Vysoké učení technické v Brně
Fakulta informačních technologií

MSP – projekt

Cvičenie: Streda 8:00

11. decembra 2021 Natália Holková (xholko02)

Príklad č. 1

Zadanie

Nový poskytovateľ innternetového pripojenia na Vašej adrese Vám ponúka svoje pripojenie "na skúšku" na jeden mesiac. Rozhodujúce kritérium pre výber poskytovateľa pripojenia je rýchlosť odozvy (ping) počas hrania Vašej obľúbenej online hry. Zadanie obsahuje priemernú odozvu [ms] počas hodinovej hernej seansy pri použitú aktuálneho pripojenia (X) a pri použití pripojenia od nového poskytovateľa (Y). Pomocou vhodnej štatistickej analýzy rozhodnite, ktorý z poskytovateľov internetového pripojenia je pre Vás vhodnejší. Svoje rozhodnutie zdôvodnite.

Dáta

X				
22.43	22.39	24.18	22.22	25.52
20.27	24.88	22.26	24.97	19.27
27.77	20.52	23.25	21.37	25.30
19.03	23.41	18.64	23.78	18.74
24.94	19.34	23.58	20.42	24.98
23.06	25.54	21.12	25.82	20.02
26.95	21.73	24.47	21.36	25.90
22.38	23.55	20.01	25.98	19.69
24.96	21.91	24.89	17.81	24.72
21.17	24.72	19.32	25.97	19.10

Y				
23.13	24.81	28.78	25.42	25.51
24.10	23.00	22.42	20.86	23.73
20.88	24.75	25.43	22.81	23.40
24.14	25.63	22.34	24.41	23.33
20.06	20.87	20.88	24.17	25.91
22.28	22.29	24.78	25.53	23.47
24.39	24.05	22.30	23.62	23.91
26.13	24.85	23.71	22.23	24.75
21.50	20.73	25.28	22.25	24.90
24.77	25.61	23.38	25.02	24.55

Riešenie

Prvým krokom pri riešení je otestovanie, či $\mathbf{X}$ a $\mathbf{Y}$ majú normálne rozdelenie. Pre toto použijeme *Lilliefors* test, ktorý je založený na *Kolmogorov-Smirnov* teste. Nulovou hypotézou v tomto teste je, že dáta majú normálne rozloženie. Pri výpočtoch sme použili programovací jazyk `python` a jeho knižnicu `statmodels`, ktorá obsahuje implementáciu tejto štatistickej metódy vo funkcii `lilliefors`.

Pracovali sme s hodnotou $\alpha = 0.05$. Pri testovaní, či $\mathbf{X}$ má normálne rozloženie nám výšlo $p-value = 0.055$, čo je väčšie ako α a preto nulovú hypotézu, že $\mathbf{X}$ má normálne rozloženie *nezamietame*. Pri testovaní, či $\mathbf{Y}$ má normálne rozloženie nám výšlo $p-value = 0.572$, čo je väčšie ako α a preto nulovú hypotézu, že $\mathbf{Y}$ má normálne rozloženie taktiež *nezamietame*.

Nakoľko vyšlo, že $\mathbf{X}$ aj $\mathbf{Y}$ majú normálne rozloženie, využijeme ďalej *Studentov* test. V programe využívame knižnicu `scipy`, ktorá obsahuje tento test vo funkcii `stat.ttest_ind`, nakoľko sa jedná o dva nezávislé súbory dát. Nulovou hypotézou je tu predpoklad, že odovzva v $\mathbf{X}$ je väčšia alebo rovná odozve v $\mathbf{Y}$. Alternatívna hypotéza je, že odovza v $\mathbf{Y}$ je väčšia ako v $\mathbf{X}$. Vyšlo nám $p-value = 0.021$ a pretože $p-value/2 <= \alpha$, nulovú hypotézu *zamietame*. Platí teda, že odozva v $\mathbf{Y}$ je väčšia ako odozva v $\mathbf{X}$.

Záverom celého riešenia je, že je pre nás výhodnejšie zostať u pôvodného poskytovateľa internetového pripojenia.

