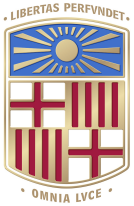
2. Estadística descriptiva univariant

Universitat de Barcelona



UNIVERSITAT DE BARCELONA

Classificació

- Variables qualitatives.
  - No tenen caràcter numèric. Etiquetes.
  - Altres noms: categòriques, nominals, factors.
  - Exemples: dades binàries i dades ordinals.
- Variables quantitatives.
  - Discretes
    - Permeten operacions aritmètiques, com calcular el promig.
    - Prenen valors en un conjunt finit.
  - Contínues
    - Permeten operacions aritmètiques, com calcular el promig.
    - Prenen valors en un interval numèric. Qualsevol valor de l'interval és possible.

Com classifiquem aquestes variables?

Quin tipus de variables són?

Números de telèfon, Grup Sanguini, Nivell d'estudis, Talles de roba, Temperatura diària a BCN, Categoria dels Hotels, Classe social, Número de fills, IBEX-35, Opinió sobre el turisme a BCN.

Freqüències

Per variables categòriques i numèriques

- Freqüència absoluta (n_i)
  - Nombre de vegades que apareix aquest valor en determinat conjunt de dades.
  - La suma de les freqüències absolutes és el total n de dades.
- Freqüència relativa (f_i)
  - Resultat de dividir la freqüència absoluta per n.
  - Proporció o tant per u d'un valor en el conjunt de dades. La suma de les freqüències relatives és 1.

$$\boxed{f_i = \frac{n_i}{n}}$$

Freqüències acumulades

Per variables numèriques

- Freqüència absoluta acumulada (N_i) (fins un valor donat): La suma de les freqüències absolutes corresponents als valors $\leq$ el valor donat.

$$\boxed{N_i = n_1 + n_2 + \ldots + n_i}$$

- Freqüència relativa acumulada (F_i) (fins un valor donat): La suma de les freqüències relatives corresponents als valors $\leq$ el valor donat.

$$\boxed{f_i = \frac{N_i}{n} = \frac{n_1 + n_2 + \ldots + n_i}{n} = f_1 + f_2 + \ldots + f_i}$$

Què passa si tenim molts valors diferents?

1. Dividim l'interval de valors d'una variable contínua en intervals més petits $I_i = (L_i, L_{i+1}]$.
2. Posem una etiqueta a cadascun (per exemple el punt mig $\frac{L_i + L_{i+1}}{2}$). Aquesta etiqueta és la marca de classe.
3. Obtenim una nova variable discreta. Podem fer així una taula de freqüències de la nova variable.

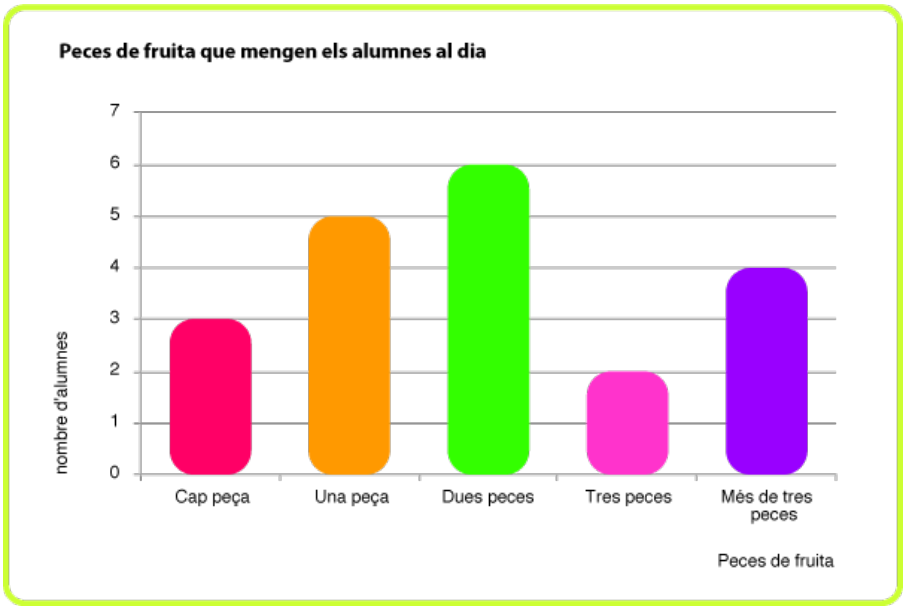
Consell: Triar entre 6 i 25 intervals. Normalment, s'agafen tots de la mateixa mida.

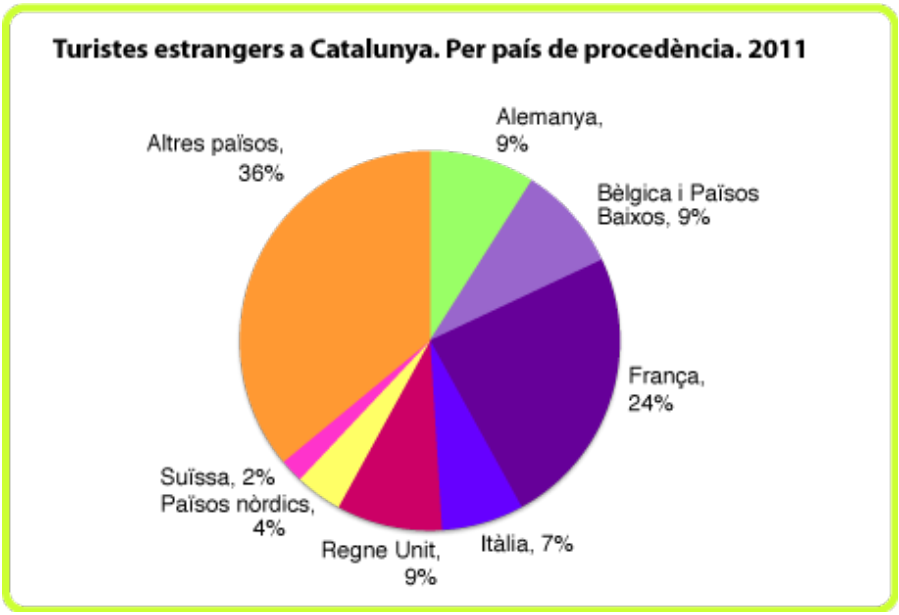
Exercici: Temperatures dels últims xx dies.

