Departament d'Estadística i Investigació Operativa

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Models estadístics i ciència de dades
Annex

Bloc D – Probabilitat i Estadística

2023

Índex

1. Proves de significació o d'hipòtesis (PH)

Una prova d'hipòtesis (PH) parteix d'una afirmació (una **hipòtesi** de partida, o **nul·la, H_0**), i es vol estudiar si les dades (una mostra finita) proporcionen proves en contra seu (una repetició intensa, una mostra infinita que representaria la població, seria definitiva):

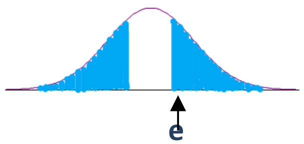
$$\mathbf{H_0}: \boldsymbol{\theta = valor}$$ (la hipòtesi és un possible *valor* del paràmetre)

La mostra es *concentra* en un estadístic (e), que segueix una distribució de probabilitat de referència coneguda (Ref) si s'assumeix certa H_0 (i altres assumpcions, com que la mostra és aleatòria)

L'**estadístic permet obtenir un IC** (veure a bloc C) o **avaluar-lo amb un valor concret del paràmetre** donant lloc a un punt de la distribució de referència. A partir d'aquest punt resultant, es pot calcular la probabilitat d'obtenir valors més extrems. El **p-value** és aquesta probabilitat de, sota H_0, obtenir resultats igual o més *extrems* que l'observat (veure annex de bloc C)

Addicionalment a H_0 afegim **la hipòtesi alternativa (H_1)**, totalment complementària a la nul·la (enfocament bilateral):

$$\mathbf{H_0}: \boldsymbol{\theta = valor}$$
$$\mathbf{H_1}: \boldsymbol{\theta \neq valor}$$



P(|Ref|>e)

Tal com hem vist al bloc C, l'estadístic pot ser de la forma "senyal/soroll" i també "quocient de variàncies"

PH i P-value

- El *P*-value diu amb quina freqüència poden passar esdeveniments com el de la mostra (o més extrems) quan la hipòtesi H_0 és correcta, i poder concloure * :
  - Si el *P-value* és petit $\rightarrow$ tenim evidència en contra de H_0

    Valor a prova fora del IC: la diferència del valor observat i la hipòtesis és *sorprenent*, no explicable per atzar)
  - Si el *P*-value no és petit, **NO** demostra la "veritat" de H_0

    Valor a prova dins del IC: la diferència del valor observat i la hipòtesis és menor que l'esperable per atzar)
- *P-value* **NO** és cap de les següents probabilitats:
  - la probabilitat d'"haver-se equivocat"
  - la probabilitat que la hipòtesi nul·la sigui certa
  - la probabilitat d'haver rebutjat erròniament la hipòtesi nul·la
  - 1-P-value NO és la probabilitat que la hipòtesis alternativa sigui certa
- Podeu trobar més informació i possibles males interpretacions a Wikipedia ("p-value")

Per a qualsevol estadístic (rati senyal/soroll, quocient de variàncies, ...), el valor de l'estadístic sota H_0 i el seu p-value permeten informar de quant inversemblant ("grau de sorpresa") és la hipòtesi només donant per certes totes les premisses. Cal evitar els mal usos i males interpretacions

* El càlcul d'IC i de PH venen d'un mateix estadístic que fa que es puguin relacionar les seves conclusions. En el cas de l'estadístic senyal/soroll, el senyal ve de l'estimació puntual observada i el soroll del "se", i per tant tenint l'estimador i "se" podem calcular l'estadístic sota H_0 o l'IC corresponent

Exemple (del bloc C) amb PH

Exemple de 2 mostres on comparar μ_1 i μ_2 amb l'IC de l'efecte diferencial (μ_1-μ_2)

```
Y1 <- c(1,1,2,2.0,2,2.5,4,5,5.5,6,7.5,8,8,9.5,9,9.5)
Y2 <- c(1.5,1,2,1.0,3,3,3.5,5,6,6,8.5,8.5,9.5,8.5,9.1,9)
```

(per exemple X1 i X2 dues mostres de notes d'uns mateixos estudiants)

```
t.test(Y1,Y2,paired=T)
```

```
t = -0.92936, df = 15, p-value = 0.3674
  alternative hypothesis: true difference in means is not equal to 0
  95 percent confidence interval:
   -0.5351864  0.2101864
  sample estimates: mean of the differences  -0.1625
```

H_0: $\mu_1-\mu_2=\mu_D = 0$
H_1: $\mu_1-\mu_2=\mu_D \neq 0$

L'IC ens diu que la **diferència de mitjanes s'espera entre -0.5 i +0.2**.

