12 Jednoduchá korelační analýza

12.1 Analýza závislosti dvou veličin ordinálního typu

Je dán náhodný výběr $(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$ z dvourozměrného rozložení, jehož Spearmanův koeficient pořadové korelace je ρ_S. Označíme R_i pořadí náhodné veličiny X_i a Q_i pořadí náhodné veličiny Y_i, $i = 1, 2, \ldots, n$. Dále označíme r_i realizaci pořadí náhodné veličiny X_i a q_i realizaci pořadí náhodné veličiny Y_i, $i = 1, 2, \ldots, n$. Bodovým odhadem ρ_S je (výběrový) Spearmanův koeficient pořadové korelace: $r_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (r_i - q_i)^2$.
Nabývá hodnot mezi –1 a 1. S jeho pomocí zjišťujeme, jak dobře odpovídá vztah veličin X, Y nějaké monotónní funkci, která může být nelineární. Čím je bližší 1, tím je silnější přímá pořadová závislost mezi veličinami X a Y, čím je bližší –1, tím je silnější nepřímá pořadová závislost mezi veličinami X a Y.

12.1.1 Asymptotický interval spolehlivosti pro ρ_S

Předpokládejme, že $n \geq 10$. Označme $z = \frac{1}{2} \ln \frac{1+r_S}{1-r_S}$ (tzv. Fisherova Z-transformace) a definujme následující intervaly spolehlivosti:

- oboustranný:

$$(d; h) = \left(\frac{e^{2\left(z - \sqrt{\frac{1}{n-3}} u_{1-\alpha/2}\right)} - 1}{e^{2\left(z - \sqrt{\frac{1}{n-3}} u_{1-\alpha/2}\right)} + 1}; \frac{e^{2\left(z + \sqrt{\frac{1}{n-3}} u_{1-\alpha/2}\right)} - 1}{e^{2\left(z + \sqrt{\frac{1}{n-3}} u_{1-\alpha/2}\right)} + 1} \right), \tag{12.1}$$

- levostranný:

$$(d; 1) = \left(\frac{e^{2\left(z - \sqrt{\frac{1}{n-3}} u_{1-\alpha}\right)} - 1}{e^{2\left(z - \sqrt{\frac{1}{n-3}} u_{1-\alpha}\right)} + 1}; 1 \right), \tag{12.2}$$

- pravostranný:

$$(-1; h) = \left(-1; \frac{e^{2\left(z + \sqrt{\frac{1}{n-3}} u_{1-\alpha}\right)} - 1}{e^{2\left(z + \sqrt{\frac{1}{n-3}} u_{1-\alpha}\right)} + 1} \right). \tag{12.3}$$

12.1.2 Test hypotézy o pořadové nezávislosti

Na hladině významnosti α testujeme hypotézu H_0: *X, Y jsou pořadově nezávislé náhodné veličiny* (tj. $\rho_S = 0$) proti oboustranné alternativě H_1: *X, Y jsou pořadově závislé náhodné veličiny* (tj. $\rho_S \neq 0$), resp. proti levostranné alternativě H_1: *Mezi X a Y existuje nepřímá pořadová závislost* (tj. $\rho_S < 0$), resp. proti pravostranné alternativě H_1: *Mezi X a Y existuje přímá pořadová závislost* (tj. $\rho_S > 0$).
Jako testová statistika slouží Spearmanův koeficient pořadové korelace r_S. Stanovíme kritický obor W. Pokud $r_S \in W$, H_0 zamítáme na hladině významnosti α a přijímáme H_1.
Pro oboustranný test má kritický obor tvar: $W = (-1; -r_{S,1-\alpha/2}(n)\rangle \cup \langle r_{S,1-\alpha/2}(n); 1)$, pro levostranný test $W = (-1; -r_{S,1-\alpha}(n)\rangle$ a pro pravostranný test $W = \langle r_{S,1-\alpha}(n); 1)$. Kritické hodnoty $r_{S,1-\alpha/2}(n)$, resp. $r_{S,1-\alpha}(n)$ vypočítáme pomocí softwaru R, nebo je najdeme ve statistických tabulkách, ovšem pouze pro hladiny významnosti $\alpha = 0,05$ a $\alpha = 0,01$ a pro $5 \leq n \leq 30$.

a) Pro $n > 20$ lze použít testovou statistiku $T_0 = \frac{r_S\sqrt{n-2}}{\sqrt{1-r_S^2}}$. Platí-li H_0, pak $T_0 \approx t(n-2)$. Stanovíme kritický obor W. Pokud $T_0 \in W$, H_0 zamítáme na asymptotické hladině významnosti α a přijímáme H_1. Pro oboustranný test má kritický obor tvar: $W = (-\infty; -t_{1-\alpha/2}(n-2)\rangle \cup \langle t_{1-\alpha/2}(n-2), \infty)$, pro levostranný test $W = (-\infty; -t_{1-\alpha}(n-2)\rangle$ a pro pravostranný test $W = \langle t_{1-\alpha}(n-2); \infty)$.

b) Pro $n > 30$ lze použít testovou statistiku $T_0 = r_S\sqrt{n-1}$. Platí-li H_0, pak $T_0 \approx N(0, 1)$. Na rozdíl od předešlé situace tedy budou v kritických oborech kvantily $u_{1-\alpha/2}$, resp. $u_{1-\alpha}$.

Upozornění: Popsané metody, které jsou určené pro náhodné veličiny ordinálního typu, se používají i pro veličiny intervalového a poměrového typu, pokud daný náhodný výběr nepochází z dvourozměrného normálního rozložení.