Núm. vegades	n_i	N_i	f_i	F_i
10	2	2	0.05	0.05
13	4	6	0.10	0.15
16	10	16	0.25	0.40
19	15	31	0.375	0.775
22	6	37	0.15	0.925
25	3	40	0.075	1

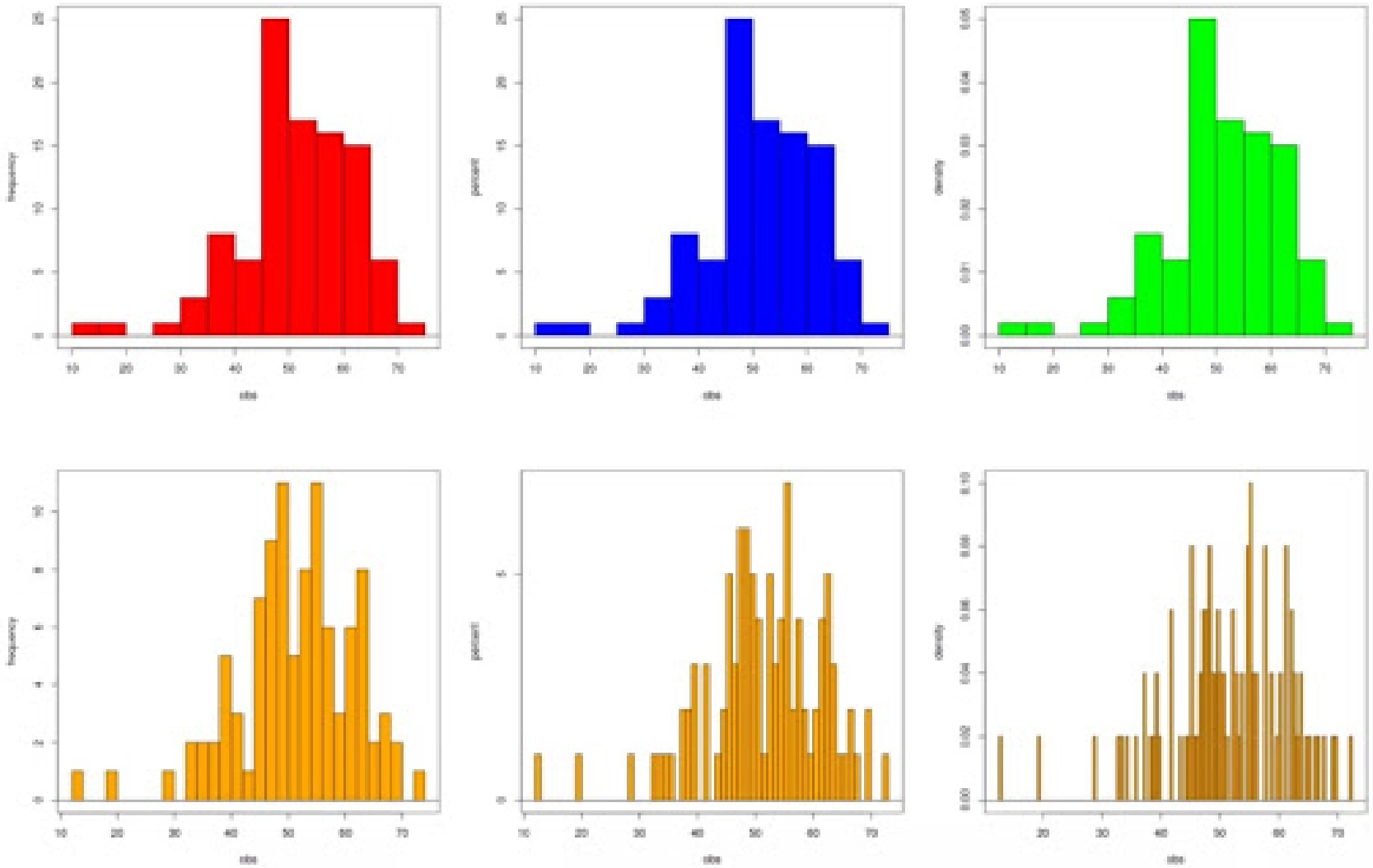
1. Quina és la mida de la mostra?
2. Quants de dies hi hagut una temperatura de més de 19 graus? I de 16 o menys graus?
3. Quin percentatge de dies hi hagut una temperatura de 22 graus?
4. El 77.5% dels dies hem tingut una temperatura de . . . . . graus.

