Cvičení z matematické statistiky 02

Poslední úprava dokumentu: 6. března 2024.

Vztah dvou veličin - popisné statistiky a grafické znázornění

---

1 Začátek práce (podrobně)

1) BUĎ: chci pokračovat ve skriptovém souboru z minula:

- ✧ Na svém disku `J:` nalezněte příslušný soubor a klikněte na něj (měl by se automaticky otevřít v RStudiu, pokud ne, tak klikněte pravým tlačítkem myši → `Otevřít v programu` → RStudio)
- ✧ nebo spusťte RStudio z nabídky a v Menu nahoře zvolte
  File ➥ **Open File** a vyberte příslušný soubor.
- ✧ Nastavení pracovního adresáře: v Menu nahoře zvolte
  Session ➥ **Set Working Directory** ➥ **To Source File Location**

2) NEBO: chci si otevřít nový skriptový soubor:

- ✧ Spusťte RStudio z nabídky.
- ✧ Nastavte si pracovní adresář.
  - Buď v Menu nahoře:
    Session ➥ **Set Working Directory** ➥ **Choose Directory ...**
    a vyberte složku `matstat`, kterou jste si vytvořili minule.
  - Nebo příkazem:
    ```
    setwd("popis_cesty/matstat")
    ```
- ✧ Otevřte si nový skriptový soubor
  - Buď v Menu nahoře:
    File ➥ **New File** ➥ **R Script**
  - nebo kliknutím na ikonku papíru se zeleným plus vlevo nahoře a zvolením `R Script`,
  - nebo klávesovou zkratkou `Ctrl+Shift+N`.
- ✧ Nový skript si nezapomeňte uložit:
  - **File** ➥ **Save as...**
  - nebo kliknutím na symbol diskety,
  - nebo klávesovou zkratkou `Ctrl+S`.

2 Rozcvička

☞ Vytvořte si libovolný vektor o pěti hodnotách a spočtěte průměr a směrodatnou odchylku.

3 Načtení dat

1) Načtěte data studenti do RStudia:

Máte-li data z minula uložená ve formátu `RData`:

- ✧ Buď v Menu nahoře:

  File ➥ **Open file...**

  a potvrdit soubor `studenti.RData` ze složky `matstat`.
  Do you want to load the R data file into the global environment?

  Yes

- ✧ Nebo příkazem:

```
load("studenti.RData")
```

Máte-li data k dispozici pouze v původním formátu `csv`:

- ✧ Buď v pravém dolním okně v záložce `Files` klikněte na příslušná data a zvolte **Import Dataset** ➥ **From Text**, vyberte soubor `studenti.csv` a nastavte:

Name	studenti
Heading	yes
Separator	Comma
Delimiter	Period

- ✧ Nebo použijte příkaz:

```
studenti <- read.csv2("studenti.csv", header=TRUE, sep=",", dec=".")
```

2) Prohlédněte si data a ujistěte se, že jsou správně načtena. Data si můžete zobrazit buď poklepáním na jejich název v pravém horním okně, nebo příkazem

```
View(studenti)
```

3) Pomocí příkazu `summary(studenti)` si nechte vypsat základní popisné statistiky pro všechny veličiny. Připomeňte si, co jednotlivé proměnné znamenají.

- ✧ Veličiny mesic.naroz a kraj nastavte jako faktory:

```
studenti$mesic.naroz <- as.factor(studenti$mesic.naroz)
studenti$kraj <- as.factor(studenti$kraj)
```

- ✧ Dále vytvořte veličinu fpohlavi:

```
studenti$fpohlavi <- factor(studenti$pohlavi, labels = c("zena","muz"))
```

- ✧ Zaveďte novou veličinu, pomocí které zjistíte, kolik procent studentů se narodilo na jaře, v létě, na podzim a v zimě. Vypište si procentuální zastoupení jednotlivých ročních období. namalujte si vhodný obrázek.

```
pom.obdobi <- 0*(studenti$mesic.naroz %in% c(12,1,2)) +
                + 1*(studenti$mesic.naroz %in% c(3,4,5)) +
                + 2*(studenti$mesic.naroz %in% c(6,7,8)) +
                + 3*(studenti$mesic.naroz %in% c(9,10,11))
studenti$obdobi <-  factor(pom.obdobi,labels = c("zima","jaro","leto","podzim"))
```

