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

BAKALÁŘSKÁ PRÁCE

TESTOVÁNÍ HYPOTÉZ V KONTINGENČNÍCH TABULKÁCH

HYPOTHESIS TESTING IN CONTINGENCY TABLES

Vedoucí bakalářské práce:
RNDr. et PhDr. Ivo Müller Ph.D.

Rok odevzdání: 2015

Vypracovala:
Sára Mihalová
MATEKO, IV. ročník

PROHLÁŠENÍ

Prohlašuji, že jsem práci vypracovala samostatně pod vedením RNDr. et PhDr. Ivo Müller Ph.D. a uvedla v seznamu literatury všechny použité zdroje.

V Olomouci dne 7. 5. 2015 Podpis: .....

PODĚKOVÁNÍ

Děkuji RNDr. et PhDr. Ivo Müller Ph.D. za cenné rady, připomínky a odborné vedení mé bakalářské práce.

V Olomouci dne 7. 5. 2015 Podpis: .....

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Sára Mihalová

Název práce: Testování hypotéz v kontingenčních tabulkách

Typ práce: Bakalářská práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: RNDr. et PhDr. Ivo Müller Ph.D.

Rok obhajoby práce: 2015

Abstrakt: Tématem bakalářské práce je testování hypotéz v kontingenčních tabulkách. Jelikož testy v kontingenčních tabulkách vycházejí z testů dobré shody, je jim věnována úvodní kapitola. V dalších částech práce jsou na vlastních datech, která byla získána různými metodami, ať už formou dotazníku, tak vlastním měřením, provedeny vybrané testy. Postupně test nezávislosti, test homogenity, McNemarův test a v poslední řadě také Fisherův faktoriálový test. Cílem práce je shrnutí nejpoužívanějších testů v kontingenčních tabulkách a jejich následná aplikace.

Klíčová slova: Fisherův faktoriálový test, homogenita, kontingenční tabulky, McNemarův test, multinomické rozdělení, nezávislost, testy dobré shody

Počet stran: 70

Počet příloh: 7

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Sára Mihalová

Title: Hypothesis testing in Contingency tables

Type of thesis: Bachelor's

Department: Department of Mathematical Analysis and Applications of Mathematics

Supervisor: RNDr. Dr. et. Ivo Müller Ph.D.

The year of presentation: 2015

Abstract: Topic of the bachelor's thesis is hypothesis testing in contingency tables. Since the tests in contingency tables are based on tests of goodness of fit, they are given in introductory chapter. In other parts of the thesis, there are performed selected tests on own data, which was obtained by various methods, whether in the form of a questionnaire, so the actual measurement. Gradually test of independence, homogeneity test, McNemar's test and the last also Fisher's exaxt test. The aim of this thesis is to summary the most used tests in contingency tables and their subsequent application.

Keywords: contingency tables, Fisher's exact test, homogeneity, independence, McNemar's test, multinomial distribution, tests of goodness of fit

Number of pages: 70

Number of appendices: 7

Language: Czech

OBSAH

SEZNAM GRAFŮ

SEZNAM TABULEK

ÚVOD

Tématem této bakalářské práce je testování hypotéz v kontingenčních tabulkách. Pro tyto účely byly různými způsoby, ať už dotazníkem, pověřením záznamu dat blízkého člověka, či vlastním měřením získána data, se kterými se může kdokoliv setkat v běžném životě. Na těchto datech, shrnutých v kontingenčních tabulkách, jsou následně testovány různé hypotézy.

Testování v kontingenčních tabulkách vychází z testů dobré shody, kterým je věnována první kapitola práce. Základem pro takové testování je multinomické rozdělení vycházející z rozdělení binomického. Tato část práce obsahuje mimo základní myšlenku a charakteristiky také princip testování při známých či neznámých parametrech. Další kapitoly, v nichž jsou uvedeny jednotlivé typy testů v kontingenčních tabulkách, jsou vždy v úvodu doprovázeny teoretickými východisky těchto metod.

Nosným pilířem celé práce jsou praktické příklady, týkající se do jisté míry mé osoby, využití testování v tabulkách. Témata jsou volena tak, aby byla pro čtenáře zajímavými a některá tak, aby si z nich mohl něco do života vzít. Pro budoucí rodiče například může být zajímavá závislost kouření rodičů a jejich dětí, popřípadě závislost kouření marihuany u jejich dětí a okolí, ve kterém se pohybují. Pro všechny, kteří chodí častěji popř. pravidelně nakupovat, myšleno především potraviny, bude jistě využitelné testování, zda nákup ovlivňuje míra sytosti člověka, a jestli náhodou hladový člověk nenakupuje více zbytečných potravin, které původně nezamýšlel koupit. V dnešní době je celkem aktuálním trendem tetování a toho se týká testování, zda má potetovaný uchazeč stejnou šanci k přijetí do nového zaměstnání jako člověk nepotetovaný.

Práce je doprovázena řešením v programu Excel, za použití jednoduchých funkcí pro výpočet chí-kvadrát testu a $p - value$, ve kterém jsou vytvořeny i grafické zpracování naměřených datových souborů. Pro názornost je ukázáno, jak ovlivňuje rozsah výběru testovací statistiku a následné rozhodnutí o nulové hypotéze a změnu hodnoty $p - value$.

Cílem této bakalářské práce je shrnutí různých používaných testů v kontingenčních tabulkách a jejich následná aplikace na vlastních nasbíraných datových souborech.

Práce by měla sloužit jako informační zdroj či k prohloubení popř. lepšímu pochopení látky části pravděpodobnosti a statistiky pro studenty zajímavější formou díky použitým příkladům ze života.

1 TESTY DOBRÉ SHODY

Při řešení statistických úloh v různých oblastech života se často zkoumá, zda naměřená data, data získaná experimentem nebo z historických zkušeností jsou ve shodě s předpokládanou strukturou či pravděpodobnostním modelem. Testy o parametrech multinomického rozdělení pravděpodobností, které je jedním z nejdůležitějších mnohorozměrných rozdělení, se používají při ověřování shody skutečných a očekávaných četností jednotlivých tříd nebo ověřování, zda má náhodná veličina určité, předem dané, rozdělení pravděpodobnosti, např. normální nebo Poissonovo rozdělení.

1.1 MULTINOMICKÉ ROZDĚLENÍ

Mějme osudí a v něm kuličky k různých barev, přičemž $k \geq 2$. Pravděpodobnost vytažení kuličky j-té barvy je rovna p_j, $j = 1, \ldots, k$, a platí $p_j > 0, p_1 + \cdots + p_k = 1$. Provedeme-li n-krát náhodný výběr, vždy po jedné kuličce s následným vracením zpět do osudí. Symbolem X_i označíme počet kuliček j-té barvy, které jsme vybrali v n pokusech. Vzniklý náhodný vektor $\boldsymbol{X} = (X_1, \ldots, X_k)'$ má multinomické rozdělení s parametry $p_1, \ldots, p_k, n$.

Toto rozdělení se značí symbolem $M(p_1, \ldots, p_k, n)$ a je dáno vzorcem

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1!\, x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

pro $x_j = 0, 1, \ldots, n$ $(j = 1, \ldots, k)$, $x_1 + x_2 + \cdots + x_k = n$, jinak je pravděpodobnost rovna 0. [1]

Jednotlivé náhodné veličiny X_j mají binomické rozdělení s parametry p_j, n, tj. $X_j \sim Bi(n, p_j)$, a pro číselné charakteristiky platí

$$E(X_j) = np_j,$$
$$var(X_j) = np_j(1 - p_j).$$

Závislost dvou náhodných veličin vektoru $\boldsymbol{X}$, tedy X_i a X_j, se vyjadřuje pomocí kovariance. Má-li náhodný vektor $\boldsymbol{X}$ multinomické rozdělení s parametry $p_1, \ldots, p_k, n$, platí $X_i + X_j \sim Bi(n, p_i + p_j)$, $i, j = 1, \ldots, k, \forall i \neq j$.

Za použití vztahu pro výpočet rozptylu součtu dvou náhodných veličin

$$var(X + Y) = var(X) + var(Y) + 2cov(X, Y)$$

platí

$$var(X_i + X_j) = n(p_i + p_j)\left(1 - (p_i + p_j)\right) =$$
$$= np_i(1 - p_i) + np_j(1 - p_j) + 2cov(X_i, X_j).$$

Odtud pak vyjádřeníme kovarianci

$$cov(X_i, X_j) = \frac{1}{2}\left\{n(p_i + p_j)\left(1 - (p_i + p_j)\right) - np_i(1 - p_i) - np_j(1 - p_j)\right\} =$$
$$= \frac{1}{2}(np_i - np_i^2 - np_ip_j + np_j - np_ip_j - np_j^2 - np_i + np_i^2 - np_j + np_j^2) =$$
$$= \frac{1}{2}(-2np_ip_j) = -np_ip_j.$$

Nechť má vektor $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)'$ multinomické rozdělení $M(p_1, p_2, \ldots, p_k, n)$. Pak varianční matice má tvar

$$\boldsymbol{V} = var\boldsymbol{X} = \begin{pmatrix} np_1(1 - p_1) & \cdots & -p_1p_k \\ \vdots & \ddots & \vdots \\ -np_kp_1 & \cdots & np_k(1 - p_k) \end{pmatrix}$$

a ukážeme dále, že její hodnost je rovna $k - 1$.

Pro další výpočty označíme

$$\boldsymbol{D} = diag\left\{\sqrt{np_1}, \cdots, \sqrt{np_k}\right\},$$
$$\boldsymbol{p} = \left(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_k}\right)'$$
$$\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{pp}'.$$

Prostudujme vlastnosti matice $\boldsymbol{Q}$:

$$\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{pp'} = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix} - \begin{pmatrix} \sqrt{p_1} \\ \sqrt{p_2} \\ \vdots \\ \sqrt{p_k} \end{pmatrix} \left(\sqrt{p_1}\, \sqrt{p_2} \cdots \sqrt{p_k}\right) =$$

$$= \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix} - \begin{pmatrix} \sqrt{p_1 p_1} & \cdots & \sqrt{p_1 p_k} \\ \vdots & \ddots & \vdots \\ \sqrt{p_k p_1} & \cdots & \sqrt{p_k p_k} \end{pmatrix} =$$

$$= \begin{pmatrix} 1 - p_1 & \cdots & -\sqrt{p_1 p_k} \\ \vdots & \ddots & \vdots \\ -\sqrt{p_k p_1} & \cdots & 1 - p_k \end{pmatrix}. \qquad (1)$$

$$= \begin{pmatrix} 1 - p_1 & \cdots & -\sqrt{p_1 p_k} \\ \vdots & \ddots & \vdots \\ -\sqrt{p_k p_1} & \cdots & 1 - p_k \end{pmatrix}.$$

Matice **Q** je idempotentní, tj. $\boldsymbol{Q}^2 = \boldsymbol{Q}$**.** Dle obecného tvrzení (viz [2], str. 321) víme, že hodnost idempotentní matice se rovná její stopě Tr. Odtud plyne

$$h(\boldsymbol{Q}) = Tr\, \boldsymbol{Q} = Tr\, \boldsymbol{I} - Tr\, \boldsymbol{pp'} = k - 1.$$

V následujících krocích ověříme, že je matice **Q** idempotentní a dále že $Tr\, \boldsymbol{pp'} = 1$.

$$\boldsymbol{Q^2} = \boldsymbol{QQ} = (\boldsymbol{I} - \boldsymbol{pp'})(\boldsymbol{I} - \boldsymbol{pp'}) =$$

$$= \begin{pmatrix} 1 - p_1 & \cdots & -\sqrt{p_1 p_k} \\ \vdots & \ddots & \vdots \\ -\sqrt{p_k p_1} & \cdots & 1 - p_k \end{pmatrix} \begin{pmatrix} 1 - p_1 & \cdots & -\sqrt{p_1 p_k} \\ \vdots & \ddots & \vdots \\ -\sqrt{p_k p_1} & \cdots & 1 - p_k \end{pmatrix}.$$

Po roznásobení je prvek v prvním řádku a prvním sloupci matice $\boldsymbol{Q^2}$, za použití podmínky $p_1 + p_2 + \cdots + p_k = 1$, roven

$$q_{11} = (1 - p_1)^2 + \left(-\sqrt{p_1 p_2}\right)\left(-\sqrt{p_2 p_1}\right) + \left(-\sqrt{p_1 p_3}\right)\left(-\sqrt{p_3 p_1}\right) +$$

$$+ \cdots + \left(-\sqrt{p_1 p_k}\right)\left(-\sqrt{p_k p_1}\right) =$$

$$= (1 - p_1)^2 + \left(-\sqrt{p_1 p_2}\right)^2 + \left(-\sqrt{p_1 p_3}\right)^2 + \cdots + \left(-\sqrt{p_1 p_k}\right)^2 =$$

$$= 1 - 2p_1 + {p_1}^2 + p_1 p_2 + p_1 p_3 + \cdots + p_1 p_k =$$

$$= 1 - 2p_1 + p_1(p_1 + p_2 + \cdots + p_k) = 1 - 2p_1 + p_1 = 1 - p_1.$$

Prvek k-tého řádku prvního sloupce

$$q_{k1} = \left(-\sqrt{p_k p_1}\right)(1 - p_1) + \left(-\sqrt{p_k p_2}\right)\left(-\sqrt{p_2 p_1}\right) + \left(-\sqrt{p_k p_3}\right)\left(-\sqrt{p_3 p_1}\right) +$$
$$+ \cdots + (1 - p_k)\left(-\sqrt{p_k p_1}\right) =$$
$$= -\sqrt{p_k p_1} + p_1\left(-\sqrt{p_k p_1}\right) + p_2\left(-\sqrt{p_k p_1}\right) + p_3\left(-\sqrt{p_k p_1}\right) + \cdots - \sqrt{p_k p_1} +$$
$$+ p_k\left(-\sqrt{p_k p_1}\right) = -\sqrt{p_k p_1} + \sqrt{p_k p_1}(p_1 + p_2 + \cdots + p_k) - \sqrt{p_k p_1} =$$
$$= -\sqrt{p_k p_1} + \sqrt{p_k p_1} - \sqrt{p_k p_1} = -\sqrt{p_k p_1}.$$

Obdobně dostaneme i ostatní prvky, které jsou shodné s prvky matice $\boldsymbol{Q}$.

Stopa matice

$$\begin{pmatrix} \sqrt{p_1 p_1} & \cdots & \sqrt{p_1 p_k} \\ \vdots & \ddots & \vdots \\ \sqrt{p_k p_1} & \cdots & \sqrt{p_k p_k} \end{pmatrix}$$

je rovna součtu prvků na hlavní diagonále, tj. za použití stejné podmínky multinomického rozdělení platí

$$Tr = \sqrt{p_1 p_1} + \sqrt{p_2 p_2} + \sqrt{p_3 p_3} \cdots + \sqrt{p_k p_k} = p_1 + p_2 + \cdots + p_k = 1.$$

Dále také platí

$$\boldsymbol{V} = \boldsymbol{DQD}. \qquad (\mathbf{2})$$

Protože ve vztahu (2) násobíme matici $\boldsymbol{Q}$ zleva i zprava regulární, tj. čtvercovou maticí, jejíž determinant je nenulový, hodnost matice $\boldsymbol{Q}$ se nemění a

$$h(\boldsymbol{V}) = k - 1.$$

Pseudoinverzní matice k matici $\boldsymbol{A}_{m\times n}$ se definuje jako taková matice $\boldsymbol{A}^{-}_{n\times m}$, jestliže platí

$$\boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{A} = \boldsymbol{A}.$$

Hledejme nyní pseudoinverzní matici k matici $\boldsymbol{V}$. Ukážeme, že volba $\boldsymbol{V^{-}} = \boldsymbol{D^{-2}}$ vyhovuje. Dle (2) a dokázané idempotenci matice $\boldsymbol{Q}$ platí

$$\boldsymbol{VV^{-}V} = \boldsymbol{DQDD^{-2}DQD} = \boldsymbol{DQ^{2}D} = \boldsymbol{DQD} = \boldsymbol{V}.$$

Tím je také dokázáno, že $\boldsymbol{D^{-2}}$ je pseudoinverzní matice k $\boldsymbol{V}$.

