FIB
Departament d'Estadística
i Investigació Operativa
UNIVERSITAT POLITÈCNICA DE CATALUNYA
UPC

Bases de l'estadística

Bloc C – Probabilitat i Estadística

2024

Índex

1. Inferència. Introducció i conceptes bàsics. Paràmetres
2. Estimació puntual. Estimadors. Propietats dels estimadors
3. Estimació per Interval de Confiança (IC)
   a. Estadístics.
   b. Confiança i risc
   c. Premisses
4. Interval de Confiança
   a. IC d'1 paràmetre
      - IC de μ Cas de σ coneguda i cas de σ desconeguda
      - IC de π
      - IC de σ
   b. IC comparant 2 paràmetres
      - IC de μ_1-μ_2 en mostres aparellades
      - IC de μ_1-μ_2 en mostres independents
      - IC de π_1-π_2 (en mostres independents)
      - IC de σ^2_1/σ^2_2 (en mostres independents)
   c. Funcions en R per a IC
5. Bloc T: disseny (com obtenim les dades)
6. Bloc T: descriptiva de les dades. Estadística Descriptiva Uni- i Bi-variant

1. Inferència estadística.

- Hem d'aportar evidència basada en dades

  per exemple, dir *el meu programa funciona* requereix proves/dades

- De forma **reproduïble**: només resultats predictibles tenen interès

  per exemple, una curació **miraculosa** no serà útil per futurs pacients

- I **transparent**

  per permetre la seva replicació per a altres

- **Inferim** les característiques de la **població** a partir de les observacions d'una **mostra aleatòria (m.a.)**

  per exemple, puc inferir la velocitat de connexió a tota la població a partir d'una mostra aleatòria de velocitats

1. Inferència estadística. Riscos

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

1. Inferència estadística. Tipus de variables

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

- **Resposta Y**. Mesura l'assoliment de l'objectiu -de vegades pot ser una mesura indirecta

  Ex: rendiment **Y** mesurat en les notes de certa assignatura

- **Decisions X.** Assignem els seus valors en els estudis experimentals

  Representen el potencial per canviar el futur: volem mesurar l'**efecte** de **X** en **Y**

  Un disseny experimental permet la seva independència de la resta de variables.

  Ex: un mètode docent basat en **llistes impreses** d'exercicis **(X=1)** comparat amb un mètode basat en **e-status (X=2).**

- **Co-variables Z**. Representen les condicions observades en dades *reals*

  Podem usar les **Z** per reduir la incertesa de **Y** (haurem de quantificar el seu encert)

  Podem obtenir les **Z** tant en estudis experimentals com observacionals

  Les **Z** solen estar interrelacionades (*col·lineals* o *no ortogonals*)

  Ex: les notes **(Z_1, Z_2)** de dues assignatures prèvies solen tenir certa relació

1. Inferència estadística. Tipus d'estudis

- **FER: Estudis experimentals**

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

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

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

X representa una causa **assignable** ben definida

La clau per intervenir és ser **propietaris** de **X**

Per garantir la independència amb tota Z, assignem **X** a l'atzar

Assignem respectant drets ètics i legals.

- **VEURE: Estudis observacionals**

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

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

  Ex: comparem notes **Y** segons el grup ($\mathbf{Z_1}$), o segons la nota d'una altra assignatura ($\mathbf{Z_2}$), o en funció d'un cert **model m** de les dues variables [$\mathbf{m=f(Z_1, Z_2)}$].

  → *El grup* $\boldsymbol{Z_1}$ *per ell mateix redueix un 10% la incertesa; la nota* $\boldsymbol{Z_2}$*, un 20%; i el model* ***m****, amb les dues, un 25%.*

No som **propietaris** de les variables Z (les unitats ja venen amb el valor de les Z)

Podem establir **relacions** entre Z i Y, que podem utilitzar per **predir** els valor de Y a partir de Z.

Però les covariables **Z** poden estar relacionades (ser **col·lineals**) i per tant poden tenir **confosos** els seus *efectes* en **Y.**

Establir **causalitat** requereix moltes premisses (fora d'un curs introductori)

1. Inferència estadística. Conceptes bàsics

- **Paràmetre**: indicador de la població que estem interessats en conèixer o estimar. Per exemple la μ (esperança) de les alçades dels estudiants de la FIB

- **Estadístic**: qualsevol indicador que s'obtingui com a funció de les dades d'una mostra. Per exemple la suma de les alçades dels estudiants recollits en una mostra

- **Estimador**: estadístic d'una mostra que s'utilitza per conèixer el valor d'un paràmetre de la població. Per exemple la mitjana de les alçades en una mostra a l'atzar d'alumnes de la FIB és una estimador de la μ (esperança) de les alçades dels estudiants de la FIB

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

2. Estimació puntual

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

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

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

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

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

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

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: $\widehat{se} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n}}$ (amb s estimador puntual de σ)

A la referència de la bibliografia (Estadística per a enginyers informàtics) trobareu més detalls als capítol 1 i 2

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

2. Estimació puntual. Casos

Per als paràmetres utilitzem lletres de l'alfabet grec

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

El cas de la MITJANA:

