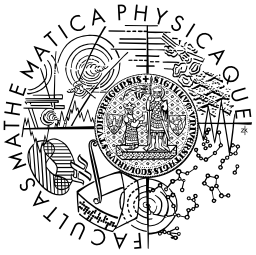

MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Noemi Kuželová

Transformace stabilizující rozptyl

Katedra pravděpodobnosti a matematické statistiky

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

Praha 2019

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

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

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

Ráda bych poděkovala svému vedoucímu bakalářské práce doc. Ing. Marku Omelkovi, Ph.D. za vstřícný přístup, veškerou ochotu i pomoc při vypracovávání této práce.

Název práce: Transformace stabilizující rozptyl

Autor: Noemi Kuželová

Katedra: Katedra pravděpodobnosti a matematické statistiky

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

Abstrakt: Mnohdy zkoumáme data, jejichž výběrový průměr konverguje k normálnímu rozdělení, jehož rozptyl však obecně závisí na neznámém parametru. K tomu, abychom se této závislosti zbavili, lze někdy využít metodu tak zvané transformace stabilizující rozptyl. Tato práce nejprve metodu detailně vysvětlí a najde obecný postup, jak vhodné transformace hledat. Poté se zaměří na data pocházející z Poissonova a binomického rozdělení s neznámými parametry. Pro tato data najde transformace, jež stabilizují (asymptotický) rozptyl, a porovná je s ještě „vylepšenými“ transformacemi z článku Anscombe (1948). Právě tvaru těchto transformací je věnována většina práce. Nakonec na simulaci pro výběr z Poissonova rozdělení ukážeme, že je opravdu vhodné tuto metodu využívat a srovnáme odvozenou transformaci s její Anscombeovou verzí.

Klíčová slova: transformace stabilizující rozptyl, delta metoda, Poissonovo rozdělení, binomické rozdělení

Title: Variance stabilizing transformation

Author: Noemi Kuželová

Department: Department of Probability and Mathematical Statistics

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

Abstract: We often examine data whose sample mean converges to a normal distribution, but the variance generally depends on an unknown parameter. To get rid of this dependence, we can sometimes use the so-called variance-stabilizing transformation method. Firstly, this thesis explains the method in detail and finds a general procedure to find suitable transformations. Then it will focus on data from Poisson and binomial distributions with unknown parameters. For these data, it finds transformations that stabilize (asymptotic) variance, and compares them with the "improved" transforms from the article Anscombe (1948). Most of the thesis is devoted to the shape of these transformations. Finally, we show in the Poisson distribution simulation that it is really appropriate to use this method and compare the derived transformation with its Anscombe version.

Keywords: variance-stabilizing transformation, delta method, Poisson distribution, binomial distribution

Obsah

Úvod

Mnoho statistických metod využívá předpoklad, že data, která analyzují, pochází z normálního rozdělení s konstantním rozptylem, jež nezávisí na střední hodnotě. Leckdy platí, že rozdělení výběrového průměru k normálnímu rozdělení konverguje, ovšem jeho asymptotický rozptyl mnohdy závisí na neznámém parametru. Jednou z možností, jak požadavku na konstantní rozptyl dostát, je tzv. transformace stabilizující rozptyl, tedy metoda, která výběrový průměr z dat transformuje s cílem zbavit rozptyl závislosti na střední hodnotě. Existuje více možností, jakou proto vhodnou transformaci zvolit. Tato práce se zabývá zejména transformacemi, které jsou uvedeny ve článku Anscombe (1948), a to pro data z Poissonova a binomického rozdělení.

V první kapitole detailněji vysvětlíme, v čem transformace stabilizující rozptyl spočívá a odvodíme obecný postup, jak ji hledat. Ve druhé kapitole se zaměříme na situaci, kdy data pocházejí z Poissonova rozdělení s neznámou střední hodnotou. Poté, co odvodíme tvar transformace stabilizující rozptyl klasickým přístupem, zaměříme se na Anscombeem navrhovanou transformaci, jejímž vlastnostem se budeme podrobněji věnovat. Ve třetí kapitole se budeme zaobírat transformacemi pro binomické rozdělení, ne všechny výpočty však již budeme provádět tak poctivě jako v případě Poissonova rozdělení. Poslední kapitolu využijeme k ilustraci praktického použití transformací stabilizujících rozptyl, a to konkrétně k intervalovým odhadům. Na Poissonově rozdělení budeme pomocí simulací zkoumat, jak výhodné je využití metody transformace stabilizující rozptyl.

Mým cílem v této práci je teoreticky popsat metodu transformace stabilizující rozptyl. Dále podrobně odvodit již zmíněné Anscombeovy transformace, a to pro náhodné výběry, jež pochází z Poissonova a binomického rozdělení. A nakonec ilustrovat využití této metody, k čemuž jsem zvolila simulaci pokrývání parametru intervaly spolehlivosti. Tato simulace navíc porovná, zda je výhodnější využít tradiční či Anscombeovu transformaci stabilizující rozptyl.

1. Co je transformace stabilizující rozptyl

Pojďme si nejprve vysvětlit, co název této práce znamená.

1.1 Stabilizace asymptotického rozptylu výběrového průměru

Mějme $X_1, \ldots, X_n$ náhodný výběr z takového rozdělení, jehož hustota závisí na parametru θ. Obvykle je na tomto parametru závislá jak střední hodnota $\mathbb{E}\, X_1$ (předpokládáme, že je reálná), tak i rozptyl $\operatorname{var} X_1$ (ten uvažujme konečný nenulový). Pro jednoduchost předpokládejme, že přímo $\mathbb{E}\, X_1 = \theta$.

Mnoho rozdělení lze při velkém rozsahu výběru aproximovat normálním rozdělením, což popisují centrální limitní věty. My si vystačíme s Lévy-Lindebergovou verzí centrální limitní věty pro nezávislé a stejně rozdělené posloupnosti náhodných veličin.

Věta 1 (Centrální limitní věta)**.** *Buď $\{X_n\}_{n=1}^{\infty}$ posloupnost reálných náhodných nezávislých veličin ze stejného rozdělení se střední hodnotou $\mathbb{E}\, X_1$ a rozptylem $\operatorname{var} X_1 \in (0,\, \infty)$. Potom platí $\sqrt{n}(\bar{X}_n - \mathbb{E}\, X_1) \xrightarrow[n\to\infty]{d} N(0,\, \operatorname{var} X_1)$.*

(Anděl, 1985, kap. X, věta 6)

V našem případě, kdy střední hodnotu uvažujeme jako neznámý parametr θ, označme $\sigma^2(\theta) = \operatorname{var} X_1$ rozptyl závislý na parametru. Potom platí vztah

$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow[n\to\infty]{d} N(0,\, \sigma^2(\theta)).$$

Cílem metody zvané transformace stabilizující rozptyl je najít funkci g takovou, aby po její aplikaci na výběrový průměr jeho asymptotický rozptyl již na parametru θ nezávisel, nebo se alespoň s rostoucím rozsahem výběru blížil ke konstantě. Klasický způsob, jak takovou transformaci hledat, je pomocí delta metody. Ani tu nebudeme uvádět v její nejobecnější podobě, bude nám stačit v následujícím tvaru.

Věta 2 (Delta metoda)**.** *Buď $X_1, \ldots, X_n$ náhodný výběr z rozdělení se střední hodnotou θ a rozptylem $\sigma^2(\theta) \in (0,\, \infty)$. Nechť $\sqrt{n}(\bar{X}_n - \theta) \xrightarrow[n\to\infty]{d} N(0,\, \sigma^2(\theta))$. Buď zobrazení $g : \mathbb{R} \to \mathbb{R}$ spojitě diferencovatelné na okolí θ. Potom*

$$\sqrt{n}(g(\bar{X}_n) - g(\theta)) \xrightarrow[n\to\infty]{d} N\left(0,\, [g'(\theta)]^2\sigma^2(\theta)\right).$$

(Anděl, 2007, speciální jednorozměrný případ věty B.7 pro $T_n = \bar{X}_n$)

Jelikož chceme konvergenci k normálnímu rozdělení s konstantním, obvykle jednotkovým, rozptylem, tak díky delta metodě nalezneme takové g, které řeší

diferenciální rovnici $[g'(\theta)]^2\sigma^2(\theta) = 1$, již vyřešíme následovně:

$$\begin{aligned} [g'(\theta)]^2\sigma^2(\theta) &= 1 \\ g'(\theta)\sigma(\theta) &= 1 \\ g'(\theta) &= \frac{1}{\sigma(\theta)} \\ g(\theta) &= \int \frac{1}{\sigma(\theta)} d\theta. \end{aligned} \tag{1.1}$$

1.2 Transformace náhodné veličiny

Delta metoda nám dává dobrý návod, jak hledat transformaci stabilizující rozptyl, ke kterému rozdělení výběrového průměru v distribuci konverguje. Našim cílem je však tuto myšlenku posunout a hledat stabilizující transformace přímo pro zkoumanou náhodnou veličinu.

Buď X náhodná veličina, jejíž hustota, tak jako výše, závisí na parametru θ a má střední hodnotu $\mathbb{E}\, X = \theta \in \mathbb{R}$ a na θ závislý nenulový rozptyl $\sigma^2(\theta)$.

Klasický přístup, jak nalézt vhodnou funkci g takovou, aby $\sigma^2(\theta)$ byl pokud možno konstantní, předpokládá blízkost X a její střední hodnoty. Jelikož za X často budeme volit výběrový průměr, který podle Kolmogorovova zákonu velkých čísel (jež je jako věta 3 uveden níže), konverguje ke střední hodnotě, je tento požadavek oprávněný. Budeme-li předpokládat spojitou diferencovatelnost g na okolí θ, můžeme k jejímu nalezení využít Taylorův rozvoj 1. řádu funkce g v bodě θ. Po zanedbání ostatních členů Taylorova polynomu odhadneme

$$g(X) \doteq g(\theta) + g'(\theta)(X - \theta).$$

Potom z linearity střední hodnoty a vlastností rozptylu

$$\mathbb{E}\, g(X) \doteq g(\theta), \ \operatorname{var} g(X) \doteq [g'(\theta)]^2\sigma^2(\theta).$$

Stejně jako když jsme zkoumali konvergenci k normálnímu rozdělení, i zde chceme najít g takovou, aby byl výraz $[g'(\theta)]^2\sigma^2(\theta)$ roven jedné. Stejným výpočtem jako v (1.1) dostáváme obecný předpis hledané transformace, který je rovný

$$g(\theta) = \int \frac{1}{\sigma(\theta)} d\theta. \tag{1.2}$$

Takto nalezená g ovšem není jediná funkce, která nám může pomoci. Pro některá diskrétní rozdělení J. F. Anscombe uvažoval Taylorův rozvoj vyššího řádu, a tak ve svém článku navrhl jiné transformace, které vycházejí ze vzorce (1.2), ovšem v některých případech se ukazují být vhodnější (viz Anscombe, 1948).

Věta 3 (Kolmogorovův zákon velkých čísel)**.** *Nechť $X_1, \ldots, X_n$ je posloupnost stejně rozdělených nezávislých náhodných veličin s konečnou střední hodnotou, pak*

$$\bar{X_n} \xrightarrow[n\to\infty]{s.j.} \mathbb{E}\, X_1.$$

(Anděl, 2003, věta B.4)

2. Poissonovo rozdělení

Nejprve budeme zkoumat vhodnou transformaci pro Poissonovo rozdělení.

2.1 Charakterizace Poissonova rozdělení

Definice 1. *Buď X náhodná veličina, pro níž platí $\mathsf{P}[X \in \mathbb{N} \cup \{0\}] = 1$. Řekneme, že X má **Poissonovo rozdělení** s parametrem $\lambda > 0$, $X \sim Po(\lambda)$, jestliže $\mathsf{P}[X = k] = \frac{\lambda^k}{k!}e^{-\lambda}$, $k \in \mathbb{N} \cup \{0\}$.*

Střední hodnota i rozptyl Poissonova rozdělení jsou rovny λ. Další důležitou vlastností, která nás bude u tohoto rozdělení zajímat, je jeho vytvořující funkce.

