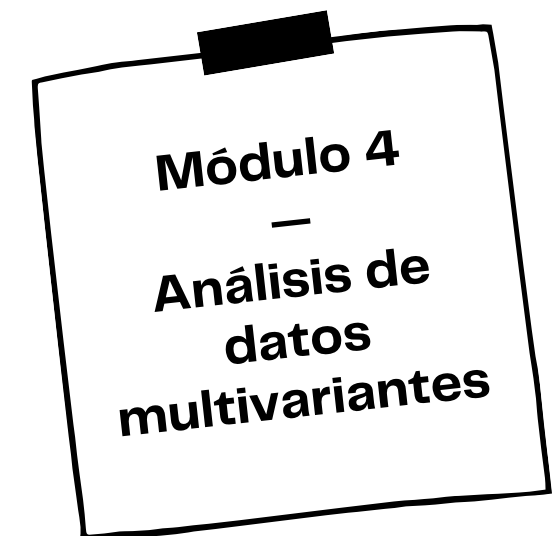


Análisis de datos multivariantes

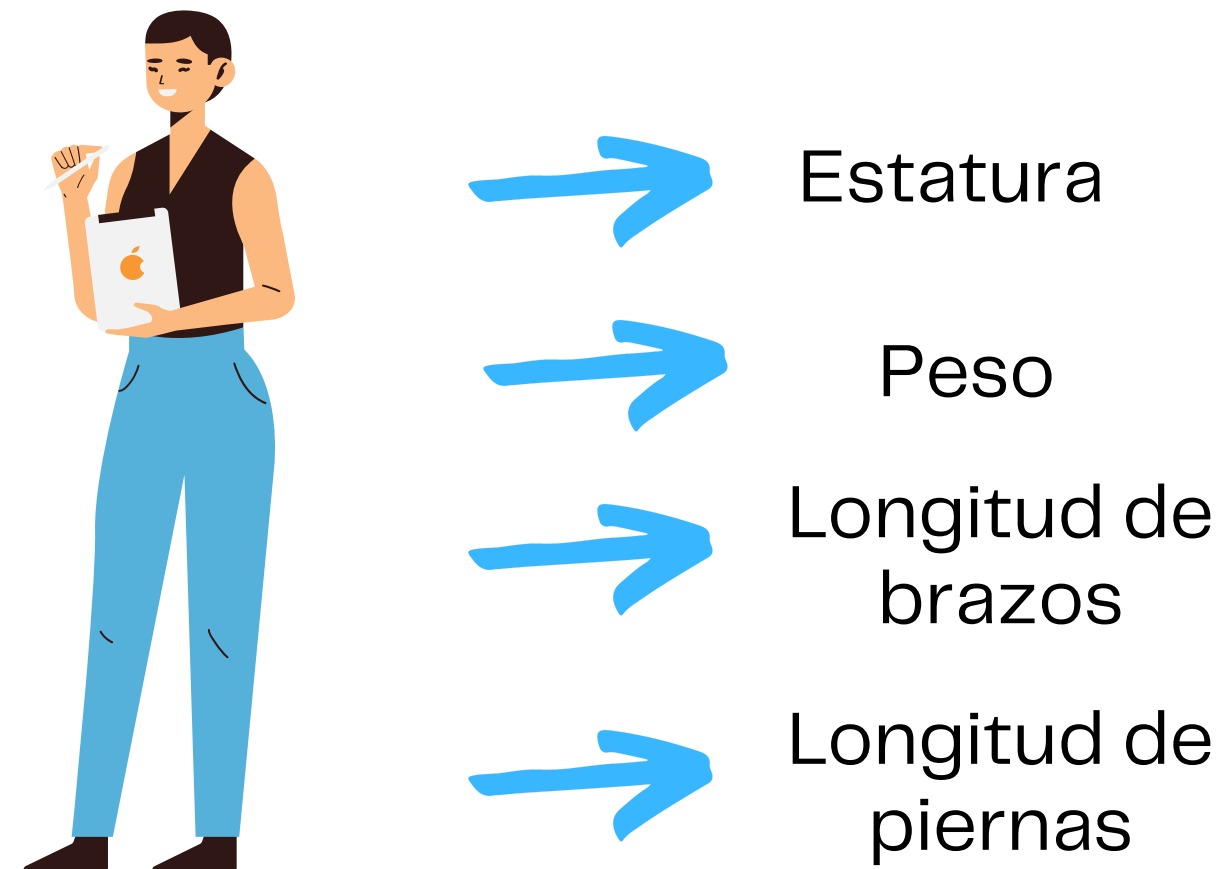


Q Agenda de hoy

- | | | | |
|---|-------------------------------------|---|---|
| 1 | Datos multivariantes | 5 | Medidas de variabilidad |
| 2 | Descripción de datos multivariantes | 6 | Medidas de dependencia lineal |
| 3 | Medidas de centralización | 7 | Gráficos útiles para el análisis multivariado |
| 4 | Matriz de varianzas y covarianzas | 8 | El concepto de distancia |

