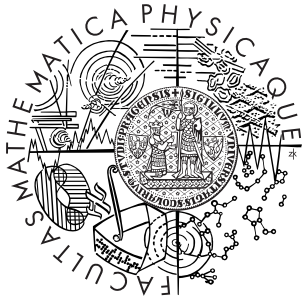
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

# BAKALÁŘSKÁ PRÁCE




Vera Djordjilovič

## Poissonovo rozdělení a příbuzné modely

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Prof. RNDr. Jiří Anděl, DrSc.
Studijní program: matematika, obecná matematika

2009

Děkuji panu Prof. RNDr. Jiřímu Andělovi, DrSc. za odborné vedení mé práce, za rady a za čas, který mi během vypracovávání této práce věnoval.

Prohlašuji, že jsem svou bakalářskou práci napsal(a) samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 28.5.2009. Vera Djordjilovič

# Obsah

**Úvod** **6**

**1 Testy Poissonova rozdělení** **8**
1.1 Test metodou $\chi^2$ . . . . . 8
1.2 Podmíněný test . . . . . 12
1.3 Modifikovaný Kolmogorovův-Smirnovův test . . . . . 14

**2 Intervaly spolehlivosti pro parametr $\lambda$** **16**
2.1 Standardní interval . . . . . 16
2.2 Skórový interval . . . . . 17
2.3 Clopper-Pearsonův interval . . . . . 17

**3 Testy rovnosti parametrů Poissonových rozdělení** **20**
3.1 Asymptotický test . . . . . 20
3.2 Podmíněný test . . . . . 22
3.3 Test založený na zobecněných lineárních modelech . . . . . 24

**4 Alternativní modely** **27**
4.1 Poissonovo rozdělení s nadhodnocenou nulou . . . . . 27
4.1.1 Volba modelu . . . . . 28
4.2 Směs Poissonových rozdělení . . . . . 30
4.3 Useknuté Poissonovo rozdělení . . . . . 33
4.4 Poissonovo rozdělení s modifikovanou nulou . . . . . 35

**A Poznámky k výpočtům** **37**

**Literatura** **39**

Název práce: Poissonovo rozdělení a příbuzné modely
Autor: Vera Djordjilovič
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: prof. RNDr. Jiří Anděl, DrSc
e-mail vedoucího: andel@karlin.mff.cuni.cz

Abstrakt: V předložené práci studujeme Poissonovo rozdělení a jeho modifikace. Popíšeme intervaly spolehlivosti pro parametr Poissonova rozdělení a na příkladě je porovnáme. Popíšeme metody testování hypotézy, že náhodný výběr pochází z Poissonova rozdělení, a v případě zamítnutí nulové hypotézy navrhneme alternativní modely. Mezi ně patří zejména Poissonovo rozdělení s nadhodnocenou nulou, Poissonovo rozdělení s modifikovanou nulou, useknuté Poissonovo rozdělení a směs Poissonových rozdělení. Popíšeme metody odhadu parametrů a testy dobré shody. Pojednáme o testech rovnosti parametrů několika Poissonových rozdělení. Postupy budou ilustrovány na numerických datech pomocí knihoven programu R.

Klíčová slova: Poissonovo rozdělení; Poissonovo rozdělení s modifikovanou nulou; Poissonovo rozdělení s nadhodnocenou nulou; useknuté Poissonovo rozdělení.

Title: Poisson distribution and related models
Author: Vera Djordjilovič
Department: Department of Probability and Mathematical Statistics
Supervisor: Prof. RNDr. Jiří Anděl, DrSc.
Supervisor's e-mail address: andel@karlin.mff.cuni.cz

Abstract: In the present work we study Poisson distribution and some its modifications. Confidence intervals for the parameter of Poisson distribution are constructed. Goodness-of-fit tests for Poisson distrubution are described and in case of rejecting the Poisson model alternative models are presented. These include Zero-Inflated Poisson Model, Zero-Modified Poisson Model, Truncated Poisson Model and Poisson Mixture Model. Methods of estimating parameters and tests of fit are described. We include some tests for equality of parameters of several Poisson distributions. The procedures are demonstrated on numerical examples using packages of program R.

Keywords: Poisson distribution; Zero-modified Poisson Model; Zero-Inflated Poisson Model; Truncated Poisson Model

# Úvod

V roce 1838 Siméon Denis Poisson (1781–1840) ve své práci *Recherches sur la probabilité des jugements en matières criminelles et matière civile* vyšetřoval binomické rozdělení v případě velkého počtu pokusů a malé pravděpodobnosti úspěchu v jednotlivých pokusech. Dokázal, že při $n\to\infty$, $p\to 0$ a $np \to \lambda > 0$ platí

$$\lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}.$$

Od té doby se rozdělení mající pravděpodobnosti

$$\mathsf{P}(X = k) = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots,$$

nazývá Poissonovo a značí $\mathsf{Po}(\lambda)$. Výpočtem dostáváme $\mathsf{E}X = \lambda$, $\mathsf{var}\, X = \lambda$.

Poissonovo rozdělení se často používá jako model pro počet výskytů určitého jevu, např. pro počet volání přicházejících na telefonní ústřednu během jedné minuty, počet zákazníků, kteří přijdou do nějakého systému obsluhy, počet kazů v jednom metru nějaké tkaniny atd.

V případě náhodného výběru, ve kterém se vyskytují nezáporné celočíselné hodnoty, jeden z nejjednodušších modelů je právě Poissonovo rozdělení. V první kapitole pojednáme o různých metodách testování hypotézy, že náhodný výběr pochází z Poissonova rozdělení. V případě zamítnutí nulové hypotézy máme k dispozici různé modifikace základního modelu, zejména Poissonovo rozdělení s modifikovanou nulou, Poissonovo rozdělení s nadhodnocenou nulou a useknuté Poissonovo rozdělení. Tyto modely jsou popsané ve čtvrté kapitole. V druhé kapitole ukážeme, jak se konstruují standardní, skórový a exaktní interval spolehlivosti pro parametr Poissonova rozdělení, a na příkladě je porovnáme.

Někdy se zajímáme o to, zda náhodné výběry pocházejí ze stejného Poissonova rozdělení, někdy jde o vztah mezi parametry dvou Poissonových rozdělení. V třetí kapitole je uvedeno jak postupovat v těchto případech.

Všechny výpočty k příkladům byly provedeny pomocí programu R. Disk, který je v příloze, obsahuje všechna data, použité funkce a vyřešené příklady. V příloze A jsou rovněž uvedeny návody k výpočtům. Použité funkce byly v některých případech převzaty z knihy Simonoff, 2003 a z webových stránek

`http://www.stern.nyu.edu/~jsimonof/AnalCatData/`.

# Kapitola 1

# Testy Poissonova rozdělení

## 1.1 Test metodou $\chi^2$

Nechť $X_1, X_2, \ldots, X_n$ je náhodný výběr z nějakého rozdělení na množině nezáporných celých čísel. Budeme testovat nulovou hypotézu, že jde o výběr z Poissonova rozdělení $\mathsf{Po}(\lambda)$, kde $\lambda$ je neznámý parametr. Test $\chi^2$ se provádí následujícím způsobem. Nejprve se vhodně zvolí čísla $r \geq 0$ a $k \geq 3$ a výběr se rozdělí do $k$ tříd tak, že se do první třídy zařadí pozorování menší nebo rovna $r$, do poslední třídy pozorování větší nebo rovna $r+k-1$ a prostřední třídy jsou tvořeny samostatnými hodnotami $r+1, \ldots, r+k-2$. Označme četnosti jednotlivých tříd $Y_r, Y_{r+1}, \ldots, Y_{r+k-1}$. Označme dále

$$q_i = \frac{\lambda^i \mathrm{e}^{-\lambda}}{i!}, \qquad i = 0, 1, 2, \ldots$$

Pak pravděpodobnosti odpovídající jednotlivým třídám jsou $p_r, \ldots, p_{r+k-1}$, kde

$$p_r = \sum_{i=0}^{r} q_i, \qquad p_i = q_i \text{ pro } i = r+1, \ldots, r+k-2, \qquad p_{r+k-1} = \sum_{i=r+k-1}^{\infty} q_i.$$

Definujme náhodnou veličinu

$$\chi^2 = \sum_{i=r}^{r+k-1} \frac{(Y_i - np_i)^2}{np_i}.$$

Ta dává do souvislosti teoretické a pozorované četnosti, a to takovým způsobem, že čím menší jsou jednotlivé rozdíly, tím menší je $\chi^2$, jinými slovy, malé

hodnoty veličiny $\chi^2$ svědčí ve prospěch nulové hypotézy a velké hodnoty ve prospěch alternativní hypotézy. Veličinu $\chi^2$ zavedl K. Pearson a dokázal, že za předpokladu platnosti nulové hypotézy při $n \to \infty$ má asymptoticky rozdělení $\chi^2_{k-1}$. V našem případě ale $\lambda$ je neznámý parametr. Proto jako testovou statistiku použijeme veličinu

$$\chi^2(\hat{\lambda}) = \sum_{i=r}^{r+k-1} \frac{[Y_i - np_i(\hat{\lambda})]^2}{np_i(\hat{\lambda})},$$

kde $\hat{\lambda}$ je odhad $\lambda$ modifikovanou metodu minimálního $\chi^2$. Odhad $\hat{\lambda}$ je řešením rovnice

$$\sum_{i=r}^{r+k-1} \frac{Y_i}{p_i(\lambda)} \frac{\partial p_i(\lambda)}{\partial \lambda} = 0.$$

