

# Pràctica 1: Tests de permutacions exactes

Ferran Garcia

2024-02-23

## Contents

<b>Exercici 1: Antropologia</b>	<b>2</b>
Pas 1: Plantejament d'hipòtesis . . . . .	2
Pas 2: Què i com permutem? . . . . .	2
Pas 3: Escollim un estadístic . . . . .	2
Pas 4: Càlcul del p-valor . . . . .	3
Pas 5: Conclusios . . . . .	4
<b>Exercici 2: Psicologia (cal canviar el text i raonament)</b>	<b>5</b>
Pas 1: Plantejament d'hipòtesis . . . . .	5
Pas 2: Què i com permutem? . . . . .	5
Pas 3: Escollim un estadístic . . . . .	5
Pas 4: Càlcul del p-valor . . . . .	6
Pas 5: Conclusios . . . . .	7

## Exercici 1: Antropologia

Volem contrastar si els etruscs eren originaris de la península italiana a través de les seves restes. Per tal de fer-ho compararem les amplades màximes del crani de restes d'etruscs amb la d'italians. Disposem de les dades següents:

```
etrusc <- c(141, 132, 154, 142, 141, 150, 134, 140)
italia <- c(133, 138, 136, 125, 135, 130, 127, 131, 116, 128)
cranis <- c(etrusc, italia)
```

### Pas 1: Plantejament d'hipòtesis

De partida, suposem que no hi ha diferència entre les mesures etrusques i les italianes. És a dir: Els etruscs provenen de la península italiana.

La nostra hipòtesi nul·la és:  $H_0 : F_e = F_i$  La nostra hipòtesi alternativa és:  $H_1 : F_e \neq F_i$

On  $F_e$  denota la distribució de les amplades màximes dels cranis etruscs i  $F_i$  denota la distribució de les amplades màximes dels cranis italians.

### Pas 2: Què i com permutem?

Permutarem les amplades de la mostra i calcularem mesures de discrepància per contrastar la nostra hipòtesi. Ens interessa trobar diferències entre els grups *etrusc* i *italià*. Per tant només les permutacions que intercanviïn mesures entre grups són informatives. És a dir, el nombre de permutacions informatives són:  $C_{n+m}^n = C_{18}^8 = \frac{18!}{8!10!} = 43758$ , on  $n$  és la mida mostral etrusca i  $m$  és la mida mostral italiana.

```
n <- 8
m <- 10
nperm <- choose(n + m, m); nperm
```

```
## [1] 43758
```

```
perm <- combn(1:(n+m), m) |>
  as.data.frame()
```

### Pas 3: Escollim un estadístic

En aquest apartat justifiquem l'elecció d'estadístics que ens ajudin a identificar diferències entre els dos grups: El de les 8 primeres observacions i el de les 10 últimes observacions.

#### T-Statistic

$$T = \frac{\bar{E} - \bar{I}}{\sqrt{\frac{S\check{s}_e}{n} + \frac{S\check{s}_i}{m}}}$$

On  $\bar{E}$  és la mitjana de les dades etrusques (les 8 primeres) i  $\bar{I}$  és la mitjana de les dades italianes (les últimes 10). Les variàncies mostrals de les dades etrusques i italianes són  $S\check{s}_e$  i  $S\check{s}_i$ , respectivament.

```
TStat = function(vec, ind){
  # estadístic t
  r <- t.test(vec[ind], vec[-ind],
              var.equal = TRUE)[["statistic"]]
  return(r)
}
```

## Pas 4: Càlcul del p-valor

Com que el nombre de permutacions informatives és relativament baix ( $4.3758 \times 10^4$ ) utilitzarem un test exacte. Per tant calcularem l'estadístic escollit per totes les permutacions.

```
perm_t <- vapply(
  X = perm,
  FUN = TStat,
  vec = cranis,
  FUN.VALUE = .0
)
```

