Teoria estimaţiei. Teste statistice (continuare)

Teste referitoare la două populaţii

Selecţii dependente şi independente

O **sursă** poate fi o persoană, un obiect sau orice altceva ce poate produce o dată. Când comparăm două populaţii avem nevoie de două selecţii, câte una pentru fiecare populaţie. Dacă se utilizează aceeaşi mulţime de surse pentru a obţine date reprezentând ambele populaţii avem de-a face cu **selecţii dependente**. Dacă se utilizează două mulţimi de surse nelegate, câte una pentru fiecare populaţie, avem **selecţii independente**.

Exemplu *Se elaborează un test pentru a vedea dacă studenţii îşi îmbunătăţesc pregătirea fizică în urma participării la orele de Educaţie fizică. Să presupunem că avem 500 de studenţi care participă la ore şi pentru verificare alegem 50 de studenţi la începutul anului şi 50 la terminare. Avem două proceduri:*

– *Procedura A – se selectează aleator 50 de studenţi şi se dă un pre-test la începutul anului. La sfârşitul anului se selectează aleator alţi 50 de studenţi şi se dă un post-test.*

– *Procedura B – se selectează 50 de studenţi la începutul anului şi se dă pre-testul, iar la sfârşit aceiaşi studenţi dau post-testul.*

Procedura A conduce la selecţii independente, iar B la selecţii dependente.

Exemplu *Se compară calitatea a două tipuri de pneuri.*

– *Procedura A – se selectează aleator n maşini şi se echipează cu pneuri de tipul 1 şi se conduc timp de o lună şi apoi alte m maşini cu pneuri de tipul 2 şi se conduc tot o lună.*

– *Procedura B – n maşini se selectează aleator, li se pune un pneu de tipul 1 şi un pneu de tipul 2 şi se conduc timp de o lună .*

Selecţiile din procedura A sunt independente, iar cele din B dependente.

Teste pentru diferenţa a două medii – selecţii independente

Se consideră două populaţii independente P_1 şi P_2 cercetate din punct de vedere al aceleiaşi caracteristici, notate cu X_1 pentru P_1, având distribuţia $N(m_1, \sigma_1^2)$ şi X_2 pentru P_2, având distribuţia $N(m_2, \sigma_2^2)$. Relativ la mediile teoretice ale celor două caracteristici se face ipoteza nulă

$$H_0 : m_1 \neq m_2$$

(sau echivalent $m_1 - m_2 = 0$), cu una din alternativele:

$H_1 : m_1 \neq m_2$ (test bilateral);
$H_1 : m_1 > m_2$ (test unilateral dreapta);
$H_1 : m_1 < m_2$ (test unilateral stînga).
Distribuţia lui $\bar{X}_1 - \bar{X}_2$ are proprietăţile:

1. este aproximativ normală;
2. are media $m_1 - m_2$;
3. are dispersia $\sigma^2 = \frac{\sigma_1^2}{m_1} + \frac{\sigma_2^2}{m_2}$.

Cazul 1. Dispersii cunoscute sau selecţii mari

Statistica

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad \boxed{1}$$

este $N(0,1)$, deci se poate aplica testul Z pentru media teoretică. Pentru $\alpha \in (0,1)$ se obţin regiunile critice corespunzătoare celor trei alternative:

$$U = \left\{ (u_1, \ldots, u_{n_1}, v_1, \ldots v_{n_2}) \in \mathbb{R}^{n_1+n_2} \left| \frac{|\bar{u} - \bar{v}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq z_{1-\alpha/2} \right. \right\};$$

$$U = \left\{ (u_1, \ldots, u_{n_1}, v_1, \ldots v_{n_2}) \in \mathbb{R}^{n_1+n_2} \left| \frac{\bar{u} - \bar{v}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq z_{1-\alpha} \right. \right\};$$

$$U = \left\{ (u_1, \ldots, u_{n_1}, v_1, \ldots v_{n_2}) \in \mathbb{R}^{n_1+n_2} \left| \frac{\bar{u} - \bar{v}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha} \right. \right\}.$$

Exemplu *Se consideră concentraţia de substanţă S din sângele a două populaţii, A şi B. Se poate afirma că media populaţiei A este mai mare decât media populaţiei B la un nivel de semnificaţie de 0.02? Valorile de selecţie pentru cele două populaţii se dau în tabelul de mai jos*

Selecţia	n	$\bar{x}$	s'
A	50	57.5	6.2
B	60	54.4	10.6

Soluţie.
Ambele selecţii fiind mai mari decât 30 se poate aplica testul Z.

(C)

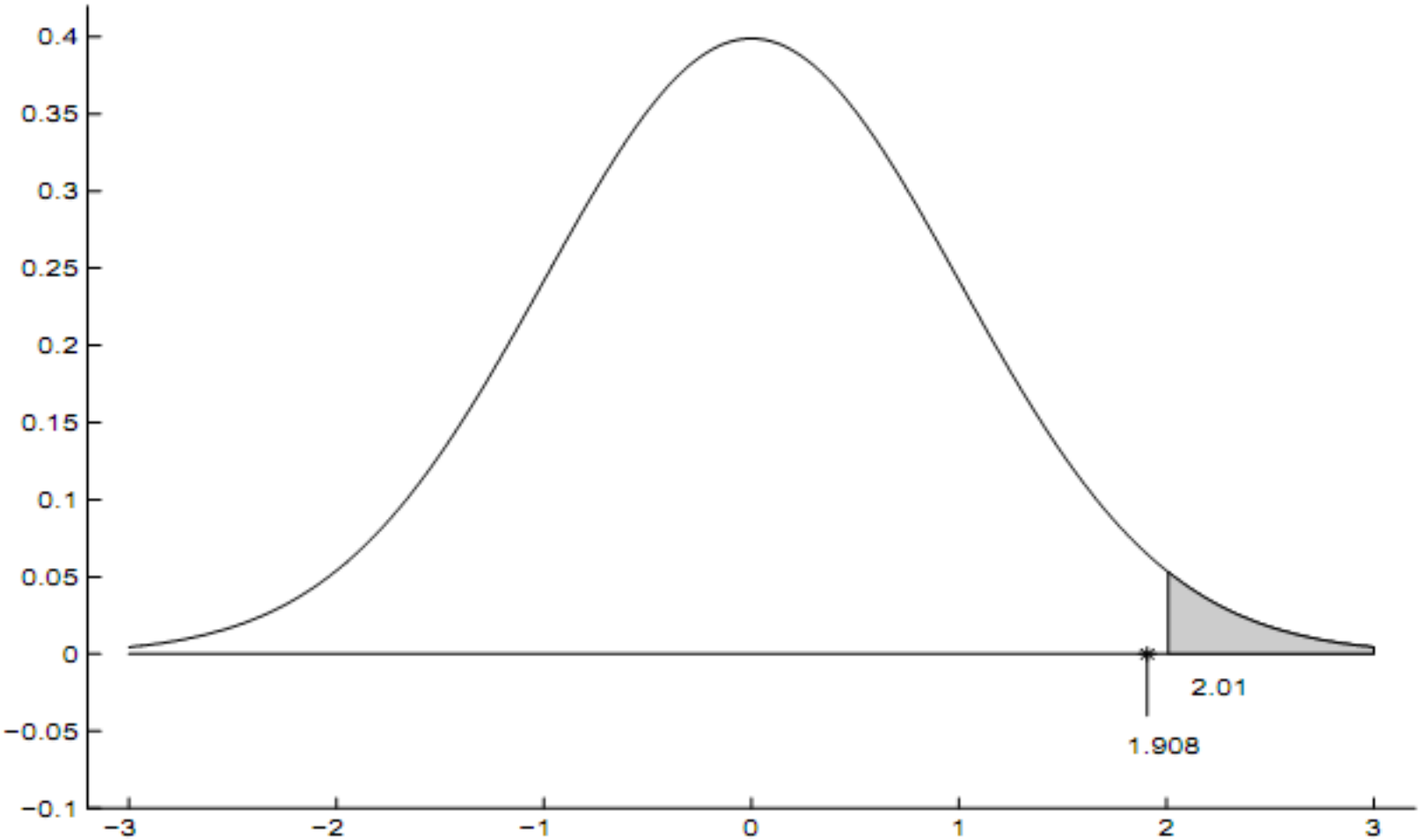
P1. $H_0 : m_A - m_B = 0 (\leq)$
P2. $H_1 : m_A - m_B > 0$
P3. Statistica testului:
$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
$\alpha = 0.02$, $z_{0.98} = 2.05$

Regiunea critică:
figura
P4. $Z = \frac{57.5 - 54.4}{\sqrt{\frac{6.2^2}{50} + \frac{10.6^2}{60}}} = 1.9074 = z^*$
P5. $z^* < z_{0.98}$, deci H_0 se acceptă.
Nu se poate afirma că $m_A > m_B$

(P)

P1. $H_0 : m_A - m_B = 0 (\leq)$
P2. $H_1 : m_A - m_B > 0$
P3. $\alpha = 0.02$
P4. $Z = \frac{57.5 - 54.4}{\sqrt{\frac{6.2^2}{50} + \frac{10.6^2}{60}}} = 1.9074 =$
$= z^*$
P5. $\Phi(1.9074) = 0.9761$
$P = P(Z > z^*) = 1 - 0.9761 =$
$= 0.0\,239$
P6. $P > 0.02$, deci H_0 se acceptă.