Užitím vztahu $\frac{\partial q_i}{\partial \lambda} = (\frac{i}{\lambda} - 1)q_i$ se po úpravě získá rovnice

$$\lambda = \frac{1}{n} \left[ Y_r \frac{\sum_{i=0}^{r} iq_i}{\sum_{i=0}^{r} q_i} + \sum_{i=r+1}^{r+k-2} iY_i + Y_{r+k-1} \frac{\sum_{i=r+k-1}^{\infty} iq_i}{\sum_{i=r+k-1}^{\infty} q_i} \right], \tag{1.1}$$

která se řeší iteračně. Za počáteční aproximaci se bere

$$\lambda_0 = \frac{1}{n} \sum_{i=r}^{r+k-1} iY_i.$$

Statistika $\chi^2(\hat{\lambda})$ také má asymptoticky $\chi^2$ rozdělení, ale jelikož jsme místo skutečné hodnoty $\lambda$ použili odhad, je to $\chi^2$ rozdělení s $k-2$ stupňů volnosti (viz Anděl, 2007, str. 273). Jakmile tedy dostaneme $\chi^2(\hat{\lambda}) \geq \chi^2_{k-2}(\alpha)$, zamítneme hypotézu, že výběr pochází z Poissonova rozdělení. Vzhledem k tomu, že jde o asymptotický test, je hladina testu jen přibližně rovna $\alpha$. K tomu, aby shoda s limitním rozdělením byla dobrá, je zapotřebí, aby teoretické četnosti $np_i(\hat{\lambda})$ byly dostatečně velké (nejméně 5).

**Poznámka.** Jedna z nevýhod tohoto testu spočívá v tom, že při velkém rozsahu výběru i malé (prakticky bezvýznamné) odchylky od Poissonova rozdělení vedou k velké hodnotě statistiky $\chi^2$, a tedy k zamítnutí nulové hypotézy. Přitom použití veličin $(Y_i - np_i)/np_i$ k posouzení věcné významnosti odchylek není vhodné, protože se dává příliš velká váha třídám s malými teoretickými četnostmi. Proto se jako celková míra bere *index nepodobnosti*

(index of dissimilarity)

$$D = \sum_{i=r}^{r+k-1} \frac{|Y_i - n\hat{p}_i|}{2n},$$

kde $\hat{p}_i = p_i(\hat{\lambda})$. Hodnota $D$ je nezáporná a nemůže nabýt hodnoty větší než 1 (viz Simonoff, 2003, str. 78). Není jednoznačně určeno jak velká hodnota $D$ znamená věcnou odchylku ale uvádí se, že hodnoty 0,1 až 0,15 jsou prakticky bezvýznamné.

**Příklad 1.** Stejný postup byl popsán v knize Cramér, 1946, až na řešeni rovnice (1.1). Místo iteračního postupu autor použil aproximaci $\bar{\lambda} = \bar{X}$ a zdůvodnil to následovně. Máme

$$\sum_{i=r+1}^{r+k-2} iY_i = \sum_{\{i;r<X_i<r+k-1\}} X_i,$$

zatímco první a poslední člen v závorce přibližně udávají součet $X_i$ pro které platí $X_i \leq r$ resp. $X_i \geq r + k - 1$. Stačí si uvědomit následující rovnost

$$\mathrm{E}(X|X \leq r) = \frac{\sum_{i=0}^{r} iq_i}{\sum_{i=0}^{r} q_i},$$

kde $X$ je náhodná veličina s rozdělením $\{q_i, i \geq 0\}$.

Podívejme se na příklady tam uvedené a srovnejme výsledky.

V tabulce 1.1 jsou výsledky experimentu, v kterém byl sledován počet vyzařovaných $\alpha$-částic v časovém intervalu délky 7,5 vteřin. Položíme $r = 0$ a sloučíme poslední tři třídy, čímž dostáváme $k = 11$. Modifikovaná metoda minimálního $\chi^2$ dává $\hat{\lambda} = 3{,}8703$, výsledek, který se jen nepatrně liší od $\bar{\lambda} = \bar{X} = 3{,}8700$, což se vysvětluje velkým počtem pozorování ($n = 2608$). Závěr je v obou případech stejný, shoda s Poissonovym rozdělením je dobrá, tj. na žádné „rozumné“ hladině bychom nulovou hypotézu nezamítli.

Tabulka 1.2 udává počet květů rostliny *Primula veris* (prvosenka jarní). Tentokrát máme 200 pozorování, sloučením prvních 3 a posledních 4 tříd dostáváme $r = 0$ a $k = 9$. Shoda s Poissonovym rozdělením už není tak dobrá (na hladině 5 procent bychom ji zamítli), což se vysvětluje zejména velkým počtem rostlin s osmi květy.

Cramér dále uvádí příklad N. G. Holmberga, který zkoumal koncentraci červených krvinek v krvi pomocí speciálního přístroje. Ten je ve tvaru

Tabulka 1.1: $\alpha$-částice

| | $i$ | počet period s $i$ $\alpha$ částicemi | teoretické četnosti | |
|---|---|---|---|---|
| 1 | 0 | 57 | 54,399 | 54,382 |
| 2 | 1 | 203 | 210,523 | 210,476 |
| 3 | 2 | 383 | 407,361 | 407,302 |
| 4 | 3 | 525 | 525,496 | 525,461 |
| 5 | 4 | 532 | 508,418 | 508,423 |
| 6 | 5 | 408 | 393,515 | 393,550 |
| 7 | 6 | 273 | 253,817 | 253,859 |
| 8 | 7 | 139 | 140,325 | 140,359 |
| 9 | 8 | 45 | 67,882 | 67,904 |
| 10 | 9 | 27 | 29,189 | 29,201 |
| 11 | 10 | 10 | 17,075 | 17,084 |
| 12 | 11 | 4 | | |
| 13 | 12 | 2 | | |
| | | | $\bar{\lambda} = 3{,}8700$ | $\hat{\lambda} = 3{,}8703$ |
| | | $\chi^2$ | 12,885 | 12,882 |
| | | $p$–hodnota | 0,16788 | 0,16801 |

obdélníka a obsahuje přihrádky různých (známých) velikostí. Pomoci mikroskopu pak sledoval počet krvinek v jedné z nich a tím získal odhad jejich koncentrace. V tomto případě máme 169 pozorování, odhady $\bar{\lambda} = 11{,}911$ a $\hat{\lambda} = 11{,}921$ dávají stejný výsledek, shoda s Poissonovym rozdělením je velmi dobrá.

**Poznámka.** Z našich příkladů vyplývá, že odhad získaný modifikovanou metodou minimálního $\chi^2$ se neliší o moc od výběrového průměru. Chernoff a Lehmann (1954) však ukázali, že za určitých předpokladů regularity při $n \to \infty$ platí

$$\chi^2(\bar{\lambda}) = \frac{[Y_i - np_i(\bar{\lambda})]^2}{np_i(\bar{\lambda})} \xrightarrow{d} \sum_{i=1}^{k-2} Z_i^2 + \gamma Z_{k-1}^2,$$

kde $Z_i$ jsou nezávislé stejně rozdělené náhodné veličiny, $i = 1, 2, \ldots, k-1$, $Z_1 \sim \mathsf{N}(0,1)$ a $0 \leq \gamma \leq 1$ obecně závisí na $\lambda$. Odtud za platnosti nulové hypotézy plyne

$$\mathsf{P}\left[\chi^2(\bar{\lambda}) \geq \chi^2_{k-2}(\alpha)\right] \overset{n\to\infty}{\longrightarrow} \beta > \alpha,$$

Tabulka 1.2: počet květů *Primula veris*

| | $i$ | počet rostlin s $i$ květy | teoretické četnosti | |
|---|---|---|---|---|
| 1 | 3 | 5 | | |
| 2 | 4 | 2 | | |
| 3 | 5 | 10 | 25,022 | 25,143 |
| 4 | 6 | 19 | 19,136 | 19,193 |
| 5 | 7 | 20 | 24,193 | 24,240 |
| 6 | 8 | 42 | 26,764 | 26,788 |
| 7 | 9 | 27 | 26,318 | 26,314 |
| 8 | 10 | 25 | 23,291 | 23,263 |
| 9 | 11 | 23 | 18,739 | 18,696 |
| 10 | 12 | 11 | 13,820 | 13,774 |
| 11 | 13 | 5 | 22,717 | 22,589 |
| 12 | 14 | 6 | | |
| 13 | 15 | 4 | | |
| 14 | 16 | | | |
| 15 | 17 | | | |
| 16 | 18 | | | |
| 17 | 19 | | | |
| 18 | 20 | 1 | | |
| | | | $\bar{\lambda} = 8{,}850$ | $\hat{\lambda} = 8{,}841$ |
| | | $\chi^2$ | 15,647 | 15,638 |
| | | $p$–hodnota | 0,03 | 0,03 |

tj. asymptotická pravděpodobnost chyby prvního druhu je větší než požadované $\alpha$. Uvádí se však, že aspoň v případě Poissonova rozdělení rozdíl $|\beta - \alpha|$ je příliš malý, než aby výrazně ovlivnil průběh testu, což souhlasí s našimi výsledky.

