Prírodovedecká fakulta UPJŠ Košice

Pravdepodobnosť a štatistika

(**poznámky z prednášok letného semestra predmetu Pravdepodobnosť a štatistika**)

prednáša: RNDr. Valéria Skřivánková, CSc.

Verzia 11. júla 2003 12:11
Zostavil Róbert Novotný
`http://skmi.science.upjs.sk/~novotnyr`
Typeset by LaTeX. Illustrations by jPicEdt.
Function plots by gnuplot.

Obsah

(a) ak $\sigma_1^2 = \sigma_2^2 = \sigma^2$ je neznáme Potom testovým kritériom je

$$g = \frac{\frac{(\bar{\mathbb{X}}_1-\bar{\mathbb{X}}_2)-(a_1-a_2)}{\sqrt{n_2 S_{1_1}^2 + n_1 S_{1_2}^2}} \cdot \sqrt{n_1 n_2}}{\sqrt{\frac{\frac{(n_1-1)S_{1_1}^2}{\sigma^2} + \frac{(n_2-1)S_{1_2}^2}{\sigma^2}}{(n_1-1)+(n_2-1)}}} = \frac{(\bar{\mathbb{X}}_1-\bar{\mathbb{X}}_2)-(a_1-a_2)}{\sqrt{(n_1-1)S_{1_1}^2 + (n_2-1)S_{1_2}^2}} \cdot \sqrt{n_1 n_2},$$

pričom uvedené $g \sim t(n_1+n_2-2)$.

(b) ak $\sigma_1^2 \neq \sigma_2^2$. A opäť rozlíšime dva prípady:

i. Ak $n_1 > 30$ a $n_2 > 30$, môžeme t-rozdelenie aproximovať normovaným normálnym rozdelením. Testové kritérium a kritický obor bude:

$$\begin{aligned} g &= \frac{(\bar{\mathbb{X}}_1-\bar{\mathbb{X}}_2)-(a_1-a_2)}{\sqrt{n_2 S_{1_1}^2 + n_1 S_{1_2}^2}} \cdot \sqrt{n_1 n_2} \sim \mathsf{N}(0,1) \\ W_0 &= \{\boldsymbol{x}_1, \boldsymbol{x}_2 : |g| \geq u_{1-\frac{\alpha}{2}}\} \end{aligned}$$

ii. Ak $n_1 \leq 30$, $n_2 \leq 30$, použijeme testové kritérium:

$$g = \frac{(\bar{\mathbb{X}}_1-\bar{\mathbb{X}}_2)-(a_1-a_2)}{\sqrt{n_2 S_{1_1}^2 + n_1 S_{1_2}^2}} \cdot \sqrt{n_1 n_2} \sim t(\gamma).$$

kde γ je „neznámy“ počet stupňov voľnosti t-rozdelenia. Určíme ho nasledovne:

$$t_\alpha(\gamma) = \frac{t_\alpha(n_1-1)\cdot\frac{S_{1_1}^2}{n_1} + t_\alpha(n_2-1)\cdot\frac{S_{1_2}^2}{n_2}}{\frac{S_{1_1}^2}{n_1}+\frac{S_{1_2}^2}{n_2}}$$

Kritický obor:

$$W_0 = \{\boldsymbol{x}_1, \boldsymbol{x}_2 : |g| \geq u_{1-\frac{\alpha}{2}}\}$$

9.3.2 Testy zhody dvoch rozptylov

V tomto prípade ide o tzv. *F-testy*. Hypotézy položíme nasledovne:

$$H_0 : \sigma_1^2 = \sigma_2^2, \qquad H_1 : \sigma_1^2 \neq \sigma_2^2$$

1. ak a_1, a_2 sú známe, použijeme nasledovné testové kritérium a kritický obor:

$$\begin{aligned} g &= \frac{\frac{n_1 S_{0_1}^2}{\sigma_1^2}}{\frac{n_2 S_{0_2}^2}{\sigma_2^2}} = \frac{S_{0_1}^2\sigma_2^2}{S_{0_2}^2\sigma_1^2} \sim F(n_1,n_2) \\ W_0 &= \{\boldsymbol{x}_1, \boldsymbol{x}_2 : g \leq F_{\alpha/2}(n_1,n_2) \vee g \geq F_{1-\alpha/2}(n_1,n_2)\} \end{aligned}$$

2. ak a_1, a_2 sú neznáme, použijeme toto testové kritérium a kritický obor:

$$\begin{aligned} g &= \frac{\frac{(n_1-1)S_{1_1}^2}{\sigma_1^2}}{\frac{(n_2-1)S_{1_2}^2}{\sigma_2^2}} = \frac{S_{1_1}^2\sigma_2^2}{S_{1_2}^2\sigma_1^2} \sim F(n_1-1,n_2-1) \\ W_0 &= \{\boldsymbol{x}_1, \boldsymbol{x}_2 : g \leq F_{\alpha/2}(n_1-1,n_2-1) \vee g \geq F_{1-\alpha/2}(n_1-1,n_2-1)\} \end{aligned}$$

Pri testovaní zhody dvoch stredných hodnôt pre σ_1^2, σ_2^2 neznáme musíme najprv otestovať zhodu dvoch rozptylov!

2. Nech $V_n \in \mathsf{Ex}(\delta)$. Budeme testovať parameter δ. Hypotézy položíme nasledovne:

$$H_0 : \delta = \delta_0, \qquad H_1 : \delta = \delta_1 < \delta_0, \quad \text{alebo} \quad H_1 : \delta_1 > \delta_0, \quad \text{alebo} \quad H_1 : \delta_1 \neq \delta_0$$

Potom použitá štatistika a príslušné kritické obory budú:

$$g = \frac{2n\bar{\mathbb{X}}}{\delta} \sim \chi^2(2n) \qquad W_0 = \left\{\mathbb{X} : \frac{2n\bar{\mathbb{X}}}{\delta} \leq \chi^2_\alpha(2n)\right\}, \quad \text{ak } \delta < \delta_0^2$$

$$W_0 = \left\{\mathbb{X} : g \geq \chi^2_{1-\alpha}(2n)\right\}, \quad \text{ak } \delta^2 > \delta_0^2$$

$$W_0 = \left\{\mathbb{X} : g \leq \chi^2_{\frac{\alpha}{2}}(2n) \vee g \geq \chi^2_{1-\frac{\alpha}{2}}(2n)\right\}, \quad \text{ak } \delta^2 \neq \delta_0^2$$

Tu je potrebné dopísať nasledovné časti:

- Odvodenie najlepšieho kritického oboru W_0
- Neparametrické testy (znamienkový test, Dixonov test)
- Párový t-test

9.3 Testy zhody pre dva nezávislé výbery

Uvažujme dva nezávislé výbery z normálneho rozdelenia.

$$V_{n_1} = (\mathbb{X}_{11}, \mathbb{X}_{12}, \ldots, \mathbb{X}_{1n_1}) \in \mathsf{N}\left(a_1, \sigma_1^2\right), \bar{\mathbb{X}}_1 \sim \mathsf{N}\left(a_1, \tfrac{\sigma_1^2}{n_1}\right)$$

$$V_{n_2} = (\mathbb{X}_{21}, \mathbb{X}_{22}, \ldots, \mathbb{X}_{2n_2}) \in \mathsf{N}\left(a_2, \sigma_2^2\right), \bar{\mathbb{X}}_2 \sim \mathsf{N}\left(a_2, \tfrac{\sigma_2^2}{n_2}\right)$$

Potom $\bar{\mathbb{X}}_1 - \bar{\mathbb{X}}_2 \sim \mathsf{N}\left(a_1 - a_2, \frac{\sigma_1^2}{n_1} - \frac{\sigma_2^2}{n_2}\right)$. (kovarianciu poznáme, je rovná 0, pretože výbery sú nezávislé).

Rozlišujeme dva typy testovanej zhody:

A.) test zhody dvoch stredných hodnôt

B.) test zhody dvoch rozptylov

9.3.1 Testy zhody dvoch stredných hodnôt

Hypotézy položíme nasledovne:

$$H_0 : a_1 = a_2, \qquad H_1 : a_1 \neq a_2$$

Rozlišujeme dva prípady:

1. ak σ_1^2, σ_2^2 sú známe – ide o *u-test*.

   Testovým kritériom a kritickým oborom sú:

$$g = \frac{(\bar{\mathbb{X}}_1 - \bar{\mathbb{X}}_2) - (a_1 - a_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{\mathbb{X}}_1 - \bar{\mathbb{X}}_2) - (a_1 - a_2)}{\sqrt{n_2\sigma_1^2 + n_1\sigma_2^2}} \cdot \sqrt{n_1 n_2} \sim \mathsf{N}(0,1)$$

$$W_0 = \{\boldsymbol{x}_1, \boldsymbol{x}_2 : |g| \geq u_{1-\frac{\alpha}{2}}\}$$

2. ak σ_1^2, σ_2^2 sú známe – ide o *t-testy*. Opäť máme dve možnosti:

Literatúra

[1] Riečan a kol.: Pravdepodobnosť a matematická štatistika, Bratislava 1984

[2] Potocký a kol.: Zbierka úloh z pravdepodobnosti a matematickej štatistiky, Bratislava 1986

[3] Skřivánková: Pravdepodobnosť v príkladoch, Košice 1999

[4] Anděl: Matematika náhody, Praha 2000

Tento materiál pokrýva látku z letného semestra predmetu Pravdepodobnosť a štatistika, ktorý prednáša RNDr. Valéria Skřivánková na Prírodovedeckej fakulte UPJŠ v Košiciach. Jeho obsahom sú definície, vety a dôkazy, ktoré odzneli na prednáškach v akademickom roku 2002/2003.

Materiál bol vytvorený výhradne pre internú potrebu študentov PrírF UPJŠ Košice.

Text nebol autorizovaný a môže obsahovať chyby, preklepy, či chýbajúce časti (budem však rád, keď ich oznámite na adrese `novotnyr@skmi.science.upjs.sk`). Na tento materiál sa nevzťahuje žiadna záruka.

Zostavené dňa 11. júla 2003, 12:11.
Zostavil Róbert Novotný, `http://skmi.science.upjs.sk/~novotnyr`

4 Niektoré špeciálne rozdelenia

4.1 Niektoré diskrétne typy rozdelení

4.1.1 Binomické rozdelenie $\mathsf{Bi}(n,p)$

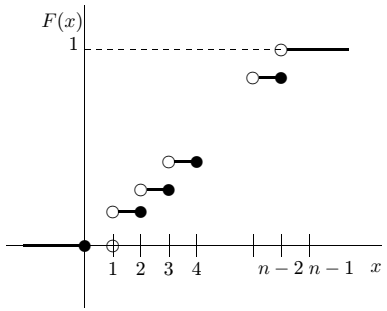
Definícia 4.1
Hovoríme, že diskrétna náhodná veličina $\mathbb{X}$ má *binomické rozdelenie* s parametrami n, p, ak nadobúda hodnoty $x_k = k$, $k = 0, 1, \ldots, n$ s pravdepodobnosťami

$$p_k = P(\mathbb{X} = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad p \in (0,1), n \in \mathbb{N}, k = 0, 1, \ldots, n$$

Interpretácia binomického rozdelenia. Bernoulliho schéma – realizujeme n nezávislých pokusov s možnými výsledkami ω, $\omega^{\complement}$, pričom $P(\{\omega\}) = p$, $p \in (0,1)$. Priraďme situácii „nastal jav ω“ hodnotu 1 a „nenastal jav ω“ hodnotu 0. Potom náhodná veličina $\mathbb{X}$ majúca binomické rozdelenie s parametrami n, p reprezentuje počet úspešných pokusov z n pokusov.

1. distribučná funkcia
$$F(x) = \sum_{k<x} \binom{n}{k} p^k (1-p)^{n-k}$$

2. charakteristická funkcia
$$\varphi(t) = \sum_{k=0}^{n} \mathrm{e}^{\mathrm{i}tk} \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=0}^{n} \binom{n}{k} (p\mathrm{e}^{\mathrm{i}t})^k \cdot (1-p)^{n-k} = (p\mathrm{e}^{\mathrm{i}t} + 1 - p)^n$$



Obr. 1: Distribučná funkcia binomického rozdelenia

3. charakteristiky polohy a variability
$$\begin{aligned} E(\mathbb{X}) &= np \\ D(\mathbb{X}) &= np \cdot (1-p) \end{aligned}$$

Dôkaz:

$$\begin{aligned} \varphi'(t) &= n(p\mathrm{e}^{\mathrm{i}t} + 1 - p)^{n-1} \cdot \mathrm{i}p\mathrm{e}^{\mathrm{i}t} \\ \varphi'(0) &= \mathrm{i} \cdot np = \mathrm{i} \cdot m_1 \Rightarrow m_1 = E(\mathbb{X}) = np \\ \varphi''(t) &= inp\left[(n-1)(p\mathrm{e}^{\mathrm{i}t} + 1 - p)^{n-2}) \cdot p\mathrm{i}\mathrm{e}^{\mathrm{i}t} \cdot e^{\mathrm{i}t} + (p\mathrm{e}^{\mathrm{i}t} + 1 - p)^{n-1} \cdot i\mathrm{e}^{\mathrm{i}t}\right] \end{aligned}$$

9.2.2 Príklady kritických oborov W_0 pre normálne a exponenciálne rozdelenie

1. Nech $V_n \in N(a, \sigma^2)$.
   (a) test parametra a, ak σ^2 je známe (v ďalšom bude používaná „nekanonická“ prologovská notácia $\texttt{?-}a|\sigma^2$ známe — pozn. sadzač)
   (b) test parametra a, ak σ^2 je neznáme; $\texttt{?-}a|\sigma^2$ neznáme
   (c) test parametra σ^2, ak a je známe; $\texttt{?-}\sigma^2|a$ známe
   (d) test parametra σ^2, ak a je neznáme; $\texttt{?-}\sigma^2|a$ neznáme

(a) $\texttt{?-}a|\sigma^2$ známe

$$H_0 : a = a_0, \qquad H_1 : a = a_1 < a_0 \text{ alebo } H_1 : a_0 > a_1 = a_1 \text{ alebo } H_1 : a \neq a_0$$

σ^2 je známe, preto použijeme štatistiku

$$g = \frac{\bar{\mathbb{X}} - a}{\sigma} \cdot \sqrt{n} \sim \mathsf{N}(0,1)$$

Kritické obory budú

$$\begin{aligned} W_0 &= \left\{\mathbb{X} : \frac{\bar{\mathbb{X}} - a}{\sigma} \cdot \sqrt{n} \leq u_\alpha\right\}, \text{ ak } a < a_0 \\ W_0 &= \left\{\mathbb{X} : \frac{\bar{\mathbb{X}} - a}{\sigma} \cdot \sqrt{n} \geq u_{1-\alpha}\right\}, \text{ ak } a > a_0 \\ W_0 &= \left\{\mathbb{X} : \left|\frac{\bar{\mathbb{X}} - a}{\sigma} \cdot \sqrt{n}\right| \geq u_{1-\frac{\alpha}{2}}\right\}, \text{ ak } a \neq a_0 \end{aligned}$$

(b) $\texttt{?-}a|\sigma^2$ neznáme

$$g = \frac{\bar{\mathbb{X}} - a}{S_1} \cdot \sqrt{n} \sim t(n-1)$$

$$\begin{aligned} W_0 &= \{\mathbb{X} : g \leq t_\alpha(n-1)\}, && \text{ak } a < a_0 \\ W_0 &= \{\mathbb{X} : g \geq t_{1-\alpha}(n-1)\}, && \text{ak } a > a_0 \\ W_0 &= \left\{\mathbb{X} : |g| \geq t_{1-\frac{\alpha}{2}}(n-1)\right\}, && \text{ak } a \neq a_0 \end{aligned}$$

(c) $\texttt{?-}\sigma^2|a$ známe

$$g = \frac{nS_0^2}{\sigma^2} \sim \chi^2(n)$$

$$\begin{aligned} W_0 &= \left\{\mathbb{X} : g \leq \chi^2_\alpha(n)\right\}, && \text{ak } \sigma^2 < \sigma_0^2 \\ W_0 &= \left\{\mathbb{X} : g \geq \chi^2_{1-\alpha}(n)\right\}, && \text{ak } \sigma^2 > \sigma_0^2 \\ W_0 &= \left\{\mathbb{X} : g \leq \chi^2_{\frac{\alpha}{2}}(n) \vee g \geq \chi^2_{1-\frac{\alpha}{2}}(n)\right\}, && \text{ak } \sigma^2 \neq \sigma_0^2 \end{aligned}$$

(d) $\texttt{?-}\sigma^2|a$ neznáme

$$g = \frac{(n-1)\,S_1^2}{\sigma^2} \sim \chi^2(n-1)$$

$$\begin{aligned} W_0 &= \left\{\mathbb{X} : g \leq \chi^2_\alpha(n-1)\right\}, && \text{ak } \sigma^2 < \sigma_0^2 \\ W_0 &= \left\{\mathbb{X} : g \geq \chi^2_{1-\alpha}(n-1)\right\}, && \text{ak } \sigma^2 > \sigma_0^2 \\ W_0 &= \left\{\mathbb{X} : g \leq \chi^2_{\frac{\alpha}{2}}(n-1) \vee g \geq \chi^2_{1-\frac{\alpha}{2}}(n-1)\right\}, && \text{ak } \sigma^2 \neq \sigma_0^2 \end{aligned}$$

4. zistiť, či realizácia testovacieho kritéria g je z W_0 a urobiť záver pre prax
   (a) ak $g_0 \in W_0$, potom H_0 zamietame a prijímame H_1
   (b) ak $g_0 \notin W_0$, potom H_0 nezamietam na danej hladine významnosti (slabší záver). Môžeme urobiť nový výber alebo zmeníme hladinu významnosti.

