Intervals de confiança

El problema


ENTRE EL 12% Y EL 16% PADECE OBESIDAD

Sanidade estima que entre un 25% y un 30% de la población infantil gallega tiene sobrepeso

Amb un estimador, estimam el paràmetre amb una certa precisió, que depèn:

- De la variabilitat de l'estimador
- De la mida de la mostra
- Del nivell de confiança de l'estimació: com de segurs volem estar que l'estimació és correcta

El problema

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

Set de cada deu estudiants de la UIB (76,6 per cent) accepten haver copiat i aferrat fragments d'una web o un altre recurs obtingut a Internet i, sense esmentar-ne la procedència, haver-lo fet servir amb altres textos fets per ells per elaborar un

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.

Per tant (per ara):

Amb 95% de confiança podem afirmar que entre un mínim d'un 73.1% i un màxim d'un 80.1% dels estudiants de la UIB accepten...

El problema

EL PAÍS

PORTADA | INTERNACIONAL | PO

ECONOMÍA

ECONOMÍA | EMPRESAS | MERCADOS | BOLSA | FINANZAS PERSONALES | VIVIENDA | TECI

ESTÁ PASANDO

MERCADO LABORAL

El paro baja en 72.800 personas por el empleo temporal del verano

- La tasa de desempleo baja ligeramente en el tercer trimestre hasta el 25,98%
- El empleo avanza en 39.500 personas, aunque se desploman los indefinidos
- Solo se crean puestos de trabajo en el sector servicios
- Radiografía del mercado laboral español en 10 titulares

MANUEL V. GÓMEZ | Madrid | 24 OCT 2013 - 21:29 CET 476

Definicions bàsiques

A l'Encuesta de Población Activa (EPA):

Errores de muestreo relativos, de la població de 16 y más años por comunidad autónoma y relación con la actividad económica

Unidades:Porcentaje

	Ocupados	Parados
	2013TIII	2013TIII
Total Nacional	0,37	0,87

http://www.ine.es/jaxi/tabla.do?per=03&type=db&divi=EPA&idtab=313

estimación ± 1 vez el error de muestreo = intervalo de confianza del 67%.

estimación ± 2 veces el error de muestreo = intervalo de confianza del 95%.

estimación ± 3 veces el error de muestreo = intervalo de confianza del 99,7%.

http://www.ine.es/docutrab/eval_epa/evaluacion_epa04.pdf

El problema

EPA d'octubre de 2013:

- El nombre estimat d'aturats a nivell nacional va ser de 5 904 700
- L'error de mostreig va ser d'un 0.87%
- Per tant, estam bastant segurs (nivell de confiança del 95%) que el nombre d'aturats estava entre

$$\begin{aligned} 5\,904\,700 - 2 \cdot 0.0087 \cdot 5\,904\,700 &= 5\,904\,700 - 102\,742 \\ &= 5\,801\,958 \quad \text{i} \\ 5\,904\,700 + 2 \cdot 0.0087 \cdot 5\,904\,700 &= 5\,904\,700 + 102\,742 \\ &= 6\,007\,442 \end{aligned}$$

- L'EPA de juny 2013 havia estimat el nombre d'aturats en 5 977 500
- No hi ha evidència que l'atur baixàs

Definicions bàsiques

Una estimació per intervals d'un paràmetre poblacional és una regla per calcular, a partir d'una mostra, un interval on, amb una certa probabilitat (nivell de confiança), es troba el valor vertader del paràmetre

Exemples

Exemple: Hem triat a l'atzar 50 estudiants de grau de la UIB, hem calculat les seves notes mitjanes de les assignatures del primer semestre, i la mitjana d'aquestes mitjanes ha estat un 6.3, amb una variància mostral de 1.8

Determinau un interval que puguem afirmar amb probabilitat 95% que conté la mitjana real de les notes mitjanes dels estudiants de grau de la UIB aquest primer semestre

Exemples

Exemple: En un experiment s'ha mesurat el percentatge d'augment d'alcohol en sang a 40 persones després de prendre 4 canyes de cervesa. La mitjana i la desviació típica mostral d'aquests percentatges d'increment han estat

$$\overline{x} = 41.2, \quad \widetilde{s} = 2.1$$

Determinau un interval que puguem afirmar amb probabilitat 95% que conté el percentatge d'augment mitjà d'alcohol en sang (vertader) d'una persona després de beure quatre canyes de cervesa.

Definicions bàsiques

Donat un paràmetre θ, l'interval $]A, B[$ és un interval de confiança del $(1-\alpha)\cdot 100\%$ per al paràmetre θ quan

$$P(A < \theta < B) = 1 - \alpha$$

El valor $(1-\alpha)\cdot 100\%$ (o també només el $1-\alpha$) rep el nom de nivell de confiança

