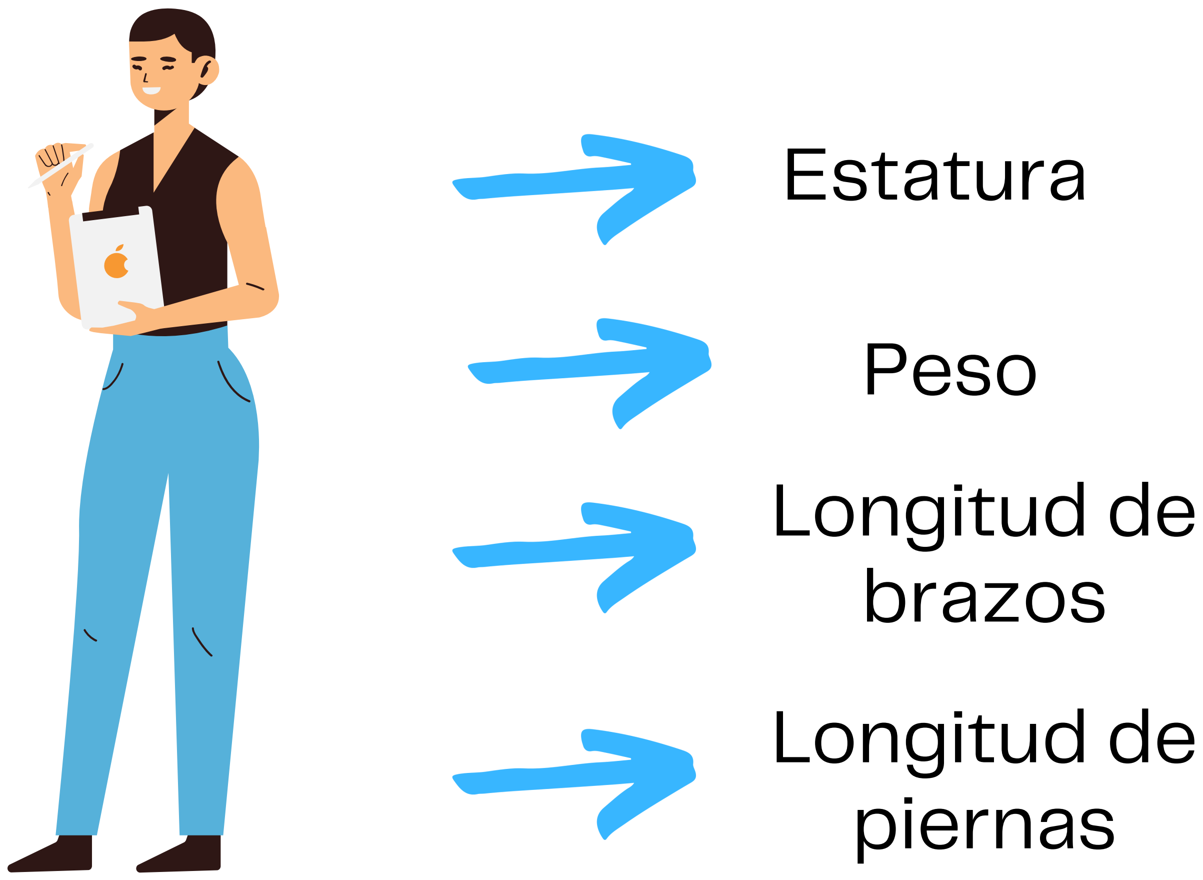
https://www.researchgate.net/publication/40944325_Analisis_de_Datos_Multivariantes

Agenda de hoy

1. Datos multivariantes
2. Descripción de datos multivariantes
3. Medidas de centralización
4. Matriz de varianzas y covarianzas
5. Medidas de variabilidad
6. Medidas de dependencia lineal

Datos multivariantes

Cuando queremos describir las características físicas de una persona, el rendimiento de un proceso, las características del comprador de un producto, entre otros, se requiere tener en cuenta varias variables de forma simultánea



