Teoria estimaţiei. Teste statistice

1. Introducere (Metode statistice, Radu Trîmbiţaş)

Statistica

- metoda de a colecta şi afişa volume mari de informaţie numerică
- modalitatea de a lua decizii în condiţii de incertitudine

Statistica

- descriptivă (colectarea, prezentarea şi descrierea datelor)
- inferenţială (sau analitică)- se referă la tehnica de interpretare a valorilor rezultate din tehnicile descriptive şi apoi utilizarea lor la luarea deciziilor.

Are drept rol

- estimarea parametrilor unei populații
- testarea ipotezelor de cercetare

	Indicatori	**Parametri**
Media	m	μ
Dispersia	s^2	σ^2
Abaterea standard	s	σ
Coeficientul de corelație Pearson	r	ρ
Coeficientul de corelație Spearman	r_s	ρ_s

Fiecare indicator ”reprezintă” (estimează) parametrul corespunzător la nivelul populației

Terminologie

1. **Populaţia** – o mulţime de indivizi, obiecte sau măsurători ale căror proprietăţi urmează a fi analizate. Pentru a forma o populaţie o mulţime de elemente trebuie să aibă o caracteristică comună. Conceptul de populaţie este una dintre noţiunile fundamentale ale statisticii. Populaţia în cauză trebuie să fie foarte atent definită şi este considerată complet definită numai atunci când se poate da lista tuturor elementelor ei. Mulţimea studenţilor unei universităţi este un exemplu de populaţie bine definită. În mod tipic gândim o populaţie ca o colecţie de oameni. Totuşi în statistică populaţia poate fi o colecţie de animale, de obiecte manufacturate sau de măsurători. De exemplu, mulţimea valorilor numerice care sunt înălţimi ale plopilor din judeţul Cluj constituie o populaţie. Un element al unei populaţii se numeşte **individ**.

2. **Eşantion** sau **selecţie** – o submulţime a unei populaţii. O selecţie trebuie să îndeplinească următoarele condiţii:

   (a) să fie *aleatoare* (orice selecţie să aibă şansa de a fi aleasă – şansa poate fi calculată);

(b) toate elementele colectivităţii să aibă aceiaşi probabilitate de a fi alese;

(c) structura selecţiei să fie cât mai apropiată de structura populaţiei, adică selecţia să fie *reprezentativă*;

(d) volumul selecţiei să fie suficient de mare.

3. **Variabilă** – o caracteristică cantitativă de interes a fiecărui element al unei populaţii sau selecţii. Ca exemple, am putea da vârsta unui student la intrarea în facultate, înălţimea ş.a.m.d. Variabilele pot fi discrete sau continue.

4. **Atribut** - o caracteristică calitativă de interes a fiecărui element al unei populaţii sau selecţii. Culoarea părului sau a ochilor studenţilor de la o facultate, calitatea unor piese de a fi corespunzătoare sau necorespunzătoare sunt exemple de atribute.

5. **Dată** – valoarea unei variabile asociate cu un element al unei populaţii sau selecţii.

6. **Date** – mulţimea valorilor colectate ale unei variabile pentru fiecare element din selecţie. Exemplu: mulţimea înălţimilor fiecăruia din cei 25 de studenţi ai unei grupe de 25 de studenţi.

7. **Experiment** – o activitate planificată al cărei rezultat este o mulţime de date.

8. **Parametru** – o caracteristică numerică a unei întregi populaţii. Vârsta medie la admitere a studenţilor sau proporţia celor peste 21 de ani dintre cei admişi sunt exemple de parametri ai unei populaţii. Un parametru este o valoare ce descrie întreaga populaţie.

9. **Statistică** – o caracteristică numerică a unei selecţii.

Măsurabilitate şi variabilitate. Într-o mulţime de date experimentale ne aşteptăm întotdeauna să apară variaţii. Dacă apar foarte puţine variaţii sau deloc, ne gândim că dispozitivul de măsurare este defect sau insuficient de precis. Dacă luăm o cutie de carton cu table de ciocolată de 100 de grame şi cântărim fiecare ciocolată, constatăm o abatere de, să zicem, ± 2 grame. Greutatea (masa) unei table de ciocolată va fi o variabilă. Nu contează ce este sau ce reprezintă variabila; va fi variabilitate dacă instrumentele de măsură sunt suficient de precise. Un obiectiv primar în analiza statistică va fi acela al măsurării variabilităţii.

Comparaţie între Calculul probabilităţilor şi Statistică. Calculul probabilităţilor şi Statistica sunt două domenii separate ale matematicii, dar strâns înrudite. Calculul probabilităţilor este vehiculul statisticii, căci fără legi de probabilitate statistica nu ar fi posibilă.

Urna probabilistică	Urna statistică
5 albe, 5 roşii, 5 albastre	???

Să ilustrăm relaţia dintre cele două ramuri ale matematicii printr-un exemplu. Avem două urne (una probabilistică şi una statistică, vezi figura 6.1). Urna probabilistică conţine 5 bile albastre, 5 roşii şi 5 albe. Subiectul Calculul probabilităţilor încearcă să răspundă la întrebări de genul: dacă se extrage o bilă sau mai multe din urnă, care este probabilitatea să avem o anumită configuraţie de culori? Pe de altă parte urna statistică are o configuraţie necunoscută. Extragem o selecţie de bile şi facem afirmaţii despre ceea ce credem că ar fi în urnă. Observaţi diferenţa: calculul probabilităţilor calculează şansa ca ceva (o selecţie) să se întâmple când se cunoaşte populaţia. Statistica cere să se extragă o selecţie, descrie selecţia (statistică descriptivă) şi apoi face inferenţe asupra populaţiei bazându-se pe informaţia găsită în selecţie (statistica inferenţială).

Selecţie

Fie datele de selecţie $x_1, x_2, \ldots, x_n$. Numărul n se va numi volumul selecţiei.

Datele de selecţie vor fi considerate valori ale unor variabile aleatoare $X_1, \ldots, X_n$, numite variabile de selecţie; în cazul unei selecţii repetate ele sunt identic repartizate cu caracteristica de studiat X.

Variabila aleatoare

$$Z_n = h_n(X_1, X_2, \ldots, X_n),$$

unde $h_n : \mathbb{R}^n \to \mathbb{R}$ *este măsurabilă se va numi* funcţie de selecţie *sau* statistică, *iar* $z_n = h_n(x_1, \ldots, x_n)$ *se va numi* valoarea funcţiei de selecţie.

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

se va numi medie de selecţie, *iar valoarea ei* valoarea mediei de selecţie.

Dacă X *are media* $m = M(X)$ *şi dispersia* $\sigma^2 = D^2(X)$ *atunci*

$$M(\overline{X}) = m \quad \text{şi} \quad D^2(\overline{X}) = \frac{\sigma^2}{n}.$$

Dacă $m = M(X)$ *şi* $\sigma^2 = D^2(X)$ *atunci*

$$Z_n = \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}$$

converge în repartiţie către legea normală $N(0,1)$ *când* $n \to \infty$, *iar când* $X \in N(n, \sigma)$ *afirmaţia are loc pentru orice* n.

Numim moment de selecţie *de ordinul k funcţia de selecţie*

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad \textit{iar} \quad m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

se numeşte valoarea momentului de selecţie.

$$m_1 = \overline{X}.$$

Fie caracteristica X pentru care există momentul teoretic de ordinul $2k$, $M_{2k} = M(X^{2k})$. Atunci

$$M(m_k) = M_k, \quad D^2(m_k) = \frac{1}{n}(M_{2k} - M_k^2),$$

iar pentru $n \to \infty$,

$$Z_n = \frac{m_k - M_k}{\sqrt{\frac{M_{2k} - M_k}{n}}}$$

este asimptotic $N(0,1)$.

Numim moment centrat de selecţie *de ordin* k *statistica*

$$\overline{m}_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^n.$$

$\overline{m}_1 = 0$, *iar* $\overline{m}_2 = m_2 - m_1^2$.

