

Poslední úprava dokumentu: 3. dubna 2024.

---

## P-hodnota, jednovýběrový a párový t-test, Wilcoxonův a znaménkový test, korelace

---

Dnes budeme pokračovat v testování hypotéz u výběrů z normálního rozdělení a zopakujeme si pojmy, které s tím souvisejí. Dále se dostaneme k tzv. pořadovým (neparametrickým) testům, které použijeme v situaci, kdy předpoklad normálního rozdělení není splněn. K lepšímu pochopení teorie vám může pomoci text *Párové testy a Neparametrické jednovýběrové testy* (dále jako *NJT*), které najdete na mých stránkách. Na závěr se podrobněji podíváme na téma korelace - osvěžte si prosím své znalosti o korelaci buď v poznámkách z přednášky, nebo textu *Korelace* (dále jako *KRL*), který najdete na mých stránkách.

Ještě jednou pro jistotu zdůrazňuji, že populační průměr znaku  $X$  = střední hodnota  $X$ . Označuje se většinou jako  $\mu$  (popř.  $\mu_x$ ) nebo  $E(X)$ . Všechno to znamená totéž, a to průměrnou hodnotu znaku  $X$  v populaci.

### Rozcvička



Spočtete 99% interval spolehlivosti pro střední hodnotu (tj. populační průměr) porodní hmotnosti chlapců. Nezapomeňte ověřit předpoklad normálního rozdělení!

## 1 Úvod

Budeme pokračovat v analýze dat *Kojeni*, s nimiž jsme pracovali na minulém cvičení.

- 1) Načtete si data do RStudia.
- 2) Vytvořte podmnožiny dat obsahující pouze chlapce, resp. pouze dívky. Výsledné datové tabulky nazvěte *KojeniH*, *KojeniD*.

```
KojeniH <- subset(Kojeni, Hoch == "hoch")  
KojeniD <- subset(Kojeni, Hoch == "divka")
```

## 2 Jednovýběrový t-test - oboustranná alternativa (dokončení z minula)

Za náhodnou veličinu  $X$  si označíme hmotnost náhodně vybraného chlapce. Na hladině  $\alpha = 5\%$  otestujte hypotézu  $H_0 : \mu = 8200$  g. Jako alternativní hypotézu uvažujte  $H_1 : \mu \neq 8200$  g.

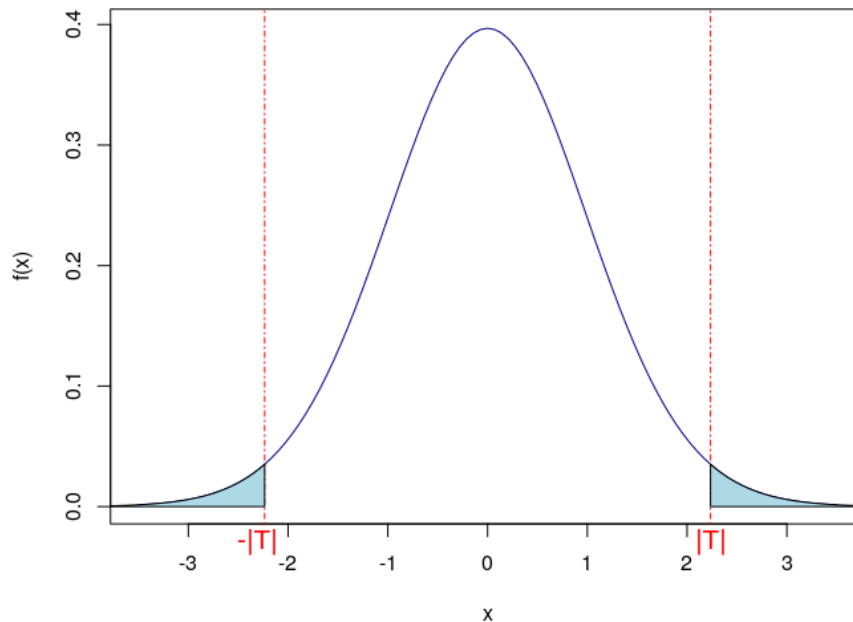
```
t.test(KojeniH$hmotnost, mu=8200)
```

- 1) Pomocí **p-hodnoty** rozhodněte, zda na dané hladině testu zamítáme/nezamítáme  $H_0$ ?  
Na základě výstupu z R můžeme o platnosti nulové hypotézy  $H_0$  rozhodnout třemi způsoby: pomocí hodnoty testové statistiky, pomocí intervalu spolehlivosti a pomocí p-hodnoty.

