VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

Statistika a pravděpodobnost

Semestrální projekt

10. prosince 2021

Tomáš Kocourek

1. úloha

Vstupní data pro zadání číslo **12** jsou shrnuta v následujících tabulkách. Netříděný statistický soubor X zahrnuje 50 naměřených hodnot průměrné odezvy v milisekundách při hodinovém hraní s připojením od původního poskytovatele, netříděný statistický soubor Y zahrnuje 50 naměřených hodnot průměrné odezvy v milisekundách při hodinovém hraní s připojením od nového poskytovatele.

X	26.33	20.99	25.35	20.97	22.79	21.88	24.03	20.84	23.93	22.15
Y	26.28	22.92	25.55	24.6	25.02	26.25	22.91	21.24	24.76	22.28

X	23.9	19.74	25.52	20.75	25.25	21.88	23.41	21.94	29.02	20.0
Y	21.03	23.63	26.61	21.81	23.73	24.02	21.88	24.92	21.5	25.4

X	25.15	20.6	22.77	21.33	24.78	21.4	24.29	22.09	24.03	21.94
Y	22.9	22.48	22.58	24.9	23.83	25.68	23.29	24.21	22.25	29.22

X	23.97	19.3	26.55	20.42	25.44	20.96	24.99	20.42	24.86	19.78
Y	26.36	24.28	23.91	25.16	24.8	22.92	20.65	25.17	22.67	25.6

X	25.06	20.38	24.51	19.53	23.77	19.29	22.79	20.2	25.19	22.12
Y	26.73	22.35	25.71	24.88	25.45	25.01	27.04	23.87	25.63	24.9

Test na normalitu

Nejprve otestujme oba statistické soubory na normalitu pomocí χ^2 testu dobré shody. Začneme se souborem X. Sestavíme nulovou hypotézu $H_0 : X \sim N(\mu, \sigma^2)$, kde μ a σ^2 jsou neznámé parametry. Volíme $\alpha = 0.05$.

Provedeme bodový odhad parametru $\mu : \mu \overset{odhad}{=} \overline{x}$.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 22.7716$$

Nyní provedeme bodový odhad parametru $\sigma^2 : \sigma^2 \overset{odhad}{=} s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = 5.1387$$

Z naměřených dat je zjevné, že variační obor činí $\langle 19.29; 29.02\rangle$ a rozpětí je $29.02 - 19.29 = 9.73$. Pro volbu počtu tříd využijeme Sturgessovo pravidlo. Počet tříd označme m.

$$m = \left\lceil 1 + \frac{\log(n)}{\log(2)} \right\rceil = \left\lceil 1 + \frac{\log(50)}{\log(2)} \right\rceil = 7$$

Nyní sestavíme tříděný statistický soubor s četnostmi jednotlivých tříd. Symbol f_k zastupuje skutečnou četnost k-té třídy. Označme symbolem x_k^- infimum k-té třídy a symbolem x_k^+ supremum k-té třídy. Pak teoretickou četnost $\hat{f}_k$ třídy k spočteme pomocí vzorce

$$\hat{f}_k = n \cdot \left(\lim_{x \to (x_k^+)^-} \Phi\left(\frac{x - \overline{x}}{s}\right) - \lim_{x \to (x_k^-)^+} \Phi\left(\frac{x - \overline{x}}{s}\right) \right),$$

kde Φ je distribuční funkce náhodné veličiny s normovaným normálním rozdělením $U \sim N(0,1)$. Jelikož distribuční funkce Φ je definovaná na celé množině $\mathbb{R}$, zvolíme $x_1^- = -\infty$ a $x_7^+ = \infty$.

k	x^-	x^+	f_k	$\hat{f}_k$	$f_k^2/\hat{f}_k$
1	$-\infty$	20.68	11	8.904	13.589
2	20.68	22.07	11	10.019	12.077
3	22.07	23.46	7	12.042	4.069
4	23.46	24.85	9	10.054	8.057
5	24.85	26.24	9	5.830	13.893
6	25.24	27.63	2	2.348	1.1704
7	27.63	∞	1	0.802	1.246
Σ	-	-	50	50	54.634

Upravíme krajní třídy, aby byla splněna podmínka teoretické četnosti.

k	x^-	x^+	f_k	$\hat{f}_k$	$f_k^2/\hat{f}_k$
1	$-\infty$	20.68	11	8.904	13.589
2	20.68	22.07	11	10.019	12.077
3	22.07	23.46	7	12.042	4.069
4	23.46	24.85	9	10.054	8.057
5	24.85	∞	12	8.980	16.035
Σ	-	-	50	50	53.827

