Estimació puntual

Estadística inferencial

El problema típic de l'estadística inferencial és:

- Volem conèixer el valor d'una característica en el global d'una població
- No podem mesurar aquesta característica en tots els individus de la població
- Extraiem una mostra aleatòria de la població, mesuram la característica en els individus d'aquesta mostra i inferim el valor de la característica en el global de la població
  - Com ho hem de fer?
  - Com ha de ser la mostra?
  - Quina informació podem inferir de veritat?

Exemples

Set de cada deu estudiants de la UIB practica el ciberplagi a l'hora de confeccionar els treballs acadèmics

Exemples

Set de cada deu estudiants de la UIB practica el ciberplagi a l'hora de confeccionar els treballs acadèmics

Fixa tècnica de la mostra de la UIB

Univers: alumnat de primer i segon cicle de la UIB (N = 11.797 estudiants)
Punts de mostreig: 38 unitats/aules (una per cada estudi oficial)
Mostreig: mixt i polietàpic, estratificat per centres amb selecció de les unitats primàries (assignatures) de forma aleatòria amb afixació proporcional i de les unitats secundàries (alumnes) mitjançant mostreig incidental a l'aula.
Mostra: 727 unitats d'anàlisi (qüestionaris), amb un error per al conjunt de la mostra del 3,52 per cent estimat per a un nivell de confiança del 95 per cent i sota la condició més desfavorable de p = q = 0.05.

Exemples

Evaluación de la efectividad de una nueva vacuna contra la leptospirosis humana en grupos en riesgo

RESUMEN *Se realizó un estudio de cohorte prospectivo cuasi experimental que incluyó a los grupos en riesgo de enfermar de leptospirosis en la provincia de Holguín para evaluar la efectividad de la vacuna contra la leptospirosis humana. Se incluyeron 118 018 personas de 15 a 65 años que presentaban un riesgo permanente o temporal de contraer la enfermedad; de estas, 101 137 fueron inmunizadas con dos dosis de 0,5 mL por vía intramuscular profunda en el músculo deltoides del brazo no dominante, con un intervalo de 6 semanas, constituyendo la cohorte de vacunados, mientras que 16 881 personas no inmunizadas pasaron a integrar la cohorte de no vacunados. A los 21 días de aplicada la segunda dosis, el universo de estudio (previamente censado en un registro de modelo) fue seguido por el sistema local de vigilancia epidemiológica con el objetivo de detectar la enfermedad. El criterio de caso sospechoso y confirmado se conservó*

http://www.scielosp.org/pdf/rpsp/v8n6/3956.pdf

Definicions bàsiques

Mostra aleatòria simple (m.a.s.) de mida n: D'una població de N individus, repetim n vegades el procés d'escollir equiprobablement un individu; *els individus triats es poden repetir*

Exemple: Escollim a l'atzar n estudiants de la UIB (amb reposició) per midar-los l'alçada

D'aquesta manera, totes les mostres possibles de n individus (possiblement repetits: multiconjunts) tenen la mateixa probabilitat

Llegiu-vos aviat la lliçó 1 de R sobre Mostreig

Definicions bàsiques

Estadístic (Estimador puntual): Una funció que aplicada a una mostra ens permet estimar un valor que vulguem saber de tota la població

Exemple: La mitjana de les alçades d'una mostra d'estudiants de la UIB ens permet estimar la mitjana de les alçades de tots els estudiants de la UIB

Formalment

Una m.a.s. de mida n (d'una v.a. X) és

- un conjunt de n còpies independents de X, o
- un conjunt de n variables aleatòries independents $X_1, \ldots, X_n$, totes amb la distribució de X

Exemple: Sigui X la v.a. "triam un estudiant de la UIB i li mesuram l'alçada". Una m.a.s. de X de mida n seran n còpies independents $X_1, \ldots, X_n$ d'aquesta X.

Una realització d'una m.a.s. són els n valors $x_1, \ldots, x_n$ que prenen les v.a. $X_1, \ldots, X_n$

Formalment

Un estadístic T és una funció aplicada a la mostra $X_1, \ldots, X_n$:

$$T = f(X_1, \ldots, X_n)$$

Aquest estadístic s'aplica a les realitzacions de la mostra

Exemple: La mitjana mostral de una m.a.s. $X_1, \ldots, X_n$ de mida n és

$$\overline{X} := \frac{X_1 + \cdots + X_n}{n}$$

Estima $E(X)$

Exemple: La mitjana mostral de les alçades d'una realització d'una m.a.s. d'alçades d'estudiants estima l'alçada mitjana d'un estudiant de la UIB

Formalment

Un estadístic T és una funció aplicada a la mostra $X_1, \ldots, X_n$

$$T = f(X_1, \ldots, X_n)$$

Per tant, un estadístic és una (nova) variable aleatòria, amb distribució, esperança, etc.

La distribució mostral de T és la distribució d'aquesta variable aleatòria

Del coneixement d'aquesta distribució mostral, podrem estimar propietats de X a partir de les d'una mostra

Error estàndard de T: desviació típica de T

Conveni

ELS ESTADÍSTICS, EN MAJÚSCULES; les realitzacions, en minúscules

Exemple:

- $X_1, \ldots, X_n$ una m.a.s. i

$$\overline{X} := \frac{X_1 + \cdots + X_n}{n}$$

la mitjana mostral

- $x_1, \ldots, x_n$ una realització d'aquesta m.a.s. i

$$\overline{x} := \frac{x_1 + \cdots + x_n}{n}$$

la mitjana (mostral) d'aquesta realització

La vida real

A la vida real, les mostres aleatòries se solen prendre sense reposició (sense repeticions). No són mostres aleatòries simples. Però:

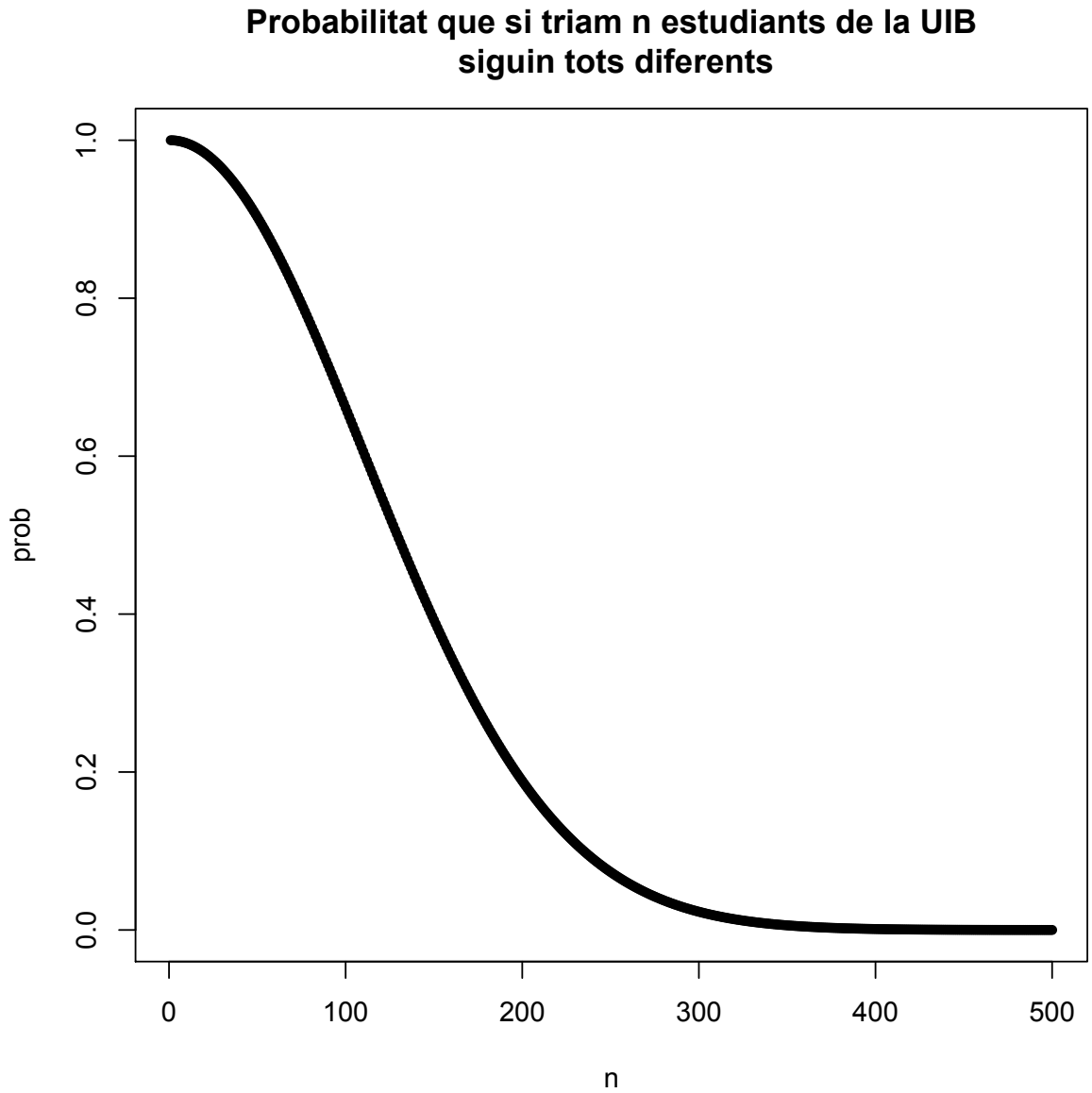
- Si N és molt més gran que n, els resultats per a m.a.s. valen (aproximadament) en aquest cas, perquè les repeticions són improbables i les variables aleatòries que formen la mostra són gairebé independents

  Farem l'abús de llenguatge de dir que en aquest cas també tenim una m.a.s.

- Si n és relativament gran, sovint es poden donar versions corregides dels estadístics

La vida real

Exemple: La UIB té uns 12000 estudiants



Mitjana mostral

Sigui $X_1, \ldots, X_n$ una m.a.s. de mida n d'una v.a. X d'esperança μ_X i desviació típica σ_X

La mitjana mostral és

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

Teorema

En aquestes condicions

$$E(\overline{X}) = \mu_X, \quad \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$$

$\sigma_{\overline{X}}$ és l'error estàndard de $\overline{X}$

Mitjana mostral

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

- És un estimador puntual de μ_X
- $E(\overline{X}) = \mu_X$
  - El valor esperat de $\overline{X}$ és μ_X
  - Si prenem moltes vegades una m.a.s. i en calculam la mitjana mostral, el valor mitjà d'aquestes mitjanes tendeix molt probablement a ser μ_X
- $\sigma_{\overline{X}} = \sigma_X/\sqrt{n}$: la variabilitat dels resultats de $\overline{X}$ tendeix a 0 quan prenem mostres grans

Mitjana mostral

```
> # tests.txt=notes dels tests de BL i BQ
> tests=scan("tests.txt")
Read 185 items
> mean(tests)
[1] 55.43243
> set.seed(100)
> mitjanes=replicate(10^4,
    mean(sample(tests,40,rep=TRUE)))
> mean(mitjanes)
[1] 55.45814
> #sd, per fer via
> c(sd(tests)/sqrt(40),sd(mitjanes))
[1] 3.390031 3.420459
```

Exemple

S'ha pres una m.a.s. de 10 estudiants de la UIB, i les seves alçades han estat

$$1.62, 1.75, 1.64, 1.69, 1.83, 1.85, 1.72, 1.61, 1.93, 1.62$$

Podem estimar l'alçada mitjana dels estudiants de la UIB:

$$\overline{x} = \frac{1.62 + 1.75 + 1.64 + \cdots + 1.62}{10} = 1.726$$

Com de "fina" és aquesta estimació? No us perdeu el proper tema!

Combinació lineal de normals és normal

Teorema

Si $Y_1, \ldots, Y_n$ *son v.a. normals independents, cada* $Y_i \sim N(\mu_i, \sigma_i)$, *i* $a_1, \ldots, a_n, b \in \mathbb{R}$ *aleshores*

$$Y = a_1 Y_1 + \cdots + a_n Y_n + b$$

és una v.a. $N(\mu, \sigma)$ *amb* μ *i* σ *les que toquen:*

- $E(Y) = a_1 \mu_1 + \cdots + a_n \mu_n + b$
- $\sigma(Y)^2 = a_1^2 \sigma_1^2 + \cdots + a_n^2 \sigma_n^2$

Cas X normal

Teorema

Sigui $X_1, \ldots, X_n$ una m.a.s. d'una v.a. X d'esperança μ_X i desviació típica σ_X. Si X és $N(\mu_X, \sigma_X)$, aleshores

$$\overline{X} \text{ és } N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$$

i per tant

$$Z = \frac{\overline{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \text{ és } N(0,1)$$

Z és l'expressió tipificada de la mitjana mostral

Teorema Central del Límit

Teorema

Sigui $X_1, \ldots, X_n$ una m.a.s. d'una v.a. X qualsevol d'esperança μ_X i desviació típica σ_X. Quan $n \to \infty$,

$$\overline{X} \to N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$$

i per tant

$$Z = \frac{\overline{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \to N(0,1)$$

(Aquestes convergències refereixen a les distribucions.)