Definice 2. *Reálnou funkci reálné proměnné $M_X(t) = \mathbb{E}\, e^{tX}$ nazveme **momentovou vytvořující funkcí** náhodné veličiny X.*

Spočteme vytvořující funkci $X \sim Po(\lambda)$:

$$M_X(t) = \mathbb{E}\, e^{tX} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} e^{tk} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda\left(e^t - 1\right)}. \tag{2.1}$$

2.2 Transformace dle delta metody

Začněme zkoumáním asymptotického chování výběrového průměru z náhodného výběru $X_1, \ldots, X_n$ z Poissonova rozdělení s neznámou střední hodnotou $\lambda > 0$. Z centrální limitní věty (věta 1) víme, že $\sqrt{n}(\bar{X}_n - \lambda) \xrightarrow[n\to\infty]{d} N(0,\ \lambda)$. Využitím delta metody (věta 2) a postupu (1.1) popsaného výše nalezneme vhodnou transformaci, jejíž aplikací získáme asymptoticky jednotkový rozptyl:

$$g_0(\lambda) = \int \frac{1}{\sqrt{\lambda}} d\mu = 2\sqrt{\lambda} + const,\ const \in \mathbb{R},$$

tedy

$$\sqrt{n}\left(2\sqrt{\bar{X}_n} - 2\sqrt{\lambda}\right) \xrightarrow[n\to\infty]{d} N\,(0{,}1)\,. \tag{2.2}$$

2.3 Anscombeova transformace

Nyní se podíváme na transformaci Poissonova rozdělení uvedenou v již dříve zmíněném článku Anscombe (1948). Tou je funkce $g(t) = 2\sqrt{t + c}$ pro vhodné reálné c.

2.3.1 Motivace

Podívejme se, co dostaneme, pokud transformaci $g_0(t) = 2\sqrt{t}$ aplikujeme dle delta metody:

$$\sqrt{n}\left(2\sqrt{\bar{X}_n} - 2\sqrt{\lambda}\right) \xrightarrow[n\to\infty]{d} N(0,1),$$

neboli

$$\sqrt{n}\left(\sqrt{\bar{X}_n} - \sqrt{\lambda}\right) \xrightarrow[n\to\infty]{d} N\left(0,\frac{1}{4}\right).$$

Úpravou levé části výrazu výše dostáváme

$$\sqrt{n}\left(\sqrt{\bar{X}_n} - \sqrt{\lambda}\right) = \sqrt{n}\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i} - \sqrt{\lambda}\right) = \sqrt{\sum_{i=1}^{n} X_i} - \sqrt{n\lambda}.$$

Jelikož součet n náhodných veličin z Poissonova rozdělení má opět Poissonovo rozdělení, a to s n-násobnou střední hodnotou, označíme $X^{(n)} = \sum_{i=1}^{n} X_i$ a platí $X^{(n)} \sim Po(n\,\lambda)$, tedy

$$\left(\sqrt{X^{(n)}} - \sqrt{\mathbb{E}\,X^{(n)}}\right) \xrightarrow[n\to\infty]{d} N\left(0,\frac{1}{4}\right).$$

Anscombe navrhuje upravit transformaci z pouhé odmocniny náhodné veličiny na $g(t) = \sqrt{t+c}$, $c \in \mathbb{R}$, všimneme si, že g je zřejmě spojitě diferencovatelná na intervalu $(-c, \infty)$ a dále má smysl uvažovat pouze c nezáporná, jelikož X s nenulovou pravděpodobností nabývá nulové hodnoty. Přičemž by transformace g měla být vhodná pro náhodnou veličinu z Poissonovo rozdělení s velkou střední hodnotou. Výpočet výše dává tomuto předpokladu dobrou motivaci.

2.3.2 Odvození vhodné konstanty c

Mějme náhodnou veličinu $X \sim Po(\lambda)$, $\lambda > 0$ velké. Buď g funkce X daná předpisem $g(t) = \sqrt{t+c}$, $c \geq 0$. V příští sekci ukážeme, že pro takto definovanou transformaci platí

$$\operatorname{var} g(X) = \frac{1}{4} + \frac{-8c+3}{32\lambda} + \frac{32c^2 - 52c + 17}{128\lambda^2} + O(\lambda^{-2,5}),$$

a proto, aby byl tento rozptyl co nejstabilnější (vůči neznámému parametru λ), je při velkém λ nejvhodnější zvolit $c = \frac{3}{8}$. Potom

$$\operatorname{var} g(X) = \frac{1}{4} + \frac{1}{64\lambda^2} + O(\lambda^{-2,5}).$$

Pro velké hodnoty λ tudíž $\operatorname{var} g(X)$ konverguje k $\frac{1}{4}$.

Notaci velké O budeme chápat v následujícím smyslu.

Definice 3. *Buďte f, g reálné funkce. Řekneme, že $f(x) = O(g(x))$, jestliže existuje $A > 0$ takové, že $\lim_{x\to\infty} \frac{|f(x)|}{|g(x)|} \leq A$.*

2.4 Rozptyl Anscombeovy transformace

Jelikož určení rozptylu transformované náhodné veličiny je poměrně zdlouhavé, rozdělíme výpočet do několika menších částí.

2.4.1 Taylorův polynom $g(X)$

K hledání vhodné konstanty c využijeme Taylorův rozvoj funkce g. Označme $T_s^{g,\lambda}$ Taylorův polynom funkce g v bodě λ řádu s. Potom pro $X \in (-c, \infty)$ a g dostatečně hladkou, jest

$$T_s^{g,\lambda}(X) = \sum_{k=0}^{s} \frac{1}{k!} g^{(\mathbf{k})}(\lambda)(X-\lambda)^k,\ s \in \mathbb{N}.$$

Podívejme se, jak vypadají derivace funkce g:

$$\begin{aligned}
g'(t) &= \frac{1}{2}(t+c)^{-\frac{1}{2}} \\
g''(t) &= \frac{1}{2}\frac{-1}{2}(t+c)^{\frac{1}{2}-1} \\
g^{(3)}(t) &= \frac{1}{2}\frac{-1}{2}\frac{-3}{2}(t+c)^{\frac{1}{2}-2} \\
&\vdots \\
g^{(k)}(t) &= \frac{1\cdot(-1)\cdot(-3)\cdots\left(\frac{1}{2}-(k-1)\right)}{2^{k-1}}(t+c)^{\frac{1}{2}-k} \\
&= \frac{1\cdot(-1)\cdot(-3)\cdots(3-2k)}{2^k}(t+c)^{\frac{1}{2}-k},\ k \in \mathbb{N}.
\end{aligned}$$

Po dosazení do Taylorovy formule:

$$T_s^{g,\lambda}(X) = \sum_{k=0}^{s} \frac{1\cdot(-1)\cdot(-3)\cdots(3-2k)}{2^k k!} \frac{(X-\lambda)^k}{(\lambda+c)^{k-\frac{1}{2}}}$$

Pro zjednodušení zápisu označme $a_0 = 1, a_k = \frac{1\cdot(-1)\cdot(-3)\cdots(3-2k)}{2^k k!}$ pro $k \geq 1$, $m = \lambda + c$ a $T = X - \lambda$. Neboli

$$T_s^{g,\lambda}(X) = \sqrt{m} \sum_{k=0}^{s} a_k \left(\frac{T}{m}\right)^k,$$

budeme využívat

$$T_5^{g,\lambda}(X) = \sqrt{m} + \frac{T}{2\sqrt{m}} - \frac{T^2}{8m^{\frac{3}{2}}} + \frac{T^3}{16m^{\frac{5}{2}}} - \frac{5T^4}{128m^{\frac{7}{2}}} + \frac{7T^5}{256m^{\frac{9}{2}}}.$$

2.4.2 Zbytkový člen rozvoje

Podíváme se, jak je Taylorův polynom „vzdálený“ od $g(X)$. Jest

$$g(X) = T_{s-1}^{g,\lambda}(X) + R_6,$$

konkrétně pro s=5:

$$g(X) = \sqrt{m} + \frac{T}{2\sqrt{m}} - \frac{T^2}{8m^{\frac{3}{2}}} + \frac{T^3}{16m^{\frac{5}{2}}} - \frac{5T^4}{128m^{\frac{7}{2}}} + \frac{7T^5}{256m^{\frac{9}{2}}} + R_6.$$

Velikost zbytku R_6 se budeme snažit omezit, abychom ukázali, že Taylorův polynom náhodnou veličinu $g(X)$ dobře aproximuje. Nejdůležitější pro nás při tom bude, jaké mocniny nabývá tento výraz v λ.

Platí $g(X) = \sqrt{X+c} = \sqrt{m}\sqrt{1+\frac{T}{m}}$. Výpočet horní meze rozdělíme na dva případy. Za prvé uvážíme T taková, že $\left|\frac{T}{m}\right| < 1$. Potom využijeme vztah

$$\begin{aligned} (1+x)^a &= \sum_{k=0}^{\infty} \binom{a}{k} x^k, \text{ pro } x \in (-1,1), \\ \text{kde } \binom{a}{k} &= \frac{a(a-1)(a-2)\dots(a-k+1)}{k!}. \end{aligned} \tag{2.3}$$

Dosazením do vzorce (2.3) dostaneme

$$\begin{aligned} R_s = g(X) - T_{s-1}^{g,\lambda}(X) &= \sqrt{m}\sqrt{1+\frac{T}{m}} - \sqrt{m}\sum_{k=0}^{s-1} a_k \left(\frac{T}{m}\right)^k \\ &= \sqrt{m}\left[\sum_{k=0}^{\infty} \frac{1\cdot(-1)\cdot(-3)\cdots(3-2k)}{2^k k!}\left(\frac{T}{m}\right)^k \right. \\ &\quad \left. - \sum_{k=0}^{s-1} \frac{1\cdot(-1)\cdot(-3)\cdots(3-2k)}{2^k k!}\left(\frac{T}{m}\right)^k\right] \\ &= \sqrt{m}\sum_{k=s}^{\infty} \frac{1\cdot(-1)\cdot(-3)\cdots(3-2k)}{2^k k!}\left(\frac{T}{m}\right)^k. \end{aligned}$$

Jelikož nyní předpokládáme, že $\frac{T}{m} < 1$, můžeme zbytkový člen omezit

$$\begin{aligned} |R_s| &\le \sqrt{m}\sum_{k=s}^{\infty} \frac{1\cdot 1\cdot 3\cdots|3-2k|}{2^k k!}\left|\frac{T}{m}\right|^k \le \sqrt{m}\left|\frac{T}{m}\right|^s \sum_{k=s}^{\infty} \frac{1\cdot 1\cdot 3\cdots|3-2k|}{2^k k!} \\ &\le \sqrt{m}\left|\frac{T}{m}\right|^s \sum_{k=1}^{\infty} \frac{1\cdot 1\cdot 3\cdots|3-2k|}{2^k k!}. \end{aligned} \tag{2.4}$$

Konvergenci řady výše zaručuje Raabeovo kritérium.

