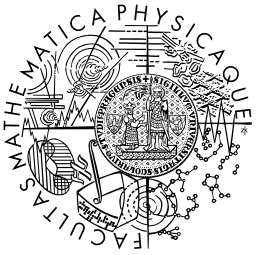


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Julie Váňová

Testy dobré shody s Poissonovým rozdělením založené na nulovém indexu

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Šárka Hudecová, Ph.D.

Studijní program: Obecná matematika

Praha 2023

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

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

Na tomto místě bych chtěla poděkovat vedoucí své práce RNDr. Šárce Hudecové, Ph.D. za velkou ochotu, vstřícnost a také čas, který mi věnovala, ať už během konzultací, nebo při pečlivém pročítání jednotlivých verzí této práce.

Název práce: Testy dobré shody s Poissonovým rozdělením založené na nulovém indexu

Autor: Julie Váňová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Šárka Hudecová, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato bakalářská práce se zabývá metodou testování dobré shody s Poissonovým rozdělením, která je založená na takzvaném nulovém indexu. Nejprve je zaveden pojem Poissonova nulového indexu a jsou diskutovány jeho základní vlastnosti. Následně se práce soustředí na odvození asymptotického rozdělení nulových indexů a jeho využití ke konstrukci testů dobré shody, přičemž jsou uvedeny i konkrétní příklady nulových indexů a následných testů. V další části jsou potom stručně popsány některé další způsoby, kterými lze testovat dobrou shodu s Poissonovým rozdělením, konkrétně jde o χ^2-testy dobré shody a testy založené na indexu disperze. Zmíněné metody testování jsou poté porovnány v simulační studii.

Klíčová slova: Poissonovo rozdělení, testy dobré shody, asymptotické testy, nulový index

Title: Goodness-of-fit tests for Poisson distribution based on zero index

Author: Julie Váňová

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Šárka Hudecová, Ph.D., Department of Probability and Mathematical Statistics

Abstract: This bachelor thesis deals with goodness-of-fit tests for Poisson distribution that are based on so-called zero index. In the first part, Poisson zero index is defined and some of its basic properties are discussed. Further, asymptotic distribution of zero indexes is derived and it is used to construct asymptotic goodness-of-fit tests. Particular examples of zero indexes and related tests are included. In the following part, other types of goodness-of-fit tests for Poisson distribution are briefly described, in particular χ^2-tests and tests based on index of dispersion. All mentioned methods are then compared in a simulation study.

Keywords: Poisson distribution, goodness-of-fit tests, asymptotic tests, zero index

Obsah

Úvod

Testy dobré shody s nějakým předem určeným rozdělením patří mezi nejznámější statistické testy. V této práci se budeme zabývat testováním dobré shody s Poissonovým rozdělením, tedy rozdělením, které nachází mnoho využití i v praxi (např. v modelování počtu pojistných událostí). Zaměříme se při tom především na metodu, která je založena na tzv. nulovém indexu. Její hlavní myšlenka spočívá v ověřování, zda platí určitý vztah mezi relativním počtem nul v realizacích nějakého náhodného výběru a jeho výběrovým průměrem.

Idea tohoto způsobu testování je popsána v článku Weiß, Homburg a Puig (2019), který je hlavním výchozím bodem této práce. V souladu s tímto článkem také definujeme již zmíněný nulový index. Výše uvedený článek ovšem zavádí tento index v trochu odlišném kontextu (INAR(1) modely časových řad). My se zaměříme na situaci, kdy máme k dispozici náhodný výběr z nějakého diskrétního rozdělení, a zajímá nás, zda se jedná o rozdělení Poissonovo.

V první kapitole formalizujeme zkoumaný problém jako testování určité nulové hypotézy a uvedeme základní vlastnosti Poissonova rozdělení. Ve druhé se potom budeme podrobně zabývat testy dobré shody založenými na nulových indexech. Odvodíme si asymptotické rozdělení těchto indexů (za platnosti hypotézy) a ukážeme, jak lze pomocí nich testovat dobrou shodu s Poissonovým rozdělením. To vše provedeme nejdříve obecně, a následně i pro konkrétní případy. Ve třetí kapitole stručněji popíšeme i další typy testů dobré shody. Jednotlivé způsoby testování poté porovnáme v simulační studii – podíváme se, jak dobře dané testy dodržují stanovenou hladinu významnosti a „spočteme“ jejich sílu proti konkrétní alternativě (negativně binomickému rozdělení).

Vlastní přínos této práce spočívá především v uvádění vysvětlujících komentářů, formulování vlastních vět a tvrzení v kapitole týkající se nulových indexů (kapitola 2) a podrobném sepsání jejich důkazů. Dále potom v odvození asymptotického rozdělení konkrétních nulových indexů, z nichž dva ($\hat{\mathrm{I}}_{3,n}$ a $\hat{\mathrm{I}}_{4,n}$) jsme si sami zkonstruovali. Částečně také v zamyšlení se nad různými úskalími testování založeného na nulových indexech (sekce 2.4). Za vlastní příspěvek lze rovněž považovat simulační studii ve čtvrté kapitole.

1. Uvedení do problému

Počet dopravních nehod v Praze za jeden týden, množství gólů během jednoho fotbalového zápasu, počet nových případů nákazy danou nemocí za den – to vše jsou příklady dat, na něž můžeme nahlížet jako na realizace určitých náhodných veličin s diskrétním rozdělením. Přestože jejich konkrétní rozdělení neznáme, máme k dispozici metody, které nám umožní udělat si o něm lepší představu.

1.1 Testy dobré shody

Jednou z možností je testovat, do jaké míry naše data odpovídají hypotetickým realizacím nějakého předem stanoveného rozdělení. K tomu slouží tzv. testy dobré shody.

Uvedený problém nyní trochu zúžíme. Budeme předpokládat, že data, která máme k dispozici, představují náhodný výběr z nějakého diskrétního rozdělení. To znamená, že události, které zkoumáme, nastávají nezávisle na sobě a pravděpodobnost výskytu je pro všechny z nich stejná. Zde se navíc nebudeme zabývat obecnými testy dobré shody, ale pouze jejich verzí pro Poissonovo rozdělení.

Matematicky formulováno, budeme uvažovat náhodný výběr $X_1, X_2, ..., X_n$ pocházející z určitého diskrétního rozdělení. V naší situaci dává navíc smysl omezit se pouze na rozdělení, jejichž nosič je podmnožina nezáporných celých čísel (jak později uvidíme, takový nosič má i Poissonovo rozdělení). Budeme tedy pracovat s neparametrickým modelem

$$\mathcal{F} = \{\text{všechna diskrétní rozdělení, jejichž nosič je podmnožina } \mathbb{N}_0\}.$$

Předmětem testování bude celé pravděpodobnostní rozdělení našeho náhodného výběru. Zajímá nás, jak moc se toto rozdělení shoduje s Poissonovým rozdělením s parametrem λ, pro nějaké blíže nespecifikované (pevné) $\lambda > 0$. Budeme tedy testovat hypotézu H_0 proti alternativě H_1, kde

$$\begin{aligned} \mathsf{H}_0 &: \ \exists\, \lambda \in (0{,}\infty) \text{ takové, že } X_1 \sim \mathsf{Po}(\lambda), \\ \mathsf{H}_1 &: \ \nexists\, \lambda \in (0{,}\infty) \text{ takové, že } X_1 \sim \mathsf{Po}(\lambda). \end{aligned} \tag{1.1}$$

Místo $X_1 \sim \mathsf{Po}(\lambda)$ lze ekvivalentně psát $X_i \sim \mathsf{Po}(\lambda)$, $i = 1, ..., n$, neboť uvažované náhodné veličiny jsou dle předpokladu stejně rozdělené.

Nulová hypotéza říká, že pro nějaké λ má náhodný výběr $X_1, ..., X_n$ požadované rozdělení. Konkrétní hodnota parametru λ je ovšem neznámá a její určení ani není předmětem tohoto testu. Naopak podle alternativy žádné λ, pro které by platilo $X_1 \sim \mathsf{Po}(\lambda)$, neexistuje. Zamítnutím nulové hypotézy ve prospěch alternativy potom pouze tvrdíme, že daný náhodný výběr z Poissonova rozdělení nepochází, o jeho skutečném rozdělení toho však stále mnoho nevíme.

1.2 Poissonovo rozdělení

Jak již bylo zmíněno, budeme se zabývat především testováním dobré shody s Poissonovým rozdělením. V této sekci si připomeneme definici tohoto pravděpodobnostního rozdělení a podíváme se na některé jeho vlastnosti.

Definice 1 (Poissonovo rozdělení)**.** *Řekneme, že náhodná veličina X má* **Poissonovo rozdělení** *s parametrem $\lambda > 0$, jestliže platí*

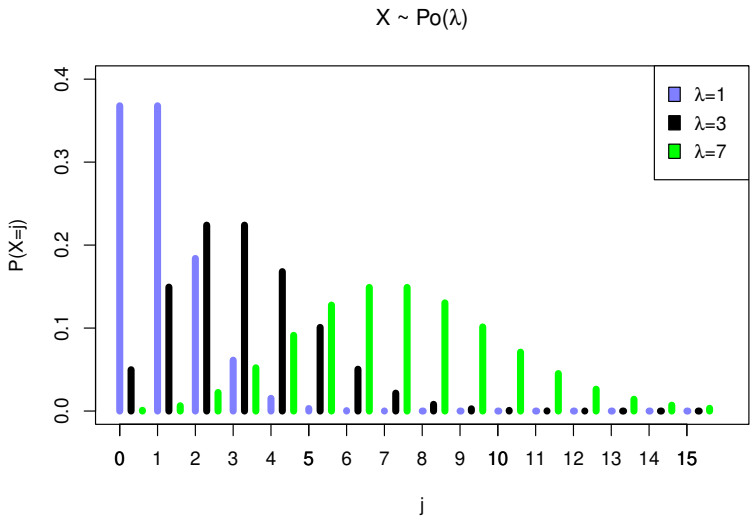
$$\mathsf{P}(X = j) = \mathrm{e}^{-\lambda} \cdot \frac{\lambda^j}{j!}, \quad j \in \mathbb{N}_0.$$

Značíme $X \sim \mathsf{Po}(\lambda)$.

Můžeme si všimnout, že $\mathsf{P}(X = j) > 0$ pro všechna $j \in \mathbb{N}_0$ a

$$\sum_{j=0}^{\infty} \mathrm{e}^{-\lambda} \cdot \frac{\lambda^j}{j!} = \mathrm{e}^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \mathrm{e}^{-\lambda} \cdot \mathrm{e}^{\lambda} = 1,$$

tedy se skutečně jedná o pravděpodobnostní rozdělení. Z definice je dále vidět, že náhodná veličina se zmíněným rozdělením má kladnou pravděpodobnost výskytu u nekonečně mnoha různých hodnot. Poissonovo rozdělení je diskrétní rozdělení, neboť jeho nosič $\mathbb{N}_0$ je spočetná množina. Obrázek 1.1 zachycuje toto rozdělení pro vybrané hodnoty parametru λ.



Obrázek 1.1: Poissonovo rozdělení pro vybrané hodnoty parametru λ

Při práci s daným rozdělením se rovněž vyplatí znát některé jeho momenty, především střední hodnotu a rozptyl. Pro náhodnou veličinu X $\sim \mathsf{Po}(\lambda)$, $\lambda > 0$, spočteme její střední hodnotu snadno:

$$\mathsf{E}X = \sum_{k=0}^{\infty} k \cdot \mathsf{P}(X = k) = \sum_{k=0}^{\infty} k \cdot \mathrm{e}^{-\lambda} \cdot \frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} \lambda \mathrm{e}^{-\lambda} \cdot \frac{\lambda^{k-1}}{(k-1)!} = \lambda \sum_{j=0}^{\infty} \mathrm{e}^{-\lambda} \cdot \frac{\lambda^j}{j!} = \lambda,$$

kde v poslední rovnosti využíváme, že Poissonovo rozdělení je pravděpodobnostní rozdělení, tedy je daná suma rovna jedné. Dále platí

$$\begin{aligned} \mathsf{E}X^2 &= \sum_{k=0}^{\infty} k^2 \cdot \mathsf{P}(X = k) = \sum_{k=0}^{\infty} k^2 \cdot \mathrm{e}^{-\lambda} \cdot \frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} \lambda k \cdot \mathrm{e}^{-\lambda} \cdot \frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda \sum_{j=0}^{\infty} (j+1) \cdot \mathrm{e}^{-\lambda} \cdot \frac{\lambda^j}{j!} = \lambda \left(\sum_{j=0}^{\infty} j \cdot \mathrm{e}^{-\lambda} \cdot \frac{\lambda^j}{j!} + \sum_{j=0}^{\infty} \mathrm{e}^{-\lambda} \cdot \frac{\lambda^j}{j!} \right) = \lambda(\lambda + 1). \end{aligned}$$

Potom

$$\mathsf{var}X = \mathsf{E}X^2 - (\mathsf{E}X)^2 = \lambda(\lambda+1) - \lambda^2 = \lambda.$$

Vidíme tedy, že střední hodnota a rozptyl náhodné veličiny s Poissonovým rozdělením si jsou rovny. Později uvidíme, že tuto speciální vlastnost lze použít i ke konstrukci testů dobré shody založených na tzv. indexu disperze.

Poissonovým rozdělením se řídí zejména veličiny, které udávají počty nějakých událostí (např. počty nehod, pojistných událostí, ...) za určitou časovou jednotku.

1.3 Některé pomocné věty a tvrzení

V této části uvedeme znění několika důležitých vět a tvrzení, na něž se budeme odkazovat v nadcházejících kapitolách.

Tvrzení 1. *Nechť X_1, X_2, X_3, ..., X_n je náhodný výběr z libovolného rozdělení, které má konečný první moment.*
(i) *Výběrový průměr $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ splňuje*

$$\mathsf{E}(\overline{X}_n) = \mathsf{E}X_1, \quad \overline{X}_n \xrightarrow[n\to\infty]{\mathsf{P}} \mathsf{E}X_1.$$

(ii) *Předpokládejme navíc, že $\mathsf{var}X_1 \in (0, \infty)$. Výběrový rozptyl*

$$S_n^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$$

splňuje

$$\mathsf{E}(S_n^2) = \mathsf{var}X_1, \quad S_n^2 \xrightarrow[n\to\infty]{\mathsf{P}} \mathsf{var}X_1.$$

(iii) *Nechť $A \subset \mathbb{R}$ je borelovská množina. Označme $Y_i := \mathbf{1}_{[X_i \in A]}$, $i = 1, ..., n$, tj.*

$$Y_i = \begin{cases} 1, & \text{pokud } X_i \in A, \\ 0, & \text{pokud } X_i \notin A. \end{cases}$$

Potom Y_i má alternativní rozdělení s pravděpodobností úspěchu $p := \mathsf{P}(X_1 \in A)$. Dále pro relativní četnost jevu $[X_i \in A]$ definovanou jako $\overline{Y}_n = \frac{1}{n}\sum_{j=1}^n \mathbf{1}_{[X_j \in A]}$ platí

$$\mathsf{E}(\overline{Y}_n) = p, \quad \overline{Y}_n \xrightarrow[n\to\infty]{\mathsf{P}} p.$$

Důkaz: Části (i) a (ii) jsou dokázány v knize Dupač a Hušková (2013, Věta 5.1) s tím, že ta je oproti našemu tvrzení formulovaná pro konvergenci skoro jistě. Kýžený výsledek tak plyne například z (i) ve větě 2.7 v knize van der Vaart (1998).

