Analiza datelor în R

Curs 7

Statistici de selecţie

Fie $X_1, X_2, \ldots, X_n$ variabile aleatoare independente, identic distribuite (i.i.d.). Definim:

- media de selecţie $\overline{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$
- dispersia de selecţie $S^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$

Principii generale

- Se formulează cu privire la parametrul sau parametrii de interes ai populaţiei o ipoteză nulă H_0 şi o ipoteză alternativă H_a;
- Pe baza datelor din eşantion se poate lua decizia nerespingerii lui H_0 sau a respingerii lui H_0, i.e., a acceptării lui H_a.

- $P(\text{resping } H_0 \mid H_0 \text{ e adevărată}) = \alpha =$ nivelul de semnificaţie al testului
- $P(\text{nu resping } H_0 \mid H_0 \text{ e falsă}) = \beta$
- $1 - \beta = P(\text{accept } H_a \mid H_a \text{ e adevărată}) =$ puterea testului

- R. Fisher - Lady tasting tea experiment

Principii generale

- Se alege o statistică test a cărei distribuţie este cunoscută când H_0 este adevărată şi se calculează valoarea ei pe datele din eşantion;
- Dacă probabilitatea de a obţine valori cel puţin "la fel de extreme" sub ipoteza nulă (**p-valoarea testului**) este foarte mică (mai mică decât α), vom respinge H_0 în favoarea lui H_a. În caz contrar, nu respingem H_0.

Testul t - un eşantion

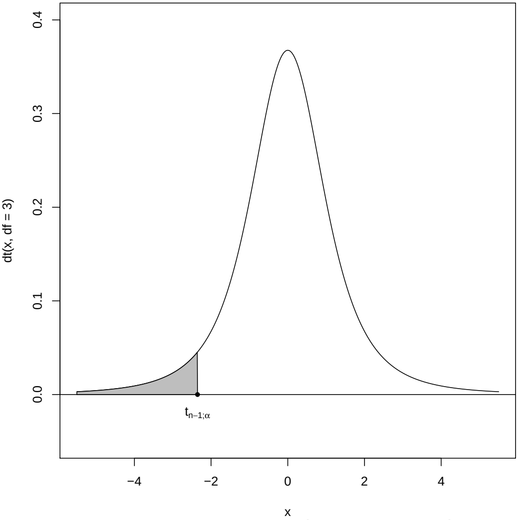
Este un test pentru media unei populaţii cu volum de eşantion mare ($n > 30$) sau distribuţie (aproximativ) normală şi volum de eşantion mic, cu dispersie necunoscută.

- $H_0 : \mu = \mu_0$
- $H_a : \mu \neq \mu_0$ sau $H_a : \mu < \mu_0$ sau $H_a : \mu > \mu_0$
- Statistica test: $T = \dfrac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}}$

  În ipoteza că H_0 e adevărată, $T \sim t(n-1)$.
- $t_{obs} = \dfrac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$
- Dacă p-valoarea corespunzătoare lui t_{obs} este $< \alpha$, se respinge H_0; în caz contrar, nu se respinge H_0 la nivelul de semnificaţie ales.

Testul t - un eşantion

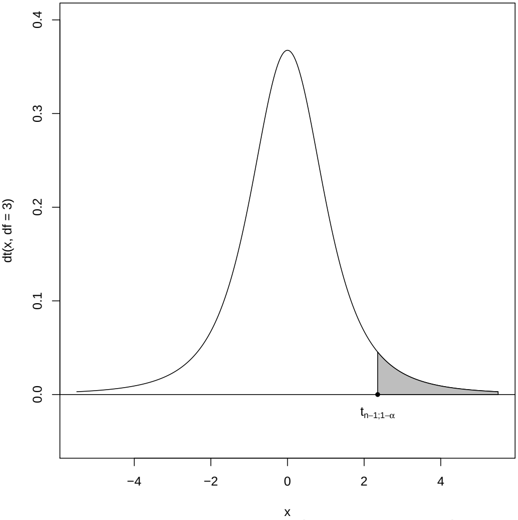
$H_a : \mu < \mu_0$



Regiunea critică: $(-\infty, t_{n-1;\alpha})$

Testul t - un eşantion

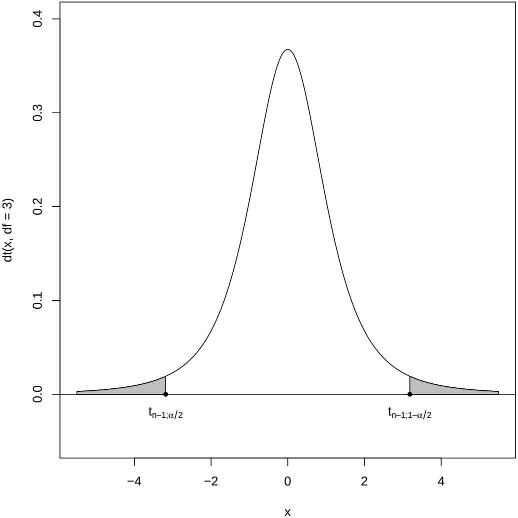
$H_a : \mu > \mu_0$



Regiunea critică: $(t_{n-1;1-\alpha}, \infty)$

Testul t - un eşantion

$H_a : \mu \neq \mu_0$



Regiunea critică: $(-\infty, t_{n-1;\frac{\alpha}{2}}) \cup (t_{n-1;1-\frac{\alpha}{2}}, \infty)$

Testul t - un eşantion

În R: `t.test(x, mu=mu0, alternative="two.sided" /"less"/"greater")`

Exemplu:

O fabrică producătoare de baterii susţine că durata medie de funcţionare a produselor sale este de 180 ore. Pentru verificare, se aleg la întâmplare şi se analizează 50 baterii. Duratele lor de funcţionare se găsesc în dataframe-ul *Battery* din *PASWR*, în variabila *facilityA*. Vom testa la nivelul de semnificaţie 0.05 cele susţinute de firmă.