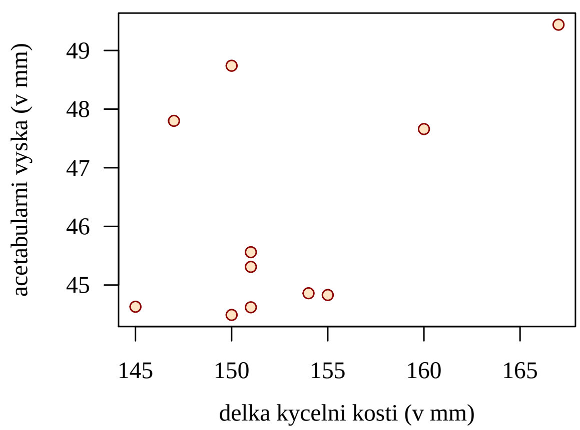
Příklad 12.1. Řešený příklad

Načtěte datový soubor 31-goldman-alaska.csv obsahující údaje o délce kyčelní kosti z pravé strany (iblade.R) a acetabulární výšce (acetab.R) z pravé strany u žen z kmene Ipituaq. (a) Vykreslete dvourozměrný tečkový diagram a okometricky zhodnoťte, zda mezi délkou kyčelní kosti a acetabulární výškou z pravé strany u žen z kmene Ipituaq neexistuje jiný než lineární trend; (b) na hladině významnosti $\alpha = 0,10$ testujte hypotézu, že délka kyčelní kosti a acetabulární výška z pravé strany u žen z kmene Ipituaq nejsou kladně korelované. Testování proveďte (1) kritickým oborem; (2) p-hodnotou. Míru závislosti mezi oběma znaky kvantifikujte pomocí vhodného koeficientu korelace. Jeho hodnotu řádně interpretujte.

Řešení příkladu 12.1

Tečkový diagram vykreslíme pomocí příkazu plot(). Argumentem pch = 21 specifikujeme kulatý tvar bodů, které mohou mít jinou barvu výplně a jinou barvu obrysu. Barvu výplně specifikujeme pomocí argumentu bg, barvu obrysu pomocí argumentu col. Diagram je zobrazen na obrázku 12.1.

```
data <- read.delim('31-goldman-alaska.csv', sep = ';', dec = '.')
data.IF <- data[data$sex == 'f' & data$pop == 'Ipituaq', c('iblade.R', 'acetab.R')]
data.IF <- na.omit(data.IF)
iblade.RIF <- data.IF$iblade.R
acetab.RIF <- data.IF$acetab.R
n <- length(iblade.RIF) # 11
plot(iblade.RIF, acetab.RIF, pch = 21, bg = 'bisque', col = 'darkred', las = 1,
     xlab = 'delka kycelni kosti (v mm)', ylab = 'acetabularni vyska (v mm)')
```



Obrázek 12.1: Dvourozměrný tečkový diagram délky kyčelní kosti a acetabulární výšky z pravé strany žen z kmene Ipituaq

Z tečkového diagramu na obrázku 12.1 je zřejmé, že mezi délkou kyčelní kosti a acetabulární výškou není patrný jiný než lineární trend.

Zadání příkladu vede na test o nezávislosti. Nejprve je třeba ověřit dvourozměrnou normalitu dvourozměrného náhodného výběru žen z kmene Ipituaq. Na hladině významnosti $\alpha = 0,10$ testujeme H_0: *Data pochází z dvourozměrného normálního rozložení.* oproti H_1: *Data nepochází z dvourozměrného normálního rozložení.* K otestování předpokladu dvourozměrné normality použijeme Roystonův test.

```
MVN::mvn(data.IF, mvnTest = 'royston')$multivariateNormality # 0,01562034
```

Náhodný výběr délky kyčelní kosti a acetabulární výšky z pravé strany žen z kmene Ipituaq nepochází z dvourozměrného normálního rozložení (p-hodnota = 0,01562).

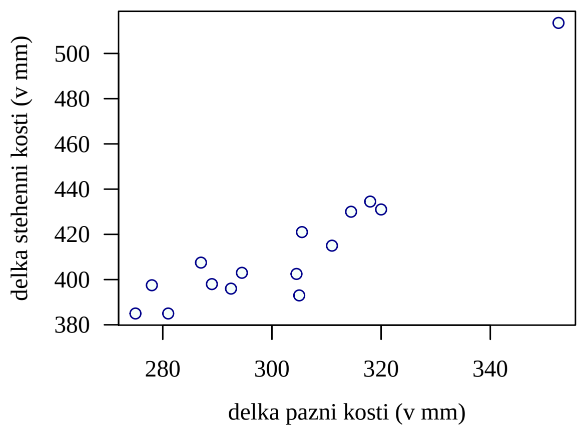
Protože náhodný výběr nepochází z dvourozměrného normálního rozložení, použijeme na ověření zadané hypotézy test o pořadové nezávislosti. Na hladině významnosti $\alpha = 0,10$ testujeme H_0: $\rho_S = 0$ oproti H_1: $\rho_S > 0$ (pravostranný test). Vzhledem k rozsahu náhodného výběru ($n = 11$) zvolíme exaktní variantu tohoto testu. Hodnotu Spearmanova koeficientu pořadové korelace r_S, která je rovněž realizací testové statistiky, vypočítáme příkazem cor() s argumentem method = 'spearman'. Dolní hranici kritického oboru vypočítáme pomocí funkce qSpearman(), p-hodnotu vypočítáme pomocí funkce pSpearman(). Obě funkce pochází z knihovny SuppDists.

```
rS <- cor(iblade.RIF, acetab.RIF, method = 'spearman') # 0,2804783
alpha <- 0.10
SuppDists::qSpearman(1 - alpha, n) # 0,4363636
1 - SuppDists::pSpearman(rS, n) # 0,1931685
```

Realizace testové statistiky $r_S = 0,2805$, kritický obor $W = \langle 0,4364; 1)$. Protože $r_S \notin W$, H_0 nezamítáme na hladině významnosti $\alpha = 0,10$. Protože p-hodnota $= 0,1932$ je větší než $\alpha = 0,10$, H_0 nezamítáme na hladině významnosti $\alpha = 0,10$. Mezi délkou kyčelní kosti a acetabulární výškou z pravé strany u žen z kmene Ipituaq neexistuje statisticky významná přímá pořadová závislost. Na základě hodnoty Spearmanova koeficientu pořadové korelace uvádíme, že mezi délkou kyčelní kosti a acetabulární výškou existuje (statisticky nevýznamný) nízký stupeň přímé pořadové závislosti ($r_S = 0,2805$). ★

