MAS10c: Druhý zápočtový test (sk. B)

2022-12-15

Vypracujte následující úkoly. Své **textové odpovědi** nebo **výběr z možností** zapisujte do vytisknutého odpovědního archu. Důležitou součástí testu je také vypracování **R-skriptu** s výpočty podkládajícími vaše odpovědi. Svůj postup můžete doplnit o komentáře v kódu **bez** použití diakritiky.

Příklad 1 [50 b.]

Pracujte s datovým souborem `B-pr-1-data.txt`, obsahujícím záznamy o krevním tlaku v mužské populaci. Soubor obsahuje následující znaky:

- `systolic` – systolický tlak (mm Hg);
- `diastolic` – diastolický tlak (mm Hg).

Zadání: Pracujte s proměnnou `systolic` popisující naměřené hodnoty systolického tlaku v mužské populaci. Na hladině významnosti $\alpha = 0.05$ zjistěte, zda je možné se domnívat, že systolický tlak u mužů je větší než pro ženskou populaci ($m.F = 118.47$, $s.F = 14.33$, $n.F = 20\,291$).

a. [4 b.] Načtěte datový soubor do proměnné `data`, vyfiltrujte znak `systolic` a odstraňte záznamy s chybějícími hodnotami.

```
data <- read.delim(...)
data.MS <- ...
```

b. [3 b.] Náhodný výběr obsahuje …….. záznamů systolického tlaku z mužské populace. Naměřené hodnoty se pohybují v rozmezí ……–…… mm Hg.

```
n <- ...
(tab <- data.frame(...))
```

```
##    n min max
## 1 49  90 160
```

c. [3 b.] Ze zadání máme za úkol porovnat střední hodnotu systolického tlaku mužské populace s hodnotou pro ženskou populaci, použijeme tedy test o střední hodnotě / test o rozptylu / párový test / test o korelačním koeficientu. Primárně bychom chtěli použít **parametrický** test. Nutným předpokladem parametrického testu je **normalita naměřených hodnot**.

d. [10 b.] **Test normality** naměřených hodnot proměnné `systolic`

- H_0 : Data …… z normálního rozdělení.
- H_1 : Data …… z normálního rozdělení.

Hladina významnosti $\alpha =$ …….. . $n =$ …… je větší než 30 a menší / větší než 100 → Shapirův–Wilkův / Andersonův–Darlingův / Lillieforsův test.

```
##       p.value
## 1 0.008740651
```

Náhodný výběr systolického tlaku u mužské populace …… z normálního rozdělení (p-hodnota = …… je menší / větší než $\alpha = 0.05$). Protože naměřené hodnoty

systolického tlaku u mužské populace nepochází z normálního rozdělení, použijeme na otestování hypotézy ze zadání neparametrický test. Vhodný neparametrický test vybereme podle výsledku testu symetrie.

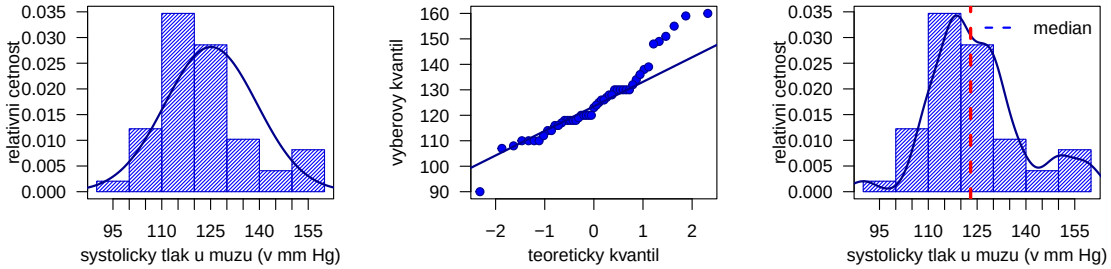
e. [10 b.] **Test symetrie**

- H_0 : Data ..... z rozdělení symetrického okolo mediánu.
- H_1 : Data ..... z rozdělení symetrického okolo mediánu.