El valor 0 (valor a prova) està DINS l'IC. El p-value (36.74%) és més gran que el risc del 5%

Conclusió: no hi ha prou evidència per rebutjar o dubtar de **H_0: $\mu_1-\mu_2=0$**,
per tant no hem trobat evidència per contradir que les mitjanes poblacionals d'aquestes dues notes siguin iguals (és **raonable pensar que no hi ha diferència en les mitjanes**)

Exemple (del bloc C) amb PH

Exemple de 2 mostres on comparar σ_1 i σ_2 amb l'IC de l'efecte diferencial (σ^2_1/σ^2_2)

(com el cas dels exercicis de comparar la variabilitat en la duració dels recanvis dels cartutxos de tinta de dues marques)

```
A <- c(350,361.9,365,365,365,370,372,377)
# mean(A)=365.7375  sd(A)=8.00231 var(A)=64.03696
B <- c(390,391.7,410,412,414,418)
# mean(B)=405.95  sd(B)=12.00396  var(B)=144.095
```

Assumim normalitat (o qqnorm(A i qqline(A))

Assumim normalitat (o qqnorm(B) i qqline(B))

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

```
F test to compare two variances
data:  B and A
F = 2.2502, num df = 5, denom df = 7, p-value = 0.3199
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.4257491 15.4206862
sample estimates:
ratio of variances
          2.250185
```

H_0: $\sigma^2_B / \sigma^2_A = 1$ ($\sigma^2_1 = \sigma^2_2$)

H_1: $\sigma^2_B / \sigma^2_A \neq 1$ ($\sigma^2_1 \neq \sigma^2_2$)

L'IC indica que el rati de variàncies s'espera entre 0.4 i 15.4. Un valor de 1 (valor a prova) està DINS l'IC.

El p-value (31.99%) és més gran que el risc del 5%

Conclusió: no hi ha prou evidència per rebutjar H_0 ($\sigma^2_B / \sigma^2_A = 1$), per tant no hem trobat evidència per contradir que **les variàncies poblacionals siguin iguals**.

2. Tipus d'errors en una prova d'hipòtesi. Tipus I

- En proves d'hipòtesis (PH) expressem les conclusions com "rebutgem H_0" o "no rebutgem H_0", però pot interessar més optar entre dues hipòtesis acotant els riscos.
- Si l'objectiu és prendre una decisió, un criteri simple seria definir a priori un llindar α per sota del qual el P-valor és vist com "petit".
- Ara, si repetim la decisió 'n' vegades, com en un repetit procés de control de qualitat, α ens donarà la freqüència d'errors determinada:

En un 100α % dels casos que rebutgem H_0, aquesta és certa.

- **Error de tipus I.** Quan utilitzem dades mostrals per posar a prova una hipòtesi sobre els paràmetres poblacionals, es pot cometre l'error de actuar com si la hipòtesi fos falsa quan no ho és realment. La probabilitat d'aquest error és podria expressar com:

$$\alpha = P(\text{concloure } H_1 \mid H_0 \text{ certa})$$

[Per procediment, aquesta prob està fixada igual a α]

Tipus d'errors en una prova d'hipòtesi. Tipus II

- La situació complementària també es pot produir.
- **Error de tipus II.** En la mateixa situació, es pot cometre l'error de no trobar evidència en contra de la hipòtesi quan realment és falsa. És a dir, no rebutjar una hipòtesi que no és certa. La probabilitat d'aquest error es pot expressar com:

$$\beta = P(\text{concloure } H_0 \mid H_1 \text{ certa}),$$

- Aquest valor, en general, no és controlable i normalment no es pot saber quant val perquè depèn del valor real del paràmetre testejat (que és desconegut).

Tipus d'error (risc)		**Decisió o Acció**	
		A_0	A_1
Realitat	H_0	**Decisió correcta**	**Error Tipus I (risc α)**
	H_1	**Error Tipus II (risc β)**	**Decisió correcta**

Tipus d'errors en una prova d'hipòtesi. Exemple 1

- Control de qualitat. Un processador ha de funcionar a certa velocitat μ_0 però el sistema de fabricació pot desestabilitzar-se i baixar-la a μ_1. Estudiades les conseqüències, l'equip directiu demana a l'estadístic que dissenyi un estudi al que sotmetre cada nou processador abans de instal·lar-lo i vendre-ho.
- Després de uns quants càlculs, l'estadístic:
  - posa μ_0 a H_0 i μ_1 a H_1
  - fixa $\alpha = 0.05$ i $\beta = 0.10$
  - proposa fer **'n' proves amb cada processador**
  - acceptar-ho si queda per damunt de un cert llindar L i rebutjar en cas contrari
- Quan posem en marxa l'estudi,
  1. Quina proporció de processadors correctes seran rebutjats?
  2. Quina proporció d'incorrectes arribaran al mercat?