Příklad 12.2. Neřešený příklad

Načtěte datový soubor 31-goldman-alaska.csv obsahující údaje o délce pažní kosti z levé strany (humer.L) a délce stehenní kosti (femur.L) z levé strany u mužů z kmene Ipituaq. (a) Vykreslete dvourozměrný tečkový diagram a okometricky zhodnoťte, zda mezi délkou pažní kosti a délkou stehenní kosti z levé strany u mužů z kmene Ipituaq neexistuje jiný než lineární trend; (b) na hladině významnosti $\alpha = 0,01$ testujte hypotézu, že délka pažní kosti a délka stehenní kosti z levé strany u mužů z kmene Ipituaq jsou nezávislé. Testování proveďte (1) kritickým oborem; (2) p-hodnotou. Míru závislosti mezi oběma znaky kvantifikujte pomocí vhodného koeficientu korelace. Jeho hodnotu řádně interpretujte.



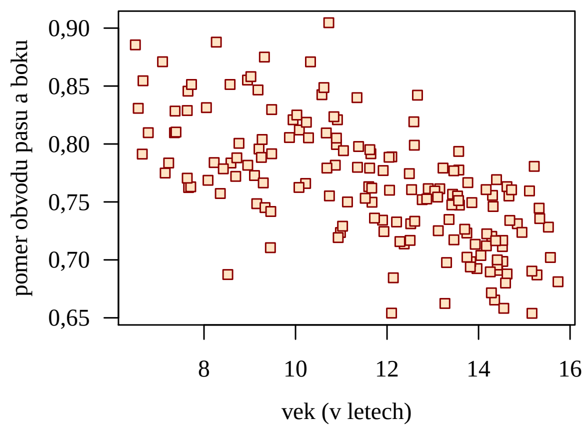
Obrázek 12.2: Dvourozměrný tečkový diagram délky pažní kosti a délky stehenní kosti z levé strany u mužů z kmene Ipituaq

Výsledky: (a) Dvourozměrný tečkový diagram viz obrázek 12.2; mezi oběma proměnnými není patrný jiný než lineární trend; (b) Roystonův test: p-hodnota $= 0,0035$, $\alpha = 0,01$; data nepochází z dvourozměrného normálního rozložení; exaktní varianta ($n \leq 20$) testu o pořadové nezávislosti: $r_S = 0,8508$, $\mathrm{W} = (-1; -0,6375\rangle \cup \langle 0,6446; 1)$; p-hodnota $< 0,0001$, $\alpha = 0,01$; H_0 zamítáme na hladině významnosti $\alpha = 0,01$; Spearmanův koeficient pořadové korelace: $r_S = 0,8508$, vysoký stupeň přímé pořadové závislosti. ★

Příklad 12.3. Řešený příklad
Načtěte datový soubor 32-two-samples-whr-mf.csv obsahující údaje o věku (age) a poměru obvodu pasu a boků (WHR) u dětí ve věku do 16 let (včetně). (a) Vykreslete dvourozměrný tečkový diagram a okometricky zhodnoťte, zda mezi věkem a poměrem obvodu pasu a boků u dívek neexistuje jiný než lineární trend; (b) na hladině významnosti $\alpha = 0,05$ testujte hypotézu, že věk a poměr obvodu pasu a boků u dívek nejsou záporně korelované. Testování proveďte (1) kritickým oborem; (2) intervalem spolehlivosti; (3) p-hodnotou. Míru závislosti mezi oběma znaky kvantifikujte pomocí vhodného koeficientu korelace. Jeho hodnotu řádně interpretujte.

Řešení příkladu 12.3
Tečkový diagram vykreslíme pomocí příkazu plot(). Argumentem pch = 22 specifikujeme čtvercový tvar bodů, které mohou mít jinou barvu výplně a jinou barvu obrysu. Diagram je zobrazen na obrázku 12.3.

```
data <- read.delim('32-two-samples-whr-mf.csv', sep = ';', dec = '.')
data.F <- data[data$sex == 'f', c('age', 'WHR')]
age.F <- data.F$age
WHR.F <- data.F$WHR
n <- length(age.F) # 166
plot(age.F, WHR.F, pch = 22, bg = 'bisque', col = 'darkred', las = 1,
     xlab = 'vek (v letech)', ylab = 'pomer obvodu pasu a boku')
```



Obrázek 12.3: Dvourozměrný tečkový diagram věku a poměru obvodu pasu a boků u dívek

Z tečkového diagramu na obrázku 12.3 je zřejmé, že mezi věkem a poměrem obvodu pasu a boků není patrný jiný než lineární trend.

Zadání příkladu vede na test o nezávislosti. Nejprve ověříme dvourozměrnou normalitu dvourozměrného náhodného výběru dívek. Na hladině významnosti $\alpha = 0,05$ testujeme H_0: *Data pochází z dvourozměrného normálního rozložení.* oproti H_1: *Data nepochází z dvourozměrného normálního rozložení.* K otestování předpokladu dvourozměrné normality použijeme Roystonův test.

```
MVN::mvn(data.F, mvnTest = 'royston')$multivariateNormality # 0,0001821173
```

Náhodný výběr věku a poměru obvodu pasu a boků u dívek nepochází z dvourozměrného normálního rozložení (p-hodnota = 0,0002).