Delenie testov

1. podľa toho, čo testujú
   (a) parametrické testy
   (b) neparametrické testy
2. podľa počtu zrealizovaných výberov
   (a) jednovýberové
   (b) dvoj- a viacvýberové

9.2 Niektoré parametrické testy (jednovýberové)

Budeme predpokladať, že výber pochádza buď z normálneho alebo exponenciálneho rozdelenia, teda $V_n \in \mathsf{N}(a, \sigma^2)$, resp. $V_n \in \mathsf{Ex}(\delta)$.

Parametrický test testuje neznámy parameter θ rozdelenia $F(x, \theta)$ pre $\theta \in \Theta$, z ktorého výber V_n pochádza. Uvažujme hypotézu

$$H_0 : \theta = \theta_0,$$

kde θ_0 je skutočná hodnota parametra (alebo hodnota, o ktorej si myslím, že je skutočná :-)). Potom pre H_1 prichádza do úvahy jedna z nasledovných hypotéz:

$$H_1 : \theta \neq \theta_0, \qquad H_1 : \theta > \theta_0, \qquad H_1 : \theta < \theta_0$$

Prvá z uvedených hypotéz je tzv. *obojstranná alternatívna hypotéza*, zvyšné dve sú *jednostranné alternatívne hypotézy*. Hypotéza H_0 je tu nazývaná *jednoduchou hypotézou*, ostatné zase *zloženými alternatívnymi hypotézami*.

9.2.1 Metódy hľadania najlepšieho kritického oboru W_0

1. *Neymann-Pearsson* – pri jednostranných alternatívnych hypotézach
2. *test podielom (pomerom) vierohodností* – pri obojstranných alternatívnych hypotézach

Neymannova-Pearssonova metóda. Za kritický obor vezmeme:

$$W_0 = \left\{ \mathbb{X} : \frac{L(\mathbb{X}, \theta_1)}{L(\mathbb{X}, \theta_0)} \geq c(\alpha) \right\},$$

kde θ_0 je skutočná hodnota parametra θ, θ_1 je hodnota θ v alternatívnej hypotéze a $c(\alpha)$ je konštanta závislá iba od hladiny významnosti (malo by byť $c(\alpha) \geq 0$, inak nie je čo odhadovať). V dôkaze sa ukáže, že W_0 v tomto tvare zaručí minimálne β.

Podiel vierohodností. Kritický obor položíme:

$$W_0 = \left\{ \mathbb{X} : \frac{L(\mathbb{X}, \widehat{\theta})}{L(\mathbb{X}, \theta_0)} \geq L^*(\alpha) \right\},$$

kde $\widehat{\theta}$ je maximálny vierohodný odhad parametra θ, θ_0 je skutočná hodnota θ a $L^*(\alpha)$ je konštanta závislá iba od hladiny významnosti. $L^*(\alpha) \geq 1$, lebo podľa definície $L(\mathbb{X}, \bar{\theta}) > L(\mathbb{X}, \theta_0)$.

$$\begin{aligned} \varphi''(0) &= i^2 np((n-1)p+1) \\ E(\mathbb{X}^2) &= n(n-1)p^2 + np \\ D(\mathbb{X}) &= n^2p^2 - np^2 - np - (n^2p^2) = np(1-p) \end{aligned}$$

□

4.1.2 Poissonovo rozdelenie $\mathsf{Po}(\lambda)$

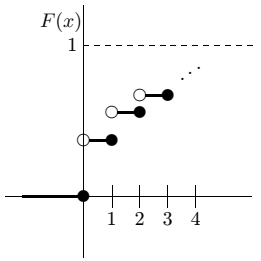
Definícia 4.2
Hovoríme, že diskrétna náhodná veličina $\mathbb{X}$ má *Poissonovo rozdelenie* s parametrom λ, ak nadobúda hodnoty x_k, $k = 0, 1, \ldots$ s pravdepodobnosťami

$$p_k = P(\mathbb{X} = k) = \frac{\lambda^k}{k!} \cdot \mathrm{e}^{-\lambda}, \quad \lambda > 0, k = 0, 1, 2, \ldots$$

Interpretácia Poissonovho rozdelenia. Náhodná veličina $\mathbb{X}$ majúca Poissonovo rozdelenie s parametrom λ reprezentuje počet prípadov, v ktorých nastal sledovaný jav pri neobmedzenej realizácii daného pokusu za jednotku času. Napr. počet zákazníkov v obchode za časovú jednotku.

1. distribučná funkcia

$$F(x) = \sum_{k<x} \frac{\lambda^k}{k!} \cdot \mathrm{e}^{-\lambda} = \mathrm{e}^{-\lambda} \cdot \sum_{k<x} \frac{\lambda^k}{k!}$$



Obr. 2: Distribučná funkcia Poissonovho rozdelenia. Pre x idúce do nekonečna sa bude „výška schodíkov“ zmenšovať, úroveň 1 sa nedosiahne v žiadnom konečnom bode.

2. charakteristická funkcia

$$\varphi(t) = \sum_{k=0}^{\infty} \mathrm{e}^{itk} \frac{\lambda^k}{k!} \cdot \mathrm{e}^{-\lambda} = \mathrm{e}^{-\lambda} \cdot \sum_{k=0}^{\infty} \left(\frac{\lambda \mathrm{e}^{it}}{k!} \right)^k$$

   Posledná suma je vlastne Taylorov rozvoj výrazu $\mathrm{e}^{(\cdot)}$. Teda máme

$$\varphi(t) = \mathrm{e}^{-\lambda} \cdot \mathrm{e}^{\lambda \mathrm{e}^{it}} = \mathrm{e}^{\lambda(\mathrm{e}^{it}-1)}$$

3. charakteristiky polohy a variability

$$\begin{aligned} E(\mathbb{X}) &= \lambda \\ D(\mathbb{X}) &= \lambda \end{aligned}$$

Dôkaz:

$$\begin{aligned}
\varphi'(t) &= \mathrm{e}^{\lambda(\mathrm{e}^{it}-1)} \cdot \lambda i \mathrm{e}^{it} \\
\varphi'(0) &= i \cdot \lambda = i \cdot m_1 \Rightarrow E(\mathbb{X}) = \lambda \\
\varphi''(t) &= i\lambda \left(\mathrm{e}^{\lambda(\mathrm{e}^{it}-1)} \cdot \lambda i \mathrm{e}^{it} \cdot \mathrm{e}^{it} + \mathrm{e}^{\lambda(\mathrm{e}^{it}-1)} \cdot i \mathrm{e}^{it} \right) \\
\varphi''(0) &= i^2 \lambda(\lambda+1) \Rightarrow E(\mathbb{X}^2) = \lambda(\lambda+1) \\
D(\mathbb{X}) &= \lambda^2 + \lambda - \lambda^2 = \lambda
\end{aligned}$$

□

4.1.3 Geometrické rozdelenie $\mathsf{Geo}(p)$

Definícia 4.3
Hovoríme, že diskrétna náhodná veličina $\mathbb{X}$ má *geometrické rozdelenie* s parametrom p, ak nadobúda hodnoty $x_k = k$, $k = 0, 1, \ldots$ s pravdepodobnosťami

$$p_k = P(\mathbb{X} = k) = p \cdot (1-p)^k, \quad p \in (0,1), k = 0, 1, \ldots$$

Interpretácia geometrického rozdelenia. Náhodná veličina $\mathbb{X}$ majúca geometrické rozdelenie s parametrom p vyjadruje počet „neúspechov“ pred prvým úspechom pri neobmedzenej realizácii pokusov v Bernoulliho schéme.

1. distribučná funkcia
$$F(x) = p \cdot \sum_{k<x} (1-p)^k$$

2. charakteristická funkcia
$$\varphi(x) = \sum_{k=0}^{\infty} \mathrm{e}^{itk} \cdot p(1-p)^k = p \cdot \sum_{k=0}^{\infty} \left((1-p) \cdot \mathrm{e}^{it} \right)^k$$
Pozrime sa na členy. Z definície je $p \in (0,1)$, teda aj $(1-p) \in (0,1)$. Už sme si ukázali, že $\mathrm{e}^{it} \leq 1$. Teda aj ich súčin je v absolútnej hodnote menší ako 1. To ale znamená, že suma je konvergentný geometrický rad. Teda.
$$\varphi(x) = \frac{p}{1-(1-p)\mathrm{e}^{it}}$$

3. charakteristika polohy a variability
$$\begin{aligned}
E(\mathbb{X}) &= \frac{1-p}{p} \\
D(\mathbb{X}) &= \frac{1-p}{p^2}
\end{aligned}$$

Dôkaz: Dôkaz nie je náročný, prenechávame ho čitateľovi. □

4.2 Niektoré spojité typy rozdelení

4.2.1 Rovnomerné rozdelenie $\mathsf{R}(a,b)$

Definícia 4.4
Hovoríme, že spojitá náhodná veličina $\mathbb{X}$ má *rovnomerné rozdelenie* na intervale (a,b), ak má hustotu

$$f(x) \begin{cases} \dfrac{1}{b-a} & \text{pre } x \in (a,b) \\ 0 & \text{inak} \end{cases} \qquad a < b;\ a, b \in \mathbb{R}$$

3. Opäť vyjdeme z obojstranného intervalu spoľahlivosti, ktorý si však predstavíme v tvare
$$\bar{\mathbb{X}} - \Delta < a < \bar{\mathbb{X}} + \Delta, \qquad \Delta = \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}$$
Podľa zadania chceme, aby $\Delta < 1$ a hľadáme n. Po dosadení a výpočte získame $n > 96$.

9 Testovanie štatistických hypotéz

9.1 Základné pojmy a metódy

Štatistická hypotéza H – každý predpoklad týkajúci sa rozdelenia $F(x,\theta)$, z ktorého V_n pochádza.
Testovanie štatistických hypotéz – overovanie správnosti nášho predpokladu
Testovacie kritérium $g(\mathbb{X}_1, \ldots, \mathbb{X}_n)$ je vhodne zvolená štatistika
Kritický obor W je tá časť množiny všetkých realizácií náhodného výberu V_n, ktoré vedie k zamietnutiu testovanej hypotézy.
Kritická hodnota k_α – tá hodnota, ktorá delí množiny všetkých realizácií V_n na kritickú oblasť a jej doplnok.
Nulová hypotéza H_0 – testovaná hypotéza.
Alternatívna hypotéza H_1 – hypotéza, ktorú staviame proti nulovej hypotéze (H_1 nemusí byť doplnok H_0[9]).
Chyby pri testovaní štatistických hypotéz

1. chyba 1. druhu – testovanú hypotézu H_0 zamietame, hoci je správna
2. chyba 2. druhu – hypotézu H_0 nezamietame, hoci je nesprávna

Definícia 9.1
- Pravdepodobnosť chyby 1. druhu je číslo α
$$\alpha = P(g_0 \in W | H_0)$$
a nazýva sa *hladina významnosti testu*.
- Pravdepodobnosť chyby 2. druhu je číslo β
$$\beta = P(g_0 \notin W | H_1)$$
a číslo $1-\beta$ sa nazýva *sila testu* ($1-\beta = P(q_0 \in W | H_1)$ – pravdepodobnosť, že správne zamietam nulovú hypotézu)

Poznámka 9.1
Ideálne by bolo, keby sa dalo α, β súčasne minimalizovať. Dá sa ukázať, že znižovaním α sa zvyšuje β a naopak. Preto sa zvolí α ľubovoľne malé a hľadá sa kritický obor, ktorý zabezpečí pre dané α minimálne β. Dostaneme najlepší kritický obor na hladine α – ozn. W_α, W_0. Najčastejšou voľbou je $\alpha = 0{,}05$, príp. 0,1, resp. 0,01.

Postup pri testovaní štatistických hypotéz

1. vysloviť hypotézy H_0, H_1
2. zvoliť testové kritérium g a hladinu významnosti α
3. nájsť najlepší kritický obor W_0

[9] Napr. môžeme položiť nulovú hypotézu „stredná hodnota je 5“ a alternatívnu hypotézu „stredná hodnota nie je 5“ – v tomto prípade je alternatívna hypotéza doplnkom nulovej. V prípade „stredná hodnota je 5“ a „stredná hodnota je väčšia ako 5“, už alternatívna hypotéza doplnkom nie je.

iv. Má platiť:

$$P\left(\chi^2_{\alpha_1}(2n) < \frac{2n\cdot\bar{\mathbb{X}}}{\delta} < \chi^2_{1-\alpha_2}(2n)\right) = 1-\alpha$$

Výsledný interval spoľahlivosti:

$$P\left(\frac{2n\cdot\bar{\mathbb{X}}}{\chi^2_{1-\alpha_2}(2n)} < \delta < \frac{2n\cdot\bar{\mathbb{X}}}{\chi^2_{\alpha_1}(2n)}\right) = 1-\alpha$$

Poznámka 8.4

1. Interval spoľahlivosti má byť čo najkratší, pre symetrické rozdelenia typu $\mathsf{N}(0,1)$, $t(n)$ to nastáva pre
$$\alpha_1 = \alpha_2 = \frac{\alpha}{2},$$
Takéto hodnoty sa však používajú aj pri nesymetrickom rozdelení $\chi^2(n)$.

2. Ak chceme odhad len z jednej strany, použijeme taký istý postup. Odvodené odhady sú obojstranné. Možno však uvažovať aj jednostranný odhad (resp. jednostranný interval spoľahlivosti), kedy neznámy parameter je odhad iba zdola (príp. iba zhora). Potom rozoznávame
   - ľavostranný[8] interval spoľahlivosti (zdola) – položíme $\alpha_1 = \alpha$, $\alpha_2 = 0$
   - pravostranný interval spoľahlivosti (zhora) – položíme $\alpha_1 = 0$, $\alpha_1 = \alpha$

Príklad 8.5
Podnik dodáva do obchodu balíčky sušienok, ktorých hmotnosť má rozdelenie $\mathsf{N}(a, 25)$. Náhodným výberom 25 balíčkov sa zistila priemerná hmotnosť 150 g. Určte:

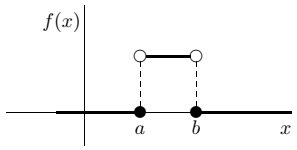
1. 95%-ný interval spoľahlivosti pre strednú hmotnosť
2. hornú medzu strednej hmotnosti, ktorá z pravdepodobnosťou 0,95 nebude prekročená
3. aký by mal byť minimálny rozsah výberu, ak chceme zaručiť chybu odhadu strednej hmotnosti menšiu ako 1 g s pravdepodobnosťou 0,95

Riešenie:
Zo zadania $V_n \in \mathsf{N}(a, 25)$, $n = 25$.

1. Chceme interval spoľahlivosti pre parameter a, ak $\sigma^2 = 25$ (je známe) – ten je takýto:
$$\bar{\mathbb{X}} - \frac{\sigma}{\sqrt{n}}\cdot u_{1-\alpha_2} < a < \bar{\mathbb{X}} + \frac{\sigma}{\sqrt{n}}\cdot u_{1-\alpha_1}$$
Bodovým odhadom a je $\bar{\mathbb{X}} = 150$, $\sigma^2 = 25$, $n = 25$. Chceme 95%-nú spoľahlivosť, teda $\alpha = 1 - 0{,}95 = 0{,}05$, $\alpha/2 = 0{,}025$. Môžeme dosadiť:
$$150 - u_{1-\frac{\alpha}{2}} < a < 150 + u_{1-\frac{\alpha}{2}}$$
Kvantily nájdeme v tabuľkách a po dosadení nám vyjde požadovaný interval spoľahlivosti:
$$148{,}04 < a < 151{,}96$$

2. Na výpočet hornej medze potrebujeme vlastne určiť pravostranný interval spoľahlivosti. Využijeme výpočet z predchádzajúceho bodu:
$$a < \bar{\mathbb{X}} + \frac{\sigma}{\sqrt{n}}u_{1-\alpha_1}$$
Teraz však položíme $\alpha_1 = \alpha = 0{,}05$. Dosadíme, kvantily nájdeme v tabuľkách a získame
$$a < 151{,}64$$

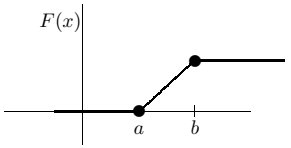
[8] Pozor! V publikácii Potocký a kol. je to naopak!



