Lucian MATICIUC

INTRODUCERE ÎN STATISTICĂ MATEMATICĂ

Universitatea ,,Alexandru Ioan Cuza''

Iaşi – 2017

Cuprins

Capitolul 1

Statistică descriptivă

1.1 Variabile empirice

"Statistica este arta învăţării din date" (Sheldon M. Ross, 2010).

"Statistica este disciplina care se ocupă cu colectarea, analiza şi interpretarea datelor obţinute din observarea unui experiment. Această disciplină are o structură coerentă bazată pe Teoria Probabilităţilor" (Karl Pearson, 1936).

Definiţia 1.1 *Partea statisticii care se ocupă cu culegerea, înregistrarea, gruparea, descrierea şi sumarizarea datelor se numeşte* ***statistică descriptivă****.*

Definiţia 1.2 *Partea statisticii care se ocupă cu interpretarea şi obţinerea concluziilor din datele colectate în cadrul unei experienţe se numeşte* ***statistică inferenţială****.*

Statistica este interesată în obţinerea de informaţii despre o colecţie (mulţime) de elemente. La baza statisticii stă noţiunea de probabilitate.

Definiţia 1.3 *O mulţime de elemente ce posedă o trăsătură comună, şi care se cercetează în statistică, poartă numele de* ***populaţie statistică*** *(colectivitate statistică). Elementele care alcătuiesc populaţia statistică se numesc* ***unităţi statistice*** *sau* ***indivizi****. Numărul de indivizi care alcătuiesc populaţia statistică determină* ***volumul populaţiei****.*

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

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

Definiţia 1.4 *Se numeşte **selecţie (eşantion, sondaj)** o submulţime a populaţiei, i.e. o colectivitate parţială de elemente extrase la întâmplare din cadrul populaţiei.*

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

Selecţia spunem că este repetată (cu întoarcere) dacă individul extras este reintrodus în colectivitate înainte de a se extrage următorul; în caz contrar, selecţia este nerepetată (fără întoarcere). Dacă volumul selecţiei este foarte mic în raport cu volumul populaţiei atunci nu se mai face distincţia între cele două tipuri de selecţie (aceasta se va considera repetată).

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

Se numeşte **serie statistică**, asociată unei selecţii 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 repezintă frecvenţele absolute corespunzătoare valorii x_i (adică numărul care arată de câte ori apare valoarea x_i în timpul selecţiei).

Definiţia 1.5 *Numim **variabilă empirică (de selecţie)**, notată pe scurt v.e.,*

asociată unei selecţii de volum n, un tablou de forma

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

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

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

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

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

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 selecţiei este mare se va face o grupare a acesor valori pe intervale disjuncte (sau clase), de obicei egale, intervale închise la stânga şi 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

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

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

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

Definiţia 1.6 *Se numeşte **frecvenţă absolută cumulată crescător**, respectiv **descrescător**, corespunzătoare valorii x_i, valorile*

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

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

Definiţia 1.7 *Se numeşte* ***frecvenţă relativă cumulată crescător****, respectiv* ***descrescător****, corespunzătoare valorii* x_i*, valorile*

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

Funcţia empirică de repartiţie a v.e. X^* se notează cu $F_n(x)$ şi este definită astfel:

Dacă X^* este dată de (1.1) atunci

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

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

Dacă X^* este dată de (1.2) atunci

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

Dacă X^* este dată de (1.3) atunci

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

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 numeşte diagramă. Reprezentarea grafică se poate face în diverse moduri.

Poligonul frecvenţelor absolute. Se iau pe abscisa Ox valorile x_i şi pe Oy frecvenţele absolute n_i corespunzătoare valorilor x_i. Unind aceste puncte vom obţine poligonul frecvenţelor 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 frecvenţelor absolute n_i (sau relative f_i) corespunzătoare lui x_i. Menţionăm că dacă unim vârfurile acestor perpendiculare prin segmente vom obţine poligonul frecvenţelor absolute (sau respectiv relative). Evident, suma lungimilor segmentelor obţinute în cazul folosirii frecvenţelor relative trebuie să fie 1.

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

Histograma. Această reprezentare se aseamănă cu **Reprezentarea cu bare**, dar se foloseşte 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 împărţi intervalul de valori într-o serie de intervale şi apoi să numărăm câte valori intră în fiecare interval. Clasele de valori sunt de obicei specificate ca intervale consecutive, de pe axa absciselor, şi care nu se suprapun. Intervalele nu trebuie să fie neapărat de dimensiuni egale. Un dreptunghi este ridicat peste acele intervale cu o înălţime egală cu frecvenţa relativă, adică proporţională cu numărul de cazuri din fiecare acea clasă. Evident, suma ariilor dreptunghiurilor obţinute trebuie să fie 1.

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

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

Obţinem 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 nişte date numerice care reprezintă fidel o variabilă aleatoare sau o caracteristică avută în vedere. Cunoaşterea momentelor statistice este utilă în practică în compararea a două variabile aleatoare sau a două populaţii statistice pe care este definită aceeaşi caracteristică, precum şi 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

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

Acestei serii îi asociem variabila empirică (de selecţie)

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

unde $f_i = \dfrac{n_i}{n}$ (frecvenţele 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)

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

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

Parametrii tendinţei centrale: $\bar{x},\ m_e,\ m_0$

Aceşti parametrii au rolul de a evidenţia poziţia în jurul căreia se grupează ansamblul valorilor unei v.e.. Această poziţie exprimată printr-un număr se numeşte poziţie centrală. Ea poate fi evidentiata 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)

şi 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 $r = 1$, obţinem

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

respectiv

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

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

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

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

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 frecvenţa relativă cea mai mare.

Cuantilele de ordin n sunt valorile caracteristicii X^* care împart volumul selecţiei în n părţi egale. Cuantila de ordin 2 este chiar mediana şi este acel număr $x_{1/2}$ care verifică ecuaţia

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

unde F_n este funcţia empirică de repartiţie asociată v.e. X^*.

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

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

Parametrii variabilităţii (ai împrăştierii): $R,\ s^2,\ s,\ (s^*)^2,\ s^*$

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

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

Momentul centrat empiric de ordin r este

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

În particular, pentru $r = 2$ obţinem
Dispersia (sau **varianţa**) **empirică**:

$$s^2 \overset{def}{=\!=} \nu'_2 = \frac{1}{n}\sum\nolimits_{i=1}^{k} n_i\,(x_i - \bar{x})^2 = \sum\nolimits_{i=1}^{k} f_i\,(x_i - \bar{x})^2\ . \tag{1.6}$$

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

Propoziţia 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.* $(X^*)^2$.

Demonstraţie. Într-adevăr,

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

■

Mărimea

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

se numeşte **abaterea medie pătratică empirică** (sau **deviaţia standard empirică**).

Dispersia (sau **varianţa**) **empirică modificată** este numărul

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

unde n este volumul selecţiei.

Avem deci şi formula de calcul

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

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

$$s^* \overset{def}{=\!=} \sqrt{(s^*)^2}\,.$$

O altă caracteristică importantă este

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

Caracteristicile formei

Aceste caracteristici se referă la forma poligonului frecvenţelor absolute şi relative (sau a curbei de repartiţie 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 şi cu $\bar{\gamma}_1$ şi măsoară gradul de turtire al poligonului frecvenţelor relative sau al curbei repartiţie faţă de repartiţia normală.

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

$$\nu_4' \overset{def}{=\!=} \mathbb{E}\left[X - \mathbb{E}\left(X\right)\right]^4 = 3\sigma^4$$

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

$$E_X = \frac{\nu_4'}{s^4} - 3 = \frac{3\sigma^4}{\left(\sigma^2\right)^2} - 3 = 0.$$

Din acest motiv graficul repartiţiei normale este curba cu care se compară toate repartiţiile.

Dacă $E_X > 0$ sau echivalent $\alpha_4 > 3$ atunci curba este mai ascuţită decât curba corespunzătoare densităţii repartiţiei normale.

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

Asimetria se măsoară prin coeficientul de asimetrie

$$\bar{\gamma}_2 = \frac{\nu_3'}{s^3},$$

adică $\bar{\gamma}_2 = \alpha_3$. Menţionăm 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. Repartiţia statistică normală are $\bar{\gamma}_2 = 0$ sau echivalent $\alpha_3 = 0$.

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

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

Exerciţiul 1.10 *Să presupunem că un aparat de măsurare este utilizat pentru a citi o distanţă de 12 de ori. Se obţin 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 in tabelul de mai jos:

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

Obţinem 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 şasea şi a şaptea, adică media aritmetică $\frac{0.20+0.25}{2} = 0.225$.

Moda (valoarea modală) este 0.20.

Media de selecţie (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 varianţa) empirică este dată de formula

$$s^2 \overset{def}{=\!=} \frac{1}{12}\sum\nolimits_{i=1}^{6} n_i \left(x_i - \bar{x}\right)^2 = \sum\nolimits_{i=1}^{6} f_i \left(x_i - \bar{x}\right)^2 = 0.00643,$$

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

Pe de altă parte dispersia empirică modificată este numărul

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

În plus abaterea empirică modificată este

$$s^* \overset{def}{=\!=} \sqrt{\left(s^*\right)^2} \simeq 0.083811.$$

Mai trebuie făcut graficul poligonului frecvenţelor relative.

Se poate scrie şi funcţia empirică de repartiţie $F\left(x\right)$ *(care este o funcţie în scară).*

Exerciţiul 1.11 *Să presupunem că un aparat de măsurare este utilizat pentru a citi o distanţă de 20 de ori. Datele sunt colectate in tabelul de mai jos:*

(1.9)

22.7	25.4	22.0	20.5	22.5
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 mulţimea 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.