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



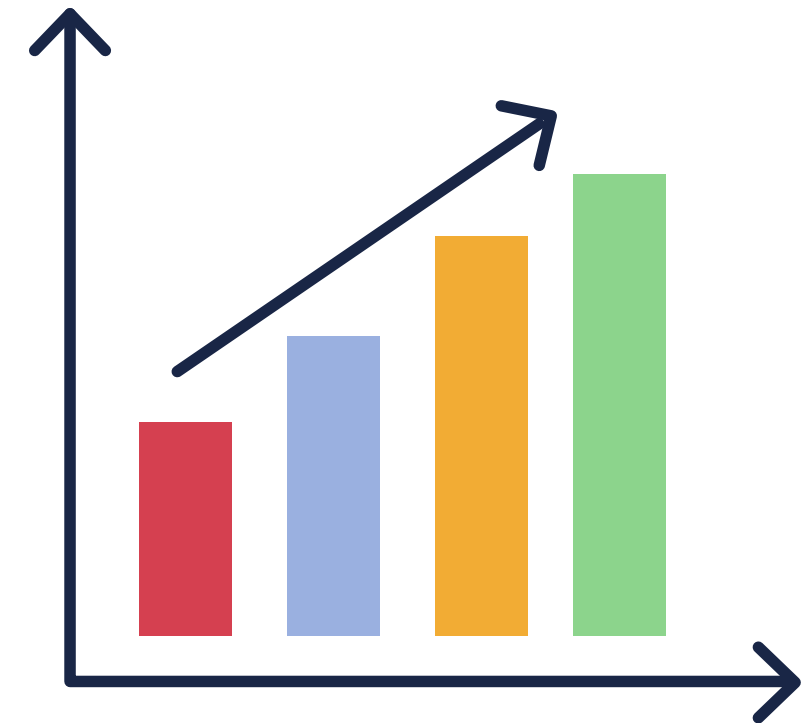
Clasificar nuevas observaciones en grupos definidos



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

Cualitativas

Continuas

Toman cualquier
valor real
(estatura)

Discretas

Solo toman valores
enteros
(número de hijos)

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 \mathbf{X} está conformada por los valores de las 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 x_j implica calcular:

Media

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


Desviación estándar

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

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

Análisis univariante

El estudio univariante de una variable cuantitativa x_j implica calcular:

Coeficiente de variación



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

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 x_j implica calcular:

Coeficiente de asimetría



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

Mide la simetría de los datos
respecto a su centro

Análisis univariante

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


$$A_j < 0$$

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

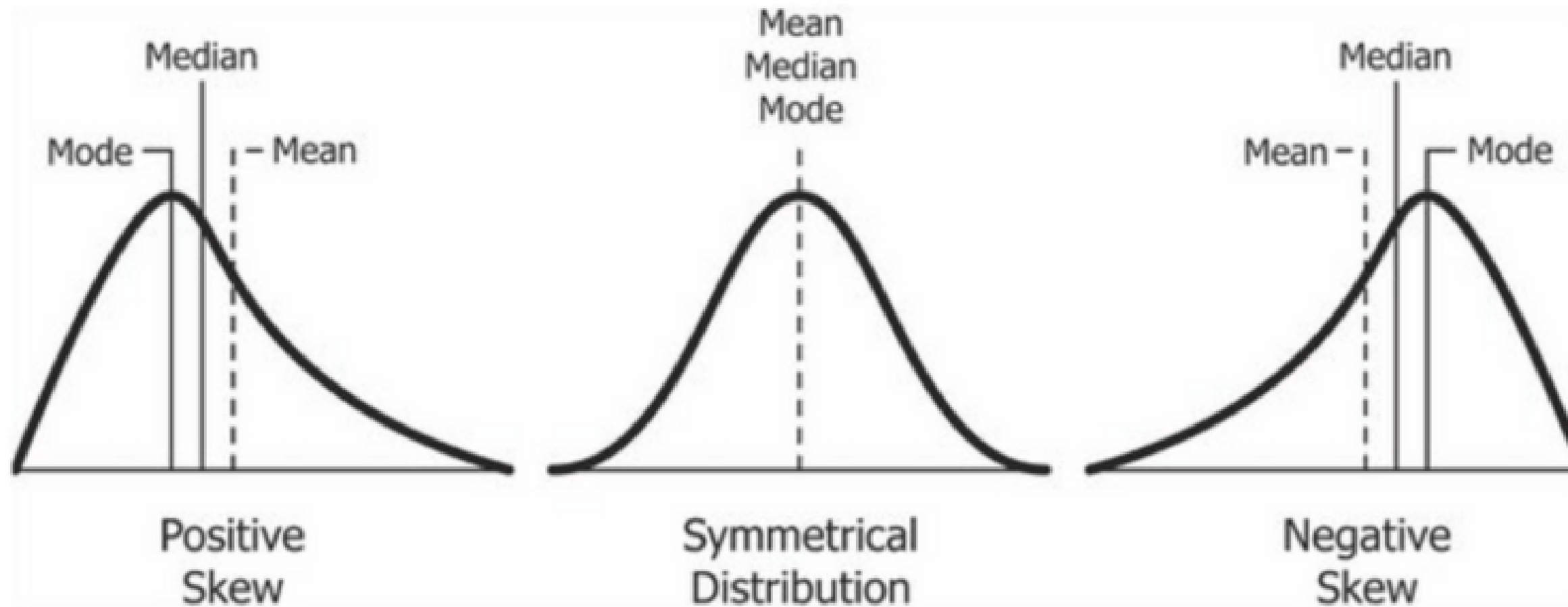

$$A_j = 0$$

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



Análisis univariante

El estudio univariante de una variable cuantitativa x_j implica calcular:

Coeficiente de Curtosis

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



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



$$K_j - 3 < 0$$

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

$$K_j - 3$$



$$K_j - 3 = 0$$

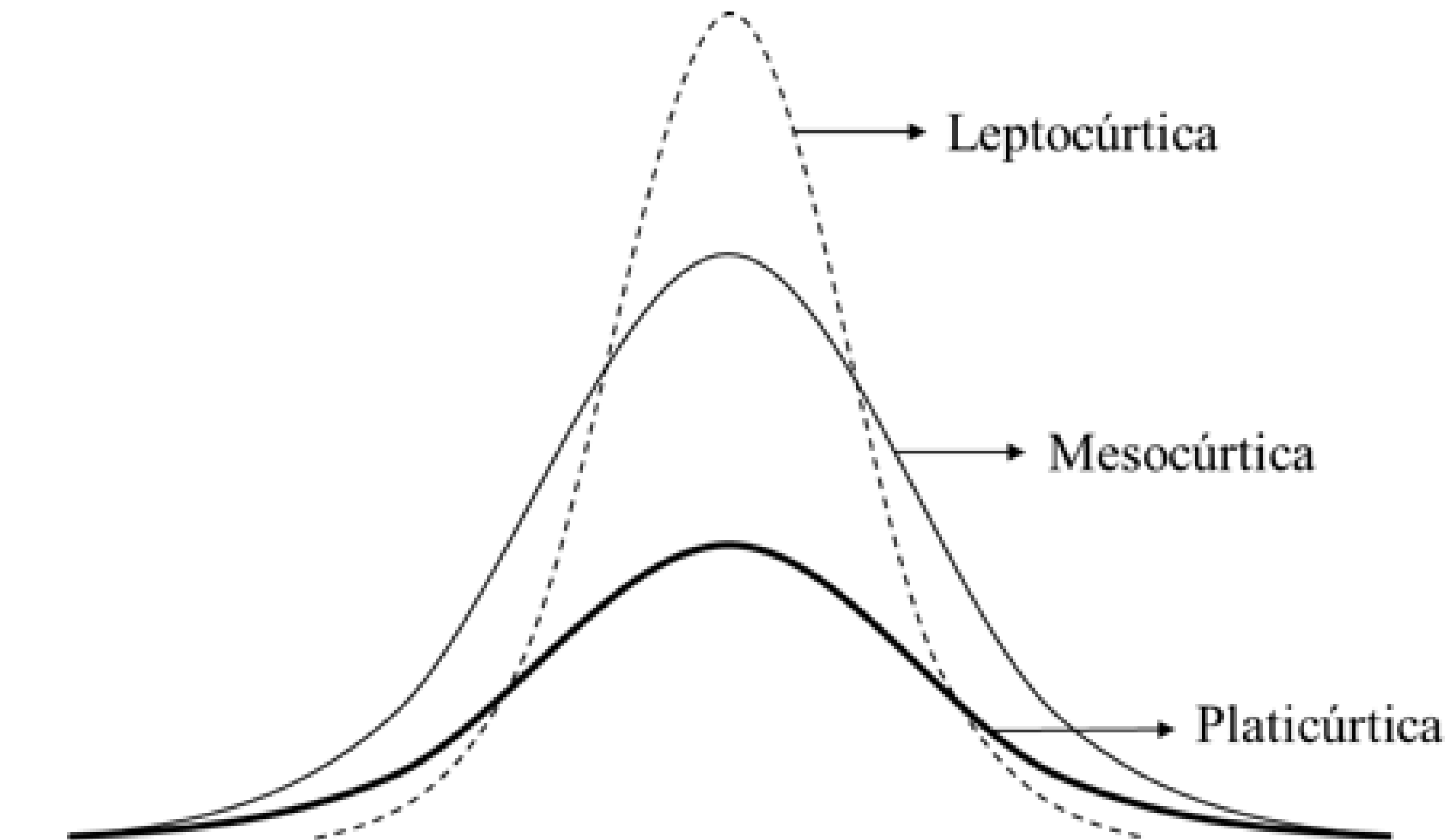
Distribución mesocúrtica: distribución normal.



$$K_j - 3 > 0$$

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

Análisis univariante



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.



