Pràctica Permutacions-Bootstrap

Víctor Navarro Garcés i Laura Julià Melis

8 d' Abril de 2019

Contents

Enunciat, lectura de dades i aproximació descriptiva.

L'any 1961 es va realitzar un assaig clínic en un hospital psiquiàtric, amb una droga anomenada Stelazine (trifluoperazine) per al tractament de l'esquizofrènia crònica. Amb la finalitat de definir blocs per fer millor les comparacions dins d'ells, es van formar 24 parelles de pacients, procurant que dins de cada parella les característiques dels pacients fossin molt similars. A un membre de la parella es va tractar amb Stelazine durant 3 mesos mentre que l'altre membre va servir de control, va rebre la medicació habitual en aquella època.

L'assignació de tractament dins de cada parella es va realitzar a l'atzar. La finalitat del tractament (tant Stelazine com a control) era pal · liar els símptomes d'ansietat associats a les fases agudes de la malaltia. Abans i després del tractament, es va mesurar el grau d'ansietat de cada pacient, segons una escala basada en observar el seu comportament dins de l'hospital. Els resultats estan resumits en el fitxer "Ansiedad.txt". Cada fila representa una parella de pacients, la primera columna representa la variació (després - abans) en el grau d'ansietat en el membre tractat amb Stelazine, i la segona columna la variació en el membre "control". Valors negatius (o propers a zero) són els desitjables, indiquen que l'ansietat ha disminuït o, com a mínim, s'ha mantingut estable. A causa dels efectes secundaris de la medicació i a altres causes, només 16 parelles van completar l'assaig.

Lectura de dades:

```
dd <- read.table("Ansiedad.txt", header=T)
dd
```

```
##    Pareja Stelazine Control
## 1       1      0.80   -0.40
## 2       2      0.40    0.10
## 3       3      0.55   -0.10
## 4       4     -0.90    0.60
## 5       5      0.34    0.20
## 6       6      1.42    0.78
```

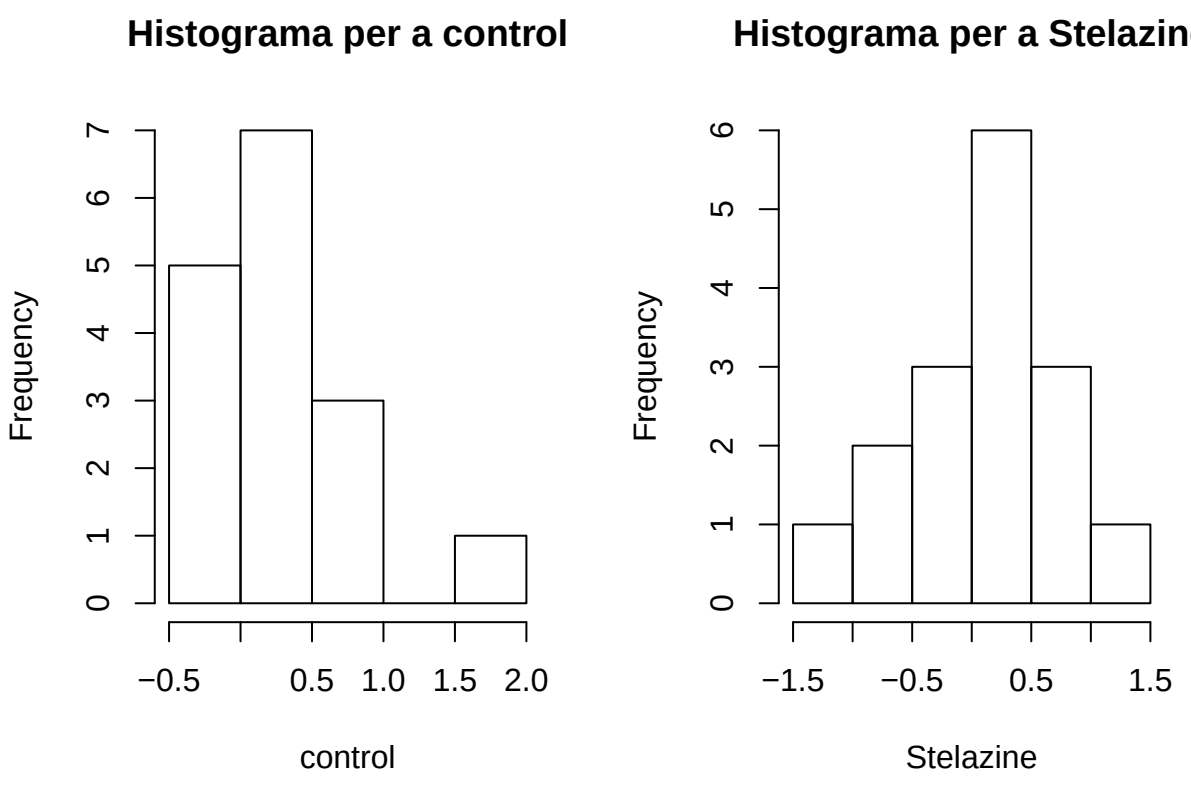
```
## 7       7      0.30    1.74
## 8       8     -0.29    0.64
## 9       9      0.53    0.42
## 10     10     -0.45    0.01
## 11     11      0.15    0.14
## 12     12     -0.55   -0.19
## 13     13      0.12   -0.05
## 14     14      0.03    0.04
## 15     15     -1.16   -0.15
## 16     16     -0.16    0.48
```

```r
control <- dd$Control # Vector de valors del grup control
Stelazine <- dd$Stelazine # Vector de valors del grup Stelazine
auc <- c(control, Stelazine)

# Tamanys mostrals
n1 <- length(control)
n2 <- length(Stelazine)
N <- n1 + n2
```

