Testarea ipotezelor statistice pentru un parametru de populatie θ necunoscut

• **Punerea problemei**

Formulele pentru lab 6 sunt in fisierul

`tests.pdf`.

Spre deosebire de metodele de estimatie de pana acum (estimare punctuala sau cu un interval de incredere), acum pornim de la ceva cunoscut, pornim cu o anumita idee predeterminata asupra parametrului θ. Aceasta provine fie din considerente teoretice, fie din experiente anterioare, etc. Deci se fac doua ipoteze, *ipoteza nula*, H_0, care contine ceea ce se stia/credea inainte si *ipoteza alternativa (sau de cercetare, reaserch hypothesis)*, H_1, ceva nou cu care vine cercetatorul (cel ce face experimentul), care, din punct de vedere al problemei este, cumva, opusul ipotezei nule. Ideea este sa "le batem cap la cap" cele doua ipoteze si, daca, pe baza datelor de selectie, se poate *respinge* ipoteza nula, in favoarea alternativei, sau *nu se respinge* (pe undeva, se accepta, sau mai degraba, ramane valabila, fiindca nu poate fi respinsa). Deci raspunsul dupa un test statistic este "resping sau nu resping ipoteza nula", NICIODATA nu spunem ca respingem sau nu ipoteza alternativa, aceea e ceva nou cu care incercam sa inlocuim ce era stiut dinainte, daca putem.

Totdeauna se considera o ipoteza nula *simpla*, adica egalitatea cu o valoare. Iar H_1 poate fi una din trei variante:

$$\begin{array}{rcl} \mathrm{H}_0 & : & \theta = \theta_0, \text{ versus} \\ \mathrm{H}_1 & : & \begin{cases} \theta < \theta_0 \ (\text{left} - \text{tailed test, test la stanga}) \\ \theta > \theta_0 \ (\text{right} - \text{tailed test, test la dreapta}) \\ \theta \neq \theta_0 \ (\text{two} - \text{tailed test, test bilateral}) \end{cases} \end{array}$$

Notiuni importante: Se da de la inceput $\alpha \in (0,1)$, *significance level* (nivelul de semnificatie), care reprezinta probabilitatea de eroare de tipul I, adica eroarea de a respinge H_0, cand aceasta este adevarata. (In general, nu acceptam nivel de semnificatie mai mare de 10%, adica prob. de eroare de tipul I mai mare de 0.1, asa cum nu acceptam nici nivel de incredere mai mic de 90% la intervale de incredere. In general, in statistica, se folosesc 3 nivele de semnificatie, $5\%, 1\%$ sau 0.1%, sigur sunt cazuri cand se pot folosi nivele de semnificatie si mai mici.) Se foloseste *test statistic (statistica de test)*, TS, care va fi aceeasi ca statistica pivot de la intervale de incredere (si cazurile vor fi exact aceleasi), careia i se cunoaste pdf-ul, mai avem *observed value of the test statistic* $TS_0 = TS(\theta = \theta_0)$ (care este un numar, θ fiind singura necunoscuta in expresia lui TS, cand acesteia i-am dat o valoare numerica, totul devine un numar), se determina o *rejection (critical) region* (regiunea critica), RR, care contine valorile statisticii de test pentru care se respinge H_0 (cum spune si numele) si se calculeaza $P-$ *value* (valoarea P, sau valoarea critica a testului). $P-$ ul, ca definitie, este probabilitatea ca statistica de test sa ia valori cel putin la fel de extreme ca si valoarea observata TS_0 (valori "extreme" in sensul testului pe care il facem, adica valori spre stanga $TS < TS_0$, pentru test la stanga, la dreapta $TS > TS_0$, pentru test la dreapta si in ambele parti $TS > |TS_0|$, pentru test bilateral). Ca relevanta, $P-$ul reprezinta pragul

minim (ca nivel de semnificatie) de respingere, adica cel mai mic α pentru care inca se respinge H_0 (daca $\alpha < P$, nu se mai respinge H_0). Cum stabilim daca respingem sau nu ipoteza nula? Pe doua cai:

– **hypothesis testing:**
if $TS_0 \in RR$, reject H_0 , otherwise, for $TS_0 \notin RR$, do not reject H_0;

– **significance testing**: testam nivelul de semnificatie
if $\alpha \geq P$, reject H_0 , otherwise, for $\alpha < P$, do not reject H_0;

Evident ca pe ambele cai obtinem acelasi raspuns. De aceea le cer clar sa afiseze **toate** cele 3 lucruri, TS_0, RR, P, ca sa vedem clar, pe ambele cai, de ce dam un raspuns sau altul la problema.