El valor α rep el nom de nivell de significació

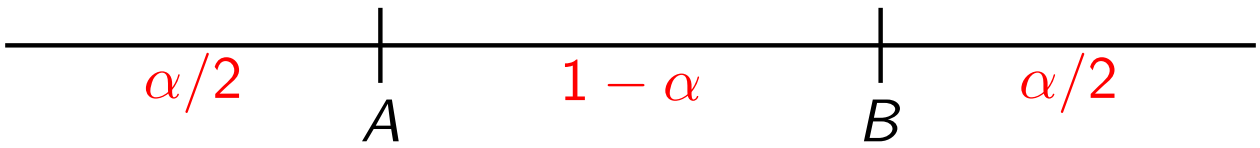
Exemple: $]A, B[$ és un interval de confiança del 95% (o de nivell de significació de 0.05) si

$$P(A < \theta < B) = 0.95$$

Definicions bàsiques

Per defecte, cercarem intervals tals que la cua de probabilitat sobrant α es reparteixi per igual a cada costat de l'interval:

$$P(\theta < A) = P(\theta > B) = \frac{\alpha}{2}$$



Exemple: Per cercar un interval de confiança $]A, B[$ del 95%, cercarem A, B de manera que

$$P(\theta < A) = 0.025 \quad \text{i} \quad P(\theta > B) = 0.025$$

Exemple: μ de població normal amb σ coneguda

Sigui X una v.a. normal amb mitjana poblacional μ desconeguda i desviació típica poblacional σ coneguda (a la pràctica, usualment, estimada en un experiment anterior)

Sigui $X_1, \ldots, X_n$ una m.a.s. de X, amb mitjana mostral $\overline{X}$

Volem determinar un interval de confiança per a μ amb un cert nivell de confiança (posem, 97.5%): un interval $]A, B[$ tal que

$$P(A < \mu < B) = 0.975$$

Exemple: μ de població normal amb σ coneguda

Sota aquestes condicions, sabem que

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

segueix una distribució normal estàndard

Comencem calculant un interval centrat en 0 on Z hi tingui probabilitat 0.975:

$$0.975 = P(-\delta < Z < \delta) = F_Z(\delta) - F_Z(-\delta) = 2F_Z(\delta) - 1$$

$$F_Z(\delta) = \frac{1.975}{2} = 0.9875 \Rightarrow \delta = \texttt{qnorm(0.9875)} = 2.24$$

Exemple: μ de població normal amb σ coneguda

Per tant

$$P(-2.24 < Z < 2.24) = 0.975$$

Substituint $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$

$$P\left(-2.24 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 2.24\right) = 0.975$$

$$P\left(\overline{X} - 2.24\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 2.24\frac{\sigma}{\sqrt{n}}\right) = 0.975$$

Exemple: μ de població normal amb σ coneguda

Per tant, la probabilitat que la μ de la X es trobi dins l'interval

$$\left]\overline{X} - 2.24\frac{\sigma}{\sqrt{n}}, \overline{X} + 2.24\frac{\sigma}{\sqrt{n}}\right[$$

és 0.975: és un interval de confiança del 97.5%

A més:

- Centrat en $\overline{X}$
- El 0.025 de probabilitat restant està repartit per igual als dos costats de l'interval

Exemple: μ de població normal amb σ coneguda

- Com a estimador: Un 97.5% de les ocasions que prenguem una mostra de mida n de X, el vertader valor de μ es trobarà dins d'aquest interval
- Per a una mostra concreta: La probabilitat que, si una μ ha produït aquesta mostra, aleshores estigui dins aquest interval concret, és del 97.5%
- Ho entendrem com a: "La probabilitat que μ estigui dins aquest interval és del 97.5%"
- Però és mentira (abús de llenguatge): La μ concreta és un valor fix, per tant que pertanyi a aquest interval concret té probabilitat 1 (si hi pertany) i 0 (si no hi pertany)

I.C. per a μ de població normal amb σ coneguda

Teorema

Sigui $X \sim N(\mu, \sigma)$ *amb* μ *desconeguda i* σ *coneguda.*

Prenem una m.a.s. de X *de mida* n*, amb mitjana* $\overline{X}$*.*

Un interval de confiança del $(1-\alpha)\cdot 100\%$ *per a* μ *és*

$$\left]\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right[$$

on $z_{1-\frac{\alpha}{2}}$ *és el* $(1-\frac{\alpha}{2})$*-quantil de la normal estàndard* Z *(és a dir,* $z_{1-\frac{\alpha}{2}} = F_Z^{-1}(1-\frac{\alpha}{2})$*, o* $P(Z \leqslant z_{1-\frac{\alpha}{2}}) = 1-\frac{\alpha}{2}$*)*

I.C. per a μ de població normal amb σ coneguda

Si X és normal amb σ coneguda, un interval de confiança per a μ del $(1-\alpha)\cdot 100\%$ és

$$\overline{X} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} := \left]\overline{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right[$$

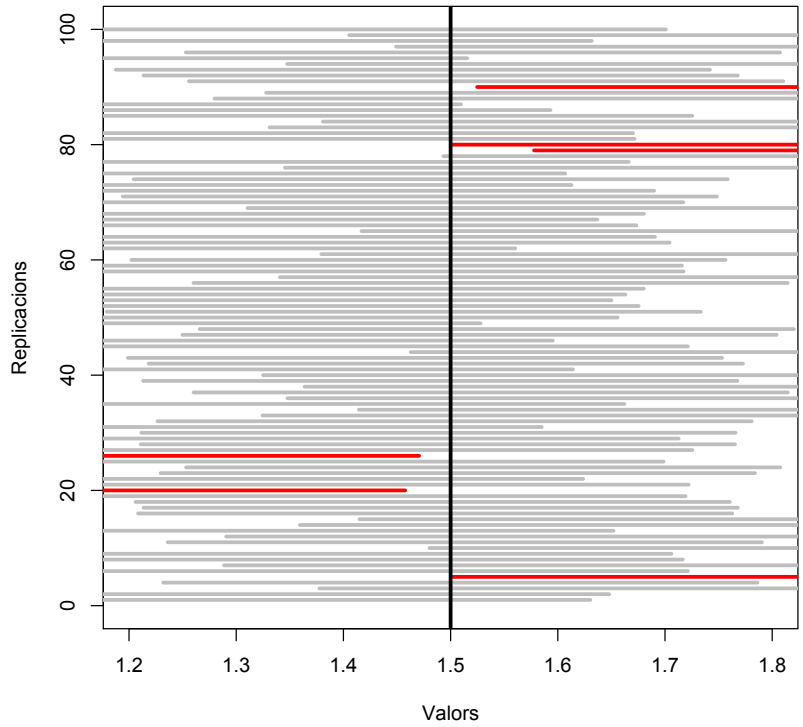
Observau que està centrat en $\overline{X}$

