Univerzita Karlova v Praze

Filozofická fakulta

BAKALÁŘSKÁ PRÁCE

FILOZOFICKÁ FAKULTA
UNIVERZITY KARLOVY
V PRAZE

Filip Děchtěrenko

Analýza vícerozměrných kontingenčních tabulek

Katedra psychologie

Vedoucí práce: Petr Boschek, RNDr., CSc.

Studijní program: Psychologie

Praha 2015

Tuto práci bych chtěl věnovat všem lidem dobré vůle a všem potenciálním zájemcům o statistiku. Děkuji svému polštáři, že mi dopřával dobrého spánku a své posteli, že nechala spočinout mé tělo.

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně, že jsem řádně citoval všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 27.7.2015 podpis autora

Název práce: Analýza vícerozměrných kontingenčních tabulek

Autor: Mgr. Filip Děchtěrenko

Katedra: Katedra Psychologie

Vedoucí diplomové práce: RNDr. Petr Boschek, CSc.

Abstrakt: Vícerozměrné kontingenční tabulky jsou vhodným nástrojem pro zápis počtu pozorování kategorických proměnných. Ve vícerozměrných tabulkách vznikají složitější vztahy mezi proměnnými, které nástroje pro dvourozměrné tabulky nezachytí. V této práci jsme představili log-lineární modely a ukázali jejich použití na trojrozměrných tabulkách. Zaměřili jsme se na statistický program SPSS, ve kterém jsme předvedli analýzu na ukázkových datech včetně její interpretace. Analýzu jsme poté zopakovali na dalších umělých datech popisující možné případy, se kterými se výzkumník může v praxi setkat. Analýzu ukázkových dat jsme zopakovali i ve statistickém programu R.

Klíčová slova: vícerozměrné kontingenční tabulky, log-lineární modely, statistika

Title: Analysis of multidimensional contingency tables

Author: Mgr. Filip Děchtěrenko

Department: Department of psychology

Supervisor: RNDr. Petr Boschek, CSc.

Abstract: Multidimensional contingency tables are suitable tool for capturing the count of observations of multiple categorical variables. There are more complex relationships amongst multiple variables which cannot be captured by analytical tools for two variables. In this work, we introduced log-linear models and showed their application on three dimensional tables. We focused on statistical program SPSS, in which we showed analysis on the sample data including the interpretation. We redone the analysis on another artificial data capturing the possible situations which could researcher encounter in the real life. The sample data were also analyzed in statistical program R.

Keywords: multidimensional contingency tables, log-linear models, statistics

Obsah

Úvod

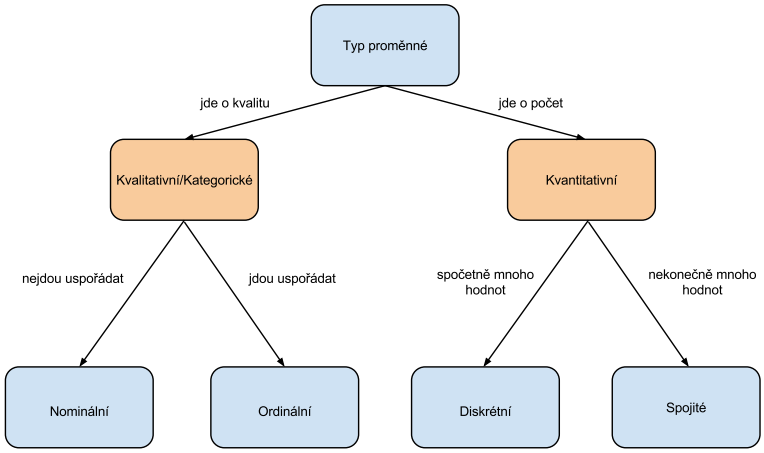
V každém kvantitativním výzkumu se musí výzkumník zabývat analýzou dat a jejich interpretací. V psychologické vědě nás zejména zajímají vztahy mezi jevy a procesy, což vyžaduje vyšší nároky na analytický přístup. Multivariační analýza dat se zabývá vztahy více proměnných. Tyto analýzy bývají mnohem složitější a ve standardních kurzech statistiky se multivariační metody téměř nepokrývají, případně se vyučují již překonané metody, které mají nedostatečnou sílu testu a v případě, že nejsou splněny předpoklady, jsou výsledky silně zkreslené. Ačkoli bylo vyvinuto mnoho pokročilých metod, které překonávají tyto omezení, např. robustní statistické metody (Wilcox, 2012) či Bayesovská statistika (Kruschke, 2014), do širokého podvědomí psychologické obce se zatím nedostaly. Sharpe (2013) poukazuje na velké rozdíly mezi matematickou statistikou jako vědou a statistikou jako nástrojem. Mezi těmito dvěma pohledy na statistiku vzniká velká mezera a je třeba mnoha prací, které se pokusí rozdíly mezi přístupy minimalizovat. Mnohorozměrných metod je velmi mnoho, v této práci si klademe za cíl přiblížit analýzu vícerozměrných kontingenčních tabulek pomocí log-lineárních modelů výzkumníkům tak, aby zvládali analyzovat a interpretovat data ze svých výzkumů a neredukovali analýzy pouze na dvě proměnné. Protože pro velkou část psychologické obce je programování spíše obtížné, zaměříme se práci na statistický program SPSS a nástroj R vyžadující znalosti programování probereme pouze na základní úrovni. V práci ukážeme několik situací, které mohou nastat během výzkumů, včetně interpretace výsledků. V textu budeme zobrazovat výstupy SPSS jako obrázky, přestože jde o tabulky. Rozhodli jsme se takto z důvodu, aby pro čtenáře byly výstupy co nejshodnější s výstupem, co dostane ze své verze SPSS. V práci budeme používat tečku pro označení desetinné čárky, pouze u výstupů SPSS bude kvůli lokálnímu nastavení použita čárka.

1. Význam multivariační analýzy dat v psychologickém výzkumu

Při každodenním životě se setkáváme s mnoha jevy a situacemi. Vztahy mezi těmito jevy nejsou jednoduché, a pokud chceme tyto jevy správně popsat, nesmíme naše uvažování omezit pouze na dva jevy, ale musíme do analýz zapojit i vztahy k ostatním objektům. V této kapitole si představíme problematiku analýzy více proměnných. Dále si popíšeme možné vztahy tří proměnných. Z těchto vztahů budeme vycházet v dalších kapitolách.

1.1 Typy proměnných dle měřítka

Objekty ve světě nemají stejné vlastnosti. Některé proměnné můžeme považovat za čísla a podle toho s nimi pracovat, jiné proměnné jsou neporovnatelné, protože nabývají různých kvalit. Abychom mohli aplikovat správnou statistickou metodu, musíme nejprve vědět, jakého typu daná proměnná je. Jednu z možných klasifikací proměnných můžeme vidět na Obrázku 1.1, nicméně v této práci se budeme zabývat pouze nominálními, ordinálními a diskrétními (s omezenou doménou), souhrnně označované jako kategoriální data. Je potřeba zdůraznit, že diskrétní



Obrázek 1.1: Rozdělení proměnných v závislosti na jejich měřítku. V této práci se budeme zabývat pouze kategorickými proměnnými.

proměnné se považují za kategoriální data pouze v případě, pokud nabývají několika málo hodnot a nepředpokládáme normalitu dat (Agresti, 2002). Kategoriální proměnné můžeme dělit podle počtu kategorií na dichotomické (alternativní) a vícekategoriální (množné).

1.2 Multivariační analýza dat

V psychologických výzkumech se často dostáváme do situací, kde potřebujeme variovat více než jednu proměnnou. Pro statistickou analýzu těchto výzkumů potřebujeme adekvátní nástroje, které by nám umožnily rozhodnout o významnosti nálezu. Podle typu proměnné používáme parametrické nebo neparametrické testy (Hendl, 2006). Ačkoli analýza kvantitativních proměnných pomocí parametrických testů vede k vyšším silám testu (Wilcox, 2012), v mnoha experimentálních designech nemáme jinou možnost, než použít neparametrické testy. Mezi velmi rozšířené metody pro kvantitativní data patří analýza rozptylu (ANOVA) či vícerozměrné regresní metody (Hendl, 2006; Aiken & West, 1991; Rutherford, 2001). Mezi novější metody vycházející ze zobecněných lineárních modelů patří lineární smíšené modely (Stroup, 2012). V mnoha případech ale nemáme k dispozici kvantitativní proměnné, ale musíme pracovat s kategorickými daty. Pro kategorická data máme k dispozici několik nástrojů, mimo jiné i vícerozměrné kontingenční tabulky a log-lineární modely. Kontingenční tabulky jsou oblíbeným nástrojem mnoha výzkumníků, protože práce s nimi je jednodušší, než s jinými složitějšími nástroji. Běžně se pracuje s dvourozměrnými tabulkami, ve kterých zachycujeme počty pozorování u všech kombinací úrovní pro jednotlivé proměnné. Tyto tabulky jdou snadno rozšířit na vícerozměrné, nicméně analýza použitelná pro dva rozměry je pro více rozměrů nedostatečná, protože vztahem více proměnných nám mohou vznikat složitější vztahy nezachytitelné dvourozměrnou analýzou. Jedná se především o problém třetí proměnné, v kontingenčních tabulkách známý jako Simpsonův paradox (Simpson, 1951), tedy jako vznik falešného vztahu mezi dvěma proměnnými, který zmizí, když do analýzy zahrneme třetí proměnnou. Více o tomto problému píšeme v sekci o kontingenčních tabulkách. Protože s rostoucí složitostí tabulek roste i složitost analýzy, používají se log-lineární modely, které umožňují automatizovat zkoumání vztahů mezi proměnnými. Výhoda log-lineárních modelů je jich blízká podobnost s modely pro kvantitativní data, které vycházejí ze zobecněných lineárních modelů. Výzkumník tedy může snadno přenášet získané dovednosti analýzy i na proměnné s jiným měřítkem. Ačkoli mohou být kontingenční tabulky lákavou alternativou k náročné analýze kvantitativních dat, je třeba se snažit vždy o zachování co nejobecnějšího měřítka, neboť každou kategorizací se ztrácí část informace (Maxwell & Delaney, 1993; MacCallum, Zhang, Preacher & Rucker, 2002).

1.3 Vztahy mezi proměnnými

V této části si popíšeme všechny možné vztahy mezi třemi proměnnými[1]. V této části nerozlišujeme mezi měřítkem jednotlivých proměnných, tyto vztahy platí obecně. Nejprve se zaměříme na dvě proměnné, poté přidáme do vztahu i třetí proměnnou.

Nejjednodušším vztahem dvou proměnných je jejich *nezávislost*. V tomto případě je hodnota jedné proměnné nezávislá na hodnotách druhé proměnné. Příkladem takového vztahu může být známka z testu a barva ponožek. Tento vztah budeme zapisovat jako $X \perp Y$. Opakem nezávislosti je závislost, tedy hodnota jedné proměnné ovlivňuje hodnotu druhé (pozor, ovlivňováním není nutně myšlena kauzalita, jedná se čistě o vztahy z hlediska pravděpodobnosti). Závislé proměnné mohou být známka z testu a čas věnovaný studiu. Závislost pro proměnné X a Y budeme značit jako XY

Přidáme-li třetí proměnnou, máme obdobně nezávislost všech tří proměnných (někdy označované jako *úplná nezávislost*) budeme označovat jako $X \perp Y \perp Z$). Příkladem může být známka z testu, barva ponožek a počasí na druhé straně zeměkoule. Taktéž můžeme mít závislost všech tří proměnných, např. čas věnovaný domácím pracím, čas věnovaný studiu a čas věnovaný koníčkům. Všechny tři proměnné se navzájem ovlivňují a při každé úrovni jednotlivé proměnné se míra vztahu liší, tedy když věnuji málo času domácím pracím, můžu věnovat dvakrát tolik času koníčkům na úkor učení, narozdíl od případu, kdy věnuji domácím pracím hodně času a závislost bude obrácená. Toto je pouze ukázka, jak může vypadat úplná závislost mezi proměnnými. V praxi není tento vztah moc užitečný, neboť nám neumožňuje redukovat vztahy mezi proměnnými na jednodušší. Tento vztah budeme označovat pro tři proměnné X, Y a Z jako XYZ. Novým vztahem mezi třemi proměnnými je *sdružená nezávislost*. U tohoto vztahu platí, že pokud sdružíme hodnoty dvou proměnných v jednu (tedy vytvoříme novou proměnnou, která kóduje všechny kombinace hodnot druhých dvou proměnných), tak tato nová proměnná bude nezávislá na původní třetí proměnné. Jinými slovy vztah dvou proměnných není závislý na hodnotě třetí proměnné. Tento vztah budeme značit jako $XY \perp Z$, pokud je Z nezávislá na sdružené proměnné X a Y. Příkladem může být známka z testu, pohlaví a barva ponožek studenta. Mezi známkou a pohlavím může a nemusí být závislost, ale vztah těchto dvou proměnných nebude záviset na barvě ponožek. Stejně tak nebude závislost mezi barvou ponožek a výsledkem a mezi barvou ponožek a časem stráveným

[1]Pro více proměnných vypadají vztahy obdobně, pouze narůstá počet kombinací

studiem. Sdružená nezávislost je speciálním případem úplné nezávislosti.
Dalším vztahem je *podmíněná nezávislost*. U podmíněné nezávislosti jsou dvě proměnné nezávislé, zafixujeme-li třetí proměnnou. Tento vztah budeme značit $(X \perp Y) \mid Z$. Příkladem podmíněné nezávislosti je známka z testu na počtu hodin strávených hraním počítačových her den předtím zafixujeme-li pohlaví, tedy budeme-li analyzovat závislost mezi známkou a počtem hodin pro obě pohlaví zvlášť. Posledním možným vztahem mezi proměnnými je *homogenní asociace*. Homogenní asociace je jednodušší než úplná závislost, ale složitější než podmíněná asociace. Pokud platí homogenní asociace, tak pro jednotlivé úrovně proměnné X musí být vztah mezi Y a Z identický (u podmíněné nezávislosti musí být pro jednotlivé úrovně X proměnné Y a Z nezávislé). Homogenní asociaci budeme značit jako $XY \perp YZ \perp XZ$).
V případě vyššího počtu proměnných nám roste počet možných vztahů od úplné nezávislosti po úplnou závislost. V této práci se těmito vztahy zabývat nebudeme a zaměříme se pouze na trojrozměrný případ.

