Introducción

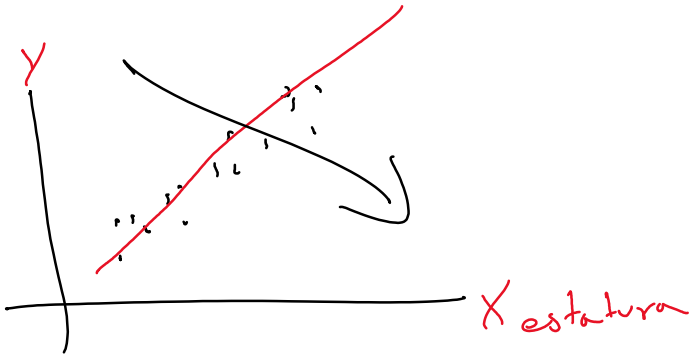
miércoles, 1 de febrero de 2023 7:15 a. m.

$F_X(2)$

$P(X \leq 2)$

$P_X(2) \rightarrow P(X=2)$

$\mu_1 > \mu_2$



$r = 0.9$

- La realidad es que la mayoría de casos de estudio tienen varias variables

	Var 1,	Var 2	...	Var K
Obs 1	—	—		
⋮				⋮
Obs n	—	—	—	—

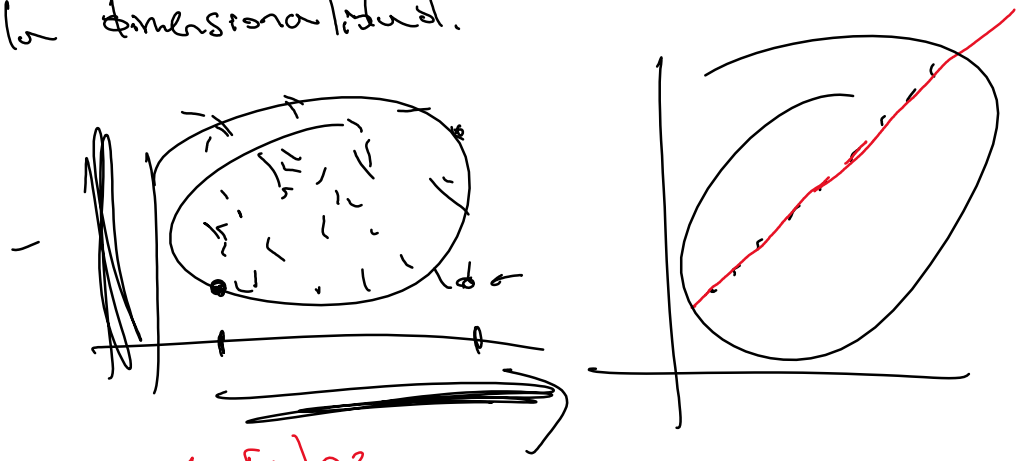
PyE1 y PyE2

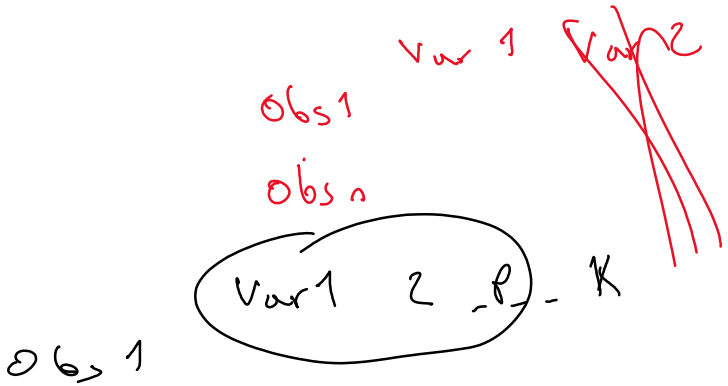
- Exploremos los paquetes [illegible] y veremos algunos nuevos para estudiar estos casos.

- Utilizaremos R y R markdown.

Principales cosas que se hacen en investigación o en análisis multivariante son:

- Reducción de la dimensionalidad.