Obr. 3: Hustota rovnomerného rozdelenia spojitej náhodnej veličiny

Interpretácia rovnomerného rozdelenia. Náhodná veličina $\mathbb{X}$ s rovnomerným rozdelením $\mathsf{R}(a,b)$ reprezentuje dobu čakania na pravidelne sa opakujúcu udalosť. Napr. doba čakania na MHD – ak prídeme na zastávku v náhodnom okamihu, čas čakania má rovnomerné rozdelenie – minimálne čakáme 0 minút, maximálne n minút, kde n je časový interval medzi príchodmi spojov.

1. distribučná funkcia
$$F(x) = \int_{-\infty}^{0} f(t)\,\mathrm{d}t = \begin{cases} 0 & \text{ak } x < a \\ 1 & \text{ak } x \le b \\ \displaystyle\int_a^x \frac{1}{b-a}\,\mathrm{d}t = \frac{x-a}{b-a} & \text{ak } x \in (a,b) \end{cases}$$



Obr. 4: Distribučná funkcia rovnomerného rozdelenia

2. charakteristická funkcia
$$\begin{aligned} \varphi(t) &= \int_{-\infty}^{\infty} \mathrm{e}^{itx} f(x)\,\mathrm{d}x = \frac{1}{b-a}\int_a^b \mathrm{e}^{itx}\,\mathrm{d}x = \frac{1}{b-a}\left[\frac{\mathrm{e}^{itx}}{it}\right]_a^b \\ &= \frac{\mathrm{e}^{itb} - \mathrm{e}^{ita}}{i(b-a)t} \quad \forall t \in \mathbb{R}_1 - \{0\} \end{aligned}$$

Poznámka 4.1
$E(\mathbb{X})$, $D(\mathbb{X})$ sa počítajú z definície, nie podľa vzťahu $\varphi_0^{(k)} = i^k \cdot m_k$, lebo bod 0 nemôžeme v tomto prípade dosadiť.

3. charakteristika polohy a variability
$$\begin{aligned} E(\mathbb{X}) &= \int_{-\infty}^{\infty} x\cdot f(x)\,\mathrm{d}x = \frac{1}{b-a}\int_a^b x\,\mathrm{d}x = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{b^2-a^2}{2(b-a)} \Rightarrow E(\mathbb{X}) = \frac{a+b}{2} \\ E(\mathbb{X}^2) &= \frac{1}{b-a}\cdot\left[\frac{x^3}{3}\right]_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{b} \\ D(\mathbb{X}) &= \frac{a^2+ab+b^2}{3}\cdot\left(\frac{a+b}{2}\right)^2 = \frac{a^2-2ab+b^2}{12} = \frac{(a-b)^2}{12} \end{aligned}$$

4.2.2 Exponenciálne rozdelenie $\mathsf{Ex}(\delta)$

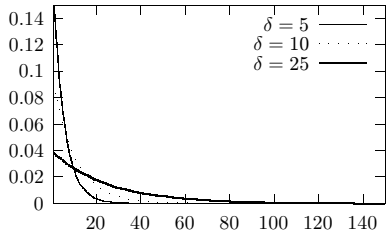
Definícia 4.5
Hovoríme, že spojitá náhodná veličina má exponenciálne rozdelenie s parametrom δ ak má hustotu

$$f(x) = \begin{cases} \frac{1}{\delta} \cdot \mathrm{e}^{-x/\delta} & \text{ak } x > 0 \\ 0 & \text{ak } x \leq 0 \end{cases}$$

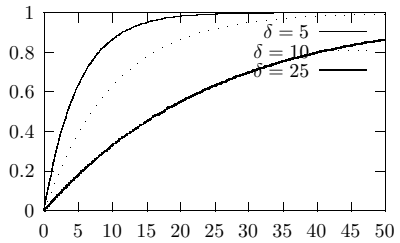
Interpretácia exponenciálneho rozdelenia Náhodná veličina $\mathbb{X}$ s exponenciálnym rozdelením $\mathsf{Ex}(\delta)$ reprezentuje[1]dobu čakania na náhodne sa vyskytujúce udalosti (dobu čakania na obsluhu, doba životnosti súčiastky).

1. distribučná funkcia

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \begin{cases} 0 & \text{ak } x \leq 0 \\ \int_0^x \frac{1}{\delta} \cdot \mathrm{e}^{-t/\delta}\,\mathrm{d}t = \frac{1}{\delta}\left[\frac{\mathrm{e}^{-t/\delta}}{-\frac{1}{\delta}}\right]_0^x = -[\mathrm{e}^{-x/\delta} - 1] = 1 - \mathrm{e}^{-x/\delta} & \text{ak } x > 0 \end{cases}$$



Obr. 5: Hustota rozdelenia $\mathsf{Ex}(\delta)$



Obr. 6: Distribučná funkcia rozdelenia $\mathsf{Ex}(\delta)$

2. charakteristická funkcia

$$\begin{aligned} \varphi(t) &= \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}tx} \cdot \frac{1}{\delta}\mathrm{e}^{-x/\delta}\,\mathrm{d}x = \frac{1}{\delta}\int_{-\infty}^{\infty} \mathrm{e}^{\frac{(\delta \mathrm{i}t-1)\cdot x}{\delta}}\,\mathrm{d}x = \frac{1}{\delta}\cdot\left[\frac{\mathrm{e}^{\frac{(\delta \mathrm{i}t-1)\cdot x}{\delta}}}{\frac{\delta \mathrm{i}t-1}{\delta}}\right]_0^{\infty} = \\ &= \frac{1}{\delta \mathrm{i}t}\left[\mathrm{e}^{\mathrm{i}t-\frac{1}{\delta}\cdot x}\right]_0^{\infty} = \frac{1}{\delta \mathrm{i}t - 1}(0-1) = \frac{1}{1 - \mathrm{i}t\delta}, \quad \forall t \in \mathbb{R} \end{aligned}$$

[1]V ďalšom budeme pojem „náhodná veličina majúca exponenciálne (resp. iné) rozdelenie“ označovať ako $\mathbb{X} \sim \mathsf{Ex}(\delta)$.

(b) Hľadáme interval spoľahlivosti pre parameter a, ak σ^2 je neznáme.

i. Vhodným bodovým odhadom pre a je $\bar{\mathbb{X}}$.
ii. Vhodnou štatistikou je $g = \frac{\bar{\mathbb{X}}-a}{S_1} \cdot \sqrt{n} \sim t(n-1)$ (S_1 slúži ako odhad pre neznámy parameter σ^2)
iii. $g_1 = t_{\alpha_1}(n-1)$, $g_2 = t_{1-\alpha_2}(n-1)$
iv. Má platiť:

$$P\left(t_{\alpha_1}(n-1) < \frac{\bar{\mathbb{X}} - a}{S_1} \cdot \sqrt{n} < t_{1-\alpha_2}(n-1)\right) = 1 - \alpha$$

Výsledný interval spoľahlivosti (určíme ho podobne ako v predchádzajúcom prípade):

$$\bar{\mathbb{X}} - \frac{S_1}{\sqrt{n}} t_{1-\alpha_2}(n-1) < a < \bar{\mathbb{X}} + \frac{S_1}{\sqrt{n}} t_{1-\alpha_1}(n-1)$$

(c) Hľadáme interval spoľahlivosti pre parameter σ^2, ak a je známe.

i. Vhodným bodovým odhadom pre σ^2 je $S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbb{X}_i - a)^2$, kde $a = E(\mathbb{X}_i) = \text{konšt}$
ii. Vhodnou štatistikou je $g = \frac{n \cdot S_0^2}{\sigma^2} \sim \chi^2(n)$
iii. $g_1 = \chi^2_{\alpha_1}(n)$, $g_2 = \chi^2_{1-\alpha_2}(n)$
iv. Má platiť:

$$P\left(\chi^2_{\alpha_1}(n) < \frac{n \cdot S_0^2}{\sigma^2} < \chi^2_{1-\alpha_2}(n)\right) = 1 - \alpha$$

V tomto prípade však musíme dať pozor nato, že rozdelenie χ^2 nie je symetrické. Výsledný interval spoľahlivosti bude teda po úpravách:

$$P\left(n \cdot S_0^2 \cdot \chi^2_{1-\alpha_2}(n) < \sigma^2 < n \cdot S_0^2 \cdot \chi^2_{\alpha_1}(n)\right) = 1 - \alpha$$

(d) Hľadáme interval spoľahlivosti pre parameter σ^2, ak a je neznáme.

i. Vhodným bodovým odhadom pre σ^2 je S_1^2
ii. Vhodnou štatistikou je $g = \frac{(n-1)\cdot S_1^2}{\sigma^2} \sim \chi^2(n-1)$
iii. $g_1 = \chi^2_{\alpha_1}(n-1)$, $g_2 = \chi^2_{1-\alpha_2}(n-1)$
iv. Má platiť:

$$P\left(\chi^2_{\alpha_1}(n-1) < \frac{(n-1)\cdot S_1^2}{\sigma^2} < \chi^2_{1-\alpha_2}(n-1)\right) = 1 - \alpha$$

Opäť musíme vziať do úvahy asymetriu rozdelenia χ^2. Výsledný interval spoľahlivosti po úpravách:

$$P\left(\frac{(n-1)\cdot S_1^2}{\chi^2_{1-\alpha_2}(n-1)} < \sigma^2 < \frac{(n-1)\cdot S_1^2}{\chi^2_{\alpha_1}(n-1)}\right) = 1 - \alpha$$

2. $V_n \in \mathsf{Ex}(\delta)$

Teraz máme len jeden prípad - budeme odhadovať parameter δ

i. Vhodným bodovým odhadom δ je $\bar{\mathbb{X}}$
ii. Vhodnou štatistikou je $g = \frac{2n\mathbb{X}}{\delta} \sim \chi^2(2n)$
iii. $g_1 = \chi^2_{\alpha_1}(2n)$, $g_2 = \chi_{1-\alpha_2}(2n)$

g_1 je α_1-kvantil štatistiky g.
Z podmienky $P(g \geq g_2) = \alpha_2$ vyplýva, že $1 - P(g < g_2) = \alpha_2$, čo je ekvivalentné tomu, že

$$F(g_2) = P(g < g_2) = 1 - \alpha_2.$$

Ďalej opäť použijeme spojitosť a rýdzu monotónnosť funkcie F, na obe strany rovnosti aplikujeme F^{-1}.

$$\begin{aligned} F^{-1}(F(g_2)) &= F^{-1}(1-\alpha_2) \\ g_2 &= F^{-1}(1-\alpha_2) \end{aligned}$$

g_2 je teda $(1-\alpha_2)$-kvantil štatistiky g.

Príklady konštrukcie pre intervaly spoľahlivosti

1. $V_n \in \mathsf{N}(a, \sigma^2)$. Pre intervaly spoľahlivosti rozdelenia $\mathsf{N}(a, \sigma^2)$ máme 4 prípady:
   (a) interval spoľahlivosti pre parameter a, ak σ^2 je známe
   (b) interval spoľahlivosti pre parameter a, ak σ^2 je neznáme
   (c) interval spoľahlivosti pre parameter σ^2, ak a je známe
   (d) interval spoľahlivosti pre parameter σ^2, ak a je neznáme
2. $V_n \in \mathsf{Ex}(\delta)$

---

1. $V_n \in \mathsf{N}(a, \sigma^2)$
   (a) Hľadáme interval spoľahlivosti pre parameter a, ak σ^2 je známe.
   Postup:
   i. Vhodným bodovým odhadom pre a je $\bar{\mathbb{X}}$.
   ii. Vhodnou štatistikou je $g = \frac{\bar{\mathbb{X}}-a}{\sigma} \cdot \sqrt{n} \sim \mathsf{N}(0,1)$
   iii. $g_1 = u_{\alpha_1}$, $g_2 = u_{1-\alpha_2}$ (u-kvantily sú kvantily normovaného normálneho rozdelenia)
   iv. Má platiť:

$$\begin{aligned} P(g_1 < &\quad g \quad &< g_2) = 1-\alpha \\ P\,(u_{\alpha_1} < &\quad \tfrac{\bar{\mathbb{X}}-a}{\sigma} \cdot \sqrt{n} \quad &< u_{1-\alpha_2}) = 1-\alpha \end{aligned}$$

   Naším cieľom je vyjadriť parameter a:

$$\begin{aligned} P\left(\frac{\sigma}{\sqrt{n}} \cdot u_{\alpha_1} < &\quad \bar{\mathbb{X}} - a \quad &< \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha_2}\right) = 1-\alpha \\ P\left(\frac{\sigma}{\sqrt{n}} \cdot u_{\alpha_1} - \bar{\mathbb{X}} < &\quad -a \quad &< \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha_2} - \bar{\mathbb{X}}\right) = 1-\alpha \\ P\left(\bar{\mathbb{X}} - \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha_2} < &\quad a \quad &< \bar{\mathbb{X}} - \frac{\sigma}{\sqrt{n}} \cdot u_{\alpha_1}\right) = 1-\alpha \end{aligned}$$

   Teraz využijeme vlastnosť, že $u_\alpha = -u_{1-\alpha}$ (normované normálne rozdelenie je symetrické):

$$P\left(\bar{\mathbb{X}} - \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha_2} < \quad a \quad < \bar{\mathbb{X}} + \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha_1}\right) = 1-\alpha$$

   Interval spoľahlivosti je teda

$$\left(\bar{\mathbb{X}} - \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha_2};\ \bar{\mathbb{X}} + \frac{\sigma}{\sqrt{n}} \cdot u_{1-\alpha_1}\right)$$

3. charakteristiky polohy a variability

$$E(\mathbb{X}) = \delta, \qquad D(\mathbb{X}) = \delta^2$$

Dôkaz:

$$\begin{aligned} \varphi'(t) &= \frac{-1}{1-\mathrm{i}t\delta}^{2} \cdot (-\mathrm{i}\delta) = \frac{\mathrm{i}\delta}{1-2\mathrm{i}t\delta+\mathrm{i}^2t^2\delta^2} = \frac{\mathrm{i}\delta}{t^2\delta^2-2\mathrm{i}t+1} \\ \varphi'(0) &= \frac{\mathrm{i}\delta}{1} \Rightarrow E(\mathbb{X}) = \delta \\ \varphi''(t) &= \frac{(-2)\cdot \mathrm{i}\delta}{(1-\mathrm{i}t\delta)^3} \cdot (-\mathrm{i}\delta) = \frac{2 \cdot i^2\delta^2}{(1-\mathrm{i}t\delta)^3} \\ \varphi''(0) &= \mathrm{i}^2\left(\frac{2\delta^2}{(1-0)^3}\right) \Rightarrow E(\mathbb{X}^2) = \frac{2\delta^2}{(1-0)^3} = 2\delta^2 \\ D(\mathbb{X}) &= E(\mathbb{X}^2) - E^2(\mathbb{X}) = 2\delta^2 - \delta^2 = \delta^2 \end{aligned}$$

□

4.2.3 Normálne rozdelenie $\mathsf{N}(a, \sigma^2)$

Definícia 4.6
Hovoríme, že spojitá náhodná veličina $\mathbb{X}$ má *normálne (Gaussovo) rozdelenie* s parametrami a, σ^2, ak má hustotu[2]

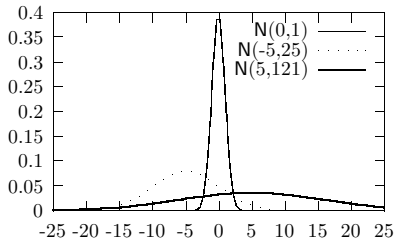
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \mathrm{e}^{-\frac{(x-a)^2}{2\sigma^2}} \qquad \text{pre } x \in \mathbb{R}_1, a \in (-\infty, \infty), \sigma > 0$$

Interpretácia normálneho rozdelenia. Náhodná veličina $\mathbb{X} \sim \mathsf{N}(a, \sigma^2)$ reprezentuje napr. náhodnú chybu v meraní.

1. distribučná funkcia

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{(t-a)^2}{2\sigma^2}}\,\mathrm{d}t$$

Tento integrál ale nemá primitívnu funkciu medzi elementárnymi funkciami. Preto sa hodnoty $F(x)$ aproximujú pre špeciálny prípad $a = 0$, $\sigma^2 = 1$, čím dostaneme tzv. *normované (štandardizované) normálne rozdelenie*.



Obr. 7: Hustota rozdelenia $\mathsf{N}(a, \sigma^2)$ pre rôzne hodnoty a, σ^2

2. charakteristická funkcia

$$\varphi(t) = \mathrm{e}^{\mathrm{i}ta - \frac{t^2\sigma^2}{2}}$$

[2]Parameter σ^2 čítame „sigma kvadrát“.

