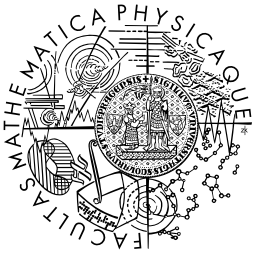


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Marek Štěpán

Oprava na spojitost

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

Na tomto místě bych rád poděkoval vedoucímu mé práce, doc. Ing. Marku Omelkovi, Ph.D. za čas, který mi věnoval, za ochotu, cenné rady a odborné připomínky, které mi pomohly při tvorbě této práce.

Název práce: Oprava na spojitost

Autor: Marek Štěpán

Katedra: Katedra pravděpodobnosti a matematické statistiky

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

Abstrakt: Pro aproximaci rozdělení náhodné veličiny, která je součtem n nezávislých, stejně rozdělených diskrétních náhodných veličin můžeme využít centrální limitní větu. Ukazuje se však, že pro konečná n umíme tuto aproximaci zpřesnit použitím opravy na spojitost. Tento pojem je v práci vysvětlen a také je v ní ilustrováno, jak může být oprava na spojitost odvozena. V práci je také numericky porovnána chyba aproximace binomického rozdělení rozdělením normálním s opravou na spojitost a aproximace bez opravy. Dále jsou zde popsány intervalové odhady a χ^2 test nezávislosti v kontingenčních tabulkách, ve kterých se používá oprava na spojitost. Na simulacích pro různé parametry vyzkoušíme vlastnosti těchto intervalů (skutečnou spolehlivost a délku) a testů (skutečnou hladinu a sílu).

Klíčová slova: oprava na spojitost, centrální limitní věta, interval spolehlivosti, χ^2 test nezávislosti

Title: Continuity correction

Author: Marek Štěpán

Department: Department of Probability and Mathematical Statistics

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

Abstract: For an approximation of discrete random variable, which is the sum of n independent, identically distributed discrete random variables, we can use the central limit theorem. However, it turns out that we can refine this approximation by applying continuity correction. This term is explained in the thesis, and it is illustrated several ways how the continuity correction can be derived. There is also a numerical comparison of the approximation error for the binomial distribution approximation by the normal distribution with the correction for continuity and approximation without the correction. There are also described confidence intervals and χ^2 test of independence in contingency tables in which continuity correction are used. On simulations for various parameters, we will test the properties of these intervals (true confidence level and length) and tests (actual significance level and power).

Keywords: continuity correction, central limit theorem, confidence interval, χ^2 test of independence

Obsah

Úvod

Tématem této práce je oprava na spojitost, která se využívá v situaci, kdy rozdělení diskrétní náhodné veličiny aproximujeme spojitým rozdělením. Pro takovou aproximaci známe centrální limitní větu, která však pro náhodnou veličinu s po částech konstantní distribuční funkcí nedává jednoznačnou aproximaci.

Cílem této bakalářské práce je popsat pojem opravy na spojitost včetně situací, kdy opravu na spojitost můžeme použít, a kdy to naopak vhodné není.

Nejčastěji se s opravou na spojitost setkáváme v souvislosti aproximace binomického rozdělení rozdělením normálním, proto budeme vše v této práci ukazovat právě na tomto rozdělení a uvedeme, jak by se výsledky daly analogicky odvodit pro ostatní diskrétní rozdělení.

V první kapitole vysvětlíme pojem opravy na spojitost, ve druhé pro binomické rozdělení naznačíme, jak byl odvozen tvar aproximace, a to jak z pozorování, tak analyticky. V závěru této kapitoly pak provedeme porovnání chyby aproximace binomického rozdělení s opravou na spojitost a bez ní. Okomentujeme také, zda by aproximace s opravou na spojitost byla dobrá pro další diskrétní rozdělení.

V kapitole 3 nejprve popíšeme tvary intervalových odhadů s opravou na spojitost a poté je budeme porovnávat s intervalovými odhady bez opravy na spojitost. Provedeme simulaci sledující důležité vlastnosti takových intervalů – jejich spolehlivost a délku.

Ve čtvrté kapitole se zaměříme na nejznámější test s opravou na spojitost - test nezávislosti v kontingenčních tabulkách. Nejprve test popíšeme a představíme Yatesovu opravu na spojitost. Poté provedeme simulace skutečné hladiny testů a budeme také diskutovat sílu takových testů.

Značení

Ačkoli je to nestandardní, budeme v této práci používat pro desetinná čísla tečkovou notaci, tedy například $\frac{1}{2} = 0.5$. Důvodem tohoto značení je sjednocení značení desetinných čísel u grafů, výsledků simulací a čísel v samotném textu.

1. Aproximace a oprava na spojitost

Nejjednodušším aproximačním nástrojem je centrální limitní věta, pro naše účely bude použitelná zejména její Lévy-Lindebergova verze. Ta nám říká, že rozdělení součtu nezávislých stejně rozdělených náhodných veličin můžeme aproximovat normálním rozdělením. Větu najdeme například v (Anděl, 2011, str. 331), kde je také odkaz na její důkaz.

Věta 1 (Lévy-Lindebergova CLV)
Buď $X_n, n \in \mathbb{N}$ náhodný výběr z rozdělení s nenulovým konečným rozptylem. Pak platí:

$$\frac{\sum_{i=1}^{n} X_i - nEX_1}{\sqrt{n\ var X_1}} \xrightarrow[n\to\infty]{D} Z,\ kde\ Z \sim N(0,1).$$

Mějme tedy $X_1, \ldots, X_n, n \in \mathbb{N}$ náhodný výběr z nějakého rozdělení s konečným rozptylem a označme $Y_n = \sum_{i=1}^{n} X_i$. Pak hodnotu F_{Y_n} distribuční funkce náhodné veličiny Y_n můžeme v každém bodě aproximovat právě pomocí Lévy-Lindebergovy CLV.

Víme totiž, že díky nezávislosti a stejně rozděleným náhodným veličinám X_i platí:

$$\mathsf{E}\, Y_n = \mathsf{E} \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \mathsf{E}\, X_i = n\ \mathsf{E}\, X_1$$

$$\mathrm{var} Y_n = \mathrm{var} \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \mathrm{var} X_i = n\ \mathrm{var} X_1.$$

Poté pro náhodnou veličinu Y_n dle věty 1 platí:

$$\frac{Y_n - \mathsf{E}\, Y_n}{\sqrt{\mathrm{var}\, Y_n}} \xrightarrow[n\to\infty]{D} Z, \text{ kde } Z \sim N(0,1).$$

Zavedeme symbol $\dot{\approx}$, který bude představovat významově „aproximujeme“. Přesněji využijeme Lévy-Lindebergovu CLV, která říká:

$$\mathsf{P}\Big(\frac{Y_n - \mathsf{E}\, Y_n}{\sqrt{\mathrm{var}\, Y_n}} \le z\Big) \xrightarrow[n\to\infty]{} \Phi(z), z \in \mathbb{R}$$

a použijeme ji pro aproximaci pro konkrétní $n \in \mathbb{N}$:

$$\mathsf{P}\Big(\frac{Y_n - \mathsf{E}\, Y_n}{\sqrt{\mathrm{var}\, Y_n}} \le z\Big) \mathrel{\dot{\approx}} \Phi(z), z \in \mathbb{R}.$$

Jelikož je nyní n pevné, z může být závislé na n a symbol $\dot{\approx}$ bude stále dávat dobrý smysl. Zároveň díky této skutečnosti můžeme upravovat nerovnost $\frac{Y_n - \mathsf{E}\, Y_n}{\sqrt{\mathrm{var}\, Y_n}} \le z$ a dostat tím i tvary aproximací přímo pro náhodnou veličinu Y_n.

Označíme-li $Z_n = \frac{Y_n - \mathsf{E}\, Y_n}{\sqrt{\mathrm{var}\, Y_n}}$ a $z_y = \frac{y - \mathsf{E}\, Y_n}{\sqrt{\mathrm{var}\, Y_n}}$, pak hodnotu distribuční funkce F_{Y_n} v bodě $y \in \mathbb{R}$ můžeme aproximovat:

$$\begin{aligned} F_{Y_n}(y) &= \mathsf{P}(Y_n \le y) = \mathsf{P}\left(\frac{Y_n - \mathsf{E}\, Y_n}{\sqrt{\mathrm{var}\, Y_n}} \le \frac{y - \mathsf{E}\, Y_n}{\sqrt{\mathrm{var}\, Y_n}}\right) \\ &= \mathsf{P}(Z_n \le z_y) \mathrel{\dot{\approx}} \Phi(z_y), \end{aligned} \tag{1.1}$$

kde Φ je distribuční funkce $Z \sim N(0,1)$.

Pro tuto aproximaci bude dle Lévy-Lindebergovy CLV (věta 1) platit, že velikost chyby se zmenšuje s rostoucím n, neboli:

$$|\mathsf{P}(Y_n \leq y) - \Phi(z_y)| \xrightarrow[n\to\infty]{} 0.$$

Tedy pro aproximaci distribuční funkce libovolné náhodné veličiny Y_n, která je součtem konečně mnoha nezávislých stejně rozdělených náhodných veličin, můžeme využít výše odvozené a aproximovat:

$$\widehat{F_{Y_n}(y)} = \Phi\left(\frac{y - \mathsf{E}\, Y_n}{\sqrt{\operatorname{var} Y_n}}\right), y \in \mathbb{R}. \tag{1.2}$$

Poznamenejme ještě, že aproximujeme konkrétní hodnotu distribuční funkce, nejedná se o odhad celé distribuční funkce.

1.1 Oprava na spojitost

Chceme se zabývat případem, kdy Y_n je diskrétní náhodná veličina.

Nechť tedy Y_n je diskrétní náhodná veličina taková, že $Y_n \in \mathbb{N}_0$ skoro jistě. Pro takovou náhodnou veličinu pak platí:

$$\mathsf{P}(Y_n \leq k) = \mathsf{P}(Y_n < k+1), \forall k \in \mathbb{N}_0.$$

Protože distribuční funkce Y_n je po částech konstantní a tyto části mají vždy délku 1, pak platí dokonce:

$$\mathsf{P}(Y_n \leq k) = \mathsf{P}(Y_n \leq k+a), \forall a \in [0,1), \forall k \in \mathbb{N}_0.$$

Pokud tedy Y_n je součtem $n \in \mathbb{N}$ nezávislých stejně rozdělených náhodných veličin, označíme $Z_n = \frac{Y_n - \mathsf{E}\, Y_n}{\sqrt{\operatorname{var} Y_n}}$ a $z_y^a = \frac{y+a-\mathsf{E}\, Y_n}{\sqrt{\operatorname{var} Y_n}}$ a při odvození (1.1) dostaneme aproximaci:

$$\begin{aligned} F_{Y_n}(y) =& \mathsf{P}(Y_n \leq y) = \mathsf{P}(Y_n \leq y+a) = \mathsf{P}\left(\frac{Y_n - \mathsf{E}\, Y_n}{\sqrt{\operatorname{var} Y_n}} \leq \frac{y+a-\mathsf{E}\, Y_n}{\sqrt{\operatorname{var} Y_n}}\right) \\ =& \mathsf{P}(Z_n \leq z_y^a) \approxdot \Phi(z_y^a), a \in [0,1), y \in \mathbb{N}_0. \end{aligned}$$

Máme tedy obecnější tvar aproximace:

$$\widehat{F_{X_n}(x)} = \Phi\left(\frac{x + a - \mathsf{E}\, X_n}{\sqrt{\operatorname{var} X_n}}\right), a \in [0,1).$$

Konstantu $a \in (0,1)$ nazýváme opravou na spojitost a budeme se dále snažit ukázat, jaká hodnota a je optimální a vylepšit tak odhad (1.2). Pokud $a = 0$, budeme říkat, že používáme aproximaci bez opravy na spojitost.

