VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Semestrálny projekt

Štatistika a pravdepodobnosť (MSP)

Bc. Ján Jakub Kubík (xkubik32)

11. decembra 2021

1 Príklad 1

Zadanie

Nový poskytovatel internetového připojení na Vaší adrese Vám nabízí svoje připojení „na zkoušku“ na jeden měsíc. Rozhodujícím kritériem pro výběr poskytovatele připojení je rychlost odezvy (ping) během hraní Vaší oblíbené online hry. Zadání obsahuje průměrnou odezvu [ms] během hodinové herní seance při použití stávajícího připojení (X) a při použití připojení od nového poskytovatele (Y). Pomocí vhodné statistické analýzy rozhodněte, který z poskytovatelů internetového připojení je pro Vás vhodnější. Své rozhodnutí zdůvodněte

Riešenie

Použitá štatistická metóda - Mann-Whitney test

Hladina významonosti: $\alpha = 0.05$.

H_0: Ping od poskytovateľa X a Y sú zhodné.

H_{A_1}: Ping od poskytovateľa X a Y nie sú zhodné.

H_{A_2}: Ping od poskytovateľa X trvá dlhšie ako ping od poskytovatela Y.

H_{A_3}: Ping od poskytovateľla X trvá kratšie ako ping od poskytovatela Y.

zadání číslo 17									
X [ms]	Y[ms]	X [ms]	Y[ms]	X [ms]	Y[ms]	X [ms]	Y[ms]	X [ms]	Y[ms]
25,75	24,42	25,32	22,52	24,53	21,26	25,66	24,19	25,97	24,83
23,85	23,08	22,19	24,87	21,32	22,77	20,2	21,98	21,01	24,71
25,31	23,7	26,69	19,9	24,46	24,3	24,9	23,33	25,77	26,02
22,33	23,3	19,13	26,77	20,3	23,78	21,18	24,62	21,54	25,85
25,06	23,57	24,5	24,48	24,91	23,89	25,36	23,7	25,67	25,37
20,94	26,79	20,62	25,42	20,11	27,14	20,51	23,34	19,18	20,47
24,05	25,51	26,24	24,96	24,9	23,32	24,58	18,96	24,21	22,83
20,64	25,75	22,12	23,28	20,75	25,75	19,99	25,28	20,11	25,47
25,87	25,37	27,65	21,04	24,2	24,67	28,22	23,34	25,69	24,6
17,54	23,68	21,43	25,21	22,02	25,74	19,86	24,89	21,75	22,77

Obr. 1.1: Data pre príklad.

Postup

Príklad 1 som celý riešil pomocou programovaieho jazyku Python. Najskôr som musel skontrolovať normalitu dát. Na to som použil Shapiro-Wilk test (funkicia shapiro z Python knižnice scipy.stats). Nulová hypotéza pre tento test je, že dáta sú normálne rozložené. Ak výsledná hodnota pvalue je menšia ako zvolené alfa, tak sa zamieta, že dáta sú normálne rozdelené.

```
pvalue of X_ms: 0.03673752769827843
pvalue of Y_ms: 0.0219634510576725
```

V tomto príklade sa normalita dát zamieta pre X aj pre Y.

Čiže som musel použit neparametrický test. Rozhodol som sa použiť Mann-Whitney test. Mann-Whitney test je tiež v Python knižnici scipy.stats ako funkcia mannwhitneyu. Keďže dát X aj Y je viac ako 20, tak som musel použiť asymptotickú verziu Mann-Whitney testu. Vzorce na výpočet testovacieho kritéria:

$t = \frac{U_x - \frac{n_x * n_y}{2}}{\sqrt{\frac{n_x * n_y * (n_x + n_y + 1))}{12}}}$ pre $n_x \geq n_y$, $t = \frac{U_y - \frac{n_x * n_y}{2}}{\sqrt{\frac{n_x * n_y * (n_x + n_y + 1))}{12}}}$ pre $n_x < n_y$

n_x je pocet prvkov skupiny X, n_y je pocet prvkov skupiny Y

$U_x = n_x * n_y + \frac{n_x * (n_x + 1))}{2} - T_x$, $U_y = n_x * n_y + \frac{n_y * (n_y + 1))}{2} - T_y$