3. charakteristiky polohy a variability

$$\begin{aligned}
\varphi'(t) &= \mathrm{e}^{\mathrm{i}ta-\frac{t^2\sigma^2}{2}}\cdot(\mathrm{i}a-2\frac{t\sigma^2}{2}) = \mathrm{e}^{\mathrm{i}ta-\frac{t^2\sigma^2}{2}}\cdot(\mathrm{i}a-t\sigma^2)\\
\varphi'(0) &= \mathrm{i}a \Rightarrow E(\mathbb{X}) = a\\
\varphi''(t) &= \mathrm{e}^{\mathrm{i}ta-\frac{t^2\sigma^2}{2}}\cdot(\mathrm{i}a-t\sigma^2)\cdot(\mathrm{i}a-t\sigma^2)+\mathrm{e}^{\mathrm{i}ta-\frac{t^2\sigma^2}{2}}(-\sigma^2)\\
\varphi''(0) &= \mathrm{i}^2(a^2+\sigma^2) \Rightarrow E(\mathbb{X}^2) = a^2+\sigma^2\\
D(\mathbb{X}) &= E(\mathbb{X}^2)-E^2(\mathbb{X}) = a^2+\sigma^2-a^2 = \sigma^2
\end{aligned}$$

Normované normálne rozdelenie. Nech $\mathbb{X}\sim\mathsf{N}(a,\sigma^2)$. Potom náhodná veličina

$$\mathbb{U} = \frac{\mathbb{X}-E(\mathbb{X})}{\sqrt{D(\mathbb{X})}} = \frac{\mathbb{X}-a}{\sigma}\sim\mathsf{N}(0,1)$$

Dôkaz: Ak $\mathbb{X}\sim\mathsf{N}(a,\sigma)$ práve vtedy, keď $\varphi_{\mathbb{X}}(t)=\mathrm{e}^{\mathrm{i}ta-\frac{t^2\sigma^2}{2}}$. Chceme dokázať, že $\mathbb{U}\sim\mathsf{N}(0,1)$ práve vtedy, keď $\varphi_{\mathbb{U}}(t)=\mathrm{e}^{-\frac{t^2}{2}}$. Počítajme:

$$\begin{aligned}
\varphi_{\mathbb{U}}(t) &= \varphi_{\frac{\mathbb{X}-a}{\sigma}}(t) = \varphi_{\frac{1}{\sigma}\mathbb{X}+\left(-\frac{a}{\sigma}\right)}(t) \overset{\text{v. ??}}{=\!=} \mathrm{e}^{\mathrm{i}t\left(-\frac{a}{\sigma}\right)}\cdot\varphi_{\mathbb{X}}\left(\frac{1}{\sigma}\cdot t\right)\\
&= \mathrm{e}^{\mathrm{i}t\left(-\frac{a}{\sigma}\right)}\cdot\mathrm{e}^{\mathrm{i}\cdot\left(\frac{t}{\sigma}\right)a-\frac{(1/\sigma)^2t^2}{2}} = \mathrm{e}^{\frac{\mathrm{i}ta}{\sigma}}\cdot\mathrm{e}^{-\frac{\mathrm{i}ta}{\sigma}-\frac{(1/\sigma)^2t^2}{2}}\\
&= \mathrm{e}^{-\frac{1}{2}t^2}
\end{aligned}$$

□

Poznámka 4.2
Distribučná funkcia normovaného normálneho rozdelenia sa zvykne označovať $\Phi(u)$.

Veta 4.1 (pravidlo 3σ)
Nech $\mathbb{X}\sim\mathsf{N}(a,\sigma^2)$. *Potom* $P(|\mathbb{X}-a|<3\sigma)=0{,}9973$.

Dôkaz:

$$P(|\mathbb{X}-a|<3\sigma) = P\left(\left|\frac{\mathbb{X}-a}{\sigma}\right|<3\right)$$

Z vlastností normovaného normálneho rozdelenia má náhodná veličina $\left|\frac{\mathbb{X}-a}{\sigma}\right| = \mathbb{U}\sim\mathsf{N}(0,1)$. Teda máme

$$\begin{aligned}
P(|\mathbb{U}|<3) &= P(\mathbb{U}\in(-3,3)) = \Phi(3)-\Phi(-3) = \Phi(3)-(1-\Phi(3)) = 2\Phi(3)-1 =\\
&= 2\cdot 0{,}99865-1 = 0{,}9973
\end{aligned}$$

□

Poznámka 4.3 (Význam rozdelenia $\mathsf{N}(0,1)$)

- Z normovaného normálneho rozdelenia sa dajú odvodiť tri špeciálne typy rozdelení – χ^2, t a F-rozdelenie, ktoré sú dôležité v matematickej štatistike.
- Súčet veľkého počtu nezávislých náhodných veličín má za veľmi všeobecných podmienok približne normované normálne rozdelenie. To je podstatou centrálnych limitných viet.

4.2.4 Chí-kvadrát rozdelenie (χ^2-rozdelenie)

Definícia 4.7
Hovoríme, že náhodné veličiny $\mathbb{X}_1,\mathbb{X}_2,\ldots,\mathbb{X}_n$ sú *nezávislé*, ak sú nezávislé im odpovedajúce javy, t. j. platí

$$P(\mathbb{X}_1<x_1,\mathbb{X}_2<x_2,\mathbb{X}_3<x_3) = \prod_{i=1}^{n}P(\mathbb{X}_i<x_i)$$

Na nájdenie maxima tejto funkcie je potrebné nájsť body, v ktorej nadobúda prvá derivácia nulovú hodnotu a z nich vybrať bod, v ktorom druhá derivácia je záporná.

$$\begin{aligned}
\frac{\partial\ln L(\boldsymbol{x},\lambda)}{\partial\lambda} &= \frac{1}{\lambda}\sum_{i=1}^{n}x_i-n\\
\frac{1}{\lambda}\sum_{i=1}^{n}x_i-n\Big|_{\lambda=\widehat{\lambda}} = 0 &\Rightarrow \widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n}x_i = \bar{\mathbb{X}}
\end{aligned}$$

Overenie druhej derivácie:

$$\frac{\partial^2\ln L(\boldsymbol{x},\lambda)}{\partial^2\lambda}\Big|_{\lambda=\widehat{\lambda}} = -\frac{1}{\lambda^2}\sum_{i=1}^{n}x_i\Big|_{\lambda=\widehat{\lambda}} < 0$$

čiže naozaj: $\mathbb{X}$ je maximálnym vierohodným odhadom parametra λ.

8.2 Intervalové odhady

Cieľom je na základe realizácie náhodného výberu V_n skonštruovať taký interval (θ_1,θ_2), ktorý s vopred danou pravdepodobnosťou obsahuje neznámy parameter θ.

Definícia 8.8
Nech náhodný výber $V_n\in F(x,\theta)$, kde $\theta\in\Theta$. Interval (θ_1,θ_2), kde (θ_1,θ_2), pre ktorý platí

$$P(\theta_1<\theta<\theta_2|_{\theta=\theta_0}) = 1-\alpha, \tag{8.6}$$

kde $\alpha\in(0,1)$, θ_0 je skutočnou hodnotou parametra θ a $\theta,\theta_0\in\Theta$; sa nazýva $100\cdot(1-\alpha)\%$-*ný interval spoľahlivosti* pre parameter θ. Číslo $1-\alpha$ sa nazýva *koeficient spoľahlivosti*.

Poznámka 8.3
Číslo α si volíme najčastejšie $\alpha=0{,}05$ (príp. 0,01 alebo 0,1), tzn. dostaneme 95% (90%, 99%) interval spoľahlivosti.

Postup pri konštrukcii intervalu spoľahlivosti

1. Vychádzame z nejakého vhodného bodového odhadu parametra θ
2. Doplníme bodový odhad na vhodnú štatistiku g
3. Nájdeme čísla g_1, g_2 také, že

$$P(g\le g_1) = \alpha_1, \qquad P(g\ge g_2) = \alpha_2, \tag{8.7}$$

   kde $\alpha_1+\alpha_2=\alpha$, $g_1<g_2$, $\alpha_1,\alpha_2,\alpha\in(0,1)$, $g_1,g_2\in\mathbb{R}$.

4. Sčítaním rovníc (8.6) a (8.7) dostaneme

$$\begin{aligned}
P(g\le g_1)+P(g\ge g_2) &= \alpha_1+\alpha_2 = \alpha\\
1-P(g_1<g<g_2) &= \alpha\\
P(g_1<g<g_2) &= 1-\alpha
\end{aligned} \tag{8.8}$$

   Zo vzťahu (8.8) ekvivalentnými úpravami získame tvar (8.6).

Čísla g_1, g_2 vo vzťahu (8.7) sú vlastne kvantily, ak distribučná funkcia štatistiky g je spojitá a rastúca. Zo spojitosti distribučnej funkcie tiež vyplýva, že $P(g\le g_1)=\alpha_1=P(g<g_1)$, čo je vlastne inverzná funkcia F^{-1}. Teda

$$F(g_1)=\alpha_1 \Rightarrow g_1 = F^{-1}(\alpha),$$

Definícia 8.4
Hovoríme, že bodový odhad $g = g(\mathbb{X}_1, \ldots, \mathbb{X}_n)$ *je najlepším nestranným odhadom* parametra θ rozdelenia $F(x, \theta)$, ak platí:

1. $E(g) = \theta$
2. $D(g) \leq D(g')$, pre g' ľubovoľný nestranný odhad

Definícia 8.5
Hovoríme, že bodový odhad $g = g(\mathbb{X}_1, \ldots, \mathbb{X}_n)$ *je konzistentným odhadom* parametra θ rozdelenia $F(x, \theta)$, ak platí:

$$\forall \varepsilon > 0 : \lim_{n\to\infty} P(|E(g) - \theta| \geq \varepsilon) = 0$$

Poznámka 8.1
Na určenie konzistentnosti bodového odhadu sa nepoužíva definícia, ale postačujúca podmienka:

$$\lim_{n\to\infty} E(g) = \theta \;\&\; \lim_{n\to\infty} D(g) = 0 \Rightarrow g \text{ je konzistentný odhad}$$

Metóda hľadania vhodného bodového bodového odhadu – metóda maximálnej vierohodnosti

Definícia 8.6
Nech náhodný výber V_n pochádza z rozdelenia daného hustotou (zákonom rozdelenia) $f_i(x_i, \theta)$, kde $\theta \in \Theta$. *Vierohodnostnou funkciou* nazývame funkciu

$$L(\boldsymbol{x}, \theta) = L(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} f_i(x_i, \theta)$$

Definícia 8.7
Maximálne vierohodným odhadom parametra θ rozdelenia $F(x, \theta)$ nazývame taký bod $\widehat{\theta}$, v ktorom vierohodnostná funkcia nadobúda maximum, t. j.

$$\forall \theta \in \Theta : L(\boldsymbol{x}, \widehat{\theta}) \geq L(\boldsymbol{x}, \theta)$$

Hľadanie $\bar{\theta}$

1. $\left.\frac{\partial L(\boldsymbol{x},\theta)}{\partial \theta}\right|_{\theta=\widehat{\theta}} = 0$
2. $\left.\frac{\partial^2 L(\boldsymbol{x},\theta)}{\partial \theta^2}\right|_{\theta=\widehat{\theta}} < 0$

V prípade, že hustota je exponenciálneho typu, tak $\widehat{\theta}$ nájdeme ako bod, v ktorom nadobúda maximum funkcia $\ln L(\boldsymbol{x}, \theta)$.

Príklad 8.2
Nech $V_n \in \mathsf{Po}(\lambda)$. Nájdite maximálny vierohodný odhad parametra λ.

Riešenie:
$\mathbb{X} \sim \mathsf{Po}(\lambda) \Leftrightarrow p_k = P(\mathbb{X} = x_k) = \frac{\lambda^k}{k!} \cdot \mathrm{e}^{-\lambda}$. Počítajme vierohodnostnú funkciu:

$$L(\boldsymbol{x}, \lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!}\mathrm{e}^{-\lambda} = \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=i}^{n}(x_i!)} \cdot \mathrm{e}^{n(-\lambda)}$$

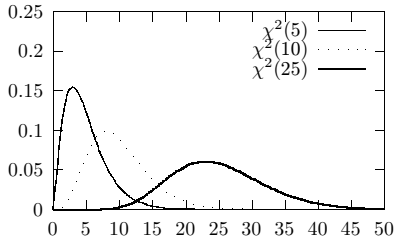
V tomto prípade bude výhodnejšie počítať maximum logaritmu vierohodnostnej funkcie.

$$\ln(L(\boldsymbol{x}, \lambda)) = \ln \lambda^{\sum_{i=1}^{n} x_i} - \ln \prod_{i=1}^{n}(x_i!) - n\lambda = \sum_{i=1}^{n} x_i \cdot \ln \lambda - \ln \prod_{i=1}^{n}(x_i!) - n\lambda$$

Definícia 4.8
Hovoríme, že spojitá náhodná veličina $\mathbb{Y}_n$ má chí-kvadrát rozdelenie o n stupňoch voľnosti, ak má hustotu

$$f_n(y) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \cdot y^{\frac{n}{2}-1} \cdot \mathrm{e}^{-\frac{y}{2}} & \text{ak } y > 0 \\ 0 & \text{ak } y \leq 0 \end{cases}$$

Označujeme $\mathbb{Y}_n \sim \chi^2(n)$.



Obr. 8: Hustota rozdelenia $\chi^2(n)$

Vlastnosti rozdelenia chí-kvadrát.

1. Rozdelenie $\chi^2(n)$ nie je symetrické[3]. Kvantily sa tabelizujú pre $n = 1, \ldots, 100$. Pre $n > 100$ sa toto rozdelenie aproximuje normálnym rozdelením $\mathsf{N}(n, 2n)$.

2. Charakteristická funkcia, charakteristiky polohy a variability.

$$\varphi(t) = \frac{1}{(1 - 2\mathrm{i}t)^{\frac{n}{2}}}, \qquad E(\mathbb{Y}) = n, \qquad D(\mathbb{Y}) = 2n$$

3. Platí nasledovná vlastnosť:

$$\mathbb{Y}_n \sim \chi^2(n) \Leftrightarrow \mathbb{Y}_n = \sum_{i=1}^{n} \mathbb{X}_i^2,$$

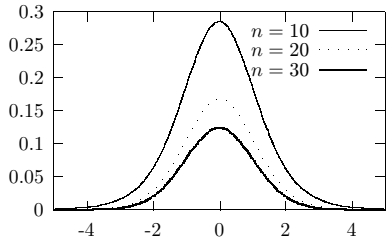
kde $\mathbb{X}_i \sim \mathsf{N}(0, 1)$ a veličiny $\mathbb{X}_i$ sú nezávislé.

4.2.5 Studentovo rozdelenie (t-rozdelenie)

Definícia 4.9
Hovoríme, že spojitá náhodná veličina $\mathbb{T}$ má *Studentovo rozdelenie* (t-*rozdelenie*) o n stupňoch voľnosti, ak má hustotu

$$f_n(t) = \frac{1}{\beta\left(\frac{n}{2}, \frac{1}{2}\right)} \cdot \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \quad \text{pre } t \in (-\infty, \infty), n \in \mathbb{N}$$

[3] Čím väčšie je n, tým má rozdelenie bližšie k symetrickému.



Obr. 9: Hustota rozdelenia $t(n)$

Vlastnosti t-rozdelenia

1. Rozdelenie t je symetrické. Kvantily sú tabelované pre $n \leq 30$. Pre väčšie n sa toto rozdelenie aproximuje pomocou rozdelenia $\mathsf{N}\left(0, \frac{n}{n-2}\right)$.

2. Charakteristiky polohy a variability.

$$E(\mathbb{T}) = 0, \qquad D(\mathbb{T}) = \frac{n}{n-2}, \quad \text{pre } n > 2$$

3. Platí:

$$\mathbb{T} \sim t(n) \Leftrightarrow \mathbb{T} = \frac{\mathbb{X}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} \mathbb{X}_i^2}} \tag{4.1}$$

kde $\mathbb{X}_i \sim \mathsf{N}(0,1)$ a veličiny $\mathbb{X}_i$ sú nezávislé.