2. Aproximace diskrétního rozdělení

Nejčastější situací, kdy chceme aproximovat diskrétní rozdělení spojitým, je případ aproximace binomického rozdělení normálním rozdělením. Proto se v této kapitole budeme více zajímat o binomické rozdělení. Aproximaci dalších rozdělení okomentujeme v části 2.3.

2.1 Binomické rozdělení

Definice 1 (Binomické rozdělení)
Řekneme, že diskrétní náhodná veličina X_n má binomické rozdělení s parametry $n \in \mathbb{N}$ a $p \in (0,1)$, pokud

$$\mathsf{P}[X_n = k] = \binom{n}{k} p^k (1-p)^{n-k}, \text{ pro } k \in \{0,...,n\}.$$

Značíme $X_n \sim Bi(n,p)$.

Uvedeme ještě několik vlastností binomického rozdělení.

Poznámka. *Jsou-li $Y_i \sim Alt(p)$, $p \in (0,1)$, nezávislé, $i = 1,...,n$, $n \in \mathbb{N}$, pak $X_n = \sum_{i=1}^{n} Y_i$ má binomické rozdělení s parametry n a p.*

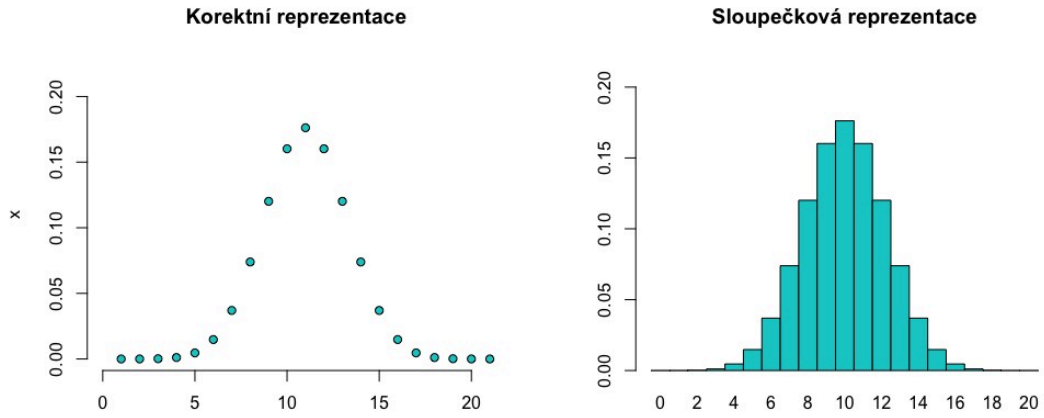
Poznámka (vlastnosti binomického rozdělení)**.** *Nechť $X_n \sim Bi(n,p)$. Potom:*

1. $\mathsf{E} X_n = np$,
2. $var X_n = np(1-p)$.

Tvrzení 2
Nechť $X_n \sim Bi(n,p)$. Potom $\frac{X_n}{n}$ je nestranným a konzistentním odhadem parametru p.

2.1.1 Znázornění binomického rozdělení v grafu

Jelikož budeme chtít porovnávat binomické rozdělení s rozdělením spojitým, potřebujeme si reprezentovat binomické rozdělení v grafu. Korektní reprezentace by mohla vypadat podobně jako na levé straně obrázku 2.1, jelikož dostáváme vyjádření hustoty binomického rozdělení vůči čítací míře na $\mathbb{Z}$. My však budeme pracovat s reprezentací na pravé straně, kde k-tý „sloupeček“ reprezentuje pravděpodobnost nabytí hodnoty k. Jelikož je šířka tohoto sloupce rovna 1 a výška rovna pravděpodobnosti nabytí hodnoty k, tak zároveň i plocha sloupečku je rovna pravděpodobnosti nabytí hodnoty k.



Obrázek 2.1: Reprezentace binomického rozdělení v grafu

2.2 Aproximace distribuční funkce binomického rozdělení

Lévy-Lindebergova centrální limitní věta nám dává tvar aproximace distribuční funkce pro binomické rozdělení. Jelikož je binomické rozdělení součtem nezávislých stejně rozdělených alternativních rozdělení, odvodili jsme tuto aproximaci v kapitole 1.1.

Hodnotu $F_{X_n}(x)$ tedy pro $x \in \{0,1,\ldots,n\}$ můžeme odhadnout:

$$\widehat{F_{X_n}(x)} = \Phi\left(\frac{x + a - \mathsf{E}\, X_n}{\sqrt{\operatorname{var} X_n}}\right), a \in [0,1).$$

2.2.1 Odhad z grafu

Budeme se snažit najít optimální hodnotu a, nejprve si však situaci pro $a = 0$ vykreslíme do grafu. Použijeme při tom sloupečkovou reprezentaci binomického rozdělení (obrázek 2.1 vpravo).

Na grafu vlevo na obrázku 2.2 je zakresleno binomické rozdělení X_{20} pro $p = 0.5$ a hustota normálního rozdělení se stejnou střední hodnotou a rozptylem jako X_{20}. A protože platí:

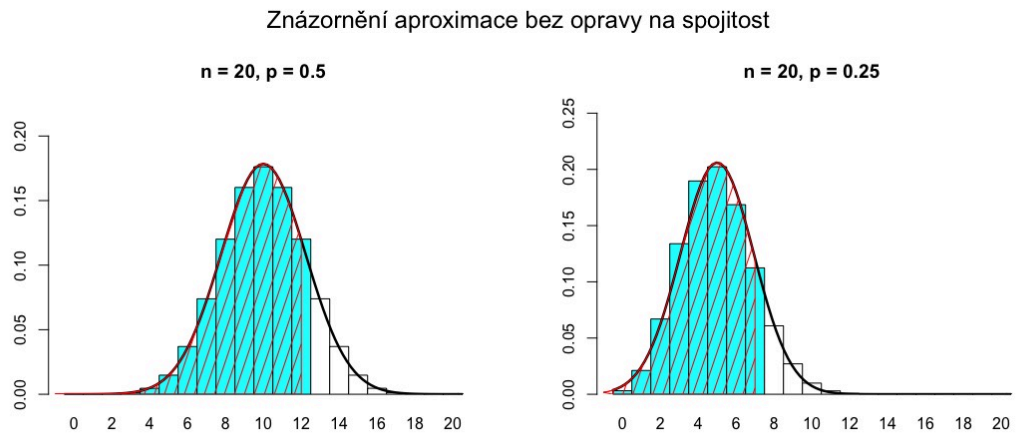
$$\Phi\left(\frac{x + a - \mathsf{E}\, X_n}{\sqrt{\operatorname{var} X_n}}\right) = F_{V_n}(x + a), \text{ kde } V_n \sim N\Big(\mathsf{E}\, X_n, \operatorname{var} X_n\Big),$$

tak se vlastně jedná o normální rozdělení, kterým aproximujeme.

Modře vybarvená plocha sloupečků na grafu vlevo je tedy odhadovaná hodnota $F_{X_{20}}(12)$, červeně vyšrafovaná plocha je hodnota $\widehat{F_{X_{20}}(12)}$ pro $a = 0$.

Z obrázku je patrné, že $a = 0$ zřejmě nebude optimální hodnota, jelikož celá polovina 12. sloupečku zůstává nepokrytá. Podobně odhadneme, že také a blízké 1 nebude optimální, jelikož bude přebývat téměř polovina 13. sloupečku.

Můžeme si také všimnout, že pro každý sloupeček plocha, kde se červeně vyšrafovaná oblast nepřekrývá s modrým sloupečkem, přibližně odpovídá ploše,



Obrázek 2.2: Znázornění aproximace distribuční funkce binomického rozdělení rozdělením normálním bez opravy na spojitost

kde naopak modrý sloupeček není překryt červenou oblastí. To nás vede k úvaze, že plochy $F_{X_n}(x)$ a $F_{V_n}(v+0.5)$ by mohly být přibližně stejné.

Podobně to bude vypadat i pro $p \neq 0.5$, jak je vykresleno vpravo na obrázku 2.2 pro $p = 0.25$. Vidíme však, že pro p vzdálená od 1/2 by mohla být aproximace s opravou na spojitost 0.5 horší. Jak ukážeme dále, bude ve většině případů stále lepší, než aproximace bez opravy na spojitost. Pouze pro extrémně nízká p je přínos opravy na spojitost diskutabilní.

2.2.2 Analytické odvození

Analytické odvození aproximace binomického rozdělení normálním můžeme najít v (Feller, 1968, str. 174 - 186), kde je na straně 184 důležité tvrzení:

Tvrzení 3
Mějme $X_n \sim Bi(n,p)$. Nechť $K_n \geq 0$ je takové, že $\frac{K_n^3}{n^2} \to \infty$. Nechť m je nejmenší přirozené číslo splňující $m \geq \mathsf{E}X_n$ a $a_k = \mathsf{P}(X_n = m+k)$. Pak pro každé $\epsilon > 0$ najdeme n_0 tak, že pro všechna $n \geq n_0$ bude platit:

$$1 - \epsilon < \frac{a_k}{h\phi(kh)} < 1 + \epsilon, \forall k \in \{0,...,K_n\},$$

kde $h = \frac{1}{\sqrt{var\, X_n}}$ a $\phi(x)$ je hustota normovaného normálního rozdělení.

Tvrzení nám vlastně dává stejnoměrnou konvergenci a_k k $h\phi(kh)$ pro všechna k, která nejsou příliš blízko n nebo 0. Ze symetrie $Bi(n,p)$ a $Bi(n,(1-p))$ totiž dostaneme, že věta bude platit i pro $a_{-k}, k \in \{0,...,K\}$. Z toho plyne, že $a_k \doteq h\phi(kh)$, tudíž pro $\alpha, \beta \in \mathbb{N}_0$:

$$\mathsf{P}(\alpha \leq X_n \leq \beta) = \sum_{k=\alpha-m}^{\beta-m} a_k \doteq \sum_{k=\alpha-m}^{\beta-m} h\phi(kh).$$

A tedy tuto pravděpodobnost aproximujeme sumou hodnot hustoty normálního rozdělení, kde je každý člen násobený h, což můžeme chápat, jako plochu obdélníků s šířkou h a výškou $\phi(kh)$.

