УНИВЕРЗИТЕТ “СВ. КИРИЛ И МЕТОДИЈ”– СКОПЈЕ
Природно-математички Факултет
Институт за математика

Ирена Стојковска

ОСНОВИ НА СТАТИСТИКА

Предавања

Скопје, 2013

Содржина

1

Елементи од теорија на веројатност

1.1 Случајни променливи

Нека $(\Omega, \mathcal{F}, P)$ е простор на веројатност.

- **Случајна променлива** дефинирана на $(\Omega, \mathcal{F}, P)$ е реално вредносна функција $\xi : \Omega \to \mathbb{R}$ која е мерлива во однос на $\mathcal{F}$ и $\mathcal{B}$ т.е. за сите Борелови множества $B \in \mathcal{B}$ важи $\xi^{-1}(B) = \{w : \xi(w) \in \mathcal{B}\} \in \mathcal{F}$.

- Функцијата $P_\xi : \mathcal{B} \to \mathbb{R}$ дефинирана со $P_\xi(B) = P\{w : \xi(w) \in B\} = P(\xi^{-1}(B))$ се нарекува **закон на случајната променлива** ξ.

- Функцијата $F_\xi : \mathbb{R} \to \mathbb{R}$ дефинирана со

  $$F_\xi(x) = P_\xi((-\infty, x]) = P\{w : \xi(w) \leq x\} = P(\xi \leq x)$$

  се нарекува **функција на распределба** на случајната променлива ξ.

  Основни својства: F_ξ е непрекината од десно, F_ξ е неопаѓачка функција, $\lim_{x\to-\infty} F_\xi(x) = 0$ и $\lim_{x\to\infty} F_\xi(x) = 1$.

- Случајната променлива ξ е **дискретна** ако множеството од сите можни вредности на ξ, односно $\xi(\Omega)$, е конечно или пребројиво.

  Случајната променлива I_A дефинирана со

  $$I_A(w) = \begin{cases} 1, & w \in A \\ 0, & w \notin A \end{cases},$$

  е пример за дискретна случајна променлива со $I_A(\Omega) = \{0, 1\}$ и се нарекува **индикатор на настанот** $A \in \mathcal{F}$.

- Случајната променлива ξ е **апсолутно непрекината** ако нејзината функција на распределба е апсолутно непрекината функција т.е. постои ненегативна функција $p_\xi : \mathbb{R} \to \mathbb{R}$ така да

$$F_\xi(x) = \int_{-\infty}^{x} p_\xi(u)du$$

за сите $x \in \mathbb{R}$. Функцијата p_ξ се нарекува **густина на распределба** на ξ.

Основни својства: $p_\xi(x) \geq 0$, $\int_{-\infty}^{\infty} p_\xi(x)dx = 1$, и $P\{\xi \in B\} = \int_B p_\xi(x)dx$, за произволно Борелово множество $B \in \mathcal{B}$.

Теорема 1.1. *Нека $\xi_1, \xi_2, ..., \xi_n$ се случајни променливи дефинирани на просторот на веројатност $(\Omega, \mathcal{F}, P)$ и $f : \mathbb{R}^n \to \mathbb{R}$ е Борелова функција. Тогаш секоја функција $f(\xi_1, \xi_2, ..., \xi_n) : \Omega \to \mathbb{R}$ е исто така случајна променлива.*

* * *

Нека $(\Omega, \mathcal{F}, P)$ е простор на веројатност и нека $n > 1$ е цел број.

- Функцијата $\zeta : \Omega \to \mathbb{R}^n$ се нарекува **повеќедимензионална случајна променлива** или **случаен вектор** ако за сите Борелови множества $B \in \mathcal{B}^n$ важи $\zeta^{-1}(B) = \{w : \zeta(w) \in B\} \in \mathcal{F}$.

Покомпонентно, $\zeta(w) = (\zeta_1(w), \zeta_2(w), ..., \zeta_n(w))$, $w \in \Omega$, каде $\zeta_i : \Omega \to \mathbb{R}$, $i = 1, 2, ..., n$. Се покажува дека $\zeta_1, \zeta_2, ..., \zeta_n$ се случајни променливи ако и само ако $\zeta = (\zeta_1, \zeta_2, ..., \zeta_n)$ е n-димензионален случаен вектор.

- Функцијата $P_\zeta : \mathcal{B}^n \to \mathbb{R}$ дефинирана со $P_\zeta(B) = P\{w : \zeta(w) \in B\} = P(\zeta^{-1}(B))$ се нарекува **закон на случајниот вектор** ζ.

- Функцијата $F_\zeta : \mathbb{R}^n \to \mathbb{R}$ дефинирана со

$$F_\zeta(x_1, x_2, ..., x_n) = P(\zeta_1 \leq x_1, \zeta_2 \leq x_2, ..., \zeta_n \leq x_n)$$

се нарекува **функција на распределба** на $\zeta = (\zeta_1, \zeta_2, ..., \zeta_n)$.

- Случајниот вектор $\zeta = (\zeta_1, \zeta_2, ..., \zeta_n)$ е од **апсолутно непрекинат тип**, ако F_ζ е апсолутно непрекината т.е. постои ненегативна функција $p_\zeta : \mathbb{R}^n \to \mathbb{R}$ така да

$$F_\zeta(x_1, x_2, ..., x_n) = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} p_\zeta(u_1, ..., u_n)du_1...du_n$$

за сите $(x_1, x_2, ..., x_n) \in \mathbb{R}^n$. Функцијата p_ζ се нарекува **густина на распределба** на ζ.

Ирена Стојковска

- Случајните променливи $\xi_1, \xi_2, ..., \xi_n$ велиме дека се **независни** ако за секој избор на Борелови множества $B_1, B_2, ..., B_n \in \mathcal{B}$ имаме

$$P(\xi_1 \in B_1, \xi_2 \in B_2, ..., \xi_n \in B_n) = P(\xi_1 \in B_1)P(\xi_2 \in B_2)...P(\xi_n \in B_n).$$

  Случајни променливи $\{\xi_i\}_{i \in I}$, каде I е произволно индексно множество, велиме дека се **независни** ако за секое конечно множество од различни индекси $i_1, ..., i_k \in I$ случајните променливи $\xi_{i_1}, ..., \xi_{i_k}$ се независни.

Теорема 1.2. *Нека $\xi_1, \xi_2, ..., \xi_n$ се независни апсолутно непрекинати случајни променливи дефинирани на просторот на веројатност $(\Omega, \mathcal{F}, P)$ со густини на распределба $p_{\xi_1}, p_{\xi_2}, ..., p_{\xi_n}$ соодветно. Тогаш, $\xi = (\xi_1, \xi_2, ..., \xi_n)$ е апсолутно непрекинат случаен вектор со густина на распределба*

$$p_\xi(x_1, x_2, ..., x_n) = p_{\xi_1}(x_1)p_{\xi_2}(x_2)...p_{\xi_n}(x_n)$$

за сите $(x_1, x_2, ..., x_n) \in \mathbb{R}^n$.

* * *

Нека $\zeta = (\zeta_1, \zeta_2, ..., \zeta_n)$ е случаен вектор дефиниран на просторот на веројатност $(\Omega, \mathcal{F}, P)$.

- Ако $\zeta_1, \zeta_2, ..., \zeta_n$ се дискретни случајни променливи, **условна распределба на** $(\zeta_1, ..., \zeta_k)$ **при услов** $\zeta_{k+1} = x_{k+1}, ..., \zeta_n = x_n$ се дефинира како

$$\begin{aligned} &P(\zeta_1 = x_1, ..., \zeta_k = x_k \mid \zeta_{k+1} = x_{k+1}, ..., \zeta_n = x_n) = \\ &= \frac{P(\zeta_1 = x_1, ..., \zeta_k = x_k, \zeta_{k+1} = x_{k+1}, ..., \zeta_n = x_n)}{P(\zeta_{k+1} = x_{k+1}, ..., \zeta_n = x_n)}, \end{aligned}$$

  за $P(\zeta_{k+1} = x_{k+1}, ..., \zeta_n = x_n) > 0$, при тоа условната распределба е функција од $x_1, ..., x_k$ при фиксни $x_{k+1}, ..., x_n$.

- Ако $\zeta = (\zeta_1, \zeta_2, ..., \zeta_n)$ е од апсолутно непрекинат тип со густина на распределба $p_\zeta(x_1, x_2, ..., x_n)$, **условна распределба на** $(\zeta_1, ..., \zeta_k)$ **при услов** $\zeta_{k+1} = x_{k+1}, ..., \zeta_n = x_n$ се дефинира како

$$p_{\zeta_1,...,\zeta_k}(x_1, ..., x_k \mid x_{k+1}, ..., x_n) = \frac{p_\zeta(x_1, ..., x_k, x_{k+1}, ..., x_n)}{p_{\zeta_{k+1},...,\zeta_n}(x_{k+1}, ..., x_n)},$$

  за $p_{\zeta_{k+1},...,\zeta_n}(x_{k+1}, ..., x_n) > 0$, и при тоа условната распределба е функција од $x_1, ..., x_k$ при фиксни $x_{k+1}, ..., x_n$.

Ирена Стојковска

1.1.1 Математичко очекување

Нека $\xi : \Omega \to \mathbb{R}$ е случајна променлива дефинирана на просторот на веројатност $(\Omega, \mathcal{F}, P)$. Нека P_ξ е закон на случајната променлива ξ и F_ξ е функција на распределба на ξ.

- **Математичко очекување** на ξ е **Лебеговиот интеграл**

$$E\xi = \int_\Omega \xi(w)P(dw) = \int_\Omega \xi dP.$$

- Нека $f : \mathbb{R} \to \mathbb{R}$ е Борелова функција. Тогаш, $f(\xi)$ е случајна променлива и за нејзиното математичко очекување важи

$$Ef(\xi) = \int_\Omega f(\xi)dP = \int_{\mathbb{R}} f(x)P_\xi(dx) = \int_{\mathbb{R}} f(x)dF_\xi(x).$$

  Последниот интеграл е познат како **Лебег-Стилтјесов интеграл**.

Теорема 1.3. *Нека ξ е случајна променлива дефинирана на просторот на веројатност $(\Omega, \mathcal{F}, P)$ и нека $f : \mathbb{R} \to \mathbb{R}$ е Борелова функција.*

(а) Ако ξ е дискретна случајна променлива со закон на распределба $P\{\xi = x_i\} = p_i$, $i \in I$, и ако $\sum_i |f(x_i)|p_i < +\infty$, тогаш математичкото очекување на $f(\xi)$ постои како конечен број и се пресметува според

$$Ef(\xi) = \sum_i f(x_i)p_i.$$

(б) Ако ξ е апсолутно непрекината случајна променлива со густина на распределба p_ξ, и ако $\int_{\mathbb{R}} |f(x)|p_\xi(x)dx < +\infty$, тогаш математичкото очекување на $f(\xi)$ постои како конечен број и се пресметува според

$$Ef(\xi) = \int_{\mathbb{R}} f(x)p_\xi(x)dx.$$

При специјални избори на функцијата f во Теорема 1.3 добиваме:

- Ако $f(x) = x$ имаме дека $E\xi = \sum_i x_i p_i$ за дискретна случајна променлива, и $E\xi = \int_{\mathbb{R}} xp_\xi(x)dx$ за апсолутно непрекината случајна променлива.
- За $f(x) = x^k$ се добива k**-тиот момент** на ξ, односно $E\xi^k$.
- Ако $E\xi$ постои и е конечно, тогаш $E(\xi - E\xi)^k$ е k**-тиот централен момент** на ξ.

Ирена Стојковска

- Вториот централен момент на ξ се нарекува **дисперзија** (или **варијанса**), и се означува со $D\xi = E(\xi - E\xi)^2$. Квадратниот корен од дисперзијата се нарекува **стандардна девијација**, и се означува со $\sigma = \sqrt{D\xi}$.

Теорема 1.4 (**Основен облик на неравенството на Чебишев**). *Нека ξ е ненегативна случајна променлива дефинирана на просторот на веројатност $(\Omega, \mathcal{F}, P)$. Тогаш, за секој позитивен реален број $\varepsilon > 0$,*

$$P\{\xi \geq \varepsilon\} \leq \frac{E(\xi)}{\varepsilon}. \tag{1.1}$$

Теорема 1.5 (**Неравенство на Чебишев**). *Нека ξ е случајна променлива дефинирана на просторот на веројатност $(\Omega, \mathcal{F}, P)$ со математичко очекување $E\xi < +\infty$ и дисперзија $var(\xi)$, и нека $\varepsilon > 0$ е позитивен реален број. Тогаш,*

$$P\{|\xi - E\xi| \geq \varepsilon\} \leq \frac{D\xi}{\varepsilon^2}. \tag{1.2}$$

Својство 1.1. *Нека ξ и μ се случајни променливи дефинирани на просторот на веројатност $(\Omega, \mathcal{F}, P)$, $E\xi$ и $E\mu$ постојат, и нека c е реален број. Тогаш важат следните тврдења:*

(а) $E(c\xi) = cE\xi$,

(б) ако $\xi(w) \leq \mu(w)$ *за секој* $w \in \Omega$, *тогаш* $E\xi \leq E\mu$,

(в) $|E\xi| \leq E|\xi|$,

(г) $E(\xi + \mu) = E\xi + E\mu$,

(д) ако ξ *и* μ *се независни, тогаш* $E(\xi\mu) = E\xi E\mu$,

(ѓ) $D\xi = E(\xi^2) - (E\xi)^2$,

(е) $D\xi \geq 0$,

(ж) $D(c) = 0$ *и* $D(c\xi) = c^2 D\xi$,

(з) ако $D(\xi) = 0$, *тогаш* $P\{\xi = c\} = 1$,

(ѕ) ако ξ *и* μ *се независни, тогаш* $D(\xi + \mu) = D\xi + D\mu$.

* * *

Нека ξ и μ се случајни променливи дефинирани на просторот на веројатност $(\Omega, \mathcal{F}, P)$ со $0 < D\xi < +\infty$ и $0 < D\mu < +\infty$.

Ирена Стојковска

- **Коваријанса** на ξ и μ е бројот дефиниран со

$$cov(\xi,\mu) = E(\xi - E\xi)(\mu - E\mu) = E(\xi\mu) - E\xi E\mu.$$

- За случајните променливи ξ и μ важи $D(\xi \pm \mu) = D\xi \pm 2cov(\xi,\mu) + D\mu$.

- **Коефициент на корелација** на ξ и μ е бројот дефиниран со

$$\rho(\xi,\mu) = \frac{cov(\xi,\mu)}{\sqrt{var(\xi)}\sqrt{var(\mu)}}.$$

- Ако ξ и μ се независни, тогаш $cov(\xi,\mu) = 0$ и $\rho(\xi,\mu) = 0$.

- За коефициентот на корелација важи неравенството $|\rho(\xi,\mu)| \leq 1$.

* * *

Нека $(\Omega, \mathcal{F}, P)$ е простор на веројатност, и $\xi : \Omega \to \mathbb{R}$ е случајна променлива дефинирана на него.

- **Проширена случајна променлива** е функцијата $\mu : \Omega \to \mathbb{R} \cup \{\infty\}$ за која $\mu^{-1}(B) \in \mathcal{F}$ за сите Борелови множества $B \in \mathcal{B}$.

- **Условно математичко очекување** $E(\xi|\mathcal{F}_1)$ на ξ во однос на σ-подалгебрата $\mathcal{F}_1 \subseteq \mathcal{F}$ може да се дефинира ако еден од броевите $E\xi^+$ и $E\xi^-$ е конечен (каде $\xi^+ = \max\{\xi, 0\} \geq 0$ и $\xi^- = \max\{-\xi, 0\} \geq 0$ се позитивниот и негативниот дел на ξ соодветно), и тогаш $E(\xi|\mathcal{F}_1)$ е проширена случајна променлива таква да

  (i) $E(\xi|\mathcal{F}_1)$ е $\mathcal{F}_1$-мерлива, и

  (ii) за секој $A \in \mathcal{F}_1$

$$\int_A \xi dP = \int_A E(\xi|\mathcal{F}_1) dP \text{ a.s.},$$

  каде ако $E\xi$ постои, тогаш по дефиниција $\int_A \xi dP \stackrel{def}{=} \int_\Omega \xi I_A dP$.

- Нека ξ_1 и ξ_2 се случајни променливи. **Условно математичко очекување** $E(\xi_1|\xi_2)$ **на** ξ_1 **при услов** ξ_2 се дефинира како

$$E(\xi_1|\xi_2) \stackrel{def}{=} E(\xi_1|\sigma(\xi_2)),$$

  каде $\sigma(\xi_2)$ е **σ-алгебра генерирана од** ξ_2 т.е. најмалата σ-алгебра која ги содржи сите множества $\{w : \xi_2(w) \leq x\}$ за сите $x \in \mathbb{R}$.

Ирена Стојковска

- Нека $A \in \mathcal{F}_1 \subseteq \mathcal{F}$ е настан. **Условна веројатност** $P(A|\mathcal{F}_1)$ **на** A **во однос на** $\mathcal{F}_1$ се дефинира како

$$P(A|\mathcal{F}_1) \stackrel{def}{=} E(I_A|\mathcal{F}_1).$$

Својство 1.2. *Нека $(\Omega, \mathcal{F}, P)$ е простор на веројатност, $\mathcal{F}_1 \subseteq \mathcal{F}$ е σ-подалгебра од σ-алгебрата $\mathcal{F}$ и нека ξ и μ се случајни променливи такви да $E\xi$ и $E\mu$ постојат. Тогаш важат следните тврдења:*

(а) *ако C е константа и $\xi = C$ a.s., тогаш $E(\xi|\mathcal{F}_1) = C$ a.s.,*

(б) *ако $\xi \leq \mu$ a.s., тогаш $E(\xi|\mathcal{F}_1) \leq E(\mu|\mathcal{F}_1)$ a.s.,*

(в) *$|E(\xi|\mathcal{F}_1)| \leq E(|\xi||\mathcal{F}_1)$ a.s.,*

(г) *за сите $a, b \in \mathbb{R}$, $E(a\xi + b\mu|\mathcal{F}_1) = aE(\xi|\mathcal{F}_1) + bE(\mu|\mathcal{F}_1)$ a.s.,*

(д) *нека $\mathcal{F}_0 = \{\emptyset, \Omega\}$, тогаш $E(\xi|\mathcal{F}_0) = E\xi$ a.s.,*

(ѓ) *$E(\xi|\mathcal{F}) = \xi$ a.s.,*

(е) *$E(E(\xi|\mathcal{F}_1)) = E\xi$,*

(ж) *ако $\mathcal{F}_1$ и $\mathcal{F}_2$ се σ-подалгебри од σ-алгебрата $\mathcal{F}$ такви што $\mathcal{F}_2 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}$, тогаш $E(E(\xi|\mathcal{F}_1)|\mathcal{F}_2) = E(\xi|\mathcal{F}_2)$ a.s.,*

(з) *ако μ е $\mathcal{F}_1$-мерлива, $E|\xi| < +\infty$ и $E|\xi\mu| < +\infty$, тогаш $E(\xi\mu|\mathcal{F}_1) = \mu E(\xi|\mathcal{F}_1)$ a.s.*

* * *

Нека ξ е случајна променлива дефинирана на просторот на веројатност $(\Omega, \mathcal{F}, P)$ со функција на распределба F_ξ.

- **Карактеристична функција** на случајната променлива ξ е функцијата $\varphi_\xi : \mathbb{R} \to \mathbb{C}$ дефинирана со

$$\varphi_\xi(t) = E(e^{it\xi}) = \int_{-\infty}^{+\infty} e^{itx} dF_\xi(x).$$

Основни својства:

(а) $|\varphi_\xi(t)| \leq \varphi_\xi(0) = 1$,

(б) $\varphi_{a\xi+b}(t) = \varphi_\xi(at)e^{itb}$, за $a, b = const$,

(в) $E\xi^k = \frac{\varphi_\xi^{(k)}(0)}{i^k}$, за $k = 0, 1, ..., n$.

(г) Ако $\xi_1, ..., \xi_n$ се независни случајни променливи, тогаш $\varphi_{\xi_1+\xi_2+...+\xi_n}(t) = \varphi_{\xi_1}(t)\varphi_{\xi_2}(t)...\varphi_{\xi_n}(t)$,

Ирена Стојковска

1.2 Некои поважни распределби на веројатност

1.2.1 Дискретни распределби

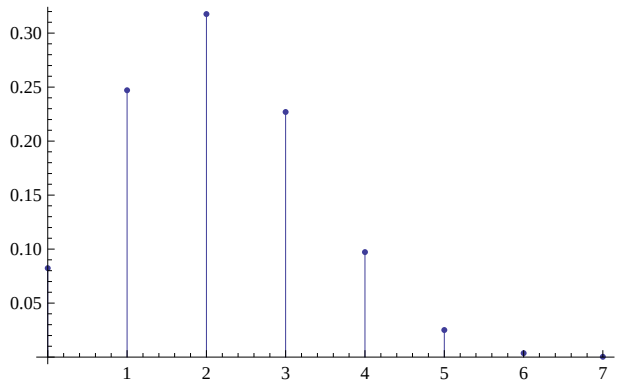
1. Биномна распределба $X \sim \mathcal{B}(n,p)$, $n \in \mathbb{N}$, $0 < p < 1$

$$p_k = P\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k},\ k = 0, 1, ..., n$$

$EX = np$, $DX = np(1-p)$, $\varphi_X(t) = (1 - p + pe^{it})$

За $n = 1$, распределбата $\mathcal{B}(1,p)$ се нарекува **Бернулиева распределба**, додека случајната променлива $Y \sim \mathcal{B}(1,p)$ се нарекува **индикатор на настанот** A чија веројатност за успех е p т.е. важи $P(A) = p$, и се означува со $Y = I_A$. Тогаш,

$$P\{I_A = 0\} = 1 - P(A) = 1 - p,\ P\{I_A = 1\} = P(A) = p.$$



Слика 1.1: Биномна распределба $\mathcal{B}(7; 0, 3)$

За $p = 0,5$ Биномната распределба е симертична, за $p < 0,5$ таа е позитивно асиметрична (Слика 1.1), а за $p > 0,5$ таа е негативно асиметрична.

2. Поасонова распределба $X \sim \mathcal{P}(a)$, $a > 0$

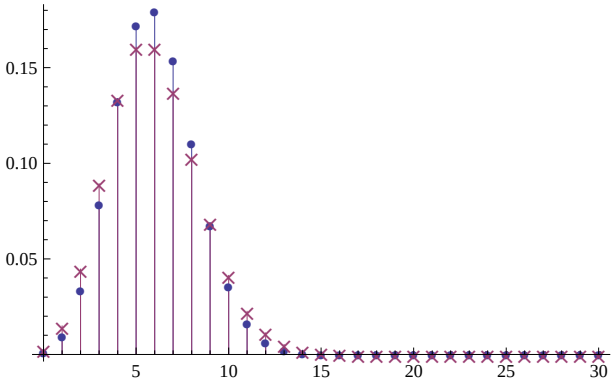
$$p_k = P\{X = k\} = \frac{a^k}{k!} e^{-a},\ k = 0, 1, 2, ...$$

$EX = a$, $DX = a$, $\varphi_X(t) = e^{a(e^{it}-1)}$

Ирена Стојковска

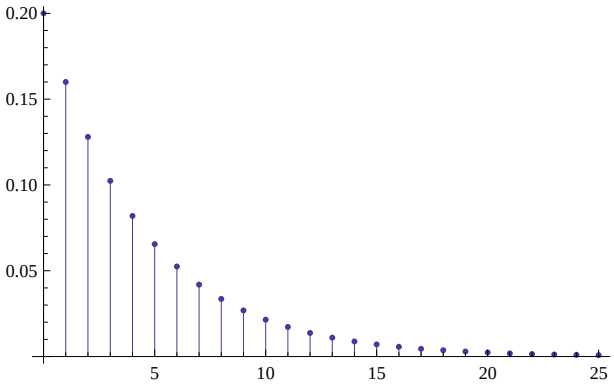
Својство 1.3. *Врската меѓу Поасонова и Биномна распределба е следната*

$$\lim_{n\to\infty,\, np\to a} \binom{n}{k} p^k (1-p)^{n-k} = \frac{a^k}{k!} e^{-a}.$$



Слика 1.2: Биномна распределба $\mathcal{B}(30; 0, 2)$ (кругчиња) и Поасонова распределба $\mathcal{P}(6)$ (крстчиња)

Претходното својство покажува дека за големи вредности на n и мали вредности на p, Биномната распредедба се апроксимира со Поасонова распределба (Слика 1.2).



Слика 1.3: Геометриска распределба $Geo(0, 2)$

3. Геометриска распределба $X \sim Geo(p),\ 0 < p < 1$

$$p_k = P\{X = k\} = p(1-p)^k,\ k = 0, 1, 2, ...$$

$EX = \frac{1-p}{p},\ DX = \frac{1-p}{p^2},\ \varphi_X(t) = \frac{p}{1-(1-p)e^{it}}$

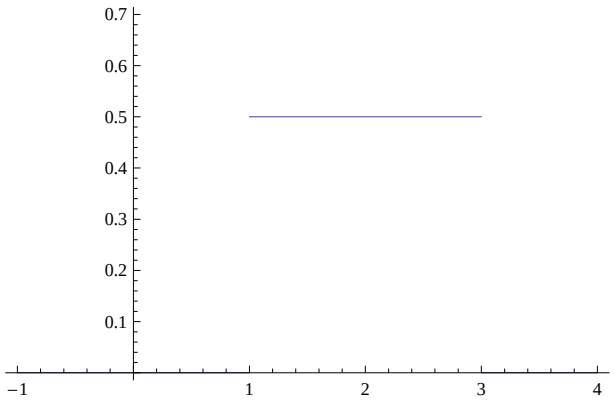
Ирена Стојковска

1.2.2 Непрекинати распределби

1. Рамномерна распределба $X \sim \mathcal{U}(a,b)$, $a<b$

$$p_X(x) = \frac{1}{b-a},\ a<x<b$$

$EX = \frac{a+b}{2}$, $DX = \frac{(b-a)^2}{12}$, $\varphi_X(t) = \frac{e^{itb}-e^{ita}}{it(b-a)}$



Слика 1.4: Рамномерна распределба $\mathcal{U}(1,3)$

2. Нормална (Гаусова) распределба $X \sim \mathcal{N}(\mu,\sigma^2)$

$$p_X(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma}\ \cdot\ e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ x\in\mathbb{R}$$

$EX = \mu$, $DX = \sigma^2$, $\varphi_X(t) = e^{ita-\frac{\sigma^2t^2}{2}}$

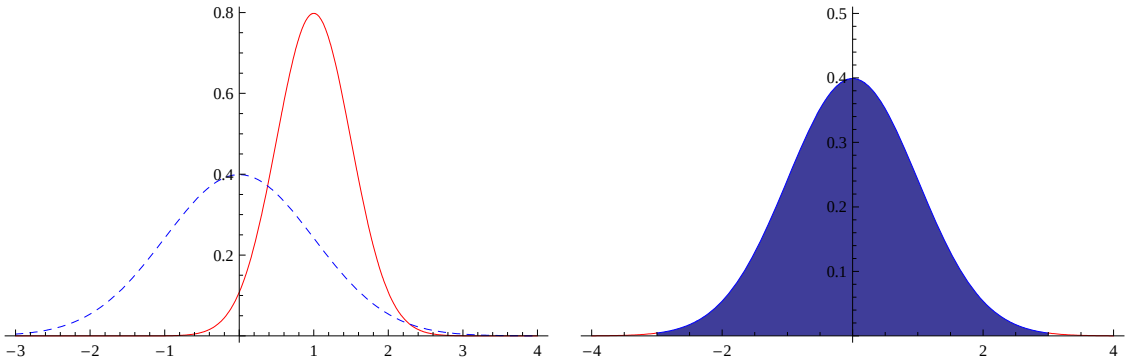
За $\mu = 0$ и $\sigma = 1$ нормалната распределба $\mathcal{N}(0,1)$ се нарекува **стандардна нормална распределба** или **нормална нормирана распределба** (Слика 1.5(а)).

Својство 1.4. *Ако случајната променлива X има $\mathcal{N}(\mu,\sigma^2)$ распределба, тогаш случајната променлива $Y = \frac{X-\mu}{\sigma}$ има $\mathcal{N}(0,1)$ распределба.*

Својство 1.5. *За случајната променлива $X \sim \mathcal{N}(\mu,\sigma^2)$ важи* **правилото на три сигми**, *односно*

$$P\{\mu-3\sigma \le X \le \mu+3\sigma\} = 0,9973 = 99,7\%.$$

Ирена Стојковска



(а) (б)

Слика 1.5: (а) Нормална (Гаусова) распределба $\mathcal{N}(1, 0.5^2)$ (црвена полна линија) и $\mathcal{N}(0, 1)$ (сина испрекината линија), (б) Правило на три сигми илустрирано за $\mathcal{N}(0, 1)$ распределба (обоениот син дел определен со интервалот $[-3, 3]$ е $99,7\%$ од плоштината под кривата)

Интерпретација на последното Својство 1.5 е такво да при нормална распределба $\mathcal{N}(\mu, \sigma^2)$ со интервалот $[\mu - 3\sigma, \mu + 3\sigma]$ е опфатено скоро целокупното веројатносно оптеретување $(99,7\%)$, Слика 1.5(б).

Својство 1.6. *Ако* $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $i = 1, 2, ..., n$ *се независни случајни променливи, тогаш* $Y = a_1X_1 + ... + a_nX_n \sim \mathcal{N}(\mu, \sigma^2)$ *каде* $\mu = a_1\mu_1 + ... + a_n\mu_n$ *и* $\sigma^2 = a_1^2\sigma_1^2 + ... + a_n^2\sigma_n^2$.

3. Гама распределба $X \sim \Gamma(\alpha, \beta)$, $\alpha > 0$, $\beta > 0$

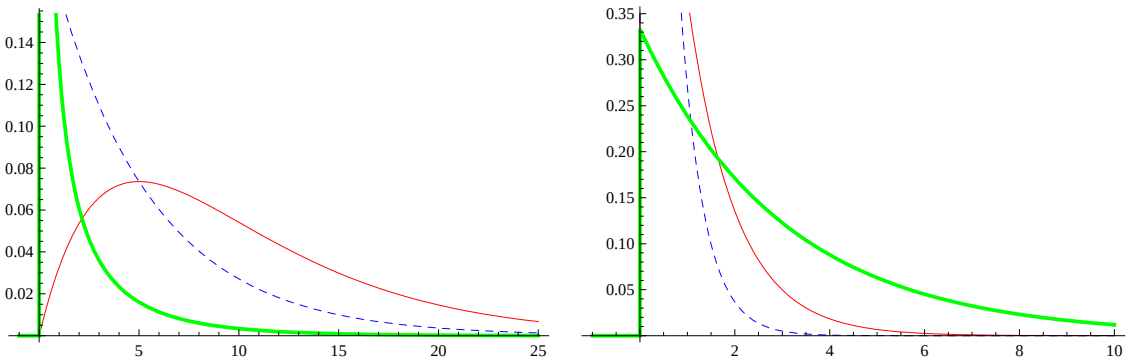
$$p_X(x) = \frac{x^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} \cdot e^{-\frac{x}{\beta}}, \ x \geq 0,$$

каде $\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1}e^{-x}dx$, $\alpha > 0$ е **Гама функција** со својства

1) $\forall \alpha > 0, \ \Gamma(\alpha + 1) = \alpha \cdot \Gamma(\alpha)$
2) $\forall n \in \mathbb{N}, \ \Gamma(n + 1) = n!, \ \Gamma(n + \frac{1}{2}) = \frac{(2n-1)!!}{2^n}\sqrt{n}$
3) $\forall \alpha \in (0, 1), \ \Gamma(\alpha) \cdot \Gamma(1 - \alpha) = \frac{\pi}{\sin \alpha\pi}$

$EX = \alpha\beta$, $DX = \alpha\beta^2$, $\varphi_X(t) = (1 - it\beta)^{-\alpha}$

Својство 1.7. *Ако* X_i, $i = 1, 2, ..., n$ *се независни случајни променливи така што* $X_i \sim \Gamma(\alpha_i, \beta)$, $i = 1, 2, ..., n$, *тогаш* $Y = X_1 + ... + X_n \sim \Gamma(\alpha, \beta)$ *каде* $\alpha = \alpha_1 + ... + \alpha_n$.



(а) (б)

Слика 1.6: (а) Гама распределба $\Gamma(2,5)$ (црвена тенка линија), $\Gamma(1,5)$ (сина испрекината линија) и $\Gamma(0.2,5)$ (зелена дебела линија), (б) Експоненцијална распределба $\mathcal{E}(1)$ (црвена тенка линија), $\mathcal{E}(0.5)$ (сина испрекината линија) и $\mathcal{E}(3)$ (зелена дебела линија)

4. **Експоненцијална распределба** $X \sim \mathcal{E}(\beta),\ \beta > 0$

$$p_X(x) = \frac{1}{\beta} \cdot e^{-\frac{x}{\beta}},\ x \geq 0$$

$EX = \beta,\ DX = \beta^2,\ \varphi_X(t) = 1 - it\beta$

Да забележиме дека експоненцијалната распределба $\mathcal{E}(\beta)$ се добива од Гама распределбата $\Gamma(\alpha,\beta)$ за $\alpha = 1$ т.е. $\Gamma(1,\beta) \equiv \mathcal{E}(\beta),\ \beta > 0$.

5. **Хи-квадрат распределба** $X \sim \chi^2_n,\ n \in \mathbb{N}$

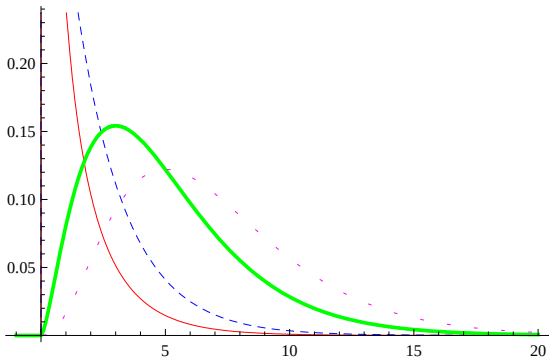
$$p_X(x) = \frac{x^{\frac{n}{2}-1}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} \cdot e^{-\frac{x}{2}},\ x \geq 0$$

$EX = n,\ DX = 2n,\ \varphi_X(t) = (1-2it)^{-n/2}$

Да забележиме дека хи-квадрат распределбата χ^2_n се добива од Гама распределбата $\Gamma(\alpha,\beta)$ за $\alpha = \frac{n}{2}$ и $\beta = 2$ т.е. $\Gamma(\frac{n}{2},2) \equiv \chi^2_n,\ n \in \mathbb{N}$.

Својство 1.8. *Ако* $X \sim \mathcal{N}(0,1)$, *тогаш* $Y = X^2 \sim \chi^2_1$.

Својство 1.9. *Ако* $X_i,\ i = 1,2,...,n$ *се независни и еднакво распределени случајни променливи со* $\mathcal{N}(0,1)$ *распределби, тогаш* $Y = X_1^2 + ... + X_n^2 \sim \chi^2_n$.



Слика 1.7: Хи-квадрат распределба χ^2_n, за $n = 1$ (црвена тенка линија), за $n = 2$ (сина испрекината линија), за $n = 5$ (зелена дебела линија) и за $n = 7$ (виолетова точкаста линија)

Забелешка 1.1. Бројот n во хи-квадрат распределбата χ^2_n се нарекува **број на степени на слобода**. Со други зборови, бројот на степени на слобода го означува бројот на линеарно независни случајни променливи меѓу случајните промеливи $X_1, X_2, ..., X_n$ во изразот за случајната променлива $Y = X_1^2 + ... + X_n^2$. Така на пример, ако меѓу случајните променливи $X_1, X_2, ..., X_n$ постои една линеарна врска, на пример $X_1 + X_2 + ... + X_n = 0$, тогаш бројот на степени на слобода се намалува за еден, т.е $Y = X_1^2 + ... + X_n^2 \sim \chi^2_{n-1}$.

6. Студентова распределба $X \sim t_n$, $n \in \mathbb{N}$

$$p_X(x) = \frac{(1 + \frac{x^2}{n})^{-\frac{n+1}{2}}}{B(\frac{1}{2}, \frac{n}{2})\sqrt{n}}, x \in \mathbb{R},$$

каде $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$, $\alpha > 0$, $\beta > 0$ е **Бета функција**. Позната е следната врска помеѓу Бета и Гама функцијата

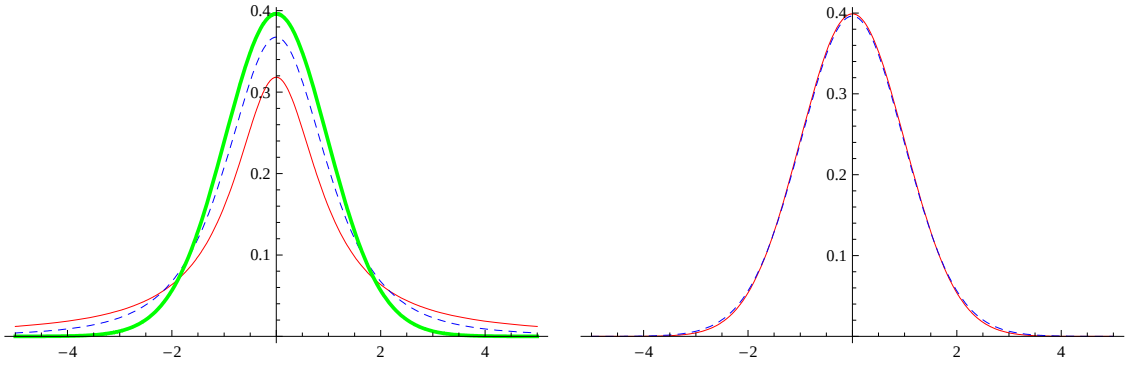
$$\forall \alpha, \beta > 0, \ B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.$$

$EX = 0$ за $n > 1$, $DX = \frac{n}{n-2}$ за $n > 2$

И овде, бројот n во студентовата t_n распределба се нарекува **број на степени на слобода**. Да забележиме дека за $n = 1$ се добива Кошиева распределба, односно густина на распределба $p_X(x) = \frac{1}{\pi(1+x^2)}$. Додека пак во пракса за $n > 30$, студентовата распределба може да се апроксимира со стандардна нормална распределба, имено важи $p_X(x) \to \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ кога $n \to \infty$ (Слика 1.8(б)).

Ирена Стојковска

Својство 1.10. *Ако* $X \sim \mathcal{N}(0,1)$ *и* $Y \sim \chi^2_n$ *се независни случајни променливи, тогаш* $Z = \frac{X}{\sqrt{\frac{Y}{n}}} \sim t_n$.



(а) (б)

Слика 1.8: (а) Студентова распределба t_n, за $n = 1$ (црвена тенка линија), за $n = 3$ (сина испрекината линија) и за $n = 36$ (зелена дебела линија), (б) Стандардна нормална распределба $\mathcal{N}(0, 1^2)$ (црвена тенка линија) и студентова таспределба t_n за $n = 36$ (сина испрекината линија)

7. Фишерова распределба $X \sim F_{n_1,\, n_2},\ n_1,\ n_2 \in \mathbb{N}$

$$p_X(x) = \frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1}}{B(\frac{n_1}{2}, \frac{n_2}{2})(1 + \frac{n_1 x}{n_2})^{\frac{n_1+n_2}{2}}}, x > 0$$

$EX = \frac{n_2}{n_2-2}$ за $n_2 > 2$, $DX = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$ за $n_2 > 4$

Својство 1.11. *Ако* $X_1 \sim \chi^2_{n_1}$ *и* $X_2 \sim \chi^2_{n_2}$ *се независни случајни променливи, тогаш* $Y = \frac{X_1/n_1}{X_2/n_2} \sim F_{n_1,n_2}$.

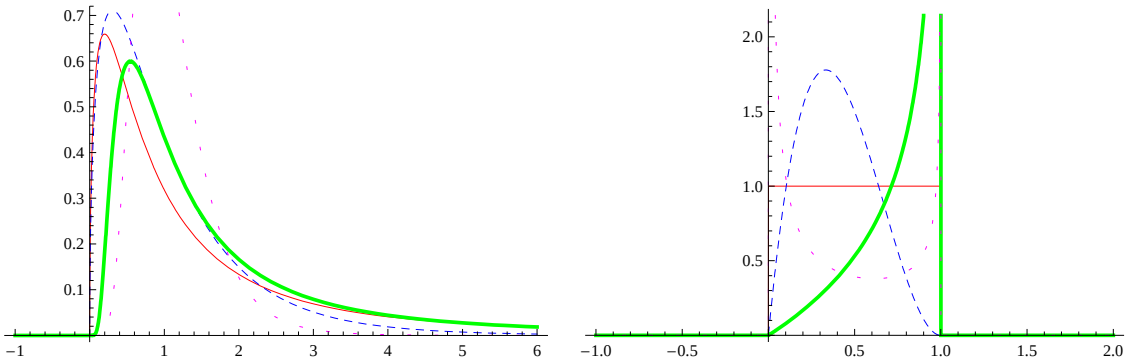
8. Бета распределба $X \sim Bet(\alpha, \beta),\ \alpha > 0,\ \beta > 0$

$$p_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)},\ 0 \le x \le 1$$

$EX = \frac{\alpha}{\alpha+\beta}$, $DX = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$

Да забележиме дека рамномерната распределбата $\mathcal{U}(0,1)$ се добива од Бета распределбата $Bet(\alpha,\beta)$ за $\alpha = \beta = 1$ т.е. $Bet(1,1) \equiv \mathcal{U}(0,1)$.

Ирена Стојковска



(а) (б)

Слика 1.9: (а) Фишерова распределба F_{n_1,n_2}, за $n_1 = n_2 = 3$ (црвена тенка линија), за $n_1 = 3$, $n_2 = 22$ (сина испрекината линија), за $n_1 = 22$, $n_2 = 3$ (зелена дебела линија) и за $n_1 = n_2 = 22$ (виолетова точкаста линија), (б) Бета распределба $Bet(1,1)$ (црвена тенка линија), $Bet(2,3)$ (сина испрекината линија), $Bet(2,0.5)$ (зелена дебела линија) и $Bet(0.2,0.5)$ (виолетова точкаста линија)

Својство 1.12. *Ако $X_1 \sim \mathcal{E}(\beta_1)$ и $X_2 \sim \mathcal{E}(\beta_2)$ се независни случајни променливи, тогаш $Y = \frac{1}{X_1+X_2} \sim Bet(\beta_1, \beta_2)$.*

Својство 1.13. *Ако $X \sim F_{n_1,n_2}$, тогаш $Y = \frac{(n_1/n_2)X}{1+(n_1/n_2)X} \sim Bet(\frac{n_1}{2}, \frac{n_2}{2})$.*

1.3 Низи од случајни променливи

Нека $\xi_1, \xi_2, \ldots$ е низа од случајни променливи дефинирани на просторот на веројатност $(\Omega, \mathcal{F}, P)$. Нека ξ е случајна променлива дефинирана на истиот простор на веројатност.

