## Cvičení z biostatistiky 03

Poslední úprava dokumentu: 28. února 2024.

---

## Popisná statistika pro dvě proměnné

---

Nemáte-li, načtěte data Deti23 do RStudia:

```
load("data/Deti23.RData")                          # jsou-li data ve formatu .RData
Deti23 <- read.csv2("data/Deti23.csv", header=TRUE)   # jsou-li data ve formatu .csv
```

a zajistěte si přímý přístup k jednotlivým proměnným datového souboru Deti23 pomocí příkazu

```
attach(Deti23)
```

# 1 Dvě kvantitativní proměnné

## 1.1 Grafické znázornění

**1)** Bodový graf váhy proti délce.

```
plot(delka, vaha)
```

✧ Úpravou příkazu lze obrázek dále zkrášlovat:

```
plot(delka, vaha, xlab="Delka (cm)", ylab="Vaha (kg)", col="seagreen", pch=13)
```

✧ V argumentu pch si můžete zkusit změnit třináctku za libovolné číslo od 0 do 25.

✧ V argumentu col můžete použít i jinou barvu. Seznam předdefinovaných barev se objeví, spustíte-li příkaz *colors()*

**2)** Dále obrázek vylepšeme přidáním nadpisu:

```
plot(delka, vaha, main="Závislost váhy na délce",
   xlab="Delka (cm)", ylab="Vaha (kg)", col="seagreen", pch=13)
```

**3)** V dalším kroku si pomocí barev a symbolů rozlišíme chlapce a dívky. Připomínám, že pohlaví udávají proměnná hoch (1 pro chlapce, 0 pro dívku), respektive proměnná Pohlavi ("M" pro chlapce, "F" pro dívky).

✧ Na naší úrovni je asi nejprůhlednější následující konstrukce:

```
divky <- which(Pohlavi=="F")   # kde jsou v datech dívky
hosi <- which(Pohlavi=="M")   # kde jsou v datech hoši
plot(delka[divky], vaha[divky], col="purple", pch=16,
   main="Závislost váhy na délce", xlab="Delka (cm)", ylab="Vaha (kg)")
points(delka[hosi], vaha[hosi], col="darkblue", pch=17)
```

Příkaz `plot` otevírá nové grafické okno a vykresluje do něj (v našem případě data pro dívky). Příkaz `points` přidává body do již existujícícho grafu (v našem případě body pro chlapce).

✧ Následující dvě konstrukce jsou méně průhledné, ale oceníte je u proměnných s velkým počtem kategorií, kdy by bylo postupné vykreslování pomocí `points` zdlouhavé.

✧ barevné odlišení s využitím číselné proměnné hoch

```
barvy <- c("purple", "darkblue")
symboly <- c(16, 17)
plot(delka, vaha, col=barvy[hoch+1], pch=symboly[hoch+1],
   main="Závislost váhy na délce", xlab="Delka (cm)", ylab="Vaha (kg)")
```

✧ barevné odlišení s využitím faktorové proměnné Pohlavi

```
barvy <- c(F="purple", M="darkblue")
symboly <- c(F=16, M=17)
plot(delka, vaha, col=barvy[Pohlavi], pch=symboly[Pohlavi],
   main="Závislost váhy na délce", xlab="Delka (cm)", ylab="Vaha (kg)")
```

Příkazy z bodu 5) si rozhodně nemusíte pamatovat! Stačí pouze vědět, že existují a v případě potřeby si je dohledáte.

**4)** Úplně nakonec přidáme do obrázku též legendu.

```
legend("topleft", legend=c("Divka", "Chlapec"), col=c("purple","darkblue"),
       pch=c(16,17))
```

**5)** Nakreslete vhodný obrázek pro popis závislosti váhy dítěte na věku matky a interpretujte ho.

✧ Nejspíš vás napadne použít příkaz

```
plot(vekMatky, vaha)
```

✧ Úplně stejný obrázek dostanete také tak, použijete-li příkaz `plot` s „vlnkovou“ notací, tj.

```
plot(vaha ~ vekMatky)
```

✧ Později budeme zkoumat závislost dvou kvantitativních veličin pomocí lineární regrese.

### 1.2 Kovariance a korelace

**6)** **Výběrová kovariance** mezi váhou a délkou dítěte.

✧ Ze skriptového okna spusťte příkaz:

```
cov(vaha, delka)
```

**7)** **Korelační koeficient** mezi váhou a délkou.

```
cor(vaha, delka)
```

**8)** Vypočtěte korelaci mezi věkem matky a otce a vykreslete tyto veličiny do bodového grafu.

**9)** Spočtěte korelaci mezi váhou dítěte a věkem matky a interpretujte výsledek.

## 2 Souvislost mezi bodovým grafem a korelačním koeficientem

Podívejme se nyní, jak spolu souvisí mrak bodů v bodovém grafu a korelační koeficient. Na stránce `https://rpsychologist.com/correlation/` manipulujte s hodnotou korelačního koeficientu v okénku nad grafem (po nastavení konkrétní hodnoty musíte stisknout tlačítko „play“ vedle okénka) a dívejte se, co se stane s mrakem bodů na obrázku.

