Lucian MATICIUC

Introducere în Statistică Matematică

Universitatea „Alexandru Ioan Cuza”

Iaşi – 2017

Cuprins

Capitolul 1

Elemente de statistică. Variabile de selecţie (empirice)

1.1 Fundamentele teoriei selecţiei

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 **indivizi** sau **unităţi statistice**. Numărul de indivizi care alcătuiesc populaţia statistică determină **volumul populaţiei**.

Caracteristica este o anumită proprietate urmărită la indivizii unei colectivităţi statistice. Există caracteristici cantitative (cele care se pot măsura, ca vârsta, greutatea, etc.) şi caracteristici calitative.

Se numeşte **selecţie** (**eşantion, sondaj**) o colectivitate parţială de elemente extrase la întâmplare. Notăm aceste elemente ale colectivităţii parţiale cu x_i, $i = \overline{1, N}$, iar N este volumul selecţiei (numărul indivizilor din selecţie).

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).

Se numeşte **serie statistică**, asociată unei selecţii de volum N, un tablou de forma

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

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).

Numim **variabilă empirică** (**de selecţie**), notată pe scurt v.e., asociată unei selecţii de volum N, un tablou de forma

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

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

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

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, iar 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

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

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

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

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

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

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

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

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

$$F_n(x) = \begin{cases} 0 & , \; x < x_1, \\ \sum\limits_{j=1}^{i-1} f_j & , \; x_{i-1} \le x < x_i, \; i = \overline{2,k}\,, \\ 1 & , \; x_k \le x. \end{cases} = \begin{cases} 0 & , \; x < x_1, \\ f_1 & , \; x_1 \le x < x_2, \\ f_1 + f_2 & , \; x_2 \le x < x_3, \\ \ldots & \\ f_1 + f_2 + \cdots + f_{k-1} & , \; x_{k-1} \le x < x_k, \\ 1 & , \; x_k \le x. \end{cases}$$

Dacă X^* este dată de (1.2) atunci

$$F_n(x) = \begin{cases} 0 & , \; x < x_1, \\ \frac{i-1}{N} & , \; x_{i-1} \le x < x_i, \; i = \overline{2,N}\,, \\ 1 & , \; x_N \le x. \end{cases}$$

Dacă X^* este dată de (1.3) atunci

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

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 trei moduri:

Reprezentarea în batoane. În planul xOy se trec pe axa absciselor valorile x_i iar în dreptul fiecărei valori x_i se ridică câte o perpendiculară de lungime egală cu valoare f_i (sau cu n_i) corespunzătoare lui x_i. Menţionăm că dacă unim vârfurile acestor perpendiculare prin segmente vom obţine poligonul frecvenţelor cumulate.

Histograma. Această reprezentare se foloseşte la o v.e. ale cărei valori sunt grupate sunt grupate pe clase. Pe axa absciselor se iau segmente egale cu amplitudinea claselor (intervalelor). Pe fiecare segment de acest fel, considerat ca bază, se ridică câte un dreptunghi a cărui înălţime este egală cu frecvenţa corespunzătoare (relativă sau absolută) acelei clase. Menţionăm că dacă unim mijloacele laturilor superioare ale acestor dreptunghiuri vom obţine poligonul frecvenţelor cumulate.

Poligonul frecvenţelor cumulate. Se măsoară pe Ox valorile x_i şi pe Oy frecvenţelor cumulate corespunzătoare valorilor x_i. Unind aceste puncte rezultate vom obţine poligonul frecvenţelor cumulate.

1.3 Valori caracteristice ale unei variabile empirice şi ale unei variabile aleatoare

Valorile caracteristice sunt nişte date numerice care reprezintă fidel o v.a. sau o caracteristică avută în vedere. Cunoaşterea valorilor caracteristice este utilă în practică la compararea a două v.a. sau a două populaţii statistice pe care este definită aceeaşi caracteristică, precum şi la deducerea legii teoretice urmată de o v.a. sau de caracteristica considerată.

1.3.1 Caracteristici ale unei variabile empirice

Să considerăm o caracteristică cantitativă reprezentată de seria statistică $\begin{pmatrix} x_1 & x_2 & \cdots & x_k \\ n_1 & n_2 & \cdots & n_k \end{pmatrix}$ cu $\sum\limits_{i=1}^{k} n_i = N$ (volumul selecţiei). Acestei serii îi asociem variabila empirică (de selecţie)

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

unde $f_i = \frac{n_i}{N}$ (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)

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

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

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

Momentul empiric de ordin r este valoarea

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

dacă X^* este dată de (1.4), respectiv

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

dacă X^* este dată de (1.5).

În particular, pentru $r = 1$, obţinem

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

respectiv

$$\bar{x} \stackrel{def}{=} m'_1 = \frac{1}{N}\sum_{i=1}^{k} n_i c_i = \sum_{i=1}^{k} f_i c_i\,.$$

Valoarea $\bar{x}$ se va numi **media de selecţie** sau media aritmetică.

Mediana (notată cu m_e) este valoarea caracteristicii X^* care împarte volumul selecţiei în două părţi egale. Dacă N este par, $N = 2k$, atunci $m_e = x_k$ sau $m_e = x_{k+1}$ sau $m_e = \frac{x_k + x_{k+1}}{2}$. Dacă N este impar, $N = 2k + 1$, atunci $m_e = x_{k+1}$.

