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. 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 rozložení

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

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 stanovte 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) $F_{0,76}(9;12)$, $F_{0,11}(15;40)$, $F_{0,03}(100;87)$. Všechny vypočítané hodnoty řádně interpretujte.

Řešení příkadu 6.1
Hodnoty α-kvantilů standardizovaného normálního rozložení, tj. u_α, vypočítáme pomocí příkazu qnorm(alpha).

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

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

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

```
4 qt(0.99, 15) # 2,60248
5 qt(0.28, 143) # -0,5842093
6 qt(0.75, 44) # 0,6801065
```

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

Hodnoty α-kvantilů χ^2 rozložení o n stupních volnosti, tj. $\chi^2_\alpha(n)$, vypočítáme pomocí příkazu qchisq(alpha, n).

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

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

Hodnoty α-kvantilů Fisherova-Snedecorova rozložení o n_1 a n_2 stupních volnosti, tj. $F_\alpha(n_1, n_2)$, vypočítáme pomocí příkazu qf(alpha, n1, n2).

```
qf(0.76, 9, 12) # 1,535992
qf(0.11, 15, 40) # 0,5563822
qf(0.03, 100, 87) # 0,6774145
```

Za předpokladu, že náhodná veličina X pochází z Fisherova-Snedecorova rozložení o 9 a 12 stupních volnosti, je 76 % hodnot menších nebo rovných 1, 5360. Za předpokladu, že náhodná veličina X pochází z Fisherova-Snedecorova rozložení o 15 a 40 stupních volnosti, je 11 % hodnot menších nebo rovných 0, 5564. Za předpokladu, že náhodná veličina X pochází z Fisherova-Snedecorova rozložení o 100 a 87 stupních volnosti, jsou 3 % hodnot menší nebo rovné 0, 6774. ★

Příklad 6.2. Řešený příklad
Pomocí softwaru R (a) vypočítejte $u_{0,10}$, $u_{0,90}$ a ověřte, že platí vztah $u_\alpha = -u_{1-\alpha}$; (b) vypočítejte $t_{0,65}(18)$, $t_{0,35}(18)$ a ověřte, že platí vztah $t_\alpha(n) = -t_{1-\alpha}(n)$; (c) vypočítejte $F_{0,48}(13; 1)$, $F_{0,52}(1; 13)$ a ověřte, že platí vztah $F_\alpha(n_1, n_2) = \frac{1}{F_{1-\alpha}(n_2,n_1)}$.

Řešení příkadu 6.2
Hodnoty α-kvantilů standardizovaného normálního rozložení, tj. u_α, vypočítáme pomocí příkazu qnorm(alpha).

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

Kvantil $u_{0,10} = -1, 2816$, kvantil $u_{0,90} = 1, 2816$. Z obou výsledků vidíme, že platí rovnost $u_{0,10} = -u_{0,90}$.

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

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

Kvantil $t_{0,65}(18) = 0, 3915$, kvantil $t_{0,35}(18) = -0, 3915$. Z obou výsledků vidíme, že platí rovnost $t_{0,65}(18) = -t_{0,35}(18)$.

Hodnoty α-kvantilů Fisherova-Snedecorova rozložení o n_1 a n_2 stupních volnosti, tj. $F_\alpha(n_1, n_2)$, vypočítáme pomocí příkazu qf(alpha, n1, n2).

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

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

★

Příklad 6.3. Neřešený příklad
Pomocí softwaru R stanovte 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) vypočítejte $u_{0,76}$, $u_{0,24}$ a ověřte, že platí vztah $u_\alpha = -u_{1-\alpha}$; (b) vypočítejte $t_{0,04}(315)$, $t_{0,96}(315)$ a ověřte, že platí vztah $t_\alpha(n) = -t_{1-\alpha}(n)$; (c) vypočítejte $F_{0,31}(180;248)$, $F_{0,69}(248;180)$ a ověřte, že platí vztah $F_\alpha(n_1,n_2) = \frac{1}{F_{1-\alpha}(n_2,n_1)}$.
Výsledky: (a) $u_{0,76} = 0,7063$, $u_{0,24} = -0,7063$; (b) $t_{0,04}(315) = -1,7564$, $t_{0,96}(315) = 1,7564$; (c) $F_{0,31}(180;248) = 0,9325$, $F_{0,69}(248;180) = 1,0724$, $\frac{1}{F_{0,69}(248;180)} = 0,9325$. ★

6.2 Číselné charakteristiky náhodných veličin intervalového a poměrového typu

6.2.1 Charakteristika polohy

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 či integrál vpravo konečný nebo absolutně konverguje. Jinak střední hodnota neexistuje.