Confiança $1-\alpha$	Significació α	$z_{1-\frac{\alpha}{2}}$
0.900	0.100	1.64
0.950	0.050	1.96
0.975	0.025	2.24
0.990	0.010	2.58

I.C. per a μ de població normal amb σ coneguda

```
ICZ=function(x,sigma,alpha){
  c(mean(x)-qnorm(1-alpha/2)*sigma/sqrt(length(x)),
  mean(x)+qnorm(1-alpha/2)*sigma/sqrt(length(x)))
}
set.seed(5)
mu=1.5; sigma=1; alpha=0.05
Poblacio=rnorm(10^6,mu,sigma)
M=replicate(100,ICZ(sample(Poblacio,50,replace=T),
 sigma,alpha))
plot(1:10,type="n",xlim=c(1.2,1.8),ylim=c(0,100),
xlab="Valors",ylab="Replicacions")
seg.int=function(i){color="grey";
  if((mu<M[1,i]) | (mu>M[2,i])){color = "red"}
  segments(M[1,i],i,M[2,i],i,col=color,lwd=3)}
invisible(sapply(1:100,FUN=seg.int))
abline(v=mu,lwd=3)
```

I.C. per a μ de població normal amb σ coneguda



Alerta!

De mitjana, un $\alpha 100\%$ de les vegades, un interval de confiança del $(1-\alpha)100\%$ no contindrà el valor real del paràmetre

Exemple

$$\left]\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right[$$

Prenem una m.a.s. de mida $n = 16$ d'una v.a. normal amb $\sigma = 4$ i μ desconeguda. La mitjana de la m.a.s. és $\overline{x} = 20$.

Calculau un interval de confiança del 97.5% per a μ

Exemple

Tenim un aparell per mesurar volums de líquid. Per saber si està ben calibrat, prenem 10 mostres consistents a emplenar un recipient d'un litre exacte i mesurar el seu contingut amb el nostre aparell. Obtenim els resultats de la taula següent:

Volum mesurat (en litres)	Freq. Absoluta
1.000	1
1.002	2
1.004	1
1.006	2
1.008	1
1.010	2
1.012	1

$$\overline{x} = 1.006$$

Exemple

Suposem que les mesures amb el nostre aparell del contingut d'aquest recipient segueixen una distribució normal amb variància poblacional coneguda $\sigma^2 = 0.01$. Calculau un interval de confiança del 90% per al resultat mitjà de mesurar un litre exacte amb el nostre aparell.

Amplada

L'amplada A de l'interval de confiança

$$\left]\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right[$$

és

$$\begin{aligned} A &= \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} - \left(\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \\ &= 2z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \end{aligned}$$

L'error màxim, al nivell de confiança $(1-\alpha)$, que cometem en estimar μ per mitjà de $\overline{X}$ és la meitat d'aquesta amplada,

$$z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

Amplada

L'amplada A de l'interval de confiança

$$\left]\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right[$$

és

$$A = 2z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

Observacions

- Per a n i α fixos, si σ creix, A creix
- Per a σ i α fixos, si n creix, A decreix
- Per a σ i n fixos, si $1-\alpha$ creix, A creix

Amplada

Si volem calcular la mida n de la mostra per assegurar-nos que l'interval de confiança per μ al nivell $(1-\alpha)$ té amplada prefixada màxima A_0 (o un error màxim $A_0/2$), podem aïllar la n:

$$A_0 \geqslant 2z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \Rightarrow n \geqslant \left(2z_{1-\frac{\alpha}{2}} \frac{\sigma}{A_0}\right)^2$$

Donada A_0, prendrem

$$n = \left\lceil \left(2z_{1-\frac{\alpha}{2}} \frac{\sigma}{A_0}\right)^2 \right\rceil$$

Exemple

Recordau que les mesures amb el nostre aparell del contingut del recipient d'1 litre exacte segueixen una distribució normal amb variància poblacional coneguda $\sigma^2 = 0.01$

Quantes mesures hauríem d'efectuar per obtenir la mesura mitjana amb un error màxim de 0.05 al nivell de confiança del 90%?

Distribució t de Student

Sigui $X \sim N(\mu, \sigma)$

Sigui $X_1, \ldots, X_n$ una m.a.s. de X, amb mitjana $\overline{X}$ i desviació típica mostral $\widetilde{S}_X$

Teorema

En aquestes condicions, la v.a.

$$t = \frac{\overline{X} - \mu}{\widetilde{S}_X / \sqrt{n}}$$

segueix una distribució t de Student amb $n-1$ graus de llibertat, t_{n-1}

$\widetilde{S}_X / \sqrt{n}$: l'error mostral, estima l'error estàndard $\sigma / \sqrt{n}$

Distribució t de Student

La distribució t de Student amb ν graus de llibertat, t_ν:

- Té densitat

$$f_{t_\nu}(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

  on $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt$ si $x > 0$.

