5 Bodové a intervalové odhady parametrů

5.1 Úvod do matematické statistiky

V rámci kapitol 2 a 3 jsme se seznámili se základními metodami popisné statistiky. Připomeňme si, že tyto metody slouží výhradně k seznámení se s datovým souborem, k pochopení podstaty předložených dat a zjištění jejich základních vlastností. Je důležité si uvědomit, že výsledky a závěry metod popisné statistiky se vztahují pouze a jedině k předloženému datovému souboru a jeho hranice nikdy nepřekročí.

Snahou každého výzkumníka je však naopak poznat a používat metody, které jsou schopné hranice datového souboru překročit a umožnit mu rozšíření informací získaných na základě datového souboru na celou zkoumanou populaci. V praxi totiž častokrát nemáme možnost zkoumat výskyt nějaké vlastnosti v celé populaci, neboť zkoumaná populace může být velmi rozsáhlá a nasbírání hodnot od každého subjektu z této populace by bylo časově i finančně velmi náročné. Z tohoto důvodu je pro nás mnohem jednodušší sestavit pouze reprezentativní vzorek subjektů ze zkoumané populace, který svým složením jednak dostatečně pokrývá celou populaci a jednak dostatečně reprezentuje její stěžejní rysy. Tento reprezentativní vzorek potom vyhodnotíme pomocí vhodných statistických metod a závěry platné pro reprezentativní vzorek následně rozšíříme na celou populaci (tento krok si v případě, že vybraný vzorek je skutečně reprezentativním vzorkem celé populace, můžeme dovolit).

Reprezentativní vzorek, ve statistické terminologii nazývaný jako *náhodný výběr*, je soubor n stochasticky nezávislých náhodných veličin $X_1, \ldots, X_n$, které se řídí stejným modelem L s parametry θ, tj. $X_1 \sim L(\theta)$, $\ldots$, $X_n \sim L(\theta)$. Protože každá dílčí náhodná veličina se řídí stejným modelem $L(\theta)$, můžeme předpokládat, že celý náhodný výběr $X_1, \ldots, X_n$ se také řídí modelem $L(\theta)$. V praxi může být modelem $L(\theta)$ například alternativní model $\mathrm{Alt}(p)$, kde $\theta = p$, binomický model $\mathrm{Bin}(N,p)$, kde $\theta = (N,p)^T$, normální model $N(\mu, \sigma^2)$, kde $\theta = (\mu, \sigma^2)^T$, apod. Konkrétní číselné realizace náhodného výběru $X_1, \ldots X_n$ (značíme je malými písmeny $x_1, \ldots, x_n$), tvoří *datový soubor*.

V souvislosti s náhodným výběrem definujeme také pojem *statistika*, jako libovolnou funkci $T = T(X_1, \ldots, X_n)$ náhodného výběru, která žádným způsobem nezávisí na parametru θ. *Realizací statistiky* t potom označujeme statistiku T vyhodnocenou v realizaci náhodného výběru, tj. $t = T(x_1, \ldots, x_n)$.

Příklad 5.1. Repezentativní vzorek
Předpokládejme, že chceme provést studii zkoumající výšku žen ve věku 25–35 let v Jihomoravském kraji. V ideálním případě bychom oslovili všechny ženy v požadovaném věku s trvalým pobytem v Jihomoravském kraji, změřili jejich výšku, zaznamenali ji do tabulky a nasbíraná data statisticky vyhodnotili. Takový výzkum by byl však časově i finančně náročný a navíc není pravděpodobné, že bychom do studie dokázali zahrnout úplně všechny ženy Proto raději vytvoříme reprezentativní vzorek žen z Jihomoravského kraje o rozsahu například $n = 1000$. Aby byl vzorek reprezentativní, měl by rovnoměrně pokrývat ženy z celého Jihomoravského kraje. S využitím multihypergeometrického modelu modelu popsaného v kapitole 4 můžeme vypočítat, že pro zachování rovnoměrného pokrytí celého Jihomoravského kraje bychom měli oslovit přibližně 92 žen okresu Blansko, 320 žen z okresu Brno-město, 187 žen z okresu Brno-venkov, 98 žen z okresu Břeclav, 130 žen z okresu Hodonín, 77 žen z okresu Vyškov a 96 žen z okresu Znojmo. Volba žen v každém okrese by měla být čistě náhodná a měla by pokrývat celou věkovou kategorii 25–35 let. Reprezentativní vzorek, neboli náhodný výběr bude sestávat z $n = 1000$ náhodných veličin $X_1, \ldots, X_{1000}$, kde veličina X_1 bude popisovat výšku první ženy, $\ldots$, X_{1000} bude popisovat výšku tisící ženy. O každé náhodné veličině předpokládáme, že se řídí normálním modelem, tj. $X_1 \sim N(\mu, \sigma^2)$, $\ldots$, $X_{1000} \sim N(\mu, \sigma^2)$, kde střední hodnota μ i rozptyl σ^2 jsou shodné pro všechny náhodné veličiny. Potom tedy také o celém náhodném výběru předpokládáme, že se řídí normálním modelem, tj. $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Nyní se dostáváme do fáze, kdy všechny ženy změříme a zjistíme například, že první změřená žena měří 165 cm, druhá žena měří 168 cm, $\ldots$, tisící žena měří 163 cm. Zaznamenáním naměřených hodnot do tabulky získáme realizace náhodnách veličin $x_1 = 165$, $x_2 = 168$, $\ldots$, $x_{1000} = 163$, které společně tvoří datový soubor. ★

Příklad 5.2. Jednorozměrné statistiky
Mějme jeden náhodný výběr $X_1, \ldots, X_n$ o rozsahu $n \geq 2$. Příkladem statistiky pro tento náhodný výběr může být například výběrový průměr

$$M = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{1}$$

Všimněme si, že ve vzorci 1 výběrového průměru vystupují pouze hodnoty náhodného výběru $X_1, \ldots, X_n$ a rozsah náhodného výběru n. Libovolný parametr θ (např. μ, σ^2, p, apod.) se ve vzorci nevyskytuje.

Dalším příkladem jednorozměrné statistiky je výběrový rozptyl

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M)^2. \tag{2}$$

Opět si všimněme, že ve vzorci 2 výběrového rozptylu se vyskytují pouze hodnoty náhodného výběru $X_1, \ldots, X_n$, rozsah náhodného výběru n a výběrový průměr M, který, jakožto statistika, je funkcí náhodného výběru. Žádný parametr se ve vzorci nevyskytuje. Výběrový rozptyl je tedy opět pouze funkcí náhodného výběru. Posledním příkladem statistiky, který si uvedeme, je výběrová směrodatná odchylka definovaná jako odmocnina z výběrového rozptylu, tj.

$$S = \sqrt{S^2}. \tag{3}$$

Výběrová směrodatná odchylka je statistikou, neboť jde pouze o odmocninou statistiky nazývané výběrový rozptyl. ★

Příklad 5.3. Dvourozměrné statistiky
Nechť $(X_1, Y_1), \ldots, (X_n, Y_n)$ je náhodný výběr z dvourozměrného rozdělení, M_1 a M_2 jsou výběrové průměry a S_1^2 a S_2^2 jsou výběrové rozptyly. Příkladem dvourozměrné statistiky je výběrová kovariance

$$S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_1)(Y_i - M_2). \tag{4}$$

Všimněme si, že ve vzorci 4 výběrové kovariance vystupují kromě hodnot dvourozměrného náhodného výběru pouze výběrové průměry M_1 a M_2, které jsou statistikami a tedy funkcemi náhodného výběru, a rozsah n. Proto je výběrová kovariance rovněž statistika. Druhým příkladem dvourozměrné statistiky je výběrový korelační koeficient

$$R_{12} = \frac{S_{12}}{\sqrt{S_1^2 S_2^2}} = \frac{S_{12}}{S_1 S_2}. \tag{5}$$

Výběrový korelační koeficient je definován jako podíl výběrové kovariance, která je statistikou, a odmocniny ze součinu výběrových rozptylů S_1 a S_2, které jsou rovněž statistikami. Žádný parametr ve vzorci 5 nefiguruje, proto je výběrový korelační koeficient též statistikou. ★

5.2 Bodové odhady parametrů

Předpokládejme nyní, že náhodný výběr $X_1, \ldots, X_n$ se řídí nějakým modelem L s parametrem θ, tj. $X_1, \ldots, X_n \sim L(\theta)$. Skutečnou hodnotu parametru θ neznáme a bohužel ji nikdy znát nebudeme. Jde o teoretickou hodnotu, kterou není možné přesně stanovit. Hodnotu parametru θ můžeme ale na základě datového souboru alespoň odhadnout, přičemž můžeme stanovit buď bodový nebo intervalový odhad parametru θ. Bodovým odhadem parametru θ je statistika $T = T(X_1, \ldots, X_n)$, která nabývá hodnot blízkých hodnotě parametru θ, ať je hodnota tohoto parametru jakákoli. Neformálně vzato, bodovým odhadem parametru θ je jedno konkrétní číslo, které získáme jako realizaci nějaké statistiky.

V praxi se můžeme setkat s různými typy bodových odhadů. Nejlepším odhadem je tzv. *nestranný* bodový odhad parametru θ. Tento odhad skutečnou hodnotu parametru θ ani nepodhodnocuje, ani nenadhodnocuje a proto je nejlepším možným typem bodového odhadu. Opakem nestranného odhadu je *vychýlený* odhad. Takový odhad skutečnou hodnotu parametru θ buď systematicky podhodnocuje, nebo systematicky nadhodnocuje. Třetím typem odhadu je tzv *asymptoticky nestranný* odhad. Asymptoticky nestranný odhad parametru θ, stanovený na základě náhodného výběru s malým rozsahem n, je vychýlený, ale s rostoucím rozsahem náhodného výběru n jeho vychýlení klesá. Čím je tedy rozsah náhodného výběru použitého ke stanovení asymptotického odhadu parametru θ větší, tím více se stanovený odhad blíží k nestrannému odhadu.

Příklad 5.4. Bodový odhad parametru μ a parametru σ^2
Předpokládejme nyní, že $X_1, \ldots, X_n$, $n \geq 2$ je náhodný výběr řídící se modelem L se střední hodnotou μ a rozptylem σ^2, tj. μ a σ^2 jsou parametry rozdělení L, jejichž přesnou hodnotu nebudeme nikdy znát. Nechť dále M je výběrový

průměr a S^2 je výběrový rozptyl, tj. M a S^2 jsou statistiky vypočítané na základě náhodného výběru $X_1, \dots, X_n$. Potom výběrový průměr M je nestranným odhadem parametru μ a výběrový rozptyl je nestranným odhadem parametru σ^2. ★

Příklad 5.5. Bodový odhad parametru σ_{12} a parametru ρ
Nechť $(X_1, Y_1), \dots, (X_n, Y_n)$ je náhodný výběr řídící se dvourozměrným modelem L_2 s kovariancí σ_{12} a koeficientem korelace ρ, tj. σ_{12} a ρ jsou parametry rozdělení L_2, jejichž skutečnou hodnotu nebudeme nikdy znát. Nechť dále S_{12} je výběrové kovariance a R je výběrvý korelační koeficient, tj. S_{12} a R jsou statistiky vypočítané na základě dvourozměrného náhodného výběru. Potom výběrové kovariance je nestranným odhadem parametru σ_{12}, zatímco výběrový korelační koeficient je asymptoticky nestranným odhadem parametru ρ. ★

Příklad 5.6. Bodové odhady parametrů μ, σ^2 a σ normálního rozdělení
Načtěte datový soubor `01-one-sample-mean-skull-mf.txt` a odstraňte z načtených dat NA hodnoty. Mějme náhodnou veličinu X popisující *největší šířku mozkovny* u skeletů mužského pohlaví. Za předpokladu, že se náhodná veličina X řídí normálním modelem se střední hodnotou μ a rozptylem σ^2, tj. $X \sim N(\mu, \sigma^2)$, stanovte nestranný bodový odhad (a) střední hodnoty μ; (b) rozptylu σ^2; (c) směrodatné odchylky σ.

