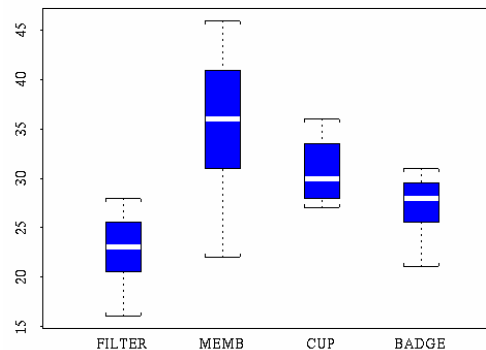
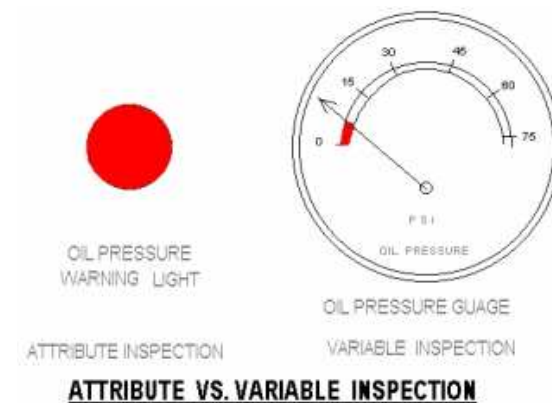
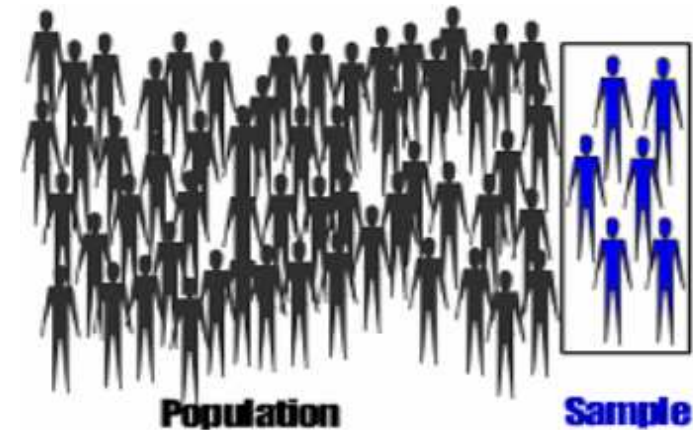


ESTADÍSTICA APLICADA

T1 Anàlisi Exploratòria de Dades



- **Població:** conjunt d'elements objecte de l'estudi
 - Poblacions finites i infinites
- **Mostra:** subconjunt de la població
 - Mostra aleatòria i grandària mostral
- **Variable:** propietat dels elements de la població
 - Variable qualitativa
 - Variable quantitativa (discretes i contínues)
- **Observació:** valor que agafa la variable en un element de la població



T1 – 1.3: Distribuciones unidimensionales

- **Freqüència absoluta** (n_i): nombre de vegades que surt un valor determinat x_i a les n observacions de la variable X
- **Distribució de freqüències**: $\{(x_i, n_i) / i = 1, 2, \dots, k\}$, i.e.: conjunt format pels diferents valors de la variable, x_i , i la freqüència associada a cada valor, n_i
- **Freqüència relativa**, f_i : $f_i = n_i / n$
- **Freqüències acumulades**, N_i i F_i

Exemple 5

Variable número de hermanos	Frecuencia n_i	Frecuencia acumulada N_i	Frecuencia relativa f_i	Frecuencia relativa acumulada F_i
0	36	36	0.40	0.40
1	25	61	0.28	0.68
2	18	79	0.20	0.88
3	7	86	0.08	0.96
4	3	89	0.03	0.99
5	1	90	0.01	1.00

Suma = 90

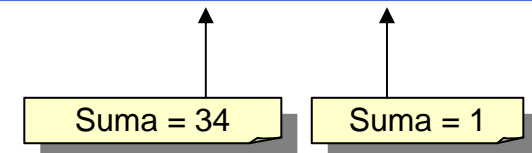
Suma = 1

(Continua)