- La distribució està tabulada (Teniu les taules a Campus Extens), i amb R és `t`

Distribució t de Student

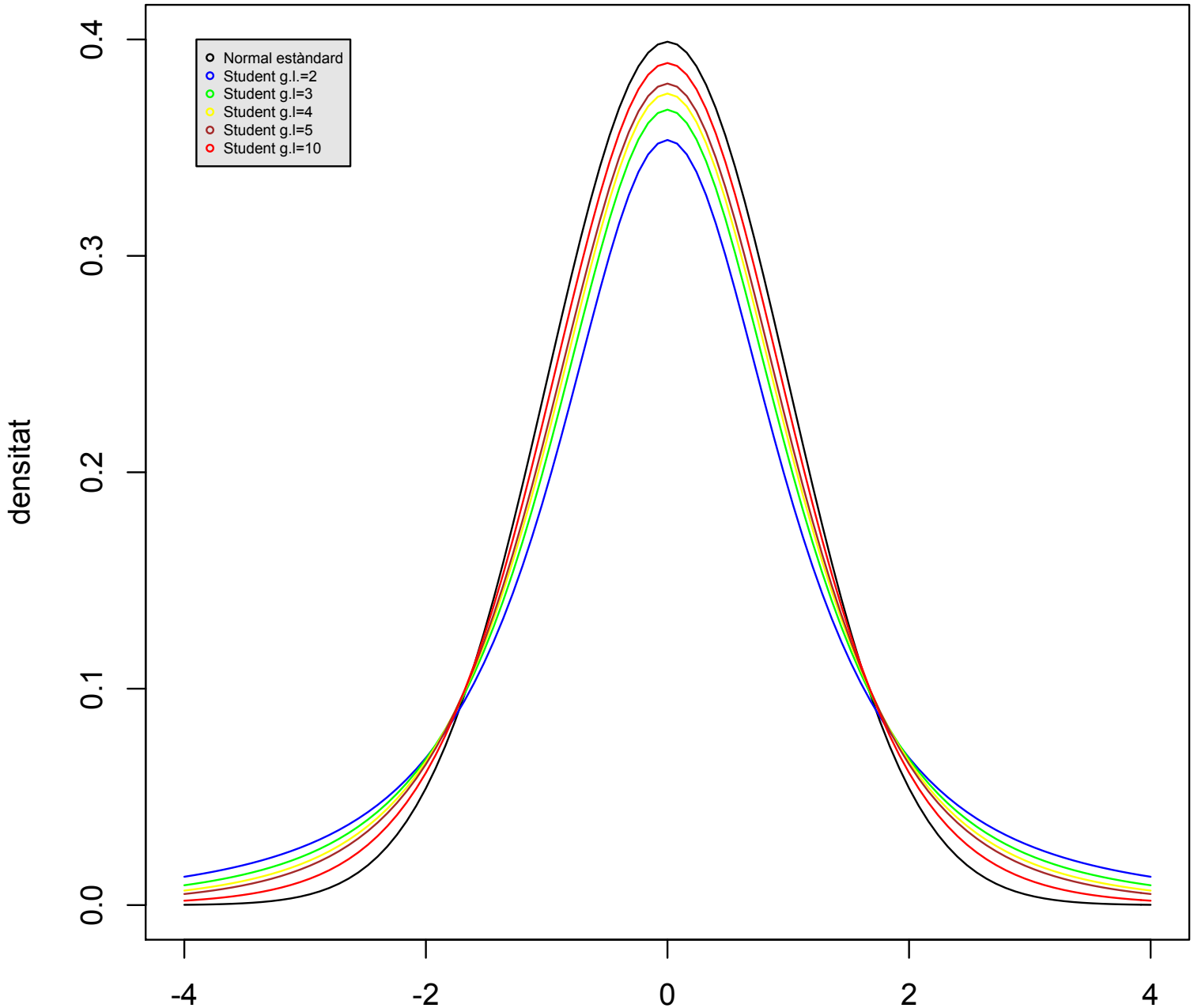
Sigui t_ν una v.a. que segueix la distribució t de Student amb ν graus de llibertat

- $E(t_\nu) = 0$ si $\nu > 1$ i $Var(t_\nu) = \dfrac{\nu}{\nu - 2}$ si $\nu > 2$
- La seva funció de distribució és simètrica respecte de $E(t_\nu) = 0$ (com la d'una $N(0, 1)$):

$$P(t_\nu \leqslant -x) = P(t_\nu \geqslant x) = 1 - P(t_\nu \leqslant x)$$

- Si ν és gran, la seva distribució és aproximadament la de $N(0, 1)$ (però amb més variància: un poc més aplatada)

