Capitolul 1

Testarea ipotezelor

1.1 Problematica testării ipotezelor

În capitolul precedent am arătat în ce fel sunt folosite valorile unei selecţii pentru a estima parametrii unei populaţii ca, media, dispersia, şi cum să asociem parametrilor intervale de încredere. Deseori însă un experimentator este interesat în a verifica doar afirmaţii, ipoteze, relativ la valoarea parametrilor şi nu este neapărat interesat de localizarea sau lungimea intervalului de încredere. Testarea ipotezelor este o metodă folosită în ştiinţă. Un cercetător observă natura, formulează o teorie şi apoi o testează pe baza unor noi observaţii. În contextul statisticii experimentatorul teoretizează că parametrul unei populaţii ia o anumită valoare sau un set de valori. Apoi selectează un eşantion din populaţie şi compară teoria cu observaţiile asupra eşantionului. Dacă observaţiile sunt într-un dezacord serios cu teoria, atunci experimentatorul poate respinge teoria (ipoteza). Dacă însă observaţiile sunt compatibile cu teoria, atunci doar din investigarea unui singur eşantion nu se justifică acceptarea ipotezei. Prin urmare testarea ipotezei trebuie să fie urmată de o decizie. Cum decidem dacă rezultatele observării unui eşantion sunt în dezacord cu ipoteza, când trebuie respinsă ipoteza şi când nu, care este probabilitatea de a lua o decizie nepotrivită, ce funcţie de valorile de selecţie folosim în procesul de decizie, sunt întrebări la care răspunde teoria statistică a testării ipotezelor.

Contextul general şi terminologia standard în problemele de testare a ipotezelor este următoarea: O populaţie este caracterizată de distribuţia de probabilitate ce depinde de parametrul θ. Se investighează un eşantion aleator şi se obţin valorile de selecţie $x_1, x_2, \ldots, x_n$, ca realizări ale selecţiei aleatoare $X_1, X_2, \ldots, X_n$.

Propoziţia care formulează o ipoteză despre valoarea parametrului θ al populaţiei în studiu se numeşte *ipoteza nulă* şi se notează cu H_0. Ipoteza nulă este *simplă*, dacă susţine că parametrul are o valoare fixată şi *compusă* în caz contrar.
Exemple:

$$\begin{aligned} H_0 &: m = 10; \text{ media populaţiei este } 10 \\ H_0 &: p = 0.45; \text{ o proporţie de } 45\% \text{ din populatie are opinia O} \end{aligned}$$

sunt ipoteze simple, în timp ce

$$H_0 : m \leq 10; \text{ media populaţiei este cel mult } 10, \tag{1.1}$$

este o ipoteză compusă.

Pe lângă ipoteza nulă se formulează o *ipoteză alternativă* notată H_a, care este o propoziţie ce susţine că valorile parametrului se găsesc într-o submulţime a complementarei valorilor specificate de ipoteza nulă.

Exemple de ipoteze alternative:

$$
\begin{aligned}
H_0 : m = 10; &\quad H_a : m \neq 10 \\
H_0 : m = 15; &\quad H_a : m < 15; \\
H_0 : m \leq 12; &\quad H_a : m > 12
\end{aligned}
$$

Pe baza valorilor de selecţie se obţine o informaţie care poate conduce la respingerea ipotezei nule în favoarea celei alternative sau nu. Se mai spune că ipoteza nulă beneficiază de prezumpţia de a fi adevărată, până ce avem suficientă ”dovadă” statistică pentru contrariu. Investigarea dacă datele de selecţie $x_1, x_2, \ldots x_n$ sprijină ipoteza nulă se numeşte *testarea ipotezei*. Testarea ipotezei se face cu ajutorul unei *funcţii test*, $t : \mathbb{R}^n \to \mathbb{R}$ şi a valorilor selecţiei. Funcţia test se alege astfel încât variabila aleatoare $t(X_1, X_2, \ldots, X_n)$ să aibă o distribuţie de probabilitate cunoscută, când H_0 este adevărată. Apoi se alege o submulţime $C \subset \mathbb{R}$ (un interval de orice tip sau o reuniune finită de intervale), numită regiune de respingere a ipotezei H_0 sau regiune critică.

`Dacă` $t(x_1, x_2, \ldots, x_n) \in C$`, ipoteza` H_0 `este respinsă;`

`Dacă` $t(x_1, x_2, \ldots, x_n) \notin C$`, ipoteza` H_0 `nu este respinsă.`

Decizia astfel luată este susceptibilă de două tipuri de erori:

1) să respingem ipoteza H_0 când ea este adevărată, eroare numită de ordin unu ;

2) să nu respingem ipoteza H_0 când este falsă – eroare, numită de ordin doi.

Calitatea testului statistic este măsurată de probabilităţile celor două erori:

$$
\begin{aligned}
\alpha &= P(t(X_1, X_2, \ldots, X_n) \in C | H_0) \text{ probabilitatea erorii de ordin 1} \\
\beta &= P(t(X_1, X_2, \ldots, X_n) \notin C | H_a) \text{ probabilitatea erorii de ordin 2}
\end{aligned}
$$

Notaţia $|H_0$, $|H_a$ în definirea probabilităţilor celor două erori semnifică faptul că se presupune că ipoteza H_0, respectiv H_a este adevărată şi deci în calculul probabilităţii se foloseşte distribuţia de probabilitate a statisticii t, corespunzătoare unei valori a parametrului din mulţimea precizată în ipoteza H_0, respectiv H_a.

$P(t(X_1, X_2, \ldots, X_n | H_0) \in C)$ este probabilitatea ca ipoteza H_0 să fie respinsă, când este adevărată. α se numeşte *nivel de semnificaţie*.

În practică se folosesc valori apropiate de 0 pentru nivelul de semnificaţie α, $\alpha = 0.05, 0.01$ sau 0.001 şi se alege regiunea critică astfel încât probabilitatea erorii de ordin 1 asociată, să fie egală cu α. Se aleg aceste valori, pentru că dorim ca eroarea ce s-ar comite respingând ipoteza nulă, atunci când este adevărată, să fie cât mai mică.

Valoarea $t(x_1, x_2, \ldots, x_n)$ asociată valorilor de selecţie este o realizare a variabilei aleatoare $t(X_1, X_2, \ldots, X_n)$. Ea se numeşte *test statistic*.

Definiţia 1.1.1 *Dintre două teste statistice definite de funcţiile test t_1, t_2, asociate ipotezelor H_0, H_a, având erorile de ordin 1 şi 2, α_1, β_1, respectiv α_2, β_2, este mai puternic acela care are ambele erori mai mici decât erorile corespunzătoare ale celuluilalt.*

Fiind fixată eroarea α, de ordin 1, dintre toate regiunile $C \subset \mathbb{R}$, astfel încât $P(t(X_1, X_2, \ldots, X_n) \in C|H_0) = \alpha$ se alege aceea pentru care eroarea corespunzătoare de ordin 2, $\beta = P(t(X_1, X_2, \ldots, X_n) \notin C|H_a)$, este minimă. $1 - \beta$, numit puterea testului, este astfel maxim.

1.2 Teste pentru parametrii legii normale

1.2.1 Testarea ipotezei $H_0 : m = m_0$, când σ este cunoscut

Fie P o populaţie având valorile numerice, ce cuantifică o caracteristică a sa, normal distribuite $N(m, \sigma^2)$. Presupunem că dispersia populaţiei este cunoscută. Ne propunem să analizăm următoarele perechi de ipoteze nule şi ipoteze alternative relativ la media populaţiei:

$$\begin{aligned} H_0 &: \quad m = m_0; \\ H_a &: \quad m < m_0; \end{aligned} \tag{1.2}$$

$$\begin{aligned} H_0 &: \quad m = m_0; \\ H_a &: \quad m > m_0; \end{aligned} \tag{1.3}$$

$$\begin{aligned} H_0 &: \quad m = m_0; \\ H_a &: \quad m \neq m_0; \end{aligned} \tag{1.4}$$

În statistica matematică se demonstrează că pentru fiecare astfel de pereche testul cel mai puternic este $t = \dfrac{\overline{x} - m}{\sigma/\sqrt{n}}$, adică

$$t(x_1, x_2, \ldots, x_n) = \frac{\dfrac{x_1 + x_2 + \ldots, x_n}{n} - m}{\sigma/\sqrt{n}} \tag{1.5}$$

Reamintim că (vezi Observaţia **??**) dacă $\alpha \in (0, 1/2)$ atunci cvantila $z_\alpha < 0$, iar $z_{1-\alpha} > 0$.

1) Fie $\alpha \in (0, 1/2)$ fixat ca nivel de semnificaţie ($\alpha = 0.05, 0.01$ sau 0.001). Să determinăm regiunea de respingere a ipotezei nule H_0 în cazul în care ipoteza alternativă este $H_a : m < m_0$. Această formă a ipotezei alternative ne sugerează că ipoteza H_0 trebuie respinsă când estimatorul mediei, $\overline{x}$, este cu mult mai mic decât m_0, în sensul că $\overline{x}$ plus un număr $a > 0$ de abateri standard, $a\sigma/\sqrt{n}$, este mai mic decât m_0 (Fig.1.1) , adică $\dfrac{\overline{x} - m_0}{\sigma/\sqrt{n}} < -a < 0$. Astfel regiunea de respingere a ipotezei H_0 este de forma:

$$C = \{x \in \mathbb{R}^n \mid \frac{\overline{x} - m_0}{\sigma/\sqrt{n}} < -a\}, \tag{1.6}$$

unde a se determină din condiţia:

$$P\left(\frac{\overline{X} - m_0}{\sigma/\sqrt{n}} < -a\right) = \alpha \tag{1.7}$$

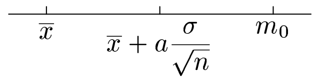

Fig.1.1: Poziţia mediei de selecţie $\overline{x}$ şi a lui $\overline{x} + a\sigma/\sqrt{n}$ faţă de m_0, care sugerează respingerea ipotezei H_0.

Cum selecţia se face dintr-o populaţie ce are distribuţia normală $N(m, \sigma^2)$, cu m necunoscut, $\overline{X} \sim N(m, \sigma^2/n)$ şi deci $\dfrac{\overline{X} - m}{\sigma/\sqrt{n}} \sim N(0,1)$. Probabilitatea (1.7) este o probabilitate asociată variabilei aleatoare $\dfrac{\overline{X} - m}{\sigma/\sqrt{n}}$, ştiind că $m = m_0$. Astfel,

$$P\left(\frac{\overline{X} - m_0}{\sigma/\sqrt{n}} < -a\right) = \Phi(-a) = \alpha. \tag{1.8}$$

Prin urmare $-a = \Phi^{-1}(\alpha) = z_\alpha < 0$ este cvantila distribuţiei $N(0,1)$. În concluzie, dacă

$$\frac{\overline{x} - m_0}{\sigma/\sqrt{n}} < z_\alpha, \tag{1.9}$$

atunci ipoteza H_0 este respinsă cu probabilitatea de eroare egală cu $\alpha < 1/2$ sau testul este semnificativ la nivelul α. (Fig. 1.2); din punct de vedere geometric α este aria domeniului haşurat, deoarece dacă $Z \sim N(0,1)$, $\alpha = P(Z < z_\alpha) = \int_{-\infty}^{z_\alpha} \Phi(t)\,dt$).