$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ mitjana mostral és una **estimació puntual del paràmetre μ** de tendència central

El cas de la DESVIACIÓ:

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

desviació tipus mostral és una **estimació puntual del paràmetre σ** de dispersió

El cas de la PROPORCIÓ:

$p = \sum_{i, x_i=1} 1/n$ proporció mostral és una **estimació puntual del paràmetre π**

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

2. Propietats dels estimadors. Estadística descriptiva

- Propietats dels estimadors

  Un paràmetre pot tenir **diversos estimadors** (aproximacions al valor puntual desconegut),

  i tota **estimació puntual té marge d'error** (ja que depèn de la mostra),

  per tant convé poder comparar-los en base a algunes **propietats dels estimadors**:

  - **Biaix** (o "sesgo" o "bias") que interessa que sigui el més proper a 0, per assegurar que el valor esperat s'ajustarà al màxim al paràmetre
  - **Eficiència** que interessa alta, indicant més precisió, menys dispersió
  - altres propietats com consistència, ...

  A l'annex d'aquest bloc C trobareu més informació de propietats dels estimadors
  A la referència de la bibliografia (Estadística per a enginyers informàtics) trobareu més informació de propietats dels estimadors al capítol 2

- Els anteriors estimadors puntuals es corresponen amb funcions d'**Estadística Descriptiva** per resumir numèricament unes dades.

  A la pàgina web de l'assignatura trobareu més informació d'Estadística Descriptiva en R
  Al final d'aquest bloc trobareu més informació d'Estadística Descriptiva lligada al bloc T

3. Estimació per Interval de confiança

- Sabem com calcular un "interval" que contingui $\bar{x}$ a partir de μ. Però el problema real és **aproximar 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

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

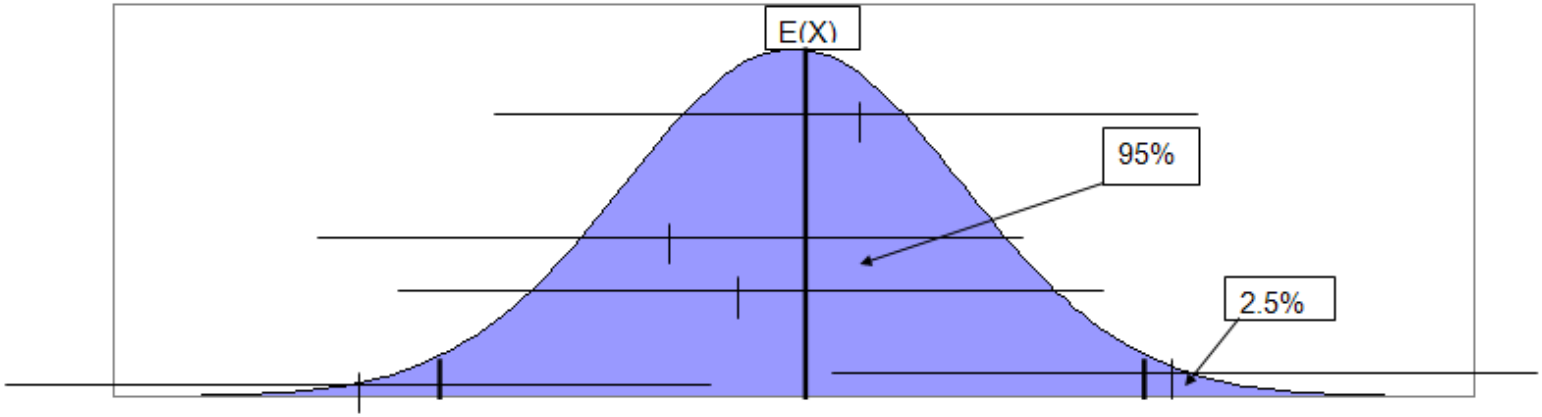
- I reordenant obtenim **l'interval de confiança 1-α del paràmetre μ**

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

A la referència de la bibliografia (Estadística per a enginyers informàtics) trobareu més detalls a la pg 48 i al capítol 3

3. Estimació per Interval de confiança

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



- *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:

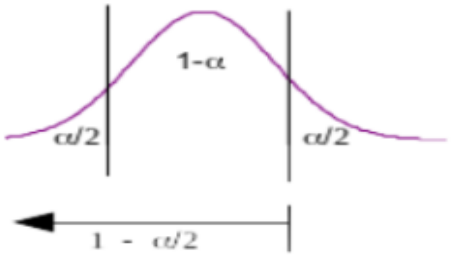
$$IC(\mu, 1-\alpha) = \bar{x} \mp z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

($z_{\alpha/2} = -z_{1-\alpha/2}$ ja que la Z és simètrica)

