6 Číselné charakteristiky náhodných veličin, centrální limitní věta

6.1 Číselné charakteristiky náhodných veličin alespoň ordinálního typu

6.1.1 Charakteristika polohy

Číslo $K_\alpha(X)$ se nazývá α-kvantil náhodné veličiny X, jestliže splňuje nerovnosti: $\Pr(X \leq K_\alpha(X)) \geq \alpha \wedge \Pr(X \geq K_\alpha(X)) \geq 1-\alpha$. Přitom $\alpha \in (0;1)$. Jde o teoretický protějšek kvantilu x_α zavedeného v popisné statistice (viz kapitola 3). Pro některá vybraná α jsou názvy kvantilů v počtu pravděpodobnosti stejné jako v popisné statistice. Vzhledem k tomu, že kvantily diskrétních náhodných veličin nejsou určeny jednoznačně, budeme se dále zabývat jen kvantily spojitých náhodných veličin. Pro spojitou náhodnou veličinu X platí: $\alpha = F(K_\alpha(X)) = \int_{-\infty}^{K_\alpha(X)} f(x)\mathrm{d}x$.

Označení pro kvantily speciálních rozdělení

- $X \sim N(0,1) \Rightarrow K_\alpha(X) = u_\alpha$,
- $X \sim t(n) \Rightarrow K_\alpha(X) = t_\alpha(n)$,
- $X \sim \chi^2(n) \Rightarrow K_\alpha(X) = \chi^2_\alpha(n)$,
- $X \sim F(n_1, n_2) \Rightarrow K_\alpha(X) = F_\alpha(n_1, n_2)$.

Převodní vztahy:

- $u_\alpha = -u_{1-\alpha}$,
- $t_\alpha(n) = -t_{1-\alpha}(n)$,
- $F_\alpha(n_1, n_2) = \frac{1}{F_{1-\alpha}(n_2, n_1)}$.

6.1.2 Charakteristika variability

Interkvartilové rozpětí $IQR = K_{0,75}(X) - K_{0,25}(X)$.

Příklad 6.1. Řešený příklad
Pomocí softwaru R zjistěte hodnotu kvantilů (a) $u_{0,85}$, $u_{0,60}$, $u_{0,13}$; (b) $t_{0,99}(15)$, $t_{0,28}(143)$, $t_{0,75}(44)$; (c) $\chi^2_{0,25}(80)$, $\chi^2_{0,64}(37)$, $\chi^2_{0,31}(2)$; (d) $F_{0,76}(9;12)$, $F_{0,11}(15;40)$, $F_{0,05}(100;87)$. Všechny vypočítané hodnoty řádně interpretujte.

Řešení příkadu 6.1
Hodnoty α-kvantilů standardizovaného normálního rozdělení, tj. u_α, vypočítáme příkazem `qnorm()`. Vstupním argumentem příkazu bude pouze hodnota α. Vstupní argumenty odpovídající hodnotám parametrů $\mu = 0$ (argument `mean = 0`) a $\sigma = 1$ (argument `sd = 1`) v příkazu specifikovat nemusíme, protože jde o výchozí hodnoty, které jsou ve funkci `qnorm()` automaticky přednastavené.

```
qnorm(0.85) # 1,036433
qnorm(0.60) # 0,2533471
qnorm(0.13) # -1,126391
```

Za předpokladu, že náhodná veličina X pochází ze standardizovaného normálního rozdělení, je 85 % hodnot menších nebo rovných $1,0364$, 60 % hodnot je menších nebo rovných $0,2533$ a 13 % hodnot je menších nebo rovných $-1,1264$.

Hodnoty α-kvantilů Studentova rozdělení o n stupních volnosti, tj. $t_\alpha(n)$, vypočítáme příkazem `qt()`. Vstupními argumenty příkazu budou hodnota α a počet stupňů volnosti (argument `df`).

```
qt(0.99, df = 15) # 2,60248
qt(0.28, df = 143) # -0,5842093
qt(0.75, df = 44) # 0,6801065
```

Za předpokladu, že náhodná veličina X pochází ze Studentova rozdělení o 15 stupních volnosti, je 99 % hodnot menších nebo rovných 2, 6025. Za předpokladu, že náhodná veličina X pochází ze Studentova rozdělení o 143 stupních volnosti, je 28 % hodnot menších nebo rovných $-0, 5842$. Za předpokladu, že náhodná veličina X pochází ze Studentova rozdělení o 44 stupních volnosti, je 75 % hodnot menších nebo rovných 0, 6801.

Hodnoty α-kvantilů χ^2 rozdělení o n stupních volnosti, tj. $\chi^2_\alpha(n)$, vypočítáme příkazem qchisq(). Vstupními argumenty příkazu budou hodnota α a počet stupňů volnosti (argument df).

```
qchisq(0.25, df = 80) # 71,14451
qchisq(0.64, df = 37) # 39,47272
qchisq(0.31, df = 2) # 0,7421274
```

Za předpokladu, že náhodná veličina X pochází z χ^2 rozdělení o 80 stupních volnosti, je 25 % hodnot menších nebo rovných 71, 1445. Za předpokladu, že náhodná veličina X pochází z χ^2 rozdělení o 37 stupních volnosti, je 64 % hodnot menších nebo rovných 39, 4727. Za předpokladu, že náhodná veličina X pochází z χ^2 rozdělení o 2 stupních volnosti, je 31 % hodnot menších nebo rovných 0, 7421.