✧ Pomocí p-hodnoty: Připomeňme, že p-hodnota je nejmenší hladina, na které bychom  $H_0$  zamítli. Zde tedy bychom zamítli pro hladinu 0.03011 a vyšší. Naše zvolená hladina ze zadání příkladu je  $\alpha = 0.05$  a jelikož  $0.05 > 0.03011$ , tak naše rozhodnutí opět je zamítnout  $H_0$ . Jednoduché pravidlo zní: „Je-li  $\alpha > p$ -hodnota  $\Rightarrow$  zamítáme  $H_0$ .“

Ručně bychom p-hodnotu spočítali takto:

Z definice p-hodnoty (text *ÚTH*, str. 4-5) je jasné, že p-hodnota je součtem ploch vymezených vypočtenými hodnotami testové statistiky (viz obrázek níže). Tyto modré plochy odpovídají



pravděpodobnostem  $P(\text{testová statistika} \leq -|T|)$  a  $P(\text{testová statistika} \geq |T|)$ . V R tyto pravděpodobnosti počítáme pomocí distribuční funkce t-rozdělení:

```
(P.vlevo <- pt(-abs(T), df=n-1))      # = P(testová statistika <= -|T|)
(P.vpravo <- 1 - pt(abs(T), df=n-1))  # = P(testová statistika >= |T|)
(P <- P.vlevo + P.vpravo)             # p-hodnota je součtem obou ploch
```

✧ Jak je vidět, obě modré plochy jsou stejné, a tak p-hodnotu lze spočítat najednou:

```
(P <- 2*pt(-abs(T), df=n-1))
```

Ověřte si, že nám vyšla stejná hodnota jako ve výstupu procedury `t.test`.

- 2) Jak byste formulovali svůj závěr bez použití výrazů „zamítáme/nezamítáme  $H_0$ “?  
„Data neprokázala, že by střední hmotnost chlapců ve 24. týdnu byla odlišná od 8 100 g.“  
(Pokud tam nějaká odlišnost je, tak není statisticky významná. Na její případné prokázání bychom potřebovali více dat.)
- 3) Jaká jiná hodnota  $\mu$  při nulové hypotéze by vedla k opačnému závěru?  
No každá, která neleží v našem intervalu spolehlivosti. Tak třeba 9 000 g.
- 4) Je potřeba znovu počítat, jestliže bychom požadovali přísnější 1% hladinu významnosti, resp. benevolentnější 10% hladinu významnosti?  
Ne. Stačí příslušnou hladinu porovnat s vypočtenou p-hodnotou 0.08226.

### 3 Testování předpokladu normálního rozdělení: Shapiro-Wilkův test

Ověření předpokladu normálního rozdělení našeho výběru jsme doposud prováděli pomocí QQ diagramu. Tento předpoklad lze ale též formálně otestovat pomocí Shapiro-Wilkova testu.

```
shapiro.test(hmotnost)
```

### 4 Jednovýběrový t-test - jednostranná alternativa

- 1) Pokusme se prokázat, že populační hmotnost chlapce ve 24. týdnu je **menší** než 8 100 g. Budeme požadovat, aby k chybnému prokázání tohoto tvrzení došlo s pravděpodobností nejvýše 0.05.
- 2) Za  $X$  označíme hmotnost náhodně vybraného chlapce a budeme předpokládat, že  $X \sim \mathcal{N}(\mu, \sigma^2)$ , potom můžeme opět pomocí *jednovýběrového t-testu* testovat  $H_0 : \mu = 8\,100$  g proti  $H_1 : \mu < 8\,100$  g. Testujeme na hladině 5 %, tj.  $\alpha = 0.05$ .  
✧ **Důležité připomenutí.** Znaménko (tj.  $>$ ,  $<$ ,  $\neq$ ) v alternativní hypotéze je nutno zvolit podle toho, co nás zajímá a v každém případě ještě před tím než vidíme data! Jinými slovy, volba alternativní hypotézy podle toho, „jak to vyšlo v datech“ je hrubou chybou.
- 3) Normalitu už jsme ověřovali před chvílí, tak to můžeme přeskočit.
- 4) Provedeme jednovýběrový t-test, kde do argumentu „alternative=“ musíme specifikovat tvar alternativní hypotézy, zde „menší než 8 100 g“, tj. `less`.

```
t.test(KojeniH$hmotnost, mu=8100, alternative="less")
```