2. Teoretický popis metod pro vícerozměrné kontingenční tabulky

Kontingenční tabulky používáme pro analýzu kategoriálních dat. Nejprve se zaměříme na dvourozměrný případ, poté přejdeme na analýzu případů s více proměnnými.

2.1 Dvourozměrné tabulky

Tato kapitola si klade za cíl představit základní terminologii u dvourozměrných tabulek, aby se čtenář rychle seznámil s problematikou, a pak mohl snadno pokračovat s vícerozměrnými tabulkami.

2.1.1 Tabulky 2 × 2

Dvourozměrná kontingenční tabulka zachycuje vztah dvou proměnných. V nejjednodušším případě může každá z proměnných nabývat 2 hodnot. Kontingenční tabulku pro případ 2×2 můžeme vidět v Tabulce 2.1. Hodnoty n_{11} až n_{22} označují počty výskytů jednotlivých kombinací znaků. Dále jsou v tabulce zobrazeny částečné součty řádkové (n_{1*} a n_{2*}) a sloupcové (n_{*1} a n_{*2}) a součet celkový (n). Pro tabulku platí, že $n_{11} + n_{12} = n_{1*}$ (obdobně pro ostatní částečné součty) a $n_{11} + n_{12} + n_{21} + n_{22} = n_{**}$. V moderních statistických programech jsou tyto součty kontrolovány automaticky, nicméně při ruční tvorbě je doporučeno tyto vztahy zkontrolovat, aby se zamezilo chybám. Tabulka se někdy zapisuje

	$Y = 1$	$Y = 2$	Součet řádků
$X = 1$	n_{11}	n_{12}	n_{1*}
$X = 2$	n_{21}	n_{22}	n_{2*}
Součet sloupců	n_{*1}	n_{*2}	n

Tabulka 2.1: Základní schéma kontingenční tabulky 2×2

ve formě relativních četností, tedy namísto hodnoty n_{ij} budeme mít hodnotu $p_{ij} = \frac{n_{ij}}{n}$. Obdobně můžeme vyjádřit řádkové (resp. sloupcové) součty pomocí

výrazů $p_{i*} = \frac{n_{i*}}{n}$ (resp. $p_{*j} = \frac{n_{*j}}{n}$). Tato forma nám umožní data zobrazit poměrně vůči celku, což ze samotných frekvencí nemusí být patrné. Tyto relativní četnosti nám vypovídají o pravděpodobnosti, s jakou dostaneme hodnotu s danými hodnotami kategorií X a Y. Poslední z často používaných vyjádření vztahů mezi proměnnými je podmíněná relativní četnost, kterou spočítáme pomocí vzorce $p_{i|j} = \frac{n_{ij}}{n_{*j}}$. Podmíněná pravděpodobnost popisuje pravděpodobnost, že proměnná Y nabyde hodnoty i, za předpokladu, že proměnná X bude rovna j.

2.1.2 Tabulky $r \times s$

V případě, že proměnné X a Y nabývají více hodnot, vypadá situace obdobně. Obecně se tyto tabulky značí jako $r \times s$ tabulky (Hendl, 2006) a v Tabulce 2.2 můžeme vidět konkrétní příklad ze článku v impaktovaném časopise od Galmarini, Symoneaux, Chollet a Zamora (2013).

	Argentina	France	Total
Women 18–30 years old	93	84	177
Men 18–30 years old	49	56	105
Women 31–70 years old	99	103	102
Men 31–70 years old	70	68	137
Total	311	311	621

Tabulka 2.2: 4×2 kontingenční tabulka zachycující počet pokusných osob rozdělených dle pohlaví, věku a příslušnosti ke státu.

2.1.3 Latentní versus pozorovaná proměnná

Při práci s kontingenčními tabulkami je důležité pochopit rozdíl mezi latentními a pozorovanými proměnnými. Pokud dichotomizujeme normálně rozdělenou proměnnou, dostáváme kategorickou proměnnou jiné podstaty, než když pracujeme se skutečně binární proměnnou. Příklad dichotomizované proměnné může být například rozdělení pokusných osob na vysoko skórující a nízko skórující, příklad skutečně binární proměnné je například rozdělení podle pohlaví. Při dichotomizaci se ztrácí část informace z dat a může to vést ke zkresleným závěrům (MacCallum et al., 2002). Aiken a West (1991) doporučují pracovat s proměnnými obsahující maximální informaci, aby nedocházelo k chybám prvního a druhého druhu. My

se v práci zaměříme na ty proměnné, které pocházejí ze skutečně nominálních hodnot.

2.1.4 Statistické testy v kontingenčních tabulkách

Analýzy kontingenčních tabulek odpovídají na otázku, zda jsou na sobě dvě proměnné nezávislé, tedy zda se na základě příslušnosti do jedné z kategorií v první proměnné můžeme předpovídat příslušnost do kategorie druhé proměnné. V případě, že proměnné jsou nominální, můžeme testovat nezávislost v kontingenčních tabulkách pomocí Chí-kvadrát testu (Field, Miles & Field, 2012). Chí-kvadrát test funguje následovně. Nejprve spočítáme testovou statistiku podle následujícího vzorce:

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

kde n_{ij} je počet dat z dané buňky kontingenční tabulky a m_{ij} jsou očekávané hodnoty, pokud by proměnné byly nezávislé. Tyto hodnoty spočítáme ze vzorce $\frac{n_{*i}n_{j*}}{n}$. Poté tuto testovou statistiku srovnáme s kritickou hodnotou Chí-kvadrát rozdělení, a pokud je větší, zamítáme nulovou hypotézu o nezávislosti proměnných. Testová statistika pří Chí kvadrát testu nezávislosti má Chí-kvadrát rozdělení pouze přibližně. Jedna alternativa je provést Yatesovou korekci (Yates, 1934), ale takto korigovaný test dává příliš konzervativní závěry. V případě malých vzorků můžeme použít Fisherův exaktní test (Fisher, 1922) nebo silnější Barnardův exaktní test (Barnard, 1947). Ačkoli nám Chí-kvadrát test může rozhodnout, zda jsou dané proměnné nezávislé, neříká nám nic o síle vztahu proměnných (velikosti efektu, (Cohen, 1988)). V následující části popíšeme ty koeficienty, které jsou dostupné v dostupných statistických softwarech. Detailnější popis dalších koeficientů asociace popisuje (Warrens, 2008). Pro popis velikosti efektu se standardně používají následující koeficienty.

Phí koeficient

Známy koeficient pro míru asociace v 2×2 kontingenčních tabulkách je koeficient ϕ definován jako

$$\phi = \sqrt{\frac{\chi^2}{n}},$$

kde χ^2 je testování statistika a n je celkový počet pozorování. Alternativně se taktéž zapisuje jako $\phi = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n}}$. Koeficient ϕ popisuje sílu vztahu latentních binárních proměnných a ačkoli teoreticky může nabývat hodnot mezi 0 (žádná

asociace) a 1 (úplná asociace), často jsou extrémní hodnoty nižší, pokud se obě proměnné nevyskytují se stejnou pravděpodobností.

Tetrachorický korelační koeficient

V případě, že proměnná byla dichotomizovaná na dvě kategorie, můžeme vypočítat sílu asociace latentních proměnných pomocí Tetrachorického korelační koeficientu. Tetrachorický korelační koeficient spočítáme ze vztahu

$$\phi_{tet} = \cos 180/(1 + \sqrt{\frac{n_{12}n_{21}}{n_{11}n_{22}}})$$

Tetrachorický korelační koeficient nabývá hodnot z rozmezí -1 až 1 a existuje vztah mezi tetrachorickým korelačním koeficientem a koeficientem ϕ (Ekström, 2011) přesahující rozsah této práce. Tetrachorický koeficient lze rozšířit na polychorický korelační koeficient pro určení míry asociace v $r \times s$ tabulkách, kde jsou obě proměnné ordinální (Olsson, 1979)

Cramerovo V

Pro obecné tabulky $r \times s$ můžeme vyjádřit sílu vztahu pomocí Cramerova V (Hendl, 2006). Cramerovo V spočítáme ze vzorce

$$V = \sqrt{\frac{\chi^2}{n(m-1)}},$$

kde n a m jsou počty řádků a sloupců kontingenční tabulky, přičemž m je větší z obou čísel. Pro čtyřpolní tabulky se oba koeficienty rovnají. Cramerovo V nabývá hodnot mezi 0 a 1 a obdobně jako u ostatních měr asociace označují hodnoty kolem 0 žádnou míru asociace a kolem 1 úplnou asociaci.

2.1.5 Poměr šancí

Často používanou mírou asociace je poměr šancí (odds ratio), který se značí jako θ. Tato míra asociace se často používá u log-lineárních modelů a zároveň má poměrně intuitivní interpretaci. Šancí rozumíme podíl pravděpodobnosti, že pro danou řádku nastane daný jev ku pravděpodobnosti. že daný jev nenastane. Máme-li tedy čtyřpolní tabulku, jde o poměr $\frac{P(Y=1|X=1)}{P(Y=2|X=1)} = \frac{p_{1|1}}{p_{2|1}}$ pro první řádku a $\frac{P(Y=1|X=2)}{P(Y=2|X=2)} = \frac{p_{1|2}}{p_{2|2}}$ pro řádku druhou. Poměr šancí je podílem těchto dvou výrazů, tedy jednoduchou úpravou dostaneme pro čtyřpolní tabulku vzorec

$$\theta = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

. Poměr šancí interpretujeme následovně:

- $\theta = 1$ Proměnná Y se chová stejně pro obě hodnoty proměnné X, tedy není asociace mezi proměnnými.
- $\theta > 1$ Mezi proměnnými je pozitivní asociace, tedy v případě $X = 1$ je θ-krát vyšší šance, že $Y = 1$ a obráceně.
- $\theta < 1$ Mezi proměnnými je negativní asociace, tedy v případě $X = 1$ je θ-krát vyšší šance, že $Y = 2$ a obráceně.

Pro $r \times s$ tabulky, počítáme poměr šancí vždy na redukované 2×2 tabulky, a tedy se asociace týká vždy daných dvou proměnných. Protože poměr šancí není symetrický kolem hodnoty 1, používá se někdy logaritmická varianta (log-odds ratio), pro kterou platí, že obě možnosti jsou stejně pravděpodobné, pokud je logaritmus poměru šancí roven 0. V log-lineárních modelech jsou jednotlivé odhady parametrů přímo logaritmickým poměrem šancí.
S poměrem šancí souvisí i relativní risk. Relativní risk je poměrem relativní četnosti v rámci jedné kategorie. Tedy v rámci čtyřpolní tabulky je relativní risk pro $Y = 1$ roven $rr = \frac{\frac{n_{11}}{n_{1*}}}{\frac{n_{21}}{n_{2*}}}$ a obdobně pro $Y = 2$.

2.2 Vícerozměrné tabulky

Vícerozměrné tabulky nám zachycují počty pozorování u třech a více proměnných. Zaměříme se detailněji na trojrozměrné tabulky, neboť pro více proměnných je dají použité metody snadno zobecnit a navíc se trojrozměrné tabulky dají vizualizovat, což je u vyšších dimenzí problém. V celém zbytku této teoretické části budeme označovat tři proměnné jako X,Y a Z. V praktické části práce budeme tyto proměnné označovat smysluplnými názvy.

2.2.1 Popis tabulek

Vícerozměrné tabulky se zapisují podobně jako dvourozměrné tabulky. Součty jednotlivých proměnných budeme značit podobně jako u dvourozměrných tabulek, tedy například n_{3*2} značí součet proměnných $n_{312} + n_{322} + n_{332}$. Pro větší přehlednost nebudeme jednotlivé součty do tabulky zapisovat. V Tabulce 2.3 můžeme vidět tabulku pro tři proměnné (dvě binární a jedna se třemi kategoriemi) Obdobně jako u dvou rozměrné varianty budeme zapisovat teoretické relativní četnosti pomocí zápisu $p_{ijk} = \frac{n_{ijk}}{n}$ (obdobně i pro součty v jednotlivých dimenzích). Ačkoli by výzkumník mohl analyzovat pouze dvourozměrné tabulky a použít nástroje známe z dvourozměrných tabulek, tato analýza by mohla vést

	$Z=1$		$Z=2$		$Z=3$	
	$Y=1$	$Y=2$	$Y=1$	$Y=2$	$Y=1$	$Y=2$
$X=1$	n_{111}	n_{112}	n_{211}	n_{212}	n_{311}	n_{312}
$X=2$	n_{121}	n_{122}	n_{221}	n_{222}	n_{321}	n_{322}

Tabulka 2.3: Základní schéma kontingenční tabulky pro 3 proměnné

k chybným závěrům. Při redukci tabulky sečteme počty pro všechny úrovně redukované proměnné, např. při redukci proměnné Z bychom do buňky n_{11} označující počty $X = 1$ a $Y = 1$ zapsali číslo $n_{111} + n_{211} + n_{311}$. Problém redukce proměnné si můžeme přiblížit na následujícím hypotetickém experimentu.