Cazurile pe care le studiem sunt exact aceleasi cu cele de la intervale de invredere (si conditiile teoretice tot aceleasi):

– $\theta = \mu$, media de populatie, cu doua cazuri, σ cunoscut sau nu;
– $\theta = \sigma^2$, varianta de populatie;
– $\theta = \mu_1 - \mu_2$, diferenta mediilor de populatie, cu trei cazuri, σ_1, σ_2 cunoscute, $\sigma_1 = \sigma_2$, necunoscute, $\sigma_1 \neq \sigma_2$, necunoscute,;
– $\theta = \frac{\sigma_1^2}{\sigma_2^2}$, raportul variantelor de populatie.

Pentru toate cazurile (cu subcazuri), formulele pentru cele trei variabile (TS_0, RR, P) sunt date in fisierul tests.pdf.

• Setarea celor doua ipoteze si rezolvarea problemei

Prima chestiune importanta la orice problema este sa setam corect cele doua ipoteze. Am spus ca ipoteza nula e **totdeauna** sub forma de egalitate, dar si ca cele doua ipoteze sunt opuse. Atunci cum facem, ca se vede clar mai sus, ca doar a treia alternativa e opusa reala a ipotezei nule. Pai, "opuse" din punct de vedere al problemei.

Probl. 1

1. a) Daca, in medie, nr-ul de fisiere stocate e 9, e ok. Daca e mai mare decat 9, e si mai ok. Problema e daca e mai mic. Deci, din punctul de vedere al acestei probleme, $\mu > 9$ intra impreuna cu $\mu = 9$ si opusul e $\mu < 9$. Deci facem test la stanga pentru medie:

$$\begin{aligned} H_0 : \mu &= 9 \\ H_1 : \mu &< 9 \end{aligned}$$

Bun, acum ca am vazut clar pentru ce parametru fac testul si ce fel de test fac, am toate formulele. Doar ca nu trebuie sa le calculez cu mana, face Matlabul asta in locul meu. Aici suntem in cazul 1. primul subcaz (σ cunoscut). Statistica de test are o distributie $N(0,1)$ (si variabilele aleatoare cu distributie normala standard se noteaza in statistica cu Z), deci fac un "test Z". Acesta e implementat in Matlab, `ztest`. Dam un `help ztest` sa vedem cum functioneaza, ce ne returneaza si ce mai trebuie calculat.

```
[H,P,CI,ZVAL] = ztest(X,M,SIGMA,ALPHA,TAIL)
```

Ce date de intrare are:

– X e vectorul continand datele de selectie;
– M e valoarea de test (cu care compar), in notatia noastra θ_0, sau, in cazul acesta, fiind

vorba de medie, μ_0, adica 9;
– σ e deviatia standard de populatie, fiindca suntem in cazul Z, cand aceasta este cunoscuta, in cazul nostru 5;
Astea primele trei sunt obligatorii. In rest, optional:
– α e nivelul de semnificatie, daca nu il dam, valoarea default e $\alpha = 0.05$ (sau nivel de semnificatie 5%), noi il vom cere de la user cu `input`, pentru a face comparatie;
– $Tail$ specifica tipul testului (stanga, dreapta, bilateral, default "bilateral"), acolo zice sa-l dai cu cuvinte, "left", etc, e mai simplu cu numere: -1 pentru test la stanga, 0 pentru test bilateral (valoarea default, daca nu il specificam, asta face) si 1 pentru test la dreapta.

Ce date de iesire returneaza:
– H prin valoarea sa ne spune daca se respinge H_0 sau nu, si anume $H = 0$, nu se respinge, $H = 1$, se respinge ipoteza nula. Vom face un `if` asupra valorii lui, pentru a da raspunsul;
– P e valoarea P a testului, una din cele 3 valori care trebuie afisate;
– CI e intervalul de incredere de $100(1-\alpha)\%$, deci puteam la laboratorul anterior sa folosim asta, in loc sa le calculam noi cu mana, dar nu avea niciun sens, pana nu discutam teste statistice. De acum incolo, pot folosi asta. Dar **atentie** la doua lucruri: 1. E vorba de CI **in sensul testului facut**, adica CI la stanga, la dreapta, sau bilateral. Noi n-am discutat decat intervale de incredere bilaterale, cu doua margini, inferioara, superioara. Le-am povestit ca pot fi si intervale de incredere unilaterale, (θ_L, ∞) sau $(-\infty, \theta_U)$, dar n-am avut probleme de laborator cu astea. 2. Le spun la curs despre legatura dintre intervale de incredere si regiuni critice. La testul Z bilateral, $RR = (-\infty, z_{\alpha/2}) \cup (z_{1-\alpha/2})$, iar pentru intervalul de incredere pornim exact cu complementara multimii $Z \in (z_{\alpha/2}, z_{1-\alpha/2})$, dar intervalul de incredere e pentru $\mu \in (\overline{X} - (z_{1-\alpha/2})\frac{\sigma}{\sqrt{n}}, \overline{X} - (z_{\alpha/2})\frac{\sigma}{\sqrt{n}})$, adica sa nu aiba impresia ca intervalul de incredere e pur si simplu complementarul regiunii critice. Regiunea critica e pentru statistica de test (sau pivotul) Z, ir intervalul de incredere e pentru parametrul necunoscut μ;
– $Zval$ e valoarea observata a statisticii de test, ce am notat noi cu TS_0 (in cazul asta Z_0), deci al doilea din cele 3 lucruri pe care le cer afisate.