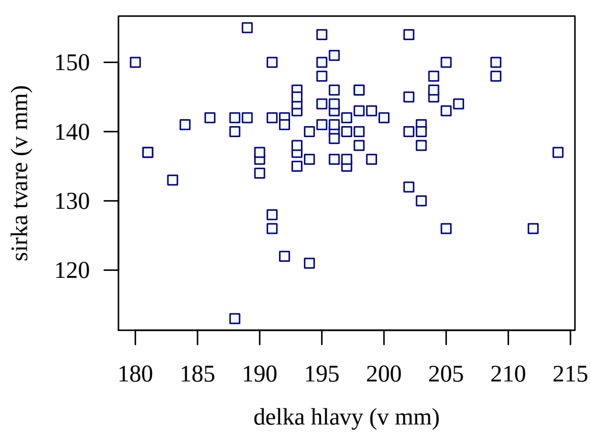
Protože náhodný výběr nepochází z dvourozměrného normálního rozložení, použijeme na ověření zadané hypotézy test o pořadové nezávislosti. Na hladině významnosti $\alpha = 0,05$ testujeme H_0: $\rho_S = 0$ oproti H_1: $\rho_S < 0$ (levostranný test). Vzhledem k rozsahu náhodného výběru ($n = 166$) zvolíme asymptotickou variantu tohoto testu ($n > 30$). Hodnotu Spearmanova koeficientu pořadové korelace vypočítáme pomocí příkazu cor() s argumentem method =

'spearman'. Hodnotu testové statistiky získáme dosazením do vzorce $T_0 = r_S\sqrt{n-1}$. Horní hranici kritického oboru vypočítáme příkazem qnorm(). Horní hranici 95% empirického pravostranného intervalu spolehlivosti vypočítáme dosazením do vzorce 12.3. Výslednou p-hodnotu vypočítáme příkazem pnorm().

```
rS <- cor(age.F, WHR.F, method = 'spearman') # -0,6828958
t0 <- rS * sqrt(n - 1) # -8,771956
alpha <- 0.05
- qnorm(1 - alpha) # -1,644854
z <- 1 / 2 * log((1 + rS) / (1 - rS)) # -0,8345205
hh <- (exp(2 * (z + sqrt(1 / (n - 3)) * qnorm(1 - alpha))) - 1) /
  (exp(2 * (z + sqrt(1 / (n - 3)) * qnorm(1 - alpha))) + 1) # -0,6079643
p.hodnota <- pnorm(t0) # 8,779472e-19
```

Realizace testové statistiky $t_0 = -8,7720$, kritický obor $W = (-\infty; -1,6449\rangle$. Protože $t_0 \in W$, H_0 zamítáme na hladině významnosti $\alpha = 0,05$. Interval spolehlivosti $IS = (-1; -0,6080)$. Protože $c = 0 \notin IS$, H_0 zamítáme na hladině významnosti $\alpha = 0,05$. Protože p-hodnota $< 0,0001$ je menší než $\alpha = 0,05$, H_0 zamítáme na hladině významnosti $\alpha = 0,05$. Mezi věkem a poměrem obvodu pasu a boků u dívek existuje statisticky významná nepřímá pořadová závislost. Na základě hodnoty Spearmanova koeficientu pořadové korelace uvádíme, že mezi věkem a poměrem obvodu pasu a boků u dívek existuje (statisticky významný) význačný stupeň nepřímé pořadové závislosti ($r_S = -0,6829$). ★

Příklad 12.4. Neřešený příklad
Načtěte datový soubor 16-anova-head.txt obsahující údaje o délce hlavy (head.L) a šířce tváře (bizyg.W) mladých dospělých mužů a žen, převážně studentů vysokých škol z Brna a Ostravy. (a) Vykreslete dvourozměrný tečkový diagram a okometricky zhodnoťte, zda mezi délkou hlavy a šířkou tváře mužů neexistuje jiný než lineární trend; (b) na hladině významnosti $\alpha = 0,10$ testujte hypotézu, že délka hlavy a šířka tváře mužů jsou nezávislé. Testování proveďte (1) kritickým oborem; (2) intervalem spolehlivosti; (3) p-hodnotou. Míru závislosti mezi oběma znaky kvantifikujte pomocí vhodného koeficientu korelace. Jeho hodnotu řádně interpretujte.



Obrázek 12.4: Dvourozměrný tečkový diagram délky hlavy a šířky tváře mužů

Výsledky: (a) Dvourozměrný tečkový diagram viz obrázek 12.4; mezi oběma proměnnými není patrný jiný než lineární trend; (b) Roystonův test: p-hodnota $= 0,0066$, $\alpha = 0,10$; data nepochází z dvourozměrného normálního rozložení; asymptotická varianta ($n > 30$) testu o pořadové nezávislosti: $t_0 = 1,4099$, $W = (-\infty; -1,6449\rangle \cup \langle 1,6449; \infty)$; $IS = (-0,0284; 0,3445)$, $c = 0$; p-hodnota $= 0,1586$, $\alpha = 0,10$; H_0 nezamítáme na hladině významnosti $\alpha = 0,10$; Spearmanův koeficient pořadové korelace: $r_S = 0,1639$, nízký stupeň přímé pořadové závislosti. ★

12.2 Analýza závislosti dvou veličin intervalového a poměrového typu

Je dán náhodný výběr $(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$ z dvourozměrného normálního rozložení s koeficientem korelace ρ. Bodovým odhadem ρ je výběrový koeficient korelace $R_{12} = \begin{cases} \frac{S_{12}}{S_1 S_2}, \; S_1, S_2 \neq 0, \\ 0 \text{ jinak}, \end{cases}$ (viz sekce 7.2). Nabývá hodnot mezi –1 a 1. Čím je bližší 1, tím je silnější přímá lineární závislost mezi veličinami X a Y, čím je bližší –1, tím je silnější nepřímá lineární závislost mezi veličinami X a Y.

12.2.1 Asymptotický interval spolehlivosti pro ρ

Předpokládejme, že $n \geq 10$. Označme $Z = \frac{1}{2} \ln \frac{1+R_{12}}{1-R_{12}}$. Vzorce pro meze intervalů spolehlivosti jsou stejné jako vzorce 12.1, 12.2 a 12.3.