(iii) je důsledkem části (i), neboť z definice střední hodnoty

$$\mathsf{E}(Y_1) = \mathsf{E}(\mathbf{1}_{[X_i \in A]}) = \mathsf{P}(X_1 \in A) = p.$$

□

Nadcházející věta je převzatá z Anděl (2007, Věta B.5). Její důkaz lze najít například v knize Anděl (1978).

Věta 2 (Mnohorozměrná Lindebergova centrální limitní věta)**.**

Nechť $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3, \ldots$ *jsou* k*-rozměrné nezávislé a stejně rozdělené náhodné vektory, které mají rozdělení se střední hodnotou* $\boldsymbol{\mu}$ *a konečnou varianční maticí* $\boldsymbol{\Sigma}$*. Potom pro* $n \to \infty$ *platí*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\boldsymbol{X}_i - \boldsymbol{\mu}) \xrightarrow{\mathsf{D}} \mathsf{N}_k(\mathbf{0},\, \boldsymbol{\Sigma}).$$

Pro potřeby následující věty připomeneme, že pro $\mathbf{x} \in \mathbb{R}^k$ značí $\mathbb{D}g(\mathbf{x})$ Jacobiho matici, tj.

$$\mathbb{D}g(\mathbf{x}) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_p}{\partial x_1} & \cdots & \frac{\partial g_p}{\partial x_k} \end{pmatrix}.$$

Věta 3 (Δ-věta)**.**

Nechť $\{\boldsymbol{T}_n\}_{n=1}^{\infty}$ *jsou náhodné vektory splňující* $\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\mu}) \xrightarrow[n \to \infty]{\mathsf{D}} \mathsf{N}_k(\mathbf{0},\, \boldsymbol{\Sigma})$ *pro nějaký vektor konstant* $\boldsymbol{\mu} \in \mathbb{R}^k$ *a pozitivně semi-definitní matici* $\boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$*. Nechť dále zobrazení* $g : \mathbb{R}^k \to \mathbb{R}^p$ *je spojitě diferencovatelné na nějakém okolí bodu* $\boldsymbol{\mu}$*. Potom*

$$\sqrt{n}(g(\boldsymbol{T}_n) - g(\boldsymbol{\mu})) \xrightarrow{\mathsf{D}} \mathsf{N}_p\Big(\mathbf{0},\ \mathbb{D}g(\boldsymbol{\mu})\boldsymbol{\Sigma}(\mathbb{D}g(\boldsymbol{\mu}))^{\mathsf{T}}\Big) \text{ pro } n \to \infty.$$

Důkaz: Viz van der Vaart (1998, Věta 3.1).

□

2. Testy založené na nulovém indexu

V předchozí kapitole jsme představili obecný koncept testů dobré shody. Nyní se již podíváme na konkrétní případ těchto testů, který je založen na tzv. nulovém indexu. Při definování nulového indexu budeme vycházet z článku Weiß a kol. (2019), kde je ovšem tento index zaveden v o něco obecnějším kontextu. Velká část vět, tvrzení a komentářů, které budou následovat, se tak v článku nevyskytuje.

2.1 Nulový index – definice a vlastnosti

V sekci 1.2 jsme odvodili, že pokud $X_1 \sim \mathsf{Po}(\lambda)$, pro nějaké $\lambda > 0$, splňuje střední hodnota rovnost $\mathsf{E}X_1 = \lambda$. Z definice Poissonova rozdělení potom vyplývá, že pro náhodnou veličinu X s tímto rozdělením platí

$$\mathsf{P}(X_1 = 0) = \mathrm{e}^{-\lambda} = \mathrm{e}^{-\mathsf{E}X_1}.$$

Jedna z možností, jak testovat, zda rozdělení náhodného výběru odpovídá Poissonovu rozdělení, spočívá právě v ověřování platnosti tohoto vztahu.

Skutečná hodnota parametru λ i pravděpodobnosti na levé straně rovnosti je pro nás neznámá, budeme tedy potřebovat výrazy na obou stranách nějak odhadnout.

Označme $p_0 := \mathsf{P}(X_1 = 0)$. Jako odhad pravděpodobnosti se nabízí použít relativní četnost

$$\hat{p}_0 := \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{[X_j = 0]}$$

jevu $[X_1 = 0]$, což je podle (ii) v tvrzení 1 nestranný a konzistentní odhad p_0.

Poznámka. Konzistencí odhadu $\hat{p}_0$ rozumíme platnost vztahu $\hat{p}_0 \xrightarrow[n \to \infty]{\mathsf{P}} p_0$. Někdy se v definici konzistence místo konvergence v pravděpodobnosti uvažuje konvergence skoro jistě, v této práci však budeme vždy chápat konzistenci odhadu v souvislosti s prvním zmíněným typem konvergence.

Vhodným odhadem střední hodnoty je výběrový průměr

$$\overline{X}_n := \frac{1}{n} \sum_{j=1}^{n} X_j,$$

neboť na základě tvrzení 1 opět disponuje žádanými vlastnostmi nestrannosti a konzistence. Zároveň se jedná i o nejlepší nestranný odhad parametru λ (viz Anděl, 2007, Příklad 7.62), tedy odhad s nejmenším rozptylem mezi všemi nestrannými odhady λ.

Budeme-li opět uvažovat vztah $p_0 = \mathrm{e}^{-\mathsf{E}X_1}$, můžeme levou stranu rovnosti odhadnout pomocí $\hat{p}_0$ a pravou stranu jako $\mathrm{e}^{-\overline{X}_n}$, což je podle věty o spojité transformaci (viz van der Vaart (1998), Věta 2.3) konzistentní odhad $\mathrm{e}^{-\mathsf{E}X_1}$. Obecně bude naše testová statistika nějaká funkce $f(\overline{X}_n, \hat{p}_0)$, u níž chceme, aby nějakým způsobem reflektovala, jak moc se hodnoty p_0 a $\mathrm{e}^{-\mathsf{E}X_1}$ liší. To vede k zavedení následujícího pojmu, jenž bude jeden z klíčových pojmů této práce.

Definice 2 (Poissonův nulový index)**.** *Nechť* $(0,\,\infty) \times (0,\,1) \subseteq G$ *a* $f : G \to \mathbb{R}$ *je funkce, jejíž předpis nezávisí na neznámých parametrech, splňující*

(i) $f(x,\,\mathrm{e}^{-x}) = 0$ *pro* $\forall\ x > 0$*;*

(ii) $f(x,\,y) \neq 0$*, pokud* $y \neq \mathrm{e}^{-x}$*.*

Nechť $\overline{X}_n$ *a* $\hat{p}_0$ *jsou definované jako výše. Potom statistika*

$$f(\overline{X}_n,\,\hat{p}_0),$$

je-li definovaná, se nazývá ***Poissonův nulový index****.*

Věta „je-li definovaná“ se v předchozí definici vyskytuje kvůli tomu, že bod $(\overline{X}_n,\,\hat{p}_0)^\mathsf{T}$ může v některých případech ležet mimo definiční obor funkce f. Jak ale uvidíme v sekci 2.2, pokud je nosičem našeho náhodného výběru podmnožina $\mathbb{N}_0$ a platí $\mathsf{P}(X_1 = 0) \in (0,\,1)$, pravděpodobnost, že je $f(\overline{X}_n,\,\hat{p}_0)$ dobře definovaná, se pro $n \to \infty$ blíží k 1.

Pro funkci f splňující (i) a (ii) v definici 2 dostaneme, že pro libovolné $\lambda > 0$:

$$f(\lambda,\,p_0) = 0 \quad \Leftrightarrow \quad p_0 = \mathrm{e}^{-\lambda}.$$

Dále je dobré si uvědomit, že Poissonův nulový index je skutečně statistika, neboť $\overline{X}_n$ ani $\hat{p}_0$ nezávisí na neznámých parametrech. Některé jeho vlastnosti nyní zformulujeme v následujícím tvrzení.

Tvrzení 4. *Mějme náhodný výběr* $X_1, X_2, ..., X_n$ *z libovolného diskrétního rozdělení, jehož nosič je podmnožina* $\mathbb{N}_0$*, s konečnou střední hodnotou. Označme* $p_0 := \mathsf{P}(X_1 = 0)$*. Nechť* f *je funkce splňující* (i) *a* (ii) *v definici 2, která je spojitá v bodě* $(\mathsf{E}X_1,\,p_0)^\mathsf{T} \in \mathbb{R}^2$ *.*

1. *Platí*
$$f(\overline{X}_n,\,\hat{p}_0) \xrightarrow[n\,\to\,\infty]{\mathsf{P}} f(\mathsf{E}X_1,\,p_0),$$
tj. $f(\overline{X}_n,\,\hat{p}_0)$ *je konzistentní odhad* $f(\mathsf{E}X_1,\,p_0)$*.*

2. *Nechť navíc* $\mathsf{E}X_1 > 0$ *a* $p_0 = \mathrm{e}^{-\mathsf{E}X_1}$*. Potom*
$$f(\overline{X}_n,\,\hat{p}_0) \xrightarrow[n\,\to\,\infty]{\mathsf{P}} 0.$$
Speciálně, výše uvedený vztah platí pro $X_1 \sim \mathsf{Po}(\lambda)$*,* $\lambda > 0$*.*

3. *Nechť* $p_0 \neq \mathrm{e}^{-\mathsf{E}X_1}$*. Potom pro* $n \to \infty$
$$f(\overline{X}_n,\,\hat{p}_0) \xrightarrow{\mathsf{P}\ \not}\ 0.$$

Důkaz:

1. $\overline{X}_n$ a $\hat{p}_0$ jsou konzistentní odhady $\mathsf{E}X_1$ a p_0. První tvrzení tak plyne z věty o spojité transformaci, kterou lze použít díky předpokladu spojitosti f v bodě $(\mathsf{E}X_1,\,p_0)^\mathsf{T}$.

2. Jedná se o důsledek předchozího bodu. Označíme-li $\lambda := \mathsf{E}X_1$, dostaneme $p_0 = \mathrm{e}^{-\lambda}$. Tudíž podle (i) v definici nulového indexu
$$f(\mathsf{E}X_1,\,p_0) = f(\lambda,\,\mathrm{e}^{-\lambda}) \overset{\text{(i)}}{=} 0.$$
Dále pro Poissonovo rozdělení s parametrem λ platí $p_0 = \mathrm{e}^{-\lambda}$ a $\mathsf{E}X_1 = \lambda$, a tedy v tomto případě je rovnost $p_0 = \mathrm{e}^{-\mathsf{E}X_1}$ zřejmě splněna.

3. Opět plyne z 1. Z části (ii) definice 2 totiž za této situace

$$f(\mathsf{E}X_1,\, p_0) \neq 0.$$

Z jednoznačnosti limity tak dostaneme požadovaný výsledek.

□

Za platnosti (1.1) a pro dostatečně velký rozsah náhodného výběru bude tedy podle 2. části předchozího tvrzení hodnota statistiky $f(\overline{X}_n,\, \hat{p}_0)$ blízká nule. Z posledního bodu naopak vidíme, že pro náhodné výběry, které splňují

$$\mathsf{P}(X_1 = 0) \neq \mathrm{e}^{-\mathsf{E}X_1},$$

a tedy nepocházejí z Poissonova rozdělení, $f(\overline{X}_n,\, \hat{p}_0)$ k nule v pravděpodobnosti nekonverguje. V případě, že hodnota tohoto Poissonova nulového indexu bude výrazně vyšší (nebo nižší) než nula, budeme tedy nejspíš schopni zamítnout nulovou hypotézu. Zde je ovšem důležité podotknout, že platnost $f(\mathsf{E}X_1,\, p_0) = 0$ nutně neimplikuje, že X_1 má Poissonovo rozdělení (více v sekci 2.4).

Abychom mohli provést test nulové hypotézy, potřebujeme znát rozdělení naší testové statistiky. Klíčovým nástrojem k určování asymptotického rozdělení nějaké transformace náhodného vektoru je Δ-věta (věta 3 v předchozí kapitole). Při pohledu na předpoklady Δ-věty vidíme, že abychom ji mohli použít k získání asymptotického rozdělení $f(\overline{X}_n,\, \hat{p}_0)$, potřebujeme znát asymptotické rozdělení náhodného vektoru $(\overline{X}_n,\, \hat{p}_0)^\mathsf{T}$. Tento výsledek nám dává následující tvrzení.

Tvrzení 5. *Nechť $X_1, X_2, ..., X_n$ je náhodný výběr z rozdělení $\mathsf{Po}(\lambda)$, $\lambda > 0$. Potom platí*

$$\sqrt{n}\left(\begin{pmatrix}\overline{X}_n\\ \hat{p}_0\end{pmatrix} - \begin{pmatrix}\lambda\\ \mathrm{e}^{-\lambda}\end{pmatrix}\right) \xrightarrow[n\,\to\,\infty]{\mathsf{D}} \mathsf{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\lambda & -\lambda\mathrm{e}^{-\lambda}\\ -\lambda\mathrm{e}^{-\lambda} & \mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda})\end{pmatrix}\right).$$

Důkaz: Předpokládáme, že $X_i \sim \mathsf{Po}(\lambda)$, $i = 1, ..., n$. Použijeme mnohorozměrnou verzi centrální limitní věty (jejíž znění jsme si uvedli ve větě 2) na náhodné vektory $\left(X_i,\, \mathbf{1}_{[X_i=0]}\right)^\mathsf{T}$, $i = 1, ..., n$.

Z předpokladu nezávislosti a stejného rozdělení X_i, $i = 1, ..., n$, dostáváme, že i náhodné vektory $\left(X_i,\, \mathbf{1}_{[X_i=0]}\right)^\mathsf{T}$, $i = 1, ..., n$, jsou nezávislé a mají stejné rozdělení, neboť obě jejich složky jsou měřitelné funkce X_i. Zbývá tedy spočítat střední hodnotu těchto vektorů a určit jejich varianční matici.

Pro libovolné $i \in \{1, ..., n\}$ platí

$$\mathsf{E}X_i = \lambda, \qquad \mathsf{var}X_i = \lambda.$$

Podle tvrzení 1 má $\mathbf{1}_{[X_i=0]}$ alternativní rozdělení s pravděpodobností úspěchu $\mathsf{P}(X_1 = 0) = \mathrm{e}^{-\lambda}$, a tedy

$$\mathsf{E}(\mathbf{1}_{[X_i=0]}) = \mathrm{e}^{-\lambda}, \quad \mathsf{var}(\mathbf{1}_{[X_i=0]}) = \mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda}).$$

Nyní spočteme kovarianci X_i a $\mathbf{1}_{[X_i=0]}$ použitím známého vzorce $\mathsf{cov}(X,Y) = \mathsf{E}(XY) - \mathsf{E}X \cdot \mathsf{E}Y$.

$$\mathsf{cov}(X_i, \mathbf{1}_{[X_i=0]}) = \mathsf{E}(X_i \cdot \mathbf{1}_{[X_i=0]}) - \mathsf{E}X_i \cdot \mathsf{E}(\mathbf{1}_{[X_i=0]}) = \mathsf{E}(0) - \lambda \cdot \mathrm{e}^{-\lambda} = -\lambda\mathrm{e}^{-\lambda}.$$

Zde využíváme, že pro libovolné $\omega \in \Omega$ buď $X_i(\omega) = 0$, nebo $X_i(\omega) \neq 0$. Druhý případ implikuje $\mathbf{1}_{[X_i=0]}(\omega) = 0$, tedy jeden člen v součinu $X_i \cdot \mathbf{1}_{[X_i=0]}$ je vždy nulový.