(1.10)

20.5	22.0	22.0	22.3	22.5
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 şi a unsprezecea valoare, adică media aritmetică $\frac{23.5+23.7}{2}$ *(se poate considera drept mediană şi una dintre cele două valori).*

Moda este valoarea 23.8 *(valoarea cu frecvenţa cea mai mare).*

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

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

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 varianţa) empirică este dată de formula

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

Este util să scriem mai întâi un tabel cu diferenţele[1] $x_i - \bar{x}$ *şi* $(x_i - \bar{x})^2$:

x_i	*Frecvenţa abs.* n_i	*Frecvenţa 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 obţinem 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

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

[1] Diferenţele de tipul $x_i - \bar{x}$ se numesc **deviaţia 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 şi formulele de mai sus putem lăsa toate valorile x_i chiar dacă se repetă (deci $n = 20$ în acest caz). Atunci frecvenţa relativă a fiecărei valori va fi aceeaşi $f_i = 1/20 = 0.05$ şi frecvenţa absolută a fiecărei valori va fi aceeaşi $n_i = 1$. 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 obţinem următoarea ***formulă de calcul a dispersiei empirice modificate*** *(vezi şi formula (1.8)):*

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

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

Clasa	*Frecvenţa absolută a clasei*	*Frecvenţa 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 definirii dispersiei empirice modificate*** *(vezi (1.8) şi (1.12) precum şi Teorema 2.13)**.***

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

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

Să considerăm un eşantion 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

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

Obţinem

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

Pe de o parte, avem că termenul $\sum_{i=1}^{n} (x_i - \mu)^2$ va fi, pentru n foarte mare (apropiat 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 $\bar{X} \overset{def}{=\!=} \frac{\sum_{i=1}^{n} X_i}{n}$ 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 populaţiei (decât dispersia empirică s^2 dată de definiţia (1.6)).**

Exerciţiul 1.14 *Presupunem că au fost obţinute următoarele valori în urma a* 15 *citiri a unei distanţe 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* $\mathcal{N}\left(0, \sigma^2\right)$*, menţionăm că în acest caz avem relaţia* $E'_\alpha = E_\alpha \, s^*$ *, cu* $\alpha \in (0,1)$ *, unde* E'_α *este dat de relaţia*

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

(i) *Calculaţi media empirică, dispersia empirică modificată, abaterea empirică modificată precum şi* $E_{0.5}$ *,* $E_{0.95}$ *.*

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

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

Rezolvare:

(i) *Obţinem*

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

şi

$$E'_{0.50} = 0.6745 \cdot s^* = 0.6745 \cdot 0.05483 = 0.03698$$
$$E'_{0.95} = 1.960 \cdot s^* = 1.960 \cdot 0.05483 = 0.10747.$$

(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 eşantion, adică* $\frac{11}{15} \cdot 100 = 73.33\%$ *.*

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

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

Diferenţa vine deoarece avem un număr mic de date. Cu cât volumul eşantionului 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*

$$\left(\bar{x} - E'_{0.50}, \bar{x} + E'_{0.50}\right) = (212.23 - 0.03698, 212.23 + 0.03698)$$
$$= (212.1930, 212.2669)\,.$$

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

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

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 repartiţie

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

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

Numărul

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

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 repartiţie

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

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

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

Propoziţia 1.15 (Proprietăţi ale mediei) *Pentru orice* $a, c \in \mathbb{R}$ *şi orice v.a.* X, Y *au loc următoarele:*
(i) *valoarea medie a unei constante este egală cu constanta:*

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

(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.*

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

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

$$\mathbb{E}(XY) = \sum\nolimits_{i,j} x_i y_j p_{ij} = \sum\nolimits_{i,j} x_i y_j \mathbb{P}(X = x_i, Y = y_j)$$
$$= \sum\nolimits_{i,j} x_i y_j \mathbb{P}(\{X = x_i\} \cap \{Y = y_j\})$$
$$= \sum\nolimits_{i,j} x_i y_j \mathbb{P}(X = x_i)\,\mathbb{P}(Y = y_j) = \sum\nolimits_{i,j} x_i y_j p_i q_j = \sum\nolimits_i x_i p_i \sum\nolimits_j y_j q_j$$
$$= \mathbb{E}(X)\,\mathbb{E}(Y).$$

■

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

$$\mu_r \overset{def}{=\!=} \mathbb{E}(X^r) = \sum\nolimits_i p_i x_i^r.$$

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

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

$$\lambda_r \overset{def}{=\!=} \mathbb{E}(|X|^r) = \sum\nolimits_i p_i |x_i|^r.$$

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

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

În particular, obţinem

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

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

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

(iv) $m_0 \overset{def}{=\!=} \lim_{r\to 0} m_r = \lim_{r\to 0} \left[p_1 x_1^r + p_2 x_2^r + \cdots + p_n x_n^r\right]^{1/r}$.

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

$$\ln(m_0) = \lim_{r\to 0} \frac{\ln(p_1x_1^r + p_2x_2^r + \cdots + p_nx_n^r)}{r}$$
$$= \lim_{r\to 0} \frac{[\ln(p_1x_1^r + p_2x_2^r + \cdots + p_nx_n^r)]'}{1} = \lim_{r\to 0} \frac{[p_1x_1^r + p_2x_2^r + \cdots + p_nx_n^r]'}{p_1x_1^r + p_2x_2^r + \cdots + p_nx_n^r}$$
$$= \lim_{r\to 0} \frac{p_1x_1^r \ln x_1 + p_2x_2^r \ln x_2 + \cdots + p_nx_n^r \ln x_n}{p_1x_1^r + p_2x_2^r + \cdots + p_nx_n^r}$$
$$= \frac{p_1 \ln x_1 + p_2 \ln x_2 + \cdots + p_n \ln x_n}{p_1 + p_2 + \cdots + p_n} = \frac{\ln[(x_1)^{p_1}(x_2)^{p_2}\cdots(x_n)^{p_n}]}{1}$$
$$= \ln[(x_1)^{p_1}(x_2)^{p_2}\cdots(x_n)^{p_n}],$$

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

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

$$\nu_r \overset{def}{=\!=} \mathbb{E}[(X-\mu_1)^r] = \sum\nolimits_i p_i (x_i - \mu_1)^r .$$

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

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

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

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

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.

Proprietăţile mediei sunt aceleaşi din cazul discret (vezi Propoziţia 1.15).

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

$$\mu_r \overset{def}{=\!=} \mathbb{E}\left(X^r\right) = \int_{-\infty}^{\infty} x^r f\left(x\right) dx\,.$$

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

În particular,

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

Se numeşte **momentul centrat de ordin** r al v.a. X, media v.a. $(X-\nu_1)^r$. Vom nota

$$\nu_r \overset{def}{=\!=} \mathbb{E}\left[\left(X-\mu_1\right)^r\right] = \int_{-\infty}^{\infty} \left(x-\mu_1\right)^r f\left(x\right) dx\,.$$

În particular,

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

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

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

Remarca 1.17 *Având în vedere calculul*

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

obţinem o formulă foarte utilă în calcule:

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

Propoziţia 1.18 (Proprietăţi ale dispersiei) *Pentru orice $a, c \in \mathbb{R}$ şi orice v.a. X, Y au loc următoarele:*
(i) dispersia unei constante este nulă,

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

(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,

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

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

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

Aceasta are avantajul că se exprimă prin aceleaşi unităţi de măsură ca şi valorile v.a. X.

Propoziţia 1.19 (Proprietăţi ale abaterii medii pătratice) *Pentru orice $a, c \in \mathbb{R}$ şi orice v.a. X au loc următoarele:*
(i) $D(c) = 0$.
(ii) $D(aX) = |a| D(X)$.

Teorema 1.20 (Inegalitatea lui Cebâşev) *Fie X o v.a. care admite media m şi 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 şi 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 aplicaţii este următoarea consecinţă:

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

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

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

Capitolul 2

Elemente de teoria selecţiei şi a estimaţiei

Acest capitol face tranziţia spre statistica inferenţială. Dacă avem un eşantion de n observaţii din cadrul unei populaţii, dorim să obţinem estimări pentru media μ a întregii populaţii, dispersia σ^2 a întregii populaţii, deviaţia standard σ a întregii populaţii etc. Dar având în vedere că eşantionul de n date este extras la întâmplare din populaţie este evident că media empirică $\bar{x}$, dispersia empirică s^2 şi deviaţia standard empirică s vor fi diferite de la un eşantion la altul. Ne interesează comportamentul acestor estimatori la eşantioane diferite.

Presupunem că o populaţie 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) şi dispersia (notată σ^2) acestei populaţii sunt 26.1 respectiv 17.5.

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

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

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

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 $\bar{x}$ şi s^2 sunt calculate plecând de la nişte variabile aleatoare (deoarece $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ iar x_i sunt alese aleator), obţinem că $\bar{x}$ şi s^2 sunt şi ele, la rândul lor, nişte variabile aleatoare Deci chiar dacă volumul n este menţinut constant pot exista variaţii ale mediei şi 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:	$\bar{x} = 27.2$	$s^2 = 23.0\,.$

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

Important în acest sens vor fi estimatorii (vezi şi definiţia dată de relaţia (2.1)), adică funcţii care să depindă de fiecare selecţie în parte.

În cadrul teoriei estimaţiei foarte importante vor fi trei distribuţii:

- distribuţia "chi pătrat" de parametrii n şi σ, notată $\chi^2(n, \sigma)$,
- distribuţia Student de parametru n, notată $t(n)$ şi
- distribuţia Fisher de parametrii m şi n, notată $F(m, n)$.

Pentru definiţii ale acestora şi diverse legături între distribuţii vezi cursul de *Teoria Probabilităţilor*.

În general vorbind, fie $\mathcal{P}$ o populaţie statistică şi X o caracteristică cantitativă relativă la $\mathcal{P}$. În cele mai multe cazuri repartiţia teoretică a lui X nu este cunoscută. Scopul statisticii este acela de a determina pe baza experienţelor cu elemente din $\mathcal{P}$ (selecţiilor din $\mathcal{P}$) a legii de repartiţie a lui X precum şi a anumitor momente ale lui X (de exemplu, media şi dispersia). Acest lucru este posibil aplicând metoda selecţiei sau a eşantioanelor.

Dacă numărul de elemente al mulţimii $\mathcal{P}$ este notat N, atunci în urma a n experienţe obţinem rezultatele $x_1, \ldots, x_n$, unde n este mult mai mic decât N şi reprezintă volumul selecţiei.

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

Pentru a reflecta fidel proprietăţile întregii populaţii, o selecţie trebuie să îndeplinească următoarele condiţii: $\mathcal{P}$ să fie cât mai omogenă; n să fie cât mai mare; unităţile selecţiei să fie extrase la întâmplare; fiecare unitate din $\mathcal{P}$ să aibă aceaşi probabilitate (şansă) de a face parte din selecţie.

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

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

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

Atunci valorile $x_1, \ldots, x_n$, observate în urma selecţiei, constitue valoarea observată a vectorului aleator n-dimensional $(X_1, \ldots, X_n)$. Repetând selecţia vom obţine diferite valori ale vectorului $(X_1, \ldots, X_n)$.

Remarca 2.1 *Fiecare moment empiric obţinut pe baza selecţiei este valoarea unei anumite v.a. teoretice, valoare care variază odată cu selecţia.*

De exemplu, media empirică

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

reprezintă valoarea v.a.

$$\bar{X} \stackrel{def}{=\!=} \frac{X_1 + \cdots + X_n}{n},$$

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

De asemenea, dispersia empirică

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

este valoarea v.a.

$$S^2 \stackrel{def}{=\!=} \frac{1}{n} \sum\nolimits_{i=1}^{n} \left(X_i - \bar{X}\right)^2.$$

Problema esenţială 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 direcţii de studiu: determinarea tipului de repartiţie teoretică şi a parametrilor corespunzători; şi de determinarea unor indicatori numerici pentru v.a. teoretică X.

Să menţionăm că uneori tipul de repartiţie teoretică se cunoaşte din experienţele anterioare. Alteori tipul de repartiţie teoretică se intuieşte din reprezentarea grafică a lui X^*, după care se face o ipoteză asupra legii, lucrând în această ipoteză (se vor verifica ipotezele statistice).

Definiţia 2.2 *Numim **statistică** (sau **funcţie de selecţie**) orice cantitate a cărei valoare poate fi calculată din datele unui eşantion. Având în vedere că valoarea unei statistici depinde de alegerea eşantionului, vedem statistica ca pe o variabilă aleatoare (şi prin urmare va fi notată cu literă mare).*

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

Statistica, ca v.a., depinde nu doar de tipul de distribuţie al populaţiei, de volumul eşantionului ales dar şi de metoda de alegere a eşantionului.