Řešení příkladu 5.6
Celkem máme k dispozici $n = 216$ náhodných veličin $X_1, \dots, X_{216}$, přičemž veličina X_1 popisuje největší šířku mozkovny u prvního skeletu, ..., X_{216} popisuje největší šířku mozkovny u dvěstěšestnáctého skeletu. Předpokládáme, že všechny náhodné veličiny se řídí normálním modelem se střední hodnotou μ a rozptylem σ^2, tj. $X_1 \sim N(\mu, \sigma^2)$,..., $X_{216} \sim N(\mu, \sigma^2)$. Protože se všechny náhodné veličiny řídí stejným modelem $N(\mu, \sigma^2)$, předpokládáme, že celý datový soubor se řídí týmž modelem $N(\mu, \sigma^2)$. Naměřením hodnoty největší šířky mozkovny každého skeletu jsme získali celkem 216 realizací náhodných veličin, konkrétně $x_1 = 145, \dots, x_{216} = 137$. Těchto 216 realizací tvoří společně datový soubor. Skutečnou hodnotu parametrů μ a σ^2 (resp. σ) nebudeme nikdy znát. Jejich hodnoty ale můžeme odhadnout pomocí nestranných bodových odhadů. Bodový odhad parametru μ stanovíme pomocí výběrového průměru, tj.

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{216}(124 + 127 + \dots + 149 + 149) = \frac{29\,632}{216} = 137.1852.$$

Bodový odhad parametru σ^2 stanovíme pomocí výběrového rozptylu, tj.

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2 \\ &= \frac{1}{215}\left((124 - 137.1852)^2 + (127 - 137.1852)^2 + \dots + (149 - 137.1852)^2 + (149 - 137.1852)^2\right) \\ &= \frac{1}{215}\left((-13.1852)^2 + (-10.1852)^2 + \dots + 11.8148^2 + 11.8148^2\right) \\ &= 23.27717 \doteq 23.2772. \end{aligned}$$

Konečně bodový odhad parametru σ stanovíme pomocí výběrové směrodatné odchylky, neboli jako odmocninu z výběrového rozptylu, tj.

$$s = \sqrt{s^2} = \sqrt{23.27717} = 4.824642 \doteq 4.8246.$$

Datový soubor načteme příkazem `read.delim()` a NA hodnoty odstraníme příkazem `na.omit()`. Pomocí operátoru `[]` vybereme z tabulky `data` pouze ty řádky, které se vztahují k mužským skeletům (`data$sex == 'm'`) a sloupec obsahující údaje o největší šířce mozkovny `'skull.B'`. Hodnotu výběrového průměru, resp. výběrového rozptylu můžeme dopočítat pomocí softwaru R přepisem vzorce 1, resp. 2 s použitím funkce `sum()`. Hodnotu výběrové směrodatné odchylky získáme odmocněním výběrového rozptylu s využitím funkce `sqrt()`. Druhou možností je vypočítat výběrový průměr pomocí funkce `mean()`, výběrový rozptyl pomocí funkce `var()` a výběrovou směrodatnou odchylku pomocí funkce `sd()`.

```r
data <- read.delim('01-one-sample-mean-skull-mf.txt')
data <- na.omit(data)
# head(data)
skull.BM <- data[data$sex == 'm', 'skull.B']
n <- length(skull.BM)
m.BM  <- 1 / n * sum(skull.BM)
s2.BM <- 1 / (n - 1) * sum((skull.BM - m.BM) ^ 2)
s.BM  <- sqrt(s2.BM)

mm.BM  <- mean(skull.BM)
ss2.BM <- var(skull.BM)
ss.BM  <- sd(skull.BM)

(tab <- data.frame(prumer = m.BM, rozptyl = s2.BM, sm.odch = s.BM))
```

```
    prumer  rozptyl  sm.odch
1 137.1852 23.27717 4.824642
```

Interpretace výsledků: Nestranný odhad střední hodnoty největší šířky mozkovny pro skelety mužského pohlaví je 137.19 mm. Nestranný odhad rozptylu (resp. směrodatné odchylky) největší šířky mozkovny pro skelety mužského pohlaví je 23.28 mm^2 (resp. 4.82 mm). To znamená, že největší šířka mozkovny skeletů mužského pohlaví se pohybuje okolo hodnoty 137.19 mm se směrodatnou odchylkou 4.82 mm.

Poznámka: Všimněme si, že hodnota výběrového průměru vypočítaná v příkladu 5.6 je totožná s hodnotou aritmetického průměru vypočítanou v příkladu **??**. Rozdíl je však v přístupu k výsledné hodnotě. V příkladu **??** jsme aritmetický průměr uvažovali jako hodnotu vztahující se pouze k datovému souboru. V příkladu 5.6 již pracujeme s informací, že výběrový průměr je nestranným odhadem střední hodnoty μ normálního rozdělení a tedy je možné ji brát jako výsledek relevantní pro celou populaci skeletů mužského pohlaví starověké egyptské populace.

Naopak srovnáme-li hodnotu výběrového rozptylu vypočítanou v příkladu 5.6 s hodnotou rozptylu vypočítanou v příkladu **??**, vidíme, že výsledky se mírně liší. Konkrétně hodnoty vypočítané v příkladu 5.6 jsou nepatrně vyšší než hodnoty vypočítané v příkladu **??**. Rozdíly v hodnotách jsou způsobeny použitím odlišných vzorců v obou příkladech. V příkladu **??** jsme k výpočtu rozptylu použili vzorec $\frac{1}{n}\sum(X_i - M)^2$, který má sice lepší interpretaci (jde o aritmetický průměr kvadrátů odchylek naměřených hodnot X_i od průměrné hodnoty M), ale není nestranným odhadem parametru σ^2. Jde o odhad vychýlený, který systematicky skutečnou hodnotu parametru σ^2 podhodnocuje. Naopak v příkladu 5.6 jsme k výpočtu rozptylu použili vzorec $\frac{1}{n-1}\sum(X_i - M)^2$, který je nestranným odhadem parametru σ^2.

Analogicky vidíme, že hodnota výběrové směrodatné odchylky vypočítané v příkladu 5.6 je nepatrně vyšší než hodnota směrodatné odchylky vypočítaná v příkladu **??**. Směrodatná odchylka vypočítaná v příkladu **??** je opět vychýleným odhadem parametru σ, který skutečnou hodnotu parametru systematicky podhodnocuje. Naopak výběrová směrodatná odchylka vypočítaná v příkladu 5.6 je nestranným odhadem parametru σ. ★

Příklad 5.7. Bodové odhady parametrů σ_{12} a ρ dvourozměrného normálního rozdělení
Načtěte datový soubor 01-one-sample-mean-skull-mf.txt a odstraňte z načtených dat NA hodnoty. Mějme náhodnou veličinu X popisující *největší šířku mozkovny* a náhodnou veličinu Y popisující *největší délku mozkovny* u skeletů mužského pohlaví. Za předpokladu, že se náhodný vektor $(X, Y)^T$ řídí dvourozměrným normálním modelem, tj. $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kde $\boldsymbol{\mu}$ je vektor středních hodnot a $\boldsymbol{\Sigma}$ je varianční matice, stanovte (a) nestranný bodový odhad kovariance σ_{12}; (b) asymptoticky nestranný bodový odhad korelačního koeficientu ρ.

Řešení příkladu 5.7

Celkem máme k dispozici $n = 216$ dvojic náhodných veličin $(X_1, Y_1) \ldots, (X_{216}, Y_{216})$, přičemž veličina X_1 popisuje největší šířku mozkovny u prvního skeletu, ..., X_{216} popisuje největší šířku mozkovny u dvěstěšestnáctého skeletu a veličina Y_1 popisuje největší délku mozkovny u prvního skeletu, ..., Y_{216} popisuje největší délku mozkovny u dvěstěšestnáctého skeletu. Předpokládáme, že všechny dvojice náhodných veličin se řídí dvourozměrným normálním modelem s vektorem středních hodnot $\boldsymbol{\mu}$ a varianční maticí $\boldsymbol{\Sigma}$, tj. $(X_1, Y_1) \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,..., $(X_{216}, Y_{216}) \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Protože se všechny dvojice náhodných veličin řídí stejným modelem $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, předpokládáme, že celý datový soubor se řídí týmž modelem $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Naměřením hodnot největší šířky a největší délky mozkovny každého skeletu jsme získali celkem 216 dvojic realizací náhodných veličin, konkrétně $(x_1, y_1) = (145, 188), \ldots, (x_{216}, y_{216}) = (137, 186)$. Těchto 216 dvojic realizací tvoří společně datový soubor. Skutečnou hodnotu parametrů σ_{12} a ρ nebudeme nikdy znát. Jejich hodnoty ale můžeme odhadnout pomocí bodových odhadů. Nestranný bodový odhad parametru σ_{12} stanovíme pomocí výběrové kovariance, tj.

$$s_{12} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m_1)(y_i - m_2),$$

kde $m_1 = 137.1851$ je výběrový průměr největší šířky mozkovny (viz příklad 5.6) a $m_2 = 182.0324$ je výběrový průměr největší délky mozkovny. Hodnotu výběrového průměru m_2 získáme analogickým postupem uvedeným v příkladu 5.6. Výběrovou kovarianci potom dopočítáme jako

$$\begin{aligned}
s_{12} &= \frac{1}{215}((145 - 137.1851)(188 - 182.0324) + (139 - 137.1851)(172 - 182.0324) + \ldots \\
&\qquad \cdots + (142 - 137.1851)(183 - 182.0324) + (137 - 137.1851)(186 - 182.0324)) \\
&= \frac{1}{215}(46.6356 - 18.2068 + \cdots + 4.6588 - 0.7348) \\
&= \frac{1113.7037}{215} = 5.1800172265 \doteq 5.1800.
\end{aligned}$$

Asymptoticky nestranný bodový odhad parametru ρ stanovíme pomocí výběrového korelačního koeficientu, tj.

$$r_{12} = \frac{s_{12}}{\sqrt{s_1^2 s_2^2}} = \frac{s_{12}}{s_1 s_2}, \tag{6}$$

kde s_{12} je výběrové kovariance (viz výše), $s_1^2 = 23.2772$ je výběrový rozptyl největší šířky mozkovny (viz příklad 5.6) a $s_2^2 = 40.7664$ je výběrový rozptyl největší délky mozkovny. Hodnotu výběrového rozptylu s_2^2 získáme analogickým postupem uvedeným v příkladu 5.6. Výběrový korelační koeficient potom dopočítáme jako

$$r_{12} = \frac{5.1800}{\sqrt{23.2772}\sqrt{40.7664}} = \frac{5.1800}{4.8246 \times 6.3849} = \frac{5.1800}{30.8046} = 0.1682.$$

Hodnotu výběrové kovariance, resp. výběrového korelačního koeficientu můžeme dopočítat pomocí softwaru R přepisem vzorce 4, resp. 5 s použitím funkcí `sum()` a `sqrt()`. Druhou možností je vypočítat výběrovou kovarianci pomocí funkce `cov()` a výběrový korelační koeficient pomocí funkce `cor()`.

```
  kovariance korelacni_koeficient
1   5.180017             0.168157
```

Interpretace výsledků: Nestranný odhad kovariance největší šířky a délky mozkovny pro skelety mužského pohlaví je 5.18 mm^2. Asymptoticky nestranný odhad korelačního koeficientu největší šířky a délky mozkovny pro skelety mužského pohlaví je 0.1682 . To znamená, že mezi největší šířkou a délkou mozkovny u skeletů mužského pohlaví existuje nízký stupeň přímé lineární závislosti.

