Zápočtové domácí úkoly

z Matematické statistiky 1 NMSA331

1 Obecné pokyny

- Řešení úloh se **odevzdávají na cvičení**. Odevzdání emailem je možné jen výjimečně, a to pouze text zpracovaný na počítači ve formátu pdf.
- Statistické testy provádějte na 5% hladině významnosti, intervalové odhady konstruujte se spolehlivostí 95 %.
- Ve svém řešení nezapomeňte uvést svůj vlastní čtyřmístný kód pro generování vlastního náhodného výběru, viz níže.

Úloha č. 1 (do 28.11. 2019 a 2.12.2019) — 40 bodů

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);
indexy <- sample(1:1851, size=n);
data <- soccer[indexy,];
```

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

1. Spočtěte BMI (body mass index) fotbalistů daný vztahem

$$\text{BMI} = \frac{\text{hmotnost [kg]}}{(\text{výška [m]})^2}$$

   a popište tuto veličinu ve Vašem datovém souboru pomocí vhodných statistických charakteristik a grafů.
   Popis provádíte pro člověka nestatistika, který Váš datový soubor nikdy neviděl a měl by si tudíž z Vašeho popisu udělat dostatečnou představu o Vašich datech. [10 bodů]

2. Uveďte horní intervalový odhad pro medián BMI hráčů.
   Návod: Intervalový odhad odvoďte na základě popisu konstrukce intervalu spolehlivosti pro obecný kvantil, který byl na přednášce. [10 bodů]

3. Potvrzují Vaše data domněnku, že je střední BMI profesionálních fotbalistů nižší než 23? Kromě odpovědi uveďte také vhodný bodový a intervalový odhad, který bude doplňovat Vaše závěry. [10 bodů]

Dále vyřešte následující problém:

4. Pro náhodně vybraných 30 fotbalistů z české ligy vyšla průměrná výška 182.13 cm a směrodatná odchylka 6.11 cm. Lze předpokládat, že výška českých fotbalistů se řídí normálním rozdělením $\mathsf{N}(\mu, \sigma^2)$.
   - Spočtěte p-hodnotu testu $H_0 : \mu \leq 180$ proti $H_1 : \mu > 180$. Výpočet vysvětlete a výsledek interpretujte.
   - Spočtěte p-hodnotu testu $H_0 : \mu = 180$ proti $H_1 : \mu \neq 180$. Výpočet vysvětlete a výsledek interpretujte.

[10 bodů]

Úloha č. 2 (do 12.12. a 16.12. 2019) – 60 bodů

Použijte opět data soccer.csv a stejný vlastní náhodný výběr jako v předchozí úloze.

1. Souvisí střední BMI hráče s tím, zda je hráč starší než 30 let? [10 bodů]
2. Liší se rozdělení výšky obránců od útočníků? [10 bodů]
3. Odborníci tvrdí, že hráči fotbalu měří méně než je jejich hmotnost v kg plus 100. Ověřte tuto domněnku. [15 bodů]

Návod: S chybějícími daty si poradíte např. pomocí volby `na.rm` *v příslušné funkci nebo pomocí funkce* `na.omit`*. U všech otázek nezapomeňte uvést kromě výpočtů i vhodné grafy, které ilustrují danou situaci.*

Dále teoreticky vyřešte následující problém:

4. Nechť $X_1, \ldots, X_n$ je náhodný výběr z alternativního rozdělení s parametrem $p \in (0, 1)$. Nechť F_0 odpovídá distribuční funkci alternativního rozdělení s parametrem $p_0 \in (0, 1)$. Uvažujme testovou statistiku Kolmogorovova-Smirnovova testu

$$D_n = \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F_0(x) \right|.$$

   (a) Vyjádřete D_n (pro tento uvažovaný případ alternativního rozdělení) pomocí $\overline{X}_n$.
   (b) Na základě vyjádření z (a) popište asymptotické rozdělení $\sqrt{n}D_n$ za platnosti nulové hypotézy $H_0 : p = p_0$. Speciálně vyjádřete limitu $\mathsf{P}(\sqrt{n}D_n > u)$ pro $u > 0$ a $n \to \infty$.
   (c) Ukažte, že D_n nelze počítat jako

$$\max \left\{ \max_{i=1,\ldots,n} \left(\frac{i}{n} - F_0(X_{(i)}) \right), \max_{i=1,\ldots,n} \left(F_0(X_{(i)}) - \frac{i-1}{n} \right) \right\}.$$

   (d) Pro spojité rozdělení je známo, že $P(\sqrt{n}D_n > u)$ konverguje k $1 - G(u)$, kde

$$G(y) = \left(1 - 2 \sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2 y^2} \right) \mathrm{I}_{(0,\infty)}(y).$$

Vykreslete do jednoho grafu funkci $1-G(u)$ a limitu $\mathsf{P}(\sqrt{n}D_n > u)$ odvozenou v (b). V případě, že druhý zmíněný výraz závisí na nějakých neznámých parametrech, uveďte graf pro několik různých voleb.

Na základě tohoto grafického porovnání diskutujte, co se děje v situaci, kdy pro alternativní rozdělení spočteme statistiku D_n a pro výpočet p-hodnoty použijeme mylně funkci G namísto správného limitního rozdělení z (b). Speciálně objasněte, zda dostáváme test konzervativní nebo antikonzervativní nebo obecně nelze říci. Svoje tvrzení řádně zdůvodněte.

(e) Uvažujte $\widetilde{D}_n$, kde namísto F_0 použijeme $F_{\hat{p}}$, kde F_p je distribuční funkce alternativního rozdělení s parametrem p a $\hat{p}$ je momentový odhad p zkonstruovaný na základě $X_1, \ldots, X_n$. Jaké hodnoty $\widetilde{D}_n$ dostáváme a proč?

[25 bodů]