Definiţia 2.3 *Spunem că v.a. $X_1, \ldots, X_n$ formează o **selecţie aleatoare** (sau **eşantion aleator**) de mărime n dacă v.a. $X_1, \ldots, X_n$ sunt independente şi dacă ele urmează aceaşi distribuţie (acelaşi tip de repartiţie). Mai precis, cerem ca familia $(X_i)_{i=\overline{1,n}}$ să fie **i.i.d. (independentă şi identic distribuită)**.*

Remarca 2.4 *În cazul în care avem o selecţie aleatoare formată din v.a. $X_1, \ldots, X_n$ putem defini similar ca în Secţiunea 1.3.1 conceptele:*
momentul de selecţie de ordin r

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

media de selecţie

$$\bar{X} \overset{def}{=\!=} \frac{X_1 + \cdots + X_n}{n},$$

momentul centrat de selecţie de ordin r

$$\nu'_r \overset{def}{=\!=} \frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)^r}{n},$$

dispersia (sau varianţa) de selecţie

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

abaterea (sau deviaţia standard) de selecţie

$$S \overset{def}{=\!=} \sqrt{S^2},$$

dispersia (sau varianţa) modificată de selecţie

$$(S^*)^2 \overset{def}{=\!=} \frac{n}{n-1} S^2 = \frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}{n-1},$$

abaterea (sau deviaţia standard) modificată de selecţie

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

2.1 Problema estimaţiei

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

Fie o selecţie aleatoare formată din v.a. $X_1, \ldots, X_n$.

Să presupunem în continuare că avem o repartiţie 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 **funcţie de selecţie**)

$$\lambda_n^* = \varphi(X_1, \ldots, X_n) \tag{2.1}$$

astfel încât

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

în diferite sensuri.

Definiţia 2.5 *λ_n^* se numeşte **estimator consistent** pentru λ dacă*

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

Remarca 2.6 *Deci λ_n^* este estimator consistent dacă*

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

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

Definiţia 2.8 *λ_n^* se numeşte **estimator corect** pentru λ dacă*

$$\lim_{n\to\infty} \mathbb{E}(\lambda_n^*) = \lambda \quad \textit{şi} \quad \lim_{n\to\infty} D^2(\lambda_n^*) = 0.$$

Definiţia 2.9 *Diferenţa $\delta = \mathbb{E}(\lambda_n^*) - \lambda$ se numeşte **deplasarea** (sau **distorsiunea**) estimatorului λ_n^* .*

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

Definiţia 2.11 *λ_n^* se numeşte **estimator absolut corect** pentru λ dacă*

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

Deci $\delta = 0$ şi estimatorul absolut corect este nedeplasat.

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

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

2.1.1 Estimări punctuale ale momentelor repartiţiei teoretice

Se demonstrează că au loc următoarele estimări:

Teorema 2.13 *Să presupunem că populaţia statistică admite medie $\mu = \mathbb{E}(X)$ şi dispersie $\sigma^2 = D^2(X)$ (deci şi selecţia aleatoare $X_1, \ldots, X_n$ admite medie şi dispersie de selecţie).*

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

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

(iii) Momentul de selecţie de ordin r este un estimator absolut corect pentru momentul de ordin r al populaţiei $\mathcal{P}$.

(iv) Pentru un x fixat, funcţia de repartiţie de selecţie $F_n(x)$ este un estimator absolut corect pentru funcţia de repartiţie $F(x)$.

Demonstraţie. (i) Fie $\bar{X} \overset{def}{=\!=} \frac{X_1+\cdots+X_n}{n}$. Dar X_i sunt independente şi au aceeaşi distribuţie cu X, deci

$$\mathbb{E}(X_i) = \mathbb{E}(X) \overset{not}{=\!=} \mu \quad şi \quad D^2(X_i) = D^2(X) \overset{not}{=\!=} \sigma^2.$$

Atunci

$$\mathbb{E}\left(\bar{X}\right) = \frac{\sum_{i=1}^n \mathbb{E}(X_i)}{n} = \frac{n\mathbb{E}(X)}{n} = \mu,$$
$$D^2\left(\bar{X}\right) = \frac{\sum_{i=1}^n D^2(X_i)}{n^2} = \frac{nD^2(X)}{n^2} = \frac{\sigma^2}{n} \to 0, \quad \text{pentru } n \to \infty.$$

(ii) Fie

$$S^2 \overset{def}{=\!=} \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(X_i-\bar{X}\right)^2 = \overline{X^2}-\bar{X}^2 = \frac{1}{n}\sum\nolimits_{i=1}^{n}X_i^2-\bar{X}^2.$$

Dar X_i sunt independente şi au aceeaşi distribuţie cu X, deci

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

Utilizând relaţia

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

obţinem că

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

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

$$\left(S^*\right)^2 \overset{def}{=\!=} \frac{n}{n-1}S^2.$$

Astfel

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

Dacă prespunem în plus că populaţia statistică admite momentul centrat ν_4' de ordin 4, atunci se poate demonstra următoarea identitate

$$D^2\left(S^2\right)=\frac{\nu_4'-\left(\nu_2'\right)^2}{n}-\frac{2\left(\nu_4'-2\left(\nu_2'\right)^2\right)}{n^2}+\frac{\nu_4'-3\left(\nu_2'\right)^2}{n^3},$$

deci

$$D^2\left(S^2\right)\to 0, \text{ pentru } n\to\infty.$$

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

Exerciţiul 2.14 *Arătaţi că dispersia de selecţie S^2 este un estimator deplasat al dispersiei teoretice σ^2 şi determinaţi deplasarea.*

Exerciţiul 2.15 *Arătaţi că dispersia de selecţie S^2 este un estimator corect (dar nu absolut corect) al dispersiei teoretice.*

2.1.2 Estimări punctuale ale parametrilor repartiţiei teoretice

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

Metoda verosimilităţii maxime Fie selecţia aleatoare $(X_1, \ldots, X_n)$ cu densitatea $f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_k)$ (în cazul v.a. continue sau funcţia de frecvenţă, în cazul v.a. discrete), unde parametrii $\lambda_1, \ldots, \lambda_k$ au valori necunoscute.

Definiţia 2.16 *Funcţia*

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

*unde $x_1, \ldots x_n$ sunt observaţiile în urma unei selecţii de volum n, se numeşte **funcţia de verosimilitate**.*

Remarca 2.17 *Deoarece $X_1, \ldots, X_n$ formează o selecţie aleatoare, v.a. sunt independente şi identic repartizate, i.e. $f_{X_i} = f_X$, deci funcţia de verosimilitate este, în cazul continuu,*

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

unde $f(x_i, \lambda_1, \ldots, \lambda_k) := f_X(x_i, \lambda_1, \ldots, \lambda_k)$,
iar în cazul discret este

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

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

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

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

Etapele sunt următoarele:

(i) *calculăm* $V(\lambda_1, \ldots, \lambda_k)$ *şi* $\ln V(\lambda_1, \ldots, \lambda_k)$ *cu ajutorul variabilelor de selecţie;*

(ii) *rezolvăm sistemul de* k *ecuaţii şi* k *necunoscute*

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

(iii) *pentru soluţia* $\left(\lambda_1^0, \ldots, \lambda_k^0\right)$ *găsită mai sus verificăm condiţia suficientă de extrem*

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

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 condiţii destul de generale, estimatorii* λ_i^* *sunt consistenţi pentru* λ_i, $i = \overline{1,k}$.

Exerciţiul 2.22 *Într-un eşantion de 10 produse fabricate de o companie se observă că primul, al treilea şi ultimul nu sunt conform standardelor. Ştim că probabilitatea ca un produs să nu fie conform standardelor este* p*.*

Având în vedere eşantionul observat, să se obţină o estimare p^* *a valori parametrului* p*.*

Selecţia aleatoare $(X_1, \ldots, X_n)$ *are densitatea dată de* $f(p) := p{\cdot}q{\cdot}p{\cdot}q{\cdot}\ldots{\cdot}q{\cdot}p$ *(fiecare* X_i *sunt distribuite de tip Bernoulli, i.e.* $X_i \sim \mathcal{B}(1,p)$, $i = \overline{1,10}$*), deci funcţia de verosmilitate asociată eşantionului observat este dată de*

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

Să studiem maximele funcţiei V *care coincid cu maximele funcţiei*

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

Deoarece

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

obţinem punctul critic $p^* = 3/10$.

Acesta e punct de maxim deoarece

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

Valoarea $p^* = 3/10$ *este deci estimatorul parametrului* p *al distribuţiei populaţiei şi este valoarea care, pentru selecţia*

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

dată maximizează funcţia de verosmilitate $V\left(p\right)$.

Să remarcăm că în cazul în care se observa că într-un eşantion de 10 produse 3 sunt neconforme standardelor (nu ştim ordinea), obţinem că funcţia de verosmilitate este dată de

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

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

Exerciţiul 2.23 *Să se estimeze, folosind metoda verosimilităţii maxime, parametrul* λ *al repartiţiei exponenţiale cu densitatea* $f\left(x\right) = \lambda e^{-\lambda x}$, $\lambda > 0, x > 0$, *ştiind că rezultatele obţinute în urma efectuării unei selecţii de volum* 5 *sunt* $x_1 = 7$, $x_2 = 6.5$, $x_3 = 6.9$, $x_4 = 6.7$, $x_5 = 6.8$. *Generalizaţi rezultatul şi studiaţi dacă estimatorul găsit este sau nu deplasat.*

Fiecare v.a. a selecţiei aleatoare $(X_1, \ldots, X_5)$ *are densitatea dată de* $f\left(x\right) = \lambda e^{-\lambda x},\ \lambda > 0, x > 0$, *deci funcţia de verosmilitate asociată eşantionului observat este dată de*

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

Să studiem maximele funcţiei V *care coincid cu maximele funcţiei*

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

Deoarece

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

obţinem punctul critic

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

Acesta e punct de maxim deoarece

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

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

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

dată maximizează funcţia de verosmilitate $V(\lambda)$.

În general, pentru un eşantion aleator $(X_1, \ldots, X_n)$, *funcţia de verosmilitate asociată eşantionului 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 obţine că estimatorul parametrului λ *al distribuţiei populaţiei este dat de punctul de maxim* $\lambda^* = \dfrac{1}{\bar{X}}$.

Deoarece $\mathbb{E}\left(\bar{X}\right) = \mathbb{E}(X) = \dfrac{1}{\lambda}$, *obţinem că*

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

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

Exerciţiul 2.24 *Să se estimeze, folosind metoda verosimilităţii maxime, parametrii* μ *şi* σ^2 *ai repartiţiei normale, cu densitatea*

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

Fiecare v.a. a selecţiei aleatoare $(X_1, \ldots, X_n)$ *are densitatea dată de* $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, *deci funcţia de verosmilitate asociată eşantionului observat este dată de*

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

Să studiem maximele funcţiei V *care coincid cu maximele funcţiei*

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

Deoarece

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

obţinem punctul critic

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

Acesta e punct de maxim deoarece matricea Hessiană are forma

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

care în punctul critic are valoarea

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

$$=\begin{pmatrix} -\dfrac{n}{\left(\sigma^*\right)^2} & 0 \\ 0 & -\dfrac{2n}{\left(\sigma^*\right)^2}^{\,2} \end{pmatrix}$$

care este negativ definită.

