Laborator 7 - Statistică inferenţială

I. Testarea ipotezelor statistice - Testul χ^2 pentru dispersia unei populaţii

Se consideră o populaţie statistică normal distribuită căreia nu i se cunoaşte dispersia σ^2. Pentru un eşantion aleator simplu cu dimensiunea n şi dispersia s^2, scorul $\chi^2 = (n-1)\dfrac{s^2}{\sigma^2}$ este distribuit $\chi^2(n-1)$. Testul decurge astfel:

1. se formulează ipoteza nulă, care susţine că dispersia populaţiei ia o valoare particulară:

$$\boxed{H_0 : \ \sigma^2 = \sigma_0^2}$$

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

$$\boxed{H_a : \ \sigma^2 < \sigma_0^2} \quad \text{(ipoteză asimetrică la stânga) sau}$$

$$\boxed{H_a : \ \sigma^2 > \sigma_0^2} \quad \text{(ipoteză asimetrică la dreapta) sau}$$

$$\boxed{H_a : \ \sigma^2 \neq \sigma_0^2} \quad \text{(ipoteză simetrică)}$$

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

4. se calculează scorul testului:

$$\boxed{\chi^2 = (n-1) \cdot \frac{s^2}{\sigma_0^2}}$$

5. se determină valorile critice:

$$\boxed{\chi^{2*} = qchisq(\alpha, n-1)} \quad \text{pentru ipoteză } H_a \text{ asimetrică la stânga,}$$

$$\boxed{\chi^{2*} = qchisq(1-\alpha, n-1)} \quad \text{pentru ipoteză } H_a \text{ asimetrică la dreapta,}$$

$$\boxed{\chi_s^{2*} = qchisq\left(\frac{\alpha}{2}, n-1\right), \ \chi_d^{2*} = qchisq\left(1-\frac{\alpha}{2}, n-1\right)} \quad \text{pentru } H_a \text{ simetrică.}$$

6. **ipoteza nulă H_0 este respinsă dacă**

$$\boxed{\chi^2 < \chi^{2*}} \quad \text{pentru ipoteză } H_a \text{ asimetrică la stânga sau}$$

$$\boxed{\chi^2 > \chi^{2*}} \quad \text{pentru ipoteză } H_a \text{ asimetrică la dreapta sau}$$

$$\boxed{\chi^2 \notin (\chi_s^{2*}, \chi_d^{2*})} \quad \text{pentru ipoteză } H_a \text{ simetrică,}$$

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

Exerciţiu rezolvat. Rezultatele unui test de inteligenţă efectuat pe un eşantion aleator simplu de 120 de indivizi dintr-o populaţie distribuită normal oferă o deviaţie standard de 13 puncte. Se poate trage concluzia că dispersia populaţiei este mai mică decât 225? (1%)

```
> alfa = 0.01
> n = 120
> s_square = 169
> sigma _square = 225
> critical_Chi_square= qchisq(alfa, s - 1)
> Chi_square_score = (n - 1)*s_square / sigma _square
> critical_Chi_square
> Chi_square_score
```

Rezultatul va fi $\chi^2 = 89.3822 > \chi^{2*} = 86.07383$, deci ipoteza nulă nu poate fi respinsă.

Exerciţii propuse

I.1 Scrieţi o funcţie (numită **ChiSq_Variance _test**) care să calculeze şi să returneze valoarea critică şi scorul testului χ^2 pentru dispersie (parametrii funcţiei vor fi: α, n, s^2, σ_0^2). Funcţia aceasta va fi utilizată apoi pentru rezolvarea exerciţiilor care urmează.

I.2 O companie de cosmetice afirmă că dispersia cantităţii de parfum într-un flacon este aproximativ egală cu 0.2 . Se consideră un eşantion de dimensiune 100, pentru care se calculează o dispersie $s^2 = 0.22$. Să se testeze cu 5% nivel de semnificaţie dacă dispersia este mai mare decât 0.2.

I.3 Pentru un eşantion de dimensiune 35, se măsoară o deviaţie standard de 10.5. Să se testeze cu nivel de semnificaţie de 1% dacă deviaţia standard a populaţiei este diferită de 11.

I.4 Să se testeze ipoteza că dispersia unei populaţii este diferită de 12.8, dacă pentru un eşantion dat de dimensiune 49, dispersia masurata este mai mică decât 13.

