Cvičení z biostatistiky 03

Poslední úprava dokumentu: 6. března 2024.

Vliv posunutí a změny měřítka, plot of means, šikmost a špičatost, tvorba podmnožiny dat, pravděpodobnost, výpočet EX a $varX$

---

1 Úvod

1) Spusťte RStudio z nabídky a nastavte si pracovní adresář:

✧ buď v Menu nahoře:

Session ➥ **Set Working Directory** ➥ **Choose Directory ...**

a vyberte příslušnou složku, ve které chcete pracovat.

✧ nebo pomocí příkazu (kde do uvozovek vyplníte cestu k danému adresáři)

```
setwd("path/to/folder")
```

2) Vytvořte si nový skriptový soubor a uložte si jej pomocí kliknutí na ikonu diskety, nebo klávesovou zkratkou `Ctrl+s`, nebo v Menu nahoře:

File ➥ **Save as...**

2 Nový datový soubor

Datový soubor Kojeni.csv obsahuje údaje (hodnoty oddělené středníky) o 99 matkách a jejich dětech (data získala Mgr. P. Hajná při přípravě své diplomové práce). Najdeme zde následující proměnné:

`trvani`	délka kojení v prvních 24 týdnech po porodu (týdny);
`pocet.deti`	o kolikáté dítě matky jde;
`Vzdelani`	vzdělání matky (*základní/maturita/VŠ*);
`por.hmotnost`	porodní hmotnost dítěte (g);
`por.delka`	porodní délka dítětete (cm);
`Prs`	dítě bylo do půl hodiny po porodu přiloženo k prsu (0=*ne*, 1=*ano*);
`hmotnost`	hmotnost dítěte ve 24. týdnu po porodu (g);
`delka`	délka dítěte ve 24. týdnu po porodu (cm);
`vyska.m`	výška matky (cm);
`vyska.o`	výška otce (cm);
`Otec`	otec u porodu (0=*ne*, 1=*ano*);
`vek.m`	věk matky (roky);
`vek.o`	věk otce (roky);
`Dudlik`	dítě mělo dudlík (0=*ne*, 1=*ano*);
`Plan`	dítě bylo podle matky plánované (0=*ne*, 1=*ano*);
`Porodnice`	rozlišení dvou porodnic, v nichž byla pořizována data (1=*Praha*, 2=*okresní město*);
`Koj24`	matka kojila ještě ve 24. týdnu po porodu (0=*ne*, 1=*ano*);
`HochL`	dítě je hoch (FALSE=*ne*, TRUE=*ano*);
`hoch`	dítě je hoch (0=*ne*, 1=*ano*);
`Hoch`	pohlaví dítěte (*dívka/hoch*).

1) Tento datový soubor si uložte do složky, kterou jste si v úvodu nastavili v R jako pracovní adresář.

2) Načtěte data do R, datovou tabulku nazvěte Kojeni. Nejlepší je použít příkaz:

```
Kojeni <- read.csv2("Kojeni.csv", header=TRUE)
```

3) Zatím si nezajišťujte přímý přístup k jednotlivým proměnným v datech, tj. nespouštějte příkaz `attach(Kojeni)`.

3 Vliv posunutí a měřítka

Vytvořte veličiny

$$\text{dobaPlnolet} = \text{vek.m} - 18$$
$$\text{hmotnost.kg} = \text{hmotnost}/1000$$

✧ **Vliv posunutí:** Jaký je vztah mezi popisnými statistikami pro proměnné vek.m a dobaPlnolet?

✧ **Vliv změny měřítka:** Jaký je vztah mezi popisnými statistikami pro proměnné hmotnost a hmotnost.kg?

4 Graf průměrů - Plot of means

1) Úprava veličiny Ujistěte se, že veličina `Vzdelani` je chápána jako faktor:

```
class(Kojeni$Vzdelani)   # podivame se, zda je chapana jako faktor, pokud ne:
Kojeni$Vzdelani <- as.factor(Kojeni$Vzdelani)
```

Kategorie každého faktoru, jsou ve vnitřních chápání R vždy seřazené podle abecedy, zde tedy: *maturita, VŠ, základní*. To zde není přirozené a změníme tedy toto uspořádání:

```
Kojeni <- transform(Kojeni,
      Vzdelani = factor(Vzdelani, levels=c("zakladni", "maturita", "VS")))
```

Nyní si upravená data uložte a můžete si zajistit přímý přístup k jednotlivým proměnným.

