Cvičení z matematické statistiky 02

Poslední úprava dokumentu: 25. února 2025.

Popisná statistika pro jednu proměnnou

Rozcvička

Vytvořte si libovolný vektor o pěti hodnotách a spočtěte součet první až třetí a poslední hodnoty.

1 Úvod

1) Spusťte RStudio z nabídky.

2) 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")
```

3) 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`.

Nebo použijte skript z minula:

- ✧ V Menu nahoře:
  File ➥ **Open File**
  a vyberte příslušný soubor.

4) Pokud jste se rozhodli pro vytvoření nového skriptového souboru, nezapomeňte si ho uložit:

- ✧ **File** ➥ **Save as...**
- ✧ nebo kliknutím na symbol diskety,
- ✧ nebo klávesovou zkratkou `Ctrl+S`.

5) Nemáte-li z minula uložen datový soubor Sleep v R formátu (soubor `Sleep.RData`), vytvořte si ho dle návodu v předchozím pracovním listu „Seznámení s R“. Máte-li data již načtena, můžete tento i další krok přeskočit a přejít rovnou k bodu 7).

6) Načtěte data Sleep do RStudia:

✧ Buď v Menu nahoře:

File ➥ **Open file...**

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

Yes

✧ Nebo příkazem:

```
load("data/Sleep.RData")
```

7) 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(Sleep)
```

8) Zajistěte si přímý přístup k jednotlivým proměnným datového souboru Sleep. Ze skriptového okna spust'te příkaz:

```
attach(Sleep)
```

2 Jedna kvantitativní proměnná

2.1 Numerické hodnoty popisných statistik

1) Připomeňte si z přednášky pojmy: průměr, směrodatná odchylka, střední chyba průměru medián, dolní a horní kvartil, kvantily. Uměli byste je pro konkrétní data vypočítat ručně pomocí kalkulačky?

2) Spočtěte výše uvedené popisné statistiky pro délku spánku.

```
summary(sleep_length)        # prumer a nektere dulezite kvantily

mean(sleep_length)           # vyberovy prumer
sd(sleep_length)             # vyberova smerodatna odchylka
var(sleep_length)            # vyberovy rozptyl
median(sleep_length)         # median
min(sleep_length)            # minimum
max(sleep_length)            # maximum
quantile(sleep_length, prob=c(0, 0.25, 0.50, 0.75, 1.00))
                             # vybrane kvantily
length(sleep_length)         # delka vektoru (pocet pozorovani)
```

3) Zkusme zcela ručně (dle vzorečku z přednášky) spočítat medián pro dobu cvičení.

✧ Bude potřeba setřídit pozorované hodnoty. Toho lze docílit příkazem:

```
sort(exercise)
```

✧ Kolik máme pozorvání?

```
length(exercise)
```

✧ Jak se určí medián v případě lichého a sudého počtu pozorování?

4) Spočtěme dle vzorečku z přednášky směrodatnou odchylku pro délku spánku.

```
n <- length(sleep_length)
prumer <- sum(sleep_length) / n
sqrt(sum((sleep_length - prumer)^2)/(n-1))
```

5) Spočtěte střední chybu průměru (můžete přitom využít příkaz `sd`) pro délku spánku a připomeňte si její význam.

```
n <- length(sleep_length)
sd(sleep_length)/sqrt(n)
```

6) Základní popisné statistiky pro všechny proměnné datového souboru lze získat následujícím příkazem:

```
summary(Sleep)
```

✧ Povšimněte si jiného tvaru výstupu u proměnné gender. Jedná se o kvalitativní proměnnou, kterými se budeme zabývat v některém z příštích cvičení.

✧ Doporučuji po každém načtení dat spočítat vždy tyto statistiky a alespoň rychlým pohledem si je prohlédnout. Odhalíte tak hrubé chyby vzniklé při přepisu dat, nebo případné problémy s načtením dat.

2.2 Grafické znázornění popisných statistik

1) Mnohé popisné statistiky kvantitativní proměnné lze přehledně znázornit pomocí krabičkového grafu. Nakresleme krabičkový graf pro věk.

```
boxplot(age)
```

Příkaz lze modifikovat a zkrášlovat obrázek.

```
boxplot(age, ylab="Věk", col="steelblue")
boxplot(age, xlab="Věk", col="sandybrown", horizontal=TRUE)
```

2) Samostatně nakreslete krabičkový graf pro dobu cvičení. Jsou v grafu nějaká odlehlá pozorování?

