MATEMATICKÁ STATISTIKA 1 (2016–17)

Zápočtové domácí úkoly (společné pro obě paralelky)

Obecné pokyny

- Úlohy se odevzdávají vždy **na začátku** Vašeho cvičení (Vašemu cvičícímu, případně jeho záskoku).
- Úlohy není zapotřebí psát v TeXu či jiném editoru. Stačí, když budou **čitelné**. Je přípustné kombinovat psaný text a vytištěný obrázek.
- Odevzdání řešení úlohy emailem je možné **ve formátu pdf**. Pouze ve v výjimečných případech bude akceptován jiný formát. V tomto případě však nemáte zaručeno, jak bude vypadat výsledek po vytisknutí. Soubory posílané emailem, prosím, pojmenujte svým **příjmením a číslem úlohy**, např. `novak-2.pdf`. Do předmětu (Subject) emailu napište `NMSA331`.
- Pokud úloha vyžaduje číslo Vaší studentské karty, uveďte na začátku úlohy také toto číslo (**a samozřejmě také svoje jméno!**).
- **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.
- **Okopírování výstupu z programu R nelze považovat za řešení!** R je pouze výpočetní prostředek, který spočítá čísla a případně nakreslí obrázky, které používáte ve Vašem textu.
- Pokud není řečené jinak, statistické testy provádějte na 5% hladině významnosti.
- Pokud není řečené jinak, intervalové odhady konstruujte se spolehlivostí 95 %.
- Při prezentaci výsledků používejte pouze **rozumný počet desetinných míst**.
- **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 %.*
- **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 zbytečně čtenáře grafy a tabulkami.
- **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.

Postup

U každé úlohy dbejte na následující (nemusí to být v uvedeném pořadí a v případě, že neprovádíte testy, tak ne všechno je relevantní[1]):

(a) Zformulujte vhodný **pravděpodobnostní model** a statistické **hypotézy**.

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

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

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

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

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

(g) **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.

(h) Zamyslete se nad daty a pokuste se zhodnotit, v čem by mohly Váš model použitý v (a) mohl pokulhávat za realitou a jak závažné důsledky by to mohlo mít. Případně se pokuste o alternativní řešení. *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.*

[1] Pokud sestavujete pouze interval spolehlivosti, tak body (c), (f), (g) lze úplně vynechat. Dále by nedávalo smysl v (a) specifikovat hypotézy.

Úloha č. 1 (do 7. 12. 2016 nebo 8. 12. 2016)

Načtěte si data ze cvičení Hosi.RData pomocí (modifikace) následujícího příkazu.

```
load("Hosi.RData")
```

Znaky `AAA` v příkazu `set.seed` nahraďte číslem své studentské karty a spusťte následující příkazy.

```
set.seed(AAA);
n <- sample(200:300, size=1);
X <- sample(Hosi0$hmotnost, size=n);
```

Proměnná `X` nyní obsahuje Vaše data o rozsahu n o hmotnosti chlapců ve 12 měsících.

Úkoly:

(i) Popište data `X` pomocí vhodných charakteristik polohy a variability.

(ii) Bodově i intervalově odhadněte, kolikrát je větší/menší pravděpodobnost, že chlapec má ve 12 měsících hmotnost menší než 10 kg než pravděpodobnost, že má více než 10 kg.

(iii) Bodově i intervalově odhadněte medián hmotnosti chlapců. (*Vzpomeňte na Příklad 38 ze cvičebnice.*)

Úloha č. 2 (do 4. 1. 2017 nebo 5. 1. 2017)

Podobně jako na cvičení načtěte soubor Kojeni.csv. Stejně jako v Úloze č. 1 znaky `AAA` v příkazu `set.seed` nahraďte číslem své studentské karty a spusťte následující příkazy.

```
set.seed(AAA);
n <- sample(70:90, size=1);
indexy <- sample(1:99, size=n);
subKojeni <- Kojeni[indexy,];
```

Na základě dat `subKojeni` odpovězte následující otázky.

(i) Dá se říci, že matky jsou v průměru o dva roky mladší než otcové?

(ii) Dá se říci, že více než 75 % dětí svou váhu během prvních 24 týdnů alespoň zdvojnásobí?

(iii) Dá se říci, že věk matky souvisí s tím, zda dítě bylo či nebylo plánováno? Pokud ano, pokuste se tuto souvislost nějak blíže kvantifikovat.

Úloha č. 3 (do 4. 1. 2017 nebo 5. 1. 2017)

Předpokládejte, že sledujete nezávislé stejně rozdělené náhodné vektory $(X_1, Y_1)^\mathsf{T}, \ldots, (X_n, Y_n)^\mathsf{T}$ s nenulovou a konečnou varianční maticí. Předpokládejme, že místo párového t-testu použijte testovou statistiku asymptotického dvouvýběrového t-testu, tj.

$$T_n^W = \frac{\overline{X}_n - \overline{Y}_n}{\sqrt{\frac{S_{n,X}^2}{n} + \frac{S_{n,Y}^2}{n}}},$$

kde $\overline{X}_n$, $S^2_{n,X}$ je výběrový průměr a výběrový rozptyl spočtený z veličin $X_1, \ldots, X_n$ a podobně $\overline{Y}_n$, $S^2_{n,Y}$. Za předpokladu, že $\mathsf{E}\, X_1 = \mathsf{E}\, Y_1$ odvoďte asymptotické rozdělení (pro $n \to \infty$) statistiky T_n^W. Za jakých předpokladů je toto limitní rozdělení normované normální, tj. $\mathsf{N}(0,1)$? 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?