Cvičení z matematické statistiky 04

Poslední úprava dokumentu: 11. března 2025.

Tvorba podmnožiny dat, zkoumání tvaru rozdělení (šikmost, špičatost, histogram), klasická definice pravděpodobnosti, výpočet EX a $var(X)$

Spusťte RStudio, otevřete si nový skript, uložte si ho a nastavte pracovní adresář.

Rozcvička

Připomeňte si, jak lze použít R jako kalkulačku, a spočítejte následující výrazy:

$$\left(\frac{1}{2}\right)^{10}, \quad \exp(-5), \quad 5!, \quad \frac{5^3}{3!}, \quad \binom{6}{2}.$$

```
(1/2)^(10)
exp(-5)
factorial(5)
5^3/factorial(3)
choose(6,2)
```

Data

Na dnešním cvičení využijeme datovou tabulku `iris`, která je k dispozici v R-kové knihovně `datasets`. Data získáte pomocí těchto příkazů:

```
install.packages("datasets")
library(datasets)
View(iris)
```

Tato data byla nasbírána Edgarem Andersonem v roce 1935 a obsahují naměřené údaje o třech druzích kosatců (iris setosa, iris versicolor a iris virginica). Naměřeny byly tyto veličiny:

- Sepal.Length (délka lístku kalicha v cm)
- Sepal.Width (šířka lístku kalicha v cm)
- Petal.Length (délka okvětního lístku v cm)
- Petal.Width (šířka okvětního lístku v cm)
- Species (druh kosatce)

1 Vytvoření podmnožiny dat

Někdy je potřeba zpracovávat pouze podmnožinu dat, jež splňuje nějakou podmínku (např. zajíma nás pouze určitý druh kosatce, nebo pouze jedinci s kališním lístkem delším 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 odpovídají odrůdě „virginica“:

```
which(iris$Species == "virginica")
```

2) Řekněme, že dále budeme chtít pracovat pouze s jedinci odrůdy virginica. Vytvořenou podmnožinu si můžeme uložit do datové tabulky Virginica.

```
Virginica <- subset(iris, Species == "virginica")
```

Poznámka: *V případě, že máte data připojená (tj. provedli jste příkaz* `attach(iris)`*) a chcete podmnožinu původních dat ukládat a dále s ní pracovat, doporučuji opět odpojit přístup k proměnným původních dat, tj. proveďte* `detach(iris)` *(vyhnete se tak možným nedorozumněním plynoucím ze shodných názvů proměnných ve dvou datech – původních a podmnožiny).*

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

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

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

(a) Jedinici odrůdy *virginica* a *versicolor* (obě odrůdy dohromady).

(b) Jedinci, kteří mají okvětní lístek delší než 5 cm.

(c) Jedinici odrůdy *setosa*, kteří mají kališní lístek kratší než 5 cm.

(d) Jedinci s okvětním lístkem délky 5 - 6 cm.

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

- `Species == "setosa"`
- `Petal.Length > 5`
- `Species != "setosa"`
- `!(Petal.Length <= 5)`
- `Petal.Length <= 5 | Petal.Length>= 6`
- `Petal.Length > 5 & Petal.Length< 6`
- `!(Petal.Length > 5 & Petal.Length < 6)`
- `Species == "setosa"& Sepal.Length < 5`

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

Nadále již v datech `iris` nebudeme provádět žádné úpravy. Pro větší pohodlí si proto připojme všechny proměnné této datové tabulky:

```
attach(iris)
```

2 Šikmost a špičatost aneb tvar rozdělení kvantitativního znaku

Kromě charakteristik polohy a variability se může hodit vědět více o tvaru rozdělení dané kvantitativní veličiny (tj. mít představu, které hodnoty se objevují více a které méně často). K tomu lze využít následující charakteristiky.

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 šířky kališního lístku (uvažujme všechny odrůdy dohromady)

```
hist(Sepal.Width)
hist(Sepal.Width, breaks = 12)  # lze nastavit počet intervalů
```

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(Sepal.Width, prob=TRUE)
```

✧ Histogram si lze dále vylepšit pomocí známých grafických argumentů:

```
hist(Sepal.Width, prob=TRUE, col="slateblue",
     xlab="šířka kališního lístku", ylab="Hustota", main="Kosatec (3 odrůdy)")
```

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

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

Šikmost a špičatost

Skutečné teoretické rozdělení daného znaku (např. šířky kališního lístku) 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 šířku kališního lístku (`Sepal.Width`).
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(Sepal.Width)^3)          # sikmost
mean(scale(Sepal.Width)^4) - 3      # spicatost
```

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

✧ Je hodnota výběrové šikmosti v souladu s tvarem histogramu pro šířku kališního lístku z předchozí části?

3 Klasická definice pravděpodobnosti

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?

4 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))

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

✧ Někdy je výhodnější vypočítat rozptyl podle vzorečku

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

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

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

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 $\mathrm{E}\, X$.

5 Konec práce

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