Poznámka: Všimněme si, že hodnota výběrového korelačního koeficientu vypočítaná v příkladu 5.7 je totožná s hodnotou korelačního koeficientu vypočítanou v příkladu **??**. Rozdíl je však v přístupu k výsledné hodnotě. V příkladu **??** jsme korelační koeficient uvažovali jako hodnotu vztahující se pouze k datovému souboru. V příkladu 5.7 již pracujeme s informací, že výběrový korelační koeficient je asymptoticky nestranným odhadem parametru ρ dvourozměrného normálního rozdělení a tedy je možné jej brát jako výsledek relevantní pro celou populaci skeletů mužského pohlaví starověké egyptské populace. Neměli bychom však zapomínat na to, že výběrový korelační koeficient je pouze asymptoticky nestranným odhadem parametru ρ a tedy jeho vychýlení klesá s rozsahem náhodného výběru. Rozsah náhodného výběru mužských skeletů, $n = 216$, je však dostatečně vysoký a tedy odhad parametru ρ můžeme považovat za nestranný. ★

Příklad 5.8. Bodový odhad vektoru středních hodnot $\boldsymbol{\mu}$ a varianční matice $\boldsymbol{\Sigma}$ dvourozměrného normálního rozdělení

Načtěte datový soubor 01-one-sample-mean-skull-mf.txt a odstraňte z načtených dat NA hodnoty. Mějme náhodnou veličinu X popisující *největší šířku mozkovny* a náhodnou veličinu Y popisující *největší délku mozkovny* u skeletů mužského pohlaví. Za předpokladu, že se náhodný vektor $(X, Y)^T$ řídí dvourozměrným normálním modelem, tj. $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kde $\boldsymbol{\mu}$ je vektor středních hodnot a $\boldsymbol{\Sigma}$ je varianční matice, stanovte (a) nestranný bodový odhad vektoru středních hodnot $\boldsymbol{\mu}$; (b) asymptoticky nestranný bodový odhad varianční matice $\boldsymbol{\Sigma}$.

Řešení příkladu 5.8

Předpokládáme, že náhodný vektor $(X, Y)^T$ se řídí dvourozměrným normálním modelem, tj. $(X, Y)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ s vektorem středních hodnot $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ a varianční maticí

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix},$$

kde μ_1 je střední hodnota náhodné veličiny X, μ_2 je střední hodnota náhodné veličiny Y, σ_1^2 je rozptyl náhodné veličiny X, σ_2^2 je rozptyl náhodné veličiny Y a ρ je korelační koeficient popisující vztah mezi veličinami X a Y.

Z příkladů 5.6 a 5.7 známe výběrové průměry $m_1 = 137.1851$ a $m_{=}182.0324$, které jsou nestrannými odhady středních hodnot μ_1 a μ_2, výběrové rozptyly $s_1^2 = 23.2772$ a $s_2^2 = 40.7664$, které jsou nestrannými odhady rozptylů σ_1^2 a σ_2^2, výběrové směrodatné odchylky $s_1 = 4.8246$ a $s_2 = 6.3842$, které jsou nestrannými odhady odchylek σ_1 a σ_2 a výběrový korelační koeficient $r_{12} = 0.1681$, který je nestranným odhadem korelačního koeficientu ρ. Nestranným odhadem vektoru středních hodnot $\boldsymbol{\mu}$ je potom vektor $(137.1851, 182.0324)^T$. Asymptoticky nestranným odhadem varianční matice $\boldsymbol{\Sigma}$ je matice

$$\begin{pmatrix} 23.2772 & 0.1681 \times 4.8246 \times 6.3842 \\ 0.1681 \times 4.8246 \times 6.3842 & 40.7664 \end{pmatrix} = \begin{pmatrix} 23.2772 & 5.1800 \\ 5.1800 & 40.7664 \end{pmatrix}.$$

★

Příklad 5.9. Bodový odhad parametru p alternativního modelu

Načtěte datový soubor 17-anova-newborns.txt a odstraňte z načtených dat NA hodnoty. Mějme náhodnou veličinu X popisující *ženské pohlaví* novorozenců. Za předpokladu, že náhodná veličina X pochází z alternativního rozdělení s parametrem p, tj. $X \sim Alt(p)$, kde p je pravděpodobnost narození holčičky, stanovte bodový odhad parametru p.

Řešení příkladu 5.9

Celkem máme k dispozici 1382 náhodných veličin $X_1, \ldots, X_{1382}$, přičemž veličina X_1 popisuje výskyt události narození holčičky ($X_1 = 1$; úspěch), nebo výskyt události narození chlapečka ($X_1 = 0$; neúspěch) u první matky, ..., X_{1382} popisuje výskyt události narození holčičky ($X_{1382} = 1$; úspěch) nebo chlapečka ($X_{1382} = 0$; neúspěch) u tisící třísté osmdesáté druhé matky. Za předpokladu, že všechny náhodné veličiny se řídí alternativním modelem se stejným parametrem p, tj. $X_1 \sim \mathrm{Alt}(p)$, ..., $X_{1382} \sim \mathrm{Alt}(p)$, se také celý náhodný výběr řídí týmž alternativním modelem, tj. $X \sim \mathrm{Alt}(p)$. Parametr p určuje pravděpodobnost narození holčičky u jedné matky. Skutečnou hodnotu parametru p nebudeme nikdy znát, můžeme ji ale odhadnout pomocí nestranného bodového odhadu.

Vektor X obsahující údaje o pohlaví novorozenců je soubor 1 (narození holčičky) a 0 (narození chlapečka). Odhad parametru p získáme opět pomocí výběrového průměru, tj.

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{1382}\,(0 + 0 + \cdots + 1 + 0) = \frac{663}{1382} = 0.4797. \tag{7}$$

Všimněme si, že odhad parametru p není nic jiného, než celkový počet narozených holčiček (čitatel vzorce 7) ku celkovému počtu všech novorozenců (jmenovatel vzorce 7), což je vlastně relativní četnost výskytu holčiček v datovém souboru. Výběrový průměr sestavený nad vektorem nul a jedniček je tedy roven relativní četnosti.

Nejprve načteme datový soubor příkazem read.delim() a odstraníme NA hodnoty příkazem na.omit(). Dále do proměnné sex vložíme údaje o pohlaví. Bližším prozkoumáním vektoru sex zjistíme, že jde o proměnnou typu factor, která nabývá dvou úrovní, a sice úrovně 1 (s popiskem 'f' (female)) a úrovně 2 (s popiskem 'm' (male)).

```
data <- read.delim('17-anova-newborns.txt')
data <- na.omit(data)
sex  <- data$sex.C
head(sex)
```

```
[1] m m f m m m
Levels: f m
```

Protože chceme vektor sex použít k odhadu parametru p, upravíme si jej nejprve do vhodné číselné podoby. Pomocí funkce as.numeric() převedeme faktor na číselný vektor a vložíme jej do proměnné pohlavi. Vidíme, že vektor pohlavi si zachoval původní kódování 1 = female, 2 = male. Ve vektoru pohlavi tedy změníme všechny hodnoty 2 na hodnoty 0, čímž dostaneme požadované kódování 0 = male, 1 = female.

```
pohlavi <- as.numeric(sex)
pohlavi[pohlavi == 2] <- 0
head(pohlavi)
```

```
[1] 0 0 1 0 0 0
```

Odhad parametru p nyní získáme buď přepisem vzorce 7, nebo funkcí mean(). Nakonec si ověříme že hodnota výběrového průměru je shodná s hodnotou relativní četnosti vypočítané pomocí původního faktoru sex.

```
N   <- length(pohlavi)
m   <- 1 / N * sum(pohlavi)
mm  <- mean(pohlavi)
p   <- sum(sex == 'f') / N
tab <- data.frame(m, mm, p)
round(tab, 4)
```

```
       m     mm      p
1 0.4797 0.4797 0.4797
```

Interpretace výsledků: Nestranný odhad pravděpodobnosti narození holčičky je 0.4797. To znamená, že k narození holčičky u jedné matky dojde s pravděpodobností 47.97 % ★

5.3 Intervalové odhady parametrů

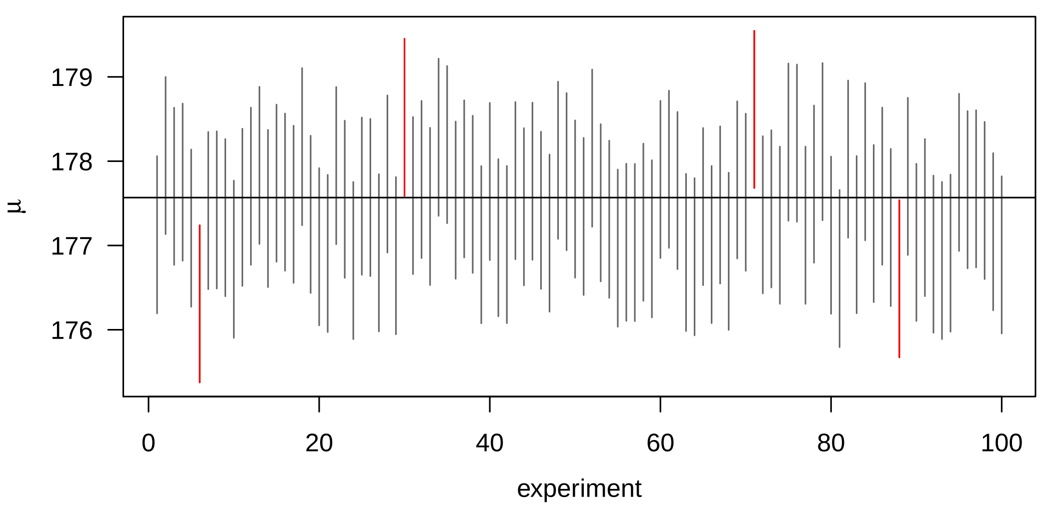
Hodnotu parametru θ modelu L, ze kterého pochází náhodný výběr $X_1, \ldots, X_n$ zkusíme nyní odhadnout pomocí tzv. intervalového odhadu. Zatímco bodový odhad parametru θ je jedno číslo (vypočítané na základě vhodné statistiky), intervalový odhad parametru θ je interval (D, H), který s dostatečně velkou pravděpodobností pokrývá hodnotu parametru θ. Hranice intervalového odhadu tvoří opět vhodné statistiky, neboli funkce náhodného výběru, tj. $D = D(x_1, \ldots, X_n)$ a $H = H(X_1, \ldots, X_n)$. Intervalový odhad nazýváme ve statistické terminologii jako interval spolehlivosti. Všechny zde prezentované intervaly spolehlivosti jsou intervaly spolehlivosti Waldova typu, nazývané zkráceně Waldovy intervaly spolehlivosti.

