Analýza kategorizovaných dat v sociologii
Úloha č. 1

Kateřina Škařupová **Přemysl Maršík**
43275 **13477**

Do analýzy jsme vstoupili s daty Leo A. Goodmana z článku *A Modified Multiple Regression Approach to the Anylysis of Dichotomous Variables* (1972). Goodman zkoumal souvislost mezi původem (sever – jih), barvou pleti a lokací jednotky u amerických vojáků bojujících ve Druhé světové válce a preferencemi lokací vojenských jednotek. Pokusili jsme se částečně jeho postup zopakovat.

Nejprve jsme sestrojili všechny dvojrozměrné kontingenční tabulky, v nichž je vysvětlovaná proměnná (preferovaná lokace jednotky) ve sloupci a vysvětlující proměnné (barva pleti, původ a současná lokace) v řádcích.

Tab. 1. Souvislost mezi preferovanou jednotkou a současnou lokací.

	preference	
lokace	S	J	Total
S	1,829	644	2,473
Row%	73.96	26.04	100.00
J	2,222	3,341	5,563
Row%	39.94	60.06	100.00
Total	4,051	3,985	8,036
	50.41	49.59	100.00

Z řádkových procent bychom mohli usuzovat na to, že proměnná „současná lokace" ovlivňuje postoj vojáka na k preferovanému umístění v jednotce. Na první pohled se zdá, že vojáci právě umístění na severu budou spíše preferovat lokaci na severu a vojáci právě umístění v jižní jednotce si pravděpodobně zvolí opět jižní jednotku. Případný vliv další proměnné zatím nemůžeme vyloučit.

Tab. 2. Souvislost mezi preferovanou jednotkou a původem vojáka.

	preference	
původ	S	J	Total
S	3,092	958	4,050
Row%	76.35	23.65	100.00
J	959	3,027	3,986
Row%	24.06	75.94	100.00
Total	4,051	3,985	8,036
	50.41	49.59	100.00

Druhá tabulka poskytuje hrubý pohled na vztah mezi původem vojáka a preferovanou jednotkou. Tady je situace podobná vztahu mezi současnou a preferovanou lokací.

Tab. 3. Souvislost mezi preferovanou jednotkou a barvou pleti.

	preference		
barva pleti	S	J	Total
černá	2,027	2,268	4,295
Row%	47.19	52.81	100.00
bílá	2,024	1,717	3,741
Row%	54.10	45.90	100.00
Total	4,051	3,985	8,036
	50.41	49.59	100.00

Ve třetí tabulce jsme už nedostali zřetelnou souvislost mezi vysvětlující a vysvětlovanou proměnnou – barvou pleti a preferovanou lokalitou. Nemůžeme zatím posoudit, zda jsou rozdíly v preferencích statisticky významné.

Následně jsme nechali Statu, aby za nás pro předchozí tabulky vypočítala očekávané četnosti a adjustované reziduály a provedli jsme test nezávislosti. V tabulkách neuvádíme marginální četnosti, neboť se neliší od tabulek 1, 2, a 3.

Tab. 4. Souvislost mezi preferovanou jednotkou a současnou lokací. Očekávané frekvence a adjustované reziduály.

	preference	
lokace	S	J
S	1829	644
Exp. Freq.	1246.655	1226.345
Adj. Res.	28.150	-28.150
J	2222	3341
Exp. Freq.	2804.345	2758.655
Adj. Res.	-28.150	28.150

Pro proměnné uvedené v tabulce Tab.4 jsme provedli test nezávislosti (chí kvadrát), jeho hodnota byla 792,4212, statisticky významná na 95% hladině významnosti. Zamítáme tedy nulovou hypotézu o nezávislosti preference lokalce na současném umístění vojáka.

Tab. 5. Souvislost mezi preferovanou jednotkou a původem vojáka. Očekávané frekvence a adjustované reziduály.

	preference	
původ	S	J
S	3092	958
Exp. Freq.	2041.631	2008.369
Adj. Res.	46.872	-46.872
J	959	3027
Exp. Freq.	2009.369	1976.631
Adj. Res.	-46.872	46.872

Stejný postup jsme zopakovali u proměnné „původ". I tady zamítáme nulovou hypotézu o nezávislosti proměnných, neboť chí kvadrát nabyl hodnoty 2200 a opět byl statisticky významný.

