Inferència estadística

Bloc 4 – Probabilitat i Estadística

Novembre 2015

Índex

Inferència estadística. Guió

Guió de la part d'Estadística de PE:

- **B4**: Tècnica general de la inferència [estadística]
  - estimar un paràmetre (***Intervals de Confiança***)
  - refutar un paràmetre (***Proves d'Hipòtesis***)
- **B5**: Aplicació (I): Avaluació de millores
  - ***Disseny d'experiments:*** comparació de dues poblacions.
- **B6**: Aplicació (II): Predicció
  - ***Previsió*** d'una var. resposta, en funció d'una var. explicativa.

Inferència estadística. Introducció

- **La ciència vol ser refutable:** El criteri de refutabilitat sosté que per ser col·locats en el rang de científics, els enunciats han de poder entrar en conflicte amb observacions possibles. [Ex: "Els marcians existeixen" : no hi ha mitjans per contradir-ho → no és científic]

- **La tècnica vol ser documentable:** S'ha de aportar evidència basada en dades [Ex: "El meu programa funciona": si no aportes proves / dades → no és tècnic]

- La metodologia estadística permet la inducció: **inferir les característiques de la població a partir de les observacions d'una mostra** [Ex: Per conèixer la velocitat mitjana real de connexió amb un determinat proveidor recullo una mostra de velocitats]

Inferència estadística. Introducció

- La Inferència Estadística defineix i **quantifica els riscos** d'aquest procés [Ex: No es pot conèixer la mitjana de la vel. de connexió en tota la població a no ser que es tingui dades de tota la població. Però la estadística, em permet quantificar l'error en l'estimació en una mostra concreta]
- Mètode científic i tècnic (estadístic):
  - per **deducció** → disseny de la recollida de dades (Població → Mostra)
  - per **inducció** → inferir (estimar) resultats (Mostra → Població)

Exemples:

- Vull dir"El meu programa funciona bé"
  - → 1) recollida de dades (*proves o evidència*)
  - → 2) anàlisis: estimar una mesura (p.e.: mitjana del rendiment)
- "El meu programa millora els resultats de ..."
  - → 1) recollida de dades (*proves o evidència*)
  - → 2) anàlisis: poder refutar la igualtat de rendiments

Inferència estadísta. Mostra Aleatòria Simple (MAS)

Sigui la v.a. $X: \Omega \rightarrow R$

$\omega_i \rightarrow X(\omega_i) = x_i$

Direm que M.A.S. de grandària **n** de la v.a. X

a la funció vectorial $M = (X_1, X_2, ..., X_n)$

$M: \Omega^n \rightarrow R^n$

$\omega = (\omega_1, \omega_2, ... \omega_n) \rightarrow M(\omega_i) = (X_1, X_2, ..., X_n)$

Direm que és una MAS si i només si es compleixen les dues condicions següents:

(1) **Tots els elements** de la població tenen la **mateixa probabilitat** de pertànyer a la mostra.

(2) **Qualsevol combinació** de *n* elements té la **mateixa probabilitat** de pertànyer a la mostra.

La informació aportada per les diferents unitats ha de ser **independent** entre sí:

- les X_i han de ser v.a. independents i idènticament distribuïdes: v.a.i.i.d.

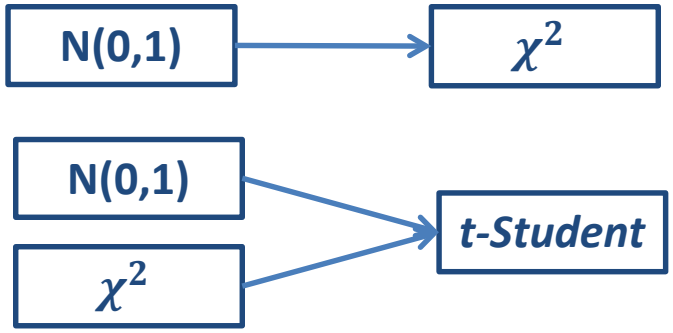
Inferència estadística. Conceptes bàsics

- Els indicadors de la població que estem interessats en conèixer reben el nom de **paràmetres**. [Ex: La μ (esperança) de les alçades dels estudiants de la FIB]
- Un **estadístic** és qualsevol indicador que s'obtingui com a funció de les dades d'una mostra. [Ex: La suma de les alçades dels estudiants recollits en una mostra]
- Quan un estadístic (d'una mostra) s'utilitza per conèixer el valor d'un paràmetre (de la població), rep el nom d'**estimador**. [Ex: La mitjana de les alçades en una mostra d'alumnes de la FIB és una estimador de la μ (esperança) de les alçades dels estudiants de la FIB]

Atenció: La paraula ***mitjana*** pot voler dir ***paràmetre esperança*** quan parlem del centre de gravetat d'una distribució poblacional, o ***estadístic mitjana*** quan ens referim al promig d'una sèrie de valors obtinguts d'una mostra.

Models derivats de la Normal: χ^2 i *t-Student*

- Hi ha un parell de distribucions noves que ens serviran per abordar el contingut d'aquest tema: χ^2 i **t-Student**
- Aquestes distribucions provenen de fer operacions amb v.a. provinents d'altres distribucions, entre elles la Normal estàndard.



- A diferència de les distribucions vistes en el tema anterior NO modelen fenòmens de la vida real, sinó el comportament dels estadístics entre les possibles mostres.

Distribució χ^2 (chi-quadrat)

- **Definició**: Siguin $X_i \sim N(0,1)$. Llavors:

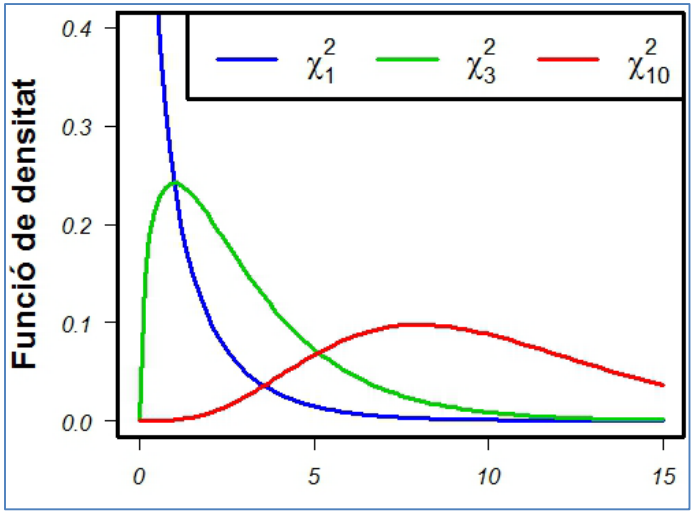
  $$X_1^2 + X_2^2 + \cdots + X_n^2 \sim \chi_n^2$$

  [Concretament, per n = 1 → $X_1^2 \sim \chi_1^2$]

- **Notació**: $X \sim \chi_n^2$
- **Paràmetres**: *n* (graus de llibertat)
- **Funció de probabilitat i distribució**:

$$f(x) = \frac{x^{k/2-1} \cdot e^{-x/2}}{2^{k/2} \cdot \Gamma(k/2)} \quad per\ x > 0$$

$$F(x) = \frac{\gamma(k/2, x/2)}{\Gamma(k/2)} \quad per\ x > 0$$



Γ : funció Gamma
γ : funció Gamma incomplerta
n: graus de llibertat

R: dchisq, pchisq, qchisq

Script per veure que la suma de Normals estàndard al quadrat és una χ^2

```
M = 500 # Mostres de normals
n = 7   # Graus de llibertat
sample = array(rnorm(M*n, 0, 1), dim=c(M,n)) # n mostres de N(0,1)
sample2 = sample*sample                      # n mostres de (N(0,1))^2
sum = apply(sample2, 1, sum)                 # suma de les les mostres al^2
hist(sum, breaks="Scott", freq=FALSE)
curve(dchisq(x, n),add=TRUE,col=2,lwd=2)
quantile(sum, c(0.25, 0.50, 0.75))      # q1, median i q3 de la suma de Normals
qchisq(c(0.25, 0.50, 0.75),n)           # q1, median i q3 de la chi-quadrat
```

Distribució t-Student

- **Definició**: Siguin dues v.a independents, $Z \sim N(0,1)$ i $Y_n \sim \chi^2_n$. Llavors:

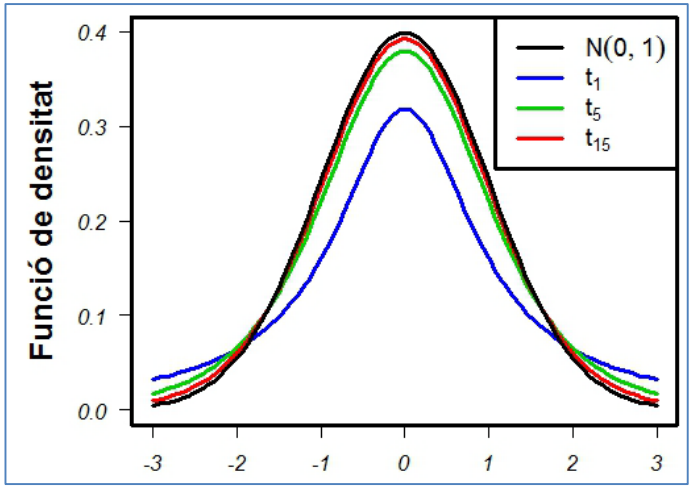
$$\frac{Z}{\sqrt{Y_n/n}} \sim t_n$$

[Quan n→∞ (n>30), llavors $t_n \rightarrow N(0,1)$]

- **Notació**: $X \sim t_n$
- **Paràmetres**: *n* (graus de llibertat)
- **Funció de probabilitat i distribució**:

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\cdot\Gamma(n/2)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad per\ x > 0$$

$$F(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n}\cdot B(1/2,n/2)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad per\ x > 0$$



Γ : funció Gamma
B: funció Beta
n: graus de llibertat

R: dt, pt, qt

Script per veure que a partir de una Z i una Y_n s'obté una t

```
M = 500; n = 7 # Mostres de normals i Graus de llibertat
samplez = rnorm(M, 0, 1)
samplechi2 = rchisq(M,n)
samplechi2n = sqrt(samplechi2/n)
t = samplez / samplechi2n
hist(t, breaks="Scott", freq=FALSE)
curve(dt(x, n),add=TRUE,col=2,lwd=2)
quantile(t, c(0.25, 0.50, 0.75))        # q1, median i q3 de Z/sqrt(Yn/n)
qt(c(0.25, 0.50, 0.75),n)               # q1, median i q3 de la chi-quadrat
```

Estimació puntual

- Un estimador $\widehat{\theta}$ del paràmetre desconegut θ, a partir de la mostra $M(\omega_i)$ $(X_1,X_2,\ldots,X_n)$ és una funció de les v.a:

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

- **Estimació puntual:** valor que l'estimador $\hat{\theta}$ pren en una mostra concreta. [Ex: $\bar{x} = \frac{\sum x_i}{n}$ és la mitjana mostral i és una estimació puntual de μ]

Nota: Distingiu entre el valor ($\bar{x}$) i la variable ($\bar{X}$) mitjana mostral (no s'extreu de cap mostra)

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

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

Nota: Generalment, la **σ** serà desconeguda i l'error tipus l'haurem d'aproximar emprant l'estimador pertinent ($\hat{\sigma}$) amb les dades de la mostra: $\widehat{se} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n}} = \sqrt{\sum \frac{(x_i-\bar{x})^2}{n-1}} \cdot \frac{1}{\sqrt{n}}$

Estimació puntual. Exemples

Nota: per als paràmetres s'utilitzen lletres de l'alfabet grec

Paràmetre (θ) (**POBLACIÓ**)	Estimador ($\hat{\theta}$) (**MOSTRA**)
μ (esperança, mitjana poblacional)	$\bar{x}$ (mitjana mostral)
σ^2 (variança poblacional) σ (desviació tipus poblacional)	s^2 (variància mostral) s (desviació tipus mostral)
π (probabilitat)	p (proporció)

Exemple:

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 [$\Sigma x = 4981$; $\Sigma x^2 = 2860855$]

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

Una estimació puntual del nombre esperat (μ) de terminals diaris connectats és:

`mean(nterm)` → $\bar{x} = 553.44$ o $\bar{x} = (\sum x_i)/n = 553.44$

Una estimació puntual de la desviació tipus (σ) del nombre de terminals connectats és:

`sd(nterm)` → $s = 114.0988$ o $s = \sqrt{(\sum(x_i - \bar{x})^2)/(n-1)} = 114.0988$

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

`sd(nterm)/sqrt(length(nterm))` → $se = 38.03 = \sqrt{(\sum(x_i - \bar{x})^2)/(n-1)} \cdot 1/\sqrt{n} = 38.03$

Estimació puntual. Propietats dels estimadors

- Inevitablement, les estimacions puntuals **fallen** o, millor dit, com depenen de la mostra que "ens ha tocat", **fluctuen** (encara que usualment tan sols observem un valor)
- Les 2 obsessions de l'Estadística són:
  - **quantificar** els errors d'estimació
  - **minimitzar** aquests errors
- L'error tipus o típic informa de l'**error esperat** a l'equiparar el valor de l'estimador obtingut en l'estudi amb el valor del paràmetre poblacional.
- Com l'estimador és "qualsevol" estadístic que s'utilitzi amb fins inferencials, hem de definir les propietats que permeten definir els "millors".

Nota: *l'error exacte en una mostra concreta roman desconegut, podent ser inferior o superior que l'error típic o esperat.*

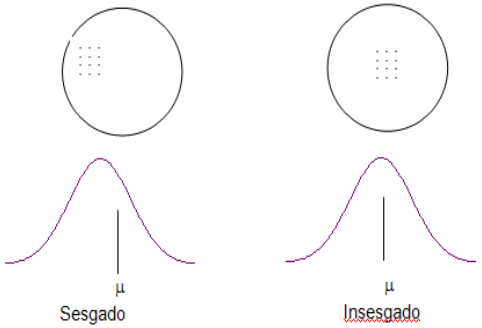
Estimació puntual. Propietats dels estimadors

Propietats desitjables

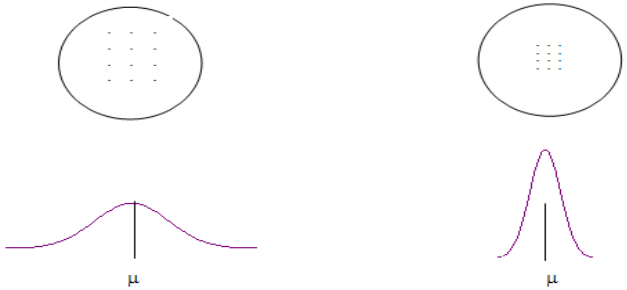
- **No tenir biaix** (= *sesgo*, *bias*)
  - Biaix és la diferència entre el centre de la distribució del estimador $[E(\hat{\theta})]$ i el valor del paràmetre a estimar $[\theta]$

$$Biaix = E(\hat{\theta}) - \theta$$

  - Un estimador $\hat{\theta}$ del paràmetre θ és NO esbiaixat si Biaix = 0

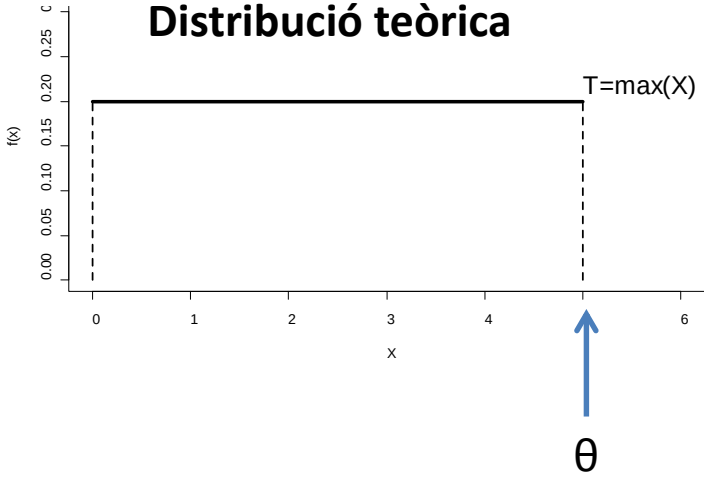


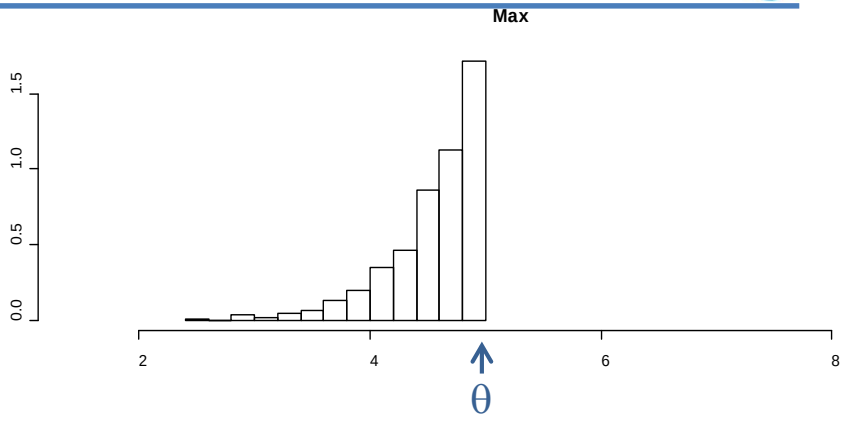
- **Ser Eficient**
  - Entre dos estimadors NO esbiaixats, es diu que és més eficient el que té una variància menor.

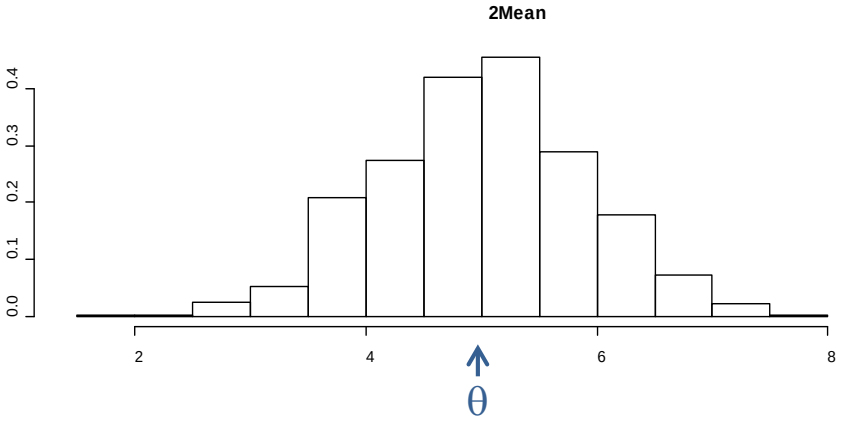


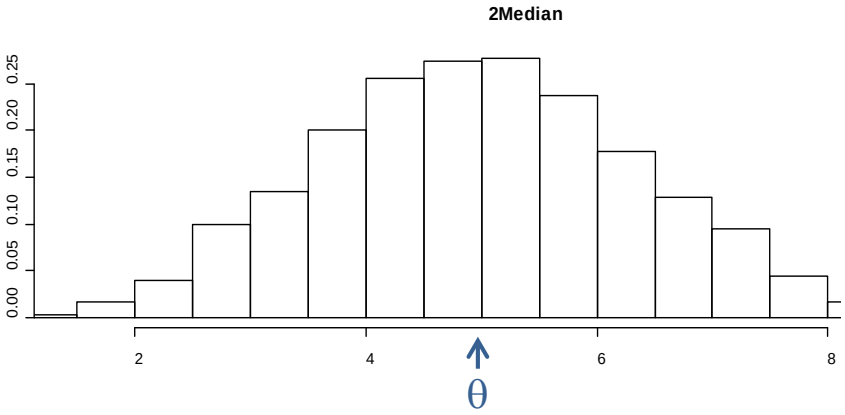
Estimació puntual. Exemple

Estimar el límit superior θ d'una llei uniforme, amb mínim zero (U[0 , θ])









Tres estimadors possibles:

- el màxim de la mostra (esbiaixat)
- el doble de la mitjana mostral (no esbiaixat)
- el doble de la mediana mostral (no esbiaixat)

Teorema del Límit Central (repàs)

- El T.L.C. estableix que, si s'agafen mostres de grandària *n* d'una població de mitjana μ i desviació típica σ, a mesura que creix *n*, la distribució de la mitjana mostral $\bar{X}$ s'aproxima a la d'una normal de mitjana μ i desviació típica $\sigma/\sqrt{n}$:

$$\bar{X}_n \xrightarrow[n\ gran]{} N(\mu, \frac{\sigma}{\sqrt{n}})$$

- ***¿Quina grandària ha de tenir n per a que es compleixi el TLC?***
  - Si X ~ N → $\bar{X}_n$~ N **∀n** [Les combinacions lineals de normals i.i.d. són normals]
  - Si X és quantitativa → $\bar{X}_n$ ~ N si **n > 30** [Com més s'assembli X a la normal abans passarà aquesta similitud]
  - Si X té una distribució **discreta** i/o **asimètrica**, la convergència requereix una grandaria mostral (*n*) més gran

Quantils en la Z~N(0,1) (repàs)

Definició: El quantil α és el valor fins al qual s'acumula una probabilitat α

Notació: Aquell a tal que $F_Z(a) = P(Z<a) = \alpha$ l'indicarem per $\mathbf{z_\alpha}$

Exemples:

$\mathbf{-1.960 = z_{0.025}}$ $\mathbf{1.960 = z_{0.975}}$ $\mathbf{1.645 = z_{0.95}}$

En la Normal, al ser simètrica, es compleix que:

$$z_\alpha = - z_{1-\alpha}$$

$$z_{\alpha/2} = - z_{1-\alpha/2}$$

Interval per a la mitjana mostral. Exemple

- Situació:

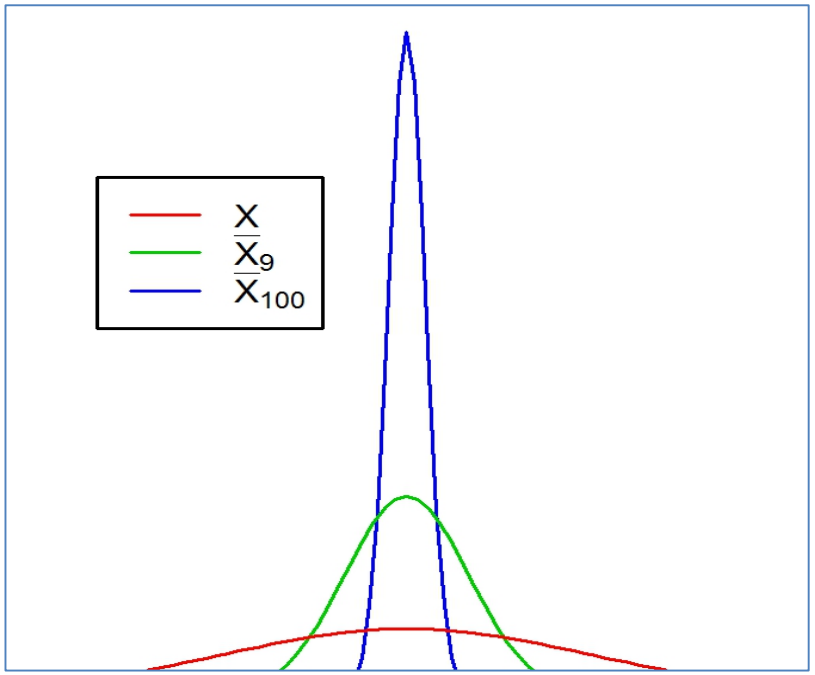
  X: “temps d’execució d’un algoritme” ~ N(μ=100ms, σ=10ms).

- Plantegem les distribucions de les v.a.: X , $\overline{X}_9$ i $\overline{X}_{100}$
- Calcularem, per *n*=1, *n*=9 i per *n*=100, els intervals amb probabilitats *grans* (95%) d’assegurar que les possibles $\bar{x}$ hi pertanyeran [deixem fora només una petita proporció α: 0.05 o 5%]
- Distribucions:

$$X \sim N(\mu = 100\ ms, \sigma = 10\ ms)$$

$$\bar{X}_9 \sim N\left(\mu = 100\ ms, \sigma = \frac{10}{\sqrt{9}} = 3.33\ ms\right)$$

$$\bar{X}_{100} \sim N\left(\mu = 100\ ms, \sigma = \frac{10}{\sqrt{100}} = 1\ ms\right)$$



Interval per a la mitjana mostral. Exemple

Els límits v, w dels intervals els podem calcular utilitzant les taules de la N(0,1):

$$z_{0.975} = 1.96 \qquad z_{0.025} = -1.96$$

- Rang que conté el 95% de les infinites execucions de l'algoritme (*X*)

$$v,w = \mu \pm z_{0.975} \cdot \sigma = 100 \pm 1.96 \cdot 10 = 100 \pm 19.60 = [80.40, 119.60]$$

- Rang que conté el 95% de les mitjanes de les infinites mostres de *n* = 9 execucions ($\bar{X}_9$)

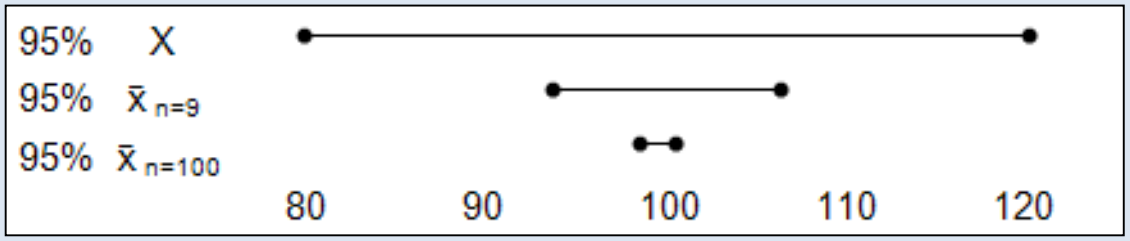
$$v,w = \mu \pm z_{0.975}\, \sigma/\sqrt{n} = 100 \pm 1.96 \cdot 10/3 = 100 \pm 6.53 = [93.47, 106.53]$$

- Rang que conté el 95% de les mitjanes de les infinites mostres de *n* = 100 execucions ($\bar{X}_{100}$)

$$v,w = \mu \pm z_{0.975}\, \sigma/\sqrt{n} = 100 \pm 1.96 \cdot 10/10 = 100 \pm 1.96 = [98.04, 101.96]$$

Representació esquemàtica:

Com l'amplada de l'interval depèn inversament de l'arrel de n, passar de n=1 a n=9, fa l'interval 3 vegades més estret



Estimació per interval de μ

- Hem vist que sabem calcular un "interval" que contingui $\bar{x}$ a partir de μ. Però el problema real és **calcular interval per μ, coneixent $\bar{x}$** (és a dir, passar d'un interval per a la mitjana mostral $\bar{x}$ a un per a la mitjana poblacional μ)
- A partir d'una probabilitat 1- α entre dos valors a i b (simètrics): *(amb σ 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-α

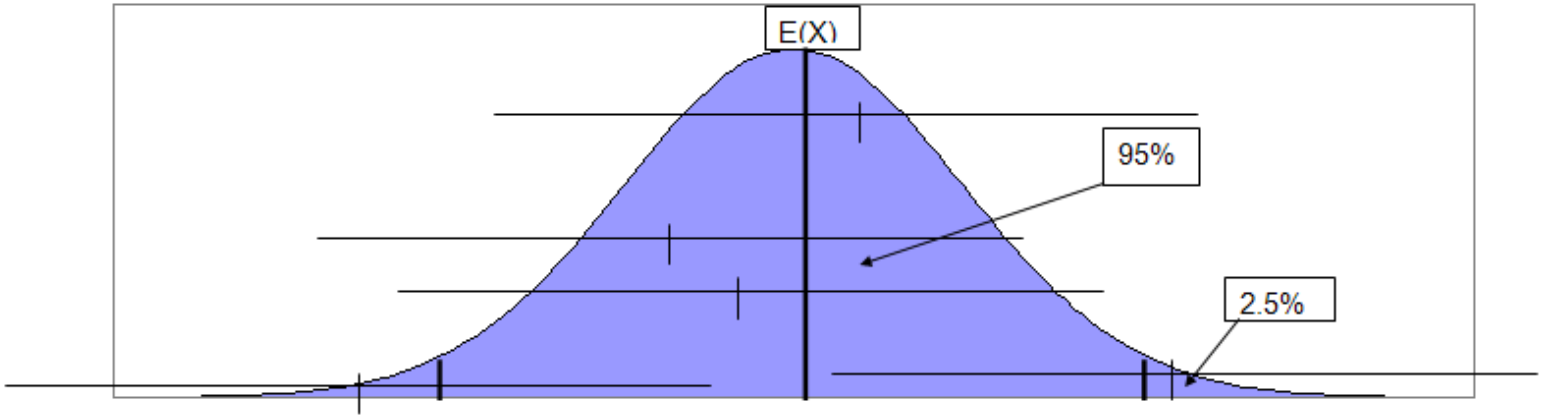
$$P\left(\mu + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

- I reordenant obtenim **l'interval de confiança 1-α del paràmetre μ**

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

Estimació per interval de μ

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



- *Aquest procediment encerta el 100·(1-α)% de les vegades!*
- Denotem **IC(μ, 1-α)** a l'**INTERVAL DE CONFIANÇA** 1- α de μ, i l'expressem:

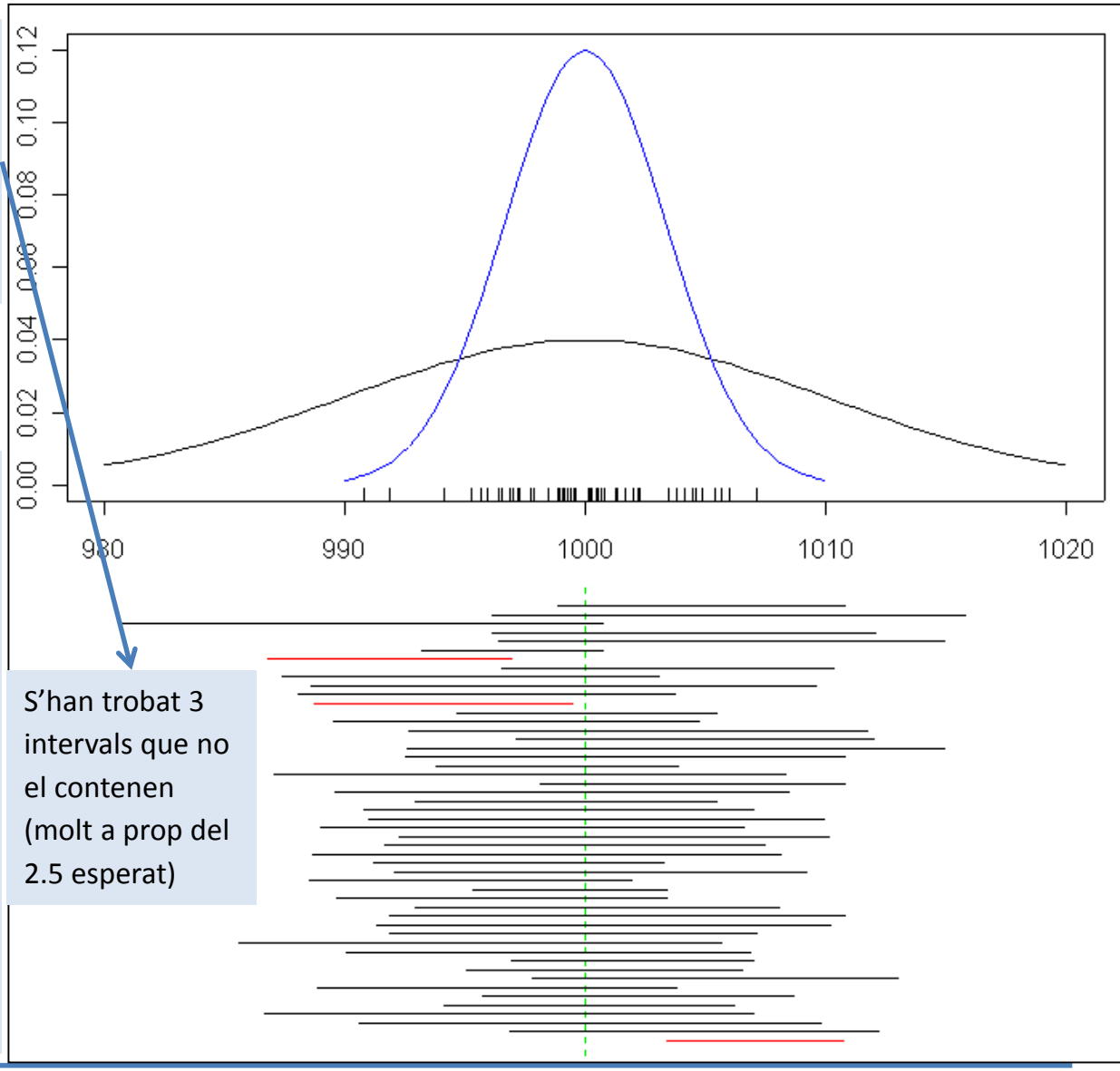
$$IC(\mu, 1-\alpha) = \bar{x} \mp z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Com que la Normal és simètrica, llavors $z_{\alpha/2} = -z_{1-\alpha/2}$

Atenció: nosaltres només observarem una mostra, i no sabrem si l'IC trobat conté o no μ, però sí sabem que aquest procediment a la llarga dóna un 100·(1-α)% d'encerts.

