

ZÁPOČTOVÉ DOMÁCÍ ÚKOLY

Z MATEMATICKÉ STATISTIKY 1 NMSA331

1 OBECNÉ POKYNY

- Přečtěte si samostatný soubor s **pokyny k vypracování řešení**.
- **Řešení odevzdávejte buď v papírové podobě na svém cvičení nebo ve formě jednoho souboru v pdf formátu v Moodle v příslušném modulu.**
- Statistické testy provádějte na 5% hladině významnosti, intervalové odhady konstruujte se spolehlivostí 95 %.
- Ve svém řešení nezapomeňte uvést své studentské číslo (UKČO) pro generování vlastního náhodného výběru, viz níže.

ÚLOHA Č. 1 (ELEKTRONICKY V MOODLE DO 11. 12. 2022 23:59, PAPIROVĚ NA ZAČÁTKU CVIČENÍ VAŠÍ PARALELKY 12. 12. 2022, RESP. 13. 12. 2022) — 45 BODŮ

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

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

Načtěte si nejprve data pro proměnné `soccer`. Následně provedete svůj osobní náhodný výběr z tohoto souboru, a to následujícím způsobem: Znak `AAA` v příkazu `set.seed` níže nahraďte svým číslem ze studentského průkazu (UKČO) a spusťte následující příkazy:

```
set.seed(AAA);  
n <- sample(400:600, 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. Spočtete 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 dolní intervalový odhad pro 25%-ní kvantil 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]
4. Potvrzují Vaše data domněnku, že je střední BMI profesionálních fotbalistů různé od 23? Kromě odpovědi uveďte také vhodný bodový a intervalový odhad, který bude doplňovat Vaše závěry. [5 bodů]
5. Zaměřte se nyní na výšku fotbalistů. Potvrzují Vaše data domněnku, že jsou brankáři zpravidla (tj. s pravděpodobností větší než 1/2) vyšší než 185 cm?
Návod: Nezapomeňte si prohlédnout data a vysvětlíte, jak jste se vyřešili problém, že výška je zaokrouhlena na cm. [10 bodů]

ÚLOHA Č. 2 (ELEKTRONICKY V MOODLE DO 1. 1. 2023 23:59, PAPIROVĚ NA ZAČÁTKU CVIČENÍ VAŠÍ PARALELKY 2. 1. 2023, RESP. 3. 1. 2023) – 55 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 většina hráčů fotbalu měří méně než je jejich hmotnost v kg plus 100. Ověřte tuto domněnku. [10 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, \dots, 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 $P(\sqrt{n}D_n > u)$ pro $u > 0$ a $n \rightarrow \infty$.
- (c) Ukažte, že D_n nelze počítat jako

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

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

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

Vykreslete do jednoho grafu funkci $1-G(u)$ a limitu $P(\sqrt{n}D_n > u)$ 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 \tilde{D}_n , kde namísto F_0 použijeme $F_{\hat{p}}$, kde \hat{p} je momentový odhad p zkonstruovaný na základě X_1, \dots, X_n . Jaké hodnoty \tilde{D}_n dostáváme a proč?

[25 bodů]