Estudiem la distribució de les dades:

```r
# Indicis de normalitat (o no) de les dades
par(mfrow=c(1,2))
hist(control, main="Histograma per a control")
hist(Stelazine, main="Histograma per a Stelazine")
```



Comparant els dos histogrames, s'observa com les dades de la variable "Control" no semblen seguir una distribució normal mentre que les dades de "Stelazine" sí. Per tant, ens comencem a fer una idea de que els dos grups pot ser no provenguin d'una mateixa distribució. Ho comprovem a continuació mitjançant un test de permutacions.

Pregunta 1.

Fixant un nivell de significació del 5% i mitjançant un test de permutacions, determina si Stelazine controla millor l'ansietat que el tractament habitual.

Hipòtesi.

Si indiquem com a F_C la distribució de la variable aleatòria Y="variació en el grau d'ansietat" dels membres que van rebre la medicació habitual en aquella època i com F_S la distribució del tractament amb la droga Stelazine, la hipòtesis nul · la és:

$$H_o : F_C = F_S$$

Valor estadístic mostral.

Es tracta d'un cas balancejat de mostres independents en el que d'un total de N=32 pacients, n1=16 van rebre la medicació habitual en aquella època (Control) i n2=16 la droga Stelazine. Per tant, té sentit fer servir com a estadístic la suma de valors d'un dels grups, per exemple, la del grup Control.

```
# Calculem l'estadístic suma de valors del grup "Control"
sum <- sum(control)
sum
```

```
## [1] 4.26
```

Nombre de permutacions.

Cal permutar lliurement les N=32 observacions, pero $32! = 2.631308e^{+35}$ són un nombre enorme de permutacions possibles.

```
# Nombre de permutacions.
factorial(N)  # enorme
```

```
## [1] 2.631308e+35
```

Com que no és necessari tenir en compte de les observacions, en realitat podem permutar fent $\frac{32!}{16!16!} = 601080390$, però segueixen essent massa computacionalment de manera que sembla més raonable realitzar un enfoc de prova de permutacions de Montecarlo.

```
choose(N, n1) # també és massa gran
```

```
## [1] 601080390
```

Càlcul del pvalor estimat.

Amb una única operació (`sum.perm`) s'han generat `nperm=99999` possibles permutacions de les 32 observacions i, per a cadascuna d'aquestes permutacions, s'ha calculat l'estadístic en el grup de "control", que té grandària $n_1 = 16$.

```
# Generem nperm permutacions aleatòries i calculem l'estadístic per a cadascuna de les combinacions
nperm <- 99999
set.seed(123)
sum.perm <- replicate(nperm, sum(auc[sample(1:N, size=n1)]))
sum.perm[1:10] # Valor de l'estadístic per a les 10 primeres remostres.
```

```
##  [1]  0.63  2.78  4.48  3.14  2.78  4.87 -0.28  4.61  2.72  2.90
```

I finalment es calcula el pvalor unilateral de que el grup "control" és superior a l'altre grup.

```
# Estimador proposat per Dwass com a estimador del p-valor.
(sum(sum.perm >= sum)+1)/ (nperm+1) # test unilateral
```

```
## [1] 0.17806
```

Conclusió.

