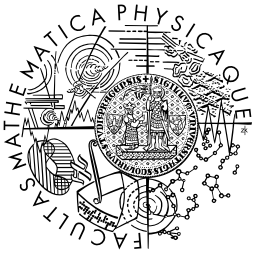


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Martin Jex

Skórové testy v kontingenčních tabulkách

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	doc. Ing. Marek Omelka, Ph.D.
Studijní program:	Matematika
Studijní obor:	Obecná matematika

Praha 2019

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Děkuji mému vedoucímu práce doc. Ing. Marku Omelkovi, Ph.D. za vedení práce, věcné připomínky, vstřícnost, ochotu, trpělivost a cenné rady. Děkuji Fakultě jaderné a fyzikálně inženýrské za zapůjčení Galtonovy desky.

Název práce: Skórové testy v kontingenčních tabulkách

Autor: Martin Jex

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. Ing. Marek Omelka, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Bakalářská práce se zabývá testováním hypotéz v multinomickém rozdělení. Využívá dvou přístupů, Pearsonova přístupu známého jako test dobré shody a přístupu vycházejícího z teorie maximální věrohodnosti. V práci jsou odvozeny testy založené na maximální věrohodnosti. Oba přístupy jsou uplatněny na multinomické rozdělení a to pro případ bez a s rušivými parametry. Také je uvedena souvislost obou přístupů. Dále jsou přístupy použity na reálná data k lepšímu pochopení probírané problematiky.

Klíčová slova: Multinomické rozdělení, maximální věrohodnost, test dobré shody, asymptotické testy

Title: Score tests in contingency tables

Author: Martin Jex

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. Ing. Marek Omelka, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The thesis deals with testing of hypotheses in multinomial distribution. It utilizes two approaches, Pearson's approach known as the of goodness of fit test and the approach stemming from theory of maximum likelihood. The thesis presents derivations of tests based on maximum likelihood. Both approaches are used on the multinomial distribution and for both cases with and without nuisance parameters. The links between both approaches are presented as well. Furthermore both approaches are illustrated on real data to facilitate better understanding of the discussed problems.

Keywords: Multinomial distribution, maximum likelihood, goodness of fit test, asymptotic tests

Obsah

Úvod

Multinomické rozdělení patří ke skupině velmi důležitých modelů v matematické statistice díky jeho širokému uplatnění k popisu různých situací. Jedná se o zobecnění binomického modelu na více než dvě výsledné skupiny, ať už jde o model pravděpodobností výsledků u hry šach (výhra, prohra a remíza) nebo model pro popis výhry určité politické strany při volbách za předpokladu, že máme více než dvě politické strany. V zásadě jakýkoliv pokus, při kterém máme nějakou množinu nezávislých vzájemně se vylučujících výsledných stavů, můžeme popsat pomocí multinomického rozdělení. Pearson na začátku 20. století publikoval test, který se poté začal nazývat test dobré shody (Pearson, 1900). Tento test úzce souvisí s multinomickým rozdělením. Jeho přístup není zdaleka jediný způsob, jak se dopracovat k testu pro toto rozdělení. Jiný postup, který vede k široce používaným testům, využívá teorii maximální věrohodnosti. O téměř půl století později díky teorii maximální věrohodnosti publikoval Rao v roce 1948 Raův skórový test jako alternativu k Neyman Pearsonově testu podílem věrohodností publikovném v roce 1928 a Waldovu testu publikovaném v roce 1943 (Rao, 2005). Tyto tři testové statistiky jsou někdy ve statistické literatuře o testování hypotéz nazývány jako Holy Trinity. V některých situacích jsou dva různé testy sobě velice podobné a někdy dokonce ekvivalentní. V této práci se budeme zabývat otázkou, zda tomu tak je v případě multinomického rozdělení u testu dobré shody a testů Holy Trinity (Rao, 2005). V kapitolách 1.1-1.4 se budeme zabývat testy bez rušivých parametrů a jejich aplikacemi na multinomické rozdělení. Obdobně v kapitolách 2.1-2.3 budeme rozebírat testy s rušivými parametry. Nakonec budeme v kapitolách 3.1-3.3 všechny prezentované přístupy ilustrovat testováním na reálných datech.

1. Základní testy v multinomickém rozdělení bez rušivých parametrů

V této kapitole se budeme zabývat testy pro multinomické rozdělení. Nejdříve představíme test dobré shody. Dále odvodíme potřebné výrazy pro sestavení testových statistik a následně budeme formulovat testy založené na metodě maximální věrohodnosti. Nakonec objasníme souvislost mezi těmito dvěma přístupy.

Multinomické rozdělení lze definovat následovně:

Definice 1. *(Anděl, 2007, str. 267) Nechť* $K \geq 2$, $n \in \mathbb{N}$ *a* $\boldsymbol{p} = (p_1, \ldots, p_K)^T$ *je vektor čísel splňující* $\sum_{k=1}^{K} p_k = 1$ *a* $p_k \in (0,1)$, $k = 1, \ldots, K$. *Náhodný vektor* $\boldsymbol{X} = (X_1, \ldots, X_K)^T$ *má multinomické rozdělení* $M_K(n; \boldsymbol{p})$ *právě tehdy, když je jeho hustota vzhledem k součinové čítací míře na* $\mathbb{Z}^K$ *tvaru:*

$$P(X_1 = x_1, \ldots, X_K = x_K) = \begin{cases} \frac{n!}{x_1! \ldots x_K!} p_1^{x_1} \ldots p_K^{x_K}, & \text{když } \sum_{k=1}^{K} x_k = n \\ 0, & \text{jinak.} \end{cases}$$

Rozdělení si lze představit jako provedení n nezávislých stejných experimentů s K možnými výsledky. Náhodná veličina X_k označuje počet, kolikrát skončil pokus k-tým výsledkem. Vektor $\boldsymbol{X}$ udává, kolikrát nastal jaký výsledek.

1.1 Test dobré shody bez rušivých parametrů

Karl Pearson představil v roce 1900 test, který se následně proslavil jako test dobré shody (Pearson, 1900). Používá se k testování, zda jsou pravděpodobnosti v multinomickém rozdělení rovny právě $(p_1^0, \ldots, p_K^0)$, kde K je počet skupin a p_k, $k = 1, \ldots, K$ jsou pravděpodobnosti, že pozorování bude v k-té skupině. Modelem je zde $\mathcal{F} = \{M_K(n; \boldsymbol{p}),\ 0 < p_k < 1, k \in \{1, \ldots, K\}\}$.

Testujeme nulovou hypotézu

$$H_0 : \boldsymbol{p} = \boldsymbol{p}^0$$

proti alternativní hypotéze

$$H_1 : \boldsymbol{p} \neq \boldsymbol{p}^0,$$

kde $\boldsymbol{p}^0$ je předem daný vektor s vlastnostmi $0 < p_k^0 < 1, k \in \{1, \ldots, K\}$ a $\sum_{k=1}^{K} p_k^0 = 1$.

Testová statistika je tvaru

$$\chi^2 = \sum_{k=1}^{K} \frac{(X_k - np_k^0)^2}{np_k^0},$$

kde K je počet kategorií, X_k, $k = 1, \ldots, K$ jsou empirické četnosti (reálná data) a np_k^0, $k = 1, \ldots, K$ jsou teoretické četnosti (očekávané hodnoty za platnosti nulové hypotézy). Statistika χ^2 má za platnosti nulové hypotézy podle věty 1, která bude uvedena na konci kapitoly, asymptoticky rozdělení χ^2_{K-1}. Proti H_0 svědčí velké hodnoty testové statistiky:

$$H_0 \quad \text{zamítáme} \Leftrightarrow \chi^2 \geq \chi^2_{K-1}(1-\alpha),$$

kde $\chi^2_{K-1}(1-\alpha)$ je $(1-\alpha)$-kvantil rozdělení χ^2_{K-1}.

Jedná se o asymptotický test, takže má smysl ho použít, jen když máme k dispozici dostatek pozorování. V literatuře se kvantifikuje tento požadavek vzorcem $np_k^0 \geq 5$, $k = 1, \ldots, K$. Lze také použít novější Yarnoldovo kritérium:

$$np_k^0 \geq 5q, \qquad k = 1, \ldots, K, \qquad K \geq 3,$$

kde q je podíl počtu kategorií nesplňujících $np_k \geq 5$ (Anděl, 2007, str. 271).

Věta 1. *(Anděl, 2007, str. 270) Jestliže náhodný vektor* $\boldsymbol{X} = (X_1, \ldots, X_K)^T \sim M_K(n; p_1, \ldots, p_K)$*, pak Pearsonova statistika*

$$\chi^2 = \sum_{k=1}^{K} \frac{(X_k - np_k)^2}{np_k} \xrightarrow[n\to\infty]{d} \chi^2_{K-1}.$$

1.2 Odvození maximálně věrohodných odhadů a Fisherovy informační matice pro multinomické rozdělení

Nejprve odvodíme maximálně věrohodné odhady parametrů $\boldsymbol{p}$ v $M_K(n; \boldsymbol{p})$ společně s rozptylovou maticí a Fisherovou informační maticí. Z definice 1 vidíme, že logaritmická věrohodnost je tvaru

$$\ell_n(\boldsymbol{p}) = \ln\left(\frac{n!}{X_1!...X_K!}\right) + X_1 \ln(p_1) + ... + X_K \ln(p_K).$$

První člen nezávisí na neznámém parametru, tedy po parciální derivaci podle p_k, $k = 1, \ldots, K$ vypadne. Dále využijeme, že pravděpodobnost p_K lze vyjádřit pomocí zbývajících jako $p_K = 1 - \sum_{k=1}^{K-1} p_k$. Dále je potřeba přeparametrizovat rozdělení, protože $M_K(n; \boldsymbol{p})$ má ve skutečnosti o jeden parametr méně. Pak už pracujeme jen s $K-1$ parametry. Označme je $\boldsymbol{p}_{-K} = (p_1, \ldots, p_{K-1})^T$. Věrohodnostní rovnice mají tvar

$$\frac{\partial \ell_n(\boldsymbol{p}_{-K})}{\partial p_k} = \frac{X_k}{p_k} - \frac{X_K}{1 - \sum_{k=1}^{K-1} p_k}, \qquad k = 1, \ldots, K-1.$$

Vektor skóre má tvar:

$$\boldsymbol{S}_n(p_1, \ldots, p_{K-1}) = \left(\frac{X_1}{p_1} - \frac{X_K}{p_K}, \ldots, \frac{X_{K-1}}{p_{K-1}} - \frac{X_K}{p_K}\right)^T.$$

Složky tohoto vektoru položíme rovny nule. Obdržíme soustavu rovnic ve tvaru

$$X_k = \frac{p_k}{p_K} X_K, \qquad k = 1, \ldots, K-1. \tag{1.1}$$

Jejich sečtením dostáváme

$$\sum_{k=1}^{K-1} X_k = \frac{\sum_{k=1}^{K-1} p_k}{p_K} X_K,$$

s využitím $\sum_{k=1}^{K-1} p_k = 1 - p_K$ a $\sum_{k=1}^{K-1} X_k = n - X_K$ můžeme předchozí rovnici upravit na tvar

$$n - X_K = \frac{1 - p_K}{p_K} X_K,$$

což nám drobnou úpravou dává maximálně věrohodný odhad $\hat{p}_K = \frac{X_K}{n}$. Dosazením $\hat{p}_K$ do rovnic (1.1) za p_K dostáváme maximálně věrohodný odhad

$$\hat{\boldsymbol{p}}_{-K} = \left(\frac{X_1}{n}, \ldots, \frac{X_{K-1}}{n} \right)^T.$$

