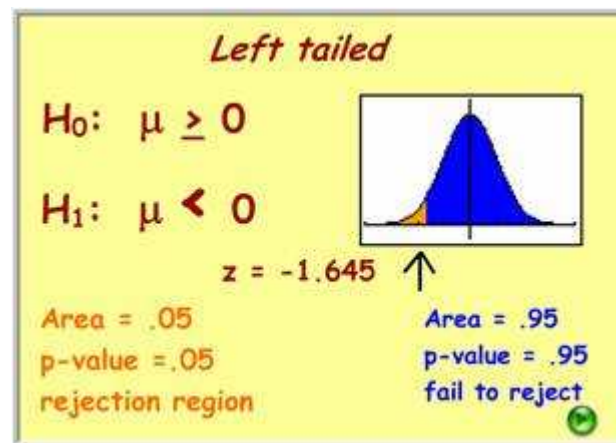
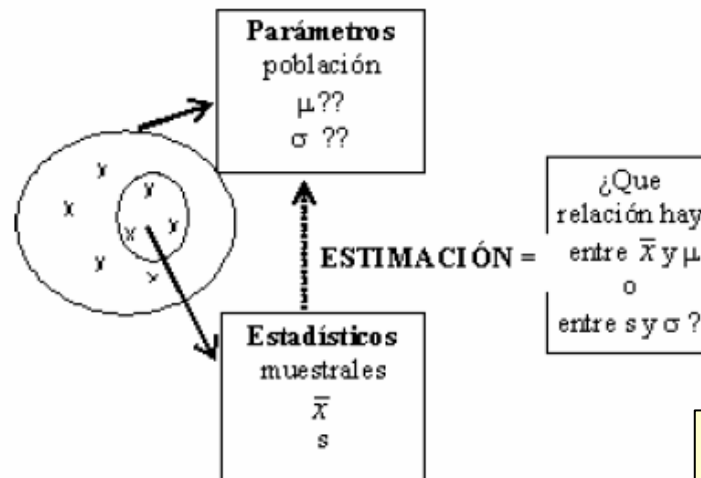


ESTADÍSTICA APLICADA

T6 Inferència



- Objectiu: calcular determinats paràmetres d'una població
- Problema: generalment, no és possible mesurar cadascun dels individus d'una població degut a: nombre d'individus, temps disponible, cost, prova destructiva, ...
- Solució: treballar amb una *mostra aleatòria* i representativa, X_1, X_2, \dots, X_n , i *estimar* els paràmetres poblacionals a partir dels estadístics



Obs.: la grandària de la mostra, n , jugarà un paper important la l'hora de fer inferència

- Sigui X una v.a. i θ un paràmetre (desconegut) associat a X
 → representarem per $\tilde{\theta}$ a tot **estimador** de θ :

paràmetre (població) θ : $\mu \quad \mu \quad \dots \quad \sigma \quad \sigma \quad \dots$
 $\downarrow \quad \downarrow \quad \quad \downarrow \quad \downarrow$

estimador (mostra) $\tilde{\theta}$: $\bar{x} \quad Q_2 \quad \dots \quad s_n \quad s_{n-1} \quad \dots$

Obs.: un estimador és una v.a.

Obs.: un paràmetre pot tenir múltiples estimadors, quin és "el millor"?

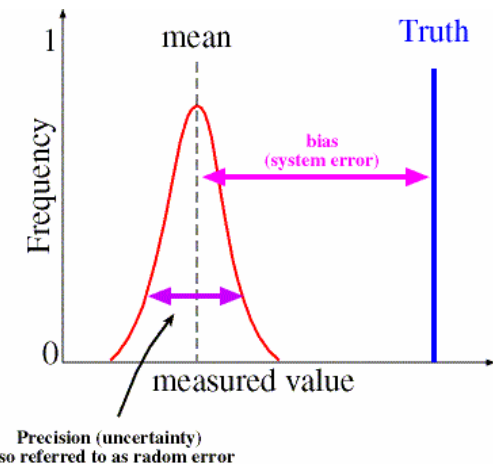
- Propietats (desitjables) dels estimadors:

