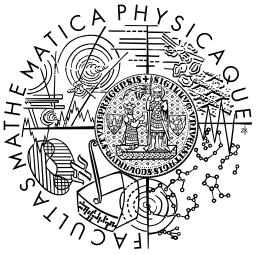


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Jan Schmidtmayer

Test shody s binomickým rozdělením založený na faktoriálních momentech

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Šárka Hudecová, Ph.D.
Studijní program: Finanční matematika
Studijní obor: MFMP

Praha 2024

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

Rád bych vyjádřil upřímné poděkování RNDr. Šárce Hudecové, Ph.D., za její cenný čas, odbornou znalost a rady, které mi pomohly s vypracováním této práce.

Název práce: Test shody s binomickým rozdělením založený na faktoriálních momentech

Autor: Jan Schmidtmayer

Katedra: Katedra pravděpodobnosti a matematické statistiky

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

Abstrakt: Bakalářská práce podrobně představuje test dobré shody s binomickým rozdělením založený na faktoriálních momentech. Nejdříve jsou v úvodu práce zavedeny základní pojmy a následně ve druhé kapitole postupně odvozeno asymptotické rozdělení testové statistiky. Diskutovány jsou také případy, kdy testová statistika není definovaná. Další částí práce je simulační studie, ve které je ukázáno chování popisovaného testu a jeho síla proti alternativám, ke kterým může běžně v praxi docházet. Všechny tyto vlastnosti jsou pro kontext porovnány s chí-kvadrát testem dobré shody. Závěr práce představuje ukázka využití tohoto testu na reálných datech z finančního prostředí.

Klíčová slova: test dobré shody, binomické rozdělení, faktoriální momenty

Title: Goodness-of-fit tests for binomial distributions based on factorial moments

Author: Jan Schmidtmayer

Department: Department of Probability and Mathematical Statistics

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

Abstract: The bachelor thesis deals with a goodness-of-fit test with a binomial distribution based on factorial moments. In the introductory section, the basic concepts are introduced, followed by the part, where the asymptotic distribution of test statistic is derived. In the next part of the thesis, we introduce a simulation study, showcasing the properties of the described test and its power against alternatives commonly encountered in practical scenarios. All these attributes are compared with those of the chi-squared test for goodness-of-fit. In the final part, the thesis presents an application of this test on real-world data, illustrating its utility in practical situations.

Keywords: goodness of fit test, binomial distribution, factorial moments

Obsah

Úvod

Binomické rozdělení přirozeně vzniká jako počet úspěchů v daném počtu nezávislých pokusů, z nichž každý skončí úspěchem se stejnou pravděpodobností. V praktických situacích ale není jisté, zda jsou jednotlivé pokusy opravdu nezávislé a zda úspěchy nastávají se stejnou pravděpodobností. Právě v takových situacích využíváme testy dobré shody s binomickým rozdělením, které ověřují, zda je binomický model pro naše data vhodný.

Tato práce se zaměří právě na jeden z typů takových testů, konkrétně na testy, které jsou založené na faktoriálních momentech. Nejdříve odvodíme a zadefinujeme potřebný aparát a následně testy podrobně představíme. V závěrečné fázi bude zpracována simulační studie a na úplném závěru bude představen příklad jejich využití na reálných datech.

Předchozí texty, které se zabývaly touto tématikou, zkoumaly v simulačních studiích sílu proti rozdělením, u kterých můžeme rozdíl poznat už například z explorativní analýzy. Tato práce se v simulační studii zaměří především na alternativy, kde rozdělení sice vzniká jako součet náhodných veličin s alternativním rozdělením, ale narozdíl od binomického rozdělení nejsou tyto náhodné veličiny nezávislé nebo stejně rozdělené. Se situacemi, kdy náhodné veličiny nejsou nezávislé, nejsou stejně rozdělené, anebo není ani jedno apriori jisté, se běžně setkáváme i v praxi. Výsledky simulací budou doplněné o obrázky a porovnány s testy v dnešní době běžně používanými.

Vlastní přínos této práce spočívá především v podrobném zpracování dané tématiky, která zahrnuje podrobné odvození testů nebo zhotovení příslušných důkazů, vypracování simulační studie doplněné o názorné obrázky a tabulky, a v neposlední řadě doplnění předešlých prací, dodefinování pro případy, kdy původní testovou statistiku nešlo použít, a opravu chyb, které se v předešlých pracích vyskytly.

1. Základní věty a definice

V úvodní kapitole si definujeme nástroje, které budou nezbytné pro testy dobré shody s binomickým rozdělením založené na faktoriálních momentech. Nejdůležitější pojmy vysvětlíme a případně ukážeme na příkladech nebo na obrázcích. Hlavními zdroji pro tuto kapitolu budou Anděl (2019) a Zvára a Štěpán (2019).

1.1 Základní definice

Na začátku vyjděme z alternativního rozdělení, ze kterého budeme v další sekci odvozovat rozdělení binomické. Tento postup bude v dalších částech práce nezbytný pro pochopení, co nastane, pokud budou porušeny některé podmínky vztahu mezi binomickým a alternativním rozdělením. Podrobněji o této věci referuje část 1.4.

Definice 1 (Alternativní (Bernoulliho) rozdělení)**.** *Nechť X je náhodná veličina, $p \in (0,1)$. Řekneme, že náhodná veličina X má alternativní rozdělení, pokud nabývá hodnot z $\mathbb{N}_0$ s pravděpodobností*

$$P(X = x) = \begin{cases} p^x(1-p)^{1-x} & \text{pro } x \in \{0,1\}, \\ 0 & \text{jinak.} \end{cases}$$

Značíme

$$X \sim Alt(p).$$

Náš test se od zbylých testů dobré shody s binomickým rozdělením liší tím, že je založen na faktoriálních momentech. Dále budeme uvažovat následující značení

$$X_{(k)} = X(X-1)(X-2)\dots(X-k+1). \tag{1.1}$$

Pro takové $X_{(k)}$ zaveďme střední hodnotu, kterou budeme nazývat k-tý faktoriální moment.

Definice 2 (Faktoriální k-tý moment)**.** *Nechť X je náhodná veličina, pro kterou platí $X \geq 0$ a $k \in \mathbb{N}$. Pak definujeme k-tý faktoriální moment náhodné veličiny X vztahem*

$$\mathsf{E}[X_{(k)}] = \mathsf{E}[X(X-1)(X-2)\dots(X-k+1)] = \mu_{(k)}.$$

Poslední definicí této sekce bude momentová vytvořující funkce.

Definice 3 (Momentová vytvořující funkce)**.** *Reálnou funkci reálné proměnné*

$$M_X(t) = \mathsf{E}[e^{tX}]$$

nazveme momentovou vytvořující funkcí náhodné veličiny X.

Momentová vytvořující funkce jednoznačně určuje rozdělení a také slouží jako funkce, jejímž derivováním můžeme dojít ke všem momentům. Tyto a mnohé další vlastnosti uvádí (Zvára a Štěpán, 2019, strana 112). Tvrzení, které ukazuje důvod jejího slovního označení, si uvedeme hned v další větě.

1.2 Základní věty

Věty uvedené v této sekci nám budou sloužit převážně jako opěrný bod netriviálních důkazů v dalších částech.

Věta 1. *Je-li momentová vytvořující funkce $M_X(t)$ náhodné veličiny X taková, že $M(b) < \infty, M(-b) < \infty$ pro některé $b > 0$ a platí $\mathsf{E}\,|X|^z < \infty$ pro $z \geq 1$, pak*

$$\mathsf{E}[X^z] = \left[\frac{d^z M_X(t)}{dt^z}\right]_{t=0}.$$

Důkaz. Důkaz lze najít ve skriptech (Dupač a a Hušková, 2001, strana 25). □

V pozdější fázi odvodíme z této věty vyšší momenty binomického rozdělení. Další věty, které uvedeme, už budou jedny ze základních vět v matematické statistice, které použijeme k odvození asymptotického rozdělení naší testové statistiky. První takovou větou je věta o Δ-metodě.

Věta 2 (Δ-metoda)**.** *Nechť $\{\boldsymbol{Y}_n\}_{n=1}^{\infty}$ splňuje*

$$\sqrt{n}(\boldsymbol{Y}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\mathbf{0}, \boldsymbol{\Sigma})$$

pro nějaký vektor konstant $\boldsymbol{\mu} \in \mathbb{R}^k$ a matici $\boldsymbol{\Sigma}$. Nechť g je spojitě diferencovatelná funkce $\mathbb{R}^k \xrightarrow{D} \mathbb{R}^p$ Označme $D(x) = \frac{\partial g(\boldsymbol{x})}{\partial \boldsymbol{x}}$. Pak platí

$$\sqrt{n}(g(\mathbf{Y}_n) - g(\boldsymbol{\mu})) \xrightarrow{D} N_p(\mathbf{0}, D(\boldsymbol{\mu})\boldsymbol{\Sigma}D(\boldsymbol{\mu})^{\top}).$$

Důkaz. Důkaz lze najít v knize (Van der Vaart, 2000, strana 25). □

Druhou podobnou větou, kterou si uvedeme, je Cramérova-Sluckého věta.

Věta 3 (Cramérova-Sluckého věta)**.** *Nechť $X_n \xrightarrow{D} X$, $A_n \xrightarrow{P} a$ a $B_n \xrightarrow{P} b$, kde X_n, X, A_n, B_n jsou náhodné veličiny a a,b jsou konstanty. Pak platí*

$$A_n X_n + B_n \xrightarrow{D} aX + b.$$

Důkaz. Důkaz můžeme dohledat v knize (Serfling, 1980, strana 19). □

1.3 Binomické rozdělení

Binomické rozdělení $\mathsf{Bi}(m, p)$ je diskrétní rozdělení se dvěma parametry, kde parametr m udává počet náhodných pokusů. Speciálně pro $m = 1$ se jedná o alternativní rozdělení. Parametr p určuje pravděpodobnost dichotomického náhodného jevu. Formálně je definováno následujícím způsobem.

Definice 4 (Binomické rozdělení)**.** *Nechť X je náhodná veličina, $p \in (0,1), m \in \mathbb{N}$. Řekneme, že náhodná veličina X má binomické rozdělení, pokud nabývá hodnot z $\mathbb{N}_0$ s pravděpodobností*

$$\mathsf{P}(X = x) = \begin{cases} \binom{m}{x} p^x (1-p)^{m-x} & \text{pro } x \in \{0,1,\ldots,m\}, \\ 0 & \text{jinak,} \end{cases}$$

kde $\binom{m}{x}$ je kombinační číslo. Značíme

$$X \sim \mathsf{Bi}(m,p).$$

Poznámka. Pro binomické rozdělení z binomické věty platí

$$\sum_{x=0}^{m} P(X = x) = 1. \tag{1.2}$$

Binomické rozdělení přirozeně vzniká jako počet úspěchů v daném počtu nezávislých pokusů, z nichž každý skončí úspěchem se stejnou pravděpodobností. Jinými slovy tedy vzniká jako součet náhodných veličin s alternativním rozdělením, které jsou nezávislé a stejně rozdělené. Právě tento vztah mezi alternativním a binomickým rozdělením formálně odvoďme.

Věta 4 (Vztah mezi alternativním a binomickým rozdělením)**.** *Nechť $m \in \mathbb{N}$ a $Y_1, Y_2, \ldots, Y_m$ jsou nezávislé stejně rozdělené náhodné veličiny s alternativním rozdělením s parametrem $p \in (0,1)$. Potom náhodná veličina $X = \sum_{i=1}^{m} Y_i$ má binomické rozdělení s parametry m a p.*

Důkaz. Nejdříve spočítejme momentovou vytvořující funkci alternativního rozdělení

$$M_Y(t) = \mathsf{E}[\exp(tY)] = e^{t\cdot 1} \cdot \mathsf{P}[Y = 1] + e^{t\cdot 0} \cdot \mathsf{P}[Y = 0] = (1-p) + pe^t.$$

Také si připomeňme, že součet m nezávislých a stejně rozdělených náhodných veličin má momentovou vytvořující funkci rovnu m-té mocnině momentové vytvořující funkce jednotlivé veličiny, tedy

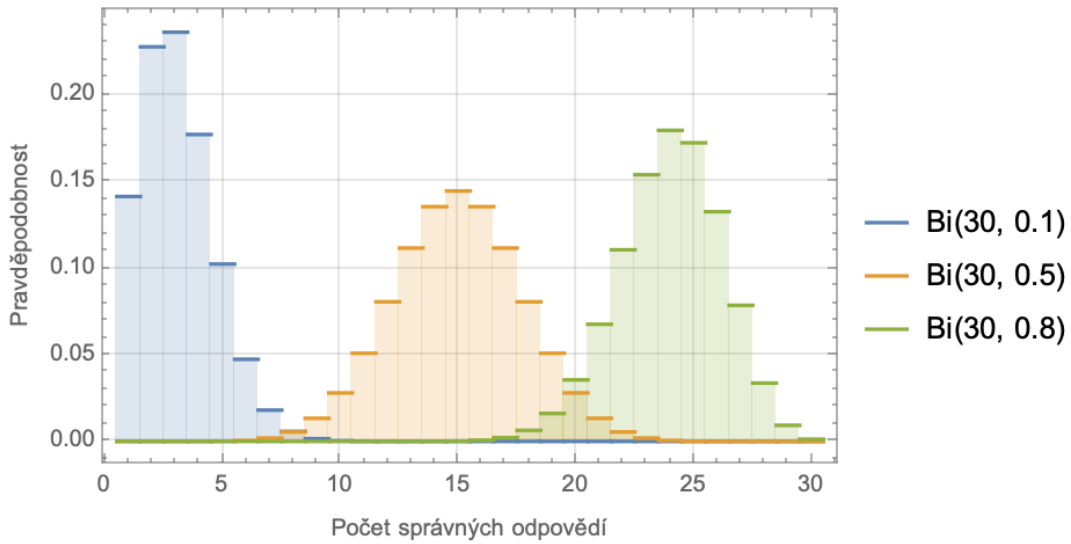
$$M_X(t) = \mathsf{E}[\exp(tX)] = \mathsf{E}\left[\exp\left(t\sum_{i=1}^{m} Y_i\right)\right] = \prod_{i=1}^{m} M_{Y_i}.$$

Ve třetí rovnosti jsme využili nezávislosti náhodných veličin. Nyní vypočtěme momentovou vytvořující funkci binomického rozdělení

$$M_X(t) = \sum_{x=0}^{m} \binom{m}{x} (pe^t)^x (1-p)^{m-x} = ((1-p) + pe^t)^m.$$

Ve druhé rovnosti jsme použili binomickou větu. Vidíme, že obě momentové vytvořující funkce jsou stejné, a totéž tedy platí pro obě rozdělení, což vyplývá z vlastností uvedených pod definicí 3.