Testové kritérion t spočítáme pomocí vzorce

$$t = \left(\sum_{k=1}^{7} \frac{f_k^2}{\hat{f}_k} \right) - n = 53.827 - 50 = 3.827.$$

Doplněk kritického oboru odpovídá $\overline{W}_\alpha = \langle 0; \chi^2_{1-\alpha} \rangle$, kde $\chi^2_{1-\alpha}$ je $(1-\alpha)$-kvantil Pearsonova rozdělení s $m - q - 1$ stupni volnosti, kde m je počet tříd a q je počet parametrů spočítaných pomocí bodového odhadu.

$$\overline{W}_{0.05} = \langle 0; \chi^2_{0.95}(2) \rangle = \langle 0; 5.991 \rangle$$

Jelikož $t \in \overline{W}_{0.05}$, **nezamítáme** nulovou hypotézu $H_0 : X \sim N(\mu, \sigma^2)$.

Nyní otestujeme soubor Y rovněž pomocí testu dobré shody. Sestavíme nulovou hypotézu $H_0 : Y \sim N(\mu, \sigma^2)$ Provedeme bodový odhad parametru $\mu : \mu \overset{odhad}{=} \overline{y}$.

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = 24.215$$

Nyní provedeme bodový odhad parametru $\sigma^2 : \sigma^2 \overset{odhad}{=} s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2 = 3.162$$

Z naměřených dat je zjevné, že variační obor činí $\langle 20.65; 29.22 \rangle$ a rozpětí je $29.22 - 20.65 = 8.57$. Pro volbu počtu tříd využijeme Sturgessovo pravidlo. Počet tříd označme m.

$$m = \left\lceil 1 + \frac{\log(n)}{\log(2)} \right\rceil = \left\lceil 1 + \frac{\log(50)}{\log(2)} \right\rceil = 7$$

Nyní sestavíme tříděný statistický soubor s četnostmi jednotlivých tříd. Symbol f_k zastupuje skutečnou četnost k-té třídy. Označme symbolem y_k^- infimum k-té třídy a symbolem y_k^+ supremum k-té třídy. Pak teoretickou četnost $\hat{f}_k$ třídy k spočteme pomocí vzorce

$$\hat{f}_k = n \cdot \left(\lim_{y \to y_k^+} \Phi\left(\frac{y - \overline{y}}{s}\right) - \lim_{y \to y_k^-} \Phi\left(\frac{y - \overline{y}}{s}\right) \right),$$

kde Φ je distribuční funkce náhodné veličiny s normovaným normálním rozdělením $U \sim N(0,1)$. Jelikož distribuční funkce Φ je definovaná na celé množině $\mathbb{R}$, zvolíme $y_1^- = -\infty$ a $y_{10}^+ = \infty$.

k	y^-	y^+	f_k	$\hat{f}_k$	$f_k^2/\hat{f}_k$
1	$-\infty$	22.04	5	4.700	5.319
2	21.874	23.43	11	8.549	14.153
3	23.099	24.82	9	12.955	6.252
4	24.323	26.21	13	12.447	13.577
5	25.547	27.60	10	7.583	13.188
6	26.771	28.99	1	2.927	0.342
7	27.996	$+\infty$	1	0.838	1.193
Σ	-	-	50	50	54.025

Upravíme krajní třídy, aby byla splněna podmínka teoretické četnosti.

k	y^-	y^+	f_k	$\hat{f}_k$	$f_k^2/\hat{f}_k$
1	$-\infty$	22.04	5	4.700	5.319
2	21.874	23.43	11	8.549	14.153
3	23.099	24.82	9	12.955	6.252
4	24.323	26.21	13	12.447	13.577
5	25.547	27.60	12	11.348	12.689
Σ	-	-	50	50	51.990

Testové kritérion t spočítáme pomocí vzorce

$$t = \left(\sum_{k=1}^{7} \frac{f_k^2}{\hat{f}_k} \right) - n = 51.990 - 50 = 1.990.$$

Doplněk kritického oboru odpovídá $\overline{W}_\alpha = \langle 0; \chi^2_{1-\alpha} \rangle$, kde $\chi^2_{1-\alpha}$ je $(1-\alpha)$-kvantil Pearsonova rozdělení s $m - q - 1$ stupni volnosti, kde m je počet tříd a q je počet parametrů spočítaných pomocí bodového odhadu.

$$\overline{W}_{0.05} = \langle 0; \chi^2_{0.95}(2) \rangle = \langle 0; 5.991 \rangle$$

Jelikož $t \in \overline{W}_{0.05}$, **nezamítáme** nulovou hypotézu $H_0 : Y \sim N(\mu, \sigma^2)$.

