

Poslední úprava dokumentu: 5. března 2025.

Popisná statistika pro dvě proměnné

Nemáte-li, načtěte data Sleep do RStudio:

```
load("data/Sleep.RData") # jsou-li data ve formátu .RData
```

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

```
attach(Sleep)
```

1 Dvě kvantitativní proměnné

1.1 Grafické znázornění

- 1) Bodový graf délky spánku proti věku.

```
plot(age, sleep_length)
```

✧ Úpravou příkazu lze obrázek dále zkrášlovat:

```
plot(age, sleep_length, xlab="Věk", ylab="Spánek (hodiny)", 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 vylepšíme přidáním nadpisu:

```
plot(age, sleep_length, main="Délka spánku vs. věk",  
      xlab="Věk", ylab="Spánek (hodiny)", col="seagreen", pch=13)
```

- 3) V dalším kroku si pomocí barev a symbolů rozlišíme muže a ženy. Pohlaví udává proměnná gender ("male" pro muže, "female" pro ženy).

```
muzi <- which(gender=="Male") # kde jsou v datech muzi  
zeny <- which(gender=="Female") # kde jsou v datech zeny  
plot(age[zeny], sleep_length[zeny], col="purple", pch=16,  
      main="Délka spánku vs. věk", xlab="Věk", ylab="Spánek (hodiny)")  
points(age[muzi], sleep_length[muzi], col="darkblue", pch=17)
```

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

- 4) Úplně nakonec přidáme do obrázku též legendu.

```
legend("topleft", legend=c("Žena", "Muž"), col=c("purple", "darkblue"), pch=c(16, 17))
```



- 5) Nakreslete vhodný obrázek pro popis závislosti nálady na množství kofeinu a interpretujte ho.

✧ Nejspíš vás napadne použít příkaz

```
plot(cafeine_intake, mood_score)
```

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

```
plot(mood_score ~ cafeine_intake)
```

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

1.2 Kovariance a korelace

- 6) **Výběrová kovariance** mezi náladou a příjmem kofeinu.

✧ Ze skriptového okna spusťte příkaz:

```
cov(mood_score, cafeine_intake)
```

- 7) **Korelační koeficient** mezi náladou a příjmem kofeinu.

```
cor(mood_score, cafeine_intake)
```



- 8) Vypočtete korelaci mezi věkem a délkou práce a vykreslete tyto veličiny do bodového grafu.



- 9) Spočtete korelaci mezi příjmem kofeinu a úrovní stresu 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

Ujistěte se, že má v datech z minula vytvořenou proměnnou `fmood3` a `fstress3`, rozdělující náladu do tří kategorií. Pokud ne, proveďte příkazy:

```
detach(Sleep)
pom1 <- 0*(Sleep$mood_score <= 3) + 1*(Sleep$mood_score %in% c(4,5,6))
      + 2*(Sleep$mood_score >= 7)
Sleep <- transform(Sleep, fmood3 = factor(pom1, labels=c("bad", "neutral", "good")))
pom2 <- 0*(Sleep$stress_level <= 3) + 1*(Sleep$stress_level %in% c(4,5,6))
      + 2*(Sleep$stress_level >= 7)
Sleep <- transform(Sleep, fstress3 = factor(pom2, labels=c("low", "moderate", "high")))
attach(Sleep)
```

Zjistěme, jak spolu souvisí úroveň stresu (proměnná `fstress3`) a nálada (proměnná `fmood3`).

- 1) Spočteme kontingenční tabulku pro `fstress3` a `fmood3` a s ní související (podmíněné) relativní četnosti (podmíněno stresem).

```
table(fstress3, fmood3)
prop.table(table(fstress3, fmood3), margin=1) * 100
```

- ✧ Zamyslete se, jak by (ideálně – v populaci, v datech při vyšším rozsahu výběru) měly vypadat podmíněné relativní četnosti, jestliže nálada nesouvisí s úrovní stresu.
- ✧ Zdá se vám, že nálada spíše souvisí, či spíše nesouvisí s úrovní stresu?
- ✧ Svoje závěry v žádném případě nepovažujte za definitivní. Nezapomeňte, že máme k dispozici pouhých 194 pozorování.
- ✧ Později budeme zkoumat závislost dvou kvalitativních veličin pomocí χ^2 (chí kvadrát) testu nezávislosti.

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

```
plot(fmood3~fstress3, col=terrain.colors(3), xlab="Stres", ylab="Relativní četnost")
```



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

```
prop.table(table(fstress3, fmood3), margin=2) * 100
```

A co celková procenta?