Druhé derivace $\ell_n(\boldsymbol{p}_{-K})$ podle téhož parametru mají tvar

$$\frac{\partial^2 \ell_n(\boldsymbol{p}_{-K})}{\partial p_k^2} = \frac{-X_k}{p_k^2} - \frac{X_K}{\left(1 - \sum_{k=1}^{K-1} p_k\right)^2}, \qquad k = 1, \ldots, K-1.$$

Křížové druhé derivace $\ell_n(\boldsymbol{p}_{-K})$ vypadají následovně

$$\frac{\partial^2 \ell_n(\boldsymbol{p}_{-K})}{\partial p_k p_l} = -\frac{X_K}{\left(1 - \sum_{k=1}^{K-1} p_k\right)^2}, \qquad k \neq l \in \{1, \ldots, K-1\}.$$

Příslušné prvky Fisherovy informační matice multinomického rozdělení o rozsahu výběru n vypočteme jako mínus střední hodnotu z druhých parciálních derivací logaritmické věrohodnosti. Nejprve spočteme prvky mimo diagonálu Fisherovy informační matice

$$-\mathrm{E}\left[\frac{\partial^2 \ell_n(\boldsymbol{p}_{-K})}{\partial p_k p_l}\right] = -\frac{-\mathrm{E}[X_K]}{\left(1 - \sum_{k=1}^{K-1} p_k\right)^2} = \frac{np_K}{\left(1 - \sum_{k=1}^{K-1} p_k\right)^2} = \frac{np_K}{p_K^2} = \frac{n}{p_K},$$

kde jsme využili toho, že $\mathrm{E}[X_K] = np_K$. Vyplývá to z vlastnosti, že marginalní rozdělení $X_k \sim Bi(n, p_k)$, $k = 1, \ldots, K$. Dále vypočteme diagonální prvky Fisherovy informační matice

$$-\mathrm{E}\left[\frac{\partial^2 \ell_n(\boldsymbol{p}_{-K})}{\partial p_k^2}\right] = -\frac{-\mathrm{E}[X_k]}{p_k^2} - \frac{-\mathrm{E}[X_K]}{\left(1 - \sum_{k=1}^{K-1} p_k\right)^2} = \frac{np_k}{p_k^2} + \frac{np_K}{\left(1 - \sum_{k=1}^{K-1} p_k\right)^2} =$$

$$= n\left(\frac{p_k}{p_k^2} + \frac{p_K}{p_K^2}\right) = n\left(\frac{p_K + p_k}{p_k p_K}\right).$$

Fisherovu informační matici lze maticově zapsat následujícím způsobem

$$\boldsymbol{J}_n(\boldsymbol{p}_{-K}) = n\left[diag\left(\frac{1}{p_1}, \ldots, \frac{1}{p_{K-1}}\right) + \frac{1}{p_K}\mathbf{1}_{K-1}\mathbf{1}_{K-1}^T\right],$$

kde $\mathbf{1}_{K-1}^T$ je řadkový vektor $K-1$ jedniček.
Pro vyjádření testových statistik budeme ještě potřebovat inverzi této matice. Abychom ji nemuseli počítat přímo, můžeme si pomoci tím, že víme, že za určitých podmínek regularity (které multinomické rozdělení splňuje, viz Anděl, 2007, str. 159) platí, že

$$\sqrt{n}(\hat{\boldsymbol{p}}_{-K} - \boldsymbol{p}_{-K}) \xrightarrow[n\to\infty]{d} N(\mathbf{0}, \boldsymbol{J}^{-1}(\boldsymbol{p}_{-K})),$$

kde $\boldsymbol{J}(\boldsymbol{p}_{-K}) = \frac{1}{n}\boldsymbol{J}_n(\boldsymbol{p}_{-K})$, což je Fisherova matice informace pro $M_K(1;\boldsymbol{p})$. Rovnost platí, protože $\sum_{k=1}^{n} Y_k \sim M_K(n;\boldsymbol{p})$, kde $Y_1, \ldots, Y_n$ jsou nezávislé stejně rozdělené náhodné veličiny z rozdělení $M_K(1;\boldsymbol{p})$. Zároveň s využitím centrální limitní věty pro nezávislé stejně rozdělené náhodné vektory (viz např. Anděl, 2007, věta B.5 , str. 331) platí

$$\sqrt{n}(\hat{\boldsymbol{p}}_{-K} - \boldsymbol{p}_{-K}) = \frac{1}{\sqrt{n}}(\boldsymbol{X} - n\boldsymbol{p}_{-K}) \xrightarrow[n\to\infty]{d} N(\mathbf{0}, diag(\boldsymbol{p}_{-K}) - (\boldsymbol{p}_{-K})(\boldsymbol{p}_{-K})^T).$$

Z toho vyplývá, že

$$\boldsymbol{J}^{-1}(\boldsymbol{p}_{-K}) = [diag(\boldsymbol{p}_{-K}) - (\boldsymbol{p}_{-K})(\boldsymbol{p}_{-K})^T].$$

Ověříme, zda je matice skutečně inverzní, tj. počítejme

$$\boldsymbol{J}(\boldsymbol{p}_{-K})\boldsymbol{J}^{-1}(\boldsymbol{p}_{-K}) =$$

$$= \left[diag\left(\frac{1}{\boldsymbol{p}_{-K}}\right) + \frac{1}{p_K}\mathbf{1}_{K-1}\mathbf{1}_{K-1}^T\right]\left[diag(\boldsymbol{p}_{-K}) - (\boldsymbol{p}_{-K})(\boldsymbol{p}_{-K})^T\right].$$

Matice roznásobíme člen po členu, tím dostaneme 4 členy

$$\boldsymbol{J}(\boldsymbol{p}_{-K})\boldsymbol{J}^{-1}(\boldsymbol{p}_{-K}) = \mathbb{I}_{K-1} - \begin{bmatrix} p_1 & p_2 & \cdots & p_{K-1} \\ p_1 & p_2 & \cdots & p_{K-1} \\ \vdots & \vdots & \ddots & \vdots \\ p_1 & p_2 & \cdots & p_{K-1} \end{bmatrix} + \begin{bmatrix} \frac{p_1}{p_K} & \frac{p_2}{p_K} & \cdots & \frac{p_{K-1}}{p_K} \\ \frac{p_1}{p_K} & \frac{p_2}{p_K} & \cdots & \frac{p_{K-1}}{p_K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{p_1}{p_K} & \frac{p_2}{p_K} & \cdots & \frac{p_{K-1}}{p_K} \end{bmatrix} -$$

$$- \begin{bmatrix} \frac{p_1 \sum_{i=1}^{K-1} p_i}{p_K} & \frac{p_2 \sum_{i=1}^{K-1} p_i}{p_K} & \cdots & \frac{p_{K-1} \sum_{i=1}^{K-1} p_i}{p_K} \\ \frac{p_1 \sum_{i=1}^{K-1} p_i}{p_K} & \frac{p_2 \sum_{i=1}^{K-1} p_i}{p_K} & \cdots & \frac{p_{K-1} \sum_{i=1}^{K-1} p_i}{p_K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{p_1 \sum_{i=1}^{K-1} p_i}{p_K} & \frac{p_2 \sum_{i=1}^{K-1} p_i}{p_K} & \cdots & \frac{p_{K-1} \sum_{i=1}^{K-1} p_i}{p_K} \end{bmatrix} = \mathbb{A}.$$

S využitím rovnosti $\sum_{k=1}^{K-1} p_k = 1 - p_K$ si nejdříve rozebereme nediagonální členy výsledné matice $\mathbb{A}$. Označme je $a_{kl}, k \neq l \in \{1, \ldots, K-1\}$. Platí

$$a_{kl} = -p_k + \frac{p_k}{p_K} - \frac{p_k - p_k p_K}{p_K} = 0.$$

Vypočteme ještě diagonální prvky matice $\mathbb{A}$

$$a_{kk} = 1 - p_k + \frac{p_k}{p_K} - \frac{p_k - p_k p_K}{p_K} = 1,$$

to znamená, že $\mathbb{A} = \mathbb{I}_{K-1}$. Ověřili jsme, že $\boldsymbol{J}(\boldsymbol{p}_{-K})\boldsymbol{J}^{-1}(\boldsymbol{p}_{-K}) = \mathbb{I}_{K-1}$. Matice $\boldsymbol{J}^{-1}(\boldsymbol{p}_{-K})$ je skutečně inverzí $\boldsymbol{J}(\boldsymbol{p}_{-K})$.

1.3 Testy založené na maximální věrohodnosti bez rušivých parametrů

Budeme se věnovat třem testovým statistikám, které jsou obecně ve tvaru popsaném níže. Předpokládejme, že $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$ je m-rozměrný parametr.

Testujeme nulovou hypotézu

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}^0$$

proti alternativní hypotéze

$$H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}^0.$$

Raova testová statistika:

$$LM_n(\boldsymbol{\theta}_0) = \frac{1}{n}[\boldsymbol{S}_n(\boldsymbol{\theta}_0)]^T[\boldsymbol{J}(\boldsymbol{\theta}_0)]^{-1}[\boldsymbol{S}_n(\boldsymbol{\theta}_0)],$$

kde $\boldsymbol{S}_n(\boldsymbol{\theta}_0)$ je vektor skóre za nulové hypotézy.

Waldova statistika:

$$W_n(\boldsymbol{\theta}_0) = n(\hat{\boldsymbol{\theta}_n} - \boldsymbol{\theta}_0)^T \boldsymbol{J}(\hat{\boldsymbol{\theta}_n})(\hat{\boldsymbol{\theta}_n} - \boldsymbol{\theta}_0).$$

Statistika založena na podílu věrohodností:

$$LR_n(\boldsymbol{\theta}_0) = 2[\ell_n(\hat{\boldsymbol{\theta}_n}) - \ell_n(\boldsymbol{\theta}_0)].$$

Tyto testové statistiky mají za určitých podmínek regularity (viz Anděl, 2007, str. 159) asymptoticky χ^2_m rozdělení.

Aplikace na multinomické rozdělení

Pro vyjádření těchto testových statistik v případě multinomického rozdělení, využijeme spočtených výrazů z předešlé kapitoly. Model je stejný jako je v testu dobré shody. Budeme dělat testy o vektoru parametrů $\boldsymbol{p} = (p_1, \ldots, p_K)^T$.

Nulová a alternativní hypotéza jsou stejné jako v kapitole 1.1.