Hodnoty α-kvantilů Fisherova-Snedecorova rozdělení o n_1 a n_2 stupních volnosti, tj. $F_\alpha(n_1, n_2)$, vypočítáme příkazem qf(). Vstupními argumenty příkazu budou hodnota α, počet stupňů volnosti n_1 (argument df1) a počet stupňů volnosti n_2 (argument df2).

```
qf(0.76, df1 = 9, df2 = 12) # 1,535992
qf(0.11, df1 = 15, df2 = 40) # 0,5563822
qf(0.05, df1 = 100, df2 = 87) # 0,7114649
```

Za předpokladu, že náhodná veličina X pochází z Fisherova-Snedecorova rozdělení o 9 a 12 stupních volnosti, je 76 % hodnot menších nebo rovných 1, 5360. Za předpokladu, že náhodná veličina X pochází z Fisherova-Snedecorova rozdělení o 15 a 40 stupních volnosti, je 11 % hodnot menších nebo rovných 0, 5564. Za předpokladu, že náhodná veličina X pochází z Fisherova-Snedecorova rozdělení o 100 a 87 stupních volnosti, je 5 % hodnot menších nebo rovných 0, 7115. ★

Příklad 6.2. Řešený příklad
Pomocí softwaru R (a) zjistěte hodnoty kvantilů $u_{0,10}$, $u_{0,90}$ a ověřte, že platí vztah $u_{0,10} = -u_{0,90}$; (b) zjistěte hodnoty kvantilů $t_{0,65}(18)$, $t_{0,35}(18)$ a ověřte, že platí vztah $t_{0,65}(18) = -t_{0,35}(18)$; (c) zjistěte hodnoty kvantilů $F_{0,48}(13; 1)$, $F_{0,52}(1; 13)$ a ověřte, že platí vztah $F_{0,48}(13; 1) = \frac{1}{F_{0,52}(1;13)}$.

Řešení příkadu 6.2
Hodnoty α-kvantilů standardizovaného normálního rozdělení, tj. u_α, vypočítáme příkazem qnorm().

```
qnorm(0.10) # -1,281552
qnorm(0.90) # 1,281552
```

Kvantil $u_{0,10} = -1, 2816$, kvantil $u_{0,90} = 1, 2816$. Vidíme tedy, že rovnost $u_{0,10} = -u_{0,90}$ platí.

Hodnoty α-kvantilů Studentova rozdělení o n stupních volnosti, tj. $t_\alpha(n)$, vypočítáme příkazem qt().

```
qt(0.65, df = 18) # 0,3915326
qt(0.35, df = 18) # -0,3915326
```

Kvantil $t_{0,65}(18) = 0, 3915$, kvantil $t_{0,35}(18) = -0, 3915$. Vidíme tedy, že rovnost $t_{0,65}(18) = -t_{0,35}(18)$ platí.

Hodnoty α-kvantilů Fisherova-Snedecorova rozdělení o n_1 a n_2 stupních volnosti, tj. $F_\alpha(n_1, n_2)$, vypočítáme příkazem qf().

```
qf(0.48, df1 = 13, df2 = 1) # 1,891045
qf(0.52, df1 = 1, df2 = 13) # 0,5288082
1 / qf(0.52, df1 = 1, df2 = 13) # 1,891045
```

Kvantil $F_{0,48}(13; 1) = 1, 8910$, kvantil $F_{0,52}(1; 13) = 0, 5288$, hodnota $\frac{1}{F_{0,52}(1;13)} = \frac{1}{0,5288} = 1, 8910$. Vidíme tedy, že rovnost $F_{0,48}(13; 1) = \frac{1}{F_{0,52}(1;13)}$ platí. ★

Příklad 6.3. Neřešený příklad
Pomocí softwaru R zjistěte hodnotu kvantilů (a) $u_{0,21}$, $u_{0,92}$, $u_{0,50}$; (b) $t_{0,14}(136)$, $t_{0,47}(9)$, $t_{0,26}(62)$; (c) $\chi^2_{0,38}(12)$, $\chi^2_{0,07}(70)$, $\chi^2_{0,66}(425)$; (d) $F_{0,83}(83;83)$, $F_{0,59}(10;7)$, $F_{0,14}(140;79)$. Všechny vypočítané hodnoty řádně interpretujte.
Výsledky: (a) $u_{0,21} = -0,8064$, $u_{0,92} = 1,4051$, $u_{0,50} = 0,0000$; (b) $t_{0,14}(136) = -1,0846$, $t_{0,47}(9) = -0,0774$, $t_{0,26}(62) = -0,6470$; (c) $\chi^2_{0,38}(12) = 9,9540$, $\chi^2_{0,07}(70) = 53,3893$, $\chi^2_{0,66}(425) = 436,4614$; (d) $F_{0,83}(83;83) = 1,2340$, $F_{0,59}(10;7) = 1,2147$, $F_{0,14}(140;79) = 0,8106$. ★