Ani v jednom případě jsme nezamítli hypotézu o normalitě statistických souborů, pročež budeme normalitu nadále předpokládat.

Test rovnosti rozptylů

Ukázali jsme, že data obou souborů můžeme považovat za normální. Nyní pomocí F-testu rozhodneme, zda mají tyto náhodné výběry stejný rozptyl. Předpokladem tedy je $X \sim N(\mu_x, \sigma_x^2), Y \sim N(\mu_y, \sigma_y^2)$, kde σ_x^2 a σ_y^2 jsou neznámé.
Nulová hypotéza: $H_0 : \sigma_x^2 = \sigma_y^2$, alternativní hypotéza $H_A : \sigma_x^2 \neq \sigma_y^2$. Spočtěme testovací kritérion t:

$$t = \frac{s^2(X)}{s^2(Y)} = \frac{5.1387}{3.162} = 1.625.$$

Hodnoty $s^2(X)$ a $s^2(Y)$ jsme spočítali v předchozí části příkladu. Doplněk kritického oboru odpovídá

$$\overline{W}_{0.05} = \langle F_{0.025}(n-1, m-1); F_{0.975}(n-1, m-1)\rangle,$$

kde $F_{0.025}(n-1, m-1)$ je (0.025)-kvantil Fisherova-Snedecorova rozdělení s $k_1 = n-1$ a $k_2 = m-1$ stupni volnosti, kde $m = n = 50$.

$$\overline{W}_{0.05} = \langle F_{0.025}(49, 49); F_{0.975}(49, 49)\rangle = \langle 0.567; 1.762\rangle$$

Jelikož $1.625 \in \langle 0.567; 1.762\rangle$, nulovou hypotézu nezamítáme. Nezamítli jsme tedy hypotézu o rovnosti rozptylů, pročež budeme rovnost rozptylů nadále předpokládat.

Studentův dvouvýběrový test

Ukázali jsme, že data obou souborů můžeme považovat za normální se stejnými rozptyly. Nyní pomocí Studentova dvouvýběrového testu rozhodneme, zda je některý z poskytovatelů pro naše potřeby lepší.

Nejprve tedy budeme testovat nulovou hypotézu $H_0 : \mu_x - \mu_y = 0 \Rightarrow \mu_x = \mu_y$ oproti alternativní hypotéze $H_A : \mu_x < \mu_y$. Spočteme testovací kritérion

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{(n-1)\cdot s^2(X) + (m-1)\cdot s^2(Y)}} \cdot \sqrt{\frac{n \cdot m \cdot (n+m-2)}{n+m}}$$

Hodnoty $\overline{x}, \overline{y}, s^2(X), s^2(Y)$ využijeme z předchozích výpočtů. Dále $n = m = 50$.

$$t = \frac{22.7716 - 24.215}{\sqrt{49 \cdot 5.1387 + 49 \cdot 3.162}} \cdot \sqrt{\frac{2500 \cdot 98}{100}} = -3.543$$

Doplněk kritického oboru spočteme jako

$$\overline{W}_{0.05} = \langle -t_{0.95}; +\infty),$$

kde $t_{0.95}$ je (0.95)-kvantil Studentova rozdělení s $n+m-2$ stupni volnosti.

$$\overline{W}_{0.05} = \langle -1.661; +\infty)$$

Zjistili jsme, že $-3.543 \notin \langle -1.661; +\infty)$, tudíž **zamítáme** nulovou hypotézu $H_0 : \mu_x - \mu_y = 0$ a **přijímáme** alternativní hypotézu $H_A : \mu_x < \mu_y$.
Jelikož střední hodnota odezvy u původního poskytovatele je nižší než střední hodnota u nového poskytovatele, zjistili jsme, že původní poskytovatel je pro nás výhodnější.