Amb aquest test de permutacions es desitja conéixer si la droga Stelazine controla millor l'ansietat que el tractament habitual. Si Stelazine controla millor l'ansietat, significa que les variacions "després - abans" del grup d'Stelazine haurien de ser inferiors a les del grup de Control. S'ha obtingut un pvalor de 0.17806, superior al nivell de significació utilitzat ($\alpha = 0.05$) de manera que no hi ha suficients evidències per poder rebutjar H_o i per tant, no hi ha motius per pensar que Stelazine sigui millor per controlar el grau d'ansietat.

Pregunta 2

Quantifica la diferència entre Stelazine i el tractament control mitjançant un interval de confiança al 95% per a la diferència de mitjanes. Calcula l'interval de confiança associat al test de permutacions anterior, i els següents intervals de confiança bootstrap: percentil, BCa, bootstrap-t i bootstrap-t simetrizado.

Interval de confiança per a la diferència de mitjanes.

A. Test de permutacions per a la diferència de mitjanes.

En primer lloc cal crear una funció que calculi la diferència de mitjanes entre dos grups.

```
# Creem funció que calcula la diferència de mitjanes.
diff.means <- function(index, dades){
  mean(dades[index])-mean(dades[-index])
}
```

Llavors es calcula l'estadístic **diferència de mitjanes** entre els grups "control" i "Stelazine" per les dades originals, les de la mostra.

```
# Calculem l'estadístic sobre la mostra real.
dmeansReal <- diff.means(1:n1, auc)
dmeansReal
```

```
## [1] 0.195625
```

Tal com s'ha comentat en l'apartat anterior, amb aquestes grandàries mostrals és més realitzable simular una mostra aleatòria de permutacions. Així doncs, s'han generat `nperm=99999` permutacions de les 32 observacions i per a cadascuna d'elles, s'ha calculat l'estadístic `diff.means`.

```
# Generem nperm permutacions aleatòries i calculem l'estadístic per a cadascuna de les combinacions
set.seed(123)
dmeansPerm <- replicate(nperm, diff.means(sample(1:N, size=n1),auc))
dmeansPerm[1:10] # Valor de l'estadístic per a les 10 primeres remostres.
```

```
##  [1] -0.258125  0.010625  0.223125  0.055625  0.010625  0.271875 -0.371875
##  [8]  0.239375  0.003125  0.025625
```

Per a l'estimació del pvalor s'ha fet servir l'estimador proposat per Dwass el qual, com era d'esperar, ens ha donat el mateix valor que en el pvalor obtingut mitjançant la suma de valors en l'apartat anterior.

```r
# Estimador proposat per Dwass com a estimació del p-valor.
(sum(dmeansPerm >= dmeansReal)+1)/(nperm+1)
```

```
## [1] 0.17806
```

B. Interval de confiança.

S'ha obtingut l'IC a partir de la funció `permTS`, a la qual se li ha indicat que realitzi el test unilateral de la diferència de mitjanes pel mètode de Montecarlo fent 99999 permutacions.

```r
library(perm)
permTS(control, Stelazine, alternative = "greater", method = "exact.mc", control = permControl(nmc = 99
```

```
##
##  Exact Permutation Test Estimated by Monte Carlo
##
## data:  control and Stelazine
## p-value = 0.179
## alternative hypothesis: true mean control - mean Stelazine is greater than 0
## sample estimates:
## mean control - mean Stelazine
##                      0.195625
##
## p-value estimated from 99999 Monte Carlo replications
## 95 percent confidence interval on p-value:
##  0.1765804 0.1813418
```

L'interval de confiança és $[0.1765804, 0.1813418]$, el qual conté el valor de l'estadístic sobre les dades de la mostra (0.195625).

Interval de confiança bootstrap-t.

Sigui $\hat{\delta} = \bar{X_C} - \bar{X_S}$ la diferència entre les mitjanes mostrals dels grups "control" i "Stelazine" i $\delta = \mu_C - \mu_S$ la corresponent diferència de mitjanes poblacionals, l'estadístic t es defineix com:

$$t = \frac{\hat{\delta} - \delta}{s\hat{e}\delta},$$

on $s\hat{e}\delta$ representa l'estimació de l'error estàndard de $\hat{\delta}$.

