

Lucian MATICIUC

**INTRODUCERE ÎN
STATISTICĂ MATEMATICĂ**

Universitatea „Alexandru Ioan Cuza”

Iași – 2018

Cuprins

1	Statistică descriptivă	1
1.1	Variabile empirice	1
1.2	Reprezentarea grafică a variabilelor empirice	5
1.3	Momente statistice asociate unei variabile	6
1.3.1	Momente statistice asociate unei variabile empirice	6
1.3.2	Momente ale unei variabile aleatoare discrete	17
1.3.3	Momente ale unei variabile aleatoare continue	19
2	Elemente de teoria selecției și a estimației	23
2.1	Problema estimației	28
2.1.1	Estimări punctuale ale momentelor	29
2.1.2	Estimări punctuale ale parametrilor	31
2.1.3	Estimări prin intervale de încredere ale momentelor	41
3	Verificarea ipotezelor statistice	55
3.1	Teste pentru media unei selecții	56
3.1.1	Testul Z pentru media unei selecții	57
3.1.2	Testul Z pentru diferența mediilor a două selecții	59
3.1.3	Testul T pentru media unei selecții	63
3.2	Teste pentru dispersia unei selecții	67
3.2.1	Testul χ^2 pentru dispersia unei selecții	67
3.2.2	Testul F pentru raportul dispersiilor a două selecții	70
	Bibliografie	76

Capitolul 1

Statistică descriptivă

1.1 Variabile empirice

“Statistica este arta învățării din date” (Sheldon M. Ross, 2010).

“Statistica este disciplina care se ocupă cu colectarea, analiza și interpretarea datelor obținute din observarea unui experiment. Această disciplină are o structură coerentă bazată pe Teoria Probabilităților” (Karl Pearson, 1936).

Definiția 1.1 *Partea statisticii care se ocupă cu culegerea, înregistrarea, gruparea, descrierea și sumarizarea datelor se numește **statistică descriptivă**.*

Definiția 1.2 *Partea statisticii care se ocupă cu interpretarea și obținerea concluziilor din datele colectate în cadrul unei experiențe se numește **statistică inferențială**.*

Statistica este interesată în obținerea de informații despre o colecție (mulțime) de elemente. La baza statisticii stă noțiunea de probabilitate.

Definiția 1.3 *O mulțime de elemente ce posedă o trăsătură comună, și care se cercetează în statistică, poartă numele de **populație statistică** (colectivitate statistică). Elementele care alcătuiesc populația statistică se numesc **unități statistice** sau **indivizi**. Numărul de indivizi care alcătuiesc populația statistică determină **volumul populației**.*

Caracteristica (sau **variabila**) este o anumită proprietate urmărită la indivizii unei colectivități statistice și a cărei valoare se poate schimba de la un individ la altul în cadrul populației. Există caracteristici cantitative (cele care se pot măsura, ca vârsta, greutatea, etc.) și caracteristici calitative.

Datele pot proveni din observațiile unei singure caracteristici (sau variabile) sau, simultan, a două sau mai multor caracteristici. O **mulțime univariată de date** reprezintă datele obținute prin observarea unei singure variabile; de exemplu, ne interesează timpul de viață al unui tip de baterii utilizate într-un anumit fel. Avem o **mulțime bivariată de date** atunci când observațiile sunt făcute pentru două variabile simultan; de exemplu, ne interesează înălțimea și greutatea pentru fiecare jucător al echipelor de baschet, deci fiecare caracteristică este o pereche de date. **Date multivariate** avem atunci când observațiile sunt făcute simultan pentru mai mult de două variabile; de exemplu, se analizează sângele și ne interesează, pentru fiecare pacient, mai mulți indicatori simultan.

Definiția 1.4 Se numește **selecție** (*eșantion, sondaj*) o submulțime a populației, i.e. o colectivitate parțială de elemente extrase la întâmplare din cadrul populației.

Notăm valorile caracteristicii măsurate pe fiecare element al colectivității parțiale cu x_i , $i = \overline{1, n}$, unde n este volumul selecției (numărul indivizilor din selecție). Se presupune că alegerea celor n indivizi ai unui eșantion este făcută astfel încât toate subgrupurile de n indivizi din întreaga populație sunt egal probabile de a fi alese.

Selecția spunem că este repetată (cu întoarcere) dacă individul extras este reintrodus în colectivitate înainte de a se extrage următorul; în caz contrar, selecția este nerepetată (fără întoarcere). Dacă volumul selecției este foarte mic în raport cu volumul populației atunci nu se mai face distincția între cele două tipuri de selecție (aceasta se va considera repetată).

Să remarcăm faptul că statistica trebuie să se ocupe și cu dezvoltarea tehnicilor potrivite de colectare a datelor. Dacă aceasta nu este făcută corect, atunci analiza datelor nu poate oferi răspunsuri cu un nivel de încredere crescut.

Se numește **serie statistică**, asociată unei selecții de volum n , un tablou de forma

$$X : \begin{pmatrix} x_1 & x_2 & \cdots & x_k \\ n_1 & n_2 & \cdots & n_k \end{pmatrix}, \quad \text{cu } \sum_{i=1}^k n_i = n,$$

unde x_i reprezintă valorile caracteristicii măsurate (scrise în ordine crescătoare) iar n_i reprezintă frecvențele absolute corespunzătoare valorii x_i (adică numărul care arată de câte ori apare valoarea x_i în timpul selecției).

Definiția 1.5 Numim **variabilă empirică** (*de selecție*), notată pe scurt *v.e.*,

asociată unei selecții de volum n , un tablou de forma

$$(1.1) \quad X^* : \begin{pmatrix} x_1 & x_2 & \cdots & x_k \\ f_1 & f_2 & \cdots & f_k \end{pmatrix}, \quad \text{cu } \sum_{i=1}^k f_i = 1,$$

unde $f_i = \frac{n_i}{n}$ reprezintă frecvențele relative, corespunzătoare valorii x_i , ale variabilei empirice X^* , adică

$$f_i = \mathbb{P}(X^* = x_i), \quad i = \overline{1, n}.$$

Dacă volumul selecției nu este prea mare și fiecare valoare x_i apare o singură dată în timpul selecției, atunci variabila empirică mai poate fi reprezentată astfel

$$(1.2) \quad X^* : \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ 1/n & 1/n & \cdots & 1/n \end{pmatrix},$$

unde x_i reprezintă cele n valori distincte ale caracteristicii măsurate.

În cazul în care caracteristicile pot lua orice valoare dintr-un interval (mărginit) de numere reale sau volumul selecției este mare se va face o grupare a acestor valori pe intervale disjuncte (sau clase), de obicei egale, intervale închise la stânga și deschise la dreapta:

$$\begin{pmatrix} [a_0, a_1) & [a_1, a_2) & \cdots & [a_{k-1}, a_k) \\ n_1 & n_2 & \cdots & n_k \end{pmatrix}.$$

Variabila empirică X^* se va reprezenta atunci

$$(1.3) \quad X^* : \begin{pmatrix} c_1 & c_2 & \cdots & c_k \\ f_1 & f_2 & \cdots & f_k \end{pmatrix}, \quad \text{cu } \sum_{i=1}^k f_i = 1,$$

unde $f_i = \frac{n_i}{n}$, $c_i = \frac{a_{i-1} + a_i}{2}$ (valoarea centrală a clasei $[a_{i-1}, a_i)$), $i = \overline{1, k}$.

Frecvența absolută ne dă numărul indivizilor care au valoarea caracteristicii măsurate egală cu o valoare dată. Frecvența absolută cumulată ne dă numărul indivizilor care au valoarea caracteristicii măsurate mai mică decât o valoare dată.

Definiția 1.6 Se numește *frecvență absolută cumulată crescător*, respectiv *descrescător*, corespunzătoare valorii x_i , valorile

$$n_i \uparrow = \sum_{\substack{j=1 \\ x_j \leq x_i}}^i n_j, \quad n_i \downarrow = \sum_{\substack{j=i \\ x_j \geq x_i}}^k n_j, \quad i = \overline{1, k},$$

adică $n_i \uparrow = n_1 + n_2 + \cdots + n_i$, $n_i \downarrow = n_i + n_{i+1} + \cdots + n_k$.

Definiția 1.7 Se numește *frecvență relativă cumulată crescător*, respectiv *descrescător*, corespunzătoare valorii x_i , valorile

$$f_i \uparrow = \frac{n_i \uparrow}{n}, \quad f_i \downarrow = \frac{n_i \downarrow}{n}, \quad i = \overline{1, k}.$$

Funcția empirică de repartiție a v.e. X^* se notează cu $F_n(x)$ și este definită astfel:

Dacă X^* este dată de (1.1) atunci

$$F_n(x) = \begin{cases} 0, & x < x_1, \\ \sum_{j=1}^{i-1} f_j, & x_{i-1} \leq x < x_i, \quad i = \overline{2, k}, \\ 1, & x_k \leq x. \end{cases}$$

$$= \begin{cases} 0, & x < x_1, \\ f_1, & x_1 \leq x < x_2, \\ f_1 + f_2, & x_2 \leq x < x_3, \\ \dots & \\ f_1 + f_2 + \dots + f_{k-1}, & x_{k-1} \leq x < x_k, \\ 1, & x_k \leq x. \end{cases}$$

Dacă X^* este dată de (1.2) atunci

$$F_n(x) = \begin{cases} 0, & x < x_1, \\ \frac{i-1}{n}, & x_{i-1} \leq x < x_i, \quad i = \overline{2, n}, \\ 1, & x_n \leq x. \end{cases}$$

Dacă X^* este dată de (1.3) atunci

$$F_n(x) = \begin{cases} 0, & x < a_0, \\ \sum_{j=1}^{i-1} f_j + \frac{x - a_{i-1}}{h} f_i, & a_{i-1} \leq x < a_i, \quad i = \overline{2, k}, \\ 1, & a_k \leq x, \end{cases}$$

unde $h = a_{i+1} - a_i$ este amplitudinea clasei (care de obicei este constantă).

1.2 Reprezentarea grafică a variabilelor empirice

Graficul unei v.e. se numește diagramă. Reprezentarea grafică se poate face în diverse moduri.

Poligonul frecvențelor absolute. Se iau pe abscisa Ox valorile x_i și pe Oy frecvențele absolute n_i corespunzătoare valorilor x_i . Unind aceste puncte vom obține poligonul frecvențelor absolute.

Reprezentarea cu bare. Se iau pe abscisa Ox valorile x_i iar în dreptul fiecărei valori x_i se ridică câte o perpendiculară de lungime egală cu valoarea frecvențelor absolute n_i (sau relative f_i) corespunzătoare lui x_i . Menționăm că dacă unim vârfurile acestor perpendiculare prin segmente vom obține poligonul frecvențelor absolute (sau respectiv relative). Evident, suma lungimilor segmentelor obținute în cazul folosirii frecvențelor relative trebuie să fie 1.

De asemenea, reprezentarea poate fi făcută și cu ajutorul unor dreptunghiuri. Mai precis, pe axa absciselor Ox se consideră segmente de tipul $[x_i - c, x_i + c]$ și pe fiecare segment de acest fel, considerat ca bază, se ridică câte un dreptunghi a cărui înălțime este egală cu frecvența corespunzătoare (absolută sau relativă) acelei valori x_i . Menționăm că dacă unim mijloacele laturilor superioare ale acestor dreptunghiuri vom obține poligonul frecvențelor.

Histograma. Această reprezentare se aseamănă cu **Reprezentarea cu bare**, dar se folosește în cazul unei v.e. ale cărei valori sunt numeroase sau sunt de tip continuu. Pentru a construi o histogramă, primul pas este de a împărți intervalul de valori într-o serie de intervale și apoi să numărăm câte valori intră în fiecare interval. Clasele de valori sunt de obicei specificate ca intervale consecutive, de pe axa absciselor, și care nu se suprapun. Intervalele nu trebuie să fie neapărat de dimensiuni egale. Un dreptunghi este ridicat peste acele intervale cu o înălțime egală cu frecvența relativă, adică proporțională cu numărul de cazuri din fiecare acea clasă. Evident, suma ariilor dreptunghiurilor obținute trebuie să fie 1.

Exercițiul 1.8 Să se reprezinte grafic următoarele date statistice care reprezintă numărul de zile de concediu medical luate de 50 de angajați ai unei companii în ultimele 6 săptămâni:

2, 2, 0, 0, 5, 8, 3, 4, 1, 0, 0, 7, 1, 7, 1, 5, 4, 0, 4, 0, 1, 8, 9, 7, 0
1, 7, 2, 5, 5, 4, 3, 3, 0, 0, 2, 5, 1, 3, 0, 1, 0, 2, 4, 5, 0, 5, 7, 5, 1

Obținem seria

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 12 & 8 & 5 & 4 & 5 & 8 & 0 & 5 & 2 & 1 \end{pmatrix}.$$

1.3 Momente statistice asociate unei variabile

Momente statistice asociate unei variabile sunt niște date numerice care reprezintă fidel o variabilă aleatoare sau o caracteristică avută în vedere. Cunoașterea momentelor statistice este utilă în practică în compararea a două variabile aleatoare sau a două populații statistice pe care este definită aceeași caracteristică, precum și la deducerea legii teoretice urmată de o variabilă aleatoare sau de caracteristica considerată.

1.3.1 Momente statistice asociate unei variabile empirice

Să considerăm o caracteristică cantitativă reprezentată de seria statistică

$$\begin{pmatrix} x_1 & x_2 & \cdots & x_k \\ n_1 & n_2 & \cdots & n_k \end{pmatrix} \text{ cu } \sum_{i=1}^k n_i = n \text{ (volumul selecției).}$$

Acestei serii îi asociem variabila empirică (de selecție)

$$(1.4) \quad X^* : \begin{pmatrix} x_1 & x_2 & \cdots & x_k \\ f_1 & f_2 & \cdots & f_k \end{pmatrix}, \text{ cu } \sum_{i=1}^k f_i = 1,$$

unde $f_i = \frac{n_i}{n}$ (frecvențele relative ale valorii x_i), $i = \overline{1, k}$, sau (dacă datele sunt grupate în clase de forma $[a_{i-1}, a_i)$, de lungimi egale)

$$(1.5) \quad X^* : \begin{pmatrix} c_1 & c_2 & \cdots & c_k \\ f_1 & f_2 & \cdots & f_k \end{pmatrix}, \text{ cu } \sum_{i=1}^k f_i = 1,$$

unde $f_i = \frac{n_i}{n}$ iar $c_i = \frac{a_{i-1} + a_i}{2}$ (valoarea centrală a clasei $[a_{i-1}, a_i)$), $i = \overline{1, k}$.

Parametrii tendinței centrale: \bar{x} , m_e , m_0

Acești parametri au rolul de a evidenția poziția în jurul căreia se grupează ansamblul valorilor unei v.e.. Această poziție exprimată printr-un număr se numește poziție centrală. Ea poate fi evidențiată prin:

Momentul empiric de ordin r este valoarea

$$\mu'_r = \frac{1}{n} \sum_{i=1}^k n_i x_i^r = \sum_{i=1}^k f_i x_i^r,$$

dacă X^* este dată de (1.4)

și respectiv

$$\mu'_r = \frac{1}{n} \sum_{i=1}^k n_i c_i^r = \sum_{i=1}^k f_i c_i^r,$$

dacă X^* este dată de (1.5).

În particular, pentru $r = 1$, obținem

$$\bar{x} \stackrel{\text{def}}{=} \mu'_1 = \frac{1}{n} \sum_{i=1}^k n_i x_i = \sum_{i=1}^k f_i x_i,$$

respectiv

$$\bar{x} \stackrel{\text{def}}{=} \mu'_1 = \frac{1}{n} \sum_{i=1}^k n_i c_i = \sum_{i=1}^k f_i c_i.$$

Valoarea \bar{x} se va numi **media empirică** (sau media aritmetică). Într-adevăr, dacă x_i , cu $i = \overline{1, n}$, ar reprezenta toate valorile caracteristicii măsurate, chiar dacă se repetă, atunci \bar{x} reprezintă efectiv media aritmetică a tuturor celor n valori, i.e. $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$.

Mediana (notată cu m_e) este valoarea caracteristicii X^* care împarte volumul selecției în două părți egale.

Dacă n este par, $n = 2k$, atunci $m_e = \frac{x_k + x_{k+1}}{2}$.

Dacă n este impar, $n = 2k + 1$, atunci $m_e = x_{k+1}$.

Moda (sau **valoarea modală**) (notată cu m_0) este valoarea caracteristicii X^* căreia îi corespunde frecvența relativă cea mai mare.

Cuantilele de ordin n sunt valorile caracteristicii X^* care împart volumul selecției în n părți egale. Cuantila de ordin 2 este chiar mediana și este acel număr $x_{1/2}$ care verifică ecuația

$$F_n(x_{1/2}) = 1/2,$$

unde F_n este funcția empirică de repartiție asociată v.e. X^* .

Cuantilele de ordin 4 se numesc **cuartile** și sunt acele valori $x_{1/4}$, $x_{1/2}$, $x_{3/4}$ pentru care

$$F_n(x_{1/4}) = 1/4, \quad F_n(x_{1/2}) = 1/2, \quad F_n(x_{3/4}) = 3/4.$$

Parametrii variabilității (ai împrăștierii): $R, s^2, s, (s^*)^2, s^*$

Amplitudinea v.e. (sau a seriei statistice) este numărul

$$R = x_{\max} - x_{\min}.$$

Momentul centrat empiric de ordin r este

$$\nu'_r = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^r = \sum_{i=1}^k f_i (x_i - \bar{x})^r.$$

În particular, pentru $r = 2$ obținem

Dispersia (sau varianța) empirică:

$$(1.6) \quad s^2 \stackrel{\text{def}}{=} \nu'_2 = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^2 = \sum_{i=1}^k f_i (x_i - \bar{x})^2.$$

Are loc următoarea relație care reprezintă o formulă utilă de calcul a dispersiei.

Propoziția 1.9 *Avem*

$$s^2 = \overline{x^2} - \bar{x}^2,$$

unde \bar{x} reprezintă media v.e. X^* iar $\overline{x^2}$ reprezintă media v.e. $(X^*)^2$.

Demonstrație. Într-adevăr,

$$\begin{aligned} s^2 &= \sum_{i=1}^k f_i (x_i - \bar{x})^2 = \sum_{i=1}^k f_i x_i^2 - 2 \sum_{i=1}^k f_i x_i \bar{x} + \sum_{i=1}^k f_i \bar{x}^2 \\ &= \sum_{i=1}^k f_i x_i^2 - 2\bar{x} \sum_{i=1}^k f_i x_i + \bar{x}^2 \sum_{i=1}^k f_i \\ &= \sum_{i=1}^k f_i x_i^2 - 2\bar{x} \bar{x} + \bar{x}^2 = \sum_{i=1}^k f_i x_i^2 - \bar{x}^2 = \overline{x^2} - \bar{x}^2. \end{aligned}$$

■

Mărimea

$$s = \sqrt{s^2}$$

se numește **abaterea medie pătratică empirică** (sau **deviația standard empirică**).

Dispersia (sau varianța) empirică modificată este numărul

$$(1.7) \quad (s^*)^2 = \frac{n}{n-1} s^2 = \frac{\sum_{i=1}^k n_i (x_i - \bar{x})^2}{n-1},$$

unde n este volumul selecției.

Avem deci și formula de calcul

$$(1.8) \quad (s^*)^2 = \frac{\overline{nx^2} - n\bar{x}^2}{n-1} = \frac{n \frac{\sum_{i=1}^k n_i x_i^2}{n} - n\bar{x}^2}{n-1} = \frac{\sum_{i=1}^k n_i x_i^2 - n\bar{x}^2}{n-1}.$$

Abaterea empirică modificată (sau **deviația standard empirică modificată**) este

$$s^* \stackrel{def}{=} \sqrt{(s^*)^2}.$$

O altă caracteristică importantă este

$$s_{\bar{x}}^* \stackrel{def}{=} \pm \frac{s^*}{\sqrt{n}}.$$

Caracteristicile formei

Aceste caracteristici se referă la forma poligonului frecvențelor absolute și relative (sau a curbei de repartiție în cazul variabilei aleatoare continue).

Boltirea se măsoară prin coeficientul de exces

$$E_X = \alpha_4 - 3, \quad \text{unde } \alpha_4 = \frac{\nu'_4}{s^4}.$$

Acest coeficient se mai notează și cu $\bar{\gamma}_1$ și măsoară gradul de turtire al poligonului frecvențelor relative sau al curbei repartiție față de repartiția normală.

Menționăm că o variabilă aleatoare repartizată normal $X \sim \mathcal{N}(m, \sigma^2)$ are $E_X = 0$. Într-adevăr, vezi calculul de la *Teoria Probabilităților*,

$$\nu'_4 \stackrel{def}{=} \mathbb{E}[X - \mathbb{E}(X)]^4 = 3\sigma^4$$

iar dispersia este $s^2 = \sigma^2$, deci

$$E_X = \frac{\nu'_4}{s^4} - 3 = \frac{3\sigma^4}{(\sigma^2)^2} - 3 = 0.$$

Din acest motiv graficul repartiției normale este curba cu care se compară toate repartițiile.

