Cvičení z biostatistiky 07

Poslední úprava dokumentu: 3. dubna 2024.

Korelace, dvouvýběrový t-test a Wilcoxonův test

Znovu se vracíme k tématu korelace, kterým jsme se zabývali již na 2. cvičení. Osvěžte si prosím své znalosti o korelaci buď v poznámkách z přednášky, nebo textu *Korelace* (dále jako *KRL*), který najdete na mých stránkách.

Rozcvička

Vytvořte si v R dva vektory, každý po 8 hodnotách, a vykreslete je proti sobě do bodového diagramu (scatter plotu). Dále spočtěte jejich korelaci.

1 Úvod

Budeme pokračovat v analýze dat Kojeni.

1) Načtěte si data `Kojeni.RData` do RStudia.

2) Zajistěte si přímý přístup k proměnným.

```
attach(Kojeni)
```

2 Pearsonův korelační koeficient

Závislost mezi výškou matky a výškou otce

Prozkoumejme, jak silně (a zda vůbec) spolu souvisí výška matky (vyska.m) a výška otce (vyska.o) a je-li závislost mezi výškou matky a otce průkazná. Označme jako X náhodnou veličinu, jež reprezentuje výšku náhodně vybrané matky a jako Y náhodnou veličinu, jež reprezentuje výšku jejího partnera. Nechť ρ je populační korelace mezi X a Y, která v jistém smyslu měří sílu závislosti mezi X a Y.

1) Nakresleme bodový graf (*scatterplot*). Vzhledem k tomu, že výška byla měřena na celé centimetry, nachází se v datech nemálo dvojic se shodnými výškami otců i matek, jež v grafu splynou v jeden bod. Pro získání lepší představy o datech bude proto výhodnější před nakreslením posunout každou výšku o malý kousek náhodně doleva či doprava (*jittering*).

```
plot(vyska.o ~ jitter(vyska.m), xlab="vyska matky", ylab="vyska otce")
```

případně obyčejný bodový graf:

```
plot(vyska.o~vyska.m)
```

2) Spočtěme hodnotu Pearsonova korelačního koeficientu mezi výškou matky a otce.

```
cor(vyska.m, vyska.o)
```

3) V případě, že lze předpokládat dvourozměrné normální rozdělení pro (X, Y), lze testovat H_0: $\varrho = 0$ proti H_1: $\varrho \neq 0$. Zamítneme-li tedy nulovou hypotézu, prokážeme závislost mezi X a Y (výškou matky a otce). Hladinu testu budeme uvažovat 5 %.

- ✧ To, zda vektor (X, Y) má dvourozměrné normální rozdělení ověříme pohledem na scatterplot (viz text *KRL*, str. 2, Poznámka 1). Body na scatterplotu se shlukují do elipsy, proto se domníváme, že normální rozdělení lze předpokládat.
- ✧ Test nezávilosti (viz text *KRL*, str. 1, dole) provedeme příkazem:

```
cor.test(vyska.m, vyska.o)
```

```
        Pearson's product-moment correlation
                                  počet stupňů volnosti t-rozdělení, které má testová statistika za H0 (n-2)
data:  Kojeni$vyska.m and Kojeni$vyska.o
hodnota testové statistiky — t = 2.0678, df = 97, p-value = 0.04132 — p-hodnota
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:                 — 95% interval spolehlivosti pro populační
 0.008402418 0.387180104                          korelační koeficient ρ
sample estimates:
      cor
0.2054734   — hodnota Pearsonova korelačního koeficientu
```

- ✧ Jaký je závěr?
  P-hodnota je 0.04, tedy menší než zvolená hladina 0.05. Zamítáme proto nulovou hypotézu o nezávislosti veličin. Na 5% hladině jsme prokázali, že veličiny *výška otce* a *výška matky* jsou závislé.
- ✧ **UPOZORNĚNÍ:** Interval spolehlivosti pro ρ ve výstupu je pouze přibližný a nemusí tedy nutně přesně korespondovat s výsledkem testu. To jest, může se stát, že zamítneme (těsně) nulovou hypotézu a interval spolehlivosti bude přesto (těsně) překrývat nulu či naopak.
- ✧ Jak se od sebe liší právě provedený test nezávislosti výšky matky a otce od párového testu s výškou matky a otce, kterým jsme se zabývali v předchozím pracovním listu?