## 1.2 Podmíněný test

Nechť $X_1, X_2, \ldots, X_n$ je náhodný výběr z Poissonova rozdělení. Pak je sdružené rozdělení veličin $X_1, X_2, \ldots, X_n$ dáno vzorcem

$$\mathsf{P}(X_1 = x_1, \ldots, X_n = x_n) = \frac{\mathrm{e}^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{x_1! x_2! \ldots x_n!}.$$

Tabulka 1.3: Červené krvinky

| | $i$ | počet přihrádek s $i$ červenými krvinkami | teoretické četnosti | |
|---|---|---|---|---|
| 1 | 4 | 1 | | |
| 2 | 5 | 3 | | |
| 3 | 6 | 5 | 8,136 | 8,091 |
| 4 | 7 | 8 | 7,660 | 7,628 |
| 5 | 8 | 13 | 11,404 | 11,367 |
| 6 | 9 | 14 | 15,093 | 15,056 |
| 7 | 10 | 15 | 17,977 | 17,948 |
| 8 | 11 | 15 | 19,466 | 19,451 |
| 9 | 12 | 21 | 19,322 | 19,323 |
| 10 | 13 | 18 | 17,703 | 17,719 |
| 11 | 14 | 17 | 15,062 | 15,088 |
| 12 | 15 | 16 | 11,960 | 11,991 |
| 13 | 16 | 9 | 8,903 | 8,934 |
| 14 | 17 | 6 | 16,314 | 16,403 |
| 15 | 18 | 3 | | |
| 16 | 19 | 2 | | |
| 17 | 20 | 2 | | |
| 18 | 21 | 1 | | |
| | | | $\bar{\lambda} = 11{,}911$ | $\hat{\lambda} = 11{,}921$ |
| | | $\chi^2$ | 4,022 | 4,017 |
| | | $p$–hodnota | 0,946 | 0,947 |

Náhodná veličina $T = \sum_{i=1}^{n} X_i$ má Poissonovo rozdělení s parametrem $n\lambda$, a tedy

$$\mathsf{P}(T = t) = \frac{(n\lambda)^t \mathrm{e}^{-n\lambda}}{t!}, \qquad t = 0, 1, 2, \ldots$$

Potom platí

$$\begin{aligned}
\mathsf{P}(X_1 = x_1, \ldots, X_n = x_n | T = t) &= \frac{\mathsf{P}(X_1 = x_1, \ldots, X_n = x_n, T = t)}{\mathsf{P}(T = t)} \\
&= \frac{\mathrm{e}^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i}}{x_1! x_2! \ldots x_n! \frac{(n\lambda)^t \mathrm{e}^{-n\lambda}}{t!}} \\
&= \frac{t!}{x_1! \ldots x_n!} \left(\frac{1}{n}\right)^t \quad \text{pro} \sum_{i=1}^{n} x_i = t \geq 1.
\end{aligned}$$

Odtud plyne, že podmíněné rozdělení veličin $X_1, X_2, \ldots, X_n$ při daném $T = t$ je multinomické $M\left(t, \frac{1}{n}(1, 1, \ldots, 1)'\right)$. Pak Pearsonova statistika

$$Q = \sum_{i=1}^{n} \frac{(X_i - \frac{t}{n})^2}{\frac{t}{n}} \tag{1.2}$$

má asymptoticky rozdělení $\chi^2_{n-1}$ (viz Anděl, 2007). Podmíněný test je založen na této statistice. Pokud dostaneme

$$Q \leq \chi^2_{n-1}\left(1 - \frac{\alpha}{2}\right) \qquad \text{nebo} \qquad Q \geq \chi^2_{n-1}\left(\frac{\alpha}{2}\right),$$

zamítneme hypotézu, že výběr pochází z Poissonova rozdělení. Aby shoda s limitním rozdělením byla dobrá, je zapotřebí, aby průměr $\bar{X}$ nebyl příliš malý, obvykle se uvádí podmínka $\bar{X} \geq 5$.

## 1.3 Modifikovaný Kolmogorovův-Smirnovův test

Jiný test dobré shody pro Poissonovo rozdělení je založen na modifikaci standardního testu Kolmogorova-Smirnova (dále K-S test), který se používá především v případech testování, zda náhodný výběr pochází z rozdělení se známými parametry, a dává dobré výsledky pro spojitá rozdělení (viz Campbell, Oprian, 1979). V ostatních případech je velice konzervativní a jeho použití se nedoporučuje.

Obvykle při testování, zda náhodný výběr $X_1, \ldots, X_n$ pochází z $\mathsf{Po}(\lambda)$, je $\lambda$ neznámý parametr. Ve svém článku Campbell a Oprian navrhli novou tabulku kritických hodnot, speciálně pro Poissonovo rozdělení. Test se provádí následovně. Nejprve se spočte $\bar{X}$ a empirická distribuční funkci $S_n$

$$S_n(x) = \frac{\sum_{i=1}^{n} \mathbf{I}[X_i \leq x]}{n}.$$

Označme $F_0(x) = \mathrm{P}(X \leq x)$, kde $X \sim \mathsf{Po}(\bar{X})$. Pak spočtěme testovou statistiku

$$D = \sup|F_0(x) - S_n(x)|$$

a překročí-li $D$ kritickou hodnotu z tabulky, zamítne se hypotéza, že výběr pochází z Poissonova rozdělení.

Ukazuje se, že je síla modifikovaného K-S testu vyšší než síla $\chi^2$-testu, a jako další výhoda se uvádí možnost použití jíž pro malá $n$.

# Kapitola 2

# Intervaly spolehlivosti pro parametr $\lambda$

Ukážeme nejprve intervalový odhad založený na centrální limitní větě.

## 2.1 Standardní interval

Mějme $X_1, X_2, \ldots, X_n$ náhodný výběr z Poissonova rozdělení $\mathsf{Po}(\lambda)$. Z centrální limitní věty pak plyne

$$\frac{\bar{X} - \lambda}{\sqrt{\frac{\lambda}{n}}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1) \qquad \text{pro } n \to \infty.$$

Dle zákona velkých čísel platí $\bar{X} \xrightarrow{\mathrm{P}} \lambda$. Odtud dostáváme

$$\frac{\bar{X} - \lambda}{\sqrt{\frac{\bar{X}}{n}}} \sqrt{\frac{\bar{X}}{\lambda}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1)$$

a pak dle Cramérovy-Sluckého věty platí

$$\xi = \frac{\bar{X} - \lambda}{\sqrt{\frac{\bar{X}}{n}}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1).$$

Proto $\mathrm{P}(|\xi| \leq u_{\frac{\alpha}{2}}) \to 1 - \alpha$, kde $u_{\frac{\alpha}{2}}$ je kritická hodnota standardního normálního rozdělení. Pak

$$\left( \bar{X} - \sqrt{\frac{\bar{X}}{n}} u_{\frac{\alpha}{2}}, \bar{X} + \sqrt{\frac{\bar{X}}{n}} u_{\frac{\alpha}{2}} \right)$$

je interval spolehlivosti pro $\lambda$ s koeficientem spolehlivosti, který při $n \to \infty$ konverguje k $\alpha$.

## 2.2 Skórový interval

Vzhledem k tomu, že je odhad rozptylu $\bar{X}$ i sám funkcí $\bar{X}$, může se stát, že je skutečné pokrytí standardního intervalu mnohem menší než $1-\alpha$. Vylepšení dosáhneme následovně. Vyjdeme z relace

$$\mathrm{P}\left(\frac{|\bar{X}-\lambda|}{\sqrt{\frac{\lambda}{n}}} \leq u_{\frac{\alpha}{2}}\right) \to 1-\alpha$$

a pokusíme se najít interval $(\lambda_l, \lambda_p)$ takový, že platí

$$\lambda \in (\lambda_l, \lambda_p) \Leftrightarrow |\bar{X}-\lambda| \leq u_{\frac{\alpha}{2}}\sqrt{\frac{\lambda}{n}}.$$

Jinými slovy, hledáme řešení rovnice $|\bar{X}-\lambda| = u_{\frac{\alpha}{2}}\sqrt{\frac{\lambda}{n}}$. Máme

$$\begin{aligned}
\bar{X}^2 - 2\lambda\bar{X} + \lambda^2 &= \frac{\lambda}{n}u^2, \\
\lambda^2 - \left(2\bar{X} + \frac{u^2}{n}\right)\lambda + \bar{X}^2 &= 0, \\
\lambda_{1,2} &= \bar{X} + \frac{u^2}{2n} \mp \frac{u}{\sqrt{n}}\sqrt{\frac{u^2}{4n} + \bar{X}}.
\end{aligned}$$

Tato řešení nám dávají levý, resp. pravý krajní bod skórového intervalu

$$\left(\bar{X} + \frac{u^2}{2n} - \frac{u}{\sqrt{n}}\sqrt{\frac{u^2}{4n} + \bar{X}}, \quad \bar{X} + \frac{u^2}{2n} + \frac{u}{\sqrt{n}}\sqrt{\frac{u^2}{4n} + \bar{X}}\right)$$

s asymptotickým koeficientem spolehlivosti $\alpha$. V knize Hátle, Likeš je uvedeno, že je tato aproximace použitelná pro $n > 9/\lambda$.

## 2.3 Clopper-Pearsonův interval

Standardní a skórový interval dávají dobré výsledky pro velká $n$, dokonce i pro malá $n$, pokud je $\lambda$ dostatečně velké. V ostatních případech se používá

exaktní interval $(\lambda_l, \lambda_p)$, kde

$$\mathsf{P}(X \geq x|\lambda = \lambda_l) = \alpha/2, \qquad \mathsf{P}(X \leq x|\lambda = \lambda_p) = \alpha/2,$$

přičemž $X = \sum_{i=1}^{n} X_i$. Pro $x = 0$ položíme $\lambda_l = 0$.