Z předchozích výpočtů plyne, že náhodný vektor $\begin{pmatrix} X_i \\ \mathbf{1}_{[X_i=0]} \end{pmatrix}$, $i = 1, ..., n$, má vektor středních hodnot $\begin{pmatrix} \lambda \\ \mathrm{e}^{-\lambda} \end{pmatrix}$ a varianční matici $\begin{pmatrix} \lambda & -\lambda\mathrm{e}^{-\lambda} \\ -\lambda\mathrm{e}^{-\lambda} & \mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda}) \end{pmatrix}$.
Dále víme, že

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \lambda) = \sqrt{n}(\overline{X}_n - \lambda) \quad \text{a} \quad \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{1}_{[X_i=0]} - \mathrm{e}^{-\lambda}) = \sqrt{n}(\hat{p}_0 - \mathrm{e}^{-\lambda}).$$

Potom z centrální limitní věty dostáváme dokazovaný vztah.

□

Nyní již můžeme formulovat větu, která říká, jak vypadá za (1.1) asymptotické rozdělení obecného Poissonova nulového indexu.

Věta 6. *Nechť $X_1, X_2, ..., X_n$ je náhodný výběr z rozdělení $\mathsf{Po}(\lambda)$ pro nějaké $\lambda > 0$. Nechť dále $\hat{\mathrm{I}}_n := f(\overline{X}_n, \hat{p}_0)$ je Poissonův nulový index a f má spojité parciální derivace na okolí bodu $(\lambda, \mathrm{e}^{-\lambda})^\mathsf{T}$. Potom platí*

$$\sqrt{n} \cdot \hat{\mathrm{I}}_n \xrightarrow{\mathsf{D}} \mathsf{N}\Big(0,\ \mathbb{D}f(\lambda, \mathrm{e}^{-\lambda})\boldsymbol{\Sigma}(\mathbb{D}f(\lambda, \mathrm{e}^{-\lambda}))^\mathsf{T}\Big) \text{ pro } n \to \infty,$$

kde

$$\boldsymbol{\Sigma} = \begin{pmatrix} \lambda & -\lambda\mathrm{e}^{-\lambda} \\ -\lambda\mathrm{e}^{-\lambda} & \mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda}) \end{pmatrix} \quad a \quad \mathbb{D}f(x, y) = \left(\frac{\partial f(x, y)}{\partial x}, \frac{\partial f(x, y)}{\partial y}\right).$$

Důkaz: Tvrzení věty je důsledkem věty 3 (Δ-věta), předchozího tvrzení a toho, že za daných předpokladů

$$f(\lambda, \mathrm{e}^{-\lambda}) = 0$$

díky (i) v definici Poissonova nulového indexu.

□

Poznámka. Rozptyl normálního rozdělení v předchozí větě je zapsán ve tvaru lineární formy. Pomocí úpravy

$$\begin{pmatrix} A & B \end{pmatrix}\begin{pmatrix} \lambda & -\lambda\mathrm{e}^{-\lambda} \\ -\lambda\mathrm{e}^{-\lambda} & \mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda}) \end{pmatrix}\begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} A & B \end{pmatrix}\begin{pmatrix} \lambda A - \lambda\mathrm{e}^{-\lambda}B \\ -\lambda\mathrm{e}^{-\lambda}A + \mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda})B \end{pmatrix} =$$
$$\lambda A^2 - 2\lambda\mathrm{e}^{-\lambda}AB + \mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda})B^2$$

tak získáme i jeho explicitní vyjádření. Dostaneme

$$\mathbb{D}f(\lambda, \mathrm{e}^{-\lambda})\boldsymbol{\Sigma}(\mathbb{D}f(\lambda, \mathrm{e}^{-\lambda}))^\mathsf{T} = \lambda A^2 - 2\lambda\mathrm{e}^{-\lambda}AB + \mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda})B^2,$$

kde

$$A = \left.\frac{\partial f(x, y)}{\partial x}\right|_{\substack{x=\lambda \\ y=\mathrm{e}^{-\lambda}}} \qquad \text{a} \qquad B = \left.\frac{\partial f(x, y)}{\partial y}\right|_{\substack{x=\lambda \\ y=\mathrm{e}^{-\lambda}}}.$$

2.2 Příklady Poissonových nulových indexů

Uvedeme nyní několik způsobů, které vedou ke konstrukci konkrétních Poissonových nulových indexů. Nejprve se zaměříme na dva nulové indexy, které byly zavedeny ve výchozím článku Weiß a kol. (2019). Pro srovnání poté zmíníme i některé další varianty.

Vyjdeme ze vztahu

$$p_0 = \mathrm{e}^{-\lambda}, \tag{2.1}$$

který platí pro náhodnou veličinu X s rozdělením $\mathsf{Po}(\lambda)$. Podle definice Poissonova nulového indexu potřebujeme najít funkci f, jejíž definiční obor obsahuje množinu $(0,\,\infty) \times (0,\,1)$ a která pro libovolné $\lambda > 0$ splňuje

$$f(\lambda,\, p_0) = 0 \quad \Leftrightarrow \quad p_0 = \mathrm{e}^{-\lambda}.$$

Chceme tedy provést vhodné ekvivalentní úpravy (2.1), tak aby na pravé straně zůstala nula.

Pokud jsou obě strany (2.1) kladné, můžeme je zlogaritmovat. Potom dostaneme

$$\ln(p_0) = -\lambda.$$

Nyní již stačí vydělit výraz na levé i na pravé straně rovnosti konstantou $\lambda > 0$, a poté k oběma částem přičíst číslo jedna.

$$\begin{aligned} \frac{\ln(p_0)}{\lambda} &= -1 && / + 1 \\ \frac{\ln(p_0)}{\lambda} + 1 &= 0 \end{aligned}$$

Označme

$$\mathrm{I}_1 := f_1(\lambda,\, p_0),$$

kde f_1 je zobrazení definované jako

$$\begin{aligned} f_1 : \quad (0,\,\infty) \times (0,\,\infty) \quad &\to \quad \mathbb{R}, \\ (x, y) \quad &\mapsto \quad 1 + \frac{\ln(y)}{x}. \end{aligned}$$

Potom pro libovolné $\lambda > 0$ a $p_0 > 0$ platí

$$f_1(\lambda,\, p_0) = 0 \quad \Leftrightarrow \quad \frac{\ln(p_0)}{\lambda} = -1 \quad \Leftrightarrow \quad \ln(p_0) = -\lambda \quad \Leftrightarrow \quad p_0 = \mathrm{e}^{-\lambda},$$

tedy f_1 splňuje (i) a (ii) v definici 2. Z toho plyne, že statistika

$$\hat{\mathrm{I}}_{1,n} := f_1(\overline{X}_n,\, \hat{p}_0)$$

je Poissonův nulový index.

V praxi se může stát, že pro náš náhodný výběr získáme odhady $\overline{X}_n = 0$, nebo $\hat{p}_0 = 0$. První možnost nastane právě tehdy, když $X_1 = X_2 = ... = X_n = 0$, a tedy bude zároveň platit $\hat{p}_0 = 1$. Druhou rovnost naopak dostaneme v situaci, kdy ani

jedno pozorování není nulové. Pro tyto případy je vhodné „spojitě“ dodefinovat $\hat{\mathrm{I}}_{1,n}$, například jako

$$\hat{\mathrm{I}}_{1,n} = \begin{cases} 1 + \frac{\ln(\hat{p}_0)}{\overline{X}_n} & \text{pro } \overline{X}_n > 0,\ \hat{p}_0 > 0, \\ 0 & \text{pro } \overline{X}_n = 0, \\ -\infty & \text{pro } \hat{p}_0 = 0, \end{cases}$$

kde v části pro $\overline{X}_n = 0$ využíváme, že v tomto případě $\hat{p}_0 = 1 = \mathrm{e}^0 = \mathrm{e}^{-\overline{X}_n}$. Zároveň platí $\lim_{x\to 0}(1 + \frac{\ln(\mathrm{e}^{-x})}{x}) = \lim_{x\to 0}(1 + \frac{-x}{x}) = 0$.

Je ovšem dobré si uvědomit, že při dostatečném rozsahu výběru obvykle k žádné z těchto „extrémních“ situací nedojde. Protože $X_1, X_2, ..., X_n$ jsou nezávislé a stejně rozdělené, snadno spočteme, že

$$\mathsf{P}(\overline{X}_n = 0) = \mathsf{P}(X_1 = 0,\ ...,\ X_n = 0) = \Big(\mathsf{P}(X_1 = 0)\Big)^n,$$

$$\mathsf{P}(\hat{p}_0 = 0) = \mathsf{P}(X_1 \neq 0,\ ...,\ X_n \neq 0) = \Big(\mathsf{P}(X_1 \neq 0)\Big)^n = \Big(1 - \mathsf{P}(X_1 = 0)\Big)^n.$$

Pokud je tedy $\mathsf{P}(X_1 = 0) \in (0,\ 1)$, obě dvě pravděpodobnosti výše konvergují pro $n \to \infty$ k nule.

Další možnost je vynásobit obě strany (2.1) nenulovou hodnotou e^λ a následně odečíst číslo jedna.

$$\begin{aligned} p_0 &= \mathrm{e}^{-\lambda} && / \cdot \mathrm{e}^\lambda \\ p_0\mathrm{e}^\lambda &= 1 && / - 1 \\ p_0\mathrm{e}^\lambda - 1 &= 0 \end{aligned}$$

Položíme $\mathrm{I}_2 := f_2(\lambda,\ p_0)$, kde

$$\begin{aligned} f_2 :\ \mathbb{R}^2 &\to \mathbb{R}, \\ (x, y) &\mapsto y\mathrm{e}^x - 1. \end{aligned}$$

Dostaneme opět

$$f_2(\lambda,\ p_0) = 0 \quad \Leftrightarrow \quad p_0\mathrm{e}^\lambda - 1 = 0 \quad \Leftrightarrow \quad p_0\mathrm{e}^\lambda = 1 \quad \Leftrightarrow \quad p_0 = \mathrm{e}^{-\lambda},$$

tedy

$$\hat{\mathrm{I}}_{2,n} := f_2(\overline{X}_n,\ \hat{p}_0) = \hat{p}_0\mathrm{e}^{\overline{X}_n} - 1$$

je podle definice 2 rovněž Poissonův nulový index.

Můžeme si všimnout, že zde, na rozdíl od předchozího příkladu, je zobrazení f_2 dobře definované pro libovolné hodnoty $\overline{X}_n$ a $\hat{p}_0$.

V předchozím případě jsme (2.1) dělili číslem $\mathrm{e}^{-\lambda}$. Alternativně bychom mohli tuto rovnost vydělit hodnotou p_0 za předpokladu, že je kladná.

$$\begin{aligned} p_0 &= \mathrm{e}^{-\lambda} && / : p_0 \\ 1 &= \frac{\mathrm{e}^{-\lambda}}{p_0} && / - 1 \\ 0 &= \frac{\mathrm{e}^{-\lambda}}{p_0} - 1 \end{aligned}$$

Pro tyto účely definujeme funkci

$$\begin{aligned} f_3: \quad \mathbb{R} \times (0,\, \infty) &\to \mathbb{R}, \\ (x, y) &\mapsto \frac{\mathrm{e}^{-x}}{y} - 1. \end{aligned}$$

Znovu obdržíme vztah

$$\mathrm{I}_3 := f_3(\lambda,\, p_0) = 0 \quad \Leftrightarrow \quad 1 = \frac{\mathrm{e}^{-\lambda}}{p_0} \quad \Leftrightarrow \quad p_0 = \mathrm{e}^{-\lambda}.$$

Získáme tak třetí nulový index $\hat{\mathrm{I}}_{3,n} := f_3(\overline{X}_n, \hat{p}_0)$, který podobně jako $\hat{\mathrm{I}}_{1,n}$ ošetříme i pro případ $\hat{p}_0 = 0$:

$$\hat{\mathrm{I}}_{3,n} = \begin{cases} \frac{\mathrm{e}^{-\overline{X}_n}}{\hat{p}_0} - 1 & \text{pro } \hat{p}_0 > 0, \\ \infty & \text{pro } \hat{p}_0 = 0. \end{cases}$$

Jiný způsob konstrukce nulového indexu opět využívá rovnost, kterou jsme získali aplikací přirozeného logaritmu na obě strany (2.1). Pro $p_0 \in (0,\, 1)$ bude logaritmus z p_0 různý od nuly, a budeme jím tedy moci celou rovnost vydělit.

$$\begin{aligned} &\ln(p_0) = -\lambda && / : \ln(p_0) \\ &1 = \frac{-\lambda}{\ln(p_0)} && / + \frac{\lambda}{\ln(p_0)} \\ &1 + \frac{\lambda}{\ln(p_0)} = 0 \end{aligned}$$

V tomto případě budeme uvažovat funkci

$$\begin{aligned} f_4: \quad \mathbb{R} \times (0,\, 1) &\to \mathbb{R}, \\ (x, y) &\mapsto 1 + \frac{x}{\ln(y)}. \end{aligned}$$

Dostaneme opět

$$\mathrm{I}_4 := f_4(\lambda,\, p_0) = 0 \quad \Leftrightarrow \quad p_0 = \mathrm{e}^{-\lambda},$$

což nám dává další Poissonův nulový index $\hat{\mathrm{I}}_{4,n} := f_4(\overline{X}_n, \hat{p}_0)$. Stejně jako u předchozích indexů ho dodefinujeme i pro hodnoty na hranici definičního oboru.

$$\hat{\mathrm{I}}_{4,n} = \begin{cases} 1 + \frac{\overline{X}_n}{\ln(\hat{p}_0)} & \text{pro } \hat{p}_0 \in (0,\, 1), \\ 1 & \text{pro } \hat{p}_0 = 0, \\ 0 & \text{pro } \hat{p}_0 = 1. \end{cases}$$

Tvrzení 4 v sekci 2.1 nám říká, že $\hat{\mathrm{I}}_{k,n} = f_j(\overline{X}_n, \hat{p}_0)$ jsou konzistentními odhady $f_k(\mathsf{E}X_1,\, p_0)$, za předpokladu, že f_k jsou spojité v bodě $(\mathsf{E}X_1,\, p_0)^\mathsf{T}$, pro $k \in \{1,\, 2,\, 3,\, 4\}$. Protože zmíněné funkce jsou spojité na celých svých definičních oborech, stačí požadovat, aby $(\mathsf{E}X_1,\, p_0)^\mathsf{T}$ bylo prvkem jejich definičního oboru. Podle 2. bodu tvrzení 4 navíc $\hat{\mathrm{I}}_{k,n}$; $k = 1,\, 2,\, 3,\, 4$; za platnosti hypotézy konvergují v pravděpodobnosti k nule.

Tvrzení 7. *Nechť $X_1, X_2, ..., X_n$ je náhodný výběr z libovolného diskrétního rozdělení se střední hodnotou $\mathsf{E}X_1 \in (0,\, \infty)$. Pokud $p_0 = \mathsf{P}(X_1 = 0) \in (0,\, 1)$, potom*

$$\hat{\mathrm{I}}_{1,n} \xrightarrow[n\,\to\,\infty]{\mathsf{P}} 1 + \frac{\ln(p_0)}{\mathsf{E}X_1}, \qquad \hat{\mathrm{I}}_{2,n} \xrightarrow[n\,\to\,\infty]{\mathsf{P}} p_0 \mathrm{e}^{\mathsf{E}X_1} - 1,$$

$$\hat{\mathrm{I}}_{3,n} \xrightarrow[n\,\to\,\infty]{\mathsf{P}} \frac{\mathrm{e}^{-\mathsf{E}X_1}}{p_0} - 1, \qquad \hat{\mathrm{I}}_{4,n} \xrightarrow[n\,\to\,\infty]{\mathsf{P}} 1 + \frac{\mathsf{E}X_1}{\ln(p_0)},$$

kde $\hat{\mathrm{I}}_{k,n}$, $k = 1,\, 2,\, 3,\, 4$, jsou Poissonovy nulové indexy definované výše.