Estimació per interval de μ. Simulació

S'han simulat 50 mostres amb n=9 provinents d'una N(μ=1000, σ). Calculant el IC95%, esperem que aproximadament el 5% (2.5) d'ells no continguin el valor real de μ



S'han trobat 3 intervals que no el contenen (molt a prop del 2.5 esperat)

Nota tècnica: Amb un IC determinat (p.ex., [985, 1004]), s'ha de dir "**tenim un alt grau de confiança (i.e., 95%) de que el paràmetre es troba entre aquest dos valors**", però no és correcte parlar de probabilitat 95% que el paràmetre estigui entre els dos valors trobats, perquè el paràmetre no es considera un element aleatori. Serà desconegut, però no és incert!

Interval de confiança per μ (σ coneguda). Exemple

- Així doncs, l'interval de confiança 1−α de μ (amb σ coneguda) és:

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

- Recordeu que ens basem en el TCL i perquè es complís calia que la variable X inicial fos Normal o que n fos més gran de 30. Per tant, els requisits per realitzar aquest càlcul són: **n > 30 o X~N**

EXEMPLE:

Una embotelladora d'ampolles de litre té una dispersió de σ = 10cc. En una mostra a l'atzar de n = 100 ampolles d'aquesta màquina, la mitjana observada ha sigut $\bar{x}$ = 995cc. Calculeu un interval de confiança del 95% de μ.

$$IC(\mu, 0.95) = \bar{x} \mp z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 995 \mp 1.96 \cdot \frac{10}{\sqrt{100}} = 995 \mp 1.96 = [993.04\,, 996.96]$$

Resultat: amb una confiança del 95%, μ es troba entre 993.04 i 996.96

Interval de confiança per μ (σ coneguda). Exercici

1. La glicèmia en mmol/L té una desviació típica de σ = 1 en una mostra de n= 9 pacients, la mitjana $\bar{x}$ val 5. Calculeu el IC(μ,0.95).

   Amb una "força" del 95%, creiem que l'autèntic valor poblacional està entre aquests límits

2. Sense canviar la confiança, com podríem reduir l'interval a la meitat?

3. Calculeu l' IC amb una confiança del 99%

ATENCIÓ: quan *n* augmenta la precisió dels IC augmenta (interval més estret). Si augmenta la confiança (disminuint el risc *α* d'error), la precisió dels IC disminueix (interval més ample)

ATENCIÓ: En aquest cas, per estimar μ necessitem conèixer σ → *situació molt particular i infreqüent*

Interval de confiança. Mecànica

Passos	Esquema de Resolució
1	Definir l'estadístic a ser utilitzat
	Especificar la seva distribució
	Indicar les premisses necessàries per dir que segueix la distribució
	Delimitar el nivell de confiança (usualment $1-\alpha=95\%$)
2	Calcular l'interval
3	Interpretar el resultat

Interval de confiança per μ (σ desconeguda)

- Si **desconeixem σ**, la podem substituir per S, i llavors l'estadístic $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ passa a ser $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ que és el quocient de 2 v.a. i ja no es pot suposar que segueix una distribució N(0,1).
- Tal com diu el pas 2 de la mecànica de construcció de l'IC, cal conèixer la llei de probabilitats que segueix l'estadístic. En aquest cas, es tracta d'una nova distribució anomenada ***t-Student***.
- Per aquest estadístic, la distribució de probabilitat concreta és t_{n-1} (n-1 graus de llibertat). Els percentils es poden trobar a taules específiques o amb R.
- Així doncs, l'interval de confiança 1-α de μ (amb σ desconeguda) és:

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

ATENCIÓ: la situació de no conèixer la σ de la població és més freqüent

Interval de confiança per μ (σ desconeguda)

- Podem utilitzar-la per conèixer el quocient informació/soroll utilitzant S en lloc de σ

$$\hat{t} = \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}} \sim t_{n-1}$$

- Demostració:

$$\hat{t} = \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}} = \frac{(\bar{X} - \mu)/\sqrt{\sigma^2/n}}{\sqrt{\frac{s^2/n}{\sigma^2/n}}} = \frac{Z}{\sqrt{\frac{s^2}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2/\sigma^2}{n-1}}}$$

$$= \frac{Z}{\sqrt{\frac{Y_n}{n-1}}} \sim t_{n-1}$$

- Coneixem que el quocient informació/soroll segueix una 't' de Student .
- Això ens ajudarà a conèixer el IC95% per la μ amb σ desconeguda

Nota: Observeu que es requereix la Normalitat de les X's

Interval de confiança per μ (σ desconeguda). Ex.

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.

R: `nterm <- c(587,470,676,451,436,672,584,697,408)`

Recordem que havíem calculat les estimacions puntuals: Una estimació del IC al 95% de confiança de la mitjana poblacional, assumint que la desviació poblacional (σ) val 100:

1-α	σ	IC(μ,1-α)	Resolució amb R

95%	Coneguda (σ=100)	[488.11; 618.78]	`n <- 9 ; sigma <- 100`
			`mean(nterm)- qnorm(0.975)*sigma/sqrt(n)`
			`mean(nterm)+ qnorm(0.975)*sigma/sqrt(n)`

99%	Coneguda (σ=100)	[467.58 ; 639.31]	`n <- 9 ; sigma <- 100`
			`mean(nterm)- qnorm(0.995)*sigma/sqrt(n)`
			`mean(nterm)+ qnorm(0.995)*sigma/sqrt(n)`

95%	Desconeguda	[465.74; 641.15]	`n <- 9`
			`mean(nterm)- qt(0.975,8)*sd(nterm)/sqrt(n)`
			`mean(nterm)+ qt(0.975,8)*sd(nterm)/sqrt(n)`

99%	Desconeguda	[425.83 ; 681.06]	`n <- 9`
			`mean(nterm)- qt(0.995,8)*sd(nterm)/sqrt(n)`
			`mean(nterm)+ qt(0.995,8)*sd(nterm)/sqrt(n)`

Interval de confiança per μ. Premisses

Per garantir el nivell de confiança de l'IC, s'ha de complir certes premisses

- Si sigma és coneguda, exigirem una de les condicions:
  - **X~N** → la combinació lineal de Normals és Normal ($\bar{X}$~N)
  - **Tenir una mostra gran (n≥30)** → Pel TCL , $\bar{X}$~N
- Si sigma no és coneguda, exigirem una de les condicions:
  - **X~N** → $(\bar{x}-\mu)/\sqrt{s^2/n} \sim t_{n-1}$
  - **Tenir una mostra gran (n≥100)** → Pel TCL , $\bar{X}$~N

Per descomptat, la premissa que sempre s'ha de complir és que l'origen de la mostra ha de ser aleatori (v.a.i.i.d)

Dist. de referència si...	σ coneguda	σ desconeguda
X Normal	Normal *sempre*	t de *Student sempre*
X no Normal	Normal *si n gran* *(n ≥ 30)*	Normal *si n + gran* *(n ≥ 100)*

Amb grans mostres la variació de s serà limitada (s estima molt bé σ), i podem considerar que

$(\bar{x}-\mu)/\sqrt{s^2/n} \approx (\bar{x}-\mu)/\sqrt{\sigma^2/n} \sim \mathrm{N}(0,1)$ [Aplicable al cas X~N, σ desc. i n gran]

Interval de confiança per π en una Binomial (n,π)

- Sigui X ~ B(n,π) → E(X) = π·*n*
  V(X) = π·(1-π)·*n*
- Sigui P = X/n → E(P) = E(X/*n*) = E(X)/*n* = π·*n* / *n* = π
  V(P) = V(X/*n*) = V(X)/*n*^{2} = π·(1-π)·*n*/*n*^{2} = π·(1-π)/*n*
- Per construir l'IC es pot recorre a la convergència de la Binomial a la Normal [amb la premissa de *n* gran i π no extrema → **π·*n* ≥ 5 y (1-π)·*n* ≥ 5**]:

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

- Així, l' interval de confiança s'assemblaria al de μ:

$$IC(\pi, 1-\alpha) = P \mp z_{1-\frac{\alpha}{2}}\sigma_P = P \mp z_{1-\frac{\alpha}{2}}\sqrt{\frac{\pi(1-\pi)}{n}}$$

La **paradoxa** de que necessitem conèixer π per estimar π, es pot solucionar:

a) Substituint π per P→ $IC(\pi, 1-\alpha) = P \mp z_{1-\alpha/2} \cdot \sqrt{\big(P(1-P)\big)/n}$

b) Aplicant el màxim de π·(1-π) → $IC(\pi, 1-\alpha) = P \mp z_{1-\alpha/2} \cdot \sqrt{\big(0.5(1-0.5)\big)/n}$

IC per π en una Binomial(n,π). Exemple

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 π:

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

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

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

Interval de confiança per σ^2 . Distribució χ^2

- Gràcies a aquesta distribució, coneixem la distribució de S^2 (estimador de σ^2):

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n} \rightarrow (n-1)\cdot\frac{s^2}{\sigma^2} \sim \chi^2_{n-1}$$

- Demostració: si $X_i \rightarrow N$, llavors

Les x_i han de provenir d'una Normal

$$n\cdot\frac{\hat{\sigma}^2}{\sigma^2} = n\cdot\frac{\left(\sum_{i=1}^{n}(x_i-\mu)^2\right)/n}{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 \sim \chi^2_n$$

- També, si $X_i \rightarrow N$

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

- *Per tant, **coneixem la distribució de S^2!!*** i podrem relacionar la distribució χ^2_{n-1} amb s^2 per definir IC de s^2, tal com fem amb les distribucions Z i t_{n-1} amb $\bar{x}$ per definir IC de μ.

Interval de confiança per σ^2

- Hem vist que: $(n-1)\cdot\frac{S^2}{\sigma^2}\sim\chi^2_{n-1}$ sempre que **X_i ~ N (premissa)**
- Per tant:

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

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

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

Atenció: és un IC per σ^2, no per σ!!

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

Nota: no és un interval simètric, ja que la distribució no ho és.
Això implica calcular dos valors en la distribució en lloc de fer $\pm$.

IC per σ^2 en una Normal (μ,σ). Exemple

En les 25 execucions d'un mateix programa s'ha observat una variabilitat $s^2=8^2$.

$$IC(\sigma^2, 0.95) = \left[\frac{s^2(n-1)}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{s^2(n-1)}{\chi^2_{n-1,\frac{\alpha}{2}}}\right] =$$

$$= \left[\frac{8^2(25-1)}{12.401}, \frac{8^2(25-1)}{39.364}\right] = [123.86, 39.02] \rightarrow \textbf{\textit{Oops! M'he equivocat}}$$

$$= \left[\frac{8^2(25-1)}{39.364}, \frac{8^2(25-1)}{12.401}\right] = [39.02, 123.86] \rightarrow \textbf{\textit{Ara sí!}}$$

Resultat:

$$IC(\sigma^2, 0.95) = [39.02, 123.86]$$

$$IC(\sigma, 0.95) = [6.25, 11.13]$$

Fent l'arrel quadrada, obtenim un interval per σ

Formulari per IC

TCL: $X_1,\ldots, X_n$ i.i.d. ($n\to\infty$), amb $E(X_i)=\mu$ i $V(X_i)=\sigma^2$,

llavors $\dfrac{\sum_{i=1}^n X_i}{n} = \bar{X}_n \approx N(\mu, \sigma^2/n)$ **i també** $\sum_{i=1}^n X_i \approx N(n\mu, \sigma^2 n)$ **)**

Estadístic mitjana mostral ($\bar{x}$): $\dfrac{(\bar{x}-\mu)}{\sqrt{\sigma^2/n}} \approx N(0,1)$ $\quad \dfrac{(\bar{x}-\mu)}{\sqrt{s^2/n}} \approx t_{n-1}$ on $\bar{x} = \sum_{i=1}^n x_i / n$

Estadístic variància mostral (s^2): $s^2 \dfrac{n-1}{\sigma^2} \approx \chi^2_{n-1}$ on $s^2 = \dfrac{\sum_{i=1}^n (x_i-\bar{x})^2}{n-1} = \dfrac{\sum_{i=1}^n x_i^2 - n(\bar{x})^2}{n-1} = \dfrac{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}{n-1}$

Paràmetre	Estadístic	Premisses	Distribució	Interval de Confiança 1-α (Risc α)
μ	$\hat{z} = \dfrac{(\bar{x}-\mu)}{\sqrt{\sigma^2/n}}$	[X → N ò n ≥ 30] i σ coneguda	$\hat{z} \to N(0,1)$	$\mu \in (\bar{x} \pm z_{1-\alpha/2}\sqrt{\dfrac{\sigma^2}{n}})$
μ	$\hat{t} = \dfrac{(\bar{x}-\mu)}{\sqrt{s^2/n}}$	X → N	$\hat{t} \to t_{n-1}$	$\mu \in (\bar{x} \pm t_{n-1,1-\alpha/2}\sqrt{\dfrac{s^2}{n}})$
μ	$\hat{z} = \dfrac{(\bar{x}-\mu)}{\sqrt{s^2/n}}$	n ≥ 100	$\hat{z} \to N(0,1)$	$\mu \in (\bar{x} \pm z_{1-\alpha/2}\sqrt{\dfrac{s^2}{n}})$
σ (normal)	$\hat{X}^2 = \dfrac{s^2(n-1)}{\sigma^2}$	X → N	$\hat{X}^2 \to \chi^2_{n-1}$	$\sigma^2 \in \left(\dfrac{S^2(n-1)}{\chi^2_{n-1,1-\alpha/2}}, \dfrac{S^2(n-1)}{\chi^2_{n-1,\alpha/2}}\right)$
π (Binomial)	$\hat{z} = \dfrac{(p-\pi)}{\sqrt{\pi(1-\pi)/n}}$	(1-π)n ≥ 5 πn ≥ 5	$\hat{z} \to N(0,1)$	$\pi \in (P \pm z_{1-\alpha/2}\sqrt{\dfrac{\hat{\pi}(1-\hat{\pi})}{n}})$ $\hat{\pi} = P \quad o \quad \hat{\pi} = 0'5$
λ (Poisson)	$\hat{z} = \dfrac{(L-\lambda)}{\sqrt{\lambda}}$	λ ≥ 5	$\hat{z} \to N(0,1)$	$\lambda \in (L \pm z_{1-\alpha/2}\sqrt{L})$

color gris Indica IC secundari

Proves d'hipòtesi. Exemple

Afirmo que encerto el 80% dels meus tirs lliures a basket, i un amic em posa a prova. Dels 20 tirs, solament faig 8. *"Fuà! Algú que encerta un 80% gairebé mai faria 8 de 20; així que no em crec la teva afirmació".*

El raonament de l'amic es basa en demanar-se què passaria si l'afirmació fos certa i es repetís la mostra de 20 tirs moltes vegades. Segurament poques vegades encertaria un nombre tan baix com 8. Un resultat com aquest és tan poc freqüent que aporta certa *evidència* per rebutjar la meva afirmació.

De fet, **aquesta *evidència* es pot quantificar**:

$$\textbf{Sigui M = nº encerts: } M \sim B(20, 0.8) \rightarrow P(M \leq 8) = 0.000102$$

Es a dir, si jo fos tan bo, solament 1 de cada 10000 vegades obtindria una puntuació tan nefasta (o pitjor).

Dues opcions: o he tingut mala sort o l'afirmació era falsa!

Source: The Basic Practice of Statistics. David S. Moore. 4th Ed.

Proves d'hipòtesi. Raonament

- Al igual que amb els IC, hem de pensar com si l'experiència es pogués repetir un gran nombre de vegades.
- Però ara no volem l'IC que contengui el paràmetre, ara partim d'una afirmació (una **hipòtesi** de partida, o **nul·la**), i volem estudiar si les dades proporcionen proves en contra seu.
- Una repetició intensa (una mostra infinita = la població) seria definitiva.
- Però, amb una mostra finita, quina informació aporten les dades?

Nota tècnica:

- Formalment, es distingeix entre les proves de Fisher (per aportar coneixement o evidencia o inferència) i els contrastos de Neyman-Pearson (per minimitzar els errors al prendre decisions). Els primers son rellevants per la Ciència (p.e., la Física) i els segons per la Tècnica (p.e. la Arquitectura). Però en aquest curs no distingirem i ho englobarem tot sota Proves de Hipòtesi (PH).
- Read more in: http://onlinestatbook.com/chapter9/significance.html

Proves d'hipòtesi. Raonament

La hipòtesi nul·la (H_0) es planteja formalment amb un paràmetre (o varis). El paràmetre en qüestió pren un valor que representem:

$$\mathbf{H_0: \pi = 0.80}$$

π representa la probabilitat poblacional d'encertar un tir lliure, i volem comprovar si aquest valor és coherent amb les observacions.

Al igual que amb els IC, la mostra es *concentra* en un estadístic, que segueix una distribució de probabilitat coneguda si s'assumeix certa la H_0.

Addicionalment a H_0 afegim la hipòtesi **alternativa** H_1, que pot ser totalment complementària a la nul·la (enfoc bilateral), o parcialment (unilateral):

$$\mathbf{H_1: \pi \neq 0.80} \qquad \mathbf{H_1: \pi < 0.80}$$

H_1 determina el(s) sentit(s) més oposat(s) a H_0: per exemple, el nombre de encerts a la canasta és l'estadístic, i si H_1 fos "≠" serien *sospitosos* tant els nombres d'encerts que van cap a 0 com els que van cap a 20. Si H_1 fos "<" serien *sospitosos* només els que van cap a 0 (que és el que hem pres, donat que el meu amic no confia molt en les meves habilitats).

Proves d'hipòtesi. P-valor

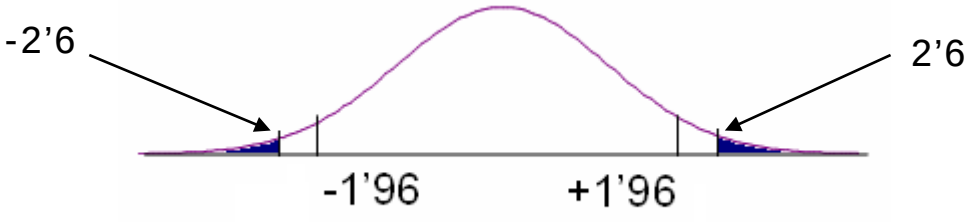
P, o **p-valor**, és la probabilitat de, sota H_0, obtenir resultats igual o més *extrems* que el observat.

Exemple:

estadístic: Z ~ N(0,1)

valor observat amb la mostra: z = 2.6

contrast: bilateral



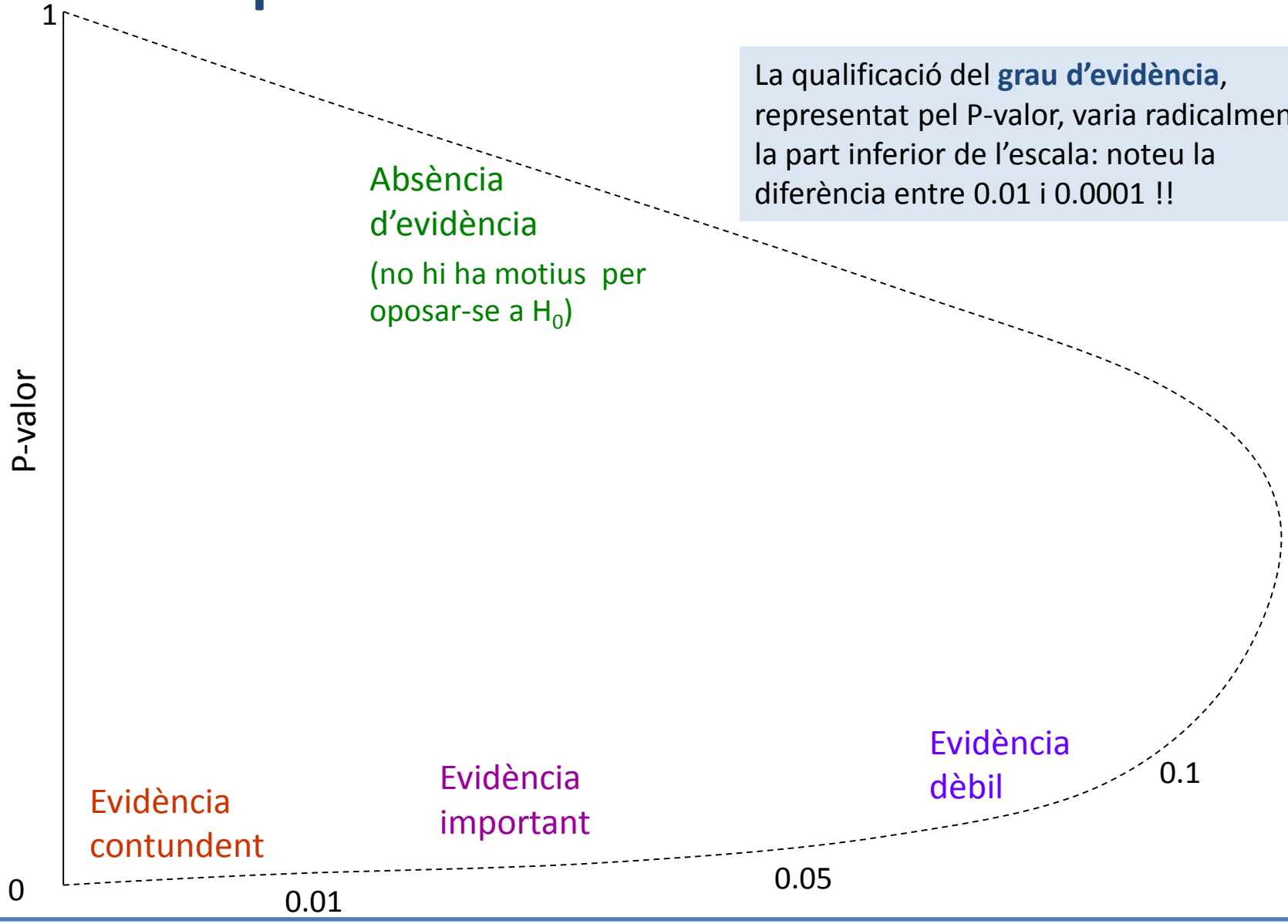
P(Z<-2.6)= 0.0047 i P(Z>2.6) = 0.0047

p-valor = P(|Z|>|z|)= 2 x 0.0047 = **0.0094** (Taules: 2·(1-0.9953); R: pnorm(-2.6)+(1-pnorm(2.6)))

P = P(v.a. Z "*més lluny*" de H_0 que el valor observat z)

RECORDEU: "sota H_0" = suposem ('temptativament') que és cert que els paràmetres poblacionals valen el que diu H_0

Proves d'hipòtesi. P-valor



La qualificació del **grau d'evidència**, representat pel P-valor, varia radicalment en la part inferior de l'escala: noteu la diferència entre 0.01 i 0.0001 !!

Malentessos amb el p-valor

- El *P*-valor diu amb quina freqüència poden passar events com el de la mostra (o més extrems) quan la hipòtesi H_0 és correcta:
  - Si el *P-valor* és petit $\rightarrow$ tenim evidència en contra de H_0
  - Si el *P*-valor no és petit, **NO** demostra la "veritat" de H_0
- *Incorrecte: P* = 0.000 $\rightarrow$ Correcte: *P*<0.001.
- *P-valor* **NO** és cap de les següents probabilitats:
  - la probabilitat d'"haver-se equivocat"
  - la probabilitat que la hipòtesi nul·la sigui certa
  - la probabilitat d'haver rebutjat erròniament la hipòtesi nul·la
- 1–*P*-valor **NO** és la probabilitat que la hipòtesi alternativa sigui certa
- Trobareu més a "Frequent misunderstandings" a Wikipedia

Proves d'hipòtesi. Resolució

1. Escollir una **variable** segons els objectius de l'estudi
2. Escollir un disseny i un **estadístic**
3. Definir una **hipòtesi** H_0 que es vol posar a prova, enfront una hipòtesi alternativa H_1
4. Especificar la **distribució** de l'estadístic si H_0 fos certa (i les premisses adients)
5. Amb les dades, calcular el **valor de l'estadístic** (z)
6. **Contrastar H_0**. Dues alternatives per fer-ho:
   a. Si $|z| > z_{1-\alpha}$ (unilateral) o $|z| > z_{1-\alpha/2}$ (bilateral) llavors rebutjar H_0 (H_0 és poc versemblant)
   b. Calcular el valor de P → Si $P < \alpha$, llavors rebutjar H_0 (H_0 és poc versemblant)
7. Afegir l'estimació per interval **IC(1-α)**

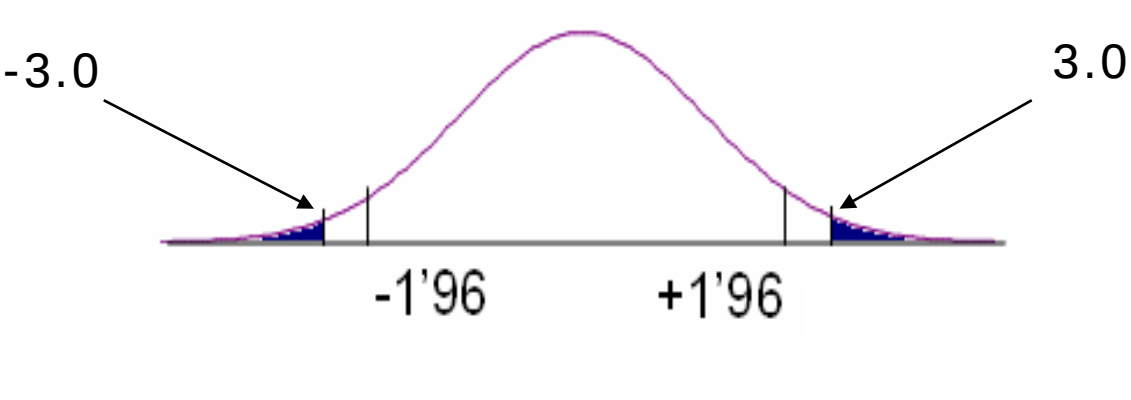
Problema: Com fixem α?

- Valorar el cost que representa una conclusió equivocada
- Definir un llindar arbitrari per fixar què es considera rellevant

Proves d'hipòtesi. Llindars d'acceptació

Es poden trobar els límits d'una regió crítica

- **Bilateral**: a l'esquerra de $z_{\alpha/2}$ (-1.96), i a la dreta de $z_{1-\alpha/2}$ (1.96)
- **Unilateral**:
  - per l'esquerra: per sota de z_{α} (-1.645)
  - per la dreta: per sobre de $z_{1-\alpha}$ (1.645)
- Si l'estadístic cau a la regió crítica (en blau al dibuix), llavors la hipòtesi nul·la (H_0) és dubtosa.
- Si l'estadístic cau a la regió d'acceptació (en blanc al dibuix), llavors no podem rebutjar H_0



Proves d'hipòtesi sobre la μ. Estadístic

- Si σ és coneguda, l'estadístic de referència és:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

- Si σ és desconeguda, l'estadístic de referència és

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

- Sota H_0, tindrà:

  - distribució Normal estàndard, si la grandària mostral és suficient:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \sim N(0,1) \qquad \text{si} \quad n \geq 100$$

  - distribució t Student amb (n-1) g.ll si la variable estudiada X és Normal

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1} \qquad \text{si} \quad X \sim N$$

Proves d'hipòtesi sobre μ. Exemple

Si en el cas d'una embotelladora de 1 litre tenim: $\bar{x}$ =997, s =10 i n=100. Podem pensar que la mitjana poblacional és μ = 1000 cc?

1. **Variable**: contingut en envasos de 1000cc
2. **Estadístic**: $\hat{t} = (\bar{x} - \mu)/(s/\sqrt{n})$
3. H_0 : μ = 1000cc vs H_1 : μ ≠ 1000cc
4. **Distribució de l'estadístic** sota H_0 : $(\bar{x} - \mu)/(s\sqrt{n}) \sim N(0,1)$ ja que n = 100;
5. **Càlculs**: $\hat{t} = (\bar{x} - \mu)/(s\sqrt{n}) = \frac{997-1000}{10/\sqrt{100}} = -3$
6. **P-valor** = Prob [(|z| > |-3|)] = 0.0027 [**Taules**: 2*(1-0.9987) ; **R**: pnorm(-3)+(1–pnorm(3)]
7. **Conclusió**: com que *P* és menor que α, es rebutja H_0: μ=1000cc

   Conclusió pràctica: ens estan estafant!

8. $\mathbf{IC(\mu, 0.95)} = \bar{x} \mp z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 997 \mp 1.96 \cdot \frac{10}{\sqrt{100}} = [995.04\,, 998.96]$

Nota: per graus llibertat propers a 100, D. t Student és molt propera a DN **R**: funció ***t.test***

Proves d'hipòtesi sobre μ. Exercici

En 9 fitxers, la diferencia D entre els temps d'execució de dos programes de compressió de fitxers ha estat de mitjana 6.71 i desviació 6.00. Acceptant que D ~ N, ¿es pot acceptar que E(D) = μ = 0? (és a dir, acceptar que els dos compressors tarden el mateix en mitjana?)

1. **Variable**: D (diferència en temps)
2. **Estadístic**: $\hat{t} = (\bar{d} - \mu)/(s/\sqrt{n})$
3. $\mathbf{H_0}: \mu = 0$ vs $\mathbf{H_1}: \mu \neq 0$
4. **Distribució de l'estadístic** sota H_0: $(\bar{d} - \mu)/(s\sqrt{n}) \sim t_{n-1}$ ja que D~N
5. **Càlculs**: $\hat{t} = (\bar{d} - \mu)/(s/\sqrt{n}) = \frac{6.71-0}{6/\sqrt{9}} = 3.355$
6. **P-valor = Prob [($|\hat{t}| > |-3.355|$)] = 0.01** **[R: `pt(-3.355) + (1-pt(3.355)]`**
7. **Conclusió**: es rebutja H_0: E(D) = μ = 0 ja que el p-valor (0.01) és < α

   Conclusió pràctica: no tarden el mateix
8. $\mathbf{IC(\mu, 0.95)} = \bar{x} \mp t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} = 6.71 \mp 2.306 \cdot \frac{6}{\sqrt{9}} = [2.1\,, 11.3]$

R: funció ***t.test***

Proves d'hipòtesi sobre π. Exemple

En el **exemple** anterior del basket, amb n=20 i 8 encerts.

1. **Variable**: resultat de cada tir (canasta o no)
2. **Estadístic**: nombre d'encerts X
3. H_0 : $\pi = 0.80$ (sóc un magnífic tirador) vs. H_1 : $\pi < 0.80$ (no sóc tan bo)
4. Si H_0 es certa: $X \sim B(n, \pi) = B(20, 0.80)$

   Premisses: mostra de tirs independents, amb la mateixa probabilitat d'encert

   [Si n és gran (i π lluny de 0 i de 1), pot ser més simple utilitzar l'aproximació de la Binomial a la Normal (en aquest cas, treballem amb P=X/n enlloc de X): Si H_0 es certa: $P \sim N(\pi, \pi(1-\pi)/n)$. Requereix una premissa addicional: $n\cdot\pi > 5$ i $n\cdot(1-\pi) > 5$ (No s'aplicarà en l'exemple)]
5. **Càlcul de l'estadístic** : x = 8
6. ***P-valor*** = P (X més lluny de H_0 que x) = 0.0001 (*unilateral*)
7. **Conclusió**: **SÍ,** hi ha (forta) evidència en contra de H_0: π=0.80
8. El càlcul de l'**IC** quan la *n* és petita no és senzill, però es pot trobar amb la instrucció de R ***binom.test***. En aquest exemple val [0, 0.61]

Si la mostra és petita obtindreu intervals molt amples... Com ha de ser!

