

Estadística descriptiva amb R

Univariant i bivariant


Estadística Descriptiva (ED)

➤ **Objectiu:** Descriure (estadísticament) ...

- DE univariant: ...les variables d'una mostra d'una en una.
 - Hi ha un 45.3% de dones a la mostra (variable gènere).
- DE bivariant: ...les relacions existents entre dues variables en una mostra
 - Entre els homes, hi ha un 8% més d'usuaris de Linux que entre les dones (variables gènere i SO).

➤ **Metodologia:** Hi ha dos tipus d'eines per fer la descriptiva

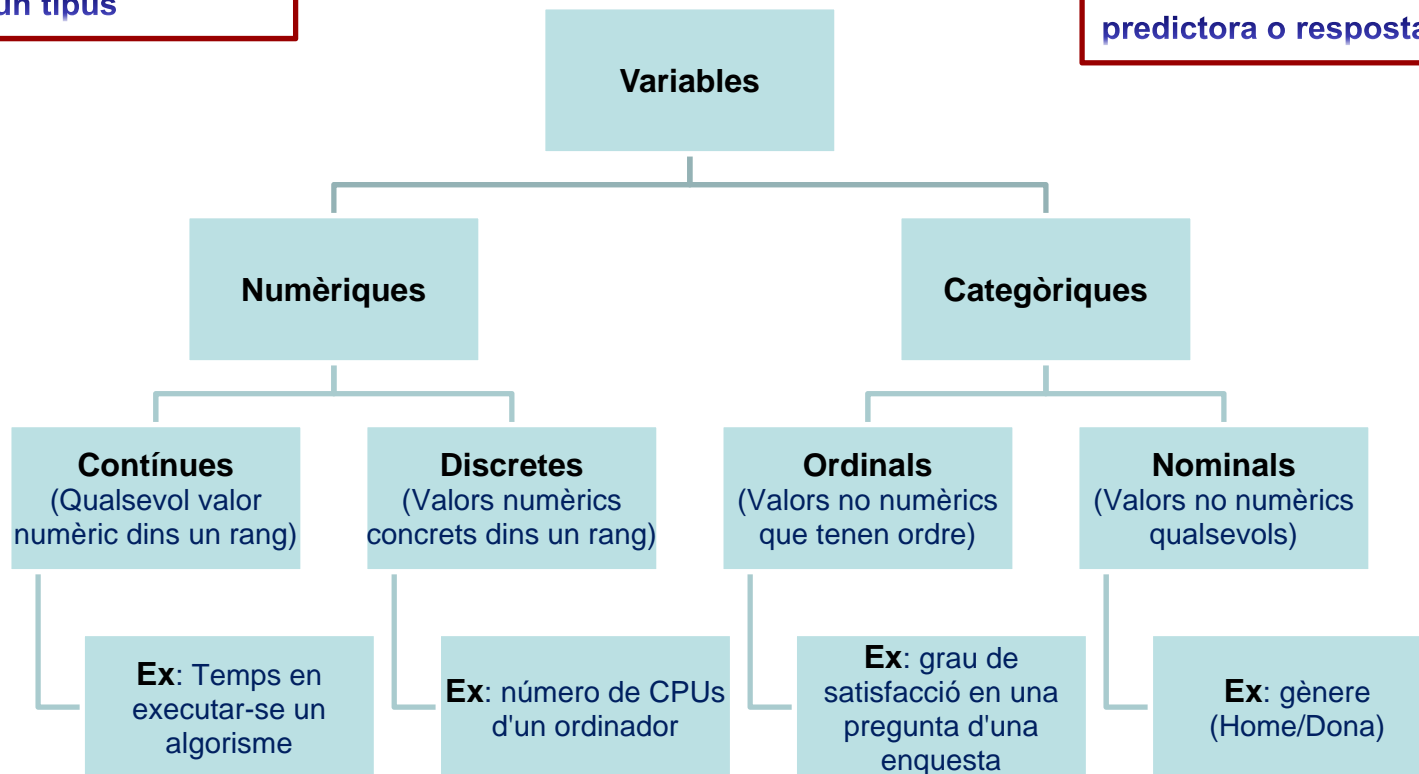
- Numèriques
 - EX: mitjana, mediana, desviació estàndard, ..
- Gràfiques
 - EX: histograma, boxplot, diagrama de barres..

 **Per saber quines
eines emprar, s'ha de
conèixer el tipus de
les variables
(diapositiva següent)**

Tipus de variables

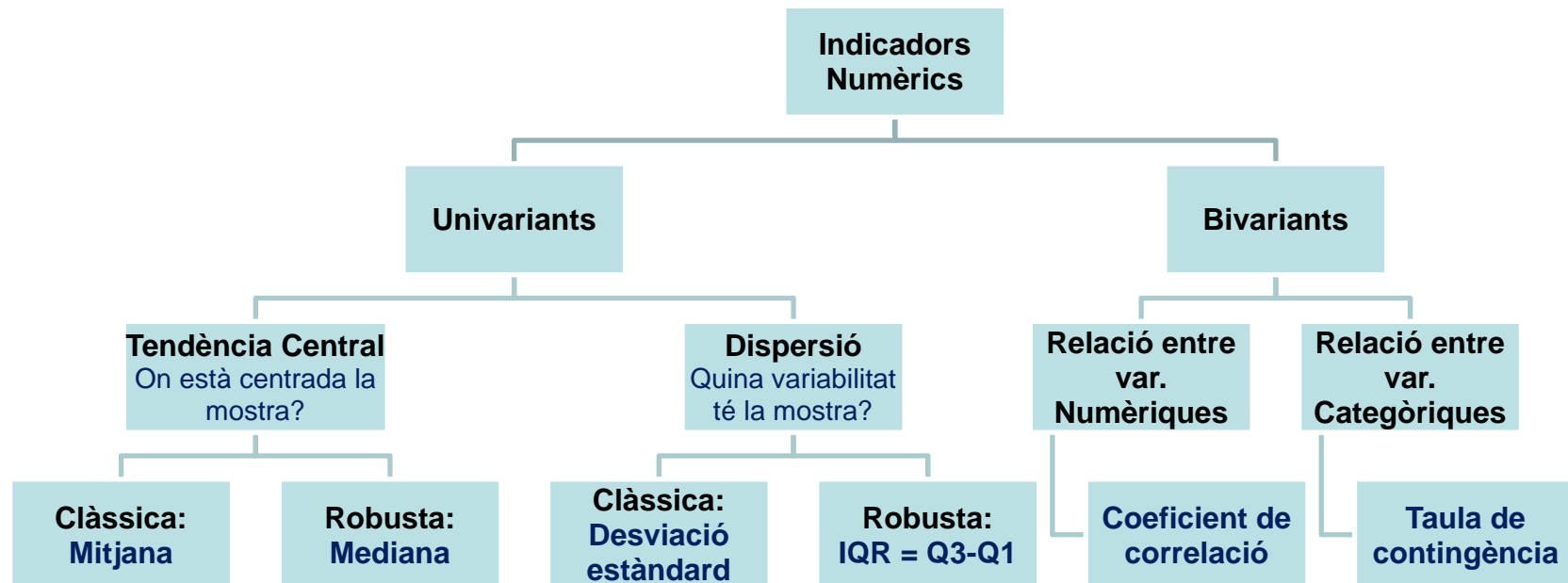
👍 Una mostra de valors pot correspondre's amb més d'un tipus

👍 Hi ha més classificacions possibles, com per exemple, segons el rol de la variable: predictora o resposta



Eines numèriques

👍 Els indicadors univariants, sobretot tenen sentit per variables numèriques contínues



Dades (I)

- En aquest joc de diapositives, usarem unes dades sobre la potència de la connexió ADSL domèstica (recollides per estudiants)
- Cada alumne participant va recollir, entre altra, la següent informació:
 - Proveïdor ADSL
 - Velocitats contractades (pujada/baixada) en Mbps
 - Velocitat real (pujada/baixada) (<http://www.internautas.org/testvelocidad/>)
 - Distància a la central (a <http://www.adslnet.es/distancia-adsl>)
 - Si l'estudiant era de barcelona
 - Tipus de connexió (Wifi/Cable)

