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. 2018)

Data soccer.csv pocházejí z roku 2011 a obsahují vybrané charakteristiky profesionálních hráčů fotbalu. Pro každého hráče známe:

`Name`	jméno,
`Position`	pozici hráče (Defender, Forward, Goalkeeper, Midfielder),
`Nationality`	národnost,
`Age`	věk (v letech),
`Height`	výška v cm,
`Weight`	hmotnost v kg.

Načtěte si nejprve data pro proměnné `soccer`. Následně provedete svůj osobní náhodný výběr z tohoto souboru, a to následujícím způsobem: 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(soccer$Height, size=n);
```

Proměnná `data` nyní obsahuje Vaše data o rozsahu n o výšce profesionálních fotbalistů. Na jejich základě vyřešte následující úkoly:

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

2. Bodově i intervalově odhadněte 10% kvantil výšky profesionálních fotbalistů.
   (*Návod: Intervalový odhad odvoďte na základě popisu konstrukce intervalu spolehlivosti pro obecný kvantil, který byl na přednášce.*)

   Na základě odvozeného intervalového odhadu posuďte, zda jsou Vaše data v souladu s tvrzením, že 10% kvantil je menší než 170 cm.

3. Potvrzují Vaše data domněnku, že je střední výška fotbalistů vyšší než 180 cm? Kromě odpovědi na otázku uveďte také příslušný bodový i intervalový odhad střední výšky hráčů.

Dále vyřešte následující problém:

4. Výrobce uvádí, že balíček mandlí v čokoládě váží 200 g, my ale máme podezření, že to je méně. Plánujeme tedy provést experiment, kterým bychom rádi prokázali domněnku, že je hmotnost balíčku statisticky významně nižší než uvádí výrobce. Test budeme provádět na testovací hladině 5 % a budeme u něho předpokládat, že hmotnost balíčku se řídí normálním rozdělením se směrodatnou odchylkou 5 g. Na základě dosavadního zkoumání si myslíme, že by hmotnost balíčku mohla být zhruba o 20 g menší než je uváděno. Kolik balíčků mandlí máme zakoupit a zvážit, aby pravděpodobnost, že se nám podaří prokázat naši domněnku, byla alespoň 90 %? Odpověď řádně zdůvodněte.

Úloha č. 2 (do 8. 1. 2019)

Použijte opět data soccer.csv. Opět proveďte nejprve vlastní náhodný výběr z dat: Znaky AAAA v příkazu set.seed nahraďte datem svých narozenin a spusťte následující příkazy:

```
set.seed(AAAA);
n <- sample(200:300, size=1);
indexy <- sample(1:1851, size=n);
data <- soccer[indexy,];
```

Váš datový soubor `data` můžete považovat za reprezentativní vzorek profesionálních hráčů fotbalu. Pomocí Vašich dat zodpovězte níže uvedené otázky. **Nezapomeňte** při sepisování řešení na jednotlivé body popsané v části **Postup**. Ve všech bodech uveďte **kromě odpovědi** také stručný **popis dat**, které k odpovědi využíváte, a také **odhad** (bodový a intervalový) dané skutečnosti.

1. Dá se říci, že jsou útočníci vyšší než obránci?
2. Je pravda, že hmotnost hráče nesouvisí s tím, zda je starší než 26 let?
3. Rozhodněte, zda je pravda, že hráči váží méně než je jejich výška v cm mínus 110 cm.
4. Je pravda, že hráči starší než 30 let tvoří méně než pětinu všech hráčů?
   (*Návod: S chybějícími daty si poradíte např. pomocí volby* `na.rm` *v příslušné funkci nebo pomocí funkce* `na.omit`.)

Dále teoreticky vyřešte následující problém:

5. Nechť $X_1, \ldots, X_n$ a $Y_1, \ldots, Y_m$ jsou dva nezávislé náhodné výběry z neznámých (obecně různých) rozdělení s konečnými a nenulovými rozptyly σ_X^2 a σ_Y^2.

   Chceme provést test hypotézy rovnosti rozptylů těchto dvou rozdělení. K tomu použijeme asymptotickou verzi dvouvýběrového F-testu shody rozptylů, jehož testová statistika je

   $$F_{n,m} = \frac{S_X^2}{S_Y^2},$$

   pro S_X^2 a S_Y^2 výběrové rozptyly našich dvou náhodných výběrů. Předpokládejme, že pro rozsahy náhodných výběrů n a $m = m_n$ platí $n/m_n \to \lambda$ pro $n \to \infty$, kde $\lambda \in (0, \infty)$ je neznámá hodnota.

   A. Najděte asymptotické rozdělení statistiky $F_{n,m}$ pro $n, m \to \infty$. Uveďte případné další předpoklady na rozdělení obou výběrů, které pro odvození rozdělení potřebujete.
   Návod: Jedná se o podíl dvou nezávislých veličin, jejichž asymptotické rozdělení znáte.

   B. Pokud je potřeba, upravte statistiku $F_{n,m}$ tak, abyste na základě tohoto rozdělení mohli konstruovat test hypotézy $H_0 : \sigma_X^2 = \sigma_Y^2$ proti alternativě $H_1 : \sigma_X^2 \neq \sigma_Y^2$ na hladině α.
   Návod: Závisí-li asymptotické rozdělení $F_{n,m}$ na neznámých parametrech, vhodně tyto parametry odhadněte.

   C. Jak vypadá asymptotické rozdělení z A. pro speciální případ, kdy oba výběry pochází ze dvou (obecně různých) normálních rozdělení?

   D. Za jakých podmínek je přípustné využít pro test hypotézy $H_0 : \sigma_X^2 = \sigma_Y^2$ proti alternativě $H_1 : \sigma_X^2 \neq \sigma_Y^2$ kritický obor zkonstruovaný za předpokladu normality obou výběrů (tj. založený na asymptotickém rozdělení z C.)?