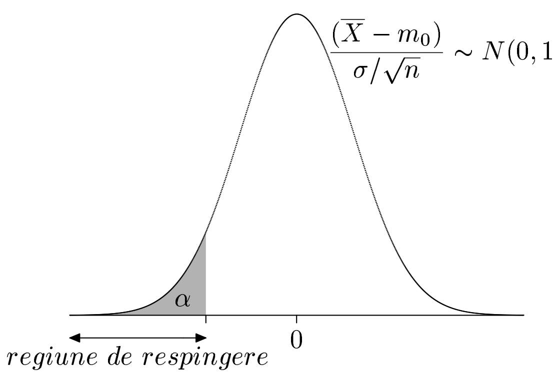

Fig.1.2: Regiunea de respingere pentru ipoteza $H_0 : m = m_0$ cu alternativa $H_a : m < m_0$.

2) În cazul în care ipoteza alternativă este $H_a : m > m_0$, ipoteza H_0 trebuie respinsă când estimatorul mediei, $\overline{x}$, este cu mult mai mare decât m_0, în sensul că $\overline{x}$ minus un număr $a > 0$ de abateri standard, $a\sigma/\sqrt{n}$, este mai mai mare decât m_0 (Fig.1.3), adică $\dfrac{\overline{x} - m_0}{\sigma/\sqrt{n}} > a > 0$. Astfel

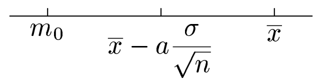

Fig.1.3: Poziţia mediei de selecţie $\overline{x}$ şi a lui $\overline{x} - a\sigma/\sqrt{n}$ faţă de m_0, care sugerează respingerea ipotezei H_0.

regiunea de respingere (regiunea critică) a ipotezei H_0 este de forma:

$$C = \left\{ x \in \mathbb{R}^n \mid \frac{\overline{x} - m_0}{\sigma/\sqrt{n}} > a \right\}, \tag{1.10}$$

unde a se determină din condiţia

$$P\left(\frac{\overline{X} - m_0}{\sigma/\sqrt{n}} > a \right) = \alpha. \tag{1.11}$$

Dar pentru o v.a. $Z \sim N(0,1)$, $P(Z > a) = P(C(Z \leq a)) = 1 - P(Z \leq a) = 1 - \Phi(a)$. În cazul nostru,

$$P\left(\frac{\overline{X} - m_0}{\sigma/\sqrt{n}} > a \right) = 1 - \Phi(a) = \alpha \tag{1.12}$$

şi deci $\Phi(a) = 1 - \alpha$, iar $a = \Phi^{-1}(1-\alpha) = z_{1-\alpha} > 0$. Prin urmare, dacă

$$\frac{\overline{x} - m_0}{\sigma/\sqrt{n}} > z_{1-\alpha}, \tag{1.13}$$

atunci ipoteza H_0 este respinsă cu o eroare de ordin 1, egală cu α (Fig. 1.4; domeniul haşurat are aria $1-\alpha$) .

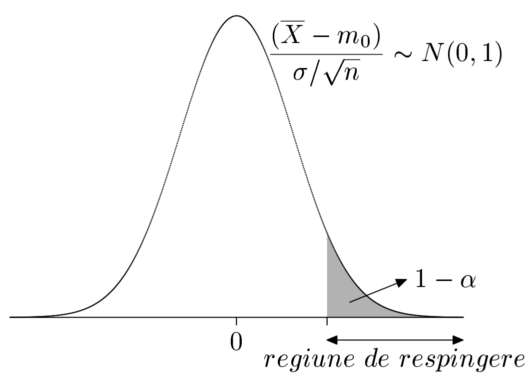

Fig.1.4: Regiunea de respingere pentru ipoteza $H_0 : m = m_0$ cu alternativa $H_a : m > m_0$.

3) Dacă ipoteza alternativă a ipotezei nule $H_0 : m = m_0$ este $H_a : m \neq m_0$, regiunea critică este reuniunea regiunilor critice de la 1) şi 2), adică :

$$C = \left\{ x \in \mathbb{R}^n \mid \left| \frac{\overline{x} - m}{\sigma/\sqrt{n}} \right| > a \right\} \tag{1.14}$$

unde a se determină, la fel, din condiţia ca:

$$P\left(\left|\frac{\overline{X}-m_0}{\sigma/\sqrt{n}}\right|\right)=\alpha \tag{1.15}$$

Dar pentru o v.a. $Z\sim N(0,1)$, $P(|Z|>a)=P(Z<-a)+P(Z>a)=P(Z<-a)+1-P(Z\leq a)=\Phi(-a)+1-\Phi(a)=1-\Phi(a)+1-\Phi(a)=2(1-\Phi(a))$. Prin urmare (1.15) devine:

$$P\left(\left|\frac{\overline{X}-m_0}{\sigma/\sqrt{n}}\right|>a\right)=2(1-\Phi(a))=\alpha. \tag{1.16}$$

Din $2(1-\Phi(a))=\alpha$, rezultă că $\Phi(a)=1-\alpha/2$, adică $a=\Phi^{-1}(1-\alpha/2)=z_{1-\alpha/2}$.

Deci, dacă

$$\left|\frac{\overline{x}-m_0}{\sigma/\sqrt{n}}\right|>z_{1-\alpha/2}, \tag{1.17}$$

atunci ipoteza $H: m=m_0$ este respinsă cu probabilitatea de eroare α (Fig. 1.5).

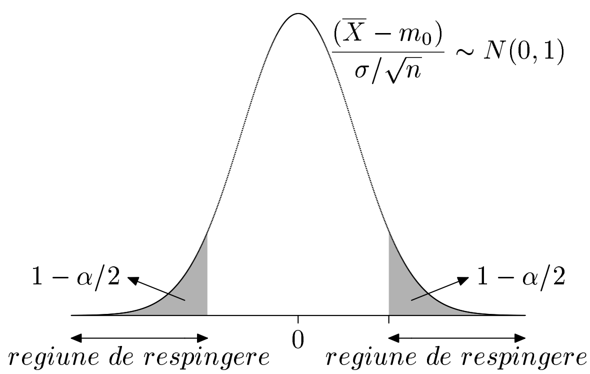

Fig.1.5: Regiunile de respingere pentru ipoteza $H_0: m=m_0$, cu alternativa $H_a: m\neq m_0$; domeniile haşurate au ariile egale cu $1-\alpha/2$.

Centralizăm în Tabelul 1.1 condiţiile în care, pe baza valorilor de selecţie, ipoteza $H_0: m=m_0$, asupra mediei distribuţiei normale $N(m,\sigma^2)$, cu σ cunoscut, este respinsă.

După această prezentare teoretică putem trage primele concluzii utile în abordarea problematicii testării ipotezelor. Şi anume, pentru a rezolva o problemă ce implică testarea ipotezelor se parcurg următoarele etape:

1) formularea ipotezei nule şi a celei alternative;
2) alegerea nivelului de semnificaţie $\alpha=0.05, 0.01$ sau 0.001. Cel mai adesea acesta se specifică în enunţul problemei;
3) identificarea funcţiei test adecvată şi calcularea valorii sale în vectorul datelor de selecţie $(x_1,x_2,\ldots,x_n)$;
4) identificarea valorii critice pentru testul statistic;
5) compararea valorii calculate cu valoarea critică şi formularea deciziei de respingere sau nu a ipotezei nule.

$H_0 : m = m_0$	Condiţia
$H_a : m < m_0$	$\dfrac{\overline{x} - m_0}{\sigma/\sqrt{n}} < z_\alpha$
$H_a : m > m_0$	$\dfrac{\overline{x} - m_0}{\sigma/\sqrt{n}} > z_{1-\alpha}$
$H_a : m \neq m_0$	$\left\lvert\dfrac{\overline{x} - m_0}{\sigma/\sqrt{n}}\right\rvert > z_{1-\alpha/2}$

Tabelul 1.1: Tabel conţinând condiţiile de respingere a ipotezei nule $H_0 : m = m_0$, cu ipoteza alternativă H_a, pe baza unei selecţii din legea $N(m, \sigma^2)$, cu σ^2 cunoscut.

Exemplul 1. BG Grup şi-a propus să vândă zilnic 1000 unităţi din produsul P. Din experienţa anterioară se ştie că volumul vânzărilor din acest produs are distribuţia normală de dispersie $\sigma = 100$ unităţi. Pentru a vedea dacă ţinta propusă poate fi atinsă, au fost monitorizate vânzările din produsul respectiv 100 zile şi s-a obţinut o medie de $\overline{x} = 1021$ unităţi pe zi. Evidenţiază aceasta la un nivel de semnificaţie de 5% că media reală a vânzărilor este diferită de ţinta propusă de 1000 unităţi?

Să ilustrăm etapele enumerate mai sus:
1) Ipoteza nulă H_0: m=1000 unităţi; ipoteza alternativă H_a: $m \neq 1000$;
2) nivelul de semnificaţie este precizat $\alpha = 5\% = 0.05$;
3) Conform Tabelului 1.1 ipoteza nulă este respinsă dacă $\left\lvert\dfrac{\overline{x} - m_0}{\sigma/\sqrt{n}}\right\rvert > z_{1-\alpha/2}$. Prin urmare funcţia test este $t(x_1, x_2, \dots, x_n) = \left\lvert\dfrac{\overline{x} - m_0}{\sigma/\sqrt{n}}\right\rvert$, unde $m_0 = 1000$, $n = 100$, $\overline{x} = (x_1 + x_2 + \cdots x_{100})/100 = 1021$, $\sigma = 100$. Deci $t(x_1, x_2, \dots, x_n) = \left\lvert\dfrac{1021 - 1000}{100/10}\right\rvert = 2.1$;
4) valoarea critică este $z_{1-\alpha/2}$. $1 - \alpha/2 = 1 - 0.025 = 0.975$ şi apelând `norminv(0.0975)` obţinem $z_{1-\alpha/2} = 1.96$.
5) Deoarece $2.1 > 1.96$ ipoteza H_0 ce afirma că media vânzărilor este 1000 se respinge (vezi Tabelul 1.1).

