12 Dvouvýběrové neparametrické testy

12.1 Wilcoxonův dvouvýběrový exaktní test

Nechť $X_{11}, \dots, X_{1n_1}$, $n_1 \geq 2$ je náhodný výběr ze spojitého rozdělení s distribuční funkcí $F_1(x)$ a $X_{21}, \dots, X_{2n_2}$, $n_2 > 2$ je na něm nezávislý náhodný výběr ze spojitého rozdělení s distribuční funkcí $F_2(x)$. O distribučních funkcích $F_1(x)$ a $F_2(x)$ předpokládáme, že se navzájem liší pouze posunutím. Nechť dále $\tilde{x}_1$ je medián náhodného výběru $X_{11}, \dots, X_{1n_1}$ a $\tilde{x}_2$ je medián náhodného výběru $X_{21}, \dots, X_{2n_2}$ a $\tilde{x}_0$ je konstanta. Na hladině významnosti α testujeme jednu z následujících tří hypotéz oproti příslušné alternativní hypotéze.

$$
\begin{array}{lcll}
H_{01} : \tilde{x}_1 - \tilde{x}_2 = \tilde{x}_0 & \text{oproti} & H_{11} : \tilde{x}_1 - \tilde{x}_2 \neq \tilde{x}_0 & \text{(oboustranná alt.)} \\
H_{02} : \tilde{x}_1 - \tilde{x}_2 \leq \tilde{x}_0 & \text{oproti} & H_{12} : \tilde{x}_1 - \tilde{x}_2 > \tilde{x}_0 & \text{(pravostranná alt.)} \\
H_{03} : \tilde{x}_1 - \tilde{x}_2 \geq \tilde{x}_0 & \text{oproti} & H_{13} : \tilde{x}_1 - \tilde{x}_2 < \tilde{x}_0 & \text{(levostranná alt.)}
\end{array}
$$

Test nazýváme Wilcoxonův dvouvýběrový exaktní test o rozdílu mediánů $\tilde{x}_1 - \tilde{x}_2$. Testovací statistika má tvar

$$
S_E = \sum_{j=1}^{n_2} S_j \tag{12.1}
$$

kde n_2 je rozsah druhého náhodného výběru a S_j, $j = 1, \dots, n_2$, je pořadí náhodných veličin $X_{21}, \dots, X_{2n_2}$ druhého náhodného výběru v seřazeném vektoru všech $n_1 + n_2$ hodnot z obou náhodných výběrů. Statistika W je tedy součet pořadí hodnot $X_{21}, \dots, X_{2n_2}$, v seřazeném vektoru všech $n_1 + n_2$ hodnot. Kritický obor podle zvolené alternativní hypotézy má tvar

$$
\begin{array}{ll}
H_{11} : \tilde{x}_1 - \tilde{x}_2 \neq \tilde{x}_0 & W = (-\infty\,;\, w_{n_1,n_2}(\alpha/2)\rangle \cup \langle w_{n_1,n_2}(1-\alpha/2)\,;\, \infty) \\
H_{12} : \tilde{x}_1 - \tilde{x}_2 > \tilde{x}_0 & W = \langle w_{n_1,n_2}(1-\alpha)\,;\, \infty) \\
H_{13} : \tilde{x}_1 - \tilde{x}_2 < \tilde{x}_0 & W = (-\infty\,;\, w_{n_1,n_2}(\alpha)\rangle
\end{array}
$$

kde $w_{n_1,n_2}(\alpha/2)$, $w_{n_1,n_2}(1-\alpha/2)$, $w_{n_1,n_2}(\alpha)$, $w_{n_1,n_2}(1-\alpha)$ jsou tabelované kvantily pro dvouvýběrový Wilcoxonův test, jejichž hodnoty získáme příkazem `qwilcox()`.

Interval spolehlivosti má podle zvolené alternativní hypotézy jeden z následujících tvarů

$$
\begin{array}{ll}
H_{11} : \tilde{x}_1 - \tilde{x}_2 \neq \tilde{x}_0 & (d, h) = \left(U^{(w_{n_1,n_2}(\alpha/2))}\,;\, U^{(w_{n_1,n_2}(1-\alpha/2)+1)}\right) \\
H_{12} : \tilde{x}_1 - \tilde{x}_2 > \tilde{x}_0 & (d, \infty) = \left(U^{(w_{n_1,n_2}(\alpha))}\,;\, \infty\right) \\
H_{13} : \tilde{x}_1 - \tilde{x}_2 < \tilde{x}_0 & (-\infty, h) = \left(-\infty\,;\, U^{(w_{n_1,n_2}(1-\alpha)+1)}\right)
\end{array}
$$

kde $U^{(1)} \leq \dots \leq U^{(n_1 n_2)}$ značí vzestupně seřazené rozdíly $Y_j - X_i$, $i = 1, \dots, n_1$, $j = 1, \dots, n_2$, a $U^{(x)}$ značí x-tý seřazený rozdíl.

p-hodnota má v závislosti na zvolené alternativní hypotéze jeden z následujících tvarů

$$
\begin{array}{ll}
H_{11} : \tilde{x}_1 - \tilde{x}_2 \neq \tilde{x}_0 & p\text{-hodnota} = 2\min\{\Pr(S_E \leq s_E)\,,\, \Pr(S_E > s_E)\} \\
H_{12} : \tilde{x}_1 - \tilde{x}_2 > \tilde{x}_0 & p\text{-hodnota} = \Pr(S_E > s_E) = 1 - \Pr(S_E \leq s_E) \\
H_{13} : \tilde{x}_1 - \tilde{x}_2 < \tilde{x}_0 & p\text{-hodnota} = \Pr(S_E \leq s_E)
\end{array}
$$

kde S_E je náhodná veličina, s_E je realizace testovací statistiky S_E (viz vzorec 12.1), tedy konkrétní číslo, a $\Pr(S_E \leq s_E)$ je distribuční funkce tabelovaného rozdělení pro Wilcoxonův dvouvýběrový exaktní test, jejíž hodnotu získáme pomocí R a implementované funkce `pwilcox()`.

```
# ZKONTROLOVAT BARVY
```

Příklad 12.1. Wilcoxonův dvouvýběrový exaktní test

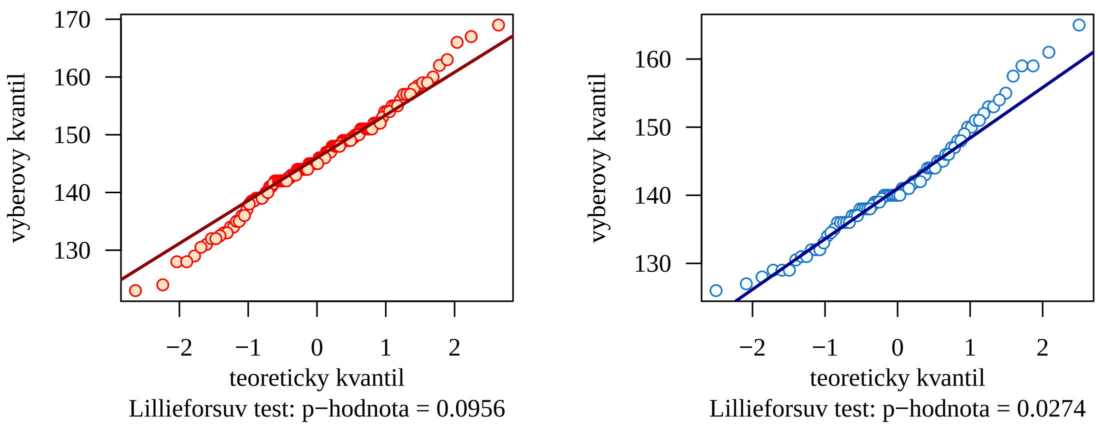
Řešení příkladu 12.1

```
data <- read.delim('00-Data//18-more-samples-variances-clavicle.txt')
head(data)
cla.LI1 <- data[data$pop == 'ind1', 'cla.L']
cla.LI1 <- na.omit(cla.LI1)
cla.LI2 <- data[data$pop == 'ind2', 'cla.L']
cla.LI2 <- na.omit(cla.LI2)
n1 <- length(cla.LI1)
n2 <- length(cla.LI2)
```