Deci mai ramane ca ei sa determine RR, dupa formulele din fisier (le trebuie cuantilele, aceleasi ca la intervale de incredere pentru cazul corespunzator) si sa fie afisata **sub forma de interval**, cu paranteze, cu virgula, cu semnul de $\cup$ daca e cazul (`Inf` sau `inf` il stie Matlabul ca e ∞, nu are nevoie de format la tiparire cu `fprintf`).

Deci, se afiseaza cele 3 lucruri (TS_0, RR, P), se face testul asupra lui h, daca e 0 sau 1 si se da raspunsul la problema, **in doua feluri**, raspunsul statistic (resping/nu resping H_0) si interpretarea in cuvinte a acestuia pentru problema. Aici problema intreba "does the data suggest that the standard is met?", deci la asta raspund.

Se executa programul cu doua nivele de semnificatie diferite, $\alpha = 0.05$ si $\alpha = 0.01$ (primul e mai mare decat $P-$ul testului, al doilea mai mic), se compara rezultatele si se vede pe ambele cai de ce raspunsul e unul sau altul.

Si cam asta e o problema tipica de testare a ipotezelor statistice. De aici incolo, se schimba distributiile (T, χ^2, F) dar ideile si functiile Matlab merg la fel.

1. b) Aici se spune clar ce sa se testeze (tot pentru media de populatie), si anume daca e

mai mare decat 5.5. Deci e test la dreapta tot pentru media de populatie, dar in cazul al doilea (σ necunoscut).

$$\begin{aligned} H_0 : \mu &= 5.5 \\ H_1 : \mu &> 5.5 \end{aligned}$$

Nu se mai cunoaste σ, se schimba distributia, e distributie Student $T(n-1)$, deci folosim `ttest`.

```
[H,P,CI,STATS] = ttest(X,M,,ALPHA,TAIL)
```

functioneaza ca si ztest, cu doua diferente: nu se mai da σ (se foloseste s in loc, dar nu trebuie sa il, isi calculeaza singur), iar ultima variabila `stats` cuprinde mai multe informatii, e o structura cu 3 campuri
'tstat' – the value of the test statistic, TS_0
'df' – the degrees of freedom of the test, adica parametrul legii Student folosite, $n-1$
'sd' – the estimated population standard deviation, o aproximare a lui σ.

Extragem campul care ne intereseaza, cu `stats.tstat`, sau `stats.df`, etc. Daca in apelul lui tstat am folosit alta variabila pe ultima pozitie,

```
[h,p,ci, moscraciun] = ttest...,
```

nu-i nicio problema, la fel extragem ce ne intereseaza, moscraciun.tstat, etc.

Le fel, ca la punctul a), se afiseaza cele 3 lucruri, se executa cu diferite nivele de semnificatie si se comenteaza in fiecare caz.

Urmatorul punct, c), l-am scos din problema, fiindca am comasat laboratoare, dar il las la explicatii. E vorba despre test pentru varianta de populatie.

1. c) Aici e destul de clar. Ma intereseaza daca intr-adevar, ma pot baza pe presupunerea ca $\sigma = 5$. Deci, test bilateral pentru deviatia standard:

$$\begin{aligned} H_0 : \sigma &= 5 \\ H_1 : \sigma &\neq 5 \end{aligned}$$

Problema e ca, teoria e pentru *varianta*, nu pentru *deviatia standard*. Nu-i nimic, avand in vedere ca deviatia standard e totdeauna pozitiva, testul de mai sus e echivalent cu testul pentru varianta:

$$\begin{aligned} H_0 : \sigma^2 &= 25 \\ H_1 : \sigma^2 &\neq 25 \end{aligned}$$

De la Matlab 2008 incoace, sunt implementate si teste pentru variante (pana atunci le implementam noi). Avem

```
[H,P,CI,STATS] = vartest(X,V,,ALPHA,TAIL)
```

cu aceelasi semniicatii ale variabilelor ca la celelelte teste. Structura `stats` contine doua campuri, 'chisqstat' – the value of the test statistic, 'df' – the degrees of freedom of the test (adica $n-1$).