- Ако за секој $w \in \Omega$, $\xi_n(w) \to \xi(w)$, кога $n \to \infty$, тогаш велиме дека низата од случајни променливи $\{\xi_n\}$ **конвергира по точки** кон случајната променлива ξ.

- Низата од случајни променливи $\{\xi_n\}$ **конвергира скоро сигурно** или **конвергира со веројатност еден** кон случајната променлива ξ ако

$$P\{w : \lim_{n\to\infty} \xi_n(w) = \xi(w)\} = 1.$$

Ознаки: $\xi_n \overset{a.s.}{\to} \xi$ или $\xi_n \to \xi$ a.s. или $\xi_n \to \xi$ w.p.1.

Ирена Стојковска

- Низата од случајни променливи $\{\xi_n\}$ **конвергира по веројатност** кон случајната променлива ξ ако

$$\forall \varepsilon > 0,\ \lim_{n\to\infty} P\{w : |\xi_n(w) - \xi(w)| > \varepsilon\} = 0.$$

  Ознаки: $\xi_n \overset{P}{\to} \xi$ или $\xi_n \to \xi$ in prob.

- Низата од случајни променливи $\{\xi_n\}$ **конвергира средно квадратно** кон случајната променлива ξ ако

$$\lim_{n\to\infty} E|\xi_n - \xi|^2 = 0.$$

  Ознаки: $\xi_n \overset{MS}{\to} \xi$ или $\xi_n \to \xi$ in m.s.

- Низата од случајни променливи $\{\xi_n\}$ **конвергира по распределба** кон случајната променлива ξ ако за функциите на распределба F_n на ξ_n и функцијата на распределба F на ξ важи

$$\lim_{n\to\infty} F_n(x) = F(x),\ \forall x \in C,$$

  каде $C \subseteq \mathbb{R}$ е множеството од точки на непрекинатост на F.

  Ознаки: $\xi_n \overset{dist.}{\to} \xi$ или $\xi_n \to \xi$ in dist. или $\xi_n \Rightarrow \xi$.

Теорема 1.6. *Нека $\{\xi_n\}$ е низа од независни случајни променливи. Тогаш,*

$$\xi_n \to 0 \text{ a.s. } \Leftrightarrow \forall \varepsilon > 0, \sum_{n=1}^{\infty} P\{|\xi_n| \geq \varepsilon\} < +\infty.$$

Својство 1.14. *Нека $(\Omega, \mathcal{F}, P)$ е простор на веројатност и нека $\xi, \xi_1, \xi_2, ... : \Omega \to \mathbb{R}$ се случајни променливи. Тогаш важат следните тврдења:*

(а) ако $\xi_n \to \xi$ a.s., тогаш $\xi_n \to \xi$ in prob.,

(б) ако $\xi_n \to \xi$ in m.s., тогаш $\xi_n \to \xi$ in prob.,

(в) ако $\xi_n \to \xi$ in prob., тогаш $\xi_n \to \xi$ in dist.

Ирена Стојковска

1.4 Гранични теореми

Теорема 1.7 (**Слаб закон на големите броеви**). *Нека* $\{\xi_n\}$ *е низа од независни случајни променливи дефинирани на просторот на веројатност* $(\Omega, \mathcal{F}, P)$ *со* $E\xi_k = m < +\infty$ *и* $var(\xi_k) = \sigma^2$ *за сите* $k = 1, 2, ...$ *Тогаш,*

$$\frac{1}{n}\sum_{k=1}^{n}\xi_k \xrightarrow{P} m \quad \textit{кога} \quad n \to \infty. \tag{1.3}$$

Теорема 1.8 (**Силен закон на големите броеви**). *Нека* $\{\xi_n\}$ *е низа од независни случајни променливи дефинирани на просторот на веројатност* $(\Omega, \mathcal{F}, P)$ *со* $E\xi_k = 0$ *и* $E\xi^4 \leq c$ *за сите* $k = 1, 2, ...$ *и* c *е некоја позитивна константа. Тогаш,*

$$\frac{1}{n}\sum_{k=1}^{n}\xi_k \xrightarrow{a.s.} 0 \quad \textit{кога} \quad n \to \infty. \tag{1.4}$$

Понекогаш силниот закон на големите броеви е познат како само **закон на големите броеви**.

Теорема 1.9 (**Централна гранична теорема**). *Нека* $\{\xi_n\}$ *е низа од независни и еднакво распределени случајни променливи дефинирани на просторот на веројатност* $(\Omega, \mathcal{F}, P)$ *со* $E\xi_k = m < +\infty$ *и* $var(\xi_k) = \sigma^2 > 0$ *за сите* $k = 1, 2, ...$ *Тогаш, за произволен* $x \in \mathbb{R}$

$$P\{\frac{\sum_{k=1}^{n}\xi_k - nm}{\sigma\sqrt{n}} \leq x\} \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{u^2}{2}}du \quad \textit{кога} \quad n \to \infty, \tag{1.5}$$

што значи дека

$$\frac{\sum_{k=1}^{n}\xi_k - nm}{\sigma\sqrt{n}} \xrightarrow{dist.} N(0,1) \quad \textit{кога} \quad n \to \infty. \tag{1.6}$$

2

Дескриптивна статистика

2.1 Податоци. Видови на податоци

Составен дел на секое истражување е собирање и анализа на податоци поврзани со одредени појави кои се предмет на истажувањето. Така на пример, при испитување на оптеретеност на раскрсниците од возила, се собираат податци за бројот на возила кои поминуваат за единица време во различни временски пресеци во текот на еден ден. Собирањето на податоците може да се спроведе преку **набљудување** (мерење на карактеристиките без да се влијае на условите) или **експериментирање** (свесно поставување на одредени услови за да набљудува рекцијата на соодветниот услов).

Кога податоците се собрани од набљудувања, односно експерименти на кои им делуваат случајни фактори (протокот на возила е случаен, не е однапред одреден) познати се под името **статистички податоци**. При тоа, мерените карактеристики за појавите ги нарекуваме **обележја** (број на возила за единица време кои ја поминуваат раскрсницата).

Видот на податоците е тесно сврзан со видот на обележјата кои се мерат. Ако обележјето ги распоредува во групи или категории измерените вредности (пол, националност, верска определеност), тогаш така добиените податоци се нарекуваат **категориски (квалитативни) податоци**. Доколку измерените вредности се нумерички, податоците се нарекуваат **квантитативни (нумерички) податоци**. Соодветните квантитативни обележја може да се од **дискретен тип** (ако множеството вредности е конечно или пребројиво) и од **непрекинат тип** (ако прима било која вредност од некој интервал).

Статистиката е наука која проучува научни методи за собирање, средување, прикажување и анализа на податоците (**дескриптивна статистика**), како и извлекување на заклучоци и донесување на соодветни решенија (**математичка статистика**).

2.2 Прикажување на податоците

1) Табела на честоти

При набљудувањата или експериментитањата вниманието се насочува на една или повеќе величини (карактеристики). Ако се разгледува само една величина, и ако истата ја означиме со X, тогаш резултатот од едно мерење на истата е реален број x (доколку станува збор за нумеричко обележје) или категорија x (ако разгледуваме категориско обележје). Ако се спроведат n мерења, се добива конечна низа од вредности (нумерички, односно катекатириски) $x_1, x_2, ..., x_n$ кои се нарекуваат **статитички податоци** кои одговараат на разгледуваното **статистичко обележје** X.

Понатаму, статистичките податоци $x_1, x_2, ..., x_n$ заради прегледност се запишуваат во **табела на честоти** на следниот начин. Нека r е бројот на различни вредности на обележјето X опфатени со податоците $x_1, x_2, ..., x_n$. Нека тие вредности ги означиме со $a_1, a_2, ..., a_r$. Доколку се обележјето X е нумеричко овие вредности се подредуваат по големина, односно $a_1 < a_2 < ... < a_r$. Со f_i ја означуваме **честотата** на вредноста a_i, $i = 1, 2, ..., r$, број кој покажува колку пати се среќава вредноста a_i во податоците $x_1, x_2, ..., x_n$. Тогаш, вкупниот збир на честотите да е еднаков на вкупниот број изведени мерења, односно важи

$$f_1 + f_2 + ... + f_r = n.$$

Доколку табелата на честоти се дополни со соодветните количници f_i/n, $i = 1, 2, ..., n$ наречени **релативни честоти** се добива **табела на релативни честоти**. Честа ознака за релативните честоти е $p_i = f_i/n$, $i = 1, 2, ..., n$, и при тоа важи

$$0 \leq p_i \leq 1,\ i = 1, 2, ..., r \text{ и } p_1 + p_2 + ... + p_r = 1.$$

Вредност на обележјето X	a_1	a_2	$\cdots$	a_r
Честота	f_1	f_2	$\cdots$	f_r
Релативна честота	p_1	p_2	$\cdots$	p_r

Табела 2.1: Табела на честоти и релативни честоти

2) Категориско обележје - столбести графици, пити

Обележјата кои се категориски најчесто примаат описни вредности - имиња на категориите.

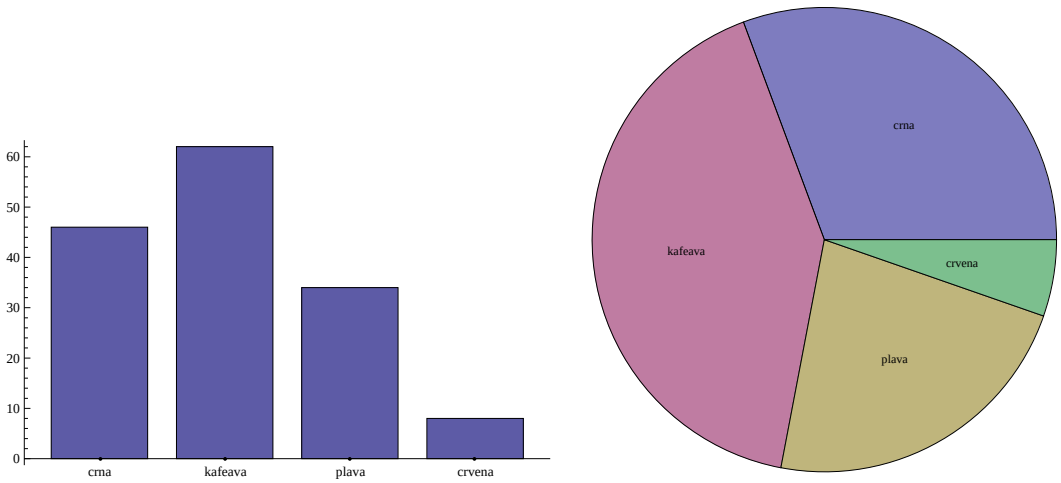
Ирена Стојковска

Пример 2.1. Така на пример, доколку обележето X е боја на косата, тогаш тоа може да ги прими вредностите: "црна", "кафеава", "плава" и "црвена". Во следната Табела 2.2 дадени се резултатите од истражувањето направено на 150 испитанци за одредување на застапеноста на бојата на косата.

Боја на косата	црна	кафеава	плава	црвена
Честота	46	62	34	8
Релативна честота	0,3067	0,4133	0,2267	0,0533
%	30,67%	41,33%	22,67%	5,33%

Табела 2.2: Застапеност на бојата на косата

Соодветен графички приказ на табелите на честота за категориските статистички податоци е со **столбести графици** (*bar chart*) и **пити** (*pie chart*). Со столбестиот график (Слика 2.1а)) се прави брза споредба на честотите на различните категории, додека графикот пита (Слика 2.1б)) овозможува полесно согледување на тоа колкав дел од целината зафаќа секоја од категориите.



(а) (б)

Слика 2.1: (а) Столбест график за застапеноста на бојата на косата, (б) График пита за застапеноста на бојата на косата

Столбестите графици и питите овозможуваат брзо согледување на распределбата на податоците, но тие се со ограничена употреба во анализата на податоците. Графиците за нумеричките податоци се попогодни за разбирање на распределбата и анализа на истата.

Ирена Стојковска

3) Нумеричко дискретно обележје - полигон на честоти, функција на кумулативни честоти

Соодветни графички прикази за распределбите на нумеричките податоци кои одговараат на дисктерните нумерички обележја се полигоните на честоти. **Полигонот на честоти** претставува искршена линија која ги поврзува точките $(a_i, f_i),\ i = 1, 2, ..., r$, додека **полигонот на релативни честоти** е искршена линија која ги поврзува точките $(a_i, p_i),\ i = 1, 2, ..., r$. Ознаките $a_i,\ f_i$ и $p_i,\ i = 1, 2, ..., r$ се воведени претходно. Разликата меѓу овие два полигона е само во мерната единица на ординатната оска. Полигоните на честоти во потполност и еднозначно ја определуваат распределбата на податоците, што не случај со истите за податоците кои одговараат на непрекинати нумерички обележја, како што ќе биде објаснето подоцна.

Друг графички приказ на распределбата на податоците е со графикот на **функцијата на кумулативните честоти** $K_n : \mathbb{R} \to \mathbb{R}$ дефинирана со

$$K_n(x) = \sum_{a_j < x} f_j,\ x \in \mathbb{R},$$

односно со графикот на **функцијата на кумулативните релативни честоти** $F_n : \mathbb{R} \to \mathbb{R}$ дефинирана со

$$F_n(x) = \sum_{a_j < x} p_j,\ x \in \mathbb{R},$$

позната уште како **емпириска функција на распределба**.

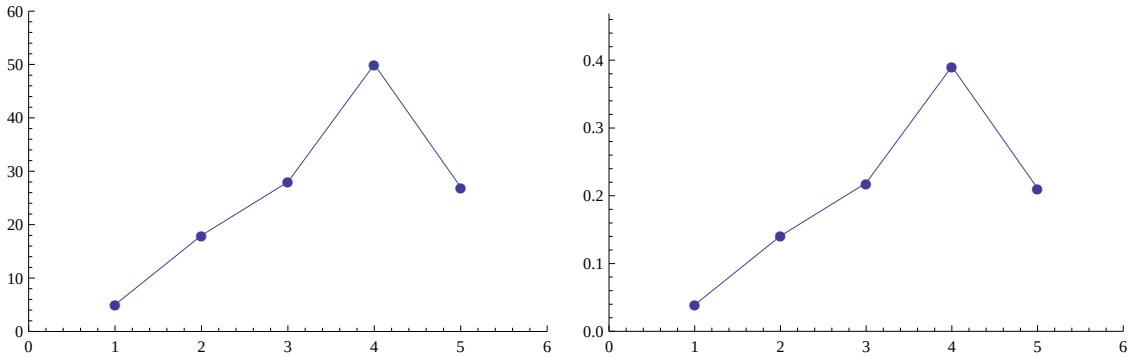
Пример 2.2. На пример, графички приказ со полигони на податоците од Табела 2.3, кои се однесуваат на оценката по математика која ја добиле 128 ученици од 5 одделение од едно училиште при завршната проверка на знаењата, е даден со Слика 2.2.

Оценка по математика	1	2	3	4	5
Честота	5	18	28	50	27
Релативна честота	0,0391	0,1406	0,2188	0,3906	0,2109

Табела 2.3: Оценка по математика на завршната проверка на знаењата

Од полигонот на честоти се согледува која вредност на обележјето има најголема честота (таму каде е највисокот врв), додека полигонот на релативни честоти може да даде оценка за симетричноста на распределбата на податоците.

Ирена Стојковска

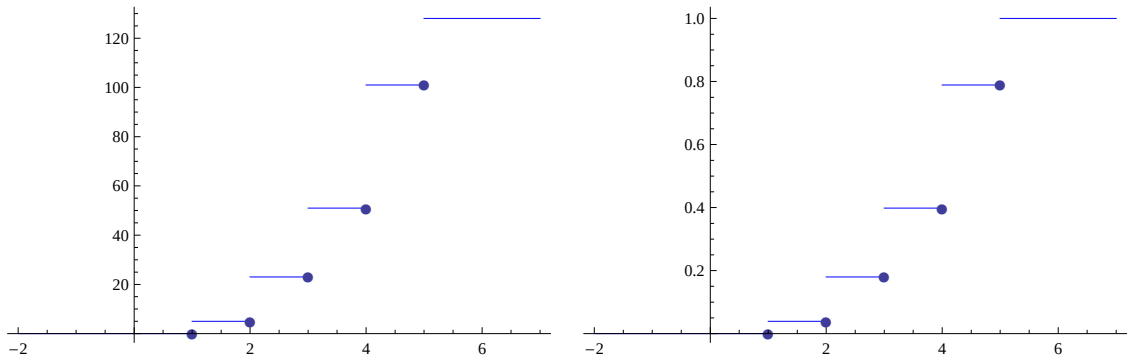


(а) (б)

Слика 2.2: (а) Полигон на честотата на податоците дадени со Табела 2.3, (б) Полигон на релативната честота на податоците дадени со Табела 2.3

Графиците на функцијата на кумулативните честоти $K_n(x)$ и функцијата на кумулативните релативни честоти $F_n(x)$ за податоците дадени со Табела 2.3 се на Слика 2.3.

$$K_n(x) = \begin{cases} 0 & , x \leq 1 \\ 5 & , 1 < x \leq 2 \\ 23 & , 2 < x \leq 3 \\ 51 & , 3 < x \leq 4 \\ 101 & , 4 < x \leq 5 \\ 128 & , x > 5 \end{cases} \qquad F_n(x) = \begin{cases} 0 & , x \leq 1 \\ 0,0391 & , 1 < x \leq 2 \\ 0,1797 & , 2 < x \leq 3 \\ 0,3985 & , 3 < x \leq 4 \\ 0,7891 & , 4 < x \leq 5 \\ 1 & , x > 5 \end{cases}$$



(а) (б)

Слика 2.3: (а) Функција на кумулативните честоти $K_n(x)$ на податоците дадени со Табела 2.3, (б) Функција на кумулативните релативни честоти $F_n(x)$ на податоците дадени со Табела 2.3

4) Нумеричко непрекинато обележје - групирање во интервали, полигон и хистограм на честоти, функција на кумулативни честоти

При собирање на податоци кои одговаарат на обележја кои примаат било која вредност од некој интервал може да се случи да нема измерени две исти вредности или пак бројот на измерени исти вредности да е многу мал во споредба со вкупниот број на измерени вредности. Во тој случај, портебно е податоците да се групираат во интервали за да се состави табелата на честоти која ќе служи како основа за графичкиот приказ на распределбата на податоците.

Групирањето во интервали не е еднозначно одредено. Најчесто се земаат во предвид најмалата $x_{\min} = \min\{x_1, x_2, ..., x_n\}$ и најголемата $x_{\max} = \max\{x_1, x_2, ..., x_n\}$ вредност на податоците $x_1, x_2, ..., x_n$, потоа интервалот $[x_{\min}, x_{\max}]$ се проширува од лево и од десно и така проширениот интервал $[a, b] \supseteq [x_{\min}, x_{\max}]$ се дели на одреден број на дисјунктни подинтервали. Најчесто подинтервалите се со иста должина h. Бројот r на подинтервали се одредува според некоја од следните формули

$$r \approx \sqrt{n}, \quad r \approx 1 + 3,21 \log n \quad \text{или} \quad r \approx 5 \log n,$$

или пак интуитивно се со цел да и после групирањето во интервали добиената распределба да биде што е можно поблизу до вистинската распределба на податоците. Препораките за избор на бројот r на подинтервали се тие да r биде 5 - 10% од n и не повеќе од 30% од n. Целта е да по групирањето на податците да може да се воочат важните својства на разгледуваното статистичко обележје.

По одредување на бројот r на подинтервали, должината h на секој од подинтервалите се пресметува според

$$h = \frac{b - a}{r},$$

додека точките $a_0 < a_1 < a_2 < ... < a_r$ кои ги определуваат границите на подинтервалите се добиваат според

$$a_0 = a,\ a_i = a_0 + ih,\ i = 1, 2, ..., r - 1,\ a_r = b.$$

Табелата на честоти која одговара на групираните податоците во интервали се пополнува на тој начин што на местото од вредности на статистичкото обележје X се наоѓаат подинтервалите $(a_{i-1}, a_i]$, $i = 1, 2, .., r$, а честотата f_i се однесува на бројот на вредности меѓу податоците $x_1, x_2, ..., x_n$ кои припаѓаат во интервалот $(a_{i-1}, a_i]$, а соодветните релативи честоти се пресметуваат според $p_i = f_i/n$ (Табела 2.4). Првиот интервал е затворен интервал од двете страни, т.е. $[a_0, a_1]$.

Ирена Стојковска

Интервал на вредности на обележјето X	$[a_0, a_1]$	$(a_1, a_2]$	$\cdots$	$(a_{r-1}, a_r]$
Честота	f_1	f_2	$\cdots$	f_r
Релативна честота	p_1	p_2	$\cdots$	p_r

Табела 2.4: Табела на честоти за групираните податоци во интервали

За потребите на графичкото претставување на податоците поделени во интервали, Табела 2.4 се дополнува со вредностите на средините на интервалите

$$\overline{a}_i = \frac{a_{i-1} + a_i}{2},\ i = 1, 2, ..., r,$$

и вредностите на функцијата на кумулативни честоти $K_n(x)$ и функцијата на кумулативни релативни честоти $F_n(x)$ на секој од подинтервалите.

Соодветен графички приказ на групираните податоци во интервали е со **хистограм на честоти** кој се состои од слепени правоаголници со ширина h и висина f_i поставени над интервалот $(a_{i-1}, a_i]$, $i = 1, 2, .., r$. Додека пак **хистограмот на релативни честоти** се разликува само во висината на правоаголниците која изнесува p_i. За хистограмот на релативни честоти важи дека збирот на плоштината правоаголниците е 1.

Групираните податоци се претставуваат графички и со **полигон на честоти**, односно **полигон на релативни честоти**, така што се поврзуваат точките $(\overline{a}_i, f_i)$, односно точките $(\overline{a}_i, p_i)$, $i = 1, 2, .., r$. Ова покажува дека при групирањето на податците во интервали сите вредности од i-тиот интервал се апроксимираат со средината $\overline{a}_i$ на тој интервал. На тој начин се губат одреден број на информации во врска со разгледуваната појава содржани во измерените статистички податоци, но од друга стана се овозможува да се одделат важните својства на разгледуваното статистичко обележје од неважните.

Графиците на **функцијата на кумулативни честоти**, односно **функцијата на кумулативни релативни честоти** се разликуваат од истите кај податоците кои одговараат на дискретни статитички обележја во тоа што овој пат се прикажуваат во вид на крива со спојување на точките $(a_i, K_n(a_i))$, односно точките $(a_i, F_n(a_i))$, $i = 0, 1, 2, ..., r$. Наклонот на така добиената крива покажува како се натрупуваат (кумулираат) дадените статистички податоци по должината на апцисната оска. Пострмен наклон одговара на поголема брзина на натрупување на податоците. Лево од a_0 и десно од a_r кривата е паралелна со апцисната оска, што значи дека во тој дел нема вредности од дадените статистички податоци. Доколку на некој од подинтервалите кривата е нормална на апцисната оска, тоа значи дека во тој интервал нема вредности од статистичките податоци.

Ирена Стојковска

Пример 2.3. Горе споменатите графички прикази ќе ги илустрираме на еден пример. По мерењето на висината кај 70 ученици во трета година во една гимназија добиени се следните резултати:

177.5, 165.0, 178.3, 169.0, 168.9, 165.5, 167.7, 157.2, 165.8, 179.0,
172.0, 172.9, 184.9, 175.4, 179.9, 181.3, 173.5, 187.2, 174.5, 166.8,
163.6, 183.2, 184.0, 158.0, 180.6, 172.5, 167.2, 167.4, 171.5, 186.5,
166.2, 166.3, 159.3, 181.6, 170.6, 167.6, 173.4, 161.8, 165.2, 180.7,
177.5, 161.7, 180.2, 168.5, 160.6, 181.8, 170.2, 163.8, 181.2, 193.1,
181.3, 168.4, 185.9, 151.1, 148.6, 182.8, 151.3, 174.7, 163.0, 170.0,
178.8, 158.5, 177.1, 186.9, 159.2, 181.4, 168.0, 161.5, 163.7, 164.8

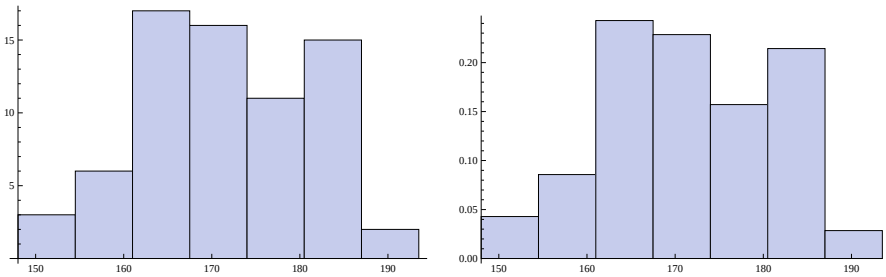
За да ги групираме овие податоци во интервали го одредуваме интервалот на нивното простирање. Бидејќи најмалата вредност е 148.6, а најголемата вредност е 193.1 значи дека вредностите на овие статистички податоци се од интервалот $[148.6, 193.1]$. Вкупниот број на податоците е $n = 70$, па препорачлив број на подинтервали е

$$r \approx \sqrt{70} = 8,366 \approx 8,$$
$$r \approx 1 + 3,21 \log 70 = 6,923 \approx 7$$
$$\text{или } r \approx 5 \log 70 = 9,225 \approx 9.$$

Друга препорачлива референца за бројот на подинтервали е 5-10% од вкупниот број на податоци. Затоа се одлучуваме за $r = 7$ подинтервали. Интервалот $[148.6, 193.1]$ го прошируваме и од лево и од десно, и така го добиваме интервалот $[148.0, 193.5]$ чија должина е $193.5 - 148.0 = 45.5$. Тогаш, должината на секој од 7-те подинтервали е $h = (193.5 - 148.0)/7 = 45.5/7 = 6.5$. Значи границите на подинтервалите се 148.0, 154.5, 161.0, 167.5, 174.0, 180.5, 187.0, 193.5. Следно вршиме распределување на податоците во вакадобиените подинтервали и ја добиваме Табела 2.5 на честотите и реалативните честоти на групираните податоци во интервали.

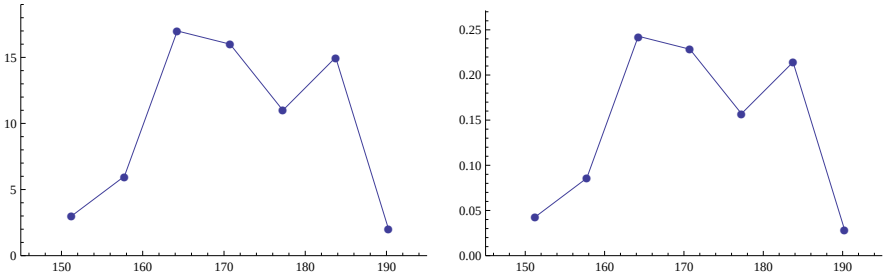
Табела 2.5 ја дополнуваме со средините на интервалите, кумулативните честоти и кумулативните релативни честоти. Врз база на вредностите од оваа табела ги цртаме хистограмите на честоти и релативни честоти (Слика 2.4), полигоните на честоти и релативни честоти (Слика 2.5) и графиците на функциите на кумулативни честоти и кумулативни релативни честоти (Слика 2.6).

Ирена Стојковска



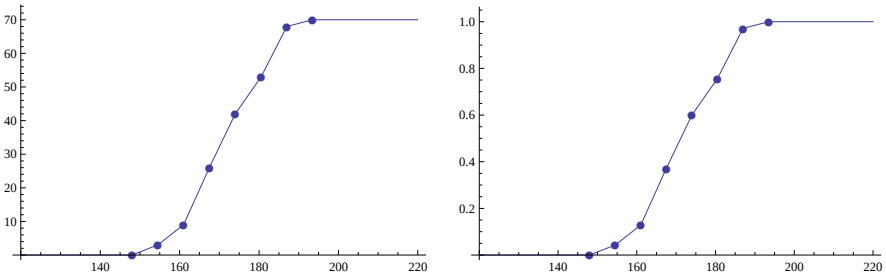
(а) (б)

Слика 2.4: (а) Хистограм на честоти на податоците дадени со Табела 2.5, (б) Хистограм на релативни честоти на податоците дадени со Табела 2.5



(а) (б)

Слика 2.5: (а) Полигон на честоти на податоците дадени со Табела 2.5, (б) Полигон на релативни честоти на податоците дадени со Табела 2.5



(а) (б)

Слика 2.6: (а) Графички приказ на функцијата на кумулативни честоти $K_n(x)$ за податоците дадени со Табела 2.5, (б) Графички приказ на функцијата на кумулативни релативни честоти $F_n(x)$ за податоците дадени со Табела 2.5

Висина (во cm)	Средина на интерв.	Честота	Релативна честота	Кумул. честота	Кумул. релативна честота
[148.0, 154.5]	151.25	3	0.0428	3	0.0428
(154.5, 161.0]	157.75	6	0.0857	9	0.1285
(161.0, 167.5]	164.25	17	0.2429	26	0.3714
(167.5, 174.0]	170.75	16	0.2286	42	0.6
(174.0, 180.5]	177.25	11	0.1571	53	0.7571
(180.5, 187.0]	183.75	15	0.2143	68	0.9714
(187.0, 193.5]	190.25	2	0.0286	70	1

Табела 2.5: Распределба на висината на учениците во интервали

Под **интерпретација на графичките прикази** се подразбира опишување на распределбата на податоците така што се бараат главните елементи на распределбата. Најнапред се бара да се најде општиот модел или однесување на распределбата, како и очигледните отстапувања од овој облик. Кај хистограмите, се бара центарот и простирањето на податоците, како и индивидуалните вредности кои се надвор од општиот модел (outlier). Главните елементи на распределбата ги опфаќаат највисоките врвови, симетричноста или искривеноста на хистограмите.

2.3 Бројни карактеристики на распределбата на податоците

1) Аритметичка средина или просек на податоците

При анализа на распределбата на статистичките податоци $x_1, x_2, ..., x_n$, покрај графичките прикази со полигини и хистограми се користат и бројни карактеристики за да се овозможи важните својства на разгледуваното статистичко обележје X да се изразат што попрецизно.

Груб показател на местото на податоците на бројната оска е **аритметичката средина** или **просекот** на податоците даден со формулата (2.1).

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.1}$$

Својство 2.1. *Збирот од сите отстапувања на податоците од нивната аритметичка средина е еднаков на нула т.е.*

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0.$$

Доказ. $\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - n\overline{x} = n\overline{x} - n\overline{x} = 0.$ ∎

Својство 2.1 ја дава смислата според која треба аритметичката средина на податоците да се сфати како некоја одредена средина на статистичките податоци.

Својство 2.2. *Збирот на квадратите на сите отстапувања на податоците од нивната аритметичка средина е помал од збирот на квадратите на сите отстапувања на податоците од било кој друг број $c \in \mathbb{R}$ т.е.*

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 \leq \sum_{i=1}^{n}(x_i - c)^2.$$

Равенство се достигнува за $c = \overline{x}$.

Доказ. Нека $c \in \mathbb{R}$ е произволен број. Тогаш, бидејќи

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

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

за следната разлика имаме

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 - \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \overline{x}^2 - 2c\overline{x} + c^2 = (\overline{x} - c)^2 \geq 0,$$

од каде следи тврдењето што требаше да се докаже. ∎

Ако статистичките податоци се веќе средени во табела на честоти (види Табела 2.1), тогаш аритметичката средина се пресметува според формулата

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{r} f_i a_i.$$

Доколку станува збор за податоци кои одговараат на непрекинато статистичко обележје (дадени со Табела 2.4), групрани во интервали, аритметичката средина се пресметува според формулата

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{r} f_i \overline{a}_i,$$

каде $\overline{a}_i$ е средина на подинтервалот $[a_{i-1}, a_i]$, $i = 1, 2, ..., r$.

Така на пример, за податоците дадени со Пример 2.2 ја имаме табелата со честоти Табела 2.3 од каде пресметуваме дека

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{r} f_i a_i = \frac{1}{128}(5 \cdot 1 + 18 \cdot 2 + 28 \cdot 3 + 50 \cdot 4 + 27 \cdot 5) = 3,59375.$$

Додека пак, за податоците дадени со Пример 2.3 ја имаме табелата со честоти Табела 2.5 од каде пресметуваме дека

$$\begin{aligned}\overline{x} &= \frac{1}{n}\sum_{i=1}^{r} f_i \overline{a}_i = \\ &= \frac{1}{70}(3 \cdot 151.25 + 6 \cdot 157.75 + 17 \cdot 164.25 + 16 \cdot 170.75 + 11 \cdot 177.25 + \\ &\quad +15 \cdot 183.75 + 2 \cdot 190.25) = 171.5857.\end{aligned}$$

Аритметичката средина има улога на "тежиште" на распределбата на податоците. Кај симетричните или скоро симетрични распределби, таа се наоѓа на средината или многу близу до средината на интервалот на простирање на податоците. Додека кај искривените распределби таа е повеќе наклонета кон "опашката". Важно е да се напомене дека аритметичката средина е осетлива на отстапувањата од распределбата, односно екстремните вредности. Затоа, како мерка за центрираност на податоците се земаат и други параметри кои би дале подобра слика за центарот на податоците.

2) Медијана и мода

Друг показател за локацијата на податоците е **медијаната** која се означува со m и се дефинира како средина на подредената низа од статистички податоци $x'_1 \leq x'_2 \leq ... \leq x'_n$. Имено важи

$$m = \begin{cases} \frac{1}{2}(x'_{\frac{n}{2}} + x'_{\frac{n}{2}+1}) & \text{, за } n \text{ парно} \\ x'_{\frac{n+1}{2}} & \text{, за } n \text{ непарно} \end{cases}$$

Својство 2.3. *Збирот на апсолутните вредности на сите отстапувања на податоците од медијаната е помал од збирот на апсолутните вредности на сите отстапувања на податоците од било кој друг број $c \in \mathbb{R}$ т.е.*

$$\sum_{i=1}^{n} |x_i - m| \leq \sum_{i=1}^{n} |x_i - c|.$$

Равенство се достигнува за $c \in [x'_{\frac{n}{2}}, x'_{\frac{n}{2}+1}]$, за n парно, односно за $c = m = x'_{\frac{n+1}{2}}$, за n непарно.

Ирена Стојковска

Медијаната претставува точка на бројната оска која даденатa низа од податоци ја дели на два еднаквобројни дела, односно лево и десно од неа има еднаков број на податоци. На вредноста на медијаната m влијаат само средните податоци, таа нема да се измени доколку на пример најмалиот од податоците (во случај кога $n \geq 3$) произволно се смали, или најголемиот податок произволно се зголеми, што не еслучај со аритметичката средина.

Така на пример, за податоците дадени во Пример 2.2 со Табела 2.3 имаме дека $n = 128$. Бидејќи, $x'_{\frac{128}{2}} = x'_{64} = 4$ и $x'_{\frac{128}{2}+1} = x'_{65} = 4$ заклучуваме дека медијаната $m = \frac{1}{2}(x'_{64} + x'_{65}) = \frac{1}{2}(4+4) = 4$. За потсетување, за аритметичката средина добивме $\overline{x} = 3,59375$. Разликата во вредностите на аритметичката средина и медијаната обично дава оценка за искривеноста на распределбата.

Во случај на **податоци групирани во интервали** (нумеричко непрекинато обележје), Табела 2.4, пресметувањето на медијаната е малку поразлично. Најнапред се наоѓа местото на медијаната според формулата $n_m = (n-1)/2+1$, каде n е вкупниот број на податоци. Потоа се наоѓа подинтервалот $(a_{j-1}, a_j]$ кој го содржи податокот со тој реден број. Откако ќе се земат во предвид вкупниот број на податоци $n_{j-1} = f_1 + ... + f_{j-1}$ во подинтервалите пред подинтервалот $(a_{j-1}, a_j]$ и вкупниот број на податоци $n_j = f_1 + ... + f_{j-1} + f_j$ во подинтервалите пред и заедно со подинтервалот $(a_{j-1}, a_j]$, медијаната се пресметува според формулата

$$m = a_{j-1} + (n_m - n_{j-1}) \cdot \frac{a_j - a_{j-1}}{f_j}.$$

На пример, за податоците дадени во Пример 2.3 со Табела 2.5 имаме дека $n = 70$, па медијаната се наоѓа на $n_m = (70-1)/2+1 = 35.5$-тото место. Од колоната со кумулативни честоти согледуваме дека податокот кој се наоѓа на 35.5-тото место е во 4-тиот подинтервал $(a_3, a_4] = (167.5, 174.0]$, значи $j = 4$. Бидејќи $n_3 = 26$ и $n_4 = 42$, имаме дека медијаната е $m = 167.5 + (35.5 - 26) \cdot \frac{174.0-167.5}{16} = 171.359375$. За потсетување, за аритметичката средина добивме $\overline{x} = 171.5857$. Близината на аритметичката стредина и медијаната се показател на симетричноста на распределбата на податоците.

И аритметичката средина и медијаната нема слисла да се пресметуваат кај категориските податоци. Кај нив оценка за центарот на распределбата се дава со параметарот **мода** кој се дефинира како податокот со најголема честота. Модата може да не постои и не мора да е единствена.

Ирена Стојковска

Кај категориските обележја таа одговара на податокот кој има највисок столб во столбестиот график, додека кај нумеричките обележја модата одговара на податокот каде се наоѓа највисокиот врв на полигонот, односно хистограмот (кај податоците групирани во интервали тоа е средината на интервалот на кој одговара највисокиот правоаголник од хистограмот).

Така на пример, за податоците дадени во Пример 2.1, мода е кавеавата боја на коса (има најголема честота, 62), за податоците дадени во Пример 2.2, мода е оценката 4 (има најголема честота, 50) и за податоците дадени во Пример 2.3, мода е средината на третиот интервал, односно висината 164.25 (имено интервалот $(161.0, 167.5]$ ја содржи најголемата честота 17).

3) Обсег, квартали, процентили и правоаголен дијаграм

Како мерка за простирање на нумеричките податоци се пресметува **обсегот** или **рангот** на статистичките податоци, кој се дефинира како разлика меѓу најголемата и најмалата вредност на низта од податоци $x_1, x_2, ..., x_n$, односно

$$R = \max\{x_1, x_2, ..., x_n\} - \min\{x_1, x_2, ..., x_n\}.$$

Главниот недостаток на обсегот при анализа на распредлбата на податоците е тоа што зависи само од екстремните вредности, и средината на податоците не влијае на големината на обсегот. Затоа, се пресметуваат други праметри кои ја дооформуваат сликата за простирањето на податоците, како што се **кварталите** и **процентилите**.

Нека е дадена подредената низа од статистички податоци $x'_1 \leq x'_2 \leq ... \leq x'_n$. Првиот квартал Q_1 е онаа вредност на податокот така да четвртина (25%) од податоците се наоѓаат лево од него, а три четвртини десно од него, вториот квартал $Q_2 = m$ се поклопува со вредноста на медијаната, додека третиот квартал Q_3 е онаа вредност на податокот така да три четвртини (75%) од податоците се наоѓаат лево од него, а четвртина десно од него.

Ако n е вкупниот број на податоци, тогаш за наоѓање на кварталите Q_1 и Q_3 важат следните формули

$$Q_1 = \begin{cases} \frac{1}{4}(x'_{\frac{n}{4}} + 3x'_{\frac{n}{4}+1}) & , n = 4k \\ x'_{\frac{n+3}{4}} & , n = 4k+1 \\ x'_{\frac{n+2}{4}} & , n = 4k+2 \\ x'_{\frac{n+1}{4}} & , n = 4k+3 \end{cases} \qquad Q_3 = \begin{cases} \frac{1}{4}(3x'_{\frac{3n}{4}} + x'_{\frac{3n}{4}+1}) & , n = 4k \\ x'_{\frac{3n+1}{4}} & , n = 4k+1 \\ x'_{\frac{3n+2}{4}} & , n = 4k+2 \\ x'_{\frac{3n+3}{4}} & , n = 4k+3 \end{cases}$$

Ирена Стојковска

На сличен начин се дефинираат и процентилите. Имено p-тиот процентил е онаа вредност на податокот така да $p\%$ од податоците се наоѓаат лево од него, а $(100-p)\%$ десно од него. Така првиот квартал Q_1 е 25-ти процентил, вториот квартал $Q_2 = m$ е 50-ти процентил и третиот квартал Q_3 е 75-ти процентил.

Кај **податоци групирани во интервали**, p-тиот процентил се наоѓа на $(n-1)\cdot p\%+1$-то место, каде n е вкупниот број на податоци. Вредноста на податокот на тоа место се бара слично како и вредноста на медијаната кај ваквиот тип на податоци.

За да се добие подобра слика на распределбата на нумеричките податоците обично се комбинираат **петте карактеристични броеви**: минимумот, првиот квртал Q_1, медијаната m, третиот квартал Q_3 и максимумот. Графичкиот приказ на овие броеви е со **правоаголен дијаграм** (box plot), така што централниот правоаголник е помеѓу кварталите Q_1 и Q_3, линијата во правоаголникот одговара на медијаната m, додека крајните линии одговараат на минималната, односно максималната вредност (Слика 2.7).

Правоаголните дијаграми се користат главно за истовремено споредување на повеќе распределби на податоци. Со нив може да се процени симетричноста на распределбата.

Така на пример, за податоците дадени во Пример 2.2, имаме дека најмалиот податок е $x_{\min} = 1$, најголемиот податок е $x_{\max} = 5$, значи обсегот е $R = 5 - 1 = 4$, и бидејќи $n = 128$, вредностите на првиот и третиот квартал се

$$Q_1 = \frac{1}{4}(x'_{\frac{128}{4}} + 3x'_{\frac{128}{4}+1}) = \frac{1}{4}(x'_{32} + 3x'_{33}) = \frac{1}{4}(3 + 3\cdot 3) = 3 \text{ и}$$

$$Q_3 = \frac{1}{4}(3x'_{\frac{3\cdot 128}{4}} + x'_{\frac{3\cdot 128}{4}+1}) = \frac{1}{4}(3x'_{96} + x'_{97}) = \frac{1}{4}(3\cdot 4 + 4) = 4$$

соодветно. Приказот на соодветниот правоаголен дијаграм е даден со Слика 2.7а).