□



Obrázek 1.1: Pravděpodobnostní funkce binomického rozdělení se stejným parametrem m, ale různými parametry p.

Ukažme si binomické rozdělení na konkrétním příkladě. Řešíme problém různě naučených studentů na test s 30 vzájemně nezávislými odpověďmi na dané otázky. První ze studentů má u každé otázky pravděpodobnost správné odpovědí 0.8, druhý má pravděpodobnost 0.5 a třetí 0.1. Správnost odpovědi u každé otázky reprezentuje jednu náhodnou veličinu s alternativním rozdělením. Dohromady můžeme vidět, jaká je pravděpodobnost jednotlivých počtů správně zodpovězených otázek na obrázku 1.1.

Některé konkrétní vlastnosti binomického rozdělení budeme potřebovat znát v dalších částech práce, a proto si je nyní odvoďme.

Věta 5 (Vlastnosti binomického rozdělení)**.** *Nechť* $X \sim \mathsf{Bi}(m, p)$. *Poté platí*

i) $\mathsf{E}(X) = mp$,

ii) $var(X) = mp(1-p)$,

iii) $M_X(t) = [(1-p) + p \cdot e^t]^m$,

iv) $\mathsf{E}(X - \mathsf{E}[X])^3 = mp(1-p)(1-2p)$,

v) $\mathsf{E}(X - \mathsf{E}[X])^4 = mp(1-p) + 3(mp)^2(1-p)^2 - 6p^2[m(1-p)^2]$.

Důkaz.
i)+ii) Nejprve si připomeňme, že pro náhodnou veličinu Y, která má alternativní rozdělení s paramterem p, platí $\mathsf{E}(Y) = p$ a $\mathrm{var}(Y) = p(1-p)$. Využijme vztahu mezi alternativním a binomickým rozdělením, který jsme si odvodili v rámci věty 4, kde binomické rozdělení vzniká jako součet m alternativních. Nejprve pro střední hodnotu z linearity dostáváme

$$\mathsf{E}[X] = \mathsf{E}\left[\sum_{i=1}^{m} Y_i\right] = \sum_{i=1}^{m} \mathsf{E}[Y_i] = \sum_{i=1}^{m} p = mp.$$

Podobně pro rozptyl

$$\mathrm{var}[X] = \mathrm{var}\left[\sum_{i=1}^{m} Y_i\right] = \sum_{i=1}^{m} \mathrm{var}[Y_i] = \sum_{i=1}^{m} p(1-p) = mp(1-p).$$

Ve druhé rovnosti jsme využili nezávislosti Y_i.

iii) Tento vztah jsme odvodili v rámci věty 4.
iv) Nejdříve spočítáme třetí necentrální moment. Za pomoci věty 1

$$\begin{aligned}\frac{d^3 M_X(t)}{dt^3} &= e^t mp\left[1 + (-1+e^t)p\right]^{m-3} \\ &\quad \times \left\{1 + p\left[p - 2 + e^t\left(p - 1 + 3m + (e^t m - 3)mp\right)\right]\right\}.\end{aligned}$$

Pro $t = 0$ dostáváme

$$\mathsf{E}[X^3] = mp\left[1 + (m-1)(3 + mp - 2p)p\right].$$

Nyní už spočítejme třetí centrální moment binomického rozdělení. Nejprve z definice třetího centrálního momentu platí, že

$$\mathsf{E}\left[X - \mathsf{E}X\right]^3 = \mathsf{E}\left[X^3\right] - 3\mathsf{E}\left[X^2\right]\mathsf{E}[X] + 2\mathsf{E}[X]^3$$

a po dosazení a úpravě výrazu máme

$$\mathsf{E}\left[X - \mathsf{E}X\right]^3 = mp(1-p)(1-2p).$$

v) Stejným postupem spočítáme i čtvrtý centrální moment binomického rozdělení

$$\begin{aligned}&\frac{d^4 M_X(t)}{dt^4} = e^t mp\left[1 + (-1+e^t)p\right]^{m-4} \\ &\times \left[e^t(7m-4)(p-1)^2 p - (p-1)^3 - e^{2t}(1 - 4m + 6m^2)(p-1)p^2 + e^{3t}m^3 p^3\right].\end{aligned}$$

Pro $t = 0$ dostáváme

$$\mathsf{E}\left[X^4\right] = mp(1-p)\left[6p(p-1) + 1\right].$$

Z definice čtvrtého centrálního momentu a po úpravě výrazu máme

$$\begin{aligned}\mathsf{E}\left[X - \mathsf{E}X\right]^4 &= \mathsf{E}\left[X^4\right] - 4\mathsf{E}\left[X^3\right]\mathsf{E}[X] + 6\mathsf{E}\left[X^2\right]\mathsf{E}[X]^2 - 3\mathsf{E}[X]^4 \\ &= mp(1-p) + 3(mp)^2(1-p)^2 - 6p^2\left[m(1-p)^2\right].\end{aligned}$$

□

Poslední vlastnost binomického rozdělení, kterou v této části uvedeme, jsou jeho faktoriální momenty. Jak už je z názvu naší práce zřejmé, právě na nich bude založena naše testová statistika.

Věta 6 (Faktoriální k-tý moment binomického rozdělení)**.** *Nechť náhodná veličina* $X \sim \mathsf{Bi}(m, p)$*. Potom pro k-tý faktoriální moment X platí*

$$\mathsf{E}\left[X_{(k)}\right] = \frac{m!}{(m-k)!}p^k = m_{(k)}p^k,$$

kde $m_{(k)} = \frac{m!}{(m-k)!}$.

Důkaz. Počítejme z definice:

$$\mathsf{E}[X(X-1)(X-2)\dots(X-k+1)] = \mathsf{E}\left[\frac{X!}{(X-k)!}\right] = \sum_{x=k}^{m}\frac{x!}{(x-k)!}p(x),$$

kde $p(x) = \binom{m}{m-x}p^x(1-p)^{m-x}$. Druhá rovnost plyne z definice střední hodnoty pro diskrétní rozdělení. Postupnými úpravami dostáváme

$$\begin{aligned}
&\sum_{x=k}^{m}\frac{x!m!}{x!(x-k)!(m-x)!}p^x(1-p)^{m-x} \\
&\quad= \sum_{x=k}^{m}\frac{m!}{(x-k)!(m-x)!}p^x(1-p)^{m-x} \\
&\quad= \frac{m!}{(m-k)!}p^k\sum_{x=k}^{m}\frac{(m-k)!}{(x-k)!(m-x)!}p^{x-k}(1-p)^{m-x}.
\end{aligned}$$

Nyní $y := x - k$

$$\frac{m!}{(m-k)!}p^k\sum_{y=0}^{m-k}\binom{m-k}{y}p^y(1-p)^{m-k-y}.$$

V tomto kroku si všimněme, že s použitím binomické věty máme $(p+1-p)^{m-k} = 1$ a dostáváme

$$\mathsf{E}\left[X_{(k)}\right] = \frac{m!}{(m-k)!}p^k.$$

□

1.4 Jiné modely součtu alternativních rozdělení

Binomické rozdělení není jediným rozdělením, které vzniká součtem alternativních rozdělení. V této kapitole se budeme zabývat rozšířeným modelem, ve kterém apriori nepředpokládáme žádný vztah mezi náhodnými veličinami s alternativním rozdělením.

Nechť $Y_1, \dots, Y_m$ jsou náhodně veličiny, pro které platí $Y_i \sim \mathsf{Alt}(p_i)$. Uvažujme náhodnou veličinu

$$X = \sum_{i=1}^{m} Y_i.$$

Taková náhodná veličina X má diskrétní rozdělení na $0, 1, \dots, m$, které může být binomickým rozdělením, směsí, či nějakým jiným obecným rozdělením. Podívejme

se na základní charakteristiky náhodné veličiny X. Střední hodnota je linearní operátor, a platí tedy

$$\mathsf{E}(X) = \mathsf{E}\left[\sum_{i=1}^{m} Y_i\right] = \mathsf{E}[Y_1] + E[Y_2] + \ldots + \mathsf{E}[Y_m]. \tag{1.3}$$

U rozptylu je už situace komplikovanější

$$\mathrm{var}(X) = \mathrm{var}\left[\sum_{i=1}^{m} Y_i\right] = \sum_{i=1}^{m} \mathrm{var}[Y_i] + \sum_{i=1,\, j=1, i\neq j}^{m}\sum^{m} \mathrm{cov}(Y_i, Y_j).$$

Jak bude vypadat náhodná veličina X, pokud přidáme další předpoklady ke vztahu mezi jednotlivými náhodnými veličinami Y_i?

1.4.1 Stejně rozdělené závislé náhodné veličiny

Nejprve se podívejme, co se stane, pokud přidáme předpoklad stejného rozdělení náhodných veličin (tj. $p_i = p$, pro všechna $i = 1, \ldots, m$). Pro tento model se obecně nebude jednat o binomické rozdělení. Některé charakteristiky ovšem budou podobné či dokonce stejné. Hodnota prvního centrálního momentu bude z linearity stejná jako při nezávislosti, což je zřejmé i z rovnosti (1.3). Druhý centrální moment už bude odlišný a platí následující vztah

$$\mathrm{var}(X) = \sum_{i=1}^{m} \mathrm{var}(Y_i) + \sum_{i=1,\, j=1, i\neq j}^{m}\sum^{m} \mathrm{cov}(Y_i, Y_j) = mp(1-p) + \sum_{i=1,\, j=1, i\neq j}^{m}\sum^{m} \mathrm{cov}(Y_i, Y_j).$$

Další charakteristiky se budou lišit na základě toho, jak na sobě budou náhodné veličiny závislé. Podrobněji o této problematice referuje například Vellaisamy a Punnen (2001) a Van Der Geest (2005).

1.4.2 Různě rozdělené nezávislé náhodné veličiny

Druhou alternativou je k situaci z úvodu této sekce přidat předpoklad nezávislosti náhodných veličin. Náhodné veličiny tedy budou různě rozdělené. Pod pojmem různě rozdělené náhodné veličiny budeme v dalších částech textu rozumět takové náhodné veličiny, které nesplňují předpoklad stejného rozdělení (tj. existuje alespoň jedna dvojice s nestejným rozdělením). Součet různě rozdělených nezávislých náhodných veličin lze brát jako zobecnění klasického binomického rozdělení. Jde o součet m veličin s alternativním rozdělením, které však všechny nemají stejnou pravděpodobnost úspěchu.

Definice 5 (Poissonovo binomické rozdělení)**.** *Nechť $i, j \in (1, \ldots, m)$,*

$$X = \sum_{i=1}^{m} Y_i,$$

kde $Y_i \sim Alt(p_i)$ a zároveň $Y_i \perp\!\!\!\perp Y_j$ pro všechny $i \neq j$. Pak řekneme, že X má Poissonovo binomické rozdělení a značíme $X \sim \mathsf{PBi}(m, \boldsymbol{p})$.

Poissonovo binomické rozdělení si jednoznačně určíme momentovou vytvořující funkcí.

Věta 7 (Momentová vytvořující funkce Poissonova binomického rozdělení)**.** *Nechť* $X \sim \mathsf{PBi}(m, \boldsymbol{p})$*, pak*

$$M_X(t) = \prod_{i=1}^{m}(1 - p_i + p_i e^t), \quad t \in \mathbb{R}$$

je momentovou vytvořující funkcí Poissonova binomického rozdělení.

Důkaz. Nejdříve použijeme definici momentové vytvořující funkce a následně definici Poissonova binomického rozdělení

$$M_X(t) = \mathsf{E}\left[\exp(tX)\right] = E\left[\exp\left(\sum_{i=1}^{n} tY_i\right)\right] = \mathsf{E}\left[\prod_{i=1}^{n}\exp(tY_i)\right].$$

Nyní můžeme využít nezávislosti náhodných veličin a také připomeňme rovnost z důkazu věty 4, kde $\mathsf{E}[\exp(tY)] = 1 - p + pe^t$, a tedy

$$\prod_{i=1}^{n}\mathsf{E}[\exp(tY_i)] = \prod_{i=1}^{m}(1 - p_i + p_i e^t).$$

□

Věta 8. *Nechť* $X \sim \mathsf{PBi}(m,\boldsymbol{p})$ *a* A *je množina přirozených čísel* $\{1, 2, 3, \ldots, m\}$*. Potom*

$$P(X = x) = \sum_{A \in \mathcal{B}_m}\left(\prod_{i \in A} p_i\right)\prod_{i \in A^c}(1 - p_i),$$

kde $\mathcal{B}_m$ *je množina všech podmnožin s* x *prvky, které můžeme vybrat z* $\{1, \ldots, m\}$*,* x *je tedy mohutností podmnožin a celková mohutnost množiny* $\mathcal{B}_m$ *je* $\binom{n}{x}$*.*

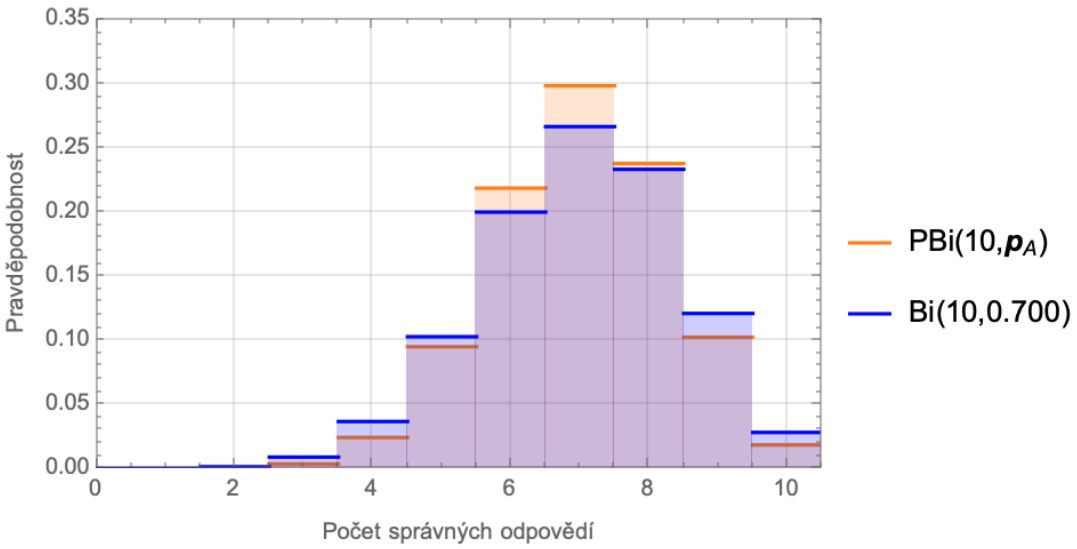
Důkaz. Odvození lze najít například v článku Wang (1993).