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

2) Pro připomenutí nakreslete krabičkové grafy hmotnost podmíněné vzděláním matky (jedna krabička pro každou skupinu dle vzdělání). Povšimněte si „vlnkové“ notace!

```
plot(hmotnost ~ Vzdelani, xlab="Vzdělání matky", ylab="Hmotnost (g)")
```

✧ Domníváte se, že se váha dětí liší dle vzdělání matky?

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

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 na disku `V`. Skript si uložte do své pracovní složky (`biostat`) a ze skriptového okna postupně spusťte příkazy:

```
source("plot_of_means.R")
plot_of_means(hmotnost, Vzdelani, "sd")
```

✧ Chybové úsečky v tomto případě představují směrodatné odchylky.

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

```
tapply(hmotnost, Vzdelani, mean)
tapply(hmotnost, Vzdelani, sd)
```

✧ Do obrázku lze přidat nadpis a popisky os:

```
plot_of_means(hmotnost, Vzdelani, "sd", xlab = "vaha", ylab = "nemocnost",
              main = "Graf prumeru")
```

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

```
plot_of_means(hmotnost, Vzdelani, "se")
```

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

```
plot_of_means(hmotnost, Vzdelani, "ci")
```

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

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

5 Šikmost a špičatost aneb tvar rozdělení kvantitativního znaku

Histogram

Histogram graficky znázorňuje četnosti jednotlivých hodnot v datech. Reálnou osu rozdělí na malé intervaly vhodné délky a zaznamenává, kolik hodnot z dat se nachází v jakém intervalu. Tuto četnost pak znázorní výškou příslušného sloupce.

1) Nakresleme histogram věku matek

```
hist(vek.m)
```

2) Histogram je vlastně odhadem hustoty daného rozdělení. Chceme-li ho porovnat s křivkou této hustoty, je vhodné ho přeškálovat tak, aby jeho plocha byla 1 (tak jako u hustoty). To se provede pomocí argumentu `prob=TRUE`. Tvar histogramu pak zůstane stejný, ale změní se měřítko na y-ové ose.

```
hist(vek.m, prob=TRUE)
```

✧ Histogram si lze dále vylepšit pomocí známých grafických argumentů:

```
hist(vek.m, prob=TRUE, col="slateblue",
     xlab="Vek matky", ylab="Hustota", main="Histogram (vek matek)")
```

Pro porovnání lze pak do histogramu přidat hustotu normálního rozdělení, jehož parametry odhadneme z dat

```
curve(dnorm(x,mean(vek.m),sd(vek.m)), col="red", add=TRUE)
```

Šikmost a špičatost

Skutečné teoretické rozdělení daného znaku (např. věku matek) samozřejmě neznáme, a neznáme tudíž ani jeho skutečnou šikmost a špičatost. Můžeme si je ale odhadnout pomocí jejich výběrových protějšků - výběrové šikmosti a špičatosti.

3) Spočtěme **výběrovou šikmost a špičatost** pro věk matek. Pro připomenutí:

$$a_3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^3 \qquad \text{(výběrová šikmost)},$$

$$a_4 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^4 - 3 \quad \text{(výběrová špičatost)}.$$

✧ Veličiny

$$z_i = \frac{x_i - \bar{x}}{s_x} \quad (i = 1, \ldots, n)$$

se někdy nazývají z-skóry a v R je lze získat pomocí funkce `scale`. Výběrovou šikmost a špičatost proto snadno spočteme pomocí následujících příkazů:

```
mean(scale(vek.m)^3)          # sikmost
mean(scale(vek.m)^4) - 3      # spicatost
```

✧ Jakých hodnot (přibližně) by výběrová šikmost a špičatost měla nabývat, jestliže bychom mohli předpokládat normalitu rozdělení věku matek?