Nyní můžeme sumu zaměnit za integrál dané funkce, jelikož pokud $n \to \infty$, pak také $h \to 0$. Jedná se tedy vlastně o Riemannův integrál. Takovou aproximaci provedeme právě tak, že budeme integrovat od $\alpha - m - 0.5$ do $\beta - m + 0.5$. Důvod pro právě tyto meze plyne z vylepšení chyby při numerické aproximaci integrálu při použití metody midpoint[1], které dává chybu takové aproximace $o(h^2)$[1]. Pokud bychom použili původní aproximaci integrálem s mezemi od $\alpha - m$ do $\beta - m$ (tedy metodu left-point[1]), bude chyba jen $o(h)$[1]. Proto, jak odvodil Feller (1968), je vhodné aproximovat:

$$\mathsf{P}(\alpha \leq X_n \leq \beta) \doteq \int_{\alpha-m-0.5}^{\beta-m+0.5} \phi(th)dt = \Phi\Big((\beta - m + 0.5)h\Big) - \Phi\Big((\alpha - m - 0.5)h\Big).$$

Z toho už jednoduchou úvahou, kdy dolní mez pošleme do $-\infty$ dostáváme naši myšlenku:

$$F_{X_n}(\beta) \doteq \Phi\Big((\beta - m + 0.5)h\Big).$$

2.2.3 Numerická ilustrace správnosti aproximace s opravou na spojitost

Nyní se pokusíme numericky ukázat, že hodnota $a = 0.5$ zajišťuje mnohem lepší aproximaci než $a = 0$. Budeme tyto dva případy nazývat aproximace s opravou na spojitost, resp. bez opravy na spojitost. Mějme tedy $X_n \sim Bi(n,p)$. Označíme pro zvolené $p \in (0.1)$:

$$D_n(x) = \left| F_{X_n}(x) - \Phi\left(\frac{x + a - np}{\sqrt{np(1-p)}}\right)\right|.$$

$D_n(x)$ spočítáme pro $n = 10, 100, 1000$, $p = 0.5, 0.25, 0.05, 0.95$ a pro

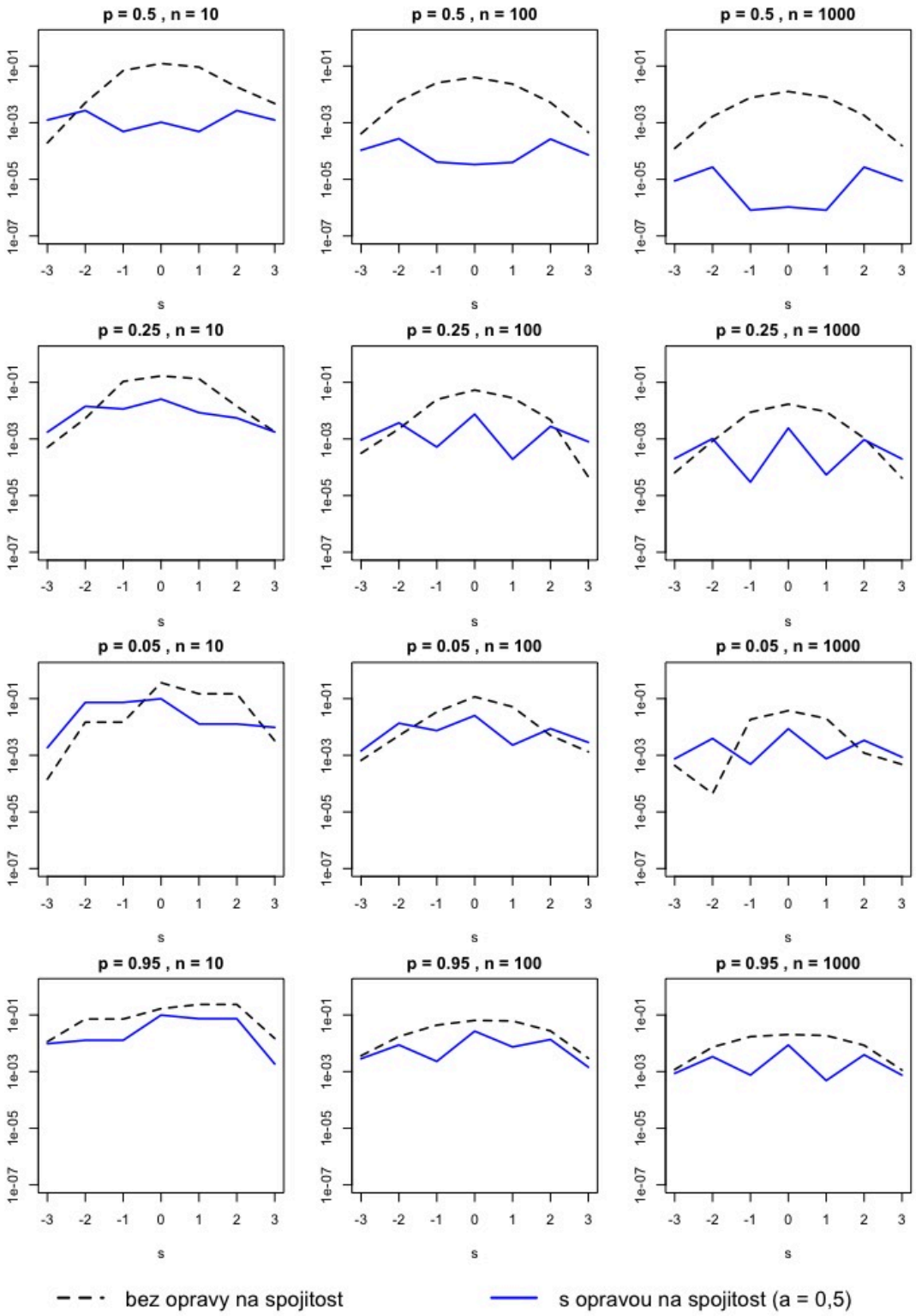
$$x = \left\lfloor \mathsf{E}\, X_n + s\sqrt{\mathrm{var} X_n} \right\rfloor, \text{ kde } s = -3, -2, -1, 0, 1, 2, 3. \tag{2.1}$$

Na těchto dvanácti grafech (obrázek 2.3), které na ose x mají koeficient s podle tvaru (2.1) a na ose y vykreslují $D_n(x)$ v logaritmickém měřítku, ilustrujeme rozdíl mezi chybou pro aproximaci s opravou na spojitost (modře) a bez opravy na spojitost (černě). Logaritmické měřítko používáme, protože rozdíl mezi aproximací s použitím opravy na spojitost a aproximací bez opravy může dosahovat až čtyř řádů (např. graf vpravo nahoře), ale také může být velmi malý (grafy v posledním řádku).

Vidíme, že pro $p = 0.5$ je aproximace s opravou na spojitost o mnoho lepší pro dostatečně velké n. Pro $n = 1000$ dostáváme pro x z okolí střední hodnoty X_n dokonce rozdíl čtyř řádů, tedy tam už je aproximace s opravou na spojitost výrazně lepší, než aproximace bez opravy na spojitost.

Pro $p = 0.25$ platí, že pro x hodně vzdálená od $\mathsf{E}\, X_n$ bude aproximace s opravou na spojitost horší. V situaci, kdy $p = 0.05$ už vidíme, že pro malá n je aproximace s opravou na spojitost lepší jen přibližně v polovině případů, což se ale se zvětšujícím n zlepšuje. Na posledním řádku vidíme, že rozdíl mezi aproximací se nezhoršuje jen pro snižující se p, ale i pro p větší než 0.5. Avšak pro taková p je aproximace s $a = 0.5$ pořád lepší, než pro $a = 0$.

[1]Popis metod midpoint, left-point a odvození řádu chyby najdeme v (Nagel, 2012, str. 3-6).



Obrázek 2.3: Porovnání chyby aproximace distribuční funkce binomického rozdělení s opravou na spojitost (a = 0.5) a bez ní

Celkově můžeme porovnáním grafů vysledovat:

- Černé čáry jdou se zvyšujícím n pro všechna $p \in (0,1)$ směrem k 0. To je v souladu s centrální limitní větou.
- Modré čáry se zvyšujícím se n klesají ještě rychleji než černé, což naznačuje, že pro velká n je aproximace s opravou na spojitost výrazněji lepší než bez opravy na spojitost.
- Rozdíl obou aproximací je největší pro p blízká 0.5. Pro p blízké 0 je méně spolehlivá, zejména pro velmi nízké n už není obecně lepší než aproximace bez opravy na spojitost. Naopak pro p blízké 1 dostáváme, že aproximace s opravou je stále lepší než bez opravy na spojitost.
- Hodnota $D_n(x)$ pro aproximaci s opravou na spojitost nejeví známky žádného pravidelného chování - pouze pro x vzdálená od $\mathsf{E}\, X_n$ právě jednu směrodatnou odchylku (tj. $s = -1, 1$) je chyba nižší, než pro $x = \mathsf{E}\, X_n$.

2.3 Aproximace ostatních diskrétních rozdělení

Dosud jsme popisovali pouze aproximaci binomického rozdělení, při kterém se používá oprava na spojitost nejčastěji. Avšak oprava na spojitost bude fungovat i pro další diskrétní náhodné veličiny, které jsou součtem nezávislých diskrétních náhodných veličin.

Pokud bychom pozorovali chyby aproximace například pro Poissonovo a negativně binomické rozdělení, ukazuje se, že výhoda při aproximaci ostatních rozdělení s opravou na spojitost je menší než v případě binomického rozdělení. Závěry jsou však stejné – pokud nemáme extrémní hodnoty parametrů, bude aproximace s opravou na spojitost lepší než bez ní.

Kromě binomického rozdělení tak můžeme použít aproximaci hodnot distribuční funkce s opravou na spojitost například u Poissonova rozdělení nebo u negativně binomického rozdělení. Tato aproximace pak bude mít stejný tvar, jako jsme odvodili u binomického rozdělení, tedy

$$\widehat{F_{X_n}(x)} = \Phi\left(\frac{x + 0.5 - \mathsf{E}\, X_n}{\sqrt{\operatorname{var} X_n}}\right), x \in \mathbb{N}_0.$$

2.4 Shrnutí aproximace distribuční funkce s opravou na spojitost

V této podkapitole nejdříve rozšíříme aproximaci i pro $x \notin \mathbb{N}_0$ a poté shrneme dosavadní poznatky.

