MAS10c: První zápočtový test (sk. A)

2022-10-24

Příklad 1 [40 b.]

Pracujte s datovým souborem `pr-1-data.csv`, obsahujícím záznamy o barvě vlasů a očí u mužů a žen. Soubor obsahuje následující znaky:

- `sex` – identifikátor pohlaví (`Female` – žena, `Male` – muž);
- `hair` – barva vlasů (`Black` – černá, `Blond` – blond, `Brown` – hnědá, `Red` – zrzavá);
- `eye` – barva očí (`Blue` – modrá, `Brown` – hnědá, `Green` – zelená, `Hazel` – oříškově hnědá).

Vypracujte následující úkoly, své textové odpovědi zapisujte do komentářů v kódu **bez** použití diakritiky.

a. [1 b.] Načtěte datový soubor do proměnné `data`.

b. [2 b.] Vypiště prvních *pět* záznamů proměnné `data`.

```
##      sex  hair   eye
## 1   <NA> Brown Hazel
## 2 Female Brown Hazel
## 3   Male Brown  Blue
## 4   Male Brown Brown
## 5   Male   Red  Blue
```

c. [2 b.] Jaké rozměry má datová tabulka `data`? Rozměry interpretujte.

d. [5 b.] Zjistěte počet chybějících hodnot v datové tabulce `data`. Chybějící hodnoty odstraňte.

e. [4 b.] Nadále budeme uvažovat pouze záznamy o můžích (`Male`). Tyto záznamy vyfiltrujte a uložte do proměnné `data.m`.

f. [10 b.] Vytvořte kontingenční tabulky absolutních a relativních simultánních četností pro znaky indikující barvu očí (`hair`) a vlasů (`eye`).

```
k.table.abs
```

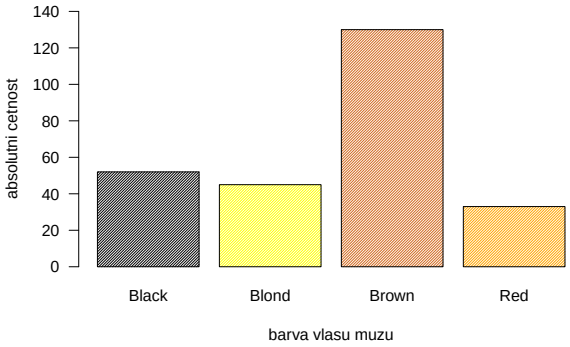
```
##       Blue Brown Green Hazel suma
## Black   10    31     2     9   52
## Blond   29     3     8     5   45
## Brown   45    47    14    24  130
## Red     10    10     6     7   33
## suma    94    91    30    45  260
```

```
k.table.rel
```

```
##         Blue  Brown  Green  Hazel   suma
## Black 0.0385 0.1192 0.0077 0.0346 0.2000
## Blond 0.1115 0.0115 0.0308 0.0192 0.1731
## Brown 0.1731 0.1808 0.0538 0.0923 0.5000
## Red   0.0385 0.0385 0.0231 0.0269 0.1269
## suma  0.3615 0.3500 0.1154 0.1731 1.0000
```

g. [6 b.] Obdržené výsledky z předchozího úkolu **interpretujte**. Datový soubor `data.m` obsahuje záznamy o barvě vlasů a očí u . . . . . mužů. Modrookých blonďatých mužů je zde celkem . . . . . ., což odpovídá . . . . . % záznamů datového souboru `data.m`. 50 % můžu má . . . . . barvu vlasů. Nejčastější barva očí mezi vybranými muži je . . . . . ., kterou nalezneme v . . . . % záznamů.

h. [10 b.] Vykreslete sloupcový graf znaku `hair` z datové tabulky `data.m` dle následujícího vzoru. Hodnoty některých atributů naleznete v kódu.



i. **BONUS:** [10 b.] Rozhodněte, o jaký typ kontingenční tabulky se jedná, svou odpověď zdůvodněte. Tuto kontingenční tabulku vytvořte. Interpretujte dvě čísla z dané tabulky.

```
##
##           Blue  Brown  Green  Hazel
##   Black 0.1923 0.5962 0.0385 0.1731
##   Blond 0.6444 0.0667 0.1778 0.1111
##   Brown 0.3462 0.3615 0.1077 0.1846
##   Red   0.3030 0.3030 0.1818 0.2121
```

Příklad 2 [60 b.]