Sota les suposicions de normalitat i igualtat de variàncies, la distribució t no depèn de paràmetres desconeguts i és coneguda. Per tant, es possible deteminar $t_{\alpha/2}$ i $t_{1-\alpha/2}$ tals que:

$$P\left[t_{\alpha/2} \leq \frac{\hat{\delta} - \delta}{s\hat{e}\delta} \leq t_{1-\alpha/2}\right] = 1 - \alpha$$

$$P\left[\hat{\delta} - t_{1-\alpha/2} \cdot s\hat{e}\delta \leq \delta \leq \hat{\delta} - t_{\alpha/2} \cdot s\hat{e}\delta\right] = 1 - \alpha$$

Garantint així l'expressió de l'interval de confiança de nivell $1 - \alpha$:

$$\left[\hat{\delta} - t_{1-\alpha/2} \cdot s\hat{e}\delta, \quad \hat{\delta} - t_{\alpha/2} \cdot s\hat{e}\delta\right]$$

Des de la perspectiva bootstrap, el que s'ha hagut de fer és aproximar la distribució de t generant `B=10000` remostres i calculant l'estadístic sobre cada una d'elles, obtenint t^*:

$$t^* = \frac{\hat{\delta^*} - \hat{\delta}}{s\hat{e}\delta^*}$$

on $\hat{\delta}^*$ i $s\hat{e}\delta^*$ són la diferència de mitjanes i l'error estàndard calculats sobre la remostra.

D'aquesta manera, a partir de t^*, ja s'han pogut aproximar els valors de $t_{\alpha/2}$ i $t_{1-\alpha/2}$ i, conseqüentment, calcular l'interval de confiança demanat. A continuació s'expliquen les passes seguides.

Primer s'han creat les funcions necessàries per al càlculs que s'hauran de fer posteriorment: `tStat` per calcular l'estadístic t per la diferència de mitjanes:

```
# Funció que calcula l'estadístic t per la diferència de mitjanes
tStat <- function(x1, x2, delta = 0, var.equal = FALSE){
  t.test(x1, x2, mu = delta, var.equal = var.equal)$statistic
}
```

I `se.diffMeans` per al càlcul de l'error estàndard de la diferència de mitjanes mostrals:

```
# Funció que calcula l'error estàndard de la diferència de mitjanes mostrals
se.diffMeans <- function(x1, x2, var.equal = FALSE){
  if(var.equal){
    m1 <- mean(x1)
    m2 <- mean(x2)
    result <- sqrt((length(x1)-1)*(sum((x1-m1)^2))+(length(x2)-1)*(sum((x2-m2)^2)))/(length(x1)+length(x
    return(result)
  } else{
    return(sqrt(var(x1)/length(x1)+var(x2)/length(x2)))
  }
}
```

S'han calculat la diferència de mitjanes i l'error estàndard per a las dades mostrals.

```
# Càlcul de la diferència de mitjanes i l'error estàndard per a las dades mostrals.
deltaEstim <- mean(control)-mean(Stelazine)
deltaEstim
```

```
## [1] 0.195625
```

```
seEstim <- se.diffMeans(control, Stelazine, var.equal = T) # suposem igualtat de variàncies
seEstim
```

```
## [1] 0.4146137
```

Després, s'han generat les `B=10000` remostres i, per a cada una d'elles, l'estadístic t^*

```
# Generació de remostres i càlcul de l'estadístic t per a cada una.
alpha <- 0.05
B=10000 # nombre de remostres
t.boot<- replicate(B,
                   tStat(
                     sample(control, replace=TRUE),
                     sample(Stelazine, replace=TRUE),
                     delta = deltaEstim, var.equal = TRUE
                   )
          )
t.boot[1:10] # Valor de l'estadístic per a les 10 primeres remostres.
```

```
##          t          t          t          t          t          t
## -0.3728742 -2.1189701 -0.1545476 -1.0191627  0.6068481 -0.1370912
##          t          t          t          t
## -0.5046663  0.2312815  0.8171677  0.0686474
```

Ja es pot obtenir l'interval bootstrap-t:

```r
# Interval bootstrap-t:
IC <- deltaEstim - quantile(t.boot, probs = c(1 - alpha/2, alpha/2)) * seEstim
names(IC) = NULL
attr(IC, "conf.level") = 1 - alpha
IC
```

```
## [1] -0.5908289  1.1086476
## attr(,"conf.level")
## [1] 0.95
```

L'interval és $[-0.5978565, 1.1104068]$, el qual inclou el valor 0. Això conduiria a la conclusió de que existeixen diferències entre Stelazine i el tractament control.