Fie X *o caracteristică pentru care există momentul teoretic* $\overline{M}_4$. *Atunci pentru momentul centrat de ordinul al doilea avem*

$$M(\overline{m}_2) = \frac{n-1}{n}\sigma^2,$$
$$D^2(\overline{m}_2) = \frac{n-1}{n^3}[(n-1)\overline{M}_4 - (n-3)\sigma^4]$$

unde $\sigma^2 = D^2(X)$ *şi*

$$Cov(\overline{X}, \overline{m}_2) = \frac{n-1}{n^2}\overline{M}_3.$$

Estimaţia-metoda intervalelor de încredere

A da un **interval de încredere** pentru parametrul unidimensional θ cu **coeficientul de încredere** $1-\alpha$ revine la construirea pe baza unei selecţii $x_1, \ldots, x_n$ a unui interval

$$[\underline{\theta}(x_1, \ldots, x_n), \overline{\theta}(x_1, \ldots, x_n)]$$

cu proprietăţile

(i) $\underline{\theta}(x_1, \ldots, x_n) \leq \overline{\theta}(x_1, \ldots, x_n)$;

(ii) $P(\underline{\theta}(x_1, \ldots, x_n) \leq \theta \leq \overline{\theta}(x_1, \ldots, x_n)) = 1 - \alpha$.

Pentru determinarea statisticilor $\underline{\theta}$ şi $\bar{\theta}$ se caută o statistică $Z_n = Z(X_1, \ldots, X_n)$ care urmează o lege de probabilitate cunoscută (independentă de θ), dar în a cărei expresie intervine parametrul necunoscut θ. Pentru $\alpha \in [0, 1]$, mic, se determină un interval numeric (z_1, z_2) astfel încât $P(Z_n \in (z_1, z_2)) = 1 - \alpha$. De aici, prin operaţii algebrice, se obţine o relaţie de tipul celei din condiţia (ii) de mai sus. Cu cât intervalul $(\underline{\theta}, \bar{\theta})$ este mai mic, cu atât estimaţia este mai bună.

Intervale de încredere pentru medie

Dacă caracteristica X urmează legea normală $N(m, \sigma^2)$, cu $m \in \mathbb{R}$ necunoscut şi $\sigma^2 > 0$ cunoscut, atunci statistica

$$Z = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}}$$

urmează legea normală standard. Fie z_α cuantila de ordin α a repartiţiei normale reduse. Deoarece

$$P(z_{\alpha/2} < Z < z_{1-\alpha/2}) = 1 - \alpha$$

(vezi figura 8.1), se obţine pentru m următorul interval de încredere $100(1-\alpha)\%$:

$$\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < m < \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

sau ţinând cont că $z_{\alpha/2} = -z_{1-\alpha/2}$, putem spune că intervalul are forma $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

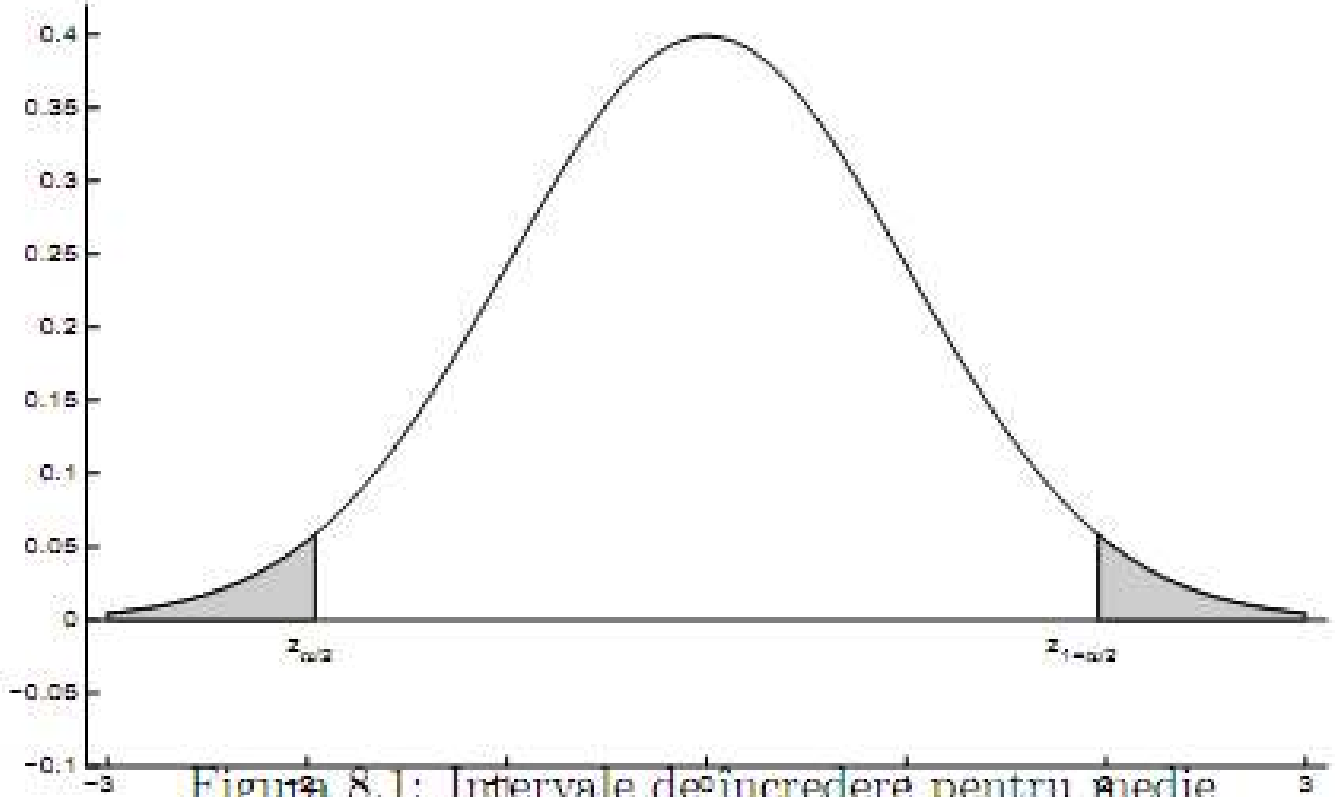


Figura 8.1: Intervale de încredere pentru medie

Exemplu *Să presupunem că avem* $n = 25$ *de date,* $\bar{x} = 20$ *şi că* $\sigma = 5$. *Să se determine un interval de încredere de 95% pentru medie.*

Soluţie. Deoarece $\alpha = 5\%$, avem $z_{\alpha/2} = -1.96$, şi se obţine

$$\bar{x} - 1.96 \cdot \frac{5}{\sqrt{25}} < m < \bar{x} + 1.96 \cdot \frac{5}{\sqrt{25}},$$

adică $m \in (18.04, 21.96)$. Rezultatul obţinut se poate interpreta astfel: în 95% din cazuri intervalul $(18.04, 21.96)$ va acoperi media m, sau probabilitatea ca m să cadă în intervalul $(18.04, 21.96)$ este de 95%.

Dacă σ este necunoscut. vom înlocui σ cu estimatia absolut corectă a sa s', dată de formula

$$s'^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = s^2 \frac{n}{n-1}$$

$$T = \frac{\bar{X} - m}{\frac{s'}{\sqrt{n}}}$$

urmează legea Student cu $n-1$ grade de libertate $T(n-1)$. Dacă $t_{n,\alpha}$ este cuantila de ordinul α a repartiţiei $T(n)$, raţionând ca în cazul precedent se obţine un interval de încredere pentru medie de forma

$$\bar{x} + t_{n-1,\alpha/2}\frac{s'}{\sqrt{n}} < m < \bar{x} + t_{n-1,1-\alpha/2}\frac{s'}{\sqrt{n}}.$$

Intervale de încredere pentru diferenţa a două medii

Fie două populaţii cu caracteristicile $X_1 \in N(m_1, \sigma_1^2)$ şi $X_2 \in N(m_2, \sigma_2^2)$. Se consideră două selecţii repetate de volume n_1 şi respectiv n_2. Mediile şi dispersiile lor de selecţie sunt

$$\bar{X}_2 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i}, \quad \bar{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i}$$

şi

$$s_1'^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(X_{1i} - \bar{X}_1\right)^2, \quad s_2'^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(X_{2i} - \bar{X}_1\right)^2.$$

(a) Dacă σ_1 şi σ_2 sunt cunoscuţi, atunci statistica

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

urmează legea normală standard. Obţinem următorul interval de încredere pentru diferenţa $m_1 - m_2$ a mediilor

$$\bar{X}_1 - \bar{X}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < m_1 - m_2 < \bar{X}_1 - \bar{X}_2 + z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

(b) Dacă σ_1 şi σ_2 sunt necunoscuţi şi $\sigma_1 = \sigma_2 = \sigma$, atunci σ poate fi înlocuit prin

$$S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}},$$

iar statistica

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (m_1 - m_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

este repartizată Student cu $n_1 + n_2 - 2$ grade de libertate. Expresia intervalului de încredere este analoagă celei de mai sus, cu precizarea că în locul cuantilelor legii normale standard se iau cuantilele repartiţiei $T(n_1 + n_2 - 2)$.