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

Parametrii variabilităţii (ale împrăştierii): R, S^2, S, S^*

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

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

Momentul central empiric de ordin r este

$$\mu'_r = \frac{1}{N}\sum_{i=1}^{k} n_i \left(x_i - \bar{x}\right)^r = \sum_{i=1}^{k} f_i \left(x_i - \bar{x}\right)^r\,.$$

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

$$S^2 \overset{def}{=} \mu_2' = \frac{1}{N}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2 = \sum_{i=1}^{k} f_i (x_i - \bar{x})^2 \; . \tag{1.6}$$

Propoziţia 1.1 *Are loc următoarea relaţie (care reprezintă o formulă utilă de calcul) a dispersiei:*

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

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

Demonstraţie. Într-adevăr,

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

■

Mărimea

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

se numeşte **abaterea medie pătratică empirică**.
Dispersia empirică modificată este numărul

$$(S^*)^2 = \frac{N}{N-1} S^2, \tag{1.7}$$

unde cu N este notat 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 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 v.a. continue).

Boltirea se măsoară prin coeficientul de exces

$$E_X = \alpha_4 - 3, \text{ unde } \alpha_4 = \frac{\mu_4'}{S^4}.$$

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 v.a. repartizată normal $X \sim N\left(m, \sigma^2\right)$ are $E_X = 0$. Într-adevăr, $\mu_4' \overset{def}{=} \mathbb{E}\left[X - \mathbb{E}\left(X\right)\right]^4 = (f\breve{a}r\breve{a}\ demonstratie) = 3\sigma^4$ iar dispersia este $S^2 = \sigma^2$, deci $E_X = \frac{\mu_4'}{S^4} - 3 = \frac{3\sigma^4}{(\sigma^2)^2} - 3 = 0$ (am notat prin $\mathbb{E}\left(X\right)$ media v.a. X).

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

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

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

Asimetria se măsoară prin coeficientul de asimetrie

$$\bar{\gamma}_2 = \frac{\mu_3'}{S^3},$$

adică $\bar{\gamma}_2 = \alpha_3$. Menţionăm că S este intotdeauna pozitiv (fiind radical din dispersie) iar μ_3' poate fi pozitiv sau negativ după cum abaterile $x_i - \bar{x}$ care predomină sunt pozitive, respectiv negative. Repartiţia statistică normală are $\bar{\gamma}_2 = 0 \Leftrightarrow \alpha_3 = 0$.

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 $X \sim N\left(m, \sigma^2\right)$ are drept axă de simetrie dreapta $x = \mathbb{E}\left(X\right) = m$ (media v.a. normale X).

Exerciţiul 1.2 *Să presupunem că un aparat de măsurare este utilizat pentru a citi o distanţă de 12 de ori. 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 este valoarea* 0.20.

Frecvenţele cumulate sunt date în tabelul.....

Media de selecţie sau media aritmetică este dată de

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

sau echivalent

$$\bar{x} = \frac{\sum\limits_{i=1}^{6} n_i x_i}{20}$$

Dispersia empirică (sau varianţa) este dată de formula

$$\begin{aligned} S^2 \overset{def}{=} \frac{1}{12}\sum_{i=1}^{6} n_i \left(x_i - \bar{x}\right)^2 \quad &= \quad \sum_{i=1}^{6} f_i \left(x_i - \bar{x}\right)^2 = \cdots = 0.00643\,. \\ &\Rightarrow \quad S = \sqrt{S^2} \simeq 0.0802. \end{aligned}$$

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.

Trebuie scrisă şi funcţia empirică de repartiţie $F(x) = \ldots.$*(este o funcţie în scară).*

Exerciţiul 1.3 *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*

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

Momentul empiric de ordin 1 *(sau media de selecţie 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

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

unde x_i *sunt valorile citite în tabloul (1.11)*

sau ***echivalent*** $\bar{x}$ ***este media aritmetică a tuturor valorilor citite (valori ce se pot repeta)****,* $\bar{x} = \frac{\sum\limits_{i=1}^{20} x_i}{20}$*, unde* x_i *sunt valorile citite în tabelul (1.9).*

Dispersia empirică (sau varianţa) 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 $x_i - \bar{x}$ *şi* $\left(x_i - \bar{x}\right)^2$ *:*

x_i	*Frecvenţa abs.* n_i	*Frecvenţa rel.* f_i	$x_i - \bar{x}$	$\left(x_i - \bar{x}\right)^2$
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	$0.205 = 1/20$	-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*

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

În plus ***abaterea empirică modificată*** *este*

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

Remarca 1.4 *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

$$\begin{aligned}(S^*)^2 &= \frac{\sum\limits_{i=1}^{N}(x_i-\bar{x})^2}{N-1} = \frac{\sum\limits_{i=1}^{N}x_i^2 - 2\sum\limits_{i=1}^{N}x_i\bar{x} + \sum\limits_{i=1}^{N}\bar{x}^2}{N-1} = \frac{\sum\limits_{i=1}^{N}x_i^2 - 2\bar{x}\sum\limits_{i=1}^{N}x_i + N\bar{x}^2}{N-1} \\ &= \frac{\sum\limits_{i=1}^{N}x_i^2 - 2N\bar{x}\cdot\bar{x} + N\bar{x}^2}{N-1} = \frac{\sum\limits_{i=1}^{N}x_i^2 - N\bar{x}^2}{N-1} = \frac{\sum\limits_{i=1}^{20}x_i^2 - 20\bar{x}^2}{N-1},\end{aligned}$$

adică obţinem următoarea ***formulă de calcul a dispersiei empirice modificate*** *(vezi şi formula (1.8)):*

$$(S^*)^2 = \frac{\sum\limits_{i=1}^{N}x_i^2 - N\bar{x}^2}{N-1}\,. \tag{1.12}$$

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 1.5 *(**Justificarea definirii dispersiei empirice modificate** (vezi (1.8) şi (1.12) precum şi Propoziţia 2.11)*.