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

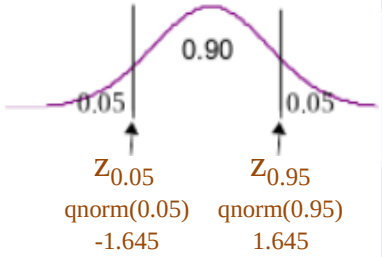
3.a. Confiança i risc

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

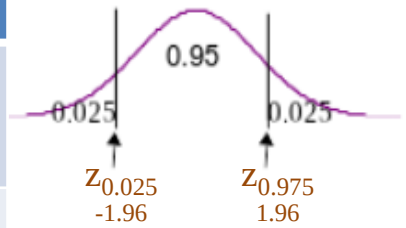


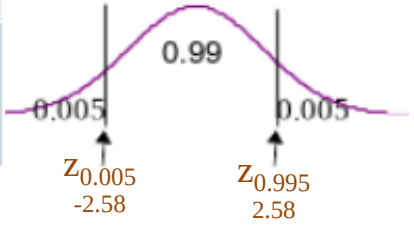
I podem relacionar el valor de confiança amb el quantil que necessitarem per construir l'IC:

(com a exemple estan indicats els quantils per una Z Normal(0,1) on sabem que $z_{\alpha} = -z_{1-\alpha}$ o $z_{\alpha/2} = -z_{1-\alpha/2}$)



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





3.b. Estadístics per fer inferència

Per a un paràmetre θ (pel que podem calcular un estimador $\hat{\theta}$ i un Interval de Confiança) podem considerar **estadístics** que, avaluats per a un valor concret del paràmetre, θ_0, **segueixen una distribució coneguda**.

Veurem estadístics "pivots" de dos tipus (veure resum en forma de taula al formulari):

- **Rati de "senyal" o "informació"** (diferència entre un valor del paràmetre, ex. μ_0 , i el mostral, $\bar{x}$) **respecte "soroll" o "error"** (estàndard error, ***se***)

  (a vegades parlem de "t-rati" per quantificar quantes vegades és més gran el senyal que el soroll)

  Aquests estadístics segueixen el model Z o t-Student*

  Estadístic $z = \frac{(\bar{x}-\mu_0)}{\sigma/\sqrt{n}} = \frac{(\bar{x}-\mu_0)}{se}$ $z \sim Z = N(0,1)$ (dona lloc al IC $\mu \in \bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ o bé $\bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot se$)

  Estadístic $t = \frac{(\bar{x}-\mu_0)}{s/\sqrt{n}} = \frac{(\bar{x}-\mu_0)}{\widehat{se}}$ $t \sim$ t**–Student** amb ν graus de llibertat (usa *s* enlloc de *σ*)

- **Quocient de variàncies**

  Aquests estadístics segueixen el model F o *Khi quadrat**

  Estadístic $\frac{s^2(n-1)}{\sigma^2}$ χ^2_ν és **Khi quadrat** amb ν graus de llibertat

El valor de l'estadístic (indicat en R per ex. *z*, *t* o "t value") permet dir si el valor concret θ_0 avaluat, **és versemblant o no** d'acord amb l'evidència de les dades,
mirant si el valor de l'estadístic és dels valors centrals o extrems de la distribució

Les distribucions t-Student (t_ν), F ($F_{\nu 1,\nu 2}$) i Khi quadrat (χ^2_ν) estan definides en el Bloc B. Son models derivats de la Normal, i estan **parametritzades amb **ν**, que anomenem "graus de llibertat" i que **depèn de** la mida **n** de la mostra*

3.c. Premisses

- La premissa fonamental és partir d'una **mostra aleatòria**

Diem que els valors venen de

Variables **A**leatòries **I**ndependents i **I**dènticament **D**istribuïdes, **v.a.i.i.d**.

app app

- Si la "n" és gran, la premissa de normalitat de la variable en estudi NO és necessària perquè els IC es basen en el TCL. En cas contrari, necessitem que la variable sigui Normal

En mostres petites (¿n<30?), sustentarem la **premissa de normalitat** en:

el **coneixement previ** de la variable resposta;

i amb el anàlisis gràfic **amb R.**

Veure l'apartat 7 de funcions de R i

l'Anàlisi gràfica de la normalitat a l'annex del bloc B.

4.a. Estimació per a IC d'1 paràmetre

Ara veurem les **fórmules d'IC** quan tenim **UN SOL PARÀMETRE** d'interès, distingint els següents casos:

- interès en la mitjana μ (amb variància poblacional coneguda o no)
  *(per exemple la mitjana de la nota d'una assignatura *)*
- interès en una proporció π
  (per exemple la proporció d'aprovats d'una assignatura)
- interès en la variabilitat σ^2
  (per exemple la desviació de la nota d'una assignatura)

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

Interval de confiança de μ amb σ coneguda

- L'interval de confiança 1−α de μ (amb σ coneguda) es calcula com

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

- Aquest IC es pot relacionar amb l'estadístic $\frac{(\bar{x}-\mu)}{\sigma/\sqrt{n}}$ que per a un valor concret (μ_0) del paràmetre, segueix la distribució *Normal*: $z = \frac{(\bar{x}-\mu_0)}{\sigma/\sqrt{n}} = \frac{(\bar{x}-\mu_0)}{se} \sim N(0,1)$

(tal com hem vist, els estadístics ens permeten contrastar si el valor del paràmetre avaluat és versemblant o no, a partir de si el valor de l'estadístic està en la zona central o extrema de la distribució)

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

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)