Dades (II)

➤# Lectura de les dades (les guardem en un objecte anomenat adsl)

```
➤adsl<-read.table(url("http://www-eio.upc.es/teaching/pe/Dades/dades_ADSL.txt"),
                  header=T,na.strings=c("00",NA),dec=',')
```

➤dim(adsl) # Nombre d'observacions (41) i de variables (18)

```
[1] 41 18
```

➤names(adsl) # Noms de les variables

```
[1] "id" "grp" "down.speed" "up.speed" "latency" "dia"
[7] "hora" "dist.central" "proveedor" "veloc.cont.up" "veloc.cont.down" "cable.o.wifi"
[13] "ciudad" "is.BCN" "Ratio.up" "Ratio.down" "log.obs.down" "log.cont.down"
```

➤head(adsl) # Capçalera de les dades (6 primeres observacions)

	id	grp	down.speed	up.speed	latency	dia	hora	dist.central	proveedor	veloc.cont.up	veloc.cont.down
1	1	12	2589	125	105	22/03/10	16:53	1058	Ono	NA	NA
2	2	11	2522	256	154	22/03/10	18:08	774	Telefónica	256	3
3	3	41	411	211	157	22/03/10	20:32	3698	Telefónica	256	1
4	4	13	1088	313	128	23/03/10	17:06	4709	Orange	256	1
5	5	43	849	193	128	23/03/10	21:04	871	Telefónica	256	1
6	6	11	5027	261	126	23/03/10	21:31	835	Telefónica	320	6

Càlcul dels indicadors clàssics

➤ `adsl$down.speed`

Visualitzar la variable `down.speed` dins d'`adsl`

```
[1] 2589 2522 411 1088 849 5027 2563 5095 2546 2560 2444
[12] 2581 4190 3875 12808 4656 7839 6845 2144 2544 2619 5071
[23] 2559 2596 5123 6393 3751 6126 4930 2546 1886 8688 2436
[34] 8613 2526 2495 7018 2308 3457 10589 5233
```

Mitjana

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 4198.5$$



n: longitud de la mostra

x_i : observació i-èsima

Variància

$$S_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = 716778$$

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

Desviació tipus o estàndard

$$S_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = 2677.26$$

$$S_x = \sqrt{\frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - n(\bar{x})^2 \right]}$$



La variància i la desviació tipus són mesures de la variabilitat que té la mostra. La primera és el quadrat de la segona.

Càlcul dels indicadors clàssics amb R

```
➤ x <- adsl$down.speed           # La nostra variable d'interès l'anomenem x
➤ n <- length(x)                 # Longitud de la mostra

➤ sum(x)/n                       # Càlcul de la mitjana
[1] 4198.512

➤ mean(x)                       # Càlcul directe de la mitjana
[1] 4198.512

➤ (sum(x^2)-n*mean(x)^2)/(n-1)   # Càlcul de la variància
[1] 7167785

➤ var(x)                        # Càlcul directe de la variància
[1] 7167785

➤ sqrt((sum(x^2)-n*mean(x)^2)/(n-1)) # Càlcul de la desviació tipus
[1] 2677.272

➤ sd(x)                         # Càlcul directe de la desviació tipus
[1] 2677.272
```



sum: suma d'un conjunt de valors
sqrt: arrel quadrada (square root)

Càlcul dels indicadors robusts

```
➤ sort(adsl$down.speed)    # Visualitzar la variable down.speed endreçada
```

[1]	411	849	1088	1886	2144	2308	2436	2444	2495	2522	2526
[12]	2544	2546	2546	2559	2560	2563	2581	2589	2596	2619	3457
[23]	3751	3875	4190	4656	4930	5027	5071	5095	5123	5233	6126
[34]	6393	6845	7018	7839	8613	8688	10589	12808			

Quartil 1

Posició Q1 = $(n+1)/4 = (41+1)/4 = 10.5$

$$Q1 = (X_{(10)} + X_{(11)})/2 = (2522 + 2526)/2 = 2524$$

Mediana (Quartil 2)

Posició Q2 = $(n+1)/2 = (41+1)/2 = 21$

$$Q2 = X_{(21)} = 2619$$

Quartil 3

Posició Q3 = $3 \cdot (n+1)/4 = 31.5$

$$Q3 = (X_{(31)} + X_{(32)})/2 = (5123 + 5233)/2 = 5178$$

Rang Interquartílic

$$IQR = Q3 - Q1 = 5178 - 2524 = 2654$$

👍 La mediana (Q2) és el valor de la mostra que deixa el 50% de les observacions per sota i l'altre 50% per sobre. El Q1 és el valor que deixa el 25% de les observacions per sota i el Q3 és el que deixa el 75% de les observacions per sota

👍 Si el càlcul de la posició dóna un nombre enter, llavors el quartil corresponent serà el nombre que ocupi aquella posició en la llista de valors endreçats (de menor a major).

👍 Si el càlcul dóna un nombre no enter, es ponderaran els valors corresponents a la posició entera anterior i posterior segons la part decimal de la posició.

Càlcul dels indicadors robusts amb R


```
➤ x.ord <- sort(x)                # x.ord seran els valors endreçats de x


➤ pos.Q1 <- (n+1)/4                # Càlcul de la posició del Q1
➤ Q1 <- (x.ord[10]+x.ord[11])/2    # Càlcul del Q1
➤ quantile(x,0.25,type=6)         # Càlcul directe del Q1
[1] 2524

➤ pos.Q2 <- (n+1)/2               # Càlcul de la posició de la mediana (Q2)
➤ Q2 <- x.ord[21]                 # Càlcul de la mediana (Q2)
➤ quantile(x,0.50,type=6)         # Càlcul directe de la mediana (Q2)
[1] 2619

➤ pos.Q3 <- 3*(n+1)/4             # Càlcul de la posició del Q3
➤ Q3 <- (x.ord[31]+x.ord[32])/2    # Càlcul del Q3
➤ quantile(x,0.75,type=6)         # Càlcul directe del Q3
[1] 5178

➤ iqr <- Q3 - Q1                  # Càlcul del IQR
➤ IQR(x,type=6)                  # Càlcul directe del Q3
[1] 2654
```

 La mediana també es pot calcular directament amb *median(x)*

 R té 9 formes diferents de calcular els quantils. La que s'explica en les diapositives es correspon amb el *type=6*

Càlcul dels indicadors numèrics per grups amb R

➤ De vegades, és vol descriure una variable estratificada segons una altra variable categòrica. Això es pot fer amb la instrucció ***tapply***

➤ ***Sintaxi:*** `tapply(var_int, var_cat, fun, ...)`

- `var_int`: variable numèrica o categòrica d'interès per la qual es vol fer alguna descriptiva
- `var_cat`: variable categòrica per la qual és vol estratificar
- `fun`: funció que es vol aplicar a la variable d'interès (`var_int`)
- `...`: altres paràmetres de la funció

➤ **Exemples:**

➤# Mitjana de la velocitat de baixada segons tipus de connexió

➤`tapply(ads1$down.speed, ads1$cable.o.wifi, mean)`

```
Cable      Wifi
4202.870 4192.944
```

➤# Descriptiva global de la velocitat de baixada segons si s'és de Barcelona

➤`with(ads1,tapply(down.speed, is.BCN, summary))`

```
$`0`
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
411	2524	2619	4157	5049	12810

```
$`1`
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1088	2557	3846	4329	5903	8688