Věta 4 (Raabeovo limitní kritérium)**.** *Nechť $\sum_{k=0}^{\infty} b_k$ řada s kladnými členy, nechť existuje $\lim_{k\to\infty} k\left(1-\frac{b_{k+1}}{b_k}\right) = q, q \in \mathbb{R}$. Jestliže $q > 1$, pak řada $\sum_{k=1}^{\infty} b_k$ konverguje. Je-li $q < 1$, pak řada $\sum_{k=0}^{\infty} b_k$ diverguje.*

(Dížková, 2012, věta 1.14)

Pojďme tedy ukázat, že řada $\sum_{k=1}^{\infty} \frac{1\cdot 1\cdot 3\cdots|3-2k|}{2^k k!}$ splňuje podmínku konvergence z věty 4:

$$\begin{aligned} \lim_{k\to\infty} k\left(1-\frac{b_{k+1}}{b_k}\right) &= \lim_{k\to\infty} k\left(1-\frac{\frac{1\cdot 1\cdot 3\cdots|3-2(k+1)|}{2^{k+1}(k+1)!}}{\frac{1\cdot 1\cdot 3\cdots|3-2k|}{2^k k!}}\right) = \lim_{k\to\infty} k\frac{2(k+1)-|1-2k|}{2(k+1)} \\ &= \lim_{k\to\infty} k\frac{3}{2k+2} = \frac{3}{2} > 1. \end{aligned}$$

Tudíž existuje konečná funkce H závislá pouze na s, že $\sum_{k=1}^{\infty} \frac{1\cdot 1\cdot 3\cdots|3-2k|}{2^k k!} \leq H(s)$ pro všechna s přirozená, a tedy z odhadu (2.4) můžeme zbytkový člen omezit:

$$|R_s| \leq H(s)\frac{|T|^s}{m^{s-\frac{1}{2}}} \text{ pro } \forall s \in \mathbb{N}. \tag{2.5}$$

V druhém případě se podíváme na T splňující $\left|\frac{T}{m}\right| \geq 1$ (a zároveň $|T| > 0$, jinak by platilo $X \leq -c$, což nepatří do definičního oboru). Tudíž pro velká m dostáváme podmínku $\frac{T}{m} \geq 1$. Pro ně můžeme využít následující větu o Lagrangeově tvaru zbytku.

Věta 5 (Lagrangeův tvar zbytku)**.** *Nechť f je reálná funkce a $T_{s-1}^{f,a}$ Taylorův polynom funkce f v bodě $a \in \mathbb{R}$ řádu $s \in \mathbb{N}$. Nechť f je spojitá na uzavřeném intervalu J krajními body a, x. Nechť navíc v každém jeho vnitřním bodě má f vlastní s-tou derivaci. Potom existuje ξ vnitřní bod J takový, že*

$$f(x) - T_{s-1}^{f,a}(x) = \frac{f^{(s)}(\xi)}{s!}(x-a)^s.$$

(Kopáček, 2004, věta 5.23)

Získáváme tak horní mez pro R_s:

$$\begin{aligned} R_s = g(X) - T_{s-1}^{g,\lambda}(X) &= \frac{g^{(s)}(\xi)}{(s)!}(X-\lambda)^s \\ &= \frac{1\cdot(-1)\cdot(-3)\cdots(3-2s)}{2^s s!}\xi^{\frac{1}{2}-s}(X-\lambda)^s. \end{aligned}$$

Ještě jednou rozlišíme dva případy. Nejprve buď $\xi \geq \lambda + c$, potom

$$\begin{aligned} |R_s| &= \left|\frac{1\cdot(-1)\cdot(-3)\cdots(3-2s)}{2^s s!}\xi^{\frac{1}{2}-s}(X-\lambda)^s\right| \\ &\leq \left|\frac{1\cdot(-1)\cdot(-3)\cdots(3-2s)}{2^s s!}(\lambda+c)^{\frac{1}{2}-s}(X-\lambda)^s\right| \\ &= \left|\sqrt{m}\frac{1\cdot(-1)\cdot(-3)\cdots(3-2s)}{2^s s!}\left(\frac{T}{m}\right)^s\right|. \end{aligned}$$

Pro $\xi \in (\lambda\,, \lambda + c)$ platí

$$\begin{aligned} |R_s| &= \left|\frac{1\cdot(-1)\cdot(-3)\cdots(3-2s)}{2^s s!}\xi^{\frac{1}{2}-s}(X-\lambda)^s\right| \\ &\leq \left|\frac{1\cdot(-1)\cdot(-3)\cdots(3-2s)}{2^s s!}(\lambda+c)^{\frac{1}{2}}\frac{1}{(\lambda+c)^s}\frac{(\lambda+c)^s}{\xi^s}(X-\lambda)^s\right| \\ &= \left|\frac{1\cdot(-1)\cdot(-3)\cdots(3-2s)}{2^s s!}(\lambda+c)^{\frac{1}{2}}\frac{1}{(\lambda+c)^s}\left(1+\frac{c}{\lambda}\right)^s(X-\lambda)^s\right| \\ &= \left|\frac{1\cdot(-1)\cdot(-3)\cdots(3-2s)}{2^s s!}(\lambda+c)^{\frac{1}{2}}\frac{1}{(\lambda+c)^s}2^s(X-\lambda)^s\right| \\ &= \left|\sqrt{m}\frac{1\cdot(-1)\cdot(-3)\cdots(3-2s)}{s!}\left(\frac{T}{m}\right)^s\right|. \end{aligned}$$

Využili jsme odhadu $\frac{c}{\lambda} < 1$, neboť c je pevná nezáporná konstanta, zatímco λ uvažuje dostatečně velké, a proto lze předpokládat, že je větší než c. Celkově tudíž pro takové náhodné T, že $\frac{T}{m} \geq 1$, platí

$$|R_s| < \frac{1 \cdot 1 \cdot 3 \cdots |3-2s|}{s!} \frac{T^s}{m^{s-\frac{1}{2}}}.$$

Aby se nám se zbytkem později lépe pracovalo, označme ještě funkci

$$G(s) = \max\{\frac{1 \cdot 1 \cdot 3 \cdots |3-2s|}{s!}, H(s)\}.$$

Potom pro všechna T platí

$$|R_s| \leq G(s) \frac{|T|^s}{m^{s-\frac{1}{2}}}. \tag{2.6}$$

2.4.3 Výpočet rozptylu $g(X)$

S využitím Taylorova rozvoje spočítáme rozptyl transformované veličiny $g(X)$.

$$\begin{aligned}
\operatorname{var} g(X) = & \operatorname{var}\left(\sqrt{m} + \frac{T}{2\sqrt{m}} - \frac{T^2}{8m^{\frac{3}{2}}} + \frac{T^3}{16m^{\frac{5}{2}}} - \frac{5T^4}{128m^{\frac{7}{2}}} + \frac{7T^5}{256m^{\frac{9}{2}}} + R_6\right) \\
= & \frac{\operatorname{var} T}{4m} + \frac{\operatorname{var} T^2}{64m^3} + \frac{\operatorname{var} T^3}{256m^5} + \frac{25 \operatorname{var} T^4}{128^2 m^7} + \frac{49 \operatorname{var} T^5}{256^2 m^9} \\
& - 2\frac{\operatorname{cov}(T,T^2)}{2 \cdot 8m^2} + 2\frac{\operatorname{cov}(T,T^3)}{2 \cdot 16m^3} - 2\frac{5 \operatorname{cov}(T,T^4)}{2 \cdot 128m^4} + 2\frac{7 \operatorname{cov}(T,T^5)}{2 \cdot 256m^5} \\
& - 2\frac{\operatorname{cov}(T^2,T^3)}{8 \cdot 16m^4} + 2\frac{5 \operatorname{cov}(T^2,T^4)}{8 \cdot 128m^5} - 2\frac{7 \operatorname{cov}(T^2,T^5)}{8 \cdot 256m^6} - 2\frac{5 \operatorname{cov}(T^3,T^4)}{16 \cdot 128m^5} \\
& + 2\frac{7 \operatorname{cov}(T^3,T^5)}{16 \cdot 256m^7} - 2\frac{5 \cdot 7 \operatorname{cov}(T^4,T^5)}{128 \cdot 256m^7} + \operatorname{var} R_6 + \sum_{k=1}^{5} 2a_k \frac{\operatorname{cov}(T^k, R_6)}{m^{k-\frac{1}{2}}} \\
= & \frac{\operatorname{var} T}{4m} + \frac{\operatorname{var} T^2}{64m^3} + + \frac{\operatorname{var} T^3}{256m^5} + \frac{25 \operatorname{var} T^4}{128^2 m^7} + \frac{49 \operatorname{var} T^5}{256^2 m^9} \\
& - \frac{\operatorname{cov}(T,T^2)}{8m^2} + \frac{\operatorname{cov}(T,T^3)}{16m^3} - \frac{5 \operatorname{cov}(T,T^4)}{128m^4} + \frac{7 \operatorname{cov}(T,T^5)}{256m^5} \\
& - \frac{\operatorname{cov}(T^2,T^3)}{4 \cdot 16m^4} + \frac{5 \operatorname{cov}(T^2,T^4)}{4 \cdot 128m^5} - \frac{7 \operatorname{cov}(T^2,T^5)}{4 \cdot 256m^6} - \frac{5 \operatorname{cov}(T^3,T^4)}{8 \cdot 128m^5} \\
& + \frac{7 \operatorname{cov}(T^3,T^5)}{8 \cdot 256m^7} - \frac{35 \operatorname{cov}(T^4,T^5)}{64 \cdot 256m^7} + \operatorname{var} R_6 + \sum_{k=1}^{5} 2a_k \frac{\operatorname{cov}(T^k, R_6)}{m^{k-\frac{1}{2}}}
\end{aligned}$$

Abychom mohli snáze určit momenty náhodné veličiny T, spočteme její vytvořující funkci: $M_T(t) = \mathbb{E}\, e^{tT} = \mathbb{E}\, e^{t(X-\lambda)} = e^{-t\,\lambda}\, \mathbb{E}\, e^{tX} = e^{-t\,\lambda} e^{-\lambda(e^t-1)} = e^{\lambda(e^t-t-1)}$, přičemž jsme využili tvar vytvořující funkce Poissonova rozdělení (2.1). Podle následující věty 6 snadno spočteme momenty náhodné veličiny T.

Věta 6 (Výpočet momentů vyššího řádu)**.** *Buď $M_X(t)$ momentová vytvořující funkce náhodné veličiny X konečná na nějakém okolí nuly, $s \in \mathbb{N}$. Potom s-tý moment X je roven $\frac{d^s M_X(0)}{dt^s}$ a tento výraz je dobře definován.*

(Zvára a Štěpán, 2006, věta 7.7)

Spočteme tedy potřebných 12 momentů T:

$$\begin{aligned}
\mathbb{E}T = 0,\ \mathbb{E}T^2 = \lambda,\ \mathbb{E}T^3 = \lambda,\ \mathbb{E}T^4 = 3\lambda^3 + \lambda,\\
\mathbb{E}T^5 = 10\lambda^2 + \lambda,\ \mathbb{E}T^6 = 15\lambda^3 + 25\lambda^2 + \lambda,\\
\mathbb{E}T^7 = 105\lambda^3 + 56\lambda^2 + \lambda,\ \mathbb{E}T^8 = 105\lambda^4 + 490\lambda^3 + 119\lambda^2 + \lambda\\
\mathbb{E}T^9 = 1260\,\lambda^4 + 1918\,\lambda^3 + 246\,\lambda^2 + \lambda\,,\\
\mathbb{E}T^{10} = 945\,\lambda^5 + 9450\,\lambda^4 + 6825\,\lambda^3 + 501\,\lambda^2 + \lambda\,,\\
\mathbb{E}T^{11} = 17325\,\lambda^5 + 56980\,\lambda^4 + 22935\,\lambda^3 + 1012\,\lambda^2 + \lambda\,,\\
\mathbb{E}T^{12} = 10395\,\lambda^6 + 190575\,\lambda^5 + 302995\,\lambda^4 + 74316\,\lambda^3 + 2035\,\lambda^2 + \lambda\,.
\end{aligned}$$

