UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

B A K A L Á Ř S K Á P R Á C E

Analýza kontingenčních tabulek

Vedoucí bakalářské práce:
Mgr. Ondřej Vencálek, Ph.D.
Rok odevzdání: 2013

Vypracovala:
Julie Rendlová
ME, III. ročník

Prohlášení:

Prohlašuji, že jsem bakalářskou práci zpracovala samostatně pod vedením pana Mgr. Ondřeje Vencálka, Ph.D. a že jsem v seznamu literatury uvedla všechny zdroje použité při jejím zpracování.

V Olomouci dne 26. 4. 2013

Poděkování:

Ráda bych na tomto místě poděkovala vedoucímu mé bakalářské práce Mgr. Ondřeji Vencálkovi, Ph.D. za obětavou spolupráci i čas, který mi věnoval při konzultacích, a také za cenné rady a připomínky, které pomohly tuto práci dovést do zdárného konce. Poděkování patří i mé rodině a přátelům za podporu během celého studia.

OBSAH

1. ÚVOD

Jako téma své bakalářské práce jsem si zvolila analýzu kontingenčních tabulek. Protože je však toto téma velmi široké a jeho základy byly mimo jiné součástí předmětu KMA/PMS2 Pravděpodobnost a matematická statistika 2, chtěla bych se zaměřit především na aplikace jednodušších či známějších testů na atraktivních praktických příkladech a na uvedení náročnějších testů v souvislosti se zajímavými motivačními příklady v pozadí.

V první části této práce budou provedeny především testy s hypotézou o nezávislosti znaků na datech získaných z Československé filmové databáze (www.csfd.cz). Předpokládám, že hodnocení filmů na těchto stránkách nemusí být vždy objektivní, protože diváci mohou mít v dnešní době zahlcené americkými filmy tendence podhodnocovat například ruské nebo asijské filmy, stejně tak by mohly mít nižší hodnocení objektivně nezpůsobené nižší kvalitou starší filmy; animované filmy mohou být nadhodnocené. Tyto své předpoklady se pokusím ověřit a dokázat na základě použití statistických metod nastudovaných během mého bakalářského studia.

Druhá část práce bude věnována použití Cochran-Mantel-Haenszelovy statistiky pro trojrozměrné kontingenční tabulky v reálném příběhu o nesprávném rozhodnutí kanadského soudu v případu týkajícím se možné diskriminace žen při povyšování pracovníků.

Ve třetí kapitole navážu na kapitolu religiozity z loňské bakalářské práce Víc než mrtví: Analýza postoje veřejnosti k pacientům v permanentním vegetativním stavu, kterou vypracovala Bc. Jana Dvořáková. Budu testovat hypotézu, zda se dvě otázky v dotazníku ptají na totéž. V původní práci je toto provedeno za použití Spearmanova korelačního koeficientu, což bych chtěla doplnit nějakým dalším testem a vzájemně porovnat výsledky.

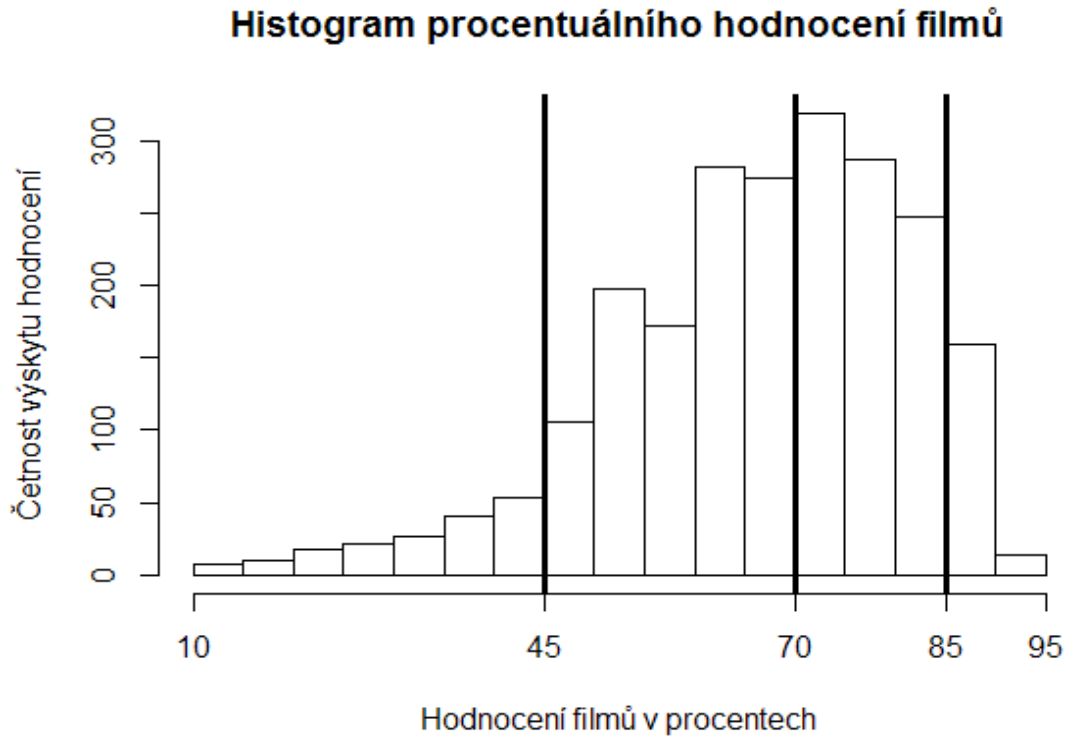
Použité metody a teoretické pozadí jednotlivých problémů uvedu vždy v rámci příslušných kapitol. Vzhledem k rozsahu dat jsem k výpočtům použila softwarové prostředí statistického programu R. Kompletní zdrojové kódy k jednotlivým příkladům budou uvedeny v přílohách bakalářské práce na disku CD.

2. NEZÁVISLOST HODNOCENÍ FILMŮ NA ROKU A ZEMI VÝROBY, DÉLCE FILMU A HRANÉ NEBO ANIMOVANÉ PODOBĚ FILMU

V dnešní se době se naprostá většina mladých lidí informuje o kvalitě filmů na stránkách Československé filmové databáze www.csfd.cz. Procentuální hodnocení filmů je zde jednoduše rozděleno do barevných pásem šedé (tj. špatné filmy), modré (tj. průměrně dobré filmy, které ničím nenadchnou) a červené (tj. výborné filmy). V rámci poslední kategorie jsou ještě filmy od 85 % hodnocení řazeny do žebříčku nejlepších filmů. Uživatelé se podle příslušného pásma rozhodují, zda na film stojí za to jít do kina, je to průměrná věc pro nenáročný večer strávený doma, nebo snímek ani nestojí za shánění. Návštěvníci stránek berou uvedené hodnocení jako objektivní a většinou jediný zdroj informací.

Problémem u objektivnosti hodnocení jednotlivých filmů může být fakt, že jde o průměr z hodnocení všech uživatelů. Vyvstává tedy otázka, zda je hodnocení nezávislé na dalších dostupných informacích o filmu. Některé filmy mohou být nedoceněné například kvůli veliké odlišnosti kultur mezi našimi zeměmi a zeměmi distribuce, mohou být pro nás zvláštní a nepochopené. Toto je bohužel domněnka, kterou nejsme schopni otestovat z dostupných dat, ale můžeme se zaměřit na možné závislosti mezi hodnocením a rokem výroby, hodnocením a zemí distribuce, hodnocením a délkou filmu, nebo také na závislost hodnocení na animované či hrané podobě filmu.

Data z ČSFD, která budou využita v této kapitole, se týkají 2236 filmů, které byly uvedeny v českých kinech v letech 1996 až 2011 (příloha A). Po vykreslení histogramu pro data ze sloupce obsahujícího procentuální hodnocení filmů (obrázek 2.1) je evidentní, že budeme muset upravit intervaly hodnocení, protože nemáme dostatečné zastoupení filmů pro každé barevné pásmo. Naše filmy se kvalitativně pohybují od 10 do 95 %, medián je roven 68 %. Bude tedy možné zachovat alespoň hranici 70 % pro výborné filmy a 85 % pro žebříčkové nejlepší, ale spodní hranici modrého pásma je třeba z 30 % navýšit alespoň na 45 %, abychom si zajistili dostatečné četnosti i v nejhorší kategorii.



Obrázek 2.1 – histogram procentuálního hodnocení filmů

2.1 Pearsonův test nezávislosti

Čerpáno z [1],[2]:

Tento test slouží k testování nulové hypotézy, že v dvojrozměrné kontingenční tabulce jsou na sobě veličiny X, Y nezávislé. Protože platí, že veličiny X a Y jsou nezávislé právě tehdy, když $p_{ij} = p_{i\bullet} p_{\bullet j}$ pro všechna (i, j), můžeme hypotézu o nezávislosti zapsat jako

$$H_0 : p_{ij} = p_{i\bullet} p_{\bullet j}, \; i = 1, ..., r, j = 1, ..., s. \tag{2.1}$$

Na souboru dat o rozsahu n sledujeme dva znaky X a Y, které jsou diskrétní povahy a nabývají konečně mnoha hodnot. Takovou situaci lze zapsat do kontingenční tabulky (tabulka 2.1), což je matice (n_{ij}), $i = 1, ..., r, j = 1, ..., s$, kde n_{ij} je počet případů, kdy se ve výběru vyskytla dvojice (i, j). V kontingenční tabulce můžeme psát

$n_{i\bullet} = \sum_{j=1}^{s} n_{ij}$, $n_{\bullet j} = \sum_{i=1}^{r} n_{ij}$, kde $n_{i\bullet}$ a $n_{\bullet j}$ jsou marginální četnosti.

Tabulka 2.1 – Dvourozměrná kontingenční tabulka				
Y / X	1	...	s	Σ
1	n_{11}	...	n_{1s}	$n_{1\bullet}$
...	...	...	...	...
r	n_{r1}	...	n_{rs}	$n_{r\bullet}$
Σ	$n_{\bullet 1}$	...	$n_{\bullet s}$	n

Jestliže je rozsah výběru n předem dán, tak má každá náhodná veličina n_{ij} binomické rozdělení s rozsahem výběru n jako parametrem, pravděpodobností p_{ij} z matice pravděpodobností (p_{ij}), $i = 1, ..., r$, $j = 1, ..., s$ a střední hodnotou $m_{ij} = np_{ij}$. Jelikož $n = \sum_{i=1}^{r} n_{i\bullet} = \sum_{j=1}^{s} n_{\bullet j} = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}$, tak četnosti n_{ij} nejsou vzájemně nezávislé a rozdělení v dvourozměrné kontingenční tabulce je multinomické s parametry n a p_{ij}, $i = 1, ..., r$, $j = 1, ..., s$. Odchylky od nezávislosti ve všech polích kontingenční tabulky lze potom shrnout Pearsonovou statistikou

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}, \tag{2.2}$$

která má za platnosti nulové hypotézy asymptoticky χ^2- rozdělení s $(rs - 1)$ stupni volnosti. Jestliže jsou proměnné nezávislé, lze m_{ij} odhadnout jako

$$\hat{m}_{ij} = n\,\hat{p}_{i\bullet}\,\hat{p}_{j\bullet} = n \cdot \frac{n_{i\bullet}}{n} \cdot \frac{n_{\bullet j}}{n},$$

přičemž dojde ke snížení stupňů volnosti statistiky (2.2) na $(r - 1)(s - 1)$ a statistiku lze přepsat do tvaru vhodného pro výpočty:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet} n_{\bullet j}}{n}} \tag{2.3}$$

Hypotézu (2.1) zamítáme ve prospěch alternativy, že mezi veličinami X, Y existuje nějaký stupeň závislosti, jestliže hodnota testové statistiky (2.3) vyjde v kritickém intervalu $\langle \chi^2_{(r-1)(s-1)}(1-\alpha);\infty)$, kde $\chi^2_{(r-1)(s-1)}(1-\alpha)$ značí $(1-\alpha)\cdot 100\,\%$ kvantil χ^2- rozdělení o *(r – 1)(s – 1)* stupních volnosti.