3 Spearmanův korelační koeficient

Závislost mezi věkem matky a věkem otce

Prozkoumejme, jak silně (a zda vůbec) spolu souvisí věk matky (vek.m) a věk otce (vek.o) a je-li závislost mezi věkem matky a otce průkazná. Označme jako X náhodnou veličinu, jež reprezentuje věk náhodně vybrané matky a jako Y náhodnou veličinu, jež reprezentuje věk jejího partnera.

1) Nakreslete bodový graf (*scatterplot*). Opět se bude hodit *jittering* (proč?).

```
plot(vek.o ~ jitter(vek.m), xlab="vek matky", ylab="vek otce")
```

2) Zejména díky tomu, že jen velice zřídka kdy je otec mladší než matka, vyskytuje se většina mraku v bodovém grafu nad přímkou $y = x$. K tomu, aby bylo možné předpokládat pro (X, Y) dvourozměrné normální rozdělení, by bodový graf měl připomínat elipsu. Zkoumání závislosti mezi X (věk matky) a Y (věk otce) bude proto vhodnější založit na Spearmanově korelačním koeficientu.

3) Spočtěme hodnotu Spearmanova korelačního koeficientu (v textu *KRL* je to vzorec (3)).

```
cor(vek.m, vek.o, method="spearman")
```

4) Otestujme dále H_0: „X, Y jsou nezávislé“ proti H_1: „X, Y jsou závislé“. Zamítneme-li tedy nulovou hypotézu, prokážeme závislost mezi věkem matky a otce.

```
cor.test(vek.m, vek.o, method="spearman")
```

```
        Spearman's rank correlation rho

data:  vek.m and vek.o
S = 43624, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7302145

Warning message:
In cor.test.default(vek.m, vek.o, method = "spearman") :
  Cannot compute exact p-value with ties
```

hodnota testové statistiky — S = 43624

p-hodnota — p-value < 2.2e-16

Spearmanův korelační koeficient — rho 0.7302145

Chybová hláška upozorňující na to, že některé dvojice se v datech vyskytují vícekrát, což způsobuje problémy při výpočtu přesné p-hodnoty. — Cannot compute exact p-value with ties

R nepoužívá testovou statistiku $|r^s_{X,Y}|\sqrt{n-1}$, která byla uvedena na přednášce (či v textu *KRL*), ale používá statistiku S definovanou jako

$$S = (n^3 - n)\frac{1 - r^s_{X,Y}}{6},$$

kde n je počet pozorování a $r^s_{X,Y}$ je Spearmanův korelační koeficient.

✧ Jaký je závěr?
P-hodnota vyšla menší než 2.2×10^{-16}, tedy velmi malá. Zamítáme tedy nulovou hypotézu. Na 5% hladině jsme prokázali, že věk matky a otce jsou silně závislé veličiny.

Ilustrace rozdílu mezi Pearsonovým a Spearmanovým korelačním koeficientem

Jak už jste asi slyšeli, Pearsonův korelační koeficient měří sílu **lineární** závislosti veličin, kdežto Spearmanův korelační koeficient měří sílu **monotónní** (= buď rostoucí, nebo klesající, ale ne nutně lineární) závislosti veličin. Pearsonův koeficient je tedy největší (nabývá hodnoty ± 1), když jedna z veličin je lineární funkcí druhé (tj. např. $Y = a + bX$). Aby Spearmanův koeficient nabýval hodnoty ± 1, stačí mu k tomu, aby jedna z veličin byla rostoucí/klesající funkcí druhé veličiny. Pojd'me si to ilustrovat na příkladu:

✧ Vytvořme si vektor pěti hodnot, představující naměřené hodnoty veličiny X

```
x <- c(1,3,8,4,6)
```

✧ Nyní vytvořme vektor y jako lineární funkci x (já si vezmu třeba $y = 2x + 4$, ale klidně zkuste něco jiného)

```
y <- 2*x + 4
```

✧ Nyní spočtěme hodnoty Pearsonova a Spearmanova korelačního koeficientu

```
cor(x, y)
cor(x, y, method="spearman")
```

Oba koeficienty vyšly 1.

