Solucions als exercicis pràctics metodologia Bootstrap

Jordi Ocaña , Sergi Civit

14 de maig de 2018

1 Exercici 1

1.1 Enunciat

L'origen de la civilització etrusca encara és un misteri per als antropòlegs. En concret, una qüestió que es planteja és si eren originaris de la península italiana o si procedien d'un altre lloc. Es va pensar que una forma de respondre a aquesta pregunta seria comparar els actuals italians amb les restes etrusques, mitjançant un estudi antropomètric. Aquí utilitzarem una part petita d'aquest estudi. En concret, es va mesurar, en mil.límetres, l'amplada màxima del crani de 8 restes de barons homes i la de 10 italians, tots ells escollits a l'atzar. Els resultats obtinguts van ser:

etruscs	141	132	154	142	141	150	134	140		
italians	133	138	136	125	135	130	127	131	116	128

A partir d'aquestes dades, es podria concloure l'existència de diferències significatives entre italians actuals i etruscos a un nivell de significació 0,05? Pots suposar que, si hi ha diferències, són degudes al paràmetre de localització (mitjana, mediana...) i no al paràmetre de dispersió, però no hauries de suposar cap forma concreta per a la distribució de l'amplada màxima del crani. Realitza les següents proves i compara'n els resultats:

```
> # Les dades en dos vectors:
> etruscos = c(141, 132, 154, 142, 141, 150, 134, 140)
> italians = c(133, 138, 136, 125, 135, 130, 127, 131, 116, 128)
> cranis = c(etruscos, italians)
> n1 = length(etruscos)
> n2 = length(italians)
> n = c(n1, n2)
> N = sum(n)
```

- Prova de permutacions: calcula el p-valor exacte si utilitzem com a estadístic de test la diferència de mitjanes mostrals. Per aquestes dades,

¿tindria sentit utilitzar com a estadístic de test solament la suma dels valors dels etruscos? (JA FET)

- **Mètode bootstrap**: Ens interessaria tenir una idea aproximada de com de diferents eren els homes dels actuals italians. Determina un interval de confiança bootstrap-t simetritzat, bootstrap-t (cues equiprobables) i boostrap-p per a la diferència de mitjanes. Comparal's. Aquests intervals, contradiu o no els resultats dels tests anteriors?

1.2 Solució

En aquest tipus d'estudis, com a mínim en una primera aproximació, el més freqüent és que els investigadors vulguin fer una prova bilateral, els interessa la presència de diferències i prou.

1.2.1 Prova de permutacions

Amb aquestes dades no seria correcte utilitzar com a estadístic pel test de permutacions la suma dels valors d'un dels grups ja que els grups no són balancejats (8 i 10 dades).

```
> # Funció pel càlcul de la diferència de mitjanes:
> diff.means <- function(indexs, vector.dades) {
+   mean(vector.dades[indexs]) - mean(vector.dades[-indexs])
+ }
> # Diferència de mitjanes per les dades originals:
> dMeans = diff.means(1:n1, cranis)
> # Indexos de les combinacions de N elements agafats en grups de n1:
> combs = combn(N, n1)
> # Càlcul de la diferència de mitjanes per cada combinació:
> dMeansPerm = apply(combs, 2, diff.means, vector.dades = cranis)
> # p-valor del test bilateral:
> sum(abs(dMeansPerm) >= abs(dMeans)) / ncol(combs)

[1] 0.001348325

>
```

Novament la conclusió final és que rebutgem H_0.

1.2.2 Interval de confiança bootstrap-t

Sigui $\hat{\delta} = \bar{X}_1 - \bar{X}_2$ la diferència entre les mitjanes mostrals d'etruscs i italians, i $\delta = \mu_1 - \mu_2$ la corresponent diferència de mitjanes poblacionals. $\acute{s}e_{\hat{\delta}}$ representa l'estimació de l'error estàndard de $\hat{\delta}$.

L'estadístic t es defineix com:

$$t = \frac{\hat{\delta} - \delta}{\acute{s}e_{\hat{\delta}}}.$$

Si fos possible determinar dues constants $t_{\alpha/2}$ i $t_{1-\alpha/2}$ tals que:

$$Pr\{t_{\alpha/2} \leq \frac{\hat{\delta} - \delta}{\hat{se}_{\hat{\delta}}} \leq t_{1-\alpha/2}\} = 1 - \alpha$$

seria possible obtenir:

$$Pr\{\hat{\delta} - t_{1-\alpha/2}\hat{se}_{\hat{\delta}} \leq \delta \leq \hat{\delta} - t_{\alpha/2}\hat{se}_{\hat{\delta}}\} = 1 - \alpha,$$

és a dir, quedaria garantit que

$$[\hat{\delta} - t_{1-\alpha/2}\hat{se}_{\hat{\delta}}, \hat{\delta} - t_{\alpha/2}\hat{se}_{\hat{\delta}}]$$

seria un interval de confiança de nivell $1 - \alpha$.