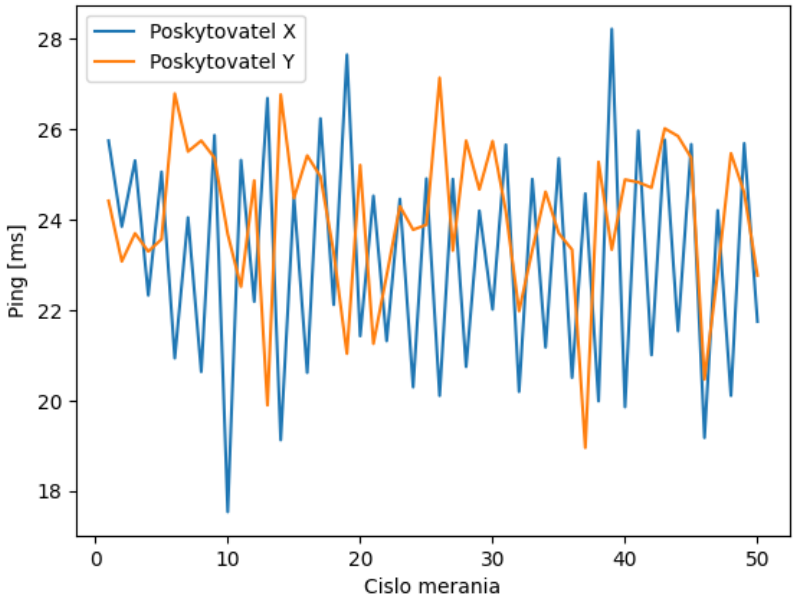
$T_x = \sum_{Z_i \in (X_1 \ldots X_m)} R_i$, $T_y = \sum_{Z_i \in (Y_1 \ldots Y_n)} R_i$

R_i pre Tx je celkové poradie jednotlivých prvkov v rámci skupiny X, R_i pre Ty je celkové poradie jednotlivých prvkov v rámci skupiny Y.

Funkcia mannwhitneyu zo scipy knižnice to počíta trochu inak, ale principiálne to je to isté. Podrobný opis toho ako to táto funkcia počita je na wikipedii sekcia Normal approximation and tie correction. Mannwhitneyu funkcia vracia priamo pvalue, ktorá ak je menšia ako alfa, tak sa H_0 zamieta v opačnom prípade nie.

```
H0, HA1: pvalue=1.0
H0, HA2: pvalue=0.9370502694009104
H0, HA3: pvalue=0.06294973059908954
```

Pre H_{A_1}, H_{A_2} aj H_{A_3} je pvalue väčšie ako alfa (0.05), čiže ani v jednom prípade sa nezamieta H_0 (Ping od poskytovatela X a Y sú zhodné). Ale pre prípad H_{A_3} sa pvalue najviac približuje k zamietnutiu H_0 (to by znamenalo, že Ping od poskytovatela X trvá kratšie ako ping od poskytovatela Y).



Obr. 1.2: Graf pingov

Taktiež z grafu pingov pre poskytovatelov je vidieť, že z nameraných hodnôt má poskytovateľ X kratšiu dobu pingov. **Z týchto dôvodov by som si vybral poskytovateľa X.**

2 Príklad 2

Zadanie

Byl proveden průzkum, zda čas [min] potřebný k vyřešení určité úlohy závisí na denní době nebo na hlučnosti okolí. Denní doba (faktor 1) nabývá tří hodnot: ráno, v poledne a večer. Hlučnost okolí (faktor 2) nabývá čtyři hodnoty: tiché prostředí, reprodukovaná hudba, pouliční hluk, křik (dítěte, studentů, kteří ve vedlejším pokoji slaví úspěšné absolvování zkoušky z MSP). Počet studentů, kteří řešili úkol za určitých podmínek, byl různý. Čas v minutách potřebný k vyřešení úlohy je uveden v tabulce. Do tabulky si každý student ke každé hodnotě faktoru 1 připíše jím zvolené hodnoty. (Zvolí si číslo a zvolí si, do které hodnoty faktoru 2 ho připíše. Tedy v tabulce přibudou celkem tři hodnoty.) Zjistěte, zda doba potřebná k vyřešení úlohy závisí na denní době nebo na hlučnosti okolí nebo na kombinaci obou faktorů. Předpokládejte rovnost rozptylů v jednotlivých kategoriích.