Pseudoinverzní matici $\boldsymbol{V^{-}}$ můžeme tedy vyjádřit ve tvaru

$$\boldsymbol{V^{-}} = \boldsymbol{D^{-2}} = \begin{pmatrix} np_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & np_k \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{np_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{np_k} \end{pmatrix}. \qquad (3)$$

V literatuře ([2], str. 269) je ukázáno, že vhodně transformovaný vektor s multinomickým rozdělením má asymptoticky pro $n \to \infty$ normální rozdělení a kvadratická forma tohoto vektoru má navíc asymptoticky chí-kvadrát rozdělení o $k - 1$ stupních volnosti.

Definujeme náhodnou veličinu

$$Y_i = \frac{X_i - EX_i}{\sqrt{np_i}} = \frac{X_i - np_i}{\sqrt{np_i}}, \qquad i = 1, \ldots, k.$$

Pak náhodný vektor $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_k)' = \boldsymbol{D^{-1}}(\boldsymbol{X} - E\boldsymbol{X})$ s varianční matici (1), dle centrální limitní věty konverguje pro $n \to \infty$ ke k-rozměrnému normálnímu rozdělení $N_k(\boldsymbol{0}, \boldsymbol{Q})$.

Obecně platí, že pokud vektor $\boldsymbol{W}$ konverguje k normální rozdělení $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ s varianční maticí, jejíž $h(\boldsymbol{V}) \geq 1$, potom náhodná veličina $(\boldsymbol{W} - E\boldsymbol{W})'\boldsymbol{V^{-}}(\boldsymbol{W} - E\boldsymbol{W})$ konverguje k rozdělení chí-kvadrát o stejném počtu stupňů volnosti, jako je hodnost této varianční matice. Tvrzení platí při libovolném zvolení pseudoinverzní matice $\boldsymbol{V^{-}}$.

V našem případě $\boldsymbol{\mu} = \boldsymbol{0}, \boldsymbol{\Sigma} = \boldsymbol{Q}, \boldsymbol{W} = \boldsymbol{Y}$. Tedy veličina

$$\boldsymbol{Z} = \boldsymbol{Y'Q^{-}Y} = \boldsymbol{Y'Y}$$

má rozdělení χ^2 o $k - 1$ stupních volnosti (blíže viz [2], str. 270).

Náhodnou veličinu

$$Z = (\boldsymbol{X} - E\boldsymbol{X})'\boldsymbol{D^{-1}}\boldsymbol{D^{-1}}(\boldsymbol{X} - E\boldsymbol{X}) = (\boldsymbol{X} - E\boldsymbol{X})'\boldsymbol{D^{-2}}(\boldsymbol{X} - E\boldsymbol{X}) =$$
$$= (\boldsymbol{X} - n\boldsymbol{p})'\boldsymbol{V}^{-}(\boldsymbol{X} - n\boldsymbol{p}),$$

kde $\boldsymbol{X} \sim M(p_1, p_2, \ldots, p_k, n)$ a $\boldsymbol{V}^{-}$ je pseudoinverzní matice daná vztahem (3), lze potom jednoduše upravit

$$Z = (X_1 - np_1, \cdots, X_k - np_k)\begin{pmatrix} \frac{1}{np_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{np_k} \end{pmatrix}\begin{pmatrix} X_1 - np_1 \\ \vdots \\ X_k - np_k \end{pmatrix} =$$

$$= \left(\frac{X_1 - np_1}{np_1}, \cdots, \frac{X_k - np_k}{np_k}\right)\begin{pmatrix} X_1 - np_1 \\ \vdots \\ X_k - np_k \end{pmatrix} =$$

$$= \frac{(X_1 - np_1)^2}{np_1} + \cdots + \frac{(X_k - np_k)^2}{np_k} = \sum_{j=1}^{k} \frac{\left(X_j - np_j\right)^2}{np_j}.$$

Nakonec tedy můžeme říci, že pro náhodný vektor $\boldsymbol{X}$ mající multinomické rozdělení s parametry $p_1, \ldots, p_k, n$, má náhodná veličina

$$Z = \sum_{j=1}^{k} \frac{\left(X_j - np_j\right)^2}{np_j} \qquad (4)$$

asymptoticky pro $n \to \infty$ rozdělení χ^2 o $k - 1$ stupních volnosti. Kritickým oborem je interval $\langle \chi^2_{k-1;1-\propto} \rangle$, v případě, že realizace padne do toho intervalu, nulovou hypotézu zamítáme.

Pro testování je potřeba mít dostatečně velký rozsah výběru n, přičemž pro každou třídu j musí platit $np_j \geq 5$. Při nesplnění této podmínky je možné třídy, které spolu jsou v příbuzném vztahu, souvisí spolu nebo jsou jen okrajové, slučovat. Se sloučením se zároveň mění i počet stupňů volnosti.

1.2 TESTY DOBRÉ SHODY PŘI ZNÁMÝCH PARAMETRECH

Testujeme, zda naměřené empirické četnosti, tj. realizace binomických náhodných veličin $X_1, \dots, X_k$ jsou ve shodě s četnostmi očekávanými. Teoretické četnosti jsou dány jako střední hodnoty těchto náhodných veličin, $np_1^0, np_2^0, \dots, np_k^0$. Nulová hypotéza tvrdí, že pravděpodobnosti v modelu multinomického rozdělení jsou rovny pravděpodobnostem $p_1^0, \dots, p_k^0$, tj.

$$H_0: p_1 = p_1^0, p_2 = p_2^0, \dots, p_k = p_k^0,$$

proti alternativě, že aspoň jedna z rovností neplatí. [1]

Jako testovací statistika se používá náhodná veličina Z daná vzorcem (4). Upravený vztah pro testovací statistiku Z má následující tvar

$$Z = \sum_{j=1}^{k} \frac{\left(X_j - np_j\right)^2}{np_j} = \sum_{j=1}^{k} \left(\frac{X_j}{\sqrt{np_j}} - \frac{np_j}{\sqrt{np_j}} \right)^2 =$$

$$= \sum_{j=1}^{k} \left(\frac{{X_j}^2}{np_j} - 2\frac{X_j}{\sqrt{np_j}}\frac{np_j}{\sqrt{np_j}} + \frac{n^2{p_j}^2}{np_j} \right) =$$

$$= \sum_{j=1}^{k} \frac{{X_j}^2}{np_j} - 2\sum_{j=1}^{k} X_j + n\sum_{j=1}^{k} p_j = \sum_{j=1}^{k} \frac{{X_j}^2}{np_j} - 2n + n =$$

$$= \sum_{j=1}^{k} \frac{{X_j}^2}{np_j} - n,$$

kde jsme využili vlastnosti pravděpodobností $\sum_{j=1}^{k} p_j = 1$ a dále pak $\sum_{j=1}^{k} X_j = n$.

1.3 TESTY DOBRÉ SHODY PŘI NEZNÁMÝCH PARAMETRECH

Často se stává, že pravděpodobnosti $p_1, \ldots, p_k$ uvažovaného multinomického rozdělení závisí na nějakém neznámém vektorovém parametru $\boldsymbol{a} = (a, \ldots, a_m)'$. Tyto parametry je potřeba odhadnout za použití obdobné metody jako v případě regrese, tj. metody nejmenších čtverců, a to metodou minimálního chí-kvadrátu, která spočívá v tom, že jako odhad bereme takovou hodnotu $\boldsymbol{a}$, která při pevných multinomických veličinách $X_1, \ldots, X_k$ minimalizuje funkci $Z(\boldsymbol{a})$, která je funkcí m proměnných $a_1, \ldots, a_m$.

Pro pravděpodobnosti předpokládáme, že platí

$$p_1 = p_1(\boldsymbol{a}), \cdots, p_k = p_k(\boldsymbol{a})$$

a dále

$$p_1(\boldsymbol{a}) + \cdots p_k(\boldsymbol{a}) = 1.$$

Testovací statistika je tedy nyní ve tvaru

$$Z(\boldsymbol{a}) = \sum_{i=1}^{k} \frac{\left(X_i - np_i(\boldsymbol{a})\right)^2}{np_i(\boldsymbol{a})}.$$

Upravený tvar pak

$$Z(\boldsymbol{a}) = \sum_{i=1}^{k} \frac{{X_i}^2}{np_i(\boldsymbol{a})} - n.$$

Pro výpočet odhadu parametru $\boldsymbol{a}$ řešíme soustavu m normálních rovnic o m neznámých. Derivujeme po částech testovací statistiku Z podle jednotlivých proměnných $a_1, \ldots, a_m$ a výsledné derivace položíme rovny nule. Derivováním upraveného tvaru bychom dostali soustavu rovnic

$$\frac{\partial Z(a)}{\partial a_j} = -\sum_{i=1}^{k} \frac{{X_i}^2}{np_i{}^2(\boldsymbol{a})} \frac{\partial p_i(\boldsymbol{a})}{\partial a_j} = 0, \qquad j = 1,2, \cdots, m,$$

která se obtížně řeší, proto se používá zjednodušený tvar

$$\sum_{i=1}^{k} \frac{X_i}{p_i(\boldsymbol{a})} \frac{\partial p_i(\boldsymbol{a})}{\partial a_j} = 0, \qquad j = 1,2, \cdots, m, \tag{5}$$

který dává dostatečně srovnatelný výsledek.

Takto zadanou soustavu lze dostat tak, že nejprve derivujeme po částech kritérium $Z(\boldsymbol{a})$ podle p_i

$$\sum_{i=1}^{k} \frac{(-2)n^2\big(X_i - np_i(\boldsymbol{a})\big)p_i(\boldsymbol{a}) - n\big(X_i - np_i(\boldsymbol{a})\big)^2}{n^2 {p_i}^2(\boldsymbol{a})} =$$

$$= (-2)\sum_{i=1}^{k}\left\{\frac{X_i - np_i(\boldsymbol{a})}{p_i(\boldsymbol{a})} + \frac{\big(X_i - np_i(\boldsymbol{a})\big)^2}{2n{p_i}^2(\boldsymbol{a})}\right\}.$$

Soustava normálních rovnic má tvar

$$-\frac{1}{2}\frac{\partial Z(\boldsymbol{a})}{\partial a_j} = \sum_{i=1}^{k}\left\{\frac{X_i - np_i(\boldsymbol{a})}{p_i(\boldsymbol{a})} + \frac{\big(X_i - np_i(\boldsymbol{a})\big)^2}{2n{p_i}^2(\boldsymbol{a})}\right\}\frac{\partial p_i(\boldsymbol{a})}{\partial a_j} = 0, \qquad j = 1, 2, \cdots, m.$$

Vliv výrazu $\frac{\big(X_i - np_i(\boldsymbol{a})\big)^2}{2n{p_i}^2(\boldsymbol{a})}$ je s roustoucím $n \to \infty$ čím dál tím menší, proto jej lze vynechat. Dále víme, že pokud parciálně zderivujeme $p_1(\boldsymbol{a}) + \cdots p_k(\boldsymbol{a}) = 1$ dostaneme rovnost

$$\frac{\partial p_1(\boldsymbol{a})}{\partial a_j} + \cdots \frac{\partial p_k(\boldsymbol{a})}{\partial a_j} = 0, \qquad j = 1{,}2, \cdots, m.$$

díky které můžeme upravit soustavu

$$\sum_{i=1}^{k} \frac{X_i - np_i(\boldsymbol{a})}{p_i(\boldsymbol{a})}\frac{\partial p_i(\boldsymbol{a})}{\partial a_j} = 0 \qquad j = 1{,}2, \cdots, m.$$

na výsledný jednodušší tvar.

Řešením výsledné zjednoušené soustavy rovnic dostaneme odhad parametru $\boldsymbol{a}$ a značíme jej $\widehat{\boldsymbol{a}}$**.** Dosazením takto obdržených odhadů $(\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_k)'$ do výrazu pro $Z(\boldsymbol{a})$ bude mít náhodná veličina

$$Z(\widehat{\boldsymbol{a}}) = \sum_{j=1}^{k} \frac{\big(X_j - np_j(\widehat{\boldsymbol{a}})\big)^2}{np_j(\widehat{\boldsymbol{a}})}$$

asymptoticky pro $n \to \infty$ rozdělení χ^2 o $k - 1 - m$ stupních volnosti, kde m označuje počet parametrů, které odhadujeme. Pokud se statistika realizuje hodnotou

větší než $\chi^2_{k-1-m;1-\alpha}$, nulovou hypotézu o očekávaném rozdělení zamítáme na hladině α. [2]

I zde při nedostatečných četnostech, tj. při nesplnění podmínky $np_j \geq 5$, je možno slučovat třídící intervaly.

2 TEST NEZÁVISLOSTI

Jedním z nejčastějších typů statistických úloh je zjištění, zda určité dvě kvalitativní náhodné veličiny, které mají ordinální nebo nominální charakter, spolu vzájemně souvisí a existuje mezi nimi tedy závislost.

2.1 TEORETICKÁ VÝCHODISKA

Nechť dvourozměrný náhodný vektor je tvořen náhodnou veličinou X, která nabývá hodnot $1,\dots,r$, a Y nabývající hodnot $1,\dots,s$ s pravděpodobnostmi $p_{ij} = (X = i, Y = j)$. Dále označíme

$$p_{i.} = P(X = i) = \sum_j p_{ij}\,,$$

$$p_{.j} = P(Y = j) = \sum_j p_{ij}.$$

Číslům $p_{i.}$ a $p_{.j}$ se říká marginální (okrajové) pravděpodobnosti a jednotlivé pravděpodobnosti p_{ij} je vhodné zapisovat ve tvaru matice, viz tabulka 2.1.

Tabulka 2.1 Matice pravděpodobností

$X\backslash Y$	1	2	⋯	s	Σ
1	p_{11}	p_{12}	⋯	p_{1s}	$p_{1.}$
2	p_{21}	p_{22}	⋯	p_{2s}	$p_{2.}$
⋮	⋮	⋮	⋯	⋮	⋮
r	p_{r1}	p_{r2}	⋯	p_{rs}	$p_{r.}$
Σ	$p_{.1}$	$p_{.2}$	⋯	$p_{.s}$	1

Empirické četnosti, při provedeném výběru o rozsahu n, kdy se vyskytla dvojice (i, j), se označují n_{ij}. Pro příslušné marginální četnosti, řádkové součty $n_{i.}$ a sloupcové součty $n_{.j}$ pak

$$n_{i.} = \sum_j n_{ij},$$

$$n_{.j} = \sum_i n_{ij}.$$

Dále pro rozsah souboru n platí

$$n = \sum_i n_{i.} = \sum_j n_{.j} = \sum_i \sum_j n_{ij}.$$

Tyto četnosti jednotlivých tříd všech možných kombinací hodnot náhodných veličin X, Y se zapisují ve tvaru matice, které se říká kontingenční tabulka, tabulka 2.2.

Tabulka 2.2 Kontingenční tabulka

$X \backslash Y$	1	2	⋯	s	Σ
1	n_{11}	n_{12}	⋯	n_{1s}	$n_{1.}$
2	n_{21}	n_{22}	⋯	n_{2s}	$n_{2.}$
⋮	⋮	⋮	⋯	⋮	⋮
r	n_{r1}	n_{r2}	⋯	n_{rs}	$n_{r.}$
Σ	$n_{.1}$	$n_{.2}$	⋯	$n_{.s}$	n

Jelikož při testování vycházíme z testů dobré shody, viz kapitola 1, máme nyní místo vektoru náhodných veličin $X_i, i = 1, \ldots, k$, majícího multinomické rozdělení, matici četností n_{ij}, $i = 1, \ldots, r$, $j = 1, \ldots, s$. Budeme porovnávat empirické (pozorované) četnosti s očekávanými (hypotetickými) četnostmi $\frac{n_{i.} n_{.j}}{n}$, které jsou výsledkem odhadů teoretických pravděpodobností pomocí modifikované metody minimálního χ^2.