□

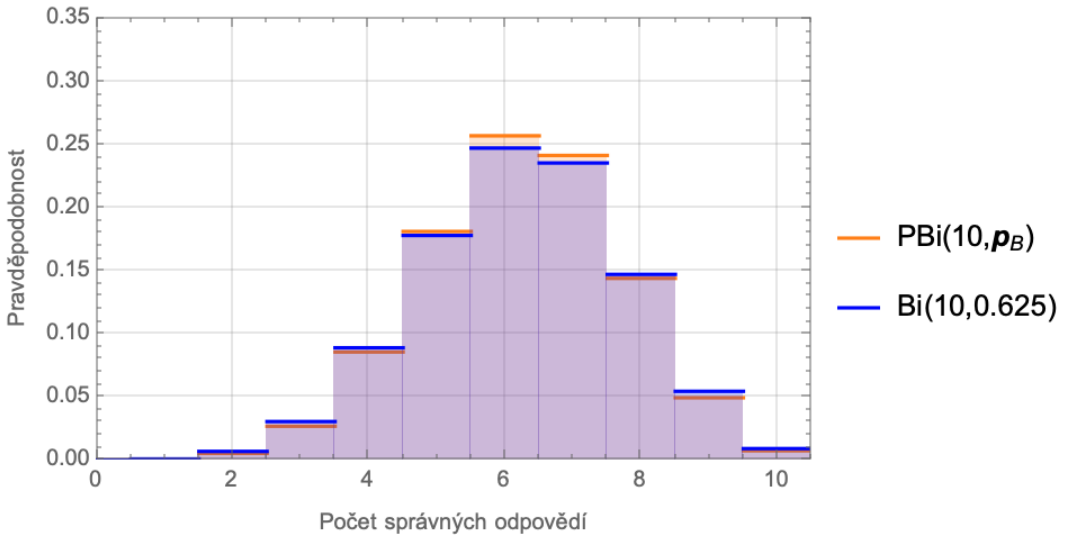
Výpočet pravděpodobnostní nebo distribuční funkce z této věty je početně nepraktický, a proto se používají různé alternativní metody výpočtu, které jsou popsány například v článku Hong (2013) nebo také v práci Tang a Tang (2023).

Ukažme si podobně jako u binomického rozdělení konkrétní příklad. Tentokrát uvažujme test, který sestává ze dvou tématických bloků, kde oba mají po pěti otázkách. Student se více naučil na druhý tématický blok, ve kterém má u každé otázky pravděpodobnost správné odpovědi 0.90. Na první tématický blok mu už nezbylo tolik času, a tak má pravděpodobnost správné odpovědi jen 0.5. Celkově tedy máme $\boldsymbol{p} = (0.50, 0.50, 0.50, 0.50, 0.50, 0.90, 0.90, 0.90, 0.90, 0.90)^\top \equiv \boldsymbol{p}_A, m = 10$. Student odpovídá na jednotlivé otázky nezávisle na ostatních, ale protože má test dva tématické bloky, tak všechny odpovědi nejsou stejně rozdělené. Na obrázku 1.2 vidíme porovnáni s rozdělením $\mathsf{Bi}(10, 0.700)$, které má stejnou střední hodnotu.

Pokud budeme rozdíl mezi jednotlivými bloky snižovat, budou si obě rozdělení stále podobnější. Například na obrázku 1.3 vidíme Poissonovo binomické rozdělení s paramtery $\boldsymbol{p} = (0.50, 0.50, 0.50, 0.50, 0.50, 0.75, 0.75, 0.75, 0.75, 0.75)^\top \equiv \boldsymbol{p}_B$, $m = 10$ a $\mathsf{Bi}(10, 0.625)$. Obě rozdělení mají opět stejné střední hodnoty, ale rozdíl v pravděpodobnostních funkcích je minimální.



Obrázek 1.2: Porovnání $\mathsf{PBi}(10, \boldsymbol{p}_A)$ a $\mathsf{Bi}(10, 0.700)$.



Obrázek 1.3: Porovnání $\mathsf{PBi}(10, \boldsymbol{p}_B)$ a $\mathsf{Bi}(10, 0.625)$.

2. Testy dobré shody

Tato kapitola podrobně popíše testy dobré shody s binomickým rozdělením založené na faktoriálních momentech. V úvodní kapitole jsme si ukázali všechny nezbytné nástroje, ze kterých si test odvodíme. Také jsme si podrobně představili binomické rozdělení a faktoriální momenty, tedy už z názvu signifikantní věci pro náš test. Nyní se můžeme posunout k testování dobré shody.

Testování dobré shody pro nás bude znamenat, že dostaneme náhodný výběr $X_1, \ldots, X_n$ a naším cílem bude zjistit, zda-li pochází z modelu $\mathcal{F}$. Konrétně budeme mít náhodný výběr z diskrétního rozdělení s distribuční funkcí F a rozsahem výběru n. Budeme uvažovat následující hypotézy

$$H_0 : F \in \mathcal{F}_m,$$
$$H_1 : F \notin \mathcal{F}_m,$$

kde $\mathcal{F}_m = \{\mathsf{Bi}(m, p), p \in (0,1)\}$.

Parametr m je nenulové přirozené číslo (budeme uvažovat $m > 1$), které zpravidla při testování známe, a proto není součástí našeho modelu $\mathcal{F}_m$. Parametr p není v nulové hypotéze přesně specifikován, a z tohoto důvodu se jedná o hypotézu složenou. Kdyby p bylo součástí specifikace H_0, jednalo by se o hypotézu jednoduchou, ale tímto případem se zabývat nebudeme.

Testováním dobré shody s binomickým rozdělením se zabýval článek Kyriakoussis a kol. (1998) a následně na něj navázal článek Aleksandrov a kol. (2022), který původní test zobecnil. V této kapitole si oba tyto testy odvodíme, podrobně ukážeme, jak fungují, a doplníme chybějící důkazy.

Předtím si ještě definujme základní statistiky, abychom se vyhnuli případným nejasnostem.

Definice 6 (Výběrový průměr a výběrový rozptyl)**.** *Nechť $X_1, \ldots, X_n$ je náhodný výběr.*

i) Veličina $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ se nazývá výběrový průměr náhodného výběru $\boldsymbol{X} = (X_1, \ldots, X_n)$.

ii) Veličina $S_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2$ se nazývá výběrový rozptyl náhodného výběru $\boldsymbol{X} = (X_1, \ldots, X_n)$.

Naše definice S_n^2 se mírně liší od standardní definice o přenásobení faktorem $\frac{n-1}{n}$, aby platil vztah s druhým empirickým faktoriálním momentem, který bude ukázán v rámci rovnosti (2.2). Nejdřívě definujme obecně výběrový k-tý faktoriální moment a následně si tento vztah odvoďme.

Definice 7 (Výběrový k-tý faktoriální moment)**.** *Nechť $n \in \mathbb{N}$ a $X_1, \ldots, X_n$ je náhodný výběr z diskrétního rozdělení. Výběrový k-tý faktoriální moment definujeme vztahem*

$$\widehat{\mu}_{(k)} = \frac{1}{n}\sum_{i=1}^{n} (X_i)_{(k)} = \frac{1}{n}\sum_{i=1}^{n} X_i(X_i - 1)\ldots(X_i - k + 1).$$

Ukažme si vztah mezi základními statistikami a hodnotami prvních dvou výběrových faktoriálních momentů.

Příklad. Pro první faktoriální moment máme

$$\widehat{\mu}_{(1)} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n \tag{2.1}$$

a podobně pro druhý faktoriální moment platí

$$\begin{aligned}\widehat{\mu}_{(2)} + \overline{X}_n - \overline{X}_n^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i(X_i-1) + \overline{X}_n - \overline{X}_n^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n + \overline{X}_n - \overline{X}_n^2 = S_n^2. \end{aligned} \tag{2.2}$$

2.1 Test dobré shody založený na prvním a druhém faktoriálním momentu

Prvním testem, který budeme uvažovat, je test navržený ve článku Kyriakoussis a kol. (1998). Všimněme si, že pokud zvolíme vhodný podíl faktoriálních momentů, dojde ke zkrácení parametru p, a tento paramter nám z rovnosti úplně vypadne. Můžeme tedy operovat pouze s paramtrem m, který je v námi uvažovaném modelu známý. Uvažujme nejjednodušší takový podíl

$$\frac{\mu_{(2)}}{\mu_{(1)}^2} = \frac{m_{(2)}p^2}{m_{(1)}^2 p^2} = \frac{m-1}{m}$$

a emprický protějšek $\frac{\mu_{(2)}}{\mu_{(1)}^2}$ je

$$T_n \equiv \frac{\widehat{\mu}_{(2)}}{\widehat{\mu}_{(1)}^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i(X_i-1)}{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2}.$$

Věta 9. *Nechť $X_1,\ldots,X_n$ je náhodný výběr z $\mathsf{Bi}(m,p)$, $p \in (0,1)$ a $m > 1$, pak*

$$T_n \xrightarrow{P} \frac{m-1}{m} \quad \textit{pro} \quad n \to \infty.$$

Důkaz. Počítejme

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i(X_i-1)}{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\sum_{i=1}^{n} X_i}{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2}.$$

Víme, že X_i jsou stejně rozdělené a nezávislé náhodné veličiny s konečným k-tým momentem, a tak ze zákona velkých čísel platí

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mathsf{E}[X] \quad \text{a} \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{P} \mathsf{E}[X^2] \quad \text{pro} \quad n \to \infty$$

a zároveň $\frac{1}{x^2}$ je spojitá transformace, takže dle věty o spojité transformaci platí

$$\frac{1}{\overline{X}_n^2} \xrightarrow{P} \frac{1}{(\mathsf{E}[X])^2} \quad \text{pro} \quad n \to \infty.$$

Dostáváme

$$\frac{E[X^2]-E[X]}{(E[X])^2} = \frac{\mathrm{var}[X]+(\mathsf{E}[X])^2-E[X]}{(E[X])^2} = \frac{mp(1-p)+(mp)^2-mp}{(mp^2)},$$

kde jsme v první rovnosti využili vztahu $\mathsf{E}[X^2] = \mathrm{var}[X] + (\mathsf{E}[X])^2$, který plyne z definice rozptylu a následně dosadili příslušné momenty binomického rozdělení. Ve výsledku po sérii algebraických úprav dostáváme

$$\frac{\frac{1}{n}\sum_{i=1}^n X_i(X_i-1)}{\left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2} \xrightarrow{P} \frac{m-1}{m}.$$

□

Abychom mohli provést test hypotézy H_0 pomocí T_n, potřebujeme odvodit její asymptotické rozdělení. K tomu nám bude sloužit následující pomocné tvrzení.

Věta 10. *Nechť $X_1,\ldots,X_n$ je náhodný výběr se střední hodnotou μ, rozptylem σ^2 a konečným centrálním momentem $\mu_k = \mathsf{E}(X-\mathsf{E}[X])^k$ řádu $k=4$. Pak*

$$\sqrt{n}\left(\begin{pmatrix}\overline{X}_n\\ S_n^2\end{pmatrix} - \begin{pmatrix}\mu\\ \sigma^2\end{pmatrix}\right) \xrightarrow{D} \mathsf{N}_2\left(0, \begin{pmatrix}\sigma^2 & \mu_3\\ \mu_3 & \mu_4-\sigma^4\end{pmatrix}\right) \quad pro \quad n\to\infty.$$

Důkaz. Důkaz můžeme dohledat například v knize (Serfling, 1980, strana 72). □

Nyní můžeme tuto větu použít v našem konkrétním případě s binomickým rozdělením, pro odvození asymptotického rozdělení T_n.

Věta 11. *Nechť $X_1,\ldots,X_n$ je náhodný výběr z $\mathsf{Bi}(m,p)$, $p\in(0,1)$ a $m>1$. Pak platí*

$$\sqrt{n}\left(T_n - \frac{m-1}{m}\right) \xrightarrow{D} N(0,V_B^2) \ pro \quad n\to\infty,$$

kde

$$V_B^2 = \frac{2(m-1)(p-1)^2}{m^3p^2}.$$

Důkaz. Všimněme si podobnosti věty 10 s příkladem 2. Výběrový průměr je přímo roven prvnímu výběrovému faktoriálnímu momentu dle (2.1) a pro druhý faktoriální moment platí z (2.2) po jednoduché úpravě

$$\widehat{\mu}_{(2)} = S_n^2 - \overline{X}_n + \overline{X}_n^2.$$

Potřebujeme tedy dostat podíl $\frac{\mu_{(2)}}{\mu_{(1)}^2}$ nějakou transformací z věty 10. Tato úvaha nás přímo vede k použití Δ-metody s funkcí $g(x,y) = \frac{y-x+x^2}{x^2}$. Napočítejme jednotlivé členy Δ-metody dle věty 2

$$g\left(\begin{pmatrix}\overline{X_n}\\ S_n^2\end{pmatrix}\right) = \frac{S_n^2-\overline{X}_n+\overline{X}_n^2}{\overline{X}_n^2} = \frac{\frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}_n^2 - \overline{X}_n + \overline{X}_n^2}{(\frac{1}{n}\sum_{i=1}^n X_i)^2}$$
$$= \frac{\frac{1}{n}\sum_{i=1}^n X_i(X_i-1)}{(\frac{1}{n}\sum_{i=1}^n X_i)^2},$$

$$g\left(\begin{pmatrix}\mu\\ \sigma^2\end{pmatrix}\right) = \frac{\sigma^2 - \mu + \mu^2}{\mu^2} = \frac{mp(1-p) - mp + (mp)^2}{(mp)^2} = \frac{mp - p}{mp} = \frac{m-1}{m}.$$

Derivjme dle jednotlivých proměnných

$$\frac{\partial g(x,y)}{\partial x} = \frac{1}{x^2} - \frac{2y}{x^3},$$
$$\frac{\partial g(x,y)}{\partial y} = \frac{1}{x^2}$$

a dosaďme $x = \mu$ a $y = \sigma^2$. Konkrétní hodnotu rozptylu normálního rozdělení dostaneme maticovým součinem

$$\frac{1}{\mu^6}\left(\mu - 2\sigma^2, \mu\right)\begin{pmatrix}\sigma^2 & \mu_3\\ \mu_3 & \mu_4 - \sigma^4\end{pmatrix}\begin{pmatrix}\mu - 2\sigma^2\\ \mu\end{pmatrix}.$$

Po sérii algebraických úprav dostáváme

$$V_B^2 = \frac{4\sigma^6 + \mu^2\sigma^2 - 4\mu\sigma^4 - \mu^2\sigma^4 + \mu^2\mu_4 + 2\mu\mu_3(\mu - 2\sigma^2)}{\mu^6}.$$

Připoměňme si hodnoty potřebných momentů binomického rozdělení, které jsme si spočítali v rámci věty 5. Nejdříve střední hodnota a rozptyl

$$\mu = mp, \qquad \sigma^2 = mp(1-p).$$

A pro třetí a čtvrtý centrální moment platí

$$\mu_3 = mp(1-p)(1-2p),$$
$$\mu_4 = mp(1-p) + 3(mp)^2(1-p)^2 - 6p^2(m(1-p)^2).$$

Po dosazení dostáváme

$$\begin{aligned} V_B^2 = {} & \frac{4(mp(1-p))^3 + (mp)^2(mp(1-p))}{(mp)^6} \\ & - \frac{4mp(mp(1-p))^2 + (mp)^2(mp(1-p))^2}{(mp)^6} \\ & + \frac{(mp)^2(mp(1-p) + 3(mp)^2(1-p)^2 - 6p^2m(1-p)^2)}{(mp)^6} \\ & + \frac{2mp(mp(1-p)(1-2p))(mp - 2mp(1-p))}{(mp)^6}. \end{aligned}$$

Tento výraz lze nadále zjednodušit až do finální formy, která je uvedená ve znění věty.