Додека за податоците дадени во Пример 2.3 групирани во интервали, за најмалиот податок се зема левата граница на најлевиот подинтервал, односно $x_{\min} = 148.0$, а за најголемиот податок се зема десната граница на најдесниот подинтервал, односно $x_{\max} = 193.5$, значи обсегот е $R = 193.5 - 148.0 = 45.5$. Бидејќи $n = 70$, првиот квартал Q_1 е вредноста на податоците која се наоѓа на $(70-1)\cdot 0.25+1 = 18.25$-тото место, значи некоја вредност од третиот подинтервал $(161.0, 167.5]$. Имено, слично како и

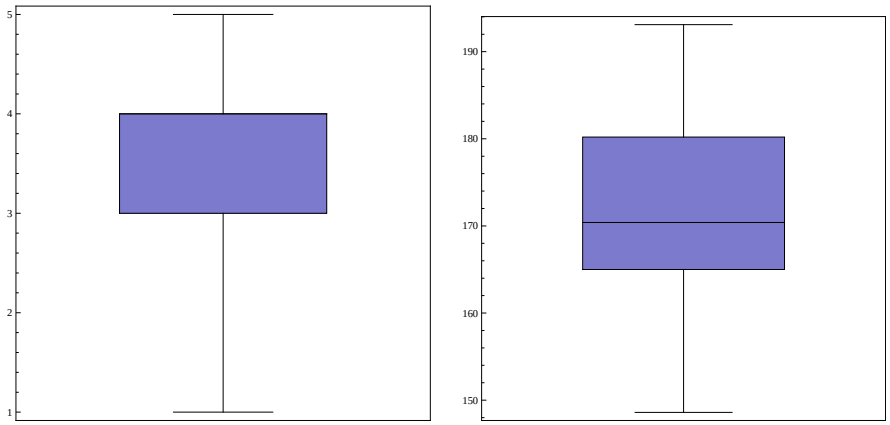
медијаната, за вредноста на Q_1 се наоѓа според

$$Q_1 = 161.0 + (18.25 - 9) \cdot \frac{167.5 - 161.0}{17} = 164.53676.$$

Слично добиваме дека третиот квартал Q_3 е вредноста на $(70-1)\cdot 0.75+1 = 52.75$-тото место, значи е вредност од петиот подинтервал $(174.0, 180.5]$ или

$$Q_3 = 174.0 + (52.75 - 42) \cdot \frac{180.5 - 174.0}{11} = 180.35227.$$

Приказот на соодветниот правоаголен дијаграм е даден со Слика 2.7б).



(а) (б)

Слика 2.7: (а) Правоаголен дијаграм за распределбата на податоците дадени во Пример 2.2, (б) Правоаголен дијаграм за распределбата на податоците дадени во Пример 2.3

4) Дисперзија или варијанса на податоците, стандардна девијација

Најчеста користена мерка за расејување на податоците е **дисперзијата** или **варијансата на податоците**, која го мери расејувањето преку оддалеченоста на податоците од нивната аритметичка средина. За дадена низа податоци $x_1, x_2, ..., x_n$, ако со $\overline{x}$ ја означиме аритметичката средина, тогаш дисперзијата на податоците се дефинира како

$$\overline{s}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2. \tag{2.2}$$

Може да се каже и дека дисперзијата $\overline{s}^2$ е просечно квадратно отстапување од аритметичката средина. Величината $\overline{s} = \sqrt{\overline{s}^2}$ се нарекува **стандардна**

девијација или **стандардно отстапување** на низата статистички податоци. Ако $\overline{s} = 0$ тоа значи дека нема никакво расејување на податоците, односно дека станува збор за константана низа од податоци.

Од Својство 2.2 заклучуваме дека важи

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

за произволен $c \in \mathbb{R}$, односно дека расејувањето на податоците мерено со квадратните отстапувања е минимално ако се земат во предвид отстапувањата на податоците од аритметичката средина и се мери со дисперзијата на податоците.

Следното Својство 2.4 го опишува попрецизно значењето на стандардната девијација s како параметар на расејување.

Својство 2.4. *Барем 89% од сите податоци од низата статистички податоци $x_1, x_2, ..., x_n$ се наоѓаат внатре во интервалот $[\overline{x} - 3\overline{s}, \overline{x} + 3\overline{s}]$, каде $\overline{x}$ е аритметичката средина на податоците и $\overline{s}$ е стандардната девијација на податоците.*

Доказ. За $k \geq 0$ имаме

$$\begin{aligned} n\overline{s}^2 &= \sum_{i=1}^{n}(x_i - \overline{x})^2 = \\ &= \sum_{x_i < \overline{x} - k\overline{s}}(x_i - \overline{x})^2 + \sum_{\overline{x} - k\overline{s} \leq x_i \leq \overline{x} + k\overline{s}}(x_i - \overline{x})^2 + \sum_{x_i > \overline{x} + k\overline{s}}(x_i - \overline{x})^2 \geq \\ &\geq \sum_{x_i < \overline{x} - k\overline{s}}(x_i - \overline{x})^2 + \sum_{x_i > \overline{x} + k\overline{s}}(x_i - \overline{x})^2 \geq \\ &\geq \sum_{x_i < \overline{x} - k\overline{s}} k^2\overline{s}^2 + \sum_{x_i > \overline{x} + k\overline{s}} k^2\overline{s}^2 = \sum_{|x_i - \overline{x}| > k\overline{s}} k^2\overline{s}^2. \end{aligned}$$

Нека m е бројот на податоци од низата $x_1, x_2, ..., x_n$ за кои важи $|x_i - \overline{x}| > k\overline{s}$, тогаш

$$n\overline{s}^2 \geq \sum_{|x_i - \overline{x}| > k\overline{s}} k^2\overline{s}^2 = mk^2\overline{s}^2.$$

Ако $\overline{s} > 0$, тогаш последното неравенство е еквивалентно на

$$m \leq \frac{1}{k^2}n.$$

Ирена Стојковска

За $k = 3$ добиваме дека $m \leq \frac{1}{9}n \approx 11\%n$, што значи дека барем 89% од сите податоци се внатре во интервалот $[\overline{x} - 3\overline{s}, \overline{x} + 3\overline{s}]$, што требаше да се докаже.

Слично се добива дека барем 75% од податоците се внатре во интервалот $[\overline{x} - 2\overline{s}, \overline{x} + 2\overline{s}]$, односно барем 93% од податоците се внатре во интервалот $[\overline{x} - 4\overline{s}, \overline{x} + 4\overline{s}]$. ∎

При нумерички пресметувања најчесто се користи следниот облик на формулата за дисперзија на податоците, тоа е

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

Кај податоци дадени во табела со честоти (види Табела 2.1), дисперзијата се пресметува според

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

Дисперзијата кај податоци групирани во интервали (види Табела 2.4) се пресметува според

$$\overline{s}^2 = \frac{1}{n}\sum_{i=1}^{r} f_i(\overline{a}_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{r} f_i \overline{a}_i^2 - \overline{x}^2,$$

каде $\overline{a}_i$ е средина на интервалот $[a_{i-1}, a_i]$, $i = 1, 2, ..., r$.

За податоците дадени со Пример 2.2, дисперзијата изнесува

$$\overline{s}^2 = \frac{1}{128}(5 \cdot 1^2 + 18 \cdot 2^2 + 28 \cdot 3^2 + 50 \cdot 4^2 + 27 \cdot 5^2) - 3,59375^2 = 1,17871,$$

од каде за стандардната девијација имаме $s = 1,08568$.

Додека пак, за податоците дадени со Пример 2.3, дисперзијата изнесува

$$\begin{aligned}\overline{s}^2 \;=\; & \frac{1}{70}(3 \cdot 151.25^2 + 6 \cdot 157.75^2 + 17 \cdot 164.25^2 + 16 \cdot 170.75^2 + 11 \cdot 177.25^2 + \\ & +15 \cdot 183.75^2 + 2 \cdot 190.25^2) - 171.5857^2 = 94.067198,\end{aligned}$$

па стандардната девијација е $s = 9.6988$.

Ирена Стојковска

2.4 Дводимензионални обележја

1) Табела на контингенција, графички приказ на дводимензионалната распределба на честотите, график на распределба на податоците

При проучување на некоја појава, често се разгледуваат повеќе од една карактеристика и се воспоставува зависност меѓу нив. Нека X и Y се две обележја кои се разгледуваат истовремено. Тогаш, при извршени n мерења се добива низа од **статистички дводимензионални податоци**, односно низа од подредени парови $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, каде првата компонента одговара на обележјето X, а втората на обележјето Y, односно податоците одговаарат на **дводимензионалното статистичко обележје** (X, Y).

Нека X и Y се дискретни обележја. Да ги означиме со $a_1 < a_2 < ... < a_r$ различните вредности на обележјето X опфатени со првата координата на податоците $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ и да ги означиме со $b_1 < b_2 < ... < b_s$ различните вредности на обележјето Y обфатени со втората координата на истите податоци. Тогаш, табелата која ги содржи честотите $f_{i,j}$ на појавување на вредноста (a_i, b_j), $i = 1, 2, ..., r$, $j = 1, 2, ..., s$ во дводимензионалните податоци се нарекува **табела на контингенција** (Табела 2.6). При тоа важи дека вкупниот збир од честоти е еднаковнна вкупниот број на изведени мерења

$$\sum_{i=1}^{r} \sum_{j=1}^{s} f_{ij} = n.$$

X \ Y	b_1	b_2	$\cdots$	b_j	$\cdots$	b_s	Σ
a_1	f_{11}	f_{12}	$\cdots$	f_{1j}	$\cdots$	f_{1s}	g_1
a_2	f_{21}	f_{22}	$\cdots$	f_{2j}	$\cdots$	f_{2s}	g_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
a_i	f_{i1}	f_{i2}	$\cdots$	f_{ij}	$\cdots$	f_{is}	g_i
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
a_r	f_{r1}	f_{r2}	$\cdots$	f_{rj}	$\cdots$	f_{rs}	g_r
Σ	h_1	h_2	$\cdots$	h_j	$\cdots$	h_s	n

Табела 2.6: Табела на контингенција на честотите

Табела 2.6 се дополнува и со честотите на секоја од вредностите a_i и b_j,

$i = 1, 2, ..., r,\ j = 1, 2, ..., s$ поединечно, односно се пресметуваат броевите

$$g_i = \sum_{j=1}^{s} f_{ij},\ i = 1, 2, ..., r, \quad h_j = \sum_{i=1}^{r} f_{ij},\ j = 1, 2, ..., s.$$

X \ Y	b_1	b_2	$\cdots$	b_j	$\cdots$	b_s	Σ
a_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$	p_{1s}	q_1
a_2	p_{21}	p_{22}	$\cdots$	p_{2j}	$\cdots$	p_{2s}	q_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
a_i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$	p_{is}	q_i
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
a_r	p_{r1}	p_{r2}	$\cdots$	p_{rj}	$\cdots$	p_{rs}	q_r
Σ	r_1	r_2	$\cdots$	r_j	$\cdots$	r_s	1

Табела 2.7: Табела на контингенција на релативните честоти

Ако во табелата на контингенција на местото од честотите f_{ij} ги ставиме релативните честоти

$$p_{ij} = \frac{f_{ij}}{n},\ i = 1, 2, ..., r,\ j = 1, 2, ..., s,$$

ја добиваме Табела 2.7. При тоа за релативните честоти важи

$$\sum_{i=1}^{r} \sum_{j=1}^{s} p_{ij} = 1.$$

Соодветните релативни честоти кои одговараат на вредностите a_i и b_j, $i = 1, 2, ..., r,\ j = 1, 2, ..., s$ поединечно се дадени со формулите

$$q_i = \sum_{j=1}^{s} p_{ij},\ i = 1, 2, ..., r, \quad r_j = \sum_{i=1}^{r} p_{ij},\ j = 1, 2, ..., s.$$

Пример 2.4. Во една работна организација спроведено е истражување во врска со оптеретеноста на вработените со работни обврски. За секој од 165-те вработени забележани се просечниот број на слободни денови во текот на еден месец кои работниикот ги зел надвор од деновите предвидени за годишен одмор (обележје X) и просечниот број на денови месечно кога тој работел прекувремено (обележје Y). На тој начин добиени се следните дводимензионални податоци:

(0, 0), (0, 0), (0, 0), (0, 0), (0, 1), (0, 1), (0, 1), (0, 1), (0, 1), (0, 1),
(0, 1), (0, 1), (0, 1), (0, 1), (0, 1), (0, 1), (0, 1), (0, 1), (0, 1), (0, 1),
(0, 1), (0, 1), (0, 1), (0, 1), (0, 1), (0, 1), (0, 1), (0, 2), (0, 2), (0, 2),
(0, 2), (0, 2), (0, 2), (0, 2), (0, 2), (0, 2), (0, 2), (0, 2), (0, 2), (1, 2),
(1, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2),
(1, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2),
(1, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2),
(1, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2), (1, 2),
(1, 2), (1, 2), (1, 2), (1, 3), (1, 3), (1, 3), (1, 3), (1, 3), (1, 3), (1, 3),
(1, 3), (1, 3), (1, 3), (1, 3), (1, 3), (1, 3), (1, 3), (1, 3), (1, 3), (1, 3),
(1, 3), (1, 3), (1, 3), (1, 3), (1, 3), (1, 3), (1, 3), (1, 3), (1, 3), (1, 3),
(1, 3), (1, 3), (1, 3), (1, 3), (1, 3), (1, 3), (1, 3), (2, 3), (2, 3), (2, 3),
(2, 3), (2, 3), (2, 3), (2, 3), (2, 3), (2, 3), (2, 3), (2, 3), (2, 3), (2, 3),
(2, 3), (2, 3), (2, 3), (2, 3), (2, 3), (2, 3), (2, 3), (2, 3), (2, 4), (2, 4),
(2, 4), (2, 4), (2, 4), (2, 4), (2, 4), (2, 4), (2, 4), (2, 4), (2, 4), (2, 4),
(2, 4), (2, 4), (2, 4), (2, 4), (2, 4), (2, 4), (2, 4), (2, 4), (2, 5),
(2, 5), (2, 5), (2, 5), (2, 5), (2, 5)

Табела 2.8 дава приказ на честотите на вредностите на дводимензионалното обележје (X, Y) опфатени со добиените податоци. Додека во Табела 2.9 дадени се соодветните релативни честоти.

X \ Y	0	1	2	3	4	5	Σ
0	4	23	12	0	0	0	39
1	0	0	44	34	0	0	78
2	0	0	0	21	21	6	48
Σ	4	23	56	55	21	6	165

Табела 2.8: Табела на контингенција на честотите за обележјето (X, Y), каде X е просечен број на слободни денови и Y е просечен број на денови со прекувремена работа

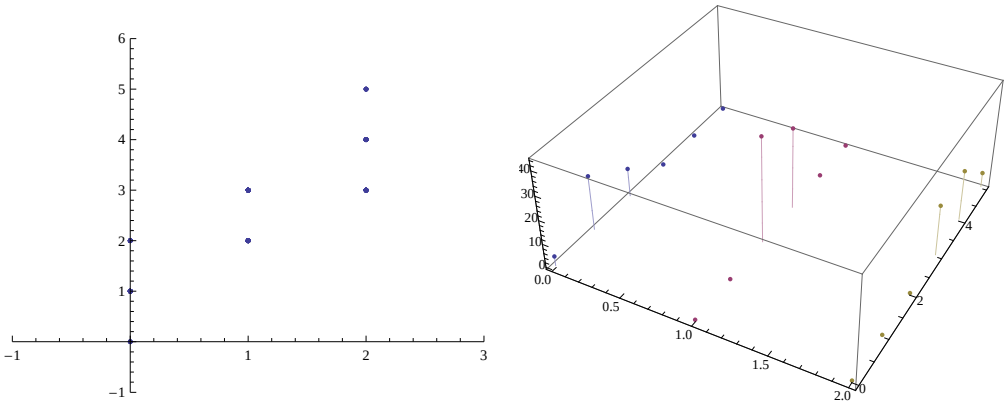
Соодветниот **график на распредедбата на податоците** (scatter plot) е даден со Слика 2.8а). Додека **графичкиот приказ на дводимензионалната распределба на честотите** дадени со Табела 2.8 е на Слика 2.8б).

2) Криви на регресија, прави на регресија, коефициент на корелација, степен на статистичка зависност

Статистичките податоци $x_1, x_2, ..., x_n$ за обележјето X и статистичките податоци $y_1, y_2, ..., y_n$ за обележјето Y превземени од дводимензионалните статистички податоци $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ може да се обработуваат

X \ Y	0	1	2	3	4	5	Σ
0	4/165	23/165	12/165	0	0	0	39/165
1	0	0	44/165	34/165	0	0	78/165
2	0	0	0	21/165	21/165	6/165	48/165
Σ	4/165	23/165	56/165	55/165	21/165	6/165	1

Табела 2.9: Табела на контингенција на релативните честотите за обележјето (X, Y), каде X е просечен број на слободни денови и Y е просечен број на денови со прекувремена работа



(а) (б)

Слика 2.8: (а) График на распределба на податоците дадени со Табела 2.8, (б) Графички приказ на честотите на податоците дадени со Табела 2.8

и посебно. Значи, ако честотите на дводимензионалните податоци се дадени со Табела 2.6, тогаш нивните аритметички средини се

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{r} g_i a_i,$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{s} h_i b_i,$$

односно нивните дисперзии се

$$\overline{s}_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 = \frac{1}{n}\sum_{i=1}^{r} g_i(a_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{r} g_i a_i^2 - \overline{x}^2,$$

$$\overline{s}_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2 = \frac{1}{n}\sum_{i=1}^{s} h_i(b_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{s} h_i b_i^2 - \overline{y}^2.$$

Тогаш, точката $(\overline{x}, \overline{y})$ може да се интерпретира како "тежиште" на соодветната распределба на дводимензионалните податоци, и се нарекува **средина на дводимензионалната распределба на честоти**.

За податоците дадени со Пример 2.4, се добива

$$\overline{x} = 1.05455, \overline{s}_x^2 = 0.524298, \overline{s}_x = 0.724084,$$

$$\overline{y} = 2.50909, \overline{s}_y^2 = 1.14689, \overline{s}_y = 1.07093.$$

Зависноста меѓу обележјата X и Y се толкува преку вредностите на одредени параметри и нивната геометриска интерпретација. Имено, може да сметаме дека со j-тата основна колона од Табела 2.6 се определува условна распределба на честоти за оние податоци на обележјето X кај кои обележјето Y прима вредност b_j, $j = 1, 2, ..., s$, односно ја имаме условната распределба на честоти дадена со Табела 2.10.

Вредност на X при услов $Y = b_j$	a_1	a_2	$\cdots$	a_i	$\cdots$	a_r	Σ
Честота	f_{1j}	f_{2j}	$\cdots$	f_{ij}	$\cdots$	f_{rj}	h_j

Табела 2.10: Условна распределба на честоти на податоците од обележјето X кој кои обележјето $Y = b_j$

Аритметичката средина на податоците од условните распределби на честоти дадени со Табела 2.10 се означува со $\overline{x}(b_j)$ и се пресметува според

$$\overline{x}(b_j) = \frac{1}{h_j} \sum_{i=1}^{r} f_{ij} a_i, \quad j = 1, 2, ..., s.$$

На сличен начин се добиваат и условните распределби на честоти на оние податоци од обележјето Y кај кои обележјето X прима вредност a_i, $i = 1, 2, ..., r$, па соодветните аритметички средини се пресметуваат според

$$\overline{y}(a_i) = \frac{1}{g_i} \sum_{j=1}^{s} f_{ij} b_j, \quad i = 1, 2, ..., r.$$

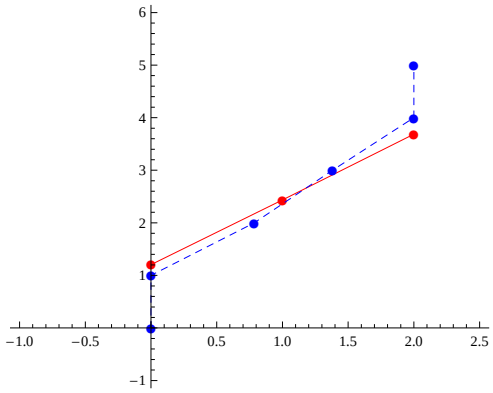
Тогаш, функциите $a_i \mapsto \overline{y}(a_i)$, $i = 1, 2, ..., r$ и $b_j \mapsto \overline{x}(b_j)$, $j = 1, 2, ..., s$ се нарекуваат **функции на регресија** на дадената низа статистички податоци за дводимензионалното обележје (X, Y). Првата функција $a_i \mapsto \overline{y}(a_i)$, $i = 1, 2, ..., r$ ја покажува зависноста на аритметичките средини на улсовните распределби на честоти во i-тата редица од Табела 2.6 од

вредностите a_i на обележјето X, додека втората функција $b_j \mapsto \overline{x}(b_j)$, $j = 1, 2, ..., s$ ја покажува зависноста на аритметичките средини на улсовните распределби на честоти во j-тата колона од Табела 2.6 од вредностите b_j на обележјето Y. Графичките прикази на овие функции ги определуваат **кривите на регресија** на аритметичките средини на условните распределби во зависност од вредностите на обележјата во условот.

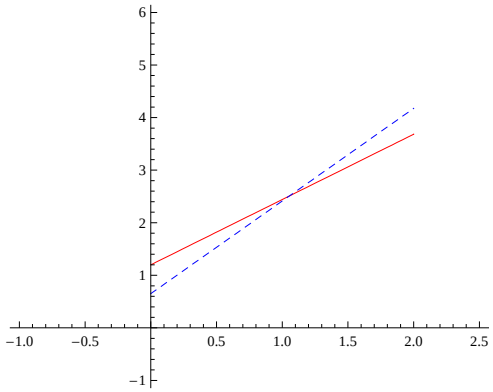
Така, за податоците дадени во Пример 2.4 со Табела 2.8 имаме

$$\begin{aligned}
&\overline{x}(0) = \tfrac{1}{4}(4 \cdot 0 + 0 \cdot 1 + 0 \cdot 2) = 0,\\
&\overline{x}(1) = \tfrac{1}{23}(23 \cdot 0 + 0 \cdot 1 + 0 \cdot 2) = 0,\\
&\overline{x}(2) = \tfrac{1}{56}(12 \cdot 0 + 44 \cdot 1 + 0 \cdot 2) = 0,7857,\\
&\overline{x}(3) = \tfrac{1}{55}(0 \cdot 0 + 34 \cdot 1 + 21 \cdot 2) = 1,3818,\\
&\overline{x}(4) = \tfrac{1}{21}(0 \cdot 0 + 0 \cdot 1 + 21 \cdot 2) = 2,\\
&\overline{x}(5) = \tfrac{1}{6}(0 \cdot 0 + 0 \cdot 1 + 6 \cdot 2) = 2,\\
&\overline{y}(0) = \tfrac{1}{39}(4 \cdot 0 + 23 \cdot 1 + 12 \cdot 2 + 0 \cdot 3 + 0 \cdot 4 + 0 \cdot 5) = 1,2051,\\
&\overline{y}(1) = \tfrac{1}{78}(0 \cdot 0 + 0 \cdot 1 + 44 \cdot 2 + 34 \cdot 3 + 0 \cdot 4 + 0 \cdot 5) = 2,4359,\\
&\overline{y}(2) = \tfrac{1}{48}(0 \cdot 0 + 0 \cdot 1 + 0 \cdot 2 + 21 \cdot 3 + 21 \cdot 4 + 6 \cdot 5) = 3,6875,
\end{aligned}$$

што значи, на пример, дека месечниот просечен број на земени слободни денови за вработените кои работеле во просек по 4 дена во месецот прекувремено е 2 дена (од $\overline{x}(4) = 2$). Соодветните криви на регресија дадени се со Слика 2.9.



Слика 2.9: График на кривите на регресија за податоците дадени со Табела 2.8 (црвената полна линија е за аритметичките средини на условните распределби на честотите на Y во зависност од X, сината испрекината линија е за аритметичките средини на условните распределби на честотите на X во зависност од Y)



Слика 2.10: График на правите на регресија за податоците дадени со Табела 2.8 (црвената полна линија е за правата на регресија на податоците за Y во зависност од податоците за X, сината испрекината линија е правата на регресија на податоците за X во зависност од податоците за Y)

Апроксимирањето на кривите на регресија со прави е со помош на методот на најмали квадрати. Така, при апроксимација на кривата $y = \overline{y}(a_i)$ со права $y = Ax+B$, коефициентите A и B се наоѓаат за да се минимизира сумата

$$S = \sum_{i=1}^{r}\sum_{j=1}^{s}(Aa_i + B - b_j)^2 f_{ij}.$$

Имено, се решава системот равенки

$$\frac{\partial S}{\partial A} = 0, \quad \frac{\partial S}{\partial B} = 0. \tag{2.3}$$

По спроведеното диференцирање системот (2.3) преминува во

$$\begin{cases} A\sum_{i=1}^{r}\sum_{j=1}^{s} a_i^2 f_{ij} + B\sum_{i=1}^{r}\sum_{j=1}^{s} a_i f_{ij} = \sum_{i=1}^{r}\sum_{j=1}^{s} a_i b_j f_{ij} \\ B\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} + A\sum_{i=1}^{r}\sum_{j=1}^{s} a_i f_{ij} = \sum_{i=1}^{r}\sum_{j=1}^{s} b_j f_{ij} \end{cases} \tag{2.4}$$

Се зема во предвид дека

$$\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} = n, \ \sum_{i=1}^{r}\sum_{j=1}^{s} a_i f_{ij} = n\overline{x}, \ \sum_{i=1}^{r}\sum_{j=1}^{s} b_j f_{ij} = n\overline{y},$$

$$\sum_{i=1}^{r}\sum_{j=1}^{s} a_i^2 f_{ij} = n(\overline{s}_x^2 + \overline{x}^2).$$

Воведуваме нова ознака

$$\overline{s}_{xy} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}(a_i - \overline{x})(b_j - \overline{y})f_{ij} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}a_i b_j f_{ij} - \overline{x}\,\overline{y},$$

од каде добиваме дека

$$\sum_{i=1}^{r}\sum_{j=1}^{s}a_i b_j f_{ij} = n(\overline{s}_{xy} + \overline{x}\,\overline{y}).$$

Тогаш системот (2.4) преминува во

$$\begin{cases} An(\overline{s}_x^2 + \overline{x}^2) + Bn\overline{x} = n(\overline{s}_{xy} + \overline{x}\,\overline{y}) \\ Bn + An\overline{x} = n\overline{y} \end{cases} \tag{2.5}$$

односно во

$$\begin{cases} A(\overline{s}_x^2 + \overline{x}^2) + B\overline{x} = (\overline{s}_{xy} + \overline{x}\,\overline{y}) \\ B + A\overline{x} = \overline{y} \end{cases} \tag{2.6}$$

Од втората равенка во (2.6) се увидува дека точката $(\overline{x}, \overline{y})$ лежи на правата $y = Ax + B$. Со решавање на овој систем (2.6) се добива дека

$$A = \frac{\overline{s}_{xy}}{\overline{s}_x^2}, \quad B = \overline{y} - \frac{\overline{s}_{xy}}{\overline{s}_x^2}\overline{x}.$$

За да покажеме дека овие две вредности за A и B ја минимизираат сумата S, потребно е да ги испитаме вторите парцијални изводи, па имаме

$$\frac{\partial^2 S}{\partial A^2} = 2\sum_{i=1}^{r}\sum_{j=1}^{s}a_i^2 f_{ij} = 2n(\overline{s}_x^2 + \overline{x}^2),$$

$$\frac{\partial^2 S}{\partial A \partial B} = 2\sum_{i=1}^{r}\sum_{j=1}^{s}a_i f_{ij} = 2n\overline{x},\ \frac{\partial^2 S}{\partial B^2} = 2\sum_{i=1}^{r}\sum_{j=1}^{s}f_{ij} = 2n.$$

Потоа, бидејќи за секои A и B важи

$$\frac{\partial^2 S}{\partial A^2} \geq 0,$$

$$\frac{\partial^2 S}{\partial A^2}\cdot\frac{\partial^2 S}{\partial B^2} - (\frac{\partial^2 S}{\partial A \partial B})^2 = 4n^2(\overline{s}_x^2 + \overline{x}^2) - (2n\overline{x})^2 = 4n^2\overline{s}_x^2 \geq 0,$$

имаме дека S е конвексна функција по променливите A и B, што значи дека сумата S достигнува минимум во вредностите за A и B добиени со изедначување на првите парцијални изводи на нула. Па, добиваме

дека **правата на регресија** која најдобро ја апроксимира кривата на регресија $y = \overline{y}(a_i)$ е

$$y = \frac{\overline{s}_{xy}}{\overline{s}_x^2}x + \overline{y} - \frac{\overline{s}_{xy}}{\overline{s}_x^2}\overline{x},$$

или во среден облик

$$y - \overline{y} = \frac{\overline{s}_{xy}}{\overline{s}_x^2}(x - \overline{x}). \tag{2.7}$$

На сличен начин, со минимизирање на сумата

$$T = \sum_{i=1}^{r}\sum_{j=1}^{s}(Cb_j + D - a_i)^2 f_{ij},$$

се добива дека

$$C = \frac{\overline{s}_{xy}}{\overline{s}_y^2}, \quad D = \overline{x} - \frac{\overline{s}_{xy}}{\overline{s}_y^2}\overline{y},$$

па правата на регресија $x = Cy+D$ која најдобро ја апроксимира кривата на регресија $x = \overline{x}(b_j)$ во нејзиниот среден облик е

$$x - \overline{x} = \frac{\overline{s}_{xy}}{\overline{s}_y^2}(y - \overline{y}). \tag{2.8}$$

Со замена на добиените вредности за коефициентите A и B во сумата S, односно вредностите на C и D во сумата T се добиваат минималните вредности на овие суми кои изнесуваат

$$S_{\min} = ns_y^2\Big(1 - \frac{\overline{s}_{xy}^2}{\overline{s}_x^2\overline{s}_y^2}\Big), \quad T_{\min} = n\overline{s}_x^2\Big(1 - \frac{\overline{s}_{xy}^2}{\overline{s}_x^2\overline{s}_y^2}\Big).$$

Бројот

$$r = \frac{\overline{s}_{xy}}{\overline{s}_x\overline{s}_y}, \tag{2.9}$$

се нарекува **коефициент на корелација**. Изразени преку него минималните вредности на сумите S и T се

$$S_{\min} = n\overline{s}_y^2(1 - r^2), \quad T_{\min} = n\overline{s}_x^2(1 - r^2),$$

и бидејќи S и T се ненегативни величини добиваме дека $1-r^2 \geq 0$, односно дека $-1 \leq r \leq 1$.

Понатаму, имајќи ги во предвид равенките на правите на регресија (2.7) и (2.8) за аголот φ меѓу овие две прави имаме

$$\tan\varphi = \frac{1 - r^2}{r}\,\frac{\overline{s}_x\overline{s}_y}{\overline{s}_x^2 + \overline{s}_y^2}. \tag{2.10}$$

Ирена Стојковска

Толкувањето на значењето на коефициентот на корелација е следното. Ако $r^2 = 1$, тогаш аголот φ меѓу правите на регресија е $\varphi = 0$ или $\varphi = \pi$, што значи дека правите се поклопуваат, и тогаш $S = T = 0$ што значи дека податоците $(x_1, y_1), ..., (x_n, y_n)$ лежат на заедничка права определена со равенката

$$y - \overline{y} = \frac{\overline{s}_y}{\overline{s}_x}(x - \overline{x}). \tag{2.11}$$

Тогаш велиме дека податоците за обележјето Y **линеарно зависат** од податоците за обележјето X (и обратно), т.е.

$$y_i = \frac{\overline{s}_y}{\overline{s}_x}(x_i - \overline{x}) + \overline{y},\ i = 1, 2, ..., n. \tag{2.12}$$

Ако $r = 0$, тогаш $\overline{s}_{xy} = 0$ и станува збор за **некорелирани статистички податоци** x_i и y_i, $i = 1, 2, ..., n$.

Во општ случај, ако $|r| < 0,5$, тогаш податоците x_i и y_i, $i = 1, 2, ..., n$ се **слабо корелирани**, ако $|r| \geq 0,5$, тогаш станува збор за **статистички сигнификантна корелација**.

Ако $r > 0$, тогаш имаме **позитивна корелација**, односно со зголемување на вредностите на едното обележје се зголемуваат вредностите на другото во рамките на дводомензионалните податоци, и соодветно за $r < 0$ имаме **негативна корелација**, односно со зголемување на вредностите на едното обележје се намалуваат вредностите на другото во рамките на дводомензионалните податоци.

Значи, коефициентот на корелација кој одговара на статаистичките податоци за дводимензионалното обележје (X, Y), го опишува степенот на афина зависност меѓу податоците за обелезјето X и податоците за обележјето Y.

На пример, за податоците дадени со Пример 2.4 од претходно добивме дека

$$\overline{x} = 1.05455, \overline{s}_x^2 = 0.524298, \overline{s}_x = 0.724084,$$

$$\overline{y} = 2.50909, \overline{s}_y^2 = 1.14689, \overline{s}_y = 1.07093.$$

Исто така се пресметува и дека

$$\overline{s}_{xy} = 0,651019.$$

Тогаш, равенките на правите на регресија се

$$y = 1,2417x + 1,19966, \quad x = 0,56764y - 0.369716,$$

Ирена Стојковска

а нивните графици дадени се со Слика 2.10. За коефициентот на корелација имаме

$$r = 0,839546,$$

што значи дека просечниот број на слободни денови во текот на еден месец кои вработениот ги користи надвор од деновите предвидени за одмор и просечниот број на денови во текот на еден месец во кои работникот работи прекувремено се позитивно корелирани (со зголемување на бројот на слободни дена кои ги зема работникот, се зголемува и бројот не денови кога останува да работи прекувремено). Исто така може да кажеме дека постои статистички сигнификантна корелација меѓу податоците за овие две обележја.

Претходно спомнавме дека кога коефициентот на корелација е $r = 0$ станува збор за некорелирани статистички податоци, што не значи дека не постои никаква статистичка поврзаност меѓу нив. Статистичката зависност се изразува преку **отстапувањето од статистичката независност** дефнирано со

$$f^2 = \frac{1}{n^2} \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(nf_{ij} - g_i h_j)^2}{g_i h_j} = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{f_{ij}^2}{g_i h_j} - 1, \tag{2.13}$$

каде f_{ij}, g_i, h_j се честотите превземени од Табела 2.6. За отстапувањето од статистичката независност важи релацијата

$$0 \leq f^2 \leq \min\{r, s\} - 1. \tag{2.14}$$

Вредноста $f = 0$ се достигнува тогаш кога важи

$$nf_{ij} = g_i h_j,\ i = 1, ..., r,\ j = 1, ..., s,$$

што значи дека условните распределби на честотите во редиците (соодветно и во колоните) меѓусебно се разликуваат само во одреден коефициент на пропорционалност, па според тоа може да се каже дека вредностите на едното обележје важно не влијаат на условната распределба на честотите на податоците за другото обележје, и велиме дека **податоците се статистички независни**. Додека, максималната вредност $f^2 = \min\{r, s\} - 1$ за $r \geq s$ се достигнува кога секоја редица од табелата на контингенција Табела 2.6 содржи точно една вредност различна од нула, и за $r \leq s$ се достигнува кога секоја колона од табелата на контингенција содржи точно една вредност различна од нула. Тоа значи дека меѓу податоците кои одговараат на обележјата X и Y постои **функционална зависност**, односно на секоја вредност a_i на обележето X и е придружена

само една вредност од обележјето Y (кога $r \geq s$), односно на секоја вредност b_j на обележето Y и е придружена само една вредност од обележјето X (кога $r \leq s$).

Доколку отстапувањето од статистичката независност го поделиме со неговата максимала можна вредност така добиениот количник се нарекува **степен на статистичката зависност** на податоците кои одговаат на обележјата X и Y соодветно, односно

$$o = f^2/(\min\{r, s\} - 1). \tag{2.15}$$

Степенот на статистичката зависност се изразува во проценти.

Важно е да се напомене дека вредностите на f^2 и o не заисат од вредностите кои ги примаат обележјата X и Y туку само од честотите од табелата на контингенција. Затоа, **параметрите f^2 и o може да се користат и при анализа на статистичката зависност и во случај на ненумерички (категориски) податоци.**

Врз основа на податоците дадени во Пример 2.4, при анализа на статистичката зависност меѓу податоците кои одговараат на обележјето X-просечен месечен број на слободни денови, и оние кои одговараат на обележјето Y-просечен месечен број на денови со прекумерна работа, добиваме дека

$$f^2 = 1,20047, \quad o = 0,600237 \approx 60\%,$$

што значи дека меѓу бројот на слободни денови и денови прекувремена работи постои статистичка зависност од 60%.

3) Непрекинати дводимензионални обележја

Ако обележјата X и Y се непрекинати, тогаш дводимензионалното обележје (X, Y) е исто така непрекинато, односно може да прима било која вредност од одреден производ на интервали. За да се состави табелата на контингенција за податоците $(x_1, y_1), ..., (x_n, y_n)$ кои одговараат на непрекинатото обележје (X, Y) најнапред треба се изврши групирање на податоците во интервали за секое од еднодимензионалните обележја X и Y, на пример во r, односно s подинтервали, со граници во точките $a_0 < a_1 < ... < a_r$ и $b_0 < b_1 < ... < b_s$ соодветно. Тогаш, честотата f_{ij} ќе го означува бројот на подредени парови од низата $(x_1, y_1), ..., (x_n, y_n)$ за кои важи да првата компонента е од i-тиот подинтервал за вредностите на X, т.е. подинтервалот $(a_{i-1}, a_i]$, а втората компонента е од j-тиот подинтервал за вредностите на Y, т.е. подинтервалот $(b_{j-1}, b_j]$. На тој начин се добива **табелата на контингенција на честотите** на податоците групирани во интервали, Табела 2.11.

Ирена Стојковска

	Y	$[b_0, b_1]$	$(b_1, b_2]$	$\cdots$	$(b_{j-1}, b_j]$	$\cdots$	$(b_{s-1}, b_s]$	
X	средина на интервал	$\overline{b}_1$	$\overline{b}_2$	$\cdots$	$\overline{b}_j$	$\cdots$	$\overline{b}_s$	Σ
$[a_0, a_1]$	$\overline{a}_1$	f_{11}	f_{12}	$\cdots$	f_{1j}	$\cdots$	f_{1s}	g_1
$(a_1, a_2]$	$\overline{a}_2$	f_{21}	f_{22}	$\cdots$	f_{2j}	$\cdots$	f_{2s}	g_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$(a_{i-1}, a_i]$	$\overline{a}_i$	f_{i1}	f_{i2}	$\cdots$	f_{ij}	$\cdots$	f_{is}	g_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$(a_{r-1}, a_r]$	$\overline{a}_r$	f_{r1}	f_{r2}	$\cdots$	f_{rj}	$\cdots$	f_{rs}	g_r
	Σ	h_1	h_2	$\cdots$	h_j	$\cdots$	h_s	n

Табела 2.11: Табела на контингенција на честотите на податоците групирани во интервали

Соодветната **табела на конингенција на релативните честоти** се добива кога на меатото од честотите f_{ij} се стават релативните честоти $p_{ij} = f_{ij}/n$, $i = 1, ..., r$, $j = 1, ..., s$, на местото од честотите g_i и h_j се стават релатицните честоти $q_i = \sum_{j=1}^{s} p_{ij}$ и $r_j = \sum_{i=1}^{r} p_{ij}$ соодветно. Графичкиот приказ на податоците дадени со табела на контингенција на честоти е со помош на **дводимензионален хистограм на честоти**, односно **дводимензионален хистограм на релативни честоти**.

Од дефинициите за $\overline{x}$, $\overline{y}$, s_x^2, s_y^2 и s_{xy} кои не зависат од честотите, заклучуваме дека правите на регресија и коефициентот на корелација може да се пресметаат и за податоци кои одговараат на дводимензионално непрекинато обележје без да бидат претходно групирани во интервали. За разлика од овие параметри, за пресметување на отстапувањето од статистичката зависност f^2 и степенот на статистичката зависност o потребни се честотите. И бидејќи групирањето во интервали не е еднозначно, користењето на параметрите f^2 и o при анализа на зависноста на податоците кои одговараат на непрекинати обележја се избегнува.

Знаејќи ја само распределбата на податоците во интервали дадена со Табела 2.11 вредностите $\overline{x}$, $\overline{y}$, s_x^2, s_y^2 и s_{xy} приближно може да се пресметаат според

$$\overline{x} \approx \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}\overline{a}_i f_{ij},\ \overline{y} \approx \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}\overline{b}_j f_{ij},$$
$$\overline{s}_x^2 \approx \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}\overline{a}_i^2 f_{ij} - \overline{x}^2,\ \overline{s}_y^2 \approx \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}\overline{b}_j^2 f_{ij} - \overline{y}^2,$$
$$\overline{s}_{xy} \approx \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}\overline{a}_i\overline{b}_j f_{ij} - \overline{x}\,\overline{y}.$$

3

Основни поими на математичката статистика

3.1 Статистички модел

При проучување на соберените статистички податоци во врска со некоја појава, се наметнува барањето да се донесе одреден заклучок во врска со разгледуваната појава. Токму ова е главната задача на математичката статистика. Донесувањето на заклучоци во врска со разгледуваната појава врз основа на конечен број на статистички податоци е познато како **статистичко заклучување**. Постојат главно два пристапа при донесувањето на статистички заклучоци. Првиот, **пристап на честоти** предложен од Кендел, е можеби и најкористениот во пракса и со него се претпоставува дека секој експеримент е бесконечно многу пати повторлив и дека мора да ги земеме во предвид сите можни исходи од експериментот за да може да донесеме некој статистички заклучок. Спротивно од овој пристап, **Бајесовиот пристап** ги темели заклучоците само на набљудуваните податоци и неодреденостите во врска со непознатите параметри се опушуваат преку распределби на веројатност кои зависат од овие податоци. Како и да е постојат и методи на статистичко заклучување кои претставуваат комбинација на овие два пристапа.

Во теоријата на статистичкото заклучување се конструираат математички модели кои овозможуваат егзактно дефинирање на проблемот, потоа со математички методи се пристапува кон решавање на проблемот и целта е да добиените резултати се применат во пракса и во другите научни дисциплини. Општа претпоставка при изградбата на теориските модели е дека низата статистички податоци $x_1, x_2, ..., x_n$ е некоја вредност на одреден случаен вектор $X = (X_1, X_2, ..., X_n)$. Претпоставуваме уште дека распределбата на веројатност на случајниот вектор X е непозната, но припаѓа на некоја фамилија $\mathcal{P}$

од **допустливи распределби на веројатност** за случајниот вектор X. Тогаш, просторот Ω од сите можни исходи на експериментот, σ-алгебрата $\mathcal{F}$ од подмножества од Ω и фамилијата $\mathcal{P}$ од допустливи распределби на веројатност (веројатносни мери) ја формираат тројката $(\Omega, \mathcal{F}, \mathcal{P})$ која се нарекува **статистички модел** за разгледуваниот експеримент. Понекогаш, терминот статистички модел се однесува и на самата фамилија $\mathcal{P}$ од допустливи распределби на веројатност.