Mějme nyní riziko α, což je koeficient nabývající hodnoty z intervalu $(0, 1)$. Tento koeficient určuje pravděpodobnost, s jakou interval spolehlivosti nepokrývá hodnotu parametru θ. Doplňkem k riziku α je tzv. koeficient spolehlivosti $(1 - \alpha)$ určující pravděpodobnost, s jakou interval spolehlivosti pokrývá hodnotu parametru θ. Podle potřeby volíme nejčastěji hodnotu rizika $\alpha = 0.1$ (koeficient spolehlivosti $1 - \alpha = 0.90$, tj. pravděpodobnost, že interval spolehlivosti pokrývá hodnotu parametru θ je 90 %), $\alpha = 0.05$ (koeficient spolehlivosti $1 - \alpha = 0.95$, tj. pravděpodobnost, že interval spolehlivosti pokrývá hodnotu parametru θ je 95 %) nebo $\alpha = 0.01$ (koeficient spolehlivosti $1 - \alpha = 0.99$, tj. pravděpodobnost, že interval spolehlivosti pokrývá hodnotu parametru θ je 99 %).

$(1 - \alpha)$% pravděpodobnost, že interval spolehlivosti pokrývá hodnotu parametru θ chápeme ve smyslu, že kdybychom nasbírali n náhodných výběrů a na základě každého z nich vypočítali intervalový odhad zkoumaného parametru θ, potom alespoň v $(1 - \alpha) \times n$ případech by vypočítaný interval spolehlivosti pokrýval (obsahoval) skutečnou hodnotu parametru θ a ve zbylých ($\alpha \times n$ a méně) případech by interval spolehlivosti skutečnou hodnotu parametru θ nepokrýval (neobsahoval).

Příklad 5.10. Pravděpodobnost pokrytí $100 \times (1 - \alpha)\%$ **intervalu spolehlivosti**
Předpokládejme, že náhodná veličina X popisující největší šířku lebky novověké egyptské mužské populace, se řídí normálním modelem se střední hodnotou $\mu = 177.568$ a rozptylem $\sigma^2 = 7.526^2$, tj. $X \sim N(177.568, 7.526^2)$. Zde tedy na chvíli připusťme, že známe skutečnou hodnotu parametru μ i skutečnou hodnotu parametru σ^2.

Představme si nyní, že jsme v rámci jednoho experimentu vybrali náhodný vzorek 250 mužů z novodobé egyptské populace a změřili největší šířku jejich lebky. Získali jsme náhodný výběr $\mathbf{X_1} = (X_{1,1}, \ldots, X_{1,250})$. Na základě náhodného výběru jsme stanovili 95% Waldův empirický interval spolehlivosti pro parametr μ a následně jsme zkontrolovali, zda skutečná hodnota parametru $\mu = 177.568$ náleží do vypočítaného intervalu spolehlivosti, nebo nikoli. Analogický experiment jsme následně zopakovali stokrát. V alespoň 95 případech ze 100 experimentů vypočítaný Waldův empirický interval spolehlivosti pokrývá skutečnou hodnotu parametru $\mu = 177.568$, zatímco v pěti a méně případech ze 100 experimentů vypočítaný interval spolehlivosti skutečnou hodnotu parametru μ nepokrývá (viz obrázek 1). ★

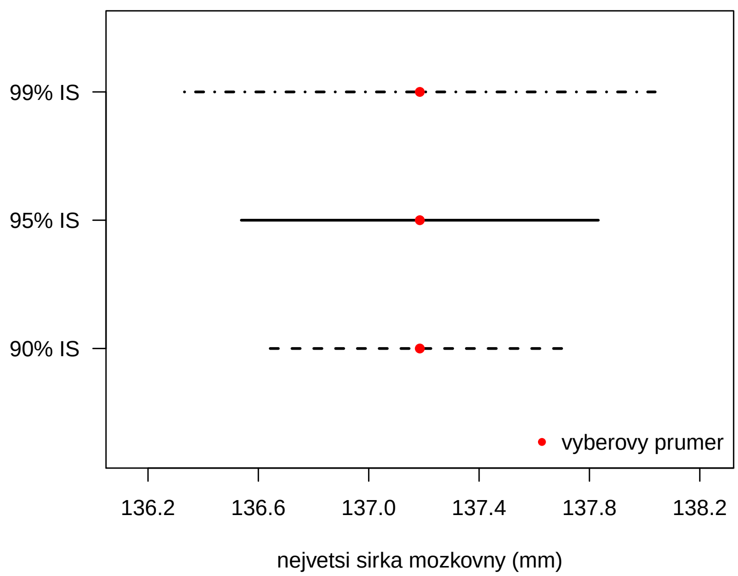


Obrázek 1: Pravděpodobnost pokrytí 95 % Waldova empirického intervalů spolehlivosti pro parametr μ normálního rozdělení při známém rozptylu σ^2

Naším cílem je nalézt takový interval spolehlivosti, který je jednak rozumně široký a který pokrývá skutečnou hodnotu parametru θ s co největší pravděpodobností. Bohužel s rostoucí pravděpodobností pokrytí parametru θ intervalem spolehlivosti roste také šířka tohoto intervalu. Proto volba výše pokrytí parametru θ intervalem spolehlivosti je vždy otázka kompromisu. Představme si, že bychom chtěli sestrojit 100% interval spolehlivosti, tedy interval, který pokrývá parametr θ se 100% pravděpodobností. Takový interval existuje a má tvar $(-\infty\,;\,\infty)$ (se 100% pravděpodobností bude skutečná hodnota parametru θ nabývat hodnoty mezi $-\infty$ a ∞). Takový interval spolehlivosi nám však příliš platný není. Kdybychom se naopak rozhodli, že sestrojíme co nejpřesnější interval spolehlivosti, tj. interval, který bude mít co nejmenší šířku, sestrojili bychom 1% interval spolehlivosti. Šířka tohoto intervalu by byla tak malá, že by se intervalový odhad blížil bodovému odhadu. Ovšem pravděpodobnost, že skutečná hodnota parametru θ náleží do toho úzkého intervalu by byla pouhé 1% (s 1% pravděpodobností skutečná hodnota parametru θ náleží do intervalu spolehlivosti, ale s 99% pravděpodobností ne). Tento interval je tedy také ne příliš užitečný. Proto volíme hodnotu rizika α jako 0.1, 0.05 nebo 0.01, protože odpovídající koeficient spolehlivosti $1-\alpha$ (0.9, 0.95 nebo 0.99) zajišťuje přijatelnou šířku intervalu spolehlivosti při zachování velmi vysoké pravděpodobnosti pokrytí parametru θ.

Jak je uvedeno výše, šířka intervalu spolehlivosti roste s rostoucím koeficientem spolehlivosti $1-\alpha$, neboli s rostoucí pravděpodobností pokrytí parametru θ. Porovnáme-li tedy navzájem 90%, 95% a 99% interval spolehlivosti, bude šířka 90% intervalu spolehlivosti menší než šířka 95% intervalu spolehlivosti a ta bude menší než šířka 99% intervalu spolehlivosti.

Příklad 5.11. Porovnání šířky Waldových empirických intervalů spolehlivosti
Načtěte datový soubor `01-one-sample-mean-skull-mf.txt` a odstraňte z načtených dat NA hodnoty. Mějme náhodnou veličinu X popisující *největší šířku mozkovny* u skeletů mužského pohlaví. Za předpokladu, že náhodná veličina X pochází z normálního rozdělení se střední hodnotou μ a rozptylem σ^2, tj. $X \sim N(\mu, \sigma^2)$, vypočítejte 90%, 95% a 99% Waldův empirický interval spolehlivosti pro parametr μ. Následně porovnejte šířky těchto intervalů.



Obrázek 2: Porovnání šířky 90%, 95% a 99% Waldova empirického intervalu spolehlivosti pro parametr μ normálního modelu

Poznámka: Přesný postup výpočtu intervalů spolehlivosti si názorně ukážeme později, v příkladu 5.12.

Z obrázku 2 vidíme, že skutečně nejméně široký je 90% Waldův empirický interval spolehlivosti. Naopak nejširší je 99% Waldův empirický interval spolehlivosti. 95% Waldův empirický interval spolehlivosti má šířku větší než 90% interval spolehlivosti, ale menší než 99% interval spolehlivosti.

★

Rozlišujeme tři základní typy intervalů spolehlivosti, a sice $100 \times (1-\alpha)\%$ oboustranný interval spolehlivosti (D, H)

pro parametr θ, $100\times(1-\alpha)\%$ levostranný interval spolehlivosti (D,∞) pro parametr θ a $100\times(1-\alpha)\%$ pravostranný interval spolehlivosti $(-\infty,H)$ pro parametr θ. Ve všech třech případech je pravděpodobnost, že parametr θ náleží do intervalu spolehlivosti, alespoň $(1-\alpha)\times 100\%$, tj. $\Pr(\theta\in IS)\geq(1-\alpha)\times 100\%$. V rámci tohoto textu se zaměříme na konstrukci (oboustranného / levostranného / pravostranného) intervalu spolehlivosti pro parametry μ a σ^2 normálního modelu a parametru p alternativního modelu. Pro parametr ρ dvourozměrného normálního modelu existuje několik typů intervalů spolehlivosti. Jejich konstrukcí se budeme zabývat v kapitole **??**.

Nyní si představíme konkrétní tvary jednotlivých intervalů spolehlivosti. Na níže uvedené vzorce se potom budeme odkazovat v řešených příkladech.

Předpokládejme nejprve, že náhodný výběr $X_1,\dots,X_n$ je náhodný výběr, který se řídí normálním modelem se střední hodnotou μ a rozptylem σ^2, tj. $X\sim N(\mu,\sigma^2)$, kde parametr σ^2 známe. Nechť m značí realizaci výběrového průměru, σ směrodatnou odchylku vypočítanou jako $\sigma=\sqrt{\sigma^2}$, n je rozsah náhodného výběru a u_α (resp.$u_{\alpha/2}$, $u_{1-\alpha/2}$, $u_{1-\alpha}$) je α-kvantil (resp. $\frac{\alpha}{2}$-kvantil, $(1-\frac{\alpha}{2})$-kvantil, $(1-\alpha)$-kvantil) standardizovaného normálního modelu. $100\times(1-\alpha)\%$ oboustranný Waldův empirický interval spolehlivosti pro parametr μ, když σ^2 známe, má tvar

$$(d,h)=(m-\frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}\,,\,m-\frac{\sigma}{\sqrt{n}}u_{\alpha/2}). \tag{8}$$

$100\times(1-\alpha)\%$ levostranný Waldův empirický interval spolehlivosti pro parametr μ, když σ^2 známe, má tvar

$$(d,\infty)=(m-\frac{\sigma}{\sqrt{n}}u_{1-\alpha}\,,\,\infty). \tag{9}$$

$100\times(1-\alpha)\%$ pravostranný Waldův empirický interval spolehlivosti pro parametr μ, když σ^2 známe, má tvar

$$(-\infty,h)=(-\infty\,,\,m-\frac{\sigma}{\sqrt{n}}u_{\alpha}). \tag{10}$$

Hodnotu α-kvantilu standardizovaného normálního rozdělení u_α vypočítáme pomocí softwaru R příkazem `qnorm(alpha)`. Analogicky můžeme získat hodnoty kvantilů $u_{\alpha/2}$, $u_{1-\alpha}$ a $u_{1-\alpha/2}$.