6.2.2 Charakteristika variability

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

pokud střední hodnoty vpravo 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 rozložení

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

6.2.3 Charakteristika společné variability dvou náhodných veličin

Kovariance

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

pokud střední hodnoty vpravo existují. Je-li $\mathrm{C}(X,Y) > 0$, resp. < 0, znamená to, že mezi 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 nevyplývá stochastická nezávislost, avšak ze stochastické nezávislosti plyne nekorelovanost.

6.2.4 Charakteristika těsnosti lineárního vztahu dvou náhodných veličin

Koeficient korelace

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

Cauchyova-Schwarzova-Buňakovského nerovnost: $|R(X,Y)| \leq 1$, 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 svobodných snoubenců (věk ženicha; věk nevěsty: kategorizované proměnné: 17–19; 20–24; 25–29; 30–34; 35–39; 40–44; 45–49; 50–54; 55–59; 60–64) vstupujících do manželství v roce 2019 na území České republiky (zdroj dat: www.czso.cz, upravené). Za předpokladu, že náhodná veličina X popisuje věk ženicha a náhodná veličina Y popisuje věk nevěsty, (a) vypočítejte střední hodnotu a směrodatnou odchylku náhodné veličiny X, resp. Y; (b) vypočítejte kovarianci a korelační koeficient. Všechny vypočítané hodnoty řádně interpretujte.

Řešení příkladu 6.5
Datový soubor načteme příkazem read.delim() s argumentem row.names = 1, specifikujícím, že první sloupec souboru obsahuje názvy řádků datové tabulky. 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 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 pravděpodobnostní funkce $p_x(x)$ získáme jako řádkové součty v tabulce data, které vypočítáme příkazem apply() s argumenty MARGIN = 1 a FUN = sum (viz kapitola 3). Následně přepisem výše zmíněného vzorce stanovíme střední hodnotu $\mathrm{E}(X)$. Dále vypočítáme rozptyl náhodné veličiny X přepisem vzorce $\mathrm{D}(X) = \sum_{x=-\infty}^{\infty} x^2 p_x(x) - \left[\sum_{x=-\infty}^{\infty} xp_x(x)\right]^2$. Nakonec pomocí příkazu sqrt() získáme hodnotu směrodatné odchylky $\sqrt{\mathrm{D}(X)}$.

```
data <- read.delim('30-fiances-single-marital-status.csv', sep = ',', dec = '.',
                   row.names = 1)
x <- c(18, 22, 27, 32, 37, 42, 47, 52, 57, 62)
px <- apply(data, 1, 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 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 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 pravděpodobnostní funkce $p_y(y)$ získáme jako sloupcové součty v tabulce data prostřednictvím příkazu apply() s argumenty MARGIN = 2 a FUN = sum. Dále vypočítáme rozptyl náhodné veličiny Y přepisem vzorce $\mathrm{D}(Y) = \sum_{y=-\infty}^{\infty} y^2 p_y(y) - \left[\sum_{y=-\infty}^{\infty} yp_y(y)\right]^2$ a směrodatnou odchylku $\sqrt{\mathrm{D}(Y)}$.

```
y <- c(18, 22, 27, 32, 37, 42, 47, 52, 57, 62)
py <- apply(data, 2, 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 svobodné nevěsty je 28, 90 let se směrodatnou odchylkou 5, 09 let.

Hodnotu kovariance mezi znaky X a Y vypočítáme dosazením do 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 ženicha, resp. 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 $(X,Y)^T$. Matici součinů xy všech kombinací středů třídicích intervalů pro znaky X a Y vypočítáme pomocí *maticového násobení* (viz kapitola 3). Simultánní pravděpodobnostní funkci $p(x,y)$ máme vloženou v proměnné data. Kovarianci $\mathrm{C}(X,Y)$ můžeme nyní dopočítat přímým dosazením do výše uvedeného vzorce. Nakonec stanovíme hodnotu korelačního koeficientu přepisem vzorce $\mathrm{R}(X,Y) = \frac{\mathrm{C}(X,Y)}{\sqrt{\mathrm{D}(X)}\sqrt{\mathrm{D}(Y)}}$.