Gràfics per variables qualitatives

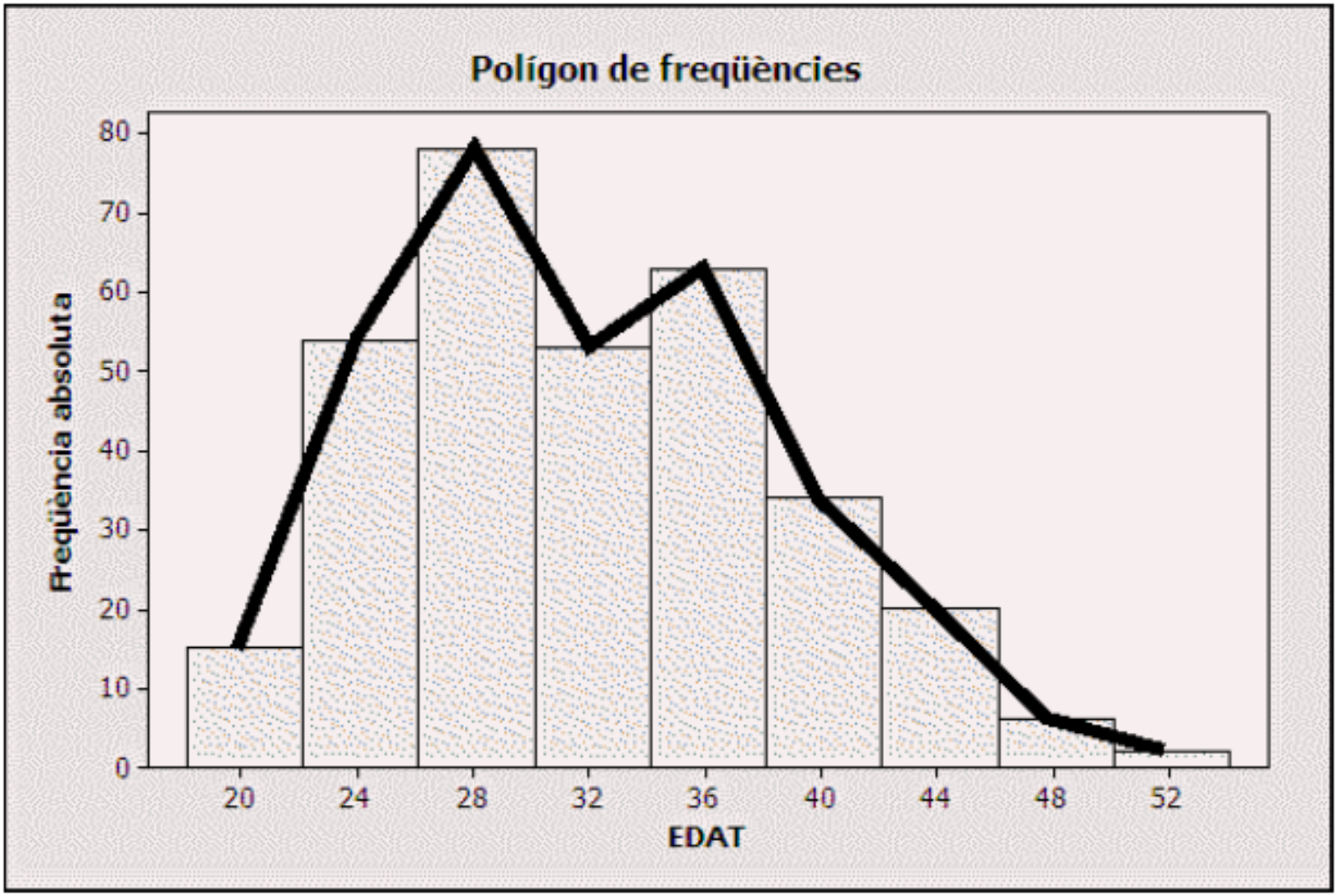




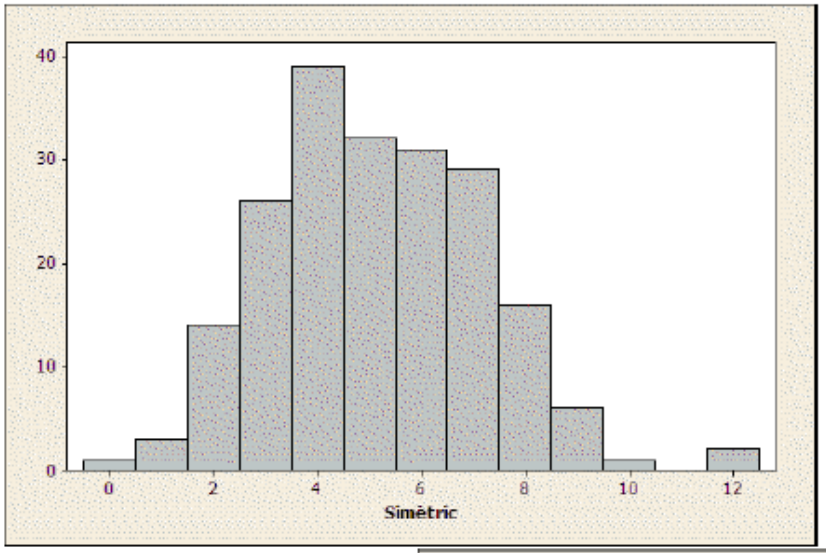
Gràfics per variables quantitatives

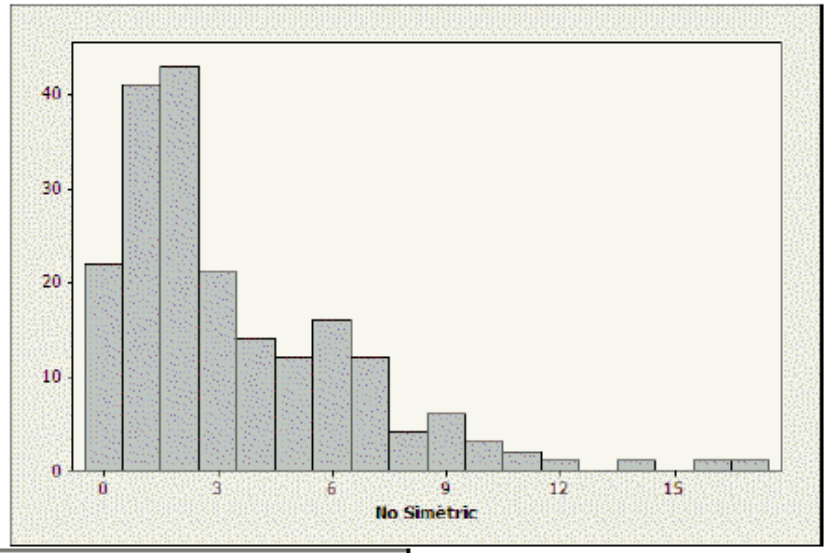


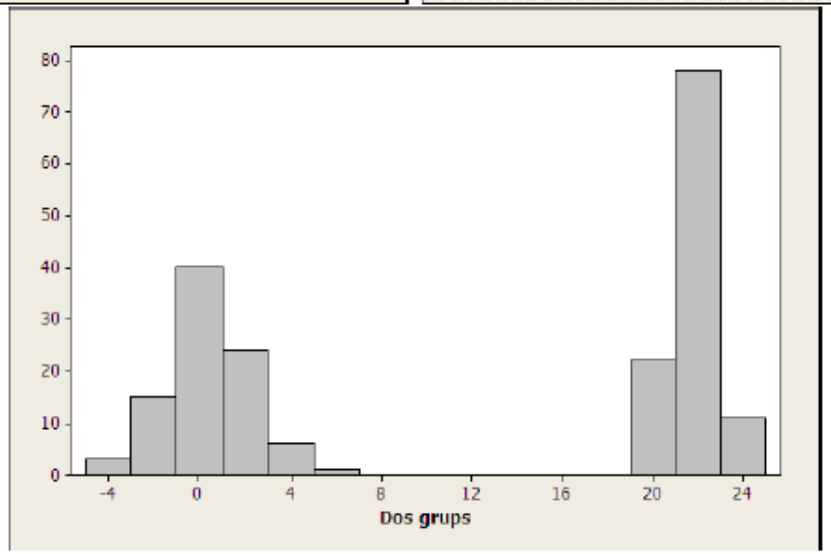
Gràfics per variables quantitatives



Gràfics per variables quantitatives







Resums numèrics de dades

Objectiu

Volem trobar un o uns valors que representin un conjunt de dades donat.

Tenim diferents tipus de mesures

- Mesures de centre
- Mesures de diversitat
- Mesures de posició
- Mesures de forma

La Moda

Definició És el valor més freqüent.

Notació **Mo**

Tipus de dades Per dades qualitatives o quantitatives discretes.

Observacions Pot no ser única, si la màxima freqüència correspon amb 2 o més valors.

Exemple càlcul **Moda**

Exemple (Casaments)

Dades cens EUA 2004. Invididus 20-24 anys. Variable: Número de vegades que han estat casats (en milers).

	Homes	*Dones*
0	*7350*	*8418*
1	*2587*	*1594*
2	*80*	*10*
Total	*10017*	*10022*

La mitjana

Definició Per obtenir la mitjana de n nombres $x_1, x_2, \ldots, x_n$

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

Notació $\bar{x}$

Tipus de dades Per dades quantitatives.

Observacions És sempre representativa?

Exemples càlcul **Mitjana**

Exemple (Cereals)

	Sodium(mg)	*Sugar(g)*	*Type*
Frosted Mini Wheats	*0*	*11*	*A*
Raisin Bran	*340*	*18*	*A*
All Bran	*70*	*5*	*A*
Apple Jacks	*140*	*14*	*C*
Cap'n Crunch	*200*	*12*	*C*
Cheerios	*180*	*1*	*C*
Cinnamon Toast Crunch	*210*	*10*	*C*
Crackling Oat Bran	*150*	*16*	*A*
Fiber One	*100*	*0*	*A*
Frosted Flakes	*130*	*12*	*C*
Froot Loops	*140*	*14*	*C*
Honey Bunches of Oats	*180*	*7*	*A*
Honey Nut Cheerios	*190*	*9*	*C*
Life	*160*	*6*	*C*
Rice Krispies	*290*	*3*	*C*
Honeys Smacks	*50*	*15*	*A*
Special K	*220*	*4*	*A*
Wheaties	*180*	*4*	*A*
Corn Flakes	*200*	*3*	*A*
HoneyComb	*210*	*11*	*C*

$$\bar{x} = \frac{1}{20}(0 + 340 + 70 + 140 + 200 + 180 + 210 + 150 + 100 + 130 + 140 + 180 + 190 + 160 + 290 + 50 + 220 + 180 + 200 + 210) = 167$$

Exemples càlcul **Mitjana**

Exemple (Casaments)

Homes

$$\bar{x}_h = \frac{7350 \cdot 0 + 2587 \cdot 1 + 80 \cdot 2}{10017} = 0.2742338$$

Dones

$$\bar{x}_d = \frac{8418 \cdot 0 + 1594 \cdot 1 + 10 \cdot 2}{10022} = 0.1610457$$

Per tant, si les dades estan agrupades en freqüències

