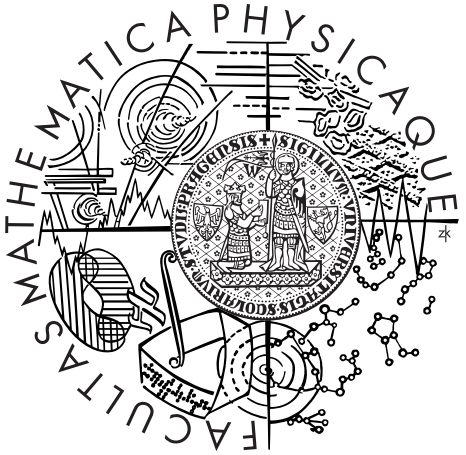
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Tibor Mach

Iterace úspěchů v posloupnosti Bernoulliových pokusů

Katedra pravděpodobnosti a matematické statistiky (32-KPMS)

Vedoucí bakalářské práce:	prof. RNDr. Jiří Anděl, DrSc.
Studijní program:	Matematika
Studijní obor:	obecná matematika (MOM)

Praha 2011

Děkuji prof. RNDr. Jiřímu Andělovi, DrSc. za zapůjčení některých materiálů, cenné rady pro sestavení obsahové i formální stránky bakalářské práce a pomoc při korektuře obsahových i formálních chyb během psaní tohoto textu.

Prohlašuji, že jsem tuto bakalářskou práci vypracoval 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 v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

V Plzni dne 25. května 2011 Podpis autora

Název práce: Iterace úspěchů v posloupnosti Bernoulliových pokusů

Autor: Tibor Mach

Katedra: Katedra pravděpodobnosti a matematické statistiky (32-KPMS)

Vedoucí bakalářské práce: prof. RNDr. Jiří Anděl, DrSc.

Abstrakt: Tato práce se zabývá vybranými pravděpodobnostními charakteristikami iterací v posloupnosti Bernoulliových pokusů a některými na nich založenými testy nezávislosti náhodných veličin. Na základě Markovových řetězců je zde odvozen explicitní vzorec pro výpočet pravděpodobnosti, že k prvnímu výskytu iterace úspěchů délky k v posloupnosti nezávislých Bernoulliových pokusů dojde v n-tém pokusu, a dále jsou zmíněny další vzorce pro tuto pravděpodobnost. Práce se dále zaměřuje na aproximace přesných hodnot této pravděpodobnosti (především Fellerovu aproximaci), meze těchto aproximací a jejich numerické srovnání. Nakonec je odvozen test nezávislosti založený délce nejdelší iterace v posloupnosti n Bernoulliových pokusů a test na základě celkového počtu iterací.

Klíčová slova: iterace, geometrické rozdělení řádu k, Bernoulliovy pokusy

Title: Success runs in a sequence of Bernoulli trials

Author: Tibor Mach

Department: Department of Probability and Mathematical Statistics (32-KPMS)

Supervisor: prof. RNDr. Jiří Anděl, DrSc.

Abstract: This work is focused on selected probability characteristics of runs in a sequence of Bernoulli trials and on some randomness tests based on these runs. Based on Markov chains, an explicit formula is derived for the probability that the first success run of a lenght k in a sequence of independent Bernoulli trials occurs in the n-th trial and other formulas for this probability are mentioned. Furthermore, approximations of the exact value of this probability (particularly the Feller approximation), bounds of these approximations, and their numeric relations are examined. Lastly, a test of randomness based on the lenght of the longest run in a sequence of n Bernoulli trials and a test based on the total amount of runs are derived.