2. úloha

Budeme zjišťovat, zda doba plnění určité úlohy závisí na době, kdy je tato úloha plněna, na míře hluku v okolí nebo na obou těchto faktorech současně. Data shrnuje následující tabulka. Modře jsou podbarvená ta data, která byla v souladu se zadáním doplněna.

	faktor2			
faktor1	ticho	hudba	hluk	křik
ráno	6	7	8	13
	8	8	7	21
	11	12	20	18
		10		
poledne	8	5	10	14
	13	11	17	19
	7	7	11	
			13	
večer	7	6	12	13
	8	8	17	17
	6	16	11	15
		15		22
				18

Dle zadání předpokládáme rovnost rozptylů. Dále budeme počítat nevyváženou dvoufaktorovou ANOVU, neboť počet měření v jednotlivých skupinách se napříč tabulkou různí. Zavedeme si následující notaci:

počet skupin faktoru 1: $A = 3$
počet skupin faktoru 2: $B = 4$
počet všech měření: $n = 40$
$\alpha = 0.05$
Sestavíme nulové a alternativní hypotézy.

$$H_{0_1} : \mu_{a_1} = \mu_{a_2} = \mu_{a_3}, H_{A_1} : \exists i, j \in \{1, 2, 3\} : \mu_{a_i} \neq \mu_{a_j}$$

$$H_{0_2} : \mu_{b_1} = \mu_{b_2} = \mu_{b_3} = \mu_{b_4}, H_{A_2} : \exists i, j \in \{1, 2, 3, 4\} : \mu_{b_i} \neq \mu_{b_j}$$

$$H_{0_3} : \text{Mezi faktory není interakce}, H_{A_3} : \text{Mezi faktory je interakce},$$

kde a_1, a_2, a_3 jsou hodnoty faktoru 1 (po řadě ráno, poledne, večer) a b_1, b_2, b_3, b_4 jsou hodnoty faktoru 2 (po řadě ticho, hudba, hluk, křik).

	Stupně volnosti	Součty čtverců	Střední hodnota čtverců	Spočtené F	Kritické F
Faktor1	2				
Faktor2	3				
Interakce	6				
Chyba	28				
Souhrn	39				

Nyní budeme vyplňovat uvedenou tabulku. Sloupec *stupně volnosti* odpovídá pro první faktor hodnotě $A - 1 = 2$ a pro druhý faktor hodnotě $B - 1 = 3$. V případě interakce zapíšeme $(A - 1) \cdot (B - 1) = 6$ a v případě chyby $n - A \cdot B = 28$. Počet stupňů volnosti pro řádek souhrn odpovídá hodnotě $n - 1 = 39$.

Pro výpočet součtu čtverců potřebujeme najít *faktor korekce CF* pomocí vzorce

$$CF = \frac{1}{n} \cdot \left(\sum_{i=1}^{n} y_i \right)^2 = \frac{1}{40} \cdot 225625 = 5640.625,$$

kde y_i odpovídá $i-$tému měření z tabulky. Nyní spočteme hodnoty *součty čtverců* pro oba faktory (SS_A, SS_B):

$$SS_A = \frac{1}{n_{a_1}} \cdot \left(\sum_{i=1}^{n_{a_1}} a_{1,i} \right)^2 + \frac{1}{n_{a_2}} \cdot \left(\sum_{i=1}^{n_{a_2}} a_{2,i} \right)^2 + \frac{1}{n_{a_3}} \cdot \left(\sum_{i=1}^{n_{a_3}} a_{3,i} \right)^2 - CF = 17.96$$

$$SS_B = \frac{1}{n_{b_1}} \cdot \left(\sum_{i=1}^{n_{b_1}} b_{1,i} \right)^2 + \frac{1}{n_{b_2}} \cdot \left(\sum_{i=1}^{n_{b_2}} b_{2,i} \right)^2 + \frac{1}{n_{b_3}} \cdot \left(\sum_{i=1}^{n_{b_3}} b_{3,i} \right)^2 + \frac{1}{n_{b_4}} \cdot \left(\sum_{i=1}^{n_{b_4}} b_{4,i} \right)^2 - CF = 447.6917,$$

kde a_1, a_2, a_3 jsou hodnoty faktoru 1 (po řadě ráno, poledne, večer) a b_1, b_2, b_3, b_4 jsou hodnoty faktoru 2 (po řadě ticho, hudba, hluk, křik).
Součet čtverců pro interakci spočteme následovně:

$$SS_{int} = \sum_{i=1}^{A} \sum_{j=1}^{B} \left(\frac{1}{n_{i,j}} \left(\sum_{k=1}^{n_{i,j}} y_{i,j,k} \right)^2 \right) - SS_A - SS_B - CF = 17.9733,$$

kde $n_{i,j}$ představuje počet měření ve shodě dvou skupin obou faktorů (např. ráno a ticho) a $y_{i,j,k}$ odpovídá $k-$tému měření z této skupiny.
Součet čtverců pro souhrn spočítáme pomocí vzorce:

$$SS_{souhrn} = \left(\sum_{i=1}^{n} y_i^2 \right) - CF = 872.375,$$

kde y_i zastupuje $i-$té měření v tabulce.
Součet čtverců pro chybu nakonec spočteme pomocí vzorce

$$SS_{chyba} = SS_{souhrn} - SS_1 - SS2 - SSint = 388.75$$

Střední hodnoty čtverců (MS) spočteme pro první čtyři řádky tak, že podělíme příslušnou hodnotu SS počty stupňů volnosti. Hodnoty shrnuje následující tabulka:

	Stupně volnosti	Součty čtverců	Střední hodnota čtverců	Spočtené F	Kritické F
Faktor1	2	17.96	8.98		
Faktor2	3	447.6917	149.23		
Interakce	6	17.9733	2.99		
Chyba	28	388.75	13.88		
Souhrn	39	872.375	-		

Hodnoty *Spočtené F* pro první 3 řádky spočteme následovně:

$$\text{Spočtené } \mathrm{F}_1 = MS_1/MS_{chyba} = 0.647$$

$$\text{Spočtené } \mathrm{F}_2 = MS_2/MS_{chyba} = 10.751$$

$$\text{Spočtené } \mathrm{F}_{int} = MS_{int}/MS_{chyba} = 0.215$$

Hodnoty *Kritické F* pro první 3 řádky vyčteme z tabulek distribuční funkce Fisherova-Snedecorova rozdělení následovně:

$$\text{Kritické F}_1 = F_{0.95}(A-1, n-A\cdot B) = F_{0.95}(2, 28) = 3.34$$

$$\text{Kritické F}_2 = F_{0.95}(B-1, n-A\cdot B) = F_{0.95}(3, 28) = 2.947$$

$$\text{Kritické F}_{int} = F_{0.95}((A-1)\cdot(B-1), n-A\cdot B) = F_{0.95}(6, 28) = 2.445$$

	Stupně volnosti	Součty čtverců	Střední hodnota čtverců	Spočtené F	Kritické F
Faktor1	2	17.96	8.98	0.647	3.34
Faktor2	3	447.6917	149.23	10.751	2.947
Interakce	6	17.9733	2.99	0.215	2.445
Chyba	28	388.75	13.88	-	-
Souhrn	39	872.375	-	-	-

Jelikož u faktoru 1 platí, že *Spočtené F* $<$ *Kritické F* (tedy $0.647 < 3.34$), **nezamítáme** nulovou hypotézu $\mu_{a_1} = \mu_{a_2} = \mu_{a_3}$.
Jelikož u faktoru 2 platí, že *Spočtené F* $>$ *Kritické F* (tedy $10.751 > 2.947$), **zamítáme** nulovou hypotézu $H_{0_2} : \mu_{b_1} = \mu_{b_2} = \mu_{b_3} = \mu_{b_4}$ a přijímáme alternativní hypotézu $H_{A_2} : \exists i, j \in \{1, 2, 3, 4\} : \mu_{a_i} \neq \mu_{a_j}$.
Jelikož pro interakci platí, že *Spočtené F* $<$ *Kritické F* (tedy $0.215 < 2.445$), **nezamítáme** nulovou hypotézu H_{0_3} : Mezi faktory není interakce.