Pomocí vzorců $\operatorname{var}(Z) = \mathbb{E}\,Z^2 - (\mathbb{E}\,Z)^2$, $\operatorname{cov}(Z_1,Z_2) = \mathbb{E}[Z_1 Z_2] - \mathbb{E}\,Z_1\,\mathbb{E}\,Z_2$ vypočteme potřebné rozptyly a kovariance.

$$\begin{aligned}
\operatorname{var} T &= \lambda\\
\operatorname{var} T^2 &= 2\lambda^2 + \lambda\\
\operatorname{var} T^3 &= 15\lambda^3 + 24\lambda^2 + \lambda\\
\operatorname{var} T^4 &= 96\,\lambda^4\,484\lambda^3 + 118\lambda^2 + \lambda\\
\operatorname{var} T^5 &= 945\,\lambda^5 + 9350\,\lambda^4\,6805\lambda^3 + 500\lambda^2 + \lambda\\
\operatorname{var} T^6 &= 10170\,\lambda^6 + 189825\,\lambda^5 + 302340\,\lambda^4 + 74266\,\lambda^3 + 2034\,\lambda^2 + \lambda
\end{aligned}$$

$$\begin{aligned}
\operatorname{cov}(T,T^2) &= \lambda\\
\operatorname{cov}(T,T^3) &= 3\,\lambda^2 + \lambda\\
\operatorname{cov}(T,T^4) &= 10\,\lambda^2 + \lambda\\
\operatorname{cov}(T,T^5) &= 15\,\lambda^3 + 25\,\lambda^2 + \lambda\\
\operatorname{cov}(T^2,T^3) &= 9\,\lambda^2 + \lambda\\
\operatorname{cov}(T^2,T^4) &= 12\,\lambda^3 + 24\,\lambda^2 + \lambda\\
\operatorname{cov}(T^2,T^5) &= 95\,\lambda^3 + 55\,\lambda^2 + \lambda\\
\operatorname{cov}(T^3,T^4) &= 102\,\lambda^3 + 55\,\lambda^2 + \lambda\\
\operatorname{cov}(T^3,T^5) &= 105\,\lambda^4 + 480\,\lambda^3 + 118\,\lambda^2 + \lambda\\
\operatorname{cov}(T^4,T^5) &= 1230\,\lambda^4 + 1905\,\lambda^3 + 245\,\lambda^2 + \lambda
\end{aligned}$$

Takže pro λ tak velké, aby $|\frac{c}{\lambda}| < 1$, pak využitím vzorců pro součet geometrické řady a jejích derivací můžeme přistoupit k výpočtu rozptylu $g(X)$. Pro přehlednost budeme počítat po členech.

$$\begin{aligned}
\frac{\operatorname{var} T}{4m} = \frac{\lambda}{4\,\lambda(1+\frac{c}{\lambda})} &= \frac{1}{4}\sum_{j=0}^{\infty}(-1)^j\left(\frac{c}{\lambda}\right)^j\\
&= \frac{1}{4}\left(1 - \frac{c}{\lambda} + \frac{c^2}{\lambda^2}\right) + O(\lambda^{-3})\\
&= \frac{1}{4} - \frac{c}{4\lambda} + \frac{c^2}{4\lambda^2} + O(\lambda^{-3})
\end{aligned}$$

Podobným postupem spočteme i další potřebné členy.

$$\frac{\operatorname{var} T^2}{64m^3} = \frac{2\lambda^2+\lambda}{64\lambda^3(1+\frac{c}{\lambda})^3} = \frac{1}{32\lambda} - \frac{3c}{32\lambda^2} + \frac{1}{64\lambda^2} + O(\lambda^{-3})$$
$$-\frac{\operatorname{cov}(T,T^2)}{8m^2} = \frac{-\lambda}{8\,\lambda^2(1+\frac{c}{\lambda})^2} = \frac{-1}{8\,\lambda} + \frac{c}{4\,\lambda^2} + O(\lambda^{-3})$$

$$\frac{\operatorname{cov}(T,T^3)}{16m^3} = \frac{3\,\lambda^2+\lambda}{16\,\lambda^3(1+\frac{c}{\lambda})^3} = \frac{3}{16\,\lambda} + \frac{1}{16\,\lambda^2} - \frac{9c}{16\,\lambda^2} + O(\lambda^{-3})$$

Tím jsme vypočítali všechny členy řádu λ^{-1}. Další figurují v koeficientu u λ^{-2}.

$$\frac{\operatorname{var} T^3}{256m^5} = \frac{15\,\lambda^3+24\,\lambda^2+\lambda}{256\,\lambda^5(1+\frac{c}{\lambda})^5} = \frac{15}{256\,\lambda^2} + O(\lambda^{-3})$$
$$-\frac{5\operatorname{cov}(T,T^4)}{128m^4} = \frac{-5(10\,\lambda^2+\lambda)}{128\,\lambda^4(1+\frac{c}{\lambda})^4} = \frac{-25}{64\,\lambda^2} + O(\lambda^{-3})$$
$$\frac{7\operatorname{cov}(T,T^5)}{256m^5} = \frac{7(15\,\lambda^3+25\,\lambda^2+\lambda)}{256\,\lambda^5(1+\frac{c}{\lambda})^5} = \frac{105}{256\,\lambda^2} + O(\lambda^{-3})$$
$$-\frac{\operatorname{cov}(T^2,T^3)}{64m^4} = \frac{-9\,\lambda^2-\lambda)}{64\,\lambda^4(1+\frac{c}{\lambda})^4} = \frac{-9}{64\,\lambda^2} + O(\lambda^{-3})$$
$$\frac{5\operatorname{cov}(T^2,T^4)}{4\cdot 128m^5} = \frac{15}{128\,\lambda^2} + O(\lambda^{-3})$$

Zbylé členy už jsou jen $O(\lambda^{-3})$.

$$\frac{25\operatorname{var} T^4}{128^2m^7} = \frac{25(96\,\lambda^4\,484\lambda^3+118\lambda^2+\lambda)}{128^2(\lambda+c)^7} = O(\lambda^{-3})$$
$$\frac{49\operatorname{var} T^5}{256^2m^9} = \frac{49945\,\lambda^5+9350\,\lambda^4\,6805\lambda^3+500\lambda^2+\lambda}{256^2(\lambda+c)^9} = O(\lambda^{-3})$$
$$-\frac{7\operatorname{cov}(T^2,T^5)}{4\cdot 256m^6} = -\frac{795\,\lambda^3+55\,\lambda^2+\lambda\,,T^5)}{4\cdot 256(\lambda+c)^6} = O(\lambda^{-3})$$
$$-\frac{5\operatorname{cov}(T^3,T^4)}{8\cdot 128m^6} = \frac{-5(102\,\lambda^3+55\,\lambda^2+\lambda)}{8\cdot 128(\lambda+c)} = O(\lambda^{-3})$$
$$\frac{7\operatorname{cov}(T^3,T^5)}{8\cdot 256m^7} = \frac{7(105\,\lambda^4+480\,\lambda^3+118\,\lambda^2+\lambda)}{8\cdot 256(\lambda+c)^7} = O(\lambda^{-3})$$
$$-\frac{35\operatorname{cov}(T^4,T^5)}{64\cdot 256m^8} = -\frac{35(1230\,\lambda^4+1905\,\lambda^3+245\,\lambda^2+\lambda)}{64\cdot 256m^8} = O(\lambda^{-3})$$

Nyní se podíváme na členy závislé na R_6. K tomu využijeme horní mez, kterou jsme vypočítali výše, viz (2.6).

$$\begin{aligned}|\operatorname{var} R_6| &\le \operatorname{var}\left(G(6)\frac{T^6}{m^{6-\frac{1}{2}}}\right) = \frac{G^2(6)\operatorname{var} T^6}{m^{11}}\\ &= \frac{G^2(6)(10170\,\lambda^6+189825\,\lambda^5+302340\,\lambda^4+74266\,\lambda^3+2034\,\lambda^2+\lambda)}{(\lambda+c)^{11}}\\ &= O(\lambda^{-5})\end{aligned}$$

Dále budeme počítat kovariance tohoto členu s ostatními. Využijeme Hölderovu nerovnost, z níž pro náhodné veličiny X_1 a X_2 s konečnými druhými absolutními momenty plyne $|\mathrm{cov}(X_1, X_2)| \leq \sqrt{\mathrm{var}\, X_1 \,\mathrm{var}\, X_2}$ (Novovičová, 2006, kap. 4.2.4).

$$\left|\frac{\mathrm{cov}(T^k,R_6)}{m^{k-\frac{1}{2}}}\right| \leq \frac{\sqrt{\mathrm{var}\, T^k}\sqrt{\mathrm{var}\, R_6}}{m^{k-\frac{1}{2}}} \leq \frac{\sqrt{\mathrm{var}\, T^k}\sqrt{O(\lambda^{-5})}}{m^{k-\frac{1}{2}}} = \frac{\sqrt{\mathrm{var}\, T^k}}{m^{k-\frac{1}{2}}}O(\lambda^{-2,5}).$$

Například pro $k = 1$ dostáváme

$$\left|\frac{\mathrm{cov}(T,R_6)}{m^{1-\frac{1}{2}}}\right| \leq \frac{\sqrt{\mathrm{var}\, T}}{m^{\frac{1}{2}}}O(\lambda^{-2,5}) = \frac{\sqrt{\lambda}}{(\lambda+c)^{\frac{1}{2}}}O(\lambda^{-2,5}) = O(\lambda^{-2,5}).$$

Pro ostatní $k \in \{2, \dots 5\}$ můžeme výpočet provést obdobně a vždy dostaneme $\left|\frac{\mathrm{cov}(T^k,R_6)}{m^{k-\frac{1}{2}}}\right| = O(\lambda^{-2,5})$. Celkem tedy odhadneme, že členy závisející v rozptylu $g(X)$ na R_6 jsou členy $O(\lambda^{-2,5})$:

$$\left|\mathrm{var}\, R_6 + \sum_{k=1}^{5} 2a_k \frac{\mathrm{cov}(T^k, R_6)}{m^{k-\frac{1}{2}}}\right| \leq |\mathrm{var}\, R_6| + \sum_{k=1}^{5} 2a_k \left|\frac{\mathrm{cov}(T^k, R_6)}{m^{k-\frac{1}{2}}}\right| = O(\lambda^{-2,5}).$$

Můžeme tedy dopočítat kýžený rozptyl.

$$\begin{aligned}
\mathrm{var}\, g(X) &= \frac{1}{4} - \frac{c}{4\lambda} + \frac{c^2}{4\lambda^2} + \frac{1}{32\lambda} - \frac{3c}{32\lambda^2} + \frac{1}{64\lambda^2} + \frac{15}{256\,\lambda^2} - \frac{1}{8\,\lambda} + \frac{c}{4\,\lambda^2} \\
&\quad + \frac{3}{16\,\lambda} + \frac{1}{16\,\lambda^2} - \frac{9c}{16\,\lambda^2} + \frac{-25}{64\,\lambda^2} + \frac{-9}{64\,\lambda^2} + \frac{15}{128\,\lambda^2} + \frac{105}{256\,\lambda^2} + O(\lambda^{-2,5}) \\
&= \frac{1}{4} + \frac{-8c+3}{32\lambda} + \frac{64c^2 - 104c + 34}{256\lambda^2} + O(\lambda^{-2,5}) \\
&= \frac{1}{4} + \frac{-8c+3}{32\lambda} + \frac{32c^2 - 52c + 17}{128\lambda^2} + O(\lambda^{-2,5})
\end{aligned} \tag{2.7}$$

3. Binomické rozdělení

Druhým rozdělením, na které se zaměříme, je rozdělení binomické.