Příklad 6.4. Neřešený příklad
Pomocí softwaru R (a) zjistěte hodnoty kvantilů $u_{0,76}$, $u_{0,24}$ a ověřte, že platí vztah $u_{0,76} = -u_{0,24}$; (b) zjistěte hodnoty kvantilů $t_{0,04}(315)$, $t_{0,96}(315)$ a ověřte, že platí vztah $t_{0,04}(315) = -t_{0,96}(315)$; (c) zjistěte hodnoty kvantilů $F_{0,31}(180;248)$, $F_{0,69}(248;180)$ a ověřte, že platí vztah $F_{0,31}(180;248) = \frac{1}{F_{0,69}(248;180)}$.
Výsledky: (a) $u_{0,76} = 0,7063$, $u_{0,24} = -0,7063$, rovnost platí; (b) $t_{0,04}(315) = -1,7564$, $t_{0,96}(315) = 1,7564$, rovnost platí; (c) $F_{0,31}(180;248) = 0,9325$, $F_{0,69}(248;180) = 1,0724$, $\frac{1}{F_{0,69}(248;180)} = 0,9325$, rovnost platí. ★

6.2 Číselné charakteristiky náhodných veličin intervalového a poměrového typu

6.2.1 Charakteristika polohy

Předpokládejme, že $p(x)$ je pravděpodobnostní funkce náhodné veličiny X (v případě, že X je diskrétní náhodná veličina) a $f(x)$ je hustota náhodné veličiny X (v případě, že X je spojitá náhodná veličina).

Střední hodnota $\mathrm{E}(X) = \begin{cases} \sum_{x=-\infty}^{\infty} xp(x), \\ \int_{-\infty}^{\infty} xf(x)\mathrm{d}x, \end{cases}$

pokud je suma, resp. integrál na pravé straně rovnosti konečný nebo absolutně konverguje. Jinak střední hodnota neexistuje.

6.2.2 Charakteristika variability

Rozptyl $\mathrm{D}(X) = \mathrm{E}\left([X - \mathrm{E}(X)]^2\right) = \mathrm{E}\left(X^2\right) - [\mathrm{E}(X)]^2 = \begin{cases} \sum_{x=-\infty}^{\infty} x^2 p(x) - \left[\sum_{x=-\infty}^{\infty} xp(x)\right]^2, \\ \int_{-\infty}^{\infty} x^2 f(x)\mathrm{d}x - \left[\int_{-\infty}^{\infty} xf(x)\mathrm{d}x\right]^2, \end{cases}$

pokud střední hodnoty $\mathrm{E}\left(X^2\right)$ a $\mathrm{E}\left(X\right)$ existují.

Směrodatná odchylka: $\sqrt{\mathrm{D}(X)}$. Centrovaná náhodná veličina: $Z = X - \mathrm{E}(X)$. Standardizovaná náhodná veličina: $U = \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{D}(X)}}$. Pro centrovanou a standardizovanou náhodnou veličinu platí: $\mathrm{E}(Z) = 0$, $\mathrm{D}(Z) = \mathrm{D}(X)$, $\mathrm{E}(U) = 0$, $\mathrm{D}(U) = 1$. Pro konstantu k platí: $\mathrm{E}(k) = k$, $\mathrm{D}(k) = 0$.

Střední hodnoty a rozptyly vybraných diskrétních a spojitých rozdělení

- $X \sim \mathrm{A}(\vartheta) \Rightarrow \mathrm{E}(X) = \vartheta$, $\mathrm{D}(X) = \vartheta(1-\vartheta)$,
- $X \sim \mathrm{Bi}(n, \vartheta) \Rightarrow \mathrm{E}(X) = n\vartheta$, $\mathrm{D}(X) = n\vartheta(1-\vartheta)$,
- $X \sim \mathrm{Hg}(N, M, k) \Rightarrow \mathrm{E}(X) = \frac{Mk}{N}$, $\mathrm{D}(X) = \frac{Mk}{N}\left(1 - \frac{M}{N}\right)\frac{N-k}{N-1}$,
- $X \sim \mathrm{Po}(\lambda) \Rightarrow \mathrm{E}(X) = \lambda$, $\mathrm{D}(X) = \lambda$,
- $X \sim N(\mu, \sigma^2) \Rightarrow \mathrm{E}(X) = \mu$, $\mathrm{D}(X) = \sigma^2$.

6.2.3 Charakteristika společné variability dvou náhodných veličin

Předpokládejme, že $p(x, y)$ je simultánní pravděpodobnostní funkce náhodného vektoru $(X, Y)^T$, $p_x(x)$ je pravděpodobnostní funkce náhodné veličiny X a $p_y(y)$ je pravděpodobnostní funkce náhodné veličiny Y (v případě, že X a Y jsou diskrétní náhodné veličiny). Dále předpokládejme, že $f(x, y)$ je simultánní hustota náhodného vektoru $(X, Y)^T$, $f_x(x)$ je hustota náhodné veličiny X a $f_y(y)$ je hustota náhodné veličiny Y (v případě, že X a Y jsou spojité náhodné veličiny). Kovariance

$$\begin{aligned}\mathrm{C}(X,Y) &= \mathrm{E}\left([X-\mathrm{E}(X)][Y-\mathrm{E}(Y)]\right)\\ &= \mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)\\ &= \begin{cases}\sum_{x=-\infty}^{\infty}\sum_{y=-\infty}^{\infty} xy\,p(x,y)-\sum_{x=-\infty}^{\infty} xp_x(x)\sum_{y=-\infty}^{\infty} yp_y(y),\\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)\mathrm{d}x\mathrm{d}y-\int_{-\infty}^{\infty} xf_x(x)\mathrm{d}x\int_{-\infty}^{\infty} yf_y(y)\mathrm{d}y,\end{cases}\end{aligned}$$

pokud střední hodnoty $E(XY)$, $E(X)$ a $E(Y)$ existují. Je-li $\mathrm{C}(X, Y) > 0$, resp. < 0, znamená to, že mezi náhodnými veličinami X a Y existuje určitý stupeň přímé, resp. nepřímé lineární závislosti. Je-li $\mathrm{C}(X, Y) = 0$, pak řekneme, že náhodné veličiny X a Y jsou nekorelované, a znamená to, že mezi nimi není žádný lineární vztah.