□

Tento výsledek lze také nalézt v článku Kyriakoussis a kol. (1998) ve větě 3. Došlo pouze na opravu chyby ve vzorci pro čtvrtý centrální moment, který je správně odvozen ve větě 5.

Asymptotické rozdělení uvedené ve větě 11 ještě nelze na testování použít. Proto si ještě uveďme důsledek této věty.

Důsledek 12. *Nechť $X_1, \ldots, X_n$ je náhodný výběr z $\mathsf{Bi}(m,p)$, $p \in (0,1)$ a $m > 1$. Pak platí*

$$T_B \equiv \frac{\sqrt{n}\left(T_n - \frac{m-1}{m}\right)}{\widehat{V}_B} \xrightarrow{D} N(0,1) \quad \textit{pro} \quad n \to \infty,$$

kde

$$\widehat{V}_B = \frac{(\overline{X}_n - m)\sqrt{2(m-1)}}{m^{3/2}\overline{X}_n}.$$

Důkaz. Vyjděme ze znění věty 11 a nejdříve po přeškálování dostáváme

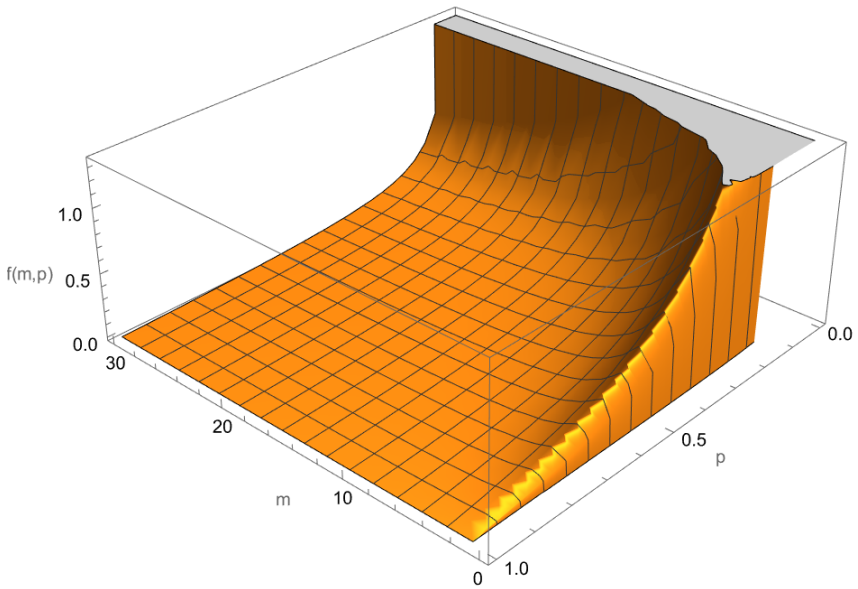
$$\frac{\sqrt{n}\left(T_n - \frac{m-1}{m}\right)}{V_B} \xrightarrow{D} N(0,1).$$

Nyní ve výrazu V_B nahradíme p pomocí momentového odhadu $\widehat{p} = \frac{\overline{X}_n}{m}$. Tento výraz nazveme $\widehat{V}_B$ a rozšíříme jím naše asymptotické rozdělení. Dostáváme

$$\frac{\widehat{V}_B}{V_B} \frac{\sqrt{n}\left(T_n - \frac{m-1}{m}\right)}{\widehat{V}_B} \xrightarrow{D} N(0,1).$$

$\widehat{V}_B$ je zároveň konzistentním odhadem V_B. Z věty o spojité transformaci je zřejmé, že $\frac{1}{\widehat{V}_B} \xrightarrow{P} \frac{1}{V_B}$. Po použití Cramérovy-Sluckého věty dostáváme znění důsledku. □

Podívejme se na to, jak vypadá graf funkce asymptotického rozptylu V_B v závislosti na parametrech m a p. Na obrázku 2.1 můžeme vidět, že pro volby parametru blízko $p = 1$ nebo pro vysoké hodnoty m se dostáváme blízko nule, a můžeme se tak při dělení dostávat do případných numerických problémů.



Obrázek 2.1: Graf funkce V_B v závislosti na proměnných p a m.

Dalším případným problémem mohou být hraničních hodnoty $\widehat{p} \in \{0,1\}$, konkrétně zda-li se nemůže stát, že by některé z členů nebyly definované. Nejprve se podívejme, co se stane, pokud $\widehat{p} = 1$. To nastane právě tehdy, když

$$X_i = m \quad \forall i = 1,\ldots,n \iff \overline{X}_n = m \iff \widehat{p} = \frac{m}{m} = 1 \iff \widehat{V}_B = 0.$$

Můžeme spočítat i pravděpodobnost, že tato situace nastane

$$\mathsf{P}[\overline{X}_n = m] = \mathsf{P}[X_i = m]^n = \left[\binom{m}{m} p^m (1-p)^0\right]^n = p^{m\cdot n}.$$

V tomto případě pro větu 11 dostáváme

$$\sqrt{n}\left(T_n - \frac{m-1}{m}\right) \xrightarrow[n\to\infty]{D} N(0,0),$$

z čehož plyne

$$T_n \xrightarrow{P} \frac{m-1}{m},$$

jelikož $N(0,0) \overset{\text{s.j.}}{=} 0$. Je tedy logické dodefinovat pro tento případ $T_B = 0$.

Podívejme se i na druhou zmiňovanou situaci, pro kterou platí

$$X_i = 0 \ \ \forall i = 1,\ldots,n \iff \overline{X}_n = 0 \iff \widehat{p} = 0 \implies T_n \text{ a } \widehat{V}_B \text{ nejsou definovány.}$$

Pravděpodobnost tohoto jevu spočítáme jako

$$\mathsf{P}[\overline{X}_n = 0] = \mathsf{P}[X_i = 0]^n = \left[\binom{m}{0} p^0 (1-p)^m\right]^n = (1-p)^{m\cdot n}.$$

Pro tuto situaci také dodefinujeme $T_B = 0$.

V praktické části tedy musíme počítat s těmito situacemi a dodefinujeme naší testovou statistiku následujícím způsobem

$$\widetilde{T}_B = \begin{cases} 0 & \text{pokud } (X_i = 0 \ \ \forall i = 1,\ldots,n) \text{ nebo } (X_i = m \ \ \forall i = 1,\ldots,n), \\ T_B & \text{jinak.} \end{cases}$$

Zaveďme si náhodnou veličinu

$$V_n = \mathbf{1}\left[(X_i = 0 \ \ \forall i = 1,\ldots,n) \text{ nebo } (X_i = m \ \ \forall i = 1,\ldots,n)\right],$$

pro kterou platí

$$\mathsf{P}[V_n = 1] = \mathsf{P}[\overline{X}_n = 0] + \mathsf{P}[\overline{X}_n = m] = (1-p)^{m\cdot n} + p^{m\cdot n} \xrightarrow[n\to\infty]{} 0$$

a z definice konvergence v pravděpodobnosti dostáváme

$$\forall \varepsilon \in (0,1) : \lim_{n\to\infty} \mathsf{P}[|V_n| > \varepsilon] = \lim_{n\to\infty} \mathsf{P}[V_n = 1] = 0 \implies V_n \xrightarrow{P} 0.$$

Nakonec při použití Cramérovy-Slutského věty a důsledku 12 máme

$$\widetilde{T}_B = T_B(1 - V_n) + 0 \cdot V_n \xrightarrow{D} \mathsf{N}(0,1) \quad \text{pro} \quad n \to \infty.$$

Ve výsledku vidíme, že dodefinování testové statistiky nemá vliv na asymptotické vlastnosti.

Z důsledku 12 dostáváme kromě testové statistiky $\widetilde{T}_B$ také její kritický obor a p-hodnotu.

Kritický obor:

$$H_0 \quad \text{zamítneme na hladině} \quad \alpha \iff \left|\widetilde{T}_B\right| > u_{1-\alpha/2},$$

kde $u_{1-\alpha/2}$ značí $1-\frac{\alpha}{2}$ kvantil normálního normovaného rozdělení.

P-hodnotu lze spočíst jako

$$2\left[1-\Phi(|\widetilde{T}_B|)\right],$$

kde Φ je distribuční funkce normovaného normálního rozdělení.

2.2 Test dobré shody založený na obecném faktoriálním momentu

Druhým testem, který budeme uvažovat, je test navržený v článku Aleksandrov a kol. (2022). Zatímco předešlý test byl založen na prvních dvou faktoriálních momentech, tento test uvažuje faktoriální moment obecně vyššího řádu. Obsahem této sekce bude také ukázka vztahu mezi testem uvažovaným v této části a testem uvažovaným v minulé sekci.

Idea testu je obdobná jako u testu předešlého, ale zde si ukážeme, že lze najít obecné pravidlo, které najde podíl faktoriálních momentů tak, aby vždy došlo ke zkrácení parametru p. Hledáme podíl r-tého a s-tého faktoriálního momentu a všimněme si, že pokud do jmenovatele přidáme $(r-s)$-tý faktoriální moment, tak dostáváme

$$\Psi_{(r,s)} \equiv \frac{\mu_{(r)}}{\mu_{(r-s)}\mu_{(s)}} = \frac{m_{(r)}p^r}{m_{(r-s)}p^{r-s}m_{(s)}p^s} = \frac{m_{(r)}}{m_{(r-s)}m_{(s)}} \qquad \text{pro } 1 \le s < r \le m.$$

Příslušným empirickým protějškem $\Psi_{(r,s)}$ je

$$T_{(r,s)} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i)_{(r)}}{\frac{1}{n}\sum_{i=1}^{n}(X_i)_{(r-s)}\frac{1}{n}\sum_{i=1}^{n}(X_i)_{(s)}}.$$

Navíc platí následující vztah mezi $\Psi_{(r,s)}$ a $T_{(r,s)}$.

Věta 13. *Nechť $X_1,\ldots,X_n$ je náhodný výběr z $\mathsf{Bi}(m,p), p \in (0,1)$, pak*

$$T_{(r,s)} \xrightarrow{P} \Psi_{(r,s)}.$$

Důkaz. Důkaz je analogický důkazu věty 9.

□

Uveďme si pomocnou větu, která nám pomůže odvodit asymptotické rozdělení $T_{(r,s)}$.

Věta 14. *Nechť $X_1, \ldots, X_n$ je náhodný výběr s $\mathsf{Bi}(m,p)$, $p \in (0,1)$, $m > 1$ a $1 \leq s < r \leq m$, označme*

$$\mathbf{Z}_i = \left((X_i)_{(r)}, (X_i)_{(r-s)}, (X_i)_{(s)}\right)^\top,$$
$$\overline{\mathbf{Z}}_n = \left(\frac{1}{n}\sum_{i=1}^{n}(X_i)_{(r)}, \frac{1}{n}\sum_{i=1}^{n}(X_i)_{(r-s)}, \frac{1}{n}\sum_{i=1}^{n}(X_i)_{(s)}\right)^\top.$$

Pak

$$\sqrt{n}\left(\overline{\mathbf{Z}}_n - \mathsf{E}[\mathbf{Z}_1]\right) \xrightarrow{D} \mathsf{N}_3(\mathbf{0},\boldsymbol{\Sigma}) \ \textit{pro} \quad n \to \infty,$$

kde

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{(r,r)} & \sigma_{(r,r-s)} & \sigma_{(r,s)} \\ \sigma_{(r,r-s)} & \sigma_{(r-s,r-s)} & \sigma_{(r-s,r)} \\ \sigma_{(r,s)} & \sigma_{(r-s,s)} & \sigma_{(s,s)} \end{pmatrix} \tag{2.3}$$

a

$$\sigma_{(k,l)} = n_{(k)} n_{(l)} p^{k+l} A_{k,l},$$

kde

$$A_{k,l} = \sum_{i=1}^{\min\{k,l\}} \frac{\binom{k}{i}\binom{m-k}{l-i}}{\binom{m}{l}} \frac{1-p^i}{p^i}. \tag{2.4}$$

Důkaz. Pomocí centrální limitní věty najděme asymptotické rozdělení $\overline{\mathbf{Z}}_n$. Náhodné vektory $\mathbf{Z}_i$ jsou nezávislé a stejně rozdělené a varianční matice je konečná, což platí triviálně z omezenosti, a jsou tak splněny předpoklady mnohorozměrné centrální věty. Z tvrzení 13 zřejmě platí

$$\mathsf{E}\left[\mathbf{Z}_1\right] = \left(\mu_{(r)}, \mu_{(r-s)}, \mu_{(s)}\right)^\top.$$

Hodnoty členů varianční matice $\mathrm{var}[\mathbf{Z}_1] = \boldsymbol{\Sigma}$ jsou podrobněji odvozeny v článku Aleksandrov a kol. (2022).

□

Máme tedy sdružené asymptotické rozdělení vektoru empirických faktoriálních momentů. Odvoďme podobně jako u předchozího testu asymptotické rozdělení $T_{(r,s)}$.

Věta 15. *Mějme náhodný výběr $X_1, \ldots, X_n$ s $\mathsf{Bi}(m,p)$, $p \in (0,1)$, $m > 1$ a nechť platí $1 \leq s < r \leq m$. Pak*

$$\sqrt{n}\left(T_{(r,s)} - \Psi_{(r,s)}\right) \xrightarrow{D} \mathsf{N}(0, V^2_{(r,s)}) \quad \textit{pro} \quad n \to \infty,$$

kde asymptotický rozptyl je

$$V^2_{(r,s)} = \frac{m^2_{(r)}}{m^2_{(r-s)} m^2_{(s)}} \left(A_{r,r} + A_{r-s,r-s} + A_{s,s} - 2A_{r,r-s} - 2A_{r,s} + 2A_{r-s,s}\right)$$

a $A_{k,l}$ je definováno v (2.4).