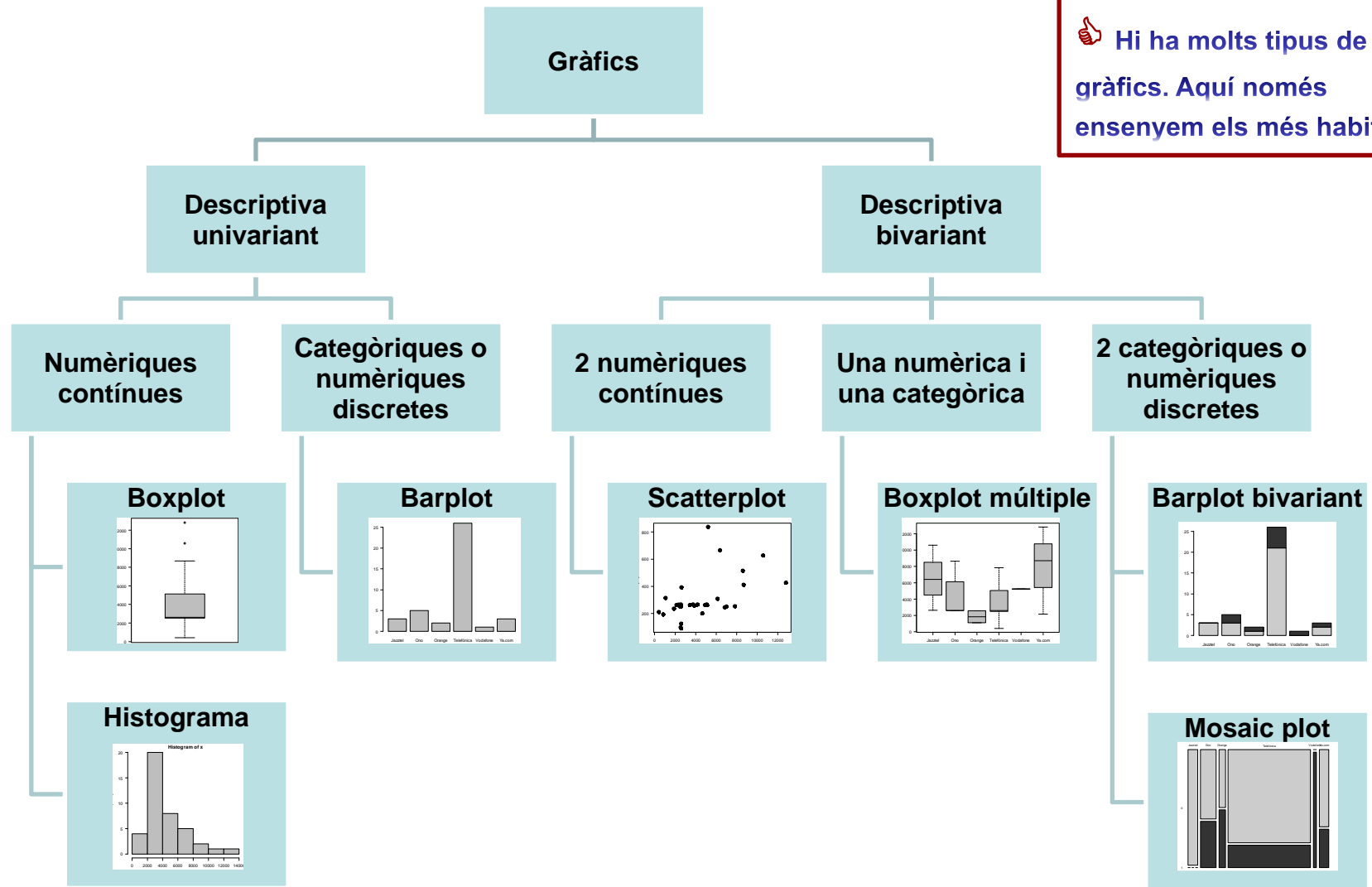


La instrucció *summary* proporciona els indicadors robustos y la mitjana



La instrucció *with* permet escriure els noms de les variables directament

Eines gràfiques



👍 Hi ha molts tipus de gràfics. Aquí només ensenyem els més habituals

Diagrama de caixa (Boxplot)

- Representa els indicadors **robustos** i els **outliers**
- Elements:
 - **Caixa**. Està delimitada pel Q1 i pel Q3 i té una línia interior que representa la mediana.
 - **Bigotis**. surten des de la caixa i tenen una longitud de 1.5 vegades el IQR (IQR=longitud de la caixa). En cas de que el mínim de la mostra sigui major que el final del bigoti esquerra, aquest bigoti només arribarà fins al mínim. De forma anàloga es procedeix amb l'altra bigoti.
 - **Outliers**. Són els punts que queden més enllà dels bigotis del box-plot. Es consideren dades anòmales.

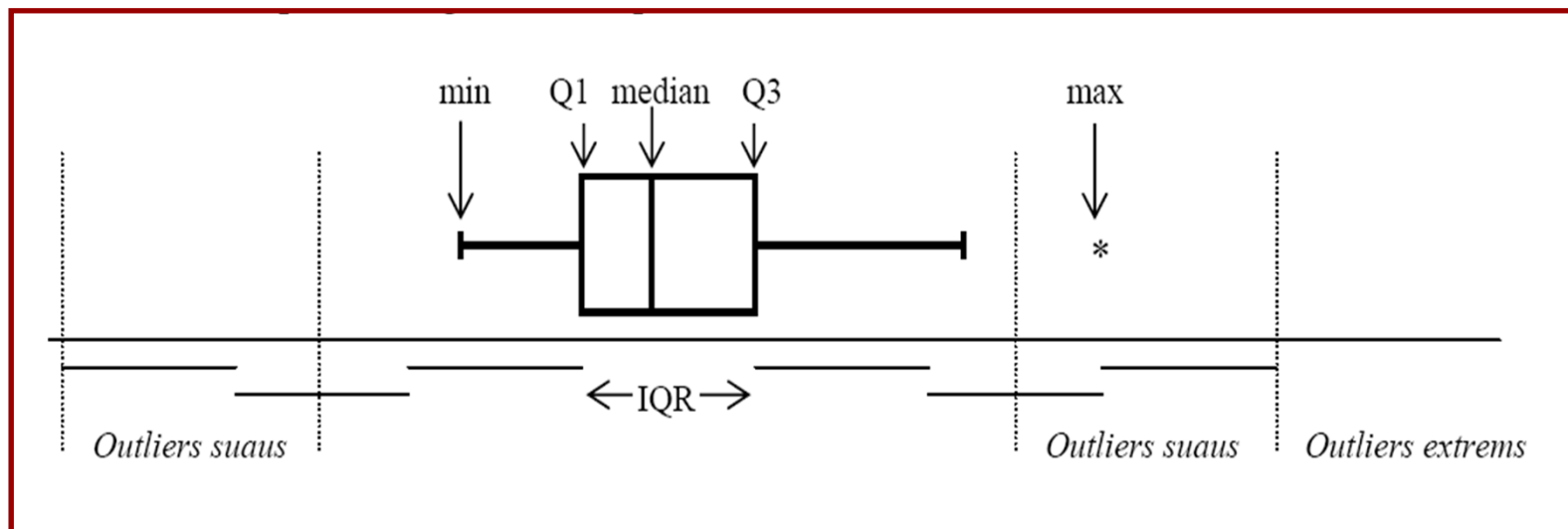


Diagrama de caixa (Boxplot) amb R

Números dels eixos sempre en horitzontal

➤ `par(las=1)`

Boxplot univariant de la variable `x` (`adsl$down.speed`)