3.1 Charakterizace binomického rozdělení

Definice 4. *Buď X náhodná veličina. Řekneme, že X má **binomické rozdělení** s parametry $n \in \mathbb{N}$ a $p \in (0,\, 1)$, píšeme $X \sim Bi(n,\, p)$, jestliže $\mathsf{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$ pro $k \in \{0, 1, \ldots, n\}$.*

Střední hodnota $X \sim Bi(n,\, p)$ je rovna np a rozptyl $np(1-p)$. Z definice 2 spočítáme vytvořující funkci X:

$$\begin{aligned} M_X(t) = \mathbb{E}\, e^{tX} &= \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k (1-p)^{n-k} \\ &= (pe^t + (1-p))^n = (p(e^t - 1) + 1)^n. \end{aligned} \tag{3.1}$$

3.2 Transformace dle delta metody

Nejprve podle klasického přístupu využívajíce delta metodu určíme, jakou funkci bychom měli zvolit, pokud máme náhodný výběr $Y_1, \ldots, Y_n$, $Y_1 \sim Alt(p)$, a chceme transformovat jeho součet $X^{(n)} = \sum_{j=1}^{n} Y_j$. Pak tedy $X^{(n)} \sim Bi(n,\, p)$. Ze vzorce (2) vypočteme vhodnou transformaci g_0, a to pro parametr p (n je známé). Na to lze nahlížet jako na hledání transformace pro veličinu $\frac{X^{(n)}}{n}$, jejíž rozptyl je roven $\frac{p(1-p)}{n}$. Potom

$$g_0(p) = \int \frac{1}{\sqrt{\operatorname{var} \frac{X^{(n)}}{n}}} dp = \sqrt{n} \int \frac{1}{\sqrt{p(1-p)}} dp = 2\sqrt{n} \sin^{-1} \sqrt{p} + k,\ k \in \mathbb{R}.$$

Tato funkce je spojitě diferencovatelná a dobře definovaná na intervalu $\left(0,\, \frac{\pi^2}{4}\right]$. Pro náhodnou veličinu $X^{(n)}$ pak tedy využijeme transformaci

$$g_0(t) = \sqrt{n} \sin^{-1} \sqrt{\frac{t}{n}}.$$

3.3 Anscombeova transformace

Uvažme náhodnou veličinu $X \sim Bi(n,\, p)$, kde je np velké a také $n(1-p)$ má velkou hodnotou, tedy pravděpodobnost úspěchu p není příliš extremální. Potom F. J. Anscombe navrhuje využití transformace dané předpisem

$$g(t) = \sqrt{n + d_2} \sin^{-1} \sqrt{\frac{t + c}{n + d_1}}$$

pro vhodné konstanty c, d_1, a d_2. Funkce g je spojitě diferencovatelná na intervalu $\left(-c,\, \frac{\pi^2 (n + d_1)}{4} - c\right]$. Pomocí Taylorova rozvoje ukážeme, že

$$\operatorname{var} g(X) = \frac{1}{4} + \frac{2d_2 - 1}{8} + \frac{3 - 8c}{32np} + \frac{3 + 8c - 8d_1}{4(n - np)} + O\left(n^{-2}\right).$$

Jelikož chceme tento rozptyl stabilizovat, je žádoucí zvolit $c = \frac{3}{8}$, $d_1 = \frac{3}{4}$ a $d_2 = \frac{1}{2}$. Potom

$$\operatorname{var} g(X) = \frac{1}{4} + O\left(n^{-2}\right)$$

a vhodná transformace bude tvaru

$$g(t) = \sqrt{n + \frac{1}{2}} \sin^{-1} \sqrt{\frac{t + \frac{3}{8}}{n + \frac{3}{4}}}.$$

3.4 Rozptyl Anscombeovy transformace

3.4.1 Taylorův polynom $g(X)$

Buď X náhodná veličina, $X \sim Bi(n,\, p)$ s neznámým parametrem $p \in (0,\, 1)$. Označme její střední hodnotu $np = \theta$. Zadefinujme funkci g předpisem $g(t) = \sqrt{n + d_2} \sin^{-1} \sqrt{\frac{t+c}{n+d_1}}$, kde $c \geq 0$ (neboť za t budeme dosazovat veličinu X, která je s kladnou pravděpodobností nulová), $d_1 > -n$ a $d_2 > -n$ jsou konstanty. Abychom nalezli vhodné hodnoty konstant, které budou stabilizovat rozptyl náhodné veličiny $g(X)$, opět využijeme Taylorův rozvoj funkce g v bodě střední hodnoty.

První tři derivace g vypadají následovně:

$$\begin{aligned}
g'(t) &= \frac{\sqrt{n + d_2}}{2} \frac{1}{\left[(n + d_1 - (t + c))\,(t + c)\right]^{\frac{1}{2}}}, \\
g''(t) &= -\frac{\sqrt{n + d_2}}{2^2} \frac{-2t - 2c + n + d_1}{\left[(n + d_1 - (t + c))\,(t + c)\right]^{\frac{3}{2}}}, \\
g^{(3)}(t) &= \frac{\sqrt{n + d_2}}{2^3} \frac{8c^2 + 3(d_1 + n)^2 - 8c(d_1 + n - 2t) - 8(d_1 + n)t + 8t^2}{\left[(n + d_1 - (t + c))\,(t + c)\right]^{\frac{5}{2}}}.
\end{aligned}$$

Taylorův polynom g řádu 3 v bodě θ je tudíž tvaru

$$\begin{aligned}
T_3^{g,\,\theta} &= \sqrt{n + d_2} \sin^{-1} \sqrt{\frac{\theta + c}{n + d_1}} + \frac{\sqrt{n + d_2}}{1! \cdot 2} \frac{X - \theta}{\left[n + d_1 - (\theta + c))(\theta + c)\right]^{\frac{1}{2}}} \\
&\quad - \frac{\sqrt{n + d_2}}{2! \cdot 2^2} \frac{(-2\theta - 2c + n + d_1)(X - \theta)^2}{\left[n + d_1 - (\theta + c))(\theta + c)\right]^{\frac{3}{2}}} \\
&\quad + \frac{\sqrt{n + d_2}}{3! \cdot 2^3} \frac{8c^2 + 3(d_1 + n)^2 - 8c(d_1 + n - 2\theta) - 8(d_1 + n)\theta + 8\theta^2}{\left[(n + d_1 - (\theta + c))\,(\theta + c)\right]^{\frac{5}{2}}} (X - \theta)^3.
\end{aligned}$$

Pokud navíc označíme T vycentrovanou náhodnou veličinu X, tedy $T = X - \theta$, můžeme psát:

$$\begin{aligned}
T_3^{g,\,\theta} &= \sqrt{n + d_2} \sin^{-1} \sqrt{\frac{\theta + c}{n + d_1}} + \frac{\sqrt{n + d_2}}{2} \frac{T}{\left[n + d_1 - (\theta + c))(\theta + c)\right]^{\frac{1}{2}}} \\
&\quad - \frac{\sqrt{n + d_2}}{8} \frac{(-2\theta - 2c + n + d_1)T^2}{\left[n + d_1 - (\theta + c))(\theta + c)\right]^{\frac{3}{2}}} \\
&\quad + \frac{\sqrt{n + d_2}}{48} \frac{(8c^2 + 3(d_1 + n)^2 - 8c(d_1 + n - 2\theta) - 8(d_1 + n)\theta + 8\theta^2)T^3}{\left[(n + d_1 - (\theta + c))\,(\theta + c)\right]^{\frac{5}{2}}}.
\end{aligned} \tag{3.2}$$

3.4.2 Zbytkový člen rozvoje

Stejně jako v článku Anscombe (1948) budeme i zde předpokládat, že zbytkový člen $R_s = g(X) - T_{s-1}^{g,\theta}$ je omezený. Konkrétně, pro každé $s \in \mathbb{N}$ existuje kladná konečná konstanta $K(s)$, že

$$|R_s| \leq K(s) \frac{\sqrt{n+d_2}\,|T|^s}{(n+d_1)^s}.$$

Budeme uvažovat $s = 4$. Potom pro rozptyl a kovarianci platí

$$\begin{aligned} |\operatorname{var} R_4| &\leq K^2(s) \frac{(n+d_2)\operatorname{var} T^4}{(n+d_1)^8}, \\ \left|\operatorname{cov}(a_k T^k, R_4)\right| &\leq K(s)\,|a_k|\,(n+d_2) \frac{\sqrt{\operatorname{var} T^k}\sqrt{\operatorname{var} T^4}}{(n+d_1)^4},\ k = 1,\ 2,\ 3. \end{aligned} \tag{3.3}$$

3.4.3 Výpočet rozptylu $g(X)$

Nyní se detailně zaměříme na výpočet rozptylu náhodné veličiny $g(X)$.

Momenty T

Pro výpočet rozptylu náhodné transformované veličiny $g(X)$ potřebujeme znát momenty $T = X - \theta$. Budeme je vyjadřovat pomocí vytvořující funkce. K jejímu výpočtu využijeme hodnotu vytvořující funkce binomického rozdělení, již jsme spočetli výše (3.1).

$$M_T(t) = \mathbb{E}\, e^{tT} = \mathbb{E}\, e^{t(X-\theta)} = e^{-t\theta}\,\mathbb{E}\, e^{tX} = e^{-t\theta} M_X(t) = e^{-t\theta}(p(e^t - 1) + 1)^n$$

Dle věty 6 o výpočtu momentů pomocí derivací vytvořující funkce, pak můžeme určit potřebné momenty náhodné veličiny T.

$$\begin{aligned} \mathbb{E}\, T &= 0 \\ \mathbb{E}\, T^2 &= np(1-p) \\ \mathbb{E}\, T^3 &= np(1-3p+2p^2) \\ \mathbb{E}\, T^4 &= np(1-p)(1-6p+3np+6p^2-3np^2) \\ \mathbb{E}\, T^5 &= np(1-p)(1-2p)(1-12p+10np+12p^2-10np^2) \\ \mathbb{E}\, T^6 &= np(1-p)(1-30p+25np+150p^2-155np^2+15n^2p^2-240p^3 \\ &\quad +260np^3-30n^2p^3+120p^4-130np^4+15n^2p^4) \end{aligned}$$

Rozptyly a kovariance mocnin T

Nyní můžeme vyjádřit i potřebné rozptyly a kovariance, opět využívajíce rovností $\operatorname{var}(Z) = \mathbb{E}\, Z^2 - (\mathbb{E}\, Z)^2$ a $\operatorname{cov}(Z_1,Z_2) = \mathbb{E}[Z_1 Z_2] - \mathbb{E}\, Z_1\, \mathbb{E}\, Z_2$.

$$\begin{aligned} \operatorname{var} T &= np(1-p) \\ \operatorname{var} T^2 &= np(1-p)(1-6p+2np+6p^2-2np^2) \\ \operatorname{var} T^3 &= np(1-p)(1-30p+24np+150p^2-150np^2+15n^2p^2-240p^3 \\ &\quad +252np^3-30n^2p^3+120p^4-126np^4+15n^2p^4) \\ &= n^3p(1-p)(15p^2-30p^3+15p^4)+O(n^2) \end{aligned}$$

$$\begin{aligned}
\operatorname{cov}(T, T^2) &= np(1-p)(1-2p)\\
\operatorname{cov}(T, T^3) &= np(1-p)(1-6p+3np+6p^2-3np^2)\\
\operatorname{cov}(T^2, T^3) &= np(1-p)(1-2p)(1-12p+9np+12p^2-9np^2)
\end{aligned}$$