- 5) Jaký je závěr (zamítáme/nezamítáme  $H_0$  na požadované hladině)?  
P-hodnota vychází 0.08226, tj. je větší než 0.05, a proto nezamítáme  $H_0$ . Stejný závěr bychom učinili i na základě příslušného intervalu<sup>1</sup> spolehlivosti  $(-\infty; 8132.261)$ , který obsahuje hodnotu 8 100 g (ta tedy má šanci být skutečným populačním průměrem hmotnosti).
- 6) Jak byste formulovali svůj závěr bez použití výrazů „zamítáme/nezamítáme  $H_0$ “?  
„Data neprokázala, že by střední hmotnost chlapců ve 24. týdnu byla odlišná od 8 100 g.“ (Pokud tam nějaká odlišnost je, tak není statisticky významná. Na její případné prokázání bychom potřebovali více dat.)
- 7) Jaká jiná hodnota  $\mu$  při nulové hypotéze by vedla k opačnému závěru?  
No každá, která neleží v našem intervalu spolehlivosti. Tak třeba 9 000 g.
- 8) Je potřeba znovu počítat, jestliže bychom požadovali přísnější 1% hladinu významnosti, resp. benevolentnější 10% hladinu významnosti?  
Ne. Stačí příslušnou hladinu porovnat s vypočtenou p-hodnotou 0.08226.
- 9) Ruční výpočet p-hodnoty by v tomto případě vypadal následovně:

```
pt(-1.4116, df=48)
```

kde  $-1.4116$  je vypočtená hodnota t-statistiky.

---

<sup>1</sup>Inf ve výstupu je zkratka pro Infinity, tj. nekonečno.

## 5 Párový t-test (= převlečený jednovýběrový t-test)

Pokusíme se zjistit, zda jsou otcové v průměru o více jak 14 cm vyšší než matky.

Označme jako  $X$  výšku náhodně vybraného otce a jako  $Y$  výšku jeho partnerky. Data lze tedy považovat za realizaci náhodného výběru z dvourozměrného rozdělení náhodného vektoru  $\begin{pmatrix} X \\ Y \end{pmatrix}$ .

Jak víme, k zodpovězení naší otázky stačí zabývat se náhodnou veličinou  $Z = X - Y$ , kde  $Z$  reprezentuje výškový rozdíl otce a matky náhodně vybraného dítěte. Testujeme pak  $H_0 : E(Z) = 14$  cm proti  $H_1 : E(Z) > 14$  cm, kde  $E(Z)$  značí střední hodnotu veličiny  $Z$ , čili její populační průměr. Jestliže lze navíc předpokládat normální rozdělení náhodné veličiny  $Z$ , lze k testu těchto hypotéz použít nám dobře známý jednovýběrový t-test (aplikovaný na výškové rozdíly). Hladinu testu uvažujte 5 %.

- 1) Připravte novou proměnnou `vyskaDif`, jež bude udávat výškový rozdíl mezi otcem a jeho partnerkou.

```
Kojeni$vyskaDif <- Kojeni$vyska.o - Kojeni$vyska.m
```

- 2) Spočtete základní popisné statistiky pro výškové rozdíly mezi otcem a matkou.

```
summary(Kojeni$vyskaDif) # charakteristiky polohy  
sd(Kojeni$vyskaDif)     # charakteristika variability
```

- 3) Pomocí obrázků i formálním testem zjistěte, zda je smysluplné předpokládat normální rozdělení pro výškové rozdíly mezi otcem a matkou.

```
hist(Kojeni$vyskaDif)           # histogram  
qqnorm(Kojeni$vyskaDif)        # normalni diagram  
qqline(Kojeni$vyskaDif)  
shapiro.test(Kojeni$vyskaDif) # Shapiro-Wilkuv test normality
```

Normální rozdělení zde můžeme předpokládat (p-hodnota Shapiro-Wilkova testu je 0.4836).

- 4) Pomocí párového t-testu (pro výšky otců a matek), resp. jednovýběrového t-testu pro výškové rozdíly, otestujte  $H_0 : E(Z) = 14$  cm proti  $H_1 : E(Z) > 14$  cm.

```
t.test(Kojeni$vyskaDif, mu=14, alternative="greater") # nebo:  
t.test(Kojeni$vyska.o, Kojeni$vyska.m, mu=14, alternative="greater", paired=TRUE)
```

✧ Jaký je závěr (zamítáme/nezamítáme  $H_0$  na požadované hladině)?