Tipus d'errors en una prova d'hipòtesi. Exemple 2

- Ch és un navegador amb fama de ràpid, i la marca dominant MD no vol perdre la seva hegemonia. **Suposem que la velocitat mitjana de Ch per carregar una pàgina patró és 700 u., i la de MD és 600 u**. La desviació típica és 150 u. Fixem α = 0.025 (unilateral)
- Si fem 10 proves <u>independents</u> de càrrega per cada navegador:

Ch	$\bar{y}_A = 680$	$S_A = 89$	$n = 10$
MD	$\bar{y}_B = 597$	$S_B = 147$	$n = 10$

- La igualtat de mitjanes poblacionals no és rebutjada i MD proclama ('testat científicament') **que el seu navegador és tan ràpid com Ch**.
- Però aquesta conclusió no és correcta: no poder rebutjar la hipòtesi nul·la (Ch és igual a MD) no implica demostrar la seva veritat. Com ja hem dit, hi ha un cert risc de que, sent diferents els dos navegadors, no puguem trobar-ne l'evidència.
- Com en aquest cas coneixem la diferència real, anem a calcular el risc 'beta' β.

Tipus d'errors en una prova d'hipòtesi. Exemple 2

- La clau és estudiar la distribució de l'estadístic de referència
- Sota H_0, la diferència de mitjanes mostrals és:

$$\hat{z} = \frac{\bar{y}_A - \bar{y}_B}{\sigma\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}} \sim N(0,1) \rightarrow \bar{y}_A - \bar{y}_B \sim \mathrm{N}\left(0, \sigma\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}\right) = N\left(0{,}150\sqrt{\frac{1}{10}+\frac{1}{10}}\right) = N(0{,}67.08)$$

- Definint α = 2.5% unilateral, la regió crítica es troba per diferències de les mitjanes mostrals més grans que 1.96·67.08 = 131.48 u.
- En realitat, la diferència entre mitjanes és de 100. Com les mostres provenen d'aquesta situació, comprovem com de probable és que NO puguem rebutjar H_0:

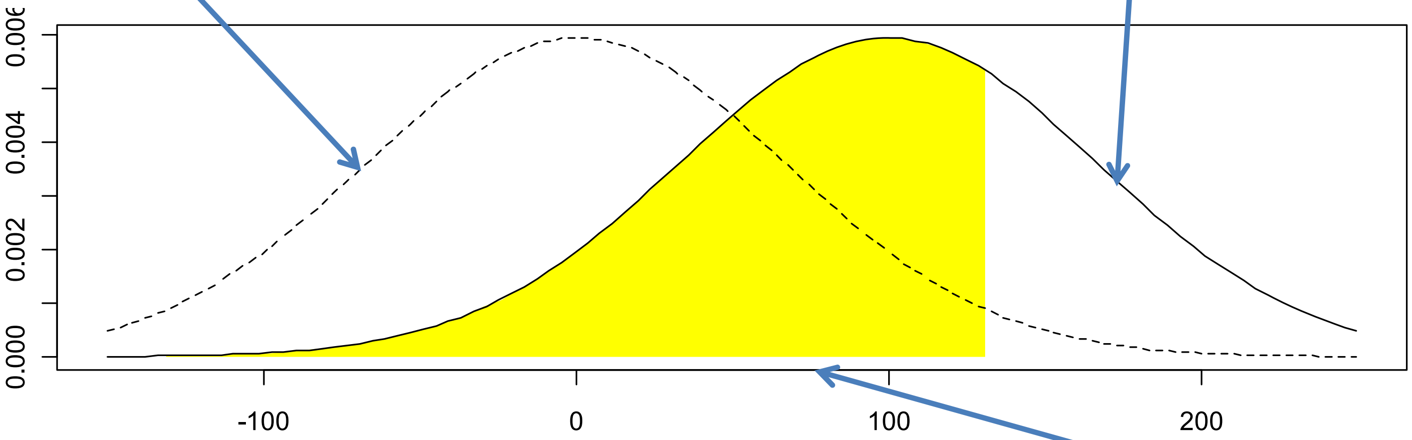
P("Error tipus II") = P($\bar{y}_A - \bar{y}_B$ < 131.48 | H_1) = P(Z < (131.48-100)/67.08) = P(Z<0.47) = 0.68

- Veiem que MD tenia molt fàcil resoldre la prova a la seva conveniència: era molt probable no trobar cap diferència significativa. La prova és **poc potent** (**potència = 1–β**).

