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

Bases de l'estadística
Exemples i exercicis

Bloc C – Probabilitat i Estadística

2024

Índex

Exemple. Nombre de terminals (assumint premissa de normalitat)

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

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

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

`mean(X)` → $\bar{x} = 553.44$ o $\bar{x} = (\sum x_i)/n = \frac{4981}{9} = 553.44$

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

`sd(X)` → $s = 114.0988$ o $s = \sqrt{(2860855 - (4981)^2/9)/(8)} = 114.0988$

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

`sd(X)/sqrt(length(X))` → $se = 38.03$ $se = \frac{s}{\sqrt{n}} = \frac{114.0988}{3} = 38.03$

** Quant valdria l'error tipus (o estàndard error) si el mateix valor de mitjana i de desviació provinguessin de n=100 valors:* $\frac{s}{\sqrt{n}} = \frac{114.0988}{10} = 11.4$

Exemple. Nombre de terminals (assumint premissa de normalitat)

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

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

1-α	σ	IC(μ,1-α)	Resolució amb R

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

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

95%	Desconeguda	[465.74; 641.15]	`mean(X)- qt(0.975,n-1)*sd(X)/sqrt(n)` `mean(X)+ qt(0.975,n-1)*sd(X)/sqrt(n)`
99%	Desconeguda	[425.83 ; 681.06]	`mean(X)- qt(0.995,n-1)*sd(X)/sqrt(n)` `mean(X)+ qt(0.995,n-1)*sd(X)/sqrt(n)`

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

Exercici. Mesures en cm (assumint premissa de normalitat)

Es prenen unes mesures en cm, amb una primera mostra y1 de 4 valors: 90, 100, 100, 110
($\Sigma y_1=400$ $\Sigma y_1^2=40200$) (En R: y1 <- c(90,100,100,110))

Donar una estimació puntual per a la mesura mitjana i un IC per aquesta mitjana al 95%:

$\overline{y_1}$ =100 cm ((90+100+100+110) / 4 o $\Sigma y_1 / 4 = 400 / 4$ o mean(y1) en R)

s1= 8.16 cm (sqrt(66.67) = sqrt($((90-100)^2+(100-100)^2+(100-100)^2+(110-100)^2)/3$)

o bé sqrt($(\Sigma y1^2 - (\Sigma y1)^2/4) / 3) = sqrt((40200-400^2/4) / 3)$

IC(μ1, 0.95)= $\bar{y}1 \pm t_{3,1-\alpha/2} \, s1/\sqrt{n1} = 100 \pm 3.1824 \cdot 4.0825 = 100 \pm 12.99 = [87.01, 112.99]$

Repetir per una altra mostra y2 de 16 valors: ($\Sigma y_2=1600$ $\Sigma y_2^2=161000$)

(En R: y2 <- c(90,90,90,90,90,100,100,100,100,100,100,110,110,110,110,110))

$\overline{y_2}$=100 cm ($\Sigma y_2 / 16 = 1600 / 16$ o mean(y2) en R)

s2= 8.16 cm (sqrt(66.67) = sqrt($(\Sigma y2^2 - (\Sigma y2)^2/16) / 15) = sqrt((161000-1600^2/16) / 15)$

IC(μ2, 0.95)= $\bar{y}2 \pm t_{15,1-\alpha/2} \, s2/\sqrt{n2} = 100 \pm 2.131 \cdot 2.04 = 100 \pm 4.35 = [95.65, 104.35]$

** Observeu els IC segons n: si n augmenta la precisió dels IC també augmenta (interval més estret)*

Exemple. Embotelladora (assumint premissa de normalitat)

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.

Calculem un interval de confiança del 95% de μ.

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

(l'amplada de l'interval és 1.96*2=3.92)

(amb una confiança del 95%, μ es troba entre 993.04 i 996.96, per sota dels 1000cc esperables)

* Si volguéssim una amplada de l'interval de 3.5 (meitat de l'interval = 1.75):

quina hauria de ser n sense canviar la confiança?