4.2.6 Fischerovo-Snedecorovo rozdelenie (F-rozdelenie)

Definícia 4.10
Hovoríme, že spojitá náhodná veličina $\mathbb{Z}$ má *Fischerovo-Snedecorovo rozdelenie* (F-*rozdelenie*) s n_1, n_2 stupňami voľnosti, ak má hustotu

$$f(z) = \begin{cases} \dfrac{1}{\beta\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \cdot \left(\dfrac{n_1}{n_2}\right)^{\frac{n_1}{2}} \cdot z^{\frac{n_1}{2}-1} \cdot \left(1 + \dfrac{n_1}{n_2} \cdot z\right)^{-\frac{n_1+n_2}{2}} & \text{ak } z > 0 \\ 0 & \text{ak } z \leq 0 \end{cases}$$

Vlastnosti F-rozdelenia

1. Rozdelenie $F(n_1, n_2)$ nie je symetrické. Kvantily sú tabelované pre $n_1 \leq 100$, $n_2 \leq 100$. Pre $n_1 > 100$ alebo $n_2 > 100$ odhadujeme toto rozdelenie normálnym rozdelením $\mathsf{N}(E(\mathbb{Z}), D(\mathbb{Z}))$. Pre interpoláciu kvantilov v tabuľkách sa používa vzťah

$$F_\alpha(n_1, n_2) = \frac{1}{F_{1-\alpha}(n_2, n_1)}$$

2. Charakteristiky polohy a variability.

$$E(\mathbb{Z}) = \frac{n_2}{n_2 - 2}, \qquad D(\mathbb{Z}) = \frac{2n_2^2(n_1 + n_2 + 2)}{n_1(n_2-2)^2(n_2-4)}, \quad n_2 > 4$$

Platí: $\varphi_g(t) = \varphi_{\frac{2n\mathbb{X}}{\sigma}}(t) = \varphi_{\frac{2n}{\sigma}\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbb{X}_i} = \varphi_{\frac{2}{\sigma}\sum_{i=1}^{n}\mathbb{X}_i}(t)$. Predstavme si členy v dolnom indexe v tvare $a\mathbb{X} + b$ ($a = 2/\sigma$, $\mathbb{X}$ bude tvoriť suma, a b bude nulové). Veta **??**, bod 3.) hovorí, že $\varphi_{a\mathbb{X}+b} = \mathrm{e}^{\mathrm{i}tb} \cdot \varphi_{\mathbb{X}}(at)$. Použijeme ju na náš prípad: $\varphi_g(t) = \mathrm{e}^0 \cdot \varphi_{\sum_{i=1}^{n}\mathbb{X}_i}\left(\frac{2}{\sigma}\cdot t\right) \overset{\text{vl. } \varphi}{=\!=} \prod_{i=1}^{n}\varphi_{\mathbb{X}_i}\left(\frac{2}{\delta}\cdot t\right)$. Použijeme teraz predpoklad o exponenciálnom rozdelení a teda posledný

$$\varphi_g(t) \;=\; \prod_{i=1}^{n}\frac{1}{1 - \mathrm{i}\left(\frac{2}{\delta}t\right)\delta} = \prod_{i=1}^{n}\underbrace{\frac{1}{1-2\mathrm{i}t}}_{\text{konšt. pre } i} = \left(\frac{1}{1-2\mathrm{i}t}\right)^n = \frac{1}{(1-2\mathrm{i}t)^n} \qquad \square$$

Ak sa pozrieme na posledný člen a na rovnosť v (7.5), zistíme, že $\varphi_g(t)$ je práve v tomto tvare, až na parameter n, ktorý je v poslednom člene dvojnásobný. Preto

$$g = \frac{2n\mathbb{X}}{\sigma} \sim \chi^2(2n).$$

8 Teória odhadov

Úlohou teórie odhadov je na základe náhodného výberu V_n čo najlepšie odhadnúť neznámy parameter θ rozdelenia $F(x,\theta)$. Rozoznávame dva typy:

- bodový odhad
- intervalový odhad

8.1 Bodové odhady

Úlohou bodového odhadu je nahradiť neznámu hodnotu parametra θ hodnotou vhodne zvolenej štatistiky.

Definícia 8.1
Nech $V_n \in F(x,\theta)$, kde $\theta \in \Theta$. *Bodovým odhadom parametra* θ nazývame ľubovoľnú „vhodne zvolenú“ funkciu náhodného výberu (štatistiku) g, takú, že

$$g = g(\mathbb{X}_1, \ldots, \mathbb{X}_n)$$

Kritéria vhodnosti bodového odhadu

- nestrannosť (nevychýlenosť)
- konzistentnosť
- výdatnosť...

Definícia 8.2
Nech $V_n \in F(x,\theta)$, kde $\theta \in \Theta$. Hovoríme, že bodový odhad $g = g(\mathbb{X}_1, \ldots, \mathbb{X}_n)$ *je nestranným* (nevychýleným) odhadom parametra θ, ak platí:

$$E(g) = E(g(\mathbb{X}_1, \ldots, \mathbb{X}_n)) = \theta$$

Definícia 8.3
Hovoríme, že bodový odhad $g = g(\mathbb{X}_1, \ldots, \mathbb{X}_n)$ *je asymptoticky nestranným odhadom* parametra θ rozdelenia $F(x,\theta)$, ak platí:

$$\lim_{n\to\infty} E(g(\mathbb{X}_1, \ldots, \mathbb{X}_n)) = \theta$$

3. Počítajme:

$$\begin{aligned} g &= \frac{(n-1)\cdot S_1^2}{\sigma^2} \overset{\text{def.}}{=\!=} \frac{(n-1)}{\sigma^2}\cdot\frac{1}{n-1}\sum_{i=1}^{n}(\mathbb{X}_i-\bar{\mathbb{X}})^2 = \frac{1}{\sigma^2}\cdot\sum_{i=1}^{n}((\mathbb{X}_i-a)-(\bar{\mathbb{X}}-a))^2 \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[(\mathbb{X}_i-a)^2+(\bar{\mathbb{X}}-a)^2-2(\mathbb{X}_i-a)(\bar{\mathbb{X}}-a)\right] \\ &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(\mathbb{X}_i-a)^2\right)-2(\bar{\mathbb{X}}-a)\sum_{i=1}^{n}(\mathbb{X}_i-a)+n\cdot(\bar{\mathbb{X}}-a)^2 \end{aligned}$$

Použijeme teraz trik: upravíme sumu $\sum_{i=1}^{n}(\mathbb{X}_i-a)$:

$$\sum_{i=1}^{n}(\mathbb{X}_i-a)=\left(\sum_{i=1}^{n}\mathbb{X}_i\right)-na=n\cdot\left(\frac{1}{n}\left(\sum_{i=1}^{n}\mathbb{X}_i\right)-a\right)=n\cdot(\bar{\mathbb{X}}-a)$$

a pokračujeme vo výpočte g

$$\begin{aligned} g &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(\mathbb{X}_i-a)^2\right)-2(\bar{\mathbb{X}}-a)\cdot n\cdot(\mathbb{X}_i-a)+n\cdot(\bar{\mathbb{X}}-a)^5 \\ &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(\mathbb{X}_i-a)^2\right)-2n\cdot(\bar{\mathbb{X}}-a)^2+n\cdot(\bar{\mathbb{X}}-a)^5 \\ &= \sum_{i=1}^{n}\left(\frac{\mathbb{X}_i-a}{\sigma}\right)^2-\frac{n(\bar{\mathbb{X}}-a)^2}{\sigma^2} \\ &= \sum_{i=1}^{n}\left(\frac{\mathbb{X}_i-a}{\sigma}\right)^2-\left(\frac{\bar{\mathbb{X}}-a}{\sigma^2}\cdot\sqrt{n}\right) \end{aligned}$$

V posledne uvedenom rozdieli má menšenec rozdelenie $\chi^2(n)$ (pozri začiatok dôkazu bodu 2). Ak sa pozrieme do menšiteľa na výraz pod mocninou, tak vidíme, že tento výraz má rozdelenie $\mathsf{N}(0,1)$. Jeho mocnina má však rozdelenie $\chi^2(n)$[6]?. Celý rozdiel má teda tiež rozdelenie $\chi^2(n)$[7], čo sme chceli dokázať.

4. Z vlastností t-rozdelenia – vzťah (4.1) – vieme, že

$$\mathbb{T}=\frac{\mathbb{X}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathbb{X}_i^2}}\sim t(n),$$

ak $\mathbb{X}\sim\mathsf{N}(0,1)$ a sú $\mathbb{X}_i$ sú nezávislé. (— je potrebné dopísať dôkaz) □

Veta 7.4
Nech $V_n\in\mathsf{Ex}(\delta)$. Potom $g=\frac{2n\bar{\mathbb{X}}}{\delta}\sim\chi^2(2n)$.

Dôkaz: Dôkaz vykonáme metódou charakteristických funkcií. Z vlastností rozdelení platí:

$$\mathbb{X}\sim\mathsf{Ex}(\delta)\quad\text{akk}\quad\varphi_{\mathbb{X}}(t)=\frac{1}{1-\mathrm{i}t\delta} \tag{7.4}$$

$$\mathbb{Y}\sim\chi^2(n)\quad\text{akk}\quad\varphi_{\mathbb{Y}}(t)=\frac{1}{(1-2\mathrm{i}t)^{n/2}} \tag{7.5}$$

[6] Je niekde vpredu taká veta
[7] Je veta o tom, že súčet zachováva rozdelenie?

3. Náhodná veličina $\mathbb{Z}\sim F(n_1,n_2)$ práve vtedy, keď

$$\mathbb{Z}=\frac{\frac{1}{n_1}\cdot\mathbb{Y}_1}{\frac{1}{n_2}\cdot\mathbb{Y}_2}=\frac{\frac{1}{n_1}\cdot\sum_{i=1}^{n_1}\mathbb{X}_{1_i}^2}{\frac{1}{n_2}\cdot\sum_{i=1}^{n_2}\mathbb{X}_{2_i}^2},$$

kde $\mathbb{Y}_i\sim\chi^2(n_i)$, pre $i=1,2$ a $\mathbb{X}_{11},\ldots,\mathbb{X}_{1n_1}$, $\mathbb{X}_{21},\ldots,\mathbb{X}_{2n_2}\sim\mathsf{N}(0,1)$ a sú navyše nezávislé.

5 Centrálne limitné vety

Podstatou centrálnych limitných viet je fakt, že súčet veľkého počtu nezávislých náhodných veličín za veľmi všeobecných podmienok má asymptoticky normálne rozdelenie. Tieto podmienky spresnia nasledovné tri vety.

Veta 5.1 (Moivre-Laplace)
Nech $\mathbb{S}_n=\sum_{i=1}^{n}\mathbb{X}_i$, kde $\mathbb{X}_i$ sú nezávislé náhodné veličiny s rozdelením $\mathbb{X}_i\sim\mathsf{Bi}(1,p)=A(p)$ (vykonávame len jeden pokus). Potom normovaná veličina $\widehat{\mathbb{S}}_n$ má približne normované normálne rozdelenie:

$$\widehat{\mathbb{S}}_n=\frac{\mathbb{S}_n-np}{\sqrt{np(1-p)}}\dot{\sim}\,\mathsf{N}(0,1),$$

t. j.

$$\lim_{n\to\infty}F_{\widehat{\mathbb{S}}_n}(s)=\Phi(s)$$

Veta 5.2 (Feller-Lindeberg)
Nech $\mathbb{S}_n=\sum_{i=1}^{n}\mathbb{X}_i$, kde $\mathbb{X}_i$ sú nezávislé náhodné veličiny s identickým rozdelením a s konečnou strednou hodnotou $E(\mathbb{X}_i)=a<\infty$ a konečnou disperziou $D(\mathbb{X}_i)=\sigma^2<\infty$ pre $i=1,\ldots,n$. Potom náhodná veličina $\widehat{\mathbb{S}}_n$ má približne normované normálne rozdelenie:

$$\widehat{\mathbb{S}}_n=\frac{\mathbb{S}_n-na}{\sqrt{n\sigma^2}}\dot{\sim}\,\mathsf{N}(0,1),$$

Veta 5.3 (Ljapunov)
Nech $\mathbb{S}_n=\sum_{i=1}^{n}\mathbb{X}_i$, kde $\mathbb{X}_i$ sú nezávislé náhodné veličiny s konečnou strednou hodnotou $E(\mathbb{X}_i)<\infty$ pre $i=1,\ldots,n$ a konečnou disperziou $D(\mathbb{X}_i)<\infty$) pre $i=1,\ldots,n$. Nech platí Ljapunovova podmienka

$$\lim_{n\to\infty}\frac{\sqrt[3]{\sum_{i=1}^{n}E(|\mathbb{X}_i-E(\mathbb{X}_i)|^3)}}{\sqrt{\sum_{i=1}^{n}D(\mathbb{X}_i)}}=0$$

Potom náhodná veličina $\widehat{\mathbb{S}}_n$ má približne normované normálne rozdelenie:

$$\widehat{\mathbb{S}}_n=\frac{\mathbb{S}_n-\sum_{i=1}^{n}E(\mathbb{X}_i)}{\sqrt{\sum_{i=1}^{n}D(\mathbb{X}_i)}}\dot{\sim}\,\mathsf{N}(0,1)$$

6 Náhodné vektory – viacrozmerné náhodné veličiny

6.1 Združené a marginálne rozdelenie

Nech je daný pravdepodobnostný priestor $(\Omega, \mathcal{A}, P)$. Uvažujme kartézsky súčin intervalov $I_n = (-\infty, x_1) \times (-\infty, x_2) \times \ldots \times (-\infty, x_n)$, kde $x_i \in \mathbb{R}$, pre $i = 1, \ldots, n$.

Definícia 6.1
Zobrazenie

$$\mathbb{X} = (\mathbb{X}_1, \mathbb{X}_2, \ldots \mathbb{X}_n) : \Omega \to \mathbb{R}_n$$

sa nazýva *náhodným vektorom* v $\mathbb{R}_n$, ak vzorom ľubovoľného intervalu v $\mathbb{R}_n$ typu I_n je jav, t. j. platí

$$\mathbb{X}^{-1}(I_n) = \{\omega \in \Omega : \mathbb{X}_1(\omega) < x_1, \mathbb{X}_2(\omega) < x_2, \ldots, \mathbb{X}_n(\omega) < x_n\} \in \mathcal{A}$$

Poznámka 6.1
Vektor $(\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_n)$ je náhodný vektor práve vtedy, ak zložky $\mathbb{X}_i$ sú náhodné veličiny.

Definícia 6.2
Reálna funkcia $F_{\mathbb{X}} : \mathbb{R}_n \to \langle 0, 1\rangle$ definovaná vzťahom

$$F_{\mathbb{X}}(x_1, x_2, \ldots, x_n) = P(\mathbb{X}_1 < x_1, \mathbb{X}_2 < x_2, \ldots, \mathbb{X}_n < x_n)$$

sa nazýva *združenou distribučnou funkciou* náhodného vektora $(\mathbb{X}_1, \ldots, \mathbb{X}_n)$. Distribučné funkcie zložiek náhodného vektora $F_i(x_i)$ pre $i = 1, \ldots, n$ nazývame *marginálne distribučné funkcie*.

Veta 6.1
Nech $F_{\mathbb{X}}(x_1, \ldots, x_n)$ je združenou distribučnou funkciou náhodného vektora $\mathbb{X} = (\mathbb{X}_1, \ldots, \mathbb{X}_n)$. Potom platí:

- $\lim\limits_{\forall i: x_i \to \infty} F(x_1, \ldots, x_n) = 1$ a $\lim\limits_{\exists i: x_i \to -\infty} F(x_1, \ldots, x_n) = 0$
- *$F(x_1, \ldots, x_n)$ je neklesajúca vzhľadom na každú premennú*
- *$F(x_1, \ldots, x_n)$ je zľava spojitá vzhľadom na každú premennú*

Dôkaz: Dôkaz je podobný ako v prípade $\mathbb{R}_1$ (pozri minulý semester). □

Veta 6.2
Nech $F(x_1, \ldots, x_n)$ je združenou distribučnou funkciou náhodného vektora $(\mathbb{X}_1, \ldots, \mathbb{X}_n)$. Potom pre marginálne distribučné funkcie zložiek platí:

$$F_{\mathbb{X}_i} = F_i(x_i) = \lim_{\substack{x_j \to \infty \\ j \neq i}} F(x_1, \ldots, x_n)$$

Dôkaz: Dôkaz urobíme pre $n = 2$. Bez ujmy na všeobecnosti chceme dokázať, že $F_1(x) = \lim\limits_{x_2 \to \infty} F(x_1, x_2)$. Uvažujme postupnosť reálnych čísel $\{x_{2_n}\}_{n=1}^{\infty}$ takú, že pre $n \to \infty$ ide $\{x_{2_n}\} \to \infty$. Potom

$$\lim_{x_2 \to \infty} F(x_1, x_2) = \lim_{x_{2_n} \to \infty} F(x_1, x_{2_n}) = \lim_{x_{2_n} \to \infty} P(\mathbb{X}_1 < x_1, \mathbb{X}_2 < x_{2_n}) = \Delta$$

Označme $A_n = \{\omega \in \Omega : \mathbb{X}_1(\omega) < x_1 \wedge \mathbb{X}_2(\omega) < x_{2_n}\}$. Postupnosť javov $\{A_n\}_{n=1}^{\infty}$ je rastúca (nezabudnite si to premyslieť!) a preto podľa poznámky **??** je $\lim\limits_{n\to\infty} A_n = \bigcup\limits_{n=1}^{\infty} A_n$. Pokračujeme vo výpočte výrazu (Δ):

$$\Delta \;=\; \lim_{x_{2_n} \to \infty} P(A_n) \overset{\text{spojitosť } P(\cdot)}{=\!=} P\left(\lim_{x_{2_n} \to \infty} A_n\right) = P\left(\bigcup_{n=1}^{\infty} A_n\right)$$