Nyní ukážeme, že pravděpodobnosti $\mathsf{P}(Y \leq y)$, kde $Y \sim \mathsf{Po}(\lambda)$, lze vyjádřit pomocí rozdělení $\chi^2_{2(y+1)}$. Připomeňme hustotu $\chi^2$ rozdělení o $n$ stupních volnosti

$$f(x) = \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} x^{n/2-1}\mathrm{e}^{-x/2}, \qquad x > 0.$$

Potom při $n = 2(y+1)$ máme

$$\Gamma(y+1)\mathsf{P}\left(\chi^2 > 2\lambda\right) = \frac{1}{2^{y+1}} \int_{2\lambda}^{\infty} \mathrm{e}^{-\frac{\chi^2}{2}} (\chi^2)^y \,\mathrm{d}\chi^2.$$

Integrací per partes dostáváme

$$\begin{aligned}
\frac{1}{2^{y+1}} \int_{2\lambda}^{\infty} \mathrm{e}^{-\frac{\chi^2}{2}} (\chi^2)^y \,\mathrm{d}\chi^2 &= \int_{\lambda}^{\infty} \mathrm{e}^{-z} z^y \,\mathrm{d}z \\
&= \mathrm{e}^{-\lambda}\lambda^y + y\mathrm{e}^{-\lambda}\lambda^{y-1} + \cdots + y(y-1)\ldots 1 \cdot \mathrm{e}^{-\lambda} \\
&= \sum_{t=0}^{y} \frac{y!}{t!}\mathrm{e}^{-\lambda}\lambda^t \\
&= \Gamma(y+1) \sum_{t=0}^{y} \frac{\lambda^t}{t!}\mathrm{e}^{-\lambda} \\
&= \Gamma(y+1)\mathsf{P}(Y \leq y).
\end{aligned}$$

Pak pro $X \sim \mathsf{Po}(n\lambda)$ platí

$$\begin{aligned}
\frac{\alpha}{2} &= \mathsf{P}(X \geq x|\lambda = \lambda_l) \\
&= 1 - \mathsf{P}(X \leq x - 1|\lambda = \lambda_l) \\
&= 1 - \mathsf{P}(\chi^2_{2x} > 2n\lambda_l) \qquad \text{pro } x > 0.
\end{aligned}$$

Odtud máme $2n\lambda_l = \chi^2_{2x,1-\frac{\alpha}{2}}$ a podobně postupujeme pro $\lambda_p$. Konečně dostáváme

$$\lambda_l = \begin{cases} 0 & \text{pro } x = 0, \\ \frac{1}{2n}\chi^2_{2x,1-\frac{\alpha}{2}} & \text{pro } x > 0, \end{cases} \qquad \lambda_p = \frac{1}{2n}\chi^2_{2x+2,\frac{\alpha}{2}}.$$

Tabulka 2.1: vstřelené a obdržené branky

| Počet branek | Vstřelené | Obdržené |
|---|---|---|
| 0 | 3 | 6 |
| 1 | 9 | 21 |
| 2 | 24 | 17 |
| 3 | 18 | 16 |
| 4 | 14 | 12 |
| 5 | 7 | 9 |
| 6 | 3 | 1 |
| 7 | 2 | |
| 8 | 1 | |
| 9 | 1 | |

**Příklad.** V tabulce 2.1 je zaznamenán počet vstřelených a obdržených branek hokejového týmu New Jersey Devils v sezóně 1999-2000 (Simonoff, 2003, str. 72). Testujme hypotézu, že jde o výběry z Poissonova rozdělení. V případě vstřelených branek máme $\bar{x} = 3{,}06$, $\chi^2 = 3{,}16$ (5 stupňů volnosti) a příslušnou $p$-hodnotu 0,675. V případě obdržených branek máme $\bar{x} = 2{,}46$, $\chi^2 = 2{,}32$ (4 stupně volnosti) a $p$-hodnotu 0,677. Nezamítáme tedy hypotézu, že jde o výběry z Poissonova rozdělení.

Standardní 95% interval spolehlivosti pro počet vstřelených branek je (2,68; 3,44) a skórový interval je (2,71; 3,46). V případě obdržených branek máme (2,12; 2,80) a (2,15; 2,83). Vidíme, že v těchto příkladech tyto intervaly vycházejí velmi podobně. Přesné intervaly spolehlivosti jsou (2,69; 3,46) resp. (2,14; 2,83).

# Kapitola 3

# Testy rovnosti parametrů Poissonových rozdělení

## 3.1 Asymptotický test

Mějme náhodný výběr $\mathbf{X}_1 = (X_{11}, \ldots, X_{1n_1})$ z rozdělení $\mathsf{Po}(\lambda_1)$, náhodný výběr $\mathbf{X}_2 = (X_{21}, \ldots, X_{2n_2})$ z rozdělení $\mathsf{Po}(\lambda_2), \ldots$, náhodný výběr $\mathbf{X}_I = (X_{I1}, \ldots, X_{In_I})$ z rozdělení $\mathsf{Po}(\lambda_I)$, $I \geq 2$, a nechť výběry $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_I$ jsou vzájemně nezávislé.

Testujme hypotézu $H_0 : \lambda_1 = \lambda_2 = \cdots = \lambda_I$ proti alternativě, že aspoň dva z těchto parametru jsou rozdílné. Označme

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \quad i = 1, \ldots, I, \qquad N = \sum_{i=1}^{I} n_i, \quad \bar{x} = \frac{1}{N} \sum_{i=1}^{I} n_i \bar{x}_i.$$

Náhodné veličiny

$$u_i = \frac{\bar{x}_i - \lambda_i}{\sqrt{\lambda_i}} \sqrt{n_i}, \qquad i = 1, \ldots, I,$$

jsou nezávislé a z centrální limitní věty plyne, že všechny mají asymptoticky normální rozdělení $\mathsf{N}(0, 1)$.

Za platnosti nulové hypotézy je $u_i = (\bar{x}_i - \lambda_1)\sqrt{n_i}/\sqrt{\lambda_1}$, $i = 1, \ldots, I$, a statistika $\bar{x}$ je konsistentním odhadem společného rozptylu všech $I$ rozdělení. Pak $\sqrt{\bar{x}}$ je konsistentním odhadem směrodatné odchylky (viz Hátle, Likeš 1972, str. 215).

Dle Cramérovy-Sluckého věty náhodné veličiny

$$\xi_i = (\bar{x}_i - \lambda_1)\sqrt{n_i}/\sqrt{\bar{x}}, \qquad i = 1, \ldots, I$$

také mají asymptoticky standardní normální rozdělení. Pak veličina

$$\chi^2 = \sum_{i=1}^{I} \left( \xi_i - \sqrt{\frac{n_i}{\bar{x}}} \bar{\xi} \right)^2,$$

kde $\bar{\xi} = \sum_{i=1}^{I} \xi_i / I$, má asymptoticky $\chi^2$ rozdělení o $I - 1$ stupních volnosti (viz Hátle, Likeš 1972, str. 320) a po úpravě dostáváme

$$\chi^2 = \frac{1}{\bar{x}} \left[ \sum_{i=1}^{I} n_i \bar{x}_i{}^2 - \frac{1}{N} \left( \sum_{i=1}^{I} n_i \bar{x}_i \right)^2 \right]. \tag{3.1}$$

Za platnosti $H_0$ bude $\mathsf{P}\left[\chi^2 \geq \chi^2_{I-1,\alpha}\right]$ přibližně $\alpha$. Nulovou hypotézu tedy zamítneme, když bude platit $\chi^2 \geq \chi^2_{I-1,\alpha}$.

**Poznámka.** Pro $n_1 = \cdots = n_I = n$ máme $N = nI$ a $\bar{x} = \sum_{i=1}^{I} \bar{x}_i / I$, takže statistika 3.1 má tvar

$$\begin{aligned} \chi^2 &= \frac{I}{\sum_{i=1}^{I} \bar{x}_i} \left[ n \sum_{i=1}^{I} \bar{x}_i{}^2 - \frac{1}{nI} n^2 \left( \sum_{i=1}^{I} \bar{x}_i \right)^2 \right] \\ &= \frac{N \sum_{i=1}^{I} \bar{x}_i{}^2}{\sum_{i=1}^{I} \bar{x}_i} - n \sum_{i=1}^{I} \bar{x}_i. \end{aligned}$$

V případě $I = 2$ můžeme postupovat jinak. Využijeme asymptotické normality výběrových průměrů

$$\bar{x}_i \xrightarrow{d} \mathsf{N}\left( \lambda_i, \frac{\lambda_i}{n_i} \right), \qquad i = 1, 2.$$

Pak máme

$$\bar{x}_1 - \bar{x}_2 \xrightarrow{d} \mathsf{N}\left( \lambda_1 - \lambda_2, \frac{\lambda_1}{n_1} + \frac{\lambda_2}{n_2} \right).$$

Za platnosti nulové hypotézy náhodná veličina

$$u = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\bar{x}\left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

má asymptoticky $\mathsf{N}(0,1)$ rozdělení.

V případě pravostranné alternativy $H_1 : \lambda_1 - \lambda_2 > 0$ nulovou hypotézu zamítneme, když bude platit $u \geq u_\alpha$, a v případě levostranné alternativy $H_1 : \lambda_1 - \lambda_2 < 0$ dostaneme-li $u \leq -u_\alpha$, kde $u_\alpha$ je kritická hodnota $\mathsf{N}(0,1)$. V případě oboustranné alternativy nulovou hypotézu zamítneme platí-li, $|u| \geq u_{\alpha/2}$.

## 3.2 Podmíněný test

Nechť $X_1$ má Poissonovo rozdělení s parametrem $\lambda_1$ a $X_2$ Poissonovo rozdělení s parametrem $\lambda_2$. Chceme testovat hypotézu $H_0 : \gamma = \lambda_1/\lambda_2 = \gamma_0$, kde $\gamma_0$ je dané kladné číslo.