Hladina významnosti $\alpha =$ ..... . Miaové test.

```
##     p.value
## 1 0.1407898
```

Náhodný výběr hodnot systolického tlaku u mužské populace ..... z rozdělení symetrického okolo mediánu (p-hodnota = ... je menší / větší než $\alpha = 0.05$). Protože data pochází ze symetrického rozdělení, použijeme **Wilcoxonův jednovýběrový test.**



f. [20 b.] **Wilcoxonův jednovýběrový test**

- H_0 : ..... .
- H_1 : ..... (..... alternativa).
- Hladina významnosti $\alpha =$ ..... .
- Stručně odůvodněte volbu H_0 a H_1.

```
x0 <- ...
alpha <- ...
m <- sum(...)
wilcox.test(..., correct = F)
q <- qsignrank(...)
(tab <- data.frame(q))
```

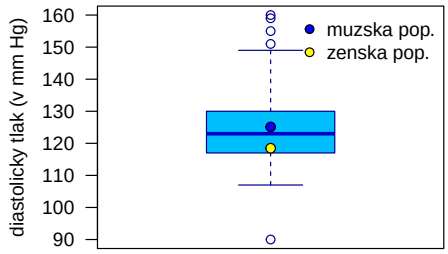
```
##       p.value
## 1 0.000966659
```

1. **Test kritickým oborem**
   Hodnota testovací statistiky $t_W =$ ....., kritický obor W má tvar .....
   Protože ....., H_0 ..... na hladině významnosti $\alpha =$ ..... .

2. **Test intervalem spolehlivosti**
   Interval spolehlivosti má tvar ..... Protože ....., H_0 ..... na hladině významnosti $\alpha =$ ..... .

3. **Test p-hodnotou**
   P-hodnota = ... Protože ....., H_0 ..... na hladině významnosti $\alpha =$ ..... .

Interpretace výsledků: Systolický tlak u mužů je / není statisticky významně větší než hodnota pro ženskou populaci. Jak podkládá následující graf vaše obdržené výsledky?



Příklad 2 [50 b.]

Pracujte s datovým souborem `B-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`).

Zadání: Pracujte s proměnnou `Sepal.Width` popisující šířku sepálu. Na hladině významnosti $\alpha = 0.01$ zjistěte, zda je důvodné se domnívat, že šířka sepálu u druhu `iris virginica` je větší než u druhu `iris versicolor`.

a. [4 b.] Načtěte datový soubor do proměnné `data` a odstraňte záznamy s chybějícími hodnotami.

```
data <- read.delim(...)
vir.SW <- ...
ver.SW <- ...
```

b. [6 b.] První náhodný výběr obsahuje ..... záznamů o šířce sepálů druhu `iris virginica`. Naměřené hodnoty se pohybují v rozmezí .....–..... cm. Druhý náhodný výběr obsahuje ..... záznamů o šířce sepálů druhu `iris versicolor`. Naměřené hodnoty se pohybují v rozmezí .....–..... cm.

```
n1 <- ...
n2 <- ...
(tab <- data.frame(...))
```

```
##   n1 n2 min1 max1 min2 max2
## 1 47 49  2.2  3.8    2  3.4
```

c. [2 b.] Ze zadání máme za úkol porovnat střední hodnoty šířky sepálů dvou druhů irisů, použijeme tedy párový test / test o rozdílu středních hodnot / test o rozdílu korelačních koeficientů. Primárně bychom chtěli použít **parametrický** test. Nutným předpokladem parametrického testu je **normalita naměřených hodnot** (zvlášť v každém výběru).

d. [7 b.] **Test normality** naměřených hodnot u `iris virginica`

- H_0 : Data ..... z normálního rozdělení.
- H_1 : Data ..... z normálního rozdělení.