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

Kovariance mezi znaky X a Y nabývá hodnoty $17,94\,\mathrm{let}^2$. Mezi věkem svobodného ženicha a věkem svobodné nevěsty existuje význačný stupeň přímé lineární závislosti ($R(X,Y) = 0,62$; 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 matky v letech (kategorizovaná proměnná: 14 a méně; 15–19; 20–24; 25–29; 30–34; 35–39; 40–44; 45–49; 50 a více) a počtu živě narozených potomků (1, 2, 3, 4, 5, 6 a více) v roce 2019 v České republice (zdroj dat: www.czso.cz, upravené). Za předpokladu, že náhodná veličina X popisuje věk 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 korelační koeficient. Všechny vypočítané hodnoty řádně interpretujte.
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 rozložením se střední hodnotou μ a rozptylem σ^2. Pak pro velká n ($n \geq 30$) lze rozložení součtu $\sum_{i=1}^{n} X_i$ aproximovat normálním rozložením $N(n\mu, n\sigma^2)$. 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 $\frac{Z_n - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}} \approx N(0,1)$. Aproximativní vzorec: $\Pr(Z_n \leq z) \approx \Phi\left(\frac{z - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}}\right)$, kde Φ je distribuční funkce rozložení $N(0,1)$.

Příklad 6.7. Řešený příklad
Pravděpodobnost výskytu dermatoglyfického vzoru *vír* u mužů z populace Valmikis je $p_m = 0,4780$ (Rajendra, 1972), pravděpodobnost výskytu dermatoglyfického vzoru *vír* u žen z populace Valmikis je $p_f = 0,3500$ (Mrunalini, 1972). Za předpokladu, že náhodná veličina X popisuje výskyt dermatoglyfického vzoru *vír* u jednoho muže z populace Valmikis, vypočítejte pravděpodobnost, že v náhodném výběru 1 200 mužů z populace Valmikis se vzor *vír* vyskytne (a) vícekrát než kterýkoli jiný vzor; (b) nejvýše u 575 mužů; (c) nejméně u 360 a nejvýše u 560 mužů. Zadané pravděpodobnosti vypočítejte (i) přesně; (ii) aproximaticky pomocí Moivreovy-Laplaceovy věty. Výsledky

vzájemně porovnejte.

Řešení příkladu 6.7
Počet mužů se vzorem *vír* je diskrétní znak, k jeho přesnému 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). O náhodné veličině $Z_{1200} = \sum_{i=1}^{1200} X_i$ tedy předpokládáme, že pochází z binomického rozložení s parametry $n = 1200$ a $\vartheta = 0,4780$. Zadané pravděpodobnosti (a), (b) a (c) vypočítáme přesně pomocí binomického rozložení.

Pravděpodobnost, že v náhodném výběru 1200 mužů z populace Valmikis se vzor *vír* vyskytne vícekrát než kterýkoli jiný vzor, odpovídá pravděpodobnosti, že vzor *vír* se vyskytne alespoň u 601 mužů, tj. $\Pr(Z_{1200} \geq 601)$. Tuto pravděpodobnost vypočítáme tak, že od jedné odečteme pravděpodobnost, že vzor *vír* se 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 příkazu pbinom().

```
36 theta <- 0.4780
37 n <- 1200
38 1 - pbinom(600, n, 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().

```
39 pbinom(575, n, theta) # 0,5438802
```

Pravděpodobnost, že v náhodném výběru 1200 mužů z populace Valmikis se vzor *vír* vyskytne nejméně u 360 a nejvýše u 560 mužů, tj. $\Pr(360 \leq Z_{1200} \leq 560)$, vypočítáme tak, že od pravděpodobnosti, že vzor *vír* se vyskytne u nejvýše 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().

```
40 pbinom(560, n, theta) - pbinom(359, n, theta) # 0,2245673
```

Pravděpodobnost, že v náhodném výběru 1200 mužů z populace Valmikis se vzor *vír* vyskytne vícekrát než kterýkoli jiný vzor, je 6,01 %. Pravděpodobnost, že v náhodném výběru se vzor *vír* vyskytne nejvýše u 575 mužů, je 54,39 %. Pravděpodobnost, že v náhodném výběru se vzor *vír* vyskytne nejméně u 360 a nejvýše u 560 mužů, je 22,46 %.

Nyní si zadané pravděpodobnosti (a), (b) a (c) vypočítáme aproximaticky pomocí normálního rozložení. Nejprve je třeba ověřit splnění podmínek dobré aproximace. První podmínka dobré aproximace je splněna, neboť

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

```
41 1 / (n + 1) # 0,0008326395
42 n / (n + 1) # 0,9991674
```

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

```
43 n * p * (1 - p) # 299,4192
```

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

Aproximovanou pravděpodobnost, že v náhodném výběru 1200 mužů z populace Valmikis se vzor *vír* vyskytne vícekrát než kterýkoli jiný vzor, tj. $\Pr(Z_{1200} \geq 601)$, vypočítáme tak, že od jedné odečteme aproximovanou pravděpodobnost, že se vzor *vír* vyskytne nejvýše u 601 mužů, což tentokrát odpovídá hodnotě distribuční funkce $F(x)$ rozdělení $N(0,1)$ v bodě $z = \frac{601-573,6}{17,3037} = 1,5835$. Výpočet provedeme pomocí příkazu pnorm().