Označme $S = X_1 + X_2$. Pak $S \sim \mathsf{Po}[(\gamma+1)\lambda_2]$. Potom je podmíněné rozdělení $X_1, X_2$ při daném $S = s$ dáno vzorcem

$$\begin{aligned} \mathsf{P}\left(X_1 = x_1, X_2 = x_2 | S = s\right) &= \frac{\mathrm{e}^{-\gamma\lambda_2}\frac{(\gamma\lambda_2)^{x_1}}{x_1!}\mathrm{e}^{-\lambda_2}\frac{\lambda_2^{x_2}}{x_2!}}{\mathrm{e}^{-(\gamma\lambda_2+\lambda_2)}\frac{(\gamma\lambda_2+\lambda_2)^s}{s!}} \\ &= \binom{s}{x_1}\left(\frac{\gamma}{\gamma+1}\right)^{x_1}\left(1-\frac{\gamma}{\gamma+1}\right)^{x_2}, \end{aligned}$$

při $s = x_1 + x_2$. Odtud vyplývá, že podmíněné rozdělení $X_1$ při daném $S = s$ je $\mathsf{Bi}(s, \gamma/\gamma+1)$ a za platnosti nulové hypotézy $\mathsf{Bi}(s, \gamma_0/\gamma_0+1)$.

Testujme nulovou hypotézu proti pravostranné alternativě $H_1 : \gamma > \gamma_0$. Pak ve prospěch alternativní hypotézy svědčí velká hodnota $x_1$, jinými slovy nulovou hypotézu zamítneme, dostaneme-li

$$\sum_{t=x_1}^{s}\binom{s}{t}\left(\frac{\gamma_0}{\gamma_0+1}\right)^t\left(1-\frac{\gamma_0}{\gamma_0+1}\right)^{s-t} \leq \alpha. \tag{3.2}$$

Výpočet se zjednoduší užitím rovnosti

$$\sum_{t=0}^{x}\binom{n}{t}p^t(1-p)^{n-t} = G\left(\frac{x+1}{n-x}\frac{1-p}{p}\right),$$

kde $G$ je distribuční funkce $\mathsf{F}$ rozdělení o $2(n-x)$ a $2(x+1)$ stupních volnosti (viz Hátle, Likeš 1972, str. 141). Podmínka 3.2 je tedy ekvivalentní podmínce

$$\frac{x_1}{x_2+1} \geq \gamma_0 F_{2x_2+2,2x_1}(\alpha),$$

kde $F_{m,n}(\alpha)$ je kritická hodnota rozdělení $\mathsf{F}_{m,n}$.

Podobně postupujeme v případě levostranné alternativy $H_1 : \gamma < \gamma_0$. Hypotézu zamítneme, jestliže

$$\sum_{t=0}^{x_1} \binom{s}{t} \left(\frac{\gamma_0}{\gamma_0+1}\right)^t \left(1-\frac{\gamma_0}{\gamma_0+1}\right)^{s-t} \leq \alpha$$

nebo pokud

$$\frac{x_1+1}{x_2} \leq \gamma_0 F_{2x_2,2x_1+2}\left(1-\alpha\right).$$

V případě oboustranné alternativy $H_1 : \gamma \neq \gamma_0$ nulovou hypotézu zamítneme, jestliže

$$\begin{aligned}
\sum_{t=0}^{x_1} \binom{s}{t} \left(\frac{\gamma_0}{\gamma_0+1}\right)^t \left(1-\frac{\gamma_0}{\gamma_0+1}\right)^{s-t} &\leq \frac{\alpha}{2} \quad \text{nebo} \\
\sum_{t=x_1}^{s} \binom{s}{t} \left(\frac{\gamma_0}{\gamma_0+1}\right)^t \left(1-\frac{\gamma_0}{\gamma_0+1}\right)^{s-t} &\leq \frac{\alpha}{2},
\end{aligned}$$

tj. jestliže

$$\frac{x_1}{x_2-1} \geq \gamma_0 F_{2x_2+2,2x_1}\left(\frac{\alpha}{2}\right) \quad \text{nebo} \quad \frac{x_1+1}{x_2} \leq \gamma_0 F_{2x_2,2x_1+2}\left(1-\frac{\alpha}{2}\right).$$

**Příklad.** Následující příklad je uveden v knize Hátle, Likeš (1972) na str. 334. Ze zkušenosti je známo, že se počet poruch určitého zařízení za 100 hodin provozu řídí Poissonovým rozdělením s parametrem $\lambda$. Testujme hypotézu, že dvě nová zařízení $A$ a $B$ mají stejný parametr $\lambda$, proti alternativě, že $A$ má vyšší střední počet poruch za 100 hodin provozu. U zařízení $A$ se během 100 hodin provozu vyskytlo 8 poruch a u zařízení $B$ 5 poruch.

Testujeme nulovou hypotézu $H_0 : \lambda_1/\lambda_2 = 1$ proti pravostranné alternativě $H_1 : \lambda_1/\lambda_2 > 1$. Nulovou hypotézu bychom zamítli na takové hladině $\alpha$, která splňuje

$$\frac{8}{6} \geq F_{12,16}\left(\alpha\right).$$

Jak známo, nejmenší takové $\alpha$ je $p$-hodnota, a v našem případě statistický software dává $\alpha \geq 0.29$, tedy nulovou hypotézu nezamítáme.

## 3.3 Test založený na zobecněných lineárních modelech

Ukazuje se, že lineární regresní model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, kde $\mathbf{X} = \{x_{ij}\}$ je matice daných čísel typu $n \times p$, kde $n > p$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ je vektor neznámých parametrů a $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ je náhodný vektor mající rozdělení $\mathsf{N}(0, \sigma^2\mathbf{I})$ při neznámém $\sigma^2 > 0$, nedává dobré výsledky v situacích, kdy je silně porušen předpoklad normality veličin $Y_i$. Navíc bychom chtěli zobecnit tento model pro případ, kdy $Y_i$ nabývají diskrétních celočíselných hodnot. Pro takové případy byla vyvinuta teorie zobecněných lineárních modelů.

Regresní model musí specifikovat distribuci $Y_i$ (náhodná komponenta), způsob, kterým $x_{ki}$, $k = 1, \ldots, p$ ovlivňují $Y_i$ (systematická komponenta) a vztah mezi náhodnou a systematickou komponentou (tzv. linkovou funkci). Zobecněný lineární model je určen speciální volbou náhodné a systematické komponenty a jejich vztahu.

Náhodná komponenta vyžaduje, aby hustota $Y_i$ byla hustota exponenciálního typu

$$f\left(y_i, \theta_i, \phi\right) = \exp\left[\frac{y_i\theta_i - b\left(\theta_i\right)}{a\left(\phi\right)} + c\left(y_i, \phi\right)\right],$$

kde $a, b, c$ jsou známé funkce splňující

$$\mu_i = \mathsf{E}Y_i = b'\left(\theta_i\right)$$

a

$$\mathsf{var}Y_i = a\left(\phi\right) b''\left(\theta_i\right).$$

Parametr $\theta_i$ závisí na $x_{ji}$, zatímco $\phi$ je obvykle známé.

Systematická komponenta má být lineární funkcí parametrů

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{pi}.$$

Linková funkce $g$ pak určuje vztah mezi $\eta_i$ a $\mu_i = \mathsf{E}Y_i$ následovně

$$g(\mu_i) = \eta_i.$$

Vektor parametrů se nejčastěji odhaduje metodou maximální věrohodnosti. Logaritmická věrohodnostní funkce vektoru $\mathbf{Y}$ je dána vzorcem

$$L = \sum_{i=1}^{n}\left[\frac{y_i\theta_i - b\left(\theta_i\right)}{a_i\left(\phi\right)} + c\left(y_i, \phi\right)\right]. \tag{3.3}$$

Jak uvádí Simonoff (2003) na str. 127, tato metoda vede na soustavu

$$\mathbf{X}'\mathbf{W}\mathbf{r} = \mathbf{0}, \tag{3.4}$$

kde

$$\mathbf{W} = \operatorname{diag}\left[\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 / V(y_i)\right] \text{ a } \quad r_i = (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}.$$

Odtud máme

$$\mathbf{X}'\mathbf{W}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{W}\mathbf{z},$$

kde $\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{r}$. Konečně dostáváme rovnici

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z},$$

která se řeší iteračně.

Chceme-li testovat hypotézu $H_0 : \beta_1 = \cdots = \beta_p = 0$ proti alternativě, že aspoň jeden z těchto parametrů je nenulový, použijeme test založený na věrohodnostním poměru. Máme

$$LR = 2(L_1 - L_2), \tag{3.5}$$

kde $L_1$ a $L_2$ jsou maximální hodnoty logaritmických věrohodnostních funkcí v obecném resp. zjednodušeném modelu. Za platnosti nulové hypotézy má statistika $LR$ asymptoticky rozdělení $\chi^2_p$.