За дефинирањето на фамилијата $\mathcal{P}$ не постојат прецизни критериуми, најчесто се изведува врз база на искуство и интуиција. Ако се земе претесна фамилија на допустливи распределби на веројатност, постои можност да вистинската распределба на веројатност остане надвор од таа фамилија, надвор од моделот. Ако за $\mathcal{P}$ се земе преширока класа, тогаш практично ништо нема да може да се заклучи за вистинската распределба врз основа на дадените податоци.

Ако фамилијата $\mathcal{P}$ од допустливи распределби на веројатност може да се опише преку конечен број на параметри $\theta = (\theta_1, ..., \theta_r)$, тогаш станува збор за **параметарски модел** на статистичко заклучување. Тогаш, може да запишеме дека

$$X = (X_1, X_2, ..., X_n) \sim F_\theta,\ \theta \in \Theta,$$

каде F_θ е функцијата на распределба на X и Θ е множеството од сите можни вредности за параметарот θ, познато како **простор на параметри**. Моделите пак чии распределби не можат да се опишат преку конечен број на параметри или начинот на изразување не е едноставен се нарекуваат **непараметарски модели**.

Многу често се претпоставува дека случајните променливи $X_1, X_2, ..., X_n$ се независни и еднакво распределени со заедничка распределба на веројатност P, односно се претопоставува дека податоците $x_1, x_2, ..., x_n$ се добиени како резултат на n независни мерења подложени на една иста статистичка законитост. Тогаш, фамилијата $\mathcal{P}$ од допустливи распределби на веројатност се стеснува на сите можни еднодимензионални распределби на веројатност.

Пример 3.1. (Биномен модел) Несиметрична монета се фрла n пати. Ако со 1 означуваме појава на "писмо", а со 0 појава на "грб", тогаш просторот од сите можни исходи од овој експримент е

$$\Omega = \{(x_1, x_2, ..., x_n) : x_i = 0 \text{ или } x_i = 1,\ i = 1, 2, ..., n\}.$$

Фамилијата од допустливи распределби на веројатност е $\mathcal{P} = \{P_\theta : 0 \leq \theta \leq 1\}$, каде веројатноста P_θ е дадена со

$$P_\theta(x) = \theta^{x_1+x_2+...+x_n}(1-\theta)^{n-(x_1+x_2+...+x_n)},\ x = (x_1, x_2, ..., x_n) \in \Omega,$$

Ирена Стојковска

каде θ е непознатата веројатност за појавување на "писмо" при едно фрлање на монетата. Просторот на параметри е $\{\theta : 0 \leq \theta \leq 1\}$. Овој статистички модел е познат како Биномен модел.

Пример 3.2. (Поасонов модел) Ако треба да се опише случајна променлива која претставува број на реализирани настани во фиксиран временски интервал, како на пример, бројот на телефонски разговори, или бројот на сообраќајни незгоди, или бројот на посетители на еден маркет, се користи Поасоновиот модел определен со множеството $\Omega = \{0, 1, 2, ...\}$ и фамилијата од допустливи распределби на веројатност $\mathcal{P} = \{P_\lambda : \lambda > 0\}$, каде веројатноста P_λ е дадена со

$$P_\lambda(x) = \frac{\lambda^x}{x!} e^{-\lambda},\ x = 0, 1, 2, ...$$

Просторот на параметри е $\{\lambda : \lambda > 0\}$.

Пример 3.3. (Гаусов (нормален) модел) Ако при мерење на некоја физичка величина со непозната вредност m, апаратот за мерење не прави системска грешка, туку на резултатот на мерењето влијаат голем број на случајни фактори, со незначително поединечно влијание, тогаш резултатот на мерењето се опишува со Гаусов (нормален) модел. Тогаш, може да се земе $\Omega = \mathbb{R}$ (сите можни резултати од мерењата), а за фамилијата од сите допустливи распределби на веројатност да се земе

$$\mathcal{P} = \{P_{m,\sigma^2} : -\infty < m < \infty,\ 0 < \sigma^2 < +\infty\},$$

каде веројатноста P_{m,σ^2} е дадена со

$$P_{m,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}},\ x \in \mathbb{R}.$$

Просторот на параметри е $\{(m, \sigma^2) : -\infty < m < \infty,\ 0 < \sigma^2 < +\infty\}$. Доколку претпоставиме дека дисперзијата σ^2 е позната и изнесува σ_0^2, тогаш фамилијата од сите допустливи распределби на веројатност е

$$\mathcal{P} = \{P_{m,\sigma_0^2} : -\infty < m < \infty\},$$

додека просторот на параметри е $\{m : -\infty < m < \infty\}$.

Пример 3.4. Нека $X_1, X_2, ..., X_n$ се независни и еднакво распределнеи случајни променливи со непрекината функција на распределба F која е непозната. Просторот на параметри за овој модел се состои од сите можни непрекинати распределби. И бидејќи овие распределби неможе да се индексираат со коечнодимензионален параметар, затоа овој модел се смета за **непараметарски модел**.

Ирена Стојковска

Исто така може да претпоставиме дека $F(x)$ има густина на распределба $p(x-\theta)$ каде θ е непознат параметар и p е непозната густина на распределба која го задоволува условот $p(x) = p(-x)$. Тогаш, овој модел е исто така непараметарски, но зависи од реално вредносен параметар θ. Затоа, тој може да се смета за **полупараметарски модел**.

Основни проблеми на статистичкото заклучување се **оценувањето на параметри**, односно наоѓање на нумерички вредности со кои се апроксимираат непознатите параметри на претпоставената распределба на веројатност и одредување на точност на таа апроксимација, и **тестирањето на хипотези**, односно дефинирање на постапки за донесување на одлуки за прифаќање, односно отфрлање на однапред поставената хипотеза за непознатите параметри или распределбата на веројатност.

Пример 3.5. Да претпоставиме дека некоја физичка величина m ја мериме n пати и нека резултатот на тие мерења се меѓусебно независни случајни променливи $X_1, X_2, ..., X_n$ со иста распределба која припаѓа на фамилијата допустливи распределби $\mathcal{P}$ дадена во Пример 3.3. Согласно законот на големите броеви имаме дека

$$\frac{X_1 + X_2 + ... + X_n}{n} \xrightarrow{\text{с.с.}} m.$$

Ако $x_1, x_2, ..., x_n$ се конкретните вредности добиени при n-те независни мерења, тогаш природно се наметнува да при големи вредности на n аритметичката средина $m_0 = \frac{x_1+x_2+...+x_n}{n}$ се смета за добра апроксимација на непознатиот параметар m. Оваа апроксимација се смета за оценка на параметарот m. Понатаму може да се применат разни критериуми за проценка колку е добра оваа оценка.

Пример 3.6. Од некои априорни причини, претпоставуваме дека вредноста на непознатиот параметар $\theta = \theta_0$ од Пример 3.1, каде $\theta_0 \in [0, 1]$ е даден број. Со помош на релативната честота $\frac{m}{n}$, каде m е бројот на појавувања на "писмо" при n-те независни фрлања т.е. $m = x_1 + x_2 + ... + x_n$, треба да се одлучиме дали ќе ја прифатиме хипотезата $\theta = \theta_0$ или ќе ја отфрлиме. Бидејќи при големи вредности на n релативната честота $\frac{m}{n}$ е блиска до веројатноста θ, добиваме еден критериум за проверка на хипотезата $\theta = \theta_0$ кој се базира на оценка на растојанието на релативната честота од вистинската вредност на веројатноста т.е. $|\frac{m}{n} - \theta|$. Ако ова растојание е големо, тогаш хипотезата ја отфрламе, ако е мало ја прифаќаме.

Ирена Стојковска

3.2 Популација, обележје и примерок

Доколку при изведување на некој експеримент секој елемент од одредено множество го избираме на случаен начин, тогаш тоа множесто може да се сфати како множество од сите можни исходи на експериментот, се означува со Ω и се нарекува **популација**. Популација (или **целна популација**) се дефинира уште и како целокупноста од еднородни елементи кои се предмет на истражување и за кои потребро е да се набави одредена информација. Набљудуваната заедничка карактеристика за елементите од популацијата која е предмет на истражување се нарекува **обележје**, се означува со X и се смета за случајна променлива чија расподелба на веројатност е непозната.

Пример 3.7. а) Се спроведува испитување за квалитетот на произведените сијалици во една фабрика. Севкупноста на сите произведени сијалици во фабриката ја претставува популацијата, додека обележје може да биде "должината на животот" на сијалицата во часови.

б) Ако се изведува испитување за успехот по математика во едно училиште, популацијата ја претставуваат учениците во училиштето, додека обележје може да биде нивната оценка по математика на завршниот испит.

в) При истражување за климатските услови во една земја, сите календарски години ја сочинуваат популацијата, додека обележје може да бидат вкупните врнежи на единица површина во текот на една година во земјата.

Нека експериментот се состои во избор на елемент од популацијата Ω и забележување на вредноста на обележето X која одговара на избраниот елемент. Резултат од овој експеримент е случајна променлива X. Ако експериментот се повтори n пати, како резултат се добива подредена n-торка од случајни променливи, односно случаен вектор $(X_1, X_2, ..., X_n)$ кој се нарекува **случаен примерок**. Бројот n се нарекува **големина на примерокот** или **обем на примерокот**. Ако случајните променливи $X_1, X_2, ..., X_n$ се независни и еднакво распределени со распределба еднаква на обележјето X, тогаш станува збор за **прост случаен примерок**, кој често се нарекува само **примерок**.

Примерокот треба да биде **репрезентативен** т.е. на правилен начин да ја претставува целата популација. Репрезентативноста може да се постигне ако секој елемент од популацијата има еднакви шанси да биде избран, и изборот на секој елемент да биде **случаен** и **независен**. Постојат повеќе методи, начини за избирање на репрезентативен примерок.

Ако за примерокот $(X_1, X_2, ..., X_n)$ ги забележиме вредностите на набљуваното обележје X за секоја од компонентите на примерокот, добиваме **реализација на примерокот** $(x_1, x_2, ..., x_n)$ која одговара на обележјето X, каде x_i е вредноста на случајната променлива X_i, $i = 1, 2, ..., n$. Често реализацијата на примерокот се нарекува примерок.

3.3 Емпириска функција на распределба

Основен проблем при статистичките истражувања е одредување на распределбата на обележјето X врз основа на даден примерок. Во математичката статистика постои теорема (позната како централна теорема на математичката статистика) која дава потврден одговор на прашањето дали простиот случаен примерок може да даде комплетна информација за распределбата на обележјето X. При тоа, точното одредување на распределбата на обележјето X бара големината n на примерокот неограничено да расте. Бидејќи во пракса може да се работи само со конечна големина на примерокот, распределбата на обележјето X може да се одреди само приближно, и колку е n поголемо толку е апроксимацијата поточна.

Начинот на кој примерокот ја одредува распределбата на обележјето X е даден преку неговата емпириска функција на распределба.

Дефиниција 3.1. Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележје со функција на распределба $F(x)$. Функцијата $F_n : \mathbb{R} \to \chi$, каде χ е множеството од случајни променливи дефинирана со

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I\{X_i < x\},$$

каде $I\{\cdot\}$ е индикатор на настан, се нарекува **емпириска функција на распределба** на примерокот $(X_1, X_2, ..., X_n)$.

Бидејќи сумата на n независни и еднакво распределени случајни променлви со Бернулиеви 0,1 распределби со параметар p, $0 < p < 1$ има $\mathcal{B}(n, p)$ распределба, заклучуваме дека случајната променлива $nF_n(x)$ има $\mathcal{B}(n, p)$ распределба каде $p = P\{X_i < x\} = F(x)$. Од тука следува дека распределбата на веројатност на емпириската функција на распределба е

$$P\{F_n(x) = \frac{k}{n}\} = \binom{n}{k} F^k(x)(1 - F(x))^{n-k},\ k = 0, 1, ..., n.$$

Нека $(x_1, x_2, ..., x_n)$ е реализација на примерокот $(X_1, X_2, ..., X_n)$. Ги подредуваме броевите $x_1, x_2, ..., x_n$ во растечки редослед, и добиената низа ја означуваме со $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$. Тогаш, **реализација на емпириската функција на распределба** на примерокот $(X_1, X_2, ..., X_n)$ која одговара на реализацијата на примерокот $(x_1, x_2, ..., x_n)$ е дадена со

$$F_n(x) = \begin{cases} 0 & , x \leq x_{(1)} \\ \frac{k}{n} & , x_{(k)} < x \leq x_{(k+1)},\ k = 1, 2, ..., n-1 \\ 1 & , x > x_{(n)} \end{cases}$$

Ирена Стојковска

Се согледува дека реализацијата на емпириската функција на распределба се поклопува со функцијата на кумулативните релативни честоти на статистичките податоци $x_1, x_2, ..., x_n$.

Значењето на емпириската функција на распределба F_n се состои во тоа што за големи вредности на n таа претставува добра апроксимација на функцијата на распределба F на обележјето X, што е покажано со следните две тврдења.

Теорема 3.1. *Нека F е функцијата на распределба на обележјето X и F_n е емпириската функција на распределба на примерокот $(X_1, X_2, ..., X_n)$ кој одговара на обележјето X. Тогаш, за секој реален број x важи*

$$P\{\lim_{n\to\infty} F_n(x) = F(x)\} = 1.$$

Доказ. Нека $x \in \mathbb{R}$ е произволен реален број. Дефинираме случајни променливи $Y_i = I\{X_i < x\}$, $i = 1, 2, ..., n$, кои се независни и еднакво распределени (од $X_1, ..., X_n$ независни и еднакво распределени) со конечни математички очекувања $E(Y_i) = P\{X_i < x\} = F(x) < +\infty$. Следи дека за низата $\{Y_i\}$ важи силниот закон на големите броеви, т.е.

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \overset{a.s.}{\to} \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = F(x),$$

и имајќи ја во предвид дефиницијата на емпириската функција на распределба на промерокот $(X_1, X_2, ..., X_n)$, $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i < x\} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, следи тврдењето на теоремата т.е. $P\{\lim_{n\to\infty} F_n(x) = F(x)\} = 1$. ∎

Теорема 3.2. (Централна теорема на математичката статистика на Гливенко - Кантели) *Нека F е функцијата на распределба на обележјето X и F_n е емпириската функција на распределба на примерокот $(X_1, X_2, ..., X_n)$ кој одговара на обележјето X. Тогаш, важи*

$$P\{\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0\} = 1.$$

Доказ. Нека F е непрекината функција на распределба. Нека $\varepsilon > 0$. Тогаш, постои $m \in \mathbb{N}$ така што $\frac{1}{m} \leq \varepsilon$. Нека $-\infty = t_0 < t_1 < ... < t_{m-1} < t_m = +\infty$ се такви да $F(t_k) = \frac{k}{m}$, $k = 0, 1, 2, ..., m$ (постојат затоа што $F(x)$ е неопаѓачка). Тогаш, за $x \in (t_k, t_{k+1}]$ важи

$$F_n(x) - F(x) \leq F_n(t_{k+1}) - F(t_k) \leq F_n(t_{k+1}) - F(t_{k+1}) + \varepsilon, \tag{3.1}$$

$$F_n(x) - F(x) \geq F_n(t_k) - F(t_{k+1}) \geq F_n(t_k) - F(t_k) - \varepsilon, \tag{3.2}$$

затоа што $F(t_{k+1}) = \frac{k+1}{m} = \frac{k}{m} + \frac{1}{m} \leq F(t_k) + \varepsilon$.

Нека $A_k = \{w \in \Omega : \lim_{n\to\infty} F_n(t_k) = F(t_k)\}$, $k = 0, 1, 2, ..., m$ и $A = \cap_{k=0}^{m} A_k$. Тогаш, од Теорема 3.1 следи дека $P(A_k) = 1$, $k = 0, 1, 2, ..., m$, од каде и $P(A) = 1$. Тогаш, $\forall w \in A$, $\exists n_0 = n_0(w)$, така што $\forall n \geq n_0(w)$,

$$|F_n(t_k) - F(t_k)| \leq \varepsilon,\ k = 0, 1, 2, ..., m. \tag{3.3}$$

Од (3.1)-(3.3) следи дека $\forall w \in A$, $\exists n_0 = n_0(w)$, така што $\forall n \geq n_0(w)$,

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \leq 2\varepsilon, \tag{3.4}$$

од каде добиваме дека важи тврдењето на теоремата кога F е непрекината функција.

Нека сега, F е произволна функција на распределба (F е неопаѓачка и непрекината од лево). Нека $\varepsilon > 0$. Тогаш, постои $m \in \mathbb{N}$ и низа $-\infty = t_0 < t_1 < ... < t_{m-1} < t_m = +\infty$ така што

$$F(t_{k+1}) - F(t_k + 0) \leq \varepsilon. \tag{3.5}$$

Б.Г.О нека $\{t_1, t_2, ..., t_{m-1}\}$ ги содржи сите точки во кои F прави скок не помал од $\varepsilon/2$. Тогаш, за $x \in (t_k, t_{k+1}]$ важи

$$F_n(x) - F(x) \leq F_n(t_{k+1}) - F(t_k + 0) \leq F_n(t_{k+1}) - F(t_{k+1}) + \varepsilon, \tag{3.6}$$

$$F_n(x) - F(x) \geq F_n(t_k + 0) - F(t_{k+1}) \geq F_n(t_k + 0) - F(t_k + 0) - \varepsilon. \tag{3.7}$$

Нека $A_k = \{w \in \Omega : \lim_{n\to\infty} F_n(t_k + 0) = F(t_k + 0)\}$, $B_k = \{w \in \Omega : \lim_{n\to\infty} F_n(t_k) = F(t_k)\}$, $k = 0, 1, 2, ..., m$ и $A = \cap_{k=0}^{m} A_k B_k$. Од Теорема 3.1 следи дека $P(A_k) = P(B_k) = 1$, $k = 0, 1, 2, ..., m$, од каде и $P(A) = 1$. Значи, $\forall w \in A$, $\exists n_1 = n_1(w)$, така што $\forall n \geq n_1(w)$,

$$F_n(t_k + 0) - F(t_k + 0) \geq -\varepsilon,\ k = 0, 1, 2, ..., m, \tag{3.8}$$

$$F_n(t_k) - F(t_k) \leq \varepsilon,\ k = 0, 1, 2, ..., m. \tag{3.9}$$

Од (3.5)-(3.9) следи (3.4), од каде се добива дека важи тврдењето на теоремата и кога F е произволна функција на распределба. ∎

3.4 Статистики

Решавањето на проблемот на наоѓање на непознатата распределба на обележјето X, подразбира и одредување на некои карактеристики на обележјето. За таа цел се служиме со случајни променливи кои се функции од примерокот.

Ирена Стојковска

Дефиниција 3.2. Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележје X со функција на распределба $F(x)$ и нека $f : \mathbb{R}^n \to \mathbb{R}$, е Борелова функција. Тогаш, случајната променлива $f(X_1, X_2, ..., X_n)$ се нарекува **статистика**.

Пример 3.8. Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележје X со функција на распределба $F(x)$.

а) Статистиката $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ се нарекува **средина на примерокот**.

б) Статистиката $\overline{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2$ се нарекува **дисперзија на примерокот**, додека $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{n}{n-1}\overline{S}_n^2$ е позната како **корегирана дисперзија на примерокот**.

в) Статистиката $Z_{n,k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ се нарекува **k-ти момент на примерокот**, додека $M_{n,k} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^k$ се нарекува **k-ти централен момент на примерокот**. Лесно се воочува дека $Z_{n,1} = \overline{X}_n$ и $M_{n,2} = \overline{S}_n^2$

г) За примерокот $(X_1, X_2, ..., X_n)$ ја формираме соодветната **варијациона низа** $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ за која важи дека за секоја реализација $(x_1, x_2, ..., x_n)$ на примерокот $(X_1, X_2, ..., X_n)$ имаме

$$X_{(1)} = x_{(1)}, X_{(2)} = x_{(2)}, ..., X_{(n)} = x_{(n)},$$

каде $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ е реализацијата $(x_1, x_2, ..., x_n)$ подредена по големина. Тогаш, секој елемент од варијационата низа $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ се нарекува **подредена статистика**. Најмалата и најголемата подредена статистика може да се изразат како функции од примерокот на следниот начин

$$X_{(1)} = \min\{X_1, X_2, ..., X_n\}, \quad X_{(n)} = \max\{X_1, X_2, ..., X_n\}.$$

Пример 3.9. Кога разгледуваме дводимензионално обележје (X, Y), со непознати распределби F_X и F_Y за секое од обележјата X и Y соодветно, тогаш испитувањата ги вршиме со помош на **дводимензионален примерок**

$$(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n).$$

Ирена Стојковска

Пример за статистика (функција од дводимензионалниот примерок) е

$$R_{X,Y} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}_n\,\overline{Y}_n}{\sqrt{\overline{S}^2_{Xn}\,\overline{S}^2_{Yn}}},$$

каде

$$\overline{S}^2_{Xn} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2,\ \overline{S}^2_{Yn} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2,$$

и се нарекува **коефициент на корелација на примерокот**.

3.5 Карактеристики на некои статистики

Бидејќи статистиките се користат при статистичкото заклучување, потребно е да се знаат нивните бројни карактеристики, точните распределби кога n е конечно (за примерок со мал обем или **мал примерок**) и асимптотското однесување на распределбите при $n \to \infty$ (за примерок со голем обем или **голем примерок**).

Нека X е обележје со непозната функција на распределба F. Да ги означиме бројните карактеристики на обележјето X со

$$EX = m,\ DX = \sigma^2,\ EX^k = m_k,\ E(X-m)^k = \mu_k.$$

Да забележиме дека $m_1 = m$, $\mu_1 = 0$ и $\mu_2 = \sigma^2$. Нека сега $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X. Тогаш важи следното својство за бројните карактеристики на некои статистики изразени преку бројните карактеристики на обележјето X.

Својство 3.1. *Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X со $EX = m,\ DX = \sigma^2,\ EX^k = m_k,\ E(X-m)^k = \mu_k$. Тогаш,*

а) $E(\overline{X}_n) = m,\ D(\overline{X}_n) = \frac{\sigma^2}{n}$,

б) $E(\overline{S}^2_n) = \frac{n-1}{n}\sigma^2,\ D(\overline{S}^2_n) = \frac{(n-1)^2}{n^3}\left(\mu_4 - \frac{n-3}{n-1}\mu_2^2\right)$,

в) $E(Z_{n,k}) = m_k,\ D(Z_{n,k}) = \frac{1}{n}(m_{2k} - m_k^2)$.

Доказ. а) $E(\overline{X}_n) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\cdot nm = m$,
$D(\overline{X}_n) = D(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}$.

б) Најнапред, $E(X_i^2) = DX_i + (EX_i)^2 = \sigma^2 + m^2$.
$E(\overline{S}^2_n) = E(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\overline{X}_n^2) =$
$\frac{1}{n}\sum_{i=1}^{n}(\sigma^2 + m^2) - (D(\overline{X}_n) + (E(\overline{X}_n))^2) = (\sigma^2 + m^2) - (\frac{\sigma^2}{n} + m^2) = \frac{n-1}{n}\sigma^2$.

Ирена Стојковска

в) $E(Z_{n,k}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i^k) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^k) = \frac{1}{n}\sum_{i=1}^{n} m_k = m_k$,
$D(Z_{n,k}) = D(\frac{1}{n}\sum_{i=1}^{n} X_i^k) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i^k) = \frac{1}{n^2}\sum_{i=1}^{n}(EX_i^{2k} - (EX_i^k)^2) =$
$\frac{1}{n^2}\sum_{i=1}^{n}(m_{2k} - (m_k)^2) = \frac{1}{n}(m_{2k} - (m_k)^2)$. ∎

Нормалната распределба има важна улога при проучување на случајните појави. Затоа, важно е да се знаат точните распределби на некои статистики, меѓу кои и основните статистики $\overline{X}_n$ и $\overline{S}_n^2$, во случај кога обележјето X има нормална распределба.

Теорема 3.3. *Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X кое има $\mathcal{N}(m, \sigma^2)$ распределба. Тогаш,*

а) Случајната променлива $\overline{X}_n$ има $\mathcal{N}(m, \frac{\sigma^2}{n})$ распределба, што значи дека случајната променлива $\frac{\overline{X}_n - m}{\sigma}\sqrt{n}$ има $\mathcal{N}(0,1)$ распределба,

б) Случајната променлива $\frac{n\overline{S}_n^2}{\sigma^2}$ има χ^2_{n-1} распределба,

в) Случајните променливи $\overline{X}_n$ и $\overline{S}_n^2$ се независни.

Доказ. а) Случајните променливи X_i, $i = 1, 2, ..., n$ се независни и еднакво распределени т.е. $X_i \sim \mathcal{N}(m, \sigma^2)$, $i = 1, 2, ..., n$, и ако во Својство 1.6 земеме $a_1 = ... = a_n = 1/n$ добиваме дека

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(m, \frac{\sigma^2}{n}). \tag{3.10}$$

Потоа, според Својство 1.4 имаме

$$\frac{\overline{X}_n - m}{\sigma}\sqrt{n} = \frac{\sum_{i=1}^{n}(X_i - m)}{\sqrt{n}\cdot\sigma} \sim \mathcal{N}(0,1).$$

б) Претходно покажавме дека $\overline{X}_n \sim \mathcal{N}(m, \frac{\sigma^2}{n})$, види (3.10), и затоа $\frac{X_i - \overline{X}_n}{\sigma} \sim \mathcal{N}(0,1)$, според Својство 1.6. Сега, бидејќи меѓу случајните променливи $\frac{X_i - \overline{X}_n}{\sigma}$, $i = 1, 2, ..., n$ постои една линеарна врска

$$\sum_{i=1}^{n} \frac{X_i - \overline{X}_n}{\sigma} = \frac{1}{\sigma}(\sum_{i=1}^{n} X_i - n\overline{X}_n) = 0,$$

заклучуваме дека (види ја Забелешка 1.1)

$$\frac{n\overline{S}_n^2}{\sigma^2} = \sum_{i=1}^{n}(\frac{X_i - \overline{X}_n}{\sigma})^2 \sim \chi^2_{n-1}. \tag{3.11}$$

Ирена Стојковска

в) Дефинираме $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ со

$$Y_k = X_k - \overline{X}_n,\ k = 1, 2, ..., n-1,\ Y_n = \overline{X}_n.$$

Тогаш, постои несингуларна матрица M така што

$$\mathbf{Y} = \mathrm{M}\mathbf{X},$$

каде $\mathbf{X} = (X_1, X_2, ..., X_n)$ е случаен вектор од независни и еднакво распределени случајни променливи со нормални распределби. Следи дека и $\mathbf{Y}$ е случаен вектор кој има n-димензионална нормална распределба.

Сега, за $k = 1, 2, ..., n-1$ имаме

$$\begin{aligned} cov(Y_k, Y_n) &= E(Y_k Y_n) - E(Y_k)E(Y_n) = E(Y_k\overline{X}_n) - E(X_k - \overline{X}_n)E(\overline{X}_n) = E(Y_k\overline{X}_n) = \\ &= E((X_k - \overline{X}_n)\overline{X}_n) = E(X_k\overline{X}_n - \overline{X}_n^2) = \frac{1}{n}\sum_{i=1}^{n} E(X_k X_i) - E(\overline{X}_n^2) = \\ &= \frac{1}{n}((n-1)m^2 + (\sigma^2 + m^2)) - (\frac{\sigma^2}{n} + m^2) = 0, \end{aligned}$$

од каде следи дека $Y_n = \overline{X}_n$ не зависи од $Y_1, Y_2, ..., Y_{n-1}$. Потоа,

$$n\overline{S}_n^2 = \sum_{k=1}^{n}(X_k - \overline{X}_n)^2 = \sum_{k=1}^{n-1} Y_k^2 + (X_n - \overline{X}_n)^2 = \sum_{k=1}^{n-1} Y_k^2 + (\sum_{k=1}^{n-1} Y_k)^2,$$

што значи дека $\overline{S}_n^2$ зависи само од $Y_1, Y_2, ..., Y_{n-1}$ за кои претходно покажавме дека се независни од $Y_n = \overline{X}_n$. Значи, $\overline{X}_n$ и $\overline{S}_n^2$ се независни случајни променливи, што требаше да се докаже. ∎

Теорема 3.4. *Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X кое има $\mathcal{N}(m, \sigma^2)$ распределба. Тогаш, случајната променлива*

$$\frac{\overline{X}_n - m}{\overline{S}_n}\sqrt{n-1} \sim t_{n-1}.$$

Доказ. Од Теорема 3.3 в), следи дека случајните променливи $\overline{X}_n$ и $\overline{S}_n^2$ се независни. Тогаш, независни се и случајните променливи $\frac{n\overline{S}_n^2}{\sigma^2}$ и $\frac{\overline{X}_n - m}{\sigma/\sqrt{n}}$. Од друга страна заради (3.10) и Својство 1.4 имаме дека

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1). \tag{3.12}$$

Ирена Стојковска

Сега, од (3.11) и (3.12), според Својство 1.10, имаме дека

$$\frac{\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{n\overline{S}_n^2}{\sigma^2}/(n-1)}}=\frac{\overline{X}_n-\mu}{\overline{S}_n}\sqrt{n-1}\sim t_{n-1}, \tag{3.13}$$

што требаше да се покаже. ∎

Теорема 3.5. *Нека се дадени два независни примероци, првиот* $(X_1, X_2, ..., X_n)$ *кој одговара на обележјето* $X \sim \mathcal{N}(m_X, \sigma^2)$ *и вториот* $(Y_1, Y_2, ..., Y_k)$ *кој одговара на обележјето* $Y \sim \mathcal{N}(m_Y, \sigma^2)$. *Тогаш,*

а) Случајната променлива $\frac{\overline{X}_n-m_X}{\sigma}$ *има* $\mathcal{N}(0, \frac{1}{n})$ *распределба,*

б) Случајната променлива $\frac{\overline{Y}_k-m_Y}{\sigma}$ *има* $\mathcal{N}(0, \frac{1}{k})$ *распределба,*

в) Случајните променливи $\frac{n\overline{S}_{Xn}^2}{\sigma^2}\sim\chi^2_{n-1}$ *и* $\frac{k\overline{S}_{Yk}^2}{\sigma^2}\sim\chi^2_{k-1}$,

г) Случајната променлива $\frac{\overline{X}_n-m_X}{\sigma}-\frac{\overline{Y}_k-m_Y}{\sigma}$ *има* $\mathcal{N}(0, \frac{1}{n}+\frac{1}{k})$ *распределба,*

д) Случајната променлива $\frac{1}{\sigma^2}(n\overline{S}_{Xn}^2+k\overline{S}_{Yk}^2)$ *има* χ^2_{n+k-2} *распределба од каде следува дека*

$$\frac{(\overline{X}_n-m_X)-(\overline{Y}_k-m_Y)}{\sqrt{n\overline{S}_{Xn}^2+k\overline{S}_{Yk}^2}}\sqrt{\frac{nk}{n+k}(n+k-2)}\sim t_{n+k-2}.$$

Доказ. а)-в) следат директно од Теорема 3.3, г) следи од Својство 1.6.

д) Од в) и Својство 1.7 следи дека

$$\frac{n\overline{S}_{Xn}^2}{\sigma^2}+\frac{k\overline{S}_{Yk}^2}{\sigma^2}\sim\chi^2_{(n-1)+(k-1)}\equiv\chi^2_{n+k-2},$$

потоа од г) следи дека

$$\frac{\frac{\overline{X}_n-m_X}{\sigma}-\frac{\overline{Y}_k-m_Y}{\sigma}}{\sqrt{\frac{1}{n}+\frac{1}{k}}}\sim\mathcal{N}(0,1),$$

и конечно од Својство 1.10 (дефиниција на студентова распределба) имаме

$$\frac{\frac{\frac{\overline{X}_n-m_X}{\sigma}-\frac{\overline{Y}_k-m_Y}{\sigma}}{\sqrt{\frac{1}{n}+\frac{1}{k}}}}{\sqrt{\frac{\frac{n\overline{S}_{Xn}^2}{\sigma^2}+\frac{k\overline{S}_{Yk}^2}{\sigma^2}}{n+k-2}}}\sim t_{n+k-2},$$

од каде следи бараното тврдење. ∎

Ирена Стојковска

Теорема 3.6. *Нека се дадени два независни примероци, првиот* $(X_1, X_2, ..., X_n)$ *кој одговара на обележјето* $X \sim \mathcal{N}(m_X, \sigma_X^2)$ *и вториот* $(Y_1, Y_2, ..., Y_k)$ *кој одговара на обележјето* $Y \sim \mathcal{N}(m_Y, \sigma_Y^2)$. *Тогаш, случајната променлива*

$$\frac{n(k-1)\sigma_Y^2 \overline{S}_{Xn}^2}{k(n-1)\sigma_X^2 \overline{S}_{Yk}^2} \sim F_{n-1,k-1}.$$

Доказ. Од Теорема 3.5 в) следи дека $\frac{n\overline{S}_{Xn}^2}{\sigma^2} \sim \chi_{n-1}^2$ и $\frac{k\overline{S}_{Yk}^2}{\sigma^2} \sim \chi_{k-1}^2$, и од нивната независност, според Својство 1.11 (дефиниција на Фишерова распределба) имаме

$$\frac{\frac{n\overline{S}_{Xn}^2}{\sigma_X^2}/(n-1)}{\frac{k\overline{S}_{Yk}^2}{\sigma_Y^2}/(k-1)} \sim F_{n-1,k-1},$$

од каде следи бараното тврдење. ∎

При големи вредности на обемот n на примерокот $(X_1, X_2, ..., X_n)$ кој одговара на произволно обележје X со конечни математичко очекување m и дисперзија σ^2, распределбата на средината на примерокот $\overline{X}_n$ се стреми кон распределбата на средината на примерокот кој одговара на обележје со нормална $\mathcal{N}(m, \sigma^2)$ распределба. Во пракса доволно е да $n > 30$ па да се користи резултатот од следната теорема.

Теорема 3.7. *Нека* $(X_1, X_2, ..., X_n)$ *е примерок кој одговара на обележјето* X *со конечни математичко очекување* $EX = m < +\infty$ *и дисперзија* $DX = \sigma^2 < +\infty$. *Тогаш,*

$$\overline{X}_n \xrightarrow{\text{c.c.}} m,\ n \to \infty,$$
$$\frac{\sqrt{n}}{\sigma}(\overline{X}_n - m) \xrightarrow{dist.} \mathcal{N}(0,1),\ n \to \infty.$$

Доказ. Првото тврдење, $\overline{X}_n \xrightarrow{\text{c.c.}} m,\ n \to \infty$, следи од законот на големите броеви. Второто тврдење, $\frac{\sqrt{n}}{\sigma}(\overline{X}_n - m) \xrightarrow{\text{dist.}} \mathcal{N}(0,1),\ n \to \infty$, следи од Централната гранична теорема. ∎

4

Оценување на параметри

4.1 Точкасти оценувачи

Нека обележјето X има функција на распределба F која припаѓа на фамилијата допустливи функции на распределби

$$\mathcal{P} = \{F(x, \theta) : \theta \in \Theta\},$$

каде θ е векторот од непознати параметри и Θ е просторот од параметри (види Пример 3.1-3.3). Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X. Еден начин за оценување на вредноста на непознатиот параметар θ е со помош на точкасти оценувачи.

Дефиниција 4.1. Нека $\widehat{\theta} = U(X_1, X_2, ..., X_n)$ е статистика која како функција не зависи од оценуваниот параметар или други непознати параметри, туку зависи само од случајните променливи од примерокот и познати константи, и при тоа за секоја реализација $(x_1, x_2, ..., x_n)$ на примерокот $(X_1, X_2, ..., X_n)$ важи $U(x_1, x_2, ..., x_n) \in \Theta$. Тогаш, велиме дека $\widehat{\theta}$ е **точкаст оценувач** или само **оценувач** за непознатиот параметар θ.

Секоја вредност $U(x_1, x_2, ..., x_n)$ која се добива за конкретна реализација $(x_1, x_2, ..., x_n)$ на примерокот $(X_1, X_2, ..., X_n)$ се нарекува **точкаста оценка** или само **оценка** за непознатиот параметар θ.

Пример 4.1. Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X кое има $\mathcal{N}(m, \sigma^2)$ распределба. Статистиката $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ може да се земе за оценувач на параметарот m, но не може да биде оценувач за σ^2 затоа што просторот од параметри за σ^2 се состои од позитивни броеви, додека $\overline{X}_n$ може да прими и негативна вредност за конкретна реализација на примерокот. Но зато статистиката $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ може да биде оценувач за σ^2.

Еден начин за добивање на оценувачи е со **принципот на замена** кој претпоставува дека непознатиот параметар θ може да се претстави во облик на функционал од функцијата на распределба F на обележјето X, т.е. $\theta = \theta(F)$. На пример, ако математичкото очекување $m = EX$ е непознатиот параметар кој сакаме да го оцениме, може да го запишеме како функционал од F на следниот начин

$$m(F) = \int_{-\infty}^{\infty} x dF(x).$$

Тогаш, зависноста на θ од функцијата на распределба F ни сугерира дека проблемот на наоѓање на оценувач за θ може да се сведе на наоѓање на добар оценувач $\widehat{F}$ на F и потоа да се замени овој оценувач на местото од F и на тој начин би се добил оценувач за θ, $\widehat{\theta} = \theta(\widehat{F})$. Принципот на замена подразбира за оценувач на F да се земе емпириската фунција на распределба на примерокот F_n (види Дефиниција 3.1), т.е. $\widehat{F} = F_n$.

Пример 4.2. Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X кое има функција на распределба F. Со принципот на замена оценувач за математичкото очекување на X, $m(F) = \int_{-\infty}^{\infty} x dF(x)$ е

$$\widehat{m} = m(\widehat{F}) = \int_{-\infty}^{\infty} x d\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n,$$

оценувач за дисперзијата на X, $\sigma^2(F) = \int_{-\infty}^{\infty} x^2 dF(x) - \Big(\int_{-\infty}^{\infty} x dF(x)\Big)^2$ е

$$\widehat{\sigma}^2 = \sigma^2(\widehat{F}) = \int_{-\infty}^{\infty} x^2 d\widehat{F}(x) - \Big(\int_{-\infty}^{\infty} x d\widehat{F}(x)\Big)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big)^2 = \overline{S}_n^2.$$

И покрај фактот дека емпириската функција на распределба на примерокот F_n рамномерно конвергира скоро сигурно кон функцијата на распределба F на обележјето X (Централна теорема на математичката статистка, Теорема 3.2), не може да се каже дека F_n е добар оценувач за F. На пример, F_n секогаш има дискретна распределба и ако вистинската распределба на F е непрекината, тогаш F_n нема да може да опфати одредени особини на F. Така, ако сакаме да ја оцениме густината на распределба p на обележјето X, тогаш ја користиме врската меѓу гистината на распределба и финкцијата на распределба, т.е.

$$F(x) = \int_{-\infty}^{x} p(u)du,$$

па природно се наметнува да оценувачот $\widehat{p}$ за густината p добиен со принципот на замена го задоволува равенството

$$\widehat{F}(x) = \int_{-\infty}^{x} \widehat{p}(u)du,$$

Ирена Стојковска

но таков оценувач не постои затоа што $\widehat{F} = F_n$ е скалеста функција. Во тој случај се применува "поглаток" оценувач за F и за него се применува принципот на замена.

Исто така може да се случи оценувачот добиен со принципот на замена и во случај кога не се работи за непрекинато обележје X да не може експлицитно да се дефинира.

Пример 4.3. Нека $\theta(F)$ го задоволува равенството

$$\int_{-\infty}^{\infty} g(x, \theta(F))dF(x) = 0,$$

за некоја функција $g(x, u)$. Тогаш, оценувачот добиен со принципот на замена $\widehat{\theta} = \theta(\widehat{F})$, каде $\widehat{F} = F_n$ го задоволува равенството

$$\int_{-\infty}^{\infty} g(x, \theta(F))d\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} g(X_i, \widehat{\theta}) = 0.$$

Па, во овој случај $\widehat{\theta}$ не е задолжително експлицитно дефиниран.

4.1.1 Непристрасни оценувачи

Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X кое има функција на распределба F која припаѓа на фамилијата допустливи распределби $\mathcal{P} = \{F(x, \theta) : \theta \in \Theta\}$. Нека $\widehat{\theta} = U(X_1, X_2, ..., X_n)$ е оценувач за параметарот θ.

Идеално сакаме распределбата на $\widehat{\theta}$ да биде концентрирана во околина на вистинската вредност на оценуваниот параметар θ. Постојат неколку едноставни мерки за квалитетот на еден оценувач базирани на неговата распределба. Првата мерка е пристрасноста на $\widehat{\theta}$, која е показател дали распределбата на $\widehat{\theta}$ е центрирана околу θ.

Дефиниција 4.2. Пристрасност (bias) на оценувачот $\widehat{\theta}$ се дефинира како

$$b(\widehat{\theta}) = E(\widehat{\theta}) - \theta. \tag{4.1}$$

За еден оценувач $\widehat{\theta}$ велиме дека е **непристрасен** (unbiased) или **центриран**, ако $b(\widehat{\theta}) = 0$, односно ако $E(\widehat{\theta}) = \theta$.

Пример 4.4. Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X со $EX = m$ и $DX = \sigma^2$. Тогаш,

а) статистиката $\overline{X}_n$ е непристрасен оценувач за m затоа што $E(\overline{X}_n) = m$ (види Својство 3.1 а)).

б) статистиката $\overline{S}_n^2$ не е непристрасен оценувач за σ^2 затоа што $E(\overline{S}_n^2) = \frac{n-1}{n}\,\sigma^2$ (види Својство 3.1 б)).

в) статистиката $S_n^2 = \frac{n}{n-1}\overline{S}_n^2$ е непристрасен оценувач за σ^2 затоа што $E(S_n^2) = E(\frac{n}{n-1}\,\overline{S}_n^2) = \frac{n}{n-1}\,E(\overline{S}_n^2) = \frac{n}{n-1}\cdot\frac{(n-1)}{n}\,\sigma^2 = \sigma^2$.

Едни од поголемите проблеми поврзани со поимот на непристрасност се тие дека непристрасните оценувачи не постојат секогаш (Пример 4.5) и дека непристрасните оценувачи не се инваријантни за трансформациите на непознатите параметри, односно ако $\widehat{\theta}$ е непристрасен оценувач за θ и $g : \Theta \to \Theta$ е некое пресликување од просторот на параметри во самиот себе, тогаш оценувачот $\vartheta = g(\theta)$ може да не е непристрасен оценувач за θ (Пример 4.6), освен ако g не е линеарна функција (Својство 4.1).