✧ Je hodnota výběrové šikmosti v souladu s tvarem histogramu pro věk matek z předchozí části?

6 Tvorba 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(vek.m < vek.o)
```

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

```
subset(Kojeni, vek.m < vek.o)
```

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(Kojeni)
```

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(Kojeni, vek.o > vek.m)
```

5) 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 ve 24. týdnu nejvýše 63 cm nebo nejméně 68 cm.

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

- ✧ `Hoch == "divka"`
- ✧ `vek.o - vek.m > 4`
- ✧ `hoch != 1`
- ✧ `hoch == 0`
- ✧ `Hoch != "hoch"`
- ✧ `vek.o != vek.m`
- ✧ `delka <= 63 | delka >= 68`
- ✧ `vek.o - vek.m >= 5`
- ✧ `!(delka > 63 & delka < 68)`
- ✧ `Hoch == "divka" & vek.o > vek.m`
- ✧ `abs(vek.o - vek.m) == 1`

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

7 Pravděpodobnost

Náhodný pokus je pokus konaný za přesně definovaných podmínek, jehož výsledek je předem nejistý (např. hod kostkou). Nejjemněji rozlišované možné výsledky náhodného pokusu, které se vzájemně vylučují (nemohou nastat dva současně) a jeden vždy musí nastat, se nazývají **elementární jevy** (např. pro hod kostkou je množina elementárních jevů rovna $\{1, 2, 3, 4, 5, 6\}$). Množina všech elementárních jevů daného pokusu se obvykle značí písmenem Ω. **Náhodný jev** A je pak libovolné tvrzení o výsledku náhodného pokusu. Náhodné jevy značíme zpravidla velkými písmeny ze začátku abecedy. Korektně pak můžeme náhodný jev A popsat tak, že vyjmenujeme, ze kterých elementárních jevů se skládá, tj. vyjmenujeme všechny elementární jevy, které jsou příznivé jevu A (jev A, že na kostce padne sudé číslo, se skládá z elementárních jevů $\{2, 4, 6\}$).

Jsou-li všechny elementární jevy stejně pravděpodobné (což je příklad třeba spravedlivé kostky, kde všechny stěny mají pravděpodobnost $1/6$), pak pravděpodobnost jevu A, který je složen z m_A elementárních jevů, spočítáme jednoduše jako

$$P(A) = \frac{m_A}{m},$$

kde m je celkový počet všech elementárních jevů (tj. celkový počet prvků množiny Ω). Tímto vzorečkem se dříve pravděpodobnost definovala, a tak dnes říkáme, že jde o tzv. **klasickou definici pravděpodobnosti**. Například pro jev A, že na kostce padne sudé číslo je $m_A = 3$ a $m = 6$, a tudíž $P(A) = \frac{3}{6} = \frac{1}{2}$.

1) Kolika způsoby si může 15 studentů *Základů biostatistiky* sednout k 15 počítačům v učebně B5? Jaká je pravděpodobnost, že si sednou (v učebně zleva doprava) přesně v pořadí, v jakém jsou uvedeni v prezenční listině?

✧ Množina Ω je tvořena všemi permutacemi, které lze vytvořit z 15 prvků (studentů). Celkový počet těchto permutací je $m = 15!$. V R to spočítáme jako

```
factorial(15)
```

✧ Uspořádání, které vyhovuje prezenční listině, je pouze jediné. Tedy počet příznivých uspořádání je $m_A = 1$, a tudíž pravděpodobnost této jediné permutace je dle klasické definice pravděpodobnosti rovna

