Curs 13: Introducere în Statistică. Estimatori

1.1 Problematica statisticii

Dezvoltarea tehnologiei a condus la generarea unui volum imens de date. Pe WEB se generează date în format text, imagine sau alt format multimedia. A crescut cantitatea de date numerice generate în experimentele din fizica energiilor înalte, în astronomie, explorarea spaţiului, biologie etc. Pe de altă parte, se generează date în banking, telekom şi în tranzacţiile mediului de afaceri. Aceste date de volum uriaş ascund informaţie care trebuie extrasă şi utilizată pentru a facilita avansul în domeniile respective.

Există mai multe domenii care dezvoltă tehnici şi proceduri de înregistrare a datelor de analiză şi extragere a informaţiei şi a cunoştiîntelor din date. Pe de o parte, *statistica* are o istorie îndelungată în această direcţie, iar pe de altă parte *machine learning*, *data mining* şi, mai nou, *data science* sunt domenii noi care au apărut şi s-au dezvoltat pe măsură ce a avansat tehnologia sistemelor de calcul.

Machine learning este ştiinţa care se ocupă cu proiectarea, analiza, implementarea şi aplicaţiile programelor ce învaţă din experienţă sau îşi îmbunătăţesc performanţele automat, prin experienţă. Acest domeniu are o intersecţie mare cu statistica. În timp ce *statistica* are ca scop formularea inferenţelor pe baza informaţiei extrase dintr-un eşantion de date, *machine learning* încorporează aspecte adiţionale relativ la analiza algoritmilor ce pot fi folosiţi pentru a capta, stoca, indexa, extrage şi combina aceste date în scopul de a îndeplini sarcini greu de realizat prin mijloace algoritmice clasice.

Printre domeniile care folosesc instrumentele oferite de *machine learning* amintim: *computer vision* (recunoaşterea feţelor, detectarea şi localizarea obiectelor), *information retrieval* (extragerea informaţiei din documentele indexate de către motoarele de căutare, identificarea topicilor din *feed*-uri, regăsirea imaginilor), navigare autonomă, controlul roboţilor, traducere automată (Google translate), bioinformatică şi multe altele. *Machine learning* este unul din domeniile cu cea mai ridicată rată de dezvoltare. Companii mari, ca Google, Yahoo, IBM, Microsoft, investesc enorm în programe de cercetare–dezvoltare în *machine learning*.

Studiile experimentale de laborator, efectuate în diverse domenii din inginerie, fizică, chimie sau experimentele pe calculator din diverse domenii ale ştiinţei şi tehnologiei, constând din simulări ale unor sisteme simple sau complexe, implică investigaţii statistice

ale datelor observate, măsurate sau simulate. Investigarea statistică constă în a studia o caracteristică comună a unei mulţimi de elemente de aceeaşi natură, numită *populaţie*. Elementele unei populaţii se numesc, generic, *indivizi*. Scopul investigaţiei statistice este de a extrage informaţii despre caracteristica populaţiei, investigând doar un eşantion constând din n indivizi, selectaţi la întâmplare. Numărul n al indivizilor din eşantion se numeşte *volumul eşantionului*.

Caracteristica comună a indivizilor populaţiei este cuantificată de o variabilă aleatoare X, pentru care fie nu se cunoaşte distribuţia de probabilitate (densitatea de probabilitate f sau funcţia de repartiţie F), fie se cunoaşte doar parţial, în sensul că se cunoaşte tipul de distribuţie de probabilitate a caracteristicii investigate, dar densitatea de probabilitate f_θ a variabilei aleatoare continue X sau distribuţia de probabilitate p_θ, când X este variabilă aleatoare discretă ($p_\theta(x) = P(X = x)$), depinde de un parametru necunoscut $\theta \in \Theta \subseteq \mathbb{R}^d$, $d \geq 1$.

Observând sau măsurând caracteristica indivizilor dintr-un eşantion, se obţine un şir de valori $x_1, x_2, \ldots, x_n$, interpretate ca valori de observaţie asupra variabilei aleatoare X. Din acestea "se estimează" parametrii de interes, cum ar fi media caracteristicii investigate, dispersia sau parametrii necunoscuţi, de care depinde distribuţia de probabilitate.

Mai precis, statistica dezvoltă metode bazate pe rezultate din teoria probabilităţilor, care permit estimarea parametrului necunoscut $\theta \in \mathbb{R}^d$ al modelului probabilist f_θ sau p_θ.

În procesul de observare sau măsurare a caracteristicii indivizilor dintr-un eşantion se consideră că rezultatul investigării unui individ este independent de cel al investigării celorlalţi. De aceea valorile înregistrate, $x_1, x_2, \ldots, x_n$, sunt interpretate ca valori de observaţie asupra unui şir de variabile aleatoare $X_1, X_2, \ldots, X_n$, independente şi identic distribuite ca variabila aleatoare X, ce modelează caracteristica investigată. Practic, înţelegem prin X_k ca fiind caracteristica individului k din eşantion, $k = \overline{1, n}$. Cele n variabile aleatoare, având aceeaşi distribuţie ca şi X, au atât media $m = M(X_k)$, cât şi dispersia $\sigma^2 = \sigma^2(X_k)$ egale cu cele ale lui X.

Considerăm $\mathcal{P}$ o populaţie supusă investigării statistice din punctul de vedere al unei caracteristici X, ce ia valori discrete sau continue. Perechea (X, f_θ), unde X este o variabilă aleatoare reală de densitate f_θ, dacă X este continuă, respectiv $f_\theta(x) = p_\theta(x)$, adică $f_\theta(x) = P(X = x)$, dacă X este discretă, se numeşte *model statistic*.