La nivelul de semnificaţie
$\alpha = 0.02$ mediile nu diferă.



Test Z pentru compararea a două medii

Cazul 2. Dispersii necunoscute şi selecţii mici

(2a) $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Se consideră statistica

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$ unde $$S_p = \sqrt{\frac{(n_1 - 1)s_1'^2 + (n_2 - 1)s_2'^2}{n_1 + n_2 - 2}}.$$

Statistica T urmează legea Student cu $n = n_1 + n_2 - 2$ grade de libertate, deci putem aplica testul t.

Exemplul *Datele de mai jos dau conţinutul în mg/100g de vitamina C a două sucuri de fructe:*

Sucul A (x_i)	*16*	*20*	*23*	*17*	*19*	
Sucul B (y_j)	*22*	*20*	*13*	*18*	*25*	*28*

Se poate afirma că unul dintre sucuri este mai bogat în vitamina C decât celălalt, la nivelul $\alpha = 0.05$ *?*

Soluţie. (C)

P1. $m_1 = m_2$.

P2. $m_1 < m_2$.

P3. Statistica testului este dată de formula (2) , numărul de grade de libertate este 9, iar $\alpha = 0.05$. Regiunea critică este dată în figura

$t_{9,0.05} = -2.262$.

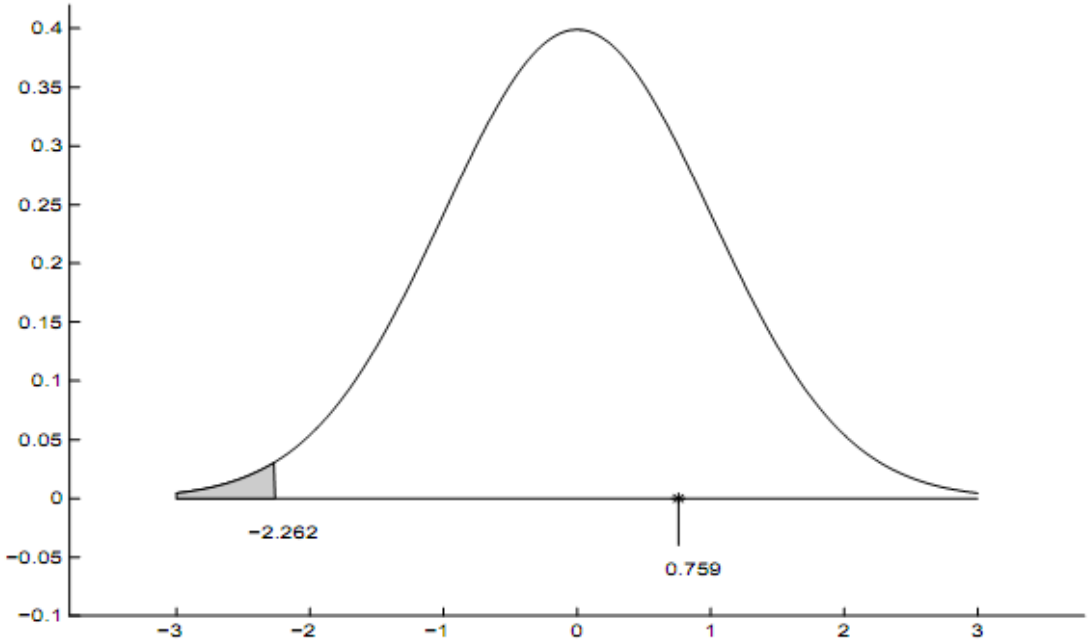


Figura 9.8: Regiune critică pentru un test t unilateral stânga

P4. Avem $\bar{x}_1 = 19$, $\bar{x}_2 = 21$, iar

$$S_p = \sqrt{\frac{\sum(x_i - \bar{x}_1)^2 + \sum(y_j - \bar{x}_2)^2}{5+6-2}} = \sqrt{\frac{30+140}{9}} = 4.3461.$$

Pentru T obţinem

$$T = \frac{19-21}{\sqrt{\frac{30+140}{9}}\sqrt{\frac{1}{5}+\frac{1}{6}}} = -0.75996 = t^*.$$

P5. Deoarece $t_{9,0.05} = -2.262 < t^*$, H_0 se acceptă, deci nu se poate afirma că unul din sucuri este mai bogat în vitamina C decât celălalt.

(2b) $\sigma_1^2 \neq \sigma_2^2$.
Statistica folosită este

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{\sqrt{\frac{s_1'^2}{n_1} + \frac{s_2'^2}{n_2}}},$$

3

care urmează legea Student cu n grade de libertate, unde n se obţine din

$$\frac{1}{n} = \frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1},$$

4

iar

$$c = \frac{\frac{s_1'^2}{n_1}}{\frac{s_1'^2}{n_1} + \frac{s_2'^2}{n_2}}.$$

5

O aproximaţie grosieră pentru n este $n = \min(n_1 - 1, n_2 - 1)$

Exemplu *Se consideră două selecţii din două populaţii normale, independente, cu dispersii diferite, care conduc la valorile*

Selecţia	n	$\bar{x}$	s'
A	*10*	*5.38*	*1.89*
B	*12*	*5.92*	*0.83*

Pentru $\alpha = 0.05$, *se poate trage concluzia că media lui A este mai mică decât media lui B?*

Soluţie. (C)

P1. $H_0 : m_A = m_B$.

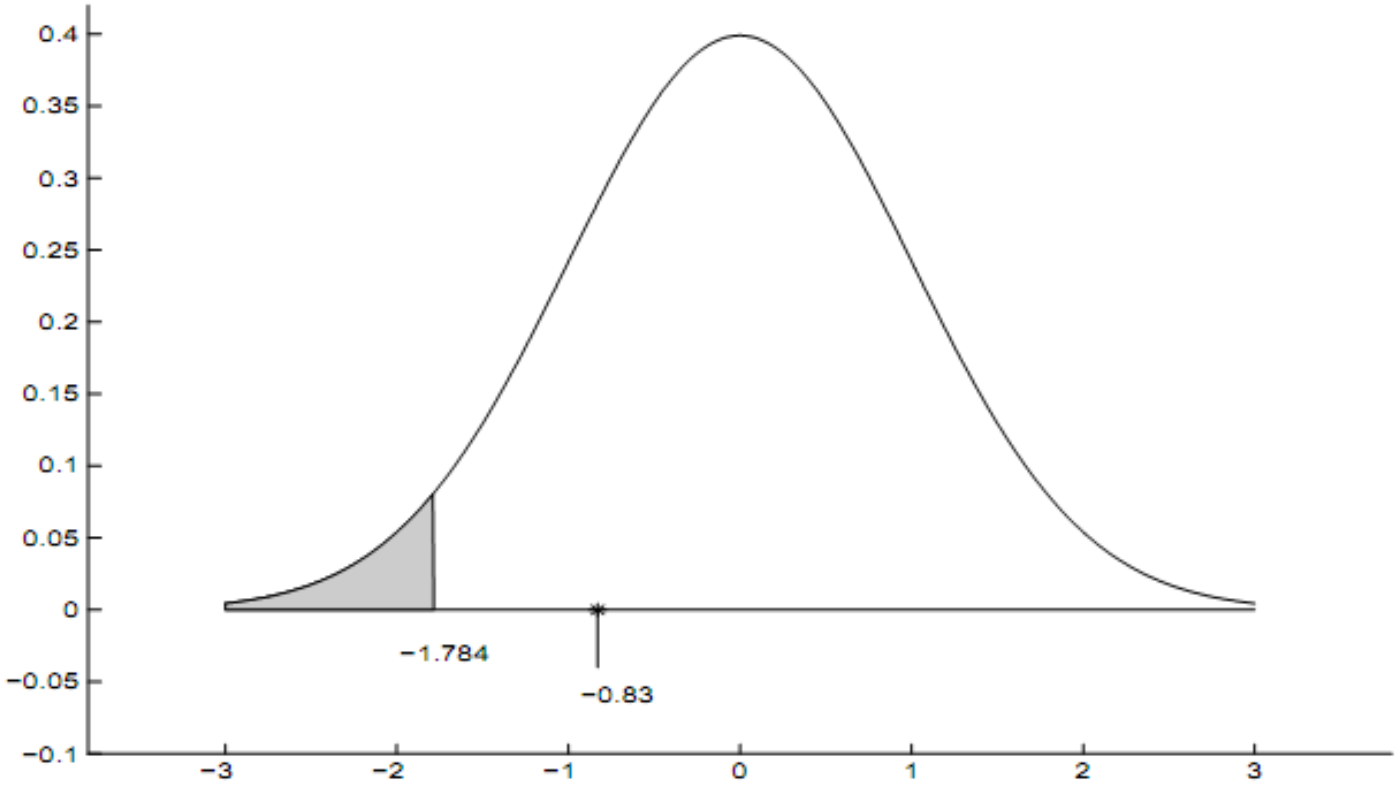
P2. $H_1 : m_A < m_B$.