Speciálně, pokud $X_1 \sim \mathsf{Po}(\lambda)$ pro nějaké $\lambda > 0$, platí

$$\hat{\mathrm{I}}_{k,n} \xrightarrow[n\,\to\,\infty]{\mathsf{P}} 0, \qquad\qquad k = 1,\, 2,\, 3,\, 4.$$

Pro další práci s nulovými indexy se hodí znát jejich asymptotické rozdělení za platnosti hypotézy (1.1). V následující větě se proto budeme zabývat jeho odvozením.

Věta 8. *Uvažujme náhodný výběr $X_1, X_2, ..., X_n$ z rozdělení $\mathsf{Po}(\lambda)$ pro nějaké $\lambda > 0$. Potom*

1.
$$\sqrt{n} \cdot \hat{\mathrm{I}}_{1,n} \xrightarrow[n\,\to\,\infty]{\mathsf{D}} \mathsf{N}\!\left(0,\, \frac{\mathrm{e}^{\lambda} - \lambda - 1}{\lambda^2}\right),$$

2.
$$\sqrt{n} \cdot \hat{\mathrm{I}}_{2,n} \xrightarrow[n\,\to\,\infty]{\mathsf{D}} \mathsf{N}\big(0,\, \mathrm{e}^{\lambda} - \lambda - 1\big),$$

3.
$$\sqrt{n} \cdot \hat{\mathrm{I}}_{3,n} \xrightarrow[n\,\to\,\infty]{\mathsf{D}} \mathsf{N}\big(0,\, \mathrm{e}^{\lambda} - \lambda - 1\big),$$

4.
$$\sqrt{n} \cdot \hat{\mathrm{I}}_{4,n} \xrightarrow[n\,\to\,\infty]{\mathsf{D}} \mathsf{N}\!\left(0,\, \frac{\mathrm{e}^{\lambda} - \lambda - 1}{\lambda^2}\right).$$

Důkaz:

1.

Víme, že $\hat{\mathrm{I}}_{1,n} = f_1(\overline{X}_n,\, \hat{p}_0)$ je nulový index. Funkce

$$f_1 :\ (x, y) \mapsto 1 + \frac{\ln(y)}{x}$$

je spojitě diferencovatelná na množině $(0,\, \infty) \times (0,\, \infty)$, a tedy i v okolí bodu $\boldsymbol{\mu} := (\lambda,\, \mathrm{e}^{-\lambda})^{\mathsf{T}}$. Potom podle věty 6 platí pro $n \to \infty$

$$\sqrt{n} \cdot \hat{\mathrm{I}}_{1,n} \xrightarrow{\mathsf{D}} \mathsf{N}\big(0,\ \mathbb{D}f_1(\boldsymbol{\mu})\boldsymbol{\Sigma}(\mathbb{D}f_1(\boldsymbol{\mu}))^{\mathsf{T}}\big), \tag{2.2}$$

kde

$$\boldsymbol{\Sigma} = \begin{pmatrix} \lambda & -\lambda\mathrm{e}^{-\lambda} \\ -\lambda\mathrm{e}^{-\lambda} & \mathrm{e}^{-\lambda}(1 - \mathrm{e}^{-\lambda}) \end{pmatrix} \quad \text{a} \quad \mathbb{D}f_1(x,\, y) = \left(-\frac{\ln(y)}{x^2},\, \frac{1}{xy}\right).$$

Nyní již stačí dosadit konkrétní hodnoty.

$$\mathbb{D}f_1(\boldsymbol{\mu}) = \left(-\frac{\ln(\mathrm{e}^{-\lambda})}{\lambda^2},\, \frac{1}{\lambda\mathrm{e}^{-\lambda}}\right) = \left(\frac{1}{\lambda},\, \frac{1}{\lambda\mathrm{e}^{-\lambda}}\right)$$

Dále platí

$$\mathbb{D}f_1(\boldsymbol{\mu})\boldsymbol{\Sigma}(\mathbb{D}f_1(\boldsymbol{\mu}))^{\mathsf{T}} = \begin{pmatrix} \frac{1}{\lambda} & \frac{1}{\lambda \mathrm{e}^{-\lambda}} \end{pmatrix} \begin{pmatrix} \lambda & -\lambda \mathrm{e}^{-\lambda} \\ -\lambda \mathrm{e}^{-\lambda} & \mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda}) \end{pmatrix} \begin{pmatrix} \lambda^{-1} \\ \frac{1}{\lambda \mathrm{e}^{-\lambda}} \end{pmatrix} =$$

$$\begin{pmatrix} \frac{1}{\lambda} & \frac{1}{\lambda \mathrm{e}^{-\lambda}} \end{pmatrix} \begin{pmatrix} 0 \\ -\mathrm{e}^{-\lambda} + \frac{1-\mathrm{e}^{-\lambda}}{\lambda} \end{pmatrix} = \frac{1-\mathrm{e}^{-\lambda}}{\lambda^2 \mathrm{e}^{-\lambda}} - \frac{\mathrm{e}^{-\lambda}}{\lambda \mathrm{e}^{-\lambda}} = \frac{1-\mathrm{e}^{-\lambda}-\lambda \mathrm{e}^{-\lambda}}{\lambda^2 \mathrm{e}^{-\lambda}} = \frac{\mathrm{e}^{\lambda}-1-\lambda}{\lambda^2}.$$

Dosazením do (2.2) dostaneme $\sqrt{n} \cdot \hat{\mathrm{I}}_{1,n} \xrightarrow[n \to \infty]{\mathsf{D}} \mathsf{N}(0, \frac{\mathrm{e}^{\lambda}-\lambda-1}{\lambda^2})$, což jsme chtěli dokázat.

2.

Postupujeme analogicky jako v předchozí části, pouze uvažujeme zobrazení f_2. Budeme se tedy držet značení $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$.

Funkce

$$f_2 : \ (x, y) \ \mapsto \ \mathrm{e}^x y - 1$$

má spojité parciální derivace na $\mathbb{R} \times \mathbb{R}$, a tedy i v okolí bodu $\boldsymbol{\mu}$. Platí

$$\mathbb{D}f_2(x,\, y) = (\mathrm{e}^x y,\, \mathrm{e}^x).$$

Podle věty 6 tedy pro $n \to \infty$ máme

$$\sqrt{n} \cdot \hat{\mathrm{I}}_{2,n} \xrightarrow{\mathsf{D}} \mathsf{N}\Big(0,\, \mathbb{D}f_2(\boldsymbol{\mu})\boldsymbol{\Sigma}(\mathbb{D}f_2(\boldsymbol{\mu}))^{\mathsf{T}}\Big). \tag{2.3}$$

Dále

$$\mathbb{D}f_2(\boldsymbol{\mu}) = (\mathrm{e}^{\lambda} \cdot \mathrm{e}^{-\lambda},\, \mathrm{e}^{\lambda}) = (1,\, \mathrm{e}^{\lambda}).$$

Díky explicitnímu vyjádření $\mathbb{D}f_2(\boldsymbol{\mu})\boldsymbol{\Sigma}(\mathbb{D}f_2(\boldsymbol{\mu}))^{\mathsf{T}}$ (viz poznámka pod větou 6) tak stačí spočítat

$$\lambda \cdot 1^2 - 2\lambda \mathrm{e}^{-\lambda} \cdot 1 \cdot \mathrm{e}^{\lambda} + \mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda})(\mathrm{e}^{\lambda})^2 = \lambda - 2\lambda + \mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda})\mathrm{e}^{2\lambda} =$$
$$-\lambda + \mathrm{e}^{\lambda}(1-\mathrm{e}^{-\lambda}) = \mathrm{e}^{\lambda} - \lambda - 1.$$

Potom dosazením do (2.3) dostaneme $\sqrt{n}(\hat{\mathrm{I}}_{2,n} - 0) \xrightarrow[n \to \infty]{\mathsf{D}} \mathsf{N}(0,\, \mathrm{e}^{\lambda} - \lambda - 1)$.

3.

Zobrazení

$$f_3 : \ (x, y) \ \mapsto \ \frac{\mathrm{e}^{-x}}{y} - 1$$

je spojitě diferencovatelné na intervalu $\mathbb{R} \times (0,\, \infty)$, a tedy opět i v okolí bodu $\boldsymbol{\mu}$. Dále

$$\mathbb{D}f_3(x,\, y) = \left(-\frac{\mathrm{e}^{-x}}{y},\ -\frac{\mathrm{e}^{-x}}{y^2} \right) \qquad \text{a} \qquad \mathbb{D}f_3(\boldsymbol{\mu}) = (-1,\ -\mathrm{e}^{\lambda}).$$

Všimněme si, že

$$\mathbb{D}f_3(\boldsymbol{\mu}) = -\mathbb{D}f_2(\boldsymbol{\mu}),$$

tudíž rozptyl asymptotického rozdělení bude stejný jako v předchozím bodě.

4.

Znovu vidíme, že funkce

$$f_4 : \ (x, y) \ \mapsto \ 1 + \frac{x}{\ln(y)}$$

má spojité parciální derivace na $\mathbb{R} \times (0,\, 1)$, a tedy zřejmě i na okolí bodu $\boldsymbol{\mu}$. Dostaneme

$$\mathbb{D}f_4(x,\, y) = \left(\frac{1}{\ln(y)},\ -\frac{x}{y\ln^2(y)}\right)$$

Potom

$$\mathbb{D}f_4(\boldsymbol{\mu}) = \left(\frac{1}{\ln(\mathrm{e}^{-\lambda})},\ -\frac{\lambda}{\mathrm{e}^{-\lambda}\ln^2(\mathrm{e}^{-\lambda})}\right) = \left(-\frac{1}{\lambda},\ -\frac{1}{\mathrm{e}^{-\lambda}\lambda}\right),$$

což nám dá

$$\begin{aligned}
&\mathbb{D}f_4(\boldsymbol{\mu})\boldsymbol{\Sigma}(\mathbb{D}f_4(\boldsymbol{\mu}))^{\mathsf{T}} = \lambda\Big(-\frac{1}{\lambda}\Big)^2 - 2\lambda\mathrm{e}^{-\lambda}\Big(-\frac{1}{\lambda}\Big)\Big(-\frac{1}{\mathrm{e}^{-\lambda}\lambda}\Big) + \mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda})\Big(-\frac{1}{\mathrm{e}^{-\lambda}\lambda}\Big)^2 \\
&= \frac{\lambda}{\lambda^2} - \frac{2\lambda\mathrm{e}^{-\lambda}}{\mathrm{e}^{-\lambda}\lambda^2} + \mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda})\cdot\frac{1}{\mathrm{e}^{-2\lambda}\lambda^2} = \frac{\lambda}{\lambda^2} - \frac{2\lambda}{\lambda^2} + \mathrm{e}^{-\lambda}(1-\mathrm{e}^{-\lambda})\cdot\frac{\mathrm{e}^{2\lambda}}{\lambda^2} \\
&= \frac{\lambda - 2\lambda + \mathrm{e}^{\lambda}(1-\mathrm{e}^{-\lambda})}{\lambda^2} = \frac{\mathrm{e}^{\lambda} - \lambda - 1}{\lambda^2}.
\end{aligned}$$

□

Věta 8 nám říká, že za nulové hypotézy (1.1) a pro velké hodnoty n mají veličiny $\sqrt{n}\cdot\hat{\mathrm{I}}_{k,n}$ (pro $k \in \{1,\, 2,\, 3,\, 4\}$) rozdělení, které je blízké normálnímu rozdělení s nulovou střední hodnotou a danými rozptyly.

Můžeme si ještě všimnout, že pro libovolné $0 < \lambda < 1$ bude $\frac{1}{\lambda^2} < 1$, tudíž

$$\mathsf{avar}(\hat{\mathrm{I}}_{1,n}) = \mathsf{avar}(\hat{\mathrm{I}}_{4,n}) > \mathsf{avar}(\hat{\mathrm{I}}_{2,n}) = \mathsf{avar}(\hat{\mathrm{I}}_{3,n}),$$

kde $\mathsf{avar}(\hat{\mathrm{I}}_{k,n})$ značí asymptotický rozptyl nulového indexu $\hat{\mathrm{I}}_{k,n}$ za platnosti (1.1). Naopak je-li $\lambda > 1$, platí $\frac{1}{\lambda^2} > 1$, a tedy asymptotické rozptyly nulových indexů splňují nerovnost

$$\mathsf{avar}(\hat{\mathrm{I}}_{1,n}) = \mathsf{avar}(\hat{\mathrm{I}}_{4,n}) < \mathsf{avar}(\hat{\mathrm{I}}_{2,n}) = \mathsf{avar}(\hat{\mathrm{I}}_{3,n}).$$

Speciálně, pro $\lambda = 1$ budou všechny čtyři asymptotické rozptyly stejně velké.

2.3 Testování pomocí nulových indexů

V předchozích částech této kapitoly jsme zavedli pojem Poissonova nulového indexu a odvodili jeho asymptotické rozdělení, přičemž jsme se podívali i na čtyři konkrétní případy. Nyní se zaměříme na to, jak tyto poznatky využít k našemu stanovenému cíli – k testování nulové hypotézy (1.1).

Z předešlých dvou sekcí víme, že asymptotické rozdělení nulového indexu (za platnosti hypotézy) obecně závisí na hodnotě λ, která je pro nás neznámá. Abychom mohli testovat (1.1), bude tedy nutné upravit statistiku $\hat{\mathrm{I}}_n$ tak, aby její rozdělení již na neznámých parametrech nezáviselo.

Připomeňme, že věta 6 nám říká, že za určitých předpokladů má nulový index za platnosti hypotézy (1.1) asymptotické rozdělení (používáme-li značení zavedené ve zmíněné větě)

$$\sqrt{n}\cdot\hat{\mathrm{I}}_n \xrightarrow[n\,\to\,\infty]{\mathsf{D}} \mathsf{N}\Big(0,\ \mathbb{D}f(\boldsymbol{\mu})\boldsymbol{\Sigma}(\mathbb{D}f(\boldsymbol{\mu}))^{\mathsf{T}}\Big).$$

Jednou z vlastností normálního rozdělení je, že

$$Z \sim \mathsf{N}(\mu,\, \sigma^2) \quad \Longrightarrow \quad c \cdot Z \sim \mathsf{N}(c\mu,\, c^2\sigma^2),$$

kde $c \in \mathbb{R}$ je libovolná konstanta a Z je nějaká náhodná veličina. Díky Cramér-Sluckého větě, jejíž znění lze nalézt například v knize van der Vaart (1998, Věta 2.8), tento vztah platí i v případě, kdy symbol $\sim$ nahradíme konvergencí v distribuci. Je-li $\mathbb{D}f(\boldsymbol{\mu})\boldsymbol{\Sigma}(\mathbb{D}f(\boldsymbol{\mu}))^\mathsf{T} > 0$, můžeme položit

$$c := \left(\mathbb{D}f(\boldsymbol{\mu})\boldsymbol{\Sigma}(\mathbb{D}f(\boldsymbol{\mu}))^\mathsf{T}\right)^{-\frac{1}{2}} > 0,$$

a dostaneme tak

$$\sqrt{n} \cdot \left(\mathbb{D}f(\boldsymbol{\mu})\boldsymbol{\Sigma}(\mathbb{D}f(\boldsymbol{\mu}))^\mathsf{T}\right)^{-\frac{1}{2}} \hat{\mathrm{I}}_n \xrightarrow[n\,\to\,\infty]{\mathsf{D}} \mathsf{N}(0,\, 1).$$

Vidíme, že za (1.1) konverguje výše uvedený výraz v distribuci k náhodné veličině s normovaným normálním rozdělením, tedy rozdělením nezávisejícím na neznámých parametrech. Stále však nulový index $\hat{\mathrm{I}}_n$ dělíme odmocninou z jeho asymptotického rozptylu, kde se objevuje neznámá konstanta λ. K získání testové statistiky budeme tedy muset tento rozptyl nějak odhadnout.