Nesmíme ještě zapomenout na podmínku shody s limitním rozdělením, která vyžaduje, aby všechny očekávané četnosti $\frac{n_{i\bullet}n_{\bullet j}}{n}$ byly větší než 5. *„Praktickými obtížemi dodržení této podmínky se v minulosti zabývala řada studií, jejichž výsledky vedly k mírnému změkčení, nicméně menších než 5 by mělo být maximálně 20 % z očekávaných četností (a každá v takovém případě musí být alespoň jednotková).“*[2]

V případě, že není podmínka dostatečně velkých očekávaných četností dodržena, lze využít některou z modifikací testových statistik pro testování hypotézy o nezávislosti. Bylo dokázáno, že tyto modifikace jsou jednoho typu a liší se pouze jedním argumentem λ:

$$2nI^{\lambda} = \frac{2}{\lambda\cdot(\lambda+1)}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}\left[\left(\frac{n_{ij}}{m_{ij}}\right)^{\lambda}-1\right], \text{ kde } -\infty<\lambda<\infty. \qquad (2.4)$$

Statistika (2.4) je označována jako *power divergence statistics* a pro $\lambda = 0$ nebo $\lambda = -1$ není definovaná. Její asymptotické rozdělení je stejné jako u Pearsonovy statistiky, kterou také obdržíme, pokud $\lambda = 1$. Vzhledem k různé citlivosti na očekávané četnosti je doporučováno volit $\lambda = 2/3$. Pokud tedy odhadujeme m_{ij}, budeme ve výpočtech používat statistiku ve tvaru

$$2nI^{2/3} = \frac{9}{5}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}\left[\left(\frac{n_{ij}n}{n_{i\bullet}n_{\bullet j}}\right)^{2/3}-1\right]. \qquad (2.5)$$

Statistika (2.5) má za platnosti nulové hypotézy χ^2- rozdělení o *(r – 1)(s – 1)* stupních volnosti.

2.2 Standardizovaná Pearsonova rezidua

Čerpáno z [3]:

Zamítnutí testu nezávislosti z předchozí kapitoly ukazuje na závislost mezi sledovanými veličinami, ale neposkytuje informace o její povaze nebo síle. V této kapitole se budeme zabývat tím, co může následovat po χ^2 testu.

Srovnávání pozorovaných a očekávaných četností buňku po buňce nám pomůže ukázat povahu závislosti u testovaných dat. Za platnosti nulové hypotézy o nezávislosti se větší rozdíly $(n_{ij} - \hat{m}_{ij})$ objevují v buňkách s většími teoretickými četnostmi m_{ij}. Směrodatná odchylka z četností n_{ij} a tím pádem i z $(n_{ij} - m_{ij})$ je $\sqrt{m_{ij}}$. Pro rozdíly $(n_{ij} - \hat{m}_{ij})$ bude odchylka menší, ale bude úměrná k $\sqrt{m_{ij}}$. Pearsonovo reziduum definované pro každou buňku kontingenční tabulky má tvar

$$e_{ij} = \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n} \right)}{\sqrt{\frac{n_{i\bullet} n_{\bullet j}}{n}}}. \tag{2.6}$$

Pokud sečteme druhé mocniny všech Pearsonových reziduí v kontingenční tabulce, získáme statistiku (2.3).

Abychom byli schopni Pearsonova rezidua správně vyhodnotit, je potřeba je ještě znormovat. Standardizovaná Personova rezidua ve tvaru

$$\frac{n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}}{\sqrt{\frac{n_{i\bullet} n_{\bullet j}}{n} \cdot \left(1 - \frac{n_{i\bullet}}{n}\right) \cdot \left(1 - \frac{n_{\bullet j}}{n}\right)}} \tag{2.7}$$

mají za platnosti nulové hypotézy o nezávislosti asymptoticky normované normální rozdělení. Pokud reziduum (2.7) v absolutní hodnotě přesáhne hodnotu 2 nebo 3 (97,5% kvantil normovaného normálního rozdělení má hodnotu 1,96), můžeme již usuzovat na zamítnutí nulové hypotézy v buňce *(i, j)*.

2.3 Testy na datech z ČSFD

2.3.1 Test nezávislosti hodnocení filmu na zemi distribuce

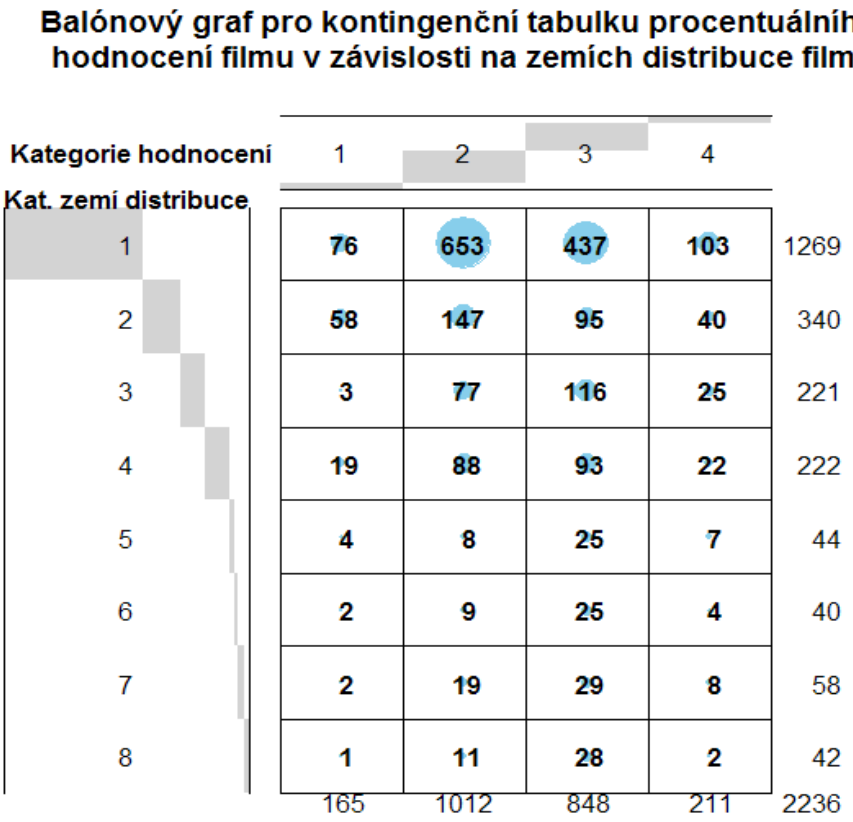
Filmy z našich dat pocházejí ze 47 různých zemí, přičemž některé země mají velmi nízké filmové zastoupení, takže bude nutné je slučovat, abychom se vyhnuli nulovým hodnotám n_{ij} v kontingenční tabulce. Úpravu je asi nejvhodnější provést ze zeměpisného hlediska, případně podobného filmařského rukopisu některých zemí. Potom můžeme vytvořit kontingenční tabulku, kde veličina X jsou kategorizované země distribuce a Y kategorie procentuálního hodnocení. Tzv. balónový graf kromě kontingenční tabulky včetně marginálních četností také graficky znázorňuje zastoupení jednotlivých kategorií i četností – velikost modrých „balónků“ je přímo úměrná hodnotě n_{ij}/n (obrázek 2.2).

Statistika (2.3) vyjde 153,359, což je hodnota, pro kterou bychom hypotézu o nezávislosti zamítali (95% kvantil χ^2 - rozdělení o 21 stupních volnosti je roven 32,671), ale je ještě potřeba ověřit splnění podmínky dostatečně velkých očekávaných četností. Z tabulky 2.2 vidíme, že teoretické četnosti jsou menší než 5 v 7 případech z 32, což je více než 20 % a změkčení tedy nelze využít.

Po dalším slučování v kategorii zemí distribuce (sloučení skupin 5 a 6 – „Rusko“ a „Asie“, sloučení skupin 7 a 8 – „Sever“ a „Ostatní“) jsou již očekávané četnosti v pořádku. Statistika (2.3) vyjde 147,684 a hypotéza o nezávislosti hodnocení na zemích distribuce filmů se pro $\chi^2_{15}(0{,}95) = 24{,}996$ zamítá.

Tabulka 2.2 – očekávané četnosti

Země / Hodnocení	1	2	3	4	5	6	7	8
1	93,64	25,09	16,31	16,38	**3,25**	**2,95**	**4,28**	**3,1**
2	574,34	153,88	100,02	100,48	19,91	18,1	26,25	19,01
3	481,27	128,94	83,81	84,19	16,69	15,17	22	15,93
4	119,75	32,08	20,85	20,95	**4,15**	**3,77**	5,47	**3,96**

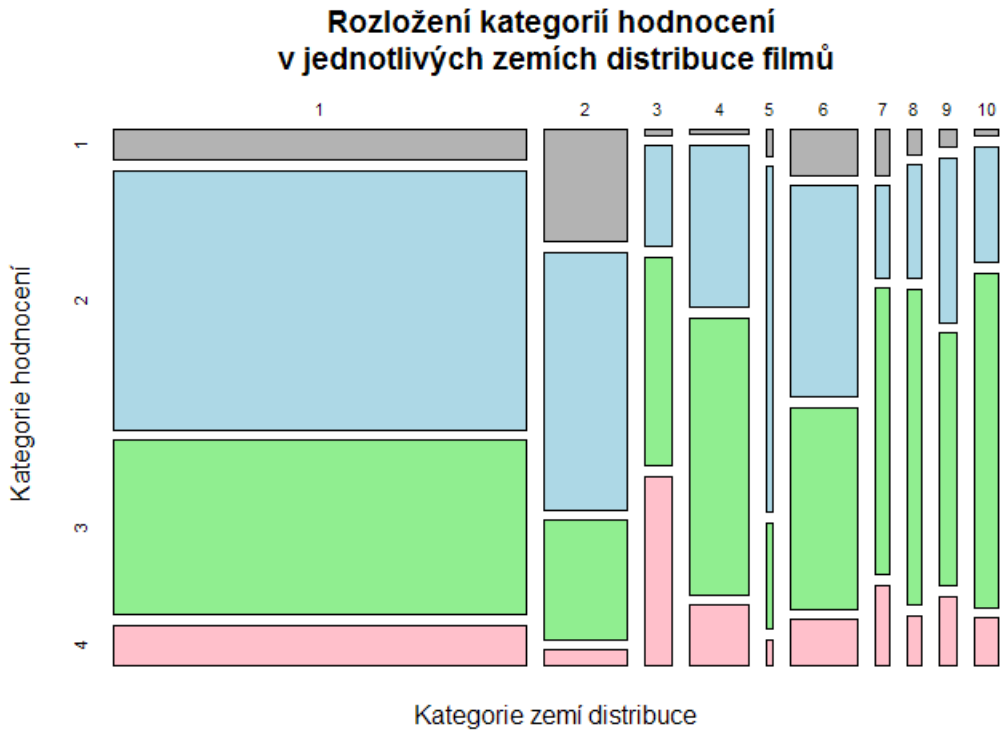


Obrázek 2.2 – balónový graf pro kontingenční tabulku procentuálního hodnocení filmu v závislosti na zemích distribuce filmu, kde kategorie hodnocení **1** je 10 – 44 %, **2** je 45 – 69 %, **3** je 70 – 84 % a **4** je 85 – 95 %; kategorie zemí distribuce filmu je **1** pro USA, **2** pro Českou republiku, Slovensko a Československo, **3** pro Velkou Británii, Austrálii, Kanadu a Nový Zéland, **4** pro Francii, Itálii, Německo, Rakousko, Španělsko a Západní Německo, **5** pro Rusko, Sovětský svaz, Východní Německo, Jugoslávii, Bosnu a Hercegovinu, Srbsko a Černou Horu a Slovinsko, **6** pro Japonsko, Čínu, Jižní Koreu, Izrael, Írán, Hong Kong, Indii, Bhútán, Kazachstán, Palestinu, Taiwan, Thajsko a Vietnam, **7** pro Polsko, Dánsko, Švédsko, Finsko, Norsko a Estonsko a **8** pro ostatní.