Souvislosti mezi obecnou teorií kontingenčních tabulek a situací při testu nezávislosti shrnuje tabulka 2.3. Poslední dva řádky se vztahují k testu nezávislosti.

Tabulka 2.3 Tabulka souvislostí pro test nezávislosti

Testy dobré shody	**Kontingenční tabulky**
X_i	n_{ij}
$i = 1,\dots,k$	$i = 1,\dots,r, \quad j = 1,\dots,s$
$X_1,\dots,X_k$	$n_{11}, n_{12},\dots,n_{rs}$
vektor $\boldsymbol{X}$	matice n_{ij}
n	n
k	rs
p_i	p_{ij}
$\sum_{i=1}^{k} p_i = 1$	$\sum_{i=1}^{r}\sum_{j=1}^{s} p_{ij} = 1$
$\sim\chi^2_{k-1}$	$\sim\chi^2_{rs-1}$
$Z = \sum_{j=1}^{k} \frac{(X_j - np_j)^2}{np_j}$	$Z = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$
$\sim\chi^2_{k-1-m}$	$\sim\chi^2_{rs-1-(r+s-2)}$

Hypotéza nezávislosti v případě kontingenčních tabulek má tvar

$$H_0: p_{ij} = p_{i.}p_{.j}, \qquad i = 1,\dots,r, \qquad j = 1,\dots,s,$$

proti alternativě

$$H_A: \exists i,j: p_{ij} \neq p_{i.}p_{.j},$$

Vektor volných parametrů

$$\boldsymbol{a} = (p_{1.}, p_{2.},\dots,p_{r-1.}, p_{.1}, p_{.2},\dots,p_{.s-1},),$$

obsahuje všechny marginální pravděpodobnosti kromě

$$p_{r.} = 1 - p_{1.} - p_{2.} - \dots - p_{r-1.}$$

a

$$p_{.s} = 1 - p_{.1} - p_{.2} - \dots - p_{.s-1},$$

které nejsou volnými parametry, neboť se vážou na podmínku součtů jednotlivých marginálních pravděpodobností

$$\sum_{i=1}^{r} p_{i.} = 1$$

a

$$\sum_{j=1}^{s} p_{.j} = 1,$$

kterými jsou jednoznačně určeny.

Počet volných parametrů je tedy roven

$$m = r - 1 + s - 1.$$

Odhady volných parametrů po řešení zjednodušené soustavy rovnic (5) jsou rovny (viz Anděl, str. 281)

$$\hat{p}_{i.} = \frac{n_{i.}}{n},$$

$$\hat{p}_{.j} = \frac{n_{.j}}{n}.$$

Proto za platnosti nulové hypotézy platí

$$\hat{p}_{ij} = \hat{p}_{i.}\hat{p}_{.j} = \frac{n_{i.}}{n}\frac{n_{.j}}{n} = \frac{n_{i.}n_{.j}}{n^2}.$$

Veličina

$$Z(\hat{\boldsymbol{a}}) = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - n\frac{n_{i.}n_{.j}}{n^2}\right)^2}{\frac{n_{i.}n_{.j}}{n^2}} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

má asymptoticky pro $n \to \infty$ rozdělení χ^2 s počtem stupňů volnosti $rs - (r + s - 2) - 1 = (r - 1)(s - 1)$. Nulovou hypotézu následně zamítáme v případě, že $z \geq \chi^2_{(r-1)(s-1);1-\alpha}$.

Ke shodě s limitním rozdělením se vyžaduje, aby aspoň pro 80 % případů platila pro teoretické četnosti nerovnost $\frac{n_{i.}n_{.j}}{n} \geq 5$ a pro zbylých 20 % neklesne hodnota pod 2. Při nesplnění podmínky lze jednotlivé sousední sloupce či řádky sloučit do jednoho,

eventuálně lze některou málo četnou hodnotu zcela vypustit. Řádky ani sloupce se však nesmí zredukovat na jeden. [3]

2.2 VLIV KOUŘENÍ RODIČŮ NA KOUŘENÍ DĚTÍ

Při doprovázení mladší sestry (15 let) do školy jsem byla překvapená, kolik z místních dětí ze základní školy před vyučováním kouří cigarety. Byla jsem obeznámena, že jde zcela o standardní jev, a že kdo nekouří, je bohužel „out“. A kdo nekouří „trávu“, je „out“ ještě více.

Prostřednictvím mé blízké osoby vyučující na druhém stupni základní školy ve Frýdku-Místku byl proveden výzkum formou dotazníku. Pomocí několika jednoduchých otázek, pro žáky osmých a devátých tříd, jsem získala informace k provedení testu, jehož cílem bylo určit, zda existuje závislost mezi kouřením cigaret u dětí a jejich rodičů. Celkem bylo dotázáno 44 žáků, kteří odpovídali na následující otázky.

- Kouří cigarety vaši rodiče? [Ano, ne.]
- Kouříš pravidelně? [Ano, ne.]

Výsledky byly zpracovány do kontingenční tabulky 2.4. Znak X= kouření rodičů, ANO, NE, znak Y= kouření dítěte, ANO, NE.

Tabulka 2.4 Údaje o kouření rodičů a dětí

$X\backslash Y$	ANO	NE	Σ
ANO	12	5	17
NE	9	18	27
Σ	21	23	44

Provedeme oboustranný test a testujeme hypotézu H_0, že kouření rodičů a jejich dětí jsou nezávislé, proti alternativě, že tato rovnost neplatí. Výpočtem očekávaných

četností, tabulka 2.4, zjistíme, že všechny splňují podmínku a jsou větší než 5, proto můžeme použít statistiku Z.

$$n\hat{p}_{11} = \frac{n_{1.}n_{.1}}{n} = \frac{17.21}{44} = 8{,}114,$$
$$n\hat{p}_{12} = \frac{n_{1.}n_{.2}}{n} = \frac{17.23}{44} = 8{,}886,$$
$$n\hat{p}_{21} = \frac{n_{2.}n_{.1}}{n} = \frac{27.21}{44} = 12{,}886,$$
$$n\hat{p}_{22} = \frac{n_{2.}n_{.2}}{n} = \frac{27.23}{44} = 14{,}114.$$

Dosazením do testovací statistiky

$$z = \frac{\left(12 - \frac{21.17}{44}\right)^2}{\frac{21.17}{44}} + \frac{\left(5 - \frac{23.17}{44}\right)^2}{\frac{23.17}{44}} + \frac{\left(9 - \frac{21.17}{44}\right)^2}{\frac{21.17}{44}} + \frac{\left(18 - \frac{23.27}{44}\right)^2}{\frac{23.27}{44}} = 5{,}803.$$

Srovnáním s kvantilem $\chi^2_{1;0,95} = 3{,}84$ dojdeme k závěru, že hypotézu o nezávislosti zamítáme. Tento výsledek je ve shodě s mnohými studiemi, které již byly zpracovány. Rodič kuřák jistě není dobrým příkladem pro své děti a měl by proto zvážit, zda právě vlastní děti nejsou tím správným důvodem s takovou závislostí přestat a jestli právě ony nejsou dostatečným hnacím motorem pro lepší a zdravější život.

2.5 Údaje o kouření EXCEL

		Dítě		
Rodiče	data	Kouří	Nekouří	Celkem
Kouří	počet	12	5	17
	%	27%	11%	39%
Nekouří	data	9	18	27
	%	20%	41%	61%
celkem počet		21	23	44
celkem %		48%	52%	100%

Z takto vytvořené tabulky 2.5 lze vyčíst, že skoro polovina dotázaných dětí v počtu 21 kouří a u 27 % z nich je to v domácnosti běžné, neboť kouří i jejich rodiče.

U nekouřících dětí z 41 % nekouří ani doma. Lze proto již nyní usuzovat, že spolu tyto dva znaky souvisí.

Pro provedení chí-kvadrát testu v programu Excel je potřeba vytvořit dvě tabulky, jejichž hodnoty se budou následně dosazovat do funkce CHITEST(actual_range;expected_range). Hodnoty pozorovaných četností a četností očekávaných jsou uvedeny v tabulce 2.6.

Tabulka 2.6 Pozorované a očekávané četnosti

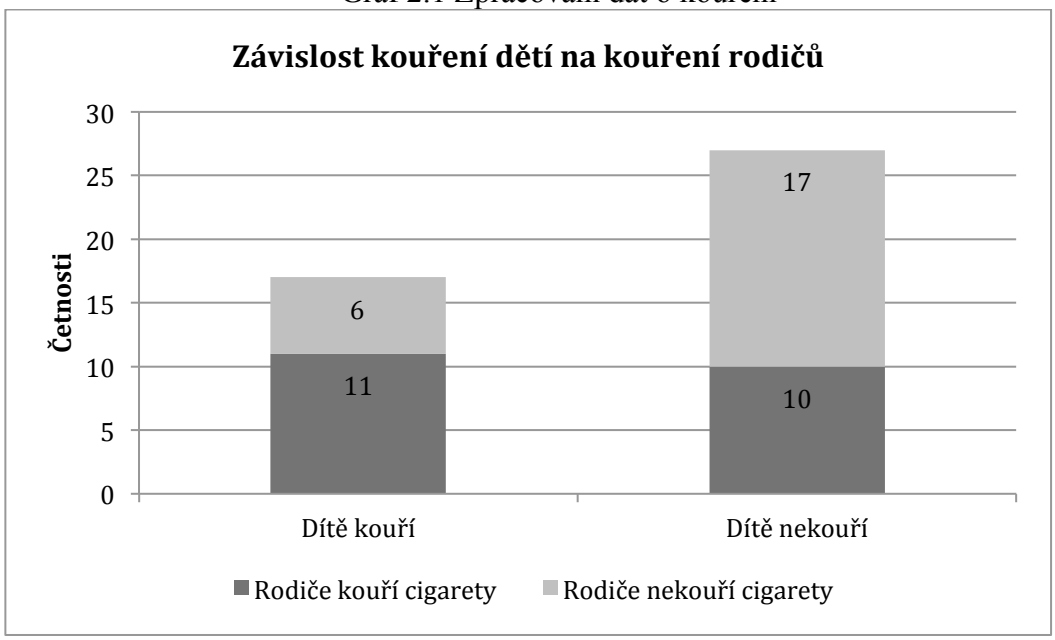
12	5
8,114	8,886
9	18
12,886	14,114

Výsledkem chí testu je hodnota $p - value$, která se srovnává s hladinou významnosti $\alpha = 0{,}05$. Jelikož

$$p - value = 0{,}016 < 0{,}05,$$

nulová hypotéza se zamítá a kouření rodičů má bohužel vliv i na kouření dítěte. Naměřená data jsou zobrazena pomocí skládananého sloupcového grafu 2.1.

Graf 2.1 Zpracování dat o kouření



2.3 KOUŘENÍ MARIHUANY MEZI KAMARÁDY

Na stejném vzorku jako v kapitole 2.2, tj. 14 a 15 letých dětí, bylo zkoumáno, zda existuje souvislost mezi kouřením marihuany u jednotlivých dětí a okolím, ve kterém se pohybují. Odpovídali opět na dvě otázky.

- Zkoušeli jste někdy kouřit marihuanu?
  - Ano, kouřím pravidelně,
  - již jsem to zkusil/a,
  - ne, nikdy.
- Zkoušeli kouřit tví nejbližší kamarádi?
  - Ano,
  - ne.

Dle výsledných četností je zpracována tabulka 2.7 a skládaný sloupcový graf 2.2. Veličina X značí odpověď na první, výše zmiňovanou otázku, a hodnoty pak Ano, kouřím pravidelně, již jsem to zkusil/a a ne, nikdy. Veličina Y označuje otázku druhou, kouří, nekouří.

Tabulka 2.7 Údaje o kouření marihuany

$X \backslash Y$	ANO	NE	Σ
PRAVIDELNĚ	9	3	12
ZKUSIL/A	13	4	17
NIKDY	3	12	15
Σ	25	19	44

Testujeme H_0: kouření marihuany dítěte nezáví na kouření jeho nejbližšího okruhu přátel proti oboustranné alternativě, že tato rovnost neplatí.

Realizace testovací statistiky Z

$$z = \frac{\left(9 - \frac{25.12}{44}\right)^2}{\frac{25.12}{44}} + \frac{\left(3 - \frac{19.12}{44}\right)^2}{\frac{19.12}{44}} + \frac{\left(13 - \frac{25.17}{44}\right)^2}{\frac{25.17}{44}} + \frac{\left(4 - \frac{19.17}{44}\right)^2}{\frac{19.17}{44}}$$
$$+ \frac{\left(3 - \frac{25.15}{44}\right)^2}{\frac{25.15}{44}} + \frac{\left(12 - \frac{19.15}{44}\right)^2}{\frac{19.15}{44}} = 12{,}580.$$

Kritická hodnota je rovna $\chi^2_{2;0,95} = 5{,}99$. Hypotézu na hladině významnosti $\propto = 0{,}05$ zamítáme. Z výsledku testování můžeme potvrdit, že to, zda dítě vyzkoušelo, popř. pravidelně kouří marihuanu, souvisí s jeho okolím. Domnívám se, že ti, kteří kouří nebo to zkoušeli, to dělají hlavně proto, aby byli zajímaví a mohli se chlubit fotografiemi na sociálních sítích a mnohdy si ani neuvědomují, jaký vliv tato látka má na zdraví a chování. Spousta z nich to ani pořádně neumí a hrají si, jak na ně látka neúčinkuje. Tyto informace byly poskytnuty z několika zdrojů, ať už užívajících, tak i těch, kteří se tomu vyhýbají. Je poměrně zarážející, co jsou děti v tomto věku schopny dělat, aby byli pro okolí zajímavějšími, a aby „zapadli“ mezi určité skupiny lidí. Závislost na kamarádech je patrná i u alkoholu, se kterým začínají již v tak brzkém věku. Podobně je tomu i v aktivitách během dne i noci, navštěvování obchodních center a nočních klubů, kde by vůbec neměli mít přístup.

Zpracováním dat v programu Excel dostaneme nejprve tabulku 2.8.

Tabulka 2.8 Údaje o kouření marihuany EXCEL

		Okolí kamarádů		
Dítě	data	ANO	NE	celkem
Pravidelně	počet	9	3	12
	%	20%	7%	27%
Zkusil/a	počet	13	4	17
	%	30%	9%	39%
Nikdy	počet	3	12	15
	%	7%	27%	34%
celkem počet		25	19	44
celkem %		57%	43%	100%

Shrnutím je patrné, že celých 66 % dotazovaných dětí ve věku 13-15 už má zkušenosti s kouřením marihuany. U více než poloviny dotazovaných toto praktikují jejich kamarádi, dá se tedy předpokládat, že i oni sami budou mít nebo mají tendence zapadnout, vyrovnat se či být lepším taktéž.

Pozorované a očekávané četnosti jsou zaznamenány v tabulkách 2.9.