Análisis de datos multivariantes

El análisis de datos multivariantes tiene como objetivo el estudio estadístico de múltiples variables medidas en una población

Resumir el conjunto de variables en unas pocas nuevas variables, transformando las originales

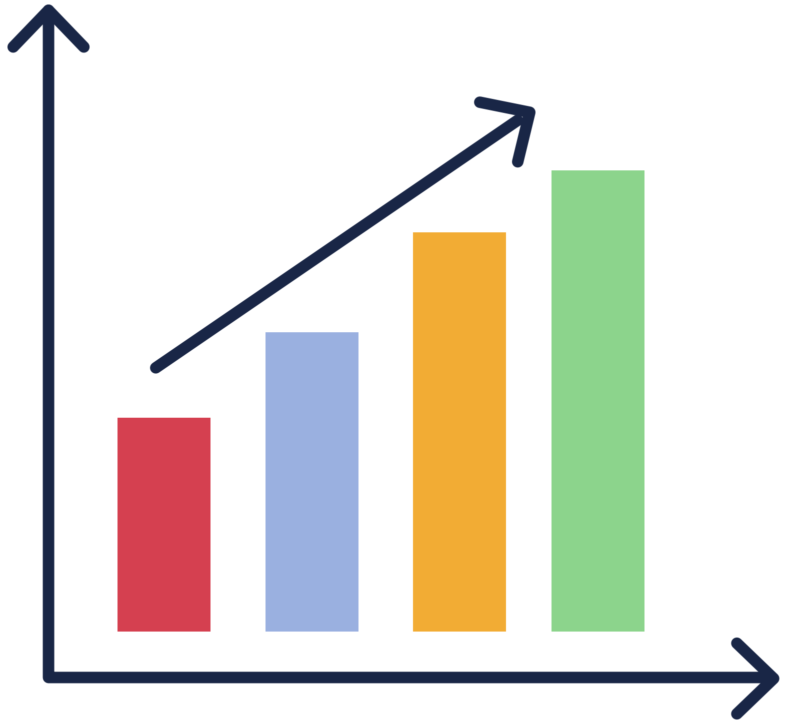
Encontrar grupos en los datos si existen

Clasificar nuevas observaciones en grupos definidos

Relacionar dos conjuntos de variables

Descripción de datos multivariantes

El primer paso en el análisis de datos multivariantes es el análisis descriptivo de los datos que nos permite comprender su estructura y extraer información relevante. El objetivo es estudiar cada variable aisladamente y además las relaciones entre ellas.



Tipos de variables

Cuantitativas

Su valor se expresa numéricamente
(edad, estatura, salario)

Cualitativas

Su valor es un atributo o categoría
(género, color de ojos, estrato)

Tipos de variables

Se pueden codificar numéricamente

Cuantitativas

Continuas

Toman cualquier valor real (estatura)

Discretas

Solo toman valores enteros (número de hijos)

Cualitativas

Binarias

Toman dos valores posibles (género)

Generales

Toman muchos valores posibles (nacionalidad)

La matriz de datos

Supongamos que observamos p variables numéricas en un conjunto de n elementos o individuos

Cada una de las variables se denomina variable **escalar o univariante**

El conjunto de todas las variables forman una variable **vectorial o multivariante**

La matriz de datos

La matriz de datos $\boldsymbol{X}$ está conformada por los valores de las $\boldsymbol{p}$ variables escalares en cada uno de los n elementos

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & \dots & \dots & x_{2p} \\ \vdots & & & \\ x_{n1} & \dots & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}$$

La matriz de datos

En 100 estudiantes de una universidad medimos la edad, el género (1 mujer, 0 hombre), el promedio, el municipio de residencia y el estrato (1, 2, 3, 4, 5, 6).

En este caso, la matriz de datos tendrá 100 filas y 5 columnas. De las 5 variables, 2 son cuantitativas, una es binaria (género) y 2 cualitativas generales.