Důkaz. Důkaz je velmi podobný důkazu věty 11. Vyjděme z tvrzení věty 14 a aplikujme na ní takovou transformaci, abychom opět postupovali dle idey uvedené v úvodu této sekce. Můžeme si všimnout, že Δ-metoda je tentokrát přímočará $g(x,y,z) = \frac{x}{yz}$. Spočtěme jednotlivé členy

$$g(\overline{\mathbf{Z}}) = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i)_{(r)}}{\frac{1}{n}\sum_{i=1}^{n}(X_i)_{(r-s)}\frac{1}{n}\sum_{i=1}^{n}(X_i)_{(s)}} = T_{(r,s)},$$

$$g\left(\mathsf{E}\left[\overline{\mathbf{Z}}\right]\right) = \frac{\mu_{(r)}}{\mu_{(r-s)}\mu_{(s)}} = \frac{m_{(r)}p^r}{m_{(r-s)}p^{r-s}m_{(s)}p^s} = \frac{m_{(r)}}{m_{(r-s)}m_{(s)}} = \Psi_{(r,s)}.$$

Pro získání rozptylu musíme napočítat vektor $\mathbf{D}(\Psi_{(r,s)})$

$$\frac{\partial g(x,y,z)}{\partial x} = \frac{1}{yz} = \begin{bmatrix} y = \mu_{(r-s)} \\ z = \mu_{(s)} \end{bmatrix} = \frac{1}{\mu_{(r-s)}\mu_{(s)}} = \frac{1}{m_{(r-s)}m_{(s)}p^r},$$

$$\frac{\partial g(x,y,z)}{\partial y} = -\frac{x}{y^2z} = \begin{bmatrix} x = \mu_{\mu_{(r)}} \\ y = \mu_{(r-s)} \\ z = \mu_{(s)} \end{bmatrix} = -\frac{\mu_{(r)}}{\mu_{(r-s)}^2\mu_{(s)}} = -\frac{m_{(r)}}{m_{(r-s)}^2m_{(s)}p^{r-s}},$$

$$\frac{\partial g(x,y,z)}{\partial z} = -\frac{x}{yz^2} = \begin{bmatrix} x = \mu_{\mu_{(r)}} \\ y = \mu_{(r-s)} \\ z = \mu_{(s)} \end{bmatrix} = -\frac{\mu_{(r)}}{\mu_{(r-s)}\mu_{(s)}^2} = -\frac{m_{(r)}}{m_{(r-s)}m_{(s)}^2p^s}.$$

Dohromady dostáváme

$$\mathbf{D}(\Psi_{(r,s)}) = \left(\frac{1}{m_{(r-s)}m_{(s)}p^r}, -\frac{m_{(r)}}{m_{(r-s)}^2m_{(s)}p^{r-s}}, -\frac{m_{(r)}}{m_{(r-s)}m_{(s)}^2p^s}\right)^\top.$$

Rozptyl $V_{(r,s)}^2$ získáme jako maticový součin $\mathbf{D}(\Psi_{(r,s)})\mathbf{\Sigma}\mathbf{D}^\top(\Psi_{(r,s)})$, kde $\mathbf{\Sigma}$ je matice (2.3). Po sérii algebraických úprav dostáváme $V_{(r,s)}^2$ uvedené ve znění věty, což kompletuje důkaz věty.

□

Důsledek 16. *Mějme náhodný výběr* $X_1, \ldots, X_n$ *s* $\mathsf{Bi}(m,p)$, $p \in (0,1)$, $m > 1$ *a nechť platí* $1 \leq s < r \leq m$. *Pak*

$$T(r,s) \equiv \sqrt{n}\frac{\left(T_{(r,s)} - \Psi_{(r,s)}\right)}{\widehat{V}_{(r,s)}} \xrightarrow{D} \mathsf{N}(0,1) \;\; pro \quad n \to \infty.$$

Důkaz. Důkaz se provede použitím Cramérovy-Slutského věty a je analogický důkazu věty 12.

□

Poznámka. Podívejme se, na volbu parametrů $r = 2$ a $s = 1$ ve větě 15. Postupně počítejme

$$\Psi_{(2,1)} = \frac{m_{(2)}}{m_{(2-1)}m_{(1)}} = \frac{m-1}{m},$$

$$T_{(2,1)} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i)_{(2)}}{\frac{1}{n}\sum_{i=1}^{n}(X_i)_{(2-1)}\frac{1}{n}\sum_{i=1}^{n}(X_i)_{(1)}} = \frac{\frac{1}{n}\sum_{i=1}^{n}X_i(X_i-1)}{(\frac{1}{n}\sum_{i=1}^{n}X_i)^2} = T_n,$$

$$\begin{aligned} V^2_{(2,1)} &= \left(\frac{m_{(2)}}{m_{(1)}m_{(1)}}\right)^2 \left(A_{2,2} + A_{1,1} + A_{1,1} - 2A_{2,1} - 2A_{2,1} + 2A_{(1,2)}\right) \\ &= \frac{m-1}{m}\left(A_{2,2} + 2A_{1,1} - 2A_{2,1}\right) = \frac{2(m-1)(p-1)^2}{m^3p^2} = V_B^2. \end{aligned}$$

Můžeme vidět, že všechny členy jsou stejné jako ve větě 11. Je tedy zřejmé, že statistika představená v sekci 2.1 je speciálním případem testové statistiky z této sekce.

Dalšími zajímavými volbami parametrů testové statistiky $T(r,s)$ jsou například (3,1), tento výběr paramterů se dá z určitého pohledu brát jako statistika založená na šikmosti nebo (4,1) či (4,2), což jsou formy statistiky založené na špičatosti.

Poznámka. Všimněme si, že stejně jako v minulé sekci můžeme za určitých okolností dostat nedefinované výrazy $T_{(r,s)}$ a $V_{(r,s)}$. V takovém případě lze podobným způsobem jako pod důsledkem 12 ukázat, že dodefinováním výrazu nijak nenarušíme asymptotické chování testu a podobně také dodefinujeme testovou statistiku

$$\widetilde{T}_{(r,s)} = \begin{cases} 0 & \text{pokud } (X_i \le s-1 \text{ nebo } X_i \le r-s-1) \quad \forall i = 1,\dots,n \\ & \text{nebo } X_i = m \quad \forall i = 1,\dots,n, \\ T_{(r,s)} & \text{jinak.} \end{cases}$$

Na základě důsledku 16 si uveďme kritický obor a p-hodnotu pro naší testovou statistiku.

Kritický obor:

$$H_0 \quad \text{zamítneme na hladině} \quad \alpha \iff \left|\widetilde{T}(r,s)\right| > u_{1-\alpha/2},$$

kde $u_{1-\alpha/2}$ značí $1-\frac{\alpha}{2}$ kvantil normálního normovaného rozdělení.

P-hodnotu lze spočíst jako

$$2\left[1 - \Phi\left(|\widetilde{T}(r,s)|\right)\right],$$

kde Φ je distribuční funkce normovaného normálního rozdělení.

3. Simulace

V této kapitole si na simulacích ukážeme některé vlastnosti testů dobré shody, které jsme si popsali ve druhé kapitole. Parametry (r, s) budeme volit (2,1) a (3,1). Tyto volby jsou nejjednodušší z hlediska výpočetní náročnosti a jsou definovány pro nejvíce možností paramteru m, což plyne z podmínky $1 \leq s < r \leq m$. Další důvody těchto voleb si uvedeme dále v textu.

Jako referenční test pro nás bude sloužit χ^2 test dobré shody. Tento test je určen k testování dobré shody s multinomickým rozdělením a je podrobněji popsán v příloze A.1. Pro náš případ s binomickým rozdělením budeme uvažovat jeho adaptaci, a to χ^2 test dobré shody s binomickým rozdělením s odhadnutými parametry. Pro tento test máme přirozený počet kategorií $K = m+1$ a odhadujeme jeden parametr $\theta = p$, tedy počet odhadnutých parametrů je $b = 1$. Náhodná veličina W_k značí absolutní četnost kategorie binomického rozdělení s k úspěchy. Dohromady dostáváme

$$\chi^2 \equiv \sum_{k=1}^{m+1} \frac{\left(W_k - np_k(\widehat{\theta})\right)^2}{np_k(\widehat{\theta})} \xrightarrow[n\to\infty]{D} \chi^2_{m-1},$$

kde $p_k = \mathsf{P}(X = k+1)$ pro $X \sim \mathsf{Bi}(m, p)$, $np_k(\widehat{\theta})$ je odhad očekávané četnosti za platnosti nulové hypotézy a $n = \sum_{k=1}^{m+1} W_k$.

Aby byla aproximace χ^2rozdělením dostatečně dobrá, je potřeba aby $np_k \geq 5$ pro všechna $k = 1, \ldots, m+1$, viz (Anděl, 2019, strana 155). Pro některé volby parametrů p a m budeme nuceni spojovat krajní kategorie. Podrobněji budeme počet spojených kategorii diskutovat u konkrétních alternativ a pro pořádek budeme χ^2 test se spojenými h prvními a l posledními kategoriemi značit $\chi^2(h, l)$.

Simulace budeme provádět prostřednictvím softwaru R Core Team (2023) a také pomocí doplňkových balíčků PoissonBinomial Junge (2023) a mvtnorm Genz a Bretz (2009). Kódy pro testy navržené v minulé kapitole, upravení χ^2 rozdělení, odhady síly a hladiny testů i tabulky a obrázky se dají považovat za vlastní přínos práce.

3.1 Hladina testů

Nejprve se podívejme na empirický odhad hladiny testu. Předpokládejme, že data opravdu pocházejí z binomického rozdělení s parametry m a p. Z tohoto rozdělení budeme náhodně generovat jednotlivé realizace, čímž dostaneme náhodný výběr o rozsahu n. Pro tento náhodný výběr spočítáme hodnotu testové statistiky a p-hodnotu. Pro všechny testy budeme uvažovat hladinu významnosti $\alpha = 0.05$, kterou porovnáme s p-hodnotou náhodného výběru a rozhodneme o zamítnutí/nezamítnutí nulové hypotézy. Celý tento proces budeme opakovat tisíckrát a výsledným podílem zamítnutých nulových hypotéz dostaneme empirický odhad hladiny testu.

Pro hladinu testu při námi zvolené hladině významnosti očekáváme přibližně 0.05. V tabulce 3.1 můžeme sledovat empirický odhad hladiny testu v závislosti na parametru $m \in \{5, 10, 20\}$ (horizontálně) a rozsahu výběru n (vertikálně) při konstantní volbě $p = 0.5$. Empirické hladiny porovnáváme pro testy

	$T(2,1)$			$T(3,1)$			$\chi^2(0,0)$	$\chi^2(3,3)$	$\chi^2(7,7)$
n \ m	5	10	20	5	10	20	5	10	20
10	0.033	0.032	0.038	0.031	0.032	0.035	0.042	0.048	0.034
20	0.047	0.037	0.046	0.034	0.045	0.041	0.039	0.042	0.043
50	0.044	0.033	0.036	0.041	0.034	0.045	0.027	0.053	0.035
100	0.041	0.039	0.040	0.044	0.045	0.040	0.042	0.052	0.041
200	0.046	0.044	0.035	0.059	0.041	0.036	0.060	0.045	0.055
500	0.054	0.042	0.048	0.053	0.038	0.046	0.042	0.062	0.053
1000	0.046	0.052	0.050	0.044	0.052	0.053	0.046	0.049	0.045

Tabulka 3.1: Empirický odhad hladiny při volbách parametrů $m \in \{5,10,20\}$, $p = 0.5$ v závislosti na rozsahu výběru n.

$T(2,1), T(3,1)$ a χ^2 test dobré shody. U empirických odhadů hladin testu budeme spojovat kategorie tak, aby pro rozsah výběru $n = 100$ byla očekávaná četnost v každé kategorii námi známého binomického rozdělení byla alespoň 5. V praxi nebudeme přesně vědět, kolik kategorií χ^2 spojit tak, aby byla podmínka splněna. Ale abychom v tomto případě dostali co nejlepší referenci o hladině testů i při extrémních volbách p, a dostali tak dobré porovnání pro naše testy, budeme kategorie spojovat přesně tak, aby byla podmínka splněna.

Důležité je sledovat především, jaký má vliv rozsah náhodného výběru na vlastnosti testových statistik založených na faktoriálních momentech. Jelikož tyto testy jsou testy asymptotickými, můžeme očekávat, že pro malé rozsahy výběru bude aproximace normálním rozdělením nepřesná. Vidíme ale, že ani aproximace χ^2 testu pro malá n není dobrá.

Podobné asymptotické chování můžeme vidět i pro volbu $p = 0.1$. V tabulce 3.1 ale dochází také k situaci, kde je původní testová statistika nedefinovaná, což jsme podrobně rozdebírali na konci sekce 2.1. Tuto situaci můžeme například vidět při volbě parametrů $n = 10$, $m = 5$ a $p = 0.1$. U testů $T(2,1)$ a $T(3,1)$ můžeme kvůli způsobu dodefinování vidět nižší hodnoty emprického odhadu hladiny. Především pro testovou statistiku $T(3,1)$ je podstatné množství hodnot (438 z 1000) dodefinováno (jedná se o všechny náhodné výběry obsahující jen 0 a 1). Pro stejné volby paramterů jsou i některé výsledky χ^2 testové statistiky nedefinované, jelikož $np_k(\widehat{\theta}) = 0$ pokud $X_i = 0$ pro všechna $i = 1, \ldots, n$. Těchto nedefinovaných členů je 69 a výpočet empirického odhadu je tedy v tomto případě počítán jen z 931 opakování.

3.2 Síla testů

Další důležitou charakteristikou je síla testu vůči různým alternativám. Empirický dohad síly budeme počítat podobným způsobem jako hladinu, ale náhodné realizace nebudeme generovat z binomického rozdělení, ale z jiného rozdělení. V článcích, které se už zabývaly testy dobré shody s binomickým rozdělením založenými na faktoriálních momentech, se jako alternativa bralo například rovnoměrné a Poissonovo rozdělení, viz článek Kyriakoussis a kol. (1998). U takových rozdělení můžeme apriori předpokládat, že už z principu odhalíme, že se nejedná o binomické rozdělení.

	$T(2,1)$			$T(3,1)$			$\chi^2(0,4)$	$\chi^2(0,8)$	$\chi^2(0,17)$
n \ m	5	10	20	5	10	20	5	10	20
10	0.028	0.024	0.025	0.002	0.003	0.008	0.004*	0.038	0.047
20	0.035	0.035	0.034	0.010	0.020	0.015	0.037	0.039	0.042
50	0.047	0.037	0.038	0.021	0.013	0.032	0.047	0.039	0.058
100	0.051	0.054	0.047	0.024	0.026	0.036	0.044	0.058	0.057
200	0.044	0.047	0.037	0.036	0.035	0.051	0.047	0.050	0.063
500	0.045	0.045	0.052	0.044	0.033	0.049	0.039	0.055	0.060
1000	0.049	0.051	0.045	0.050	0.050	0.041	0.044	0.048	0.047

Tabulka 3.2: Empirický odhad hladiny při volbách parametrů $m \in \{5,10,20\}$ a $p = 0.1$ v závislosti na rozsahu výběru n. *Počítano z 931 opakování.