✧ Nyní zkusme y vytvořit pomocí funkce, která bude rostoucí, ale ne lineárně. Například zvolme logaritmus (ale vyzkoušet můžete třeba také x^2, která je na kladných číslech též rostoucí).

```
y <- log(x)
```

✧ Spočtěme opět hodnoty Pearsonova a Spearmanova korelačního koeficientu

```
cor(x, y)
cor(x, y, method="spearman")
```

Pearsonův koeficient, který měří sílu lineární závislosti, již nevyšel 1 (protože logaritmus není lineární funkce), ale Spearmanův koeficient vyšel opět 1 (protože logaritmus je ryze rostoucí, a tudíž monotónní).

✧ Na závěr si můžeme zkusit vytvořit vektor y pomocí funkce, která není lineární ani monotónní... například sinus.

```
y <- sin(x)
cor(x, y)
cor(x, y, method="spearman")
```

Ani jeden z koeficientů již nedá hodnotu 1.

✧ Pro lepší představu si můžete vždy vykreslit i bodový graf

```
plot(x,y)
```

4 Korelační matice

Prozkoumejme, jak spolu souvisí tělesné míry dítěte ve 24. týdnu a výška rodičů (delka, hmotnost, vyska.m, vyska.o).

1) Nakresleme bodové grafy pro všechny uvažované dvojice proměnných, jež nás nyní zajímají.

```
plot(Kojeni[, c("delka", "hmotnost", "vyska.m", "vyska.o")], col="red", pch=16)
```

2) Spočtěme hodnoty výběrového korelačního koeficientu pro všechny uvažované dvojice proměnných, jež nás nyní zajímají.

```
cor(Kojeni[, c("delka", "hmotnost", "vyska.m", "vyska.o")])
```

✧ Výsledná tabulka čísel se nazývá *korelační matice*. V každém poli je uvedena hodnota Pearsonova korelačního koeficientu pro příslušnou dvojici veličin. Matice je symetrická, neboť hodnota korelace nezávisí na pořadí veličin (tj. $cor(X, Y) = cor(Y, X)$). Na diagonále této matice jsou jedničky, protože korelace veličiny s ní samou je 1 (tj. $cor(X, X) = 1$).

3) Otestujte, zda je významná závislost mezi délkou a hmotností dítěte ve 24. týdnu.

4) Otestujte, zda je významná závislost mezi délkou dítěte ve 24. týdnu a výškou matky.

5 Dvouvýběrový t-test: závislost hmotnosti ve 24. týdnu na pohlaví

Nyní se zaměříme na testy sloužící k porovnání středních hodnot dvou populací, ze kterých máme k dispozici výběry. Prostudujte si prosím příslušnou teorii v textu *Dvouvýběrové testy* (dále jako *DVT*), který je k dispozici na mých stránkách.

Pokusíme se zjistit, zda hmotnost ve 24. týdnu závisí na pohlaví. Jedním z projevů závislosti hmotnosti na pohlaví je rozdílná střední hodnota hmotnosti pro chlapce a dívky. Pokusíme se tedy prokázat, že E(hmotnost chlapce) $\neq \mathsf{E}$(hmotnost dívky).

Označme jako X hmotnost náhodně vybraného chlapce a jako Y hmotnost náhodně vybrané dívky. Předpokládejme, že $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ a $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$, přičemž X a Y jsou vzájemně nezávislé. Za platnosti těchto předpokladů můžeme pomocí *dvouvýběrového t-testu* testovat $\mathrm{H}_0 : \mu_x = \mu_y$ proti $\mathrm{H}_1 : \mu_x \neq \mu_y$. Hladinu testu uvažujme $\alpha = 0.05$.

1) Začněte tím, že spočítáte základní popisné statistiky pro hmotnost v závislosti na pohlaví.

```
tapply(hmotnost, Hoch, summary)
```

2) Pokračujte tím, že si uděláte představu o platnosti testovaných hypotéz pomocí krabičkových grafů hmotnosti dle pohlaví.

```
boxplot(hmotnost ~ Hoch, col=c("pink", "blue"), data=Kojeni, ylab="Hmotnost (g)")
```

3) Nakreslete graf průměrů (plot of means) s „anténami“ odvozenými ze směrodatných chyb průměrů a dále z intervalů spolehlivosti pro střední hodnotu.

```
install.packages("RcmdrMisc")  # nainstaluje knihovnu
library(RcmdrMisc)        # otevre knihovnu
plotMeans(hmotnost, Hoch, error.bars = "conf.int", level=0.95)
plotMeans(hmotnost, Hoch, error.bars = "se")
```

Ověřování předpokladů dvouvýběrového t-testu

4) Je smysluplné předpokládat v této situaci, že X a Y jsou nezávislé náhodné veličiny?