I.5 Pentru 36 de pachete de ţigarete este găsită o deviaţie standard a concentraţiei de nicotină $s = 0.3$ mg. Să se testeze cu 1% nivel de semnificaţie dacă dispersia este mai mare decât 0.0625.

II. Testarea ipotezelor statistice - Testul χ^2 pentru experimente multinomiale

Se consideră un experiment cu s rezultate posibile, cu probabilităţi $p_1, p_2, \ldots, p_s$. Experimentul se repetă în mod independent de n ori şi se notează cu O_i numărul de realizări ale rezultatului i (O_i este o variabilă aleatoare distribuită binomial $B(n, p_i)$):

$$\sum_{i=1}^{s} O_i = n.$$

Pentru rezultatul i numărul mediu de realizări este (media variabilei O_i) $E_i = np_i$.

Statistica $\sum_{i=1}^{s} \frac{(O_i - Ei)^2}{E_i}$ urmează o distribuţie $\chi^2(s-1)$. Testul multinomial inferează asupra probabilităţilor $p_1, p_2, \ldots, p_s$ şi decurge astfel:

1. se formulează ipoteza nulă, care susţine că probabilităţile iau valori fixate.

$$\boxed{H_0 : \ p_i = \pi_i, \forall i}$$

2. se formulează o ipoteză alternativă:

$$\boxed{H_a : \ p_i \neq \pi_i, \text{ pentru măcar un } i}$$

3. se fixează nivelul de semnificaţie: α (1% sau 5%);
4. se calculează scorul testului:
$$\chi^2 = \sum_{i=1}^{s} \frac{(O_i - Ei)^2}{E_i}$$
5. se calculează valoarea critică:
$$\chi^{2*} = qchisq(1-\alpha, s-1)$$
6. **ipoteza nulă H_0 este respinsă dacă $\chi^2 > \chi^{2*}$**, altfel se spune că **nu există suficiente dovezi pentru a respinge ipoteza nulă H_0 şi a accepta ipoteza alternativă H_a.**

Exerciţiu rezolvat. Studenţii doresc o cât mai mare libertate în alegerea cursurilor. Şapte cursuri similare, predate de profesori diferiţi, au fost alese de 119 studenţi astfel (ordinea este aleatoare):

18 12 25 23 8 19 14

Cu5% nivel de semnificaţie, se poate trage concluzia că studenţii au avut preferinţe în alegerea cursurilor (i. e., există măcar un curs care este ales cu probabilitate diferită de 1/7)?

```
> alfa = 0.01
> O = c(18, 12, 25, 23, 8, 19, 14)
> p = c(1/7, 1/7, 1/7, 1/7, 1/7, 1/7, 1/7)
> s = length(O)
> n = sum(O)
> E = n*p
> critical_Chi_square= qchisq(1 - alfa, s - 1)
> Chi_square_score = sum((O - E)^2/E)
> critical_Chi_square
> Chi_square_score
```

Rezultatul va fi $\chi^2 = 12.94118 > \chi^{2*} = 12.59159$, deci ipoteza nulă poate fi respinsă, măcar unul dintre cursuri este preferat faţă de un altul.

Exerciţii propuse

II.1 Scrieţi o funcţie (numită **multinomial_test**) care să calculeze şi să returneze valoarea critică şi scorul testului multinomial (parametrii funcţiei vor fi: α, O şi p). Funcţia aceasta va fi utilizată apoi pentru rezolvarea exerciţiilor care urmează.

II.2 Compania Mars Inc. pretinde că bomboanele M&M sunt distribuite pe culori astfel: 35% sunt maro, 20% galbene, 15% roşii, 10% portocalii, 10% verzi şi 10% albastre. Pentru 100 de bomboane, se observă următoarea distribuţie pe culori:

	Maro	Galben	Roşu	Portocaliu	Verde	Albastru
Frecvenţa observată	33	26	18	9	7	7

Să se decidă cu 1% şi 5% nivel de semnificaţie dacă dispersiile celor două populaţii sunt diferite.

II.3 Dintre şoferii ce au suferit un accident de circulaţie, au fost selectaţi 88 de şoferi, în mod aleatoriu, şi au fost împărţiţi în funcţie de vârstă.

Vârstă	sub 25 de ani	între 25 şi 40	între 40 şi 65	peste 65 de ani
Frecvenţa observată	21	36	12	19

Testaţi ipoteza că distribuţia accidentelor pe categorii de vârstă este următoarea: 16%, 44%, 27%, 13%.