

Laborator 6 - Statistică inferențială

I. Inferență asupra mediei - Testul Z pentru media unei populații cu dispersia cunoscută

Se consideră o populație statistică căreia i se cunoaște dispersia σ^2 . Pentru un eșantion aleator simplu cu media de selecție \bar{x}_n , dacă populația urmează o lege normală sau dimensiunea eșantionului este suficient de mare, scorul $z = \frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}}$ este distribuit normal standard: $N(0, 1)$.

Testul Z decurge astfel:

1. se formulează ipoteza nulă, care susține că media populației ia o valoare particulară:

$$H_0 : \mu = \mu_0$$

2. se formulează o ipoteză alternativă care poate fi de trei feluri:

$$H_a : \mu < \mu_0 \quad (\text{ipoteză asimetrică la stânga}) \text{ sau}$$

$$H_a : \mu > \mu_0 \quad (\text{ipoteză asimetrică la dreapta}) \text{ sau}$$

$$H_a : \mu \neq \mu_0 \quad (\text{ipoteză simetrică})$$

3. se fixează nivelul de semnificație: α (care uzual poate fi 1% sau 5%);
4. se calculează scorul testului:

$$z = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}$$

5. se determină valoarea critică z^* :

$$z^* = qnorm(\alpha, 0, 1) \quad \text{pentru ipoteză } H_a \text{ asimetrică la stânga } (z^* < 0),$$

$$z^* = qnorm(1 - \alpha, 0, 1) \quad \text{pentru ipoteză } H_a \text{ asimetrică la dreapta } (z^* > 0),$$

$$z^* = -qnorm(\alpha/2, 0, 1) = qnorm(1 - \alpha/2, 0, 1) \quad \text{pentru ipoteză } H_a \text{ simetrică } (z^* > 0).$$

6. ipoteza nulă H_0 este respinsă dacă

$$z < z^* \quad \text{pentru ipoteză } H_a \text{ asimetrică la stânga sau}$$

$$z > z^* \quad \text{pentru ipoteză } H_a \text{ asimetrică la dreapta sau}$$

$$|z| > |z^*| \quad \text{pentru ipoteză } H_a \text{ simetrică,}$$

dacă nu suntem într-una din aceste situații, atunci se spune că **nu există suficiente dovezi pentru a respinge ipoteza nulă H_0 și a accepta ipoteza alternativă H_a .**

Exercițiu rezolvat. Un producător de becuri dorește să testeze cu 5% nivel de semnificație afirmația că media de viață a acestora este de cel puțin 810 de ore (se știe că deviația standard a populației este $\sigma = 50$ de ore). Se alege un eșantion de 200 de becuri a căror medie de viață este găsită 816 ore. Poate fi acceptată ipoteza producătorului?

```

> alfa = 0.05
> population_mean = 810
> sample_mean = 816
> n = 200
> sigma = 50
> critical_z = qnorm(1- alfa)
> z_score = (sample_mean - population_mean)/(sigma/sqrt(n))
> critical_z
> z_score

```

Scorul va fi $z = 1.69705 > z^* = 1.64485$ și ipoteza nulă poate fi respinsă, se acceptă ipoteza că media populației este mai mare decât 810.

Exerciții propuse

- I.1 Scrieți o funcție (numită **z_test**) care să calculeze și să returneze valoarea critică și scorul testului (parametrii funcției vor fi: tipul ipotezei alternative, n , μ_0 , \bar{x}_n , α , σ etc). Funcția aceasta va fi utilizată apoi la rezolvarea exercițiilor de mai jos.
- I.2 Din experiență se știe că rezultatele studenților la un test de matematică urmează o lege normală cu media 75 și dispersia 17. Catedra de matematică dorește să afle dacă studenții din anul curent au un comportament atipic. Media rezultatelor unui grup de 36 studenți este 85 de puncte. Cu 1% nivel de semnificație se poate trage concluzia că studenții din anul curent sunt atipici?
- I.3 Pe cutiile de un anumit tip de detergent este indicată o greutate de 21oz. O agenție de protecție a consumatorilor dorește să testeze această greutate cu 1% nivel de semnificație. Pentru 100 de cutii găsește o greutate medie de 20.5oz. Dacă se știe că deviația standard a greutății este 2.5oz, agenția poate pretinde mărirea cantității de detergent dintr-o cutie?
- I.4 O firma producătoare de tuburi fluorescente dorește să știe dacă poate pretinde că media de viață a acestora este 1000 de ore. Pentru aceasta fabrică 100 de tuburi și măsoară pentru ele o medie de viață de 970 de ore. Firma respectivă cunoaște că deviația standard a vieții tuburilor este 85 de ore. Cu 5% nivel de semnificație se poate trage concluzia că media de viață este mai mică de 1000 de ore? Dar cu 1%?
- I.5 Se cere ca media de viață a unui tip de baterii să fie 220 de ore. Se știe (din procesul de fabricație) că durata de viață a bateriilor urmează o lege normală cu deviația standard 9 ore. Un eșantion de 36 baterii are o medie de viață măsurată de 218 de ore. Se poate trage concluzia că media de viață a bateriilor este diferită de 220 de ore?