1.2.2 Testarea ipotezei $H_0 : m = m_0$, când σ este necunoscut

Fie $x_1, x_2, \dots x_n$ o selecţie din distribuţia normală $N(m, \sigma^2)$, cu m şi σ^2 necunoscute. Asociem ipotezei nule $H_0 : m = m_0$, aceleaşi ipoteze alternative ca mai sus. Ca funcţie test folosim funcţia t definită astfel:

$$t(x_1, x_2, \dots, x_n) = \frac{\overline{x} - m}{s/\sqrt{n}}, \tag{1.18}$$

unde $\overline{x}$ este media de selecţie din distribţia $N(m, \sigma^2)$, iar s^2 este dispersia valorilor de selecţie. Interpretând ca de obicei x_i ca valori de observare aupra v.a. $X_i \sim N(m, \sigma^2)$, $i = \overline{1, n}$, variabila aleatoare

$$t(X_1, X_2, \ldots, X_n) = \frac{\overline{X} - m}{s/\sqrt{n}} \sim t(n-1) \tag{1.19}$$

are distribuţia Student cu $n-1$ grade de libertate. Cum densitatea de probabilitate a distribuţiei Student este simetrică faţa de dreapta $x = 0$, funcţia de repartiţie Student $\mathcal{T}$ are aceeaşi proprietate ca şi repartiţia Φ a distribuţiei $N(0,1)$:

$$\mathcal{T}(-a) = 1 - \mathcal{T}(a). \tag{1.20}$$

Astfel printr-un raţionament analog cu cel corespunzător cazului σ^2 cunoscut, obţinem condiţiile în care respingem ipoteza H_0 cu o probabilitate de eroare α (Tabelul 1.2), unde t_q reprezintă q–cvantila

$H_0 : m = m_0$	Condiţia
$H_a : m < m_0$	$\frac{\overline{x} - m_0}{s/\sqrt{n}} < t_\alpha$
$H_a : m > m_0$	$\frac{\overline{x} - m_0}{s/\sqrt{n}} > t_{1-\alpha}$
$H_a : m \neq m_0$	$\left\lvert \frac{\overline{x} - m_0}{s/\sqrt{n}} \right\rvert > t_{1-\alpha/2}$

Tabelul 1.2: Tabel conţinând condiţiile de respingere a ipotezei nule $H_0 : m = m_0$, cu ipoteza alternativă H_a, pe baza unei selecţii din legea $N(m, \sigma^2)$, cu σ^2 necunoscut.

distribuţiei Student, $t(n-1)$, cu $n-1$ grade de libertate, n fiind volumul selecţiei.

1.3 Testarea ipotezelor folosind o selecţie de volum mare

Fie $x_1, x_2, \ldots, x_n$ o selecţie aleatoare de volum $n \geq 30$, dintr-o distribuţie arbitrară de medie m şi dispersie σ^2. Dorim să precizăm condiţiile de respingere a ipotezei $H_0 : m = m_0$ în cazul în care ipoteza alternativă este una din următoarele trei: $H_a : m < m_0$, $H_a : m > m_0$, $H_a : m \neq m_0$. Exploatând faptul că, pentru n suficient de mare, media aritmetică $\overline{X}$ a variabilelor aleatoare de selecţie $X_1, X_2, \ldots, X_n$ are distribuţia aproximativ normală de medie m şi disperie $D^2 = \sigma^2/n$, putem repeta raţionamentul de mai sus şi obţinem următoarele condiţii de respingere a ipotezei nule (Tabelul 1.3):

1.4 p-valoarea unui test statistic

Pe lângă metoda prezentată mai sus, care împlică regiunea critică, în testarea unei ipoteze, se mai foloseşte şi metoda p-valorilor [Scheaffer,1995], [Swift, 1997]. Şi anume, fie $x_1, x_2, \ldots, x_n$ o realizare a selecţiei aleatoare $X_1, X_2, \ldots, X_n$, din populaţia ce are distribuţia de probabilitate f_θ, $\theta \in \Theta$.

$H_0 : m = m_0$	Condiţia (σ cunoscut)	Condiţia (σ necunoscut)
$H_a : m < m_0$	$\dfrac{\overline{x} - m_0}{\sigma/\sqrt{n}} < z_\alpha$	$\dfrac{\overline{x} - m_0}{s\sqrt{n}} < z_\alpha$
$H_a : m > m_0$	$\dfrac{\overline{x} - m_0}{\sigma/\sqrt{n}} > z_{1-\alpha}$	$\dfrac{\overline{x} - m_0}{s/\sqrt{n}} > z_{1-\alpha}$
$H_a : m \neq m_0$	$\left\lvert\dfrac{\overline{x} - m_0}{\sigma/\sqrt{n}}\right\rvert > z_{1-\alpha/2}$	$\left\lvert\dfrac{\overline{x} - m_0}{s/\sqrt{n}}\right\rvert > z_{1-\alpha/2}$

Tabelul 1.3: Tabel conţinând condiţiile de respingere a ipotezei nule $H_0 : m = m_0$, cu ipoteza alternativă H_a, pe baza unei selecţii de volum mare dintr-o lege arbitrară

Pentru a testa ipoteza nulă $H_0 : \theta \in \Theta_0$, $\Theta_0 \subset \Theta$, contra unei ipoteze H_a, de forma

$$H_a : \quad \theta < \theta_0; \tag{1.21}$$
$$H_a : \quad \theta > \theta_0; \tag{1.22}$$
$$H_a : \quad \theta \neq \theta_0 \tag{1.23}$$

cu nivelul de semnificaţie α, se alege funcţia test t, se stabileşte distribuţia de probabilitate a variabilei aleatoare $T = t(X_1, X_2, \dots, X_n)$, în condiţiile în cate ipoteza H_0 este adevărată şi se calculează din valorile de selecţie, realizarea $t(x_1, x_2, \dots, x_n)$ a v.a. T. După cum ipoteza H_a are una din formele de mai sus, se calculează probabilităţile:

$$P(T < t(x_1, x_2, \dots, x_n)) \tag{1.24}$$
$$P(T > t(x_1, x_2, \dots, x_n)) \tag{1.25}$$
$$P(|T| > t(x_1, x_2, \dots, x_n) \tag{1.26}$$

care se numesc p–valori ale testului, şi reprezintă probabilitatea de a obţine atunci când H_0 este adevărată, o valoare a lui T în relaţia precizată ($<, >, \neq$) cu valoarea calculată din datele de selecţie, sau încă probabilitatea de a obţine o valoare a testului statistic atât de extremă sau mai extremă decât valoarea obţinută din valorile de selecţie, când H_0 este adevărată. p–valoarea unui test statistic se mai numeşte *nivel de semnificaţie observat*.

Dacă p–valoarea asociată testului este mai mică eventual egală cu nivelul de semnificaţie α, stabilit, atunci ipoteza nulă este respinsă, deoarece $p < \alpha$ implică faptul că un astfel de rezultat extrem este improbabil să se producă când ipoteza nulă este adevărată.

Dacă p–valoarea este mai mare decât α, înseamnă că valori de selecţie ca cele implicate în calculul p–valorii pot apărea destul de probabil când H_0 este adevărată. Astfel nu putem conchide dacă să respingem sau nu ipoteza H_0.

Exemplul 2. Un ghid turistic conţine informaţia că preţul mediu al unui meniu la restaurantele din staţiunea S este de[1] 120000 lei. O broşură editată de o firmă ce desfăşoară activităţi turistice conţine

[1] Preţurile sunt cele din anul 2001, când a apărut prima ediţie a acestei cărţi.

informaţia că preţul mediu al unui meniu este mai mic. Pentru a testa ipoteza $H_0 : m = 120000$ contra ipotezei $H_a : m < 120000$, o agenţie de investigaţii statistice, selectează 80 de restaurante din staţiunea S şi obţine media preţului unui meniu $\overline{x} = 110000$ şi abaterea standard a eşantionului de $s = 30000$. Nivelul de semnificaţie ales este de $\alpha = 0.05$.

Deoarece volumul selecţiei este mare, putem considera că media aritmetică a variabilelor de selecţie are distribuţia normală şi deci dacă ipoteza H_0 este adevărată, variabila aleatoare

$$T = \frac{\overline{X} - 120000}{s/\sqrt{80}} \sim \text{ApN}(0, 1). \tag{1.27}$$

Realizarea v.a. T dedusă din valorile de selecţie este:

$$\frac{\overline{x} - 120000}{s/\sqrt{80}} = \frac{110000 - 120000}{30000/\sqrt{80}} = -2.9814 \tag{1.28}$$

Cum $T \sim \text{ApN}(0, 1)$,

$$P(T < -2.9814) \approx \Phi(-2.9814) = 1 - \Phi(2.9814) = 0.0014 < 0.05. \tag{1.29}$$

În concluzie ipoteza H_0 se respinge.

Calculul p–valorii unui test şi compararea lui cu nivelul de semnificaţie propus, aduce mai multă informaţie decât simpla concluzie: ”ipoteza nulă a fost respinsă la nivelul de semnificaţie α” sau ”rezultatul este nesemnificativ la nivelul α”, deoarece p–valoarea indică şi cât de aproape este rezultatul obţinut de valoarea α, care decide respingerea sau nu a ipotezei H_0 sau în ce măsură datele sunt în dezacord cu ipoteza nulă. Softurile statistice au implementate rutine de calcul a p–valorii cu patru cinci zecimale exacte, pentru testele importante.

Observaţia 1.4.1 *Când nu avem la dispoziţie un software statistic, ci doar valorile tabelate ale funcţiei de repartiţie Φ a distribuţiei normale standard $N(0, 1)$, putem aproxima p–valorile unei statistici având distribuţia Student cu d grade de libertate folosind aproximarea valorilor funcţiei de repartiţie Student cu d grade de libertate:[2]:*

$$\mathcal{T}(x) = \Phi(y), \textit{unde } \ y = \frac{(4d - x^2 - 1)x}{4d + 2x^2}, \textit{dacă } d \geq 3 \tag{1.30}$$

De exemplu presupunem că valoarea t–testului $\dfrac{\overline{x} - m_0}{s/\sqrt{n}}$ obţinută din valorile de selecţie $x_1, x_2, \ldots x_6$ este 2.21. Pentru a evalua p–valoarea

$$P\left(\frac{\overline{X} - m_0}{s/\sqrt{n}} > 2.21\right)$$

[2] B. Li, B. de Moor, A corrected normal approximation for the Student distribution, *Computational Statistics and Data Analysis*, 29 (1999) 213–216.

o exprimăm cu ajutorul funcţiei de repartiţie a distribuţiei Student cu $d = 6 - 1$ grade de libertate:

$$P\left(\frac{\overline{X} - m_0}{s/\sqrt{n}} > 2.21\right) = 1 - P\left(\frac{\overline{X} - m_0}{s/\sqrt{n}} \leq 2.21\right) = 1 - \mathcal{T}(2.21)$$

Acum $\mathcal{T}(2.21)$ va fi aproximată de $\Phi(y)$ conform relaţiei (1.30):

$$y = \frac{(4 \cdot 5 - 2.21^2 - 1)(2.21)}{4 \cdot 5 + 2 \cdot 2.21^2} = 1.77$$

Căutând în tabele obţinem $\Phi(1.77) = 0.96164$ şi deci p–valoarea căutată este $1 - \mathcal{T}(1.77) \approx 1 - 0.96164 \approx 0.0384$. Pentru a evalua aproximaţia calculăm în `MATLAB` $1 - \mathcal{T}(1.77)$, apelând 1-tcdf(2.21,5) şi obţinem 0.039, adică aproximaţia este foarte bună.