U pokusných osob sledujeme přítomnost/nepřítomnost 3 znaků (označíme X, Y, Z). Odpovědi „ano“ budeme kódovat jako 1, odpovědi „ne“ jako 2. Pro $n = 1000$ dostaneme tabulku 2.4

	$Z=1$		$Z=2$	
	$Y=1$	$Y=2$	$Y=1$	$Y=2$
$X=1$	180	70	70	180
$X=2$	70	180	180	70

Tabulka 2.4: Ukázka konkrétního problému s redukcí proměnné

Pokud agregujeme proměnnou Z, tabulka se nám redukuje na Tabulku 2.5

	$Y=1$	$Y=2$
$X=1$	250	250
$X=2$	250	250

Tabulka 2.5: Ukázka konkrétního problému s redukcí proměnné

Pokud bychom popsali asociaci proměnných pomocí Cramerova V, dostaneme pro redukovanou tabulku hodnotu $V = 0$, přitom pokud spočítáme Cramerovo V pro dílčí 2×2 tabulky při hodnotách $Z = 1$ a $Z = 2$, dostaneme hodnoty 0.44 pro oba případy. Tento rozdíl ve výsledcích nastává, pokud funguje některá z proměnných jako moderátor (tedy mění úroveň závislosti při rozdílných hodnotách moderující proměnné) a musíme tedy pro analýzu nasadit komplexnější nástroje (Field et al., 2012). Tento jev se někdy nazývá Simpsonův paradox (Simpson, 1951). Z pohledu pravděpodobnosti nás zajímá, zda jsou proměnné X a Y podmíněně nezávislé na proměnné Z. Tato varianta nezávislosti je jen

jedna z mnoha a obdobně bychom mohli zkoumat nezávislosti pro ostatní kombinace proměnných X, Y a Z, ať už podmíněnou nebo nepodmíněnou nezávislost. Přestože by tyto dílčí analýzy šly dělat pomocí Chí-kvadrát testů, pro systematický přístup se používají log-lineární modely nebo logistická regrese.

2.3 Log-lineární modely

Logartimcko-linární modely (Agresti, 2002) spadají do kategorie zobecněných lineárních modelů (Field et al., 2012) s Poissonovskou transformační funkcí. Díky tomu jsou ideálním nástrojem pro analýzu kontingenčních tabulek, neboť počty v jednotlivých buňkách pochází z Poissonovéského rozdělení (Agresti, 2002). Pro analýzu vztahů v kontingenční tabulce lze použít i logistickou regresi. Tento nástroj analýzy je vhodný v případě, že máme jasně dané závislé a nezávislé proměnné. Pokud nás zajímá obecný vztah proměnných, jsou vhodnější log-lineární modely. Log-lineární modely vycházejí z definice nezávislosti v kontingenční tabulky. U log-lineárních modelů zapisujeme závislost ve tvaru

$$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ} + \lambda_{ijk}^{XYZ},$$

kde μ_{ijk} jsou očekávané počty pozorování v buňce $X = i$, $Y = j$ a $Z = k$, kde λ odpovídá celkovému efektu stejnému pro všechny buňky, λ_i^X,λ_j^Y a λ_k^Z odpovídají hlavním efektům jednotlivých proměnných X, Y a Z, tedy jak ovlivňují tyto proměnné logaritmus očekávaných počtů v buňce ijk. Zbývající proměnné popisují interakci mezi proměnnými, konkrétně λ_{ij}^{XY}, λ_{jk}^{YZ} a λ_{ik}^{XZ} odpovídají interakcím mezi dvojicí proměnných za předpokladu, že kontrolujeme třetí proměnnou (jde tedy o podmíněnou pravděpodobnost). Proměnná λ odpovídá interakci všech tří proměnných [1]. Pro jednotlivé úrovně parametrů platí omezující podmínky. Ty mohou nabývat různých tvarů, my budeme pracovat s podmínky podobnými v analýze rozptylu, konkrétně $\sum_i \lambda_i^A = \sum_j \lambda_j^B = \sum_{i,j} \lambda_{i,j}^{AB} \sum_{i,j} \lambda_{i,j,k}^{ABC} = 0$ a obdobně pro další kombinace faktorů. Z těchto vztahů tedy vyplývá, že některé parametry jsou redundantní, např. má-li proměnná A dvě úrovně, stačí určit parametr λ_1^A, protože λ_2^A bude mít stejnou hodnotu, ale pouze opačné znaménko. Koeficienty λ^{AB} odpovídají rovnou logaritmickému poměru šancí v 2×2 kontingenční tabulce s proměnnými A a B [2].

[1] Tento zápis silně připomíná způsob, jakým se zapisuje trojcestná analýza rozptylu. Je to tím, že všechny tyto nástroje vycházejí ze zobecněných lineárních modelů

[2] Odvození např. v (Agresti, 2002)

Tento model obsahující všechny kombinace proměnných se nazývá *saturovaný* a vychází se z něj při procesu výběru modelu či při testování konkrétních hypotéz daným návrhem výzkumu. Pokud ze saturovaného modelu vynecháme různé proměnné, vzniknou nám podmodely odpovídající závislostem (resp. podmíněným závislostem) mezi proměnnými a nás zajímá, zda tyto podmodely nevysvětlují vztahy v kontingenční tabulce přesněji, než saturovaný model.
U jednotlivých modelů můžeme jejich vhodnost pomocí testů dobré shody[3]. Konkrétně můžeme použít Chí-kvadrát test dobré shody nebo testy poměru věrohodnosti (likelihood tests). Oba testy počítají testovou statistiku, která má asymptoticky χ^2 rozdělení. Pro menší vzorky bývá přesnějším odhadem testové statistiky test poměru věrohodnosti (Quine & Robinson, 1985; McDonald, John, 2014). Protože saturovaný model obsahuje parametry odpovídající všem možným vztahům mezi proměnnými, při výpočtu parametrů modelu dosáhneme vždy úplné shody s daty, tedy testy dobré shody budou ukazovat nulovou odchylku od modelu. U dílčích modelů jsou již odchylky nenulové a při procesu výběru modelu s těmito odchylkami pracujeme.

2.3.1 Jednotlivé podmodely

V této části si popíšeme všechny podmodely, které můžeme v trojrozměrném případě uvažovat. Tyto modely odpovídají možným vztahům mezi proměnnými, které jsme popisovali v části 1.3.

Úplná nezávislost

Tento model můžeme zapsat ve tvaru

$$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z,$$

obsahuje tedy pouze hlavní efekty pro jednotlivé proměnné. Jak vyplývá z názvu, tento model neobsahuje žádné vztahy mezi proměnnými.

Sdružená nezávislost

Pro tři proměnné máme tři možné modely popisující sdruženou nezávislost. V souladu s notací z části o vztazích mezi proměnnými můžeme mít $XY \perp Z$, $XZ \perp Y$ nebo $YZ \perp X$. Pokud bychom chtěli zkoumat nezávislost proměnné Z na sdružené proměnné XY, dostali bychom model

$$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY},$$

[3]V anglické literatuře mluvíme o *Goodness of fit* pro konkrétní modely.

a obdobně pro další varianty. Tento model je méně obecný než model pro úplnou nezávislost. Platí tedy, že pokud testy dobré shody ukazují, že model pro úplnou nezávislost vystihuje data, není třeba zkoušet sdruženou nezávislost.

Podmíněná nezávislost

Při podmíněné nezávislosti máme opět tři možné modely: $(X \perp Y) \mid Z$, $(X \perp Z) \mid Y$ a $(Y \perp Z) \mid X$. Pro třetí variantu bude tedy model následovně

$$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ij}^{XZ}$$

Tento model předpokládá, že asociace mezi proměnnými Y a Z je nulová nezávisle na úrovni X.

Homogenní asociace

Poslední z možných modelů vystihuje situaci $XY \perp YZ \perp XZ$, což odpovídá modelu bez interakce všech tří proměnných. Zapisujeme jej tedy ve tvaru

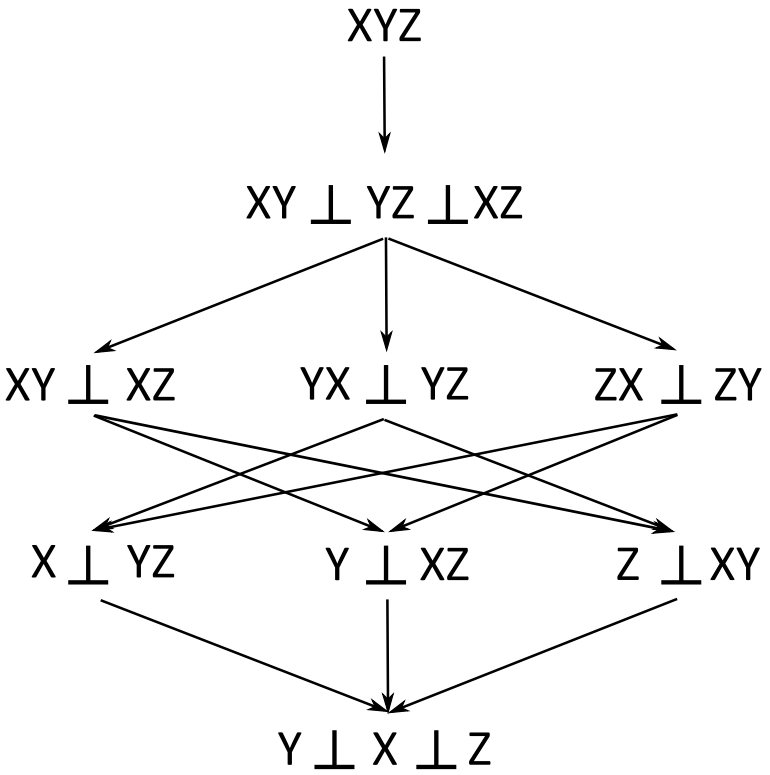
$$log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ij}^{XZ} + \lambda_{ij}^{YZ}$$

V tomto případě je úroveň asociace mezi proměnnými X a Y stejná pro všechny úrovně Z (a obdobně pro další kombinace proměnných).

2.3.2 Výběr modelu

V předchozí části jsme popsali několik modelů, které můžeme použít pro popis našich dat. Otázkou je, který model vybrat jako nejvhodnější. Ačkoli by šel použít saturovaný model pro popis dat (má přesnou shodu s daty), nebývá tento model prakticky užitečný, protože přepokládá závislost všech proměnných. Častokrát lze situaci popsat jednodušším modelem, přestože se data budou mírně od tohoto modelu odchylovat. Jednotlivé modely jsou na sebe hierarchicky navázány, jak je vidět na obrázku 2.1. Modely na stejné úrovni jsou mají stejnou složitost, modely zobrazené v hierarchii výše jsou více komplexnější než modely zobrazené níže. Při porovnávání modelů máme několik kritérií, podle kterých se můžeme rozhodnout. U každého modelu můžeme otestovat, jak dobře vysvětluje data pomocí testu dobré shody. Pokud vyjde test významně, popisuje model data špatně[4]. Tímto způsobem dostaneme pro jednotlivé jejich shodu s daty. Toto pravidlo nám často nemusí stačit pro případy, kdy některé modely dávají větší smysl z hlediska

[4]Hledáme tedy nevýznamnou p-hodnotu. Při běžných statistických testech náš naopak zajímá významná p hodnota.



Obrázek 2.1: Hierarchie log-lineárních modelů. Modely výše položené jsou komplexnější než modely níže.

výzkumného problému nebo teorie. V tomto případě máme k dispozici několik dalších kritérií:

- Podobnost mezi pozorovanými a očekávanými poměry věrohodností - Pro modely spočítáme poměry šancí pro všechny kombinace proměnných, a pokud jsou podobné, vybereme model, který je v hierarchii níže (tedy jednodušší model). Otázka podobnosti šancí je bohužel subjektivní a záleží vždy na konkrétním případu.

- Index nepodobnosti δ (Kuha & Firth, 2011) - Tato statistika nám popisuje, kolik procent dat je odchýleno od svých očekávaných buněk. Tuto statistiku spočítáme pomocí vzorce $\delta = \frac{\sum_i |n_i - \tilde{\mu}_i|}{2n}$, kde $\tilde{\mu}_i$ jsou residua od daného modelu. Index nepodobnosti nabývá vždy hodnot mezi 0 a 1 a pravidlo palce říká, že pro menší hodnoty než 0.03 (tedy méně než 3%) je rozdíl zanedbatelný. Máme-li velké vzorky, budou nám testy dobré shody vycházet významně i při malých rozdílech mezi očekávanými a naměřenými četnostmi. Tento index nám popisuje prakticky použitelnou odchýlenost od modelu.