Análisis univariante

El estudio univariante de una variable cuantitativa $\boldsymbol{x_j}$ implica calcular:

Media

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

En el caso de una variable binaria, la media es igual a la proporción de unos en los datos

Desviación estándar

$$s_j = \sqrt{\frac{\sum_{i=1}^{n} d_{ij}}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}{n}}$$

Análisis univariante

En el caso de una variable binaria,

Media $p = proporción\ de\ unos\ en\ los\ datos$

Análisis univariante

El estudio univariante de una variable cuantitativa $\boldsymbol{x_j}$ implica calcular:

Coeficiente de variación

$$CV_j = \sqrt{\frac{s_j^2}{\overline{x}_j^2}}$$

Útil para comparar la variabilidad de distintas variables cuantitativas ya que no depende de las unidades de medida

Análisis univariante

El estudio univariante de una variable cuantitativa $\boldsymbol{x_j}$ implica calcular:

<u>Coeficiente de asimetría</u>

$$A_j = \frac{1}{n} \frac{\sum \left(x_{ij} - \overline{x}_j\right)^3}{s_j^3}.$$

Mide la simetría de los datos respecto a su centro

Análisis univariante

$$A_j = \frac{1}{n} \frac{\sum (x_{ij} - \overline{x}_j)^3}{s_j^3}.$$

$$A_j < 0$$

Distribución asimétrica negativa: existe mayor concentración de valores a la izquierda de la media que a su derecha

$$A_j = 0$$

Distribución simétrica: existe la misma concentración de valores a la derecha y a la izquierda de la media

$$A_j > 0$$

Distribución asimétrica positiva: existe mayor concentración de valores a la derecha de la media que a su izquierda