Tabulka 2.9 Pozorované a očekávané četnosti

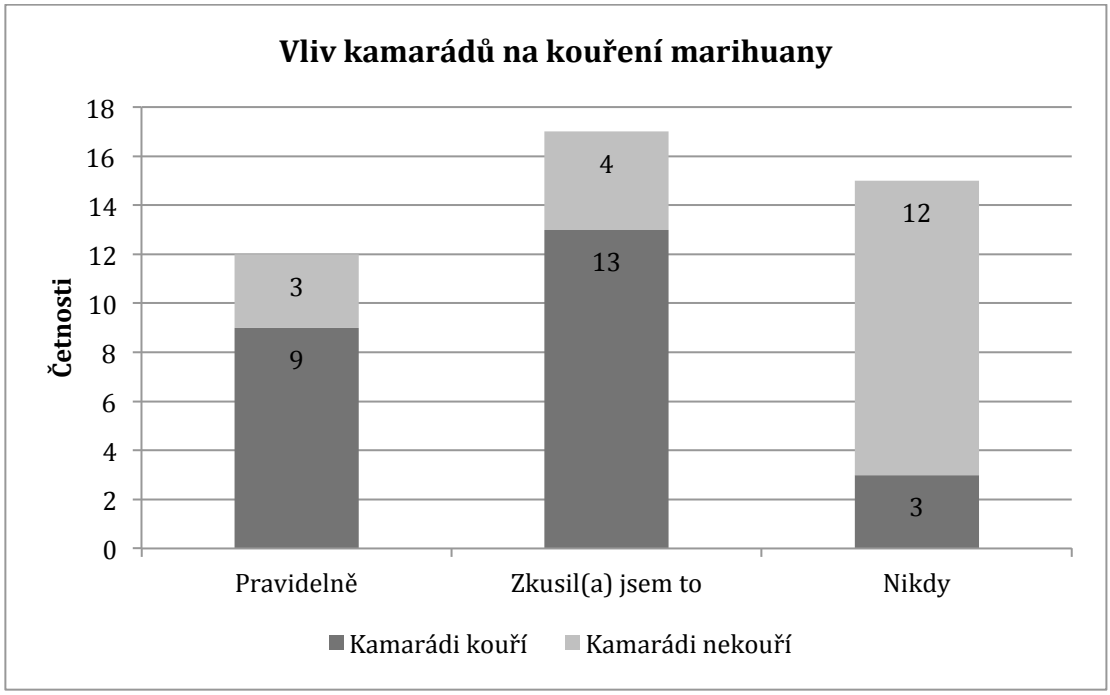
9	3
6,818	5,182
13	4
9,659	7,341
3	12
8,523	6,477

Za použití funkce CHITEST v Excelu dostaneme hodnotu

$$p - value = 0{,}002,$$

která signalizuje, že H_0 zamítáme na hladině významnosti $\alpha = 0{,}05$.

Graf 2.2 Zpracování dat o kouření marihuany



3 TEST HOMOGENITY

Testem homogenity rozumíme testování shodnosti struktury jednoho ze sledovaných znaků za různých podmínek, které jsou roztřízeny do jednotlivých kategorií.

3.1 TEORETICKÁ VÝCHODISKA

Jsou-li řádkové četnosti $n_{i.}$ v kontingenční tabulce předem stanoveny, lze tyto řádky považovat za r výběrů z multinomického rozdělení s parametry $n_{1.}, \ldots, n_{r.}$. Většinou je pak třeba testovat hypotézu homogenity, která říká, že příslušná multinomická rozdělení mají stejné pravděpodobnosti.

Testovací hypotéza má tvar $H_0: p_{i1} = p_1, \ldots, p_{is} = p_s, \forall i \in 1, \ldots, r,$ proti alternativě, že aspoň jedna z rovností neplatí. Za platnosti nulové hypotézy má testovací statistika

$$Z = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i.} n_{.j}}{n}\right)^2}{\frac{n_{i.} n_{.j}}{n}}$$

asymptoticky χ^2 rozdělení o $(r-1)(s-1)$ stupních volnosti. Hypotézu zamítáme v případě, že je realizace testovací statistiky větší než hodnota příslušného kvantilu, tj. $z \geq \chi^2_{(r-1)(s-1);1-\alpha}$. [1], [4]

3.2 VÝKONOSTNÍ ROZLOŽENÍ RANNÍHO A ODPOLEDNÍHO BĚHU

Běhání je jedním z mých největších koníčků, průměrně naběhám za měsíc 130 km v 16 bězích, 4x týdně a především v odpoledních či večerních hodinách. Pokud vím, že následující den nebudu mít k večeru čas, zařadím běh ještě ráno před snídaní. Problém je v tom, že mi delší dobu trvá, než své tělo rozhýbu a dostanu se na své běžné rychlostní tempo. Zajímalo mě proto, zda mé výkony jsou ráno rozdílné oproti těm odpoledním, nebo se neliší a neexistuje mezi dobou běhu a výslednými časy žádná závislost.

Během dvou měsíců jsem byla celkem 20x běhat ráno a 36x odpoledne. Nejčastěji běhám 6 km do 36 min, 8 km vzdálenost do 48 min a 10 km pod jednu hodinu. Ze subjektivního pocitu jsem předpokládala, že výsledky běhů dopadnou tak, že bude hypotézu třeba zamítnout, avšak tato domněnka se ukáže později jako mylná.

V následující tabulce 3.1 jsou zaznamenány četnosti jednotlivých běhů, kdy byly tyto časy překročeny a kdy byly lepší nebo stejné. X označuje čas běhu, ráno a odpoledne. Veličina Y pak výsledné časy, horší čas a čas lepší nebo v limitu.

Tabulka 3.1 Údaje o běhání

$X\backslash Y$	HORŠÍ ČAS	V LIMITU	Σ
RÁNO	11	9	20
ODPOLEDNE	12	24	36
Σ	23	33	56

Na takto získaných datech testujeme hypotézu o symetrii, protože nás zajímá, zda jsou výsledky symetrické, tj. dobré běžecké výsledky ráno i odpoledne jsou stejně pravděpodobné. Testujeme hypotézu $H_0: p_{11} = p_{21} = p_1, p_{12} = p_{22} = p_2$ proti oboustranné alternativě, že aspoň jedna z rovností neplatí.

V případě nezávislosti by nulová hypotéza byla formulovaná následovně $H_0: p_{ij} = p_{i.}p_{.j}$. Testovali bychom, zda existuje závislost mezi dobou běhu a výkonností.

Dosazením do příslušné statistiky testování homogenity

$$z = \frac{\left(11 - \frac{20.23}{56}\right)^2}{\frac{20.23}{56}} + \frac{\left(9 - \frac{20.33}{56}\right)^2}{\frac{20.33}{56}} + \frac{\left(12 - \frac{36.23}{56}\right)^2}{\frac{36.23}{56}} + \frac{\left(24 - \frac{36.33}{56}\right)^2}{\frac{36.33}{56}} = 2{,}494$$

Srovnáním s příslušným kvantilem $\chi^2_{1;0,95} = 3{,}84$ vidíme, že realizace testovací statistiky je menší a nulovou hypotézu tedy nelze zamítnout. Běhání ráno a odpoledne vzhledem k výsledným časům tedy nejsou rozdílné.

Jak je vidět v tabulce 3.2, více než polovina, 59%, všech běhu bylo ve standardním času, kdežto 41% bylo horších. Nejlepších časů bylo častěji dosahováno v odpoledních hodinách. Horší časy ráno i odpoledne byly zhruba srovnatelné.

Tabulka 3.2 Údaje o běhání EXCEL

		Výkon		
Čas běhu	data	Horší čas	Lepší/v limitu	celkem
Ráno	počet	11	9	20
	%	20%	16%	36%
Odpoledne	data	12	24	36
	%	21%	43%	64%
celkem počet		23	33	56
celkem %		41%	59%	100%

Pozorované četnosti využité do funkce CHITEST, actual_range, jsou uvedeny v tabulce 3.3 v bílých kolonkách, očekávané četnosti pak pro kolonku expected_range jsou označeny tmavě.

Tabulka 3.3 Pozorované a očekávané četnosti

11	9
8,214	11,786
12	24
14,786	21,214

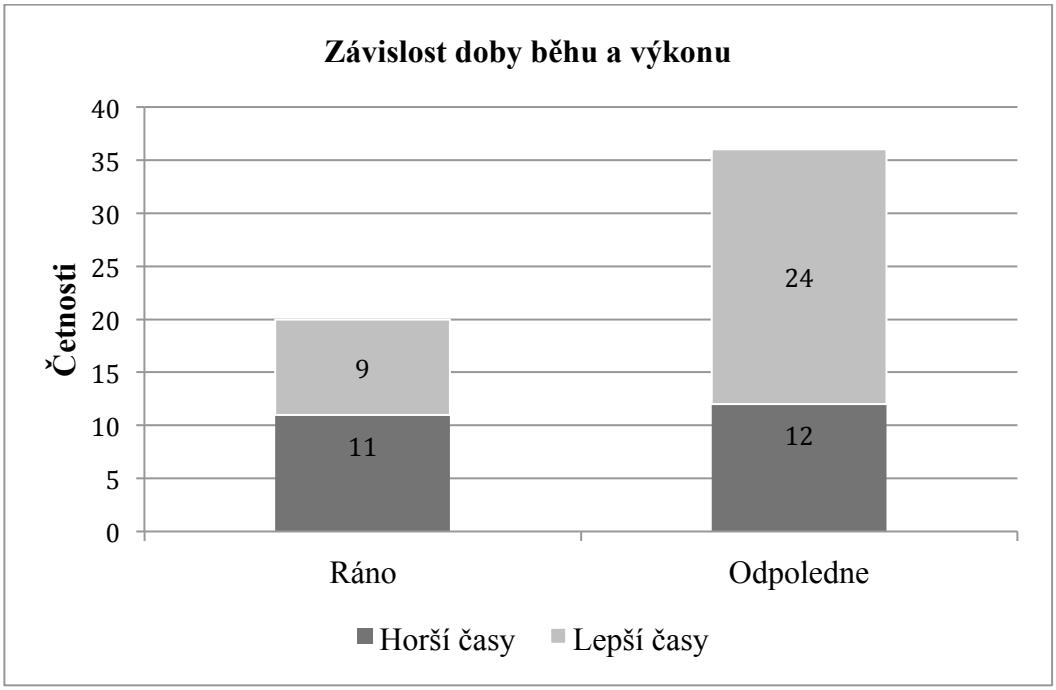
Srovnáním vypočtené hodnoty

$$p - value = 0{,}144$$

s hladinou $\alpha = 0{,}05$ vidíme, že rozdílné časy mezi běhy ráno a odpoledne jsou způsobeny náhodou a nebyl nalezen signifikantní rozdíl.

Grafickým výstupem pro analýzu homogenity ranních a odpoledních běhů je kumulativní sloupcový graf 3.1.

Graf 3.1. Zpracování dat o běhání



3.2.1 TESTOVÁNÍ PŘI VĚTŠÍM ROZSAHU VÝBĚRU

Výsledky testování běhů během dvou měsíců byly překvapením, vzhledem k pocitu, který při ranních bězích mám. Proto byly k datům z kapitoly 3.2 přidáno ještě 12 běhů z dalšího měsíce, ve kterém bylo naběháno z důvodu nemoci jen 96 km. Záměrně bylo uskutečněno více běhů ráno, i přes častou nechuť pohybu. Výsledky jsou zaznamenány v tabulce 3.4.

Tabulka 3.4 Údaje o běhání s přidanými běhy

$X\backslash Y$	HORŠÍ ČAS	V LIMITU	Σ
RÁNO	16	12	28
ODPOLEDNE	13	27	40
Σ	29	39	68

Na takto získaných datech testujeme znova hypotézu $H_0: p_{11} = p_{21} = p_1, p_{12} = p_{22} = p_2$ proti alternativě, že aspoň jedna z rovností neplatí.

Dosazením do příslušné statistiky testování homogenity

$$z = \frac{\left(16 - \frac{29.28}{68}\right)^2}{\frac{29.28}{68}} + \frac{\left(12 - \frac{39.28}{68}\right)^2}{\frac{29.28}{68}} + \frac{\left(13 - \frac{29.40}{68}\right)^2}{\frac{29.40}{68}} + \frac{\left(27 - \frac{39.40}{68}\right)^2}{\frac{39.40}{68}} = 4{,}089$$

a srovnáním s příslušným kvantilem $\chi^2_{1;0,95} = 3{,}84$ vidíme, že realizace testovací statistiky je větší a nulovou hypotézu zamítáme. Běhání ráno a odpoledne vzhledem k výsledným časům tedy jsou statisticky rozdílné.

Srovnáním hodnoty

$$p - value = 0{,}043,$$

s hladinou $\alpha = 0{,}05$, kdy $0{,}043 < 0{,}05$, hypotézu rovněž zamítáme.

Při sestavování tabulky bylo zjištěno, že v případě zaznamenání 11 běhů, namísto 12, kdy by ubyl jeden ranní běh s horším časem, byla by hodnota testovací statistiky rovna

$$z = 3{,}522,$$

a hypotézu bychom v tomto případě nemohli opět zamítnout a běhaní ráno a odpoledne by stále bylo považováno za statisticky nerozdílné.

Na základě srovnání hodnoty $p - value$

$$p - value = 0{,}061,$$

s hladinou $\alpha = 0{,}05$, kdy $0{,}061 > 0{,}05$, hypotézu nelze zamítnout.

Zde jde názorně vidět, jaký vliv má i malá změna rozsahu souboru na hodnotu realizace testovací statistiky a s tím i zamítnutí či nezamítnutí nulové hypotézy.

3.3 PODÍL OBJEDNANÝCH DRUHŮ PIV U MUŽŮ A ŽEN

Ve spolupráci s jednou pražskou hospodou byl proveden gendrově zaměřený výzkum mezi místními hosty, jaký druh piva si obvykle dávají při večerním posezení. Výčepní postupně zaznamenával, na mnou předpřipravený arch, objednávky zvlášť u mužů a u žen. Cílem bylo zjistit, zda má pohlaví X, žena, muž, vliv na druh objednaného piva Y, třetinka, šnyt, mlíko a hladinka.

Šnytem se rozumí pivo čepované najednou do půllitru. Míra není v tomto případě podstatná, zbytek tvoří bohatá pěna. Jde o kompromis mezi malým třetinkovým pivem a hladinkou, čili klasickým půllitrem s hustou krémovou pěnou. Mlíko tvoří pouze smetanová pěna načepovaná do půllitru až po okraj a jeho kouzlo tkví právě v rychlém vypití po čepování Tento druh piva je oblíbený především u znalců a pokud člověk, který pivo má rád mlíko nezná, měl by jej určitě vyzkoušet. Bohužel ne ve všech hospodách mlíko znají a umějí správně načepovat.

Bylo zaznamenané celkem 200 mužů a 100 žen, zjištěny jejich objednávky a výsledky shrnuty do přehledné tabulky 3.5.

Tabulka 3.5 Údaje o objednaných pivech

$X\backslash Y$	TŘETINKA	ŠNYT	MLÍKO	HLADINKA	Σ
ŽENA	17	26	13	44	100
MUŽ	9	48	20	123	200
Σ	26	74	33	167	300

Testujeme nulovou hypotézu o homogenitě $H_0: p_{11} = p_{21} = p_1, p_{12} = p_{22} = p_2, p_{13} = p_{23} = p_3, p_{14} = p_{24} = p_4$, zda podíl objednávek piva u mužu je stejný jako u žen, proti oboustranné alternativě, že aspoň jedna z rovností neplatí.

Dosazením do testovací statistiky

$$z = \frac{\left(17 - \frac{26.100}{300}\right)^2}{\frac{26.100}{300}} + \frac{\left(26 - \frac{74.100}{300}\right)^2}{\frac{74.100}{300}} + \cdots + \frac{\left(119 - \frac{167.200}{300}\right)^2}{\frac{167.200}{300}} =$$

$$= 8{,}013 + 0{,}072 + 0{,}364 + 2{,}445 + 0{,}013 + 4{,}006 + 0{,}036 + 0{,}0182 + 1{,}223 =$$

$$= 16{,}190.$$

Srovnáním s příslušným kvantilem $\chi^2_{3;0,95} = 7{,}815$ je nutno nulovou hypotézu zamítnout. Podíl druhů piv je tedy u mužů a žen rozdílný.

V případě řešení v programu Excel se vytvoří tabulka 3.6, která je stejně jako u předchozích příkladů doplněná o procentuální vyjádření zastoupení jednotlivých piv u žen i mužů.

Tabulka 3.6 Údaje o objednaných pivech EXCEL

		Druh piva				
Pohlaví	data	třetinka	šnyt	mlíko	hladinka	celkem
žena	počet	17	26	13	44	100
	%	17%	26%	13%	44%	100%
muž	data	9	48	20	123	200
	%	5%	24%	10%	62%	100%
celkem počet		26	74	33	167	300
celkem %		9%	25%	11%	56%	100%

Z tabulky 3.6 je vidět, že podíl třetinky je daleko vyšší u žen a podíl hladinek je daleko vyšší u mužů. Podíl šnytů a mlíka je u obou pohlaví srovnatelný. Chí-kvadrát testem dojdeme k ověření, zda jsou tyto rozdíly způsobené náhodou, či jde o signifikantní rozdíl.