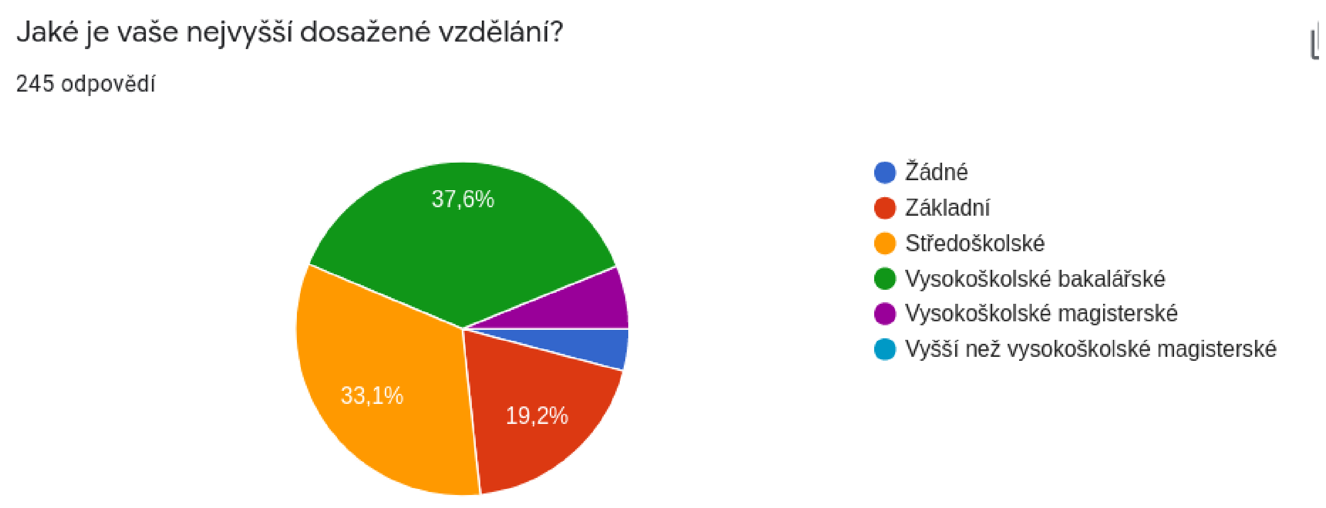
3. úloha

V této úloze testujeme nezávislost dvou kvalitativních proměnných. Konkrétně zjišťujeme, zda existuje závislost mezi nejvyšším dosaženým vzděláním a průměrným počtem přečtených knih za rok.

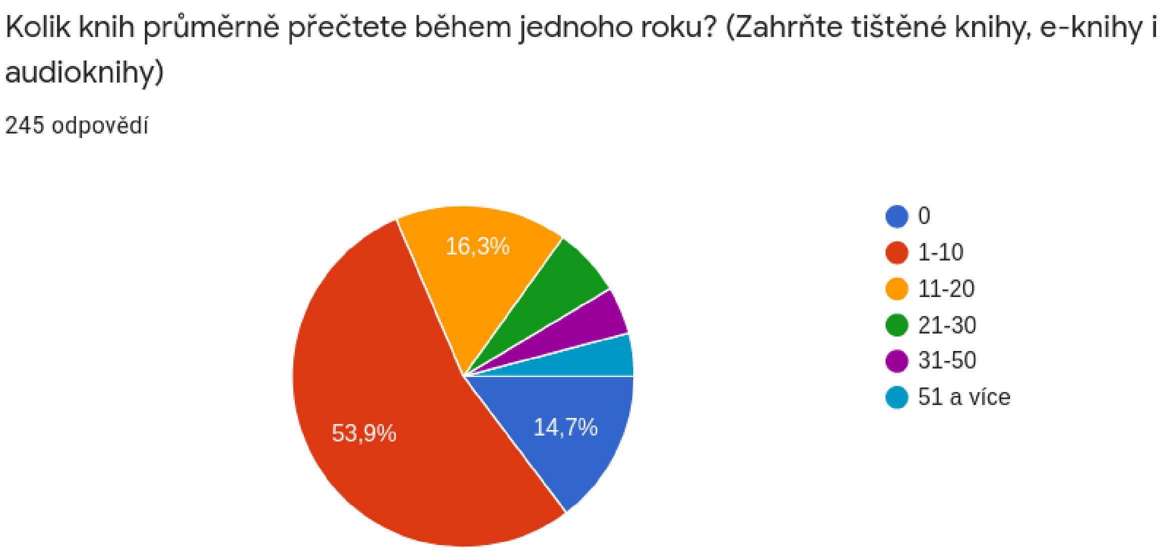
V rámci nulové hypotézy tedy předpokládáme nezávislost $H_0 : \forall i, j : p_{i,j} = p_{\cdot,j} \cdot p_{i,\cdot}$, zatímco v rámci alternativní hypotézy připouštíme existující závislost $H_A : \exists i, j : p_{i,j} \neq p_{\cdot,j} \cdot p_{i,\cdot}$.
Pro potřeby sběru dat byl sestaven dotazník s následujícími otázkami a možnostmi odpovědi:

1. Jaké je vaše nejvyšší dosažené vzdělání?
    - Žádné
    - Základní
    - Středoškolské
    - Vysokoškolské bakalářské
    - Vysokoškolské magisterské
    - Vyšší než vysokoškolské magisterské
2. Kolik knih průměrně přečtete během jednoho roku? (Zahrňte tištěné knihy, e-knihy i audioknihy)
    - 0
    - 1-10
    - 11-20
    - 21-30
    - 31-50
    - 51 a více

Sběr dat byl uskutečněn mezi 11. a 28. listopadem 2021 prostřednictvím sociálních sítí Facebook a Discord. Dotazník byl vytvořen pomocí systému Google forms. Dotazník vyplnilo dohromady 245 respondentů. Odpovědi shrnují následující koláčové grafy:



Otázka 1



Otázka 2

Kategoriální analýza

Na základě sesbíraných dat sestavíme kontingenční tabulku.