Keywords: run, geometric distribution of order k, Bernoulli trials

Obsah

1. Úvod

Pojem iterace úspěchů v posloupnosti Bernoulliových pokusů je natolik jednoduchý, že ho lze snadno ilustrovat i naprostému laikovi pomocí příměru s opakovaným hodem mincí. Výpočet pravděpodobností jevů s těmito iteracemi spojených je přitom nejen výrazně složitější, ale často nám dává překvapivé a až téměř kontraintuitivní výsledky. Podívejme se na dva záznamy sto dvaceti hodů mincí. Číslo 0 značí, že na minci padl orel[1], hodnota 1 znamená, že padla panna. První záznam vypadá takto

1	1	0	0	1	0	0	1	0	1	1	0	0	1	0	0	0	1	1	0
1	0	1	0	0	1	1	0	1	0	0	1	0	1	0	1	1	0	1	1
0	0	1	1	0	1	1	1	0	1	0	0	1	0	0	1	1	0	1	0
0	1	1	0	1	0	0	1	1	0	1	0	1	1	0	0	1	1	1	0
0	1	0	1	0	1	0	0	0	1	0	1	0	1	0	1	0	1	0	1
1	0	0	1	0	0	1	0	1	1	0	0	1	0	0	1	1	0	1	1

a druhý takto

1	1	1	0	0	0	1	1	1	0	1	0	1	1	1	1	1	1	0	1
0	0	0	1	1	0	0	1	1	0	1	0	1	0	0	0	1	1	0	1
0	0	1	1	1	0	1	0	0	0	0	1	0	1	1	1	0	1	1	0
0	1	1	1	0	1	1	0	0	1	1	1	1	1	1	0	1	1	0	1
0	1	1	1	0	0	0	0	0	0	0	0	1	1	0	1	1	1	0	1
1	1	1	0	1	1	1	1	0	1	0	1	1	0	1	1	0	1	0	1.

Než budete číst dál, zkuste uhodnout, který z těchto dvou záznamů vznikl skutečným házením mince a který je vymyšlený. Jak uvádí Révész [1], podobný pokus byl poprvé proveden T. Vargou. Učitel na základní škole rozdělí žáky do dvou skupin. Každý žák v první skupině dostane minci a má za úkol stodvacetkrát touto mincí hodit a pokaždé zapsat výsledek hodu na papír. Žáci v druhé skupině naproti tomu žádnou minci nedostanou. Jejich úkolem je vymyslet „náhodnou“ posloupnost výsledků tak, aby co nejvěrněji odpovídala výsledkům skutečných hodů mincí. Poté, co jsou žáci hotovi, zamíchají mezi sebou listy zaznamenávající průběhy skutečných hodů s těmi vymyšlenými a odevzdají je takto zamíchané učiteli. Je-li učitel obeznámen s problematikou iterací v Bernoulliových pokusech (a žádné ze školních dětí nikoliv), bude schopen s velkou přesností opět rozdělit listy s výsledky do původních skupin. Během stodvaceti hodů nestrannou mincí totiž s pravděpodobností téměř 0.987 dojde k výskytu „iterace pann“ délky alespoň 4, tedy s pravděpodobností 0.987 padne za sebou ve stodvaceti hodech alespoň čtyřikrát panna. Vzhledem k tomu, že pravděpodobnost, že padne panna a pravděpodobnost, že padne orel, jsou shodné, stejná úvaha platí i pro „iteraci orlů“. Děti, které „náhodné“ hody mincí vymýšlely se ale většinou budou bát zapsat víc než tři shodné výsledky za sebou. Učitel tedy prostě rozřadí listy do skupin podle toho, zda se v nich vyskytuje iterace délky alespoň 4. Intuice dětem

[1] „Orel“ je označení pro rubovou stranu mince. Zajímavé je, že na českých mincích (s výjimkou kovové padesátikoruny) je rubová strana ta, která označuje nominální hodnotu mince. Jako „panna“ se označuje lícová strana mince. Na českých mincích je to ta strana, kde je vyražen český lev.

říká, že pravděpodobnost, že padne čtyřikrát za sebou stejná hodnota je velmi malá. Jak je ovšem vidět z předchozího, při větším počtu hodů je výskyt takové iterace jev naopak velmi pravděpodobný. Vraťme se ke dvěma záznamům hodů mincí ze začátku tohoto odstavce. V prvním záznamu najdeme nejvýše 3 stejné výsledky za sebou. Přitom pravděpodobnost, že během sto dvaceti hodů nevznikne iterace délky alespoň 4, je 0.013. První záznam je tedy téměř určitě vymyšlený. Nyní připomeneme termín Bernoulliův pokus.