Això és possible sota la suposició de normalitat i d'igualtat de variàncies, cas al qual la distribució de t no depèn de paràmetres desconeguts i és coneguda.

Sota l'enfoc bootstrap aproximarem la distribució de t generant B mostres (remostres) a partir d'una estimació de la distribució de les dades i calculant t sobre cada remostra. Així obtenim una mostra gran de valors de t simulats, $t_1^*, t_2^*, \ldots, t_B^*$. Per cada réplica de la simulació bootstrap es calcularà una valor t^*,

$$t^* = \frac{\hat{\delta}^* - \hat{\delta}}{\hat{se}_{\hat{\delta}}^*}.$$

A la simulació bootstrap $\hat{\delta}^*$ i $\hat{se}_{\hat{\delta}}^*$ corresponen a la diferència de mitjanes mostrals i a l'error estàndard, respectivament, calculats sobre cada remostra. $\hat{\delta}$ fa el paper del paràmetre poblacional.

Un cop obtinguts els B valors de t^*, es poden aproximar els valors de $t_{\alpha/2}$ i $t_{1-\alpha/2}$ a partir dels seus corresponents quantils mostrals.

Definirem algunes funcions necessàries pels càlculs que hem de fer. Quan fem bootstrap (o quan fem una prova de permutacions) casi sempre podem pensar indistintament en termes dels valors de la mostra directament o en termes dels seus índexos, és a dir, es pot enfocar el càlcul dels estadístics pensant que de la mostra $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ hem agafat a l'atzar els elements $x_3, x_1, \ldots$ o que hem seleccionat els índexs $3, 1$ etc. Només cal que el càlcul dels estadístics estigui definit d'acord amb l'enfoc concret utilitzat. Aquí potser seria més clar i directe pensar directament en termes dels valors de la mostra:

COMPTE: Hi havia un petit error en la funció se.diffMeans en l'estimació quan var.equal =TRUE que us vaig passar, I QUE EN AQUEST PDF ja està arreglat

```
> # Funció que calcula l'estadistic 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
+ }
> # Càlcul de 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)
+     return(sqrt((length(x1)-1)*(sum((x1 - m1)^2)) + (length(x2)-1)*(sum((x2 - m2)^2))) /
+       (length(x1) + length(x2) - 2))
+   } else {
+     return(sqrt(var(x1)/length(x1) + var(x2)/length(x2)))
+   }
+ }
```

La funció *tStat* calcula l'estadístic t. Espera rebre directament dues mostres, *x1* i *x2* i el valor "teòric" de la diferència de mitjanes, *delta*, que es restarà de la diferència de les mitjanes mostrals de les dades *x1* i *x2*. Aquest valor teòric de la diferència de mitjanes correspon a l'argument *mu* de t.test.

Dins cada rèplica de la simulació bootstrap, *x1* i *x2* reben una "remostra" formada agafant a l'atzar i amb reemplaçament 8 elements del vector dels valors "de veritat" dels etruscs i una remostra formada agafant a l'atzar i amb reemplaçament 10 valors del vector dels italians, respectivament. Com a valor teòric de la diferència de mitjanes, *delta* rebrà l'estimació feta sobre les dades reals, $\hat{\delta}$.

```
> deltaEstim = mean(etruscos) - mean(italians)
> seEstim = se.diffMeans(etruscos, italians, var.equal = TRUE)
> set.seed(213)
> alpha = 0.05
> B = 10000
> set.seed(213)  # Per fer reproduïble aquesta simulació...
> t.star = replicate(B,
+   tStat(
+     sample(etruscos, replace = TRUE),
+     sample(italians, replace = TRUE),
+     delta = deltaEstim, var.equal = TRUE
+   )
+ )
> tCrit = quantile(t.star, probs = c(1 - alpha/2, alpha/2))
> # Interval bootstrap-t:
> ic = deltaEstim - tCrit * seEstim
> names(ic) = NULL
> attr(ic, "conf.level") = 1 - alpha
> ic

[1]  2.422278 23.844165
attr(,"conf.level")
[1] 0.95

> # COMPTE aqui els resultats són diferents ja que NO donem per segura la condició d'igualta
> seEstim = se.diffMeans(etruscos, italians)
```

```
> alpha = 0.05
> B = 10000
> set.seed(213)  # Per fer reproduïble aquesta simulació...
> t.star = replicate(B,
+   tStat(
+     sample(etruscos, replace = TRUE),
+     sample(italians, replace = TRUE),
+     delta = deltaEstim
+   )
+ )
> tCrit = quantile(t.star, probs = c(1 - alpha/2, alpha/2))
> # Interval bootstrap-t:
> ic = deltaEstim - tCrit * seEstim
> names(ic) = NULL
> attr(ic, "conf.level") = 1 - alpha
> ic

[1]  5.430498 19.961850
attr(,"conf.level")
[1] 0.95

>
```