➤ `boxplot(x)`

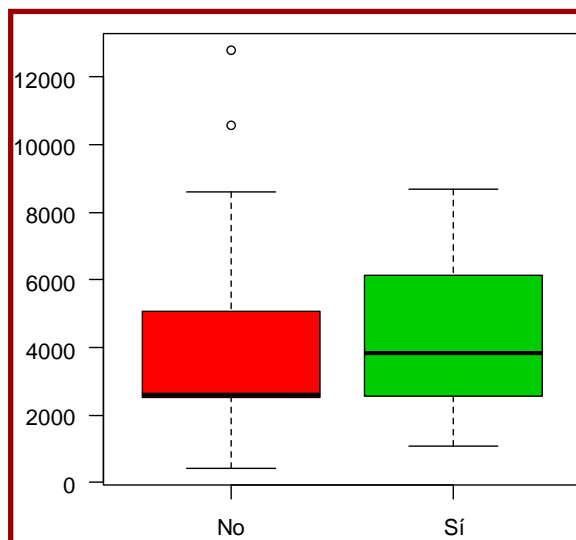
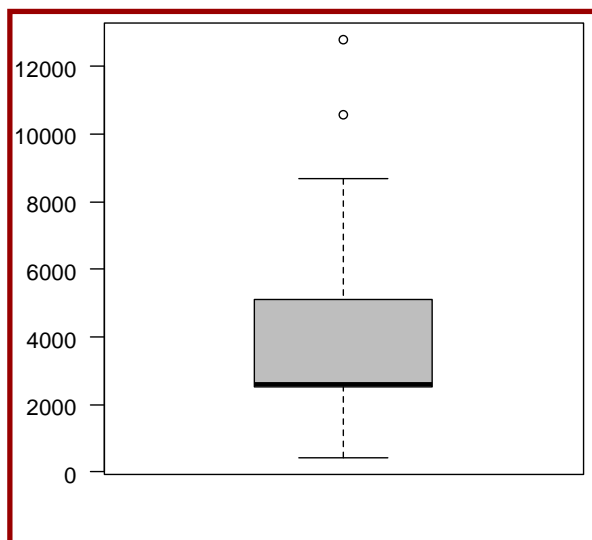
➤ `boxplot(x,col="grey",main="Vel.Baixada")` # Li afegim un color i un títol


Boxplot bivariant de la variable `down.speed` en funció de la variable `is.BCN`

➤ `boxplot(down.speed~is.BCN,data=adsl)`

➤ `boxplot(down.speed~is.BCN,data=adsl,col=2:3,names=c("No","Sí"))`

 La instrucció *par* serveix per fixar certs paràmetres gràfics
El paràmetre *las* indica la direcció dels nombres dels eixos



 Comprova si la representació del boxplot de l'esquerra quadra amb els valors calculats prèviament.

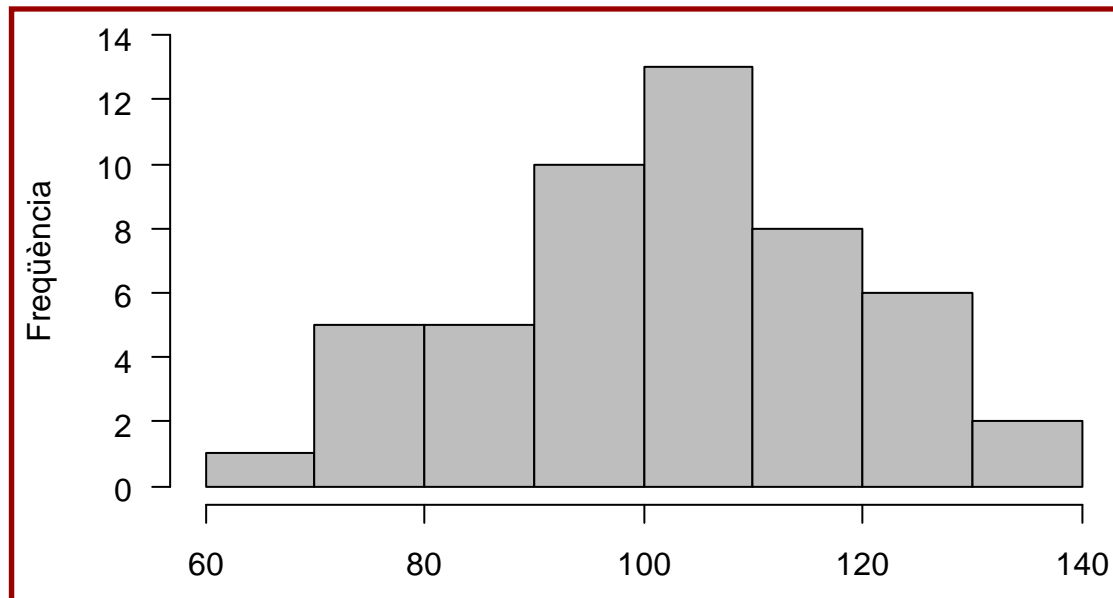
 Hi ha 2 outliers representats amb punts

 Els colors 2 i 3 fan referència als colors vermells i verd

 Per veure més opcions del gràfic fer *?boxplot*

Histograma

- Representa la **distribució** de la variable estudiada
- L'**eix horitzontal** conté els **valors** de la variable i l'**eix vertical** les **freqüències** (o proporcions)
- La **superfície de cada barra** és proporcional a la freqüència dels valors representats



👍 En aquest histograma,
entre 90 i 100 hi ha 10 valors

Histograma amb R

Obrim una finestra preparada per col·locar 2 gràfics

➤ `par(mfrow=c(1,2))`



mfrow serveix per posar més d'un gràfic en la mateixa finestra. El primer valor és el nº de files i el segon el nº de columnes

Histograma univariant de les velocitats de baixada i pujada

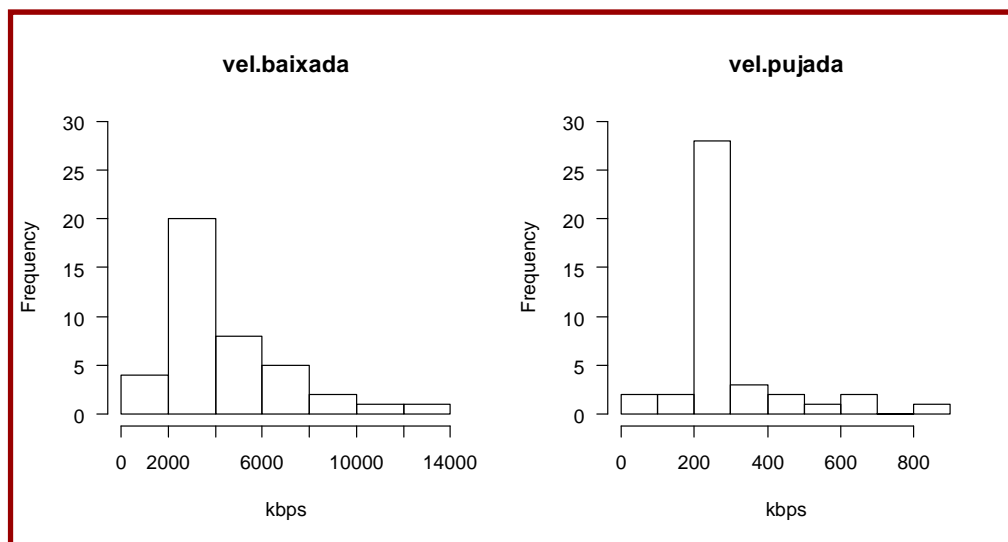
➤ `hist(adsl$down.speed)`

➤ `hist(adsl$up.speed)`

Histogrames millorats (títols, etiquetes, mateixa escala)

➤ `hist(adsl$down.speed, main="vel.baixada", xlab="kbps", ylim=c(0,30))`

➤ `hist(adsl$up.speed, main="vel.pujada", xlab="kbps", ylim=c(0,30))`



xlab defineix l'etiqueta de l'eix horitzontal



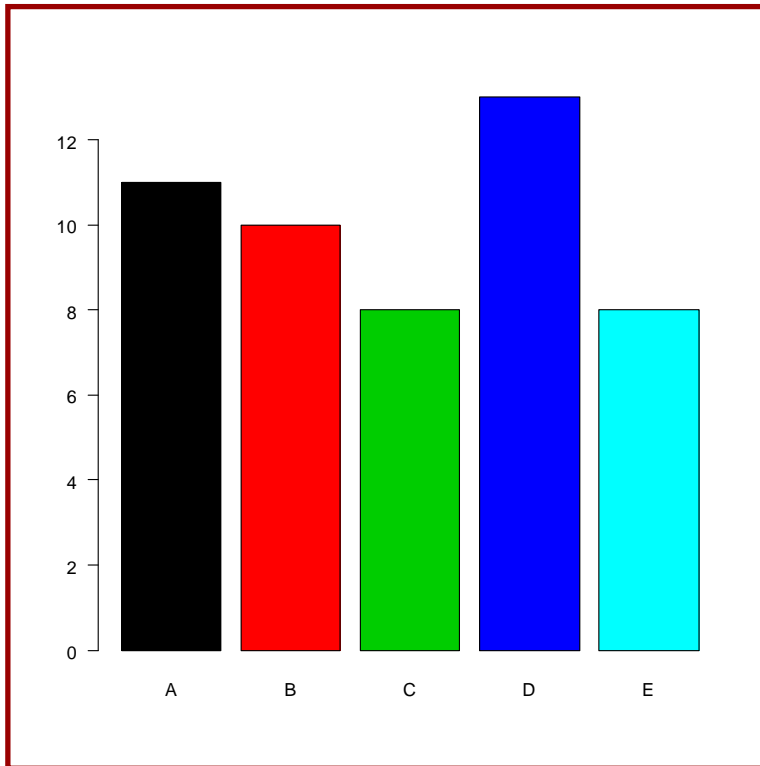
ylim defineix els límits de l'eix vertical