Z tohoto důvodu se v této části podíváme na rozdělení, která jsme si uvedli v rámci sekce 1.4, tedy rozdělení, která stejně jako binomické rozdělení vznikají jako součet náhodných veličin s alternativním rozdělením, tyto alternativní veličiny však nejsou stejně rozdělené, nejsou nezávislé, anebo nejsou ani nezávislé ani stejně rozdělené. Tyto případy už bývají v praxi obtížněji odhalitelné a například jen z explorativní analýzy nemusíme odhalit odchylku od binomického rozdělení. Tuto podobnost jsme mohli vidět například na obrázku 1.3, a právě v těchto situacích se v praxi uchylujeme k testům dobré shody. Ve všech případech budeme volit hladinu významnosti $\alpha = 0.05$.

Jelikož v praxi nevíme, z jakého rozdělení naše data pochází a jaký je přesně parametr úspěchu p, nemůžeme ani přesně vědět, jaké máme volit spojování kategorií χ^2 testu. U testování síly budeme tedy volit spojování kategorii na základě paramteru m, který je v našem případě známý vždy. Konkrétně budeme spojování volit tak, aby pro naše známé m a rozsah výberu $n = 100$ splňovaly podmínku všechny kategorie rozdělení $\mathsf{Bi}(m, 0.5)$.

3.2.1 Síla testu vůči alternativě stejně rozdělených závislých náhodných veličin

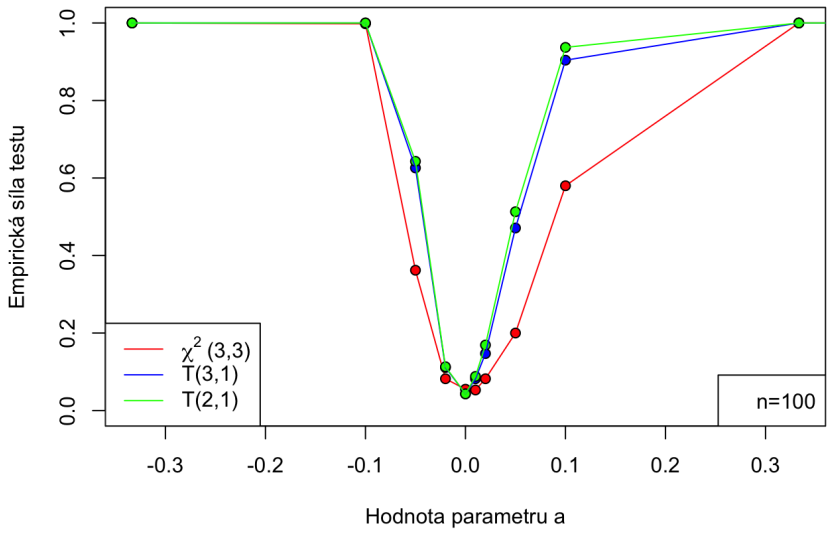
Touto situací jsme se zabývali v části 1.4.1, kde jsme si také uvedli, že záleží na tom, jak na sobě budou náhodné veličiny závislé. My budeme konkrétně uvažovat náhodný vektor $(Z_1, \ldots, Z_m)^\top = \boldsymbol{Z} \sim \mathsf{N}_m(\mathbf{0}, \boldsymbol{\Sigma}_a)$, kde

$$\boldsymbol{\Sigma}_a = \begin{pmatrix} 1 & a & \cdots & a \\ a & 1 & \cdots & a \\ \vdots & \vdots & \ddots & \vdots \\ a & a & \cdots & 1 \end{pmatrix}$$

a parametr a volíme tak, aby $\boldsymbol{\Sigma}_a$ byla pozitivně semidefinitní. Všechny složky vektoru mají stejné marginální rozdělení $Z_i \sim \mathsf{N}(0,1)$ a jsou tak stejně rozdělené, ale pro volby paramteru $a \neq 0$ nejsou složky vektoru $\boldsymbol{Z}$ nezávislé. Z i-té složky náhodného vektoru $\boldsymbol{Z}$ vytvoříme dichotomickou proměnnou $Y_i = \mathbf{1}(Z_i > c)$, kde $i = 1, \ldots, m$ a volba $c \in \mathbb{R}$ ovlivňuje korelaci a pravděpodobnost úspěchu. Tímto způsobem dostaneme jednu realizaci $(Y_1, \ldots, Y_m)^\top$. Potom $X = \sum_{i=1}^m Y_i$ je součet stejně rozdělených závislých náhodných veličin s alternativním rozdělením.

Tento proces opakujeme nezávisle n-krát tak, abychom dostali náhodný výběr $X_1, \ldots, X_n$.

Jaký má vliv volba parametru a na sílu testů pro $c = 0$, $n = 100$ a $m = 10$, můžeme sledovat na obrázku 3.1. Síla testů $T(2,1)$ a $T(3,1)$ je téměř stejná, ale oproti χ^2 testu vidíme výrazně vyšší citlivost na změnu parametru a. Hlavně v oblasti, kde testy už zpozorují změnu v závislosti, ale ještě nezamítají všechny případy, je rozdíl v síle výrazný (o 0.3 ve prospěch testů založených na faktoriálních momentech). Pro vyšší rozsahy výběru jsou výsledky velmi podobné s tím, že všechny testy na vyšším vzorku odhalí změnu v parametru a rychleji. Další varianty s různými volbami paramteru m lze viděte na obrázcích A.1 a A.2.



Obrázek 3.1: Síla testů vůči alternativě závislých náhodných veličin, kde $m = 10$ v závislosti na volbě parametru a v matici Σ_a.

3.2.2 Síla testu vůči alternativě nezávislých různě rozdělených náhodných veličin

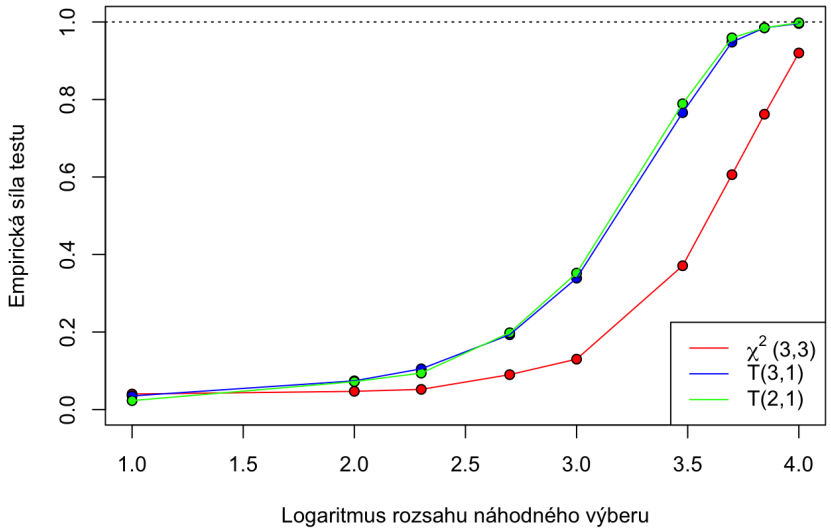
Veličinou, která vzniká jako součet alternativních veličin, jež jsou nezávislé, ale nejsou stejně rozdělené, jsme se zabývali v části 1.4.2. Připomeňme, že takové veličiny mají dle definice 5 Poissonovo binomické rozdělení s parametry m a $\boldsymbol{p}$. V této sekci budeme uvažovat $\boldsymbol{p}$ jako vektor m stejných složek a rozdíl od binomického rozdělení bude určovat vektor $\boldsymbol{\delta}$. V takovém případě můžeme model, ze kterého budeme veličiny generovat zapsat jako

$$\mathcal{F}_{ALT} = \left\{ \mathsf{PBi}(m, \boldsymbol{p} + \boldsymbol{\delta}), \boldsymbol{p} = (p, \ldots, p)^\top, \boldsymbol{\delta} = (0, \ldots, 0, d, \ldots, d)^\top \right\}$$

přičemž H_0 platí právě tehdy, když $\boldsymbol{\delta} = \mathbf{0}$. Označme počet složek prvního bloku q, druhý blok má v takovém případě $m - q$ složek.

Vraťme se k příkladu, který jsme si v rámci kapitoli 1.4.2 uváděli. Šlo o příklad s dvěma bloky zkouškových otázek, u kterých měl student různou pravděpodob-

nost správné odpovědi. Studentovy odpovědi měly Poissonovo binomické rozdělení s parametry $\boldsymbol{p} = (0.50, 0.50, 0.50, 0.50, 0.50, 0.75, 0.75, 0.75, 0.75, 0.75)^\top \equiv \boldsymbol{p}_B$ a $m = 10$. Na obrázku 3.2 se můžeme podívat, jak si testy vedou při různých rozsazích výběru. Můžeme vidět, že na intervalu (100, 10 000) testy dobré shody založené na faktoriálních momentech poráží χ^2 test v počtu správně zamítnutých H_0 na hladině $\alpha = 0.05$. Další volby parametrů p a d můžeme vidět na obrázku A.4.

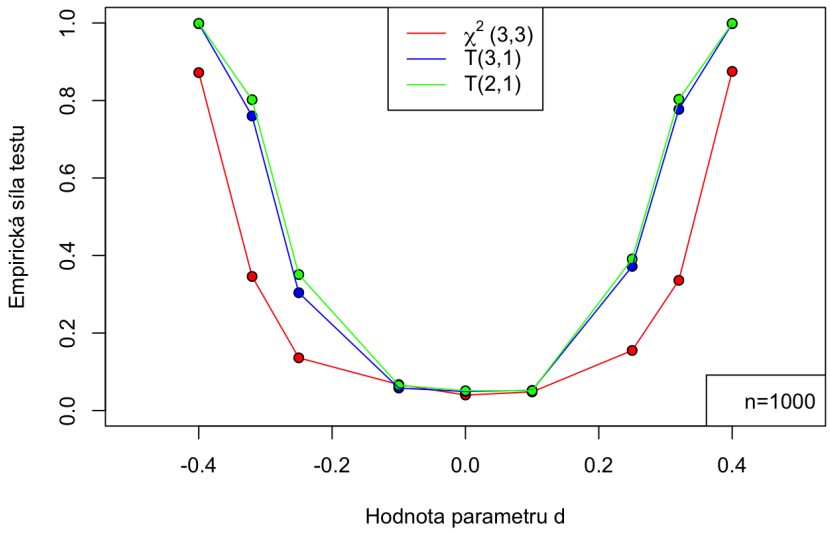


Obrázek 3.2: Síla testů proti PBi(10, $\boldsymbol{p}_B$) v závislosti na rozsahu výběru.

Podívejme se, jak testy reagují, pokud bychom měnili v příkladu se studentem pravděpodobnost úspěchu ve druhém bloku při konstantním rozsahu výběru $n = 1000$. Rozdíl mezi bloky určujeme pomocí paramteru d, který bude postupně nabývat hodnot $d \in \{-0.40, -0.25, -0.10, 0.00, 0.10, 0.25, 0.40\}$.

Na grafu 3.3 můžeme vidět, že testy založené na faktoriálních momentech jsou citlivější na velikost parametru d. Při změně o 0,40 zamítají takřka ve všech případech. Na změnu v řádu jedné čtvrtiny narozdíl od χ^2 reagují, když zamítají nulovou hypotézu ve $\frac{2}{5}$ případů a na změnu v řádu jedné desetiny stejně jako χ^2 nereagují. Aby testy takto malý rozdíl v pravděpodobnosti úspěchu zpozorovaly, je potřeba násobně více pozorování, jak je vidět na obrázku A.3. Stejný princip s jinými parametry, ale podobnými výsledky, lze vidět na obrázku A.5.

Sledujme ještě variantu, kde bude různý počet otázek v jednotlivých blocích, ale počet otázek, pravděpodobnosti správných odpovědí i počet bloků bude stejný. Počet otázek, pro které je pravděpodobnost úspěchu 0.50, budeme značit q, počet otázek s pravděpodobností úspěchu 0.75 je pak $10 - q$. Na obrázku 3.4 můžeme vidět příslušné síly testu vzhledem k volbě parametru q. Vidíme, že testy založené na faktoriálnich momentech správně zamítají nulovou hypotézu ve výrazně více případech oproti χ^2 testu, který takřka nereaguje. Při vyšší volbě rozdílu parametrem d na obrázku A.6 už vidíme reakci obou testů, ale testy založené na faktoriálních momentech opět disponují podstatně větší sílou.

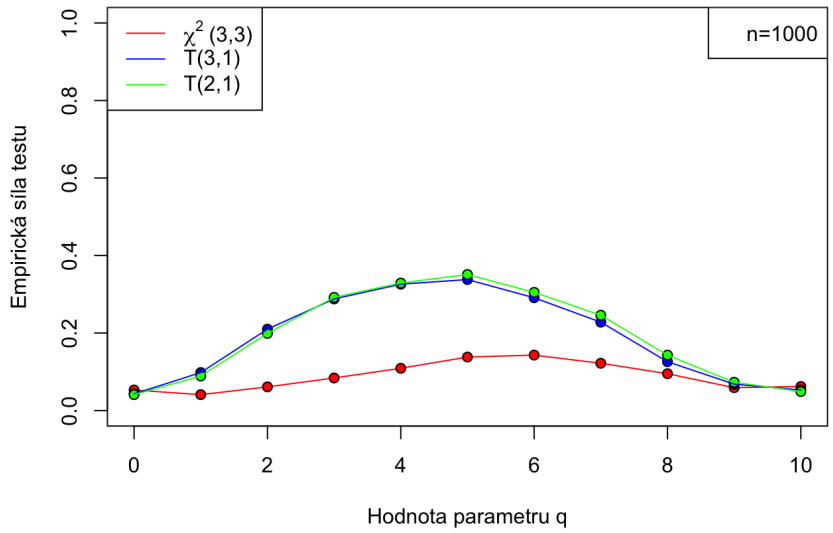


Obrázek 3.3: Síla testů proti alternativě $\mathsf{PBi}(10, \boldsymbol{p}+\boldsymbol{\delta})$, $p = 0.5$, $q = 5$ v závislosti na parametru d.

3.3 Závěr simulační studie

Simulační studie nám ukázala, že hladiny u vybraných testů dobré shody s binomickým rozdělením se nijak zásadně neliší. Pro všechny testy potřebujeme dostatečně velké rozsahy výběrů, aby fungovaly jejich asymptotické vlastnosti. χ^2 testy musíme i při vyšších hodnotách n upravovat tak, aby byla aproximace dostatečně dobrá. Pokud nastane situace, kdy je zároveň p blízko 0 nebo 1 a m a n jsou malé, může docházet k situacím, které jsou v testech založených na faktoriálních momentech speciálně dodefinovány a u χ^2 testu definovány vůbec nejsou. U statistik, které jsou založené na faktoriálních momentech vyššího řádu, dochází k těmto situacím častěji, a mají tak pro tyto specifické hodnoty parametrů nižší hodnoty emprického odhadu hladiny kvůli zvolenému způsobu dodefinování.

