Statistică - Cursul 9

Olariu E. Florentin

Aprilie, 2016

Table of contents

Momentele

Definition 0.1

Momentul central de ordin k *al populaţiei este*

$$\mu_k = M\left[(X-\mu)^k\right].$$

Momentul central de ordin k *al eşantionului X este*

$$m_k = \frac{\sum_{i=1}^{n}(x_i-\overline{x}_n)^k}{n}.$$

- Momentul central al eşantionului, m_k, este statistica corespunzătoare momentului central μ_k.

Forma distribuţiei

Definition 0.2

Asimetria[a] *(***skewness***) al unei distribuţii X este*

$$\mu_k = \frac{M\left[(X-\mu)^3\right]}{\left(M\left[(X-\mu)^2\right]\right)^{3/2}}.$$

Asimetria eşantionului *este* $a_3 = \dfrac{m_3}{s^3}.$

[a] Sau coeficientul lui Pearson de asimetrie.

- a_3 este o măsură a asimetriei: dacă $a_3 = 0$ avem o distribuţie simetrică, dacă $a_3 < 0$, distribuţia prezintă o asimetrie la stânga (modul este deplasat către dreapta), iar dacă $a_3 > 0$, distribuţia prezintă o asimetrie la dreapta (modul este deplasat către stânga).

Forma distribuţiei

Definition 0.3

Turtirea (**kurtosis**) *unei distribuţii* X *este*

$$\mu_k = \frac{M\left[(X-\mu)^4\right]}{\left(M\left[(X-\mu)^2\right]\right)^2}.$$

Turtirea eşantionului *este* $a_4 = \dfrac{m_4}{s^4}$.

- Turtirea măsoară înălţimea distribuţiei: dacă $a_4 = 3$ avem o distribuţie mesokurtică (e. g. normal law), dacă $a_4 < 3$, distribuţia este platykurtică (comparată cu legea normală este mai degrabă "scundă"), iar dacă $a_4 > 3$, distribuţia este leptokurtică (este "înaltă" prin comparaţie cu legea normală).

Inegalităţile lui Markov şi a lui Cebâşev

Proposition 1.1

Fie $X \geqslant 0$ *o variabilă aleatoare. Dacă* $a > 0$, *atunci*

$$P(X \geqslant a) \leqslant \frac{M[X]}{a}.$$

proof:

$$M[X] = \int_0^{+\infty} tf(t)\,dt = \int_0^a tf(t)\,dt + \int_a^{+\infty} tf(t)\,dt \geqslant$$

$$\int_a^{+\infty} tf(t)\,dt \geqslant a\int_a^{+\infty} f(t)\,dt = aP(X \geqslant a).$$

∎

Inegalitatea lui Cebâşev

Proposition 1.2

Fie X o variabilă aleatoare cu media μ şi dispersia σ^2. Atunci

$$P(|X-\mu| \geqslant k) \leqslant \frac{\sigma^2}{k^2}.$$

proof: Considerăm variabila $Y = (X-\mu)^2$ şi $a = k^2$ în inegalitatea lui Markov

$$P(|X-\mu| \geqslant k) = P\left[(X-\mu)^2 \geqslant k^2\right] \leqslant \frac{\mathbb{E}\left[(X-\mu)^2\right]}{k^2} = \frac{\sigma^2}{k^2}.$$

■

Teorema lui Cebâşev

Theorem 1.1

Fie $(X_n)_{n\geqslant 1}$ un şir de variabile aleatoare independente având dispersii finite, uniform mărginite, i. e. $D^2[X_n] \leqslant c$, pentru orice $n \geqslant 1$. Atunci

$$\lim_{n\to\infty}\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i - \frac{1}{n}\sum_{i=1}^{n}M[X_i]\right| < \epsilon\right) = 1.$$

proof: Ştim că

$$M\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right] = \frac{1}{n}\sum_{i=1}^{n}M[X_i] \text{ şi}$$

$$D^2\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}D^2[X_i] < \frac{c}{n}.$$