Upozornění: Z nekorelovanosti obecně nevyplývá stochastická nezávislost, avšak ze stochastické nezávislosti plyne nekorelovanost.

6.2.4 Charakteristika těsnosti lineárního vztahu dvou náhodných veličin

Koeficient korelace

$$\begin{aligned}\mathrm{R}(X,Y) &= \begin{cases}\mathrm{E}\left(\frac{X-\mathrm{E}(X)}{\sqrt{\mathrm{D}(X)}}\frac{Y-\mathrm{E}(Y)}{\sqrt{\mathrm{D}(Y)}}\right) & \text{pro } \sqrt{\mathrm{D}(X)}\sqrt{\mathrm{D}(Y)}>0,\\ 0 \text{ jinak,}\end{cases}\\ &= \begin{cases}\frac{\mathrm{C}(X,Y)}{\sqrt{\mathrm{D}(X)}\sqrt{\mathrm{D}(Y)}} & \text{pro } \sqrt{\mathrm{D}(X)}\sqrt{\mathrm{D}(Y)}>0,\\ 0 \text{ jinak.}\end{cases}\end{aligned}$$

Cauchyho-Schwarzova-Bunjakovského nerovnost: $|\mathrm{R}(X, Y)| \leq 1$, přičemž rovnost nastane tehdy a jen tehdy, když mezi veličinami X a Y existuje s pravděpodobností 1 úplná lineární závislost, tj. existují konstanty a, b, pro které $\Pr(Y = a + bX) = 1$.

Příklad 6.5. Řešený příklad
Načtěte datový soubor 30-fiances-single-marital-status.csv obsahující hodnoty simultánní pravděpodobnostní funkce věku v letech u svobodných snoubenců (věk v letech u ženicha, věk v letech u nevěsty) vstupujících do manželství v roce 2019 na území České republiky. Za předpokladu, že náhodná veličina X popisuje věk v letech u svobodného ženicha a náhodná veličina Y popisuje věk v letech u svobodné nevěsty, (a) vypočítejte střední hodnotu a směrodatnou odchylku náhodné veličiny X, resp. Y; (b) vypočítejte kovarianci a koeficient korelace. Všechny vypočítané hodnoty řádně interpretujte.

Poznámka: Datový soubor 30-fiances-single-marital-status.csv obsahuje kompletní údaje o všech sňatcích uskutečněných na území České republiky v roce 2019 a o věku v letech u každého ženicha i u každé nevěsty, kteří byli v roce 2019 sezdáni. V tomto případě tedy nepočítáme odhady, ale přímo hodnoty $\mathrm{E}(X)$, $\mathrm{E}(Y)$, $\sqrt{\mathrm{D}(X)}$, $\sqrt{\mathrm{D}(Y)}$, $\mathrm{C}(X, Y)$ a $\mathrm{R}(X, Y)$. V praxi však většinou máme pouze reprezentativní vzorek populace, na jehož základě střední hodnoty, rozptyly, kovarianci a koeficient korelace pouze odhadujeme. V takovém případě bychom počítali odhady $\widehat{\mathrm{E}(X)}$, $\widehat{\mathrm{E}(Y)}$, $\sqrt{\widehat{\mathrm{D}(X)}}$, $\sqrt{\widehat{\mathrm{D}(Y)}}$, $\widehat{\mathrm{C}(X,Y)}$ a $\widehat{\mathrm{R}(X,Y)}$ náhodných veličin X a Y.

Řešení příkladu 6.5
Datový soubor načteme příkazem read.delim() s argumentem row.names = 1, specifikujícím, že hodnoty vložené v prvním sloupci mají být vnímány jako názvy řádků načtené tabulky. Sloupce tabulky pojmenujeme stejně jako

řádky, protože věk u ženicha i u nevěsty je rozdělen do stejných věkových kategorií. Datovou tabulku si následně zobrazíme. Příkazem format() s argumentem scientific = F nastavíme, aby se čísla v tabulce zobrazila v klasickém desetinném zápisu (formát 0.0003) namísto v přednastaveném vědeckém zápisu čísel (formát 3×10^{-4}). Převod na desetinný zápis čísel provádíme pouze pro lepší přehlednost vypsané tabulky. Poznamenejme, že v řádcích tabulky je uveden věk u ženicha, ve sloupcích věk u nevěsty (viz kapitola 13, sekce 13.20).

```
data <- read.delim("30-fiances-single-marital-status.csv", sep = ",", dec = ".",
    row.names = 1)
names(data) <- row.names(data)
format(data, scientific = F)
```

```
       17-19  20-24  25-29  30-34  35-39  40-44  45-49  50-54  55-59  60-64
17-19 0.0010 0.0007 0.0001 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
20-24 0.0051 0.0435 0.0182 0.0033 0.0005 0.0002 0.0000 0.0000 0.0000 0.0000
25-29 0.0024 0.0777 0.2238 0.0410 0.0050 0.0007 0.0001 0.0000 0.0000 0.0000
30-34 0.0008 0.0248 0.1553 0.1283 0.0187 0.0027 0.0003 0.0000 0.0000 0.0000
35-39 0.0002 0.0061 0.0411 0.0677 0.0364 0.0071 0.0007 0.0000 0.0000 0.0000
40-44 0.0000 0.0017 0.0089 0.0197 0.0209 0.0125 0.0013 0.0002 0.0001 0.0000
45-49 0.0000 0.0003 0.0015 0.0034 0.0039 0.0054 0.0014 0.0001 0.0001 0.0000
50-54 0.0000 0.0001 0.0003 0.0005 0.0007 0.0009 0.0006 0.0003 0.0001 0.0000
55-59 0.0000 0.0000 0.0001 0.0001 0.0002 0.0003 0.0001 0.0002 0.0001 0.0000
60-64 0.0000 0.0000 0.0000 0.0000 0.0001 0.0001 0.0000 0.0001 0.0001 0.0001
```