Să notăm cu μ valoarea exacta a cantităţii măsurate (o valoare teoretică ce nu poate fi determinată de fapt) iar σ^2 dispersia întregii populaţii, adică $\sigma^2 = \frac{\sum\limits_{i=1}^{n}(x_i-\mu)^2}{n}$.

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

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

Pe de o parte avem că termenul $\sum\limits_{i=1}^{n}(x_i-\mu)^2$ va fi, pentru n foarte mare, egal cu $n\sigma^2$. Pe de altă parte avem că termenul $n(\bar{x}-\mu)^2$ va fi, pentru n foarte mare, egal cu $n\frac{\sigma^2}{n} = \sigma^2$, deoarece $(\bar{x}-\mu)^2$ aproximează dispersia v.a. $\bar{X} \overset{def}{=} \frac{\sum\limits_{i=1}^{n}X_i}{n}$ care este dată de

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

Deci, pentru n foarte mare,

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

adică **dispersia empirică modificată aproximează mai bine dispersia σ^2 a populaţiei (decât dispersia empirică S^2 dată de definiţia (1.6)).**

Exerciţiul 1.6 *Presupunem că au fost obţinute următoarele valori în urma a* 15 *citiri a unei distanţe:*

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

Calculaţi media de selecţie, dispersia empirică modificată, abaterea empirică modificată, $E_{0.5}$ *şi* $E_{0.95}$.
Ce procent din cele 15 *observaţii sunt în intervalul* $(\bar{x}-S^*, \bar{x}+S^*)$*?*
Având în vedere că, pentru $\alpha \in (0,1)$,

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

găsiţi care este procentul de valori aflat în intervalul $(\bar{x}-E'_{0.5}, \bar{x}+E'_{0.5})$ *(având în vedere că media* $\bar{X}$ *este repartizată* $\mathcal{N}\left(0,\sigma^2\right)$*, menţionăm că în acest caz* $E'_\alpha = E_\alpha\, S^*$*.). Interpretaţi rezultatul.*
Găsiţi care este procentul de valori aflat în intervalul $(\bar{x}-E'_{0.95}, \bar{x}+E'_{0.95})$*. Interpretaţi rezultatul.*

1.3.2 Caracteristici ale unei variabile aleatoare discrete: medii şi momente

Fie X o 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

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

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

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

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

Propoziţia 1.7 (Proprietăţi ale mediei)

1. *Valoarea medie a unei constante este egală cu constanta:*

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

2.
$$\mathbb{E}\left(a + X\right) = a + \mathbb{E}\left(X\right).$$

3.
$$\mathbb{E}\left(aX\right) = a\mathbb{E}\left(X\right).$$

4.
$$\mathbb{E}\left(X + Y\right) = \mathbb{E}\left(X\right) + \mathbb{E}\left(Y\right).$$

5. *Media produsului a două v.a. independente este produsul mediilor variabilelor considerate*

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

Demonstraţie. 5. Avem că

$$\begin{aligned}
\mathbb{E}\left(XY\right) &= \sum_{i=\overline{1,n},\, j=\overline{1,m}} x_i y_j p_{ij} = \sum_{i,j} x_i y_j \mathbb{P}\left(X = x_i, Y = y_j\right) \\
&= \sum_{i,j} x_i y_j \mathbb{P}\left(\{X = x_i\} \cap \{Y = y_j\}\right) = (\text{deoarece } X, Y \text{ independente}) \\
&= \sum_{i,j} x_i y_j \mathbb{P}\left(X = x_i\right)\mathbb{P}\left(Y = y_j\right) = \sum_{i,j} x_i y_j p_i q_j = \sum_{i} x_i p_i \sum_{,j} y_j q_j \\
&= \mathbb{E}\left(X\right)\mathbb{E}\left(Y\right).
\end{aligned}$$

■

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

$$\nu_r \overset{def}{=} \mathbb{E}\left(X^r\right) = \sum_i p_i x_i^r \,.$$

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

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

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

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

În particular, obţinem

(i) $m_1 = \nu_1 = \mathbb{E}(X)$.

(ii) $m_2 = \sqrt{\nu_2} = \sqrt{\sum\limits_i p_i x_i^2}$ =valoarea medie pătratică a lui X.