Teorema Central del Límit

"Teorema"

Si n és gran ($n \geqslant 30$ o $\underline{\mathbf{40}}$), $\overline{X}$ és aproximadament normal, amb esperança μ_X i desviació típica $\dfrac{\sigma_X}{\sqrt{n}}$

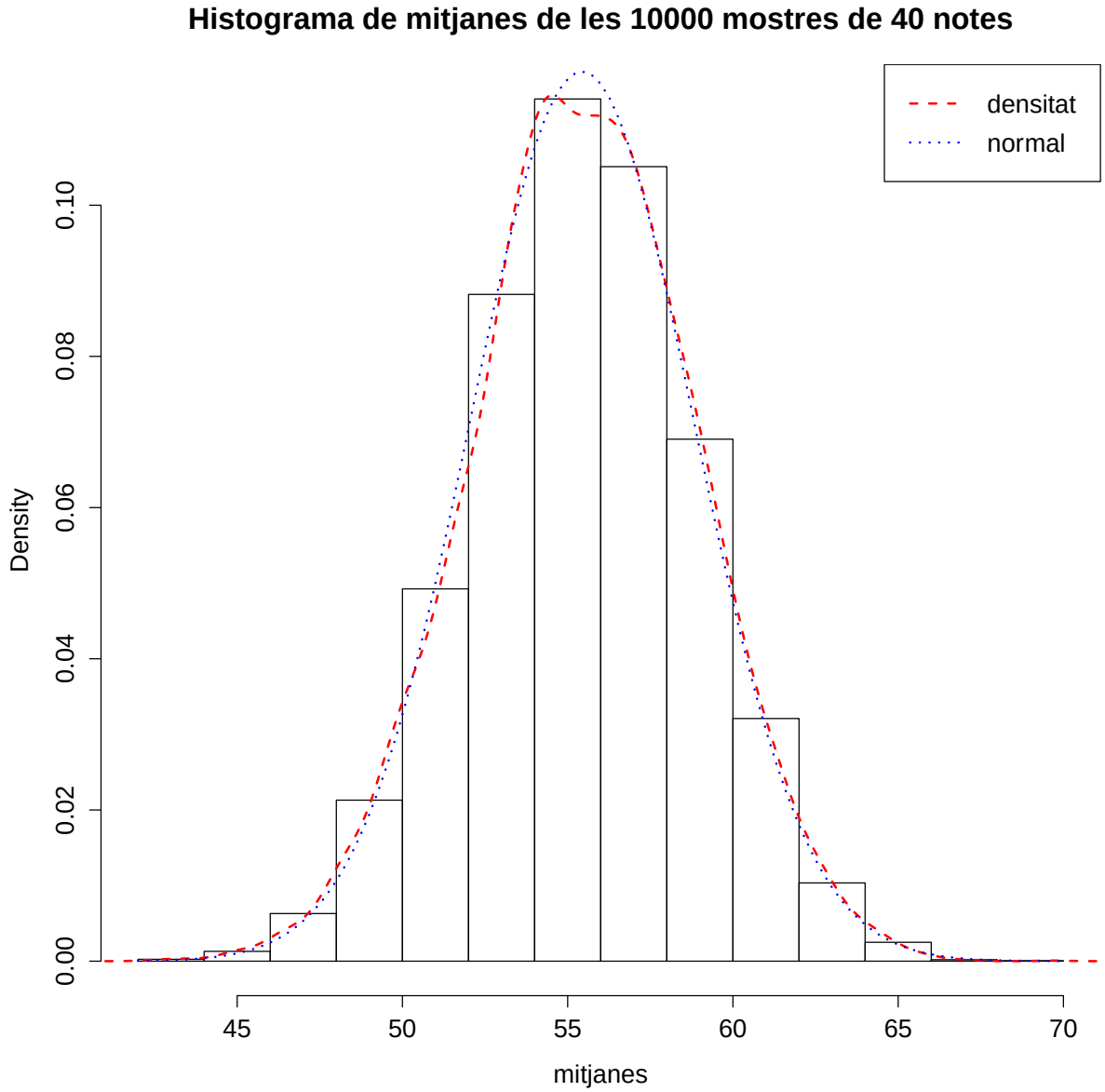
Exemple: Tenim una v.a. X de mitjana $\mu_X = 3$ i desv. típ. $\sigma_X = 0.2$. Prenem mostres aleatòries simples de mida 50. La distribució de la mitjana mostral $\overline{X}$ és aproximadament

$$N\left(3, \frac{0.2}{\sqrt{50}}\right) = N(3, 0.0283)$$

Teorema Central del Límit

```
> hist(mitjanes,freq=FALSE, main="Histograma
 de les mitjanes de 10000 mostres de 40 notes")
> lines(density(mitjanes),lty=2,lwd=2,col="red")
> curve(dnorm(x,mean(tests),sd(tests)/sqrt(40)),
 lty=3,lwd=2,col="blue",add=TRUE)
> legend("topright",legend=c("densitat","normal"),
 lwd=c(2,2),lty=c(2,3),col=c("red","blue"))
```

Teorema Central del Límit



Exemple

L'alçada d'una espècie de matolls té valor mitjà 115 cm, amb una desviació típica de 25. Prenem una m.a.s. de 100 matolls d'aquesta espècie.

Quina és la probabilitat que la mitjana mostral de les alçades sigui $\leqslant 110$ cm?

$$Z = \frac{\overline{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} = \frac{\overline{X} - 115}{2.5} \text{ és (aproximadament) } N(0,1)$$

$$\begin{aligned} P(\overline{X} \leqslant 110) &= P\left(Z \leqslant \frac{110 - 115}{2.5}\right) = P(Z \leqslant -2) \\ &= 0.0228 \end{aligned}$$

Exemple

L'alçada d'una espècie de matolls té valor mitjà 115 cm, amb una desviació típica de 25. Prenem una m.a.s. de 100 matolls d'aquesta espècie.

Quina és la probabilitat que la mitjana mostral de les alçades estigui entre 113 cm i 117 cm?

Exemple

L'alçada d'una espècie de matolls té valor mitjà 115 cm, amb una desviació típica de 25. Prenem una m.a.s. de 100 matolls d'aquesta espècie.

Quina és la probabilitat que la mitjana mostral de les alçades estigui entre 113 cm i 117 cm?

$$Z = \frac{\overline{X} - 115}{2.5} \text{ és } N(0,1)$$

$$P(113 \leqslant \overline{X} \leqslant 117) = P\left(\frac{113-115}{2.5} \leqslant Z \leqslant \frac{117-115}{2.5}\right)$$

$$= P(-0.8 \leqslant Z \leqslant 0.8) = F_Z(0.8) - F_Z(-0.8)$$

$$= 2F_Z(0.8) - 1 = 2 \cdot 0.7881 - 1 = 0.5763$$

Mitjana mostral sense reposició

Sigui $X_1, \ldots, X_n$ una m.a. sense reposició de mida n d'una v.a. X d'esperança μ_X i desviació típica σ_X.

Si n és petit en relació a la mida N de la població, tot funciona (per aproximació) com fins ara

Si n és gran en relació a N, aleshores

$$E(\overline{X}) = \mu_X, \quad \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

(factor de població finita)

El T.C.L. ja no val tal qual en aquest darrer cas

Proporció mostral

Sigui X una v.a. Bernoulli de paràmetre p_X (1 èxit, 0 fracàs). Sigui $X_1, \ldots, X_n$ una m.a.s. de mida n de X.

$S = \sum_{i=1}^{n} X_i$ és el nombre d'èxits observats. És $B(n, p)$.