Teorema lui Cebâşev

Folosind inegalitatea lui Cebâşev pentru variabila $\frac{1}{n}\sum_{i=1}^{n} X_i$ obţinem

$$1 \geqslant \left(\left| \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} M[X_i] \right| < \epsilon \right) \geqslant 1 - \frac{D^2\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]}{\epsilon^2} \geqslant 1 - \frac{c}{n\epsilon^2}.$$

Trecând la limită,

$$\lim_{n\to\infty} \left(\left| \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} M[X_i] \right| < \epsilon \right) = 1. \blacksquare$$

Legea numerelor mari - varianta slabă

- Legea numerelor mari spune că pe măsură ce creşte numărul de variabile independente, identic distribuite, media lor de selecţie se apropie de media lor comună.

Theorem 1.2

(Legea slabă numerelor mari , legea lui Khintchine) Fie $(X_n)_{n\geqslant 1}$ *un şir de variabile aleatoare independente şi identic distribuite cu media* μ *şi dispersia* σ^2*. Atunci*

$$\lim_{n\to\infty}\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu]\right|<\epsilon\right)=1 \ \textit{sau} \ \lim_{n\to\infty}\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu]\right|\geqslant\epsilon\right)=0.$$

proof: O consecinţă a teoremei anterioare, deoarece $\frac{1}{n}\sum_{i=1}^{n}M[X_i]=\mu$. ■

Legea numerelor mari - varianta tare

Theorem 1.3

(Legea tare numerelor mari) Fie $(X_n)_{n\geqslant 1}$ un şir de variabile aleatoare independente şi identic distribuite cu media μ şi dispersia σ^2. Atunci

$$P\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}X_i=\mu\right)=1.$$

proof: Fiind mai complicată este omisă.

Un exemplu cu frecvenţe

- Bernoulli este primul care a demonstrat legea slabă numerelor mari dar doar pentru distribuţii Bernoulli.
- Să presupunem că avem o experienţă aleatoare şi un eveniment aleator asociat A cu $P(A) = p$.
- Repetăm în mod independent experienţa şi considerăm următorul şir de variabile aleatoare : $X_i = 1$ dacă A se produce la a i-a repetare şi 0 altfel.
- Variabilele sunt independente şi distribuite Bernoulli cu media p.

Un exemplu cu frecvenţe

- The law of large numbers says that, with probability 1,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to p.$$

- $\sum_{i=1}^{n} X_i$ is the number of occurrences of A after n performings.
- In other words the law of large numbers says that the A occurs with frequency p.

Istorie

- James Bernoulli A demonstrat legea slabă a numerelor mari în 1700; Poisson i-a generalizat rezultatul în 1800.
- Cebâşev a descoperit inegalitatea care-i poartă numele în 1866, iar Markov a extins rezultatul lui Bernoulli la variabile aleatoare dependente.
- În 1909 the Émile Borel a demonstrat teorema care astăzi este cunoscută sub numele de legea tare a numerelor mari (care generalizează o dată în plus teorema lui Bernoulli).
- În 1926 Kolmogorov a obţinut o condiţie mai generală, suficientă pentru ca un şir de variabile aleatoare independente să respecte legea numerelor mari. Condiţia este

$$\sum_{n \geqslant 1} \frac{D^2[X_n]}{n^2} < +\infty.$$

Teorema limită centrală

Theorem 2.1

(Teorema limită centrală, Lindeberg-Lévy) Fie $(X_n)_{n \geqslant 1}$ *un şir de variabile aleatoare independenteşi identic distribuite cu media* μ *şi dispersia* σ^2. *Atunci*

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\frac{\sigma}{\sqrt{n}}} \to N(0,1) \ sau$$

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \leqslant a\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} \exp -t^2/2 dt.$$

Teorema limită centrală

- Teorema limită centrală permite estimarea unor probabilităţi asociate sumelor de variabile (independente şi identic distribuite).
- Pe de altă parte, teorema explică de ce atât de multe procese (din ştiinţele sociale, biologie, psihologie etc) urmează o lege normală.
- în esenţa ei teorema limită centrală spune că, pentru eşantioane suficient de mari ($n \geqslant 30$), variabila

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma/\sqrt{n}}$$

urmează o lege normală standard, $N(0,1)$.
- Teorema limită centrală are lco chiar şi pentru variabile dependente, dacă au corelaţia foarte scăzută.