Hodnoty pozorovaných (bílá pole) a teoretických (tmavá pole) četností jsou uvedeny v tabulce 3.7.

Tabulka 3.7 Pozorované a očekávané četnosti

17	26	13	44
8,667	24,667	11,000	55,667
9	48	20	123
17,333	49,333	22,000	111,333

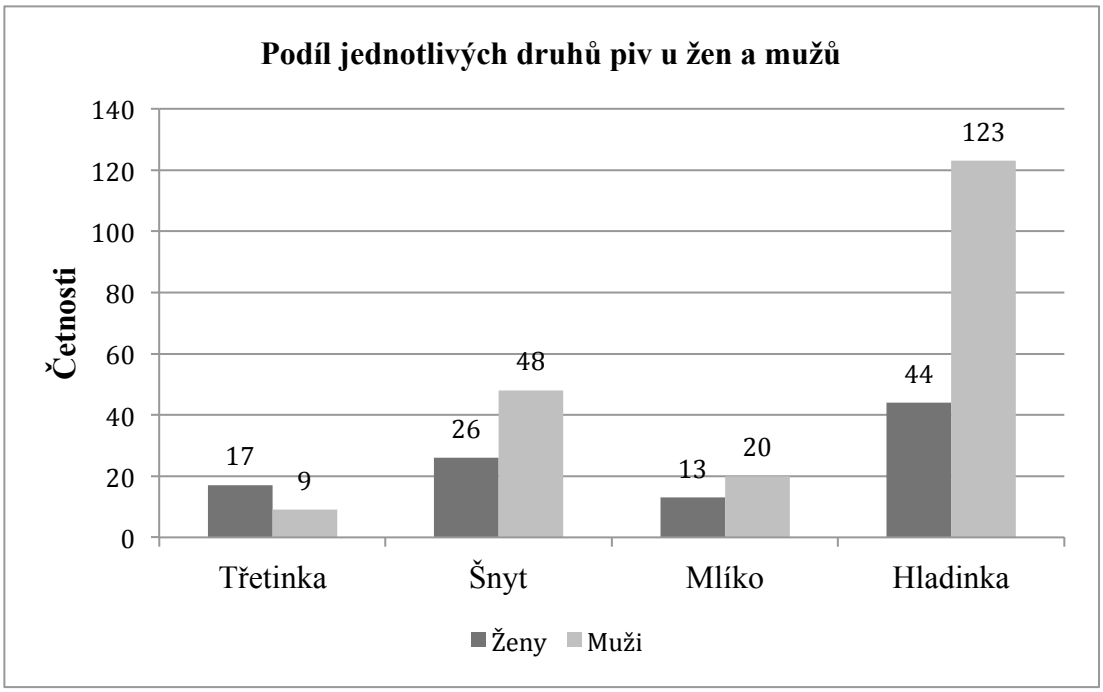
Výsledkem použití funkce CHITEST je hodnota

$$p - value = 0{,}000966,$$

hypotéza se zamítá na hladině významnosti $\alpha = 0{,}05$ a podíl objednaných piv u mužů a žen se signifikantně liší, čímž je potvrzen výsledek za použití testovací statistiky Z.

Data jsou graficky zpracována do shlukového sloupcového grafu 3.2.

Graf 3.2 Zaznamenání objednaných piv



3.3.1 TESTOVÁNÍ PŘI MENŠÍM ROZSAHU VÝBĚRU

V případě, že bychom vzali v úvahu menší, čtvrtinový, vzorek žen i mužů, tedy celkem 50 mužů a 25 žen, dostali bychom následující počty objednaných piv, viz tabulka 3.8.

Tabulka 3.8 Údaje o objednaných pivech při menším rozsahu

$X\backslash Y$	TŘETINKA	ŠNYT	MLÍKO	HLADINKA	Σ
ŽENA	8	5	2	10	25
MUŽ	3	10	5	32	50
Σ	11	15	7	42	75

Testujeme nulovou hypotézu o homogenitě $H_0: p_{11} = p_{21} = p_1, p_{12} = p_{22} = p_2, p_{13} = p_{23} = p_3, p_{14} = p_{24} = p_4$, zda podíl objednávek piva u mužu je stejný jako u žen, proti alternativě, že aspoň jedna z rovností neplatí.

Z hodnot pozorovaných (bílá pole) a teoretických (tmavá pole) četností, které jsou uvedeny v tabulce 3.9, vidíme, že není splněna podmínka dostatečných četností. V tomto případě se nabízí možnost sloučení dvou druhů piv, třetinky a mlíka, které jsou sice rozdílné, ale pro názornost je lze označovat jako méně objednávané druhy.

Tabulka 3.9 Pozorované a očekávané četnosti

8	5	2	10
3,667	5	2,333	14
3	10	5	32
7,333	10	4,677	28

Sloučením prvního a třetího sloupce dostaneme novou kontingenční tabulku 3.10.

Tabulka 3.10 Údaje o objednaných pivech po sloučení

$X\backslash Y$	OSTATNÍ	ŠNYT	HLADINKA	Σ
ŽENA	10	5	10	25
MUŽ	8	10	32	50
Σ	18	15	42	75

Pozorované a očekávané četnosti jsou uvedeny v tabulce 3.11.

Tabulka 3.11 Pozorované a očekávané četnosti

10	5	10
6	5	14
8	10	32
12	10	28

Dosazením do testovací statistiky

$$z = \frac{\left(10 - \frac{18.25}{75}\right)^2}{\frac{18.25}{75}} + \frac{\left(5 - \frac{15.25}{75}\right)^2}{\frac{15.25}{75}} + \cdots + \frac{\left(32 - \frac{42.50}{75}\right)^2}{\frac{42.50}{75}} =$$

$$= 2{,}667 + 0 + 1{,}143 + 1{,}333 + 0 + 0{,}571 =$$

$$= 5{,}714.$$

Srovnáním s příslušným kvantilem $\chi^2_{2;0,95} = 5{,}991$ docházíme k závěru, že hypotézu nelze zamítnout. Podíl druhů piv je tedy u mužů a žen stejný. Z hodnoty realizace testovací statistiky, která je dost blízká kritické hodnotě, můžeme ale předpokládat, že při malém zvýšení rozsahu výběru už budeme muset hypotézu zamítnout a došli bychom ke stejnému závěru, jako při testování celého naměřeného souboru 100 žen a 200 mužů.

Hodnota $p - value$ je v tomto případě

$$p - value = 0{,}057$$

a srovnáním s hladinou významnosti $\alpha = 0{,}05$ je $0{,}057 > 0{,}05$ a hypotézu nelze zamítnout. Podíl objednaných piv u mužů a žen se signifikantně liší, čímž je potvrzen výsledek za použití testovací statistiky Z.

Pokud bychom opět srovnali výsledky při různých rozsazích, víme, že čím nižší nám vyjde hodnota $p - value$, tím spíš jsme přesvědčení o nepravdivosti nulové hypotézy a je třeba ji zamítnout. Při větším rozsahu (300) je hodnota $p - value$ více než 50x menší než hladina významnosti $\propto$ a máme proto dostatečný důvod být přesvědčení o nesprávnosti nulové hypotézy o symetrii. Když vezmeme v potaz jen čtvrtinový výběr, který však může být ovlivněn různými vlivy, např. časem objednání aj., hodnota $p - value$ je hodně blízká hladině $\propto$ a rozhodnutí o nulové hypotéze není tak jednoznačné, jako v prvním případě.

4 FISHERŮV FAKTORIÁLOVÝ TEST

Na základě nákladnosti některých prováděných pokusů, například v medicíně nebo při provádění různých destruktivních zkoušek, byl odvozen R. A. Fisherem tzv. Fisherův faktoriálový test.

4.1 TEORETICKÁ VÝCHODISKA

Důležitou podmínkou pro aproximaci rozdělením χ_1^2 je splnění dostatečných teoretických četností $\frac{n_{i.}n_{.j}}{n} > 6$. V případě nesplnění lze sloupce či řádky spojovat. U čtyřpolní tabulky toto ovšem nelze a proto se používá Fisherův faktoriálový test, který umožňuje testovat hypotézu o nezávislosti i při malých četnostech. I přes výtky ohledně dosažené hladiny testu vyjádřené součtem jednotlivých pravděpodobností, která je nižší než hladina α a k zamítnutí hypotézy tak dochází až při větším odchýlení, je tento test hojně využíván. Je založen na výpočtu podmíněné pravděpodobnosti p toho, že při daných marginálních četnostech $n_{1.}, n_{2.}, n_{.1}, n_{.2}$ vznikne tabulka s četnostmi $n_{11}, n_{12}, n_{21}, n_{22}$. [1]

Celkový počet možností, že při třídění do čtyřpolní tabulky dostaneme tabulku s marginálními četnostmi $n_{1.}, n_{2.}, n_{.1}, n_{.2}$, je roven

$$\binom{n}{n_{1.}}\binom{n}{n_{.1}} = \frac{n!\,n!}{n_{1.}!\,n_{2.}!\,n_{.1}!\,n_{.2}!},$$

protože počet možností, jak rozdělit všechny prvky podle zkoumaného znaku X na četnosti $n_{1.}$ a $n_{2.}$, je dán

$$\binom{n}{n_{1.}} = \frac{n!}{n_{1.}!\,(n-n_{1.})!} = \frac{n!}{n_{1.}!\,n_{2.}!}$$

a obdobně pak počet možností, jak rozdělit všechny prvky podle zkoumaného znaku Y na četnosti $n_{.1}$ a $n_{.2}$, je dán

$$\binom{n}{n_{.1}} = \frac{n!}{n_{.1}!\,(n-n_{.1})!} = \frac{n!}{n_{.1}!\,n_{.2}!}.$$

Celkový počet možností, že při třídění n prvků do čtyřpolní tabulky dostaneme jednotlivé četnosti $n_{11}, n_{12}, n_{21}, n_{22}$, je roven

$$\binom{n}{n_{11}}\binom{n-n_{11}}{n_{12}}\binom{n-n_{11}-n_{12}}{n_{21}}1 = \frac{n!}{n_{11}!\,n_{12}!\,n_{21}!\,n_{22}!}$$

protože nejprve se budou realizovat četnost n_{11} a to tak, že počet všech možností je

$$\binom{n}{n_{11}} = \frac{n!}{n_{11}!\,(n-n_{11})!},$$

dále četnost n_{12}, pro kterou ve druhém políčku zbývá $n - n_{11}$ prvků k realizaci, proto

$$\binom{n-n_{11}}{n_{12}} = \frac{(n-n_{11})!}{n_{12}!\,(n-n_{11}-n_{12})!}.$$

Pro četnost n_{21} zbývá už jen $n - n_{11} - n_{12}$ prvků a existuje právě

$$\binom{n-n_{11}-n_{12}}{n_{21}} = \frac{(n-n_{11}-n_{12})!}{n_{21}!\,(n-n_{11}-n_{12}-n_{21})!} = \frac{(n-n_{11}-n_{12})!}{n_{12}!\,n_{22}!}$$

možností, jak se bude při třídění výběru o rozsahu n realizovat. Pro četnost n_{22} zbývá nakonec už jen jediná možnost.

Výsledná podmíněná pravděpodobnost p, že při daných marginálních četnostech $n_{1.}, n_{2.}, n_{.1}, n_{.2}$ vznikne tabulka s četnostmi $n_{11}, n_{12}, n_{21}, n_{22}$, je pak rovna podílu

$$p = \frac{\dfrac{n!\,n!}{n_{1.}!\,n_{2.}!\,n_{.1}!\,n_{.2}!}}{\dfrac{n!}{n_{11}!\,n_{12}!\,n_{21}!\,n_{22}!}},$$

po drobných úpravách je výsledná pravděpodobnost ve tvaru

$$p = \frac{n_{1.}!\,n_{2.}!\,n_{.1}!\,n_{.2}!}{n!\,n_{11}!\,n_{12}!\,n_{21}!\,n_{22}!}.$$

[1]

Provedení testu závisí na tom, zda jde o test jednostranný nebo oboustranný. V literatuře [5] se lze setkat s několika variantami řešení oboustranného testu.

Při provedení jednostranného testu se nejprve vypočítá pravděpodobnost p pro danou tabulku a následně pro všechny tabulky, které vzniknou postupným snižováním nejmenší z četností o jedničku, při zachování hodnot marginálních četností. Výsledná pravděpodobnost je rovna součtu jednotlivých pravděpodobností. V případě, že hodnota $p \leq \alpha$, hypotéza se zamítá.

Pro oboustranný test hypotézy se nejprve provádí jednostranný test. Poté porovnáme marginální četnost v řádku i v sloupci u nejmenší z četností a zvolíme jako tzv. linii tu možnost, která má menší marginální četnost. V dalším kroku prohodíme ve zvolené linii četnosti a postupujeme obdobně jako při jednostranném testu, kdy snižujeme menší četnost o jedničku až k nule, přičemž se zachovávají marginální četnosti. Výsledná pravděpodobnost je rovna součtu jednotlivých pravděpodobností vzniklých tabulek a pokud $\leq \alpha$, hypotézu zamítáme.

Při druhé možností provedení oboustranného testu nejprve v linii na jednom místě zvolíme nulu a při zachování původních marginálních četností zvyšujeme hodnotu četnosti na tomto místě o jedničku až do chvíle, kdy součet pravděpodobností je $\leq \alpha/2$. Tento postup opakujeme stejně při zvolení nuly na druhém místě vybraného řádku či sloupce. Nulovou hypotézu zamítneme v případě, že se některá z vytvořených tabulek při zvyšování četnosti shoduje s původní kontingenční tabulkou.

Jinou variantou je pak postupné zvyšování jednoho místa v linii o jedničku až do chvíle, kdy se na druhém místě neobjeví nula. Jednotlivé pravděpodobnosti takto vzniklých tabulek seřadíme vzestupně a sečteme všechny nejmenší pravděpodobnosti, jejichž součet nepřekročí hodnotu $\propto$. Nulovou hypotézu zamítneme v případě, že do tohoto součtu je zahrnuta i pravděpodobnost původní tabulky.

Čtvrtou variantou je metoda, kdy v linii snižujeme nejmenší četnost o jedničku, jako v případě jednostranného testu, a výsledný součet pravděpodobností vzniklých tabulek se označí S. Dalším krokem je zvolení hodnoty nula na druhém místě v linii a postupné zvyšování této četnosti o jedničku. Končíme ve chvíli, kdy je součet S' pravděpodobností těchto tabulek $\leq S$. Nulovou hypotézu zamítáme, pokud $S + S' \leq \alpha$. [5]

4.2 ZÁVISLOST PLNOSTI ŽALUDKU S MNOŽSTVÍM NAKOUPENÝCH NEPOTŘEBNÝCH POTRAVIN

Říká se, že najezený člověk nakoupí v supermarketu daleko míň potravin než člověk nenajezený. Ten totiž není schopný objektivně odhadnout, jaké množství potravin potřebuje a mnohdy nakoupí víc, než je reálně schopný zkonzumovat do data spotřeby. Najezený člověk v obchodě ušetří i několik stovek korun.

Po dobu dvou měsíců jsme se sestrou každá udělaly větší nákup v obchodě, z toho 5x jsme byly najezené a 5x hladové a sledovaly jsme, kolikrát jsme nakoupily více jak 5 nepotřebných potravin, které bychom normálně nekoupily, a utratily celkově víc peněz. Cílem zkoumání bylo zjistit, zda opravdu spolu souvisí nakoupené množství jídla s prázdným či plným žaludkem, zda nakoupíme více nepotřebných potravin hladoví a aplikovat následně výsledky v životě. Znak *X* značí míru plnosti žaludku, najezená nebo hladová před cestou do supermarketu. Znak *Y* pak označuje, zda byly nakoupeny nadbytečné potraviny. Případ, kdy množství nepotřebných potravin byl menší než 5 a kdy množství nakoupených nepotřebných potravin bylo 5 a více.