P3. Statistica testului este dată de (3) $\alpha = 0.05$, iar regiunea critică apare în figura . Numărul de grade de libertate se obţine cu ajutorul formulelor (4) şi (5)

$$c = \frac{\frac{1.89^2}{9}}{\frac{1.89^2}{9} + \frac{0.83^2}{11}} = .86371$$

$$n = \frac{1}{\frac{.86371^2}{10-1} + \frac{(1-.86371)^2}{12-1}} = 11.824$$

Cuantila $t_{n,0.05} = t_{11.824,0.05} = -1.7844$, iar cu aproximarea din observaţia de mai sus $t_{n,0.05} = t_{9,0.05} = -1.83358$.



Testul t pentru compararea a două medii – dispersii diferite

P4. Statistica testului are valoarea

$$T = \frac{5.38 - 5.92}{\sqrt{\frac{1.89^2}{10} + \frac{0.83^2}{12}}} = -0.\,83863 = t^*.$$

P5. Deoarece $t_{n,0.05} < t^*$ se acceptă H_0. Nu putem afirma că media lui A este mai mică decât media lui B.

Teste pentru medii dependente (observaţii perechi)

Pentru selecţii dependente procedura de verificare este diferită de cea din cazul observaţiilor independente. Datorită faptului că datele provin din aceeaşi sursă, ele vor trebui să apară în perechi. Perechile se vor compara considerând diferenţele dintre valorile lor numerice. Utilizarea datelor în perechi are proprietatea de a înlătura anumiţi factori necontrolabili care ar putea să afecteze experimentul.

Statistica testului va fi

$$T = \frac{\bar{d} - m_d}{\frac{s_d}{\sqrt{n}}}, \qquad \text{unde } d_i = x_{i,1} - x_{i,2},$$

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i, \qquad s_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}}.$$

Ea este repartizată Student cu $n - 1$ grade de libertate.

Exemplu *Un medicament se testează pentru efectul său asupra tensiunii arteriale. Pentru 12 oameni se ia tensiunea înainte şi după consumarea medicamentului şi se găseşte că media diferenţelor este 5 şi abaterea medie pătratică este 5.83.*

a) Să se determine un interval de încredere de 99% pentru media diferenţei.

b) Folosind rezultatul de la punctul a) să se decidă dacă există o diferenţă semnificativă între folosirea şi nefolosirea medicamentului ($\alpha = 1\%$).

Soluţie. a) Avem $t_{11,0.01} = -3.106$, $t_{11,0.99} = 3.106$. Intervalul de încredere se obţine punând condiţia

$$-3.106 < T = \frac{\bar{d} - m_d}{\frac{s_d}{\sqrt{n}}} < 3.106,$$

$$5 - 3.106 \cdot \frac{5.83}{\sqrt{12}} < m_d < 5 + 3.106 \cdot \frac{5.83}{\sqrt{12}},$$

adică $m_d \in (-0.\,22732, 10.\,227)$.

b) Ipoteza nulă este $H_0 : m_d = 0$, iar ipoteza alternativă $H_1 : m_d \neq 0$. Deoarece 0 aparţine intervalului de încredere, ipoteza nulă se acceptă şi se poate trage concluzia că nu există diferenţe semnificative între folosirea şi nefolosirea medicamentului.

Teste pentru două proporţii

Adesea suntem interesaţi să facem verificări de ipoteze privind proporţii, procentaje sau probabilităţi asociate cu două populaţii. Reamintim că:

1. frecvenţa observată este $p_i' = \frac{x_i}{n_i}$, unde n_i este numărul de observaţii, iar x_i numărul de succese, $i = 1, 2$;

2. p_i este probabilitatea de succes într-un experiment binomial cu n_i probe, $i = 1, 2$.

Variabila aleatoare $p_1' - p_2'$ este aproximativ normală cu media $m = p_1 - p_2$ şi abaterea medie pătratică

$$\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}.$$

Vom folosi statistica

$$Z = \frac{(p_1' - p_2') - (p_1 - p_2)}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \tag{6}$$

unde

$$p = \frac{x_1 + x_2}{n_1 + n_2}.$$

Ipoteza nulă este $H_0 : p_1 = p_2$; sunt posibile şi ipoteze alternative bilaterale şi unilaterale.

Exemplu *Un nou tratament al unei boli este comparat cu tratamentul folosit în mod obişnuit. Materialul clinic obţinut sub un contrul atent este trecut în tabelul*

Tratamentul\Rezultatul	*Vindecat*	*Nevindecat*	*Total*
Vechi	$5(x_1)$	$8(n_1 - x_1)$	$13(n_1)$
Nou	$9(x_2)$	$3(n_2 - x_2)$	$12(n_2)$
Total	$14(x_1 + x_2)$	$11(n_1 + n_2 -$	$25(n_1 + n_2)$
		$x_1 - x_2)$	

Datele atestă superioritatea noului tratament? $(\alpha = 1\%)$

Soluţie. (C)

P1. $H_0 : p_1 = p_2$.

P2. $H_1 : p_1 \neq p_2$.

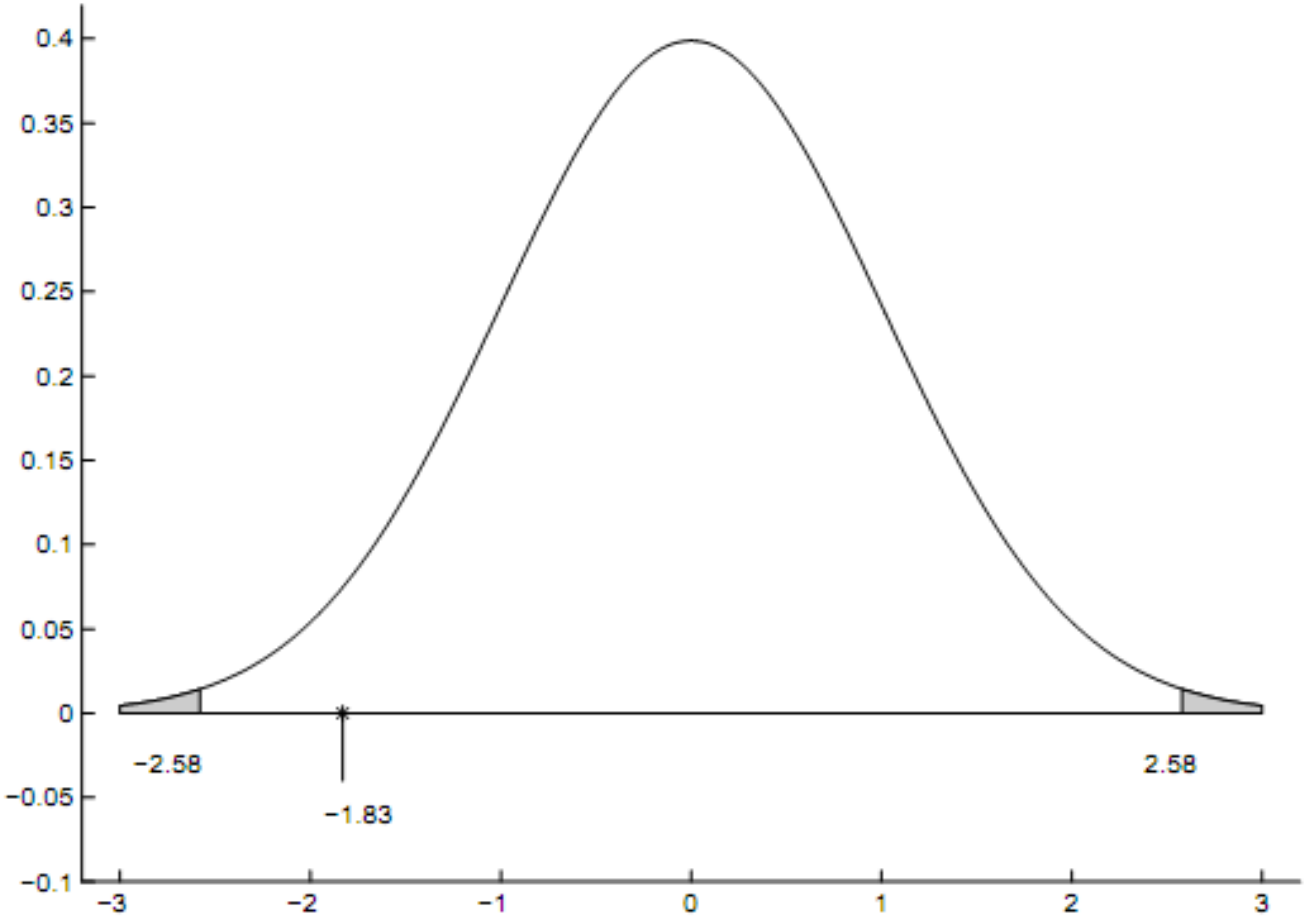
P3. Statistica testului este dată de (6), $\alpha = 0.01$, $z_\alpha = -2.58$, iar regiunea critică apare în figura

