2.Estadística Descriptiva Univariant

Estadística
Grau en Matemàtiques

Josep A. Sanchez
Dept. Estadística i I.O.(UPC)

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Estadística Descriptiva Univariant

En funció de la resposta, les variables aleatòries es classifiquen en:

- **Qualitatives o categòriques**
  - **Nominals** (Tipus de dieta)
  - **Ordinals** (Grau d'una determinada malaltia)
- Quantitatives
  - **Discretes** (nombre d'errades en compilar un programa, nombre dedefectes d'una peça)
  - **Continues** (pes, temperatura, temps)

Alerta: Les variables qualitatives a voltes poden semblar quantitatives perquè es codifiquen amb un nombre.

Observació: Una variable categòrica amb pocs nivells s'anomena **Factor**.

Exercici

De que tipus són les següents variables?

- *Speed of cars registered at a certain point on a highway*
- *Number of cigarettes smoked by someone on one day*
- *Type of enterprise (S.L. or S.A.)*
- *Number of employees in a company*
- *Type of job at a university*
- *Weight of an animal*
- *Annual revenue of an enterprise*
- *Gender of a person*
- *Bank account number*

Estadística Descriptiva

L'Estadística Descriptiva té dues parts:

- **Resums Numèrics**
  - Taules de freqüència per a variables categòriques
  - Càlcul d'Estadístics per a variables numèriques
- **Representacions gràfiques**

Resums Numèrics (v. categòriques)

Taula de freqüències:

- Freqüència absoluta de la categoria C_i: $n_i = \#\{x_j | x_j = C_i\}$
- Freqüència relativa de la categoria C_i: $f_i = \frac{n_i}{n}$

En cas de variables categòriques ordinals, té sentit calcular també:

- Freqüència absoluta acumulada de la categoria C_i: $N_i = \#\{x_j | x_j \leq C_i\}$
- Freqüència relativa acumulada de la categoria C_i: $F_i = \frac{N_i}{n}$

Exemple de taula de freqüències

```
## Sample: X={ D,B,E,D,A,E,D,B,B,D,D,D,E,D,C,D,E,B,E,B }
```

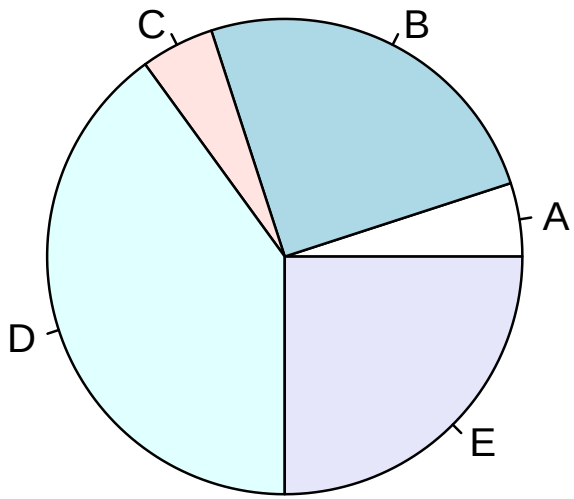
```
data.frame(cbind(ni=table(X),
      fi=prop.table(table(X)),
      Ni=cumsum(table(X)),
      Fi=cumsum(prop.table(table(X)))))
```

```
##   ni   fi Ni   Fi
## A  1 0.05  1 0.05
## B  5 0.25  6 0.30
## C  1 0.05  7 0.35
## D  8 0.40 15 0.75
## E  5 0.25 20 1.00
```

Representacions Gràfiques (v. categòriques)

- Diagrama de Sectors o de Pastís (*Pie Chart*): per variables nominals amb poques categories