1.5 Testarea ipotezelor compuse relativ la medie

Discutăm mai întâi cazul

$$\begin{aligned} H_0 :& \quad m \leq m_0 \\ H_a :& \quad m > m_0 \end{aligned} \tag{1.31}$$

unde m este media unei distribuţii normale $N(m, \sigma^2)$ sau a unei distribuţii arbitrare. Pentru a testa H_0 contra H_a considerăm o selecţie de volum n din $N(m, \sigma^2)$ sau o selecţie de volum mare din distribuţia arbitrară. Calculăm media de selecţie, $\overline{x}$, şi determinăm un interval de încredere de nivel $1 - \alpha$ pentru m de forma $(\theta_1, \infty) = (\overline{x} - z_{1-\alpha}D, \infty)$, unde D este abaterea standard a mediei aritmetice a variabilelor de selecţie sau un estimator al ei. Acest interval conţine media cu o probabilitate de $1-\alpha$. De aceea, dacă $m_0 < \overline{x} - z_{1-\alpha}D$ sau echivalent $\frac{\overline{x}-m_0}{D} > z_{1-\alpha}$, respingem ipoteza $H_0 : m \leq m_0$. Regiunea critică asociată nivelului de semnificaţie α este :

$$C = \left\{(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \mid \frac{\overline{x} - m_0}{D} > z_{1-\alpha}\right\},$$

deoarece

$$\alpha = P(m_0 < \overline{x} - z_{1-\alpha}D) = P\left(\frac{\overline{x} - m_0}{D} > z_{1-\alpha}\right) \tag{1.32}$$

Pentru a testa ipoteza $H_0 : m \geq m_0$ contra $H_a : m < m_0$ procedăm astfel:

Verificăm poziţia lui m_0 faţa de un intervalul de încredere pentru m, cu nivelul de ncredere $1 - \alpha$, interval de forma $(-\infty, \theta_2(x_1, x_2, \ldots, x_n)) = (-\infty, \overline{x} + z_{1-\alpha}D)$. Dacă $m_0 > \overline{x} + z_{1-\alpha}D$ respingem ipoteza H_0. Prin urmare regiunea critică este

$$C = \left\{(x_1, x_2, \ldots, x_n) \mid \frac{\overline{x} - m_0}{D} < -z_{1-\alpha.}\right\}$$

Dar cum $-z_{1-\alpha} = z_\alpha$ (vezi Cap.4, pag. **??**), regiunea critică este:

$$C = \left\{(x_1, x_2, \ldots, x_n) \mid \frac{\overline{x} - m_0}{D} < z_\alpha\right\} \tag{1.33}$$

În ambele cazuri, dacă populaţia are distribuţia $N(m, \sigma^2)$ cu σ^2 cunoscut, atunci $D = \sigma/\sqrt{n}$, iar dacă σ^2 este necunoscut $D = s/\sqrt{n}$, iar în locul cvantilei distribuţiei $N(0, 1)$, se ia cvantila corespunzătoare aceleiaşi valori, dar a distribuţiei Student cu $n - 1$ grade de libertate.

Dacă însă selecţia este de volum mare dintr-o distribuţie arbitrară, atunci cvantila este cea a distribuţiei normale, ori se cunoaşte σ^2 ori nu, iar D se ia ca în cazul selecţiei din distribuţia normală. Toate aceste cazuri sunt centralizate în Tabelele 1.4, 1.5.

Exemplul 3. O corporaţie vinde echipamentul ce-l produce oferind service un an după vânzare. Directorul Departamentului Service susţine că numărul solicitărilor de service pe săptămână nu este mai mare de 15. Pentru a verifica ipoteza sa se analizează cererile de service în 36 de săptămâni selectate aleator cu un rezultat de $\overline{x} = 17$ şi $s^2 = 9$ pentru datele eşantionului. Contrazic aceste rezultate supoziţia directorului la un nivel de semnificaţie de 5%?

Avem de testat ipoteza $H_0 : m \leq 15$, contra ipotezei $H_a : m > 15$. Pentru aceasta se evaluează funcţia test $t(x_1, \ldots, x_{36}) = \frac{\overline{x}-15}{s/\sqrt{36}}$ şi se compară cu $z_{1-\alpha} = z_{0.95} = \Phi^{-1}(0.95) = 1.64$.

Perechea de ipoteze	Legea de prob. (σ cunoscut)
$H_0 : m \leq m_0; H_a : m > m_0$	$N(m, \sigma^2)$; $\dfrac{\overline{x} - m_0}{\sigma/\sqrt{n}} > z_{1-\alpha}$
$H_0 : m \geq m_0; H_a : m < m_0$	$N(m, \sigma^2)$; $\dfrac{\overline{x} - m_0}{\sigma/\sqrt{n}} < z_{\alpha}$
$H_0 : m \leq m_0; H_a : m > m_0$	lege arbitrară, $n \geq 30$; $\dfrac{\overline{x} - m_0}{\sigma/\sqrt{n}} > z_{1-\alpha}$
$H_0 : m \geq m_0; H_a : m < m_0$	lege arbitrară, $n \geq 30$; $\dfrac{\overline{x} - m_0}{\sigma/\sqrt{n}} < z_{\alpha}$

Tabelul 1.4: Tabel conţinând condiţiile de respingere a ipotezei nule H_0 compuse, cu ipoteza alternativă H_a, pe baza unei selecţii din $N(m, \sigma^2)$ sau a unei selecţii de volum mare dintr-o lege arbitrară, când σ este cunoscut.

1.6 Teste pentru diferenţa dintre mediile a două populaţii

În numeroase experimente sunt necesare comparaţii între mediile m_1, m_2, a două populaţii, pe baza datelor obţinute prin selecţii independente din ambele populaţii:

$$x_1, x_2, \ldots, x_{n_1} \tag{1.34}$$

$$y_1, y_2, \ldots, y_{n_2} \tag{1.35}$$

Perechea de ipoteze	Legea de prob. (σ necunoscut)
$H_0 : m \leq m_0; H_a : m > m_0$	$N(m, \sigma^2)$; $\dfrac{\overline{x} - m_0}{s/\sqrt{n}} > t_{1-\alpha}$
$H_0 : m \geq m_0; H_a : m < m_0$	$N(m, \sigma^2)$; $\dfrac{\overline{x} - m_0}{s/\sqrt{n}} < t_{\alpha}$
$H_0 : m \leq m_0; H_a : m > m_0$	lege arbitrară, $n \geq 30$; $\dfrac{\overline{x} - m_0}{s/\sqrt{n}} > z_{1-\alpha}$
$H_0 : m \geq m_0; H_a : m < m_0$	lege arbitrară, $n \geq 30$; $\dfrac{\overline{x} - m_0}{s/\sqrt{n}} < z_{\alpha}$

Tabelul 1.5: Tabel conţinând condiţiile de respingere a ipotezei nule H_0 compuse, cu ipoteza alternativă H_a, pe baza unei selecţii din $N(m, \sigma^2)$ sau a unei selecţii de volum mare dintr-o lege arbitrară, când σ este necunoscut.

Se pot formula următoarele perechi de ipoteze:

$$H_0 : m_1 - m_2 = 0;\ \ H_a : m_1 - m_2 \neq 0 \tag{1.36}$$
$$H_0 : m_1 - m_2 \geq 0;\ \ H_a : m_1 - m_2 < 0 \tag{1.37}$$
$$H_0 : m_1 - m_2 \leq 0;\ \ H_a : m_1 - m_2 > 0 \tag{1.38}$$

Testarea fiecărei ipoteze trebuie să ţină seama de contextul de lucru care poate fi:
(1) ambele populaţii au distribuţia normală $N(m_1, \sigma_1^2)$, respectiv $N(m_2, \sigma_2^2)$, cu dispersiile cunoscute;
(2) ambele populaţii au distribuţia normală $N(m_1, \sigma_1^2)$, respectiv $N(m_2, \sigma_2^2)$, cu dispersiile necunoscute;
(3) selecţiile pentru testarea ipotezei au volum foarte mare şi distribuţiile populaţiilor sunt arbitrare.

În fiecare caz în parte se precizează distribuţia de probabilitate exactă sau aproximativă a variabilei standardizate, asociate diferenţei $\overline{X}-\overline{Y}$ dintre mediile aritmetice ale variabilelor de selecţie din prima, respectiv a doua populaţie, în condiţiile în care H_0 este adevărată.

Apoi în funcţie de perechea de ipoteze H_0, H_a se determină regiunea critică ca şi în cazul corespunzător al unei singure populaţii (vezi secţiunea 1.5 şi tabelele 1.1, 1.3, 1.4 în care m_0 se înlocuieşte cu 0, $\overline{x}$ cu $\overline{x} - \overline{y}$, $\sigma/\sqrt{n}$ se înlocuieşte cu D, unde D^2 este dispersia diferenţei $\overline{X} - \overline{Y}$, $D^2 = \sigma_1^2/n_1 + \sigma_2^2/n_2$ şi în final $s/\sqrt{n}$ se înlocuieşte cu $S = \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$, s_1^2 fiind dispersia primei selecţii, iar s_2^2, dispersia celei de-a doua selecţii).

O discuţie specială necesită cazul în care selecţiile independente se fac din legile normale $N(m_1, \sigma^2)$, $N(m_2, \sigma^2)$, cu dispersiile egale şi necunoscute. Să testăm ipoteza:

$$H_0 \ : \ m_1 - m_2 = 0 \tag{1.39}$$
$$H_a \ : \ m_1 - m_2 \neq 0 \tag{1.40}$$

În acest caz un estimator al dispersiei teoretice comune este conform (**??**):

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_i-\overline{x})^2 + \sum_{j=1}^{n_2}(y_j-\overline{y})^2}{(n_1-1)+(n_2-1)} \tag{1.41}$$

Astfel testul statistic este:

$$t(x_1,\dots,x_{n_1};y_1,\dots,y_{n_2}) = \frac{\overline{x}-\overline{y}-0}{\sqrt{\dfrac{s^2}{n_1}+\dfrac{s^2}{n_2}}} \tag{1.42}$$

iar v.a. a cărei realizare este această valoare, obţinută din selecţii, are distribuţia Student $t(n_1+n_2-2)$. Ipoteza H_0 este respinsă în cazul în care:

$$\left|\frac{\overline{x}-\overline{y}-0}{\sqrt{\dfrac{s^2}{n_1}+\dfrac{s^2}{n_2}}}\right| > t_{1-\alpha/2}, \tag{1.43}$$

unde $t_{1-\alpha/2}$ este cvantila distribuţiei Student $t(n_1+n_2-2)$.

Exemplul 4. O firmă producătoare de încălţaminte organizează cursuri de perfecţionare/specializare pentru noii angajaţi. Pentru a se specializa în executarea unei operaţii de asamblare a unui tip de pantof, un nou angajat necesită aproximativ o lună de cursuri de pregătire. Pentru a testa o nouă metodă folosită în cursurile de specializare, faţă de metoda standard s-au selectat aleator două grupuri de câte 9 cursanţi, un grup fiind instruit conform noii metode, iar al doilea conform procedurii standard. Timpul (în minute) în care fiecare cursant reuşeşte după trei săptămâni de cursuri, să asambleze un pantof din tipul precizat este dat în tabelul următor:

Metoda standard, x_i, $i=\overline{1,9}$:	32	37	35	28	41	44	35	31	34
Metoda nouă, y_i, $i=\overline{1,9}$	35	31	29	25	34	40	27	32	31

Indică aceste date că timpul mediu necesar unui angajat să asambleze un pantof după trei saptămâni de perfecţionare este mai mic pentru cursanţii instruiţi conform noii proceduri, ştiind că distribuţia măsurătorilor timpului de execuţie este normală pentru fiecare populaţie şi cu aceeaşi abatere standard? Să se determine şi p valoarea testului folosit.