Máme $V_n = (\mathbb{X}_1, \ldots, \mathbb{X}_n)$. Jednotlivé zložky $\mathbb{X}_i \sim \mathsf{N}(a, \sigma^2)$ práve vtedy, keď $\varphi_{\mathbb{X}_i}(t) = \mathrm{e}^{\mathrm{i}t\boxed{a} - \frac{1}{2}t^2\cdot\boxed{\sigma^2}}$ „Zaškatuľkované“ sú práve $E(\mathbb{X}_i) = a$ a $D(\mathbb{X}_i) = \sigma^2$. Pre $\bar{\mathbb{X}}$ by teda malo platiť: $\varphi_{\bar{\mathbb{X}}}(t) = \mathrm{e}^{\mathrm{i}t\boxed{a} - \frac{1}{2}t^2\cdot\boxed{\frac{\sigma^2}{n}}}$. Overme to:

$$\varphi_{\bar{\mathbb{X}}}(t) = \varphi_{\frac{1}{n}\sum\limits_{i=1}^{n}\mathbb{X}_i}(t) \overset{\text{vl. } \varphi_{\mathbb{X}}(t)}{=\!=} \varphi_{\sum\limits_{i=1}^{n}\mathbb{X}_i}\left(\frac{1}{n}\cdot t\right)$$

Jednotlivé veličiny $\mathbb{X}_i$ sú nezávislé, môžeme teda použiť vetu 6.5, bod 3.

$$\varphi_{\bar{\mathbb{X}}}(t) = \prod_{i=1}^{n}\varphi_{\mathbb{X}_i}\left(\frac{t}{n}\right) = \prod_{i=1}^{n}\mathrm{e}^{\mathrm{i}\left(\frac{t}{n}\right)a - \frac{1}{2}\left(\frac{t}{n}\right)^2\sigma^2} = \mathrm{e}^{n\cdot\mathrm{i}\left(\frac{t}{n}\right)a - \frac{1}{2}\left(\frac{t}{n}\right)^2\sigma^2} = \mathrm{e}^{\mathrm{i}ta - \frac{1}{2}\left(\frac{t^2}{n}\right)\sigma^2},$$

čo sme chceli overiť. □

7.3 Štatistika a jej rozdelenie

Definícia 7.2
Nech $V_n \in F(x, \theta)$, $\theta \in \Theta$. *Štatistikou* nazývame takú funkciu náhodného výberu V_n, rozdelenie ktorej nezávisí od parametra θ rozdelenia $F(x, \theta)$, z ktorého výber pochádza.

$$g = g(\mathbb{X}_1, \ldots, \mathbb{X}_n)$$

Veta 7.3
Nech $V_n \in \mathsf{N}(a, \sigma^2)$. Potom pre nasledovné štatistiky platí:

1. $g = \dfrac{\bar{\mathbb{X}} - a}{\sigma} \cdot \sqrt{n} \sim \mathsf{N}(0, 1)$
2. $g = \dfrac{n \cdot S_0^2}{\sigma^2} \sim \chi^2(n)$
3. $g = \dfrac{(n-1) \cdot S_1^2}{\sigma^2} \sim \chi^2(n-1)$
4. $g = \dfrac{\bar{\mathbb{X}} - a}{S_1} \cdot \sqrt{n} \sim t(n-1)$

Dôkaz:

1. Overme, či $g = \frac{\bar{\mathbb{X}} - a}{\sigma} \cdot \sqrt{n} \sim \mathsf{N}(0, 1)$.
   $V_n = (\mathbb{X}_1, \ldots, \mathbb{X}_n) \sim \mathsf{N}(a, \sigma^2)$, teda aj jednotlivé zložky $\mathbb{X}_i \sim \mathsf{N}(a, \sigma^2)$. Z vety 7.2 vieme, že $\bar{\mathbb{X}} \sim \mathsf{N}\left(a, \frac{\sigma^2}{n}\right)$. Teda $E(\bar{\mathbb{X}}) = a$, $D(\bar{\mathbb{X}}) = \frac{\sigma^2}{n}$ a môžeme rozdelenie normovať, čím dostávame

   $$\frac{\bar{\mathbb{X}} - a}{\sqrt{\sigma^2/n}} = \frac{\bar{\mathbb{X}} - a}{\sigma} \cdot \sqrt{n} \sim \mathsf{N}(0, 1)$$

2. Platí[5], že $\mathbb{Y} = \sum\limits_{i=1}^{n} \mathbb{X}_i^2 \sim \chi^2(n)$ práve vtedy, keď $\mathbb{X}_i \sim \mathsf{N}(0, 1)$ a náhodné veličiny $\mathbb{X}_i$ sú nezávislé. Chceme ukázať, že uvedená štatistika g sa dá napísať v tomto tvare.

   $$g = \frac{n \cdot S_0^2}{\sigma^2} = \frac{n}{\sigma^2} \cdot \frac{1}{n}\sum_{i=1}^{n}(\mathbb{X}_i - a)^2 = \sum_{i=1}^{n}\left(\frac{\mathbb{X}_i - a}{\sigma^2}\right)^2$$

   Z predpokladu $\mathbb{X}_i \sim \mathsf{N}(a, \sigma^2)$, teda $\frac{\mathbb{X}_i - a}{\sigma^2} \sim \mathsf{N}(0, 1)$. Potrebujeme overiť ešte nezávislosť, ale tá je zaručená už z definície náhodného výberu (V_n je tvorený nezávislými náhodnými veličinami). Overili sme oba predpoklady a teda z ekvivalencie vyplýva požadované rozdelenie $\chi^2(n)$.

[5] ale skäde?

2. charakteristiky variability – výberové rozptyly

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbb{X}_i - \bar{\mathbb{X}})^2$$

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbb{X}_i - a)^2, \quad a = E(\mathbb{X}_i) \text{ je konšt. pre všetky } i$$

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbb{X}_i - \bar{\mathbb{X}})^2$$

Neskôr ukážeme, že S^2 [čítame: S kvadrát] nie je dobrou charakteristikou variability. Výberový rozptyl S_0^2 [S_0 kvadrát] budeme používať, ak poznáme strednú hodnotu. Ak ju nepoznáme, použijeme výberový rozptyl S_1^2 [S_1 kvadrát].

3. charakteristiky závislosti – výberový korelačný koeficient

$$R_{\mathbb{X},\mathbb{Y}} = \frac{K_{\mathbb{X},\mathbb{Y}}}{S_{\mathbb{X}} \cdot S_{\mathbb{Y}}},$$

kde $K_{\mathbb{X},\mathbb{Y}} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{X}_i \cdot \mathbb{Y}_i\right) - \bar{\mathbb{X}} \cdot \bar{\mathbb{Y}}$ a $S_{\mathbb{X}}^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbb{X} - \bar{\mathbb{X}})^2$.

Veta 7.1
Nech náhodný výber V_n pochádza z rozdelenia $F(x,\theta)$, ktoré má konečnú strednú hodnotu $E(\mathbb{X}_i) = a$ pre $i = 1,\ldots,n$ a konečnú disperziu $D(\mathbb{X}) = \sigma^2$, pre $i = 1,\ldots,n$. Potom

1. $E(\bar{\mathbb{X}}) = a$
2. $D(\bar{\mathbb{X}}) = \frac{\sigma^2}{n}$
3. $E(S_0^2) = E(S_1^2) = \sigma^2$
4. $E(S^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$

Dôkaz:

1. $E(\mathbb{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{X}_i\right) = \frac{1}{n}\sum_{i=1}^{n}E(\mathbb{X}_i) \overset{E(\mathbb{X}_i)=a}{=\!=} \frac{1}{n}(n\cdot a) = a$
2. $D(\mathbb{X}) = D\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{X}_i\right) = \frac{1}{n^2}\cdot D(\sum_{i=1}^{n}\underbrace{\mathbb{X}_i}_{\text{nezávislé}}) = \frac{1}{n^2}\sum_{i=1}^{n}D(\mathbb{X}_i) \overset{D(\mathbb{X}_i)=a}{=\!=} \frac{1}{n^2}(n\cdot\sigma^2) = \frac{\sigma^2}{n}$
3. $E(S_0^2) = E\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbb{X}_i - a)^2\right)^2 = \frac{1}{n}\sum_{i=1}^{n}E(\mathbb{X}_i - a)^2 = \frac{1}{n}\sum_{i=1}^{n}E(\mathbb{X}_i - E(\mathbb{X}))^2 = \frac{1}{n}\sum_{i=1}^{n}D(\mathbb{X}) =$
   $= \frac{1}{n}(n\cdot\sigma^2) = \sigma^2$
4. Druhú nerovnosť v 3) a dôkaz 4) (zatiaľ) ponechávame na pozorného čitateľa □

Veta 7.2
Nech $V_n \in \mathsf{N}(a,\sigma^2)$. Potom $\bar{\mathbb{X}} \sim N\left(a, \frac{\sigma^2}{n}\right)$.

Dôkaz: Dôkaz urobíme metódou charakteristických funkcií. Už vieme, že parametrami normálneho rozdelenia, z ktorého pochádza $\mathbb{X}$, sú a, $\frac{\sigma^2}{n}$ (pozri predošlá veta). Chceme ukázať, že $\bar{\mathbb{X}}$ má práve normálne rozdelenie.

$$= P\left(\bigcup_{n=1}^{\infty}\{\omega\in\Omega : \mathbb{X}_1(\omega) < x_1 \wedge \mathbb{X}_2(\omega) < x_{2_n}\}\right)$$

$$= P\left(\bigcup_{n=1}^{\infty}\{\omega\in\Omega : \mathbb{X}_1(\omega) < x_1\} \cap \{\omega\in\Omega : \mathbb{X}_2(\omega) < x_{2_n}\}\right)$$

$$\text{(distr. zákon)} = P\left(\{\omega\in\Omega : \mathbb{X}_1(\omega) < x_1\} \cap \bigcup_{n=1}^{\infty}\{\omega\in\Omega : \mathbb{X}_2(\omega) < x_{2_n}\}\right) =$$

□

Pozrime sa teraz na prienik pod pravdepodobnostnou funkciou. Prvý člen tvorí vlastne $\mathbb{X}_1 < x_1$, druhý člen je celý pravdepodobnostný priestor Ω (pretože $x_{2_n} \to \infty$ a zjednocujeme rastúcu postupnosť). Ich prienik je teda práve prvý člen prieniku a teda môžeme pokračovať vo výpočte:

$$= P(\mathbb{X}_1 < x_1) = F_1(x_1) = F_{\mathbb{X}_1}(x_1)$$

Poznámka 6.2
Podľa vety 6.2, ak poznáme združenú distribučnú funkciu, vieme určiť jednoznačne všetky marginálne distribučné funkcie. Vo všeobecnosti to naopak neplatí; ak poznáme marginálne distribučné funkcie, nevieme jednoznačne skonštruovať združenú distribučnú funkciu. Výnimku tvorí prípad nezávislých náhodných veličín. Ak $P(\mathbb{X}_1 < x_1, \ldots \mathbb{X}_n < x_n) = \prod_{i=1}^{n} P(\mathbb{X}_i < x_i)$, potom

$$F(x_1,\ldots,x_n) = \prod_{i=1}^{n} F_i(x_i).$$

6.2 Diskrétne a absolútne spojité rozdelenie v $\mathbb{R}_2$

Definícia 6.3
Hovoríme, že náhodný vektor $(\mathbb{X},\mathbb{Y})$ má *diskrétne rozdelenie*, ak existujú postupnosti reálnych čísel $\{x_i\}_{i\in I}$, $\{y_j\}_{j\in J}$ a odpovedajúca postupnosť kladných čísel $\{p_{ij}\}_{i\in I, j\in J}$ tak, že platí:

$$p_{ij} = P(\mathbb{X} = x_i, \mathbb{Y} = y_j) \qquad \& \qquad \sum_{i\in I}\sum_{j\in J} p_{ij} = 1$$

a

$$F(x,y) = \sum_{x_i<x}\sum_{y_j<y} p_{ij}$$

Pravdepodobnosť p_{ij} sa nazýva *združený zákon rozdelenia* náhodného vektora $(\mathbb{X},\mathbb{Y})$.

Definícia 6.4
Hovoríme, že náhodný vektor $(\mathbb{X},\mathbb{Y})$ má *absolútne spojité rozdelenie*, ak existuje nezáporná, v $\mathbb{R}_2$ integrovateľná funkcia $f(x,y)$ taká, že platí:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y\,\mathrm{d}x = 1 \quad \& \quad F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\,\mathrm{d}v\,\mathrm{d}u$$

Funkcia $f(x,y)$ sa nazýva *združenou hustotou náhodného vektora* $(\mathbb{X},\mathbb{Y})$ a platí

$$f(x,y) = \frac{\partial F(x,y)}{\partial x \partial y}$$

Veta 6.3
Nech p_{ij} je združený zákon rozdelenia diskrétneho náhodného vektora $(\mathbb{X},\mathbb{Y})$. Potom pre marginálne zákony zložiek platí:

$$p_{i\bullet} = \sum_j p_{ij} \qquad \text{a} \qquad p_{\bullet j} = \sum_i p_{ij}$$

Dôkaz: Bez ujmy na všeobecnosti dokážeme prvý vzťah (druhý vzťah sa dokáže analogicky). Vyjdeme z marginálnej distribučnej funkcie

$$F_1(x) \overset{\text{v. 6.2}}{=\!=} \lim_{y\to\infty} F(x,y) \overset{\text{def.}}{=\!=} \lim_{y\to\infty} \sum_{x_i<x}\sum_{y_i<y} p_{ij} = \sum_{x_i<x} \lim_{y\to\infty} \sum_{y_i<y} p_{ij} =$$

Ak $y \to \infty$ a počítame sumu cez všetky $y_j < y$, je to v podstate to isté ako výpočet sumy pre všetky j. Máme teda

$$F_1(x) = \sum_{x_i<x} \underbrace{\sum_{\forall j} p_{ij}}_{p_{i\bullet}}, \qquad \text{z čoho } p_{i\bullet} = \sum_{\forall j} p_{ij}$$

Veta 6.4
Nech $f(x,y)$ je združenou hustotou spojitého náhodného vektora $(\mathbb{X},\mathbb{Y})$. Potom pre marginálne hustoty zložiek platí:

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y \qquad \text{a} \qquad f_2(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x$$

Dôkaz: Opäť bez ujmy na všeobecnosti dokážeme prvý vzťah a opäť vyjdeme z marginálnej distribučnej funkcie:

$$\begin{aligned} F_1(x) &\overset{\text{v. 6.2}}{=\!=} \lim_{y\to\infty} F(x,y) \overset{\text{def.}}{=\!=} \lim_{y\to\infty} \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\,\mathrm{d}v\,\mathrm{d}u \\ &= \int_{-\infty}^{x} \lim_{y\to\infty} \int_{-\infty}^{y} f(u,v)\,\mathrm{d}v\,\mathrm{d}u = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f(u,v)\,\mathrm{d}v\,\mathrm{d}u \end{aligned}$$

Označme v poslednom dvojnom integráli $f(u) = \int_{-\infty}^{\infty} f(u,v)\,\mathrm{d}v$ a to sa nám hodí do definície, pretože $F_1(x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u$. Stačí už len „premenovať“ premenné vo vyjadrení $f(u)$ a dostaneme požadované tvrdenie. □

Poznámka 6.3
V časti 6.1 sme vybudovali aparát potrebný na dokázanie vety **??**, bod 3.)

$$E(a\mathbb{X} \pm b\mathbb{Y}) = aE(\mathbb{X}) \pm bE(\mathbb{Y})$$

Dôkaz: Podľa vety o prenose integrácie pre funkciu dvoch premenných platí:

$$\begin{aligned} E(\underbrace{a\mathbb{X} \pm b\mathbb{Y}}_{g(\mathbb{X},\mathbb{Y})}) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (ax \pm by)\cdot f(x,y)\,\mathrm{d}y\,\mathrm{d}x \\ &= a\cdot\int_{-\infty}^{\infty} x\left(\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y\right)\mathrm{d}x \pm b\cdot\int_{-\infty}^{\infty} y\left(\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x\right)\mathrm{d}y \\ &\overset{\text{v. 6.3}}{=\!=} a\int_{-\infty}^{\infty} x\cdot f_1(x)\,\mathrm{d}x \pm b\int_{-\infty}^{\infty} y\cdot f_2(y)\,\mathrm{d}y \\ &= a\cdot E(\mathbb{X}) \pm b\cdot E(\mathbb{Y}) \end{aligned}$$

□

6.3 Podmienené rozdelenie v $\mathbb{R}_2$

Definícia 6.5
Nech $(\mathbb{X},\mathbb{Y})$ je diskrétny náhodný vektor so združeným zákonom rozdelenia

$$p_{ij} = P(\mathbb{X} = x_i, \mathbb{Y} = y_j), i \in I, j \in J$$

1. určíme variačné rozpätie $R = x_{\max} - x_{\min}$
2. určíme počet intervalov k (typicky $5 \le k \le 15$)
3. určíme dĺžku intervalu $(a_i, b_i\rangle$: $h \doteq \dfrac{R}{k}$, pričom h vhodne zaokrúhlime nahor (ak by sme zaokrúhľovali nadol, posledná hodnota by nemusela patriť do žiadneho intervalu)
4. zostrojíme tabuľku početností pre intervaly, pričom početnosť intervalu bude počet hodnôt, ktoré padnú do intervalu. Za reprezentanta intervalu berieme (považujeme) stred intervalu x_i^*.
5. nakreslíme histogram (stĺpčeky šírky h)
6. vypočítame charakteristiky znaku

$$\begin{aligned} \bar{x} &= \frac{1}{N}\sum_{i=1}^{k} x_i^* \cdot m_i \\ \widetilde{x}_{\text{kor}} &= a_i + h\cdot\frac{\frac{N}{2} - \sum_{j\le i} m_j}{m_i} \\ \widehat{x}_{\text{kor}} &= x_i^* + \frac{h}{2}\cdot\frac{m_{i+1} - m_{i-1}}{2m_i - m_{i+1} - m_{i-1}} \end{aligned}$$

7.2 Náhodný výber a výberové charakteristiky

Nevýhodou popisnej štatistiky je nutnosť vyčerpávajúceho zisťovania, čo v praxi často znamená potrebu financií, času atď. Rovnako meranie môže v niektorých prípadoch spôsobiť zničenie meraného prvku, čo opäť znemožňuje opakované zisťovanie.
Upustíme teda od tohto spôsobu zisťovania a budeme realizovať náhodné (reprezentatívne) výbery o rozsahu $n \ll N$. Prvky reprezentatívnej vzorky sú nositeľmi hodnôt sledovaného znaku x_i, možno ich považovať za náhodné veličiny $\mathbb{X}_i$.