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} x_i \cdot n_i = \sum_{i=1}^{k} x_i \cdot f_i$$

Exemples càlcul **Mitjana**

Exemple (Sous)

Hi ha una empresa amb 4 programadors que cobren 1000 euros i el propietari que cobra 6000 euros.

Si calculem la mitjana

$$\bar{x} = \frac{4 \cdot 1000 + 1 \cdot 6000}{5} = 2000$$

Ens serveix en aquest cas la mitjana per fer-nos una idea del que es cobra en aquesta empresa?

La Mediana

Definició És el valor que queda al centre una vegada s'ha ordenat la llista de més petit a gran.
Per calcular-la tenim casos diferents:

- Si n és senar: agafem el valor central.
- Si n és parell: agafem els dos valors centrals i fem el promig.
- Si les dades estan agrupades per intervals $(L_i, L_{i+1}]$

$$Me = L_i + (L_{i+1} - L_i)\frac{\frac{n}{2} - N_{i-1}}{n_i}$$

Notació **Me**

Tipus de dades Per dades quantitatives o qualitatives ordinals.

Exemples càlcul **Mediana**

Exemple (Casaments)

La mediana del nombre de casaments tant per homes com per dones és 0.

Exemple (Cereals)

Per calcular la mediana hauríem d'ordenar totes les dades segons els mil·ligrams de sodi. I mirem les dades que ocupen les posicions 10 i 11. En els dos llocs tenim 180, per tant la mediana és 180.

Exemple (Sous)

La mediana dels sous és 1000 euros.

Què opineu quan comparem aquestes medianes amb les mitjanes?

Altres mesures de centralitat

Mitjana retallada al 5% S'eliminen el 5% de les observacions més grans i el 5% de les més petites i es calcula la mitjana amb el 90% restant.

Mitjana harmònica El recíproc de la mitjana aritmètica dels recíprocs.

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

o bé

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

si tenim les dades agrupades en freqüències.

Exemple

Exemple

Un cotxe va tots els dies de la setmana (excepte diumenge) de BCN a Castelldefels (20km). El dissabte va a 70 km/h i els dies laborables a 40 km/h. Quina és la velocitat mitjana d'una setmana?

$$H_x = \frac{6}{\frac{5}{40} + \frac{1}{70}} = 43.07 km/h$$

Variància i desviació típica

Definició

Variància

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Variància corregida

$$\widetilde{s_x}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Desviació típica

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

Tipus de dades Per dades quantitatives.