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

Interval de confiança de μ amb σ desconeguda

- L'interval de confiança 1−α de μ (amb σ desconeguda) es calcula com:

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

- Aquest IC es pot relacionar amb l'estadístic $\frac{(\bar{x}-\mu)}{s/\sqrt{n}}$ que per a un valor concret (μ_0) del paràmetre, segueix la distribució ***t-Student***: $t = \frac{(\bar{x}-\mu_0)}{s/\sqrt{n}} = \frac{(\bar{x}-\mu_0)}{se} \sim t_{n-1}$

  (tal com hem vist, l'estadístic ens permet contrastar si el valor del paràmetre avaluat és versemblant o no, i en alguns resultats en R s'anomena ***t-value)***

- En aquest cas cal que la variable X inicial fos Normal (premissa de normalitat) ja que la definició de la ***t-Student*** parteix de variables normals

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

t i N(0,1) són similars, més quan *n* creix: $t_{n\text{-->}\infty}$ → N(0,1)
Per valors de *n* petits, t té més variabilitat reflectint més incertesa (relacionat amb que aproximem σ per s)

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

Interval de confiança de μ. Premisses

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

La premissa fonamental és que l'origen de la mostra sigui aleatori (v.a.i.i.d.)

A més a més:

- Si σ és coneguda, exigirem una de les condicions:
  - **X~N** → la combinació lineal de Normals és Normal ($\bar{X}$~N)
  - **Tenir una mostra "gran"** → Pel TCL , $\bar{X}$~N

- Si σ 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 "gran")** → Pel TCL , $\bar{X}$~N

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

Dist. de referència	σ coneguda	**σ desconeguda**
X Normal	Usar la Normal	Usar **t de *Student***
X no Normal i *n* "gran"		

4.b. Interval de confiança de π

- Sigui X ~ B(n,π) → E(X) = π·n
  V(X) = π·(1-π)·n
- Aleshores, 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 (com a guia comprovar que **π·n ≥ 5 i (1-π)·n ≥ 5**]:

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

L'estadístic $\frac{(P-\pi)}{se}$ per a un valor concret (π_0) del paràmetre, segueix una N: $z = \frac{(P-\pi_0)}{se} \sim N(0,1)$

- L'IC amb confiança 1-α és:

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

amb error estàndard $se = \sqrt{\pi(1-\pi)/n}$ $(\pi \to P\ o\ \pi \to 0.5)$

La **paradoxa** de que necessitem conèixer π per estimar el IC de π, es pot solucionar de 2 maneres:

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

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

Interval de confiança de σ^2

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

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

Per tant:

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

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

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

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

És un IC per σ^2, no per σ !!

No és un interval simètric. Implica calcular els dos quantils en lloc de sumar i restar a partir d'un sol quantil

4.b. Estimació per a IC comparant 2 paràmetres

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

- Comparar μ_1 i μ_2

  (per ex IC de l'efecte diferencial (μ_1-μ_2) comparant mitjanes entre dues assignatures)*

  Cal diferenciar entre:

  - **mostres aparellades**** (cada cas dóna lloc a dues mesures, parells de mesures)
    (els mateixos estudiants en les dues assignatures, μ_1-μ_2 = $\mu_{Diferència}$ = μ_D)
  - **mostres independents** (cada cas és una mesura independent)
    (estudiants diferents en les dues assignatures)

- Comparar π_1 i π_2

  (per ex IC de l'efecte diferencial (π_1-π_2) comparant aprovats entre dues assignatures)*

- Comparar σ^2_1 i σ^2_2

  (per ex IC de l'efecte diferencial (σ^2_1 / σ^2_2) comparant desviacions entre dues assignatures)*

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

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

IC de $\mu_1 - \mu_2$ (o de μ_D) en mostres aparellades

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

- L'estadístic $\frac{(\bar{d}-\mu_D)}{s_d/\sqrt{n}}$ que per a un valor concret (μ_0) del paràmetre, segueix la distribució

t-Student: $t = \frac{(\bar{d}-\mu_0)}{s_d/\sqrt{n}} = \frac{(\bar{d}-\mu_0)}{se} \sim t_{n-1}$

on $(\bar{d} - \mu_0)$ és el "senyal" i $\boldsymbol{se} = s_d/\sqrt{n}$ l'error estàndard

- L'IC de la diferencia amb confiança 1-α és:

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

Pot tenir interès pràctic avaluar el "t-rati": $\boldsymbol{t = \bar{d} / \frac{s_d}{\sqrt{n}} = \bar{d}/se}$
que diu quantes vegades és més gran el senyal que el soroll (assumint m.a. aparellada)

IC de $\mu_1 - \mu_2$ en mostres independents

Siguin Y_1 ($E(Y_1)=\mu_1$ $V(Y_1)=\sigma_1^2$) i Y_2 ($E(Y_2)=\mu_2$ $V(Y_2)=\sigma_2^2$) amb distribucions Normals (**σ_1 i σ_2 seran valors desconeguts però cal poder assumir-los iguals***)
de les que obtenim dues mostres aleatòries simples de grandària n_1 i n_2 **independents,** amb mitjanes $\bar{y}_1$, $\bar{y}_2$ i desviacions s_1, s_2 (com a estimadors d'un paràmetre comú σ)