Definícia 7.1
n-rozmerný náhodný vektor $V_n = (\mathbb{X}_1, \ldots, \mathbb{X}_n)$, kde $\mathbb{X}_i$ pre $i = 1, \ldots, n$ sú nezávislé náhodné veličiny s identickým náhodným rozdelením $F_i(x_i,\theta) \sim F(x,\theta)$, sa nazýva *náhodný výber* o rozsahu n z rozdelenia $F(x,\theta)$.

$$\text{ozn. } V_n \in F(x,\theta)$$

Poznámka 7.2
1. θ je neznámy parameter s hodnotami z parametrického priestoru Θ.
2. Keďže zložky V_n sú nezávislé, pre distribučnú funkciu náhodného vektora platí:

$$F(x_1,\ldots,x_n,\theta) = \prod_{i=1}^{n} F_i(x_i,\theta)$$

Charakteristiky náhodného výberu sú dobrými odhadmi skutočných charakteristík základného štatistického súboru.

1. charakteristika polohy – výberový priemer

$$\bar{\mathbb{X}} = \frac{1}{n}\cdot\sum_{i=1}^{n} \mathbb{X}_i$$

– *medián* $\widetilde{x}$ – prostredná hodnota. Hodnoty usporiadame podľa veľkosti a nájdeme prostrednú hodnotu.

$$\widetilde{x} = \begin{cases} x_{\frac{N+1}{2}} & \text{ak } N \text{ je nepárne} \\ \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2} & \text{ak } N \text{ je párne} \end{cases}$$

V prípade, že N je párne, v podstate „umelo“ vytvoríme prostredný člen.

- Charakteristiky variability štatistického znaku
  - analógia k $D(\mathbb{X})$ – s^2 [s-kvadrát].

$$s^2 = \frac{1}{N}\sum_{i=1}^{K}(x_i - \bar{x})^2 \cdot m_i = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Analogicky podľa výpočtového tvaru $D(\mathbb{X}) = m_2 - m_1^2$ máme

$$s^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2$$

  - analógia k $Q(\mathbb{X})$

$$Q(\mathbb{X}) = \frac{x_{0,75} - x_{0,25}}{2}$$

Hodnoty $x_{0,25}$ a $x_{0,75}$ určíme podobne ako pri mediáne. Dohoda: ak hodnota $\widetilde{x}$ existuje, zarátame ju dvakrát, inak nie.

  - variačné rozpätie

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

  - variačný koeficient znaku x

$$V(x) = \frac{s_x}{\bar{x}} \cdot 100\%$$

Ak $V(x) < 30\%$, hovoríme o dobrej charakteristike. V prípade, že $V(x) > 50\%$, je potrebné použiť iné charakteristiky polohy.

- charakteristiky závislosti 2 znakov. Na každom prvku štatistického súboru sledujeme dva znaky
  - korelačný koeficient

$$r(x,y) = \frac{k(x,y)}{s_x \cdot s_y}, \text{ kde } k(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

Výraz $k(x,y)$ je kovariancia (odhad).

  - rovnica regresnej priamky

$$\begin{aligned} \text{1. regresná priamka:} \quad & y - \bar{y} &=& \; r(x,y) \cdot \frac{s_y}{s_x}(x - \bar{x}) \\ \text{2. regresná priamka:} \quad & x - \bar{x} &=& \; r(x,y) \cdot \frac{s_x}{s_y}(y - \bar{y}) \\ \text{resp. iný tvar 2. regresnej priamky:} \quad & y - \bar{y} &=& \; \frac{1}{r(x,y)} \cdot \frac{s_y}{s_x}(x - \bar{x}) \end{aligned}$$

Poznámka 7.1
Ak je rôznych hodnôt štatistického znaku veľa ($\gg 20$), potom dáta triedime do intervalov typu $\langle a_i, b_i)$ alebo $(a_i, b_i\rangle$ (tak, že každý krajný bod je tam práve raz). Postup:

Potom *podmienené rozdelenie* náhodného vektora $\mathbb{X}$ za predpokladu $\mathbb{Y}$ definujeme ako

$$P(\mathbb{X} = x_i | \mathbb{Y} = y_j) = \frac{P(\mathbb{X} = x_i, \mathbb{Y} = y_j)}{P(\mathbb{Y} = y_j)}, \forall i \in I, j \in J, \quad \text{pričom } P(\mathbb{Y} = y_j) > 0$$

Odpovedajúca podmienená distribučná funkcia je daná vzťahom

$$F(x|y) = \sum_{x_i < x} P(\mathbb{X} = x_i | \mathbb{Y} = y_j)$$

Definícia 6.6
Nech $(\mathbb{X}, \mathbb{Y})$ je spojitý náhodný vektor, ktorého rozdelenie je dané združenou hustotou $f(x,y)$. Potom *podmienená hustota* náhodnej veličiny $\mathbb{X}$ za podmienky $\mathbb{Y}$ definujeme ako

$$f(x|y) = \frac{f(x,y)}{f_2(y)}$$

Odpovedajúca podmienená distribučná funkcia je definovaná vzťahom

$$F(x|y) = \int_{-\infty}^{x} f(t|y)\,\mathrm{d}t$$

Poznámka 6.4
1. Analogicky definujeme funkciu $F(y|x)$.
2. Môžeme definovať aj podmienené charakteristiky podľa vzťahu

$$E(\mathbb{X}^k|\mathbb{Y}) \overset{\text{spoj.}}{=\!=} \int_{-\infty}^{\infty} x^k f(x|y)\,\mathrm{d}x.$$

6.4 Charakteristiky náhodného vektora

Charakteristika polohy náhodného vektora $(\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_n)$ je definovaná ako vektor stredných hodnôt

$$E(\mathbb{X}_1, \ldots, \mathbb{X}_n) = (E(\mathbb{X}_1), \ldots, E(\mathbb{X}_n)).$$

Definícia 6.7
Kovariančnou maticou $K_{\mathbb{X}}$ náhodného vektora $\mathbb{X} = (\mathbb{X}_1, \ldots, \mathbb{X}_n)$ nazývame symetrickú maticu danú prvkami

$$\begin{aligned} K_{ii} &= D(\mathbb{X}_i), \quad \forall i = 1, \ldots, n \\ K_{ij} &= E[(\mathbb{X}_i - E(\mathbb{X}_i))(\mathbb{X}_j - E(\mathbb{X}_j))] = \mathrm{cov}(\mathbb{X}_i, \mathbb{X}_j), \quad \forall i = 1, \ldots, n; i \neq 1 \end{aligned}$$

Číslo $K_{ij} = \mathrm{cov}(\mathbb{X}_i, \mathbb{X}_j)$ nazývame *kovarianciou náhodných vektorov* $\mathbb{X}_i$, $\mathbb{X}_j$.

Definícia 6.8
Korelačnou maticou náhodného vektora $\mathbb{X} = (\mathbb{X}_1, \ldots, \mathbb{X}_n)$ nazývame symetrickú maticu $R_{\mathbb{X}} = (\varrho_{ij})_{i,j=1}^{n}$ s prvkami

$$\begin{aligned} \varrho_{ii} &= 1, \forall i = 1, \ldots, n \\ \varrho_{ij} &= \frac{\mathrm{cov}(\mathbb{X}_i, \mathbb{Y}_j)}{\sqrt{D(\mathbb{X}_i) \cdot D(\mathbb{X}_j)}}, \quad \forall i, j \in 1, \ldots, n, \text{pričom } D(\mathbb{X}_i) > 0, i = 1, \ldots, n \end{aligned}$$

Číslo $\varrho_{ij} = \varrho(\mathbb{X}_i, \mathbb{X}_j)$ sa nazýva *korelačný koeficient náhodných vektorov* $\mathbb{X}_i$, $\mathbb{X}_j$, pričom $i \neq j$.

Poznámka 6.5
1. Disperzia (variancia, rozptyl) je špeciálnym prípadom kovariancie pre $i = j$.

2. Výpočtový tvar kovariancie je

$$\begin{aligned}\operatorname{cov}(\mathbb{X}_i,\mathbb{X}_j) &= E[(\mathbb{X}_i - E(\mathbb{X}_i))(\mathbb{X}_j - E(\mathbb{X}_j))] \\ &= E[\mathbb{X}_i\cdot\mathbb{X}_j - \mathbb{X}_i\cdot\underbrace{E(\mathbb{X}_j)}_{\text{konšt.}} - \mathbb{X}_j\cdot\underbrace{E(\mathbb{X}_i)}_{\text{konšt.}} + E(\mathbb{X}_i)\cdot E(\mathbb{X}_j)] \\ &= E(\mathbb{X}_i\cdot\mathbb{X}_j) - E(\mathbb{X}_i)\cdot E(\mathbb{X}_j) - E(\mathbb{X}_j)\cdot E(\mathbb{X}_i) + E(\mathbb{X}_i)\cdot E(\mathbb{X}_j) \\ &= E(\mathbb{X}_i\cdot\mathbb{X}_j) - E(\mathbb{X}_i)\cdot E(\mathbb{X}_j).\end{aligned}$$

Veta 6.5
Nech náhodné veličiny $\mathbb{X}$, $\mathbb{Y}$ sú nezávislé. Potom platí:

1. $E(\mathbb{X}\cdot\mathbb{Y}) = E(\mathbb{X})\cdot E(\mathbb{Y})$
2. $D(\mathbb{X}\pm\mathbb{Y}) = D(\mathbb{X}) + D(\mathbb{Y})$
3. $\varphi_{\mathbb{X}+\mathbb{Y}}(t) = \varphi_{\mathbb{X}}(t)\cdot\varphi_{\mathbb{Y}}(t)$

Dôkaz:

1. Nech $\mathbb{X}$, $\mathbb{Y}$ sú nezávislé. To je však práve vtedy, ak zodpovedajúce javy sú nezávislé a teda $F(x,y) = F_1(x)\cdot F_2(y)$, ale aj $f(x,y) = f_1(x)\cdot f_2(y)$. Podľa vety o prenose integrácie:

$$E(\underbrace{\mathbb{X}\cdot\mathbb{Y}}_{g(\mathbb{X})}) = \int_{-\infty}^{\infty} x\cdot y\cdot f(x,y)\,\mathrm{d}y\,\mathrm{d}x \overset{\text{nezávislosť}}{=\!=} \int_{-\infty}^{\infty} xy\cdot f_1(x)\cdot f_2(y)\,\mathrm{d}y\,\mathrm{d}x$$

   Podľa vety z matematickej analýzy možno posledný člen napísať ako:

$$\int_{-\infty}^{\infty} x\cdot f_1(x)\,\mathrm{d}x\cdot\int_{-\infty}^{\infty} y\cdot f_2(y)\,\mathrm{d}y = E(\mathbb{X})\cdot E(\mathbb{Y})$$

2. Pre kovarianciu platí, že $\operatorname{cov}(\mathbb{X},\mathbb{Y}) = E(\mathbb{X}\cdot\mathbb{Y}) - E(\mathbb{X})\cdot E(\mathbb{Y}) \overset{\text{a)}}{=\!=} E(\mathbb{X})\cdot E(\mathbb{Y}) - E(\mathbb{X})\cdot E(\mathbb{Y}) = 0$. Podľa vety **??**, bodu 3.) o vlastnostiach disperzie, platí rovnosť $D(\mathbb{X}\pm\mathbb{Y}) = D(\mathbb{X}) + D(\mathbb{Y}) \pm 2\operatorname{cov}(\mathbb{X},\mathbb{Y}) = D(\mathbb{X}) + D(\mathbb{Y})$. Člen $\operatorname{cov}(\mathbb{X},\mathbb{Y})$ je však rovný nule, preto dostávame požadovanú rovnosť.

3. Vyjdime z definície charakteristickej funkcie: $\varphi_{\mathbb{X}+\mathbb{Y}}(t) = E(\mathrm{e}^{it(\mathbb{X}+\mathbb{Y})}) = E(\mathrm{e}^{it\mathbb{X}}\cdot\mathrm{e}^{it\mathbb{Y}})$. Pozrime sa bližšie na obidva členy vo vnútri. Premenné i, t v exponente sú konštanty, náhodné veličiny $\mathbb{X}$ a $\mathbb{Y}$ sú nezávislé. Potom sú ale nezávislé aj členy $it\mathbb{X}$ a $it\mathbb{Y}$ a dokonca aj členy $\mathrm{e}^{it\mathbb{X}}$ a $\mathrm{e}^{it\mathbb{Y}}$. Ďalej podľa už dokázaného bodu 1) máme po úprave $E(\mathrm{e}^{it\mathbb{X}})\cdot E(\mathrm{e}^{it\mathbb{Y}}) = \varphi_{\mathbb{X}}(t)\cdot\varphi_{\mathbb{Y}}(t)$, čo sme chceli dokázať. □

Poznámka 6.6
Poslednú vetu možno zovšeobecniť: strednú hodnotu súčinu nezávislých veličín možno spočítať ako súčin ich stredných hodnôt, disperziu súčtu nezávislých veličín možno vyrátať ako súčet ich disperzií, a charakteristickú funkciu súčtu nezávislých náhodných veličín možno vypočítať ako súčin charakteristických funkcií jednotlivých náhodných veličín. Skrátene zapísané:

$$E\left(\prod_{i=1}^{n}\mathbb{X}_i\right) = \prod_{i=1}^{n}E(\mathbb{X}_i), \qquad D\left(\sum_{i=1}^{n}\mathbb{X}_i\right) = \sum_{i=1}^{n}D(\mathbb{X}_i), \qquad \varphi_{\sum_{i=1}^{n}\mathbb{X}_i}(t) = \prod_{i=1}^{n}\varphi_{\mathbb{X}_i}(t)$$

Veta 6.6 (vlastnosti $\varrho(\mathbb{X},\mathbb{Y})$)
Nech $\varrho(\mathbb{X},\mathbb{Y})$ je korelačný koeficient náhodného vektora $(\mathbb{X},\mathbb{Y})$. Potom platí:

1. *ak $\mathbb{X}$, $\mathbb{Y}$ sú nezávislé, potom $\varrho(\mathbb{X},\mathbb{Y}) = 0$*
2. *ak $\mathbb{X}$, $\mathbb{Y}$ sú lineárne závislé, potom $|\varrho(\mathbb{X},\mathbb{Y})| = 1$*

Tabuľkové spracovanie dát. Usporiadame dáta do neklesajúcej postupnosti

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(N)}$$

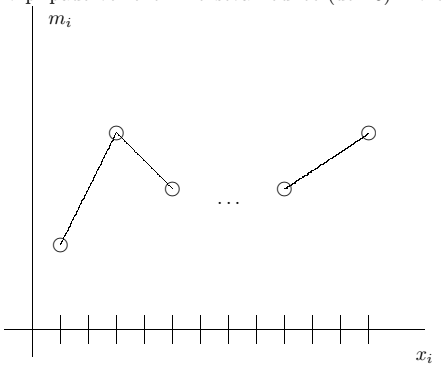
Ak sa údaje opakujú, vzniká *tabuľka početností*, obsahuje absolútne a relatívne početnosti, kumulatívne početnosti a kumulatívne relatívne početnosti.

	x_i	x_1	x_2	$\ldots$	x_K	
absolútna početnosť	m_i	m_1	m_2	$\ldots$	m_K	$\sum_{i=1}^{K} m_i = N$
relatívna početnosť	$\frac{m_i}{N}$	$\frac{m_1}{N}$	$\frac{m_2}{N}$	$\ldots$	$\frac{m_K}{N}$	
kumulatívna početnosť	$\sum_{i=1}^{K} m_i$	m_1	m_1+m_2	$\ldots$	$\sum_{i=1}^{K} m_i$	$\sum_{i=1}^{K} m_i = N$
kumulatívna relatívna početnosť	$\sum_{i=1}^{K} \frac{m_i}{N}$	$\frac{m_1}{N}$	$\frac{m_1+m_2}{N}$	$\ldots$	$\sum_{i=1}^{K} \frac{m_i}{N}$	$\sum_{i=1}^{K} \frac{m_i}{N} = 1$

Keďže podľa zákona veľkých čísel $\frac{\mathrm{m}_i}{N} \to p_i$ pre $n\to\infty$, môžeme definovať *empirickú distribučnú funkciu* (z nameraných hodnôt). Distribučnú funkciu odhadneme z tabuľky kumulatívnych relatívnych početností.

$$F_N(x) = \sum_{x_i<x}\frac{m_i}{N}$$

Grafické metódy. *Polygón* je spojnicový diagram spájajúci (najčastejšie) body $[x_i, m_i]$. *Histogram* je stĺpcový diagram. Používa sa v prípade veľkého množstva hodnôt (do 20) – v tomto prípade zadelíme hodnoty do intervalov.