Fie m_1, m_2 timpul mediu de asamblare necesar unui individ instruit conform procedurii standard, respeciv noii proceduri. Testăm ipoteza $H_0 : m_1 = m_2$ contra ipotezei $H_a : m_1 > m_2$.

Mediile de selecţie şi sumele pătratelor abaterilor de la medii sunt:

$$\overline{x} = 35.22, \quad \sum_{i=1}^{9}(x_i-\overline{x})^2 = 195.56$$

$$\overline{y} = 31.56, \quad \sum_{i=1}^{9}(y_i-\overline{y})^2 = 160.22.$$

Estimatorul dispersiei teoretice comune este:

$$s^2 = \frac{\sum_{i=1}^{9}(x_i - \overline{x})^2 + \sum_{i=1}^{9}(y_i - \overline{y})^2}{9+9-2} = \frac{195.56 + 160.22}{16} = 22.24. \tag{1.44}$$

Astfel abaterea standard este $s = 4.72$.

Deoarece ipoteza alternativă este $H_a : m_1 > m_2$, avem de testat dacă:

$$\frac{\overline{x} - \overline{y} - 0}{\sqrt{\dfrac{s^2}{n_1} + \dfrac{s^2}{n_2}}} > t_{1-\alpha}, \tag{1.45}$$

unde $t_{1-\alpha}$ este cvantila distribuţiei Student cu 16 grade de libertate, iar α este eroarea de ordin 1, aleasă, să zicem $\alpha = 0.05$. $1 - \alpha = 1 - 0.05 = 0.95$. Prin urmare apelând funcţia `MATLAB`, `tinv`, obţinem: $t_{0.95} =$`tinv`$(0.95, 16) = 1.7459$. Testul statistic este: $t = \dfrac{35.22 - 31.56}{4.72\sqrt{\dfrac{1}{9} + \dfrac{1}{9}}} = 1.64$. Deoarece $1.64 < 1.7459$, ipoteza H_0 nu se respinge. Prin urmare conchidem că nu există suficientă dovadă că noua metodă este superioară celei standard.

Pentru a determina p valoarea testului statistic, calculăm

$$P\left(\frac{\overline{X} - \overline{Y}}{s\sqrt{\dfrac{1}{9} + \dfrac{1}{9}}} > 1.64\right). \tag{1.46}$$

Cum variabila aleatoare

$$\frac{\overline{X} - \overline{Y}}{s\sqrt{\dfrac{1}{9} + \dfrac{1}{9}}}, \tag{1.47}$$

are distribuţia Student $t(16)$, probabilitatea cerută este egală cu $1 - \mathcal{T}(1.64)$, unde $\mathcal{T}$ este funcţia de repartiţie a legii $t(16)$.

$\mathcal{T}(1.64) =$`tcdf`$(1.64, 16) = 0.9397$ şi deci $p = 1 - 0.9397 = 0.0603$. Deoarece $p > \alpha$, ipoteza H_0 nu este respinsă.

Un caz special şi de interes pentru practica statistică este testul pentru verificarea ipotezei

$$H_0 \;:\; m_1 - m_2 = 0 \tag{1.48}$$

$$H_a \;:\; m_1 - m_2 \neq 0 \tag{1.49}$$

pe baza selecţiilor din $N(m_1, \sigma_1^2)$, $N(m_2, \sigma_2^2)$, cu $\sigma_1 \neq \sigma_2$ şi necunoscute. Se demonstrează că regiunea critică pentru testarea ipotezei nule este:

$$\left|\frac{\overline{x} - \overline{y} - 0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}\right| > \frac{\dfrac{s_1^2}{n_1}t_1 + \dfrac{s_2^2}{n_2}t_2}{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}, \tag{1.50}$$

unde $t_1 = t_{1-\alpha/2}$ este cvantila distribuţiei $t(n_1 - 1)$, iar $t_2 = t_{1-\alpha/2}$ este cvantila distribuţiei $t(n_2 - 1)$.

1.7 Teste pentru parametrul p al distribuţiei binomiale

Considerăm o populaţie în care o caracteristică a indivizilor săi are două calităţi exclusive A_1 şi A_2. Selectând la întâmplare n indivizi, observăm că x au calitatea A_1. Considerând x o realizare a unei v.a. $X \sim$ Bin(n, p) cu p necunoscut, dorim să testăm, ipoteza H_0 contra H_a, relativ la proporţia p din populaţie care are atributul A_1:

$$\begin{aligned} H_0 :&\quad p = p_0; \\ H_a :&\quad p \neq p_0; \end{aligned} \tag{1.51}$$

$$\begin{aligned} H_0 :&\quad p \leq p_0; \\ H_a :&\quad p > p_0; \end{aligned} \tag{1.52}$$

$$\begin{aligned} H_0 :&\quad p \geq p_0; \\ H_a :&\quad p < p_0; \end{aligned} \tag{1.53}$$

În acest scop, reamintim ca pentru n foarte mare, variabila aleatoare $X \sim$ ApN$(np, np\,(1-p))$ are aproximativ distribuţia normală. Conform Propoziţiei **??**, variabila $\dfrac{X}{n}$ are de asemenea distribuţia aproximativ normală. Mai precis, ţinând seama că $M(X/n) = M(X)/n$ şi $\sigma^2(X/n) = \sigma^2(X)/n^2$, obţinem $\dfrac{X}{n} \sim$ ApN$(p, p\,(1-p)/n)$. Pe baza rezultatelor deduse deja privind testarea ipotezelor relativ la media unei populaţii de volum mare, deducem că trebuie să aplicăm testul Z selecţiei x:

$$Z = \frac{p^* - p_0}{d}, \tag{1.54}$$

unde $p^* = x/n$ este un estimator al parametrului p, interpretat ca medie a variabilei X/n, d este un estimator al abaterii standard a v.a. X/n, şi anume: $d = \sqrt{p^*(1-p^*)/n}$. Ipoteza H_0 este respinsă, în funcţie de ipoteza alternativă asociată, conform Tabelului 1.6, probabilitatea aproximativă de eroare fiind α.

Exemplul 5. O agenţie de testare a opiniei publice a selectat la întâmplare 457 utilizatori internet, pentru a testa informaţia dată de Ministerul Comerţului, potrivit căreia cel puţin 20% dintre cei ce navighează pe internet au făcut în anul 2004 cumpărături la un magazin virtual. 393 dintre cei chestionaţi, au afirmat că nu au cumpărat niciodată vreun produs de la un magazin virtual. Evidenţiază această informaţie, la nivelul de semnificaţie de 0.01, că mai puţin de 20% dintre utilizatorii internet fac cumpărături de la magazine virtuale?

Rezolvare: Populaţia investigată este cea a utilizatorilor internet. Aceştia se împart în două clase disjuncte: cei ce efecturează cumpărături de la magazine virtuale şi cei care nu. Deci vom avea de efectuat test privind proporţia p din populaţie care efectuează cumpărături pe internet.
(1) Formulăm ipoteza nulă şi alternativă:

$$H_0 : \ p \geq 0.20 \quad H_a : \ p < 0.20$$

Ipoteza H_0	Ipoteza H_a	Condiţia de respingere H_0
$p = p_0$	$p \neq p_0$	$\left\lvert \dfrac{p^* - p_0}{d} \right\rvert > z_{1-\alpha/2}$
$p \leq p_0$	$p > p_0$	$\dfrac{p^* - p_0}{d} > z_{1-\alpha}$
$p \geq p_0$	$p < p_0$	$\dfrac{p^* - p_0}{d} < z_{\alpha}$

Tabelul 1.6: Tabel conţinând condiţiile de respingere a ipotezei nule $H_0 : p = p_0$, cu ipoteza alternativă H_a, pe baza unei selecţii din legea $Bin(n, p)$, cu n suficient mare.

(2) nivelul de semnificaţie este $\alpha = 0.01$;
(3) Conform Tabelului 1.6 ipoteza nulă este respinsă dacă $\dfrac{p^* - p_0}{d} < z_\alpha$. Valoarea de selecţie ce indică numărul celor ce au făcut cumpărături pe internet este $x = 457 - 393 = 64$ şi testul $Z(x) = \dfrac{p^* - 0.2}{d}$, unde $p^* = 64/457 = 0.14004$, şi $d = \sqrt{p^*(1-p^*)/n} = 0.016233$. Deci $Z(64) = -3.6934$;
(4) valoarea critică este z_α. Apelând `norminv(0.01)`:=z_α obţinem -2.3263.
(5) Deoarece $-3.6934 < -2.3263$, respingem ipoteza nulă. Deci datele rezultate din chestionaarea unui eşantion de navigatori pe internet nu susţin informaţia dată de Ministerul Comerţului.

Fie acum x_1, x_2 două valori de selecţie, independente, x_1 realizare a v.a. $X_1 \sim \text{Bin}(n_1, p_1)$, iar x_2 realizare a v.a. $X_2 \sim \text{Bin}(n_2, p_2)$. Testăm ipoteza $H_0 : p_1 = p_2$ contra ipotezei $H_a : p_1 \neq p_2$.

În ipoteza că $p_1 = p_2 = p$, distribuţia de probabilitate a v.a. $X_1 + X_2$ este $\text{Bin}(n_1 + n_2, p)$, deoarece X_1, X_2 sunt v.a. independente. Aplicând metoda celor mai mici pătrate obţinem că în acest caz un estimator al lui p este: $p^* = \dfrac{x_1 + x_2}{n_1 + n_2}$. Pentru n_1 şi n_2 foarte mari, $Y = \dfrac{X_1}{n_1} + \dfrac{X_2}{n_2} \sim \text{ApN}(p, D^2)$, unde $D = \sqrt{\dfrac{p\,(1-p)}{n_1 + n_2}}$. Astfel testarea ipotezei $H_0 : p_1 = p_2$ contra ipotezei $H_a : p_1 \neq p_2$ revine la testarea ipotezei:

$$\begin{aligned} H_0 &: \quad p = 0; \\ H_a &: \quad p \neq 0; \end{aligned} \tag{1.55}$$

pe baza realizărilor x_1, x_2 ale v.a. X_1, X_2. Faptul că $Y \sim \text{ApN}(p, D^2)$ sugerează alegerea statisticii $Z = \dfrac{p^* - 0}{d}$, unde d este un estimator al lui D,

$$d = \sqrt{\frac{p^*(1-p^*)}{n_1 + n_2}}. \tag{1.56}$$

Dacă

$$\left|\frac{p^*}{\sqrt{\frac{p^*(1-p^*)}{n_1+n_2}}}\right| > z_{1-\alpha/2}, \tag{1.57}$$

atunci ipoteza $H_0 : p = 0$ este respinsă.

