Cvičení z matematické statistiky 01

Poslední úprava dokumentu: 27. února 2024.

Seznámení s R a popisná statistika pro jednu proměnnou

Pracovní listy k tomuto cvičení vznikly s využitím materiálů Dr. Hudecové a doc. Zváry.

1 Úvod

1) Postup prací pro cvičení se bude postupně objevovat na webu:

`https://web.natur.cuni.cz/uamvt/turcm6am`

Tam jsou také uvedeny podmínky zápočtu a kontakt na cvičícího (pro případ dotazů či žádosti o konzultaci).

2) Ve svém domovském adresáři na disku `J:` si vytvořte složku `matstat` k tomuto cvičení.

3) Do složky `matstat` zkopírujte soubor `studenti.csv` a `cvMS01.R`, které jsou k dispozici na disku `V:` ve složce `turcicova`.

4) Spust'te RStudio (najdete ho v nabídce `Start`).

5) V RStudio nastavte coby pracovní adresář složku `matstat`, a to jedním z následujících způsobů:

(a) Napište a pomocí Enter odešlete v okně Console příkaz

```
setwd("popis_cesty/matstat")
```

kde `popis_cesty` nahrad'te za popis cesty ke složce `matstat`. (Lomítka v popisu cesty musejí být dopředná, tj. "/", nikoli zpětná.)

(b) V horní nabídce postupně zvolte

Session ➥ **Set Working directory** ➥ **Choose directory...**

a pomocí myši najděte složku `matstat`.

6) Zjistěte, zda se vše povedlo. Po zadání příkazu

```
getwd()
```

by se mělo vypsat popis_cesty/matstat.

7) Pomocí horní nabídky zaved'te do skriptového okna (alias Script Window, vlevo nahoře) skriptový soubor `cv01.R`:

File ➥ **Open File...**

Několik poznámek k práci se skriptovým souborem

- ✧ Při psaní příkazů v Script Window jsou vám automaticky nabízeny příkazy začínající na daná písmena. Pokud se vám některý z těchto příkazů zamlouvá, vyberte ho pomocí šipky na klávesnici a poté potvrďte klávesou `Enter`.
- ✧ Znak `#` odděluje poznámky, tj. něco, co chcete v souboru vidět vy, ale nechcete, aby to vidělo R.
- ✧ Různé příkazy musí být na různých řádcích (doporučuji), nebo na stejném řádku oddělené středníkem.
- ✧ Když chceme, aby se provedl příkaz na konkrétním řádku, nastavíme kurzor na příslušný řádek a zmáčkneme klávesovou zkratku `Ctrl-Enter`, nebo klikneme na tlačítko **Run** (v pravém horním rohu Script Window).
- ✧ Když chceme, aby se provedlo více příkazů najednou, označíme je myší jako blok a zmáčkneme klávesovou zkratku `Ctrl-Enter`, nebo klikneme na tlačítko **Run** (v pravém horním rohu Script Window).
- ✧ Čas od času si (doplňovaný) skriptový soubor uložte (stačí sem tam stisknout klávesovou zkratku `Ctrl-S`) nebo kliknout na ikonu diskety vlevo nahoře.
- ✧ Nápovědu k libovolnému příkazu vyvoláme pomocí `help(prikaz)` nebo `?prikaz`. (Nápověda se pak zobrazí v pravém dolním okně.)

Další poznámky pro práci s R / RStudiem

- ✧ Všechny požadované příkazy je nutné napsat do Script Window nebo do Console.
- ✧ Po odeslání požadovaného příkazu se příkaz v tichosti provede a v Console se objeví výsledky výpočtu (jsou-li nějaké). V případě chybného příkazu se objeví červená chybová hláška.
- ✧ Předchozí příkazy je možné v Consoli vyvolávat pomocí "šipky nahoru" na klávesnici. Vyvolaný příkaz lze dále upravit a odeslat klávesou `Enter`.
- ✧ Chceme-li provést ve výpočtu drobnou změnu (např. zpracovat proměnnou vaha místo proměnné vyska), můžeme si příslušnou část ve skriptovém okně zkopírovat níže, upravit a odeslat.
- ✧ Pro pozdější použití vytvořeného skriptu je dobré si do něj sem tam napsat nějaký komentář (za znak `#`).

Úkol označený symbolem vlevo značí samostatnou práci, které se můžete věnovat na cvičení, pokud bude čas, nebo se na ni můžete podívat doma.