1.75 = $1.96 \cdot \frac{10}{\sqrt{n}}$ → n=125.44 (mínim una n de 126)

quin hauria de ser el risc i la confiança sense canviar la n?

1.75 = $z_{1-\frac{\alpha}{2}} \cdot \frac{10}{\sqrt{100}}$ → 1-α/2 = 0.9599 -> α=8% (risc del 8%,confiança del 92%)

Exercici. Glicèmia (assumint premissa de normalitat)

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).

$$IC(\mu, 0.95) = \bar{x} \mp z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 5 \mp 1.96 \cdot \frac{1}{\sqrt{9}} = 5 \mp 0.653 = [4.35\ , 5.65]$$

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?

$$IC(\mu, 0.95) = \bar{x} \mp z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \rightarrow 1.96 \cdot \frac{1}{\sqrt{n}} = \frac{0.653}{2} \rightarrow n \approx 36 \rightarrow n\ ha\ de\ ser\ 4\ vegades\ major$$

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

$$IC(\mu, 0.99) = \bar{x} \mp z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 5 \mp 2.576 \cdot \frac{1}{\sqrt{9}} = 5 \mp 0.857 = [4.14\ , 5.86]$$

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*

Exemple. Moneda (IC per π)

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

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

Exemple. Execucions (IC per σ) (assumint premissa de normalitat)

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

Exemple. Diferència temps mitjà en mostres aparellades

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

							Mean	Variances	Var. "pooled"
A	23.05	39.06	21.72	24.47	28.56	27.58	**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 amb s pooled $\sqrt{42.009}$=6.48):

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

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

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

$$\bar{d} = 2.946 \qquad s_D = \sqrt{0.996} = 0.998 \qquad se = \frac{s_D}{\sqrt{n}} = \frac{0.998}{\sqrt{6}} = 0.41$$

$$\boldsymbol{IC(\mu_D, 0.95)} = \bar{d} \mp qt(0.975,5) \cdot se = [2.946 \mp 2{,}571 \cdot 0.41 = [\mathbf{1.89, 4.0}]$$

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

Exemple. Diferència temps mitjà en mostres aparellades

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

- Numerador: La **mitjana de les diferencies** (2.946) coincideix amb la **diferència de les mitjanes**. Per tant, el numerador és el mateix en les 2 solucions
- Denominador: 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 (senyal) és el mateix, el **denominador** (soroll) **és molt inferior en la solució aparellada**
- Recorda: **el control de les condicions** (en aquest cas, la variabilitat dels bancs de dades) **augmenta l'eficiència** per trobar diferències

IC i estadístic:

- Càlcul de l'estadístic $t = \frac{(\bar{d}-\theta_0)}{se} \sim t_{n-1}$ $\quad t = \frac{(2.946-0)}{0.41} = 7.185$
- L'estadístic avaluat per un valor de 0 pel paràmetre ($\mu_D = \mu_0 = 0$ indicant una diferència nul·la) dona un valor de 7.185, que en la distribució t_5 (95% entre -2.57 i 2.57) seria un valor extrem; per tant és **poc versemblant que no hi hagi diferència** en mitjana entre els programes A i B i l'**IC quantifica la diferència** (entre 1.89 i 4 unitats)

Exercici. Diferència temps mitjà compressors (aparellat)

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, per saber si triguen el mateix, podem quantificar per interval de confiança la diferència mitjana de temps, i plantejar-se si és raonable pensar que els dos compressors tarden el mateix en mitjana

Estadístic: $\hat{t} = (\bar{d} - \mu)/(s/\sqrt{n}) = \frac{6.71-0}{6/\sqrt{9}} = 3.355$

IC(μ, 0.95): $\bar{x} \mp t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} = 6.71 \mp 2.306 \cdot \frac{6}{\sqrt{9}} = [2.1\,, 11.3]$