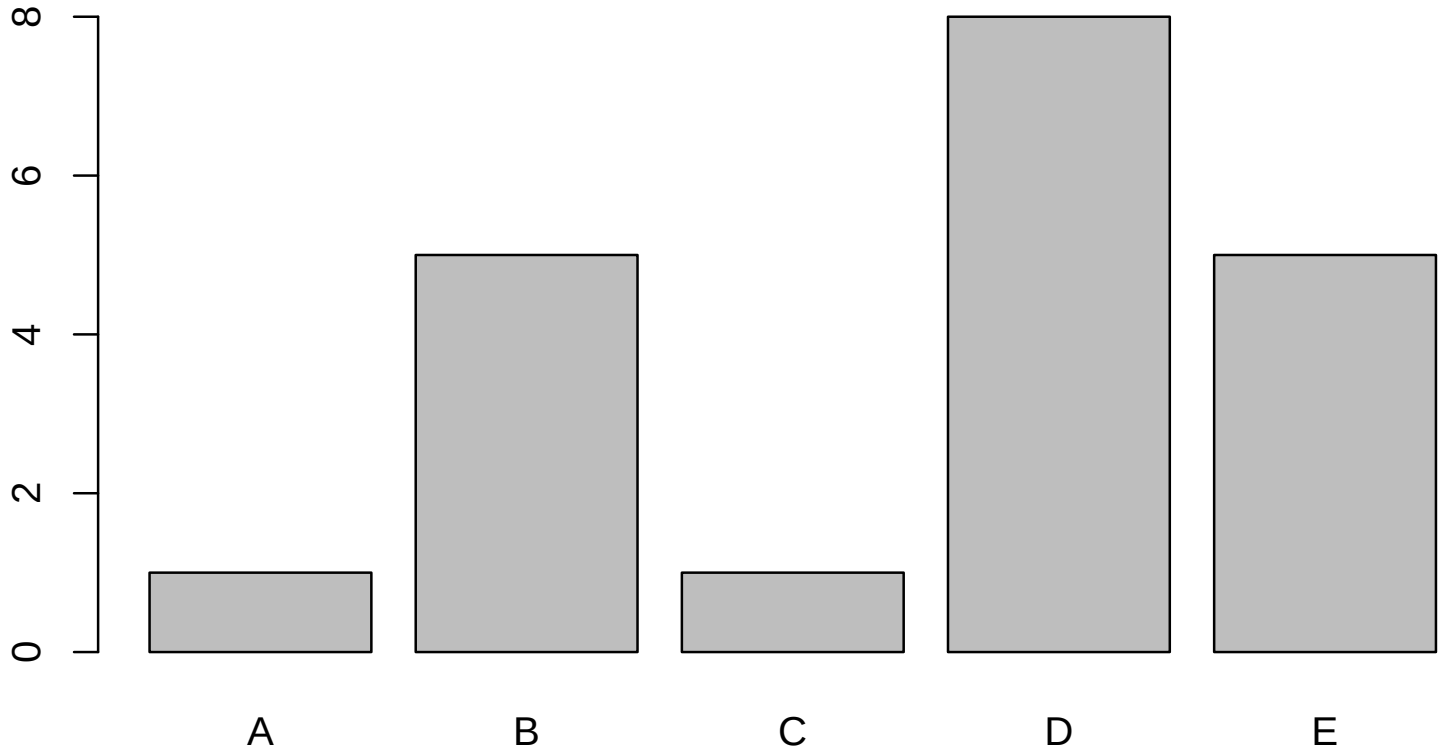
```
pie(table(X))
```



Representacions Gràfiques (v. categòriques)

- Diagrama de Barres horizontal o vertical (*Bar plot*): per ordinals o nominals amb moltes categories

```
barplot(table(X))
```



Resums Numèrics (v. numèriques)

Càlcul d'estadístics a partir de la mostra

- Tipus d'estadístics:
  - de Tendència Central
    - Mitjana, mediana, mitjana retallada (*trimmed mean*),...
  - de Posició
    - Quartils, decils, percentils...
  - de Dispersió
    - Desviació Estàndar, Variància, Rang, IQR, Coeficient de Variació,...
  - de Forma
    - Coeficient d'Asimetria, Kurtosi,...

Estadístics de Tendència Central

```
## Sample:X={ 15,4,11,14,8,4,5,15,3,1,6,5,14,13,19,13,3,3,20,7 }
```

Mitjana (*mean, average*): $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$

```
mean(X)
```

```
## [1] 9.15
```

Mediana (*median*): $X_{((n+1)/2)}$ n imparell, $\frac{X_{(n/2)}+X_{(n/2+1)}}{2}$ n parell

```
median(X)
```

```
## [1] 7.5
```

Mitjana retallada al α%($\alpha \in [0, 0.5]$): Es treu els $(\alpha/2)$% inferior i superior de la mostra i es calcula la mitjana.

```
mean(X,trim=0.1)
```

```
## [1] 8.75
```

Estadístics de Tendència Central

- La mitjana és sensible a la presència de dades atípiques, però fa servir tota la informació de la mostra
- La mediana és molt robusta si hi ha dades atípiques, però no fa servir massa informació de la mostra (només dels valors centrals)
- La mitjana retallada és un compromís entre les dues anteriors mesures: és robusta i fa servir més informació de la mostra que la mediana
  - La mitjana retallada al 0% $= \bar{X}$
  - La mitjana retallada al 50% $=$ Median(X)