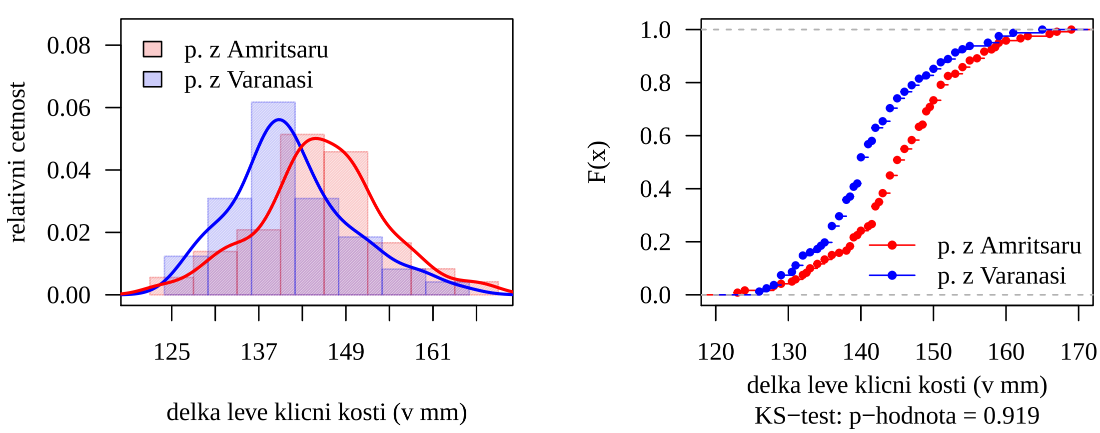
```
par(mar = c(5, 4, 1, 2), family = 'Times')
qqnorm(cla.LI1, main = '', xlab = '', pch = 21, col = 'red', bg = 'bisque',
       ylab = 'vyberovy kvantil', las = 1)
mtext('teoreticky kvantil', side = 1, line = 2)
mtext(bquote(paste('Lillieforsuv test: p-hodnota = ', .(round(nortest::lillie.test(cla.LI1)$p.val, 4)))),
      side = 1, line = 3.2)
qqline(cla.LI1, lwd = 2, col = 'darkred')

qqnorm(cla.LI2, main = '', xlab = '', pch = 21, col = 'dodgerblue3', bg = 'mintcream',
       ylab = 'vyberovy kvantil', las = 1)
mtext('teoreticky kvantil', side = 1, line = 2)
mtext(bquote(paste('Lillieforsuv test: p-hodnota = ', .(round(nortest::lillie.test(cla.LI2)$p.val, 4)))),
      side = 1, line = 3.2)
qqline(cla.LI2, lwd = 2, col = 'darkblue')
```

★



Obrázek 1: Kvantilové diagramy délky klíční kosti z levé strany u mužů indické populace z Amritsaru (vlevo) a indické populace z Varanasi (vpravo)



Obrázek 2: Porovnání histogramů a graf distribučních funkcí délky klíční kosti z levé strany u mužů indické populace z Amritsaru a indické populace z Varanasi

```
par(mar = c(5, 4, 1, 2), family = 'Times')
hist(cla.LI1, prob = T, breaks = seq(122, 170, by = 6), main = '', axes = F,
     col = rgb(0.9, 0, 0, 0.2), density = 80, xlim = c(120, 170),
     xlab = 'delka leve klicni kosti (v mm)', ylim = c(0, 0.085),
     ylab = 'relativni cetnost')
hist(cla.LI2, prob = T, breaks = seq(124, 166, by = 6), main = '', axes = F,
     col = rgb(0, 0, 0.9, 0.2), density = 80, add = T)
box(bty = 'o')
axis(1, seq(125, 167, by = 6))
axis(2, las = 1)
lines(density(cla.LI2), lwd = 2, col = 'blue')
lines(density(cla.LI1), lwd = 2, col = 'red')
legend('topleft', fill = c(rgb(0.9, 0, 0, 0.2), rgb(0, 0, 0.9, 0.2)),
       legend = c('p. z Amritsaru', 'p. z Varanasi'), bty = 'n')

plot(ecdf(cla.LI1), col = 'red', main = '', xlim = c(120, 170), ylab = 'F(x)',
     xlab = '', las = 1, cex = 0.6)
plot(ecdf(cla.LI2), add = T, col = 'blue', cex = 0.6)
ks.val <- round(ks.test(cla.LI2-mean(cla.LI2), cla.LI1-(mean(cla.LI1)))$p.val, 4)
mtext('delka leve klicni kosti (v mm)', side = 1, line = 2.1)
mtext(bquote(paste('KS-test: p-hodnota = ', .(ks.val))), side = 1, line = 3.3)
legend('bottomright', col = c('red', 'blue'), lty = 1, pch = 20,
       legend = c('p. z Amritsaru', 'p. z Varanasi'), bty = 'n')
```

Příklad 12.2. Wilcoxonův dvouvýběrový exaktní test

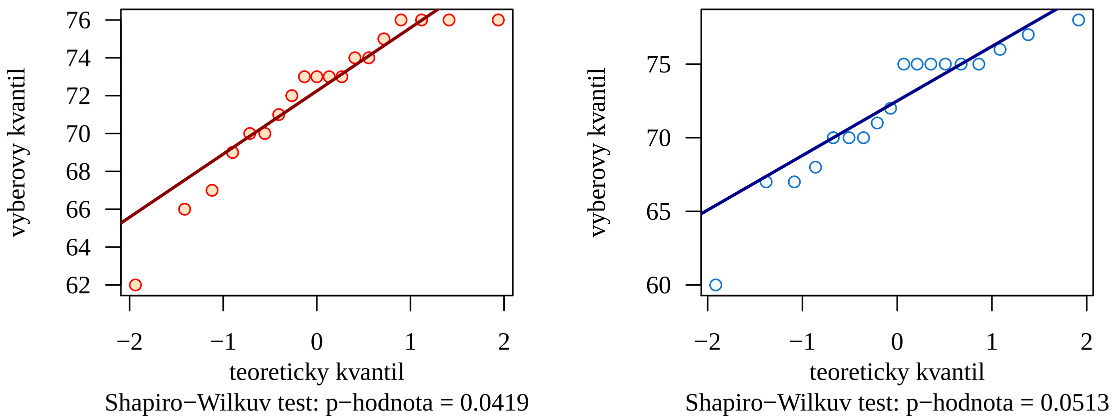
Řešení příkladu 12.2

```
data <- read.delim('00-Data//15-anova-means-skull.txt')
head(data)
upface.HN <- data[data$pop == 'nem', 'upface.H']
upface.HN <- na.omit(upface.HN)
upface.HC <- data[data$pop == 'cin', 'upface.H']
upface.HC <- na.omit(upface.HC)
n1 <- length(upface.HN)
n2 <- length(upface.HC)
```