Dacă $E_X > 0$ sau echivalent $\alpha_4 > 3$ atunci curba este mai ascuțită decât curba corespunzătoare densității repartiției normale.

Dacă $E_X < 0$ sau echivalent $\alpha_4 < 3$ atunci curba este mai turtită decât curba corespunzătoare densității repartiției normale.

Asimetria se măsoară prin coeficientul de asimetrie

$$\bar{\gamma}_2 = \frac{\nu'_3}{s^3},$$

adică $\bar{\gamma}_2 = \alpha_3$. Menționăm că abaterea s este întotdeauna pozitivă (fiind radicalul dispersiei) iar ν'_3 poate fi pozitiv sau negativ după cum abaterile $x_i - \bar{x}$ care predomină sunt pozitive, respectiv negative. Repartiția statistică normală are $\bar{\gamma}_2 = 0$ sau echivalent $\alpha_3 = 0$.

Dacă $\bar{\gamma}_2 < 0$ atunci repartiția este cu asimetrie negativă (curba prezintă asimetrie spre stânga), iar dacă $\bar{\gamma}_2 > 0$ atunci repartiția este cu asimetrie pozitivă.

Evident simetria curbei este dată de raportarea la dreapta $x = \bar{x}$. Curba repartiției normale $X \sim \mathcal{N}(m, \sigma^2)$ are drept axă de simetrie dreapta $x = m = \mathbb{E}(X)$.

Exercițiul 1.10 Să presupunem că un aparat de măsurare este utilizat pentru a citi o distanță de 12 de ori. Se obțin valorile:

0.20, 0.10, 0.35, 0.25, 0.13, 0.20, 0.10, 0.20, 0.25, 0.20, 0.30, 0.35.

Datele sunt colectate în tabelul de mai jos:

$$\begin{pmatrix} 0.10 & 0.13 & 0.20 & 0.25 & 0.30 & 0.35 \\ 2 & 1 & 4 & 2 & 1 & 2 \end{pmatrix}$$

Obținem deci

$$X^* : \begin{pmatrix} 0.10 & 0.13 & 0.20 & 0.25 & 0.30 & 0.35 \\ 2/12 & 1/12 & 4/12 & 2/12 & 1/12 & 2/12 \end{pmatrix}$$

Amplitudinea este $0.35 - 0.10 = 0.25$.

Mediana este o valoare situată între a șasea și a șaptea, adică media aritmetică $\frac{0.20+0.25}{2} = 0.225$.

Moda (valoarea modală) este 0.20.

Media de selecție (sau media aritmetică) este dată de

$$\bar{x} = \sum_{i=1}^6 f_i x_i = 0.22$$

sau echivalent

$$\bar{x} = \frac{\sum_{i=1}^6 n_i x_i}{20}.$$

Dispersia (sau varianța) empirică este dată de formula

$$s^2 \stackrel{\text{def}}{=} \frac{1}{12} \sum_{i=1}^6 n_i (x_i - \bar{x})^2 = \sum_{i=1}^6 f_i (x_i - \bar{x})^2 = 0.00643,$$

deci abaterea empirică este $s = \sqrt{s^2} \simeq 0.0802$.

Pe de altă parte dispersia empirică modificată este numărul

$$(s^*)^2 = \frac{12}{11} s^2 = \frac{12}{11} 0.00643 = 0.0070244.$$

În plus abaterea empirică modificată este

$$s^* \stackrel{\text{def}}{=} \sqrt{(s^*)^2} \simeq 0.083811.$$

Mai trebuie făcut graficul poligonului frecvențelor relative.

Se poate scrie și funcția empirică de repartiție $F(x)$ (care este o funcție în scară).

Exercițiul 1.11 Să presupunem că un aparat de măsurare este utilizat pentru a citi o distanță de 20 de ori. Datele sunt colectate în tabelul de mai jos:

	22.7	25.4	22.0	20.5	22.5
(1.9)	22.3	24.2	24.7	23.5	23.1
	25.5	24.7	23.1	22.0	23.8
	23.8	24.4	23.7	23.8	22.6

Aceste citiri reprezintă mulțimea de date. O primă analiză a lor din punct de vedere numeric poate fi făcută calculând amplitudinea. Vedem din tabel ca amplitudinea este $25.5 - 20.5 = 5.0$.

Să considerăm în continuare datele de mai sus puse în ordine crescătoare.

	20.5	22.0	22.0	22.3	22.5
(1.10)	22.6	22.7	23.1	23.1	23.5
	23.7	23.8	23.8	23.8	24.2
	24.4	24.7	24.7	25.4	25.5

Putem determina imediat mediana. În cazul nostru mediana este dată de o valoare situată între a zecea și a unsprezecea valoare, adică media aritmetică $\frac{23.5+23.7}{2}$ (se poate considera drept mediană și una dintre cele două valori).

Moda este valoarea 23.8 (valoarea cu frecvența cea mai mare).

Variabila empirică (de selecție) X^* va avea tabloul

$$(1.11) \quad X^* : \begin{pmatrix} 20.5 & 22.0 & 22.3 & 22.5 & 22.6 & 22.7 & 23.1 & 23.5 & 23.7 \\ 0.05 & 0.1 & 0.05 & 0.05 & 0.05 & 0.05 & 0.1 & 0.05 & 0.05 \\ & & & 23.8 & 24.2 & 24.4 & 24.7 & 25.4 & 25.5 \\ & & & 0.15 & 0.05 & 0.05 & 0.1 & 0.05 & 0.05 \end{pmatrix}$$

Momentul empiric de ordin 1 (sau media empirică sau media aritmetică) este dată de

$$\begin{aligned} \bar{x} &= \sum_{i=1}^{15} f_i x_i \\ &= 0.05 \cdot 20.5 + 0.1 \cdot 22.0 + 0.05 \cdot 22.3 + 0.05 \cdot 22.5 + 0.05 \cdot 22.6 + 0.05 \cdot 22.7 \\ &\quad + 0.1 \cdot 23.1 + 0.05 \cdot 23.5 + 0.05 \cdot 23.7 + 0.15 \cdot 23.8 + 0.05 \cdot 24.2 + 0.05 \cdot 24.4 \\ &\quad + 0.1 \cdot 24.7 + 0.05 \cdot 25.4 + 0.05 \cdot 25.5 \\ &= 23.415 \end{aligned}$$

sau echivalent

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^{15} n_i x_i}{20} \\ &= \frac{20.5 + 2 \cdot 22.0 + 22.3 + 22.5 + 22.6 + 22.7 + 2 \cdot 23.1 + 23.5 + 23.7 + 3 \cdot 23.8}{20} \\ &\quad + \frac{24.2 + 24.4 + 2 \cdot 24.7 + 25.4 + 25.5}{20} = \frac{468.3}{20} = 23.415 \end{aligned}$$

unde x_i sunt valorile citite din tabloul (1.11).

Echivalent \bar{x} este media aritmetică a tuturor valorilor citite (valori ce se pot repeta),

$$\bar{x} = \frac{\sum_{i=1}^{20} x_i}{20},$$

unde x_i sunt valorile din tabelul (1.9).

Dispersia (sau varianța) empirică este dată de formula

$$s^2 \stackrel{\text{def}}{=} \frac{1}{20} \sum_{i=1}^{15} n_i (x_i - \bar{x})^2 = \sum_{i=1}^{15} f_i (x_i - \bar{x})^2.$$

Este util să scriem mai întâi un tabel cu diferențele¹ $x_i - \bar{x}$ și $(x_i - \bar{x})^2$:

x_i	Frecvența abs. n_i	Frecvența rel. f_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
20.5	1	0.05 = 1/20	-2.915	8.4972
22.0	2	0.1 = 2/20	-1.415	2.0022
22.3	1	0.05 = 1/20	-1.115	1.2432
22.5	1	0.05 = 1/20	-0.915	0.8372
22.6	1	0.05 = 1/20	-0.815	0.6642
22.7	1	0.05 = 1/20	-0.715	0.5112
23.1	2	0.1 = 2/20	-0.315	0.0992
23.5	1	0.05 = 1/20	0.085	0.0072
23.7	1	0.05 = 1/20	0.285	0.0812
23.8	3	0.15 = 3/20	0.385	0.1482
24.2	1	0.05 = 1/20	0.785	0.6162
24.4	1	0.05 = 1/20	0.985	1.97
24.7	2	0.1 = 2/20	1.285	1.6512
25.4	1	0.05 = 1/20	1.985	3.9402
25.5	1	0.05 = 1/20	2.085	4.3472
	20	1 = 20/20		

Deci, calculând obținem valoarea dispersiei empirice

$$s^2 = 1.4832$$

iar abaterea medie pătratică empirică este

$$s = \sqrt{s^2} = \sqrt{1.4832} = 1,2178.$$

Pe de altă parte dispersia empirică modificată este numărul

$$(s^*)^2 = \frac{n}{n-1} s^2 = \frac{20}{19} 1.4832 = 1.5612.$$

¹Diferențele de tipul $x_i - \bar{x}$ se numesc **deviația de la medie** iar suma tuturor, pentru $i = \overline{1, n}$, este nulă, deoarece

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} = n\bar{x} - n\bar{x} = 0.$$

În plus abaterea empirică modificată este

$$s^* \stackrel{def}{=} \sqrt{(s^*)^2} = 1.2494.$$

Remarca 1.12 În toate tabele și formulele de mai sus putem lăsa toate valorile x_i chiar dacă se repetă (deci $n = 20$ în acest caz). Atunci frecvența relativă a fiecărei valori va fi aceeași $f_i = 1/20 = 0.05$ și frecvența absolută a fiecărei valori va fi aceeași $n_i = 1$. Formula pentru s^2 devine

$$s^2 \stackrel{def}{=} \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{20} \sum_{i=1}^{20} (x_i - \bar{x})^2$$

iar

$$\begin{aligned} (s^*)^2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - 2 \sum_{i=1}^n x_i \bar{x} + \sum_{i=1}^n \bar{x}^2}{n-1} \\ &= \frac{\sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n\bar{x}^2}{n-1} = \frac{\sum_{i=1}^n x_i^2 - 2n\bar{x} \cdot \bar{x} + n\bar{x}^2}{n-1} \\ &= \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1} = \frac{\sum_{i=1}^{20} x_i^2 - 20\bar{x}^2}{n-1}, \end{aligned}$$

adică obținem următoarea formulă de calcul a dispersiei empirice modificate (vezi și formula (1.8)):

$$(1.12) \quad (s^*)^2 = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1}.$$

Dacă grupăm datele în clase de lungimi egale atunci lungimea clasei va fi amplitudinea împărțită la numărul claselor. Să împărțim datele în 5 clase. Atunci lungimea va fi de $5/5 = 1$. deci vom avea intervalele

Clasa	Frecvența absolută a clasei	Frecvența relativă a clasei
[20.5; 21.5)	1	0.05 = 1/20
[21.5; 22.5)	4	0.2 = 4/20
[22.5; 23.5)	5	0.25 = 5/20
[23.5; 24.5)	6	0.3 = 6/20
[24.5; 25.5)	4	0.25 = 4/20
		1 = 20/20

Remarca 1.13 (Justificarea definiției dispersiei empirice modificate (vezi (1.8) și (1.12) precum și Teorema 2.13).

Să notăm cu μ media întregii populații (o valoare teoretică ce, în general, nu poate fi determinată de fapt) iar σ^2 dispersia întregii populații care are volumul N , adică

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}.$$

Să considerăm un eșantion de volum n .

Are loc evident

$$(x_i - \mu)^2 = (x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - \mu) + (\bar{x} - \mu)^2$$

deci

$$\begin{aligned} \sum_{i=1}^n (x_i - \mu)^2 &= \sum_{i=1}^n (x_i - \bar{x})^2 + 2 \sum_{i=1}^n (x_i - \bar{x})(\bar{x} - \mu) + \sum_{i=1}^n (\bar{x} - \mu)^2 \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + 2(\bar{x} - \mu) \sum_{i=1}^n (x_i - \bar{x}) + n(\bar{x} - \mu)^2 \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + 2(\bar{x} - \mu) \left(\sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} \right) + n(\bar{x} - \mu)^2 \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + 2(\bar{x} - \mu)(n\bar{x} - n\bar{x}) + n(\bar{x} - \mu)^2 \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2. \end{aligned}$$

Obținem

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i - \mu)^2 - n(\bar{x} - \mu)^2, \quad i = \overline{1, n}.$$

Pe de o parte, avem că termenul $\sum_{i=1}^n (x_i - \mu)^2$ va fi, pentru n foarte mare (aproape de valoarea N), aproximat de $n\sigma^2$, adică $\sigma^2 \simeq \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$.

Pe de altă parte, $(\bar{x} - \mu)^2$ aproximează dispersia variabila aleatoare $\bar{X} \stackrel{def}{=} \frac{\sum_{i=1}^n X_i}{n}$ care este dată de

$$D^2(\bar{X}) = D^2\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{1}{n^2} \sum_{i=1}^n D^2(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n},$$

deci termenul $n(\bar{x} - \mu)^2$ va fi, pentru n foarte mare, aproximat de numărul $n \frac{\sigma^2}{n} = \sigma^2$.

Deci, pentru n foarte mare,

$$(s^*)^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \simeq \frac{n\sigma^2 - \sigma^2}{n-1} = \sigma^2,$$

adică **dispersia empirică modificată aproximează mai bine dispersia σ^2 a populației (decât dispersia empirică s^2 dată de definiția (1.6)).**

Exercițiul 1.14 Presupunem că au fost obținute următoarele valori în urma a 15 citiri a unei distanțe cu ajutorul unui aparat de măsurare:

212.22	212.25	212.23	212.15	212.23
212.11	212.29	212.34	212.22	212.24
212.19	212.25	212.27	212.20	212.25

Având în vedere că media \bar{X} este repartizată $\mathcal{N}(0, \sigma^2)$, menționăm că în acest caz avem relația $E'_\delta = E_\delta s^*$, cu $\delta \in (0, 1)$, unde E'_δ este dat de relația

$$\mathbb{P}(|X - \bar{x}| \leq E'_\delta) = \delta \quad \Leftrightarrow \quad \mathbb{P}(\bar{x} - E'_\delta \leq X \leq \bar{x} + E'_\delta) = \delta.$$

(i) Calculați media empirică, dispersia empirică modificată, abaterea empirică modificată precum și $E_{0.5}$, $E_{0.95}$.

(ii) Ce procent din cele 15 observații sunt în intervalul $(\bar{x} - s^*, \bar{x} + s^*)$? Interpretați rezultatul.

(iii) Ce procent din cele 15 observații se află în intervalul $(\bar{x} - E'_{0.5}, \bar{x} + E'_{0.5})$? dar în intervalul $(\bar{x} - E'_{0.95}, \bar{x} + E'_{0.95})$? Interpretați rezultatul.

Rezolvare:

(i) Obținem

$$\begin{aligned} \bar{x} &= \frac{1}{15} \sum_{i=1}^{15} x_i = \frac{3183.439}{15} = 212.23, \\ s^2 &= \frac{1}{15} \sum_{i=1}^{15} (x_i - \bar{x})^2 = \frac{0.0421}{15} = 0.002806, \quad s = \sqrt{s^2} = 0.05297, \\ (s^*)^2 &= \frac{15}{15-1} s^2 = \frac{1}{14} \sum_{i=1}^{15} (x_i - \bar{x})^2 = \frac{0.0421}{14} = 0.003007, \\ s^* &= \sqrt{(s^*)^2} = 0.05483 \end{aligned}$$

și

$$E'_{0.50} = 0.6745 \cdot s^* = 0.6745 \cdot 0.05483 = 0.03698$$

$$E'_{0.95} = 1.960 \cdot s^* = 1.960 \cdot 0.05483 = 0.10747.$$

(ii) Intervalul este

$$(\bar{x} - s^*, \bar{x} + s^*) = (212.23 - 0.0548, 212.23 + 0.0548) = (212.175, 212.284)$$

în care se găsesc 11 valori din eșantion, adică $\frac{11}{15} \cdot 100 = 73.33\%$.

Dar, teoretic, în intervalul $(\bar{x} - s^*, \bar{x} + s^*)$ se găsesc 68.26% dintre valori, deoarece

$$\mathbb{P}(|X| \leq \sigma) = 0.6826.$$

Diferența vine deoarece avem un număr mic de date. Cu cât volumul eșantionului va fi mai mare, cu atât numărul de valori din intervalul $(\bar{x} - s^*, \bar{x} + s^*)$ va fi mai aproape de procentul de 68.26% dintre valorile citite.

(iii) Intervalul este

$$\begin{aligned}(\bar{x} - E'_{0.50}, \bar{x} + E'_{0.50}) &= (212.23 - 0.03698, 212.23 + 0.03698) \\ &= (212.1930, 212.2669).\end{aligned}$$

Teoretic, în intervalul $(\bar{x} - E'_{0.50}, \bar{x} + E'_{0.50})$ se găsesc 50% dintre valori, deoarece

$$\mathbb{P}(|X| \leq E'_{0.50}) = 0.50.$$

1.3.2 Momente ale unei variabile aleatoare discrete

Fie X o variabilă aleatoare (v.a.) discretă cu un număr finit de valori, având tabloul de repartiție

$$X : \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}$$

unde $p_i \geq 0$, $i = \overline{1, n}$, și $\sum_{i=1}^n p_i = 1$.

Numărul

$$\mathbb{E}(X) = \sum_{i=1}^n p_i x_i$$

este **valoarea medie a v.a. X** sau **media v.a. X** .

Dacă X este o v.a. discretă cu un număr infinit de valori, având tabloul de repartiție

$$X : \begin{pmatrix} x_1 & x_2 & \cdots & x_n & \cdots \\ p_1 & p_2 & \cdots & p_n & \cdots \end{pmatrix}$$

unde $p_i \geq 0$, $i \in \mathbb{N}^*$, $\sum_{i=1}^{\infty} p_i = 1$, atunci media v.a. X este definită de

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} p_i x_i.$$

Propoziția 1.15 (Proprietăți ale mediei) Pentru orice $a, c \in \mathbb{R}$ și orice v.a. X, Y au loc următoarele:

(i) valoarea medie a unei constante este egală cu constanta:

$$X : \begin{pmatrix} c \\ 1 \end{pmatrix} \Rightarrow \mathbb{E}(X) = c;$$

(ii)

$$\mathbb{E}(a + X) = a + \mathbb{E}(X);$$

(iii)

$$\mathbb{E}(aX) = a\mathbb{E}(X);$$

(iv)

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y);$$

(v) *media produsului a două v.a. independente este produsul mediilor variabilelor considerate, i.e.*

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

Demonstrație. (v) Deoarece X, Y sunt independente avem că

$$\begin{aligned} \mathbb{E}(XY) &= \sum_{i,j} x_i y_j p_{ij} = \sum_{i,j} x_i y_j \mathbb{P}(X = x_i, Y = y_j) \\ &= \sum_{i,j} x_i y_j \mathbb{P}(\{X = x_i\} \cap \{Y = y_j\}) \\ &= \sum_{i,j} x_i y_j \mathbb{P}(X = x_i) \mathbb{P}(Y = y_j) = \sum_{i,j} x_i y_j p_i q_j = \sum_i x_i p_i \sum_j y_j q_j \\ &= \mathbb{E}(X)\mathbb{E}(Y). \end{aligned}$$

■

Se numește **moment inițial de ordin r** al v.a. X , media v.a. X^r . Vom nota

$$\mu_r \stackrel{def}{=} \mathbb{E}(X^r) = \sum_i p_i x_i^r.$$

Evident momentul inițial de ordin 1 este exact media v.a. X .

