Universitat Autònoma de Barcelona
Facultat de Ciències
Estadística
M. Saucedo

Seminari 4: Contrasts d'hipòtesis

En aquesta pràctica veurem com resoldre tests d'hipòtesis sobre la mitjana, la variància i la proporció d'una mostra i sobre la comparació de mitjanes i de variàncies de dues mostres.

L'`R` té incorporades instruccions per resoldre tests d'hipòtesis analitzant una mostra donada, donades certes condicions. En altres casos, haurem d'aplicar altres tècniques.

En qualsevol cas, per resoldre un test d'hipòtesis comparem el nivell de significació α amb el p-valor i apliquem la següent regla:

Si p-valor$\leq \alpha$ decidim a favor de H_1. En cas contrari, suposem H_0 certa.

Tots els fitxers de dades que utilitzarem els podeu trobar a

`https://mat.uab.cat/~mbarcelona/est/`

que són fitxers `.RData` que es carreguen amb la funció `load()`.

1. Contrasts sobre la mitjana μ

Suposem en tots els casos que tenim una mostra normal.

1.1. Amb σ coneguda

Considerem les dades del fitxer `dades1.RData` que conté $n = 20$ nombres. Les dades provenen d'una distribució normal amb desviació estàndard $\sigma = 0{,}9$. Posem les dades en un vector que anomenen `x`.

Contrast unilateral: Volem fer el contrast de les hipòtesis

$$\begin{cases} H_0 : \mu = 3{,}5 \\ H_1 : \mu > 3{,}5 \end{cases}$$

amb nivell de significació $\alpha = 0{,}05$.

Pel que hem vist a teoria,

$$Z = \frac{\overline{X} - 3{,}5}{0{,}9}\sqrt{20} \sim N(0,1).$$

Per al vector `x` calculem la mitjana empírica i la mitjana empírica estandarditzada:

```
xm <- mean(x)
z <- (xm-3.5)/0.9*sqrt(20)
```

Obtenim $z = 1{,}328225$. És un valor raonable? Per poder contestar aquesta pregunta calculem el p-valor $P(Z > z)$:

```
pvalor <- 1 - pnorm(z)
```

Resulta p-valor= 0,092052, que representa la probabilitat que en una mostra de mida $n = 20$ d'una $N(\mu = 3{,}5, \sigma = 0{,}9)$ obtinguem un valor de la mitjana mostral $\overline{x}$ igual o superior al valor obtingut a partir de la mostra `x`. Per tant, si aquesta probabilitat és gran suposem que H_0 és certa i si és petita considerem H_1 certa. Utilitzant la regla de decisió introduïda a principi de la pràctica, observem que p-valor$> \alpha$, per tant, considerem H_0 certa.

Exercici 1. Per $\alpha = 0{,}1$ rebutjem H_0?

Contrast bilateral: Per fer el contrast

$$\begin{cases} H_0 : \mu = 3{,}5 \\ H_1 : \mu \neq 3{,}5 \end{cases}$$

amb nivell de significació $\alpha = 0{,}05$ seguim el mateix procediment que abans, però ara el p-valor serà $P(|Z| > z)$ que es calcula

```
p_valor <- 2*(1-pnorm(abs(z)))
```

i dóna 0.184104. Seguint els mateixos passos que abans comparem el p-valor amb α. En aquest cas, per $\alpha = 0{,}05$ o $\alpha = 0{,}1$ considerem H_0 certa.

1.2. Amb σ desconeguda

En aquest cas l'única informació que tenim sobre la població ens ve donada per la mostra. Com anticipat a principi de la pràctica, l'`R` té instruccions per poder treballar directament amb aquesta. Emprenem la funció `t.test` amb les següents opcions:

`x`	Vector que conté les dades.
`mu`	Valor numèric que indica la mitjana en la hipòtesi nul·la.
`alternative`	`"two.sided"` (per defecte) per tests bilaterals.
	`"less"` per tests unilaterals esquerre.
	`"greater"` per tests unilaterals drets.

Observació: La funció `t.test` *té moltes més opcions disponibles. Les podeu trobar amb la comanda* `?t.test`.

Exemple 1. Considerem les dades del fitxer `dades2.RData`. Suposem que les dades provenen d'una distribució normal amb σ desconeguda. Volem comprovar si podem afirmar que la mitjana és inferior a 12 amb un 5 % de significació.