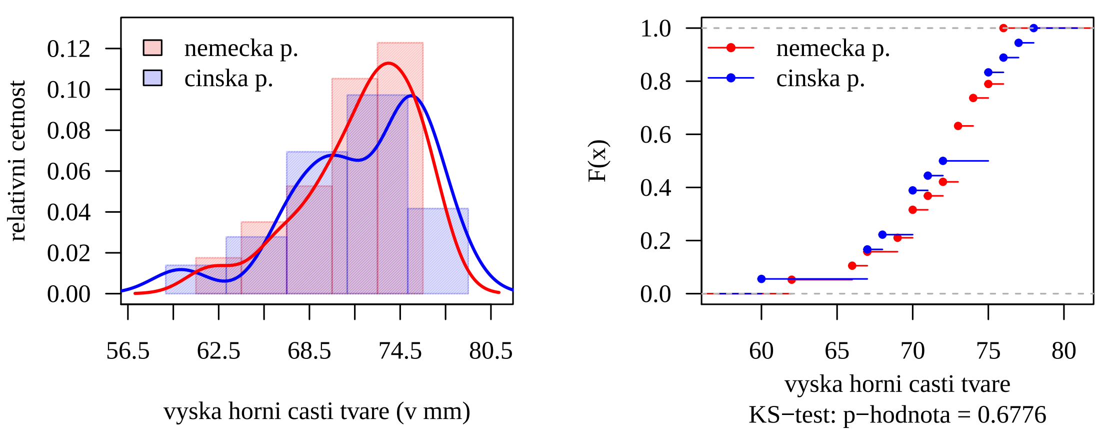
```
par(mar = c(5, 4, 1, 2), family = 'Times')
qqnorm(upface.HN, main = '', xlab = '', pch = 21, col = 'red', bg = 'bisque',
       ylab = 'vyberovy kvantil', las = 1)
mtext('teoreticky kvantil', side = 1, line = 2)
mtext(bquote(paste('Shapiro-Wilkuv test: p-hodnota = ', .(round(shapiro.test(upface.HN)$p.val, 4)))),
      side = 1, line = 3.2)
qqline(upface.HN, lwd = 2, col = 'darkred')

qqnorm(upface.HC, main = '', xlab = '', pch = 21, col = 'dodgerblue3', bg = 'mintcream',
       ylab = 'vyberovy kvantil', las = 1)
mtext('teoreticky kvantil', side = 1, line = 2)
mtext(bquote(paste('Shapiro-Wilkuv test: p-hodnota = ', .(round(shapiro.test(upface.HC)$p.val, 4)))),
      side = 1, line = 3.2)
qqline(upface.HC, lwd = 2, col = 'darkblue')
```

★



Obrázek 3: Kvantilové diagramy výšky horní části tváře mužů německé populace (vlevo) a čínské populace (vpravo)



Obrázek 4: Porovnání histogramů a graf distribučních funkcí výšky horní části tváře mužů německé a čínské populace

```
par(mar = c(5, 4, 1, 2), family = 'Times')
hist(upface.HN, prob = T, breaks = seq(61, 76, by = 3), main = '', axes = F,
     col = rgb(0.9, 0, 0, 0.2), density = 80, ylim = c(0, 0.13), xlim = c(57, 81),
     xlab = 'vyska horni casti tvare (v mm)',
     ylab = 'relativni cetnost')

hist(upface.HC, prob = T, breaks = seq(59, 79, by = 4), main = '', axes = F,
     col = rgb(0, 0, 0.9, 0.2), density = 80, add = T)
box(bty = 'o')
axis(1, seq(53.5, 80.5, by = 3))
axis(2, las = 1)
lines(density(upface.HC), lwd = 2, col = 'blue')
lines(density(upface.HN), lwd = 2, col = 'red')
legend('topleft', fill = c(rgb(0.9, 0, 0, 0.2), rgb(0, 0, 0.9, 0.2)),
       legend = c('nemecka p.', 'cinska p.'), bty = 'n')

plot(ecdf(upface.HN), col = 'red', main = '', xlim = c(57, 81), ylab = 'F(x)',
     xlab = '', las = 1, cex = 0.6)
plot(ecdf(upface.HC), add = T, col = 'blue', cex = 0.6)
ks.val <- round(ks.test(upface.HC - mean(upface.HC), upface.HN - (mean(upface.HN)))$p.
   val, 4)
mtext('vyska horni casti tvare', side = 1, line = 2.1)
mtext(bquote(paste('KS-test: p-hodnota = ', .(ks.val))), side = 1, line = 3.3)
legend('topleft', col = c('red', 'blue'), lty = 1, pch = 20,
       legend = c('nemecka p.', 'cinska p.'), bty = 'n')
```

Příklad 12.3. Wilcoxonův dvouvýběrový exaktní test

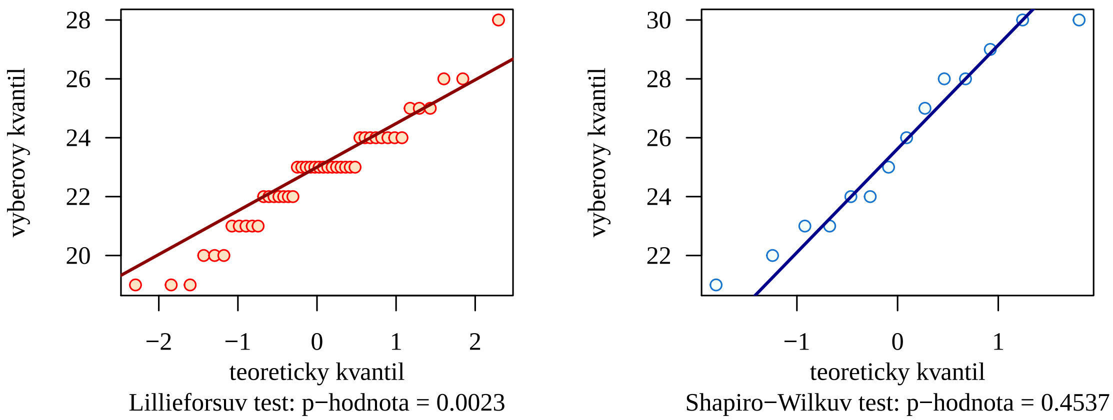
Řešení příkladu 12.3

```
data <- read.delim('00-Data//19-more-samples-correlations-skull.txt')
head(data)
intorb.BP <- data[data$pop == 'per', 'intorb.B']
intorb.BP <- na.omit(intorb.BP)
intorb.BB <- data[data$pop == 'ban', 'intorb.B']
intorb.BB <- na.omit(intorb.BB)
n1 <- length(intorb.BP) # 46
n2 <- length(intorb.BB) # 14
```