Pro zbytkový člen ještě využijeme, že

$$\begin{aligned}
\mathbb{E}\,T^8 = np(1-p)(&1-126p+119np+1806p^2-2275np^2+490n^2p^2-8400p^3\\
&+11620np^3-3360n^2p^3+105n^3p^3+16800p^4-24080np^4+7630n^2p^4\\
&-315n^3p^4-15120p^5+21924np^5-7140n^2p^5+315n^3p^5+5040p^6\\
&-7308np^6+2380n^2p^6-105n^3p^6),
\end{aligned}$$

a tudíž

$$\begin{aligned}
\operatorname{var} T^4 &= np(1-p)(1-126p+118np+1806p^2-2262np^2+484n^2p^2-8400p^3\\
&\quad+11560np^3-3312n^2p^3+96n^3p^3+16800p^4-23960np^4+7516n^2p^4\\
&\quad-288n^3p^4-15120p^5+21816np^5-7032n^2p^5+288n^3p^5+5040p^6\\
&\quad-7272np^6+2344n^2p^6-96n^3p^6)\\
&= n^4p(1-p)(96p^3-288p^4+288p^5-96p^6)+O(n^3).
\end{aligned}$$

Jednotlivé členy Taylorova rozvoje

Označme koeficienty u náhodných veličin T, T^2 a T^3 v Taylorově rozvoji podle výše spočteného (3.2): $a_0 = \sqrt{n+d_2}\sin^{-1}\sqrt{\frac{\theta+c}{n+d_1}}$, $a_1 = \frac{\sqrt{n+d_2}}{2[n+d_1-(\theta+c))(\theta+c)]^{\frac{1}{2}}}$, $a_2 = -\frac{\sqrt{n+d_2}}{2!\cdot 2^2}\frac{(-2\theta-2c+n+d_1)}{[n+d_1-(\theta+c))(\theta+c)]^{\frac{3}{2}}}$ a $a_3 = \frac{\sqrt{n+d_2}}{48}\frac{(8c^2+3(d_1+n)^2-8c(d_1+n-2\theta)-8(d_1+n)\theta+8\theta^2)}{[(n+d_1-(t+c))(t+c)]^{\frac{5}{2}}}$. Potom můžeme psát

$$\begin{aligned}
\operatorname{var} g(X) = {}& \operatorname{var}\left(a_0+a_1T+a_2T^2+a_3T^3+R_4\right)\\
&= a_1^2\operatorname{var}T + a_2^2\operatorname{var}T^2 + a_3^2\operatorname{var}T^3 + 2a_1a_2\operatorname{cov}(T, T^2)\\
&\quad + 2a_1a_3\operatorname{cov}(T, T^3) + 2a_2a_3\operatorname{cov}(T^2, T^3) + \widetilde{R_4},
\end{aligned} \tag{3.4}$$

kde $\widetilde{R_4} = \operatorname{var}R_4 + 2\sum_{k=1}^{3}\operatorname{cov}(a_kT^k, R_4)$, tedy $\widetilde{R_4}$ značí součet členů závislých na zbytku R_4.

Výpočet rozptylu $g(X)$ provedeme zvlášť pro jednotlivé členy. Jak jsme předeslali výše, čísla $n(1-p)$ a np uvažujeme velké, a tudíž můžeme předpokládat, že $\frac{d_1-c}{n(1-p)} < 1$ a také $\frac{c}{np} < 1$. Ještě připomeňme, že $\theta = np$.

$$\begin{aligned}
a_1^2\operatorname{var}T &= \frac{n+d_2}{4(n+d_1-np-c)(np+c)}np(1-p)\\
&= \frac{n^2(1+\frac{d_2}{n})p(1-p)}{4n^2p(1-p)(1+\frac{d_1-c}{n(1-p)})(1+\frac{c}{np})}\\
&= \frac{1}{4}\left(1+\frac{d_2}{n}\right)\left[\sum_{j=0}^{\infty}(-1)^j\left(\frac{d_1-c}{n(1-p)}\right)^j\right]\left[\sum_{j=0}^{\infty}(-1)^j\left(\frac{c}{np}\right)^j\right]\\
&= \frac{1}{4}\left(1+\frac{d_2}{n}\right)\left(1-\frac{d_1-c}{n(1-p)}+O\left(n^{-2}\right)\right)\left(1-\frac{c}{np}+O\left(n^{-2}\right)\right)\\
&= \frac{1}{4}\left(1+\frac{d_2}{n}-\frac{c}{np}-\frac{d_1-c}{n(1-p)}\right)+O\left(n^{-2}\right)
\end{aligned}$$

Obdobně počítáme i další členy rozvoje.

$$\begin{aligned}
a_2^2 \operatorname{var} T^2 &= \frac{(n+d_2)(-2np-2c+n+d_1)^2}{8^2(n+d_1-np-c)^3(np+c)^3} np(1-p)(1-6p+2np+6p^2-2np^2) \\
&= \frac{(1-2p)^2\left(1+\frac{d_1-2c}{n(1-2p)}\right)^2\left(1+\frac{d_2}{n}\right)\left(\frac{1-6p+6p^2}{np(1-p)}+2\right)}{64np(1-p)}\left(1+O(n^{-1})\right)^2 \\
&= \frac{(1-2p)^2\left(1+\frac{d_1-2c}{n(1-2p)}\right)^2\left(1+\frac{d_2}{n}\right)\left(\frac{1-6p+6p^2}{np(1-p)}+2\right)}{64np(1-p)}\left(1+O(n^{-1})\right)^2 \\
&= \frac{2(1-2p)^2}{64np(1-p)} + O\left(n^{-2}\right) = \frac{(1-2p)^2}{32np(1-p)} + O\left(n^{-2}\right)
\end{aligned}$$

Dále bude ve výpočtu konstant hrát roli kovariance T, T^2 a T, T^3.

$$\begin{aligned}
a_1a_2 \operatorname{cov}(T,\, T^2) &= -\frac{(n+d_2)(-2np-2c+n+d_1)}{16(n+d_1-np-c)^2(np+c)^2} np(1-p)(1-2p) \\
&= -\frac{(1-2p)(1+\frac{d_2}{n})(1-2p+\frac{d_1-2c}{n})}{16np(1-p)}\left(1+O(n^{-1})\right)^2 \\
&= -\frac{(1-2p)^2}{16np(1-p)} + O\left(n^{-2}\right) \\
a_1a_3 \operatorname{cov}(T,\, T^3) &= \frac{8c^2+3(d_1+n)^2-8n(d_1+n)p+8n^2p^2-8c(d_1+n-2np)}{96(n+d_1-np-c)^3(np+c)^3} \\
&\quad \cdot (n+d_2)np(1-p)(1-12p+9np+12p^2-9np^2) \\
&= \frac{(3-8p+8p^2)}{96np^3(1-p)^3}\left(1+O(n^{-1})\right)^3 p(1-p)\left(3p-3p^2+O(n^{-1})\right) \\
&= \frac{3-8p+8p^2}{32np(1-p)} + O\left(n^{-2}\right)
\end{aligned}$$

Nahlédneme, že zbylé dva členy už jsou řádu menšího než n^{-1}, neboli

$$\begin{aligned}
a_3^2 \operatorname{var} T^3 &= O\left(n^{-2}\right) \\
a_2a_3 \operatorname{cov}(T^2,\, T^3) &= O\left(n^{-2}\right).
\end{aligned}$$

Podobně pro zbytkové členy využitím výpočtu (3.3) dostáváme

$$|\operatorname{var} R_4| \le O(n^{-4})$$

a dále

$$\begin{aligned}
|\operatorname{cov}(a_1T,\, R_4)| &\le O(n^{-2}) \\
\left|\operatorname{cov}(a_2T^2,\, R_4)\right| &\le O(n^{-2,5}) \\
\left|\operatorname{cov}(a_3T^3,\, R_4)\right| &\le O(n^{-3}).
\end{aligned}$$

Tudíž pro celkový zbytkový člen platí $\widetilde{R_4} = O(n^{-2})$.

Konečný výpočet rozptylu

Můžeme tedy dosadit spočtené členy do (3.4) a vyčíslit rozptyl náhodné veličiny $g(X)$:

$$\begin{aligned}\operatorname{var} g(X) = \frac{1}{4} + \frac{d_2}{4n} - \frac{c}{4np} + \frac{c-d_1}{4n(1-p)} + \frac{(1-2p)^2}{32np(1-p)} - 2\frac{(1-2p)^2}{16np(1-p)} \\ + 2\frac{3-8p+8p^2}{32np(1-p)} + O\left(n^{-2}\right).\end{aligned} \tag{3.5}$$

Tento tvar můžeme ještě upravit, neboť platí

$$\begin{aligned}\frac{(1-2p)^2}{32np(1-p)} - 2\frac{(1-2p)^2}{16np(1-p)} &= -\frac{3(1-p-p)^2}{32np(1-p)} \\ &= -\frac{3(1-p)^2}{32np(1-p)} + \frac{6(1-p)p}{32np(1-p)} - \frac{3p^2}{32np(1-p)} \\ &= -\frac{3(1-p)}{32np} + \frac{6}{32n} - \frac{3(p-1+1)}{32n(1-p)} \\ &= -\frac{3}{32n(1-p)} + \frac{3}{8n} - \frac{3}{32np}\end{aligned}$$

a dále

$$\begin{aligned}2\frac{3-8p+8p^2}{32np(1-p)} &= \frac{6(1-p)+6p}{32np(1-p)} - \frac{16}{32n(1-p)} + \frac{16(p-1+1)}{32n(1-p)} \\ &= \frac{6}{32np} + \frac{6}{32n(1-p)} - \frac{16}{32n(1-p)} - \frac{16}{32n} + \frac{16}{32n(1-p)} \\ &= \frac{3}{16np} + \frac{3}{16n(1-p)} - \frac{1}{2n}.\end{aligned}$$

Dosazením dílčích výpočtů do (3.5) konečně dopočítáme rozptyl transformované veličiny.

$$\begin{aligned}\operatorname{var} g(X) &= \frac{1}{4} + \frac{d_2}{4n} - \frac{c}{4np} + \frac{c-d_1}{4n(1-p)} - \frac{3}{32n(1-p)} + \frac{3}{8n} - \frac{3}{32np} \\ &\quad + \frac{3}{16np} + \frac{3}{16n(1-p)} - \frac{1}{2n} + O\left(n^{-2}\right) \\ &= \frac{1}{4} + \frac{2d_2-1}{8n} + \frac{3-8c}{32np} + \frac{3+8c-8d_1}{32n(1-p)} + O\left(n^{-2}\right)\end{aligned}$$

4. Simulace vhodnosti transformace pro Poissonovo rozdělení

V poslední části této práce ilustrujeme vhodnost využití transformace stabilizující rozptyl při odhadování střední hodnoty pomocí intervalových odhadů náhodného výběru z Poissonova rozdělení. Navíc zde porovnáme, jaké přesnosti na náhodně generovaných datech z Poissonova rozdělení dosahuje klasická transformace a jaké ta Anscombeova.

Mějme $X_1, \ldots, X_n$ posloupnost náhodných stejně rozdělených veličin s konečným a nenulovým rozptylem. Označíme $\mu = \mathbb{E}\, X_1$ a $\sigma^2 = \operatorname{var} X_1$. Budeme zmíněnými způsoby počítat intervaly spolehlivosti pro střední hodnotu, tedy neznámý parametr, jak jej uvažujeme v celé této práci.

Pro $\alpha \in (0,\, 1)$ buď u_α α-kvantil normovaného normálního rozdělení.