- Akaikovo informační kritéria (AIC; Burnham, 2004) - Jedním z alterna-

tivních pohledů na odchylky od modelu jsou informační kritéria, která berou v potaz složitost modelu a velikost vzorku. Akaikovo informační kritérium spočítáme ze vzorce $AIC = G^2 - 2 \cdot k$, kde k je počet parametrů v modelu a G^2 je testová statistika testu poměru věrohodností pro dané dva modely. Čím nižší hodnota AIC, tím lépe model vysvětluje data. Toto informační kritérium se často využívá při automatickém výběru modelu.

Kromě těchto kritérií existují i další, například Bayesovské informační kritérium (BIC; Schwarz, 1978) související s Bayesovským přístupem k analýze dat.
Proces výběru modelu bývá časově náročný, protože se musí postupně otestovat jednotlivé modely. V mnoha statistických programech existuje automatický proces výběru modelu, bohužel pro log-lineární modely bývá velmi omezený, a proto je lepší rozhodovat o výběru nejlepšího modelu ručně na základě výše uvedených kritérií.

2.3.3 Log-lineární modely pro čtyř a vícerozměrné tabulky

Log-lineární modely jdou snadno rozšířit do vícerozměrných případů. Vždy máme saturovaný model (obsahuje všechny n-cestné interakce včetně všech efektů nižších řádů) a totální nezávislost všech n-proměnných. Počet možných vztahů mezi proměnnými exponenciálně narůstá a pro počáteční analýzy je tedy vhodné použít automatické hledání modelu (byť výsledky je potřeba ověřit výše zmíněnými přístupy).
U čtyřrozměrných tabulek máme další možné vztahy mezi proměnnými, nicméně dají se interpretovat pomocí sdružených nezávislostí a podmíněných závislostí. Obdoby homogenní asociace pro více proměnných jsou náročné na interpretaci. Pro analýzu vícerozměrných tabulek se často používají grafické modely (Darroch, Lauritzen & Speed, 1980; Edwards, 2012), jejich popis ale přesahuje svoji odbornostní tuto práci.

3. Analýza kontingenčních tabulek v SPSS a R

V této kapitole se zaměříme na analýzu vícerozměrných kontingenčních tabulek pomocí programů SPSS 20.0 a R 3.1.0. Program SPSS je komerční software společnosti IBM pro statistickou analýzu dat v grafickém editoru. Pro tento program jsme se rozhodli pro jeho přístupnost těm psychologům, kteří nejsou zdatní v programování. Nevýhodou používání SPSS je jeho cena a také neaktuálnost použitých statistických metod pro analýzu. Program R je jazyk vhodný spíše pro lidi se znalostí programování, na druhou stranu vzhledem k jeho modulárnímu přístupu v něm můžeme používat i moderní statistické metody. V této práci nebudou popsány základy práce s SPSS nebo R, předpokládáme základní znalost obou nástrojů. Pro účely této kapitoly jsme vytvořili umělá data, na nichž vysvětlíme jednotlivé analýzy. Pro zjednodušení popisu některých operací v SPSS budeme používat následující značení. Otevření jednotlivých položek v menu budu značit šipkami, tedy zápis File → Open → Data... značí otevření menu File, výběr položky Open a poté položky Data.... Většinu operací se budeme snažit zachytit i obrázkem z konkrétních formulářů, u některých triviálních operací pouze popíšeme postup a samotné obrázky dáme do přílohy pro redukci textu.

3.1 Popis dat

Pro ukázkovou analýzu v obou programech budeme umělá data z následujícího experimentu. U pokusných osob jsme si zapisovali jejich pohlaví, informaci, zda kouří, a zda vyrůstali na vesnici nebo ve městě. Data jsou zobrazena v Tabulce 3.1. Programy SPSS i R umí pracovat s touto agregovanou formou, tedy na každém řádku nemáme jednotlivá pozorování, ale pouze počet pozorování pro kombinaci úrovní ostatních proměnných.

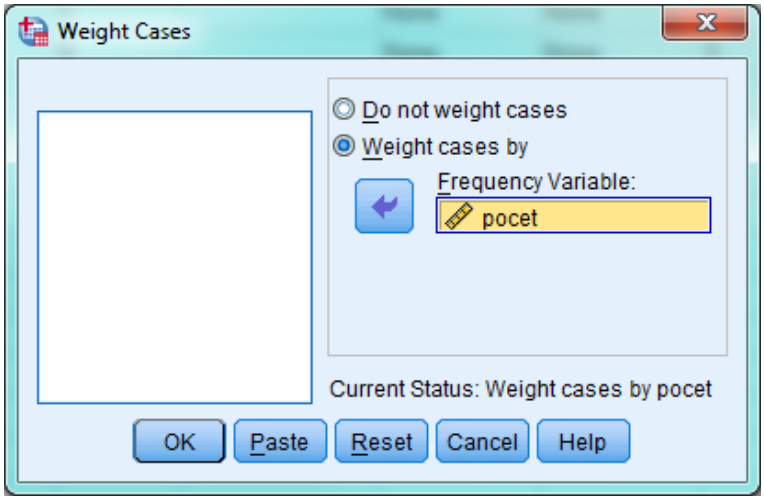
3.2 Analýza pomocí SPSS

Při analýze pomocí SPSS si nejprve ukážeme přístup přes kontingenční tabulky, poté se zaměříme na log-lineární modely. Data načteme klasickým způsobem (File → Open → Data...). Protože máme agregovaná data, musíme ještě přidat váhu jednotlivým řádkům podle jejich počtu (jinak by SPSS považovalo každý řádek

pohlaví	kouří	lokalita	počet
muž	ano	vesnice	10
muž	ne	vesnice	52
žena	ano	vesnice	4
žena	ne	vesnice	303
muž	ano	město	101
muž	ne	město	22
žena	ano	město	476
žena	ne	město	521

Tabulka 3.1: Data používané v této kapitole. Jednotlivé řádky popisují počet pozorování pro dané podmínky. První tři sloupce jsou nominální proměnné, poslední sloupec je intervalová proměnná.

za jedno pozorování), tedy vybereme Data → Weight cases a v nově otevřeném okně vybereme Weight cases by a klikneme na šipku (obrázek 3.1).

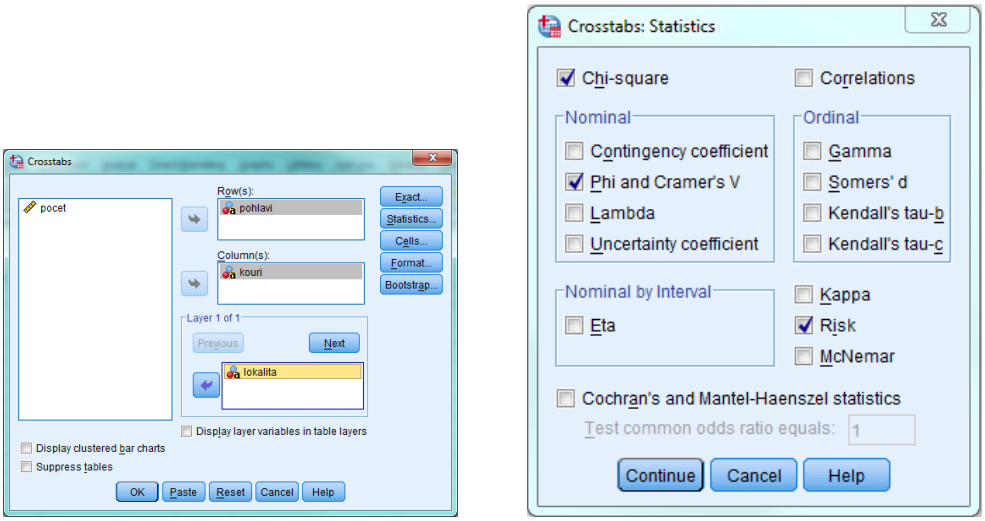


Obrázek 3.1: Formulář pro vážení jednotlivých pozorování při agregovaném zápisu dat.

3.2.1 Kontingeční tabulky

Pro analýzu kontingenčních tabulek otevřeme formulář pomocí Analyze → Descriptive Statistics → Crosstabs. Otevře se nám okno zobrazené na obrázku 3.2a. Zde přiřadíme proměnné, které budou v řádcích tabulky (pohlaví), ve sloupcích (kouri) a dále ve vrstách (lokalita). Vrstva je označení pro další rozměry tabulky a při výpočtu testových statistik se pracuje v jednotlivých vrstvách (dochází

tedy k redukci na menší tabulky). V záložkách na straně můžeme nastavit další parametry. Záložka Exact... umožňuje nastavit parametry pro přesný výpočet testových statistik. Toto nastavení se používá u malých vzorků nebo u vzorků kde některá z buněk má malé počty. Záložka Statistics... obsahuje zaškrtávací políčka pro jednotlivé testy a koeficienty asociace. Pro účely této práce se budeme věnovat jen těm základním: Chí-kvadrát testu, Cramerovo V, koeficient Phí a poměry šancí. V záložce Cells... se dají nastavit další výstupy, např. procentuální zobrazení počtů, očekávané počty v jednotlivých buňkách při platnosti nezávislosti pozorování či odchylky skutečných počtů od očekávaných. Poslední dvě záložky se týkají způsobu řazení řádků tabulky (Format...) a bootstrapovými testy, které se používají v případě malých vzorků a bývají méně konzervativní než Fisherův přesný test (Lin, Chang & Pal, 2014). Po spuštění analýzy se nám



(a) Základní okno (b) Testové statistiky

Obrázek 3.2: Formuláře pro práci s kontingnečními tabulkami v SPSS: (a) Hlavní okno pro kontingenční tabulky. Proměnné jsou již přiřazeny do řádků, sloupců a vrstev. (b) Nastavení použitých testových statistik. V tomto případě je vybrán Chí-kvadrát test nezávislosti, dva koeficienty asociace a poměry šancí.

ve výstupním okně zobrazí několik tabulek 3.3. První je analýza proměnných z hlediska chybějících hodnot. Jde o standardní výstup SPSS nespecifický kontingenčním tabulkám, detailní popis je k nalezení učebnici statistiky pomocí SPSS (Field, 2013). Následuje deskriptivní statistika pro jednotlivé kombinace úrovní zadaných proměnných (3.2a). SPSS reportuje dílčí kontingenční tabulky pro jednotlivé úrovně proměnných zahrnutých ve vrstvách (v tomto případě pro proměnnou lokalita). Pokud by nás zajímaly dvourozměrné tabulky pro jiné kombinace proměnných, musíme tyto proměnné přiřadit do sloupců a řádků ve

formuláři Analyze → Descriptive Statistics → Crosstabs. Následující tabulka obsahuje výsledky testů nezávislosti pro jednotlivé vrstvy tabulky (3.2b). Konkrétně je v ní zobrazen Chí-kvadrát test (řádek označen jako Pearson Chi-Square) včetně použití korekce na spojitost (Continuity correction), poměru věrohodnosti (řádek označen jako Likelihood Ratio), Fisherův přesný test a počet řádků tabulky použitých pro testy (N of Valid Cases).

Předposlední tabulka obsahuje koeficienty asociace, v našem případě Phí koeficient a Cramerovo V (3.4a). V tomto případě jde o malé až střední efekty (Cohen, 1988).

Poslední tabulka obsahuje poměry šancí pro jednotlivé dílčí tabulky (rozdělené

pohlavi * kouri * lokalita Crosstabulation

Count

lokalita			kouri		Total
			ano	ne	
město	pohlavi	muž	101	22	123
		žena	476	521	997
	Total		577	543	1120
vesnice	pohlavi	muž	10	52	62
		žena	4	303	307
	Total		14	355	369
Total	pohlavi	muž	111	74	185
		žena	480	824	1304
	Total		591	898	1489

(a) Počty pozorování pro jednotlivé kombinace proměnných.

Chi-Square Tests

lokalita		Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
město	Pearson Chi-Square	51,787[a]	1	,000		
	Continuity Correction[b]	50,420	1	,000		
	Likelihood Ratio	55,977	1	,000		
	Fisher's Exact Test				,000	,000
	N of Valid Cases	1120				
vesnice	Pearson Chi-Square	31,064[c]	1	,000		
	Continuity Correction[b]	27,135	1	,000		
	Likelihood Ratio	21,615	1	,000		
	Fisher's Exact Test				,000	,000
	N of Valid Cases	369				
Total	Pearson Chi-Square	36,399[a]	1	,000		
	Continuity Correction[b]	35,437	1	,000		
	Likelihood Ratio	35,529	1	,000		
	Fisher's Exact Test				,000	,000
	N of Valid Cases	1489				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 73,43.
b. Computed only for a 2x2 table
c. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 59,63.
d. 1 cells (25,0%) have expected count less than 5. The minimum expected count is 2,35.

(b) Výsledky testů dobré shody.

Obrázek 3.3: Výstup pro kontingenční tabulky: (a) Pro každou úroveň proměnné z vrstvy (v tomto případě pro proměnnou lokalita). Z tabulky je vidět, že u vesnice a města jsou počty jiné. (b) Výsledky jednotlivých testů pro nezávislost. Chí-kvadrát test i poměr věrohodností (likelihood ratio) ukazují, že je zde významná závislost mezi pohlavím a kouřením pro lokalitu město i vesnici.

dle proměnné ve vrstvách, tady proměnná lokalita). Pro každou tabulku je zobrazen v prvním řádku poměr šancí, v druhém relativní risk pro případ, kdy proměnná nabývá první hodnoty a relativní risk pro případ, kdy proměnná nabývá druhé hodnoty. V tomto případně máme poměr šancí v lokalitě město 14.567 ve prospěch ženy (vždy první z kategorií), což znamená, že je přibližně 14krát vyšší šance, že ve městě bude kouřit žena než muž.