1.2.3 Interval de confiança bootstrap-t simetritzat

En lloc de considerar "cues simètriques" (totes dues amb la mateixa probabilitat), podem considerar la possibilitat de simetria respecte de zero: a la distribució de t buscarem una constant $t_{1-\alpha} > 0$ tal que entre $-t_{1-\alpha}$ i $+t_{1-\alpha}$ hi quedi una probabilitat $1-\alpha$, encara que si aquesta distribució no és simètrica a les cues hi quedaran probabilitats desiguals.

Per un nivell de confiança $1-\alpha$, si $t_{1-\alpha}$ és un valor tal que

$$Pr\{|t| \leq t_{1-\alpha}\} = 1-\alpha,$$

l'interval de confiança simetritzat es defineix com $[\hat{\delta} - t_{1-\alpha}\hat{se}_{\hat{\delta}}, \hat{\delta} + t_{1-\alpha}\hat{se}_{\hat{\delta}}]$.

A la pràctica estimarem $t_{1-\alpha}$ mitjançant remostratge bootstrap:

```
> # (COMPTE "a partir" resultats obtinguts a partir dels valors t bootstrap PRESSUPOSANT igu
> deltaEstim = mean(etruscos) - mean(italians)
> seEstim = se.diffMeans(etruscos, italians, var.equal = TRUE)
> set.seed(213)
> alpha = 0.05
> B = 10000
> set.seed(213)  # Per fer reproduïble aquesta simulació...
> t.star = replicate(B,
+   tStat(
+     sample(etruscos, replace = TRUE),
```

```
+     sample(italians, replace = TRUE),
+     delta = deltaEstim, var.equal = TRUE
+   )
+ )
> t1_alpha = quantile(abs(t.star), probs = 1 - alpha)
> # Interval bootstrap-t simetritzat:
> ic = deltaEstim - c(t1_alpha, -t1_alpha) * seEstim
> names(ic) = NULL
> attr(ic, "conf.level") = 1 - alpha
> ic
```

```
[1]  1.097689 22.602311
attr(,"conf.level")
[1] 0.95
```

Aquest interval no inclou el valor 0. **A la DARRERA SECCIÓ calcularem el p-valor bootstrap del test d'hipòtesis d'estudi** i completarem la conclusió.

1.2.4 Interval de confiança bootstrap-p

```
> # (aprofitem els valors t bootstrap obtinguts abans,
> # sense presuposar igualtat de variàncies)
>
> # Interval bootstrap-p:
> icBoot.perc = quantile(t.star, probs = c(alpha/2, 1 - alpha/2))
> names(icBoot.perc) = NULL
> attr(icBoot.perc, "conf.level") = 1 - alpha
> icBoot.perc
```

```
[1] -2.485262  1.953480
attr(,"conf.level")
[1] 0.95
```

1.2.5 Contrast d'hipòtesis bootstrap. Determinació del p-valor

Finalment, els valors bootstrap t^* obtinguts abans també ens podrien servir per obtenir el p-valor associat al contrast on comparem una hipòtesi nul·la d'igualtat de mitjanes vs una hipòtesi bilateral de diferència. Tal com l'hem obtinguda, fixem-nos que la mostra de valors t^* reproduïa la distribució de l'estadístic t centrat, propi de la situació representada per la hipòtesi nul·la: en calcular t^* sobre cada remostra, a la diferència de mitjanes sobre la remostra li restàvem la diferència de mitjanes de la mostra original. Per tant, aprofitant el resultat de la darrera simulació bootstrap, sense presuposar igualtat de variàncies, l'estimació bootstrap del p-valor de la prova bilateral serà:

```
> t.obs = t.test(etruscos, italians, var.equal=TRUE)$statistic
> t.obs
```

```
       t
3.657643
```

```
> sum(abs(t.star) >= abs(t.obs)) / B
```

```
[1] 0.0043
```

Per tant podem **CONCLOURE que HI HAN diferencies estadísticament significatives** pel que fà al valor mig de l'amplada màxima (mesurat en en mil.límetres) de cranis de barons estruscos i italians.