K odvození příslušných intervalů spolehlivosti se nám bude hodit Cramér-Sluckého věta:

Věta 7 (Cramér-Slucký)**.** *Nechť $\{X_n\}$, $\{Y_n\}$ a $\{Z_n\}$ jsou posloupnosti náhodných veličin, X, Z náhodné veličiny, c konstanta a platí $X_n \xrightarrow[n\to\infty]{d} X$, $Z_n \xrightarrow[n\to\infty]{d} Z$ a $Y_n \xrightarrow[n\to\infty]{\mathsf{P}} c$. Potom platí $X_nY_n + Z_n \xrightarrow[n\to\infty]{d} cX + Z$.*

(Anděl, 2007, věta B. 10)

4.1 Interval spolehlivosti bez transformace

Využijeme centrální limitní věty (věta 1), podle které platí:

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow[n\to\infty]{d} N(0,\, 1),$$

navíc mějme $\widehat{\sigma}_n$ konzistentní odhad σ, tj. $\frac{\sigma}{\widehat{\sigma}_n} \xrightarrow[n\to\infty]{\mathsf{P}} 1$, pak podle věty 7 a centrální limitní věty platí

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\widehat{\sigma}_n} = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}\frac{\sigma}{\widehat{\sigma}_n} \xrightarrow[n\to\infty]{d} N(0,\, 1). \tag{4.1}$$

Potom z definice konvergence v distribuci dostáváme

$$\mathsf{P}\left[u_{\alpha/2} < \sqrt{n}\frac{\bar{X}_n - \mu}{\widehat{\sigma}_n} < u_{1-\alpha/2}\right] \xrightarrow[n\to\infty]{} 1 - \alpha,$$

tedy

$$\mathsf{P}\left[\bar{X}_n - u_{1-\alpha/2}\frac{\widehat{\sigma}_n}{\sqrt{n}} < \mu < \bar{X}_n - u_{\alpha/2}\frac{\widehat{\sigma}_n}{\sqrt{n}}\right] \xrightarrow[n\to\infty]{} 1 - \alpha.$$

Abychom zachovali obvyklý zápis, můžeme ještě využít symetrie normálního normovaného rozdělení, z níž plyne, že $u_\alpha = -u_{1-\alpha}$, $\alpha \in (0,\, 1)$. Tudíž interval $\left(\bar{X}_n - u_{1-\alpha/2}\frac{\widehat{\sigma}_n}{\sqrt{n}}, \bar{X}_n + u_{1-\alpha/2}\frac{\widehat{\sigma}_n}{\sqrt{n}}\right)$ je asymptotický interval spolehlivosti pro střední hodnotu o pravděpodobnosti pokrytí $1-\alpha$, kde $\alpha \in (0,\, 1)$.

Speciálně pro náhodný výběr z Poissonova rozdělení položíme $\widehat{\sigma}_n^2 = \bar{X}_n$, což je konzistentní odhad σ^2, neboť pro toto rozdělení jest $\mu = \sigma^2$. A zároveň platí, že výběrový průměr je konzistentním odhadem střední hodnoty, jak plyne z Čebyševova zákonu velkých čísel (věta 3).

Budeme tudíž využívat interval spolehlivosti

$$\left(\bar{X}_n - u_{1-\alpha/2}\frac{\sqrt{\bar{X}_n}}{\sqrt{n}}, \bar{X}_n + u_{1-\alpha/2}\frac{\sqrt{\bar{X}_n}}{\sqrt{n}}\right). \tag{4.2}$$

Jelikož odhadujeme kladný parametr λ, tak pokud by výraz $\bar{X}_n - u_{1-\alpha/2}\frac{\sqrt{\bar{X}_n}}{\sqrt{n}}$ vyšel záporný, uvážíme místo něj jako dolní mez intervalu nulu. Délka tohoto intervalu je

$$\bar{X}_n - u_{1-\alpha/2}\frac{\sqrt{\bar{X}_n}}{\sqrt{n}} - \left(\bar{X}_n + u_{1-\alpha/2}\frac{\sqrt{\bar{X}_n}}{\sqrt{n}}\right) = 2u_{1-\alpha/2}\frac{\sqrt{\bar{X}_n}}{\sqrt{n}}. \tag{4.3}$$

4.2 Interval spolehlivosti transformací

Mějme transformaci stabilizující rozptyl g, pro niž platí vztah

$$\sqrt{n}\left(g(\bar{X}_n) - g(\mu)\right) \xrightarrow[n\to\infty]{d} N(0,\, 1),$$

z toho pro $\alpha \in (0,\, 1)$ plyne následující:

$$\mathsf{P}\left[u_{\alpha/2} < \sqrt{n}\left(g(\bar{X}_n) - g(\mu)\right) < u_{1-\alpha/2}\right] \xrightarrow[n\to\infty]{} 1-\alpha, \text{ a proto}$$
$$\mathsf{P}\left[g(\bar{X}_n) - \frac{u_{1-\alpha/2}}{\sqrt{n}} < g(\mu) < g(\bar{X}_n) + \frac{u_{1-\alpha/2}}{\sqrt{n}}\right] \xrightarrow[n\to\infty]{} 1-\alpha,$$

což pro g prosté a rostoucí můžeme dále upravit jako

$$\mathsf{P}\left[g^{-1}\left(g(\bar{X}_n) - \frac{u_{1-\alpha/2}}{\sqrt{n}}\right) < \mu < g^{-1}\left(g(\bar{X}_n) + \frac{u_{1-\alpha/2}}{\sqrt{n}}\right)\right] \xrightarrow[n\to\infty]{} 1-\alpha.$$

Získáváme tedy interval spolehlivosti pro střední hodnotu μ o asymptotické pravděpodobnosti pokrytí $1-\alpha$, a to ve tvaru

$$\left(g^{-1}\left(g(\bar{X}_n) - \frac{u_{1-\alpha/2}}{\sqrt{n}}\right), g^{-1}\left(g(\bar{X}_n) + \frac{u_{1-\alpha/2}}{\sqrt{n}}\right)\right). \tag{4.4}$$

V našem konkrétním případě, kdy náhodný výběr pochází z Poissonova rozdělení, budeme uvažovat dvě transformující funkce, a sice funkci, kterou jsme odvodili pomocí delta metody a (v lehce modifikované podobě) tu, již navrhl J. F. Anscombe.

4.2.1 Delta metoda

Nejprve uvážíme transformaci odvozenou pomocí delta metody, tj. $g(t) = 2\sqrt{(t)}$. Funkce g je zřejmě prostá a rostoucí na celém svém definičním oboru. Jest $g^{-1}(t) = \frac{t^2}{4}$, $t \geq 0$. Dosazením do (4.2) tedy dostáváme interval spolehlivosti

$$\left(\frac{1}{4} \left(2\sqrt{\bar{X}_n} - \frac{u_{1-\alpha/2}}{\sqrt{n}} \right)^2, \frac{1}{4} \left(2\sqrt{\bar{X}_n} + \frac{u_{1-\alpha/2}}{\sqrt{n}} \right)^2 \right). \tag{4.5}$$

Mohlo by se stát, že výraz $2\sqrt{\bar{X}_n} - \frac{u_{1-\alpha/2}}{\sqrt{n}}$ v intervalu spolehlivosti (4.5) vyjde záporný. V tom případě by nedávalo dobrý smysl uvažovat jeho druhou mocninu (respektive její násobek) a místo toho položíme dolní mez tohoto intervalu spolehlivosti rovnu nule.

Pokud ještě umocníme krajní meze intervalu (4.5), dostaneme jej ve tvaru

$$\left(\bar{X}_n + \frac{(u_{1-\alpha/2})^2}{4n} - \sqrt{\frac{\bar{X}_n}{n}} u_{1-\alpha/2},\ \bar{X}_n + \frac{(u_{1-\alpha/2})^2}{4n} + \sqrt{\frac{\bar{X}_n}{n}} u_{1-\alpha/2} \right).$$

Délka tohoto intervalu je $2\sqrt{\bar{X}_n} \frac{u_{1-\alpha/2}}{\sqrt{n}}$, tedy je rovna délce intervalu spolehlivosti bez využití transformace stabilizující rozptyl, jak jsme spočítali výše v (4.3). Tudíž se tento interval liší tím, že má střed namísto v bodě $\bar{X}_n$ v bodě $\bar{X}_n + \frac{(u_{1-\alpha/2})^2}{4n}$.

4.2.2 Anscombe

Nyní budeme zkoumat, jak vypadají intervaly spolehlivosti, za využití transformaci podle Anscombea. V kapitole 2 jsme odvodili hodnotu rozptylu pro transformovanou náhodnou veličinu z Poissonova rozdělení. Nyní se však takovou veličinou nezabýváme přímo, ale chystáme se transformovat výběrový průměr náhodného výběru z tohoto rozdělení. Proto budeme muset transformaci $g(t) = 2\sqrt{t + \frac{3}{8}}$ nejprve drobně modifikovat.

Tvar transformace

Mějme náhodný výběr $X_1, \ldots, X_n$ takový, že $X_1 \sim Po(\lambda)$. Chceme zjistit, jakou transformaci využít při vyšetřování asymptotického výběrového průměru $\bar{X}_n$. Označme součet nezávislých náhodných veličin z tohoto výběru jako $X^{(n)} = \sum_{i=1}^{n} X_i$. Potom $X^{(n)} \sim Po(n\,\lambda)$.

V kapitole 2, konkrétně ve výpočtu (2.7), jsme dokázali, že pro funkci $g(t) = 2\sqrt{t + c}$ platí $\operatorname{var} g\left(X^{(n)}\right) = 1 + O\left((n\,\lambda)^{-1}\right)$. Můžeme ovšem psát

$$\begin{aligned} \operatorname{var} g\left(X^{(n)}\right) &= \operatorname{var} 2\sqrt{X^{(n)} + c} = \operatorname{var} 2\sqrt{\sum_{i=1}^{n} X_i + c} = n \operatorname{var} 2\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i + \frac{c}{n}} \\ &= n \operatorname{var} 2\sqrt{\bar{X}_n + \frac{c}{n}}. \end{aligned}$$

Tudíž jest

$$n \operatorname{var} 2\sqrt{\bar{X}_n + \frac{c}{n}} = 1 + O\left((n\,\lambda)^{-1}\right).$$

Tedy pro zkoumání asymptotického rozdělení výběrového průměru využijeme transformaci $\bar{g}_n(t) = 2\sqrt{t + \frac{c}{n}}$. Dokážeme, že pro ni platí

$$\sqrt{n}\left(\bar{g}_n(\bar{X}_n) - \bar{g}_n(\lambda)\right) \xrightarrow[n\to\infty]{d} N(0,\,1). \tag{4.6}$$

Nejprve upravíme výraz nalevo.

$$\begin{aligned}
\sqrt{n}\left(\bar{g}_n(\bar{X}_n) - \bar{g}_n(\lambda)\right) &= \sqrt{n}\left(2\sqrt{\bar{X}_n + \frac{3}{8n}} - 2\sqrt{\lambda + \frac{3}{8n}}\right) \\
&= 2\sqrt{n}\left(\sqrt{\bar{X}_n}\sqrt{1 + \frac{3}{8n\bar{X}_n}} - \sqrt{\lambda}\sqrt{1 + \frac{3}{8n\,\lambda}}\right) \\
&= 2\sqrt{n}\left(\sqrt{\bar{X}_n} - \sqrt{\lambda}\right)\sqrt{1 + \frac{3}{8n\bar{X}_n}} \\
&\quad + 2\sqrt{n\,\lambda}\left(\sqrt{1 + \frac{3}{8n\bar{X}_n}} - \sqrt{1 + \frac{3}{8n\,\lambda}}\right)
\end{aligned}$$

Z delta metody (věta 2) víme, že $2\sqrt{n}\left(\sqrt{\bar{X}_n} - \sqrt{\lambda}\right) \xrightarrow[n\to\infty]{d} N(0,\,1)$. A zároveň $\sqrt{1 + \frac{3}{8n\bar{X}_n}} = \sqrt{1 + \frac{3}{8X^{(n)}}} \xrightarrow[n\to\infty]{P} 1$ (neboť součet n nezáporných náhodných veličin roste s pravděpodobností jedna pro n jdoucí do nekonečna nade všechny meze), tudíž podle věty 7

$$2\sqrt{n}\left(\sqrt{\bar{X}_n} - \sqrt{\lambda}\right)\sqrt{1 + \frac{3}{8n\bar{X}_n}} \xrightarrow[n\to\infty]{d} N(0,\,1).$$

Dále

$$2\sqrt{n\,\lambda}\left(\sqrt{1 + \frac{3}{8n\bar{X}_n}} - \sqrt{1 + \frac{3}{8n\,\lambda}}\right) = 2\sqrt{n\,\lambda}\frac{1 + \frac{3}{8n\bar{X}_n} - 1 - \frac{3}{8n\,\lambda}}{\sqrt{1 + \frac{3}{8n\bar{X}_n}} + \sqrt{1 + \frac{3}{8n\,\lambda}}} \xrightarrow[n\to\infty]{P} 0.$$

Ukázali jsme tedy, že platí $\sqrt{n}\left(\bar{g}_n(\bar{X}_n) - \bar{g}_n(\lambda)\right) \xrightarrow[n\to\infty]{d} N(0,\,1)$, opět díky větě 7.