Pokud by nás zajímaly komplexnější vztahy mezi proměnnými, musíme použít nástroje pro log-lineární modely.

Symmetric Measures

Lokalita			Value	Approx. Sig.
město	Nominal by Nominal	Phi	,290	,000
		Cramer's V	,290	,000
	N of Valid Cases		369	
vesnice	Nominal by Nominal	Phi	,215	,000
		Cramer's V	,215	,000
	N of Valid Cases		1120	
Total	Nominal by Nominal	Phi	,156	,000
		Cramer's V	,156	,000
	N of Valid Cases		1489	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

(a) Koeficienty asociace pro jednotlivé proměnné.

Risk Estimate

Lokalita		Value	95% Confidence Interval Lower	Upper
město	Odds Ratio for Pohlaví (žena / muž)	14,567	4,404	48,185
	For cohort Kouří? = ano	12,379	4,011	38,205
	For cohort Kouří? = ne	,850	,761	,949
	N of Valid Cases	369		
vesnice	Odds Ratio for Pohlaví (žena / muž)	5,025	3,117	8,101
	For cohort Kouří? = ano	1,720	1,549	1,910
	For cohort Kouří? = ne	,342	,233	,502
	N of Valid Cases	1120		
Total	Odds Ratio for Pohlaví (žena / muž)	2,575	1,879	3,528
	For cohort Kouří? = ano	1,630	1,421	1,870
	For cohort Kouří? = ne	,633	,528	,759
	N of Valid Cases	1489		

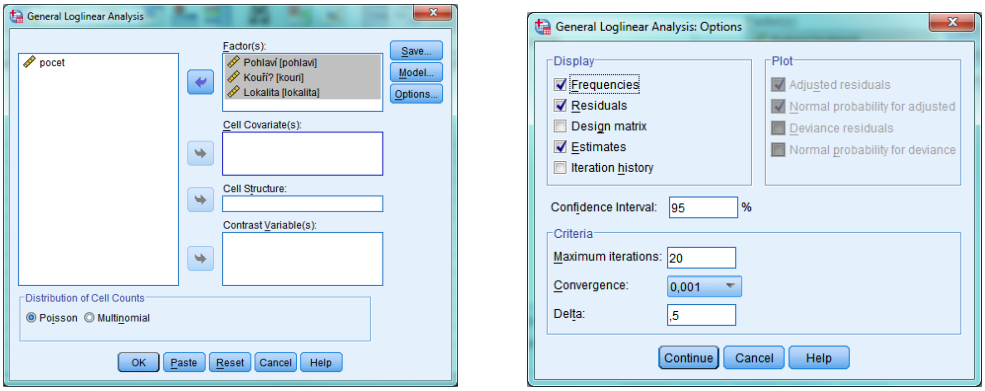
(b) Poměry věrohodností a relativní risk.

Obrázek 3.4: Další výstupy pro kontingenční tabulky: (a) Koeficienty asociace pro jednotlivé vrstvy tabulky. V našem případě jsou zobrazeny koeficient Phí a Cramerovo V. (b) Výsledky jednotlivých testů pro nezávislost. Chí-kvadrát test i poměr věrohodností (likelihood ratio) ukazují, že je rozdíl pro dílčí dvou rozměrné tabulky.

3.2.2 Log-lineární modely - popis nástrojů v SPSS

Nyní se podíváme na analýzu pomocí log-lineárních modelů. SPSS nabízí několik možností. Pokud chceme otestovat konkrétní model, vybereme v menu Analyze → Loglinear → General. Další možnosti v sekci je logit analýza (Analyze → Loglinear → Logit) vhodná pro případy, kdy máme danou závislou a nezávislou proměnnou a výběr modelu (Analyze → Loglinear → Hierarchical model selection), který automatizuje proces výběru modelu na základě p-hodnot testů dobré shody. Ačkoli tento způsob výběru není nejlepší a je lepší použít jiná kritéria jako např. AIC nebo BIC (Miller, 1984; Raftery, 2013), v SPSS nemáme jinou možnost, než použít automatický výběr modelu. SPSS neumožňuje automatické porovnávání hierarchických modelů. Mohli bychom sice porovnávat modely ručně pomocí dekompozice testu poměru věrohodností, ale v tom případě by již bylo lepší provádět analýzu v lepších nástrojích jako například v programu R. Nejprve si ukážeme základní analýzu konkrétního modelu.
Formulář pro zadání log-lineárních modelů je zobrazen na obrázku 3.5.

Stejně jako u kontingenčních tabulek vybereme pomocí šipky proměnné pohlavi, kouri a lokalita. Kdybychom měli další proměnné, které bychom chtěli přidat do analýzy jako kovariát (druhá šipka), přidáme je do příslušné sekce. Další část se používá pro rozlišování tzv. strukturní nebo výběrové nuly (Bishop, Fienberg & W, 2007). Výběrová nula označuje klasicky nulový počet pozorování, a tedy při větším výběru existuje nenulová šance, že do dané buňky budou spadat nějaká



(a) Základní okno (b) Nastavení modelu

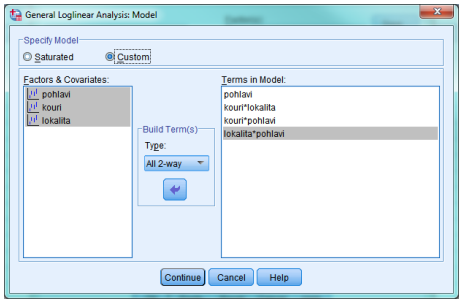
Obrázek 3.5: : (a) Hlavní okno pro log-lineární modely. Proměnné jsou již přiřazeny jako faktory modelu. Jako distribuce bylo vybráno Poissonovo rozdělení (není apriorně nastavená požadovaná velikost vzorku) (b) Nastavení použitých testových statistik. Kromě četností a reziduí byly vybrány i odhady parametrů modelu.

pozorování. Strukturní nula ovšem jiné hodnoty, než nula nabývat nemůže, protože chybějící pozorování jsou daná strukturou dat. Příkladem mohou být data ze studie zjišťující, kolikrát týdně měl člověk sex se svým partnerem. U lidí, kteří nemají stálého partnera, bude toto číslo vždy nula, což se liší od nulového počtu u lidí, kteří stálého partnera mají, ale pouze tento týden k milostnému aktu nedošlo. Pokud bychom měli v datech binární proměnnou označující strukturní nuly, můžeme ji přidat třetí šipkou. Poslední část se používá při výpočtu zobecněných poměru šancí, ale v této práce se jimi nebudeme zabývat. V poslední části tohoto formuláře vybereme, zda byly počty v kontingenční tabulce generovány z Poissonova rozdělení nebo z Multinomiálního. V případě, že při sběru dat nebyl fixován celkový počet pozorování, vycházíme z Poissonova rozdělení, v případě, že byla velikost vzorku dopředu stanovena, použijeme Multinomiální rozdělení (Agresti, 2002).

Tlačítko Save umožňuje uložit výstup jako další proměnné do dat. To se může hodit, pokud chceme s daty dělat další analýzu nebo vizualizaci. Tlačítko Options otevře formulář zobrazený na obrázku 3.5b. V tomto formuláři můžeme nastavit zobrazení jednotlivých statistik, zejména je užitečné zobrazení četností, reziduí a odhadů jednotlivých parametrů (*Frequencies*, *Residuals* a *Estimates*). Dále můžeme nastavit konvergenci algoritmu hledající parametry modelů. Pro většinu případů můžeme nechat standardní nastavení. Pro nesaturované modely můžeme ještě upravit zobrazení grafů.

Standardně pracujeme se saturovaným modelem, pokud nás zajímá jiný model,

vybereme si ho přes tlačítko *Model*. Toto tlačítko nám otevře formulář zobrazený na obrázku 3.6. Pro výběr modelu nejprve zaškrtneme přepínač Custom a



Obrázek 3.6: Formulář pro výběr modelu. Vybereme proměnné, které přidáme do modelu jako hlavní efekty či jako interakce.

poté vždy označíme proměnné, které chceme přidat do modelu. Pokud chceme přidat i jejich interakci, označíme obě proměnné a rozbalovacího seznamu vybereme Interaction. V témže seznamu je již několik před vybraných modelů, což urychlí zadávání. Kdybychom chtěli zadat model homogenní asociace, vybereme pro všechny kombinace proměnných možnost All 2-way. V seznamu napravo se interakce mezi proměnnými zobrazují pomocí hvězdičky. Nyní nechme vybraný saturovaný model.

Po nastavení všech parametrů klikneme na OK a na výstupu se nám zobrazí výsledky analýzy. V první tabulce jsou zobrazeny počty jednotlivých pozorování, včetně chybějících dat a případně strukturních/výběrových nul. Také tam je zobrazen počet úrovní kategorií pro jednotlivé proměnné. V další tabulce nalezneme informace o konvergenci algoritmu, hledající odhady jednotlivých parametrů modelu. Pro nás je důležitá informace, že algoritmus úspěšně dokonvergoval k výsledku, což poznáme z poznámky *c*. V případě, že by algoritmus nedokonvergoval, musíme zkontrolovat data, zda tam nemáme lineárně závislé proměnné. Možností, jak tento problém vyřešit, je odstranit korelované proměnné (Hendl, 2006). Další tabulka obsahuje výsledky testů dobré shody, konkrétně Chí-kvadrát test dobré shody a test poměru věrohodností. V tomto případě jsou obě testové statistiky rovné nule, protože saturovaný model má shodu s daty vždy úplnou. Poznámky *a* a *b* popisují distribuci pro model a testovaný model.

V další tabulce vidíme pro jednotlivé kombinace úrovní proměnných pozorované a očekávané počty (3.7, včetně procentuálního vyjádření). V každém řádku jsou vidět i rezidua, včetně standardizovaných a upravených standardizovaných reziduí (Cook & Weisberg, 1982). Standardizovaná rezidua (někdy taky označovaná jako Pearsonova rezidua) korigují rozdíly pomocí očekávaného průměru, a tím

Cell Counts and Residuals

Pohlaví	Kouří?	Lokalita	Observed		Expected		Residuals	Std. Residuals
			Count[a]	%	Count	%		
žena	ano	město	10,500	0,7%	10,500	0,7%	,000	,000
		vesnice	101,500	6,8%	101,500	6,8%	,000	,000
	ne	město	52,500	3,5%	52,500	3,5%	,000	,000
		vesnice	22,500	1,5%	22,500	1,5%	,000	,000
muž	ano	město	4,500	0,3%	4,500	0,3%	,000	,000
		vesnice	476,500	32,0%	476,500	32,0%	,000	,000
	ne	město	303,500	20,4%	303,500	20,4%	,000	,000
		vesnice	521,500	35,0%	521,500	35,0%	,000	,000

a. For saturated models, ,500 has been added to all observed cells.

Obrázek 3.7: Očekávané a pozorované počty v jednotlivých buňkách, včetně reziduí. Protože jde o saturovaný model, byla ke všem buňkám přidána hodnota 0.5.

dostaneme normované rezidua vhodná k rozpoznání buněk, které se významně vychylují od modelu. Upravená standardizovaná rezidua jsou vhodnější metrikou pro odchylky buněk od modelu. Pokud absolutní hodnota buňky přesáhne hodnotu 2 až 3, můžeme buňku považovat za odchýlenou od nulového modelu. V tomto případě jsou všechny odchylky rovny nule, protože jak jsme již zmínili výše, saturovaný model má úplnou shodu s daty. Následuje důležitá část výstupu - odhady jednotlivých parametrů modelu (Obrázek 3.8. V tabulce jsou popsány pouze ty parametry, které nejsou redundantní[1]. Pro jednotlivé parametry máme vyjádřeny jejich odhady včetně standardní chyby, testové statistiky, p-hodnoty a konfidenčního intervalu pro odhady parametrů. Pokud dosadíme tyto odhady do rovnice log-lineárního modelu, dostaneme očekávané počty z předchozí tabulky. Z-skóry a p-hodnoty nám mohou napovědět, které proměnné vyhodit z modelu pro další testování, nicméně v samotném testováním modelů se spíše častěji používá přístup přes testy dobré shody. Posledními hodnotami pro každou proměnnou jsou dolní a horní limity konfidenčního intervalu. Z nich můžeme vyčíst, jak variabilní je skutečný efekt v populaci. Toto je zejména užitečné při velkých vzorcích, při kterých i malé odchylky od modelu vycházejí významně. V posledních dvou tabulkách nalezneme korelaci a kovarianci jednotlivých koeficientů.

