MS
(curs 13 - S13)

Problema statisticii matematice

Statistica se ocupă cu studiul datelor numerice generate în experimentele din diverse domenii: fizica, astronomie, explorarea spațiului, biologie, banking, telekom, tranzacții financiare, machine learning...

- **Investigarea statistică** =studiul unei caracteristici comune a unei mulțimi de elemente de aceeași natură, numită **populație** (de exemplu: un anume tip de produs:chip, iar caracteristica: timpul de viață).
- Elementele unei populații se numesc, generic, **indivizi**.
- Scopul investigației statistice este de a extrage informații despre caracteristica populației, investigând doar **un eșantion** constând din n indivizi, selectați la întâmplare.
- Numărul n al indivizilor din eșantion se numește **volumul eșantionului**.

Scenariul matematic al statisticii

V.a. X =caracteristica comună a indivizilor populației
(X v.a. ce înregistrează durata de viața a chip-ului studiat)

- f_X =? sau F_X =?, i.e. nu se cunoaște distribuția/densitatea de probabilitate și nici funcția de repartiție F_X
- fie se cunoaște doar parțial, în sensul că se cunoaște tipul de distribuție/densității de probabilitate a caracteristicii investigate, dar depinde de un parametru necunoscut $\theta \in \Theta \subseteq \mathbb{R}^d$, $d \geq 1$.
  (Notăm f_θ sau p_θ.)
  - Observând/ măsurând caracteristica indivizilor dintr-un eșantion, se obține un șir de valori $x_1, x_2, \dots, x_n$, interpretate ca valori de observație asupra variabilei aleatoare X.
  - Din acestea "se estimează" parametrii de interes: media caracteristicii investigate, dispersia sau parametrii necunoscuți (i.e. θ), de care depinde distribuția de probabilitate.
- valorile înregistrate, $x_1, x_2, \dots, x_n$, sunt interpretate ca valori de observație asupra unui șir de variabile aleatoare $X_1, X_2, \dots, X_n$, independente și identic distribuite ca variabila aleatoare X

Model statistic

Definitie

Considerăm $\mathcal{P}$ o populație supusă investigării statistice din punctul de vedere al unei caracteristici X, ce ia valori discrete sau continue.
Perechea (X, f_θ) se numește **model statistic**, unde

- dacă X este continuă, aunci X are densitate f_θ;
- dacă X este discretă, atunci $f_\theta(x) = p_\theta(x) = P(X = x)$.

- populația=un tip de chip, caracteristica investigată X = durata de viață(in teorie avem : durata de viață a dispozitivelor/ circuitelor este exponențial distribuită);
- modelul statistic: (X, f_θ) cu $X \sim \text{Exp}(\theta)$, θ - parametru necunoscut.
- din înregistrarea duratei de viață a n chipuri selectate la întâmplare din producția dintr-o anumită perioadă, se va estima parametrul θ
- θ- media v.a. X a modelului exponențial, se poate folosi pentru ca firma producătoare să stabilească garanția pentru acest tip de chip-uri.

Definiție

Fie (X, f_θ) un model statistic asociat unei populații.
Un vector aleator $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \ldots \mathbf{X_n})$, ale cărui coordonate sunt independente și identic distribuite după legea modelului f, se numește **selecție aleatoare** de volum n.
În urma investigării prin sondaj a populației, se înregistrează n valori numerice $(x_1, x_2, \ldots, x_n)$, numite *valori de selecție* sau **valori de realizare** a selecției aleatoare $\mathbf{X}$.

- **Statistică**="orice ce poate fi calculat din datele colectate"
  $Y = T(X_1, X_2, \ldots, X_n)$- o funcție reală continuă de variabile $X_1, X_2, \ldots, X_n$;
- Distribuția de probabilitate a statisticii Y se numește **distribuția de selecție** a statisticii (poate fi dedusă sau aproximată)

Având valorile de selecție $x_1, \ldots, x_n$, primele informații ce se extrag sunt:

- **media de selecție** sau **media experimentală**, notată cu $\bar{x}$:
  $$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$
  Este o realizare a statisticii $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$,
- **dispersia de selecție** (**dispersia experimentală**), s^2, definită prin:
  $$s^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2.$$
  Este o realizare a statisticii $S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - X)^2$.
- **abaterea standard** a eșantionului=$\sqrt{s^2}$ (se notează cu s).

