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 Moodlu v příslušném modulu.
- 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.

2 Data

Data `movies.csv` obsahují údaje o filmech uvedených do amerických kin od roku 2010 do roku 2020 a jejich hodnocení ve filmové databázi IMDb. Pro každý film známe jeho název, žánr, rok uvedení, délku, průměrné hodnocení diváků, počet hodnocení, tržby (v USA a celosvětově). Dále pak máme k dispozici informaci o tom, zda je film z USA produkce a zda je film natočen v angličtině.

Kódování proměnných je následující:

`title`	název filmu,
`year`	rok uvedení,
`genre`	žánr,
`duration`	délka filmu v minutách,
`avg_vote`	průměrné hodnocení,
`votes`	počet hodnocení,
`usa_gross_income`	tržby v USA v mil. dolarů,
`worlwide_gross_income`	tržby celosvětově v mil. dolarů,
`usa`	film z produkce USA: 0 = ne, 1= USA v koprodukci s jinými státy 2= výhradní produkce USA,
`english`	film v angličtině: 1= ano, 0 = ne.

Načtěte si nejprve data pro proměnné `movies`. 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(200:400, size=1)
indexy <- sample(1:nrow(movies), size=n)
data <- movies[indexy,]
```

Proměnná `data` nyní obsahuje Vaše data o rozsahu n. Na základě těchto svých dat vyřešte následující úlohy.

Úloha č. 1 (do 6.12. 2021 resp. 9.12. 2021) — 45 bodů

1. Popište vhodně Váš datový soubor, a to jak číselně tak pomocí vhodných grafických nástrojů. Relevantní proměnné pro nás budou následující: průměrné hodnocení, celosvětová tržba, kategorie dle podílu produkce USA a jazyk filmu (angličtina nebo jiný). Ze zbylých proměnných vyberte jednu dle své preference a přidejte i její popis. [10 bodů]
2. Odhadněte bodově i intervalově 80% kvantil celosvětové tržby filmů z výhradní produkce USA. Interpretujte tyto odhady. [10 bodů]
3. Je možné tvrdit, že průměrné hodnocení filmů, na kterých se nepodílela produkce USA, je nižší než 6,5? [10 bodů]
4. Spočítejte pro každý film jeho tržbu mimo USA. Lze tvrdit, že je střední tržba filmu v USA a mimo USA stejná? [15 bodů]

V otázkách 3. a 4. kromě odpovědi uveďte i vhodný intervalový odhad.

Úloha č. 2 (do 3.1.2022 resp. 6.1.2022) – 55 bodů

Použijte opět data movies.csv a stejný vlastní náhodný výběr jako v předchozí úloze. Kromě odpovědi na otázku uveďte vždy i vhodné grafy a intervalové odhady a slovní interpretaci.

1. Je možné tvrdit, že střední celosvětová tržba závisí na tom, zda je film natočen v angličtině? [15 bodů]
2. Za dobře hodnocený film budeme považovat film s průměrným hodnocením alespoň 7. Je pravda, že podíl těchto filmů je menší než 25 %? [10 bodů]
3. Liší se statisticky významně podíl dobře hodnocených filmů mezi filmy z výhradní produkce USA a mezi ostatními filmy (kde není USA ve výhradní produkci)? [15 bodů]

Dále teoreticky vyřešte následující problém:

4. 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.)?

[15 bodů]