Střední hodnotu náhodné veličiny X vypočítáme pomocí vzorce $\mathrm{E}(X) = \sum_{x=-\infty}^{\infty} xp_x(x)$, kde x je vektor středů třídicích intervalů věkových kategorií u ženicha, tj. posloupnost hodnot 18, 22, 27, 32, 37, 42, 47, 52, 57, 62, a $p_x(x)$ je pravděpodobnostní funkce náhodné veličiny X. Hodnoty funkce $p_x(x)$ získáme jako řádkové součty v tabulce data příkazem apply(). Rozptyl náhodné veličiny X vypočítáme přepisem vzorce $\mathrm{D}(X) = \sum_{x=-\infty}^{\infty} x^2 p_x(x) - \left[\sum_{x=-\infty}^{\infty} xp_x(x)\right]^2$. Směrodatnou odchylku získáme odmocněním rozptylu $D(X)$.

```
x <- c(18, 22, 27, 32, 37, 42, 47, 52, 57, 62)
px <- apply(data, MARGIN = 1, FUN = sum)
EX <- sum(x * px) # 31,3168
DX <- sum(x ^ 2 * px) - (sum(x * px)) ^ 2 # 32,40604
sqrt(DX) # 5,69263
```

Střední hodnota věku u svobodného ženicha je $31,32$ let se směrodatnou odchylkou $5,69$ let.

Střední hodnotu náhodné veličiny Y vypočítáme pomocí vzorce $\mathrm{E}(Y) = \sum_{y=-\infty}^{\infty} yp_y(y)$, kde y je vektor středů třídicích intervalů věkových kategorií u nevěsty, shodou okolností opět posloupnost hodnot 18, 22, 27, 32, 37, 42, 47, 52, 57, 62, a $p_y(y)$ je pravděpodobnostní funkce náhodné veličiny Y. Hodnoty funkce $p_y(y)$ získáme jako sloupcové součty v tabulce data. Rozptyl náhodné veličiny Y vypočítáme přepisem vzorce $\mathrm{D}(Y) = \sum_{y=-\infty}^{\infty} y^2 p_y(y) - \left[\sum_{y=-\infty}^{\infty} yp_y(y)\right]^2$. Odmocněním rozptylu $\mathrm{D}(Y)$ získáme směrodatnou odchylku.

```
y <- c(18, 22, 27, 32, 37, 42, 47, 52, 57, 62)
py <- apply(data, MARGIN = 2, FUN = sum)
EY <- sum(y * py) # 28,9035
DY <- sum(y ^ 2 * py) - (sum(y * py)) ^ 2 # 25,92119
sqrt(DY) # 5,091285
```

Střední hodnota věku u svobodné nevěsty je $28,90$ let se směrodatnou odchylkou $5,09$ let.

Kovarianci mezi náhodnými veličinami X a Y vypočítáme pomocí vzorce $\mathrm{C}(X,Y) = \sum_{x=-\infty}^{\infty} \sum_{y=-\infty}^{\infty} xyp(x,y) - \sum_{x=-\infty}^{\infty} xp_x(x) \sum_{y=-\infty}^{\infty} yp_y(y)$, kde x, resp. y je vektor středů třídicích intervalů věkových kategorií u ženicha, resp. u nevěsty, $p_x(x)$, resp. $p_y(y)$ je pravděpodobnostní funkce náhodné veličiny X, resp. Y a $p(x,y)$ je simultánní pravděpodobnostní funkce náhodného vektoru $(X,Y)^T$. Hodnoty funkce $p(x,y)$ máme vloženy v proměnné data. Matici součinů xy všech kombinací středů třídicích intervalů náhodných veličin X a Y vypočítáme pomocí operátoru maticového násobení %*% (viz kapitola 3). Koeficient korelace dopočítáme přepisem vzorce $\mathrm{R}(X,Y) = \frac{\mathrm{C}(X,Y)}{\sqrt{\mathrm{D}(X)}\sqrt{\mathrm{D}(Y)}}$.

```
pxy <- data
xy <- x %*% t(y)
CXY <- sum(xy * pxy) - sum(x * px) * sum(y * py) # 17,94097
RXY <- CXY / (sqrt(DX) * sqrt(DY)) # 0,6190212
```