1.3 Comentari final

Atès que la variable de resposta estudiada és contínua, tots els mètodes que hem utilitzat tenen sentit.

2 Exercici 2. OPCIONAL

2.1 Enunciat

(Extret de la plana web del prof. A. Pitarque, Univ. València.) Comparamos dos muestras aleatorias de 10 hombres y de 10 mujeres de edades comprendidas entre los 18 y los 22 años en un ítem que mide su autoestima (escala ordinal de 0 a 10 puntos). Los datos fueron:

HOMBRES	8	7	6	8	7	5	6	4	9	9
MUJERES	8	6	5	6	5	4	4	4	6	4

```
> # Les dades en dos vectors:
> homes = c(8, 7, 6, 8, 7, 5, 6, 4, 9, 9)
> dones = c(8, 6, 5, 6, 5, 4, 4, 4, 6, 4)
> autoestima = c(homes, dones)
> n1 = length(homes)
> n2 = length(dones)
> n = c(n1, n2)
> N = sum(n)
```

a) ¿Podemos afirmar que ambas muestras difieren significativamente en autoestima? b) ¿Podemos afirmar que la autoestima de los hombres es significativamente mayor que la de las mujeres? Tant la pregunta a) com la b) has de respondre-la utilitzant els mètodes permutacions exacte per la diferència de mitjanes i IC bootstrap, i comentar si els resultats de les diverses proves són coherents entre ells.

2.2 Solució

Clarament la variable de resposta no és contínua. Ens queda el dubte de si es tracta d'una variable quantitativa però discreta (que inclouria el cas purament "ordinal") o una escala únicament ordinal. En aquest darrer cas les diferències entre valors no tindrien sentit, per exemple no estaria clar si entre un valor de 8 i un de 5 hi ha la mateixa diferència que entre un valor de 6 i un de 3. Això faria que molts dels estadístics habituals no fossin aplicables.

2.2.1 Interval de confiança bootstrap-t simetritzat

Novament, aquest planteig tindria sentit per una variable discreta.

Definirem en primer lloc algunes funcions necessàries pels càlculs que hem de fer:

```
> # Funció que calcula l'estadistic 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
+ }
> # Càlcul de 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)
+     return(sqrt((length(x1)-1)*(sum((x1 - m1)^2)) + (length(x2)-1)*(sum((x2 - m2)^2))) /
+       (length(x1) + length(x2) - 2))
+   } else {
+     return(sqrt(var(x1)/length(x1) + var(x2)/length(x2)))
+   }
+ }
```

Sigui $\hat{\delta} = \bar{X}_1 - \bar{X}_2$ la diferència entre les mitjanes mostrals d'homes i dones, i $\delta = \mu_1 - \mu_2$ la corresponent diferència de mitjanes poblacionals. L'estadístic t es defineix com:

$$t = \frac{\hat{\delta} - \delta}{\hat{se}_{\hat{\delta}}}$$

Per un nivell de confiança $1 - \alpha$, si $t_{1-\alpha}$ és un valor tal que

$$Pr\{|t| \leq t_{1-\alpha}\} = 1 - \alpha,$$

l'interval de confiança simetritzat es defineix com $[\hat{\delta} - t_{1-\alpha}\hat{se}_{\hat{\delta}}, \hat{\delta} + t_{1-\alpha}\hat{se}_{\hat{\delta}}]$.

A la pràctica estimarem $t_{1-\alpha}$ mitjançant remostratge bootstrap:

```
> deltaEstim = mean(homes) - mean(dones)
> seEstim = se.diffMeans(homes, dones)
```

```
> alpha = 0.05
> B = 10000
> t.star = replicate(B,
+   tStat(
+     sample(homes, replace = TRUE),
+     sample(dones, replace = TRUE),
+     delta = deltaEstim
+   )
+ )
> t1_alpha = quantile(abs(t.star), probs = 1 - alpha)
> # Interval bootstrap-t simetritzat:
> ic = deltaEstim - c(t1_alpha, -t1_alpha) * seEstim
> names(ic) = NULL
> attr(ic, "conf.level") = 1 - alpha
> ic

[1] 0.2464227 3.1535773
attr(,"conf.level")
[1] 0.95
```

Aquest interval no inclou el valor 0, per tant conduiria a la conclusió de desigualtat de mitjanes poblacionals.