2 Základní operace a funkce v R

1) Nejprve si pomocí příkazů předepsaných ve skriptu `cvMS01.R` vyzkoušejte, že R lze použít jako kalkulačku.

2) Práce s proměnnými

```
a <- 2    # ulozeni hodnoty do promenne
a = 2     # dalsi moznost ulozeni hodnoty do promenne
a         # vypsani hodnoty promenne
print(a)  # dalsi zpusob vypsani hodnoty promenne
b <- 3    # ulozeni do jine promenne
b
a + b     # secteni ulozenych hodnot
a - b     # odecteni ulozenych hodnot
a * b     # vynasobeni ulozenych hodnot
a / b     # vydeleni ulozenych hodnot
c <- log(b)         # vytvoreni promenne na zaklade jine promenne
c
```

3) Vytvoření vektoru hodnot: Do vektoru nazvaného `N` si uložíme 5 měření obsahu dusíku v ovzduší ($\mu g/m^3$) v okolí slévárny, které jsme naměřili:

```
N <- c(10.53, 22.40, 16.34, 13.07, 18.31)
```

Spočítejte minimální, maximální a průměrný obsah dusíku, který jsme naměřili.

```
min(N)
max(N)
mean(N)
```

4) „Úklid“ (před začátkem jiné práce)

```
ls()          # zjistime, jake promenne mame nadefinovane
rm(list=ls()) # vycistime Rko (vsechny promenne se smazou)
ls()          # zjistime, jake promenne mame nactene (snad nic)
```

3 Práce s daty

✧ Data (nějaká vlastní měření apod.) si většinou ukládáte v nějakém tabulkovém procesoru (LibreOffice Calc, MS Excel apod.) Z tabulkového procesoru lze data uložit ve formě textového souboru (např. ve formátu `csv`), obvykle nabídkou **Save As**/**Uložit jako**. Z formátu `csv` lze již data snadno načíst do R.

✧ Další formát vhodný pro načítání do R je `txt` (k jeho načtení by se použil příkaz `read.table`).

✧ Podívejme se na data uložená v souboru `studenti.csv`, která obsahují následující proměnné:

`ID`	identifikační číslo studenta
`Rok`	rok přednášky
`mesic.naroz`	měsíc narozní
`rok.naroz`	rok narození
`pohlavi`	pohlaví (0 žena, 1 muž)
`vyska`	výška v cm
`vaha`	hmotnost v kg
`boty`	velikost bot
`pocet.souroz`	počet sourozenců
`vek.otce`	věk otce
`vek.matky`	věk matky
`kraj`	bydliště (kraj)

1) Načtěte data do R.

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) Data si můžete prohlédnout pomocí

```
View(studenti)
```

nebo

```
print(studenti)
head(studenti)   # vypise pouze zacatek souboru
```

4 Veličiny měřené na různých měřítkách

Je zřejmé, že data jsou zaznamenávána v různých měřítkách. Jakými charakteristikami a obrázky by bylo vhodné je popsat?

3) Představu o datech získáme pomocí popisných statistik. Nechte si vypsat základní charakteristiky polohy a variability všech veličin zahrnutých v datech.

```
summary(studenti)
```

Připomeňte si význam jednotlivých charakteristik polohy.
Zkratka `NA` (*not available*) označuje chybějící hodnotu.
Dále si všimněte rozdílu mezi tím, co R vypsalo pro výšku a kraje. Je popis veličiny udávající pohlaví smysluplný?

4) Jsou-li některé veličiny kategoriální a jsou kódovány pomocí čísel, musíte R sdělit, že je má chápat jako tzv. faktory. To je případ veličiny `pohlavi`, která je v nula-jedničkovém měřítku, tedy v kategoriálním měřítku, což R samo o sobě nemůže poznat. Proto z této veličiny ručně vytvoříme faktor.

```
studenti$pohlavi <- as.factor(studenti$pohlavi)
```

Poté zopakujte `summary(studenti)` a podívejte se, co se změnilo.

5 Popis kvalitativních veličin

5) Zjistěte, jaké je v datech zastoupení mužů a žen.

```
table(studenti$pohlavi)                # absolutni cetnosti
proportions(table(studenti$pohlavi))   # relativni cetnosti
```

Nebo se podívejte do výstupů funkce `summary` výše.