Distribució t de Student

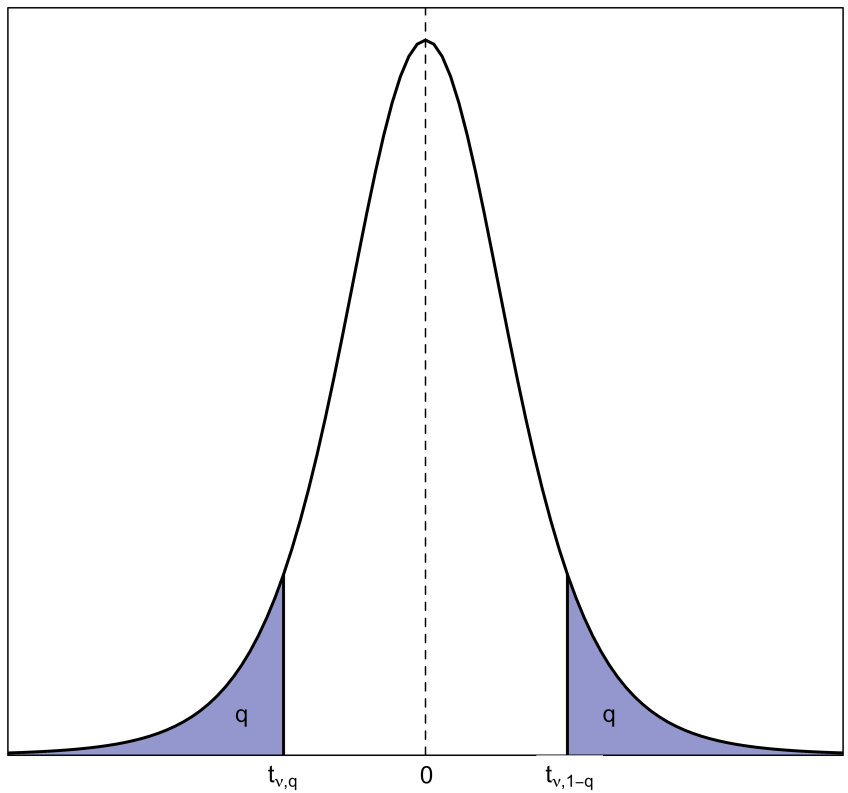


Distribució t de Student

Indicarem amb $t_{\nu,q}$ el q-quantil d'una v.a. X_{t_ν} que segueix una distribució t_ν:

$$P(X_{t_\nu} \leqslant t_{\nu,q}) = q$$

Per simetria, $t_{\nu,q} = -t_{\nu,1-q}$



μ de població normal amb σ desconeguda

Considerem la situació següent:

- X una v.a. normal amb μ i σ desconegudes
- $X_1, \ldots, X_n$ una m.a.s. de X de mida n, amb mitjana $\overline{X}$ i variància mostral $\widetilde{S}_X^2$

Teorema

En aquestes condicions, un interval de confiança del $(1-\alpha)\cdot 100\%$ *per a* μ *és*

$$\left]\overline{X} - t_{n-1,1-\frac{\alpha}{2}}\frac{\widetilde{S}_X}{\sqrt{n}}, \overline{X} + t_{n-1,1-\frac{\alpha}{2}}\frac{\widetilde{S}_X}{\sqrt{n}}\right[$$

Exemple

L'empresa *RX-print* ofereix una impressora de radiografies d'altíssima qualitat. En la seva publicitat afirma que els seus cartutxos imprimeixen una mitjana de 500 radiografies amb l'especificació:

`Dades tècniques: Mostra mensual de mida` $n = 25$, `població suposada normal, nivell de confiança del 90%`

Uns radiòlegs desitgen comprovar aquestes afirmacions i prenen una mostra a l'atzar de mida $n = 25$, obtenint una mitjana de $\overline{x} = 518$ radiografies i una desviació típica mostral $\widetilde{s} = 40$

Amb aquesta mostra, la mitjana poblacional anunciada pel fabricant cau dins de l'interval de confiança del 90%?

Exemple

Cal calcular l'interval de confiança per a μ amb

$$n = 25, \overline{x} = 518, \widetilde{s} = 40, \alpha = 0.1$$

Serà

$$\left]\overline{x} - t_{24,0.95}\frac{\widetilde{s}}{\sqrt{n}}, \overline{x} + t_{24,0.95}\frac{\widetilde{s}}{\sqrt{n}}\right[$$

Mirant en les taules de la t de Student, obtenim $t_{24,0.95} = 1.71$

```
> qt(0.95,24)
[1] 1.710882
```

Operant: $]504.32, 531.68[$, i no conté el 500 (però s'equivoca a favor del consumidor!)

Observacions

- L'interval de confiança obtingut està centrat en $\overline{X}$
- La fórmula

$$\left]\overline{X} - t_{n-1,1-\frac{\alpha}{2}}\frac{\widetilde{S}_X}{\sqrt{n}}, \overline{X} + t_{n-1,1-\frac{\alpha}{2}}\frac{\widetilde{S}_X}{\sqrt{n}}\right[$$

  per a l'interval de confiança del $(1-\alpha)\cdot 100\%$ es pot fer servir quan X és normal i n qualsevol

- Si n és gran $t_{n-1,1-\frac{\alpha}{2}} \approx z_{1-\frac{\alpha}{2}}$ i podem aproximar-lo amb

$$\left]\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\widetilde{S}_X}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\widetilde{S}_X}{\sqrt{n}}\right[$$

μ per a mostres grans

Considerem ara la situació següent:

- X una v.a. qualsevol amb mitjana poblacional μ desconeguda i desv. típ. σ coneguda
- $X_1, \ldots, X_n$ una m.a.s. de X, amb mitjana $\overline{X}$
- n és gran (posem, $n \geqslant 40$)

Teorema

En aquestes condicions, podem prendre com a interval de confiança del $(1-\alpha)\cdot 100\%$ *per a* μ

$$\left]\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right[$$

μ per a mostres grans

Considerem ara la situació següent:

- X una v.a. qualsevol amb mitjana poblacional μ desconeguda i desv. típ. σ desconeguda
- $X_1, \ldots, X_n$ una m.a.s. de X, amb mitjana $\overline{X}$ i desviació típica mostral $\widetilde{S}_X$
- n és gran (posem, $n \geqslant 40$)

"Teorema"

En aquestes condicions, es recomana prendre com a interval de confiança del $(1-\alpha)\cdot 100\%$ per a μ

$$\left]\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\widetilde{S}_X}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\widetilde{S}_X}{\sqrt{n}}\right[$$

Exemple

En un experiment s'ha mesurat el percentatge d'augment d'alcohol en sang a 40 persones després de prendre 4 canyes de cervesa. La mitjana i la desviació típica mostral d'aquests percentatges d'increment han estat

$$\overline{x} = 41.2, \quad \widetilde{s} = 2.1$$

Determinau un interval que puguem afirmar amb probabilitat 95% que conté el percentatge d'augment mitjà d'alcohol en sang d'una persona després de beure quatre canyes de cervesa.