Пример 4.5. Нека обележјето X има $\mathcal{P}(\theta)$ распределба, односно

$$P\{X = k\} = \frac{\theta^k}{k!}e^{-\theta},\ k = 0, 1, 2, ...,$$

каде $\theta > 0$ е непознат параметар. Тогаш, врз основа на примерок со обем 1 не може да се дефинира непристрасна оценка за $\vartheta = 1/\theta$. Имено, да претпоставиме дека статистиката $\widehat{\vartheta} = U(X_1)$ е непристрасна оценка за ϑ. Тогаш, за секој $\theta > 0$ важи

$$\frac{1}{\theta} = E(U(X_1)) = \sum_{k=0}^{\infty} U(k)\frac{\theta^k}{k!}e^{-\theta},$$

од каде добиваме дека $e^{\theta} = \sum_{k=0}^{\infty} U(k)\frac{\theta^{k+1}}{k!}$. Но, од друга страна $e^{\theta} = \sum_{k=0}^{\infty}\frac{\theta^k}{k!}$, од каде заклучуваме дека $\sum_{k=0}^{\infty} U(k)\frac{\theta^{k+1}}{k!} = \sum_{k=0}^{\infty}\frac{\theta^k}{k!}$, што не е можно. Значи не постои непристрасен оценувач за $\vartheta = 1/\theta$.

Пример 4.6. Нека обележјето X има конечни $EX = m$ и $DX = \sigma^2$. Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X. Покажавме дека $E(\overline{X}_n) = m$ и $D(\overline{X}_n) = \frac{\sigma^2}{n}$ (Својство 3.1). Од тука следува дека

$$E(\overline{X}_n^2) = D(\overline{X}_n) + E(\overline{X}_n)^2 = \frac{\sigma^2}{n} + m^2.$$

Значи, $\overline{X}_n$ е непристрасен оценувач за m, но $\overline{X}_n^2$ не е непристрасем оценувач за m^2.

Ирена Стојковска

Својство 4.1. *Ако $\widehat{\theta}$ е непристрасен оценувач за θ, тогаш $\widehat{\vartheta} = a\widehat{\theta} + b$ е непристрасен оценувач за $\vartheta = a\theta + b$, каде a и b се познати константи.*

Доказ. Од услов имаме дека $E(\widehat{\theta}) = \theta$. Па,

$$E(\widehat{\vartheta}) = E(a\widehat{\theta} + b) = aE(\widehat{\theta}) + b = a\theta + b = \vartheta,$$

од каде следи дека $\widehat{\vartheta}$ е непристрасен оценувач за ϑ, што требаше да се покаже. ■

Во случај на голем примерок (кога $n \to \infty$) има смисла да се зборува за асимптотски непристрасен оценувач.

Дефиниција 4.3. Ако за оценувачот $\widehat{\theta}_n$ важи

$$\lim_{n\to\infty} E(\widehat{\theta}_n) = \theta \tag{4.2}$$

тогаш велиме дека тој е **асимптотски непристрасен оценувач**.

Понекогаш, самиот облик на асимптотскиот непристрасен оценувач дозволува тој да може да се корегира во непристрасен.

Пример 4.7. Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X со $EX = m$ и $DX = \sigma^2$.

а) Во Пример 4.4 покажавме дека статистиката $\overline{S}_n^2$ не е непристрасен оценувач за σ^2, но бидејќи $E(\overline{S}_n^2) = \dfrac{(n-1)\sigma^2}{n} \to \sigma^2$, кога $n \to \infty$, следи дека $\overline{S}_n^2$ е асимптотски непристрасен оценувач за σ^2. Неговата корекција до непристрасност е статистиката $S_n^2 = \frac{n}{n-1}\overline{S}_n^2$.

б) Ако тргнеме од непристрасниот оценувач S_n^2 за σ^2 и сакаме да добиеме оценувач за σ кој би го задржал својството за непристрасност, некако ни се наметнува идејата дека тоа може да биде оценувачот $S_n = \sqrt{S_n^2}$. Меѓутоа, од Теорема 3.3 имаме

$$Y = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1},$$

од каде следи дека

$$E(S_n) = \frac{\sigma}{\sqrt{n-1}}\, E(\sqrt{Y}) = \frac{\sigma}{\sqrt{n-1}}\, \frac{\sqrt{2}\Gamma(n/2)}{\Gamma((n-1)/2)} \neq \sigma$$

(покажи!), од каде заклучуваме дека S_n не е непристрасен оценувач за σ (уште еден пример дека непристрасните оценувачи не се инваријантни за трансформациите на непознатите параметри). Но, бидејќи $E(S_n) \to \sigma$, кога $n \to \infty$ (покажи!), следи дека S_n е асимптотски непристрасен оценувач за σ и тој лесно може да се корегира до непристрасност.

Ирена Стојковска

4.1.2 Оценувачи со минимална дисперзија

Постојат случаи кога во потрага за подобар оценувач за θ може да одбереме пристрасен оценувач отколку непристрасен, затоа што првиот имал помала дисперзија. Едни од мерките за расејување на распределбата на $\widehat{\theta}$ околу θ се средната апсолутна грешка (MAE - mean absolute error) и средната квадратна грешка (MSE - mean square error), погодни за споредба на различни оценувачи за θ.

Дефиниција 4.4. Средна апсолутна грешка (MAE) на оценувачот $\widehat{\theta}$ се дефинира како

$$MAE(\widehat{\theta}) = E|\widehat{\theta} - \theta|. \tag{4.3}$$

Дефиниција 4.5. Средна квадратна грешка (MSE) на оценувачот $\widehat{\theta}$ се дефинира како

$$MSE(\widehat{\theta}) = E(\widehat{\theta} - \theta)^2. \tag{4.4}$$

Бидејќи сакаме $\widehat{\theta}$ да биде близу до θ, природно е да прифатиме оценувач со помала вредности за MAE или MSE. При споредби на различни оценувачи на θ обично повеќе се користи MSE отколку MAE. Ова се должи на следната декомпозиција на $MSE(\widehat{\theta})$,

$$MSE(\widehat{\theta}) = E(\widehat{\theta}^2) - 2\theta E(\widehat{\theta}) + \theta^2 - E(\widehat{\theta})^2 + E(\widehat{\theta})^2 = D(\widehat{\theta}) + (b(\widehat{\theta}))^2. \tag{4.5}$$

За непристрасен оценувач $\widehat{\theta}$ за параметарот θ имаме дека $MSE(\widehat{\theta}) = D(\widehat{\theta})$.

Често важи тоа дека распределбата на еден оценувач $\widehat{\theta}$ е приближно нормална со математичко очекување θ и дисперзија $\sigma^2(\theta)/n$. Во тие случаи, дисперзијата се апроксимира добро со $MSE(\widehat{\theta})$, бидејќи компонентата на дисперзија на MSE е многу поголема отколку компонентата на пристрасност (во равенството (4.5)), па затоа $MSE(\widehat{\theta}) \approx D(\widehat{\theta})$. Но, исто така важно е да се забележи дека MSE на еден оценувач може да биде бесконечна дури и кога неговата распределба е приближно нормална.

Дефиниција 4.6. Нека $\widehat{\theta} = U(X_1, X_2, ..., X_n)$ и $\widetilde{\theta} = V(X_1, X_2, ..., X_n)$ се оценувачи за параметарот θ. Тогаш, велиме дека оценувачот $\widehat{\theta}$ е **подобар оценувач** од $\widetilde{\theta}$, ако за секој $\theta \in \Theta$ важи $MSE(\widehat{\theta}) < MSE(\widetilde{\theta})$.

Заради дискусијата по равенството (4.5), во случај на непристрасни оценувачи, претходната дефиниција може да се запише во следниот облик.

Дефиниција 4.6а. Нека $\widehat{\theta} = U(X_1, X_2, ..., X_n)$ и $\widetilde{\theta} = V(X_1, X_2, ..., X_n)$ се два непристрасни оценувачи за параметарот θ. Тогаш, велиме дека оценувачот $\widehat{\theta}$ е **подобар оценувач** од $\widetilde{\theta}$, ако за секој $\theta \in \Theta$ важи $D(\widehat{\theta}) < D(\widetilde{\theta})$.

Ирена Стојковска

Пример 4.8. Нека обележјето X има $\mathcal{P}(\theta)$ распределба, каде $\theta > 0$ е непознат параметар. Тогаш, статистиките $\overline{X}_n$ и S_n^2 се непристрасни оценувачи за θ (покажи!). За нивните дисперзии имаме

$$D(\overline{X}_n) = \frac{\theta}{n}, \quad D(S_n^2) = \frac{\theta}{n(n-1)^2}(n^2(2\theta+1) - 2n(\theta+1) + 1)$$

(покажи!), од каде се гледа дека за секој $n \in \mathbb{N}$ и секој $\theta > 0$ важи $D(\overline{X}_n) < D(S_n^2)$ (покажи!), значи $\overline{X}_n$ е подобар оценувач за θ од S_n^2.

Се поставува прашањето дали во класата на непристрасни оценувачи за некој параметар θ постои оценувач чија дисперзија не е поголема од истата кај останатите непристрасни оценувачи од класата. Имено, таков оценувач не секогаш постои, но ако постои тогаш е единствен со веројатност 1.

Дефиниција 4.7. Нека со $\mathcal{N}(\theta)$ ја означиме класата од сите непристрасни оценувачи за параметарот θ. Оценувачот $\widehat{\theta} \in \mathcal{N}(\theta)$ за кој важи

$$D(\widehat{\theta}) \leq D(\widetilde{\theta}), \text{ за секој } \theta \in \Theta \text{ и } \widetilde{\theta} \in \mathcal{N}(\theta)$$

велиме дека е непристрасен **оценувач со минимална дисперзија.**

Теорема 4.1. *Нека $\widehat{\theta} = U(X_1, X_2, ..., X_n)$ и $\widetilde{\theta} = V(X_1, X_2, ..., X_n)$ се два оценувачи за θ со минимална дисперзија. Тогаш, за секој $\theta \in \Theta$ важи $P\{\widehat{\theta} = \widetilde{\theta}\} = 1$.*

Доказ. Означуваме со $d = D\widehat{\theta} = D\widetilde{\theta}$ и $\overline{\theta} = (\widehat{\theta} + \widetilde{\theta})/2$. Ќе покажеме дека $\overline{\theta}$ е оценувач за θ со минимална дисперзија. Од

$$E(\overline{\theta}) = E((\widehat{\theta} + \widetilde{\theta})/2) = (E(\widehat{\theta}) + E(\widetilde{\theta}))/2 = (\theta + \theta)/2 = \theta,$$

значи $\overline{\theta}$ е непристрасен оценувач за θ. Потоа, бидејќи d е минималната дисперзија за еден оценувач за θ следи дека

$$D(\overline{\theta}) \geq d. \tag{4.6}$$

Од друга страна,

$$\begin{aligned} D(\overline{\theta}) &= D(\frac{\widehat{\theta} + \widetilde{\theta}}{2}) = \frac{1}{4}\,(D\widehat{\theta} + 2cov(\widehat{\theta}, \widetilde{\theta}) + D\widetilde{\theta}) \leq \\ &\leq \frac{1}{4}\,(D\widehat{\theta} + 2\sqrt{\widehat{\theta} \cdot \widetilde{\theta}} + D\widetilde{\theta}) = \frac{1}{4}\,(d + 2\sqrt{d \cdot d} + d) = d, \end{aligned}$$

што значи

$$D(\overline{\theta}) \leq d. \tag{4.7}$$

Ирена Стојковска

Од (4.6) и (4.7) имаме дека $D(\overline{\theta}) = d$ и важи равенство во (4.7), од каде имаме дека $cov(\widehat{\theta}, \widetilde{\theta}) = d$, и затоа

$$D(\widehat{\theta} - \widetilde{\theta}) = D\widehat{\theta} - 2cov(\widehat{\theta}, \widetilde{\theta}) + D\widetilde{\theta} = d - 2d + d = 0.$$

Од последното равенство имаме дека $P\{\widehat{\theta} - \widetilde{\theta} = c\} = 1$ за некоја константата c. Од $E\widehat{\theta} = E\widetilde{\theta} = \theta$ следи дека $c = 0$, па $P\{\widehat{\theta} = \widetilde{\theta}\} = 1$, што требаше да се докаже. ■

4.1.3 Конзистентни оценувачи

Нека $\widehat{\theta}_n = U(X_1, X_2, ..., X_n)$ е оценувач за параметарот θ. Пожелно е распределбата на $\widehat{\theta}_n$ да се концентрира околу вистинската вредност на параметарот θ со зголемување на n. Ова својство на оценувачот $\widehat{\theta}_n$ е познато како конзистентност.

Дефиниција 4.8. Оценувачот $\widehat{\theta}_n$ велиме дека е **конзистентен оценувач** или **стабилен оценувач** за θ, ако за секој $\varepsilon > 0$,

$$\lim_{n\to\infty} P\{|\widehat{\theta}_n - \theta| < \varepsilon\} = 1, \tag{4.8}$$

што значи дека $\widehat{\theta}_n \xrightarrow{P} \theta$.

Да забележиме дека индексот n во $\widehat{\theta}_n$ ја истакнува улогата на големината n на примерокот.

Пример 4.9. Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X со конечни $EX = m$ и $DX = \sigma^2$. Тогаш, непристрасниот оценувач S_n^2 за σ^2 е и конзистентен оценувач за σ^2 (покажи!).

Својство 4.2. *Нека $\widehat{\theta}_n = U(X_1, X_2, ..., X_n)$ е конзистентен оценувач за θ и $\{c_n\}$ е низа реални броеви така што $\lim_{n\to\infty} c_n = 0$. Тогаш, и оценувачот $\widetilde{\theta}_n = \widehat{\theta}_n + c_n = U(X_1, X_2, ..., X_n) + c_n$ е конзистентен оценувач за θ.*

Доказ. Нека $\varepsilon, \eta > 0$. Од $\widehat{\theta}_n \xrightarrow{P} \theta$ и $\lim_{n\to\infty} c_n = 0$, имаме дека $\exists n_0$, така што $\forall n \geq n_0$,

$$P\{|\widehat{\theta}_n - \theta| < \varepsilon/2\} > 1 - \eta, \tag{4.9}$$

$$|c_n| < \varepsilon/2. \tag{4.10}$$

Да забележиме дека

$$|\widetilde{\theta}_n - \theta| = |\widehat{\theta}_n + c_n - \theta| \leq |\widehat{\theta}_n - \theta| + |c_n|. \tag{4.11}$$

Ирена Стојковска

Тогаш, ако $|\widehat{\theta}_n-\theta|<\varepsilon/2$, од (4.10) и (4.11) ќе следи дека $|\widetilde{\theta}_n-\theta|<\varepsilon/2+\varepsilon/2=\varepsilon$. Значи,

$$P\{|\widehat{\theta}_n-\theta|<\varepsilon/2\}\leq P\{|\widetilde{\theta}_n-\theta|<\varepsilon\}\leq 1. \tag{4.12}$$

Сега, од (4.9) и (4.12) следи $\widetilde{\theta}_n \xrightarrow{P} \theta$, што требаше да се докаже. ∎

Во случај кога оценувачот $\widehat{\theta}_n$ има ограничена дисперзија, неговата конзистентност може да се провери со помош на основниот облик на неравенството на Чебишев, односно

$$P\{|\widehat{\theta}_n-\theta|<\varepsilon\}=P\{|\widehat{\theta}_n-\theta|^2<\varepsilon^2\}\geq 1-\frac{E(\widehat{\theta}_n-\theta)^2}{\varepsilon^2}=1-\frac{MSE(\widehat{\theta}_n)}{\varepsilon^2}.$$

Следствено, ако $MSE(\widehat{\theta}_n)\to 0$ тогаш важи (4.8). Потоа, користејќи ја декомпозицијата (4.5) за $\widehat{\theta}=\widehat{\theta}_n$, се добива дека $MSE(\widehat{\theta}_n)\to 0$ ако $D(\widehat{\theta}_n)\to 0$ и $b(\widehat{\theta}_n)\to 0$. Заклучуваме дека во случајот кога $\widehat{\theta}_n$ има ограничена дисперзија, неговата конзистентност може да се провери со проверка на следните услови

$$\text{(а)}\ \lim_{n\to\infty} E(\widehat{\theta}_n)=\theta, \quad \text{(б)}\ \lim_{n\to\infty} D(\widehat{\theta}_n)=0. \tag{4.13}$$

Имено, важи следната теорема.

Теорема 4.2. *Нека $\widehat{\theta}_n = U(X_1, X_2, ..., X_n)$ е оценувач за кој важат условите (4.13). Тогаш, $\widehat{\theta}_n$ е конзистентен оценувач за θ.*

Конзистентноста базирана на условите (4.13) понекогаш се нарекува **средноквадратна конзистентност**.

Дефиниција 4.9. Оценувачот $\widehat{\theta}_n$ се нарекува **силно конзистентен оценувач** за θ, ако

$$P\{\lim_{n\to\infty}\widehat{\theta}_n=\theta\}=1, \tag{4.14}$$

што значи дека $\widehat{\theta}_n \xrightarrow{\text{c.c.}} \theta$.

Пример 4.10. Според Теорема 3.1 имаме дека емпириската функција на распределба на примерокот F_n е силно конзистентен оценувач за функцијата на распределба F на обележјето X. Од таму и оправдувањето на идејата за принципот на замена како метод за наоѓање на оценувачи.

4.1.4 Најефикасни оценувачи

Општиот критериум за споредување на оценувачи за ист непознат параметар θ се темели врз споредбата на нивните средно квадратни грешки (MSE), односно нивните дисперзии доколку тие се непристрасни оценувачи (Дефиниција 4.6-4.6а). Подобриот оценувач се смета за **поефикасен оценувач**. Затоа, природно се наметнува задачата да се најде (ако постои) **најефикасен оценувач** меѓу сите непристрасни оценувачи, односно оценувачот со најмала дисперзија или барем да се одреди долната граница за вредностите на дисперзијата на сите можни непристрасни оценувачи за параметарот θ. Постоењето на таков оценувач се разгледува при одрдедени **услови за регуларност**.

Нека $\mathcal{P} = \{p(x,\theta) : \theta \in \Theta\}$ е допустлива фамилија од распределби за обележјето X, каде $p(x,\theta)$ е допустлива густина на распределба (ако X е непрекинато обележје) или допустлива распределба на веројатност (ако X е дискретно обележје). Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X. Функцијата

$$L(x,\theta) = p(x_1,\theta)p(x_2,\theta)...p(x_n,\theta),\ x = (x_1, x_2, ..., x_n),\ \theta \in \Theta, \tag{4.15}$$

се нарекува **функција на подобност**. Со неа (при фиксирана вредност на θ) е одредена распределбата на примерокот $(X_1, X_2, ..., X_n)$.

Дефиниција 4.10. Нека $f : \Theta \to \mathbb{R}$. Оценувачот $U = U(X_1, X_2, ..., X_n)$ е **регуларен оценувач** за $f(\theta)$ ако важат следните **услови за регуларност**:

(i) множеството $A = \{x = (x_1, x_2, ..., x_n) : L(x,\theta) > 0\}$ не зависи од θ,

(ii) за сите $x \in A$, $L(x,\theta)$ е диференцијабилна по θ,

(iii) може да се диференцира (по θ) под интегралот $\int_{\mathbb{R}^n} L(x,\theta)dx$, односно под сумата $\sum_{x\in\mathbb{R}^n} L(x,\theta)$,

(iv) може да се диференцира (по θ) под интегралот $\int_{\mathbb{R}^n} U(x)L(x,\theta)dx$, односно под сумата $\sum_{x\in\mathbb{R}^n} U(x)L(x,\theta)$,

(v) функцијата $f : \Theta \to \mathbb{R}$ е диференцијабилна.

Теорема 4.3 (**Теорема на Рао-Крамер**). *Нека* $\mathbf{X} = (X_1, X_2, ..., X_n)$ *е примерок и* $U = U(X_1, X_2, ..., X_n)$ *е непристрасен регуларен оценувач за* $f(\theta)$ *и нека* U *има конечен втор момент. Тогаш, за секој* $\theta \in \Theta$ *важи*

$$D(U) \geq \frac{(f'(\theta))^2}{E(\frac{\partial}{\partial\theta}\ln L(\mathbf{X},\theta))^2} = D_0. \tag{4.16}$$

Равенство важи ако и само ако $\frac{\partial}{\partial\theta}\ln L(\mathbf{X},\theta) = h(\theta)(U - f(\theta))$, *за некоја функција* $h : \Theta \to \mathbb{R}$.

Ирена Стојковска

Доказ. Бидејќи $L(x,\theta)$ е функција на подобност и U е непристрасен оценувач за $f(\theta)$, следи дека за секој $\theta \in \Theta$ важат следните равенства

$$\int_{\mathbb{R}^n} L(x,\theta)dx = 1 \text{ и } \int_{\mathbb{R}^n} U(x)L(x,\theta)dx = f(\theta). \tag{4.17}$$

Ако равенствата (4.17) ги диференцираме по θ и користејќи ги условите за регуларност ќе добиеме

$$0 = \frac{\partial}{\partial\theta}\int_{\mathbb{R}^n} L(x,\theta)dx = \int_{\mathbb{R}^n} \frac{\partial L}{\partial\theta}\,dx = \int_{\mathbb{R}^n} (\frac{1}{L}\,\frac{\partial L}{\partial\theta})Ldx = \int_{\mathbb{R}^n} \frac{\partial \ln L}{\partial\theta}\,Ldx, \tag{4.18}$$

$$f'(\theta) = \frac{\partial}{\partial\theta}\int_{\mathbb{R}^n} U(x)L(x,\theta)dx = \int_{\mathbb{R}^n} U\,\frac{\partial L}{\partial\theta}\,dx = \int_{\mathbb{R}^n} U\,\frac{\partial \ln L}{\partial\theta}\,Ldx. \tag{4.19}$$

Од (4.18) и (4.19) имаме

$$\begin{aligned} f'(\theta) &= \int_{\mathbb{R}^n} U\,\frac{\partial \ln L}{\partial\theta}\,Ldx = \\ &= \int_{\mathbb{R}^n} U\,\frac{\partial \ln L}{\partial\theta}\,Ldx - f(\theta)\int_{\mathbb{R}^n} \frac{\partial \ln L}{\partial\theta}\,Ldx = \\ &= \int_{\mathbb{R}^n} (U - f(\theta))\,\frac{\partial \ln L}{\partial\theta}\,Ldx. \end{aligned}$$

Од непристрасноста на U и (4.18) имаме дека $E(U - f(\theta)) = 0$ и $E(\frac{\partial \ln L}{\partial\theta}) = 0$, па затоа

$$\begin{aligned} (f'(\theta))^2 &= \Big(\int_{\mathbb{R}^n} (U - f(\theta))\,\frac{\partial \ln L}{\partial\theta}\,Ldx\Big)^2 \leq \\ &\leq \int_{\mathbb{R}^n} (U - f(\theta))^2 Ldx \cdot \int_{\mathbb{R}^n} (\frac{\partial \ln L}{\partial\theta})^2 Ldx = DU \cdot E(\frac{\partial \ln L}{\partial\theta})^2. \end{aligned}$$

Значи,

$$DU \geq \frac{(f'(\theta))^2}{E(\frac{\partial \ln L}{\partial\theta})^2}$$

т.е. важи (4.16), што требаше да се покаже. ∎

Неравенството (4.16) е познато како **неравенство на Рао-Крамер**. При тоа важи

$$E(\frac{\partial}{\partial\theta}\ln L(\mathbf{X},\theta))^2 = n\,I(\theta), \tag{4.20}$$

каде

$$I(\theta) = E(\frac{\partial}{\partial\theta}\ln p(x,\theta))^2, \tag{4.21}$$

е **информација на Фишер**.

Ирена Стојковска

Дефиниција 4.11. Регуларниот непристрасен оценувач U за $f(\theta)$ е **најефикасен** ако $D(U) = D_0$, односно ако за неговата дисперзија важи равенство во неравенството на Рао-Крамер.

Дефиниција 4.12. Нека U е регуларен непристрасен оценувач за $f(\theta)$ со конечна дисперзија, тогаш количникот

$$e(U) = \frac{D_0}{D(U)}$$

се нарекува **ефкасност** на оценувачот U.

Јасно е дека важи $0 \leq e(U) \leq 1$. За најефикасен оценувач U важи $e(U) = 1$. Ако $\lim_{n\to\infty} e(U_n) = 1$, тогаш за оценувачот U_n велиме дека е **асимптотски најефикасен оценувач**.

Пример 4.11. Нека $\mathbf{X} = (X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето $X \sim \mathcal{N}(\theta, \sigma^2)$, каде $\theta > 0$ е непознат параметар, додека σ^2 е познат параметар. Непознатиот параметар го оценуваме со средината на примерокот $\overline{X}_n$. Важат условите за регуларност (покажи!). Да испитаме дали за овој оценувач важи равенство во неравенството на Рао-Крамер. Имаме,

$$L(\mathbf{X}, \theta) = \Big(\frac{1}{2\pi\sigma^2}\Big)^{n/2} \exp\Big\{ -\frac{1}{2\sigma^2}\Big(\sum_{i=1}^{n} X_i^2 - 2\theta \sum_{i=1}^{n} X_i + n\theta^2 \Big)\Big\},$$

од каде се добива дека

$$\ln L(\mathbf{X}, \theta) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\Big(\sum_{i=1}^{n} X_i^2 - 2\theta \sum_{i=1}^{n} X_i + n\theta^2 \Big),$$

$$\frac{\partial}{\partial\theta}\ \ln L(\mathbf{X}, \theta) = -\frac{1}{2\sigma^2}\Big(-2\sum_{i=1}^{n} X_i + 2n\theta \Big) = \frac{n}{\sigma^2}(\overline{X}_n - \theta),$$

$$E\Big(\frac{\partial}{\partial\theta}\ \ln L(\mathbf{X}, \theta)\Big)^2 = \frac{n^2}{\sigma^4} E(\overline{X}_n - \theta)^2 = \frac{n^2}{\sigma^4} E(\overline{X}_n^2 - 2\theta\overline{X}_n + \theta^2).$$

Бидејќи $E(\overline{X}_n) = \theta$ и $D(\overline{X}_n) = \frac{\sigma^2}{n}$, следи дека $E(\overline{X}_n^2) = \frac{\sigma^2}{n} + \theta^2$. Затоа,

$$E\Big(\frac{\partial}{\partial\theta}\ \ln L(x, \theta)\Big)^2 = \frac{n^2}{\sigma^4} \cdot (\frac{\sigma^2}{n} + \theta^2 - 2\theta^2 + \theta^2) = \frac{n}{\sigma^2}.$$

Функцијата $f(\theta) = \theta$ (во неравенството на Рао-Крамер), па затоа $f'(\theta) = 1$, и

$$D_0 = \frac{(f'(\theta))^2}{E(\frac{\partial}{\partial\theta}\ln L(\mathbf{X}, \theta))^2} = \frac{1}{\frac{n}{\sigma^2}} = \frac{\sigma^2}{n} = D(\overline{X}_n),$$

од каде заклучуваме дека $\overline{X}_n$ е најефикасен оценувач за θ.

Ирена Стојковска

Од Теорема 4.1 следи следното својство за единственост на најефикасен оценувач.

Својство 4.3. *Најефикасниот оценувач за $f(\theta)$, ако постои, е единствен со веројатност 1.*

Доказ. Нека U и V се два најефикасни оценувачи за $f(\theta)$, тогаш тие се непристрасни оценувачи за $f(\theta)$ и $DU = DV = D_0$, каде D_0 е дефинирано со (4.16). Значи, U и V се оценувачи со минимална дисперзија, па од Теорема 4.1 следи дека $P\{U = V\} = 1$, што требаше да се докаже. ∎

Од Теорема 4.3 следи следното својство за линеарна зависност меѓу две функции $f(\theta)$ и $g(\theta)$, ако постојат најефикасни оценувачи за двете од нив.

Својство 4.4. *Ако помеѓу функциите $f(\theta)$ и $g(\theta)$ не постои линеарна зависност, тогаш може да постои најефикасен оценувач за само една од нив.*

Доказ. Нека U е најефикасен оценувач за $f(\theta)$ и V е најефикасен оценувач за $g(\theta)$. Тогаш, од Теорема 4.3 следи дека

$$\frac{\partial}{\partial\theta}\ln L(\mathbf{X},\theta) = h_1(\theta)(U - f(\theta)),\ \frac{\partial}{\partial\theta}\ln L(\mathbf{X},\theta) = h_2(\theta)(V - g(\theta)),$$

од каде

$$h_1(\theta)(U - f(\theta)) = h_2(\theta)(V - g(\theta)),$$

односно

$$h_1(\theta)U - h_2(\theta)V - h_1(\theta)f(\theta) + h_2(\theta)g(\theta) = 0,$$

па U и V се линеарно зависни. ∎

4.2 Доволни статистики

Нека $\mathcal{P} = \{F(x,\theta) : \theta \in \Theta\}$ е фамилијата од допустливи распределби за обележјето X, и нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X. Нека $L(x,\theta)$ е распределбата на примерокот $(X_1, X_2, ..., X_n)$, односно функцијата на подобност.

Целта ни е да најдеме статистика $U = U(X_1, X_2, ..., X_n)$ која содржи иста информација за непознатиот параметар θ како и примерокот $(X_1, X_2, ..., X_n)$, односно при оценување на непознатиот параметар со помош на таа статистика да не изгубиме дел од информацијата која со себе ја носи примерокот. Во прилог на ова размислување е поимот за доволност.

Ирена Стојковска

Прв кој ја вовел доволноста е Фишер (1920) при оценување на дисперзијата σ^2 кај обележје $X \sim \mathcal{N}(m, \sigma^2)$. Тој ја оценувал σ^2 врз база примерокот $(X_1, X_2, ..., X_n)$ со помош на статистиките

$$U_1 = \sum_{i=1}^{n} |X_i - \overline{X}_n| \text{ и } U_2 = \sum_{i=1}^{n} (X_i - \overline{X}_n)^2,$$

каде $\overline{X}_n$ е средината на примерокот $(X_1, X_2, ..., X_n)$. Покажал дека условната распределба на U_1 при услов $U_2 = t$ не зависи од параметарот σ^2, додека условната распределба на U_2 при услов $U_1 = t$ зависи од σ^2. Со тоа, тој заклучил дека сета информација за σ^2 од примерокот е содржана во статистиката U_2, што значи дека оценувањето на σ^2 со U_1 може да се подобри со користење на информацијата во U_2, додека оценувањето со U_2 не може да биде подобрено користејќи се со U_1. Значи, оценувањето на σ^2 треба да биде базирано на U_2.

Дефиниција 4.13. Статистика $U = U(X_1, X_2, ..., X_n)$ е **доволна статистика** за параметарот θ, ако за секој $\theta \in \Theta$ условната распределба на примерокот $(X_1, X_2, ..., X_n)$ при услов $U = t$ т.е. $L(x, \theta | U = t)$ не зависи од θ.

Пример 4.12. Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето $X \sim \mathcal{B}(m, \theta)$, каде $0 < \theta < 1$ е непознат параметар. Тогаш, распределбата на примерокот $(X_1, X_2, ..., X_n)$ е

$$L(x, \theta) = P((X_1, ..., X_n) = (x_1, ..., x_n)) = \prod_{i=1}^{n} \binom{m}{x_i} \theta^{x_i} (1 - \theta)^{m - x_i},$$

за $x_i \in \{0, 1, ..., m\}$, $i = 1, 2, ..., n$. Да провериме дали статистиката $U = X_1 + X_2 + ... + X_n$ е доволна статистика за θ. Од независноста на $X_1, X_2, ..., X_n$ следи дека $U \sim \mathcal{B}(nm, \theta)$.

Нека $(x_1, ..., x_n)$ е произволен елемент од доменот и нека $t = x_1 + ... + x_n$. Тогаш, условната распределба на примерокот при услов $U = t$ е следната.

$$P((X_1, ..., X_n) = (x_1, ..., x_n) | U = t) = \frac{P((X_1, ..., X_n) = (x_1, ..., x_n), U = t)}{P(U = t)} =$$

$$= \frac{P((X_1, ..., X_n) = (x_1, ..., x_n), X_1 + ... + X_n = x_1 + ... + x_n)}{P(U = t)} =$$

$$= \frac{P((X_1, ..., X_n) = (x_1, ..., x_n))}{P(U = t)} = \frac{\prod_{i=1}^{n} \binom{m}{x_i} \theta^{x_i} (1 - \theta)^{m - x_i}}{\binom{nm}{t} \theta^t (1 - \theta)^{nm - t}} =$$

$$= \frac{\left(\prod_{i=1}^{n} \binom{m}{x_i} \right) \theta^{x_1 + ... + x_n} (1 - \theta)^{nm - (x_1 + ... + x_n)}}{\binom{nm}{t} \theta^t (1 - \theta)^{nm - t}} = \frac{\prod_{i=1}^{n} \binom{m}{x_i}}{\binom{nm}{t}},$$

Ирена Стојковска

што значи не зависи од θ. Во случај кога $t \neq x_1 + ... + x_n$ имаме дека условната распределба е 0, па тривијално следи дека не зависи од θ. Значи, според дефиницијата за доволна статистика, заклучуваме дека $U = X_1 + ... + X_n$ е доволна статитика за θ.

Проверката со помош на дефиницијата дали една статистика е доволна наидува на проблем при наоѓање на условната распределба посебно кога обележјето X е непрекинато. Постои поедноставен критериум за проверка дали една статистика е доволна - Теоремата за факторизација.

Теорема 4.4 (**Теорема за факторизација**). *Статистика $U = U(X_1, X_2, ..., X_n)$ е доволна статистика за параметарот θ ако и само ако функцијата на подобност $L(x, \theta)$ може да се запише во облик*

$$L(x, \theta) = g(U(x), \theta) \cdot h(x),$$

каде функцијата h не зависи од параметарот θ.

Доказ. Доказот ќе го споведене за случајна променлива од дискретен тип. Нека U е доволна статистика и нека $x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$ е таков што $U(x_1, x_2, ..., x_n) = t$. Тогаш,

$$\begin{aligned} L(x, \theta) &= P\{(X_1, X_2, ..., X_n) = x\} = P\{(X_1, X_2, ..., X_n) = x,\ U = t\} = \\ &= P\{U = t\} \cdot P\{(X_1, X_2, ..., X_n) = x \mid U = t\} = \\ &= g(t, \theta) \cdot L(x, \theta \mid U = t) = g(U(x), \theta) \cdot h(x), \end{aligned}$$

затоа што $L(x, \theta \mid U = t)$ не зависи од θ.

Нека сега $L(x, \theta) = g(U(x), \theta) \cdot h(x)$ и нека $x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$ е таков што $U(x_1, x_2, ..., x_n) = t$. Тогаш,

$$\begin{aligned} L(x, \theta \mid U = t) &= P\{(X_1, X_2, ..., X_n) = x \mid U = t\} = \\ &= \frac{P\{(X_1, X_2, ..., X_n) = x\ U = t\}}{P\{U = t\}} = \frac{L(x, \theta)}{\sum_{y:\, U(y)=t} L(y, \theta)} = \\ &= \frac{g(U(x), \theta) \cdot h(x)}{\sum_{y:\, U(y)=t} g(U(y), \theta) \cdot h(y)} = \frac{h(x)}{\sum_{y:\, U(y)=t} h(y)}, \end{aligned}$$

не зависи од θ, па селди дека U е доволна статистика за θ.

Ако $U(x) \neq t$, тогаш $P\{(X_1, X_2, ..., X_n) = x \mid U = t\} = 0$, и доказот е тривијален. ∎

Пример 4.13. Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето $X \sim \mathcal{U}(0, \theta)$, каде $\theta > 0$ е непознат параметар. Да се обидеме со критериумот за факторизација да најдеме една доволна статистика за θ. Распределбата на примерокот $(X_1, X_2, ..., X_n)$ е

$$L(x, \theta) = \frac{1}{\theta^n},\ 0 < x_i < \theta,\ i = 1, ..., n,$$

па таа може да се запише како

$$L(x, \theta) = \frac{1}{\theta^n} I\{0 < x_1 < \theta, ..., 0 < x_n < \theta\} =$$

$$= \frac{1}{\theta^n} I\{\max_{1 \le i \le n} x_i < \theta\} I\{\min_{1 \le i \le n} x_i > 0\} =$$

$$= g(\max\{x_1, ..., x_n\}, \theta) h(x),$$

од каде, според теоремата за факторизација, следи дека подредената статистика $X_{(n)} = \max\{X_1, ..., X_n\}$ е доволна статистика за θ.

Според дефиницијата за доволност, доволните статистики не се единствени. Се покажува дека ако U е доволна статистика и g е биекција, тогаш и $g(U)$ е доволна статистика. Со ова се потенцира поголемото значење на просторот на примерокот индуциран од доволната статистика U (т.е. просторот составен од множествата $\{x : U(x) = t\}$), отколку самата доволна статистика.

Својство 4.5. *Ако $U = U(X_1, X_2, ..., X_n)$ е доволна статистика за параметарот θ и g е биекција, тогаш и статистиката $V = g(U)$ е доволна статистика за θ.*

Доказ. Нека U е доволна статистика за θ. Од Теорема 4.4, следи дека

$$L(x, \theta) = g(U(x), \theta) \cdot h(x).$$

Нека $V = g(U)$, при што g е биекција. Нека s е инверзната функција на g т.е. $U = s(V)$. Тогаш,

$$L(x, \theta) = g(s(V(x)), \theta) \cdot h(x) = f_1(V(x), \theta) \cdot h(x),$$

па според Теорема 4.4 следи дека V е доволна статистика за θ. ∎

Доволните статистики ни помагаат да најдеме непристрасен оценувач со помала дисперзија од дисперзијата на некој даден непристрасен оценувач, имено тој оценувач е функција од некоја доволна статистика. Исто така важи дека, ако постои оценувач со минимална дисперзија, тогаш тој е функција од некоја доволна статистика.

Ирена Стојковска

Теорема 4.5. *Нека T е доволна статистика за параметарот θ и U е непристрасен оценувач за θ. Тогаш, условното математичко очекување $Q(T) = E(U|T)$ е непристрасен оценувач за θ и за секој $\theta \in \Theta$ важи $D(Q(T)) \leq D(U)$. Равенство важи ако и само ако $P\{U = Q(T)\} = 1$.*

Доказ. Од тоа што T е доволна статистика за θ следи дека условната распределба на U при услов $T = t$ не зависи од θ, а со тоа и $E(U|T) = Q(T)$ не зависи од θ и може да се смета за оценувач за θ. Од непристрасноста на U, имаме

$$E(Q(T)) = E(E(U|T)) = E(U) = \theta,$$

од каде следи дека $Q(T)$ е непристрасен оценувач за θ. Понатаму,

$$D(Q(T)) = D(E(U|T) \leq D(U),$$

а равенство се достигнува акоми само ако $Q(T)$ и U се оценувачи со минимална дисперзија, што според Теорема 4.1 е еквивалентно со $P\{U = Q(T)\} = 1$, што требаше да се докаже. ∎

4.3 Методи за наоѓање на оценки

4.3.1 Метод на моменти

Нека $(X_1, X_2, ..., X_n)$ е случаен примерок кој одговара на обележјето X со функција на распределба F која зависи од непознатите праметри $\theta_1, ..., \theta_r$. Се прашуваме дали може принципот на замена да се примени при оценување на непознатите параметри $\theta_1, ..., \theta_r$ во еден параметарски модел.

Еден пристап (за кој зборевме претходно) е ако може да се изразат $\theta_1, ..., \theta_r$ како функционали од F, а потоа со замена на оценувач $\widehat{F}$ на местото од F се добиваат оценувачи $\widehat{\theta}_1, ..., \widehat{\theta}_r$ за $\theta_1, ..., \theta_r$ соодветно.

Обопштување на овој пристап е да при оценување на $\theta = (\theta_1, ..., \theta_r)$ успееме да најдеме r функционални параметри од F, $\eta_1(F), ..., \eta_r(F)$ кои зависат од θ, односно

$$\eta_k(F) = g_k(\theta),\ k = 1, ..., r, \tag{4.22}$$

каде $g_1, ..., g_r$ се познати функции. На овој начин се добива систем од r равенки со r непознати $\theta_1, ..., \theta_r$. Користејќи го примципот на замена, може да ги оцениме $\eta_1(F), ..., \eta_r(F)$ со $\eta_1(\widehat{F}), ..., \eta_r(\widehat{F})$ соодветно. Ако за секои $\eta_1(F), ..., \eta_r(F)$ постои единствено решение на системот (4.22), тогаш оценувачот $\widehat{\theta}$ на θ го дефинираме така да го задоволува системот

$$\eta_k(\widehat{F}) = g_k(\widehat{\theta}),\ k = 1, ..., r. \tag{4.23}$$

Еден избор за $\eta_k(F),\ k=1,...,r$ е k-тиот момент на обележјето X, односно $\eta_k(F)=E(X^k),\ k=1,...,r$. Тогаш, според принципот на замена оценувачи за $\eta_k(F),\ k=1,...,r$ се k-тите моменти на примерокот $(X_1,X_2,...,X_n)$, односно $\eta_k(\widehat{F})=\frac{1}{n}\sum_{i=1}^{n}X_i^k=Z_{n,k},\ k=1,...,r$ соодветно. Затоа, овој метод за наоѓање на оценувач е познат како **метод на моменти**. При тоа значи, оценувачот $\widehat{\theta}$ на θ најден со методот на моменти го задоволува системот

$$Z_{n,k}=g_k(\widehat{\theta}),\ k=1,...,r. \tag{4.24}$$

Пример 4.14. Нека обележјето X има $\mathcal{N}(m,\sigma^2)$, каде m и σ^2 се непознати параметри. Нека $(X_1,X_2,...,X_n)$ е примерок кој одговара на ова обележје. Тогаш, за наоѓање на оценувачи на параметрите m и σ^2 со метод на моменти, прво го составуваме следниот систем равенки по m и σ^2.

$$\begin{cases} E(X)=m \\ E(X^2)=D(X)+(E(X))^2=\sigma^2+m^2 \end{cases}$$

Потоа, според принсипот на замена, $E(X)$ и $E(X^2)$ ги заменуваме со нивните оценувачи $Z_{n,1}$ и $Z_{n,2}$ соодветно, додека непознатите параметри m и σ^2 ги заменуваме со нивните оценувачи $\widehat{m}$ и $\widehat{\sigma^2}$ соодветно. Така го добиваме следниот систем.

$$\begin{cases} Z_{n,1}=\widehat{m} \\ Z_{n,2}=\widehat{\sigma^2}+\widehat{m}^2 \end{cases}$$

Решавајќи го последниот систем по $\widehat{m}$ и $\widehat{\sigma^2}$ ги добиваме оценувачите за m и σ^2 со метод на моменти. Тоа се,

$$\begin{cases} \widehat{m}=Z_{n,1}=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}X_i=\overline{X}_n \\ \widehat{\sigma^2}=Z_{n,2}-\widehat{m}^2=Z_{n,2}-Z_{n,1}^2=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}X_i^2-\Big(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}X_i\Big)^2=\overline{S}_n^2 \end{cases}$$

Со методот на моменти се добиваат конзистентни оценувачи за оценуваните параметри.