Se compară două procedee de montaj pentru un dispozitiv, unul clasic şi unul nou, care necesită pentru aplicarea corectă o perioadă de instruire de o lună şi respectiv 3 săptămâni. Au fost instruite două grupuri de câte 9 muncitori, unul cu metoda clasică şi celălat cu metoda nouă. S-a înregistrat timpul de montaj (în minute) pentru fiecare muncitor, obţinânduse rezultatele din tabela de mai jos:

Procedura	*Timpul*								
Clasică	32	37	35	28	41	44	35	31	34
Nouă	35	31	29	25	34	40	27	32	31

Determinaţi un interval de încredere de 95% pentru diferenţa mediilor în ipoteza că timpii au distribuţia normală şi dispersiile sunt egale.

Soluţie. Pentru datele din tabelul de mai sus avem

$$\begin{array}{ll} \bar{x}_1 = 35.22 & \bar{x}_2 = 31.56 \\ \sum_{i=1}^{9}(x_{1i} - \bar{x}_1)^2 = 195.56 & \sum_{i=1}^{9}(x_{2i} - \bar{x}_2)^2 = 160.22. \end{array}$$

Deci

$$S_p = \sqrt{\frac{195.56 + 160.22}{9 + 9 - 2}} = 4.7155.$$

Cum numărul de grade de libertate este $n_1 + n_2 - 2 = 16$ şi $t_{16,0.975} = 2.120$ se obţine un interval de forma

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2,1-\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

de unde prin înlocuire avem

$$(35.22 - 31.56) \pm 2.120 \cdot 4.7155 \cdot \sqrt{\frac{1}{9} + \frac{1}{9}} = 3.66 \pm 4.7126 =$$
$$= (-1.0526, 8.3726).$$

(c) Dacă σ_1 şi σ_2 sunt necunoscuţi şi $\sigma_1 \neq \sigma_2$, atunci statistica

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

urmează legea Student cu n grade de libertate, unde n este soluţia ecuaţiei

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

iar c este dat de

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

Estimarea unei proporţii

Metodele folosite pentru estimarea valorii medii pot fi folosite şi pentru a estima proporţia p de indivizi dintr-o populaţie care au o anumită caracteristică (calitativă), de exemplu pentru a estima cu ajutorul unei selecţii proporţia de alegători care au votat în favoarea unui anumit candidat. Proporţia indivizilor dintr-o selecţie care au o anumită caracteristică poate fi tratată ca un caz special de medie, introducând o variabilă aleatoare X care ia valoarea 1 pentru indivizii care au caracteristica respectivă şi 0 pentru ceilalţi indivizi. Media acestor variabile aleatoare $\bar{X}_n$ are, pentru selecţii de volum mare, o repartiţie aproximativ normală cu media egală cu p şi abaterea medie pătratică $\frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$. Faptul că abaterea medie pătratică depinde de parametrul p necunoscut îngreunează calculele, dar totuşi putem să afirmăm că:

a) oricare ar fi valoare lui $p \in [0,1]$, $p(1-p) \leq \frac{1}{4}$; aşadar dacă folosim $0.5/\sqrt{n}$ în loc de $\frac{\sigma}{\sqrt{n}}$, vom folosi un număr care nu este mai mic decât media pătratică;

b) dacă n este suficeient de mare, eroare ce provine din înlocuirea cantităţii $\sqrt{\frac{p(1-p)}{n}}$ cu $\sqrt{\bar{x}_n(1-\bar{x}_n)/n}$ este mică.

Folosind unul din aceste procedee, putem forma intervale de încredere pentru p în acelaşi mod ca şi pentru m.

Exemplu *La un sondaj organizat în timpul alegerilor, din 1000 de persoane chestionate, 300 s-au pronunţat în favoarea unui anumit candidat. Intervalul de încredere de 95% pentru procentul de alegători care este pentru acel candidat este dat de*

$$\frac{300}{1000} - 1.96\sqrt{\frac{0.3 \cdot 0.7}{1000}} < p < \frac{300}{1000} + 1.96\sqrt{\frac{0.3 \cdot 0.7}{1000}},$$

adică $0.2716 < p < 0.3284$.

Intervale de încredere pentru dispersie şi raportul a două dispersii

Estimarea lui σ^2 cu ajutorul intervalelor de încredere se bazează pe repartiţia dispersiei de selecţie s^2 (sau s'^2).

$$X^2 = \frac{ns^2}{\sigma^2} = \frac{(n-1)s'^2}{\sigma^2}$$

urmează legea hi-pătrat standard cu $n-1$ grade de libertate $\chi^2(n-1,1)$. Pentru a determina un interval de încredere $1-\alpha$ pentru σ^2, vom determina valorile χ_1^2 şi χ_2^2 astfel încât

$$P\left(\chi_1^2 < X^2 < \chi_2^2\right) = 1-\alpha.$$

Dacă $\chi_{n,\alpha}^2$ este cuantila de ordin α a repartiţiei $\chi^2(n,1)$, atunci putem lua $\chi_1^2 = \chi_{n-1,\alpha/2}^2$ şi $\chi_2^2 = \chi_{n-1,1-\alpha/2}^2$, aşa cum se arată în figura 8.2. Avem

$$\chi_{n-1,\alpha/2}^2 < X^2 < \chi_{n-1,1-\alpha/2}^2 \Leftrightarrow \chi_{n-1,\alpha/2}^2 < \frac{ns^2}{\sigma^2} < \chi_{n-1,1-\alpha/2}^2,$$

de unde se obţine

$$\frac{ns^2}{\chi^2_{n-1,1-\alpha/2}} < \sigma^2 < \frac{ns^2}{\chi^2_{n-1,\alpha/2}}.$$