12.3 Test hypotézy o nezávislosti

Na hladině významnosti α testujeme hypotézu H_0: *X, Y jsou nezávislé náhodné veličiny* (tj. $\rho = 0$) proti oboustranné alternativě H_1*: X, Y nejsou nezávislé náhodné veličiny* (tj. $\rho \neq 0$), resp. proti levostranné alternativě H_1: *Mezi X a Y existuje nepřímá závislost* (tj. $\rho < 0$), resp. proti pravostranné alternativě H_1: *Mezi X a Y existuje přímá závislost* (tj. $\rho > 0$).

Testová statistika: $T_0 = \frac{R_{12}\sqrt{n-2}}{\sqrt{1-R_{12}^2}}$. Platí-li H_0, pak $T_0 \sim t(n-2)$. Stanovíme kritický obor W. Pokud $T_0 \in W$, H_0 zamítáme na hladině významnosti α a přijímáme H_1. Pro oboustranný test má kritický obor tvar: $W = (-\infty; -t_{1-\alpha/2}(n-2)\rangle \cup \langle t_{1-\alpha}/2(n-2); \infty)$, pro levostranný test $W = (-\infty; -t_{1-\alpha}(n-2)\rangle$ a pro pravostranný test $\langle t_{1-\alpha}(n-2); \infty)$.

Příklad 12.5. Řešený příklad

Načtěte datový soubor 19-more-samples-correlations-skull.txt obsahující údaje o výšce nosu (nose.H) a šířce nosu (nose.B) mužů bantuské, čínské, malajské, německé a peruánské populace. (a) Vykreslete dvourozměrný tečkový diagram a okometricky zhodnoťte, zda mezi výškou nosu a šířkou nosu mužů peruánské populace neexistuje jiný než lineární trend; (b) na hladině významnosti $\alpha = 0,01$ testujte hypotézu, že výška nosu a šířka nosu mužů peruánské populace jsou nezávislé. Testování proveďte (1) kritickým oborem; (2) intervalem spolehlivosti; (3) p-hodnotou. Míru závislosti mezi oběma znaky kvantifikujte pomocí vhodného koeficientu korelace. Jeho hodnotu řádně interpretujte.

Řešení příkladu 12.5

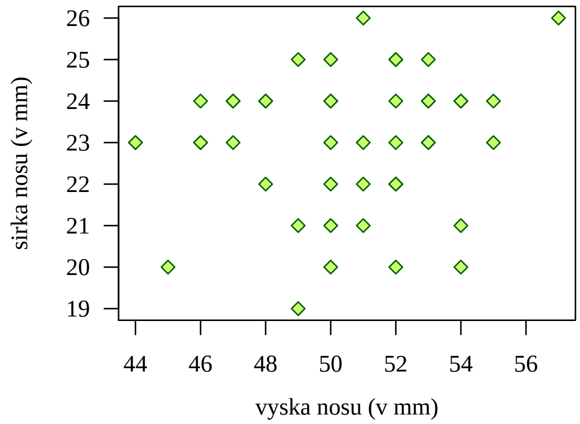
Tečkový diagram vykreslíme pomocí příkazu plot(). Argumentem pch = 23 specifikujeme kosočtvercový tvar bodů, které mohou mít jinou barvu výplně a jinou barvu obrysu. Diagram je zobrazen na obrázku 12.5.

```
data <- read.delim('19-more-samples-correlations-skull.txt', sep = '\t', dec = '.')
data.P <- data[data$pop == 'per', c('nose.H', 'nose.B')]
data.P <- na.omit(data.P)
nose.HP <- data.P$nose.H
nose.BP <- data.P$nose.B
n <- length(nose.HP) # 46
plot(age.F, WHR.F, pch = 23, bg = 'darkolivegreen1', col = 'darkgreen',
     xlab = 'vyska nosu (v mm)', ylab = 'sirka nosu (v mm)', las = 1)
```

Z tečkového diagramu na obrázku 12.5 je zřejmé, že mezi výškou nosu a šířkou nosu není patrný jiný než lineární trend.

Zadání příkladu vede na test o nezávislosti. Nejprve ověříme dvourozměrnou normalitu dvourozměrného náhodného výběru mužů peruánské populace. Na hladině významnosti $\alpha = 0,01$ testujeme H_0: *Data pochází z dvourozměrného normálního rozložení.* oproti H_1: *Data nepochází z dvourozměrného normálního rozložení.* K otestování předpokladu dvourozměrné normality použijeme Roystonův test.

```
MVN::mvn(data.P, mvnTest = 'hz')$multivariateNormality # 0,1305902
```



Obrázek 12.5: Dvourozměrný tečkový diagram výšky nosu a šířky nosu mužů peruánské populace

Náhodný výběr výšek nosu a šířek nosu mužů peruánské populace pochází z dvourozměrného normálního rozložení (p-hodnota = 0,1306).

Protože náhodný výběr pochází z dvourozměrného normálního rozložení, použijeme na ověření zadané hypotézy test o nezávislosti. Na hladině významnosti $\alpha = 0,01$ testujeme H_0: $\rho = 0$ oproti H_1: $\rho \neq 0$ (oboustranný test). Test hypotézy o nezávislosti provedeme pomocí funkce cor.test() s argumentem method = 'pearson'. Výstupem funkce je realizace výběrového koeficientu korelace R_{12}, realizace testové statistiky $T_0 = \frac{R_{12}\sqrt{n-2}}{\sqrt{1-R_{12}^2}}$, interval spolehlivosti a p-hodnota. Hranice kritického oboru dopočítáme příkazem qt().