Definice 1. Náhodnou veličinu X nazveme Bernoulliovým pokusem právě tehdy, když má X alternativní rozdělení s parametrem p.

V uvedeném příkladu jsme měli posloupnost stodvaceti nezávislých Bernoulliových pokusů s pravděpodobností úspěchu 0.5. Obecně máme počet pokusů roven n a pravděpodobnost úspěchu rovnu p. V posloupnosti n takových Bernoulliových pokusů se můžeme zajímat o to, jaká je pravděpodobnost výskytu iterace délky k, jaká je pravděpodobnost, že nejdelší iterace úspěchů bude délky k, nebo například chceme znát pravděpodobnost výskytu l iterací délky k. Přitom, ačkoliv je pojem iterace veskrze intuitivní, k odvození explicitních vyjádření většiny těchto pravděpodobností se dospělo až během posledních dvaceti až třiceti let. Tyto teoretické poznatky nacházejí následně široké uplatnění v různých oblastech lidské činnosti, například ve statistice, při testování kvality, v teorii spolehlivosti, v biologii, psychologii a mnohých dalších oborech.

V této práci se zaměříme především na geometrické rozdělení řádu k a na jeho aplikaci při testování hypotézy náhodnosti. Ve druhé kapitole se budeme teoreticky zabývat geometrickým rozdělením řádu k. Odvodíme přesný vzorec tohoto rozdělení založený na Markovových řetězcích a uvedeme další explicitní vzorce získané jinými metodami. Protože je přesný výpočet tohoto rozdělení pro některé hodnoty velmi náročný, zaměříme se dále na některé jeho aproximace. Podobně jako Feller [2] získáme vytvořující funkci pravděpodobnosti a pomocí ní pak Fellerovu aproximaci geometrického rozdělení řádu k. Dále uvedeme ještě některé další používané aproximace a numericky je srovnáme. Na závěr kapitoly se zmíníme o podmíněných pravděpodobnostech v případě veličiny s geometrickým rozdělením řádu k. Třetí kapitola se věnuje testům nezávislosti založených na iteracích Bernoulliových pokusů a geometrickém rozdělení řádu k. Konkrétně půjde o test na základě délky nejdelší iterace a test založený na celkovém počtu iterací.

Pro naše účely budeme potřebovat dvě různé definice iterace úspěchů. Uvažujme posloupnosti n Bernoulliových pokusů.

Definice 2 (Iterace úspěchů prvního druhu)**.** Iterací úspěchů prvního druhu v posloupnosti n Bernoulliových pokusů rozumíme nepřerušenou posloupnost úspěchů ohraničenou z obou stran neúspěchem, případně začátkem respektive koncem celé posloupnosti pokusů.

Definice 3 (Iterace úspěchů druhého druhu)**.** Posloupnost n Bernoulliových pokusů obsahuje l iterací úspěchů délky k právě tehdy, když obsahuje právě l nepřerušených, vzájemně se nepřekrývajících bloků úspěchů, z nichž každý obsahuje právě k úspěchů.

Jinými slovy, v případě, že se vyskytne iterace úspěchů délky k v n-tém pokusu, další iteraci uvažujeme při definici 3 až od $(n+1)$-ního pokusu. Ilustrujme

si rozdíl mezi definicemi na příkladu. Mějme posloupnost Bernoulliových pokusů, kde hodnota 1 značí úspěch a hodnota 0 neúspěch. Nechť hodnoty prvních 12 pokusů jsou například

$$111111001111. \tag{1.1}$$

V závislosti na použité definici máme

$$\underbrace{111111}_{6}00\underbrace{1111}_{4} \qquad \text{nebo} \qquad \underbrace{111}_{3}\underbrace{111}_{3}00\underbrace{111}_{3}1,$$

kde vlevo uvažujeme definici iterace úspěchů prvního druhu a vpravo definici iterace úspěchů druhého druhu při $k = 3$. Definice 2 nám tedy v (1.1) dává nejdříve iteraci úspěchů délky 6, iteraci 2 neúspěchů a nakonec iteraci 4 úspěchů. Z definice 3 dostáváme 2 iterace úspěchů délky 3 oddělené dvěma neúspěchy od další iterace délky 3. V rámci definice 3 tedy uvažujeme pouze iterace délky k. V dalším textu budeme uvažovat definici iterace úspěchů prvního druhu (dále stručně jen iterace úspěchů), nebude-li výslovně uvedeno jinak. Uvedeme definice některých symbolů používaných v dalším textu.