P4.

$$p_1' = \frac{x_1}{n_1} = \frac{5}{13},\ p_2' = \frac{x_2}{n_2} = \frac{9}{12},\ p = \frac{x_1 + x_2}{n_1 + n_2} = \frac{14}{25}.$$

$$Z = \frac{\frac{5}{13} - \frac{9}{12}}{\sqrt{\frac{14}{25}\frac{11}{25} \cdot \frac{25}{13 \cdot 12}}} = -1.8387 = z^*.$$

P5. Deoarece $z_\alpha = -2.58 < z^* < z_{1-\alpha} = 2.58$, ipoteza nulă se acceptă; datele nu ne permit să afirmăm superioritatea noului tratament.



Testul Z pentru două proporţii

Teste asupra dispersiilor a două populaţii

Să considerăm două populaţii independente, cercetate din punct de vedere al aceleiaşi caracteristici notate cu X_1 pentru prima populaţie şi X_2 pentru a doua populaţie şi care urmează legea $N(m_1, \sigma_1^2)$ şi respectiv $N(m_2, \sigma_2^2)$. Relativ la cele două populaţii se face ipoteza nulă

$H_0 : \sigma_1^2 = \sigma_2^2$ (sau echivalent $\frac{\sigma_1^2}{\sigma_2^2} = 1$)

cu alternativele

$H_1 : \sigma_1^2 \neq \sigma_2^2$ (sau echivalent $\frac{\sigma_1^2}{\sigma_2^2} \neq 1$, test bilateral)

$H_1 : \sigma_1^2 > \sigma_2^2$ (sau echivalent $\frac{\sigma_1^2}{\sigma_2^2} > 1$, test unilateral dreapta)

$H_1 : \sigma_1^2 < \sigma_2^2$ (sau echivalent $\frac{\sigma_1^2}{\sigma_2^2} < 1$, test unilateral stânga) .

Pentru verificarea ipotezei nule H_0 cu una din alternativele considerate se efectuează câte o selecţie repetată de volume n_1 şi n_2 din cele două populaţii. Statistica

$$F = \frac{s_1'^2}{s_2'^2} \qquad (7)$$

urmează legea F (Snedecor-Fisher) cu $n_1 - 1$ şi $n_2 - 1$ grade de libertate Testul corespunzător se numeşte **testul F** sau **testul Snedecor-Fisher**.

Exemplu *O firmă de îmbuteliat băuturi răcoritoare trebuie să decidă dacă va achiziţiona o maşină nouă de îmbuteliat sau o va folosi pe cea veche. Un criteriu de decizie este egalitatea dispersiilor corespunzătoare volumelor celor două maşini. Informaţiile de mai jos ne permit să respingem afirmaţia producătorului maşinii noi că maşina modernă are o dispersie mai mică decâ a maşinii actuale, pentru* $\alpha = 1\%$*?*

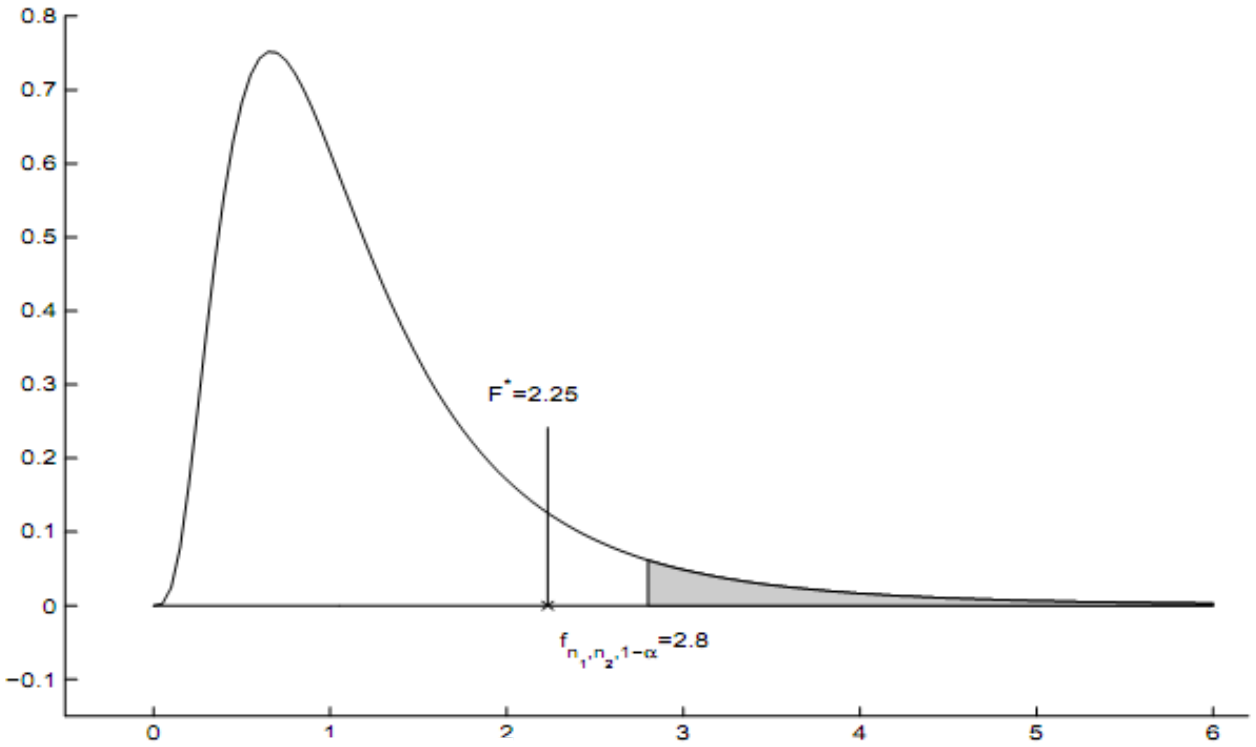
Selecţia	n	s'^2
maşina actuală, p	22	0.0008
maşina modernă, m	25	0.0018

Soluţie.

P1. $H_0 : \sigma_m^2 = \sigma_p^2$ sau $\frac{\sigma_m^2}{\sigma_p^2} = 1 (\leq)$.

P2. $H1 : \sigma_m^2 > \sigma_p^2$ sau $\frac{\sigma_m^2}{\sigma_p^2} > 1$.

P3. Statistica testului este F dată de (9.9), $n_1 = 24$, $n_2 = 21$, $\alpha = 0.01$, cuantila este $f_{24,21,0.99} = 2.80$, iar regiunea critică apare în figura



Diagramă pentru un test F

P4.

$$F = \frac{s_m'^2}{s_p'^2} = \frac{0.0018}{0.0008} = 2.25 = F^*.$$

P5. Deoarece $F^* < f_{24,21,0.99}$ ipoteza nulă se acceptă, deci respingem afirmaţia producătorului maşinii noi.

Testul χ^2

Până acum testul χ^2 a fost utilizat pentru a verifica ipoteze asupra dispersiei teoretice a unei populaţii. Distribuţia χ^2 se poate de asemenea utiliza la teste asupra experimentelor multinomiale şi tabelelor de contingenţă (analiza categorială a datelor - Categorical Data Analysis). Aceste tipuri de teste vor fi utilizate la compararea rezultatelor experimentale cu cele teoretice pentru a determina: (1) preferinţele, (2) independenţa, (3) omogenitatea. Datele care vor fi utilizate în cadrul acestor tehnici vor fi *enumerative*, adică vor rezulta din numărarea apariţiilor.

Statistica χ^2

Există multe probleme în care datele sunt grupate în clase sau categorii, iar rezultatele sunt ilustrate prin numărare. Să presupunem că avem un număr k de **clase** sau **celule,** în care sunt înregistrate n observaţii. **Frecvenţele observate** din fiecare celulă sunt notate cu $O_1, O_2, \ldots, O_k$.