Proves d'hipòtesi sobre π. Exercici

Llencem una moneda 100 vegades i obtenim 63 cares. Està "trucada" la moneda?

1. **Variable**: resultat de cada llançament(cara o creu)
2. **Estadístic**: $\mathrm{Z} = (P - \pi)/\sqrt{\pi \cdot (1-\pi)/n}$
3. **H_0 :** $\pi = 0.50$ (moneda equilibrada vs. **H_1 :** $\pi \neq 0.50$ (moneda trucada
4. Si H_0 es certa: $Z \sim N\left(\pi, \sqrt{\pi \cdot (1-\pi)/n}\right) = N(0.5\,, 0.05)$

   Premisses: 1) llançaments independents, amb la mateixa prob. de cara i (2) $n\pi>5$; $n(1-\pi)>5$
5. **Càlcul de l'estadístic**: $\mathrm{Z} = \frac{0.63-0.5}{\sqrt{0.5\cdot(1-0.5)/100}} = 2.6$
6. ***P-valor*** = P (X més lluny de H_0 que x) = 0.0094 (*bilateral*
7. **Conclusió**: **SÍ,** hi ha (certa) evidència en contra de H_0: $\pi = 0.50$
8. ~~El càlcul de l'**IC** quan la *n* és petita no és senzill, però es pot trobar amb la instrucció de R ***binom.test***. En aquest exemple val:~~

$$IC(\pi, 0.95) = p \mp z_{0.975} \cdot \sqrt{\left(\pi \cdot (1-\pi)\right)/n} = 0.63 \mp 1.96 \cdot \sqrt{(0.63 \cdot 0.37)/100} = [0.53, 0.73]$$

Proves d'hipòtesi vs IC

- Per exemple, sobre el valor del paràmetre π en la població:
  - **PH** fa una pregunta "tancada": **¿és $\pi = 0.5$?**
  - **IC** fa una pregunta "oberta": **¿quin es el valor de π ?**
- Donar els resultats sempre amb IC implica:
  - Si es rebutja H_0, dir on es troba el paràmetre
  - Si no es rebutja H_0, quantificar la informació de que es disposa
  - L'IC proporciona informació més fàcil d'interpretar que el P-valor.

Formulari. Proves d'hipòtesi

Paràmetre	Hipòtesi	Estadístic	Premisses	Distribució sota H_0	Criteri Decisió (Risc α)
μ	H : μ = μ_0	$\hat{z} = \frac{\bar{y} - \mu_0}{\sqrt{\sigma^2/n}}$	Y→N ò n≥30 i σ coneguda	$\hat{z} \rightarrow$ N(0,1)	Rebutjar H si $\lvert \hat{z} \rvert > z_{1-\alpha/2}$ ($\lvert \hat{z} \rvert$>1.96 amb α=5%)
μ	H : μ = μ_0	$\hat{t} = \frac{\bar{y} - \mu_0}{\sqrt{s^2/n}}$	Y → N	$\hat{t} \rightarrow t_{n-1}$	Rebutjar H si $\lvert \hat{t} \rvert > t_{n-1,1-\alpha/2}$ ($\lvert \hat{t} \rvert > t_{n-1,0.975}$ amb α=5%)
μ	H : μ = μ_0	$\hat{z} = \frac{\bar{y} - \mu_0}{\sqrt{s^2/n}}$	n ≥ 100	$\hat{z} \rightarrow$ N(0,1)	Rebutjar H si $\lvert \hat{z} \rvert > z_{1-\alpha/2}$ ($\lvert \hat{z} \rvert$>1.96 amb α=5%)
π (Binomial)	H : π = π_0	$\hat{z} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$	$(1-\pi_0)n \geq 5$ $\pi_0 n \geq 5$	$\hat{z} \rightarrow$ N(0,1)	Rebutjar H si $\lvert \hat{z} \rvert > z_{1-\alpha/2}$ ($\lvert \hat{z} \rvert$>1.96 amb α=5%)
Anexe: λ (Poisson)	H : λ = λ_0	$\hat{z} = \frac{f - \lambda_0}{\sqrt{\lambda_0}}$	$\lambda_0 \geq 5$	$\hat{z} \rightarrow$ N(0,1)	Rebutjar H si $\lvert \hat{z} \rvert > z_{1-\alpha/2}$ ($\lvert \hat{z} \rvert$>1.96 amb α=5%)
σ (normal)	H : σ = σ_0	$\hat{\chi}^2 = \frac{s^2 \cdot (n-1)}{\sigma^2}$	Y → N	$\hat{\chi}^2 \rightarrow \chi^2_{n-1}$	Rebutjar H si $\hat{\chi}^2 < \chi^2_{n-1,\alpha/2}$ o $\hat{\chi}^2 > \chi^2_{n-1,1-\alpha/2}$
En les proves unilaterals s'acumula el valor de P-valor a un sol costat:				H: μ ≤ μ_0 → Rebutjar H si $\hat{z} > z_{1-\alpha}$ H: μ ≥ μ_0 → Rebutjar H si $\hat{z} < -z_{1-\alpha}$	

Annexe: Premissa de Normalitat

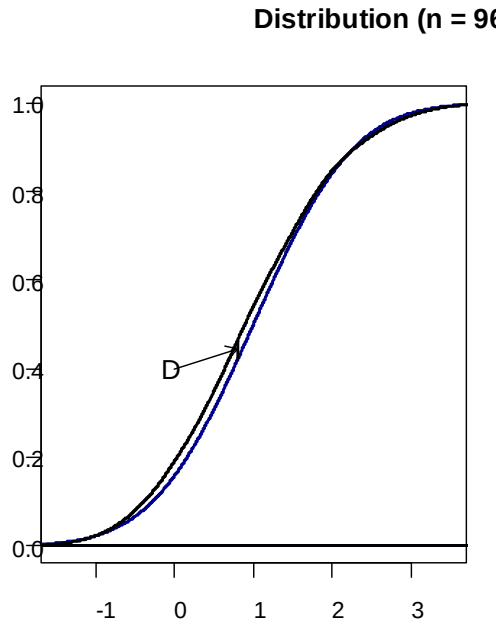
- Hi ha dues mesures que ajuden a valorar el grau d'ajustament, afinitat o similitud a una certa distribució de referència

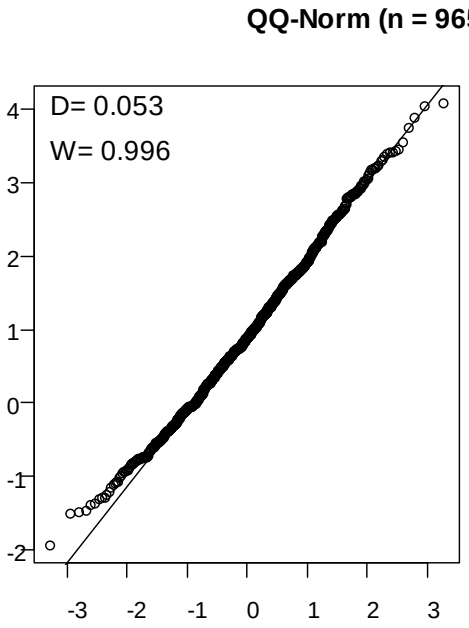
Kolmogorov-Smirnov (Estadístic D)	Shapiro-Wilk (Estadístic W)
Distància màxima entre la funció de distribució empírica i la teòrica. Valors elevats indiquen No Normalitat Entre 0 i 1 (usualment, prop de 0). Valors alts indiquen des-ajustament	Mesura la correlació entre els quantils observats i els teòrics. Valors elevats indiquen Normalitat Entre 0 i 1 (usualment, prop de 1). Valors alts indiquen bon ajustament

- Tots dos estadístics fluctuen a les mostres i han de interpretar-se amb prudència. Proporcionen **P-valors**, però farem només un anàlisi descriptiu i visual (qqplot)
- A continuació mostrem alguns exemples generats a partir de distribucions conegudes i amb diferents mides mostrals.

Annexe: Premissa de Normalitat

- Els paquets solen proporcionar P-valors de la H_0: X ~ N
- No ens hem de fiar massa dels P-valors obtinguts d'aquestes proves
- Es millor emprar les mesures i les eines gràfiques (QQ-Norm, PP-Norm)

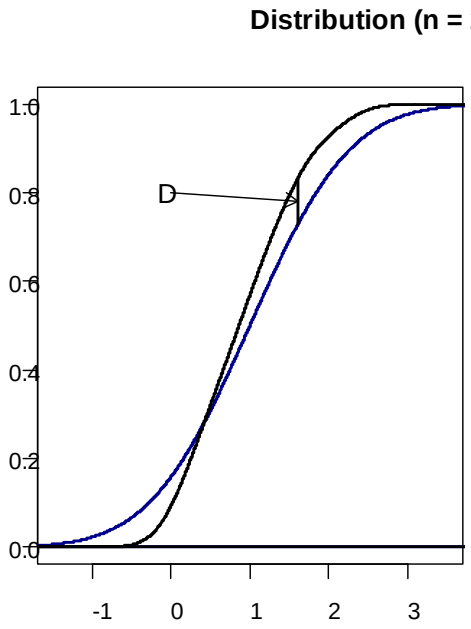


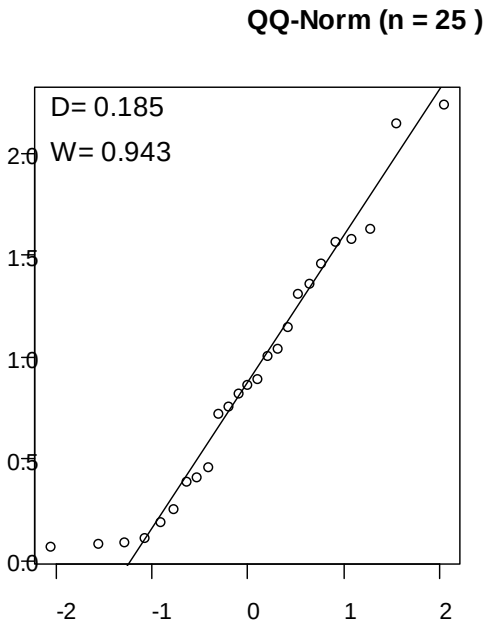


- Aquestes 965 observacions van estar generades seguint una distribució Normal
- D, W i el QQ-Norm mostren que les dades s'ajusten prou bé a una Normal
- No obstant, els p-valors de les proves (0.008 i 0.013) ens farien rebutjar la hipòtesi de Normalitat

Annexe: Premissa de Normalitat

- Aquí tenim un exemple del cas contrari

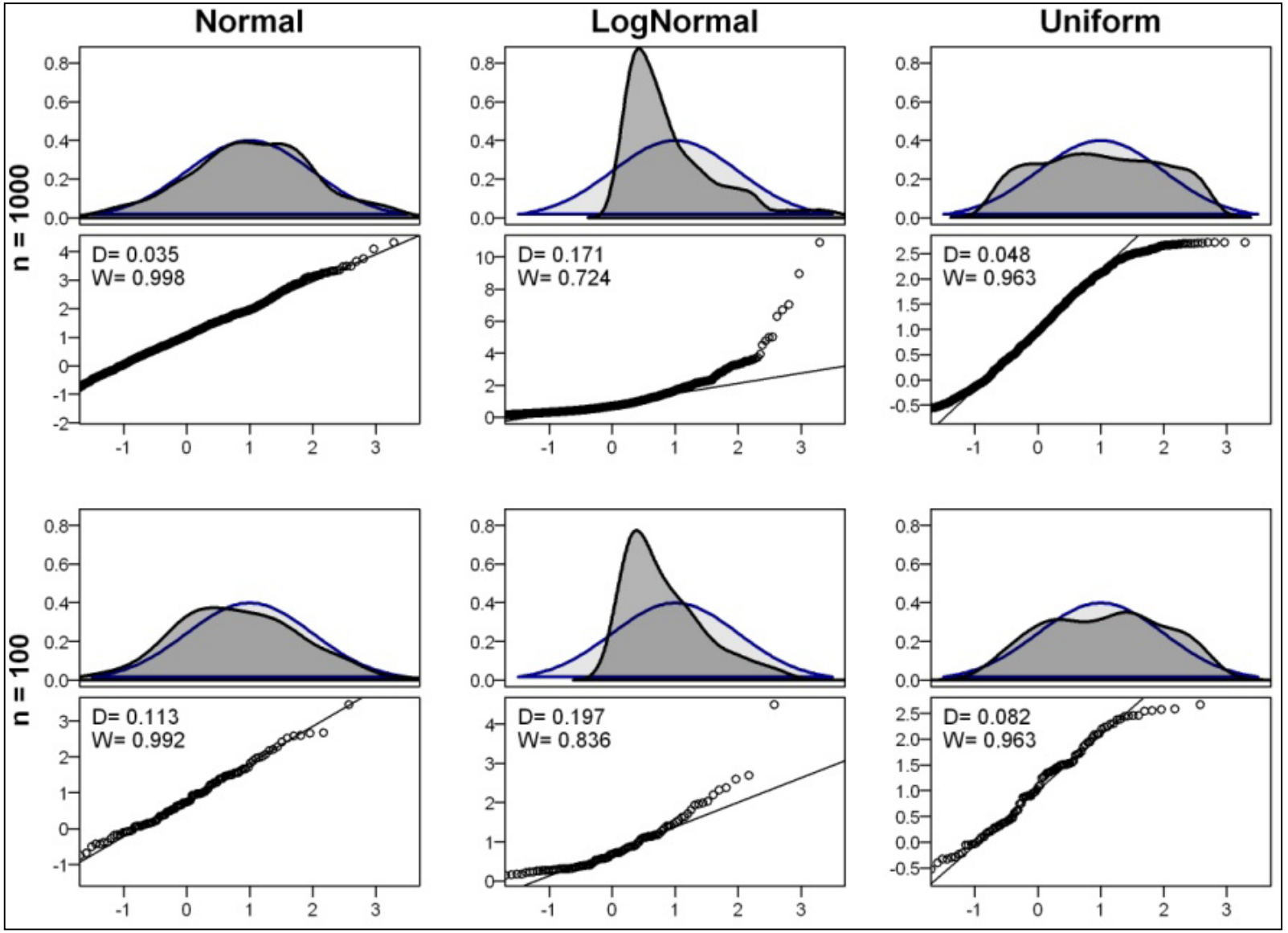




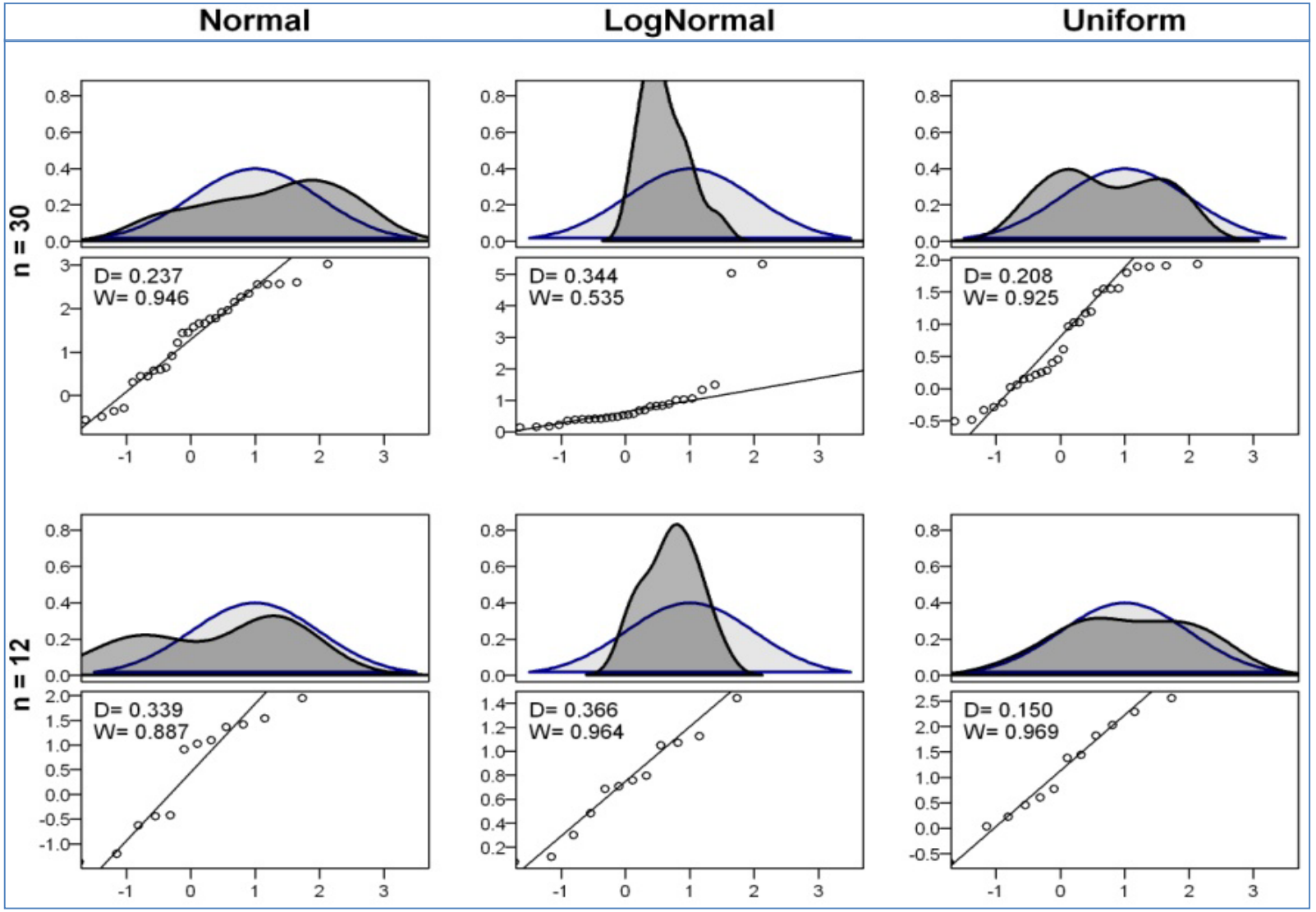
- Aquestes 25 observacions van estar generades seguint una distribució Exponencial
- F_X i D mostren la distància amb la teòrica, encara que W té una bona correlació.
- No obstant, els P-valors de les proves (0.32 i 0.17) no ens fan rebutjar la hipòtesi de Normalitat

Noteu la paradoxa: quan la mostra és petita, és important detectar la No Normalitat, però els P valors fallen. En canvi, quan la mostra és gran, és poc important detectar-la (pel bon comportament asimptòtic dels estimadors) però el P-valor detecta desviacions irrellevants.

Annexe: Premissa de Normalitat (n gran)



Annexe: Premissa de Normalitat (n petita)



Disseny d'experiments

Bloc 5 – Probabilitat i Estadística

Abril 2016

Índex

Inferència estadística. Guió

Guió de la part d'Estadística de PE:

- B4: Tècnica general de la inferència [estadística]
  - estimar un paràmetre (*Intervals de Confiança*)
  - refutar un paràmetre (*Proves d'Hipòtesis*)
- **B5: Aplicació (I): Avaluació de millores**
  - ***Disseny d'experiments:* comparació de dues poblacions.**
- B6: Aplicació (II): Predicció
  - *Previsió* d'una var. resposta, en funció d'una var. explicativa.

Introducció al disseny d'experiments

Distribució F de Fisher-Snedecor

- **Definició**: Siguin $X_1 \sim \chi^2_n \quad i \quad X_2 \sim \chi^2_m$. Llavors:

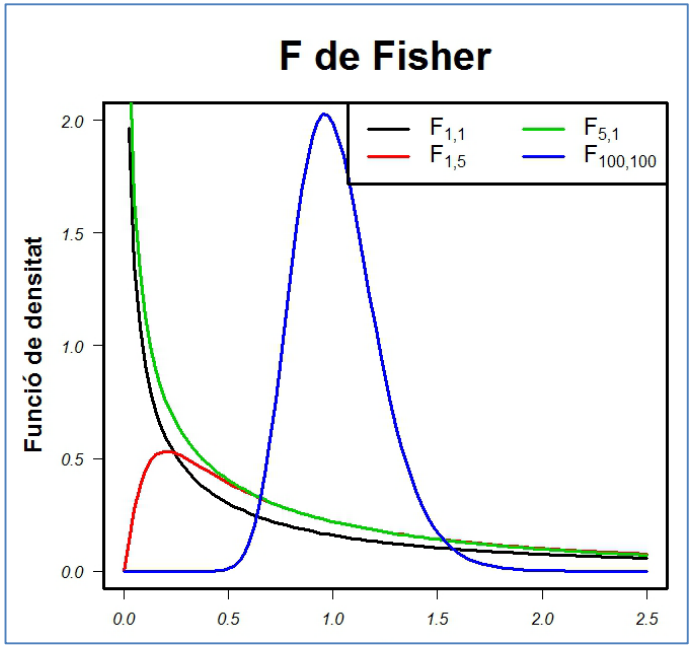
$$Y = \frac{X_1/n}{X_2/m} \sim F_{n,m}$$

1/Y~$F_{m,n}$

- **Notació**: $F \sim F_{n,m}$
- **Paràmetres**: *n* (graus de llibertat numerador)

  m (graus de llibertat denominador)
- **Funció de probabilitat i distribució**:

[F distribution at Wikipedia]



NOTA: La distribució F de Fisher, la farem servir per comparar variàncies de 2 poblacions

R: df, pf, qf

Script per veure que el quocient de χ^2 dividits pels g.ll és una F

```
M=500 ; n=5; m=7
samplechi2n = rchisq(M,n)
samplechi2m = rchisq(M,m)
F = (samplechi2n/n) / (samplechi2m/m)
hist(F, breaks="Scott", freq=FALSE)
curve(df(x, n, m),add=TRUE,col=2,lwd=2)
quantile(F, c(0.25, 0.50, 0.75))
qf(c(0.25, 0.50, 0.75), n, m)
```

Distribució F de Fisher-Snedecor. Càlcul de probabilitats

- Sigui $Y_1 \sim F_{1,\infty}$

  $P(Y_1 > 3.84) = 0.05$ [Taula F(0.95): 1ª columna, última fila]

- Sigui $Y_2 \sim F_{10,5}$

  $P(Y_2 > 10.05) = 0.01$ [Taula F(0.99): 10a columna, 5a fila]

  `R: qf(0.99,10,5) = 10.05`

  `R: 1-pf(10.05,10,5) = 0.01`

- Sigui $Y_3 \sim F_{5,10}$ (Recordeu: $Y_3 = 1/Y_2$)

  $P(Y_3 > 1/10.05) = P(1/Y_2 > 1/10.05) = P(Y_2 < 10.05) = 0.99$

  `R: 1-pf(1/10.05,5,10) = 0.99`

  `R: pf(10.05,10,5) = 0.99`

  [Cas general: Si $Y_1 \sim F_{n,m}$ i $Y_2 \sim F_{m,n}$ llavors $P(Y_1 > y) = P(Y_2 < 1/y)$]

Tipus de variables

- **Objectiu**: estimar l'efecte (causal)
  - d'una **intervenció (X)**
  - en una **resposta (Y)**
  - donades unes **condicions (Z)**
- La resposta Y ha de mesurar el nostre objectiu [Ex: temps d'execució d'un algoritme]
- La intervenció X és el nostre potencial per canviar el futur [Ex: algoritme A o B]
- Les condicions Z 'predeterminen' el futur (resposta Y) [Ex: CPU emprada]
  - Haurien de ser idèntiques pels diferents valors de X
  - Veurem que baixen σ_Y^2 i permeten augmentar la potència
  - Veurem al B6 com utilitzar-les per anticipar Y

Observar versus assignar

- **Objectiu**: Observar els canvis a la distribució de **Y** quan **X** canvia [Ex: estudiar un canvi en un paràmetre de Y com la esperança $\mathbf{E(Y) = \mu_Y}$ si es passa de $\mathbf{X_A}$ a $\mathbf{X_B}$]

- **Observar**: Sobre les altres condicions Z no tenim control i no podem evitar que canviïn conjuntament amb X i, en conseqüència, no sabrem si els canvis a Y es deuen a X o a alguna altra Z. [Ex: nombre d'errors de sintaxi en un codi R segons el SO que cada alumne ha escollit lliurament. Els que empren Linux són millor programadors i faran menys errors → En Linux es fan menys errors?]

- **Assignar**: si assignem les X, podrem controlar les Z per que siguin independents de X, i no puguin ser una explicació de les variacions a la resposta Y. [Ex: nombre d'errors de sintaxi en un codi R segons el SO que cada alumne ha emprat segons assignació. En teoria, no s'observaran diferències]

Tècniques estadístiques

[Els problemes es resoldran d'acord amb les tècniques del B4]

- **IC**: permet estimar la magnitud de l'efecte de X [Ex: si comparem $\bar{y}_A$ i $\bar{y}_B$ podem trobar una estimació de la diferència de mitjanes poblacionals $\mu_A - \mu_B$]
  - L'IC és el resultat definitiu d'un estudi experimental: és directe i simple d'interpretar
- **PH**: permet rebutjar valors rellevants de l'efecte [Ex: si comparem $\bar{y}_A$ i $\bar{y}_B$ podem decidir si la diferència de mitjanes poblacionals $\mu_A - \mu_B$ és nul·la]
  - habitualment, la hipòtesi nul·la (H_0) equival a un "punt mort" o a l'estat actual (que es voldria rebutjar)
  - la hipòtesi alternativa seria un nou estat, superior a l'actual [Ex: H_0: $\mu_A = \mu_B$; H_1: $\mu_A > \mu_B$ o, mes neutre, H_1: $\mu_A \neq \mu_B$]
  - en funció de l'evidència disponible amb la mostra, una decisió o l'altre condueix a posar en marxa diferents accions

Atenció: Les decisions poden ser equivocades, cal balancejar el risc dels errors i el cost de les seves conseqüències (es veurà més endavant)

Tipus de dissenys

- **Segons la resposta**. Per comparar **2 opcions X** estudiarem proves segons la **resposta Y** sigui numèrica o dicotòmica:

Numèrica: comparem mitjanes	*Dicotòmica:* comparem proporcions
Cas: Comparació de 2 mitjanes	Cas: Comparació de 2 proporcions
Pregunta: E(Y) depèn de la opció *X*?	Pregunta: π(Y) depèn de la opció *X*?
Hipòtesi formal: H_0: $\mu_A = \mu_B$	Hipòtesi formal: H_0: $\pi_A = \pi_B$
IC95% de l'efecte diferencial: $\mu_A - \mu_B$	IC95% de l'efecte diferencial: $\pi_A - \pi_B$
Ex: Comparar temps de 2 algoritmes	Ex: Comparar proporció d'aprovats en 2 cursos

- **Segons la recollida de dades**:
  - *Mostres independents:* cada cas és una **mesura independent** [Ex: Per comparar els temps de càrrega de 2 navegadors, els hi faig carregar 30 pàgines web diferents a cadascun]
  - *Mostres aparellades:* cada cas dóna lloc a dues mesures, **parells de mesures** [Ex: Per comparar els temps de càrrega de 2 navegadors, faig carregar les mateixes 30 pàgines web a cadascun]

Nota: Sempre que es pugui, s'ha de fer un disseny amb dades aparellades ja que és més eficient

Mostres Independents

Comparació de mitjanes ($\mu_1=\mu_2$). Mostres independents

- La prova formal que volem posar a prova és:

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases} \Leftrightarrow \begin{cases} H_0: \mu_1 - \mu_2 = 0 \\ H_1: \mu_1 - \mu_2 \neq 0 \end{cases}$$

- Els indicadors (Esperança i Variància) de la diferència són:

$$E(\bar{y}_1 - \bar{y}_2) = E(\bar{y}_1) - E(\bar{y}_2) = \mu_1 - \mu_2$$

$$\mathrm{V}(\bar{y}_1 - \bar{y}_2) = V(\bar{y}_1) + V(\bar{y}_2) - 2Cov(\bar{y}_1, \bar{y}_2) = (m.a.s) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

- Per tant, sota H_0 es compleix:

$$\hat{z} = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = (si \;\; H_0 \; és \; igualtat) = \frac{(\bar{y}_1 - \bar{y}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

[Premisses: $Y_1 \sim N(\mu_1, \sigma_1^2)$ i $Y_2 \sim N(\mu_2, \sigma_2^2)$ i m.a.s. independents]

Nota: Normalment, σ_1 i σ_2 seran **desconegudes**, però les podem suposar iguals (malgrat ser desconegudes). El que farem serà combinar les estimacions s_1 i s_2 donant lloc a una única estimació "pooled" (veure següent diapositiva)

Comparació de mitjanes ($\mu_1=\mu_2$). Mostres independents

- Si podem assumir igualtat de variàncies: $s_1^2 = s_2^2 = s^2$, llavors, s_1^2 i s_2^2 seran estimadors del mateix paràmetre s^2. En aquesta condició, una bona estimació de s^2 s'obté ponderant s_1^2 i s_2^2 amb els seus respectius graus de llibertat donant lloc a la **s "pooled"**:

$$\hat{\sigma}^2 = s^2 = \frac{(n_1-1)\cdot s_1^2 + (n_2-1)\cdot s_2^2}{(n_1-1)+(n_2-1)} = \frac{\sum_{i=1}^{n_1}(y_{1i}-\bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i}-\bar{y}_2)^2}{n_1+n_2-2}$$

- Al substituir **σ** pel seu estimador ***s*** *pasem a tenir una t-Student:*

$$\hat{z} = \frac{(\bar{y}_1-\bar{y}_2)-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1) \rightarrow \hat{t} = \frac{(\bar{y}_1-\bar{y}_2)-(\mu_1-\mu_2)}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Nota: Aquest estadístic representa la "senyal" proporcionada per la distància entre mitjanes relativa al "soroll" aleatori que porti aquesta senyal

Comparació de mitjanes ind. ($\mu_1=\mu_2$). Exemple

Els temps mitjans d'execució de dos programes provats en diferents bancs de dades independents (n_1=50 i n_2=100) han sigut: $\bar{y}_1 = 24$ *i* $\bar{y}_2 = 21$ amb s_1= 8 i s_2= 6. Suposem $\sigma^2_1 = \sigma^2_2$. Es desitja decidir quin programa es posa al mercat. Tenen rendiments diferents?