↓ Vzdělání, Počty knih →	0	1-10	11-20	21-30	31-50	51 a více	Σ
Žádné	2	2	3	1	1	1	10
Základní	5	23	6	6	5	2	47
Středoškolské	11	42	14	6	2	6	81
Vysokoškolské bakalářské	15	62	12	1	1	1	92
Vysokoškolské magisterské	3	3	5	2	2	0	15
Vyšší než vysokoškolské magisterské	0	0	0	0	0	0	0
Σ	36	132	40	16	11	10	245

Jelikož pro takto sesbíraná data nelze použít kategoriální analýzu, je třeba sjednotit některé třídy. Seskupíme konkrétně třídy Žádné a Základní do jedné třídy, do další třídy sjednotíme Vysokoškolské bakalářské, Vysokoškolské magisterské a Vyšší než vysokoškolské magisterské. Co se týče počtu přečtených knih na rok, sjednotíme do jedné třídy $21-30, 31-50$ a 50 a více.

↓ Vzdělání, Počty knih →	0	1-10	11-20	21 a více	Σ
Základní a nižší	7	25	9	16	57
Středoškolské	11	42	14	14	81
Vysokoškolské bakalářské a vyšší	18	65	17	7	107
Σ	36	132	40	37	245

Nyní spočítáme hodnoty

$$\frac{n_{\cdot,j} \cdot n_{i,\cdot}}{n_{i,j}}.$$

Lze snadno nahlédnout, že platí

$$\forall i, j : \frac{n_{\cdot,j} \cdot n_{i,\cdot}}{n_{i,j}} > 5$$

↓ Vzdělání, Počty knih →	0	1-10	11-20	21 a více	Σ
Základní a nižší	8.376	30.710	9.306	8.608	57
Středoškolské	11.902	42.641	13.224	12.233	81
Vysokoškolské bakalářské a vyšší	15.722	57.649	17.469	16.159	107
Σ	36	132	40	37	245

Nyní spočteme hodnoty

$$n_{i,j} - \frac{n_{\cdot,j} \cdot n_{i,\cdot}}{n_{i,j}}$$

↓ Vzdělání, Počty knih →	0	1-10	11-20	21 a více	Σ
Základní a nižší	-1.376	-5.710	-0.306	7.392	0
Středoškolské	-0.902	-1.641	0.776	1.767	0
Vysokoškolské bakalářské a vyšší	2.278	7.351	-0.469	-9.159	0
Σ	0	0	0	0	0

V dalším kroku spočteme hodnoty

$$\frac{\left(n_{i,j} - \frac{n_{\cdot,j} \cdot n_{i,\cdot}}{n_{i,j}}\right)^2}{\frac{n_{\cdot,j} \cdot n_{i,\cdot}}{n_{i,j}}}$$

↓ Vzdělání, Počty knih →	0	1-10	11-20	21 a více	Σ
Základní a nižší	0.226	1.062	0.010	6.347	7.645
Středoškolské	0.068	0.062	0.045	0.255	0.431
Vysokoškolské bakalářské a vyšší	0.330	0.937	0.013	5.192	6.471
Σ	0.624	2.061	0.068	11.794	14.547

Nalezli jsme naše testovací kritérion

$$\chi^2 = \sum_{i=1}^{3} \sum_{j=1}^{4} \frac{\left(n_{i,j} - \frac{n_{\cdot,j} \cdot n_{i,\cdot}}{n_{i,j}}\right)^2}{\frac{n_{\cdot,j} \cdot n_{i,\cdot}}{n_{i,j}}} = 14.547$$

Nyní nalezneme doplněk kritického oboru $\overline{W}_{0.05} = \langle 0; \chi^2_{0.95}((3-1) \cdot (4-1)) = \langle 0; \chi^2_{0.95}(6) \rangle$, kde $\chi^2_{0.05}(6)$ je (0.05)-kvantil Pearsonova rozdělení s 6 stupni volnosti.

$$\chi^2_{0.05}(6) = 12.592, \text{tedy platí, že } 14.547 \notin \langle 0; 12.592 \rangle,$$

pročež **zamítáme** nulovou hypotézu $H_0 : \forall i, j : p_{i,j} = p_{\cdot,j} \cdot p_{i,\cdot}$ a naopak **přijímáme** alternativní hypotézu $H_A : \exists i, j : p_{i,j} \neq p_{\cdot,j} \cdot p_{i,\cdot}$. Na dané hladině významnosti $\alpha = 0.05$ tedy existuje závislost mezi nejvyšším dosaženým vzděláním a počtem přečtených knih za rok.