Ens demanen un interval de confiança del 95% per a μ de la v.a. X "percentatge d'augment d'alcohol en sang d'una persona després de beure quatre canyes de cervesa"

Exemple

No sabem com és X, però $n = 40$ és gran

Podem emprar

$$\left]\overline{x} - z_{1-\frac{\alpha}{2}}\frac{\widetilde{s}}{\sqrt{n}}, \overline{x} + z_{1-\frac{\alpha}{2}}\frac{\widetilde{s}}{\sqrt{n}}\right[$$

on

$$n = 40, \overline{x} = 41.2, \widetilde{s} = 2.1,$$

$$\alpha = 0.05 \Rightarrow z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$$

$$]40.55, 41.85[$$

Podem afirmar amb un 95% de confiança que l'augment mitjà d'alcohol en sang d'una persona després de beure quatre canyes de cervesa està entre el 40.55% i el 41.85%

Exemple

S'ha pres una mostra de sang a 1000 adults sans i s'hi ha mesurat la quantitat de calci (en mg per dl de sang). S'ha obtingut una mitjana mostral de 9.5 mg/dl amb una desviació típica mostral de 0.5 mg/dl.

Trobau un interval de confiança del 95% per a la quantitat mitjana de calci en sang en un adult sa

Amplada

L'amplada de

$$\left]\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\widetilde{S}_X}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\widetilde{S}_X}{\sqrt{n}}\right[$$

és

$$A = 2z_{1-\frac{\alpha}{2}}\frac{\widetilde{S}_X}{\sqrt{n}}$$

Per determinar n (gran) que doni com a màxim una amplada A prefixada, ens cal $\widetilde{S}_X$, que depèn de la mostra.

Solucions:

- Si sabem la desv. típ. poblacional σ, l'empram per aproximar $\widetilde{S}_X$
- Si hem pres una mostra prèvia (pilot), n'empram la desviació típica mostral per estimar σ

Amplada

D'una població X n'hem pres una m.a.s. pilot que ha tingut una desviació típica mostral $\widetilde{s}_{pilot}$.

Estimarem que la mida mínima n d'una m.a.s. de X que doni un interval de confiança per a μ_X de nivell de confiança $1-\alpha$ i amplada màxima A_0 és

$$n = \left\lceil \left(2 z_{1-\frac{\alpha}{2}} \frac{\widetilde{s}_{pilot}}{A_0} \right)^2 \right\rceil$$

Exemple

Volem estimar l'alçada mitjana dels estudiants de la UIB. Cercam un interval de confiança del 99% amb una precisió màxima de 1 cm. En una mostra pilot de 25 estudiants, obtinguérem

$$\overline{x} = 170 \text{ cm}, \widetilde{s} = 10 \text{ cm}$$

Basant-nos en aquestes dades, quina mida hauria de tenir la mostra per assolir el nostre objectiu?

p per a mostres petites

Considerem la situació següent:

- X una v.a. Bernoulli amb p desconeguda
- $X_1, \ldots, X_n$ una m.a.s. de X, amb nombre d'èxits x i per tant freqüència relativa d'èxits $\widehat{p}_X = x/n$

Recordau que x és $B(n, p)$

Mètode "exacte" de Clopper-Pearson

Un interval de confiança $]p_0, p_1[$ del $(1-\alpha)100\%$ per a p s'obté trobant el p_0 més gran i el p_1 més petit tals que

$$\sum_{k=x}^{n} p_0^k (1-p_0)^{n-k} \leqslant \frac{\alpha}{2}, \qquad \sum_{k=0}^{x} p_1^k (1-p_1)^{n-k} \leqslant \frac{\alpha}{2}$$

A mà (consultant taules) és una feinada.

p per a mostres petites

El paquet `epitools` porta

`binom.exact(èxits,mida,conf.)`

per calcular-ho.

De 10 pacients tractats amb un medicament, 2 s'han curat. Donau un interval de confiança del 95% per a la proporció p de pacients que aquest medicament cura.

```
> install.packages("epitools",dep=TRUE)
> library(epitools)
> round(binom.exact(2,10,0.95),3)
  x  n proportion lower upper conf.level
1 2 10        0.2 0.025 0.556       0.95
```

Dóna $]0.025, 0.556[$

p per a mostres grans I

Considerem ara la situació següent:

- X una v.a. Bernoulli amb p desconeguda
- $X_1, \ldots, X_n$ una m.a.s. de X, amb n gran (per exemple, $n \geqslant 40$) i freqüència relativa d'èxits $\widehat{p}_X$

