Zápočtové domácí úkoly

z Matematické statistiky 1 NMSA331

1 Obecné pokyny

- Úlohy není zapotřebí psát v TeXu či jiném editoru. Stačí, když budou **čitelné**. Je přípustné kombinovat psaný text a vytištěný obrázek.
- Řešení úloh se odevzdávají na cvičení. Odevzdání emailem je možné jen výjimečně (v případě absence na daném cvičení), a to **ve formátu pdf**.
- **Okopírování výstupu z programu R nelze považovat za řešení!** R je pouze výpočetní prostředek, který spočítá čísla a případně nakreslí obrázky, které použijete ve svém vlastním textu.
- **Méně může znamenat více**. Při psaní zpráv se omezte na to podstatné. Není cílem ukázat, co všechno umíte spočítat. Naopak, čtenáři Vašich zpráv ocení, že je nezahlcujete čísly, která nejsou pro daný problém nezbytná. Podobně nezahlcujte čtenáře zbytečnými grafy a tabulkami. Uvádějte vždy jen ty podstatné.
- Při prezentaci výsledků používejte pouze **rozumný počet desetinných míst**.
- Statistické testy provádějte na 5% hladině významnosti, intervalové odhady konstruujte se spolehlivostí 95 %.
- **Interpretace výsledků statistických procedur musí být srozumitelná a správná**. Je třeba rozlišovat mezi tím, co je a co není náhodné (v klasické statistické indukci). Náhodná jsou data (která si představujeme jako realizace náhodných veličin) a vše, co je od nich odvozeno (testová statistika, rozhodnutí o platnosti nulové hypotézy, P-hodnota, meze intervalového odhadu). Skutečnost, že hypotéza platí však náhodná není. Stejně tak není náhodná skutečná hodnota parametru. Není tedy správné například říkat, že: *Pravděpodobnost, že nulová hypotéza platí, je menší než 5 %.*
- **Nepouštějte se do postupů, kterým nerozumíte! Nepište věty, kterým nerozumíte!** I když máte pocit, že něco takového se říkalo na přednášce nebo cvičení. Lépe je napsat toho méně a jenom to, čemu rozumím, než psát nesmysly.
- **Plagiátorství** zjištěné v kterémkoliv z dodaných souborů bude mít za následek nulový počet bodů. Změna formátování, resp. překlad z jednoho jazyka do druhého nevede k práci, kterou nelze považovat za plagiát! V případě obdržení dvou prací, z nichž jednu lze považovat za plagiát, si cvičící vyhrazuje právo nezjišťovat, kdo je primárním a kdo sekundárním autorem.

2 Postup

U každé úlohy dbejte na následující (nemusí to být v uvedeném pořadí a v případě, že neprovádíte testy, tak ne všechno je relevantní):

(a) Zformulujte vhodný **pravděpodobnostní model** a statistické **hypotézy** (pokud nějaké uvažujete).

(b) Vytvořte (alespoň jednu) **tabulku**, která vhodně numericky shrnuje data tak, aby bylo možné si z uvedených čísel udělat představu o problému, který nás zajímá.

(c) Pokud uvádíte intervalový odhad, nezapomeňte vždy uvádět, zda se jedná o přesný nebo asymptotický intervalový odhad.

(d) Nakreslete (alespoň jeden) vhodný **obrázek**, pomocí něhož si lze udělat **představu o platnosti testovaných hypotéz**. Obrázek okomentujte.

(e) Uveďte metodu (včetně **vzorce**), pomocí níž testujete, resp. počítáte interval spolehlivosti.

(f) Uveďte **rozdělení použité statistiky** (v případě testování stačí za platnosti nulové hypotézy). Jde o přesné či asymptotické rozdělení?

(g) Uveďte **hodnotu testové statistiky**, **P-hodnotu testu** (a vzorec, jak byla P-hodnota spočtena).

(h) **Závěr vyjádřete slovně** ve formě srozumitelné pro nestatistika (zejména pak bez použití spojení typu „(ne)zamítáme H_0“). Uvědomte si, že pokud zamítnete nějaké tvrzení, tak klienta zpravidla zajímá, co tedy místo zamítnutého tvrzení platí. Tedy pokud např. zamítnete hypotézu o tom, že střední hodnota je nějaké μ_0, tak klienta zpravidla zajímá, jestli je tedy střední hodnota větší nebo menší než μ_0. Podobně, pokud zamítnete nezávislost, tak klienta zajímá, jak by „závislost“ veličin mohla vypadat a jak by se dala charakterizovat.

(i) Zamyslete se nad daty a pokuste se zhodnotit, v čem by mohl Váš model použitý v (a) pokulhávat za realitou a jak závažné důsledky by to mohlo mít. *Pokud např. předpokládáte normální rozdělení, tak zhodnoťte, jak moc by vadilo, pokud by tento předpoklad nebyl splněn. Dejte si také pozor, že některé procedury vyžadují spojité rozdělení.*

Úloha č. 1 (do 4. 12. 2017, resp. 7. 12. 2017)

Do proměnné Hosi si načtěte data Hosi.txt. Znaky AAAA v příkazu set.seed níže nahraďte datem svých narozenin ve tvaru DDMM a spusťte následující příkazy.

```
set.seed(AAAA);
n <- sample(200:300, size=1);
data <- sample(Hosi$delka, size=n);
```

Proměnná data nyní obsahuje Vaše data o rozsahu n o výšce chlapců v jednom roce. Na jejich základě vyřešte následující úkoly:

1. Popište Váš datový soubor výšky chlapců pomocí vhodných charakteristik polohy a variability. Rozdělení dat ilustrujte vhodným grafem.

