

## 1. Inferència estadística.

- Hem d'aportar evidència basada en dades
  - per exemple, dir *el meu programa funciona* requereix proves/dades
- De forma **reproduïble**: només resultats predictibles tenen interès
  - per exemple, una curació **miraculosa** no serà útil per futurs pacients
- I **transparent**
  - per permetre la seva replicació per a altres
- **Inferim** les característiques de la **població** a partir de les observacions d'una **mostra aleatòria (m.a.)**
  - per exemple, puc inferir la velocitat de connexió a tota la població a partir d'una mostra aleatòria de velocitats



## 1. Inferència estadística. Riscos

- Mètode científic i tècnic (estadístic):
  - per **deducció** → disseny de la recollida de dades (Població → m.a.)
  - per **inducció** → inferir (estimar) resultats (m.a. → Població)
- La Inferència Estadística defineix i **quantifica els riscos** d'aquest procés [per exemple, no es pot conèixer la mitjana de la vel. de connexió a tota la població a no ser que es tingui dades de tota la població, però l'estadística permet estimar i **quantificar l'error** a partir d'una **mostra a l'atzar** concreta]
- L'**evidència** aportada per les **dades** termina amb **l'anàlisi**, com per exemple:
  - “El meu programa funciona bé”  
→ estimar una mesura (ex: **mitjana** del rendiment) i **el seu error**
  - “El meu programa millora els resultats de ...”  
→ estimar la millora de rendiment (ex: **diferència mitjanes**) i **el seu error**

## 1. Inferència estadística. Tipus de variables

Per analitzar la relació entre variables, cal establir el paper de cadascuna d'elles:

- **Resposta Y.** Mesura l'assoliment de l'objectiu -de vegades pot ser una mesura indirecta  
Ex: rendiment **Y** mesurat en les notes de certa assignatura
- **Decisions X.** Assignem els seus valors en els estudis experimentals  
Representen el potencial per canviar el futur: volem mesurar l'**efecte de X en Y**  
Un disseny experimental permet la seva independència de la resta de variables.  
Ex: un mètode docent basat en **llistes impreses** d'exercicis (**X=1**) comparat amb un mètode basat en **e-status (X=2)**.
- **Co-variables Z.** Representen les condicions observades en dades *reals*  
Podem usar les **Z** per reduir la incertesa de **Y** (haurem de quantificar el seu encert)  
Podem obtenir les **Z** tant en estudis experimentals com observacionals  
Les **Z** solen estar interrelacionades (*col-lineals o no ortogonals*)  
Ex: les notes (**Z<sub>1</sub>, Z<sub>2</sub>**) de dues assignatures prèvies solen tenir certa relació

## 1. Inferència estadística. Tipus d'estudis

### • FER: Estudis experimentals

Volem **canviar** el futur **Y** a partir d'intervencions en **X**

A l'anàlisi estimem els **efectes de X en Y**.

Ex: Per intentar millorar les notes **Y**, assignem a l'atzar els alumnes a diferents entorns de treball **X**

**X** representa una causa **assignable** ben definida  
La clau per intervenir és ser **proprietaris de X**  
Per garantir la independència amb tota **Z**, assignem **X** a l'atzar  
Assignem respectant drets ètics i legals.

### • VEURE: Estudis observacionals

Permeten **predir Y** a partir dels valors observats **Z**

Quantificarem la **capacitat** de **Z** per **reduir la incertesa** en la predicció de **Y**

Ex: comparem notes **Y** segons el grup (**Z<sub>1</sub>**), o segons la nota d'una altra assignatura (**Z<sub>2</sub>**), o en funció d'un cert **model m** de les dues variables [**m=f(Z<sub>1</sub>, Z<sub>2</sub>)**].

→ **El grup Z<sub>1</sub> per ell mateix redueix un 10% la incertesa; la nota Z<sub>2</sub>, un 20%; i el model m, amb les dues, un 25%**.

No som **proprietaris** de les variables **Z** (les unitats ja venen amb el valor de les **Z**)  
Podem establir **relacions** entre **Z i Y**, que podem utilitzar per **predir** els valor de **Y** a partir de **Z**.  
Però les covariables **Z** poden estar relacionades (ser *col-lineals*) i per tant poden tenir **confosos** els seus **efectes** en **Y**.  
Establir **causalitat** requereix moltes premisses (fora d'un curs introductorí)

## 1. Inferència estadística. Conceptes bàsics

- Paràmetre:** indicador de la població que estem interessats en conèixer o estimar. Per exemple la  $\mu$  (esperança) de les alçades dels estudiants de la FIB
- Estadístic:** qualsevol indicador que s'obtingui com a funció de les dades d'una mostra. Per exemple la suma de les alçades dels estudiants recollits en una mostra
- Estimador:** estadístic d'una mostra que s'utilitza per conèixer el valor d'un paràmetre de la població. Per exemple la mitjana de les alçades en una mostra a l'atzar d'alumnes de la FIB és una estimador de la  $\mu$  (esperança) de les alçades dels estudiants de la FIB

**Mitjana** pot voler dir **paràmetre esperança** quan parlem del centre de gravetat de la distribució poblacional, o **estadístic mitjana** quan ens referim al valor mitjà d'una sèrie de valors obtinguts d'una mostra

## 2. Estimació puntual

- Un estimador  $\hat{\theta}$  del paràmetre desconegut  $\theta$ , a partir de la mostra  $M(\omega)$  ( $X_1, X_2, \dots, X_n$ ) (*mostra aleatòria simple definida a l'annex del bloc B*), és una funció de les VA :

$$\hat{\theta} = f(X_1, X_2, \dots, X_n)$$

- Estimació puntual:** valor que l'estimador  $\hat{\theta}$  pren en una mostra concreta.

Per exemple  $\bar{x} = \frac{\sum x_i}{n}$  és la mitjana mostra i és una estimació puntual de  $\mu$

Distingiu entre el valor  $\bar{x}$  d'una m.a.s. concreta i la variable aleatòria mitjana mostra  $\bar{X}$

- Error tipus o error estàndard:** variabilitat de l'estimador. En el cas anterior de la MITJANA, l'**error tipus (o estàndard) de la mitjana (o mean standard error o se)** és:

$$se = \sqrt{V(\bar{X}_n)} = \sqrt{E[(\bar{X}_n - \mu)^2]} = \frac{\sigma}{\sqrt{n}}$$

Generalment, la  $\sigma$  serà desconeguda i l'error tipus l'haurem d'aproximar emprant l'estimador pertinent ( $\hat{\sigma}$ ) amb les dades de la mostra:  $\hat{se} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n}}$  (amb  $s$  estimador puntual de  $\sigma$ )

Llegiu les consideracions de la secció 5 d'aquest bloc "Dissenys (com obtenim les dades)" per aplicar al bloc T

## 2. Estimació puntual. Casos

Per als paràmetres utilitzem lletres de l'alfabet grec

Paràmetre ( $\theta$ ) (POBLACIÓ)	Estimador ( $\hat{\theta}$ ) (MOSTRA)
$\mu$ (esperança, mitjana poblacional)	$\bar{x}$ (mitjana mostra)
$\sigma^2$ (variància poblacional)	$s^2$ (variància mostra)
$\sigma$ (desviació tipus poblacional)	$s$ (desviació tipus mostra)
$\pi$ (probabilitat)	$p$ (proporció)

El cas de la MITJANA:

$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$  mitjana mostra és una estimació puntual del paràmetre  $\mu$  de tendència central

El cas de la DESVIACIÓ:

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

desviació tipus mostra és una estimació puntual del paràmetre  $\sigma$  de dispersió

El cas de la PROPORCIÓ:

$$p = \sum_{i,x_i=1} 1/n \text{ proporció mostra és una } \text{estimació puntual del paràmetre } \pi$$

Cal tenir en compte les propietats dels estimadors (veure a l'annex, juntament amb altres possibles estimadors)

## Exemple. Nombre de terminals (assumint premissa de normalitat)

En 9 dies consecutius s'ha observat el nombre de terminals en una Universitat connectats a internet: 587, 470, 676, 451, 436, 672, 584, 697 i 408       $[\sum x_i = 4981 ; \sum x_i^2 = 2860855]$

$$x <- c(587, 470, 676, 451, 436, 672, 584, 697, 408)$$

Una estimació puntual del nombre esperat ( $\mu$ ) de terminals diaris connectats és:

$$\text{mean}(x) \rightarrow \bar{x} = 553.44 \quad \text{o} \quad \bar{x} = (\sum x_i)/n = \frac{4981}{9} = 553.44$$

Una estimació puntual de la desviació tipus ( $\sigma$ ) del nombre de terminals connectats és:

$$\text{sd}(x) \rightarrow s = 114.0988 \quad \text{o} \quad s = \sqrt{(2860855 - (4981)^2/9)/(8)} = 114.0988$$

L'estimació de l'error tipus o variabilitat de la mitjana és:

$$\text{sd}(x) / \text{sqrt}(\text{length}(x)) \rightarrow se = 38.03 \quad se = \frac{s}{\sqrt{n}} = \frac{114.0988}{\sqrt{9}} = 38.03$$

\* Quant valdria l'error tipus (o estàndard error) si el mateix valor de mitjana i de desviació provinguessin de  $n=100$  valors:  $\frac{s}{\sqrt{n}} = \frac{114.0988}{\sqrt{100}} = 11.4$

## 2. Estimadors i Estadística Descriptiva

Els anteriors estimadors puntuals es corresponen a les funcions d'Estadística Descriptiva per resumir numèricament unes dades (veure'n més a l'apartat de R de la pàgina web)

En la següent taula hi ha algunes funcions (bàsiques) en R per Estadística Descriptiva en variables **numèriques i categòriques** de forma **univariant o bivariant**:

	UNIVARIANT (num)	UNIVARIANT (categ)	BIVARIANT
INDICADORS	<b>length()</b> * <b>mean( )</b> <b>var( )</b> <b>sd( )</b> <b>summary( )</b> <b>median( )</b>	<b>table( )</b>	<b>cov( , )</b> <b>cor( , )</b>
GRÀFIQUES	<b>hist( )</b> <b>boxplot( )</b>	<b>barplot(table( ))</b>	<b>plot( , )</b>

\* La mida de la mostra ( $n$ ) no és un estimador, però l'incluem a la llista per sentit pràctic

## 3. Estimació per Interval

- Sabem com calcular un “interval” que contingui  $\bar{x}$  a partir de  $\mu$ . Però el problema real és **aproximar  $\mu$ , coneixent  $\bar{x}$**  (és a dir, passar d'un interval per a la mitjana mostral  $\bar{x}$ , a un per a la mitjana poblacional  $\mu$ )
- A partir d'una probabilitat  $1-\alpha$  entre dos valors  $a$  i  $b$  (simètrics): (amb  $\sigma$  coneguda)

$$P(a \leq \bar{X}_n \leq b) = 1 - \alpha \rightarrow P\left(\frac{a-\mu}{\sigma/\sqrt{n}} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq \frac{b-\mu}{\sigma/\sqrt{n}}\right) = 1 - \alpha \rightarrow P\left(z_{\frac{\alpha}{2}} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

- Obtenim l'interval de la v. a.  $\bar{X}_n$  amb **probabilitat  $1-\alpha$**

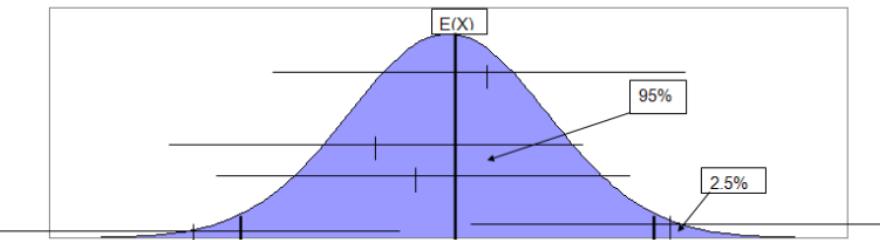
$$P\left(\mu + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

- I reordenant obtenim **l'interval de confiança  $1-\alpha$  del paràmetre  $\mu$**

$$P\left(\bar{X}_n + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

## 3. Estimació per Interval de confiança

- $P\left(\bar{X}_n + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$  significa que podem assegurar que  $E(X) = \mu$  estarà (amb una confiança de  $1 - \alpha$ ) en el rang calculat
- Si  $1-\alpha$  és 95% ( $\alpha = 5\%$ ): **el 95% dels intervals (IC) contindran  $\mu$**  (veure una simulació a l'annex)

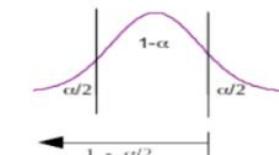


- Aquest procediment encerta el  $100 \cdot (1-\alpha)\%$  de les vegades!
- Denotem  $IC(\mu, 1-\alpha)$  a l'**INTERVAL DE CONFIANÇA**  $1-\alpha$  de  $\mu$ , i l'expressem:

Nosaltres només observarem una mostra, i no sabrem si l'IC trobat conté o no  $\mu$ , però sí sabem que aquest procediment a la llarga dóna un  $100 \cdot (1-\alpha)\%$  d'encerts

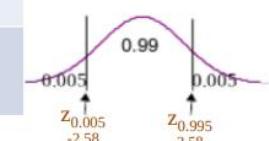
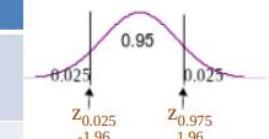
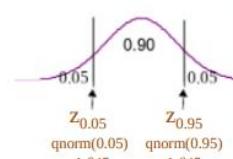
### 3.a. Confiança i risc

El càlcul d'un IC implica una confiança  $1-\alpha$  (i per tant un risc  $\alpha$ ) que podem representar:



I podem relacionar el valor de confiança amb el quantil que necessitarem per construir l'IC: (com a exemple estan indicats els quantils per una Z Normal(0,1) on sabem que  $z_\alpha = -z_{1-\alpha}$  o  $z_{\alpha/2} = -z_{1-\alpha/2}$ )

Confiança $1-\alpha$	Risc $\alpha$	$\alpha/2$	$1 - \alpha/2$
0.95	0.05	0.025	0.975
0.90	0.10	0.05	0.95
0.99	0.01	0.005	0.995



### 3.b. Estadístics per fer inferència

- Veurem estadístics de dos tipus:

- Rati de "senyal" o "informació" (diferència entre el valor del paràmetre i el mostra) respecte "soroll" o "error" (estàndard error,  $se$ )

Aquests estadístics es modelen seguint el model Z o T Student\* (en alguns casos avaluem el "t-rati" que quantifica quantes vegades és més gran el senyal que el soroll)

$$\text{estadístic } \hat{Z} = \frac{(\bar{x}-\mu)}{\sigma/\sqrt{n}} = \frac{(\bar{x}-\mu)}{se} \quad \hat{Z} \sim N(0,1) \quad (\text{dóna lloc al IC } \mu \in \bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \text{ o bé } \bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot se)$$

$$\text{estadístic } \hat{t} = \frac{(\bar{x}-\mu)}{S/\sqrt{n}} = \frac{(\bar{x}-\mu)}{se} \quad \hat{t} \sim \text{T de Student amb } v \text{ graus de llibertat}$$

- Quocient de variàncies. Aquests estadístics es modelen seguint el model F

$$\text{estadístic } \hat{F} = \frac{S_A^2}{S_B^2} \quad \hat{F} \sim F \text{ de Fisher-Snedecor amb } v_1 \text{ i } v_2 \text{ graus de llibertat}$$

Les distribucions T-Student ( $t_v$ ),  $F (F_{v1,v2})$  i  $\chi^2$  quadrat ( $\chi^2_v$ ) estan definides en el Bloc B (annex). Son models derivats de la Normal, i estan parametritzades amb  $v$ , que anomenem "graus de llibertat" i que depèn de les mides  $n$  de les mostres

Ara veurem les **fórmules d'IC** quan tenim **UN SOL PARÀMETRE** ➤ interès en una proporció  $\pi$   
d'interès, distingint els següents casos:

- interès en la mitjana  $\mu$  (amb variància poblacional coneiguda o no)  
(per exemple la mitjana de la nota d'una assignatura)
- interès en la variabilitat  $\sigma^2$   
(per exemple la desviació de la nota d'una assignatura)

### Exemple. Nombre de terminals (assumint premissa de normalitat)

A partir de les estimacions puntuals calculades, podem calcular una estimació per interval de  $\mu$  (nombre esperat de terminals diaris connectats en mitjana), amb diversos nivells de confiança, i en el supòsit de desconèixer el valor de la variabilitat poblacional o el d'assumir-ne un valor conegit (per exemple  $\sigma=100$ ). També assumim la premissa de normalitat

```
X <- c(587, 470, 676, 451, 436, 672, 584, 697, 408)
n <- 9
```