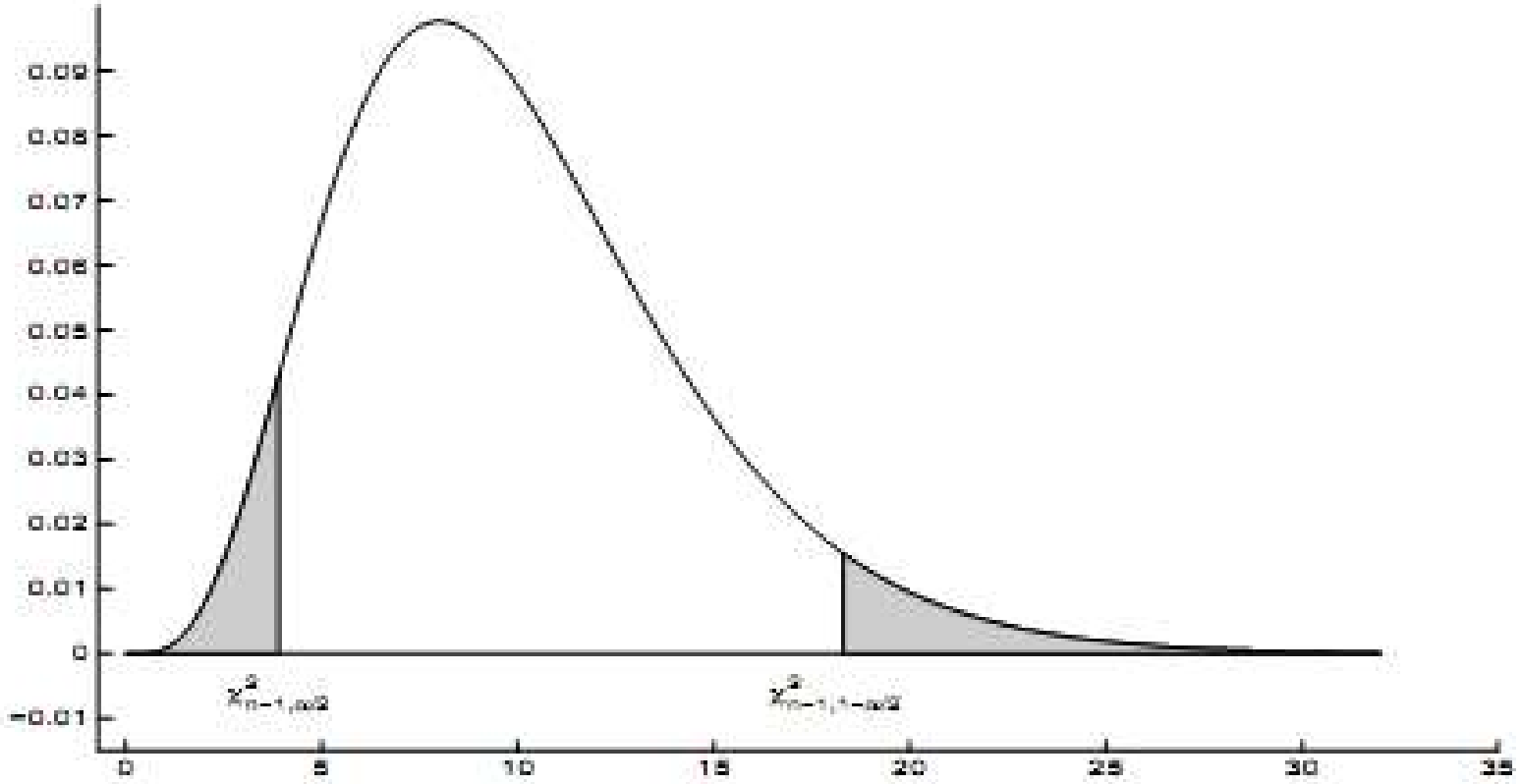


Figura 8.2: Interval de încredere pentru σ^2

Exemplu *Un experimentator doreşte să determine variabilitatea echipamentului pentru măsurarea volumului unei surse audio. S-au efectuat trei măsurători independente pentru acelaşi sunet şi s-au obţinut valorile 4.1, 5.2 şi 10.2. Daţi un interval de încredere de 90% pentru* σ^2.

Soluţie. Presupunând că datele sunt normale, avem $s'^2 = 10.57$, $\alpha/2 = 0.05$, numărul de grade de libertate este $n-1 = 2$, iar cuantilele sunt $\chi^2_{2,0.05} = 0.103$ şi $\chi^2_{2,0.95} = 5.991$.

$$\left(\frac{(n-1)s'^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)s'^2}{\chi^2_{n-1,\alpha/2}} \right) = (3.53, 205.4).$$

Pentru estimarea raportului a două dispersii, deducem că statistica

$$F = \frac{\frac{s_1'^2}{\sigma_1^2}}{\frac{s_2'^2}{\sigma_2^2}}$$

urmează legea F cu $n_1 - 1$ şi $n_2 - 1$ grade de libertate (notaţiile sunt cele din secţiunea 8.6.2). Raţionând ca mai sus se obţine

$$\frac{1}{f_{n_1-1,n_2-1;1-\alpha/2}}\frac{s_1'^2}{s_2'^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{f_{n_1-1,n_2-1;\alpha/2}}\frac{s_1'^2}{s_2'^2}.$$

Exemplu *Se compară două procedee de montaj pentru un dispozitiv, unul clasic şi unul nou, care necesită pentru aplicarea corectă o perioadă de instruire de o lună şi respectiv 3 săptămâni. Au fost instruite două grupuri de câte 9 muncitori, unul cu metoda clasică şi celălat cu metoda nouă. S-a înregistrat timpul de montaj (în minute) pentru fiecare muncitor, obţinânduse rezultatele din tabela de mai jos:*

Procedura	*Timpul*								
Clasică	32	37	35	28	41	44	35	31	34
Nouă	35	31	29	25	34	40	27	32	31

Dorim să determinăm un interval de încredere 95 % pentru raportul dispersiilor.

Soluţie. Avem $n_1 = 8$, $n_2 = 8$, cuantilele sunt $f_{8,8,0.025} = 0.224707$, $f_{8,8,0.975} = 4.45023$, iar dispersiile de selecţie $s_1'^2 = 24.4444$, $s_2'^2 = 20.0278$. Raportul dispersiilor de selecţie este $\frac{s_1'^2}{s_2'^2} = 1.22053$, iar intervalul de încredere $(0.274262, 5.43163)$.

Verificarea ipotezelor statistice

Procedeul de a folosi o selecţie pentru a verifica dacă o ipoteză este adevărată (sau falsă) este numit **test statistic** asupra valabilităţii (sau falsităţii) ipotezei.

Nu există nici o certitudine că nu vom comite o eroare. Intr-adevăr există două tipuri de erori pe care le putem face. Dacă se întâmplă ca ipoteza cercetată să fie adevărată şi noi decidem că este falsă, facem o **eroare de ordinul I** (speţa, genul, tipul I). Probabilitatea acestei erori se notează cu α; dacă dimpotrivă, ipoteza este falsă şi noi decidem că este adevărată, facem o **eroare de ordinul II**, probabilitatea acestei erori notându-se cu β.

Frecvenţa cu care facem o greşeală este, desigur, foarte importantă şi vom vedea că această frecvenţă poate fi controlată până la un anumit grad.

Decizia dacă ipoteza va fi acceptată sau respinsă se va baza pe informaţia pe care o deţinem făcând observaţii şi pe riscul pe care suntem dispuşi să-l acceptăm în a lua o decizie greşită.

Numim ipoteză statistică o presupunere relativă la legea pe care o urmează o caracteristică X.

Când ipoteza statistică se referă la parametrii de care depinde legea de probabilitate a caracteristicii X se obţine un **test parametric**, iar în caz contrar un **test neparametric**.

Pentru testele parametrice vom considera că $\theta \in A = A_0 \cup A_1$, unde $A_0 \cap A_1 = \varnothing$. Ipoteza H_0: $\theta \in A_0$ o vom numi **ipoteză nulă**, iar ipoteza H_1 (sau H_a): $\theta \in A_1$ o vom numi **ipoteză alternativă**. Dacă ipoteza nulă are forma H_0: $\theta = \theta_0$ (ipoteză simplă), putem avea pentru ipoteza alternativă una din formele: $H_1 : \theta \neq \theta_0$ (test bilateral), $H_1 : \theta < \theta_0$ (test unilateral stânga), $H_1 : \theta > \theta_0$ (test unilateral dreapta). Ipoteza nulă este cea asupra căreia ne focalizăm atenţia. În general ea este o propoziţie de forma ,,un parametru al unei populaţii are o valoare specificată".

Ipoteza alternativă este o propoziţie despre acelaşi parametru al populaţiei care este utilizat şi în ipoteza nulă. În general ea ne spune că parametrul populaţiei are o valoare diferită de cea dată în ipoteza nulă. Respingerea ipotezei nule va implica acceptarea ipotezei alternative.

Decizia	Ipoteza nulă este	
	adevărată	falsă
Acceptăm H_0	decizie corectă	eroare de ordinul II
Respingem H_0	eroare de ordinul I	decizie corectă

Construirea unui test revine la obţinerea regiunii critice $U \subset \mathbb{R}^n$ pentru un nivel de semnificaţie (probabilitate de risc) α dat astfel încât

$$P\left((X_1, \ldots, X_n) \in U | H_0\right) = \alpha,$$

unde $X_1, \ldots, X_n$ sunt variabilele de selecţie corespunzătoare selecţiei de volum n considerate. Dacă $(x_1, \ldots, x_n) \notin U$, H_0 va fi acceptată şi dacă $(x_1, \ldots, x_n) \in U$, H_1 va fi respinsă.

Există două abordări pentru testele statistice: abordarea clasică şi abordarea bazată pe probabilitate.

a) Abordarea clasică are urmatorii paşi:

Pasul 1. Formularea ipotezei nule.

Pasul 2. Formularea ipotezei alternative.

Pasul 3. Determinarea criteriului de test – constă în

i. determinarea unei statistici a testului;

ii. specificarea unui nivel de semnificaţie α;

iii. determinarea regiunii critice.

Pasul 4. Calcularea valorii statisticii.

Pasul 5. Luare unei decizii şi interpretarea ei.

Decizia. Dacă valoarea statisticii testului cade în interiorul regiunii critice se respinge H_0, iar în caz contrar se acceptă.

b) Abordarea bazată pe probabilităţi. Valoarea de probabilitate, sau nivelul de semnificaţie P, asupra unei ipoteze este cel mai mic nivel α pentru care informaţiile din selecţia observată sunt semnificative, cu condiţia ca ipoteza nulă să fie adevărată. La luarea unei decizii se va compara P cu valoarea statisticii.

Pasul 1. Formularea ipotezei nule.

Pasul 2. Formularea ipotezei alternative.

Pasul 3. Se determină α.

Pasul 4. Se calculează valoarea statisticii z^* ca la pasul 4 anterior.

Pasul 5. Calculul valorii P. Avem trei cazuri în funcţie de tipul testului (bilateral sau unilateral).

i. Dacă H_1 este unilaterală dreapta ($>$), atunci $P = P(Z > z^*)$ (aria din dreapta lui z^*);

ii. Dacă H_1 este unilaterală stânga ($<$), atunci $P = P(Z < z^*)$ (aria din stânga lui z^*);

iii. Dacă H_1 este bilaterală ($\neq$), atunci $P = 2P(Z > |z^*|)$.