La proporció mostral és

$$\widehat{p}_X = \frac{S}{n}$$

i és un estimador de p_X

Fixau-vos que $\widehat{p}_X$ és un cas particular de $\overline{X}$, per tant val tot el que hem dit fins ara per a mitjanes mostrals

Proporció mostral

$$\widehat{p}_X = \frac{S}{n}$$

- $E(\widehat{p}_X) = p_X$
- $\sigma_{\widehat{p}_X} = \sqrt{\frac{p_X(1-p_X)}{n}}$, l'error estàndard de la proporció mostral
- Si la mostra és sense reposició i n és relativament gran, $\sigma_{\widehat{p}_X} = \sqrt{\frac{p_X(1-p_X)}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$

Proporció mostral

Pel T.C.L.:

"Teorema"

Si n és gran ($n \geqslant 30$ o $\underline{\mathbf{40}}$) i la mostra és aleatòria simple,

$$\frac{\widehat{p}_X - p_X}{\sqrt{\frac{p_X(1-p_X)}{n}}} \approx N(0,1)$$

Exemple

En una mostra aleatòria de 60 estudiants de la UIB del curs 2013-14, 37 varen ser dones. Estimau la fracció de dones entre els estudiants de la UIB

Exemple

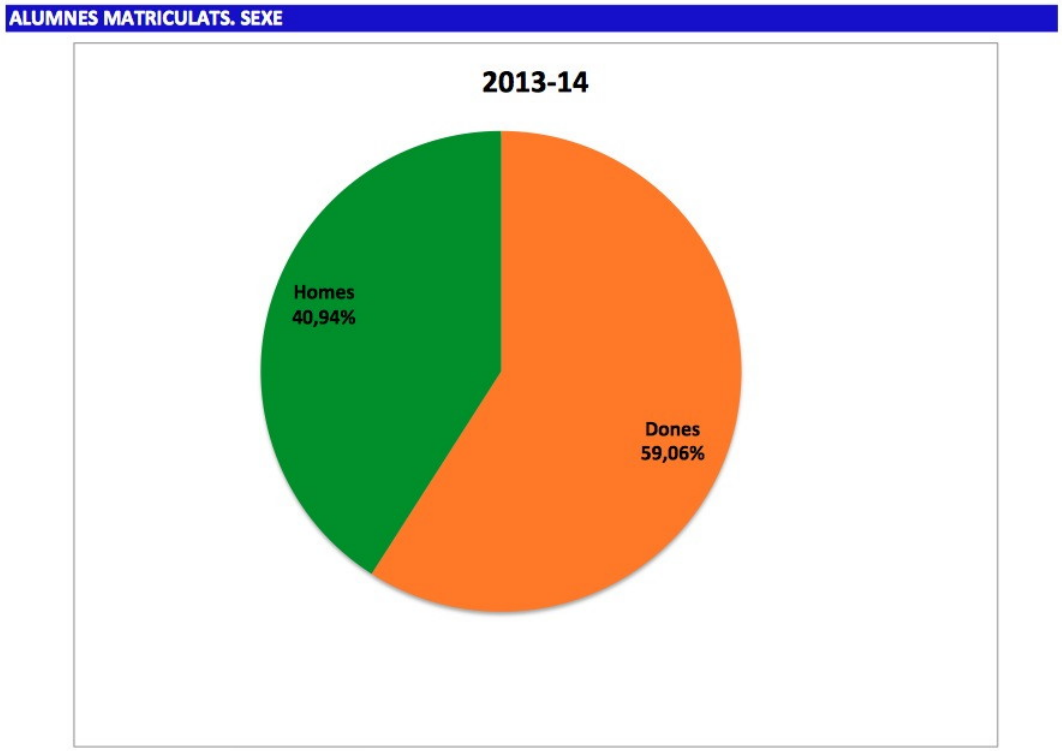
En una mostra aleatòria de 60 estudiants de la UIB del curs 2013-14, 37 varen ser dones. Estimau la fracció de dones entre els estudiants de la UIB

$$\frac{37}{60} = 0.6167$$

Exemple

En una mostra aleatòria de 60 estudiants de la UIB del curs 2013-14, 37 varen ser dones. Estimau la fracció de dones entre els estudiants de la UIB

$$\frac{37}{60} = 0.6167$$



Exemple

Un 59% dels estudiants de la UIB són dones. Si prenem una m.a.s. de 60 estudiants, quina és la probabilitat que la proporció mostral de dones sigui superior al 61%?

Exemple

Un 59% dels estudiants de la UIB són dones. Si prenem una m.a.s. de 60 estudiants, quina és la probabilitat que la proporció mostral de dones sigui superior al 61%?

$$p_X = 0.59$$

$$\widehat{p}_X \sim N\left(0.59, \sqrt{\frac{0.59(1-0.59)}{60}}\right) = N(0.59, 0.0635)$$

$$Z = \frac{\widehat{p}_X - 0.59}{0.0635} \sim N(0,1)$$

$$P(\widehat{p}_X > 0.61) = P\left(\frac{\widehat{p}_X - 0.59}{0.0635} > \frac{0.61 - 0.59}{0.0635}\right)$$

$$= P(Z > 0.315) = P(Z < -0.315) = 0.3764$$

Variància mostral

Sigui $X_1, \ldots, X_n$ una m.a.s. de mida n d'una v.a. X d'esperança μ_X i desviació típica σ_X

La variància mostral és

$$\widetilde{S}_X^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}$$

La desviació típica mostral és

$$\widetilde{S}_X = +\sqrt{\widetilde{S}_X^2}$$

A més, escriurem

$$S_X^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n} = \frac{(n-1)}{n}\widetilde{S}_X^2 \quad \text{i} \quad S_X = +\sqrt{S_X^2}$$

Variància mostral: Propietats

- $S_X^2 = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n} = \left(\dfrac{\sum_{i=1}^{n} X_i^2}{n} - \overline{X}^2\right)$
- $\widetilde{S}_X^2 = \dfrac{n}{n-1}\left(\dfrac{\sum_{i=1}^{n} X_i^2}{n} - \overline{X}^2\right)$

Variància mostral: Propietats

- $S_X^2 = \dfrac{\sum_{i=1}^n (X_i - \overline{X})^2}{n} = \left(\dfrac{\sum_{i=1}^n X_i^2}{n} - \overline{X}^2 \right)$
- $\widetilde{S}_X^2 = \dfrac{n}{n-1} \left(\dfrac{\sum_{i=1}^n X_i^2}{n} - \overline{X}^2 \right)$

Teorema

Si la v.a. X és normal, aleshores $E(\widetilde{S}_X^2) = \sigma_X^2$ i la v.a.

$$\frac{(n-1)\widetilde{S}_X^2}{\sigma_X^2}$$

té distribució χ_{n-1}^2

La distribució χ^2_n

La distribució χ^2_n (χ: en català, khi; en castellà, ji; en anglès, chi), on n són els graus de llibertat:

- És la de
$$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$
on $Z_1, Z_2, \ldots, Z_n$ son v.a. independents $N(0,1)$
- Té densitat
$$f_{\chi^2_n}(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{(n/2)-1} e^{-x/2} \qquad \text{si } x \geqslant 0$$
on $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt$ si $x > 0$