Definice 4 (dolní celá část)**.** Pro reálné číslo x definujeme dolní celou část $\lfloor x \rfloor$ čísla x jako $\lfloor x \rfloor = \max\{n \in \mathbb{Z} | n \leq x\}$.

Poznámka 1. Buďte $\{a_n\}$ a $\{b_n\}$ reálné posloupnosti. Pak budeme psát $a_n \sim b_n$ právě tehdy, když platí, že buď $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$, nebo existuje $n_0 \in \mathbb{N}$ takové, že pro všechna $n \geq n_0, n \in \mathbb{N}$ je $a_n = b_n = 0$.

2. Geometrické rozdělení řádu k

Definice 5. Mějme posloupnost nezávislých stejně rozdělených náhodných veličin $X_1, X_2, \ldots$ nabývajících hodnoty 1 s pravděpodobností p a hodnoty 0 s pravděpodobností $q = 1 - p$. Jde tedy o alternativní rozdělení s parametrem p. Jako T_k označme veličinu udávající čas prvního výskytu iterace úspěchů délky k, tedy

$$T_k := \min\{n \in \mathbb{N}; X_{n-k+1} = 1, X_{n-k+2} = 1, \ldots, X_n = 1\}.$$

Rozdělení takto definované náhodné veličiny nazveme **geometrické rozdělení řádu k s parametrem p.**

Toto rozdělení lze interpretovat jako čekání na první výskyt k úspěchů za sebou, tedy na čas prvního výskytu iterace úspěchů délky k. Jde vlastně o zobecnění geometrického rozdělení, jelikož pro $k = 1$ právě geometrické rozdělení dostáváme.

2.1 Distribuční funkce

Nejprve uvedeme, jakou definici distribuční funkce budeme dále uvažovat.

Definice 6. Nechť X je reálná náhodná veličina. Pak její distribuční funkci F definujeme předpisem $F(x) := \mathsf{P}(X \leq x)$.

Z definice geometrického rozdělení řádu k plyne zřejmě (pro $n \in \mathbb{N}$)

$$\mathsf{P}(T_k = n) = \begin{cases} 0 & \text{pro } 0 \leq n < k, \\ p^k & \text{pro } n = k, \\ qp^k & \text{pro } k < n \leq 2k. \end{cases} \tag{2.1}$$

Skutečně, pravděpodobnost, že v n pokusech dojde k iteraci délky větší než n, je samozřejmě nulová. Pravděpodobnost výskytu iterace délky právě n v n pokusech je rovna p^n, neboť pokusy jsou nezávislé a stejně rozdělené s pravděpodobností úspěchu p. Pokud je $n > k$ a k prvnímu výskytu iterace délky k dojde v n-tém pokusu, musí tento a jemu předcházejících $k-1$ pokusů skončit úspěchem, přičemž $(n-k)$-tý pokus naopak neúspěchem. Pokud by totiž i $(n-k)$-tý pokus byl úspěšný, došlo by k prvnímu výskytu iterace už v $(n-1)$-ním pokusu. Je-li navíc $n \leq 2k$, je $m < k$, kde $m := n - k - 1$, čili v prvních m pokusech nemůže dojít k výskytu iterace délky k bez ohledu na jejich výsledky. V případě $k < n \leq 2k$ tedy dostáváme pravděpodobnost prvního výskytu iterace délky k v n pokusech rovnou qp^k. Pro $n > 2k$ dostáváme snadno rekurentní formuli

$$\mathsf{P}(T_k = n) = \sum_{i=1}^{k} qp^{i-1}\mathsf{P}(T_k = n - i).$$

Dojde-li k prvnímu výskytu iterace úspěchů délky k v n-tém pokusu, musí nutně mezi prvními k pokusy být alespoň jeden, který skončí neúspěchem. V opačném

případě by došlo k výskytu iterace už v k-tém pokusu. Podle věty o úplné pravděpodobnosti platí rovnost

$$\begin{aligned}\mathsf{P}(T_k = n) &= \sum_{i=1}^{k} \mathsf{P}(T_k = n | \text{první neúspěch nastal v } i\text{-tém pokusu}) \\ &\quad \times \mathsf{P}(\text{první neúspěch nastal v } i\text{-tém pokusu}). \qquad (2.2)\end{aligned}$$