Se numește **moment absolut de ordin r** al v.a. X , media v.a. $|X|^r$. Vom nota

$$\lambda_r \stackrel{def}{=} \mathbb{E}(|X|^r) = \sum_i p_i |x_i|^r.$$

Se numește **valoarea medie de ordin r** al v.a. X , numărul

$$m_r = [\mathbb{E}(X^r)]^{1/r}.$$

În particular, obținem

$$(i) \quad m_1 = \mu_1 = \mathbb{E}(X).$$

$$(ii) \quad m_2 = \sqrt{\mu_2} = \sqrt{\sum_i p_i x_i^2} \quad (\text{este valoarea medie pătratică a lui } X).$$

$$(iii) \quad m_{-1} = [\mathbb{E}(X^{-1})]^{-1} = \frac{1}{\sum_i p_i \frac{1}{x_i}} = \frac{1}{\sum_i \frac{p_i}{x_i}} \quad (\text{este media armonică}).$$

$$(iv) \quad m_0 \stackrel{def}{=} \lim_{r \rightarrow 0} m_r = \lim_{r \rightarrow 0} [p_1 x_1^r + p_2 x_2^r + \dots + p_n x_n^r]^{1/r}.$$

Remarca 1.16 Calculăm valoarea $\ln m_0$ aplicând *L'Hospital*:

$$\begin{aligned} \ln(m_0) &= \lim_{r \rightarrow 0} \frac{\ln(p_1 x_1^r + p_2 x_2^r + \cdots + p_n x_n^r)}{r} \\ &= \lim_{r \rightarrow 0} \frac{[\ln(p_1 x_1^r + p_2 x_2^r + \cdots + p_n x_n^r)]'}{1} = \lim_{r \rightarrow 0} \frac{[p_1 x_1^r + p_2 x_2^r + \cdots + p_n x_n^r]'}{p_1 x_1^r + p_2 x_2^r + \cdots + p_n x_n^r} \\ &= \lim_{r \rightarrow 0} \frac{p_1 x_1^r \ln x_1 + p_2 x_2^r \ln x_2 + \cdots + p_n x_n^r \ln x_n}{p_1 x_1^r + p_2 x_2^r + \cdots + p_n x_n^r} \\ &= \frac{p_1 \ln x_1 + p_2 \ln x_2 + \cdots + p_n \ln x_n}{p_1 + p_2 + \cdots + p_n} = \frac{\ln[(x_1)^{p_1} (x_2)^{p_2} \cdots (x_n)^{p_n}]}{1} \\ &= \ln[(x_1)^{p_1} (x_2)^{p_2} \cdots (x_n)^{p_n}], \end{aligned}$$

deoarece $(x_1^r)'_r = (e^{\ln x_1^r})'_r = (e^{r \ln x_1})'_r = e^{r \ln x_1} \ln x_1 = x_1^r \ln x_1$.

Se numește **momentul centrat de ordin r** al v.a. X , media v.a. $(X - \mu_1)^r$.
Vom nota

$$\nu_r \stackrel{def}{=} \mathbb{E}[(X - \mu_1)^r] = \sum_i p_i (x_i - \mu_1)^r.$$

Se numește **dispersia** v.a. X , momentul centrat de ordin 2 al v.a. X , adică

$$\sigma^2 = D^2(X) \stackrel{def}{=} \nu_2 = \mathbb{E}(X - \mu_1)^2, \quad \text{unde } \mu_1 = \mathbb{E}(X).$$

Mediana v.a. X (notată cu m_e) este valoarea v.a. X care împarte valorile lui X în două părți egale:

$$m_e = \begin{cases} x_{k+1}, & \text{dacă } n = 2k + 1, \\ \frac{x_k + x_{k+1}}{2}, & \text{dacă } n = 2k. \end{cases}$$

Moda este valoarea pe care o ia X cu probabilitatea cea mai mare.

1.3.3 Momente ale unei variabile aleatoare continue

Fie X o v.a. continuă cu densitatea de probabilitate $f(x)$.

Numărul

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx$$

este **media v.a.** X .

Proprietățile mediei sunt aceleași din cazul discret (vezi Propoziția 1.15).

Se numește **moment inițial de ordin** r al v.a. X , media v.a. X^r . Vom nota

$$\mu_r \stackrel{\text{def}}{=} \mathbb{E}(X^r) = \int_{-\infty}^{\infty} x^r f(x) dx.$$

Evident momentul inițial de ordin 1 este exact media v.a. X .

În particular,

$$\mu_2 = \mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx.$$

Se numește **momentul centrat de ordin** r al v.a. X , media v.a. $(X - \mu_1)^r$. Vom nota

$$\nu_r \stackrel{\text{def}}{=} \mathbb{E}[(X - \mu_1)^r] = \int_{-\infty}^{\infty} (x - \mu_1)^r f(x) dx.$$

În particular,

$$\nu_2 = \mathbb{E}[(X - \mu_1)^2] = \int_{-\infty}^{\infty} (x - \mu_1)^2 f(x) dx.$$

Se numește **dispersia** v.a. X , momentul centrat de ordin 2 al v.a. X , adică

$$\sigma^2 = D^2(X) \stackrel{\text{def}}{=} \nu_2 = \mathbb{E}(X - \mu_1)^2, \text{ unde } \mu_1 = \mathbb{E}(X).$$

Remarca 1.17 Având în vedere calculul

$$\begin{aligned} D^2(X) &= \mathbb{E}(X - \mu_1)^2 = \int_{\mathbb{R}} (x - \mu_1)^2 f(x) dx \\ &= \int_{\mathbb{R}} x^2 f(x) dx - 2 \int_{\mathbb{R}} \mu_1 x f(x) dx + \int_{\mathbb{R}} \mu_1^2 f(x) dx \\ &= \int_{\mathbb{R}} x^2 f(x) dx - 2\mu_1 \int_{\mathbb{R}} x f(x) dx + \mu_1^2 \int_{\mathbb{R}} f(x) dx \\ &= \mathbb{E}(X^2) - 2\mu_1 \mathbb{E}(X) + \mu_1^2 = \mathbb{E}(X^2) - \mu_1^2, \end{aligned}$$

obținem o formulă foarte utilă în calcule:

$$D^2(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

Propoziția 1.18 (Proprietăți ale dispersiei) Pentru orice $a, c \in \mathbb{R}$ și orice v.a. X, Y au loc următoarele:

(i) dispersia unei constante este nulă,

$$D^2(c) = \mathbb{E}(c^2) - (\mathbb{E}(c))^2 = c^2 - c^2 = 0;$$

(ii)

$$D^2(aX) = a^2 D^2(X),$$

deoarece

$$D^2(aX) = \int_{\mathbb{R}} [ax - \mathbb{E}(aX)]^2 f(x) dx = a^2 \int_{\mathbb{R}} [x - \mathbb{E}(X)]^2 f(x) dx = a^2 D^2(X);$$

(iii) dispersia sumei a două v.a. independente este suma dispersiei variabilelor considerate

$$D^2(X + Y) = D^2(X) + D^2(Y).$$

Într-adevăr,

$$\begin{aligned} D^2(X + Y) &= \mathbb{E} \left((X + Y)^2 \right) - (\mathbb{E}(X + Y))^2 \\ &= \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}(Y^2) - (\mathbb{E}X)^2 - 2\mathbb{E}(X)\mathbb{E}(Y) - (\mathbb{E}Y)^2 \\ &= \mathbb{E}(X^2) + \mathbb{E}(Y^2) - (\mathbb{E}X)^2 - (\mathbb{E}Y)^2 = D^2(X) + D^2(Y). \end{aligned}$$

De obicei gradul de împrăștiere a valorilor unei v.a. X se exprimă nu prin dispersie ci prin **abaterea medie pătratică** notată $\sigma \stackrel{def}{=} D(X)$, și definită de

$$\sigma = D(X) = \sqrt{D^2(X)}.$$

Aceasta are avantajul că se exprimă prin aceleași unități de măsură ca și valorile v.a. X .

Propoziția 1.19 (Proprietăți ale abaterii medii pătratice) Pentru orice $a, c \in \mathbb{R}$ și orice v.a. X au loc următoarele:

(i) $D(c) = 0$.(ii) $D(aX) = |a| D(X)$.

Teorema 1.20 (Inegalitatea lui Cebâșev) Fie X o v.a. care admite media m și dispersia σ^2 finite. Atunci oricare ar fi $\varepsilon > 0$, are loc inegalitatea

$$\mathbb{P}(|X - m| < \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2}.$$

Evident, următoarea formă este echivalentă.

Corolarul 1.21 Fie X o v.a. care admite media m și dispersia σ^2 finite. Atunci oricare ar fi $\varepsilon > 0$, are loc inegalitatea

$$\mathbb{P}(|X - m| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

O formă des întâlnită în aplicații este următoarea consecință:

Corolarul 1.22 *Fie X o v.a. care admite media m și dispersia σ^2 finite. Atunci luând în inegalitatea lui Cebâșev $\varepsilon = k\sigma$, obținem inegalitatea*

$$\mathbb{P}(|X - m| < k\sigma) \geq 1 - \frac{1}{k^2}, \quad \text{pentru orice } k > 0$$

(sau echivalent $\mathbb{P}(|X - m| \geq k\sigma) \leq \frac{1}{k^2}$).

Capitolul 2

Elemente de teoria selecției și a estimății

Acest capitol face tranziția spre statistica inferențială. Dacă avem un eșantion de n observații din cadrul unei populații, dorim să obținem estimări pentru media μ a întregii populații, dispersia σ^2 a întregii populații, deviația standard σ a întregii populații etc. Dar având în vedere că eșantionul de n date este extras la întâmplare din populație este evident că media empirică \bar{x} , dispersia empirică s^2 și deviația standard empirică s vor fi diferite de la un eșantion la altul. Ne interesează comportamentul acestor estimatori la eșantioane diferite.

Presupunem că o populație are următoarele 100 de valori:

18.2	26.4	20.1	29.9	29.8	26.6	26.2	25.7	25.2	26.3
26.7	30.6	22.6	22.3	30.0	26.5	28.1	25.6	20.3	35.5
22.9	30.7	32.2	22.2	29.2	26.1	26.8	25.3	24.3	24.4
29.0	25.0	29.9	25.2	20.8	29.0	21.9	25.4	27.3	23.4
38.2	22.6	28.0	24.0	19.4	27.0	32.0	27.3	15.3	26.5
31.5	28.0	22.4	23.4	21.2	27.7	27.1	27.0	25.2	24.0
24.5	23.8	28.2	26.8	27.7	39.8	19.8	29.3	28.5	24.7
22.0	18.4	26.4	24.2	29.9	21.8	36.0	21.3	28.8	22.8
28.5	30.9	19.1	28.1	30.3	26.5	26.9	26.6	28.2	24.2
25.5	30.2	18.9	28.9	27.6	19.6	27.9	24.9	21.3	26.7

Media (notată μ) și dispersia (notată σ^2) acestei populații sunt 26.1 respectiv 17.5.

Alegând la întâmplare 10 valori din tabelul de mai sus putem obține o estimare a mediei și dispersiei (adică media și dispersia empirică, notate \bar{x} și s^2). Evident aceste valori vor fi estimatori ale valorilor teoretice μ și σ^2 (nu vor coincide cu acestea). De asemenea, prin selectarea altor 10 valori vom obține altă medie și dispersie \bar{x}, s^2 .

Dacă volumul selecției crește atunci este de așteptat ca \bar{x}, s^2 să se apropie de valorile teoretice μ și σ^2 (cu cât volumul se apropie mai mult de 100, cu atât \bar{x} și s^2 se apropie mai mult de $\mu = 26.1$ și $\sigma^2 = 17.5$).

În tabelul de mai jos putem vedea acest lucru (s-au luat la întâmplare selecții de volum 10, 20, etc. iar aceste selecții de diverse volume nu mai sunt menționate):

No.	\bar{x}	s^2
10	26.9	28.1
20	25.9	21.9
30	25.9	20.0
40	26.5	18.6
50	26.6	20.0
60	26.4	17.6
70	26.3	17.1
80	26.3	18.4
90	26.2	17.8
100	26.1	17.5

Având în vedere că \bar{x} și s^2 sunt calculate plecând de la niște variabile aleatoare (deoarece $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ iar x_i sunt alese aleator), obținem că \bar{x} și s^2 sunt și ele, la rândul lor, niște variabile aleatoare. Deci chiar dacă volumul n este menținut constant pot exista variații ale mediei și dispersiei empirice. Pentru aceasta vezi tabelul de mai jos:

Set 1:	29.9	18.2	30.7	24.4	36.0	25.6	26.5	29.9	19.6	27.9
Set 2:	26.9	28.1	29.2	26.2	30.0	27.1	26.5	30.6	28.5	25.5
Set 3:	32.2	22.2	23.4	27.9	27.0	28.9	22.6	27.7	30.6	26.9
Set 4:	24.2	36.0	18.2	24.3	24.0	28.9	28.8	30.2	28.1	29.0

cu

Set 1:	$\bar{x} = 26.9$	$s^2 = 28.1$
Set 2:	$\bar{x} = 27.9$	$s^2 = 2.9$
Set 3:	$\bar{x} = 26.9$	$s^2 = 10.9$
Set 4:	$\bar{x} = 27.2$	$s^2 = 23.0$

Fluctuațiile care se văd în tabelul de mai sus ridică următoarea problemă: în ce măsură valorile \bar{x} și s^2 estimează corect valorile reale ale mediei și dispersiei? Remarcăm că în primul și al treilea set media este mai apropiată de 26.1 dar dispersia este mare. În setul al doilea dispersia este mai mică dar în schimb media de 27.9 este destul de departe de valoarea 26.1. Evident sunt de preferat datele obținute în urma unei selecții de volum cât mai mare posibil.

Important în acest sens vor fi estimatorii (vezi și definiția dată de relația (2.1)), adică funcții care să depindă de fiecare selecție în parte.

În cadrul teoriei estimației foarte importante vor fi trei distribuții:

- distribuția “chi pătrat” de parametrii n și σ , notată $\chi^2(n, \sigma)$,
- distribuția Student de parametru n , notată $t(n)$ și
- distribuția Fisher de parametrii m și n , notată $F(m, n)$.

Pentru definiții ale acestora și diverse legături între distribuții vezi cursul de *Teoria Probabilităților*.

În general vorbind, fie \mathcal{P} o populație statistică și X o caracteristică cantitativă relativă la \mathcal{P} . În cele mai multe cazuri repartiția teoretică a lui X nu este cunoscută. Scopul statisticii este acela de a determina pe baza experiențelor cu elemente din \mathcal{P} (selecțiilor din \mathcal{P}) a legii de repartiție a lui X precum și a anumitor momente ale lui X (de exemplu, media și dispersia). Acest lucru este posibil aplicând metoda selecției sau a eșantioanelor.

Dacă numărul de elemente al mulțimii \mathcal{P} este notat N , atunci în urma a n experiențe obținem rezultatele x_1, \dots, x_n , unde n este mult mai mic decât N și reprezintă volumul selecției.

O selecție poate fi repetată (cu întoarcere) sau fără întoarcere (adică dacă elementul cercetat se pune la loc în populație sau nu).

Pentru a reflecta fidel proprietățile întregii populații, o selecție trebuie să îndeplinească următoarele condiții: \mathcal{P} să fie cât mai omogenă; n să fie cât mai mare; unitățile selecției să fie extrase la întâmplare; fiecare unitate din \mathcal{P} să aibă aceeași probabilitate (șansă) de a face parte din selecție.

În cadrul unei populații \mathcal{P} considerăm, mai întâi, o variabilă aleatoare teoretică X , necercetată direct, care se referă la \mathcal{P} în totalitate, și apoi o variabilă aleatoare empirică (de selecție) X^* , ce ia valorile x_1, \dots, x_n , adică

$$X^* : \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ 1/n & 1/n & \cdots & 1/n \end{pmatrix}.$$

Dar valorile x_1, \dots, x_n sunt, teoretic vorbind, diferite de la un eșantion la altul. Deci, având în vedere că pentru un eșantion de volum n ales la întâmplare valorile x_1, \dots, x_n sunt și ele incerte (înainte de citi efectiv datele), putem considera x_1, \dots, x_n ca valori pe care le iau n variabilele aleatoare independente X_1, \dots, X_n , ce au aceeași repartiție și anume repartiția lui X . Deci, de exemplu, X_1 este variabila aleatoare care are drept valori caracteristica x_1 a primei unități statistice dintr-un eșantion ales la întâmplare și de volum n .

Atunci valorile x_1, \dots, x_n , observate în urma selecției, constituie valoarea observată a vectorului aleator n -dimensional (X_1, \dots, X_n) . Repetând selecția vom obține diferite valori ale vectorului (X_1, \dots, X_n) .

Remarca 2.1 Fiecare moment empiric obținut pe baza selecției este valoarea unei anumite v.a. teoretice, valoare care variază odată cu selecția.

De exemplu, media empirică

$$\bar{x} = \frac{x_1 + \dots + x_n}{n}$$

reprezintă valoarea v.a.

$$\bar{X} \stackrel{\text{def}}{=} \frac{X_1 + \dots + X_n}{n},$$

și de aceea putem vorbi de \bar{x} ca de o v.a.

De asemenea, dispersia empirică

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

este valoarea v.a.

$$S^2 \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Problema esențială este următoarea: în ce măsură anumite momente ale lui X^* (media, dispersia etc.) pot reprezenta momente corespunzătoare pentru v.a. X . Astfel putem vorbi de două direcții de studiu: determinarea tipului de repartiție teoretică și a parametrilor corespunzători; și de determinarea unor indicatori numerici pentru v.a. teoretică X .

Să menționăm că uneori tipul de repartiție teoretică se cunoaște din experiențele anterioare. Alteori tipul de repartiție teoretică se intuiește din reprezentarea grafică a lui X^* , după care se face o ipoteză asupra legii, lucrând în această ipoteză (se vor verifica ipotezele statistice).

Definiția 2.2 Numim *statistică* (sau *funcție de selecție*) orice cantitate a cărei valoare poate fi calculată din datele unui eșantion. Având în vedere că valoarea unei statistici depinde de alegerea eșantionului, vedem statistica ca pe o variabilă aleatoare (și prin urmare va fi notată cu literă mare).

De exemplu, media empirică se va nota cu \bar{X} iar valoarea ei corespunzătoare unui anume eșantion este \bar{x} ; similar, dispersia empirică se va nota cu S^2 iar valoarea ei corespunzătoare unui anume eșantion se notează cu s^2 .

Statistica, ca v.a., depinde nu doar de tipul de distribuție al populației, de volumul eșantionului ales dar și de metoda de alegere a eșantionului.

Definiția 2.3 Spunem că v.a. X_1, \dots, X_n formează o *selecție aleatoare* (sau *eșantion aleator*) de mărime n dacă v.a. X_1, \dots, X_n sunt independente și dacă ele urmează aceeași distribuție (aceleași tip de repartiție). Mai precis, cerem ca familia $(X_i)_{i=1, n}$ să fie *i.i.d.* (*independentă și identic distribuită*).

Remarca 2.4 În cazul în care avem o selecție aleatoare formată din v.a. X_1, \dots, X_n putem defini similar ca în Secțiunea 1.3.1 conceptele:
momentul de selecție de ordin r

$$\mu'_r \stackrel{\text{def}}{=} \frac{\sum_{i=1}^n X_i^r}{n},$$

media de selecție

$$\bar{X} \stackrel{\text{def}}{=} \frac{X_1 + \dots + X_n}{n},$$

momentul centrat de selecție de ordin r

$$\nu'_r \stackrel{\text{def}}{=} \frac{\sum_{i=1}^n (X_i - \bar{X})^r}{n},$$

dispersia (sau varianța) de selecție

$$S^2 \stackrel{\text{def}}{=} \nu'_2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n},$$

abaterea (sau deviația standard) de selecție

$$S \stackrel{\text{def}}{=} \sqrt{S^2},$$

dispersia (sau varianța) modificată de selecție

$$(S^*)^2 \stackrel{\text{def}}{=} \frac{n}{n-1} S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1},$$

abaterea (sau deviația standard) modificată de selecție

$$S^* \stackrel{\text{def}}{=} \sqrt{(S^*)^2}.$$

2.1 Problema estimației

Dacă avem în vedere un parametru al populației, cum ar fi media întregii populații statistice μ sau dispersia ei σ^2 (care sunt parametri în cazul unei caracteristici X distribuite normal), ne interesează estimarea acestuia folosind informațiile oferite de un eșantion.

Fie o selecție aleatoare formată din v.a. X_1, \dots, X_n .

Să presupunem în continuare că avem o repartiție teoretică cu densitatea $f(x, \lambda)$ cunoscută, unde λ este un parametru legat de ea.

A estima parametrul λ înseamnă a determina o **statistică** (sau **funcție de selecție**)

$$(2.1) \quad \lambda_n^* = \varphi(X_1, \dots, X_n)$$

astfel încât

$$\lambda_n^* \simeq \lambda$$

în diferite sensuri.

Definiția 2.5 λ_n^* se numește **estimator consistent** pentru λ dacă

$$\lim_{n \rightarrow \infty} \mathbb{P}(|\lambda_n^* - \lambda| \leq \varepsilon) = 1, \quad \text{pentru orice } \varepsilon > 0.$$

Remarca 2.6 Deci λ_n^* este estimator consistent dacă

$$\lambda_n^* \xrightarrow{\mathbb{P}} \lambda, \quad \text{pentru } n \rightarrow +\infty.$$

Remarca 2.7 O estimarea a lui λ va fi atunci valoarea estimatorului consistent λ_n^* calculată într-o selecție (x_1, \dots, x_n) fixată.

Definiția 2.8 λ_n^* se numește **estimator corect** pentru λ dacă

$$\lim_{n \rightarrow \infty} \mathbb{E}(\lambda_n^*) = \lambda \quad \text{și} \quad \lim_{n \rightarrow \infty} D^2(\lambda_n^*) = 0.$$

Definiția 2.9 Diferența $\delta = \mathbb{E}(\lambda_n^*) - \lambda$ se numește **deplasarea** (sau **distorsiunea**) estimatorului λ_n^* .

Definiția 2.10 Dacă $\delta \neq 0$ spunem că estimatorul este deplasat. Estimatorul corect este un estimator deplasat.