Estadístics de Posició

Es basen en els estadístics d'ordre $x_{(i)}$

Quartils: valors que separen la mostra ordenada en quatre parts

- Q_1: Primer quartil, valor on un 25% de la mostra és inferior i un 75% superior

$$Q_1 = x_{([n/4])}$$

- Q_2: Segon quartil, valor on un 50% de la mostra és inferior i un 50% superior (coincideix amb la Mediana)

$$Q_2 = x_{([n/2])}$$

- Q_3: Tercer quartil, valor on un 75% de mostra és inferior i un 25% superior

Estadístics de Dispersió

Variància: $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$; **Desviació Estàndard**: $S = \sqrt{S^2}$

```
c(var(X),sd(X))
```

```
## [1] 34.028947  5.833434
```

Rang: $R = x_{(n)} - x_{(1)}$

```
diff(range(X))
```

```
## [1] 19
```

Rang interquartílic: $IQR = Q_3 - Q_1$

```
IQR(X)
```

```
## [1] 10
```

Coeficient de Variació: $CV = \frac{S}{\bar{X}}$

Estadístics de Forma

Coeficient d'asimetria: $\gamma = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{S^3}$

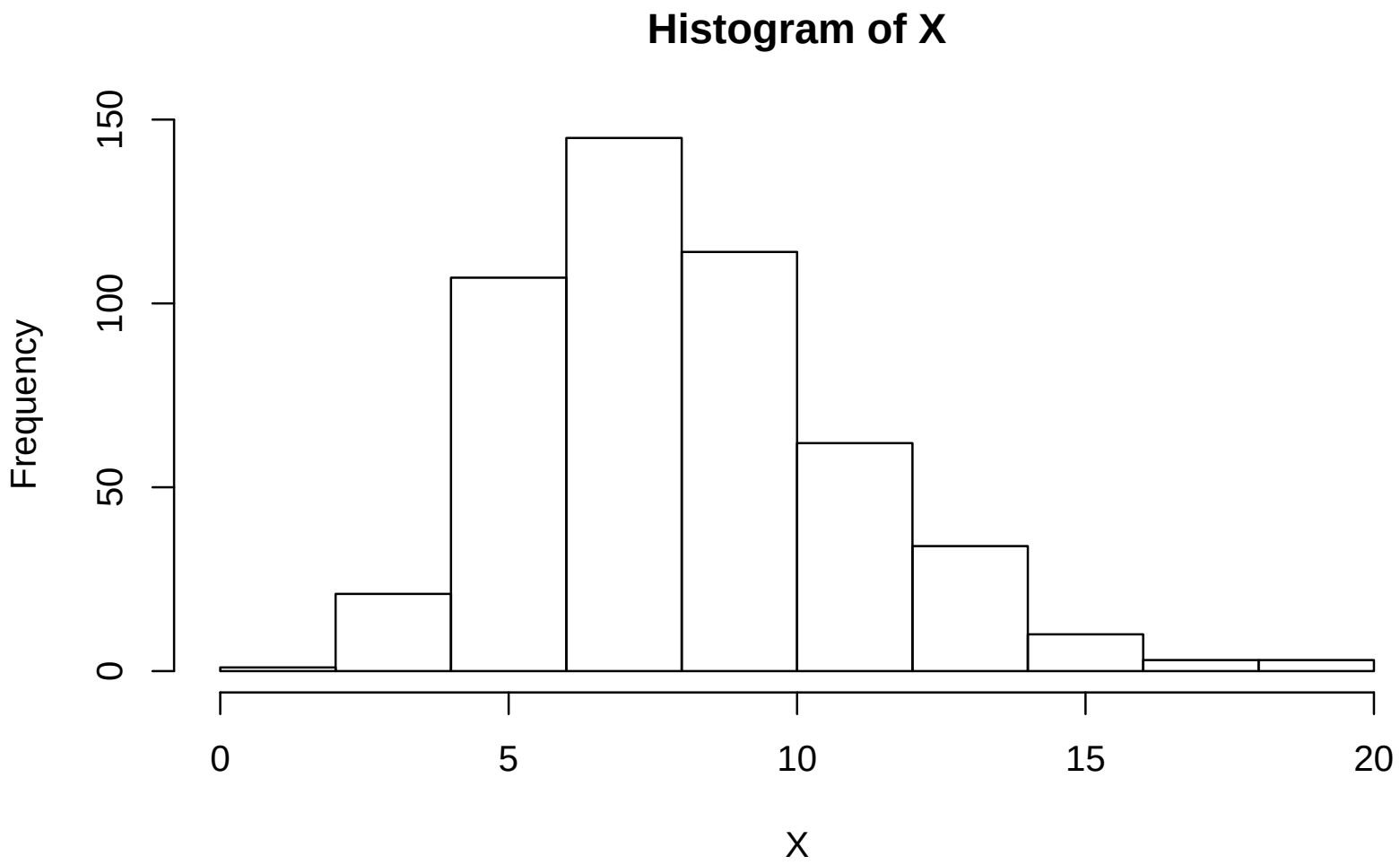
Kurtosi: $\kappa = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{S^4}$