1- $\alpha$	$\sigma$	IC( $\mu, 1-\alpha$ )	Resolució amb R
95%	Coneguda ( $\sigma=100$ )	[488.11; 618.78]	<code>sigma &lt;- 100 mean(X) - qnorm(0.975) * sigma/sqrt(n) mean(X) + qnorm(0.975) * sigma/sqrt(n)</code>
99%	Coneguda ( $\sigma=100$ )	[467.58 ; 639.31]	<code>sigma &lt;- 100 mean(X) - qnorm(0.995) * sigma/sqrt(n) mean(X) + qnorm(0.995) * sigma/sqrt(n)</code>
95%	Desconeguda	[465.74; 641.15]	<code>mean(X) - qt(0.975, n-1) * sd(X) / sqrt(n) mean(X) + qt(0.975, n-1) * sd(X) / sqrt(n)</code>
99%	Desconeguda	[425.83 ; 681.06]	<code>mean(X) - qt(0.995, n-1) * sd(X) / sqrt(n) mean(X) + qt(0.995, n-1) * sd(X) / sqrt(n)</code>

\* Observeu que, a més confiança (menys risc d'error), la precisió dels IC disminueix (interval més ample) i que els IC amb  $\sigma$  desconeguda són més amples que els equivalents assumint el verdader valor de  $\sigma$ , ja que hi ha més incertesa i usem  $t$  enlloc de  $N(0,1)$

### 4.a. Interval de confiança de $\mu$ (amb $\sigma$ coneiguda)

- L'interval de confiança  $1-\alpha$  de  $\mu$  (amb  $\sigma$  coneiguda) es calcula com

$$IC(\mu, 1-\alpha) = \bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

- Recordeu que ens basem en el TCL i perquè es complís calia que la variable X inicial fos Normal o que n fos "gran". Per tant, els requisits per realitzar aquest càlcul són: n "gran" o  $X \sim N$
- Aquest IC es pot obtenir aïllant el paràmetre  $\mu$  de l'estadístic:  $\hat{\mu} = \frac{(\bar{x}-\mu)}{\sigma/\sqrt{n}} = \frac{(\bar{x}-\mu)}{se}$  del qual en coneixem la seva distribució que és  $N(0,1)$  ja que el IC es pot veure com  $\bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot se$

Quan  $n$  augmenta la precisió dels IC augmenta (interval més estret)

Si augmenta la confiança (disminuint el risc  $\alpha$  d'error), la precisió dels IC disminueix (interval més ample)

Per estimar  $\mu$  necessitem conèixer  $\sigma$ , que és una situació poc realista doncs  $\sigma$  acostuma a ser un paràmetre desconegut (també podem assumir un valor raonable, pel coneixement previ)

### 4.a. Interval de confiança de $\mu$ amb $\sigma$ desconeguda

- L'interval de confiança  $1-\alpha$  de  $\mu$  (amb  $\sigma$  desconeguda) es calcula com:

$$IC(\mu, 1-\alpha) = \bar{x} \pm t_{n-1, 1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

- Aquest IC es pot obtenir aïllant el paràmetre  $\mu$  de l'estadístic:  $\hat{\mu} = \frac{(\bar{x}-\mu)}{s/\sqrt{n}} = \frac{(\bar{x}-\mu)}{se}$  on al desconèixer  $\sigma$ , es substitueix per  $s$  i es distribueix segons la llei **t-Student** amb  $n-1$  graus de llibertat, i el IC es pot veure com  $\bar{x} \pm t_{n-1, 1-\frac{\alpha}{2}} \cdot se$
- En aquest cas cal que la variable X inicial fos Normal (premissa de normalitat) ja que la definició de la **t-Student** parteix de variables normals

La situació de desconèixer  $\sigma$  és més realista i freqüent: no se n'assumeix cap valor sinó que s'aproxima per la seva estimació puntual  $s$

$t$  i  $N(0,1)$  són similars, més quan  $n$  creix:  $t_{n \rightarrow \infty} \rightarrow N(0,1)$

Per valors de  $n$  petits,  $t$  té més variabilitat reflectint més incertesa (relacionat amb que aproximem  $\sigma$  per  $s$ )

A l'IC amb  $\sigma$  desconeguda li correspondrà ser més ample que l'equivalent assumint el verdader valor de  $\sigma$  ja que hi ha més incertesa i usem  $t$  enlloc de  $N(0,1)$

## 4.a Interval de confiança de $\mu$ . Premisses

Per garantir el nivell de confiança de l'IC, s'ha de complir certes premisses  
La premissa fonamental és que l'origen de la mostra sigui aleatori (v.a.i.i.d.)

A més a més:

- Si sigma és coneguda, exigirem una de les condicions:

- $X \sim N \rightarrow$  la combinació lineal de Normals és Normal ( $\bar{X} \sim N$ )
- Tenir una mostra "gran"  $\rightarrow$  Pel TCL,  $\bar{X} \sim N$

- Si sigma no és coneguda, exigirem una de les condicions:

- $X \sim N \rightarrow (\bar{x} - \mu) / \sqrt{s^2/n} \sim t_{n-1}$
- Tenir una mostra gran ( $n$  "gran")  $\rightarrow$  Pel TCL,  $\bar{X} \sim N$

Amb grans mostres la variació de "s" serà més petita (s estima bé  $\sigma$ ), i podem considerar que  $(\bar{x} - \mu) / \sqrt{s^2/n} \approx (\bar{x} - \mu) / \sqrt{\sigma^2/n} \sim N(0,1)$

Dist. de referència	$\sigma$ coneguda	$\sigma$ desconeguda
X Normal		Usar <b>t de Student</b>
X no Normal i n "gran"	Usar la Normal	

## 4.b. Interval de confiança de $\pi$

- Sigui  $X \sim B(n, \pi) \rightarrow E(X) = \pi \cdot n$   
 $V(X) = \pi \cdot (1-\pi) \cdot n$
- Aleshores,  $P = X/n \rightarrow E(P) = E(X/n) = E(X)/n = \pi \cdot n / n = \pi$   
 $V(P) = V(X/n) = V(X)/n^2 = \pi \cdot (1-\pi) \cdot n / n^2 = \pi \cdot (1-\pi) / n$
- Per construir l'IC es pot recoure a la convergència de la Binomial a la Normal [amb la premissa de  $n$  "gran" i  $\pi$  no extrema (com a guia comprovar que  $\pi \cdot n \geq 5$  i  $(1-\pi) \cdot n \geq 5$ ):

$$P \rightarrow N \left( \mu_P = \pi, \sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}} \right)$$

- Així, l'interval de confiança s'assemlaria al de  $\mu$ , aïllant el paràmetre  $\pi$  de l'estadístic  $\hat{\pi} = \frac{(P-\pi)}{\sigma_P} = \frac{(P-\pi)}{se}$  del qual en coneixem la seva distribució que és  $N(0,1)$ :

$$IC(\pi, 1 - \alpha) = P \pm z_{1-\frac{\alpha}{2}} \text{ se} = P \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

La **paradoxa** de que necessitem conèixer  $\pi$  per estimar el IC de  $\pi$  es pot solucionar de 2 maneres:

a) Substituint  $\hat{\pi}$  per  $P$ :  $IC(\pi, 1 - \alpha) = P \pm z_{1-\alpha/2} \cdot \sqrt{(P(1-P))/n}$

b) Aplicant el màxim de  $\hat{\pi} \cdot (1 - \hat{\pi})$  que correspon a fer  $\hat{\pi}$  igual a 0.5:  $IC(\pi, 1 - \alpha) = P \pm z_{1-\alpha/2} \cdot \sqrt{(0.5(1-0.5))/n}$

## Exemple. Moneda (IC per $\pi$ )

Llencem 100 vegades una moneda a l'aire i observem 56 cares ( $P = 56/100 = 0.56$ ).

Les dues solucions per l'IC segons com estimem  $\pi$ :

$$IC(\pi, 0.95) = P \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{P \cdot (1-P)}{n}} = 0.56 \pm 1.96 \sqrt{\frac{0.56 \cdot 0.44}{100}} \approx 0.56 \pm 0.10 = [0.46, 0.66]$$

$$IC(\pi, 0.95) = P \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\pi_m \cdot (1-\pi_m)}{n}} = 0.56 \pm 1.96 \sqrt{\frac{0.50 \cdot 0.50}{100}} \approx 0.56 \pm 0.10 = [0.46, 0.66]$$

Donen el mateix IC fins al 2n decimal. El motiu és que la probabilitat estimada (0.56) és molt similar a la probabilitat de màxima indeterminació (0.50).

Es podria contrastar un possible valor del paràmetre (per exemple  $\pi=0.50$  indicant una moneda equilibrada), amb confiança del 95%: com que el valor 0.50 cau dins l'IC, és versemblant que la moneda sigui equilibrada d'acord amb l'evidència empírica que les dades aporten.

## 4.c. Interval de confiança de $\sigma^2$

$$\text{Si } X_i \sim N \quad (n-1) \cdot \frac{s^2}{\sigma^2} = (n-1) \cdot \frac{(\sum_{i=1}^n (x_i - \bar{x})^2) / (n-1)}{\sigma^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} = \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{\sigma} \right)^2 \sim \chi^2_{n-1}$$

Podem relacionar l'estadístic de quocient de variàncies (amb el qual definirem l'IC) amb una  $\chi^2$  per ser suma de Normals al quadrat (veure models derivats de la Normal a l'annex del bloc B)

Per tant:

$$P \left( \chi^2_{n-1, \frac{\alpha}{2}} \leq \frac{s^2 \cdot (n-1)}{\sigma^2} \leq \chi^2_{n-1, 1-\frac{\alpha}{2}} \right) = 1 - \alpha$$

$$P \left( \frac{1}{\chi^2_{n-1, 1-\frac{\alpha}{2}}} \leq \frac{\sigma^2}{s^2 \cdot (n-1)} \leq \frac{1}{\chi^2_{n-1, \frac{\alpha}{2}}} \right) = 1 - \alpha$$

$$P \left( \frac{s^2 \cdot (n-1)}{\chi^2_{n-1, 1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{s^2 \cdot (n-1)}{\chi^2_{n-1, \frac{\alpha}{2}}} \right) = 1 - \alpha$$

$$IC(\sigma^2, 1 - \alpha) = \left[ \frac{s^2 \cdot (n-1)}{\chi^2_{n-1, 1-\frac{\alpha}{2}}}, \frac{s^2 \cdot (n-1)}{\chi^2_{n-1, \frac{\alpha}{2}}} \right]$$

És un IC per  $\sigma^2$ , no per  $\sigma$  !!

No és un interval simètric, ja que  $\chi^2$  no ho és. Implica calcular els dos quantils (inferior i superior) en lloc de fer  $\pm$

## 5. IC per a comparar 2 paràmetres

Ara veurem les **fòrmules d'IC** quan tenim **DOS PARÀMETRES** d'interès, distingint els següents casos:

- Comparar  $\mu_1$  i  $\mu_2$   
(per ex IC de l'efecte diferencial ( $\mu_1 - \mu_2$ ) comparant mitjanes entre dues assignatures\*)
- Cal diferenciar entre:
  - **mostres aparellades\*\*** (cada cas dóna lloc a dues mesures, parells de mesures)  
(els mateixos estudiants en les dues assignatures,  $\mu_1 - \mu_2 = \mu_{Diferència} = \mu_D$ )
  - **mostres independents** (cada cas és una mesura independent)  
(estudiants diferents en les dues assignatures)
- Comparar  $\pi_1$  i  $\pi_2$   
(per ex IC de l'efecte diferencial ( $\pi_1 - \pi_2$ ) comparant aprovats entre dues assignatures\*)
- Comparar  $\sigma^2_1$  i  $\sigma^2_2$   
(per ex IC de l'efecte diferencial ( $\sigma^2_1 / \sigma^2_2$ ) comparant desviacions entre dues assignatures\*)

(\* cal que l'origen de la mostra sigui aleatori (v.a.i.i.d). Per tant, en principi, no unes notes observades)

(\*\*) Si és possible, un disseny amb dades aparellades serà més eficient (com veurem més endavant)

## 5.b. IC de $(\mu_1 - \mu_2)$ mostres independents

Siguin  $Y_1$  ( $E(Y_1) = \mu_1$   $V(Y_1) = \sigma_1^2$ ) i  $Y_2$  ( $E(Y_2) = \mu_2$   $V(Y_2) = \sigma_2^2$ ) amb distribucions Normals ( $\sigma_1$  i  $\sigma_2$  seran valors desconeguts però cal poder assumir-los iguals\*) de les que obtenim dues mostres aleatòries simples de grandària  $n_1$  i  $n_2$  independents, amb mitjanes  $\bar{y}_1$ ,  $\bar{y}_2$  i desviacions  $s_1$ ,  $s_2$  (com a estimadors d'un paràmetre comú  $\sigma$ )

- L'estadístic  $\hat{t} = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{se}$  segueix la distribució  $t_{n_1+n_2-2}$  amb error estàndard  $se = s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$  on  $s$  és arrel de la variància "pooled"  $s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$
- L'IC de la diferència amb confiança  $1-\alpha$  és:

$$IC(\mu_1 - \mu_2, 1 - \alpha) = (\bar{y}_1 - \bar{y}_2) \pm t_{n_1+n_2-2, 1-\frac{\alpha}{2}} \cdot se = \\ (\bar{y}_1 - \bar{y}_2) \pm t_{n_1+n_2-2, 1-\frac{\alpha}{2}} \cdot s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

\* Apart de la premissa de normalitat de  $Y_1$  i  $Y_2$ , comprovarem (gràficament) que tenen variabilitats semblants

## 5.c. IC de $\pi_1 - \pi_2$

Siguin  $P_1$  i  $P_2$  les proporcions mostrals de 2 poblacions binomials amb  $\pi_1$ ,  $\pi_2$  de les que obtenim dues mostres aleatòries simples de grandària  $n_1$  i  $n_2$  independents

- L'estadístic  $\hat{z} = \frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{se}$  segueix la distribució  $N(0,1)$  amb error estàndard  $se = \sqrt{P_1(1-P_1)/n_1 + P_2(1-P_2)/n_2}$

- L'IC de la diferència amb confiança  $1-\alpha$  és:

$$IC(\pi_1 - \pi_2, 1 - \alpha) = (P_1 - P_2) \pm z_{1-\frac{\alpha}{2}} \cdot se = \\ (P_1 - P_2) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{P_1(1-P_1)/n_1 + P_2(1-P_2)/n_2}$$

## 5.a. IC de $\mu_1 - \mu_2$ (o de $\mu_D$ ) en mostres aparellades

Siguin  $Y_1$  ( $E(Y_1) = \mu_1$   $V(Y_1) = \sigma_1^2$ ) i  $Y_2$  ( $E(Y_2) = \mu_2$   $V(Y_2) = \sigma_2^2$ ) de les que obtenim una mostra aleatòria simple **aparellada** de grandària  $n$ , Definim  $D = Y_1 - Y_2$  (o bé  $Y_2 - Y_1$ ) on  $D$  és normal amb  $E(D) = \mu_D$  i  $V(D) = \sigma_D^2$  i els  $n$  valors de la diferència tenen mitjana  $\bar{d}$  i desviació  $s_d$

- L'estadístic  $\hat{t} = \frac{(\bar{d} - \mu_D)}{s_d/\sqrt{n}} = \frac{(\bar{d} - \mu_D)}{se}$  segueix la distribució  $t_{n-1}$

on  $(\bar{d} - \mu_D)$  és el "senyal" i  $se = s_d/\sqrt{n}$  l'error estàndard

- L'IC de la diferència amb confiança  $1-\alpha$  és:

$$IC(\mu_1 - \mu_2, 1 - \alpha) = IC(\mu_D, 1 - \alpha) = \bar{d} \pm t_{n-1, 1-\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}} = \bar{d} \pm t_{n-1, 1-\frac{\alpha}{2}} \cdot se$$

Pot tenir interès pràctic avaluar el "t-rati":  $t = \bar{d} / \frac{s_d}{\sqrt{n}} = \bar{d}/se$

que diu quantes vegades és més gran el senyal que el soroll (assumint m.a. aparellada)

En aquest cas, la convergència requereix mostres "grans": usualment que  $P \cdot n$  i  $(1-P) \cdot n$  siguin superiors a 5

## 5.d. IC de $\sigma_1^2/\sigma_2^2$

Siguen  $s_1$  i  $s_2$  les desviacions mostrals de dues mostres aleatòries simples de grandària  $n_1$  i  $n_2$  independents, de dues variables Normals

- L'estadístic  $\hat{F} = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$  segueix la distribució  $F_{(n_1-1, n_2-1)}$

- L'IC del quocient de variàncies amb confiança  $1-\alpha$  és:  
(seguint el mateix raonament de l'IC de  $\sigma^2$ )

$$IC(\sigma_1^2/\sigma_2^2, 1-\alpha) = \left[ \frac{s_1^2/s_2^2}{F_{(n_1-1, n_2-1), 1-\frac{\alpha}{2}}}, \frac{s_1^2/s_2^2}{F_{(n_1-1, n_2-1), \frac{\alpha}{2}}} \right]$$

o bé (atenció a l'intercanvi dels graus de llibertat de la F)

$$IC(\sigma_1^2/\sigma_2^2, 1-\alpha) = \left[ \frac{s_1^2/s_2^2}{F_{(n_2-1, n_1-1), \frac{\alpha}{2}}}, \frac{s_1^2/s_2^2}{F_{(n_2-1, n_1-1), 1-\frac{\alpha}{2}}} \right]$$

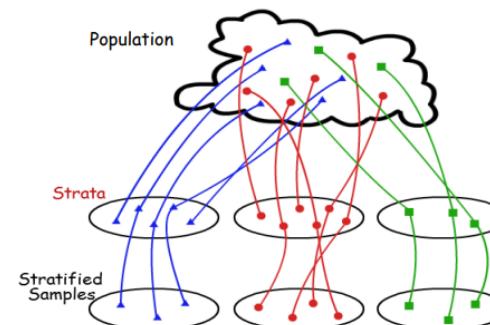
## 6. Dissenys (com obtenim les dades)

El disseny escollit condiciona l'anàlisi estadística posterior.