De exemplu, populaţia poate fi un anumit tip de chip-uri şi caracteristica pe care dorim s-o investigăm este durata de viaţă. În general, durata de viaţă a dispozitivelor şi circuitelor este exponenţial distribuită. Astfel, modelul statistic al caracteristicii durată de viaţă este (X, f_θ) cu $X \sim \text{Exp}(\theta)$, θ fiind un parametru necunoscut. Înregistrând durata de viaţă a n chipuri selectate la întâmplare din producţia dintr-o anumită perioadă, se va estima parametrul θ, care se ştie că reprezintă media variabilei aleatoare X a modelului exponenţial. Având un estimator al lui θ, firma producătoare poate stabili garanţia pe care o dă pentru buna funcţionare a tipului respectiv de chip-uri.

Nu întotdeauna populaţia constă din obiecte fizice, palpabile. De exemplu, pentru a deduce distribuţia de probabilitate a intervalului dintre două pachete de informaţie pe un canal de comunicaţie, populaţia investigată constă din astfel de intervale. În acest caz eşantionul nu se alege apriori şi apoi să se facă măsurătorile, ci se observă direct într-o

perioadă dată, într-un anumit tip de canal de comunicaţie, pachetele de informaţie şi se înregistrează lungimea intervalelor dintre două pachete succesive. Numele de populaţie şi indivizi vine din biologie şi demografie, domenii în care s-au făcut pentru prima dată investigaţii statistice.

Definiţia 1.1.1 Fie (X, f_θ) un model statistic asociat unei populaţii. Un vector aleator $\xi = (X_1, X_2, \dots X_n)$, ale cărui coordonate sunt independente şi identic distribuite după legea modelului f, se numeşte *selecţie aleatoare* de volum n. În urma investigării prin sondaj a populaţiei, se înregistrează n valori numerice $(x_1, x_2, \dots, x_n)$, numite *valori de selecţie* sau *valori de realizare* a selecţiei aleatoare ξ.

Fie $\xi = (X_1, X_2, \dots, X_n)$ o selecţie aleatoare de volum n asociată modelului statistic (X, f_θ). O funcţie reală continuă de variabile $X_1, X_2, \dots, X_n$,

$$Y = T(X_1, X_2, \dots, X_n),$$

este o variabilă aleatoare, numită *statistică*. Dacă $(x_1, x_2, \dots, x_n)$ este o realizare a selecţiei aleatoare $(X_1, X_2, \dots, X_n)$, atunci $T(x_1, x_2, \dots, x_n)$ este o realizare a lui Y. Distribuţia de probabilitate a statisticii Y se numeşte *distribuţia de selecţie* a statisticii. Această distribuţie poate fi dedusă pe baza unor rezultate de teoria probabilităţilor sau poate fi aproximată.

Având valorile de selecţie $x_1, x_2, \dots, x_n$, primele informaţii ce se extrag din acestea sunt *media de selecţie* sau *media experimentală*, notată cu $\overline{x}$, care este media lor aritmetică:

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

şi *dispersia de selecţie* (*dispersia experimentală*), s^2, definită prin:

$$s^2 = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2.$$

Radical din dispersia de selecţie, $\sqrt{s^2}$, se notează cu s şi se numeşte *abaterea standard* a eşantionului.

Media de selecţie $\overline{x}$ este realizare a statisticii $\overline{X}$, unde

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

iar dispersia de selecţie este o realizare a statisticii

$$S^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \overline{X})^2.$$

Propoziţia 1.1.1 *Dacă $(X_1, X_2, \dots, X_n)$ este o selecţie aleatoare de volum n dintr-o populaţie modelată statistic de (X, f), cu $M(X) = m$ şi $\sigma^2(X) = \sigma^2$, atunci statistica medie $\overline{X}$ are şi ea aceeaşi valoare medie m, iar dispersia sa este $D^2 = \sigma^2/n$.*

Demonstraţie: Variabilele X_k au media $m = M(X_k)$ şi dispersia $\sigma^2 = \sigma^2(X_k)$. Astfel,

$$M(\overline{X}) = M\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{M(X_1) + \cdots + M(X_n)}{n} = \frac{n\,m}{n} = m,$$

iar

$$\sigma^2(\overline{X}) = \sigma^2\left(\frac{X_1}{n} + \cdots + \frac{X_n}{n}\right) = \frac{1}{n^2}\sigma^2(X_1) + \cdots + \frac{1}{n^2}\sigma^2(X_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

În calculul dispersiei am aplicat formula de calcul a dispersiei unei combinaţii liniare cu coeficienţi reali de variabile aleatoare independente:

$$\sigma^2(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1^2\sigma^2(X_1) + a_2^2\sigma^2(X_2) + \cdots + a_n^2\sigma^2(X_n).$$

□

1.2 Estimatori ai parametrilor modelelor statistice

Fie (X, f_θ) un model statistic şi $(x_1, x_2, \ldots, x_n)$ o realizare a unei selecţii aleatoare de volum n, $(X_1, X_2, \ldots, X_n)$.

Un estimator punctual al parametrului θ este o funcţie $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$. Este evident că un estimator punctual este o realizare a variabilei aleatoare $\hat{\theta}(X_1, X_2, \ldots, X_n)$. Deoarece există o infinitate de funcţii $\hat{\theta}$, înseamnă că există o infinitate de estimatori ai parametrului θ. Este însă rezonabil să alegem estimatori care să aproximeze parametrul distribuţiei f_θ cu o probabilitate suficient de mare.

Definiţia 1.2.1 Estimatorul $\hat{\theta}(x_1, x_2, \ldots, x_n)$ cu proprietatea că pentru orice $\varepsilon > 0$ are loc

$$\lim_{n\to\infty} P(|\,\hat{\theta}(X_1, \ldots, X_n) - \theta| > \varepsilon) = 0 \tag{1}$$

se numeşte *estimator consistent* al parametrului θ.