6) Předchozí výsledek si znázorníme také graficky (k tomu je vhodné si předchozí tabulky uložit do proměnných):

```
cetnosti <- table(studenti$pohlavi)
rel.cetnosti <- proportions(table(studenti$pohlavi))
pie(rel.cetnosti)          # kolacovy graf
barplot(cetnosti, xlab="pohlaví", ylab = "četnost")     # sloupcovy graf
```

7) Uložte si jeden z předchozích obrázků do svého adresáře `matstat`: v okně s obrázkem klikněte na `Export` → `Save as Image` → zadejte název → `Save`.

8) Podívejte se na popis veličiny udávající kraj.

9) Zjistěte, jaké je v datech procentuální zastoupení Pražáků.

6 Popis kvantitativních veličin

10) Vypište si popisné statistiky veličiny udávající výšku studentů (bez rozdílu pohlaví).

```
summary(studenti$vyska)
```

Dále vypočtěte některé charakteristiky variability (musíme ošetřit, že v datech se nacházejí chybějící pozorování):

```
sd(studenti$vyska, na.rm=TRUE)    # smerodatna odchylka (standard deviation)
var(studenti$vyska, na.rm=TRUE)   # rozptyl (variance)
IQR(studenti$vyska, na.rm=TRUE)   # mezikvartilove rozpeti (interquartile range)
```

11) Vykreslete si krabicový diagram a histogram a připomeňte si, co tyto grafy znázorňují. Co z nich umíte vyčíst?

```
boxplot(studenti$vyska)        # krabicovy diagram
hist(studenti$vyska)           # histogram
```

✧Pomocí argumentu `breaks` můžeme volit počet intervalů, které jsou v histogramu uvažovány. Podívejte se, jak se histogram mění, zvyšujeme-li a snižujeme-li jejich počet.

```
hist(studenti$vyska, breaks=20)
```

7 Změny nastavení v obrázcích aneb zkrášlování obrázků

12) Vezměte si například obrázek histogramu a pokuste se ho trochu vylepšit.

✧ změňte název os:

```
hist(studenti$vyska, xlab = "výška (cm)", ylab = "četnost")
```

✧ změňte barvu:

```
hist(studenti$vyska, xlab = "výška (cm)", ylab = "četnost", col = "blue")
```

✧ přidejte nadpis

```
hist(studenti$vyska, xlab = "výška (cm)", ylab = "četnost", col = "blue",
main = "Histogram výšky studentů")
```

8 Počítání nových proměnných

13) Zaveďte novou veličinu, která bude udávat výšku v metrech.

```
studenti$vyska.m <-  studenti$vyska/100
```

Nechte si vypsat základní popisné statistiky pro tuto novou veličinu. Porovnejte je s charakteristikami výšky v cm. Jak se změnil průměr a jak směrodatná odchylka?

14) Vytvořte novou proměnnou udávající věk studentů v letech a zjistěte, jaký byl nejmladší a nejstarší student na přednášce. Nakreslete si krabicový diagram věku studentů.

15) Spočtěte rozdíl věku rodičů studentů. Jaký je průměrný rozdíl věků rodičů? Jaká je minimální a maximální hodnota v datech?

16) Zaveďte novou veličinu udávající BMI (body mass index) studentů.

```
studenti$BMI <-  studenti$vaha/(studenti$vyska.m)^2
```

Prohlédněte si základní popisné statistiky BMI studentů. Vykreslete si vhodné popisné grafy. Identifikujte „odlehlá“ pozorování a prohlédněte si jejich záznamy.

17) 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$obdobi), col=c("blue","green","red","yellow"),
    main="Koláčový graf")
barplot(table(studenti$obdobi), xlab="hmotnost", ylab="četnost",
        col=c("blue","green","red","yellow"), main = "Roční období narození")
```

9 Uložení dat

Uložíme si současnou formu datové tabulky studenti:

1) bud' ve formátu csv

```
write.table(studenti, file = "studenti_upr.csv",row.names=FALSE,
                                        col.names=TRUE, sep=";", dec=",")
```

2) nebo ve formě R datového formátu (přípona RData)

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

10 Konec práce

Než zavřete všechna okna, nezapomeňte si uložit skriptový soubor:

File ➥ Save as

nebo klávesovou zkratkou Ctrl+s. (Při standardním zavírání otevřených oken budete tak jako tak dotázáni, zda chcete tak učinit).