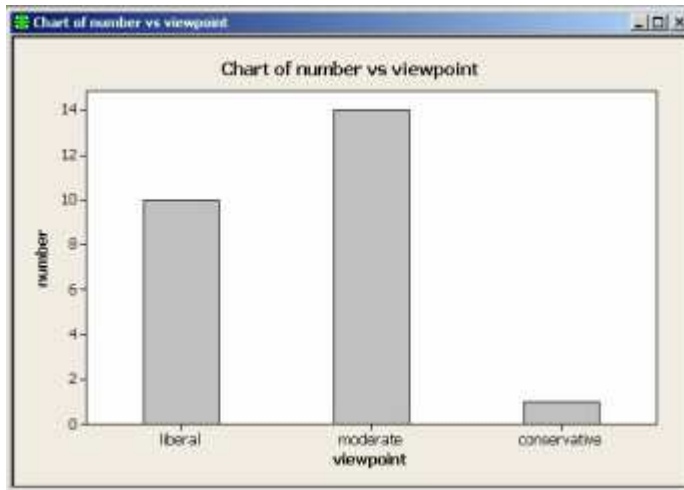
- Si hi ha moltes observacions o la variable és contínua, s'agrupen les dades en **interval de classe**
- **Freqüència d'una classe**
- **Marca de classe**: valor mitjà de cada interval de classe
- **Distribució de freqüències**

Exemple 6

x_i	Marcas de classe	n_i	N_i	f_i	F_i
[2.90-3.00)	2.95	1	1	0.03	0.03
[3.00-3.10)	3.05	1	2	0.03	0.06
[3.10-3.20)	3.15		2	0.00	0.06
[3.20-3.30)	3.25	8	10	0.24	0.29
[3.30-3.40)	3.35	4	14	0.12	0.41
[3.40-3.50)	3.45	5	19	0.15	0.56
[3.50-3.60)	3.55	3	22	0.09	0.65
[3.60-3.70)	3.65	6	28	0.18	0.82
[3.70-3.80)	3.75	1	29	0.03	0.85
[3.80-3.90]	3.85	5	34	0.15	1.00



T1 – 1.4: Representacions gràfiques



← Diagrama de barres (bar chart)

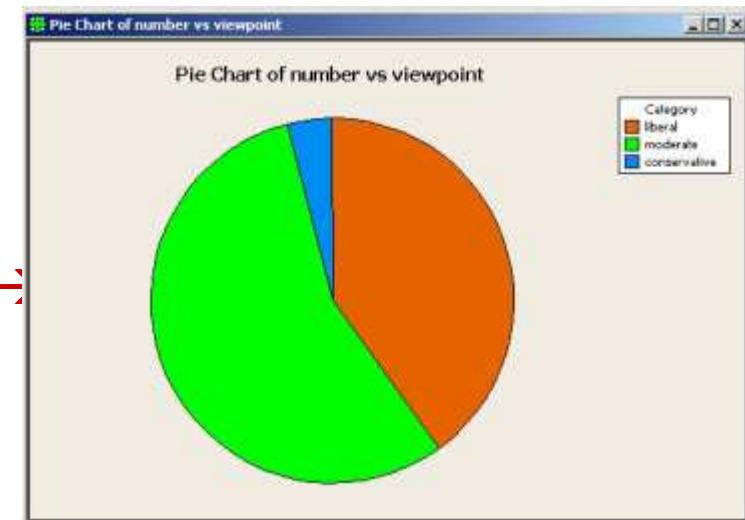
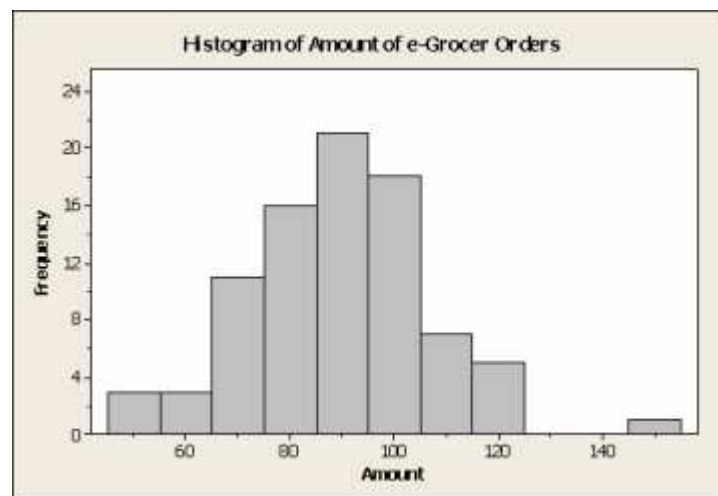