Plantegem el contrast d'hipòtesis:

$$\begin{cases} H_0 : \mu = 12 \\ H_1 : \mu < 12 \end{cases}$$

És un test unilateral esquerre. Executem en `R` la següent comanda:

```
t.test(dades2, mu = 12, alternative = "less")
```

Aquesta instrucció dóna el següent output:

```
        One Sample t-test

data:  dades2
t = -1.1984, df = 24, p-value = 0.1212
```

```
alternative hypothesis: true mean is less than 12
95 percent confidence interval:
   -Inf 12.09879
sample estimates:
mean of x
11.76902
```

Interpretem-ho:

`One Sample t-test`	Tipus de test.
`data`	Vector de dades que s'està analitzant.
`t`	Valor de l'estadístic de contrast amb les dades de la mostra.
`dt`	Graus de llibertat de la t-Student.
`p-valor`	p-valor associat al test.
`alternative hypothesis`	Hipòtesis alternativa (útil per saber si hem fet el test que volíem).
`confidence interval`	Interval de confiança associat al test.
`sample estimates`	Mitjana mostral.

Observació: En aquest cas, l'interval de confiança és 'a la dreta", concordement al contrast. Per calcular l'interval de confiança de la mitjana hem de posar l'opció `alternative = "two.sided"`, *que ve per defecte si no s'explicita aquest argument. Per canviar el nivell de confiança, hem d'afegir l'opció* `conf.level` *a la funció* `t.test`. *(Això s'ha vist a la Pràctica 4).*

En el nostre cas, el p-valor associat al test és 0.1212, que és superior a α. Per tant, acceptem H_0. Amb 5 % de significació, no tenim prou evidències estadístiques per afirmar que la mitjana és inferior a 12.

Exercici 2. Feu uns tests per comprovar si

a) La mitjana és superior a 11.

b) La mitjana és igual a 12.

Treballeu amb un nivell de significació del 10 %.

2. Contrasts sobre una proporció

Per fer un contrast sobre la proporció, s'utilitza la funció `prop.test` amb els següents arguments:

`N`	Nombre d'observacions en les quals s'ha produït un determinat esdeveniment.
`n`	Nombre total d'observacions.
`p`	Proporció que es contrasta a la hipòtesis nul·la.
`alternative`	`"two.sided"` (per defecte) per tests bilaterals.
	`"less"` per tests unilaterals esquerre.
	`"greater"` per tests unilaterals drets.
`conf.level`	Nivell de confiança.
`correct`	Indica si s'aplica la correcció de Yates `(TRUE/FALSE)`.

Observació: Als resultats que hem vist a teoria no hem aplicat la correcció de Yates. Per tant, si volem que els resultats obtinguts amb `R` *i els teòrics coincideixin hem de posar l'argument* `correct = FALSE` *(la funció té* `TRUE` *per defecte).*

L'output de la funció `prop.test` és similar al de la funció `t.test`.

Exemple 2. En una mostra de $n = 32$ observacions, s'ha observat que en $N = 14$ d'elles s'ha produït un cert esdeveniment A. Amb nivell de significació $\alpha = 0{,}1$, podem afirmar que el nombre d'observacions en que es produeix l'esdeveniment és inferior a la meitat?

Realitzem el següent contrast sobre la probabilitat $p = P(A)$:

$$\begin{cases} H_0 : p = 0{,}5 \\ H_1 : p < 0{,}5 \end{cases}$$

Observem que en el nostre cas $p_0 = 0{,}5$. Amb `R`:

```
N <- 14
n <- 32
prop.test(N, n, alternative = "less", correct = FALSE)
```

Observació: La funció `prop.test` *agafa per defecte* `p=0.5`*, per tant, en aquests cas podem ometre'l.*

El p-valor és 0.2398. És superior al nivell de significació, per tant, acceptem la hipòtesis nul·la. Amb 1 % de significació, no hi ha evidències estadístiques per afirmar que la proporció d'observacions en les quals s'ha produït l'esdeveniment A és inferior al 50 %.

Exercici 3. A una empresa de residus químics, s'ha agafat una mostra de 300 empleats i s'ha observat que 135 d'ells presenten evidència d'excés de radiació. Amb 5 % de significació, podem concloure que la proporció de treballadors amb excés de radiació és superior al 40 %? Canvia la resposta si treballem amb 1 % de significació? Proveu a calcular el p-valor utilitzant la funció de llibreria que acabem de veure amb l'opció `correct=FALSE` i utilitzant la funció `pnorm` adequadament per veure que obteniu els mateixos resultats.

3. Contrasts sobre la variància σ^2

Per fer un contrast sobre la variància, s'utilitza la funció `varTest`, que es troba al package `EnvStats`[1]. Les seves opcions són:

`x`	Vector que conté les dades.
`alternative`	`"two.sided"` (per defecte) per tests bilaterals.
	`"less"` per tests unilaterals esquerres.
	`"greater"` per tests unilaterals drets.
`conf.level`	Nivell de confiança.
`sigma.squared`	Valor numèric que indica la variància en la hipòtesi nul·la.