Se obţine că estimatorul parametrilor μ şi σ^2 ai distribuţiei populaţiei este dat de punctul de maxim

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

Deoarece

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

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

Deoarece

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

obţinem că $\left(\sigma^*\right)^2=S^2$ este un estimator deplasat al lui σ^2.

Exerciţiul 2.25 *Să se estimeze, folosind metoda verosimilităţii maxime, parametrul λ al repartiţiei cu densitatea $f(x) = (1+\lambda)\, x^{\lambda}$, $\lambda > 0, x \in (0,1)$, ştiind că rezultatele obţinute în urma efectuării unei selecţii de volum 4 sunt $x_1 = 0.4$, $x_2 = 0.6$, $x_3 = 0.85$, $x_4 = 0.9$.*

Funcţia de verosmilitate asociată eşantionului 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 funcţiei $\ln V(\lambda) = 4\ln(1+\lambda) + \lambda \sum_{i=1}^{4} \ln x_i$ *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

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

Acesta e punct de maxim deoarece

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

Exerciţiul 2.26 *Să se estimeze, folosind metoda verosimilităţii maxime, parametrul λ al repartiţiei Poisson $\mathcal{P}(\lambda)$, considerându-se o selecţie de volum n. Arătaţi că estimatorul găsit este absolut corect.*

Fiecare v.a. a selecţiei aleatoare $(X_1, \ldots, X_n)$ are densitatea dată de $f(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{k_i}}{(k_i)!} e^{-\lambda}$, deci funcţia de verosmilitate asociată eşantionului observat este dată de

$$V(\lambda) := \prod_{i=1}^{n} \frac{\lambda^{k_i}}{(k_i)!} e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^{n} k_i}}{k_1! \cdot \ldots \cdot k_2!} e^{-n\lambda} .$$

Să studiem maximele funcţiei

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

Luând

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

obţinem punctul critic

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

Acesta e punct de maxim deoarece

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

Se obţine că estimatorul parametrului λ al distribuţiei populaţiei este dat de punctul de maxim

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

Deoarece

$$\mathbb{E}\left(\lambda^*\right) = \lambda \quad \Leftrightarrow \quad \mathbb{E}\left(\bar{X}\right) = \mathbb{E}\left(X\right),$$

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

Pe de altă parte,

$$D^2\left(\lambda^*\right) = D^2\left(\bar{X}\right) = \frac{1}{n}D^2\left(X\right) = \frac{1}{n}\lambda \to 0, \quad \textit{pentru } n \to +\infty,$$

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

Metoda momentelor Fie selecţia aleatoare $(X_1, \ldots, X_n)$ asociată caracteristicii X a populaţiei, cu densitatea $f_X\left(x, \lambda_1, \ldots, \lambda_k\right)$ (în cazul v.a. continue sau funcţia de frecvenţă, în cazul v.a. discrete), unde parametrii $\lambda_1, \ldots, \lambda_k$ au valori necunoscute.

Presupunem că există şi sunt finite momentele de ordin $1, 2, \ldots, k$ (notate $\mu_1, \mu_2, \ldots, \mu_k$). Dacă avem o selecţia aleatoare, din faptul ca momentele de selecţie de ordin r (notate μ_r') sunt estimatori absoluţi corecţi pentru momentele teoretice de ordin r, putem scrie că

$$\mu_r\left(\lambda_1, \ldots, \lambda_k\right) \simeq \mu_r', \quad r = \overline{1,k}\,.$$

Deoarece momentele teoretice depind de parametrii $\lambda_1, \ldots, \lambda_k$, se pot gasi estimatori pentru aceştia rezolvând sistemul

$$\mu_r\left(\lambda_1, \ldots, \lambda_k\right) = \mu_r', \quad r = \overline{1,k}\,.$$

Acesta este un sistem de k ecuaţii cu k necunoscute: $\lambda_1, \ldots, \lambda_k$.

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

Remarca 2.27 *Pentru estimarea punctuală a parametrilor unei repartiţii există şi* ***metoda celor mai mici pătrate****.*

Exerciţiul 2.28 *Să se estimeze, folosind metoda momentelor, parametrul λ al repartiţiei Poisson discrete şi infinite $\mathcal{P}(\lambda)$, considerându-se o selecţie de volum n. Arătaţi că estimatorul găsit este absolut corect.*

Exerciţiul 2.29 *Fie selecţia aleatoare $X_1, \ldots, X_n$ dată de timpul de servire a n clienţi la un anumit ghişeu. Caracteristica X a populaţiei se presupune că este distribuită de tip exponenţial de parametru λ. Să se estimeze, folosind metoda momentelor, parametrul λ al repartiţiei.*

Calculăm momentul teoretic de ordin 1, adică media teoretică şi obţinem

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

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

$$\mu_1' = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}.$$

Egalând obţinem

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

deci estimatorul lui λ folosind metoda momentelor este

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

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

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

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

Pe de altă parte, momentele de selecţie de ordin 1 şi 2 sunt respectiv

$$\mu_1' = \frac{\sum_{i=1}^{5} x_i}{5} = 2.44 \quad şi \quad \mu_2' = \frac{\sum_{i=1}^{5} x_i^2}{5} = 8.73.$$

Obţinem deci sistemul

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

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 soluţiile

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

deci

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

şi apoi

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

Obţinem estimatorii $a^* = -0.4461$ *şi* $b^* = 5.3261$.

În cazul general al unei selecţii aleatoare $(X_1, \ldots, X_n)$ *momentele de selecţie de ordin* 1 *şi* 2 *sunt respectiv*

$$\mu_1' = \frac{\sum_{i=1}^n x_i}{n} \quad \text{şi} \quad \mu_2' = \frac{\sum_{i=1}^n x_i^2}{n}\,.$$

Obţinem deci sistemul

$$\begin{cases} \dfrac{a+b}{2} = \mu_1'\,, \\ \dfrac{a^2 + ab + b^2}{3} = \mu_2'\,. \end{cases}$$

Deci

$$\begin{aligned} & a^2 + a\left(2\mu_1' - a\right) + \left(2\mu_1' - a\right)^2 = 3\mu_2' \\ \Leftrightarrow\ & a^2 - 2\mu_1'\,a + 4\left(\mu_1'\right)^2 - 3\mu_2' = 0, \end{aligned}$$

care are soluţiile

$$a_{1,2} = \frac{2\mu_1' \pm \sqrt{(2\mu_1')^2 - 4 \cdot \left(4\,(\mu_1')^2 - 3\mu_2'\right)}}{2} = \mu_1' \pm \sqrt{3}\sqrt{\mu_2' - (\mu_1')^2}\,,$$

deci alegem $a = \mu_1' - \sqrt{3}\sqrt{\mu_2' - (\mu_1')^2}$ *şi apoi* $b = \mu_1' + \sqrt{3}\sqrt{\mu_2' - (\mu_1')^2}$.

Obţinem estimatorii

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

deoarece media de selecţie

$$\bar{X} = \mu_1'$$

iar dispersia de selecţie

$$S^2 = \nu_2' = \mu_2' - \left(\mu_1'\right)^2 .$$

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

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

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

în urma obţinerii eşantionului $x_1 = 2.4$, $x_2 = 2.7$, $x_3 = 2.8$, $x_4 = 2.2$, $x_5 = 2.4$.

Conform calculelor din cadrul cursului de "Teoria Probabilităţilor" momentele teoretice de ordinul 1 şi 2 sunt date de

$$\mathbb{E}\left(X\right) = \frac{p}{\lambda} \quad şi \quad \mathbb{E}\left(X^2\right) = \frac{p\left(p+1\right)}{\lambda^2} .$$

Pe de altă parte, momentele de selecţie de ordin 1 şi 2 sunt respectiv

$$\mu_1' = \frac{\sum_{i=1}^5 x_i}{5} = 2.5 \quad şi \quad \mu_2' = \frac{\sum_{i=1}^5 x_i^2}{5} = 6.298.$$

Obţinem deci sistemul

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

Deci

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

iar

$$p = 130.2075.$$

Obţinem estimatorii $p^* = 130.2075$ şi $\lambda^* = 52.083$.

2.1.3 Estimări prin intervale de încredere ale momentelor repartiţiei teoretice

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

Fie selecţia aleatoare $(X_1, \ldots, X_n)$ asociată caracteristicii X a populaţiei şi fie λ valoarea teoretică ce dorim să o estimăm prin această metodă.

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

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

Numărul

$$\delta \simeq 1$$

se va numi **nivel de încredere** sau siguranţa estimaţiei.

Numărul

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

se numeşte **prag de semnificaţie.**

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

Conform[1] Propoziţiei 3.51 din cadrul *Teoriei Probabilităţilor*, variabila aleatoare standardizată $\frac{\bar{X} - \mathbb{E}(\bar{X})}{\sqrt{D^2(\bar{X})}}$ va fi atunci repartizată normal standard, i.e.

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

De asemenea, mai ştim că $\bar{X} \sim \mathcal{N}(m, \sigma^2/n)$ (vezi[2] Exerciţiul 11, Capitolul 4 din cadrul *Teoriei Probabilităţilor*).

---

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

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

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

Pentru a determina intervalul de încredere pentru m punem condiţia

$$\mathbb{P}\left(|Z| < z_{\alpha/2}\right) = \delta \quad \Leftrightarrow \quad \mathbb{P}\left(\left|\frac{\bar{X} - m}{\sigma/\sqrt{n}}\right| < z_{\alpha/2}\right) = \delta \tag{2.3}$$

cu δ cunoscut, adică vom obţine

$$2\Phi\left(z_{\alpha/2}\right) - 1 = \delta \quad \Leftrightarrow \quad \Phi\left(z_{\alpha/2}\right) = \frac{1+\delta}{2}\,,$$

deci valoarea $z_{\alpha/2}$ se va obţine din tabele.