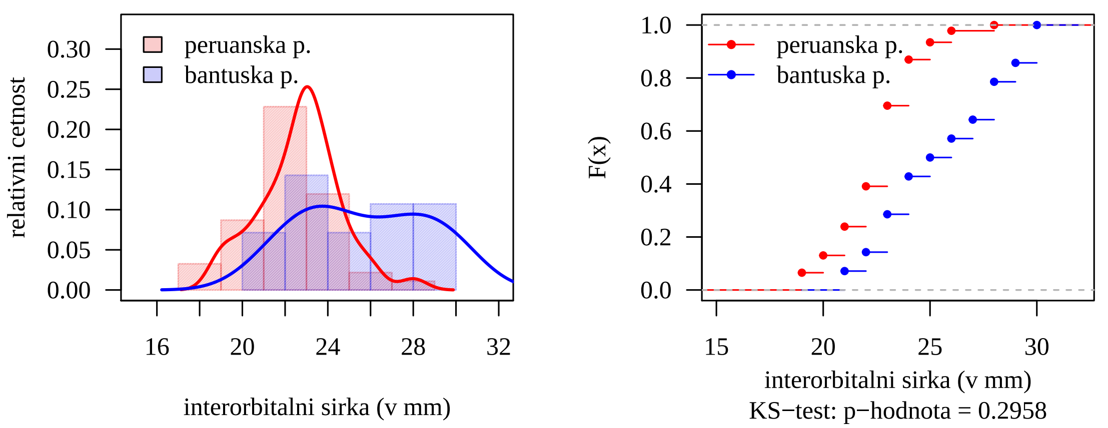
```
par(mar = c(5, 4, 1, 2), family = 'Times')
qqnorm(intorb.BP, main = '', xlab = '', pch = 21, col = 'red', bg = 'bisque',
       ylab = 'vyberovy kvantil', las = 1)
mtext('teoreticky kvantil', side = 1, line = 2)
mtext(bquote(paste('Lillieforsuv test: p-hodnota = ', .(round(nortest::lillie.test(
    intorb.BP)$p.val, 4)))),
      side = 1, line = 3.2)
qqline(intorb.BP, lwd = 2, col = 'darkred')

qqnorm(intorb.BB, main = '', xlab = '', pch = 21, col = 'dodgerblue3', bg = 'mintcream',
       ylab = 'vyberovy kvantil', las = 1)
mtext('teoreticky kvantil', side = 1, line = 2)
mtext(bquote(paste('Shapiro-Wilkuv test: p-hodnota = ', .(round(shapiro.test(intorb.BB)$
    p.val, 4)))),
      side = 1, line = 3.2)
qqline(intorb.BB, lwd = 2, col = 'darkblue')
```

★



Obrázek 5: Porovnání histogramů a graf distribučních funkcí interorbitální šířky mužů peruánské a bantuské populace



Obrázek 6: Porovnání histogramů a graf distribučních funkcí interorbitální šířky mužů peruánské a bantuské populace

```
par(mar = c(5, 4, 1, 2), family = 'Times')
hist(intorb.BP, prob = T, breaks = seq(17, 29, by = 2), main = '', axes = F,
     col = rgb(0.9, 0, 0, 0.2), density = 80, xlim = c(15, 32),  ylim = c(0, 0.33),
     xlab = 'interorbitalni sirka (v mm)',
     ylab = 'relativni cetnost')
hist(intorb.BB, prob = T, breaks = seq(20, 30, by = 2), main = '', axes = F,
     col = rgb(0, 0, 0.9, 0.2), density = 80, add = T)
box(bty = 'o')
axis(1, seq(12, 36, by = 2))
axis(2, las = 1)
lines(density(intorb.BP), lwd = 2, col = 'red')
lines(density(intorb.BB), lwd = 2, col = 'blue')
legend('topleft', fill = c(rgb(0.9, 0, 0, 0.2), rgb(0, 0, 0.9, 0.2)),
       legend = c('peruanska p.', 'bantuska p.'), bty = 'n')

plot(ecdf(intorb.BP), col = 'red', main = '', xlim = c(15, 32), ylab = 'F(x)',
     xlab = '', las = 1, cex = 0.6)
plot(ecdf(intorb.BB), add = T, col = 'blue', cex = 0.6)
ks.val <- round(ks.test(intorb.BP - mean(intorb.BP), intorb.BB - (mean(intorb.BB)))$p.
    val, 4)
mtext('interorbitalni sirka (v mm)', side = 1, line = 2.1)
mtext(bquote(paste('KS-test: p-hodnota = ', .(ks.val))), side = 1, line = 3.3)

legend('topleft', col = c('red', 'blue'), lty = 1, pch = 20,
       legend = c('peruanska p.', 'bantuska p.'), bty = 'n')
```

12.2 Wilcoxonův dvouvýběrový asymptotický test

Pro náhodné výběry o rozsazích $n_1 > 30$ a $n_2 > 30$ máme možnost použít k otestování nulové hypotézy asymptotickou variantu testu. Tuto variantu nazýváme Wilcoxonův dvouvýběrový asymptotický test o rozdílu mediánů $\tilde{x}_1 - \tilde{x}_2$. Testovací statistika asymptotického testu má tvar

$$S_A = \frac{S_E - \frac{n_2(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}} \tag{12.2}$$

kde S_E je testovací statistika definovaná vztahem 12.1, n_1 je rozsah prvního náhodného výběru, n_2 je rozsah druhého náhodného výběru. Za platnosti nulové hypotézy pochází statistika U_A ze standardizovaného normálního rozdělení, tj.

$$S_A = \frac{S_E - \frac{n_2(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}} \overset{H_0}{\sim} N(0,1).$$

Kritický obor podle zvolené alternativní hypotézy má tvar

$$\begin{array}{ll}
H_{11}: \tilde{x}_1 - \tilde{x}_2 \neq \tilde{x}_0 & W = \left(-\infty\,;\, u_{\alpha/2}\right\rangle \cup \left\langle u_{1-\alpha/2}\,;\, \infty\right) \\
H_{12}: \tilde{x}_1 - \tilde{x}_2 > \tilde{x}_0 & W = \langle u_{1-\alpha}\,;\, \infty) \\
H_{13}: \tilde{x}_1 - \tilde{x}_2 < \tilde{x}_0 & W = (-\infty\,;\, u_{\alpha}\rangle
\end{array}$$

kde $u_{\alpha/2}$, $u_{1-\alpha/2}$, u_α, $u_{1-\alpha}$ jsou kvantily standardizovaného normálního rozdělení, jejichž hodnoty získáme pomocí R a implementované funkce `qnorm()`.

Interval spolehlivosti má podle zvolené alternativní hypotézy jeden z následujících tvarů

$$\begin{array}{ll}
H_{11}: \tilde{x}_1 - \tilde{x}_2 \neq \tilde{x}_0 & (d,h) = \left(U^{(C_{\alpha/2})}\,;\, U^{(n_1 n_2+1-C_{\alpha/2})}\right) \\
H_{12}: \tilde{x}_1 - \tilde{x}_2 > \tilde{x}_0 & (d,\infty) = \left(U^{(C_{\alpha})}\,;\, \infty\right) \\
H_{13}: \tilde{x}_1 - \tilde{x}_2 < \tilde{x}_0 & (-\infty,h) = \left(-\infty\,;\, U^{n_1 n_2+1-C_{\alpha/2}}\right)
\end{array}$$

kde $C_{\alpha/2} = \frac{n_1 n_2}{2} - u_{\alpha/2}\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}$, $C_\alpha = \frac{n_1 n_2}{2} - u_\alpha\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}$, $U^{(1)} \leq \cdots \leq U^{(n_1 n_2)}$ značí vzestupně seřazené rozdíly $Y_j - X_i$, $i = 1,\ldots,n_1$, $j = 1,\ldots,n_2$, a $U^{(x)}$ značí x-tý seřazený rozdíl.

p-hodnota má v závislosti na zvolené alternativní hypotéze jeden z následujících tvarů

$$\begin{array}{ll}
H_{11}: \tilde{x}_1 - \tilde{x}_2 \neq \tilde{x}_0 & p\text{-hodnota} = 2\min\{\Pr(S_A \leq s_A)\,,\, \Pr(S_A > s_A)\} \\
H_{12}: \tilde{x}_1 - \tilde{x}_2 > \tilde{x}_0 & p\text{-hodnota} = \Pr(S_A > s_A) = 1 - \Pr(S_A \leq s_A) \\
H_{13}: \tilde{x}_1 - \tilde{x}_2 < \tilde{x}_0 & p\text{-hodnota} = \Pr(S_A \leq s_A)
\end{array}$$

kde S_A je náhodná veličina, s_A je realizace testovací statistiky S_A (viz vzorec 12.2), tedy konkrétní číslo, a $\Pr(S_A \leq s_A)$ je distribuční funkce standardizovaného normálního rozdělení, jejíž hodnotu získáme pomocí R a implementované funkce `pnorm()`.