Riešenie

Použitá štatistická metóda - 2-faktorová ANOVA s interakciou

Hladina významonosti: $\alpha = 0.05$.

H_{0_1}: doba potrebná k vyriešeniu úlohy nezávysí na dennej dobe (faktor 1). H_{A_1}: opak H_{0_1} (závysí).

H_{0_2}: doba potrebná k vyriešeniu úlohy nezávysí na hlučnosti (faktor 2). H_{A_2}: opak H_{0_2} (závysí).

H_{0_3}: doba potrebná k vyriešeniu úlohy nezávysí na kombinácii dennej doby (faktor 1) a hlučnosti (faktor 2). H_{A_3}: opak H_{0_3} (závysí).

	faktor 2			
faktor 1	ticho	hudba	hluk	krik
rano	6	7	8	13
	8	8	7	21
	11	12	20	
	7	10		
obed	8	5	10	14
	13	11	17	45
	7	7	11	
			13	
vecer	7	6	12	13
	8	8	17	17
	6	16	30	15
		15		22
				18

Obr. 2.1: Data pre príklad.

Postup

Do tabuľky som doplnil 3 hodnoty. Sú zvýraznené červenou farbou. Celý príklad som riešil pomocou Pythonu. V zadaní je uvedené, že mám predpokladať rovnosť rozptylov v jednotlivých kategóriach. To znamená, že som nemusel robiť žiadne overovanie rovnosti rozptylov (napr. Bartlettov test), ale mohol som sa rovno pustiť do počítania ANOVA.

Na vytvorenie modelu pre riešenie 2-faktorovej ANOVA som použil funkciu ols z Python knižnice statsmodels.formula.api. Formula pre zostavenie modelu v Python kóde:

```
model = ols('riesenie_min ~ C(faktor_1) + C(faktor_2) + C(faktor_1):C(faktor_2)', data=df).fit()
```

Na riešenie modelu som použil funkciu sm.stats.anova_lm typu 3 z Python knižnice statsmodels.api:

```
result_table = sm.stats.anova_lm(model, typ=3)
```

Typ 3 je pre vyriešenie 2-faktorovej nevyvážená ANOVA s interakciou.

	sucet stvorcov	stupen volnosti	testovacia statistika	p-hodnota
C(faktor_1)	116.416667	2.0	1.614629	0.216945
C(faktor_2)	668.333333	3.0	6.179587	0.002335
C(faktor_1):C(faktor_2)	360.089171	6.0	1.664740	0.166764

Obr. 2.2: Tabuľka v výsledným riešením

Ak výsledná hodnota pvalue je väčšia ako zvolené alfa (0.05), tak sa nezamieta H_{0_x}. V opačnom prípade (pvalue je menšia ako zvolené alfa) sa H_{0_x} zamieta a platí H_{A_x}

H_{0_1}: nezamietam (nezamietam, že doba potrebná k vyriešeniu úlohy nezávysí na dennej dobe) Nezameitam to pretože p-hodnota pre faktor_1 je väčšia ako alfa.

H_{0_2}: zamietam (zamietam,že doba potrebná k vyriešeniu úlohy nezávysí na hlučnosti, čiže platí H_{A_2} - doba potrebná k vyriešeniu úlohy závysí na hlučnosti) Zamietam to pretože p-hodnota pre faktor_2 je menšia ako alfa.

H_{0_3}: nezamietam (nezamietam, že doba potrebná k vyriešeniu úlohy nezávysí na kombinácii dennej doby a hlučnosti). Nezamietam to pretože p-hodnota pre faktor_1:faktor_2 je väčšia ako alfa.

3 Príklad 3

Zadanie

Tento úkol je na testování nezávislosti dvou kvalitativních proměnných (faktorů, pojmů). Tyto proměnné si každý student zvolí sám. Každá kvalitativní proměnná bude popsána minimálně 4 typy hodnot. Pak každý student:

1. navrhne nulovou hypotézu (tvrzení o nezávislosti zvolených proměnných)
2. sestaví formulář pro dotazník
3. **provede anketu** (ve svém okolí, pomocí internetu,...). Pomocí dotazníku osloví vybrané respondenty. Počet respondentů by měl být dostatečný pro splnění podmínky pro teoretickou četnost. Uveďte, jak, kde a kdy byla provedena.
4. odpovědi přepíše do tabulky pro kategoriální analýzu
5. pomocí vhodného statistického testu vyhodnotí závislost (nezávislost)
6. zformuluje závěr.