Deci

$$|Z| < z_{\alpha/2} \quad \Leftrightarrow \quad \left|\tfrac{\bar{X}-m}{\sigma/\sqrt{n}}\right| < z_{\alpha/2} \quad \Leftrightarrow \quad \left|\bar{X} - m\right| < \frac{\sigma}{\sqrt{n}}\, z_{\alpha/2}$$

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

Propoziţia 2.34 ***În cazul în care populaţia are caracteristica*** $X \sim \mathcal{N}\left(m, \sigma^2\right)$ ***iar*** σ^2 ***este cunoscut, intervalul de încredere pentru medie este dat de***

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

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

$$\Phi\left(z_{\alpha/2}\right) = \frac{1+\delta}{2} \quad \textit{(sau } \Phi\left(z_{\alpha/2}\right) = 1 - \frac{\alpha}{2}\ \textit{)}\,.$$

Din graficul repartiţiei normale standard se poate gasi interpretarea lui $z_{\alpha/2}$. Astfel

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

Remarca 2.35 *Fie* a *valoarea exactă a unei mărimi. În acest caz* $m = \mathbb{E}\left(a\right) = a$ *şi* $\bar{a}$ *valoarea aproximativă a acestei mărimi (obţinută cu ajutorul unui aparat). Faptul că* σ *este cunoscut reprezintă precizia măsurătorilor (siguranţa aparatului). Intervalul de încredere pentru* a *este*

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

deoarece

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

O problemă comună care intervine în practică este aceea de a determina **numărul minim de observaţii necesare pentru a obţine o anumită precizie a rezultatelor**. În acest sens utilizăm tot relaţia (2.3). Presupunem că sunt date δ, siguranţa estimaţiei, şi Δ (eroarea absolută). Atunci

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

Deci din

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

obţinem

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

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

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

Exemplul 2.36 *Care este numărul de măsurători necesare pentru a obţine un interval de încredere de* 95% *cu o eroare absolută de* 2*, ştiind că abaterea empirică modificată a fost obţinută şi este de* 2.6 *?*

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

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

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

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*).*

Propoziţia 2.37 ***În cazul în care volumul selecţiei*** $n > 30$, ***populaţia are caracteristica*** X ***care urmează o distribuţie oarecare, nu neapărat de tip normal, iar*** σ^2 ***este cunoscut, intervalul de încredere pentru medie este dat de***

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

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

$$\Phi\left(z_{\alpha/2}\right) = \frac{1+\delta}{2} \quad \left(\text{sau } \Phi\left(z_{\alpha/2}\right) = 1 - \frac{\alpha}{2}\right).$$

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 selecţiei $n > 30$, X repartiţie oarecare şi σ necunoscut considerăm

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

unde

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

Propoziţia 2.38 *Relaţie (2.4) ne dă acum* ***intervalul de încredere pentru medie în cazul în care volumul selecţiei*** $n > 30$, ***populaţia are caracteristica*** X ***care urmează o distribuţie oarecare, nu neapărat de tip normal, iar*** σ^2 ***este necunoscut:***

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

Exerciţiul 2.39 *În urma efectuării unei selecţii de volum* 250 *s-au obţinut* $\bar{x} = 126.18$ *şi* $S^* = 4.05$. *Determinaţi intervalul de încredere pentru media teoretică* m *corespunzătoare pragului de semnificaţie* $\alpha = 0.01$.

În cazul în care volumul selecţiei $n \leq 30$, X este repartizată $\mathcal{N}(m, \sigma^2)$ şi σ necunoscut trebuie să reamintim, mai întâi, legătura dintre distribuţia normală şi distribuţia χ^2 precum şi legătura dintre distribuţia normală, distribuţia χ^2 şi distribuţia Student (pentru demonstraţii vezi Propoziţia 3.129 şi Propoziţia 3.132 din cadrul *Teoriei Probabilităţilor*).

Propoziţia 2.40 *Dacă* A_i *sunt v.a. independente şi normale de tip* $\mathcal{N}(0, \sigma^2)$, *atunci*

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

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

Propoziţia 2.41 *Dacă* $X \sim \mathcal{N}(0, \sigma^2)$ *şi* $Y \sim \chi^2(a, \sigma)$ *sunt două v.a. independente, atunci distribuţia*

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

adică T *este distribuită Student de parametru* a.

Aplicând acum Propoziţia 2.40 şi[3] Propoziţia 3.130 din cadrul *Teoriei Probabilităţilor*, deducem că

$$X_i \sim \mathcal{N}\left(m, \sigma^2\right) \quad \Leftrightarrow \quad \left(X_i - m\right) \sim \mathcal{N}\left(0, \sigma^2\right)$$
$$\Rightarrow \quad \sum_{i=1}^{n} \left(X_i - m\right)^2 \sim \chi^2\left(n, \sigma\right)$$

iar

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

Pe de altă parte, avem că

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

deci, utilizând (2.6), (2.7) şi[4] Propoziţia 3.82 din cadrul *Teoriei Probabilităţilor*,

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

Pe de altă parte,

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

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

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

Vom obţinem astfel că variabila

$$Y := \frac{\sqrt{n}}{S^*}\left(\bar{X} - m\right) = \frac{\sqrt{n}\left(\bar{X} - m\right)}{\sqrt{\frac{nS^2}{n-1}}} \sim t\left(n-1\right) \tag{2.9}$$

este distribuită Student de parametru $n-1$.

Acum avem din (2.3) şi (2.9)

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

deci

$$\mathbb{P}\left(Y > t\right) = \alpha/2$$

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

Propoziţia 2.42 ***În cazul în care volumul selecţiei*** $n \leq 30$, ***populaţia are caracteristica*** $X \sim \mathcal{N}\left(m, \sigma^2\right)$ ***iar*** σ^2 ***este necunoscut, intervalul de încredere pentru medie este dat de***

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

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

Exerciţiul 2.43 *În urma efectuării unei selecţii de volum* 20 *s-au obţinut* $\bar{x} = 0.149$ *şi* $S^* = 0.048$. *Determinaţi intervalul de încredere pentru media teoretică* m *corespunzătoare pragului de semnificaţie* $\alpha = 0.05$.

Exerciţiul 2.44 *Acelaşi enunţ pentru:* $n = 16$, $\bar{x} = 25.4$, $S^* = 1.3$, $\delta = 95\%$. *Determinaţi intervalul de încredere pentru medie utilizând valoarea* E_{95} *(calculată pentru o v.a distribuită normal). Care interval de încredere este mai mic şi de ce? Citiţi tabelul repartiţiei Student corespunzător la* $n = \infty$ *grade de libertate şi comparaţi cu* E_α *de la distribuiţia normală.*

Interval de încredere pentru dispersia teoretică Fie X o caracteristică considerată asociată unei populaţii şi considerăm că $X \sim \mathcal{N}\left(m, \sigma^2\right)$. Deci media $\mathbb{E}\left(X\right) = m$ şi $D^2\left(X\right) = \sigma^2$. Considerăm statistica (sau funcţia de selecţie)

$$Y = \frac{nS^2}{\sigma^2}. \tag{2.11}$$

Din proprietatea (2.8) deducem că $nS^2 \sim \chi^2\left(n-1, \sigma\right)$, deci Y va fi repartizată atunci (vezi Propoziţia 3.130 din cadrul *Teoriei Probabilităţilor*)

$$Y \sim \chi^2\left(n-1, 1\right). \tag{2.12}$$

Din condiţia

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

se vor putea determina, utilizând tabelul distribuţiei χ^2, valorile χ_1^2 şi χ_2^2, unde δ este nivelul de încredere (sau siguranţa estimaţiei) 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 ecuaţie cu două necunoscute.

Cantitatea χ_1^2 se va determina din relaţia

$$\mathbb{P}\left(Y > \chi_1^2\right) = q_1 = 1 - \frac{\alpha}{2} = 1 - \frac{1-\delta}{2} \tag{2.14}$$

iar cantitatea χ_2^2 se va determina din relaţia

$$\mathbb{P}\left(Y > \chi_2^2\right) = q_2 = \frac{\alpha}{2} = \frac{1-\delta}{2}, \tag{2.15}$$

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

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

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

unde $f\left(x\right)$ reprezintă densitatea de repartiţie a unei v.a. Y repartizată $\chi^2\left(n-1, 1\right)$.

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

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

adică obţinem următorul rezultat.

Propoziţia 2.45 ***În cazul în care volumul selecţiei*** $n \leq 30$, ***populaţia are caracteristica*** $X \sim \mathcal{N}\left(m, \sigma^2\right)$ ***iar*** m ***este necunoscut, intervalul de încredere pentru dispersia teoretică*** σ^2 ***este dat de***

$$\frac{nS^2}{\chi_2^2} < \sigma^2 < \frac{nS^2}{\chi_1^2}\,, \tag{2.16}$$

unde valorile χ_1^2 *şi* χ_2^2 *sunt citite din tabelul distribuţiei* $\chi^2\left(n-1,1\right)$ *folosind relaţiile (2.14-2.15).*

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

$$\frac{S\sqrt{n}}{\chi_2} < \sigma < \frac{S\sqrt{n}}{\chi_1}\,. \tag{2.17}$$

Remarca 2.46 *De fapt, folosind notaţii intuitive, avem că cele două valori* χ_1^2 *şi* χ_2^2 *pot fi scrise sub forma*

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

Remarca 2.47 *Când* $n > 30$ *folosim faptul că*

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

(vezi[5] *Remarca 5.72 din cadrul Teoriei Probabilităţilor).*

În cazul nostru, pentru n *suficient de mare,*

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

Atunci vom putea folosi funcţia de repartiţie Φ *şi deducem, având în vedere că* $\dfrac{nS^2}{\sigma^2} \sim \chi^2\left(n-1,1\right)$,

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

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

(vezi Propoziţia 3.47 din cadrul Teoriei Probabilităţilor).

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

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

şi 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 funcţiei de repartiţie Φ*, asociată unei v.a.* $Z\sim\mathcal{N}(0,1)$*, vom obţine valoarea* χ_2^2 *iar apoi valoarea* χ_1^2*.*

Exerciţiul 2.48 *În urma măsurării unei distanţe obţinem o colecţie de* 20 *de date. Abaterea empirică modificată obţinută este* 1.8*. Care este intervalul de incredere pentru dispersie cu* $\delta=95\%$*.*

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

Luând

$$X=\frac{n_1S_1^2}{\sigma_1^2} \quad \text{şi} \quad Y=\frac{n_2S_2^2}{\sigma_2^2}$$

vom obţine că

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

deci X are (n_1-1) grade de libertate şi 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

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

Considerăm statistica (sau funcţia de selecţie)

$$T\overset{def}{=}\frac{\frac{X}{n_1-1}}{\frac{Y}{n_2-1}}.$$

În cazul nostru,

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

deci, folosind[6] Propoziţia 3.135 din cadrul *Teoriei Probabilităţilor*,

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

Din condiţia

$$\mathbb{P}\left(F_1 < T < F_2\right) = \delta$$

se vor putea determina, utilizând tabelul distribuţiei Fisher, valorile F_1 şi F_2.

Avem evident că

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

care va reprezenta o ecuaţie cu două necunoscute.

Cantitatea F_1 se va determina din relaţia

$$\mathbb{P}\left(T > F_1\right) = q_1 = 1 - \frac{\alpha}{2} = 1 - \frac{1-\delta}{2} \tag{2.18}$$

iar cantitatea F_2 se va determina din relaţia

$$\mathbb{P}\left(T > F_2\right) = q_2 = \frac{\alpha}{2} = \frac{1-\delta}{2}. \tag{2.19}$$

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

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

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

[6] Dacă $X \sim \chi^2\left(a, 1\right)$ şi $Y \sim \chi^2\left(b, 1\right)$ atunci v.a.

$$T \overset{def}{=\!=} \frac{X/a}{Y/b} \sim F\left(a, b\right),$$

adică fracţia $\frac{X/a}{Y/b}$ este distribuită Fisher de parametrii a şi b.

Acum având cunoscute valorile F_1 şi 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 obţinem următorul rezultat.