Příklad 12.4. Wilcoxonův dvouvýběrový asymptotický test

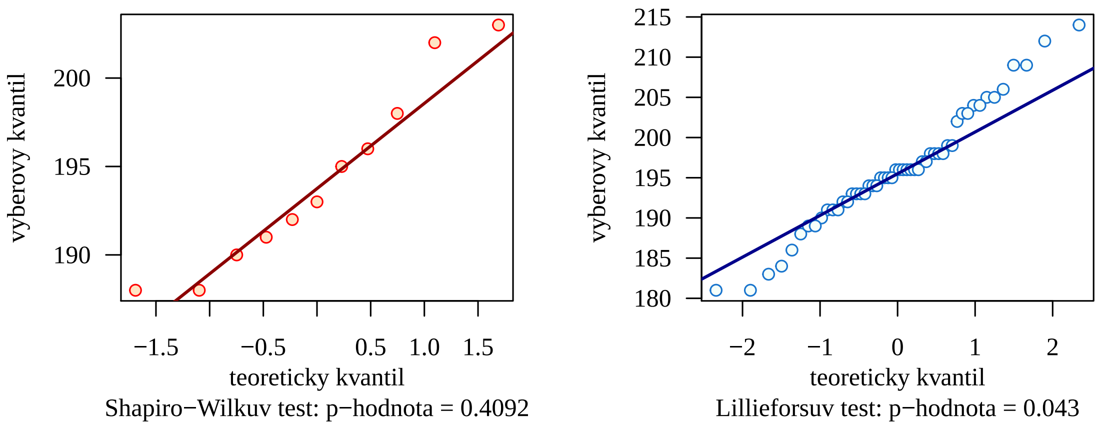
Řešení příkladu 12.4

```
data <- read.delim('00-Data//16-anova-head.txt')
head(data)
head.LMSa <- data[data$sex == 'm' & data$sexor == 'sa', 'head.L']
head.LMSa <- na.omit(head.LMSa)
head.LMOp <- data[data$sex == 'm' & data$sexor == 'op', 'head.L']
head.LMOp <- na.omit(head.LMOp)
n1 <- length(head.LMSa)
n2 <- length(head.LMOp)
```

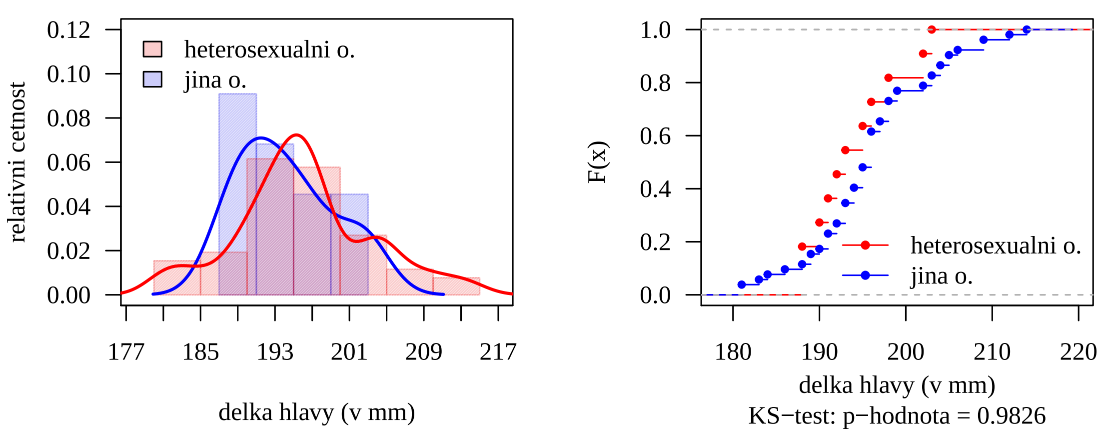
```
par(mar = c(5, 4, 1, 2), family = 'Times')
qqnorm(head.LMSa, main = '', xlab = '', pch = 21, col = 'red', bg = 'bisque',
       ylab = 'vyberovy kvantil', las = 1)
mtext('teoreticky kvantil', side = 1, line = 2)
mtext(bquote(paste('Shapiro-Wilkuv test: p-hodnota = ', .(round(shapiro.test(head.LMSa)$
    p.val, 4)))),
      side = 1, line = 3.2)
qqline(head.LMSa, lwd = 2, col = 'darkred')

qqnorm(head.LMOp, main = '', xlab = '', pch = 21, col = 'dodgerblue3', bg = 'mintcream',
       ylab = 'vyberovy kvantil', las = 1)
mtext('teoreticky kvantil', side = 1, line = 2)
mtext(bquote(paste('Lillieforsuv test: p-hodnota = ', .(round(nortest::lillie.test(head.
    LMOp)$p.val, 4)))),
      side = 1, line = 3.2)
qqline(head.LMOp, lwd = 2, col = 'darkblue')
```

★



Obrázek 7: Kvantilové diagramy délky hlavy mužů orientovaných výhradně heterosexuálně (vlevo) a mužů orientovaných jinak než výhradně heterosexuálně (vpravo)



Obrázek 8: Porovnání histogramů a graf distribučních funkcí délky hlavy mužů orientovaných výhradně heterosexuálně a mužů orientovaných jinak než výhradně heterosexuálně

```
par(mar = c(5, 4, 1, 2), family = 'Times')
hist(head.LMSa, prob = T, breaks = seq(187, 203, by = 4), main = '', axes = F,
     col = rgb(0, 0, 0.9, 0.2), density = 80, xlim = c(178, 217),
     xlab = 'delka hlavy (v mm)', ylim = c(0, 0.12),
     ylab = 'relativni cetnost')

hist(head.LMOp, prob = T, breaks = seq(180, 215, by = 5), main = '', axes = F,
     col = rgb(0.9, 0, 0, 0.2), density = 80, add = T)
box(bty = 'o')
axis(1, seq(169, 221, by = 4))
axis(2, las = 1)
lines(density(head.LMSa), lwd = 2, col = 'blue')
lines(density(head.LMOp), lwd = 2, col = 'red')
legend('topleft', fill = c(rgb(0.9, 0, 0, 0.2), rgb(0, 0, 0.9, 0.2)),
       legend = c('heterosexualni o.', 'jina o.'), bty = 'n')

plot(ecdf(head.LMSa), col = 'red', main = '', xlim = c(178, 220),
     ylab = 'F(x)', xlab = '', las = 1, cex = 0.6)
plot(ecdf(head.LMOp), add = T, col = 'blue', cex = 0.6)
ks.val <- round(ks.test(head.LMOp - mean(head.LMOp), head.LMSa - (mean(head.LMSa)))$p.
   val, 4)
mtext('delka hlavy (v mm)', side = 1, line = 2.1)
mtext(bquote(paste('KS-test: p-hodnota = ', .(ks.val))), side = 1, line = 3.3)
legend('bottomright', col = c('red', 'blue'), lty = 1, pch = 20,
       legend = c('heterosexualni o.', 'jina o.'), bty = 'n')
```