- L'estadístic $\frac{(\bar{y}_1-\bar{y}_2)-(\mu_1-\mu_2)}{se}$ per a un valor concret ($\mu_1=\mu_2$ o $\mu_1-\mu_2=0$) de la diferència de paràmetres, segueix la distribució *t-Student*: $t = \frac{(\bar{y}_1-\bar{y}_2)-(0)}{se} \sim t_{n_1+n_2-2}$ amb error estàndard $\boldsymbol{se} = s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ on s és arrel de la variància "pooled" $s^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$

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

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

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

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

IC de $\pi_1 - \pi_2$

Siguin P_1 i P_2 les proporcions mostrals de 2 poblacions binomials amb π_1, π_2 de les que obtenim dues mostres aleatòries simples de grandària n_1 i n_2 i independents

- L'estadístic $\frac{(P_1-P_2)-(\pi_1-\pi_2)}{se}$ per a un valor concret ($\pi_1 = \pi_2$ o $\pi_1 - \pi_2 = 0$) de la diferència de paràmetres, segueix la distribució N(0,1): $z = \frac{(P_1-P_2)-(0)}{se} \sim \mathrm{N}(0,1)$

- L'IC de la diferència amb confiança 1-α és:

$$\boldsymbol{IC(\pi_1 - \pi_2, 1-\alpha) = (P_1 - P_2) \pm z_{1-\frac{\alpha}{2}} \cdot se =}$$

$$\boldsymbol{(P_1 - P_2) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{P_1(1-P_1)/n_1 + P_2(1-P_2)/n_2}}$$

amb error estàndard $\boldsymbol{se} = \sqrt{P_1(1-P_1)/n_1 + P_2(1-P_2)/n_2}$

En aquest cas, la convergència requereix mostres "grans": usualment $\pi_i n_i$ i $(1-\pi_i)n_i$ superiors a 5 amb $\pi_i \to Pi$

IC de σ_1^2/σ_2^2

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

- L'estadístic $\hat{F} = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2}$ segueix la distribució $F_{(n_1-1,n_2-1)}$
- L'IC del quocient de variàncies amb confiança 1-α és:
  (seguint el mateix raonament de l'IC de σ^2)

$$IC(\sigma^2{}_1/\sigma^2{}_2, 1-\alpha) = \left[\frac{s_1^2/s_2^2}{F_{(n_1-1,n_2-1),1-\frac{\alpha}{2}}} , \frac{s_1^2/s_2^2}{F_{(n_1-1,n_2-1),\frac{\alpha}{2}}} \right]$$

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

$$IC(\sigma^2{}_1/\sigma^2{}_2, 1-\alpha) = \left[s_1^2/s_2^2 \cdot F_{(n_2-1,n_1-1),\frac{\alpha}{2}} , s_1^2/s_2^2 \cdot F_{(n_2-1,n_1-1),1-\frac{\alpha}{2}} \right]$$

4.c. Funcions en R per a IC. Exemples

Comprovar la premissa de normalitat: *(a Anàlisis gràfica de la normalitat, annex de Bloc B)*

```
qqnorm(X)
qqline(X)
```

app

IC de μ amb σ coneguda (per aquesta funció cal la llibreria BSDA):

```
library(BSDA)
z.test(X,sigma.x=  )        # per a una mostra amb un valor conegut per a la sigma
```

IC de μ (o μ's) amb σ (o σ's) desconeguda:

```
t.test(X)                                   # per a una mostra
t.test(X-Y) o t.test(X,Y,paired=T)  # per a dues mostres aparellades
t.test(X,Y,var.equal=T)  # per a dues mostres independents amb variàncies iguals
t.test(X,Y,var.equal=F)  # per a dues mostres independents amb variàncies diferents
```

IC del rati de σ^2's en dues mostres independents:

```
var.test(X,Y)
```

IC de π:

`prop.test` (mostres grans) i `binom.test` (mostres petites)

Exemple d'una mostra. IC de μ

Exemple de 9 valors amb positius i negatius (mesures per sota o sobre un llindar)

```
X <- c(-4,-2,-1,0,0,4,8,8,9)   # mean=2.4 sd=4.9 Assumim normalitat (o qqnorm(X) i qqline(X))
library(BSDA)
z.test(X,sigma.x=4)            # IC suposant una σ poblacional de 4
```

```
z = 1.8333, p-value = 0.06675
  alternative hypothesis: true mean is not equal to 0
  95 percent confidence interval:
   -0.1688409  5.0577298
  sample estimates: mean of x  2.444444
```

```
t.test(X)                      # IC si no coneixem el valor poblacional de σ sinó que usem la s mostral
```

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

Es pot comprovar la coincidència dels límits dels IC amb els que es calcularien amb les fórmules

Apart dels IC, aquestes funcions en R ofereixen el resultat d'un p-value : probabilitat que indica si l'estadístic (z o t) usat per calcular l'IC, i avaluat en un valor a prova del paràmetre, és "extrem" o no en la distribució del model de referència (*veure més a l'annex i a bloc D amb més funcions que proporcionen p-values)*

