Téma 3 – Příklad 1

Zadání příkladu

Na základě průzkumu provedeného u čtenářů časopisů A, B, C byla sestavena kontingenční tabulka. Rozhodněte, zda výběr časopisu a bydliště čtenáře jsou znaky závislé, neboli zda se struktura čtenářů z hlediska bydliště u časopisů A, B, C liší.

		Časopis		
Bydliště	A	B	C	Celkem
Liberec	75	75	50	200
Jablonec	40	70	40	150
Česká Lípa	35	5	10	50
Celkem	150	150	100	400

Vypracování příkladu

Jelikož se jedná o dvě slovní proměnné, kdy závislá proměnná je nominální, bude k řešení využito chí-kvadrát testu o nezávislosti dvou kategoriálních proměnných. Tento test je oboustranný, zkoumá závislost proměnné *a* na proměnné *b* a naopak. Postup je obdobný jako u jiného statistického testování hypotéz a probíhá v několika krocích.

1. H_0: výběr časopisu je nezávislý na bydlišti respondenta (obecně *a* a *b* jsou nezávislé proměnné) – v tomto případě formuluje hypotézu jednostranně, jelikož nepředpokládáme, že opačný směr závislosti je smysluplný
   H_1: non H_0
2. Testové kritérium G (někdy označované jako chí-kvadrát) má tuto podobu:
   $$G = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n_{ij} - n'_{ij}\right)^2}{n'_{ij}}$$
   Pro jeho určení je třeba vypočítat tzv. teoretické (hypotetické, očekávané) četnosti za pomoci vzorce:
   $$n'_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$$
   Např. pro četnost n_{11} platí, že (150*200)/400 = 75. Teoretické četnosti jsou v následující tabulce vyznačeny modře a vyjadřují, jak by měla teoreticky kontingenční tabulka vypadat, kdyby mezi proměnnými byla nezávislost.

		Časopis		
Bydliště	A	B	C	Celkem
Liberec	75 75	75 75	50 50	200
Jablonec	40 56,25	70 56,25	40 37,5	150
Česká Lípa	35 18,75	5 18,75	10 12,5	50
Celkem	150	150	100	400

Testové kritériu má potom podobu:
0 + 0 + 0 + 4,694 + 3,361 + 0,167 + 14,083 + 10,083 + 0,5 = 32,889

3. $W \equiv \left\{ G,\ G > \chi^2_{1-\alpha}\left[(r-1)(s-1)\right] \right\}$

   W ≡ (G, G ≥ $X^2(4)$)
   W ≡ (G, G ≥ 9,488)

4. 32,889 ≥ 9,488
   Testové kritérium je prvkem kritického oboru. Nulovou hypotézu zamítáme, přijímáme hypotézu alternativní.

Jelikož byla závislost prokázána, pokusíme se určit i její těsnost. K tomu nám slouží koeficienty kontingence:
Pearsonův:

$$C_P = \sqrt{\frac{G}{G+n}}$$

$C_P = 0{,}276$

nebo
Cramérův:

$$C_C = \sqrt{\frac{G}{n \cdot h}}$$, kde h … min (r-1);(s-1)

$C_C = 0{,}203$

Na 5% hladině významnosti jsme prokázali, že **existuje statisticky významná závislost** mezi četností časopisů a bydlištěm čtenářů.

Řešení v SGP

V tomto případě je výjimečně možné v případě programu Statgraphics užití dat ve formě tabulky (sdružených) četností. Pro správnou funkci stačí zadat pouze sdružené četnosti,

nikoliv okrajové (součtové) hodnoty. Stejně tak proměnná *Bydliste* je v tomto případě pouze pro označení řádků v tabulce. Z výpočetního hlediska nemá vliv.

<untitled>

	Bydliste	A	B	C
1	Liberec	75	75	50
2	Jablonec	40	70	40
3	Ceska Lipa	35	5	10
4				
5				
6				
7				
8				

Procedura v SGP: Describe - Categorical Data - Contingency Tables
(Frequency Table, Chi-Square Test, Summary Statistics)

Při vstupním dialogu zadáme jako *Colums A, B* a *C*, tedy sloupce kontingenční tabulky. Do políčka *Labels* je možné zadat *Bydliste,* aby kontingenční tabulka vypadala stejně, jako naše vstupní tabulka.