Příklad 12.5. Wilcoxonův dvouvýběrový asymptotický test

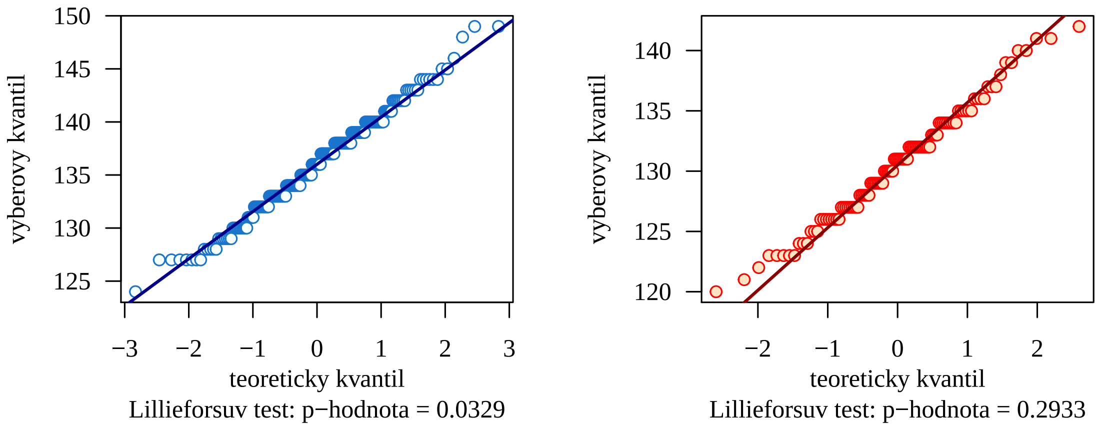
Řešení příkladu 12.5

```r
data <- read.delim('00-Data//05-one-sample-correlation-skull-mf.txt')
head(data)
skull.pHM <- data[data$sex == 'm', 'skull.pH']
skull.pHM <- na.omit(skull.pHM)
skull.pHF <- data[data$sex == 'f', 'skull.pH']
skull.pHF <- na.omit(skull.pHF)
n1 <- length(skull.pHM) # 216
n2 <- length(skull.pHF) # 107
```

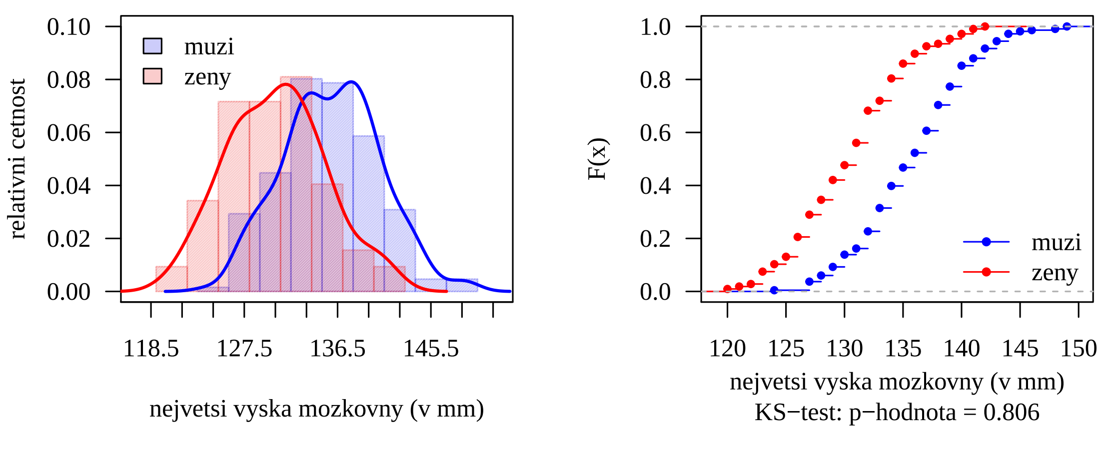
```r
par(mar = c(5, 4, 1, 2), family = 'Times')
qqnorm(skull.pHM, main = '', xlab = '', pch = 21, col = 'dodgerblue3', bg = 'mintcream',
       ylab = 'vyberovy kvantil', las = 1)
mtext('teoreticky kvantil', side = 1, line = 2)
mtext(bquote(paste('Lillieforsuv test: p-hodnota = ', .(round(nortest::lillie.test(skull.pHM)$p.val, 4)))),
      side = 1, line = 3.2)
qqline(skull.pHM, lwd = 2, col = 'darkblue')

qqnorm(skull.pHF, main = '', xlab = '', pch = 21, col = 'red', bg = 'bisque',
       ylab = 'vyberovy kvantil', las = 1)
mtext('teoreticky kvantil', side = 1, line = 2)
mtext(bquote(paste('Lillieforsuv test: p-hodnota = ', .(round(nortest::lillie.test(skull.pHF)$p.val, 4)))),
      side = 1, line = 3.2)
qqline(skull.pHF, lwd = 2, col = 'darkred')
```

★



Obrázek 9: Kvantilové diagramy největší výšky mozkovny mužů (vlevo) a žen (vpravo)



Obrázek 10: Porovnání histogramů a graf distribučních funkcí největší výšky mozkovny mužů a žen

```
par(mar = c(5, 4, 1, 2), family = 'Times')
hist(skull.pHM, prob = T, breaks = seq(123, 150, by = 3), main = '', axes = F,
     col = rgb(0, 0, 0.9, 0.2), xlim = c(117, 152), density = 80, ylim = c(0, 0.10),
     xlab = 'nejvetsi vyska mozkovny (v mm)',
     ylab = 'relativni cetnost')

hist(skull.pHF, prob = T, breaks = seq(119, 143, by = 3), main = '', axes = F,
     col = rgb(0.9, 0, 0, 0.2), density = 80,
     xlab = 'vyska horni casti tvare (v mm)',
     ylab = 'relativni cetnost', add = T)
box(bty = 'o')
axis(1, seq(115.5, 157.5, by = 3))
axis(2, las = 1)
lines(density(skull.pHM), lwd = 2, col = 'blue')
lines(density(skull.pHF), lwd = 2, col = 'red')
legend('topleft', fill = c(rgb(0, 0, 0.9, 0.2), rgb(0.9, 0, 0, 0.2)),
       legend = c('muzi', 'zeny'), bty = 'n')

plot(ecdf(skull.pHM), col = 'blue', main = '', xlim = c(119, 150), ylab = 'F(x)', xlab =
    '',
     cex = 0.6, las = 1)
plot(ecdf(skull.pHF), add = T, col = 'red', cex = 0.6)
ks.val <- round(ks.test(skull.pHM - mean(skull.pHM), skull.pHF - (mean(skull.pHF)))$p.
   val, 4)
mtext('nejvetsi vyska mozkovny (v mm)', side = 1, line = 2.1)
mtext(bquote(paste('KS-test: p-hodnota = ', .(ks.val))), side = 1, line = 3.3)
legend('bottomright', col = c('blue', 'red'), lty = 1, pch = 20,
       legend = c('muzi', 'zeny'), bty = 'n')
```