Príklad č. 2

Zadanie

Bol vykonaný prieskum, či čas [min] potrebný k vyriešeniu určitej úlohy závisí na dennej dobe alebo na hlučnosti okolia. Denná doba (faktor 1) nadobúda troch hodnôt: ráno, popoludnie a večer. Hlučnosť okolia (faktor 2) nadobúda štyri hodnoty: tiché prostredie, reprodukovaná hudba, pouličný hluk, krik (dieťaťa, študentov, ktorí vo vedľajšej izbe oslavujú úspešné absolvovanie skúšky z MSP).

Počet študentov, ktorí riešili úlohu za určitých podmienok, bol rôzny. Čas v minútach potrebný k vyriešeniu úlohy je uvedený v tabuľke. Do tabuľky si každý študent ku každej hodnote faktoru 1 pripíše jeho zvolené hodnoty. (Zvolí si číslo a zvolí si, do ktorej hodnoty faktoru 2 ho pripíše. Teda v tabuľke pribudnú celkovo tri hodnoty.) Zistite, či doba potrebná k vyriešeniu úlohy závisí na dennej dobe alebo na hlučnosti okolia alebo na kombinácii oboch faktorov. Predpokladajte rovnosť rozptylov v jednotlivých kategóriách.

Dáta

	faktor 2			
faktor 1	ticho	hudba	hluk	krik
ráno	6	7	8	13
	8	8	7	21
	11	12	20	
	9	10		
popoludnie	8	5	10	14
	13	11	17	
	7	7	11	
		10	13	
večer	7	6	12	13
	8	8	17	17
	6	16	**18**	15
		15		22
				18

Riešenie

Pre riešenie tohto zadania je potrebné využiť nevyváženú dvojfaktorovú ANOVU. Knižnica `pingouin` poskytuje priamo funkciu `anova` na jej riešenie. Dáta je najskôr potrebné transformovať z tabuľky do `pandas.DataFrame`, ktorý bude obsahovať záznam pre každú bunku tabuľky - teda hodnotu 1. faktora, hodnotu 2. faktora a čas potrebný na vyriešenie.

Funkcia `anova` nám vráti ako výsledok tabuľku:

	Source	SS	DF	MS	F	p-unc	n2
0	Faktor 1	16.191	2.0	8.095	0.597	0.558	0.021
1	Faktor 2	342.907	3.0	114.302	8.422	0.000	0.437
2	Faktor 1 * Faktor 2	44.801	6.0	7.467	0550	0.766	0.057
3	Residual	380.000	28.0	13.571	NaN	NaN	NaN

Z tejto tabuľky vieme priamo vyčítať *phodnoty* (p-unc), ktoré použijeme pre rozhodovanie o závislosti doby potrebnej na vyriešenie úlohy na jednotlivých faktoroch a ich kombinácii. Pri $\alpha = 0.05$ dospejeme k nasledujúcim záverom:

- doba potrebná na vyriešenie úlohy **závisí** na dennej dobe pretože $0.558 > \alpha$
- doba potrebná na vyriešenie úlohy **nezávisí** na hlučnosti okolia pretože $0.000 <= \alpha$
- doba potrebná na vyriešenie úlohy **závisí** na kombinácii oboch faktorov pretože $0.766 > \alpha$

Príklad č. 3

Zadanie

Táto úloha je na testovanie nezávislosti dvoch kvalitatívnych premenných (faktorov, pojmov). Tieto premenné si každý študent zvolí sám. Každá kvalitatívna premenná bude popísaná minimálne 4 typmi hodnôt. Potom každý študent:

1. navrhne nulovú hypotézu (tvrdenie o nezávislosti zvolených premenných)
2. zostaví formulár pre dotazník
3. vykoná anketu (vo svojom okolí, pomocou internetu,...). Pomocou dotazníku osloví vybraných respondentov. Počet respondentov by mal byť dostatočný pre splnenie podmienky pre teoretickú četnosť. Uveďte ako, kde a kedy bola vykonaná.
4. odpovede prepíše do tabuľky pre kategoriálnu analýzu
5. pomocou vhodného štatistického testu vyhodnotí závislosť (nezávislosť)
6. zformuluje záver

Riešenie

Cieľom dotazníka bolo zistiť, či existuje závislosť medzi dosiahnutou známkou zo štátnej záverečnej skúšky a počtom jazykov, ktoré človek ovláda. Dotazník bol vytvorený pomocou Google Forms a obsahoval nasledujúce otázky:

1. Akú známku ste dostali na štátniciach?
   odpoveď: A/B/C/D/E
2. Koľko jazykov ovládate (vrátane materinského)?
   odpoveď: 1/2/3/4/5 a viac

Dotazní vyplnilo celkovo 194 ľudí. Po prepise do tabuľky sme zaznamenali nasledovné počty odpovedí v jednotlivých kategóriách:

	1	**2**	**3**	**4**	**5 a viac**	$\sum$
A	7	5	7	10	9	38
B	8	12	15	5	5	45
C	9	14	10	8	8	49
D	6	7	6	7	5	31
E	7	7	7	5	5	31
$\sum$	37	45	45	35	32	194

Graficky môžme vidieť rozloženie odpovedí na obrázku 1.

Ďalej je potrebné overiť podmienku, že teoretická četnosť každej bunky tabuľky $n_{i,j}$, ktorá je definovaná ako:

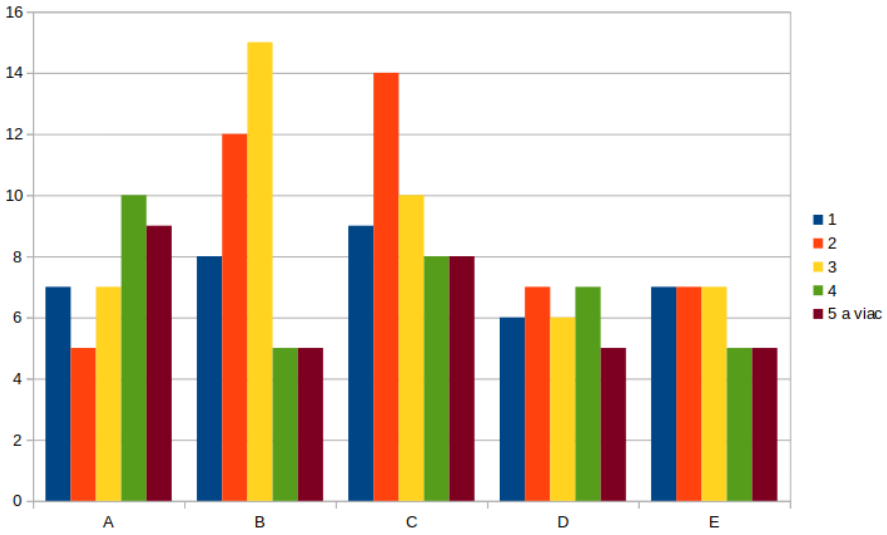
$$n_{i,j} = \frac{n_{i,\bullet} * n_{\bullet,j}}{n}, \forall i,j \in \{1,2,3,4,5\}$$

nadobúda hodnotu > 5. Vypočítané teoretické četnosti je možné vidieť v tabuľke nižšie. Je vidno, že spĺňajú podmienku o minimálnej hodnote teoretickej četnosti. Je preto možné pokračovať ďalej vo výpočte.

	1	**2**	**3**	**4**	**5 a viac**	$\sum$
A	7.25	8.81	8.81	6.86	6.27	38
B	8.58	10.44	10.44	8.12	7.42	45
C	9.35	11.37	11.37	8.84	8.08	49
D	5.91	7.19	7.19	5.59	5.11	31
E	5.91	7.19	7.19	5.59	5.19	31
$\sum$	37	45	45	35	32	194

Nulová hypotéza teda znie, že *známka zo štátnej záverečnej skúšky nezávisí počtu jazykov, ktoré človek ovláda.*

Vyšlo nám $p-value = 0.999$, čo je väčšie ako $\alpha = 0.05$ a teda nulovú hypotézu *nezamietame - nie je závislosť medzi známkou zo štátnic a počtom jazykov, ktoré človek ovláda.*



Obr. 1: Počet jazykov vzhladom na známku zo štátnic

Spustenie

Výpočty boli realizované v jazyku `python` s využitím knižníc `numpy`, `scipy`, `pandas`, `pingouin` a `statmodels` a na operačnom systéme Ubuntu. V prípade, že nie sú tieto knižnice prítomné, je možné vytvoriť virtuálne prostredie a nainštalovať ich príkazmi:

```
python3 -m venv ./env-msp
source ./env-msp/bin/activate
pip3 install -r requirements.txt
```

Po splnení všetkých závislostí je možné spustiť program príkazom:

```
python3 msp_project.py
```