```
z <- (601 - n * theta) / sqrt(n * theta * (1 - theta)) # 1,583473
1 - pnorm(z) # 0,05665682
```

Aproximovaná 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 $F(z)$ rozdělení $N(0,1)$ v bodě $z = \frac{575-573,6}{17,3037} = 0,08091$. Tuto hodnotu vypočítáme příkazem pnorm().

```
z <- (575 - n * theta) / sqrt(n * theta * (1 - theta)) # 0,08090739
pnorm(z) # 0,5322422
```

Aproximovanou pravděpodobnost, že v náhodném výběru 1200 mužů z populace Valmikis se vzor *vír* vyskytne nejméně u 360 a nejvýše u 560 mužů, tj. $\Pr(360 \leq Z_{1200} \leq 560)$, vypočítáme tak, že od aproximované pravděpodobnosti, že se vzor *vír* vyskytne u nejvýše 560 mužů, odečteme aproximovanou pravděpodobnost, že se vzor *vír* vyskytne nejvýše u 360 mužů. Obě pravděpodobnosti jsou hodnotami distribuční funkce $F(z)$ rozložení $N(0,1)$ v bodě $z = \frac{560-573,6}{17,3037} = -0,7860$, resp. v bodě $z = \frac{360-573,6}{17,3037} = -12,3442$ a vypočítáme je pomocí příkazu pnorm().

```
z1 <- (560 - n * theta) / sqrt(n * theta * (1 - theta)) # -0,7859575
z2 <- (360 - n * theta) / sqrt(n * theta * (1 - theta)) # 0,2159462
pnorm(z1) - pnorm(z2) # 0,2159462
```

Aproximovaná pravděpodobnost, že v náhodném výběru 1200 mužů z populace Valmikis se vzor *vír* vyskytne vícekrát než kterýkoli jiný vzor, je 5,67 %. Aproximovaná pravděpodobnost, že v náhodném výběru se vzor *vír* vyskytne nejvýše u 575 mužů, je 53,22 %. Aproximovaná pravděpodobnost, že v náhodném výběru se vzor *vír* vyskytne nejméně u 360 a nejvýše u 560 mužů, je 21,59 %.

Přesné a aproximované výsledky si nyní vzájemně porovnáme (viz tabulka 6.1).

Tabulka 6.1: Porovnání přesných a aproximovaných výsledků

	přesné výsledky	aproximované výsledky
(a)	0,0601	0,0567
(b)	0,5439	0,5322
(c)	0,2246	0,2159

Z tabulky 6.1 vidíme, že aproximované pravděpodobnosti vypočítané za předpokladu normálního rozložení s využitím Moivreovy-Laplaceovy věty se liší od přesných pravděpodobností vypočítaných za předpokladu binomického rozložení na druhém desetinném místě.

★

Příklad 6.8. Neřešený příklad
Pravděpodobnost výskytu epigenetického znaku *sutura metopica* u mužů a žen z moderní japonské populace je $p_j = 0,0910$ (Mouri, 1976). Za předpokladu, že náhodná veličina X popisuje výskyt epigenetického znaku *sutura metopica* u jednoho jedince z japonské populace, vypočítejte pravděpodobnost, že v náhodném výběru 9 000 jedinců z japonské populace se epigenetický znak *sutura metopica* vyskytne (a) u méně než desetiny jedinců; (b) u 850 až 1000 jedinců; (c) alespoň u 825 jedinců. Zadané pravděpodobnosti vypočítejte (i) přesně; (ii) aproximaticky pomocí Moivreovy-Laplaceovy věty. Výsledky vzájemně porovnejte.
Výsledky: první podmínka dobré aproximace je splněna ($\frac{1}{n+1} < \vartheta < \frac{n}{n+1} = 0,0001111 < 0,0910 < 0,9999$); druhá podmínka dobré aproximace je splněna ($n\vartheta(1-\vartheta) = 744,471 > 9$); (i-a) $\Pr(Z_{9000} \leq 899) = 0,9982$; (i-b) $\Pr(850 \leq Z_{9000} \leq 1000) = 0,1321$; (i-c) $\Pr(Z_{9000} \geq 825) = 0,4183$; (ii-a) $\Pr(Z_{9000} \leq 899) = 0,9983$; (ii-b) $\Pr(850 \leq Z_{9000} \leq 1000) = 0,1279$; (ii-c) $\Pr(Z_{9000} \geq 825) = 0,4130$. ★