Všechny tři testové statistiky pro multinomické rozdělení vypočteme. Jako první vypočteme testovou statistiku založenou na podílu věrohodností

$$LR_n(\boldsymbol{p}^0) = 2[\ell_n(\hat{\boldsymbol{p}}_n) - \ell_n(\boldsymbol{p}^0)] =$$
$$= 2\left[\left\{\ln\frac{n!}{X_1!\ldots X_K!} + \sum_{k=1}^{K} X_k \ln\frac{X_k}{n}\right\} - \left\{\ln\frac{n!}{X_1!\ldots X_K!} + \sum_{k=1}^{K} X_k \ln p_k^0\right\}\right] =$$
$$= 2\sum_{k=1}^{K} X_k \ln\frac{X_k}{np_k^0}.$$

Nezáleží na tom, zda vezmeme celý (přeparametrizovaný) vektor $\hat{\boldsymbol{p}}_n$ nebo $\hat{\boldsymbol{p}}_{-K}$. Jediný rozdíl bude v posledním členu sumy, v prvním případě má tvar $X_K \ln\frac{X_K}{np_K^0}$ a v druhém je ekvivalentního tvaru $X_K \ln\frac{n-\sum_{k=1}^{K-1} X_k}{n\left(1-\sum_{k=1}^{K-1} p_k^0\right)}$, což je totéž.

Výpočet Waldovy statistiky už je poněkud složitější, proto uvedeme následující lemma. Pro zjednodušení výrazů budeme často využívat dvou rovností:

$$p_K^0 = 1 - \sum_{k=1}^{K-1} p_k^0, \tag{1.2}$$

$$X_K = n - \sum_{k=1}^{K-1} X_k. \tag{1.3}$$

Lemma 2 (O tvaru $W_n(\boldsymbol{p}^0)$)**.** *Waldova statistika má v případě multinomického rozdělení tvar:*

$$W_n(\boldsymbol{p}^0) = \sum_{k=1}^{K} \frac{(X_k - np_k^0)^2}{X_k}.$$

Důkaz. Waldova statistika používá Fisherovu informační matici v maximálně věrohodném odhadu. Vyjádřeme si ji

$$\boldsymbol{J}(\hat{\boldsymbol{p}}_{-K}) = \left[diag\left(\frac{1}{\frac{X_1}{n}}, \ldots, \frac{1}{\frac{X_{K-1}}{n}}\right) + \frac{1}{\frac{X_K}{n}}\mathbf{1}_{K-1}\mathbf{1}_{K-1}^T\right] =$$
$$= \left[diag\left(\frac{n}{X_1}, \ldots, \frac{n}{X_{K-1}}\right) + \frac{n}{X_K}\mathbf{1}_{K-1}\mathbf{1}_{K-1}^T\right].$$

Nyní dosadíme tuto matici do Waldovy statistiky

$$W_n(\boldsymbol{p}^0) = n\left(\hat{\boldsymbol{p}}_{-K} - \boldsymbol{p}_{-K}^0\right)^T \boldsymbol{J}(\hat{\boldsymbol{p}}_{-K})\left(\hat{\boldsymbol{p}}_{-K} - \boldsymbol{p}_{-K}^0\right) =$$

$$= n \begin{pmatrix} \frac{X_1}{n} - p_1^0 \\ \vdots \\ \frac{X_{K-1}}{n} - p_{K-1}^0 \end{pmatrix}^T \begin{bmatrix} \frac{n}{X_1} + \frac{n}{X_K} & \frac{n}{X_K} & \cdots & \frac{n}{X_K} \\ \frac{n}{X_K} & \frac{n}{X_2} + \frac{n}{X_K} & \cdots & \frac{n}{X_K} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{n}{X_K} & \frac{n}{X_K} & \cdots & \frac{n}{X_{K-1}} + \frac{n}{X_K} \end{bmatrix} \begin{pmatrix} \frac{X_1}{n} - p_1^0 \\ \vdots \\ \frac{X_{K-1}}{n} - p_{K-1}^0 \end{pmatrix}.$$

Nejdříve budeme násobit levý vektor s maticí. Výsledkem bude řádkový vektor, který si označíme $V_n(\boldsymbol{p}^0)$, jehož k-tá souřadnice $k \in \{1, \ldots, K-1\}$ vznikne vynásobením řádkového vektoru a k-tého sloupce matice. Vypočteme první složku

$$V_n^1(\boldsymbol{p}^0) = \left(\frac{X_1}{n} - p_1^0, \ldots, \frac{X_{K-1}}{n} - p_{K-1}^0\right) \left(\begin{pmatrix} \frac{n}{X_K} \\ \frac{n}{X_K} \\ \vdots \\ \frac{n}{X_K} \end{pmatrix} + \begin{pmatrix} \frac{n}{X_1} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \right) =$$

$$= \sum_{k=1}^{K-1} \frac{X_k}{X_K} - \sum_{k=1}^{K-1} \frac{np_k^0}{X_K} + 1 - \frac{np_1^0}{X_1}.$$

S využitím rovností (1.2) a (1.3) dostaneme výraz

$$V_n^1(\boldsymbol{p}^0) = \frac{n - X_K}{X_K} - \frac{n(1 - p_K^0)}{X_K} + 1 - \frac{np_1^0}{X_1}.$$

Úpravou dostaneme

$$V_n^1(\boldsymbol{p}^0) = \frac{X_1 - np_1^0}{X_1} + \frac{np_K^0 - X_K}{X_K}.$$

Složky vektoru $k = 2, \ldots, K-1$ vypočteme analogicky. Dostaneme obdobný výsledek tvaru

$$V_n^k(\boldsymbol{p}^0) = \frac{X_k - np_k^0}{X_k} + \frac{np_K^0 - X_K}{X_K}.$$

Výsledný vektor $V_n(\boldsymbol{p}^0) = \left(V_n^1(\boldsymbol{p}^0), \ldots, V_n^{K-1}(\boldsymbol{p}^0)\right)^T$ je

$$\left(\frac{X_1 - np_1^0}{X_1} + \frac{np_K^0 - X_K}{X_K}, \ldots, \frac{X_{K-1} - np_{K-1}^0}{X_{K-1}} + \frac{np_K^0 - X_K}{X_K}\right).$$

K vyjádření Waldovy statistiky zbývá vynásobit s pravým vektorem a pronásobit n

$$W_n(\boldsymbol{p}^0) = n \begin{pmatrix} \frac{X_1 - np_1^0}{X_1} + \frac{np_K^0 - X_K}{X_K} \\ \vdots \\ \frac{X_{K-1} - np_{K-1}^0}{X_{K-1}} + \frac{np_K^0 - X_K}{X_K} \end{pmatrix}^T \begin{pmatrix} \frac{X_1}{n} - p_1^0 \\ \vdots \\ \frac{X_{K-1}}{n} - p_{K-1}^0 \end{pmatrix}.$$

Zaměříme se na k-tou složku $k \in \{1, \ldots, K-1\}$ výsledného součtu

$$n \left(\frac{X_k - np_k^0}{X_k} + \frac{np_K^0 - X_K}{X_K}\right) \left(\frac{X_k - np_k^0}{n}\right) =$$

$$= n \left(\frac{(X_k - np_k^0)^2}{nX_k} + \frac{X_k(np_K^0 - X_K)}{nX_K} - \frac{p_k^0(np_K^0 - X_K)}{X_K}\right).$$

Tento výraz upravíme na

$$\frac{(X_k - np_k^0)^2}{X_k} + \frac{X_k(np_K^0 - X_K)}{X_K} - \frac{p_k^0(n^2 p_K^0 - nX_K)}{X_K}.$$

Nyní sečteme přes k a tím dostaneme Waldovu testovou statistiku

$$W_n(\boldsymbol{p}^0) = \sum_{k=1}^{K-1} \frac{(X_k - np_k^0)^2}{X_k} + \frac{\left(\sum_{k=1}^{K-1} X_k\right)(np_K^0 - X_K)}{X_K} - \frac{\left(\sum_{k=1}^{K-1} p_k^0\right)(n^2 p_K^0 - nX_K)}{X_K}.$$

S využitím rovností (1.2) a (1.3) dostáváme

$$W_n(\boldsymbol{p}^0) = \sum_{k=1}^{K-1} \frac{(X_k - np_k^0)^2}{X_k} + \frac{(n - X_K)(np_K^0 - X_K)}{X_K} - \frac{(1 - p_K^0)(n^2 p_K^0 - nX_K)}{X_K} =$$

po roznásobení

$$= \sum_{k=1}^{K-1} \frac{(X_k - np_k^0)^2}{X_k} + \frac{n^2 p_K^0 - nX_K - 2np_K^0 X_K + X_K^2 - n^2 p_K^0 + nX_K + n^2(p_K^0)^2}{X_K}.$$

Což nám po vyrušení členů s opačným znaménkem dává

$$W_n(\boldsymbol{p}^0) = \sum_{k=1}^{K-1} \frac{(X_k - np_k^0)^2}{X_k} + \frac{X_K^2 - 2np_K^0 X_K + n^2(p_K^0)^2}{X_K} = \sum_{k=1}^{K} \frac{(X_k - np_k^0)^2}{X_k}.$$

□

Zbývá Raova testová statistika. Výpočet je opět delší a pro realizaci výpočtu využijeme následující lemma. Znovu budeme využívat rovnosti (1.2) a (1.3).

Lemma 3 (O tvaru $LM_n(\boldsymbol{p}^0)$)**.** *Raova statistika má v případě multinomického rozdělení tvar:*

$$LM_n(\boldsymbol{p}^0) = \sum_{k=1}^{K} \frac{(X_k - np_k^0)^2}{np_k^0}.$$

Důkaz.

$$LM_n(\boldsymbol{p}^0) = \frac{1}{n} \boldsymbol{S}_n(\boldsymbol{p}^0{}_{-K})^T \left(\left[\boldsymbol{J}(\boldsymbol{p}^0{}_{-K}) \right]^{-1} \right) \boldsymbol{S}_n(\boldsymbol{p}^0{}_{-K})$$

Maticový zápis součinu:

$$\frac{1}{n} \begin{pmatrix} \frac{X_1}{p_1^0} - \frac{X_K}{p_K^0} \\ \vdots \\ \frac{X_{K-1}}{p_{K-1}^0} - \frac{X_K}{p_K^0} \end{pmatrix}^T \begin{bmatrix} p_1^0(1-p_1^0) & -p_1^0 p_2^0 & \dots & -p_1^0 p_{K-1}^0 \\ -p_1^0 p_2^0 & p_2^0(1-p_2^0) & \dots & -p_2^0 p_{K-1}^0 \\ \vdots & \vdots & \ddots & \vdots \\ -p_1^0 p_{K-1}^0 & -p_2^0 p_{K-1}^0 & \dots & p_{K-1}^0(1-p_{K-1}^0) \end{bmatrix} \begin{pmatrix} \frac{X_1}{p_1^0} - \frac{X_K}{p_K^0} \\ \vdots \\ \frac{X_{K-1}}{p_{K-1}^0} - \frac{X_K}{p_K^0} \end{pmatrix}$$

Nejdříve budeme násobit levý vektor s maticí. Výsledkem bude řádkový vektor, jehož k-tá souřadnice $k \in \{1, \dots, K-1\}$ vznikne vynásobením řádkového vektoru a k-tého sloupce matice. Vypočteme první složku

$$LM_n^1(\boldsymbol{p}^0) = \left(\frac{X_1}{p_1^0} - \frac{X_K}{p_K^0}, \dots, \frac{X_{K-1}}{p_{K-1}^0} - \frac{X_K}{p_K^0} \right) \left(\begin{pmatrix} -p_1^0 p_1^0 \\ -p_1^0 p_2^0 \\ \vdots \\ -p_1^0 p_{K-1}^0 \end{pmatrix} + \begin{pmatrix} p_1^0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \right) =$$

$$= X_1 - \frac{p_1^0 X_K}{p_K^0} - p_1^0 \sum_{k=1}^{K-1} X_k + \frac{p_1^0 X_K}{p_K^0} \sum_{k=1}^{K-1} p_k^0.$$