Z věty 6 víme, že za nulové hypotézy závisí rozptyl asymptotického rozdělení $\hat{\mathrm{I}}_n$ na λ a $\mathrm{e}^{-\lambda}$, tedy celkově jde tento rozptyl vyjádřit jako $h(\lambda)$, kde

$$h : (0,\, \infty) \to (0,\, \infty),$$

a λ je parametr daného Poissonova rozdělení. Jednou z možností je tedy nahradit neznámou proměnnou λ výběrovým průměrem $\overline{X}_n$. Platí totiž

$$\overline{X}_n \xrightarrow[n\,\to\,\infty]{\mathsf{P}} \mathsf{E}X_1 \overset{(1.1)}{=} \lambda.$$

Pokud je funkce h spojitá v bodě $\mathsf{E}X_1$, dostaneme z věty o spojité transformaci

$$h(\overline{X}_n) \xrightarrow[n\,\to\,\infty]{\mathsf{P}} h(\mathsf{E}X_1) \overset{(1.1)}{=} h(\lambda),$$

a $h(\overline{X}_n)$ je tak konzistentní odhad $h(\lambda)$.

Nyní již můžeme zformulovat klíčovou větu, která ukazuje, jakým způsobem lze pomocí nulových indexů testovat (1.1). Pro připomenutí je dobré zmínit, že pro dané $\alpha \in (0,\, 1)$ budeme symbolem $u_{1-\frac{\alpha}{2}}$ značit $(1 - \frac{\alpha}{2})$-tý kvantil normovaného normálního rozdělení. Platí $u_{1-\frac{\alpha}{2}} = \Phi^{-1}(1 - \frac{\alpha}{2})$, kde Φ je distribuční funkce $\mathsf{N}(0,\, 1)$,

Věta 9. *Nechť $X_1, X_2, ..., X_n$ je náhodný výběr z nějakého diskrétního rozdělení, jehož nosič je podmnožina $\mathbb{N}_0$. Nechť $\alpha \in (0,\, 1)$ je předem stanovené číslo. Uvažujme Poissonův nulový index*

$$\hat{\mathrm{I}}_n := f(\overline{X}_n,\, \hat{p}_0),$$

pro nějž za nulové hypotézy (1.1) platí

$$\sqrt{n} \cdot \hat{\mathrm{I}}_n \xrightarrow[n\,\to\,\infty]{\mathsf{D}} \mathsf{N}\Big(0,\, h(\mathsf{E}X_1)\Big),$$

kde h je nějaká funkce, $h : (0, \infty) \to (0, \infty)$. Nechť $\hat{h}_n$ je konzistentní odhad $h(\mathsf{E}X_1)$, který nezávisí na neznámých parametrech.
Potom za platnosti hypotézy (1.1) platí

$$\mathrm{T}_n := \sqrt{n} \cdot \frac{\hat{\mathrm{I}}_n}{\sqrt{\hat{h}_n}} \xrightarrow[n \to \infty]{\mathsf{D}} \mathsf{N}(0,\, 1).$$

Položíme-li

$$C(\alpha) := \Big(-\infty, -u_{1-\frac{\alpha}{2}}\Big\rangle \cup \Big\langle u_{1-\frac{\alpha}{2}}, \infty\Big),$$

potom test (1.1) řídící se rozhodovacím pravidlem:

$$\textit{zamítneme } (\mathsf{H}_0) \quad \Longleftrightarrow \quad \mathrm{T}_n \in C(\alpha)$$

má asymptotickou hladinou významnosti α.

Důkaz: T_n lze rozložit na součin dvou výrazů jako

$$\mathrm{T}_n = \sqrt{n} \cdot \frac{\hat{\mathrm{I}}_n}{\sqrt{\hat{h}_n}} = \sqrt{n}\frac{\hat{\mathrm{I}}_n}{\sqrt{h(\mathsf{E}X_1)}} \cdot \frac{\sqrt{h(\mathsf{E}X_1)}}{\sqrt{\hat{h}_n}},$$

přičemž z vlastností normálního rozdělení diskutovaných před touto větou máme

$$\sqrt{n}\frac{\hat{\mathrm{I}}_n}{\sqrt{h(\mathsf{E}X_1)}} \xrightarrow[n \to \infty]{\mathsf{D}} \mathsf{N}(0,\, 1).$$

Dále

$$\frac{\sqrt{h(\mathsf{E}X_1)}}{\sqrt{\hat{h}_n}} \xrightarrow[n \to \infty]{\mathsf{P}} 1,$$

což plyne z konzistence $\hat{h}_n$ a věty o spojité transformaci (druhá odmocnina je spojitá funkce). Potom podle Cramérovy-Sluckého věty

$$\mathrm{T}_n \xrightarrow[n \to \infty]{\mathsf{D}} 1 \cdot Z = Z, \quad \text{kde } Z \sim \mathsf{N}(0,\, 1).$$

K důkazu zbytku věty si nejdříve uvědomme, že

$$\mathsf{P}\Big(|\mathrm{T}_n| < u_{1-\frac{\alpha}{2}}\Big) = \mathsf{P}\Big(-u_{1-\frac{\alpha}{2}} < \mathrm{T}_n < u_{1-\frac{\alpha}{2}}\Big) = \mathsf{P}\Big(\mathrm{T}_n < u_{1-\frac{\alpha}{2}}\Big) - \mathsf{P}\Big(\mathrm{T}_n \le -u_{1-\frac{\alpha}{2}}\Big).$$

Z definice konvergence v distribuci tedy máme (za platnosti hypotézy)

$$\lim_{n\to\infty} \mathsf{P}\Big(\mathrm{T}_n < u_{1-\frac{\alpha}{2}}\Big) = \Phi(u_{1-\frac{\alpha}{2}}{}^{-}) = \Phi(u_{1-\frac{\alpha}{2}}),$$

kde poslední rovnost plyne ze spojitosti funkce Φ na celé reálné ose, a podobně

$$\lim_{n\to\infty} \mathsf{P}\Big(\mathrm{T}_n \le -u_{1-\frac{\alpha}{2}}\Big) = \Phi(-u_{1-\frac{\alpha}{2}}).$$

Díky tomu, že normované normální rozdělení je symetrické kolem nuly, dostaneme

$$\Phi(-u_{1-\frac{\alpha}{2}}) = 1 - \Phi(u_{1-\frac{\alpha}{2}}).$$

Potom za platnosti (1.1)

$$\lim_{n\to\infty} \mathsf{P}\Big(|\mathrm{T}_n| < u_{1-\frac{\alpha}{2}}\Big) = \Phi(u_{1-\frac{\alpha}{2}}) - \Phi(-u_{1-\frac{\alpha}{2}}) = \Phi(u_{1-\frac{\alpha}{2}}) - (1 - \Phi(u_{1-\frac{\alpha}{2}}))$$
$$= 1 - \tfrac{\alpha}{2} - \Big(1 - (1 - \tfrac{\alpha}{2})\Big) = 1 - \tfrac{\alpha}{2} - \tfrac{\alpha}{2} = 1 - \alpha.$$

Za (H_0) tedy

$$\lim_{n\to\infty} \mathsf{P}\Big(\mathrm{T}_n \in C(\alpha)\Big) = \lim_{n\to\infty} \mathsf{P}\Big(|\mathrm{T}_n| \geq u_{1-\frac{\alpha}{2}}\Big) = 1 - \lim_{n\to\infty} \mathsf{P}\Big(|\mathrm{T}_n| < u_{1-\frac{\alpha}{2}}\Big) = \alpha,$$

což odpovídá tomu, že test $\Big(\mathrm{T}_n,\, C(\alpha)\Big)$ má hladinu α asymptoticky.

□

Zde je dobré připomenout, že Poissonův nulový index $\hat{\mathrm{I}}_n$, který se vyskytuje v čitateli T_n, je za (1.1) blízký nule (tvrzení 4). Ve prospěch alternativy tak svědčí vysoké záporné a kladné hodnoty statistiky T_n.

V souvislosti s testováním je rovněž dobré zmínit p-hodnotu, která je definovaná jako

$$p = \inf\{\alpha \in (0,\, 1\rangle :\ t \in C(\alpha)\},$$

kde t značí realizovanou hodnotu testové statistiky T_n. Dosazením kritických hodnot a drobnou úpravou získáme, že p-hodnotu výše uvedeného testu jde vyjádřit ve tvaru

$$p = \inf\{\alpha \in (0,\, 1\rangle :\ |t| \geq u_{1-\frac{\alpha}{2}}\} = \inf\{\alpha \in (0,\, 1\rangle :\ |t| \geq \Phi^{-1}(1-\tfrac{\alpha}{2})\} =$$
$$\inf\{\alpha \in (0,\, 1\rangle :\ \Phi(|t|) \geq 1 - \tfrac{\alpha}{2}\} = \inf\{\alpha \in (0,\, 1\rangle :\ \alpha \geq 2(1 - \Phi(|t|))\} =$$
$$2(1 - \Phi(|t|)).$$

U třetí rovnosti zde využíváme, že distribuční funkce Φ je neklesající.

P-hodnota podle své definice vyjadřuje nejmenší možnou hladinu, na které bychom zamítali nulovou hypotézu. Pokud tedy pro naše data vyjde p-hodnota, která je menší než předem stanovená hladina $\alpha \in (0,\, 1)$, budeme moci zamítnout (1.1). Alternativně tak lze při testování hypotézy rozhodovat o jejím zamítnutí pomocí pravidla

$$\text{zamítneme } (\mathsf{H}_0) \quad \Longleftrightarrow \quad p \leq \alpha.$$

Analogii věty 9 můžeme podobně jako jiné obecné věty zformulovat ještě pro nulové indexy zavedené v sekci 2.2.

Věta 10. *Nechť $X_1, X_2, ..., X_n$ je náhodný výběr z nějakého diskrétního rozdělení, jehož nosič je podmnožina $\mathbb{N}_0$, a $\alpha \in (0,\, 1)$ je předem stanovené číslo.*
Označme

$$\mathrm{T}_{1,n} := \frac{\hat{\mathrm{I}}_{1,n}}{\sqrt{\frac{\mathrm{e}^{\overline{X}_n} - \overline{X}_n - 1}{n(\overline{X}_n)^2}}}, \qquad \mathrm{T}_{2,n} := \sqrt{n} \cdot \frac{\hat{\mathrm{I}}_{2,n}}{\sqrt{\mathrm{e}^{\overline{X}_n} - \overline{X}_n - 1}},$$

$$\mathrm{T}_{3,n} := \sqrt{n} \cdot \frac{\hat{\mathrm{I}}_{3,n}}{\sqrt{\mathrm{e}^{\overline{X}_n} - \overline{X}_n - 1}}, \qquad \mathrm{T}_{4,n} := \frac{\hat{\mathrm{I}}_{4,n}}{\sqrt{\frac{\mathrm{e}^{\overline{X}_n} - \overline{X}_n - 1}{n(\overline{X}_n)^2}}}.$$

Potom za platnosti nulové hypotézy (1.1) platí pro $k \in \{1, 2, 3, 4\}$

$$\mathrm{T}_{k,n} \xrightarrow[n \to \infty]{\mathsf{D}} \mathsf{N}(0, 1).$$

Dále statistický test (1.1) řídící se pravidlem:

$$\textit{zamítneme } (\mathsf{H}_0) \quad \Longleftrightarrow \quad |\mathrm{T}_{k,n}| \geq u_{1-\frac{\alpha}{2}}$$

má asymptotickou hladinu významnosti α, $k = 1, 2, 3, 4$.

2.4 Úskalí testování založeného na nulových indexech

Testy dobré shody popsané v této kapitole jsou takzvané asymptotické testy, což znamená, že dosahují požadované hladiny významnosti α pouze asymptoticky. Skutečná hladina při daném rozsahu výběru tak může být výrazně vyšší.

Z tohoto hlediska by variantou na možné vylepšení mohlo být nahrazení kvantilu normovaného normálního rozdělení kvantilem Studentova t-rozdělení s $n - 1$ stupni volnosti, kde n je rozsah výběru. Pro zamítání nulové hypotézy bychom potom používali pravidlo:

$$\text{zamítneme } (\mathsf{H}_0) \quad \Longleftrightarrow \quad |\mathrm{T}_n| \geq t_{n-1}(1 - \tfrac{\alpha}{2}),$$

kde $t_{n-1}(1 - \frac{\alpha}{2})$ značí $(1 - \frac{\alpha}{2})$-tý kvantil t-rozdělení s $n - 1$ stupni volnosti. Platí totiž

$$t_{n-1}(1 - \tfrac{\alpha}{2}) \quad \searrow \quad u_{1-\frac{\alpha}{2}},$$

pro $n \to \infty$ a libovolné $\alpha \in (0, 1)$. Protože podle předchozího je pro každé přirozené $n \geq 2$

$$t_{n-1}(1 - \tfrac{\alpha}{2}) \quad \geq \quad u_{1-\frac{\alpha}{2}},$$

zamítali bychom tak nulovou hypotézu pro menší škálu hodnot statistiky T_n.

Výhodou tohoto přístupu je, že daný test by byl konzervativnější než test využívající kvantily $\mathsf{N}(0, 1)$, a tedy by měl, obzvlášť pro malé hodnoty rozsahu n, menší skutečnou hladinu významnosti a s tím související pravděpodobnost chyby I. druhu (tj. pravděpodobnost zamítnutí platné hypotézy). Naproti tomu, test pracující s kvantily normovaného normálního rozdělení má vyšší sílu, a tedy i menší pravděpodobnost chyby II. druhu (pravděpodobnost nezamítnutí neplatné hypotézy).

Jedna z dalších nevýhod testů dobré shody založených na nulovém indexu je, že testují pouze jeden aspekt pravděpodobnostního rozdělení. Z tohoto důvodu je lze použít pouze k odhalení toho, že dané rozdělení neodpovídá Poissonovu. Naopak nezamítnutí nulové hypotézy rozhodně nestačí k ukázání, že nějaký náhodný výběr pochází z Poissonova rozdělení. Důvodem je, že existují případy (viz příklad níže), kdy $\mathsf{P}(X = 0) = \mathrm{e}^{-\mathsf{E}X}$, ale náhodná veličina X nemá Poissonovo rozdělení.