Si la recollida és complexa, el model estadístic emprat també:

- Cas amb dades **niades** (*clusters*): primer es seleccionen a l'atzar grups del nivell superior (p.e., escola); després, de l'inferior (*classe*); fins arribar a l'individu (*alumne*).
- Cas amb dades **estratificades**: es veuen tots els estrats però, dins de cada estrat, es seleccionen a l'atzar els individus.

Llavors, el grup d'alumnes escollits no és pròpiament una m.a.; s'han d'analitzar amb tècniques apropiades (que no veurem).



TAMPOC és habitual disposar de la població completa i poder accedir a qualsevol unitat en les mateixes condicions (requisit per a ser m.a.s.).

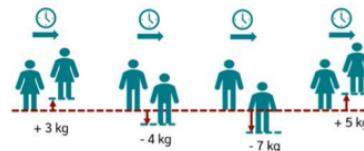
### Atenció:

Normalment, unes unitats seran més "visibles" que altres i tindran més probabilitats de ser escollides (p.e.: resultats que només es poden obtenir ordenats).

## 6. Dissenys (com obtenim les dades)

### Disseny aparellat:

- de cada unitat es treu 1 variable i 2 observacions (les dues mesures o respostes): precisa que la primera observació en una "parella" no alteri l'estat de la unitat i, per tant, de la segona observació



### Mostres independents

- per a cada unitat, es treu 1 observació i 2 variables (la mesura o resposta i la categoria per comparar):
  - precisa que la categoria es pugui assignar a la unitat (no pot ser una condició, com el gènere)
  - en estudis observacionals, quan el grup no és assignable, es seleccionen les mostres per separat



Això és una simple aproximació. El món del disseny d'experiments és molt més ampli.

La clau es "atzar":

Recollir dades de qualsevol manera no garanteix una m.a. "**al tuntún**" ≠ a l'atzar

És imprescindible **planificar** la selecció a l'atzar de les unitats que mesurarem.

I **executar** correctament l'experiment, sense valors mancats. I **documentar-lo** de forma **reproduïble**.

## Exemple. Diferència temps mitjà en mostres aparellades

En 6 bancs de dades s'ha obtingut els temps de 2 programes. Es desitja saber si B millora A, estimant l'efecte diferencial en mitjana

							Mean	Variances	Var. "pooled"
	A	23.05	39.06	21.72	24.47	28.56	27.406	39.428	42.009
	B	20.91	37.21	19.29	19.95	25.32	24.07	24.460	44.591

**SOLUCIÓ INCORRECTA:** (tractar com a dades de mostres independents):

$$\hat{t} = \frac{(\bar{Y}_1 - \bar{Y}_2)}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(27.406 - 24.460)}{6.48 \sqrt{\frac{1}{6} + \frac{1}{6}}} = 0.787$$

**SOLUCIÓ CORRECTA:** (com a dades de mostres aparellades). El mètode comença per calcular la diferència D (A-B) per cada parella (i assumint normalitat):

							Mean	Variances
	D = A-B	2.13	1.85	2.43	4.51	3.24	3.51	2.946

**Estadístic:**  $\hat{t} = \frac{(\bar{D}-\mu_D)}{S_D/\sqrt{n}} = \frac{\bar{D}}{S_D/\sqrt{n}} = 7.229$

$$IC(\mu_D, 0.95) = 2.946 \mp 2,571 \cdot 0.41 = [1.89, 4.0]$$

**Conclusió:** En mitjana A triga 2.946 unitats de temps més. B millora A, trigant entre 1.89 i 4.0 unitats de temps menys, amb una confiança del 95%

# Estudis estadístics

Un estudi estadístic treu **informació de les dades** analitzant la relació entre variables.

L'estudi ha de descriure (**Estadística Descriptiva**) les dades. Totes.

Com hem vist al bloc C, cal establir:

➤ l'obtenció de les dades	Només una m.a. justifica mesures d'incertesa com SE o ICs
➤ el paper de cada variable	Resposta Y, causa assignable X, i covariables Z
➤ el tipus d'estudi	Experimental o observacional
➤ el disseny	Aparellat, grups independents, ...

Per poder incrementar el coneixement, un estudi **científic** ha de ser **reproducible**

Si no es poden replicar els resultats, no és una investigació científica. Una investigació fracassa si no pot ser reproduïda

La incapacitat de reproduir un experiment és un problema **científic i social**:

Les investigacions metodològicament pobres són un malbaratament de recursos

En el Bloc D, veurem **eines i models** estadístics per representar relacions entre variables

## Eines estadístiques. IC, PH i NP

- La inferència estadística permet inferir o estimar característiques de la població (paràmetres) a partir de les observacions d'una mostra, per quantificar i documentar unes possibles conclusions.
- Un **interval de confiança (IC)** permet estimar un paràmetre informant dels seus valors versemblants.
- Les proves de significació (**provees d'hipòtesis o PH**) plantegen avaluar l'evidència en contra d'un valor concret d'interès del paràmetre (hipòtesis) a partir de les dades. És una metodologia conservadora que planteja una hipòtesi nul·la ( $H_0$ ) assignant un valor del paràmetre com a punt de partida i estudia si les dades proporcionen provees en contra seva. *No refutar  $H_0$  no vol dir haver demostrat que  $H_0$  és certa.*
- Els **contrastos de Neyman Pearson (NP)** serveixen per prendre decisions acotant riscos a través de dues hipòtesis. L'avantatge sobre les PH és que permeten afitar els errors al decidir per una de les dues opcions confrontades.

A l'annex d'aquest bloc D trobareu més informació sobre PH i sobre els tipus d'errors al contrastar dues hipòtesis

A la referència de la [bibliografia](#) (Estadística per a enginyers informàtics) trobareu més detalls de les provees de significació i contrast de dues hipòtesis al capítol 4

## Eines estadístiques i p-valor (p-value)

- Un IC indica, amb un cert risc, els **valors versemblants d'un paràmetre** d'acord amb l'evidència empírica que les dades aporten, implicant un **estadístic** del qual en coneixem la seva distribució
- El **P-valor** és la probabilitat d'obtenir per azar un resultat més "extrem" que el de la mostra observada **avaluant l'estadístic per a un valor concret del paràmetre** a través d'una PH

Un P-valor "petit" indica poca probabilitat que el valor concret contrastat sigui versemblant d'acord amb l'evidència de les dades. Es pot comparar el P-valor amb el risc, o per ex:

- P-valor "petit" (ex.  $< 0.001$ ) indica que és poc probable trobar una altra mostra més "extrema" (és un resultat que s'observaria menys d'una vegada cada 1000 intents), i per tant que hi ha evidència per dubtar que el valor concret contrastat sigui versemblant (és difícil justificar que les diferències entre la mostra observada i el valor contrastat es deuen només a l'atzar).
- P-valor "gran" (ex.  $> 0.001$ ) indica que és bastant probable trobar una altra mostra més "extrema" i, per tant no hi ha evidència per dubtar que el valor concret sigui versemblant (l'atzar pot explicar les diferències entre la mostra observada i el valor contrastat)

A la referència de la [bibliografia](#) (Estadística per a enginyers informàtics) trobareu més detalls en el capítol 4

R usualment indica els **estimadors** ("Estimate") complementats amb el seu **error estàndard** ("Std.Error"), i l'**estadístic** (per ex. "t" o "t value") i el **P-valor** associats a un possible valor concret del paràmetre (per ex. "Pr(>|t|)" o "p-value")

## MODELS ESTADÍSTICS. Notació

Un model estadístic explica la variabilitat d'una resposta (Y) separant una part determinista (una **fórmula amb paràmetre/s θ**) i una part aleatòria ( $\varepsilon$ , amb una distribució adequada):

$$Y_i = f(\theta) + \varepsilon_i$$

$f(\theta)$  és la part determinista que explica el valor de la resposta Y en funció dels paràmetres  $\theta$  dels quals obtindrem estimacions puntuals i intervals de confiança.  
Per exemple:

- el valor del paràmetre de tendència central (mitjana) en una mostra de la variable de resposta Y
- la diferència de mitjanes en dues mostres (aparellades o independents) de la variable de resposta Y
- l'equació d'una recta que relaciona dues variables Y i X en una mostra

$\varepsilon_i$  (soroll, error, residu,...) representa la part individual no recollida pel model determinista

- No informa sobre la relació entre les variables, i és diferent per a cada observació
- Un model és millor si té  $\varepsilon$  petites. Interessa  $V(\varepsilon)$  petita, i amb esperança 0
- Molt habitualment es modela amb una distribució Normal:  $\varepsilon \sim N(0, \sigma)$  on interessa  $\sigma$  mínima

Per a una observació  $i$ , a partir del model  $i$  i les característiques de la observació, podem obtenir:

- una **predicció**,  $\hat{Y}_i = f(\hat{\theta})$ , aplicant l'estimació de la part determinista del model
- un **error o residu** ( $\varepsilon_i$ ) fent la diferència entre l'observat  $Y_i$  i la predicció del model ( $\hat{Y}_i$ )

## MODELS ESTADÍSTICS. Casos

### a) Model sobre el valor del paràmetre $\mu$

$$Y_i = \mu + \varepsilon_i$$

Objectiu: estimació puntual i per IC de  $\mu$  ( $\hat{\mu} = \bar{y}$ )

Cas particular: igualtat del paràmetre  $\mu$  en mostres aparellades ( $Y_1, Y_2$ )

$$D_i = \mu_D + \varepsilon_i \quad D = Y_1 - Y_2 \quad (\text{la diferència com a variable de resposta})$$

Objectiu: estimació puntual i per IC de  $\mu_D$  ( $\hat{\mu}_D = \bar{d}$ )

(predicció d'un cas  $i$ )

$$\hat{Y}_i = \hat{\mu} = \bar{y}$$

### b) Model comparant el paràmetre $\mu$ en $k$ mostres independents ( $k \geq 2$ grups)

$$Y_i = \mu + \vartheta_j + \varepsilon_i$$

Objectiu: estimació puntual i per IC de cada  $\mu_j$  ( $\hat{\mu}_j = \bar{y} + \hat{\vartheta}_j$ )

El model contempla  $\mu$  com a mitjana de referència i  $\vartheta_j$  com a canvi de la mitjana del grup  $j$

### c) Model lineal simple o múltiple

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

$$\hat{Y}_i = \hat{\beta}_0 + \dots + \hat{\beta}_k X_{ki}$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i$$

$$= b_0 + b_k X_{ki}$$

Objectiu: estimació puntual i per IC dels coeficients  $\beta_j$  ( $\hat{\beta}_0 = b_0$ ,  $\hat{\beta}_1 = b_1, \dots$ ) de l'equació d'una recta

Determine una relació entre la resposta i les variables explicatives a través d'una relació lineal

- + Altres tècniques (ciència de dades: visualització multidimensional, clustering, machine learning,...)  
Per exemple: PCA [Y1, Y2, ... Yk] ---- transformació geomètrica ----> [\Psi1, \Psi2 ...] [... \Psik]

## MODELS ESTADÍSTICS. Funció lm() de R

(lm de "linear model" però només el cas c) és el que anomenarem estrictament model lineal)

En els casos anteriors de models, la funció lm() en R que li correspon és:

a)  $Y_i = \mu + \varepsilon_i$

`lm(Y~1)`

$D_i = \mu_D + \varepsilon_i$

`lm(D~1)`



b)  $Y_i = \mu + \vartheta_k + \varepsilon_i$

`lm(Y~G)`

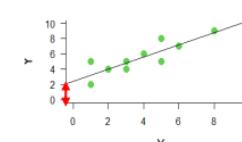
(G columna separant k grups)



c)  $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

`lm(Y~X)`

(o `lm(Y~X1+X2+...+Xk)`)



En R, els resultat de les estimacions puntuals dels paràmetres són:

➢ "Coefficients" on trobarem "Estimate" per als diferents paràmetres:

➢  $\hat{\mu}$  o  $\hat{\mu}_D$  o  $\hat{\beta}_0$  com a valors base o de referència de la resposta  $Y$  (↑)

➢  $\hat{\vartheta}_j$  com a canvi en les mitjanes entre grups (↑)

➢  $\hat{\beta}_1$  com a pendent de l'equació de la recta

També, per a cada estimació puntual, podem trobar:

➢ "Std Error" és l'error estàndard de l'estimador que permet el càlcul de IC

➢ "p-value" per avaluar la versemblança de valors concrets del paràmetre (per defecte el valor 0)

## MODELS ESTADÍSTICS. Avaluació

Avaluarem els models en dos aspectes:

- **Validació de les premisses** per comprovar i assegurar que té sentit aplicar el model  
(en principi, les comunes a tots els models seran **mostra aleatòria** i **normalitat** de la resposta)

- **Anàlisi dels resultats** que pot implicar:

- Indicadors per valorar la "qualitat" dels resultats i, si escau, la **capacitat predictiva del model**
- **Interpretabilitat** dels resultats

L'avaluació ha de permetre detectar si un model és adequat, o si pot presentar problemes com per exemple de **sobre-ajustament** o **overfitting** (més detalls al final d'aquest bloc D)

En els següents apartats es presentaran cadascun dels models indicats prèviament, amb les seves particularitats i concretant la **validació de les premisses** i l'**anàlisi dels resultats**

Abans de presentar els models, veurem el cas d'haver de transformar les dades, per exemple amb logaritmes per aconseguir complir la premissa de normalitat de la resposta. Veurem com desfer la transformació en les resultats

## MODELS ESTADÍSTICS. Transformacions

En qualsevol dels models indicats, a vegades cal una **transformació** (molt sovint **logarítmica**) de la variable resposta o de les variables explicatives per complir les premisses del model (per exemple la premissa de Normalitat)

Després cal **desfer\*** la transformació a les estimacions (puntuals o per IC) obtingudes:

- En el cas de models pel paràmetre  $\mu$ , si la normalitat es compleix per  $\ln(Y)$ , llavors el model serà  $\ln(Y) = Y' = \mu' + \varepsilon$  amb estimacions  $\hat{Y}'$  i  $IC = [\inf', \sup']$

i desfent logaritmes obtenim l'estimador puntual i l'IC buscat:  $\exp(\hat{Y}')$  i  $[\exp(\inf'), \exp(\sup')]$

- Si és el cas de mostres aparellades i la diferència  $D=Y_1-Y_2$  no és normal, no és aconsellable fer-ne el logaritme  $\ln(D)$ , ja que pot tenir valors negatius, sinó fer la diferència de logaritmes o logaritme del rati  $\ln(Y_1) - \ln(Y_2) = \ln(Y_1/Y_2) = Y'' = \mu'' + \varepsilon$  amb estimacions  $\hat{Y}''$  i  $IC = [\inf'', \sup'']$

I desfent logaritmes obtenim estimadors del valor esperat del rati  $Y_1/Y_2$ :  $\exp(\hat{Y}'')$  i  $[\exp(\inf''), \exp(\sup'')]$

- En el cas del model lineal simple o múltiple, s'ajusta una recta amb les variables transformades i després es poden desfer les transformacions a les prediccions

\* Al desfer una transformació logarítmica, ja no fem un IC de  $\mu$ , sinó de  $\mu'$  ( $\mu$  és **mitjana aritmètica** i  $\mu'$  és **mitjana geomètrica**)

## Model per estimar $\mu$

### Notació del model:

$$Y_i = \mu + \varepsilon_i$$

### Funcions de R: `lm(Y~1)` i `summary(lm(Y~1))`

(la constant 1 a la dreta de ~ és la sintaxi que usa R per indicar que només volem estimar el paràmetre  $\mu$ )