2.4.1 Aproximace distribuční funkce pro $x \in \mathbb{R}$

Doposud jsme uvažovali aproximace distribuční funkce v bodech $k \in \mathbb{N}$. Chceme rozšířit aproximaci i pro $x \in \mathbb{R}$, pro která to dává smysl. Pro náhodnou veličinu Y_n s hodnotami v $\mathbb{N}_0$ platí $\mathsf{P}(Y_n \geq 0) = 1$, tedy i když by bylo možné

tuto aproximaci použít pro všechna $x \in \mathbb{R}$, budou nás zajímat pouze $x \geq 0$. Distribuční funkci v bodě $x < 0$ je totiž zjevně nejlepší odhadnout 0. Pro takové $x \geq 0$ bude platit:

$$\begin{aligned} \mathsf{P}(Y_n \leq x) &= \mathsf{P}(Y_n \leq \lfloor x \rfloor) \\ \mathsf{P}(Y_n < x) &= \mathsf{P}(Y_n < \lceil x \rceil), \end{aligned} \tag{2.2}$$

kde $\lfloor x \rfloor$ je dolní celá část x a $\lceil x \rceil$ je horní celá část x.

2.4.2 Počítání pravděpodobností s opravou na spojitost

V této kapitole jsme aproximovali hodnotu distribuční funkce náhodné veličiny X_n, která je realizována na $\mathbb{N}_0$ a je součtem nezávislých stejně rozdělených diskrétních náhodných veličin. Ukázali jsme, že v případech, které nejsou extrémní, je hodnota opravy na spojitost $a = 0.5$ vhodná.

Pokud tedy budeme chtít pro takovou X_n počítat pravděpodobnosti s ostrou či neostrou nerovností pomocí aproximace normálním rozdělením, je vhodné pro $x, x_1, x_2 \geq 0$, $x_1 \neq x_2$ počítat pomocí těchto pravidel:

$$\mathsf{P}(X_n \leq x) \doteq \Phi\left(\frac{\lfloor x \rfloor + 0.5 - \mathsf{E}\, X_n}{\sqrt{\operatorname{var} X_n}}\right) \tag{2.3}$$

$$\mathsf{P}(X_n < x) \doteq \Phi\left(\frac{\lceil x \rceil - 0.5 - \mathsf{E}\, X_n}{\sqrt{\operatorname{var} X_n}}\right) \tag{2.4}$$

$$\mathsf{P}(x_1 \leq X_n \leq x_2) \doteq \Phi\left(\frac{\lfloor x_2 \rfloor + 0.5 - \mathsf{E}\, X_n}{\sqrt{\operatorname{var} X_n}}\right) - \Phi\left(\frac{\lceil x_1 \rceil - 0.5 - \mathsf{E}\, X_n}{\sqrt{\operatorname{var} X_n}}\right) \tag{2.5}$$

$$\mathsf{P}(x_1 < X_n \leq x_2) \doteq \Phi\left(\frac{\lfloor x_2 \rfloor + 0.5 - \mathsf{E}\, X_n}{\sqrt{\operatorname{var} X_n}}\right) - \Phi\left(\frac{\lfloor x_1 \rfloor + 0.5 - \mathsf{E}\, X_n}{\sqrt{\operatorname{var} X_n}}\right) \tag{2.6}$$

$$\mathsf{P}(x_1 \leq X_n < x_2) \doteq \Phi\left(\frac{\lceil x_2 \rceil - 0.5 - \mathsf{E}\, X_n}{\sqrt{\operatorname{var} X_n}}\right) - \Phi\left(\frac{\lceil x_1 \rceil - 0.5 - \mathsf{E}\, X_n}{\sqrt{\operatorname{var} X_n}}\right) \tag{2.7}$$

$$\mathsf{P}(x_1 < X_n < x_2) \doteq \Phi\left(\frac{\lceil x_2 \rceil - 0.5 - \mathsf{E}\, X_n}{\sqrt{\operatorname{var} X_n}}\right) - \Phi\left(\frac{\lfloor x_1 \rfloor + 0.5 - \mathsf{E}\, X_n}{\sqrt{\operatorname{var} X_n}}\right). \tag{2.8}$$

Vztahy (2.3), (2.4) plynou jako důsledek rovností (2.2), vztah (2.5) byl odvozen z vlastností distribuční funkce: $\mathsf{P}(x_1 \leq X_n \leq x_2) = \mathsf{P}(X_n \leq x_2) - \mathsf{P}(X_n < x_1)$, (2.6), (2.7), (2.8) obdobně.

Poznámka. *Povšimněme si také, že pro $x \notin \mathbb{N}_0$ bude platit:*

$$P(X_n \leq x) = P(X_n < x),$$

a také tedy:

$$\Phi\left(\frac{\lfloor x \rfloor + 0.5 - E\, X_n}{\sqrt{var\, X_n}}\right) = \Phi\left(\frac{\lceil x \rceil - 0.5 - E\, X_n}{\sqrt{var\, X_n}}\right).$$

3. Využití opravy na spojitost v konstrukci intervalových odhadů

V této kapitole představíme intervalové odhady s využitím opravy na spojitost. Nejprve popíšeme odkud vycházejí a jak jsou rozdílné oproti klasickým intervalovým odhadům bez opravy na spojitost. V závěru této kapitoly pak provedeme simulace, kde tyto předpoklady ověříme pro různé hodnoty parametrů.

V celé kapitole budeme pracovat pouze s binomickým rozdělením, ale snadno lze analogicky odvodit příslušné intervalové odhady pro další náhodné veličiny, které jsou součtem nezávislých náhodných veličin a mají hodnoty v $\mathbb{N}_0$.

Budeme chtít sestavit intervalové odhady pro neznámý parametr p_x, parametr n nám bude známý. Označme tedy ještě

$$\widehat{p_n} = \frac{X_n}{n}.$$

Z tvrzení 2 víme, že se jedná o nestranný a konzistentní odhad p.

3.1 Dolní intervalový odhad

Nechť $X_n \sim Bi(n,p_x)$. Pro sestavení dolního intervalu spolehlivosti klasickou asymptotickou metodou vyjdeme standardně z tvaru, který nám dává centrální limitní věta, navíc upravený s pomocí Cramér-Sluckého věty[1] (víme, že dle věty o spojité transformaci[2] a tvrzení 2 je $\sqrt{(\widehat{p_n}(1-\widehat{p_n})}$ nestranný odhad $\sqrt{(p_x(1-p_x))}$):

$$\mathsf{P}\left(\frac{\sqrt{n}(\widehat{p_n}-p_x)}{\sqrt{(\widehat{p_n}(1-\widehat{p_n})}} < u_{1-\alpha}\right) \xrightarrow[n\to\infty]{} 1-\alpha.$$

Do tvaru, se kterým jsme pracovali dosud, se dostaneme jednoduše úpravou:

$$\begin{aligned}\mathsf{P}\left(\frac{\sqrt{n}(\widehat{p_n}-p_x)}{\sqrt{\widehat{p_n}(1-\widehat{p_n})}} < u_{1-\alpha}\right) &= \mathsf{P}\left(\widehat{p_n} < p_x + u_{1-\alpha}\sqrt{\frac{\widehat{p_n}(1-\widehat{p_n})}{n}}\right)\\ &= \mathsf{P}\Big(X_n < np_x + u_{1-\alpha}\sqrt{n\widehat{p_n}(1-\widehat{p_n})}\Big).\end{aligned}$$

V tomto tvaru můžeme použít navrhovanou opravu na spojitost. Použijeme pravidlo (2.3), což můžeme podle poznámky na konci kapitoly 2.4.2, protože pravá strana nebude téměř nikdy číslo z $\mathbb{N}_0$. A pokud by shodou okolností bylo, pak tím pravděpodobnost zvýšíme, tedy dodržíme předepsanou spolehlivost.

Na pravé straně tedy přidáme $+0.5$ jako jakousi opatrnost:

$$\mathsf{P}\Big(X_n < np_x + 0.5 + u_{1-\alpha}\sqrt{n\widehat{p_n}(1-\widehat{p_n})}\Big).$$

[1] Můžeme najít například na straně 333 v (Anděl, 2011).

[2] Najdeme formulovanou na straně 332 v (Anděl, 2011).

Pak stačí upravit a dostáváme:

$$\mathsf{P}\left(p_x \geq \widehat{p_n} - \frac{0.5}{n} - \frac{u_{1-\alpha}\sqrt{n\widehat{p_n}(1-\widehat{p_n})}}{n}\right).$$

Máme tedy dolní intervalový odhad s opravou na spojitost pro parametr p v binomickém rozdělení:

$$\left(\widehat{p_n} - \frac{0.5}{n} - \frac{u_{1-\alpha}\sqrt{n\widehat{p_n}(1-\widehat{p_n})}}{n}, 1\right).$$

Vidíme také rovnou rozdíl oproti intervalovému odhadu bez opravy na spojitost:

$$\left(\widehat{p_n} - \frac{u_{1-\alpha}\sqrt{n\widehat{p_n}(1-\widehat{p_n})}}{n}, 1\right).$$

Z tohoto rozdílu vyplývá, že interval s opravou na spojitost bude delší o $\frac{0.5}{n} = \frac{1}{2n}$ než interval bez opravy na spojitost. Diskuzi, zda má poté i vyšší spolehlivost, provedeme v kapitole 3.4.1.

3.2 Horní intervalový odhad

Vyjdeme ze standardního tvaru podobně jako v odvození dolního intervalového odhadu:

$$\mathsf{P}\left(\frac{\sqrt{n}(\widehat{p_n} - p_x)}{\sqrt{(\widehat{p_n}(1-\widehat{p_n})}} > u_\alpha\right) \xrightarrow[n\to\infty]{} 1-\alpha.$$

Obdobným postupem jako u dolního intervalového odhadu dostaneme:

$$\mathsf{P}\Big(p_x < \frac{1}{n}\Big(X_n + 0.5 + u_{1-\alpha}\sqrt{n\widehat{p_n}(1-\widehat{p_n})}\Big)\Big).$$

Tedy horní intervalové odhady s opravou na spojitost a bez ní mají tvar:

$$\left(0, \widehat{p_n} + \frac{0.5}{n} + \frac{u_{1-\alpha}\sqrt{n\widehat{p_n}(1-\widehat{p_n})}}{n}\right),$$

$$\left(0, \widehat{p_n} + \frac{u_{1-\alpha}\sqrt{n\widehat{p_n}(1-\widehat{p_n})}}{n}\right).$$

3.3 Oboustranný intervalový odhad

Zde také vyjdeme ze standardního výsledku centrální limitní věty upraveného pomocí Cramér-Sluckého věty:

$$\mathsf{P}\left(u_{\alpha/2} < \frac{\sqrt{n}(\widehat{p_n} - p_x)}{\sqrt{(\widehat{p_n}(1-\widehat{p_n})}} < u_{1-\alpha/2}\right) \xrightarrow[n\to\infty]{} 1-\alpha.$$

Převedeme na tvar:

$$\mathsf{P}\Big(np_x - u_{1-\alpha/2}\sqrt{n\widehat{p_n}(1-\widehat{p_n})} < X_n < np_x + u_{1-\alpha/2}\sqrt{n\widehat{p_n}(1-\widehat{p_n})}\Big).$$

A použijeme opravu na spojitost analogicky jako v dolním a horní odhadu, tentokrát na obou stranách:

$$\mathsf{P}\Big(np_x - 0.5 - u_{1-\alpha/2}\sqrt{n\widehat{p_n}(1-\widehat{p_n})} < X_n < np_x + 0.5 + u_{1-\alpha/2}\sqrt{n\widehat{p_n}(1-\widehat{p_n})}\Big).$$