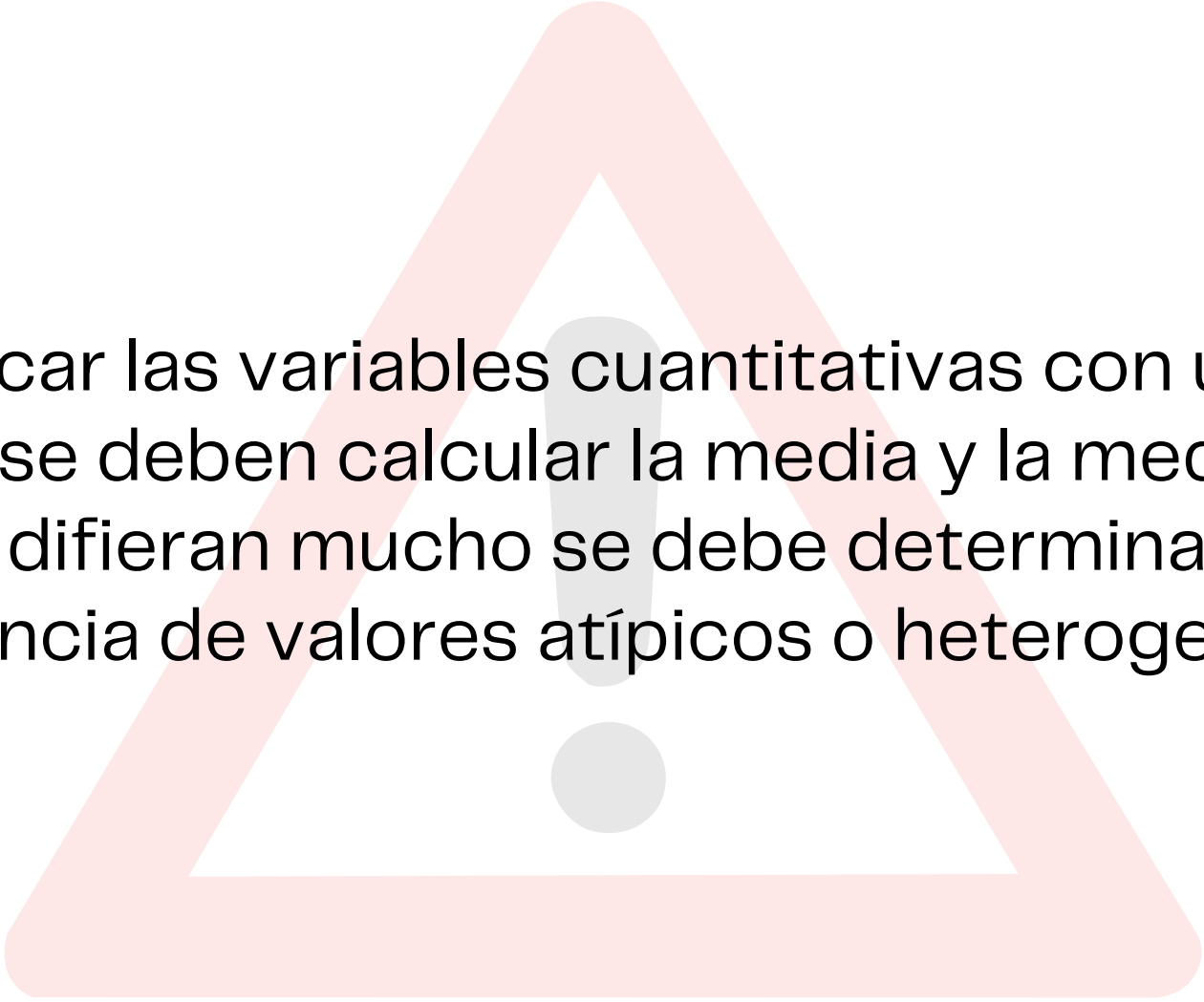
En el caso de medidas de centralización, conviene calcular la **mediana** (valor que se encuentra en la posición central al ordenar los datos de menor a mayor)



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

Arreglos basados en medidas descriptivas

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

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

Arreglos basados en medidas descriptivas

- 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:

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

Arreglos basados en medidas descriptivas

- 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**. Esta es una matriz cuadrada y simétrica que cuya diagonal contiene las varianzas de las p variables y fuera de la diagonal contiene las covarianzas entre las variables.



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

Arreglos basados en medidas descriptivas

$$\mathbf{S} = \begin{bmatrix} s_1^2 & \dots & s_{1p} \\ \vdots & & \vdots \\ s_{p1} & \dots & s_p^2 \end{bmatrix} = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \dots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \dots & \text{Var}(X_n) \end{bmatrix}$$

Medidas globales de variabilidad

Variabilidad total

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

Varianza generalizada

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



La varianza generalizada es un solo número y por lo tanto se pierde información al calcularla

