

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 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:

<code>title</code>	název filmu,
<code>year</code>	rok uvedení,
<code>genre</code>	žánr,
<code>duration</code>	délka filmu v minutách,
<code>avg_vote</code>	průměrné hodnocení,
<code>votes</code>	počet hodnocení,
<code>usa_gross_income</code>	tržby v USA v mil. dolarů,
<code>worldwide_gross_income</code>	tržby celosvětově v mil. dolarů,
<code>usa</code>	film z produkce USA: 0 = ne, 1= USA v koprodukcí s jinými státy 2= výhradní produkce USA,
<code>english</code>	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 — 45 BODŮ (V MOODLE DO 12.12. 23:59 NEBO NA ZAČÁTKU 11. CVIČENÍ, TJ. VE STŘEDU 13.12.)

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 u filmu, na kterém se nepodílela produkce USA, je možné očekávat hodnocení spíše 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 – 55 BODŮ (V MOODLE DO 2.1. 23:59 NEBO 3.1. 8:50-9:00 V K4)

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.

1. Dá se říct, že hodnocení filmů z výhradní produkce USA je (v nějakém smyslu) vyšší než hodnocení ostatních filmů? [10 bodů]
2. Je možné tvrdit, že střední celosvětová tržba závisí na tom, zda je film natočen v angličtině? [10 bodů]
3. Za dobře hodnocený film budeme považovat film s průměrným hodnocením alespoň 7. Je pravda, že podíl těchto filmů je menší než 20 %? [10 bodů]
4. Je podíl dobře hodnocených filmů vyšší pro filmy z výhradní produkce USA než pro ostatní? [10 bodů]

Dále teoreticky vyřešte následující problém:

5. Nechť X_1, \dots, X_n a Y_1, \dots, 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 $n/m_n \rightarrow \lambda$ pro $n \rightarrow \infty$, kde $\lambda \in (0, \infty)$ je neznámá hodnota.

- A. Najděte asymptotické rozdělení statistiky $F_{n,m}$ pro $n, m \rightarrow \infty$. Uveďte případné další předpoklady na rozdělení obou výběrů, které pro odvození rozdělení potřebujete.
- 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 .
- 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]