- $\tilde{\theta}$ és centrat o **sense biaix** si $E[\tilde{\theta}] = \theta$

- $\tilde{\theta}$ és **consistent** si $\lim_{n \rightarrow \infty} E[(\tilde{\theta} - \theta)^2] = 0$

Error d'estimació al quadrat

- $\tilde{\theta}_1$ és **més eficient** que $\tilde{\theta}_2$ si $E[(\tilde{\theta}_1 - \theta)^2] \leq E[(\tilde{\theta}_2 - \theta)^2]$



- Un estimador no esbiaixat i consistent de la mitjana poblacional, μ , ve donat per la **mitjana mostral**:

$$\bar{X}_n = \frac{1}{n} \sum X_i$$

- Teorema (distribució mostral de la mitjana mostral):**

$$(i) \quad \mu_{\bar{X}_n} = \mu_X \quad \sigma_{\bar{X}_n} = \frac{\sigma_X}{\sqrt{n}}$$

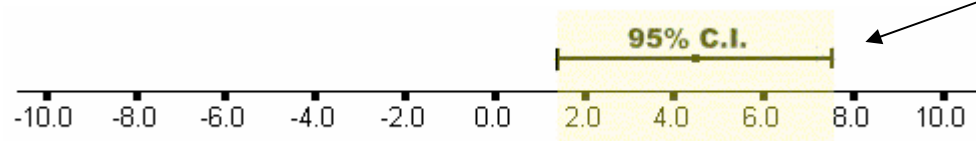
$$(ii) \quad \text{Si } X \sim N(\mu, \sigma) \Rightarrow \bar{X}_n \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Si X és normal \rightarrow la mitjana mostral també ho serà!

- Un estimador no esbiaixat i consistent de la variància poblacional, σ^2 , ve donat per la **variància mostral corregida**, s_{n-1}^2

Obs.: la variància mostral, s_n^2 , és un estimador esbiaixat!

- Quan es dona un valor concret (p.e.: $\tilde{\theta}_0 = 3.5$) com a estimació del valor real del paràmetre θ , s'està fent una **estimació puntual**
- Problema: l'estimació puntual no ens proporciona cap informació sobre l'error que s'està cometent a l'estimar el valor real θ per $\tilde{\theta}$
- L'**estimació per interval** consisteix a donar un interval $]r_1, r_2[$ tal que $P(r_1 < \theta < r_2) = 1 - \alpha$
- $(1 - \alpha)$ s'anomena **nivell de confiança** i α s'anomena **nivell de significació** (habitualment, $\alpha = 0.05$, $\alpha = 0.01$, ó $\alpha = 0.001$)



En l'exemple, θ estarà en l'interval $]1.3, 7.7[$ amb probabilitat 0.95

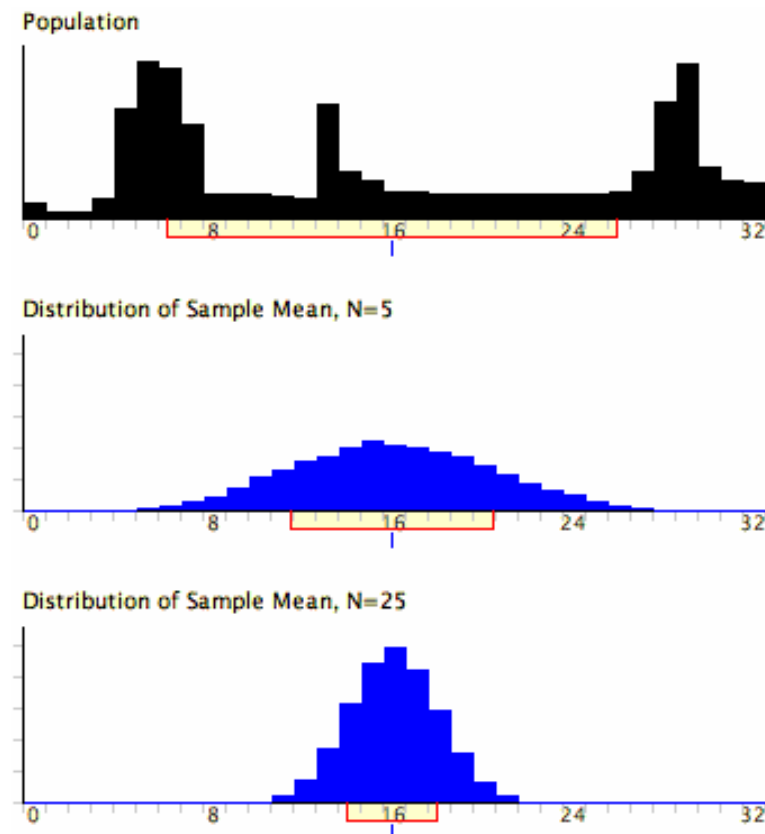
- Segons el Tma. de la distribució mostral de la mitjana mostral:

$$\text{Si } X \sim N(\mu, \sigma) \Rightarrow \bar{X}_n \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Però... què passa si X no és normal?

- Tma. Central del Límit (TCL):** encara que X no es distribueixi segons una normal, si n “és gran” ($n \gg 30$), llavors la mitjana mostral serà normal:

$$\text{Si } n \gg 30 \Rightarrow \bar{X}_n \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$



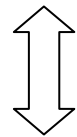
T6 – 6.7: Aplicació del TCL als IC

- Si es satisfà la **condició clau**: $\bar{X}_n \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$
- Lavors: $Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- I, per tant:

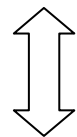
Recordar que si **X** era **normal**, aquesta condició sempre serà certa. Fins i tot si X no és normal, la condició segueix sent certa quan **n** "és gran" (**TCL**)

z_α és aquell valor que en una $N(0,1)$ deixa a la seva **dreta** un àrea de valor $\alpha/2$

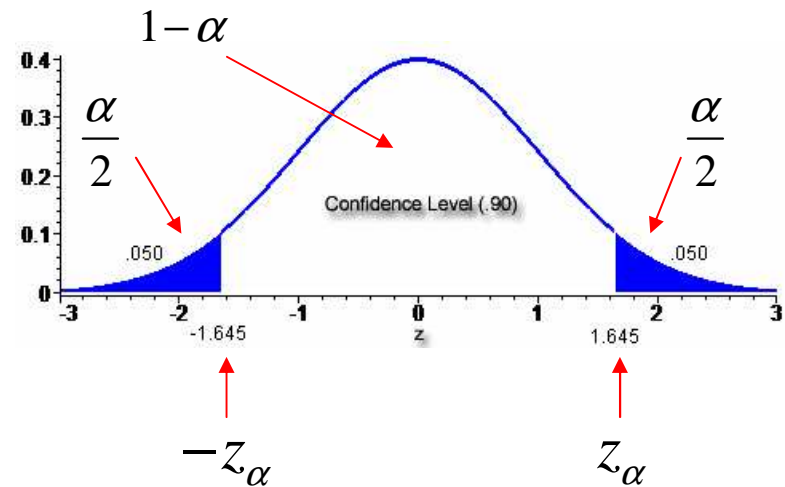
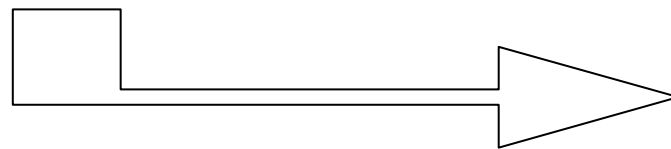
$$P(-z_\alpha < Z < z_\alpha) = 1 - \alpha$$



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



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



$$\bar{X}_n \pm z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$$

IC per a μ a nivell de confiança de **(1- α)**

T6 – 6.8 i 6.10: Int. de Confiança per a μ i p

MA1

ICs per a μ :

z_α és aquell valor que en una $N(0,1)$ deixa a la seva dreta un àrea de valor $\alpha/2$

t_α és aquell valor que en una t-Student amb $(n-1)$ graus de llibertat deixa a la seva dreta un àrea de valor $\alpha/2$

- Condició prèvia: X normal o bé $n \gg 30$ (TCL)