Per veure més opcions del gràfic fer *?hist*

Diagrama de barres (Barplot)

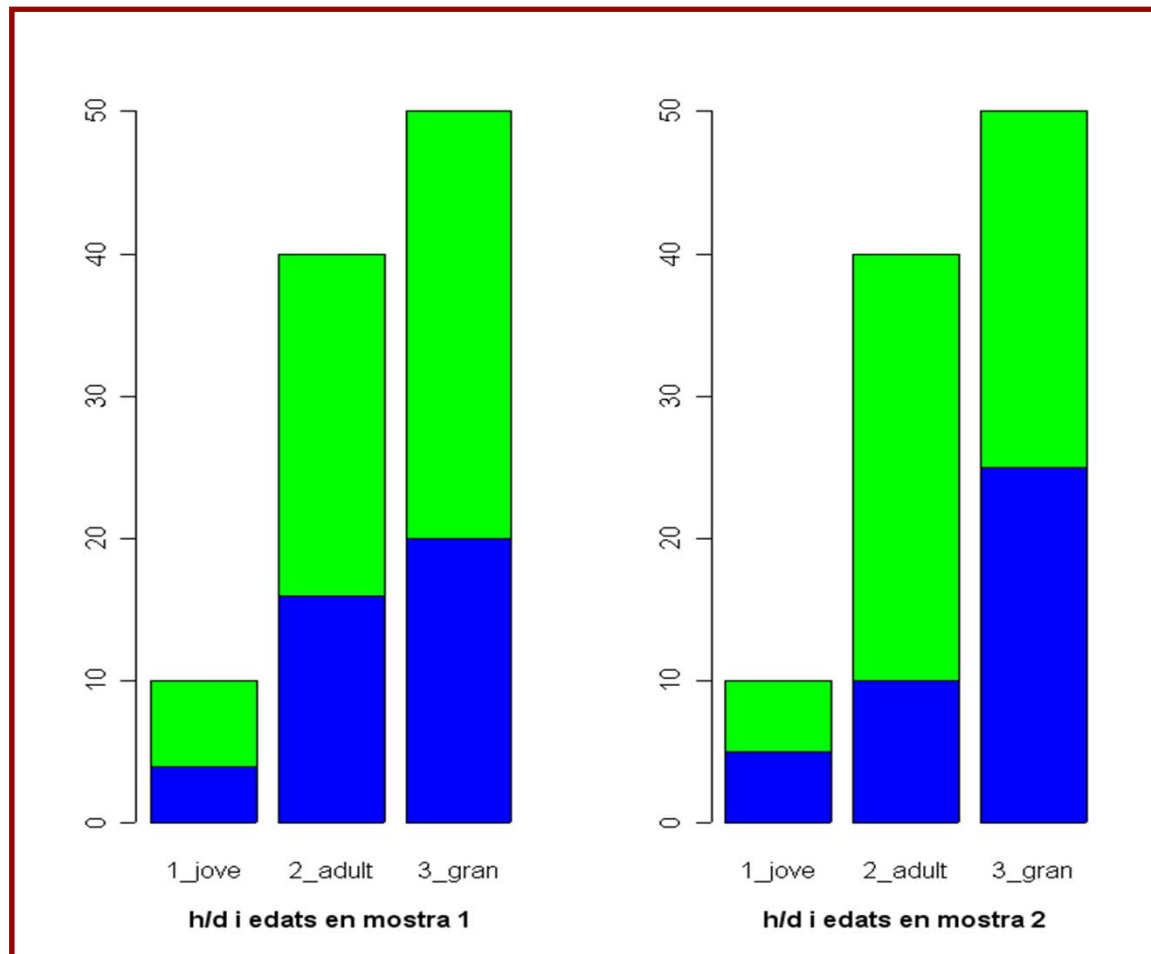
- Representa el nombre d'efectius per a cada categoria
- L'eix horitzontal conté les categories de la variable i l'eix vertical les freqüències (o proporcions)



A diferencia de l'histograma, el *barplot*, conté un espai entre cada barra

Diagrama de barres (Barplot) - Interpretació

Exemple de dos barplots bivariants en dues mostres, d'una variable d'edat amb 3 categories relacionada amb el gènere



👍 A l'esquerra, es manté la mateixa proporció de h i d en les tres franges d'edat

👍 A la dreta, no es manté la mateixa proporció de h i d en la barra d'adults respecte les altres dues

Diagrama de barres (Barplot) amb R

```
# Fixem finestra amb 2 gràfics (mfrow=c(1,2)), ticks perpendiculars
```

```
# als eixos (las=2) i en negreta i cursiva (font.axis=4)
```

```
➤ par(mfrow=c(1,2), las=2, font.axis=4)
```

 *font.axis* defineix la font dels 'ticks' dels eixos

```
# Barplot univariant dels proveïdors
```

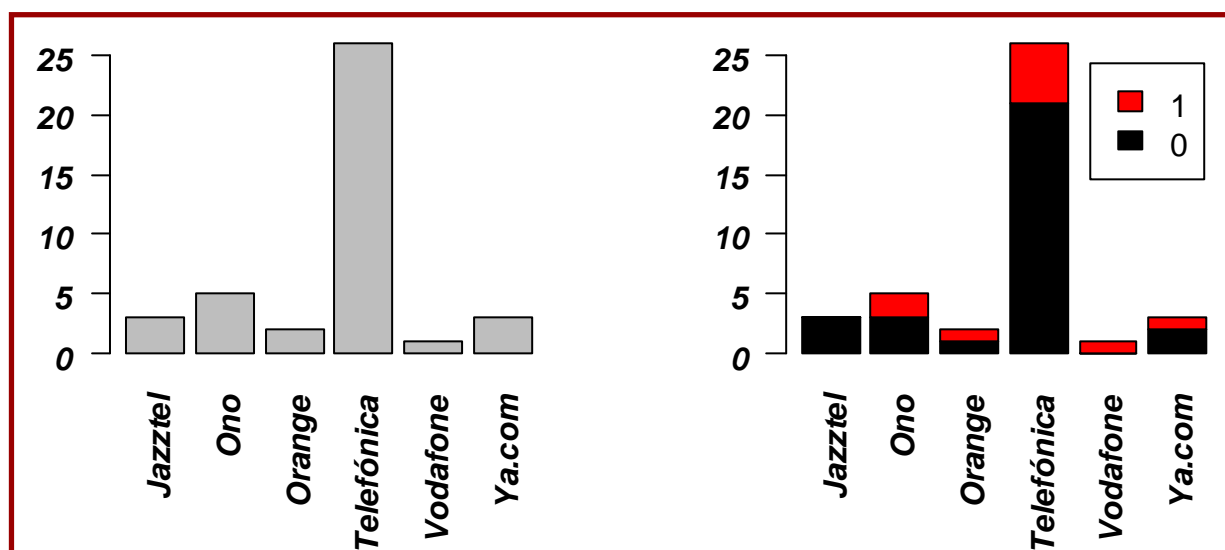
```
➤ (t.prov <- table(adsl$proveedor))
```


```
➤ barplot(t.prov)
```

```
# Barplot bivariant dels proveïdors i si els alumnes són de Barcelona
```

```
➤ (t.prov2 <- table(adsl$sis.BCN, adsl$proveedor))
```

```
➤ barplot(t.prov2, col=1:2, legend=TRUE)
```