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

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

Remarca 1.8 *Calculăm valoarea* $\ln m_0$:

$$\begin{aligned}
\ln(m_0) &= \lim_{r\to 0}\frac{\ln\left(p_1 x_1^r + p_2 x_2^r + \cdots + p_n x_n^r\right)}{r} = \textit{nedeterminarea } \frac{0}{0}\textit{, aplic L'Hospital}\\
&= \lim_{r\to 0}\frac{\left[\ln\left(p_1 x_1^r + p_2 x_2^r + \cdots + p_n x_n^r\right)\right]'}{1} = \lim_{r\to 0}\frac{\left[p_1 x_1^r + p_2 x_2^r + \cdots + p_n x_n^r\right]'}{p_1 x_1^r + p_2 x_2^r + \cdots + p_n x_n^r}\\
&= \lim_{r\to 0}\frac{p_1 x_1^r \ln x_1 + p_2 x_2^r \ln x_2 + \cdots + p_n x_n^r \ln x_n}{p_1 x_1^r + p_2 x_2^r + \cdots + p_n x_n^r}\\
&= \frac{p_1 \ln x_1 + p_2 \ln x_2 + \cdots + p_n \ln x_n}{p_1 + p_2 + \cdots + p_n} = \frac{\ln\left[(x_1)^{p_1}(x_2)^{p_2}\cdots(x_n)^{p_n}\right]}{1}\\
&= \ln\left[(x_1)^{p_1}(x_2)^{p_2}\cdots(x_n)^{p_n}\right],
\end{aligned}$$

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

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

$$\mu_r \overset{def}{=} \mathbb{E}\left[(X-\nu_1)^r\right] = \sum_i p_i \left(x_i - \nu_1\right)^r .$$

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

$$\sigma^2 = D^2(X) \overset{def}{=} \mu_2 = \mathbb{E}\left(X-\nu_1\right)^2, \text{ unde } \nu_1 = \mathbb{E}(X).$$

Mediana v.a. X este

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

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

1.3.3 Caracteristici ale unei variabile aleatoare continue: medii şi momente

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 1.7).

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

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

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

În particular,

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

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

$$\mu_r \stackrel{def}{=} \mathbb{E}\left[(X-\nu_1)^r\right] = \int_{-\infty}^{\infty} (x-\nu_1)^r f(x)\, dx\,.$$

În particular,

$$\mu_2 = \mathbb{E}\left[(X-\nu_1)^2\right] = \int_{-\infty}^{\infty} (x-\nu_1)^2 f(x)\, dx\,.$$

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

$$\sigma^2 = D^2(X) \stackrel{def}{=} \mu_2 = \mathbb{E}(X-\nu_1)^2\,, \text{ unde } \nu_1 = \mathbb{E}(X)\,.$$

Remarca 1.9 *Având în vedere calculul*

$$\begin{aligned}
D^2(X) &= \mathbb{E}(X-\nu_1)^2 = \int_{\mathbb{R}} (x-\nu_1)^2 f(x)\,dx = \int_{\mathbb{R}} \left(x^2 - 2\nu_1 x + \nu_1^2\right) f(x)\,dx \\
&= \int_{\mathbb{R}} x^2 f(x)\,dx - 2\int_{\mathbb{R}} \nu_1 x f(x)\,dx + \int_{\mathbb{R}} \nu_1^2 f(x)\,dx \\
&= \int_{\mathbb{R}} x^2 f(x)\,dx - 2\nu_1 \int_{\mathbb{R}} x f(x)\,dx + \nu_1^2 \int_{\mathbb{R}} f(x)\,dx \\
&= \mathbb{E}\left(X^2\right) - 2\nu_1 \mathbb{E}(X) + \nu_1^2 = \mathbb{E}\left(X^2\right) - \nu_1^2,
\end{aligned}$$

obţinem o formulă foarte utilă în calcule:

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

Propoziţia 1.10 (Proprietăţi ale dispersiei)
1. *Dispersia unei constante este nulă:*

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

2.

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

deoarece

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

3.*Dispersia sumei a două v.a. independente este suma dispersiei variabilelor considerate*

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

Într-adevăr,

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

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 1.11 (Proprietăţi ale abaterii medii pătratice)
1. $D(c) = 0.$
2. $D(aX) = |a| D(X).$

Teorema 1.12 (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*

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

(fără demonstraţie).

O formă des întâlnită în aplicaţii este următoarea consecinţă:

Corolarul 1.13 *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*

$$\mathbb{P}(|X - m| < k\sigma) \geq 1 - \frac{1}{k^2}, \forall k > 0.$$

Capitolul 2

Elemente de teoria selecţiei şi a estimaţiei

Presupunem că o populaţie discretă conţine 100 de valori. Fie astfel tabelul

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 m) 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 m 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 m şi σ^2 (cu cât volumul se apropie mai mult de 100, cu atât $\bar{x}$ şi S^2 se apropie mai mult de $m = 26.1$ 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 v.a. (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 v.a. 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	$\bar{x} = 26.9$	$S^2 = 28.1$
Set 2:	26.9	28.1	29.2	26.2	30.0	27.1	26.5	30.6	28.5	25.5	$\bar{x} = 27.9$	$S^2 = 2.9$
Set 3:	32.2	22.2									$\bar{x} = 26.9$	$S^2 = 10.9$
Set 4:	24.2	36.0									$\bar{x} = 27.2$	$S^2 = 23.0$

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 (2.1)), adică funcţii care să depindă de fiecare selecţie în parte. În cadrul teoriei estimaţiei foarte importante vor fi trei distribuţii/repartiţii: distribuţia "hi pătrat" de parametrii n şi σ (notată $\chi^2(n, \sigma)$), distribuţia Student de parametru n (notată $t(n)$), şi distribuţia Fisher-Snedecor de parametrii a şi b (notată $F(a, b)$). Pentru acestea (definiţii şi legături) 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 caracteristici ale lui X (media şi dispersia, de exemplu). 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 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.