Aici mai trebuie facuta o observatie: Observam ca avem $P = 0.1598$, adica muuuult mai mare decat orice nivel de semnificatie rezonabil. Deci orice α am da (pana in 10%), rezultatul va fi acelasi: NU se respinge H_0, adica varianta ESTE egala cu 25. Si fiind un P atat de mare, inseamna ca datele sugereaza *puternic* ca $\sigma = 5$. La fel, vor fi cazuri cand P e foarte mic, comparat cu orice nivel de semnificatie am putea seta, atunci datele vor sugera puternic ca trebuie RESPINSA ipoteza nula.$\Big)$

Probl. 2

Totul merge la fel pentru compararea mediilor sau variantelor a doua populatii. Aici ne intereseaza sa comparam mediile (sa vedem daca e mai bun gas mileage cand dam mai multi bani pe benzina premium, merita sau nu?). Dar pentru a compara mediile, trebuie sa stim ce caz folosim, sunt egale variantele de populatie, sau nu (pentru primul caz, cand se cunosc efectiv valorile lor, nu avem aplicatie, ca nu e realist). In loc sa presupunem $\sigma_1 = \sigma_2$ sau $\sigma_1 \neq \sigma_2$, cum am facut la intervale de incredere, aici testam asta:

$$\begin{aligned} H_0 : \sigma_1 &= \sigma_2, \text{ adica } \sigma_1^2 = \sigma_2^2, \text{ adica } \frac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1 : \sigma_1 &\neq \sigma_2, \text{ adica } \sigma_1^2 \neq \sigma_2^2, \text{ adica } \frac{\sigma_1^2}{\sigma_2^2} \neq 1, \end{aligned}$$

din moment ce teoria e pentru parametrul $\theta = \dfrac{\sigma_1^2}{\sigma_2^2}$. Avem

```
[H,P,CI,STATS] = vartest2(X,Y,ALPHA,TAIL)
```

cu semnificatiile de mai inainte, doar ca acum dam datele a doua selectii (vectorii X si Y) si nu mai dam o valoare de test (θ_0), ci totdeauna se compara cu 1. Din nou, structura `stats` contine 3 campuri, 'fstat' – the value of the test statistic 'df1' – the numerator degrees of freedom of the test, adica $n_1 - 1$, 'df2' – the denominator degrees of freedom of the test, adica $n_2 - 1$.

In functie de raspunsul la punctul a), daca respingem ipoeza nula (asta insemnand $\sigma_1 \neq \sigma_2$) sau nu ($\sigma_1 = \sigma_2$), putem trece la punctul b), compararea mediilor de populatie. La noi a iesit ca variantele de populatie sunt egale (nu s-a respins H_0), deci asta vom folosi. Ce fel de test avem? Zice "does gas mileage seem to be higher, on average, when premium gasoline is used?", deci e media de la premium mai mare decat de la regular?

$$\begin{aligned} H_0 : \mu_1 &= \mu_2, \text{ adica } \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 &> \mu_2, \text{ adica } \mu_1 - \mu_2 > 0, \end{aligned}$$

caci teoria e pentru parametrul $\theta = \mu_1 - \mu_2$. Deci, test la dreapta. Avem:

```
[H,P,CI,STATS] = ttest2(X,Y,ALPHA,TAIL,VARTYPE)
```

care compara parametrul $\theta = \mu_1 - \mu_2$ cu 0. In plus fata de ce aveam inainte, mai apare `vartype`, prin care ii spunem daca variantele de populatie sunt egale (ii dam `vartype='equal'` si asta e default-ul, deci putem sa nu-l dam) sau nu (ii dam `vartype='unequal'`. Ultima variabila, `stats`, contine 3 campuri: 'tstat' – the value of the test statistic 'df' – the degrees of freedom of the test 'sd' – the pooled estimate of the population standard deviation (for the equal variance case) or a vector containing the unpooled estimates of the population standard deviations (for the unequal variance case). Extragem de acolo ce dorim, la fel ca mai sus.

Aici avem cazul extrem in directia opusa, $P = 4.944095e{-}007$ (sa le si spui sa-l afiseze cu format %e, altfel apare doar 0), foaaaaaaaaarte mic, deci aici nu vom seta niciodata un α mai mic decat acest P. Deci totdeauna respingem H_0, adica datele sugereaza *foarte puternic* ca μ_1 e mai mare decat μ_2, adica ca, da, e mai bine cu benzina premium, merita bani mai multi...

O ultima observatie: daca la punctul a) iesea ca variantele de populatie difera (respingeam ipoteza nula), atunci la testul pentru medii dadeam optiunea "unequal":

```
[H,P,CI,STATS] = ttest2(X,Y,ALPHA,TAIL,"unequal")
```