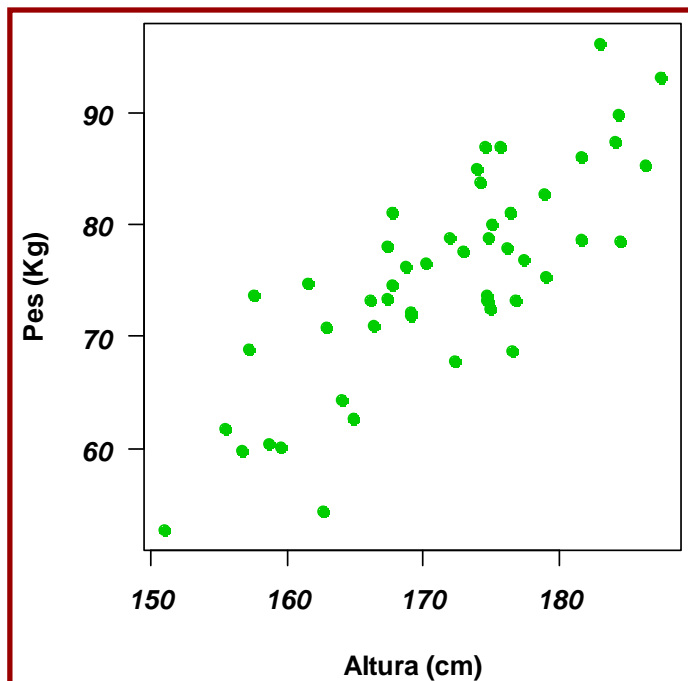
 Els parèntesis a l'inici i final d'una assignació permeten visualitzar el resultat de l'assignació per pantalla.

 A la funció barplot sempre se li passa una taula (uni o bivariant)

 Per veure més opcions del gràfic fer `?barplot`

Diagrama de dispersió (Scatterplot)

- Representa la distribució bivariant de dues variables numèriques.
- L'eix horitzontal conté els valors de la variable explicativa i l'eix vertical, els de la variable resposta.

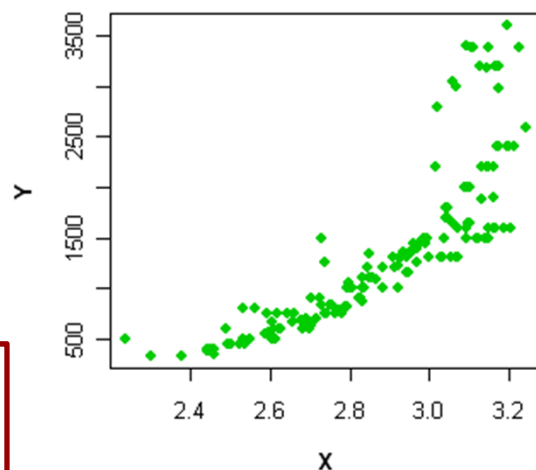


Ens interessa veure quin tipus de relació
existeix entre ambdues variables

Diagrama de dispersió (Scatterplot) - Interpretació

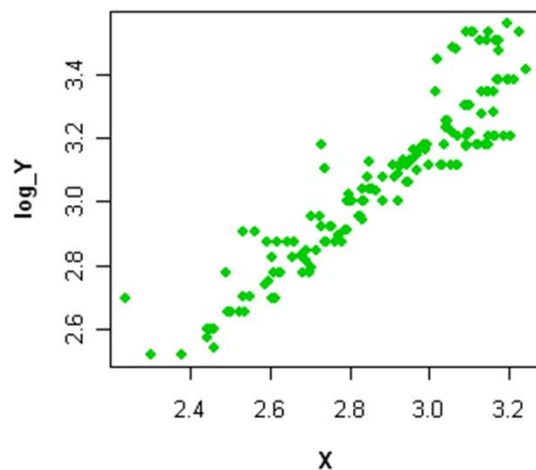
Exemples de relacions entre una variable explicativa (X) y les variables resposta: Y, Y', log_Y i Y'')

*Relació no lineal,
creixent i
força intensa*

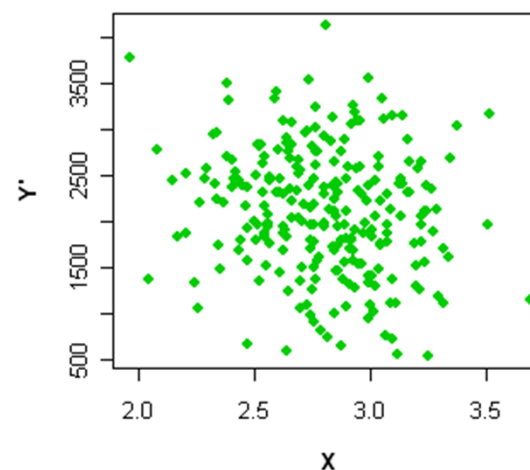


👍 Transformacions com
fer el logaritme permeten
linealitzar.

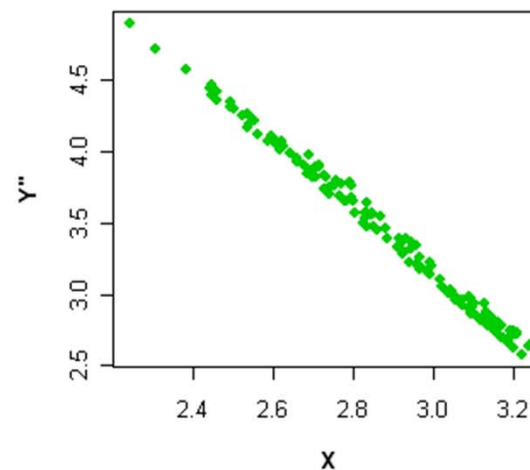
*Relació lineal,
creixent i
força intensa*



*Absència de
relació*




*Relació lineal,
decreixent
i molt intensa*



Coeficient de correlació lineal (r)

- Permet determinar **la direcció i la intensitat** de una **relació lineal** entre dues variables numèriques
- S'obté fent la divisió de la variació conjunta de X i Y (S_{XY}) pel producte de les desviacions estàndards de X (S_X) i de Y (S_Y)

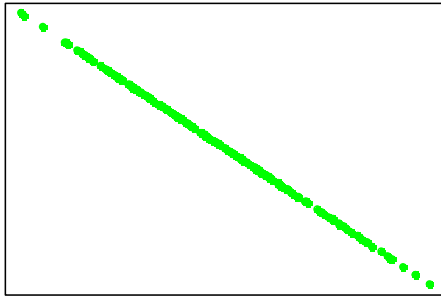
$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y}) / (n-1)}{S_X S_Y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

 **n:** longitud de la mostra
 x_i, y_i : observació i-èsima de la variable x o y
 \bar{x}, \bar{y} : mitjana de la variable x o y
 S_x, S_y : desviació tipus de la variable x o y

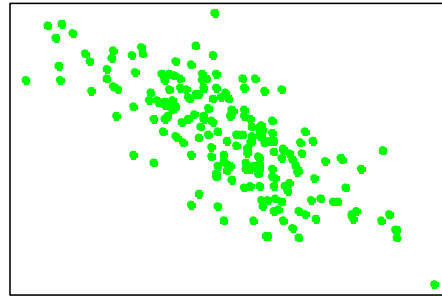
- Propietats:
 - És un valor entre -1 i +1
 - El signe indica la direcció de la relació: directa (si és positiu) o inversa (si és negatiu)
 - La magnitud en valor absolut mesura la intensitat de la relació:
 - $r_{xy} = 0$ indica absència de relació lineal
 - $r_{xy} = 1$ ó $r_{xy} = -1$ indica una relació lineal perfecta que podem representar amb una recta $Y = a + bX$