Kovariance mezi náhodnými veličinami X a Y nabývá hodnoty $17,94\,\text{let}^2$. Mezi věkem u svobodného ženicha a věkem u svobodné nevěsty existuje význačný stupeň přímé lineární závislosti ($\mathrm{R}(X,Y) = 0,62$; stupeň míry závislosti viz kapitola 3, tabulka 3.2). ★

Příklad 6.6. Neřešený příklad
Načtěte datový soubor 29-live-births.csv obsahující hodnoty simultánní pravděpodobnostní funkce věku v letech u matky v roce 2019 a počtu jejích živě narozených potomků do téhož roku na území České republiky. Za předpokladu, že náhodná veličina X popisuje věk v letech u matky a náhodná veličina Y popisuje počet živě narozených potomků, (a) vypočítejte střední hodnotu a směrodatnou odchylku náhodné veličiny X, resp. Y; (b) vypočítejte kovarianci a koeficient korelace. Všechny vypočítané hodnoty řádně interpretujte.

Poznámka: Datový soubor 29-live-births.csv obsahuje kompletní údaje o věku v letech a o počtu živě narozených potomků u každé matky žijící na území České republiky v roce 2019. V tomto případě tedy znovu nepočítáme odhady středních hodnot, směrodatných odchylek, kovariance a koeficientu korelace náhodných veličin X a Y, ale přímo hodnoty $\mathrm{E}(X)$, $\mathrm{E}(Y)$, $\sqrt{\mathrm{D}(X)}$, $\sqrt{\mathrm{D}(Y)}$, $\mathrm{C}(X,Y)$ a $\mathrm{R}(X,Y)$.

Výsledky: (a) $\mathrm{E}(X) = 30,41$, $\sqrt{\mathrm{D}(X)} = 5,41$, $\mathrm{E}(Y) = 1,73$, $\sqrt{\mathrm{D}(Y)} = 0,90$; (b) $\mathrm{C}(X,Y) = 1,49$, $\mathrm{R}(X,Y) = 0,31$, mírný stupeň přímé lineární závislosti. ★

6.3 Centrální limitní věta

$X_1, \ldots, X_n$ jsou stochasticky nezávislé náhodné veličiny se stejným rozdělením se střední hodnotou μ a rozptylem σ^2. Pak pro velká n ($n \geq 30$) lze rozdělení součtu $\sum_{i=1}^{n} X_i$ aproximovat normálním rozdělením $N(n\mu, n\sigma^2)$. Zkráceně píšeme $\sum_{i=1}^{n} X_i \approx N(n\mu, n\sigma^2)$. Standardizací tohoto součtu vytvoříme náhodnou veličinu $U_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \approx N(0,1)$.

Důsledek: Moivreova-Laplaceova věta
$X_1, \ldots, X_n$ jsou stochasticky nezávislé náhodné veličiny, $X_i \sim \mathrm{A}(\vartheta)$, $i = 1, 2, \ldots, n$. Pak $Z_n = \sum_{i=1}^{n} X_i \sim \mathrm{Bi}(n, \vartheta)$ a za splnění podmínek dobré aproximace $\frac{1}{n+1} < \vartheta < \frac{n}{n+1}$ a $n\vartheta(1-\vartheta) > 9$ platí, že $U = \frac{Z_n - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}} \approx N(0,1)$.
Vzorec aproximace: $\Pr(Z_n \leq z) \approx \Phi\left(\frac{z-n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}}\right) = \Phi(u)$, kde $\Phi(u)$ je distribuční funkce rozdělení $N(0,1)$ v bodě $u = \frac{z-n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}}$.

Příklad 6.7. Řešený příklad
Pravděpodobnost výskytu dermatoglyfického vzoru *vír* u mužů z populace Valmikis je $p_m = 0,4780$ (Rajendra Rao, 1972), pravděpodobnost výskytu dermatoglyfického vzoru *vír* u žen z populace Valmikis je $p_f = 0,3500$ (Mrunalini Devi, 1972). Za předpokladu, že náhodná veličina X popisuje výskyt dermatoglyfického vzoru *vír* u jednoho muže z populace Valmikis, vypočítejte pravděpodobnost, že v náhodném výběru 1200 mužů z populace Valmikis se vzor *vír* vyskytne (a) vícekrát než kterýkoli jiný vzor; (b) nejvýše u 575 mužů; (c) alespoň u 360 a nejvýše u 560 mužů. Zadané pravděpodobnosti vypočítejte (i) pomocí vhodného rozdělení; (ii) přibližně na základě aproximace tohoto rozdělení pomocí Moivreovy-Laplaceovy věty. Výsledky vzájemně porovnejte.

Řešení příkladu 6.7
Počet mužů s dermatoglyfickým vzorem *vír* v náhodném výběru 1200 mužů z populace Valmikis je diskrétní znak, k jeho popisu tedy použijeme diskrétní náhodnou veličinu. Výskyt vzoru *vír* byl zkoumán v 1200 Bernoulliho pokusech $X_1, \ldots, X_{1200}$, přičemž v každém pokusu mohlo dojít k nastání sledované události ($X_i = 1$; výskyt vzoru *vír*), nebo k nenastání sledované události ($X_i = 0$; výskyt libovolného jiného vzoru). Počet mužů se vzorem *vír* v náhodném

výběru 1200 mužů potom popisuje náhodná veličina $Z_{1200} = \sum_{i=1}^{1200} X_i$, která asymptoticky pochází z binomického rozdělení s parametry $n = 1200$ a $\vartheta = 0,4780$. Zadané pravděpodobnosti (a), (b) a (c) tedy vypočítáme pomocí binomického rozdělení.