✧ Ano. Předpokládáme, že děti svou hmotnost vzájemně nijak neovlivňovaly...

5) Lze předpokládat, že rozdělení hmotnosti je jak u chlapců tak u dívek normální?

✧ Normalitu je třeba ověřit zvlášť u chlapců i dívek.

✧ Oba normální QQ grafy lze příkazy nakreslit následovně:

```
qqnorm(hmotnost[Hoch=="divka"], pch=16, col="red", main="Divky",
       xlab="Kvantily N(0, 1)", ylab="Hmotnost (g)")
qqline(hmotnost[Hoch=="divka"], col="darkblue")
qqnorm(hmotnost[Hoch=="hoch"], pch=16, col="red", main="Chlapci",
       xlab="Kvantily N(0, 1)", ylab="Hmotnost (g)")
qqline(hmotnost[Hoch=="hoch"], col="darkblue")
```

✧ Shapirův-Wilkův test normality zvlášť pro chlapce a dívky

```
shapiro.test(hmotnost[Hoch=="divka"])
shapiro.test(hmotnost[Hoch=="hoch"])
```

P-hodnoty obou testů jsou výrazně větší než zvolená hladina 0.05, a tudíž u obou výběrů můžeme normalitu předpokládat.

6) Základní varianta dvouvýběrového t-testu předpokládá shodné rozptyly hmotnosti mezi chlapci i dívkami.

✧ Co si myslíte o platnosti tohoto předpokladu na základě krabičkových grafů?

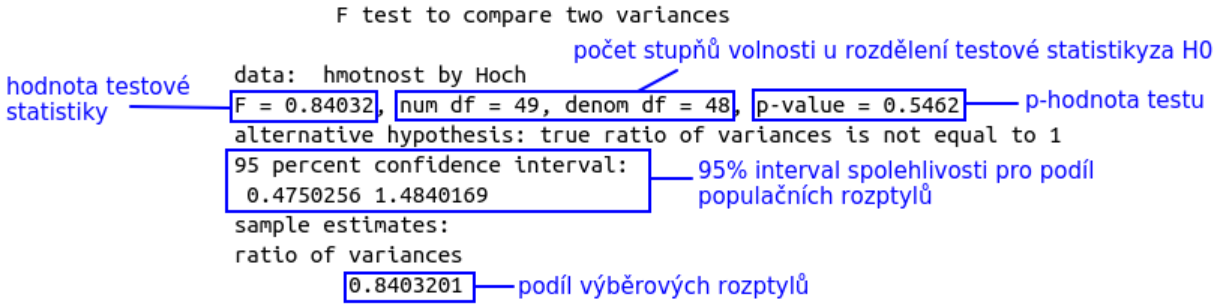
```
boxplot(hmotnost ~ Hoch)
```

Výška krabičky odpovídá mezikvartilovému rozpětí, které je taky určitou charakteristikou variability. Krabičky se zdají být přibližně stejně vysoké, shodu rozptylů bychom tedy mohli předpokládat.

✧ Otestujme nyní pomocí F-testu hypotézu $H_0 : \sigma_x^2 = \sigma_y^2$ proti $H_1 : \sigma_x^2 \neq \sigma_y^2$.

```
var.test(hmotnost ~ Hoch, data=Kojeni)    # nebo
var.test(KojeniD$hmotnost, KojeniH$hmotnost)
```

✧ Výstup vypadá následovně:

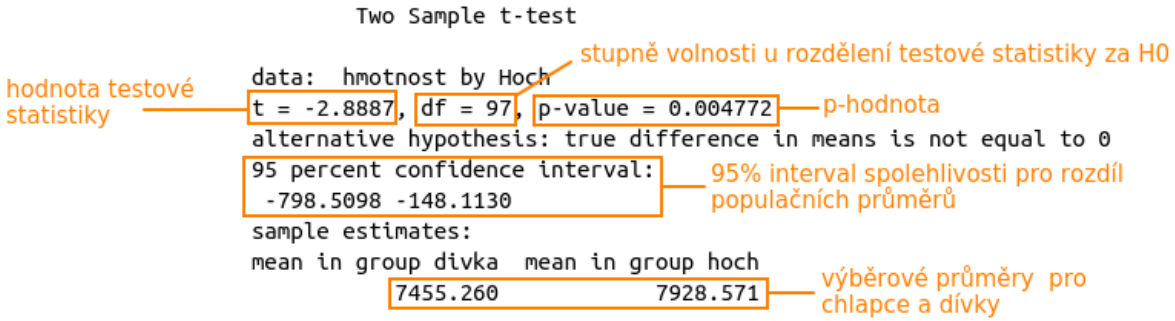


✧ Podporují tato data hypotézu o shodných rozptylech?
P-hodnota vychází 0.5462, takže shodu rozptylů nezamítáme.