Per dades agrupades

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2 = \sum_{i=1}^{k} f_i(x_i - \bar{x})^2$$

Alternativa

$$\boxed{s_x^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - (\bar{x})^2}$$

Propietats

- Translació de les dades: $y_i = x_i + a$

$$\bar{y} = \bar{x} + a, \qquad s_y^2 = s_x^2$$

- Canvi d'escala: $y_i = bx_i$

$$\bar{y} = b\bar{x}, \qquad s_y^2 = b^2 s_x^2 \qquad s_y = |b| s_x$$

Dispersió Relativa

Objectiu Comparar la dispersió de dues variables. Obtenir mesures que no tinguin unitats.

Coeficient de variació

$$CV_x = \frac{s_x}{\bar{x}}$$

S'acostuma a demanar que $\bar{x} > 0$.

Problema

Quan $\bar{x}$ és molt propera a 0, CV_x perd significat.

Altres mesures de dispersió

Rang

$$Rang_x = \max(x_1, \ldots, x_n) - \min(x_1, \ldots, x_n).$$

Desviació mitjana

$$D_m = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$$

Desviació mediana

$$D_{Me} = \frac{1}{n} \sum_{i=1}^{n} |x_i - Me|$$

Estandardització d'una variable

Objectiu Poder comparar dos valors de dues mostres diferents.

Per qualsevol conjunt de dades $x_1, \ldots, x_n$ amb mitjana $\bar{x}$ i desviació típica s_x, la **mostra estandaritzada** és

$$z = (z_1, \ldots, z_n) \qquad \text{on } z_i = \frac{x_i - \bar{x}}{s_x},\ 1 \leq i \leq n.$$

La mostra z té mitjana 0 i desviació típica 1.

Desigualtat de Txebitxev

Per qualsevol conjunt de dades $x_1, \ldots, x_n$ amb mitjana $\bar{x}$ i desviació típica s_x, aleshores per $K > 0$

$$[\bar{x} - K \cdot s_x, \bar{x} + K \cdot s_x]$$

conté el $\left(1 - \frac{1}{K^2}\right) \cdot 100\%$ o més de les dades.

Exemple

Suposem que tenim unes dades amb mitjana 72.1 i desviació típica 8.3. Aleshores,

1. *Quin és el % d'observacions que trobarem a l'interval* $[30.6, 113.6]$*?*
2. *Entre quins valors trobem el 85% de les observacions?*

Estadístic Ordinal

Donades unes dades $x = (x_1, \ldots, x_n)$, la llista ordenada

$$x_{(1)}, \ldots, x_{(n)}$$

s'anomena **estadístic ordinal** de x.

Exemple (Emissions CO2)

Dades de l'emissions de CO2 en tonelades per càpita en 27 països d'Europa.

Dades originals $x_1, \ldots, x_{27}$

8.3	10.4	6.4	9.5	11.2	8.8	13.1	10.7	5.7	9.8	8.3	5.3	7.2	3.5
4.2	21.3	6.2	10.8	7.8	4.9	9.9	4.2	6.7	8.3	7.0	5.0	8.3	

Llista ordenada $x_{(1)}, \ldots, x_{(27)}$

3.5	4.2	4.2	4.9	5.0	5.3	5.7	6.2	6.4	6.7	7.0	7.2	7.8	8.3
8.3	8.3	8.3	8.8	9.5	9.8	9.9	10.4	10.7	10.8	11.2	13.1	21.3	

Mínim, màxim, mediana i rang interquartílic

Mínim $x_{(1)}$

Màxim $x_{(n)}$

Mediana Valor situat al centre de la llista ordenada, $x_{(n/2)}$

Quartils

- Primer quartil: $Q_1 = x_{(n/4)}$
- Segon quartil: $Q_2 = Me$
- Tercer quartil: $Q_3 = x_{(3n/4)}$

Rang Interquartílic Mesura de dispersió

$$\boxed{IQR = Q_3 - Q_1}$$

Quantils i percentils

Per cada valor q entre 0 i 1, el quantil q de les dades és aquell valor situat en la posició qn de la llista ordenada. Aquest valor també s'anomena el percentil $100q$. Denotarem per $Pe_x(q)$.

Percentils per dades agrupades

$$Pe_x(q) = L_i + (L_{i+1} - L_i)\frac{qn - N_{i-1}}{n_i}$$

Exemple (Emissions CO2)