En aquestes condicions (pel T.C.L.),

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

p per a mostres grans I

Per tant

$$P\left(-z_{1-\frac{\alpha}{2}} \leqslant \frac{\widehat{p}_X - p}{\sqrt{\frac{p(1-p)}{n}}} \leqslant z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

i aïllant la p obtenim:

p per a mostres grans I

Mètode de Wilson

En aquestes condicions, un interval de confiança del $(1-\alpha)\cdot 100\%$ per a p és (posant $\widehat{q}_X = 1-\widehat{p}_X$)

$$\left]\frac{\widehat{p}_X + \frac{z^2_{1-\alpha/2}}{2n} - z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_X\widehat{q}_X}{n} + \frac{z^2_{1-\alpha/2}}{4n^2}}}{1+\frac{z^2_{1-\alpha/2}}{n}}, \frac{\widehat{p}_X + \frac{z^2_{1-\alpha/2}}{2n} + z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_X\widehat{q}_X}{n} + \frac{z^2_{1-\alpha/2}}{4n^2}}}{1+\frac{z^2_{1-\alpha/2}}{n}}\right[$$

`binom.wilson` del paquet `epitools`

p per a mostres grans II

Considerem ara la situació següent:

- X una v.a. Bernoulli amb p desconeguda
- $X_1, \ldots, X_n$ una m.a.s. de X, amb n més gran i $\widehat{p}_X$ enfora de 0 i 1. Per exemple, tal que:

$$n \geqslant 100, n\widehat{p}_X \geqslant 10, n(1-\widehat{p}_X) \geqslant 10$$

Fórmula de Laplace (1812)

En aquestes condicions, es pot prendre com a interval de confiança del $(1-\alpha)\cdot 100\%$ per a p

$$\left]\widehat{p}_X - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}_X(1-\widehat{p}_X)}{n}}, \widehat{p}_X + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}_X(1-\widehat{p}_X)}{n}}\right[$$

Exemple

En una mostra aleatòria de 500 famílies amb nins en edat escolar es va trobar que 340 introduïen fruita de forma diària en la dieta dels seus fills

Cercau un interval de confiança del 95% per a la proporció real de famílies d'aquesta ciutat amb nins en edat escolar que incorporen fruita fresca de forma diària en la dieta dels seus fills

Exemple

$X =$ "Aportar diàriament fruita a la dieta dels fills"
és $Be(p)$, i cercam interval de confiança del 95% per a p

Com que $n = 500 \geqslant 100$, $n\widehat{p}_X = 340 \geqslant 10$ i
$n(1 - \widehat{p}_X) = 160 \geqslant 10$, podem emprar

$$\left]\widehat{p}_X - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}_X(1-\widehat{p}_X)}{n}}, \widehat{p}_X + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}_X(1-\widehat{p}_X)}{n}}\right[$$

amb

$$n = 500, \widehat{p}_X = \frac{340}{500} = 0.68,$$
$$\alpha = 0.05 \Rightarrow z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$$

Dóna

$$]0.639, 0.721[$$

Exemple

Amb els altres mètodes:

```
> round(binom.exact(340,500,0.95),3)
    x   n proportion lower upper conf.level
1 340 500       0.68 0.637 0.721       0.95
> round(binom.wilson(340,500,0.95),3)
    x   n proportion lower upper conf.level
1 340 500       0.68 0.638 0.719       0.95
```

Donen:

- Clopper-Pearson: $]0.637, 0.721[$
- Wilson: $]0.638, 0.719[$
- Laplace: $]0.639, 0.721[$

Exemple

En un assaig d'un nou tractament de quimioteràpia, en una mostra de n (gran) malalts tractats, cap desenvolupà càncer testicular com a efecte secundari. Trobau un interval de confiança al 95% per a la proporció de malalts tractats amb aquesta quimio que desenvolupen càncer testicular.

Els metges empren $\left]0, \frac{3}{n}\right[$ (la regla del 3)

Observacions

- El mètode de Wilson dóna un I.C. centrat en

$$\frac{\widehat{p}_X + \frac{z^2_{1-\alpha/2}}{2n}}{1 + \frac{z^2_{1-\alpha/2}}{n}} = \frac{2n\widehat{p}_X + z^2_{1-\frac{\alpha}{2}}}{2n + 2z^2_{1-\frac{\alpha}{2}}}$$

- No es coneix una fórmula per al centre de l'I.C. de Clopper-Pearson.
- La fórmula de Laplace dóna un I.C. centrat en $\widehat{p}_X$
- Quan n creix es redueix l'amplada de l'interval de confiança

Amplada

L'amplada de l'interval de confiança de Laplace és

$$A = 2z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}_X(1-\widehat{p}_X)}{n}}$$

No podem determinar la mida de la mostra a fi que l'interval de confiança tingui una certa amplada màxima sense conèixer $\widehat{p}_X$, que no coneixem sense una mostra

El màxim de $\sqrt{\widehat{p}_X(1-\widehat{p}_X)}$ s'assoleix a $\widehat{p}_X = 0.5$

Per tant, calcularem n per obtenir una A màxima fixada suposant el pitjor dels casos ($\widehat{p}_X = 0.5$):

$$A \geqslant 2z_{1-\frac{\alpha}{2}}\sqrt{\frac{0.5^2}{n}} = \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \Rightarrow n = \left\lceil \frac{z^2_{1-\frac{\alpha}{2}}}{A^2} \right\rceil$$

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.

Error =

Exemple

Volem estudiar quina fracció de les morts per càncer corresponen a morts per càncer d'estómac. Per determinar aquesta fracció a un nivell de confiança del 95% i garantir un error màxim de 0.05, de quina mida ha de ser la mostra en el pitjor dels casos?