V případech, kdy máme testy s podobnou hladinou, je logické pracovat s testem, který má větší sílu. V rámci simulační studie jsme ukázali, že testy založené na faktoriálních momentech mají výrazně vyšší sílu proti alternativám závislých náhodných veličiny i proti alternativám různě rozdělených náhodných veličin a to při různých způsobech závislosti a také při rozličných variantách různě rozdělených náhodných veličin. Zároveň jsme v simulační studii nepozorovali výraznější rozdíly v síle mezi testovými statistikami s různými řády faktoriálních momentů.



Obrázek 3.4: Síla proti $\mathsf{PBi}(10, \boldsymbol{p}+\boldsymbol{\delta})$, kde $p = 0.5$ a $d = 0.25$ v závislosti na volbě parametru q.

4. Aplikace na reálných datech

Binomické rozdělení přirozeně vzniká jako počet úspěchů v m nezávislých pokusech se stejnou pravděpodobností úspěchu. V některých praktických situacích však nemusí být zcela jasné, zda jsou dané pokusy nezávislé a zda mají stejnou pravděpodobnost úspěchu. Případně mohou být oba předpoklady zřejmě porušené, ale stále se můžeme ptát, zda by binomické rozdělení mohlo být rozumným (přibližným) modelem pro naše data.

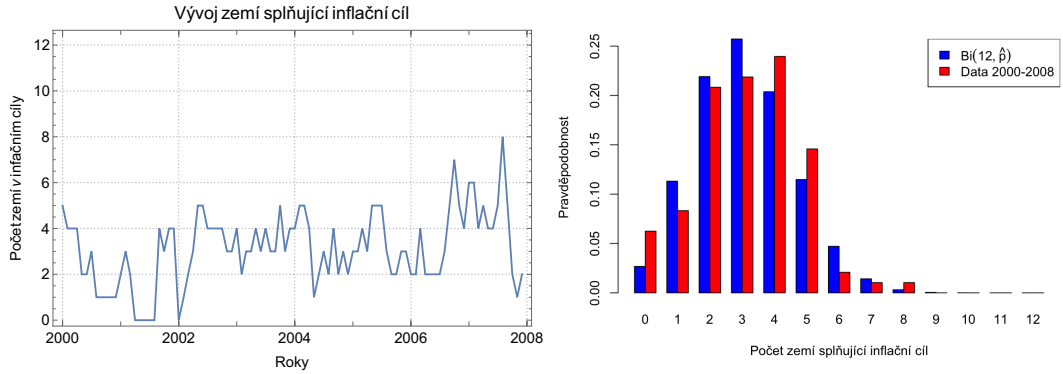
4.1 Inflace v zemích eurozóny

Jako příklad sledujme naplňování inflačního cíle v zemích eurozóny. Konkrétně budeme pracovat s daty z Eurostat (2024a) a Eurostat (2024b). Evropská centrální banka si už při svém vzniku v roce 1998 dala za cíl udržovat cenovou stabilitu. Tento cíl byl formalizován Radou guvernérů v roce 2003 Bank (2003) a konkrétně říká: „Cenová stabilita je definovaná jako meziroční růst harmonizovaného indexu spotřebitelských cen (HICP) pro eurozónu těsne pod 2 %. Cenová stabilita by měla být udržována na střednědobém horizontu.“ A zároveň dodává: „Toto prohlášení potvrzuje závazek ECB pro udržování bezpečného pásma od hrozby deflace.“ Tento cíl byl na zasedání Rady guvernérů 8.7.2021 revidován Bank (2021) a ECB si dala za cíl přesně 2 %, přičemž hodnoty vyšší i nižší jsou stejně nepřijatelné. Sledujme tedy vývoj v období 2000-2021, označme míru mezoričního růstu harmonizovaného indexu spotřebitelských cen π a sledujme, kolik zemí splňovalo podmínku $0 \leq \pi \leq 2\%$. Data o π jsou vydávána měsíčně, a můžeme tak v různých obdobích sledovat, zda-li je, nebo bylo, binomické rozdělení smysluplným modelem pro tato data.

Nejprve se podívejme na stejné období a země, které byly sledované v článku Aleksandrov a kol. (2022). Jedná se o případ sedmnácti zemí eurozóny (EA17) a jejich inflační vývoj mezi rokem 2000 a koncem roku 2006. Pro tato data dostáváme hodnotu testové statistiky $T(2{,}1) = -0.818$ a p-hodnotu 0.412, tedy na hladině $\alpha = 0.05$ nemůžeme zamítnout H_0, že tato data pochází z binomického rozdělení a mohli bychom pro ně považovat binomické rozdělení za dostatečně dobrý model. Je nutné dodat, že eurozóna měla do vstupu Slovinska 1.1.2007 12 členů, což ztěžuje interpretaci výsledků.

Vezměme nyní jen země (EA12), které opravdu byly v celém časovém období součástí eurozóny (i s Řeckem, které oficiálně vstoupilo v roce 2001) a období, které můžeme považovat z ekonomického hlediska za stabilní. Mějme tedy období od začátku milénia až po konec roku 2007, kdy začala celosvětová finanční krize. Pro tento výběr dat dostáváme hodnotu testové statistiky $T(2{,}1) = 0.672$ a p-hodnotu 0.501. Tato data tedy ještě lépe odpovídají binomickému rozdělení.

Pokud bychom chtěli stejný postup aplikovat na novější data i se státy, které se přidaly do eurozóny (19 států do vstupu Chorvatska v roce 2023), tak zjišťujeme, že nehledě na to, jaké období zvolíme, zamítáme H_0 na hladině $\alpha = 0.05$ s velkou rezervou. Pokud vynecháme období s divokým inflačním vývojem po začátku pandemie COVID-19 v dubnu 2020 a budeme volit různá data začátku měření, tak vždy dostáváme podobné výsledky. Například od doby kdy všech 19 zemí bylo alespoň součástí ERM II (duben 2007) $T(2{,}1) = 30.352$, od konce



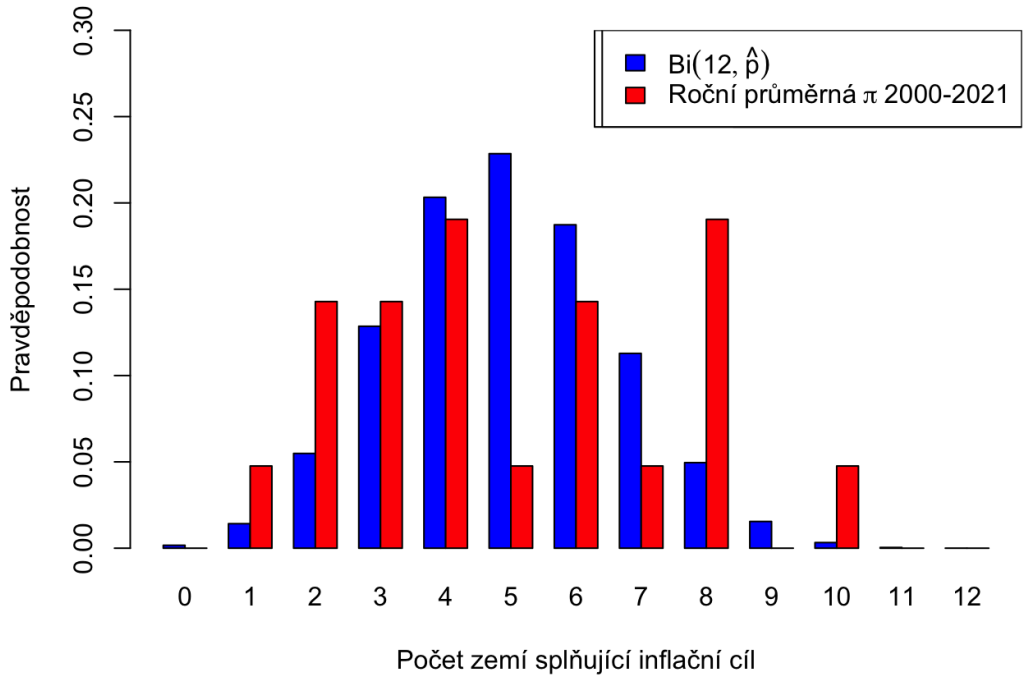
Obrázek 4.1: Počet zemí eurozóny, které splňovaly inflační cíl ECB mezi lety 2000 a 2008.

Velké recese a vstupu Slovenska do eurozóny (leden 2009) $T(2,1) = 21.965$ nebo od vstupu Lotyšska do eurozóny (leden 2014) $T(2,1) = 7.251$. Pro mnohé další období můžeme vidět příslušné p-hodnoty v tabulce 4.1. V tabulce můžeme také vidět rozdíly mezi jednotlivými testy a také rozdíly mezi původními zeměmi eurozóny a zeměmi, které se postupně členy eurozóny staly (bez Chorvatska, které se stalo členem až 1.1.2023).

Dalším problémem může být, že kromě toho, že jsou na sobě závislé jednotlivé státy, tak jsou na sobě závislé i jednotlivé hodnoty π v rámci jednotlivých států, jelikož každá měsíční hodnota meziroční inflace závisí z $\frac{11}{12}$ na vývoji již v předešlých měsících. Tuto závislost můžeme ošetřit tím, že budeme brát průměrnou roční hodnotu π, čímž zároveň zásadně snížíme rozsah našeho výběru. Pro země EA12 a období 2000–2022 dostáváme $T(2,1) = 5.843$, pokud vyřadíme roky 2008 a 2022, ve kterých proběhly krize a žádný stát nesplňoval inflační cíl ECB, máme $T(2,1) = 3.650$ a p-hodnotu < 0.001, tedy ve všech případech zamítáme H_0 na hladině $\alpha = 0.05$. Poslední zmiňovaný případ můžeme vidět na obrázku 4.2. V tabulce 4.1 si také můžeme všimnout, že s menším rozsahem výběru dostáváme vyšší p-hodnoty a s vyšším rozsahem výběru dostáváme hodnoty testových statistik vzdálenější od 0, což svědčí proti nulové hypotéze.

	$T(2,1)$		$T(3,1)$		$\chi^2(4,4)$	
období	EA12	EA19	EA12	EA19	EA12	EA19
2000–2006	0.795	0.610	0.290	0.699	0.590	0.753
2000–2007	0.501	0.093	0.945	0.155	0.641	0.188
2000–2008	0.002	<0.001	0.140	<0.001	0.281	0.002
2009–2022	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
2014–2022	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
2000–2022	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Tabulka 4.1: P- hodnoty testů dobré shody s binomickým rozdělením vzhledem k časovému období a výběru zemí.



Obrázek 4.2: Počet zemí eurozóny, které splňovaly inflační cíl ECB mezi lety 2000 a 2022, počítáno s průměrnou roční inflací v porovnání s empirickou pravděpodobnostní funkcí

Závěr

Cílem práce bylo popsat test dobré shody s binomickým rozdělením, který je založený na faktoriálních momentech.

Nejprve jsme zavedli všechny důležité pojmy, tedy faktoriální momenty společně s definicemi a větami, které jsme následné použili k odvození testu. Následně jsme zavedli binomické rozdělení jako součet nezávislých a stejně rozdělených náhodných veličin s alternativním rozdělením a popsali jsme, co se stane, pokud je porušena nezávislost nebo náhodné veličiny nejsou stejně rozdělené.

V další části jsme už podrobně popsali ideu testu jako podílu druhého faktoriálního momentu a druhé mocniny prvního faktoriálního momentu. Následně test podrobně odvodili, provedli diskuzi, za jakých podmínek může případně docházet k problémům, a určili příslušný kritický obor a p-hodnotu. V navazující sekci jsme tento test zobecnili pro faktoriální momenty libovolného řádu.

Třetí kapitola se věnovala simulacím. Nejprve jsme zkoumali asymptotické vlastnosti testů a následně porovnávali jejich sílu s χ^2 testem dobré shody za různých alternativ. Jako alternativy byla volena taková rozdělení, která vznikají jako součet veličin s alternativním rozdělením, jestliže tyto veličiny nejsou i.i.d. V simulacích se ukázalo, že testy založené na faktoriálních momentech mají oproti χ^2 testu výrazně větší sílu proti většině alternativ anebo se v některých případech chovají podobně. Další výhodou testů, které jsou založené na faktoriálních momentech, je, že odpadá nutnost spojování kategorií tak, aby byly splněny podmínky. Tímto spojováním už zároveň předjímáme, jak by měl test asi dopadnout. Všechna výše uvedená zjištění, ke kterým jsme v průběhu práce došli, ukazují, že testy založené na faktoriálních momentech jsou lepší variantou, co se týče zkoumání dobré shody s binomickým rozdělením oproti χ^2 testu dobré shody.

Na závěr práce jsme na příkladu zemí eurozóny, které splňovaly za určité období inflační cíl určovaný ECB, ukázali praktické využití námi uvedených testů.

Seznam použité literatury

Aleksandrov, B., Weiss, C. H., Jentsch, C. a Faymonville, M. (2022). Novel goodness-of-fit tests for binomial count time series. *Statistics*, **56**(5), 957–990. URL https://doi.org/10.1080/02331888.2022.2134384.

Anděl, J. (2011). *Základy matematické statistiky*. Matfyzpress, Praha. ISBN 978-80-7378-162-0.

Anděl, J. (2019). *Statistické metody*. Matfyzpress, Praha. ISBN 978-80-7378-381-5.

Bank, E. C. (2003). The ECB's monetary policy strategy. URL https://www.ecb.europa.eu/press/pr/date/2003/html/pr030508_2.en.html. Přístup 2024-04-25.

Bank, E. C. (2021). ECB's governing council approves its new monetary policy strategy. URL https://www.ecb.europa.eu/press/pr/date/2021/html/ecb.pr210708~dc78cc4b0d.en.html. Přístup 2024-04-25.

Dupač, V. a a Hušková, M. (2001). Pravděpodobnost a matematická statistika. *Skripta MFF UK Praha. Karolinum, Praha*.

Eurostat (2024a). Hicp - monthly data (annual rate of change) (prc_hicp_manr). URL https://ec.europa.eu/eurostat/web/products-datasets/product?code=prc_hicp_manr. Přístup: 2024-04-25, poslední update datasetu: 2024-04-17.

Eurostat (2024b). Hicp - annual data (average index and rate of change) (prc_hicp_aind). URL https://ec.europa.eu/eurostat/web/products-datasets/product?code=prc_hicp_aind. Přístup 2024-04-25, poslední update datasetu 2024-04-17.

Genz, A. a Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg. ISBN 978-3-642-01688-2.