Poznámka: V praxi se se situací, kdy odhadujeme parametr střední hodnoty μ a přitom známe skutečnou hodnotu rozptylu σ^2, spíše nesetkáme. Jak jsme si uvedli, skutečná hodnota parametru σ^2 nám není známá. Intervalové odhady 8, 9 a 10 se tedy spíše využívají při simulačních studiích, v rámci kterých zkoumáme různé vlastnosti těchto odhadů (např. změnu polohy intervalu spolehlivosti při měnící se hodnotě parametru μ, nebo změnu šířky intervalu spolehlivosti při měnící se hodnotě parametru σ^2, rozsahu náhodného výběru n nebo koeficientu spolehlivosti $1-\alpha$, apod.). Dále budeme tyto intervaly spolehlivosti využívat v sekci **??** zabývající se testováním hypotéz. Zde budeme interval spolehlivosti pro parametr μ když σ^2 známe používat v případě, kdy budeme porovnávat střední hodnotu našeho náhodného výběru se střední hodnotou publikovanou, společně se svým rozptylem, v literatuře.

Předpokládejme nyní, že náhodný výběr $X_1,\dots,X_n$ je náhodný výběr, který se řídí normálním modelem se střední hodnotou μ a rozptylem σ^2, tj. $X\sim N(\mu,\sigma^2)$, kde parametr σ^2 neznáme. Nechť m značí realizaci výběrového průměru, s značí realizaci výběrové směrodatné odchylky, n je rozsah náhodného výběru a $t_{n-1}(\alpha)$ (resp.$t_{n-1}(\alpha/2)$, $t_{n-1}(1-\alpha/2)$, $t_{n-1}(1-\alpha)$) je α-kvantil (resp. $\frac{\alpha}{2}$-kvantil, $(1-\frac{\alpha}{2})$-kvantil, $(1-\alpha)$-kvantil) Studentova modelu o $n-1$ stupních volnosti. $100\times(1-\alpha)\%$ oboustranný Waldův empirický interval spolehlivosti pro parametr μ, když σ^2 neznáme, má tvar

$$(d,h)=(m-\frac{s}{\sqrt{n}}t_{n-1}(1-\alpha/2)\,,\,m-\frac{s}{\sqrt{n}}t_{n-1}(\alpha/2)). \tag{11}$$

$100\times(1-\alpha)\%$ levostranný Waldův empirický interval spolehlivosti pro parametr μ, když σ^2 neznáme, má tvar

$$(d,\infty)=(m-\frac{s}{\sqrt{n}}t_{n-1}(1-\alpha)\,,\,\infty). \tag{12}$$

$100\times(1-\alpha)\%$ pravostranný Waldův empirický interval spolehlivosti pro parametr μ, když σ^2 neznáme, má tvar

$$(-\infty,h)=(-\infty\,,\,m-\frac{s}{\sqrt{n}}t_{n-1}(\alpha)). \tag{13}$$

Hodnotu α-kvantilu Studentova modelu s $n-1$ stupni volnosti $t_{n-1}(\alpha)$ vypočítáme pomocí softwaru R příkazem qt(alpha, n-1). Analogicky můžeme získat hodnoty kvantilů $t_{n-1}(\alpha/2)$, $t_{n-1}(1-\alpha)$ a $t_{n-1}(1-\alpha/2)$.

Předpokládejme dále, že $X_1, \dots, X_n$ je náhodný výběr, který se řídí normálním modelem se střední hodnotou μ a rozptylem σ^2, tj. $X \sim N(\mu, \sigma^2)$, kde parametr μ neznáme. Nechť s^2 značí realizaci výběrového rozptylu, n je rozsah náhodného výběru a $\chi^2_{n-1}(\alpha)$ (resp.$\chi^2_{n-1}(\alpha/2)$, $\chi^2_{n-1}(1-\alpha/2)$, $t\chi^2_{n-1}(1-\alpha)$) je α-kvantil (resp. $\frac{\alpha}{2}$-kvantil, $(1-\frac{\alpha}{2})$-kvantil, $(1-\alpha)$-kvantil) χ^2 modelu o $n-1$ stupních volnosti. $100\times(1-\alpha)\%$ oboustranný Waldův empirický interval spolehlivosti pro parametr σ^2, když μ neznáme, má tvar

$$(d,h) = \left(\frac{(n-1)s^2}{\chi^2_{n-1}(1-\alpha/2)}\,,\, \frac{(n-1)s^2}{\chi^2_{n-1}(\alpha/2)} \right). \tag{14}$$

$100\times(1-\alpha)\%$ levostranný Waldův empirický interval spolehlivosti pro parametr σ^2, když μ neznáme, má tvar

$$(d,\infty) = \left(\frac{(n-1)s^2}{\chi^2_{n-1}(1-\alpha)}\,,\, \infty \right). \tag{15}$$

$100\times(1-\alpha)\%$ pravostranný Waldův empirický interval spolehlivosti pro parametr σ^2, když μ neznáme, má tvar

$$(0,h) = \left(0\,,\, \frac{(n-1)s^2}{\chi^2_{n-1}(\alpha)} \right). \tag{16}$$

Hodnotu α-kvantilu χ^2 modelu s $n-1$ stupni volnosti $\chi^2_{n-1}(\alpha)$ vypočítáme pomocí softwaru R příkazem qchisq(alpha, n-1). Analogicky můžeme získat hodnoty kvantilů $\chi^2_{n-1}(\alpha/2)$, $\chi^2_{n-1}(1-\alpha)$ a $\chi^2_{n-1}(1-\alpha/2)$.

Protože parametr rozptylu σ^2 je z definice vždy větší než 0, stanovuje se hodnota dolní hranice $100\times(1-\alpha)\%$ pravostranného Waldova empirického intervalu spolehlivosti jako nula namísto nekonečna.

Konečně předpokládejme, že $X_1, \dots, X_n$ je náhodný výběr, který se řídí alternativním modelem s parametrem p, tj. $X \sim \mathrm{Alt}(p)$. Nechť m značí realizaci výběrového průměru, N je rozsah náhodného výběru a u_α (resp.$u_{\alpha/2}$, $u_{1-\alpha/2}$, $u_{1-\alpha}$) je α-kvantil (resp. $\frac{\alpha}{2}$-kvantil, $(1-\frac{\alpha}{2})$-kvantil, $(1-\alpha)$-kvantil) standardizovaného normálního modelu. $100\times(1-\alpha)\%$ oboustranný Waldův empirický interval spolehlivosti pro parametr p má tvar

$$(d,h) = \left(m - \sqrt{\frac{m(1-m)}{N}}u_{1-\alpha/2}\,,\, m - \sqrt{\frac{m(1-m)}{N}}u_{\alpha/2} \right). \tag{17}$$

$100\times(1-\alpha)\%$ levostranný Waldův empirický interval spolehlivosti pro parametr p má tvar

$$(d,1) = \left(m - \sqrt{\frac{m(1-m)}{N}}u_{1-\alpha}\,,\, 1 \right). \tag{18}$$

$100\times(1-\alpha)\%$ pravostranný Waldův empirický interval spolehlivosti pro parametr p má tvar

$$(0,h) = \left(0\,,\, m - \sqrt{\frac{m(1-m)}{N}}u_{\alpha} \right). \tag{19}$$

Hodnotu α-kvantilu standardizovaného normálního rozdělení u_α vypočítáme pomocí softwaru R příkazem qnorm(alpha). Analogicky můžeme získat hodnoty kvantilů $u_{\alpha/2}$, $u_{1-\alpha}$ a $u_{1-\alpha/2}$.

Protože parametr p značí pravděpodobnost úspěchu v jednom pokusu, platí, že $p \in (0,1)$, a tedy horní hranice levostranného Waldova empirického intervalu spolehlivosti je 1 (viz vzorec 18). Analogicky dolní hranice pravostranného Waldova empirického intervalu spolehlivosti je 0 (viz vzorec 19).

Příklad 5.12. Intervalový odhad parametru μ normálního modelu

Načtěte datový soubor 01-one-sample-mean-skull-mf.txt a odstraňte z načtených dat NA hodnoty. Mějme náhodnou veličinu X popisující *největší šířku mozkovny* u skeletů mužského pohlaví. Za předpokladu, že náhodná veličina X

pochází z normálního rozdělení se střední hodnotou μ a rozptylem σ^2, tj. $X \sim N(\mu, \sigma^2)$, vypočítejte 90%, 95% a 99% Waldův empirický interval spolehlivosti pro parametr μ.

Řešení příkladu 5.12

Není-li v textu specifikován typ požadovaného intervalu spolehlivosti, počítáme vždy oboustranný interval spolehlivosti. Ze zadání víme, že chceme spočítat interval spolehlivosti pro parametr μ. Dále si všimněme, že ze zadání příkladu není známá skutečná hodnota rozptylu σ^2. Naším úkolem je tedy vypočítat 90% (resp. 95%, či 99%) Waldův empirický oboustraný interval spolehlivosti pro parametr μ když σ^2 neznáme.

Při výpočtu intervalů spolehlivosti budeme vycházet ze vzorce 11. Z příkladu 5.6 víme, že realizace výběrového průměru $m = 137.1852$ a rozsah náhodného výběru $n = 216$. Hodnotu směrodatné odchylky σ odhadhneme pomocí výběrové směrodatné odchylky $s = 4.8246$ (viz příklad 11). Zbývá stanovit hodnotu kvantilu $t_{n-1}(\alpha/2)$ a kvantilu $t_{n-1}(1-\alpha/2)$ Studentova modelu. K tomu je potřeba nejprve dopočítat koeficient α. Ten vyjádříme, v případě výpočtu 90% Waldova empirického intervalu spolehlivosti, postupnými kroky z rovnice $100 \times (1-\alpha)\,\% = 90\,\%$.

$$\begin{aligned}
100 \times (1-\alpha)\,\% &= 90\,\% \\
100 \times (1-\alpha) &= 90 \\
1-\alpha &= 0.90 \\
1-0.90 &= \alpha \\
\alpha &= 0.10
\end{aligned}$$

Pomocí softwaru R a funkce $\mathsf{qt()}$ nyní stanovíme hodnotu kvantilu $t_{n-1}(\alpha/2) = t_{215}(0.10/2) = t_{215}(0.05) = \mathsf{qt}(0.05, 215) = -1.6520$ a kvantilu $t_{n-1}(1-\alpha/2) = t_{215}(1-0.10/2) = t_{215}(0.95) = \mathsf{qt}(0.95, 215) = 1.6520$. Nyní již známe všechny potřebné hodnoty a můžeme dosadit do vzorce 11.

$$\begin{aligned}
(d,h) &= \left(m - \frac{s}{\sqrt{n}} t_{n-1}(1-\alpha/2)\,,\, m - \frac{s}{\sqrt{n}} t_{n-1}(\alpha/2)\right) \\
&= \left(137.1852 - \frac{4.8246}{\sqrt{215}} 1.6520\,,\, 137.1852 - \frac{4.8246}{\sqrt{215}}(-1.6520)\right) \\
&= (137.1852 - 0.5436\,,\, 137.1852 - (-0.5436)) \\
&= (136.6416\,,\, 137.7288)
\end{aligned}$$

V případě výpočtu 95% Waldova empirického intervalu spolehlivosti, vyjádříme koeficient α z rovnice $100 \times (1-\alpha)\,\% = 95\,\%$.