Riešenie

Použitá štatistická metóda - Test dobrej zhody

Hladina významonosti: $\alpha = 0.05$.

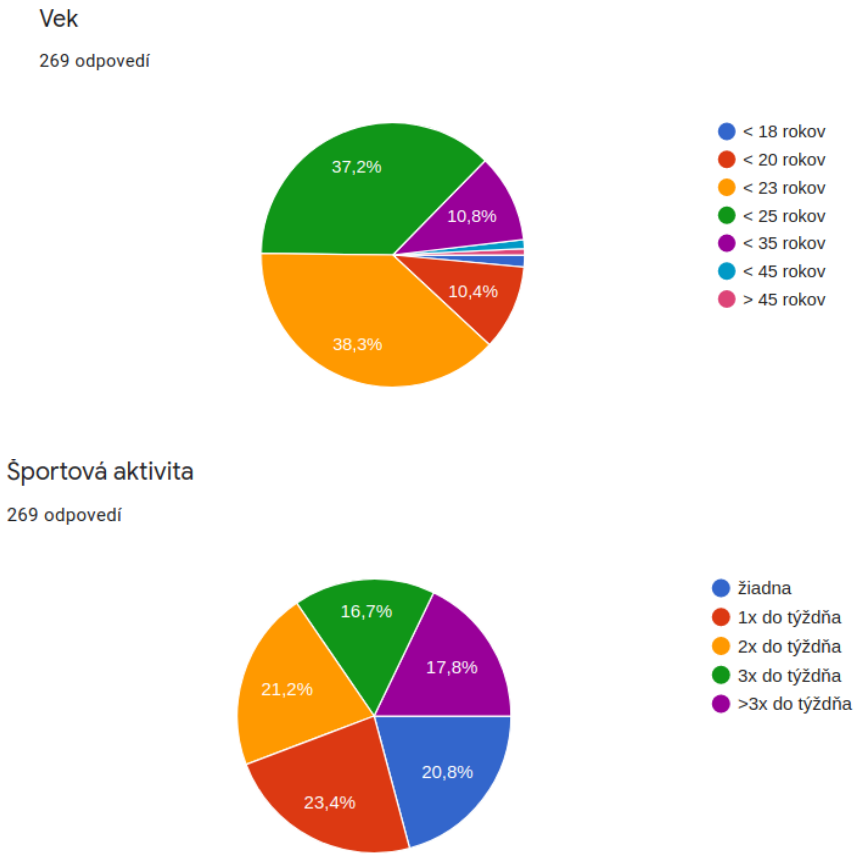
H_0: nezávyslosť medzi frekvenciou športovej aktivity a vekom.

H_A: existuje závyslosť medzi frekvenciou športovej aktivity a vekom.

Postup

Najskôr som si cez Google Forms zostavil dotazník pre skúmanie požadovanej hypotézy. Formulár som poslal všetkým členom rodiny, rodinným známym, kamarátom, spolužiakov a do Facebookovej skupiny Koleje pod Palackého vrchem.

Výsledky prieskumu



Výsledky prieskumu so si agregoval v Pythone pre jednotlivé skupiny cez požadované vekové kategórie pomocou Python kódu. Z agregácii som vytvoril tabuľku v Google Sheets a nad tabuľkou som otestoval motodou Test dobrej zhody hypotezou H_0 (tiež v Google Sheets).

	0x	1x	2x	3x	>3x	SUMA
<18	1	2	0	1	0	4
<20	3	7	4	8	6	28
<23	19	25	24	20	15	103
<25	23	24	23	12	18	100
<35	8	4	6	4	7	29
<45	2	1	0	0	0	3
>45	0	0	0	0	2	2
SUMA	56	63	57	45	48	269

Obr. 3.1: TAB1.