R proporciona:

- l'estimació puntual de  $\mu$  ( $\hat{\mu} = \bar{y}$ ). R ho indica com a "Estimate" de l'intercept
- l'estimació de l'error tipus (se) assumint m.a. de l'estimador anterior  
R ho indica com a "Std. Error" (se) de l'intercept ( $\bar{y}$ ), que permetria construir un IC per a  $\mu$   

$$IC(1 - \alpha\%, \mu) = \bar{y} \pm t_{n-1,1-\alpha/2} \cdot se$$
- el valor de l'estadístic ("t value") i el p-value ("Pr(>|t|)") per avaluar la versemblança d'un possible valor concret del paràmetre (per defecte el valor 0)
- la desviació residual ("Residual Standard error") o desviació de la part aleatòria que no recull el model

## Model per estimar $\mu$

### Validació de les premisses

Les **premisses** són: **mostra aleatòria i normalitat de la resposta**

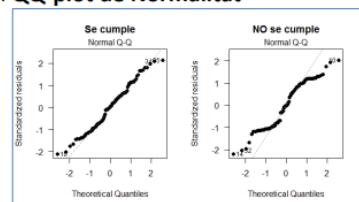
- La premissa de **mostra aleatòria** (m.a.) **no es pot verificar**. Depèn de que el disseny de recollida de dades s'hagi realitzat de forma correcta. Es podria únicament verificar la independència respecte a l'ordre de recollida de les dades.
- La premissa de **normalitat** l'avallarem gràficament amb un **QQ-plot de Normalitat**

Aquest gràfic representa els quantils empírics de les dades enfront dels quantils teòrics

Si els punts queden (aproximadament) sobre la recta no tindrem evidència per no assumir la normalitat

En R es construeix amb `qqnorm(Y)` `qqline(Y)`

(més informació a l'annex del bloc B)



### Anàlisi dels resultats

El resultat bàsic és l'estimació puntual (i per IC) de la mitjana poblacional

La desviació residual és l'indicador de la variabilitat que el model no recull

## Model per estimar $\mu$ . Exemple

Ex: mostra de 9 valors positius i negatius (per exemple errors en unes mesures)

```
> x = c(-4, -2, -1, 0, 0, 4, 8, 8, 9) # mean=2.4 sd=4.9
```

Solució per Bloc C (IC)

```
> t.test(x)
```

```
t = 1.496, df = 8, p-value = 0.173
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-1.323423 6.212312
sample estimates: mean of x 2.444444
```

Solució per Bloc D (model estadístic)

```
> (mod <- lm(Y~1))
> summary(mod)
```

Min	1Q	Median	3Q	Max
-6.444	-3.444	-2.444	5.556	6.556

Coefficients: Estimate Std.Error t value Pr(> t )
(Intercept) 2.444 1.634 1.496 0.173

Residual standard error: 4.902 on 8 degrees of freedom

$$IC(\mu, 95\%) = 2.444 \pm t_{8,0.975} 1.634 = [-1.32, 6.21]$$

L'estadístic:  $(2.444-0)/1.634 = 1.496$  amb p-value  $P(|t_8| > 1.496) = pt(-1.496, 8) + (1-pt(1.496, 8)) = 0.173$

Les premisses que s'assumeixen són mostra aleatòria i Normalitat de X

- Els resultats mostren una **estimació puntual** de 2.44 unitats de mitjana de la discrepància en les mesures. Amb **confiança** del 95% estarà entre -1.32 i 6.21.
- El p-value "gran" indica que el valor 0 és versemblant com a mitjana poblacional
- La part residual que el model no recull s'indica amb la **desviació dels residus** (4.902), o desviació tipus de la diferència en la mostra entre el valor observat i el valor de mitjana estimada

## Model per estimar $\mu$ . Cas de $\mu_D$ en mostres aparellades

### Notació del model:

$$D_i = \mu_D + \varepsilon_i$$

(amb  $D=Y_1-Y_2$ )

### Funcions de R:

```
lm(D~1) i summary(lm(D~1))
```

(la constant 1 a la dreta de ~ és la sintaxi que usa R per indicar que només volem estimar el paràmetre  $\mu$ )

R proporciona: (com el cas anterior per  $\mu$  d'una mostra)

- l'estimació puntual de  $\mu$  ( $\hat{\mu}_D = \bar{d}$ ). R ho indica com a "Estimate" de l'intercept
- l'estimació de l'error tipus (se) assumint m.a. de l'estimador anterior  
R ho indica com a "Std. Error" (se) de l'intercept ( $\bar{d}$ ), que permetria construir l'IC per a  $\mu_D$   

$$IC(1 - \alpha\%, \mu_D) = \bar{d} \pm t_{n-1,1-\alpha/2} \cdot se$$
- El valor de l'estadístic ("t value") i el p-value ("Pr(>|t|)") per avaluar la versemblança d'un possible valor concret del paràmetre (per defecte el valor 0)
- la desviació residual ("Residual Standard error") o desviació de la part aleatòria que no recull el model

### Anàlisi dels resultats

El resultat bàsic és l'estimació puntual (i per IC) de la mitjana poblacional de la diferència

En aquest cas, és interessant avaluar la versemblança del valor 0 per a la diferència de mitjanes, indicant que no es diferencien i que representen mostres aparellades amb comportament mitjà equivalent.

La desviació residual és l'indicador de la part residual que el model no recull

## Model per estimar $\mu$ . Cas de $\mu_D$ en mostres aparellades

### Validació de les premisses

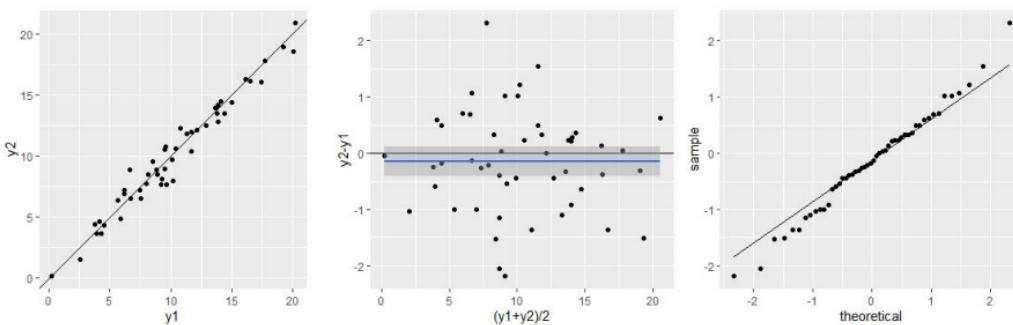
Les **premisses** són: **mostra aleatòria i normalitat de la diferència i diferència constant entre parelles (efecte additiu)**

- La premissa de **mostra aleatòria** (m.a.) **no es pot verificar** (com en el cas anterior) (únicament verificar la independència respecte a l'ordre de recollida de les dades)
- La premissa de **normalitat** l'avaluarem gràficament (com en el cas anterior, amb QQ-plot)
- En una comparació de mitjanes s'assumeix un **efecte additiu** (**un algorisme redueix el temps d'execució en 2 segons**) però a vegades es té un **efecte multiplicatiu** (**un algorisme redueix la meitat els temps d'execució**)
  - El **gràfic de Bland-Altman** (introduït al final del bloc C) representa les diferències de les respostes per cada individu en funció de les seves mitjanes. Permet estudiar si hi ha o no un efecte additiu o multiplicatiu, i decidir si convindria una transformació a les dades
  - Veurem 3 situacions especials:
    - Cas 1: sense efecte lineal (additiu)
    - Cas 2: amb efecte multiplicatiu
    - Cas 3: amb efecte multiplicatiu i transformació logarítmica:  $(\ln Y_1, \ln Y_2)$

## Model per estimar $\mu$ . Cas de $\mu_D$ en mostres aparellades

### Validació de les premisses. Cas 1: sense efecte lineal (additiu)

#### Exemple de 50 observacions aparellades ( $D = Y_1 - Y_2$ )



Cas amb diferència de mitjanes puntual estimada de 0.146 (i p-value "gran")

L'IC per la diferència de mitjanes és  $IC_{95\%}(\mu_{Y_2} - \mu_{Y_1}) = [-1.70 \text{ a } 1.99]$  (0 és valor del IC)

És a dir,  $Y_2 = Y_1 + 0.146$  en mitjana, o bé  $Y_2 = Y_1 + [-1.70 \text{ a } 1.99]$

Per tant, **no hi ha evidència de què ambdues mitjanes siguin diferents**

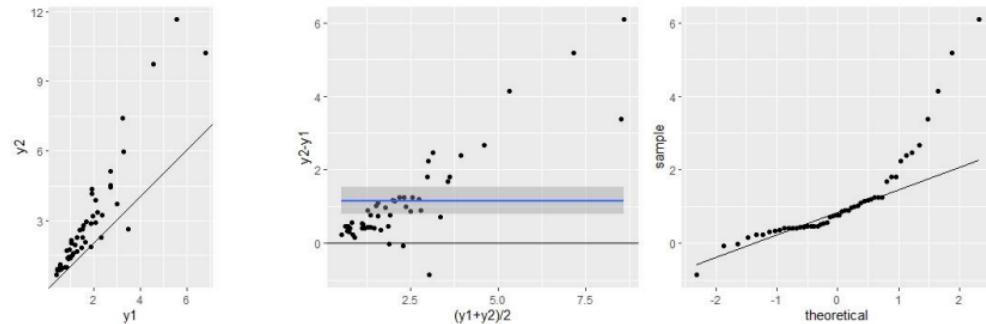
(el valor 0 per a la diferència de mitjanes és versemblant)

Es pot assumir Normalitat de la variable diferència ja que tots els quantils observats s'ajusten força bé als quantils teòrics de la Normal.

## Model per estimar $\mu$ . Cas de $\mu_D$ en mostres aparellades

### Validació de les premisses. Cas 2: amb efecte multiplicatiu

#### Nou exemple de 50 observacions aparellades



La diferència de mitjanes estimada és 1.16 amb un  $IC_{95\%}(\mu_{Y_2} - \mu_{Y_1}) = [0.38 \text{ a } 1.92]$

Però aquest valor no ens informa bé, ja que **l'efecte no és constant**

**Per a valors grans, l'efecte és més gran** i té mes variabilitat

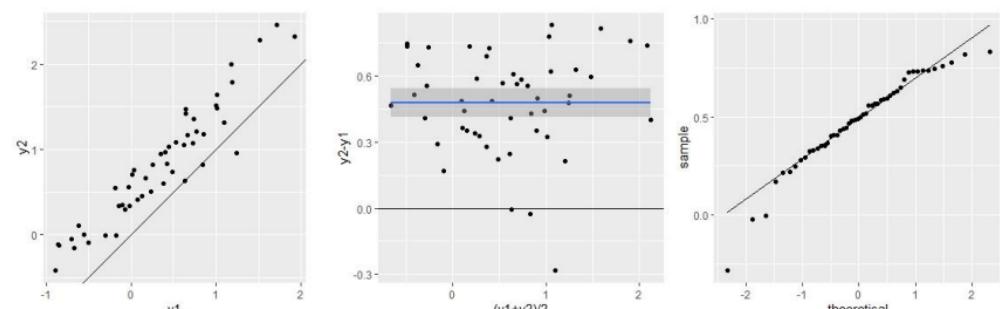
Una transformació sobre les variables que solucioni aquests problemes, farà la interpretació més fàcil → Provarem fent la transformació logarítmica (natural)

La distribució de les diferències NO és Normal

## Model per estimar $\mu$ . Cas de $\mu_D$ en mostres aparellades

### Validació de les premisses. Cas 3: amb efecte multiplicatiu i transformació log

#### Aplicar logaritmes a cada variable del exemple anterior de 50 observacions aparellades



La diferència mitjana estimada dels logaritmes és 0.48 amb  $IC_{95\%}(\mu_{Y'_2} - \mu_{Y'_1})$  de 0.21 a 0.75

Si  $Y_1, Y_2$  són les variables originals i  $Y'_1, Y'_2$  són les log-transformades:

$$Y'_2 = \ln(Y_2) \quad \rightarrow \quad Y'_2 = Y_1 + 0.48 \quad (\text{en mitjana})$$

$$Y'_1 = \ln(Y_1) \quad \rightarrow \quad \ln(Y_2) = \ln(Y_1) + 0.48 \quad \rightarrow \quad Y_2 = e^{\ln(Y_1) + 0.48} \rightarrow Y_2 = e^{\ln(Y_1)} \cdot e^{0.48} = 1.62 \cdot Y_1$$

**Interpretació:**  $Y_2$  és en mitjana 1.62 ( $IC_{95\%}$  de 1.23 a 2.12) vegades més gran que  $Y_1$

(ja que  $\exp(0.21) = e^{0.21} = 1.23$  i  $\exp(0.75) = e^{0.75} = 2.12$ )

Es pot assumir Normalitat de la variable diferència de logaritmes (= logaritme del quotient) ja que tots els quantils observats s'ajusten prou bé als quantils teòrics de la Normal.

## Model per estimar $\mu$ . Cas de $\mu_D$ en mostres aparellades

Exemple de 50 observacions aparellades ( $D = Y_1 - Y_2$ )

Solució per Bloc C (IC)	<pre>t = 0.159 , df = 49, p-value = 0.874 alternative hypothesis: true mean is not equal to 0 95 percent confidence interval: -1.70  1.99 sample estimates: mean of x  0.146</pre>
Solució per Bloc D (model estadístic)	<pre>Coefficients: Estimate Std. Error t value Pr(&gt; t ) (Intercept) 0.146 0.918 0.159 0.874 Residual standard error: 0.9054 on 49 degrees of freedom</pre>

$$IC(\mu, 95\%) = 0.146 \pm t_{49,0.975} \cdot 0.918 = [-1.70, 1.99]$$

L'estadístic:  $(0.146 - 0) / 0.918 = 0.159$  amb p-valor  $P(|t_{49}| > 0.159) = pt(-0.159, 49) + (1 - pt(0.159, 49)) = 0.874$

Les premisses que s'assumeixen són mostra aleatòria i Normalitat de  $Y_1 - Y_2$  i diferència constant

- Els resultats indiquen una **estimació puntual** de 0.146 com a mitjana esperada de la diferència, i amb **confiança del 95%** la diferència estarà entre -1.70 i 1.99
- El p-valor "gran" indica que el valor 0 és versemblant, per tant no hi ha evidència per dubtar que la diferència mitjana sigui nul·la i que les mitjanes en les dues mostres siguin equivalents (l'atzar pot explicar les discrepàncies observades en les diferències)
- La part no recollida pel model és la **desviació dels residus** o "Residual Standard error" (0.9054)

## Exemple. Compressor

Compleixen que es comprimeixin en 5 sg en mitjana? (inferència sobre la mitjana)

`lm(temp~1)`

`summary(lm(temp~1))`

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.0287	0.2379	21.14	1.33e-07 ***

Residual standard error: 0.6728 on 7 degrees of freedom

$$IC(5.0287 \pm t_{7,0.975} 0.2379 = [4.47, 5.59]$$

PH  $H_0: \mu=5$   $H_1: \mu \neq 5$  estadístic:  $(5.0287 - 5) / 0.2379 = 0.12$  (p-value  $P(|t| > 0.12) = pt(-0.12, 7) + (1 - pt(0.12, 7)) = 0.908$

5 cau dins del IC (el p-value és superior a un risc del 5%), per tant 5 segons de mitjana és un valor versemblant, la diferència entre la mitjana mostra i l'esperada és deguda a l'atzar

La part residual que el model no recull o desviació dels residus és 0.6728

Si la variància és superior a 0.22 segons<sup>2</sup> el compressor té una qualitat insufficient.

