PRAVDEPODOBNOSŤ a ŠTATISTIKA

Štatistický súbor a náhodný výber

doc. RNDr. Štefan Peško, CSc.

Katedra matematických metód a operačnej analýzy, FRI ŽU

22. apríla 2019

Motivácia

Štatistika skúma náhodné udalosti na rozsiahlom súbore prípadov a hľadá tie vlastnosti, ktoré sa prejavia vo veľkom súbore, nie v jednotlivých prípadoch.

V praxi potrebujeme analyzovať náhodné premenné, pričom ich rozdelenie nie je úplne známe.

Často predpokladáme určitý typ rozdelenia, ale nepoznáme hodnoty parametrov. Jediným spôsobom ako tieto informácie doplniť je vychádzať z výsledkov opakovania pokusov pri rovnakých podmienkach. Hľadané parametre sú potom funkciami týchto výsledkov.

Štatistický súbor

Základným pojmom matematickej štatistiky je štatistický súbor, ktorým rozumieme konečnú neprázdnu množinu prvkov (predmetov alebo jednotiek), ktoré majú z daného hľadiska určité spoločné vlastnosti.

Na prvkoch štatistického súboru, štatistických jednotkách, sledujeme štatistické znaky, napr. farbu očí, národnosť, pohlavie – kvalitatívny znak, hmotnosť, výšku, vek dĺžku – kvantitatívny znak.

Poznámka 1:
V štatistickom súbore sa môžu, na rozdiel od množiny hodnôt, hodnoty štatistického znaku opakovať.

Poznámka 2:
Pri výpočtoch budeme využívať vstavané programy a datasety štatistických súborov v štatistickej knižnici *scipy.stats* v jazyku *python*.

Miery polohy a variability

Predpokladajme, že sme na n štatistických jednotkách namerali súbor hodnôt

$$x_1, x_2, \ldots, x_n$$

daného štatistického znaku. Celkovému počtu prvkov súboru n budeme hovoriť rozsah súboru.

Aritmetický priemer a rozptyl znakov v súbore hodnôt definujeme vzťahmi

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s_*^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{1}$$

Namiesto rozptylu s_*^2 je niekedy vhodnejšie používať rozptyl

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

Ak sú niektoré hodnoty znaku $x_1, x_2, \ldots, x_n$ rovnaké, má zmysel ich napísať do tabuľky rozdelenia početností

x	x_1	$\ldots$	x_i	$\ldots$	x_k
$f(x)$	f_1	$\ldots$	f_i	$\ldots$	f_k

Tabuľka 1: Tabuľka rozdelenia početností k znakov

kde f_i je (absolútna) početnosť znaku x_i.

Potom aritmetický priemer a rozptyl vypočítame pomocou vzťahov

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i f_i, \quad s_*^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 f_i. \tag{3}$$

Empirická distribučná funkcia

Hodnota znaku	x_1	$\ldots$	x_i	$\ldots$	x_k
Početnosť znaku	f_1	$\ldots$	f_i	$\ldots$	f_k
Relatívna početnosť	$\frac{f_1}{n}$	$\ldots$	$\frac{f_i}{n}$	$\ldots$	$\frac{f_k}{n}$
Kumulatívna početnosť	f_1	$\ldots$	$\sum_{t=1}^{i} f_t$	$\ldots$	$\sum_{t=1}^{k} f_t$
Kumulatívna realtívna početnosť	$\frac{f_1}{n}$	$\ldots$	$\sum_{t=1}^{i} \frac{f_t}{n}$	$\ldots$	$\sum_{t=1}^{k} \frac{f_t}{n}$