Nyní se nabízí otázka, zda výsledek testu nemohlo výrazně ovlivnit provedené slučování zemí distribuce. Test nezávislosti je tedy vhodné provést ještě v každé z šesti skupin zemí. U některých skupin sice není možné dodržet požadavek na dostatečně velké očekávané četnosti, ale získáme těmito testy alespoň představu o tom, které země jsme mohli sloučit naprosto nevhodně. Když je toto hotovo, získáváme vhodnější uskupení zemí distribuce rozdělených nyní do deseti skupin a na nově uspořádaných datech můžeme opět

provést test nezávislosti (2.3) s výsledkem rovným 309,433, pro který musíme hypotézu o nezávislosti zamítnout ($\chi^2_{27}(0{,}95) = 40{,}113$). Podmínka shody s limitním rozdělením je splněna z 82,5 %, což je sice dostačující, pokud vezmeme v úvahu změkčení podmínky, ale stejně je vhodné ověřit výsledek pomocí jiné statistiky méně citlivé právě na požadavek dostatečně velkých očekávaných četností, jako je například statistika (2.5). Její hodnota vyjde nižší, $2nI^{2/3} = 289{,}69$, ale stále dostatečně vysoká pro zamítnutí hypotézy o nezávislosti.

V tuto chvíli bychom se měli zaměřit na provedení takového testu, který by ukázal, v kterých polích kontingenční tabulky došlo k největším odchylkám od nezávislosti. Z obrázku 2.3, jenž znázorňuje tzv. mozaikovým grafem rozložení filmů do jednotlivých kategorií hodnocení pro každou zemi (případně skupinu zemí), se můžeme domnívat, že v zemích ze skupin 3, 4, 7, 8, 9, 10 se točí lepší filmy než jinde. Tuto domněnku je vhodné podpořit výpočtem standardizovaných Pearsonových reziduí (2.7), jehož výsledky jsou uvedeny v tabulce 2.3. Barevně označená rezidua svědčí o závislosti v daných buňkách. Pokud reziduum v buňce *(i, j)* dosáhlo hodnoty vyšší než 2, je označeno zeleně a znamená, že skutečná hodnota n_{ij} je významně větší, než bychom očekávali. Pokud dosáhlo hodnoty nižší než -2, je označeno červeně a ukazuje na situaci, kdy je hodnota n_{ij} významně nižší, než odpovídá očekávání. Pro USA, ČR a SR můžeme výsledky interpretovat tak, že z těchto zemí pochází méně výborných a nejlepších filmů, než lze očekávat, což je vykompenzováno nečekaně velkým množstvím filmů v kategorii průměrně dobrých či špatných filmů. Naopak ČSR a VB má mnohem větší zastoupení v kategorii nejlepších či výborných filmů a mnohem menší v kategoriích špatných a průměrně dobrých filmů, než by odpovídalo očekávání plynoucímu z dat v naší kontingenční tabulce. Výsledky bychom mohli mylně interpretovat tak, že se například ve Velké Británii natáčí více kvalitních filmů než v České republice. Musíme si ale uvědomit, že zde dochází k určitému „výběrovému zkreslení“ – naše data obsahují pouze filmy, které se dostaly do českých kin, což zdaleka nebudou všechny filmy z produkce VB, pouze ty lepší. Zajímavé, i když poměrně očekávané je, že podobné tvrzení neplatí o filmech z USA. Americká produkce nás zahrnuje sice velikým množstvím filmů, ale jde o snímky daleko průměrnější.



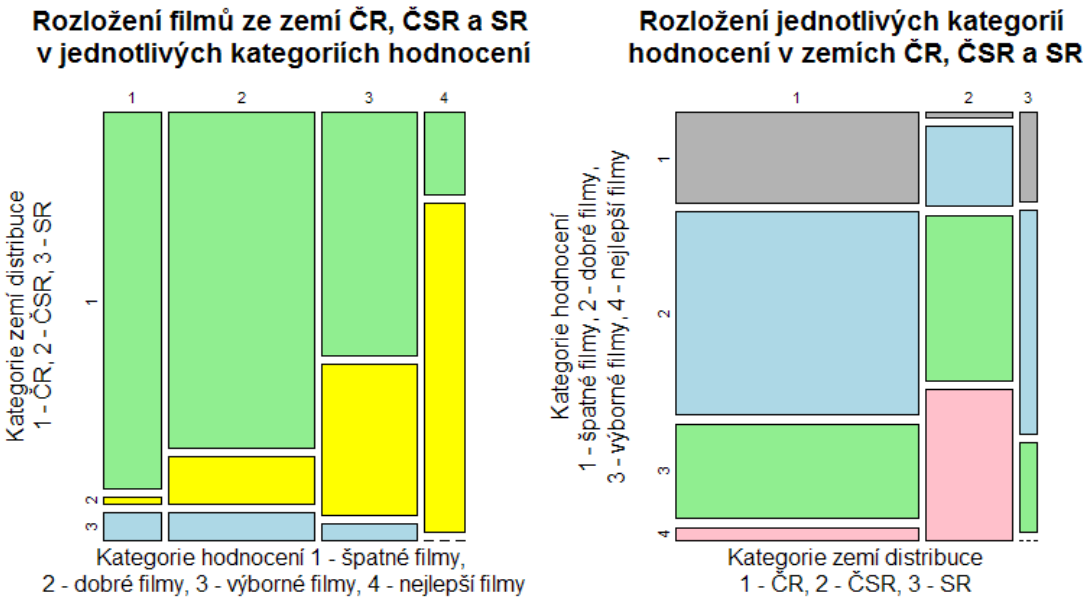
Obrázek 2.3 – mozaikový graf pro kontingenční tabulku procentuálního hodnocení filmu v závislosti na zemích distribuce filmu, kde kategorie hodnocení **1** je 10 – 44 %, **2** je 45 – 69 %, **3** je 70 – 84 % a **4** je 85 – 95 %; kategorie zemí distribuce filmu je **1** pro USA, **2** pro Českou republiku a Slovensko, **3** pro Československo, **4** pro Velkou Británii, **5** pro Kanadu, **6** pro Francii, Itálii, Německo a Španělsko, **7** pro Rusko, Sovětský svaz, Východní Německo, Jugoslávii, Bosnu a Hercegovinu, Srbsko a Černou Horu a Slovinsko, **8** pro Japonsko, Čínu, Jižní Koreu, Izrael, Írán, Hong Kong, Indii, Bhútán, Kazachstán, Palestinu, Taiwan, Thajsko a Vietnam, **9** pro Polsko, Dánsko, Švédsko, Finsko, Norsko a Estonsko a **10** pro ostatní.

Tabulka 2.3 – standardizovaná Pearsonova rezidua

Země / Hodnocení	1	2	3	4	5	6	7	8	9	10
1	-2,88	9,72	-2,23	-3,41	-0,35	1,02	0,44	-0,58	-1,16	-2,02
2	6,75	1,95	-4,77	-3,75	2,04	-1,04	-3,64	-2,92	-1,94	-3,92
3	-3,89	-5,03	0,63	4,95	-1,52	0,62	2,61	3,23	1,92	5,1
4	-2,45	-3,66	9,07	1,22	-0,62	-0,16	1,48	0,12	1,15	0,01

Pokud shrneme informace, které nám standardizovaná Pearsonova rezidua dala, lze říci, že filmy ze skupin 1 a 2 jsou spíše špatné a průměrně dobré, skupiny 3, 4, 7, 8 a 10 jsou zastoupeny spíše v kategoriích výborných a nejlepších filmů, zatímco filmy ze skupin 5, 6 a 9 víceméně odpovídají očekávání.

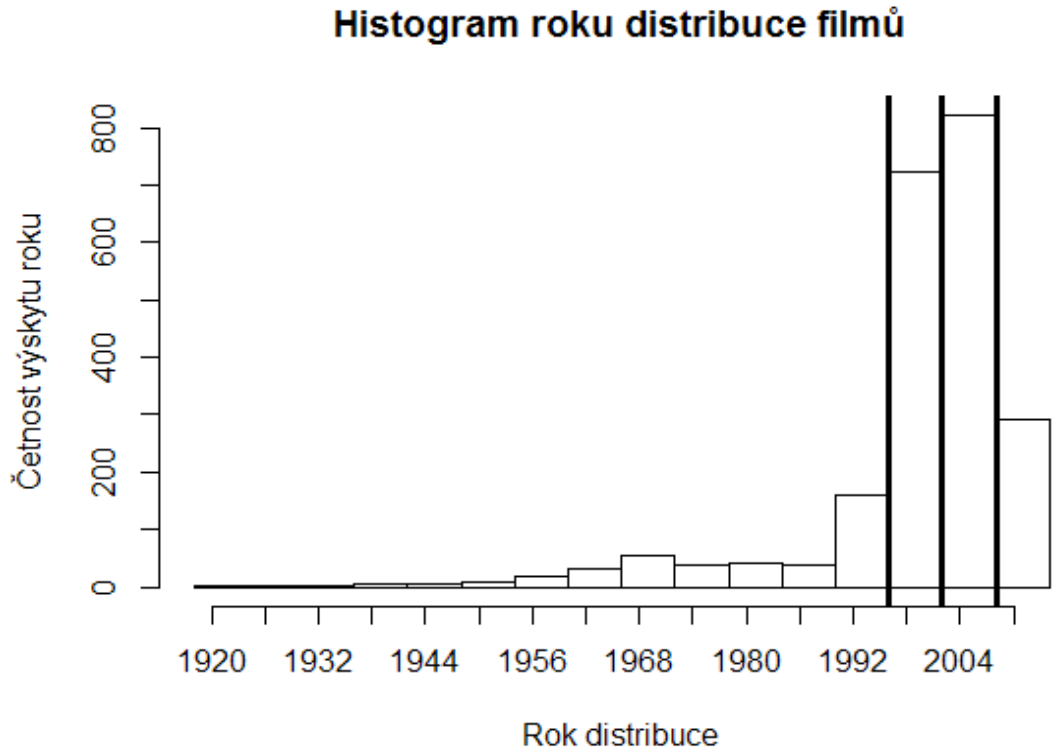
Vraťme se ještě k momentu, kdy jsme prováděli testy nezávislosti v jednotlivých sloučených skupinách zemí. U skupiny zemí 2, tj. ČR, SR a ČSR jsme zjistili, že je sice nelze sloučit dohromady všechny, ale pouze pro Českou republiku a Slovensko se závislost neprokázala (výpočty v příloze B). Odlišnost filmové tvorby Československa můžeme pozorovat na obrázku 2.4, který zachycuje dva mozaikové grafy. Z levého grafu vidíme, že skupina špatných filmů je tvořena především českou tvorbou, zatímco nejlepší filmy pocházejí hlavně z ČSR. Z pravého grafu pak můžeme získat domněnku, že tvorba ČSR byla mnohem lepší, než je tvorba ČR a SR, což potvrdil i výpočet standardizovaných Pearsonových reziduí (sloupce 2 a 3 v tabulce 2.3) – rozložení filmů v jednotlivých kategoriích hodnocení je pro Českou republiku se Slovenskem téměř opačné než pro ČSR. Vzhledem k tomu, že se ale Československo nacházelo na stejném území, jako nyní ČR a SR, země tedy mají shodné i některé režiséry, scénáristy a herce, vyvstává další otázka, kterou lze z našich dat otestovat, a sice jestli není hodnocení filmů závislé na roku výroby.