1. **Variables**: Y_1 i Y_2 de temps d'execució

2. **Estadístic**: $\hat{t} = \frac{(\bar{y}_1-\bar{y}_2)}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ amb $s^2 = \frac{(n_1-1)\cdot s_1^2+(n_2-1)\cdot s_2^2}{(n_1-1)+(n_2-1)}$

3. **Hipòtesi**: $\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases}$ (bilateral)

4. **Distribució estadístic**: t_{148} **Premisses**: m.a.s. i Y_1 i Y_2 Normals amb $\sigma_1 = \sigma_2 = \sigma$

5. **Càlculs estadístic**: $s^2 = \frac{(50-1)\cdot s_1^2+(100-1)\cdot s_2^2}{(50-1)+(100-1)} = 45.27 \rightarrow \hat{t} = \frac{(24-21)}{6.72\sqrt{\frac{1}{50}+\frac{1}{100}}} = 2.58$

6. **P-valor** = $P(|t_{148}| > |2.58|) = 0.011$

[R: pt(-2.58,148)+(1-pt(2.58,148)] [(Taules ($t_{148} \approx Z$): 2·(1-0.9951)]

R: t.test(Y_A, Y_B, var.equal=TRUE)

Comparació de mitjanes ind. ($\mu_1=\mu_2$). Exemple (cont)

7. **Conclusió**:
   - *P-valor = 0.011 < 0.05 → Rebutgem H0 (descartem que $\mu_1=\mu_2$) amb un p-valor de 0.011*
   - *$\hat{t}$ = 2.58 > 1.976 = $t_{148,0.975}$ → Rebutgem H_0 (descartem que $\mu_1=\mu_2$) amb risc α = 0.05*

   Conclusió pràctica:
   - *(PH): Aquestes dades (o més extremes) són poc probables (P=0.011) si els rendiments fossin iguals: Rebutgem H_0*

8. **Inferència amb Interval de Confiança**:

$$IC(\mu_1-\mu_2, 95\%) = (\bar{y}_1-\bar{y}_2) \mp t_{n_1+n_2-2,1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s^2}{n_1}+\frac{s^2}{n_2}} = 3 \mp 1.976 \cdot 1.165$$

$$= [0.70\,, 5.30]$$

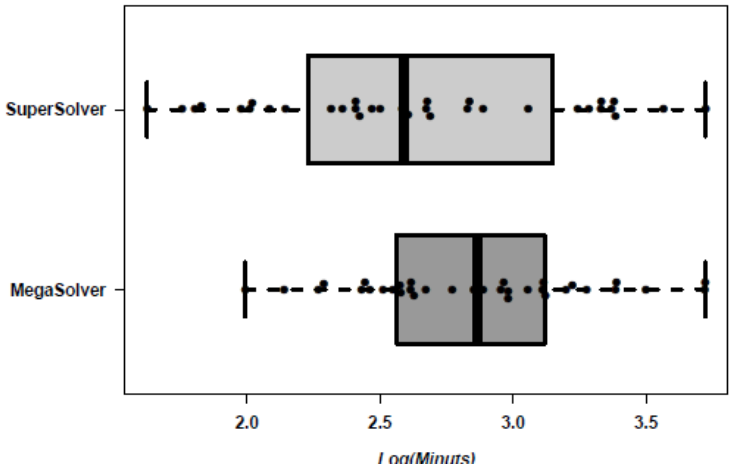
   - *IC95%: El programa "2" triga en mitjana entre 0.7 i 5.3 segons menys amb una confiança del 95%*

Comparació de mitjanes ind. ($\mu_1=\mu_2$). Exercici

Per comparar la velocitat amb la qual resolen dos servidors diferents, *SuperSolver* i *MegaSolver*, problemes d'optimització s'envia un total de 70 problemes de maximització diferents als dos servidors, 35 a cadascun. Pel fet que el temps que triguen els servidors per resoldre els problemes, és asimètrica cap a la dreta, treballem a continuació amb els logaritmes dels temps. Siguin X el logaritme del temps que triga el *SuperSolver* i Y el del *MegaSolver* .

Els valors descriptius a cada mostra són els següents i a més a més es mostra una representació gràfica:

	Mitjana	Mediana	Desv. est.	Mínim	Màxim
SuperSolver	2,63	2,59	0,57	1,63	3,72
MegaSolver	2,85	2,86	0,44	1,99	3,72



Comparació de mitjanes ind. ($\mu_1=\mu_2$). Exercici

1. **Variables**:
2. **Estadístic**:
3. **Hipòtesi**:
4. **Distribució estadístic**: **Premisses**:
5. **Càlculs estadístic**:
6. **P-valor** =
7. **IC95%**

Conclusió:

Comparació de var ($\sigma_1^2 = \sigma_2^2$). Mostres independents

- La prova formal que volem posar a prova és:

$$\begin{cases} H_0: \sigma_1 = \sigma_2 \\ H_1: \sigma_1 \neq \sigma_2 \end{cases}$$

- Sota H_0, tenim l'estadístic: amb la següent distribució

$$\hat{F} = \frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1}$$ [Premisses: m.a.s independents i $Y_1 \sim N(\mu_1, \sigma^2_1)$ i $Y_2 \sim N(\mu_2, \sigma^2_2)$]

- Es rebutjarà H_0 sempre que:

$$\hat{F} < F_{n1-1,n2-1,\alpha/2} \text{ o bé que } \hat{F} > F_{n1-1,n2-1,1-\alpha/2}$$

- Però si posem la S més gran (S_M) al numerador, l'estadístic: $\hat{F} = \frac{S_M^2}{S_m^2} \sim F_{n_M-1, n_m-1} > 1$, llavors només cal mirar la cua dreta a les taules, i es rebutjarà H_0 sempre que

$$\hat{F} > F_{n_M-1, n_m-1}$$

- No obstant, si la prova és **unilateral**, tot el risc α s'acumula al costat adient (depenent del sentit de H_1 i de quin grup presenta la major variància mostral)

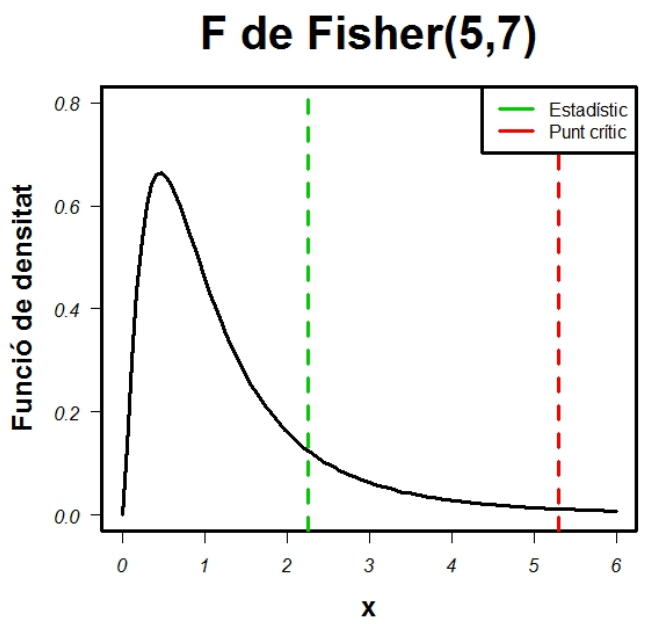
Nota tècnica: Pel càlcul del p-valor, si la prova és bilateral, sempre s'haurà d'agafar la cua més petita (esquerra o dreta) i multiplicar-la per 2

Comprar. de var ($\sigma_1^2 = \sigma_2^2$). Mostres independents. Ex.

Estem interessats en comparar la duració dels recanvis dels cartutxos de tinta de dos marques: "ORIGINAL, S.A." (**A**) y "YO_TAMBIÉN_LO_HAGO, S.L." (**B**). Hem comprovat que la marca B té una mitjana major de duració que la marca A. Però sospitem que la variabilitat pot ser diferent (provarem si **són iguals o no**). En dues mostres, hem obtingut els següents resultats:

A	$\bar{y}_A = 363$	S_A^2=64	$n_A = 8$
B	$\bar{y}_B = 407$	S_B^2=144	$n_B = 6$

1. **Variable**: Y_A i Y_B (duracions)
2. **Estadístic**: S^2_{Major}/S^2_{Menor}
3. **Hipòtesis**: $H_0: \sigma_B^2 = \sigma_A^2$ vs $H_1: \sigma_B^2 \neq \sigma_A^2$
4. **Distribució estadístic**: $F_{5,7}$
5. **Càlculs**: $S_B^2/S_A^2 = 144/64 = 2.25$
6. **P-valor**: $2 \cdot P(F_{5,7} > 2.25) = 0.32$
7. **Conclusió**: Com 2.25 < 5.29 = $F_{5,7,0.975}$ i p-valor>0.05 llavors no podem rebutjar H_0. No hem trobat evidència per contradir que les variàncies poblacionals siguin iguals

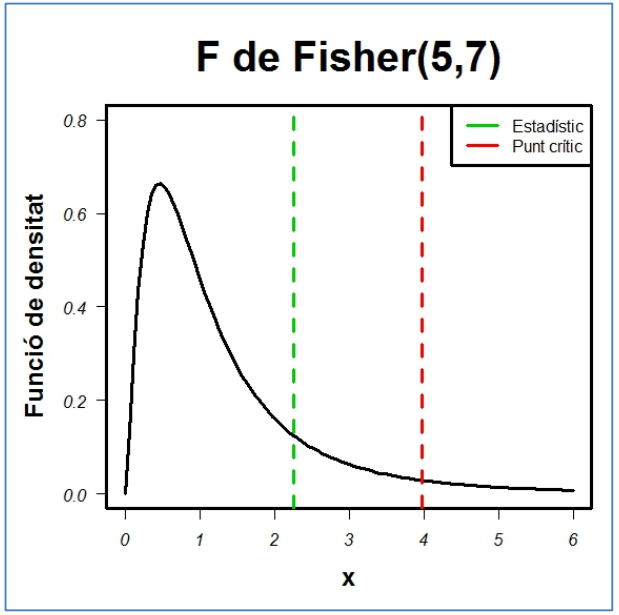


Compar. de var ($\sigma_1^2 = \sigma_2^2$). Mostres independents. Ex.

I si posem a prova si la variabilitat **és igual versus superior** en B?

R: `var.test(`Y_A`, `Y_B`)`

1. **Variable**: Y_A i Y_B (duracions)
2. **Estadístic**: S^2_{Major}/S^2_{Menor}
3. **Hipòtesis**: $H_0: \sigma_B^2 = \sigma_A^2$ vs $H_1: \sigma_B^2 > \sigma_A^2$
4. **Distribució estadístic sota H_0**: $F_{5,7}$
5. **Càlculs**: $S_B^2/S_A^2 = 144/64 = 2.25$
6. **P-valor:** $P(F > 2.25) = 0.16$
7. **Conclusió**: No podem rebutjar H_0



Ja que 2.25 < 3.97 = $F_{5,7,0.95}$ [**R:** `qf(0.95,5,7)= 3.97`]

O bé *P*-valor = 0.16 > 0.05 = α [**R:** `1-pf(2.25,5,7)= 0.16`]

No hem trobat evidència per contradir que les variàncies poblacionals siguin iguals.

Nota: si la prova és bilateral, posem al numerador el valor més gran segons les dades; però si és unilateral, segons H_1

Compar. de var ($\sigma_1^2 = \sigma_2^2$). Mostres independents. Exer.

Posa a prova la igualtat de variàncies en l'exercici dels optimitzadors

	Mitjana	Mediana	Desv. est.	Mínim	Màxim
SuperSolver	2,63	2,59	0,57	1,63	3,72
MegaSolver	2,85	2,86	0,44	1,99	3,72

1. **Variable**:
2. **Estadístic**:
3. **Hipòtesis**:
4. **Distribució estadístic sota H_0**:
5. **Càlculs**:
6. **P-valor:**
7. **Conclusió**:

Comparació de prop ($\pi_1 = \pi_2$). Mostres independents

- Ara desitgem posar a prova: $H_0 : \pi_1 = \pi_2$ vs $H_1 : \pi_1 \neq \pi_2$

[Ex: existeix la mateixa proporció d'alumnes que utilitzin PC (vs MAC) en la Facultat d'Informàtica i en l'Escola de Telecomunicacions?]

- Si el disseny de l'estudi, comporta **2 mostres**, una en la FIB i un altre en ETSETB, el plantejament és "**homogeneïtat**" entre 2 poblacions:

$$¿P(PC|FIB) = \pi_1 = \pi_2 = P(PC|ETSETB)?$$

- Mentre que si es tracta d' **1 única mostra**, en la que s'han recollit ambdues variables, el plantejament és "**independència**" de 2 variables aleatòries:

$$¿P(PC \cap FIB) = P(PC) \cdot P(FIB)?$$

- Tant l'objectiu d' "**homogeneïtat**" com el de "**independència**" es resolen de la mateixa manera. Per tant, no farem distinció entre un i altre.

Comparac. de prop ($\pi_1 = \pi_2$). Mostres independents. Ex

- L'exemple dels PC's podem presentar-ho en forma de taula 2x2. Suposem que hem obtingut 2 m.a.s. de 100 casos en cada facultat [situarem el recompte o freqüència $\mathbf{f_{ij}}$ de casos observats en la fila "i", columna "j"]

freqüència observada: f_{ij}	FIB	Telecos	Total
Utilitzen PC	77	63	140
Utilitzen MAC	23	37	60
Total	100	100	N = 200

- Quines serien les freqüències que es podria esperar ($\mathbf{e_{i,j}}$) de ser certa H_0? (Si fos certa H_0, llavors el ús de PC o MAC seria independent de la facultat i passaria, per exemple, que $P(PC \cap FIB) = P(PC) \cdot P(FIB)$
- Per calcular els $\mathbf{e_{i,j}}$, primer estimem les distribucions marginals:

  P(FIB)= 100/200 ; P(Telecos)= 100/200 ; P(PC)= 140/200 ; P(MAC)= 60/200
- Llavors, es poden calcular les freqüències esperades (e_{ij}) sota H_0. P.ex:

  $\mathbf{e_{FIB,PC}} = N \cdot P(PC \cap FIB) = N \cdot P(PC) \cdot P(FIB) = 200 \cdot (140/200) \cdot (100/200) = 70$

Comparac. de prop ($\pi_1 = \pi_2$). Mostres independents. Ex

freqüència esperada: e_{ij}	FIB	Telecos	Total
Utilitzen PC	70	70	140
Utilitzen MAC	30	30	60
Total	100	100	N = 200

Noteu que:

(1) $e_{ij} = \frac{e_{i\cdot} \cdot e_{\cdot j}}{e_{\cdot\cdot}} = \frac{(total\ de\ fila\ i)\cdot(total\ de\ columna\ j)}{(total\ de\ totals)}$

(2) calculat el primer (**e_{11}**) , en una taula 2x2, els altres s'obtenen per diferència (e_{12}= 140-70 = 70)

- Ara tenim el resultat empíric (**f_{ij}**) i l'esperat sota H_0 (**e_{ij}**) i podem estudiar les seves divergències a través de l'estadístic de Pearson (chi-quadrat):

$$\hat{X}^2 = \sum_{\forall ij} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

- **Sota H_0** → $\hat{X}^2 \sim \chi^2_{(I-1)(J-1)}$ on I,J són el número de files i columnes
- **Premissa**: **$e_{ij} \geq 5$** $\forall$**i,j**
- En el nostre cas, I=J=2, i la distribució χ^2 de referència té 1 g.d.l ja que (I-1)·(J-1) = 1·1 = 1

Comp. de prop ($\pi_1 = \pi_2$) en mostres indep. Exemple

1. **Variable**: PC o MAC

2. **Estadístic**: $\hat{X}^2 = \sum_{\forall ij} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$

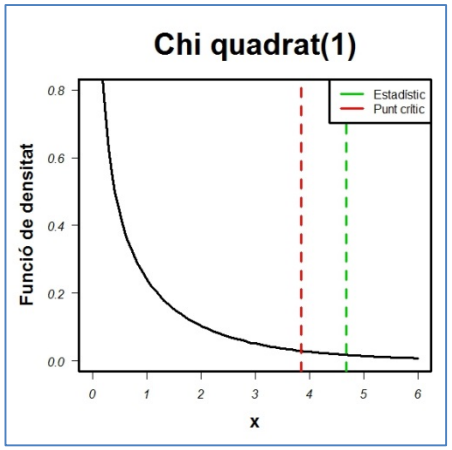
$(f_{ij} - e_{ij})^2/e_{ij}$	FIB	Telecos	Total
Utilitzen PC	0.7	0.7	
Utilitzen MAC	1.633	1.633	
Total			4.667

3. **Hipòtesis H_0** : P(PC|FIB) = P(PC|ETSETB) ≡ $\pi_1 = \pi_2$ ≡ Independència escola i ordinador

4. **Distribució estadístic**: χ^2_1

5. **Càlculs**: $\hat{X}^2 = 4.667$

6. ***P*-valor**= P($\chi^2_1 > 4.667$) = 0.0308 (punt crític = $\chi^2_{1,0.95} = 3.841$)



7. **Conclusió**: rebutgem H_0 (*P*-valor<0.05 o que 4.667>3.841)

   Aquestes dades (o més extremes) son poc probables si fos certa la independència

   Conclusió pràctica: Els Fibers prefereixen més el PC

R: `chisq.test`

NOTA: l'estadístic de Pearson ($\hat{X}^2$) tendirà a cero quan s'apropin f_{ij} i e_{ij} (és a dir, quan les dades millor reprodueixin H_0) per la qual cosa la regió crítica se situa només a la dreta, per l'efecte del "quadrat" (regió crítica **intrínsecament unilateral**).

Comp. de prop ($\pi_1 = \pi_2$) en mostres indep. Exercici

1. **Variable**:

2. **Estadístic**:

3. **Hipòtesi H_0**:

4. **Distribució estadístic**:

5. **Càlculs**:

6. ***P*-valor**:

   Punt crític:

7. **Conclusió**:

   Conclusió pràctica:

f_{ij}	noi	noia	Total
aprova	68	73	141
suspèn	32	27	59
Total	100	100	200

e_{ij}	noi	noia	Total
aprova	70.5	70.5	141
suspèn	29.5	29.5	59
Total	100	100	200

$(f_{ij} - e_{ij})^2/e_{ij}$	noi	noia	Total
aprova	0.089	0.089	
suspèn	0.212	0.212	
Total			0.601

Comp. de prop ($\pi_1 = \pi_2$). Mostres indep. i grans

- Si tenim mostres grans, podem utilitzar un estadístic alternatiu amb una altra distribució sota H_0. Sigui H_0: $\pi_1 = \pi_2$, llavors l'estadístic serà:

$$\hat{z} = \frac{(P_1 - P_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{P(1-P)}{n_1} + \frac{P(1-P)}{n_2}}} \sim N(0,1) \quad \text{on} \quad P = \frac{n_1 \cdot P_1 + n_2 \cdot P_2}{n_1 + n_2} \; ; \; P_1 = \frac{X_1}{n_1} \; ; P_2 = \frac{X_2}{n_2}$$

[Aquest test, a diferència del de l'estadístic $\hat{X}^2$, sí que pot ser bilateral i permet calcular ICs]

Exemple: Volem comparar la proporció d'aprovats segons el gènere (cont.). Disposem de dues mostres de 100 individus (prou grans)

$P_1 = 0.68$ i $P_2 = 0.73$ → $P = 0.705$ (P representa la proporció comuna)

$$SE^2 = P(1-P) \cdot (1/100 + 1/100) = 0.00416$$

$$\hat{z} = (0.68 - 0.73)/\sqrt{0.00416} = -0.775$$

`R: prop.test`

Com és bilateral → ***P-valor*** = $P(|Z| > |\hat{z}|) = 2 \cdot 0.219 = 0.438$ [amb la χ^2 donava el mateix]

$$IC(\pi_1 - \pi_2, 95\%) = (0.68 - 0.73) \mp 1.96 \cdot \sqrt{(0.68 \cdot 0.32 + 0.73 \cdot 0.27)/100} = [-0.18, 0.08]$$

NOTA: El resultat és el mateix que si ho haguèssim resolt amb X^2, amb idèntica conclusió: no hi ha evidència per dir que el gènere porta diferències en quant a la proporció d'aprovats.

Formulari de comparació de proporcions

Hipòtesi	Estadístic	Premisses	Distrib.(H_0)	Decisió α=0.05
H_0: $\pi_1 = \pi_2 = \pi$ H_1: $\pi_1 \neq \pi_2$	$\hat{z} = \frac{(P_1 - P_2)}{\sqrt{\frac{P(1-P)}{n_1} + \frac{P(1-P)}{n_2}}}$ $P = \frac{n_1 \cdot P_1 + n_2 \cdot P_2}{n_1 + n_2}$	n_1, n_2 grans m.a.s. indep	$\hat{z} \rightarrow N(0,1)$	Rebutjar si $\lvert\hat{z}\rvert >$ 1.96
	$\hat{X}^2 = \sum_{\forall ij} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$	$e_{ij} \geq 5 \ \forall$ ij m.a.s indep	$\hat{X}^2 \rightarrow \chi^2_{(I-1)(J-1)}$	Rebutjar si $\hat{X}^2 > \chi^2_{(I-1)(J-1),0.95}$
Les corresponents proves unilaterals es fan acumulant el risc α a un costat.				

Mostres Aparellades

Comp. de mitjanes ($\mu_1 = \mu_2$). Mostres aparellades

- Suposem que, per comparar 2 programes, A i B, les **dades o unitats** experimentals en que els provem **són les mateixes**.
- Ara, cada unitat ens proporciona informació sobre la diferència del rendiment de tots dos programes. Per això, definir una nova variable diferència **D** entre els dos rendiments que permet fer la **comparació "dins" de cada unitat**.
- Ara l'estudi de si els dos programes són iguals és pot fer a partir de la variable D:

$$H_0: \mu_A - \mu_B = 0 \rightarrow H_0: \mu_D = 0$$

- Per tant, l'estadístic és com el de la PH de μ per una mostra:

$$\hat{t} = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}} = (\text{si } \mu_D = 0) = \frac{\overline{D}}{S_D/\sqrt{n}} \sim t_{n-1}$$

  - $\overline{D}$ serà la mitjana de la variable diferencia D
  - S_D serà la desviació típica de D
  - μ_D serà el valor a contrastar: 0 per la igualtat de les 2 opcions

Comp. de mitjanes ($\mu_1 = \mu_2$). Mostres aparellades. Ex

En 6 bancs de dades s'ha obtingut els temps de 2 programes. Es desitja saber si B millora A; o decidir si al mercat canvien

							Mean	Variances	Var. "pooled"
A	23.05	39.06	21.72	24.47	28.56	27.58	**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 H_0: $\mu_1=\mu_2$ bilateral):

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

***P*-valor** = $P(|t_{10}| > 0.79) = 0.4494$ [R: pt(-0.79,10)+(1-pt(0.79,10)]

[Taules(t_{10}): 2*(1-0.9931)]

Punt crític: $t_{10,0.975} = 2.228$ [R: qt(0.975,10)]

[Taules(t_{10}): 2.228]

Conclusió: donat *P-valor*=0.4494 > 0.05 o bé, donat t=0.787 < 2.228, res s'oposa a acceptar H_0 (triguen el mateix). Sota H_0 aquest resultat es 'esperable'

Comp. de mitjanes ($\mu_1 = \mu_2$). Mostres aparellades. Ex

SOLUCIÓ CORRECTA: (com a dades de mostres aparellades). El mètode per dades dependents comença per calcular la diferencia A-B per cada parella:

							Mean	Variances
D = A-B	2.13	1.85	2.43	4.51	3.24	3.51	**2.946**	**0.996**

1. **Variable**: D diferència de temps
2. **Estadístic**: $\hat{t} = \frac{(\bar{D}-\mu_D)}{S_D/\sqrt{n}} = \frac{\bar{D}}{S_D/\sqrt{n}}$
3. **Hipòtesis**: $\{H_0: \mu_D = 0 \; vs. \; H_1: \mu_D \neq 0$
4. **Distribució estadístic**: $\hat{t} \sim t_{n-1}$ **Premissa**: m.a. aparellades D~N
5. **Càlculs**: $\hat{t} = \frac{(\bar{D}-\mu_D)}{S_D/\sqrt{n}} = \frac{2{,}946-0}{0.998/\sqrt{6}} = 7.229$
6. **P-valor** = P(t_5> 7.229) = 0.0008 (punt crític $t_{0.975,5}$= 2.571)
7. **Conclusió**: rebutgem H_0 ja que *P*-valor < 0.05 o que 7.229 > 2.571)

   Aquestes dades no són probables sota la igualtat. Resultats com aquest son molt poc esperables sota H_0
8. $\mathbf{IC(\mu_D, 0.95)} = 2.946 \mp 2{,}571 \cdot 0.41 = [1.89{,}4.0]$

R: `t.test(Y_A, Y_B, paired=TRUE)`

Comp. de mitjanes ($\mu_1 = \mu_2$). Mostres aparellades. Ex

Comparació entre el resultat aparellat (correcte) i per mostres independents (incorrecte):

- La **mitjana de les diferencies** (2.946) coincideix amb la diferència de les mitjanes.

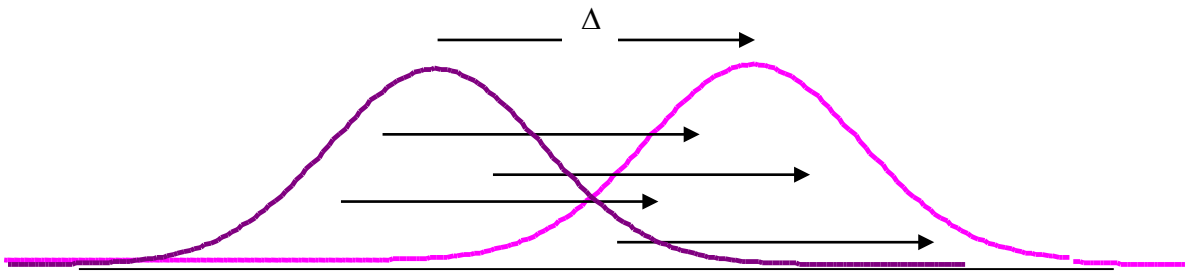
- La **variància de les diferències** (0.996) és molt inferior a la "pooled" (42.009), donat que les diferències entre les unitats (s'espera que alguns bancs de dades siguin més 'durs' que altres) han desaparegut al comparar el rendiments dels 2 programes "dins" de cada banc de dades.

- Encara que el numerador és el mateix (la senyal), el **denominador** (el soroll) **és molt inferior en la segona**. [Això és el que ha provocat els diferents resultats de les dues solucions]

- La segona prova té menys soroll: **el control de les condicions** (en aquest cas, la variabilitat dels bancs de dades) **augmenta l'eficiència** (potència)

Premissa d'efecte additiu constant

- En la prova sobre els paràmetres μ_1 i μ_2 a partir de les dues mitjanes $\bar{y}_1$ i $\bar{y}_2$ de dues mostres, al fer la diferència estem mesurant un desplaçament Δ
- Si l'efecte és lineal, **additiu**, cada cas (individu) té el mateix desplaçament (Δ)



- Així, un **efecte additiu constant** implica:
  - Idèntica forma de la distribució, i per tant: $\sigma_1^2 = \sigma_2^2$ ["***Homoscedasticitat***"]
  - Diferència que pot resumir-se en que $(\bar{y}_1 - \bar{y}_2)$ estima l'efecte en **cada individu**

Ex: Una algoritme que redueix en 2 segons el temps d'execució té un efecte additiu; però un algorisme que redueix a la meitat els temps d'execució, no ho té (de fet, té un efecte multiplicatiu)

Comprovació de premisses mostres aparellades

- Hi ha 2 premisses comprovables: **efecte additiu constant** i **Normalitat de D**
- A mostres aparellades, la **premissa d'efecte additiu constant** es pot comprovar perquè es coneixen els valors de la resposta Y en les dues condicions de la variable X
- El gràfic de Bland-Altman (BA) permet avaluar la **premissa d'efecte additiu** constant representant les diferencies de les respostes per cada individu en funció de les seves mitjanes.
- Si s'observa un efecte multiplicatiu (en comptes d'additiu), es pot solucionar traient logaritmes de la variable resposta i fent l'anàlisi igual:

$$\mathbf{D = \log(y_a) - \log(y_b) = \log(y_a / y_b)}$$

- En aquest cas, es calcularà el IC95% pel quocient fet l'exponencial (operació contraria al logaritme) de l'interval obtingut.
- En les mostres aparellades, també es requereix la **premissa de Normalitat** de la variable diferència ($D = y_a - y_b$)

Comprovació de premisses mostres aparellades. R

Siguin Y1 i Y2, les respostes obtingudes en els mateixos casos (dades aparellades).

R – Anàlisi Gràfica (3 gràfics):

```
plot(Y1,Y2,main="Y1 vs. Y2")
BlandAltman(Y1,Y2,"Diferències vs. Mitjanes")
qqnorm(Y1-Y2)
```

Aquests 3 gràfics serveixen per comprovar les premisses d'efecte additiu i Normalitat

R – Anàlisi Numèrica:

```
t.test(Y1,Y2,paired=TRUE,var.equal=TRUE)
```

R – Funció Bland-Altman:

```
BlandAltman <- function(y1,y2,tit){
   Bmean <- (y1+y2)/2
   Bdif <- y2-y1
   ymax <- max(abs(Bdif))
   plot(Bmean ,Bdif ,ylim=c(-ymax,ymax),
          xlab="Mitjanes",
          ylab="Diferències",
          main=tit,pch=19)
   abline(h=0,lty=2)
   mtext("Y2 més gran",2,line= 0.5,
          at=1.1*ymax,adj=1,cex=0.7)
   mtext("Y1 més gran",2,line= 0.5,
          at=-1.1*ymax,adj=1,cex=0.7)
}
```

El gràfic de B-A representa per a cada parella, la seva diferència (eix y) en funció de la seva mitjana (eix x). D'aquesta manera, es veu si canvia la diferència en funció de la magnitud.