- leptocúrtica, $\kappa > 3$: más apuntada y con colas más gruesas que la normal.
- platicúrtica, $\kappa < 3$: menos apuntada y con colas menos gruesas que la normal.
- mesocúrtica, $\kappa = 3$:: cuando tiene una distribución normal.

Representacions Gràfiques (v. numèriques)

- Histograma

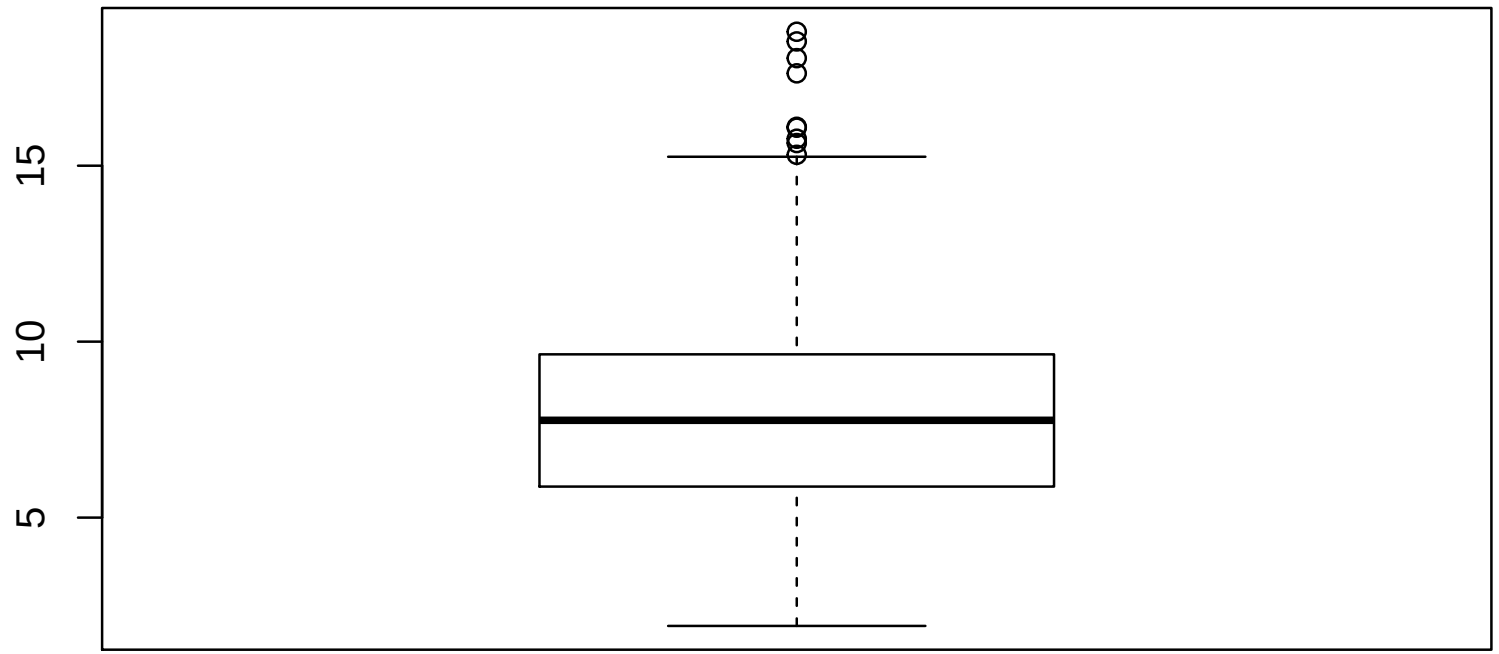
```
hist(X)
```



Representacions Gràfiques (v. numèriques)

- Diagrama de Caixa (*Box-Plot*)

```
boxplot(X)
```



Representacions Gràfiques (v. numèriques)