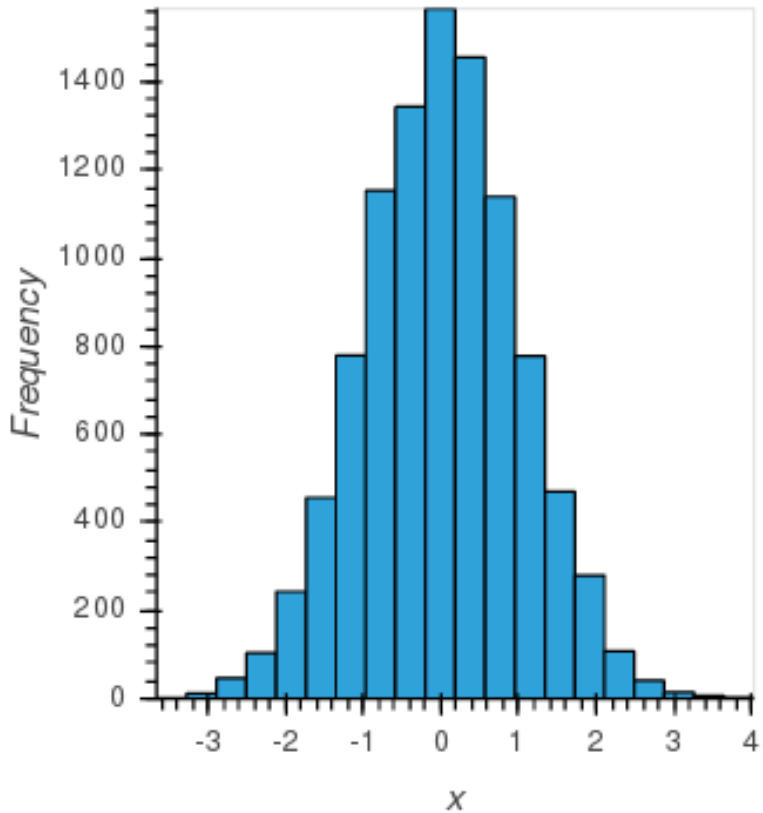
Tabuľka 2: Rozšírená tabuľka rozdelenia početností hodnôt znaku

Interpretácia výsledkov v tabuľke je jednoduchá, napr. relatívna početnosť f_i/n je odhadom pp. hodnoty štatistického znaku x_i. Empirickú distribučnú funkciu $F_n(x)$ definujeme pre $x \in \Re$ vzťahom

$$F_n(x) = \frac{\text{počet takých } x_i, \text{ kde } x_i < x}{n}. \qquad (4)$$

Histogram

Histogram početnosti je stĺpcový diagram so stĺpcami, ktorých základňa sa rovná šírke intervalu a výška i-tého stĺpca sa rovná početnosti $f_i, i = 1, 2, \ldots, k$ resp. relatívnej početnosti. Počet intervalov volíme obvykle $8, 9, \ldots, 20$ pričom za reprezentanta hodnôt znaku x_i volíme stred intervalu.



Príklad 8.1 (Skúška z Algebry)

Štatistický súbor je tvorený študentami, ktorí absolvovali skúšku z algebry. Bolo vykonaných 166 skúšok, niektorí študenti opakovali skúšku viac krát.
Pozorovaným znakom je známka A, B, C, D, E, FX, ktorá je kvalitatívnym znakom. Jednotlivým známka možno priradiť číselnú hodnotu, kvantitatívny znak X.

Známka	A	B	C	D	E	FX	$\sum$
Hodnota	1	1.5	2	2.5	3	4	
Početnosť	1	3	11	14	60	77	166

Tabuľka 3: Výsledky z 8 termínov skúšky z algebry v ZS 2015

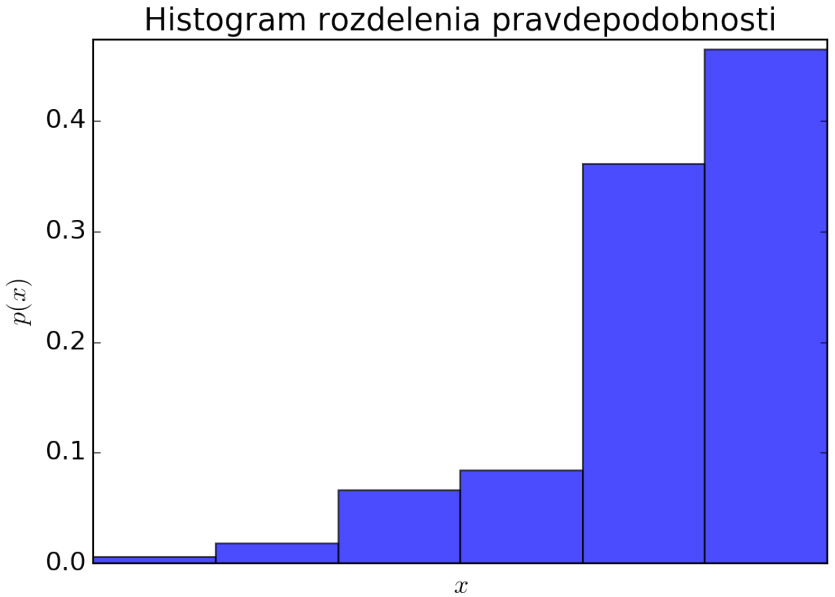
Vytvorme rozšírenú tabuľku početnosti, histogram relatívnych početností (pravdepodobností), empirickú distribučnú funkciu, aritmetický priemer a rozptyl.