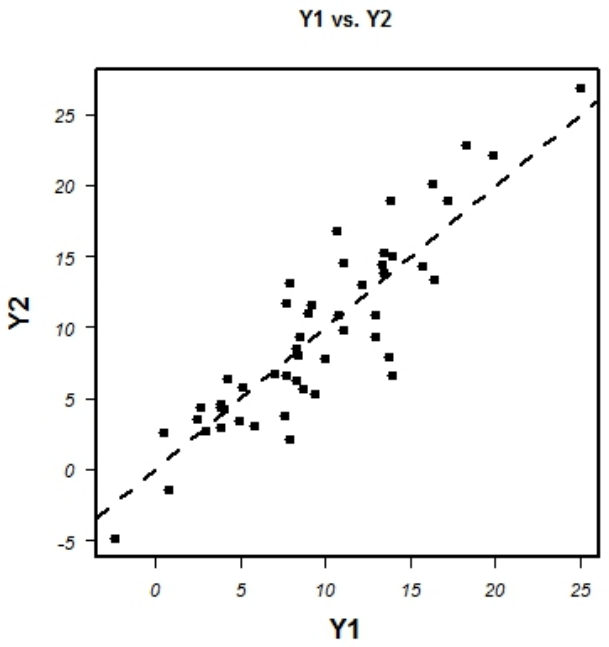
Veurem 3 situacions especials:

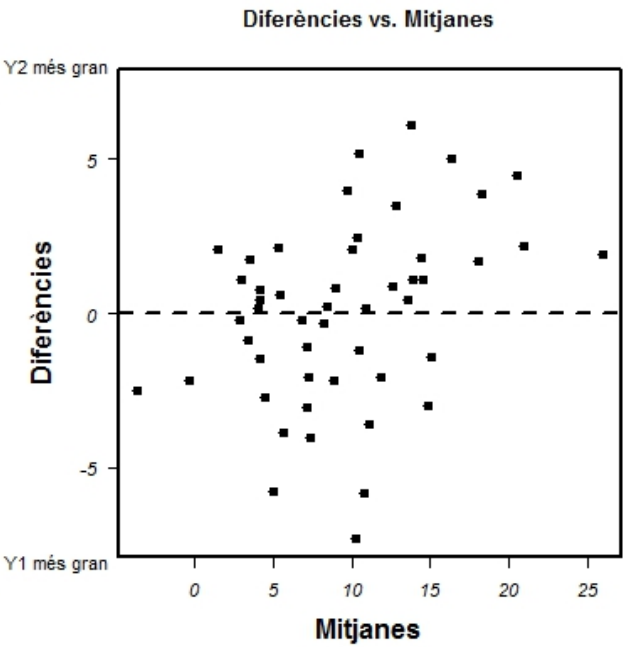
Cas 1: sense efecte lineal (additiu)

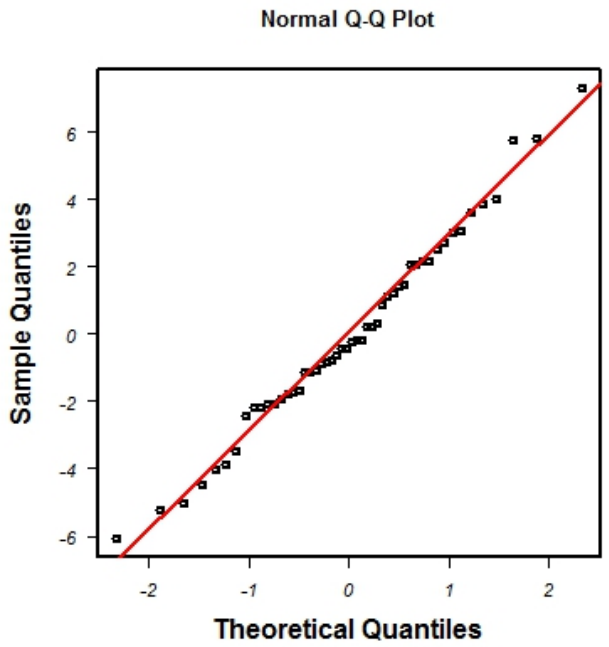
Cas 2: amb efecte multiplicatiu

Cas 3: amb efecte multiplicatiu i transformació logarítmica: `(lnY1, lnY2)`

Anàlisi de dades aparellades. No efecte







La dif. de mitjanes puntual estimada és -0.02

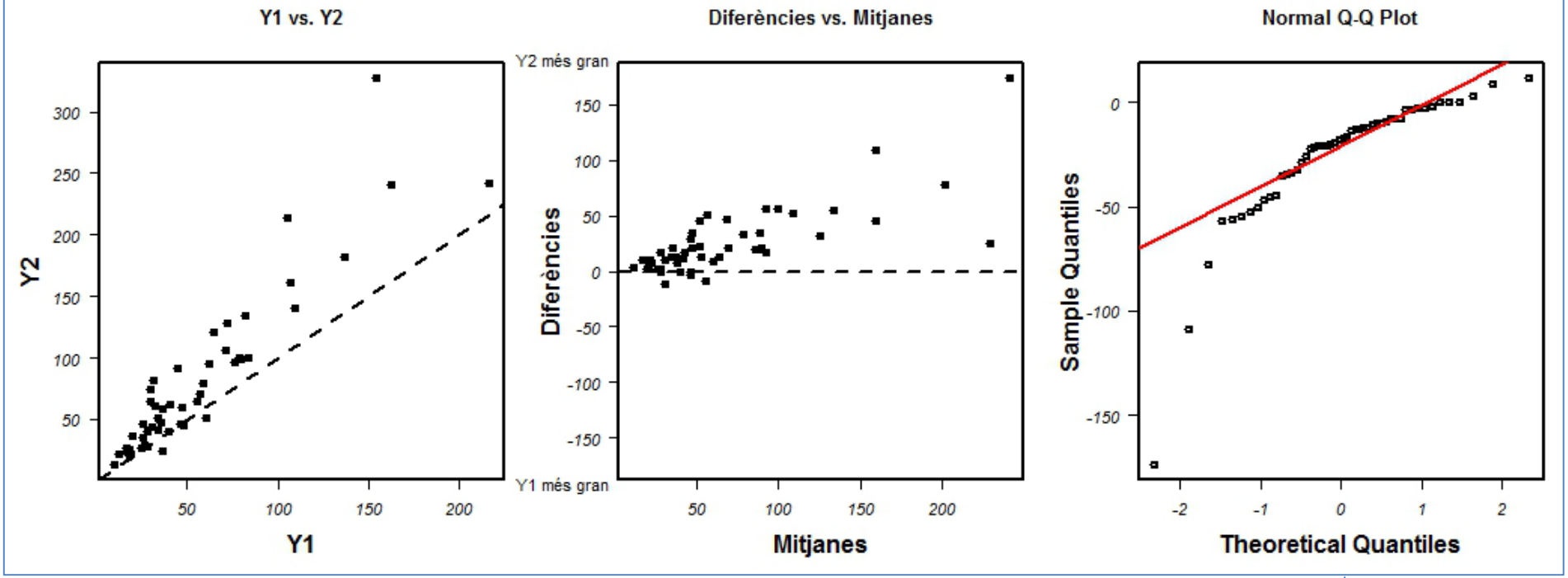
La dif. de mitjanes per interval: $IC_{95\%}(\mu_{Y2}-\mu_{Y1})$ = [-0.85 a 0.82]

És a dir, $Y_2 = Y_1 +$ [-0.85 a 0.82]

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

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

Anàlisi de dades aparellades. Efecte multiplicatiu

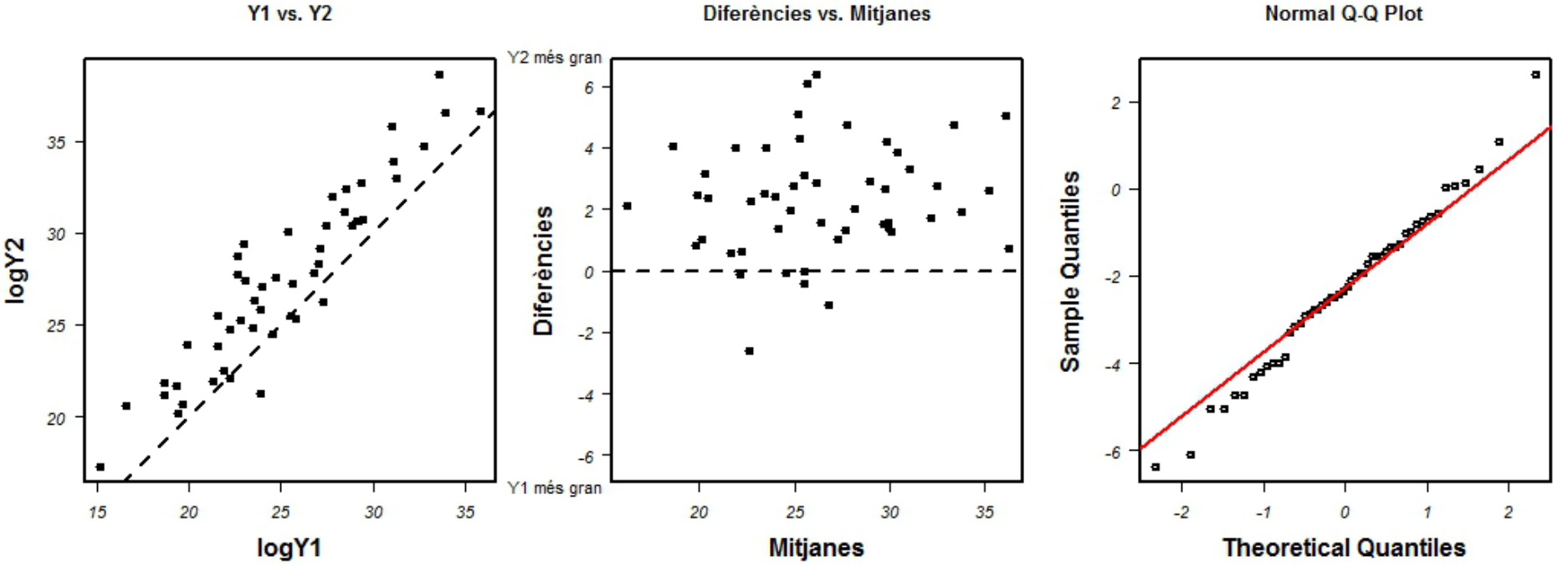


La diferència de mitjanes estimada és 25.3 amb un $IC_{95\%}(\mu_{Y2}-\mu_{Y1})$ = [16.4 a 34.2], però aquest valor no ens informa bé, ja que **l'efecte no és constant**. **Per valors grans, l'efecte és més gran** i té mes variabilitat. Si alguna transformació sobre les variables soluciona aquests problemes, la interpretació pot ser més fàcil.

Provarem solucionar-ho fent la transformació logarítmica (natural).

La distribució de les diferències NO és Normal.

Anàlisi de dades aparellades. Treure logaritmes



La diferència mitjana estimada dels logaritmes és 2.29 amb un $IC_{95\%}(\mu_{Y2}-\mu_{Y1})$ de 1.8 a 2.8. Si (Y1,Y2) són les variables originals i (Y1',Y2') són les variables log-transformades, llavors:

$$\left.\begin{matrix} Y2' = \ln(Y2) \\ Y1' = \ln(Y1) \end{matrix}\right\} \rightarrow Y2' = Y1' + 2.29 \rightarrow \ln(Y2) = \ln(Y1) + 2.29 \rightarrow Y2$$
$$= e^{\ln(Y1)+2.29} \rightarrow Y2 = e^{\ln(Y1)} \cdot e^{2.29} = 9.87 \cdot Y1$$

Interpretació: Y2 és sistemàticament 9.87 ($IC_{95\%}$ de 6.04 a 16.4) vegades més gran que Y1

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

Formulari. Proves de μ i σ en 2 mostres

Paràmetre	Hipòtesis	Estadístic	Premisses	Distrib. sota H_0	Decisió (Risc α)
μ	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$\hat{z} = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	[Y_1 $Y_2 \rightarrow$N ò n_1 i $n_2 \geq 30$] m.a.s ind. σ_1 , σ_2 coneg	$\hat{z} \rightarrow N(0,1)$	Rebutjar si $\lvert\hat{z}\rvert > Z_{1-\alpha/2}$ ($\lvert\hat{z}\rvert$>1.96 amb α=5%)
μ	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$\hat{t} = \frac{\bar{y}_1 - \bar{y}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$	Y_1 , $Y_2 \rightarrow$ N $\sigma_1 = \sigma_2$ m.a.s indep.	$\hat{t} \rightarrow t_{n_1+n_2-2}$	Rebutjar si $\lvert\hat{t}\rvert > t_{n1+n2-2,\ 1-\alpha/2}$ ($\lvert\hat{t}\rvert > t_{n1+n2-2,0.975}$ amb α=5%)
μ	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$\hat{z} = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	n_1 , $n_2 \geq 100$ m.a.s indep	$\hat{z} \rightarrow N(0,1)$	Rebutjar si $\lvert\hat{z}\rvert > Z_{1-\alpha/2}$
μ	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$\hat{t} = \frac{\bar{D} - \mu_0}{S_D\sqrt{\frac{1}{n_D}}}$	D → N m.a aparellada	$\hat{t} \rightarrow t_{n-1}$	Rebutjar si $\lvert\hat{t}\rvert > t_{n-1,\ 1-\alpha/2}$
σ	$H_0 : \sigma^2_1 = \sigma^2_2$ $H_1 : \sigma^2_1 \neq \sigma^2_2$	$\hat{F} = \frac{S_1^2}{S_2^2} \quad (S_1^2 \geq S_2^2)$	Y_1 , $Y_2 \rightarrow$ N m.a.s indep	$\hat{F} \rightarrow F_{n_1-1,n_2-1}$	Rebutjar si $\hat{F} > F_{n1-1,n2-1,\ 1-\alpha/2}$

Les corresponents proves unilaterals es fan acumulant el risc α a un costat

Tipus d'errors

Tipus d'errors en una prova d'hipòtesi. Tipus I

- Als exemples anteriors, hem expressat les conclusions com "rebutgem H_0" o "no rebutgem H_0".
- Si l'objectiu és prendre una decisió, un criteri simple seria definir a priori un llindar α per sota del qual el P-valor és vist com "petit".
- Ara, si repetim la decisió 'n' vegades, com en un repetit procés de control de qualitat, α ens donarà la freqüència d'errors determinada:

En un 100α % dels casos que rebutgem H_0, aquesta és certa.

- **Error de tipus I.** Quan utilitzem dades mostrals per posar a prova una hipòtesi sobre els paràmetres poblacionals, es pot cometre l'error de actuar com si la hipòtesi fos falsa quan no ho és realment. La probabilitat d'aquest error és podria expressar com:

$$\alpha = P(\text{concloure } H_1 \mid H_0 \text{ certa})$$

[Per procediment, aquesta prob està fixada igual a α]

Tipus d'errors en una prova d'hipòtesi. Tipus II

- La situació complementària també es pot produir.
- **Error de tipus II.** En la mateixa situació, es pot cometre l'error de no trobar evidència en contra de la hipòtesi quan realment és falsa. És a dir, no rebutjar una hipòtesi que no és certa. La probabilitat d'aquest error és pot expressar com:

$$\beta = P(\text{concloure } H_0 \mid H_1 \text{ certa}),$$

- Aquest valor, en general, no és controlable i normalment no es pot saber quant val perquè depèn del valor real del paràmetre testejat (que és desconegut).

Tipus d'error (risc)		**Decisió o Acció**	
		A_0	A_1
Realitat	H_0	**Decisió correcta**	**Error Tipus I (risc α)**
	H_1	**Error Tipus II (risc β)**	**Decisió correcta**

Tipus d'errors en una prova d'hipòtesi. Exercici

- Control de qualitat. Un processador ha de funcionar a certa velocitat μ_0 però el sistema de fabricació pot desestabilitzar-se i baixar-la a μ_1. Estudiades les conseqüències, l'equip directiu demana a l'estadístic que dissenyi un estudi al que sotmetre cada nou processador abans de instal·lar-lo i vendre-ho.
- Després de uns quants càlculs, l'estadístic:
  - posa μ_0 a H_0 i μ_1 a H_1
  - fixa $\alpha = 0.05$ i $\beta = 0.10$
  - proposa fer **'n' proves amb cada processador**
  - acceptar-ho si queda per damunt de un cert llindar L i rebutjar en cas contrari
- Quan posem en marxa l'estudi,
  1. Quina proporció de processadors correctes seran rebutjats?
  2. Quina proporció d'incorrectes arribaran al mercat?

Tipus d'errors en una prova d'hipòtesi. Exemple

- Ch és un navegador amb fama de ràpid, i la marca dominant MD no vol perdre la seva hegemonia. **Suposem que la velocitat mitjana de Ch per carregar una pàgina patró és 700 u., i la de MD és 600 u**. La desviació típica és 150 u. Fixem $\alpha = 0.025$ (unilateral)
- Si fem 10 proves independents de càrrega per cada navegador:

Ch	$\bar{y}_A = 680$	$S_A = 89$	$n = 10$
MD	$\bar{y}_B = 597$	$S_B = 147$	$n = 10$

- Com el *P*-valor resultant és 0.07, no rebutgem la hipòtesi d'igualtat i MD proclama ('testat científicament') **que el seu navegador és tan ràpid com Ch**.
- Però aquesta conclusió no és correcta: no poder rebutjar la hipòtesi nul·la (Ch és igual a MD) no implica demostrar la seva veritat. Com ja hem dit, hi ha un cert risc de que, sent diferents els dos navegadors, no puguem trobar-ne l'evidència.
- Com en aquest cas coneixem la diferència real, anem a calcular el risc 'beta' β.

Tipus d'errors en una prova d'hipòtesi. Exemple (cont)

- La clau és estudiar la distribució de l'estadístic de referència
- Sota H_0, la diferència de mitjanes mostrals és:

$$\hat{z} = \frac{\bar{y}_A - \bar{y}_B}{\sigma\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}} \sim N(0,1) \rightarrow \bar{y}_A - \bar{y}_B \sim \mathrm{N}\left(0, \sigma\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}\right) = N\left(0{,}150\sqrt{\frac{1}{10}+\frac{1}{10}}\right) = N(0{,}67.08)$$

- Definint α = 2.5% unilateral, la regió crítica es troba per diferències de les mitjanes mostrals més grans que 1.96·67.08 = 131.48 u.
- En realitat, la diferència entre mitjanes és de 100. Com les mostres provenen d'aquesta situació, comprovem com de probable és que NO puguem rebutjar H_0:

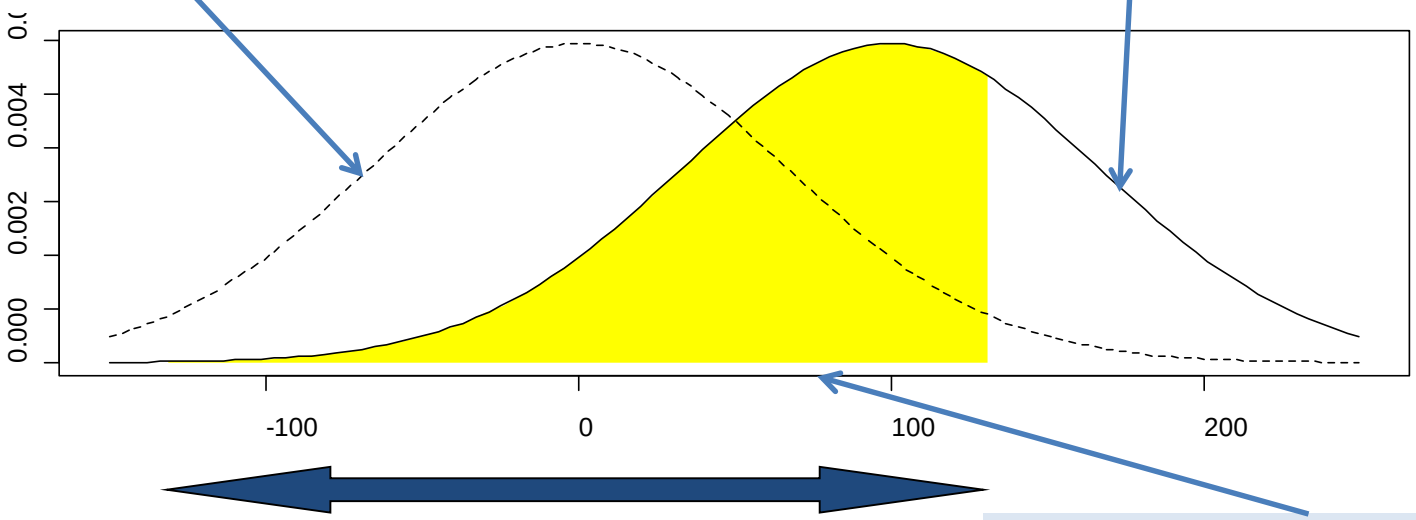
P("Error tipus II") = P($\bar{y}_A - \bar{y}_B$ < 131.48 | H_1) = P(Z < (131.48-100)/67.08) = P(Z<0.47) = 0.68

- Veiem que MD tenia molt fàcil resoldre la prova a la seva conveniència: era molt probable no trobar cap diferència significativa. La prova és **poc potent** (**potència = 1–β**).

Tipus d'errors en una prova d'hipòtesi. Exemple (cont)

Distribució "ingènua" per a la diferència de mitjanes mostrals (noteu que la campana té esperança zero)

Distribució real per a la diferència de mitjanes mostrals ($\bar{y}_A - \bar{y}_B$). Els valors típics estaran al voltant de 100, que és la diferència autèntica de les mitjanes poblacionals ($\mu_A - \mu_B$)



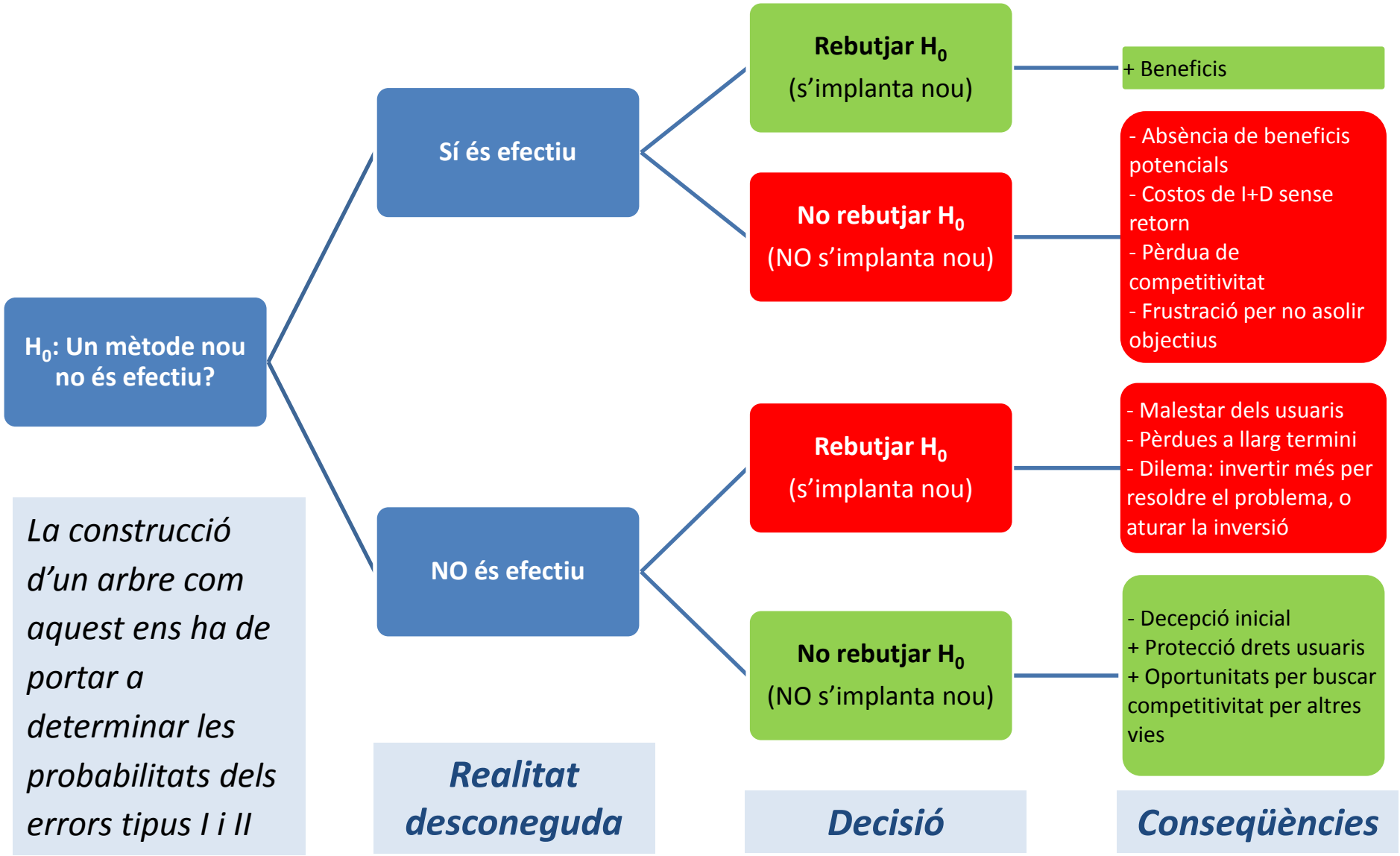
Regió d'"*acceptació*"
(no es pot rebutjar la hipòtesi nul·la)

Diferència observada:
680 – 597 = 83 u.

Tipus d'errors en una prova d'hipòtesi. Exemple (Cont)

- En el nostre exemple, podem estirar de les orelles als de MD per:
  - La conclusió ha estat incorrectament exposada
  - El disseny és defectuós: li manca potència (la *n* és petita); o, més aviat, si volien demostrar equivalència, havien d'haver plantejat un altre tipus d'estudi (que no veurem en aquest curs)
- En conseqüència, l'experimentador ha d'assumir que el seu estudi està exposat a diferents perills:
  - Mostres no aleatòries (els individus no són independents entre sí)
  - Assignació de X no aleatòria (individus no similars entre els grups)
  - Variables amagades que pertorben la resposta observada
- I estar disposat a posar mesures per evitar errors com aquests. A més a més, ha de saber que l'anàlisi d'un estudi estadístic no és una demostració matemàtica i, com a mínim, la conclusió ha de ser prudent.

To do or not to do. Arbre de decisions i conseqüències



La construcció d'un arbre com aquest ens ha de portar a determinar les probabilitats dels errors tipus I i II

Annexes

Annexe I: Comparació $\pi_1 = \pi_2$. Mostres aparellades

Al igual que en el cas independent, la variable que observem "per parelles" és una dicotomia. Per tant, la informació la posarem en forma de taula.

Per exemple, si s'ha preguntat a 100 "fibers": 1. T'agrada el PC?: molt/poc

2. T'agrada el MAC?: molt/poc

freqüència observada: f_{ij}	PC - Molt	PC - Poc	Total
MAC - Molt	61	4	65
MAC - Poc	16	19	35
Total	77	23	N = 100

Davant d'aquesta taula podem realitzar dues preguntes:

1. $P(Molt_{PC}|Molt_{MAC}) = P(Molt_{PC}|Poc_{MAC})$ [Que m'agradi el PC és independent que m'agradi el MAC? o Hi ha un caràcter comú (efecte cas) que guia ambdues respostes?]
2. $P(Molt_{PC}) = P(Molt_{MAC})$ [Existeix preferència per una màquina?]

Annexe I: Comparació $\pi_1 = \pi_2$. Mostres aparellades

- Per a contestar la **primera pregunta** podríem utilitzar l'estadístic de χ^2(Chi Quadrat) previ ja que es tracta de veure si la proporció de Molt_{PC} és la mateixa en les categories Molt_{MAC} i Poc_{MAC}. El resultat és:

$$X^2 = 50.242 > 3.841 = \chi^2_{1,0.95} \qquad o \qquad P - valor = P(\chi^2_1 > 50.242) < 0.0001$$

- Conclusió: Hi ha un "efecte cas":n'hi ha casos favorables a totes les màquines, i n'hi ha de negatius a totes

- Però la pregunta d'interès és la segona, per a la que volem comparar

$$P(\text{Molt}_{PC}) = 0.77 \quad \text{amb} \quad P(\text{Molt}_{MAC}) = 0.65$$

- Resulta que aquestes estimacions no són independents, ja que venen dels mateixos casos i la variància de la seva resta no podria ser la suma de les seves variances. [Penseu que la casella "1,1" contribueix a ambdues estimacions.]
- En realitat, comparar els efectius marginals 77 amb 65 equival a comparar els efectius 16 amb 4 que es troben fora de la diagonal havent suprimit els 61 casos comuns.
- Així, podem reestructurar la taula reunint la informació rellevant (següent diapositiva)

Annexe I: Comparació $\pi_1 = \pi_2$. Mostres aparellades

	f_k
Prefereixen PC	16
Prefereixen MAC	4
Total	20

- De ser certa H_0: $P(Molt_{PC}) = P(Molt_{MAC})$ → H_0: P(prefereixen PC) = P(prefereixen MAC) = 0.5
- I podríem realitzar la prova ja coneguda sent a = casos favorables a PC i b = casos favorables a MAC. Llavors els efectius esperats serien e = (a+b)/2
- En el nostre cas, l'estadístic valdrà:

$$\hat{X}^2 = \sum_{k=1}^{2} \frac{(f_k - e_k)^2}{e_k} = \sum_{k=1}^{2} \frac{\left(f_k - \frac{a+b}{2}\right)^2}{(a+b)/2} = \frac{(a-b)^2}{(a+b)} = \frac{(16-4)^2}{(16+4)} = 7.2$$

- Ja que el p-valor és $P(\chi_1^2 > 7.2) = 0.0073 < \alpha$ o bé que $X^2 = 7.2 > 3.84 = \chi_{1,0.95}$ podem rebutjar H_0
- **Conclusió pràctica**: Hi ha preferència pel PC.

Annexe II: Grandària Mostral. Cas I: Estimar μ

Objectiu: Estimar μ amb un IC95% d'una amplada determinada (A)

$$A = LS_{IC95\%} - LI_{IC95\%} = \left[\bar{y} + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right] - \left[\bar{y} - Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right] = 2 \cdot Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \rightarrow$$

$$\rightarrow n = \left(\frac{2 \cdot Z_{1-\frac{\alpha}{2}} \cdot \sigma}{A}\right)^2 = \left(\frac{2 \cdot Z_{1-\frac{\alpha}{2}}}{\frac{A}{\sigma}}\right)^2 \approx \frac{16}{\left(\frac{A}{\sigma}\right)^2}$$

Exemple: si volem estimar la mitjana del temps de posta en marxa d'un equip amb una incertesa total *A* (amplada de l'IC) que sigui la meitat de la desviació tipus, $A/\sigma = 0.5$, implica ***n=64 observacions***.

Annexe II: Grandària Mostral. Cas II: Estimar μ_1-μ_2

Objectiu: Estimar la diferència μ_1-μ_2 amb un IC95% d'una amplada determinada (A)

$$A = LS_{IC95\%} - LI_{IC95\%} = \left[\bar{y}_1 - \bar{y}_2 + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{2}\sigma}{\sqrt{n}}\right] - \left[\bar{y}_1 - \bar{y}_2 - Z_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{2}\sigma}{\sqrt{n}}\right] =$$

$$= 2 \cdot Z_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{2}\sigma}{\sqrt{n}} \rightarrow n = \left(\frac{2\sqrt{2} \cdot Z_{1-\frac{\alpha}{2}} \cdot \sigma}{A}\right)^2 = \left(\frac{2\sqrt{2} \cdot Z_{1-\frac{\alpha}{2}}}{\frac{A}{\sigma}}\right)^2 \approx \frac{32}{\left(\frac{A}{\sigma}\right)^2}$$

Exemple: si volem estimar la diferència de mitjanes del temps de posta en marxa de dos models d'equip amb una incertesa total *A* (amplada de l'IC) que sigui la meitat de la desviació tipus comuna , $A/\sigma = 0.5$, implica *n=128 observacions*.

Annexe II: Grandària Mostral. Cas III: Contrastar $\mu_1 = \mu_2$(independ.)

Objectiu: Contrastar $\mu_1 = \mu_2$ en 2 mostres independents

$$\begin{cases} H_0: \mu_1 - \mu_2 = 0 \\ H_1: \mu_1 - \mu_2 = \Delta \end{cases}$$

$$V(\bar{y}_1 - \bar{y}_2) = V(\bar{y}_1) + V(\bar{y}_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = (si\ \ n_1 = n_2 = n) = \frac{2\sigma^2}{n}$$

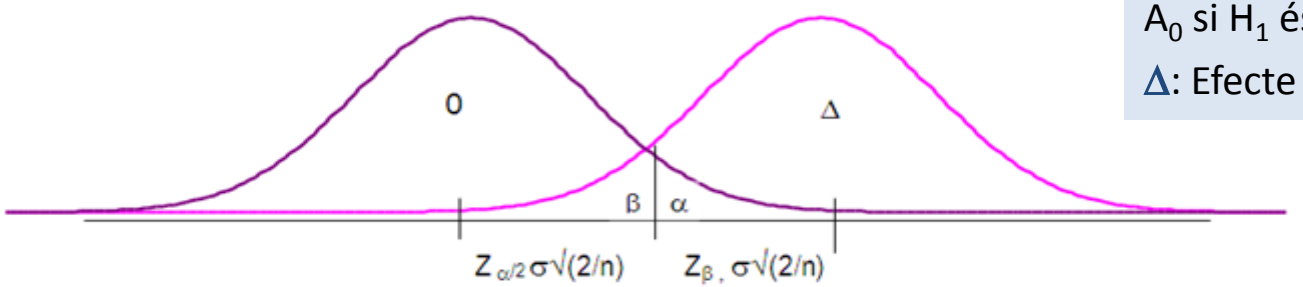
Elements:

σ: Variabilitat del fenomen en estudi

α: Freqüència admesa de A_1 si H_0 és certa

β: Freqüència admesa de A_0 si H_1 és certa

Δ: Efecte de l' intervenció



$$Z_{\alpha/2}\ \sigma\sqrt{(2/n)} + Z_{\beta}\sigma\sqrt{(2/n)} = \Delta$$

Per tan la grandària de cada grup és: $\boldsymbol{n = \frac{2\sigma^2\left(Z_{1-\frac{\alpha}{2}} + |Z_{\beta}|\right)^2}{\Delta^2}}$

Exemple: Quina grandària mostral seria necessària per detectar una diferència de 10 cm en la alçada mitjana d'homes i dones? ($\alpha = 0.05$; $\beta = 1–0.8$ i $\sigma = 8$ cm)

$$n = (2 \cdot 8^2(1.96 + 0.84)^2)/10^2 \approx 10$$

Annexe II: Grandària Mostral. Cas IV: Contrastar $\pi_1 = \pi_2$(independ.)

Objectiu: Contrastar $\pi_1 = \pi_2$ en 2 mostres independents

$$\begin{cases} H_0: \pi_1 - \pi_2 = 0 \\ H_1: \pi_1 - \pi_2 = \Delta \end{cases}$$

Apliquem la següent fòrmula per saber la grandària per grup:

$$n = \frac{4}{\Delta^2}$$

Exemple: Si volem demostrar que una nova versió deixa molt satisfets a un 20% més de clients (posem: π_1=0.5 i π_2=0.7) aleshores $(\pi_1-\pi_2)^2$ = 0.04 que implica n=100 per grup.

$$n = (2.8^2(1.96 + 0.84)^2)/10^2 \approx 10$$

Annexe II: Grandària Mostral. Exercici

Pràctica de B7: Uns companys han comparat els temps de descàrrega de 2 navegadors i la seva pràctica ha anat tan bé que un altre professor els hi suggereix fer un disseny nou per publicar un article a una revista de informàtica.

Amb les dades de la seva pràctica han observat una desviació tipus S igual a 54 i una diferència de temps igual a 22. Però, per 'garantir el tir' el professor els hi suggereix treballar amb un $\sigma = 60$.

Quines grandàries mostrals necessitem en les següents situacions?