Teorema luiBernoulli

Proposition 2.1

Fie α_n numărul de apariţii ale unui eveniment A în n repetări independente ale unei experienţe aleatoarwe gee. Dacă $f_n = \dfrac{\alpha_n}{n}$ este frecvenţa relativă a apariţiei lui A, atunci şirul $(f_n)_{n\geqslant 1}$ converge în probabilitate la $p = P(A)$.

proof: $\alpha_n = nf_n$ este o variabilă distribuită binomial, astfel $M[\alpha_n] = np$ şi $D^2[\alpha_n] = np(1-p)$. Mai mult,

$$P(|f_n - p| < \epsilon) = P(|\alpha_n - np| < n\epsilon) = P(|\alpha_n - \mathbb{E}[\alpha_n]| < n\epsilon) \geqslant$$

$$\geqslant 1 - \frac{p(1-p)}{n\epsilon^2}.$$

Evident, trecând la limită, $\lim\limits_{n\to\infty} P(|f_n - p| < \epsilon) = 1$, pentru orice $\epsilon > 0$.

■

Aproximarea normală a distribuţiei binomiale

- Fie X_n un şir de variabile Bernoulli(p) independente.
- $X = \sum_{i=1}^{n} X_i$ are o distribuţie binomială, $B(n, p)$.
- Folosind teorema limită centrală obţinem teorema de Moivre-Laplace care spune că, pentru n suficient de mare, variabila

$$Y = \frac{X - \mathbb{E}[X]}{\sqrt{Var[X]}} = \frac{X - np}{\sqrt{np(1-p)}}$$

este o variabilă normală standard $(N(0, 1))$.
- Aproximarea este bună pentru $np(1 - p) \geqslant 10$.

Aproximarea normală a distribuţiei binomiale

- Un alt mod de a vedea acest rezultat este următorul: când k este aproape de np

$$\binom{n}{k} p^k (1-p)^{n-k} \sim \frac{\exp -\frac{(k-np)^2}{2np(1-p)}}{\sqrt{2\pi np(1-p)}}.$$

- Considerăm următorul exemplu: fie X numărul de apariţii ale stemei în 40 de aruncări ale unei monede.
- Cât este $P(X = 20)$?

$$P(X = 20) = P(19.5 \leqslant X \leqslant 20.5) =$$

$$= P\left(\frac{19.5-20}{\sqrt{10}} \leqslant \frac{X-20}{\sqrt{10}} \leqslant \frac{20.5-20}{\sqrt{10}}\right) = P\left(-0.16 \leqslant \frac{X-20}{\sqrt{10}} \leqslant 0.16\right)$$

$$\sim \Phi(0.16) - \Phi(-0.16) = 0.1272,$$

unde $\Phi(\cdot)$ este funcţia de repartiţie a variabilei $N(0,1)$.

Corecţia continuă

- Corecţia continuă este o ajustare care se face ori de câte ori o distribuţie discretă este aproximată printr-una continuă.
- $P(X = 10) = P(9.5 \leqslant X \leqslant 10.5)$, $P(X > 15) = P(X \geqslant 15.5)$, $P(X < 13) = P(X \leqslant 12.5)$.

Statistică inferenţială

- Statistica inferenţială are scopul să trage concluzii relativ la o populaţie folosind rezultate ale teoriei probabilităţilor şi statistici obţinute din eşantionane.
- Fără utilizarea teoriei probabilităţilor (legea numerelor mari, teorema limită centrală ş. a.) am putea considera drept sistematic un comportament care este de fapt datorat hazardului sau, din contră, un comportament sistematic ar putea trece neobservat.
- Exemple de inferenţe statistice: *intervale de încredere* pentru estimarea parametrilor sau *teste de semnificaţie*.
- Tehnicile de inferenţă au la bază distribuţiile eşantioanelor.