Hladina významnosti $\alpha =$ ..... . $n =$ ..... je větší než 30 a menší / větší než 100 → Shapirův–Wilkův / Andersonův–Darlingův / Lillieforsův test.

```
##      p.value
## 1 0.2166587
```

Náhodný výběr šířky sepálů druhu `iris virginica` ..... z normálního rozdělení (p-hodnota = ... je menší / větší než $\alpha = 0.01$).

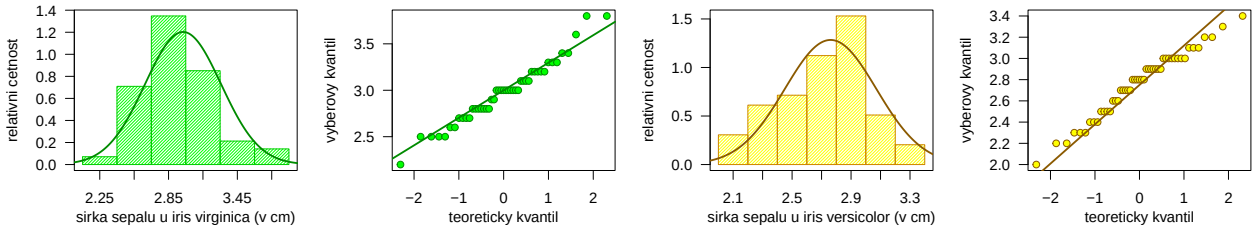
e. [7 b.] **Test normality** naměřených hodnot u `iris versicolor`

- H_0 : Data ..... z normálního rozdělení.
- H_1 : Data ..... z normálního rozdělení.

Hladina významnosti $\alpha =$ ..... . $n =$ ..... je větší než 30 a menší / větší než 100 → Shapirův–Wilkův / Andersonův–Darlingův / Lillieforsův test.

```
##      p.value
## 1 0.1286694
```

Náhodný výběr šířky sepálů druhu `iris versicolor` ..... z normálního rozdělení (p-hodnota = ... je menší / větší než $\alpha = 0.01$).



Protože oba výběry pochází z normálního rozdělení, použijeme na otestování hypotézy ze zadání **parametrický test**. Vhodný parametrický test vybereme v závislosti na výsledku testu o podílu rozptylů.

f. [9 b.] **Test o podílu rozptylů**

- H_0 : ..... → ..... .
- H_1 : ..... → ..... (..... alternativa).
- Hladina významnosti $\alpha =$ ..... .

```
##    p.value
## 1 0.654041
```

1. **Test kritickým oborem**
   Hodnota testovací statistiky $f_W =$ ....., kritický obor W má tvar .....
   Protože ....., H_0 ..... na hladině významnosti $\alpha =$ ..... .
2. **Test intervalem spolehlivosti**
   Interval spolehlivosti má tvar ..... Protože ....., H_0 ..... na hladině významnosti $\alpha =$ ..... .
3. **Test p-hodnotou**
   P-hodnota = ... Protože ....., H_0 ..... na hladině významnosti $\alpha =$ ..... .

Mezi rozptylem šířky sepálů druhu `iris virginica` a `iris versicolor` existuje / neexistuje statisticky významný rozdíl. Protože rozptyly obou výběrů jsou shodné, použijeme na otestování hypotézy ze zadání **klasický test o rozdílu středních hodnot** (rozptyly σ_1^2 a σ_2^2 jsou neznámé, ale shodné).

g. [15 b.] **Klasický test o rozdílu středních hodnot**

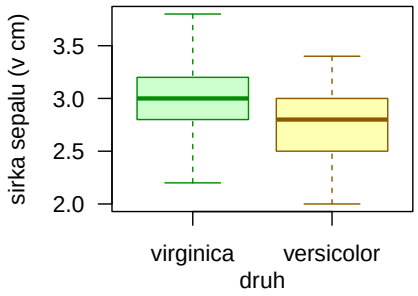
- H_0 : ..... .
- H_1 : ..... (..... alternativa).
- Hladina významnosti $\alpha =$ ..... .
- Stručně odůvodněte volbu H_0 a H_1.

```
alpha <- 0.01
t.test(...)
q <- qt(...)
(tab <- data.frame(q))
```

```
##         p.value
## 1 0.0007198713
```