V okně Frequency Table je možné vidět kompletní kontingenční tabulku. Z doplňkových možností Pane Options si můžeme zvolit různé pohledy na relativní vyjádření četností nebo také hodnoty očekávaných četností a dílčích příspěvků k testovému kritériu.

Frequency Table

	A	B	C	Row Total
Liberec	75	75	50	200
	18,75%	18,75%	12,50%	50,00%
	37,50%	37,50%	25,00%	
	50,00%	50,00%	50,00%	
	75,00	75,00	50,00	
	0,00	0,00	0,00	
Jablonec	40	70	40	150
	10,00%	17,50%	10,00%	37,50%
	26,67%	46,67%	26,67%	
	26,67%	46,67%	40,00%	
	56,25	56,25	37,50	
	4,69	3,36	0,17	
Ceska Lipa	35	5	10	50
	8,75%	1,25%	2,50%	12,50%
	70,00%	10,00%	20,00%	
	23,33%	3,33%	10,00%	
	18,75	18,75	12,50	
	14,08	10,08	0,50	
Column Total	150	150	100	400
	37,50%	37,50%	25,00%	100,00%

Cell contents:
- Observed frequency
- Percentage of table
- Percentage of row
- Percentage of column
- Expected frequency
- Contribution to chi-square

Hlavní výstup je zobrazen v okně Test of Independence. Položka *Statistic* zobrazuje hodnotu testové statistiky. Hodnota *P-Value* ukazuje maximální možnou hladinu významnosti, na které nezamítáme nulovou hypotézu. V tomto případě tedy na 5% hladině významnosti zamítáme nulovou hypotézu o nezávislosti obou sledovaných znaků.

Tests of Independence

Test	*Statistic*	*Df*	*P-Value*
Chi-Square	32,889	4	0,0000

Stejně jako v ostatních případech používá Statgraphics k vyhodnocení testu hypotézy ukazatel P-Value (v jiných programech např. Significance Level apod.), což je maximální možná hodnota hladiny významnosti, na které ještě nezamítáme nulovou hypotézu. Není tudíž nutné určovat kritický obor pro námi zvolenou hladinu významnosti.

Hodnoty koeficientů kontingence nalezneme v okně *Summary Statistics*. Pearsovou koeficientu odpovídá údaj *Contingency Coeff., Cramer's V* je koeficient Cramerův.

Summary Statistics

		With Rows	*With Columns*
Statistic	*Symmetric*	*Dependent*	*Dependent*
Lambda	0,0667	0,0000	0,1200
Uncertainty Coeff.	0,0420	0,0443	0,0399
Somer's D	-0,0550	-0,0524	-0,0579
Eta		0,0840	0,2074

Statistic	*Value*	*P-Value*	*Df*
Contingency Coeff.	0,2756		
Cramer's V	0,2028		
Conditional Gamma	-0,0870		
Pearson's R	-0,0607	0,2256	398
Kendall's Tau b	-0,0551	0,2251	
Kendall's Tau c	-0,0516		

Interpretace

Na 5% hladině významnosti bylo prokázáno, že struktura čtenářů se z hlediska bydliště u časopisů A, B a C statisticky významně liší. Vzhledem k tomu, že považujeme za smysluplný pouze jednostranný vliv, je možné formulovat odpověď takto. Tato závislost je slabá. Pearsonův kontingenční koeficient má hodnotu 0,2756 nebo Cramerův 0,2028.

Řešení v Excelu

Dle verze programu je možné použít funkce **CHITEST(aktuální;očekávané)** k provedení výše uvedeného testu. Aktuální četnosti je termín pro četnosti empirické. Výstupem funkce je hodnota P-Value. Výpočet očekávaných četností je ovšem nutné provést pomocí matematických operací (aplikací vzorců).