✧ Co udává `95 percent confidence interval` ve výstupu?
Je to 95% interval spolehlivosti pro podíl $\frac{\sigma_x^2}{\sigma_y^2}$. Jelikož tento interval obsahuje jedničku (což je teoretická hodnota tohoto podílu za platnosti H_0), tak opět docházíme k závěru, že shodu rozptylů nelze zamítnout.

7) Nyní proveďme dvouvýběrový t-test.

```
t.test(hmotnost ~ Hoch, data=Kojeni, var.equal=TRUE)    # nebo
t.test(KojeniD$hmotnost, KojeniH$hmotnost, var.equal=TRUE)
```



✧ Argumentem `var.equal=TRUE` říkáme, že jsme otestovali shodu rozptylů a došli jsme k závěru, že shodné rozptyly lze předpokládat.

✧ Jaký je závěr (zamítáme/nezamítáme H_0 na požadované hladině)?
Nulovou hypotézu zamítáme, protože p-hodnota je menší než zvolená hladina $\alpha = 0.05$.

✧ Jak byste formulovali svůj závěr bez použití výrazů „zamítáme/nezamítáme H_0“?
„Na 5% hladině jsme prokázali, že střední hmotnost dívek a chlapců v populaci není stejná.“

✧ Jak souvisí výsledek testu s intervalem spolehlivosti, jež je též uveden ve výstupu?
95% interval spolehlivosti (-798.51 ; -148.11) obsahuje s pravděpodobností 0.95 skutečnou hodnotu rozdílu $\mu_x - \mu_y$. Má-li platit nulová hypotéza (podle které je $\mu_x - \mu_y = 0$), měla by nula ležet v tomto intervalu. Ona tam ale neleží, tedy to s velkou pravděpodobností není ta skutečná hodnota rozdílu $\mu_x - \mu_y$. Docházíme tedy ke stejnému závěru jako pomocí p-hodnoty, a to k zamítnutí H_0.

8) Jak byste učinili svůj závěr, kdybyste měli k dispozici pouze hodnotu testové statistiky $T = -2.8887$ a tabulku kvantilů t-rozdělení?

✧ Tabulku kvantilů (kterým se v kontextu testování hypotéz často říká kritické hodnoty) najdete např. na mých stránkách k tomuto cvičení.

✧ Rozhodování provedeme podle standardního rozhodovacího kritéria, které je uvedeno např. v textu *DVT* na str. 2 nahoře. Víme, že $|T| = 2.8887$. V tabulkách potřebujeme najít kvantil t-rozdělení příslušející hladině $\alpha = 0.05$ s 97 stupni volnosti, tedy $qt_{97}(1-\alpha/2) = qt_{97}(0.975)$. V těchto tabulkách je tento kvantil označen jako kritická hodnota $t_{97}(\alpha) = t_{97}(0.05)$, ale v různých tabulkách se značení může lišit. Hodnota přímo pro 97 stupňů volnosti v tabulkách není, najdeme si tedy nejbližší dostupnou hodnotu, a to $t_{100}(0.05) = 1.98$. Jelikož $|T| \geq 1.98$, tj. testová statistika překročila kritickou hodnotu, zamítáme nulovou hypotézu a rovnosti populačních průměrů.

9) O platnosti předpokladů (a volbě statistického postupu) by se správně mělo rozhodovat ještě předtím, než jsou známa data. Předpoklady bychom tedy měli formulovat např. na základě historických dat apod. V případě t-testu je zejména předpoklad shody rozptylů poměrně zbytečný, neboť t-test, který shodu rozptylů nepředpokládá, je v situaci, kdy shoda rozptylů platí, zpravidla jenom nepatrně slabší (tzn. síla testu je menší) než t-test se shodou rozptylů. Obvykle tedy provádíme rovnou *t-test pro nestejné rozptyly*, který se nazývá Welchův test.

```
t.test(hmotnost ~ Hoch, data=Kojeni)   # nebo
t.test(KojeniD$hmotnost, KojeniH$hmotnost)
```

✧ Výstup je (co do uspořádání) shodný jako u základní verze dvouvýběrového t-testu.