Diagrama de sectors (pie chart) →



← Histogrames de freqüències

- Donada una mostra de n observacions, x_1, x_2, \dots, x_n , com es distribueixen les dades? (ús de descriptors de **centralització** i **dispersió**)
- **Estadístics** d'una mostra: \bar{x}, s, \dots
- **Paràmetres** d'una població: μ, σ, \dots

- Mesures de tendència central:

μ ■ **Mitjana** (mean): $\bar{x} = \frac{1}{n} \sum x_i$

Molt sensible a dades atípiques!

- **Moda**

■ **Mediana** (median): $Q2 = \begin{cases} x_{\frac{n+1}{2}} & \text{si } n \text{ senar} \\ \text{mitjana entre } x_{\frac{n}{2}} \text{ i } x_{\frac{n}{2}+1} & \text{si } n \text{ parell} \end{cases}$

- Mesures de posició:

■ **Quartil.les**: $Q1 = LQ = x_{\frac{n+1}{4}}$ $Q3 = UQ = x_{\frac{3(n+1)}{4}}$

(Continua)

Ejemplo 15: **Mediana.**

Los resultados del ejemplo 14, de un estudio de control de calidad de hormigón, ordenados de menor a menor, son:

8 dades (parell) ordenades

51; 128; 154; 174; 191; 199; 209; 280 Kp/cm²

y la mediana será:

$$\frac{174 + 191}{2} = 182.50 \text{ Kp} / \text{cm}^2$$

Con los datos del ejemplo 5 (número de hermanos de los alumnos de una clase), la mediana es **1 hermano**.

(Continua)

- Mesures de dispersió:

- Rang o amplitud: $R = M\grave{a}x. - M\grave{i}n.$

- Rang interquartílic (IQR): $IQR = Q3 - Q1$

- Desviació mitjana: $\frac{1}{n} \sum |x_i - \bar{x}|$

- Variància: $s_n^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 = \frac{\sum x_i^2}{n} - \bar{x}^2 \equiv Var(X)$

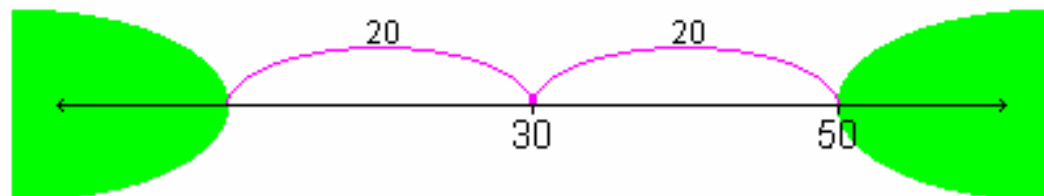
- Variància corregida (Minitab): $s_{n-1}^2 = \frac{n}{n-1} s_n^2$

- Desviació tipus (StDev): $s = +\sqrt{s^2}$

- **Desigualtat de Txebyshev:** “és molt poc freqüent trobar dades que estiguin molt allunyades de la mitjana; tant menys freqüent quan més ens allunyem d'aquesta mitjana”

Sigui k un enter, llavors:

$$\text{freqüència relativa } (|x_i - \bar{x}| \leq k \cdot s) \geq 1 - \frac{1}{k^2}$$



- **Boxplot:** permet caracteritzar la distribució (simetria, dispersió, etc.) i detectar dades atípiques (*outliers*)
- **Construcció d'un Boxplot:**
 1. Trobar: $Q1, Q2, Q3$ i $IQR = Q3 - Q1$
 2. Calcular: $LI = Q1 - 1.5 \cdot IQR$ i $LS = Q3 + 1.5 \cdot IQR$
 3. Una dada és atípica si està fora de l'interval $[LI, LS]$
 4. Dibuixar una línia des de cada extrem del rectangle central fins el valor més llunyà no atípic