Nyní využijeme rovnosti (1.2) a (1.3), čímž dostaneme výraz

$$LM_n^1(\boldsymbol{p}^0) = X_1 - \frac{p_1^0 X_K}{p_K^0} - np_1^0 + p_1^0 X_K + \frac{p_1^0 X_K}{p_K^0} - \frac{p_1^0 p_K^0 X_K}{p_K^0}.$$

Po vyrušení členů s opačným znaménkem dostáváme dostatečně jednoduchý výraz

$$LM_n^1(\boldsymbol{p}^0) = X_1 - np_1^0.$$

Složky $k = 2, \ldots, K-1$ vektoru vypočteme analogicky. Dostaneme obdobný výsledek tvaru

$$LM_n^k(\boldsymbol{p}^0) = X_k - np_k^0.$$

Výsledný vektor $\left(LM_n^1(\boldsymbol{p}^0), \ldots, LM_n^{K-1}(\boldsymbol{p}^0)\right)^T$ je

$$(X_1 - np_1^0, \ldots, X_{K-1} - np_{K-1}^0).$$

Nyní potřebujeme vynásobit tento řádkový vektor s pravým vektorem a přenásobit $\frac{1}{n}$

$$LM_n(\boldsymbol{p}^0) = \frac{1}{n}(X_1 - np_1^0, \ldots, X_{K-1} - np_{K-1}^0) \begin{pmatrix} \frac{X_1}{p_1^0} - \frac{X_K}{p_K^0} \\ \vdots \\ \frac{X_{K-1}}{p_{K-1}^0} - \frac{X_K}{p_K^0} \end{pmatrix}.$$

Zaměříme se na k-tou složku $k \in \{1, \ldots, K-1\}$ výsledného součtu

$$(X_k - np_k^0)\left(\frac{X_k}{p_k^0} - \frac{X_K}{p_K^0}\right) =$$

$$= \frac{X_k^2}{p_k^0} - \frac{X_k X_K}{p_K^0} - \frac{np_k^0 X_k}{p_k^0} + \frac{np_k^0 X_K}{p_K^0} = \frac{X_k^2 - np_k^0 X_k}{p_k^0} + \frac{np_k^0 X_K - X_k X_K}{p_K^0}.$$

Nyní sečteme přes k a vynásobíme $\frac{1}{n}$, tento výraz bude roven Raově statistice

$$LM_n(\boldsymbol{p}^0) = \sum_{k=1}^{K-1} \frac{X_k^2 - np_k^0 X_k}{np_k^0} + \frac{nX_K \sum_{k=1}^{K-1} p_k^0 - X_K \sum_{k=1}^{K-1} X_k}{np_K^0}.$$

Rovnosti (1.2) a (1.3) nám statistiku zjednoduší na

$$LM_n(\boldsymbol{p}^0) = \sum_{k=1}^{K-1} \frac{X_k^2 - np_k^0 X_k}{np_k^0} + \frac{nX_K(1 - p_K^0) - X_K(n - X_K)}{np_K^0}.$$

Úpravou dostáváme

$$LM_n(\boldsymbol{p}^0) = \sum_{k=1}^{K-1} \frac{X_k^2 - np_k^0 X_k}{np_k^0} + \frac{X_K^2 - np_K^0 X_K}{np_K^0} = \sum_{k=1}^{K} \frac{X_k^2 - np_k^0 X_k}{np_k^0}.$$

Pokud bychom chtěli dále upravit do průhlednějšího tvaru, odečteme „0“ neboli $\sum_{k=1}^{K}(X_k - np_k^0)$. Dostáváme

$$LM_n(\boldsymbol{p}^0) = \sum_{k=1}^{K}\left(\frac{X_k^2 - np_k^0 X_k}{np_k^0} - (X_k - np_k^0)\right) =$$

$$= \sum_{k=1}^{K}\frac{X_k^2 - np_k^0 X_k - np_k^0 X_k + n^2(p_k^0)^2}{np_k^0}.$$

Což je to stejné jako

$$LM_n(\boldsymbol{p}^0) = \sum_{k=1}^{K}\frac{(X_k - np_k^0)^2}{np_k^0}.$$

□

Odvodili jsme standardní tvary testových statistik z maximálně věrohodného přístupu. Shrneme výsledky

$$LM_n(\boldsymbol{p}^0) = \sum_{k=1}^{K}\frac{(X_k - np_k^0)^2}{np_k^0},$$

$$W_n(\boldsymbol{p}^0) = \sum_{k=1}^{K}\frac{(X_k - np_k^0)^2}{X_k},$$

$$LR_n(\boldsymbol{p}^0) = 2\sum_{k=1}^{K}X_k \ln\frac{X_k}{np_k^0}.$$

Všechny tři testové statistiky mají za nulové hypotézy v naší situaci asymptoticky rozdělení χ^2_{K-1}.

1.4 Vzájemná souvislost uvedených přístupů

Testová statistika v Pearsonově přístupu je tvaru

$$\chi^2 = \sum_{k=1}^{K}\frac{(X_k - np_k^0)^2}{np_k^0}.$$

První čeho si můžeme povšimnout je, že v případě multinomického rozdělení platí

$$LM_n(\boldsymbol{p}^0) = \chi^2.$$

To znamená, že v našem případě je Pearsonova statistika přímo Raova statistika.

Waldova statistika je Pearsonově statistice velice podobná. Jediný rozdíl je ve jmenovateli. Ve Waldově testu máme ve jmenovateli napozorované četnosti X_k namísto předpokládaných np_k^0. Ovšem za platnosti nulové hypotézy platí, že

$$\frac{X_k}{np_k^0} \xrightarrow[n\to\infty]{P} 1, k = 1, \ldots, K \tag{1.4}$$

navíc jsou oba testy asymptotické. Z toho plyne, že za platnosti nulové hypotézy budou testové statistiky asymptoticky ekvivalentní

$$LM_n(\boldsymbol{p}^0) - W_n(\boldsymbol{p}^0) \xrightarrow[n\to\infty]{P} 0$$

díky Cramérově-Sluckého větě (viz např. Anděl, 2007, str. 333 věta B.10). Tuto skutečnost vysvětlíme podrobněji

$$LM_n(\boldsymbol{p}^0) - W_n(\boldsymbol{p}^0) = \sum_{k=1}^{K} \frac{(X_k - np_k^0)^2}{np_k^0} - \sum_{k=1}^{K} \frac{(X_k - np_k^0)^2}{X_k} =$$

$$= \sum_{k=1}^{K} \frac{(X_k - np_k^0)^2}{np_k^0} - \sum_{k=1}^{K} \frac{(X_k - np_k^0)^2}{np_k^0} \frac{np_k^0}{X_k} = \sum_{k=1}^{K} \left(1 - \frac{np_k^0}{X_k}\right) \frac{(X_k - np_k^0)^2}{np_k^0} \xrightarrow[n\to\infty]{d} 0.$$

Kvůli tomu, že každý člen sumy konverguje v distribuci k nule, protože

$$\left(1 - \frac{np_k^0}{X_k}\right) \xrightarrow[n\to\infty]{P} 0,$$

díky platnosti 1.4. Druhý člen konverguje v distribuci k jistému rozdělení, to lze vidět přenasobením $\frac{1}{1-p_k^0}$. Přenásobením nenulovým konečným číslem dostaneme následující asymptotiku

$$\frac{(X_k - np_k^0)^2}{np_k^0(1-p_k^0)} \xrightarrow[n\to\infty]{d} \chi_1^2.$$

Přesněji využijeme centrální limitní věty aplikované na binomické rozdělení ($X_k \sim Bi(n, p_k)$), máme

$$\frac{X_k - np_k^0}{\sqrt{np_k^0(1-p_k^0)}} \xrightarrow[n\to\infty]{d} N(0,1)$$

(viz např. Anděl, 2007, str. 335, věta B.12). Umocněním na druhou dostaneme

$$\frac{(X_k - np_k^0)^2}{np_k^0(1-p_k^0)} \xrightarrow[n\to\infty]{d} \chi_1^2,$$

což je až na přenásobení konstantou člen vystupující v námi diskutovaném součinu. Celkem jsem obdrželi

$$\left(\left(1 - \frac{np_k^0}{X_k}\right) \frac{(X_k - np_k^0)^2}{np_k^0}\right) \xrightarrow[n\to\infty]{d} 0.$$

Konvergence v distribuci ke konstantě implikuje i konvergenci k téže konstantě v pravděpodobnosti.

Nyní uvažujme poslední statistiku z maximálně věrohodného přístupu test založený na podílu věrohodností, kterou můžeme psát

$$LR_n(\boldsymbol{p}^0) = 2\sum_{k=1}^{K} X_k \ln \frac{X_k}{np_k^0} = 2\sum_{k=1}^{K} np_k^0 \left(1 + \frac{X_k - np_k^0}{np_k^0}\right) \ln\left(1 + \frac{X_k - np_k^0}{np_k^0}\right).$$

Taylorovým rozvojem logaritmu do druhého řádu dostaneme následující výraz

$$LR_n(\boldsymbol{p}^0) = 2\sum_{k=1}^{K} np_k^0 \left(1 + \frac{X_k - np_k^0}{np_k^0}\right) \ln\left(1 + \frac{X_k - np_k^0}{np_k^0}\right) \doteq$$

$$\doteq 2\sum_{k=1}^{K} np_k^0 \left(1+\frac{X_k-np_k^0}{np_k^0}\right)\left[\left(\frac{X_k-np_k^0}{np_k^0}\right)-\frac{\left(\frac{X_k-np_k^0}{np_k^0}\right)^2}{2}\right].$$

Zanedbáním členu $\sum_{k=1}^{K} np_k^0\left(\frac{X_k-np_k^0}{np_k^0}\right)^3$ získáme výsledný tvar

$$LR_n(\boldsymbol{p}^0) \doteq 2\sum_{k=1}^{K} np_k^0\left(\frac{X_k-np_k^0}{np_k^0}\right)+\sum_{k=1}^{K} np_k^0\left(\frac{X_k-np_k^0}{np_k^0}\right)^2=\sum_{k=1}^{K}\frac{(X_k-np_k^0)^2}{np_k^0}.$$

Poslední rovnost platí z následujícího výpočtu, který ukáže, že první člen je roven nule

$$\sum_{k=1}^{K} np_k^0\left(\frac{X_k-np_k^0}{np_k^0}\right)=\sum_{k=1}^{K}\left(X_k-np_k^0\right)=(n-n)=0.$$

Tento postup (publikován např. Williams, 1976) ilustruje souvislost Pearsonovy statistiky se statistikou testu založeného na podílu věrohodností. Pearsonova statistika je aproximací testu založeného na podílu věrohodností za předpokladu, že platí $|X_k-np_k^0|<np_k^0$, pro $k\in\{1,\ldots,K\}$. Předpoklad je pro naší situaci splněn s pravděpodobností 1 pro $n\to\infty$. Pokud nerovnost přenásobíme $\frac{1}{n}$, dostaneme $|\frac{X_k}{n}-p_k^0|<p_k^0$. Z vlastnosti 1.4 vidíme, že je v našem případě tato podmínka splněna. Tedy celkem máme

$$P\left(\left|\frac{X_k}{n}-p_k^0\right|<p_k^0\right)\xrightarrow{n\to\infty}1.$$

Tento předpoklad zaručuje, že můžeme použít Taylorovu větu pro rozvoj $\ln(1+x)$ kolem $x=0$ v nekonečnou řadu tj.
$\ln(1+x)=\sum_{k=1}^{\infty}(-1)^{k+1}\frac{x^k}{k}=x-\frac{x^2}{2}+\frac{x^3}{3}-\ldots.$ pro každé $|x|<1$.