Interval spolehlivosti

Transformaci $\bar{g}_n(t) = 2\sqrt{t + \frac{3}{8n}}$ tedy můžeme dosadit do obecného předpisu pro interval spolehlivosti (4.2). Opět se jedná o funkci na celém svém definičním oboru rostoucí a prostou. Vyjádříme inverzní zobrazení $g^{-1}(t) = \frac{t^2}{4} - \frac{3}{8n}, t \geq -\frac{3}{8n}$. Potom získáme interval spolehlivosti

$$\left(\frac{1}{4}\left(2\sqrt{\bar{X}_n + \frac{3}{8n}} - \frac{u_{1-\alpha/2}}{\sqrt{n}}\right)^2 - \frac{3}{8n},\ \frac{1}{4}\left(2\sqrt{\bar{X}_n + \frac{3}{8n}} + \frac{u_{1-\alpha/2}}{\sqrt{n}}\right)^2 - \frac{3}{8n}\right). \tag{4.7}$$

Ani v tomto případě není g^{-1} prostá funkce na celé reálné ose. Pokud by tedy výraz $\sqrt{\bar{X}_n + \frac{3}{8n}} - \frac{u_{1-\alpha/2}}{\sqrt{2}n}$ v dolní mezi (4.7) vyšel záporný nebo by jeho druhá mocnina byla menší než $\frac{3}{8n}$, budeme tuto mez uvažovat rovnou 0.

Umocněním ještě můžeme upravit tvar intervalu následovně:

$$\left(\bar{X}_n + \frac{(u_{1-\alpha/2})^2}{4n} - \frac{u_{1-\alpha/2}}{\sqrt{n}} \sqrt{\bar{X}_n + \frac{3}{8n}}, \bar{X}_n + \frac{(u_{1-\alpha/2})^2}{4n} + \sqrt{\frac{\bar{X}_n + \frac{3}{8n}}{n}} u_{1-\alpha/2} \right).$$

I v tomto případě budeme počítat délku intervalu spolehlivosti. Jeho délka je rovná $2\frac{u_{1-\alpha/2}}{\sqrt{n}}\sqrt{\bar{X}_n + \frac{3}{8n}}$. Všimneme si, že kdykoli, kromě případu, kdy jest $2\sqrt{\bar{X}_n + \frac{3}{8n}} - \frac{u_{1-\alpha/2}}{\sqrt{n}} \geq 0$ a zároveň $2\sqrt{\bar{X}_n} - \frac{u_{1-\alpha/2}}{\sqrt{n}} \leq 0$, vychází délka intervalu spolehlivosti s využitím Anscombeovy transformace větší, než když využijeme transformaci pomocí delta věty.

4.3 Algoritmus

Pro simulaci jsme využili výpočetní prostředí R Core Team (2018). Přesný použitý skript se nachází v elektronické příloze této práce. Zvolili jsme fixní pravděpodobnost pokrytí 0,95, tedy $\alpha = 0{,}05$. Vygenerovali jsme $B = 10^5$ náhodných výběrů z Poissonova rozdělení se střední hodnotou $\lambda > 0$ o rozsahu $n \in \mathbb{N}$. Pro každý z náhodných výběrů jsme spočítali intervaly spolehlivosti (4.1), (4.5) i (4.7). Načež jsme se pro každý z těchto intervalů podívali, zda překrývá zvolenou střední hodnotu λ. Našim cílem je v rámci těchto tří metod porovnat průměrný počet intervalů spolehlivosti, do nichž λ náleží. Druhou vlastností, jež nás bude zajímat, je průměrná délka nagenerovaných intervalů spolehlivosti.

4.4 Výsledky simulace

Následuje tabulka 4.1 ukazující výsledky simulací, jež byly provedeny podle výše zmíněného algoritmu.

Podle těchto dat využití transformace stabilizující rozptyl ve většině případů vede ke zvýšení pravděpodobnosti pokrytí. V našich simulacích je výjimkou pouze případ, kdy $n = 10$, $\lambda = 0{,}2$ a využíváme klasický přístup k transformaci pomocí delta metody. Pokud jsme využili transformaci stabilizující rozptyl podle Anscombea, pak při všech volbách rozsahu výběru i střední hodnoty nejen, že se zvýšilo průměrné pokrytí oproti případu, kdy jsme transformaci nevyužili vůbec, ale i ve zmíněném případě $n = 10, \lambda = 0{,}2$, kdy transformace podle delta metody selhala, tato Anscombeova vedla ke zvýšení pokrytí.

Na druhou stranu délky intervalů spolehlivosti o výhodnosti Anscombovy transformace nesvědčí. Jak jsme nahlédli již v teoretickém odvození vlastností intervalů spolehlivosti, bez využití jakékoli transformace i transformace dle delta metody jsou si všechny průměrné délky intervalů spolehlivosti rovny. Zatímco využitím Anscombovy transformace jejich hodnotu zvýšíme. A to zejména jsou-li rozsah výběru i střední hodnota malé.

$\alpha = 0{,}05$		bez trans.	delta trans.	Anscombova trans.
$n = 10$	pokrytí	86,34	84,77	98,33
$\lambda = 0{,}2$	délka IS	0,511	0,511	0,560
$n = 10$	pokrytí	86,96	94,62	94,62
$\lambda = 0{,}5$	délka IS	0,852	0,852	0,885
$n = 10$	pokrytí	92,59	95,68	95,68
$\lambda = 1$	délka IS	1,223	1,223	1,247
$n = 10$	pokrytí	94,34	94,67	94,67
$\lambda = 10$	délka IS	3,914	3,914	3,922
$n = 100$	pokrytí	94,79	94,84	94,84
$\lambda = 10$	délka IS	1,240	1,240	1,240
$n = 10$	pokrytí	94,91	94,97	95,97
$\lambda = 100$	délka IS	12,39	12,39	12,40
$n = 100$	pokrytí	95,05	95,06	95,06
$\lambda = 100$	délka IS	3,920	3,920	3,920

Tabulka 4.1: Průměrné procentuální pokrytí parametru λ Poissonova rozdělení a průměrná délka intervalu spolehlivosti pro interval spolehlivosti počítaný bez využití transformace, s využitím transformace získané z delta metody a Anscombeovy transformace stabilizující rozptyl při rozsahu výběru n. Vše pro předepsanou pravděpodobnost pokrytí 0,95 a 10^5 generovaných náhodných výběrů o rozsahu n.

Závěr

V práci jsme čtenáře seznámili se statistickou metodou transformace stabilizující rozptyl. Vysvětlili jsme, v čem spočívá a odvodili, jak takovou transformaci s použitím delta metody nalézt.

Poté jsme se zaměřili na transformace vhodné pro data, o nichž předpokládáme, že pochází z Poissonova rozdělení, avšak s neznámou střední hodnotu. Odvodili jsme pro ně možnou transformaci ve tvaru $g(t) = 2\sqrt{t}$ a podrobně dokázali, že při Anscombeově transformaci $g(t) = 2\sqrt{t+c}$, jež předpokládá velkou střední hodnotu, je ideální zvolit $c = \frac{3}{8}$. Podobně jsme pak zkoumali binomické rozdělení. Pomocí delta metody jsme pro ně nalezli transformaci $g(t) = 2\sqrt{n}\sin^{-1}\sqrt{\frac{t}{n}}$ a dokázali jsme, že pro Anscombeovu transformaci $g(t) = 2\sqrt{n+d_2}\sin^{-1}\sqrt{\frac{t+c}{n+d_1}}$, za předpokladu velkých hodnot np a $n(1-p)$, je záhodno zvolit konstanty tak, aby $g(t) = 2\sqrt{n+\frac{1}{2}}\sin^{-1}\sqrt{\frac{t+\frac{3}{8}}{n+\frac{3}{4}}}$.

Nakonec jsme předvedli využití transformace stabilizující rozptyl při odhadování střední hodnoty pomocí intervalů spolehlivosti. Konkrétně jsme ukázali, jak vypadají v případě Poissonova rozdělení a při využití obou transformací, kterými jsme se zabývali spočetli jejich délku (v závislosti na konkrétních datech). Celou práci jsme zakončili simulací toho, jak dobře intervaly spolehlivosti pokrývají zkoumaný parametr, a to ve všech třech případech, tedy když jsme transformaci stabilizující rozptyl vůbec nevyužili, když jsme použili transformaci dle delta metody a tu Anscombeovu. Tato simulace vede k závěru, že Anscomeova transformace zvyšuje pravděpodobnost pokrytí parametru, ovšem délka intervalu spolehlivosti se jejím použitím zvyšuje. Zatímco klasická transformace ne vždy pravděpodobnost pokrytí zvětší, ovšem délku jí příslušného intervalu spolehlivosti nezvýší.

Touto prací jistě není téma transformace stabilizující rozptyl vyčerpáno. Článek Anscombe (1948) hovoří ještě o upravené transformaci pro negativně binomické rozdělení, která by si jistě také zasloužila bližší prozkoumání a pečlivé dokázání vhodnosti jejího využití.

Dalo by se však pokračovat i v dalším zkoumání vhodných transformací pro Poissonovo a binomické rozdělení. Anscombeem navrhované transformace nejsou jediné, které slibují ještě lepší stabilizaci rozptylu. Například v článku (Mosteller a Youtz, 1961) je pro data z Poissonova rozdělení navrhnuta transformace $g(t) = \sqrt{t}+\sqrt{t+1}$ a pro binomické rozdělení $g(t) = \frac{1}{2}\sin^{-1}\sqrt{\frac{t}{n+1}}+\frac{1}{2}\sin^{-1}\sqrt{\frac{t+1}{n+1}}$.

Seznam použité literatury

Anděl, J. (1985). *Matematická statistika*. 2. vydání. SNTL/Alpha, Praha.

Anděl, J. (2003). *Statistické metody*. 2. vydání. Matfyzpress, Praha. ISBN 80-85863-27-8.

Anděl, J. (2007). *Základy matematické statistiky*. 1. vydání. Matfyzpress, Praha. ISBN 80-86732-40-1.

Anscombe, F. J. (1948). The Transformation of Poisson, Binomial and Negative-binomial Data. *Biometrika*, **35**, 246–254.

Dížková, M. (2012). Konvergence číselných řad – teorie a příklady. MUNI, Brno, bakalářská práce.

Kopáček, J. (2004). *Matematická analýza nejen pro fyziky (I)*. 4. vydání. Matfyzpress, Praha. ISBN 80-86732-25-8.

Mosteller, F. a Youtz, C. (1961). Tables of the Rreeman-Tukey Transformations for the Binomial and Poisson Distributions. *Biometrika*, **84**, 433–440.

Novovičová, J. (2006). *Pravděpodobnost a matematická statistika*. 12. verze. ČVUT, Olomouc. ISBN 80-01-01980-2.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Zvára, K. a Štěpán, J. (2006). *Pravděpodobnost a matematická statistika*. 4. vydání. Matfyzpress, Praha. ISBN 80-86732-71-1.