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

$$r_{jk} = \frac{S_{jk}}{\sqrt{S_{jj}}\sqrt{S_{kk}}}$$

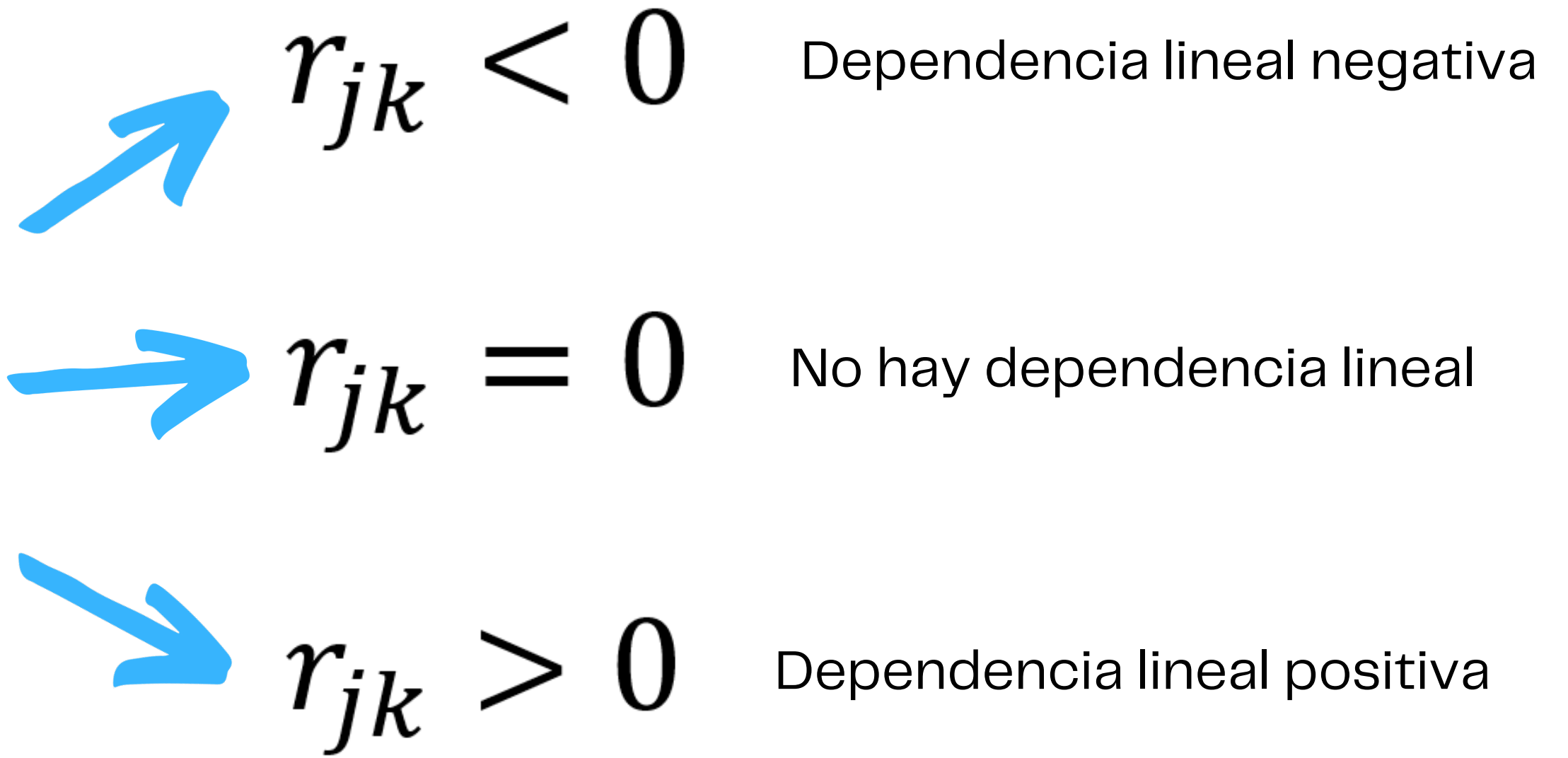


$$-1 < r_{jk} < 1$$

Mide la dependencia lineal
entre dos variables

Medidas de dependencia lineal

$$r_{jk} = \frac{S_{jk}}{\sqrt{S_{jj}}\sqrt{S_{kk}}}$$



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**

Gráficos útiles para el análisis multivariado

Es imposible graficar simultáneamente los valores de todas las variables de nuestra matriz de datos, sin embargo, podemos utilizar gráficos individuales y de pares de variables, los cuales pueden resultar muy informativos

Gráficos en R



Gráficos para variables individuales

Sirven para conocer las distribuciones marginales de los datos para cada una de las variables:

- Gráfico de puntos (recomendado para muestras pequeñas)
- Gráfico de cajas (recomendado para muestras moderadas o grandes)
- Histograma (recomendado para muestras moderadas o grandes)