Coeficient de correlació lineal (r) - Exemples

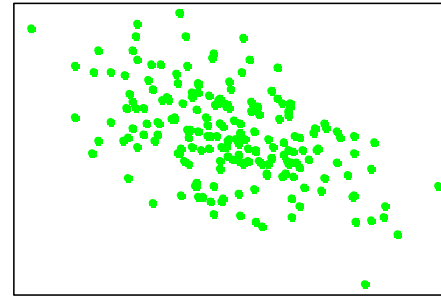
$r = -1.00$



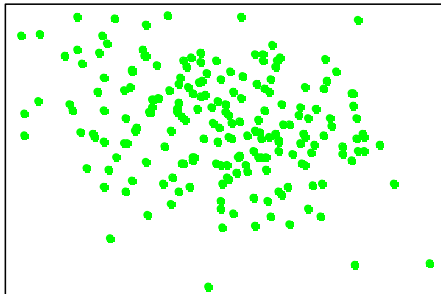
$r = -0.75$



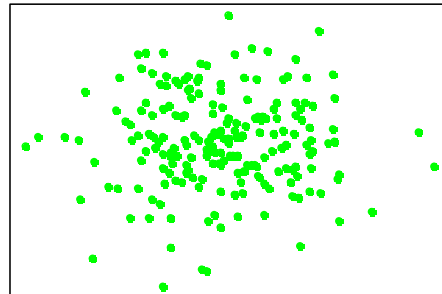
$r = -0.50$



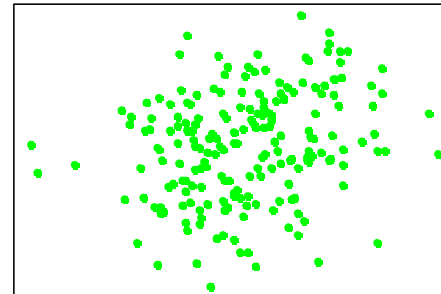
$r = -0.25$



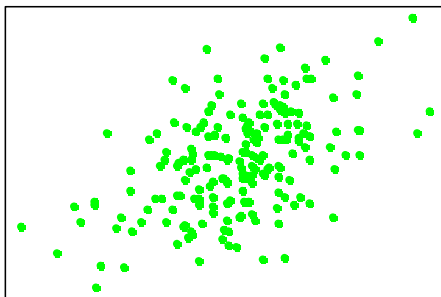
$r = 0.00$



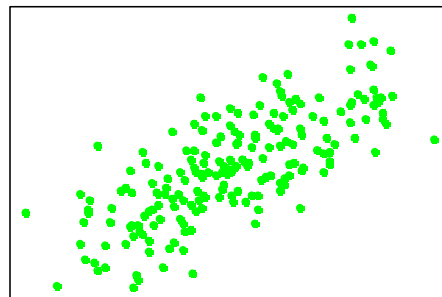
$r = 0.25$



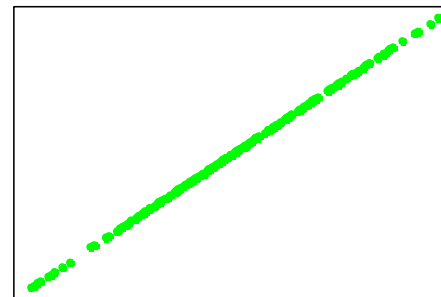
$r = 0.50$



$r = 0.75$



$r = 1.00$




 La instrucció *cor*
calcula la
correlació entre
dues variables
numèriques

Diagrama de dispersió (Scatterplot) amb R

```
# Fixem finestra amb 2 gràfics (mfrow=c(1,2)), ticks horitzontals (las=1) i amb  
# font normal (font.axis=1) i etiquetes en negreta (font.lab=2).
```

```
➤ par(mfrow=c(1,2), las=1, font.axis=1, font.lab=2)
```

```
# Scatterplots de la velocitat de baixada en funció de la  
# velocitat de pujada i la distància a la central
```

```
➤ plot(adsl$down.speed~adsl$up.speed)
```

```
➤ plot(adsl$down.speed~adsl$dist.central)
```

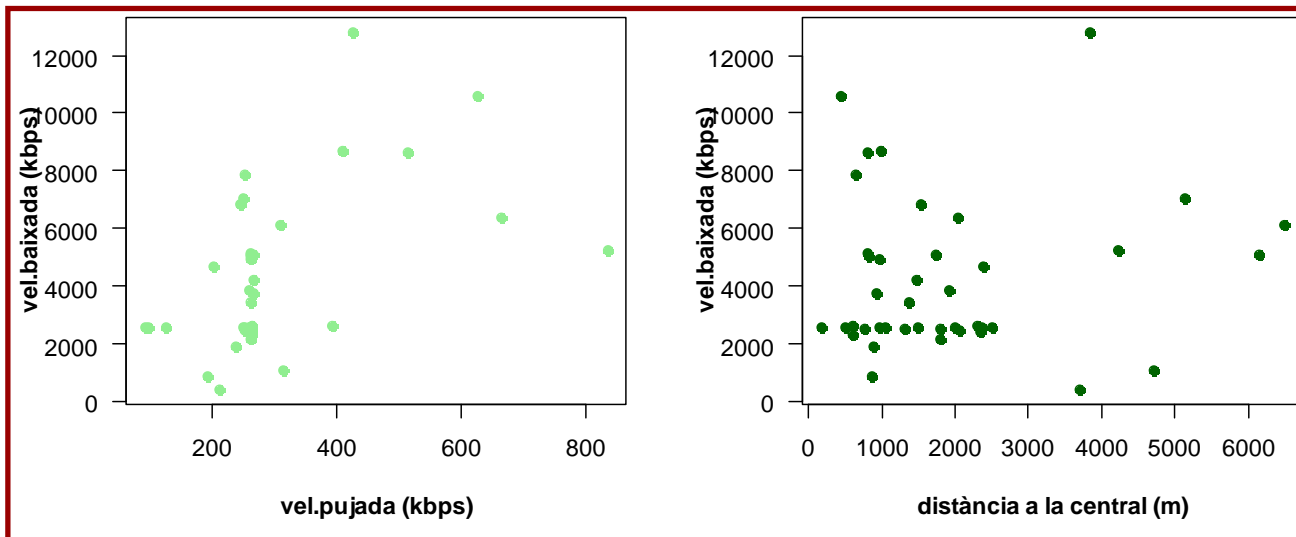
```
# Scatterplots millorats
```

```
➤ plot(down.speed~up.speed, data=adsl, pch=19, col="lightgreen",  
       xlab="vel.pujada (kbps)", ylab="vel.baixada (kbps)")
```

```
➤ plot(down.speed~dist.central, data=adsl, pch=19, col="darkgreen",  
       xlab="distància a la central (m)", ylab="vel.baixada (kbps)")
```



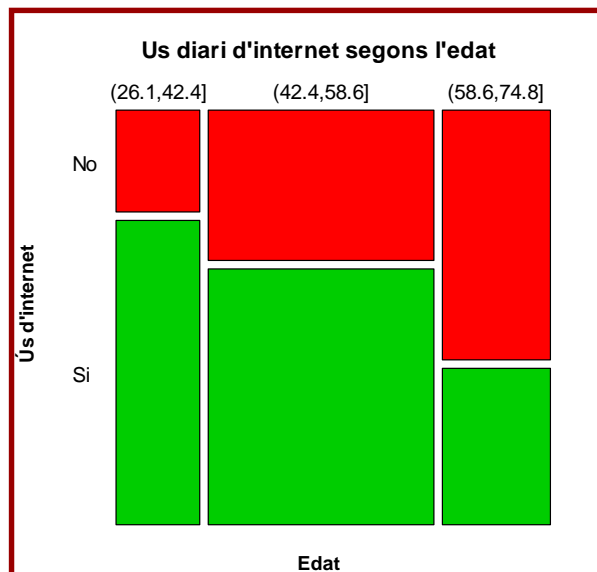
*pch=19 s'utilitza per
dibuixar punts sòlids*



*Amb quina variable està
relacionada la velocitat de
baixada?*

Diagrama de mosaic (Mosaicplot)

- És **la representació gràfica d'una taula** de contingència de dues variables categòriques.
- **Columnes:** l'amplada és proporcional al número d'efectius de cada categoria en una de les variables.
- **Files:** l'alçada de cada fila dins de cada columna és proporcional a la proporció d'efectius de la categoria de la fila dins de la categoria-columna corresponent .