Estimarea parametrilor

- Distribuţia unei anumite populaţii poate fi necunoscută în întregime; de aceea ne putem dori să aflăm cel puţin media, dispersia sau alţi parametri ai ei.
- Aceşti parametri ai unei populaţii pot fi estimaţi folosind statistici calculate din eşantion.
- Există două tipuri de estimare: *estimare punctuală* şi *estimare cu intervale* .

Proposition 1.1

Estimarea punctuală a unui parametru *constă în determinarea unui număr, de obicei valoarea unei statistici corespunzătoare, desemnat sa estimeze acel parametru.*

Estimarea parametrilor

Proposition 1.2

Estimarea cu un interval a unui parametru *constă în determinarea unui interval ale cărui limite sunt statistici calculate din eşantioane.*

Nivelul de încredere $(1-\alpha)$ *este proporţia acelor intervale (de încredere) care conţin parametrul estimat.*

Un **interval de încredere** *este un interval care are un anumit nivelul de încredere (prescris).*

Estimarea punctuală

- Exemple de estimatori punctuali sunt media de selecţie $\overline{x}_n$, dispersia eşantionului s^2, sau deviaţia standard a eşantionului s.
- Pentr un parametru dat putem avea mai mulţi estimatori punctuali: media populaţiee poate fi estimată prin media de selecţie, mediană, mod.
- Se pot ridica anumite întrebări legate de calitatea estimatorilor punctuali.
- Cât de exact este un estimator (i.e., the *încrederea*) - este în mod frecvent mai mare (*supra-estimator*) sau mai mic (*sub-estimator*) în raport cu parametrul estimat?
- Din acest punct de vedere se preferă estimatorii *nedeplasaţi*.
- Care este variabilitatea unui estimator punctual (văzut ca o variabilă aleatoare) - aceasta este *acurateţea*.

Caracteristici ale estimării punctuale

- De exemplu dispersia mediei de selec tie, care este $\sigma/\sqrt{n}$: cu cât este mai mare eşantionul cu atât mai mică este variabilitatea mediei de selecţie.

Definition 1.1

Fie θ un anumit parametru al unei populaţii, x un eşantion al acestei populaţii şi $\hat{\theta}_n = \hat{\theta}_n(x)$ un estimator punctual al lui θ.
$\hat{\theta}_n$ este o **statistică nedeplasată** *dacă $M[\hat{\theta}_n] = \theta$. Altfel $\hat{\theta}_n$ este numită statistică* **deplasată**.
$\hat{\theta}_n$ este o **statistică consistentă** *dacă, pentru orice $\epsilon > 0$*

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1.$$

Un estimator nedeplasat cu dispersie minimă, dacă există, se numeşte **statistică eficientă**.

Distribuţia mediei de selecţie

- Fie $\mathcal{P}$ o populaţie formată din N indivizi ale căror valori ale atributului sunt $a_1, a_2, \ldots, a_N$.
- Variabila care reprezintă atributul populaţiei este notată cu X.
- Pentru această populaţie, deci şi pentru X, media şi dispersia sunt

$$\frac{1}{N}\sum_{i=1}^{N} a_i = \mu, \text{ respectiv } \frac{1}{N}\sum_{i=1}^{N}(a_i - \mu)^2 = \sigma^2.$$

- Pentru a estima μ şi σ^2folosim eşantioane de dimensiune n.

Distribuţia mediei de selecţie

- Pentru un eşantion dat, $x^{(k)} = \left(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}\right)$, media de selecţie este

$$\overline{x}_n^{(k)} = \frac{1}{n}\sum_{j=1}^{n} x_j^{(k)}.$$

- Fiecare $x_j^{(k)}$ este o valoare a unei variabile aleatoare cu aceeaşi distribuţie ca X.

