Contrast d'hipòtesis

Selecció d'activitats resoltes

Jose Fco. Martínez Boscá, Arnau Mir Torres, Lluís M. Pla Aragonés,
Àngel J. Gil Estallo (autors) i Ángel A. Juan (editor)

Introducció

La utilització dels contrastos d'hipòtesis estadístics és summament freqüent en la presa de decisions en diferents camps de les ciències, aplicades a múltiples problemes: els econòmics (per a valorar, per exemple, les vendes futures d'una empresa), els biològics (per a provar l'efectivitat d'una droga nova), els industrials (controls de producció), els mèdics i fins i tot els vinculats a les ciències de l'educació. Els contrastos estadístics d'hipòtesis són un tema que pertany a la rama de l'estadística inferencial. La comprensió d'aquests contrastos és rellevant per a utilitzar-los en camps disciplinaris diversos. Segons Gardner, es pot dir que l'estadística actua com a "pont entre les ciències naturals i socials" (1997, pàg. 171).

Mitjançant la construcció de mostres es pretén explicar quin és el comportament de la població a partir d'un conjunt limitat de casos per a prendre finalment decisions sobre tot l'Univers. És així com el contrast estadístic d'hipòtesis adquireix rellevància, i ens proporciona una eina que permet establir conclusions sobre fenòmens poblacionals a partir de les dades disponibles, que habitualment són de tipus mostral. Però perquè aquests procediments puguin ser portats a terme de manera efectiva, és imprescindible conèixer la lògica del procés i els errors que es poden cometre en la conseqüent presa de decisions, a fi de fer una lectura correcta dels resultats als quals s'han arribat.

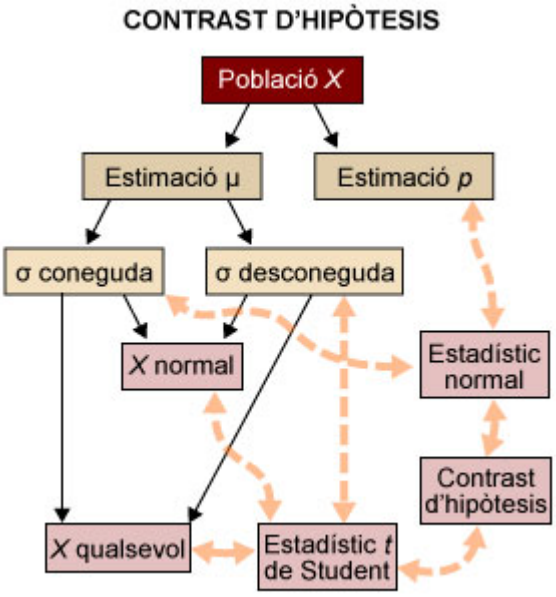
En aquest mòdul es pretén entendre què és i per a què s'utilitza un contrast d'hipòtesis, i també saber calcular i interpretar els estadístics de contrast i els *p*-valors a l'hora de fer aquests contrastos per a la mitjana poblacional, sigui o no coneguda la desviació estàndard poblacional, i per a les proporcions.

També farem contrastos sobre la mitjana de mostres aparellades. Que dues mostres siguin dependents (aparellades) o no està determinat per les fonts (persones o objectes) que ens aporten les observacions. Si en l'obtenció d'ambdues mostres s'han utilitzat les mateixes fonts o fonts associades, tindrem dues mostres dependents. Pel contrari, si s'han utilitzat fonts completament diferents parlarem de mostres independents.

Suposem que, en iniciar el semestre, seleccionem a l'atzar trenta alumnes matriculats en Estadística i els passem un test de coneixements previs. Al final del semestre, seleccionem trenta alumnes més a l'atzar i els passem un test de coneixements adquirits durant el curs. En aquest cas, consideraríem ambdues mostres com a independents. Al contrari, si el test de coneixements adquirits es fes als mateixos trenta alumnes que van fer el test inicial, llavors parlaríem de mostres dependents. Un altre exemple seria: es qüestiona si el temps d'arrencada d'un ordinador amb l'antivirus TRONX és més lent que si l'ordinador no té aquest antivirus. Amb l'objectiu de fer

aquest estudi s'ha mesurat el temps d'arrencada de deu ordinadors sense tenir cap antivirus instal·lat. Després, s'ha instal·lat l'antivirus en tots aquests ordinadors i s'ha tornat a mesurar el temps d'arrencada. Suposant que el temps d'arrencada d'un ordinador es distribueix normalment, es decideix amb $\alpha = 0,05$ si l'existència en l'ordinador de l'antivirus provoca que els temps d'arrencada siguin més grans. Trobeu el *p*-valor del contrast.