Suma frecvenţelor observate este $O_1+O_2+\cdots+O_n = n$. Dorim să comparăm aceste frecvenţe cu frecvenţele teoretice notate cu $E_1, E_2, \ldots, E_k$. Suma acestor frecvenţe este $E_1 + E_2 + \cdots + E_k = n$. Pentru a decide dacă frecvenţele observate sunt în concordanţă cu cele teoretice vom utiliza distribuţia χ^2. Statistica testului va fi:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}. \qquad (8)$$

Această valoare calculată pentru X^2 va fi suma mai multor numere nenegative, câte unul pentru fiecare categorie. O valoare mică a numărătorului înseamnă o diferenţă mică între valoarea observată şi cea teoretică; pentru a face distincţie între o valoare $O_i = 110$ şi $E_i = 115$, $|O_i - E_i| = 10$ şi una obţinută din $O_i = 15$ şi $E_i = 10$, $|O_i - E_i| = 10$ se împarte la valoarea teoretică. Observaţia aceasta sugerează ideea că valori mici ale statisticii înseamnă concordanţă, iar valori mari discordanţă. Pentru selecţii repetate mari, distribuţia statisticii X^2 poate fi aproximată bine cu ajutorul distribuţiei χ^2.

Observaţia aceasta sugerează ideea că valori mici ale statisticii înseamnă concordanţă, iar valori mari discordanţă. Pentru selecţii repetate mari, distribuţia statisticii X^2 poate fi aproximată bine cu ajutorul distribuţiei χ^2. Această aproximare este considerată adecvată când toate frecvenţele teoretice sunt ≥ 5 şi $k \geq 5$. Pentru $k < 4$, E_i trebuie să fie mult mai mare decât 5. Dacă aceste condiţii nu se realizează, se poate proceda la o regrupare a datelor.

Teste privind experimentele multinomiale

Se numeşte **experiment multinomial** un experiment cu următoarele caracteristici:

1. constă din n probe identice;
2. rezultatul unui experiment cade în exact una din cele k celule sau clase posibile;
3. fiecare celulă i are asociată o probabilitate p_i care rămâne constantă pe parcursul experimentului şi $p_1 + p_2 + \cdots + p_k = 1$;
4. experimentul conduce la o mulţime de frecvenţe observate O_1, O_2, ..., O_k, unde O_i este numărul de probe al căror rezultat cade în celula i şi $O_1 + O_2 + \cdots + O_n = n$.

Ipoteza nulă are forma:

$$H_0 : p_i = p_i^{(0)}, i = \overline{1,k},$$

unde valorile $p_i^{(0)}$ sunt date, iar ipoteza alternativă

$$H_1 : \exists i_0 \in \{1, \ldots, k\}\ p_{i_0} \neq p_{i_0}^{(0)}.$$

Frecvenţa teoretică va fi $E_i = np_i$, iar numărul gradelor de libertate $k-1$. Într-adevăr, vectorul aleator $O(O_1, O_2, \ldots, O_k)$ urmează legea multinomială, adică dacă n_i sunt valorile de selecţie corespunzătoare variabilelor aleatoare O_i, avem:

$$P\left(O_1 = n_1, O_2 = n_2, \ldots, O_k = n_k\right) = \frac{n!}{n_1! n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}, \qquad (9)$$

unde $n_1 + n_2 + \ldots + n_k = n$, $n_i \in \{0, \ldots, n\}$, $i = \overline{1,k}$, $p_1 + p_2 + \cdots + p_k = 1$. Examinând ipoteza nulă observăm că ea se referă la parametri unei legi multinomiale.

Teorema 1 *Statistica*

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - np_i)^2}{np_i} \tag{10}$$

urmează legea χ^2 *cu* $k-1$ *grade de libertate, când* $n \to \infty$.

Exemplu *Să presupunem că dorim să testăm dacă un zar este perfect sau măsluit. Se aruncă zarul de mai multe ori şi dacă fiecare faţă apare cam în* $\frac{1}{6}$ *din cazuri, se poate presupune că zarul este bun. Aruncând zarul de 60 de ori se obţin frecvenţele*

Număr	1	2	3	4	5	6
Apariţii	7	12	10	12	8	11

Să se verifice dacă zarul este corect, pentru $\alpha = 5\%$.

Soluţie.

P1. Ipoteza nulă este $H_0 : p_i = \frac{1}{6}$, $i = \overline{1,6}$ sau echivalent $p_1 = \frac{1}{6} \wedge p_2 = \frac{1}{6} \wedge p_3 = \frac{1}{6} \wedge p_4 = \frac{1}{6} \wedge p_5 = \frac{1}{6} \wedge p_6 = \frac{1}{6}$.

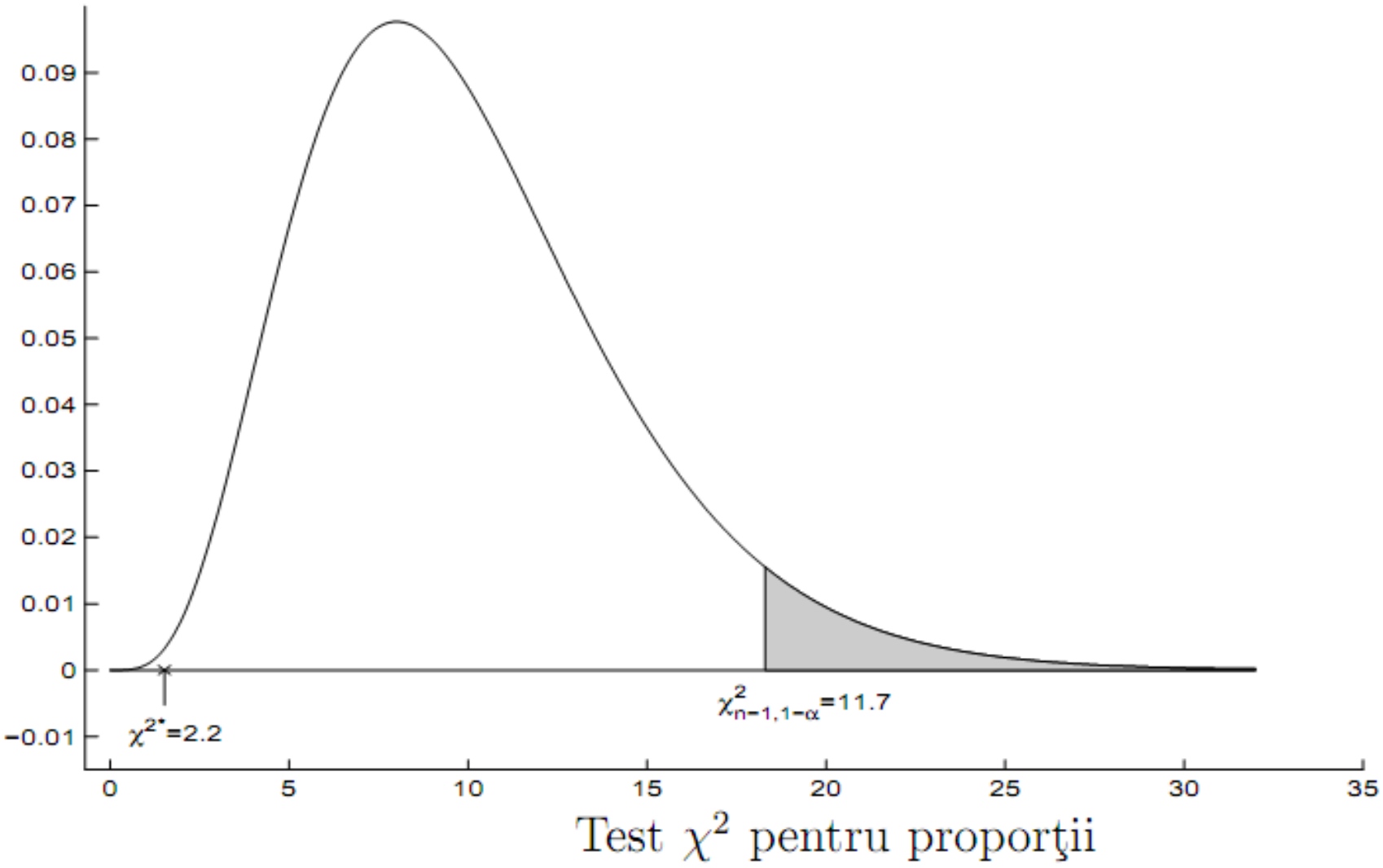
P2. Ipoteza alternativă este $H_1 : \exists j \in \{1, \ldots, 6\}\ p_j \neq \frac{1}{6}$ sau formulat echivalent $p_1 \neq \frac{1}{6} \vee\ p_2 \neq \frac{1}{6} \vee\ p_3 \neq \frac{1}{6} \vee\ p_4 \neq \frac{1}{6} \vee\ p_5 \neq \frac{1}{6} \vee\ p_6 \neq \frac{1}{6}$.