Výpočtové metódy.

- Charakteristiky polohy štatistického znaku:
  - *aritmetický priemer* $\bar{x}$ – budeme ho používať ako odhad strednej hodnoty.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N}x_i = \frac{1}{N}\sum_{i=1}^{K}x_i\cdot m_i$$

  - *modus* $\widehat{x}$ – najpočetnejšia hodnota znaku x

Časť II
Matematická štatistika

História – korene už v staroveku.

Starovek – sčítanie ľudu a majetku (vojenské a daňové účely) – Egypt, Čína, Mezopotámia

Stredovek – vznik a konsolidácia nových štátov – zisťovanie geografických údajov, hospodársky a politický popis štátu. „status“ = stav štátu.

Novovek

- 17. stor. – „politická aritmetika“ v anglosaských krajinách – Petty, Grand. Vznik zárodkov poisťovníctva a z toho vyplývajúca tvorba úmrtnostných tabuliek (Huygens). Do 20. storočia tzv. popisná štatistika, hlavný princíp je vyčerpávajúce zisťovanie (čím viac údajov, tým lepšie výsledky).
- 20. stor. – využívanie aparátu pravdepodobnosti (v jadre). Vznik matematickej (induktívnej) štatistiky – princíp spočívajúci v náhodnom výbere

7 Popisná štatistika a náhodný výber

7.1 Základné pojmy a metódy

Štatistický súbor – skupina prvkov, ktoré sú predmetom štatistického skúmania a ktoré majú spoločnú vlastnosť. Napr. skupina študentov na prednáške, skupina výrobkov vyrobených na jednom stroji

Rozsah štatistického súboru – počet prvkov štatistického súboru. Označujeme N.

Štatistický znak – sledovaná vlastnosť prvkov. Označujeme x. Napr. váha, výška, vedomosti, farba očí.

Štatistické dáta – namerané hodnoty štatistického znaku Ozn. $x_1, x_2, \ldots, x_N$.

Delenie štatistických znakov

- kvantitatívne – dajú sa jednoznačne číselne vyjadriť
- kvalitatívne – nedajú sa vyjadriť jednoznačne číslom, snaha je ich kvantifikovať

Etapy štatistickej práce

1. štatistické zisťovanie (hromadenie) dát
2. spracovanie štatistických dát
3. vyhodnocovanie výsledkov; záver pre prax

Štatistické zisťovanie (hromadenie) dát. Štatisticky sa zisťujú dáta, je potrebná dôkladná evidencia. Získame východzie dáta $x_1, x_2, \ldots, x_N$.

Spracovanie štatistických dát. Spracovanie sa koná troma spôsobmi:

1. tabuľkové
2. grafické
3. výpočtové

3. ak $\mathbb{X}$, $\mathbb{Y}$ sú ľubovoľné náhodné veličiny, tak $|\varrho(\mathbb{X},\mathbb{Y})| \le 1$

Dôkaz:

1. Ak $\mathbb{X}$, $\mathbb{Y}$ sú nezávislé, potom podľa vety 6.5, bodu 1) platí, že $E(\mathbb{X}\cdot\mathbb{Y}) = E(\mathbb{X})\cdot E(\mathbb{Y})$. Pre kovarianciu platí: $\operatorname{cov}(\mathbb{X},\mathbb{Y}) = \underbrace{E(\mathbb{X}\cdot\mathbb{Y})}_{E(\mathbb{X})\cdot E(\mathbb{Y})} - E(\mathbb{X})\cdot E(\mathbb{Y}) = 0$. Ale potom $\varrho(\mathbb{X},\mathbb{Y}) = \frac{\operatorname{cov}(\mathbb{X},\mathbb{Y})}{\sqrt{D(\mathbb{X})\cdot D(\mathbb{Y})}} = 0$.

2. Nech $\mathbb{X}$, $\mathbb{Y}$ sú lineárne závislé. Bez ujmy na všeobecnosti môžeme predpokladať, že $\mathbb{Y} = a\mathbb{X}+b$. Počítajme:

$$\varrho(\mathbb{X},\mathbb{Y}) \overset{\text{z výp. tvaru}}{=\!=} \frac{E(\mathbb{X}\cdot\mathbb{Y}) - E(\mathbb{X})\cdot E(\mathbb{Y})}{\sqrt{D(\mathbb{X})\cdot D(\mathbb{Y})}} = \frac{E[\mathbb{X}\cdot(a\mathbb{X}+b)] - E(\mathbb{X})\cdot E(a\mathbb{X}+b)}{\sqrt{D(\mathbb{X})\cdot D(a\mathbb{X}+b)}}$$

V ďalšom kroku využijeme vlastnosti strednej hodnoty a disperzie. Ďalej si všimneme člen $D(a\mathbb{X}+b)$ v menovateli. Veličiny $a\mathbb{X}$ a b sú zrejme nezávislé, môžeme teda použiť vetu 6.5, pričom však $D(b) = 0$. Teda

$$\varrho(\mathbb{X},\mathbb{Y}) \overset{\text{vl. E, D}}{=\!=} \frac{aE(\mathbb{X}^2) + b\cdot E(\mathbb{X}) - a(E(\mathbb{X}))^2 - b\cdot E(\mathbb{X})}{\sqrt{D(\mathbb{X})\cdot a^2 D(\mathbb{X})}}$$

Teraz odmocníme členy v menovateli. Keďže $D(\mathbb{X}) > 0$, máme po vynásobením oboch $D(\mathbb{X})$ a ich následnom odmocnení člen $D(\mathbb{X})$. Konštanta a^2 po odmocnení nám však dá $|a|$. V čitateli nám po sčítaní vypadnú členy $b\cdot E(\mathbb{X})$ a zo zvyšných dvoch členov vyjmeme a pred zátvorku.

$$\varrho(\mathbb{X},\mathbb{Y}) \;=\; \frac{aE(\mathbb{X}^2) - (E(\mathbb{X}))^2}{|a|D(\mathbb{X})} = \frac{a\cdot D(\mathbb{X})}{|a|\cdot D(\mathbb{X})} = \frac{a}{|a|} = \begin{cases} 1 & \text{ak } a > 0 \\ -1 & \text{ak } a < 0 \end{cases}$$

Z posledných alternatív teda vyplýva, že $|\varrho(\mathbb{X},\mathbb{Y})| = 1$.

3. Nech $\mathbb{X}$, $\mathbb{Y}$ sú ľubovoľné. Uvažujme výraz $E[t(\mathbb{X}-E(\mathbb{X})) + (\mathbb{Y}-E(\mathbb{Y}))]^2 \ge 0$. Ak výraz vo vnútri strednej hodnoty nebude záporný, tak aj stredná hodnota bude nezáporná. Upravujme postupne tento výraz:

$$\begin{aligned} E[t^2E(\mathbb{X}-E(\mathbb{X}))^2 + (\mathbb{X}-E(\mathbb{Y}))^2 + 2t(\mathbb{X}-E(\mathbb{X}))(\mathbb{Y}-E(\mathbb{Y}))] &\ge 0 \\ t^2D(\mathbb{X}) + D(\mathbb{Y}) + 2t\operatorname{cov}(\mathbb{X},\mathbb{Y}) &\ge 0 \end{aligned}$$

Uvažujme kvadratickú rovnicu s premennou t. Počítajme diskriminant za predpokladu, že rovnica má najviac jeden reálny koreň.

$$\begin{aligned} 4\operatorname{cov}^2(\mathbb{X},\mathbb{Y}) - 4\cdot D(\mathbb{X})\cdot E(\mathbb{Y}) &\le 0 \\ \operatorname{cov}^2(\mathbb{X},\mathbb{Y}) &\le D(\mathbb{X})\cdot D(\mathbb{Y}), \qquad (6.2) \\ &\text{a} \\ \operatorname{cov}^2(\mathbb{X},\mathbb{Y}) &\ge 0 \qquad (6.3) \end{aligned}$$

Postupne predelíme zlomok (výraz je nezáporný – vyplýva to z nerovností (6.2) a (6.2)), aby sme na pravej strane získali 1 a odmocníme s ohľadom na absolútne hodnoty:

$$\begin{aligned} 0 \le \;& \tfrac{\operatorname{cov}^2(\mathbb{X},\mathbb{Y})}{D(\mathbb{X})\cdot D(\mathbb{Y})} && \le 1 \\ 0 \le \;& \tfrac{|\operatorname{cov}(\mathbb{X},\mathbb{Y})|}{\sqrt{D(\mathbb{X})\cdot D(\mathbb{Y})}} && \le 1 \\ 0 \le \;& \left|\tfrac{\operatorname{cov}(\mathbb{X},\mathbb{Y})}{\sqrt{D(\mathbb{X})\cdot D(\mathbb{Y})}}\right| && \le 1 \\ 0 \le \;& |\varrho(\mathbb{X},\mathbb{Y})| && \le 1 \end{aligned}$$

□

Poznámka 6.7

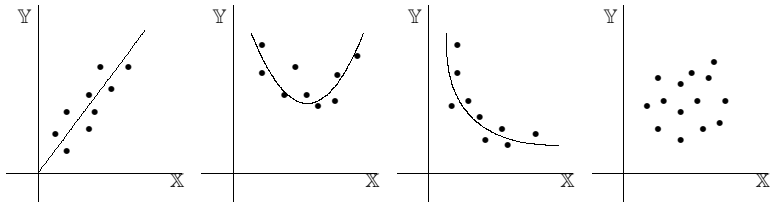
1. K tvrdeniu „ak $\mathbb{X}$, $\mathbb{Y}$ sú nezávislé, potom $\varrho(\mathbb{X},\mathbb{Y})=0$“ obrátená veta neplatí.
2. K tvrdeniu „ak $\mathbb{X}$, $\mathbb{Y}$ sú lineárne závislé, potom $|\varrho(\mathbb{X},\mathbb{Y})|=1$“ platí aj opačné tvrdenie.

Korelačný koeficient $\varrho(\mathbb{X},\mathbb{Y})$ je mierou (lineárnej) závislosti. V matematickej štatistike sa používa:

- ak $|\varrho(\mathbb{X},\mathbb{Y})|>0{,}8$, hovorieva sa o silnej (lineárnej) závislosti,
- ak $0{,}3\le|\varrho(\mathbb{X},\mathbb{Y})|\le 0{,}8$ ide o miernu (lineárnu) závislosť,
- ak $|\varrho(\mathbb{X},\mathbb{Y})|<0{,}3$, hovoríme o slabej (lineárnej) závislosti.

6.5 Regresia ako trend závislosti

Trend (smer) závislosti náhodných veličín $\mathbb{X}$, $\mathbb{Y}$ sa dá graficky znázorniť tzv. regresnou čiarou.



Obr. 10: Lineárna, parabolická, hyperbolická závislosť a nezávislosť

V praxi sa najčastejšie používa lineárna závislosť, ktorej zodpovedá *regresná priamka*[4].

Definícia 6.9
Regresnou priamkou závislosti $\mathbb{Y}$ na $\mathbb{X}$ (1. regresnou priamkou) nazývame priamku

$$y=ax+b,$$

kde koeficienty a, b spĺňajú podmienku

$$E[\mathbb{Y}-(a\mathbb{X}+b)] \text{ je minimálne}$$

(koeficienty minimalizujú strednú kvadratickú odchýlku). Koeficienty a, b nazývame *regresnými koeficientami*.

Veta 6.7
Nech $\varrho(\mathbb{X},\mathbb{Y})$ je korelačný koeficient náhodných veličín $\mathbb{X}$, $\mathbb{Y}$. Pre 1. regresnú priamku závislosti $\mathbb{Y}$ na $\mathbb{X}$ platí:

$$y-E(\mathbb{Y})=\varrho(\mathbb{X},\mathbb{Y})\cdot\sqrt{\frac{D(\mathbb{Y})}{D(\mathbb{X})}}\cdot(x-E(\mathbb{X}))$$

Potom $a=\varrho(\mathbb{X},\mathbb{Y})\cdot\sqrt{\frac{D(\mathbb{Y})}{D(\mathbb{X})}}$, $b=E(\mathbb{Y})-aE(\mathbb{X})$.

Dôkaz: Dôkaz vykonáme použitím metódy najmenších štvorcov. Chceme minimalizovať výraz $S(a,b)=E(\mathbb{Y}-(a\mathbb{X}+b))^2$. Má platiť:

[4] Dá sa jednoduchšie popísať ako napr. hyperbola. V okolí hyperboly vieme často priamkou dobre aproximovať.

- $\frac{\partial S(a,b)}{a}=0$ (stacionárny bod)
- $\frac{\partial S(a,b)}{b}=0$ (stacionárny bod)
- 2. diferenciál má byť kladný

Upravme najprv

$$\begin{aligned}
S(a,b) &= E[(\mathbb{Y}-E(\mathbb{Y}))-a(\mathbb{X}-E(\mathbb{X}))+\overbrace{(E(\mathbb{Y})-aE(\mathbb{X})-b)}^{\Delta-\text{konšt.}}]^2\\
&= E\big[(\mathbb{Y}-E(\mathbb{Y}))^2+a^2(\mathbb{X}-E(\mathbb{X}))^2+\Delta^2+\\
&\quad -2a(\mathbb{X}-E(\mathbb{X}))(\mathbb{Y}-E(\mathbb{Y}))+2\Delta(\mathbb{Y}-E(\mathbb{Y}))-2a\Delta(\mathbb{X}-E(\mathbb{X}))\big]\\
\text{Aplikujme } E: &= D(\mathbb{Y})+a^2D(\mathbb{X})+(E(\mathbb{Y})-aE(\mathbb{X})-b)^2-2a\operatorname{cov}(\mathbb{X},\mathbb{Y})+0
\end{aligned}$$

Počítajme parciálnu deriváciu podľa a:

$$\begin{aligned}
\frac{\partial S(a,b)}{\partial a} &= 2aD(\mathbb{X})+2(E(\mathbb{Y})-aE(\mathbb{X})-b)\cdot -E(\mathbb{X})-2\operatorname{cov}(\mathbb{X},\mathbb{Y})\\
\frac{\partial S(a,b)}{\partial b} &= 2(E(\mathbb{Y})-aE(\mathbb{X})-b)(-1)
\end{aligned}$$

Obe parciálne derivácie položíme rovné 0, teda ich môžeme upraviť

$$\begin{aligned}
2aD(\mathbb{X})+2(E(\mathbb{Y})-aE(\mathbb{X})-b)\cdot -E(\mathbb{X})-2\operatorname{cov}(\mathbb{X},\mathbb{Y}) &= 2(E(\mathbb{Y})-aE(\mathbb{X})-b)(-1)\\
&\overset{\text{a}}{}\\
a[D(\mathbb{X})+E^2(\mathbb{X})]+bE(\mathbb{X}) &= \operatorname{cov}(\mathbb{X},\mathbb{Y})+E(\mathbb{X})\cdot E(\mathbb{Y})
\end{aligned}$$

Postupnou úpravou dostaneme

$$\begin{aligned}
a &= \frac{\operatorname{cov}(\mathbb{X},\mathbb{Y})}{D(\mathbb{X})}=\frac{\operatorname{cov}(\mathbb{X},\mathbb{Y})}{\sqrt{D(\mathbb{X})\cdot D(\mathbb{Y})}}\cdot\sqrt{\frac{D(\mathbb{Y})}{D(\mathbb{X})}}=\varrho(\mathbb{X},\mathbb{Y})\cdot\sqrt{\frac{D(\mathbb{Y})}{D(\mathbb{X})}}\\
b &= E(\mathbb{Y})-a\cdot E(\mathbb{X})
\end{aligned}$$

Potrebujeme však ešte overiť, že druhý diferenciál je naozaj kladný. Túto úlohu však prenechávame na čitateľa. □