Intuitiv, dacă estimatorul este consistent, atunci distribuţia statisticii $\hat{\theta}(X_1, \ldots, X_n)$ este din ce în ce mai concentrată în jurul parametrului θ pe măsură ce volumul selecţiei creşte.

Definiţia 1.2.2 Un estimator $\hat{\theta}(x_1, x_2, \ldots, x_n)$ care verifică proprietatea că valoarea medie a statisticii $\hat{\theta}(X_1, X_2, \ldots, X_n)$ este chiar parametrul θ, adică

$$M(\hat{\theta}(X_1, X_2, \ldots, X_n)) = \theta, \tag{2}$$

se numeşte *estimator centrat* sau *nedeplasat*.

Definiţia 1.2.3 Fie $\hat{\theta}_1$, $\hat{\theta}_2$ doi estimatori nedeplasaţi ai parametrului θ. Dacă între dispersiile statisticilor $\hat{\theta}_1(X_1, \ldots, X_n)$ şi $\hat{\theta}_2(X_1, \ldots, X_n)$ există relaţia

$$\sigma^2(\hat{\theta}_1(X_1, \ldots, X_n)) \leq \sigma^2(\hat{\theta}_2(X_1, \ldots, X_n)), \tag{3}$$

atunci estimatorul $\hat{\theta}_1$ se zice că este *mai eficient* decât estimatorul $\hat{\theta}_2$.

Exemplul 1. Fie (x_1, x_2) o realizare a selecţiei aleatoare (X_1, X_2), distribuţia comună a variabilelor aleatoare X_i, $i = 1, 2$, fiind distribuţia exponenţială de parametru θ,

$$f_\theta(x) = \begin{cases} \frac{1}{\theta} e^{-x/\theta}, & \text{dacă } x \geq 0, \\ 0, & \text{dacă } x < 0. \end{cases}$$

Considerăm trei estimatori pentru θ:

$$\hat{\theta}_1 = x_1, \ \hat{\theta}_2 = \frac{x_1 + x_2}{2}, \ \hat{\theta}_3 = \frac{x_1 + x_2}{3}.$$

Estimatorii daţi sunt realizări ale variabilelor aleatoare:

$$\begin{aligned} \hat{\theta}_1(X_1, X_2) &= X_1, \\ \hat{\theta}_2(X_1, X_2) &= (X_1 + X_2)/2, \\ \hat{\theta}_3(X_1, X_2) &= (X_1 + X_2)/3. \end{aligned}$$

Se ştie că media unei variabile aleatoare exponenţial distribuite este θ. Prin urmare, $M(X_1) = \theta$, $M((X_1 + X_2)/2) = (\theta + \theta)/2 = \theta$, iar $M((X_1 + X_2)/3) = 2\theta/3$. Astfel, primii doi estimatori sunt nedeplasaţi, iar al treielea este deplasat. Să determinăm care este mai eficient dintre primii doi:

$$\sigma^2(X_1) = \theta^2, \sigma^2((X_1 + X_2)/2) = \sigma^2(X_1)/4 + \sigma^2(X_2)/4 = \theta^2/4 + \theta^2/4 = \theta^2/2.$$

În concluzie al doilea estimator este mai eficient.

Existând mai multe posibilităţi de a defini estimatori ai parametrului, ne întrebăm dacă există o metodă ce permite definirea unui estimator "bun". Răspunsul este pozitiv în câteva cazuri de interes.

1.2.1 Estimarea mediei

Fie (X, f) un model statistic continuu sau discret şi $x_1, x_2, \ldots, x_n$ observaţii independente din legea f.

Propoziţia 1.2.1 *Media de selecţie a observaţiilor $x_1, x_2, \ldots, x_n$,*

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}, \tag{4}$$

este un estimator nedeplasat al mediei $m = M(X)$ a modelului statistic.

Demonstraţie: Rezultă imediat din faptul că $M(\overline{X}) = m$, deci $\hat{m} = \overline{x}$ este un estimator nedeplasat al mediei m. □

În concluzie, un bun estimator al mediei $M(X)$, a distribuţiei oricărui model statistic, este media aritmetică a valorilor de selecţie.

Exemplul 2. Cererea de memorie pentru o aplicaţie, ca proporţie din memoria ce poate fi alocată de un utilizator, este o variabilă aleatoare X ce are densitatea de probabilitate

$$f(x) = \begin{cases} (\theta+1)x^{\theta}, & 0 < x < 1, \\ 0, & \text{în rest.} \end{cases}$$

a) Să se determine media teoretică $M(X)$ a variabilei aleatoare X şi apoi să se estimeze θ în funcţie de media de selecţie $\overline{x}$ a unei selecţii aleatoare de volum n.
b) Să se determine un estimator al parametrului θ din selecţia următoare:

$$0.2, 0.4, 0.5, 0.7, 0.8, 0.9, 0.9, 0.6, 0.6, 0.4,$$

rezultată în urma rulării aplicaţiei cu diferite date de intrare.

Rezolvare: a) Mai întâi calculăm media teoretică:

$$M(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_0^1 (\theta+1)x^{\theta+1}dx = (\theta+1)\frac{x^{\theta+2}}{\theta+2}\Big|_0^1 = \frac{\theta+1}{\theta+2}.$$

Dacă $m = M(X)$ şi $\overline{x}$ este media de selecţie a unui eşantion de valori $x_1, x_2, \ldots, x_n$, atunci din egalitatea impusă $\hat{m} = \overline{x}$ se determină un estimator al parametrului θ:

$$\frac{\hat{\theta}+1}{\hat{\theta}+2} = \overline{x} \quad \Leftrightarrow \quad \hat{\theta} = \frac{2\overline{x}-1}{1-\overline{x}}.$$

b) Pentru valorile înregistrate avem $\overline{x} = 0.6$, deci un estimator pentru parametrul θ este $\hat{\theta} = \dfrac{2\overline{x}-1}{1-\overline{x}} = 0.5$.