Tipuri de variabile pe o populaţie $\mathcal{P}$: v.a. teoretică X, necercetată direct, care se referă la $\mathcal{P}$ în totalitate; v.a. empirică (de selecţie) X^*, ce ia valorile $x_1, \ldots, x_n$, adică

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

Valorile $x_1, \ldots, x_n$ se pot considera ca valori pe care le iau n v.a. independente $X_1, \ldots, X_n$, ce au aceeaşi repartiţie şi anume repartiţia lui X.

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

Remarca 2.1 *Fiecare carateristică empirică obţinută pe baza selecţiei este valoarea unei anumite v.a. teoretice, valoare care variază odată cu selecţia.*
De exemplu media de selecţie $\bar{x} \overset{def}{=} \frac{x_1+\cdots+x_n}{n}$ *reprezintă valoarea v.a.* $\bar{X} \overset{def}{=} \frac{X_1+\cdots+X_n}{n}$, *şi de aceea putem vorbi de* $\bar{x}$ *ca de o v.a.*
De asemenea dispersia empirică

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

este valoarea v.a. $\frac{1}{n} \sum\limits_{i=1}^{n} \left(X_i - \bar{X}\right)^2$.

Problema esenţială este următoarea: în ce măsură anumite caracteristici ale lui X^* (media, dispersia, momentele, etc.) pot reprezenta caracteristici 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).

2.1 Problema estimaţiei

Să presupunem în continuare că legea de repartiţie teoretică este cunoscută şi fie λ un parametru legat de ea. Presupunem că $f(x, \lambda)$ este legea de repartiţie teoretică.

Definiţia 2.2 *A estima parametrul λ înseamnă a determina o **funcţie de selecţie***

$$\lambda_n^* = \varphi(x_1, \ldots, x_n) \tag{2.1}$$

astfel încât

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

în diferite sensuri.

Definiţia 2.3 *λ_n^* se numeşte **estimator consistent** pentru λ dacă*

$$\lim_{n\to\infty} \mathbb{P}(|\lambda_n^* - \lambda| \le \varepsilon) = 1, \ \forall \varepsilon > 0.$$

Definiţia 2.4 *λ_n^* se numeşte **estimator corect** pentru λ dacă*

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

Definiţia 2.5 *Diferenţa $\delta = \mathbb{E}(\lambda_n^*) - \lambda$ se numeşte distorsiunea (deplasarea) estimatorului λ_n^*.*

Definiţia 2.6 *Dacă $\delta \neq 0$ spunem că estimatorul este deplasat. Estimatorul corect este un estimator deplasat.*

Definiţia 2.7 *λ_n^* se numeşte **estimator absolut corect** pentru λ dacă*

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

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

Definiţia 2.8 *λ_n^* se numeşte **estimator eficient** pentru λ dacă $\delta = 0$ şi $D^2(\lambda_n^*)$ este minim.*

Teorema 2.9 (Rao-Cramer) *Numim expresia*

$$E(\lambda_n^*) \overset{def}{=} \frac{1}{D^2(\lambda_n^*)\; n\, \mathbb{E}\left[[\ln(f(x,\lambda))]_\lambda'\right]^2}$$

eficienţa estimatorului λ_n^*. Dacă eficienţa estimatorului λ_n^* este $E(\lambda_n^*) = 1$ atunci estimatorul nedeplasat λ_n^* este eficient (adică $D^2(\lambda_n^*)$ este minim).

Remarca 2.10 *Din inegalitatea lui Cebâşev se deduce imediat că un estimator absolut corect este un estimator consistent.*

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

2.1.1 Estimări punctuale ale valorilor caracteristice teoretice

Se demonstrează că au loc următoarele estimări:

Propoziţia 2.11
1) *Media de selecţie $\bar{X}$ este un estimator absolut corect pentru media teoretică $\mathbb{E}(X)$.*
2) *Dispersia empirică modificată $(S^*)^2$ este un estimator absolut corect al dispersiei teoretice $\sigma^2 = D^2(X)$.*
3) *Momentul empiric de ordin r (notat m_r') este un estimator absolut corect pentru momentul iniţial de ordin r al colectivităţii $\mathcal{P}$.*
4) *Pentru un x fixat, funcţia empirică de repartiţie $F_n(x)$ este un estimator absolut corect pentru funcţia de repartiţie $F(x)$.*

Demonstraţie. 1) Fie $\bar{X} \overset{def}{=} \frac{X_1+\cdots+X_n}{n}$. Dar X_i sunt independente şi au aceeaşi distribuţie cu X, deci

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

Atunci

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

2) Fie

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

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

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

Se poate arăta că

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

Pe de altă parte, utilizând relaţia

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

obţinem că

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

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

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

Astfel

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

adică S^2 nu este un estimator absolut corect al dispersiei teoretice $D^2\left(X\right)$, pe când $\left(S^*\right)^2$ este un estimator absolut corect al dispersiei teoretice $D^2\left(X\right)$.

■

2.1.2 Estimări punctuale ale parametrilor unei repartiţii 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 m şi dispersia σ^2 din repartiţia normală). Pentru estimarea acestor parametri avem la dispoziţie două metode.