Osamostatníme p_x:

$$\mathsf{P}\left(\widehat{p_n} - \frac{0.5}{n} - \frac{u_{1-\alpha}\sqrt{n\widehat{p_n}(1-\widehat{p_n})}}{n} < p_x < \widehat{p_n} + \frac{0.5}{n} + \frac{u_{1-\alpha}\sqrt{n\widehat{p_n}(1-\widehat{p_n})}}{n}\right).$$

Dostáváme tedy oboustranné intervalové odhady s opravou na spojitost a bez opravy:

$$\left(\widehat{p_n} - \frac{0.5}{n} - \frac{u_{1-\alpha}\sqrt{n\widehat{p_n}(1-\widehat{p_n})}}{n}, \widehat{p_n} + \frac{0.5}{n} + \frac{u_{1-\alpha}\sqrt{n\widehat{p_n}(1-\widehat{p_n})}}{n}\right),$$

$$\left(\widehat{p_n} - \frac{u_{1-\alpha}\sqrt{n\widehat{p_n}(1-\widehat{p_n})}}{n}, \widehat{p_n} + \frac{u_{1-\alpha}\sqrt{n\widehat{p_n}(1-\widehat{p_n})}}{n}\right).$$

3.4 Porovnání spolehlivosti intervalových odhadů s opravou a bez opravy na spojitost

V této kapitole budeme chtít ukázat vlastnosti intervalových odhadů s opravou na spojitost a bez ní. Bude nás zajímat odhad skutečné spolehlivosti a také délka intervalu.

Simulaci provedeme následovně[3]:

1. Zvolíme parametry n, p binomického rozdělení.
2. Zvolíme B - počet náhodných výběrů.
3. Pro $b = 1, \ldots, B$:
   - Generujeme náhodně $X_n^b \sim Bi(n,p)$.
   - Spočteme I_b interval spolehlivosti bez opravy na spojitost a I_b^c interval spolehlivosti s opravou na spojitost.
   - Spočteme délky D_b a D_b^c intervalů I_b,I_b^c.
4. Získáme odhady skutečného pokrytí:

$$\frac{1}{B}\sum_{b=1}^{B} \mathbb{I}[I_b \ni p],$$
$$\frac{1}{B}\sum_{b=1}^{B} \mathbb{I}[I_b^c \ni p].$$

[3]Příslušný skript bude k práci přiložen jako elektronická příloha.

5. Spočteme průměrné délky těchto intervalů.

Pro naše účely volíme:

$$\begin{aligned} n &= 10, 30, 100, 300, 1000, 3000, \\ p &= 0.5, 0.25, 0.05, 0.95, \\ B &= 100\,000, \\ \alpha &= 0.05. \end{aligned}$$

3.4.1 Výsledky simulace

Vytvoříme tři tabulky - pro oboustranný, horní a dolní intervalový odhad. V těchto tabulkách budeme ve sloupcích měnit parametr p, v řádcích pak budeme mít pro každé zvolené n nejprve odhad spolehlivosti a pod ním průměrnou délku tohoto intervalu.

	p = 0.5		p = 0.25		p = 0.05		p = 0.95	
	bez	s	bez	s	bez	s	bez	s
n = 10	0.890	0.890	0.924	0.940	0.400	0.401	0.399	0.400
	0.579	0.666	0.459	0.524	0.130	0.181	0.130	0.180
n = 30	0.958	0.958	0.941	0.955	0.784	0.786	0.782	0.785
	0.352	0.385	0.302	0.335	0.115	0.133	0.115	0.132
n = 100	0.944	0.964	0.946	0.946	0.878	0.958	0.877	0.959
	0.195	0.205	0.169	0.179	0.081	0.090	0.082	0.090
n = 300	0.942	0.956	0.944	0.960	0.926	0.957	0.926	0.957
	0.113	0.116	0.098	0.101	0.049	0.052	0.049	0.052
n = 1000	0.946	0.954	0.947	0.955	0.941	0.954	0.941	0.953
	0.062	0.063	0.054	0.055	0.027	0.028	0.027	0.028
n = 3000	0.950	0.955	0.950	0.950	0.948	0.954	0.949	0.955
	0.036	0.036	0.031	0.031	0.016	0.016	0.016	0.016

Tabulka 3.1: Odhad skutečné spolehlivosti a průměrná délka oboustranného intervalového odhadu bez a s opravou na spojitost

Oboustranný interval spolehlivosti

V tabulce 3.1 pro oboustranný intervalový odhad vidíme, že rozdíl mezi použitím opravy na spojitost a jejím nepoužitím není velký. Obecně pro všechny hodnoty platí, že pravděpodobnost pokrytí je vyšší u varianty se spojitostí, avšak délka takového intervalu je větší. Konkrétně vidíme, že je větší přibližně o $\frac{0.5+0.5}{n} = \frac{1}{n}$, což je v souladu s očekáváním. Přibližně proto, že pro intervalové odhady přesahující interval (0,1) jsme příslušný interval pronikli s intervalem (0,1).

Pro hodnotu parametru $p = 0.5$ vidíme, že jsou rozdíly ve spolehlivosti velmi malé. Avšak tyto rozdíly pozorujeme na hranici zvolené spolehlivosti – intervalový

odhad s opravou na spojitost ji dodržuje už pro menší n, odhad bez opravy na spojitost se sice svou spolehlivostí příliš neliší, jeho odhadnutá spolehlivost je ale pod námi chtěnou spolehlivostí.

U intervalového odhadu bez opravy na spojitost vidíme, že až na výjimku pro $n = 30$, která bude zřejmě způsobena malým n a příhodnou volbou parametrů, bude jeho spolehlivost alespoň 0.95 pro n vyšší než 1000.

Pro $p = 0.25$ vidíme, že rozdíly jsou již znatelné, s opravou častěji dodržíme spolehlivost, avšak za cenu delšího intervalu.

Extrémní hodnoty $p = 0.05$ a $p = 0.95$, jak plyne z posledních dvou sloupců, je potřeba odhadovat z výběru o rozsahu alespoň 100, což potvrzuje pravidlo, že aby byly intervaly použitelné, mělo by platit $np \geq 5$ a $n(1-p) \geq 5$. Rozdíly v odhadované spolehlivosti jsou pak pro $n \geq 100$ značné. Jelikož rozdíl délek intervalů se s zvětšujícím n snižuje, vyplatí se použít variantu s opravou na spojitost i pro velká n. Tam má vyšší spolehlivost a přitom velmi podobnou délku.

Celkově v tabulce vidíme, že pro opatrnost se vyplatí použít při sestavování asymptotických intervalových odhadů pro binomické rozdělení vždy intervalový odhad s opravou na spojitost. Dodržuje totiž předepsanou spolehlivost už pro mnohem menší n a pro n velká zároveň není interval výrazně delší.

Jednostranné intervaly spolehlivosti

Dále uvádíme tabulky 3.2 a 3.3 pro horní intervalový odhad a dolní intervalový odhad. Vyplývají z nich podobná pozorování, pouze rozdíly jsou pro jednostranný interval zpravidla menší. Délky takových intervalů by se měly lišit o přibližně $1/2n$, což odpovídá naměřeným hodnotám v tabulkách.

Pro horní odhad dostáváme pro $p = 0.05$ obě varianty se shodnou spolehlivostí. To nastává ze dvou důvodů: Protože se jedná o extrémní hodnotu p, nemusí se oprava na spojitost projevit a také vhodně zvolené n může výhodu opravy na spojitost smazat. Podobně pak pro dolní odhad a hodnotu parametru $p = 0.95$.

Zároveň si také můžeme všimnout, že ve třetím sloupci horního odhadu máme podobné hodnoty jako ve čtvrtém sloupci u dolního odhadu a také naopak. To není náhoda, jelikož jsou tyto hodnoty zvoleny symetricky a také tvary jednostranných intervalových odhadů mají symetrický tvar, bude skutečná spolehlivost stejná.

	p = 0.5		p = 0.25		p = 0.05		p = 0.95	
	bez	s	bez	s	bez	s	bez	s
n = 10	0.946	0.946	0.944	0.944	0.401	0.401	0.988	0.999
	0.745	0.792	0.455	0.505	0.117	0.167	0.999	1.000
n = 30	0.952	0.952	0.901	0.962	0.786	0.786	0.984	0.997
	0.647	0.664	0.377	0.394	0.106	0.122	0.996	0.999
n = 100	0.954	0.954	0.936	0.936	0.882	0.882	0.971	0.988
	0.582	0.587	0.321	0.326	0.085	0.090	0.984	0.988
n = 300	0.954	0.954	0.940	0.955	0.935	0.935	0.970	0.983
	0.547	0.549	0.291	0.293	0.070	0.072	0.971	0.972
n = 1000	0.947	0.954	0.943	0.951	0.940	0.940	0.961	0.971
	0.526	0.526	0.273	0.273	0.061	0.062	0.961	0.962
n = 3000	0.952	0.952	0.949	0.949	0.941	0.941	0.955	0.962
	0.515	0.515	0.263	0.263	0.057	0.057	0.957	0.957

Tabulka 3.2: Odhad skutečné spolehlivosti a průměrná délka horního intervalového odhadu bez a s opravou na spojitost

	p = 0.5		p = 0.25		p = 0.05		p = 0.95	
	bez	s	bez	s	bez	s	bez	s
n = 10	0.945	0.945	0.980	0.980	0.989	0.999	0.402	0.402
	0.744	0.791	0.942	0.966	0.999	1.000	0.118	0.168
n = 30	0.949	0.949	0.949	0.979	0.984	0.997	0.784	0.784
	0.647	0.664	0.877	0.893	0.996	0.999	0.105	0.122
n = 100	0.954	0.954	0.955	0.972	0.972	0.988	0.882	0.882
	0.582	0.587	0.821	0.826	0.984	0.988	0.085	0.090
n = 300	0.953	0.953	0.950	0.963	0.971	0.983	0.936	0.936
	0.547	0.549	0.791	0.793	0.971	0.972	0.071	0.072
n = 1000	0.947	0.953	0.956	0.956	0.962	0.972	0.940	0.940
	0.526	0.526	0.772	0.773	0.961	0.962	0.061	0.062
n = 3000	0.952	0.952	0.951	0.955	0.954	0.962	0.942	0.942
	0.515	0.515	0.763	0.763	0.957	0.957	0.057	0.057

Tabulka 3.3: Odhad skutečné spolehlivosti a průměrná délka dolního intervalového odhadu bez a s opravou na spojitost

4. Test nezávislosti v kontingenčních tabulkách

V této kapitole se chceme zaměřit na nejznámější test s opravou na spojitost – χ^2 test nezávislosti v kontingenčních tabulkách. Konkrétně nás bude zajímat nejčastější situace – kontingenční tabulky 2×2. Variantu tohoto testu s opravou na spojitost navrhl Yates (1934).