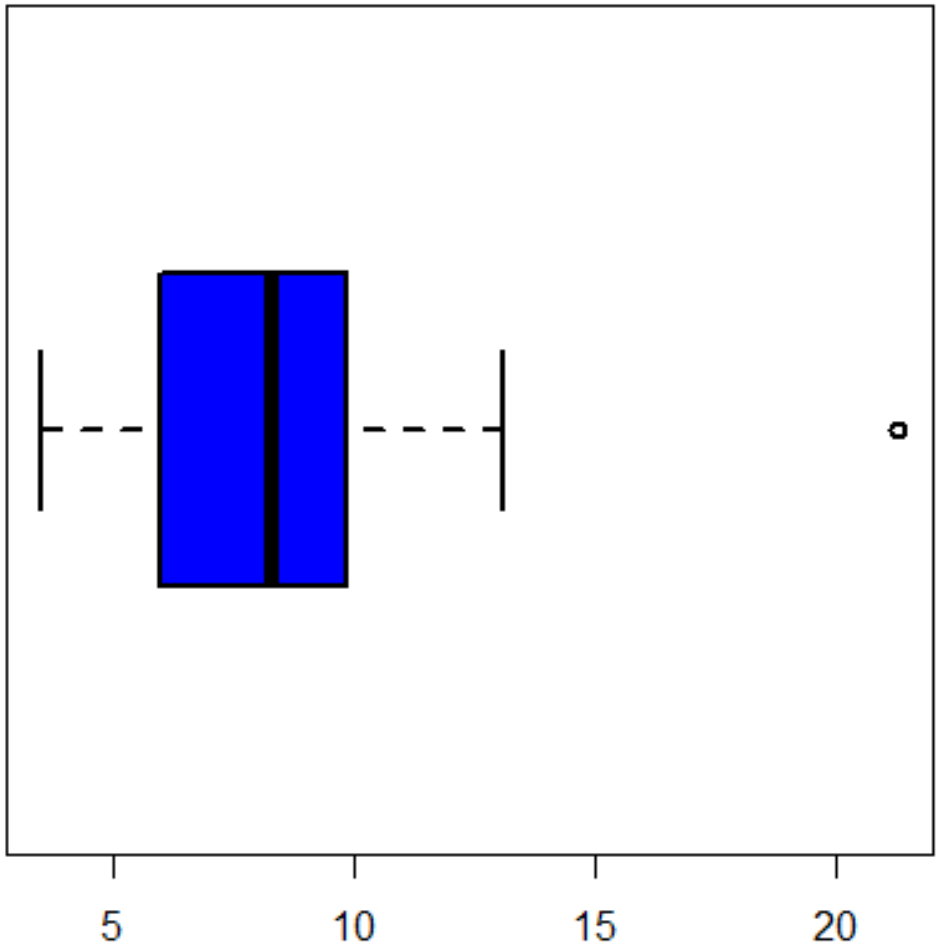
Els quartils

0%	25%	50%	75%	100%
3.50	5.95	8.30	9.85	21.30

Alguns quantils

10%	30%	50%	70%	80%	90%
4.62	6.36	8.30	9.56	10.30	10.96

Boxplot o Diagrama de caixa

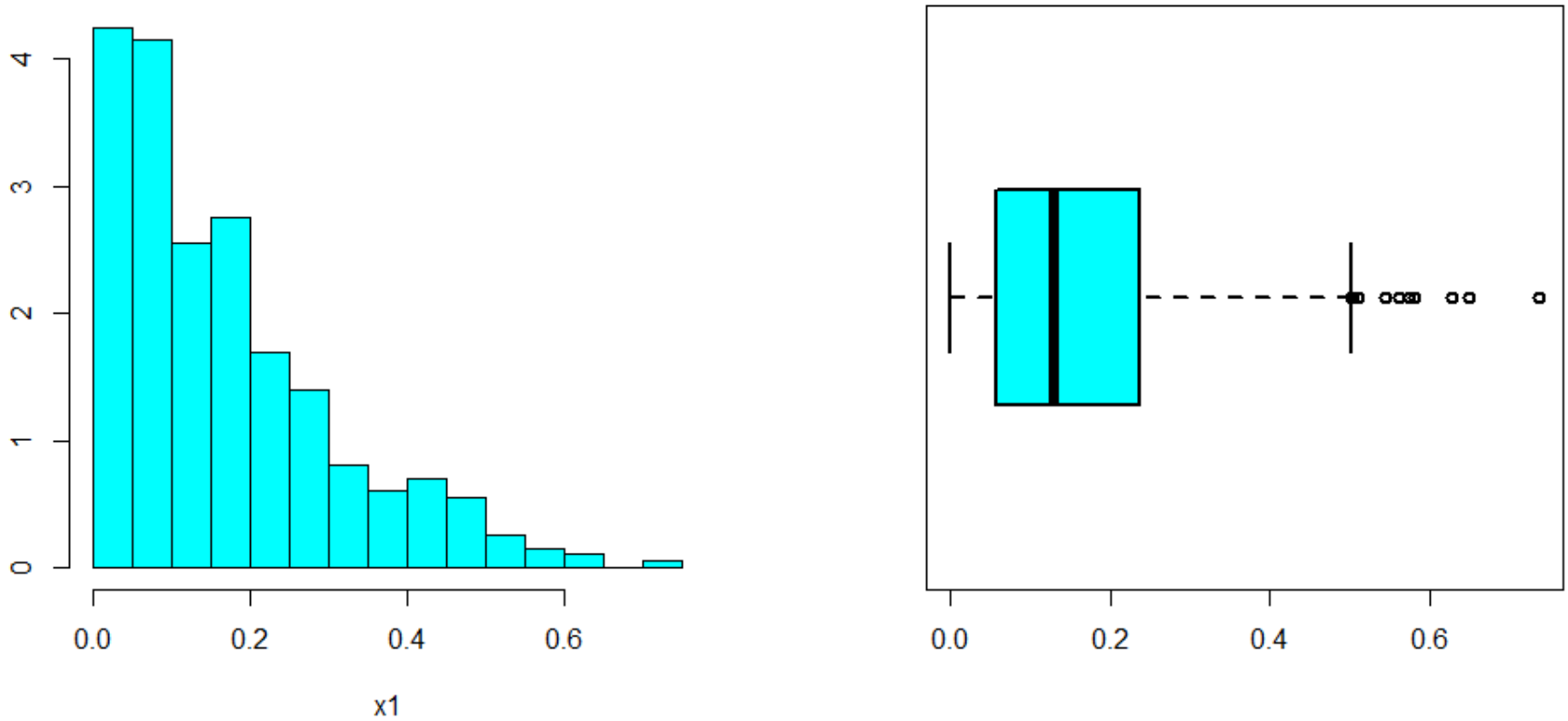


Coeficient d'asimetria

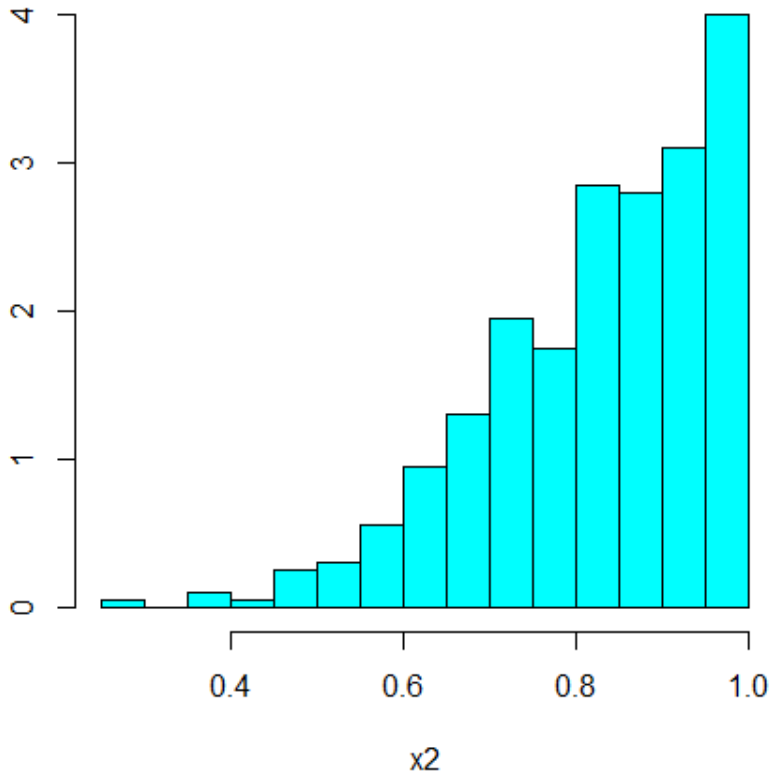
$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s_x^3}$$

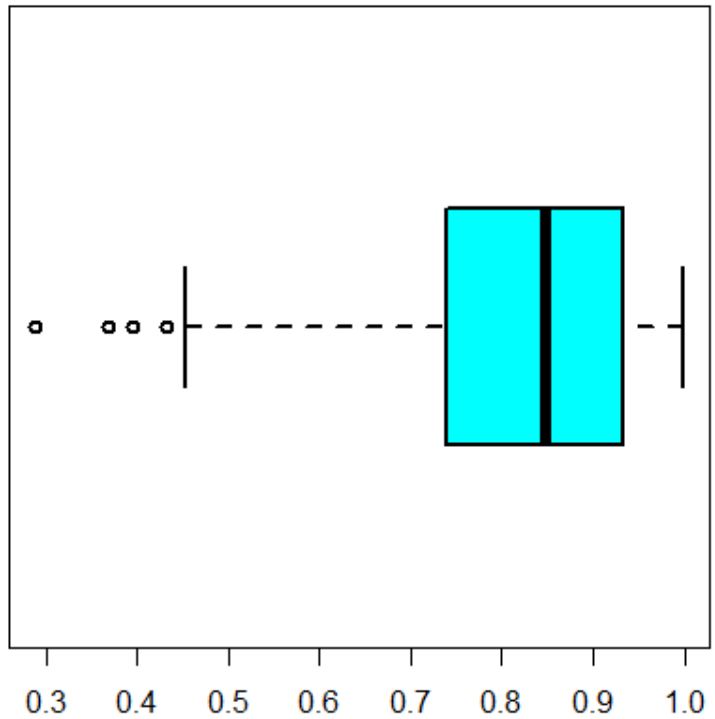
- Asimetria positiva (cap a la dreta): més proporció de valors petits.
- Asimetria negativa (cap a l'esquerra): més proporció de valors grans.