Pracujte s datovým souborem `pr-2-data.txt`, obsahujícím záznamy o rozměrech okvětních lístků trojice druhů kosatců. Soubor obsahuje následující znaky:

- `Sepal.Length` – délka sepálu (cm);
- `Sepal.Width` – šířka sepálu (cm);
- `Petal.Length` – délka petálu (cm);
- `Petal.Width` – šířka petálu (cm);
- `Species` – druh kosatce (`setosa`, `virginica`, `versicolor`).

Vypracujte následující úkoly, své textové odpovědi zapisujte do komentářů v kódu **bez** použití diakritiky.

a. [1 b.] Načtěte datový soubor do proměnné `data`.

b. [1 b.] Vypiště prvních *pět* záznamů proměnné `data`.

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 1          4.6         3.1          1.5         0.2     setosa
## 2          6.4         2.9          4.3         1.3 versicolor
## 3          4.8         3.0          1.4          NA     setosa
## 4          4.8         3.0          1.4         0.3     setosa
## 5          6.3         2.7          4.9         1.8  virginica
```

c. [2 b.] Odstraňte záznamy s chybějícími hodnotami.

d. [6 b.] Zjistěte počty záznamů s následujícími vlastnostmi:

1. druh *iris versicolor* a `Sepal.Length` alespoň 6 cm;
2. druh *iris versicolor* a `Sepal.Length` $\in (5, 6)$;
3. druh *iris versicolor* nebo *iris setosa*.

e. [5 b.] Nadále budeme pracovat pouze se znaky `Sepal.Length` a `Petal.Length` u druhu *iris setosa*. Tyto hodnoty vyfiltrujte a uložte do proměnné `data.s`. Sloupce této tabulky uložte do proměnných `data.SL` a `data.PL`.

f. [15 b.] U proměnné `Sepal.Length` vypočtěte následující číselné charakteristiky: průměr, směrodatnou odchylku, minimální a maximální hodnotu, dolní a horní kvartil, medián, interkvartilové rozpětí, šikmost a špičatost (využijte proměnnou `data.SL`). Vypočtené hodnoty zapište do tabulky `tab`:

```
(tab <- round(tab, digits = 2))
```

```
##            m    s min dolni.kv median horni.kv max IQR sikmost spicatost
## setosa 5.01 0.37 4.3      4.8      5      5.2 5.8 0.4    0.12     -0.59
```

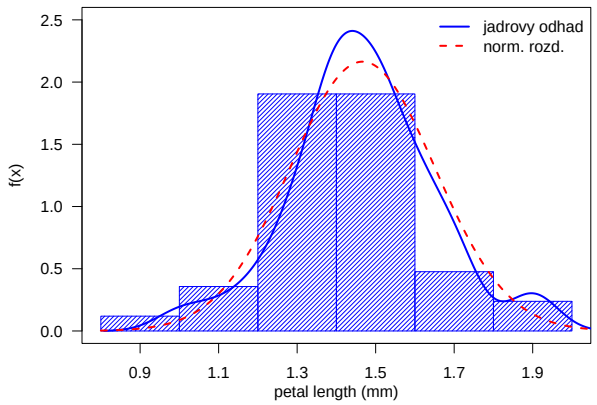
g. [6 b.] Obdržené výsledky z předchozího úkolu **interpretujte**. Délka sepálu (`Sepal.Length`) u druhu *iris setosa* se pohybovala v rozmezí od . . . . cm do . . . . cm. 25 % hodnot je menší nebo rovno . . . . cm. 75 % hodnot je menší nebo rovno . . . . cm. Medián značí hodnotu, pro kterou platí, že . . . . % hodnot délky sepálů je . . . . . této hodnotě.

h. [6 b.] Za předpokladu, že se délka petálů (`Petal.Length`) druhu *iris setosa* řídí normálním rozdělením, odhadněte parametry tohoto rozdělení. Odhadněte hodnotu výběrového koeficientu korelace mezi znaky `Sepal.Length` a `Petal.Length`.

i. [8 b.] Za stejných předpokladů jako v předchozím úkolu odhadněte pravděpodobnost, že délka petálu

1. nabývá hodnoty menší nebo rovny 1.5 cm;
2. je větší než 1.7 cm;
3. nabývá hodnoty právě 1.77 cm;
4. je větší než 1.3 cm, ale menší než 1.8 cm.

j. [10 b.] Vykreslete histogram znaku `Petal.Length` z datové tabulky `data.s` dle následujícího vzoru. Modrá křivka vyjadřuje jádrový odhad hustoty proměnné `Petal.Length`, červená křivka značí odhad hustoty na základě předpokladů v úkolu (h). Hodnoty některých atributů naleznete v kódu.



k. **BONUS:** [5 b.] Vykreslete krabicový graf znaku `Sepal.Length` z datové tabulky `data.s`.