Interval de confiança bootstrap-t simetritzat.

Amb la realització d'aquest tipus d'interval, s'està considerant que la distribució de t es simètrica respecte de zero (i no que ambdues cues de la distribució siguin equiprobables o simètriques). Així, es desitja buscar una constant $t_{1-\alpha} > 0$ tal que:

$$P\big[|t| \leq t_{1-\alpha}\big] = 1 - \alpha$$

és a dir, que entre $-t_{1-\alpha}$ i $t_{1-\alpha}$ hi quedi una probabilitat $1 - \alpha$.

L'expressió de l'interval de confiança simetritzat és:

$$\Big[\hat{\delta} - t_{1-\alpha} \cdot s\hat{e}\delta, \quad \hat{\delta} + t_{1-\alpha} \cdot s\hat{e}\delta\Big]$$

Es duu a terme l'estimació del valor $t_{1-\alpha}$ mitjançant remostratge bootstrap. Cal notar que això s'ha dut a terme a partir dels valors t^* (`t.boots`) obtinguts en la secció anterior i pressupossant, com abans, igualtat de variàncies.

```r
t1_alpha = quantile(abs(t.boot), probs = 1 - alpha)
# Interval bootstrap-t simetritzat:
ICsim <- deltaEstim - c(t1_alpha, -t1_alpha) * seEstim
names(IC) = NULL
attr(IC, "conf.level") = 1 - alpha
ICsim #[-0.6574345,  1.0486845]
```

```
##        95%        95%
## -0.6495977  1.0408477
```

L'interval és $[-0.6574345, 1.0486845]$, també inclou el valor 0.

Interval de confiança percentil.

Un altre cop, s'aprofitaran els valors t bootstrap obtinguts abans:

```r
# Interval bootstrap-p:
icBoot.perc = quantile(t.boot, probs = c(alpha/2, 1 - alpha/2))
names(icBoot.perc) = NULL
attr(icBoot.perc, "conf.level") = 1 - alpha
icBoot.perc
```

```
## [1] -2.202105  1.896836
## attr(,"conf.level")
## [1] 0.95
```

Interval de confiança BCa.

Obtenció de N rèpliques jackknife de la diferència de mitjanes.

```
dif_i <- vector()
for(i in 1:n1){
  dif_i[i] <- mean(dd[-i,"Control"])-mean(dd[-i,"Stelazine"])  #càlcul de la diferència de mitjanes eli
}
dif_i # N rèpliques jackknife de la diferència de mitjanes.
```

```
##  [1] 0.2886667 0.2286667 0.2520000 0.1086667 0.2180000 0.2513333 0.1126667
##  [8] 0.1466667 0.2160000 0.1780000 0.2093333 0.1846667 0.2200000 0.2080000
## [15] 0.1413333 0.1660000
```

Càlcul de l'acceleració de l'error estàndard.

```
# Càlcul de l'acceleració de l'error estàndard.
resta<-mean(dif_i)-dif_i
a <- sum(resta^3)/(6*sum(resta^2)^1.5)
a
```

```
## [1] 0.006637686
```

Càlcul de l'estadístic t sobre la mostra real:

```
# Estadístic t sobre la mostra real:
tReal<- t.test(control, Stelazine, mu = 0, var.equal = T)$statistic
tReal
```

```
##         t
## 0.9436496
```

Càlcul del factor de correcció del biaix.

```
# Càlcul del factor de correcció del biaix.
z0 <- qnorm(sum(t.boot>=tReal)/B)
z0
```

```
## [1] -0.987087
```

Finalment, ja podem calcular els percentils (valors crítics) per al cas de bootstrap amb acceleració i correcció de biaix, i l'interval de confiança:

```
# Valors crítics:
prob1 <-pnorm(z0 + (z0+qnorm(alpha/2))/(1- a*(z0+qnorm(alpha/2))))
prob2 <-pnorm(z0 + (z0-qnorm(alpha/2))/(1- a*(z0-qnorm(alpha/2))))

# Interval BCa:
icBoot.Bca = quantile(t.boot, probs = c(prob1, prob2))
names(icBoot.Bca) = NULL
attr(icBoot.Bca, "conf.level") = 1 - alpha
icBoot.Bca
```

```
## [1] -4.75384541 -0.01996617
## attr(,"conf.level")
## [1] 0.95
```

L'interval obtingut és: $[-4.75384541, -0.01996617]$.