- Si les dades són simètriques val 0.

Gràfics amb asimetria positiva $As = 1.073313$



Gràfics amb asimetria negativa $As = -0.9086109$



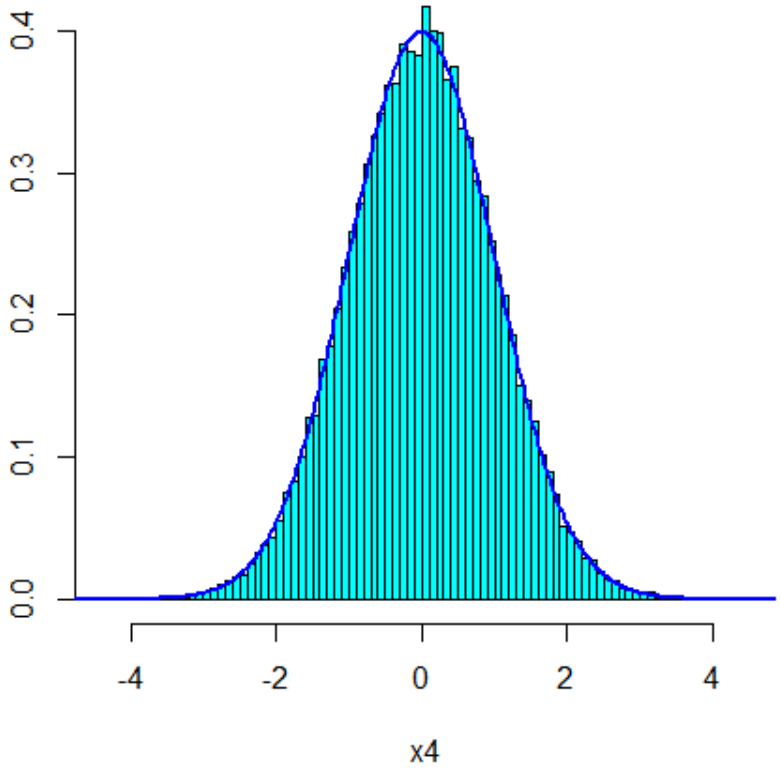


Coeficient de curtosi o mesura d'apuntament

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s_x^4} - 3$$

- Indica el grau d'apuntament de les nostres dades.
- Dades normals valor 0.
- Si la corba és més plana que una campana de Gauss serà inferior a 0.
- Si la corba és més apuntada que una campana de Gauss serà superior a 0.
- Només vàlid per a dades simètriques.