Mapa conceptual



Activitats resoltes - Contrast d'Hipòtesi

Activitats Resoltes: Contrast d'Hipòtesi

Fitxers necessaris per realitzar les activitats:

- ActR07TA.csv
- ActR07ANTUV.csv
- ActR07OPSIS.csv

A les primeres files dels fitxers consten les descripcions de les variables i al moment d'importar-los a R cal tenir en compte que aquestes línies no contenen dades (s'aconsella usar *skip*).

En altres activitats en les quals el nombre d'observacions és petit se suggereix copiar-pegar les dades o la instrucció que s'usa per introduir-los en R.

Activitat 1: Estudi sobre la velocitat d'un proveïdor d'Internet.

Contrast d'hipòtesis paramètriques. Contrast de la mitjana d'una població normal amb variància desconeguda. Estimació per intervals: Interval de confiança.

En un test per mesurar la velocitat de baixada d'un proveïdor d'internet les dades obtingudes de 10 usuaris, en kbps, van anar els següents:

```
VAR1<-c(150.8, 234, 260, 235.4, 280, 276, 200, 300, 256, 190)
```

Aquest proveïdor d'Internet publicita que la seva velocitat de baixada és de 256 kbps. Suposant que la velocitat segueix una distribució normal, contrasteu l'afirmació de l'empresa, al nivell de significació del 5%. Calculeu un interval de confiança per μ al 95% de nivell de confiança.

Per contestar a aquestes pregunta plantejar la hipòtesi nul · la i alternativa, donar la fórmula de l'estadístic de contrast així com la seva distribució de probabilitat. Trobeu el *p-valor* del contrast i l'interval de confiança corresponent. Comproveu que les conclusions són les mateixes amb el *p-valor* i amb l'interval.

Solució

$$H_0 : \mu = 256$$
$$H_1 : \mu \neq 256$$

L'estadístic del contrast seria :

$$t^* = \frac{\bar{x} - \mu}{s/n} = \frac{238.2 - 256}{46.2/\sqrt{10}} = \frac{-17.8}{14.61} = -1.22$$

Hem de calcular ara el *p-valor*, és a dir la probabilitat d'obtenir un valor igual o inferior a –1.22 o superior a 1.22 en una distribució *t-student* amb 9 graus de llibertat (10-1). Usant les taules estadístiques de la distribució *t-student* amb 9 graus de llibertat podem aproximar aquesta probabilitat per: $p - valor = 2 * 0,125 = 0,250$.

L'interval de confiança al 95% és: $\bar{x} \pm t_{9,0.025} \frac{s}{\sqrt{n}}$.

A continuació, usem *t.test*:

```
t.test(VAR1, alternative='two.sided', mu=256, conf.level=.95)
```

```
##
##  One Sample t-test
##
## data:  VAR1
## t = -1.2162, df = 9, p-value = 0.2548
## alternative hypothesis: true mean is not equal to 256
## 95 percent confidence interval:
##  205.1497 271.2903
## sample estimates:
## mean of x
##    238.22
```

Com que el *p-valor* (0.25) és més gran que el nivell de significació 0.05 acceptem que $\mu = 256$. L'interval de confiança al 95% per μ és $(205.15, 271.29)$ i com 256 està dins de l'interval acceptem també la hipòtesi nul · la. Això està d'acord amb la conclusió que hem obtingut amb el *p-valor*. ## Activitat 2: Estudi sobre les compres realitzades per Internet en un portal d'i-business.

Contrast d'hipòtesis paramètriques. Contrast per a la proporció. Estimació per intervals: Interval de confiança.