2. Testy v multinomickém rozdělení s rušivými parametry

V této kapitole se budeme zabývat o něco obecnější situací. V některých případech nás mohou zajímat hypotézy jen o části modelu. Např. v případě normálního rozdělení můžeme chtít testovat hypotézy jen o střední hodnotě a tedy rozptyl nám vstupuje do testů jen jako rušivý vliv. Nejprve uvedeme Pearsonův přístup. Poté nastíníme přístup maximální věrohodnosti a postup budeme aplikovat na $M_K(n;\boldsymbol{p})$.

2.1 Test dobré shody s rušivými parametry

Tento test je přímým zobecněním testu dobré shody bez rušivých parametrů. Toto zobecnění zahrnuje situaci, kdy je vektor parametrů $\boldsymbol{p}^0$ zavislý na neznámém vektoru parametrů $\boldsymbol{\psi}_0$. Kdybychom ho znali, bylo by možné použít přístup z kapitoly 1.1. Modelem pro nás bude následující situace: Nechť náhodný vektor $(X_1,\dots,X_K)^T \sim M_K(n,\boldsymbol{p}(\boldsymbol{\psi}_X))$, kde $\boldsymbol{\psi}_X \in \Theta_{\boldsymbol{\psi}} \subset \mathbb{R}^d$ je d-rozměrný parametr, $d < K-1$ a p je funkce, která zobrazuje $\Theta_{\boldsymbol{\psi}}$ do $(0,1)^K$ a splňuje $p(\boldsymbol{\psi})^T\mathbf{1}_K = 1$ pro všechny $\psi \in \Theta_{\boldsymbol{\psi}}$.

Testujeme nulovou hypotézu

$$H_0 : \exists \boldsymbol{\psi}_x \in \Theta_{\boldsymbol{\psi}} : \boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\psi}_X)$$

proti alternativní hypotéze

$$H_1 : \forall \boldsymbol{\psi}_x \in \Theta_{\boldsymbol{\psi}} : \boldsymbol{p} \neq \boldsymbol{p}(\boldsymbol{\psi}_X).$$

Nejdříve musíme odhadnout neznámé parametry. S použitím metody maximální věrohodnosti bude odvození vypadat následovně:
Logaritmická věrohodnost je tvaru

$$\ell_n(\boldsymbol{\psi}) = \sum_{k=1}^{K} X_k \ln p_k(\boldsymbol{\psi}) + \ln\left(\frac{n!}{X_1!\dots X_K!}\right).$$

Soustava věrohodnostních rovnic $\left.\frac{\partial \ell_n(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}_n} = \mathbf{0}_d$ vede na soustavu d rovnic o d neznámých $\hat{\boldsymbol{\psi}}_n$:

$$\sum_{k=1}^{K} \frac{X_k}{p_k(\hat{\boldsymbol{\psi}}_n)} \frac{\partial p_k(\hat{\boldsymbol{\psi}}_n)}{\partial \boldsymbol{\psi}} = \mathbf{0}_d.$$

Výsledná testová statistika bude vypadat následovně

$$\chi^{*2}_n = \sum_{k=1}^{K} \frac{(X_k - np_k(\hat{\boldsymbol{\psi}}_n))^2}{np_k(\hat{\boldsymbol{\psi}}_n)}.$$

Její asymptotické rozdělení za nulové hypotézy a za jistých předpokladů regularity je χ^2_{K-d-1}, kde d je počet odhadnutých parametrů (viz Anděl, 2007, str. 273

věta 12.9). Jedná se o stejný myšlenkový postup jako u testu v kapitole 1.1. To znamená, že porovnáváme napozorovaná data v kategorii k (empirické četnosti) s očekávaným počtem za platnosti nulové hypotézy (teoretické četnosti).

2.2 Testy založené na maximální věrohodnosti s rušivými parametry

Nejdříve uvedeme obecný přístup, který poté o něco modifikujeme, abychom neměli problémy s aplikací na multinomické rozdělení. Vektor $\boldsymbol{\theta}$ rozdělíme na dvě složky $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\tau} \\ \boldsymbol{\psi} \end{pmatrix}$. Uvažujeme $\boldsymbol{\psi} \in \boldsymbol{\Theta}_\psi \subset \mathbb{R}^d$, $\boldsymbol{\tau} \in \boldsymbol{\Theta}_\tau \subset \mathbb{R}^{p-d}$ a tedy $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$, kde $d, p \in \mathbb{N}, d < p$. Pokud je $d = 0$, jedná se o případ testu bez rušivých parametrů diskutovaný v kapitolách 1.1-1.4. Chceme testovat hypotézu jen o $\boldsymbol{\tau}$. Parametr $\boldsymbol{\psi}$ je potřebný jen k popisu modelu, ale v nulové hypotéze se nevyskytuje. Parametru $\boldsymbol{\tau}$ říkáme cílový parametr. Složkám $\boldsymbol{\psi}$ říkáme rušivé parametry.
Testujeme složenou nulovou hypotézu

$$H_0^* : \boldsymbol{\tau} = \boldsymbol{\tau}_0$$

proti složené alternativě

$$H_1^* : \boldsymbol{\tau} \neq \boldsymbol{\tau}_0.$$

Nejdříve se budeme zabývat zobecněním přístupu z kapitoly 1.3 na tento případ (D.R. Cox, D.V. Hinkley, 1974, str. 322). Nejprve zavedeme několik potřebných značení. Je dobré si uvědomit, že kromě maximálně věrohodného odhadu $\begin{pmatrix} \hat{\boldsymbol{\tau}}_n \\ \hat{\boldsymbol{\psi}}_n \end{pmatrix}$ budeme potřebovat odhad za nulové hypotézy. To znamená, že potřebujeme odhadnout $\boldsymbol{\psi}$ za podmínky $\boldsymbol{\tau} = \boldsymbol{\tau}_0$. Analogicky, jako jsme rozdělili vektor $\boldsymbol{\theta}$, si rozdělíme vektor skóre $\boldsymbol{S}_n(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{S}_n^\tau(\boldsymbol{\tau},\boldsymbol{\psi}) \\ \boldsymbol{S}_n^\psi(\boldsymbol{\tau},\boldsymbol{\psi}) \end{pmatrix}$, kde

$$\boldsymbol{S}_n^\tau(\boldsymbol{\tau},\boldsymbol{\psi}) = \left(\frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \tau_1}, \ldots, \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \tau_{p-d}} \right)^T \text{ a } \boldsymbol{S}_n^\psi(\boldsymbol{\tau},\boldsymbol{\psi}) = \left(\frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \psi_1}, \ldots, \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \psi_d} \right)^T.$$

Maximálně věrohodný odhad parametru $\boldsymbol{\psi}$ za podmínky $\boldsymbol{\tau} = \boldsymbol{\tau}_0$ dostaneme jako řešení rovnice $\boldsymbol{S}_n^\psi(\boldsymbol{\tau}_0,\boldsymbol{\psi}) = \mathbf{0}_d$, označíme jej $\tilde{\boldsymbol{\psi}}_n$. Dále potřebujeme, podobně jako jsme rozdělili vektor $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\tau} \\ \boldsymbol{\psi} \end{pmatrix}$, rozdělit Fisherovu informační matici. Matici rozdělíme na blokovou matici o blocích $[p-d] \times [p-d], [p-d] \times d, d \times [p-d]$ a $d \times d$

$$\boldsymbol{J}(\boldsymbol{\tau},\boldsymbol{\psi}) = \begin{bmatrix} \boldsymbol{J}^{\tau\tau}(\boldsymbol{\tau},\boldsymbol{\psi}) & \boldsymbol{J}^{\tau\psi}(\boldsymbol{\tau},\boldsymbol{\psi}) \\ \boldsymbol{J}^{\psi\tau}(\boldsymbol{\tau},\boldsymbol{\psi}) & \boldsymbol{J}^{\psi\psi}(\boldsymbol{\tau},\boldsymbol{\psi}) \end{bmatrix}.$$

Znovu se budeme zabývat třemi testovými statistikami, které jsou obecně ve tvaru:

$$LR_n^*(\boldsymbol{\tau}_0) = 2[\ell_n(\hat{\boldsymbol{\tau}}_n,\hat{\boldsymbol{\psi}}_n) - \ell_n(\boldsymbol{\tau}_0,\tilde{\boldsymbol{\psi}}_n)],$$

$$W_n^*(\boldsymbol{\tau}_0) = n(\hat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0)^T \boldsymbol{J}_\tau^{-1}(\hat{\boldsymbol{\tau}}_n,\hat{\boldsymbol{\psi}}_n)(\hat{\boldsymbol{\tau}}_n - \boldsymbol{\tau}_0),$$

kde

$$\boldsymbol{J}_{\tau}^{-1}(\hat{\boldsymbol{\tau}}_n,\hat{\boldsymbol{\psi}}_n) = \boldsymbol{J}^{\tau\tau}(\hat{\boldsymbol{\tau}}_n,\hat{\boldsymbol{\psi}}_n) - \boldsymbol{J}^{\tau\psi}(\hat{\boldsymbol{\tau}}_n,\hat{\boldsymbol{\psi}}_n)[\boldsymbol{J}^{\psi\psi}(\hat{\boldsymbol{\tau}}_n,\hat{\boldsymbol{\psi}}_n)]^{-1}\boldsymbol{J}^{\psi\tau}(\hat{\boldsymbol{\tau}}_n,\hat{\boldsymbol{\psi}}_n),$$

což je vlastně levý horní blok o rozměru $[p-d]$ (odpovídající parametru $\boldsymbol{\tau}$) matice $\boldsymbol{J}^{-1}(\hat{\boldsymbol{\tau}}_n,\hat{\boldsymbol{\psi}}_n)$,

$$LM_n^*(\boldsymbol{\tau}_0) = \frac{1}{n}[\boldsymbol{S}_n^{\tau}(\boldsymbol{\tau}_0,\tilde{\boldsymbol{\psi}}_n)]^T[\boldsymbol{J}_{\tau}(\boldsymbol{\tau}_0,\tilde{\boldsymbol{\psi}}_n)][\boldsymbol{S}_n^{\tau}(\boldsymbol{\tau}_0,\tilde{\boldsymbol{\psi}}_n)],$$

kde

$$\boldsymbol{J}_{\tau}(\boldsymbol{\tau}_0,\tilde{\boldsymbol{\psi}}_n) = \left(\boldsymbol{J}_{\tau}^{-1}(\boldsymbol{\tau}_0,\tilde{\boldsymbol{\psi}}_n)\right)^{-1}.$$

Všechny tři testové statistiky mají asymptoticky χ^2_{p-d} (viz např. Kulich, 2000, str. 129).