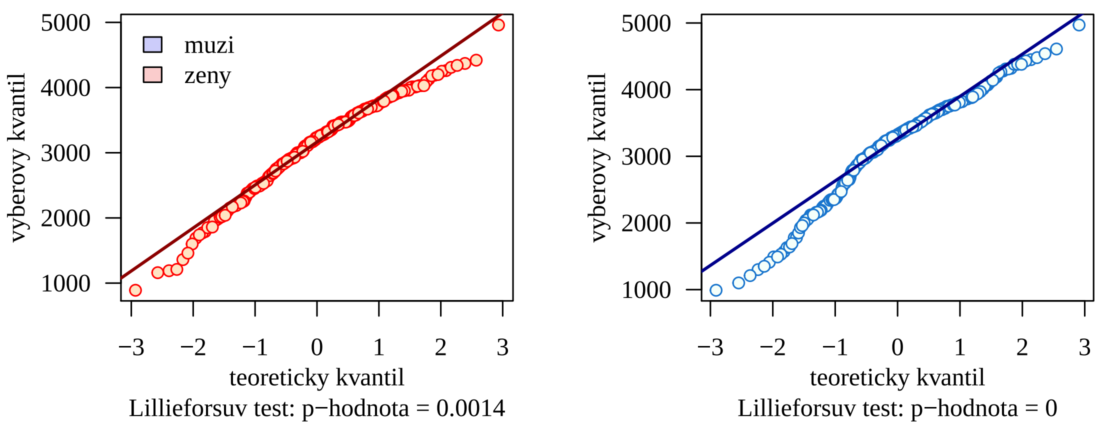
Příklad 12.6. Wilcoxonův dvouvýběrový asymptotický test

Řešení příkladu 12.6

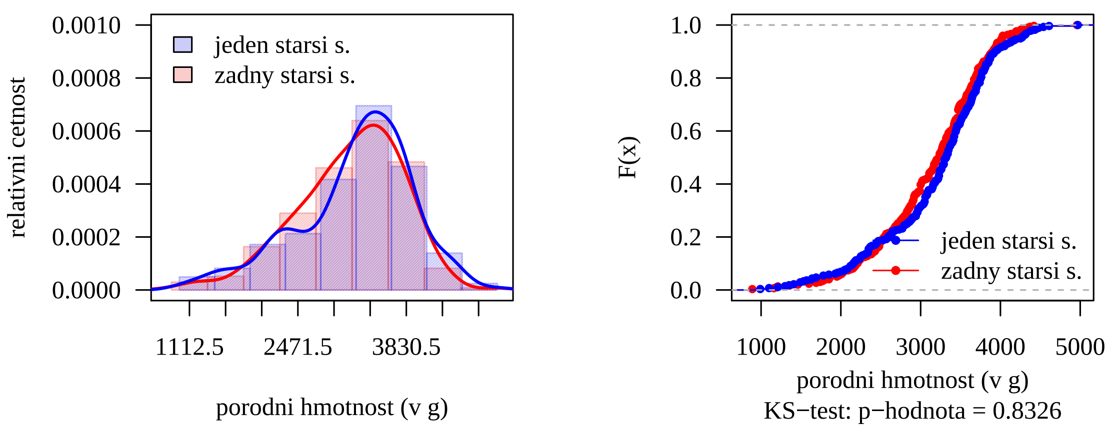
```
data <- read.delim('00-Data//10-two-samples-means-birth.txt')
head(data)
birth.W0 <- data[data$o.sib.N == '0', 'birth.W']
birth.W0 <- na.omit(birth.W0)
birth.W1 <- data[data$o.sib.N == '1', 'birth.W']
birth.W1 <- na.omit(birth.W1)
n1 <- length(birth.W0)
n2 <- length(birth.W1)
```

```
par(mar = c(5, 4, 1, 2), family = 'Times')
qqnorm(birth.W0, main = '', xlab = '', pch = 21, col = 'red', bg = 'bisque',
       ylab = 'vyberovy kvantil', las = 1)
mtext('teoreticky kvantil', side = 1, line = 2)
mtext(bquote(paste('Lillieforsuv test: p-hodnota = ', .(round(nortest::lillie.test(birth
    .W0)$p.val, 4)))),
      side = 1, line = 3.2)
qqline(birth.W0, lwd = 2, col = 'darkred')
legend('topleft', fill = c(rgb(0, 0, 0.9, 0.2), rgb(0.9, 0, 0, 0.2)),
       legend = c('muzi', 'zeny'), bty = 'n')
qqnorm(birth.W1, main = '', xlab = '', pch = 21, col = 'dodgerblue3', bg = 'mintcream',
       ylab = 'vyberovy kvantil', las = 1)
mtext('teoreticky kvantil', side = 1, line = 2)
mtext(bquote(paste('Lillieforsuv test: p-hodnota = ', .(round(nortest::lillie.test(birth
    .W1)$p.val, 4)))),
      side = 1, line = 3.2)
qqline(birth.W1, lwd = 2, col = 'darkblue')
```

★



Obrázek 11: Kvantilové diagramy porodní hmotnosti novorozenců s žádným starším sourozencem (vlevo) a s jedním starším sourozencem (vpravo)



Obrázek 12: Porovnání histogramů a graf distribučních funkcí porodní hmotnosti novorozenců s žádným starším sourozencem a s jedním starším sourozencem

```
par(mar = c(5, 5, 1, 2), family = 'Times')
hist(birth.W0, prob = T, breaks = seq(886, 4963, by = 453), main = '', axes = F,
     col = rgb(0.9, 0, 0, 0.2), density = 80, xlim = c(800, 5000),
     xlab = 'porodni hmotnost (v g)', ylim = c(0, 0.001),
     ylab = '')
hist(birth.W1, prob = T, breaks = seq(986, 4973, by = 443), main = '', axes = F,
     col = rgb(0, 0, 0.9, 0.2), density = 80, add = T)
box(bty = 'o')
axis(1, seq(1112.5, 4736.5, by = 453))
axis(2, las = 1, at = seq(0, 0.0010, by = 0.0002), labels = c('0.0000', '0.0002', '
    0.0004', '0.0006', '0.0008', '0.0010'))
lines(density(birth.W0), lwd = 2, col = 'red')
lines(density(birth.W1), lwd = 2, col = 'blue')
mtext('relativni cetnost', side = 2, line = 4)
legend('topleft', fill = c(rgb(0, 0, 0.9, 0.2), rgb(0.9, 0, 0, 0.2)),
       legend = c('jeden starsi s.', 'zadny starsi s.'), bty = 'n')

plot(ecdf(birth.W0), col = 'red', main = '', xlim = c(800, 5000),
     ylab = 'F(x)', xlab = '', cex = 0.6, las = 1)
plot(ecdf(birth.W1), add = T, col = 'blue', cex = 0.6)
ks.val <- round(ks.test(birth.W1 - mean(birth.W1), birth.W0 - (mean(birth.W0)))$p.val,
    4)
mtext('porodni hmotnost (v g)', side = 1, line = 2.1)
mtext(bquote(paste('KS-test: p-hodnota = ', .(ks.val))), side = 1, line = 3.3)
legend('bottomright', col = c('blue', 'red'), lty = 1, pch = 20,
       legend = c('jeden starsi s.', 'zadny starsi s.'), bty = 'n')
```