Přitom i-tý sčítanec na pravé straně rovnosti (2.2) představuje pravděpodobnost, že nejprve dojde k $i-1$ úspěchům, poté v i-tém pokusu k neúspěchu a následně po zbývajících $n-i$ pokusech k výskytu iterace úspěchů délky k. Vzhledem k nezávislosti pokusů je tato pravděpodobnost rovna právě $qp^{i-1} \cdot \mathsf{P}(T_k = n-i)$.

Přímé vyjádření rozdělení pravděpodobnosti (a tedy i distribuční funkce) veličiny T_k je o něco složitější a různé přístupy vedou k různým (ekvivalentním) vyjádřením. Podrobněji se budeme zabývat přístupem založený na Markovových řetězcích, u ostatních zmíníme pouze výsledky.

Tvrzení 1. Buď T_k náhodná veličina s geometrickým rozdělením řádu k s parametrem p a s distribuční funkcí F. Dále buď $\{Y_t, t \in \mathbb{N}\}$ Markovův řetězec s množinou stavů $S = \{0, 1, \ldots, k\}, k \in \mathbb{N}$, kde $Y_0 = 0$ a

$$Y_{t+1} = \begin{cases} Y_t + 1 & \text{pro } X_{t+1} = 1, \quad t < k, \\ 0 & \text{pro } X_{t+1} = 0, \quad t < k, \\ k & \text{pro } t \geq k. \end{cases}$$

Pak pro $n \in \mathbb{N}$ platí:

1. $\mathsf{P}(T_k = n) = p \cdot p_{0,k-1}^{(n-1)}$
2. $F(n) = p_{0,k}^{(n)}$,

kde $p_{i,j}^{(n)}$ je pravděpodobnost přechodu řetězce $\{Y_t, t \in \mathbb{N}\}$ ze stavu i do stavu j po n krocích.

Důkaz. Pravděpodobnosti přechodu řetězce Y_t jsou zřejmě

$$p_{i,j} = \begin{cases} p & \text{pro } j = i+1, \quad i < k, \\ q & \text{pro } j = 0, \qquad i < k, \\ 1 & \text{pro } i = j = k, \\ 0 & \text{jinak.} \end{cases}$$

Veličina T_k tedy udává čas prvního vstupu řetězce Y_t do stavu k za počáteční podmínky $Y_0 = 0$ (to je v souladu s definicí T_k, neboť stavy $0, \ldots, k-1$ jsou přechodné, stav k je trvalý absorbční a řetězec do něj vstoupí v konečném čase). Vstoupí-li řetězec Y_t do stavu k poprvé v čase n, znamená to, že v čase n došlo k prvnímu výskytu iterace úspěchů délky k. Můžeme tedy psát

$$\begin{aligned}\mathsf{P}(T_k = n) &= \mathsf{P}(Y_0 \neq k, \ldots, Y_{n-1} \neq k, Y_n = k | Y_0 = 0) \\ &= \sum_{j=0}^{k-1} \mathsf{P}(Y_0 \neq k, \ldots, Y_{n-2} \neq k, Y_{n-1} = j, Y_n = k | Y_0 = 0).\end{aligned}$$