3) Samostatně spočtěte popisné statistiky a vykreslete krabičkové grafy pro příjem kofeinu a délku pracovní doby.

4) Připomeňme, že

$$\mathsf{adulthood} = \mathsf{age} - 18$$
$$\mathsf{caffeine\_intake.g} = \mathsf{caffeine\_inkate}/1000$$

✧ **Vliv posunutí:** Jaký je vztah mezi popisnými statistikami pro proměnné age a adulthood?

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

3 Jedna kvalitativní proměnná

3.1 Tabulka absolutních a relativních četností

1) Spočítejte základní popisné statistiky pro proměnnou mood_score (nálada na škále 1-10).

```
summary(mood_score)
```

2) Veličina mood_score je sice kvantitativní, ale nabývá pouze celočíselných hodnot 1–10. Podívejme se, kolikrát se která hodnota vyskytuje.

```
table(mood_score)
```

3.2 Tvorba `factor`ů

3) Z pohledu R je veličina mood_score kvantitativní/číselná (těm se v R říká `numeric` nebo `integer`). Zkuste

```
class(mood_score)
```

Někdy se ale může hodit z ní udělat veličinu kategoriální (těm se v R říká `factor`), jejíž kategorie budou dány stupněm nálady.

Jako první nezapomeňte „odpojit“ data:

```
detach(Sleep)
```

Novou veličinu pak vytvoříte pomocí příkazu:

✧ Buď

```
Sleep <- transform(Sleep, fmood = factor(mood_score))
```

✧ nebo

```
Sleep$fmood <- as.factor(Sleep$mood_score)
```

Že se skutečně jedná o `factor`ovou veličinu si můžete ověřit pomocí:

```
class(Sleep$fmood)
```

4) Řekněme, že nás ale takto jemné dělení nálady nezajímá a stačí nám rozlišovat náladu pouze na *dobrou* (mood_score = 7 – 10), *neutrální* (mood_score = 4 – 6) a *špatnou* (mood_score = 1 – 3). Jinými slovy, chceme z proměnné mood_score vytvořit kvalitativní proměnnou o třech úrovních *dobrá/neutrální/špatná*, resp. *good/neutral/bad*.

✧ Vytvořme nyní proměnnou fmood3, jež bude `factor`em vytvořeným z mood_score výše popsaným způsobem.

```
pom <- 0*(Sleep$mood_score <= 3) + 1*(Sleep$mood_score %in% c(4,5,6)) +
                                 + 2*(Sleep$mood_score >= 7)
Sleep <- transform(Sleep, fmood3 = factor(pom, labels=c("bad", "neutral", "good")))
```

5) Spočtěme nyní absolutní a relativní četnosti jednotlivých úrovní fmood3.

```
table(Sleep$fmood3)
prop.table(table(Sleep$fmood3))
round(prop.table(table(Sleep$fmood3)) * 100, 2)
```

3.3 Grafické zobrazení absolutních a relativních četností

1) Zobrazme absolutní četnosti pomocí sloupcového grafu.

```
barplot(table(Sleep$fmood3), xlab="Nálada", ylab="Četnost",
col=terrain.colors(3))
```

2) Zobrazme relativní četnosti pomocí sloupcového grafu.

```
barplot(prop.table(table(Sleep$fmood3)),
xlab="Nálada", ylab="Relativní četnost", col=topo.colors(3))
```

3) Zobrazme relativní četnosti pomocí koláčového grafu.

```
pie(table(Sleep$fmood3), main="Nálada", col=heat.colors(3))
```

4 Samostatná práce

1) Vytvořte novou proměnnou nazvanou fstress3, jež bude odlišovat jedince s nízkou, střední a vysokou úrovní stresu.

2) Spočtěte absolutní a relativní četnosti výskytu jednotlivých úrovní stresu v datech.

3) Graficky znázorněte absolutní a relativní četnosti jednotlivých úrovní stresu.

5 Konec práce

Rozšířený datový soubor Sleep (s novými proměnnými fmood3 a fstress3) uložte na síťovém disku ve svém adresáři v R formátu jako soubor Sleep.RData (tj. přepište dřívější Sleep.RData novou verzí) do podsložky `data`.

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