Zjistěte, kolik studentů se narodilo v kterém ročním období. Vykreslete si vhodné grafy:

```
pie(table(studenti$obdobi), col=c("blue","green","red","yellow"))
barplot(table(studenti$obdobi), xlab="hmotnost", ylab="četnost")
```

✧ Uložte si aktuální verzi datové tabulky

```
save(studenti,file="studenti.RData")
```

✧ Zajistěte si přímý přístup k jednotlivým proměnným datového souboru studenti pomocí příkazu

```
attach(studenti)
```

4 Popis vztahu kvantitativní a kvalitativní veličiny

1) Budeme zkoumat zda a jak se liší výška mužů a výška žen.

✧ Nechte si vypsat zvlášť popisné statistiky pro muže a ženy a prohlédněte si, v čem jsou odlišnosti.

```
tapply(vyska, fpohlavi, summary)
```

✧ Graficky lze předchozí čísla znázornit pomocí krabicového grafu

```
boxplot(vyska ~ fpohlavi)
```

✧ Dále by nás mohl zajímat histogram výšky pro muže a ženy zvlášť

```
library(lattice)
histogram(~vyska|fpohlavi, data = studenti)
```

Nemáte-li knihovnu `lattice` nainstalovanou, pak buď zadejte `install.packages("lattice")`, nebo v pravém dolním okně běžte na záložku `Packages`, dále klikněte na `Install`, do políčka v příslušném okně napište název knihovny a potvrďte tlačítkem `Install`.
Pokud vám knihovna z nějakého důvodu nejde nainstalovat, můžete vykreslit oba histogramy vedle sebe takto:

```
par(mfrow=c(1,2))   # graficke okno se pomyslne rozdeli na dve casti
hist(vyska[fpohlavi=="zena"], main="Histogram pro ženy")
hist(vyska[fpohlavi=="muz"], main="Histogram pro muže")
par(mfrow=c(1,1))   # aby byl nadale v grafickem okne pouze jeden obrazek
```

2) Stejným způsobem si prohlédněte, jestli a jak se liší hmotnost u můžu a žen.

3) Zjistěte, zda se věk otců liší pro dívky a pro chlapce.

5 Vztah dvou kvantitativních veličin

Podíváme se na vztah výšky a váhy (společně pro muže i ženy).

4) Bodový graf (**scatterplot**) váhy proti délce. Proveďte následujícíc dva příkazy a porovnejte, která veličina je na ose x a na ose y:

```
plot(vyska, vaha)
plot(vyska ~ vaha)
```

✧ Obrázek lze dále zkrášlovat pomocí argumentů z minulého cvičení:

```
plot(delka, vaha, xlab="Výška (cm)", ylab="Váha (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()*

✧ Dále obrázek vylepšeme přidáním nadpisu:

```
plot(vyska, vaha, main="Závislost váhy na výšce",
   xlab="Výška (cm)", ylab="Vaha (kg)", col="seagreen", pch=13)
```

Číselně lze vztah dvou kvantitativních veličin popsat pomocí korelace, ale o tom si povíme až na některém z dalších cvičení. Později budeme také zkoumat závislost dvou kvantitativních veličin pomocí lineární regrese.

5) Stejným způsobem prozkoumejte vztah věku matky a věku otce.

6) Souvisí spolu velikost bot a index BMI?

6 Vztah výšky a váhy v závislosti na pohlaví (tj. vztah 3 veličin)

7) V dalším kroku si pomocí barev a symbolů rozlišíme muže a ženy. Připomínám, že pohlaví udávají proměnná pohlavi (1 pro muže, 0 pro ženu), respektive proměnná fpohlavi.

✧ Na naší úrovni je asi nejprůhlednější následující konstrukce:

```
divky <- which(fpohlavi=="zena")   # kde jsou v datech dívky
hosi <- which(fpohlavi=="muz")   # kde jsou v datech hoši
plot(vyska[divky], vaha[divky], col="pink", pch=16,
main="Závislost váhy na výšce", xlab="Výška (cm)", ylab="Váha (kg)")
points(vyska[hosi], vaha[hosi], col="blue", 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 pohlavi

```
barvy <- c("pink", "blue")
symboly <- c(16, 17)
plot(vyska, vaha, col=barvy[pohlavi+1], pch=symboly[pohlavi+1],
main="Závislost váhy na délce", xlab="Výška (cm)", ylab="Váha (kg)")
```

✧ barevné odlišení s využitím faktorové proměnné fpohlavi

```
barvy <- c(zena="pink",muz="blue")
symboly <- c(zena=16, muz=17)
plot(vyska, vaha, col=barvy[fpohlavi], pch=symboly[fpohlavi],
main="Závislost výšky na délce", xlab="Výška (cm)", ylab="Váha (kg)")
```

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

8) Úplně nakonec přidáme do obrázku též legendu.

```
legend("topleft", legend=c("Zena", "Muz"), col=c("pink","blue"), pch=c(16,17))
```

7 Vztah dvou kvalitativních veličin

9) Budeme se zajímat o to, zda roční období narození (veličina `obdobi`) nějak souvisí s pohlavím.

✧ Nechte si vypsat tabulku četností těchto dvou znaků:

```
table(obdobi, fpohlavi)                          # absolutni cetnosti
proportions(table(obdobi, fpohlavi), margin=1)   # radkove relativni cetnosti
proportions(table(obdobi, fpohlavi), margin=2)   # sloupcove relativni cetnosti
```

Zamyslete se nad interpretací uvedených hodnot. Je nějaký rozdíl mezi muži a ženami, co se týče nadváhy a podváhy?

10) Vykreslíme si sloupcový graf nadváhy zvlášť pro muže a ženy. Do skriptového okna přepište:

```
barplot(table(obdobi, fpohlavi), beside=TRUE, legend=TRUE)
```

11) Další možný popisný obrázek si vykreslíme pomocí:

```
plot(fpohlavi, obdobi, ylab="roční období")
```

Co všechno lze z obrázku vyčíst? Změňte pořadí pohlaví a ročního období narození. Jak se změní obrázek?

8 Vytvoření podmnožiny dat

Někdy je potřeba zpracovávat pouze podmnožinu dat, jež splňuje nějakou podmínku. Tomuto tématu jsme se již krátce věnovali na minulém cvičení. Níže najdete příklady, jak vybrat z dat podmnožinu splňující určitou podmínku a jak tuto podmnožinu uložit.

12) Zjistíme, pro které studenty je otec starší než matka:

```
which(vek.matky < vek.otce)
```

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

```
subset(studenti, vek.matky < vek.otce)
```

14) 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(studenti)
```

15) Řekněme, že dále budeme chtít pracovat pouze se studenty, u nichž je otec starší než matka. Vytvořenou podmnožinu si můžeme uložit do datové tabulky studentiOsM.

```
studentiOsM <- subset(studenti, vek.matky < vek.otce)
```

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

```
save(studentiOsM, file = "studentiOsM.RData")
```

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

(a) Studenti, u kterých je otec jinak starý než matka.

(b) Studenti, u kterých je otec o alespoň 5 let starší než matka.

(c) Studenti, u kterých se věk rodičů liší o právě jeden rok.

(d) Ženy.

(e) Ženy, které mají otce staršího než matku.

(f) Studenti, kteří mají výšku nejvýše 170 cm nebo nejméně 180 cm.

Nápověda: Ke specifikaci jednotlivých podmnožin si vybírejte z následujících logických výrazů (vybraným výrazem pak nahraďte výraz `vek.matky < vek.otce` v bodě 16):

- `fpohlavi == "zena"`
- `vek.otce - vek.matky > 4`
- `pohlavi != 1`
- `pohlavi == 0`
- `fpohlavi != "muz"`
- `vek.otce != vek.matky`
- `vyska <= 170 | vyska >= 180`
- `vek.otce - vek.matky >= 5`
- `!(vyska > 180 & vyska < 170)`
- `fpohlavi == "zena"& vek.otce > vek.matky`
- `abs(vek.otce - vek.matky) == 1`

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

9 Konec práce

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