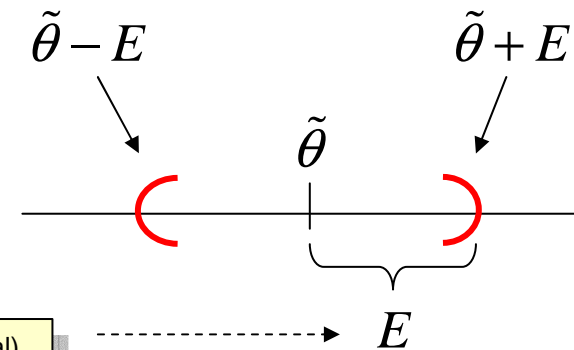
- σ coneguda \rightarrow

$$\bar{X}_n \pm z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$$

- σ desconeguda \rightarrow

$$\bar{X}_n \pm t_\alpha \cdot \frac{s_{n-1}}{\sqrt{n}}$$

Error màxim d'estimació, E (amplària de l'interval)



ICs per a la proporció d'èxits, p , en una Binomial:

- $p_0 = \#$ èxits observats / $\#$ proves

- p desconegut, dues opcions:

a) aproximar-lo per p_0

b) agafar $p = 0.5$ (màxima indeterminació)

$$p_0 \pm z_\alpha \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$$

- **Problema 1 (pàg. 192):** $n = 250$ $\bar{x} = 0.72642$ $s_n = 0.00058$

99% CI per a μ ?

$n \gg 30 \Rightarrow TCL$

$$s_{n-1} = \sqrt{\frac{n}{n-1}} \cdot s_n = 0.0005812$$

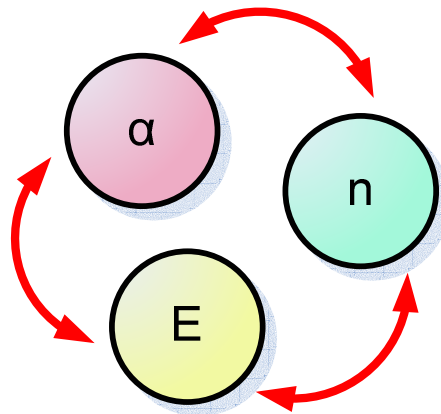
$\sigma ? \Rightarrow t(n-1)$

$1 - \alpha = 0.99 \Rightarrow \alpha = 0.01$

$$t_\alpha(n-1) = \{taules\} = 2.576$$

$$\bar{X}_n \pm t_\alpha \cdot \frac{s_{n-1}}{\sqrt{n}} \Rightarrow 0.72642 \pm 2.576 * 3.68E - 5 \Rightarrow (0.726325, 0.726515)$$

- Obs.: **Error, n i α estan “lligats”**, és a dir:
 - Si α decreix (volem més nivell de confiança) \rightarrow n creix (calen més dades per poder afinar més) ó E creix (cal un interval més ample)
 - Si E decreix (volem menys error) \rightarrow n creix (més dades per poder ser més precisos) o α creix (ens conformem amb menys nivell de confiança)
 - Si n decreix (tenim menys dades) \rightarrow E creix (cal un interval més ample) o α creix (cal conformar-nos amb menys nivell de confiança)
 - I viceversa...



T6 – 6.12: Contrast d'hipòtesi: Introducció

MA1

- Contrast d'hipòtesi: es tracta de contrastar la veracitat d'una afirmació o **hipòtesi nul·la** (H_0), la qual fa referència al valor d'un paràmetre θ de la població

Contrast bilateral

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

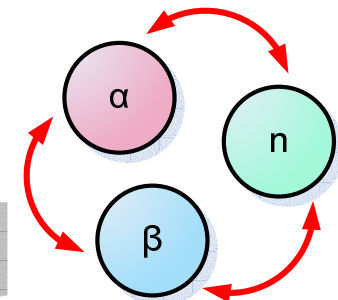
H_0 és la **hipòtesi nul·la**, que donarem per bona fins que es demostrï el contrari (tothom és innocent fins que...). H_1 és la **hipòtesi alternativa**

Contrast unilateral

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta < \theta_0 \quad (\text{ó } \theta > \theta_0) \end{cases}$$

- Idea del contrast: comparar el que hauria de ser (si H_0 fos certa) amb el que realment diuen les dades, i.e.:
 - (i) agafar una mostra aleatòria, (ii) calcular un **estadístic de contrast** a partir de la mostra, (iii) determinar quant probable era haver obtingut aquest valor de l'estadístic si H_0 fos certa, i (iv) establir conclusions sobre la certesa de H_0
- Les conclusions d'un contrast no seran mai 100% segures \rightarrow 2 tipus d'errors:
 - **Error tipus I (molt greu!)**: rebutjar H_0 / H_0 certa $\rightarrow \alpha = P(\text{error tipus I})$
 - **Error tipus II**: acceptar H_0 / H_0 falsa $\rightarrow \beta = P(\text{error tipus II})$