És suficient o no? (inferència sobre la desviació)

$$IC(\sigma^2, 0.95) = \left( \frac{s^2(n-1)}{\chi^2_{n-1, 1-\alpha/2}}, \frac{s^2(n-1)}{\chi^2_{n-1, \alpha/2}} \right) = \left( \frac{0.67^2(8-1)}{\chi^2_{7, 0.975}}, \frac{0.67^2(8-1)}{\chi^2_{7, 0.025}} \right) = \left( \frac{3.14}{16.01}, \frac{3.14}{1.69} \right) = (0.20, 1.86)$$

0.22 cau dins del IC, per tant té una qualitat suficient

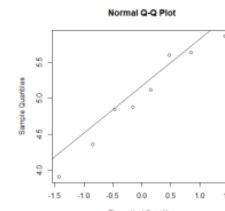
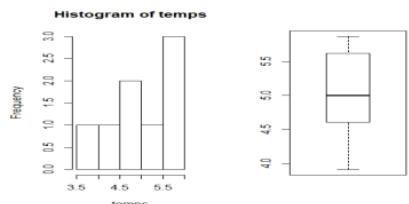
## Exemple. Compressor

El fabricant d'un determinat compressor d'arxius assegura que arxius de 500Kb són comprimits en 5 segons. Aquest compressor ha estat testejat amb 8 arxius i el temps en segons necessari per a cada compressió han estat : 4.85,4.36,5.12,5.64,5.6,5.87,3.91,4.88

R: `temp <- c(4.85, 4.36, 5.12, 5.64, 5.6, 5.87, 3.91, 4.88)`

### Estadística Descriptiva:

R: `hist(temp)`   `boxplot(temp)`   `qqnorm(temp)`   `qqline(temp)`



### Estimació puntual:

```
R: mean(temp) [1] 5.02875
sd(temp) [1] 0.6728285
var(temp) [1] 0.4526982
summary(temp)
Min. 1st Qu. Median Mean 3rd Qu. Max.
3.910 4.728 5.000 5.029 5.610 5.870
```

## Model comparant el paràmetre $\mu$ en mostres independents

### Notació del model:

$$Y_i = \mu + \theta_k + \varepsilon_i$$

El model contempla com a paràmetres:  $\mu$  com a mitjana de referència i  $\theta_k$  com a canvi de la mitjana del grup  $k$  (per tant, atenció!, no contempla les diverses  $\mu_k$ )

### Funcions de R:

`lm(Y~G)` i `summary(lm(Y~G))`

(G és una columna amb caràcters identificant els grups o factors; si és numèrica cal usar `as.factor(G)`)

R proporciona:

- l'estimació puntual de la  **$\mu$  de referència** (R ho indica com a "Estimate" de l'intercept)
- l'estimació del **canvi en les mitjanes** entre grups (indicat amb "Estimate" per a cada canvi de grup de G)
- l'estimació de l'error tipus (se) per a cada estimació anterior (indicat com a "Std. Error"). Permet calcular IC
- l'estadístic ("t value") amb el p-value ("Pr(>|t|)") per a cada estimació i que permet avaluar la versemblança d'un possible valor concret del paràmetre (per defecte el valor 0)
- la desviació residual ("Residual Standard error") o desviació de la part aleatòria que no recull el model

Altres funcions en R (veure més endavant en un exemple) permeten obtenir resultats complementaris:

<code>confint(lm(Y~G))</code>	# Interval de confiança
<code>library(emmeans)</code>	# Llibreria per calcular mitjanes
<code>emmeans(lm(Y~G), ~G)</code>	# Mitjana per cada grup
<code>plot(emmeans(lm(Y~G), ~G))</code>	# Es requereix library(ggplot2)
<code>pairs(emmeans(lm(Y~G), ~G))</code>	# Fa comparacions 2 a 2

# Model comparant el paràmetre $\mu$ en mostres independents

## Validació de les premisses

Les premisses són: **mostre aleatòria i normalitat dins de cada grup**

i **homoscedasticitat entre grups** (variabilitat semblant entre els grups)

- La premissa de **mostre aleatòria** (m.a.) no es pot verificar (com en el casos anteriors) (bàsicament verificar la independència respecte a l'ordre de recollida de les dades)
- La premissa de **normalitat** l'avaluarem gràficament dins de cada grup (amb QQ-plot)
- La premissa d'**homoscedasticitat** fa referència a igualtat de variàncies (igualtat dels paràmetres de variància desconeguts)

En general només cal una comprovació gràfica (per ex. boxplots equivalents) ja que es pretén comparar el paràmetre  $\mu$  en mostres independents que se suposen equivalents en dispersió

# Model comparant el paràmetre $\mu$ en mostres independents

## Exemple de 2 mostres independents per comparar $\mu_1$ i $\mu_2$

Solució per Bloc C (IC de la diferència)

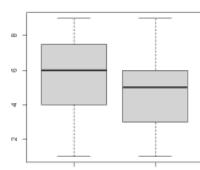
```

Y1 <- c(1,2,3,5,6,6,7,7,8,8,9)
# mean=5.6 sd=2.62
Y2 <- c(1,1,3,4,5,5,6,7,9)
# mean=4.5 sd=2.65
Y <- c(1,2,3,5,6,6,7,7,8,8,9,1,1,3...) # 2 mostres junes
G <- c(1,1,1,1,1,1,1,1,1,2,2,2...) # a lm() cal usar as.factor(G)
# mean(Y1)-mean(Y2) -> 1.080808
# o bé 5.636364 - 4.555556 = 1.080808

```

Estudiem la normalitat de Y1 i Y2 amb qqnorm(Y1) i qqline(Y1) i amb qqnorm(Y2) i qqline(Y2)

boxplot(X1, X2)



```

t.test(Y1, Y2, var.equal=T) # o t.test(Y~as.factor(G), var.equal=T)

t = 0.91335, df = 18, p-value = 0.3731
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.405312 3.566928
sample estimates: mean of x mean of y 5.636364 4.555556

```

Les premisses que s'assumeixen són mostra aleatòria, Normalitat de Y1 i Y2, i homoscedasticitat (variabilitat semblant en els dos grups comprovada gràficament al boxplot)

- Els resultats indiquen que encara que l'estimació puntual de la diferència de mitjanes és de 1.08, el rati senyal/soroll és 0.91 (amb un p-valor "gran") i per tant la diferència de mitjanes podria ser zero.
- Les dades són compatibles amb una diferència de mitjanes poblacionals des de -1.40 fins a 3.57 amb una **confiança del 95%**.
- És versemblant una diferència de 0 (igualtat de mitjanes).

# Model comparant el paràmetre $\mu$ en mostres independents

## Anàlisi dels resultats (Indicador de variabilitat explicada pels grups)

- En aquest model  $Y_i = \mu + \vartheta_k + \varepsilon_i$  podem **descomposar la variabilitat**

La variabilitat **total** de la resposta té una part **explicada** (entre els grups o factors) i una part no explicada (o **residual**)

En particular, els diferents tipus de variabilitat (*sunes de quadrats*) es defineixen d'aquesta manera:

$$\text{Variabilitat total (T): } SS_T = \sum_j \sum_i (y_{i,j} - \bar{y})^2$$

$$\text{Variabilitat residual (R) o dins dels grups: } SS_R = \sum_j \sum_i (y_{i,j} - \bar{y}_j)^2$$

$$\text{Variabilitat entre (E) els grups: } SS_E = \sum_j (\bar{y}_j - \bar{y})^2$$

- S'anomena **coeficient de determinació** o  $R^2$  al rati entre la variabilitat explicada i la total

$$R^2 = \frac{\text{variabilitat explicada}}{\text{variabilitat total}} = \frac{\text{"suma de diferències entre grups,al quadrat"}}{\text{"suma de diferències totals,al quadrat}} = \frac{SS_E}{SS_T}$$

No requereix m.a.  
ni altres premisses

Com més gran és el valor de  $R^2$ , millor representa el model la relació entre les variables:

$R^2$  és màxim (1=100%) si la relació és perfectament determinista, la part residual és 0, les mitjanes de cada grup són diferents i dins cada grup no hi ha variació, tota la variació és entre grups

$R^2$  és mínim (0) si el factor no determina res de la variació de Y, les mitjanes de cada grup són idèntiques, no hi ha variació entre grups, tota la variació és dins els grups

- R es refereix a aquest indicador com "**Multiple R-squared**" (i afegeix una variant, "Adjusted R-squared")

A la referència a [bibliografia](#) (Estadística per a enginyers informàtics) hi ha més detalls al capítol 6.6. I també veure [app](#)

# Model comparant el paràmetre $\mu$ en mostres independents

## Exemple de 2 mostres independents per comparar $\mu_1$ i $\mu_2$

Solució per Bloc D (model estadístic amb IC de la  $\mu$  de referència i de la diferència)

# els estimadors principals ara són una de les mitjanes 5.6364 com a referència, i el canvi -1.08 (5.636 - 1.08 → 4.55) lm(Y~as.factor(G))

summary(lm(Y~as.factor(G)))

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.6364	0.7938	7.100	1.28e-06 ***
G2	-1.0808	1.1833	-0.913	0.373

Residual standard error: 2.633 on 18 degrees of freedom  
Multiple R-squared: 0.04429, Adjusted R-squared: -0.008803  
F-statistic: 0.8342 on 1 and 18 DF, p-value: 0.3731

IC (intercept)  $(5.6364 \pm t_{18,0.975} 0.7938 = [3.97, 7.30])$  amb estadístic:  $(5.6364 - 0) / 0.7938 = 7.1$  i p-valor "petit"

IC  $\vartheta_k$  (G2)  $(-1.0808 \pm t_{18,0.975} 1.1833 = [-3.567, 1.405])$  amb estadístic:  $(-1.0808 - 0) / 1.1833 = -0.913$  i p-valor "gran"

- En aquests resultats també veiem que l'estimació puntual de la diferència de mitjanes és de -1.08
- El rati senyal/soroll del terme constant (intercept) és 7.1 (amb un p-valor "petit") i per tant la mitjana de la categoria de referència (G1) no és versemblant que sigui zero. I el rati senyal/soroll de la diferència de mitjanes és -0.9 (amb un p-valor "gran") i per tant la diferència de mitjanes poblacional entre G1 i G2 podria ser zero
- La part residual que el model no recull és 2.633 ("Residual Standard error" o desviació de la diferència en la mostra entre l'observat i la prediccio). És també una estimació conjunta de la desviació estàndard intragrup
- $R^2$  és 0.04429: els grups només expliquen un 4.4% de la variabilitat total

## Model comparant el paràmetre $\mu$ en mostres independents

**Exemple de 2 mostres independents per comparar  $\mu_1$  i  $\mu_2$**

Solució per Bloc D (model estadístic amb resultats de funcions complementàries)

`confint(lm(Y~as.factor(G)))`

	2.5 %	97.5 %
(Intercept)	3.968623	7.304104
G2	-3.566928	1.405312

(IC dels paràmetres  $\mu_1$  i  $\mu_2$  del model, també calculats abans)

`emmeans(lm(Y~as.factor(G)), ~G)`

G	emmean	SE	df	lower.CL	upper.CL
1	5.64	0.794	18	3.97	7.3
2	4.56	0.878	18	2.71	6.4

(Estimació i SE permeten calcular IC per la  $\mu_1$  i  $\mu_2$  dels dos grups)

`pairs(emmeans(lm(Y~as.factor(G)), ~G))`

contrast	estimate	SE	df	t.ratio	p.value
G1 - G2	1.08	1.18	18	0.913	0.3731

(Estimació i SE permeten calcular IC del canvi  $\theta_2$  entre els dos grups)

- En aquests resultats veiem, en diversos formats, els mateixos IC presentats en els resultats anteriors
- El gràfic representa els dos IC de les respectives  $\mu_1$  i  $\mu_2$  amb l'estimació puntual com un punt negre i els IC en blau. Així es poden comparar i veure si tenen solapament (és versemblant que vinguin d'un mateix valor de  $\mu$ ) o no (no seria versemblant que  $\mu_1$  i  $\mu_2$  coincideixin)

Els resultats estan afectats segons si el disseny és balancejat o no (és a dir, si els  $k$  grups tenen igual nombre d'observacions o no). El disseny balancejat és el més eficient en el sentit que proporciona un error estàndard més baix.

## Model lineal simple i múltiple

**Notació del model:**  $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$  (o  $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon_i$  si és múltiple)

El model contempla com a paràmetres els coeficients  $\beta_i$  (de l'equació d'una recta)

Els coeficients de la recta estimada en el cas **simple**  $Y = b_0 + b_1 X$  es calculen per mínims quadrats obtenint

$$b_1 = r_{XY} \cdot \frac{S_y}{S_x} = \frac{S_{XY}}{S_x^2} \quad b_0 = \bar{Y} - b_1 \bar{X}$$

i variància residual  $s^2 = \frac{\sum(e_i^2)}{n-2} = \frac{(n-1)S_y^2(1-r^2)}{n-2}$  ( $r_{XY}$  és la correlació i  $S_{XY}$  la covariància)

En el cas del model lineal simple  $b_0$  és la constant o ordenada a l'origen (**intercept**), i  $b_1$  el pendent indicant el canvi en la resposta degut a un increment en una unitat en la variable explicativa  $X$

Les variables  $X_i$  poden ser tant quantitatives com qualitatives, i les anomenem també predictores

**Funcions de R:**

`lm(Y~X)`      i      `summary(lm(Y~X))`

(o bé pel cas múltiple: `lm(Y~X1+X2+...)` i `summary(lm(Y~X1+X2+...))`)

R proporciona per a tots els coeficients de la recta ajustada obtenim

- l'estimació dels paràmetres ("Estimate") incloent la  $b_0$  (o *intercept*) com a resposta base per valors nuls o de referència de les  $X$ , i les  $b_i$  com a pendent o efecte de grup
- l'estimació de l'error tipus per a cada estimació anterior ("Std. Error"). Permet calcular IC
- estadístic senyal/soroll ("t value") amb el p-value ("Pr(>|t|)") per a cada estimació i que permet avaluar la versemblança d'un possible valor concret del paràmetre (per defecte el valor 0)
- la variància o desviació residual ("Residual Standard error") que no recull el model

## Exemple. Temps algoritme Dijkstra

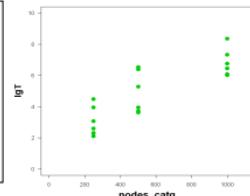
Temps i nombre de nodes de graf en l'algoritme de Dijkstra (X: nombre nodes, Y: temps amb transformació logarítmica  
 $\bar{y}_1 = 3.082 \quad s_1^2 = 0.90 \quad \bar{y}_2 = 4.910 \quad s_2^2 = 1.79 \quad \bar{y}_3 = 6.83 \quad s_3^2 = 0.79 \quad \bar{y} = 4.94 \quad s^2 = 3.5 \rightarrow s = 1.87$

El gràfic (mostres en verd) indica un canvi més gran cap a grup 3 que entre els 1 i 2.

El model ho quantifica:

Coefficients:					
	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	3.082	0.4394	7.014	4.18e-06 ***	
as.factor(nodes)500	1.8250	0.6214	2.937	0.0102 *	
as.factor(nodes)1000	3.7467	0.6214	6.030	2.31e-05 ***	

Residual standard error: 1.076 on 15 degrees of freedom  
Multiple R-squared: 0.708, Adjusted R-squared: 0.669  
F-statistic: 18.18 on 2 and 15 DF, p-value: 9.789e-05



El model recull un 70.8 % de la variabilitat total de la variable resposta. La part residual és 1.076

L'estimació de la mitjana de referència, la del grup 1, és 3.082 amb IC  $3.082 \pm qt(0.975, 15) * 0.4394$   
 $\rightarrow [2.15, 4.02]$

L'estimació del canvi de la mitjana del 1r al 2n grup és 1.825 amb IC  $1.825 \pm qt(0.975, 15) * 0.6214$   
 $\rightarrow [0.5, 3.15]$

L'estimació del canvi de la mitjana del 2n al 3r grup és 3.7467 amb IC  $3.7467 \pm qt(0.975, 15) * 0.6214$   
 $\rightarrow [2.42, 5.07]$

A partir de l'estimació de referència i de les dels canvis obtenim les estimacions de les tres mitjanes:  
**3.082**     $3.082 + 1.825 \rightarrow 4.91$      $3.082 + 3.7467 \rightarrow 6.83$

## Exemple. Temps algoritme Dijkstra

Temps i nombre de nodes de graf en l'algoritme de Dijkstra (X: nombre nodes, Y: temps amb transformació logarítmica  
 $\bar{y}_1 = 3.082 \quad s_1^2 = 0.90 \quad \bar{y}_2 = 4.910 \quad s_2^2 = 1.79 \quad \bar{y}_3 = 6.83 \quad s_3^2 = 0.79 \quad \bar{y} = 4.94 \quad s^2 = 3.5 \rightarrow S = 1.87$

Abans hem obtingut els IC dels canvis del paràmetre mitjana poblacional dels temps logarítmics entre els grafs de 250 nodes i els de 500 nodes; i entre grafs de 250 i 1000 nodes.