Gràfics exemple curtosi ($Cu_1 = 0.01978$ i $Cu_2 = 8.8324$)



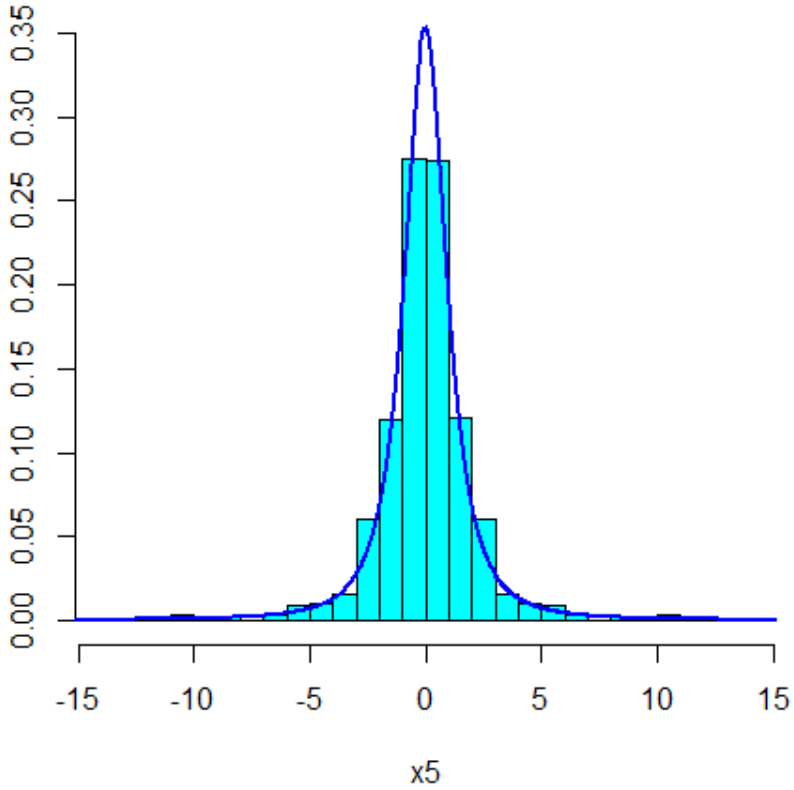
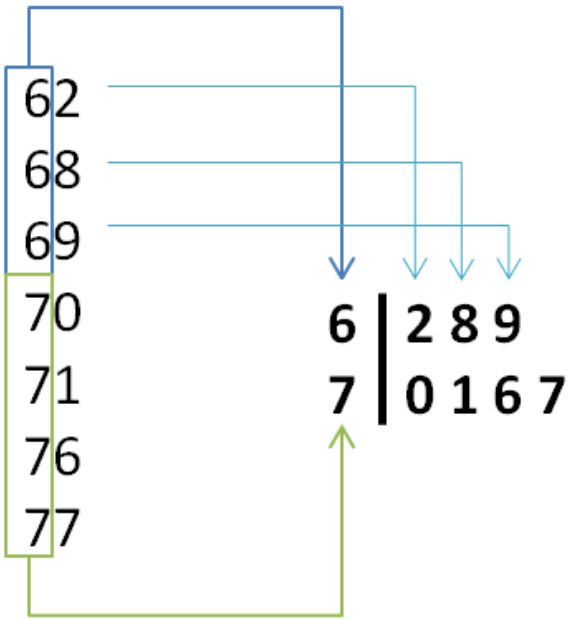
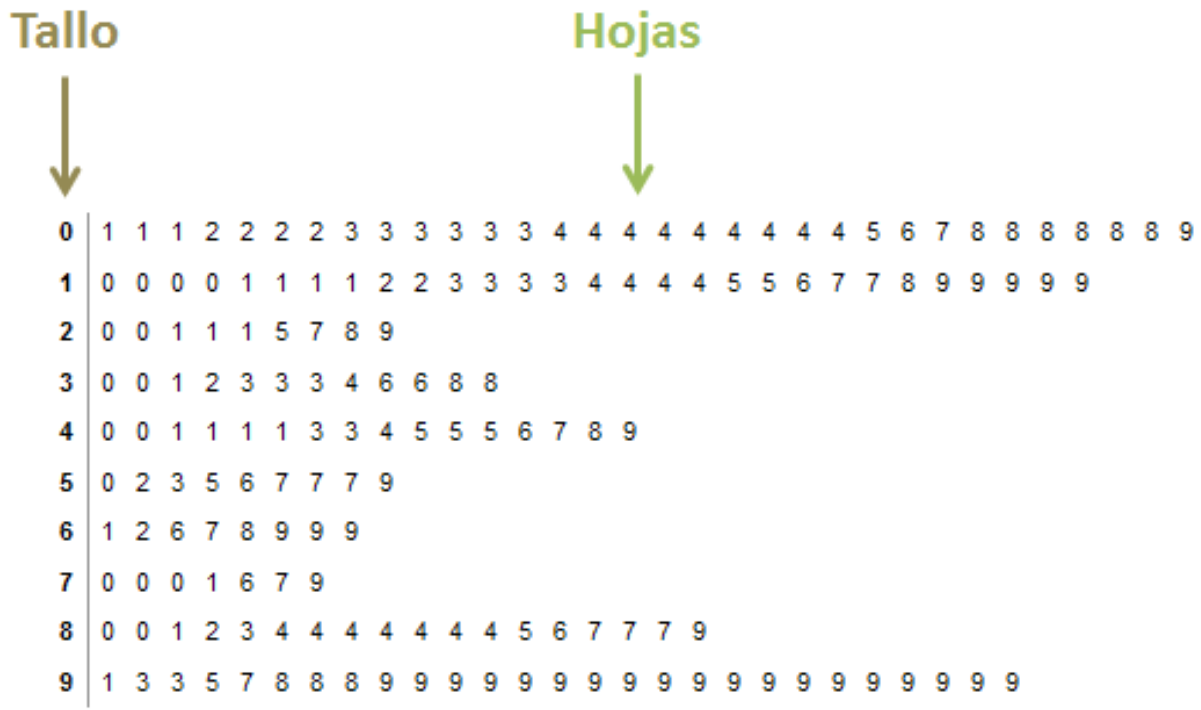


Diagrama de tija i fulles





Exercici

Per qualsevol conjunt de dades $x_1, \ldots, x_n$ amb mitjana $\bar{x}$ i desviació típica s_x, si considerem $y_i = K_1 \cdot x_i + K_2$, com queden modificats els següents estadístics:

mitjana, mediana, moda, variància, desviació típica, percentils, coeficient de curtosi i coeficient d'asimetria.