Exemple de dues mostres independents. IC de μ_1-μ_2

Exemple de 2 mostres on comparar μ_1 i μ_2 amb l'IC de l'efecte diferencial (μ_1-μ_2)

```
X1 <-c(1,2,3,5,6,6,7,7,8,8,9)
```
mean=5.6 sd=2.62 Assumim normalitat (o qqnorm(X1) i qqline(X1))

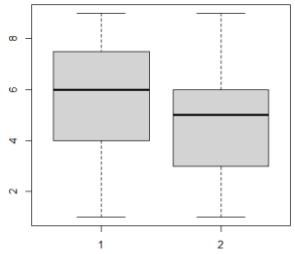
```
X2 <-c(1,1,3,4,5,5,6,7,9)
```
mean=4.5 sd=2.65 Assumim normalitat (o qqnorm(X2) i qqline(X2))

(per exemple X1 i X2 dues mostres de notes amb variabilitat equivalent)

```
t.test(X1,X2,var.equal=T)
```

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

```
boxplot(X1,X2)
```



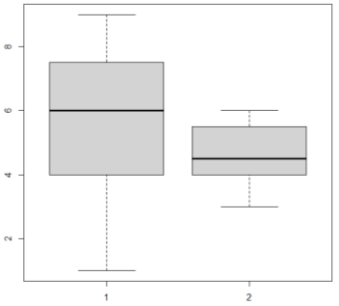
```
X3 <-c(3,4,4,4.5,4.5,5,6,6)
```
mean=4.6 sd=1.0 Assumim normalitat (o qqnorm(X3) i qqline(X3))

(per exemple X1 i X3 dues mostres de notes amb variabilitat no equivalents)

```
t.test(X1,X3,var.equal=F)
```

```
t = 1.1641, df = 13.793, p-value = 0.2641
  alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -0.8546781  2.8774054
  sample estimates: mean of x mean of y  5.636364  4.625000
```

```
boxplot(X1,X3)
```



Interpretació: la diferència de mitjanes és de fins a 0.85 punts a favor del grup 3 o de fins a 2.9 punts a favor del grup 1, amb una confiança del 95%. Les notes 1 i 3 o es diferencien clarament en mitjana

Exemple de dues mostres aparellades. IC de μ_D

Exemple de 2 mostres on comparar μ_1 i μ_2 amb l'IC de l'efecte diferencial (μ_1-μ_2)

```
Y1 <- c(1,1,2,2.0,2,2.5,4,5,5.5,6,7.5,8,8,9.5,9,9.5)
Y2 <- c(1.5,1,2,1.0,3,3,3.5,5,6,6,8.5,8.5,9.5,8.5,9.1,9)
```

(per exemple X1 i X2 dues mostres de notes d'uns mateixos estudiants)

```
t.test(Y1,Y2,paired=T)
```

```
t = -0.92936, df = 15, p-value = 0.3674
  alternative hypothesis: true difference in means is not equal to 0
  95 percent confidence interval:
   -0.5351864  0.2101864
  sample estimates: mean of the differences  -0.1625
```

En mostres aparellades, és pot treballar amb la diferència dels valors (D=Y1-Y2). Per tant és com el cas d'una mostra on enlloc de μ_1-μ_2 interessa μ_D (i equivaldria a fer `t.test(D)`)

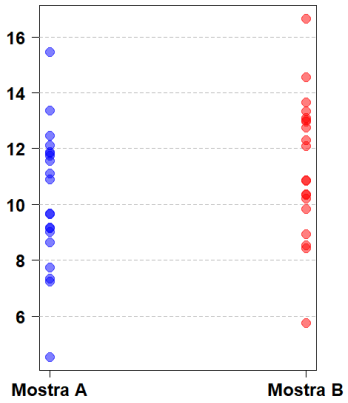
```
D <- Y1-Y2
 -0.5  0.0  0.0  1.0 -1.0 -0.5  0.5  0.0 -0.5  0.0 -1.0 -0.5 -1.5  1.0 -0.1  0.5
mean(D)
  -0.1625
sd(D)
  0.6994045
```

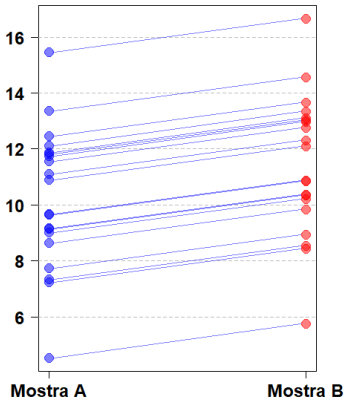
Exemple: no tractar aparellat com a idependent

És molt important no fer una anàlisi de dades aparellades com a dades independents

Aparentment, en el gràfic de l'esquerra (independer no veiem que pugui existir diferències en mitjana entre les dues poblacions

En el gràfic de la dreta (aparellat) es veu clarament que la mitjana és superior en la mostra B.