Obrázek 2.4 – mozaikové grafy pro kontingenční tabulku procentuálního hodnocení filmů ze zemí ČR, SR a ČSR.

2.3.2 Test nezávislosti hodnocení filmu na roku výroby

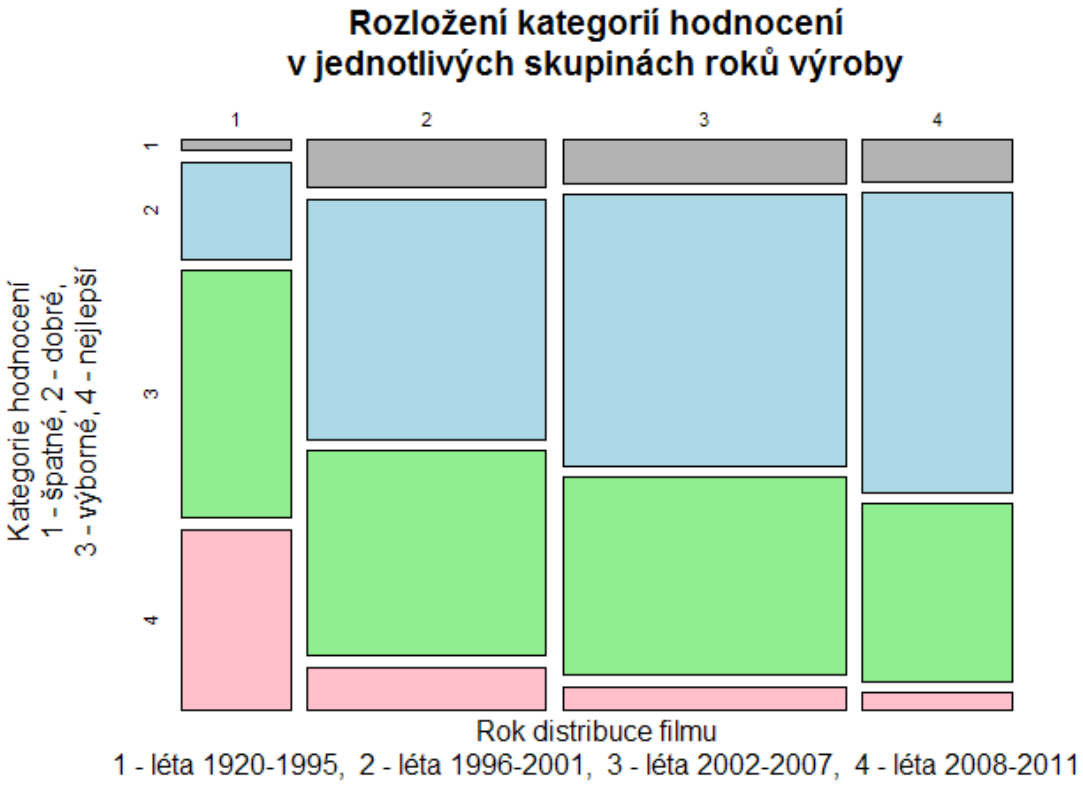
Filmy z našeho souboru dat byly natočeny v letech 1920 až 2011. Rozložení filmů v různých letech můžeme vidět na obrázku 2.5, kde jsou do histogramu zakresleny i naše dělící hranice víceméně odpovídající mediánu, dolnímu a hornímu kvartilu. Kontingenční tabulka bude mít rozměry 4 x 4, kde kategorie procentuálního hodnocení zůstávají stejné jako v předchozí kapitole a kategorie roku výroby jsou 1920 – 1995, 1996 – 2001, 2002 – 2007 a 2008 – 2011.



Obrázek 2.5 – histogram roku distribuce filmu

Pearsonova statistika pro test nezávislosti (2.3) vyjde 327,621, což je hodnota, pro kterou musíme hypotézu zamítnout ve prospěch alternativy, že zde nějaká závislost existuje ($\chi^2_9(0{,}95) = 16{,}919$). Již z obrázku 2.6, kde je znázorněn mozaikový graf pro rok distribuce filmu, můžeme usuzovat, že filmy z let 1920 až 1995 jsou mnohem častěji

hodnoceny jako výborné a nejlepší, než je tomu u novějších filmů. Výsledky standardizovaných Pearsonových reziduí (2.7) v tabulce 2.4 tuto domněnku potvrzují – starší filmy z našich dat jsou opravdu hodnoceny jako výborné a nejlepší častěji, než bychom očekávali, naopak nejnovější filmy mají v lepších kategoriích oproti očekávání nižší zastoupení. Opět je zde důležité uvědomit si možnost „výběrového zkreslení“ – pokud se v kinech v letech 1996 až 2011 promítaly filmy z let předchozích, byly to jen ty lepší snímky.



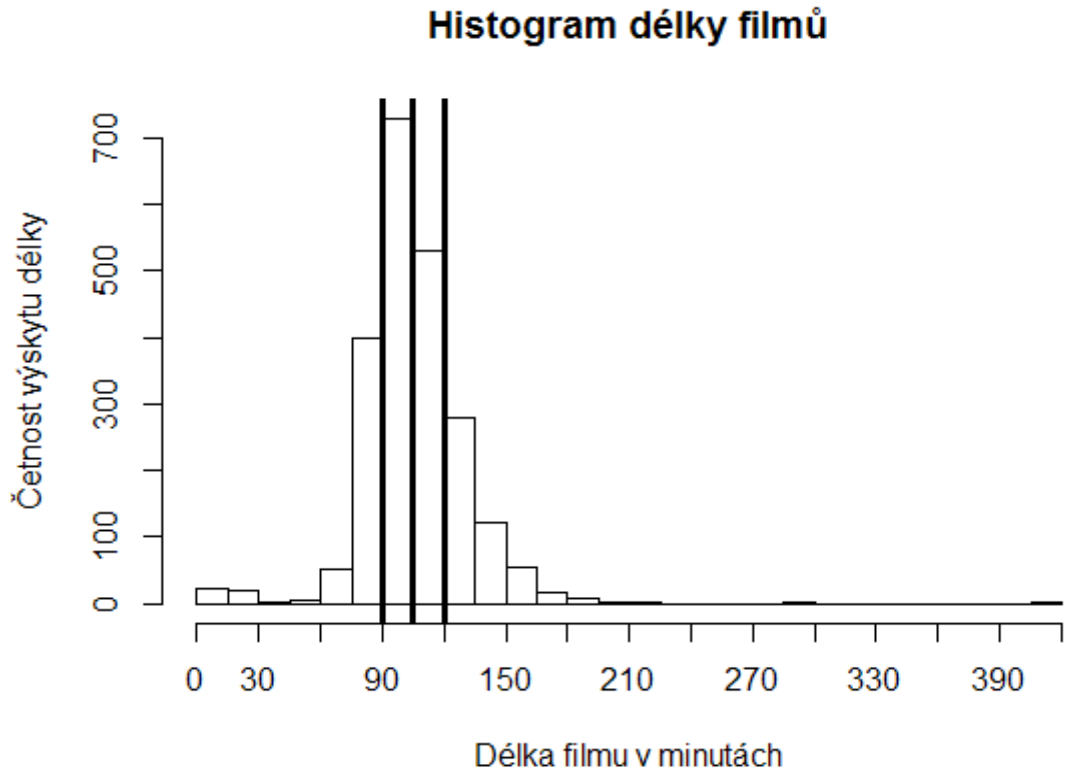
Obrázek 2.6 – mozaikový graf pro kontingenční tabulku roku výroby a procentuálního hodnocení filmů

Tabulka 2.4 – standardizovaná Pearsonova rezidua				
Rok Hodnocení	1	2	3	4
1	-4	1,88	0,85	0,29
2	-10,41	-0,34	3,84	4,89
3	3,25	0,22	-0,72	-2,25
4	15,9	-1,47	-6,11	-4,85

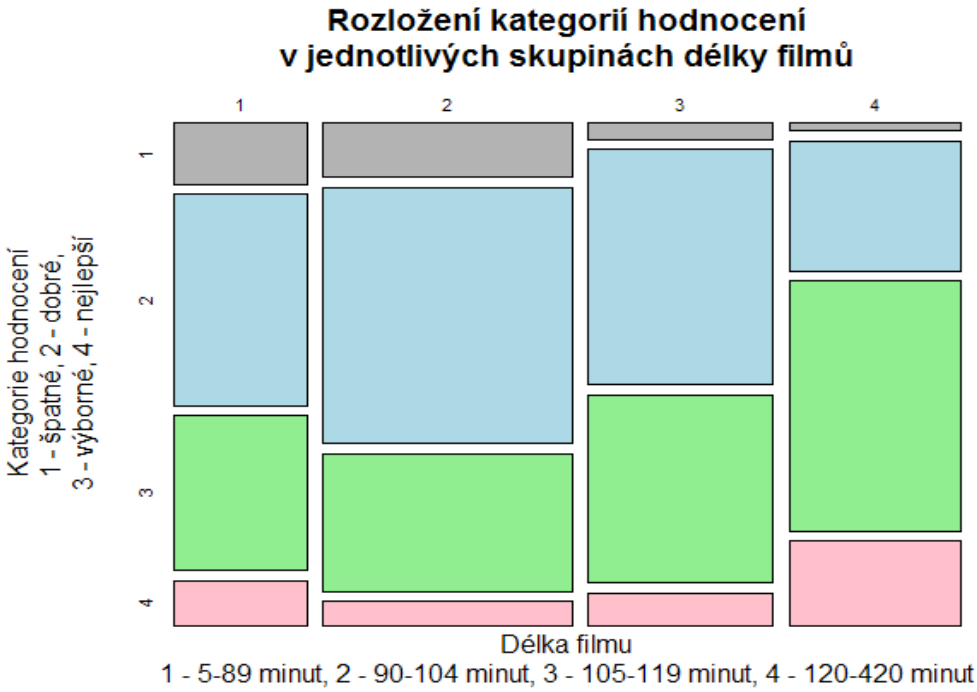
2.3.3 Test nezávislosti hodnocení filmu na jeho délce

U testování nezávislosti hodnocení filmů na jejich délce by bylo zajímavé mít jednu z kategorií délky rezervovanou pro krátkometrážní filmy, ale jak je vidět z obrázku 2.7, máme v této oblasti k dispozici jen velmi málo dat. Délky filmů budou proto rozděleny do skupin 5 až 89 minut, 90 až 104 minut, 105 až 119 minut a 120 až 420 minut. Dělící hranice odpovídají zaokrouhleným hodnotám mediánu a dolního a horního kvartilu.

Test nezávislosti musíme výsledkem Pearsonovy statistiky (2.3) 228,177 opět zamítnout ($\chi_9^2(0{,}95) = 16{,}919$). I tentokrát je z mozaikového grafu (obrázek 2.8) patrné, že dlouhé filmy budou hodnoceny lépe, než bychom očekávali a filmy z 2. kategorie budou hodnoceny oproti očekávání naopak hůře. Domněnku potvrzují i standardizovaná Pearsonova rezidua (2.7) v tabulce 2.5.