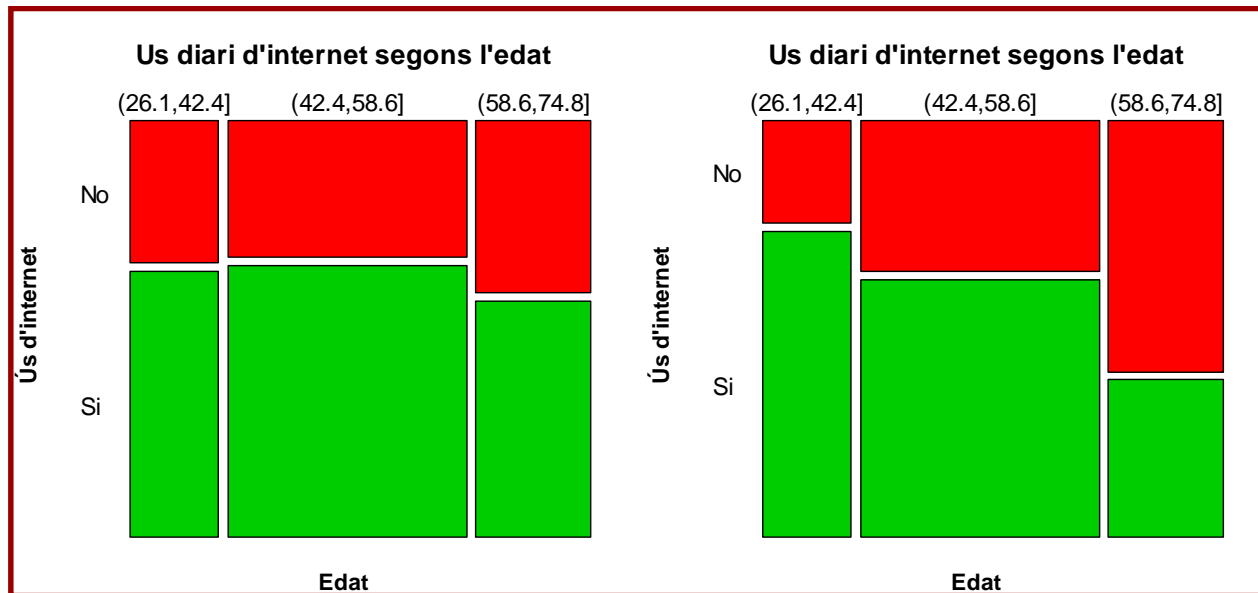


	(26.1,42.4]	(42.4,58.6]	(58.6,74.8]
No	5	20	16
Sí	15	34	10

👍 Aquesta és la taula de contingència corresponent al gràfic

Diagrama de mosaic – Interpretació (I)

Exemple de dos mosaicplots en dues mostres, d'una variable d'edat amb 3 categories relacionada amb l'ús diari d'internet



👍 L'amplada de la columna del mig indica que hi ha més individus en la franja d'edat central (més del doble) que en les altres.

👍 A l'esquerra, es manté aproximadament la mateixa proporció d'ús d'internet en les tres franges d'edat

👍 A la dreta, no es manté la mateixa proporció d'ús d'internet en les tres franges d'edat

Diagrama de mosaic – Interpretació (II)

Els mosaicplots anteriors es corresponen amb les taules mostrades a continuació (n =100)

```
> (t1 <- table(Us.internet1,Edat_C))
      Edat_C
Us.internet1 (26.1,42.4] (42.4,58.6] (58.6,74.8]
No           7           18           11
Si          13           36           15
```

```
> (t2 <- table(Us.internet2,Edat_C))
      Edat_C
Us.internet2 (26.1,42.4] (42.4,58.6] (58.6,74.8]
No           5           20           16
Si          15           34           10
```

Entre dues variables categòriques es vol saber si hi ha **independència o no entre les mateixes**. En cas de suposar independència, es poden calcular, a partir dels efectius marginals (totals per cada fila i cada columna), els efectius esperats:

```
> chisq.test(t1)$exp
      Edat_C
Us.internet1 (26.1,42.4] (42.4,58.6] (58.6,74.8]
No           7.2         19.44         9.36
Si          12.8         34.56        16.64
```

```
> chisq.test(t2)$exp
      Edat_C
Us.internet2 (26.1,42.4] (42.4,58.6] (58.6,74.8]
No           8.2         22.14        10.66
Si          11.8         31.86        15.34
```



Exemple de càlcul d'una cel·la de la taula 1 en cas d'independència:

$P(\text{No} \cap (26.1,42.4]) = P(\text{No}) \cdot P((26.1,42.4]) = (7+18+11)/100 \cdot (7+13)/100 = 0.072$

Efectius esperats en la 1a cel·la de la primera taula $\rightarrow e_{11} = n \cdot P(\text{No} \cap (26.1,42.4]) = 100 \cdot 0.072 = 7.2$

La proximitat dels resultats esperats suposant independència amb la taula 1 (t1) indica **independència**, mentre que la discrepància dels efectius de la taula 2 (t2) denota **no independència**.

Diagrama de mosaic (Mosaicplot) amb R

Fixem dues finestres i els marges del gràfic

➤ `par(mfrow=c(1,2),mar=c(1,2,0,0))`

Accedim a les variables d'adsl només escrivint el seu nom

➤ `attach(adsl)`

Mosaic plot del grup contra el proveïdor i el tipus de connexió

➤ `mosaicplot(table(grp,proveedor),col=1:5,main="",cex.axis=1.1)`

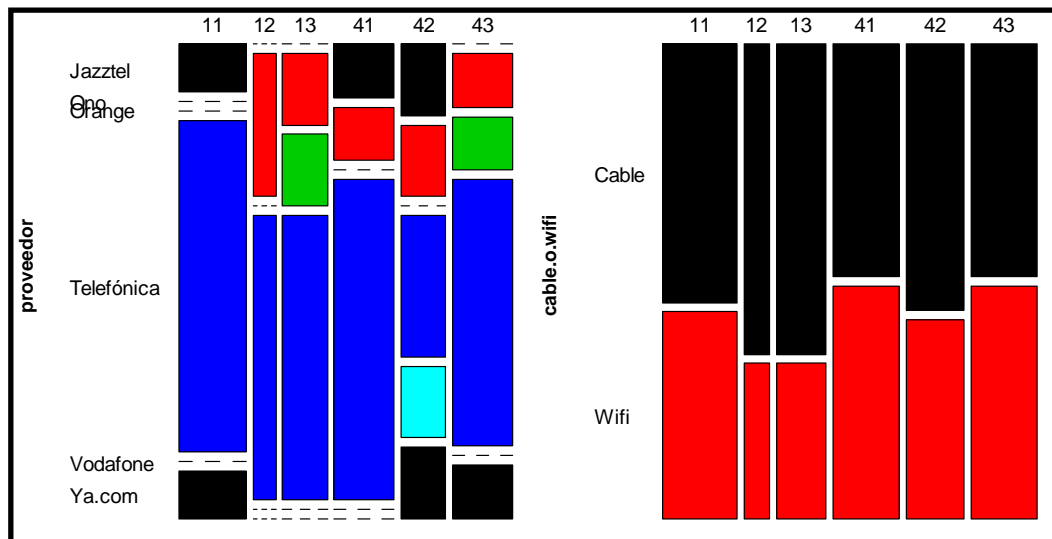
➤ `mosaicplot(table(grp,cable.o.wifi),col=c("black","red"),main="",cex.axis=1.1)`

Traiem el objecte adsl del camí de cerca

➤ `detach(adsl)`



El paràmetre *mar* rep un vector amb 4 components que representen els marges inferior, esquerra, superior i dret, respectivament



Quin és el grup que té més alumnes?

Quin és el únic proveïdor present en tots els grups?

Quin és el grup amb una major proporció d'Ono?

Quin és el grup amb més alumnes amb Ono?



El paràmetre *cex* fa referència a la grandària d'algun ítem. Per defecte, val 1



Per veure més opcions del gràfic fer
?mosaicplot