Un portal d'e-business sap que el 60% de tots els seus visitants a la Web estan interessats a adquirir els seus productes, però són poc inclinats al comerç electrònic i no realitzen finalment la compra via Internet.

a) Es demana contrastar al nivell de significació del 2% si en l'últim any s'ha reduït el percentatge de gent que no està disposada a comprar per Internet. Per a això es va agafar una mostra de 500 visitants, per conèixer la seva opinió, i es va observar que el 55% no estava disposat a realitzar compres via on-line.

b) En l'empresa anterior s'està realitzant un estudi per determinar la quantitat mensual d'hores de baixa laboral, causades per accidents en el treball. Suposant un nivell de significació del 0.05, s'ha realitzat el següent contrast d'hipòtesi, per comparar si la mitjana d'hores és igual o inferior a 18, $H_0 : \mu = 18$ vs. $H_1 : \mu < 18$. Obtenim el següent output:

```
One sample t-test
data:  variable$V1
t = 0.33,  p-value = 0.63
alternative hypothesis: true mean is less than 18
sample estimates:
mean of x
18.129
```

Quines conclusions trauries?

Solució

a) El que es demana és contrastar el següent:

$$H_0 : p = 0,6$$
$$H_A : p < 0,6$$

Podem observar que es compleixen els supòsits de normalitat, ja que la distribució del nombre de visitants del portal que no estan disposats a comprar via Internet es va aproximar a una normal a causa que $np_{H_0} \geq 5$ i $n(1 - p_{H_0}) \geq 5$. Usant R obtenim:

```
prop.test(275,  500,  alternative='less',  p=.6, conf.level=.98, correct=FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  275 out of 500, null probability 0.6
## X-squared = 5.2083, df = 1, p-value = 0.01124
## alternative hypothesis: true p is less than 0.6
## 98 percent confidence interval:
##  0.0000000 0.5950852
## sample estimates:
##    p
## 0.55
```

Del resultat anterior es desprèn que el p-valor és $0.01124 < 0.02$, llavors rebutjarem la hipòtesi nul · la a un nivell de significació del 2%.
En conclusió, existeix evidència estadística que la proporció de visitants al portal que estan disposats comprar per Internet ha augmentat, és a dir, que el percentatge de visitants que són poc inclinats a comprar via on-line ha disminuït.

b) Com que el p-valor $0.63 > 0.05$, no podem rebutjar la hipòtesi nul · la, és a dir, assumirem com a possible l'opció que el nombre mitjà de dies de baixa laboral sigui 18, ja que no tenim indicis suficients per rebutjar aquesta possibilitat.

Activitat 3: Estudi d'una empresa informàtica sobre l'eficiència d'un servidor Web.

Contrast d'hipòtesis paramètriques. Contrast de la mitjana d'una població normal amb variància desconeguda. Estimació per intervals: Interval de confiança.

En una empresa informàtica es desitja mesurar l'eficiència d'un servidor Web. Per a això, mesuren el temps d'espera del client entre la petició que aquest fa i la resposta que li dóna el servidor.

Els temps d'espera (en mil · lisegons) del servidor (TA) per 50 peticions es troben en el fitxer ActR07TA.csv.

Suposant normalitat, contesteu a les preguntes següents:

a) Podem considerar que el temps mitjà d'espera del servidor A és de 9 mil · lisegons? Raoneu la resposta. Trobeu el *p-valor* del contrast. Preneu $\alpha = 0.1$.
b) Trobar un interval de confiança per al temps mitjà d'espera del servidor Al 95% de confiança.

Solució

a) Llegim les dades del fitxer:

```
test22<-read.table("Act07TA.csv", skip = 1, sep=";" , dec=",", header=TRUE)
head(test22)
```

```
##      TA
## 1  9.67
## 2  9.62
## 3  9.50
## 4 10.88
## 5  8.94
## 6 10.59
```

Si busquem la mitjana i la desviació típica de la mostra, trobarem l'estadístic de contrast amb la instrucció següent:

```
t<-(mean(test22$TA)-9)/(sd(test22$TA)/sqrt(length(test22$TA)))
t
```

```
## [1] 7.344513
```

Ara ja podem trobar el p-valor:

```
2*pt(t, df=49,lower.tail=FALSE)
```

```
## [1] 1.944584e-09
```