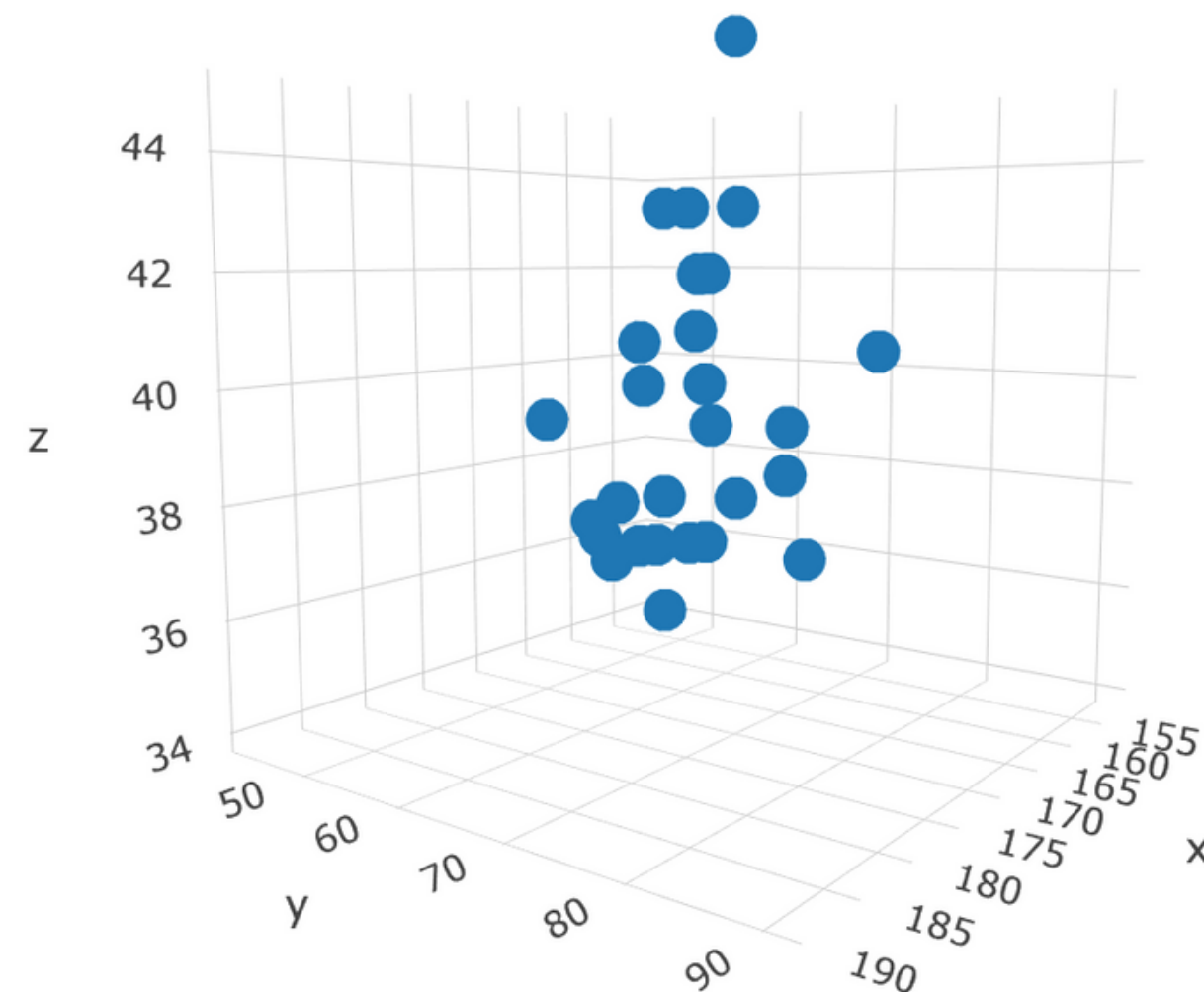
Gráficos para pares de variables

Sirven para conocer las distribuciones de los datos para dos variables. Dan indicaciones sobre la orientación de los datos en el plano cartesiano y sobre la asociación que existe entre ellos:

- Diagramas de dispersión simples
- Diagramas de dispersión con marginales (histogramas, gráficos de caja, gráficos de puntos)

Gráficos para tres variables

- Diagramas de dispersión tridimensionales



Matrices de dispersión

Presentan conjuntamente todos los diagramas de dispersión de los datos para cada par de variables. Se pueden construir varios tipos dependiendo del contenido de su diagonal:

- Matriz de dispersión con gráficos de puntos en su diagonal.
- Matriz de dispersión con gráficos de cajas en su diagonal.
- Matriz de dispersión con histogramas en su diagonal.

R Graphics Cookbook



Representaciones pictóricas de datos multivariados

Son imágenes que representan los valores de tres o más variables medidas para cada individuo u objeto. **Su objetivo es ayudar a reconocer individuos u objetos similares.**

Cuando se usan estos gráficos se recomienda que todas las variables estén medidas en la misma escala. De lo contrario, se deben usar los datos estandarizados

Representaciones pictóricas de datos multivariados

- **Gráfico de estrellas:** consiste en construir círculos de radio fijo con p rayos igualmente espaciados que salen del centro del círculo. Las longitudes de los rayos representan los valores de las variables.
- **Gráfico de caras de Chernoff:** usa varias características de la cara para representar los datos de las variables (curvatura de la boca, ángulo de la ceja, amplitud de la nariz, longitud de la nariz, altura de la cara, radio de las orejas, longitud de la ceja).

El concepto de distancia

La mayoría de las técnicas del análisis multivariado están basadas en el concepto de distancia