```
prop.table(table(fstress3, fmood3)) * 100
```



- 4) Pomocí vhodných popisných statistik a obrázků vyšetřete závislosti mezi pohlavím (proměnná `gender`) a úrovní stresu (proměnná `fstress3`).

4 Kvantitativní a kvalitativní proměnná

Podívejme se nyní, jak spolu souvisí doba cvičení (kvantitativní proměnná `exercise`) a pohlaví (kvalitativní proměnná `gender`).

Ujistěte se, že proměnnou `gender` vnímá R jako faktor:

```
class(gender)
```

Pokud ne, tak proveďte:

```
detach(Sleep)
Sleep$gender <- as.factor(Sleep$gender)
attach(Sleep)
```

- 1) Spočtěme popisné statistiky pro dobu cvičení v závislosti na pohlaví.

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

- 2) Nakresleme krabíkové grafy doby cvičení podmíněné pohlavím (jedna krabíčka pro každou skupinu pohlaví). Povšimněte si „vlnkové“ notace!

```
plot(exercise ~ gender, col="seagreen", xlab="Pohlaví", ylab="Doba cvičení (minuty)")
```

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

3) Dalším užitečným obrázkem je graf průměrů (plot of means with error bars).

```
install.packages("RcmdrMisc") # instalace knihovny
library(RcmdrMisc)           # otevření knihovny
plotMeans(exercise, gender, error.bars = "sd") # graf průměru se smer. odchylkami
plotMeans(exercise, gender, error.bars = "se") # se středními chybami průměrů
plotMeans(exercise, gender, error.bars = "conf.int") # s intervaly spolehlivosti
```

✧ Význam střední chyby průměru (standard error of the mean, zkr. SEM) jsme si vysvětlovali minule. Detailně je popsán v článku na mých stránkách.

✧ Domníváte se, že se doba cvičení liší mezi pohlavími?

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

Samostatná práce



Pomocí vhodných popisných statistik a obrázků vyšetřete závislosti mezi:

- 1) Dobou práce (work_hours) a pohlavím (gender);
- 2) Cvičením (exercise) a náladou (fmood3).

✧ 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 v různých skupinách pomocí **vícevýběrových testů** (analýza rozptylu, Kruskalův-Wallisův test).

5 Vytvoření podmnožiny dat

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

- 1) Zjistíme, kteří jedinci jsou mladší 30 let:

```
which(age < 30)
```

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

```
subset(Sleep, age < 30)
```

- 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 nedorozuměním plynoucím ze shodných názvů proměnných ve dvou datech – původních a podmnožiny).

```
detach(Sleep)
```

- 4) Řekněme, že dále budeme chtít pracovat pouze s jedinci mladšími 30 let. Vytvořenou podmnožinu si můžeme uložit do datové tabulky Sleep30.

```
Sleep30 <- subset(Sleep, age < 30)
```

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

```
save(Sleep30, file = "data/Sleep30.RData")
```



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

- (a) Jedinci, u kterých cvičili alespoň 60 minut.
- (b) Jedinci s dobrou náladou.
- (c) Ženy.
- (d) Muži, kteří pracovali víc než 8 hodin.
- (e) Jedinci mezi 25 a 35 lety.

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

✧ `gender == "female"`

✧ `exercise > 60`

✧ `gender != "male"`

✧ `!(exercise <= 59)`

✧ `age <= 35 | age >= 25`

✧ `mood_score >= 7`

✧ `!(age > 25 & age < 35)`

✧ `gender == "male" & work_hours > 8`

✧ `fmood3 == "good"`

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

6 Konec práce

Než zavřete všechna okna, nezapomeňte si uložit poslední změny ve skriptovém souboru:

File ➡ **Save**

nebo klávesovou skratkou **Ctrl-s**.