Hong, Y. (2013). On computing the distribution function for the Poisson binomial distribution. *Computational Statistics & Data Analysis*, **59**, 41–51. URL https://doi.org/10.1016/j.csda.2012.10.006.

Junge, F. (2023). *PoissonBinomial: Efficient Computation of Ordinary and Generalized Poisson Binomial Distributions*. URL https://CRAN.R-project.org/package=PoissonBinomial. R package version 1.2.6.

Kyriakoussis, A., Li, G. a Papadopoulos, A. (1998). On characterization and goodness-of-fit test of some discrete distribution families. *Journal of statistical planning and inference*, **74**(2), 215–228. URL https://doi.org/10.1016/S0378-3758(98)00102-5.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. John Wiley & Sons.

Tang, W. a Tang, F. (2023). The Poisson binomial distribution—old & new. *Statistical Science*, **38**(1), 108–119. URL https://doi.org/10.1214/22-STS852.

Van Der Geest, P. (2005). The binomial distribution with dependent Bernoulli trials. *Journal of Statistical Computation and Simulation*, **75**(2), 141–154. URL https://doi.org/10.1080/0094965042000193224.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press. ISBN 0-471-00710-2.

Vellaisamy, P. a Punnen, A. P. (2001). On the nature of the binomial distribution. *Journal of Applied Probability*, **38**(1), 36–44. URL http://www.jstor.org/stable/3215739.

Wang, Y. H. (1993). On the number of successes in independent trials. *Statistica Sinica*, **3**(2), 295–312. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24304959.

Zvára, K. a Štěpán, J. (2019). *Pravděpodobnost a matematická statistika*. Matfyzpress, Praha. ISBN 978-80-7378-388-4.

Seznam obrázků

Seznam tabulek

A. Přílohy

A.1 χ^2 test dobré shody s multinomickým rozdělením

V této části budeme čerpat z knihy (Anděl, 2011, kapitola 12).

Definice 8 (Multinomické rozdělení)**.** *Nechť $K \geq 2$ a $n > 1$ jsou přirozená číslo a $\boldsymbol{p} = (p_1, \ldots, p_K)^\top$ je vektor konstant splňující $p_k > 0 \quad \forall k$ a $\sum_{k=1}^K p_k = 1$. Náhodný vektor $\boldsymbol{W} = (W_1, \ldots, W_K)^\top$ má multinomické rozdělení $\mathsf{Mult}_K(n, \boldsymbol{p})$ právě když jeho hustota vzhledem k součinové čítací míře na $\mathbb{Z}^K$ je*

$$P[W_1 = w_1, \ldots, W_K = w_K] = \begin{cases} \frac{n!}{w_1! \ldots w_K!} p_1^{w_1} \ldots p_K^{w_K} & \sum_{k=1}^K w_k = n, w_k \geq 0 \quad \forall k, \\ 0 & \textit{jinak.} \end{cases}$$

Podívejme se na asymptotické vlastnosti multinomického rozdělení

Věta 17. *Nechť $\boldsymbol{W} \sim \mathsf{Mult}_K(n, \boldsymbol{p})$, pak*

i)

$$Z_n = \frac{1}{\sqrt{n}} diag\left(\sqrt{\boldsymbol{p}}\right)^{-1} (\boldsymbol{W} - n\boldsymbol{p}) \xrightarrow{D} \mathsf{N}_K \left(\boldsymbol{0}, \boldsymbol{I}_K - \sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}}^\top\right),$$

ii)

$$\boldsymbol{Z}_n^\top \boldsymbol{Z}_n = \sum_{k=1}^K \frac{W_k - np_k}{np_k} \xrightarrow{D} \chi^2_{K-1}.$$

Důkaz. Důkaz lze najít v knize (Anděl, 2011, věty 12.4 a 12.5). □

V našem případě však ještě pravděpodobnosti $p_1, \ldots, p_k$ závisí na neznámém parametru $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_b)^\top$, který odhadneme metodou maximální věrohodnosti a tento odhad označíme $\widehat{\boldsymbol{\theta}}_n$. Test dobré shody s multinomickým rozdělením při neznámém parametru $\boldsymbol{\theta}$ uvažuje hypotézy

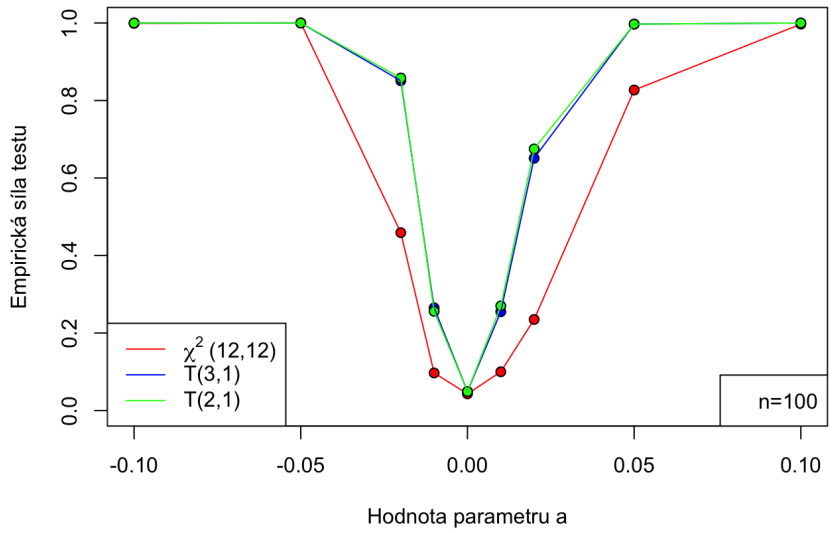
$$\begin{aligned} H_0 &: \exists \boldsymbol{\theta} \in \Theta \quad \boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\theta}), \\ H_1 &: \forall \boldsymbol{\theta} \in \Theta \quad \boldsymbol{p} \neq \boldsymbol{p}(\boldsymbol{\theta}) \end{aligned}$$

a za platnosti H_0 společně s předpoklady regularity, viz (Anděl, 2019, věta 10.4.)

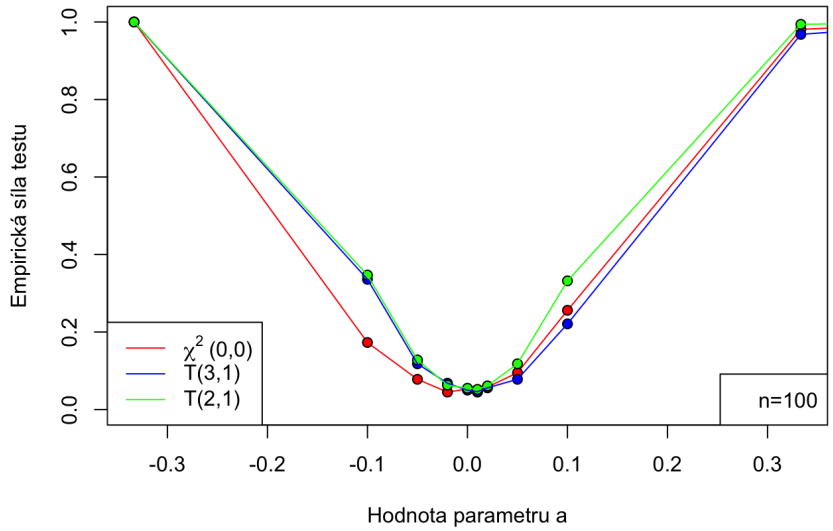
$$\chi^2 \equiv \sum_{k=1}^K \frac{\left(W_k - np_k(\widehat{\boldsymbol{\theta}}_n)\right)^2}{np_k(\widehat{\boldsymbol{\theta}}_n)} \xrightarrow[n\to\infty]{D} \chi^2_{K-b-1},$$

kde K je počet kategorií a b je počet odhadovaných parametrů.

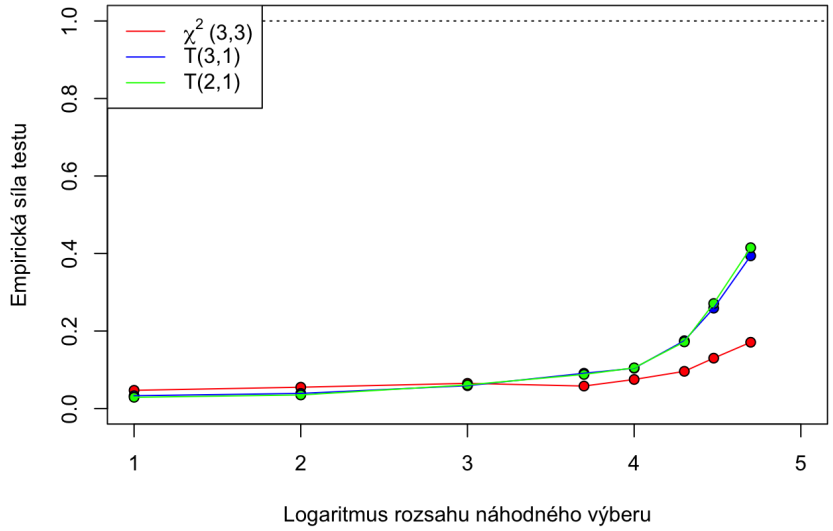
A.2 Obrázky



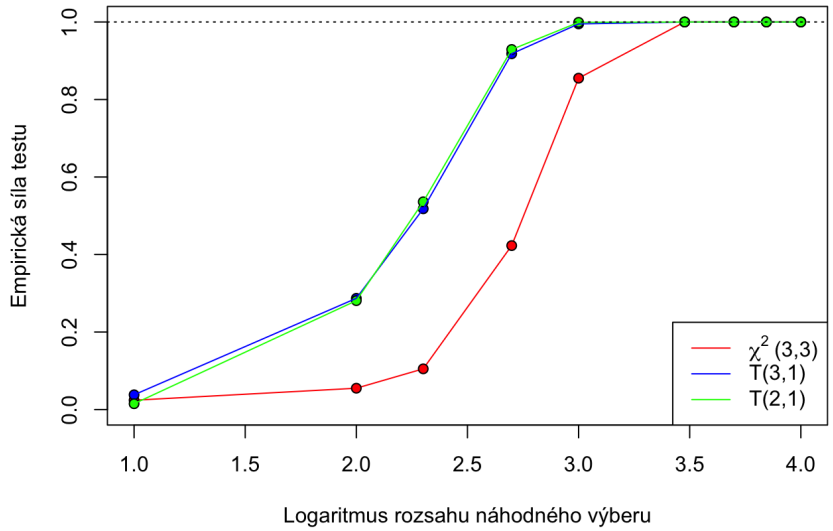
Obrázek A.1: Síla testů vůči alternativě závislých náhodných veličin $m = 30$ v závislosti na volbě parametru a v matici Σ_a.



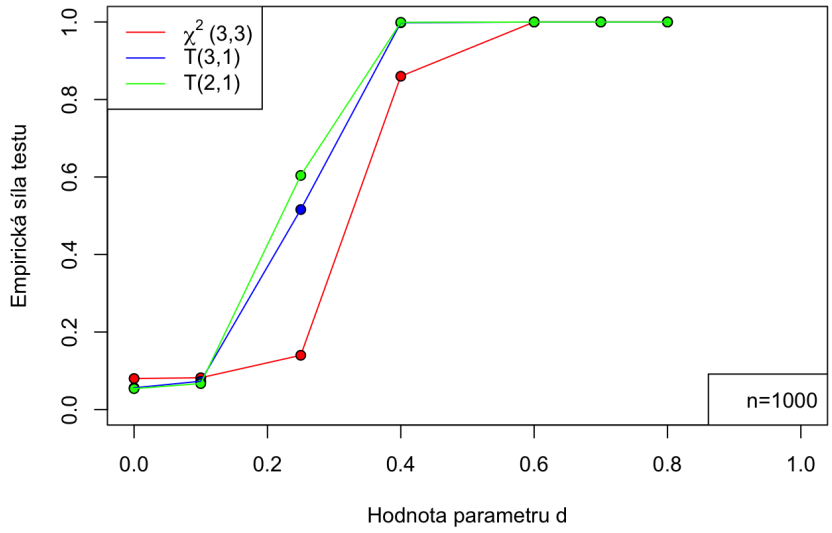
Obrázek A.2: Síla testů vůči alternativě závislých náhodných veličin $m = 4$ v závislosti na volbě parametru a v matici Σ_a.



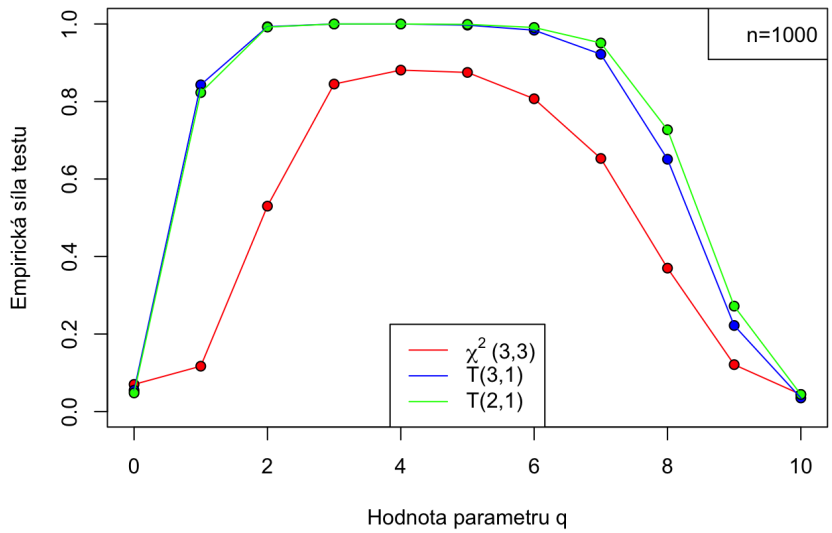
Obrázek A.3: Síla testů proti PBi(10, $\boldsymbol{p} + \boldsymbol{\delta}$), kde $p = 0.5$, $q = 5$ a $d = 0.1$ v závislosti na rozsahu výběru.



Obrázek A.4: Síla testů proti PBi(10, $\boldsymbol{p} + \boldsymbol{\delta}$), $p = 0.5$, $d = 0.4$, $q = 5$ v závislosti na rozsahu výběru.



Obrázek A.5: Síla testů proti PBi(10, $\boldsymbol{p} + \boldsymbol{\delta}$), $p = 0.1$, $q = 5$ v závislosti na paramteru $d \in \{0{,}0.1{,}0.25{,}0.4{,}0.6{,}0.7{,}0.8\}$.



Obrázek A.6: Síla testů proti PBi(10, $\boldsymbol{p} + \boldsymbol{\delta}$), $p = 0.50$, $d = 0.40$ v závislosti na paramteru $q \in \{0{,}1{,}2{,}3{,}4{,}5{,}6{,}7{,}8{,}9{,}10\}$.