Propoziţia 2.49 ***În cazul în care volumul selecţiei*** $n \leq 30$, ***populaţiile au caracteristicile*** $X_1 \sim \mathcal{N}\left(m_1, \sigma_1^2\right)$, $X_2 \sim \mathcal{N}\left(m_2, \sigma_2^2\right)$ ***iar*** m_1, m_2 ***şi*** σ_1^2, σ_2^2 ***sunt necunoscute, intervalul de încredere pentru raportul a două dispersii este dat de***

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

unde valorile F_1 şi F_2 sunt citite din tabelul distribuţiei $F\left(n_1 - 1, n_2 - 1\right)$ folosind relaţiile (2.18-2.19).

Remarca 2.50 *De fapt, folosind notaţii intuitive, avem că cele două valori F_1 şi F_2 pot fi scrise sub forma*

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

Se ştie că are loc **următoarea relaţie între valori**

$$F_{1-\alpha,a,b} = \frac{1}{F_{\alpha,b,a}}\,. \tag{2.21}$$

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

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

precum şi valoarea ariei α corespunzătoare distribuţiei $F\left(15, 5\right)$

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

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

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

Obţinem astfel **intervalul de încredere pentru raportul a două dispersii** sub forma

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

Exerciţiul 2.51 *În urma a* 31 *de observaţii se obţine abaterea empirică modificată* $S^* = 1.5$*, iar la* 25 *de observaţii se obţine abaterea empirică modificată* $S^* = 0.7$*. Să se determine intervalul de încredere pentru raportul dispersiilor considerându-se un prag de semnificaţie de* 5%*.*
Intervalul obţinut conţine valoarea 1 *? Interpretaţi rezultatul (care este deci probabilitatea ca* $\sigma_1^2 \neq \sigma_2^2$*?).*

Capitolul 3

Verificarea ipotezelor statistice

Aşa cum se poate vedea în exemplele din capitolul precedent, de cele mai multe ori nu suntem interesaţi de marginile intervalului construit, ci mai degrabă de problema când intervalul construit conţine media sau dispersia teoretică. Deci problema esenţială este aceea de a putea preciza dacă o selecţie statistică este consistentă în raport cu populaţia întreagă. Procedura utilizată de a testa validitatea unei statistici este cunoscută sub numele de **test statistic**.

Ipoteza statistică este o ipoteză care se face relativ la parametrii unei repartiţii sau la legea de repartiţie pe care o urmează o v.a.

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

$$(\mathrm{H}_0) : \theta = \theta^0$$

ipoteza conform căreia $\theta = \theta^0$, unde θ^0 este o valoare calculată a parametrului pe baza unui eşantion $x_1, \ldots, x_n$ din populaţia $\mathcal{P}$.

Ipoteza (H_0) se numeşte ipoteza nulă.

Pot interveni şi ipoteze alternative

$$\begin{array}{l} (\mathrm{H}_0) : \theta = \theta^0 \\ (\mathrm{H}_1) : \theta = \theta^1 \end{array} \quad \text{sau} \quad \begin{array}{l} (\mathrm{H}_0) : \theta = \theta^0 \\ (\mathrm{H}_1) : \theta \neq \theta^0 \end{array}$$

Definiţia 3.1 *A testa o ipoteză înseamnă a lua o decizie dacă ipoteza se respinge sau dacă ipoteza se acceptă.*

Definiţia 3.2 *Se numeşte **test statistic** orice procedură de verificare a unei ipoteze statistice.*

Definiţia 3.3 *Orice test statistic se bazează pe un criteriu de testare.* ***Citeriu de testare*** *a unei ipoteze statistice este o statistică (o funcţie de selecţie)* $u = u(x_1, \ldots, x_n)$ *satisfăcând condiţiile*

1. funcţia de selecţie depinde de ipoteza făcută;

2. dacă ipoteza făcută se acceptă atunci repartiţia teoretică este complet determinată.

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

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

Vom nota cu

$$\mathcal{U} = \{u(x_1, \ldots, x_n) \in \mathbb{R} : \text{ipoteza } (\mathrm{H}_0) \text{ este adevărată şi } \mathbb{P}(u(x_1, \ldots, x_n) \in U) = \alpha\}$$

şi cu

$$\mathcal{V} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : u(x_1, \ldots, x_n) \in U, \text{ ipoteza } (\mathrm{H}_0) \text{ este adevărată şi } \mathbb{P}((x_1, \ldots, x_n) \in V) = \alpha\}$$

Dacă $(x_1, \ldots, x_n) \in \mathcal{V}$, adică $u(x_1, \ldots, x_n) \in \mathcal{U}$, atunci ipoteza (H_0) se respinge.

Dacă $(x_1, \ldots, x_n) \notin \mathcal{V}$, adică $u(x_1, \ldots, x_n) \notin \mathcal{U}$, atunci ipoteza (H_0) se acceptă.

Mulţimea $\mathcal{V}$ se numeşte **regiunea critică** şi mulţimea $\mathbb{R}^n \setminus \mathcal{V}$ se numeşte **regiunea de acceptare**.

3.1 Ipoteze asupra mediilor

În cazul în care σ se cunoaşte vom folosi distribuiţia $\mathcal{N}(0, 1)$ şi statistica (vezi şi definiţia (2.2))

$$z \overset{def}{=\!=} \frac{\bar{x} - m}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

În cazul în care σ nu se cunoaşte vom folosi distribuiţia $t(n)$.

3.1.1 Compararea mediei unei populaţii statistice

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

$$(\mathrm{H}_0) : m = m_0.$$

Testul z bilateral Verificăm ipoteza nulă

$$(\mathrm{H}_0) : m = m_0$$

faţă de altă ipoteză

$$(\mathrm{H}_1) : m = m_1 \neq m_0.$$

Dacă ipoteza nulă este acceptată atunci

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

Pentru α fixat se va determina, din tabelul repartiţiei Normale standard, valoarea critică z_α astfel încât

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

sau echivalent

$$\delta = 1 - \alpha = \mathbb{P}(-z_\alpha \leq z_c \leq z_\alpha) = 2\Phi(z_\alpha) \Leftrightarrow \Phi(z_\alpha) = \delta/2 = (1-\alpha)/2.$$

Exerciţiul 3.5 *De făcut graficul*

Obţinem atunci că **intervalul $[-z_\alpha, z_\alpha]$ este interval de acceptare iar regiunea $|z| > z_\alpha$ este regiune critică**, adică

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

Testul z bilateral înseamnă: dacă $z_c \in [-z_\alpha, z_\alpha]$ atunci (H_0) se acceptă, şi dacă $z_c \notin [-z_\alpha, z_\alpha]$ atunci (H_0) se respinge.

Testul z unilateral stânga Verificăm ipoteza nulă

$$(\mathrm{H}_0) : m = m_0$$

faţă de altă ipoteză

$$(\mathrm{H}_1) : m = m_1 < m_0.$$

Dacă ipoteza nulă este acceptată atunci

$$z_c \overset{def}{=\!=} \frac{\bar{x} - m_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

Pentru α fixat se va determina, din tabelul repartiţiei Normale, valoarea critică z_α astfel încât

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

sau echivalent

$$\begin{aligned}\delta &= 1-\alpha = \mathbb{P}(z_c \geq -z_\alpha) = 1-\mathbb{P}(z_c < -z_\alpha) = 1-\mathbb{P}(z_c > z_\alpha)\\ &= \mathbb{P}(z_c \leq z_\alpha) = F(z_\alpha) = \frac{1}{2}+\Phi(z_\alpha) \Leftrightarrow \Phi(z_\alpha) = \delta - 1/2 = 1/2-\alpha.\end{aligned}$$

Exerciţiul 3.6 *De făcut graficul*

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

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

Testul z unilateral stânga înseamnă: dacă $z_c \geq -z_\alpha$ atunci (H_0) se acceptă, şi dacă $z_c < -z_\alpha$ atunci (H_0) se respinge.

Testul z unilateral dreapta Verificăm ipoteza nulă

$$(\mathrm{H}_0) : m = m_0$$

faţă de altă ipoteză

$$(\mathrm{H}_1) : m = m_1 > m_0.$$

Dacă ipoteza nulă este acceptată atunci

$$z_c \overset{def}{=\!=} \frac{\bar{x}-m_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

Pentru α fixat se va determina, din tabelul repartiţiei Normale, valoarea critică z_α astfel încât

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

sau echivalent

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

Exerciţiul 3.7 *De făcut graficul*

Obţinem atunci că **intervalul** $(-\infty, z_\alpha]$ **este interval de acceptare iar intervalul** (z_α, ∞) **este regiune critică**, adică

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

Testul z **unilateral dreapta înseamnă: dacă** $z_c \leq z_\alpha$ **atunci** (H_0) **se acceptă, şi dacă** $z_c > z_\alpha$ **atunci** (H_0) **se respinge.**

3.1.2 Compararea mediilor a două populaţii statistice

Fie $X \sim \mathcal{N}\left(m_1, \sigma_1^2\right)$ şi $Y \sim \mathcal{N}\left(m_2, \sigma_2^2\right)$ cu σ_1, σ_2 cunoscuţi. Verificăm ipoteza

$$(\mathrm{H}_0) : m_1 = m_2.$$

Această se foloseşte când, în condiţii diferite, de obţinere a unui produs cu aceeaşi valoare nominală a unui parametru, se constată deosebiri între valorile medii. Se pune problema dacă este vorba de calităţi diferite ale produselor sau de abateri întâmplătoare. Statistica utilizată va fi

$$z \overset{def}{=\!=} \frac{(\bar{x} - \bar{y}) - (m_1 - m_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}.$$

Testul z **bilateral** Verificăm ipoteza

$$(\mathrm{H}_0) : m_1 = m_2$$

faţă de altă ipoteză

$$(\mathrm{H}_1) : m_1 \neq m_2.$$

Deoarece X, Y sunt repartizate normal obţinem că.

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

deci (fără demonstraţie) se poate arăta că

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

Dacă ipoteza (H_0) se acceptă, atunci

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

şi deci

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

Pentru α fixat se va determina, din tabelul repartiţiei Normale, valoarea critică z_α astfel încât

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

sau echivalent

$$\delta = 1-\alpha = \mathbb{P}(-z_\alpha \leq z_c \leq z_\alpha) = 2\Phi(z_\alpha) \Leftrightarrow \Phi(z_\alpha) = \delta/2 = (1-\alpha)/2.$$

Obţinem atunci că **intervalul** $[-z_\alpha, z_\alpha]$ **este interval de acceptare iar regiunea** $|z| > z_\alpha$ **este regiune critică**.