La distribució χ^2_n

- La distribució està tabulada (Teniu les taules a Campus Extens), i amb R és `chisq`
- Si $X_{\chi^2_n}$ és una v.a. amb distribució χ^2_n,

$$E(X_{\chi^2_n}) = n, \quad Var(X_{\chi^2_n}) = 2n$$

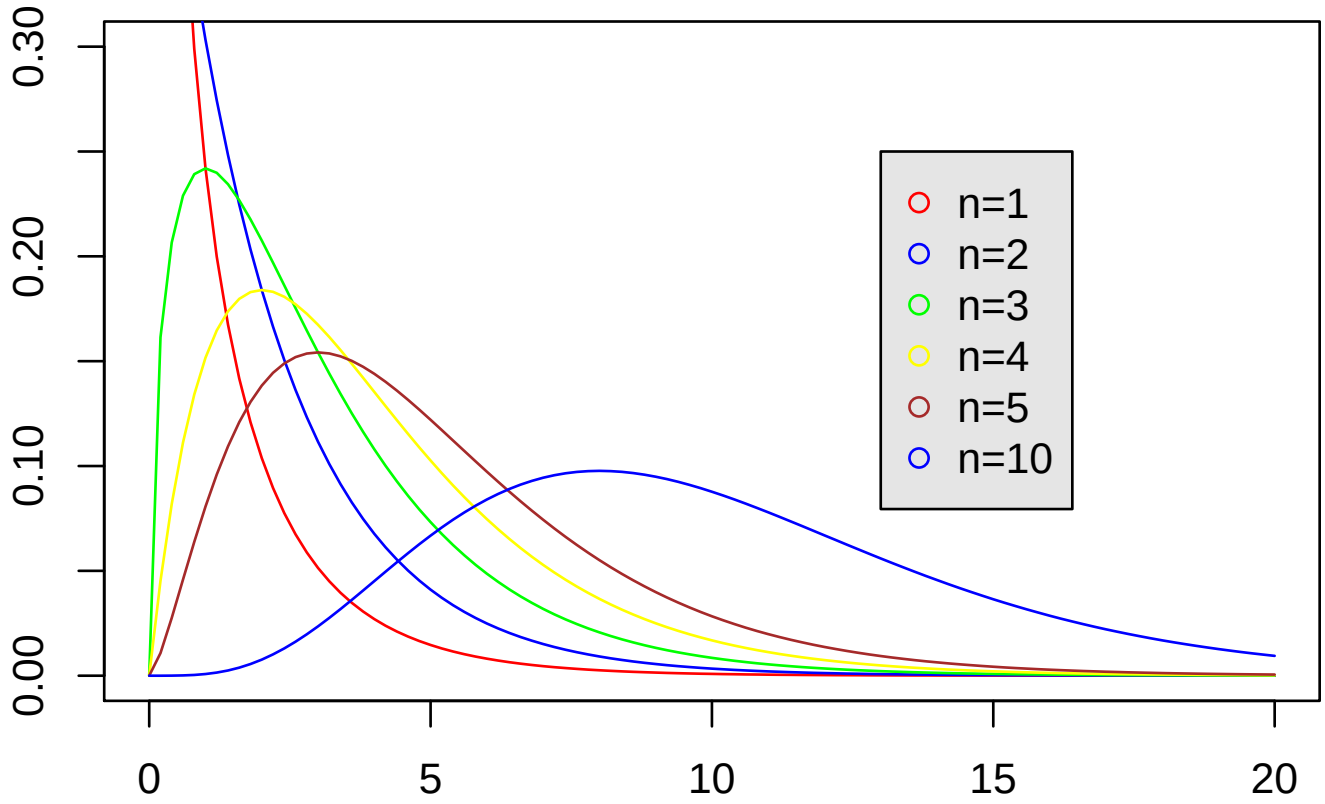
- χ^2_n s'aproxima a una distribució normal $N(n, \sqrt{2n})$ per a n gran ($n > 40$ o 50)

La distribució χ^2_n

$n\backslash\alpha$	0.995	0.99	0.975	0.95	0.9	0.85	0.8	0.75	0.7	0.65
1	7.879	6.635	5.024	3.841	2.706	2.072	1.642	1.323	1.074	0.873
2	10.597	9.21	7.378	5.991	4.605	3.794	3.219	2.773	2.408	2.1
3	12.838	11.345	9.348	7.815	6.251	5.317	4.642	4.108	3.665	3.283
4	14.86	13.277	11.143	9.488	7.779	6.745	5.989	5.385	4.878	4.438
5	16.75	15.086	12.833	11.07	9.236	8.115	7.289	6.626	6.064	5.573
6	18.548	16.812	14.449	12.592	10.645	9.446	8.558	7.841	7.231	6.695
7	20.278	18.475	16.013	14.067	12.017	10.748	9.803	9.037	8.383	7.806
8	21.955	20.09	17.535	15.507	13.362	12.027	11.03	10.219	9.524	8.909
9	23.589	21.666	19.023	16.919	14.684	13.288	12.242	11.389	10.656	10.006
10	25.188	23.209	20.483	18.307	15.987	14.534	13.442	12.549	11.781	11.097
11	26.757	24.725	21.92	19.675	17.275	15.767	14.631	13.701	12.899	12.184
12	28.3	26.217	23.337	21.026	18.549	16.989	15.812	14.845	14.011	13.266
13	29.819	27.688	24.736	22.362	19.812	18.202	16.985	15.984	15.119	14.345
14	31.319	29.141	26.119	23.685	21.064	19.406	18.151	17.117	16.222	15.421
15	32.801	30.578	27.488	24.996	22.307	20.603	19.311	18.245	17.322	16.494
16	34.267	32	28.845	26.296	23.542	21.793	20.465	19.369	18.418	17.565
17	35.718	33.409	30.191	27.587	24.769	22.977	21.615	20.489	19.511	18.633
18	37.156	34.805	31.526	28.869	25.989	24.155	22.76	21.605	20.601	19.699
19	38.582	36.191	32.852	30.144	27.204	25.329	23.9	22.718	21.689	20.764
20	39.997	37.566	34.17	31.41	28.412	26.498	25.038	23.828	22.775	21.826
21	41.401	38.932	35.479	32.671	29.615	27.662	26.171	24.935	23.858	22.888
22	42.796	40.289	36.781	33.924	30.813	28.822	27.301	26.039	24.939	23.947
23	44.181	41.638	38.076	35.172	32.007	29.979	28.429	27.141	26.018	25.006

$F_{\chi^2_{10}}(25.188) = 0.995,\ F_{\chi^2_{20}}(26.5) \approx 0.85$ Feu el test!

La distribució χ^2_n



Funció de densitat de χ^2_n per a alguns n

Exemple

L'augment diari del pes d'un pollastre d'una granja segueix una distribució normal amb desviació típica 1.7. Es pren una mostra de 12 pollastres. Suposam que aquesta mostra és petita respecte del total de la població de la granja.

Probabilitat que la desviació típica mostral sigui $\leqslant 2.5$?

Sigui $X=$ l'augment diari del pes d'un pollastre. Sabem que $\sigma_X^2 = (1.7)^2 = 2.89$. Com que X és normal i $n = 12$, tenim que

$$\frac{11 \cdot \widetilde{S}_X^2}{2.89} = \frac{(n-1)\widetilde{S}_X^2}{\sigma_X^2} \sim \chi_{11}^2$$

Exemple

L'augment diari del pes d'un pollastre d'una granja segueix una distribució normal amb desviació típica 1.7. Es pren una mostra de 12 pollastres. Suposam que aquesta mostra és petita respecte del total de la població de la granja.