Tab. 6. Souvislost mezi preferovanou jednotkou a barvou pleti. Očekávané frekvence a adjustované reziduály.

	preference	
barva pleti	S	J
černá	2027	2268
Exp. Freq.	2165.138	2129.862
Adj. Res.	-6.179	6.179
bílá	2024	1717
Exp. Freq.	1885.862	1855.138
Adj. Res.	6.179	-6.179

Překvapivě jsme museli zamítnout i nulovou hypotézu o nezávislosti proměnné „preference" na proměnné „barva pleti". Hodnota chí kvadrátu zde byla ze všech případů nejnižší (38,1770), nicméně také statisticky významná.

Pro všechny tři první tabulky jsme vypočítali relativní riziko.

Tab. 7. Relativní rizika k tabulkám 1, 2, 3 uvedená pro oba sloupce.

	preference	
rel. risk.	S	J
lokace(S/J)	1,852	0,429
původ(S/J)	3,298	0,311
barva(Č/B)	0,872	1,151

Ve zkoumaném problému nelze jednoznačně popsat jeden z možných výsledků jako úspěch, proto uvádíme v tabulce Tab. 7. relativní rizika obou odpovědí. Každý řádek tabulky obsahuje hodnoty relativního rizika pro příslušnou dvojici proměnných.

Hodnota 1,852 nám říká, že 1,852krát více vojáků umístěných v severní jednotce, než těch v jižní, preferuje lokaci na severu. Naopak preference vojáků umístěných na jihu pro jižní jednotku jsou 2,331krát (1/0,429) vyšší než u vojáků ze severní jednotky.

Podle původu jsou preference vojáků rozděleny ještě silněji. 3,3krát více seveřanů než jižanů by volilo severní jednotku a naopak 3,22krát (1/0,311) více jižanů než seveřanů volí jih.

Barva pleti ovlivňuje preference (s ohledem na relativní riziko) jen málo. Pouze 1,151krát více bělochů než Afroameričanů by rádo severní jednotku. A také jen 1,147krát (1/0,872) více Afroameričanů než bělochů volí jižní umístění.

S takto vypočtenými relativními riziky úzce souvisí i hodnoty poměru šancí. Poměr šancí v jednotlivých tabulkách je podílem relativních rizik v příslušných řádcích, tedy

$$ODR = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1}{\pi_2}\cdot\frac{(1-\pi_2)}{(1-\pi_1)} = RR_S \cdot \frac{1}{RR_J} \quad ,$$

kde RR_S značí relativní riziko pro volbu severní jednotky a RR_J pro volbu jižní.

Poměr šancí pro tabulku Tab. 1. tedy činí 4,317, což ukazuje, že šance že náhodně vybraný voják ze severní jednotky preferuje svou jednotku je více než 4krát větší než šance, že tutéž jednotku preferuje náhodný voják umístěný na jihu.

V tabulce Tab. 2. jest poměr šancí roven 10,6. Tuto veličinu bychom mohli použít, kdybychom chtěli někoho přesvědčit, jak silně je ovlivněna preference jednotky původem. Je desetkrát větší šance, že voják ze severu preferuje severní jednotku, než šance, že ji preferuje voják z jihu.

Jak už bylo patrno z předchozích závěrů, barva pleti není rozhodujícím faktorem pro volbu oblíbené jednotky. Poměr šancí v tabulce Tab. 3. je 0,758, tedy šance, že Afroameričan preferuje severní jednotku, je jen o málo menší než šance, že ji preferuje běloch.

Kontrastní podíly šancí v tabulkách jsou: 0,232 pro Tab. 1., 0,094 pro Tab. 2. a 1,319 pro Tab. 3. Lze snadno ukázat, že kontrastní podíly šancí jsou převrácené hodnoty prostých podílů šancí.

$$\frac{\pi_2/(1-\pi_2)}{\pi_1/(1-\pi_1)} = \frac{\pi_2}{\pi_1}\cdot\frac{(1-\pi_1)}{(1-\pi_2)} = 1/ODR$$

Pak ovšem

$$\ln(1/ODR) = \ln(1) - \ln(ODR) = -\ln(ODR) \quad .$$

Hodnoty logaritmů prostého a kontrastního poměru šancí jsou navzájem opačné.

Tab. 8. Pokus o smysluplnou projekci 4D tabulky do dvou rozměrů:

AFROAM.:	lokace			
	S		J	
	preference			
původ	S	J	S	J
S	387	36	876	250
J	383	270	381	1,712

BĚLOŠI:	lokace			
	S		J	
	preference			
původ	S	J	S	J
S	955	162	874	510
J	104	176	91	869

Bohužel se nám nepodařilo přesvědčit Statu, aby nám vytvořila smysluplnou kontingenční tabulku, v níž by byly obsaženy všechny čtyři proměnné. Pokusili jsme se o ni tedy sami.