P-hodnota vyšla 0.9802, což je víc než zvolená 5% hladina. Nezamítáme tedy  $H_0$ . Na základě našich dat nelze zamítnout hypotézu, že rozdíl populačních průměrů výšek je 14 cm.

✧ Jak byste formulovali svůj závěr bez použití výrazů „zamítáme/nezamítáme  $H_0$ “?

Data neprokázala, že by rozdíl populačních průměrů výšek byl různý od 14 cm. (Pokud ten rozdíl není 14 cm, tak ale ne statisticky významně. Na detekci případné odchylky bychom potřebovali více dat.)

✧ Jak souvisí výsledek testu s intervalem spolehlivosti, jež je též uveden ve výstupu?

Je to 95% interval spolehlivosti pro  $E(Z)$ , čili je to interval, který s pravděpodobností 0.95 obsahuje skutečnou hodnotu rozdílu populačních průměrů výšek otce a matky. Jelikož hodnota 14 leží v tomto intervalu, má šanci být tím skutečným rozdílem výšek. Docházíme tedy ke stejnému závěru, a to nezamítnutí  $H_0$ .

Výsledný interval spolehlivosti má tvar  $(10.93, \infty)$ , je tedy jednostranný. Je to proto, že jsme zvolili jednostrannou alternativní hypotézu.

## 6 Jednovýběrové pořadové testy (Wilcoxonův a znaménkový)

Pokusíme se zjistit, zda jsou otcové typicky o více jak 2 roky starší než matky.

Označme jako  $X$  věk náhodně vybraného otce a jako  $Y$  věk jeho partnerky. Data lze tedy opět považovat za realizaci náhodného výběru z dvourozměrného rozdělení náhodného vektoru  $\begin{pmatrix} X \\ Y \end{pmatrix}$ .

K zodpovězení naší otázky však opět stačí zabývat se náhodnou veličinou  $Z = X - Y$ , kde  $Z$  reprezentuje věkový rozdíl otce a matky náhodně vybraného dítěte. Chceme testovat hypotézu  $H_0 : E(Z) = 2$  roky proti  $H_1 : E(Z) > 2$  roky, kde  $E(Z)$  značí střední hodnotu veličiny  $Z$ . Hladinu testu uvažujeme opět 5 %.

- 1) Připravte novou proměnnou `vekDif`, jež bude udávat věkový rozdíl mezi otcem a jeho partnerkou.

```
Kojeni$vekDif <- Kojeni$vek.o - Kojeni$vek.m
```



- 2) Spočtete základní popisné statistiky pro věkové rozdíly mezi otcem a matkou.



- 3) Obrázky i formálním testem zjistíte, zda je smysluplné předpokládat normální rozdělení pro věkové rozdíly mezi otcem a matkou. Pro posouzení normality nakreslete mimo jiné též histogram.

✧ Patrně jste zjistili, že rozdělení věkových rozdílů mezi otcem a matkou není normální, ale je značně zešikmené. K posouzení věkové odlišnosti mezi otcem a matkou musíme tedy použít nějaký neparametrický/pořadový test.

### 6.1 Jednovýběrový Wilcoxonův test

Pokud chceme použít neparametrický test, musíme přeformulovat testované hypotézy pomocí populačního mediánu. Budeme nyní testovat nulovou hypotézu

$H_0$  : populační medián  $Z$  je roven 2 roky,      proti alternativě:

$H_1$  : populační medián  $Z$  není roven 2 roky.

- 4) Jednovýběrový Wilcoxonův test zjišťující, zda jsou otcové typicky o více jak 2 roky starší než matky provedeme pomocí příkazu

```
wilcox.test(Kojeni$vekDif, mu=2, alternative="greater")
```

```
Wilcoxon signed rank test with continuity correction
```

```
data: Kojeni$vekDif
V = 2690, p-value = 0.001054
alternative hypothesis: true location is greater than 2
```

Hodnota  $V$  je testová statistika Wilcoxonova testu (před vynormováním) a udává součet pořadí od těch hodnot  $Z_i$ , které byly větší než 2 (po vyloučení hodnot rovných 2). V textu *NJT* na str. 1 je tato statistika označena jako  $W$ .

✧ Jaký je závěr (zamítáme/nezamítáme  $H_0$  na požadované hladině)?

P-hodnota vyšla 0.001, což je výrazně menší hodnota než hladina testu 0.05. Zamítáme tedy nulovou hypotézu.