Exemplul 3. Pentru a estima rata sosirii λ a cererilor de acces la o bază de date s-au monitorizat intervalele de timp dintre 10 cereri consecutive şi s-au înregistrat valorile:

$$0.2, 0.1, 0.1, 0.05, 0.05, 0.2, 0.8, 0.5, 0.2, 0.8.$$

Care este estimatorul ratei sosirilor, $\hat{\lambda}$?

Sosirile cererilor la o bază de date este un proces Poisson (N_t), $t \geq 0$, de rată $\lambda > 0$. Intervalul inter–sosirilor are distribuţia exponenţială, $X \sim Exp(\theta = 1/\lambda)$. Dar parametrul θ pentru distribuţia exponenţială este valoarea medie a variabilei X.

Prin urmarem din datele înregistrate putem calcula un estimator al lui θ, adică a mediei lungimii intervalelor inter-sosirilor: $\hat{\theta} = 0.3$. Din relaţia $\theta = 1/\lambda$, adică $\lambda = 1/\theta$, obţinem un estimator al ratei sosirilor ca fiind $\hat{\lambda} = 1/\hat{\theta} = 1/0.3 = 3.33$.

1.2.2 Estimarea dispersiei

Propoziţia 1.2.2 *Dacă (X, f) este un model statistic şi m, σ^2 sunt media şi dispersia variabilei aleatoare X, atunci dispersia valorilor de selecţie $x_1, x_2, \ldots, x_n$ din legea de probabilitate definită de f,*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2, \tag{5}$$

este un estimator nedeplasat al dispersiei $\sigma^2(X)$.

Demonstraţie: Statistica S^2, a cărei realizare este s^2, este dată de relaţia:

$$S^2 = \hat{\theta}(X_1, X_2, \ldots, X_n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \tag{6}$$

unde $\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$. Media acestei statistici este

$$M(S^2) = \frac{1}{n-1} M\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right). \tag{7}$$

Să explicităm $\sum_{i=1}^{n}(X_i - \overline{X})^2$:

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \overline{X})^2 &= \sum_{i=1}^{n}[(X_i - m) - (\overline{X} - m)]^2 \\
&= \sum_{i=1}^{n}\left((X_i - m)^2 - 2(\overline{X} - m)(X_i - m) + (\overline{X} - m)^2\right) \\
&= \sum_{i=1}^{n}(X_i - m)^2 - 2(\overline{X} - m)\sum_{i=1}^{n}(X_i - m) + n(\overline{X} - m)^2 \\
&= \sum_{i=1}^{n}(X_i - m)^2 - 2n(\overline{X} - m)^2 + n(\overline{X} - m)^2 \\
&= \sum_{i=1}^{n}(X_i - m)^2 - n(\overline{X} - m)^2,
\end{aligned} \tag{8}$$

ceea ce implică

$$M\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = M\left(\sum_{i=1}^{n}(X_i - m)^2\right) - nM((\overline{X} - m)^2). \tag{9}$$

Variabilele aleatoare $X_1, X_2, \ldots, X_n$ sunt identic distribuite şi, prin urmare, au aceeaşi medie m şi dispersie σ^2. Deci, $M\left((X_i - m)^2\right) = \sigma^2$, $\forall\, i = \overline{1, n}$, iar

$$M\left(\sum_{i=1}^{n}(X_i - m)^2\right) = \sum_{i=1}^{n} M\left((X_i - m)^2\right) = n\sigma^2.$$

Astfel,

$$M\left(\sum_{i=1}^{n}(X_i-\overline{X})^2\right)=n\sigma^2-nD^2(\overline{X})=n\sigma^2-n\frac{\sigma^2}{n}=(n-1)\sigma^2. \tag{10}$$

Împărţind la $n-1$, obţinem rezultatul dorit. □

Propoziţia 1.2.2 afirmă că media statisticii S^2 este $M(S^2)=\sigma^2$, unde σ^2 este dispersia legii de probabilitate a modelului statistic.

1.3 Estimatorul verosimilităţii maxime

Estimatorii nedeplasaţi pentru medie şi dispersie s-au determinat simplu, din valorile de selecţie dintr-o lege de probabilitate total necunoscută, prin metode ce nu sunt explicit formulate. În cazul în care legea de probabilitate a modelului statistic este cunoscută, în sensul că se cunoaşte densitatea de probabilitate, dar aceasta depinde de unul sau mai mulţi parametrii necunoscuţi, estimatorii parametrilor pot fi determinaţi prin metode bine fundamentate ştiinţific. În acest scop considerăm că se supune investigaţiei statistice o caracteristică a unei populaţii, măsurată de o variabilă aleatoare X a cărei densitate de probabilitate f_θ depinde de un parametru necunoscut θ. Se investighează un eşantion de volum n din populaţia respectivă şi se înregistrează valorile (de observaţie asupra lui X) $x_1,x_2,\ldots,x_n$. Apoi se determină un estimator pentru parametrul θ, care maximizează probabilitatea înregistrării unor valori foarte apropiate de acestea.

Mai precis, pentru fiecare parametru θ, probabilitatea ca variabila aleatoare X să ia valori apropiate de x_i este probabilitatea ca X să ia valori intr-un interval de forma $[x_i,x_i+h)$, cu h foarte mic. Această probabilitate este

$$P(X\in[x_i,x_i+h))=\int_{x_i}^{x_i+h}f_\theta(x)\,dx\approx f_\theta(x_i)h.$$

Cu alte cuvinte, aria trapezului curbiliniu de baza h (valoarea integralei) se poate aproxima cu aria dreptunghiului de bază h şi înălţime $f_\theta(x_i)$.