Nyní přístup o něco málo upravíme (Shao, 1999, str. 384-387). Předpokládejme, že $\boldsymbol{\Theta}_0$ je určen nulovou hypotézou tvaru

$$H_0^1 : \boldsymbol{\theta} = \boldsymbol{g}(\boldsymbol{\psi}) \text{ pro nějaké } \boldsymbol{\psi} \in \boldsymbol{\Theta}_{\psi} \subset \mathbb{R}^d,$$

kde je $\boldsymbol{\psi}$ je d složkový vektor a g je spojitě diferencovatelná funkce z $\mathbb{R}^d$ do $\mathbb{R}^p$, kde $d, p \in \mathbb{N}, d < p$. Poté platí, že

$$LR_n^* = 2[\ln \ell_n(\hat{\boldsymbol{\theta}}_n) - \ln \ell_n(g(\hat{\boldsymbol{\psi}}_n))] \xrightarrow[n\to\infty]{d} \chi^2_{p-d}.$$

Hypotéza H_0^1 je ekvivalentní hypotéze

$$H_0^2 : \boldsymbol{R}(\boldsymbol{\theta}) = \boldsymbol{0}_{p-d},$$

kde $\boldsymbol{R}(\boldsymbol{\theta})$ je spojitě diferencovatelná funkce z $\mathbb{R}^p$ do $\mathbb{R}^{p-d}$ (Shao, 1999, str. 386). Waldova statistika, která lze získat z obecného případu aplikací delta-věty (např. Shao, 1999, str. 43, Theorem 1.12), je potom tvaru

$$W_n^* = n\boldsymbol{R}(\hat{\boldsymbol{\theta}}_n)^T\{\boldsymbol{C}(\hat{\boldsymbol{\theta}}_n)^T[\boldsymbol{J}(\hat{\boldsymbol{\theta}}_n)]^{-1}\boldsymbol{C}(\hat{\boldsymbol{\theta}}_n)\}^{-1}[\boldsymbol{R}(\hat{\boldsymbol{\theta}}_n)],$$

kde $\boldsymbol{C}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{R}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}$ je matice rozměrů $[p-d] \times p$.
Raova statistika je v tomto případě tvaru (Shao, 1999, Theorem 6.6, str. 386):

$$LM_n^* = \frac{1}{n}[\boldsymbol{S}_n(g(\tilde{\boldsymbol{\psi}}_n))]^T[\boldsymbol{J}(g(\tilde{\boldsymbol{\psi}}_n))]^{-1}[\boldsymbol{S}_n g((\tilde{\boldsymbol{\psi}}_n))],$$

kde $\boldsymbol{S}_n(\boldsymbol{\theta}) = \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, g je spojitě diferencovatelná funkce z $\mathbb{R}^d$ do $\mathbb{R}^p$, kde $d, p \in \mathbb{N}, d < p$ a $\tilde{\boldsymbol{\psi}}_n$ je maximálně věrohodný odhad $\boldsymbol{\psi}$ za předpokladu platnosti nulové hypotézy. Tyto tři upravené testové statistiky mají podle (Shao, 1999, str. 386-387 Theorem 6.6) za platnosti nulové hypotézy a předpokladů regularity (viz Shao, 1999, str. 249, předpoklady pro Theorem 4.16) asymptoticky rozdělení χ^2_{p-d}. Všechny tři hypotézy H_0^1,H_0^2 a H_0^* jsou na sebe převoditelné (Shao, 1999, str. 386-387 Theorem 6.7) a tedy můžeme na testování použít jakoukoliv z nich.

Aplikace na multinomické rozdělení

Nyní jsme ve stejné sitauci jako v kapitole 2.1. V případě multinomického rozdělení máme $p = K - 1$. To znamená, že vektor $\boldsymbol{p} = (p_1, \ldots, p_K)$ je za platnosti nulové hypotézy vyjádřitelný v parametrickém tvaru jako $\boldsymbol{p}(\boldsymbol{\psi})$, kde $dim(\boldsymbol{\psi}) = d < K - 1$ (D.R. Cox, D.V. Hinkley, 1974, str. 326).

Testujeme nulovou hypotézu

$$H_0 : \boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\psi})$$

proti alternativní hypotéze

$$H_1 : \boldsymbol{p} \neq \boldsymbol{p}(\boldsymbol{\psi}).$$

Začneme se statistikou založenou na poměru věrohodností, která je nejjednodušší na odvození. Stačí si uvědomit, že platí následující:

$$\sup_{\theta \in \Theta} \ell_n(\boldsymbol{\theta}) = \sum_{k=1}^{K} X_k \ln\left(\frac{X_k}{n}\right) + \ln\left(\frac{n!}{X_1!...X_K!}\right)$$

a

$$\sup_{\theta \in \Theta_0} \ell_n(\boldsymbol{\theta}) = \sum_{k=1}^{K} X_k \ln p_k\left(\tilde{\boldsymbol{\psi}}_n\right) + \ln\left(\frac{n!}{X_1!...X_K!}\right).$$

Člen $\sup_{\theta \in \Theta} \ell_n(\boldsymbol{\theta})$ je supremum přes celý parametrický prostor, tedy dosazení maximálně věrohodného odhadu odvozeného v kapitole 1.2. Druhý člen $\sup_{\theta \in \Theta_0} \ell_n(\boldsymbol{\theta})$ je supremum přes paramaterický prostor za platnosti nulové hypotézy, což je dosazení tvaru parametrické funkce společně s maximálně věrohodným odhadem rušivého parametru. Potom testovou statistiku lze vyjádřit jako

$$LR_n^*(\boldsymbol{p}_0(\boldsymbol{\psi})) = 2\sum_{k=1}^{K} X_k \ln\left(\frac{X_k}{np_k(\tilde{\boldsymbol{\psi}}_n)}\right).$$

Pro nalezení Raovy statistiky bychom mohli provést analogický výpočet jako v Lemma 3. Místo rovností (1.2) bychom využívali analogické rovnosti

$$p_K(\tilde{\boldsymbol{\psi}}_n) = 1 - \sum_{k=1}^{K-1} p_k(\tilde{\boldsymbol{\psi}}_n).$$

Dostali bychom výsledek podobného tvaru jako v případě bez rušivých parametrů

$$LM_n^*(\boldsymbol{p}_0(\boldsymbol{\psi})) = \sum_{k=1}^{K} \frac{(X_k - np_k(\tilde{\boldsymbol{\psi}}_n))^2}{np_k(\tilde{\boldsymbol{\psi}}_n)}.$$

Waldova statistika je pro obecný výpočet problematická. Problém je, že neznáme funkci $R(\boldsymbol{\theta})$. V literatuře (např. D.R. Cox, D.V. Hinkley, 1974, str. 326 nebo Shao, 1999, str. 390) se uvádí, že Waldovu statistiku lze získat ze statistiky $LR_n^*(\boldsymbol{p}_0(\boldsymbol{\psi}))$ analogicky jako v případě testu bez rušivých parametrů diskutovaném v kapitole 1.3. Výsledný tvar této statistiky je

$$W_n^*(\boldsymbol{p}_0(\boldsymbol{\psi})) = \sum_{k=1}^{K} \frac{(X_k - np_k(\tilde{\boldsymbol{\psi}}_n))^2}{X_k}.$$

O této statistice alespoň víme, že má asymptoticky stejné rozdělení jako má statistika $LM_n^*(\boldsymbol{p}_0(\boldsymbol{\psi}))$ díky faktu, že za platnosti nulové hypotézy platí

$$\frac{X_k}{np_k(\tilde{\boldsymbol{\psi}}_n))} \xrightarrow[n\to\infty]{P} 1, k = 1, \ldots, K$$

a Cramérově-Sluckého větě (viz např. Anděl, 2007, str. 333 věta B.10).

Z upraveného přístupu a poznámky za statistikou $W_n^*(\boldsymbol{p}_0(\boldsymbol{\psi}))$ vidíme, že testové statistiky budou mít v případě multinomického rozdělení za nulové hypotézy asymptoticky rozdělení χ^2_{K-d-1}, kde d je počet odhadnutých parametrů.

2.3 Vzájemná souvislost uvedených přístupů

Testová statistika v Pearsonově přístupu je tvaru

$$\chi_n^{*2} = \sum_{k=1}^{K} \frac{(X_k - np_k(\tilde{\boldsymbol{\psi}}_n))^2}{np_k(\tilde{\boldsymbol{\psi}}_n)}.$$

Tak jako v případě testu v multinomickém rozdělení bez rušivých parametrů platí

$$LM_n^*(\boldsymbol{p}_0(\boldsymbol{\psi})) = \chi_n^{*2}.$$

To znamená, že v našem případě je Pearsonova statistika přímo Raova statistika.

Waldova statistika je opět Pearsonově statistice velice podobná. Jediný rozdíl je ve jmenovateli. Ve Waldově testu máme ve jmenovateli napozorované četnosti X_k namísto předpokládaných $np_k(\tilde{\boldsymbol{\psi}}_n)$. Znovu platí obdobný závěr jako v případě bez rušivých parametrů. Testové statistiky jsou asymptoticky ekvivalentní.

Nyní uvažujme poslední statistiku z maximálně věrohodného přístupu test založený na podílu věrohodností. Postup je opět zcela analogický jako v kapitole 1.3. Proto ho zde popíšeme stručněji. Připomeneme tvar testové statistiky

$$LR_n(\boldsymbol{p}_0(\boldsymbol{\psi})) = 2\sum_{k=1}^{K} X_k \ln \frac{X_k}{np_k(\tilde{\boldsymbol{\psi}}_n)}.$$

Taylorovým rozvojem logaritmu do druhého řádu a zanedbáním členu $\sum_{k=1}^{K} np_k(\tilde{\boldsymbol{\psi}}_n) \left(\frac{X_k - np_k(\tilde{\boldsymbol{\psi}}_n)}{np_k(\tilde{\boldsymbol{\psi}}_n)}\right)^3$ dostaneme následující výraz

$$LR_n(\boldsymbol{p}_0(\boldsymbol{\psi})) \doteq 2\sum_{k=1}^{K} np_k(\tilde{\boldsymbol{\psi}}_n) \left(\frac{X_k - np_k(\tilde{\boldsymbol{\psi}}_n)}{np_k(\tilde{\boldsymbol{\psi}}_n)}\right) + \sum_{k=1}^{K} np_k(\tilde{\boldsymbol{\psi}}_n) \left(\frac{X_k - np_k(\tilde{\boldsymbol{\psi}}_n)}{np_k(\tilde{\boldsymbol{\psi}}_n)}\right)^2.$$

Po úpravě s pomocí výpočtu (opět se bude první člen rovnat nule)

$$\sum_{k=1}^{K} np_k(\tilde{\boldsymbol{\psi}}_n) \left(\frac{X_k - np_k(\tilde{\boldsymbol{\psi}}_n)}{np_k(\tilde{\boldsymbol{\psi}}_n)}\right) = \sum_{k=1}^{K} \left(X_k - np_k(\tilde{\boldsymbol{\psi}}_n)\right) = (n - n) = 0$$