Obrázek 2.7 – histogram délky filmu



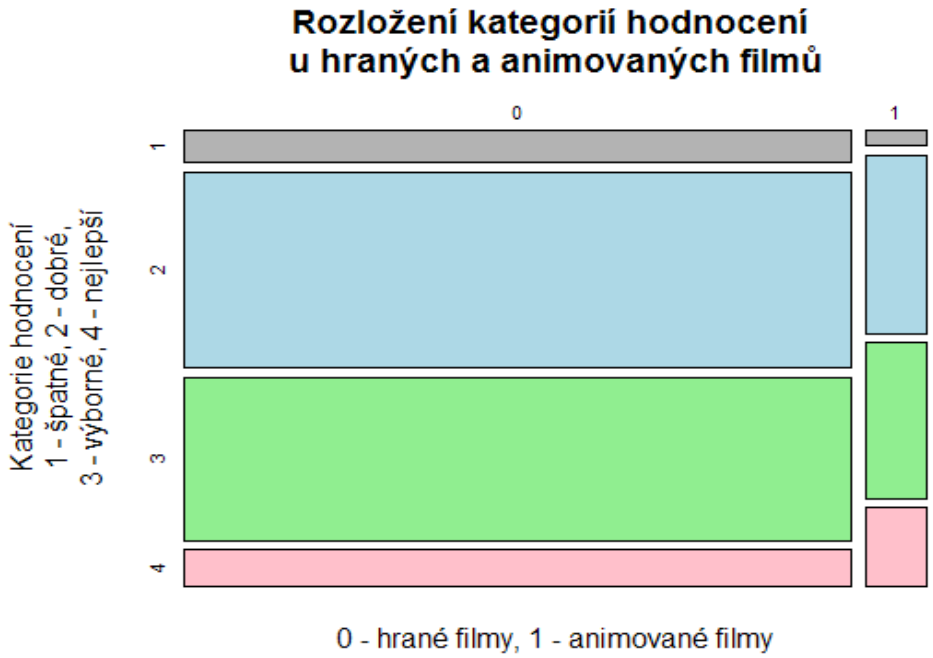
Obrázek 2.8 – mozaikový graf pro kontingenční tabulku délky filmů a jejich procentuálního hodnocení

Tabulka 2.5 – standardizovaná Pearsonova rezidua				
Délka Hodnocení	1	2	3	4
1	4,73	5,15	-4,17	-5,8
2	-0,28	6,14	2,51	-9,21
3	-2,39	-6,18	1,07	8,01
4	0,22	-4,81	-2,31	7,56

2.3.4 Test nezávislosti hodnocení filmu na jeho hrané či animované podobě

Poslední test nezávislosti, který bude na datech z ČSFD proveden, se týká vztahu hodnocení filmu k jeho hrané či animované podobě. V datech se vyskytuje 185 filmů animovaných a 2051 hraných. Pearsonova statistika (2.3) je rovna 21,3676. Protože $\chi^2_3(0{,}95) = 7{,}815$, musíme test nezávislosti na hladině $\alpha = 0{,}05$ zamítnout.

Již z obrázku 2.9 se můžeme domnívat, že animované filmy jsou oproti hraným nadhodnocené a v tabulce 2.6 standardizovaných Pearsonových reziduí (2.7) vidíme, že animované filmy mají v nejlepší kategorii opravdu větší zastoupení, než by z dat bylo možné očekávat, zatímco hrané filmy zde mají oproti očekávání naopak nižší skutečnou četnost.



Obrázek 2.9 – mozaikový graf pro kontingenční tabulku hrané či animované podoby filmů a jejich procentuálního hodnocení

Tabulka 2.6 – standardizovaná Pearsonova rezidua		
Hodnocení	Hrané	Animované
1	1,95	-1,95
2	1,04	-1,04
3	0,5	-0,5
4	-4,34	4,34

Zajímavé by určitě bylo sledovat ještě závislost hodnocení na žánru filmu, jenže filmy jsou na ČSFD většinou zařazené do několika kategorií současně (např. drama / romantický / historický / životopisný), přičemž tyto kategorie nejsou disjunktní a nevíme, jak významnou roli u nich hraje pořadí (např. zda je thriller / drama to stejné, jako drama / thriller). K testování takovéto závislosti je proto zapotřebí použití metod založených na logaritmicko-lineárních modelech, což překračuje rozsah této práce.

3. JAK SOUD PŘIJAL NEMOŽNÉ VYSVĚTLENÍ

Čerpáno z [4]:

Ačkoliv je často pravda, že závislost mezi dvěma proměnnými může být způsobena nějakou nepozorovanou třetí proměnnou, přijetí takových argumentů vyžaduje vždy pečlivé prozkoumání. V roce 1987 přijal kanadský odvolací soud vysvětlení nabídnuté National Revenue Service of Canada (tj. federální agentura mající na starosti především daně a poplatky), že při testech na pracovní povýšení bylo nižší bodové ohodnocení u žen než u mužů vysvětleno rozdíly ve vzdělání – proměnnou, která nebyla původním soudem pozorována. My se pokusíme ukázat, že v tomto konkrétním případě je správnost nabídnutého vysvětlení vyloučena.

3.1 Proměnná, která nikdy neexistovala

Závislost mezi dvěma proměnnými může být nepravá, protože může být způsobena nějakou nepozorovanou třetí proměnnou. Povykládáme si příběh o soudu, který přijal argument zahrnující právě nepozorovanou proměnnou, která, jak ukážeme, byla špatně interpretována. Příběh je řečený v této kapitole a formální demonstrace je uvedena v kapitole 3.2.

V případu Maloley proti National Revenue Service of Canada, National Revenue Service povyšovala zaměstnance na pozice úředníků – vymahačů inkasa – na základě použití psychologického testu nazvaného Všeobecný test inteligence. Dle výsledků tohoto testu uspělo 59 % mužů a 27 % žen (viz tabulka 3.1).

Tabulka 3.1 – frekvence úspěšnosti

	Uspěl(a)	Neuspěl(a)	Celkem	Procento úspěšnosti
Ženy	68	183	251	27 %
Muži	68	47	115	59%
Celkem	136	230	366	37%

Podle kanadského zákona National Revenue Service musela prokázat, že test byl spolehlivým a právoplatným způsobem pro selekci kandidátů podle jejich hodnot a schopností. Když byla National Revenue Service vyzvána u soudu, obhajovala použití testu jednoduše odvoláním se na tabulku 3.2, která ukázala, že 52 % mužů a 25 % žen mělo nějaké vysokoškolské vzdělání.

Tabulka 3.2 – frekvence vysokoškolské vzdělanosti				
	Má VŠ	Nemá VŠ	Celkem	Procento vysokoškolské vzdělanosti
Ženy	63	188	251	25 %
Muži	60	55	115	52 %
Celkem	123	243	366	34 %

Odvolací soud uznal tvrzení National Revenue Service of Canada, že rozdíl mezi mužskou a ženskou úspěšností nebyl diskriminací, protože pouze odrážel rozdíl v kognitivních schopnostech obou pohlaví, jak bylo také evidentní z dat o vysokoškolské vzdělanosti. Odvolací soud došel k závěru, že frekvence uspění v testu byla jednoduše v přímém poměru k frekvenci vysokoškolské vzdělanosti, ačkoliv National Revenue Service data o úspěšnosti v poměru pohlaví / vzdělání původně ani nezařadila mezi důkazy.

Chceme ukázat, že rozsudek soudu stanovený pouze na základě těchto důkazů byl chybný. Rozdíl v úspěšnosti mužů a žen u testu sice mohl být právě tak velký, že pouze odrážel rozdíl ve vysokoškolské vzdělanosti, ale také nemusel.

Data jsou nezvyklá. Tabulka 3.1 a tabulka 3.2 jsou dvě marginální tabulky typu 2 x 2 z kontingenční tabulky typu 2 x 2 x 2, ale odvolací soud nepožádal o kompletní 2 x 2 x 2 kontingenční tabulku a nikdy ji neviděl. Argument o lineární závislosti v těchto tabulkách, který soud přijal, není z hlediska statistiky pouze nevěrohodný; je nemožný. Tento příklad je důležitý, protože ukazuje, že některá tvrzení o nesledovaných proměnných nejsou jen vynucená na základě nějakých domýšlených souvislostí (odvolací soud automaticky předpokládal, že ti jedinci, kteří měli vysokoškolské vzdělání, zároveň uspěli u testu), některá jsou prostě špatná.

Kdyby byla respektována a sledována kontingenční tabulka 2 x 2 x 2, standardní analýza by srovnávala úspěšnost u testu pro ženy a muže očištěnou od vzdělání použitím Cochran-Mantel-Haenszelovy statistiky. Musíme vzít v úvahu všechny možné tabulky typu 2 x 2 x 2, které mohou produkovat tabulku 3.1 a tabulku 3.2 – těch je konečně mnoho – a spočítat Cochran-Mantel-Haenszelovu normovanou odchylku (tj. kladná odmocnina z CMH statistiky, viz kapitola 3.2.1) pro každou z nich. Jednodušší výpočet přinášející stejný výsledek je uveden níže v kapitole 3.2. Když je toto hotovo, vidíme, že nejmenší normovaná odchylka, která může být produkována z tabulky 2 x 2 x 2 kompatibilní s tabulkou 3.1 a tabulkou 3.2, je rovna 3,11 s p-value 0,0018 (tj. pravděpodobnost, že by za platnosti nulové hypotézy vyšlo testové kritérium stejně nebo pro nulovou hypotézu ještě nepříznivěji; je-li p-value menší než hladina testu α, zamítáme nulovou hypotézu). Tudíž neexistuje kontingenční tabulka typu 2 x 2 x 2 kompatibilní s tabulkou 3.1 a tabulkou 3.2, ve které může vysokoškolská vzdělanost objasnit rozdíly v úspěšnosti u testu mezi muži a ženami. Může a nemusí zde být diskriminace žen, ale ať už tu je nebo není, vysvětlení přijaté odvolacím soudem je jednoduše špatné. Rozdíly v úspěšnosti u testu jsou mnohem větší, než může být objasněno na základě rozdílu v poměru vysokoškolské vzdělanosti.

3.2 Cochran-Mantel-Haenszelova statistika pro kontingenční tabulku 2 x 2 x 2 s danými marginálními tabulkami

Vezměme v úvahu tabulku 2 x 2 x 2 s četnostmi n_{egp} znamenající vzdělání (e) x rod (g) x úspěšnost (p), kde $e = 1$ pro vysokoškolské vzdělání a $e = 2$ jinak, $g = 1$ pro ženy a $g = 2$ pro muže, $p = 1$ pro uspěl(a) a $p = 2$ pro neuspěl(a). Pišme m pro počet jednotlivců, kteří u testu uspěli a navštěvovali vysokou školu, tj. $m = n_{1\bullet 1}$, takže vidíme, že m nemůže být určeno z tabulek 3.1 a 3.2. Kontingenční tabulka 2 x 2 x 2 má ještě jednu další marginální tabulku, a sice tabulku 3.3.