Příklad. Nechť rozdělení náhodné veličiny X je dané vztahy

$$\mathsf{P}(X=0)=\frac{1}{\mathrm{e}},\qquad \mathsf{P}(X=1)=1-\frac{2}{\mathrm{e}},\qquad \mathsf{P}(X=2)=\frac{1}{\mathrm{e}}.$$

Potom X má diskrétní rozdělení s nosičem $\{0,\,1,\,2\}$ a střední hodnotou

$$\mathsf{E}X=\sum_{k=0}^{2}k\cdot\mathsf{P}(X=k)=1-\frac{2}{\mathrm{e}}+2\cdot\frac{1}{\mathrm{e}}=1.$$

Dále platí

$$\mathsf{P}(X=0)=\mathrm{e}^{-1}=\mathrm{e}^{-\mathsf{E}X}.$$

Zároveň však v tomto případě neexistuje konstanta $\lambda>0$, pro kterou by X mělo rozdělení $\mathsf{Po}(\lambda)$, neboť například

$$\mathsf{P}(X=3)=0,$$

ovšem pro veličinu s Poissonovým rozdělením by tato pravděpodobnost byla kladná.

Můžeme si všimnout, že v situaci, kdy by rozdělení našeho náhodného výběru $X_1, X_2, ..., X_n$ odpovídalo rozdělení náhodné veličiny X z předchozího příkladu, by nulová hypotéza (1.1) neplatila (a tedy by platila alternativa (H_1)). Zároveň však pro libovolný Poissonův nulový index

$$\hat{\mathrm{I}}_n:=f(\overline{X}_n,\,\hat{p}_0),$$

kde funkce f je spojitá v bodě $(\mathsf{E}X_1,\,p_0)^{\mathsf{T}}$, platí podle druhého bodu tvrzení 4

$$\hat{\mathrm{I}}_n\xrightarrow[n\,\to\,\infty]{\mathsf{P}}0,$$

neboť v tomto případě $\mathsf{E}X_1>0$ a

$$p_0=\mathrm{e}^{-\mathsf{E}X_1}.$$

Vzhledem k tomu, že u testů dobré shody založených na nulovém indexu svědčí proti (1.1) hodnoty statistiky $\hat{\mathrm{I}}_n$ významně vzdálené od 0, by zmíněné testy v tomto případě nejspíš porušení nulové hypotézy neodhalily.

Problém může nastat rovněž v situaci, kdy skutečné rozdělení náhodného výběru je Poissonovo s vyšší hodnotou parametru λ. Pokud totiž není rozsah našeho výběru dostatečně velký, může se stát, že žádná z napozorovaných hodnot není nulová. Potom získáme odhad

$$\hat{p}_0=0.$$

U některých nulových indexů (např. $\hat{\mathrm{I}}_{1,n}$) přitom pracujeme s logaritmem této relativní četnosti nebo jí dělíme. V takových případech by hodnota testové statistiky T_n byla $-\infty$ nebo ∞, a tedy bychom zamítli platnou hypotézu.

V tabulce níže můžeme vidět hodnoty $\mathrm{e}^{-\lambda}$ pro různé hodnoty λ. Ty odpovídají pravděpodobnostem $\mathsf{P}(X_1=0)$, kde $X_1\sim\mathsf{Po}(\lambda)$. V pravém sloupci je potom pro každé λ uveden průměrný počet pozorování, na které připadne jedna nulová realizace.

λ	$e^{-\lambda}$	e^{λ}
1	0,37	2,72
2	0,14	7,39
3	$4{,}98 \cdot 10^{-2}$	20,09
4	$1{,}83 \cdot 10^{-2}$	54,60
5	$6{,}74 \cdot 10^{-3}$	148,41
6	$2{,}48 \cdot 10^{-3}$	403,43
7	$9{,}12 \cdot 10^{-4}$	1 096,63
8	$3{,}35 \cdot 10^{-4}$	2 980,96
9	$1{,}23 \cdot 10^{-4}$	8 103,08
10	$4{,}54 \cdot 10^{-5}$	22 026,47

Lze si například všimnout, že již pro $X_1 \sim \mathsf{Po}(7)$ nabývá náhodná veličina X_1 hodnoty 0 přibližně v jednom z 1000 případů.

Obecně je tedy dobré používat testy založené na nulovém indexu spíše v případech, kdy chceme testovat hypotézu, že náhodný výběr $X_1, X_2, ..., X_n$ pochází z rozdělení $\mathsf{Po}(\lambda)$ pro nějakou malou hodnotu parametru $\lambda > 0$. V situacích, kdy očekáváme, že konstanta λ bude větší číslo, by mohl být vhodnější jiný typ testu dobré shody (viz kapitola 3).

3. Další varianty testů dobré shody

Testy dobré shody založené na nulovém indexu, kterým jsme se podrobně věnovali v kapitole 2, jsou pouze jedním z mnoha způsobů, jak lze nulovou hypotézu (1.1) testovat. Z tohoto důvodu se hodí zmínit i některé další metody, které se v této situaci také dají použít. Oproti předchozí kapitole však uvedené testy popíšeme daleko stručněji, s tím, že detaily lze nalézt v uvedené literatuře.

3.1 χ^2-testy dobré shody

Asi nejznámější test pro testování dobré shody s diskrétním rozdělením je tzv. Pearsonův χ^2-test dobré shody, kterému je věnována například velká část kapitoly 12 v knize Anděl (2007). Zde se pokusíme nastínit jeho základní princip.

Uvažujme nejprve situaci, kdy $X_1, X_2, ..., X_n$ je náhodný výběr z diskrétního rozdělení takového, že

$$\mathsf{P}\Big(X_1 \in \{1,\, 2, ...,\, K\}\Big) = 1,$$

kde K je nějaké známé přirozené číslo. Zajímá nás test nulové hypotézy

$$\mathsf{P}(X_1 = i) = p_i^0,\; i = 1, ...,\; K \tag{3.1}$$

pro nějaké předepsané pravděpodobnosti $p_1^0, ...,\; p_K^0$ splňující $\sum_{j=1}^K p_j^0 = 1$, tj. chceme testovat dobrou shodu s nějakým předem určeným diskrétním rozdělením, které má konečný nosič.

Označíme $Y_i := \sum_{j=1}^n \mathbf{1}_{[X_j=i]}$, $i = 1, ...,\; K$. Potom testová statistika bude mít tvar

$$\chi^2 = \sum_{j=1}^{K} \frac{\Big(Y_i - np_i^0\Big)^2}{np_i^0}.$$

Výraz v čitateli zde nějakým způsobem reflektuje rozdíl mezi empirickými četnostmi výskytu hodnot $1, ...,\; K$ (ty udávají veličiny Y_i) a jejich očekávanými četnostmi. Za nulové hypotézy by tyto dvě kvantity měly být podobné, tudíž proti (3.1) svědčí vysoké hodnoty testové statistiky.

Dá se ukázat, že za platnosti nulové hypotézy (3.1) má tato statistika asymptoticky χ^2-rozdělení s $K-1$ stupni volnosti (viz Anděl (2007), Věta 12.5). Pro získání asymptotického testu s hladinou významnosti α tedy budeme zamítat (3.1), pokud $\chi^2 \geq \chi^2_{K-1}(1-\alpha)$, kde $\chi^2_{K-1}(1-\alpha)$ značí $(1-\alpha)$-tý kvantil rozdělení χ^2_{K-1}.

K aplikování zmíněné metody na náš problém (testování (1.1)) brání skutečnost, že nosič Poissonova rozdělení je nekonečná množina. Tuto překážku je ovšem možné částečně odstranit tím, že si hodnoty, kterých veličina s Poissonovým rozdělením nabývá, rozdělíme do konečného počtu tříd.

Konkrétně si nejprve zvolíme konstanty $k \geq 0$ a $r \geq 0$. Dále definujeme veličiny

$$\begin{aligned} Y_1 &:= \sum_{j=1}^{n} \mathbf{1}_{[X_j \leq r]}, \\ Y_i &:= \sum_{j=1}^{n} \mathbf{1}_{[X_j = r+i-1]}, \quad i = 2, ..., k-1, \\ Y_k &:= \sum_{j=1}^{n} \mathbf{1}_{[X_j \geq r+k-1]}. \end{aligned}$$

Naše pozorování jsme tak rozdělili do k tříd. První z nich je tvořena realizacemi $\leq r$, a poslední realizacemi, které jsou $\geq r+k-1$. Do zbylých $k-2$ jsme zařadili pozorování, která jsou rovna číslům $r+1$, $r+2$, ..., $r+k-2$. Veličiny $Y_1, ..., Y_k$ budou podle své definice udávat počty realizací v jednotlivých třídách (takzvané empirické četnosti). Dále pravděpodobnost, že náhodná veličina X_1 spadne do i-té třídy ($i = 1, ..., k$) označíme p_i. Snadno ověříme, že pro $X_1 \sim \mathsf{Po}(\lambda)$ platí

$$p_1 = \sum_{j=0}^{r} q_j, \quad p_i = q_{r+i-1} \ \text{ pro } \ i = 2, ..., k-1, \quad p_k = \sum_{j=r+k-1}^{\infty} q_j,$$

kde

$$q_i = \mathsf{P}(X_1 = i) = \mathrm{e}^{-\lambda} \cdot \frac{\lambda^i}{i!}, \quad i = 0,\, 1,\, 2, ...$$

Principem testování bude opět porovnávání empirických četností Y_i s teoretickými četnostmi $n \cdot p_i$. Protože pravděpodobnosti p_i závisí na neznámém parametru λ, je potřeba tento parametr odhadnout. Podle sekce 12.3 z Anděl (2007) se odhad λ získá řešením rovnice

$$\sum_{i=1}^{k} \frac{Y_i}{p_i(\lambda)} \cdot \frac{\partial p_i(\lambda)}{\partial \lambda} = 0,$$

která se obvykle řeší numericky (iteračně). Můžeme si ještě všimnout, že

$$\begin{aligned} \frac{\partial p_1(\lambda)}{\partial \lambda} &= \sum_{j=0}^{r} \frac{\partial q_j(\lambda)}{\partial \lambda}, \qquad \frac{\partial p_k(\lambda)}{\partial \lambda} = \sum_{j=r+k-1}^{\infty} \frac{\partial q_j(\lambda)}{\partial \lambda}, \\ \frac{\partial p_i(\lambda)}{\partial \lambda} &= \frac{\partial q_{r+i-1}(\lambda)}{\partial \lambda}, \quad i = 2, ..., k-1, \end{aligned}$$

a platí

$$\frac{\partial q_i(\lambda)}{\partial \lambda} = -\mathrm{e}^{-\lambda} \cdot \frac{\lambda^i}{i!} + \mathrm{e}^{-\lambda} \cdot \frac{\lambda^{i-1}}{(i-1)!} = q_i(\lambda)\left(\frac{i}{\lambda} - 1\right).$$

Označíme-li jako $\hat{\lambda}$ získaný odhad λ, použijeme testovou statistiku

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(Y_i - np_i(\hat{\lambda})\right)^2}{np_i(\hat{\lambda})}.$$

Ta má za nulové hypotézy asymptoticky χ^2_{k-2} rozdělení (odhadem λ ztratíme navíc jeden stupeň volnosti). Abychom dostali test s asymptotickou hladinou významnosti α (pro předem zvolené $\alpha \in (0,\, 1)$), budeme se rozhodovat podle pravidla:

$$\text{zamítneme } (\mathsf{H}_0) \quad \Longleftrightarrow \quad \chi^2 \geq \chi^2_{k-2}(1-\alpha).$$

3.2 Testy založené na indexu disperze

Jak již bylo zmíněno v sekci 1.2, pro náhodný výběr $X_1, X_2, ..., X_n$ pocházející z rozdělení $\mathsf{Po}(\lambda)$, pro nějaké $\lambda > 0$, platí

$$\mathsf{E}X_1 = \lambda = \mathsf{var}X_1.$$

Z toho plyne, že pokud jsou střední hodnota a rozptyl nějaké náhodné veličiny různé, daná veličina nemá Poissonovo rozdělení. Jeden ze způsobů, který vede k zamítnutí hypotézy (1.1), je tedy ukázat, že rozdělení našeho náhodného výběru má rozdílný první a druhý centrovaný moment.

Protože skutečné hodnoty střední hodnoty a rozptylu nejsou známé, budeme se opět muset spokojit s odhady těchto kvantit. K odhadnutí střední hodnoty náhodného výběru použijeme stejně jako v předchozí části výběrový průměr $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Pro odhad rozptylu je obvyklou volbou výběrový rozptyl

$$S_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2,$$

jakožto nestranný a konzistentní odhad $\mathsf{var}X_1$ (viz (iii) ve větě 1).

Statistika používaná k testování nulové hypotézy bude v tomto případě nějaká funkce $\overline{X}_n$ a S_n^2, která dokáže odrážet, jak moc jsou jejich hodnoty odlišné. Konkrétně se bude jednat o takzvaný index disperze, jehož definici jsme převzali z článku Weiß a kol. (2019).

Definice 3 (Index disperze)**.** *Nechť funkce $f_d : H \to \mathbb{R}$, $(0, \infty)^2 \subseteq H$, jejíž předpis nezávisí na neznámých parametrech, splňuje*

(i) $f_d(\mu, \mu) = 1$ *pro* $\forall\ (\mu, \mu)^\mathsf{T} \in H$*;*

(ii) $f_d(\mu, \sigma^2) \neq 1$*, pokud* $(\mu, \sigma^2)^\mathsf{T} \in H$ *a* $\mu \neq \sigma^2$*.*

Nechť $\overline{X}_n$ a S_n^2 jsou definované jako výše. Potom statistika

$$f_d(\overline{X}_n, S_n^2),$$

je-li definovaná, se nazývá ***index disperze****.*

Platí-li (i) a (ii), potom opět pro $\forall\ (\mu, \sigma^2)^\mathsf{T} \in H$:

$$f_d(\mu, \sigma^2) = 1 \quad \Leftrightarrow \quad \mu = \sigma^2.$$

Použitím věty o spojité transformaci dostaneme, že pokud je $X_1, X_2, ..., X_n$ náhodný výběr z rozdělení s konečným druhým momentem a f_d je spojitá v bodě $(\mathsf{E}X_1, \mathsf{var}X_1)^\mathsf{T} \in H$, platí

$$f_d(\overline{X}_n, S_n^2) \xrightarrow[n \to \infty]{\mathsf{P}} f_d(\mathsf{E}X_1, \mathsf{var}X_1). \tag{3.2}$$

Dosazením $\mu = \mathsf{E}X_1$ a $\sigma^2 = \mathsf{var}X_1$ do definice indexu disperze potom díky (3.2) dostaneme vztah

$$f_d(\overline{X}_n, S_n^2) \xrightarrow[n \to \infty]{\mathsf{P}} 1 \quad \Leftrightarrow \quad \mathsf{E}X_1 = \mathsf{var}X_1.$$

Za nulové hypotézy a pro vysoká n tedy bude hodnota indexu disperze $f_d(\overline{X}_n, S_n^2)$ blízká jedné. Naopak v situaci, kdy je zmíněná hodnota významně vzdálená

od čísla 1, můžeme očekávat, že $\mathsf{E}X_1 \neq \mathsf{var}X_1$, a tedy uvažovaný náhodný výběr nepochází z Poissonova rozdělení.