L'output d'aquesta funció és similar al de la funció `t.test`.

Exercici 4. Considerem les dades del fitxer `dades2.RData`. Amb 10 % de significació, podem afirmar que la variància és 1?

4. Contrasts de dues mostres normals independents

Suposem en tots els casos que tenim mostres X i Y normals.

[1]Segurament l'haureu d'instal·lar.

4.1. Contrasts sobre comparació de variàncies

Per comparar les variàncies de dues poblacions utilitzem la funció `var.test` amb els següents arguments:

`x`	Vector que conté les dades de la primera mostra.
`y`	Vector que conté les dades de la segona mostra.
`ratio`	Valor numèric que indica el quocient entre les variàncies en la hipòtesi nul·la.
`alternative`	`"two.sided"` (per defecte) per tests bilaterals. `"less"` per tests unilaterals esquerre. `"greater"` per tests unilaterals drets.
`conf.level`	Nivell de confiança.

L'output d'aquesta funció és similar al de la funció `t.test`.

Observació: Recordeu que el test de comparació de variàncies estudia la relació $\frac{\sigma_X^2}{\sigma_Y^2}$ *(no la seva diferència!). En la majoria dels casos, fem contrastos amb hipòtesi nul·la* $\sigma_X^2 = \sigma_Y^2$*, és a dir, comparem el quocient de variàncies amb 1, que és el valor per defecte de l'argument* `ratio` *de la funció* `var.test`*.*

Exemple 3. Es varen provar dos tractaments antireumàtics, administrant-los a 10 i 15 pacients, respectivament. Els resultats després del tractament (a més puntuació, més eficàcia), varen ser:

Tractament A:	15	61	21	17	40	42	10	23	35	28					
Tractament B:	24	37	42	25	16	54	65	40	58	35	18	56	69	32	44.

Amb nivell de significació $\alpha = 0{,}1$, hi ha diferències significatives entre les variàncies dels dos tractaments?

Plantegem el test d'hipòtesis

$$\begin{cases} H_0 : \sigma_A^2 = \sigma_B^2 \\ H_1 : \sigma_A^2 \neq \sigma_B^2 \end{cases}$$

Després de definir els vectors amb les dades, executem la comanda

```
var.test(A, B)
```

El p-valor és 0.8446, que és superior a α, per tant, acceptem la hipòtesis nul·la. Amb 10 % de significació, no tenim evidències estadístiques per afirmar que hi ha diferències significatives entre les variàncies.

4.2. Contrasts sobre comparació de mitjanes

Per comparar les mitjanes de dues poblacions utilitzem la funció `t.test`. En aquest cas, hem d'especificar més arguments dels que hem utilitzat per fer contrastos sobre la mitjana d'una sola mostra:

`x`	Vector que conté les dades de la primera mostra.
`y`	Vector que conté les dades de la segona mostra.
`alternative`	`"two.sided"` (per defecte) per tests bilaterals. `"less"` per tests unilaterals esquerre. `"greater"` per tests unilaterals drets.
`mu`	Valor numèric que indica la diferència entre les mitjanes en la hipòtesi nul·la.
`paired`	Indica si les mostres són aparellades o independents (`TRUE/FALSE`).
`var.equal`	Indica si es consideren variàncies iguals o no (`TRUE/FALSE`).
`conf.level`	Nivell de confiança.

Els arguments que la funció agafa per defecte són

```
alternative = "two.sided"
mu = 0
paired = FALSE
var.equal = FALSE
conf.level = 0.95
```

Exemple 4. Considerem les dades de l'Exemple 3. Amb nivell de significació $\alpha = 0{,}1$, hi ha diferències significatives entre els dos tractaments?

Estem tractant amb mostres independents. Plantegem el test d'hipòtesis

$$\begin{cases} H_0 : \mu_A = \mu_B \\ H_1 : \mu_A \neq \mu_B \end{cases}$$

A l'Exemple 3 hem vist que no tenim evidències estadístiques per suposar les variàncies diferents. Executem la comanda

```
t.test(A, B, var.equal = TRUE)
```

El p-valor és 0.08763, que és inferior a α, per tant, acceptem la hipòtesis alternativa. Amb 10 % de significació, podem afirmar que hi diferència entre els dos tractaments.

Problemes

1. Considereu el fitxer `kidsfeet.RData`
   Aquest document conté les dades de 39 nens i nenes de 9 i 10 anys. Les variables són:

   Mes de naixement.

   Any de naixement (1987 o 1988).

   Longitud del peu més llarg (en cm).

