MAS10c: Druhý zápočtový test (sk. A)

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 `A-pr-1-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 $\alpha = 0.10$ zjistěte, zda je důvodné se domnívat, že šířka sepálu u druhu `iris virginica` je menší než u druhu `iris setosa`.

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

```
data <- read.delim(...)
vir.SW <- ...
set.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 setosa`. Naměřené hodnoty se pohybují v rozmezí .....–..... cm.

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

```
##   n1 n2 min1 max1 min2 max2
## 1 44 42  2.2  3.8  2.9  4.4
```

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. [9 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 0.2394528
```

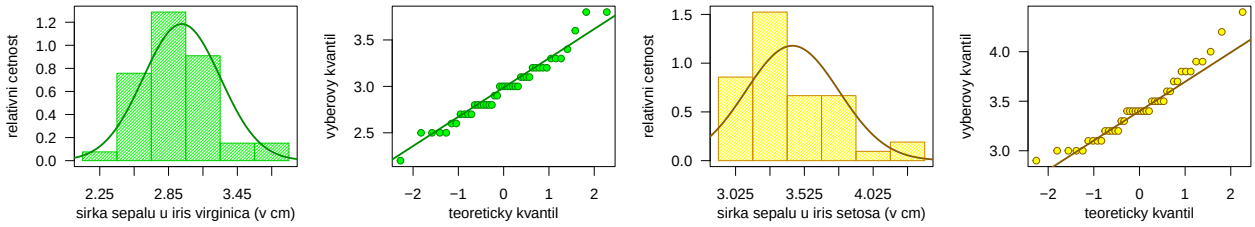
Náhodný výběr šířky sepálů druhu `iris virginica` ..... z normálního rozdělení (p-hodnota = ... je menší / větší než $\alpha = 0.10$).

e. [9 b.] **Test normality** naměřených hodnot u `iris setosa`
- 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 0.06889027
```

Náhodný výběr šířky sepálů druhu `iris setosa` ..... z normálního rozdělení (p-hodnota = ... je menší / větší než $\alpha = 0.10$). Protože naměřené hodnoty šířky sepálů druhu `iris setosa` nepochází z normálního rozdělení, použijeme na otestování hypotézy ze zadání neparametrický **Wilcoxonův dvouvýběrový test**.



f. [20 b.] **Wilcoxonův dvouvýběrový test**
- H_0 : ..... .
- H_1 : ..... (..... alternativa).
- Hladina významnosti $\alpha =$ ..... .
- Stručně odůvodněte volbu H_0 a H_1.

```
alpha <- 0.10
wilcox.test(..., exact = F, conf.int = T)
q <- qwilcox(...)
(tab <- data.frame(q))
```

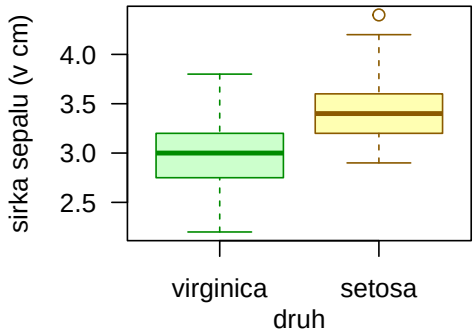
```
##      p.value
## 1 1.07732e-08
```

1. **Test kritickým oborem**
   Hodnota testovací statistiky $s_E =$ ....., 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 menší než u druhu `iris setosa`. Jak podkládá následující graf vaše obdržené výsledky?



Příklad 2 [50 b.]

Pracujte s datovým souborem `A-pr-2-data.txt`, obsahujícím záznamy o krevním tlaku v ženské populaci. Soubor obsahuje následující znaky:

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

Zadání: Pracujte s proměnnou `diastolic` popisující naměřené hodnoty diastolického tlaku v ženské populaci. Na hladině významnosti $\alpha = 0.05$ zjistěte, zda je možné se domnívat, že diastolický tlak u žen je větší nebo roven hodnotám obdrženým pro mužskou populaci ($m.M = 77.31$, $s.M = 9.44$, $n.M = 35\,401$).

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

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

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

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

```
##    n min max
## 1 37  60  90
```

c. [3 b.] Ze zadání máme za úkol porovnat střední hodnotu diastolického tlaku ženské populace s hodnotou pro mužskou 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 `diastolic`

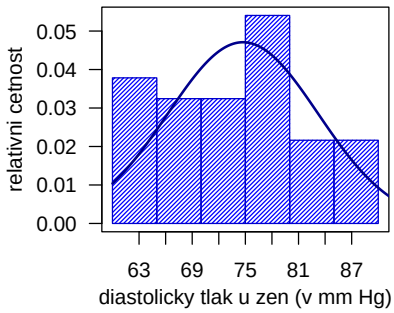
- H_0 : Data ..... z normálního rozdělení.
- H_1 : Data ..... z normálního rozdělení.

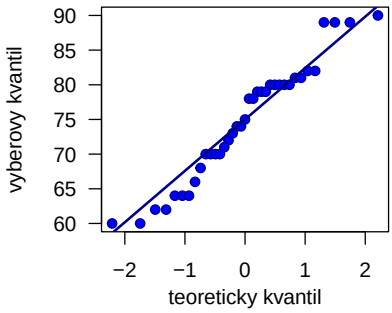
Hladina významnosti $\alpha =$ ..... . $n =$ ..... je větší než 30 a menší / větší než 100 → Shapirův–Wilkův / Andersonův–Darlingův / Lillieforsův test.

```
##     p.value
## 1 0.1212513
```

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

diastolického tlaku u ženské populace pochází z normálního rozdělení, použijeme na otestování hypotézy ze zadání parametrický **jednovýběrový test o střední hodnotě**.





e. [30 b.] **Klasický test o střední hodnotě**
- H_0 : ..... .
- H_1 : ..... (..... alternativa).
- Hladina významnosti α = ...
- Stručně odůvodněte volbu H_0 a H_1.

```
alpha <- 0.05
m.M <- ...
t.test(...)
q <- ...
(tab <- data.frame(q))
```

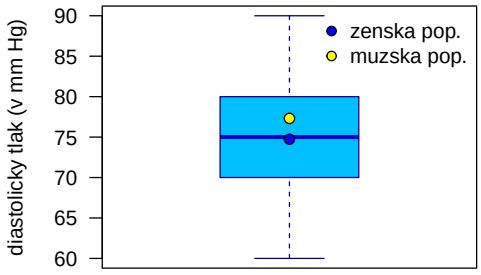
```
##      p.value
## 1 0.03606096
```

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

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

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

Interpretace výsledků: Diastolický tlak u žen je / není statisticky významně větší nebo roven hodnotě pro mužskou populaci. Jak podkládá následující graf vaše obdržené výsledky?



BONUS [15 b.]

Pracujte s datovým souborem `A-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 Pearsonova χ^2 testu 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.

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

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

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** Pearsonův χ^2 test 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)
```

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

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

d. [11 b.] **Pearsonův χ^2 test**

```
chisq.test(..., correct = F)
alpha <- ...
r <- ...
s <- ...
q <- qchisq(..., ...)
(tab <- data.frame(q))
```

```
##       p.value
## 1 2.76232e-05
```

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

Pro zjištění míry závislosti v kontingenční tabulce použijeme ..... koeficient.

```
V <- ...
```

Interpretace výsledků: Mezi pohlavím a skutečností, zda daná osoba kouří cigarety, existuje / neexistuje statisticky významná stochastická závislost. Mezi znaky `sex` a `smoking` existuje ..... stupeň závislosti ($V =$ .....).