Jednou z nejintuitivnějších možností je zvolit funkci f_d jako

$$\begin{aligned} f_d: \quad (0,\,\infty) \times \langle 0,\,\infty) \quad &\to \quad \langle 0,\,\infty), \\ (x, y) \quad &\mapsto \quad \frac{y}{x}. \end{aligned}$$

Snadno ověříme, že podmínky (i) a (ii) jsou splněny, tudíž

$$\hat{\mathrm{I}}_{d,n} := f_d(\overline{X}_n,\, S_n^2) = \frac{S_n^2}{\overline{X}_n}$$

je podle definice 3 index disperze. Náhodný výběr, jehož nosičem je podmnožina nezáporných celých čísel, navíc v případě $\overline{X}_n = 0$ má i nulový výběrový rozptyl. Použijeme-li konvenci $\frac{0}{0} := 1$, můžeme uvažovat index disperze $\hat{\mathrm{I}}_{d,n}$ i pro $\overline{X}_n = 0$. Z 3.2 máme, že pokud $\mathsf{E}X_1 > 0$, platí

$$\hat{\mathrm{I}}_{d,n} = \frac{S_n^2}{\overline{X}_n} \xrightarrow[n\,\to\,\infty]{\mathsf{P}} \frac{\mathsf{var}X_1}{\mathsf{E}X_1} \overset{(1.1)}{=} \frac{\lambda}{\lambda} = 1.$$

Dá se navíc ukázat (viz Semjonov (2020), Věta 7), že za nulové hypotézy (1.1)

$$\sqrt{n}\big(\hat{\mathrm{I}}_{d,n} - 1\big) \xrightarrow[n\,\to\,\infty]{\mathsf{D}} \mathsf{N}(0,\,2),$$

což je díky vlastnostem normálního rozdělení ekvivalentní s

$$\sqrt{\frac{n}{2}}\left(\hat{\mathrm{I}}_{d,n} - 1\right) \xrightarrow[n\,\to\,\infty]{\mathsf{D}} \mathsf{N}(0,\,1).$$

Analogicky jako v důkazu věty 9 dostaneme, že za (H_0) splňuje testová statistika $\mathrm{D}_n := \sqrt{\frac{n}{2}}\big(\hat{\mathrm{I}}_{d,n} - 1\big)$ rovnost

$$\lim_{n\to\infty} \mathsf{P}\big(|\mathrm{D}_n| \geq u_{1-\frac{\alpha}{2}}\big) = \alpha.$$

Z toho plyne, že test s rozhodovacím pravidlem

$$\text{zamítneme } (\mathsf{H}_0) \quad \Longleftrightarrow \quad |\mathrm{D}_n| \geq u_{1-\frac{\alpha}{2}}$$

má asymptotickou hladinu významnosti α, kde $\alpha \in (0,\,1)$ je nějaké předem zvolené číslo.

Podobně jako testy založené na nulovém indexu, je výše popsaný test rovněž asymptotický test, který testuje pouze jednu charakteristiku Poissonova rozdělení. Nastane-li situace, kdy $\mathsf{E}X_1 = \mathsf{var}X_1$, test dobré shody založený na indexu disperze obvykle potenciální neplatnou hypotézu nezamítne.

4. Simulační studie

V dosavadní části práce jsme teoreticky popsali řadu metod, pomocí nichž lze testovat dobrou shodu s Poissonovým rozdělením. Konkrétně se jednalo o testy založené na nulovém indexu, test využívající index disperze a χ^2-testy dobré shody. Nyní se pokusíme zmíněné způsoby testování porovnat pomocí dat nasimulovaných v programu R Core Team (2021). Všechny testy si budeme implementovat sami (nebudeme používat žádné dodatečné balíčky).

4.1 Hladina testů

Jako první budeme zkoumat, jak moc dané testy dodržují předepsanou hladinu významnosti. Bude nás tedy zajímat, v jaké míře bude v jednotlivých případech docházet k zamítání platné nulové hypotézy. Z předchozích kapitol víme, že všechny uvedené testy jsou pouze asymptotické. Můžeme tak očekávat, že zejména pro nízké hodnoty n (rozsahu náhodného výběru) stanovená hladina příliš dodržována nebude, pro různé testy se však míra jejího dodržování může lišit.

Náš postup bude nyní spočívat v tom, že si nejprve vygenerujeme realizace náhodného výběru $X_1, X_2, ..., X_n$ z rozdělení $\mathsf{Po}(\lambda)$, a následně provedeme několik dílčích testů dobré shody, přičemž si vždy zaznamenáme, pro které z nich jsme zamítli (v této situaci platnou) hypotézu (1.1). Tento proces 1000krát zopakujeme. Pro každý test tak dostaneme číslo N_i udávající počet případů, kdy došlo k zamítnutí (1.1). Pomocí něj spočteme odhad hladiny daného testu jako

$$\frac{N_i}{1000},$$

tj. poměr případů zamítnutí platné hypotézy ku celkovému počtu pokusů. Navíc rozlišíme několik různých hodnot parametru λ a rozsahu výběru n.

Mezi zkoumané testy budou patřit v první řadě testy založené na nulových indexech ze sekce 2.2 popsané ve větě 10, dále test s testovou statistikou D_n založený na indexu disperze (viz sekce 3.2) a χ^2-testy dobré shody ze sekce 3.1. U χ^2-testů budeme uvažovat případy pro 4 a 6 kategorií ($k = 4$ a $k = 6$), přičemž u obou zvolíme $r = 0$. Pro první volbu parametrů tak získáme statistiku

$$\chi^2_{1,n} := \sum_{i=1}^{4} \frac{\left(Y_i - np_i(\hat{\lambda})\right)^2}{np_i(\hat{\lambda})},$$

kde

$$Y_1 := \sum_{j=1}^{n} \mathbf{1}_{[X_j=0]}, \quad Y_2 := \sum_{j=1}^{n} \mathbf{1}_{[X_j=1]}, \quad Y_3 := \sum_{j=1}^{n} \mathbf{1}_{[X_j=2]}, \quad Y_4 := \sum_{j=1}^{n} \mathbf{1}_{[X_j\geq 3]},$$

a $\hat{\lambda}$ je odhad λ získaný řešením příslušné rovnice (viz sekce 3.1). Hypotézu (1.1) potom budeme zamítat právě tehdy, když $\chi^2_{1,n} \geq \chi^2_2(1-\alpha)$. Testovou statistiku příslušející volbě $k = 6$ označíme $\chi^2_{2,n}$. U testů, kde se v kritickém oboru vyskytuje kvantil normovaného normálního rozdělení (tj. testů založených na nulových

indexech a indexu disperze) navíc vyzkoušíme i variantu, kde tento kvantil nahradíme kvantilem t-rozdělení s $n-1$ stupni volnosti tak, jak bylo diskutováno v sekci 2.4.

Všechny zmíněné testy budeme vždy provádět na hladině $\alpha = 0{,}05$.

V následujících tabulkách jsou v jednotlivých sloupcích vypsány získané odhady hladiny pro test založený na dané testové statistice. Pokud je testová statistika označená hvězdičkou, provádíme test s kvantilem t-rozdělení namísto $\mathsf{N}(0,1)$ rozdělení. Pro přehlednost u těchto případů uvádíme pouze ty hodnoty, které se od těch původních liší o více než 0,005.

Tabulka 4.1: Odhady hladiny jednotlivých testů pro případ $\lambda = 1$

n	$\mathrm{T}_{1,n}$	$\mathrm{T}^*_{1,n}$	$\mathrm{T}_{2,n}$	$\mathrm{T}^*_{2,n}$	$\mathrm{T}_{3,n}$	$\mathrm{T}^*_{3,n}$	$\mathrm{T}_{4,n}$	$\mathrm{T}^*_{4,n}$
20	0,080	0,058	0,056	0,045	0,105	0,092	0,057	0,036
50	0,071	⋯	0,055	⋯	0,083	0,077	0,057	0,050
100	0,052	0,043	0,051	⋯	0,061	⋯	0,050	⋯
500	0,056	⋯	0,053	⋯	0,056	⋯	0,054	⋯
2 000	0,047	⋯	0,044	⋯	0,047	⋯	0,042	⋯

n	D_n	D^*_n	$\chi^2_{1,n}$	$\chi^2_{2,n}$
20	0,046	0,034	0,048	0,056
50	0,051	⋯	0,046	0,058
100	0,054	⋯	0,051	0,047
500	0,049	⋯	0,046	0,052
2 000	0,048	⋯	0,048	0,054

Tabulka 4.2: Odhady hladiny jednotlivých testů pro případ $\lambda = 2$

n	$\mathrm{T}_{1,n}$	$\mathrm{T}^*_{1,n}$	$\mathrm{T}_{2,n}$	$\mathrm{T}^*_{2,n}$	$\mathrm{T}_{3,n}$	$\mathrm{T}^*_{3,n}$	$\mathrm{T}_{4,n}$	$\mathrm{T}^*_{4,n}$
20	0,121	0,111	0,051	0,033	0,189	0,176	0,100	0,087
50	0,089	0,082	0,051	⋯	0,141	⋯	0,061	0,053
100	0,077	⋯	0,055	⋯	0,091	⋯	0,060	⋯
500	0,061	⋯	0,051	⋯	0,059	⋯	0,052	⋯
2 000	0,051	⋯	0,046	⋯	0,056	⋯	0,046	⋯

n	D_n	D^*_n	$\chi^2_{1,n}$	$\chi^2_{2,n}$
20	0,056	0,047	0,047	0,036
50	0,040	⋯	0,041	0,046
100	0,039	⋯	0,059	0,053
500	0,053	⋯	0,044	0,052
2 000	0,043	⋯	0,043	0,047

Z tabulek 4.1 a 4.2 můžeme vidět, že zatímco testy založené na indexu disperze a χ^2-testy dodržují stanovenou hladinu $\alpha = 0{,}05$ poměrně dobře i pro nízké rozsahy výběrů, u testů založených na nulových indexech (s výjimkou testů se statistikou $\mathrm{T}_{2,n}$) je pro malá n získaný odhad hladiny mnohdy významně vyšší. Podobný závěr si můžeme udělat i z tabulky 4.3 níže.

Tabulka 4.3: Odhady hladiny jednotlivých testů pro případ $\lambda = 3$

n	$\mathrm{T}_{1,n}$	$\mathrm{T}^*_{1,n}$	$\mathrm{T}_{2,n}$	$\mathrm{T}^*_{2,n}$	$\mathrm{T}_{3,n}$	$\mathrm{T}^*_{3,n}$	$\mathrm{T}_{4,n}$	$\mathrm{T}^*_{4,n}$
20	0,364	$\cdots$	0,027	0,019	0,371	$\cdots$	0,375	$\cdots$
50	0,112	$\cdots$	0,033	$\cdots$	0,213	$\cdots$	0,099	$\cdots$
100	0,095	$\cdots$	0,050	$\cdots$	0,151	0,145	0,077	$\cdots$
500	0,064	$\cdots$	0,060	$\cdots$	0,073	$\cdots$	0,063	$\cdots$
2 000	0,058	$\cdots$	0,051	$\cdots$	0,058	$\cdots$	0,052	$\cdots$

n	D_n	D^*_n	$\chi^2_{1,n}$	$\chi^2_{2,n}$
20	0,041	0,030	0,040	0,047
50	0,046	$\cdots$	0,045	0,048
100	0,052	$\cdots$	0,050	0,050
500	0,044	$\cdots$	0,064	0,052
2 000	0,041	$\cdots$	0,051	0,055

Lze si také všimnout, že u většiny zkoumaných testů založených na nulových indexech platí, že čím vyšší je parametr (skutečného) Poissonova rozdělení, tím více dat je potřeba k tomu, aby test dodržoval požadovanou hladinu. To může souviset s tím, že s rostoucím λ klesá poměr nulových pozorování a častěji dochází k případům, kdy jsou dokonce všechna data nenulová (viz sekce 2.4).

Z definice jednotlivých nulových indexů a následných testových statistik vidíme, že pokud $\hat{p}_0 = 0$, dostaneme bez ohledu na bližší podobu dat

$$\mathrm{T}_{1,n} = -\infty \quad \text{a} \quad \mathrm{T}_{3,n} = \infty,$$

tedy nulová hypotéza bude pro testy s danými testovými statistikami (a k nim náležející testy využívající kvantily t-rozdělení) zamítnuta. V situaci, kdy generujeme data z rozdělení $\mathsf{Po}(3)$ o rozsahu 20, se přitom přibližně ve třetině případů nevyskytne ani jedna nulová hodnota.

U testů založených na nulových indexech $\hat{\mathrm{I}}_{2,n}$ a $\hat{\mathrm{I}}_{4,n}$ bude v případě absence nulových pozorování platit $\hat{\mathrm{I}}_{2,n} = -1$ a $\hat{\mathrm{I}}_{4,n} = 1$, a tedy

$$\mathrm{T}_{2,n} = \frac{-1}{\sqrt{\frac{\mathrm{e}^{\overline{X}_n} - \overline{X}_n - 1}{n}}}, \qquad \mathrm{T}_{4,n} = \frac{1}{\sqrt{\frac{\mathrm{e}^{\overline{X}_n} - \overline{X}_n - 1}{n(\overline{X}_n)^2}}}.$$

Absolutní hodnoty daných testových statistik tak v této situaci závisejí pouze na velikosti odhadu asymptotického rozptylu. Toto může být jeden z důvodů, proč pro vyšší hodnoty parametru λ vycházejí některé odhady hladiny naopak výrazně nižší než stanovených 0,05.

Rovněž je vidět, že používání kvantilů Studentova t-rozdělení dává významně rozdílné výsledky pouze u nízkých rozsahů výběru (v našem případě $n = 20$). Při větším počtu dat je tak nahrazování kvantilu normovaného normálního rozdělení tímto kvantilem vcelku zbytečné.

Na základě získaných odhadů se jako nejlepší z hlediska dodržování hladiny jeví test založený na nulovém indexu $\hat{\mathrm{I}}_{2,n}$, test založený na indexu disperze se statistikou D_n a oba χ^2-testy.

4.2 Síla testů

Nyní nás bude zajímat síla zkoumaných testů proti určitým alternativám. Připomeneme, že síla testu při dané konkrétní alternativě je rovna pravděpodobnosti, se kterou zamítneme (v této situaci neplatnou) nulovou hypotézu, pochází-li naše data z rozdělení příslušného této alternativě.

Při odhadování síly uvažovaných testů budeme postupovat analogicky jako v předchozí části s jediným rozdílem – data budeme generovat z rozdělení různého od Poissonova.

Jako alternativu budeme uvažovat takzvané negativně binomické rozdělení (pro různé hodnoty parametrů). Pro připomenutí, pro $X \sim \mathsf{NB}(\mu, r)$, kde $\mu > 0$ a $r > 0$, platí

$$\mathsf{P}(X = k) = \binom{r+k-1}{k} \left(\frac{r}{r+\mu}\right)^r \left(\frac{\mu}{r+\mu}\right)^k, \quad k \in \mathbb{N}_0.$$

Zároveň

$$\mathsf{E}X = \mu \qquad \text{a} \qquad \mathsf{var}X = \mu + \frac{\mu^2}{r}.$$

Jedná se tedy o diskrétní rozdělení se stejným nosičem jako má Poissonovo rozdělení.

Poznámka. Toto je alternativní parametrizace negativně binomického rozdělení. Obvykle definujeme negativně binomické rozdělení pomocí definice uvedené např. v sekci 1.2.3 v knize Anděl (2007), tj. jako rozdělení s parametry $r > 0$ a $p \in (0, 1)$ splňující

$$\mathsf{P}(X = k) = \binom{r+k-1}{k} p^r (1-p)^k, \quad k \in \mathbb{N}_0.$$

Platí $\mathsf{E}X = \frac{r(1-p)}{p}$ a $\mathsf{var}X = \frac{r(1-p)}{p^2}$. Výše uvedený způsob parametrizace tak získáme, pokud položíme $\mu := \mathsf{E}X = \frac{r(1-p)}{p}$. Není těžké ověřit, že potom $p = \frac{r}{r+\mu}$.