Tabulka 3.3 – Nepozorovaná marginální tabulka 2x2			
	Uspěl(a)	Neuspěl(a)	Celkem
Má VŠ	m	$n_{1\bullet\bullet} - m$	$n_{1\bullet\bullet} = 123$
Nemá VŠ	$n_{\bullet\bullet 1} - m$	$n_{\bullet\bullet\bullet} - n_{\bullet\bullet 1} - n_{1\bullet\bullet} + m$	$n_{2\bullet\bullet} = 243$
Celkem	$n_{\bullet\bullet 1} = 136$	$n_{\bullet\bullet\bullet} - n_{\bullet\bullet 1} = 230$	$n_{\bullet\bullet\bullet} = 366$

3.2.1 Cochran-Mantel-Haenszelův test podmíněné nezávislosti

Čerpáno z [3]:

Mantel a Haenszel navrhli roku 1959 test s hypotézou H_0: podmíněná nezávislost v kontingenčních tabulkách typu 2 x 2 x K. Se zaměřením na retrospektivní studie onemocnění zacházeli se sloupcovými marginálními součty jako s fixními. Tudíž v každé parciální tabulce *k* s četnostmi $\{n_{ijk}\}$ jejich analýza vycházela z předpokladu pevně daných marginálních řádkových součtů $(n_{1\bullet k}, n_{2\bullet k})$ i marginálních sloupcových součtů $(n_{\bullet 1k}, n_{\bullet 2k})$. Rozdělení pro četnost v první buňce tabulky n_{11k} v každé parciální tabulce je hypergeometrické. Tyto četnosti určují $\{n_{12k}, n_{21k}, n_{22k}\}$ vzhledem k marginálním součtům.

Za platnosti H_0 mají střední hodnota a rozptyl tvar:

$$\mu_{11k} = E(n_{11k}) = n_{1\bullet k} n_{\bullet 1k} / n_{\bullet\bullet k} ;$$

$$\mathrm{var}(n_{11k}) = n_{1\bullet k} n_{2\bullet k} n_{\bullet 1k} n_{\bullet 2k} / \left[n_{\bullet\bullet k}^2 (n_{\bullet\bullet k} - 1) \right].$$

Četnosti z různých parciálních tabulek jsou nezávislé. Testová statistika kombinuje informaci srovnáváním $\Sigma_k n_{11k}$ s jeho očekávanou hodnotou za platnosti nulové hypotézy. Testová statistika má tvar:

$$\chi_{CMH}^2 = \frac{\left[\sum_k (n_{11k} - \mu_{11k}) \right]^2}{\sum_k \mathrm{var}(n_{11k})} . \qquad (3.1)$$

Tato statistika má za platnosti nulové hypotézy pro velké rozsahy přibližně χ^2 - rozdělení s jedním stupněm volnosti. Když je poměr šancí (viz kapitola 3.2.2) *OR(k) > 1* v parciální tabulce *k*, očekáváme $(n_{11k} - \mu_{11k}) > 0$. Když je *OR(k) > 1* nebo *OR(k) < 1* v každé parciální tabulce, $\Sigma_k (n_{11k} - \mu_{11k})$ má tendenci být v absolutní hodnotě relativně velký. Tento test funguje nejlépe, když je asociace mezi X a Y podobná ve všech parciálních tabulkách.

Cochran navrhl roku 1954 podobnou statistiku. Zacházel však s řádky jako s nezávislými binomickými vzorky ze dvou populací. Cochranova statistika je vztah (3.1) s $\mathrm{var}(n_{11k})$ nahrazeným $\mathrm{var}(n_{11k}) = n_{1\bullet k} n_{2\bullet k} n_{\bullet 1k} n_{\bullet 2k} / n_{\bullet\bullet k}^3$.

Kvůli podobnosti jejich přístupu, nazýváme vztah (3.1) Cochran-Mantel-Haenszelovou (CMH) statistikou a kladnou odmocninu z této statistiky

$$\chi_{CMH} = \left| \frac{\sum_k (n_{11k} - \mu_{11k})}{\sqrt{\sum_k \mathrm{var}(n_{11k})}} \right| \tag{3.2}$$

Cochran-Mantel-Haenszelovou normovanou odchylkou, která má za platnosti nulové hypotézy pro velké rozsahy přibližně normované normální rozdělení.

V našem případě kontingenční tabulky 2 x 2 x 2 můžeme zápis zjednodušit na parciální tabulky indexované *e* (vzdělání), kde *e = 1* pro vysokoškolsky vzdělané pracovníky a *e = 2* jinak (viz tabulka 3.4). Dále

$$\mu_{e21} = E(n_{e21}) = n_{e\bullet 1} n_{e2\bullet} / n_{e\bullet\bullet} \; ;$$

$$\mathrm{var}(n_{e21}) = n_{e\bullet 1} n_{e\bullet 2} n_{e2\bullet} n_{e1\bullet} / \left[n_{e\bullet\bullet}^2 (n_{e\bullet\bullet} - 1) \right]$$

a normovaná CMH odchylka bude tedy ve tvaru

$$\chi_{CMH} = \left| \frac{\sum_e \frac{n_{e21} n_{e12} - n_{e22} n_{e11}}{n_{e\bullet\bullet}}}{\sqrt{\sum_e \frac{n_{e\bullet 1} n_{e\bullet 2} n_{e2\bullet} n_{e1\bullet}}{n_{e\bullet\bullet}^2 (n_{e\bullet\bullet} - 1)}}} \right| . \tag{3.3}$$

Chceme testovat nulovou hypotézu, že při povyšování pracovníků nedocházelo v National Revenue Service of Canada k diskriminaci žen.

Tabulka 3.4 – parciální tabulka očištěná od vzdělání			
e	Uspěl(a)	Neuspěl(a)	
Muži	n_{e21}	n_{e22}	$n_{e2\bullet}$
Ženy	n_{e11}	n_{e12}	$n_{e1\bullet}$
	$n_{e\bullet 1}$	$n_{e\bullet 2}$	$n_{e\bullet\bullet}$

Po dosazení hodnot do tabulky 3.4 dostaneme dvě parciální tabulky (tabulka 3.5, tabulka 3.6) se dvěma neznámými n_{121} a *m;* viz výše $m = n_{1\bullet 1}$.

Tabulka 3.5 – parciální tabulka pro vysokoškolsky vzdělané osoby			
Má VŠ	Uspěl(a)	Neuspěl(a)	
Muži	n_{121}	60 – n_{121}	60
Ženy	m – n_{121}	63 – m + n_{121}	63
	m	123 – m	123

Tabulka 3.6 – parciální tabulka pro osoby bez vysokoškolského vzdělání			
Nemá VŠ	Uspěl(a)	Neuspěl(a)	
Muži	68 – n_{121}	n_{121} – 13	55
Ženy	68 – m + n_{121}	120 + m – n_{121}	188
	136 – m	107 + m	243

Ačkoliv *m* je neznámé (nezjistitelné z tabulek 3.1 a 3.2), je omezeno pozorovanými daty: $a \le m \le b$, kde

$$a = \max(0; n_{\bullet 11} - n_{21\bullet}) + \max(0; n_{\bullet 21} - n_{22\bullet}),$$

$$b = \min(n_{\bullet 11}; n_{11\bullet}) + \min(n_{\bullet 21}; n_{12\bullet}),$$

tedy $13 \le m \le 123$. Přímý přístup zkouší každé *m* v tomto rozsahu, dále pro každé *m* může být $13 \le n_{121} \le m$, tj. 6216 hodnot testové statistiky, výpočet je tudíž vhodné provést ve statistickém softwaru (příloha C) algoritmem pro nalezení minimální hodnoty testové statistiky χ_{CMH}:

```
zvolme proměnnou min{χ_CMH} dostatečně velkou
pro m od 13 do 123
      pro a od 13 do m
      {vypočtěme χ_CMH(a,m) dle vzorce (3.3)
       pokud χ_CMH(a,m) < min{χ_CMH}, potom změňme hodnotu
       min{χ_CMH} na χ_CMH(a,m)}
```

Výsledkem je $\min\{\chi_{CMH}\} = 3{,}11$ pro *m* = *116* a n_{121} = *32* s p-value přibližně 0,0018.

Jinými slovy, i když *m* není pozorováno, není zde žádná možná hodnota *m* taková, že by rozložení vysokoškolského vzdělání objasnilo rozdílnou úspěšnost mužů a žen v testu, H_0 se zamítá ve prospěch alternativy.

Vykreslíme-li si vektor *y* hodnot CMH normovaných odchylek do grafu, zjistíme, že se v něm prvky opakují (první hodnota jedenkrát, druhá dvakrát, ..., sto jedenáctá sto jedenáctkrát). Lze učinit závěr, že vnitřní cyklus (indexovaný přes *a*) je vlastně zbytečný, neboť bez ohledu na hodnoty *a* se do vektoru *y* ukládá vždy stále stejná hodnota až do změny indexu *m*.

Nezávislost CMH normované odchylky na hodnotě proměnné n_{121} lze dokázat úpravou čitatele ve vzorci (3.3):

$$\sum_e \frac{n_{e21}n_{e12} - n_{e22}n_{e11}}{n_{e\bullet\bullet}} = \frac{n_{121}(63 - m + n_{121}) - (60 - n_{121})(m - n_{121})}{123} +$$

$$+ \frac{(68 - n_{121})(120 + m - n_{121}) - (n_{121} - 13)(68 - m + n_{121})}{243} = m \cdot \left(\frac{55}{243} - \frac{60}{123} \right) + \frac{68 \cdot 133}{243}$$

a tedy
$$\chi_{CMH} = \frac{m \cdot \left(\frac{55}{243} - \frac{60}{123} \right) + \frac{68 \cdot 133}{243}}{\sqrt{\frac{63 \cdot 60 \cdot m \cdot (123 - m)}{123^2 \cdot 122} + \frac{188 \cdot 55 \cdot (136 - m) \cdot (107 + m)}{243^2 \cdot 242}}} . \qquad (3.4)$$

Tím se zjednoduší i algoritmus pro výpočet (příloha C) a data lze přehledně vykreslit do grafu (obrázek 3.1).

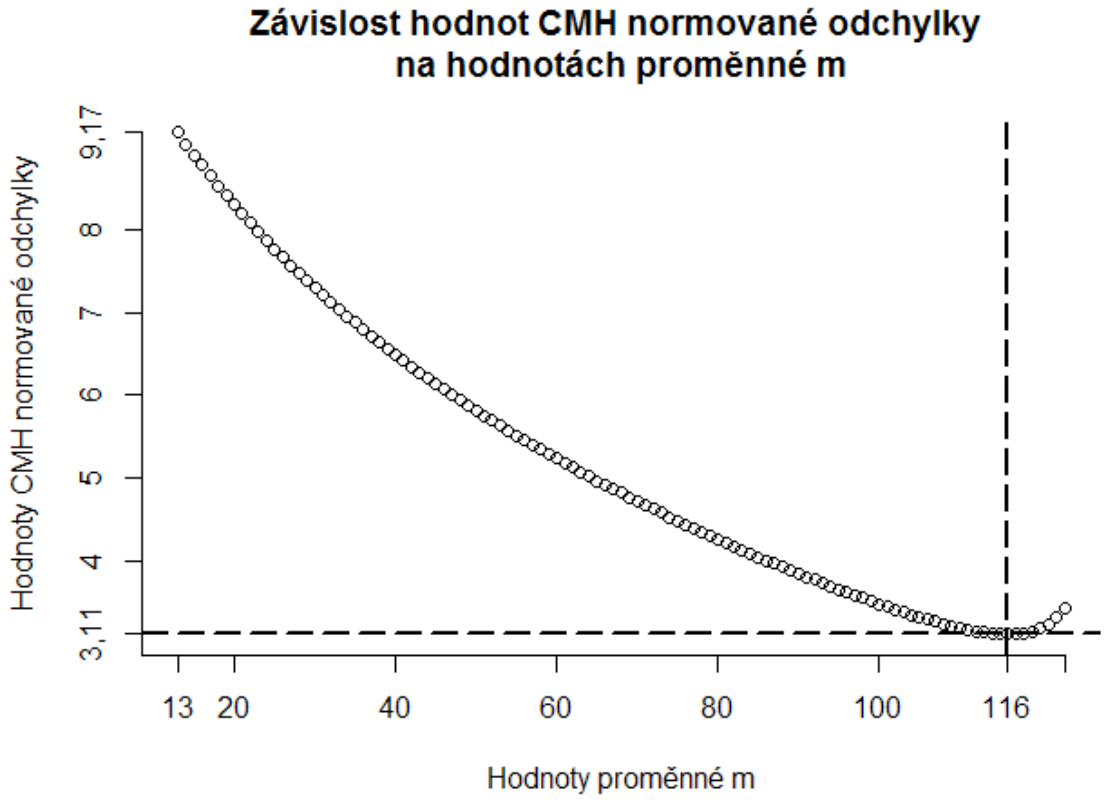
Ve skutečnosti není nutné zkoušet každou možnou hodnotu *m*. Přihlédnutím k tomu, že χ_{CMH} je funkce reálného argumentu, jejíž derivace χ'_{CMH} existuje v každém bodě jejího definičního oboru, může být dokázáno následující lemma:

Lemma. Pro $a < m < b$ je derivace χ'_{CMH} rovna nule pro nanejvýš jednu hodnotu *m*.