Výsledky jsou zpracovány do čtyřpolní kontingenční tabulky 4.1.

Tabulka 4.1 Údaje o nakoupených, nepotřebných potravinách

$X \backslash Y$	<5 potravin	>5 potravin	Σ
NAJEZENÁ	9	1	10
HLADOVÁ	3	7	10
Σ	12	8	20

Vzhledem k malým četnostem je vhodné použití oboustranného Fisherova faktoriálového testu, neboť aproximační podmínka dostatečných teoretických četností není splněna, speciálně

$$\frac{80}{20} = 4 \not\geq 5.$$

Logickým řešením je volba jednostranného testu, vzhledem k tomu, že nás zajímá, zda hladový člověk nakoupí více nepotřebných potravin než najezený. Pro názornost však ukážeme použití i různých variant oboustranného testu z kapitoly 3.4.

Dle prvního způsobu začneme nejprve jednostranným testem a určíme pravděpodobnost výchozí tabulky, která je rovna

$$p_1 = \frac{10!\,10!\,12!\,8!}{20!\,9!\,1!\,3!\,7!} = 0{,}009526.$$

Postupně snižujeme nejmenší četnost o jedničku až k nule. Po snížení dostaneme tabulku 4.2.

Tabulka 4.2. Kontingenční tabulka se sníženou četností

$X\backslash Y$	<5 POTRAVIN	>5 POTRAVIN	Σ
NAJEZENÁ	10	0	10
HLADOVÁ	2	8	10
Σ	12	8	20

Pravděpodobnost je rovna

$$p_2 = \frac{10!\,10!\,12!\,8!}{20!\,10!\,0!\,2!\,8!} = 0{,}000357.$$

Po záměně četností v druhém sloupci, dostaneme tabulku 4.3.

Tabulka 4.3 Kontingenční tabulka po záměně četnosti

$X\backslash Y$	<5 POTRAVIN	>5 POTRAVIN	Σ
NAJEZENÁ	3	7	10
HLADOVÁ	9	1	10
Σ	12	8	20

Hodnota pravděpodobnosti je pak

$$p_3 = \frac{10!\,10!\,12!\,8!}{20!\,3!\,7!\,9!\,1!} = 0{,}009526.$$

Dále snižujeme opět četnost o jedničku k nule, tabulka 4.4.

Tabulka 4.4 Kontingenční tabulka po záměně a snížení četnosti

$X\backslash Y$	<5 POTRAVIN	>5 POTRAVIN	Σ
NAJEZENÁ	2	8	10
HLADOVÁ	10	0	10
Σ	12	8	20

Odpovídající pravděpodobnost je rovna

$$p_4 = \frac{10!\,10!\,12!\,8!}{20!\,2!\,8!\,10!\,0!} = 0{,}000357.$$

Součtem všech čtyř pravděpodobností $p_1 + p_2 + p_3 + p_4$, po zaokrouhlení na dvě desetinná místa, je hodnota $P = 0{,}02 < 0{,}05$, tedy nulovou hypotézu o nezávislosti míry plnosti žaludku a množství nakupených nadbytečných potravin na uvedené hladině testu zamítáme. Výsledek provedeného testování je tedy ve shodě s již dříve provedenými studiemi.

Postupně ukážeme i alternativní způsoby řešení oboustranných testů, postupně, jak byly zmiňovány v kapitole 3.4.

V linii je zvolena nula na prvním řádku a vypočtena pravděpodobnost p_1

$$p_1 = \frac{10!\,10!\,12!\,8!}{20!\,10!\,0!\,2!\,8!} = 0{,}000357.$$

Už při prvním zvýšení dostáváme původní kontingenční tabulku 4.1, při druhém zvýšení pak příslušná pravděpodobnost překračuje v součtu hodnotu $\alpha/2$ a již se nezapočítává, viz tabulka 4.5.

Tabulka 4.5 Soubor kontingenčních tabulek 2. varianty

X\Y	<5	>5	**Σ**
N	10	0	10
H	2	8	10
Σ	12	8	20
$p_1 = 0{,}000357$			

X\Y	<5	>5	**Σ**
N	9	1	10
H	3	7	10
Σ	12	8	20
$p_2 = 0{,}009526$			

X\Y	<5	>5	**Σ**
N	8	2	10
H	4	6	10
Σ	12	8	20
$p_3 = 0{,}07502$			

$$p_1 + p_2 = 0{,}000357 + 0{,}009526 = 0{,}009883$$

Jelikož se při zvyšování četnosti jedna z vytvořených tabulek shoduje s původní tabulkou 4.1, nulovou hypotézu zamítáme.

V třetí možnosti zvyšujeme v linii v prvním řádku četnost o 1 a dostaneme postupně tabulky, které jsou shrnuty v tabulce 4.6.

Tabulka 4.6 Soubor kontingenčních tabulek 3. varianty

$X\backslash Y$	<5	>5	**Σ**
N	8	2	10
H	4	6	10
Σ	12	8	20
$p_1 = 0{,}07502$			

$X\backslash Y$	<5	>5	**Σ**
N	7	3	10
H	5	5	10
Σ	12	8	20
$p_2 = 0{,}24006$			

$X\backslash Y$	<5	>5	**Σ**
N	6	4	10
H	6	4	10
Σ	12	8	20
$p_3 = 0{,}35008$			

$X\backslash Y$	<5	>5	**Σ**
N	5	5	10
H	7	3	10
Σ	12	8	20
$p_4 = 0{,}24006$			

$X\backslash Y$	<5	>5	**Σ**
N	4	6	10
H	8	2	10
Σ	12	8	20
$p_5 = 0{,}07502$			

$X\backslash Y$	<5	>5	**Σ**
N	3	7	10
H	9	1	10
Σ	12	8	20
$p_6 = 0{,}0009526$			

$X\backslash Y$	<5	>5	**Σ**
N	10	0	10
H	2	8	10
Σ	12	8	20
$p_7 = 0{,}000357$			

Seřazením pravděpodobností

$$p_7 = 0{,}000357$$
$$p_6 = 0{,}009526$$
$$\vdots$$
$$p_3 = 0{,}35008$$

vidíme, že v součtu prvních dvou pravděpodobností

$$p_7 + p_6 = 0{,}000357 + 0{,}009526 = 0{,}009883,$$

které nepřekročí hodnotu α, je zahrnuta i pravděpodobnost původní kontingenční tabulky 4.1, proto nulovou hypotézu zamítáme.

Čtvrtý způsob vede k vytvoření souboru tabulek 4.7. Hodnoty součtů jsou rovny

$$S = 0{,}000357$$
$$S' = 0{,}000357$$

a jejich součet

$$S + S' = 0{,}000714 \leq 0{,}05,$$

proto nulovou hypotézu zamítáme.

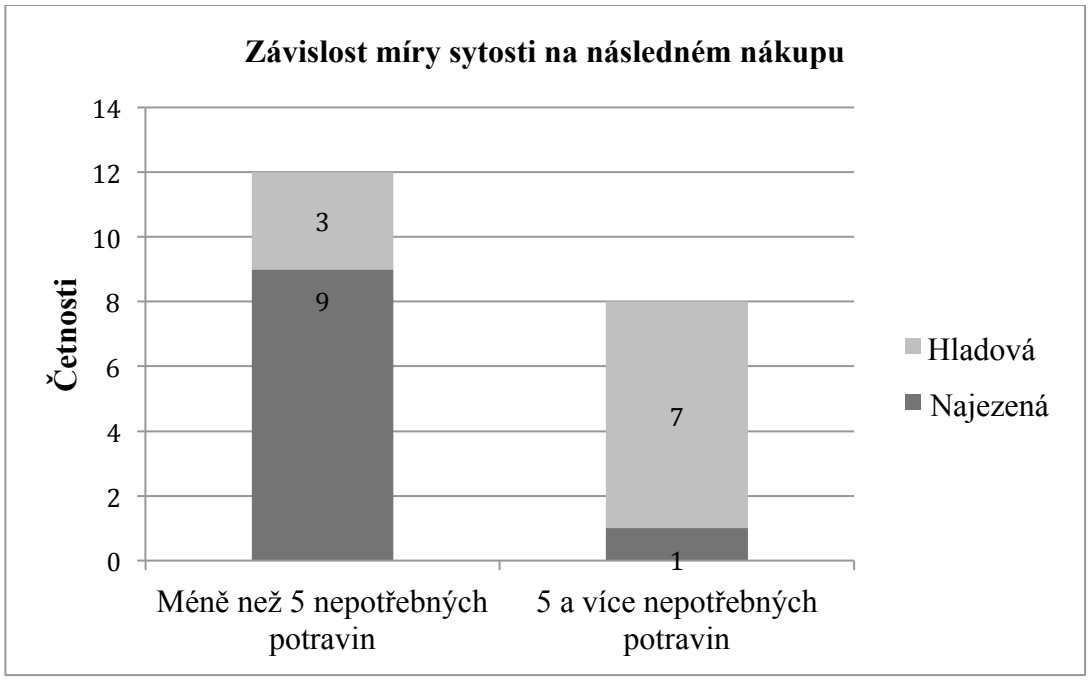
Tabulka 4.7 Soubor kontingenčních tabulek 4. varianty

X\Y	<5	>5	**Σ**
N	10	0	10
H	2	8	10
Σ	12	8	20

X\Y	<5	>5	**Σ**
N	8	2	10
H	4	6	10
Σ	12	8	20

Naměřená data jsou zpracována do kumulativního sloupcového grafu 4.1.

Graf 4.1 Zaznamenání nakoupených, nepotřebných potravin



5 McNEMARŮV TEST

McNemarův test je testem symetrie pro čtyřpolní tabulku. Zaměřuje se na pozorování, u kterých se při opakovaném měření vyskytují rozdílné výsledky. Testovat takto můžeme např. zda proběhlá politická kampaň ovlivnila rozhodnutí voličů.

5.1 TEORETICKÁ VÝCHODISKA

V případě, kdy sledujeme, zda provedený zásah má nebo nemá vliv na pravděpodobnost výskytu určitého znaku, se používá McNemarův test. Vybere se náhodně soubor n statistických jednotek a zjistí se podíl těch jednotek, u kterých se vyskytoval sledovaný znak. Následně se provede určitý zásah, který by mohl pravděpodobnost výskytu tohoto znaku ovlivnit, a znova se na celém souboru zjistí, jaký je podíl těch jednotek, u kterých se znak vyskutuje. Na základě této změny je třeba rozhodnout, zda provedený zákrok ovlivňuje pravděpodobnost výskytu či nikoliv.

Je nutné mít získaná data uspořádaná do tvaru čtyřpolní tabulky 5.1, kde symbolem **(+)** označuje výskyt sledovaného znaku a symbol **(-)** případ, kdy se znak nevyskystl.

Tabulka 5.1 Četnosti pro McNemarův test

Před zásahem / Po zásahu	**+**	**-**	**Σ**
+	n_{11}	n_{12}	$n_{1.}$
-	n_{21}	n_{22}	$n_{2.}$
Σ	$n_{.1}$	$n_{.2}$	n

Jednotlivé četnosti můžeme považovat za výběr o rozsahu n z multinomického rozdělení o parametrech $n, p_{11}, p_{12}, p_{21}, p_{22}$. Pravděpodobnosti jsou vypsány v tabulce 5.2.

Tabulka 5.2 Pravděpodobnosti pro McNemarův test

Před zásahem / Po zásahu	**+**	**-**	**Σ**
+	p_{11}	p_{12}	$p_{1.}$
-	p_{21}	p_{22}	$p_{2.}$
Σ	$p_{.1}$	$p_{.2}$	1

Je třeba testovat hypotézu $H_0: p_{12} = p_{21}$ proti oboustranné alternativě $H_A: p_{12} \neq p_{21}$. Nulová hypotéza tedy tvrdí, že pravděpodobnost toho, že před zásahem se daný znak vyskytl a po již nikoliv je rovna pravděpodobnosti, že před zásahem se znak nevyskytl a po zásahu ano.

V důsledku symetrie je také shoda i u marginálních pravděpodobností, tedy shody pravděpodobnosti výskytu před zásahem a výskytu po zásahu. Hypotézu lze proto ekvivalentně napsat i ve tvaru $H_0: p_{1.} = p_{.1}$ nebo také $H_0: p_{2.} = p_{.2}$ [3]

Obdobně jako v kapitole 2 Test nezávislosti, vycházíme při testování z testů dobré shody, viz kapitola 1. Vzhledem k malé, čtyřpolní, tabulce máme však o to sníženou náročnost provedení následujících úvah. Namísto vektoru náhodných veličin $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$, kde jednotlivé veličiny $X_i, i = 1, \ldots, k$, mají multinomické rozdělení, máme nyní matici četností n_{ij}, $i = 1, 2$, $j = 1,2$.

Vektor volných parametrů

$$\boldsymbol{a} = (p_{1.}, p_{.1})$$

obsahuje jen dvě marginální pravděpodobnosti, neboť zbylé dvě $p_{2.}, p_{.2}$ jsou jednoznačně určeny podmínkou součtů marginálních pravděpodobností

$$p_{2.} = 1 - p_{1.}$$

a

$$p_{.2} = 1 - p_{.1}$$

Souvislosti mezi obecnou teorií kontingenčních tabulek a situací při McNemarově testu shrnuje tabulka 2.3. Poslední dva řádky se vztahují k testu nezávislosti.

Tabulka 5.3 Tabulka souvislostí s McNemarovým testem

Testy dobré shody	Kontingenční tabulky
X_i	n_{ij}
$i = 1, \dots, k$	$i = 1,2 \;\; j = 1,2$
$X_1, \dots, X_k$	$n_{11}, n_{12}, n_{21}, n_{22}$
vektor $\boldsymbol{X}$	matice n_{ij}
n	n
k	4
p_i	p_{ij}
$\sum_{i=1}^{k} p_i = 1$	$\sum_{i=1}^{2}\sum_{j=1}^{2} p_{ij} = 1$
$\sim\chi^2_{k-1}$	$\sim\chi^2_{4-1} = \chi^2_3$
$Z = \sum_{j=1}^{k} \frac{(X_j - np_j)^2}{np_j}$	$\chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$
$\sim\chi^2_{k-1-m}$	$\sim\chi^2_{4-1-2} = \chi^2_1$

Řešením zjednodušené soustavy rovnic (5) dostaneme (blíže [2], str. 273)

$$\frac{n_{11}}{p_{11}} - \frac{n_{22}}{p_{22}} = 0$$

$$\frac{n_{12}}{p_{12}} + \frac{n_{21}}{p_{21}} - 2\frac{n_{22}}{p_{22}} = 0.$$

Odtud vyjádřením n_{11} z první rovnice

$$n_{11} = \frac{n_{22}}{p_{22}} p_{11}$$

a úpravou druhé rovnice

$$n_{12} + n_{21} = 2\frac{n_{22}}{p_{22}} p_{12}.$$

Připojením triviální rovnice $n_{22} = \frac{n_{22}}{p_{22}} p_{22}$ a následným sečtením se dostane

$$n_{11} + n_{12} + n_{21} + n_{22} = \frac{n_{22}}{p_{22}} p_{11} + 2\frac{n_{22}}{p_{22}} p_{12} + \frac{n_{22}}{p_{22}} p_{22}$$

$$n = \frac{n_{22}(p_{11} + 2p_{12} + p_{22})}{p_{22}} = \frac{n_{22}(p_{11} + 2p_{12} + 1 - p_{11} - 2p_{12})}{p_{22}}$$

$$n = \frac{n_{22}}{p_{22}}.$$

Pro odhady jednotlivých pravděpodobností pak platí

$$\hat{p}_{22} = \frac{n_{22}}{n},$$

$$\hat{p}_{11} = \frac{n_{11}}{n},$$

$$\hat{p}_{12} = \frac{n_{12} + n_{12}}{2n},$$

$$\hat{p}_{21} = \hat{p}_{12}.$$

Dosazením odhadů pravděpodobností je veličina χ^2

$$\chi^2 = \sum_{1}^{2} \frac{(n_{11} - n\hat{p}_{11})^2}{n\hat{p}_{11}} + \frac{(n_{12} - n\hat{p}_{12})^2}{n\hat{p}_{12}} + \frac{(n_{21} - n\hat{p}_{21})^2}{n\hat{p}_{21}} + \frac{(n_{22} - n\hat{p}_{22})^2}{n\hat{p}_{22}} =$$

$$= \frac{\left(n_{11} - n\frac{n_{11}}{n}\right)^2}{n\hat{p}_{11}} + \frac{\left(n_{12} - n\frac{n_{12} + n_{12}}{2n}\right)^2}{n\frac{n_{12} + n_{12}}{2n}} + \frac{\left(n_{21} - n\frac{n_{12} + n_{12}}{2n}\right)^2}{n\frac{n_{12} + n_{12}}{2n}} +$$

$$+ \frac{\left(n_{22} - n\frac{n_{22}}{n}\right)^2}{n\frac{n_{22}}{n}}.$$

První a poslední výraz je roven nule. Následnými úpravami pak

$$\frac{{n_{12}}^2 - n_{12}(n_{12} + n_{12}) + \frac{(n_{12} + n_{12})^2}{4}}{\frac{n_{12} + n_{12}}{2}} + \frac{{n_{21}}^2 - n_{21}(n_{12} + n_{12}) + \frac{(n_{12} + n_{12})^2}{4}}{\frac{n_{12} + n_{12}}{2}} =$$

$$= \frac{-n_{12}n_{21} + \frac{{n_{12}}^2 + n_{12}n_{21} + {n_{21}}^2}{4} - n_{12}n_{21} + \frac{{n_{12}}^2 + n_{12}n_{21} + {n_{21}}^2}{4}}{\frac{n_{12} + n_{12}}{2}} =$$

$$= \frac{-4n_{12}n_{21} + {n_{12}}^2 + 2n_{12}n_{21} + {n_{21}}^2}{n_{12} + n_{12}} = \frac{{n_{12}}^2 - 2n_{12}n_{21} + {n_{21}}^2}{n_{12} + n_{12}}.$$

$$\frac{{n_{12}}^2 - 2n_{12}n_{21} + {n_{21}}^2}{n_{12} + n_{12}} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}.$$

Veličina

$$\chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

má za platnosti nulové hypotézy asymptoticky χ^2 rozdělení o jednom stupni volnosti. Kritický obor $W = \langle \chi^2_{1,1-\alpha}, \infty)$ pro test na hladině α.