Platí ovšem

$$\mathsf{P}(Y_0 \neq k, \ldots, Y_{n-2} \neq k, Y_{n-1} = j, Y_n = k | Y_0 = 0) = 0,\ j \neq k-1,$$

neboť pravděpodobnost přechodu řetězce ze stavu i do stavu j je pro $j \neq i+1$ rovna nule. Máme tedy

$$\begin{aligned}\mathsf{P}(T_k = n) &= \mathsf{P}(Y_0 \neq k, \ldots, Y_{n-2} \neq k, Y_{n-1} = k-1, Y_n = k | Y_0 = 0) \\ &= \mathsf{P}(Y_n = k | Y_{n-1} = k-1, Y_{n-2} \neq k, \ldots, Y_1 \neq k, Y_0 = 0) \\ &\quad \times \mathsf{P}(Y_0 \neq k, \ldots, Y_{n-2} \neq k, Y_{n-1} = k-1 | Y_0 = 0).\end{aligned}$$

S využitím markovské vlastnosti dostáváme

$$\begin{aligned}\mathsf{P}(T_k = n) &= \mathsf{P}(Y_n = k | Y_{n-1} = k-1) \\ &\quad \times \mathsf{P}(Y_0 \neq k, \ldots, Y_{n-2} \neq k, Y_{n-1} = k-1 | Y_0 = 0) \\ &= \mathsf{P}(Y_n = k | Y_{n-1} = k-1) \times \mathsf{P}(Y_{n-1} = k-1 | Y_0 = 0),\end{aligned}$$

neboť $[Y_j = k], j = 0, \ldots, n-2$ a $[Y_{n-1} = k-1]$ jsou disjunktní jevy, a tedy pravděpodobnost jejich průniku je 0. Konečně

$$\mathsf{P}(T_k = n) = \mathsf{P}(Y_n = k | Y_{n-1} = k-1) \times \mathsf{P}(Y_{n-1} = k-1 | Y_0 = 0) = p \cdot p_{0,k-1}^{(n-1)}.$$

Vyjádření distribuční funkce z výše uvedeného již snadno plyne, uvědomíme-li si, že stav k je absorbční a řetězec v něm setrvává, pokud do něj jednou vstoupí. Vzhledem k tomu přejde řetězec ze stavu 0 do stavu k po n krocích právě tehdy, když do stavu k vstoupí v některém z kroků $n, n-1, \ldots, 1$. Potom

$$F(n) = \sum_{r=1}^{n} p \cdot p_{0,k-1}^{(r-1)} = \mathsf{P}(X_n = k | X_0 = 0) = p_{0,k}^{(n)}. \qquad \square$$

Poznámka 2. Symbol $\binom{n}{k_1,k_2,\ldots,k_m}$ značí tzv. multinomický koeficient a je definován jako

$$\binom{n}{k_1, k_2, \ldots, k_m} := \frac{n!}{k_1! k_2! \cdots k_m!}.$$

Dále uvedeme bez důkazu některá (nerekurentní) vyjádření rozdělení a distribuční funkce veličiny T_k založená na kombinatorických výpočtech. Všechna jsou převzata z knihy Balakrishnan, Koutras [3], kde se autoři odkazují na literaturu uvedenou u jednotlivých vzorců.

Na základě kombinatorických výpočtů odvozují Philippou a Muwafi [4] následující vyjádření rozdělení T_k:

$$\mathsf{P}(T_k = n) = \sum \binom{n_1 + n_2 + \cdots + n_k}{n_1, n_2, \ldots, n_k} p^n \left(\frac{q}{p}\right)^{n_1 + \cdots + n_k}, \quad n \geq k,$$

kde sčítáme přes všechna nezáporná celá čísla $n_1, n_2, \ldots, n_k$ splňující $\sum_{i=1}^{k} i n_i = n - k$. Z tohoto vzorce pak Philippou a Makri [5] získávají formuli pro distribuční funkci F veličiny T_k

$$F(n) = 1 - \frac{p^{n+1}}{q} \sum \binom{n_1 + n_2 + \cdots + n_k}{n_1, n_2, \ldots, n_k} \left(\frac{q}{p}\right)^{n_1 + \cdots + n_k}, \quad n \geq k,$$

kde opět sčítáme přes všechna nezáporná celá čísla $n_1, n_2, \ldots, n_k$, která splňují $\sum_{i=1}^{k} i n_i = n - k$.