1.8 Teste de concordanţă

În analiza unor date de selecţie se pune problema de a deduce dacă acestea provin sau nu dintr-o anumită lege de probabilitate, reprezentată prin funcţia de repartiţie. De exemplu, în analiza reţelei trafic rutier se poate ipotetiza că numărul accidentelor cea au loc pe o anumită arteră are distribuţia Poisson de medie λ accidente pe săptămână. O astfel de ipoteză este testată statistic, luând date din trafic şi analizând dacă supoziţia privind legea Poisson este în concordanţă cu datele. Testele de acest tip se numesc *teste de concordanţă* (*goodness of fit tests*, în limba engleză).

Având datele de selecţie $x_1, x_2, \dots x_n$, interpretate ca fiind valori de observare asupra n v.a. independente şi identic distribuite după legea de probabilitate având funcţia de repartiţie necunoscută F, un estimator al acestei funcţii este *funcţia empirică de repartiţie* definită astfel:

$$F_e(x) = \frac{\text{nr. valorilor}\{x_i \mid x_i \leq x\}}{n} \tag{1.58}$$

Propoziţia 1.8.1 *Funcţia empirică de repartiţie asociată datelor de selecţie $x_1, x_2, \dots, x_n$ este un estimator nedeplasat (centrat) al funcţiei de repartiţie a legii din care provin datele.*

Demonstraţie: Funcţia empirică de repartiţie poate fi exprimată cu ajutorul funcţiei caracteristice, χ[3], a mulţimii $D = (-\infty, x]$:

$$\chi_D(x) = \begin{cases} 1 & \text{dacă } x \in D \\ 0 & \text{în rest} \end{cases} \tag{1.59}$$

Şi anume:

$$F_e(x) = \frac{1}{n}\sum_{i=1}^{n} \chi_{(-\infty,x]}(x_i) \tag{1.60}$$

Notând cu $X = X_i$ variabila aleatoare ce are distribuţia comună a variabilelor de selecţie X_i, $i = \overline{1,n}$ şi cu $F(x) = P(X \leq x)$ funcţia de repartiţie a lui X, estimatorul $F_e(x)$ este un estimator nedeplasat pentru $F(x)$ dacă avem:

$$M\left(\frac{1}{n}\sum_{i=1}^{n} \chi_{(-\infty,x]}(X_i)\right) = M\left(\frac{1}{n}\sum_{i=1}^{n} \chi_{(-\infty,x]}(X)\right) = F(x) \tag{1.61}$$

[3]Numele acestei funcţii se notează, clasic, cu χ şi nu are nici o legătură cu numele testului χ^2 ce va fi prezentat în continuare.

Calculând a doua medie din relaţia de mai sus obţinem:

$$M\left(\frac{1}{n}\sum_{i=1}^{n}\chi_{(-\infty,x]}(X)\right)=\frac{1}{n}\sum_{i=1}^{n}M\left(\chi_{(-\infty,x]}(X)\right) \tag{1.62}$$

Dar se ştie că media unei variabile $Y = g(X)$, când există, este egală cu: $M(Y) = \int_{-\infty}^{\infty} g(x)f(x)dx$, unde f este densitatea de probabilitate a v.a. X. În cazul nostru, deoarece $\chi_{(-\infty,x]}(u) = 0$, pentru $u \in (x, \infty)$ avem:

$$\begin{aligned} M\left(\chi_{(-\infty,x]}(X)\right) &= \int_{-\infty}^{\infty}\chi_{(-\infty,x]}(u)f(u)du = \int_{-\infty}^{x} f(u)du = P(X \le x) = \\ &= F(x) \end{aligned} \tag{1.63}$$

□

Datorită acestei proprietăţi funcţia empirică de repartiţie este folosită în testele de concordanţă. Un astfel de test are ipoteza nulă:

$$H_0: \text{ datele provin din legea de funcţie de repartiţie } F_0.$$

1.8.1 Testul χ^2 pentru distribuţii discrete

Testul χ^2 datorat lui K. Pearson este cel mai vechi şi mai bun test de concordanţă pentru distribuţii discrete (Poisson, binomială, etc).

Testarea ipotezei revine la a compara funcţia de repartiţie ipotetică F_0 cu funcţia empirică de repartiţie asociată datelor statistice $x_1, x_2, \ldots x_n$. La un nivel de semnificaţie prescris, ipoteza nulă este respinsă sau nu, în funcţie de abaterea repartiţiei F_e faţă de repartiţia teoretică F_0.

Pentru a investiga concordanţa datelor cu funcţia de repartiţie F_0 se împart datele în k clase. Mai precis, se determină a=$\min\{x_i \mid i = \overline{1,n}\}$ şi $b = \max\{x_i \mid i = \overline{1,n}\}$. Se divizează intervalul $[a-\epsilon, b]$ cu $\epsilon > 0$ suficient de mic, prin punctele de diviziune $y_0 = a - \epsilon < y_1 < \cdots < y_j < \cdots < y_k = b$ în k subintervale disjuncte, astfel încât în fiecare subinterval $(y_{j-1}, y_j]$, $j = \overline{1,k}$ să cadă cel puţin 5 valori de selecţie. Notăm cu N_j numărul de valori din selecţie ce cad în intervalul $I_j = (y_{j-1}, y_j]$, $j = \overline{1,k}$.

Valorea x_i este valoare de observare asupra unei variabile aleatoare discrete X în a i-a repetare a unui experiment aleator, $i = \overline{1,n}$. Dacă ipoteza H_0 este adevărată atunci variabila X are funcţia de repartiţie F_0. Notăm cu $p_j = P(X \in I_j)$, probabilitatea ca X să ia valori în intervalul I_j:

$$\begin{aligned} p_1 &= P(X \le y_1) = F_0(y_1); \\ p_j &= P(y_{j-1} < X \le y_j) = F(y_j) - F(y_{j-1}), j = \overline{2, k-1}; \\ p_k &= P(X > y_{k-1}) = 1 - F(y_{k-1}); \end{aligned} \tag{1.64}$$

Evident că $p_1 + p_2 + \cdots p_k = 1$. Pe de altă parte variabilele aleatoare discrete N_j ce dau numărul de valori ce cad în intervalul I_j sunt variabile binomiale $N_j \sim \text{Bin}(n, p_j)$. Deci media lor este $M(N_j) = np_j$. Concordanţa dintre valorile de selecţie şi funcţia de repartiţie teoretică (ipotetizată) depinde de diferenţele $D_j = N_j - np_j$. Cu cât diferenţele sunt mai mici cu atât mai bună este aproximarea funcţiei de repartiţie F_0 prin funcţia empirică de repartiţie asociată datelor. Dacă n este suficient de mare atunci probabilităţile p_j sunt mici şi astfel distribuţia binomială poate fi aproximată de distribuţia Poisson, adică $N_j \sim \text{Poiss}(\lambda = np_j)$. Prin urmare, $\sigma^2(N_j) = \lambda = np_j$, iar variabila aleatoare standardizată are distribuţia aproximativ normală

$$Z_j = \frac{N_j - np_j}{\sqrt{np_j}} \sim \text{ApN}(0,1) \tag{1.65}$$

Conform Propoziţiei **??**, variabila:

$$T = \sum_{j=1}^{k} Z_j^2 = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j} \tag{1.66}$$

are distribuţia $\chi^2(k-1)$ cu $k-1$ grade de libertate.

Aşa cum am menţionat mai sus, valori mici ale funcţiei test T sunt în favoarea ipotezei nule, în timp ce valori mari indică falsitatea ei. Fixând un nivel de semnificaţie $\alpha = 0.05, 0.01, 0.001$, putem determina valoarea critică t_c astfel încât atunci când ipoteza H_0 este adevărată, să avem $P(T > t_c) = \alpha$, sau echivalent $1 - P(T \leq t_c) = \alpha$, adică $P(T \leq t_c) = 1 - \alpha$ şi deci valoarea crtică $t_c = \chi^2_{1-\alpha}(k-1)$ este $(1-\alpha)$–cvantila distribuţiei χ^2 cu $k-1$ grade de libertate.

Valoarea testului χ^2 de concordanţă depinde de numărul k, al claselor alese. Pentru ca aproximaţia să fie validă trebuie ales un eşantion de volum, n, mare, iar subdivizarea să fie astfel încât în fiecare clasă (subinterval I_j) să cadă cel puţin 5 valori [**?**].

Avem deci următorul algoritm de testare a concordanţei datelor cu repartiţia F_0:

- Fixăm nivelul de semnificaţie α;
- se determină valoarea minimă, respectiv maximă dintre valorile de selecţie $x_1, x_2, \ldots, x_n$. Fie $a < b$ aceste valori;
- Se alege numărul k al claselor;
- Se fixează punctele de diviziune y_j, $j = \overline{0,k}$, ale intervalului $[a - \epsilon, b]$;
- Se determină numărul N_j al valorilor ce cad în intervalul $I_j = (y_{j-1}, y_j]$, $j = \overline{1,k}$;
- Se calculează probabilităţile $p_1 = F_0(y_1)$, $p_j = F_0(y_j) - F_0(y_{j-1})$, $j = 2, \ldots k-1$, $p_k = 1 - F_0(y_{k-1})$.
- Se calculează valoarea funcţiei test $t := \displaystyle\sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}$;
- dacă $t > \chi^2_{1-\alpha}(k-1)$ se respinge ipoteza că datele provin dintr-o lege de probabilitate de funcţie de repartiţie F_0.

În consideraţiile de mai sus s-a presupus că repartiţia teoretică este complet definită (nu depinde de parametri). În cazul în care repartiţia depinde de parametrii $\theta_1, \ldots, \theta_m$, aceştia se estimează din datele cercetării şi astfel probabilităţile p_j sunt reprezentate de estimatori ai lor p_j^*, $j = \overline{1,k}$, iar variabila aleatoare

$$T = \sum_{j=1}^{k} \frac{(N_j - np_j^*)^2}{np_j^*} \tag{1.67}$$

are aproximativ distribuţia $\chi^2(k-1-m)$ cu $k-1-m$ grade de libertate.

Exemplul 6. Numărul de întreruperi ale servirii apelurilor într-o reţea de comunicaţie în 200 de zile selectate aleator dintr-un an este sumarizat în următorul tablou:

Nr. întreruperi	0	1	2	3	4	≥ 5
Frecv. observată	64	71	42	18	4	1

Să testăm modelul Poisson pentru numărul de întreruperi X într-o singură zi.

Analizând tabelul ne sunt sugerate următoarele puncte de diviziune ale intervalului $[-1,5]$ $y_j = j-1$, $j=\overline{0,4}$, $y_5 = 5$. Astfel numărul N_j de valori ce întră în intervalul $I_j = (y_{j-1}, y_j]$, $j = \overline{1,4}$ sunt respectiv $64, 71, 42, 18$, iar $N_5 = 5$.