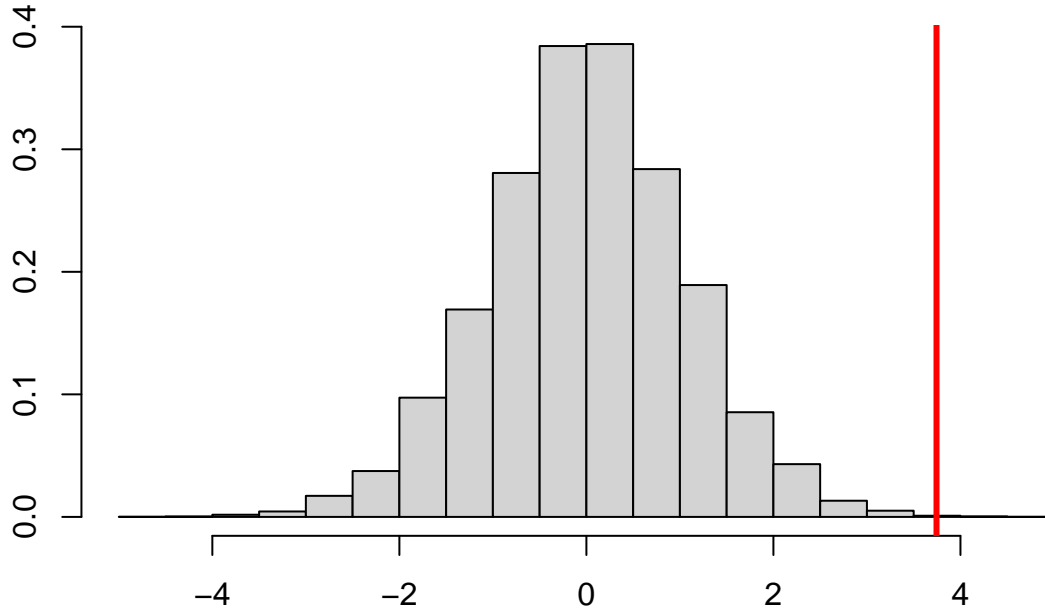
També calculem l'estadístic corresponent a la mostra real i el visualitzem en un histograma sobre la distribució permutada:

```
real_t <- TStat(vec = cranis,
               ind = 1:m); real_t
```

```
##          t
## 3.744035
```

```
hist(perm_t,
     probability = TRUE,
     breaks = 30,
     main = "Distribució permutada i observació real",
     xlab = NULL, ylab = NULL)
abline(v = real_t, lwd = 3, col = "red")
```

## Distribució permutada i observació real



Notem que l'estadístic calculat a partir de la mostra real és bastant extrem en la distribució de dades permutades que sosté l'hipòtesi nul·la. Ara calculem el *p-valor* com la proporció de mostres permutades més extremes que la real 3.7440347:

```
perm_extr <- sum(abs(perm_t) > real_t)
pval <- perm_extr/nperm
round(pval, digits = 4)
```

```
## [1] 9e-04
```

No cal fer els càlculs per adonar-se que si enlloc de fer un contrast bilateral el fèssim a una cua el *p-valor* seria encara més petit sempre i quant l'hipòtesi alternativa fos  $H_1 : F_e > F_i$

### Pas 5: Conclusios

Hem fet un test de permutacions exacte utilitzant l'estadístic T. Podem concloure que els etruscs tenen una amplada de crani més gran que els italians. Per tant és difícil sostenir que els etruscs provenen de la península italiana basant-se en la similitud dels seus cranis.

## Exercici 2: Psicologia (cal canviar el text i raonament)

Volem contrastar si les dones i els homes tenen nivells d'autoestima diferents. Per tal de fer-ho compararem les respostes a una enquesta sobre l'autoestima mesurada amb una escala de Likert. Suposem que la diferència entre dos nivells equival a la diferència numèrica entre els mateixos.

```
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)
```

### Pas 1: Plantejament d'hipòtesis

De partida, suposem que no hi ha diferència entre les mesures dels homes i de les dones. És a dir: Els nivells d'autoestima són els mateixos en homes i en dones.

La nostra hipòtesi nul·la és:  $H_h : F_e = F_d$  La nostra hipòtesi alternativa és:  $H_1 : F_h \neq F_d$

On  $F_h$  denota la distribució de les respostes sobre l'autoestima dels homes i  $F_d$  denota la distribució de les respostes sobre l'autoestima de les dones.

### Pas 2: Què i com permutem?

Permutarem les escales de Likert de la mostra i calcularem mesures de discrepància per contrastar la nostra hipòtesi. Ens interessa trobar diferències entre els grups *dones* i *homes*. Per tant només les permutacions que intercanviïn mesures entre grups són informatives. És a dir, el nombre de permutacions informatives son:  $C_{n+m}^n = C_{20}^{10} = \frac{20!}{10!10!} = 184756$ , on  $n$  és la mida mostral d'homes i  $m$  és la mida mostral de dones.

```
n <- 10
m <- 10
nperm <- choose(n + m, m); nperm
```

```
## [1] 184756
```

```
perm <- combn(1:(n+m), m) |>
  as.data.frame()
```

### Pas 3: Escollim un estadístic

En aquest apartat justifiquem l'elecció d'estadístics que ens ajudin a identificar diferències entre els dos grups: El de les 10 primeres observacions i el de les 10 últimes observacions.