Definiția 2.11 λ_n^* se numește **estimator absolut corect** pentru λ dacă

$$\mathbb{E}(\lambda_n^*) = \lambda \quad \text{și} \quad \lim_{n \rightarrow \infty} D^2(\lambda_n^*) = 0.$$

Deci $\delta = 0$ și estimatorul absolut corect este nedepășat.

Remarca 2.12 Din inegalitatea lui Cebășev se deduce imediat că un estimator absolut corect este un estimator consistent.

Se cunosc două tipuri de estimare: estimări punctuale și estimări prin intervale de încredere.

2.1.1 Estimări punctuale ale momentelor repartiției teoretice

Se demonstrează că au loc următoarele estimări:

Teorema 2.13 Să presupunem că populația statistică admite medie $\mu = \mathbb{E}(X)$ și dispersie $\sigma^2 = D^2(X)$ (deci și selecția aleatoare X_1, \dots, X_n admite medie și dispersie de selecție).

(i) Media de selecție \bar{X} este un estimator absolut corect al mediei teoretice μ .

(ii) Dispersia de selecție modificată $(S^*)^2$ este un estimator absolut corect al dispersiei teoretice σ^2 .

(iii) Pentru $r \in \mathbb{N}^*$ avem că momentul de selecție de ordin r (notat μ_r') este un estimator absolut corect pentru momentul de ordin r (notat μ_r) al populației \mathcal{P} .

(iv) Pentru un $x \in \mathbb{R}$ fixat, funcția de repartiție de selecție $F_n(x)$ este un estimator absolut corect pentru funcția de repartiție $F(x)$.

Demonstrație. (i) Fie $\bar{X} \stackrel{def}{=} \frac{X_1 + \dots + X_n}{n}$. Dar X_i sunt independente și au aceeași distribuție cu X , deci

$$\mathbb{E}(X_i) = \mathbb{E}(X) \stackrel{not}{=} \mu \quad \text{și} \quad D^2(X_i) = D^2(X) \stackrel{not}{=} \sigma^2.$$

Atunci

$$\begin{aligned} \mathbb{E}(\bar{X}) &= \frac{\sum_{i=1}^n \mathbb{E}(X_i)}{n} = \frac{n\mathbb{E}(X)}{n} = \mu, \\ D^2(\bar{X}) &= \frac{\sum_{i=1}^n D^2(X_i)}{n^2} = \frac{nD^2(X)}{n^2} = \frac{\sigma^2}{n} \rightarrow 0, \quad \text{pentru } n \rightarrow \infty. \end{aligned}$$

(ii) Fie

$$S^2 \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \overline{X^2} - \bar{X}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2.$$

Dar X_i sunt independente și au aceeași distribuție cu X , deci

$$\mathbb{E}(X_i) = \mathbb{E}(X) = \mu \quad \text{și} \quad D^2(X_i) = D^2(X) = \sigma^2.$$

Utilizând relația

$$D^2(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \quad \Leftrightarrow \quad \mathbb{E}(X^2) = D^2(X) + [\mathbb{E}(X)]^2,$$

obținem că

$$\begin{aligned} \mathbb{E}(S^2) &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i^2) - \mathbb{E}(\bar{X}^2) \\ &= \frac{1}{n} \sum_{i=1}^n [D^2(X_i) + [\mathbb{E}(X_i)]^2] - [D^2(\bar{X}) + [\mathbb{E}(\bar{X})]^2] \\ &= \frac{1}{n} \sum_{i=1}^n [\sigma^2 + \mu^2] - \left[\frac{\sigma^2}{n} + \mu^2 \right] = [\sigma^2 + \mu^2] - \left[\frac{\sigma^2}{n} + \mu^2 \right] \\ &= \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n} \sigma^2 = \frac{n-1}{n} D^2(X). \end{aligned}$$

Se poate justifica acum de ce s-a definit și disperia empirică modificată

$$(S^*)^2 \stackrel{\text{def}}{=} \frac{n}{n-1} S^2.$$

Astfel

$$\mathbb{E}[(S^*)^2] = \frac{n}{n-1} \mathbb{E}(S^2) = \frac{n}{n-1} \frac{n-1}{n} \sigma^2 = \sigma^2.$$

Dacă presupunem în plus că populația statistică admite momentul centrat ν'_4 de ordin 4, atunci se poate demonstra următoarea identitate

$$D^2[(S^*)^2] = \frac{n^2}{(n-1)^2} \left(\frac{\nu'_4 - (\nu'_2)^2}{n} - \frac{2(\nu'_4 - 2(\nu'_2)^2)}{n^2} + \frac{\nu'_4 - 3(\nu'_2)^2}{n^3} \right)$$

sau echivalent

$$D^2[(S^*)^2] = \frac{1}{n} \left(\mu'_4 - \frac{n-3}{n-1} (\nu'_2)^2 \right),$$

deci

$$D^2 \left[(S^*)^2 \right] \rightarrow 0, \text{ pentru } n \rightarrow \infty.$$

Am obținut că dispersia empirică modificată $(S^*)^2$ este un estimator absolut corect al dispersiei teoretice $\sigma^2 = D^2(X)$ (iar S^2 nu este un estimator absolut corect al dispersiei teoretice). ■

Exercițiul 2.14 Să se arate că dispersia de selecție S^2 este un estimator deplasat al dispersiei teoretice σ^2 și determinați deplasarea.

Exercițiul 2.15 Să se arate că dispersia de selecție S^2 este un estimator corect (dar nu absolut corect) al dispersiei teoretice.

2.1.2 Estimări punctuale ale parametrilor repartiției teoretice

Presupunem că repartiția teoretică este cunoscută (din cunoașterea în ansamblu a fenomenului studiat sau în urma unei ipoteze sugerată de grafice). Pentru ca repartiția teoretică să fie complet determinată este necesar să cunoaștem și valorile parametrilor de care depinde (de exemplu media μ și dispersia σ^2 care sunt parametri ai repartiției normale). Pentru estimarea acestor parametri avem la dispoziție două metode.

Metoda verosimilității maxime Fie selecția aleatoare (X_1, \dots, X_n) cu densitatea $f_{(X_1, \dots, X_n)}(x_1, \dots, x_n, \lambda_1, \dots, \lambda_k)$ (în cazul v.a. continue sau funcția de frecvență, în cazul v.a. discrete), unde parametrii $\lambda_1, \dots, \lambda_k$ au valori necunoscute.

Definiția 2.16 Funcția

$$V(\lambda_1, \dots, \lambda_k) \stackrel{\text{def}}{=} f_{(X_1, \dots, X_n)}(x_1, \dots, x_n, \lambda_1, \dots, \lambda_k),$$

unde x_1, \dots, x_n sunt observațiile în urma unei selecții de volum n , se numește **funcția de verosimilitate**.

Remarca 2.17 Deoarece X_1, \dots, X_n formează o selecție aleatoare, v.a. sunt independente și identic repartizate, i.e. $f_{X_i} = f_X$, deci funcția de verosimilitate este, în cazul continuu,

$$V(\lambda_1, \dots, \lambda_k) \stackrel{\text{def}}{=} \prod_{i=1}^n f(x_i, \lambda_1, \dots, \lambda_k),$$

unde $f(x_i, \lambda_1, \dots, \lambda_k) := f_X(x_i, \lambda_1, \dots, \lambda_k)$,

iar în cazul discret este

$$V(\lambda_1, \dots, \lambda_k) \stackrel{\text{def}}{=} \prod_{i=1}^n p(x_i, \lambda_1, \dots, \lambda_k),$$

unde $p(x_i, \lambda_1, \dots, \lambda_k) = \mathbb{P}(X = x_i)$.

Definiția 2.18 Valorile $\lambda_i^* = \varphi_i(x_1, \dots, x_n)$, $i = \overline{1, k}$, asociate lui λ_i pentru care funcția V ia valori maxime se numesc **estimatori de maximă verosimilitate**.

Remarca 2.19 Cum $V(\lambda_1, \dots, \lambda_k) > 0$ iar maximele funcției V coincid cu maximele funcției $\ln V$ (deoarece \ln este crescătoare) metoda va consta în determinarea punctelor de maxim pentru $\ln V$.

Etapele sunt următoarele:

- (i) calculăm $V(\lambda_1, \dots, \lambda_k)$ și $\ln V(\lambda_1, \dots, \lambda_k)$ cu ajutorul variabilelor de selecție;
- (ii) rezolvăm sistemul de k ecuații și k necunoscute

$$\frac{\partial}{\partial \lambda_i} [\ln V(\lambda_1, \dots, \lambda_k)] = 0, \quad i = \overline{1, k};$$

(iii) pentru soluția $(\lambda_1^0, \dots, \lambda_k^0)$ găsită mai sus verificăm condiția suficientă de extrem

$$d^2 [\ln V(\lambda_1^0, \dots, \lambda_k^0)] \quad \text{este formă pătratică negativ definită.}$$

Remarca 2.20 Când avem mai multe puncte de maxim se ia cel mai mare dintre maxime.

Remarca 2.21 Se poate demonstra că, în condiții destul de generale, estimatorii λ_i^* sunt consistenți pentru λ_i , $i = \overline{1, k}$.

Exercițiul 2.22 Într-un eșantion de 10 produse fabricate de o companie se observă că primul, al treilea și ultimul nu sunt conform standardelor. Știm că probabilitatea ca un produs să nu fie conform standardelor este p .

Având în vedere eșantionul observat, să se obțină o estimare p^* a valorii parametrului p .

Selecția aleatoare (X_1, \dots, X_n) are densitatea dată de $f(p) := p \cdot q \cdot p \cdot q \cdot \dots \cdot q \cdot p$ (fiecare X_i sunt distribuite de tip Bernoulli, i.e. $X_i \sim \mathcal{B}(1, p)$, $i = \overline{1, 10}$), deci funcția de verosimilitate asociată eșantionului observat este dată de

$$V(p) := p^3 q^7, \quad q := 1 - p.$$

Să studiem maximele funcției V care coincid cu maximele funcției

$$\ln V(p) = 3 \ln p + 7 \ln(1-p).$$

Deoarece

$$\frac{\partial}{\partial p} [\ln V(p)] = 0 \Leftrightarrow \frac{3}{p} - \frac{7}{1-p} = 0 \Leftrightarrow p = \frac{3}{10},$$

obținem punctul critic $p^* = 3/10$.

Acesta e punct de maxim deoarece

$$\frac{\partial^2}{\partial p^2} [\ln V(p^*)] = -\frac{3}{(p^*)^2} - \frac{7}{(1-p^*)^2} < 0$$

(sau echivalent $d^2 [\ln V(p^*)]$ este formă pătratică negativ definită).

Valoarea $p^* = 3/10$ este deci estimatorul parametrului p al distribuției populației și este valoarea care, pentru selecția

$$(x_1, \dots, x_{10}) = (1, 0, 1, 0, \dots, 0, 1)$$

dată maximizează funcția de verosimilitate $V(p)$.

Să remarcăm că în cazul în care se observa că într-un eșantion de 10 produse 3 sunt neconforme standardelor (nu știm ordinea), obținem că funcția de verosimilitate este dată de

$$V(p) := C_{10}^3 p^3 q^7, \quad q := 1-p.$$

Maximul funcției V este în acest caz tot $p^* = 3/10$.

Exercițiul 2.23 Să se estimeze, folosind metoda verosimilității maxime, parametrul λ al repartiției exponențiale cu densitatea $f(x) = \lambda e^{-\lambda x}$, $\lambda > 0, x > 0$, știind că rezultatele obținute în urma efectuării unei selecții de volum 5 sunt $x_1 = 7$, $x_2 = 6.5$, $x_3 = 6.9$, $x_4 = 6.7$, $x_5 = 6.8$. Generalizați rezultatul și studiați dacă estimatorul găsit este sau nu deplasat.

Fiecare v.a. a selecției aleatoare (X_1, \dots, X_5) are densitatea dată de $f(x) = \lambda e^{-\lambda x}$, $\lambda > 0, x > 0$, deci funcția de verosimilitate asociată eșantionului observat este dată de

$$V(\lambda) := \prod_{i=1}^5 f(x_i, \lambda) = \lambda^5 e^{-\lambda \sum_{i=1}^5 x_i}.$$

Să studiem maximele funcției V care coincid cu maximele funcției

$$\ln V(\lambda) = 5 \ln \lambda - 33.9 \cdot \lambda.$$

Deoarece

$$\frac{\partial}{\partial \lambda} [\ln V(\lambda)] = 0 \Leftrightarrow \frac{5}{\lambda} - 33.9 = 0 \Leftrightarrow \lambda = \frac{5}{33.9} = 0.1475,$$

obținem punctul critic

$$\lambda^* = 0.1475 = \frac{1}{\frac{33.9}{5}} = \frac{1}{\bar{x}}.$$

Acesta e punct de maxim deoarece

$$\frac{\partial^2}{\partial \lambda^2} [\ln V(\lambda^*)] = -\frac{5}{(\lambda^*)^2} < 0$$

(sau echivalent $d^2 [\ln V(\lambda^*)]$ este formă pătratică negativ definită).

Numărul $\lambda^* = 0.1475 = \frac{1}{\bar{x}}$ este deci valoarea care, pentru selecția

$$(x_1, \dots, x_{10}) = (7, 6.5, 6.9, 6.7, 6.8)$$

dată maximizează funcția de verosimilitate $V(\lambda)$.

În general, pentru un eșantion aleator (X_1, \dots, X_n) , funcția de verosimilitate asociată eșantionului observat este dată de

$$V(\lambda) := \prod_{i=1}^n f(x_i, \lambda) = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}.$$

Se obține că estimatorul parametrului λ al distribuției populației este dat de punctul de maxim $\lambda^* = \frac{1}{\bar{X}}$.

Deoarece $\mathbb{E}(\bar{X}) = \mathbb{E}(X) = \frac{1}{\lambda}$, obținem că

$$\mathbb{E}(\lambda^*) \neq \lambda \Leftrightarrow \mathbb{E}\left(\frac{1}{\bar{X}}\right) \neq \frac{1}{\mathbb{E}(\bar{X})},$$

deci $\lambda^* = \frac{1}{\bar{X}}$ este un estimator deplasat.

Exercițiul 2.24 Să se estimeze, folosind metoda verosimilității maxime, parametrii μ și σ^2 ai repartiției normale, cu densitatea

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Fiecare v.a. a selecției aleatoare (X_1, \dots, X_n) are densitatea dată de $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, deci funcția de verosimilitate asociată eșantionului observat este dată de

$$V(\mu, \sigma) := \prod_{i=1}^n f(x_i, \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2}.$$

Să studiem maximele funcției V care coincid cu maximele funcției

$$\ln V(\mu, \sigma) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

Deoarece

$$\begin{cases} \frac{\partial}{\partial \mu} [\ln V(\mu, \sigma)] = 0, \\ \frac{\partial}{\partial \sigma} [\ln V(\mu, \sigma)] = 0 \end{cases} \Leftrightarrow \begin{cases} \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0, \\ -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (x_i - \mu)^2 = 0 \end{cases}$$

obținem punctul critic

$$(\mu^*, \sigma^*) = \left(\frac{\sum_{i=1}^n x_i}{n}, \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} \right) = (\bar{x}, s).$$

Acesta e punct de maxim deoarece matricea Hessiană (care este matricea asociată formei pătratice $d^2 [\ln V(\mu, \sigma)]$) are forma

$$\begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{2}{\sigma^3} \sum_{i=1}^n (x_i - \mu) \\ -\frac{2}{\sigma^3} \sum_{i=1}^n (x_i - \mu) & \frac{n}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 \end{pmatrix}$$

care în punctul critic are valoarea

$$\begin{aligned} & \begin{pmatrix} -\frac{n}{(\sigma^*)^2} & -\frac{2}{(\sigma^*)^3} \sum_{i=1}^n (x_i - \mu^*) \\ -\frac{2}{(\sigma^*)^3} \sum_{i=1}^n (x_i - \mu^*) & \frac{n}{(\sigma^*)^2} - \frac{3}{(\sigma^*)^4} \sum_{i=1}^n (x_i - \mu^*)^2 \end{pmatrix} \\ &= \begin{pmatrix} -\frac{n}{(\sigma^*)^2} & 0 \\ 0 & -\frac{2n}{(\sigma^*)^2} \end{pmatrix} \end{aligned}$$

care este negativ definită.

Se obține că estimatorul parametrilor μ și σ^2 ai distribuției populației este dat de punctul de maxim

$$\mu^* = \bar{X} \quad \text{și respectiv} \quad (\sigma^*)^2 = S^2.$$

Deoarece

$$\mathbb{E}(\mu^*) = \mu \quad \Leftrightarrow \quad \mathbb{E}(\bar{X}) = \mathbb{E}(X)$$

obținem că $\mu^* = \bar{X}$ este un estimator nedepășat al lui μ .

Deoarece

$$\mathbb{E}\left((\sigma^*)^2\right) = \mathbb{E}(S^2) = \frac{n-1}{n} \sigma^2 \neq \sigma^2 \quad \Leftrightarrow \quad \mathbb{E}\left((\sigma^*)^2\right) \neq \sigma^2$$

obținem că $(\sigma^*)^2 = S^2$ este un estimator depășat al lui σ^2 .

Exercițiul 2.25 Să se estimeze, folosind metoda verosimilității maxime, parametrul λ al repartiției cu densitatea $f(x) = (1 + \lambda)x^\lambda$, $\lambda > 0$, $x \in (0, 1)$, știind că rezultatele obținute în urma efectuării unei selecții de volum 4 sunt $x_1 = 0.4$, $x_2 = 0.6$, $x_3 = 0.85$, $x_4 = 0.9$.

Funcția de verosimilitate asociată eșantionului observat este dată de

$$V(\lambda) := \prod_{i=1}^4 f(x_i, \lambda) = (1 + \lambda)^4 (x_1 x_2 x_3 x_4)^\lambda.$$

Maximele funcției $\ln V(\lambda) = 4 \ln(1 + \lambda) + \lambda \sum_{i=1}^4 \ln x_i$ sunt date de

$$\frac{\partial}{\partial \lambda} [\ln V(\lambda)] = 0 \quad \Leftrightarrow \quad \frac{4}{1 + \lambda} + \sum_{i=1}^4 \ln x_i = 0 \quad \Leftrightarrow \quad \lambda = -1 - \frac{4}{\sum_{i=1}^4 \ln x_i}.$$

Deci punctul critic este

$$\lambda^* = -1 - \frac{1}{\ln x}.$$

Acesta e punct de maxim deoarece

$$\frac{\partial^2}{\partial \lambda^2} [\ln V(\lambda^*)] = -\frac{4}{(1 + \lambda^*)^2} < 0.$$

Exercițiul 2.26 Să se estimeze, folosind metoda verosimilității maxime, parametrul λ al repartiției Poisson $\mathcal{P}(\lambda)$, considerându-se o selecție de volum n . Arătați că estimatorul găsit este absolut corect.

Fiecare v.a. a selecției aleatoare (X_1, \dots, X_n) are densitatea dată de $f(\lambda) = \prod_{i=1}^n \frac{\lambda^{k_i}}{(k_i)!} e^{-\lambda}$, deci funcția de verosimilitate asociată eșantionului observat este dată de

$$V(\lambda) := \prod_{i=1}^n \frac{\lambda^{k_i}}{(k_i)!} e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^n k_i}}{k_1! \cdot \dots \cdot k_n!} e^{-n\lambda}.$$

Să studiem maximele funcției

$$\ln V(\lambda) = \left(\sum_{i=1}^n k_i \right) \ln \lambda - \sum_{i=1}^n \ln(k_i!) - n\lambda.$$

Luând

$$\frac{\partial}{\partial \lambda} [\ln V(\lambda)] = \left(\sum_{i=1}^n k_i \right) \frac{1}{\lambda} - n = 0,$$

obținem punctul critic

$$\lambda^* = \frac{\sum_{i=1}^n k_i}{n} = \bar{x}.$$

Acesta e punct de maxim deoarece

$$\frac{\partial^2}{\partial \lambda^2} [\ln V(\lambda^*)] = - \left(\sum_{i=1}^n k_i \right) \frac{1}{(\lambda^*)^2} < 0.$$

Se obține că estimatorul parametrului λ al distribuției populației este dat de punctul de maxim

$$\lambda^* = \bar{X}.$$

Deoarece

$$\mathbb{E}(\lambda^*) = \lambda \iff \mathbb{E}(\bar{X}) = \mathbb{E}(X),$$

obținem că $\lambda^* = \bar{X}$ este un estimator nedepășat al lui μ .

Pe de altă parte,

$$D^2(\lambda^*) = D^2(\bar{X}) = \frac{1}{n} D^2(X) = \frac{1}{n} \lambda \rightarrow 0, \quad \text{pentru } n \rightarrow +\infty,$$

deci $\lambda^* = \bar{X}$ este un estimator absolut corect pentru parametrul teoretic λ .