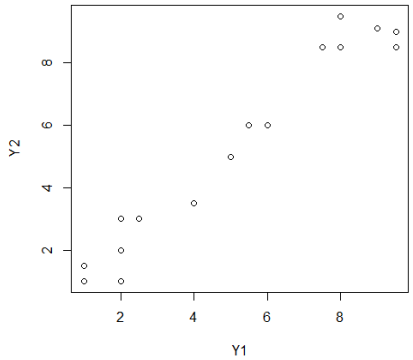




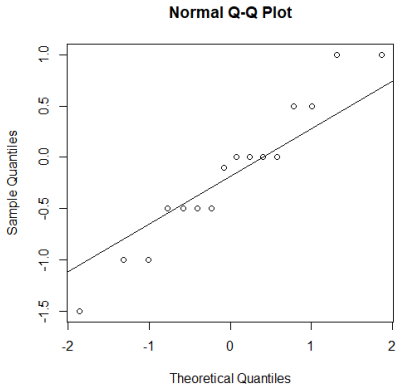
app

En mostres aparellades, podem veure la relació entre les dues mostres (gràfic esquerra) **i la normalitat de la diferència** (gràfic dreta), amb funcions R de descriptiva gràfica:

```
plot(Y1,Y2)
```




```
qqnorm(Y1-Y2)
qqline(Y1-Y2)
```



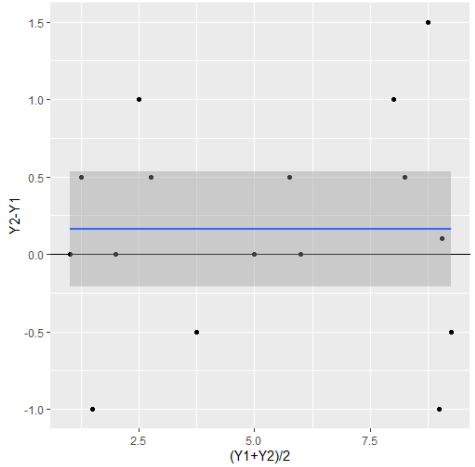
Però convé **relacionar les diferències amb les mitjanes** (com en el següent gràfic Bland-Altman)

Exemple de gràfic *diferències vs mitjanes* (aparellades)

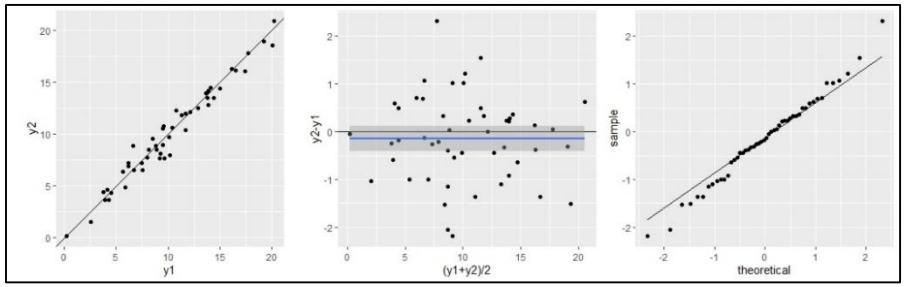
Hi ha funcions R específiques per mostres aparellades: **gràfic de Bland-Altman (BA),** que representa les diferències de les respostes per cada individu en funció de les seves mitjanes

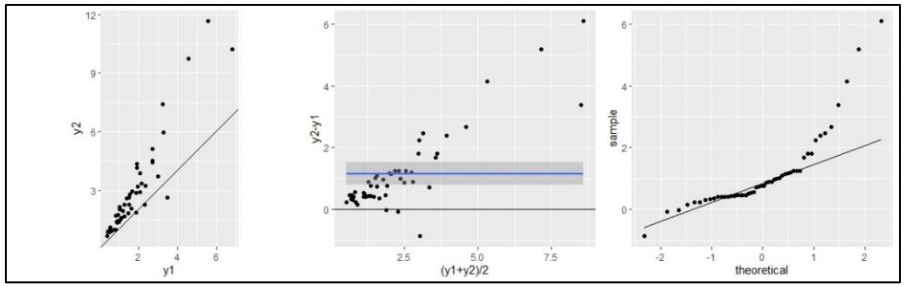
```
install.packages("PairedData")
library(PairedData)
p <- paired(Y1,Y2)
plot(p,type='BA')
```

(o bé `plot((Y1+Y2)/2,Y2-Y1)`)



El gràfic diferències en front de mitjanes (complementant el plot i el qqnorm) permet veure si hi ha efecte additiu (o multiplicatiu), i decidir si convindria una transformació a les dades (es veurà al bloc D)





Exemple de funcions R per comparar dues σ^2

Exemple de 2 mostres on comparar σ_1 i σ_2 amb l'IC de l'efecte diferencial (σ^2_1/σ^2_2)

(com el cas dels exercicis de comparar la variabilitat en la duració dels recanvis dels cartutxos de tinta de dues marques)

```
A <- c(350,361.9,365,365,365,370,372,377)
```

mean(A)=365.7375 sd(A)=8.00231 var(A)=64.03696 Assumim normalitat (o qqnorm(A i qqline(A))

```
B <- c(390,391.7,410,412,414,418)
```

mean(B)=405.95 sd(B)=12.00396 var(B)=144.095 Assumim normalitat (o qqnorm(B) i qqline(B))

```
var.test(B,A)
```

```
F test to compare two variances
data:  B and A
F = 2.2502, num df = 5, denom df = 7, p-value = 0.3199
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.4257491 15.4206862
sample estimates:
ratio of variances
          2.250185
```