Notând cu $X_1,X_2,\ldots,X_n$ variabilele aleatoare independente şi identic distribuite ca X, avem că probabilitatea ca variabilele $X_1,X_2,\ldots,X_n$ să ia valori foarte apropiate de valorile înregistrate $x_1,x_2,\ldots,x_n$ este:

$$\begin{aligned}P(X_1\in[x_1,x_1+h),X_2\in[x_2,x_2+h),\ldots,X_n\in[x_n,x_n+h))\\ =P(X_1\in[x_1,x_1+h))P(X_2\in[x_2,x_2+h))\cdots P(X_n\in[x_n,x_n+h))\\ =f_\theta(x_1)f_\theta(x_2)\cdots f_\theta(x_n)h^n.\end{aligned}$$

Deoarece h^n nu depinde de θ, rezultă că parametrul θ ce maximizează probabilitatea

$$P(X_1\in[x_1,x_1+h),X_2\in[x_2,x_2+h),\ldots,X_n\in[x_n,x_n+h))$$

este parametrul ce maximizează produsul

$$f_\theta(x_1)f_\theta(x_2)\cdots f_\theta(x_n).$$

Definiţia 1.3.1 Funcţia $L : \mathbb{R} \to \mathbb{R}$, de variabilă θ, asociată eşantionului $x_1, x_2, \ldots, x_n$, definită prin

$$L(\theta; x_1, x_2, \ldots, x_n) = f_\theta(x_1) \cdot f_\theta(x_2) \cdots f_\theta(x_n),$$

se numeşte *funcţia de verosimilitate*.

Valorile $x_1, x_2, \ldots, x_n$ fiind cunoscute (fiind valorile rezultate din observaţii), funcţia de verosimilitate este o funcţie de o singură variabilă, θ.

Dacă eşantionul s-a extras dintr-o populaţie a cărei distribuţie de probabilitate este discretă şi depinde de un parametru θ, adică $p_X(x; \theta) = P(X = x)$, atunci definim funcţia de verosimilitate astfel

$$L(\theta; x_1, x_2, \ldots, x_n) = p_X(x_1; \theta) p_X(x_2; \theta) \cdots p_X(x_n; \theta).$$

În acest caz, $L(\theta; x_1, x_2, \ldots, x_n)$ reprezintă, datorită independenţei variabilelor discrete $X_1, X_2, \ldots, X_n$, probabiliatea ca X_1 să ia valoarea x_1, X_2 să ia valoarea $x_2, \ldots, X_n$ să ia valoarea x_n.

Atât în cazul continuu, cât şi în cel discret, **estimatorul verosimilităţii maxime** a parametrului θ este

$$\hat{\theta} = \text{argmax}(L(\theta; x_1, x_2, \ldots, x_n)),$$

unde prin $\text{argmax}(L(\theta; x_1, x_2, \ldots, x_n))$ se înţelege argumentul θ care maximizează funcţia L.

Exemplul 4. Fie populaţia $\mathcal{P}$ formată dintr-un tip de circuite. Caracteristica ce dorim să o investigăm prin sondaj statistic este durata de viaţă a acestor circuite, ştiind că aceasta este exponenţial distribuită, cu parametrul θ necunoscut. Măsurând timpul de viaţă (în ani) a 10 circuite, se obţin valorile:

$$0.8830,\ 1.96511,\ 1.9189,\ 4.8448,\ 0.9208\ 3.4377,\ 1.7162,\ 4.2327,\ 5.9435,\ 8.3128.$$

Să determinăm estimatorul de verosimilitate maximă pentru θ (adică pentru media duratei de viaţă a acestui tip de circuite). Densitatea de probabilitate a distribuţiei exponenţiale este

$$f_\theta(x) = \begin{cases} 0, & \text{dacă } x < 0, \\ \frac{1}{\theta} e^{-x/\theta}, & \text{dacă } x \geq 0. \end{cases} \tag{11}$$

Astfel, funcţia de verosimilitate este

$$L(\theta; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-x_i/\theta} = \frac{1}{\theta^n} e^{-\sum_{i=1}^{n} x_i/\theta}. \tag{12}$$

Funcţia logaritmică cu bază mai mare ca 1 are derivata întâi pozitivă. Notând cu $h = \ln$ şi cu $\ell(\theta) = h(L(\theta))$, avem că funcţia $\ell'(\theta) = h(L(\theta))L'(\theta)$ are acelaşi semn ca derivata lui L, deci ℓ şi L au aceleaşi puncte de extrem şi de aceeaşi natură. Pentru simplitatea calculelor, determinăm punctul de maxim absolut (dacă acesta există) pentru ℓ şi acesta va fi punct de maxim absolut şi pentru L:

$$\ell(\theta) = \ln L(\theta; x_1, x_2, \ldots, x_n) = -n \ln \theta - \frac{\sum_{i=1}^{n} x_i}{\theta}. \tag{13}$$

Avem

$$\ell'(\theta) = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2}. \tag{14}$$

Rezolvând ecuaţia $\ell'(\theta) = 0$ în raport cu θ, obţinem punctul $\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$, care este maxim absolut pentru ℓ, deci şi pentru L. În concluzie,

$$\text{argmax} L(\theta; x_1, x_2, \ldots, x_n) = \overline{x},$$

estimatorul de verosimilitate maximă a parametrului θ a distribuţiei exponenţiale este media de selecţie. În cazul exemplului dat, estimatorul verosimilităţii maxime a mediei de viaţă a circuitelor este media selecţiei:

$$\hat{\theta} = \frac{x_1 + x_2 + \ldots + x_{10}}{10} = 5.1861.$$

Observaţie: Valorile de selecţie au fost generate simulând o variabilă $X \sim Exp(\theta = 5)$, deci estimatorul verosimilitaţii maxime $\hat{\theta} = 5.1861$ este "destul de bun".