Теорема 4.6. *Нека обележјето X има функција на распределба F која зависи од r непознати параметри $\theta_1,...,\theta_r$, нека X има конечни моменти $m_k=E(X^k)$, $k=1,...,r$ и нека параметрите $\theta_1,...,\theta_r$ може да се запишат како непрекинати функции од моментите $m_1,...,m_r$, $\theta_k=\theta_k(m_1,...,m_r)$, $k=1,...,r$. Тогаш, оценувачите $\widehat{\theta}_k=\theta_k(Z_{n,1},...,Z_{n,r})$, $k=1,...,r$ добиени со методот на моменти се конзистентни оценувачи за параметрите θ_k, $k=1,...,r$ соодветно.*

Ирена Стојковска

Доказ. Од Законот на големите броеви имаме дека $\forall \varepsilon > 0$,

$$\lim_{n\to\infty} P\{|Z_{n,k} - m_k| < \varepsilon\} = 1,\ k = 1, 2, ..., r.$$

Од непрекинатоста на $\theta_k = \theta_k(m_1, ..., m_r),\ k = 1, ..., r$, имаме дека $\forall \varepsilon > 0,\ \exists \delta > 0$,

$$\Big(|x_1 - m_1| < \delta, ..., |x_r - m_r| < \delta\Big) \Longrightarrow |\theta_k(x_1, ..., x_r) - \theta_k(m_1, .., m_r)| < \varepsilon,\ k = 1, ..., r.$$

Ако ставиме $x_k = Z_{n,k},\ k = 1, ..., r$ добиваме дека

$$\bigcap_{k=1}^{r} \{|Z_{n,k} - m_k| < \delta\} \subseteq \{|\widehat{\theta}_k - \theta_k| < \varepsilon\},$$

односно

$$\{|\widehat{\theta}_k - \theta_k| \geq \varepsilon\} \subseteq \bigcup_{k=1}^{r} \{|Z_{n,k} - m_k| \geq \delta\},$$

од каде

$$P\{|\widehat{\theta}_k - \theta_k| \geq \varepsilon\} \leq P\Big(\bigcup_{k=1}^{r} \{|Z_{n,k} - m_k| \geq \delta\}\Big) \leq \sum_{k=1}^{r} P\{|Z_{n,k} - m_k| \geq \delta\} \to 0,$$

па следи дека $\widehat{\theta}_k$ е конзистентен оценувач за $\theta_k,\ k = 1, ..., r$. ∎

4.3.2 Метод на максимална подобност

Како што видовме, со принципот на замена се наоѓаат задоволително добри оценувачи за непознатите параметри. Квалитетот на најдените оценувачи со принципот на замена може многу да варира во зависност од изборот на оценувачот за функцијаата на распределба F или изборот на функционалните параметри од F. Исто така принципот на замена тешко може да се примени кај нееднакво распределени случајни променливи од примерокот. Овие проблеми се на некој начин надминати со методот на максимална подобност кој многу често наоѓа не само задоволително добри оценувачи, туку и оценувачи со оптимални својства. Оценувачот најден со методот на максимална подобност се дефинира така да ја максимизита вредноста на одредена функција наречена функција на подобност. Всушност, овој метод се базира на еден важен принцип во статистиката, принципот на подобност, кој вели дека функцијата на подобност ги содржи во себе сите информации за непознатиот параметар, предмет на оценување.

Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X со распределба на веројатност $p(x, \theta)$ која припаѓа на фамилијата од допустливи распределби на веројатност $\mathcal{P} = \{p(x, \theta) : \theta \in \Theta\}$, каде θ е непознат параметар.

Тогаш, функцијата на распределба на примерокот $(X_1, X_2, ..., X_n)$ се нарекува **функција на подобност**, се означува со $L(x, \theta)$, $x = (x_1, x_2, ..., x_n)$. Во случај на прост случаен примерок, имаме

$$L(x, \theta) = p(x_1, \theta)p(x_2, \theta)...p(x_n, \theta),\ x = (x_1, x_2, ..., x_n),\ \theta \in \Theta. \tag{4.25}$$

Често, за дадена реализација $x = (x_1, x_2, ..., x_n)$ на примерокот $(X_1, X_2, ..., X_n)$, функцијата на подобност се третира како функција од параметарот θ т.е.

$$\mathcal{L}(\theta) = L(x, \theta),\ \theta \in \Theta. \tag{4.26}$$

На овој начин функцијата на подбност може да се "прочита" како веројатноста дека случајниот вектор $(X_1, X_2, ..., X_n)$ ќе ја прими вредноста $(x_1, x_2, ..., x_n)$. При тоа, за различни вредности на $\theta \in \Theta$ се добиваат различни веројатности. Па, може да се каже дека функцијата на подобност мери колку е параметарот θ погоден за добивање на конкретната реализација $(x_1, x_2, ..., x_n)$ на примерокот $(X_1, X_2, ..., X_n)$.

Според **методот на максимална подобност** (maximum likelihood - ML), за **максимално подобен оценувач** (maximum likelihood estimator - MLE) $\widehat{\theta}$ на параметарот θ се зема статистиката $\widehat{\theta} = \widehat{\theta}(X_1, X_2, ..., X_n)$ за која важи

$$L(X_1, X_2, ..., X_n, \widehat{\theta}) = \max_{\theta} L(X_1, X_2, ..., X_n, \theta). \tag{4.27}$$

Постојат главно два начина за барање на ML оценувачите (решавање на задачата (4.27)). Едниот е со директна максимизација на функцијата на подобност $L(X_1, X_2, ..., X_n, \theta)$ за дадена реализација $(x_1, x_2, ..., x_n)$ на примерокот $(X_1, X_2, ..., X_n)$, и овој начин е корисен кога доменот на податоците зависи од непознатиот параметар.

Вториот начин се применува кога доменот на податоците не зависи од непознатиот параметар и функцијата на подобност е диференцијабилна по θ на множеството Θ, и тогаш ML оценувачот се наоѓа како решение на **равенката на подобност**

$$\frac{\partial \ln L(X_1, X_2, ..., X_n, \theta)}{\partial \theta} = 0. \tag{4.28}$$

Потврда за овој начин ни дава фактот дека $\ln : \mathbb{R}^+ \to \mathbb{R}$ е растечка функција, па затоа функциите $L(X_1, X_2, ..., X_n, \theta)$ и $\ln L(X_1, X_2, ..., X_n, \theta)$ имаат исти точки на максимум. Равенката на подобност може да има повеќе решенија, па во тој случај треба да се провери кое од нив навистина ја максимизира функцијата на подобност. Исто така, ако просторот од параметри Θ не е отворено множество по решавањето на равенката на подобност ќе треба да се провериат и граничните услови, односно дали може максимумот да се достигне на границата од просторот на параметри.

Ирена Стојковска

Пример 4.15. Нека обележјето X има $\mathcal{U}(0,\theta)$ распределба, каде $\theta > 0$ е непознат параметар. Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X. Тогаш, функцијата на подобност е

$$L(x,\theta) = \left(\frac{1}{\theta}\right)^n = \left(\frac{1}{\theta}\right)^n I\{0 \le x_1, ..., x_n \le \theta\} = \left(\frac{1}{\theta}\right)^n I\{\max\{x_1, ..., x_n\} \le \theta\},$$

за $0 \le x_1, ..., x_n \le \theta$. Кога $\max\{x_1, ..., x_n\} > \theta$, тогаш $L(x,\theta) = 0$. Бидејќи, за $\max\{x_1, ..., x_n\} \le \theta$, $L(x,\theta)$ е опаѓачка функција по θ, заклучуваме дека ML оценувач за θ е

$$\widehat{\theta} = \max\{X_1, ..., X_n\} = X_{(n)}.$$

Пример 4.16. Нека обележјето X има $\mathcal{P}(\lambda)$ распределба, каде $\lambda > 0$ е непознат параметар. Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X. Тогаш, функцијата на подобност е

$$L(x,\lambda) = \prod_{i=1}^{n} \left(\frac{\lambda^{x_i}}{x_i!}\, e^{-\lambda}\right) = \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}\, e^{-n\lambda}.$$

Логоритам од функцијата на подобност е

$$\ln L(x,\lambda) = \ln\lambda \sum_{i=1}^{n} x_i - \ln\Big(\prod_{i=1}^{n} x_i!\Big) - n\lambda.$$

Бараме извод по λ, и добиваме

$$\frac{\partial}{\partial\lambda} \ln L(x,\lambda) = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n.$$

Решавајќи ја равенката на подобност $\frac{\partial}{\partial\lambda} \ln L(X_1, ..., X_n, \lambda) = 0$, добиваме ML оценувач $\widehat{\lambda} = \frac{1}{n}\sum_{i=1}^n X_i = \overline{X}_n$ за λ. Лесно се проверува дека при $\sum_{i=1}^n x_i > 0$ овој оценувач навистина ја максимизира функцијата на подобност. Имено,

$$\frac{\partial^2}{\partial\lambda^2} \ln L(x,\lambda) = -\frac{1}{\lambda^2}\sum_{i=1}^{n} x_i < 0.$$

Значи, $\widehat{\lambda} = \overline{X}_n$ е ML оценувач за λ, ако $\sum_{i=1}^n X_i > 0$. Ако пак $\sum_{i=1}^n X_i = 0$, тогаш не постои ML оценувач за λ, затоа што логоритамот од функцијата на подобност $\ln L(x,\lambda) = -n\lambda$ нема максимум на интервалот $(0, +\infty)$, а и доменот на податоците не зависи од непознатиот параметар, па не може да се примени директна максимизација.

Ирена Стојковска

Со методот на максимална подобност не мора да се добие само еден максимално подобен оценувач.

Пример 4.17. Нека обележјето X има $\mathcal{U}(\theta-1,\theta+1)$ распределба, каде $\theta\in\mathbb{R}$ е непознат параметар. Нека $(X_1,X_2,...,X_n)$ е примерок кој одговара на обележјето X. Тогаш, секоја статистика $U=U(X_1,...,X_n)$ за која важи

$$X_{(n)}-1<U<X_{(1)}+1$$

е максимално подобен оценувач за θ, а такви статистики ги има повеќе. На пример, за секој $t\in(0,1)$, статистиките

$$U_t=X_{(n)}-1+t(X_{(1)}-X_{(n)}+2)$$

се ML оценувачи за θ.

Максимално подобните оценувачи поседуваат одредени убави својства. Конкретно, секој најефикасен оценувач е максимално подобен оценувач.

Својство 4.6. *Нека U е најефикасен оценувач за параметарот θ. Тогаш, U е единствено решение на равенката на подобност.*

Доказ. Според Теоремата на Рао-Крамер (Теорема 4.3), U е најефикасен оценувач за θ ако и само ако важи равенство во неравенството на Рао-Крамер, т.е. ако

$$\frac{\partial}{\partial\theta}L(x,\theta)=h(\theta)(U-\theta)$$

за некоја функција $h(\theta)$. Тогаш, равенката на подобност

$$\frac{\partial L(X_1,X_2,...,X_n,\theta)}{\partial\theta}=0$$

преминува во

$$h(\theta)(U(X_1,...,X_n)-\theta)=0$$

чие решение е $\theta=U$. Единствеността следи од единствеността (со веројатност 1) на најефикасен оценувач. ∎

Максимално подобните оценувачи се функција од доволна статистика.

Својство 4.7. *Нека U е доволна статитика за параемтарот θ, тогаш секое решение на равенката на подобност е функција од U.*

Ирена Стојковска

Доказ. Нека U е доволна статистика за параметарот θ. Од Теоремата за факторизација (Теорема 4.4), следи

$$L(x,\theta) = g(U(x),\theta)\cdot h(x).$$

Тогаш, равенката на подобност

$$\frac{\partial L(X_1,X_2,...,X_n,\theta)}{\partial\theta} = 0$$

преминува во

$$\frac{\partial g(U(X_1,X_2,...,X_n),\theta)}{\partial\theta} = 0,$$

од каде следи дека секое решение (по θ) е функција од U. ∎

Меѓутоа, секој ML оценувач не мора да биде доволна статистика.

Пример 4.18. Нека обележјето X има $\mathcal{U}[\theta,\theta+1]$ распределба, каде $\theta\in\mathbb{R}$ е непознат параметар. Нека $(X_1,X_2,...,X_n)$ е примерок кој одговара на обележјето X. Тогаш, функцијата на подобност на примерокот е

$$L(x,\theta) = 1,\ \text{за}\ \theta\le x_1,...,x_n\le\theta+1,$$

па секоја статистика $U=U(X_1,...,X_n)$ за која важи

$$X_{(n)}-1\le U\le X_{(1)}$$

е максимално подобен оценувач за θ. Според тоа и $X_{(1)}$ е *ML* оценувач за θ, но $X_{(1)}$ не е доволна статитика. Имено, густината на распределба на $X_{(1)}$ е

$$p(t,\theta) = n(\theta+1-t)^{n-1},\ \theta\le t\le\theta+1,$$

па условната распределба на примерокот $(X_1,X_2,...,X_n)$ при услов $X_{(1)}=t$ е

$$\frac{L(x,\theta)}{p(t,\theta)} = \frac{1}{n(\theta+1-t)^{n-1}},\ \text{за}\ \theta\le x_1,...,x_n\le\theta+1\ \text{и}\ \theta\le t\le\theta+1$$

зависи од θ.

Друго убаво својство на ML оценувачите е дека тие се инваријантни во однос на трансформациите.

Својство 4.8. *Нека U е максимално подобен оценувач за параметарот θ, и нека f е монотона функција (или поопшто, биекција), тогаш $f(U)$ е максимално подобен оценувач за $f(\theta)$.*

Ирена Стојковска

Доказ. Од f биекција, постои нејзина инверзна функција f^{-1}. Нека $f(\theta) = \theta^*$, тогаш $\theta = f^{-1}(\theta^*)$, па

$$L(x, \theta) = L(x, f^{-1}(\theta^*)) = L^*(x, \theta^*).$$

Тогаш,

$$\max\{L(x, \theta) : \theta \in \Theta\} = \max\{L^*(x, \theta^*) : \theta^* \in \Theta^* = f(\Theta)\},$$

и бидејќи левата страна се максимизира за U, следи дека десната страна се максимизира за $f(U)$. Значи, $f(U)$ е максимално подобен оценувач за $f(\theta)$, што требаше да се докаже. ∎

Под дополнителни услови за регуларност може да се покаже конзистентноста и асимптотската нормалност на ML оценувачите. На следните примери се испитани некои од асимптотските својства кај ML оценувачите.

Пример 4.19. Нека обележјето X има $Geo(\theta)$ распределба, каде $0 < \theta < 1$ е непознат параметар, односно распределба на веројатност дадена со

$$P(x, \theta) = \theta(1 - \theta)^x, \quad x = 0, 1, 2, ...$$

Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X. Тогаш, максимално подобниот оценувач за θ е

$$\widehat{\theta} = \frac{1}{\overline{X}_n + 1}.$$

Да ја испитаме конзистентноста на оценувачот $\widehat{\theta}$ за θ. Имаме за $0 < \varepsilon < \theta$,

$$P\{|\widehat{\theta} - \theta| < \varepsilon\} = P\{\theta - \varepsilon < \widehat{\theta} < \theta + \varepsilon\} = P\{\frac{1}{\theta + \varepsilon} - 1 < \overline{X}_n < \frac{1}{\theta - \varepsilon} - 1\} =$$

$$= P\{-\frac{\varepsilon}{(\theta + \varepsilon)\theta} < \overline{X}_n - \frac{1 - \theta}{\theta} < \frac{\varepsilon}{(\theta - \varepsilon)\theta}\} \geq P\{|\overline{X}_n - \frac{1 - \theta}{\theta}| < \frac{\varepsilon}{(\theta + \varepsilon)\theta}\} =$$

$$= P\{|\overline{X}_n - E(\overline{X}_n)| < \frac{\varepsilon}{(\theta + \varepsilon)\theta}\} \geq 1 - \frac{D(\overline{X}_n)}{(\frac{\varepsilon}{(\theta+\varepsilon)\theta})^2} = 1 - \frac{\frac{1-\theta}{n\theta^2}}{(\frac{\varepsilon}{(\theta+\varepsilon)\theta})^2} \to 1,$$

кога $n \to \infty$. Слично се покажува и во случај кога $\varepsilon \geq \theta$. Заклучуваме дека *ML* оценувачот $\widehat{\theta}$ е конзистентен оценувач за θ.

Потоа, од Централаната гранична теорема, имаме дека

$$\sqrt{n}(\overline{X}_n - \frac{1 - \theta}{\theta}) \xrightarrow{dist.} \mathcal{N}(0, \frac{1 - \theta}{\theta^2}),$$

Ирена Стојковска

од каде добиваме дека

$$\sqrt{n}(\widehat{\theta}-\theta)=\sqrt{n}(g(\overline{X}_n)-g(\frac{1-\theta}{\theta}))\xrightarrow{dist.}\mathcal{N}(0,g'(\frac{1-\theta}{\theta})\frac{1-\theta}{\theta^2})=\mathcal{N}(0,\theta^2(1-\theta)),$$

при што е користен Делта методот за монотоната функција $g(x)=\frac{1}{1+x}$. Значи, *ML* оценувачот $\widehat{\theta}$ за θ е асимптотски нормален оценувач.

Пример 4.20. Нека $(X_1,X_2,...,X_n)$ е примерок кој одговара на обележјето X кое има $\mathcal{U}[0,\theta]$ распределба, каде $\theta>0$ е непознат параметар. Максимално подобен оценувач за θ е

$$\widehat{\theta}=X_{(n)}=\max\{X_1,...,X_n\},$$

чија функција на распределба е

$$F_{\widehat{\theta}}(x)=P\{\widehat{\theta}<x\}=\Big(\frac{x}{\theta}\Big)^n,\ 0\le x\le\theta.$$

Тогаш, за произволен $\varepsilon>0$ имаме,

$$P\{|\widehat{\theta}-\theta|>\varepsilon\}=P\{\widehat{\theta}<\theta-\varepsilon\}=\Big(\frac{\theta-\varepsilon}{\theta}\Big)^n\to 0,\ \text{кога } n\to\infty.$$

Значи, $\widehat{\theta}$ е конзистентен оценувач за θ. Од друга страна имаме

$$P\{n(\theta-\widehat{\theta})<x\}=P\{\widehat{\theta}>\theta-\frac{x}{n}\}=1-(1-\frac{x}{\theta n})^n\to 1-e^{-x/\theta},\ \text{за } x>0.$$

Значи, $n(\theta-\widehat{\theta})$ конвергира по распределба кон експоненцијална случајна променлива.

4.4 Интервали на доверба

Нека $\mathcal{P}=\{F(x,\theta):\theta\in\Theta\}$ е фамилијата од допустливи распределби за обележјето X, каде θ е непознат параметар. Нека $(X_1,X_2,...,X_n)$ е примерок кој одговара на обележјето X. Нека $\widehat{\theta}=U(X_1,X_2,...,X_n)$ е оценувач за θ. Најголемиот проблем кај точкастите оценувачи е тоа што веројатноста $P\{\widehat{\theta}=\theta\}$ е многу мала (таа е нула во случај кога $\widehat{\theta}$ има апсолутно непрекината распределба), што доведува до потешкотии при анализа на грешката, односно отстапувањето, при користење на оценувачот $\widehat{\theta}$ за оценување на непознатиот параметар θ.

Алтернативен пристап за оценување на непознатите параметри е интервалното оценување преку интервали на доверба.

Дефиниција 4.14. Нека $L(X_1, X_2, ..., X_n)$ и $U(X_1, X_2, ..., X_n)$ се две статистики така што за секој $\theta \in \Theta$ важи $P\{L(X_1, X_2, ..., X_n) \leq U(X_1, X_2, ..., X_n)\} = 1$ и

$$P\{L(X_1, X_2, ..., X_n) < \theta < U(X_1, X_2, ..., X_n)\} \geq 1 - \alpha, \tag{4.29}$$

каде $0 < \alpha < 1$ е даден број. Тогаш, интервалот $(L(X_1, X_2, ..., X_n), U(X_1, X_2, ..., X_n))$ се нарекува $100(1-\alpha)\%$ **интервал на доверба** за непознатиот параметар θ. Бројот $1 - \alpha$ се нарекува **ниво на доверба** или **веројатност на доверба**.

Интервалите на доверба се дефинирани преку распределбата на случајниот примерок $(X_1, X_2, ..., X_n)$, но во пракса, тие се интерпретираат преку набљудуваните вредности на овие случајни променливи оставајќи впечаток дека веројатносното тврдење е во врска со θ, а не со случајниот интервал. Значи, за дадена реализација на примерокот $(x_1, x_2, ..., x_n)$, интервалот $(L(x_1, x_2, ..., x_n), U(x_1, x_2, ..., x_n))$ или ќе ја содржи вистинската вредност на θ или нема да ја содржи вистинската вредност на θ, но при повторување на експериментот голем број пати, $100(1-\alpha)\%$ од добиените интервали ќе ја содржат вистинската вредност на θ.

Некогаш ќе може да се најде интервал $(L(X_1, X_2, ..., X_n), U(X_1, X_2, ..., X_n))$ со особина

$$P\{L(X_1, X_2, ..., X_n) < \theta < U(X_1, X_2, ..., X_n)\} = 1 - \alpha, \tag{4.30}$$

за сите $\theta \in \Theta$, или со особина

$$P\{L(X_1, X_2, ..., X_n) < \theta < U(X_1, X_2, ..., X_n)\} \approx 1 - \alpha, \tag{4.31}$$

за сите $\theta \in \Theta$, кој се нарекува **приближно** $100(1-\alpha)\%$ **интервал на доверба** за θ.

Пример 4.21. Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето $X \sim \mathcal{N}(\mu, 1)$. Тогаш, $\sqrt{n}(\overline{X}_n - \mu) \sim \mathcal{N}(0, 1)$, па затоа

$$P\{-1.96 < \sqrt{n}(\overline{X}_n - \mu) < 1.96\} = 0.95,$$

што е еквивалентно со

$$P\{\overline{X}_n - \frac{1.96}{\sqrt{n}} < \mu < \overline{X}_n + \frac{1.96}{\sqrt{n}}\} = 0.95,$$

каде $\overline{X}_n$ е средината на примерокот. Следствено, интервалот со крајни точки $\overline{X}_n \pm \frac{1.96}{\sqrt{n}}$ е 95% интервал на доверба за μ.

Да забележиме дека, ако обележјето X има математичко очекување μ и дисперзија 1 (не е неопходно да е нормално распределено), тогаш од централната гранична теорема имаме,

$$P\{-1.96 < \sqrt{n}(\overline{X}_n - \mu) < 1.96\} \approx 0.95,$$

Ирена Стојковска

ако n е доволно големо. Според горниот начин на расудување, следи дека интервалот со крајни точки $\overline{X}_n \pm \frac{1.96}{\sqrt{n}}$ е приближно 95% интервал на доверба за μ.

Може да се покаже дека ако обележјето X има математичко очекување μ и непозната дисперзија σ^2, тогаш за n доволно големо (во пракса, за $n > 30$), интервалот

$$(\overline{X}_n - \frac{1.96\ S_n}{\sqrt{n}}, \overline{X}_n + \frac{1.96\ S_n}{\sqrt{n}})$$

е приближно 95% интервал на доверба за μ, каде S_n^2 е корегираната дисперзијата на примерокот (тогаш S_n е **корегирана стандардна девијација на примерокот**).

Изборот на статистиките $L(X_1, X_2, ..., X_n)$ и $U(X_1, X_2, ..., X_n)$ кои ќе бидат лева, односно десна граница на интервалот на доверба не е еднозначно одреден. Една постапка за одредување на овие граници е така нареченият **пивот метод**, со кој најнапред се одбира **централна статитика** $T(X_1, ..., X_n, \theta)$ за параметарот θ така што

(1) распределбата на $T(X_1, ..., X_n, \theta)$ не зависи од оценуваниот параметар θ,

(2) за секој $(x_1, ..., x_n) \in \mathbb{R}^n$ функцијата $T(x_1, ..., x_n, \theta)$ е непрекината и строго монотона функција по θ.

Се покажува дека кога равенката $T(x_1, ..., x_n, \theta) = t$ е решлива за секој $(x_1, ..., x_n) \in \mathbb{R}^n$ и t од множеството вредности на $T(x_1, ..., x_n, \theta)$, тогаш за секој $0 < \alpha < 1$ може да се конструира интервал на довреrба за θ со ниво на доверба $1 - \alpha$. Имено, постојат броеви a и b така што

$$1 - \alpha = P\{a < T(X_1, ..., X_n, \theta) < b\},$$

за сите $\theta \in \Theta$. Заради горните услови, последното равенство ќе може да се трансформира во облик

$$1 - \alpha = P\{L(X_1, ..., X_n) < \theta < U(X_1, ..., X_n)\},$$

па интервалот $(L(X_1, ..., X_n), U(X_1, ..., X_n))$ е $100(1-\alpha)\%$ интервал на доверба за параметарот θ.

Пример 4.22. Нека обележјето X има $\mathcal{U}[0, \theta]$ распределба, каде $\theta > 0$ е непознат параметар. Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X. Максимално подобен оценувач за θ е подредената статистика $X_{(n)} = \max\{X_1, ..., X_n\}$. Тогаш, функцијата на распределба на статистиката $T = X_{(n)}/\theta$ е

$$F_T(x) = x^n,\ 0 \leq x \leq 1.$$

Ирена Стојковска

Заклучуваме дека статистиката T е централна статистика за θ. За да најдеме $100(1-\alpha)\%$ интервал на доверба за θ, треба да најдеме броеви a и b така што

$$P\{a < \frac{X_{(n)}}{\theta} < b\} = 1 - \alpha.$$

Очигледно постојат бесконечно многу избори за a и b. Може да се покаже дека за $a = \alpha^{1/n}$ и $b = 1$ се добива најкраткиот можен интервал на доверба. Тој интервал на доверба е

$$(X_{(n)}, \frac{X_{(n)}}{\alpha^{1/n}}).$$

Пример 4.23. Нека обележјето $X \sim \mathcal{P}(\lambda)$, каде $\lambda > 0$ е непознат параметар. Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X. Максимално подобен оценувач за λ е $\overline{X}_n$. При тоа, од $EX = DX = \lambda$ имаме дека $E(\overline{X}_n) = \lambda$ и $D(\overline{X}_n) = \lambda/n$.

Во случај на голем примерок, од законот на големите броеви и централната гранична тепрема, имаме дека

$$\overline{X}_n \xrightarrow{P} \lambda \quad \text{и} \quad \frac{\sqrt{n}}{\sqrt{\lambda}}(\overline{X}_n - \lambda) \xrightarrow{dist.} \mathcal{N}(0, 1^2).$$

Од $\overline{X}_n \xrightarrow{P} \lambda$ следи дека $\sqrt{\overline{X}_n} \xrightarrow{P} \sqrt{\lambda}$. Имено за $0 < \varepsilon < 2\sqrt{\lambda}$,

$$P\{|\sqrt{\overline{X}_n} - \sqrt{\lambda}| < \varepsilon\} = P\{(-\varepsilon + \sqrt{\lambda})^2 < \overline{X}_n < (\varepsilon + \sqrt{\lambda})^2\} =$$

$$= P\{\varepsilon^2 - 2\varepsilon\sqrt{\lambda} < \overline{X}_n - \lambda < \varepsilon^2 + 2\varepsilon\sqrt{\lambda}\} \geq P\{|\overline{X}_n - \lambda| < -\varepsilon^2 + 2\varepsilon\sqrt{\lambda}\} \geq$$

$$\geq 1 - \frac{\lambda/n}{(-\varepsilon^2 + 2\varepsilon\sqrt{\lambda})^2} \longrightarrow 1, \text{ кога } n \to \infty.$$

Слично се покажува и кога $\varepsilon \geq 2\sqrt{\lambda}$. Понатаму, од $\frac{\sqrt{n}}{\sqrt{\lambda}}(\overline{X}_n - \lambda) \xrightarrow{dist.} \mathcal{N}(0, 1^2)$ и $\sqrt{\overline{X}_n} \xrightarrow{P} \sqrt{\lambda}$, користејќи ја теоремата на Slutsky заклучуваме дека

$$\frac{\sqrt{n}}{\sqrt{\overline{X}_n}}(\overline{X}_n - \lambda) \xrightarrow{dist.} \mathcal{N}(0, 1^2).$$

Тогаш, за централна статистика за параметарот λ може да ја земеме статистиката $T = \frac{\sqrt{n}}{\sqrt{\overline{X}_n}}(\overline{X}_n - \lambda)$. Сега, бараме броеви a и b така што

$$P\{a < \frac{\sqrt{n}}{\sqrt{\overline{X}_n}}(\overline{X}_n - \lambda) < b\} = 1 - \alpha. \tag{4.32}$$

Ирена Стојковска

Бидејќи статистиката T е асимптотски нормална следи дека

$$P\{a < \frac{\sqrt{n}}{\sqrt{\overline{X}_n}}(\overline{X}_n - \lambda) < b\} \approx \Phi_0(b) - \Phi_0(a), \tag{4.33}$$

каде $\Phi_0(x) = \int_0^x \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$ е Лапласовиот интеграл, па може да најдеме само приближен $100(1-\alpha)\%$ интервал на доверба за λ. Понатаму, имаме дека

$$P\{a < \frac{\sqrt{n}}{\sqrt{\overline{X}_n}}(\overline{X}_n - \lambda) < b\} = P\{\overline{X}_n - \frac{b\sqrt{\overline{X}_n}}{\sqrt{n}} < \lambda < \overline{X}_n - \frac{a\sqrt{\overline{X}_n}}{\sqrt{n}}\}, \tag{4.34}$$

значи приближниот $100(1-\alpha)\%$ интервал на доверба за λ е

$$(\overline{X}_n - \frac{b\sqrt{\overline{X}_n}}{\sqrt{n}}, \overline{X}_n - \frac{a\sqrt{\overline{X}_n}}{\sqrt{n}}). \tag{4.35}$$

Броевите a и b ги определуваме од условот да интервалот на доверба е со минимална должина, односно бараме броеви a и b така да разликата $b-a$ е минимална при услов $\Phi_0(b) - \Phi_0(a) = 1-\alpha$. Од симетријата на стандардната нормална распределба околу нулата заклучуваме дека за $a = -b$ се добива интервал на доверба со минимална должина. Тогаш,

$$1 - \alpha = \Phi_0(b) - \Phi_0(a) = \Phi_0(b) - \Phi_0(-b) = 2\Phi_0(b).$$

Значи, $\Phi_0(b) = (1-\alpha)/2$, и тогаш означуваме $b = u_{(1-\alpha)/2}$, $a = -u_{(1-\alpha)/2}$, каде $\Phi_0(u_\alpha) = \alpha$.

4.4.1 Интервал на доверба за веројатност на настан

Нека настанот A се реализира со непозната веројатност $p = P(A)$. За да најдеме интервал на доверба за p, го разгледуваме обележјето $X = I_A$ кое има Бернулиева $0,1$ распределба, односно има $\mathcal{B}(1,p)$ распределба. Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X. Тогаш, од централната гранична теорема имаме дека статистиката

$$T = \frac{\sum_{i=1}^n X_i - np}{\sqrt{npq}} \xrightarrow{dist.} \mathcal{N}(0, 1^2),$$

каде $q = 1-p$, има асимптотски стандардна нормална распределба $\mathcal{N}(0,1^2)$. Па, во случај на голем примерок, за централна статистика за p може да ја земеме статистиката T и нејзината асимптотска распределба да ја користиме при изведувањето на интервалот на доверба за p.

Ирена Стојковска

Сега, треба да најдеме броеви a и b така што

$$1-\alpha = P\{a < \frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} < b\} \approx \Phi_0(b) - \Phi_0(a),$$

каде $\Phi_0(x) = \int_0^x \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$ е Лапласовиот интеграл. Заради симетричноста на стандардната нормална распределба околу нилата земаме $a = -b$, и од $1-\alpha \approx \Phi_0(b) - \Phi_0(a) = 2\Phi_0(b)$, може приближно да земеме дека $b = u_{(1-\alpha)/2}$ и $a = -u_{(1-\alpha)/2}$, каде $\Phi_0(u_\alpha) = \alpha$.

Откако ги одредивме a и b, обликот на приближниот $100(1-\alpha)\%$ интервал на доверба за p го добиваме на следниот начин. Имено,

$$1-\alpha = P\{|\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}}| < u_{(1-\alpha)/2}\} = P\{\frac{(\sum_{i=1}^{n} X_i - np)^2}{np(1-p)} < u^2_{(1-\alpha)/2}\} =$$

$$= P\{(n^2 - nu^2_{(1-\alpha)/2})p^2 - (2n\sum_{i=1}^{n} X_i + nu^2_{(1-\alpha)/2})p + (\sum_{i=1}^{n} X_i)^2 < 0\} = P\{\widehat{p}_1 < p < \widehat{p}_2\},$$

каде $\widehat{p}_1$ и $\widehat{p}_2$ се соодвтно помалиот и поголемиот корен на квадратната рнвенка по p

$$(n^2 - nu^2_{(1-\alpha)/2})p^2 - (2n\sum_{i=1}^{n} X_i + nu^2_{(1-\alpha)/2})p + (\sum_{i=1}^{n} X_i)^2 = 0,$$

и бараниот интервал на доверба е $(\widehat{p}_1, \widehat{p}_2)$, односно

$$\Big(\frac{n}{n+u^2_{(1-\alpha)/2}}\Big(\frac{\sum_{i=1}^{n} X_i}{n} + \frac{u^2_{(1-\alpha)/2}}{2n} - u_{(1-\alpha)/2}\sqrt{\frac{\sum_{i=1}^{n} X_i(n-\sum_{i=1}^{n} X_i)}{n} - \frac{u^2_{(1-\alpha)/2}}{4n^2}}\Big),$$
$$\frac{n}{n+u^2_{(1-\alpha)/2}}\Big(\frac{\sum_{i=1}^{n} X_i}{n} + \frac{u^2_{(1-\alpha)/2}}{2n} + u_{(1-\alpha)/2}\sqrt{\frac{\sum_{i=1}^{n} X_i(n-\sum_{i=1}^{n} X_i)}{n} - \frac{u^2_{(1-\alpha)/2}}{4n^2}}\Big)\Big).$$

4.4.2 Интервали на доверба за параметрите на нормална распределба

Во случај на мал примерок, претпоставуваме дека обележјето X има нормална распределба $\mathcal{N}(m, \sigma^2)$. Тогаш, се користат точните распределби на централните статистики за наоѓање на интервалите на доверба за параметрите m и σ^2.

Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X.

(1) Интервал на доверба за m кога σ^2 е позната

Видовме дека статистиката $\overline{X}_n$ има $\mathcal{N}(m, \sigma^2/n)$ распределба, од каде заклучуваме дека за централна статистика може да ја земеме статистиката

$$T = \frac{\overline{X}_n - m}{\sigma}\sqrt{n} \sim \mathcal{N}(0, 1^2).$$

Ирена Стојковска

Бараме броеви a и b така што

$$1-\alpha = P\{a < \frac{\overline{X}_n - m}{\sigma}\sqrt{n} < b\} = \Phi_0(b) - \Phi_0(a),$$

каде $\Phi_0(x) = \int_0^x \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$ е Лапласовиот интеграл. Од равенството

$$P\{a < \frac{\overline{X}_n - m}{\sigma}\sqrt{n} < b\} = P\{\overline{X}_n - b\frac{\sigma}{\sqrt{n}} < m < \overline{X}_n - a\frac{\sigma}{\sqrt{n}}\},$$

заклучуваме дека $100(1-\alpha)\%$ интервал на доверба за m е

$$(\overline{X}_n - b\frac{\sigma}{\sqrt{n}}, \overline{X}_n - a\frac{\sigma}{\sqrt{n}}).$$

Броевите a и b ги определуваме од условот да интервалот на доверба е со минимална должина, односно бараме броеви a и b така да разликата $b-a$ е минимална при услов $\Phi_0(b) - \Phi_0(a) = 1-\alpha$. Од симетријата на стандардната нормална распределба околу нулата заклучуваме дека за $a = -b$ се добива интервал на доверба со минимална должина. Тогаш,

$$1-\alpha = \Phi_0(b) - \Phi_0(a) = \Phi_0(b) - \Phi_0(-b) = 2\Phi_0(b).$$

Значи, $\Phi_0(b) = (1-\alpha)/2$, и тогаш означуваме $b = u_{(1-\alpha)/2}$, $a = -u_{(1-\alpha)/2}$, каде $\Phi_0(u_\alpha) = \alpha$.

(2) Интервал на доверба за m кога σ^2 не е позната

Во овој случај за централна статистика ја земаме статистиката

$$T = \frac{\overline{X}_n - m}{\overline{S}_n}\sqrt{n-1} \sim t_{n-1}.$$

Бараме броеви a и b така што

$$1-\alpha = P\{a < \frac{\overline{X}_n - m}{\overline{S}_n}\sqrt{n-1} < b\}. \tag{4.36}$$

Од равенството

$$P\{a < \frac{\overline{X}_n - m}{\overline{S}_n}\sqrt{n-1} < b\} = P\{\overline{X}_n - b\frac{\overline{S}_n}{\sqrt{n-1}} < m < \overline{X}_n - a\frac{\overline{S}_n}{\sqrt{n-1}}\},$$

заклучуваме дека $100(1-\alpha)\%$ интервал на доверба за m е

$$(\overline{X}_n - b\frac{\overline{S}_n}{\sqrt{n-1}}, \overline{X}_n - a\frac{\overline{S}_n}{\sqrt{n-1}}).$$

Броевите a и b ги определуваме од условот да интервалот на доверба е со минимална должина, односно бараме броеви a и b така да разликата $b-a$ е минимална при услов (4.36). Од симетријата на студентова распределба околу нулата заклучуваме дека за $a=-b$ се добива интервал на доверба со минимална должина. Тогаш, равенството (4.36) преминува во

$$1-\alpha = P\{|\frac{\overline{X}_n - m}{\overline{S}_n}\sqrt{n-1}| < b\} = 1 - P\{|\frac{\overline{X}_n - m}{\overline{S}_n}\sqrt{n-1}| \geq b\}.$$

Означуваме со $t_{n,\alpha}$ број за кој важи $P\{|t_n| > t_{n,\alpha}\} = \alpha$, каде t_n е случајна променлива со студентова распределба со n степени на слобода. Според тоа, $b = t_{n-1,\alpha}$ и $a = -t_{n-1,\alpha}$.

(3) Интервал на доверба за σ^2 кога m е познато

Поаѓајќи од фактот дека $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - m)^2$ е непристрасен оценувач за σ^2, може да не "инспирира" да за централна статистика во овој случај ја земеме статистиката

$$T = \frac{\sum_{i=1}^n (X_i - m)^2}{\sigma^2} = \sum_{i=1}^n (\frac{X_i - m}{\sigma})^2 \sim \chi_n^2.$$

Бараме броеви a и b $(a, b > 0)$ така што

$$1-\alpha = P\{a < \frac{\sum_{i=1}^n (X_i - m)^2}{\sigma^2} < b\}.$$

Од равенството

$$P\{a < \frac{\sum_{i=1}^n (X_i - m)^2}{\sigma^2} < b\} = P\{\frac{\sum_{i=1}^n (X_i - m)^2}{b} < \sigma^2 < \frac{\sum_{i=1}^n (X_i - m)^2}{a}\},$$

заклучуваме дека $100(1-\alpha)\%$ интервал на доверба за σ^2 е

$$(\frac{\sum_{i=1}^n (X_i - m)^2}{b}, \frac{\sum_{i=1}^n (X_i - m)^2}{a}).$$

Броевите a и b ги определуваме од условите

$$P\{\frac{\sum_{i=1}^n (X_i - m)^2}{\sigma^2} < a\} = \alpha/2 \text{ и } P\{\frac{\sum_{i=1}^n (X_i - m)^2}{\sigma^2} > b\} = \alpha/2.$$

Ако означиме со $\chi^2_{n,\alpha}$ број за кој важи $P\{\chi_n^2 > \chi^2_{n,\alpha}\} = \alpha$, каде χ_n^2 е случајна променлива со χ^2 распределба со n степени на слобода, тогаш имаме дека $a = \chi^2_{n,1-\alpha/2}$ и $b = \chi^2_{n,\alpha/2}$.

Ирена Стојковска

(4) Интервал на доверба за σ^2 кога m не е познато

За централна статистика за σ^2 ја земаме статистиката

$$T = \frac{n\overline{S}_n^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Тогаш, бараме броеви a и b $(a, b > 0)$ така што

$$1 - \alpha = P\{a < \frac{n\overline{S}_n^2}{\sigma^2} < b\}.$$

Од равенството

$$P\{a < \frac{n\overline{S}_n^2}{\sigma^2} < b\} = P\{\frac{n\overline{S}_n^2}{b} < \sigma^2 < \frac{n\overline{S}_n^2}{a}\},$$

заклучуваме дека $100(1-\alpha)\%$ интервал на доверба за σ^2 е

$$(\frac{n\overline{S}_n^2}{b}, \frac{n\overline{S}_n^2}{a}).$$

Броевите a и b ги определуваме од условите

$$P\{\frac{n\overline{S}_n^2}{\sigma^2} < a\} = \alpha/2 \text{ и } P\{\frac{n\overline{S}_n^2}{\sigma^2} > b\} = \alpha/2.$$

Тогаш, имаме дека $a = \chi^2_{n-1,1-\alpha/2}$ и $b = \chi^2_{n-1,\alpha/2}$.

5

Тестирање на хипотези

5.1 Основни поими

Често при забележување на вредностите на обележјето X при некое истражување, обележето X не е потполно определено. Во тој случај се јавува потреба од поставување на одредени претпоставки во врска со некои својства на случајната променлива X. Овие претпоставки се нарекуваат **статистички хипотези**, а постапката за донесување на одлука за прифаќање или отфрлање на статистичката хипотеза се нарекува **тестирање на статистичка хипотеза**.

Нека $\mathcal{P} = \{F(x,\theta) : \theta \in \Theta\}$ е фамилијата од допустливи распределби на обележјето X. Хипотезите кои ги фиксираат вредностите на непознатите параметри θ или го намалуваат просторот на параметри Θ се нарекуваат **параметарски хипотези**. Хипотезите кои не зависат од непознатите параметри се нарекуваат **непараметарски хипотези**.

Пример 5.1. Примери за параметарски хипотези:

(а) Обележјето X има математичко очекување $m > m_0$.

(б) Обележјето X кое е распределено според Поасонов закон на распределба има математичко очекување енакво на даден број λ_0.