Pasul 6. Decizia se ia comparând P cu valoarea stabilită anterior pentru α :

i. dacă $P \leq \alpha$, se respinge H_0;

ii. dacă $P > \alpha$, se acceptă H_0.

Teste asupra unei populaţii

Testul Z privind media teoretică

Se consideră caracteristica X care urmează legea normală $N(m, \sigma^2)$, unde $m \in \mathbb{R}$ este necunoscut, iar $\sigma > 0$ este cunoscut. Relativ la media teoretică $m = M(X)$, avem ipoteza nulă

$$H_0 : \ m = m_0,$$

în raport cu una din alternativele:

$H1 : m \neq m_0$ (testul Z bilateral);

$H1 : m > m_0$ (testul Z unilateral dreapta);

$H1 : m < m_0$ (testul Z unilateral stânga).

Pentru verificarea ipotezei H_0, în raport cu una din alternativele de mai sus, se consideră o selecţie repetată de volum n şi un nivel de semnificaţie $\alpha \in (0,1)$. Se ştie că statistica

$$Z = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}}$$

urmează legea normală $N(0,1)$. Prin urmare, pentru $\alpha \in (0,1)$ putem determina un interval numeric (z_1, z_2) astfel încât

$$P(z_1 < Z < z_2) = \Phi(z_2) - \Phi(z_1) = 1 - \alpha.$$

Intervalul (z_1, z_2) nu este determinat în mod unic, dar având în vedere alternativa H_1 considerată, adăugăm una din condiţiile suplimentare:

(i) $z_1 = -z_2$ dacă $H1 : m \neq m_0$, adică $\Phi(z_{1-\alpha/2}) = 1 - \alpha/2$;

(ii) $z_1 = -\infty$, $z_2 = z_{1-\alpha}$, unde $\Phi(z_{1-\alpha}) = 1 - \alpha$, pentru $H1 : m > m_0$;

(iii) $z_1 = z_\alpha$, $z_2 = +\infty$, unde $\Phi(z_\alpha) = \alpha$, pentru $H1 : m < m_0$.

Corespunzător celor trei alternative definim regiunea critică:

$$U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n | \frac{|\bar{u} - m_0|}{\frac{\sigma}{\sqrt{n}}} \geq z_{1-\alpha/2} \right\} \qquad \text{(pentru (i))}$$

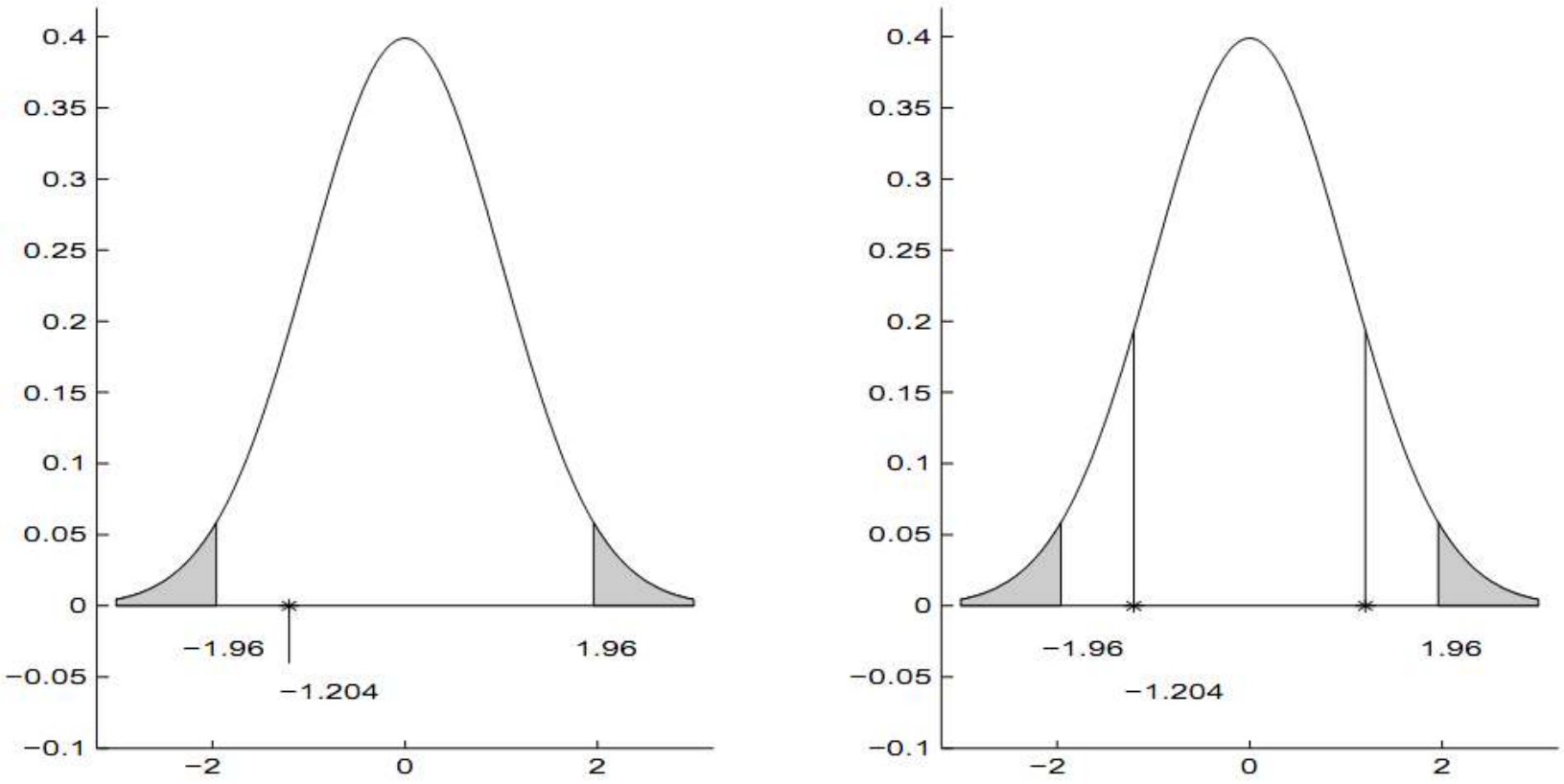
$$U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n | \frac{\bar{u} - m_0}{\frac{\sigma}{\sqrt{n}}} \geq z_{1-\alpha} \right\} \qquad \text{(pentru (ii))}$$

$$U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n | \frac{\bar{u} - m_0}{\frac{\sigma}{\sqrt{n}}} \leq z_{\alpha} \right\}. \qquad \text{(pentru (iii))}$$

unde $\bar{u} = \frac{1}{n} \sum_{k=1}^{n} u_k$.

Ipoteza nulă va fi admisă dacă datele de selecţie satisfac condiţia $(x_1, \ldots, x_n) \notin U$, iar în caz contrar va fi respinsă.

Testul Z se poate aplica şi pentru caracteristici care nu sunt normale, dacă volumul selecţiei este mare ($n > 30$) şi dacă media este necunoscută, iar dispersia cunoscută.



Testul Z – abordarea clasică (stânga) şi abordarea bazată pe probabilităţi (dreapta)

Exemplu *Biroul de internări al unui spital afirmă că vârsta medie a pacienţilor săi este de 42 de ani. O selecţie aleatoare de 120 de vârste obţinute din înregistrările bolnavilor dă o medie de selecţie de 44.2 ani. Este selecţia semnificativă pentru a afirma că media este mai mare de 42 de ani, pentru* $\alpha = 5\%$ *şi* $\sigma = 20$*?*

Soluţie.

(C)

P1. $H_0 : m = 42$

P2. $H_1 : m > 42$

P3. Statistica testului:

$Z = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}}$

$\alpha = 0.05$

Regiunea critică:

vezi figura

P4. $Z = \frac{44.2 - 42}{\frac{20}{\sqrt{120}}} = 1.205$

P5. $z^* < 1.64$ deci H_0 se acceptă.

Nu putem spune că $m > 42$.