$$\begin{aligned}
100 \times (1-\alpha)\,\% &= 95\,\% \\
100 \times (1-\alpha) &= 95 \\
1-\alpha &= 0.95 \\
1-0.95 &= \alpha \\
\alpha &= 0.05
\end{aligned}$$

Pomocí softwaru R a funkce $\mathsf{qt()}$ stanovíme hodnotu kvantilu $t_{n-1}(\alpha/2) = t_{215}(0.05/2) = t_{215}(0.025) = \mathsf{qt}(0.025, 215) = -1.9711$ a kvantilu $t_{n-1}(1-\alpha/2) = t_{215}(1-0.05/2) = t_{215}(0.975) = \mathsf{qt}(0.975, 215) = 1.9711$. Hodnoty m, s a n jsme stanovili výše, zbývá tedy dosadit do vzorce 11.

$$\begin{aligned}
(d,h) &= \left(m - \frac{s}{\sqrt{n}} t_{n-1}(1-\alpha/2)\,,\, m - \frac{s}{\sqrt{n}} t_{n-1}(\alpha/2)\right) \\
&= \left(137.1852 - \frac{4.8246}{\sqrt{215}} 1.9711\,,\, 137.1852 - \frac{4.8246}{\sqrt{215}}(-1.9711)\right) \\
&= (137.1852 - 0.6486\,,\, 137.1852 - (-0.6486)) \\
&= (136.5366\,,\, 137.8338)
\end{aligned}$$

Konečně, v případě výpočtu 99% Waldova empirického intervalu spolehlivosti, vyjádříme koeficient α z rovnice $100 \times (1-\alpha)\,\% = 99\,\%$.

$$
\begin{aligned}
100 \times (1-\alpha)\,\% &= 99\,\% \\
100 \times (1-\alpha) &= 99 \\
1-\alpha &= 0.99 \\
1-0.99 &= \alpha \\
\alpha &= 0.01
\end{aligned}
$$

Pomocí softwaru R stanovíme hodnotu kvantilu $t_{n-1}(\alpha/2) = t_{215}(0.01/2) = t_{215}(0.005) = \mathsf{qt}(0.005, 215) = -2.5989$ a kvantilu $t_{n-1}(1-\alpha/2) = t_{215}(1-0.01/2) = t_{215}(0.995) = \mathsf{qt}(0.995, 215) = 2.5989$. Hranice intervalu spolehlivosti dopočítáme analogicky jako v předchozích dvou případech dosazením do vzorce 11.

$$
\begin{aligned}
(d,h) &= \left(m - \frac{s}{\sqrt{n}} t_{n-1}(1-\alpha/2)\,,\; m - \frac{s}{\sqrt{n}} t_{n-1}(\alpha/2)\right) \\
&= \left(137.1852 - \frac{4.8246}{\sqrt{215}} 2.5989\,,\; 137.1852 - \frac{4.8246}{\sqrt{215}}(-2.5989)\right) \\
&= (137.1852 - 0.8551\,,\; 137.1852 - (-0.8551)) \\
&= (136.3301\,,\; 138.0403)
\end{aligned}
$$

Datový soubor načteme příkazem `read.delim()` a NA hodnoty odstraníme příkazem `na.omit()`. Pomocí operátoru `[]` vybereme z tabulky `data` pouze ty řádky, které se vztahují k mužským skeletům (`data$sex == 'm'`) a sloupec obsahující údaje o největší šířce mozkovny `'skull.B'`. Hodnotu výběrového průměru a výběrové směrodatné odchylky dopočítáme pomocí funkcí `mean()` a `sd()`, rozsah náhodného výběru stanovíme funkcí `length()`. Do proměnné `alpha` si vložíme všechny tři hodnoty koeficientů α, tj. 0.1, 0.05 i 0.01, najednou. Nyní přepisem vzorce 11, kde funkci `qt()` využijeme na výpočet $\alpha/2$, resp. $1-\alpha/2$ kvantilů Studentova rozdělení, a to současně pro všechny tři koeficienty α, získáme dolní, resp. horní hranice všech tří Waldových empirických intervalů spolehlivosti.

```
data <- read.delim('01-one-sample-mean-skull-mf.txt')
data <- na.omit(data)

skull.BM <- data[data$sex == 'm', 'skull.B']
m        <- mean(skull.BM)
s        <- sd(skull.BM)
n        <- length(skull.BM)
alpha    <- c(0.1, 0.05, 0.01)

dh       <- m - s / sqrt(n) * qt(1 - alpha / 2, n - 1)
hh       <- m - s / sqrt(n) * qt(alpha / 2, n - 1)

tab <- data.frame(d = dh, h = hh, row.names = c('90% DIS', '95% DIS', '99% DIS'))
round(tab, 4)
```

```
               d        h
90% DIS 136.6429 137.7275
95% DIS 136.5381 137.8322
99% DIS 136.3320 138.0383
```

Interpretace výsledků: 90% Waldův empirický interval spolehlivosti pro parametr μ má tvar $(136.64\,,\, 137.73)$ mm. To znamená, že $136.64\,\text{mm} < \mu < 137.73\,\text{mm}$ s pravděpodobností 90 %. V 90 případech ze sta bude střední hodnota největší šířky mozkovny u skeletů mužského pohlaví nabývat hodnoty z intervalu $(136.64\,,\, 137.73)$ mm.

95% Waldův empirický interval spolehlivosti pro parametr μ má tvar $(136.54\,,\, 137.83)$ mm. To znamená, že $136.54\,\text{mm} < \mu < 137.83\,\text{mm}$ s pravděpodobností 95 %. V 95 případech ze sta bude střední hodnota největší šířky mozkovny u skeletů mužského pohlaví nabývat hodnoty z intervalu $(136.54\,,\, 137.83)$ mm.

99% Waldův empirický interval spolehlivosti pro parametr μ má tvar $(136.33\,,\,138.04)$ mm. To znamená, že $136.33\,\text{mm} < \mu < 138.04\,\text{mm}$ s pravděpodobností 99 %. V 99 případech ze sta bude střední hodnota největší šířky mozkovny u skeletů mužského pohlaví nabývat hodnoty z intervalu $(136.33\,,\,138.04)$ mm. ★

Příklad 5.13. Intervalový odhad parametru σ^2 normálního modelu
Načtěte datový soubor 01-one-sample-mean-skull-mf.txt a odstraňte z načtených dat NA hodnoty. Mějme náhodnou veličinu X popisující *největší délku mozkovny* u skeletů ženského pohlaví. Za předpokladu, že náhodná veličina X pochází z normálního rozdělení se střední hodnotou μ a rozptylem σ^2, tj. $X \sim N(\mu, \sigma^2)$, vypočítejte (a) 99%, oboustranný intervalový odhad; (b) 99% levostranný intervalový odhad; (c) 99% pravostranný intervalový odhad rozptylu σ^2.

Řešení příkladu 5.13
Celkem máme k dispozici $n = 109$ náhodných veličin $X_1, \ldots, X_{109}$, přičemž veličina X_1 popisuje největší délku mozkovny u prvního ženského skeletu, ..., veličina X_{109} popisuje výšku největší délku mozkovny u sto devátého ženského skeletu. Předpokládáme, že všechny náhodné veličiny pochází z normálního rozdělení, tj. $X_1 \sim N(\mu, \sigma^2)$, ..., $X_{109} \sim N(\mu, \sigma^2)$, a tedy i celý náhodný výběr pochází ze stejného normálního rozdělení, tj. $X_1, \ldots, X_{109} \sim N(\mu, \sigma^2)$. Naměřením hodnoty největší délky mozkovny každého skeletu jsme získali celkem 109 realizací náhodných veličin $X_1, \ldots, X_{109}$. Těchto 109 realizací tvoří dohromady datový soubor. Skutečné hodnoty parametrů μ a σ^2 nebudeme nikdy znát, ale můžeme je odhadnout pomocí bodových nebo intervalových odhadů.

V rámci tohoto příkladu se máme zaměřit na parametr rozptylu σ^2 a odhadnout jej pomocí všech tří typů intervalových odhadů. Všimněme si, že ze zadání neznáme skutečnou hodnotu parametru μ. Budeme tedy sestrojovat 99% Waldovy empirické intervaly spolehlivosti pro rozptyl σ^2 když parametr μ neznáme, a tedy budeme vycházet ze vzorců 14, 15 a 16.

Hodnotu parametru σ^2 odhadneme pomocí výběrohého rozptylu. K jeho výpočtu potřebujeme dopočítat výběrový průměr, tj.

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{109}(168 + 174 + \cdots + 162 + 170) = \frac{19\,024}{109} = 174.5321.$$

Dosazením výběrového průměru $m = 174.53$ do vzorce 2 získáme hodnotu výběrového rozptylu, tj.

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1^n}(x_i - m)^2 \\ &= \frac{1}{108}\left((168 - 174.5321)^2 + (174 - 174.5321)^2 + \cdots + (162 - 174.5321)^2 + (170 - 174.5321)^2\right) \\ &= \frac{1}{108}\left((-6.5321)^2 + (-0.5321)^2 + \cdots + (-12.5321)^2 + (-4.5321)^2\right) \\ &= 38.6772 \end{aligned}$$

Výběrový rozptyl $s^2 = 38.6772$, rozsah náhodného výběru $n = 109$. Zbývá stanovit hodnotu kvantilu $\chi^2_{n-1}(\alpha/2)$ a kvantilu $\chi^2_{n-1}(1-\alpha/2)$ χ^2 modelu. K tomu je potřeba nejprve dopočítat koeficient α, a to vyjádřením z rovnice $100 \times (1-\alpha)\% = 99\%$.

$$\begin{aligned} 100 \times (1-\alpha)\,\% &= 99\,\% \\ 100 \times (1-\alpha) &= 99 \\ 1-\alpha &= 0.99 \\ 1 - 0.99 &= \alpha \\ \alpha &= 0.01 \end{aligned}$$

Pomocí softwaru R a funkce qchisq() nyní stanovíme hodnotu kvantilu $\chi^2_{n-1}(\alpha/2) = \chi^2_{108}(0.01/2) = \chi^2_{108}(0.005) =$ $\mathsf{qchisq}(0.005, 108) = 73.8989$ a kvantilu $\chi^2_{n-1}(1-\alpha/2) = \chi^2_{108}(1-0.01/2) = \chi^2_{108}(0.995) = \mathsf{qchisq}(0.995, 108) =$ 149.5994.