Estimatori

Ideea de bază = folosirea unei singure valori calculată din datele coleționate, de exemplu media/dispersia de selecție

Fie (X, f_θ) un model statistic și $(x_1, x_2, \ldots, x_n)$ o realizare a unei selecții aleatoare de volum n, $(X_1, X_2, \ldots, X_n)$
Estimator punctual al parametrului θ este o funcție $\hat{\theta}(x_1, x_2, \ldots, x_n)$ (este o realizare a variabilei aleatoare $\hat{\theta}(X_1, X_2, \ldots, X_n)$).
Există o infinitate de estimatori ai parametrului θ, deci alegem estimatorii care să aproximeze θ cu o probabilitate suficient de mare.

Estimatorul $\hat{\theta}(x_1, x_2, \ldots, x_n)$ cu proprietatea că pentru orice $\varepsilon > 0$ are loc
$$\lim_{n\to\infty} P(|\,\hat{\theta}(X_1, X_2, \ldots, X_n) - \theta| > \varepsilon) = 0$$
se numește **estimator consistent** al parametrului θ.

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
$$M(\hat{\theta}(X_1, X_2, \ldots, X_n)) = \theta$$
se numește **estimator centrat** sau **nedeplasat**.

Fie $\hat{\theta}_1, \hat{\theta}_2$ doi estimatori nedeplasați ai parametrului θ. Dacă între dispersiile statisticilor $\hat{\theta}_1(X_1, X_2, \ldots, X_n)$ și $\hat{\theta}_2(X_1, X_2, \ldots, X_n)$ are loc
$$\sigma^2(\hat{\theta}_1(X_1, X_2, \ldots, X_n)) \leq \sigma^2(\hat{\theta}_2(X_1, X_2, \ldots, X_n)),$$
atunci estimatorul $\hat{\theta}_1$ se zice că este **mai eficient** decât estimatorul $\hat{\theta}_2$.

Estimarea mediei

Fie (X, f) un model statistic continuu sau discret și $x_1, x_2, \dots, x_n$ observații independente din legea f.

Proprietate

Media de selecție a observațiilor $x_1, x_2, \dots, x_n$,

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n},$$

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

Cererea de memorie pentru o aplicație, ca proporție din memoria ce poate fi alocată de un utilizator, este o variabilă aleatoare X ce are densitatea de probabilitate

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

a) Să se determine media teoretică $M(X)$ a variabilei aleatoare X și apoi să se estimeze θ în funcție de media de selecție $\bar{x}$ a unei selecții aleatoare de volum n.
b) Să se determine un estimator al parametrului θ din selecția 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 aplicației cu diferite date de intrare.
Rezolvare:

a) Mai întâi calculăm media teoretică:

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

Dacă $m = M(X)$ și $\bar{x}$ este media de selecție a unui eșantion de valori $x_1, x_2, \dots, x_n$, atunci din egalitatea impusă $\hat{m} = \bar{x}$ se determină un estimator al parametrului θ:

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

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

Estimarea dispersiei

Proprietate

Dacă (X, f) este un model statistic și m, σ^2 sunt media și dispersia variabilei aleatoare X, atunci dispersia valorilor de selecție $x_1, x_2, \dots, x_n$ din legea de probabilitate definită de f,

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

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

Observație: Media statisticii S^2 este $M(S^2) = \sigma^2$, unde σ^2 este dispersia legii de probabilitate a modelului statistic.

Estimatorul verosimilității maxime

Cazul de studiu: legea de probabilitate a modelului statistic este cunoscută, adică se cunoaște densitatea de probabilitate, dar aceasta depinde de unul sau mai mulți parametrii necunoscuți= **distribuții parametrice**: exponențială, binomială, normală.

Există diverse metode de estimare a acestor parametrii, **metoda verosimiltății maxime="maximum likelihood estimates(MLE)"** răspunde la întrebarea:

Pentru care valoare a parametrului valorile observate au cea mai mare probabilitate sa fie observate?