(P)

P1. $H_0 : m = 42$

P2. $H_1 : m > 42$

P3. $\alpha = 0.05$

P4. $Z = \frac{44.2 - 42}{\frac{20}{\sqrt{120}}} = 1.205$

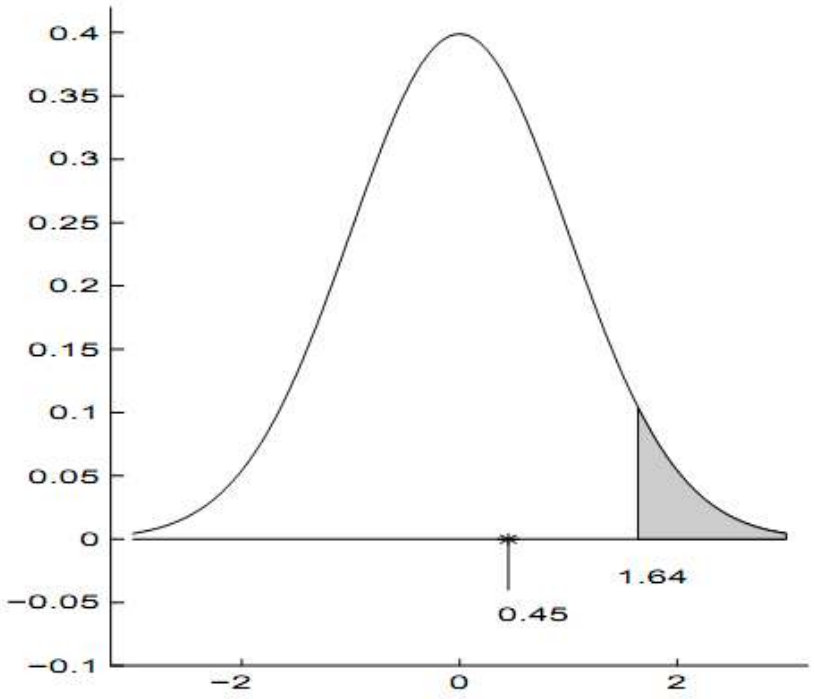
P5. $P = 1 - \Phi(1.205) =$

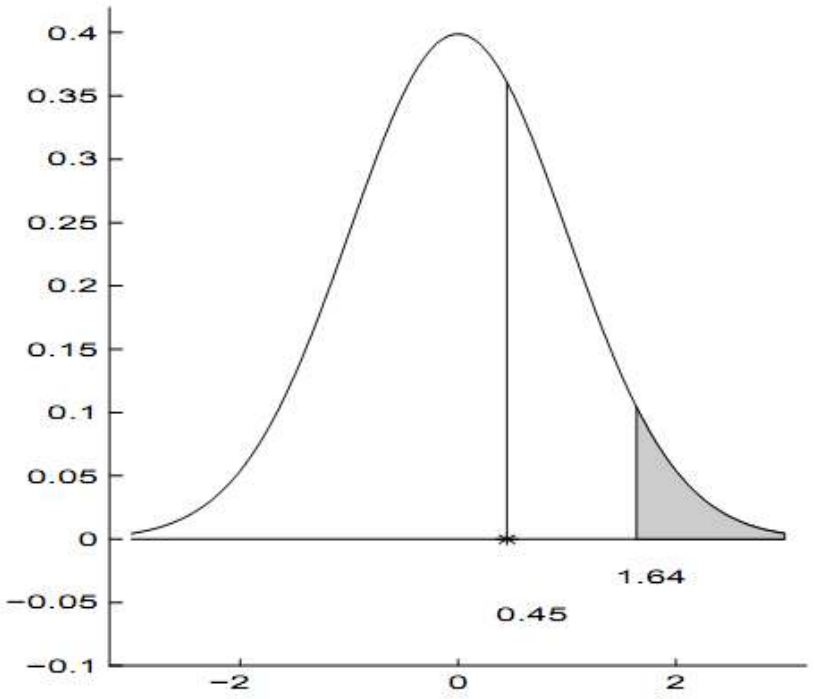
$= 1 - 0.8849 = 0.1151$

(figura dreapta)

P6. $P > 0.05$, deci se acceptă H_0.

Nu putem spune că $m > 42$.


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−0.05
−0.1
−2
0
2
1.64
0.45


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−0.05
−0.1
−2
0
2
1.64
0.45

Testul t (Student) privind media teoretică

Se consideră caracteristica X ce urmează legea normală $N(m, \sigma^2)$, parametrii $m \in \mathbb{R}$ şi $\sigma > 0$ fiind necunoscuţi. Dorim să verificăm ipoteza nulă $H0 : m = m_0$ în raport cu una din alternativele

- $H_1 : m \neq m_0$ (testul t bilateral),
- $H_1 : m > m_0$ (testul t unilateral dreapta),

Pentru verificarea acestei ipoteze se consideră o selecţie repetată de volum n, cu datele de selecţie $x_1, \ldots, x_n$. Statistica

$$T = \frac{\bar{X} - m}{\frac{s'}{\sqrt{n}}} = \frac{\bar{X} - m}{\sqrt{\frac{\bar{m}_2}{n-1}}}$$

urmează legea Student cu $n - 1$ grade de libertate

Prin urmare, pentru nivelul de semnificaţie $\alpha \in (0,1)$ dat, se poate determina intervalul (t_1, t_2) astfel încât

$$P(T \in (t_1, t_2)) = F_{n-1}(t_2) - F_{n-1}(t_1) = 1 - \alpha,$$

unde $F_m(t)$ este funcţia de repartiţie Student cu m grade de libertate. Intervalul (t_1, t_2) nu este determinat unic din statistica de mai sus. În funcţie de alternativa aleasă H_1 se consideră suplimentar:

(1) $t_1 = -t_2$, $t_2 = t_{n-1,1-\alpha/2}$, dacă $H_1 : m \neq m_0$;

(2) $t_1 = -\infty$, $t_2 = t_{n-1,1-\alpha}$, dacă $H_1 : m > m_0$;

(3) $t_1 = t_{n-1,\alpha}$, $t_2 = +\infty$, dacă $H_1 : m < m_0$,

unde $t_{m,\gamma}$ este cuantila de ordin γ a repartiţiei Student cu m grade de libertate, adică $F_m(t_{m,\gamma}) = \gamma$. Corespunzător celor trei ipoteze alternative de mai sus avem regiunile critice:

(1)

$$U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n | \frac{|\bar{u} - m_0|}{\frac{s'}{\sqrt{n}}} \geq t_{n-1,1-\alpha/2} \right\};$$

(2)

$$U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n | \frac{\bar{u} - m_0}{\frac{s'}{\sqrt{n}}} \geq t_{n-1,1-\alpha} \right\};$$

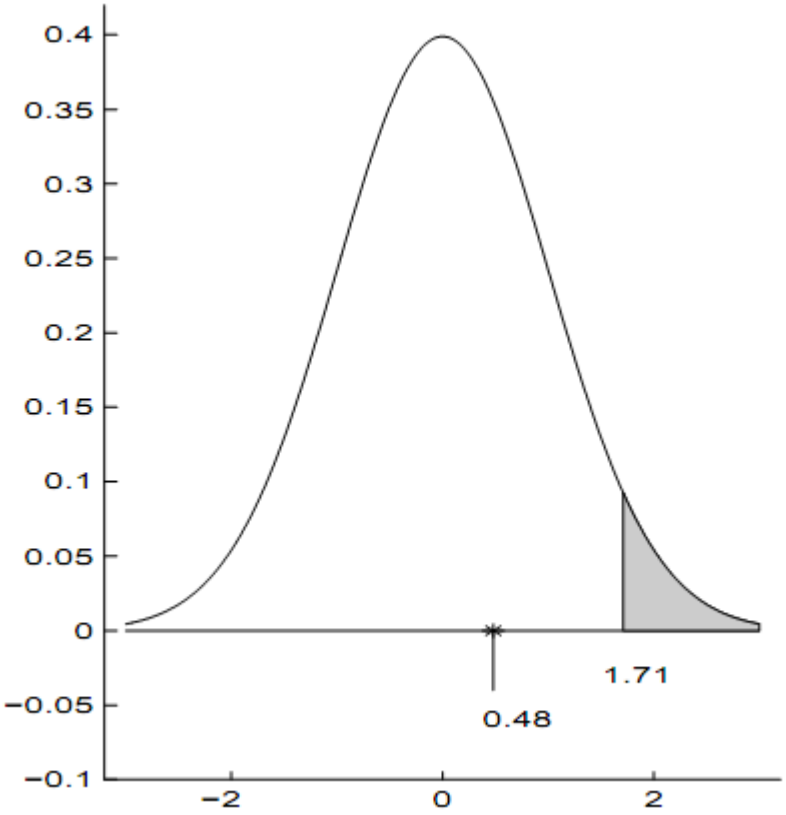
(3)