II. Inferență asupra mediei - Testul t pentru media unei populații cu dispersia necunoscută

Se consideră o populație statistică distribuită normală căreia nu i se cunoaște dispersia. Pentru un eșantion aleator simplu cu media de selecție \bar{x}_n și deviația standard s , scorul $t = \frac{\bar{x}_n - \mu}{s/\sqrt{n}}$ este distribuit Student cu $n - 1$ grade de libertate: $t(n - 1)$.

Testul t decurge astfel:

1. se formulează ipoteza nulă, care susține că media populației ia o valoare particulară:

$$H_0 : \mu = \mu_0$$

2. se formulează o ipoteză alternativă care poate fi de trei feluri:

$$H_a : \mu < \mu_0 \quad (\text{ipoteză asimetrică la stânga}) \text{ sau}$$

$$H_a : \mu > \mu_0 \quad (\text{ipoteză asimetrică la dreapta}) \text{ sau}$$

$$H_a : \mu \neq \mu_0 \quad (\text{ipoteză simetrică})$$

3. se fixează nivelul de semnificație: α (care uzual poate fi 1% sau 5%);

4. se calculează scorul testului:

$$t = \frac{\bar{x}_n - \mu_0}{s/\sqrt{n}}$$

5. se determină valoarea critică t^* :

$$t^* = qt(\alpha, n - 1) \quad \text{pentru ipoteză } H_a \text{ asimetrică la stânga } (t^* < 0),$$

$$t^* = qt(1 - \alpha, n - 1) \quad \text{pentru ipoteză } H_a \text{ asimetrică la dreapta } (t^* > 0),$$

$$t^* = -qt(\alpha/2, n - 1) = qt(1 - \alpha/2, n - 1) \quad \text{pentru ipoteză } H_a \text{ simetrică } (t^* > 0).$$

6. ipoteza nulă H_0 este respinsă dacă

$$t < t^* \quad \text{pentru ipoteză } H_a \text{ asimetrică la stânga sau}$$

$$t > t^* \quad \text{pentru ipoteză } H_a \text{ asimetrică la dreapta sau}$$

$$|t| > |t^*| \quad \text{pentru ipoteză } H_a \text{ simetrică,}$$

altfel se spune că **nu există suficiente dovezi pentru a respinge ipoteza nulă H_0 și a accepta ipoteza alternativă H_a .**

Exercițiu rezolvat. Pentru un experiment asupra metabolismului 5 insecte sunt hrănite cu zahăr. Valorile nivelului de glucoză (care urmează o lege normală) obținute din măsurători sunt:

55.95 68.24 52.73 21.5 23.78

Sa se testeze cu 5% nivel de semnificație ipoteza că media nivelului de glucoză este mai mare de 40.

```
> alfa = 0.05
> x = c(55.95, 68.24, 52.73, 21.5, 23.78)
> population_mean = 40
> sample_mean = mean(x)
> n = 5
> s = sd(x)
> se = s/sqrt(n)
> critical.t = qt(1 - alfa, n - 1)
> t.score = (sample_mean - population_mean)/se
> critical.t
> t.score
```

Rezultatul va fi $t^* = 2.13184 > t = 0.47867$, ipoteza nulă nu poate fi respinsă.

Exerciții propuse

II.1 Scrieți o funcție (de tipul `t.test`) care să calculeze și să returneze valoarea critică și scorul testului `t` (funcția va primi ca argumente tipul ipotezei alternative, media de selecție, deviația standard a eșantionului etc). Funcția aceasta va fi utilizată pentru rezolvarea exercițiilor de mai jos.

II.2 Se măsoară pentru un esantion provenit dintr-o populație normală următoarele valori

36 32 28 33 41 28 31 26 29 34

Cu 1% nivel de semnificație să se testeze ipoteza că media are o valoare diferită de 34.