Metoda verosimilităţii maxime Fie $f(x, \lambda_1, \ldots, \lambda_k)$ funcţia de frecvenţă (în cazul v.a. discrete) sau densitatea de probabilitate (în cazul v.a. continue) şi $x_1, \ldots, x_n$ observaţiile în urma selecţiei de volum n.

Definiţia 2.12 *Funcţia*

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

*unde x_i sunt variabilele de selecţie, se numeşte **funcţia de verosimilitate**.*

Definiţia 2.13 *Valorile $\lambda_i^* = \varphi_i(x_1, \ldots, x_n)$, $i = \overline{1,k}$ asociate lui λ_i pentru care V este maximă se numesc estimatori de maximă verosimilitate.*

Remarca 2.14 *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:

1. calculăm $V(\lambda_1, \ldots, \lambda_k)$ şi $\ln V(\lambda_1, \ldots, \lambda_k)$ cu ajutorul variabilelor de selecţie

2. rezolvăm sistemul de k ecuaţii şi k necunoscute

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

3. pentru soluţia $(\lambda_1^0, \ldots, \lambda_k^0)$ găsită mai sus verificăm condiţia suficientă de extrem

$$d^2 \left[\ln V\left(\lambda_1^0, \ldots, \lambda_k^0\right)\right] < 0.$$

Remarca 2.15 *Când avem mai multe puncte de maxim se ia cel mai mare dintre maxime.*

Remarca 2.16 *Se poate demonstra că, în condiţii destul de generale, estimatorii λ_i^* sunt consistenţi pentru λ_i, $i = \overline{1,k}$.*

Exerciţiul 2.17 *Să se estimeze, folosind metoda verosimilităţii maxime, parametrul* λ *al repartiţiei (exponenţiale) cu densitatea* $f(x) = \lambda e^{-\lambda x}$, $\lambda > 0, x > 0$. *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.*

Exerciţiul 2.18 *Acelaşi enunţ pentru:* $f(x) = (1+\lambda)\, x^{\lambda}$, $\lambda > 0, x \in (0,1)$. *Rezultatele obţinute în urma efectuării unei selecţii de volum* 4 *sunt* $x_1 = 0.4$, $x_2 = 0.6$, $x_3 = 0.85$, $x_4 = 0.9$.

Exerciţiul 2.19 *Să se estimeze, folosind metoda verosimilităţii maxime, 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 şi eficient.*

Metoda momentelor Fie $f(x, \lambda_1, \ldots, \lambda_k)$ densitatea de probabilitate (sau funcţia de frecvenţă în cazul unei v.a. discrete) pentru caracteristica X relativă la întreaga populaţie şi $\lambda_1, \ldots, \lambda_k$ parametrii necunoscuţi. Presupunem că există şi sunt finite momentele de ordin $1, 2, \ldots, k$ (notate m_r). Dacă avem o selecţie $x_1, \ldots, x_n$ din populaţie, din faptul ca momentele empirice de ordin r (notate m_r') sunt estimatori absoluţi corecţi pentru momentele teoretice de ordin r, putem scrie că

$$m_r(\lambda_1, \ldots, \lambda_k) \simeq m_r', r = \overline{1, k}.$$

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

$$m_r(\lambda_1, \ldots, \lambda_k) = m_r',\ r = \overline{1, k}.$$

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 2.20 *Şi metoda celor mai mici pătrate este tot o metodă pentru estimarea punctuală a parametrilor unei repartiţii.*

Exerciţiul 2.21 *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 şi eficient.*

Exerciţiul 2.22 *Să se estimeze, folosind metoda momentelor, parametrii v.a. uniform distribuite pe intervalul* $[a, b]$, *folosind datele* $x_1 = 3.1$, $x_2 = 0.2$, $x_3 = 1.6$, $x_4 = 5.2$, $x_5 = 2.1$.

Exerciţiul 2.23 *Să se estimeze, folosind metoda momentelor, parametrul m al v.a. X distribuite normal $\mathcal{N}\left(m,\sigma^2\right)$.*

2.1.3 Estimări prin intervale de încredere ale valorilor caracteristice teoretice

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

Fie λ valoarea teoretică ce dorim să o estimăm prin această metodă şi $x_1,\ldots,x_n$ observaţiile în urma unei selecţii de volum n.

Definiţia 2.24 *Intervalul de încredere (λ_1,λ_2) unde $\lambda_i=\varphi_i(x_1,\ldots,x_n)$, $i=\overline{1,2}$, cu proprietatea*

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

se numeşte interval de încredere pentru λ.

Numărul δ se va numi nivel de încredere sau siguranţa estimaţiei. De obicei $\delta=0.95$, 0.98, 0.99.

Numărul $\alpha=1-\delta$ se numeşte prag de semnificaţie.

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

Variabila

$$Z\overset{def}{=}\frac{\bar{X}-\mathbb{E}\left(\bar{X}\right)}{D\left(\bar{X}\right)}=\frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}}=\frac{\sqrt{n}}{\sigma}\left(\bar{X}-m\right)\tag{2.2}$$

va fi atunci repartizată normal standard $Z\sim\mathcal{N}(0,1)$.

Ştim că $\bar{X}\sim\mathcal{N}\left(m,\sigma^2/n\right)$.