Ipoteza nulă este:

$$H_0 : \text{datele au distribuţia Poisson } p(x;\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}, \text{pentru } x \in \mathbb{N}$$

Conform Problemei **??**, Capitolul **??**, un estimator al parametrului λ al distribuţiei Poisson este media de selecţie. Deci:

$$\lambda^* = \frac{64\cdot 0 + 71\cdot 1 + 42\cdot 2 + 18\cdot 3 + 4\cdot 4 + 1\cdot 5}{200} = 1.15$$

Astfel avem distribuţia de probabilitate estimată:

$$p(x) = e^{-1.15}\frac{1.15^x}{x!} \quad \text{pentru } x \in \mathbb{N} \text{ şi } 0 \text{ în rest.}$$

Să estimăm şi probabilităţile p_j^*: $p_1^* = P(X \leq 0) = P(X=0) = e^{-1.15} = 0.31664$, $p_2^* = P(0 < X \leq 1) = P(X=1) = e^{-1.15}1.15 = 0.36413$, $p_3^* = P(1 < X \leq 2) = e^{-1.15}1.15^2/2 = 0.20938$, $p_4^* = P(2 < X \leq 3) = 0.080261$, $p_5^* = P(3 < X)$=$1 - P(X \leq 3)$= 0.32315. Valoarea funcţiei test:

$$t = \frac{(64-200p_1^*)^2}{200p_1^*} + \frac{(71-200p_2^*)^2}{200p_2^*} + \frac{(42-200p_3^*)^2}{200p_3^*} + \frac{(18-200p_4^*)^2}{200p_4^*} +$$

$$\frac{(5-200p_5^*)^2}{200p_5^*} = 1.0696 + 20.255 + 41.944 + 38.821 + 55.017 = 157.11$$

Fixând nivelul de semnificaţie la $\alpha = 0.05$, trebuie să calculăm $(1-\alpha)$–cvantila distribuţiei $\chi^2(k-1-m)$, unde $k=5$ şi numărul m al parametrilor este egal cu 1. 0.95–cvantila distribuţiei $\chi^2(3)$ calculată în `MATLAB` este
`chi2inv(0.95,3)`=7.8147. Cum $t > \chi^2_{0.95}(3)$ ipoteza nulă se respinge, deci datele nu provin din distribuţia Poisson.

1.8.2 Testul Kolmogorov–Smirnov

Dacă datele $x_1, x_2, \ldots, x_n$ sunt valori de observaţie asupra v.a. continue, independente şi identic distribuite $X_1, X_2, \ldots, X_n$ atunci ipoteza nulă este:

$$H_0 : \quad \text{funcţia de repartiţie a legii de probabilitate a datelor este} \quad F(x),$$

unde F este o funcţie continuă. Testul statistic este în acest caz maximumul dintre valoarea absolută a diferenţei dintre funcţia empirică de repartiţie, asociată datelor şi funcţia teoretică de repartiţie:

$$D_n = \max_{x\in\mathbb{R}} |F_e(x) - F(x)| \tag{1.68}$$

D_n se numeşte testul statistic Kolmogorov–Smirnov. Pentru a-l calcula pe D_n se ordonează datele crescător. Fie $x'_1, x'_2, \ldots, x'_n$ datele în ordine crescătoare. Funcţia empirică de repartiţie asociată datelor va fi:

$$F_e(x) = \begin{cases} 0 & \text{pt. } x < x'_1 \\ \dfrac{1}{n} & \text{pt. } x'_1 \le x < x'_2 \\ \vdots & \\ \dfrac{j}{n} & \text{pt. } x'_j \le x < x'_{j+1} \\ \vdots & \\ 1 & \text{pt. } x'_n \le x \end{cases} \tag{1.69}$$

Observăm că funcţia F_e este constantă pe intervalele $[x'_j, x'_{j+1})$ şi efectuează un salt de $1/n$ în punctele x'_j. Deoarece repartiţia teoretică este crescătoare şi mărginită de 1, valoarea maximă a diferenţei $F_e(x) - F(x)$ este nenegativă şi este atinsă în unul din punctele x'_j, $j = \overline{1,n}$ (Fig.1.6). Prin urmare

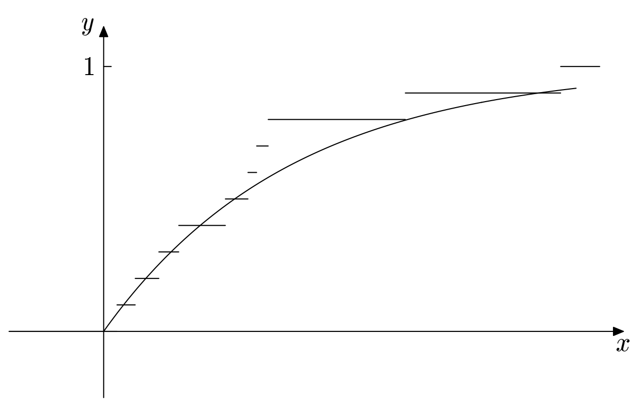

Fig.1.6: Graficele funcţiei de repartiţie teoretică (funcţia continuă) şi empirică (funcţia în scară).

$$\max_x\{F_e(x) - F(x)\} = \max_{j=\overline{1,n}} \left\{\frac{j}{n} - F(x'_j)\right\} \tag{1.70}$$

Similar, valoarea maximă $\max_x\{F(x) - F_e(x)\}$ este de asemenea nenegativă şi coincide cu:

$$\max_{j=\overline{1,n}} \left\{F(x'_j) - \frac{j-1}{n}\right\} \tag{1.71}$$

În concluzie avem:

$$\begin{aligned} &D_n = \max_x |F_e(x) - F(x)| = \\ &\max\{\max_x\{F_e(x) - F(x)\}, \max_x\{F(x) - F_e(x)\}\} = \\ &\max_{j=\overline{1,n}} \left\{ \frac{j}{n} - F(x'_j), F(x'_j) - \frac{j-1}{n} \right\} \end{aligned} \tag{1.72}$$

Dacă ipoteza H_0 este adevărată, D_n nu va depăşi o valoare d_c pe care o fixăm astfel încât eroarea de ordin 1, α, să fie foarte mică:

$$P(D_n > d_c | H_0) = \alpha \tag{1.73}$$

Pentru a determina această valoare critică se foloseşte o teoremă a lui Kolmogorov care afirmă că pentru orice funcţie continuă de repartiţie are loc relaţia:

$$\lim_{n\to\infty} P\left(D_n \leq \frac{\lambda}{\sqrt{n}}\right) = k(\lambda), \tag{1.74}$$

unde $\lambda > 0$, este arbitrar, iar funcţia $k(\lambda)$ este definită prin:

$$k(\lambda) = 1 - 2\sum_{m+1}^{\infty}(-1)^{m-1}e^{-2m^2\lambda^2} \tag{1.75}$$

Astfel luând pragul critic de forma $d_c = \lambda/\sqrt{n}$, rezultă că atunci când ipoteza nulă este adevărată avem:

$$P(D_n > d_c \,|\, H_0) = 1 - P(D_n \leq d_c \,|\, H_0) = 1 - P(D_n \leq \frac{\lambda}{\sqrt{n}}) \approx 1 - k(\lambda) = \alpha. \tag{1.76}$$

Unui nivel de semnificaţie α îi corespunde prin relaţia $k(\lambda) = 1 - \alpha$, un prag critic $d_c = \dfrac{\lambda_\alpha}{\sqrt{n}}$. Valoarea λ_α sau $\lambda_\alpha/\sqrt{n}$ se poate găsi fie din tabele, fie apelând o funcţie corespunzătoare într-un software statistic.

Dacă valoarea testului D_n asociată datelor de selecţie este mai mare decât $\lambda_\alpha/\sqrt{n}$ atunci nu există concordanţă între date şi repartiţia teoretică şi astfel ipoteza nulă este respinsă.

`MATLAB` pune la dispoziţie funcţia `kstest` (abreviere pentru Kolmogorov–Smirnov test) care pe baza datelor, a repartiţiei teoretice şi a nivelului de semnificaţie α poate returna valoarea testului D_n şi valoarea critică d_c.

Exemplul 7. Timpul de viaţă, în ore, a 10 componente electronice de acelaşi tip, selectate aleator din producţia unei luni este:

$$145, 81, 112, 66, 155, 140, 72, 94, 124, 116$$

Testaţi dacă aceste date sunt în concordanţa cu distribuţia exponenţială de parametru $\theta = 100$ ore.

Ipoteza $H_0 : F(x) = 1 - e^{-x/100}$;

Ipoteza $H_a : F(x) \neq 1 - e^{-x/100}$;

Următoarea secvenţă `MATLAB` va conduce la o decizie:

```
x=[145,81,112,66,155,140,72,94,124,116];
y=sort(x);
P=expcdf(y,100);
Vcdf=[y',P'];
[H,Q,Dn,dc]=kstest(y,Vcdf,0.05,0);
```

Observăm că mai întâi datele sunt sortate. Apoi este evaluată funcţia de repartiţie teoretică în punctele $y_1 < y_2 \ldots < y_{10}$. Matricea de 10 linii şi 2 coloane `Vcdf` furnizează perechile $(y_j, F(y_j))$, $j = \overline{1, 10}$, funcţiei `kstest` care calculează valoarea testului `Dn`, determină pragul critic `dc` şi ia decizia codificată prin 0 sau 1 şi o atribuie lui `H`. Dacă `H`=1, ipoteza nulă este respinsă, iar dacă `H`=0 datele sunt în concordanţă cu repartiţia teoretică. În cazul dat obţinem: `H`=1, `Q`=0.011752 (p-valoarea testului), `Dn`= 0.48315 şi valoarea critică `dc`=0.40925. Deoarece `Dn`$>$`dc` ipoteza nulă este respinsă.

Mai precizăm că argumentele funcţiei `kstest` sunt în ordinea scrisă: vectorul valorilor sortate crescător, matricea cu două coloane `Vcdf`, nivelul de semnificaţie 0.05 şi codul 0 pentru testul K-S descris teoretic mai sus.

1.9 Probleme

1. O fabrică producătoare de produse farmaceutice adaugă un colorant învelişului protector al unor tablete, ce ar putea micşora turbiditatea tabletei. Pentru a testa influenţa colorantului asupra acestei caracteristici chimice, sunt alese la întâmplare 10 cutii ce sunt păstrate un timp într-un depozit şi apoi este măsurată turbiditatea tabletelor din cutii. Sunt înregistrate pentru fiecare cutie următoarele valori:

$$3.9, 4.1, 4.4, 4.0, 3.8, 4.0, 3.9, 4.3, 4.2, 4.4,$$

ce sunt valori de observaţie asupra unor v.a. $X_i \in N(0, 0.2^2)$, $i = 1, 2, \ldots, 10$. Fāra adăugarea colorantului turbiditatea este în medie de 4.0. Indică rezultatele sondajului că turbiditatea a crescut, la un nivel de semnificaţie $\alpha = 0.05$?

2. Pentru un laborator s-a achiziţionat un număr mare de unităţi dintr-un material. Firma producătoare susţine că, caracteristica de interes a materialului are distribuţia normală de medie 3.5 şi abatere standard necunoscută. După recepţia întregului lot sunt selectate la întâmplare 15 unităţi din materialul adus şi supuse probelor. Valorile înregistrate sunt:

$$\begin{array}{l} 2.72, 0.50, 3.72, 4.01, 1.43, 5.64, 5.64, 3.43, \\ 4.08, 3.81, 3.16, 4.80, 2.44, 7.42, 3.25 \end{array}$$

Evidenţiază aceste date, la nivelul de semnificaţie de 0.01 că media caracteristicii este diferită de 3?