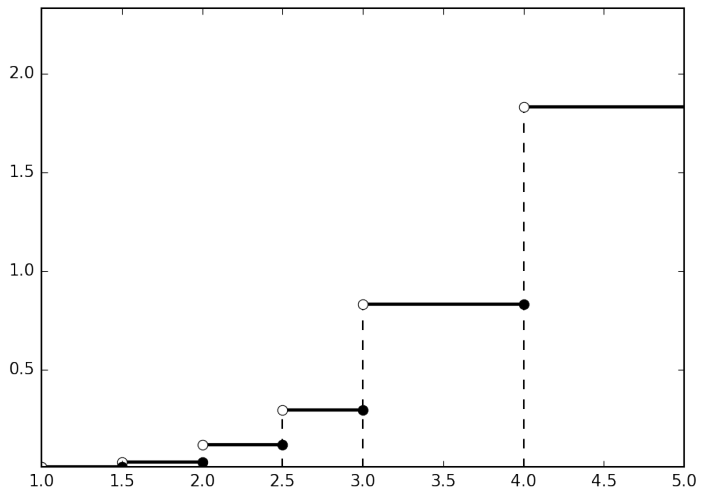
Rozsah súboru je $n = 166$, počet znakov $k = 6$.

Známka	A	B	C	D	E	FX	$\sum$
i	1	2	3	4	5	6	
x_i	1	1.5	2	2.5	3	4	
f_i	1	3	11	14	60	77	166
f_i/n	0.006	0.018	0.066	0.084	0.361	0.464	1
$\sum_{t=1}^{i} f_t$	1	4	15	29	89	166	
$\sum_{t=1}^{i} f_t/n$	0.006	0.024	0.090	0.175	0.536	1	

Tabuľka 4: Rozšírená tabuľka početností hodnotenia študentov

V súbore vyskúšaných študentov je aritmetický priemer známok $\overline{x} = 3.316$ a rozptyl je $s_*^2 = 0.516$.


Histogram rozdelenia pravdepodobnosti
0.4
0.3
0.2
0.1
0.0
$p(x)$
x


2.0
1.5
1.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

Medián a modus

Ak jednotlivé hodnoty znakov usporiadame do neklesajúcej postupnosti

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

dostaneme usporiadaný súbor hodnôt. Indexy v zátvorkách udávajú poradie zistených hodnôt.

Mediánom usporiadaného súboru n hodnôt rozumieme hodnotu

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{ak je } n \text{ nepárne,} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{ak je } n \text{ párne.} \end{cases} \tag{5}$$

Jedná sa o hodnotu, ktorá rozdelí hodnoty usporiadaného súboru na dve rovnako veľké časti.

Modusom súboru n hodnôt znaku rozumieme jeho najpočetnejšiu hodnotu. Nemusí byť určený jednoznačne. Má zmysel ak je počet znakov k podstatne menší než rozsah súboru n.