TAB1 je tabuľka nameraných hodnôt. Vzorec: $n_{i,j}$

	0x	1x	2x	3x	>3x	SUMA
<18	0,8327137546	0,936802974	0,8475836431	0,6691449814	0,7137546468	4
<20	5,828996283	6,557620818	5,933085502	4,68401487	4,996282528	28
<23	21,44237918	24,12267658	21,82527881	17,23048327	18,37918216	103
<25	20,81784387	23,42007435	21,18959108	16,72862454	17,84386617	100
<35	6,037174721	6,791821561	6,144981413	4,851301115	5,17472119	29
<45	0,624535316	0,7026022305	0,6356877323	0,5018587361	0,5353159851	3
>45	0,4163568773	0,468401487	0,4237918216	0,3345724907	0,3568773234	2
SUMA	56	63	57	45	48	269

Obr. 3.2: TAB2.

TAB2 som zostavil z TAB1 pomocou vzorca: $\frac{n_{i,.} * n_{.,j}}{n}$

podmienka p: $\frac{n_{i,.} * n_{.,j}}{n} > 5; \forall i, j$

Pre splnenie podmienky p som musel zlúčiť dokopy riadky < 18, < 20 a riadky < 35, < 45, > 45.

	0x	1x	2x	3x	>3x	SUMA
<20	4	9	4	9	6	32
<23	19	25	24	20	15	103
<25	23	24	23	12	18	100
>25	10	5	6	4	9	34
SUMA	56	63	57	45	48	269

Obr. 3.3: TAB1 po uprave.

TAB1 po úprave je výsledná tabuľka po zlúčení.

	0x	1x	2x	3x	>3x	SUMA
<20	6,661710037	7,494423792	6,780669145	5,353159851	5,710037175	32
<23	21,44237918	24,12267658	21,82527881	17,23048327	18,37918216	103
<25	20,81784387	23,42007435	21,18959108	16,72862454	17,84386617	100
>35	7,078066914	7,962825279	7,204460967	5,687732342	6,066914498	34
SUMA	56	63	57	45	48	269

Obr. 3.4: TAB2.

Ďalej som vyrátal **TAB2** z TAB1 po úprave. Počítal som to rovnako ako prvé TAB2, len som vychádzal z TAB1 po úprave. TAB2 už splňuje podmienku p.

	0x	1x	2x	3x	>3x	SUMA
<20	-2,661710037	1,505576208	-2,780669145	3,646840149	0,2899628253	0
<23	-2,442379182	0,8773234201	2,17472119	2,769516729	-3,379182156	0
<25	2,182156134	0,5799256506	1,810408922	-4,728624535	0,156133829	0
>35	2,921933086	-2,962825279	-1,204460967	-1,687732342	2,933085502	0
SUMA	0	0	0	0	0	0

Obr. 3.5: TAB3.

TAB3 som vyrátal ako $TAB1 - TAB2$

Vzorec: $n_{i,j} - \frac{n_{i,.} * n_{.,j}}{n}$

	0x	1x	2x	3x	>3x	SUMA
<20	1,063495751	0,3024595061	1,140318268	2,484409851	0,01472467472	5,005408051
<23	0,2781974901	0,03190758624	0,2166942422	0,445154253	0,6212938066	1,593247378
<25	0,2287367233	0,01436006373	0,1546787974	1,336624535	0,001366171004	1,735766291
>35	1,206218175	1,102414448	0,2013649913	0,5008042374	1,418017439	4,428819291
SUMA	2,77664814	1,451141604	1,713056299	4,766992877	2,055402092	**12,76324101**

Obr. 3.6: TAB4.

TAB4 som som vyrátal ako $TAB3^2/TAB2$

Vzorec: $\frac{(n_{i,j} - \frac{n_{i,.} * n_{.,j}}{n})^2}{\frac{n_{i,.} * n_{.,j}}{n}}$

Testovacie kritérium je zvýraznena suma z TAB4. Vzorec: $t = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{i,j} - \frac{n_{i,.} * n_{.,j}}{n})^2}{\frac{n_{i,.} * n_{.,j}}{n}}$

$t = 12.76324101$

Počet riadkov, stĺpcov v tabuľke: $r = 4, s = 5$

Doplnok kritického oboru:

$k = (r-1) * (s-1), k = 12$

$\overline{W_\alpha} =< 0, \chi^2_{1-\alpha}(k)) >$, $\overline{W_{0.05}} =< 0, \chi^2_{0.95}(12))$, $\overline{W_\alpha} =< 0, 21.026 >$

$t \in \overline{W_\alpha} \Longrightarrow H_0$ **nezamietam (nezamietam nezávyslosť medzi frekvenciou športovej aktivity a vekom).**