Análisis univariante

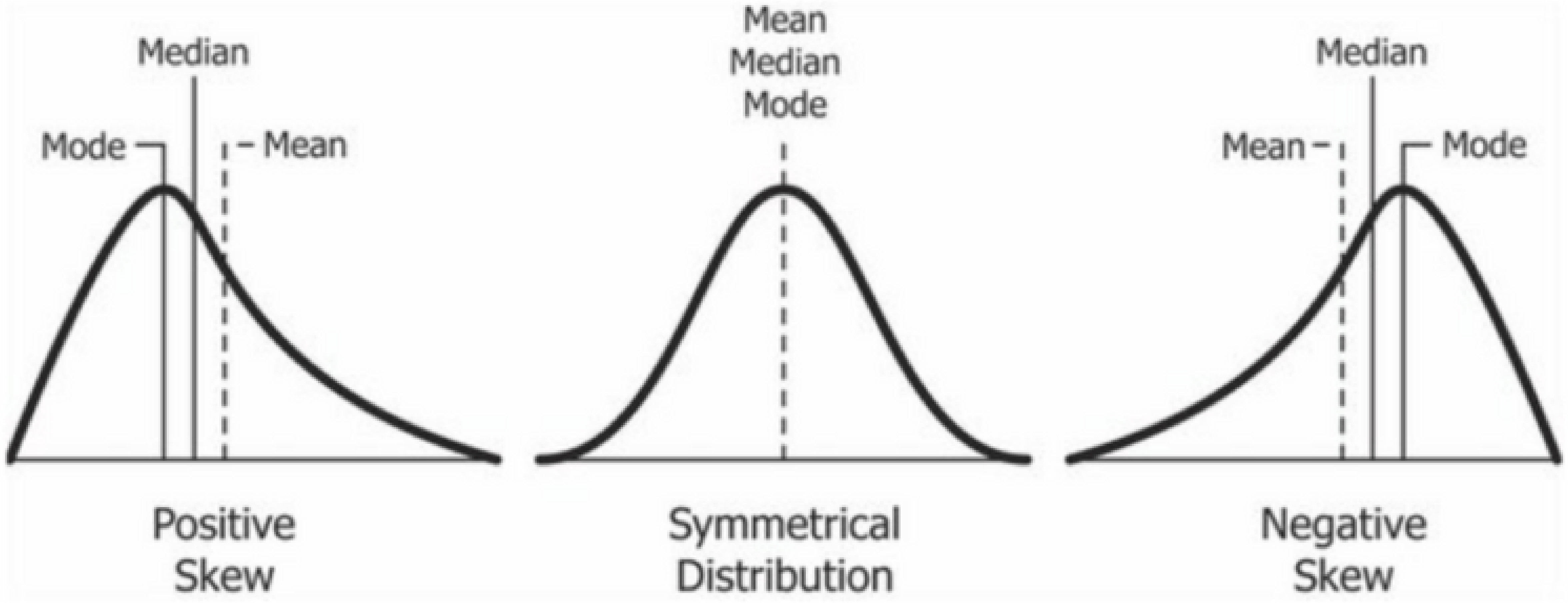


Imagen tomada de: https://machinelearningparatodos.com/como-saber-si-una-variable-sigue-una-distribucion-normal-en-python/

Análisis univariante

El estudio univariante de una variable cuantitativa x_j implica calcular:

Coeficiente de Curtosis

$$K_j = \frac{1}{n}\frac{\sum (x_{ij} - \overline{x}_j)^4}{s_j^4}$$

Mide la concentración de los valores de una variable en torno a su media

Análisis univariante

Generalmente, la curtosis se expresa como **exceso de curtosis**, es decir, se compara con respecto a la distribución normal, la cual tiene una curtosis de igual a 3.

Exceso de curtosis

$$K_j - 3$$

$$K_j - 3 < 0$$

Distribución platicúrtica: los valores se concentran poco entorno a su media.

$$K_j - 3 = 0$$

Distribución mesocúrtica: distribución normal.

$$K_j - 3 > 0$$

Distribución leptocúrtica: los valores se concentran mucho entorno a su media.