12.3 Mann-Whitneyův dvouvýběrový test

Nechť $X_{11}, \dots, X_{1n_1}$, $n_1 \geq 2$ je náhodný výběr ze spojitého rozdělení s distribuční funkcí $F_1(x)$ a $X_{21}, \dots, X_{2n_2}$, $n_2 > 2$ je na něm nezávislý náhodný výběr ze spojitého rozdělení s distribuční funkcí $F_2(x)$. O distribučních funkcích $F_1(x)$ a $F_2(x)$ předpokládáme, že se navzájem liší pouze posunutím. Nechť dále $\tilde{x}_1$ je medián náhodného výběru $X_{11}, \dots, X_{1n_1}$ a $\tilde{x}_2$ je medián náhodného výběru $X_{21}, \dots, X_{2n_2}$ a $\tilde{x}_0$ je konstanta. Na hladině významnosti α testujeme jednu z následujících tří hypotéz oproti příslušné alternativní hypotéze.

$$
\begin{array}{lcll}
H_{01}: \tilde{x}_1 - \tilde{x}_2 = \tilde{x}_0 & \text{oproti} & H_{11}: \tilde{x}_1 - \tilde{x}_2 \neq \tilde{x}_0 & \text{(oboustranná alt.)} \\
H_{02}: \tilde{x}_1 - \tilde{x}_2 \leq \tilde{x}_0 & \text{oproti} & H_{12}: \tilde{x}_1 - \tilde{x}_2 > \tilde{x}_0 & \text{(pravostranná alt.)} \\
H_{03}: \tilde{x}_1 - \tilde{x}_2 \geq \tilde{x}_0 & \text{oproti} & H_{13}: \tilde{x}_1 - \tilde{x}_2 < \tilde{x}_0 & \text{(levostranná alt.)}
\end{array}
$$

Test nazýváme Mann-Whitneyův dvouvýběrový test o rozdílu mediánů $\tilde{x}_1 - \tilde{x}_2$. Testovací statistika má tvar

$$
U = S_E + \frac{n_2(n_2+1)}{2}, \tag{12.3}
$$

kde S_E je testovací statistika definovaná vztahem 12.1, n_2 je rozsah druhého náhodného výběru. Za platnosti nulové hypotézy pochází statistika U_A ze standardizovaného normálního rozdělení, tj.

$$
U = S_E + \frac{n_2(n_2+1)}{2} \overset{H_0}{\sim} N(0,1).
$$

$$
\begin{array}{ll}
H_{11}: \tilde{x}_1 - \tilde{x}_2 \neq \tilde{x}_0 & W = \left(-\infty\,;\, w_{n_1,n_2}(\alpha/2) - \frac{n_2(n_2+1)}{2}\right\rangle \cup \left\langle w_{n_1,n_2}(1-\alpha/2) - \frac{n_2(n_2+1)}{2}\,;\, \infty\right) \\
H_{12}: \tilde{x}_1 - \tilde{x}_2 > \tilde{x}_0 & W = \left\langle w_{n_1,n_2}(1-\alpha) - \frac{n_2(n_2+1)}{2}\,;\, \infty\right) \\
H_{13}: \tilde{x}_1 - \tilde{x}_2 < \tilde{x}_0 & W = \left(-\infty\,;\, w_{n_1,n_2}(\alpha) - \frac{n_2(n_2+1)}{2}\right\rangle
\end{array}
$$

kde $w_{n_1,n_2}(\alpha/2)$, $w_{n_1,n_2}(1-\alpha/2)$, $w_{n_1,n_2}(\alpha)$, $w_{n_1,n_2}(1-\alpha)$ jsou tabelované kvantily pro dvouvýběrový Wilcoxonův test, jejichž hodnoty získáme příkazem `qwilcox()`.

Interval spolehlivosti má podle zvolené alternativní hypotézy jeden z následujících tvarů

$$
\begin{array}{ll}
H_{11}: \tilde{x}_1 - \tilde{x}_2 \neq \tilde{x}_0 & (d,h) = \left(U^{(w_{n_1,n_2}(\alpha/2) - \frac{n_2(n_2+1)}{2})}\,;\, U^{(w_{n_1,n_2}(1-\alpha/2)+1-\frac{n_2(n_2+1)}{2})}\right) \\
H_{12}: \tilde{x}_1 - \tilde{x}_2 > \tilde{x}_0 & (d,\infty) = \left(U^{(w_{n_1,n_2}(\alpha) - \frac{n_2(n_2+1)}{2})}\,;\, \infty\right) \\
H_{13}: \tilde{x}_1 - \tilde{x}_2 < \tilde{x}_0 & (-\infty,h) = \left(-\infty\,;\, U^{(w_{n_1,n_2}(1-\alpha)+1-\frac{n_2(n_2+1)}{2})}\right)
\end{array}
$$

kde $U^{(1)} \leq \dots \leq U^{(n_1 n_2)}$ značí vzestupně seřazené rozdíly $Y_j - X_i$, $i = 1, \dots, n_1$, $j = 1, \dots, n_2$, a $U^{(x)}$ značí x-tý seřazený rozdíl.

p-hodnota má v závislosti na zvolené alternativní hypotéze jeden z následujících tvarů

$$
\begin{array}{ll}
H_{11}: \tilde{x}_1 - \tilde{x}_2 \neq \tilde{x}_0 & p\text{-hodnota} = 2\min\{\Pr(U \leq u)\,,\, \Pr(U > u)\} \\
H_{12}: \tilde{x}_1 - \tilde{x}_2 > \tilde{x}_0 & p\text{-hodnota} = \Pr(U > u) = 1 - \Pr(U \leq u) \\
H_{13}: \tilde{x}_1 - \tilde{x}_2 < \tilde{x}_0 & p\text{-hodnota} = \Pr(U \leq u)
\end{array}
$$

kde U je náhodná veličina, u je realizace testovací statistiky U (viz vzorec 12.3), tedy konkrétní číslo, a $\Pr(U \leq u)$ je distribuční funkce tabelovaného rozdělení pro Wilcoxonův dvouvýběrový exaktní test, jejíž hodnotu získáme pomocí R a implementované funkce `pwilcox()`.

Příklad 12.7. Mann-Whitneyův dvouvýběrový test

Řešení příkladu 12.7

★

```
data <- read.delim('00-Data//18-more-samples-variances-clavicle.txt')
head(data)
cla.LI1 <- data[data$pop == 'ind1', 'cla.L']
cla.LI1 <- na.omit(cla.LI1)
cla.LI2 <- data[data$pop == 'ind2', 'cla.L']
cla.LI2 <- na.omit(cla.LI2)

wilcox.test(cla.LI1, cla.LI2, alt = 't')$p.val # Z
```

Příklad 12.8. Mann-Whitneyův dvouvýběrový test

Řešení příkladu 12.8

```
data <- read.delim('00-Data//15-anova-means-skull.txt')
head(data)
upface.HN <- data[data$pop == 'nem', 'upface.H']
upface.HN <- na.omit(upface.HN)
upface.HC <- data[data$pop == 'cin', 'upface.H']
upface.HC <- na.omit(upface.HC)

wilcox.test(upface.HN, upface.HC, alt = 'g')$p.val # N
```

Příklad 12.9. Mann-Whitneyův dvouvýběrový test

Řešení příkladu 12.9

★

```
data <- read.delim('00-Data//19-more-samples-correlations-skull.txt')
head(data)
intorb.BP <- data[data$pop == 'per', 'intorb.B']
intorb.BP <- na.omit(intorb.BP)
intorb.BB <- data[data$pop == 'ban', 'intorb.B']
intorb.BB <- na.omit(intorb.BB)

wilcox.test(intorb.BP, intorb.BB, alt = 'l')$p.val # Z
```