```
1/factorial(15)
```

2) Kolika způsoby si může sednout do posluchárny o 170 židlích 170 studentů?

✧ Vyjde vám nějaké ohromné číslo, protože faktoriál se zvětšuje velkou rychlostí.

3) Kolika způsoby si může sednout do posluchárny o 171 židlích 171 studentů?

✧ Zde už je odpověď `Inf` (= *infinity*, nekonečno). I R má své limity...

4) S jakou pravděpodobností vyhrajete ve Sportce 1. cenu s jedním vsazeným sloupcem?

✧ Ve Sportce se tipuje 6 ze 49 čísel. Vyhrát první cenu znamená uhodnout všech šest čísel.

✧ Princip losování přitom zaručuje, že každá šestice čísel má stejnou pravděpodobnost. Množina Ω je tedy tvořena všemi šesticemi, které lze vytvořit ze 49 čísel. Těchto šestic je $\binom{49}{6}$, což v R vypočteme jako

```
choose(49, 6)
```

✧ Počet příznivých jevů, tj. počet šestic, na něž připadne první cena, je roven 1. Proto pravděpodobnost výhry první ceny je dle klasické definice pravděpodobnosti rovna

```
1 / choose(49, 6)
```

5) S jakou pravděpodobností vyhrajete ve Sportce 1. pořadí s úplně vyplněným tiketem (10 sloupců), jestliže v každém sloupci máte uvedenu jinou kombinaci vsazených čísel?

```
10 / choose(49, 6)
```

6) Házíme dvěma kostkami. S jakou pravděpodobností bude:

✧ součet obou hodnot roven 10?

✧ součet obou hodnot roven alespoň 10?

8 Výpočet střední hodnoty a rozptylu z definice

Připomeňte si vzorec pro výpočet střední hodnoty diskrétní náhodné veličiny (např. *NVR*, str. 3, vzorec (1)).

Příklad

Náhodná veličina X nabývá pouze hodnot $2, 3, 5$ a to s pravděpodobnostmi

$$P(X = 2) = 0.6 \qquad P(X = 3) = 0.3 \qquad P(X = 5) = 0.1.$$

Vypočtěte střední hodnotu a rozptyl této náhodné veličiny.

✧ **střední hodnotu** vypočteme z definice jako

$$\begin{aligned} \mathrm{E}\, X &= 2 \cdot P(X = 2) + 3 \cdot P(X = 3) + 5 \cdot P(X = 5) \\ &= 2 \cdot 0.6 + 3 \cdot 0.3 + 5 \cdot 0.1 = 2.6 \end{aligned}$$

✧ **rozptyl** můžeme také vypočítat z definice (viz např. *NVR*, str. 4, vzorec (2))

$$\begin{aligned}\operatorname{var} X &= \mathrm{E}(X - \mathrm{E}\,X)^2 \\ &= (2 - \mathrm{E}\,X)^2 \cdot P(X = 2) + (3 - \mathrm{E}\,X)^2 \cdot P(X = 3) + (5 - \mathrm{E}\,X)^2 \cdot P(X = 5) \\ &= (2 - 2.6)^2 \cdot 0.6 + (3 - 2.6)^2 \cdot 0.3 + (5 - 2.6)^2 \cdot 0.1 \\ &= 0.84\end{aligned}$$

✧ Někdy je výhodnější vypočítat rozptyl podle vzorečku

$$\operatorname{var} X = \mathrm{E}(X^2) - (\mathrm{E}\,X)^2$$

K tomu je potřeba si vypočítat $\mathrm{E}(X^2)$ (tzv. druhý moment veličiny X), což uděláme opět z definice střední hodnoty (tentokrát jde o střední hodnotu X^2)

$$\begin{aligned}\mathrm{E}(X^2) &= 2^2 \cdot P(X = 2) + 3^2 \cdot P(X = 3) + 5^2 \cdot P(X = 5) \\ &= 4 \cdot 0.6 + 9 \cdot 0.3 + 25 \cdot 0.1 = 7.6\end{aligned}$$

a pak dosadit

$$\operatorname{var} X = \mathrm{E}(X^2) - (\mathrm{E}\,X)^2 = 7.6 - (2.6)^2 = 0.84.$$

Příklad - samostatná práce

Náhodná veličina X nabývá pouze hodnot 0 a 1, a to s pravděpodobnostmi

$$P(X = 0) = 0.8 \qquad\qquad P(X = 1) = c.$$

Určete hodnotu c a spočtěte $\mathrm{E}\,X$.

9 Konec práce

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