Per exemple, el canvi en mitjana en el temps logarítmic entre grafs de 250 i 500 nodes (el canvi entre el grup 1 i el 2) és 1.825 ( $\bar{\theta}_2 = 1.825$  del model amb resposta logarítmica)

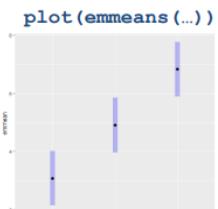
Per tant, el canvi en mitjana en el temps entre grafs de 250 i 500 és  $\exp(1.825) = e^{1.825} = 6.2$

També podem trobar l'IC per a cadascun dels paràmetres:

`emmeans(lm(lgt~as.factor(nodes)), ~nodes)`

nodes	emmean	SE	df	lower.CL	upper.CL
250	3.08	0.439	15	2.15	4.02
500	4.91	0.439	15	3.97	5.84
1000	6.83	0.439	15	5.89	7.76

Confidence level used: 0.95



Els IC de  $\mu_1$ ,  $\mu_2$  i  $\mu_3$  estan separats (veure gràfic amb IC de color blau), per tant hi ha un increment en les mitjanes esperades del logaritme del temps entre grafs de 250, 500 i 1000 nodes

Desfent logaritmes, per exemple a partir del IC al 95% del temps logarítmic per grafs de 250 nodes, obtenim l'IC del temps per grafs de 250 nodes:  
 $[\exp(2.15) = e^{2.15}, \exp(4.02) = e^{4.02}] \rightarrow [8.58, 55.7]$

## Exemple. Temps recorre arbres (preordre,inordre,postordre)

Temps <- c(392, 421, 540, 475, 427, 411, 476, 489, 499, 454, 509, 432, 539, 552, 518, 511, 438, 532, 447, 590, 566, 557, 540, 501, 575, 458, 476, 485)  
metode <- c(1,1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,3,3,3,3)  
nodes <- c(100,140,200,160,120,130,170,180,190,150,160,100,210,200,180,170,140,190,120,190,180,170,160,150,200,110,130,140)  
Per diverses mides d'arbres (en nombre de nodes) recollim el temps de recórrer-los (o linearitzar-los) usant els tres mètodes indicats. Per explicar la variable de resposta del temps, provem a estimar usant 3 models: model comparant les mitjanes dels tres mètodes, model relacionant amb el nombre de nodes, i model usant mètode i node

```
summary(lm(Temps~as.factor(metode)))
```

Coefficients:	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	458.89	16.29	28.171	<2e-16 ***
as.factor(metode)2	39.44	23.04	1.712	0.0992 .
as.factor(metode)3	60.61	22.45	2.699	0.0123 *

Residual standard error: 48.87 on 25 degrees of freedom  
Multiple R-squared: 0.2292, Adjusted R-squared: 0.1675  
F-statistic: 3.716 on 2 and 25 DF, p-value: 0.03864

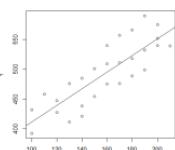
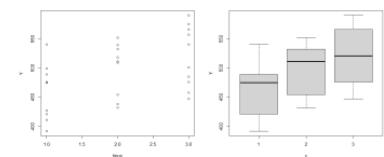
Aquest primer model només explica un 22.92 % de la variabilitat

```
summary(lm(Temps~nodes))
```

Coefficients:	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	271.2801	29.2824	9.264	1.02e-09 ***
nodes	1.3996	0.1812	7.726	3.38e-08 ***

Residual standard error: 30.06 on 26 degrees of freedom  
Multiple R-squared: 0.6966, Adjusted R-squared: 0.6849  
F-statistic: 59.69 on 1 and 26 DF, p-value: 3.379e-08

Aquest segon model explica quasi un 70% (69.66) de la variabilitat. El model és l'equació de la recta Temps = b0+b1\*nodes = 271.3+1.4\*nodes  
Per 0 nodes el temps és 271.3 (seria un temps fixe) i per cada node de més, en el temps podem esperar un augment de 1.4 unitats de temps



## Exemple. Recorregut arbre (preordre,inordre,postordre)

```
summary(lm(Temp~nodes+as.factor(metode)))
```

Coefficients:	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	239.78109	15.88211	15.098	9.43e-14 ***
nodes	1.41868	0.09699	14.628	1.87e-13 ***
as.factor(metode)2	22.10498	7.56022	2.924	0.00743 **
as.factor(metode)3	59.82295	7.27785	8.220	1.95e-08 ***

Residual standard error: 15.84 on 24 degrees of freedom  
Multiple R-squared: 0.9223, Adjusted R-squared: 0.9125  
F-statistic: 94.91 on 3 and 24 DF, p-value: 1.892e-13

El model recull un 92.23 % de la variabilitat total de la variable resposta. La part residual és 15.84 (bastant inferior a la dels models anteriors)

Ara l'intercept és 239.78 amb error estàndard 15.88, per tant l'IC  $239.78 \pm qt(0.975, 24) * 15.88211 \rightarrow [207.001, 272.56]$

L'estimació del pendent és 1.42 amb IC  $1.42 \pm qt(0.975, 24) * 0.09699 \rightarrow [1.22, 1.62]$

L'estimació del canvi de la mitjana del 1r al 2n mètode és 22.105 amb IC  $22.105 \pm qt(0.975, 24) * 7.56 \rightarrow [6.5, 37.71]$

L'estimació del canvi de la mitjana del 1r al 3r mètode és 59.823 amb IC  $59.823 \pm qt(0.975, 24) * 7.27785 \rightarrow [44.8, 74.84]$

## Model lineal simple i múltiple. Exemple

### Exemple de model amb variables explicatives quantitatives i qualitatives

El forat de gènere (*gender pay gap*) es refereix a la diferència de sou existent entre un treballador home i una treballadora dona. Una empresa de consultoria estudia en una mostra de 30 homes i 20 dones les relacions entre salari amb experiència, edat i sexe

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	44.0387	10.9254	4.031	0.000207
xp	1.1575	0.5638	2.053	0.045789
age	-0.3088	0.4683	-0.659	0.512878
Gw	-5.2430	0.8572	-6.116	1.94e-07

Residual standard error: 2.663 on 46 degr of freedom  
Multiple R-squared: 0.7647, Adjusted R-squared: 0.7494

- A "Estimate" hi ha l'**estimació puntual** del pendent per a les quantitatives, o de l'efecte de grup per a les qualitatives
- La constant ("intercept") ara es refereix a l'estimació de la resposta mitjana per a les categories de referència (aqui "home"), i per al valor 0 de les variables quantitatives (no és interpretable una edat de 0 anys)
- Els coeficients ajustats per a cada variable:
  - Cada any més de vida laboral representa un increment de salari de **1157.5 €**
  - Donada una experiència fixada, cada any més d'edad, **308.8 €** menys. Donada la correlació, tènicament col-linealitat, entre edat i experiència és molt difícil interpretar-lo: "entre persones amb la mateixa experiència, p.e., 10 anys, cada any més d'edad representa 309 € menys."
  - Amb les mateixes experiència i edat, si es tracta d'una dona, el salari mitjà es redueix en **5243 €**

## Model lineal simple i múltiple

### Validació de les premisses

Les **premisses** són:

- **linealitat** (la forma del núvol de punts s'ajusta a una recta o a un pla en el cas múltiple)
- **mostra aleatòria** (implica independència dels residus)
- **normalitat dels residus**
- **homoscedasticitat dels residus** (variabilitat homogènia dels residus)

La linealitat fa referència a la part determinista, mentre que les altres fan referència a la part aleatòria o residual (per això el que es valida és la independència, normalitat i homoscedasticitat dels residus):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \epsilon_i \rightarrow \epsilon_i \sim N(0, \sigma^2)$$

Linealitat
Normalitat
Homoscedasticitat
Independència

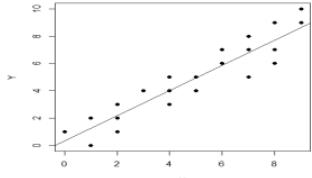
Ara veurem aquestes premisses en gràfics on podrem comprovar-les

# Model lineal simple i múltiple

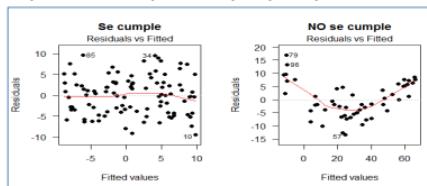
## Validació de les premisses

Les **premisses** del model lineal i els gràfics on estudiar-les són:

- **Linealitat** ( $Y$  i  $X$  s'ajusten a una recta, pel cas simple, o a un pla o hiperplà, pel cas múltiple)

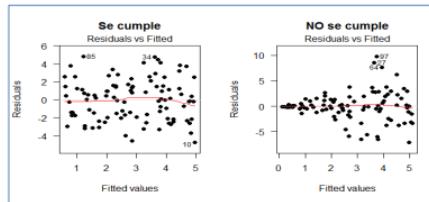


Plot  $Y$  i  $X$  on veure si s'ajusta a una recta (en aquest cas sí)



Gràfics de residus enfront les prediccions on veure si estan per sota i per sobre del 0 uniformement. Esquerra compleix linealitat; i dreta no

- **Homoscedasticitat** ( $\epsilon \sim N(0, \sigma)$  o variabilitat constant)

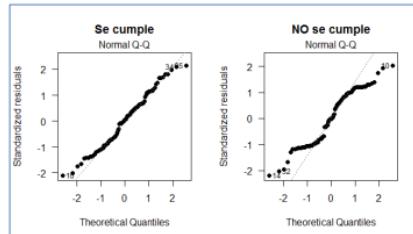


Gràfic dels residus enfront les prediccions on veure que es distancien del zero de la mateixa forma, sense zones amb més i menys dispersió.  
En aquests gràfics:  
Esquerra compleix homoscedasticitat; i dreta no

# Model lineal simple i múltiple. Validació

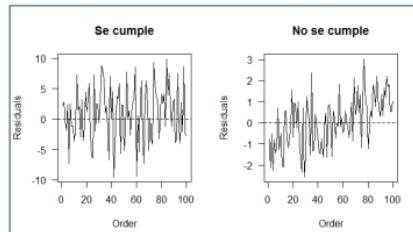
## Validació de les premisses

- **Normalitat** ( $\epsilon \sim N(0, \sigma)$  residus amb distribució Normal)



qqnorm i qqline(on veure si els residus s'ajusten al model Normal  
En aquests gràfics un compleix i l'altre no

- **Independència** (mostra aleatòria i no dependència entre observacions)



La independència es garanteix amb un bon disseny i recollida de dades  
Però observant els residus enfront l'ordre de recollida es pot veure si hi ha alguna dependència. S'espera no trobar cap patró específic

En aquest cas un compleix i l'altre no, perquè hi ha una tendència creixent dels residus al llarg del temps

# Model lineal simple i múltiple

## Anàlisi dels resultats (Indicador de capacitat predictiva)

- En el model lineal simple i múltiple (tal com hem vist en el model de comparar mitjanes en mostres independents) podem **descomposar la variabilitat**

La variabilitat **total** de la resposta  $Y$  té una part **explicada** (variacions en algun predictor  $X$  suposa canvis en la  $Y$ ) i una part no explicada, restant o **residual**

En particular, així es fa la descomposició de la variabilitat

$$\sum (y_i - \bar{Y})^2 = \sum (y_i - \hat{y})^2 + \sum (\hat{y} - \bar{Y})^2$$

$$SQ_{Total} = SQ_{Residual} + SQ_{Explicada}$$

- Anomenem **coeficient de determinació** o  $R^2$  al rati entre la variabilitat explicada pel model i la variabilitat total de la resposta

$$R^2 = \frac{\text{variabilitat explicada}}{\text{variabilitat total}} = \frac{SQ_E}{SQ_T}$$

No requereix m.a.  
ni altres premisses

Com més gran és el valor de  $R^2$ , millor representa el model la relació entre les variables

$R^2$  és màxim (1 = 100%) si la relació és perfectament determinista, la part residual és zero

$R^2$  és mínim (0) si el model no determina res de la variació de  $Y$  que prové de la part aleatòria i no de les  $X$

En regressió lineal simple,  $R^2 = (r_{XY})^2$ , és a dir, equivalent al quadrat del coeficient de correlació lineal  
 $R^2$  és indicador de "bondat" de l'ajust o **capacitat predictiva** i  $r_{XY}$  ho és de l'associació de les variables

- R es refereix a aquest indicador com "**Multiple R-squared**" (i afegeix una variant, "Adjusted R-squared")

A la referència de la [bibliografia](#) (Estadística per a enginyers informàtics) trobareu més detalls al capítol 6.6

# Model lineal simple i múltiple

## Anàlisi dels resultats (Funcions R per fer prediccions)

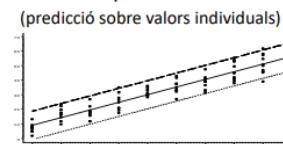
- A partir d'una predicció puntual (aplicant la fórmula de l'estimació de la part determinista del model) es pot calcular un interval a la predicció amb dos tipus d'enfocament:

➤ **Valor individual**. Estimar el **valor** de la resposta per a **una** observació amb uns valors concrets de les variables predictors [Ex: Quin és el retard *previsible* pel vol BCN-ROM de Vueling de les 8:00?] **predict(..., interval = "prediction")**

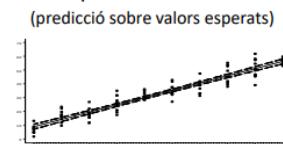
➤ **Valor esperat**. Estimar la **mitjana** de la resposta en **totes** les observacions amb uns valors concrets de les variables predictors [Ex: Quin és el retard *esperat* per als vols BCN-ROM de Vueling que surten a les 8:00?] **predict(..., interval = "confidence")**

En ambdós casos, **l'estimació puntual de la predicció és la mateixa, però no la seva incertesa**.

En el cas del valor puntual tenim un rang més ampli de valors plausibles:



(predicció sobre valors individuals)



(predicció sobre valors esperats)

- R proporciona la funció **predict()**:

```
predict(lm( )) # predicció puntual a les obs
new <- data.frame([nom_columna]=[valor on fer pred],...)
predict(lm( ), new ) # pred puntual a nova obs
predict(lm( ), new, interval=...) # pred per interval 95%
predict(lm( ), new, interval=..., level=...)
```

Indicant només el model, fa prediccions puntuals a les pròpies observacions. Però és pot indicar:

- valors de noves observacions (new) per predir
- el tipus d'interval a partir de la predicció puntual
- el nivell de confiança per l'interval

A la referència de la [bibliografia](#) (Estadística per a enginyers informàtics) trobareu més detalls al capítol 7.2

## Model lineal simple i múltiple. Exemple

### Exemple de mostra de dues variables amb relació lineal

```
Y <- c(1,0,1,3,4,5,4,6,2,6,5,2,9,7,7,5,6,9,3,6,8,7,10,5,4,6)
```

```
X <- c(0,1,2,2,3,4,4,6,1,8,7,2,9,6,8,5,6,8,4,6,7,7,9,7,5,6)
```

```
cor(X,Y) # = 0.9239073
```

```
lm(Y~X)
```

```
summary(lm(Y~X))
```

```
Residuals:
    Min      1Q Median      3Q     Max 
-1.7710 -0.8719  0.1483  0.8054  1.3903
```

```
Coefficients:
```

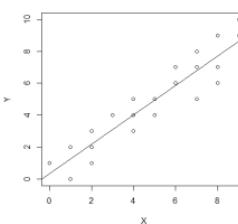
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.33581	0.44460	0.755	0.457
X	0.91931	0.07771	11.830	1.68e-11 ***

Residual standard error: 1.015 on 24 degrees of freedom

Multiple R-squared: 0.8536, Adjusted R-squared: 0.8475

F-statistic: 139.9 on 1 and 24 DF, p-value: 1.68e-11

```
plot(X,Y)
abline(lm(Y~X))
```



- El model recull un **85.36 % de la variabilitat total** de la resposta. La resta és **residual** amb desviació de 1.015
- L'**ordenada a l'origen (o terme independent)** és 0.33581 amb IC  $0.33581 \pm qt(0.975, 24) * 0.4446 \rightarrow [-0.58, 1.25]$
- El **pendent (o terme lineal)** és 0.91931 amb IC  $0.91931 \pm qt(0.975, 24) * 0.07771 \rightarrow [0.76, 1.08]$   
Per a cada unitat més a X, el pendent indica 0.919 unitats més a Y (amb **95% de confiança** entre 0.76 i 1.08)
- 0 no és un valor versemblant per al pendent (p-value "petit" i no és a l'interval). Però 1 si és versemblant: +1 a X implica +1 a Y  
0 sí és un valor versemblant per a l'ordenada a l'origen (és a l'interval i el p-value és "gran")  
Amb aquests possibles valors, la recta  $Y = b_0 + b_1 X$  seria  $Y=X$ , indicant una relació d'identitat entre X i Y

## Model lineal simple i múltiple

### Exemple de mostra de dues variables amb relació lineal

```
Y <- c(1,0,1,3,4,5,4,6,2,6,5,2,9,7,7,5,6,9,3,6,8,7,10,5,4,6)
```

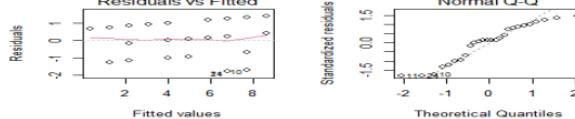
```
X <- c(0,1,2,2,3,4,4,6,1,8,7,2,9,6,8,5,6,8,4,6,7,7,9,7,5,6)
```

```
cor(X,Y) # = 0.9239073
```

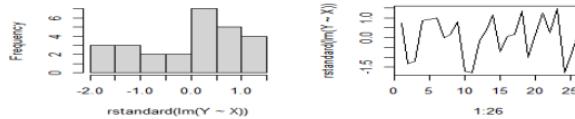
```
lm(Y~X)
```

```
par(mfrow=c(2,2))
plot(lm(Y~X), c(2,1)) # o bé plot(model, c(2,3))
```

```
hist(rstandard(lm(Y~X)))
plot(rstandard(lm(Y~X)), type="l")
```



Histogram of  $rstandard(lm(Y ~ X))$



Totes les **premisses** semblen acceptables, tot i que al QQ-plot es pot sospitar que la Normalitat falla (a la vista dels extrems, no hi ha cues llargues)

## Exemple. Benzina i velocitat (model lineal)

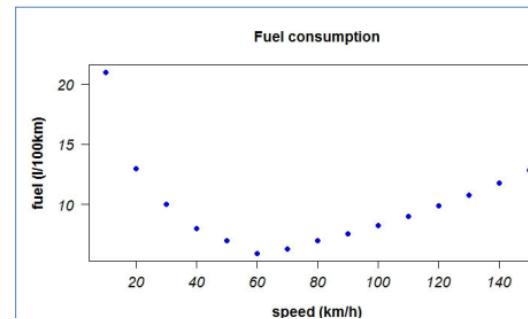
• Una equació com  $Y = b_0 + b_1 \cdot X$  pot relacionar-nos dues variables com el consum de benzina i la velocitat (dades a la taula)

• Així, tenim un model per previsions del **consum (Y)** segons la **velocitat (X)**:

$$Y = 11.058 - 0.01466 \cdot X$$

• *Què vol dir el coeficient  $-0.01466$ ? Realment podem esperar menys consum amb més velocitat veient el gràfic?*

• A més, no oblidem que el consum de benzina no depèn només de la velocitat.