- este o metodă de estimare a unui parametru ("point estimates");
- se poate folosi cu ușurință prin aplicarea unui algoritm de determinare a **"estimatorului de verosimilitate maximă"**;

- studiem o caracteristică a unei populații, măsurată prin v.a. X cu densitate de probabilitate f_θ depinde de un parametru necunoscut θ.
- Fie un eșantion de volum n din populația respectivă, cu valorile înregistrate $x_1, x_2, \dots, x_n$.
- **se determină un estimator pentru parametrul θ, care maximizează probabilitatea înregistrării unor valori ale caracteristicii X foarte apropiate de velorile x_i**
- probabilitatea ca X să ia valori apropiate de x_i este:

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

- Not and cu $X_1, X_2, \dots, X_n$ v.a. i.i.d. ca X, avem:

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

- h^n nu depinde de θ, deci parametrul θ ce maximizează probabilitatea

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

  este parametrul ce maximizează produsul $f_\theta(x_1) f_\theta(x_2) \cdots f_\theta(x_n)$.
- Funcția $L : \mathbb{R} \to \mathbb{R}$, de variabilă θ, asociată eșantionului $x_1, x_2, \dots, x_n$:

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

  se numește **funcția de verosimilitate** (este o functie de o singură variabilă, și anume θ).
- **estimatorul verosimilității maxime** a parametrului θ este

$$= \operatorname{argmax}(L(\theta; x_1, x_2, \dots, x_n)),$$

  unde prin $\operatorname{argmax}(L(\theta; x_1, x_2, \dots, x_n))$ se înțelege argumentul θ care maximizează funcția L.

Fie populația $\mathcal{P}$ formată dintr-un tip de circuite. Caracteristica ce dorim să o investigăm prin sondaj statistic este durata de viață a acestor circuite, știind că aceasta este exponențial distribuită, cu parametrul θ necunoscut. Măsurând timpul de viață (în ani) a 10 circuite, se obțin valorile:

1.8, 3.3, 0.9, 0.1, 2.9 3.5, 1.1, 2.8, 2.7, 3.3.

Să se determine estimatorul de verosimilitate maximă pentru θ (adică pentru media duratei de viață a acestui tip de circuite).

- Densitatea de probabilitate a distribuției exponențiale este $f_\theta(x) = \begin{cases} 0, & \text{dacă } x < 0, \\ \frac{1}{\theta} e^{-x/\theta}, & \text{dacă } x \geq 0. \end{cases}$
- funcția de verosimilitate este

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

- Pentru simplitatea calculelor, vom determina punctul de maxim absolut (dacă acesta există) pentru $\ln(L)$ și acesta va fi punct de maxim absolut și pentru L:

$$l(\theta) = \ln L(\theta; x_1, x_2, \ldots, x_n) = -n \ln \theta - \frac{\sum\limits_{i=1}^{n} x_i}{\theta}.$$

- Avem $l'(\theta) = -\frac{n}{\theta} + \frac{\sum\limits_{i=1}^{n} x_i}{\theta^2}$.

- Rezolvând ecuația $l'(\theta) = 0$ în raport cu θ, obținem punctul

$$\hat{\theta} = \frac{\sum\limits_{i=1}^{n} x_i}{n} = \bar{x}$$

(este maxim absolut ($l''(\bar{x}) < 0$) pentru $l = \ln(L)$, deci și pentru L).
- În concluzie,

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

estimatorul de verosimilitate maximă a parametrului θ a distribuției exponențiale este media de selecție.
- În cazul exemplului dat, estimatorul verosimilității maxime a mediei de viață a circuitelor este media selecției:

$$\hat{\theta} = (x_1 + x_2 + \ldots + x_{10})/10 \approx 2.24$$

Teorema limită centrală

TLC afirmă că "în medie totul este normal"

https://www.albany.edu/~jr853689/CentralLimitTheoremForDice.htm
https://www.youtube.com/watch?v=eqxbc7mQpTs

Teorema limită centrală

Se consideră (X_n) este un șir de variabile aleatoare i.i.d.
- media comună m
- abaterea standard σ
- $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ șirul variabilelor medie aritmetică

Atunci, pentru $n \to \infty$: $\bar{X}_n \sim ApN(m, D^2 = \sigma^2/n)$
(distribuția de probabilitate a variabilelor $\bar{X}_n$ este aproximativ normală de medie m și dispersie $D^2 = \sigma^2/n$).

În practica statistică: pentru $n \geq 30$, distribuția normală poate fi folosită ca distribuție a mediei aritmetice a n v.a. i.i.d. cu media și dispersia finită.

Teorema limită centrală prezintă interes și în următorul context:
- șirului de variabile aleatoare i.i.d. (X_n) îi asociem șirul (S_n)

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

- Evident, $S_n = n\bar{X}_n$, $M(S_n) = M(n\bar{X}_n) = nM(X_n) = nm$,
- $$\sigma^2(S_n) = \sigma^2(n\bar{X}_n) = n^2\sigma^2(\bar{X}_n) = n^2\sigma^2/n = n\sigma^2.$$

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

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