Pravděpodobnost, že v náhodném výběru 1200 mužů z populace Valmikis se vzor *vír* vyskytne vícekrát než kterýkoli jiný vzor, odpovídá pravděpodobnosti, že se vzor *vír* vyskytne alespoň u 601 mužů, tj. $\Pr(Z_{1200} \geq 601)$. Tuto pravděpodobnost vypočítáme tak, že od jedné odečteme pravděpodobnost, že se vzor *vír* vyskytne nejvýše u 600 mužů, což odpovídá hodnotě distribuční funkce $F(z)$ rozdělení $\mathrm{Bi}(1200; 0,4780)$ v bodě $z = 600$. Výpočet provedeme pomocí funkce pbinom().

```
theta <- 0.4780; n <- 1200
1 - pbinom(q = 600, size = n, prob = theta) # 0,06007658
```

Pravděpodobnost, že v náhodném výběru 1200 mužů z populace Valmikis se vzor *vír* vyskytne nejvýše u 575 mužů, tj. $\Pr(Z_{1200} \leq 575)$, odpovídá hodnotě distribuční funkce $F(z)$ v bodě $z = 575$. Tuto hodnotu vypočítáme příkazem pbinom().

```
pbinom(q = 575, size = n, prob = theta) # 0,5438802
```

Pravděpodobnost, že v náhodném výběru 1200 mužů z populace Valmikis se vzor *vír* vyskytne alespoň u 360 a nejvýše u 560 mužů, tj. $\Pr(360 \leq Z_{1200} \leq 560)$, vypočítáme tak, že od pravděpodobnosti, že se vzor *vír* vyskytne nejvýše u 560 mužů, odečteme pravděpodobnost, že se vzor *vír* vyskytne nejvýše u 359 mužů. Obě pravděpodobnosti jsou hodnotami distribuční funkce $F(z)$ v bodě $z = 560$, resp. v bodě $z = 359$ a vypočítáme je příkazem pbinom().

```
pbinom(q = 560, size = n, prob = theta) - pbinom(q = 359, size = n, prob = theta) #
    0,2245673
```

Pravděpodobnost, že v náhodném výběru 1200 mužů z populace Valmikis se vzor *vír* vyskytne vícekrát než kterýkoli jiný vzor, je $0,0601$ $(6,01\,\%)$. Pravděpodobnost, že v náhodném výběru se vzor *vír* vyskytne nejvýše u 575 mužů, je $0,5439$ $(54,39\,\%)$. Pravděpodobnost, že v náhodném výběru se vzor *vír* vyskytne alespoň u 360 a nejvýše u 560 mužů, je $0,2246$ $(22,46\,\%)$.

Nyní si zadané pravděpodobnosti (a), (b) a (c) vypočítáme přibližně s využitím Moivreovy-Laplaceovy věty pomocí standardizovaného normálního rozdělení. Nejprve je třeba ověřit splnění podmínek dobré aproximace. První podmínka dobré aproximace je splněna, neboť

$$\begin{aligned} \frac{1}{n+1} < \vartheta < \frac{n}{n+1} \\ \frac{1}{1200+1} < 0,4780 < \frac{1200}{1200+1} \\ 0,0008326 < 0,4780 < 0,9992. \end{aligned}$$

```
1 / (n + 1) # 0,0008326395
n / (n + 1) # 0,9991674
```

Rovněž druhá podmínka dobré aproximace je splněna, neboť $n\vartheta(1-\vartheta) = 1200 \times 0,4780 \times (1-0,4780) = 299,4192 > 9$.

```
n * theta * (1 - theta) # 299,4192
```

Podle Moivreovy-Laplaceovy věty, pokud náhodná veličina $Z_n = \sum_{i=1}^{n} X_i \sim \mathrm{Bi}(n, \vartheta)$, potom náhodná veličina $U = \frac{Z_n - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}} \approx N(0,1)$. V našem případě náhodná veličina $Z_{1200} = \sum_{i=1}^{1200} X_i \sim \mathrm{Bi}(1200; 0,4780)$, a tedy náhodná veličina $U = \frac{Z_{1200} - 1200 \times 0,4780}{\sqrt{1200 \times 0,4780 \times (1-0,4780)}} = \frac{Z_{1200} - 573,6}{17,3037} \approx N(0,1)$. Tohoto poznatku využijeme při přibližném výpočtu zadaných pravděpodobností (a), (b) a (c).