## 3 Dvě kvalitativní proměnné

Zjistěme, jak spolu souvisí pohlaví dítěte (proměnná Pohlavi) a nemocnost (proměnná fhcd3).

**1)** Spočtěme kontingenční tabulku pro Pohlavi a fhcd3 a s ní související (podmíněné) relativní četnosti (podmíněno pohlavím).

```
table(Pohlavi, fhcd3)
prop.table(table(Pohlavi, fhcd3), margin=1) * 100
```

✧ Zamyslete se, jak by (ideálně – v populaci, v datech při vyšším rozsahu výběru) měly vypadat podmíněné relativní četnosti, jestliže nemocnost nesouvisí s pohlavím.

✧ Zdá se vám, že nemocnost spíše souvisí, či spíše nesouvisí s pohlavím dítěte?

✧ Svoje závěry v žádném případě nepovažujte za definitivní. Nezapomeňte, že máme k dispozici pouhých 23 pozorování.

✧ Později budeme zkoumat závislost dvou kvalitativních veličin pomocí $\chi^2$ (chí kvadrát) testu nezávislosti.

**2)** Graficky lze (pouze příkazy) spočtené podmíněné relativní četnosti znázornit následovně (povšimněte si vlnkové notace).

```
plot(fhcd3 ~ Pohlavi, col=terrain.colors(3), xlab="Pohlavi", ylab="Relativni cetnost")
```

**3)** Přidejme si do tabulky sloupcová procenta. Jak se tato čísla interpretují?

```
prop.table(table(Pohlavi, fhcd3), margin=2) * 100
```

A co celková procenta?

```
prop.table(table(Pohlavi, fhcd3)) * 100
```

**4)** Pomocí vhodných popisných statistik a obrázků vyšetřete závislosti mezi pohlavím dítěte (proměnná Pohlavi) a jeho zjednodušeným pořadím (*prvorozenec/má sourozence*, proměnná fporadi2).

## 4 Kvantitativní a kvalitativní proměnná

V datech jste si vytvořili (kvalitativní) proměnnou (`factor`) fhcd3, jež udává, jak často (*nikdy/jednou/opakovaně*) prodělalo dítě během prvního roku života onemocnění horních cest dýchacích. Podívejme se nyní, jak spolu souvisí váha dítěte v jednom roce (kvantitativní proměnná vaha) a nemocnost (kvalitativní proměnná fhcd3).

**1)** Spočtěme popisné statistiky pro váhu v závislosti na nemocnosti.

```
tapply(vaha, fhcd3, summary)
tapply(vaha, fhcd3, sd)          # smer. odchylky
tapply(vaha, fhcd3, var)         # rozptyly
tapply(vaha, fhcd3, length)      # počty pozorování
```

✧ V tomto případě je věcná interpretace poněkud obtížná z důvodu dosti malého počtu hodnot.

**2)** Nakresleme krabičkové grafy váhy podmíněné nemocností (jedna krabička pro každou skupinu dle nemocnosti). Povšimněte si „vlnkové“ notace!

```
plot(vaha ~ fhcd3, col="seagreen", xlab="Nemocnost", ylab="Vaha (kg)")
```

✧ Porovnejte obrázek s vypočtenými mediány, kvartily a variabilitou (charakterizovanou pomocí kvartilového rozpětí).

✧ Domníváte se, že se váha dětí liší pro různé úrovně nemocnosti?

✧ Později budeme zkoumat odlišnost populačních průměrů v různých skupinách pomocí **vícevýběrových testů** (analýza rozptylu, Kruskalův-Wallisův test).

### Nepovinné, ale užitečné - další grafy

**3)** Dalším užitečným obrázkem je znázornění průměrů se směrodatnými odchylkami (plot of means with error bars). Bohužel jeho vykreslení bez specializovaných knihoven je poněkud komplikovanější. Proto použijeme předem vytvořený skript `plot_of_means.R`, který najdete na mých stránkách nebo v Google Classroom. Skript si uložte do své pracovní složky (`biostat`) a ze skriptového okna postupně spusťte příkazy:

```
source("plot_of_means.R")
plot_of_means(vaha, fhcd3, "sd")
```

✧ Chybové úsečky v tomto případě představují směrodatné odchylky.