**Poznámka.** V případě Poissonova rozdělení s parametrem $\lambda$ máme

$$f(y) = \exp\left[y\log(\lambda) - \lambda - \log(y!)\right],$$

tedy jde o rozdělení exponenciálního typu, stačí položit

$$\theta = \log\lambda, \quad \phi = a(\phi) = 1, \quad b(\theta) = \mathrm{e}^{\theta}, \quad c(y) = \log(y!).$$

Speciální případ zobecněného lineárního modelu je *Poissonův regresní model*, kde $Y_i$ mají Poissonovo rozdělení a linková funkce je logaritmus

$$\ln\lambda_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}.$$

Máme tedy

$$Y_i \sim \mathsf{Po}\left[\exp\left(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}\right)\right], \qquad i = 1, \ldots, n.$$

Mějme dva nezávislé výběry $X_1, X_2, \ldots, X_m$ z $\mathsf{Po}(\lambda)$ a $Y_1, Y_2, \ldots, Y_n$ z $\mathsf{Po}(\mu)$. Nechť $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ je sdružený výběr, který označíme $Z_1, Z_2, \ldots, Z_N$, kde $N = m + n$. Veličiny $Z_1, \ldots, Z_N$ jsou nezávislé a $Z_i \sim \mathsf{Po}(\lambda_i)$, kde $\lambda_i = \lambda$ pro $i = 1, \ldots, m$ a $\lambda_i = \mu$ pro $i = m+1, \ldots, N$. Pak pro vektor $u = (\underbrace{0, , \ldots, 0}_{m}, \underbrace{1, \ldots, 1}_{n})$ máme

$$\ln \lambda_i = \beta_0 + \beta_1 u_i,$$

kde $\beta_0 = \ln \lambda$ a $\beta_1 = \ln \mu - \ln \lambda$. Tedy podmínka $\lambda = \mu$ je ekvivalentní podmínce $\beta_1 = 0$. Chceme-li testovat $H_0 : \lambda = \mu$, můžeme přejít k hypotéze $\beta_1 = 0$, kterou testujeme pomocí statistiky (3.5).

# Kapitola 4

# Alternativní modely

V první kapitole jsme pojednali o testech dobré shody pro Poissonovo rozdělení, nezmínili jsme se však o tom, jak postupovat v případech zamítnutí nulové hypotézy. Popíšeme několik modifikaci základního modelu.

## 4.1 Poissonovo rozdělení s nadhodnocenou nulou

Jelikož je Poissonovo rozdělení určeno jediným parametrem, je v jistém smyslu „striktní“. Často se stává, že heterogenita populace způsobí signifikantní rozdíl mezí střední hodnotou a rozptylem a shoda s Poissonovým rozdělením je narušena. Jedna z příčin může být příliš velký počet nul v náhodném výběru. V takových případech je vhodné uvažovat následující modifikaci Poissonova rozdělení.

Řekneme, že náhodná veličina $X$ má *Poissonovo rozdělení s nadhodnocenou nulou* (Zero-inflated Poisson, ZIP), jestliže s pravděpodobností $p$ je to náhodná veličina s Poissonovým rozdělením a s pravděpodobností $1-p$ musí být $X=0$. Veličina $X$ pak má rozdělení

$$\mathsf{P}\left(X=x\right)=\begin{cases} 1-p+p\mathrm{e}^{-\lambda} & \text{je-li } x=0,\\ p\mathrm{e}^{-\lambda}\frac{\lambda^x}{x!} & \text{je-li } x>0, \end{cases}$$

a platí tedy

$$\mathsf{E}X=p\lambda, \qquad \mathsf{var}X=p\lambda\left[1+\lambda\left(1-p\right)\right].$$

Poměr rozptylu a střední hodnoty je

$$\frac{\mathsf{var}X}{\mathsf{E}X}=1+\lambda(1-p)\geq 1$$

a rovnost nastává, právě když $p = 1$ (Poissonovo rozdělení).

Parametry $\lambda$ a $p$ se obvykle odhadují metodou maximální věrohodnosti. Věrohodnostní funkce náhodného výběru $X_1, \ldots, X_n$, ve kterém se vyskytuje $n_0$ nulových hodnot je

$$f(\mathbf{x}, p, \lambda) = \left(1 - p + p\mathrm{e}^{-\lambda}\right)^{n_0} p^{n-n_0} \mathrm{e}^{-\lambda(n-n_0)} \frac{\lambda^{\sum x_i}}{x_1! x_2! \ldots x_n!},$$

kde jsme pro zjednodušení zápisu využili rovnosti $0! = 1$. Logaritmická věrohodnostní funkce pak má tvar

$$L(p, \lambda) = n_0 \ln\left(1 - p + p\mathrm{e}^{-\lambda}\right) + (n - n_0)(\ln p - \lambda) + \ln\lambda \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \ln x_i!.$$

Dostáváme systém věrohodnostních rovnic

$$\frac{\partial L(p,\lambda)}{\partial p} = -\frac{n_0\left(1 - \mathrm{e}^{-\lambda}\right)}{1 - p + p\mathrm{e}^{-\lambda}} + \frac{n - n_0}{p} = 0, \tag{4.1}$$

$$\frac{\partial L(p,\lambda)}{\partial \lambda} = -\frac{n_0 p \mathrm{e}^{-\lambda}}{1 - p + p\mathrm{e}^{-\lambda}} + (n - n_0) + \frac{\sum_{i=1}^{n} x_i}{\lambda} = 0. \tag{4.2}$$

Nemáme k dispozici explicitní vzorce pro $\hat{\lambda}$ a $\hat{p}$ a výpočty se proto provádějí pomocí knihovny VGAM programu R. Poznamenejme pouze, že z (4.1) dostáváme

$$1 - \mathrm{e}^{-\lambda} = \frac{n - n_0}{np},$$

a proto

$$n\left(1 - \hat{p} + \hat{p}\mathrm{e}^{-\hat{\lambda}}\right) = n_0,$$

z čehož vyplývá, že se teoretická a empirická četnost nulových pozorování shodují.

### 4.1.1 Volba modelu

Při volbě modelu se často řídíme dvěma požadavky: co nejlepší shoda dat s modelem a co nejmenší počet parametrů. Tyto dva principy obvykle jdou proti sobě a abychom mohli posoudit kvalitu modelu, je zapotřebí dát je do souvislosti. Jedna z možností je *AIC* kritérium (*Akaike Information Criterion*), které má tvar

$$AIC = -2L + 2k,$$

kde $L$ je hodnota logaritmické věrohodnostní funkce v bodě maximálně věrohodného odhadu a $k$ je počet odhadovaných parametrů v modelu. Při porovnání modelů se pak za nejlepší považuje ten, který má minimální hodnotu $AIC$.

Ukazuje se však, že v případech výběrů malých rozsahů $AIC$ kritérium může vést na modely s velkým počtem parametrů, tj. penalizace $2k$ není dostatečně silná, aby zabránila zavádění nadbytečných parametrů. Proto se doporučuje modifikovaný $AIC$ (Hurvich a Tsai, 1989)

$$AIC_c = -2L + 2k\left(\frac{n}{n-k-1}\right).$$

**Příklad 1.** Mějme $X_1, X_2, \ldots, X_m$ náhodný výběr z diskrétního rozdělení. Chceme porovnat Poissonovo rozdělení a ZIP model na základě $AIC$ kritéria. Definujme vektor $\mathbf{n}$, ve kterém $i$-tá složka představuje počet výskytů hodnoty $i$ v náhodném výběru. V případě, že náhodný výběr pochází z Poissonova rozdělení, máme $k = 1$ a

$$L_1 = -m\hat{\lambda} + \sum_{x=0}^{\infty} n_x \left(x\ln\hat{\lambda} - \ln x!\right).$$

V případě Poissonova rozdělení s nadhodnocenou nulou máme $k = 2$ a

$$L_2 = n_0 \ln\left(1 - \hat{p} + \hat{p}e^{-\hat{\lambda}}\right) + \sum_{x=1}^{\infty} n_x \ln\left(\hat{p}e^{-\hat{\lambda}}\frac{\hat{\lambda}^x}{x!}\right).$$

Dostaneme-li $AIC_1 = -2L_1 + 1 < -2L_2 + 2 = AIC_2$ (resp. $AIC_{c,1} < AIC_{c,2}$ pro malá $m$), rozhodneme se pro Poissonův model, v opačném případě volíme ZIP model.

**Příklad 2.** Na str. 85 Simonoff píše, že se v Belo Horizonte v Brazílii u 797 dětí zkoumala kazivost zubů. Index kazivosti (decayed, missing and filled teeth index–DMFT index) je založen na počtu zkažených, chybějících a plombovaných zubů a jelikož se u dětí vztahoval k osmi mléčným stoličkám, nabýval hodnoty od 0 do 8. Výsledky jsou uvedeny v tabulce 4.1. Hodnoty $\bar{x} = 3{,}32$ a $s^2 = 6{,}64$ implikují, že Poissonovo rozdělení v tomto případě není vhodný model. Spočítáme-li Pearsonovu statistiku, dostáváme $\chi^2 = 957{,}6$ (7 stupňů volnosti) a $p$-hodnota je prakticky nula. Uvažujme proto Poissonovo rozdělení s nadhodnocenou nulou. Maximálně věrohodné odhady parametrů jsou $\hat{\lambda} = 4{,}17$ a $\hat{p} = 0.8$. Hodnota Pearsonovy statistiky je v tomto případě

Tabulka 4.1: Kazivost zubů

| DMFT index | počet dětí |
|---:|---:|
| 0 | 172 |
| 1 | 73 |
| 2 | 96 |
| 3 | 80 |
| 4 | 95 |
| 5 | 83 |
| 6 | 85 |
| 7 | 65 |
| 8 | 48 |

$\chi^2 = 65{,}7$ (6 stupňů volnosti), což je hodnota stále vysoce signifikantní, ale implikuje, že je shoda dat s modelem mnohem lepší. Navíc hodnota AIC v případě Poissonova rozdělení je 3958,1 a v případě ZIP modelu 3471,7. Odtud vyplývá, že je v této situaci model ZIP jednoznačně lepší.