1) Volen estimar $\mu_1 - \mu_2$ amb una amplitud d'interval, A = 40?

$n = 32 / (A/\sigma)^2 = 72$

2) Volen contrastar una Δ=20 amb α=0.025 i β=0.20 unilaterals.

$$n = \frac{2\sigma^2\left(Z_{1-\frac{\alpha}{2}} + |Z_\beta|\right)^2}{\Delta^2} = \frac{2 \cdot 60^2 \cdot (1.96+0.84)^2}{20\text{^}2} \approx 141.47 \rightarrow n = 142$$

Annexe II: Grandària Mostral. Potència sense augmentar n

Podem augmentar la potencia sense augmentar la n?

Ho aconsseguirem si trobem sistemes per baixar la variabilitat **s^2** de la resposta. Aquesta variabilitat té 2 components:

$S^2 = S_B^2 + S_W^2$ - S_B^2: variabilitat **entre** (***B**etween*) casos

- S_W^2: variabilitat **intra** (***W**ithin*) casos

[Ex: la variabilitat de la variable "pes" ve donada pel fet de que les persones tenim constitucions diferents (S_B^2) i cada persona pot pesar lleugerament diferent depenent del moment del dia (S_W^2)]

La variància de la resposta la podem disminuir amb diferents estratègies:

1. Utilitzant un disseny aparellat, que només tingui $S_W^2 \rightarrow S^2 = 2 \cdot S_W^2$
2. Definint el canvi **C** respecte al valor inicial (*baseline*) $\rightarrow S^2 = 2 \cdot S_W^2$
3. Fent promitjos de K repeticions independents: $\rightarrow S^2 = S_B^2 + S_W^2/K$

Aquestes estratègies es complementen, ja que controlen dos fonts diferents de variabilitat.

Previsió

Bloc 6 – Probabilitat i Estadística

Maig 2016

Índex

1. Previsió i disseny d'experiments
2. Fases del procés de models estadístics
3. Model "quantitativa vs quantitativa": model, paràmetres i interpretació
4. Estimadors dels paràmetres: distribució, inferència
5. Anàlisi de les premisses. Anàlisi de residus
6. Predicció
7. Model "quantitativa vs categòrica":
   a. model, paràmetres i interpretació
   b. descomposició de la variabilitat

Inferència estadística. Guió

Guió de la part d'Estadística de PE:

- B4: Tècnica general de la inferència [estadística]
  - estimar un paràmetre (*Intervals de Confiança*)
  - refutar un paràmetre (*Proves d'Hipòtesis*)
- B5: Aplicació (I): Avaluació de millores
  - *Disseny d'experiments:* comparació de dues poblacions.
- **B6: Aplicació (II): Predicció**
  - ***Previsió* d'una var. resposta, en funció d'una var. explicativa.**

Previsió i disseny d'experiments

- Al B5 parlem de variables i condicions; i es defineix el disseny d'experiments com: "estimar l'efecte causal de la intervenció X en la resposta Y donades les condicions Z"
  - La resposta Y ha de mesurar el nostre objectiu
  - La intervenció X és el nostre potencial per canviar el futur
  - Les condicions Z 'predeterminen' el futur i permeten anticipar Y

- **Tipus d'estudis: *VEURE*** enfront de ***FER***:
  - Estudis observacionals: ***veiem*** i podem fer previsions, predir, anticipar,...Els individus arriben amb el valor de Z, que relacionem amb la resposta Y [Ex: comparem les notes de PE (Y) en funció del gènere (Z)]
  - Estudis experimentals: ***fem*** i podem intervenir, canviar el futur. Observem l'efecte en Y havent assignat X a les unitats [Ex: comparem les notes de PE (Y) en un experiment on a uns alumnes se'ls ha assignat emprar e-status i els altres no (X)]

Nota 1: La clau per intervenir és ser 'propietaris' de la variable X

Nota 2: El passat (Z) ens esclavitza, el futur (X) ens allibera

Previsió i disseny d'experiments

Per respondre una pregunta '**causal**' sobre una condició Z, hem de pensar un experiment on '**assignar**' aquesta condició Z.

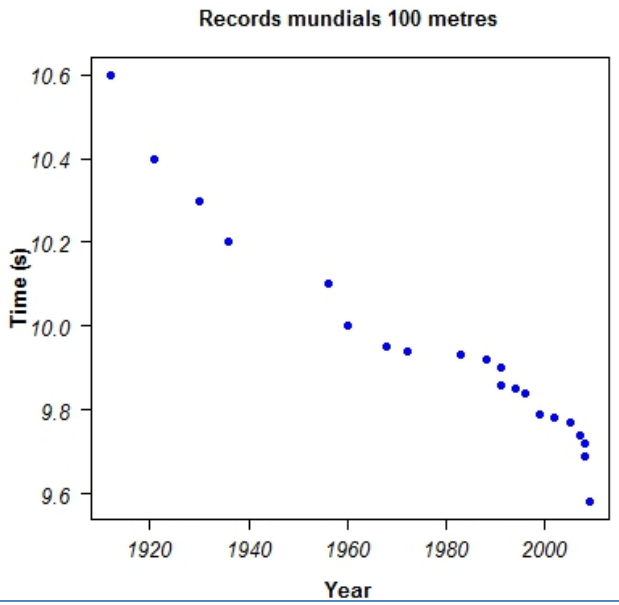
Exemple: per respondre si hi ha discriminació per gènere, podem 'assignar' a l'atzar un nom i una foto de dona/home a uns currículums i preguntar quin salari els hi pagarien. [Això permet deixar fixes o iguals ('controlar') totes les altres variables: experiència, dedicació, formació,...]. Així podríem estimar l'efecte de ser dona/home en el salari. Però, en el futur, no podrem 'assignar' el gènere a un ciutadà.

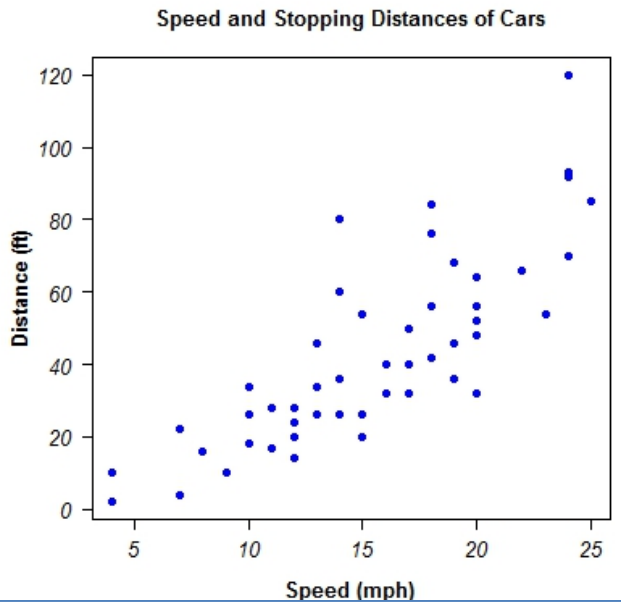
Resum:

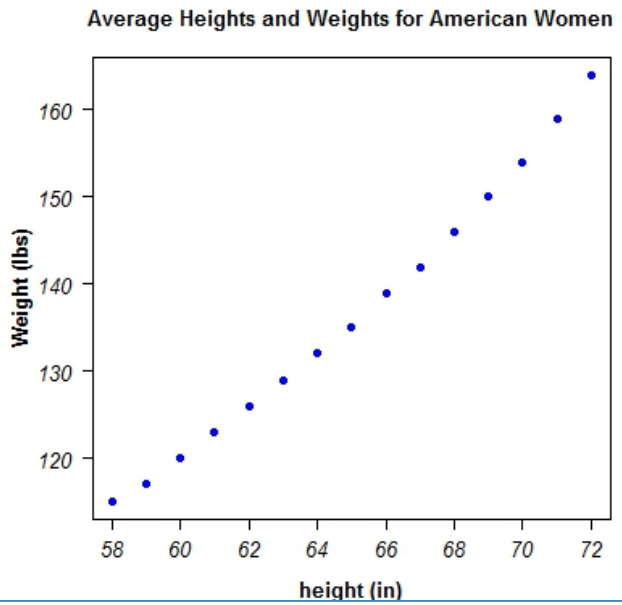
1) Un **experiment** amb **assignació** a l'atzar permet estimar '**efectes**' havent controlat totes les altres variables.

2) Convé valorar la possibilitat **d'assignar** en el futur per saber si podem utilitzar la relació només per **predir** o també per **intervenir**

Previsió i disseny d'experiments







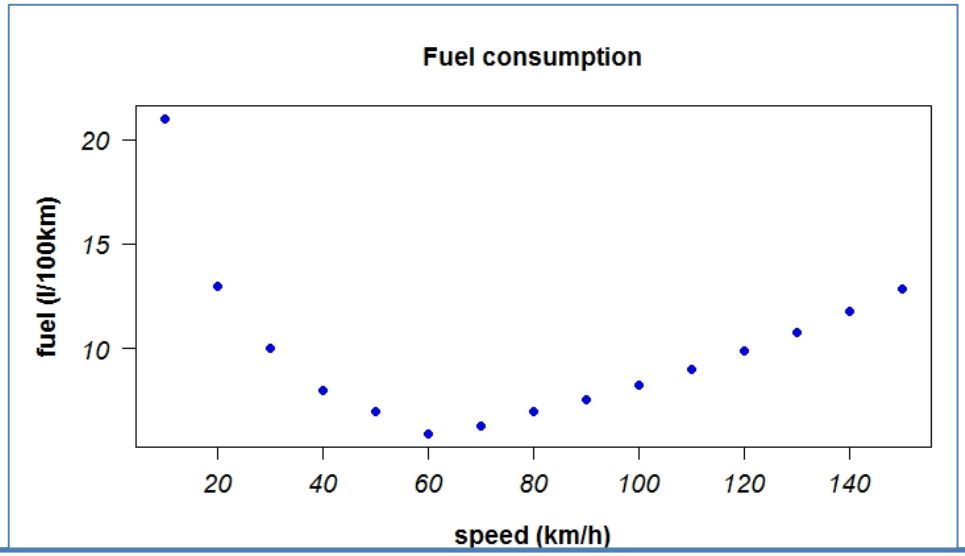
- Temps dels records mundials de 100 metres masculí: 1912-2012. Sempre serà decreixent. Previsió o Intervenció?
- Distància de frenada en funció de la velocitat. Previsió o Intervenció?
- Pes en funció d'alçada en dones. Previsió o Intervenció?

Model quantitativa vs. quantitativa. Exemple I

- 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.

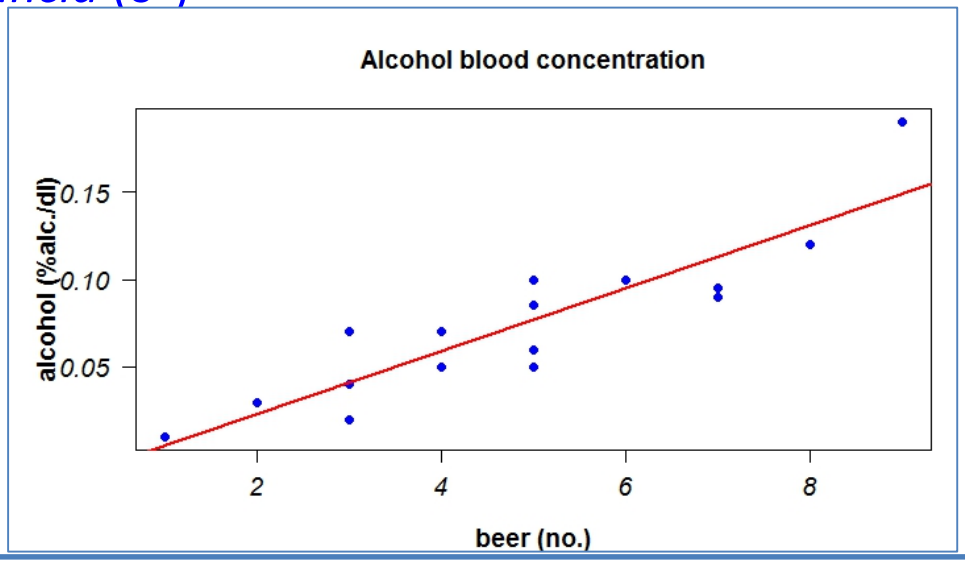


speed (km/h)	fuel (l/100 km)
10	21
20	13
30	10
40	8
50	7
60	5.9
70	6.3
80	6.95
90	7.57
100	8.27
110	9.03
120	9.87
130	10.79
140	11.77
150	12.83

Model quantitativa vs. quantitativa. Exemple II

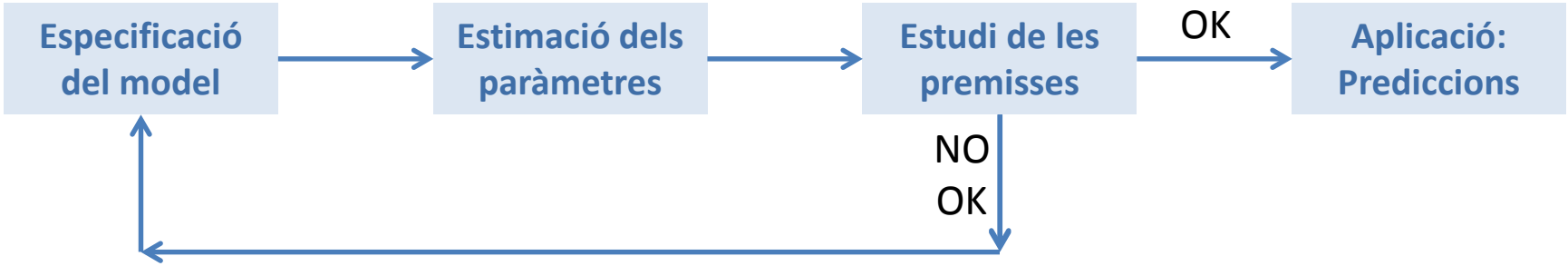
- Un estudi ha sol·licitat a 16 voluntaris que es prengui una quantitat determinada (aleatòriament) de cervesa, mesurada en llaunes, i es mesura l'alcohol a la sang trenta minuts després [%alc. /dl sang].
- Un model simple és ajustar-hi una recta, que implica dos paràmetres: *pendent* (β_1) i *constant* (β_0) a l'origen
- Al voltant tenim una certa dispersió que requereix un tercer paràmetre: la *variància* (σ^2)

Source: The Basic Practice of Statistics. 4th ed. David S. Moore. Example 24.7



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

Fases en el disseny d'un model



Source: Capítol 7 d'*Estadística per a enginyers informàtics*. Ed UPC

- Un cop especificat el model i estimats els paràmetres, perquè sigui útil (aplicar-lo i fer prediccions), cal estudiar les premisses assumides. Serà suficient una anàlisi exploratòria per confirmar que són "raonables"
- Si durant el procés de modelar, no s'aconsegueix trobar els resultats desitjats, pot ser que el model sigui millorable. En aquest cas, podem procedir a realitzar **transformacions** (ln(X), ln(Y), 1/Y, Y/X, arrels, potències,...) o buscar **altres variables predictores**

Model quantitativa vs. quantitativa. Paràmetres

- Model: $Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$ **sent** $\varepsilon_i \sim N(0, \sigma)$

  Y_i valor de la variable resposta Y en el cas i-èsim

  X_i valor que pren la condició X en el cas i-èsim

  ε_i error aleatori o distància a la recta del cas i-èsim

- Els paràmetres seran: β_0 com a **constant** a l'origen, β_1 com a **pendent** de la recta i i σ^2 com la **variància dels** ε_i o variància residual ($\beta_0+\beta_1 \cdot X_i$ serà la part determinista de Y; i ε_i serà la part aleatòria de Y)

EXEMPLE: (*Estadística per a enginyers informàtics*. Ed UPC pàg 141)**.** Homes adults i sans de Barcelona: *Y* és Pes en Kg; *X* és Alçada en cm. Suposem com a model una recta amb paràmetres:

β_0 = -100 Kg β_1 = +1 Kg/cm σ = 6Kg

Quin pes correspon a un senyor de 160 cm? 60Kg

I a un de 180 cm? 80 kg

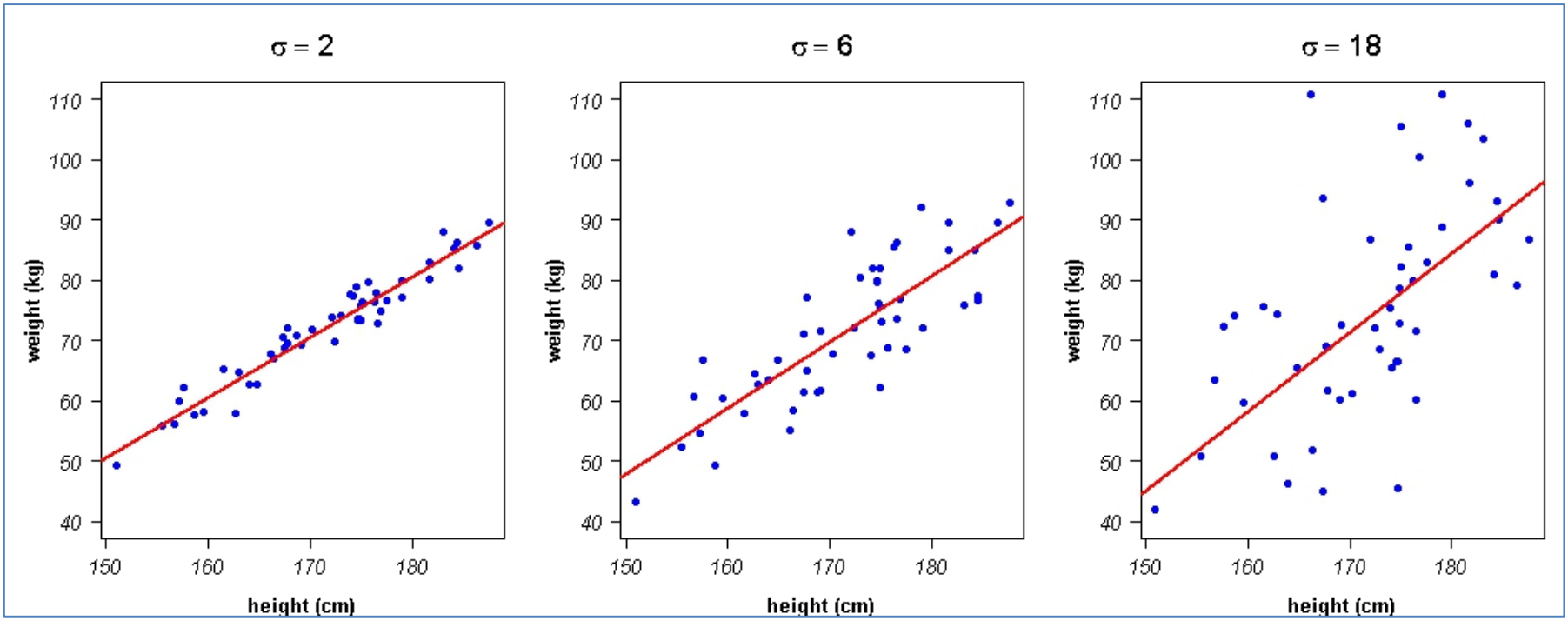
Què significa "correspon"? 'Esperat, en mitjana'

Què significa σ = 6Kg ? *Separar-se uns 6 kg del pes esperat és habitual. (Rarament és més del doble)*

Què opines de l'etiqueta 'pes ideal' en algunes farmàcies? Que ignoren la variabilitat natural

Model i Paràmetres

- El **paràmetre** més important per un estadístic és la variància σ^2 (encara que σ és més fàcil d'interpretar).
- Diferents valors de **σ** condicionaran la forma del núvol de punts



- Noms possibles per **ε**: *negatius* → error, residu, pertorbació

  positius → idiosincràsia

Estimació dels paràmetres

- β_0 , β_1 i σ^2 són valors poblacionals, *autèntics*, desconeguts, a 'estimar'. L'estimació dels dos primers, dóna lloc a la recta estimada:

$$\hat{y}_i = b_0 + b_1 \cdot X_i$$

- Aquesta permet fer prediccions per a cada observació amb el seu error de predicció:

$$e_i = y_i - \hat{y}_i \quad \text{[els } e_i \text{ són els residus del model]}$$

- L'estimació mínim quadràtica consisteix en calcular els estimadors $\mathbf{b_0}$ i $\mathbf{b_1}$ de β_0 i β_1, minimitzant la suma dels errors de predicció al quadrat: $\sum(e_i)^2 = \sum(y_i - \hat{y}_i)^2 = \sum(y_i - b_0 - b_1 x_i)^2$ [*annex 6.12 d'Estadística per a enginyers informàtics. Ed UPC*]
- La solució al problema de minimització és el següent: [*Ref: Eei.Ed.UPC pg144*]

$$\hat{\beta}_1 = b_1 = \frac{S_{XY}}{S_X^2} = r \cdot \frac{S_Y}{S_X}$$

$$\hat{\beta}_0 = b_0 = \bar{Y} - b_1 \bar{X}$$

$$\hat{\sigma}^2 = S^2 = \frac{\sum(e_i^2)}{n-2} = \frac{(n-1)S_Y^2(1-r^2)}{n-2} = \frac{(n-1)(S_Y^2 - b_1 S_{XY})}{n-2}$$

[Recordeu que $\mathbf{S_{XY}}$ és la covariància mostral, i $\mathbf{r}=\mathbf{r_{XY}}$ la correlació mostral]

Estimació dels paràmetres. Exemple

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

Recordatori:

$$\bar{y} = \frac{\sum y_i}{n} \qquad s_Y^2 = \frac{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}{n-1} \qquad s_{XY} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{n-1}$$

Càlculs dels estadístics convencionals:

$$\bar{y} = 0.07375 \qquad s_Y^2 = 0.0019483 \qquad s_{XY} = 0.08675$$

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

Resultats de la regressió:

$$b_1 = \frac{s_{XY}}{s_X^2} = r_{XY} \cdot \frac{s_Y}{s_X} = 0.01796$$

$$b_0 = \bar{Y} - b_1 \bar{X} = -0.0127$$

$$S = \sqrt{\frac{\sum(e_i^2)}{n-2}} = 0.0204$$

Model amb R:

```
> lm(alc ~ n.cerv)
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
```

Interpretació dels paràmetres

- Els **paràmetres** de la recta han de ser interpretats d'acord amb les seves unitats.
- El **pendent** s'interpreta directament com a tal:
  - Experiments: La resposta Y tindrà un canvi esperat de β_1 (unitats de Y) per cada increment de 1 unitat fet en la causa X.
  - Previsió: Una variació de 1 unitat en la variable X s'associa amb una variació de β_1 unitats en la variable Y.
- La **variància residual** s'interpreta:
  - Experiments: Variabilitat de la variable Y.
  - Previsió: Error de predicció de la variable Y, conegut el valor de X.
- La **constant**, en certs casos, es pot interpretar com el valor que pren la resposta en absència de la variable predictora. [La constant és necessària per construir el model, però secundària en sí mateixa]

Interpretació dels paràmetres. Exemple

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?

Interpretació dels paràmetres. Exemple (cont)

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

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

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

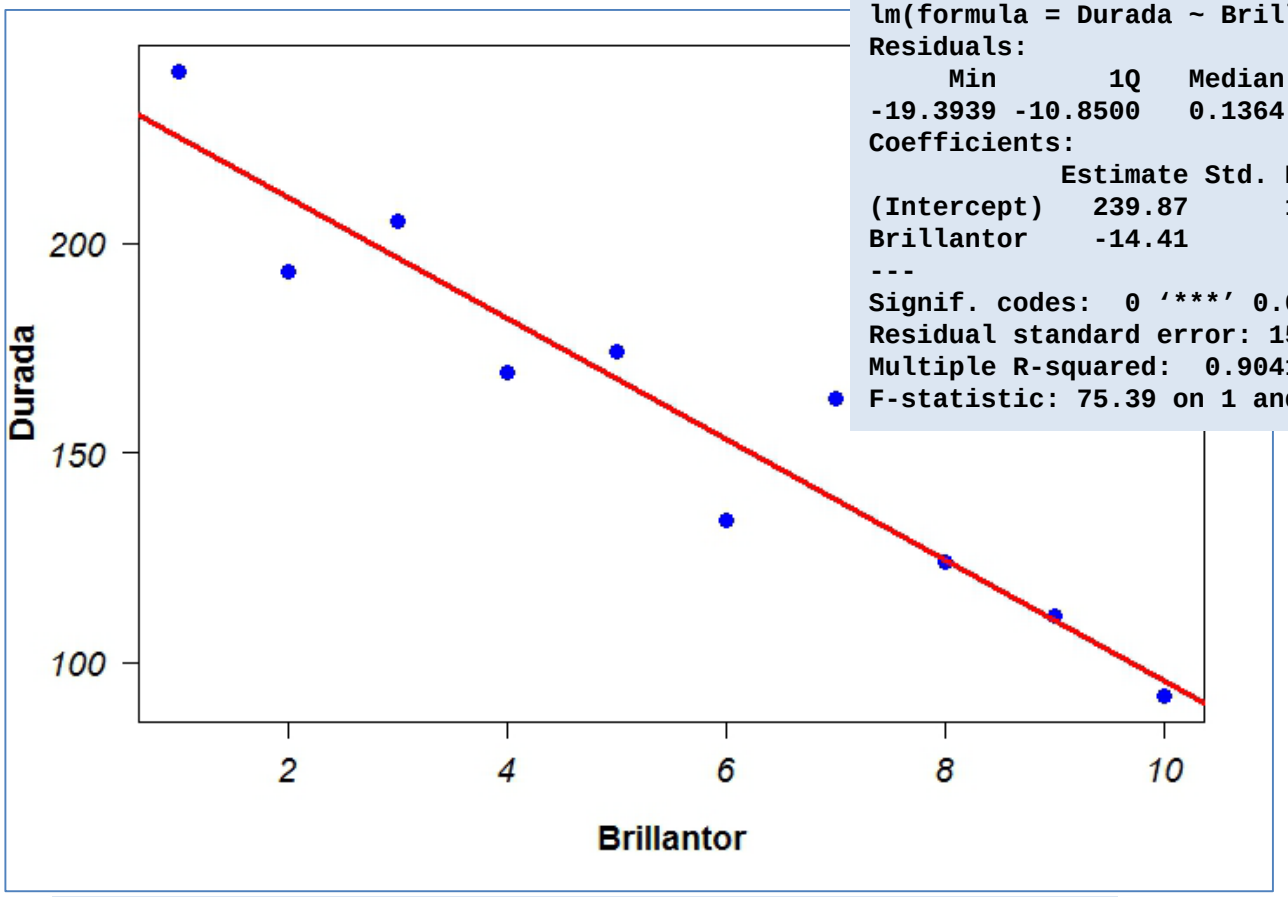
Interpretació de b_1: Per cada grau de brillantor que augmentem, la bateria dura uns 14.4 minuts menys.

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

Interpretació dels paràmetres. Exemple (cont)

```
> datos <- read.table("clipboard",header=TRUE)
> mod.lm <- lm(Durada~Brillantor,datos)
> summary(mod.lm)
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
```




```
> par(cex.lab=1.2,cex.axis=1.2,las=1,font.lab=2,font.axis=3)
> plot(Durada~Brillantor,datos,pch=19,col=4,cex=1.2)
> abline(mod.lm,col=2,lwd=3)
```

Distribució dels estimadors (mínims quadrats)

- b_1 és una combinació lineal de normals i, per tant, continuarà seguint una distribució Normal. Així, la distribució de l'estimador b_1 és:

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{(n-1)S_X^2}\right)$$

- b_0 també és una combinació lineal de normals. Així, la distribució de l'estimador b_0 és:

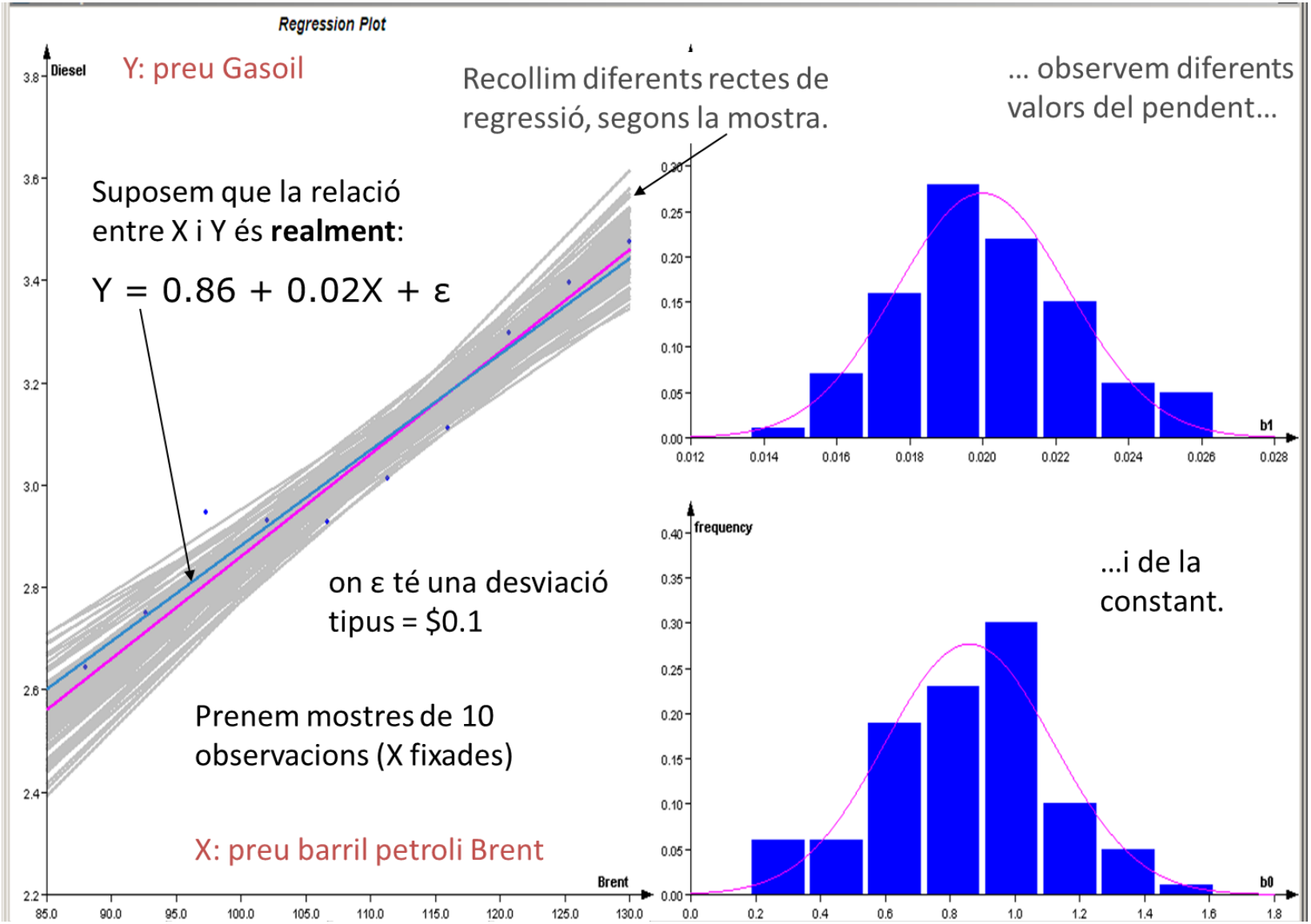
$$b_0 \sim N\left(\beta_0, \sigma^2\left(\frac{\bar{X}^2}{(n-1)S_X^2}\right)\right)$$

- $S^2 = \frac{\sum(e_i^2)}{n-2}$ és estimador no esbiaixat de σ^2, i coneixem que $\frac{\sum(e_i^2)}{\sigma^2} \sim \chi^2_{n-2}$. Així, la distribució de referència de la variància residual és:

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$$

Feu clic aquí per veure apps per entendre distribució dels estimadors

Distribució dels estimadors. Simulació



Distribució dels estimadors. Disminuir S_{b1}

Si es vol millorar una recollida de dades on s'ha observat un error estàndard massa gran en l'estimació b_1 de β_1 :

$$S_{b_1} = \sqrt{\frac{S^2}{(n-1)S_X^2}}$$

¿què es pot fer?