Probabilitat que la desviació típica mostral sigui $\leqslant 2.5$?

$$\frac{11\widetilde{S}_X^2}{2.89} \sim \chi^2_{11}$$

$$P(\widetilde{S}_X < 2.5) = P\left(\widetilde{S}_X^2 < 2.5^2\right)$$

$$= P\left(\frac{11 \cdot \widetilde{S}_X^2}{2.89} < \frac{11 \cdot 2.5^2}{2.89}\right) = P(\chi^2_{11} < 23.79)$$

Exemple

L'augment diari del pes d'un pollastre d'una granja segueix una distribució normal amb desviació típica 1.7. Es pren una mostra de 12 pollastres. Suposam que aquesta mostra és petita respecte del total de la població de la granja.

Probabilitat que la desviació típica mostral sigui $\leqslant 2.5$?

$$\frac{11\widetilde{S}_X^2}{2.89} \sim \chi^2_{11}$$

$$P(\widetilde{S}_X < 2.5) = P\left(\widetilde{S}_X^2 < 2.5^2\right)$$

$$= P\left(\frac{11 \cdot \widetilde{S}_X^2}{2.89} < \frac{11 \cdot 2.5^2}{2.89}\right) = P(\chi^2_{11} < 23.79)$$

$$= \texttt{pchisq(23.7889,11)} = 0.986$$

Exemple

L'augment diari del pes d'un pollastre d'una granja segueix una distribució normal amb desviació típica 1.7. Es pren una mostra de 12 pollastres. Suposam que aquesta mostra és petita respecte del total de la població de la granja.

Probabilitat que la desviació típica mostral sigui $\leqslant 2.5$?

$$\frac{11\widetilde{S}_X^2}{2.89} \sim \chi^2_{11}$$

$$P(\widetilde{S}_X < 2.5) = P\left(\widetilde{S}_X^2 < 2.5^2\right)$$

$$= P\left(\frac{11 \cdot \widetilde{S}_X^2}{2.89} < \frac{11 \cdot 2.5^2}{2.89}\right) = P(\chi^2_{11} < 23.79)$$

$$\approx P(\chi^2_{11} < 24.725) = 0.99$$

Estimadors no esbiaixats

Quan un estimador és bo?

Un estimador puntual $\widehat{\theta}$ d'un paràmetre poblacional θ és no esbiaixat quan el seu valor esperat és precisament el valor del paràmetre:

$$E(\widehat{\theta}) = \theta$$

Es diu aleshores que l'estimació puntual és no esbiaixada.

El biaix de $\widehat{\theta}$ és $E(\widehat{\theta}) - \theta$

Estimadors no esbiaixats

Exemples

- $E(\overline{X}) = \mu_X$: $\overline{X}$ és estimador no esbiaixat de μ_X
- $E(\widehat{p}_X) = p_X$: $\widehat{p}_X$ és estimador no esbiaixat de p_X
- $E(\widetilde{S}_X^2) = \sigma_X^2$ si X és normal: $\widetilde{S}_X^2$ és estimador no esbiaixat de σ_X^2 quan X és normal
- $E(S_X^2) = \dfrac{n-1}{n}\sigma_X^2$ si X és normal; per tant S_X^2 és esbiaixat, amb biaix

$$E(S_X^2) - \sigma_X^2 = \frac{n-1}{n}\sigma_X^2 - \sigma_X^2 = -\frac{\sigma_X^2}{n} \underset{n\to\infty}{\longrightarrow} 0$$

Estimadors eficients

Quan un estimador és millor?

Quan és no esbiaixat i té poca variabilitat (així és més probable que aplicat a una m.a.s. doni prop del valor esperat)

Error estàndard d'un estimador $\widehat{\theta}$: la seva desviació típica

$$\sigma_{\widehat{\theta}} = \sqrt{Var(\widehat{\theta})}$$

Estimadors eficients

Donats dos estimadors $\widehat{\theta}_1$, $\widehat{\theta}_2$ no esbiaixats (o amb biaix $\underset{n\to\infty}{\longrightarrow} 0$) del mateix paràmetre θ, direm que

$$\widehat{\theta}_1 \text{ és més eficient que } \widehat{\theta}_2$$

quan

$$\sigma_{\widehat{\theta}_1} < \sigma_{\widehat{\theta}_2},$$

és a dir, quan

$$Var(\widehat{\theta}_1) < Var(\widehat{\theta}_2)$$

Estimadors eficients

Exemple: Sigui X una v.a. amb mitjana μ_X i desviació típica σ_X

Considerem la mediana $Me = Q_{0.5}$ de la realització d'una m.a.s. de X com a estimador puntual de μ_X

Si X és normal,

$$E(Me) = \mu_X,$$
$$Var(Me) \approx \frac{\pi}{2}\frac{\sigma_X^2}{n} \approx \frac{1.57\sigma_X^2}{n} = 1.57 Var(\overline{X})$$

Per tant, Me és un estimador no esbiaixat de μ_X, però menys eficient que $\overline{X}$

Estimadors eficients

- Si la població és normal, la mitjana mostral és l'estimador no esbiaixat més eficient de la mitjana poblacional
- Si la població és Bernoulli, la proporció mostral és l'estimador no esbiaixat més eficient de la proporció poblacional
- Si la població és normal, la variància mostral és l'estimador no esbiaixat més eficient de la variància poblacional

Estimadors eficients

Hem dit que si la població és normal, la variància mostral és l'estimador no esbiaixat més eficient de la variància poblacional

L'estimador "variància"

$$S_X^2 = \frac{(n-1)}{n}\widetilde{S}_X^2$$

encara és més eficient, però té biaix $\underset{n\to\infty}{\longrightarrow} 0$

Si n és petit ($\leqslant 30$ o 40), és millor fer servir la variància mostral $\widetilde{S}_X^2$ per estimar la variància, ja que el biaix influeix, però si n és gran, el biaix ja no és tan important i es pot fer servir S_X^2

Exemple: Estimació de poblacions

Tenim una població numerada $1, 2, \ldots, N$

En prenem una m.a.s. $x_1, \ldots, x_n$; sigui $m = \max(x_1, \ldots, x_n)$

Teorema

L'estimador no esbiaixat més eficient de N és

$$\widehat{N} = m + \frac{m - n}{n}$$

Un problema de rellevància històrica:

`http://en.wikipedia.org/wiki/German_tank_problem`

Exemple: Estimació de poblacions

Exemple: Assegut en un bar del Passeig Marítim he apuntat les llicències dels 40 primers taxis que he vist:

```
> taxis=c(1217,600,883,1026,150,715,297,137,508,134,
  38,961,538,1154,314,1121,823,158,940,99,977,286,
  1006,1207,264,1183,1120,498,606,566,1239,860,114,
  701,381,836,561,494,858,187)
```

Suposaré que formen una m.a.s. dels taxis de Palma. Aleshores, estim que el nombre de taxis de Palma és

```
> N=max(taxis)+(max(taxis)-length(taxis))/length(taxis)
> N
[1] 1268.975
```

En realitat, n'hi ha 1246

http://www.caib.es/eboibfront/es/2014/10195/551436/departamento-de-movilidad-seccion-de-transportes-r

Estimadors màxim versemblants

Com trobam bons estimadors?