   Amplada del peu més llarg (en cm), mesurat a la part més ampla del peu.

   Sexe (nois, B, o noies, G).

   Peu que s'ha mesurat (el dret, R, o l'esquerre, L).

   Lateralitat (dretana, R, o esquerrana, L).

   Suposeu normalitat quan calgui.

   a) Podeu afirmar amb un 5 % de significació que la mitjana de longitud del peu més llarg supera els 8.8 cm? Canvia la resposta si agafeu un 1 % de significació?
   b) Podeu afirmar amb un 5 % de significació que la proporció d'esquerrans és inferior al 30 %? Canvia la resposta si agafeu un 10 % de significació?
   c) Podeu afirmar amb un 5 % de significació que la desviació típica de la amplada és inferior a 0.56 cm? Canvia la resposta si agafeu un 3 % de significació?
   d) Podeu afirmar que hi ha diferències significatives entre les variàncies de la longitud i de l'amplada? Treballeu amb un nivell de significació del 5 %.
   e) Podeu afirmar que hi ha diferències significatives per a la longitud mitjana entre esquerrans i dretans? Treballeu amb un nivell de significació del 3 %.

f) Canvia la resposta de l'apartat anterior si suposem que que les variàncies són iguals?

g) Feu un test amb nivell de significació del 3 % per comprovar si hi ha diferències significatives entre les variàncies de la longitud entre esquerrans i dretans. Mirant el resultat d'aquest test, quina resposta és millor considerar, la de l'apartat e) o la de l'apartat f)?

2. Després d'un tractament contra l'obesitat, els pesos en Kg de vuit persones eren

$$58,\ 50,\ 60,\ 65,\ 64,\ 62,\ 56,\ 57.$$

Suposeu normalitat i $\alpha = 0{,}1$.

a) Podem afirmar que la mitjana teòrica μ és 61, sabent que $\sigma = 3$?

b) Podem afirmar que la mitjana teòrica μ és 61, si σ és desconeguda?

c) Podem afirmar que la variància teòrica σ^2 és superior a 10?

3. En unes eleccions on participen dos candidats, A i B, es va dur a terme un sondeig d'opinió amb 1000 votants seleccionats a l'atzar. D'ells, 615 prefereixen A. Podem afirmar que la proporció de votants que prefereixen A és superior al 60 %?

4. El fitxer `tterreny.RData` conté les dades de 106 cotxes. Les variables són:

Marca	Marca del cotxe (de 1 a 15)
Preu	Preu en euro
Cilindres	Cilindres del motor
Consum90	Consum mitjà a 90 km/h
Consum120	Consum mitjà a 120 km/h
Consum.Urba	Consum mitjà a la ciutat
Velocitat	Velocitat màxima

Suposeu normalitat. Contrasteu les hipòtesis següents amb nivell de significació $\alpha = 0{,}05$:

a) El consum mitjà a 120 Km/h és de 12 litres.

b) La mitjana de la velocitat màxima és de 155 Km/h.

c) La mitjana del consum urbà és inferior de 12.2 litres.

d) A partir de les dades de les variables `Consum90` i `Consum120`, podem acceptar com a vàlida l'afirmació que $\mu_{120} = \mu_{90}$?

e) A partir de les dades de les variables `Consum120` i `ConsumUrba`, podem acceptar com a vàlida l'afirmació que $\mu_{120} = \mu_U$?

f) A partir de les dades de les variables `Consum120` i `ConsumUrba`, podem acceptar com a vàlida l'afirmació que $\mu_{120} = \mu_{90} + 2$?

5. Donada una mostra $X_1, \ldots, X_n \sim F$, si ens trobem en un context paramètric, podem suposar que $F \in \mathcal{F} = \{F_\theta : \theta \in \Theta\}$ una família que depèn d'un nombre finit de paràmetres. En aquest escenari, un test tal que $H_0 : \theta \in \Theta_0$ i $H_1 : \theta \in \Theta \backslash \Theta_0$ té una funció de potència que es defineix com, donat $\theta \in \Theta_1 = \Theta \backslash \Theta_0$, la probabilitat de rebutjar la hipòtesis nul·la quan la vertadera probabilitat és F_θ:

$$\Pi(\theta) = \mathbf{P}(\text{rebutjar } H_0 | \theta).$$

Suposem que tenim $X_1, \ldots, X_{100} \sim N(\mu, \sigma^2)$ amb $\sigma^2 = 1$ conegut i fixat. Dibuixeu la funció de potència dels següents tests amb $\alpha = 0{,}1$:

a) $H_0 : \mu = 0$ contra $H_1 : \mu > 0$.

b) $H_0 : \mu = 0$ contra $H_1 : \mu \neq 0$.