Pentru a determina intervalul de încredere pentru m punem condiţia

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

cu δ cunoscut, adică vom obţine că

$$2\Phi\left(z_{\alpha/2}\right)=\delta.$$

Din tabele vom obţine valoarea $z_{\alpha/2}$. Deci

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

adică **intervalul de încredere pentru medie este dat de**

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

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

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

Remarca 2.25 *Fie a valoarea exactă a unei mărimi. În acest caz $m = \mathbb{E}\,(a) = a$ 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

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

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

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

obţinem

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

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 2.26 *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* $\mathbb{P}\left(|X| < E_\delta\right) = \delta$*, unde* $X \sim \mathcal{N}\left(0, \sigma^2\right)$*).*

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*).*

Remarca 2.27 *Cazul când volumul selecţiei* $n > 30$*,* X *repartiţie oarecare şi* σ *cunoscut.*

Conform teoremei lui Liapunov avem că

$$Z = \frac{\bar{X} - \mathbb{E}\left(\bar{X}\right)}{D\left(\bar{X}\right)} = \frac{\sqrt{n}}{\sigma}\left(\bar{X} - m\right) \sim \mathcal{N}\left(0, 1\right).$$

În acest caz intervalul de încredere este dat de relaţia (2.4).

Remarca 2.28 *Cazul când volumul selecţiei* $n > 30$*,* X *repartiţie oarecare şi* σ *necunoscut.*

Avem că

$$\sigma^2 \simeq \left(S^*\right)^2$$

unde

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

Relaţie (2.4) devine

(2.5)

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

Exerciţiul 2.29 *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* $\alpha = 0.01$*.*

Remarca 2.30 *Cazul când volumul selecţiei $n \leq 30$, X este repartizată $\mathcal{N}\left(m,\sigma^2\right)$ şi σ necunoscut.*

Reamintim acum legătura dintre distribuţia normală şi distribuţia χ^2 precum şi legătura dintre distribuţia normală, distribuţia χ^2 şi distribuţia Student:

Propoziţia 2.31 *Dacă A_i sunt v.a. normale de tip $\mathcal{N}\left(0,\sigma^2\right)$ atunci*

$$\sum\nolimits_{i=1}^{n} A_i^2 \sim \chi^2\left(n,\sigma\right) \tag{2.6}$$

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

Propoziţia 2.32 *Dacă $X \sim \mathcal{N}\left(0,\sigma^2\right)$, $Y \sim \chi^2\left(a,\sigma\right)$, atunci distribuţia*

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

adică T este distribuită Student de parametru a.

Pe de altă parte, aplicând şi Propoziţia 2.31, deducem că

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

iar

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

Obţinem că

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

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) şi (2.7),

(2.8)
$$nS^2 = \sum_{i=1}^{n} (X_i - m)^2 - n\left(\bar{X} - m\right)^2 \sim \chi^2(n,\sigma) - \chi^2(1,\sigma) = \chi^2(n-1,\sigma).$$

Pe de altă parte

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

Vom obţinem astfel că variabila

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

este distribuită Student de parametru $n-1$*, notat* $t(n-1)$.

Acum avem din (2.3) şi (2.9)

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

deci

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

***iar acum (spre deosebire de cazurile precedente în care** $Y \sim N(0,1)$**) valoarea** t **se va citi din tabelul distribuţiei Student de parametru** $n-1$ (având în vedere că acum $Y \sim t(n-1)$).*

Acum, similar, 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 2.33 *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 2.34 *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 $\mathbb{E}(X) = m$ şi $D^2(X) = \sigma^2$. Considerăm statistica, sau funcţia de estimaţie,

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

Din proprietatea (2.8) deducem că $nS^2 \sim \chi^2(n-1,\sigma)$, deci Y va fi repartizată atunci

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

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.

Cantitatea $\alpha \stackrel{def}{=} 1-\delta$ se numeşte prag de semnificaţie.

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

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

iar cantitatea χ_2^2 se va determina din relaţia

$$\mathbb{P}\left(\frac{nS^2}{\sigma^2} > \chi_2^2\right) = q_2 = \frac{\alpha}{2} = \frac{1-\delta}{2}.$$

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

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

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

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

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

adică obţinem **intervalul de încredere pentru dispersia** σ^2

$$\frac{nS^2}{\chi_2^2} < \sigma^2 < \frac{nS^2}{\chi_1^2} \tag{2.14}$$

sau

$$\frac{(n-1)\left(S^*\right)^2}{\chi_2^2} < \sigma^2 < \frac{(n-1)\left(S^*\right)^2}{\chi_1^2},$$

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

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

$$\frac{S\sqrt{n}}{\chi_2} < \sigma < \frac{S\sqrt{n}}{\chi_1} \tag{2.15}$$

sau

$$\frac{S^*\sqrt{n-1}}{\chi_2} < \sigma < \frac{S^*\sqrt{n-1}}{\chi_1}.$$

Remarca 2.35 *Formulele (2.14) şi (2.15) se folosesc în general când $n \leq 30$ şi $X \sim \mathcal{N}\left(m, \sigma^2\right)$*

Remarca 2.36 *Când $n > 30$ folosim faptul că*

$$\chi^2(\nu, 1) \to \mathcal{N}(\nu, 2\nu), \text{ pentru } \nu \to \infty.$$

În cazul nostru deci, pentru n suficient de mare,

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

Atunci vom putea folosi funcţia lui Laplace Φ şi deducem, având în vedere că $\frac{nS^2}{\sigma^2} \sim \mathcal{N}(n-1, 2(n-1))$

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

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

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

şi atunci

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

Folosind tabelul funcţiei lui Laplace vom obţine atunci valoarea χ_2^2.