Solucions per disminuir S_{b_1} :

- Intentar 'controlar' les fonts de variació en S^2
- Augmentar 'n'
- Ampliar la 'finestra' de l'estudi per augmentar $\sum(x_i - \bar{x})^2$

Distribució dels estimadors. Inferència

Es pot realitzar la inferència habitual amb PH per β_0, β_1 i σ^2:

- **Prova d'hipòtesi sobre el pendent:**

$$\begin{cases} H_0: \beta_1 = \beta_1' \\ H_1: \beta_1 \neq \beta_1' \end{cases} \rightarrow \frac{b_1 - \beta_1'}{S_{b_1}} = \frac{b_1 - \beta_1'}{\sqrt{\frac{S^2}{(n-1)S_x^2}}} \sim t_{n-2} \quad \text{sent} \quad \hat{\beta}_1 = b_1 = \frac{S_{XY}}{S_X^2}$$

No ens perdem amb la notació!!

$\boldsymbol{\beta_1}$: Paràmetre que volem contrastar

$\boldsymbol{\beta_1'}$: Valor a contrastar

$\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}$ o $\boldsymbol{b_1}$: estimador del paràmetre

- **Prova d'hipòtesi sobre el terme independent:**

$$\begin{cases} H_0: \beta_0 = \beta_0' \\ H_1: \beta_0 \neq \beta_0' \end{cases} \rightarrow \frac{b_0 - \beta_0'}{S_{b_0}} = \frac{b_0 - \beta_0'}{\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_x^2}\right)}} \sim t_{n-2} \quad sent \quad \hat{\beta}_0 = b_0 = \bar{Y} - b_1\bar{X}$$

- **Prova d'hipòtesi sobre la desviació tipus residual:**

$$\begin{cases} H_0: \sigma = \sigma_0 \\ H_1: \sigma \neq \sigma_0 \end{cases} \rightarrow \frac{(n-2)S^2}{\sigma_0^2} \sim \chi_{n-2}^2 \quad sent \quad \hat{\sigma} = S^2 = \frac{\sum(e_i^2)}{n-2}$$

Els graus de llibertat són **n-2** per les dues restriccions al necessitar estimar dos paràmetres previs

Formulari

Paràmetre	β_0	β_1	σ^2
Estimador	$b_0 = \bar{Y} - b_1\bar{X}$	$b_1 = s_{XY}/(s_X^2)$	$S^2 = \left(\sum(e_i^2)\right)/(n-2)$
Esperança	$E(b_0) = \beta_0$	$E(b_1) = \beta_1$	$E(S^2) = \sigma^2$
Variància	$S_{b_0}^2 = \sigma^2\left(\frac{\bar{X}^2}{(n-1)S_X^2}\right)$	$S_{b_1}^2 = \frac{\sigma^2}{(n-1)S_X^2}$	$V(S^2) = 2\sigma^4/(n-2)$
Distribució	$b_0 \sim N$ $(b_0-\beta_0) / S_{b0} \sim t_{n-2}$.	$b_1 \sim N$ $(b_1-\beta_1) / S_{b1} \sim t_{n-2}$.	$(n-2)S^2/\sigma^2 \sim \chi^2_{n-2}$
Interval de Confiança	$IC(95\%,\beta_0) =$ $= b_0 \pm t_{n-2,0.975}\cdot S_{b0}$	$IC(95\%,\beta_1) =$ $= b_1 \pm t_{n-2,0.975}\cdot S_{b1}$	$IC(95\%,\sigma^2) =$ $\left[\frac{(n-2)S^2}{\chi^2_{n-2,0.975}}, \frac{(n-2)S^2}{\chi^2_{n-2,0.025}}\right]$
H_0 usual	$\beta_0 = 0$	$\beta_1 = 0$	
Rebutgem H_0 si	$b_0 / S_{b0} > t_{n-2,0.975}$	$b_1 / S_{b1} > t_{n-2,0.975}$	

Prova d'hipòtesi sobre el pendent (β_1). Exemple

1. **Variables**: Cervesa i contingut d'alcohol a la sang

R: `lm`

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.0127     0.0126   -1.00     0.33
n.cerv        0.0180     0.0024    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
```

2. **Estadístic**: $\hat{t} = \frac{b_1 - \beta_1'}{s_{b_1}} = \frac{b_1 - \beta_1'}{\sqrt{s^2/[(n-1)\cdot S_x^2]}}$
3. **Hipòtesis**: $H_0: \beta_1 = 0 \quad vs \quad H_1: \beta_1 \neq 0$
4. **Distr. estadístic sota H_0**: $t_{n-2} = t_{14}$
5. **Càlculs**: $\hat{t} = 0.018/0.024 = 7.48$
6. ***P*-valor:** $P(|t_{14}| > 7.48) \approx 3\cdot10^{-6}$ (punt crític = $t_{14,0.975} = 2.145$)
7. **Conclusió**: rebutgem H_0 (*P*-valor < 0.05 o que 7.48 > 2.145)

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

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

[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]

Prova d'hipòtesi sobre el terme independent (β_0). Ex.

1. **Variables**: Cervesa i contingut d'alcohol a la sang

R: `lm`

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.0127     0.0126   -1.00     0.33
n.cerv        0.0180     0.0024    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
```

2. **Estadístic**: $\hat{t} = \frac{b_0 - \beta_0'}{s_{b_0}} = \frac{b_0 - \beta_0'}{\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_x^2}\right)}}$

3. **Hipòtesis**: $H_0: \beta_0 = 0 \quad vs \quad H_1: \beta_0 \neq 0$

4. **Distr. estadístic sota H_0**: $t_{n-2} = t_{14}$

5. **Càlculs**: $\hat{t} = -0.0127/0.0126 = -1.00$

6. ***P*-valor:** $P(|t_{14}| > 1.00) \approx 0.33$ (punt crític = $t_{14,0.975} = 2.145$)

7. **Conclusió**: NO rebutgem H_0 (*P*-valor > 0.05 o que -1.00 > -2.145)

   Conclusió pràctica: 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%

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

Prova d'hipòtesi sobre el pendent (β_1). Exercici

1. **Variables**:

2. **Estadístic**:

3. **Hipòtesis**:

4. **Distr. estadístic sota H_0**:

5. **Càlculs**:

6. ***P*-valor:**

7. **Conclusió**:

   Conclusió pràctica:

8. **$IC_{95\%}$:**

R: lm

```
Call: lm(formula = Durada ~ Brillantor)
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
```

Validació del model lineal

- L'anàlisi de les premisses en la variable de resposta en regressió pot fer referència a la part determinista (recta) o a la part aleatòria (residual).
- En la part determinista (1 premissa):
  - **Linealitat** entre X i Y en el rang considerat

- En la part aleatòria (3 premisses). Com que X_i no és v.a., és constant, no està mesurada amb error, llavors $V(y_i) = V(\beta_0 + \beta_1 X_i + e_i) = V(e_i) = \sigma^2$. Així les premisses sobre la part aleatòria de y_i les analitzem sobre els residus e_i. Els e_i són v.a. i.i.d. amb una distribució Normal $N(0,\sigma^2)$ [es diu que e_i és soroll blanc]:
  - **Homoscedasticitat**: mateixa σ^2 per qualsevol *i*
  - **Independència**: un error no aporta informació sobre el valor de l'altre
  - **Normalitat:** resultat de molts fenòmens aleatoris amb pesos petits

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \rightarrow \epsilon_i \sim N(0,\sigma) \qquad \epsilon_i, \epsilon_j \ \ ind. \ \forall i,j$$

Linealitat | Normalitat | Homoscedasticitat | Independència

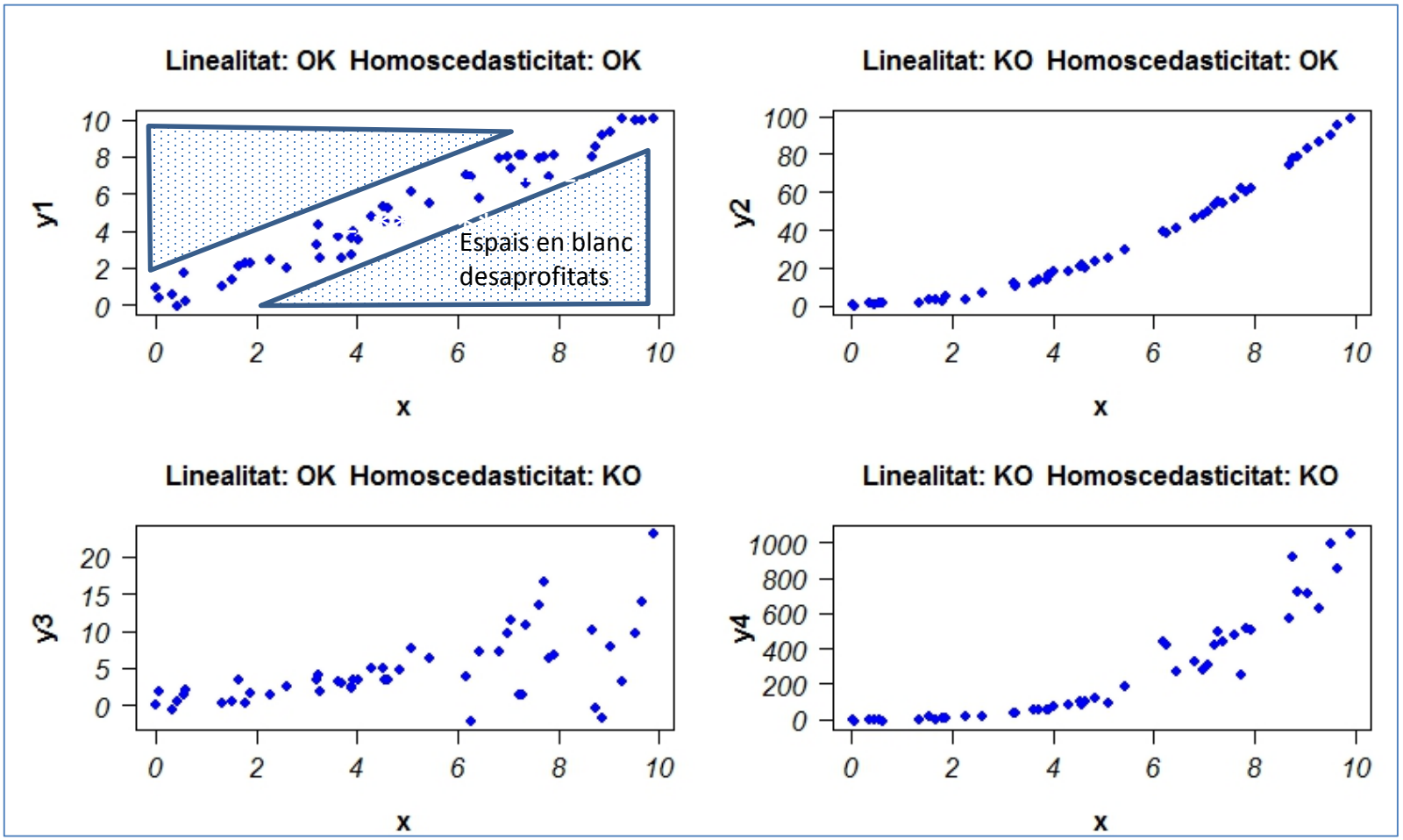
Validació model lineal. Anàlisi dels residus

- El compliment de les premisses anteriors permet:
  - poder recórrer a les distribucions de referència (per fer IC, PH)
  - garantir que el model és el millor possible
- L'anàlisi de les premisses:
  - Estudia si són raonables
  - O, en cas contrari, com trobar un model alternatiu per a que es compleixin
- La validació es realitza mitjançant gràfics dels residus. Usarem els següents gràfics per validar (o no) les premisses:

- Y_i versus X_i → Linealitat i homoscedasticitat
- **e_i versus "Fitted Values"** → Linealitat i homoscedasticitat
- **e_i versus ordre observacions** → Independència
- **Qqnorm dels residus (e_i)** → Premissa de Normalitat
- **Histograma dels residus (e_i)** → Premissa de Normalitat

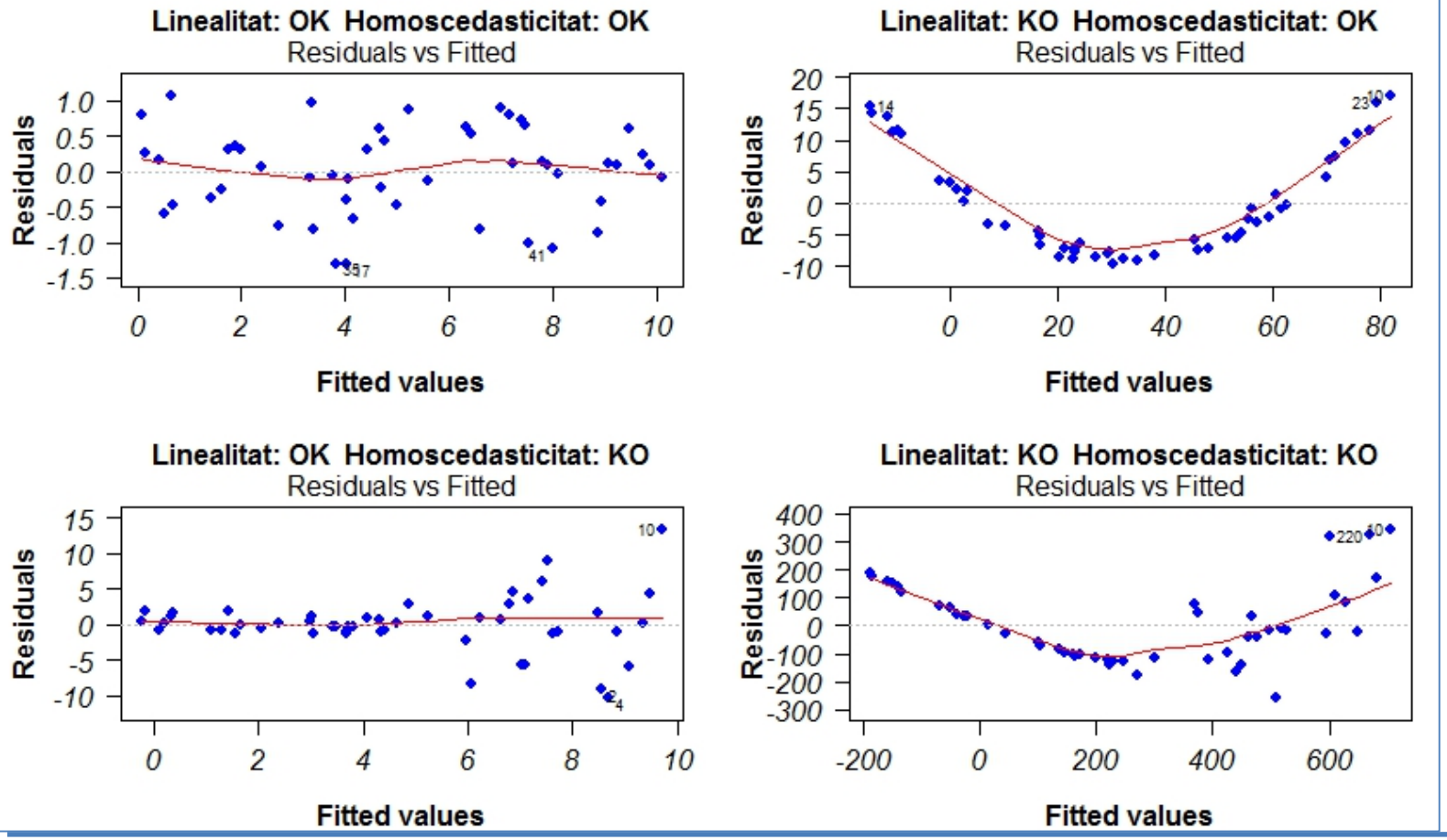
Validació model lineal. Gràfic Y_i versus X_i

- Permet estudiar la linealitat i la homoscedasticitat. És molt fàcil i intuïtiu, però ineficient: molts espais en blanc. Es pot millorar, substituint Y pels residus (e_i).



Validació model lineal. e_i versus fitted values.

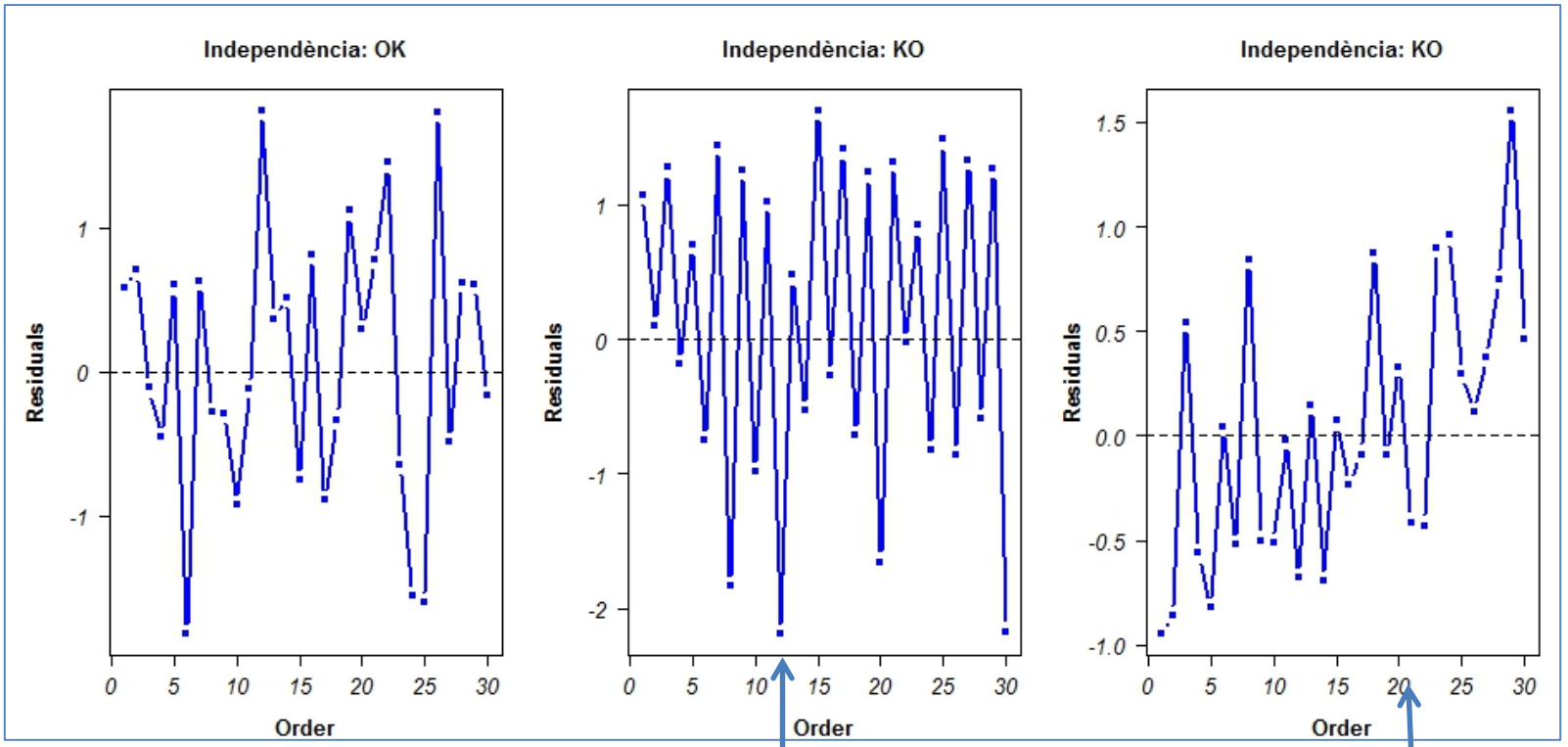
- **Linealitat**: El núvol de punts ha de mantenir sempre la mateixa alçada (aprox.).
- **Homoscedasticitat**: La variabilitat dels residus ha de mantenir-se constant independentment dels valors predits (*fitted values*).



Suggeriment: proveu amb les dades del consum de benzina i velocitat

Validació model lineal. e_i versus ordre de les observac.

- **Independència**: Els residus no han de mostrar cap patró enfront l'ordre.

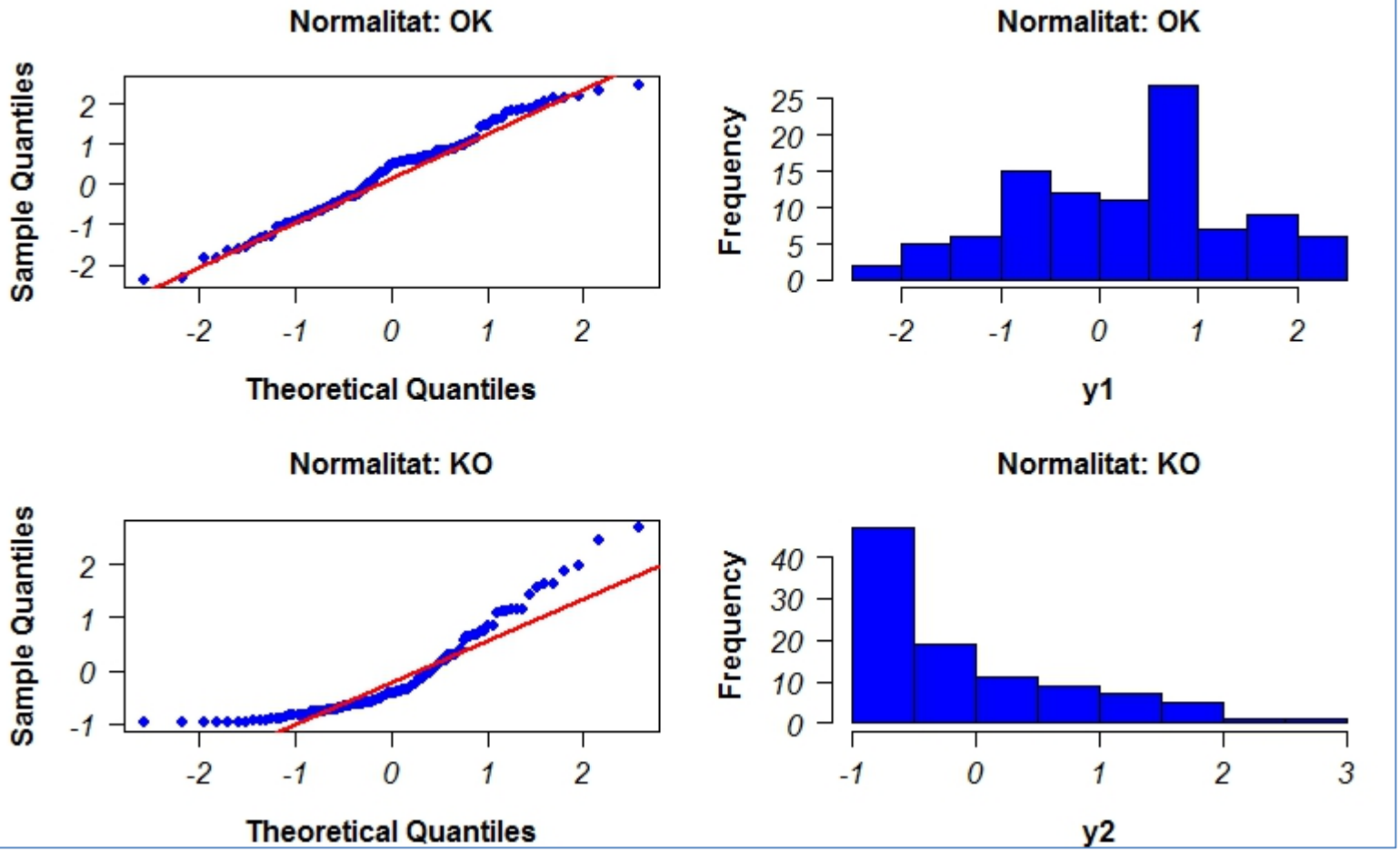


Hi ha patró: s'alternen pujades i baixades sistemàticament: no hi ha independència entre observacions consecutives. És típica de variables recollides al llarg del temps. [Ex: hores dormides en dies consecutius]

Hi ha un patró: tendència creixent dels residus. Segurament, s'ha anat canviant el criteri de recollida amb el temps.

Validació model lineal. qqnorm i histograma de residus

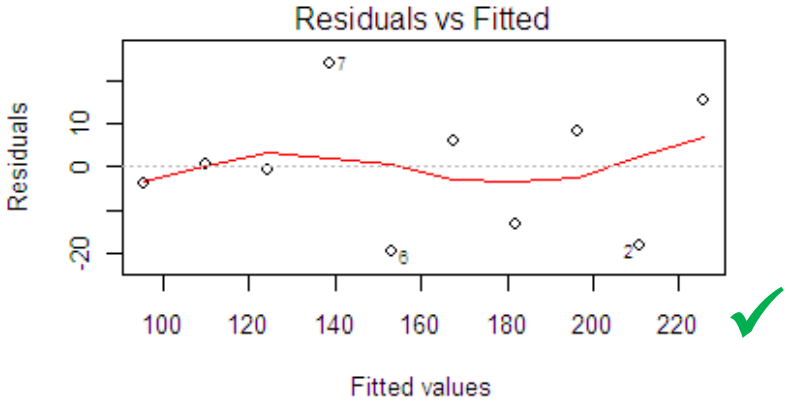
- **Normalitat**: Els residus han de ser normals: situar-se sobre la recta en el qqnorm i forma de campana a l'histograma.

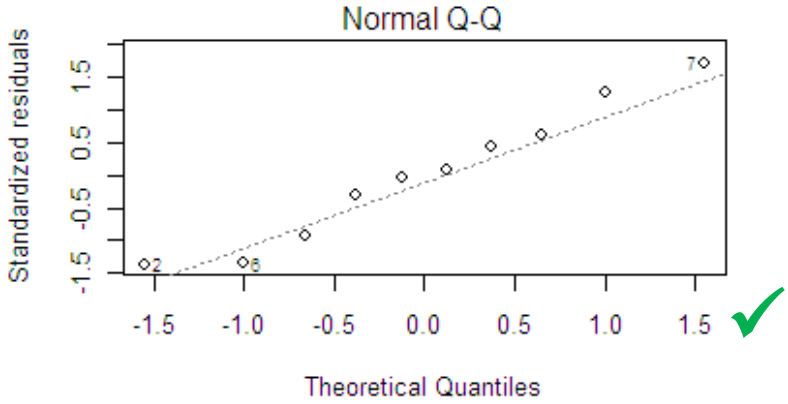


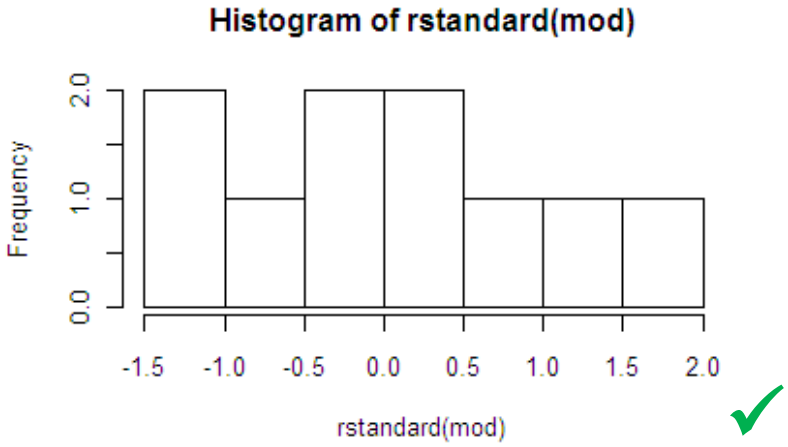
Nota: El qqnorm és molt més fiable a l'hora d'avaluar la Normalitat

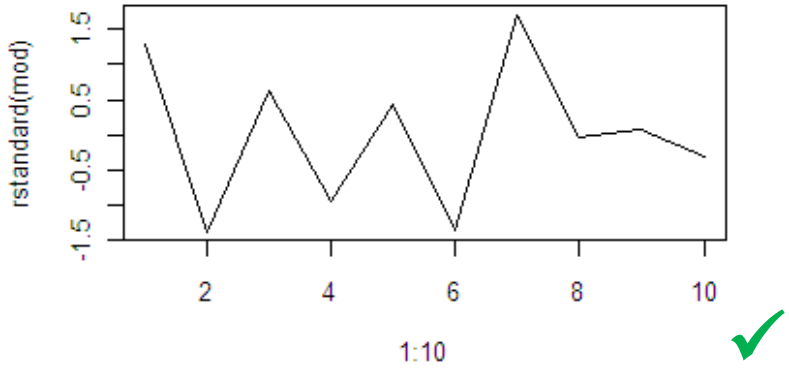
Validació model lineal. Codi R

```
##-- 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 estandaritzats
plot (1:10,rstandard(lm(Durada ~ Brill)),type="l") # Ordre dels residus
```





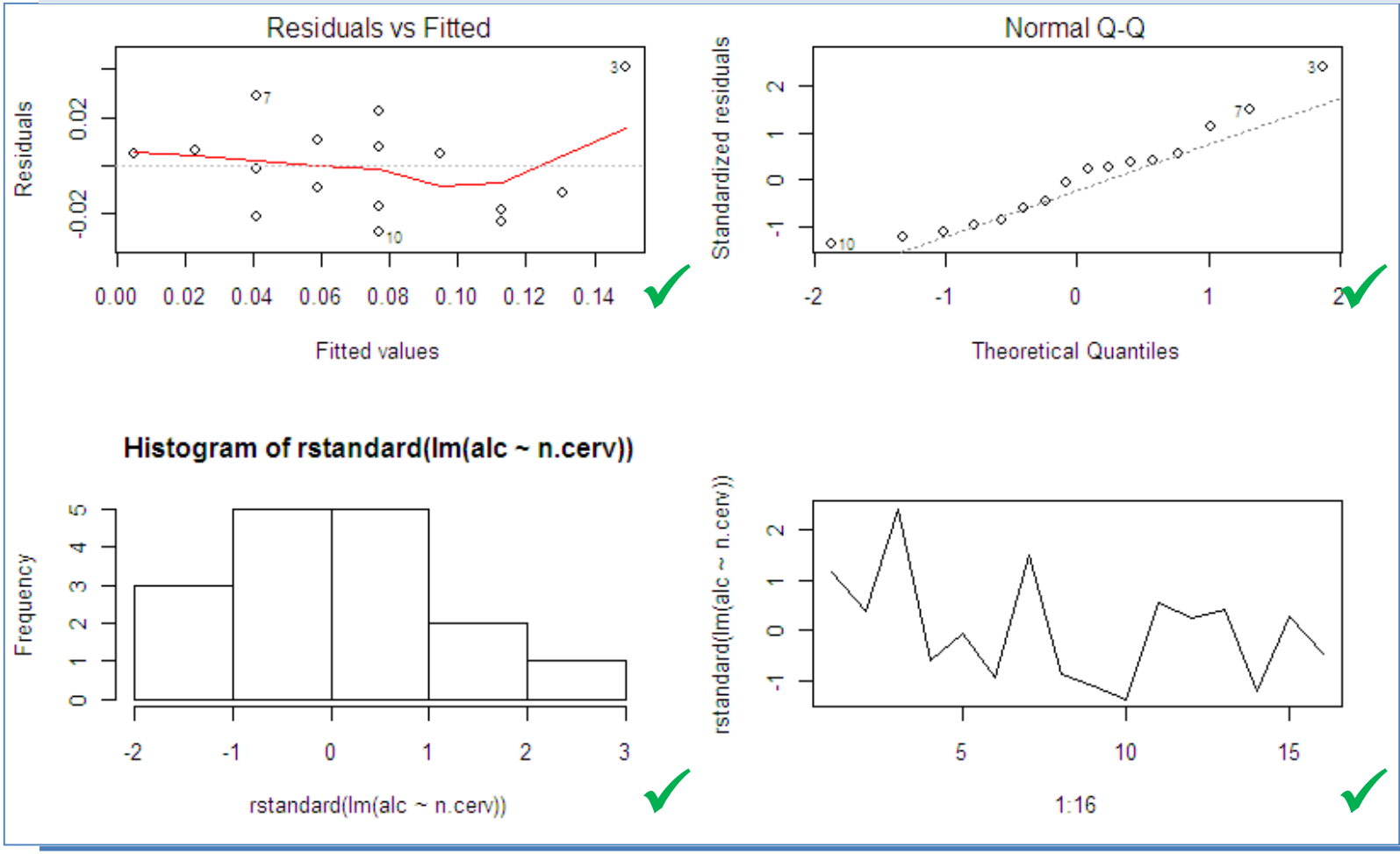




Encara que són poques dades, res s'oposa a validar cap de les 4 premisses

Validació model lineal. Codi R

```
##-- Exemple de les cerveses
par(mfrow=c(2,2))
plot(lm(alc~n.cerv),c(2,1))                    # QQ-Norm i Standard Residuals vs. Fitted
hist(rstandard(lm(alc~n.cerv)))                # Histograma dels residus estandaritzats
plot (1:16,rstandard(lm(alc~n.cerv)),type="l") # Ordre dels residus estandaritzats
```