II.3 Pe pachetele unui tip de țigări este trecută o concentrație de nicotină (care urmează o lege normală) de 11.4 mg. Datorită unor reclamații, o agenție neguvernamentală se hotărăște să testeze această concentrație. Pentru 100 de pachete de țigări este găsită o medie a concentrației de 11.9 mg cu o deviație standard $s = 0.25$ mg. Să se testeze cu 1% și 5% nivel de semnificație dacă reclamațiile primite sunt îndreptățite.

II.4 Media rezultatelor unui test la istorie este de 80 de puncte. Catedra de istorie dorește să afle dacă studenții actuali au un comportament tipic la acest test. Pentru un eșantion aleator simplu rezultatele se găsesc în fișierul `history.txt`. Să se formuleze și să se testeze ipoteza alternativă corespunzătoare (cu 1% și 5% nivel de semnificație).

I.5 Se consideră un eșantion de dimensiune 64 cu media 52 și dispersia $s^2 = 89.5$, care provine dintr-o populație distribuită normal. Să se testeze ipoteza că media populației este 49 versus ipoteza că media este diferită de 49.

III. Testul χ^2 de concordanță (goodness-of-fit)

Ca test de concordanță (goodness-of-fit) testul χ^2 își propune să determine cât de bine o distribuție teoretică se potrivește cu una empirică (obținută din eșantion). Presupunem că indivizii dintr-un anumit eșantion pot fi grupați în k categorii C_1, C_2, \dots, C_k cu frecvențele (absolute) observate o_1, o_2, \dots, o_k . Pe de altă parte se așteaptă ca frecvențele (absolute) teoretice să fie e_1, e_2, \dots, e_k . Fie $m = \sum_{i=1}^k o_i$.

Testul decurge astfel:

1. se formulează ipoteza nulă, care susține că frecvențele observate și cele teoretice sunt identice:

$$H_0 : (o_1, o_2, \dots, o_k) = (e_1, e_2, \dots, e_k)$$

2. se formulează o ipoteză alternativă care spune că frecvențele observate și cele teoretice sunt diferite:

$$H_a : (o_1, o_2, \dots, o_k) \neq (e_1, e_2, \dots, e_k)$$

3. se alege nivelul de semnificație: $\alpha \in \{1\%, 5\%\}$;

4. se calculează scorul χ^2 (statistica testului):

$$\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

5. se determină valoarea critică:

$$\chi^{2*} = qchisq(1 - \alpha, k - 1) \quad H_a \text{ este asimetrică la dreapta}$$

6. ipoteza nulă H_0 este respinsă și se acceptă H_a dacă

$$\chi^2 > \chi^{2*}$$

altfel se spune că **nu există suficiente dovezi pentru a respinge ipoteza nulă H_0 și a accepta ipoteza alternativă H_a .**

Exercițiu rezolvat. Tabelul de mai jos conține frecvențele observate la aruncarea unui zar de 120 de ori. Testați ipoteza că zarul nu este corect construit cu 5% nivel de semnificație.

fața	1	2	3	4	5	6
nr. apariții	25	17	15	23	24	16

Scorul este $\chi^2 = 5$ iar valoarea critică $\chi^{2*} = qchisq(0.95, 5) = 11.07$; deoarece $\chi^2 \not> \chi^{2*}$ nu putem respinge ipoteza nulă.

Exerciții propuse.

III.1 Scrieți o funcție (numită **chisq_goodness_of_fit_test**) care calculează și returnează valoarea critică și scorul pentru un test χ^2 (parametrii funcției vor fi: o , e și α). Funcția aceasta va fi utilizată pentru rezolvarea exercițiilor care urmează.

III.2 Gregor Mendel¹ a studiat 556 de boabe de mazare alese la întâmplare și a găsit că: 315 erau rotunde și galbene, 108 erau rotunde și verzi, 101 erau încrêțite și galbene iar 32 erau încrêțite și verzi. Conform cu teoria lui numerele acestea trebuiau să fie în proporția 9 : 3 : 3 : 1. Cu 1% nivel de semnificație se poate trage concluzia că această teorie nu este adevărată?

III.3 În 360 de aruncări a două zaruri de 37 de ori suma a fost egală cu cinci, de 74 de ori suma a fost egală cu șapte și de 24 de ori suma a fost egală cu unsprezece. Folosind un nivel de semnificație de 5%, testați ipoteza că zarurile nu sunt corect construite.