(в) Нормално распределените обележја X и Y со дисперзии $\sigma_X^2 = \sigma_Y^2 = \sigma_0^2$ имаат еднакви математички очекувања т.е. $m_X = m_Y$.

Примери за непараметарски хипотези:

(г) Обележјето X има $\mathcal{P}(2)$ распределба.

(д) Обележјето X има нормална распределба.

(ѓ) Обележјата X и Y се независни.

Ако статистичката хипотеза еднозначно ја определува распределбата на обележјето X, тогаш таа се нарекува **проста хипотеза**. Во спотивно, ако само ја намалува фамилијата од допустливи распределби, се нарекува **сложена хипотеза**. Така, хипотезите (б) и (г) од Пример 5.1 се прости, останатите се сложени.

Нека H_0 е хипотеза за обележјето X која сакаме да ја тестираме. Една од задачите на математичката статистика е да врз основа на реализацијата $(x_1, x_2, ..., x_n)$ на примерокот $(X_1, X_2, ..., X_n)$ донесе одлука за прифаќање или отфрлање на хипотезата H_0. Правилото со кое врз основа на реализацијата $(x_1, x_2, ..., x_n)$ на примерокот $(X_1, X_2, ..., X_n)$ се одлучува дали ја прифаќаме или отфрламе хипотезата H_0 се нарекува **статистички тест**. Со статистичкиот тест се проверува дали реализацијата $(x_1, x_2, ..., x_n) \in C$, при што множеството $C \subseteq \mathbb{R}^n$ се нарекува **критична област**, и ако $(x_1, x_2, ..., x_n) \in C$ тогаш **хипотезата H_0 се отфрла**, а ако $(x_1, x_2, ..., x_n) \notin C$, тогаш **хипотезата H_0 се прифаќа**.

Една карактеристика на статистичкиот тест е **ниво на значајност на тестот** $\alpha \in (0, 1)$, кое се дефинира како горна граница на веројатноста да примерокот $(X_1, X_2, ..., X_n)$ прима вредности од критичната област C, при услов хипотезата H_0 да е точна, односно

$$P\{(X_1, X_2, ..., X_n) \in C | H_0\} \leq \alpha. \tag{5.1}$$

Најмалата вредност на α за кое е исполнет условот (5.1) се нарекува **големина на критичната област** C или **големина на тестот**.

При тестирање на хипотези бројот α однапред го задаваме со тоа што за α земаме "мали" вредности, на пример $\alpha = 0.05$ или $\alpha = 0.01$. Имено, во случај на мали вредности за α природно е да се отфрли хипотезата H_0 како неточна, ако $(x_1, x_2, ..., x_n) \in C$, затоа што тогаш настанот $\{(X_1, X_2, ..., X_n) \in C | H_0\}$ ќе има мала веројатност.

Прифаќањето на хипотезата H_0 означува дека меѓу реализацијата на примерокот и теориската распределба на обележјето на дадено ниво на значајност не постои значително отстапување. Но, исто така важи и дека при прифаќање на хипотезата H_0 не мора да значи дека таа е точна, додека пак при нејзино отфрлање можно е таа да е точна.

Критичната област C обично е определна со помош на некоја **тест статистика** $T_n = T_n(X_1, X_2, ..., X_n)$. За тест статистиката претпоставуваме дека е позната нејзината точна распределба (или приближна распределба при големи вредности на n). Ако H_0 е сложена хипотеза претпоставуваме дека распределбата на тест статистиката е иста за сите распределби на обележјето обфатени со хипотезата H_0. Имено, тест статистиката го карактеризира отстапувањето на реализацијата на примерокот од природно очекуваната вредност, под претпоставка H_0 да е точна.

Ирена Стојковска

Хипотезата H_0 која се тестира обично се нарекува **нулта хипотеза**. Додека, претпоставката за обележјето X со која се претпоставува нешто кое не е обфатено со нултата хипотеза, а сепак е допустливо, се нарекува **алтернативна хипотеза** и се означува со H_1. Во случај на отфрлање на нултата хипотеза H_0, **алтернативната хипотеза** H_1 **се прифаќа**.

При тестирање на една иста статистичка хипотеза H_0 може да се користат различни тестови, односно различни критични области. Тестовите меѓусебно се споредуваат според направените грешки при отфрлање на H_0 кога таа е точна (**грешка од прв вид**) и прифаќање на H_0 кога е точна H_1 (**грешка од втор вид**). Ако веројатноста за грешка од прв вид ја означиме со α и веројатноста за грешка од втор вид ја означиме со β имаме дека

$$\alpha = P\{(X_1, X_2, ..., X_n) \in C | H_0\} = P_0\{(X_1, X_2, ..., X_n) \in C\}, \tag{5.2}$$

$$\beta = P\{(X_1, X_2, ..., X_n) \notin C | H_1\} = P_1\{(X_1, X_2, ..., X_n) \notin C\}. \tag{5.3}$$

Пожелно е грешките од прв и втор вид да бидат што е можно помали. Меѓутоа обично со смалување на грешката од прв вид се зголемува грешката од втор вид. Затоа, постапка за избирање на критичната област е следната: Најнапред се фиксира веројатноста за грешка од прв вид α, а потоа меѓу сите критични области со големина α се избира онаа за која веројатноста за грешка од втор вид β е минимална.

Веројатноста за правилно отфрлање на H_0 кога H_1 е точна се нарекува **моќ на тестот** и се означува со p т.е.

$$p = P\{(X_1, X_2, ..., X_n) \in C | H_1\} = 1 - \beta. \tag{5.4}$$

Па, заради претходната дискусија за начинот на избирање на критичната област, може да кажеме дека со таа постапка сме добиле **најмоќен тест** за однапред дадена веројатност за грешка од прв вид α.

5.2 Тестирање на параметарски хипотези

Нека $\mathcal{P} = \{F(x, \theta) : \theta \in \Theta\}$ е фамилијата од допустливи распределби на обележјето X, каде θ е непознат параметар и Θ е просторот од параметри. Една параметарска хипотеза има облик $H_0 : \theta \in \Theta_0$, каде $\Theta_0 \subseteq \Theta$. Ако Θ_0 е едноелемнтно множество, тогаш H_0 е проста хипотеза. При тестирање на нултата хипотеза H_0, за алтернативна хипотеза се зема хипотезата $H_1 : \theta \in \Theta_1$, каде $\Theta_1 \subseteq \Theta \setminus \Theta_0$.

5.2.1 Нејман-Пирсонов тест

Нека обележјето X има функција на распределба $F(x, \theta)$, каде θ е непознат параметар. Сакаме да ја тестираме нултата проста хипотеза $H_0 : \theta = \theta_0$,

наспроти алтернативната проста хипотеза $H_1 : \theta = \theta_1$. При тоа, целта ни е да најдеме најмоќен тест за однапред дадена веројатност за грешка од прв вид α. Одговор на ова прашање ни дава Лемата на Нејман-Пирсон.

Теорема 5.1 (Лема на Нејман-Пирсон). *За секои $n \in \mathbb{N}$ и $\alpha \in (0,1)$, постои реален број $c \in \mathbb{R}$, така што при тестирање на нултата хипотеза $H_0 : \theta = \theta_0$, против алтернативната хипотеза $H_1 : \theta = \theta_1$, може да се најде најмоќен тест со оптимален критичен домен C_0 од облик*

$$C_0 = \{x = (x_1, x_2, ..., x_n) : \frac{L(x;\theta_1)}{L(x;\theta_0)} \geq c\},$$

каде $L(x;\theta)$ е функцијата на подобност на примерокот $(X_1, ..., X_n)$.

Доказ. Доказот ќе го изведеме за апсолутно непрекинато обележје X со густина на распределба $p(x;\theta)$. Нека $(X_1, ..., X_n)$ е примерок со големина n кој одговара на обележјето X. Ќе покажеме дека за $\alpha \in (0,1)$, постои $c \in \mathbb{R}$ така што

$$P_0\{\frac{L(X_1, ..., X_n;\theta_1)}{L(X_1, ..., X_n;\theta_0)} \geq c\} = \alpha, \tag{5.5}$$

што ќе значи дека за дадено ниво на значајност на тестот α постои критичен домен со големина α од обликот

$$\{x : \frac{L(x;\theta_1)}{L(x;\theta_0)} \geq c\}. \tag{5.6}$$

Потоа ќе ја покажеме неговата оптималност.

Дефинираме $h(c) = P_0\{\frac{L(X_1,...,X_n;\theta_1)}{L(X_1,...,X_n;\theta_0)} \geq c\}$. Функцијата $h(c)$ е опаѓачка и важи $h(0) = 1$. За секој $c \in \mathbb{R}$ означуваме $A_c = \{x : \frac{L(x;\theta_1)}{L(x;\theta_0)} \geq c\} \subset \mathbb{R}^n$. Тогаш,

$$\begin{aligned} 1 &\geq P_1\{\frac{L(X_1, ..., X_n;\theta_1)}{L(X_1, ..., X_n;\theta_0)} = \int_{A_c} L(x;\theta_1)\, dx \geq \\ &\geq c \int_{A_c} L(x;\theta_0)\, dx = c\, P_0\{\frac{L(X_1, ..., X_n;\theta_1)}{L(X_1, ..., X_n;\theta_0)} = c\, h(c), \end{aligned}$$

од каде следи дека $h(c) \leq \frac{1}{c}$ за секој $c \neq 0$, и $\lim_{c\to\infty} h(c) = 0$. Па, ако $h(c)$ е непрекината функција, тогаш за произволен $\alpha \in (0,1)$ постои $c \in \mathbb{R}$ таков што $h(c) = \alpha$.

Оптималноста на критичниот домен C_0 со големина α од обликот (5.6) ја покажуваме на следниот начин. Нека C е произволен критичен домен со големина α. Ако $P_1\{(X_1, ..., X_n) \in C \Delta C_0\} = 0$,[1] тогаш важи дека

$$P_1\{(X_1, ..., X_n) \in C\} = P_1\{(X_1, ..., X_n) \in C_0\},$$

[1] $C \Delta C_0 = (C \cup C_0) \setminus (C \cap C_0) = (C \cap \overline{C_0}) \cup (\overline{C} \cap C_0)$

односно

$$P_1\{(X_1,...,X_n) \notin C\} = P_1\{(X_1,...,X_n) \notin C_0\}. \tag{5.7}$$

Нека $P_1\{(X_1,...,X_n) \in C\Delta C_0\} \neq 0$. Тогаш, имаме

$$P_1\{(X_1,...,X_n) \in C\} = \int_C L(x;\theta_1)\,dx = \int_{C\cap C_0} L(x;\theta_1)\,dx + \int_{C\cap\overline{C_0}} L(x;\theta_1)\,dx.$$

Слично,

$$P_1\{(X_1,...,X_n) \in C_0\} = \int_{C_0} L(x;\theta_1)\,dx = \int_{C_0\cap C} L(x;\theta_1)\,dx + \int_{C_0\cap\overline{C}} L(x;\theta_1)\,dx.$$

Со одземање на последните две равенства добиваме

$$\begin{aligned}
& P_1\{(X_1,...,X_n) \in C\} - P_1\{(X_1,...,X_n) \in C_0\} = \\
= \;& \int_{C\cap\overline{C_0}} L(x;\theta_1)\,dx - \int_{C_0\cap\overline{C}} L(x;\theta_1)\,dx \leq \\
\leq \;& c\int_{C\cap\overline{C_0}} L(x;\theta_0)\,dx - c\int_{C_0\cap\overline{C}} L(x;\theta_0)\,dx = \\
= \;& c\Big(\int_{C\cap\overline{C_0}} L(x;\theta_0)\,dx + \int_{C\cap C_0} L(x;\theta_0)\,dx - \\
& - \int_{C\cap C_0} L(x;\theta_0)\,dx - \int_{C_0\cap\overline{C}} L(x;\theta_0)\,dx\Big) = \\
= \;& c\Big(\int_C L(x;\theta_0)\,dx - \int_{C_0} L(x;\theta_0)\,dx\Big) = \\
= \;& c\,(\alpha-\alpha) = 0,
\end{aligned}$$

од каде следува дека

$$P_1\{(X_1,...,X_n) \in C\} \leq P_1\{(X_1,...,X_n) \in C_0\},$$

односно

$$P_1\{(X_1,...,X_n) \notin C\} \geq P_1\{(X_1,...,X_n) \notin C_0\}. \tag{5.8}$$

Од (5.7) и (5.8) заклучуваме дека веројатноста за грешка од втор вид β е најмала ако за критичен домен со големина α се земе C_0 од облик (5.6). ∎

Следните примери ја покажуваат примената на Лемата на Нејман-Пирсон при наоѓање на најмоќни тестови.

Пример 5.2. Нека обележјето X има $\mathcal{N}(m,1)$ распределба, каде m е непознат параметар. За дадена големина на примерокот n и ниво на значајност (веројатност за грешка од прв вид) α, со примена на Лемата на Нејман-Пирсон ќе

конструираме најмоќен тест за проверка на хипотезата $H_0 : m = m_0$, против $H_1 : m = m_1$, кога $m_0 < m_1$.

Густината на распределба на обележјето X е

$$p_X(x, m) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2}},$$

за $x \in \mathbb{R}$, од каде добиваме дека функцијата на подобност е

$$L(x, m) = (2\pi)^{-n/2} \exp\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - m)^2\},$$

за $x = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$. Да го пресметаме количникот на функциите на подобност за $m = m_1$ и $m = m_0$, односно

$$\frac{L(x; m_1)}{L(x; m_0)} = \frac{\exp\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - m_1)^2\}}{\exp\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - m_0)^2\}} =$$

$$= \exp\{-\frac{1}{2}\sum_{i=1}^{n}((x_i - m_1)^2 - (x_i - m_0)^2)\} = \exp\{(m_1 - m_0)\sum_{i=1}^{n} x_i - \frac{n}{2}(m_1^2 - m_0^2)\}.$$

Според Лемата на Нејман-Пирсон за даденото $\alpha \in (0, 1)$ постои $c \in \mathbb{R}$ така што оптималниот критичен домен има облик

$$C_0 = \{x = (x_1, x_2, ..., x_n) : \frac{L(x; m_1)}{L(x; m_0)} \geq c\}.$$

Неравенството $\frac{L(x;m_1)}{L(x;m_0)} \geq c$ е еквивалентно со

$$(m_1 - m_0)\sum_{i=1}^{n} x_i - \frac{n}{2}(m_1^2 - m_0^2) \geq \ln c,$$

па имајќи во предвид дека $\sum_{i=1}^{n} x_i = n\overline{x}$ и $m_1 - m_0 > 0$ (од условот дека $m_0 < m_1$), добиваме

$$\overline{x} \geq \frac{\ln c}{n(m_1 - m_0)} + \frac{1}{2}(m_1 + m_0) = c_0,$$

односно оптималниот критичен домен има облик

$$C_0 = \{x = (x_1, x_2, ..., x_n) : \overline{x} \geq c_0\},$$

каде $c_0 \in \mathbb{R}$ се наоѓа од условот $P\{(X_1, ..., X_n) \in C_0 | H_0\} = \alpha$. Бидејќи под претпоставка H_0 да е точна $\overline{X}_n \sim \mathcal{N}(m_0, \frac{1}{n})$, од каде $\sqrt{n}(\overline{X}_n - m_0) \sim \mathcal{N}(0, 1)$, и затоа имаме дека

$$\alpha = P\{(X_1, ..., X_n) \in C_0 | H_0\} = P_0\{\overline{X}_n \geq c_0\} =$$

$$= P_0\{\sqrt{n}(\overline{X}_n - m_0) \geq \sqrt{n}(c_0 - m_0)\} = \frac{1}{2} - \Phi_0(\sqrt{n}(c_0 - m_0)),$$

каде $\Phi_0(x) = \int_0^x \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$ е интегралот на Лаплас чии вредности се читаат од таблица. Значи,

$$c_0 = m_0 + \frac{1}{\sqrt{n}} \Phi_0^{-1}(\frac{1-2\alpha}{2}).$$

Веројатноста за грешка од втор вид е

$$\beta = P\{(X_1, ..., X_n) \notin C_0 | H_1\} = P_1\{\overline{X}_n < c_0\} =$$

$$= P_1\{\sqrt{n}(\overline{X}_n - m_1) < \sqrt{n}(c_0 - m_1)\} = \frac{1}{2} + \Phi_0(\sqrt{n}(c_0 - m_1)) =$$

$$= \frac{1}{2} + \Phi_0(\sqrt{n}(m_0 - m_1) + \Phi_0^{-1}(\frac{1-2\alpha}{2})).$$

Да забележиме дека, ако $m_0 > m_1$, тогаш оптималниот критичен домен ќе има облик

$$C_0 = \{x = (x_1, x_2, ..., x_n) : \overline{x} \leq c_0\},$$

каде

$$c_0 = m_0 + \frac{1}{\sqrt{n}} \Phi_0^{-1}(\frac{2\alpha - 1}{2}),$$

и веројатноста за грешка од втор вид ќе биде

$$\beta = \frac{1}{2} - \Phi_0(\sqrt{n}(m_0 - m_1) + \Phi_0^{-1}(\frac{2\alpha - 1}{2})).$$

На пример, при тестирање на $H_0 : m = 0$ против $H_1 : m = 1$ со ниво на значајност $\alpha = 0,05$, врз основа на дадена низа статистички податоци $x_1, ..., x_n$ со големина $n = 100$, критичната вредност c_0 е

$$\begin{aligned} c_0 &= m_0 + \frac{1}{\sqrt{n}} \Phi_0^{-1}(\frac{1-2\alpha}{2}) = 0 + \frac{1}{\sqrt{100}} \Phi_0^{-1}(\frac{1 - 2 \cdot 0,05}{2}) = \\ &= \frac{1}{10} \Phi_0^{-1}(0,45) = \frac{1}{10} \cdot 1,645 = 0,1645, \end{aligned}$$

значи оптималниот критичен домен е

$$C_0 = \{x : \overline{x} \geq 0,1645\}.$$

Доколку за дадените податоци важи $\overline{x} \geq 0,1645$, тогаш H_0 се отфрла, во спротивно, таа се прифаќа. Веројатноста за грешка од втор вид е

$$\begin{aligned} \beta &= \frac{1}{2} + \Phi_0(\sqrt{n}(m_0 - m_1) + \Phi_0^{-1}(\frac{1-2\alpha}{2})) = \frac{1}{2} + \Phi_0(\sqrt{100}(0-1) + \Phi_0^{-1}(\frac{1 - 2 \cdot 0,05}{2})) = \\ &= \frac{1}{2} + \Phi_0(-10 + \Phi_0^{-1}(0,45)) = \frac{1}{2} + \Phi_0(-10 + 1,645) = \frac{1}{2} + \Phi_0(-8,355) \approx \frac{1}{2} - \frac{1}{2} = 0. \end{aligned}$$

Ирена Стојковска

Пример 5.3. Нека обележјето X има $\mathcal{E}(\lambda)$ распределба, каде $\lambda > 0$ е непознат параметар. Нека се дадени големината на примерокот n и нивото на значајност α. Според Лемата на Нејман-Пирсон постои најмоќен тест за проверка на хипотезата $H_0 : \lambda = \lambda_0$, против $H_1 : \lambda = \lambda_1$, кога $\lambda_0 < \lambda_1$, кој се наоѓа на следниот начин. Бидејќи густината на распределба на обележјето X е

$$p(x, \lambda) = \frac{1}{\lambda} e^{-x/\lambda},$$

за $x > 0$, добиваме дека функцијата на подобност на примерокот $(X_1, ..., X_n)$ е

$$L(x, \lambda) = \frac{1}{\lambda^n} e^{-\frac{1}{\lambda}\sum_{i=1}^n x_i},$$

за $x_i > 0$, $i = 1, ..., n$. Тогаш, количникот на функциите на подобност за $\lambda = \lambda_1$ и $\lambda = \lambda_0$ е

$$\frac{L(x; \lambda_1)}{L(x; \lambda_0)} = (\lambda_0/\lambda_1)^n \exp\{(\frac{1}{\lambda_0} - \frac{1}{\lambda_1}) \sum_{i=1}^n x_i\}.$$

Според Лемата на Нејман-Пирсон за даденото $\alpha \in (0, 1)$ постои $c \in \mathbb{R}$ така што оптималниот критичен домен има облик

$$C_0 = \{x = (x_1, x_2, ..., x_n) : \frac{L(x; \lambda_1)}{L(x; \lambda_0)} \geq c\}.$$

Неравенството $\frac{L(x;\lambda_1)}{L(x;\lambda_0)} \geq c$ е еквивалентно со

$$(\frac{1}{\lambda_0} - \frac{1}{\lambda_1}) \sum_{i=1}^n x_i \geq \ln(c(\lambda_1/\lambda_0)^n),$$

и беидејќи $\frac{1}{\lambda_0} - \frac{1}{\lambda_1} > 0$, од условот $\lambda_0 < \lambda_1$, имаме дека последното неравенство е еквивалентно на

$$\sum_{i=1}^n x_i \geq \frac{\ln(c(\lambda_1/\lambda_0)^n)}{\frac{1}{\lambda_0} - \frac{1}{\lambda_1}} = c_0,$$

односно оптималниот критичен домен има облик

$$C_0 = \{x = (x_1, x_2, ..., x_n) : \sum_{i=1}^n x_i \geq c_0\}.$$

Бројот $c_0 \in \mathbb{R}$ се наоѓа од условот $P\{(X_1, ..., X_n) \in C_0 | H_0\} = \alpha$. Кога H_0 е точна, имаме дека $\sum_{i=1}^n X_i \sim \Gamma(n, \lambda_0)$ и $\frac{2}{\lambda_0}\sum_{i=1}^n X_i \sim \chi^2_{2n}$ (покажи!). Па, c_0 го наоѓаме од

$$\alpha = P\{(X_1, ..., X_n) \in C_0 | H_0\} = P_0\{\sum_{i=1}^n X_i \geq c_0\} = P_0\{\frac{2}{\lambda_0}\sum_{i=1}^n X_i \geq \frac{2c_0}{\lambda_0}\},$$

Ирена Стојковска

од каде заклучуваме дека $\frac{2c_0}{\lambda_0} = \chi^2_{2n,\alpha}$, односно $c_0 = \frac{\lambda_0}{2}\chi^2_{2n,\alpha}$. (Да се потсетиме дека со $\chi^2_{n,\alpha}$ го означувавме бројот за кој важи $P\{\chi^2_n > \chi^2_{n,\alpha}\} = \alpha$, каде χ^2_n е случајна променлива со χ^2 распределба со n степени на слобода и кој се чита од таблица)

Веројатноста за грешка од втор вид е

$$\beta = P\{(X_1, ..., X_n) \notin C_0 | H_1\} = P_1\{\sum_{i=1}^{n} X_i < c_0\} =$$

$$= P_1\{\frac{2}{\lambda_1}\sum_{i=1}^{n} X_i < \frac{2c_0}{\lambda_1}\} = 1 - P_1\{\frac{2}{\lambda_1}\sum_{i=1}^{n} X_i \geq \frac{2c_0}{\lambda_1}\} = 1 - P_1\{\frac{2}{\lambda_1}\sum_{i=1}^{n} X_i \geq \frac{\lambda_0}{\lambda_1}\chi^2_{2n,\alpha}\},$$

односно од β е такво што $\chi^2_{2n,1-\beta} = \frac{\lambda_0}{\lambda_1}\chi^2_{2n,\alpha}$.

5.2.2 Рамномерно најмоќни тестови

Нека $\mathcal{P} = \{F(x,\theta) : \theta \in \Theta\}$ е фамилијата од допустливи распределби на обележјето X, каде θ е непознат параметар и $\Theta \subseteq \mathbb{R}$ е просторот од параметри. Претпоставуваме дека за произволни $\theta_0, \theta_1 \in \Theta$ задоволени се условите од Лемата Нејман-Пирсон, т.е. може да се определи оптимален критичен домен $C_0 = C_0(\alpha, \theta_0, \theta_1) \subseteq \mathbb{R}$ со големина α за тестирање на хипотезата $H_0 : \theta = \theta_0$ против $H_1 : \theta = \theta_1$. Ако C_0 не зависи од вредноста на параметарот $\theta_1 \in \Theta \setminus \{\theta_0\}$, тогаш множеството C_0 се нарекува **рамномерно оптимален критичен домен** за тестирање на хипотезата $H_0 : \theta = \theta_0$ против $H_1 : \theta \in \Theta \setminus \{\theta_0\}$. Соодветниот статистички тест е нарекува **рамномерно најмоќен тест**. Многу често рамномерниот најмоќен тест постои само при така наречени **едностранн алтернативни хипотези** од облик $H_1^+ : \theta > \theta_0$ или $H_1^- : \theta < \theta_0$.

Пример 5.4. Нека обележјето X има $\mathcal{N}(m, 1)$ распределба, каде m е непознат параметар. Во Пример 5.2 видовме дека оптималниот критичен домен за тестирање на $H_0 : m = m_0$ против $H_1 : m = m_1$ зависеше од врската меѓу m_0 и m_1 (се добиваа различни оптимални критични домени, кога $m_0 < m_1$ и кога $m_0 > m_1$). Тоа значи дека во овој случај не постои рамномерно оптимален критичен домен при тестирање на $H_0 : m = m_0$ против $H_1 : m \neq m_0$.

Но затоа, рамномерно оптималниот критичен домен при тестирање на $H_0 : m = m_0$ против $H_1^+ : m > m_0$ е $C_0 = \{x = (x_1, x_2, ..., x_n) : \overline{x} \geq c_0\}$, каде $c_0 = m_0 + \frac{1}{\sqrt{n}}\Phi_0^{-1}(\frac{1-2\alpha}{2})$. Слично, оптималниот критичен домен при тестирање на $H_0 : m = m_0$ против $H_1^- : m < m_0$ е $C_0 = \{x = (x_1, x_2, ..., x_n) : \overline{x} \leq c_0\}$, каде $c_0 = m_0 + \frac{1}{\sqrt{n}}\Phi_0^{-1}(\frac{2\alpha-1}{2})$.

5.2.3 Тестови со коефициент на подобност

Нека $\mathcal{P} = \{F(x,\theta) : \theta \in \Theta\}$ е фамилијата од допустливи распределби на обележјето X, каде θ е непознат параметар и Θ е просторот од параметри. При тестирање на простата хипотеза $H_0 : \theta = \theta_0$ против простата хипотеза $H_1 : \theta = \theta_1$ одговор на прашањето за најмоќен тест ни дава Лемата на Нејман-Пирсон. Во случај кога барем една од хипотезите, нултата или алтернативната, е сложена за наоѓање на критична област со големина α се користи **методот со коефициент на подобност** кој претставува еден вид на обопштување на тестот на Нејман-Пирсон.

При тестирање на нултата хипотеза $H_0 : \theta \in \Theta_0$ против алтернативната хипотеза $H_1 : \theta \in \Theta_1$ со методот со коефициент на подобност се користи тест статистиката

$$\lambda(X_1, ..., X_n) = \frac{\max_{\theta\in\Theta_0} L(X_1, ..., X_n; \theta)}{\max_{\theta\in\Theta} L(X_1, ..., X_n; \theta)} \tag{5.9}$$

која се нарекува **коефициент на подобност** (LR - likelihood ratio). Да забележиме дека ако $\widehat{\theta}$ е такво да $\max_{\theta\in\Theta} L(X_1, ..., X_n; \theta) = L(X_1, ..., X_n; \widehat{\theta})$, тогаш $\widehat{\theta}$ е ML оценувач за параметарот θ. Исто така, ако H_0 е проста хипотеза т.е. $\Theta_0 = \{\theta_0\}$, тогаш $\max_{\theta\in\Theta_0} L(X_1, ..., X_n; \theta) = L(X_1, ..., X_n; \theta_0)$.

За коефициентот на подобност важи

$$0 \leq \lambda(x_1, ..., x_n) \leq 1.$$

Веројатноста дека вредноста на $\lambda(x_1, ..., x_n)$ е голема, е поголема тогаш кога хипотезата H_0 е точна и обратно, таа е релативно мала кога H_0 не е точна. Затоа, разумно е критичната област C да се одбере така да ги содржи оние вредности на $x = (x_1, ..., x_n)$ за кои соодветниот коефициент на подобност не е поголем од даден број c $(0 < c \leq 1)$ т.е. има облик

$$C = \{x = (x_1, ..., x_n) : \lambda(x_1, ..., x_n) \leq c\}.$$

Постапката за наоѓање на критичен домен со големина α со метод со коефициент на подобност е да при дадени вредности на големината на примерокот n и нивото на значајност на тестот α се наоѓа вредноста на константата c од условот да

$$\max_{\theta\in\Theta_0} P\{\lambda(X_1, ..., X_n) \leq c\} = \alpha.$$

Кога нултата хипотеза H_0 е проста, овој услов преминува во

$$P\{\lambda(X_1, ..., X_n) \leq c | \theta = \theta_0\} = P_0\{\lambda(X_1, ..., X_n) \leq c\} = \alpha.$$

Ирена Стојковска

5.2.4 Тестови за параметрите на нормална распределба

Методот со коефициент на подобност овозможува да на едноставен начин се определат некои од тестовите за параметрите на нормална распределба $\mathcal{N}(m, \sigma^2)$. Претпоставуваме дека се зададени големината на примерокот n и нивото на значајност α.

Во случај на две независни нормално распределени обележја X и Y со распределби $\mathcal{N}(m_1, \sigma_1^2)$ и $\mathcal{N}(m_2, \sigma_2^2)$ соодветно, претпоставуваме дека се зададени големините на примероците n_1 и n_2 соодветно и нивото на значајност α.

(1) Тестирање на $H_0 : m = m_0$, против $H_1 : m \neq m_0$, кога σ^2 е позната.

Кога $X \sim \mathcal{N}(m, \sigma^2)$ при σ^2 познато, функцијата на подобност е

$$L(x, m) = (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2\},\ x = (x_1, ..., x_n) \in \mathbb{R}^n,\ m \in \mathbb{R}.$$

При σ^2 познато ML оценувач за m е $\overline{X}_n$, односно $\mathcal{L}(m) = L(x, m)$ достигнува максимум во $m = \overline{x} = \frac{1}{n}(x_1 + ... + x_n)$. И бидејќи нултата хипотеза H_0 е проста (т.е. $\Theta_0 = \{m_0\}$), за коефициентот на подобност имаме

$$\lambda(x_1, ..., x_n) = \frac{L(x, m_0)}{L(x, \overline{x})} =$$

$$= \exp\{-\frac{1}{2\sigma^2}(\sum_{i=1}^{n}(x_i - m_0)^2 - \sum_{i=1}^{n}(x_i - \overline{x})^2)\} = \exp\{-\frac{n}{2\sigma^2}(\overline{x} - m_0)^2)\}.$$

Според методот со коефициент на подобност, критичната област C ги содржи сите оние $x = (x_1, ..., x_n) \in \mathbb{R}^n$ за кои

$$\exp\{-\frac{n}{2\sigma^2}(\overline{x} - m_0)^2)\} \leq c,$$

каде $0 < c \leq 1$. Последното неравенство е еквивалентно со

$$\left|\frac{\overline{x} - m_0}{\sigma}\sqrt{n}\right| \geq c_0,$$

каде $c_0 = \sqrt{-2 \ln c} \geq 0$, односно критичната област може да се запише во обликот

$$C = \{x = (x_1, ..., x_n) \in \mathbb{R}^n : \left|\frac{\overline{x} - m_0}{\sigma}\sqrt{n}\right| \geq c_0\}.$$

Вредноста на бројот c_0 ја одредуваме од даденото ниво на значајност на тестот α, односно од условот

$$P_0\{\left|\frac{\overline{X}_n - m_0}{\sigma}\sqrt{n}\right| \geq c_0\} = \alpha.$$

Ирена Стојковска

Повторно користевме дека нултата хипотеза H_0 е проста. Сега, бидејќи статистиката

$$Z = \frac{\overline{X}_n - m_0}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1),$$

имаме дека

$$\alpha = P_0\{|Z| \geq c_0\} = 1 - P_0\{-c_0 < Z < c_0\} = 1 - (\Phi_0(c_0) - \Phi_0(-c_0)) = 1 - 2\Phi_0(c_0),$$

каде $\Phi_0(c_0) = \int_0^{c_0} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$ е Лапласовиот интеграл. Значи, за c_0 имаме дека е бројот

$$c_0 = \Phi_0^{-1}(\frac{1-\alpha}{2}).$$

На сличен начин се добива дека критичните области при тестирање на

(а) $H_0 : m = m_0$, против $H_1 : m > m_0$, кога σ^2 е позната,

(б) $H_0 : m = m_0$, против $H_1 : m < m_0$, кога σ^2 е позната,

соодветно се дадени со

$$\text{(а) } C = \{x = (x_1, ..., x_n) \in \mathbb{R}^n : \frac{\overline{x} - m_0}{\sigma}\sqrt{n} \geq c_0\},\ c_0 = \Phi_0^{-1}(\frac{1-2\alpha}{2}),$$

$$\text{(б) } C = \{x = (x_1, ..., x_n) \in \mathbb{R}^n : \frac{\overline{x} - m_0}{\sigma}\sqrt{n} \leq -c_0\},\ c_0 = \Phi_0^{-1}(\frac{1-2\alpha}{2}).$$

(2) Тестирање на $H_0 : m = m_0$, против $H_1 : m \neq m_0$, кога σ^2 не е позната.

Кога дисперзијата σ^2 не е позната, нултата хипотеза може да се запише како $H_0 : (m, \sigma^2) = (m_0, \sigma^2)$, односно тогаш $\Theta_0 = \{(m_0, \sigma^2) : \sigma^2 > 0\}$, каде m_0 е фиксен реален број. Тогаш, задачата преминува во тестирање на сложена хипотеза $H_0 : t = (m, \sigma^2) \in \Theta_0$ против сложена хипотеза $H_1 : t = (m, \sigma^2) \in \Theta_1 = \Theta \setminus \Theta_0$, каде множеството допустливи вредности за параметрите на нормалната распределба е $\Theta = \{(m, \sigma^2) : m \in \mathbb{R},\ \sigma^2 > 0\}$.

Имајќи во предвид дека функцијата на подобност е

$$L(x, m, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2\},$$

за $x = (x_1, ..., x_n) \in \mathbb{R}^n,\ m \in \mathbb{R},\ \sigma^2 > 0$, и бидејќи ML оценувачи за m и σ^2 се $\overline{X}_n$ и $\overline{S}_n^2$ соодветно, имаме дека

$$\max_{t \in \Theta} L(x, t) = \max_{m \in \mathbb{R},\ \sigma^2 > 0} L(x, m, \sigma^2) = L(x, \overline{x}, s^2),$$

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

$$\max_{t\in\Theta_0} L(x,t) = \max_{\sigma^2>0} L(x, m_0, \sigma^2) = L(x, m_0, \sigma_0^2),$$

каде $\sigma_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_0)^2$. Тогаш, коефициентот на подобност е

$$\lambda(x_1, ..., x_n) = \frac{L(x, m_0, \sigma_0^2)}{L(x, \overline{x}, s^2)} = \Big(\frac{\sum_{i=1}^{n}(x_i - m_0)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\Big)^{-n/2}.$$

Според методот со коефициент на подобност, критичната област C ги содржи сите $x = (x_1, ..., x_n) \in \mathbb{R}^n$ за кои важи

$$\Big(\frac{\sum_{i=1}^{n}(x_i - m_0)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\Big)^{-n/2} \leq c,$$

каде $0 < c \leq 1$. Последното неравенство е еквивалентно со

$$\frac{(\overline{x} - m_0)^2}{s^2} \geq c^{-2/n} - 1,$$

односно со

$$\Big|\frac{\overline{x} - m_0}{s}\sqrt{n-1}\Big| \geq c_0,$$

каде $c_0 = \sqrt{(n-1)(c^{-2/n} - 1)} \geq 0$, па критичната област е

$$C = \{x = (x_1, ..., x_n) \in \mathbb{R}^n : \Big|\frac{\overline{x} - m_0}{s}\sqrt{n-1}\Big| \geq c_0\},$$

каде c_0 се одредува од даденото ниво на значајност α. Познато ни е дека статистиката

$$T = \frac{\overline{X}_n - m_0}{\overline{S}_n}\sqrt{n-1} \sim t_{n-1}.$$

Затоа од

$$\alpha = P_0\{|T| \geq c_0\},$$

имаме дека

$$c_0 = t_{n-1,\alpha},$$

каде означуваме со $t_{n,\alpha}$ број (кој се чита од таблица) за кој важи $P\{|t_n| > t_{n,\alpha}\} = \alpha$, каде t_n е случајна променлива со студентова распределба со n степени на слобода.

На сличен начин се добива дека критичните области при тестирање на

(а) $H_0 : m = m_0$, против $H_1 : m > m_0$, кога σ^2 не е позната,

(б) $H_0 : m = m_0$, против $H_1 : m < m_0$, кога σ^2 не е позната,

соодветно се дадени со

$$\text{(а) } C = \{x = (x_1, ..., x_n) \in \mathbb{R}^n : \frac{\overline{x} - m_0}{s}\sqrt{n-1} \geq c_0\},\ c_0 = t_{n-1,2\alpha},$$

$$\text{(б) } C = \{x = (x_1, ..., x_n) \in \mathbb{R}^n : \frac{\overline{x} - m_0}{s}\sqrt{n-1} \leq -c_0\},\ c_0 = t_{n-1,2\alpha}.$$

(3) Тестирање на $H_0 : \sigma^2 = \sigma_0^2$, против $H_1 : \sigma^2 \neq \sigma_0^2$.

Покрај постапката на одредување на критичната област C со помош на коефициенттот на подобност, при тестирање на хипотези за параметрите на нормална распределба, може да се примени и постапка со интервал на доверба. Тогаш, при тестирање на $H_0 : \theta = \theta_0$ против $H_1 : \theta \neq \theta_0$ со ниво на значајност α, хипотезата H_0 се прифаќа ако врз основа на податоците $x = (x_1, ..., x_n)$ вредноста θ_0 е обфатена со интервалот $(L(x), U(x))$, каде $(L(X_1, ..., X_n), U(X_1, ..., X_n))$ е $100(1-\alpha)\%$ интервал на доверба за θ, во спротивно H_0 се отфрла. Значи, критичната област има облик

$$C = \{x = (x_1, ..., x_n) \in \mathbb{R}^n : L(x) \geq \theta_0 \text{ или } U(x) \leq \theta_0\}.$$

При тоа, ако статистиките $L(X_1, ..., X_n)$ и $U(X_1, ..., X_n)$ го определуваат интервалот на доверба за θ со минимална должина, тогаш големината на вака дефинираната критична област е α, имено важи

$$P_0\{(X_1, ..., X_n) \in C\} = 1 - P_0\{L(X_1, ..., X_n) < \theta_0 < U(X_1, ..., X_n)\} = 1 - (1-\alpha) = \alpha.$$

Така, знаеме дека $100(1-\alpha)\%$ интервал на доверба за σ^2 е

$$(\frac{n\overline{S}_n^2}{b}, \frac{n\overline{S}_n^2}{a}),$$

каде $a = \chi_{n-1,1-\alpha/2}$ и $b = \chi^2_{n-1,\alpha/2}$, при што со $\chi^2_{n,\alpha}$ го означуваме бројот (кој се чита од таблица) за кој важи $P\{\chi_n^2 > \chi^2_{n,\alpha}\} = \alpha$, каде χ_n^2 е случајна променлива со χ^2 распределба со n степени на слобода. Тогаш, критичната област е

$$C = \{x = (x_1, ..., x_n) \in \mathbb{R}^n : \frac{ns^2}{\sigma_0^2} \geq b \text{ или } \frac{ns^2}{\sigma_0^2} \leq a\},$$

каде $a = \chi_{n-1,1-\alpha/2}$ и $b = \chi^2_{n-1,\alpha/2}$.

На сличен начин се добива дека критичните области при тестирање на

$$\text{(а) } H_0 : \sigma^2 = \sigma_0^2, \text{ против } H_1 : \sigma^2 > \sigma_0^2,$$

(б) $H_0 : \sigma^2 = \sigma_0^2$, против $H_1 : \sigma^2 < \sigma_0^2$,

соодветно се дадени со

$$\text{(а) } C = \{x = (x_1, ..., x_n) \in \mathbb{R}^n : \frac{ns^2}{\sigma_0^2} \geq c\},\ c = \chi^2_{n-1,\alpha},$$

$$\text{(б) } C = \{x = (x_1, ..., x_n) \in \mathbb{R}^n : \frac{ns^2}{\sigma_0^2} \leq c\},\ c = \chi^2_{n-1,1-\alpha}.$$

(4) Тестирање на $H_0 : m_1 = m_2$, против $H_1 : m_1 \neq m_2$, кога σ_1^2 и σ_2^2 се познати.

Повторно користиме метод со интервал на доверба, овој пат за разликата $m_1 - m_2$. Прво, за централна статистика ја земаме статистиката

$$Z = \frac{(\overline{X}_{n_1} - m_1) - (\overline{Y}_{n_2} - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1),$$

(покажи!). Заради симетричноста на распределбата на Z, значи дека бараме таков број c така што

$$1 - \alpha = P\{|Z| < c|H_0\} = P_0\{\left|\frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right| < c\},$$

од каде следи дека

$$c = \Phi_0^{-1}(\frac{1-\alpha}{2})$$

и критичната област е

$$C = \{(x, y) \in \mathbb{R}^{n_1+n_2} : \left|\frac{\overline{x} - \overline{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right| \geq c\}.$$

На сличен начин се добива дека критичните области при тестирање на

(а) $H_0 : m_1 = m_2$, против $H_1 : m_1 > m_2$, кога σ_1^2 и σ_2^2 се познати,

(б) $H_0 : m_1 = m_2$, против $H_1 : m_1 < m_2$, кога σ_1^2 и σ_2^2 се познати,

соодветно се дадени со

$$\text{(а) } C = \{(x, y) \in \mathbb{R}^{n_1+n_2} : \frac{\overline{x} - \overline{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq c\},\ c = \Phi_0^{-1}(\frac{1-2\alpha}{2}),$$

$$\text{(б) } C = \{(x, y) \in \mathbb{R}^{n_1+n_2} : \frac{\overline{x} - \overline{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq -c\},\ c = \Phi_0^{-1}(\frac{1-2\alpha}{2}).$$

Ирена Стојковска

(5) Тестирање на $H_0 : m_1 = m_2$, против $H_1 : m_1 \neq m_2$, кога $\sigma_1^2 = \sigma_2^2 = \sigma^2$ не се познати.