Un dels programes triga en mitjana entre 2.1 i 11.3 unitats de temps més, amb una confiança del 95%. Per tant, no triguen el mateix

Si ens plantegem saber si és raonable pensar que els dos compressors tarden el mateix en mitjana, podríem contrastar si és versemblant una diferència nul·la de les mitjanes poblacionals (μ_D=0=μ_1-μ_2 o bé μ_1=μ_2), amb confiança 95%: com que el valor 0 no pertany al IC, no és versemblant la igualtat de mitjanes poblacionals d'acord amb l'evidència empírica que les dades aporten

Exemple. Diferència temps mitjà en mostres independents

Els temps mitjans d'execució de dos programes provats en diferents bancs de dades independents (n_1=50 i n_2=100) són: $\bar{y}_1 = 24$ *i* $\bar{y}_2 = 21$ amb s_1= 8 i s_2= 6. Assumint normalitat i suposant igualtat de variàncies ($\sigma^2_1 = \sigma^2_2$), es desitja estimar si tenen rendiments diferents

Estadístic: $\hat{t} = \frac{(\bar{y}_1 - \bar{y}_2)}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ amb $s^2 = \frac{(n_1-1)\cdot s_1^2+(n_2-1)\cdot s_2^2}{(n_1-1)+(n_2-1)} = \frac{(50-1)\cdot s_1^2+(100-1)\cdot s_2^2}{(50-1)+(100-1)} = 45.27$

$$s=6.72 \quad se= 6.72\sqrt{\frac{1}{50}+\frac{1}{100}} = 1.16$$

$$\boldsymbol{IC(\mu_1 - \mu_2, 95\%)} = (\bar{y}_1 - \bar{y}_2) \mp t_{n_1+n_2-2,1-\frac{\alpha}{2}} \cdot s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = 3 \mp 1.976 \cdot 1.16 = [\mathbf{0.70}, \mathbf{5.30}]$$

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

Si es contrasta una diferència nul·la de les mitjanes poblacionals (μ_1-μ_2=0 o bé μ_1=μ_2), amb confiança 95%: com que el valor 0 no pertany al IC, no és versemblant la igualtat de mitjanes poblacionals d'acord amb l'evidència empírica que les dades aporten.

Ex. de transparència 4 amb funcions específiques de R

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

1-α	σ	IC(μ,1-α) amb fórmules	Resolució amb funcions de R

95%	Coneguda (σ=100)	[488.11; 618.78]	`> z.test(X,sigma.x=100)`
			`z = 16.603, p-value < 2.2e-16`
			`alternative hypothesis: true mean is not equal to 0`
			`95 percent confidence interval:`
			`488.1123 618.7766`
			`sample estimates: mean of x 553.4444`

99%	Coneguda (σ=100)	[467.58 ; 639.31]	`> z.test(X,sigma.x=100,conf.level=0.99)`
			`z = 16.603, p-value < 2.2e-16`
			`alternative hypothesis: true mean is not equal to 0`
			`99 percent confidence interval:`
			`467.5835 639.3054`
			`sample estimates: mean of x 553.4444`

95%	Desconeguda	[465.74; 641.15]	`> t.test(X)`
			`t = 14.552, df = 8, p-value = 4.874e-07`
			`alternative hypothesis: true mean is not equal to 0`
			`95 percent confidence interval:`
			`465.7404 641.1485`
			`sample estimates: mean of x 553.4444`

99%	Desconeguda	[425.83 ; 681.06]	`> t.test(X,conf.level=0.99)`
			`t = 14.552, df = 8, p-value = 4.874e-07`
			`alternative hypothesis: true mean is not equal to 0`
			`99 percent confidence interval:`
			`425.8293 681.0596`
			`sample estimates: mean of x 553.4444`

En els 4 casos veiem un valor "alt" de l'estadístic (***z*** o ***t***) que en la distribució corresponent (N(0,1) o t_8) no és un valor central sinó extrem (amb p-value "petit") indicant que **no és versemblant** el valor a prova, que per defecte és 0