Rozvojem vytvořující funkce pravděpodobnosti $A(s)$ (viz dále) do Taylorovy řady kolem nuly odvozují Uppuluri a Patil [6] jednodušší vyjádření, kde se místo multinomických vyskytují pouze binomické koeficienty:

$$\begin{aligned} \mathsf{P}(T_k = n) &= p^k \sum_{j=0}^{\infty} (-1)^j \binom{n-k-jk}{j} (qp^k)^j \\ &\quad - p^{k+1} \sum_{j=0}^{\infty} (-1)^j \binom{n-k-jk-1}{j} (qp^k)^j, \quad n \geq k. \end{aligned}$$

Konečně uveďme Muselliho [7] formuli

$$\mathsf{P}(T_k = n) = \sum_{j=1}^{\lfloor \frac{n+1}{k+1} \rfloor} (-1)^{j-1} p^{jk} q^{j-1} \left\{ \binom{n-jk-1}{j-2} + q \binom{n-jk-1}{j-1} \right\},$$

ve které se sčítá pouze konečně mnoho čísel.

Definujme nyní ještě veličinu L_n jako délku nejdelší iterace úspěchů v n pokusech. Později využijeme některých vztahů mezi rozdělením veličin T_k a L_n, které nyní odvodíme. Zřejmě platí

$$\mathsf{P}(T_k \leq n) = \mathsf{P}(L_n \geq k), \tag{2.3}$$

z čehož okamžitě dostáváme

$$\mathsf{P}(L_n \geq k) = F(n)$$

a

$$\mathsf{P}(L_n < k) = 1 - F(n). \tag{2.4}$$

Dále snadno odvodíme

$$\mathsf{P}(L_n = k) = \mathsf{P}(T_k \leq n) - \mathsf{P}(T_{k+1} \leq n), \quad 1 \leq k \leq n, \tag{2.5}$$

neboť vzhledem k (2.3) je

$$\mathsf{P}(L_n = k) = \mathsf{P}(L_n \geq k) - \mathsf{P}(L_n \geq k+1) = \mathsf{P}(T_k \leq n) - \mathsf{P}(T_{k+1} \leq n).$$

Pravděpodobnost $\mathsf{P}(T_k = n)$ můžeme pomocí L_n vyjádřit jako

$$\mathsf{P}(T_k = n) = qp^k \mathsf{P}(L_{n-k-1} < k), \quad n \geq k+2, \tag{2.6}$$

protože pravá strana rovnice (2.6) je rovna pravděpodobnosti, že v prvních $n-k-1$ pokusech nevznikne iterace úspěchů délky k, v $(n-k)$-tém pokusu nastane neúspěch a dalších k pokusů skončí úspěchem, což je právě pravděpodobnost výskytu první iterace úspěchů délky k v n-tém pokusu. Jednoduchou úpravou vztahu (2.6) dostáváme

$$\mathsf{P}(L_n < k) = \frac{\mathsf{P}(T_k = n+k+1)}{qp^k}, \quad n \geq 1.$$

Pomocí těchto identit lze již snadno vyjádřit rozdělení a distribuční funkci veličiny L_n.

2.2 Vytvořující funkce pravděpodobnosti

Vytvořující funkci pravděpodobnosti geometrického rozdělení řádu k lze odvodit mnoha způsoby. My zde budeme postupovat stejným způsobem jako Feller [2]. Pro toto odvození uvažujeme definici iterace úspěchů druhého druhu. Díky této definici je výskyt iterace rekurentní jev (více viz Feller [2]) a lze použít následující postup odvození vytvořující funkce.

Tvrzení 2. Buď T_k náhodná veličina s geometrickým rozdělením řádu k. Pak pro příslušnou vytvořující funkci pravděpodobnosti $A(s)$ platí

$$A(s) = \sum_{n=0}^{\infty} \mathsf{P}(T_k = n)s^n = \frac{(ps)^k(1-ps)}{1-s+qp^k s^{k+1}} = \frac{(ps)^k}{1-qs[1+ps+\ldots+(ps)^{k-1}]}.$$