Tipus d’errors en una prova d’hipòtesi. Exemple 2

Distribució “ingènua” per a la diferència de mitjanes mostrals (noteu que la campana té esperança zero)

Distribució real per a la diferència de mitjanes mostrals $(\bar{y}_A - \bar{y}_B)$. En l’exemple anterior, els valors típics estaran al voltant de 100, que és la diferència autèntica de les mitjanes poblacionals $(\mu_A - \mu_B)$



Regió d’”*acceptació*”
(no es pot rebutjar la hipòtesi nul·la)

Diferència observada:
680 – 597 = 83 u.

Tipus d'errors en una prova d'hipòtesi. Exemple 2

- En el nostre exemple, podem estirar de les orelles als de MD per:
  - La conclusió ha estat incorrectament exposada
  - El disseny és defectuós: li manca potència (la *n* és petita); o, més aviat, si volien demostrar equivalència, havien d'haver plantejat un altre tipus d'estudi (que no veurem en aquest curs)
- En conseqüència, l'experimentador ha d'assumir que el seu estudi està exposat a diferents perills:
  - Mostres no aleatòries (els individus no són independents entre sí)
  - Assignació de X no aleatòria (individus no similars entre els grups)
  - Variables amagades que pertorben la resposta observada
- I estar disposat a posar mesures per evitar errors com aquests. A més a més, ha de saber que l'anàlisi d'un estudi estadístic no és una demostració matemàtica i, com a mínim, la conclusió ha de ser prudent.

To do or not to do. Arbre de decisions i conseqüències

H_0: Un mètode nou no és efectiu?

- **Sí és efectiu**
  - **Rebutjar H_0** (s'implanta nou)
    - + Beneficis
  - **No rebutjar H_0** (NO s'implanta nou)
    - - Absència de beneficis potencials
    - - Costos de I+D sense retorn
    - - Pèrdua de competitivitat
    - - Frustració per no asolir objectius
- **NO és efectiu**
  - **Rebutjar H_0** (s'implanta nou)
    - - Malestar dels usuaris
    - - Pèrdues a llarg termini
    - - Dilema: invertir més per resoldre el problema, o aturar la inversió
  - **No rebutjar H_0** (NO s'implanta nou)
    - - Decepció inicial
    - + Protecció drets usuaris
    - + Oportunitats per buscar competitivitat per altres vies

Realitat desconeguda | ***Decisió*** | ***Conseqüències***

La construcció d'un arbre com aquest ens ha de portar a determinar les probabilitats dels errors tipus I i II

3. Estadístic F (quocient de quadrats mitjos)

- En models **Y =** $\boldsymbol{\mu + \vartheta_k + \varepsilon}$, és usual plantejar la **prova d'hipòtesis global** de que el factor X (el grup) no explica res de la resposta. S'empra l'estadístic:

$$\frac{variabilitat\ explicada/(k-1)}{variabilitat\ residual/(n-k)} = \frac{QM_E}{QM_R}$$

El numerador quantifica la variació entre grups (*explicada*)
El denominador quantifica la variabilitat interna al grup (*residual, aleatòria*), i és una estimació conjunta de la variància del soroll σ^2.
Quan els grups presenten la mateixa mitjana, el numerador també estima σ^2 de forma independent, i el quocient segueix una distribució F-Fisher (k-1 i n-k graus de llibertat).
Clàssicament, un valor gran de F indica que X sí es un factor explicatiu de Y.

- En models **Y=** $\boldsymbol{\beta_0 + \beta_1 X + \varepsilon}$ (amb 1 o més X) també podem estudiar si les X expliquen la resposta. L'estadístic és similar:

$$\frac{variabilitat\ explicada/(p-1)}{variabilitat\ residual/(n-p)} = \frac{QM_E}{QM_R}$$

on p representa el nombre de paràmetres: $\boldsymbol{\beta_0, \beta_1, \ldots \beta_{p-1}}$ o nombre de predictors X +1
Suposant que els pendents són tots 0 (X's no expliquen res de la resposta), segueix una F-Fisher (p-1 i n-p graus de llibertat).
Clàssicament, un valor gran de F indica que *alguna* X sí es un factor explicatiu de Y.

A la referència de la bibliografia (Estadística per a enginyers informàtics) trobareu més detalls al capítol 6.6. I al final de l'annex del bloc C trobareu la definició i funcions en R de la distribució F.