V průběhu 20. století byla předmětem neshod mezi statistiky a knihy podporují jednu ze tří možností: používat Yatesovu opravu, když je některá z očekávaných četností menší než 5, menší než 10 anebo používat Yatesovu opravu na spojitost vždy. Pro námi zvolené parametry budeme tuto situaci diskutovat na konci kapitoly.

Nejdříve popíšeme multinomické rozdělení a kontingenční tabulky včetně pohledů, kterými se na ně můžeme dívat. Pak popíšeme test nezávislosti pro kontingenční tabulky 2×2 bez opravy na spojitost a s opravou. V poslední části této kapitoly vyzkoušíme jejich vlastnosti.

4.1 Multinomické rozdělení

Definice 2 (Multinomické rozdělení)
Nechť $K \geq 2$ a n jsou přirozená čísla a $\boldsymbol{p} = (p_1,...,p_K)^\top$ je vektor splňující $p_k \geq 0, \forall k \in \{1,...,K\}$ a $\sum_{k=1}^K p_k = 1$. Náhodný vektor $\boldsymbol{X} = (X_1,...,X_K)^\top$ má multinomické rozdělení $Mult_K(n,\boldsymbol{p})$ právě když jeho hustota vzhledem k součinové čítací míře na $\mathbb{Z}^k$ je

$$\mathsf{P}[\boldsymbol{X} = (x_1,x_2,\ldots,x_k)^\top] = \begin{cases} \frac{n!}{x_1!x_2!\ldots x_k!}\, p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}, & \sum_{k=1}^K x_k = n, x_k \in \mathbb{N}_0, \\ 0, & \textit{jinak.} \end{cases}$$

Poznámka. *Multinomické rozdělení popisuje následující situaci:*
Mějme K přihrádek a n nezávislých pozorování, kdy v každém z těchto pozorování vybereme jednu přihrádku podle pravděpodobností určených vektorem $\boldsymbol{p}$.

4.2 Kontingenční tabulky

Mějme dvě kategoriální veličiny – X nabývající hodnot $\{1,...,J\}$ a Z nabývající hodnot $\{1,...,K\}$. Zvolme $N \in \mathbb{N}$ a nechť $\binom{X_1}{Z_1}, \binom{X_2}{Z_2}, \ldots, \binom{X_N}{Z_N}$ je náhodný

výběr z rozdělení $\binom{X}{Z}$. Označíme:

$$
\begin{aligned}
n_{jk} &= \sum_{i=1}^{N} \mathbb{I}[X_i = j, Z_i = k], j \in \{1,\ldots,J\}, k \in \{1,\ldots,K\},\\
\boldsymbol{n} &= (n_{11},\ldots,n_{JK})^\top,\\
p_{jk} &= \mathsf{P}[X = j, Z = k], j \in \{1,\ldots,J\}, k \in \{1,\ldots,K\},\\
\boldsymbol{p} &= (p_{11},\ldots,p_{JK})^\top.
\end{aligned}
$$

Náhodné veličiny n_{jk} nazýváme pozorovanými četnostmi pro kombinace kategorií j a k. Náhodný vektor $\boldsymbol{n}$ má multinomické rozdělení $Mult_{JK}(N,\boldsymbol{p})$, protože vznikl klasifikací N pozorování do JK skupin.

Potřebujeme ještě označit:

$$
\begin{aligned}
n_{j+} &= \sum_{k=1}^{K} n_{jk}, & n_{+k} &= \sum_{j=1}^{J} n_{jk},\\
p_{j+} &= \sum_{k=1}^{K} p_{jk}, & p_{+k} &= \sum_{j=1}^{J} p_{jk},
\end{aligned}
$$

a můžeme sestavit kontingenční tabulku $J \times K$:

	$Z=1$	$Z=2$	$\ldots$	$Z=K$	$\sum$
$X=1$	n_{11}	n_{12}	$\ldots$	n_{1K}	n_{1+}
$X=2$	n_{21}	n_{22}	$\ldots$	n_{2K}	n_{2+}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$X=J$	n_{J1}	n_{J2}	$\ldots$	n_{JK}	n_{J+}
$\sum$	n_{+1}	n_{+2}	$\ldots$	n_{+K}	N

Sdružené rozdělení $\binom{X}{Z}$ pak popisuje tabulka pravděpodobností:

	$Z=1$	$Z=2$	$\ldots$	$Z=K$	$\sum$
$X=1$	p_{11}	p_{12}	$\ldots$	p_{1K}	p_{1+}
$X=2$	p_{21}	p_{22}	$\ldots$	p_{2K}	p_{2+}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$X=J$	p_{J1}	p_{J2}	$\ldots$	p_{JK}	p_{J+}
$\sum$	p_{+1}	p_{+2}	$\ldots$	p_{+K}	N

4.2.1 Kontingenční tabulky 2×2

Kontingenční tabulky 2×2 jsou speciálním případem, kdy $J = 2$ a $K = 2$.

	$Z=1$	$Z=2$	$\sum$
$X=1$	n_{11}	n_{12}	n_{1+}
$X=2$	n_{21}	n_{22}	n_{2+}
$\sum$	n_{+1}	n_{+2}	N

Na kontingenční tabulky můžeme také nahlížet po sloupcích. Pokud si označíme:

$$p_{1(1)} = \frac{p_{11}}{p_{+1}} = \mathsf{P}[X = 1|Z = 1], \qquad p_{1(2)} = \frac{p_{12}}{p_{+2}} = \mathsf{P}[X = 1|Z = 2]$$
$$p_{2(1)} = \frac{p_{21}}{p_{+1}} = \mathsf{P}[X = 2|Z = 1], \qquad p_{2(2)} = \frac{p_{22}}{p_{+2}} = \mathsf{P}[X = 2|Z = 2]$$
$$\boldsymbol{p_{(1)}} = (p_{1(1)},p_{2(1)})^\top, \qquad \boldsymbol{p_{(2)}} = (p_{1(2)},p_{2(2)})^\top,$$

pak náhodný vektor $\boldsymbol{n_{(1)}} = (n_{11},n_{21})^\top$ (první sloupec) má multinomické rozdělení $Mult_2(n_{+1},\boldsymbol{p_{(1)}})$ a náhodný vektor $\boldsymbol{n_{(2)}} = (n_{12},n_{22})^\top$ (druhý sloupec) má multinomické rozdělení $Mult_2(n_{+2},\boldsymbol{p_{(2)}})$.

4.3 χ^2 test nezávislosti v kontingenčních tabulkách

V této kapitole popíšeme obecně χ^2 test nezávislosti v kontingenčních tabulkách, v následující kapitole se pak budeme zabývat analýzou případu 2×2, nejrozšířenějších kontingenčních tabulek.

Náhodné veličiny X a Z jsou nezávislé právě tehdy, když platí:

$$p_{jk} = p_{j+}p_{+k}, \forall j \in \{1,\dots,J\}, \forall k \in \{1,\dots,K\}.$$

Formulujeme tedy χ^2 test nezávislosti v kontingenčních tabulkách, neboli test nezávislosti dvou kategoriálních veličin X a Z.

Model: $\mathcal{F} = \{\boldsymbol{n} = (n_{11},\dots,n_{JK})^\top \sim Mult_{JK}(N, \boldsymbol{p} = (p_{11},\dots,p_{JK})^\top)$ rozdělení popisující kontingenční tabulku $J \times K$ náhodných veličin X, Z pro N pozorování$\}$

Testovaný parametr: nezávislost X a Z

Hypotéza a alternativa:

$$H_0 : p_{jk} = p_{j+}p_{+k}, \forall j \in \{1,\dots,J\}, \forall k \in \{1,\dots,K\}.$$

Testová statistika:

$$T_{N,JK} = \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(n_{jk} - \frac{n_{j+}n_{+k}}{N}\right)^2}{\frac{n_{j+}n_{+k}}{N}} \tag{4.1}$$

Kritický obor[1]:

$$H_0 \text{ zamítneme} \iff T_{N,JK} \geq \chi^2_{(J-1)(K-1)}(1-\alpha)$$

P-hodnota (asymptotická):

$$p(x) = 1 - G_{(J-1)(K-1)}(x),$$

kde x je napozorovaná hodnota $T_{N,JK}$ a $G_{(J-1)(K-1)}$ je distribuční funkce $\chi^2_{(J-1)(K-1)}$.

[1]Odvození kritického oboru najdeme např. v (Anděl, 2011, str. 269-274).

4.4 Yatesova oprava na spojitost

V článku, který vydal Frank Yates v roce 1934 navrhl vylepšení testové statistiky $T_{N,JK}$:

$$T^C_{N,JK} = \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(\left|n_{jk} - \frac{n_{j+}n_{+k}}{N}\right| - 0.5\right)^2}{\frac{n_{j+}n_{+k}}{N}}.$$

I pro takto upravenou testovou statistiku Yates používá stejný kritický obor, tedy:

$$H_0 \text{ zamítneme} \iff T^C_{N,JK} \geq \chi^2_{(J-1)(K-1)}(1-\alpha).$$

4.5 Porovnání hladiny a síly testů s Yatesovou opravou na spojitost a bez ní

Budeme uvažovat variantu pro $J = 2, K = 2$, tedy pro kontingenční tabulky 2×2. Testová statistika bez opravy na spojitost tedy bude:

$$T_N = \sum_{j=1}^{2}\sum_{k=1}^{2} \frac{\left(n_{jk} - \frac{n_{j+}n_{+k}}{N}\right)^2}{\frac{n_{j+}n_{+k}}{N}}. \tag{4.2}$$

Testová statistika s Yatesovou opravou na spojitost v tomto případě má tvar:

$$T^C_N = \sum_{j=1}^{2}\sum_{k=1}^{2} \frac{\left(\left|n_{jk} - \frac{n_{j+}n_{+k}}{N}\right| - 0.5\right)^2}{\frac{n_{j+}n_{+k}}{N}}. \tag{4.3}$$

Poznámka. *V statistickém výpočetním prostředí R*[1] *se pro Yatesovu opravu používá testová statistika:*

$$T^C_N = \sum_{j=1}^{2}\sum_{k=1}^{2} \frac{\left(max\left(\left|n_{jk} - \frac{n_{j+}n_{+k}}{N}\right| - 0.5; 0\right)\right)^2}{\frac{n_{j+}n_{+k}}{N}}.$$

Použití opravy na spojitost se u χ^2 testu dá volit parametrem correct = TRUE. Hodnota TRUE je nastavena jako výchozí.

My jsme však testovou statistiku s opravou na spojitost počítali pomocí tvaru (4.3)*, jak jej navrh Yates. Vzhledem k volbě hladiny $\alpha = 0.05$ nebude mít na výsledné zamítání v kontingenčních tabulkách 2×2 tato volba téměř žádný vliv.*

Ujasníme ještě, že označení T_N a T^C_N budeme používat také pro zmíněné testy s testovou statistikou T_N, resp. T^C_N, kritickým oborem a pravidlem: hypotézu nezávislosti zamítneme $\iff$ T_N, resp. $T^C_N \geq \chi^2_1(1-\alpha)$.