α , β i n estan lligats!: si fem decreïxer α llavors β creixerà (a menys que n augmenti) i viceversa



T6 – 6.12: Contrast d'hipòtesi sobre μ i p

MA1

- Condició prèvia: X és normal o bé $n \gg 30$ (per poder aplicar el TCL)
- Objectiu: fixat α , es tracta de fer un contrast (bilateral o unilateral) amb $H_0: \mu = \mu_0$ (sobre la mitjana poblacional), ó $H_0: p = p_0$ (sobre la proporció d'èxits)
- Metodologia de resolució:

- Calcular el corresponent **estadístic de contrast**:

μ (σ coneguda)

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

μ (σ desconeguda)

$$t^* = \frac{\bar{x} - \mu}{\frac{s_{n-1}}{\sqrt{n}}} \sim t(n-1)$$

p

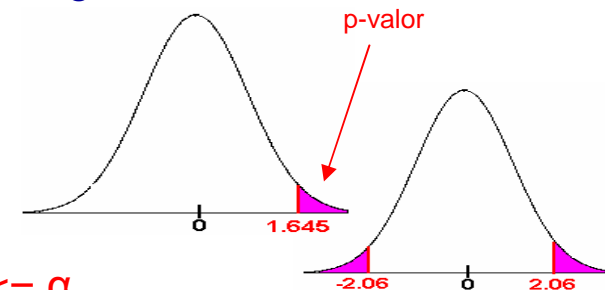
$$z^* = \frac{p_0 - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

- Calcular el **p-valor**, i.e.: la probabilitat que, essent H_0 certa, es pugui donar un estadístic de contrast tant “extrem” com el que s’ha obtingut:

contrast unilateral (1 cua) $p\text{-valor} = P(Z > |z^*|)$

contrast bilateral (2 cues) $p\text{-valor} = 2 \cdot P(Z > |z^*|)$

Un p-valor “gran” s’ha d’entendre com “és creïble que la H_0 sigui certa”



- Aplicar la “regla de decisió”: **rebutjar $H_0 \iff p\text{-valor} \leq \alpha$**
- Concloure si, al nivell de significació α prefixat, hi ha indicis raonables per a rebutjar H_0 o si, pel contrari, les dades no aporten evidències significatives contra H_0

T6 – 6.12: Contrast d'hipòtesi sobre μ i p

MA1

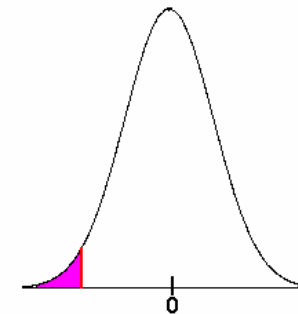
- **Problema 7 (pàg. 193):** $n = 6$ $\bar{x} = 7750$ $s_n = 145$ X és normal

Contrastar, per a $\alpha = 0.05$ i $\alpha = 0.01$: $\begin{cases} H_0 : \mu = 8000 \\ H_1 : \mu < 8000 \end{cases}$

$\sigma ? \Rightarrow t(n-1)$

$$s_{n-1} = \sqrt{\frac{n}{n-1}} \cdot s_n = 158.8395417$$

$$t^* = \frac{\bar{x} - \mu}{\frac{s_{n-1}}{\sqrt{n}}} \sim t(n-1) \Rightarrow t^* = \{taules\} = -3.8553 \sim t(5)$$



$$p\text{-valor} = P(T > |t^*|) = P(T > |3.8553|) \approx \{taules\} \approx 0.0075$$

$$p\text{-valor} = 0.0075 \leq \alpha_1 = 0.01 \leq \alpha_2 = 0.05 \Leftrightarrow \text{rebutjar } H_0$$

Hi ha evidències clares per a rebutjar la hipòtesi nul·la, tant per al cas $\alpha = 0.05$ com per al cas $\alpha = 0.01$