Com que el nivell de significació (0.1) és major que el p-valor (1.945×10^{-9}), rebutjarem la hipòtesi nul · la i acceptarem l'alternativa, és a dir, no podem considerar que el temps mitjà d'espera del servidor A és de 9 ms. També es podria haver fet amb un altre mètode, usant directament la instrucció corresponent de R:

```
t.test(test22$TA,mu=9,alternative="two.sided",conf.level=0.9)
```

```
##
##  One Sample t-test
##
## data:  test22$TA
## t = 7.3445, df = 49, p-value = 1.945e-09
## alternative hypothesis: true mean is not equal to 9
## 90 percent confidence interval:
##   9.721565 10.148435
## sample estimates:
## mean of x
##     9.935
```

b) Hem de trobar l'interval de confiança al 95% per a la mitjana de temps d'espera del servidor A. Hem d'utilitzar la següent fórmula:

$$\left(\bar{x}_A - t_{\alpha/2,n-1}\frac{s_A}{\sqrt{n_A}}, \bar{x}_A + t_{\alpha/2,n-1}\frac{s_A}{\sqrt{n_A}}\right)$$

Això el podem calcular, utilitzant R, així:

```
error_std<-sqrt(var(test22$TA))/(sqrt(length(test22$TA)))
error_std
```

```
## [1] 0.1273059
```

```
valor_critic<-qt(0.025,df=49,lower.tail=FALSE)
ci<-c(mean(test22$TA)-valor_critic*error_std,mean(test22$TA)+valor_critic*error_std)
ci
```

```
## [1]  9.679169 10.190831
```

Per tant, l'interval de confiança per al temps mitjà d'espera d'A és $(9.67917, 10.19083)$. Podem veure que el 9 no està dins de l'interval, i per tant, es confirma el que s'ha demostrat l'apartat a) (encara que amb un nivell de significació diferent). També es pot trobar aquest interval amb la instrucció:

```
t.test(test22$TA,mu=9,alternative="two.sided",conf.level=0.95)
```

```
##
##  One Sample t-test
##
## data:  test22$TA
```

```
## t = 7.3445, df = 49, p-value = 1.945e-09
## alternative hypothesis: true mean is not equal to 9
## 95 percent confidence interval:
##   9.679169 10.190831
## sample estimates:
## mean of x
##     9.935
```

Activitat 4: Estudi sobre la qualitat d'un sistema operatiu

Contrast d'hipòtesis paramètriques. Inferència sobre dades aparellades. Contrast de la mitjana sobre poblacions normals. Estimació per intervals: Interval de confiança.

Hem demanat a 10 informàtics que avaluïn, sobre la base d'uns criteris preestablerts, la qualitat d'un determinat sistema operatiu. Les puntuacions varien entre un mínim de 0 i un màxim de 15. Passats tres mesos, les mateixes 10 persones repeteixen el procés d'avaluació. Els resultats obtinguts, que introduirem en les columnes V1 i V2 i es troben en el fitxer OPSIS.csv, i són els següents:

```
test4<-read.table("ActR07OPSIS.csv", skip = 0, sep=";" , dec=",", header=TRUE)
test4
```

```
##      V1   V2
## 1  13.2 14.0
## 2   8.2  8.8
## 3  10.9 11.2
## 4  14.3 14.2
## 5  10.7 11.8
## 6   6.6  6.4
## 7   9.5  9.8
## 8  10.8 11.3
## 9   8.8  9.3
## 10 13.3 13.6
```

El nostre objectiu és doble: d'una banda, pretenem calcular un interval de confiança, a nivell del 95%, per $\mu_{V1} - \mu_{V2}$; per un altre, contrastar les hipòtesis: $H_0 : \mu_{V1} - \mu_{V2} = 0$ vs. $\mu_{V1} - \mu_{V2} \neq 0$.

Solució

Hem de fer un contrast de diferència de mitjanes aparellades. Usem *t.test*.

```
with(test4,t.test(V1,V2, alternative = 'two.sided', conf.level = .95, paired = TRUE ))
```

```
##
##  Paired t-test
##
## data:  V1 and V2
## t = -3.3489, df = 9, p-value = 0.008539
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6869539 -0.1330461
## sample estimates:
## mean of the differences
##                   -0.41
```

Els resultats obtinguts ens diuen que, sobre la base de les observacions registrades, hi ha una probabilitat de 0,95 que $\mu_A - \mu_B$ sigui un valor de l'interval $(-0.687, -0.133)$. A més, amb un p-valor de 0.009 també podem afirmar que hi ha indicis suficients per a descartar la hipòtesi nul · la. Per tant, sembla assenyat pensar que les dues mitjanes poblacionals són diferents.