P3. Statistica testului este (8) sau echivalent (10), $\alpha = 0.05$, iar numărul de grade de libertate este $k - 1 = 5$. Regiunea critică apare în figura şi $\chi^2_{5,0.95} = 11.07$.

P4. Frecvenţele teoretice sunt $E_i = np_i = 60 \cdot \frac{1}{6} = 10$, $i = \overline{1,6}$. Valoarea statisticii este

$$X^2 = \frac{1}{10}\left((7-10)^2 + (12-10)^2 + (10-10)^2 + (12-10)^2 + \right.$$
$$\left. +(8-10)^2 + (11-10)^2\right) = 2.2. = \chi^{2*}$$

P5. Deoarece χ^{2*} nu este în regiunea critică se acceptă H_0. Concluzia: selecţia nu ne permite să afirmăm că zarul este măsluit.



Test χ^2 pentru proporţii

Tabele de contingenţă

Se pune problema dependenţei sau independenţei datelor în funcţie de doi factori (variabile).

Datele unei selecţii de volum n dintr-o populaţie cu două caracteristeci se aşează într-o tabelă numită **tabelă de corelaţie** sau **tabelă de contingenţă**. De remarcat că unele caracteristici pot fi cantitative (dimensiuni, greutăţi, etc.), iar altele calitative (piesă bună sau rebut).

Presupunem că analizăm populaţia în raport cu două caracteristici cantitative X şi Y, care iau valorile $x_1, \ldots, x_r$ şi respectiv $y_1, \ldots, y_s$. Vom nota cu n_{ij} frecvenţa absolută a cazurilor pentru care $X = x_i$, $i = \overline{1, r}$, $Y = y_j$, $j = \overline{1, s}$. Dacă n este volumul selecţiei, atunci

$$\sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} = n. \qquad (11)$$

Frecvenţa relativă notată f_{ij} sau p_{ij} se defineşte prin raportul:

$$p_{ij} = f_{ij} = \frac{n_{ij}}{n}.$$

$$\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}}{n} = \sum_{i=1}^{r}\sum_{j=1}^{s} p_{ij} = 1.$$

Numărul elementelor selecţiei pentru care $X = x_i$, indiferent de valorile pe care le ia Y îl notăm cu $n_{i.}$. Analog, frecvenţa cazurilor pentru care $Y = y_j$, indiferent de valorile lui X va fi notată cu $n_{.j}$. Avem

$$n_{i.} = \sum_{j=1}^{s} n_{ij},\ n_{.j} = \sum_{i=1}^{r} n_{ij} \qquad \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{s} n_{.j} = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = n$$

$$p_{i.} = \sum_{j=1}^{s} p_{ij},\ p_{.j} = \sum_{i=1}^{r} p_{ij} \quad \sum_{i=1}^{r} p_{i.} = \sum_{j=1}^{s} p_{.j} = 1.$$

Schematic o tabelă de corelaţie se reprezintă prin

$X\backslash Y$	x_1	$\dots$	x_i	$\dots$	x_r	
y_1	n_{11}	$\dots$	n_{i1}	$\dots$	n_{r1}	$n_{.1}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_j	n_{1j}	$\dots$	n_{ij}	$\dots$	n_{rj}	$n_{.j}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_s	n_{1s}	$\dots$	n_{is}	$\dots$	n_{rs}	$n_{.s}$
	$n_{1.}$	$\dots$	$n_{i.}$	$\dots$	$n_{r.}$	n

Frecvenţele $n_{i.}$, $n_{.j}$ se numesc frecvenţe absolute marginale, iar $p_{i.}$, $p_{.j}$ se numesc frecvenţe relative marginale (din cauză că apar pe liniile şi coloanele din marginea tabelului).

Momentele de selecţie. Putem defini momente în raport cu fiecare dintre cele două caracteristici, precum şi momente mixte. Astfel momentul de selecţie de ordinul h în raport cu X este dat de

$$m_{h0} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}x_i^h = \frac{1}{n}\sum_{i=1}^{r} n_{i.}x_i^h = \sum_{i=1}^{r} p_{i.}x_i^h,$$

iar momentul de selecţie de ordin k în raport cu Y este

$$m_{0k} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}y_j^k = \frac{1}{n}\sum_{j=1}^{s} n_{.j}y_j^k = \sum_{j=1}^{s} p_{.j}y_j^k.$$

Momentul (mixt) de selecţie de ordinul h în raport cu X şi de ordinul k în raport cu Y este

$$m_{hk} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}x_i^h y_j^k = \sum_{i=1}^{r}\sum_{j=1}^{s} p_{ij}x_i^h y_j^k.$$

În particular **mediile de selecţie** sunt

$$m_{10} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}x_i = \frac{1}{n}\sum_{i=1}^{r} n_{i.}x_i,$$

$$m_{01} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}y_j = \frac{1}{n}\sum_{j=1}^{s} n_{.j}y_j.$$

Momentele centrate se definesc asemănător cu cele din cazul unidimensional. Astfel, momentul centrat de selecţie de ordinul h în raport cu X este

$$\bar{m}_{h0} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(x_i - m_{10})^h = \frac{1}{n}\sum_{i=1}^{r} n_{i.}(x_i - m_{10})^h,$$

iar momentul centrat de ordinul k în raport cu Y este

$$\bar{m}_{0k} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(y_j - m_{01})^k = \frac{1}{n}\sum_{j=1}^{s} n_{.j}(y_j - m_{01})^k.$$

Momentul centrat (mixt) de selecţie de ordinul h în raport cu X şi de ordinul k în raport cu Y este

$$\begin{aligned}\bar{m}_{hk} &= \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(x_i - m_{10})^h(y_j - m_{01})^k = \\ &= \sum_{i=1}^{r}\sum_{j=1}^{s} p_{ij}(x_i - m_{10})^h(y_j - m_{01})^k.\end{aligned}$$

Momentele centrate de ordinul al doilea

$$\bar{m}_{20} = s_1^2 = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(x_i - m_{10})^2 = \frac{1}{n}\sum_{i=1}^{r} n_{i.}(x_i - m_{10})^2$$

$$\bar{m}_{02} = s_2^2 = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(y_j - m_{01})^2 = \frac{1}{n}\sum_{j=1}^{s} n_{.j}(y_j - m_{01})^2$$

se numesc **dispersiile de selecţie** ale caracteristicii X şi respectiv Y. Au loc relaţiile

$$\bar{m}_{20} = s_1^2 = m_{20} - m_{10}^2$$
$$\bar{m}_{02} = s_2^2 = m_{02} - m_{01}^2.$$

Momentul centrat mixt de ordinul al doilea

$$\bar{m}_{11} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(x_i - m_{10})(y_j - m_{01}) = m_{11} - m_{10}m_{01}$$

se numeşte ***corelaţia*** *(****covarianţa****) lui* X *şi* Y*, iar raportul*

$$\bar{r} = \bar{r}(X,Y) = \frac{\bar{m}_{11}}{\sqrt{\bar{m}_{20}\bar{m}_{02}}} = \frac{\bar{m}_{11}}{s_1 s_2}$$

se numeşte ***coeficient de corelaţie statistic*** *sau* ***empiric****.*

Proprietăţi.

Dacă X *şi* Y *sunt independente, atunci* $\bar{r}(X,Y) = 0$*.*

$\bar{r}(X,Y) = 1$ *dacă şi numai dacă între* X *şi* Y *există o legătură liniară.*

Coeficientul de corelaţie al lui Pearson poate exprima o legătură mai generală decât coeficientul de corelaţie statistic.

Se numeşte coeficient de corelaţie al lui Pearson numărul

$$\rho^2 = \frac{1}{\sqrt{(r-1)(s-1)}} \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}}.$$

Proprietăţi.

X şi Y sunt independente dacă şi numai dacă $\rho^2 = 0$.

Are loc inegalitatea $0 \leq \rho^2 \leq 1$.

Dacă între X şi Y există o dependenţă funcţională de forma $Y = h(X)$, atunci $\rho^2 = 1$.