Exemplul 5. Un simulator al distribuţiei Bernoulli

$$X = \begin{pmatrix} 1 & 0 \\ p & 1-p \end{pmatrix},$$

de parametru p necunoscut, generează stringul de biţi:

$$1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0.$$

Să se determine estimatorul verosimilităţii maxime al parametrului p pe baza eşantionului de biţi.

Rezolvare: Notând parametrul p necunoscut cu θ, distribuţia de probabilitate a variabilei X este $p_X(\theta; b) = P(X = b)$, unde b este bitul 1 sau 0.

$$p_X(\theta; b) = \begin{cases} \theta, & \text{dacă } b = 1, \\ 1 - \theta, & \text{dacă } b = 0. \end{cases}$$

Funcţia de verosimilitate asociată eşantionului de biţi generaţi este

$$L(\theta; b_1, b_2, \ldots, b_{26}) = p_X(\theta; b1) p_X(\theta; b_2) \cdots p_X(\theta; b_{26}) = \theta^{12}(1-\theta)^{14},$$

unde 12 este numărul de biţi 1 din string, iar 14 numărul de biţi 0. Se logaritmează şi se determină punctul de maxim absolut al funcţiei $\ell(\theta) = \ln(L(\theta))$. Cum

$$\ell(\theta) = 12\ln(\theta) + 14\ln(1-\theta),$$

rezultă

$$\ell'(\theta) = \frac{12}{\theta} - \frac{14}{1-\theta} = \frac{12(1-\theta) - 14\theta}{\theta(1-\theta)}.$$

Ecuaţia $\ell'(\theta) = 0$ are soluţia $\hat{\theta} = \dfrac{12}{26}$. Se verifică că acesta este un punct de maxim pentru ℓ, deci estimatorul verosimilităţii maxime pentru parametrul p al distribuţiei Bernoulli, dedus din stringul de biţi, este egal cu numărul biţilor 1 din string supra numărul total de biţi. Acest estimator al lui p este de fapt probabilitatea intuitivă de a obţine bitul 1: numărul cazurilor favorabile din string supra numărul cazurilor posibile.

1.3.1 Estimarea mediei şi dispersiei distribuţiei normale

Distribuţia normală apare în numeroase modele statistice şi estimarea parametrilor ei intervine, de exemplu, în numeroase probleme de *machine learning*. Tocmai de aceea, prezintă interes deteminarea estimatorului de verosimilitate maximă atât pentru media, cât şi pentru dispersia lui $N(m, \sigma^2)$. Discutăm trei cazuri:

1. Fie $(X, f(x; m, \sigma^2))$ un model statistic caracterizat de distribuţia normală de **medie** m **necunoscută** şi **dispersie cunoscută**. Densitatea de probabilitate a distribuţiei normale este

$$f(x; m) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/(2\sigma^2)},\ x \in \mathbb{R}. \tag{15}$$

Funcţia de verosimilitate asociată realizării $(x_1, x_2, \ldots, x_n)$ a unei selecţii aleatoare de volum n este

$$L(m) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\dfrac{(x_i-m)^2}{2\sigma^2}} = Ce^{-\displaystyle\sum_{i=1}^{n} \frac{(x_i-m)^2}{2\sigma^2}},\ m \in \mathbb{R}, \tag{16}$$

unde C este o constantă pozitivă ce nu depinde de m. Pentru a determina punctele de extrem ale lui L considerăm funcţia ℓ, $\ell = \ln(L)$:

$$\ell(m) = \ln C - \sum_{i=1}^{n} \frac{(x_i-m)^2}{2\sigma^2}. \tag{17}$$

Calculăm

$$\ell'(m) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(x_i - m) = \frac{1}{\sigma^2}(x_1 + x_2 + \cdots + x_n - nm). \tag{18}$$

Punctul $\hat{m}(x_1, x_2, \ldots, x_n) = (x_1 + x_2 + \cdots + x_n)/n$ este punctul de maxim pentru ℓ, deci şi pentru L. Prin urmare, *media selecţiei este estimatorul de verosimilitate maximă pentru media distribuţiei normale.*

2. În cazul în care media este cunoscută şi dispersia necunoscută, spaţiul parametrului σ^2 este $\Theta = (0, \infty)$. Printr-un calcul analog, rezultă că estimatorul de verosimilitate maximă al dispersiei este

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2. \tag{19}$$

Evident că acest estimator este deplasat, adică

$$M\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2\right) \neq \sigma^2. \tag{20}$$

Un estimator nedeplasat pentru dispersie este

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{21}$$

Se spune că s^2 s-a obţinut din $\hat{\sigma}^2$ prin ajustare.

3. A treia alternativă este atunci când atât media, cât şi dispersia sunt necunoscute. În acest caz funcţia de verosimilitate depinde de doi parametrii:

$$L(m, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n}\frac{(x_i - m)^2}{2\sigma^2}}. \tag{22}$$

Logaritmând din nou, obţinem

$$\ell(m, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \sum_{i=1}^{n}\frac{(x_i - m)^2}{2\sigma^2}. \tag{23}$$

Punctele staţionare ale lui ℓ sunt soluţii ale sistemului:

$$\frac{\partial \ell}{\partial m} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - m)^2 = 0 \tag{24}$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - m)^2 = 0, \tag{25}$$

adică

$$\begin{aligned} \hat{m}(x_1, x_2, \ldots, x_n) &= \frac{x_1 + x_2 + \ldots + x_n}{n} = \overline{x}, \\ \hat{\sigma}^2(x_1, x_2 \ldots, x_n) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2. \end{aligned}$$

Se verifică apoi că $(\hat{m}, \hat{\sigma}^2)$ este punct de maxim pentru ℓ, deci şi pentru L, arătând că $\dfrac{\partial^2 \ell}{\partial m^2}(\hat{m}, \hat{\sigma}^2) < 0$ şi

$$\begin{vmatrix} \dfrac{\partial^2 \ell}{\partial m^2} & \dfrac{\partial^2 \ell}{\partial m \partial \sigma^2} \\ \dfrac{\partial^2 \ell}{\partial m \partial \sigma^2} & \dfrac{\partial^2 \ell}{\partial (\sigma^2)^2} \end{vmatrix} (\hat{m}, \hat{\sigma}^2) > 0. \tag{26}$$

Estimatorul de maximă verosimilitate pentru medie este nedeplasat, dar pentru dispersie este deplasat. Ajustând estimatorul $\hat{\sigma}^2$ la

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

obţinem estimator nedeplasat.