Poznamenejme ještě, že hodnotu $1 - p = 0{,}2$ můžeme interpretovat jako podíl dětí, které vůbec netrpí kazivostí zubů.

## 4.2 Směs Poissonových rozdělení

Další model, který můžeme uvažovat v případě velké variability dat je následující. Nechť každé pozorování $X_i$ má Poissonovo rozdělení s parametrem $\lambda Y_i$, kde $Y_i$ je nezáporná náhodná veličina splňující $\mathsf{E}Y_i = 1$. Podmíněné rozdělení $X_i$ při daném $Y_i = y$ je

$$\mathsf{P}\left(X_i = x | Y_i = y\right) = \mathrm{e}^{-\lambda y} \frac{(\lambda y)^x}{x!}.$$

Nepodmíněné rozdělení $X_i$ pak už není Poissonovo, ale je dáno vzorcem

$$\mathsf{P}\left(X_i = x\right) = \int_0^\infty \mathrm{e}^{-\lambda y} \frac{(\lambda y)^x}{x!} g\left(y\right) \, \mathrm{d}y,$$

kde $g(y)$ je hustota veličiny $Y_i$. Předpokládejme dále, že $Y_i$ má gama rozdělení s jedním parametrem, tedy

$$g(y) = \frac{\nu^\nu}{\Gamma\left(\nu\right)} y^{\nu-1} \mathrm{e}^{-y\nu}, \quad y > 0, \quad \nu > 0,$$

kde $\Gamma(a) = \int_0^\infty x^{a-1}\mathrm{e}^{-x}\,\mathrm{d}x$ pro $a > 0$.

Nepodmíněné rozdělení $X_i$ pak má tvar

$$\begin{aligned}
\mathsf{P}(X_i = x) &= \int_0^\infty \mathrm{e}^{-\lambda y}\frac{(\lambda y)^x}{x!}\frac{\nu^\nu}{\Gamma(\nu)}y^{\nu-1}\mathrm{e}^{-y\nu}\,\mathrm{d}y \\
&= \frac{\lambda^x}{x!}\frac{\nu^\nu}{\Gamma(\nu)}\int_0^\infty \mathrm{e}^{-(\nu+\lambda)y}y^{x+\nu-1}\,\mathrm{d}y \\
&= \frac{\lambda^x}{x!}\frac{\nu^\nu}{\Gamma(\nu)}\frac{1}{(\lambda+\nu)^{x+\nu}}\int_0^\infty \mathrm{e}^{-t}t^{x+\nu-1}\,\mathrm{d}t \\
&= \frac{\lambda^x}{x!}\frac{\nu^\nu}{(\lambda+\nu)^{x+\nu}}\frac{\Gamma(x+\nu)}{\Gamma(\nu)} \\
&= \frac{\Gamma(x+\nu)}{\Gamma(\nu)\,x!}\left(\frac{\lambda}{\lambda+\nu}\right)^x\left(\frac{\nu}{\lambda+\nu}\right)^\nu.
\end{aligned}$$

To je však negativně binomické rozdělení s parametry $\nu$ a $\nu/(\nu+\mu)$. Toto rozdělení se při $\nu \to \infty$ blíží Poissonovu rozdělení s parametrem $\lambda$

$$\begin{aligned}
\lim_{\nu\to\infty}\mathsf{P}(X_i = x) &= \frac{\lambda^x}{x!}\lim_{\nu\to\infty}\frac{(x+\nu-1)\cdots\nu\Gamma(\nu)}{(\lambda+\nu)^x\,\Gamma(\nu)}\left[\left(1-\frac{\lambda}{\lambda+\nu}\right)^{\frac{\lambda+\nu}{\lambda}}\right]^{\frac{\lambda\nu}{\lambda+\nu}} \\
&= \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}\lim_{\nu\to\infty}\frac{(x+\nu-1)\cdots\nu}{(\lambda+\nu)^x} \\
&= \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!},
\end{aligned}$$

kde jsme využili identity $\Gamma(a+1) = a\Gamma(a)$, $a > 0$ a limity

$$\lim_{n\to\infty}\left[\left(1-\frac{\lambda}{\lambda+\nu}\right)^{\frac{\lambda+\nu}{\lambda}}\right]^{\frac{\lambda\nu}{\lambda+\nu}} = \mathrm{e}^{-\lim_{\nu\to\infty}\frac{\lambda\nu}{\lambda+\nu}} = \mathrm{e}^{-\lambda}.$$

Střední hodnota a rozptyl veličin $X_i$ se určí snadno využitím vlastností podmíněné střední hodnoty

$$\begin{aligned}
\mathsf{E}X_i &= \mathsf{E}\left[\mathsf{E}(X_i|Y_i)\right] \\
&= \mathsf{E}(\lambda Y_i) \\
&= \lambda, \qquad i = 1,\ldots,n.
\end{aligned}$$

Podobně

$$\begin{aligned} \mathsf{E}X_i^2 &= \mathsf{E}\left[\mathsf{E}\left(X_i^2|Y_i\right)\right] \\ &= \mathsf{E}\left(\lambda^2 Y_i^2 + \lambda Y_i\right) \\ &= \lambda^2\left(1+\frac{1}{\nu}\right)+\lambda, \end{aligned}$$

a tedy

$$\mathsf{var}X_i = \lambda(1+\frac{\lambda}{\nu}), \qquad i=1,\ldots,n. \tag{4.3}$$

Jelikož $\nu > 0$, je rozptyl tohoto rozdělení skutečně větší než u Poissonova rozdělení a s rostoucím $\nu$ se blíží $\lambda$.

Maximálně věrohodný odhad parametru $\lambda$ je výběrový průměr $\bar{X}$, zatímco neexistuje explicitní vyjádření maximálně věrohodného odhadu $\nu$ (viz Simonoff 2003, str. 88). Můžeme proto použít momentovou metodu. Vyjdeme z rovnosti (4.3), $\lambda$ odhadneme pomocí $\bar{X}$ a $\mathsf{var}X_i$ pomocí $S^2$. Dostáváme odhad

$$\tilde{\nu} = \frac{\bar{X}^2}{S^2 - \bar{X}}.$$

Mějme náhodný výběr $X_1, X_2, \ldots, X_n$. Chceme-li testovat nulovou hypotézu, že náhodný výběr pochází z Poissonova rozdělení proti alternativě, že pochází z negativně binomického rozdělení, použijeme testovou statistiku

$$Z = \left(\frac{S^2}{\bar{X}} - 1\right)\sqrt{\frac{n-1}{2}}. \tag{4.4}$$

Podmíněné rozdělení veličiny $Q = (n-1)s^2/\bar{X}$ při daném $T = \sum X_i$ je asymptoticky $\chi^2_{n-1}$ (viz Anděl 2005, str. 278). Odtud máme

$$\mathsf{E}\left(\frac{1}{n-1}Q\middle|T\right) \approx 1, \qquad \mathsf{var}\left(\frac{1}{n-1}Q\middle|T\right) \approx \frac{2}{n-1},$$

a tedy $Z$ má asymptoticky $\mathsf{N}(0,1)$ rozdělení. Je-li $Z \geq u_\alpha$, zamítneme hypotézu, že náhodný výběr pochází z Poissonova rozdělení.

**Příklad.** V tabulce 4.2 jsou uvedeny počty utkání anglické 1. ligy v sezóně 1967-1968, ve kterých padlo $k$ branek, $k = 0, 1, \ldots, 8$. Zkoumejme shodu těchto dat s Poissonovým rozdělením. Vypočteme $\bar{x} = 1{,}514$ a $s^2 = 1{,}765$. Při $n = 924$ máme $Z = 3{,}56$ a $p$-hodnotu 0,0002. Spočteme dále Pearsonovu statistiku. Dostáváme $\chi^2 = 18{,}03$ (6 stupňů volnosti) a $p$-hodnotu

Tabulka 4.2: Počet bránek v anglické 1. lize v sezóně 1967 - 1968

| Počet branek | Počet utkání |
|---|---|
| 0 | 225 |
| 1 | 293 |
| 2 | 224 |
| 3 | 114 |
| 4 | 41 |
| 5 | 15 |
| 6 | 9 |
| 7 | 1 |
| 8 | 2 |

0,006. Obě statistiky prokazují špatnou shodu dat s Poissonovým rozdělením. Uvažujme proto směs Poissonových rozdělení (negativně binomické rozdělení). Pomocí knihovny MASS vypočteme maximálně věrohodné odhady parametrů $\hat{\lambda} = 1{,}514$ a $\hat{\nu} = 9{,}626$. Pearsonova statistika je v tomto případě $\chi^2 = 3,8$ (5 stupňů volnosti) a příslušná $p$-hodnota je 0,58.

Vhodnost tohoto modelu je v tomto případě průhledná i z jiného hlediska. Pravděpodobnost, že útok skončí úspěchem, závisí na tom, kteří z hráčů útočí. Navíc tato pravděpodobnost se u týmu mění v čase v závislosti na průběhu sezóny nebo dokonce utkání.

## 4.3 Useknuté Poissonovo rozdělení

Méně často se stává, že je shoda s Poissonovým rozdělením narušena z důvodu příliš malé variability dat, což můžeme nahlédnout pomocí statistiky (4.4). Jedna z příčin může být to, že veličiny tvořící náhodný výběr nabývají pouze kladných hodnot. Navrhneme proto modifikaci Poissonova rozdělení, která má tuto vlastnost.