Metoda momentelor Fie selecția aleatoare (X_1, \dots, X_n) asociată caracteristicii X a populației, cu densitatea $f_X(x, \lambda_1, \dots, \lambda_k)$ (în cazul v.a. continue sau funcția de frecvență, în cazul v.a. discrete), unde parametrii $\lambda_1, \dots, \lambda_k$ au valori necunoscute.

Presupunem că există și sunt finite momentele de ordin $1, 2, \dots, k$ (notate $\mu_1, \mu_2, \dots, \mu_k$). Dacă avem o selecția aleatoare, din faptul că momentele

de selecție de ordin r (notate μ'_r) sunt estimatori absoluți corecți pentru momentele teoretice de ordin r , putem scrie că

$$\mu_r(\lambda_1, \dots, \lambda_k) \simeq \mu'_r, \quad r = \overline{1, k}.$$

Deoarece momentele teoretice depind de parametrii $\lambda_1, \dots, \lambda_k$, se pot găsi estimatori pentru aceștia rezolvând sistemul

$$\mu_r(\lambda_1, \dots, \lambda_k) = \mu'_r, \quad r = \overline{1, k}.$$

Acesta este un sistem de k ecuații cu k necunoscute: $\lambda_1, \dots, \lambda_k$.

Se arată că soluția sistemului $(\lambda_1^*, \dots, \lambda_k^*)$ este un estimator consistent pentru $(\lambda_1, \dots, \lambda_k)$.

Remarca 2.27 Pentru estimarea punctuală a parametrilor unei repartiții există și metoda celor mai mici pătrate.

Exercițiul 2.28 Să se estimeze, folosind metoda momentelor, parametrul λ al repartiției Poisson discrete și infinite $\mathcal{P}(\lambda)$, considerându-se o selecție de volum n . Arătați că estimatorul găsit este absolut corect.

Exercițiul 2.29 Fie selecția aleatoare X_1, \dots, X_n dată de timpul de servire a n clienți la un anumit ghișeu. Caracteristica X a populației se presupune că este distribuită de tip exponențial de parametru λ . Să se estimeze, folosind metoda momentelor, parametrul λ al repartiției.

Calculăm momentul teoretic de ordin 1, adică media teoretică și obținem

$$\mathbb{E}(X) = \frac{1}{\lambda}.$$

Pe de altă parte, momentul de selecție de ordin 1 este

$$\mu'_1 = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}.$$

Egalând obținem

$$\lambda = \frac{1}{\bar{X}},$$

deci estimatorul lui λ folosind metoda momentelor este

$$\lambda^* = \frac{1}{\bar{X}}.$$

Exercițiul 2.30 Să se estimeze, folosind metoda momentelor, parametrii v.a. uniform distribuite în intervalul $[a, b]$, în urma obținerii eșantionului $x_1 = 3.1$, $x_2 = 0.2$, $x_3 = 1.6$, $x_4 = 5.2$, $x_5 = 2.1$.

Fiecare v.a. a selecției aleatoare (X_1, \dots, X_5) are densitatea dată de $f(x) = \frac{1}{b-a}$, $x \in [a, b]$. Calculăm momentele teoretice de ordinul 1 și 2 și obținem

$$\mathbb{E}(X) = \frac{a+b}{2} \quad \text{și} \quad \mathbb{E}(X^2) = \frac{a^2 + ab + b^2}{3}.$$

Pe de altă parte, momentele de selecție de ordin 1 și 2 sunt respectiv

$$\mu'_1 = \frac{\sum_{i=1}^5 x_i}{5} = 2.44 \quad \text{și} \quad \mu'_2 = \frac{\sum_{i=1}^5 x_i^2}{5} = 8.73.$$

Obținem deci sistemul

$$\begin{cases} \frac{a+b}{2} = 2.44, \\ \frac{a^2 + ab + b^2}{3} = 8.73. \end{cases}$$

Deci

$$\begin{aligned} a^2 + a(4.88 - a) + (4.88 - a)^2 &= 26.19 \\ \Leftrightarrow a^2 - 4.88a + (4.88)^2 - 26.19 &= 0 \\ \Leftrightarrow a^2 - 4.88a - 2.3756 &= 0, \end{aligned}$$

care are soluțiile

$$a_{1,2} = \frac{4.88 \pm \sqrt{(4.88)^2 - 4 \cdot (-2.3756)}}{2} = \frac{4.88 - 5.7721}{2},$$

deci

$$a_1 = 5.3261, \quad a_2 = -0.4461$$

și apoi

$$b_1 = -0.4461, \quad b_2 = 5.3261.$$

Obținem estimatorii $a^* = -0.4461$ și $b^* = 5.3261$.

În cazul general al unei selecții aleatoare (X_1, \dots, X_n) momentele de selecție de ordin 1 și 2 sunt respectiv

$$\mu'_1 = \frac{\sum_{i=1}^n x_i}{n} \quad \text{și} \quad \mu'_2 = \frac{\sum_{i=1}^n x_i^2}{n}.$$

Obținem deci sistemul

$$\begin{cases} \frac{a+b}{2} = \mu'_1, \\ \frac{a^2+ab+b^2}{3} = \mu'_2. \end{cases}$$

Deci

$$\begin{aligned} a^2 + a(2\mu'_1 - a) + (2\mu'_1 - a)^2 &= 3\mu'_2 \\ \Leftrightarrow a^2 - 2\mu'_1 a + 4(\mu'_1)^2 - 3\mu'_2 &= 0, \end{aligned}$$

care are soluțiile

$$a_{1,2} = \frac{2\mu'_1 \pm \sqrt{(2\mu'_1)^2 - 4 \cdot (4(\mu'_1)^2 - 3\mu'_2)}}{2} = \mu'_1 \pm \sqrt{3}\sqrt{\mu'_2 - (\mu'_1)^2},$$

deci alegem $a = \mu'_1 - \sqrt{3}\sqrt{\mu'_2 - (\mu'_1)^2}$ și apoi $b = \mu'_1 + \sqrt{3}\sqrt{\mu'_2 - (\mu'_1)^2}$.

Obținem estimatorii

$$a^* = \bar{X} - \sqrt{3}\sqrt{S^2} = \bar{X} - \sqrt{3}S \quad \text{și} \quad b^* = \bar{X} + \sqrt{3}\sqrt{S^2} = \bar{X} + \sqrt{3}S,$$

deoarece media de selecție

$$\bar{X} = \mu'_1$$

iar dispersia de selecție

$$S^2 = \nu'_2 = \mu'_2 - (\mu'_1)^2.$$

Exercițiul 2.31 Să se estimeze, folosind metoda momentelor, parametrul m al v.a. X distribuite normal $\mathcal{N}(m, \sigma^2)$.

Exercițiul 2.32 Să se estimeze, folosind metoda momentelor, parametrii v.a. distribuite de tip Gamma $X \sim \Gamma(p, \lambda)$, cu densitatea

$$f(x) = \frac{1}{\Gamma(p)} x^{p-1} e^{-\lambda x},$$

în urma obținerii eșantionului $x_1 = 2.4, x_2 = 2.7, x_3 = 2.8, x_4 = 2.2, x_5 = 2.4$.

Conform calculelor din cadrul cursului de "Teoria Probabilităților" momentele teoretice de ordinul 1 și 2 sunt date de

$$\mathbb{E}(X) = \frac{p}{\lambda} \quad \text{și} \quad \mathbb{E}(X^2) = \frac{p(p+1)}{\lambda^2}.$$

Pe de altă parte, momentele de selecție de ordin 1 și 2 sunt respectiv

$$\mu'_1 = \frac{\sum_{i=1}^5 x_i}{5} = 2.5 \quad \text{și} \quad \mu'_2 = \frac{\sum_{i=1}^5 x_i^2}{5} = 6.298.$$

Obținem deci sistemul

$$\begin{cases} \frac{p}{\lambda} = 2.5, \\ \frac{p(p+1)}{\lambda^2} = 6.298. \end{cases}$$

Deci

$$2.5(2.5\lambda + 1) = 6.298\lambda \quad \Leftrightarrow \quad \lambda = 52.083$$

iar

$$p = 130.2075.$$

Obținem estimatorii $p^* = 130.2075$ și $\lambda^* = 52.083$.

2.1.3 Estimări prin intervale de încredere ale momentelor repartiției teoretice

A estima prin intervale de încredere înseamnă a determina un interval (și nu o valoare) în care se găsește, cu o probabilitate dată, valoarea teoretică exprimată.

Fie selecția aleatoare (X_1, \dots, X_n) asociată caracteristicii X a populației și fie λ valoarea teoretică ce dorim să o estimăm prin această metodă.

Definiția 2.33 Se numește *interval de încredere pentru λ* un interval de tipul (λ_1, λ_2) unde $\lambda_i = \varphi_i(X_1, \dots, X_n)$, $i = \overline{1, 2}$, cu proprietatea

$$\mathbb{P}(\lambda_1 < \lambda < \lambda_2) = \delta \simeq 1.$$

Numărul

$$\delta \simeq 1$$

se va numi **nivel de încredere** sau siguranța estimației.

Numărul

$$\alpha = 1 - \delta \simeq 0$$

se numește **prag de semnificație**.

Intervale de încredere pentru media teoretică Fie X o caracteristică considerată, asociată unei populații și considerăm că $X \sim \mathcal{N}(m, \sigma^2)$. Deci media $\mathbb{E}(X) = m$ și dispersia $D^2(X) = \sigma^2$. Presupunem că σ^2 este cunoscut.

Știm că $\bar{X} \sim \mathcal{N}(m, \sigma^2/n)$ (vezi¹ Exercițiul 11, Capitolul 4 din cadrul *Teoriei Probabilităților*).

Deci, conform² Propoziției 3.51 din cadrul *Teoriei Probabilităților*, variabila aleatoare standardizată $\frac{\bar{X} - \mathbb{E}(\bar{X})}{\sqrt{D^2(\bar{X})}}$ va fi atunci repartizată normal standard, i.e.

$$(2.2) \quad Z \stackrel{\text{def}}{=} \frac{\bar{X} - \mathbb{E}(\bar{X})}{\sqrt{D^2(\bar{X})}} = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}}{\sigma} (\bar{X} - m) \sim \mathcal{N}(0, 1).$$

Pentru a determina intervalul de încredere pentru m punem condiția

$$(2.3) \quad \mathbb{P}(|Z| < z) = \delta \quad \Leftrightarrow \quad \mathbb{P}\left(\left|\frac{\bar{X} - m}{\sigma/\sqrt{n}}\right| < z\right) = \delta$$

cu δ cunoscut, adică vom obține

$$2\Phi(z) - 1 = \delta \quad \Leftrightarrow \quad \Phi(z) = \frac{1 + \delta}{2} = 1 - \frac{\alpha}{2},$$

deci valoarea z se va obține din tabelele de valori ale distribuției $\mathcal{N}(0, 1)$.

Remarca 2.34 De fapt, folosind notații intuitive, avem că valoarea z de mai sus poate fi scrisă astfel:

$$z = z_{1-\alpha/2}.$$

Deci

$$\begin{aligned} |Z| < z_{1-\alpha/2} &\Leftrightarrow \left|\frac{\bar{X} - m}{\sigma/\sqrt{n}}\right| < z_{1-\alpha/2} \Leftrightarrow |\bar{X} - m| < \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \\ &\Leftrightarrow -\frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} < \bar{X} - m < \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}. \end{aligned}$$

¹Dacă $X_k \sim \mathcal{N}(m, \sigma^2)$, $k = \overline{1, n}$, sunt v.a. independente, atunci

$$\bar{X}_n := \frac{\sum_{k=1}^n X_k}{n} \sim \mathcal{N}\left(m, \frac{\sigma^2}{n}\right).$$

² $X \sim \mathcal{N}(m, \sigma^2)$ dacă și numai dacă v.a. standardizată $\frac{X - m}{\sigma} \sim \mathcal{N}(0, 1)$.

Propoziția 2.35 În cazul în care populația are caracteristica $X \sim \mathcal{N}(m, \sigma^2)$ iar σ^2 este cunoscut, intervalul de încredere pentru medie este dat de

$$(2.4) \quad \bar{X} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} < m < \bar{X} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2},$$

unde $z_{1-\alpha/2}$ este valoarea dată de

$$\Phi(z_{1-\alpha/2}) = 1 - \frac{\alpha}{2}.$$

Din graficul repartiției normale standard se poate găsi interpretarea lui $z_{1-\alpha/2}$. Astfel

$$\mathbb{P}(|Z| < z_{1-\alpha/2}) = \delta \Leftrightarrow \int_{-z_{1-\alpha/2}}^{z_{1-\alpha/2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}} dx = \delta.$$

Remarca 2.36 Fie a valoarea exactă a unei mărimi. În acest caz $m = \mathbb{E}(a) = a$ și \bar{a} valoarea aproximativă a acestei mărimi (obținută cu ajutorul unui aparat). Faptul că σ este cunoscut reprezintă precizia măsurătorilor (siguranța aparatului). Intervalul de încredere pentru a este

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

deoarece

$$\frac{\bar{a} - a}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1).$$

O problemă comună care intervine în practică este aceea de a determina **numărul minim de observații necesare pentru a obține o anumită precizie a rezultatelor**. În acest sens utilizăm tot relația (2.3). Presupunem că sunt date δ , siguranța estimației, și Δ (eroarea absolută). Atunci

$$|\bar{a} - a| < \Delta.$$

Deci din

$$|\bar{a} - a| < \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \leq \Delta$$

obținem

$$\frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \leq \Delta \Rightarrow \sqrt{n} \geq \frac{\sigma}{\Delta} z_{1-\alpha/2} \Rightarrow n \geq \left(\frac{\sigma}{\Delta} z_{1-\alpha/2} \right)^2,$$

adică n este primul număr natural care verifică inegalitatea de mai sus.

Remarcăm că eroarea absolută Δ reprezintă și jumătate din lungimea intervalului de încredere $\left(\bar{a} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}, \bar{a} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \right)$.

Exemplul 2.37 Care este numărul de măsurători necesare pentru a obține un interval de încredere de 95% cu o eroare absolută de 2, știind că abaterea empirică modificată a fost obținută și este de 2.6 ?

În cazul v.a. normale se știe că (vezi cursul de "Teoria Probabilităților") valoarea $E_{95} = 1.960 \sigma$ (adică acea cantitate pentru care $\mathbb{P}(|X| < E_{\delta}) = \delta$, unde $X \sim \mathcal{N}(0, \sigma^2)$).

Obținem că $n \in \mathbb{N}$ trebuie să verifice

$$n \geq \left(\frac{\sigma}{\Delta} z_{1-\alpha/2} \right)^2 = \left(\frac{2.6}{2} 1.960 \right)^2 = 6.49.$$

Prin urmare vom lua $n = 8$ (vom lua de fapt un număr par de măsurători, iar primul număr par care verifică inegalitatea de mai sus este 8).

Propoziția 2.38 În cazul în care volumul selecției $n > 30$, populația are caracteristica X care urmează o distribuție oarecare, nu neapărat de tip normal, iar σ^2 este cunoscut, intervalul de încredere pentru medie este dat de

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

unde $z_{1-\alpha/2}$ este valoarea dată de

$$\Phi(z_{1-\alpha/2}) = 1 - \frac{\alpha}{2}.$$

Aceasta are loc deoarece variabila aleatoare standardizată $\frac{\bar{X} - \mathbb{E}(\bar{X})}{\sqrt{D^2(\bar{X})}}$ este, conform Teoremei Limită Centrală, repartizată normal standard $\mathcal{N}(0, 1)$, pentru n mare.

În cazul în care volumul selecției $n > 30$, X repartitie oarecare și σ necunoscut considerăm

$$\sigma^2 \simeq (S^*)^2,$$

unde

$$(S^*)^2 = \frac{n}{n-1} S^2 \quad \Leftrightarrow \quad S^* = \sqrt{\frac{n}{n-1}} S.$$

Propoziția 2.39 Relația (2.4) ne dă acum intervalul de încredere pentru medie în cazul în care volumul selecției $n > 30$, populația are caracteristica X care urmează o distribuție oarecare, nu neapărat de tip normal, iar σ^2 este necunoscut:

$$(2.5) \quad \bar{X} - \frac{S^*}{\sqrt{n}} z_{1-\alpha/2} < m < \bar{X} + \frac{S^*}{\sqrt{n}} z_{1-\alpha/2}.$$

Exercițiul 2.40 În urma efectuării unei selecții de volum 250 s-au obținut $\bar{x} = 126.18$ și $S^* = 4.05$. Determinați intervalul de încredere pentru media teoretică m corespunzătoare pragului de semnificație $\alpha = 0.01$.

În cazul în care volumul selecției $n \leq 30$, X este repartizată $\mathcal{N}(m, \sigma^2)$ și σ necunoscut trebuie să reamintim, mai întâi, legătura dintre distribuția normală și distribuția χ^2 precum și legătura dintre distribuția normală, distribuția χ^2 și distribuția Student (pentru demonstrații vezi Propoziția 3.129 și Propoziția 3.132 din cadrul *Teoriei Probabilităților*).

Propoziția 2.41 Dacă A_i sunt v.a. independente și normale de tip $\mathcal{N}(0, \sigma^2)$, atunci

$$(2.6) \quad \sum_{i=1}^n A_i^2 \sim \chi^2(n, \sigma)$$

adică este distribuită χ^2 ("hi pătrat") de parametrii n și σ .

Propoziția 2.42 Dacă $X \sim \mathcal{N}(0, \sigma^2)$ și $Y \sim \chi^2(a, \sigma)$ sunt două v.a. independente, atunci distribuția

$$T \stackrel{\text{def}}{=} \frac{X}{\sqrt{\frac{Y}{a}}} \sim t(a),$$

adică T este distribuită Student de parametru a .

Aplicând acum Propoziția 2.41 și³ Propoziția 3.130 din cadrul *Teoriei Probabilităților*, deducem că

$$\begin{aligned} X_i \sim \mathcal{N}(m, \sigma^2) &\Leftrightarrow (X_i - m) \sim \mathcal{N}(0, \sigma^2) \\ \Rightarrow \sum_{i=1}^n (X_i - m)^2 &\sim \chi^2(n, \sigma) \end{aligned}$$

iar

$$\begin{aligned} \bar{X} \sim \mathcal{N}\left(m, \frac{\sigma^2}{n}\right) &\Leftrightarrow (\bar{X} - m) \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right) \\ (2.7) \quad \Rightarrow (\bar{X} - m)^2 &\sim \chi^2\left(1, \frac{\sigma}{\sqrt{n}}\right) \\ \Leftrightarrow n(\bar{X} - m)^2 &\sim \chi^2\left(1, \sqrt{n} \frac{\sigma}{\sqrt{n}}\right) = \chi^2(1, \sigma). \end{aligned}$$

³ $X \sim \chi^2(n, \sigma)$ dacă și numai dacă $aX \sim \chi^2(n, \sqrt{a}\sigma)$, pentru orice $a > 0$.

Pe de altă parte, avem că

$$\begin{aligned}
 nS^2 &= \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n [(X_i - m) - (\bar{X} - m)]^2 \\
 &= \sum_{i=1}^n \left[(X_i - m)^2 - 2(X_i - m)(\bar{X} - m) + (\bar{X} - m)^2 \right] \\
 &= \sum_{i=1}^n (X_i - m)^2 - 2(\bar{X} - m) \sum_{i=1}^n (X_i - m) + \sum_{i=1}^n (\bar{X} - m)^2 \\
 &= \sum_{i=1}^n (X_i - m)^2 - 2(\bar{X} - m) \left(\sum_{i=1}^n X_i - nm \right) + n(\bar{X} - m)^2 \\
 &= \sum_{i=1}^n (X_i - m)^2 - 2n(\bar{X} - m)^2 + n(\bar{X} - m)^2,
 \end{aligned}$$

deci, utilizând (2.6), (2.7) și⁴ Propoziția 3.82 din cadrul *Teoriei Probabilităților*,

$$\begin{aligned}
 (2.8) \quad nS^2 &= \sum_{i=1}^n (X_i - m)^2 - n(\bar{X} - m)^2 \\
 &\sim \chi^2(n, \sigma) - \chi^2(1, \sigma) = \chi^2(n-1, \sigma).
 \end{aligned}$$

Pe de altă parte,

$$(\bar{X} - m) \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right) \Leftrightarrow \sqrt{n}(\bar{X} - m) \sim \mathcal{N}(0, \sigma^2).$$

Vom obținem astfel că variabila

$$(2.9) \quad T \stackrel{\text{def}}{=} \frac{\sqrt{n}}{S^*} (\bar{X} - m) = \frac{\sqrt{n}(\bar{X} - m)}{\sqrt{\frac{nS^2}{n-1}}} \sim t(n-1)$$

este distribuită Student de parametru $n-1$.

Acum avem din (2.3) și (2.9)

$$\begin{aligned}
 \mathbb{P}\left(\left|\frac{\sqrt{n}}{S^*}(\bar{X} - m)\right| < t\right) &= \delta \Leftrightarrow \mathbb{P}(|T| < t) = \delta \\
 \mathbb{P}(|T| \geq t) &= \alpha \Leftrightarrow 2\mathbb{P}(T > t) = \alpha.
 \end{aligned}$$

deci

$$\mathbb{P}(T > t) = \alpha/2$$

iar acum (spre deosebire de cazurile precedente în care statistica T era repartizată normal standard) valoarea t se va citi din tabelul distribuției Student de parametru $n-1$.