Desviación estándar

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

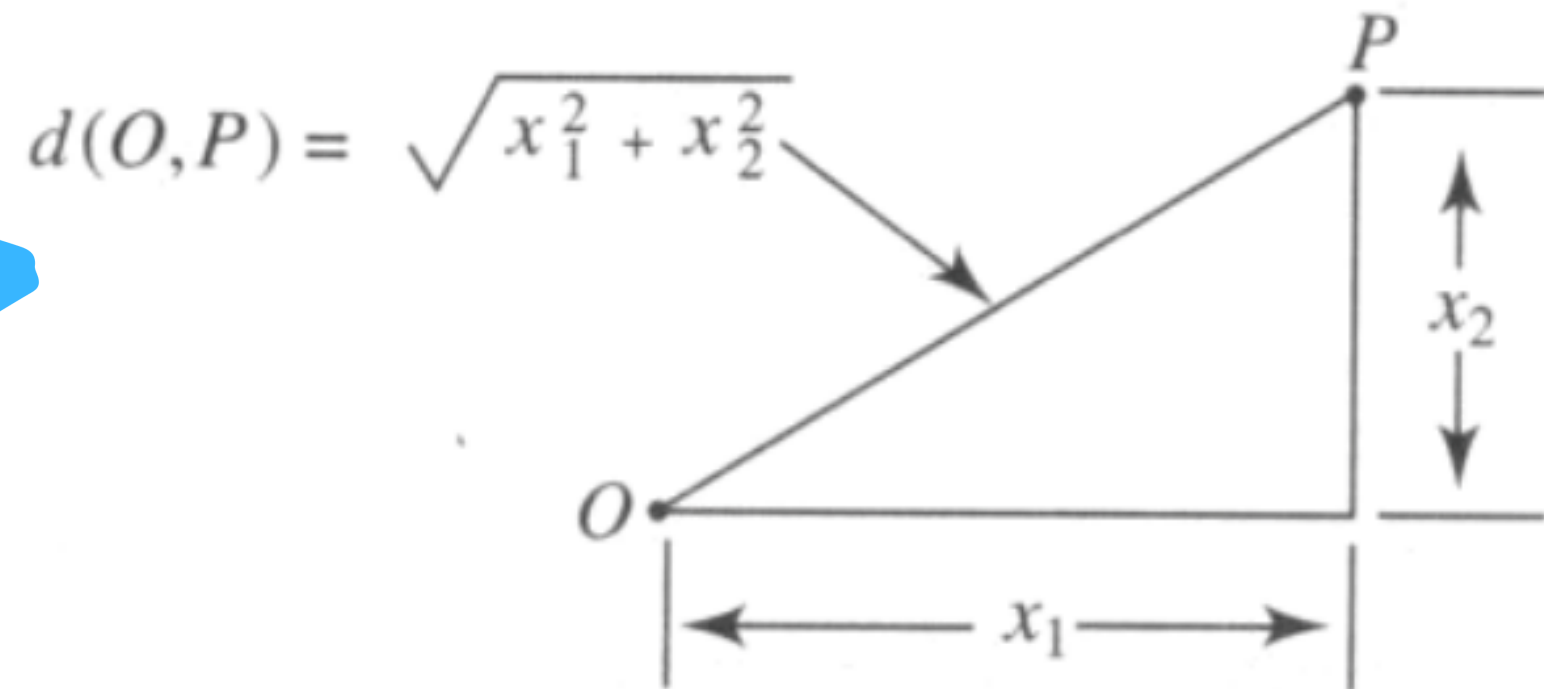


La desviación estándar es un promedio de las distancias entre el valor de una variable x en un punto x_j y la media de la variable

El concepto de distancia

$$d(O, P) = \sqrt{x_1^2 + x_2^2} \quad (\text{Teorema de Pitágoras})$$

Considere el punto $P = (x_1, x_2)$
La distancia euclidiana del
origen $O = (0,0)$ a P es:



Distancia euclidiana

Existen diferentes formas de calcular la distancia entre dos puntos. La distancia más utilizada es la distancia euclidiana, sin embargo tiene el inconveniente de depender de las unidades de medidas de las variables

$$d_{ij} = \left(\sum_{s=1}^p (x_{is} - x_{js})^2 \right)^{1/2}$$

Distancia euclidiana

Sea x la estatura de una persona en metros e y su peso en kilogramos.

Comparemos la distancia entre tres personas

Individuo	Estatura (metros)	Peso (kg)
A	1.80	80
B	1.70	72
C	1.65	81

Distancia euclidiana

$A(1.80, 80)$

$B(1.70, 72)$

$C(1.65, 81)$



$$\begin{aligned}d^2(A, B) &= (1.80 - 1.70)^2 + (80 - 72)^2 \\&= 0.1^2 + 8^2 \\&= 64.04\end{aligned}$$



$$\begin{aligned}d^2(A, C) &= (1.80 - 1.65)^2 + (80 - 81)^2 \\&= 0.15^2 + 1^2 \\&= 1.225\end{aligned}$$

Con la distancia euclidiana, el individuo A está más cerca del individuo C que del B

Distancia euclidiana

Supongamos que decidimos medir la estatura en centímetros en lugar de metros:

$A(180, 80)$

$B(170, 72)$

$C(165, 81)$



$$\begin{aligned}d^2(A, B) &= (180 - 170)^2 + (80 - 72)^2 \\&= 10^2 + 8^2 \\&= 164\end{aligned}$$