Při výběru modelu vypadá formulář podobně. Nejprve vybereme kategorické proměnné do modelu. Poté pomocí tlačítka *Define range* nastavíme rozsahy pro jednotlivé proměnné. V našem případě má každá proměnná rozsah 1 až 2. Obdobně jako při zadávání konkrétního modelu bychom nyní mohli vybrat parametry, které budou použity pro počáteční inicializaci hledání nejvhodnějšího modelu. Jelikož výběr začíná ve většině případů od saturovaného modelu, nebudeme toto nastavení potřebovat. Ostatní nastavení můžeme nechat beze změny. Po spuštění

[1]Pro každý parametr o 1 méně, než je počet úrovní daná proměnné, protože poslední hodnotu parametru můžeme vyjádřit pomocí celkového počtu a předchozích parametrů

Parameter Estimates[b,c]

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant	6,257	,044	142,881	,000	6,171	6,343
[pohlavi = 1]	-3,143	,215	-14,598	,000	-3,565	-2,721
[pohlavi = 2]	0[a]	.	.	.	.	.
[kouri = 1]	-,090	,063	-1,424	,154	-,214	,034
[kouri = 2]	0[a]	.	.	.	.	.
[lokalita = 1]	-,541	,072	-7,498	,000	-,683	-,400
[lokalita = 2]	0[a]	.	.	.	.	.
[pohlavi = 1] * [kouri = 1]	1,597	,241	6,612	,000	1,123	2,070
[pohlavi = 1] * [kouri = 2]	0[a]	.	.	.	.	.
[pohlavi = 2] * [kouri = 1]	0[a]	.	.	.	.	.
[pohlavi = 2] * [kouri = 2]	0[a]	.	.	.	.	.
[pohlavi = 1] * [lokalita = 1]	1,389	,262	5,298	,000	,875	1,902
[pohlavi = 1] * [lokalita = 2]	0[a]	.	.	.	.	.
[pohlavi = 2] * [lokalita = 1]	0[a]	.	.	.	.	.
[pohlavi = 2] * [lokalita = 2]	0[a]	.	.	.	.	.
[kouri = 1] * [lokalita = 1]	-4,121	,479	-8,602	,000	-5,060	-3,182
[kouri = 1] * [lokalita = 2]	0[a]	.	.	.	.	.
[kouri = 2] * [lokalita = 1]	0[a]	.	.	.	.	.
[kouri = 2] * [lokalita = 2]	0[a]	.	.	.	.	.
[pohlavi = 1] * [kouri = 1] * [lokalita = 1]	1,005	,631	1,593	,111	-,232	2,242
[pohlavi = 1] * [kouri = 1] * [lokalita = 2]	0[a]	.	.	.	.	.
[pohlavi = 1] * [kouri = 2] * [lokalita = 1]	0[a]	.	.	.	.	.
[pohlavi = 1] * [kouri = 2] * [lokalita = 2]	0[a]	.	.	.	.	.
[pohlavi = 2] * [kouri = 1] * [lokalita = 1]	0[a]	.	.	.	.	.
[pohlavi = 2] * [kouri = 1] * [lokalita = 2]	0[a]	.	.	.	.	.
[pohlavi = 2] * [kouri = 2] * [lokalita = 1]	0[a]	.	.	.	.	.
[pohlavi = 2] * [kouri = 2] * [lokalita = 2]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Model: Poisson

c. Design: Constant + pohlavi + kouri + lokalita + pohlavi * kouri + pohlavi * lokalita + kouri * lokalita + pohlavi * kouri * lokalita

Obrázek 3.8: Výstup s odhady parametrů saturovaného modelu včetně testů významnosti a konfidenčních intervalů.

výpočtu dostaneme stejně jako při obecném modelu výstupy týkající se popisu dat, informaci o konvergenci, pozorované a očekávané počty, a testy dobré shody. Nově je ve výstupu i tabulka obsahující výsledky testů, pokud odstraníme všechny efekty stupně K (spodní polovina tabulky) či všechny efekty stupně K a výše (horní polovina tabulky), tabulku nalezneme na obrázku 3.9.

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and Higher Order Effects[a]	1	7	1822,549	,000	1752,537	,000	0
	2	4	415,502	,000	316,828	,000	2
	3	1	2,825	,093	2,753	,097	8
K-way Effects[b]	1	3	1407,047	,000	1435,709	,000	0
	2	3	412,676	,000	314,075	,000	0
	3	1	2,825	,093	2,753	,097	0

a. Tests that k-way and higher order effects are zero.

b. Tests that k-way effects are zero.

Obrázek 3.9: Výstup obsahující testy významnosti, pokud odstraníme všechny efekty řádu K.

Pro naši trojrozměrnou tabulku je nejvyšší možné K=3 (jedná se o interakci všech proměnných). U nejvyššího možného K budou výsledky testů shodné pro obě části tabulky, u ostatních se bude lišit. Vzhledem k tomu, že začínáme ze saturovaného modelu, budou testy pro nižší K většinou významné. Proto je lepší se zaměřit až na část o zpětné eliminaci (Backward Elimination Statistics). Hned v první tabulce nalezneme výsledky postupného eliminování proměnných. V každém kroku je vždy zobrazen model, ze kterého se vychází (v našem případě se začíná ze saturovaného modelu), a poté efekt, který se algoritmus pokusil odstranit. Vždy se vybírají efekty nejvyšších řádů, v prvním kroku je tedy interakce všech tří proměnných. Pokud je p-hodnota větší než 0.05, je daný efekt odstraněn. V případě, že máme více efektů stejného řádu (např. v kroku 2), vybere se efekt s nejvyšší p-hodnotou, která musí být zároveň vyšší než 0.05. V našem případě není žádná z p-hodnot vyšší a tedy algoritmus končí s modelem homogenní asociace.
V poslední části jsou znovu vidět pozorované a očekávané počty pozorování, včetně testů dobré shody. Přestože bychom mohli si nechat zobrazit koeficienty jednotlivých parametrů při automatickém hledání modelu, SPSS nedělá konzistentní omezující podmínky na parametry a je tedy vždy lepší daný model finálně zkoumat ze základního formuláře pro práci s log-lineárními modely.

3.2.3 Příklady pro konkrétní situace

Nyní si ukážeme analýzu trojrozměrných kontingenčních tabulek pomocí log-lineárních modelů na několika konkrétních příkladech, kde dopředu známe závislost mezi daty. Chceme zde představit situace, které může výzkumník potkat a jak se tyto situace projevují do dat. V této kapitole budeme používat několik různých dat z různých oblastí psychologie. Kompletní data pro jednotlivé případy (včetně výstupů SPSS) jsou na přiloženém médiu, v textu budeme zmiňovat jen důležité výsledky.

Totální nezávislost

V tomto případě pracujeme s daty, kde platí úplná nezávislost mezi proměnnými. Tato umělá data pocházejí z experimentu, ve kterém jsme se ptali lidí, zda podstoupili terapii (nebudeme zmiňovat žádný konkrétní přístup) a zda jim to pomohlo. Data můžeme vidět v tabulce 3.2. Spustíme-li nyní výběr modelu (stejné nastavení jako u ukázkového příkladu v předchozí části), dostaneme tabulku zobrazenou na obrázku 3.10. Výsledný model tedy je ve tvaru $log(\mu_{ijk}) = \lambda +$

pohlaví	terapie	výsledek	počet
muž	ano	ano	10
muž	ano	ne	20
muž	ne	ano	31
muž	ne	ne	59
žena	ano	ano	26
žena	ano	ne	49
žena	ne	ano	75
žena	ne	ne	150

Tabulka 3.2: Data použitá v příkladu 1. Jednotlivé řádky popisují počet pozorování pro dané podmínky. První tři sloupce jsou nominální proměnné, poslední sloupec je intervalová proměnná.

$\lambda_i^{pohlavi} + \lambda_j^{terapie} + \lambda_k^{vysledek}$, což odpovídá modelu úplné nezávislosti. Spustíme-li nyní základní hledání modelu (v sekci výběr modelu přidáme pouze hlavní efekty), dostaneme výstup zobrazený na obrázku 3.11. Pokud spočítáme exponenciálu těchto koeficientů, dostaneme poměry šancí, konkrétně šanci, že se pacient uzdraví, spočítáme jako $e^{-0.672} = 0.51$, tedy je přibližně dvakrát větší šance, že se pacient neuzdraví. Protože máme model kompletní nezávislosti, je tato šance je nezávislá na pohlaví nebo na tom, zda daná osoba podstoupila terapii. Pokud bychom se na tato data podívali pomocí kontingenčních tabulek, dostali bychom přibližně stejné tabulky, kdybychom data agregovali přes libovolnou ze tří proměnných. Pro srovnání přístupu log-lineárních modelů a kontingenčních tabulek můžeme spočítat kontingenční tabulky postupně pro případy, kde bude ve vrstvách pohlaví, terapie a výsledek. Ve všech třech případech budou koeficienty asociace podobné.

Sdružená nezávislost

V dalším příkladu si ukážeme data, na kterých se vztah terapií a výsledkem neliší mezi pohlavím. Z hlediska zápisu jde tedy o model $(TV \perp P)$. Data pro tento případ jsou zobrazena v tabulce 3.3. Spustíme-li hledání modelu, automatický proces se zastaví u sdružené nezávislosti pohlaví na terapii a výsledku, tedy u modelu $log(\mu_{ijk}) = \lambda + \lambda_i^{pohlavi} + \lambda_j^{terapie} + \lambda_k^{vysledek} + \lambda_{ij}^{terapie,vysledek}$. Tento model tedy říká, že pohlaví je nezávislé na kombinacích obou proměnných a tedy můžeme data agregovat přes pohlaví a dostat dvourozměrné tabulky. Pokud si spočítáme kontingenční tabulky pro terapii a výsledek, dostaneme téměř stejný

Step Summary

Step[a]			Effects	Chi-Square[c]	df	Sig.	Number of Iterations
0	Generating Class[b]		pohlavi*terapie*vysledek	,000	0	.	
	Deleted Effect	1	pohlavi*terapie*vysledek	,043	1	,836	2
1	Generating Class[b]		pohlavi*terapie, pohlavi*vysledek, terapie*vysledek	,043	1	,836	
	Deleted Effect	1	pohlavi*terapie	,000	1	1,000	2
		2	pohlavi*vysledek	,010	1	,922	2
		3	terapie*vysledek	,014	1	,905	2
2	Generating Class[b]		pohlavi*vysledek, terapie*vysledek	,043	2	,979	
	Deleted Effect	1	pohlavi*vysledek	,010	1	,922	2
		2	terapie*vysledek	,014	1	,905	2
3	Generating Class[b]		terapie*vysledek, pohlavi	,052	3	,997	
	Deleted Effect	1	terapie*vysledek	,014	1	,905	2
		2	pohlavi	79,697	1	,000	2
4	Generating Class[b]		pohlavi, terapie, vysledek	,067	4	,999	
	Deleted Effect	1	pohlavi	79,697	1	,000	2
		2	terapie	109,882	1	,000	2
		3	vysledek	44,842	1	,000	2
5	Generating Class[b]		pohlavi, terapie, vysledek	,067	4	,999	

a. At each step, the effect with the largest significance level for the Likelihood Ratio Change is deleted, provided the significance level is larger than ,050.

b. Statistics are displayed for the best model at each step after step 0.

c. For 'Deleted Effect', this is the change in the Chi-Square after the effect is deleted from the model.

Obrázek 3.10: Výsledky zpětné eliminace. Postupně byly odstraněné všechny efekty, až zůstaly pouze hlavní efekty.

Parameter Estimates[b,c]

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant	5,003	,073	68,461	,000	4,860	5,147
[pohlavi = 1]	-,916	,108	-8,487	,000	-1,128	-,705
[pohlavi = 2]	0[a]	.	.	.	.	.
[terapie = 1]	-1,099	,113	-9,754	,000	-1,319	-,878
[terapie = 2]	0[a]	.	.	.	.	.
[vysledek = 1]	-,672	,103	-6,515	,000	-,874	-,470
[vysledek = 2]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Model: Poisson

c. Design: Constant + pohlavi + terapie + vysledek

Obrázek 3.11: Odhady jednotlivých parametrů

poměr šancí, jako koeficient interakce v log-lineárním modelu (pouze jej musíme zlogaritmovat).

pohlaví	terapie	výsledek	počet
muž	ano	ano	10
muž	ano	ne	20
muž	ne	ano	42
muž	ne	ne	30
žena	ano	ano	21
žena	ano	ne	41
žena	ne	ano	81
žena	ne	ne	61

Tabulka 3.3: Data použitá v příkladu 2.

Podmíněná nezávislost

V následujícím příkladu máme situaci, ve které jsou proměnné terapie a výsledek podmíněně nezávislé na pohlaví. Znamená to tedy, že pokud se budu na data dívat bez ohledu na pohlaví, dostanu tam závislost mezi terapií a výsledkem, zatímco budu-li analyzovat data pro muže a ženy zvlášť, budu mít mezi proměnnými nezávislost. Data jsou vidět v tabulce 3.4.

pohlaví	terapie	výsledek	počet
muž	ano	dobrý	88
muž	ano	ne	92
muž	ne	ano	7
muž	ne	ne	10
žena	ano	ano	191
žena	ano	ne	11
žena	ne	ano	202
žena	ne	ne	12

Tabulka 3.4: Data použitá v příkladu 3.

Spustíme-li výběr modelu, skončí proces u právě modelu podmíněné nezávislosti, tedy u $log(\mu_{ijk}) = \lambda + \lambda_i^{pohlavi} + \lambda_j^{terapie} + \lambda_k^{vysledek} + \lambda_{ij}^{pohlavi,terapie} + \lambda_{ik}^{pohlavi,vysledek}$. Zkusíme-li se nyní podívat na kontingenční tabulky (proměnná pohlaví bude ve vrstvách) včetně poměru šancí, měli bychom dostat pro muže i ženy poměr šancí roven přibližně 1. V tomto případě dostaneme pro muže $\theta_{pohlavi|muz} = 1.366$ a pro ženy $\theta_{pohlavi|zena} = 1.032$. Toto je přímo ukázka Simpsonova paradoxu - kdybychom analyzovali data bez ohledu na pohlaví, dostali

bychom jiné závěry.