Tento asymptotický výsledek je použitelný při splnění podmínky dostatečných četností $n_{12} + n_{21} \geq 8$.

5.2 ZÁVISLOST PŘIJETÍ DO NOVÉHO ZAMĚSTNÁNÍ A TETOVÁNÍ

Lidé v dnešní době hledají různé způsoby, jak vyjádřit sebe sama a chtějí něčím vynikat a jednou z aktuálně rozšiřovanou možností, jak tohoto docílit je tetování. Na jedné straně tetování je už dnes vnímáno jako naprosto běžná věc a lidé jsou vůči němu daleko tolerantnějšími než dřív, na druhé však stále existují určité předsudky vůči potetovaným lidem, převažně u starších generací, a vnímají jej ve spojení s kriminalitou, námořnictvem a nevzdělaností. Na základě průzkumu, dostupného na stránkách vyplnto.cz, Natálie Majvaldové [6], lze konstatovat, že tyto předsudky, vůči potetovaným lidem, avšak myšleno jen na „seriózních postech“ (doktor, právník, politik), nikoliv všeobecně, má necelých 30 % dotazovaných respondentů v počtu 412, viz koláčový graf 5.1.

Graf 5.1 Předsudky k tetování

Doba, kdy se tetovaly především tyto skupiny lidí, je však dávno pryč a v současnosti se tetování jeví spíš jako módní trend, prostředek sebevyjádření a dá se hovořit i o umění. Jedním z důvodů ale, proč se člověk někdy bojí a má respekt nechat si něco vytetovat, je pravděpodobný problém při hledání nové práce. Je všeobecně známo, že lidé považují tetování za součást neprofesionálního vzhledu, obzvlášť pokud jsou na první pohled viditelná, na rukách, krku nebo obličeji.

U 30 náhodně vybraných lidí, zaměstnanců, zaměstnavatelů i studentů, z nichž většina studuje obor Management jakosti, jsem provedla následující výzkum. Měli si představit, že jsou v pozici personalisty ve společnosti FRISCHBETON s.r.o., dceřiné společnosti STRABAG, která vyrábí betonové směsi v provozovnách v Čechách i na Moravě a hledají na pozici manažera pro jakost pro Moravskoslezský kraj nového zaměstnance. Vzhledem k osobním zkušenostem a znalostem ze studia tito lidé ví, jakých kvalit je třeba pro takovou pozici.

Na základě představeného životopisu měli rozhodnout, zda by o takového člověka stáli, či nikoliv. Tento člověk splňoval veškeré požadavky na vzdělání, praxi, znalosti a dovednosti. Existoval pouze jeden malý háček. Tento zájemce o práci měl tetování na rukách, o čemž v první chvíli respondenti nevěděli. Po zjištění, zda by na základě životopisu rozhodli o přijetí či nepřijetí, byla reprezentativnímu vzorku ukázána fotografie uchazeče a měli znova rozhodnout, zda by jej i tak přijali/nepřijali, či změnili svůj prvotní názor.

Cílem průzkumu bylo zjistit, zda ukázka fotografie potetovaného zájemce ovlivňuje pravděpodobnost přijetí na pozici, na kterou se všemi aspekty hodí. Pro takto formulovanou závislost je vhodné použití McNemarova testu.

Nejprve zapíšeme výsledky popsaného zjišťování do čtyřpolní tabulky, kde symbol $(+)$ značí, přijetí uchazače, symbol $(-)$ značí jeho nepřijetí, viz následující tabulka 5.4.

Tabulka 5.4 Přijetí uchazeče do zaměstnání

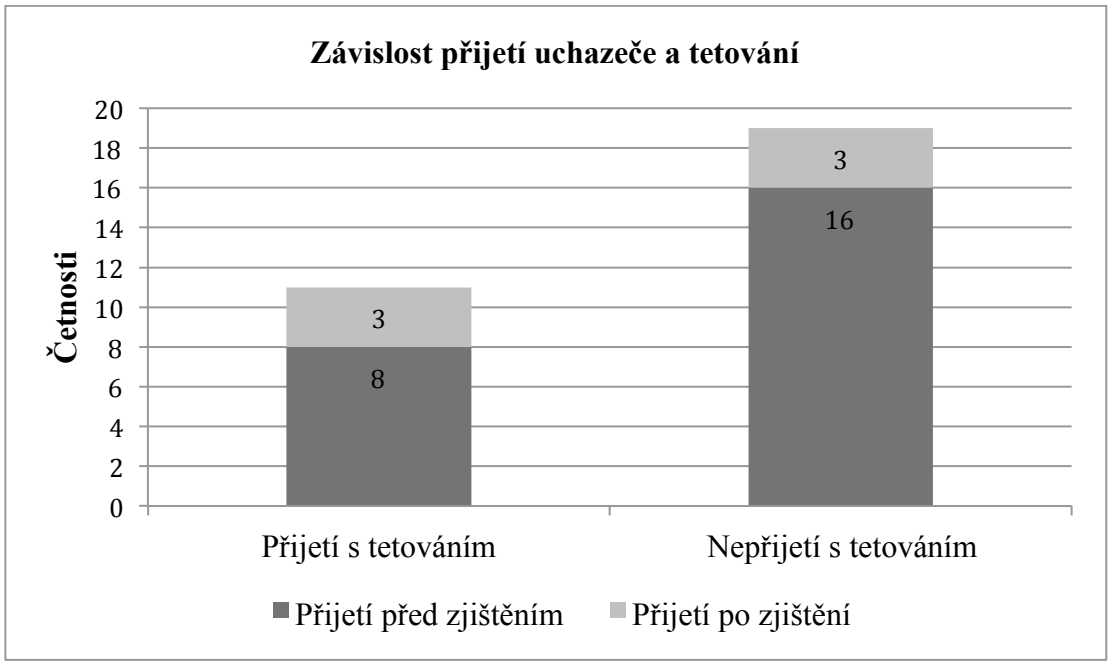
Před\Po	+	-	Σ
+	8	16	24
-	3	3	6
Σ	11	19	30

Protože podmínka dostatečné četnosti $16 + 3 \geq 8$ je splněna, hodnota testovací statistiky je rovna

$$z = \frac{(16-3)^2}{16+3} = 8{,}895.$$

Porovnáním s kvantilem $\chi^2_{1;0,95} = 3{,}84$ dojdeme k závěru, že hypotézu na základě provedeného výběru musíme zamítnout. Z výsledku plyne, že mezi tetováním a přijetím do zaměstnání existuje závislost. Je tedy pravděpodobné, že potetovaní lidé budou při hledání nové práce částečně handicapováni, pokud zaměstnavatel zjistí, že je uchazeč potetován nebo mu je předem tato skutečnost oznámena. Graficky je situace zaznamenána v grafu 5.2.

Graf 5.2 Zaznamenání dat o přijetí potetovaného uchazeče



6 SROVNÁNÍ JEDNOTLIVÝCH METOD

V kontingenčních tabulkách můžeme na základě různých faktorů, ať už to bude zvolená nulová hypotéza, či rozsah naměřených dat, testovat závislost a homogenitu pomocí testů, které jsou zpracovány v kapitolách 2-5.

K testům hypotéz, jak o nezávislosti, tak symetrii, se používá stejná testová statistika se stejným limitním rozdělením. U nezávislosti uvažujeme náhodný výběr z jedné populace, kdy sledujeme u každé zkoumané jednotky z výběru výskyt nominálního znaku. Marginální četnosti jsou proto náhodné. U testů homogenity tyto marginální četnosti jsou pevně zadány a provádí se výběr z několika populací, např. ženy a muži, u kterých se sleduje zkoumaný nominální znak nabývající několika hodnot. Zajímá nás, zda jsou pravděpodobnosti výsledků u jednotlivých populací stejné.

Testováním nezávislosti tedy například dojdeme k závěru, zda má nebo nemá pohlaví vliv na pravorukost či levorukost jedince, kdežto u homogenity pak, zda je rozložení pravorukých a levorukých žen stejné jako u mužů či nikoliv, tj. zda je struktura shodná.

Fisherův faktoriálový test používáme v situaci, kdy máme málo naměřených dat, ať už z důvodu nákladnosti či třeba nemožnosti provést další měření. Na základě alternativní hypotézy rozhodneme, zda je vhodné provést jednostranný nebo oboustranný test, který má různé varianty řešení.

Pokud nás zajímá, zda měl nějaký provedený zákrok vliv na výskyt daného znaku, např. nás zajímá, zda televizní reklama měla vliv na nákup určitého produktu, používáme McNemarův test.

ZÁVĚR

Hlavním cílem práce bylo shrnutí nejčastěji používaných testů v kontingenčních tabulkách a aplikace těchto metod na vlastních datech, která byla pro tyto účely získána různými způsoby, tj. formou dotazníku, vlastního měření aj.

Nápad o testování závislosti kouření rodičů a jejich dětí popř. vlivu kouření marihuany u kamarádů, vznikl při doprovázení mladší sestry do školy a znepokojivém pohledu na početné cigarety v rukách dětí. Výsledky testování nebyly nijak překvapující. U obou případů byla zjištěna závislost, avšak při kouření marihuany kamarádů nejsme do jisté míry schopni říci, co je v tomto případě příčinou a co následkem a kterým směrem se pohybujeme, na rozdíl od kouření rodičů, kdy je málo pravděpodobné, aby následkem kouření dítěte kouřili rodiče, tady jde o závislost jedním směrem.

Na základě vlastního pocitu o špatných výsledcích ranních běhů byla zpracována hypotéza o symetrii, která však vedla k závěru, že ranní a odpolední běhání dává z hlediska času statisticky nevýznamný rozdíl. Symetrie byla dále testována i pro data získána díky spolupráci s výčepní z jedné pražské hospody, kdy prostřednictvím poskytnutého archu zaznamenala, jaké druhy piva si objednávali muži a jaké ženy. U obou testů bylo následně provedeno zvětšení rozsahu, přidáním dalších běhů po měsíci, a zmenšení výběru mužů a žen na čtvrtinu. Rozhodnutí o nulové hypotéze vedlo díky těmto změnám k opačným, byť v případě piv k těsnému, výsledkům.

Cílem aplikace Fisherova faktoriálového testu bylo zjistit, zda náhodou není výhodnější nakupovat potraviny s plným žaludkem, což se taky na základě vlastních nákupů a nákupů mé sestry potvrdilo.

Poslední provedený test je všeobecně aktuálním tématem. Zkoumaná závislost tetování na přijetí do nového zaměstnání byla potvrzena prostým rozhovorem s vybranými lidmi znalými v oboru, kdy po předložení životopisu rozhodovali o přijetí či nepřijetí uchazeče. Ukázanou fotografií většina změnila na daného člověka názor a z výzkumu je patrné, že by potetovaní zájemci o práci mohli mít v budoucnu tímto jistý handicap.

Předem vytyčené cíle byly splněny, zjištěny a zaznamenány. Práce může sloužit jako informační zdroj pro studenty matematických oborů k doplnění znalostí nebo upřesnění mezer v rámci problematiky testů dobré shody a kontingenčních tabulek, které by mohly být díky praktickým příkladům srozumitelnější. Práce by se dala v budoucnu rozšířit o méně používáné testy, rovněž založené na praktických příkladech použití ad. Mnohé zjištěné závislosti lze v reálném životě využít, popřípadě vedou k jistému zamyšlení či změně dosavadních praktik.

POUŽITÁ LITERATURA

[1] HRON, Karel a Pavla KUNDEROVÁ. *Základy počtu pravděpodobnosti a metod matematické statistiky*. 1. vyd. Olomouc: Univerzita Palackého v Olomouci, 2013, 330 s. ISBN 978-80-244-3396-7.

[2] ANDĚL, Jiří. *Základy matematické statistiky*. Vyd. 3. Praha: Matfyzpress, 2011, 358 s. ISBN 978-80-7378-162-0

[3] PAVLÍK, Jiří. *Aplikovaná statistika*. Praha: Vysoká škola chemicko-technologická, 2005, 172 s. ISBN 80-708-0569-2.

[4] ZVÁRA, Karel. 2004. *Biostatistika*. 2. vyd. Praha: Karolinum, 213 s. ISBN 80-246-0739-5.

[5] ANDĚL, Jiří. *Matematická statistika*. Vyd. 3. Praha: SNTL- Nakladatelství technické literatury, 1978, 352 s..

[6] Majvaldová, N. – *Tetování: trend či cejch?* (výsledky průzkumu), 2012. Dostupné online na http://tetovani-trend-ci-cejch.vyplnto.cz.

PŘÍLOHY

Příloha A

Dotazník kouření žáků 8. a 9. tříd

DOTAZNÍK

Vliv kouření rodičů a kamarádů

Jmenuji se Sára Mihalová a studuji na Univerzitě Palackého v Olomouci obor Aplikovaná matematika. Ráda bych Vás požádala o vyplnění krátkého anonymního dotazníku obsahujícího 4 otázky, který bude následně sloužit jako podklad pro moji bakalářskou práci.

Děkuji.

Odpoveď označte vpravo křížkem v příslušné kolonce.