Sigui X una v.a. discreta amb densitat

$$f_X(x;\lambda)$$

que depèn d'un paràmetre desconegut λ

Sigui $X_1, \ldots X_n$ una m.a.s. de X, i sigui $x_1, x_2, \ldots, x_n$ una realització d'aquesta mostra

La funció de versemblança de la mostra és la probabilitat condicionada següent:

$$\begin{aligned} L(\lambda|x_1, x_2, \ldots, x_n) &:= P(x_1, x_2, \ldots, x_n|\lambda) \\ &= P(X_1 = x_1) \cdots P(X_n = x_n) \\ &= f_X(x_1;\lambda) \cdots f_X(x_n;\lambda) \end{aligned}$$

Estimadors màxim versemblants

Donada la funció de versemblança $L(\lambda|x_1,\ldots,x_n)$ de la mostra, indicarem per

$$\hat{\lambda}(x_1,\ldots,x_n)$$

el valor del paràmetre λ on s'aconsegueix el màxim de $L(\lambda|x_1,\ldots,x_n)$. Serà una funció de $x_1,\ldots,x_n$.

Definició

Un estimador $\hat{\lambda}$ d'un paràmetre λ és màxim versemblant (MV, en anglès EM) quan, per a cada m.a.s, la probabilitat d'observar-la és màxima quan el paràmetre pren el valor de l'estimador aplicat a la mostra, és a dir, quan la funció de versemblança

$$L(\lambda|x_1,x_2,\ldots,x_n) = P(x_1,x_2,\ldots,x_n|\lambda)$$

assoleix el seu màxim.

Estimadors màxim versemblants

Exemple: Suposem que tenim una v.a. Bernoulli X de probabilitat d'èxit p desconeguda

Per a cada m.a.s. $x_1, \ldots, x_n$ de X, siguin $\widehat{p}_x$ la seva proporció mostral i $P(x_1, \ldots, x_n \mid p)$ la probabilitat d'obtenir-la quan el paràmetre pren el valor p

Teorema

El valor de p per al qual $P(x_1, \ldots, x_n \mid p)$ és màxim és $\widehat{p}_x$.

La proporció mostral és un estimador MV de p. Vegem-ho.

Estimadors màxim versemblants

Observació

En general, com que ln és creixent, en lloc de maximitzar $L(\lambda|x_1, \ldots, x_n)$, maximitzam

$$\ln(L(\lambda|x_1, \ldots, x_n))$$

que sol ser més fàcil (productes $\rightarrow$ sumes).

Estimadors màxim versemblants

Sigui $X_1, \dots X_n$ una m.a.s. d'una v.a. Bernoulli X de paràmetre p (desconegut). Posem $q = 1 - p$

$$f_X(1; p) = P(X = 1) = p, \quad f_X(0; p) = P(X = 0) = q$$

és a dir, per $x \in \{0, 1\}$, resulta que

$$f_X(x; p) = P(X = x) = p^x q^{1-x}.$$

La funció de versemblança és:

$$\begin{aligned} L(p|x_1, \dots, x_n) &= f_X(x_1; p) \cdots f_X(x_n; p) \\ &= p^{x_1} q^{1-x_1} \cdots p^{x_n} q^{1-x_n} \\ &= p^{\sum_{i=1}^n x_i} q^{\sum_{i=1}^n (1-x_i)} = p^{\sum_{i=1}^n x_i} q^{n-\sum_{i=1}^n x_i} \\ &= p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i} \end{aligned}$$

Exemple

La funció de versemblança és

$$\begin{aligned} L(p|x_1,\ldots,x_n) &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \\ &= p^{n\overline{x}}(1-p)^{n-n\overline{x}} \end{aligned}$$

on $\overline{x} = \dfrac{\sum_{i=1}^n x_i}{n}$

Volem trobar el valor de p on s'assoleix el màxim d'aquesta funció (on $\overline{x}$ és un paràmetre: la variable és p)

Maximitzarem el seu logaritme:

$$\begin{aligned} &\ln(L(p|x_1,\ldots,x_n)) \\ &\quad = n\overline{x}\ln(p) + n(1-\overline{x})\ln(1-p) \end{aligned}$$

Exemple

Derivem respecte de p:

$$\begin{aligned}
&\ln(L(p|x_1,\ldots,x_n))' \\
&\quad = n\overline{x}\frac{1}{p} - n(1-\overline{x})\frac{1}{1-p} \\
&\quad = \frac{1}{p(1-p)}\Big((1-p)n\overline{x} - pn(1-\overline{x})\Big) \\
&\quad = \frac{1}{p(1-p)}(n\overline{x} - pn) = \frac{n}{p(1-p)}(\overline{x}-p)
\end{aligned}$$

Estudiem el signe:

$$\begin{aligned}
\ln(L(p|x_1,\ldots,x_n))' \geqslant 0 \quad &\Leftrightarrow \overline{x} - p \geqslant 0 \\
&\Leftrightarrow p \leqslant \overline{x}
\end{aligned}$$

Exemple

Per tant

$$\ln(L(p|x_1,\ldots,x_n)) \begin{cases} \text{creixent fins } \overline{x} \\ \text{decreixent a partir de } \overline{x} \\ \text{té un màxim a } \overline{x} \end{cases}$$

I el resultat queda demostrat.
$L(\widehat{p}_X|x_1,\ldots,x_n) \geqslant L(p|x_1,\ldots,x_n)$ per a qualsevol p

Alguns estimadors MV

- $\widehat{p}_x$ és l'estimador MV del paràmetre p d'una v.a. Bernoulli
- $\overline{X}$ és l'estimador MV del paràmetre λ d'una v.a. Poisson
- $\overline{X}$ és l'estimador MV del paràmetre μ d'una v.a. normal
- S_X^2 (no $\widetilde{S}_X^2$) és l'estimador MV del paràmetre σ^2 d'una v.a. normal
- El màxim (no $\widehat{N}$) és l'estimador MV de la N al problema dels taxis

Exemple: λ per a una Poisson

Sigui X una característica d'una població que segueix una llei $Po(\lambda)$, amb $\lambda > 0$ desconegut. Prenem una mostra aleatòria simple $X_1, \ldots, X_n$ d'aquesta població i obtenim els resultats $x_1, \ldots, x_n$.

Trobau l'estimador màxim versemblant de λ per a $x_1, \ldots, x_n$.

Exemple: λ per a una Poisson

Primer hem de trobar la funció de versemblança $L(\lambda \mid x_1, \ldots, x_k)$.

Si $X \sim Pois(\lambda)$, sabem que $P(X = x_i) = e^{-\lambda} \cdot \frac{\lambda^{x_i}}{x_i!}$ i per tant

$$L(\lambda \mid x_1, \ldots, x_k) = P(X = x_1) \cdot P(X = x_2) \cdots P(X = x_n)$$

$$=$$

Exemple: λ per a una Poisson

Primer hem de trobar la funció de versemblança $L(\lambda \mid x_1, \ldots, x_k)$.