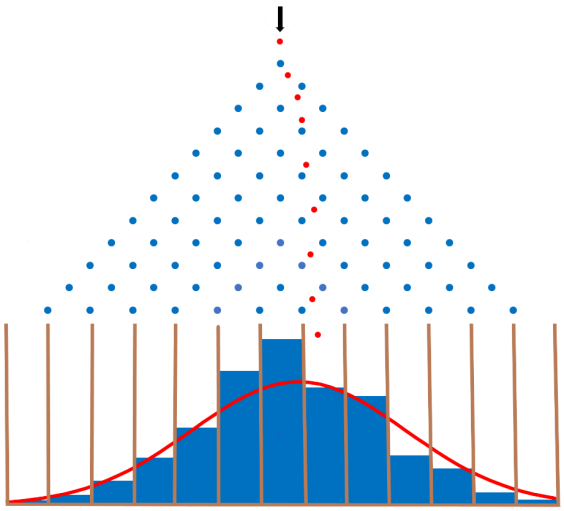
dostaneme

$$LR_n(\boldsymbol{p}_0(\boldsymbol{\psi})) \doteq \sum_{k=1}^{K} \frac{(X_k - np_k(\tilde{\boldsymbol{\psi}}_n))^2}{np_k(\tilde{\boldsymbol{\psi}}_n)}.$$

Tento postup ilustruje souvislost (i v případu s rušivými parametry) Pearsonovy statistiky se statistikou testu založeného na podílu věrohodností. Pearsonova statistika je aproximací testu založeného na podílu věrohodností za předpokladu, že platí $|X_k - np_k(\tilde{\boldsymbol{\psi}}_n)| < np_k(\tilde{\boldsymbol{\psi}}_n)$, pro $k \in \{1, \ldots, K\}$. Význam tohoto předpokladu je analogický jako v kapitole 1.3.

3. Galtonova deska

Galtonova deska má symetricky rozmístěné kolíky, od kterých se kulička odráží doleva nebo doprava a propadá do přihrádek pod kolíky. Pro představu uvádíme obrázek 3.1.



Obrázek 3.1: Galtonova deska - reálné četnosti/hustota normálního rozdělení

Následně budeme uvažovat 3 přístupy na ilustraci námi diskutovaných testů v předchozích kapitolách. Provedli jsme 1500 pozorování na desce s 13 přihrádkami. Výsledky měření jsou uvedeny v tabulce 3.1.

č. přihrádky	0	1	2	3	4	5	6	7	8	9	10	11	12
poč. kuliček	9	19	46	89	146	253	313	222	206	94	69	24	10

Tabulka 3.1: Galtonova deska - dopady kuliček

3.1 Přístup bez rušivého parametru za předpokladu binomického rozdělení s $p = 1/2$

V této kapitole budeme testovat, zda se kuličky odráží s pravděpodobností 1/2. Pravděpodobnost dopadů kuličky je v důsledku určena rozdělením $Bi(12,1/2)$. Výsledky měření/očekávání jsou uvedeny na konci kapitoly 3.3 v tabulce 3.5. Z prvního pohledu na tabulku vidíme, že binomický model dobře nepopisuje situaci. Například v přihrádce číslo 7 bychom čekali, že napozorujeme něco kolem 290 kuliček. Ovšem my jsme napozorovali 222. Těchto „neodpovídajích“ kategorií je mnoho a tedy se můžeme domnívat, že hypotézu zamítneme. Naši domněnku ověříme výpočtem. Abychom mohli porovnávat tento model s dalšími, sloučíme buňky 1, 2, 3 do jedné a symetricky 10, 11, 12 do druhé (pro tento model by bylo možné sloučit jen buňky 1 s 2 a 11 s 12). Máme tedy 8 stupňů volnosti, protože 4 stupně ubyly kvůli sloučení. Hodnoty testových statistik společne s p-hodnotami uvedeme v tabulce 3.2.

Testová statistika	hodnota	p-hodnota
$LM_n(\boldsymbol{p}_0)$	295,81	$3{,}21 \cdot 10^{-59}$
$W_n(\boldsymbol{p}_0)$	123,25	$7{,}06 \cdot 10^{-23}$
$LR_n(\boldsymbol{p}_0)$	100,07	$4{,}12 \cdot 10^{-18}$

Tabulka 3.2: Hodnoty testové statistiky a p-hodnoty

Za hladinu testu jsme zvolili $\alpha = 0{,}05$. Tedy $\chi^2_8(0{,}95) = 15{,}51$ a potom $\chi^2_8(0{,}95) < W(\boldsymbol{p}_0)$, $\chi^2_8(0{,}95) < LR(\boldsymbol{p}_0)$, $\chi^2_8(0{,}95) < LM(\boldsymbol{p}_0)$. Ve všech třech případech zamítáme nulovou hypotézu, že dopad kuličky se řídí podle binomického rozdělení $Bi(12{,}1/2)$. Z uvedeného lze vyvodit, že deska nebyla korektně zkonstruována a naší domněnku, že model neodpovídá realitě jsme si potvrdili.

3.2 Přístup s rušivým parametrem za předpokladu binomického rozdělení

Použijme nyní na příklad postup z kapitoly 2.2. Situace je stejná až na to, že neznáme pravděpodobnost, s jakou se kulička odráží doleva nebo doprava (od každého kolíku se stejnou pravděpodobností). Tato pravděpodobnost je zde rušivým parametrem. Pravděpodobnost dopadů kuličky je určena rozdělením $Bi(12{,}\psi)$ (ψ neznáme). Za nulové hypotézy se dopad kuličky řídí binomickým rozdělením, to znamená, že $p_k = \binom{12}{k}\psi^k(1-\psi)^{12-k}$, $k = 0{,}1, \ldots, 12$, kde p_k je pravděpodobnost, že se kulička odrazí po dopadu na kolík doprava. Rušivý parametr je zde jedno rozměrný parametr ψ.

Testujeme nulovou hypotézu

$$H_0 : \boldsymbol{\theta} = \boldsymbol{p}(\psi),$$

proti alternativě

$$H_1 : \boldsymbol{\theta} \neq \boldsymbol{p}(\psi).$$

Nejdříve musíme pomocí maximální věrohodnosti odhadnout rušivý parametr ψ. Potřebujeme vyřešit věrohodnostní rovnice $\left.\frac{\partial \ell_n(\boldsymbol{\psi})}{\partial \psi}\right|_{\psi=\tilde{\psi}_n} = 0$.
Máme

$$\ell_n(\psi) = \sum_{k=0}^{12} X_k \ln\left(\binom{12}{k}(\tilde{\psi}_n)^k(1-\tilde{\psi}_n)^{12-k}\right) + \ln\left(\frac{n!}{X_0!\ldots X_{12}!}\right).$$

Po zderivování podle ψ, dosazením $\psi = \tilde{\psi}_n$ a položením nule dostáváme

$$\sum_{k=0}^{12}\left(X_k\frac{k}{\tilde{\psi}_n} - X_k\frac{12-k}{1-\tilde{\psi}_n}\right) = 0.$$

Úpravou dostáváme

$$\sum_{k=0}^{12}\left((X_k k(1-\tilde{\psi}_n) - X_k(12-k))\tilde{\psi}_n\right) = 0.$$

Osamostatněním výrazů s $\tilde{\psi}_n$ dostáváme

$$\sum_{k=0}^{12}(kX_k) - \sum_{k=0}^{12}\left(12\tilde{\psi}_n X_k\right) = 0.$$

S využitím vlastnosti $\sum_{k=0}^{12} X_k = n$ dostaneme finální odhad ψ

$$\tilde{\psi}_n = \sum_{k=0}^{12}\frac{kX_k}{12n}.$$

V případě našich naměřených dat máme $\tilde{\psi}_n = 0{,}512611$.

Vypočteme test poměrem věrohodností s rušivými parametry. Tedy potřebujeme nalézt $\sup_{\theta\in\Theta}\ell_n(\boldsymbol{\theta})$ a $\sup_{\theta\in\Theta_0}\ell_n(\boldsymbol{\theta})$. Snadno se nahlédne, že supremum věrohodnosti přes celý parametrický prostor vypadá následovně

$$\sup_{\theta\in\Theta}\ell_n(\boldsymbol{\theta}) = \sum_{k=1}^{K} X_k \ln\left(\frac{X_k}{n}\right) + \ln\left(\frac{n!}{X_1!...X_K!}\right)$$

(je to dosazení maximálně věrohodného odhadu nalezeného v kapitole 1.1). Zajímavější je supremum věrohodnosti za nulové hypotézy. Supremum věrohodnosti za nulové hypotézy

$$\sup_{\theta\in\Theta_0}\ell_n(\boldsymbol{\theta}) = \sum_{k=1}^{K} X_k \ln\left(p_k(\tilde{\psi}_n)\right) + \ln\left(\frac{n!}{X_1!...X_K!}\right),$$

kde $p_k(\tilde{\psi}_n) = \binom{12}{k}(\tilde{\psi}_n)^k(1-\tilde{\psi}_n)^{12-k}$ a $\tilde{\psi}_n = \sum_{k=0}^{12}\frac{X_k k}{12n}$.
Testové statistiky dané

$$LR_n^*(\boldsymbol{\psi}_0) = 2\left(\sup_{\theta\in\Theta}\ell(\boldsymbol{\theta}) - \sup_{\theta\in\Theta_0}\ell(\boldsymbol{\theta})\right) = 2\sum_{k=0}^{12} X_k \ln\left(\frac{X_k}{np_k(\tilde{\boldsymbol{\psi}}_n)}\right),$$

$$LM_n^*(\boldsymbol{p}_0(\boldsymbol{\psi})) = \sum_{k=0}^{12}\frac{(X_k - np_k(\tilde{\boldsymbol{\psi}}_n))^2}{np_k(\tilde{\boldsymbol{\psi}}_n)},$$

$$W_n^*(\boldsymbol{p}_0(\boldsymbol{\psi})) = \sum_{k=0}^{12}\frac{(X_k - np_k(\tilde{\boldsymbol{\psi}}_n))^2}{X_k}$$

mají podle teorie z kapitoly 2.2 asymptoticky rozdělení χ^2_{11}.

Nemůžeme použít Yarnoldovo kritérium, protože se vztahuje na testy při známých parametrech $p_1,\ldots,p_K$. Zbývá nám tedy kritérium, že předpokládané četnosti jsou v každé skupině větší nebo rovné 5 (viz Anděl, 2007, str. 274).