III.4 Două monede sunt aruncate de 240 de ori iar rezultatele se găsesc mai jos. Cu 1% nivel de semnificație testați ipoteza că zarurile nu sunt corect construite.

results	{ H, H }	{ H, T }	{ T, T }
frequencies	55	115	70

IV. Testul χ^2 pentru testarea independenței statistice

Pentru testarea independenței statistice testul χ^2 inferează asupra independenței a două variabile categorice. Valorile observate ale acestor două variabile sunt date într-un așa numit tabel de contingență.

		Y				
		o_{11}	o_{12}	\dots	o_{1r}	$o_{1,}$
X		o_{21}	o_{22}	\dots	o_{2r}	$o_{2,}$
		\vdots	\vdots	\dots	\vdots	
		o_{p1}	o_{p2}	\dots	o_{pr}	$o_{p,}$
		$o_{,1}$	$o_{,2}$	\dots	$o_{,r}$	m

¹Matematician și biolog austriac, inițiatorul geneticii.

unde o_{ij} este numărul de observații aparținând simultan categoriei i a lui X și j a lui Y , $e_{ij} = \frac{o_{i.} \cdot o_{.j}}{m}$ sunt frecvențele așteptate (teoretice) și $o_{i.} = \sum_{j=1}^r o_{ij}$ and $o_{.j} = \sum_{i=1}^p o_{ij}$, $m = \sum_{i=1}^p \sum_{j=1}^r o_{ij}$.

Testul decurge astfel:

1. se formulează ipoteza nulă, care susține că cele două variabile sunt independente:

$$H_0 : \text{variabilele sunt independente}$$

2. se formulează o ipoteză alternativă care spune că cele două variabile sunt dependente:

$$H_a : \text{variabilele sunt dependente}$$

3. se alege nivelul de semnificație: $\alpha \in \{1\%, 5\%\}$;

4. se calculează scorul χ^2 (statistica testului):

$$\sum_{i=1}^p \sum_{j=1}^r \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

5. se determină valoarea critică:

$$\chi^{2*} = qchisq(1 - \alpha, (p - 1)(r - 1)) \quad H_a \text{ este asimetrică la dreapta}$$

6. ipoteza nulă H_0 este respinsă și se acceptă H_a dacă

$$\chi^2 > \chi^{2*}$$

altfel se spune că **nu există suficiente dovezi pentru a respinge ipoteza nulă H_0 și a accepta ipoteza alternativă H_a .**

Exercițiu rezolvat. Dorim să știm dacă există o relație între sex și muzica preferată. Tabelul de mai jos conține rezultatele unui studiu asupra a 110 bărbați și 115 femei. Testați această ipoteză cu 5% nivel de semnificație.

	dance	rap	jazz	rock	country	classic	
bărbați	20	25	15	10	25	15	110
femei	25	15	20	10	20	25	115
	45	40	35	20	45	40	225

Scorul calculat este $\chi^2 = 6.717$ iar valoarea critică $\chi^{2*} = qchisq(0.95, 5) = 11.07$; deoarece $\chi^2 \not> \chi^{2*}$ nu putem respinge ipoteza nulă.

Exercises to work.

- IV.1 Scrieți o funcție (numită **chisq_independence_test**) care calculează și returnează valoarea critică și scorul pentru un test χ^2 pentru independență (parametrii funcției vor fi: *contingency_table* și α). Funcția aceasta va fi utilizată pentru rezolvarea exercițiilor următoare.
- IV.2 Se presupune că există o legătură (dependență) între venit și nivelul de educație. Tabelul de mai jos conține rezultatele unui sondaj. Testați ipoteza de mai sus cu 1% nivel de semnificație.

	middle school	high school	college	master	Ph. D.	
low income	40	25	25	10	25	125
middle income	35	35	30	30	35	165
upper income	20	30	35	40	45	170
	95	90	90	80	105	460

- IV.3 Se crede că mărcile de telefon utilizate depind de aria geografică a posesorului. Tabelul de mai jos conține rezultatele unui eșantion aleator de 746 proprietari de telefoane. Testați ipoteza de mai sus cu 5% nivel de semnificație.