3. Un utilaj dintr-o fabrică trebuie reparat dacă produce mai mult de 10% unităţi defecte din unităţile ce le produce într-o zi. Un eşantion aleator de 100 de unităţi din producţia dintr-o zi conţine 15 unităţi produs defecte şi inginerul de specialitate susţine că utilajul trebuie reparat. Rezultatul investigării eşantionului susţine decizia inginerului la un nivel de semnificaţie de 0.01?

Indicaţie: $H_0 : p \leq 0.10$, $H_a : p > 0.10$.

4. O firmă rezervă fonduri în bugetul anual pentru repararea unui nou aparat achiziţionat, pe baza presupunerii că costul lunar al reparaţiei este de 1.200.000 lei. Pentru a testa dacă presupunerea este realistă, se analizează costul reparaţiei lunare pentru $n = 10$ aparate similare pe care le posedă. Selectarea celor 10 aparate pentru care se culege informaţia necesară se face în mod aleator şi se obţine o medie a costului $\overline{x} = 1.290.000$ lei şi o dispersie de selecţie $s = 110.000$ lei. Deoarece este important, pentru stabilirea bugetului, atât un cost mai mic, cât şi unul mai mare, testatîi ipoteza $H_0 : m = 1.200.000$ contra ipotezei $H_a : m \neq 1.200.000$, la un nivel de semnificaţie de $\alpha = 0.05$, ştiind că costul reparaţiilor lunare are distribuţie normală.

Indicaţie: Eşantionul fiind redus, şi dispersia σ^2 necunoscută, se foloseşte testul Student.

5. Tensiunea la ieşire pentru un tip de circuit electric este specificată ca fiind 130. Este selectat aleator un eşantion de 40 de circuite, pentru care media voltajului este de 128.6 şi dispersia $s = 2.1$. Să se testeze ipoteza $H_0 : m = 130$ contra alternativei $H_a : m < 130$. Folosiţi nivelul de semnificaţie de 5%.

6. Din înregistrările departamentului de vănzări al YXU Computers reiese că vânzările săptămânale sunt de 5775 mii euro. Pentru a creşte vânzările, firma a început o campanie publicitară. După 15 săptămâni media vânzărilor a ajuns la 6350 mii euro cu o abatere standard de 850 mii euro. Trebuie să continue campania publicitară? Testaţi cu $\alpha = 0.005$.

7. Media de vârstă a automobilelor ce circulă pe şoselele din România este de 8.4 ani. Pentru un eşantion de 34 maşini selectate la întâmplare în parcarea unei companii s-a identificat o medie de vârstă de 9.7 ani, cu o abatere standard de 3.1 ani. La nivelul de semnificaţie de 0.01 putem concluziona că media de vârstă a maşinilor din parcare este mai mare decât media pe ţară?

8. La recepţia unei mari cantităţi dintr-un produs beneficiarul selectează la întâmplare 300 de bucăţi. Testele de calitate relevă că 4% din totalul recepţionat prezintă defecte. Furnizorul susţine că doar 2% din cantitatea livrată prezintă defecte. Formulaţi ipoteza nulă şi alternativă adecvată, efectuaţi testul potrivit şi comentaţi credibilitatea furnizorului.

9. Un reprezentat al *Electrica* susţine că 65% dintre proprietarii de case preferă încălzirea pe bază de electricitate în locul celei pe gaz. Un studiu statistic relevă că 60% din 200 de proprietari de case preferă încălzirea cu curent electric în locul celei cu gaz. Relevă aceste date că ipoteza reprezentatului *Electrica* trebuie respinsă în favoarea alternativei $H_a : p \neq 0.65$, la un nivel de semnificaţie de 0.05?

10. Din înregistrările Spitalului S, rezultă că 52 de bărbaţi dintr-un eşantion de 1000 selectaţi din internaţii în ultimii 5 ani şi 23 de femei din 1000 internate în aceeaşi perioadă au probleme cardiace. Prezintă aceste date suficientă evidenţă ca să indice o rată mai mare a bolilor de inimă la bărbaţii decât la femeile internate în Spitalul S? Folosiţi nivelul de semnificaţie de 5%.

Indicaţie: $H_0 : p_1 - p_2 = 0$, $H_a : p_1 - p_2 > 0$.

11. Un producător de maşini automate de spălat produce trei sorturi de maşini A, B, C, dintr-un model. Din înregistrările maşinilor vândute, din acest model, se selectează 1000 de poziţii şi se constată că din acestea 400 sunt de tipul A. Să se calculeze p–valoarea testului statistic ce se foloseşte pentru a decide dacă datele de selecţie oferă suficientă evidenţă pentru a conchide că mai mult de $1/3$ dintre clienţi preferă tipul de maşină A.

12. Pe o porţiune de autostradă unde viteza automobilelor este limitată la 90km/h se efectuează un control al vitezei pe un eşantion de 72 de automobile, selectate la întâmplare. Informaţiile sunt înregistrate în tabelul:

Viteza(km/h)	Nr. maşini
[40,50)	1
[50,60)	3
[60,70)	6
[70,80)	16
[80,90)	24
[90,100)	13
[100,110)	9

Notăm cu m şi σ, media şi abaterea standard a vitezei automobilelor ce rulează pe autostrada în considerare.

Să se definească testul ce permite să decidem dacă cu un nivel de semnificaţie de 5% viteza medie pe autostradă este în acest moment mai mare de $90km/h$.

13. Durata medie a unei investigaţii efectuată la Centrul Medical G este de 7 minute, cu o abatere standard de 2 minute. Unui medic nou angajat la Centru pare să-i ia mai mult timp investigarea unui pacient. Pentru a se convinge dacă media duratei unei investigaţii a crescut, managerul centrului selectează aleator 3 zile de lucru în care înregistrează numărul de pacienţi investigaţi de noul medic. Rezultă că în 7 ore acesta a consultat 56 de pacienţi. Folosind un nivel de semnificaţie de 5%:
a) Enunţaţi ipoteza nulă şi ipoteza alternativă pentru testul managerului;
b) Ce probabilitate calculaţi pentru a obţine p-valoarea testului;
c) Este testul semnificativ sau nu?

14. Într-o companie se intenţionează introducerea timpului flexibil de lucru, astfel încât angajaţii să lucreze 42 de ore pe săptămână. Reprezentanţii sindicatului susţin însă că acest sistem va duce la scăderea numărului de ore lucrate de către staff-ul companiei. Într-o săptămână tipică de lucru, 8 membri ai staff-ului selectaţi aletor, au lucrat, respectiv:

$$48, 32, 37, 35, 36, 42, 41, 40 \text{ ore}$$

Susţin aceste date ipoteza sindicatului la nivelul de semnificaţie de 5%, şi ştiind că numărul de ore lucrate de un angajat este normal distribuit? Care este p–valoarea testului?

15. Un producător de fibre sintetice afirmă că produsul lui are o caracteristică mai perforamntă decât fibrele naturale. Mostre de câte 10 fibre sintetice şi 10 fibre naturale au fost selectate aleator şi testate din punctul de vedere al caracteristicii în discuţie. Mediile şi dispersiile celor două eşantioane sunt date în tabelul:

Fibre naturale	Fibre sintetice
$\overline{x}_1 = 272$	$\overline{x}_2 = 335$
s_1^2=1636	s_2^2=1892

Susţin aceste date afirmaţia producătorului de fibre sintetice? Testaţi folosind $\alpha = 0.10$.

16. Se ştie că aproximativ unul din 10 fumători preferă ţigările C. După o campanie publicitară, au fost intervievaţi 200 de fumători pentru a determina nivelul de succes al campaniei. 26 dintre cei intervievaţi şi-au exprimat preferinţa pentru ţigările C. Evidenţiază aceste date că a crescut preferinţa pentru aceste ţigări? Pentru testare folosiţi nivelul $\alpha = 0.05$.

17. O cincime din maşinile mai vechi de 7 ani eliberează o mare cantitate de bioxid de carbon. În urma unei legi se interzice accesul în localităţi a maşinilor ce depăşesc un anumit nivel de monoxid de carbon eliberat. O firmă producătoare de dispozitive care ataşate motorelor reduc cantitatea de poluant, a efectuat o campanie publicitara pentru un dispozitiv. Pentru a testa efectele campaniei s-a ales un eşantion aleator de 50 de maşini ce circulă prin oraş şi a căror vârstă depăşeste 7 ani pentru a testa nivelul de poluare al acestora. 23 dintre acestea depăseau nivelul admis de poluare. Indică acest rezultat că proporţia maşinilor vechi de peste 7 ani care eliberează o mare cantitate de poluant a scăzut în urma apariţiei legii şi a campaniei publicitare? Testaţi ipoteza folosind nivelul de semnificaţie de 0.05.

18. Pentru a compara performanţele studenţilor din învăţământul de stat şi particular au fost colectate rezultatele testelor a două grupuri de studenţi. Pentru un eşantion de 106 studenţi din învâţamântul particular media este de $\overline{x} = 282$ puncte şi abaterea standard $s = 46$ puncte, iar pentru un eşantion de 151 de studenţi din învăţământul de stat, media este $\overline{y} = 265$ puncte şi abaterea standard $s = 68$.
a) La un nivel de semnificaţie de 1% par a fi rezultatele aceleaşi?
b) Care este p–valoarea testului statistic?

19. O companie farmaceutică îşi propune să testeze dacă Aspirina 1 este la fel de eficientă ca Aspirina 2. În acest scop este selectat un eşantion de 400 de persoane ce acuză dese dureri de cap. Li se administrează Aspirina 1 şi 260 dintre ei raportează că se simt mai bine într-o oră. Un alt eşantion de 350 de persoane este ales în mod independent de primul şi persoanelor selectate li se administrează Aspirina 2. 220 raportează eficacitatea acesteia în cel puţin o oră.
a) La nivelul de semnificaţie de 1% apare o diferenţă între eficacitatea celor două tipuri de aspirină?
b) Să se calculeze p–valoarea testului statistic.

20. În opinia băncii ROBank în prima parte a unei zile de afaceri sunt procesaţi 30 de clienţi pe oră. Pentru a avea o imagine mai clară asupra timpului cât le ia clienţilor săi să-şi efectueze tranzacţiile bancare, un angajat al băncii urmăreşte şi înregistrează numărul de clienţi şi timpul de servire. Pe perioada observării procesului de servire au fost procesaţi 74 de cienţi cu un timp mediu de servire de 3 minute şi o abatere standard de 1.5 minute. Care este timpul mediu de servire ipotetizat de banca ROBank? Informaţiile obţinute din observarea timpului de servire a celor 74 de clienţi contrazic la un nivel de semnificaţie de 5% opinia oficială a băncii privind timpul mediu de servire?
Indicaţie: $H_0 : m = 2$minute, $H_a : m > 2$, $n = 74 > 30$.