Análisis univariante

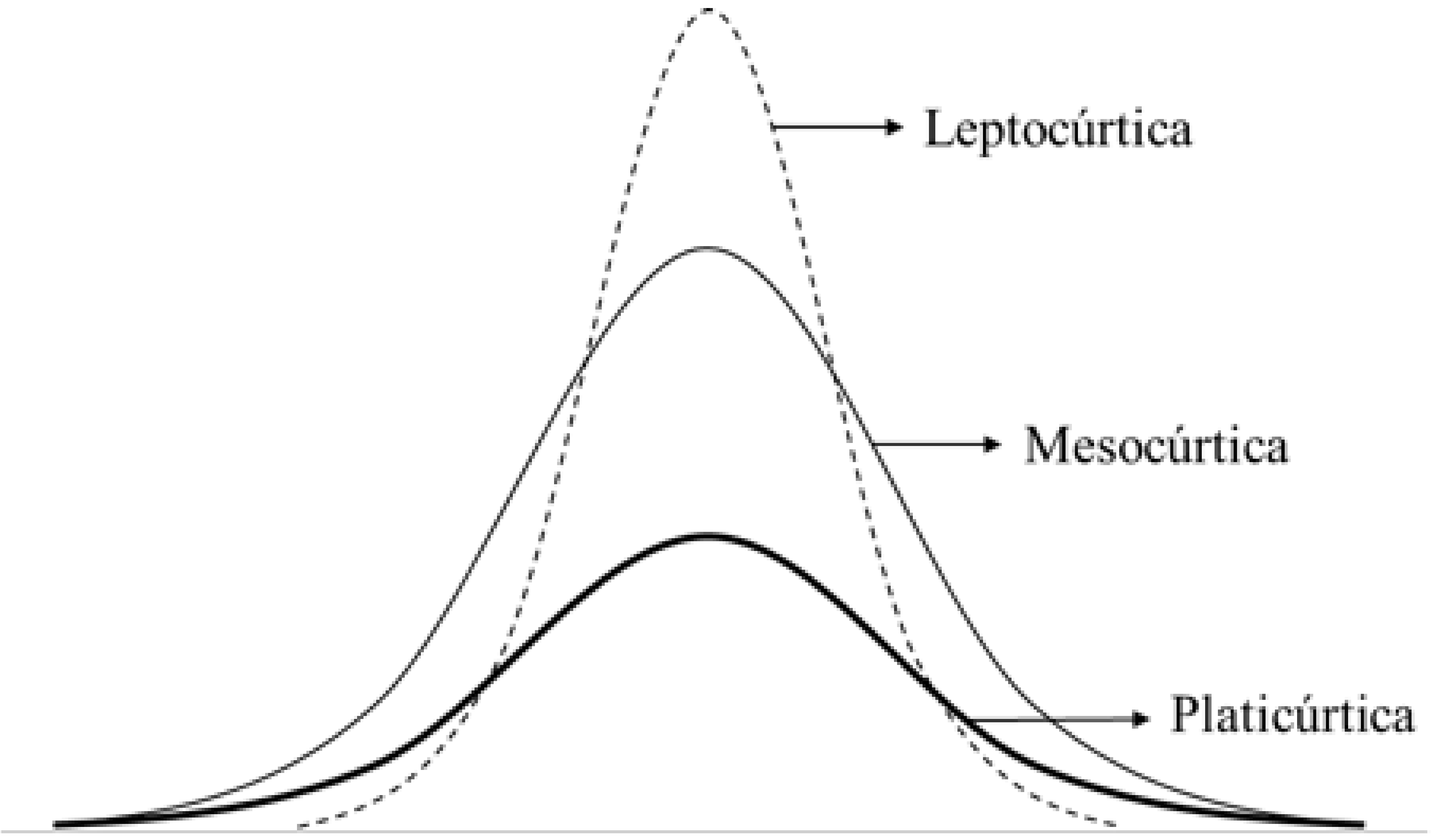


Imagen tomada de: https://revistas.tec.ac.cr/index.php/eagronegocios/article/view/4456/4046

Análisis univariante

El coeficiente de kurtosis es útil para detectar la presencia de observaciones atípicas o outliers que corresponden a datos heterogéneos con el resto. La detección de estas observaciones es fundamental ya que estos valores extremos pueden distorsionar las medidas descriptivas.

Análisis univariante

En caso de encontrar datos atípicos en el conjunto de datos, es importante calcular además de los estadísticos tradicionales mencionados anteriormente, medidas más robustas de centralización y de dispersión de los datos.

En el caso de medidas de centralización, conviene calcular la **mediana** (valor que se encuentra en la posición central al ordenar los datos de menor a mayor)

En el caso de medidas de dispersión, conviene calcular la **MEDA** (mediana de las desviaciones absolutas respecto a la mediana)

Análisis univariante

Es importante siempre graficar las variables cuantitativas con un histograma o un diagrama de cajas. Además siempre se deben calcular la media y la mediana de cada variable. En el caso de que las dos medidas difieran mucho se debe determinar si se debe a una distribución asimétrica, la presencia de valores atípicos o heterogeneidad en los datos.

Practiquemos...

Tenemos una base de datos que cuenta con 8 variables tomadas en un grupo de 27 estudiantes:

- **Sexo (sex):** 0 para mujer, 1 para hombre.
- **Estatura (est):** estatura del estudiante en centímetros.
- **Peso (pes):** peso del estudiante en kilogramos.
- **Longitud del pie (lpie):** longitud del pie del estudiante en centímetros.
- **Longitud del brazo (lbra):** longitud del brazo del estudiante en centímetros.
- **Anchura de la espalda (aes):** anchura de la espalda del estudiante en centímetros.
- **Diámetro del cráneo (dcr):** diámetro del cráneo del estudiante en centímetros.
- **Longitud entre rodilla y tobillo (lrt):** longitud entre la rodilla y el tobillo del estudiante en centímetros.