#### T-Statistic

$$T = \frac{\bar{E} - \bar{I}}{\sqrt{\frac{S\check{s}_e}{n} + \frac{S\check{s}_i}{m}}}$$

On  $\bar{E}$  és la mitjana de les dades etrusques (les 8 primeres) i  $\bar{I}$  és la mitjana de les dades italianes (les últimes 10). Les variàncies mostrals de les dades etrusques i italianes son  $S\check{s}_e$  i  $S\check{s}_i$ , respectivament.

```
TStat = function(vec, ind){
  # estadístic t
  r <- t.test(vec[ind], vec[-ind],
              var.equal = TRUE)[["statistic"]]
  return(r)
}
```

## Pas 4: Càlcul del p-valor

Com que el nombre de permutacions informatives és relativament baix ( $1.84756 \times 10^5$ ) utilitzarem un test exacte. Per tant calcularem l'estadístic escollit per totes les permutacions.

```
perm_t <- vapply(
  X = perm,
  FUN = TStat,
  vec = autoestima,
  FUN.VALUE = .0
)
```

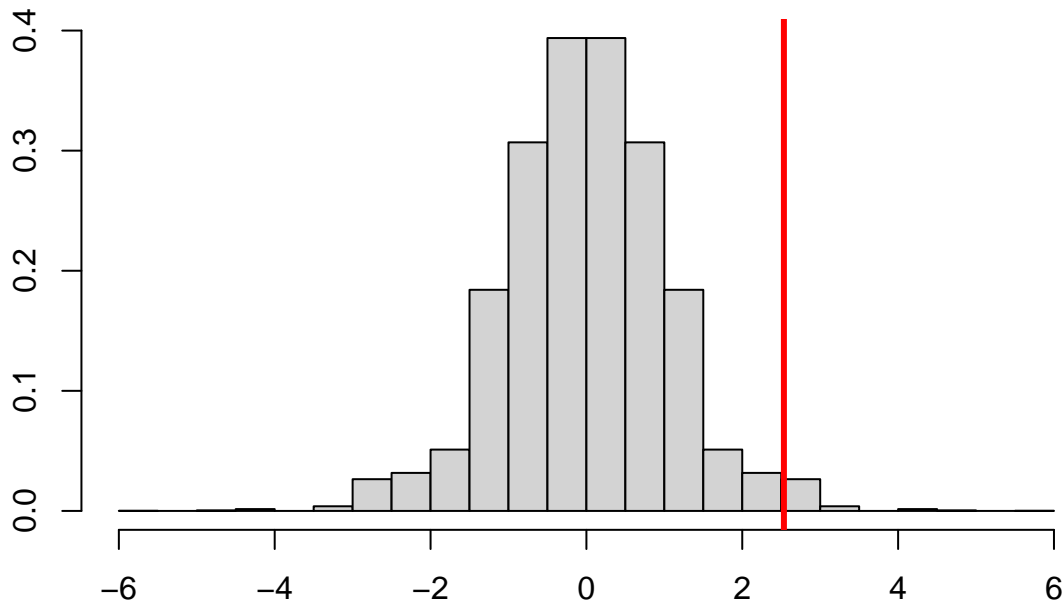
També calculem l'estadístic corresponent a la mostra real i el visualitzem en un histograma sobre la distribució permutada:

```
real_t <- TStat(vec = autoestima,
               ind = 1:m); real_t
```

```
##          t
## 2.53421
```

```
hist(perm_t,
     probability = TRUE,
     breaks = 30,
     main = "Distribució permutada i observació real",
     xlab = NULL, ylab = NULL)
abline(v = real_t, lwd = 3, col = "red")
```

## Distribució permutada i observació real



Notem que l'estadístic calculat a partir de la mostra real és bastant extrem en la distribució de dades permutades que sosté l'hipòtesi nul·la. Ara calculem el *p-valor* com la proporció de mostres permutades més extremes que la real 2.5342104:

```
perm_extr <- sum(abs(perm_t) > real_t)
pval <- perm_extr/nperm
round(pval, digits = 4)
```

```
## [1] 0.0149
```

En aquest cas no podem afirmar amb un nivell de confiança del 99% que la distribució de l'autoestima entre homes i dones és diferent. No obstant, amb menor confiança sí que podem afirmar que existeixen diferències.

### Pas 5: Conclusios

Hem fet un test de permutacions exacte utilitzant l'estadístic T. No podem concloure que existeixi cap diferència entre l'autoestima en funció del gènere.