Saturovaný model

V následujícím příkladu si ukážeme data, ve kterých data nejlépe vysvětluje saturovaný model, tedy při zafixování konkrétní úrovně u vybrané proměnné budou dílčí tabulky mít závislost různou. Pro tento příklad používáme data z teoretického experimentu, ve kterém jsme rozdělili pokusné osoby na neurotiky a stabilní (např., dichotomizací NEO-FFI dotazníku) a sledovali jsme, zda při odměně nebo trestu mají dobré nebo špatné pocity. Data jsou k nalezení v tabulce 3.5. V tomto případě automatický proces výběru modelu skončí u saturovaného mo-

ns	motivace	výsledek	počet
neurotik	odměna	dobrý	33
neurotik	odměna	špatný	20
neurotik	trest	dobrý	20
neurotik	trest	špatný	31
stabilní	odměna	dobrý	118
stabilní	odměna	špatný	30
stabilní	trest	dobrý	30
stabilní	trest	špatný	18

Tabulka 3.5: Data použitá v příkladu 4.

delu. Při kontrole jednotlivých koeficientů vidíme, že i trojcestná interakce je významná. Zkusíme-li pro jistotu data vysvětlit modelem homogenní asociace, vidíme, že celková shoda s daty je velmi nízká (všechny odchylky jsou větší než 3), nemůžeme je tedy vysvětlit jednodušším modelem. Jednotlivé asociace mezi proměnnými bude různé pro rozdílné kombinace proměnných a redukované kontingenční tabulky ponesou jinou informaci. Pokud si například necháme spočítat tabulky pro neurotiky a stabilní, vidíme, že asociace mezi proměnnými u neurotiků je kladná $\phi = 0.23$), zatímco u stabilních je záporná $\phi = -0.25$. Obdobné rozdíly vidíme i u poměru věrohodností pro obě pohlaví.

Homogenní asociace

V následujícím příkladu si ukážeme aplikaci log-lineárních modelů na data z dotazníku, ve kterém máme dichotomické proměnné P1, P2 a P3. Na rozdíl od předchozích dat máme tyto data v neagregované formě, tedy jedno pozorování na řádek. Protože máme celkem 339 pozorování, vypíšeme sem pro ilustraci jen

prvních 6 řádek (3.6), celá data budou k dispozici v elektronické příloze k práci. Na rozdíl od předchozích případů nemusíme nastavovat vážení jednotlivých řádků, ale pracujeme přímo s daty V tomto případě skončí automatický výběr modelu[2] u ho-

P1	P2	P3
0	0	0
0	0	1
0	0	0
0	1	0
0	0	0
0	0	0

Tabulka 3.6: Data použitá v příkladu 5. Protože jsou zapsaná ve formátu jedno pozorování na řádek, vypisujeme jen prvních 6 pozorování.

mogenní asociace, tedy u $log(\mu_{ijk}) = \lambda + \lambda_i^{P1} + \lambda_j^{P2} + \lambda_k^{P3} + \lambda_{ij}^{P1,P2} + \lambda_{ik}^{P1,P3} + \lambda_{jk}^{P2,P3}$. Jak již bylo zmíněno dříve, v tomto modelu platí, že pro libovolnou úroveň proměnné (tedy pro zafixovanou hodnotu některé z otázek) bude míra asociace pro ostatní proměnné stejná[3]. Vyzkoušíme-li si ručně data modelovat jednodušími modely, vyjde nám špatná shoda s daty (u jednotlivých buněk budou velká residua). Toto si můžeme ověřit na kontingenčních tabulkách. Necháme-li spočítat tabulky pro proměnné P2 a P3, pro případy, že dotazovaný odpověděl v otázce P1 hodnotu 0 nebo 1, dostaneme pro obě tabulky podobné asociace a poměry šancí ($\phi_{P1=0} = 0.238, \theta_{P1=0} = 3.249$ a $\phi_{P1=1} = 0.253, \theta_{P1=1} = 3.056$).

Trojrozměrná tabulka s více než dvěma úrovněmi

Jako poslední příklad popsaný v této práci jsme vybrali data o přijatých studentech na univerzitu v Berkeley (Bickel, Hammel & O'Connell, 1975). O přijatých osobách bylo sbírána informace o pohlaví, číslo oddělení, kam se hlásili, a informace o přijetí/odmítnutí. Data jsou k vidění v tabulce 3.7. Stejně jako v předchozích příkladech budeme nejprve hledat vhodný model. V tomto případě musíme pouze u proměnné oddělení zadat rozsah 1-6, protože máme celkem 6 oddělení. Proces hledání modelu skončí hned u saturovaného modelu, nemůžeme tedy odebrat trojcestnou interakci mezi proměnnými. Necháme-li spočítat parametry tohoto modelu, dostaneme výstup zobrazený na obrázku 3.8 Protože SPSS standardně vybírá parametry postupně od jednodušších ke složitějším a dále je počítá vze-

[2]Pozor, protože v tomto případě máme hodnoty kódované jako 0 a 1, musíme u automatického výběru modelu nastavit tento rozsah namísto 1 a 2, což jsme tam zadávali v předchozích

oddělení	pohlaví	výsledek	počet
A	muž	přijat	512
A	muž	odmítnut	313
A	žena	přijat	89
A	žena	odmítnut	19
B	muž	přijat	353
B	muž	odmítnut	207
B	žena	přijat	17
B	žena	odmítnut	8
C	muž	přijat	120
C	muž	odmítnut	205
C	žena	přijat	202
C	žena	odmítnut	391
D	muž	přijat	138
D	muž	odmítnut	279
D	žena	přijat	131
D	žena	odmítnut	244
E	muž	přijat	53
E	muž	odmítnut	138
E	žena	přijat	94
E	žena	odmítnut	299
F	muž	přijat	22
F	muž	odmítnut	351
F	žena	přijat	24
F	žena	odmítnut	317

Tabulka 3.7: Data použitá v příkladu 6.

stupně, budou v tomto případě parametry pro oddělení 6 redundantní. Tedy při interpretaci interakce pohlaví a přijetí budeme pracovat s oddělením 6. V tomto případě je poměr šancí pro přijetí muže nebo ženy do oddělení 6 přibližně stejný ($e^{-0.187} = 0.83$, $p = .536$). Pro poměry šancí pro jiná oddělení musíme sečíst parametr pro interakci pohlaví a přijetí s parametrem označující trojcestnou interakci pohlaví a přijetí pro konkrétní oddělení. Kdyby nás zajímalo např. oddělení 3, dostaneme poměr šancí $e^{-0.187+0.312} = 1.13$. Pro testování významnosti

případech.

[3]U podmíněné asociace by byla asociace nulová

Parameter	Odhad parametru	p-hodnota
Constant	5.760	0.000
oddeleni = 1	-2.790	0.000
oddeleni = 2	-3.620	0.000
oddeleni = 3	0.210	0.006
oddeleni = 4	-0.261	0.002
oddeleni = 5	-0.058	0.469
pohlavi = 1	0.102	0.189
vysledek = 1	-2.562	0.000
oddeleni = 1 * pohlavi = 1	2.676	0.000
oddeleni = 2 * pohlavi = 1	3.093	0.000
oddeleni = 3 * pohlavi = 1	-0.746	0.000
oddeleni = 4 * pohlavi = 1	0.032	0.784
oddeleni = 5 * pohlavi = 1	-0.873	0.000
oddeleni = 1 * vysledek = 1	4.086	0.000
oddeleni = 2 * vysledek = 1	3.284	0.000
oddeleni = 3 * vysledek = 1	1.903	0.000
oddeleni = 4 * vysledek = 1	1.942	0.000
oddeleni = 5 * vysledek = 1	1.408	0.000
pohlavi = 1 * vysledek = 1	-0.187	0.536
oddeleni = 1 * pohlavi = 1 * vysledek = 1	-0.845	0.034
oddeleni = 2 * pohlavi = 1 * vysledek = 1	-0.002	0.996
oddeleni = 3 * pohlavi = 1 * vysledek = 1	0.312	0.350
oddeleni = 4 * pohlavi = 1 * vysledek = 1	0.105	0.756
oddeleni = 5 * pohlavi = 1 * vysledek = 1	0.389	0.282

Tabulka 3.8: Parametry modelu pro příklad 6. Vzhledem k množství kategorií u proměnné oddělení je počet parametrů velmi velký (celkem 63 parametrů), přepsali jsme pro větší čitelnost pouze neredundantní parametry. V tabulce jsou zobrazeny pouze odhady a parametru a signifikance.

se díváme na p-hodnotu u trojcestné interakce daného oddělení, v tomto případě opět nevýznamný ($p = 0.350$). U ostatních oddělení je rozdíl taktéž nevýznamný, pouze u oddělení 1 je tento rozdíl významný ($e^{-0.187-0.845} = 0.36$, $p = 0.034$), tedy je přibližně třikrát větší šance, že do oddělení 1 bude přijat muž, než že bude přijata žena. Pokud bychom si zkusili odebrat oddělení 1 z dat a spustit na tom opětovně hledání modelu, skončili bychom u podmíněné nezávislosti

přijetí na pohlaví. Toto je další případ Simpsonova paradoxu, protože pokud zanedbáme oddělení a podíváme se pouze na vztah pohlaví a přijetí, dostaneme silnou závislost ve prospěch mužů.

3.2.4 Omezení analýzy log-lineárních modelů v SPSS

Ačkoli je práce s log-lineárními modely v SPSS přímočará, existuje zde mnoho omezení, které mohou zkreslit interpretaci. Hlavní nedostatek je používání p-hodnoty pro výběr modelu. Jak již bylo zmíněno v sekci 3.2.2, je lepší použít pro výběr modelu jiná kritéria, např. AIC nebo BIC, které berou v úvahu i složitost modelu. V SPSS můžeme AIC spočítat ručně z dat z výstupu, případně upravit přímo kódy příkazů, ale v tomto případě se ztrácí jednoduchost používání tohoto programu a je lepší přejít k silnějším statistickým nástrojům. Dále není v SPSS možné porovnávat hierarchické modely jinak než v automatickém hledání modelu. V mnoha případech je lepší porovnávat přímo konkrétní modely a rozhodování o nejvhodnějším modelu založit na více kritériích, v SPSS můžeme pouze zkoumat celkovou shodu modelu s daty. Obdobně chybí možnost spočítat indexy, které pomáhají pro rozhodování při výběru modelu. Při velkých datech vychází i drobná odchylka od modelu významně a hledání vhodného modelu často končí u saturovaného, který je obtížný na interpretaci a navíc nepoukazuje na zajímavé vztahy mezi proměnnými. V případě velkých dat si opět musíme tyto indexy spočítat ručně, případně se obrátit na silnější statistický nástroj.

3.2.5 Analýza čtyřrozměrných tabulek v SPSS

Při analýze čtyřrozměrných tabulek postupujeme stejně jako při trojrozměrných tabulkách. Začneme tedy ze saturovaného modelu, který obsahuje všechny 4 proměnné, spustíme proces hledání modelu, a poté spustíme hledání parametrů pomocí základního formuláře pro práci s log-lineárními modely. U vícerozměrných tabulek nám exponenciálně roste počet proměnných, proto bude většinou hledání vhodného modelu důležitější, než interpretace signifikance jednotlivých parametrů.

3.3 Analýza pomocí R

Alternativou k analýze dat pomocí SPSS je statistický program R. Tento program má otevřený kód a modulární architekturu, tedy každý uživatel si při práci s konkrétními daty může načíst specializované balíčky. Pro základní analýzu

kontingenčních tabulek pomocí log-lineárních modelů tabulkami nejsou potřeba specializované balíčky, pro výpočet asociačních koeficientů je třeba balíček `vcd` (Meyer, Zeileis & Hornik, 2015), konkrétně příkaz `assoc`. Protože samotná znalost jazyka R převyšuje míru odbornost této práce, popíšeme jen základní analýzu a interpretaci výše příkladu uvedeného na začátku této kapitoly (3.1), základy jazyka jsou popsány v knize Field et al. (2012).

V jazyku R máme dvě možnosti, jak analyzovat kontingenční tabulky. Jedna je pomocí příkazu `loglin()`, který pracuje přímo s kontingenčními tabulkami (ty získáme příkazem `table()`). Druhou možností je použít přímo zobecněný lineární model s Poissonovskou transformační funkcí (funkce `glm(, family = poisson())`). Tento způsob pracuje přímo s objektem `data.frame`, který je nativní v R a odpovídá tabulkám z SPSS. My se zaměříme pouze na přístup přes zobecněné lineární modely, neboť interpretace parametrů je konzistentní s výstupy dalších metod (např. analýza rozptylu). Data uložená ve formátu csv (tento soubor je na přiloženém médiu) nejprve načteme příkazem `read.csv2("data.csv")`[4]. Protože R umí pracovat přímo s textovými úrovněmi, nemusíme kódovat pomocí čísel a výstupy bývají čitelnější než v SPSS.