Hranice 99% Waldova empirického oboustranného intervalu spolehlivosti vypočítáme dosazením do vzorce 14.

$$\begin{aligned}(d, h) &= \left(\frac{(n-1)s^2}{\chi^2_{n-1}(1-\alpha/2)}\,,\ \frac{(n-1)s^2}{\chi^2_{n-1}(\alpha/2)}\right)\\ &= \left(\frac{108 \times 38.6772}{149.5994}\,,\ \frac{108 \times 38.6772}{73.8989}\right)\\ &= (27.9222\,,\ 56.5250)\end{aligned}$$

Pro výpočet levostranného intervalu spolehlivosti stanovíme nejprve hodnotu kvantilu $\chi^2_{n-1}(1-\alpha)$. Z předchozího odstavce již víme, že $\alpha = 0.01$ a tedy $\chi^2_{n-1}(1-\alpha) = \chi^2_{108}(1-0.01) = \chi^2_{108}(0.99) = \mathsf{qchisq}(0.99, 108) = 145.0988$. Hranice 99% Waldova empirického levostranného intervalu spolehlivosti vypočítáme dosazením hodnot do vzorce 15.

$$\begin{aligned}(d, \infty) &= \left(\frac{(n-1)s^2}{\chi^2_{n-1}(1-\alpha)}\,,\ \infty\right)\\ &= \left(\frac{108 \times 38.6772}{145.0988}\,,\ \infty\right)\\ &= (28.7882\,,\ \infty)\end{aligned}$$

Pro výpočet pravostranného intervalu spolehlivosti stanovíme nejprve hodnotu kvantilu $\chi^2_{n-1}(\alpha)$. Protože koeficient α je opět rovný hodnotě 0.01, bude kvantil $\chi^2_{n-1}(\alpha) = \chi^2_{108}(0.01) = \chi^2_{108}(0.01) = \mathsf{qchisq}(0.01, 108) = 76.7736$. Hranice 99% Waldova empirického pravostranného intervalu spolehlivosti vypočítáme dosazením hodnot do vzorce 16.

$$\begin{aligned}(0, h) &= \left(0\,,\ \frac{(n-1)s^2}{\chi^2_{n-1}(\alpha)}\right)\\ &= \left(0\,,\ \frac{108 \times 38.6772}{76.7736}\right)\\ &= (0\,,\ 54.4085)\end{aligned}$$

Datový soubor načteme příkazem read.delim() a NA hodnoty odstraníme příkazem na.omit(). Pomocí operátoru [] vybereme z tabulky data pouze ty řádky, které se vztahují k ženským skeletům (data$sex == 'f') a sloupec obsahující údaje o největší délce mozkovny 'skull.L'. Hodnotu výběrové směrodatné odchylky dopočítáme funkcí sd(), rozsah náhodného výběru stanovíme pomocí funkce length(). Do proměnné alpha vložíme hodnotu koeficientu $\alpha = 0.01$. Nyní přepisem vzorců 14, 15 a 16, kde funkci qchisq() využijeme na výpočet $\alpha/2$, $1-\alpha/2$, α a $1-\alpha$ kvantilů χ^2 modelu získáme dolní a horní hranice všech tří Waldových empirických intervalů spolehlivosti.

```
data <- read.delim('01-one-sample-mean-skull-mf.txt')
data <- na.omit(data)

skull.LF <- data[data$sex =='f', 'skull.L']
n        <- length(skull.LF)
s.LF     <- sd(skull.LF)
alpha    <- 0.01

dh <- (n - 1) * s.LF ^ 2 / qchisq(1 - alpha / 2, n - 1)
hh <- (n - 1) * s.LF ^ 2 / qchisq(alpha / 2, n - 1)

DH <- (n - 1) * s.LF ^ 2 / qchisq(1 - alpha, n - 1)
HH <- (n - 1) * s.LF ^ 2 / qchisq(alpha, n - 1)

(tab <- data.frame(d = round(c(dh, DH, 0), 4),
                   h = c(round(hh, 4), 'inf', round(HH, 4)),
                  row.names = c('99% DIS', '99% LIS', '99% PIS')))
```

```
                d        h
99% DIS  27.9222   56.525
99% LIS  28.7882      inf
99% PIS   0.0000  54.4085
```

Poznámka: Horní hranice pravostranného Waldova empirického intervalu spolehlivosti je nekonečno. Abychom mohli tuto hodnotu zanést do tabulky výsledků, musíme ji zadat jako textový řetězec. Datové tabulky v softwaru R však mají tu vlastnost, že hodnoty ve stejném sloupci musí být stejného typu, tedy buď jsou všechny hodnoty numerické, nebo jsou všechny hodnoty textové. Proto v okamžiku, kdy do sloupce `h` tabulky `tab` vložíme textový řetězec `inf`, software R všechny numerické hodnoty v tomto sloupci automaticky převede na textové řetězce a tak se k nim od tohoto okamžiku také chová. Protože text nejde zaokrouhlit, není potom možné tabulku `tab` zaokrouhlit na čtyři desetinná místa. Proto musíme všechny hodnoty v tabulce zaokrouhlit dříve, než je do tabulky vložíme společně s textovým řetězcem `inf`.

Interpretace výsledků: 99% Waldův empirický oboustranný interval spolehlivosti pro parametr σ^2 má tvar $(27.92\,,\,56.53)\,\mathrm{mm}^2$. To znamená, že $27.92\,\mathrm{mm}^2 < \sigma^2 < 56.53\,\mathrm{mm}^2$ s pravděpodobností 99 %. V 99 případech ze sta bude rozptyl největší délky mozkovny u skeletů ženského pohlaví větší než $27.92\,\mathrm{mm}^2$ a menší než $56.53\,\mathrm{mm}^2$.

99% Waldův empirický levostranný interval spolehlivosti pro parametr σ^2 má tvar $(28.79\,,\,\infty)\,\mathrm{mm}^2$. To znamená, že $\sigma^2 > 28.79\,\mathrm{mm}^2$ s pravděpodobností 99 %. V 99 případech ze sta bude rozptyl největší délky mozkovny u skeletů ženského pohlaví větší než $28.79\,\mathrm{mm}^2$.

99% Waldův empirický pravostranný interval spolehlivosti pro parametr σ^2 má tvar $(0\,,\,54.41)\,\mathrm{mm}^2$. To znamená, že $\sigma^2 < 54.41\,\mathrm{mm}^2$ s pravděpodobností 99 %. V 99 případech ze sta bude rozptyl největší délky mozkovny u skeletů ženského pohlaví menší než $54.41\,\mathrm{mm}^2$. ★

Příklad 5.14. Intervalový odhad parametru σ normálního modelu

Načtěte datový soubor `01-one-sample-mean-skull-mf.txt` a odstraňte z načtených dat NA hodnoty. Mějme náhodnou veličinu X popisující *největší šířku mozkovny* u skeletů ženského pohlaví. Za předpokladu, že náhodná veličina X pochází z normálního rozdělení se střední hodnotou μ a rozptylem σ^2, tj. $X \sim N(\mu, \sigma^2)$, vypočítejte (a) 95%, oboustranný intervalový odhad směrodatné odchylky σ.

Řešení příkladu 5.14

V rámci tohoto příkladu se máme zaměřit na parametr směrodatné ochylky σ. Žádný interval spolehlivosti pro směrodatnou odchylku ale neznáme. Vystačíme si tedy s tím, co máme k dispozici. Vypočítáme hranice 95% oboustranného Waldova empirického intervalu spolehlivosti pro parametr rozptylu σ^2. Jejich odmocněním získáme hranice 95% oboustranného Waldova empirického intervalu spolehlivosti pro parametr σ.

Budeme tedy vycházet ze vzorce 14. Hodnotu parametru σ^2 odhadneme pomocí výběrohého rozptylu. Nejprve tedy spočítáme výběrový průměr a následně výběrový rozptyl, tj.

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{109}(130 + 134 + \cdots + 138 + 140) = \frac{14\,622}{109} = 134.1468.$$

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1^n}(x_i - m)^2 \\ &= \frac{1}{108}\left((130 - 134.1468)^2 + (134 - 134.1468)^2 + \cdots + (138 - 134.1468)^2 + (140 - 134.1468)^2\right) \\ &= \frac{1}{108}\left((-4.1468)^2 + (-0.1468)^2 + \cdots + (3.8532)^2 + (5.8532)^2\right) \\ &\doteq 22.0523 \end{aligned}$$

Výběrový rozptyl $s^2 = 22.0523$, rozsah náhodného výběru $n = 109$. Zbývá stanovit hodnotu kvantilu $\chi^2_{n-1}(\alpha/2)$ a kvantilu $\chi^2_{n-1}(1-\alpha/2)$ χ^2 modelu. K tomu je potřeba nejprve dopočítat koeficient α, a to vyjádřením z rovnice $100 \times (1-\alpha)\% = 95\%$.

$$
\begin{aligned}
100 \times (1-\alpha)\,\% &= 95\,\% \\
100 \times (1-\alpha) &= 95 \\
1-\alpha &= 0.95 \\
1-0.95 &= \alpha \\
\alpha &= 0.05
\end{aligned}
$$

Pomocí softwaru a funkce qchisq() nyní stanovíme hodnotu kvantilu $\chi^2_{n-1}(\alpha/2) = \chi^2_{108}(0.05/2) = \chi^2_{108}(0.025) =$ qchisq(0.025, 108) $= 81.1329$ a kvantilu $\chi^2_{n-1}(1-\alpha/2) = \chi^2_{108}(1-0.05/2) = \chi^2_{108}(0.975) =$ qchisq(0.975, 108) $=$ 138.6506.

Hranice 95% Waldova empirického oboustranného intervalu spolehlivosti pro parametr σ^2 vypočítáme dosazením do vzorce 14.

$$
\begin{aligned}
(d,h)_{\sigma^2} &= \left(\frac{(n-1)s^2}{\chi^2_{n-1}(1-\alpha/2)}\,,\,\frac{(n-1)s^2}{\chi^2_{n-1}(\alpha/2)}\right) \\
&= \left(\frac{108 \times 22.0523}{138.6506}\,,\,\frac{108 \times 22.0523}{81.1329}\right) \\
&= (17.1773\,,\,29.3549)
\end{aligned}
$$

Konečně, hranice 95% Waldova empirického oboustranného intervalu spolehlivosti pro parametr σ získáme odmocněním hranic intervalu spolehlivosti (17.1773 , 29.3549), tj.

$$
\begin{aligned}
(d,h)_{\sigma} &= \sqrt{(d,h)_{\sigma^2}} \\
&= \sqrt{(17.1773\,,\,29.3549)} \\
&= (4.1446\,,\,5.4180)
\end{aligned}
$$

Datový soubor načteme příkazem read.delim() a NA hodnoty odstraníme příkazem na.omit(). Pomocí operátoru [] vybereme z tabulky data pouze ty řádky, které se vztahují k ženským skeletům (data$sex == 'f') a sloupec obsahující údaje o největší šířce mozkovny 'skull.L'. Hodnotu výběrové směrodatné odchylky dopočítáme funkcí sd(), rozsah náhodného výběru stanovíme pomocí funkce length(). Do proměnné alpha vložíme hodnotu koeficientu $\alpha = 0.05$. Nyní přepisem vzorce 14, kde pomocí funkce qchisq() vypočítáme $\alpha/2$, $1-\alpha/2$ kvantily χ^2 modelu získáme dolní a horní hranici 95% Waldova empirického oboustranného intervalu spolehlivosti pro parametr σ^2. Odmocněním obou hranic pomocí funkce sqrt() získáme hranice 95% Waldova empirického oboustranného intervalu spolehlivosti pro parametr σ.

```
data <- read.delim('01-one-sample-mean-skull-mf.txt')
data <- na.omit(data)

skull.BF <- data[data$sex =='f', 'skull.B']
n        <- length(skull.BF)
s.BF     <- sd(skull.BF)
alpha    <- 0.05

dh <- (n - 1) * s.BF ^ 2 / qchisq(1 - alpha / 2, n - 1)
hh <- (n - 1) * s.BF ^ 2 / qchisq(alpha / 2, n - 1)

dh.s <- sqrt(dh)
hh.s <- sqrt(hh)

tab <- data.frame(d = c(dh, dh.s), h = c(hh, hh.s),
                  row.names = c('95% DIS pro rozptyl', '95% DIS pro sm.odchylku'))
round(tab, 4)
```

```
                           d       h
95% DIS pro rozptyl     17.1774 29.3549
95% DIS pro sm.odchylku  4.1446  5.4180
```

Interpretace výsledků: 95% Waldův empirický oboustranný interval spolehlivosti pro parametr σ má tvar $(4.14\,,\,5.42)$ mm. To znamená, že $4.14\,\text{mm} < \sigma < 5.42\,\text{mm}$ s pravděpodobností 95 %. V 95 případech ze sta bude směrodatná odchylka největší šířky mozkovny u skeletů ženského pohlaví větší než 4.14 mm a menší než 5.42 mm. ★

Příklad 5.15. Intervalový odhad parametru σ^2 normálního modelu

Načtěte datový soubor 01-one-sample-mean-skull-mf.txt a odstraňte z načtených dat NA hodnoty. Mějme náhodnou veličinu X popisující *největší délku mozkovny* u skeletů ženského pohlaví. Za předpokladu, že náhodná veličina X pochází z normálního rozdělení se střední hodnotou μ a rozptylem σ^2, tj. $X \sim N(\mu, \sigma^2)$, vypočítejte (a) 99%, oboustranný intervalový odhad; (b) 99% levostranný intervalový odhad; (c) 99% pravostranný intervalový odhad rozptylu σ.