Nechť $X \sim \mathsf{Po}(\lambda)$. Řekneme, že náhodná veličina $Y$ má *useknuté Poissonovo rozdělení* (Zero-truncated Poisson), platí-li

$$\mathsf{P}(Y = k) = \mathsf{P}(X = k | X > 0) = \frac{\mathsf{P}(X = k)}{1 - \mathsf{P}(X = 0)}, \qquad k = 1, 2, \ldots$$

Potom platí

$$\mathsf{E}Y = \frac{\lambda}{1-\mathrm{e}^{-\lambda}}, \qquad \mathsf{var}\, Y = \left(\frac{\lambda}{1-\mathrm{e}^{-\lambda}}\right)\left(1-\frac{\lambda \mathrm{e}^{-\lambda}}{1-\mathrm{e}^{-\lambda}}\right)$$

a máme $\mathsf{E}Y > \mathsf{var}\, Y$.

Předpokládejme, že $Y_1, \ldots, Y_n$ jsou nezávislé stejně rozdělené náhodné veličiny a mají useknuté Poissonovo rozdělení. Pak pro $\mathbf{Y}=\mathbf{y}$ věrohodnostní funkce $p(\mathbf{y}, \lambda)$ je

$$p\left(\mathbf{y}, \lambda\right) = \frac{\mathrm{e}^{-n\lambda}\lambda^{\sum y_i}}{y_1! y_2! \cdots y_n! \left(1-\mathrm{e}^{-\lambda}\right)^n}$$

a logaritmická věrohodnostní funkce je

$$L\left(\mathbf{y}, \lambda\right) = -n\lambda + \ln\lambda \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \ln y_i! - n\ln\left(1-\mathrm{e}^{-\lambda}\right).$$

Položíme

$$L'\left(\mathbf{y}, \lambda\right) = -n + \frac{\sum_{i=1}^{n} y_i}{\lambda} - \frac{n\mathrm{e}^{-\lambda}}{1-\mathrm{e}^{-\lambda}} = 0$$

a po úpravě dostáváme

$$\lambda = \left(1-\mathrm{e}^{-\lambda}\right)\bar{y}. \tag{4.5}$$

V tomto případě jsme nenašli explicitní vzorec pro maximálně věrohodný odhad parametru $\lambda$, ale snadno se přesvědčíme, že funkce $g(\lambda) = (1-\mathrm{e}^{-\lambda})\bar{y} - \lambda$ má na intervalu $(0, \infty)$ právě jeden kořen, a tedy je to maximálně věrohodný odhad. Zapíšeme-li rovnici 4.5 ve tvaru

$$\bar{y} = \frac{\lambda}{1-\mathrm{e}^{\lambda}},$$

zjistíme, že v tomto případě momentová metoda a metoda maximální věrohodnosti vedou na tutéž rovnici.

**Příklad.** V tabulce 4.3 je uveden počet obyvatel v jednotlivých domech (Simonoff 2003, str. 121). Simonoff dále píše, že domy bez obyvatel nebyly do studie zahrnuty. Můžeme tedy uvažovat useknuté Poissonovo rozdělení jako model pro tato data. Maximálně věrohodný odhad je $\hat{\lambda} = 0{,}577$. Příslušná Pearsonova statistika je $\chi^2 = 1{,}332$ (1 stupeň volnosti) a $p$-hodnota je 0,25. Model useknutého Poissonova rozdělení tedy nezamítneme. Uveďme pro úplnost hodnoty výběrového průměru a rozptylu

$$\bar{X} = 1{,}32, \qquad S^2 = 0{,}37.$$

Tabulka 4.3: počet obyvatel

| počet obyvatel | počet domů |
|---:|---:|
| 1 | 436 |
| 2 | 133 |
| 3 | 19 |
| 4 | 2 |
| 5 | 1 |
| 6 | 0 |
| 7 | 1 |

## 4.4 Poissonovo rozdělení s modifikovanou nulou

V některých situacích se nuly vyskytují v náhodném výběru, ale jejích počet je natolik malý, že nemůžeme přijmout model Poissonova rozdělení. Potřebujeme tedy model, ve kterém je pravděpodobnost, že je pozorování rovno nule, menší než u Poissonova modelu. V takových případech můžeme modifikovat ZIP model tak, že připouštíme, aby $p > 1$. Pak máme

$$\mathsf{P}(X = 0) = 1 - p + p\mathrm{e}^{-\lambda}, \qquad \text{pro} \quad 1 \leq p \leq \frac{1}{1 - \mathrm{e}^{-\lambda}}.$$

Přímým výpočtem se snadno ověří, že $\mathsf{P}(X = 0) \leq \mathrm{e}^{-\lambda}$.

Parametry $p$ a $\lambda$ odhadneme následovně. Uvažujeme-li pouze kladná pozorování, parametr $\lambda$ bude parametr useknutého Poissonova rozdělení a odhadneme ho pomocí (4.5). Pak z (4.1) dostáváme maximálně věrohodný odhad parametru $p$

$$\hat{p} = \frac{1 - n_0/n}{1 - \mathrm{e}^{-\hat{\lambda}}}. \tag{4.6}$$

**Příklad.** Na str. 121 Simonoff (2003) uvádí, že v Los Angeles u 457 bezdomovců bylo zjišťováno, kolikrát v předchozím týdnu navštívili sociální jídelny nebo distribuční centra potravinových balíčků. Výsledky jsou uvedeny v tabulce 4.4.

Abychom zjistili, jestli je rozdíl mezi výběrovým průměrem $\bar{x} = 2{,}14$ a veličinou $s^2 = 1{,}54$ signifikantní, spočteme statistiku (4.4). Při $n = 457$ dostáváme $Z = -4{,}19$ a $p$-hodnotu $1{,}4 \cdot 10^{-5}$. Nemůžeme tedy přijmout Poissonovo rozdělení jako model pro tato data. Uvažujme proto Poissonovo

Tabulka 4.4: Sociální jídelny

| počet návštěv | počet osob |
|---:|---:|
| 0 | 27 |
| 1 | 137 |
| 2 | 128 |
| 3 | 96 |
| 4 | 50 |
| 5 nebo 6 | 19 |

rozdělení s modifikovanou nulou. Máme maximálně věrohodné odhady $\hat{\lambda} = 1{,}95$ a $\hat{p} = 1{,}1$. Hodnota Pearsonovy statistiky 3,47 (3 stupně volnosti) a příslušná $p$-hodnota 0,32 svědčí ve prospěch našeho modelu.

Simonoff naznačuje, že důvodem k menšímu počtu nul ve výběru může být to, že se respondenti vybírali na místech, která bezdomovci obvykle navštěvují, mimo jiné na parkovištích, v denních centrech, v distribučních centrech pro potravinové balíčky, v sociálních jídelnách atd.

# Příloha A

# Poznámky k výpočtům

V příkladech první kapitoly porovnáváme hodnoty Pearsonových statistik v případě použití odhadu modifikovanou metodou minimálního $\chi^2$ a výběrového průměru. První odhad dostáváme pomocí funkce `chiodhad(x)`, která za parametr má `data.frame` (tabulka pozorovaných četností). Před provedením testu jsou sloučeny třídy s malými pozorovanými četnostmi. Hodnotu Pearsonovy statistiky pak dostáváme pomocí funkce `chitest(x,l)`, kde druhý parametr představuje odhad parametru $\lambda$.

Funkce pro výpočet skórového a exaktního intervalu spolehlivosti jsou převzaté z

`http://www.stern.nyu.edu/~jsimonof/AnalCatData`.

Funkce `scoreintpois <- function (x, alpha)` za parametry má vektor pozorování a požadovaný koeficient spolehlivosti, kdežto funkce `exactpoisci <- function(x, t, alpha)` za první parametr má součet složek vektoru pozorování, za druhý rozsah výběru a za třetí koeficient spolehlivosti.

Maximálně věrohodný odhad parametru useknutého Poissonova rozdělení je spočítán pomocí funkce `truncpoismle <- function(x)`. Funkce pro odhad parametrů Poissonova rozdělení s nadhodnocenou nulou je `zipmle <- function(x)`. V případě Poissonova rozdělení s modifikovanou nulou nejprve uvažujeme nenulová pozorování a parametr $\lambda$ odhadneme jako parametr useknutého Poissonova rozdělení. Maximálně věrohodný odhad parametru $p$ pak z (4.6) snadno dopočítáme.

V případě zobecněných lineárních modelů se výpočty často provádějí pomocí knihovny MASS. Tam se najde funkce `glm.nb`, která určí maximálně věrohodné odhady parametrů negativně binomického rozdělení.

Shodu dat s alternativními modely testujeme pomocí Pearsonovy statistiky, která je speciálním případem mocninné divergence (viz Simonoff, 2003, str. 77). Na její výpočet máme k dispozici funkci `powdiv <- function(obs, exp, npar, lambda)`, kde jako parametr kromě pozorovaných a empirických četností máme počet odhadovaných parametrů v modelu. Zvolíme-li `lambda=1`, dostáváme Pearsonovu statistiku.

# Literatura


[1] Anděl J.: *Základy matematické statistiky*, Matfyzpress, Praha, 2007.

[2] Cambell D. B., Oprian C. A.: *On the Kolmogorov-Smirnov test for the Poisson distribution with unknown mean*, Biom. J. **21** (1979) 17-24.

[3] Chernoff H., Lehmann E. L.: *The use of maximum likelihood estimates in $\chi^2$ test for goodness of fit*, Ann. Math. Statist. **25** (1954) 579–586.

[4] Cramér H.: *Mathematical Methods of Statistics*, Princeton University Press, Princeton, 1946.

[5] Hátle J., Likeš J.: *Základy počtu pravděpodobnosti a matematické statistiky*, SNTL, Praha, 1972.

[6] Simonoff J. S.: *Analyzing Categorical Data*, Springer, New York, 2003.