Pro výběr proměnných, které mají tvořit kontingenční tabulku použijeme zápis formule. Formule je v R ve tvaru $Y \sim X1 + X2 + X3$, kde Y je závislá proměnná a proměnné $X1$, $X2$ a $X3$ jsou proměnné, které tvoří kontingenční tabulku. První proměnná tvoří řádky kontingenční tabulky, druhý řádky a zbývající proměnné tvoří vrstvy (obdobně jako v SPSS). Pro náš případ můžeme psát `xtabs(pocet pohlavi + kouri + lokalita, data = df)` pro zobrazení tabulek pro jednotlivé lokality. Obdobně si zobrazíme i ostatní kontingenční tabulky. Nyní necháme data modelovat pomocí saturovaného log-lineárního modelu. Toho docílíme dvojicí příkazů 3.1. Hvězdičky mezi proměnnými značí, že do modelu mají být zahr-

```
loglin.sat <- glm(pocet ~ pohlavi * kouri * lokalita, data = df,
 family = poisson())
summary(loglin.sat)
```

Blok kódu 3.1: Hledání parametrů saturovaného modelu.

nuté obě proměnné včetně jejich interakce. Kdyby nás zajímala pouze interakce mezi proměnnými, zapsali bychom ji do formule pomocí dvojtečky, tedy pro interakci proměnné pohlaví a kouří bychom napsali `pohlavi : kouri`.Výstup tohoto příkazu je vidět na výstupu 3.2. Na začátku bloku je použitá formule, tedy sa-

[4]Pokud bychom měli data oddělená čárkou, použijeme příkaz `read.csv("data.csv")`.

```
> summary(loglin.sat)

Call:
glm(formula = pocet ~ pohlavi * kouri * lokalita,
    family = poisson(), data = df)

Deviance Residuals:
[1]  0  0  0  0  0  0  0  0

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)             2.2947     3.3274   0.690  0.49042
pohlavi                -7.1260     2.4053  -2.963  0.00305 **
kouri                   1.0782     1.7738   0.608  0.54327
lokalita                1.9543     1.7409   1.123  0.26162
pohlavi:kouri           3.7431     1.2448   3.007  0.00264 **
pohlavi:lokalita        3.5310     1.2320   2.866  0.00416 **
kouri:lokalita         -2.1084     0.9776  -2.157  0.03103 *
pohlavi:kouri:lokalita -1.0644     0.6572  -1.620  0.10533
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1.8225e+03  on 7  degrees of freedom
Residual deviance: 6.9722e-14  on 0  degrees of freedom
AIC: 64.256

Number of Fisher Scoring iterations: 3
```

Blok kódu 3.2: Výstup analýzy saturovaného modelu na ukázkových datech.

turovaný model. Pod tím jsou vidět odchylky pro jednotlivé buňky. Protože jde o saturovaný model, jsou tyto odchylky nulové. V další části vidíme jednotlivé

koeficienty včetně jejich významnosti [5]. Vidíme, že trojcestná interakce (na posledním řádku) není významná, naznačuje nám to tedy, že bychom mohli zkusit model homogenní asociace. V poslední části je zobrazena celková odchylka vůči nulovému modelu (jde tedy o klasický test dobré shody), celková odchylka residuí (protože jde o saturovaný model, je tato hodnota nulová) a hodnota AIC.

```
loglin.homass <- glm(pocet ~ pohlavi * kouri +
    pohlavi * lokalita + kouri * lokalita,
    data = df, family = poisson())
summary(loglin.homass)
```

Blok kódu 3.3: Hledání parametrů modelu homogenní asociace.

Podíváme-li se nyní pomocí příkazu 3.3 na model homogenní asociace, zjistíme, že nyní jsou všechny parametry významné a navíc zbývající odchylka je velmi malá (2.829). Na rozdíl od SPSS můžeme v R porovnávat hierarchické modely testem poměru věrohodností pomocí příkazu `anova(loglin.sat, loglin.homass, test = "LRT")`, čímž zjistíme, zda parametry ve složitějším modelu významně zlepšují shodu s daty. Na výstupu 3.5 vidíme, že v tomto případě není rozdíl mezi modely významný ($p = 0.093$), můžeme tedy vybrat jednodušší model homogenní asociace. Pokud bychom modely porovnávaly přes AIC, je rozdíl velmi malý ($AIC_{sat} = 64.256$ a $AIC_{hom\_assoc} = 65.078$) a tedy pro praktické použití je lepší použít jednodušší model. V jazyce R můžeme hledat nejlepší model automaticky pomocí příkazu `step(loglin.sat)`. Tento proces na rozdíl od SPSS vybírá model na základě minimalizace AIC, což by v tomto případě vedlo pouze k saturovanému modelu. Je tedy lepší vždy porovnávat modely ručně na základě více kritérií.

Analýzy kontingenčních tabulek v R bývají ještě komplexnější a spoustu témat jsme sem nezahrnuli (analýzu reziduí, mozaikový graf), neboť přesahují rozsah této práce. Dobrým zdrojem pro více informací o analýze kontingenčních tabulek v jazyce R je kniha od Kateri (2014).

[5] Významnost se označuje hvězdičkami v posledním sloupci. Legenda, kolik hvězdiček označuje konkrétní hladinu významnosti je pod tabulkou.

```
> summary(loglin.homass)

Call:
glm(formula = pocet ~ pohlavi * kouri + pohlavi * lokalita +
    kouri * lokalita, family = poisson(), data = df)

Deviance Residuals:
      1       2       3       4       5       6       7       8
 0.9520 -0.3703 -1.1343  0.1565 -0.2679  0.6048  0.1247 -0.1188

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.9220     1.3966  -2.092   0.0364 *
pohlavi           -3.4639     0.6914  -5.010 5.45e-07 ***
kouri              3.9318     0.6113   6.432 1.26e-10 ***
lokalita           4.7558     0.6081   7.820 5.26e-15 ***
pohlavi:kouri      1.7837     0.2324   7.674 1.67e-14 ***
pohlavi:lokalita   1.5991     0.2453   6.520 7.03e-11 ***
kouri:lokalita    -3.6989     0.3014 -12.272  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1822.5488  on 7  degrees of freedom
Residual deviance:    2.8219  on 1  degrees of freedom
AIC: 65.078

Number of Fisher Scoring iterations: 4
```

Blok kódu 3.4: Výstup analýzy modelu homogenní asociace na ukázkových datech.

```
> anova(loglin.sat, loglin.homass, test = "LRT")
Analysis of Deviance Table

Model 1: pocet ~ pohlavi * kouri * lokalita
Model 2: pocet ~ pohlavi * kouri + pohlavi * lokalita +
                 kouri * lokalita
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         0     0.0000
2         1     2.8219 -1  -2.8219  0.09298 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Blok kódu 3.5: Porovnání saturovaného modelu a modelu homogenní asociace. V tomto případě není rozdíl mezi modely významný, můžeme tedy použít jednodušší model.

Závěr

V této práci jsme představili téma práce s vícerozměrnými daty, konkrétně s vícerozměrnými kontingenčními tabulkami. Analýza vícerozměrných kontingenčních tabulek má svá specifika a při redukci na dvourozměrné případy může docházet ke ztrátě informace a falešným závěrům. Proto jsme v práci vysvětlili logaritmicko-lineární modely, které zobecňují analýzu kontingenčních tabulek do libovolného rozměru. Aplikaci tohoto nástroje jsme ukázali na několika příkladech s různými formy vztahů mezi proměnnými, které mohou v reálných případech nastat. Ačkoli má analýza v SPSS svá omezení, pro mnoho psychologů je to snadnější nástroj na zvládnutí a tím pádem je větší šance, že se rozšíří povědomí o správném přístupu k analýze kontingenčních tabulek. Věříme, že tento text může být použit jako doplňkový materiál pro kurzy analýzy statistických dat s podporou SPSS. Pro programátorsky zdatnější čtenáře jsme přidali i stručnou kapitolu o analýze kontingenčních tabulek v nástroji R. Tento nástroj je aktivně vyvíjen a reaguje na nové poznatky ze statistické teorie, takže pomocí tohoto nástroje můžeme dosáhnout přesnějších závěrů.

Reference

Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). Hoboken, NJ, USA: John Wiley & Sons, Inc.

Aiken, L. S. & West, S. G. (1991). *Multiple Regression: Testing and Interpreting Interactions* (S. McElroy, Ed.). Thousand Oaks: SAGE Publications, Inc.

Barnard, G. A. (1947). Significance Tests for 2x2 Tables. *Biometrika*, *34*, 123–138. doi: 10.1093/biomet/34.1-2.123

Bickel, P. J., Hammel, E. a. & O'Connell, J. W. (1975). Sex Bias in Graduate Admissions: Data from Berkeley. *Science*, *187*(4175), 398–404. doi: 10.1126/science.187.4175.398

Bishop, Y. M., Fienberg, S. E. & W, H. P. (2007). Maximum Likelihood Estimation For Incomplete Tables. V *Discrete multivariate analysis: Theory and practice* (pp. 177–228). New York: Springer. doi: 10.1007/978-0-387-72806-3_5

Burnham, K. P. (2004, November). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, *33*(2), 261–304. doi: 10.1177/0049124104268644

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge.

Cook, R. D. & Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.

Darroch, J. N., Lauritzen, S. L. & Speed, T. P. (1980). Markov Fields and Log-Linear Interaction Models for Contingency Tables. *The Annals of Statistics*, *8*(3), 522–539. doi: 10.1214/aos/1176345006

Edwards, D. (2012). *Introduction to graphical modelling* (2nd ed.). Springer Science & Business Media.

Ekström, J. (2011). The Phi-coefficient, the Tetrachoric Correlation Coefficient, and the Pearson-Yule Debate. *Department of Statistics, UCLA*, 1–19.

Field, A. (2013). *Discovering Statistics using IBM SPSS Statistics*. SAGE Publications.

Field, A., Miles, J. & Field, Z. (2012). *Discovering Statistics Using R*. Thousand Oaks: SAGE Publications Inc.

Fisher, R. (1922). On the Interpretation of $\chi 2$ from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, *85*(1), 87–94. doi: 10.2307/2340521

Galmarini, M. V., Symoneaux, R., Chollet, S. & Zamora, M. C. (2013). Under-

standing Apple Consumers' Expectations in Terms of Likes and Dislikes. Use of Comment Analysis in a Cross-Cultural Study. *Appetite*, *62*, 27–36. doi: 10.1016/j.appet.2012.11.006

Hendl, J. (2006). *Přehled Statistických Metod Zpracování Dat: Analýza a Meta-analýza Dat* (3rd ed.). Praha: Portál.

Kateri, M. (2014). *Contingency Table Analysis: Methods and Implementation Using R.* New York, NY: Springer New York. doi: 10.1007/978-0-8176-4811-4

Kruschke, K. J. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* (2nd ed.). Academic Press.

Kuha, J. & Firth, D. (2011). On the index of dissimilarity for lack of fit in loglinear and log-multiplicative models. *Computational Statistics and Data Analysis*, *55*(1), 375–388. doi: 10.1016/j.csda.2010.05.005

Lin, J.-J., Chang, C.-H. & Pal, N. (2014). A Revisit to Contingency Table and Tests of Independence: Bootstrap is Preferred to Chi-Square Approximations as Well as Fisher's Exact Test. *Journal of Biopharmaceutical Statistics*, *25*(3), 438–458. doi: 10.1080/10543406.2014.920851

MacCallum, R. C., Zhang, S., Preacher, K. J. & Rucker, D. D. (2002). On the Practice of Dichotomization of Quantitative Variables. *Psychological methods*, *7*(1), 19–40. doi: 10.1037/1082-989X.7.1.19

Maxwell, S. E. & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, *113*(1), 181–190. doi: 10.1037/0033-2909.113.1.181

McDonald, John, H. (2014). *Handbook of biological statistics* (3rd ed.). Baltimore, USA: Sparky House Publishing.

Meyer, D., Zeileis, A. & Hornik, K. (2015). vcd: Visualizing Categorical Data [Manuál počítačového program].

Miller, B. A. J. (1984). Selection of Subsets of Regression Variables. *Journal of the Royal Statistical Society. Series A (General)*, *147*(3), 389–425. doi: 10.2307/2981576

Olsson, U. (1979). Maximum Likelihood Estimation of the Polychoric Correlation Coefficient. *Psychometrika*, *44*(4), 443–460. doi: 10.1007/BF02296207

Quine, M. P. & Robinson, J. (1985, June). Efficiencies of Chi-Square and Likelihood Ratio Goodness-of-Fit Tests. *The Annals of Statistics*, *13*(2), 727–742. doi: 10.1214/aos/1176349550

Raftery, A. E. (2013). Bayesian Model Selection in Social Research. *Sociological Methodology*, *25*(1995), 111–163. doi: 10.2307/271063

Rutherford, A. (2001). *Introducing ANOVA and ANCOVA: a GLM Approach.* SAGE Publications.

Schwarz, G. (1978). *Estimating the Dimension of a Model* (Vol. 6) (No. 2). doi: 10.1214/aos/1176344136

Sharpe, D. (2013). Why the resistance to statistical innovations? Bridging the communication gap. *Psychological methods*, *18*(4), 572–82. doi: 10.1037/a0034177

Simpson, E. H. (1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *13*(2), 238–241.

Stroup, W. W. (2012). *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications.* CRC Press.

Warrens, M. J. (2008). On Association Coefficients for 2 x 2 Tables and Properties that do not depend on the Marginal Distributions. *Psychometrika*, *73*, 777–789. doi: 10.1007/s11336-008-9070-3

Wilcox, R. R. (2012). *Introduction to Robust Estimation and Hypothesis Testing* (3rd ed.). Academic Press.

Yates, F. (1934). Contingency tables involving small numbers and the chiˆ2 test. *Supplement to the Journal of the Royal Statistical Society*, 217–235.