⁴Dacă $X_1, X_2 \sim \chi^2(n_i, \sigma)$, $i = \overline{1, 2}$, sunt v.a. independente, atunci $X_1 + X_2 \sim \chi^2(n_1 + n_2, \sigma)$.

Remarca 2.43 De fapt, folosind notații intuitive, avem că valoarea t de mai sus poate fi scrisă astfel:

$$t = t_{\alpha/2, n-1}.$$

Propoziția 2.44 În cazul în care volumul selecției $n \leq 30$, populația are caracteristica $X \sim \mathcal{N}(m, \sigma^2)$ iar σ^2 este necunoscut, intervalul de încredere pentru medie este dat de

$$(2.10) \quad \bar{X} - \frac{S^*}{\sqrt{n}} t_{\alpha/2, n-1} < m < \bar{X} + \frac{S^*}{\sqrt{n}} t_{\alpha/2, n-1},$$

unde valoarea $t_{\alpha/2, n-1}$ este citită din tabelul distribuției Student de parametru $n - 1$.

Exercițiul 2.45 În urma efectuării unei selecții de volum 20 s-au obținut media empirică $\bar{x} = 0.149$ și abaterea empirică modificată $S^* = 0.048$. Determinați intervalul de încredere pentru media teoretică m corespunzătoare pragului de semnificație $\alpha = 0.05$.

Exercițiul 2.46 În urma efectuării unei selecții de volum 16 s-au obținut media empirică $\bar{x} = 25.4$ și abaterea empirică (nemodificată) $S = 1.259$. Determinați intervalul de încredere pentru media teoretică m corespunzătoare nivelului de încredere $\delta = 95\%$.

Determinați apoi intervalul de încredere pentru medie utilizând valoarea E_{95} (calculată pentru o v.a distribuită normal). Care interval de încredere este mai mic și de ce?

Citiți tabelul repartiției Student corespunzător la $n = \infty$ grade de libertate și comparați cu valorile E_δ de la distribuția normală.

Determinăm intervalul de încredere pentru medie folosind tabelul repartiției Student deoarece suntem în cazul $n \leq 30$, dispersia teoretică σ^2 este necunoscută iar caracteristica $X \sim \mathcal{N}(m, \sigma^2)$.

Intervalul de încredere pentru medie este dat de:

$$\bar{x} - \frac{S^*}{\sqrt{n}} t_{\alpha/2, n-1} < m < \bar{x} + \frac{S^*}{\sqrt{n}} t_{\alpha/2, n-1},$$

unde $S^* = \sqrt{\frac{n}{n-1}} S = \sqrt{\frac{16}{15}} 1.259 = 1.3$, pragul de semnificație $\alpha = 0.05$ iar $t_{\alpha/2, n-1} = t_{0.025, 15}$ este dat de

$$\begin{aligned} \mathbb{P} \left(\left| \frac{\sqrt{n}}{S^*} (\bar{X} - m) \right| < t \right) &= \delta \quad \Leftrightarrow \quad \mathbb{P} \left(\left| \frac{\sqrt{n}}{S^*} (\bar{X} - m) \right| \geq t \right) = \alpha \\ \Leftrightarrow \quad \mathbb{P} \left(\frac{\sqrt{n}}{S^*} (\bar{X} - m) > t \right) &= \frac{\alpha}{2}, \end{aligned}$$

adică citim din tabelul distribuției Student valoarea

$$t_{0.025,15} = 2.131.$$

Obținem intervalul de încredere

$$\left(25.4 - \frac{1.3}{\sqrt{16}} 2.131, 25.4 + \frac{1.3}{\sqrt{16}} 2.131 \right) = (24.71, 26.09) .$$

Precizăm că dacă volumul selecției ar fi mai mare, atunci am putea utiliza repartiția normală în locul celei Student, deci ar fi utilă valoarea $E_{0.95} = 1.960$ (vezi Exem-
plul 3.58 din cadrul [Teoriei Probabilităților](#)) care verifică

$$\mathbb{P} \left(\left| \frac{\sqrt{n}}{S^*} (\bar{X} - m) \right| \leq E_{0.95} \right) = \delta,$$

deci obținem intervalul

$$\bar{x} - \frac{S^*}{\sqrt{n}} E_{0.95} < m < \bar{x} + \frac{S^*}{\sqrt{n}} E_{0.95}$$

adică

$$\left(25.4 - \frac{1.3}{\sqrt{16}} 1.96, 25.4 + \frac{1.3}{\sqrt{16}} 1.96 \right) = (24.76, 26.04) .$$

Deci intervalul obținut în cazul folosirii repartiției normale este mai mic decât cel obținut în cazul folosirii repartiției Student. Aceasta este posibil deoarece folosirea repartiției Student se face când numărul de date citite (volumul eșantionului) este mai mic.

Să remarcăm și faptul că valorile tabelului Student corespunzătoare lui $n = \infty$ grade de libertate sunt exact cele din tabelul repartiției normale. De exemplu,

$$t_{\alpha/2,\infty} = t_{0.025,\infty} = 1.960 \equiv E_{0.95} = E_{\delta}, \quad \text{cu } \delta = 1 - \alpha.$$

Similar avem valori identice și în celelalte cazuri, de exemplu,

$$t_{\alpha/2,\infty} = t_{0.0005,\infty} = 3.291 \equiv E_{0.999} = E_{\delta}, \quad \text{cu } \delta = 1 - \alpha.$$

Intervale de încredere pentru dispersia teoretică Fie X o caracteristică considerată asociată unei populații și considerăm că $X \sim \mathcal{N}(m, \sigma^2)$. Deci media $\mathbb{E}(X) = m$ și $D^2(X) = \sigma^2$. Considerăm statistica (sau funcția de selecție)

$$(2.11) \quad \chi^2 \stackrel{\text{def}}{=} \frac{nS^2}{\sigma^2} .$$

Din proprietatea (2.8) deducem că $nS^2 \sim \chi^2(n-1, \sigma)$, deci χ^2 va fi repartizată atunci (vezi Propoziția 3.130 din cadrul *Teoriei Probabilităților*)

$$(2.12) \quad \chi^2 \sim \chi^2(n-1, 1).$$

Din condiția

$$(2.13) \quad \mathbb{P}\left(\chi_1^2 < \frac{nS^2}{\sigma^2} < \chi_2^2\right) = \delta$$

se vor putea determina, utilizând tabelul distribuției χ^2 , valorile χ_1^2 și χ_2^2 , unde δ este nivelul de încredere (sau siguranța estimației) ales.

Avem evident că

$$\delta = \mathbb{P}\left(\chi_1^2 < \frac{nS^2}{\sigma^2} < \chi_2^2\right) = \mathbb{P}\left(\frac{nS^2}{\sigma^2} > \chi_1^2\right) - \mathbb{P}\left(\frac{nS^2}{\sigma^2} > \chi_2^2\right),$$

care va reprezenta o ecuație cu două necunoscute.

Cantitatea χ_1^2 se va determina din relația

$$(2.14) \quad \mathbb{P}(\chi^2 > \chi_1^2) = 1 - \frac{\alpha}{2} = 1 - \frac{1-\delta}{2}$$

iar cantitatea χ_2^2 se va determina din relația

$$(2.15) \quad \mathbb{P}(\chi^2 > \chi_2^2) = \frac{\alpha}{2} = \frac{1-\delta}{2},$$

unde $\alpha = 1 - \delta$ este prag de semnificație.

Reamintim că în tabelul distribuției se pot citi valorile ariile porțiunii de grafic de la χ_1^2 la ∞ , adică

$$\mathbb{P}(\chi^2 > \chi_1^2) = \int_{\chi_1^2}^{\infty} f(x) dx,$$

unde $f(x)$ reprezintă densitatea de repartiție a unei v.a. χ^2 repartizată $\chi^2(n-1, 1)$.

Acum având cunoscute valorile χ_1^2 și χ_2^2 deducem din (2.13) că

$$\chi_1^2 < \frac{nS^2}{\sigma^2} < \chi_2^2 \Leftrightarrow \frac{1}{\chi_2^2} < \frac{\sigma^2}{nS^2} < \frac{1}{\chi_1^2}$$

adică obținem următorul rezultat.

Propoziția 2.47 În cazul în care volumul selecției $n \leq 30$, populația are caracteristica $X \sim \mathcal{N}(m, \sigma^2)$ iar m este necunoscut, intervalul de încredere pentru dispersia teoretică σ^2 este dat de

$$(2.16) \quad \frac{nS^2}{\chi_2^2} < \sigma^2 < \frac{nS^2}{\chi_1^2},$$

unde valorile χ_1^2 și χ_2^2 sunt citite din tabelul distribuției $\chi^2(n-1, 1)$ folosind relațiile (2.14-2.15).

Pentru abaterea medie pătratică σ , intervalul de încredere este dat de

$$(2.17) \quad \frac{S\sqrt{n}}{\chi_2} < \sigma < \frac{S\sqrt{n}}{\chi_1}.$$

Remarca 2.48 De fapt, folosind notații intuitive, avem că cele două valori χ_1^2 și χ_2^2 pot fi scrise astfel:

$$\chi_1^2 = \chi_{1-\alpha/2, n-1}^2 \quad \text{și} \quad \chi_2^2 = \chi_{\alpha/2, n-1}^2.$$

Remarca 2.49 Când $n > 30$ folosim faptul că

$$\chi^2(m, 1) \rightarrow \mathcal{N}(m, 2m), \quad \text{pentru } m \rightarrow \infty$$

(vezi⁵ Remarca 5.72 din cadrul *Teoriei Probabilităților*).

În cazul nostru, pentru n suficient de mare,

$$\chi^2(n-1, 1) \simeq \mathcal{N}(n-1, 2(n-1)).$$

Atunci vom putea folosi funcția de repartiție Φ și deducem, având în vedere că $\frac{nS^2}{\sigma^2} \sim \chi^2(n-1, 1)$,

$$\delta = \mathbb{P}\left(\chi_1^2 < \frac{nS^2}{\sigma^2} < \chi_2^2\right) = \Phi\left(\frac{\chi_2^2 - (n-1)}{\sqrt{2(n-1)}}\right) - \Phi\left(\frac{\chi_1^2 - (n-1)}{\sqrt{2(n-1)}}\right)$$

(vezi Propoziția 3.47 din cadrul *Teoriei Probabilităților*).

⁵Dacă $X \sim \chi^2(n, 1)$, atunci, pentru n mare, obținem, aplicând Teorema Limită Centrală, că v.a. $Z_n = \frac{X-n}{\sqrt{2n}}$ este distribuită normal standard $\mathcal{N}(0, 1)$, deci, echivalent, $X = (\sqrt{2n} Z_n + n) \sim \mathcal{N}(n, 2n)$.

Reamintim că graficul repartiției normale $\mathcal{N}(n-1, 2(n-1))$ este simetric față de media $(n-1)$ deci

$$\frac{\chi_1^2 + \chi_2^2}{2} = n-1 \quad \Leftrightarrow \quad \chi_1^2 = 2(n-1) - \chi_2^2$$

și atunci

$$\delta = \Phi\left(\frac{\chi_2^2 - (n-1)}{\sqrt{2(n-1)}}\right) - \Phi\left(-\frac{\chi_2^2 - (n-1)}{\sqrt{2(n-1)}}\right) = 2\Phi\left(\frac{\chi_2^2 - (n-1)}{\sqrt{2(n-1)}}\right) - 1.$$

Folosind tabelul funcției de repartiție Φ , asociată unei v.a. $Z \sim \mathcal{N}(0, 1)$, vom obține valoarea χ_2^2 iar apoi valoarea χ_1^2 .

Exercițiul 2.50 În urma măsurării unei distanțe obținem o colecție de 20 de date. Abaterea empirică modificată obținută este 1.8. Care este intervalul de încredere pentru dispersie cu $\delta = 95\%$.

Intervale de încredere pentru raportul a două dispersii teoretice Fie X_1 și X_2 două caracteristici asociate la două populații și considerăm că $X_1 \sim \mathcal{N}(m_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(m_2, \sigma_2^2)$. Presupunem că mediile $\mathbb{E}(X_i) = m_i$ și dispersiile $D^2(X_i) = \sigma_i^2$ sunt necunoscute, $i = \overline{1, 2}$.

Luând

$$X = \frac{n_1 S_1^2}{\sigma_1^2} \quad \text{și} \quad Y = \frac{n_2 S_2^2}{\sigma_2^2}$$

vom obține că

$$X \sim \chi^2(n_1 - 1, 1) \quad \text{și} \quad Y \sim \chi^2(n_2 - 1, 1),$$

deci X are $(n_1 - 1)$ grade de libertate și Y are $(n_2 - 1)$ grade de libertate.

Dar

$$(S_i^*)^2 = \frac{n_i}{n_i - 1} S_i^2, \quad i = \overline{1, 2},$$

deci

$$X = \frac{(n_1 - 1)(S_1^*)^2}{\sigma_1^2} \quad \text{și} \quad Y = \frac{(n_2 - 1)(S_2^*)^2}{\sigma_2^2}.$$

Considerăm statistica (sau funcția de selecție)

$$F \stackrel{\text{def}}{=} \frac{\frac{X}{n_1 - 1}}{\frac{Y}{n_2 - 1}}.$$

În cazul nostru,

$$F = \frac{\frac{X}{n_1-1}}{\frac{Y}{n_2-1}} = \frac{\frac{(n_1-1)(S_1^*)^2}{(n_1-1)\sigma_1^2}}{\frac{(n_2-1)(S_2^*)^2}{(n_2-1)\sigma_2^2}} = \frac{\frac{(S_1^*)^2}{\sigma_1^2}}{\frac{(S_2^*)^2}{\sigma_2^2}} = \frac{(S_1^*)^2 \sigma_2^2}{(S_2^*)^2 \sigma_1^2}$$

deci, folosind⁶ Propoziția 3.135 din cadrul *Teoriei Probabilităților*,

$$F = \frac{(S_1^*)^2 \sigma_2^2}{(S_2^*)^2 \sigma_1^2} = \frac{\frac{X}{n_1-1}}{\frac{Y}{n_2-1}} \sim F(n_1 - 1, n_2 - 1).$$

Din condiția

$$\mathbb{P}(F_1 < F < F_2) = \delta$$

se vor putea determina, utilizând tabelul distribuției Fisher, valorile F_1 și F_2 .

Avem evident că

$$\delta = \mathbb{P}\left(F_1 < \frac{(S_1^*)^2 \sigma_2^2}{(S_2^*)^2 \sigma_1^2} < F_2\right) = \mathbb{P}\left(\frac{(S_1^*)^2 \sigma_2^2}{(S_2^*)^2 \sigma_1^2} > F_1\right) - \mathbb{P}\left(\frac{(S_1^*)^2 \sigma_2^2}{(S_2^*)^2 \sigma_1^2} > F_2\right),$$

care va reprezenta o ecuație cu două necunoscute.

Cantitatea F_1 se va determina din relația

$$(2.18) \quad \mathbb{P}(F > F_1) = 1 - \frac{\alpha}{2} = 1 - \frac{1 - \delta}{2}$$

iar cantitatea F_2 se va determina din relația

$$(2.19) \quad \mathbb{P}(F > F_2) = \frac{\alpha}{2} = \frac{1 - \delta}{2}.$$

Reamintim că din tabel se pot citi valorile ariei porțiunii de grafic de la F_1 la ∞ , adică

$$\mathbb{P}(F > F_1) = \int_{F_1}^{\infty} f(x) dx,$$

unde $f(x)$ reprezintă densitatea de repartiție a v.a. $F \sim F(n_1 - 1, n_2 - 1)$.

⁶Dacă $X \sim \chi^2(a, 1)$ și $Y \sim \chi^2(b, 1)$ atunci v.a.

$$T \stackrel{\text{def}}{=} \frac{X/a}{Y/b} \sim F(a, b),$$

adică fracția $\frac{X/a}{Y/b}$ este distribuită Fisher de parametri a și b .

Acum având cunoscute valorile F_1 și F_2 deducem din (2.13) că

$$\delta = \mathbb{P} \left(\frac{1}{F_2} \frac{(S_1^*)^2}{(S_2^*)^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{F_1} \frac{(S_1^*)^2}{(S_2^*)^2} \right),$$

adică obținem următorul rezultat.

Propoziția 2.51 În cazul în care volumul selecției $n \leq 30$, populațiile au caracteristicile $X_1 \sim \mathcal{N}(m_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(m_2, \sigma_2^2)$ iar m_1, m_2 și σ_1^2, σ_2^2 sunt necunoscute, intervalul de încredere pentru raportul a două dispersii este dat de

$$(2.20) \quad \frac{1}{F_2} \frac{(S_1^*)^2}{(S_2^*)^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{F_1} \frac{(S_1^*)^2}{(S_2^*)^2}$$

unde valorile F_1 și F_2 sunt citite din tabelul distribuției $F(n_1 - 1, n_2 - 1)$ folosind relațiile (2.18-2.19).

Remarca 2.52 De fapt, folosind notații intuitive, avem că cele două valori F_1 și F_2 pot fi scrise astfel:

$$F_1 = F_{1-\alpha/2, n_1-1, n_2-1} \quad \text{și} \quad F_2 = F_{\alpha/2, n_1-1, n_2-1}.$$

Se știe că are loc următoarea relație între valori

$$(2.21) \quad F_{1-\alpha, a, b} = \frac{1}{F_{\alpha, b, a}}.$$

De exemplu luăm $a = 5$ și $b = 15$ și $\alpha = 0.01$. Atunci din tabel putem citi valoarea ariei α corespunzătoare distribuției $F(5, 15)$

$$F_{\alpha, 5, 15} = 4.36$$

precum și valoarea ariei α corespunzătoare distribuției $F(15, 5)$

$$F_{\alpha, 15, 5} = 9.72.$$

Deci valoarea $F_{1-\alpha, 5, 15}$ este calculată folosind valoarea $F_{\alpha, 15, 5}$, conform formulei

$$F_{1-\alpha, 5, 15} = \frac{1}{F_{\alpha, 15, 5}} = \frac{1}{9.72} = 0.1029$$

Obținem astfel intervalul de încredere pentru raportul a două dispersii sub forma

$$(2.22) \quad \frac{1}{F_{\alpha/2, n_1-1, n_2-1}} \frac{(S_1^*)^2}{(S_2^*)^2} < \frac{\sigma_1^2}{\sigma_2^2} < F_{\alpha/2, n_2-1, n_1-1} \frac{(S_1^*)^2}{(S_2^*)^2}.$$

Exercițiul 2.53 În urma a 31 de observații se obține abaterea empirică modificată $S^* = 1.5$, iar la 25 de observații se obține abaterea empirică modificată $S^* = 0.7$. Să se determine intervalul de încredere pentru raportul dispersiilor considerându-se un prag de semnificație de 5%. Intervalul obținut conține valoarea 1? Interpretați rezultatul (care este deci probabilitatea ca $\sigma_1^2 \neq \sigma_2^2$?).

Capitolul 3

Verificarea ipotezelor statistice

Așa cum se poate vedea în exemplele din capitolul precedent, de cele mai multe ori nu suntem interesați de marginile intervalului construit, ci mai degrabă de problema când intervalul construit conține media sau dispersia teoretică. Deci problema esențială este aceea de a putea preciza dacă o selecție statistică este consistentă în raport cu populația întreagă.

Mai precis, folosind un eșantion de date, putem estima un parametru al unei repartiții teoretice în mod punctual sau printr-un interval de valori plauzibile. Dar altă problemă care se poate pune este aceea de a investiga și apoi de a decide dacă parametrul are sau nu o valoare anume.

Ipoteza statistică este o ipoteză care se face relativ la parametrii unei repartiții (la un singur parametru sau la doi parametri) sau la forma/tipul legii de repartiție pe care o urmează caracteristica cercetată.

Procedura utilizată pentru investigare este aceea de a testa validitatea unei decizii privind o ipoteză statistică prin folosirea, în acest sens, a unei statistici (sau funcție de selecție), adică a unui **test statistic**.

Fie X o v.a. cu densitatea de repartiție $f(x, \theta)$. Notăm cu

$$(H_0) : \theta = \theta^0,$$

ipoteza conform căreia $\theta = \theta^0$, unde θ^0 este o valoare calculată a parametru-lui pe baza unui eșantion x_1, \dots, x_n din populația \mathcal{P} .

Ipoteza (H_0) se numește **ipoteza nulă**.

Pot interveni și ipoteze alternative

$$\begin{array}{ll} (H_0) : \theta = \theta^0, & (H_0) : \theta = \theta^0, \\ (H_1) : \theta = \theta^1, & \text{sau} \\ & (H_1) : \theta \neq \theta^0. \end{array}$$

Definiția 3.1 A testa o ipoteză înseamnă a lua o decizie dacă **ipoteza nulă se respinge** sau dacă **ipoteza nulă se acceptă** (sau se admite sau nu există motive pentru respingerea ei sau se admite până la următorul eșantion).