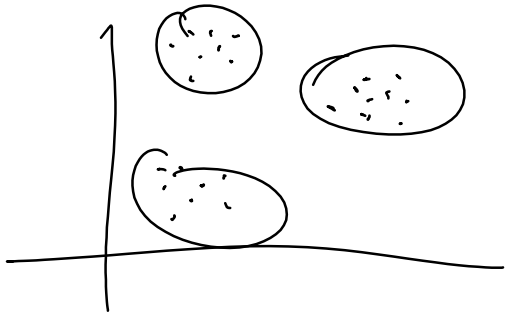


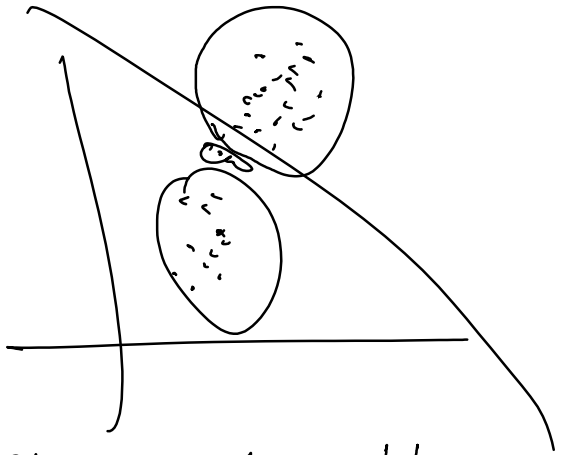
Obsn

- Ordenamiento de datos
- Dependencia de variables
- Predicción

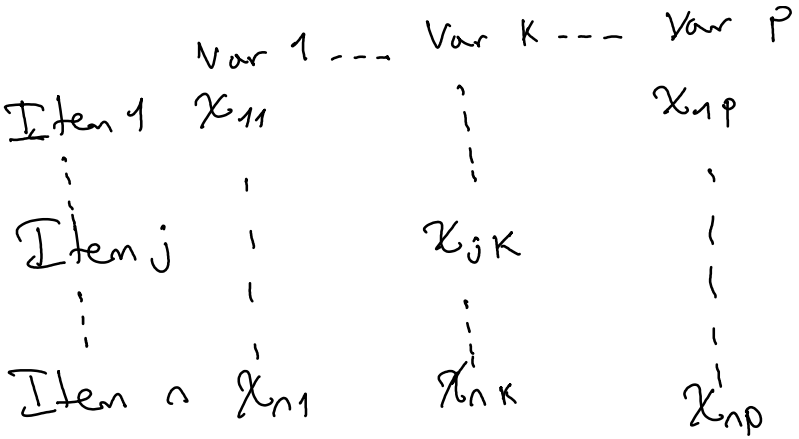
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots$$

- Pruebas de hipótesis
- Clasificación.





Vamos a usar notación matricial X_{jk} es la medida de la k-ésima variable en la j-ésima observación



$$X = \begin{bmatrix} x_{11} & \cdots & x_{1K} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & & x_{nK} & & x_{np} \end{bmatrix}$$

Ejemplo:

Variable 1:	42	55	48	58
Variable 2:	4	5	4	3

$x_{11} = 42$

$x_{21} = \cancel{4}$ ~~55~~

$$X = \begin{bmatrix} 42 & 4 \\ 55 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$$

Media $\bar{x} = \frac{1}{n} \sum x_{jK}$

$\bar{x}_K = \frac{1}{n} \sum_{j=1}^{n}$ ...

Varianza muestral $\quad S_K^2 = \frac{1}{n} \sum_{j=1}^{n} (x_{jK} - \bar{x}_K)^2$

Matriz de varianzas-covarianzas

$$S_{p \times p} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & \ddots & & \\ \vdots & & & \\ S_{p1} & & & S_{pp} \end{bmatrix}$$

Simétrica.

$S_{21} = S_{12}$

$S_K = \sqrt{S_K^2}$ → desviación estándar de K

$S_K^2 = S_{KK} \neq S_K$

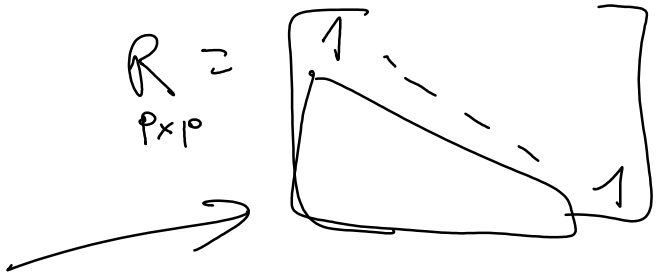
↳ varianza de K

Coeficiente de correlación muestral

$r_{iK} = \frac{S_{iK}}{\quad} = \sum (x_{ji} - \bar{x}_i)(x_{jK} - \bar{x}_K)$

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \frac{\sum (x_{ji} - \bar{x}_i)^2}{\sqrt{\sum (x_{ji} - \bar{x}_i)^2}\sqrt{\sum (x_{jk} - \bar{x}_k)^2}}$$

$R = \underset{p \times p}{}$



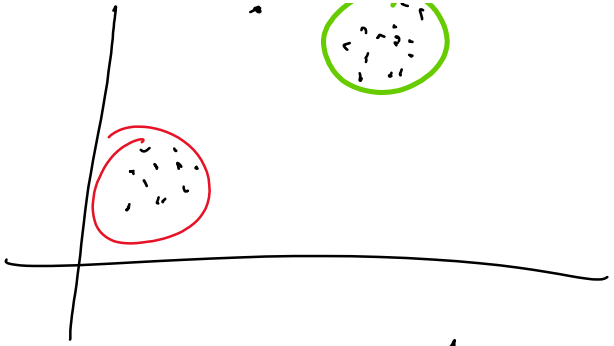
Simétrica

- Es una medida de la asociación lineal entre las variables que NO depende de las unidades de las variables.