Exerciţiul 2.37 *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 Luând

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

vom obţine că

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

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

Dar

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

deci

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

Considerăm statistica, sau funcţia de estimaţie,

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

Propoziţia 2.38 *Dacă* $X \sim \chi^2(a, 1)$ *şi* $Y \sim \chi^2(b, 1)$ *atunci v.a.*

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

adică fracţia $\dfrac{X/a}{Y/b}$ *este distribuită Fisher-Snedecor de parametrii* a *şi* b.

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

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

Din condiţia

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

se vor putea determina, utilizând tabelul distribuţiei Fisher-Snedecor $F\left(a, b\right)$, valorile F_1 şi F_2.

Cantitatea $\alpha \overset{def}{=} 1-\delta$ se numeşte prag de semnificaţie.

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

$$\mathbb{P}\left(\frac{\left(S_1^*\right)^2}{\left(S_2^*\right)^2}\frac{\sigma_2^2}{\sigma_1^2} > F_1\right) = q_1 = 1 - \frac{\alpha}{2} = 1 - \frac{1-\delta}{2}$$

iar

cantitatea F_2 se va determina din relaţia

$$\mathbb{P}\left(\frac{\left(S_1^*\right)^2}{\left(S_2^*\right)^2}\frac{\sigma_2^2}{\sigma_1^2} > F_2\right) = q_2 = \frac{\alpha}{2} = \frac{1-\delta}{2}.$$

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

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

unde $f\left(x\right)$ reprezintă densitatea de repartiţie a unei v.a. Y repartizată $F\left(n_1-1, n_2-1\right)$ de parametrii n_1-1 şi n_2-1.

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

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

adică obţinem **intervalul de încredere pentru raportul a două dispersii**

$$\frac{1}{F_2}\frac{(S_1^*)^2}{(S_2^*)^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{F_1}\frac{(S_1^*)^2}{(S_2^*)^2} \tag{2.16}$$

Pe de altă parte are loc **următoarea relaţie între valori**

$$F_{1-\alpha,a,b} = \frac{1}{F_{\alpha,b,a}} \tag{2.17}$$

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

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

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

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

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

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

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

$$\frac{1}{F_{\alpha/2,a,b}}\frac{(S_1^*)^2}{(S_2^*)^2} < \frac{\sigma_1^2}{\sigma_2^2} < F_{\alpha/2,b,a}\frac{(S_1^*)^2}{(S_2^*)^2} \tag{2.18}$$

Exerciţiul 2.39 *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

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

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 cu o valoare (σ cunoscut)

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

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

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

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

sau echivalent

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

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

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

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

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

sau echivalent

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

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 (σ_1, σ_2 cunoscuţi)

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

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

sau echivalent

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

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

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

sau echivalent

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

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** $z_c \ge -z_\alpha$ **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 cu o valoare (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

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

obţinem că

$$t \sim t\left(n-1\right)$$

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

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

Atunci $\mathcal{V} = \left\{\left(x_1, \ldots, x_n\right) \in \mathbb{R}^n : |t_c| > t_\alpha\right\}$ 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 cu o valoare

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

Verificăm ipoteza nulă

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

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

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

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

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

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

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

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

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

Pentru α fixat se va determina, din tabelul repartiţiei $\chi^2(n-1,1)$, 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

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

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

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

Pentru α fixat se va determina, din tabelul repartiţiei $\chi^2(n-1,1)$, 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** $\chi_c^2\le\chi_2^2$ **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

$$F\stackrel{def}{=}\frac{\frac{\left(S_1^*\right)^2}{\sigma_1^2}}{\frac{\left(S_2^*\right)^2}{\sigma_2^2}}=\frac{\sigma_2^2}{\sigma_1^2}\frac{\left(S_1^*\right)^2}{\left(S_2^*\right)^2}=\frac{\sigma_2^2}{\sigma_1^2}\frac{\frac{n_1}{n_1-1}S_1^2}{\frac{n_2}{n_2-1}S_2^2}=\frac{n_2-1}{n_1-1}\frac{\frac{n_1S_1^2}{\sigma_1^2}}{\frac{n_2S_2^2}{\sigma_2^2}}\,.$$

Ştim că $U_1\stackrel{def}{=}\frac{n_1S_1^2}{\sigma_1^2}$ este distribuit $\chi^2\left(n_1-1,1\right)$ iar $U_2\stackrel{def}{=}\frac{n_2S_2^2}{\sigma_2^2}\sim\chi^2\left(n_2-1,1\right)$. 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 $(S_1^*)^2>(S_2^*)^2$.

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

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

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

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

sau echivalent

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

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

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

şi

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

Deci, având în vedere ecuaţiile de mai sus, avem că (vezi desenul, precum şi relaţia (2.17))

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

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 (2.17))

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

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, Kenneth N. Berk, *Modern Mathematical Statistics with Applications* (Second Edition), series: Springer Texts in Statistics, Springer New York, 2012.

[5] Lucian Maticiuc, *Teoria probabilităţilor*, Universitatea "Alexandru Ioan Cuza", Iaşi, math.uaic.ro/ maticiuc, 2017.

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

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

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