Důkaz. Označme $\chi_{CMH} = \frac{N(m)}{\sqrt{D(m)}}$ ve vztahu (3.3), takže

$$\chi'_{CMH} = \frac{N'(m)\sqrt{D(m)} - \frac{N(m)D'(m)}{2\sqrt{D(m)}}}{D(m)} . \qquad (3.5)$$

Pro $D(m) > 0$ lze rovnici (3.5) upravit jako $\chi'_{CMH}\sqrt{(D(m))^3} = N'(m)D(m) - \frac{1}{2}N(m)D'(m)$.



Obrázek 3.1 – závislost hodnot normované odchylky na hodnotách proměnné m.

Ze vztahu (3.3) můžeme pro určité konstanty *d*, *e*, *f*, *g* a *h*, které závisí na pozorovaných marginálních tabulkách ale ne na *m*, psát

$$N(m) = d + em, \text{ kde } d = \frac{68 \cdot 133}{243},\ e = \frac{55}{243} - \frac{60}{123};$$

$$D(m) = f + gm + hm^2, \text{ kde } f = \frac{188 \cdot 55 \cdot 136 \cdot 107}{243^2 \cdot 242},$$

$$g = \frac{63 \cdot 60}{123 \cdot 122} + \frac{188 \cdot 55 \cdot (136 - 107)}{243^2 \cdot 242},\ h = -\frac{63 \cdot 60}{123^2 \cdot 122} - \frac{188 \cdot 55}{243^2 \cdot 242};$$

takže $N'(m)D(m) - \frac{1}{2}N(m)D'(m) = e \cdot \left(f + gm + hm^2\right) - \frac{1}{2}(d + em) \cdot (g + 2hm) =$

$= \left(ef - \frac{dg}{2}\right) + m \cdot \left(\frac{eg}{2} - dh\right)$, což je výraz lineární v *m* dokazující lemma. □

Lemma znamená, že pro celé číslo *m* mezi *a* a *b* je odchylka χ_{CMH} buď monotónní v *m*, nebo je χ_{CMH} monotónní od *a* do celého čísla *r*, kde $a < r < b$, a potom monotónní od *r* do *b* v opačném směru. V každém případě je maximální hodnota χ_{CMH} pro $a < m < b$ rovna $\max\{\chi_{CMH}(a); \chi_{CMH}(r); \chi_{CMH}(b)\}$ a minimální hodnota χ_{CMH} je rovna $\min\{\chi_{CMH}(a); \chi_{CMH}(r); \chi_{CMH}(b)\}$. Připomeňme si ještě jednou, že *m* je počet jednotlivců, kteří uspěli u testu a navštěvovali vysokou školu, tj. $m = n_{1\bullet 1}$ (viz tabulka 3.3).

Pokud položíme derivaci χ'_{CMH} rovnu nule, vypočteme *r*:

$$\chi'_{CMH} = \frac{\left(ef - \frac{dg}{2}\right) + m\left(\frac{eg}{2} - dh\right)}{\sqrt{D(m)^3}} = 0, \text{ odtud } m = \frac{\frac{dg}{2} - ef}{\frac{eg}{2} - dh} = 116, \text{ a tedy}$$

$$\min\{\chi_{CMH}(a); \chi_{CMH}(r); \chi_{CMH}(b)\} = \min\{\chi_{CMH}(13); \chi_{CMH}(116); \chi_{CMH}(123)\} =$$

$$= \min\{9{,}17; 3{,}11; 3{,}44\} \text{ (výpočet viz příloha C).}$$

Pro data z kapitoly 3.1 nám tudíž *m* = 116 dává minimum Cochran-Mantel-Haenszelovy odchylky (3.4) rovno 3,11 s p-value 0,0018.

3.2.2 Poměr šancí

Čerpáno z [1]:

Poměr šancí (odds ratio) je ukazatel pro kontingenční tabulky typu 2 x 2, který na rozdíl od Pearsonovy nebo CMH statistiky vypovídá v případě zamítnutí hypotézy o nezávislosti i o směru a síle závislosti mezi veličinami. Teoretický poměr šancí můžeme zapsat jako

$$OR = \frac{p_{11}p_{22}}{p_{12}p_{21}}. \qquad (3.6)$$

Za platnosti nulové hypotézy o nezávislosti nabývá poměr šancí hodnoty 1, protože $p_{ij} = p_{i\bullet} p_{\bullet j}$. Čím více se hodnota ukazatele vzdaluje od jedné, tím větší je závislost mezi sledovanými veličinami. Při interpretaci výsledků je důležité si uvědomit, že hodnoty ukazatele (3.6) jsou kolem bodu jedna nesymetrické, protože $0 \leq OR \leq \infty$.

Jestliže pracujeme s empirickými daty, můžeme (3.6) odhadnout jako

$$\hat{OR} = \frac{n_{11} n_{22}}{n_{12} n_{21}}, \text{ protože } \hat{p}_{ij} = n_{ij} / n. \tag{3.7}$$

Tabulka 3.7 ukazuje, jak by vypadala tabulka 3.3 po dosazení hodnoty m = 116. Poměr šancí (3.7) pro tabulku 3.7 je roven

$$\hat{OR} = \frac{m \cdot (n_{\bullet\bullet\bullet} - n_{\bullet\bullet 1} - n_{1\bullet\bullet} + m)}{(n_{1\bullet\bullet} - m)(n_{\bullet\bullet 1} - m)} = \frac{116 \cdot 223}{7 \cdot 20} = 185,$$

takže člověk s nějakou vysokou školou 185 krát častěji uspěje u testu než člověk bez vysokoškolského vzdělání, což je mimořádně silná souvislost. Přesto i u této tabulky, která je nejvíce příznivá obhajobě National Revenue Service of Canada, dochází k zamítnutí hypotézy, že při povyšování pracovníků nedocházelo k diskriminaci žen, na hladině testu $\alpha = 0{,}05$.

Tabulka 3.7 – Tabulka, pro kterou je CMH normovaná odchylka tak blízko nule, jak je to možné

	Uspěl(a)	Neuspěl(a)	Celkem
Má VŠ	116	7	123
Nemá VŠ	20	223	243
Celkem	136	230	366

4. KORELOVANOST OTÁZEK V DOTAZNÍKU

V poslední kapitole se zaměříme na poněkud jiný druh testování nezávislosti než doposud. Pro některá data je totiž nevhodná analýza pomocí výše uvedených statistik s χ^2 - rozdělením. Příkladem může být situace, kdy je potřeba ověřit, zda se dvě otázky v dotazníku ptají na totéž. Volba metody pro testování takovéto závislosti nemusí být tak zřejmá jako u klasických testů hypotézy o nezávislosti, jak si ukážeme na příkladu z bakalářské práce Bc. Jany Dvořákové s názvem Víc než mrtví: Analýza postoje veřejnosti k pacientům v permanentním vegetativním stavu.

Autorka ve své práci zopakovala část výzkumu z původní práce „More dead than dead: Perceptions of persons in the persistent vegetative state“ zabývající se pohledem anglosaské společnosti na pacienty v permanentním vegetativním stavu, na datech z České republiky. Na základě vyhodnocování dotazníků se snažila zjistit, jak naše společnost řadí lidi živé, mrtvé a trvale v kómatu z hlediska přisuzovaných mentálních kapacit. Část práce byla věnována závislosti odpovědí v dotazníku na náboženském založení respondenta – lze se domnívat, že věřící člověk bude hodnotit permanentní vegetativní stav mnohem hůře než smrt, protože mrtvého člověka nechápe jako tělo ležící v hrobě, ale jako duši, která se setká s Bohem. V kapitole o religiozitě byla mimo jiné testována hypotéza, že se od sebe liší otázky „Jsem duchovně založený člověk“ (1), „Jsem věřící“ (2), „Věřím na posmrtný život“ (3) a „Duše žije i poté, co člověk zemře“ (4), na které respondenti mohli odpovídat na škále od -3 (zcela nesouhlasím) do 3 (zcela souhlasím). Data s odpověďmi jsou uvedena v příloze D.

Možná závislost mezi otázkami byla testována vždy po dvojicích pomocí Spearmanova korelačního koeficientu. Autorka ovšem dospěla k velmi nejasnému závěru: *„Dle mého názoru zde nejde s jistotou říci, zda jsou otázky 1 a 2 nekorelované a že se každá ptá na něco jiného, resp. že jsou chápané rozdílně a odpovědi na ně se liší. Avšak ani nejde s jistotou říci, že jsou stejné a že bych získala stejný výsledek, kdybych se ptala jen na jednu otázku jako v původním experimentu.“*[5] Nabízí se tedy otázka, zda by nebylo vhodné hypotézu otestovat pomocí jiné statistické metody.

4.1 Vyhodnocování vztahů mezi ordinálními proměnnými korelační analýzou

Čerpáno z [6],[7],[8]:

Proměnné, které vyjadřují stupně souhlasu či nesouhlasu s určitými výroky, jsou ordinálního typu, takže u nich má smysl testovat nezávislost v souvislosti s pořadím hodnot na škále. Nehledáme tedy pouze odpověď na to, zda jsou proměnné závislé či ne, ale také v jakém směru můžeme lineární závislost najít (tj. zda se s růstem hodnot jedné proměnné hodnoty druhé proměnné také zvyšují nebo naopak snižují). Jednou z možností, jak takové vztahy hodnotit, je korelační analýza.

4.1.1 Spearmanův korelační koeficient

Jedním ze základních koeficientů, které berou v úvahu i pořadí hodnot, je Spearmanův korelační koeficient, jenž vychází z toho, že jednotlivým hodnotám proměnné X se přiřadí takové pořadí R_k, že $\sum_{k=1}^{n} R_k = \frac{n \cdot (n+1)}{2}$ a hodnotám proměnné Y obdobně pořadí Q_k tak, že $\sum_{k=1}^{n} Q_k = \frac{n \cdot (n+1)}{2}$. Pokud máme data vyjádřena kontingenční tabulkou typu r x r, vypočítáme pořadí následovně:

$$R_1 = \frac{n_{1\bullet} + 1}{2},\ R_i = \sum_{k=1}^{i-1} n_{k\bullet} + \frac{n_{i\bullet} + 1}{2} \text{ pro } 2 \le i \le r,$$

$$Q_1 = \frac{n_{\bullet 1} + 1}{2},\ Q_i = \sum_{k=1}^{i-1} n_{\bullet k} + \frac{n_{\bullet i} + 1}{2} \text{ pro } 2 \le i \le r.$$

Spearmanův korelační koeficient, který nabývá hodnot z intervalu $\langle -1;1 \rangle$, lze potom psát ve tvaru

$$r_S = 1 - \frac{6 \cdot \sum_{i=1}^{r} (R_i - Q_i)^2}{n \cdot (n^2 - 1)}. \qquad (4.1)$$

Jestliže $r_S = 1$ nebo $r_S = -1$, jsou proměnné X a Y korelované, přičemž v prvním případě jde o tzv. pozitivní korelaci (s rostoucími hodnotami proměnné X rostou i hodnoty proměnné Y) a ve druhém o korelaci negativní (s rostoucími hodnotami proměnné X získáváme sestupné pořadí hodnot proměnné Y). Jestliže $r_S = 0$, hovoříme o lineární nezávislosti.