Testul z **bilateral înseamnă: dacă** $z_c \in [-z_\alpha, z_\alpha]$ **atunci** (H_0) **se acceptă, şi dacă** $z_c \notin [-z_\alpha, z_\alpha]$ **atunci** (H_0) **se respinge.**

Testul z **unilateral stânga** Verificăm ipoteza

$$(\mathrm{H}_0) : m_1 = m_2$$

faţă de altă ipoteză

$$(\mathrm{H}_1) : m_1 < m_2.$$

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

$$z_c = \frac{\bar{x}-\bar{y}}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \sim \mathcal{N}(0,1).$$

Pentru α fixat se va determina, din tabelul repartiţiei Normale, valoarea critică z_α astfel încât

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

sau echivalent

$$\begin{aligned} \delta &= 1-\alpha = \mathbb{P}(z_c \geq -z_\alpha) = 1-\mathbb{P}(z_c < -z_\alpha) = 1-\mathbb{P}(z_c > z_\alpha) \\ &= \mathbb{P}(z_c \leq z_\alpha) = F(z_\alpha) = \frac{1}{2}+\Phi(z_\alpha) \Leftrightarrow \Phi(z_\alpha) = \delta - 1/2 = 1/2-\alpha. \end{aligned}$$

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

Testul z **unilateral stânga înseamnă: dacă** $z_c \geq -z_\alpha$ **atunci** (H_0) **se acceptă, şi dacă** $z_c < -z_\alpha$ **atunci** (H_0) **se respinge.**

Testul z unilateral dreapta Verificăm ipoteza nulă

$$(\mathrm{H}_0) : m_1 = m_2$$

faţă de altă ipoteză

$$(\mathrm{H}_1) : m_1 > m_2.$$

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

$$z_c = \frac{\bar{x} - \bar{y}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim \mathcal{N}(0,1).$$

Pentru α fixat se va determina, din tabelul repartiţiei Normale, valoarea critică z_α astfel încât

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

sau echivalent

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

Testul z unilateral dreapta înseamnă: dacă $z_c \leq z_\alpha$ atunci (H_0) se acceptă, şi dacă $z_c > z_\alpha$ atunci (H_0) se respinge.

Remarca 3.8 *Dacă $\sigma_1^2 = \sigma_2^2 = \sigma^2$ atunci $z_c = \frac{\bar{x}-\bar{y}}{\sigma\sqrt{1/n_1+1/n_2}}$.*

Remarca 3.9 *Dacă n_1 şi n_2 sunt suficienţi de mari şi σ_1^2 şi σ_2^2 necunoscuţi, atunci putem aproxima*

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

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

Remarca 3.10 *Testele prezentate mai sus pot fi utilizate şi în cazul în care X, Y nu urmează legea normală dar n_1 şi n_2 sunt suficienţi de mari, deoarece în acest caz, conform theoremei lui Liapunov, 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 Compararea mediei unei populaţii statistice (σ necunoscut şi n mic)

Fie $X \sim \mathcal{N}\left(m, \sigma^2\right)$ cu σ necunoscut. Verificăm ipoteza

$$(\mathrm{H}_0) : m = m_0.$$

Statistica utilizată va fi

$$t \overset{def}{=\!=} \frac{\bar{x} - m}{S^*/\sqrt{n}}.$$

Având în vedere relaţia (2.9) şi faptul că

$$t = \frac{\sqrt{n}\,(\bar{x} - m)}{\sqrt{\frac{nS^2}{n-1}}}$$

obţinem că

$$t \sim t\,(n-1)$$

Testul t bilateral Verificăm ipoteza

$$(\mathrm{H}_0) : m = m_0$$

faţă de altă ipoteză

$$(\mathrm{H}_1) : m = m_1 \neq m_0.$$

Dacă ipoteza este acceptată atunci

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

Atunci $\mathcal{V} = \{\,(x_1, \ldots, x_n) \in \mathbb{R}^n : |t_c| > t_\alpha\}$ şi $\mathcal{U} = (-\infty, -t_\alpha) \cup (t_\alpha, \infty)$.

Pentru α fixat se va determina, din tabelul repartiţiei Student, valoarea critică t_α astfel încât (folosim şi simetria graficului)

$$\begin{aligned} \alpha &= \mathbb{P}\left(|t_c| > t_\alpha\right) \Leftrightarrow \alpha = \mathbb{P}\left(t_c < -t_\alpha\right) + \mathbb{P}\left(t_c > t_\alpha\right) \Leftrightarrow \alpha = 2\,\mathbb{P}\left(t_c > t_\alpha\right) \\ &\Rightarrow \mathbb{P}\left(t_c > t_\alpha\right) = \alpha/2 \end{aligned}$$

sau echivalent

$$\delta = 1 - \alpha = \mathbb{P}\left(-t_\alpha \leq t_c \leq t_\alpha\right).$$

Exerciţiul 3.11 *De făcut graficul*

Obţinem atunci că **intervalul** $[-t_\alpha, t_\alpha]$ **este interval de acceptare iar regiunea** $|t| > t_\alpha$ **este regiune critică,** adică

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

Testul t **bilateral înseamnă: dacă** $t_c \in [-t_\alpha, t_\alpha]$ **atunci** (H_0) **se acceptă, şi dacă** $t_c \notin [-t_\alpha, t_\alpha]$ **atunci** (H_0) **se respinge.**

Exerciţiul 3.12 *Media teoretică a unei distanţe este* $m = 400.008$ *m. S-au făcut* $n = 20$ *de observaţii şi s-a găsit media* $\bar{x} = 400.012$ *m. şi abaterea empirică modificată* $S^* = 0.0020$ *m. Să se cerceteze, aplicându-se* Testul t bilateral, *dacă media de selecţie (media distanţelor observate) diferă semnificativ sau nu de media teoretică, considerându-se un prag de semnificaţie de* $\alpha = 0.05$.

Determinaţi (utilizând capitolul precedent) intervalul de încredere pentru medie.

Testul t **unilateral stânga** Verificăm ipoteza

$$(\mathrm{H}_0) : m = m_0$$

faţă de altă ipoteză

$$(\mathrm{H}_1) : m = m_1 < m_0.$$

Dacă ipoteza este acceptată atunci

$$t_c \overset{def}{=\!=} \frac{\bar{x} - m_0}{S^*/\sqrt{n}} \sim t\,(n-1)\,.$$

Pentru α fixat se va determina, din tabelul repartiţiei Student, valoarea critică t_α astfel încât

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

Exerciţiul 3.13 *De făcut graficul*

Obţinem atunci că **intervalul** $[-t_\alpha, \infty)$ **este interval de acceptare iar intervalul** $(-\infty, -t_\alpha)$ **este regiune critică,** adică

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

Testul t **unilateral stânga înseamnă: dacă** $t_c \geq -t_\alpha$ **atunci** (H_0) **se acceptă, şi dacă** $t_c < -t_\alpha$ **atunci** (H_0) **se respinge.**

Exerciţiul 3.14 *Media teoretică a unei distanţe este* $m = 400.008$ *m. S-au făcut* $n = 20$ *de observaţii şi s-a găsit media* $\bar{x} = 400.012$ *m. şi abaterea empirică* $S = 0.0019$ *m. Să se cerceteze, aplicându-se* Testul t *unilateral stânga, dacă media de selecţie (media distanţelor observate) diferă semnificativ sau nu de media teoretică, considerându-se un prag de semnificaţie de* $\alpha = 0.05$*.*

Determinaţi (utilizând capitolul precedent) intervalul de încredere pentru medie.

Testul t unilateral dreapta Verificăm ipoteza nulă

$$(\mathrm{H}_0) : m = m_0$$

faţă de altă ipoteză

$$(\mathrm{H}_1) : m = m_1 > m_0.$$

Dacă ipoteza nulă este acceptată atunci

$$t_c \overset{def}{=\!=} \frac{\bar{x} - m_0}{S^*/\sqrt{n}} \sim t\,(n-1)\,.$$

Pentru α fixat se va determina, din tabelul repartiţiei Student, valoarea critică t_α astfel încât

$$\alpha = \mathbb{P}\,(t_c > t_\alpha)$$

sau echivalent

$$\delta = 1 - \alpha = \mathbb{P}\,(t_c \leq t_\alpha)\,.$$

Exerciţiul 3.15 *De făcut graficul*

Obţinem atunci că **intervalul** $(-\infty, t_\alpha]$ **este interval de acceptare iar intervalul** (t_α, ∞) **este regiune critică**, adică

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

Testul t unilateral dreapta înseamnă: dacă $t_c \leq t_\alpha$ atunci (H_0) se acceptă, şi dacă $t_c > t_\alpha$ atunci (H_0) se respinge.

3.1.4 Compararea mediilor a două populaţii statistice (σ_1, σ_2 necunoscuţi)

Fie $X \sim \mathcal{N}\left(m_1, \sigma_1^2\right)$ şi $Y \sim \mathcal{N}\left(m_2, \sigma_2^2\right)$ şi presupunem că $\sigma_1 = \sigma_2 = \sigma$ necunoscuţi.

Verificăm ipoteza

$$(\mathrm{H}_0) : m_1 = m_2.$$

&&&&&&&&&&&&&&&&

3.2 Ipoteze asupra dispersiilor

3.2.1 Compararea dispersiei unei populaţii statistice

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

Verificăm ipoteza nulă

$$(\mathrm{H}_0) : \sigma^2 = \sigma_0^2 .$$

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

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

Testul χ^2 bilateral Verificăm ipoteza nulă

$$(\mathrm{H}_0) : \sigma^2 = \sigma_0^2$$

faţă de altă ipoteză

$$(\mathrm{H}_1) : \sigma^2 = \sigma_1^2 \neq \sigma_0^2 .$$

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 repartiţiei $\chi^2(n-1, 1)$, valoarile critice χ_1^2 şi χ_2^2 astfel încât

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

Pentru modul de determinare al valorilor χ_1^2 şi χ_2^2 vezi Secţiunea *Interval de încredere pentru dispersia teoretică* (pag. 66).

Exerciţiul 3.16 *De făcut graficul*

Obţinem atunci că **intervalul $\left[\chi_1^2, \chi_2^2\right]$ este interval de acceptare iar regiunea $[0, \chi_1^2) \cup \left(\chi_2^2, \infty\right)$ este regiune critică**, adică

$$\mathcal{V} = \left\{ (x_1, \ldots, x_n) \in \mathbb{R}^n : \frac{(n-1)(S^*)^2}{\sigma_0^2} \in [0, \chi_1^2) \cup \left(\chi_2^2, \infty\right) \right\} .$$

Testul χ^2 bilateral înseamnă: dacă $\chi_c^2 \in \left[\chi_1^2, \chi_2^2\right]$ atunci (H_0) se acceptă, şi dacă $\chi_c^2 \notin \left[\chi_1^2, \chi_2^2\right]$ atunci (H_0) se respinge.

Testul χ^2 unilateral stânga Verificăm ipoteza nulă

$$(\mathrm{H}_0) : \sigma^2 = \sigma_0^2$$

faţă de altă ipoteză

$$(\mathrm{H}_1) : \sigma^2 = \sigma_1^2 \ < \sigma_0^2 \, .$$

Dacă ipoteza nulă este acceptată atunci

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

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

$$\delta = 1 - \alpha = \mathbb{P}\left(\chi_c^2 \geq \chi_1^2\right) \, .$$

Exerciţiul 3.17 *De făcut graficul*

Obţinem atunci că **intervalul $[\chi_1^2, \infty)$ este interval de acceptare iar intervalul $[0, \chi_1^2)$ este regiune critică**, adică

$$\mathcal{V} = \left\{ (x_1, \ldots, x_n) \in \mathbb{R}^n : \frac{(n-1)\left(S^*\right)^2}{\sigma_0^2} \in [0, \chi_1^2) \right\}.$$

Testul χ^2 unilateral stânga înseamnă: dacă $\chi_c^2 \geq \chi_1^2$ atunci (H_0) se acceptă, şi dacă $\chi_c^2 < \chi_1^2$ atunci (H_0) se respinge.