✧ Porovnejte obrázek s vypočtenými průměry a variabilitou (charakterizovanou pomocí směrodatných odchylek).

✧ Do obrázku lze přidat nadpis a popisky os:

```
plot of means(vaha, fhcd3, "sd", xlab = "vaha", ylab = "nemocnost",
main = "Graf prumeru")
```

✧ Místo směrodatných odchylek lze do grafu vynést směrodatné chyby jednotlivých průměrů:

```
plot_of_means(vaha, fhcd3, "se")
```

✧ V neposlední řadě mohou chybové úsečky představovat i intervaly spolehlivosti pro populační průměry, jejich význam si však vysvětlíme až na příštím cvičení.

**4)** K vykreslení průměrů s intervaly spolehlivosti lze také použít funkci `plotmeans` z knihovny `gplots`.

```
install.packages("gplots")      # instalace knihovny
library(gplots)                 # otevreni knihovny
plotmeans(vaha~fhcd3,p=0.95)    # graf prumeru s 95% intervaly spolehlivosti
```

### Samostatná práce

Pomocí vhodných popisných statistik a obrázků vyšetřete závislosti mezi:

**1)** Věkem matky (vekMatky) a nemocností (fhcd3);

**2)** Váhou dítěte v jednom roce (vaha) a pořadím dítěte (*prvorozenec/má sourozence*, na minulém cvičení vytvořená proměnná fporadi2).

✧ Z obrázků a vypočtených čísel si utvořte názor, zda spíše existuje, či spíše neexistuje nějaká forma závislosti mezi zkoumanými znaky.

✧ Později budeme zkoumat odlišnost populačních průměrů u dvou skupin pomocí dvouvýběrových testů (t-test, Wilcoxonův test).

## 5 Vytvoření podmnožiny dat

Někdy je potřeba zpracovávat pouze podmnožinu dat, jež splňuje nějakou podmínku (např. zajímají nás pouze dívky, nebo pouze děti vyšší než nějaký limit apod.) Bude proto potřeba umět vybrat si z dat podmnožinu splňující určitou podmínku a poté tuto podmnožinu uložit.

**1)** Zjistíme, pro které děti je otec starší než matka:

```
which(vekMatky < vekOtce)
```

**2)** Můžeme též vypsat hodnoty všech veličin z dat, u kterých je otec starší než matka:

```
subset(Deti23, vekMatky < vekOtce)
```

**3)** V případě, že chceme podmnožinu původních dat ukládat a dále s ní pracovat, doporučuji odpojit přístup k proměnným původních dat (vyhnete se tak možným nedorozumněním plynoucím ze shodných názvů proměnných ve dvou datech – původních a podmnožiny).

```
detach(Deti23)
```

**4)** Řekněme, že dále budeme chtít pracovat pouze s dětmi, u nichž je otec starší než matka. Vytvořenou podmnožinu si můžeme uložit do datové tabulky DetiOsM.

```
DetiOsM <- subset(Deti23, vekOtce > vekMatky)
```

**5)** Tuto podmnožinu si dále můžeme uložit (ale není to nutné, nebudeme ji už dále potřebovat) pomocí známého příkazu:

```
save(DetiOsM, file = "data/DetiOsM.RData")
```

**6)** Sami si můžete zkusit vytvořit nebo se alespoň podívat (nemusíte výsledky nikam ukládat) na následující podmnožiny:

(a) Děti, u kterých je otec jinak starý než matka.
(b) Děti, u kterých je otec o alespoň 5 let starší než matka.
(c) Děti, u kterých se věk rodičů liší o právě jeden rok.
(d) Dívky.
(e) Dívky, které mají otce staršího než matku.
(f) Děti, které mají délku nejvýše 74 cm nebo nejméně 79 cm.

**Nápověda:** Ke specifikaci jednotlivých podmnožin si vybírejte z následujících logických výrazů (Subset expression):

- `Pohlavi == "F"`
- `vekOtce - vekMatky > 4`
- `hoch != 1`
- `hoch == 0`
- `Pohlavi != "M"`
- `vekOtce != vekMatky`
- `delka <= 74 | delka >= 79`
- `vekOtce - vekMatky >= 5`
- `!(delka > 74 & delka < 79)`
- `Pohlavi == "F"& vekOtce > vekMatky`
- `abs(vekOtce - vekMatky) == 1`

**Poznámka:** Jestliže s vytvořenou podmnožinou neplánujete dále pracovat (tj. jenom vás zajímá, jak vypadá), není potřeba provádět dokola `detach(Deti23)`, `attach(Deti23)`.

# 6 Konec práce

Než zavřete všechna okna, nezapomeňte si uložit poslední změny ve skriptovém souboru:
File ➥ Save
nebo klávesovou skratkou `Ctrl-s`.