Nyní budeme chtít simulací odhadnout skutečnou hladinu testů T_N a T^C_N. Použijeme při tom sloupcový pohled na kontingenční tabulky.

Simulaci provedeme následovně[2]:

1. Zvolíme parametry:

   $$n = n_{+1}, \qquad m = n_{+2}, \qquad p_1 = p_{1(1)}, \qquad p_2 = p_{1(2)}.$$

   Jelikož za H_0 bude platit $p_1 = p_2$, stačí volit jen $p = p_1 = p_2$.

[1] R Core Team (2018)

[2] Příslušný skript bude k práci přiložen jako elektronická příloha.

2. Zvolíme B - počet náhodných výběrů

3. Pro $b = 1, \ldots, B$:

   - Generujeme náhodně z $Mult_2(n, (p, 1-p)^\top)$ a $Mult_2(m, (p, 1-p)^\top)$ a uspořádáme naměřené hodnoty do kontingenční tabulky (po sloupcích).
   - Spočteme testové statistiky T_N a T_N^C.

4. Získáme odhady skutečné hladiny:

$$\frac{1}{B}\sum_{b=1}^{B} \mathbb{I}[T_N \geq \chi_1^2(1-\alpha)],$$
$$\frac{1}{B}\sum_{b=1}^{B} \mathbb{I}[T_N^C \geq \chi_1^2(1-\alpha)].$$

Pro naše účely volíme:

$n = 10, 30, 100, 300, 1000,$
$m = 10, 30, 100, 300, 1000,$ díky symetrii bude stačit vždy jen $m \geq n$,
$p = 0.5, 0.25, 0.05, 0.95,$
$B = 10\,000,$
$\alpha = 0.05.$

Poznámka. *Pokud při generování vyšla očekávaná hodnota pro některou kategorii nulová, pak bychom měli v testové statistice dělit nulou. Tento případ jsme ošetřili nezamítnutím hypotézy H_0, protože nastane v situaci, kdy máme v kontingenční tabulce řádek nulový, což znamená, že oba parametry p_1 i p_2 budou blízko 0, nebo 1, což nesvědčí proti H_0.*

4.5.1 Výsledky simulace hladiny testů

Výsledky výše popsané simulace jsme zaznamenali do tabulky 4.1, ve které ve sloupcích měníme hodnoty parametru p a v řádcích parametry n a m. Hodnoty v tabulce jsou průměrné hladiny testů, tedy odhady skutečných hladin.

Vidíme, že test T_N^C je konzervativnější, než test T_N pro všechny zvolené hodnoty parametrů. Průměrná hladina je u testu T_N^C pod hranicí chtěné hladiny (která byla v tomto případě 0.05) ve všech případech. Naopak test T_N dle odhadu hladiny v některých případech hladinu překračuje. Vidíme, že pro $p = 0.5$ se to stává v přibližně polovině případů, pravděpodobně zde bude mít také vliv simulace, protože pro $B = 10\,000$ nemusí mít simulace odhadu hladiny takovou přesnost.

Pro parametr $p = 0.25$ jsme v naší simulaci dostali výsledek, že test bez opravy na spojitost dodržuje hladinu v téměř všech případech. Test T_N^C je tedy zbytečně konzervativní.

Všimneme si, že pro extrémní parametry $p = 0.05$ a $p = 0.95$ jsme dostali podobná data, což je v souladu se sloupcovým pohledem na kontingenční tabulku – přehození řádků totiž nezmění testovou statistiku.

Při volbě parametrů $n = 10$ a $m = 10$ jsme dostali průměrnou hladinu téměř nulovou, to se děje s přispěním poznámky za popisem simulací – a sice proto, že pro situace, kde bychom měli dělit nulou, nezamítáme hypotézu.

Zajímavé výsledky dostáváme pro parametry $n = 10$, $m = 100, 300, 1000$. V tomto případě test bez opravy na spojitost nedodržuje hladinu a pro dodržení hladiny je vhodné použít test T_N^C. Kdybychom měli diskutovat myšlenku ze začátku této kapitoly, tak opravdu vidíme, že použití testu s Yatesovou opravou na spojitost je vhodné pro takové kontingenční tabulky, kde očekávaná hodnota některé kategorie bude menší než 5, avšak pokud je parametr p někde blízko hodnoty 0.5, pak by i test bez opravy na spojitost měl dodržovat stanovenou hladinu.

Celkově tedy vidíme, že pokud použijeme test s Yatesovou opravou na spojitost, nemusíme mít strach, že by nedodržoval stanovenou hladinu. Důvody proč nepoužívat opravu na spojitost se pokusíme ukázat níže.

		p = 0.5		p = 0.25		p = 0.05		p = 0.95	
		bez	s	bez	s	bez	s	bez	s
n = 10	m = 10	0.046	0.014	0.034	0.010	0.001	0.000	0.000	0.000
	m = 30	0.056	0.020	0.051	0.020	0.022	0.003	0.026	0.004
	m = 100	0.049	0.022	0.050	0.016	0.053	0.018	0.062	0.021
	m = 300	0.052	0.022	0.038	0.017	0.056	0.016	0.054	0.015
	m = 1000	0.041	0.021	0.028	0.018	0.075	0.014	0.068	0.011
n = 30	m = 30	0.053	0.029	0.049	0.023	0.028	0.002	0.030	0.001
	m = 100	0.049	0.031	0.047	0.025	0.040	0.012	0.038	0.014
	m = 300	0.050	0.034	0.047	0.027	0.037	0.017	0.040	0.022
	m = 1000	0.050	0.032	0.048	0.031	0.044	0.018	0.044	0.021
n = 100	m = 100	0.057	0.042	0.045	0.031	0.049	0.022	0.042	0.018
	m = 300	0.053	0.041	0.050	0.036	0.046	0.025	0.045	0.026
	m = 1000	0.052	0.040	0.048	0.038	0.043	0.027	0.050	0.028
n = 300	m = 300	0.053	0.043	0.052	0.043	0.050	0.031	0.052	0.031
	m = 1000	0.047	0.039	0.049	0.040	0.053	0.036	0.048	0.033
n = 1000	m = 1000	0.052	0.047	0.048	0.043	0.049	0.039	0.050	0.038

Tabulka 4.1: Tabulka odhadů hladiny testů T_N a T_N^C

4.5.2 Síla testu s opravou na spojitost

Podobně jako v kapitole o intervalových odhadech nás nezajímala pouze spolehlivost, ale i délka, tak pro statistické testy nás analogicky zajímá síla testu, neboli pravděpodobnost zamítnutí neplatné hypotézy při dané konkrétní alternativě. Obecně totiž platí, že pokud má test nižší skutečnou hladinu, bude síla takového testu nižší.

V této kapitole zvolíme několik hodnot parametrů n, m a p_1 a vykreslíme silofunkci pro proměnnou p_2.

Budeme sledovat dva případy:

Parametry n a m jsou přibližně stejně velké. Jako příklad zvolíme hodnotu $n = m = 100$.

Parametry n a m se výrazně liší. Vybrali jsme pro názornost hodnoty $n = 30$ a $m = 300$.

Pro tři různé parametry p_1 vykreslíme jednotlivé silofunkce pro testy T_N a T_N^C a budeme se snažit rozhodnout, které parametry mají na rozdíl síly T_N a T_N^C vliv.

Poznámka. *V této kapitole při vykreslování silofunkce používáme odhad silofunkce pomocí simulací*[3]*. Bereme ji jako funkci proměnné p_2 s parametry n, m, p_1, α a je simulována podobně, jako v simulaci hladiny – tedy vypočítáním průměrného počtu zamítnutí pro dané p_2. Rozdílem je volba počtu iterací, tentokrát volíme $B = 1000$ (kvůli časové náročnosti).*

Na obrázcích 4.1, 4.2, 4.3 je vykreslený odhad silofunkce testu T_N a T_N^C. Červeně je pro větší přehlednost zakreslen také rozdíl mezi těmito funkcemi. Právě tento rozdíl síly testů bez opravy na spojitost a s ní nás zajímá.

Na těchto třech grafech si nejprve všimneme, že síla testu s opravou na spojitost je vždy nižší, než síla testu bez opravy. To je v souladu se závěrem simulace skutečné hladiny, kde jsme si všimli, že průměrná hladina je pro všechny měřené parametry nižší pro T_N^C.

Dále vidíme, že rozdíl silofunkcí je vždy vyšší na pravém grafu, tedy v situaci, kdy jsou rozdílné hodnoty n a m. Z toho vyvodíme, že síla testu klesá pro všechny parametry p_1, p_2, pokud se n a m výrazně liší.

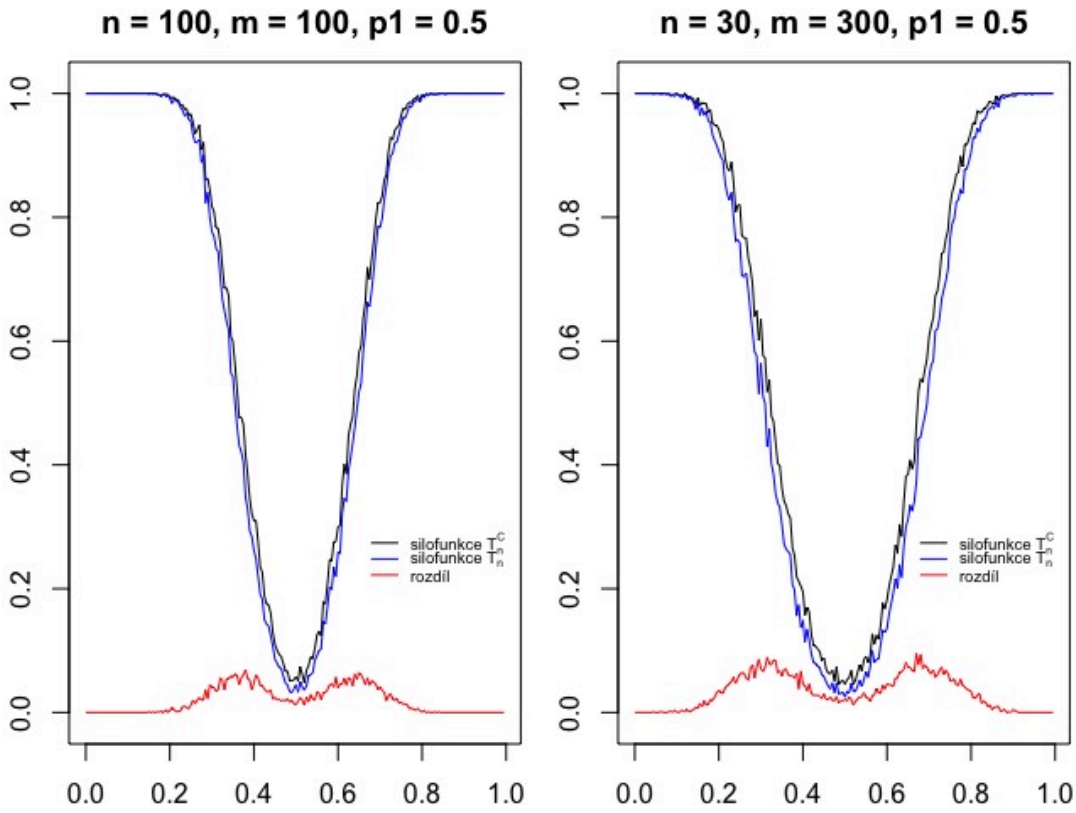
Také si můžeme všimnout, že rozdíl sil v obrázcích postupně roste (ve smyslu maxima), pokud se parametr p_1 vzdaluje od jedné poloviny. Avšak nejedná se o tak výrazný rozdíl, jako při porovnání levého a pravého grafu.