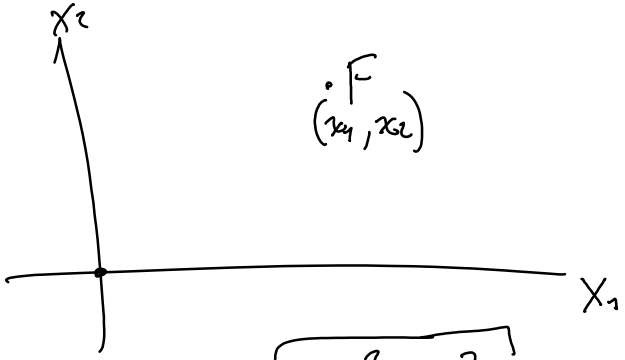
Vector de medias

$$\underset{p \times 1}{\bar{X}} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

Para estudiar/analizar datos multivariados es muy útil utilizar el concepto de distancias



Cómo calculamos la distancia?



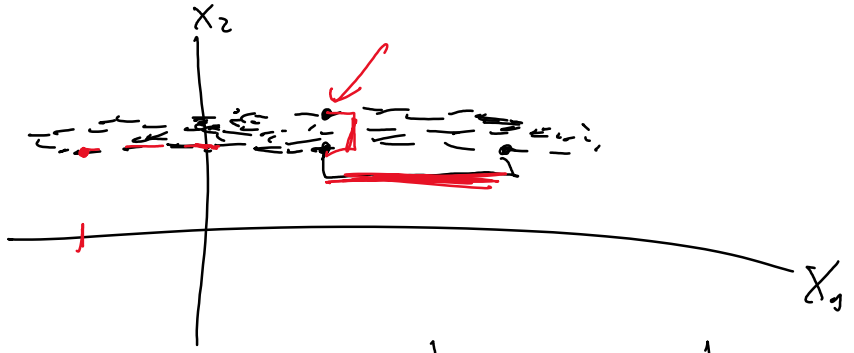
$$\text{Euclid}(O, F) = \sqrt{x_1^2 + x_2^2}$$

$$\text{Euclid}(O, F) = \sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}$$

Distancia euclidea entre dos puntos

$$P = (x_1, x_2 \cdots, x_p) \qquad Q = (y_1, y_2, \cdots y_p)$$

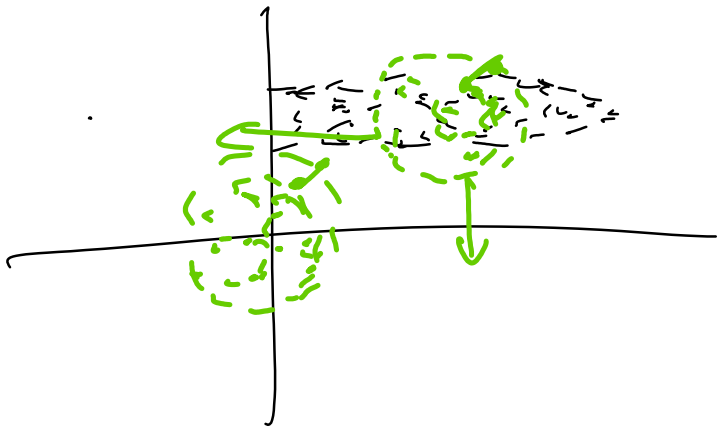
$$d(P, Q) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_p - y_p)^2}$$



Para la mayoría de aplicaciones, la distancia euclidiana no es apropiada

Calcularemos la distancia estadística

$$x_1^* = \frac{x_{1} - \bar{x}_1}{\sqrt{s_{11}}} \qquad x_2^* = \frac{x_2}{\sqrt{s_{22}}}$$



Distancia estadística de un punto P y el origen

$$\sqrt{(x_1^* - x_2^*)^2}$$

$$d(O,F) = \sqrt{(x_1^*)^2 + (x_2^*)^2}$$

$$= \sqrt{\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}}}$$

$$d(F,Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \cdots + \frac{(x_p - y_p)^2}{s_{pp}}}$$

$$= \sqrt{(x_1^* - y_1^*)^2 + \cdots + (x_p^* - y_p^*)^2}$$