Cal destacar que aquesta conclusió és coherent amb què el valor 0 no estigui inclòs en l'interval de confiança trobat per a la diferència d'ambdues mitjanes.

Activitat 5: Estudi sobre la mitjana del temps de transferència de fitxers.

Contrast d'hipòtesi. Contrast de la mitjana amb variància desconeguda. Contrast de la mitjana amb variància coneguda.

El responsable de comunicacions d'una empresa informàtica afirma que la mitjana del temps de transferència per FTP d'un fitxer de 2Mb és superior a 30 segons. Per comprovar aquesta afirmació aquest va prendre una mostra del temps de transferència de 12 fitxers de 2Mb, i es va obtenir que la mitjana i la desviació estàndard mostral valen $\bar{x} = 30.2$, $s = 1.833$ (en segons).

a) Suposant que el temps de transferència es distribueix normalment, a partir de les dades mostrals obtingudes, tenim les suficients evidències per creure l'afirmació del responsable? (Preneu $\alpha = 0.05$). Trobeu el *p-valor* del contrast.
b) Si a més de disposar d'aquestes observacions ens haguessin donat com a informació addicional (obtinguda d'experiències prèvies) que la variància del temps de transferència és de segons $\sigma^2 = 9.2^2$, hauríem arribat a la mateixa conclusió que a l'apartat anterior? Trobeu el p-valor del contrast.

A cada apartat cal seguir el següent esquema:

1. Especificar el tipus de contrast que fa.
2. Indicar la hipòtesi nul · la i alternativa.
3. Indicar la fórmula de l'estadístic de contrast així com la seva distribució de probabilitat.
4. Usar el programari per trobar els p-valoris corresponents i els valors crítics.

Solució

a) Hem de fer un contrast d'una mitjana amb variància desconeguda. Les hipòtesis nul · la i alternativa són:

$$H_0 : \mu = 30$$
$$H_1 : \mu > 30$$

on μ representa la mitjana del temps de transferència per FTP d'un fitxer de 2Mb.
L'estadístic de contrast és:

$$t = \frac{\bar{x} - 30}{s/\sqrt{12}}$$

on $\bar{x}$ és la mitjana mostral i s és la desviació estàndard mostral. La distribució de t és la de t de Student amb 11 graus de llibertat. La mitjana i la desviació estàndard mostral valen respectivament: $\bar{x} = 30.2$, $s = 1.833$. El valor de l'estadístic de contrast val:

```
t<-(30.2-30)/(1.833/sqrt(12))
t
```

```
## [1] 0.3779707
```

El *p-valor*:

```
pt(0.3779707,df=11,lower.tail=FALSE)
```

```
## [1] 0.3563222
```

```
qt(0.05,df=11,lower.tail=FALSE)
```

```
## [1] 1.795885
```

Com que $t = 0.3779707 < t_{0.05,11} = 1.795885$ acceptem la hipòtesi nul · la i concloem que l'afirmació del responsable no es pot considerar certa.

b) Hem de fer un contrast d'una mitjana amb variància coneguda. Les hipòtesis nul · la i alternativa són:

$$H_0 : \mu = 30$$
$$H_1 : \mu > 30$$

on μ representa la mitjana del temps de transferència per FTP d'un fitxer de 2Mb. L'estadístic de contrast és:

$$z = \frac{\bar{x} - 30}{\sigma/\sqrt{12}}$$

on $\bar{x}$ és la mitjana mostral i σ és la desviació estàndard poblacional. La distribució de z és la d'una normal $N(0, 1)$. La mitjana i la desviació estàndard poblacional valen respectivament: $\bar{x} = 30.2$, $\sigma = \sqrt{9.2} = 3.03$. El valor de l'estadístic de contrast val:

```
z<-(30.2-30)/(3.03/sqrt(12))
z
```

```
## [1] 0.2286536
```

El *p-valor* val:

```
pnorm(0.2286536,lower.tail=FALSE)
```

```
## [1] 0.4095691
```

I el valor crític:

```
qnorm(0.05,lower.tail=FALSE)
```

```
## [1] 1.644854
```

En aquest cas acceptem la hipòtesi nul · la i concloem que l'afirmació del responsable no és certa.