✧ O kolik se změnila P-hodnota?
P-hodnota je nyní 0.004818, tj. o $4.6 \cdot 10^{-5}$ větší než u dvouvýběrového t-testu výše. To je opravdu zanedbatelný rozdíl. Tím, že se shodou rozptylů nebudeme obtěžovat a budeme vždy automaticky používat Welchův test, neuděláme velkou chybu, ať už je to s těmi rozptyly jakkoli...

✧ Liší se závěr?
Ne. Opět bychom nulovou hypotézu zamítli.

6 Dvouvýběrový t-test (jednostranná H1)

Jsou chlapci ve 24. týdnu vyšší než dívky?

Pokusme se nyní zjistit, zda jsou chlapci ve 24. týdnu (v průměru) vyšší než dívky.

Označme jako X délku náhodně vybraného chlapce a jako Y délku náhodně vybrané dívky. Předpokládejme, že $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ a $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$, přičemž X a Y jsou vzájemně nezávislé. Za platnosti těchto předpokladů můžeme opět pomocí *dvouvýběrového t-testu* testovat $\mathrm{H}_0 : \mu_x = \mu_y$ proti $\mathrm{H}_1 : \mu_x > \mu_y$. Hladinu testu opět předpokládejte $\alpha = 0.05$.

1) Začněte tím, že spočítáte základní popisné statistiky pro délku v závislosti na pohlaví.

```
tapply(delka,Hoch,summary)
```

2) Pokračujte tím, že si uděláte představu o platnosti testovaných hypotéz pomocí krabičkových grafů délky dle pohlaví.

```
boxplot(delka ~ Hoch)
```

3) Nakreslete graf průměrů s „anténami“ odvozenými ze směrodatných chyb průměrů a potom také s „anténami“ odvozenými z intervalů spolehlivosti pro střední hodnotu.

4) Je smysluplné předpokládat v této situaci, že X a Y jsou nezávislé náhodné veličiny?
Ano. Předpokládáme, že jednotlivé děti svou délku vzájemně nijak neovlivňovaly.

5) Lze předpokládat, že rozdělení délky je jak u chlapců tak u dívek normální?

```
shapiro.test(delka[Hoch=="hoch"])
shapiro.test(delka[Hoch=="divka"])
```

6) Lze předpokládat shodu rozptylů?

```
var.test(delka ~ Hoch, data=Kojeni)    # nebo
var.test(delka[Hoch=="divka"], delka[Hoch=="hoch"])
```

Ne. P-hodnota F-testu vyšla 0.03225, což je menší číslo než hladina testu 0.05. Shodu rozptylů tedy zamítáme, ale moc nám to nevadí, protože můžeme použít *t-test pro nestejné rozptyly* (alias Welchův test).

7) Proveďme dvouvýběrový t-test, nyní s jednostrannou alternativou.

```
t.test(delka ~ Hoch, data=Kojeni, alternative="less")   # nebo
t.test(delka[Hoch=="hoch"], delka[Hoch=="divka"], alternative="greater")
```

✧ **Pozor** na volbu alternativní hypotézy v R, aby odpovídala tomu, co chcete testovat. Při použití ”vlnovkové notace” (první příkaz) je pořadí skupin při výpočtu rozdílu určeno pořadím úrovní příslušného `factor`u (zde Hoch). Pokud jsme tyto úrovně záměrně nějak nepřeuspořádaly (jako jsme to kdysi dělali u veličiny `Vzdelani`), tak budou uspořádané podle abecedy (zde tedy nejdřív `dívka` a pak `hoch`, tedy $\mu_y - \mu_x$). Naše alternativní hypotéza pak požaduje $\mu_y - \mu_x$ `is less than` 0. U druhého příkazu si pořadí skupin určujeme sami tím, v jakém pořadí výběry do příkazu napíšeme. Jako první jsme napsali chlapce a jako druhé dívky, takže tentokrát testujeme rozdíl $\mu_x - \mu_y$. Naše alternativní hypotéze odpovídá tomu, že $\mu_x - \mu_y$ `is greater than` 0. Je to možná méně pohodlná, ale určitě bezpečnější varianta.