Exemple de funcions R per a π

Exemple de llençar 100 vegades una moneda i observar 56 cares

```
prop.test(56,100)
```

```
 1-sample proportions test with continuity correction
data:  56 out of 100, null probability 0.5
X-squared = 1.21, df = 1, p-value = 0.2713
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4573588 0.6579781
sample estimates:  p   0.56
```

```
binom.test(56,100) # més apropiat si la mostra és petita
```

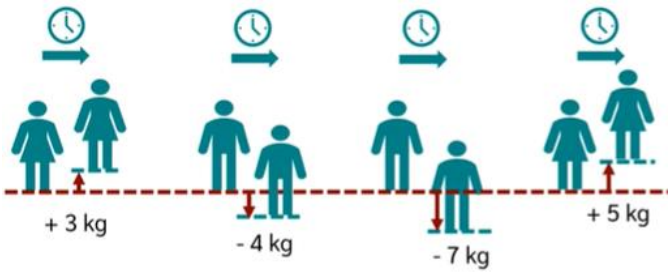
```
 Exact binomial test
data:  56 and 100
number of successes = 56, number of trials = 100, p-value = 0.2713
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4571875 0.6591640
sample estimates:
probability of success
                  0.56
```

Cap d'aquests IC coincideix *exactament* amb el calculat amb la fórmula aproximant a la Normal explicada anteriorment. La coincidència augmentaria amb grandàries mostrals majors i amb proporcions més properes a ½.

5. Bloc T: disseny (com obtenim les dades)

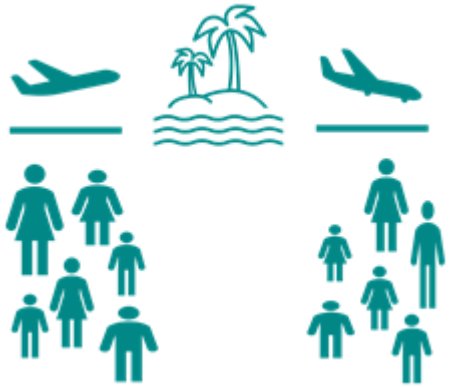
Disseny aparellat:

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



Mostres independents

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



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

La clau es "atzar":

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

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

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

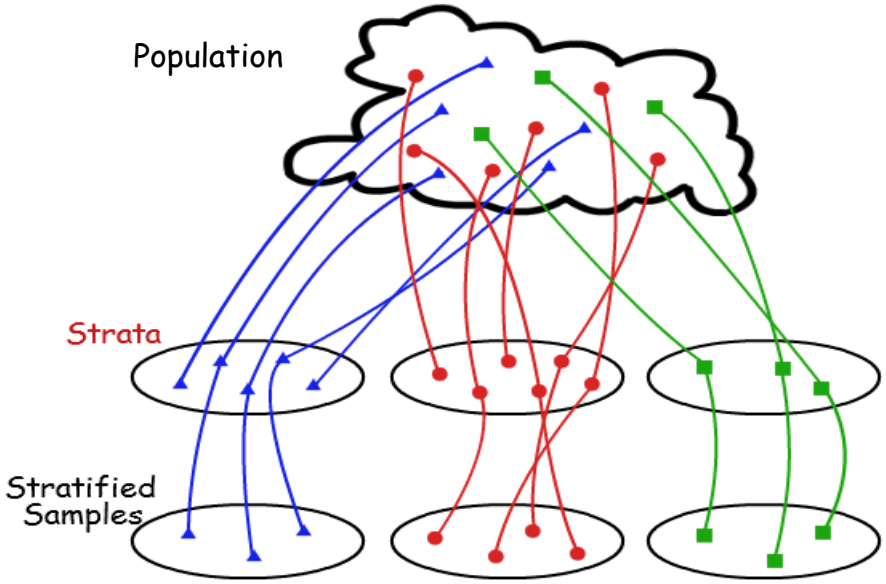
Dissenys

El disseny escollit condiciona l'anàlisi estadística posterior.

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

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

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



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

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

6. Bloc T: estadística descriptiva

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

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

	UNIVARIANT (num)	UNIVARIANT (categ)	BIVARIANT (num,num)
INDICADORS	length() * **mean( )** **var( )** **sd( )** **summary( )*** **median( )**	**table( )** **prop.table()**	**cov(,)** **cor(,)**
GRÀFIQUES	hist( ) boxplot( )	barplot(table( ))	plot(,)

* La mida de la mostra (length() -> *n*) no és un estimador, però pot ser una funció úitl per descriure conjunts de dades. I summary() és una funció que presenta diversos estimadors resumidament

Més funcions gràfiques en R: https://www.r-graph-gallery.com/

Més informació de les funcions i exemples de descriptiva usant R a https://www-eio.upc.es/teaching/pe/R/ED.pdf