Kouří cigarety vaši rodiče?	
ANO	
NE	

Kouříš pravidelně?	
ANO	
NE	

Zkoušel jsi kouřit někdy marihuanu?	
ANO, kouřím pravidelně	
již jsem to zkusil/a	
ne, nikdy	

Zkoušeli kouřit marihuanu tví nejbližší kamarádi?	
ANO	
NE	

Příloha B

Životopis uchazeče o práci manažera pro jakost

ŽIVOTOPIS - CURRICULUM VITAE

OSOBNÍ ÚDAJE

Jméno a příjmení:	**Ing. Matyáš Nováček, Ph.D., MBA**
Bytem:	Jeremenkova 1, 772 00 Olomouc
E-mail:	matyas.novacek@me.com
Tel.:	+420 737 304 959
Datum narození:	3. 7. 1976
Rodinný stav:	ženatý

PRACOVNÍ ZKUŠENOSTI

2006 – doposud	International Arai Company spol. s r. o. Senior quality manager Úvazek: HPP *Zodpovědnost za řízení a udržování systému jakosti, vedení týmu výrobní a výstupní kontroly, řízení kontroly ve výrobním procesu, analýzy neshod, řízení neshodných výrobků, nápravná a preventivní opatření, řešení reklamací, realizace nových projektů v oblasti výrobní kvality, sledování kvalitativních ukazatelů, školení kontrolních a výrobních pracovníků, analýza trhu, finanční bilance, monitoring, odpovědnost za pracovníky*
2001 - 2006	Quality Production spol. s r. o. Inženýr kvality Úvazek: HPP *Kontrola dodržování standardů kvality a interních směrnic, optimalizace kontrolních procesů, komunikace se zákazníkem, řešení zákaznických a dodavatelských reklamací, spolupráce s výrobou, sledování kvalitativních ukazatelů, reporting*

VZDĚLÁNÍ

2011 – 2013	Business Institut EDU a. s. Studijní program: Master of Business Administration Studijní obor: Management, organizace a řízení ukončeno závěrečnou zkouškou 3/2013, titul MBA
2009	Jazyková škola s právem státní jazykové zkoušky Ostrava Státní jazyková zkouška všeobecná Jazyk: Německý Úroveň: C1

Příloha C

Životopis uchazeče o práci manažera pro jakost - pokračování

2006	Jazyková škola s právem státní jazykové zkoušky Ostrava Státní jazyková zkouška všeobecná Jazyk: Anglický Úroveň: C1
2002 – 2005	Vysoká škola báňská – Technická univerzita Ostrava Fakulta metalurgie a materiálového inženýrství Doktorský studijní program Studijní program: Řízení průmyslových systémů ukončeno obhajobou doktorské disertační práce 6/2004, titul Ph.D.
1999 – 2001	Vysoká škola báňská – Technická univerzita Ostrava Fakulta metalurgie a materiálového inženýrství Magisterské prezenční studium Studijní program: Ekonomika a řízení průmyslových systémů Studijní obor: Management jakosti ukončeno státní závěrečnou zkouškou 6/2001, titul Ing.
1996 - 1999	Vysoká škola báňská – Technická univerzita Ostrava Fakulta metalurgie a materiálového inženýrství Bakalářské prezenční studium Studijní program: Ekonomika a řízení průmyslových systémů Studijní obor: Management jakosti ukončeno státní závěrečnou zkouškou 9/2014, titul Bc.
1992 - 1996	Čtyřleté všeobecné gymnázium Olomouc - Hejčín ukončeno maturitní zkouškou 6/1996

DOVEDNOSTI A ZNALOSTI

Normy/nástroje:	Normy ISO/TS 16 949, ISO 9001, ISO 14 001, nástroje kvality FMEA, APQP. 8D, PPAP, SPC,
Jazykové znalosti:	Anglický jazyk – slovem a písmem, úroveň C1 Německý jazyk – slovem a písmem, úroveň C1
Počítačové znalosti:	OSX, iOS, iWork, MS Windows, MS Office, SAP
Vlastnosti:	Pracovitost, cílevědomost, loajálnost, zodpovědnost, spolehlivost, vytrvalost, komunikativnost, učenlivost, asertivita
Řidičský průkaz:	A, B
Zájmy:	Rodina, běh, IT technologie, Apple

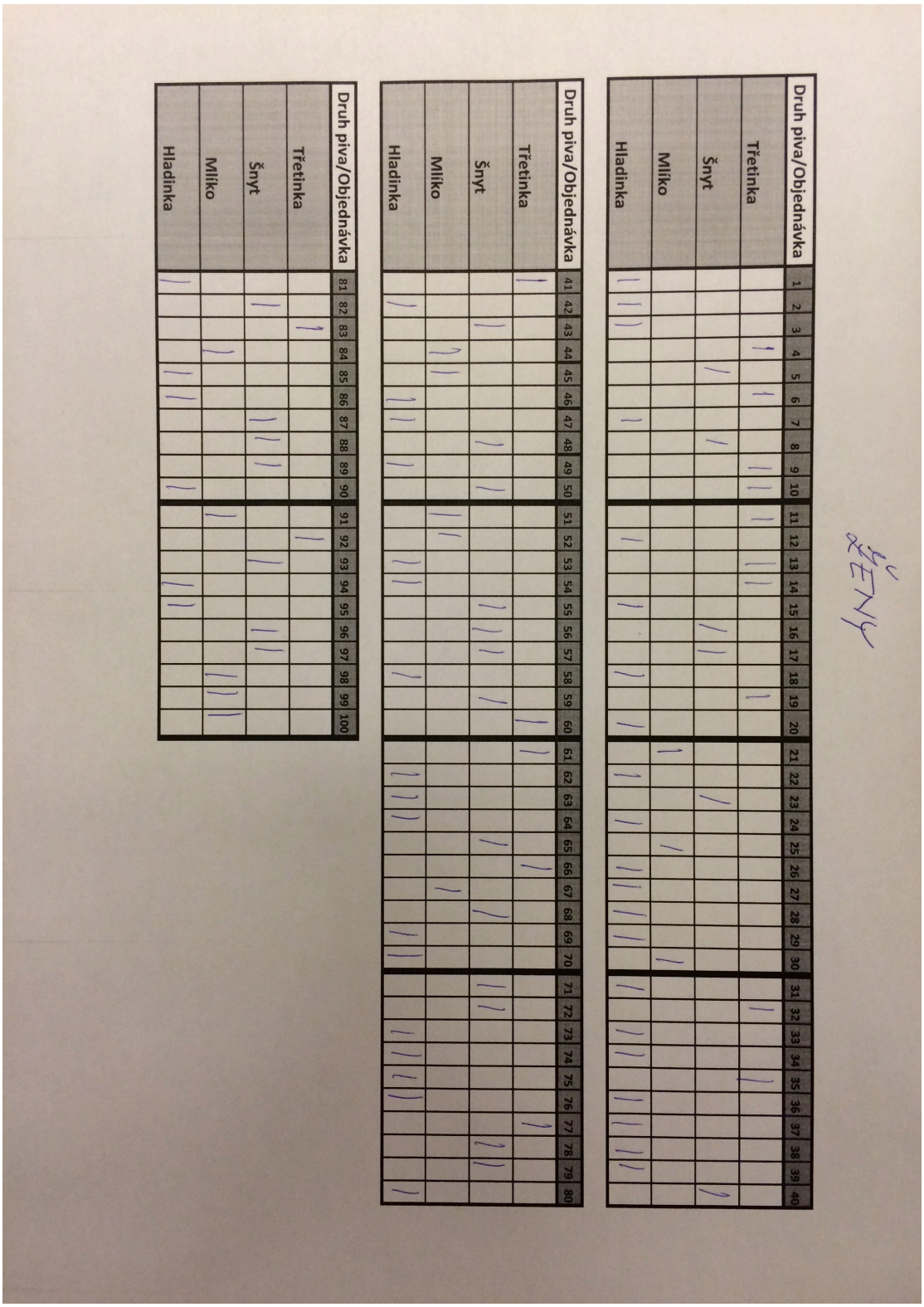
Příloha D

Fotografie uchazeče

Zdroj: http://cdn29.elitedaily.com/wp-content/uploads/2014/10/tattoo.jpg

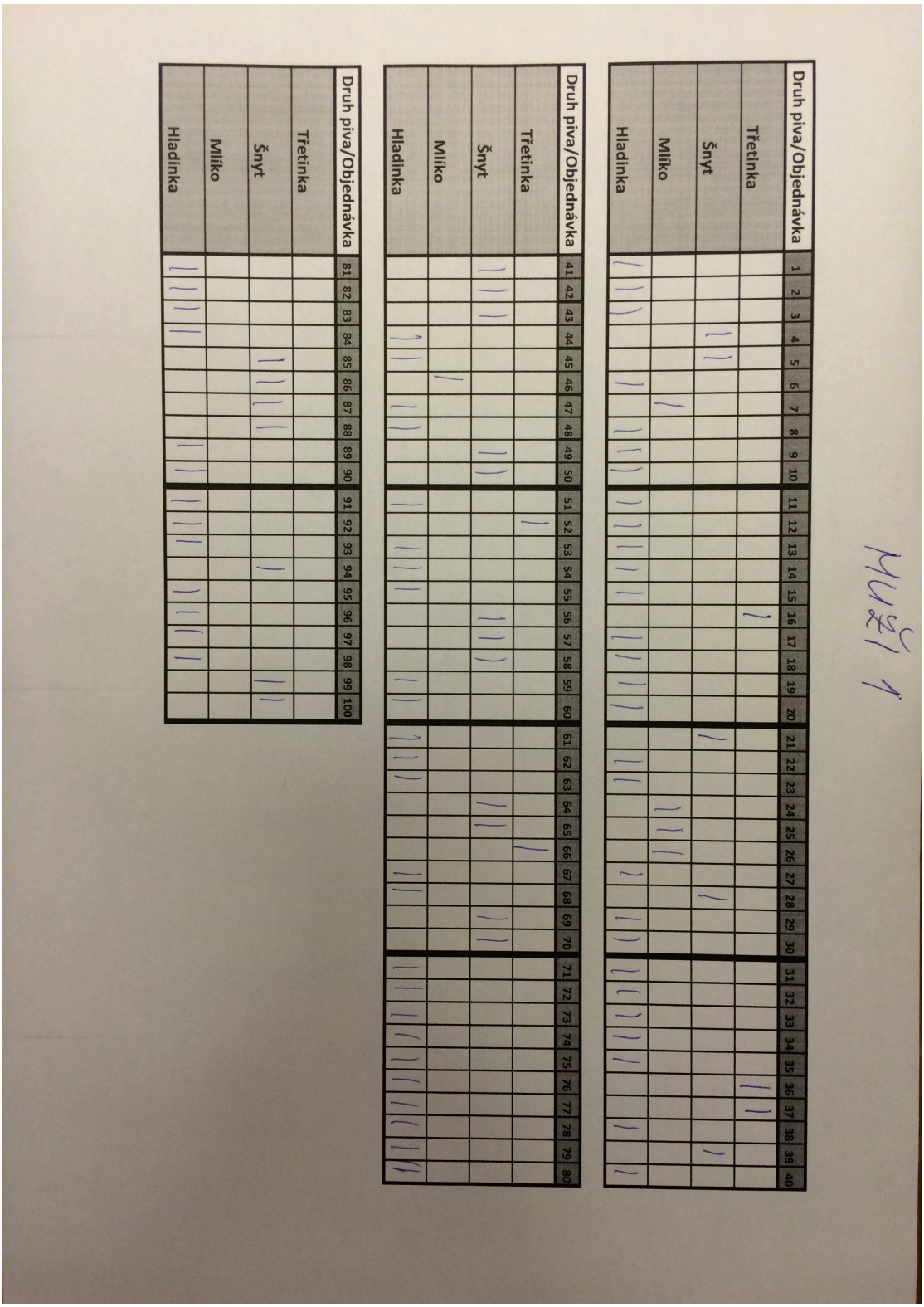
Příloha E

Záznamový arch objednávek piv - ŽENY



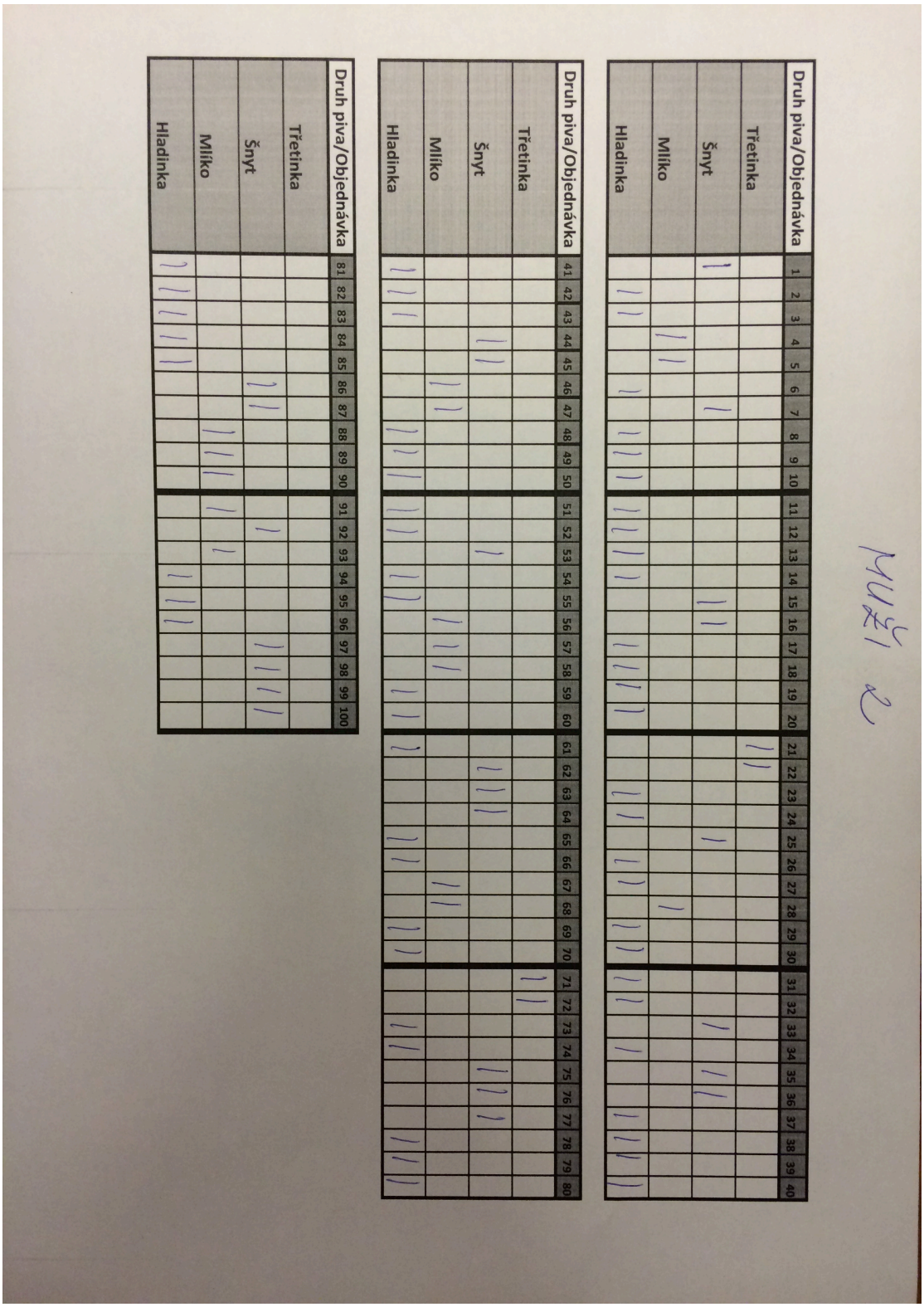
Příloha F

Záznamový arch objednávek piv - MUŽI 1



Příloha G

Záznamový arch objednávek piv - MUŽI 2



MUŽI 2

Druh piva/Objednávka	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Třetinka																					1	1																				
Šnyt	1					1									1	1								1								1	1			1					
Mlíko			1	1																							1													
Hladinka	1	1			1		1	1	1	1	1	1	1	1			1	1	1	1			1	1	1	1		1	1		1	1		1			1	1	1	1

Druh piva/Objednávka	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
Třetinka																															1	1									
Šnyt				1	1							1									1	1	1												1	1	1				
Mlíko						1	1								1	1	1									1	1														
Hladinka	1	1	1					1	1		1	1		1	1			1	1		1			1	1			1	1												

Druh piva/Objednávka	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
Třetinka																				
Šnyt					1	1					1					1	1	1		
Mlíko								1	1		1	1								
Hladinka	1	1	1	1										1	1	1				