✧ Jaký je závěr?
P-hodnota je 0.21, což je více než hladina 0.05, takže nezamítáme nulovou hypotézu, že střední délka chlapců a dívek v populaci je shodná.

✧ Jak byste interpretovali spočtený interval spolehlivosti?
95% interval spolehlviosti je $(-0.5641; \infty)$. Tento interval obsahuje 0, takže nula má šanci být skutečným rozdílem středních hodnot délek. Docházíme tedy ke stejnému závěru jako pomocí p-hodnoty, a to nezamítnout nulovou hypotézu.

✧ Poznámka: Kdybyste si nebyli jistí pořadím kategorií nějakého `factor`u, můžete si ho ověřit příkazem

```
levels(as.factor(Hoch))
```

7 Dvouvýběrový Wilcoxonův test

Souvislost mezi pohlavím dítěte a věkem matky

Nyní se pokusme zjistit, zda lze prokázat souvislost mezi pohlavím dítěte a věkem matky. Bude nás tedy zajímat, zda se střední hodnota věku matek liší v populaci dívek a chlapců.

Označme jako X věk matky náhodně vybraného chlapce a jako Y věk matky náhodně vybrané dívky.

1) Je smysluplné předpokládat v této situaci, že X a Y jsou nezávislé náhodné veličiny?
Ano.

2) Začněte opět tím, že spočítáte základní popisné statistiky pro věk matek v závislosti na pohlaví dítěte.

3) Pokračujte opět tím, že si pomocí krabičkových grafů uděláte představu o souvislostech mezi pohlavím dítěte a věkem matky.

4) Lze předpokládat, že rozdělení věku matek je jak u chlapců tak u dívek normální?

```
shapiro.test(vek.m[Hoch=="hoch"])
shapiro.test(vek.m[Hoch=="divka"])
```

Ne. U chlapců zamítáme, že by věk matek měl normální rozdělení.

5) V předcházejícím kroku jsme zjistili, že u chlapců je rozdělení věku matek prokazatelně (dokonce na hladinách nižších než 1 %) nenormální. V takové situaci můžeme sáhnout k některému z pořadových testů. Ke zjištění, zda věk matky souvisí s pohlavím dítěte, bychom mohli použít *dvouvýběrový Wilcoxonův test* (*Mannův-Whitneyův test*, *Wilcoxon rank sum test*). Testovat budeme H_0 : „rozdělení X je shodné s rozdělením Y“ proti H_1 : „rozdělení X se liší od rozdělení Y“. Vzhledem k tomu, že Wilcoxonův test je citlivý zejména na situace, kdy se rozdělení X a Y liší pouze polohou (např. mediánem), specifikují se testované hypotézy často jako $\mathrm{H}_0 : \mathsf{med}(X) = \mathsf{med}(Y)$ proti $\mathrm{H}_1 : \mathsf{med}(X) \neq \mathsf{med}(Y)$.

Proveďme dvouvýběrový Wilcoxonův test.

```
wilcox.test(vek.m ~ Hoch, data=Kojeni)   # nebo
wilcox.test(vek.m[Hoch=="hoch"], vek.m[Hoch=="divka"])
```

Tato funkce počítá testovou statistiku způsobem popsaným Mannem a Whitneyem v roce 1947. Hodnota $W = 1220.5$ ve výstupu je hodnotou testové statistiky U_y ze vzorce (6) v textu *DVT*.

✧ Jaký je závěr?
P-hodnota vyšla 0.9776. Nezamítáme tedy nulovou hypotézu, že by střední hodnota věku matek byla v populaci dívek i chlapců shodná.

8 Samostatná práce

1) Vhodným způsobem zjistěte, zda lze prokázat závislost **porodní hmotnosti**, resp. **porodní délky** dítěte na jeho **pohlaví**.

✧ Vždy začněte výpočtem vhodných popisných statistik a kreslením vhodných obrázků.

✧ Specifikujte použitý pravděpodobnostní model. (Tj. zaveďte si označení pomocí náhodných veličin X a Y a specifikujte předpokládáné rozdělení.)

✧ Specifikujte testované hypotézy.

✧ Závěr formulujte způsobem, jenž bude srozumitelný též člověku bez hlubokých znalostí statistiky.