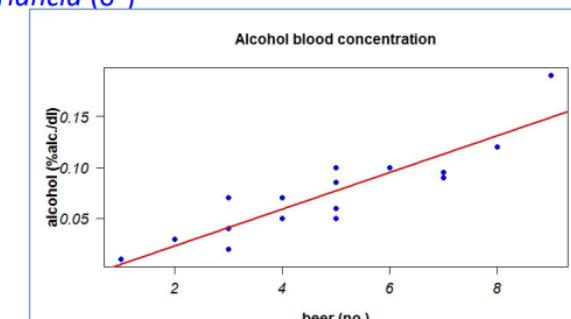
speed (km/h)	fuel (l/100 km)
10	21
20	13
30	10
40	8
50	7
60	5.9
70	6.3
80	6.95
90	7.57
100	8.27
110	9.03
120	9.87
130	10.79
140	11.77
150	12.83

## Exemple. Cervesa i alcohol (model lineal)

• Un estudi ha sol·licitat a 16 voluntaris que es prengui una quantitat determinada (aleatoriament) de cervesa, mesurada en llaunes, i es mesura l'alcohol a la sang trenta minuts després [%alc. /dl sang].

• Un model simple és ajustar-hi una recta, que implica dos paràmetres: **pendent ( $\beta_1$ )** i **constant ( $\beta_0$ )** a l'origen

• Al voltant tenim una certa dispersió que requereix un tercer paràmetre: la **variància ( $\sigma^2$ )**



cerveses	alcohol
5	0.100
2	0.030
9	0.190
8	0.120
3	0.040
7	0.095
3	0.070
5	0.060
3	0.020
5	0.050
4	0.070
6	0.100
5	0.085
7	0.090
1	0.010
4	0.05

Source: The Basic Practice

of Statistics. 4th ed.

David S. Moore.

Example 24.7

## Exemple. Cervesa i alcohol (model lineal)

cerveses	alcohol
5	0.100
2	0.030
9	0.190
8	0.120
3	0.040
7	0.095
3	0.070
5	0.060
3	0.020
5	0.050
4	0.070
6	0.100
5	0.085
7	0.090
1	0.010
4	0.05

### Càlculs dels estadístics convencionals:

$$\bar{y} = 0.07375 \quad s_y^2 = 0.0019483 \quad s_{XY} = 0.08675$$

$$\bar{x} = 4.8125 \quad s_x^2 = 4.829167 \quad r_{XY} = \frac{s_{XY}}{s_X s_Y} = 0.894338$$

### Resultats de la regressió:

$$b_1 = \frac{s_{XY}}{s_X^2} = r_{XY} \cdot \frac{s_Y}{s_X} = 0.01796 \quad b_0 = \bar{Y} - b_1 \bar{X} = -0.0127 \quad s = \sqrt{\frac{\sum(e_i^2)}{n-2}} = 0.0204$$

### Model amb R:

```
> lm(alc ~ n.cerv) # (alc és Y, n.cerv és X)
Call:
lm(formula = alc ~ n.cerv)

Coefficients:
(Intercept)      n.cerv
-0.01270       0.01796
```

### Variància de l'error amb R:

```
sum(lm(alc~n.cerv)$resid^2)/14
```

## Exemple. Cervesa i alcohol (model lineal)

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01270 0.01260 -1.00 0.33
n.cerv 0.01800 0.00240 7.48 3.0e-06 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0204 on 14 degrees of freedom
Multiple R-squared:  0.8, Adjusted R-squared:  0.786
F-statistic: 55.9 on 1 and 14 DF, p-value: 2.97e-06
```

$$IC_{95\%}: IC(\beta_1, 95\%) = b_1 \mp t_{n-2, 0.975} \cdot s_{b_1} = 0.018 \mp 2.15 \cdot 0.0024 = [0.013, 0.023]$$

(Cada cervesa de més incrementa el contingut d'alcohol per decilitre de sang en un valor que pot estar entre 0.0128% i 0.0231%, amb un 95% de confiança)

**Conclusió pràctica:** No és versemblant que el coeficient del pendent sigui 0

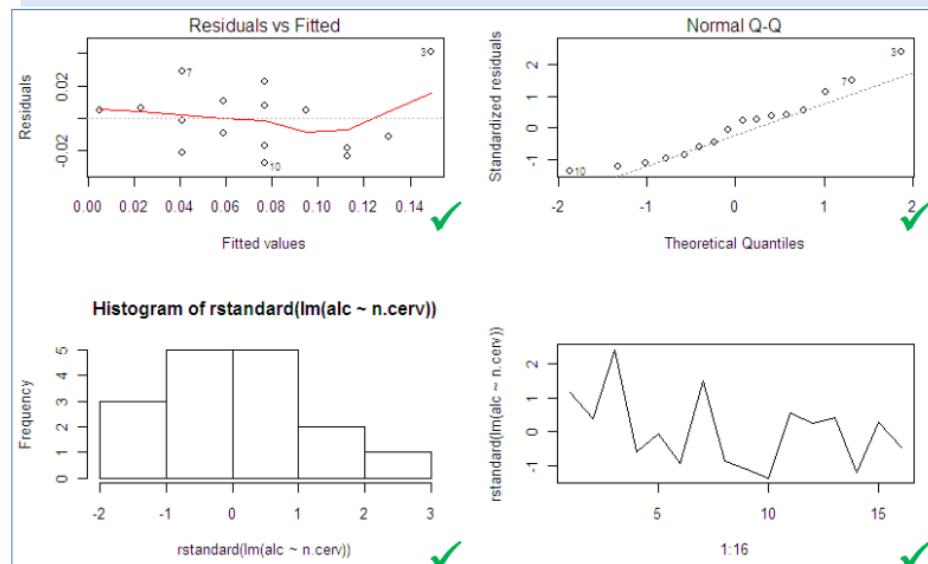
$$IC_{95\%}: IC(\beta_0, 95\%) = b_0 \mp t_{n-2, 0.975} \cdot s_{b_0} = 0.0127 \mp 2.15 \cdot 0.0126 = [-0.040, 0.014]$$

(És versemblant que el terme independent sigui 0. No es pot rebutjar que la recta passi per l'origen, pel punt (0,0). A 0 llaunes de cervesa li correspon una quantitat d'alcohol en sang de 0.0%)

## Exemple. Cervesa i alcohol (model lineal, validació)

```
##-- Exemple de les cerveses
par(mfrow=c(2,2))
plot(lm(alc~n.cerv),c(2,1))
hist(rstandard(lm(alc~n.cerv)))
plot(1:16,rstandard(lm(alc~n.cerv)),type="l")
```

# QQ-Norm i Standard Residuals vs. Fitted  
# Histograma dels residus estandarditzats  
# Ordre dels residus estandarditzats



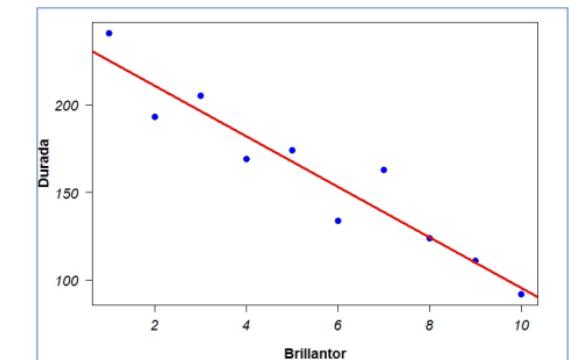
Encara que són poques dades, res s'oposa a validar cap de les 4 premisses

## Exemple. Brillantor i durada (model lineal)

La pantalla de l'ordinador portàtil és l'element que consumeix més energia del sistema. Per estudiar l'impacte que el nivell de brillantor de la pantalla (que l'usuari pot graduar) té en la durada de la bateria, treballant amb tasques quotidianes, es mesura el temps que l'ordinador triga des que arrenca amb la bateria totalment carregada fins que avisa per manca d'energia suficient per continuar. Els resultats obtinguts figuren a continuació:

Brillantor (X)	1	2	3	4	5	6	7	8	9	10
Durada (Y)	241	193	205	169	174	134	163	124	111	92

Varia la durada de la bateria segons el nivell de brillantor?



```
> plot(Durada~Brillantor)
> abline(lm(Durada~Brillantor))
```

## Exemple. Brillantor i durada (model lineal)

$$\begin{aligned} \bar{y} &= 160.6 \\ s_y^2 &= 2106.044 \\ \bar{x} &= 5.5 \\ s_x^2 &= 9.167 \\ s_{xy} &= -132.11 \\ r_{xy} &= s_{xy}/(s_x s_y) = -0.95 \end{aligned} \quad \left\{ \begin{array}{l} b_1 = \frac{s_{xy}}{s_x^2} = r_{xy} \cdot \frac{s_y}{s_x} = -14.41 \\ b_0 = \bar{y} - b_1 \bar{x} = 239.9 \\ s^2 = \frac{\sum e_i^2}{n-2} = 227.3 \end{array} \right.$$

```
> summary(lm(Durada ~ Brillantor, datos))
Call:
lm(formula = Durada ~ Brillantor, data = datos)
Residuals:
    Min      1Q  Median      3Q     Max 
-19.3939 -10.8500  0.1364  7.8258 24.0182 
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 239.87    10.30   23.290 1.23e-08 ***
Brillantor   -14.41     1.66  -8.683 2.41e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
Residual standard error: 15.08 on 8 degrees of freedom 
Multiple R-squared:  0.9041,    Adjusted R-squared:  0.8921 
F-statistic: 75.39 on 1 and 8 DF,  p-value: 2.411e-05
```

Recta resultant:  $\hat{y}_i = 239.9 - 14.41x_i$

Interpretació de  $b_1$ : Per cada grau de brillantor augmentat, la bateria dura uns 14.4 minuts menys.

IC<sub>95%</sub>:  $IC(\beta_1, 95\%) = b_1 \mp t_{n-2, 0.975} \cdot s_{b_1} = -14.41 \mp 2.3 \cdot 1.66 = [-18.23, -10.59]$

(Cada grau que pugem la brillantor de la pantalla significa entre uns 10 i 18 minuts menys de durada de la bateria)

Interpretació de  $b_0$ : Amb un grau de brillantor nul (sense usar la pantalla), la bateria durarà unes 4 hores (239.9 minuts)

Interpretació de la  $s$ : la desviació residual és 15.1. Podem esperar fluctuacions d'uns quinze minuts respecte les previsions de durada en funció de la brillantor que ens doni el model

## Exemple. Brillantor i durada (model lineal, predicció)

Les dades de l'exemple de la pantalla d'ordinador

Brillantor (X)	1	2	3	4	5	6	7	8	9	10
Durada (Y)	241	193	205	169	174	134	163	124	111	92

Havíem trobat que la recta estimada era:

$$\hat{y}_i = 239.9 - 14.41x_i$$

Quina durada podem esperar per a pantalles de brillantor 7.5?

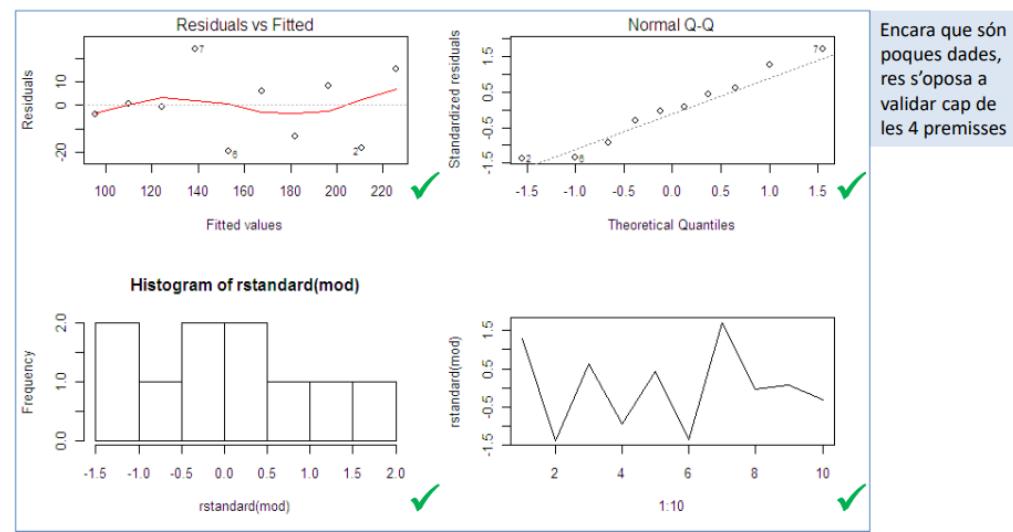
$$\begin{aligned} \bar{x} &= 5.5 \\ s_x^2 &= 9.167 \\ s^2 &= 227.3 \end{aligned}$$

	Valor esperat	Valors individuals
Estimació puntual	$\hat{y}_h = 239.9 - 14.41 \cdot 7.5 = 131.8$ new <- data.frame("X"=7.5) predict(lm(Y~X), new) $\rightarrow 131.8$	$\hat{y}_h = 239.9 - 14.41 \cdot 7.5 = 131.8$ new <- data.frame("X"=7.5) predict(lm(Y~X), new) $\rightarrow 131.8$
Estimació per interval	predict(lm(Y~X), new, int="confidence") fit lwr upr 1 131.8 118.4 145.2	predict(lm(Y~X), new, int="prediction") fit lwr upr 1 131.8 94.5 169.0
Conclusió	Per a les pantalles de brillantor de 7.5 podem esperar una durada mitjana entre 118.4 i 145.2 min. amb una confiança del 95%	Per a una pantalla de brillantor 7.5 podem esperar una durada entre 94.5 i 169.0 min. amb una confiança del 95%

Veure gràfics de pags. 191-192 a *Estadística per a enginyers informàtics*. Ed UPC

## Exemple. Brillantor i durada (model lineal, validació)

```
##-- Exemple de la pantalla d'ordinador
par(mfrow=c(2,2))
plot(lm(Durada ~ Brill), c(2,1)) # QQ-Norm i Standard Residuals vs. Fitted
hist(rstandard(lm(Durada ~ Brill))) # Histograma dels residus estandarditzats
plot(1:10, rstandard(lm(Durada ~ Brill)), type="l") # Ordre dels residus
```



## Exemple. Modem (model lineal, predicció)