	North A.	Europe	Asia	South A.	
Apple	106	74	36	28	244
Samsung	64	70	36	60	230
Google	26	12	6	4	48
Xiaomi	4	26	28	22	80
Oppo	2	6	70	10	88
Motorola	6	4	10	36	56
	208	192	186	160	746

- IV.4 Sunt nivelul de educație și sexul dependente? Un sondaj asupra unui eșantion aleator de 400 indivizi dă rezultatele de mai jos. Testați ipoteza de mai sus cu 1% nivel de semnificație.

	liceu	licență	master	doctorat	
bărbați	65	52	45	42	204
femei	41	45	52	58	196
	106	97	97	100	400

V. Inferență asupra dispersiilor a două populații - Testul F

Se consideră două populații normale ; din cele două populații se extrag două eșantioane aleatoare simple (și independente între ele) cărora li se calculează dispersiile s_1^2 și s_2^2 . Scorul

$F = \frac{s_1^2}{s_2^2}$ este distribuit $F(n_1 - 1, n_2 - 1)$.

Testul F decurge astfel:

1. se formulează ipoteza nulă, care susține că dispersiile celor două populații sunt egale:

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

2. se formulează ipoteza alternativă; putem avea două tipuri de ipoteză alternativă:

$$H_a : \frac{\sigma_1}{\sigma_2} > 1 \quad (\text{asimetrică la dreapta}) \text{ pentru un un test one-tailed}$$

$$H_a : \frac{\sigma_1}{\sigma_2} \neq 1 \quad (\text{simetrică}) \text{ pentru un un test two-tailed.}$$

3. se fixează nivelul de semnificație: α (care uzual poate fi 1% sau 5%);

4. se calculează scorul testului:

$$F = \frac{s_1^2}{s_2^2}$$

5. se determină valoarea critică (sau, după caz, valorile critice)

$$F^* = qf(1 - \alpha, n_1 - 1, n_2 - 1) \text{ pentru } H_a \text{ asimetrică la dreapta ,}$$

$$F_s^* = qf(\alpha/2, n_1 - 1, n_2 - 1), F_d^* = qf(1 - \alpha/2, n_1 - 1, n_2 - 1)$$

pentru H_a simetrică.

6. **ipoteza nulă H_0 este respinsă** și se acceptă H_a dacă

$$F > F^* \text{ pentru } H_a \text{ asimetrică la stânga,}$$

$$F < F_s^* \text{ sau } F > F_d^* \text{ pentru } H_a \text{ simetrică.}$$

altfel **nu există suficiente dovezi pentru a respinge ipoteza nulă.**

Exercițiu rezolvat. Rezultatele unui test psihologic efectuat pe două eșantioane, unul de femei și unul de bărbați sunt următoarele:

$$\text{bărbați: } n_1 = 120, s_1 = 5.05$$

$$\text{femei: } n_2 = 135, s_2 = 5.44$$

Se poate trage concluzia că dispersiile celor două populații diferă semnificativ (1%)?

```
> alfa = 0.01
> n1 = 120
> n2 = 135
> s1 = 5.05
> s2 = 5.44
> critical_F_s = qf(alfa/2, n1 - 1, n2 - 1)
> critical_F_d = qf(1 - alfa/2, n1 - 1, n2 - 1)
> critical_F_s
> critical_F_d
> F_score
```

Scorul este $F = 0.86175$, valorile critice sunt $F_s^* = 0.62843$ și $F_d^* = 1.58257$; deoarece $F \in [F_s^*, F_d^*]$ ipoteza nulă nu poate fi respinsă. În acest caz putem considera că nu există dovezi semnificative pentru a afirma că dispersiile sunt diferite.

Exerciții propuse

V.1 Scrieți o funcție (numită **F_test**) care să calculeze și să returneze valorile critice și scorul testului F (parametrii funcției vor fi: tipul ipotezei alternative, α , n_1 , n_2 , s_1 , s_2 etc.). Funcția aceasta va fi utilizată apoi la rezolvarea exercițiilor care urmează.

V.2 Cercetătorii studiază amplitudinea mișcării obținută prin stimularea nervoasă a șoarecilor. Pentru șoarecii drogați se obțin următoarele date:

12.512 12.869 19.098 15.350 13.297 15.589

Pentru șoarecii normali se obțin următoarele date:

11.074 9.686 12.164 8.351 12.182 11.489

Influența drogurilor este semnificativă în ceea ce privește cele două dispersii (5% nivel de semnificație)?

V.3 Un profesor crede că un anumit program de lectură îmbunătățește abilitățile și dorința copiilor de a citi. Pentru aceasta el alege două grupuri de elevi: unul de 22 de elevi care urmează programul prescris (A) și unul de 22 de elevi care nu urmează acest program (B). Rezultatele sunt date în fișierul *program.txt*. Să se decidă cu 1% și 5% nivel de semnificație dacă dispersiile celor două populații sunt diferite.