1.4 Teorema limită centrală

În statistică prezintă importanţă deosebită variabila medie aritmetică (media de selecţie), notată $\overline{X}$, asociată unei selecţii aleatoare $(X_1, X_2, \ldots, X_n)$,

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}. \tag{27}$$

Dacă $x_1, x_2, \ldots x_n$ sunt valori de selecţie, atunci media de selecţie $\overline{x}$ este o observaţie asupra variabilei $\overline{X}$. Variabila aleatoare medie aritmetică $\overline{X}$ are media $M(\overline{X}) = m$ şi dispersia $D^2(\overline{X}) = \sigma^2/n$. Remarcăm că dispersia mediei aritmetice descreşte invers proporţional cu n, iar abaterea standard descreşte invers proporţional cu $\sqrt{n}$. Cum abaterea standard este o măsură a împrăştierii valorilor lui $\overline{X}$, relaţia $D(\overline{X}) = \sigma/\sqrt{n}$ arată că pe măsură ce n creşte, distribuţia mediei aritmetice este din ce în ce mai concentrată în jurul valorii medii m. De exemplu, dacă $n = 100$, $m = 0$ şi $\sigma = 3$, atunci $D(\overline{X}) = 3/10 = 0.3$. Crescând n la 1000, $D(\overline{X}) = 3/\sqrt{1000} = 0.0949$.

Propoziţia 1.4.1 *Dacă $X_1, X_2, \ldots, X_n$ sunt variabile aleatoare independente şi normal distribuite, $X_i \sim N(m_i, \sigma_i^2)$, atunci pentru orice $a_i \in \mathbb{R}$, $i = \overline{1, n}$, combinaţia liniară a variabilelor aleatoare cu coeficienţii a_i este normal distribuită,*

$$X = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n \sim N\left(\sum_{i=1}^{n} a_i m_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2 \right). \tag{28}$$

Precizare: Media şi dispersia oricărei combinaţii liniare de variabile independente se calculează la fel:

$$M(a_1 X_1 + \cdots + a_n X_n) = a_1 M(X_1) + \cdots + a_n M(X_n),$$

$$\sigma^2(a_1X_1 + \cdots + a_nX_n) = a_1^2\sigma^2(X_1) + \cdots + a_n^2\sigma^2(X_n).$$

Informaţia adiţională pe care o aduce Propoziţia 1.4.1 este că dacă variabilele aleatoare ce intră în combinaţie au distribuţia normală, atunci şi combinaţia lor liniară are distribuţia normală.

Exemplul 6. Unui semnal X, transmis printr-un canal de comunicaţie, i se adaugă un zgomot N. Ştiind că variabilele aleatoare X şi N sunt independente şi $X \sim N(0,1)$, $N \sim N(0,\sigma^2)$, să se determine distribuţia de probabilitate a semnalului $Y = X + N$ detectat.

Rezolvare: Y fiind o combinaţie liniară cu coeficienţii $a_1 = a_2 = 1$ a variabilelor aleatoare independente şi normal distribuite X şi N, rezultă că $Y \sim N(0, 1+\sigma^2)$, unde $1+\sigma^2$ reprezintă dispersia variabilei Y.

În statistică este important următorul rezultat:

Corolar 1.4.1 *Dacă variabilele aleatoare i.i.d. $X_1, X_2, \ldots, X_n$ au distribuţia normală, $X_i \sim N(m,\sigma^2)$, $i = \overline{1,n}$, atunci media lor aritmetică are de asemenea distribuţia normală, cu aceeaşi medie m şi dispersie $\dfrac{\sigma^2}{n}$, adică*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(m, \sigma^2/n). \tag{29}$$

Ne întrebăm în mod natural ce distribuţie de probabilitate are media aritmetică a n variabile aleatoare i.i.d. având distribuţia comună de probabilitate absolut arbitrară, adică ne-normală (ea putând fi exponenţială, binomială etc). Răspunsul este dat de unul din cele mai remarcabile rezultate din teoria probabilităţilor, cu aplicaţii importante în statistică:

Teorema 1.4.1 (*Teorema limită centrală*) *Dacă (X_n) este un şir de variabile aleatoare independente şi identic distribuite având media comună m şi abaterea standard σ, iar*

$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \tag{30}$$

este şirul variabilelor medie aritmetică, atunci pentru $n \to \infty$ distribuţia de probabilitate a variabilelor $\overline{X}_n$ este aproximativ normală de medie m şi dispersie $D^2 = \sigma^2/n$. Notăm $\overline{X}_n \sim ApN(m, D^2 = \sigma^2/n)$.