Nulovost Spearmanova korelačního koeficientu lze testovat pomocí statistiky

$$t = r_S \sqrt{\frac{n-2}{1-r_S^2}}, \tag{4.2}$$

která má za platnosti nulové hypotézy Studentovo t - rozdělení o *(n – 2)* stupních volnosti.

4.1.2 Kendallovo tau-b

Nevýhodou Spearmanova korelačního koeficientu může být fakt, že nezohledňuje možná opakování v pozorovaných datech. Pokud se v datech vyskytují shody, je tedy vhodné zvážit použití Kendallova korelačního koeficientu *tau*, definovaného pomocí počtu konkordantních párů n_c (tj. počet dvojic respondentů, ve kterých hodnotí jeden respondent obě proměnné nižší nebo vyšší úrovní než druhý) a diskordantních párů n_d (tj. počet dvojic respondentů, kteří hodnotí proměnné rozdílně) jako

$$\tau = \frac{n_c - n_d}{\frac{n}{2}(n-1)}.$$

Kendallovou *tau* má několik variant, přičemž *tau-b* a *tau-c* zohledňují právě i počty uzlů v datech. Pro čtvercovou tabulku typu *r* x *r* je nejvhodnější použít Kendallovo *tau-b* definované jako

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_c + n_d + t_X) \cdot (n_c + n_d + t_Y)}}, \tag{4.3}$$

kde t_X je počet vázaných párů takových, které obsahují stejnou hodnotu proměnné X, ale různou hodnotu proměnné Y, a t_Y je počet vázaných párů, kde je tomu naopak.

Kendallův korelační koeficient nabývá stejně jako Spearmanův hodnot z intervalu $\langle -1;1 \rangle$ se stejnou interpretací pro $\tau_b = \pm 1$ nebo $\tau_b = 0$.

4.2 Provedení a vyhodnocení testů na konkrétních datech

Nejprve zopakujeme výpočet Spearmanova korelačního koeficientu (4.1) tak, jak byl proveden v práci Bc. Jany Dvořákové. Výsledky jsou uvedeny v tabulce 4.1 (vždy první hodnota v buňce). Dospěli jsme ke stejným číslům jako v původní práci.

Tabulka 4.1 – Spearmanův korelační koeficient				
	otázka 1	otázka 2	otázka 3	otázka 4
otázka 1	---	0,7406 / 15,7038	0,6142 / 11,0905	0,6316 / 11,6086
otázka 2	0,7406 / 15,7038	---	0,5854 / 10,2876	0,5996 / 10,676
otázka 3	0,6142 / 11,0905	0,5854 / 10,2876	---	0,8403 / 22,088
otázka 4	0,6316 / 11,6086	0,5996 / 10,676	0,8403 / 22,088	---

Významnost hodnot korelačního koeficientu lze interpretovat pomocí p-value, které vyšlo ve všech případech menší než 10^{-4}, takže zamítáme hypotézu o lineární nezávislosti ve prospěch alternativy, že se dvojice otázek ptají na totéž.

V literatuře častěji uváděným přístupem je provedení testu nulovosti Spearmanova korelačního koeficientu pomocí statistiky (4.2), jejíž výsledky jsou uvedeny taktéž v tabulce 4.1 (vždy druhá hodnota v buňce). Protože jsou všechny tyto hodnoty větší než 97,5% kvantil studentova rozdělení o 203 stupních volnosti ($t_{203}(0{,}975) = 1{,}9717$) a nacházejí se tudíž v kritickém intervalu, musíme hypotézu o nekorelovanosti otázek zamítnout opět pro všechny dvojice.

Přestože je závěr plynoucí z provedeného testu jednoznačný, musíme si uvědomit, že výsledky mohou být zkreslené vzhledem k výskytu shod v datech, které použitá metoda nezohledňuje, takže pořadí R_i a Q_i použité ve vzorcích (4.1) a (4.2) nemusí být jednoznačné. Z tohoto důvodu je vhodné data otestovat například pomocí Kendallova *tau-b* (4.3), u kterého tento problém odpadá. Výsledky jsou uvedeny v tabulce 4.2.

Tabulka 4.2 – Kendallovo tau-b				
	otázka 1	otázka 2	otázka 3	otázka 4
otázka 1	---	0,6425	0,5207	0,5313
otázka 2	0,6425	---	0,4945	0,5088
otázka 3	0,5207	0,4945	---	0,7638
otázka 4	0,5313	0,5088	0,7638	---

Hodnoty Kendallova korelačního koeficientu vyšly sice nižší než pro Spearmanův korelační koeficient, ale p-value je stále ve všech případech menší než 10^{-4}, takže pro všechny dvojice otázek získáváme stejný závěr, že se hypotéza o nezávislosti otázek zamítá, korelace jsou významné a můžeme říci, že se otázky v dotazníku ptají na totéž. Navíc, protože jsou hodnoty Kendallova *tau-b* kladné, je evidentní, že mezi dvojicemi otázek jsou korelace pozitivní – tj. s rostoucími hodnotami na škále jedné otázky rostou i hodnoty na škále druhé otázky. Je to také korelace, kterou bychom logicky očekávali – například pokud respondenti odpovídali na otázku, zda jsou věřící, „určitě ano“, vyskytovala se v kladných hodnotách škály i jejich odpověď na otázku, zda jsou duchovně založení.

5. ZÁVĚR

Ve své práci jsem se snažila provést analýzu kontingenčních tabulek na různých atraktivních datech vyžadujících vždy použití jiných statistických metod, abych ukázala šířku využití této statistické disciplíny.

Na datech o filmech z Československé filmové databáze jsem provedla testy nezávislosti známou Pearsonovou statistikou, ale také jsem uvedla a vyzkoušela metody, které lze využít v případě, že nejsou splněny podmínky pro použití této statistiky, nebo v případě, kdy nám zamítnutí nulové hypotézy neposkytuje o datech dostatečné informace.

Závislost hodnocení filmů na ostatních dostupných informacích o filmech se projevila ve všech testovaných kategoriích, i když ne vždy tak, jak jsem se původně domnívala. Uživatelé na www.csfd.cz americké filmy nijak nevyzdvihují, spíše naopak je z hodnocení evidentní jejich průměrnost. Filmy z některých pro nás exotičtějších zemí se k nám nedostávají ve velkém množství, ale když už se nějaký snímek z těchto zemí v našich kinech objeví, jsme schopni ocenit jeho kvalitu. Poněkud překvapivé je zjištění závislosti u hodnocení a délky filmu – nejlépe hodnocené jsou filmy dvouhodinové a delší, zatímco snímky mající minutáž v rozmezí 90 až 104 jsou ty nejprůměrnější. Dále se prokázala má původní domněnka, že animované filmy jsou hodnocené lépe než hrané. Naopak zjištění, že hodnocení filmu závisí na roku jeho výroby, se ukázalo v opačném směru, než jsem očekávala, ale tento závěr může být významně ovlivněn výběrem dat, na kterých jsem testy prováděla.

Kapitola o filmech však není dle mého názoru ještě zdaleka vyčerpána, protože možná závislost mezi hodnocením a žánrem filmu musí být testována složitějšími metodami, které přesahují rozsah bakalářské práce, ale dávají mi na druhou stranu možnost na toto téma navázat v práci diplomové.

V další kapitole jsem se zaměřila na netypický příklad trojrozměrné kontingenční tabulky, u které díky zanedbání jedné marginální tabulky došlo k chybnému rozhodnutí soudu. Testování nezávislosti probíhalo u tohoto příkladu poněkud netradičně, protože nebyly známy četnosti v zanedbané marginální tabulce, takže jsem musela použitou

Cochran-Mantel-Haenszelovu statistiku spočítat pro všechny možné hodnoty, které by se v této tabulce mohly objevit, a výsledky potom shrnout do jednoho vyhodnocení.

Poslední kapitolu jsem věnovala metodám z korelační analýzy pro ordinální proměnné, které lze využít například při sestavování dotazníků k otestování možnosti, že by se dvě otázky ptaly na totéž.

Na prvním příkladě jsem chtěla mimo jiné demonstrovat, jak se statistika objevuje v pozadí našeho běžného života a že je vždy dobré se zamyslet nad objektivitou různých denně využívaných dat, kterou dnes bereme jako samozřejmost. Další kapitola ukazuje, že znalost statistických metod analýzy kontingenčních tabulek je potřeba napříč celým profesním spektrem. A nakonec poslední kapitola nastínila možnost využití kontingenčních tabulek všude tam, kde je potřeba sestavování a vyhodnocování dotazníků. Zároveň jsem na tomto příkladě chtěla ukázat, že volba statistické metody nemusí být u reálných dat vždy jednoduchá věc a tím pádem mohou i výsledky velmi záviset na konkrétních rozhodnutích statistika provádějícího analýzu.

Tato práce mi pomohla ujasnit si souvislosti mezi mnoha teoretickými poznatky, které jsem nabyla během svého bakalářského studia. Ve své práci jsem se chtěla zaměřit především na praktické využití metod analýzy kontingenčních tabulek na zajímavých reálných datech, takže si myslím, že jsem cíl své bakalářské práce splnila.

6. POUŽITÁ LITERATURA A ZDROJE

[1] Anděl, J., *Základy matematické statistiky*, 1. vydání, Praha: MATFYZPRESS, 2005

[2] Pecáková, I., *Testy nezávislosti v řídkých kontingenčních tabulkách*, Statistika, ročník 87, číslo 1, rok 2007, str. 61-68

[3] Agresti, A., *Categorical Data Analysis*, 2. vydání, New York: WILEY-INTERSCIENCE, 2002

[4] Gastwirth, J., Krieger, A., Rosenbaum, P., *How a Court Accept an Impossible Explanation*, The American Statistician, ročník 48, číslo 4, listopad 1994, str. 313-315

[5] Dvořáková, J., *Víc než mrtví: Analýza postoje veřejnosti k pacientům v permanentním vegetativním stavu*, bakalářská práce, Univerzita Palackého v Olomouci, 2012

[6] Řezanková, H., *Analýza kategoriálních dat*, 1. vydání, Praha: Nakladatelství VŠE, 2005

[7] Řezanková, H., Kunstová, R., *Analýza vztahů ordinálních proměnných aplikovaná na úrovně kompetencí absolventů vysokých škol*, Informační bulletin české statistické společnosti, ročník 23, číslo 3, září 2012, str. 87-98 [online],
dostupné z: http://statspol.cz/bulletiny/ib-2012-3-web.pdf
[citováno 2.4.2013]

[8] Kendall Tau Rank Correlation Coefficient [online],
dostupné z: http://en.wikipedia.org/wiki/Kendall_tau_rank_correlation_coefficient
[citováno 23.4.2013]