Po ověření podmínek pro použití testu shodnosti jsme sloučili do jedné pozice 1,2,3 tedy $X_{0-2} = X_0 + X_1 + X_2$ a symetricky stejně 10,11,12 tzn. $X_{10-12} = X_{10} + X_{11} + X_{12}$, protože by jinak neplatilo $np_k(\tilde{\psi}_n) \geq 5$. Kvůli tomuto jsme v kapitole 3.1 slučovali 3 a 3 buňky. Sloučení nám bohužel změní logaritmickou věrohodnost. Nová logaritmická věrohodnost je tvaru:

$$\ell_n(\psi) = \sum_{k=3}^{9} X_k \ln\left(\binom{12}{k}(\tilde{\psi}_n)^k(1-\tilde{\psi}_n)^{12-k}\right) +$$

$$+\ln\left(\frac{n!}{X_3!...X_9!X_{0-2}!X_{10-12}!}\right)+X_{0-2}\ln\left(\sum_{k=0}^{2}\left(\binom{12}{k}(\tilde{\psi}_n)^k(1-\tilde{\psi}_n)^{12-k}\right)\right)+$$

$$+X_{10-12}\ln\left(\sum_{k=10}^{12}\left(\binom{12}{k}(\tilde{\psi}_n)^k(1-\tilde{\psi}_n)^{12-k}\right)\right).$$

U této věrohodnosti už bohužel nenalezneme maximum analyticky, a proto se budeme muset spokojit s numerickým řešením $\tilde{\psi}_n^1 = 0{,}511888$, což je velice blízko předešlému odhadu. Znovu vypočteme předpokládané četnosti (tabulka 3.5 na konci kapitoly 3.3).

Opět jako v kapitole 3.1 vidíme, že model nepopisuje situaci velice věrně. Například u první přihrádky (sloučené 0+1+2) bychom čekali za platnosti nulové hypotézy 23 dopadlých kuliček a napozorovali jsme 74 dopadlých kuliček. Opět ověříme výpočtem. Vypočteme hodnoty testových statistik a vyneseme je do tabulky společně s p-hodnotou. Máme 7 stupňů volnosti, protože odhadujeme jeden parametr a 4 skupiny nám ubyly díky sloučení. Hodnoty testových statistik společne s p-hodnotami uvedeme v tabulce 3.3.

Testová statistika	hodnota	p-hodnota
$LM_n^*(\boldsymbol{p}_0(\boldsymbol{\psi}))$	271,07	$9{,}27 \cdot 10^{-55}$
$W_n^*(\boldsymbol{p}_0(\boldsymbol{\psi}))$	119,27	$1{,}09 \cdot 10^{-22}$
$LR_n^*(\boldsymbol{\psi}_0)$	94,42	$1{,}53 \cdot 10^{-17}$

Tabulka 3.3: Hodnoty testové statistiky a p-hodnoty

Za hladinu testu jsme zvolili $\alpha = 0{,}05$. Tedy $\chi_7^2(0{,}95) = 14{,}07$, dále $\chi_7^2(0{,}95) < W_n^*(\boldsymbol{p}_0(\boldsymbol{\psi}))$, $\chi_7^2(0{,}95) < LR_n^*(\boldsymbol{p}_0(\boldsymbol{\psi}))$, $\chi_7^2(0{,}95) < LM_n^*(\boldsymbol{p}_0(\boldsymbol{\psi}))$. Ve všech třech případech zamítáme nulovou hypotézu, že dopad kuličky se řídí podle binomického rozdělení. Z uvedeného lze vyvodit, že kulička se nechová podle binomického rozdělení $Bi(12,p)$. Jinak řečeno všechny kolíky kuličky neodráží s neznámou stejnou pravděpodobností doleva.

3.3 Přístup s rušivým parametrem za předpokladu beta-binomického rozdělení

Jak jsme se přesvědčili v předchozích případech, binomické rozdělení se nehodí k popisu naší situace. Vyzkoušíme, zda bude tzv. beta-binomické rozdělení popisovat skutečnost lépe. Beta-binomické rozdělení je složené z binomického a beta rozdělení. Jinými slovy si lze model představit jako binomické rozdělení, kde pravděpodobnost p je náhodně vygenerovaná z beta rozdělení. Tento přístup povedeme analogicky jako v kapitole 3.2. Za nulové hypotézy platí, že $p_k = \binom{12}{k}\frac{B(k+\alpha,12-k+\beta)}{B(\alpha,\beta)}$, $k = 0{,}1,\ldots,12$, kde p_k je pravděpodobnost, že se kulička odrazí po dopadu na kolík doprava. Položíme $\boldsymbol{\psi} = \begin{pmatrix}\alpha\\ \beta\end{pmatrix}$.

Testujeme nulovou hypotézu

$$H_0 : \boldsymbol{\theta} = \boldsymbol{p}(\boldsymbol{\psi}),$$

proti alternativě

$$H_1 : \boldsymbol{\theta} \neq \boldsymbol{p}(\boldsymbol{\psi}).$$

Nejdříve musíme pomocí maximální věrohodnosti odhadnout rušivý parametr $\boldsymbol{\psi}$. Znovu maximum nepůjde nalézt analyticky a tedy uvedeme jen logaritmickou věrohodnost a maximum najdeme numericky v programu Wolfram Mathematica. Logaritmická věrohodnost je ve tvaru

$$\ell_n(\psi) = \sum_{k=0}^{12} X_k \ln\left(\frac{B(k+\alpha,12-k+\beta)}{B(\alpha,\beta)}\right) + \ln\left(\frac{n!}{X_0!...X_{12}!}\right).$$

Numerické řešení, které maximalizuje věrohodnost je $\tilde{\alpha}_n = 9{,}12246$ a $\tilde{\beta}_n = 8{,}67345$. Po zkontrolování četností zjistíme, že znovu musíme slučovat buňky. Označme $X_{0-2} = X_0 + X_1 + X_2$ a $X_{10-12} = X_{10} + X_{11} + X_{12}$. Zde by jako v případě kapitoly 3.1 stačilo sloučit buňky 0 s 1 a 11 s 12, ale abychom mohli porovnávat s ostatními modely, sloučili jsme 3 a 3. Znovu nám to upraví logaritmickou věrohodnost a dostaneme jiné odhady. Logaritmická věrohodnost je tvaru

$$\ell_n(\psi) = \sum_{k=2}^{10} X_k \ln\left(\frac{B(k+\alpha,12-k+\beta)}{B(\alpha,\beta)}\right) +$$
$$+ \ln\left(\frac{n!}{X_2!...X_{10}!X_{0-2}!X_{10-12}!}\right) + X_{0-2}\ln\left(\frac{B(\alpha,12+\beta)}{B(\alpha,\beta)} + \frac{B(1+\alpha,11+\beta)}{B(\alpha,\beta)}\right) +$$
$$+ X_{10-12}\ln\left(\frac{B(12+\alpha,\beta)}{B(\alpha,\beta)} + \frac{B(11+\alpha,1+\beta)}{B(\alpha,\beta)}\right).$$

Numerické řešení, které maximalizuje věrohodnost je $\tilde{\alpha}_n^1 = 10{,}1691$ a $\tilde{\beta}_n^1 = 9{,}658$. Znovu vyneseme do tabulky 3.5 na konci této kapitoly. Na první pohled se zdá, že tento model je pro popis Galtonovy desky lepší než předešlé. Především u malých četností tzn. na krajích desky model funguje o poznání lépe. Opět přejdeme k vypočtu hodnot testových statistik a vyneseme je do tabulky. Máme 6 stupňů volnosti, protože odhadujeme 2 parametry a 4 stupně nám ubyly díky slučování. Hodnoty testových statistik společně s p-hodnotami uvedeme v tabulce 3.4.

Testová statistika	hodnota	p-hodnota
$LM_n^*(\boldsymbol{p}_0(\boldsymbol{\psi}))$	31,76	$1{,}81 \cdot 10^{-5}$
$W_n^*(\boldsymbol{p}_0(\boldsymbol{\psi}))$	33,82	$7{,}29 \cdot 10^{-6}$
$LR_n^*(\boldsymbol{p}_0(\boldsymbol{\psi}))$	16,12	0,013

Tabulka 3.4: Hodnoty testové statistiky a p-hodnoty

Za hladinu testu jsme zvolili $\alpha = 0{,}05$. Tedy $\chi_6^2(0{,}95) = 12{,}60$, dále $\chi_6^2(0{,}95) < W_n^*(\boldsymbol{p}_0(\boldsymbol{\psi}))$, $\chi_6^2(0{,}95) < LR_n^*(\boldsymbol{p}_0(\boldsymbol{\psi}))$, $\chi_6^2(0{,}95) < LM_n^*(\boldsymbol{p}_0(\boldsymbol{\psi}))$. Ve všech třech případech zamítáme nulovou hypotézu. Ovšem už si můžeme všimnout, že p-hodnoty jsou už daleko větší než u binomického modelu v kapitole 3.1 a 3.2. Kdybychom zvolili hladinu testu $\alpha = 0{,}01$, u testu podílem věrohodností už

bychom dokonce nezamítali. Lze tedy konstatovat, že tento model popisuje naši situaci lépe.

Číslo přihrádky	Empirická četnost	Očekávaná četnost $Bi(12,1/2)$	Očekávaná četnost $Bi(12,\tilde{\psi}_n^1)$	Očekávaná četnost Beta-binomické rozdělení
0+1+2	74	28,9	23,6	64,9
3	89	80,6	69,6	102,7
4	146	181,3	164,3	172,4
5	253	290,0	275,6	234,6
6	313	338,4	337,2	265,1
7	222	290,0	303,1	250,7
8	206	181,3	198,7	196,9
9	94	80,6	92,6	125,6
10+11+12	103	28,9	35,2	87,2
Celkem	1500	1500	1500	1500

Tabulka 3.5: Odhadnuté a očekávané dopady kuliček

Závěr

V práci jsme se zabývali multinomickým rozdělením. Na první pohled se nemusí zdát, že by postupy založené na maximální věrohodnosti, byly problémové. Jak jsme viděli, práce s tímto rozdělením obnášela několik nepříjemností. Nejdříve jsme museli rozdělení přeparametrizovat, abychom mohli vypočíst Fisherovu informační matici. Maximálně věrohodné odhady parametrů je možné odvodit i jinými způsoby (např. pomocí Lagrangeových multiplikátorů). Po přeparametrizování můžeme bez úprav použít základní teorie k vybudování testů bez rušivých parametrů. Jak jsme viděli, testy s rušivými parametry jsou zde nestandardní ve významu, že obecnou teorii (jako je uvedena např. v Anděl, 2007, str. 184), tzn. rozdělení parametru na cílovou a rušivou složku je potřeba upravit na jednu z námi diskutovaných alternativ, tzn. brát cílový vektor jako funkci závislou na rušivém parametru. Přesto se nám podařilo ukázat souvislost Pearsonova přístupu a přístupu založeného na maximální věrohodnosti. Postup jsme ilustrovali na reálných datech.

Seznam použité literatury

J. Anděl (2007). *Základy matematické statistiky*. Druhé opravené vydání. Matfyzpress, Praha. ISBN 80-7378-001-1.

D.R. Cox, D.V. Hinkley (1974). *Theoretical Statistics*. První vydání. London. ISBN 978-0-412-12420-7.

M. Kulich (2000). Asymptotické testy hypotéz v modelech s rušivými parametry. JČMF *Robust 2000*, 125–134.

K. Pearson (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 157–175.

C. Rao (2005). Score Test: Historical Review and Recent Developments. in *Advances in Ranking and Selection, Multiple Comparisons, and Reliability*, 3–20 ,eds. Balakrishnan, N, Kannan, Nandini, Nagaraja, H. N. First edition. Birkhäuser Basel ISBN 978-0-8176-3232-8.

J. Shao (1999). *Mathematical Statistics*. První vydání. Springer-Verlag New York, New York. ISBN 978-0-387-22759-7.

K. Williams (1976). The failure of Pearson's goodness of fit statistic. *The Statistician*, **25**(1), 49.