Definiția 3.2 Se numește **test statistic** orice statistică (sau funcție de selecție) pe baza căreia se poate lua o decizie privind acceptarea sau respingerea unei ipoteze statistice.

Definiția 3.3 Orice test statistic se bazează pe un criteriu de testare. **Criteriul de testare** a unei ipoteze statistice este o statistică (o funcție de selecție) $u = u(x_1, \dots, x_n)$ satisfăcând condițiile:

- (i) funcția de selecție depinde de ipoteza făcută,
- (ii) dacă ipoteza făcută se acceptă atunci repartiția teoretică este complet determinată.

Definiția 3.4 Se numește **nivel de semnificație** (sau **prag de semnificație**, notat cu α), probabilitatea respingerii ipotezei făcute când în realitate ea este adevărată.

Pentru ca o ipoteză să fie respinsă cât mai greu se va alege $\alpha \simeq 0$ (adică $\alpha = 0.01$ sau $\alpha = 0.05$).

Definiția 3.5 Se numește **regiunea critică** (sau **regiunea de respingere**) mulțimea tuturor valorilor statisticii utilizate $u(x_1, \dots, x_n) \in \mathbb{R}$ pentru care ipoteza nulă (H_0) va fi respinsă.

Putem scrie **regiunea critică** cu un **nivel de semnificație** α sub forma

$$\mathcal{V} = \{(x_1, \dots, x_n) \in \mathbb{R}^n : \text{ipoteza } (H_0) \text{ este adevărată ,}$$

$$\text{prin utilizarea statisticii } u\},$$

$$\text{unde } \mathbb{P}((x_1, \dots, x_n) \in \mathcal{V}) = \alpha.$$

Deci dacă $(x_1, \dots, x_n) \in \mathcal{V}$, atunci ipoteza (H_0) se respinge.

Dacă $(x_1, \dots, x_n) \notin \mathcal{V}$, atunci ipoteza (H_0) se acceptă iar mulțimea $\mathbb{R}^n \setminus \mathcal{V}$ se va numi **regiunea de acceptare** cu un **nivel de semnificație** α .

3.1 Teste pentru media unei selecții

Fie \mathcal{P} o populație statistică și X o caracteristică cercetată. Notăm $\mathbb{E}(X) = m$ și $D^2(X) = \sigma^2$. Presupunem că $X \sim \mathcal{N}(m, \sigma^2)$.

În cazul în care dispersia teoretică σ^2 se cunoaște vom folosi distribuția $\mathcal{N}(0, 1)$ și statistica (vezi și definiția (2.2))

$$Z \stackrel{\text{def}}{=} \frac{\bar{X} - m}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

În cazul în care σ^2 nu se cunoaște vom folosi distribuția Student.

3.1.1 Testul Z pentru media unei selecții

Fie $X \sim \mathcal{N}(m, \sigma^2)$ cu σ^2 cunoscut. Ipoteza nulă propusă este

$$(H_0) : m = m_0.$$

Testul Z bilateral Verificăm ipoteza nulă

$$(H_0) : m = m_0$$

față de ipoteza alternativă

$$(H_1) : m \neq m_0.$$

Dacă ipoteza nulă este acceptată atunci

$$Z = \frac{\bar{X} - m_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Pentru α fixat se va determina, din tabelul repartiției Normale standard, valoarea critică z_c astfel încât

$$\alpha = \mathbb{P}(|Z| > z_c)$$

sau echivalent

$$1 - \alpha = \delta = \mathbb{P}(-z_c \leq Z \leq z_c) = 2\Phi(z_c) - 1 \quad \Leftrightarrow \quad \Phi(z_c) = \frac{\delta + 1}{2} = 1 - \frac{\alpha}{2}.$$

Remarca 3.6 De fapt, folosind notații intuitive, avem că valoarea critică z_c de mai sus poate fi scrisă astfel:

$$z_c = z_{1-\alpha/2}.$$

Exercițiul 3.7 Reprezentați grafic situația de mai sus.

Obținem atunci că **intervalul** $[-z_{1-\alpha/2}, z_{1-\alpha/2}]$ **este interval de acceptare** iar **regiunea** $|z| > z_{1-\alpha/2}$ **este regiune critică**, adică

$$\mathcal{V} = \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n : \left| \frac{\bar{x} - m_0}{\sigma/\sqrt{n}} \right| > z_{1-\alpha/2} \right\}.$$

Testul z bilateral înseamnă: se calculează $z_0 = \frac{\bar{x} - m_0}{\sigma/\sqrt{n}}$ (valoarea statisticii în eșantionul observat) și apoi **dacă** $z_0 \in [-z_{1-\alpha/2}, z_{1-\alpha/2}]$ **atunci** (H_0) **se acceptă, și dacă** $z_0 \notin [-z_{1-\alpha/2}, z_{1-\alpha/2}]$ **atunci** (H_0) **se respinge.**

Testul Z unilateral stânga Verificăm ipoteza nulă

$$(H_0) : m = m_0 \quad (\text{de fapt } m \geq m_0)$$

față de ipoteza alternativă

$$(H_1) : m < m_0.$$

Dacă ipoteza nulă este acceptată atunci

$$Z \stackrel{\text{def}}{=} \frac{\bar{X} - m_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Pentru α fixat se va determina, din tabelul repartiției Normale, valoarea critică z_c astfel încât

$$\alpha = \mathbb{P}(Z < -z_c)$$

sau echivalent

$$\alpha = \Phi(-z_c) = 1 - \Phi(z_c) \quad \Leftrightarrow \quad \Phi(z_c) = 1 - \alpha.$$

Remarca 3.8 De fapt, folosind notații intuitive, avem că valoarea critică z_c de mai sus poate fi scrisă astfel:

$$z_c = z_{1-\alpha}.$$

Exercițiul 3.9 Reprezentați grafic situația de mai sus.

Obținem atunci că **intervalul** $[-z_{1-\alpha}, \infty)$ **este interval de acceptare** iar **intervalul** $(-\infty, -z_{1-\alpha})$ **este regiune critică**, adică

$$\mathcal{V} = \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n : \frac{\bar{x} - m_0}{\sigma/\sqrt{n}} < -z_{1-\alpha} \right\}.$$

Testul z unilateral stânga înseamnă: se calculează $z_0 = \frac{\bar{x} - m_0}{\sigma/\sqrt{n}}$ (valoarea statisticii în eșantionul observat) și apoi **dacă** $z_0 \geq -z_{1-\alpha}$ **atunci** (H_0) **se acceptă, și dacă** $z_0 < -z_{1-\alpha}$ **atunci** (H_0) **se respinge.**

Testul Z unilateral dreapta Verificăm ipoteza nulă

$$(H_0) : m = m_0 \quad (\text{de fapt } m \leq m_0)$$

față de ipoteza alternativă

$$(H_1) : m > m_0.$$

Dacă ipoteza nulă este acceptată atunci

$$Z \stackrel{\text{def}}{=} \frac{\bar{X} - m_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

Pentru α fixat se va determina, din tabelul repartiției Normale, valoarea critică z_c astfel încât

$$\alpha = \mathbb{P}(Z > z_c)$$

sau echivalent

$$\Phi(z_c) = 1 - \alpha.$$

Remarca 3.10 De fapt, folosind notații intuitive, avem că valoarea critică z_c de mai sus poate fi scrisă astfel:

$$z_c = z_{1-\alpha}.$$

Exercițiul 3.11 Reprezentați grafic situația de mai sus.

Obținem atunci că **intervalul** $(-\infty, z_{1-\alpha}]$ **este interval de acceptare** iar **intervalul** $(z_{1-\alpha}, \infty)$ **este regiune critică**, adică

$$\mathcal{V} = \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n : \frac{\bar{x} - m_0}{\sigma/\sqrt{n}} > z_{1-\alpha} \right\}.$$

Testul z unilateral dreapta înseamnă: se calculează $z_0 = \frac{\bar{x} - m_0}{\sigma/\sqrt{n}}$ (valoarea statisticii în eșantionul observat) și apoi **dacă** $z_0 \leq z_{1-\alpha}$ **atunci** (H_0) **se acceptă, și dacă** $z_0 > z_{1-\alpha}$ **atunci** (H_0) **se respinge.**

3.1.2 Testul Z pentru diferenția mediilor a două selecții

Fie $X \sim \mathcal{N}(m_1, \sigma_1^2)$ și $Y \sim \mathcal{N}(m_2, \sigma_2^2)$ cu σ_1^2, σ_2^2 cunoscuți. Ipoteza nulă propusă este

$$(H_0) : m_1 = m_2.$$

Această se folosește când în condiții diferite de obținere a unui produs cu aceeași valoare nominală a unui parametru, se constată deosebiri între valorile medii. Se pune problema dacă este vorba de calități diferite ale produselor sau de abateri întâmplătoare. Statistica utilizată va fi

$$Z \stackrel{\text{def}}{=} \frac{(\bar{X} - \bar{Y}) - (m_1 - m_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}.$$

Testul Z bilateral Verificăm ipoteza

$$(H_0) : m_1 = m_2$$

față de ipoteza alternativă

$$(H_1) : m_1 \neq m_2.$$

Deoarece X, Y sunt repartizate normal obținem că.

$$\bar{X} \sim \mathcal{N}(m_1, \sigma_1^2/n_1) \quad \text{și} \quad \bar{Y} \sim \mathcal{N}(m_2, \sigma_2^2/n_2)$$

deci (vezi rezultatele din cadrul *Teoriei Probabilităților*) se poate arăta că

$$\bar{X} - \bar{Y} \sim \mathcal{N}(m_1 - m_2, \sigma_1^2/n_1 + \sigma_2^2/n_2).$$

Dacă ipoteza (H_0) se acceptă, atunci

$$\bar{X} - \bar{Y} \sim \mathcal{N}(0, \sigma_1^2/n_1 + \sigma_2^2/n_2)$$

și deci

$$Z = \frac{(\bar{X} - \bar{Y}) - 0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = \frac{(\bar{X} - \bar{Y})}{\sqrt{D^2(\bar{X} - \bar{Y})}} \sim \mathcal{N}(0, 1).$$

Pentru α fixat se va determina, din tabelul repartiției Normale standard, valoarea critică z_c astfel încât

$$\alpha = \mathbb{P}(|Z| > z_c)$$

sau echivalent

$$1 - \alpha = \delta = \mathbb{P}(-z_c \leq Z \leq z_c) = 2\Phi(z_c) - 1 \quad \Leftrightarrow \quad \Phi(z_c) = \frac{\delta + 1}{2} = 1 - \frac{\alpha}{2}.$$

Remarca 3.12 De fapt, folosind notații intuitive, avem că valoarea critică z_c de mai sus poate fi scrisă astfel:

$$z_c = z_{1-\alpha/2}.$$

Obținem atunci că **intervalul** $[-z_{1-\alpha/2}, z_{1-\alpha/2}]$ **este interval de acceptare** iar **regiunea** $|z| > z_{1-\alpha/2}$ **este regiune critică.**

Testul z bilateral înseamnă: se calculează $z_0 = \frac{\bar{x}-m_0}{\sigma/\sqrt{n}}$ (valoarea statisticii în eșantionul observat) și apoi

$$\begin{cases} \text{dacă } z_0 \in [-z_{1-\alpha/2}, z_{1-\alpha/2}], & \text{atunci } (H_0) \text{ se acceptă;} \\ \text{dacă } z_0 \notin [-z_{1-\alpha/2}, z_{1-\alpha/2}], & \text{atunci } (H_0) \text{ se respinge.} \end{cases}$$

Testul Z unilateral stânga Verificăm ipoteza

$$(H_0) : m_1 = m_2 \quad (\text{de fapt } m_1 \geq m_2)$$

față de ipoteza alternativă

$$(H_1) : m_1 < m_2.$$

Dacă ipoteza (H_0) este acceptată atunci, similar ca mai sus,

$$Z = \frac{(\bar{X} - \bar{Y}) - 0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = \frac{(\bar{X} - \bar{Y})}{\sqrt{D^2 (\bar{X} - \bar{Y})}} \sim \mathcal{N}(0, 1).$$

Pentru α fixat se va determina, din tabelul repartiției Normale, valoarea critică z_c astfel încât

$$\alpha = \mathbb{P}(Z < -z_c)$$

sau echivalent

$$\alpha = \Phi(-z_c) = 1 - \Phi(z_c) \quad \Leftrightarrow \quad \Phi(z_c) = 1 - \alpha.$$

Remarca 3.13 De fapt, folosind notații intuitive, avem că valoarea critică z_c de mai sus poate fi scrisă astfel:

$$z_c = z_{1-\alpha}.$$

Obținem atunci că **intervalul** $[-z_{1-\alpha}, \infty)$ **este interval de acceptare** iar **intervalul** $(-\infty, -z_{1-\alpha})$ **este regiune critică.**

Testul z unilateral stânga înseamnă: se calculează $z_0 = \frac{\bar{x}-m_0}{\sigma/\sqrt{n}}$ (valoarea statisticii în eșantionul observat) și apoi

$$\begin{cases} \text{dacă } z_0 \in [-z_{1-\alpha}, \infty), & \text{atunci } (H_0) \text{ se acceptă;} \\ \text{dacă } z_0 \notin [-z_{1-\alpha}, \infty), & \text{atunci } (H_0) \text{ se respinge.} \end{cases}$$

Testul Z unilateral dreapta Verificăm ipoteza nulă

$$(H_0) : m_1 = m_2 \quad (\text{de fapt } m_1 \leq m_2)$$

față de ipoteza alternativă

$$(H_1) : m_1 > m_2.$$

Dacă ipoteza (H_0) este acceptată atunci, similar ca mai sus,

$$Z = \frac{(\bar{X} - \bar{Y}) - 0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = \frac{(\bar{X} - \bar{Y})}{\sqrt{D^2 (\bar{X} - \bar{Y})}} \sim \mathcal{N}(0, 1).$$

Pentru α fixat se va determina, din tabelul repartiției Normale, valoarea critică z_c astfel încât

$$\alpha = \mathbb{P}(Z > z_c)$$

sau echivalent

$$\Phi(z_c) = 1 - \alpha.$$

Remarca 3.14 De fapt, folosind notații intuitive, avem că valoarea critică z_c de mai sus poate fi scrisă astfel:

$$z_c = z_{1-\alpha}.$$

Obținem atunci că **intervalul $(-\infty, z_{1-\alpha}]$ este interval de acceptare iar intervalul $(z_{1-\alpha}, \infty)$ este regiune critică.**

Testul z unilateral dreapta înseamnă: se calculează $z_0 = \frac{\bar{x} - \bar{y}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$ (valoarea statisticii în eșantionul observat) și apoi

$$\begin{cases} \text{dacă } z_0 \in (-\infty, z_{1-\alpha}], \text{ atunci } (H_0) \text{ se acceptă;} \\ \text{dacă } z_0 \notin (-\infty, z_{1-\alpha}], \text{ atunci } (H_0) \text{ se respinge.} \end{cases}$$

Remarca 3.15 Dacă $\sigma_1^2 = \sigma_2^2 = \sigma^2$ atunci $Z = \frac{\bar{X} - \bar{Y}}{\sigma \sqrt{1/n_1 + 1/n_2}}$.

Remarca 3.16 Dacă n_1 și n_2 sunt suficienți de mari și σ_1^2 și σ_2^2 necunoscuți, atunci putem aproxima

$$\sigma_1^2 \simeq (S_1^*)^2 \quad \text{și} \quad \sigma_2^2 \simeq (S_2^*)^2,$$

$$\text{unde } (S_i^*)^2 = \frac{n_i}{n_i - 1} S_i^2, \quad i = \overline{1, 2}.$$

Remarca 3.17 Testele prezentate mai sus pot fi utilizate și în cazul în care X, Y nu urmează legea normală dar n_1 și n_2 sunt suficienți de mari, deoarece în acest caz, conform T.L.C., variabila

$$Z = \frac{(\bar{X} - \bar{Y}) - \mathbb{E}(\bar{X} - \bar{Y})}{D(\bar{X} - \bar{Y})} \sim \mathcal{N}(0, 1).$$

3.1.3 Testul T pentru media unei selecții

Fie $X \sim \mathcal{N}(m, \sigma^2)$ cu σ^2 necunoscut iar volumul de observații mic, $n \leq 30$. Ipoteza nulă propusă este

$$(H_0) : m = m_0.$$

Statistica utilizată va fi

$$T \stackrel{\text{def}}{=} \frac{\bar{X} - m}{S^*/\sqrt{n}}.$$

Având în vedere relația (2.9) și faptul că

$$\frac{\bar{X} - m}{S^*/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - m)}{\sqrt{\frac{nS^2}{n-1}}}$$

obținem că

$$T \sim t(n-1).$$

Testul T bilateral Verificăm ipoteza

$$(H_0) : m = m_0$$

față de ipoteza alternativă

$$(H_1) : m \neq m_0.$$

Dacă ipoteza nulă este acceptată atunci

$$T = \frac{\bar{X} - m_0}{S^*/\sqrt{n}} \sim t(n-1).$$

Pentru α fixat se va determina, din tabelul repartiției Student, valoarea critică t_c astfel încât

$$\alpha = \mathbb{P}(|T| > t_c)$$

sau echivalent

$$\mathbb{P}(T > t_c) = \frac{\alpha}{2}.$$

Remarca 3.18 De fapt, folosind notații intuitive, avem că valoarea critică t_c de mai sus poate fi scrisă astfel:

$$t_c = t_{\alpha/2, n-1} .$$

Exercițiul 3.19 Reprezentați grafic situația de mai sus.

Obținem atunci că

intervalul $[-t_{\alpha/2, n-1}, t_{\alpha/2, n-1}]$ este interval de acceptare

iar

regiunea $|t| > t_{\alpha/2, n-1}$ este regiune critică,

adică

$$\mathcal{V} = \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n : \left| \frac{\bar{x} - m_0}{S^*/\sqrt{n}} \right| > t_{\alpha/2, n-1} \right\} .$$

Testul T bilateral înseamnă: se calculează

$$t_0 = \frac{\bar{x} - m_0}{S^*/\sqrt{n}}$$

(valoarea statisticii în eșantionul observat) și apoi

$$\begin{cases} \text{dacă } t_0 \in [-t_{\alpha/2, n-1}, t_{\alpha/2, n-1}], \text{ atunci } (H_0) \text{ se acceptă;} \\ \text{dacă } t_0 \notin [-t_{\alpha/2, n-1}, t_{\alpha/2, n-1}], \text{ atunci } (H_0) \text{ se respinge.} \end{cases}$$

Exercițiul 3.20 Media teoretică a unei distanțe este $m = 400.008$ m. S-au făcut $n = 20$ de observații și s-a găsit media $\bar{x} = 400.012$ m. și abaterea empirică modificată $S^* = 0.0020$ m. Să se cerceteze, aplicându-se Testul T bilateral, dacă media de selecție (media distanțelor observate) diferă semnificativ sau nu de media teoretică, considerându-se un prag de semnificație de $\alpha = 0.05$.

Determinați și intervalul de încredere pentru medie.

Testul T unilateral stânga Verificăm ipoteza

$$(H_0) : m = m_0 \quad (\text{de fapt } m \geq m_0)$$

față de ipoteza alternativă

$$(H_1) : m < m_0 .$$

Dacă ipoteza nulă este acceptată atunci

$$T \stackrel{\text{def}}{=} \frac{\bar{X} - m_0}{S^*/\sqrt{n}} \sim t(n-1).$$

Pentru α fixat se va determina, din tabelul repartiției Student, valoarea critică t_c astfel încât

$$\alpha = \mathbb{P}(T < -t_c) \Leftrightarrow \mathbb{P}(T > t_c) = \alpha.$$

Remarca 3.21 De fapt, folosind notații intuitive, avem că valoarea critică t_c de mai sus poate fi scrisă astfel:

$$t_c = t_{\alpha, n-1}.$$

Exercițiul 3.22 Reprezentați grafic situația de mai sus.

Obținem atunci că

intervalul $[-t_{1-\alpha, n-1}, \infty)$ **este interval de acceptare**

iar

regiunea $(-\infty, -t_{1-\alpha, n-1})$ **este regiune critică,**

adică

$$\mathcal{V} = \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n : \frac{\bar{x} - m_0}{S^*/\sqrt{n}} < -t_{1-\alpha, n-1} \right\}.$$

Testul t unilateral stânga înseamnă: se calculează

$$t_0 = \frac{\bar{x} - m_0}{S^*/\sqrt{n}}$$

(valoarea statisticii în eșantionul observat) și apoi

⎧ dacă $t_0 \in [-t_{1-\alpha, n-1}, \infty)$, **atunci** (H_0) **se acceptă;**
⎨ dacă $t_0 \notin [-t_{1-\alpha, n-1}, \infty)$, **atunci** (H_0) **se respinge.**

Exercițiul 3.23 Media teoretică a unei distanțe este $m = 400.008$ m. S-au făcut $n = 20$ de observații și s-a găsit media $\bar{x} = 400.012$ m. și abaterea empirică $S = 0.0019$ m. Să se cerceteze, aplicându-se Testul T unilateral stânga, dacă media de selecție (media distanțelor observate) diferă semnificativ sau nu de media teoretică, considerându-se un prag de semnificație de $\alpha = 0.05$.