Přibližná pravděpodobnost, že v náhodném výběru 1200 mužů z populace Valmikis se vzor *vír* vyskytne vícekrát než kterýkoli jiný vzor, tj. alespoň u 601 mužů, vypočítáme tak, že od jedné odečteme přibližnou pravděpodobnost, že se vzor *vír* vyskytne nejvýše u 601 mužů, která odpovídá hodnotě distribuční funkce $\Phi(u)$ rozdělení $N(0,1)$ v bodě $u = \frac{601 - 573,6}{17,3037} = 1,5835$. Výpočet provedeme pomocí funkce pnorm().

```
u <- (601 - n * theta) / sqrt(n * theta * (1 - theta)) # 1,583473
1 - pnorm(q = u, mean = 0, sd = 1) # 0,05665682
```

Přibližná pravděpodobnost, že v náhodném výběru 1200 mužů z populace Valmikis se vzor *vír* vyskytne nejvýše u 575 mužů, odpovídá hodnotě distribuční funkce $\Phi(u)$ rozdělení $N(0,1)$ v bodě $u = \frac{575-573,6}{17,3037} = 0,0809$. Tuto hodnotu vypočítáme příkazem pnorm().

```
u <- (575 - n * theta) / sqrt(n * theta * (1 - theta)) # 0,08090739
pnorm(q = u, mean = 0, sd = 1) # 0,5322422
```

Přibližná pravděpodobnost, že v náhodném výběru 1200 mužů z populace Valmikis se vzor *vír* vyskytne alespoň u 360 a nejvýše u 560 mužů, vypočítáme tak, že od přibližné pravděpodobnosti, že se vzor *vír* vyskytne nejvýše u 560 mužů, odečteme přibližnou pravděpodobnost, že se vzor *vír* vyskytne nejvýše u 360 mužů. Obě pravděpodobnosti jsou hodnotami distribuční funkce $\Phi(u)$ rozdělení $N(0,1)$ v bodě $u = \frac{560-573,6}{17,3037} = -0,7860$, resp. v bodě $u = \frac{360-573,6}{17,3037} = -12,3442$ a vypočítáme je příkazem pnorm().

```
u1 <- (560 - n * theta) / sqrt(n * theta * (1 - theta)) # -0,7859575
u2 <- (360 - n * theta) / sqrt(n * theta * (1 - theta)) # -12,34416
pnorm(q = u1, mean = 0, sd = 1) - pnorm(q = u2, mean = 0, sd = 1) # 0,2159462
```

Přibližná pravděpodobnost, že v náhodném výběru 1200 mužů z populace Valmikis se vzor *vír* vyskytne vícekrát než kterýkoli jiný vzor, je $0,0567$ ($5,67\,\%$). Přibližná pravděpodobnost, že v náhodném výběru se vzor *vír* vyskytne nejvýše u 575 mužů, je $0,5322$ ($53,22\,\%$). Přibližná pravděpodobnost, že v náhodném výběru se vzor *vír* vyskytne alespoň u 360 a nejvýše u 560 mužů, je $0,2159$ ($21,59\,\%$).

Přibližné výsledky si vzájemně porovnáme s výsledky vypočítanými pomocí asymptotického rozdělení (viz tabulka 6.1).

Tabulka 6.1: Výsledné pravděpodobnosti výskytu dermatoglyfického vzoru *vír* v náhodném výběru 1200 mužů z populace Valmikis vypočítané (i) pomocí asymptotického rozdělení; (ii) na základě aproximace tohoto rozdělení pomocí Moivreovy-Laplaceovy věty

	výpočet pomocí asymptotického rozdělení	přibližný výpočet
(a)	0.0601	0.0567
(b)	0.5439	0.5322
(c)	0.2246	0.2159

Z tabulky 6.1 vidíme, že pravděpodobnosti vypočítané pomocí asymptotického rozdělení se od přibližných pravděpodobností vypočítaných za předpokladu standardizovaného normálního rozdělení s využitím Moivreovy-Laplaceovy věty liší v hodnotách na druhém desetinném místě. Výsledky přibližného výpočtu jsou dostatečně blízké výsledkům získaným pomocí asymptotického rozdělení. ★

Příklad 6.8. Neřešený příklad
Pravděpodobnost výskytu epigenetického znaku *sutura metopica* v moderní japonské populaci je $p_j = 0,0910$ (Mouri, 1976). Za předpokladu, že náhodná veličina X popisuje výskyt epigenetického znaku *sutura metopica* u jednoho jedince z moderní japonské populace, vypočítejte pravděpodobnost, že v náhodném výběru 9000 jedinců z moderní japonské populace se epigenetický znak *sutura metopica* vyskytne (a) nejvýše u desetiny jedinců; (b) alespoň u 850 a nejvýše u 1012 jedinců; (c) alespoň u 825 jedinců. Zadané pravděpodobnosti vypočítejte (i) pomocí vhodného rozdělení; (ii) přibližně na základě aproximace tohoto rozdělení pomocí Moivreovy-Laplaceovy věty. Výsledky vzájemně porovnejte.
Výsledky: $Z_{9000} \sim \mathrm{Bi}(9000; 0,0910)$; (i-a) $\Pr(Z_{9000} \leq 900) = 0,9984$; (i-b) $\Pr(850 \leq Z_{9000} \leq 1012) = 0,1321$; (i-c) $\Pr(Z_{9000} \geq 825) = 0,4183$; první podmínka dobré aproximace je splněna $(0,0001111 < 0,0910 < 0,9999)$, druhá podmínka dobré aproximace je splněna $(744,471 > 9)$, $U = \frac{Z_{9000}-819}{27,285} \approx N(0,1)$; (ii-a) $\Pr(Z_{9000} \leq 900) \approx 0,9985$; (ii-b) $\Pr(850 \leq Z_{9000} \leq 1012) \approx 0,1279$; (ii-c) $\Pr(Z_{9000} \geq 825) \approx 0,4130$; výsledky přibližného výpočtu jsou dostatečně blízké výsledkům získaným pomocí asymptotického rozdělení (liší se v hodnotách na druhém až čtvrtém desetinném místě). ★