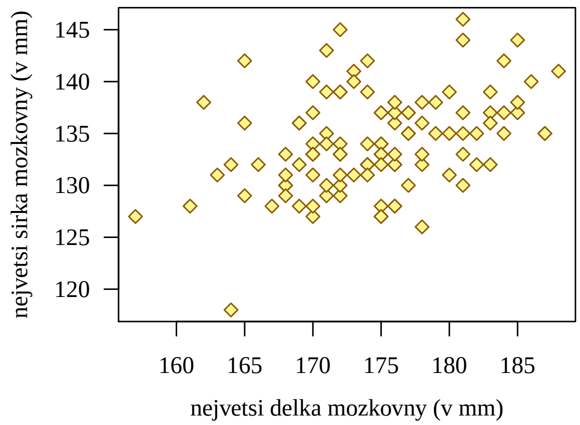
```
cor.test(nose.HP, nose.BP, method = 'pearson', conf.level = 0.99,
         alternative = 'two.sided')
alpha <- 0.01
- qt(1 - alpha / 2, n - 2) # -2,692278
qt(1 - alpha / 2, n - 2) # 2,692278
```

```

        Pearson's␣product-moment␣correlation

data:␣␣nose.HP␣and␣nose.BP
t␣=␣0,91788,␣df␣=␣44,␣p-value␣=␣0,3637
alternative␣hypothesis:␣true␣correlation␣is␣not␣equal␣to␣0
99␣percent␣confidence␣interval:
␣-0,2494938␣␣0,4859523
sample␣estimates:
␣␣␣␣␣␣cor
0,1370691
```

Realizace testové statistiky $t_0 = 0,9179$, kritický obor $W = (-\infty; -2,6923\rangle \cup \langle 2,6923; \infty)$. Protože $t_0 \notin W$, H_0 nezamítáme na hladině významnosti $\alpha = 0,01$. Interval spolehlivosti $IS = (-0.2495; 0,4860)$. Protože $c = 0 \in IS$, H_0 nezamítáme na hladině významnosti $\alpha = 0,01$. Protože p-hodnota $= 0,3637$ je větší než $\alpha = 0,01$, H_0 nezamítáme na hladině významnosti $\alpha = 0,01$. Mezi výškou nosu a šířkou nosu mužů peruánské populace neexistuje statisticky významná závislost. Na základě hodnoty výběrového koeficientu korelace uvádíme, že mezi výškou nosu a šířkou nosu mužů peruánské populace existuje (statisticky nevýznamný) nízký stupeň přímé lineární závislosti ($r_{12} = 0,1371$).

★

Příklad 12.6. Neřešený příklad
Načtěte datový soubor 01-one-sample-mean-skull-mf.txt obsahující údaje o největší délce mozkovny (skull.L) a největší šířce mozkovny (skull.B) mužů a žen starověké egyptské populace. (a) Vykreslete dvourozměrný tečkový diagram a okometricky zhodnoťte, zda mezi největší délkou a šířkou mozkovny žen starověké egyptské populace neexistuje jiný než lineární trend; (b) na hladině významnosti $\alpha = 0,05$ testujte hypotézu, že největší délka a šířka mozkovny žen starověké egyptské populace nejsou kladně korelované. Testování proveďte (1) kritickým oborem; (2) intervalem spolehlivosti; (3) p-hodnotou. Míru závislosti mezi oběma znaky kvantifikujte pomocí vhodného koeficientu korelace. Jeho hodnotu řádně interpretujte.



Obrázek 12.6: Dvourozměrný tečkový diagram největší délky a šířky mozkovny žen starověké egyptské populace

Výsledky: (a) Dvourozměrný tečkový diagram viz obrázek 12.6; mezi oběma proměnnými není patrný jiný než lineární trend; (b) Roystonův test: p-hodnota $= 0,4072$, $\alpha = 0,05$; data pochází z dvourozměrného normálního rozložení; test o nezávislosti: $t_0 = 4,2616$, $\mathrm{W} = \langle 1,6592; \infty)$; $IS = (0,2368; 1)$, $c = 0$; p-hodnota $< 0,0001$, $\alpha = 0,05$; H_0 zamítáme na hladině významnosti $\alpha = 0,05$; výběrový koeficient korelace: $r_{12} = 0,3809$, mírný stupeň přímé lineární závislosti. ★

12.4 Test o rozdílu dvou koeficientů korelace

Máme dva nezávislé náhodné výběry o rozsazích $n \geq 10$ a $n^* \geq 10$ z dvourozměrných normálních rozložení s koeficienty korelace ρ a ρ^*. R_{12} a R_{12}^* jsou výběrové koeficienty korelace prvního a druhého výběru. Položme $Z = \frac{1}{2} \ln \frac{1+R_{12}}{1-R_{12}}$ a $Z^* = \frac{1}{2} \ln \frac{1+R_{12}^*}{1-R_{12}^*}$. Testujeme H_0: $\rho = \rho^*$ proti H_1: $\rho \neq \rho^*$, resp. proti některé z jednostranných alternativ.
Testová statistika: $U_0 = \frac{Z-Z^*}{\sqrt{\frac{1}{n-3}+\frac{1}{n^*-3}}}$. Platí-li H_0, pak $U_0 \approx N(0,1)$. Stanovíme kritický obor W. Je-li $U_0 \in W$, H_0 zamítáme na asymptotické hladině významnosti α a přijímáme H_1.

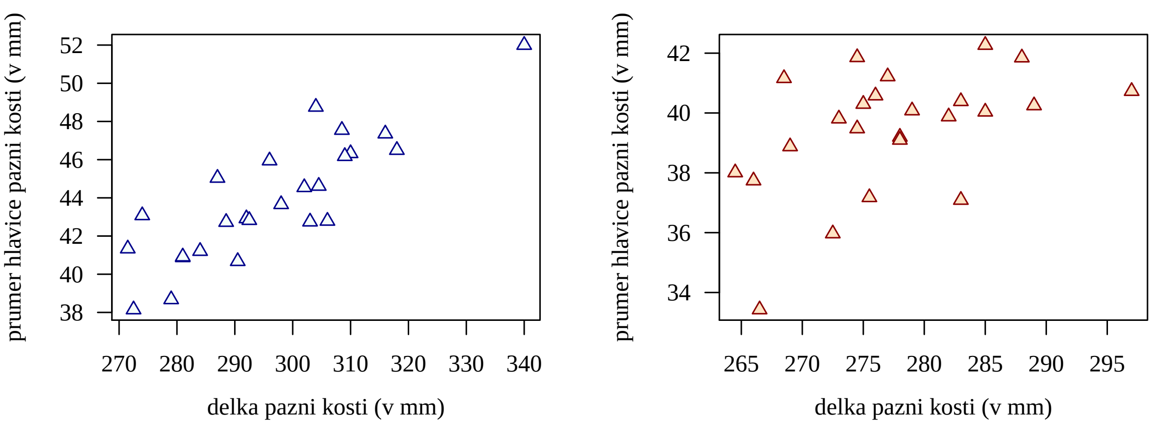
Kritický obor pro oboustranný test je $W = (-\infty; -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}; \infty)$, pro levostranný test $W = (-\infty; -u_{1-\alpha}\rangle$ a pro pravostranný test $W = \langle u_{1-\alpha}; \infty)$.