Testarea independenţei. În general, pentru a testa independenţa factorului de pe linie de cel de pe coloană se utilizează o tabelă de contingenţă $r \times s$ (r este numărul de linii, iar s numărul de coloane). Numărul de grade de libertate se va determina cu formula

$$gl = (r-1)(s-1).$$

Justificare: ţinând cont că avem în total rs căsuţe în tabel, iar totalurile pe linii şi coloane sunt fixate şi suma totalurilor pe linii şi coloane este n, ne rămân posibilităţile fixate de formulă. Formula de mai sus este valabilă numai dacă $r \neq 1$ şi $s \neq 1$. Dacă $r = 1$, $gl = s - 1$, iar dacă $s = 1$, atunci $gl = r - 1$. Deci

$$gl = \begin{cases} (r-1)(s-1) & \text{dacă} \quad r \neq 1 \wedge s \neq 1, \\ s-1 & \text{dacă} \quad r = 1, \\ r-1 & \text{dacă} \quad s = 1. \end{cases}$$

Folosind notaţiile (de mai sus) referitoare la selecţii bidimensionale, ipoteza nulă se scrie:

$H_0 : p_{ij} = p_{i.}p_{.j},\ i = \overline{1, r},\ j = \overline{1, s}.$

Frecvenţele teoretice pentru o tabelă de contingenţă $r \times s$ sunt date de

$$E_{ij} = \frac{n_{i.}n_{.j}}{n}, \tag{12}$$

unde n este volumul selecţiei. Justificare: totalurile marginale de pe linii trebuie distribuite proporţional cu cele de pe coloană. În tabela finală în căsuţe se trec în paranteză sau mai jos şi frecvenţele teoretice.

Statistica testului are expresia:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}. \tag{13}$$

Exemplu *(Clasificarea studenţilor după sex şi disciplina preferată). Fiecărui student dintr-un grup de 300 de studenţi i se identifică sexul şi apoi este întrebat dacă preferă discipline din sfera ştiinţelor naturii (SN), ştiinţelor sociale(SS) sau umaniste (SU). Tabela de mai jos ne dă frecvenţele determinate pentru aceste categorii. Ne permite această selecţie să respingem ipoteza ,,preferinţa pentru SN, SS sau SU este independentă de sex" la nivelul de semnificaţie* $\alpha = 5\%$*?*

	Disciplina favorită			
Sex	SN	SS	SU	Total
M	37	41	44	122
F	35	72	71	178
Total	72	113	115	300

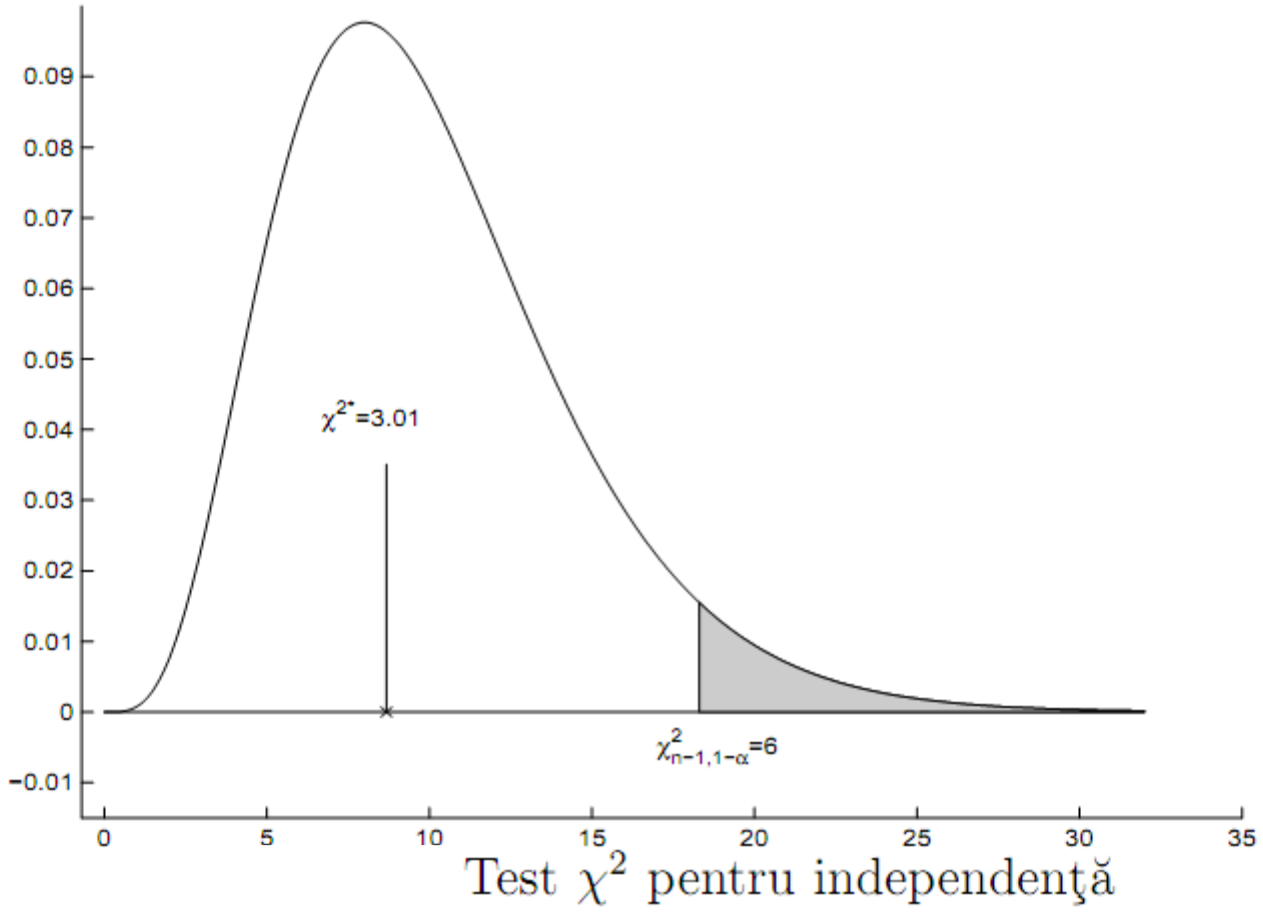
Soluţie.

P1. $H_0 : p_{ij} = p_{i.}p_{.j},\ i = \overline{1,2},\ j = \overline{1,3}$ (preferinţa pentru SN, SU, SS este independentă de sex).

P2. Preferinţa nu este independentă de sex, adică

$H_1 : \exists i_0 \in \{1,2\}\ \exists j_0 \in \{1,2,3\}\ p_{i_0 j_0} \neq p_{i_0 .} p_{.j_0}$.

P3. Statistica testului este dată de (13), $gl = 2$, $\alpha = 0.05$, $\chi^2_{2,0.95} = 6.00$. Regiunea critică apare în figura



P4.Vom determina frecvenţele teoretice, folosind formula (12)

$$E_{11} = \frac{72}{300} \cdot 122 = 29.28, \quad E_{12} = \frac{113}{300} \cdot 122 = 45.95,$$
$$E_{13} = \frac{115}{300} \cdot 122 = 46.77, \quad E_{21} = \frac{72}{300} \cdot 178 = 42.72,$$
$$E_{22} = \frac{113}{300} \cdot 178 = 67.05, \quad E_{23} = \frac{115}{300} \cdot 178 = 68.23.$$

Tabela de contingenţă completă este următoarea:

	Disciplina favorită			
Sex	SN	SS	SU	Total
M	37 (29.28)	41 (45.95)	44 (46.77)	122
F	35 (42.72)	72 (67.05)	71 (68.23)	178
Total	72	113	115	300

Valoarea statisticii este

$$X^2 = \frac{(37-29.28)^2}{29.28} + \frac{(41-45.95)^2}{45.95} + \frac{(44-46.67)^2}{46.67} + \frac{(35-42.72)}{42.72} + \\ + \frac{(72-67.05)^2}{67.05} + \frac{(71-68.23)^2}{68.23} \\ = 3.0186 = \chi^{2*}$$

P5. Deoarece χ^{2*} nu este în regiunea critică nu se poate respinge H_0, deci preferinţele sunt independente de sex.

Testarea omogenităţii. Aceste teste se utilizează când una din cele două variabile este controlată de experimentator, astfel încât totalurile marginale pe linii sau pe coloane să aibă valori predeterminate.

De exemplu să presupunem că vrem să sondăm preferinţele asupra unui proiect de lege propus de guvern. În cadrul sondajului se selectează aleator 200 de persoane din mediul urban, 200 din suburbii şi 100 din mediul rural şi acestea sunt chestionate asupra părerii faţă de iniţiativa guvernului.

Tip de reşedinţă	Propunerea guvernului		
	Da	Nu	Total
Urban	143	57	200
Suburban	98	102	200
Rural	13	87	100
Total	254	246	500

Ne permite selecţia să afirmăm că persoanele din diverse medii au opinii diferite asupra propunerii guvernamentale ($\alpha = 5\%$)?

Soluţie.

P1. $H_0 : p_{ij} = p_j,\ i = \overline{1,3},\ j = 1,2$ (proporţia în interiorul celor trei grupuri este aceeaşi, adică $p_{urban,da} = p_{da}$, $p_{suburban,da} = p_{da}$, $p_{rural,da} = p_{da}$, etc...)

P2. H_1 : Proporţia nu este aceeaşi în interiorul grupurilor (există măcar unul pentru care proporţia diferă de a celorlalte).