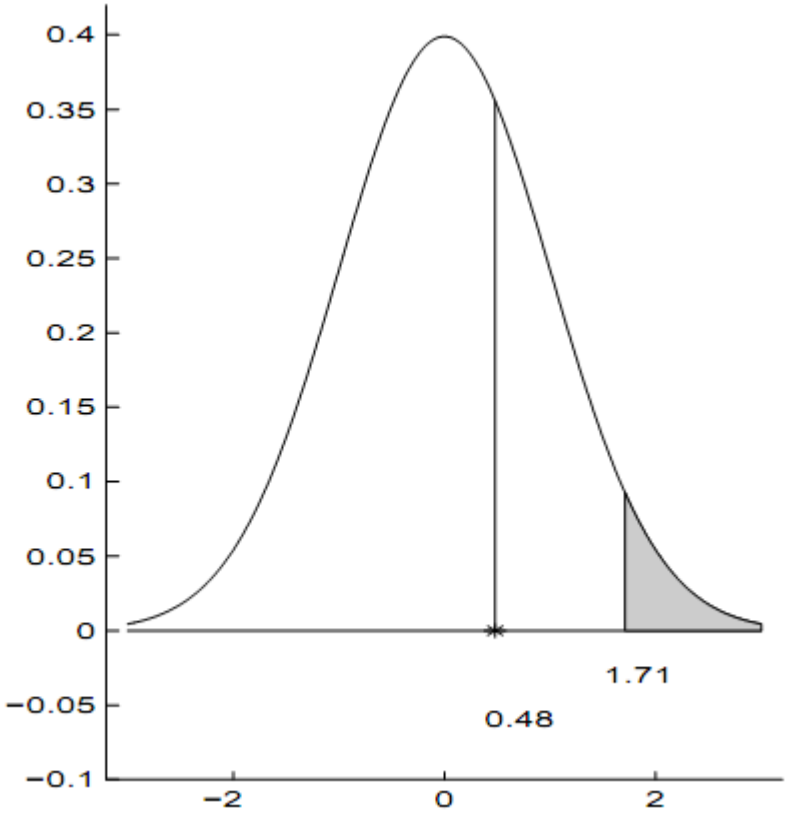
$$U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n | \frac{|\bar{u} - m_0|}{\frac{s'}{\sqrt{n}}} \leq t_{n-1,\alpha} \right\}.$$

Ipoteza H_0 este admisă dacă $(x_1, \ldots, x_n) \notin U$, iar în caz contrar este respinsă.

Exemplu *Autorităţile orăşeneşti din oraşul X afirmă că nivelul de CO din atmosferă nu este mai mare de 4.9. Ne permite o selecţie aleatoare de 25 de determinări cu media* $\bar{x} = 5.1$ *şi* $s' = 2.1$ *să respingem afirmaţia la nivelul de semnificaţie* $\alpha = 5\%$*?*

(C)	(P)
P1. $H_0 : m = 4.9$	P1. $H_0 : m = 4.9$
P2. $H_1 : m > 4.9$	P2. $H_1 : m > 4.9$
P3. Statistica testului:	P3. $\alpha = 0.05$
$T = \frac{\bar{X}-m}{\frac{s'}{\sqrt{n}}}$	P4. $T = \frac{5.1-4.9}{\frac{2.1}{\sqrt{25}}} = 0.47619$
$\alpha = 0.05$	P5.$P = P(T > 0.47619) >$
Regiunea critică:	> 0.25 (figura)
figura	P6. $P > 0.05$, deci nu se
P4. $T = \frac{5.1-4.9}{\frac{2.1}{\sqrt{25}}} = 0.47619$	respinge H_0 – nu putem
P5. Nu se respinge H_0 – nu avem suficiente dovezi să afirmăm că nivelul de CO> 4.9	afirma că nivelul de CO> 4.9


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−0.05
−0.1
−2
0
2
1.71
0.48


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−0.05
−0.1
−2
0
2
1.71
0.48

Teste asupra proporţiilor

Dacă k este numărul de realizări ale unui eveniment A în n probe independente, probabilitatea de realizare la fiecare probă a evenimentului A fiind p, atunci pentru a verifica ipoteza nulă

$$H_0 : p = p_0,$$

în raport cu alternativa unilaterală

$$H_1 : p > p_0,$$

calculăm expresia

$$P = \sum_{l=k}^{n} \binom{n}{l} p_0^l (1 - p_0)^{n-l}$$

şi respingem ipoteza nulă la pragul de semnificaţia α dacă $P > \alpha$.

Dacă n are valori mari, putem folosi statistica Z şi respingem H_0 dacă

$$Z = \frac{k - np_0}{\sqrt{np_0(1-p_0)}} = \frac{\frac{k}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > \Phi^{-1}(\alpha) = \Psi(\alpha).$$

Exemplu *Oficiul pentru protecţia consumatorilor afirmă că* 15% *din fasolea dintr-un lot supus controlului are gărgăriţe. Pentru a verifica afirmaţia la nivelul* 0.10 *se iau* 200 *de boabe şi se găsesc* 7 *cu gărgăriţe. Avem motive să ne îndoim de afirmaţia Oficiului pentru protecţia consumatorilor?*

(C)

P1. $H_0 : p = 0.15 (\geq)$

P2. $H_1 : p < 0.15$

P3. Statistica testului:

$Z = \frac{p' - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$, $p' = \frac{k}{n}$

$\alpha = 0.10$, $z_{0.15} = -1.28$

Regiunea critică:

vezi figura

P4. $p' = \frac{17}{200} = 0.0\,85$

$Z = \frac{0.085 - 0.150}{\sqrt{\frac{0.15 \cdot 0.85}{200}}} = -2.\,5744 = z^*$

P5. Se respinge H_0. Se pare că mai puţin de 15% din boabele de fasole au gărgăriţe.

(P)

P1. $H_0 : p = 0.15 (\geq)$

P2. $H_1 : p < 0.15$

P3. $\alpha = 0.10$

P4. $p' = \frac{17}{200} = 0.0\,85$

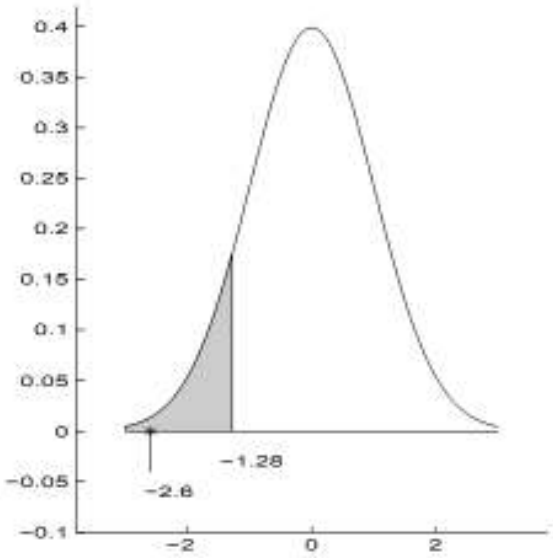
$$Z = \frac{0.085 - 0.150}{\sqrt{\frac{0.15 \cdot 0.85}{200}}} = -2.5744 = z^*$$

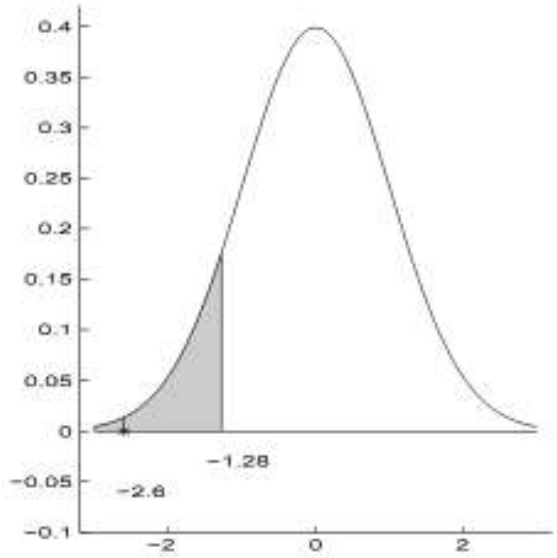
P5. $P = P(Z < z^*) = P(Z < -2.5744) = 0.0047$

figura

P6. $P < 0.10$. Se respinge H_0

Se pare că mai puţin de 15% din boabele de fasole au gărgăriţe.