```
> modem$TamlMb
[1] 1.59129 1.59129 0.51858 1.29297 0.14062 0.22461 0.66895 2.68000
> modem$Tp01Mb
[1] 23.22 14.56 6.07 13.50 1.38 2.24 5.95 23.45
> mod1 = lm(TpolMb ~ TamlMb, data=modem)
> summary(mod1)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.908      1.962     0.46  0.65995  
TamlMb      9.544      1.447     6.59  0.00058 ***
...
Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.3322     0.0679    34.3  4.1e-08 ***
Log.tamlmb  1.0083     0.0673    15.0  5.6e-06 ***
> predict(mod2, int="prediction")
  fit      lwr      upr
1 2.80061 2.30739 3.29384
2 2.80061 2.30739 3.29384
3 1.67006 1.18913 2.15100
```

$$\begin{aligned} \bar{x} &= -0.293 \\ s_x^2 &= 1.065 \\ s^2 &= 0.0338 \end{aligned}$$

## Bloque 4 – Inferencia estadística

Conceptos básicos:

- **Parámetro:** parte de la **población** que queremos hacer la estimación.
- **Muestra:** es una fracción del parámetro que se usa para obtener datos, ya que analizar a toda la parte de la población que queremos analizar se haría muy difícil.
- **Estadístico:** Cualquier indicador que se obtenga a partir de los datos de la muestra.
- **Estimador:** estadístico de una muestra que se usa para obtener el valor de un parámetro de la población.

Estimación puntual de  $\mu$  es la media muestral  $\bar{x} = \frac{\sum x_i}{n}$

Error tipo o error estándar es la variabilidad (valores muy dispersos) del estimador  $\rightarrow se = \sigma/\sqrt{n}$

Cuando la  $\sigma$  es desconocida  $\rightarrow \widehat{se} = 1/\sqrt{n}$

Parámetro ( $\theta$ ) (población)	Estimador ( $\bar{\theta}$ ) (Muestra)
$\mu$ (esperanza, media poblacional)	$\bar{x}$ (media muestral)
$\sigma^2$ (variancia poblacional)	$s^2$ (variancia muestral)
$\sigma$ (desviación tipo poblacional)	$s$ (desviación tipo muestral)
$\pi$ (probabilidad)	$p$ (proporción)

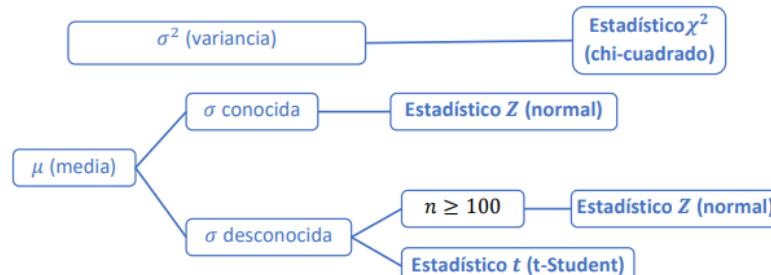
Propiedades de los estimadores:

- No tener sesgo (biaix en catalán) → Cuando la diferencia entre la esperanza y la estima es nula.
- Ser eficiente → Cuando no tiene sesgo y la variancia es menor.

### Intervalos de confianza (estimar un parámetro)

Mecánica

- 1- Definir el estadístico usado
- 2- Especificar distribución



- 3- Indicar las premisas necesarias para decir que sigue la distribución
- 4- Delimitar el nivel de confianza (usualmente  $1 - \alpha = 95\% \rightarrow \alpha = 5\%$ )  
Dónde  $\alpha = 1 - \text{porcentaje que nos dicen}$
- 5- Calcular el intervalo → Usando la distribución especificada
- 6- Interpretar el resultado → El tanto % de las veces VA estará en el intervalo dado

El **error tipo** es igual al denominador del estadístico.

- **Error tipo de la media:** es la desviación "habitual" de la media muestral  $\bar{x}$  respecto a la media de la población  $\mu$ . Se calcula  $S^2/\sqrt{n}$ .

- **Error tipo de la proporción:** es la desviación "habitual" de la proporción de la muestra  $p$  respecto a la proporción real  $\pi$ . Se calcula  $\sqrt{\frac{\pi \cdot (\pi - 1)}{n}}$

Nota:  $\pi = p$ , pero si no hay diferencia  $\rightarrow \pi = p = 0,5$ .

### Pruebas de Hipótesis (Refutar un parámetro)

- 1- Escoger una variable según los objetivos del estudio (La variable que queremos demostrar)
- 2- Escoger un diseño y un estadístico



- 3- Definir la hipótesis nula  $H_0: \mu = \text{media o } \pi = \text{proporción}$  y la hipótesis alternativa  $H_1$   
 $H_1: \mu \neq \text{media o } \pi \neq \%$  bilateral  
 $H_1: \mu \neq \text{media o } \pi < \%$  unilateral
- 4- Especificar la distribución del estadístico si  $H_0$  fuera cierto (y sus premisas)
- 5- Contrastar  $H_0$  Dos alternativas para hacerlo
  - a. Si  $|z| > z_{1-\alpha}$  (unilateral) o  $|z| > z_{1-\alpha/2}$  (bilateral)  $\rightarrow H_0$  Se rechaza (es poco fiable)
  - b. Calcular el valor P  $\rightarrow$  Si  $P < \alpha \rightarrow H_0$  Se rechaza (es poco fiable)
- 6- Añadir la estimación para el intervalo  $IC(1 - \alpha)$

### Intervalos de confianza

```

binom.test() # Binomial
# Estadístico t-Student
t.test(datos1, datos2, var.equal=TRUE, conf.level=prob)$conf.int # Independientes
t.test(Diferencia^2, var.equal=TRUE, conf.level=prob)$conf.int # Apareadas
var.test() # Fisher
prop.test() # Proporciones
chisq.test() # Proporciones Pearson
  
```

```

mean () # Calcula la media
sd() # Calcula la desviación
var() # Calcula la variancia
pt, pf, pnorm # Para calcular el p-valor
qt, qf, qnorm # Para calcular el punto crítico
  
```

## Bloque 6 – Previsión

**Estudios observacionales:** se trata de estudios dónde vemos lo que sucede y sirven para predecir, anticipar, prever...

**Estudios experimentales:** se trata de estudios en los que podemos interactuar, por lo que podemos intervenir y cambiar el futuro.

**Modelo:**  $Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$       Parte determinista de  $Y$   
Parte aleatoria de  $Y$

### Parámetros

$\beta_0$  = Constante al origen       $Y_i$  = Valor de la variable respuesta Y en el caso i-ésimo

$\beta_1$  = Pendiente de la recta       $X_i$  = Valor que toma la condición X en el caso i-ésimo

$\varepsilon_i$  = Error aleatorio o distancia de la recta del caso i-ésimo / error de predicción

$S^2 = \sigma^2$  = Variancia residual o variancia de los  $\varepsilon_i$  → cuanto mayor sea  $\sigma$  los valores estarán más dispersos y por lo tanto mayor variabilidad habrá.

$S$  = Desviación típica del término aleatorio del modelo

$\hat{\beta}_0 = b_0$  = Estimación del término independiente

$\hat{\beta}_1 = b_1$  = Estimación término lineal / pendiente estimada

} Son los estimadores de  $\beta_0$  y  $\beta_1$

$\bar{Y}$  = Media de la variable respuesta

$S_{XY}$  = Covariancia muestral

$r = r_{XY}$  = Correlación muestral / coeficiente de Pearson

$r_{XY}^2 = R^2$  = coeficiente de determinación

$0 \leq R^2 \leq 1 \rightarrow \begin{cases} \text{más cerca del 1} \rightarrow \text{más capacidad predictiva y poca variabilidad} \\ \text{más cerca del 0} \rightarrow \text{menos capacidad predictiva y mucha variabilidad} \end{cases}$

$S_Y$  = Desviación tipo de la variable respuesta / error de estimación del término independiente

$S_X$  = Desviación tipo de la condición / error de estimación del término lineal

## Interpretación de los parámetros

Los **parámetros** de la recta han de ser interpretados de acuerdo con sus unidades

La **pendiente** se interpreta:

- Experimentos: La respuesta  $Y$  tendrá un cambio esperado de  $\beta_1$  (unidades de  $Y$ ) por cada incremento de 1 unidad de la causa  $X$ .
- Previsión: una variación de 1 unidad en la variable  $X$  se asocia a una variación de  $\beta_1$  unidades en la variable  $Y$ .

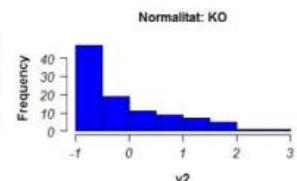
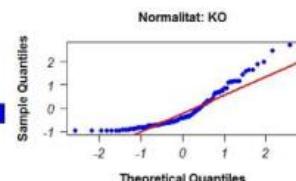
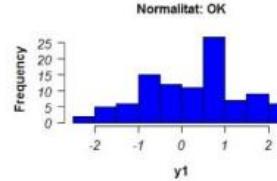
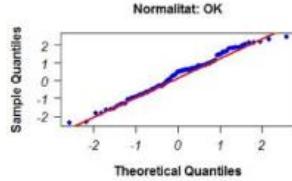
La **variancia residual** se interpreta:

- Experimentos: variabilidad de la variable  $Y$ .
- Previsión: error de predicción de la variable  $Y$ , conociendo el valor de  $X$ . (son las fluctuaciones de nuestra previsión, es decir cuanto por encima y por debajo de nuestro valor nos podríamos equivocar).

La **constante** se puede interpretar como el valor que toma la respuesta en ausencia de la variable predictoría.

Si el error estándar es demasiado grande → Para **disminuir  $S_{b1}$**  hay que aumentar  $n$  (número de observaciones).

- **Normalidad:** en el caso de un gráfico que se mantengan los datos sobre la recta qqnorm y en el caso de un histograma que forme una campana.



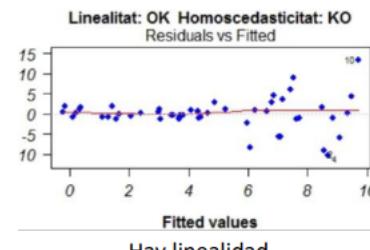
**Inferencia** se puede hacer la prueba de hipótesis para  $\beta_0$ ,  $\beta_1$  y  $\sigma^2$  (seguir el procedimiento habitual, tomando la distribución t-Student como estadístico).

La inferencia con intervalo de confianza:  $IC(\mu_1 - \mu_2, 95\%) = (\bar{y}_1 - \bar{y}_2) \pm t_{n_1+n_2-2, 1-\alpha/2} \cdot \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}$ , el resultado representaría que por ejemplo la muestra 2 tarde de media entre  $X$  e  $Y$  segundos menos con un intervalo de confianza del 95%.

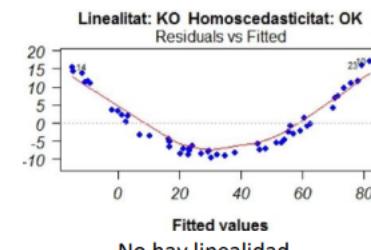
Para **validar el modelo lineal** hay que mirar las premisas:

- En la parte determinista:

- **Linealidad:** en el rango que se nos da que se mantenga igual sobre el eje de la  $y$



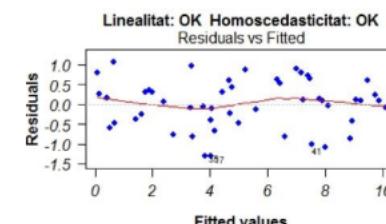
Hay linealidad



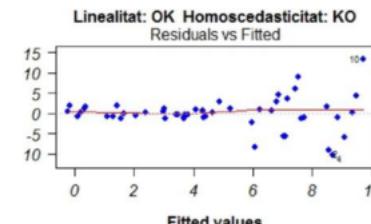
No hay linealidad

- En la parte aleatoria:

- **Homoscedasticidad:** misma  $\sigma^2$  para cualquier caso/dato, cuando miramos el gráfico vemos que todos los datos tienen el mismo error.

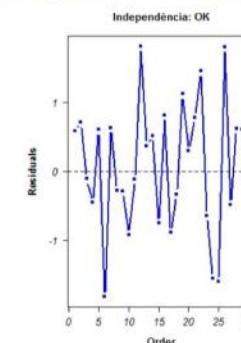


Hay homoscedasticidad

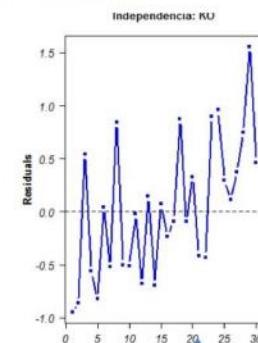


Hay heteroscedasticidad (no hay homoscedasticidad)

- **Independencia:** Detecta si existe o no dependencia entre los datos.

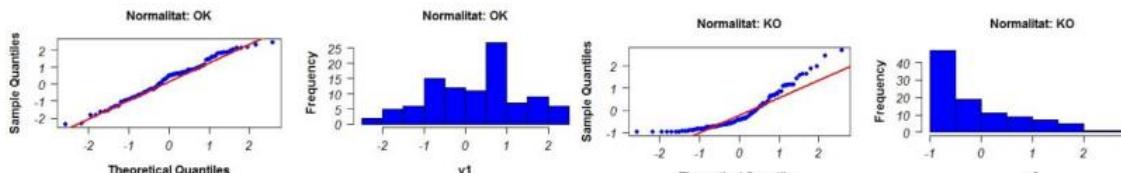


Hay independencia, porque no observamos ningún patrón en el gráfico



No hay independencia, porque observamos un patrón en el gráfico

- **Normalidad**: en el caso de un gráfico que se mantengan los datos sobre la recta qqnorm y en el caso de un histograma que forme una campana.



```
datos <- read.table("clipboard", header=TRUE) # Leemos los valores copiados
# La función lm se usa para ajustar un modelo lineal, sea un modelo de regresión lineal,
# de análisis de varianza o de análisis de covarianza
mod.lm <- lm(var_respuesta ~ condicion, datos)

# Obtenemos una salida más detallada e informativa
summary(mod.lm)

# Permite modificar distintos parámetros de la ventana gráfica
par(cex.lab = 1.2, cex.axis = 1.2, las = 1, font.lab = 2, font.axis = 3)

# Dibuja el gráfico
plot(var_respuesta ~ condicion, datos, pch = 19, col = 4, cex = 1.2)

# Permite sobreponer una recta de regresión a un gráfico de dispersión
abline(mod.lm, col = 2, lwd = 3)
```

#### summary(mod.lm) - Ejemplo e interpretación de los datos:

Call:  
lm(formula = Preu ~ Capacitat, data = datos)

#### Residuals:

Min	1Q	Median	3Q	Max
-102.32	-20.89	12.48	36.99	89.21

#### Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	$b_0 = 386.3889$	$S_{b_0} = 69.6313$	$t_{b_0} = 5.549$	$p - \text{valor}_{b_0} = 4.40e-05 ***$
Capacitat	$b_1 = 2.4133$	$S_{b_1} = 0.4097$	$t_{b_1} = 5.891$	$p - \text{valor}_{b_1} = 2.28e-05 ***$

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error:  $S = 57.94$  on *grados de libertad* = 16 degrees of freedom  
Multiple R-squared:  $R^2 = 0.6844$ , Adjusted R-squared: 0.6647  
F-statistic: 34.7 on 1 and 16 DF, p-value: 2.278e-05

$\bar{Y}$

- **Residuals**: Proporciona estadísticas sobre los residuos del modelo, como el mínimo, el 1er cuartil, la mediana, el 3er cuartil y el máximo.

- **Coefficients**: Una tabla que muestra los coeficientes estimados, sus errores estándar, estadísticas t y valores p asociados.

R				
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.2375	0.1821	1.30	0.203
X1	1.3456	0.2456	5.48	1.2e-06 ***
X2	-0.0187	0.1254	-0.15	0.879

- **Estimate**: El valor estimado del coeficiente.
- **Std. Error**: El error estándar del coeficiente.
- **t value**: El valor t, que indica cuántas desviaciones estándar está el coeficiente de su valor esperado bajo la hipótesis nula.
- **Pr(>|t|)**: El valor p asociado con la prueba t.

## Ejemplo para validación lineal  
par(mfrow=c(2,2))  
plot(lm(var\_respuesta ~ condicion),c(2,1)) # QQ-Norm y Standard Residuals vs. Fitted  
hist(rstandard(lm(var\_respuesta ~ condicion))) # Histograma de residuos estandarizados  
plot (1:10,rstandard(lm(var\_respuesta ~ condicion)),type="l") # Orden de los residuos

$p - \text{valor}_{b_x} < 0.05 \rightarrow$  Indica que es un valor significante, que es importante para nuestra muestra, cuanto más cercano a 0 mejor.

$$\begin{aligned} t_{b_0} &= b_0/S_{b_0} \\ t_{b_1} &= b_1/S_{b_1} \end{aligned}$$

- **Residual standard error**: La desviación estándar de los residuos, que es una medida de cuán dispersos están los residuos.

R				
Residual standard error: 0.878 on 97 degrees of freedom				

- **F-statistic**: La estadística F para probar la significancia conjunta de los coeficientes del modelo.

R				
F-statistic: 45.62 on 2 and 97 DF, p-value: 2.2e-16				

- **R-squared**: La proporción de la varianza total explicada por el modelo.
- **Adjusted R-squared**: Una versión ajustada del R-cuadrado que tiene en cuenta el número de predictores.