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

Validació model lineal. Consideracions generals

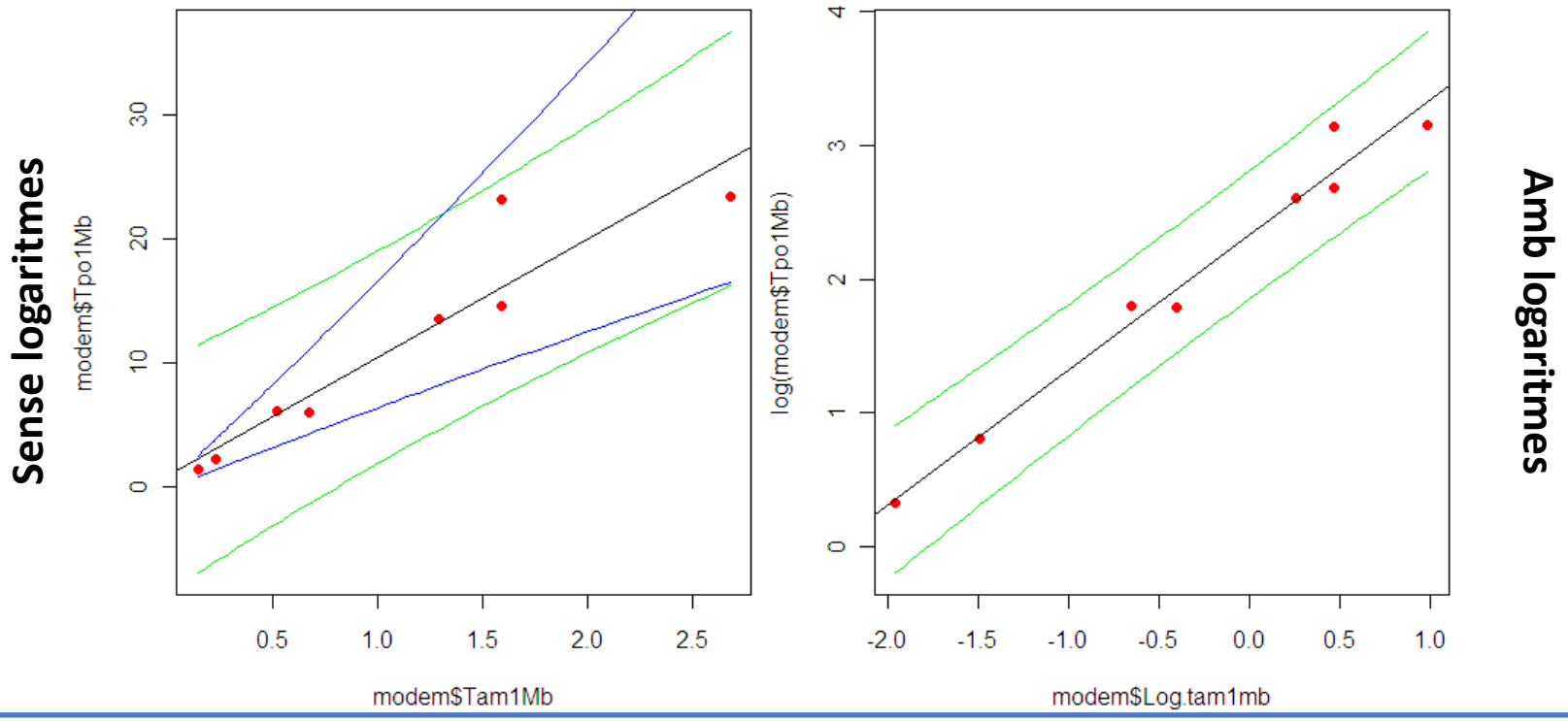
- **Totes les premisses**: Generalment, amb poques dades, és difícil avaluar les premisses i l'opció més prudent és acceptar-les a no ser que ho veiem molt clar que alguna d'elles s'infringeix.
- **Totes les premisses**: No s'ha de ser categòric i s'han d'interpretar els resultats amb cautela. S'ha d'evitar "Aquesta variable és homoscedàstica" o "Hem demostrat que la variable és Normal". Millor dir "Aquesta variable pot ser modelada assumint homoscedasticitat" o "Aquesta variable pot ser modelada amb la distribució Normal"
- **Linealitat i Homoscedasticitat**: Encara que es pot fer servir directament el gràfic de les Y_i vs X_i, es recomana emprar els residus vs "valors predits" per estudiar-les.
- **Homoscedasticitat**: cal recordar que S té molta oscil·lació mostral i, a vegades és poden observar fluctuacions al llarg dels valors predits que poden ser degudes al atzar.
- **Normalitat**: És més fiable el qqnorm que no l'histograma per avaluar-la
- **Independència**: El fet de ser independents respecte a l'ordre no garanteix del tot la indepèndencia. Aquesta ha de ser garantida amb un bon disseny de recollida.
- **Totes les premisses**: En motes ocasions, traient logaritmes (o arrel o fent la inversa) de alguna/es de les variable/s, podem solucionar l'incompliment de les premisses

Validació. Transformació logarítmica

En ocasions, fer la transformació logarítmica pot solucionar el NO compliment de les premisses. Ex: Velocitat de descarrega de fitxers amb un mòdem de 1Mbps (**Resposta**: temps [s] ; **Var. explicativa**: mida fitxer [MB])

Model #1: temps vs mida. Problema: heteroscedasticitat. Tenim prediccions negatives

Model #2: log(temps) vs log(mida). Desfem canvi amb **exp(predicció)**; ara són satisfactòries i tenen en compte que fitxers petits tenen fluctuacions petites en temps



Predicció

- En primer lloc la predicció puntual de Y per a valors concrets de X (X_h) usa la part determinista: $\hat{y}_h = b_0 + b_1 \cdot X_h$
- Però, com tenir en compte la part aleatòria? Dues situacions ben diferenciades:
  1. Estimar un interval de confiança pel **valor esperat** de les observacions X = X_h
  2. Estimar un interval de confiança pel **valor individual** corresponent a X = X_h

1. La estimació puntual per *y* donat un valor X_h és:

$$\hat{y}_h = b_0 + b_1 \cdot X_h = \bar{Y} + b_1 \cdot (X_h - \bar{X})$$

Podem estimar l'esperança i la variància d'aquest estimador:

$$E(\hat{y}_h) = E(b_0 + b_1 \cdot X_h) = \beta_0 + \beta_1 \cdot X_h = \mu_h$$ [És no esbiaxiat!!!]

$$V(\hat{y}_h) = V\left(\bar{Y} + b_1 \cdot (X_h - \bar{X})\right) = V(\bar{Y}) + (X_h - \bar{X})^2 \cdot V(b_1) = \frac{\sigma^2}{n} + \frac{(X_h - \bar{X})^2 \sigma^2}{(n-1) \cdot S_x^2} = \sigma^2 \left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{(n-1) \cdot S_x^2}\right)$$

Nota 1: Noteu que major variància a major distància entre X_h i $\bar{X}$ (més precissió al mesurar punts propers a la mitjana)

Nota 2: Substituint s per S podem fer regions de confiança per μ_h amb una t_{n-2}

Predicció

2. Per **predir** l'interval dels valors individuals y_h de Y per X=X_h utilitzarem també:

$$\hat{y}_h = b_0 + b_1 \cdot X_h$$

Calcularem esperança i variança:

$$E(y_h) = E(\hat{y}_h) = \mu_h$$

Té Error Quadràtic Mitjà de Predicció (EQMP) que es pot descomposar de forma semblant a la descomposició de sumes de quadrats:

$$EQMP = E[(\hat{y}_h - y_h)^2] = E(\hat{y}_h - \mu_h)^2 + E(y_h - \mu_h)^2$$

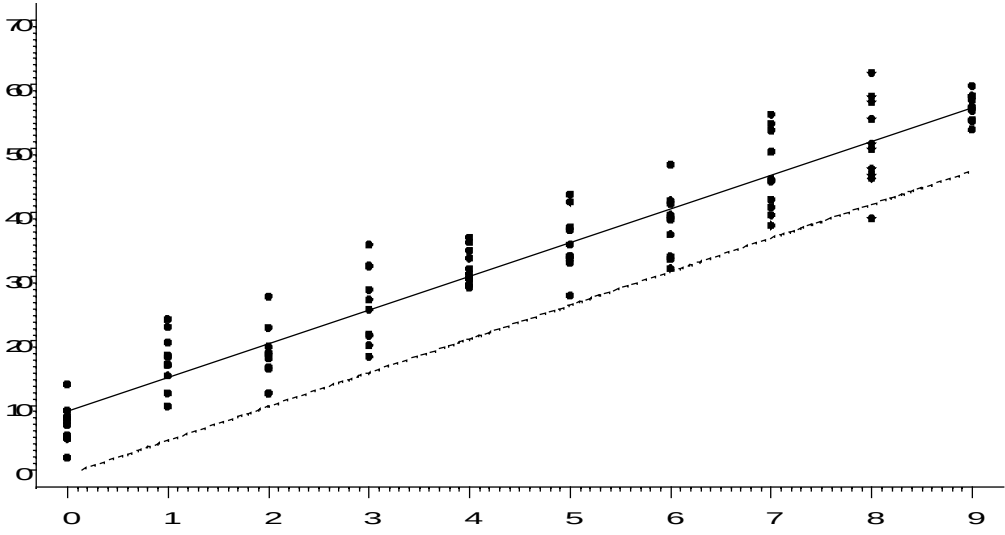
És a dir:

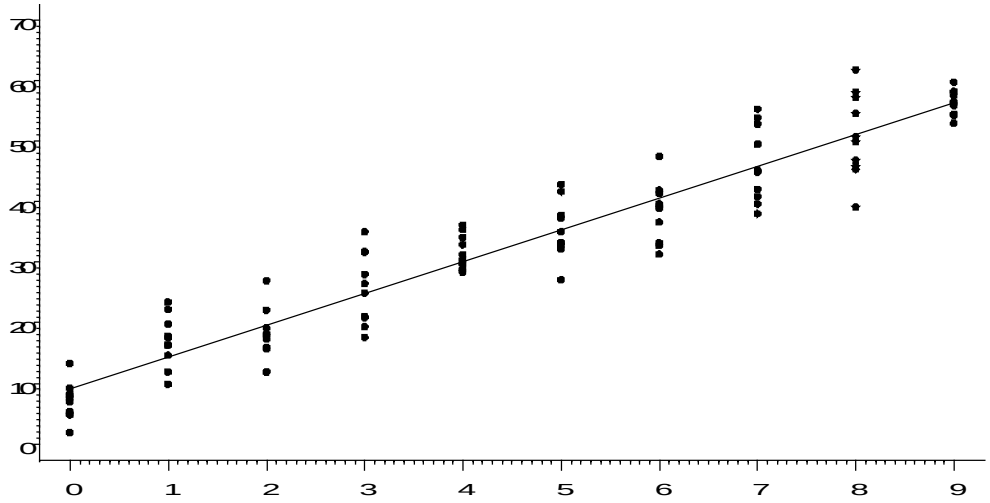
$$V(y_h) = \sigma^2 \left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{(n-1)S_x^2} \right) + \sigma^2 = \sigma^2 \left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{(n-1)S_x^2} \right)$$

Permet identificar 3 fonts de variabilitat en la predicció dels valors individuals: **Natural** (σ^2) **+ Per estimació mitjana** (σ^2/n) **+ Per estimació pendent**

Predicció. Gràfics

Predicció sobre valors individuals





Predicció sobre valors esperats

Formulari. Resum de previsions de la resposta

	Valor esperat	Valors individuals
Estimació puntual	$\hat{y}_h = b_0 + b_1 X_h$	$\hat{y}_h = b_0 + b_1 X_h$
Estimació per interval	$\hat{y}_h \mp \mathrm{t}_{\mathrm{n}-2,0.975} \cdot \mathrm{S} \sqrt{\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$	$\hat{y}_h \mp \mathrm{t}_{\mathrm{n}-2,0.975} \cdot \mathrm{S} \sqrt{1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$
R	`predict(…,interval='confidence')`	`predict(…,interval='prediction')`

Predicció. Exemple

Recuperem 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?

$\bar{x} = 5.5$
$s_x^2 = 9.167$
$s^2 = 227.3$

	Valor esperat	Valors individuals
Estimació puntual	$\hat{y}_h = 239.9 - 14.41 \cdot 7.5 = \mathbf{131.83}$	$\hat{y}_h = 239.9 - 14.41 \cdot 7.5 = \mathbf{131.83}$
Estimació per interval	$131.83 \mp 2.31 \cdot \sqrt{227.3} \cdot \sqrt{\frac{1}{10} + \frac{(7.5-5.5)^2}{9 \cdot 9.167}}$ $= [\mathbf{118.41, 145.25}]$	$131.83 \mp 2.31 \cdot \sqrt{227.3} \cdot \sqrt{1 + \frac{1}{10} + \frac{(7.5-5.5)^2}{9 \cdot 9.167}}$ $= [\mathbf{94.50, 169.16}]$
Conclusió	**Per a les pantalles** de brillantor de 7.5 podem esperar una **durada mitjana** entre 118.41 i 145.25 min. amb una confiança del 95%	**Per a una pantalla** de brillantor 7.5 podem esperar una **durada** entre 94.50 i 169.16 min. amb una confiança del 95%

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

Predicció. Exemple (Amb R)

```
> modem$Tam1Mb
[1] 1.59129 1.59129 0.51858 1.29297 0.14062 0.22461 0.66895 2.68000
> modem$Tpo1Mb
[1] 23.22 14.56  6.07 13.50  1.38  2.24  5.95 23.45
> mod1 = lm(Tpo1Mb ~ Tam1Mb, data=modem)
> summary(mod1)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.908      1.962    0.46  0.65995
Tam1Mb         9.544      1.447    6.59  0.00058 ***
> modem$Log.tam1mb = log(modem$Tam1Mb)
> mod2 = lm(log(Tpo1Mb) ~ Log.tam1mb, data=modem)
> summary(mod2)
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3322     0.0679    34.3  4.1e-08 ***
Log.tam1mb    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
...
```

Trobeu aquests valors a mà

$\bar{x} = -0.293$
$s_x^2 = 1.065$
$s^2 = 0.0338$

Capacitat predictiva. Coeficient R^2

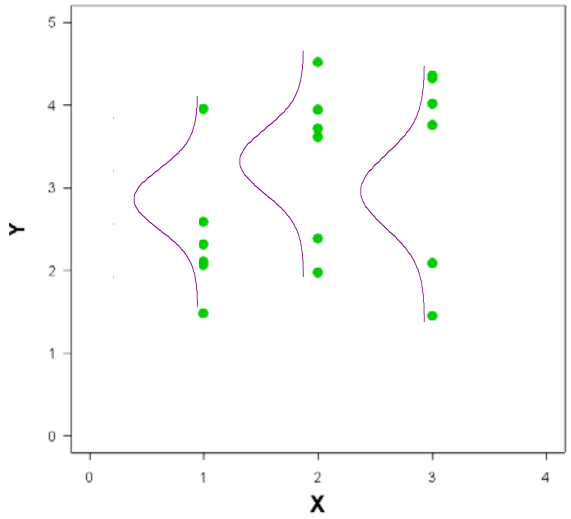
- Es va veure que la correlació entre X i Y (r_{XY}) estudia la relació (lineal) entre dues variables X i Y amb un rol simètric:

$$r_{XY} = r = \frac{S_{XY}}{S_X S_Y}$$

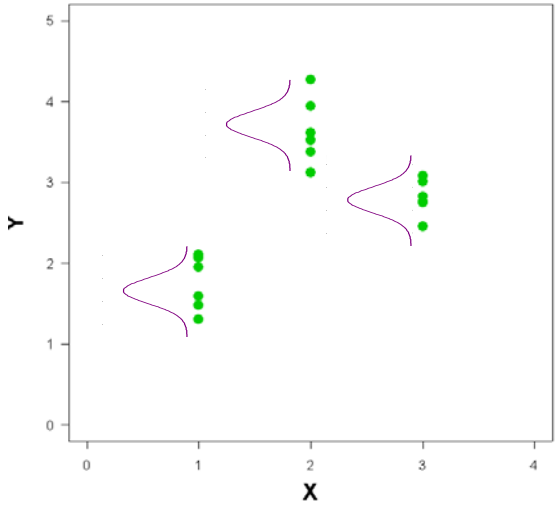
- Definim el coeficient ***R^2*** (***Coeficient de determinació, o R-squared***), com el quadrat de la correlació lineal *r*. Noteu que:
  - $R^2 = r_{XY}^2$ → $0 \leq R^2 \leq 1$
  - Ve a significar quina fracció de la variabilitat de Y s'explica per el factor X (la interpretació és asimètrica).
  - Un R^2 alt ens diu que el model lineal fa un bon ajustament de les dades :: els punts s'allunyen poc de la recta :: poca variabilitat d'origen aleatori
  - Recíprocament, amb R^2 baix, les dades no s'ajusten be :: els punts es poden allunyar molt :: gran variabilitat d'origen aleatori (no explicada per X, volem dir).
  - R^2 és un indicador de qualitat de l'ajustament, partint de que tenim un model lineal mentre que *r* és un indicador d'associació entre dues variables relacionades linealment, però no suposa cap model al darrera (caràcter descriptiu)

Annexe: quantitativa vs. categòrica.

- Quin model usar quan la intervenció **X** (o condició Z) és categòrica?



Petita variabilitat entre grups
Gran variabilitat intra grups



Gran variabilidad entre grups
Petita variabilitat intra grups

- Com a la regressió, podrem descomposar la variabilitat total en dues fonts de variació: entre-grupos (between) i intra-grupos (intra)
- Com a la regressió, podem tenir 2 objectius ben difereciats:
  - predir la resposta Y a partir (observació) dels valors Z (interesarà conneixer R^2)
  - canviar la resposta Y escollint (disseny d'experiments) els valor de X (interesaran les μ_i)

Model quantitativa vs. categòrica. Exemple

Exemple: Temps i nombre de nodes de graf en Dijkstra

Volem estudiar: - el temps de CPU empleat per l'algorisme Dijkstra

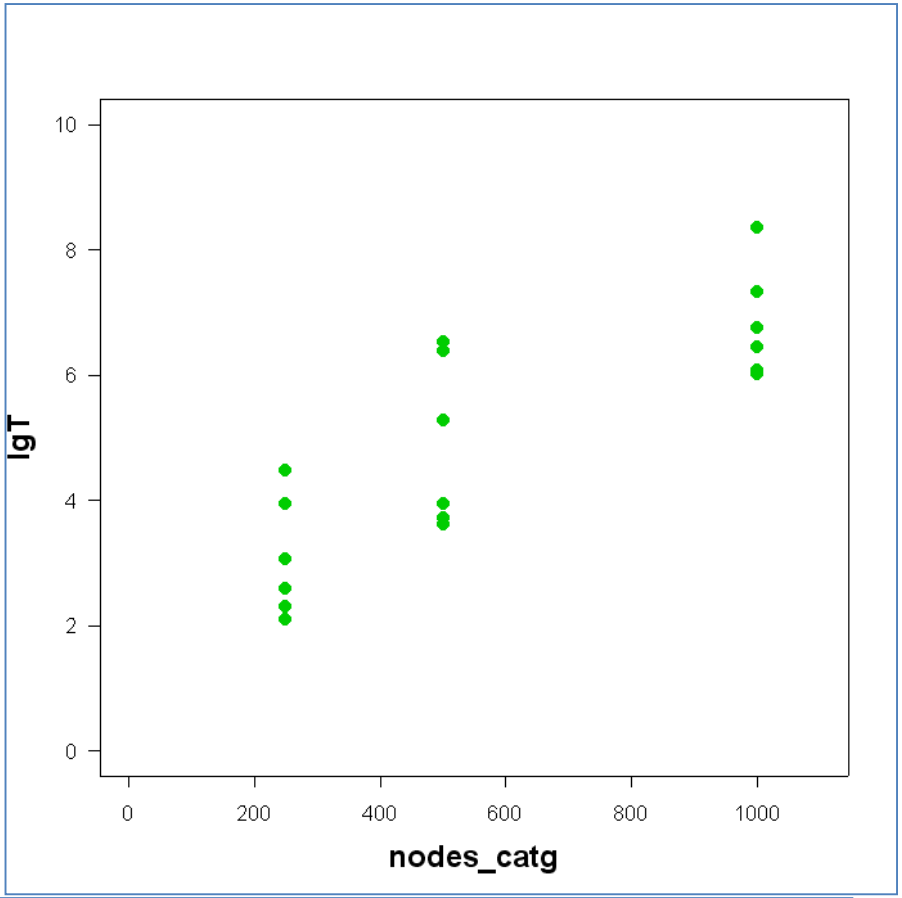
- segons el número de nodes del graf

Ho fem estudiant les característiques en grafs de 250, 500 i de 1000 nodes.

Si dubtem de la linealitat de les 3 mitjanes podem optar per un model que oblidi el número de nodes i els tracti com 3 categories, mirant únicament quin és el valor mitjà de Y per cada categoria de Z.

En aquest cas, té sentit estudiar la descomposició de la variabilitat en:

- Deguda al factor X
- Aleatòria o residual



Model quantitativa vs. categòrica. Paràmetres

Sigui el model: $\mathbf{Y_{ij} = \mu_j + e_{ij}}$

Y_{ij} valor de la Y en el cas i del grup j

μ_j esperança del grup j (de $\mathbf{n_j}$ observacions de les **N** totals)
(el paràmetre μ_j s'estima per la mitjana $\bar{y}_j$) (desviació $\mathbf{S_j}$)

e_{ij} error aleatori o diferència del cas i a la mitjana del seu grup j
(el paràmetre σ^2 és la variància de e_i o variància residual)

Aquest model es pot veure com l'extensió de la comparació de 2 μ al cas de k μ. Però també, com l'estudi de la descomposició de la variabilitat:

- **Entre els grups**: quant expliquen de la variabilitat global
- **Dins els grups**: variabilitat 'dins' que no es pot relacionar amb el grup

Model quantitativa vs. categòrica. Paràmetres

EXEMPLE: (*Estadística per a enginyers informàtics. Ed UPC pg 154 Ref: Eei.Ed.UPC pg154*). Notes en 3 grups d'una assignatura. Els paràmetres μ_1, μ_2 i μ_3 són les esperances de la nota en cadascun dels grups de grandàries n_1=32, n_2=28 i n_3=25 casos respectivament (en total N=85)

Les mitjanes i desviacions mostrals en cada grup són:

$\bar{y}_1 = 6.15; \bar{y}_2 = 5.73; \bar{y}_3 = 5.48; s_1 = 1.8; s_2 = 1.5; s_3 = 2.0$

Quant val la mitjana global? $\bar{Y} = 5.81$

Quant val la variabilitat combinada dins els grups? $S^2 = 3.136$

(com en la regressió, S^2 és l'estimació de σ^2 i es comprovarà amb resultat a taula Anova)

Model quantitativa vs. categòrica. Anàlisi: ANOVA

TAULA D'ANÀLISIS DE LA VARIANÇA (ANOVA, ANalysis Of VAriance)

R: **aov**

Posarem els termes de SQ en forma de taula (Ref: *Eei.Ed.UPC* pg 154):

Font Var.	SQ (Sum Squares)	Df*	QM$	Ratio	P-value
Between (Explicada pel model)	$SQ_E = \sum_{j=1}^{k} n_j \cdot (\bar{y}_j - \bar{Y})^2$	k-1	$QM_E = \frac{SQ_E}{k-1}$	$\hat{F} = \frac{QM_E}{QM_R}$	$P(F_{k-1,N-k} > \hat{F})$
Intra (Residual)	$SQ_R = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$	N-k	$QM_R = \frac{SQ_R}{N-k}$		
Total	$SQ_T = SQ_E + SQ_R = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{Y})^2$	N-1			

Between, Intra: Variabilitat explicada pel model (between) i aleatòria (intra)

***Df**: Degree of freedom (graus de llibertat)

$QM: Quadratic Mean

k: número de grups a comparar

N: Nombre d'observacions totals

$\hat{F}$: Aquest estadístic sota la hiòtesi (H_0) de que les mitjanes no són diferents segueix una F amb k-1 i N-k graus de llibertat

Model quantitativa vs. categòrica. Anàlisi: ANOVA

PH GLOBAL: [Ref: *Eei.Ed.UPC* pg 155]

- La hipòtesi de que X no aporta informació sobre Y (igualtat de totes les m_j) es tradueix en que tota la variabilitat entre les m_j és deguda a la fluctuació del mostreig
- PH: H_0 : Variabilitat(μ_j) = 0

  H_1: Variabilitat(μ_j) > 0 **unilateral!**
- Es resol amb la ràtio F dels quadrats mitjos de la taula de descomposició de la variabilitat:

$$\hat{F} = \frac{QM_E}{QM_R}$$

COEFICIENT DE DETERMINACIÓ:

- Com en el cas anterior, és un rati que ens permet identificar de tota la variabilitat de les Y quina part ve associada a (explicada per) X

$$R^2 = \frac{SQ_E}{SQ_T}$$

ANOVA. Exemple

Notes en 3 grups (Ref: Eei.Ed.UPC pàg. 154)

Font Variabilitat	SQ	GdL DF	QM	Raó
Explicada (entre)	6.60	2	3.3	1.052
Residual (intra)	257.19	82	3.136	
Total	263.79	84		

$$SQ_E = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y}) = 6.60 \qquad SQ_R = \sum_{j=1}^{k} (n_j - 1) \cdot s_j^2 = 257.19 \qquad SQ_T = \mathrm{SQ_E} + \mathrm{SQ_R} = 263.79$$

1. **Objectiu:** saber si el grup afecta a l'esperança de la nota
2. **Hipòtesi:** H_0: Variabilitat(μ_j) = 0 (unilateral)
3. **Estadístic**: $\hat{F} = QM_E / QM_R$
4. **Distribució sota H_0**: $\hat{F} \rightarrow F_{2,82}$ (les premisses caldrà indicar-les i analitzar-les)
5. **Càlculs**: $\hat{F} = 1.052$
6. **P-valor:** 0.354 `[(1-pf(1.052,2,82))]`
7. **Conclusió**: SÍ és versemblant que l'esperança de la nota és igual en tots els grups

   Pràctica: El rendiment mitjà no és diferent en els tres grups estudiats

ANOVA. Exemple (cont)

8. Intervals de confiança

Font Variabilitat	SQ	GdL DF	QM	Raó
Explicada (entre)	6.60	2	3.3	1.052
Residual (intra)	257.19	82	3.136	
Total	263.79	84		

Sabent que:

$\bar{y}_1 = 6.15; \bar{y}_2 = 5.73; \bar{y}_3 = 5.48; s_1 = 1.8; s_2 = 1.5; s_3 = 2.0$

$\bar{Y} = 5.81\ ; S^2 = 3.136$

Es pot calcular un IC per a la μ global: $IC(\mu, 1-\alpha) = \bar{Y} \mp t_{N-k,1-\frac{\alpha}{2}} \cdot \sqrt{\frac{QM_R}{N}}$

$IC(\mu, 0.95) = [5.432, 6.197]$

També es pot calcular un IC per a cada μ_j amb la desviació pooled (més robust que calculat amb les dades de cada grup): $IC(\mu_j, 1-\alpha) = \bar{y}_j \mp t_{N-k,1-\frac{\alpha}{2}} \cdot \sqrt{\frac{QM_R}{n_j}}$

$IC(\mu_1, 0.95) = [5.527, 6.773]$ $IC(\mu_2, 0.95) = [5.064, 6.396]$ $IC(\mu_3, 0.95) = [4.775, 6.185]$

ANOVA. Exercici

- X:nombre de nodes, Y: temps amb transformació logarítmica

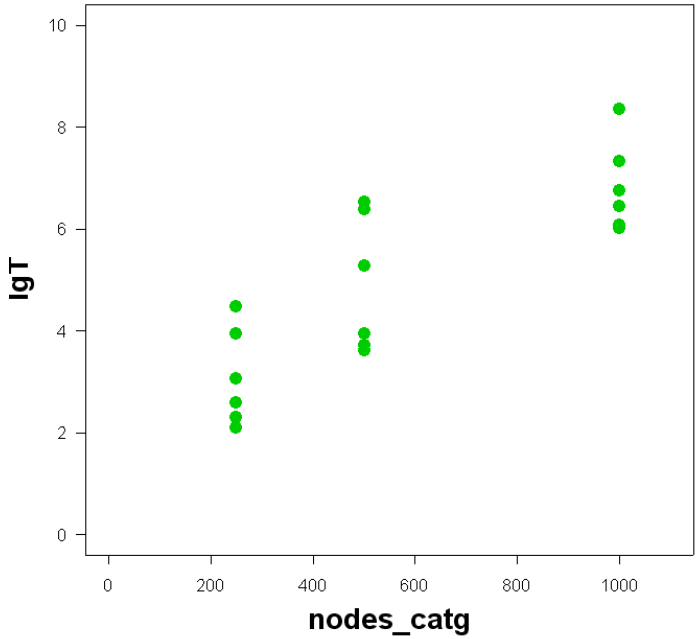
x_i (nodes)	y_i (lgt)
250	2.31
250	4.48
250	2.59
250	3.06
250	2.10
250	3.95
500	3.94
500	6.38
500	6.52
500	5.27
500	3.72
500	3.61
1000	6.45
1000	7.32
1000	6.76
1000	6.08
1000	8.35
1000	6.01

$\bar{y}_1 = 3.082 \quad s_1^2 = 0.90$

$\bar{y}_2 = 4.910 \quad s_2^2 = 1.79$

$\bar{y}_3 = 6.83 \quad s_3^2 = 0.79$

$\bar{y} = 4.94 \;\; s^2 = 1.16 \rightarrow S = 1.08$



Model amb R:

```
> aov(lgT~as.factor(nodes_catg))
Call:aov(formula=lgT~as.factor(
nodes_catg))
Terms:
                  nodes_catg Residuals
Sum of Squares     42.12188  17.37490
Deg. of Freedom           2        15
Residual standard error: 1.076256
Estimated effects may be unbalanced
```

ANOVA. Exercici

Temps i nombre de nodes de graf en Dijkstra

Font Variabilitat	SQ	GdL DF	QM	Raó
Explicada (entre)	42.12	2	21.06	18.15
Residual (intra)	17.38	15	1.16	
Total	59.5	17		

Model amb R:

```
> anova(aov(lgT~nodes_catg))
Call:aov(formula=lgT~nodes_catg)
Analysis of Variance Table
Response:  lgT
            Df Sum Sq Mean Sq F value    Pr(>F)
nodes_catg   2 42.122  21.061  18.182 9.789e-05
Residuals   15 17.375   1.158
```

$$SQ_E = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y}) = 42.12 \qquad SQ_R = \sum_{j=1}^{k} (n_j - 1) \cdot s_j^2 = 17.38 \qquad SQ_T = \mathrm{SQ_E} + \mathrm{SQ_R} = 59.50$$

1. **Objectiu:** saber si nombre de nodes afecta al temps
2. **Hipòtesi:** H_0: Variabilitat(μ_j) = 0 (unilateral)
3. **Estadístic**: $\hat{F} = QM_E / QM_R$
4. **Distribució sota H_0**: $\hat{F} \rightarrow F_{2,15}$ (les premisses caldrà indicar-les i analitzar-les)
5. **Càlculs**: $\hat{F} = 18.15$
6. **P-valor:** 0.000098 `[(1-pf(18.15,2,15))]`
7. **Conclusió**: No és versemblant que el nombre nodes no aporti informació sobre el temps

Pràctica: El logaritme del temps mitjà és diferent en els tres nivells de nombre de nodes.