Variància d'una població normal

Considerem ara la situació següent:

- X una v.a. normal amb μ i σ desconegudes
- $X_1, \ldots, X_n$ una m.a.s. de X i variància mostral $\widetilde{S}_X^2$

Teorema

En aquestes condicions

$$\frac{(n-1)\tilde{S}_X^2}{\sigma^2}$$

té distribució χ^2_{n-1}

Variància d'una població normal

Considerem ara la situació següent:

- X una v.a. normal amb μ i σ desconegudes
- $X_1, \ldots, X_n$ una m.a.s. de X i variància mostral $\widetilde{S}_X^2$

Teorema

En aquestes condicions, un interval de confiança del $(1-\alpha)\cdot 100\%$ *per a* σ^2 *és*

$$\left]\frac{(n-1)\widetilde{S}_X^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)\widetilde{S}_X^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right[,$$

on $\chi^2_{\nu,q}$ *és el q-quantil de la distribució* χ^2_ν

Variància d'una població normal

En efecte

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

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

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

I ara χ^2_{n-1} no és simètrica, així que s'han de calcular $\chi^2_{n-1,\frac{\alpha}{2}}$ i $\chi^2_{n-1,1-\frac{\alpha}{2}}$

Observació: L'interval de confiança per σ^2 no està centrat en $\widetilde{S}^2_X$

Exemple

Un índex de qualitat d'un reactiu químic és el temps que triga a actuar. L'estàndard és que aquest ha de ser $\leqslant 30$ segons. Se suposa que la distribució del temps d'actuació del reactiu és aproximadament normal.

Es realitzen 30 proves en les quals es mesura el temps d'actuació del reactiu:

12, 13, 13, 14, 14, 14, 15, 15, 16, 17, 17, 18, 18, 19, 19, 25, 25, 26, 27, 30, 33, 34, 35, 40, 40, 51, 51, 58, 59, 83

Es demana calcular un interval de confiança per a la desviació típica al nivell 95%

Exemple

$$\left]\frac{(n-1)\widetilde{S}_X^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)\widetilde{S}_X^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right[$$

```
> Temps=c(12,13,13,14,14,14,15,15,16,17,17,18,
 18,19,19,25,25,26,27,30,33,34,35,40,40,51,51,
 58,59,83)
> length(Temps) #n
[1]  30.0000
> var(Temps) # variància mostral
[1]  301.5506
```

i $\alpha = 0.05$:

$$\chi^2_{29,0.975} = 45.72,\ \chi^2_{29,0.025} = 16.05$$

Exemple

L'interval serà

$$\left]\frac{(n-1)\widetilde{S}_X^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)\widetilde{S}_X^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right[$$

Obtenim

$$\left]\frac{29 \cdot 301.5506}{45.72}, \frac{29 \cdot 301.5506}{16.05}\right[=]191.27, 544.86[$$

Aquest era per a la variància! Per a la desviació típica

$$]\sqrt{191.27}, \sqrt{544.86}[=]13.83, 23.34[$$

"Poblacions finites"

Fins ara hem emprat mostres aleatòries simples

A la pràctica, sovint es prenen mostres aleatòries sense reposició

Si la mida N de la població és molt més gran que la mida n de la mostra (posem $N \geqslant 40n$), les fórmules donades fins ara funcionen aproximadament bé

Però...

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.

"Poblacions finites"

Es dóna l'efecte de població finita quan N és relativament petit

En aquest cas, a les fórmules que hem donat per als intervals de confiança per a μ o p cal multiplicar l'error estàndard o l'error mostral pel factor corrector

$$\sqrt{\frac{N-n}{N-1}}$$

"Poblacions finites"

Considerem la situació següent:

- X una població de mida N que segueix una distribució amb mitjana poblacional μ desconeguda
- $X_1, \ldots, X_n$ una m.a. sense reposició de X, amb mitjana $\overline{X}$
- n és gran

"Teorema"

En aquestes condicions, es recomana prendre com a interval de confiança del $(1-\alpha)\cdot 100\%$ per a μ

$$\left]\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}},\ \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right[$$

"Poblacions finites"

Considerem la situació següent:

- X una població de mida N que segueix una distribució Bernoulli amb p desconeguda
- $X_1, \ldots, X_n$ una m.a. sense reposició de X, amb n molt gran i amb freqüència relativa d'èxits $\widehat{p}_X$ no extrema

"Teorema"

En aquestes condicions, es recomana prendre com a interval de confiança del $(1-\alpha)\cdot 100\%$ per a p

$$\left]\widehat{p}_X - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}_X(1-\widehat{p}_X)}{n}}\sqrt{\frac{N-n}{N-1}}\,,\ \widehat{p}_X + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}_X(1-\widehat{p}_X)}{n}}\sqrt{\frac{N-n}{N-1}}\right[$$

"Poblacions finites"

"Teorema"

En les condicions anteriors, per obtenir un interval de confiança del $(1-\alpha)\cdot 100\%$ per a p en el pitjor dels casos caldrà prendre una mostra de mida

$$n = \left\lceil \frac{N z^2_{1-\frac{\alpha}{2}}}{A^2(N-1) + z^2_{1-\frac{\alpha}{2}}} \right\rceil$$

Exemple

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 $p = q = 0.05$.

De la població total d'estudiants de grau de la UIB quants d'hem d'escollir de manera aleatòria sense reposició per estimar la proporció dels que han comès plagi, amb un error del 3.52% i un nivell de confiança del 95%?