Activitat 6: Contrast de la mitjana de la diferència en el temps d'arrencada d'un ordinador amb antivirus i sense antivirus.

Contrast d'hipòtesi. Contrast de la mitjana amb variància desconeguda. Mostres aparellades.

Es qüestiona si el temps (en segons) d'arrencada d'un ordinador amb l'antivirus TRONX és més lent que si l'ordinador no té l'antivirus. Amb l'objectiu de fer aquest estudi s'ha mesurat el temps d'arrencada de 10 ordinadors sense tenir cap antivirus instal · lat. Després s'ha instal · lat l'antivirus en tots aquests ordinadors i s'ha tornat a mesurar el temps d'arrencada. S'han obtingut els resultats recollits en ActR07ANTIV.csv.

Suposant que el temps d'arrencada d'un ordinador es distribueix normalment, decideix amb $\alpha = 0.05$ si l'existència en l'ordinador de l'antivirus fa que els temps d'arrencada sigui major. Trobeu el p-valor del contrast.

Per resoldre el problema:

a) Especificar el tipus de contrast que fa.
b) Indicar la hipòtesi nul · la i alternativa.
c) Realitzar el test corresponent.

Solució

Llegim les dades del fitxer:

```
test6<-read.table("ActRO7ANTIV.csv", skip = 2, sep=";" , dec=",",header=TRUE)
test6
```

```
##      SIN  CON
## 1  47.80 46.7
## 2  70.54 80.8
## 3  59.30 63.9
## 4  46.00 43.9
## 5  54.40 56.7
## 6  70.10 80.1
## 7  56.90 60.4
## 8  72.60 71.2
## 9  62.30 60.3
## 10 63.30 69.8
```

Efectuem el test de dades aparellades amb alternativa "$\leq$" o "less":

```
t.test(test6$SIN,test6$CON, alternative="less", mu=0, paired=TRUE, conf.level=0.95)
```

```
##
##  Paired t-test
##
## data:  test6$SIN and test6$CON
## t = -2.0278, df = 9, p-value = 0.0366
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.2934189
## sample estimates:
## mean of the differences
##                  -3.056
```

El valor de l'estadístic de contrast val $t = -2.0278$ i el p-valor $p = 0.0366$. Per tant rebutgem la hipòtesi nul · la i arribem a la conclusió que el temps d'arrencada és major amb l'antivirus.

Activitat 7: Contrast de la proporció d'usuaris d'un programari.

Contrast d'hipòtesi. Contrast de la proporció en mostres grans.

En una universitat el centre de càlcul afirma que la proporció de terminals que usen el programa estadístic R a primera hora del matí és inferior al 50%. Agafada una mostra de 700 terminals a primera hora del matí es va observar que 410 estaven usant el programa R. Amb $\alpha = 0.01$ podem estar d'acord amb l'afirmació del centre de càlcul? Trobeu el p-valor del contrast.

Solució

Hem de fer un contrast de proporcions. Sigui p la proporció de terminals que usen R a primera hora del matí. El contrast d'hipòtesi és:

$$H_0 : p = 0.5$$
$$H_1 : p < 0.5$$

L'estadístic de contrast val

$$z = \frac{\hat{p} - 0.5}{\sqrt{\frac{0.5 \cdot 0.5}{700}}}$$

on $\hat{p}$ és la proporció de terminals que usen R a primera hora del matí i val $410/700 = 0.586$. L'estadístic de contrast segueix aproximadament la distribució normal N(0,1) si la mesura de la mostra és suficientment gran com és el nostre cas. El valor de l'estadístic de contrast és.

```
pbar<-410/700
```

```
z<-(pbar-0.5)/sqrt(0.5*0.5/700)
z
```

```
## [1] 4.535574
```

Calculem el *p-valor*:

```
pnorm(4.535574)
```

```
## [1] 0.9999971
```

i el valor crític

```
qnorm(0.99)
```

```
## [1] 2.326348
```

Per tant hem d'acceptar la hipòtesi nul · la i concloure que la proporció és el 50% i no inferior al 50%.