✧ Jak byste formulovali svůj závěr bez použití výrazů „zamítáme/nezamítáme  $H_0$ “?

Na 5% hladině jsme prokázali, že rozdíl věku otců a matek v populaci je větší než 2 roky.

- 5) Nicméně ani použití Wilcoxonova testu nebylo v této situaci zrovna nejsprávnější. Jedno-výběrový Wilcoxonův test sice nepředpokládá normální rozdělení, avšak stále předpokládá symetrické rozdělení (což nebyl případ věkových rozdílů mezi otcem a matkou, jak jsme viděli na histogramu).

## 6.2 Znaménkový test

Nejobecnějším (avšak zároveň nejslabším) testem pro tuto situaci je test znaménkový.

- 1) K provedení znaménkového testu je nejprve potřeba zjistit, u kolika párů (otec-matka) nastala situace, kdy je otec o více jak 2 roky starší než matka:

```
U <- sum(Kojeni$vekDif > 2)      # v textu NJT oznaceno jako U
```

- 2) Dále je potřeba zjistit, u kolika párů je věkový rozdíl odlišný od 2 let, neboť páry s věkovým rozdílem právě 2 roky jsou pro znaménkový test bezcenné (jako by v datech nebyly).

```
n2 <- sum(Kojeni$vekDif != 2)    # n* z textu NJT
```

- 3) Znaménkový test (asymptotický, s Yatesovými korekcemi) nyní provedeme takto:

```
prop.test(U, n2, alternative="greater")
```

```
1-sample proportions test with continuity correction

data:  U out of n2, null probability 0.5
X-squared = 1.375, df = 1, p-value = 0.1205
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.4749392 1.0000000
sample estimates:
          p
0.5681818
```

- ✧ X-squared ve výstupu je druhá mocnina statistiky  $Z_2$  z textu *NJT*, vzorec (3).
- ✧ Význam ostatních hodnot ve výstupu si ozřejmíme později, v rámci cvičení o kategoriálních datech.
- ✧ Jaký je závěr (zamítáme/nezamítáme  $H_0$  na požadované hladině)?  
P-hodnota je 0.1205, tedy je větší než požadovaná hladina 0.05. Nulovou hypotézu tedy nezamítáme.
- ✧ Jak byste formulovali svůj závěr bez použití výrazů „zamítáme/nezamítáme  $H_0$ “?  
Data neprokázala, že by se populační průměr věku otců a matek neodlišoval o 2 roky.

## 7 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. Necht'  $\rho$  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*).

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

případně obyčejný bodový graf:

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

- 2) Spočtíme 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  $H_0: \rho = 0$  proti  $H_1: \rho \neq 0$ . 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 nezávislosti (viz text *KRL*, str. 1, dole) provedeme příkazem:

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

```
Pearson's product-moment correlation
data: Kojeni$vyyska.m and Kojeni$vyyska.o
t = 2.0678, df = 97, p-value = 0.04132
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.008402418 0.387180104
sample estimates:
cor
0.2054734
```

počet stupňů volnosti t-rozdělení, které má testová statistika za  $H_0$  (n-2)

hodnota testové statistiky

p-hodnota

95% interval spolehlivosti pro populační korelační koeficient  $\rho$

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  $\rho$  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?

## 8 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 p-hodnota
alternative hypothesis: true rho is not equal to 0
sample estimates:
 rho
0.7302145 Spearmanův korelační koeficient

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

```

Chybavá hláška upozorňující na to, že některé dvojice se v datech vyskytují vícekrát, což způsobuje problémy při výpočtu přesné p-hodnoty.

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

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

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

✧ Jaký je závěr?

P-hodnota vyšla menší než  $2.2 \times 10^{-16}$ , tedy velmi malá. Zamítáme tedy nulovou hypotézu. Na 5% hladině jsme prokázali, že věk matky a otce jsou silně závislé veličiny.

## 8.1 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  $\pm 1$ ), když jedna z veličin je lineární funkcí druhé (tj. např.  $Y = a + bX$ ). Aby Spearmanův koeficient nabýval hodnoty  $\pm 1$ , 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 spočteme 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)
```

- ✧ Spočteme 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)
```

## 9 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) Spočteme 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.



## 10 Konec práce

Než zavřete všechna okna, nezapomeňte si uložit poslední změny ve skriptovém souboru:

**File** ➔ **Save**

nebo klávesovou skratkou **Ctrl-s**.