Štatistická práca

Skúmanie času stráveného na Netflixe

Pravděpodobnost a statistika 1, LS 2021/2022

Získanie dát a ich spracovanie

Kompletné dáta som získala z môjho Netflix účtu, ktorý v sebe zahŕňa 4 profily, k nahliadnutiu sú v súbore `data/ViewingActivity.csv`. Tento súbor obsahuje veľa pre túto prácu nepodstatných údajov, preto som ich prečistila pomocou skriptíčok v pythone, ktoré sú v zložke `data/data_modification_scripts`. Výsledné .csv súbory rozdelené podľa jednotlivých profilov (nazývaných len ako `Student1`,..., `Student4`) doplnené o dátumy, v ktoré nebol Netflix sledovaný (manuálne zarovnané na rovnaký začiatočný a konečný dátum) sú v zložke `data_final`. Prvý stĺpec obsahuje dátum, druhý stĺpec obsahuje počet hodín strávených sledovaním Netflixu v ten deň. Pre ilustráciu:

```
2022-06-01,1.1919
2022-05-31,0
2022-05-30,0
2022-05-29,2.4011
2022-05-28,0.9231
2022-05-27,2.1519
2022-05-26,1.4458
```

Strávia študenti rovnaký čas na Netflixe počas a mimo skúškového?

Cieľom tejto práce je overiť hypotézu, či študenti trávia rovnako času na Netflixe počas skúškového, ako mimo neho. Dáta budem skúmať dvomi spôsobmi, najprv všetky štyri profily dohromady a následne každý profil zvlášť.

Pre účely tejto práce budem za skúškové obdobie pokladať:
- letné skúškové: 05-20 až 06-30
- zimné skúškové: 01-08 až 02-14

Považujem za potrebné tu podotknúť, že študenti, ktorým jednotlivé profily patria, študujú na rôznych vysokých školách, skúškové obdobia sa preto môžu jemne líšiť, tieto rozdiely budem považovať za nepodstatné.

Označme:
- $S_1, ..., S_n \sim N(\mu_s, \sigma_s^2)$ časy strávené sledovaním Netflixu počas skúškových období
- $M_1, ..., M_m \sim N(\mu_m, \sigma_m^2)$ časy strávené sledovaním Netflixu mimo skúškových období

Nulová hypotéza $H_0 : \mu_s = \mu_m$ (počas a mimo skúškového strávia študenti priemerne rovnaký čas denne na Netflixe)

Alternatívna hypotéza $H_1 : \mu_s < \mu_m$ (počas skúškového strávia študenti priemerne menej času denne na Netflixe ako mimo skúškového)

Hladina významnosti: $\alpha = 0.05$

Poznámka: Správne by alternatívna hypotéza mala byť len nerovnosť, no keďže chcem zistiť aj to, kedy je sledovanosť vyššia, používam dve alternatívne hypotézy a využívam pritom schopnosť knižnice `scipy` analyzovať dáta aj týmto spôsobom.

Na overenie hypotézy budem používať **dvojvýberový Walshov t-test**, keďže nič nenaznačuje, že rozptyly náhodných veličín by mohli byť rovnaké.

Časť 1: Všetci študenti

Po rozdelení časov jednotlivých študentov na časy z obdobia skúškového a mimo neho a následným spojením týchto zoznamov dostávam dva zoznamy: `skuskove_all` a `mimo_all`. (potrebné skriptíčka v `stats/all_stat.py`)
Následne prevediem t-test:

```
>>> stats.ttest_ind(skuskove_all, mimo_all, equal_var=False, alternative="less")

Ttest_indResult(statistic=-3.1844329401039104, pvalue=0.0007389556448342999)
```

Získaná p-value je menšia ako haldina vierohodnosti α, preto môžem H_0 zamietnuť. Platí teda alternatívna hypotéza H_1, teda že študenti strávia na Netflixe menej času počas skúškového, ako mimo neho.

Časť 2: Každý profil zvlášť

Pozrime sa teraz na jednotlivých študentov zvlášť:
Student1:

```
>>> stats.ttest_ind(skuskove, mimo, equal_var=False, alternative="less")

Ttest_indResult(statistic=-3.8022382781536277, pvalue=7.995788759218318e-05)
```

Student2:

```
>>> stats.ttest_ind(skuskove, mimo, equal_var=False, alternative="less")

Ttest_indResult(statistic=-5.185564583985727, pvalue=1.6338836876844316e-07)
```

Student3:

```
>>> stats.ttest_ind(skuskove, mimo, equal_var=False, alternative="less")

Ttest_indResult(statistic=-0.5842829109856048, pvalue=0.27967840340112504)
```

Student4:

```
>>> stats.ttest_ind(skuskove, mimo, equal_var=False, alternative="less")

Ttest_indResult(statistic=1.8422393338979608, pvalue=0.9668485613877612)
```

Z týchto výsledkov môžme usúdiť:

- Pri prvých dvoch študentoch je H_0 zamietnuteľná, keďže p-value $< \alpha$
- Pri treťom študentovi H_0 zamietnuť nemôžeme
- Pri štvrtom študentovi H_0 nemôžeme zamietnuť, no keďže hodnota štatistiky prekračuje kritickú hodnotu t-rozdelenia s $n + m - 2$ stupňami voľnosti (čo je pre počet stupňov volnosti rastúci nad všetky medze $= 1.6448$), budem sa týmto študentom zaoberať ešte trochu.

Výška p-hodnoty pri štvrtom študentovi naznačuje, že stredné hodnoty sledovania Netflixu mimo a počas skúškového sú s vysokou pravdepodobnosťou rovnaké. Výsledná hodnota štatistiky ale naznačuje, že H_0 zamietnuť môžme.
Skúsim preto sformulovať inú alternatívnu hypotézu $H_2 : \mu_s > \mu_m$.
Po overení H_0 s H_2 ako alternatívou dostávame:

```
>>> stats.ttest_ind(skuskove, mimo, equal_var=False, alternative="greater")

Ttest_indResult(statistic=1.8422393338979608, pvalue=0.03315143861223883)
```

Z toho plynie, že H_0 je zamietnuteľná v prospech H_2.