Příklad 12.7. Řešený příklad
Načtěte datový soubor 31-goldman-alaska.csv obsahující údaje o délce pažní kosti z levé strany (humer.L) a průměru hlavice pažní kosti z levé strany (humer.HDL) u mužů a žen aljašské populace z kmene Tigara. (a) Vykreslete dvourozměrný tečkový diagram a okometricky zhodnoťte, zda mezi délkou pažní kosti a průměrem hlavice pažní kosti z levé strany mužů resp. žen z kmene Tigara neexistuje jiný než lineární trend; (b) na hladině významnosti $\alpha = 0,05$ testujte hypotézu, že koeficient korelace délky pažní kosti a průměru hlavice pažní kosti z levé strany u mužů z kmene Tigara je stejný jako u žen z kmene Tigara. Testování proveďte (1) kritickým oborem; (2) p-hodnotou. Míru závislosti mezi oběma znaky u mužů, resp. u žen kvantifikujte pomocí vhodného koeficientu korelace. Hodnoty obou koeficientů korelace řádně interpretujte.

Řešení příkladu 12.7

Tečkový diagram pro muže, resp. pro ženy vykreslíme pomocí příkazu plot(). Argumentem pch = 24 specifikujeme trojúhelníkový tvar bodů, které mohou mít jinou barvu výplně a jinou barvu obrysu. Oba diagramy jsou zobrazeny na obrázku 12.7.

```
data <- read.delim('31-goldman-alaska.csv', sep = ';', dec = '.')
data.TM <- data[data$sex == 'm' & data$pop == 'Tigara', c('humer.L', 'humer.HDL')]
data.TF <- data[data$sex == 'f' & data$pop == 'Tigara', c('humer.L', 'humer.HDL')]
data.TM <- na.omit(data.TM)
data.TF <- na.omit(data.TF)
humer.LTM <- data.TM$humer.L
humer.HDLTM <- data.TM$humer.HDL
humer.LTF <- data.TF$humer.L
humer.HDLTF <- data.TF$humer.HDL

plot(humer.LTM, humer.HDLTM, pch = 24, bg = 'bisque', col = 'darkred', las = 1,
     xlab = 'delka pazni kosti (v mm)', ylab = 'prumer hlavice pazni kosti (v mm)')
plot(humer.LTF, humer.HDLTF, pch = 24, bg = 'bisque', col = 'darkred', las = 1,
     xlab = 'delka pazni kosti (v mm)', ylab = 'prumer hlavice pazni kosti (v mm)')
```



Obrázek 12.7: Dvourozměrný tečkový diagram délky pažní kosti a průměru hlavice pažní kosti z levé strany u mužů (vlevo), resp. u žen (vpravo) aljašské populace z kmene Tigara

Z tečkového diagramu na obrázku 12.7 je zřejmé, že mezi délkou pažní kosti a průměrem hlavice pažní kosti u mužů ani u žen není patrný jiný než lineární trend.

Zadání příkladu vede na test o rozdílu dvou koeficientů korelace. Nejprve ověříme dvourozměrnou normalitu dvourozměrného náhodného výběru mužů, resp. žen z kmene Tigara. Na hladině významnosti $\alpha = 0,05$ testujeme H_0: *Data pochází z dvourozměrného normálního rozložení.* oproti H_1: *Data nepochází z dvourozměrného normálního rozložení.* K otestování předpokladu dvourozměrné normality použijeme pro každý z výběrů Roystonův test.

```
MVN::mvn(data.TM, mvnTest = 'royston')$multivariateNormality # 0,6324143
MVN::mvn(data.TF, mvnTest = 'royston')$multivariateNormality # 0,09177741
```

Náhodný výběr délek pažní kosti a průměrů hlavice pažní kosti z levé strany u mužů z kmene Tigara pochází z dvourozměrného normálního rozložení (p-hodnota = 0,6324). Náhodný výběr délek pažní kosti a průměrů hlavice pažní kosti z levé strany u žen z kmene Tigara pochází z dvourozměrného normálního rozložení (p-hodnota = 0,0918).

Protože oba náhodné výběry pochází z dvourozměrných normálních rozložení, použijeme na ověření zadané hypotézy test o rozdílu dvou koeficientů korelace. Na hladině významnosti $\alpha = 0,05$ testujeme H_0: $\rho = \rho^*$ oproti H_1:

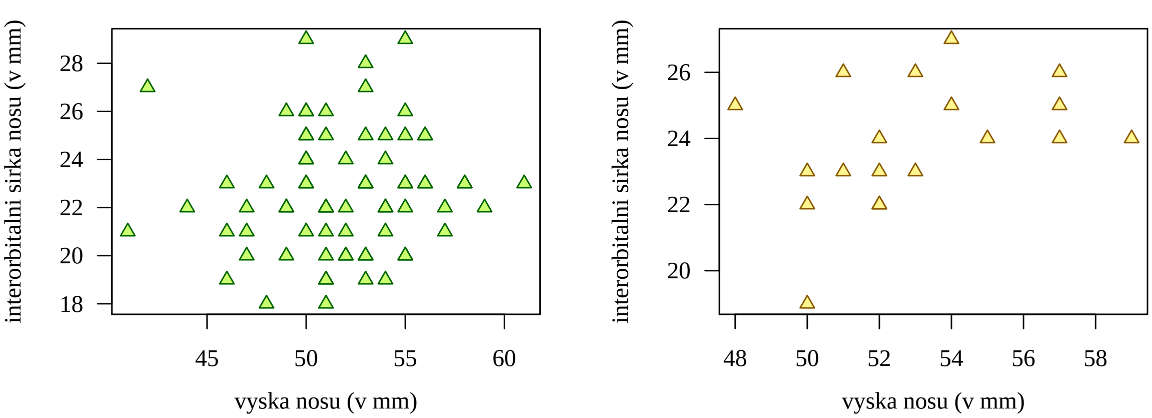
$\rho \neq \rho^*$ (oboustranný test). Rozsahy obou náhodných výběrů n a n^* vypočítáme příkazem length(). Hodnoty realizací výběrových koeficientů korelace R_{12} a R^*_{12} vypočítáme příkazem cor() s argumentem method = 'pearson'. Dále vypočítáme Fisherovy Z-transformace obou výběrových koeficientů korelace podle vzorců $Z = \frac{1}{2}\ln\frac{1+R_{12}}{1-R_{12}}$ a $Z^* = \frac{1}{2}\ln\frac{1+R^*_{12}}{1-R^*_{12}}$, a nakonec stanovíme realizaci testové testistiky $U_0 = \frac{Z-Z^*}{\sqrt{\frac{1}{n-3}+\frac{1}{n^*-3}}}$. Hranice kritického oboru dopočítáme příkazem qnorm(), p-hodnotu příkazem pnorm().

```
nM <- length(humer.LTM) # 25
nF <- length(humer.LTF) # 24
r12M <- cor(humer.LTM, humer.HDLTM, method = 'pearson') # 0,8523391
r12F <- cor(humer.LTF, humer.HDLTF, method = 'pearson') # 0,4836377
zM <- 1 / 2 * log((1 + r12M) / (1 - r12M)) # 1,264643
zF <- 1 / 2 * log((1 + r12F) / (1 - r12F)) # 0,5277218
u0 <- (zM - zF) / sqrt(1 / (nM - 3) + 1 / (nF - 3)) # 2,415505
alpha <- 0.05
- qnorm(1 - alpha / 2) # -1,959964
qnorm(1 - alpha / 2) # 1,959964
p.hodnota <- 2 * min(pnorm(u0), 1 - pnorm(u0)) # 0,0157134
```

Realizace testové statistiky $u_0 = 2,4155$, kritický obor $W = (-\infty; -1,9600\rangle \cup \langle 1,9600; \infty)$. Protože $u_0 \in W$, H_0 zamítáme na hladině významnosti $\alpha = 0,05$. Protože p-hodnota $= 0,0157$ je menší než $\alpha = 0,05$, H_0 zamítáme na hladině významnosti $\alpha = 0,05$. Mezi koeficientem korelace délky pažní kosti a průměru hlavice pažní kosti z levé strany u mužů a u žen z kmene Tigara existuje statisticky významný rozdíl. Mezi délkou pažní kosti a průměrem hlavice pažní kosti z levé strany u mužů z kmene Tigara existuje vysoký stupeň přímé lineární závislosti ($r_{12} = 0,8523$), mezi délkou pažní kosti a průměrem hlavice pažní kosti z levé strany u žen z kmene Tigara existuje mírný stupeň přímé lineární závislosti ($r^*_{12} = 0,4836$). ★

Příklad 12.8. Neřešený příklad

Načtěte datový soubor 19-more-samples-correlations-skull.txt obsahující údaje o výšce nosu (nose.H) a interorbitální šířce nosu (intorb.B) mužů bantuské, čínské, malajské, německé a peruánské populace. (a) Vykreslete dvourozměrný tečkový diagram a okometricky zhodnoťte, zda mezi výškou nosu a interorbitální šířkou nosu mužů malajské, resp. čínské populace neexistuje jiný než lineární trend; (b) na hladině významnosti $\alpha = 0,10$ testujte hypotézu, že koeficient korelace výšky a interorbitální šířky nosu mužů malajské populace je větší nebo rovný koeficientu korelace výšky a interorbitální šířky nosu mužů čínské populace. Testování proveďte (1) kritickým oborem; (2) p-hodnotou. Míru závislosti mezi oběma znaky u mužů čínské, resp. malajské populace kvantifikujte pomocí vhodného koeficientu korelace. Hodnoty obou koeficientů korelace řádně interpretujte.



Obrázek 12.8: Dvourozměrný tečkový diagram výšky nosu a interorbitální šířky nosu mužů malajské populace (vlevo), resp. čínské populace (vpravo)

Výsledky: (a) Dvourozměrný tečkový diagram pro muže malajské, resp. čínské populace viz obrázek 12.8 vlevo, resp. vpravo; mezi oběma proměnnými není v ani jedné populaci patrný jiný než lineární trend; (b) Roystonův test (malajská populace): p-hodnota $= 0,7360$, $\alpha = 0,10$; data pochází z dvourozměrného normálního rozložení; Roystonův test (čínská populace): p-hodnota $= 0,6342$, $\alpha = 0,10$; data pochází z dvourozměrného normálního rozložení; test o rozdílu dvou koeficientů korelace: $u_0 = -1,1371$, W $= (-\infty; -1,2816\rangle$; p-hodnota $= 0,1278$, $\alpha = 0,10$; H_0 nezamítáme na hladině významnosti $\alpha = 0,10$; výběrový koeficient korelace (malajská populace): $r_{12} = 0,0862$, velmi nízký stupeň přímé lineární závislosti; výběrový koeficient korelace (čínská populace): $r_{12}^* = 0,3812$, mírný stupeň přímé lineární závislosti.

★