Овој пат за одредување на интервал на доверба за $m_1 - m_2$ за централна статистика се користи статистиката

$$T = \frac{(\overline{X}_{n_1} - m_1) - (\overline{Y}_{n_2} - m_2)}{\sqrt{n_1\overline{S}_x^2 + n_2\overline{S}_y^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}(n_1 + n_2 - 2)} \sim t_{n_1+n_2-2},$$

(види Теорема 3.5). Па, заради симетричноста на распределбата на T при одредување на интервал на доверба со минимална должина бараме број c така што

$$1 - \alpha = P\{|T| < c|H_0\} = P_0\Big\{\Big|\frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{n_1\overline{S}_x^2 + n_2\overline{S}_y^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}(n_1 + n_2 - 2)}\Big| < c\},$$

од каде следи дека

$$c = t_{n_1+n_2-2,\,\alpha}$$

и критичната област е

$$C = \{(x, y) \in \mathbb{R}^{n_1+n_2} : \Big|\frac{\overline{x} - \overline{y}}{\sqrt{n_1 s_x^2 + n_2 s_y^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}(n_1 + n_2 - 2)}\Big| \geq c\}.$$

На сличен начин се добива дека критичните области при тестирање на

(а) $H_0 : m_1 = m_2$, против $H_1 : m_1 > m_2$, кога $\sigma_1^2 = \sigma_2^2 = \sigma^2$ не се познати,

(б) $H_0 : m_1 = m_2$, против $H_1 : m_1 < m_2$, кога $\sigma_1^2 = \sigma_2^2 = \sigma^2$ не се познати,

соодветно се дадени со

$$\text{(а) } C = \{(x, y) \in \mathbb{R}^{n_1+n_2} : \frac{\overline{x} - \overline{y}}{\sqrt{n_1 s_x^2 + n_2 s_y^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}(n_1 + n_2 - 2)} \geq c\},$$

$$\text{каде } c = t_{n_1+n_2-2,\,2\alpha},$$

$$\text{(б) } C = \{(x, y) \in \mathbb{R}^{n_1+n_2} : \frac{\overline{x} - \overline{y}}{\sqrt{n_1 s_x^2 + n_2 s_y^2}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}(n_1 + n_2 - 2)} \leq -c\},$$

$$\text{каде } c = t_{n_1+n_2-2,\,2\alpha}.$$

Ирена Стојковска

(6) Тестирање на $H_0 : \sigma_1^2 = \sigma_2^2$, против $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Како централни статистики при одредување на интервал на доверба за σ_1^2/σ_2^2 се користат статистиките

$$F_1 = \frac{n_1(n_2-1)\sigma_2^2\overline{S}_x^2}{n_2(n_1-1)\sigma_1^2\overline{S}_y^2} \sim F_{n_1-1,\,n_2-1}, \text{ и } F_2 = \frac{n_2(n_1-1)\sigma_1^2\overline{S}_y^2}{n_1(n_2-1)\sigma_2^2\overline{S}_x^2} \sim F_{n_2-1,\,n_1-1},$$

(види Теорема 3.6). Имено, ако $s_x^2 > s_y^2$ тогаш критичната област е

$$C = \{(x,y) \in \mathbb{R}^{n_1+n_2} : \frac{n_1(n_2-1)s_x^2}{n_2(n_1-1)s_y^2} \geq c\},\ c = F_{n_1-1,\,n_2-1,\,\alpha/2},$$

и ако $s_x^2 < s_y^2$ тогаш критичната област е

$$C = \{(x,y) \in \mathbb{R}^{n_1+n_2} : \frac{n_2(n_1-1)s_y^2}{n_1(n_2-1)s_x^2} \geq c\},\ c = F_{n_2-1,\,n_1-1,\,\alpha/2}.$$

При тоа со $F_{n_1,\,n_2,\,\alpha}$ се означува бројот (кој се чита од таблица) за која важи $P\{F_{n_1,\,n_2} > F_{n_1,\,n_2,\,\alpha}\} = \alpha$, каде $F_{n_1,\,n_2}$ е случајна променлива која има Фишерова рспределба со n_1, n_2 степени на слобода.

На сличен начин се добива дека критичните области при тестирање на

$$\text{(а) } H_0 : \sigma_1^2 = \sigma_2^2, \text{ против } H_1 : \sigma_1^2 > \sigma_2^2,$$

$$\text{(б) } H_0 : \sigma_1^2 = \sigma_2^2, \text{ против } H_1 : \sigma_1^2 < \sigma_2^2,$$

соодветно се дадени со

$$\text{(а) } C = \{(x,y) \in \mathbb{R}^{n_1+n_2} : \frac{n_1(n_2-1)s_x^2}{n_2(n_1-1)s_y^2} \geq c\},\ c = F_{n_1-1,\,n_2-1,\,\alpha},$$

$$\text{(б) } C = \{(x,y) \in \mathbb{R}^{n_1+n_2} : \frac{n_2(n_1-1)s_y^2}{n_1(n_2-1)s_x^2} \geq c\},\ c = F_{n_2-1,\,n_1-1,\,\alpha}.$$

5.3 Тестирање на непараметарски хипотези

Нека X е обележје со непозната функција на распределба F_X. Непараметарските хипотези се однесуваат на функцијата на распределба на обележјето X и не зависат од непознатите параметри. Постојат неколку типа на непараметарски тестови. **Тестовите на согласност** ја тестирааат нултата хипотеза $H_0 : F_X = F$, каде F е дадена функција на распределба, против алтернативната хипотеза $H_1 : F_X \neq F$. Ако разгледуваме две обележја X и Y со непознати функции на распределба F_X и F_Y соодветно, **тестовите за хомогеност** ја тестирааат нултата хипотеза $H_0 : F_X = F_Y$, против алтернативната хипотеза $H_1 : F_X \neq F_Y$. **Тестовите за независност** ја тестирааат нултата хипотеза $H_0 : F = F_X F_Y$, каде F е функцијата на распределба на случајниот вектор (X, Y), против алтернативната хипотеза $H_1 : F \neq F_X F_Y$.

5.3.1 Пирсонов χ^2-тест на согласност

Пирсоновиот χ^2-тест на согласност е еден од првите предложени тестови кој се темели на едноставен математички модел. Со него се тестира хипотезата $H_0 : F_X = F$, против алтернативната хипотеза $H_1 : F_X \neq F$, каде F_X е непознатата функција на распределба на обележјето X и F е дадена функција на распределба. Главните карактеристики на овој тест кои му обезбедуваат широка примена се тие што тој може да се применува за произволна функција на распределба F и реализацијата на тест статистиката е лесно пресметлива.

Нека $(X_1, X_2, ..., X_n)$ е примерок кој одговара на обележјето X. Дефинирањето на Пирсоновата χ^2 тест статистика е доста едноставно. Најнапред го разбиваме просторот од релани броеви на r дисјунктни подмножества, односно $\mathbb{R} = S_1 \cup S_2 \cup \cdots \cup S_r$, $S_i \cap S_j = \emptyset$, $i \neq j$. За секој $i \in \{1, 2, ..., r\}$ го означуваме со M_i бројот на случајни променливи од примерокот $(X_1, X_2, ..., X_n)$ кои примаат вредности од множеството S_i, и нека $p_i = P\{X \in S_i | H_0\}$. Тогаш, за секој $i \in \{1, 2, ..., r\}$, под претпоставка H_0 да е точна, случајната променлива M_i има $\mathcal{B}(n, p_i)$ распределба. И бидејќи $E(M_i) = np_i$, статистиката

$$\chi^2 = \sum_{i=1}^{r} \frac{(M_i - np_i)^2}{np_i} \tag{5.10}$$

добро го опишува отстапувањата на случајните променливи $M_1, M_2, ..., M_r$ од нивните математички очекувања. Статистиката χ^2 дефинирана со (5.10) се нарекува **Пирсонова χ^2 статистика**. Имено, се покажува дека распределбата на случајната променлива χ^2 асимптотски се стреми кон χ^2 распределба со $r - 1$ степени на слобода, односно

$$\chi^2 \xrightarrow{dist} \chi^2_{r-1}. \tag{5.11}$$

Ирена Стојковска

Знаејќи ја асимптотската распределба на тест статистиката χ^2, може да ја определиме критичната област со големина α (ниво на зачајност, веројатност за грешка од прв вид) за тестирање на хипотезата H_0. Имено, бидејќи

$$P\{\chi^2 > \chi^2_{r-1,\alpha} | H_0\} \approx \alpha, \tag{5.12}$$

каде бројот $\chi^2_{r-1,\alpha}$ се чита од таблица, имаме дека критичната област е

$$C = \{x \mid \overline{\chi}^2 = \sum_{i=1}^{r} \frac{(m_i - np_i)^2}{np_i} > \chi^2_{r-1,\alpha}\},$$

каде m_i е бројот на компоненти од реализацијата $x = (x_1, ..., x_n)$ на примерокот $(X_1, ..., X_n)$ кои припаѓаат во множеството S_i, $i \in \{1, ..., r\}$. Ова значи дека, ако врз база на податоците $x_1, ..., x_n$ добиеме дека $\overline{\chi}^2 > \chi^2_{r-1,\alpha}$, тогаш хипотезата H_0 ја отфрламе, во спротивно, ако $\overline{\chi}^2 \leq \chi^2_{r-1,\alpha}$, хипотезата H_0 ја прифаќаме. Да забележиме дека во пракса, апроксимацијата (5.12), а со тоа и самиот Пирсонов χ^2 - тест, дава задоволителни резултати за $n \geq 50$ и $m_i \geq 5$, за секој $i \in \{1, 2, ..., r\}$.

Пирсоновиот χ^2-тест може да се модифицира и за случај кога хипотезата $H_0 : F_X = F$ не е проста, односно кога функцијата на распределба F зависи од непознати праметри. Во тој случај нултата хипотеза е $H_0 : F_X \in \{F(x, \theta) | \theta \in \Theta\}$, па веројатностите $p_i(\theta) = P\{X \in S_i | H_0\}$, $i \in \{1, 2, ..., r\}$ зависат од непознатиот параметар θ. Тогаш, и Пирсоновата χ^2 статистика исто така ќе зависи од непознатиот параметар, односно

$$\chi^2(\theta) = \sum_{i=1}^{r} \frac{(M_i - np_i(\theta))^2}{np_i(\theta)}. \tag{5.13}$$

Нека $\theta = (\theta_1, ..., \theta_j)$, каде $j \leq r - 1$, односно функцијата на распределба F има j непознати параметри. Се покажува дека ако $\widehat{\theta}$ е максимално подобен оценувач за θ, тогаш важи

$$\chi^2(\widehat{\theta}) \xrightarrow{dist} \chi^2_{r-j-1}. \tag{5.14}$$

Имено, ова значи дека во случај кога функцијата на распределба F има непознати параметри (вкупно j непознати параметри), најнапред тие се заменуваат со нивни оценки врз база на дадените податоци $x_1, ..., x_n$ и со тоа распределбата F е потполно одредена. Па, заради (5.14), во овој случај критичната област ќе биде

$$C = \{x \mid \overline{\chi}^2 = \sum_{i=1}^{r} \frac{(m_i - np_i)^2}{np_i} > \chi^2_{r-j-1,\alpha}\},$$

па ако $\overline{\chi}^2 > \chi^2_{r-j-1,\alpha}$, тогаш хипотезата H_0 ја отфрламе, во спротивно, ако $\overline{\chi}^2 \leq \chi^2_{r-j-1,\alpha}$, хипотезата H_0 ја прифаќаме.

Ирена Стојковска

5.3.2 Колмогоров тест на согласност

Кога при тестирање на хипотезата $H_0 : F_X = F$, функцијата на распределба F е непрекината, за тест статистика може да се земе **Колмогоровата тест статистика** дефинирана со

$$D_n = D_n(X_1, ..., X_n) = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|, \tag{5.15}$$

каде $(X_1, ..., X_n)$ е примерок кој одговара на обележето X и $F_n(x)$ е емпириската функција на распределба на примерокот.

Се покажува дека при големи вредности на n, случајната променлива $\sqrt{n}D_n$ има асимптотска функција на распределба

$$K(x) = \lim_{n\to\infty} P\{\sqrt{n}D_n \leq x\} = 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2x^2},\ x > 0,$$

која се нарекува **Колмогорова распределба**.

Тогаш, знаејќи ја распределбата на тест статистиката, може да се определи критичната вредност c_0 од критичната област C при тестирање на $H_0 : F_X = F$ со ниво на значајност α, од условот

$$P\{D_n(X_1, ..., X_n) > c_0 | H_0\} = \alpha.$$

Се изведува дека критичната област го има обликот

$$C = \{x \mid D_n(x) > d_{n,\alpha}\},$$

каде вредноста $d_{n,\alpha}$ е таква да $P\{D_n > d_{n,\alpha}\} = \alpha$ и се чита од таблица. Додека пак, при големи вредности на $n \geq 100$, критичната област го има обликот

$$C = \{x \mid D_n(x) > c_0\},$$

каде критичната вредност c_0 може да се пресметува со примена на следната таблица.

α	0,10	0,05	0,01
c_0	$\frac{1,22}{\sqrt{n}}$	$\frac{1,36}{\sqrt{n}}$	$\frac{1,63}{\sqrt{n}}$

Во пракса, вредноста на статистиката $D_n(x) = d_n$ за дадена низа од статистички податоци $x = (x_1, ..., x_n)$ со подреден облик $x_{(1)} \leq ... \leq x_{(n)}$ се пресметува според формулата

$$d_n = \max_{1\leq i\leq n} |F_n(x_{(i)}) - F(x_{(i)})|,$$

Ирена Стојковска

односно според формулата

$$d_n = \max_{1\le i\le r} |F_n(a_i) - F(a_i)|,$$

каде a_i, $i = 1, ..., r$ се десните граници на интервалите при интервало зададени статистички податоци. Па, ако $d_n > d_{n,\alpha}$ (односно $d_n > c_0$ при вредности $n \ge 100$), тогаш хипотезата H_0 ја отфрламе, во спротивно, ако $d_n \le d_{n,\alpha}$ (односно $d_n \le c_0$ при вредности $n \ge 100$), хипотезата H_0 ја прифаќаме.

5.3.3 Тест за хомогеност на Колмгоров-Смирнов

Колмогоровата распределба наоѓа примена и при конструкција на статистичкиот тест за тестирање на хипотезата $H_0 : F_X = F_Y$, каде F_X и F_Y се функциите на распределба на непрекинатите обележја X и Y соодветно. Нека $(X_1, .., X_{n_1})$ и $(Y_1, .., Y_{n_2})$ се два независни примерока кои одговараат на обележјата X и Y соодветно. Во овој случај се користи **тест статистиката на Колмогоров-Смирнов** дефинирана со

$$D_{\frac{n_1n_2}{n_1+n_2}} = D_{\frac{n_1n_2}{n_1+n_2}}(X_1, ..., X_{n_1}, Y_1, ..., Y_{n_2}) = \sup_{x\in\mathbb{R}} |F_{n_1}(x) - F_{n_2}(x)|, \tag{5.16}$$

каде $F_{n_1}(x)$ и $F_{n_2}(x)$ се емпириските функции на распределба на примероците $(X_1, .., X_{n_1})$ и $(Y_1, .., Y_{n_2})$ соодветно. Ако хипотезата H_0 е точна, тогаш емпириските функции на распределба $F_{n_1}(x)$ и $F_{n_2}(x)$ оценуваат една иста функција на распределба $F_X = F_Y = F$, па за големи вредности на n_1 и n_2 природно е да се очекуваат мали реализирани вредности на тест статистиката $D_{\frac{n_1n_2}{n_1+n_2}}$.

Се покажува дека при големи вредности на n_1 и n_2, случајната променлива $\sqrt{\frac{n_1n_2}{n_1+n_2}} D_{\frac{n_1n_2}{n_1+n_2}}$ има асимптотски Колмогорова распределба. Тогаш, критичната вредност c_0 од критичната област C при тестирање на H_0 со ниво на значајност α, се определува од условот

$$P\{D_{\frac{n_1n_2}{n_1+n_2}}(X_1, ..., X_{n_1}, Y_1, ..., Y_{n_2}) > c_0 | H_0\} = \alpha.$$

Се добива дека критичната област го има обликот

$$C = \{(x, y) \mid D_{\frac{n_1n_2}{n_1+n_2}}(x, y) > d_{\frac{n_1n_2}{n_1+n_2},\alpha}\},$$

каде вредноста $d_{n,\alpha}$ е таква да $P\{D_n > d_{n,\alpha}\} = \alpha$ и се чита од таблица. И ако ја означиме со $d_{\frac{n_1n_2}{n_1+n_2}}$ реализацијата на статистиката $D_{\frac{n_1n_2}{n_1+n_2}}(x, y)$ за дадени низи статистички податоци $x = (x_1, ..., x_{n_1})$ и $y = (y_1, ..., y_{n_2})$, тогаш ако $d_{\frac{n_1n_2}{n_1+n_2}} > d_{\frac{n_1n_2}{n_1+n_2},\alpha}$, хипотезата H_0 ја отфрламе, во спротивно, ако $d_{\frac{n_1n_2}{n_1+n_2}} \le d_{\frac{n_1n_2}{n_1+n_2},\alpha}$, тогаш хипотезата H_0 ја прифаќаме.

Ирена Стојковска

5.3.4 χ^2-тест за независност

Претходно зборувавме за зависноста меѓу обележјата X и Y, врз основа дводимензионалните статистички податоци $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, изразена преку вредностите на одредени параметри (криви на регресија, прави на регресија, коефициент на корелација, отстапување од статистичка независност, степен на статистичка зависност) и нивната геометриска интерпретација (види Дескриптивна статистика, Дводимензионални обележја). Овде ќе се задржиме на тестирање на хипотезата $H_0 : X$ и Y се независни случајни променливи, која симболички може да се запише како $H_0 : F = F_X F_Y$, каде F_X и F_Y се непознатите функции на распределба на обележјата X и Y соодветно, и F е функцијата на распределба на случајниот вектор (X, Y).

Нека $((X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n))$ е дводимензионален примерок кој одговара на дводимензионалното обележје (X, Y). Врз основа на овој примерок ќе конструираме статистички тест за тестирање на хипотезата H_0. Нека $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ е реализација на дводимензионалниот примерок дадена со табелата на контингенција на честотите, Табела 2.6, и соодветната табела на контингенција на релативните честоти, Табела 2.7. Во случај кога X и Y се непрекинати обележја, по групирањето на податоците во интервали, ја користиме Табела 2.11.

Ако хипотезата H_0 е точна, тогаш од условот за независност на две случајни променливи од дискретен тип важи

$$p_{i,j} = q_i r_j,\ i = 1, ..., r,\ j = 1, ..., s \tag{5.17}$$

(при тоа ги користиме ознаките од горе споменатите табели). Па, ако го означиме со $t = (q_1, ..., q_r, r_1, ..., r_s)$ векторот од непознати параметри, заради (5.17) и условите

$$\sum_{i=1}^{r}\sum_{j=1}^{s} p_{ij} = 1,\ \sum_{i=1}^{r} q_i = 1,\ \sum_{j=1}^{s} r_j = 1,$$

не се сите негови компоненти независни, имено постојат две функционални зависности, и тогаш димензијата на векторот t е $r + s - 2$.

Понатаму, со M_{ij} го означуваме бројот на случајни парови од примерокот $((X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n))$ кои ја примаат вредноста (a_i, b_j), повторно се користиме со ознаките од горе споменатите табели, и нека

$$G_i = M_{i1} + M_{i2} + ... + M_{is},\ i = 1, ..., r,\ H_j = M_{1j} + M_{2j} + ... + M_{rj},\ j = 1, ..., s. \tag{5.18}$$

Тогаш, слично како Пирсоновата χ^2 статистика (5.10), под претпоставка H_0 да е точна, статистиката

$$D = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(M_{ij} - np_{ij})^2}{np_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(M_{ij} - nq_i r_j)^2}{nq_i r_j} \xrightarrow{dist} \chi^2_{rs-1}, \tag{5.19}$$

Ирена Стојковска

го карактеризира отстапувањето на случајните променливи M_{ij} од нивните математички очекувања, односно од теориската распределба и таа има χ^2 распределба со $rs-1$ степени на слобода, според (5.11).

И слично како во (5.13), бидејќи p_{ij} зависат од векторот од непознати параметри t, тогаш и статистиката (5.19) ќе зависи од t, односно

$$D(t)=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(M_{ij}-np_{ij}(t))^2}{np_{ij}(t)}=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(M_{ij}-nq_i(t)r_j(t))^2}{nq_i(t)r_j(t)}. \tag{5.20}$$

Нека $\widehat{T}$ е максимално подобен оценувач за t, и бидејќи димензијата на t е $r+s-2$, слично како во (5.14), статистиката $D(\widehat{T})$ ќе има χ^2 распределба со $rs-(r+s-2)-1=rs-r-s+1=(r-1)(s-1)$ степени на слобода, односно

$$D(\widehat{T})=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(M_{ij}-np_{ij}(\widehat{T}))^2}{np_{ij}(\widehat{T})}=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(M_{ij}-nq_i(\widehat{T})r_j(\widehat{T}))^2}{nq_i(\widehat{T})r_j(\widehat{T})}\xrightarrow{dist}\chi^2_{(r-1)(s-1)}. \tag{5.21}$$

Од равенствата

$$q_i(\widehat{T})=\frac{G_i}{n}, i=1,...,r,\ r_j(\widehat{T})=\frac{H_j}{n},\ j=1,...,s,$$

каде G_i и H_j се дефинирани со (5.18), статистиката $D(\widehat{T})$ може да ја запишеме во еквивалентен облик

$$D(\widehat{T})=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(M_{ij}-\frac{G_iH_j}{n})^2}{\frac{G_iH_j}{n}}=n\Big(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{M_{ij}^2}{G_iH_j}-1\Big). \tag{5.22}$$

Тогаш, за определување на критичната област со големина α за тестирање на хипотезата H_0 се користиме со приближното равенство

$$P\{D(\widehat{T})>\chi^2_{(r-1)(s-1),\,\alpha}\}\approx\alpha,$$

и заклучуваме дека критичната област го има обликот

$$C=\{(x,y)\mid d=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(f_{ij}-\frac{g_ih_j}{n})^2}{\frac{g_ih_j}{n}}=n\Big(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{f_{ij}^2}{g_ih_j}-1\Big)=nf^2>\chi^2_{(r-1)(s-1),\,\alpha}\},$$

каде f_{ij},g_i,h_j се честотите кои за низата податоци $(x_1,y_1),(x_2,y_2),...,(x_n,y_n)$ дадени се со Табела 2.6, и f^2 е отстапувањето од статитичката независност дефинирано со (2.13). Тоа значи дека, ако $d>\chi^2_{(r-1)(s-1),\,\alpha}$, тогаш хипотезата H_0 ја отвфрламе, и ако $d\le\chi^2_{(r-1)(s-1),\,\alpha}$, тогаш хипотезата H_0 ја прифаќаме.

Ирена Стојковска

6

Регресиона анализа

При мерење на повеќе обележја, често се поставува прашањето како тие зависат едно од друго. На пример, како зависи нивото на холестеролот во крвта од возраста на индивидуата, како зависи потрошувачката на електричната енергија во домаќинството од надворешната температура и месечните примања на членовите на домаќинството и слично. При тоа променливата која е предмет на истражувањето се нарекува **зависна променлива** и се означува со Y и таа е случајна променлива (на пример, нивото на холестеролот во крвта, односно потрошувачката на електричната енергија во домаќинството), додека променливата која може да се контролира и која влијае на промените на зависната променлива се нарекува **независна променлива** и во општ случај нив може да ги има повеќе од една и се означуваат со $x^{(1)}, ..., x^{(r)}$ и тие не се случајни променливи (на пример, возраста на индивидуата, односно надворешната температура и месечните примања на членовите на домаќинството).

Одредувањето на зависноста на променливата Y од независните променливи $x^{(1)}, ..., x^{(r)}$ е преку формирање на математички модел кој ќе ја изразува таа зависност. Бидејќи во реалните ситуации на зависноста често влијаат и некои случајни фактори (на пример, грешки при мерењата), треба и тие да бидат земени во предвид при формирањето на моделот. Општиот облик на овој модел наречен **модел на регресија** е

$$Y = f_t(x^{(1)}, ..., x^{(r)}) + \varepsilon, \tag{6.1}$$

каде функцијата f_t се нарекува **регресиона функција** и зависи од параметарот t кој многу често е векторски параметар, а ε е **случајна компонента** со $E(\varepsilon) = 0$ и $D(\varepsilon) = \sigma^2$.

Кога имаме само една независна променлива x, станува збор за **модел на еднодимензионална регресија** т.е.

$$Y = f_t(x) + \varepsilon, \tag{6.2}$$

а кога се повеќе од една независна променлива, имаме **модел на повеќедимензионална регресија**.

Изборот на соодветниот модел на регресија зависи од конкретната ситуација. Затоа, во случај на еднодимензионална регресија, најнапред се прикажуваат графички статистичките податоци $(x_1, y_1), ..., (x_n, y_n)$, каде $x_1, ..., x_n$ се вредности на независната променлива x, а $y_1, ..., y_n$ се соодветните вредности на случајната променлива Y добиени како резултат на набљудувања или мерења, а потоа се бара општиот облик и отстапувањата од овој облик, со цел да се најде математичкиот модел кој најдобро би го опишал обликот.

Основната задача на регресионата анализа е да после изборот на моделот, врз основа на низата статистички податоци, да ги процени непознатите параметри t и σ^2. Попрецизно кажано, во случај на еднодимензионална регресија, треба да се дефинираат оценувачи за непознатите параметри t и σ^2, врз основа на примерокот $(x_1, Y_1), ..., (x_n, Y_n)$, каде $x_1, ..., x_n$ се вредности на независната променлива x и $Y_1, ..., Y_n$ се соодветните случајни променливи дефинирани со

$$Y_i = f_t(x_i) + \varepsilon_i,\ i = 1, ..., n, \tag{6.3}$$

при што за случајните променливи $\varepsilon_1, ..., \varepsilon_n$ претпоставуваме дека се независни со $E(\varepsilon_i) = 0$ и $D(\varepsilon_i) = \sigma^2$, $i = 1, ..., n$.

6.1 Линеарна регресија

Ако регресионата функција f_t во (6.1) е линеарна, станува збор за **линеарен модел на регресија**. Ако кај линеарниот модел имаме само една независна променлива x, тогаш таквиот модел е **прост линеарен модел на регресија**, и тогаш $t = (a, b)$, па моделот е

$$Y = ax + b + \varepsilon, \tag{6.4}$$

каде a и b се параметри и ε е случајна променлива со $E(\varepsilon) = 0$ и $D(\varepsilon) = \sigma^2$. Да забележиме дека тогаш,

$$E(Y) = E(ax + b + \varepsilon) = ax + b + E(\varepsilon) = ax + b,$$

$$D(Y) = D(ax + b + \varepsilon) = D(\varepsilon) = \sigma^2,$$

затоа што $ax + b$ не е случајна променлива.

Смислата на моделот (6.4) е следната. Случајната променлива Y зависи од променливата x, на тој начин што за секој $x = x_i$, $Y_i = ax_i + b + \varepsilon_i$ има еден детерминистички собирок кој е линеарна функција од x_i т.е. $ax_i + b$ и

стохастички собирок ε_i кој претставува случајна осцилација околу детерминистичкиот собирок затоа што $E(\varepsilon_i) = 0$. При едно реализирано мерење, добиваме една реализација на примерокот $(x_1, Y_1), ..., (x_n, Y_n)$, која ја означуваме со $(x_1, y_1), ..., (x_n, y_n)$. Нејзиниот графички приказ има тенденција на групитање околу правата $y = ax + b$. Групирањето зависи од распределбата на $\varepsilon_1, ..., \varepsilon_n$, имено колку нивната дисперзија $D(\varepsilon_i) = \sigma^2$, $i = 1, ..., n$ е поголема, толку растурањето на тие точки околу правата $y = ax + b$ е поголемо.

6.1.1 Оценување на параметрите на регресија

Врз основа на примерокот $(x_1, Y_1), ..., (x_n, Y_n)$, треба да најдеме оценувачи $\widehat{a}$, $\widehat{b}$, $\widehat{\sigma}^2$ за непознатите параметри a, b, σ^2 на линеарниот модел на регресија (6.4). Без воведување на дополнителни претпоставки за распределбата на случајните променливи ε_i, со **методот на најмали квадрати** може да се најдат оценувачите $\widehat{a}$ и $\widehat{b}$.

Нека $(x_1, y_1), ..., (x_n, y_n)$ е реализација на примерокот $(x_1, Y_1), ..., (x_n, Y_n)$, тогаш за реализациите $\widehat{a}'$ и $\widehat{b}'$ на оценувачите $\widehat{a}$ и $\widehat{b}$, според методот на најмали квадрати треба да важи

$$\min_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{n} (y_i - ax_i - b)^2 = \sum_{i=1}^{n} (y_i - \widehat{a}'x_i - \widehat{b}')^2.$$

Па, слично како претходно (види Дескриптивна статистика, Дводимензионални обележја, оредување на коефициентите на правата на регресија), со решавање на системот

$$\frac{\partial S(a,b)}{\partial a} = 0, \ \frac{\partial S(a,b)}{\partial b} = 0,$$

каде

$$S(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2,$$

се добиваат решенијата (оценките на a и b добиени со методот на најмали квадрати)

$$\widehat{a}' = \frac{s_{xy}}{s_x^2}, \ \widehat{b}' = \overline{y} - \widehat{a}'\overline{x},$$

каде

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \ \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \ s_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2,$$

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}),$$

Ирена Стојковска

(покажи!). Ако означиме со

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

тогаш оценувачите добиени со метод на најмали квдрати, статистики кои зависат од примерокот $(x_1, Y_1), ..., (x_n, Y_n)$ се

$$\widehat{a} = \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}) = \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \overline{x})Y_i, \tag{6.5}$$

$$\widehat{b} = \overline{Y} - \widehat{a}\overline{x} = \frac{1}{n}\sum_{i=1}^{n}\Big(1 - \frac{\overline{x}}{s_x^2}(x_i - \overline{x})\Big)Y_i, \tag{6.6}$$

(покажи!). Исто така, според методот на најмали квадрати, како оценувач за σ^2 може да се земе

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - (\widehat{a}x_i + \widehat{b}))^2. \tag{6.7}$$

Бидејќи $E(Y_i) = ax_i + b$ и $D(Y_i) = \sigma^2$, за оценувачите $\widehat{a}$ и $\widehat{b}$ имаме

$$E(\widehat{a}) = E\Big(\frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \overline{x})Y_i\Big) = \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \overline{x})E(Y_i) = \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \overline{x})(ax_i + b) =$$

$$= \frac{1}{s_x^2}\Big(\frac{a}{n}\sum_{i=1}^{n}(x_i^2 - x_i\overline{x}) + \frac{b}{n}\sum_{i=1}^{n}(x_i - \overline{x})\Big) = \frac{1}{s_x^2}(as_x^2 + b\cdot 0) = a,$$

$$E(\widehat{b}) = E(\overline{Y} - \widehat{a}\overline{x}) = E(\overline{Y}) - E(\widehat{a})\overline{x} = \frac{1}{n}\sum_{i=1}^{n}E(Y_i) - a\overline{x} =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) - a\overline{x} = a\overline{x} + b - a\overline{x} = b,$$

што значи дека $\widehat{a}$ и $\widehat{b}$ се непристрасни оценувачи за a и b соодветно. Понатаму, од независноста на ε_i, $i = 1, ..., n$ следи и независноста на $Y_i = ax_i + b + \varepsilon_i$, $i = 1, ..., n$ (покажи!). Затоа, за дисперзиите на $\widehat{a}$ и $\widehat{b}$ имаме

$$D(\widehat{a}) = \frac{\sigma^2}{ns_x^2} \to 0,\ D(\widehat{b}) = \frac{\sigma^2}{n}\Big(1 + \frac{\overline{x}^2}{s_x^2}\Big) \to 0,$$

кога $n \to \infty$ (покажи!), од каде заклучуваме дека $\widehat{a}$ и $\widehat{b}$ се конзистентни оценувачи за a и b соодветно.

Ирена Стојковска

Ако сега ја разгледаме статистиката $\widehat{a}x+\widehat{b}$ како оценувач за $ax+b$, бидејќи

$$E(\widehat{a}x+\widehat{b})=ax+b,\ D(\widehat{a}x+\widehat{b})=\frac{\sigma^2}{n}\Big(1+\frac{(x-\overline{x})^2}{s_x^2}\Big)\to 0,$$

кога $n\to\infty$ (покажи!), заклучуваме дека $\widehat{a}x+\widehat{b}$ е непристрасен и конзистентен оценувач за $ax+b$.

Но, за оценувачот $\widehat{\sigma}^2$ дефиниран со (6.7) имаме

$$E(\widehat{\sigma}^2)=E\Big(\frac{1}{n}\sum_{i=1}^{n}(Y_i-(\widehat{a}x_i+\widehat{b}))^2\Big)=\frac{n-2}{n}\sigma^2$$

(покажи!), од каде заклучуваме дека тој не е непристрасен оценувач за σ^2. Затоа, неговата корекција, оценувачот

$$\breve{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}(Y_i-(\widehat{a}x_i+\widehat{b}))^2 \tag{6.8}$$

е непристрасен оценувач за σ^2.

За да ги користиме најдените оценувачи при барање на **интервали на доверба** и **тесирање на хипотези** за параметрите на регресија, потребно е да ги знаеме и нивните распределби, точни или асимптотски. Така, **во случај на голем примерок**, од централната гранична теорема, следи дека оценувачите $\widehat{a}$ и $\widehat{b}$ се асимптотски нормални оценувачи за a и b соодветно, односно

$$\widehat{a}\xrightarrow{dist}\mathcal{N}\Big(a,\ \frac{\sigma^2}{ns_x^2}\Big),\ \widehat{b}\xrightarrow{dist}\mathcal{N}\Big(b,\ \frac{\sigma^2}{n}\Big(1+\frac{\overline{x}^2}{s_x^2}\Big)\Big).$$

Истото важи и за оценувачот $\widehat{a}x+\widehat{b}$, за кој се покажува дека е асимптоцки нормален оценувач за $ax+b$, односно

$$\widehat{a}x+\widehat{b}\xrightarrow{dist}\mathcal{N}\Big(ax+b,\ \frac{\sigma^2}{n}\Big(1+\frac{(x-\overline{x})^2}{s_x^2}\Big)\Big).$$

Па, бри барање на интервали на доверба и тестирање на хипотези за a, b и $ax+b$ врз основа на голем примерок, се користиме со гореспоменатите распределби, при што непознатиот параметар σ^2 го заменуваме со реализација на оценувачот дефиниран со (6.7) или (6.8), што во пракса нема големо значење која оценка ќе се одбере затоа што за големи вредности на n разликата меѓу двете оценки е незначителна.

За да ги најдеме точните распределби на оценувачите $\widehat{a}$ и $\widehat{b}$, **во случај на мал примерок** потребни се дополнителни претпоставки за случајните променливи ε_i, $i=1,...,n$ во линеарниот модел на регресија $Y_i=ax_i+b+\varepsilon_i$, $i=1,...,n$.

Ирена Стојковска

Имено, покрај независноста се претпоставува и дека $\varepsilon_i \sim \mathcal{N}(0,\ \sigma^2),\ i = 1, ..., n$. Од тука следи дека и случајните променливи $Y_i,\ i = 1, ..., n$ се нзависни и за нивните распределби важи

$$Y_i \sim \mathcal{N}(ax_i + b,\ \sigma^2),\ i = 1, ..., n.$$

Па тогаш, за точните распределби на оценувачите $\widehat{a}$ и $\widehat{b}$ дефинирани со (6.5) и (6.6) соодветно, имаме

$$\widehat{a} \sim \mathcal{N}\Big(a,\ \frac{\sigma^2}{ns_x^2}\Big),\ \widehat{b} \sim \mathcal{N}\Big(b,\ \frac{\sigma^2}{n}\Big(1 + \frac{\overline{x}^2}{s_x^2}\Big)\Big),$$

бидејќи претставуваат линеарна комбинација од случајните променливи Y_i, $i = 1, ..., n$. Во овој случај, случајната променлива $\widehat{\sigma}^2$ дефинирана со (6.7), а исто така и корегираниот оценувач $\breve{\sigma}^2$ дефиниран со (6.8), претставуваат збир од квадрати на одредени линеарни комбинации од нормално распределени случајни променливи Y_i, $\widehat{a}$ и $\widehat{b}$, кои не се независни. Но, се покажува дека, и во тој случај, распределбата на така претставената случајна променлива е хи-квадрат распределба. Имено, важи

$$\frac{n}{\sigma^2}\,\widehat{\sigma}^2 \sim \chi^2_{n-2},\ \text{односно}\ \frac{n-2}{\sigma^2}\,\breve{\sigma}^2 \sim \chi^2_{n-2}.$$

Исто така, се покажува дека $\widehat{a}$ и $\widehat{\sigma}^2$, но и $\widehat{b}$ и $\widehat{\sigma}^2$ се независни, што помага при одредување на точните распределби на статистиките кои се користат при наоѓање на интервалите на доверба и тестирање на хипотезите за параметрите a, b и $ax + b$, имено важи

$$T_1 = \frac{(\widehat{a} - a)\sqrt{(n-2)s_x^2}}{\widehat{\sigma}} = \frac{(\widehat{a} - a)\sqrt{ns_x^2}}{\breve{\sigma}} \sim t_{n-2},$$

$$T_2 = \frac{(\widehat{b} - b)\sqrt{(n-2)s_x^2}}{\widehat{\sigma}\sqrt{s_x^2 + \overline{x}^2}} = \frac{(\widehat{b} - b)\sqrt{ns_x^2}}{\breve{\sigma}\sqrt{s_x^2 + \overline{x}^2}} \sim t_{n-2},$$

$$T_3 = \frac{(\widehat{a}x + \widehat{b} - ax - b)\sqrt{(n-2)s_x^2}}{\widehat{\sigma}\sqrt{s_x^2 + (x - \overline{x})^2}} = \frac{(\widehat{a}x + \widehat{b} - ax - b)\sqrt{ns_x^2}}{\breve{\sigma}\sqrt{s_x^2 + (x - \overline{x})^2}} \sim t_{n-2}$$

(покажи!).

Оценувачи за параметрите на регресија a, b и σ^2 може да се најдат и со **метод на максимална подобност**. Но, и во тој случај потребни се дополнителните претпоставки за нормална распределеност на случајните променливи ε_i, $i = 1, ..., n$ во линеарниот модел на регресија $Y_i = ax_i + b + \varepsilon_i$, $i = 1, ..., n$, имено покрај независноста се претпоставува и дека $\varepsilon_i \sim \mathcal{N}(0,\ \sigma^2),\ i = 1, ..., n$.

Ирена Стојковска

Го означуваме со $t = (a, b, \sigma^2)$ векторот од непознати параметри кој припаѓа на просторот од параметри

$$\Theta = \{(a, b, \sigma^2) \mid a \in \mathbb{R},\ b \in \mathbb{R},\ \sigma^2 > 0\}.$$

Тогаш, функцијата на подобност е

$$L(x, y, t) = (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - ax_i - b)^2\}.$$

Со решавање на системот равенки на подобност

$$\frac{\partial \ln L(x, y, t)}{\partial a} = 0,\ \frac{\partial \ln L(x, y, t)}{\partial b} = 0,\ \frac{\partial \ln L(x, y, t)}{\partial \sigma^2} = 0,$$

по a, b и σ^2 се добиваат решенијата

$$a = \frac{s_{xy}}{s_x^2},\ b = \overline{y} - a\overline{x},\ \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (ax_i + b - y_i)^2$$

(покажи!), од каде заклучуваме дека максимално подобни (ML) оценувачи за a, b и σ^2, се $\widehat{a}$, $\widehat{b}$ и $\widehat{\sigma}^2$, дефинирани со (6.5), (6.6) и (6.7) соодветно.

Прилог А

Делта метод. *Нека важи*

$$a_n(X_n - \theta) \xrightarrow{dist.} Z,$$

каде θ *е константа,* $\{a_n\}$ *е низа од реални броеви така што* $\lim_{n\to+\infty} = +\infty$ *и* Z *е случајна променлива. Ако* $g(x)$ *е непрекинато диференцијабилна функција во* $x = \theta$*, тогаш*

$$a_n(g(X_n) - g(\theta)) \xrightarrow{dist.} g'(\theta)Z.$$

Доказ. Од $a_n(X_n - \theta) \xrightarrow{dist.} Z$ се добива дека $X_n \xrightarrow{P} \theta$. Имено за $\varepsilon > 0$,

$$P\{|X_n - \theta| < \varepsilon\} = P\{-a_n\varepsilon < a_n(X_n - \theta) < a_n\varepsilon\} \longrightarrow F_Z(+\infty) - F_Z(-\infty) = 1,$$

кога $n \to +\infty$. Понатаму, од услов $g(x)$ е непрекината функција, па нејзиниот Тајлоров развој во $x = \theta$ е

$$g(X_n) = g(\theta) + g'(\theta_n^*)(X_n - \theta),$$

каде θ_n^* е меѓу X_n и θ, односно важи $|\theta_n^* - \theta| \leq |X_n - \theta|$. Затоа, од $X_n \xrightarrow{P} \theta$, следи дека $\theta_n^* \xrightarrow{P} \theta$, и од $g'(x)$ непрекината во $x = \theta$ селеди дека $g'(\theta_n^*) \xrightarrow{P} g(\theta)$. Сега,

$$a_n(g(X_n) - g(\theta)) = g'(\theta_n^*)a_n(X_n - \theta) \xrightarrow{dist.} g'(\theta)Z,$$

според теоремата на Slutsky за $X_n = a_n(X_n - \theta)$ и $Y_n = g'(\theta_n^*)$. **Забелешка.** Тврдењето важи и во поопшт случај, кога $g(x)$ не мора да е непрекинато диференцијабилна функција. ∎

Теорема на Slutsky. *Нека* $X_n \xrightarrow{dist.} X$ *и* $Y_n \xrightarrow{P} \theta$*, каде* θ *е константа. Тогаш,*

(i) $X_n + Y_n \xrightarrow{dist.} X + \theta$,

(ii) $X_nY_n \xrightarrow{dist.} \theta X$,

(ii) $\frac{X_n}{Y_n} \xrightarrow{dist.} \frac{X}{\theta}$.

Литература

[1] Р. Малчески, *Статистика за бизнис,* Факултет за општествени науки, Скопје (2006)

[2] П. Младеновић, *Вероватноћа и статистика,* Математички факултет, Београд (2005)

[3] Ж. Попеска, *Статистика - предавања,* Институт за информатика, Природно-математички факултет, Скопје

[4] И. Стојковска, *Основи на статистика - вежби,* Институт за математика, Природно-математички факултет, Скопје

[5] Z. A. Ivković, *Matematička statistika,* Naučna knjiga, Beograd (1975)

[6] K. Knight, *Mathematical statistics,* Chapman & Hall/CRC, Boca Raton, Florida (2000)

[7] Ž. Pauše, *Uvod u matematičku statistiku,* Školska knjiga, Zagreb (1993)