Definition 1.2

Variabila aleatoare care are drept valori toate mediile de selecţie posibile, $\overline{x}_n^{(k)}$, pentru eşantioane de dimensiune n, se numeşte **distribuţia mediei de selecţie**.

Distribuţia mediei de selecţie

- Se poate demonstra următorul rezultat

Theorem 1.1

Media şi dispersia mediei de selecţie, $\overline{x}_n$, sunt μ şi σ^2/n:

$$M[\overline{x}_n] = \mu, D^2[\overline{x}_n] = \frac{\sigma^2}{n}.$$

În plus, pentru valori mari ale lui n ($\geqslant 30$), distribuţia mediei de selecţie este normală, i. e.

$$\overline{x}_n \sim N(\mu, \sigma^2/n).$$

Estimarea mediei şi a dispersiei

- Bineînţeles, un estimator al mediei adevărate a populaţiei (μ) este media de selecţie, $\overline{x}_n$.
- Un estimator pentru adevărata dispersie a populaţiei σ^2) este dispersia eşantionului s^2 - o statistică nedeplasată.
- Pentru deviaţia standard a mediei de selecţie, $\sigma/\sqrt{n}$ numită the **eroarea standard a mediei** (standard error of the mean - SEM), un estimator este $s/\sqrt{n}$.

Estimarea cu intervale

- Un estimator de tip interval al mediei populaţiei poate fi obţinut cu inegalitatea lui Cebâşev:

$$P(|\overline{x}_n - M[\overline{x}_n]| \geqslant k \cdot Var[\overline{x}_n]) \leqslant \frac{1}{k^2} \Leftrightarrow$$

$$P\left(|\overline{x}_n - \mu| \geqslant k \cdot \frac{\sigma}{\sqrt{n}}\right) \leqslant \frac{1}{k^2} \Leftrightarrow$$

$$P\left(|\overline{x}_n - \mu| < k \cdot \frac{\sigma}{\sqrt{n}}\right) \geqslant 1 - \frac{1}{k^2} \Leftrightarrow$$

$$P\left(\overline{x}_n - k \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x}_n + k \cdot \frac{\sigma^2}{n}\right) \geqslant 1 - \frac{1}{k^2}.$$

- Astfel o estimare cu interval a lui μ este

$$\left(\overline{x}_n - k \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x}_n + k \cdot \frac{\sigma^2}{n}\right).$$

Intervale de încredere

Definition 1.3

Un interval de încredere pentru un parametru θ cu $(1-\alpha)$ nivel de încredere este definit cu ajutorul a două statistici L şi U astfel ca

$$P(L \leqslant \theta \leqslant U) = (\geqslant) 1 - \alpha.$$

- L şi U sunt de fapt variabile aleatoare ale căror valori sunt statisticia: pentru diferite eşantioane au diferite valori.
- De obicei nivelul de încredere este o probabilitate aproape de 1, cum ar fi 0.90, 0.95,sau 0.99 (care dau $\alpha \in \{0.10, 0.05, 0.01\}$).

Intervale de încredere pentru medie

- Să presupunem că avem un eşantion de dimensiune n şi un nivel de încredere $(1-\alpha)$ şi dorim să construim un interval de încredere pentru media μ.
- Ştim că media de selecţie, $\overline{x}_n$, urmează o distribuţie normală

$$N\left(\mu, \frac{\sigma^2}{n}\right).$$

- Prin standardizare variabila $Z = \dfrac{\overline{x}_n - \mu}{\sigma/\sqrt{n}}$ este distribuită $N(0,1)$.
- Căutăm o valoare z^*, numită *valoare critică*, astfel ca o variabilă normal standard să acopere sub grafic o arie de $(1-\alpha)$ pe intervalul centrat în medie şi de lungime $2z^*$ deviaţii standard.