Základný a výberový súbor

Uvažujme veľký štatistický súbor, ktorý budeme nazývať základný súbor (tiež populácia). Na každom z jeho N jednotiek môžeme zmerať hodnotu zvoleného číselného znaku X, čím by sme získali hodnoty $x_1, x_2, \ldots, x_N$. Aritmetický priemer hodnôt znaku na celej populácii označíme μ a populačný rozptyl značíme σ^2 vypočítame podľa (1)

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2.$$

Základný súbor je však tak veľký, že nie je možné alebo hospodárne zistiť hodnotu každej jeho štatistickej jednotky. Preto náhodne vyberieme skupinu $n, (n \ll N)$ štatistických jednotiek a zistíme hodnoty sledovaného znaku len pre tieto jednotky. Tento výber rozsahu n nazývame výberový súbor. Musí byť taký, aby dobre reprezentoval celý základný súbor.

Náhodný výber

Urobíme taký konečný výber zo základného súboru X s vracaním, pri ktorom vybraný prvok, po zistení jeho hodnoty, do populácie vraciame a až potom znovu vyberáme. Výber robíme tak, aby mal každý prvok z populácie rovnakú pp. $1/N$.

Označme hodnoty na takto vybraných prvkoch $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$. Sú to nezávislé náhodné premenné, vyberané z tej istej populácie, pre ktoré platí

$$E(\mathbf{X}_j) = \mu, \; D(\mathbf{X}_j) = \sigma^2 \quad j = 1, 2, \ldots, n.$$

Možno predpokladať, že v takomto náhodnom vektore $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ nezávislých náhodných premenných, majú premenné rovnaké rozdelenie dané distribučnou funkciou $F_{\mathbf{X}}(x)$. Takúto n-ticu náhodných premenných nazývame náhodný výber rozsahu n.

Funkciu jednej alebo viac náhodných premenných, ktorá nezávisí na hodnotách neznámych parametrov, sa nazývame štatistika.

Nech je $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ náhodný výber zo základného súboru X. Potom výberovým priemerom rozumieme štatistiku

$$\overline{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i. \tag{6}$$

Poznámka:
Náhodná premenná $\mathbf{Y} = \frac{\mathbf{X}_1 - \mu}{\sigma}$ je štatistikou, len ak sú známe hodnoty μ a σ.

Tvrdenie 20

Pre výberový priemer vypočítaný z náhodného výberu $\mathbf{X}$ rozsahu n s konečnou strednou hodnotou μ a konečným rozptylom σ^2 platí

$$E(\overline{\mathbf{X}}) = \mu, \quad D(\overline{\mathbf{X}}) = \frac{\sigma^2}{n}. \tag{7}$$

Dôkaz:
Z vlastnosti strednej hodnoty súčtu náhodných premenných je

$$E(\overline{\mathbf{X}}) = E\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\Big) = \frac{1}{n}\sum_{i=1}^{n}E(\mathbf{X}_i) = \mu.$$

Náhodné premenné $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ sú nezávislé a tak z vlastnosti disperzie súčtu je

$$D(\overline{\mathbf{X}}) = D\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\Big) = \frac{1}{n^2}\sum_{i=1}^{n}D(\mathbf{X}_i) = \frac{\sigma^2}{n}.$$

Výberový rozptyl

Nech je $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ náhodný výber zo základného súboru X. Potom výberovým rozptylom rozumieme štatistiku

$$\mathbf{S}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}})^2.$$

Poznámka:

Rozptyl výberového priemeru je pre $n > 1$ v prípade konečného základného súboru menší než v prípade rovnako veľkého výberu z nekonečného základného súboru. Ak je konečný základný súbor podstatne väčší než výberový súbor, považuje sa za súbor nekonečný.

Tvrdenie 21

Pre výberový rozptyl vypočítaný z náhodného výberu $\mathbf{X}$ rozsahu n s konečnou strednou hodnotou μ a konečným rozptylom σ^2 platí

$$E(\mathbf{S}^2) = \sigma^2. \tag{8}$$

Dôkaz:

Jednoduchou úpravou dostaneme

$$\sum_{i=1}^{n}(\mathbf{X}_i - \mu)^2 = \sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}} + \overline{\mathbf{X}} - \mu)^2 = \sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})^2 + n(\overline{\mathbf{X}} - \mu)^2.$$