Testul χ^2 asupra dispersiei

Fie caracteristica X ce urmează legea normală $N(m, \sigma^2)$ unde m şi σ^2 sunt necunoscuţi. Relativ la dispersia teoretică se formulează ipoteza nulă

$$H_0 : \sigma^2 = \sigma_0^2,$$

cu una din alternativele

- $H_1 : \sigma^2 \neq \sigma_0^2$ (testul χ^2 bilateral);
- $H_1 : \sigma^2 > \sigma_0^2$ (testul χ^2 unilateral dreapta);
- $H_1 : \sigma^2 < \sigma_0^2$ (testul χ^2 unilateral stânga).

Pentru a verifica ipoteza nulă H_0 în raport cu una din alternativele H_1 precizate, se consideră o selecţie repetată de volum n, cu datele de selecţie $x_1, \ldots, x_n$. Statistica

$$X^2 = \frac{1}{\sigma^2}\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2 = \frac{(n-1)s'^2}{\sigma^2}$$

urmează legea χ^2 cu $n-1$ grade de libertate.

Pentru nivelul de semnificaţie $\alpha \in (0,1)$ dat se poate determina un interval (χ_1^2, χ_2^2) astfel încât

$$P\left(X^2 \in (\chi_1^2, \chi_2^2)\right) = F_{n-1}(\chi_2^2) - F_{n-1}(\chi_1^2) = 1-\alpha,$$

unde $F_m(x)$ este funcţia de repartiţie pentru legea χ^2 cu m grade de libertate. Deoarece intervalul (χ_1^2, χ_2^2) nu este determinat unic, în funcţie de alternativa H_1 aleasă se consideră condiţia suplimentară:

(1) $\chi_1^2 = \chi^2_{n-1,\alpha/2}$, $\chi_2^2 = \chi^2_{n-1,1-\alpha/2}$, dacă $H_1 : \sigma^2 \neq \sigma_0^2$

(2) $\chi_1^2 = 0$, $\chi_2^2 = \chi^2_{n-1,1-\alpha}$, dacă $H_1 : \sigma^2 > \sigma_0^2$;

(3) $\chi_1^2 = \chi^2_{n-1,\alpha}$, $\chi_2^2 = +\infty$, dacă $H_1 : \sigma^2 < \sigma_0^2$,

unde $\chi^2_{m,\gamma}$ este cuantila de ordin γ a legii χ^2 cu m grade de libertate, adică $F_m(\chi^2_{m,\gamma}) = \gamma$.

Regiunile critice sunt

(1)

$$U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n | \frac{1}{\sigma_0^2} \sum_{k=1}^{n} (u_k - \bar{u})^2 \notin (\chi^2_{n-1,\alpha/2}, \chi^2_{n-1,1-\alpha/2}) \right\};$$

(2)

$$U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n | \frac{1}{\sigma_0^2} \sum_{k=1}^{n} (u_k - \bar{u})^2 \geq \chi^2_{n-1,1-\alpha} \right\};$$

(3)

$$U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n | \frac{1}{\sigma_0^2} \sum_{k=1}^{n} (u_k - \bar{u})^2 \leq \chi^2_{n-1,\alpha} \right\}.$$

Ipoteza nulă va fi admisă dacă $(x_1, \ldots, x_n) \notin U$; în caz contrar va fi respinsă.

Exemplu *Să presupunem că o companie de îmbuteliat băuturi răcoritoare doreşte să detecteze situaţia când variabilitatea volumului de băutură dintr-o sticlă scapă de sub control. O dispersie de* 0.0004 *se consideră acceptabilă şi se procedează la reglarea maşinii de îmbuteliat atunci când dispersia devine mai mare decât acea valoare. Să presupunem că avem o selecţie de 28 de sticle cu dispersia de selecţie (varianţa) de* 0.0010. *Ne indică aceasta că procesul este înafara controlului la un nivel de* 5%*?*

(C)

P1. $H_0 : \sigma^2 = 0.0004$

P2. $H_1 : \sigma^2 > 0.0004$

P3. Statistica testului: X^2, $n = 28$
27 grade de libertate
$\alpha = 0.05$, $\chi^2_{27,0.95} = 40.1$
Regiunea critică:
vezi figura

P4. $X^2 = \frac{27 \cdot 0.001}{0.0004} = 67.5 = \chi^{2*}$

P5. Respingem H_0.

(P)

P1. $H_0 : \sigma^2 = 0.0004$

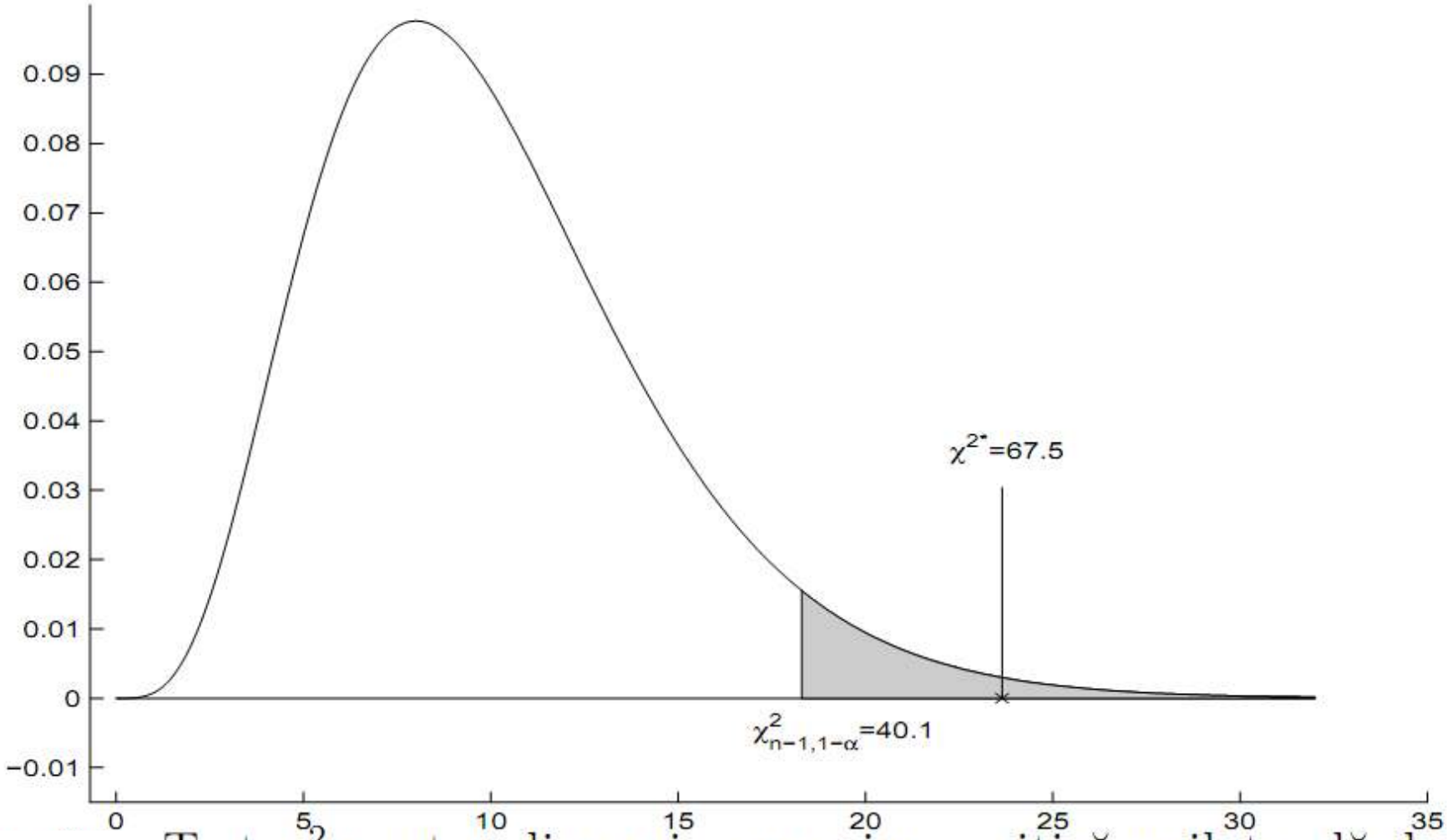
P2. $H_1 : \sigma^2 > 0.0004$

P3. $\alpha = 0.05$

P4. $X^2 = \frac{27 \cdot 0.001}{0.0004} = 67.5 = \chi^{2*}$

P5. $P = P(X^2 > \chi^{2*}) =$
$= P(X^2 > 67.5) < 0.005$
(figura)

P6. $P < 0.005 < 05$, deci
ipoteza se respinge



Test χ^2 pentru dispersie cu regiune critică unilaterală dreapta