Intervale de încredere pentru medie

- Fie $Z : N(0,1)$, valoarea critică este aleasă aşa încât
$$P(-z^* \leqslant Z \leqslant z^*) = 1 - \alpha.$$
- Definiţii echivalente ale lui z^*:
$$P(Z \leqslant -z^*) = \alpha/2 \text{ sau } P(Z \geqslant z^*) = \alpha/2.$$
- Notâm cu $\Phi(a) = P(Z \leqslant a)$, funcţia de repartiţie normală standard.
- Astfel $z^* = -\Phi^{-1}(\alpha/2)$ - valoare care poate fi găsită în tabele sau aproximată în pachetele de prelucrare statistică uzuale (R, MiniTab, SPSS etc).
- Odată ce am determinat valoarea critică, ştim că
$$P\left(-z^* \leqslant \frac{\overline{x}_n - \mu}{\sigma/\sqrt{n}} \leqslant z^*\right) = 1 - \alpha.$$

Intervale de încredere pentru medie

- Echiuivalent

$$P\left(\overline{x}_n - z^*\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \overline{x}_n + z^*\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

- Am demonstrat astfel

Theorem 1.2

Un interval de încredere cu nivelul de încredere $(1-\alpha)$ *pentru pentru media unei populaţii cu media cunsocută este*

$$\left(\overline{x}_n - z^*\frac{\sigma}{\sqrt{n}}, \overline{x}_n + z^*\frac{\sigma}{\sqrt{n}}\right),$$

unde z^* *este valoarea critică asociată cu* $\alpha/2$. *Mai mult, acest interval este exact pentru o populaţie distribuită normal şi aproximativ altfel, când eşantionul este suficient de mare* ($n \geqslant 30$).

Intervale de încredere pentru medie

Nivel de încredere	$\alpha/2$	z^*
90%	0.05	1.645
95%	0.025	1.960
99%	0.005	2.576

- $z^* \dfrac{\sigma}{\sqrt{n}}$ se numeşte *eroarea marginală*.
- Lungimea unui interval de încredere pentru medie este $2z^* \dfrac{\sigma}{\sqrt{n}}$.
- Dacă dorim o lungime anume pentru acest interval, w, atunci ne trebuie un eşantion de dimensiune $n = \dfrac{(2z^*\sigma)^2}{w^2}$. În practică această valoare poate fi nerealistă (dacă n este prea mare).
- Pe măsură ce n creşte lungimea intervalului (sau eroarea marginală) scade, ceea ce este util, dar poate fi nepractic.

Intervale de încredere pentru medie

- Să ne amintim că $\frac{\sigma}{\sqrt{n}}$ este eroarea standard a mediei. Un intervale de încredere poate fi văzut astfel

$$estimatedMean \pm z^* estimatedStdDev$$

- *Exemplu.* Un anumit medicament este analizat măsurându-i-se substanţa activă de trei ori, rezultatele sunt 0.8403, 0.8636, şi 0.8447 g/l. Se ştie că această concentraţie urmează urmează o lege normală cu deviaţia standard $\sigma = 0.0068$ g/l. Să se determine un interval de încredere de 99% pentru adevărata medie a concentraţiei, μ.
  Soluţie:

$$\overline{x}_3 = 0.8404, \alpha = 0.01, \alpha/2 = 0.005, z^* = 2.576, z^* \frac{\sigma}{\sqrt{n}} = 0.0101$$

Intervalul de încredere este $(0.8303, 0.8505)$.

Sfârşit

Bibliography

- Freedman, D., R. Pisani, R. Purves, *Statistics*, W. W. Norton & Company, 4th edition, 2007.
- Johnson, R., P. Kuby, *Elementary Statistics*, Brooks/Cole, Cengage Learning, 11th edition, 2012.
- Shao, J., *Mathematical Statistics*, Springer Verlag, 1998.
- Spiegel, M. R., L. J. Stephens, *Theory and Problems of Statistics*, Schaum's Outline Series, McGraw Hill, 3rd edition, 1999.