Odtiaľ je

$$E(S^2) = E\Big(\frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \mu)^2\Big) - E\Big(\frac{n}{n-1}\Big(\overline{\mathbf{X}} - \mu)^2\Big)$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} D(\mathbf{X}_i) - E\Big(\frac{n}{n-1}D(\overline{\mathbf{X}})\Big) = \frac{n}{n-1}\sigma^2 - \frac{n}{(n-1)n}\sigma^2$$

$$= \sigma^2.$$

Rozdelenia odvodené od normálneho

Majme n nezávislých náhodných premenných $\mathbf{X}_i \sim N(0,1)$, potom súčet $\mathbf{Y} = \sum_{i=1}^{n} \mathbf{X}_i^2$ je náhodnou premennou s hustotu pp.

$$f_{\mathbf{Y}}(y) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}, \quad x > 0.$$

Hovoríme, že $\mathbf{Y}$ má χ^2 rozdelenie s n stupňami voľnosti, čo značíme $\mathbf{Y} \sim \chi^2(n)$. Platí $E(\mathbf{Y}) = n, D(\mathbf{Y}) = 2n$.

Poznámka:
Gamma funkcia sa zavádza pre $a > 0$ vzťahom

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} \mathrm{d}x.$$

Pri výpočtoch vystačíme zo znalosťou $\Gamma(1/2) = \sqrt{\pi}, \Gamma(1) = 1$, čo spolu s vlastnosťou $\Gamma(a+1) = a\Gamma(a)$ znamená, že Γ-funkcia je zovšeobecnením pojmu faktoriál $\Gamma(n) = (n-1)!$, $n \in \mathcal{N}$.

Uvažujeme náhodnú premennú $\mathbf{T}$, ktorá je daná podielom nezávislých premenných

$$\mathbf{T} = \frac{\mathbf{Z}}{\sqrt{\frac{\mathbf{X}}{n}}}, \tag{9}$$

kde $\mathbf{Z} \sim N(0,1)$ a $\mathbf{T} \sim \chi^2(n)$. Náhodná premenná T s hustota pp.

$$f_{\mathbf{T}}(t) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(n)\sqrt{\pi n}}\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \ t \in \Re,$$

má t (Studentovo) rozdelenie s n stupňami voľnosti, čo značíme $\mathbf{T} \sim t(n)$. Platí $E(\mathbf{T}) = 0$ pre $n > 1$ a $D(\mathbf{T}) = \frac{n}{n-2}$ pre $n > 2$.

Poznámka:

So Studentovým rozdelením sa stretneme pri intervalovom odhade neznámeho parametru a pri testovaní štatistických hypotéz.

Majme dve nezávislé náhodné premenné $\mathbf{X} \sim \chi^2(k), k \in \mathcal{N}$ a $\mathbf{Y} \sim \chi^2(m), m \in \mathcal{N}$. Potom náhodná premenná

$$\mathbf{U} = \frac{\frac{\mathbf{X}}{k}}{\frac{\mathbf{X}}{m}}, \tag{10}$$

s hustotou pravdepodobnosti

$$f_{\mathbf{U}}(u) = \frac{\Gamma(\frac{k+m}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{m}{2})} \left(\frac{k}{m}\right)^{\frac{k}{2}} u^{\frac{k}{2}-1} \left(1 + \frac{k}{m}u\right)^{\frac{k+m}{2}}, \ u > 0,$$

má *F* (Fisher-Snedecerovo) rozdelenie so stupňami voľnosti k a m.

Poznámka:
Hodnoty $100p\%$ kvantilov $\chi^2_p(n), t_p(n), F_p(k, m)$ sú tabelované v štatistických tabuľkách, dnes sa na výpočet používa štatistický softvér.

Tvrdenie 22

Pre náhodný výber $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ z normálneho rozdelenia $N(\mu, \sigma^2)$ platí

a) $\overline{\mathbf{X}}$ a $\mathbf{S}^2$ sú nezávislé náhodné premenné,

b)

$$\overline{\mathbf{X}} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \tag{11}$$

c)

$$\frac{(n-1)\mathbf{S}^2}{\sigma^2} \sim \chi^2(n-1). \tag{12}$$

Príklad 8.6

Firma uvádza životnosť autobatérie 5 ± 0.5 roka. Pri overovaní kavality bolo testovaných 20 batérii. Aká je pp., že smerodajná odchylka testovaných batérií bude viac než 7 mesiacov?