Practiquemos...

observacion	sexo	est	pes	pie	lbr	aes	cdr	lrt
1	0	159	49	36	68	42	47	40
2	1	164	62	39	73	44	55	44
3	0	172	65	38	75	48	58	44
4	0	167	52	37	73	41,5	58	44
5	0	164	51	36	71	44,5	54	40
6	0	161	67	38	71	44	56	42
7	0	168	48	39	72,5	41	54,4	43
8	1	181	74	43	74	50	60	47
9	1	183	74	41	79	47,5	59,5	47
10	0	158	50	36	68,5	44	57	41

Análisis multivariante

El análisis de datos multivariantes tiene como objetivo el estudio estadístico de múltiples variables medidas en una población

Análisis multivariante

Medidas de centralización

- **Vector de medias:** vector de dimensión p que contiene las medias de cada una de las p variables.

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \begin{bmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_p \end{bmatrix}$$

Análisis multivariante

La matriz de varianzas y covarianzas

- En el caso de las variables escalares o univariantes, la variabilidad respecto a la media se mide por la varianza o la desviación estándar.
- La relación lineal entre dos variables se mide por la **covarianza** que se calcula de la siguiente forma:

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)$$

Análisis multivariante

La matriz de varianzas y covarianzas

- En el caso de una variable multivariante, la información de la relación o dependencia lineal de todas las variables que la conforman se puede presentar en la **matriz de varianzas y covarianzas** que se define como:

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$$

Análisis multivariante

La matriz de varianzas y covarianzas

- La matriz de varianzas y covarianzas es una matriz cuadrada y simétrica que cuya diagonal contiene las varianzas de las $\boldsymbol{p}$ variables y fuera de la diagonal contiene las covarianzas entre las variables:

$$\mathbf{S} = \begin{bmatrix} s_1^2 & \cdots & s_{1p} \\ \vdots & & \vdots \\ s_{p1} & \cdots & s_p^2 \end{bmatrix}$$

Una matriz simétrica cumple que la **matriz traspuesta** es igual a la matriz original.

Medidas globales de variabilidad

Cuando estudiamos variables que se miden en las mismas unidades (euros, km, kg) o son adimensionales (porcentajes, proporciones) podemos calcular medidas de la variabilidad promedio que permitan comparar distintos conjuntos de variables.

Medidas globales de variabilidad

Variabilidad total

$$T = tr(\mathbf{S}) = \sum_{i=1}^{p} s_i^2$$

Variabilidad promedio

$$\overline{s}^2 = \frac{1}{p} \sum_{i=1}^{p} s_i^2$$

Medidas globales de variabilidad

Las siguientes medidas fueron propuestas por Peña y Rodríguez (2000):

Varianza generalizada

$$VG = |\mathbf{S}|$$

Determinante de la matriz de varianzas y covarianzas: https://www.matematicas10.net/2015/12/ejemplos-de-determinante-de-una-matriz.html

Desviación típica generalizada

$$|\mathbf{S}|^{1/2}$$

Raíz cuadrada de la varianza generalizada

Medidas globales de variabilidad

Las siguientes medidas fueron propuestas por Peña y Rodríguez (2000):

Variabilidad promedio

$$VP = |\mathbf{S}|^{1/p}$$

Desviación promedio

$$DP = |\mathbf{S}|^{1/2p}$$

Medidas de dependencia lineal

Un objetivo fundamental de la descripción de los datos multivariantes es comprender la estructura de dependencias entre las variables

Coeficiente de correlación lineal

$$r_{jk} = \frac{s_{jk}}{s_j s_k}$$

$$-1 < r_{jk} < 1$$

Mide la dependencia lineal entre dos variables

Medidas de dependencia lineal

$$r_{jk} = \frac{s_{jk}}{s_j s_k}$$

$r_{jk} < 0$ Dependencia lineal negativa

$r_{jk} = 0$ No hay dependencia lineal

$r_{jk} > 0$ Dependencia lineal positiva

Medidas de dependencia lineal

La dependencia lineal por pares entre todas las variables se mide por la **matriz de correlación**

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ \vdots & \vdots & \dots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix}$$

Matriz cuadrada y simétrica