Z těchto grafů tedy můžeme shrnout:

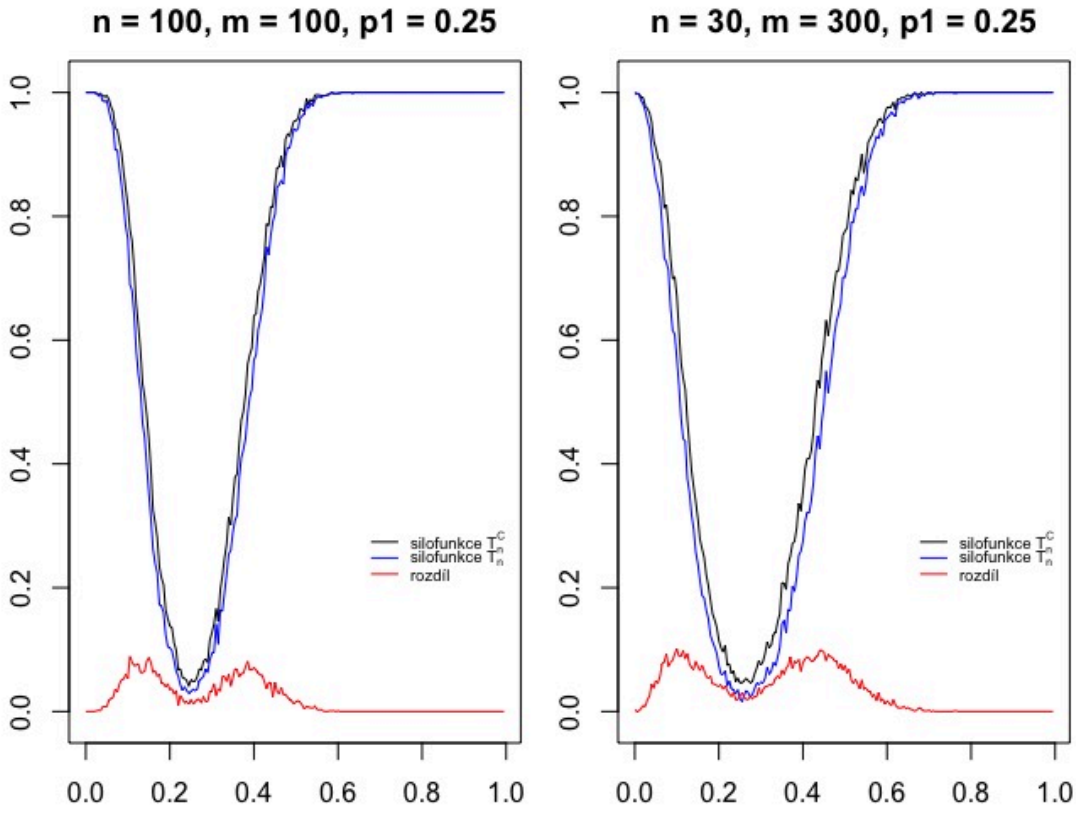
- Síla testu T_N^C je vždy nižší než síla T_N pro všechny alternativy, což je v souladu s tvary testových statistik.
- Rozdíl síly testů je nižší (ve smyslu maxima) pro hodnoty parametru p_1 okolo $\frac{1}{2}$.
- Na rozdíl síly testů má velký vliv poměr velikostí n a m. Tento rozdíl bude velký, pokud budou n a m hodně rozdílné.

Můžeme si v souvislosti se závěrem z kapitoly 4.5.1 všimnout, že pokud použijeme test s opravou na spojitost v situaci, kterou jsme doporučovali (tedy pro p blízká 0 nebo 1 a rozdílná n a m), pak výrazně ztratí test na síle. Tedy za dodržení předepsané hladiny platíme nižší sílou testu.

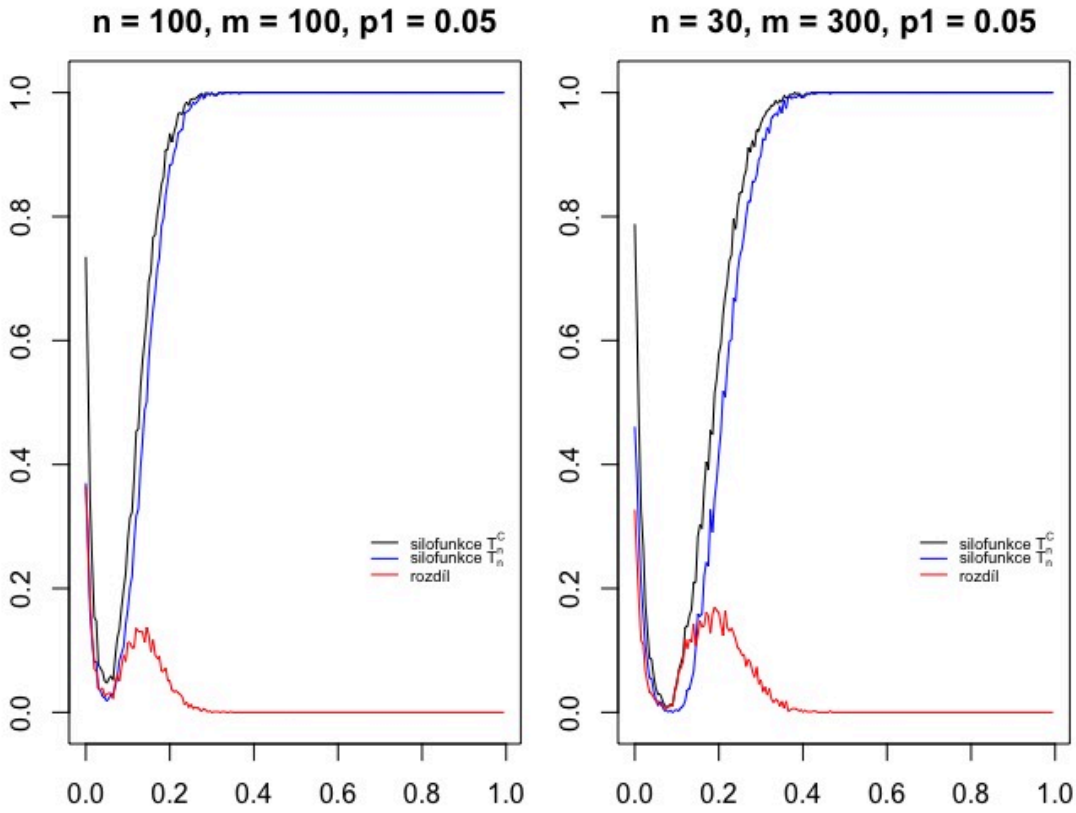
[3]Příslušný skript bude k práci přiložen jako elektronická příloha.



Obrázek 4.1: Odhadnutá silofunkce pro testy T_N a T_N^C ($p_1 = 0.5$)



Obrázek 4.2: Odhadnutá silofunkce pro testy T_N a T_N^C ($p_1 = 0.25$)



Obrázek 4.3: Odhadnutá silofunkce pro testy T_N a T_N^C ($p_1 = 0.05$)

Závěr

V práci jsme v první části vysvětlili, jak můžeme aproximovat diskrétní rozdělení, které je součtem diskrétních rozdělení s hodnotami v $\mathbb{N}_0$. Ukázali jsme také, že díky tomu, že je distribuční funkce takové náhodné veličiny po částech konstantní, není taková aproximace jednoznačná.

Ve druhé kapitole jsme ukázali pro binomické rozdělení, jak lze vhodnou hodnotu opravy na spojitost odvodit. Nejprve jsme nahlédli z grafu, že by hodnota 0.5 mohla dávat smysl, poté jsme nastínili analytické odvození popsané v (Feller, 1968). V další části této kapitoly jsme provedli numerické ověření, že použít opravu na spojitost při aproximaci hodnoty distribuční funkce náhodné veličiny s binomickým rozdělením je vhodné pro p, která nejsou příliš blízká 0. Taková aproximace dává menší chybu, v některých případech dokonce až o 5 řádů. Také jsme okomentovali aproximaci ostatních diskrétních rozdělení s použitím opravy na spojitost.

Ve třetí části jsme popsali tvary dolního, horního a oboustranného intervalového odhadu pro parametr p s využitím opravy na spojitost a porovnali je se standardními tvary těchto intervalů spolehlivosti. Poté jsme provedli simulaci, ve které jsme chtěli odhadnout spolehlivost intervalových odhadů s opravou na spojitost a bez ní. Z těchto výstupů jsme usoudili, že za cenu delšího intervalu dosahujeme častěji zvolenou spolehlivost při použití intervalu spolehlivosti s opravou na spojitost.

V poslední části jsme se zabývali studiem χ^2 testu nezávislosti v kontingenčních tabulkách 2×2. Popsali jsme kontingenční tabulky včetně sloupcového pohledu a také χ^2 test. Poté jsme prezentovali opravu na spojitost pro tento test navrženou Frankem Yatesem v roce 1934. Na simulacích jsme ověřili vlastnosti této opravy na spojitost – výsledkem našich simulací bylo, že test s opravou na spojitost je konzervativnější než bez ní a je vhodné jej použít pouze pokud by test bez opravy na spojitost nedodržoval hladinu. Z našich simulací vyplynulo, že se tak stane, pokud bude parametr p (nebo jeho odhad) blízký 0, nebo 1 a zároveň parametry n a m se budou lišit, neboli v takové kontingenční tabulce, která má v jednotlivých buňkách velmi rozdílně hodnoty.

Seznam použité literatury

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

Feller, W. (1968). *An introduction to probability theory and its applications*. Wiley, New York. ISBN 978-0-471-25708-0.

Nagel, J. R. (2012). *Introduction to Numerical Integration*. University of Utah, Salt Lake City, Utah. URL `http://www.ece.utah.edu/~ece6340/LECTURES/Jan30/`. PDF soubor Numerical Intergation, poslední přístup 8. 5. 2019.

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/`.

Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*, **1**(2), 217–235. ISSN 14666162. URL `http://www.jstor.org/stable/2983604`.

Seznam obrázků

Seznam tabulek