Cu alte cuvinte, teorema precedentă afirmă că "în medie totul este normal". Formulat riguros (matematic), teorema limită centrală asigură în condiţiile enuntaţe că funcţiile de repartiţie ale variabilelor standardizate

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

tind, când $n \to \infty$, la funcţia de repartiţie Φ a distribuţiei normale standard, $N(0,1)$, adică

$$\lim_{n\to\infty} P(Z_n \leq x) = \Phi(x).$$

În practica statistică se consideră că pentru $n \geq 30$, distribuţia normală poate fi folosită ca distribuţie a mediei aritmetice a n variabile aleatoare i.i.d. cu media şi dispersia finită. Cu alte cuvinte, dacă $x_1, x_2, \ldots, x_n$ este un eşantion de volum $n \geq 30$ dintr-o populaţie a cărei caracteristică de interes are o distribuţie de probabilitate arbitrară, de medie m şi abatere standard σ, media de selecţie $\overline{x}$ poate fi considerată ca o observaţie asupra unei variabile aleatoare normal distribuite de medie m şi abatere standard $\sigma/\sqrt{n}$.

Pentru a verifica această proprietate efectuăm următorul experiment: generăm de 100 de ori consecutiv câte n numere aleatoare din distribuţia exponenţială de parametru $\theta = 2.3$. Asociem fiecărui şir de numere aleatoare

$$x_1^k, x_2^k, \ldots x_n^k,\ k = \overline{1,100}, \tag{31}$$

media aritmetică a valorilor sale

$$\overline{x}^k = \frac{x_1^k + x_2^k + \cdots + x_n^k}{n},\ k = \overline{1,100}. \tag{32}$$

Valorile $\overline{x}^1, \overline{x}^2, \ldots \overline{x}^{100}$ le interpretăm ca observaţii asupra variabilei aleatoare

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

adică a mediei de selecţie a n variabile aleatoare i.i.d. $X_1, X_2, \ldots, X_n \sim \text{Exp}(\theta = 2.3)$. Generăm apoi funcţia de repartiţie empirică a datelor $\overline{x}^1, \overline{x}^2, \ldots, \overline{x}^{100}$ şi afişăm simultan graficul acesteia, comparativ cu graficul repartiţiei Φ a distribuţiei normale standard. Se observă că pe măsură ce volumul n al eşantioanelor generate creşte, funcţia de repartiţie empirică este din ce în ce mai bine aproximată de repartiţia Φ.

O altă metodă de a vizualiza rezultatul teoremei limită centrală este de a genera histograma datelor $\overline{x}^1, \overline{x}^2, \ldots, \overline{x}^{100}$, care pe măsură ce n creşte, aproximează clopotul lui Gauss.

Teorema limită centrală prezintă interes şi în următorul context: şirului de variabile aleatoare i.i.d. (X_n) îi asociem şirul (S_n), definit prin

$$S_n = X_1 + X_2 + \cdots + X_n.$$

Evident, $S_n = n\overline{X}_n$, $M(S_n) = M(n\overline{X}_n) = nM(\overline{X}_n) = nm$, iar

$$\sigma^2(S_n) = \sigma^2(n\overline{X}_n) = n^2\sigma^2(\overline{X}_n) = n^2\sigma^2/n = n\sigma^2.$$

Prin urmare, pentru n suficient de mare, S_n fiind combinaţie liniară a unor variabile aleatoare aproximativ normal distribuite, este şi ea aproximativ normal distribuită:

$$S_n \sim ApN(nm, D^2 = n\sigma^2). \tag{33}$$

Exemplul 7. Variabila aleatoare X, care dă numărul de zone defecte ale unui CD de un anumit tip, are următoarea distribuţie de probabilitate: $p(x) = P(X = x)$, unde

x	$p(x)$
0	0.75
1	0.15
2	0.10

a) Să se calculeze media şi abaterea standard a numărului de zone defecte ale CD-ului.
b) Ce distribuţie de probabilitate are media de selecţie a unui eşantion de 400 de CD-uri din tipul investigat? Care este media şi dispersia acestei distribuţii?
c) Care este probabilitatea ca media numărului de zone defecte/CD într-un lot de 400 de CD-uri să fie mai mică decât 0.3?

Rezolvare: a) $m = M(X) = 0.15 + 0.2 = 0.35$, iar

$$\sigma^2(X) = (0 - 0.35)^2 0.75 + (1 - 0.35)^2 0.15 + (2 - 0.35)^2 0.1 = 0.4275$$

şi deci abaterea standard este $\sigma = \sigma(X) = \sqrt{\sigma^2(X)} = 0.6538$.
b) Volumul eşantionului fiind mare, conform teoremei limită centrală

$$\overline{X}_{400} \sim \text{ApN}(m, \sigma^2/400),$$

unde $m = 0.35$, $D^2 = \sigma^2/400 = 0.00106$, iar $D = 0.032$.
c) Avem

$$\begin{aligned} P(\overline{X}_{400} < 0.3) &= F_{\overline{X}}(0.3) = \Phi\left(\frac{0.3 - m}{D}\right) = \Phi\left(\frac{0.3 - 0.35}{0.032}\right) \\ &= \Phi(-1.5625) = 1 - \Phi(1.5625) = 1 - 0.94 = 0.06. \end{aligned}$$

Exemplul 8. Se consideră o buclă `for`:

```
for(i=1;i<=n;i++) // n>30
   {
   executa blocul B;
   }
```

Ştiind că timpul de execuţie al blocului `B` este o variabilă aleatoare de distribuţie de probabilitate necunoscută, având media $m = 60ms$ şi abaterea standard de $\sigma = 8ms$, iar execuţiile succesive ale blocului sunt independente, să se determine distribuţia de probabilitate a timpului de execuţie a buclei `for`. Care este probabilitatea ca timpul de execuţie al buclei în cazul $n = 50$ să fie cuprins între $0.75s$ sec şi 1 sec?

Notând cu T_i timpul celei de-a i-a execuţii a blocului B, timpul total de execuţie

$$T = T_1 + T_2 + \cdots + T_{50}$$

este aproximativ normal distribuit.