Řešení příkladu 5.15

Hranice všech tří intervalů spolehlivosti získáme odmocněním hranic 99% Waldových empirických intervalů spolehlivosti pro rozptyl σ^2 vypočítaných v rámci příkladu 5.13. Konkrétně hranice 99% Waldova empirického oboustranného intervalu spolehlivosti pro parametr σ vypočítáme jako

$$(d, h) = \sqrt{(27.9222\,,\,56.5250)} = (5.2842\,,\,7.5183)$$

Hranice 99% Waldova empirického levostranného intervalu spolehlivosti pro parametr σ vypočítáme jako

$$\begin{aligned}(d, \infty) &= \sqrt{(28.7882\,,\,\infty)}\\ &= (5.3655\,,\,\infty)\end{aligned}$$

Hranice 99% Waldova empirického pravostranného intervalu spolehlivosti pro parametr σ vypočítáme jako

$$\begin{aligned}(0, h) &= \sqrt{(0\,,\,54.4085)}\\ &= (0\,,\,7.3762)\end{aligned}$$

Odmocniny dolních resp. horních hranici 99% Waldových empirických intervalů spolehlivosti získáme pomoc funkce sqrt().

```
dh <- sqrt(dh)
hh <- sqrt(hh)

DH <- sqrt(DH)
HH <- sqrt(HH)

(tab <- data.frame(d = round(c(dh, DH, 0), 4),
                   h = c(round(hh, 4), 'inf', round(HH, 4)),
                  row.names = c('99% DIS', '99% LIS', '99% PIS')))
```

```
             d      h
99% DIS 5.2841 7.5183
99% LIS 5.3655    inf
99% PIS 0.0000 7.3762
```

Interpretace výsledků: 99% Waldův empirický oboustranný interval spolehlivosti pro parametr σ má tvar $(5.28\,,\,7.52)$ mm. To znamená, že $5.28\,\text{mm} < \sigma < 7.52\,\text{mm}$ s pravděpodobností 99 %. V 99 případech ze sta bude rozptyl největší délky mozkovny u skeletů ženského pohlaví větší než 5.28 mm a menší než 7.52 mm.

99% Waldův empirický levostranný interval spolehlivosti pro parametr σ má tvar $(5.37\,,\,\infty)$ mm. To znamená, že $\sigma > 5.37$ mm s pravděpodobností 99 %. V 99 případech ze sta bude rozptyl největší délky mozkovny u skeletů ženského pohlaví větší než 5.37 mm.

99% Waldův empirický pravostranný interval spolehlivosti pro parametr σ má tvar $(0\,,\,7.38)$ mm. To znamená, že $\sigma < 7.38$ mm s pravděpodobností 99 %. V 99 případech ze sta bude rozptyl největší délky mozkovny u skeletů ženského pohlaví menší než 7.38 mm. ★

Příklad 5.16. Intervalový odhad parametru p alternativního modelu
Načtěte datový soubor `17-anova-newborns.txt` a odstraňte z načtených dat NA hodnoty. Mějme náhodnou veličinu X popisující *ženské pohlaví* novorozenců. Za předpokladu, že náhodná veličina X pochází z alternativního rozdělení s parametrem p, tj. $X \sim Alt(p)$, kde p je pravděpodobnost narození holčičky, stanovte (a) 95% oboustranný intervalový odhad parametru p; (b) 90% levostranný intervalový odhad parametru p; (c) 99% pravostranný intervalový odhad parametru p.

Řešení příkladu 5.16

95% oboustranný intervalový odhad získáme pomocí 95% oboustranného Waldova empirického intervalu spolehlivosti. Z příkladu 5.9 víme, že realizace výběrového průměru $m = 0.4797$ a rozsah náhodného výběru $N = 1382$. Zbývá nám tedy dopočítat hodnotu $\alpha/2$-kvantilu a $1-\alpha/2$-kvantilu standardizovaného normálního rozdělení. Z rovnice $100 \times (1-\alpha)\% = 95\%$ dopočítáme, že $\alpha = 0.05$. Pomocí softwaru zjistíme hodnoty kvantilů $u_{\alpha/2} = u_{0.05/2} = u_{0.025} = -1.9600$ a $u_{1-\alpha/2} = u_{1-0.05/2} = u_{0.975} = 1.9600$. Dosazením hodnot do vzorce 17 získáme realizaci 95% oboustranného Waldova empirického intervalu spolehlivosti pro parametr p, tj.

$$
\begin{aligned}
(d,h) &= \left(m - \sqrt{\frac{m(1-m)}{N}}u_{1-\alpha/2}\,,\, m - \sqrt{\frac{m(1-m)}{N}}u_{\alpha/2}\right)\\
&= \left(0.4797 - \sqrt{\frac{0.4797(1-0.4797)}{1382}}1.9600\,,\, 0.4797 - \sqrt{\frac{0.4797(1-0.4797)}{1382}}(-1.9600)\right)\\
&= (0.4797 - 0.0263\,,\, 0.4797 - (-0.0263))\\
&= (0.4534\,,\, 0.5060)
\end{aligned}
$$

90% levostranný intervalový odhad získáme pomocí 90% levostranného Waldova empirického intervalu spolehlivosti. Realizace výběrového průměru $m = 0.4797$ a rozsah náhodného výběru $N = 1382$. Zbývá nám tedy dopočítat hodnotu $1-\alpha$-kvantilu standardizovaného normálního rozdělení. Z rovnice $100 \times (1-\alpha)\% = 90\%$ dopočítáme, že $\alpha = 0.1$. Pomocí softwaru zjistíme hodnotu kvantilu $u_{1-\alpha} = u_{1-0.1} = u_{0.9} = 1.2816$. Dosazením hodnot do vzorce 18 získáme realizaci 90% levostranného Waldova empirického intervalu spolehlivosti pro parametr p, tj.

$$
\begin{aligned}
(d,1) &= \left(m - \sqrt{\frac{m(1-m)}{N}}u_{1-\alpha}\,,\, 1\right)\\
&= \left(0.4797 - \sqrt{\frac{0.4797(1-0.4797)}{1382}}1.2816\,,\, 1\right)\\
&= (0.4797 - 0.0172\,,\, 1)\\
&= (0.4625\,,\, 1)
\end{aligned}
$$

99% pravostranný intervalový odhad získáme pomocí 99% pravostranného Waldova empirického intervalu spolehlivosti. Realizace výběrového průměru $m = 0.4797$ a rozsah náhodného výběru $N = 1382$. Zbývá tedy dopočítat hodnotu α-kvantilu standardizovaného normálního rozdělení. Z rovnice $100 \times (1-\alpha)\% = 99\%$ dopočítáme, že $\alpha = 0.01$. Pomocí softwaru zjistíme hodnotu kvantilu $u_\alpha = u_{0.01} = -2.3263$. Dosazením hodnot do vzorce 19 získáme realizaci 99% pravostranného Waldova empirického intervalu spolehlivosti pro parametr p, tj.

$$
\begin{aligned}
(0, h) &= \left(0\,,\, m - \sqrt{\frac{m(1-m)}{N}}u_{\alpha}\right)\\
&= \left(0\,,\, 0.4797 - \sqrt{\frac{0.4797(1-0.4797)}{1382}}(-2.3263)\right)\\
&= (0\,,\, 0.4797 - (-0.03126))\\
&= (0\,,\, 0.5110)
\end{aligned}
$$

Nejprve načteme datový soubor příkazem read.delim() a odstraníme NA hodnoty příkazem na.omit(). Dále do proměnné sex vložíme údaje o pohlaví. Protože proměnná sex je typu factor s dvěma úrovněmi, úrovní 1 ('f'; (female)) a úrovní 2 ('m', (male)), převedeme jej nejprve pomocí funkce as.numeric() na číselný vektor, který vložíme do proměnné pohlavi. Následně změníme všechny hodnoty 2 na hodnoty 0, čímž dostaneme požadované kódování 0 = male, 1 = female (viz příklad 5.9.

```
data <- read.delim('17-anova-newborns.txt')
data <- na.omit(data)
sex  <- data$sex.C
pohlavi <- as.numeric(sex)
pohlavi[pohlavi == 2] <- 0
```

Hranice 95% oboustranného Waldova empirického intervalu spolehlivosti nyní získáme přepisem vzorce 17 s použitím funkcí mean(), sd(), length() a qnorm().

```
alpha <- 0.05
N     <- length(pohlavi)
m     <- mean(pohlavi)
s     <- sd(pohlavi)

dh <- m - sqrt(m * (1 - m) / N) * qnorm(1 - alpha / 2)
hh <- m - sqrt(m * (1 - m) / N) * qnorm(alpha / 2)
```

Dolní hranici 90% levostranného Waldova empirického intervalu spolehlivosti nyní získáme přepisem vzorce 18.

```
DH <- m - sqrt(m * (1 - m) / N) * qnorm(1 - alpha)
```

Horní hranici 99% pravostranného Waldova empirického intervalu spolehlivosti nyní získáme přepisem vzorce 19. Výsledné hranice všech tří intervalů spolehlivosti vložíme do jedné tabulky.

```
HH  <- m - sqrt(m * (1 - m) / N) * qnorm(alpha)
tab <- data.frame(d = c(dh, DH, 0), h = c(hh, 1, HH),
                  row.names = c('95% OIS', '90% LIS', '99% PIS'))
round(tab, 4)
```

```
             d      h
95% OIS 0.4534 0.5061
90% LIS 0.4576 1.0000
99% PIS 0.0000 0.5018
```

Interpretace výsledků: 95 % oboustranný Waldův empirický interval spolehlivosti pro parametr p má tvar $(0.4534\,,\, 0.5061)$, což znamená, že $0.4534 < p < 0.5061$ s pravděpodobností 95 %. S 95% pravděpodobností se pravděpodobnost narození holčičky pohybuje v rozmezí 45.34 % – 50.61 %.

90 % levostranný Waldův empirický interval spolehlivosti pro parametr p má tvar $(0.4576\,,\, 1)$, což znamená, že $0.4576 < p < 1$ s pravděpodobností 90 %. S pravděpodobností 90% je pravděpodobnost narození holčičky větší než 45.76 %.

99 % pravostranný Waldův empirický interval spolehlivosti pro parametr p má tvar $(0\,,\, 0.5018)$, což znamená, že $0 < p < 0.5018$ s pravděpodobností 99 %. S pravděpodobností 99% je pravděpodobnost narození holčičky menší než 50.61 %. ★