Si $X \sim Pois(\lambda)$, sabem que $P(X = x_i) = e^{-\lambda} \cdot \frac{\lambda^{x_i}}{x_i!}$ i per tant

$$\begin{aligned} L(\lambda \mid x_1, \ldots, x_k) &= P(X = x_1) \cdot P(X = x_2) \cdots P(X = x_n) \\ &= \left(e^{-\lambda} \cdot \frac{\lambda^{x_1}}{x_1!}\right) \cdot \left(e^{-\lambda} \cdot \frac{\lambda^{x_2}}{x_2!}\right) \cdots \left(e^{-\lambda} \cdot \frac{\lambda^{x_n}}{x_n!}\right) \\ &= e^{-n\lambda} \cdot \frac{\lambda^{x_1+\cdots+x_n}}{x_1! \cdots x_n!} = e^{-n\lambda} \cdot \frac{\lambda^{n\bar{x}}}{x_1! \cdots x_n!} \end{aligned}$$

Exemple: λ per a una Poisson

Ara volem trobar el valor de λ que maximitza

$$f(\lambda) := \ln(L(\lambda \mid x_1, \ldots, x_k)) = -n\lambda + n\overline{x}\ln(\lambda) - \ln(x_1! \cdots x_n!)$$

Exemple: λ per a una Poisson

Ara volem trobar el valor de λ que maximitza

$$f(\lambda) := \ln(L(\lambda \mid x_1, \ldots, x_k)) = -n\lambda + n\overline{x}\ln(\lambda) - \ln(x_1! \cdots x_n!)$$

Derivem respecte de λ

$$f'(\lambda) = -n + n\overline{x} \cdot \frac{1}{\lambda} = \frac{n(\overline{x} - \lambda)}{\lambda}$$

Com que $n, \lambda > 0$, tenim que

$$f'(\lambda) \begin{cases} > 0 & \text{si } \lambda < \overline{x} \\ < 0 & \text{si } \lambda > \overline{x} \\ = 0 & \text{si } \lambda = \overline{x} \end{cases}$$

Exemple: λ per a una Poisson

Per tant,

$$\ln(L(\lambda \mid x_1, \ldots, x_k)) \begin{cases} \text{és creixent a }]0, \overline{x}[\\ \text{és decreixent a }]\overline{x}, \infty[\\ \text{té el màxim a } \lambda = \overline{x} \end{cases}$$

Teorema

L'estimador màxim versemblant per a λ és la mitjana $\overline{x}$.

Exemple: Marca-recaptura

En una població hi ha N individus, en capturam K, els marcam i els tornam a amollar. Ara en tornam a capturar n, dels quals k estan marcats. A partir d'aquestes dades, volem estimar N.

Suposam que N i K no han canviat de la primera a la segona captura

$X =$"Un individu estigui marcat" és $Be(p)$ amb $p = \frac{K}{N}$

$X_1, \ldots, X_n$ la mostra capturada en segon lloc: $\widehat{p}_X = \frac{k}{n}$

Exemple: Marca-recaptura

$\widehat{p}_X$ és estimador màxim versemblant de p: Estimam que

$$\frac{K}{N} = \frac{k}{n} \Rightarrow N = \frac{n \cdot K}{k}$$

Per tant, l'estimador

$$\widehat{N} = \frac{n \cdot K}{k}$$

maximitza la probabilitat de l'observació "k marcats de n capturats". És l'estimador màxim versemblant de N.

Exemple: Marca-recaptura

Suposem que hem marcat 15 peixos del llac, i que en una captura de 10 peixos, n'hi ha 4 de marcats. Quants peixos estimau que conté el llac?

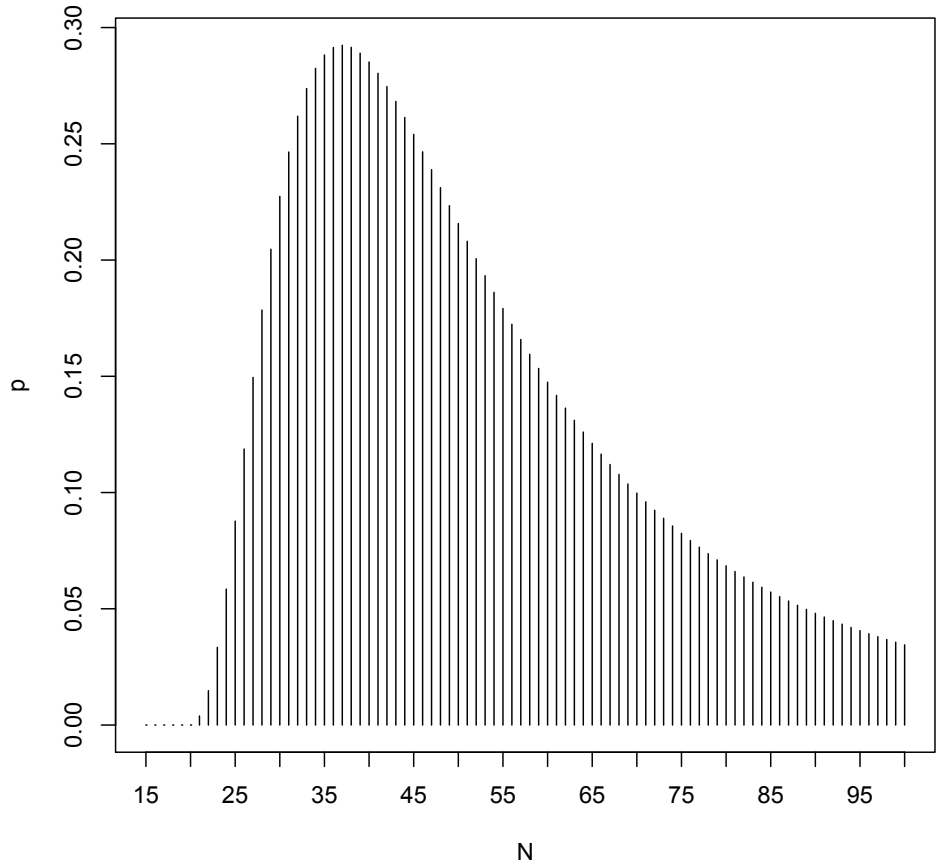
$$\widehat{N} = \frac{15 \cdot 10}{4} = 37.5$$

Per tant, estimam que hi haurà entre 37 i 38 peixos al llac

Exemple: Marca-recaptura

$$P(k \text{ marcats de } n \text{ capturats}) = \frac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}}$$

```
> N=15:100
> p=choose(15,4)*choose(N-15,6)/choose(N,10)
> plot(N,p,type="h",xaxp=c(15,100,17))
```



Exemple: Marca-recaptura

L'estimador

$$\widehat{N} = \frac{n \cdot K}{k}$$

és esbiaixat, amb biaix $\underset{n \to \infty}{\longrightarrow} 0$

L'estimador de Chapman

$$\widehat{N} = \frac{(n+1) \cdot (K+1)}{k+1} - 1$$

és menys esbiaixat per a mostres petites, i no esbiaixat si $K + n \geqslant N$ (però no màxim versemblant)