Testul χ^2 unilateral dreapta Verificăm ipoteza nulă

$$(\mathrm{H}_0) : \sigma^2 = \sigma_0^2$$

faţă de altă ipoteză

$$(\mathrm{H}_1) : \sigma^2 = \sigma_1^2 \ > \sigma_0^2 \, .$$

Dacă ipoteza nulă este acceptată atunci

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

Pentru α fixat se va determina, din tabelul repartiţiei $\chi^2 \left(n-1, 1\right)$, valoarea critică χ_2^2 astfel încât

$$\alpha = \mathbb{P}\left(\chi_c^2 > \chi_2^2\right)$$

sau echivalent

$$\delta = 1 - \alpha = \mathbb{P}\left(\chi_c^2 \leq \chi_2^2\right) \, .$$

Exerciţiul 3.18 *De făcut graficul*

Obţinem atunci că **intervalul** $\left[0, \chi_2^2\right]$ **este interval de acceptare iar intervalul** $\left(\chi_2^2, \infty\right)$ **este regiune critică**, adică

$$\mathcal{V} = \left\{ (x_1, \ldots, x_n) \in \mathbb{R}^n : \frac{(n-1)\left(S^*\right)^2}{\sigma_0^2} \in \left(\chi_2^2, \infty\right) \right\}.$$

Testul χ^2 **unilateral dreapta înseamnă: dacă** $\chi_c^2 \leq \chi_2^2$ **atunci** (H_0) **se acceptă, şi dacă** $\chi_c^2 > \chi_2^2$ **atunci** (H_0) **se respinge.**

Exerciţiul 3.19 *Ştim că dispersia reală este* $\sigma^2 = 2.25$*. În urma a* 30 *de observaţii se obţine abaterea empirică modificată* $S^* = 0.9$*. Să se cerceteze, aplicându-se* testul χ^2 unilateral dreapta*, dacă valorile obţinute diferă semnificativ sau nu de* σ^2 *dat, considerându-se un prag de semnificaţie de* 5%*.*

3.2.2 Compararea dispersiilor a două populaţii statistice

Fie $X \sim \mathcal{N}\left(m_1, \sigma_1^2\right)$ şi $Y \sim \mathcal{N}\left(m_2, \sigma_2^2\right)$. Verificăm ipoteza

$$(\mathrm{H}_0) : \sigma_1^2 = \sigma_2^2 .$$

Statistica utilizată va fi

$$F \overset{def}{=\!=} \frac{\frac{\left(S_1^*\right)^2}{\sigma_1^2}}{\frac{\left(S_2^*\right)^2}{\sigma_2^2}} = \frac{\sigma_2^2}{\sigma_1^2}\frac{\left(S_1^*\right)^2}{\left(S_2^*\right)^2} = \frac{\sigma_2^2}{\sigma_1^2}\frac{\frac{n_1}{n_1-1}S_1^2}{\frac{n_2}{n_2-1}S_2^2} = \frac{n_2-1}{n_1-1}\frac{\frac{n_1 S_1^2}{\sigma_1^2}}{\frac{n_2 S_2^2}{\sigma_2^2}} .$$

Ştim că $U_1 \overset{def}{=\!=} \frac{n_1 S_1^2}{\sigma_1^2}$ este distribuit $\chi^2\left(n_1 - 1, 1\right)$ iar $U_2 \overset{def}{=\!=} \frac{n_2 S_2^2}{\sigma_2^2} \sim \chi^2\left(n_2 - 1, 1\right)$. Atunci deducem că

$$F = \frac{\frac{U_1}{n_1-1}}{\frac{U_2}{n_2-1}} \sim F\left(n_1 - 1, n_2 - 1\right).$$

Vom presupune (fără a restrânge generalitatea) că $\left(S_1^*\right)^2 > \left(S_2^*\right)^2$.

Testul F **bilateral** Verificăm ipoteza

$$(\mathrm{H}_0) : \sigma_1^2 = \sigma_2^2$$

faţă de altă ipoteză

$$(\mathrm{H}_1) : \sigma_1^2 \neq \sigma_2^2 .$$

Dacă ipoteza (H_0) se acceptă, atunci

$$F_c = \frac{(S_1^*)^2}{(S_2^*)^2} \sim F\left(n_1 - 1, n_2 - 1\right).$$

Pentru α fixat se va determina, din tabelul repartiţiei $F\left(n_1 - 1, n_2 - 1\right)$, valoarile critice F_1 şi F_2 astfel încât

$$\delta = 1 - \alpha = \mathbb{P}\left(F_1 \leq F_c \leq F_2\right)$$

sau echivalent

$$\alpha = \mathbb{P}\left(F_c < F_1\right) + \mathbb{P}\left(F_c > F_2\right)$$

Vom determina valorile F_1 şi F_2 presupunând că cele două probabilităţi de mai sus sunt egale, adică

$$\mathbb{P}\left(F_c < F_1\right) = \alpha/2$$

şi

$$\mathbb{P}\left(F_c > F_2\right) = \alpha/2.$$

Deci, având în vedere ecuaţiile de mai sus, avem că (vezi desenul, precum şi relaţia (2.21))

$$F_2 = F_{\frac{\alpha}{2}, n_1 - 1, n_2 - 1} \text{ şi } F_1 = F_{1 - \frac{\alpha}{2}, n_1 - 1, n_2 - 1} = \frac{1}{F_{\frac{\alpha}{2}, n_2 - 1, n_1 - 1}}.$$

Exerciţiul 3.20 *De făcut graficul*

Obţinem atunci că **intervalul** $\left[F_{1-\alpha/2}, F_{\alpha/2}\right]$ **este interval de acceptare iar regiunea** $\left(0, F_{1-\alpha/2}\right) \cup \left(F_{\alpha/2}, \infty\right)$ **este regiune critică**.

Testul F **bilateral înseamnă: dacă** $F_c \in \left[F_{1-\alpha/2}, F_{\alpha/2}\right]$ **atunci** (H_0) **se acceptă, şi dacă** $F_c \notin \left[F_{1-\alpha/2}, F_{\alpha/2}\right]$ **atunci** (H_0) **se respinge.**

Exerciţiul 3.21 *În urma a* 31 *de observaţii se obţine abaterea empirică modificată* $S^* = 1.5$*, iar la* 25 *de observaţii se obţine 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 semnificaţie de* 5%*.*

Testul F unilateral stânga Verificăm ipoteza

$$(\mathrm{H}_0) : \sigma_1^2 = \sigma_2^2$$

faţă de altă ipoteză

$$(\mathrm{H}_1) : \sigma_1^2 < \sigma_2^2 .$$

Pentru α fixat se va determina, din tabelul repartiţiei $F(n_1 - 1, n_2 - 1)$, valoarea critică F_1 astfel încât

$$\delta = 1 - \alpha = \mathbb{P}(F_c \geq F_1)$$

sau echivalent

$$\alpha = \mathbb{P}(F_c < F_1) .$$

Avem că (vezi desenul, precum şi relaţia (2.21))

$$F_1 = F_{1-\alpha, n_1-1, n_2-1} = \frac{1}{F_{\alpha, n_2-1, n_1-1}}.$$

Exerciţiul 3.22 *De făcut graficul*

Obţinem atunci că **intervalul $[F_1, \infty)$ este interval de acceptare iar regiunea $(0, F_1)$ este regiune critică**.

Testul F unilateral stânga înseamnă: dacă $F_c \in [F_1, \infty)$ atunci (H_0) se acceptă, şi dacă $F_c \notin [F_1, \infty)$ atunci (H_0) se respinge.

Testul F unilateral dreapta Verificăm ipoteza

$$(\mathrm{H}_0) : \sigma_1^2 = \sigma_2^2$$

faţă de altă ipoteză

$$(\mathrm{H}_1) : \sigma_1^2 > \sigma_2^2 .$$

Pentru α fixat se va determina, din tabelul repartiţiei $F(n_1 - 1, n_2 - 1)$, valoarea critică F_2 astfel încât

$$\delta = 1 - \alpha = \mathbb{P}(F_c \leq F_1)$$

sau echivalent

$$\alpha = \mathbb{P}(F_c > F_2) .$$

Avem că (vezi desenul)

$$F_2 = F_{\alpha, n_1-1, n_2-1} .$$

Exerciţiul 3.23 *De făcut graficul*

Obţinem atunci că **intervalul** $[0, F_2)$ **este interval de acceptare iar regiunea** (F_2, ∞) **este regiune critică**.

Testul F **unilateral dreapta înseamnă: dacă** $F_c \in [0, F_2)$ **atunci** (H_0) **se acceptă, şi dacă** $F_c \notin [0, F_2)$ **atunci** (H_0) **se respinge.**

Bibliografie

[1] George Ciucu, Virgil Craiu, Ion Săcuiu, *Probleme de statistică matematică*, Editura Tehnică, Bucureşti, 1974.

[2] George Ciucu, Gabriel Sîmboan, *Teoria probabilităţilor şi statistică matematică. Culegere de probleme*, Editura Tehnică, Bucureşti, 1962.

[3] George Ciucu, Constantin Tudor, *Probabilităţi şi procese stochastice*, vol. I, Editura Academiei, Bucureşti, 1978.

[4] Jay L. Devore, *Probability and Statistics for Engineering and the Sciences* (Ninth Edition), Cengage Learning, Boston, 2016.

[5] Jay L. Devore, Kenneth N. Berk, *Modern Mathematical Statistics with Applications* (Second Edition), series: Springer Texts in Statistics, Springer New York, 2012.

[6] Lucian Maticiuc, *Teoria probabilităţilor*, Universitatea "Alexandru Ioan Cuza", Iaşi, http://www.math.uaic.ro/ maticiuc/didactic/Probability Theory.pdf, 2017.

[7] Gheorghe Mihoc, Nicolae Micu, *Teoria probabilităţilor şi statistică matematică*, Editura Didactică şi Pedagogică, Bucureşti, 1980.

[8] Elena Nenciu, *Lecţii de statistică matematică*, Editura Universităţii "Alexandru Ioan Cuza", Iaşi, 1972.

[9] Elena Nenciu, *Teoria probabilităţilor şi statistică matematică*, Editura Universităţii "Alexandru Ioan Cuza", Iaşi, 1986.

[10] Sheldon Ross, *Introductory Statistics* (Third Edition), Elsevier, Oxford, 2010.

[11] Iulian Stoleriu, *Statistica prin Matlab*, Editura Matrix ROM, Bucureşti, 2010.

[12] Pavel Talpalaru, Liliana Popa, Emilia Popovici, *Probleme de teoria probabilităţilor şi statistică matematică*, Editura Universităţii Tehnice „Gheorghe Asachi", Iaşi, 1995.