Determinați și intervalul de încredere pentru medie.

Testul T unilateral dreapta Verificăm ipoteza nulă

$$(H_0) : m = m_0 \quad (\text{de fapt } m \leq m_0)$$

față de ipoteza alternativă

$$(H_1) : m > m_0.$$

Dacă ipoteza nulă este acceptată atunci

$$T \stackrel{\text{def}}{=} \frac{\bar{X} - m_0}{S^*/\sqrt{n}} \sim t(n-1).$$

Pentru α fixat se va determina, din tabelul repartiției Student, valoarea critică t_c astfel încât

$$\alpha = \mathbb{P}(T > t_c).$$

Remarca 3.24 De fapt, folosind notații intuitive, avem că valoarea critică t_c de mai sus poate fi scrisă astfel:

$$t_c = t_{\alpha, n-1}.$$

Exercițiul 3.25 Reprezentați grafic situația de mai sus.

Obținem atunci că

intervalul $(-\infty, t_{\alpha, n-1}]$ **este interval de acceptare**

iar

regiunea $(t_{\alpha, n-1}, \infty)$ **este regiune critică,**

adică

$$\mathcal{V} = \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n : \frac{\bar{x} - m_0}{S^*/\sqrt{n}} > t_{\alpha, n-1} \right\}.$$

Testul t unilateral dreapta înseamnă: se calculează

$$t_0 = \frac{\bar{x} - m_0}{S^*/\sqrt{n}}$$

(valoarea statisticii în eșantionul observat) și apoi

$$\begin{cases} \text{dacă } t_0 \in (-\infty, t_{\alpha, n-1}), & \text{atunci } (H_0) \text{ se acceptă;} \\ \text{dacă } t_0 \notin (-\infty, t_{\alpha, n-1}), & \text{atunci } (H_0) \text{ se respinge.} \end{cases}$$

3.2 Teste pentru dispersia unei selecții

3.2.1 Testul χ^2 pentru dispersia unei selecții

Fie \mathcal{P} o populație statistică și X o caracteristică cercetată. Notăm $\mathbb{E}(X) = m$ și $D^2(X) = \sigma^2$. Presupunem că $X \sim \mathcal{N}(m, \sigma^2)$.

Ipoteza nulă propusă este

$$(H_0) : \sigma^2 = \sigma_0^2.$$

Statistica utilizată va fi (vezi (2.11) și (2.12))

$$\chi^2 \stackrel{\text{def}}{=} \frac{nS^2}{\sigma^2} = \frac{(n-1)(S^*)^2}{\sigma^2} \sim \chi^2(n-1, 1).$$

Testul χ^2 bilateral Verificăm ipoteza nulă

$$(H_0) : \sigma^2 = \sigma_0^2$$

față de ipoteza alternativă

$$(H_1) : \sigma^2 \neq \sigma_0^2.$$

Dacă ipoteza nulă este acceptată atunci

$$\chi^2 = \frac{(n-1)(S^*)^2}{\sigma_0^2} \sim \chi^2(n-1, 1).$$

Pentru α fixat se va determina, din tabelul repartiției $\chi^2(n-1, 1)$, valorile critice χ_1^2 și χ_2^2 astfel încât

$$\delta = 1 - \alpha = \mathbb{P}(\chi_1^2 \leq \chi^2 \leq \chi_2^2).$$

Pentru modul de determinare al valorilor χ_1^2 și χ_2^2 vezi Secțiunea *Intervale de încredere pentru dispersia teoretică* (pag. 48).

Remarca 3.26 Folosind notații intuitive, avem că cele două valori χ_1^2 și χ_2^2 pot fi scrise astfel:

$$\chi_1^2 = \chi_{1-\alpha/2, n-1}^2 \quad \text{și} \quad \chi_2^2 = \chi_{\alpha/2, n-1}^2.$$

Exercițiul 3.27 Reprezentați grafic situația de mai sus.

Obținem atunci că

intervalul $\left[\chi_{1-\alpha/2, n-1}^2, \chi_{\alpha/2, n-1}^2 \right]$ **este interval de acceptare**

iar

regiunea $\left[0, \chi_{1-\alpha/2, n-1}^2 \right) \cup \left(\chi_{\alpha/2, n-1}^2, \infty \right)$ **este regiune critică,**

adică

$$\mathcal{V} = \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n : \frac{(n-1)(S^*)^2}{\sigma_0^2} \in [0, \chi_{1-\alpha/2, n-1}^2) \cup (\chi_{\alpha/2, n-1}^2, \infty) \right\}.$$

Testul χ^2 bilateral înseamnă: se calculează

$$\chi_0^2 = \frac{(n-1)(S^*)^2}{\sigma_0^2}$$

(valoarea statisticii în eșantionul observat) și apoi

$$\begin{cases} \text{dacă } \chi_0^2 \in \left[\chi_{1-\alpha/2, n-1}^2, \chi_{\alpha/2, n-1}^2 \right], & \text{atunci } (H_0) \text{ se acceptă;} \\ \text{dacă } \chi_0^2 \notin \left[\chi_{1-\alpha/2, n-1}^2, \chi_{\alpha/2, n-1}^2 \right], & \text{atunci } (H_0) \text{ se respinge.} \end{cases}$$

Testul χ^2 unilateral stânga Verificăm ipoteza nulă

$$(H_0) : \sigma^2 = \sigma_0^2 \quad (\text{de fapt } \sigma^2 \geq \sigma_0^2)$$

față de ipoteza alternativă

$$(H_1) : \sigma^2 < \sigma_0^2.$$

Dacă ipoteza nulă este acceptată atunci

$$\chi^2 = \frac{(n-1)(S^*)^2}{\sigma_0^2} \sim \chi^2(n-1, 1).$$

Pentru α fixat se va determina, din tabelul repartiției $\chi^2(n-1, 1)$, valoarea critică χ_1^2 astfel încât

$$\mathbb{P}(\chi^2 \geq \chi_1^2) = 1 - \alpha.$$

Remarca 3.28 Folosind notații intuitive, avem că valoarea χ_1^2 poate fi scrisă astfel:

$$\chi_1^2 = \chi_{1-\alpha, n-1}^2.$$

Exercițiul 3.29 Reprezentați grafic situația de mai sus.

Obținem atunci că

intervalul $[\chi_{1-\alpha, n-1}^2, \infty)$ **este interval de acceptare**

iar

regiunea $[0, \chi_{1-\alpha, n-1}^2)$ **este regiune critică,**

adică

$$\mathcal{V} = \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n : \frac{(n-1)(S^*)^2}{\sigma_0^2} \in [0, \chi_{1-\alpha, n-1}^2) \right\}.$$

Testul χ^2 unilateral stânga înseamnă: se calculează

$$\chi_0^2 = \frac{(n-1)(S^*)^2}{\sigma_0^2}$$

(valoarea statisticii în eșantionul observat) și apoi

$$\begin{cases} \text{dacă } \chi_0^2 \in [\chi_{1-\alpha, n-1}^2, \infty), & \text{atunci } (H_0) \text{ se acceptă;} \\ \text{dacă } \chi_0^2 \notin [\chi_{1-\alpha, n-1}^2, \infty), & \text{atunci } (H_0) \text{ se respinge.} \end{cases}$$

Testul χ^2 unilateral dreapta Verificăm ipoteza nulă

$$(H_0) : \sigma^2 = \sigma_0^2 \quad (\text{de fapt } \sigma^2 \leq \sigma_0^2)$$

față de ipoteza alternativă

$$(H_1) : \sigma^2 > \sigma_0^2.$$

Dacă ipoteza nulă este acceptată atunci

$$\chi_c^2 = \frac{(n-1)(S^*)^2}{\sigma_0^2} \sim \chi^2(n-1, 1).$$

Pentru α fixat se va determina, din tabelul repartiției $\chi^2(n-1, 1)$, valoarea critică χ_2^2 astfel încât

$$\alpha = \mathbb{P}(\chi^2 > \chi_2^2).$$

Remarca 3.30 Folosind notații intuitive, avem că valoarea χ_2^2 poate fi scrisă astfel:

$$\chi_2^2 = \chi_{\alpha, n-1}^2.$$

Exercițiul 3.31 Reprezentați grafic situația de mai sus.

Obținem atunci că

intervalul $[0, \chi_{\alpha, n-1}^2]$ **este interval de acceptare**

iar

regiunea $(\chi_{\alpha, n-1}^2, \infty)$ **este regiune critică,**

adică

$$\mathcal{V} = \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n : \frac{(n-1)(S^*)^2}{\sigma_0^2} \in (\chi_{\alpha, n-1}^2, \infty) \right\}.$$

Testul χ^2 unilateral dreapta înseamnă: se calculează

$$\chi_0^2 = \frac{(n-1)(S^*)^2}{\sigma_0^2}$$

(valoarea statisticii în eșantionul observat) și apoi

$$\begin{cases} \text{dacă } \chi_0^2 \in [0, \chi_{\alpha, n-1}^2], & \text{atunci } (H_0) \text{ se acceptă;} \\ \text{dacă } \chi_0^2 \notin [0, \chi_{\alpha, n-1}^2], & \text{atunci } (H_0) \text{ se respinge.} \end{cases}$$

Exercițiul 3.32 Știm că dispersia reală este $\sigma^2 = 2.25$. În urma a 30 de observații se obține abaterea empirică modificată $S^* = 0.9$. Să se cerceteze, aplicându-se Testul χ^2 unilateral dreapta, dacă valorile obținute diferă semnificativ sau nu de σ^2 dat, considerându-se un prag de semnificație de 5%.

3.2.2 Testul F pentru raportul dispersiilor a două selecții

Fie $X \sim \mathcal{N}(m_1, \sigma_1^2)$ și $Y \sim \mathcal{N}(m_2, \sigma_2^2)$. Ipoteza nulă propusă este

$$(H_0) : \sigma_1^2 = \sigma_2^2.$$

Statistica utilizată va fi

$$F \stackrel{\text{def}}{=} \frac{\frac{(S_1^*)^2}{\sigma_1^2}}{\frac{(S_2^*)^2}{\sigma_2^2}} = \frac{\sigma_2^2 (S_1^*)^2}{\sigma_1^2 (S_2^*)^2} = \frac{\sigma_2^2 \frac{n_1}{n_1-1} S_1^2}{\sigma_1^2 \frac{n_2}{n_2-1} S_2^2} = \frac{n_2-1}{n_1-1} \frac{n_1 S_1^2}{n_2 S_2^2}.$$

Știm că $U_1 \stackrel{\text{def}}{=} \frac{n_1 S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1, 1)$ iar $U_2 \stackrel{\text{def}}{=} \frac{n_2 S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1, 1)$. Atunci deducem că

$$F = \frac{\frac{U_1}{n_1 - 1}}{\frac{U_2}{n_2 - 1}} \sim F(n_1 - 1, n_2 - 1).$$

Vom presupune, fără a restrânge generalitatea, că $(S_1^*)^2 > (S_2^*)^2$.

Testul F bilateral Verificăm ipoteza nulă

$$(H_0) : \sigma_1^2 = \sigma_2^2$$

față de ipoteza alternativă

$$(H_1) : \sigma_1^2 \neq \sigma_2^2.$$

Dacă ipoteza nulă este acceptată atunci

$$F = \frac{(S_1^*)^2}{(S_2^*)^2} \sim F(n_1 - 1, n_2 - 1).$$

Pentru α fixat se va determina, din tabelul repartiției $F(n_1 - 1, n_2 - 1)$, valorile critice F_1 și F_2 astfel încât

$$\delta = 1 - \alpha = \mathbb{P}(F_1 \leq F_c \leq F_2).$$

Pentru modul de determinare al valorilor F_1 și F_2 vezi Secțiunea *Intervale de încredere pentru raportul a două dispersii teoretice* (pag. 51).

Vom determina valorile F_1 și F_2 presupunând că cele două probabilități de mai sus sunt egale, adică

$$\mathbb{P}(F < F_1) = \alpha/2$$

și

$$\mathbb{P}(F > F_2) = \alpha/2.$$

Remarca 3.33 Folosind notații intuitive, avem că cele două valori F_1 și F_2 pot fi scrise astfel (vezi și relația (2.21)):

$$F_1 = F_{1-\alpha/2, n_1-1, n_2-1} = \frac{1}{F_{\alpha/2, n_2-1, n_1-1}} \quad \text{și} \quad F_2 = F_{\alpha/2, n_1-1, n_2-1}.$$

Exercițiul 3.34 Reprezentați grafic situația de mai sus.

Obținem atunci că

intervalul $\left[\frac{1}{F_{\alpha/2, n_2-1, n_1-1}}, F_{\alpha/2, n_1-1, n_2-1} \right]$ **este interval de acceptare**

iar

regiunea $\left(0, \frac{1}{F_{\alpha/2, n_2-1, n_1-1}} \right) \cup (F_{\alpha/2, n_1-1, n_2-1}, \infty)$ **este regiune critică,**

adică

$$\mathcal{V} = \left\{ (x_1, \dots, x_n) \in \mathbb{R}^n : \frac{(S_1^*)^2}{(S_2^*)^2} \in \left(0, \frac{1}{F_{\alpha/2, n_2-1, n_1-1}} \right) \cup (F_{\alpha/2, n_1-1, n_2-1}, \infty) \right\}.$$

Testul F bilateral înseamnă: se calculează

$$F_0 = \frac{(S_1^*)^2}{(S_2^*)^2}$$

(valoarea statisticii în eşantionul observat) și apoi

$$\begin{cases} \text{dacă } F_0 \in \left[\frac{1}{F_{\alpha/2, n_2-1, n_1-1}}, F_{\alpha/2, n_1-1, n_2-1} \right], & \text{atunci } (H_0) \text{ se acceptă;} \\ \text{dacă } F_0 \notin \left[\frac{1}{F_{\alpha/2, n_2-1, n_1-1}}, F_{\alpha/2, n_1-1, n_2-1} \right], & \text{atunci } (H_0) \text{ se respinge.} \end{cases}$$

Exercițiul 3.35 În urma a 31 de observații se obține abaterea empirică modificată $S^* = 1.5$, iar la 25 de observații se obține abaterea empirică modificată $S^* = 0.7$. Să se cerceteze, aplicându-se Testul F bilateral, dacă ipoteza nulă este respinsă sau nu (adică dacă cele două dispersii sunt egale) considerându-se un prag de semnificație de 5%.

Testul F unilateral stânga Verificăm ipoteza

$$(H_0) : \sigma_1^2 = \sigma_2^2 \quad (\text{de fapt } \sigma_1^2 \geq \sigma_2^2)$$

față de ipoteza alternativă

$$(H_1) : \sigma_1^2 < \sigma_2^2.$$

Dacă ipoteza nulă este acceptată atunci

$$F = \frac{(S_1^*)^2}{(S_2^*)^2} \sim F(n_1 - 1, n_2 - 1).$$

Pentru α fixat se va determina, din tabelul repartiției $F(n_1 - 1, n_2 - 1)$, valoarea critică F_1 astfel încât

$$\mathbb{P}(F \geq F_1) = 1 - \alpha.$$

Remarca 3.36 Folosind notații intuitive, avem că valoarea F_1 poate fi scrisă astfel (vezi și relația (2.21)):

$$F_1 = F_{1-\alpha, n_1-1, n_2-1} = \frac{1}{F_{\alpha, n_2-1, n_1-1}}.$$

Exercițiul 3.37 Reprezentați grafic situația de mai sus.

Obținem atunci că

intervalul $\left[\frac{1}{F_{\alpha, n_2-1, n_1-1}}, \infty \right)$ este interval de acceptare

iar

regiunea $\left(0, \frac{1}{F_{\alpha, n_2-1, n_1-1}} \right)$ este regiune critică.

Testul F unilateral stânga înseamnă: se calculează

$$F_0 = \frac{(S_1^*)^2}{(S_2^*)^2}$$

(valoarea statisticii în eșantionul observat) și apoi

$$\left\{ \begin{array}{l} \text{dacă } F_0 \in \left[\frac{1}{F_{\alpha, n_2-1, n_1-1}}, \infty \right), \text{ atunci } (H_0) \text{ se acceptă;} \\ \text{dacă } F_0 \notin \left[\frac{1}{F_{\alpha, n_2-1, n_1-1}}, \infty \right), \text{ atunci } (H_0) \text{ se respinge.} \end{array} \right.$$

Exercițiul 3.38 În urma a 31 de observații se obține abaterea empirică modificată $S^* = 1.5$, iar la 25 de observații se obține abaterea empirică modificată $S^* = 0.7$. Să se cerceteze, aplicându-se Testul F bilateral, dacă ipoteza nulă este respinsă sau nu (adică dacă cele două dispersii sunt egale) considerându-se un prag de semnificație de 5%.

Testul F unilateral dreapta Verificăm ipoteza

$$(H_0) : \sigma_1^2 = \sigma_2^2 \quad (\text{de fapt } \sigma_1^2 \leq \sigma_2^2)$$

față de ipoteza alternativă

$$(H_1) : \sigma_1^2 > \sigma_2^2.$$

Dacă ipoteza nulă este acceptată atunci

$$F = \frac{(S_1^*)^2}{(S_2^*)^2} \sim F(n_1 - 1, n_2 - 1).$$

Pentru α fixat se va determina, din tabelul repartiției $F(n_1 - 1, n_2 - 1)$, valoarea critică F_2 astfel încât

$$\mathbb{P}(F > F_2) = \alpha.$$

Remarca 3.39 Folosind notații intuitive, avem că valoarea F_2 poate fi scrisă astfel:

$$F_2 = F_{\alpha, n_1 - 1, n_2 - 1}.$$

Exercițiul 3.40 Reprezentați grafic situația de mai sus.

Obținem atunci că

intervalul $[0, F_{\alpha, n_1 - 1, n_2 - 1})$ **este interval de acceptare**

iar

regiunea $(F_{\alpha, n_1 - 1, n_2 - 1}, \infty)$ **este regiune critică.**

Testul F unilateral dreapta înseamnă: se calculează

$$F_0 = \frac{(S_1^*)^2}{(S_2^*)^2}$$

(valoarea statisticii în eșantionul observat) și apoi

$\left\{ \begin{array}{l} \text{dacă } F_0 \in [0, F_{\alpha, n_1 - 1, n_2 - 1}), \text{ atunci } (H_0) \text{ se acceptă;} \\ \text{dacă } F_0 \notin [0, F_{\alpha, n_1 - 1, n_2 - 1}), \text{ atunci } (H_0) \text{ se respinge.} \end{array} \right.$

Bibliografie

- [1] George Ciucu, Virgil Craiu, Ion Săcuiu, *Probleme de statistică matematică*, Editura Tehnică, București, 1974.
- [2] George Ciucu, Gabriel Sîmboan, *Teoria probabilităților și statistică matematică. Culegere de probleme*, Editura Tehnică, București, 1962.
- [3] Jay L. Devore, *Probability and Statistics for Engineering and the Sciences* (Ninth Edition), Cengage Learning, Boston, 2016.
- [4] Jay L. Devore, Kenneth N. Berk, *Modern Mathematical Statistics with Applications* (Second Edition), series: Springer Texts in Statistics, Springer New York, 2012.
- [5] Haralambie M. Ionescu, *Statistică matematică*, Editura Didactică și Pedagogică, București, 1962.
- [6] Marius Iosifescu, Gheorghe Mihoc, Radu Theodorescu, *Teoria probabilităților și statistică matematică*, Editura Tehnică, București, 1966.
- [7] Lucian Maticiuc, *Teoria probabilităților*, Universitatea "Alexandru Ioan Cuza", Iași, <http://www.math.uaic.ro/maticiuc/didactic/Probability Theory.pdf>, 2018.
- [8] Gheorghe Mihoc, George Ciucu, Virgil Craiu, *Teoria probabilităților și statistică matematică*, Editura Didactică și Pedagogică, București, 1970.
- [9] Gheorghe Mihoc, Dumitru Firescu, *Statistică matematică*, Editura Didactică și Pedagogică, București, 1966.
- [10] Gheorghe Mihoc, Nicolae Micu, *Teoria probabilităților și statistică matematică*, Editura Didactică și Pedagogică, București, 1980.
- [11] Octav Onicescu, Gheorghe Mihoc, *Lección de statistică matematică*, Editura Tehnică, București, 1958.
- [12] Sheldon Ross, *Introductory Statistics* (Third Edition), Elsevier, Oxford, 2010.

- [13] Iulian Stoleriu, *Statistica prin Matlab*, Editura Matrix ROM, București, 2010.
- [14] Pavel Talpalaru, Liliana Popa, Emilia Popovici, *Probleme de teoria probabilităților și statistică matematică*, Editura Universității Tehnice „Gheorghe Asachi”, Iași, 1995.