Z definice negativně binomického rozdělení dostaneme, že pro náhodný výběr $X_1, X_2, ..., X_n$ pocházející z rozdělení $\mathsf{NB}(\mu, r)$ máme

$$\mathsf{P}(X_1 = 0) = \left(\frac{r}{r+\mu}\right)^r, \tag{4.1}$$

tedy obecně neplatí $p_0 = \mathrm{e}^{-\mathsf{E}X_1}$. Můžeme si ovšem všimnout, že

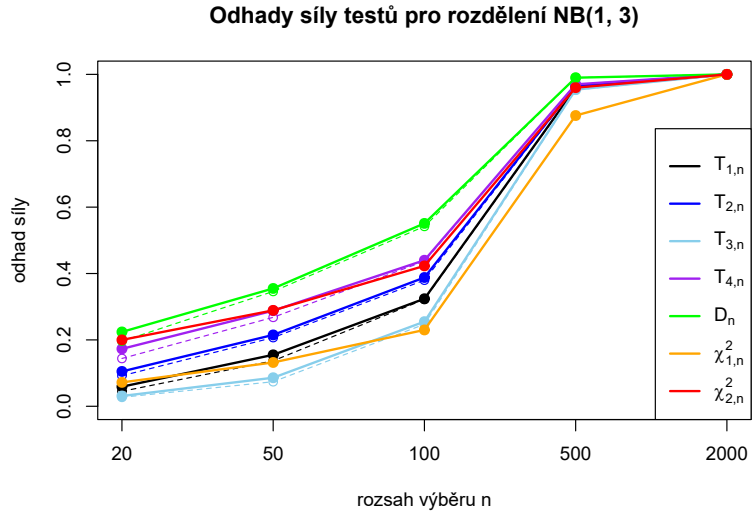
$$\lim_{r\to\infty} \mathsf{P}(X_1 = 0) = \mathrm{e}^{-\mu} = \mathrm{e}^{-\mathsf{E}X_1}.$$

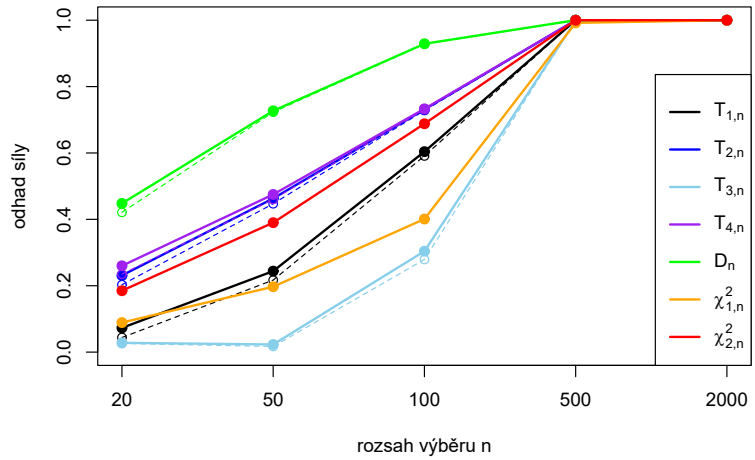
Je možné ukázat (viz Šír (2020), 2.4) i silnější vztah:

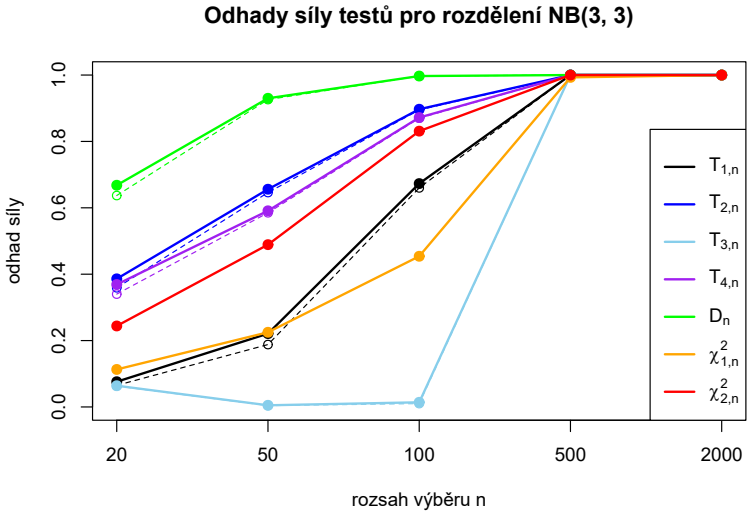
$$\lim_{r\to\infty} \mathsf{P}(X_1 = k) = \mathrm{e}^{-\mu} \cdot \frac{\mu^k}{k!}, \quad k \in \mathbb{N}_0,$$

což nám dává $\lim_{r\to\infty} \mathsf{P}(X_1 = k) = \mathsf{P}(Y = k)$ pro $Y \sim \mathsf{Po}(\mu)$. Platí rovněž $\lim_{r\to\infty} \mathsf{var}X_1 = \mu = \mathsf{E}X_1$.

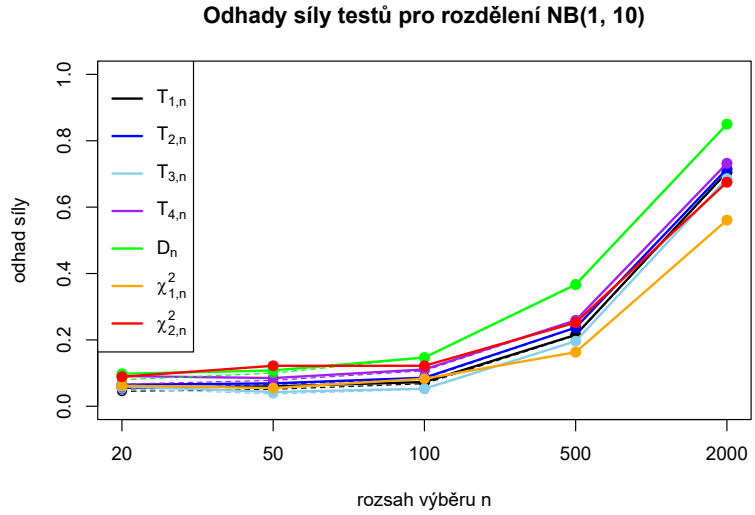
Můžeme tak očekávat, že pro testy dobré shody založené na nulovém indexu, χ^2-testy i test založený na indexu disperze bude s rostoucím parametrem r (a pevným μ) jejich síla vůči alternativě $\mathsf{NB}(\mu, r)$ klesat, neboť rozdíly mezi rozdělením $\mathsf{NB}(\mu, r)$ a Poissonovým rozdělením s parametrem μ se budou zmenšovat.

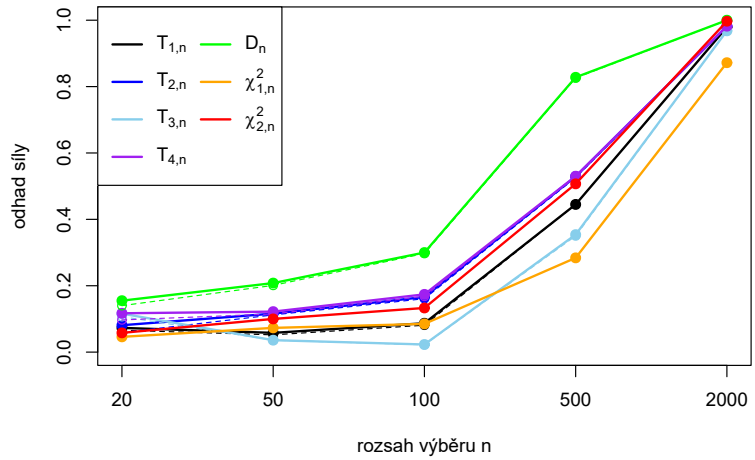


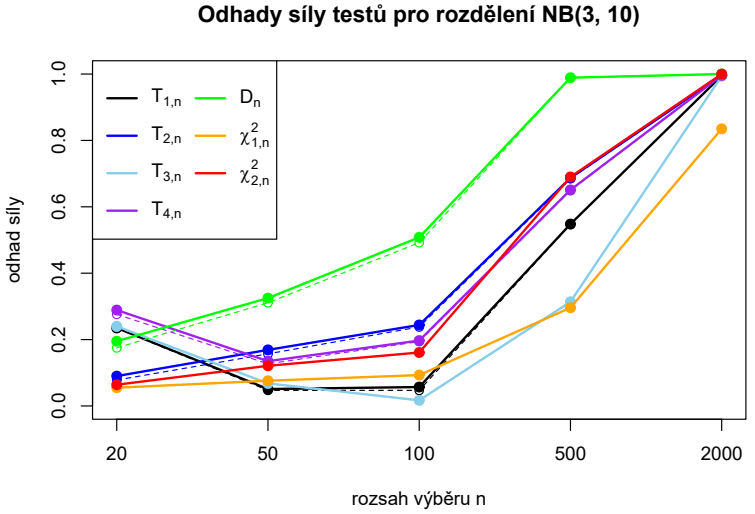




Obrázek 4.1: Odhady síly jednotlivých testů vůči alternativě $\mathsf{NB}(\mu, r)$ pro $r = 3$ a různé hodnoty μ (přerušované čáry značí odhady pro testy, kde jsou kvantily rozdělení $\mathsf{N}(0, 1)$ nahrazeny kvantily t-rozdělení)







Obrázek 4.2: Odhady síly jednotlivých testů vůči alternativě $\mathsf{NB}(\mu, r)$ pro $r = 10$ a různé hodnoty μ (přerušované čáry značí odhady pro testy, kde jsou kvantily rozdělení $\mathsf{N}(0, 1)$ nahrazeny kvantily t-rozdělení)

Na obrázcích 4.1 a 4.2 jsme mohli vidět, že u každého testu se odhad jeho síly proti určité konkrétní alternativě pro $n \to \infty$ blíží k 1. Testy tak můžeme považovat za konzistentní (vůči daným alternativám).

Porovnáním 4.1 a 4.2 si můžeme všimnout, že odhadnuté síly testů proti alternativě $\mathsf{NB}(\mu, 3)$ obvykle rostou při zvětšujícím se rozsahu výběru rychleji než odhadnuté síly proti alternativě $\mathsf{NB}(\mu, 10)$ (je-li μ v obou případech stejné). To odpovídá poznatkům diskutovaným v předchozí části, tedy že pro vyšší hodnotu r je rozdělení $\mathsf{NB}(\mu, r)$ „bližší“ Poissonovu rozdělení.

Trochu neobvyklá situace nastává u některých testů založených na nulových indexech, kde proti určitým alternativám (např. $\mathsf{NB}(3, 10)$) dochází při zvyšování rozsahu výběru k dočasnému poklesu jejich odhadnuté síly. Konkrétně to lze vidět především u testů s testovými statistikami $\mathrm{T}_{1,n}$ a $\mathrm{T}_{3,n}$. Jak ovšem víme z předchozí sekce, tyto testy mají pro nízký počet pozorování (v našem případě $n = 20$) a vyšší λ (střední hodnotu skutečného rozdělení) i velmi vysokou hladinu významnosti, která zřejmě souvisí s absencí nulových pozorování. Pomocí vztahu (4.1) můžeme nahlédnout, že pro vyšší hodnoty parametrů μ a r bude pravděpodobnost $\mathsf{P}(X_1 = 0)$, kde $X_1 \sim \mathsf{NB}(\mu, r)$, rovněž velmi malá.

Z dostupných odhadů síly jednotlivých testů proti alternativám $\mathsf{NB}(\mu, r)$ (pro dané hodnoty parametrů) lze jako nejsilnější test (vůči zmíněným alternativám) ze všech zkoumaných testů dobré shody považovat test založený na indexu disperze, tj. test s testovou statistikou D_n. Jako poměrně silné se jeví i testy založené na nulových indexech $\hat{\mathrm{I}}_{2,n}$ a $\hat{\mathrm{I}}_{4,n}$ a χ^2-test dobré shody rozlišující 6 kategorií (test s testovou statistikou $\chi^2_{2,n}$). Naopak nejhůře v tomto ohledu dopadl test, jehož testová statistika má tvar $\mathrm{T}_{3,n}$.

Závěr

K testování, zda rozdělení určitého náhodného výběru odpovídá Poissonovu rozdělení, existuje celá řada metod. Jedna z nich – testy založené na tzv. nulovém indexu – byla hlavním předmětem této práce.

Nulovým indexům a jejich využití při testování jsme věnovali celou druhou kapitolu. Nejprve jsme obecně zavedli Poissonův nulový index jakožto určitou funkci výběrového průměru a relativní četnosti nul v realizacích náhodného výběru. Poté jsme pomocí Δ-věty odvodili asymptotické rozdělení těchto indexů za platnosti hypotézy, že daný náhodný výběr pochází z Poissonova rozdělení, a využili jej ke konstrukci asymptotických testů dobré shody. Pro lepší představu jsme vše provedli i pro čtyři konkrétní příklady nulových indexů. Ve třetí kapitole jsme se podívali i na jiné způsoby testování dobré shody – variantu χ^2-testů pro Poissonovo rozdělení a testy založené na indexu disperze. Všechny zmíněné typy testů jsme potom porovnali pomocí simulací v kapitole 4.

Mezi hlavní výhody testů založených na nulových indexech patří jejich výpočetní jednoduchost (obzvlášť ve srovnání s χ^2-testy). Na druhou stranu však tyto testy nemají tak všestranné využití, neboť nejsou konzistentní vůči všem alternativám. Také se příliš nehodí používat v situacích, kdy testujeme dobrou shodu s rozdělením $\mathsf{Po}(\lambda)$ pro nějakou větší hodnotu parametru λ. V takovém případě je lepší volbou například test založený na indexu disperze se statistikou D_n či χ^2-testy dobré shody s vhodně zvolenými kategoriemi.

Ze všech zkoumaných testů lze na základě simulační studie doporučit výše zmíněný test s testovou statistikou D_n a test založený na nulovém indexu se statistikou $\mathrm{T}_{2,n}$, neboť tyto testy poměrně dobře dodržují stanovenou hladinu významnosti a jeví se i jako nejsilnější vůči zkoumaným alternativám (negativně binomickému rozdělení pro různé hodnoty parametrů). Naopak nejméně vhodný se zdá být test se statistikou $\mathrm{T}_{3,n}$ (založený na nulovém indexu), který má obzvlášť pro nižší rozsahy výběru (cca $n \leq 100$) velmi vysokou hladinu a zároveň malou sílu proti daným alternativám $\mathsf{NB}(\mu, r)$.

Seznam použité literatury

Anděl, J. (1978). *Matematická statistika*. SNTL/ALFA, Praha.

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

Dupač, V. a Hušková, M. (2013). *Pravděpodobnost a matematická statistika*. Druhé upravené vydání. Nakladatelství Karolinum, Praha. ISBN 978-80-246-2208-8.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Semjonov, V. (2020). Index disperze pro diskrétní rozdělení. Bakalářská práce, Matematicko-fyzikální fakulta, Univerzita Karlova.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge, United Kingdom.

Weiss, C. H., Homburg, A. a Puig, P. (2019). Testing for zero inflation and overdispersion in INAR(1) models. *Statist. Papers 60*, **26**, 823–848.

Šír, D. (2020). Dvourozměrné negativně binomické rozdělení. Bakalářská práce, Matematicko-fyzikální fakulta, Univerzita Karlova.