Test na normalitu

Nejprve otestujme oba statistické soubory na normalitu pomocí χ^2 testu dobré shody. Začneme se souborem X. Sestavíme nulovou hypotézu $H_0 : X \sim N(\mu, \sigma^2)$, kde μ a σ^2 jsou neznámé parametry. Volíme $\alpha = 0.05$.

Provedeme bodový odhad parametru $\mu : \mu \overset{odhad}{=} \overline{x}$.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 22.7716$$

Nyní provedeme bodový odhad parametru σ^2 : $\sigma^2 \overset{odhad}{=} s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = 5.1387$$

Z naměřených dat je zjevné, že variační obor činí $\langle 19.29; 29.02\rangle$ a rozpětí je $29.02 - 19.29 = 9.73$. Pro volbu počtu tříd využijeme Sturgessovo pravidlo. Počet tříd označme m.

$$m = \left\lceil 1 + \frac{\log(n)}{\log(2)} \right\rceil = \left\lceil 1 + \frac{\log(50)}{\log(2)} \right\rceil = 7$$

Nyní sestavíme tříděný statistický soubor s četnostmi jednotlivých tříd. Symbol f_k zastupuje skutečnou četnost k-té třídy. Označme symbolem x_k^- infimum k-té třídy a symbolem x_k^+ supremum k-té třídy. Pak teoretickou četnost $\hat{f}_k$ třídy k spočteme pomocí vzorce

$$\hat{f}_k = n \cdot \left(\lim_{x \to (x_k^+)^-} \Phi\left(\frac{x - \overline{x}}{s}\right) - \lim_{x \to (x_k^-)^+} \Phi\left(\frac{x - \overline{x}}{s}\right) \right),$$

kde Φ je distribuční funkce náhodné veličiny s normovaným normálním rozdělením $U \sim N(0,1)$. Jelikož distribuční funkce Φ je definovaná na celé množině $\mathbb{R}$, zvolíme $x_1^- = -\infty$ a $x_7^+ = \infty$.

k	x^-	x^+	f_k	$\hat{f}_k$	$f_k^2/\hat{f}_k$
1	$-\infty$	20.68	11	8.904	13.589
2	20.68	22.07	11	10.019	12.077
3	22.07	23.46	7	12.042	4.069
4	23.46	24.85	9	10.054	8.057
5	24.85	26.24	9	5.830	13.893
6	25.24	27.63	2	2.348	1.1704
7	27.63	∞	1	0.802	1.246
Σ	-	-	50	50	54.634

Upravíme krajní třídy, aby byla splněna podmínka teoretické četnosti.

k	x^-	x^+	f_k	$\hat{f}_k$	$f_k^2/\hat{f}_k$
1	$-\infty$	20.68	11	8.904	13.589
2	20.68	22.07	11	10.019	12.077
3	22.07	23.46	7	12.042	4.069
4	23.46	24.85	9	10.054	8.057
5	24.85	∞	12	8.980	16.035
Σ	-	-	50	50	53.827

Testové kritérion t spočítáme pomocí vzorce

$$t = \left(\sum_{k=1}^{7} \frac{f_k^2}{\hat{f}_k}\right) - n = 53.827 - 50 = 3.827.$$

Doplněk kritického oboru odpovídá $\overline{W}_\alpha = \langle 0; \chi^2_{1-\alpha}\rangle$, kde $\chi^2_{1-\alpha}$ je $(1-\alpha)$-kvantil Pearsonova rozdělení s $m - q - 1$ stupni volnosti, kde m je počet tříd a q je počet parametrů spočítaných pomocí bodového odhadu.

$$\overline{W}_{0.05} = \langle 0; \chi^2_{0.95}(2) \rangle = \langle 0; 5.991 \rangle$$

Jelikož $t \in \overline{W}_{0.05}$, **nezamítáme** nulovou hypotézu $H_0 : X \sim N(\mu, \sigma^2)$.