Compararea mediilor a două populaţii

Pentru populaţii cu distribuţie aproximativ normală sau volum de eşantion mare:

- testul t pentru eşantioane independente
- testul t pentru eşantioane dependente (perechi)

În cazul populaţiilor cu distribuţie non-normală şi eşantioane mici se folosesc teste neparametrice:

- eşantioane independente $\rightarrow$ testul Mann - Whitney;
- eşantioane dependente (perechi) $\rightarrow$ testul Wilcoxon.

Testul t pentru două eşantioane independente

Se foloseşte pentru compararea mediilor a două populaţii independente cu distribuţie aproximativ normală sau volume de eşantion suficient de mari.

- $H_0 : \mu_1 - \mu_2 = \mu_0$
- $H_a : \mu_1 - \mu_2 \neq \mu_0$ sau $H_a : \mu_1 - \mu_2 < \mu_0$ sau $H_a : \mu_1 - \mu_2 > \mu_0$
- Statistica test: $T = \dfrac{(\overline{X}_1 - \overline{X}_2) - \mu_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$

  În ipoteza că H_0 e adevărată, T are distribuţie t (Welch)
- $t_{obs} = \dfrac{(\overline{x}_1 - \overline{x}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
- Dacă p-valoarea corespunzătoare lui t_{obs} este $< \alpha$, se respinge H_0; altfel, nu se respinge H_0 la nivelul de semnificaţie ales.

Testul t pentru două eşantioane independente

În R: `t.test(x, y, mu=mu0, alternative="two.sided" /"less"/"greater")`

Exemplu:

Pentru setul de date *mtcars* din *datasets*, vom testa la nivelul de semnificaţie 0.05 ipoteza că maşinile cu transmisie manuală ($am = 0$) sunt mai puţin eficiente decât cele cu transmisie automată ($am = 1$) în privinţa consumului.

Testul t pentru două eşantioane dependente (perechi)

Se foloseşte pentru compararea mediilor a două populaţii cu distribuţie aproximativ normală sau volume de eşantion suficient de mari, în cazul în care acestea constituie observaţii repetate sau perechi.

- $H_0 : \mu_1 - \mu_2 = \mu_0$
- $H_a : \mu_1 - \mu_2 \neq \mu_0$ sau $H_a : \mu_1 - \mu_2 < \mu_0$ sau $H_a : \mu_1 - \mu_2 > \mu_0$
- Se consideră $D_i = X_i - Y_i$, $i = \overline{1, n}$. Atunci D_i sunt i.i.d. şi

$$\mu_1 - \mu_2 = \mu_0 \Leftrightarrow \mu_D = \mu_0.$$

$\longrightarrow$ se poate utiliza testul t pentru un singur eşantion reprezentând mulţimea diferenţelor din perechile de observaţii.

Testul t pentru două eşantioane dependente (perechi)

În R: `t.test(x, y, mu=mu0, alternative="two.sided" /"less"/"greater", paired=TRUE)`

Exemplu:

O dietă este promovată ca fiind capabilă să reducă în mod considerabil nivelul de glucoză din sânge. Zece pacienţi diabetici sunt aleşi aleator şi puşi să urmeze dieta o lună, şi rezultatele lor sunt prezentate mai jos.

înainte	268	225	252	192	307	228	246	298	231	185
după	106	236	253	110	203	101	211	176	194	203

Verificăm la nivelul de semnificaţie 0.05 dacă există suficiente dovezi care să susţină eficienţa dietei la pacienţii diabetici.

Testele Wilcoxon şi Mann-Whitney

Date două populaţii X şi Y, testele Wilcoxon şi Mann-Whitney se folosesc pentru a verifica ipoteza

$$H_0 : P(X < Y) = \frac{1}{2} \text{ (valorile din cele două grupuri sunt similare)}$$

cu una dintre alternativele

- $P(X < Y) \neq \frac{1}{2}$
- $P(X < Y) < \frac{1}{2}$
- $P(X < Y) > \frac{1}{2}$.

Testele Wilcoxon şi Mann-Whitney se pot folosi şi pentru date ordinale.

În R:

```
wilcox.test(x, y, alternative="two.sided"/"less"/
  "greater", paired=T/F)
```

Testele Wilcoxon şi Mann-Whitney

Exemplu:

Se planifică un studiu pilot care să testeze eficienţa administrării de suplimente de vitamina E pentru prevenirea bolii Alzheimer. 20 subiecţi cu vârste peste 65 ani sunt repartizaţi aleator în două grupuri. Primul grup (10 persoane) primeşte 400 UI/zi vitamina E, iar grupul al doilea primeşte un tratament placebo. Se înregistrează nivelul iniţial de vitamina E în fiecare grup, obţinându-se valorile:
$Grup\ 1$: 7.5, 12.6, 3.8, 20.2, 6.8, 403.3, 2.9, 7.2, 10.5, 205.4
$Grup\ 2$: 8.2, 13.3, 102.0, 12.7, 6.3, 4.8, 19.5, 8.3, 407.1, 10.2
Analizăm dacă există diferenţe între grupuri la momentul iniţial.