V testovanom (náhodnom) výbere $(\mathbf{X}_1, \ldots, \mathbf{X}_n), n = 20$ je sledovaným znakom životnosť autobatérie. Predpokladá sa $\mu = E(\mathbf{X}) = 62$ mesiacov a $\sigma = \sqrt{D(\mathbf{X})} = 6$ mesiacov.

Potrebujeme zistiť $\mathcal{P}(\mathbf{S} > 7)$, no nepoznáme pp. rozdelenie $\mathbf{S}$ a tak predpokladáme, že životnosť autobatérii má normálne rozdelenie $\mathbf{X} \sim N(\mu, \sigma^2)$. Z vlastností (12) vieme $\frac{(n-1)\mathbf{S}^2}{\sigma^2} \sim \chi^2(n-1)$ a tak premenná $\mathbf{Y} = \frac{19S^2}{36} \sim \chi^2(19)$. A preto dostaneme

$$\mathcal{P}(\mathbf{S} > 7) = \mathcal{P}\left(\mathbf{Y} > \frac{19 \cdot 7^2}{36}\right) = \mathcal{P}(\mathbf{Y} > 25.86)$$
$$= 1 - F_{\chi^2(19)}(25.86) = 0.134.$$

Naviac vieme, že firma vyrába autobatérie v dvoch prevádzkach, pričom sa predpokladá, že batérie majú porovnateľnú životnosť t.j. nezáleží, kde boli vyrobené. Pre overenie kvality bolo testovaných 20 batérii z 1. prevádzky a 30 batérii z 2. prevádzky. Aká je pp., že pri vzorke z 1. prevádzky bude zistených viac než dvojnásobný rozptyl než pri vzorke z 2. prevádzky?

Označme $\mathbf{S}_i^2$ zistený rozptyl na vzorke batérií z i-tej prevádzky. Môžeme predpokladať, že obe vzorky sú z to istého normálneho rozdelenie $N(\mu, \sigma^2)$. Platí

$$\frac{\mathbf{S}_1^2}{\mathbf{S}_2^2} = \frac{\frac{(n_1-1)\mathbf{S}_1^2}{\sigma^2}}{\frac{(n_2-1)\mathbf{S}_2^2}{\sigma^2}} \sim F(n_1-1, n_2-1).$$

A preto dostaneme pre $n_1 = 20, n_2 = 30$

$$\mathcal{P}\left(\frac{\mathbf{S}_1^2}{\mathbf{S}_2^2} > 2\right) = 1 - F_{F(19,29)}(2) \approx 0.045.$$

Cvičenie

8.1 Nasledujúce hodnoty udávajú rozmery v mm, ktoré sme namerali na 20-tich súčiastkách: 28.6, 28.2, 28.5, 28.25, 28.6, 28.2, 28.5, 28.6, 28.5, 28.5, 28.6, 28.2, 28.5, 28.25, 28.6, 28.3, 28.25, 28.2, 28.5, 28.5. a) Vypočítajte aritmetický priemer, rozptyl, medián a modus, b) zostrojte rozšírenú tabuľku rozdelenia početnosti c) zostrojte histogram, polygón a empirickú distribučnú funkciu.

8.2 Nech $(X_1, X_2, \ldots, X_9)$ je náhodný výber z rozdelenia $N(2, 4)$. Overte na náhodne generovaných príkladoch, či výberový priemer $\overline{X}$ má hustotu pp. $f(x) = \frac{3}{2\sqrt{\pi}} e^{\frac{-9(\overline{x}-2)^2}{8}}$.

8.3* Vytvorte tabuĺku vybraných kvantilov a) $N(0, 1)$ pre z_α, b) χ^2 pre $\chi^2_\alpha(n)$, c) t pre $t_\alpha(n)$, d) F pre $F_\alpha(k, m)$.

8.4* Nech $(X_1, X_2, \ldots, X_n)$ je náhodný výber z rozdelenia $N(\mu, \sigma^2)$. Dokážte, že $\mathbb{D}(S^2) = \frac{2\sigma^4}{n-1}$.