2. Bodově i intervalově odhadněte první kvartil výšky chlapců v jednom roce.
   (*Návod: Intervalový odhad odvoďte na základě popisu konstrukce intervalu spolehlivosti pro obecný kvantil, který byl na přednášce.*)

   Dle růstových grafů zveřejněných Státním zdravotním ústavem je první kvartil výšky chlapců v jednom roce roven 73 cm. Na základě odvozeného intervalového odhadu posuďte, zda jsou Vaše data v souladu s tímto tvrzením.

3. Potvrzují Vaše data domněnku, že střední výška chlapců v jednom roce je vyšší než 75 cm? Pokud ne, uveďte příslušný bodový i intervalový odhad střední výšky chlapců.

Úloha č. 2 (do 8. 1. 2018, resp. 4. 1. 2018)

Předběžná verze zadání. Zadání může být ještě modifikováno. Definitivní verze bude zveřejněna nejpozději do 14.12.

Do proměnné Davis si načtěte data Davis.txt, která obsahují údaje o mužích a ženách, kteří pravidelně navštěvují vybranou posilovnu. U každé osoby známe:

`Pohlavi`	pohlaví osoby,
`Vaha`	skutečná (naměřená) hmotnost osoby v kg,
`Vyska`	skutečná (naměřená) výška osoby v cm,
`RVaha`	hmotnost v kg, kterou osoba veřejně uvádí,
`RVyska`	výška v cm, kterou osoba veřejně uvádí.

Stejně jako v Úloze č. 1 znaky AAAA v příkazu set.seed nahraďte datem svých narozenin a spusťte následující příkazy:

```
set.seed(AAAA);
n <- sample(120:150, size=1);
indexy <- sample(1:200, size=n);
data <- Davis[indexy,];
```

Váš datový soubor data můžete považovat za reprezentativní vzorek populace osob navštěvujících posilovnu. Zodpovězte následující otázky o této populaci:

1. Dá se říci, že muži mají o více než 15 kg vyšší hmotnost než ženy?

2. Je pravda, že ženy si veřejně uvádějí nižší hmotnost, než ve skutečnosti mají?

3. Dá se říci, že více než 60 % mužů uvádí vyšší výšku než ve skutečnosti má?

Nezapomeňte při sepisování řešení na jednotlivé body v části Postup. Zejména, pokud v 1. nebo 2. je Vaše odpověď „ne“, tak nabídněte vlastní odhad (bodový a intervalový) dané skutečnosti.
Dále teoreticky vyřešte následující problém:

4. Nechť $(X_1, Y_1)^\mathsf{T}, \ldots, (X_n, Y_n)^\mathsf{T}$ je náhodný výběr z dvourozměrného normálního rozdělení s nenulovou a konečnou varianční maticí s korelačním koeficientem $\rho \in (-1, 1)$.

   Předpokládejme, že chceme provést test hypotézy $\mathsf{E}\, X = \mathsf{E}\, Y$ a namísto párového t-testu použijeme testovou statistiku přesného dvouvýběrového t-testu, tj.

$$T_n = \sqrt{\frac{n}{2}} \frac{\overline{X}_n - \overline{Y}_n}{\sqrt{\frac{n-1}{2n-2}\left(S^2_{n,X} + S^2_{n,Y}\right)}},$$

   kde $\overline{X}_n$, $S^2_{n,X}$ je výběrový průměr a výběrový rozptyl spočtený z veličin $X_1, \ldots, X_n$ a podobně $\overline{Y}_n$, $S^2_{n,Y}$.

   (a) Za předpokladu, že platí $\mathsf{E}\, X_1 = \mathsf{E}\, Y_1$ a $\mathsf{E}\, X_1^2 = \mathsf{E}\, Y_1^2$, odvoďte asymptotické rozdělení (pro $n \to \infty$) statistiky T_n.

   (b) Za jakých předpokladů je toto limitní rozdělení z (a) normované normální, tj. $\mathsf{N}(0, 1)$?

   (c) Co z toho plyne pro test nulové hypotézy $H_0 : \mathsf{E}\, X_1 = \mathsf{E}\, Y_1$ v případě, že mylně použijeme dvouvýběrový t-test místo párového t-testu?

   (d⋆) *Dobrovolná část:* Zvolte si nějaké $\rho \neq 0$. Nagenerujte si náhodný výběr o rozsahu $n = 100$ z normálního rozdělení $\mathsf{N}(\mathbf{0}, \mathbf{\Sigma})$, kde $\mathbf{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ pomocí příkazů

```
library(mvtnorm)
z=rmvnorm(n,mean=c(0,0),sigma=matrix(ncol=2,nrow=2,c(1,rho,rho,1)))
```

   Proměnná `z` má teď dva sloupce odpovídající X_i a Y_i. Otestujte shodu středních hodnot pomocí párového t-testu i (špatně) pomocí dvouvýběrového t-testu. Porovnejte výsledky (hodnotu testové statistiky a p-hodnotu).
   Dobrovolný dodatek: Proveďte celý experiment 1000 krát a spočtěte, v kolika procent případů jste zamítli nulovou hypotézu pro párový t-test a v kolika procentech pro dvouvýběrový t-test. Opět porovnejte výsledek s teoretickou úvahou z (c).