1. **Test kritickým oborem**
   Hodnota testovací statistiky $t_W =$ ....., kritický obor W má tvar .....
   Protože ....., H_0 ..... na hladině významnosti $\alpha =$ ..... .
2. **Test intervalem spolehlivosti**
   Interval spolehlivosti má tvar ..... Protože ....., H_0 ..... na hladině významnosti $\alpha =$ ..... .
3. **Test p-hodnotou**
   P-hodnota = ... Protože ....., H_0 ..... na hladině významnosti $\alpha =$ ..... .

Interpretace výsledků: Šířka sepálů u druhu `iris virginica` je / není statisticky významně větší než u druhu `iris versicolor`. Jak podkládá následující graf vaše obdržené výsledky?



BONUS [15 b.]

Pracujte s datovým souborem `B-bonus-data.txt`, obsahujícím záznamy o kouření cigaret u mužů a žen:

- `sex` – pohlaví (`F` – žena, `M` – muž);
- `smoking` – indikátor, zda daná osoba kouří cigarety (`yes` – ano, `no` – ne).

Zadání: Pomocí testu podílem šancí zjistěte, zda na asymptotické hladině významnosti $\alpha = 0.01$ existuje statisticky významná závislost pohlaví a skutečnosti, zda daná osoba kouří cigarety.

a. [0 b.] Načtěte datový soubor do proměnné `data` a odstraňte záznamy s chybějícími hodnotami. Načtenou datovou tabulku převeďte do vhodného formátu.

```
data <- read.delim(...)
data <- ...
(data <- t(table(...)))
```

```
##        sex
## smoking  F  M
##     no  38 22
##     yes 11 37
```

b. [3 b.] **Formulace hypotéz**

- H_0 : Znaky `sex` a `smoking` ..... stochasticky nezávislé. → ..... → .....
- H_1 : Znaky `sex` a `smoking` ..... stochasticky nezávislé. → ..... → .....
- Hladina významnosti $\alpha =$ ..... .

c. [1 b.] **Podmínka dobré aproximace** Test podílem šancí můžeme provést, je-li splněna podmínka dobré aproximace (alespoň 80 % očekávaných četností je větších nebo rovných 5 a zbylých 20 % očekávaný četností neklesne pod hodnotu 2).

```
round(chisq.test(data, correct = F)$expected, 1)
```

```
##        sex
## smoking    F    M
##     no  27.2 32.8
##     yes 21.8 26.2
```

Podmínka dobré aproximace ..... splněna. Všechny teoretické četnosti jsou ..... než 5.

d. [3 b.] **Výpočet (logaritmu) podílu šancí**

```
a <- ...; b <- ...; c <- ...; d <- ...
OR <- ... / ...
lnOR <- log(...)
```

```
##         OR     lnOR
## 1 5.809917 1.759566
```

Podíl šancí $OR =$ ..... . Logaritmus podílu šancí $\ln(OR) =$ ..... .

e. [8 b.] **Test podílem šancí**

```
source('funkce-verze-03.R')
alpha <- ...
odds.ratio.test(...)
q1 <- ...
q2 <- ...
(tab <- data.frame(q1, q2))
```

```
##        p.value
## 1 5.348677e-05
```

1. **Test kritickým oborem**
   Hodnota testovací statistiky $t_0 =$ ....., kritický obor W má tvar .....
   Protože ....., H_0 ..... na hladině významnosti $\alpha =$ ..... .
2. **Test intervalem spolehlivosti**
   Interval spolehlivosti má tvar ..... Protože ....., H_0 ..... na hladině významnosti $\alpha =$ ..... .
3. **Test p-hodnotou**
   P-hodnota = ... Protože ....., H_0 ..... na hladině významnosti $\alpha =$ ..... .

Interpretace výsledků: Mezi pohlavím a skutečností, zda daná osoba kouří cigarety, existuje / neexistuje statisticky významná stochastická závislost.