P3. Statistica testului este dată de (13), $gl = 2$, $\alpha = 0.05$, $\chi^2_{2,0.95} = 6.00$. Regiunea critică apare în figura

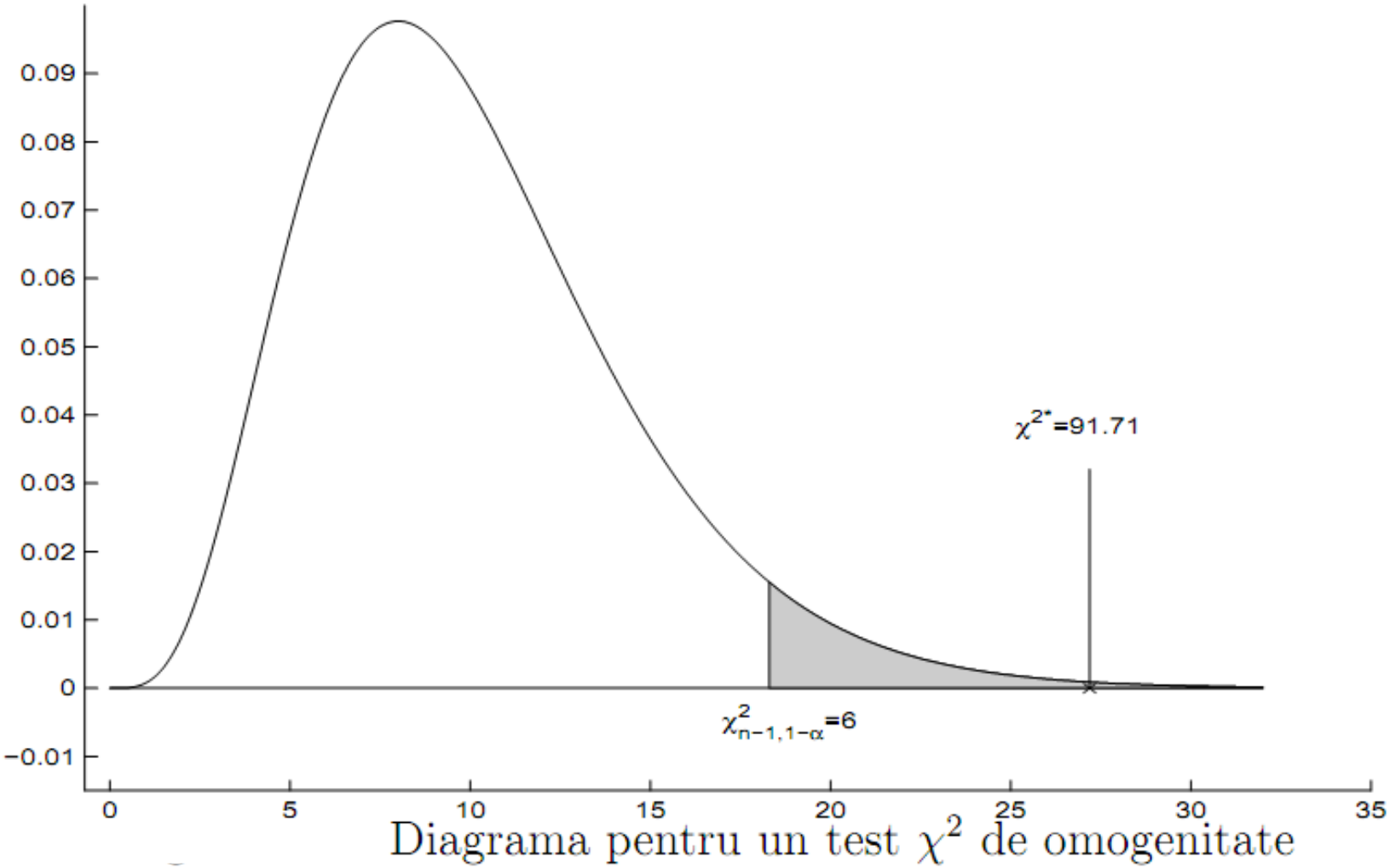


Diagrama pentru un test χ^2 de omogenitate

P4. După calculul frecvenţelor teoretice obţinem tabela

Tip de reşedinţă	Propunerea guvernului		
	Da	Nu	Total
Urban	143 (101.6)	57 (98.4)	200
Suburban	98 (101.6)	102 (98.4)	200
Rural	13 (50.8)	87 (49.2)	100
Total	254	246	500

Valoarea statisticii este

$$X^2 = \frac{(143-101.6)^2}{101.6} + \frac{(57-98.4)^2}{98.4} + \frac{(98-101.6)^2}{101.6} + \frac{(102-98.4)^2}{98.4} + \frac{(13-50.8)^2}{50.8} + \frac{(87-49.2)^2}{49.2} = 91.715 = \chi^{2*}$$

P5. H_0 se respinge. Concluzia: cele trei grupuri de oameni nu au aceeaşi proporţie de oameni favorabili iniţiativei legislative.

Teste de concordanţă

Fie caracteristica X ce urmează o lege de probabilitate cu funcţia de repartiţie F necunoscută. Dorim să verificăm ipoteza nulă $H_0 : F = F_0$ în raport cu alternativa $H_1 : F \neq F_0$, unde F_0 este dată.

Testul χ^2 de concordanţă

Dacă domeniul valorilor caracteristicii X este intervalul $[a, b]$, se consideră clasele precizate prin diviziunea:

$$a = a_0 < a_1 < a_2 < \ldots < a_k = b.$$

Fie E_i evenimentul „$X \in [a_{i-1}, a_i]$“, $i = \overline{1, k}$. Avem

$$p_i = P(E_i) = P(a_{i-1} \leq X < a_i) = F(a_i) - F(a_{i-1}).$$

Aceste probabilităţi sunt necunoscute. Fie $p_i^{(0)} = F_0(a_i) - F_0(a_{i-1})$. În acest fel se ajunge la acelaşi tip de ipoteze ca la testul χ^2 pentru proporţii. Pentru verificarea ipotezei se consideră o selecţie repetată de volum n, cu valorile $x_1, x_2, \ldots, x_n$. Se notează cu n_i frecvenţa clasei $[a_{i-1}, a_i)$. Suma acestor frecvenţe este $n_1 + n_2 + \cdots + n_k = n$. Statistica testului este

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i^{(0)}\right)^2}{np_i^{(0)}},$$

care este repartizată χ^2 cu $k-1$ grade de libertate. Testul se mai numeşte **testul χ^2 neparametric**.

Exemplu *Rezultatele măsuratorilor diametrului X pentru 1000 de piese de acelaşi tip în mm sunt cele ce urmează:*

Diametru	*frecvenţa*
$97.75 - 98.25$	21
$98.25 - 98.75$	47
$98.75 - 99.25$	87
$99.25 - 99.75$	158
$99.75 - 100.25$	181
$100.25 - 100.75$	201
$100.75 - 101.25$	142
$101.25 - 101.75$	97
$101.75 - 102.25$	41
$102.25 - 102.75$	25

Folosind nivelul de semnificaţie $\alpha = 0.05$ şi cunoscând media $m = M(X) = 100.25mm$ şi abaterea medie pătratică $\sigma = \sqrt{D^2(X)} = 1mm$ se cere verificarea normalităţii caracteristicii X

$$H0 : F(x) = F_0(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}} dt, \ x \in \mathbb{R}, \qquad \text{cu } m = 100.25, \ \sigma = 1.$$

Calculăm probabilităţile

$$p_i^{(0)} = P\left(a_{i-1} \leq X < a_i | F = F_0\right) = \Phi(a_i - m) - \Phi(a_{i-1} - m), \ i = \overline{1, 10},$$

unde $a_0 = -\infty$, $a_{10} = +\infty$. Obţinem

X	$p_i^{(0)}$
$(-\infty, a_1)$	0.0228
$[a_1, a_2)$	0.0440
$[a_2, a_3)$	0.0919
$[a_3, a_4)$	0.1498
$[a_4, a_5)$	0.1915
$[a_5, a_6)$	0.1915
$[a_6, a_7)$	0.1498
$[a_7, a_8)$	0.0918
$[a_8, a_9)$	0.0440
$[a_9, a_{10})$	0.0228

$$\chi^2 = \sum_{i=1}^{10} \frac{\left(n_i - np_i^{(0)}\right)^2}{np_i^{(0)}} =$$

$$= \frac{(21 - 22.8)^2}{22.8} + \cdots + \frac{(25 - 22.8)^2}{22.8} = 3.21 = \chi^{2*}.$$

Cuantila corespunzătoare este $\chi^2_{k-1,1-\alpha} = \chi^2_{9,0.95} = 16.9$ şi cum $\chi^{2*} = 3.21 < 16.9 = \chi^2_{9,0.95}$, ipoteza nulă se acceptă.