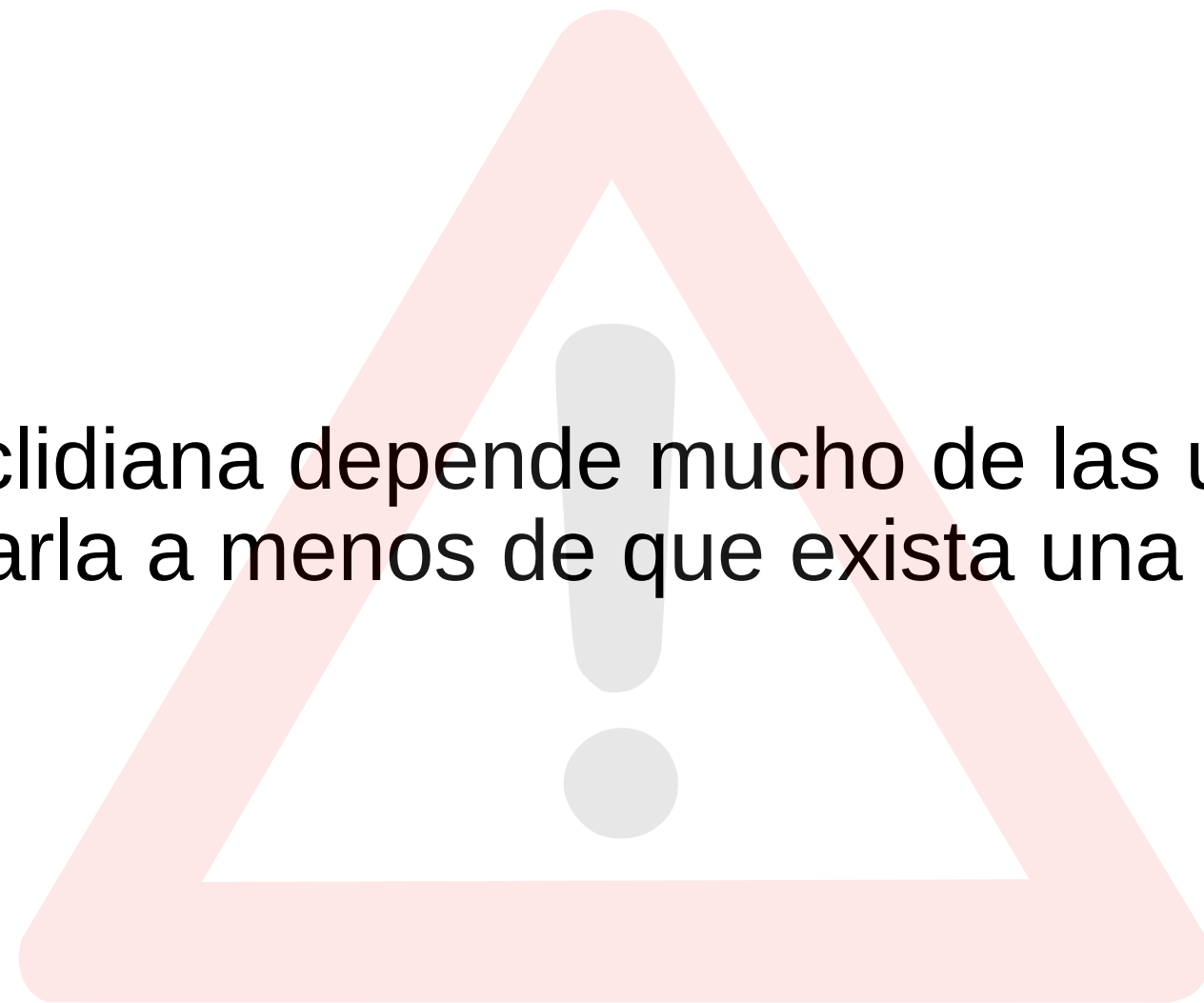


$$\begin{aligned}d^2(A, C) &= (180 - 165)^2 + (80 - 81)^2 \\&= 15^2 + 1^2 \\&= 226\end{aligned}$$

Con la distancia euclidiana, el individuo A está más cerca del individuo B que del C


Distancia euclidiana

Como la distancia euclidiana depende mucho de las unidades de medida, no se recomienda utilizarla a menos de que exista una una unidad fija natural



Distancia euclidiana ponderada

Una forma de evitar el problema de las unidades es dividir cada variable por un término que las estandarice

$$d_{ij} = \left[(\mathbf{x}_i - \mathbf{x}_j)' \mathbf{M}^{-1} (\mathbf{x}_i - \mathbf{x}_j) \right]^{1/2}$$


M es una matriz diagonal que se utiliza para estandarizar las variables

Distancia euclidiana ponderada

$$M = \begin{pmatrix} S_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & S_p \end{pmatrix}$$



$$d_{ij} = \left(\sum_{s=1}^p \left(\frac{x_{is} - x_{js}}{S_s} \right)^2 \right)^{1/2} = \left(\sum_{s=1}^p S_s^{-2} (x_{is} - x_{js})^2 \right)^{1/2}$$

Distancia de Mahalanobis

La distancia de Mahalanobis entre dos puntos está definida como:

$$d_{ij} = \left[(\mathbf{x}_i - \mathbf{x}_j)' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j) \right]^{1/2}$$

Distancia de Mahalanobis

La distancia de Mahalanobis entre un punto y su vector de medias está definida como:

$$d_i = \left[(\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \right]^{1/2}$$

Calculemos las distancias en R para nuestra base de datos de características físicas de 27 estudiantes...