Důkaz. Nechť u_n je pravděpodobnost výskytu iterace délky k v n-tém pokusu (dle definice iterace druhého druhu). Položme $u_0 := 1$. Uvažujme nyní pokusy $X_n, X_{n-1}, \ldots, X_{n-k+1}$. Pravděpodobnost, že všech těchto k pokusů skončí úspěchem, je zřejmě p^k. Pokud toto nastane, dojde nutně v některém z nich k výskytu iterace délky k. Vzhledem k definici výše dokonce platí, že se tato vyskytne právě v jednom z pokusů $X_n, X_{n-1}, \ldots, X_{n-k+1}$. Pravděpodobnost, že se iterace délky k vyskytne v $(n-r)$-tém pokusu $(r = 0, 1, \ldots, k-1)$ a dalších r pokusů skončí úspěchem, je $u_{n-r}p^r$. Díky tomu můžeme psát

$$u_n + u_{n-1}p + \cdots + u_{n-k+1}p^{k-1} = p^k, \quad n \geq k.$$

Když tuto rovnici vynásobíme výrazem s^n a sečteme přes $n = k, k+1, k+2, \ldots$, dostáváme

$$\sum_{n=k}^{\infty} u_n s^n + p\sum_{n=k-1}^{\infty} u_n s^{n+1} + \cdots + p^{k-1}\sum_{n=1}^{\infty} u_n s^{n+k-1} = p^k \sum_{n=k}^{\infty} s^n.$$

Vzhledem k tomu, že triviálně platí $u_1 = u_2 = \cdots = u_{k-1} = 0$, máme

$$\sum_{n=k}^{\infty} u_n s^n + ps\sum_{n=k}^{\infty} u_n s^n + \cdots + (ps)^{k-1}\sum_{n=k}^{\infty} u_n s^n = p^k \sum_{n=k}^{\infty} s^n, \tag{2.7}$$

a protože $u_0 = 1$, je vytvořující funkce $U(s)$ posloupnosti u_n

$$U(s) = \sum_{n=k}^{\infty} u_n s^n + u_0 s^0 = \sum_{n=k}^{\infty} u_n s^n + 1$$

čili

$$\sum_{n=k}^{\infty} u_n s^n = U(s) - 1. \tag{2.8}$$

Dosazením (2.8) do rovnice (2.7) a úpravou získáváme

$$[U(s)-1][1+ps+(ps)^2+\cdots+(ps)^{k-1}] = (ps)^k \sum_{n=0}^{\infty} s^n. \tag{2.9}$$

Mějme s takové, že $|s| < 1$. Pak na pravé straně rovnice (2.9) vidíme nekonečný součet geometrické řady a na levé straně její částečný součet, takže (2.9) můžeme upravit na

$$(U(s) - 1) \cdot \frac{1 - (ps)^k}{1 - ps} = \frac{(ps)^k}{1 - s}$$

neboli

$$U(s) = \frac{1 - s + qp^k s^{k+1}}{(1 - s)(1 - (ps)^k)}.$$

Nyní definujme ještě a_n jako pravděpodobnost prvního výskytu iterace délky k v n-tém pokusu, přičemž $a_0 := 0$. Jak je vidět, $P(T_k = n) = a_n$, čili vytvořující funkce $\sum_{n=0}^{\infty} a_n s^n$ posloupnosti a_n je právě vytvořující funkcí pravděpodobnosti geometrického rozdělení řádu k. Zbývá ukázat, že

$$U(s) = \frac{1}{1 - A(s)}, \tag{2.10}$$

z toho pak již okamžitě plyne tvrzení věty. Pravděpodobnost, že k výskytu iterace délky k dojde poprvé v m-tém a potom znovu v n-tém pokusu, je $a_m u_{n-m}$ (zde opět využíváme definice iterace druhého druhu) a pravděpodobnost prvního výskytu iterace délky k v m-tém pokusu je $f_m = f_m u_0$, neboť $u_0 = 1$. Podobně jako výše píšeme

$$u_n = a_1 u_{n-1} + a_2 u_{n-2} + \cdots + a_n u_0, \quad n \geq 1, \tag{2.11}$$

neboť tyto jednotlivé případy jsou vzájemně disjunktní, a u_n je pravděpodobnost, že nastane právě jeden z nich. Nyní si stačí všimnout konvoluce na pravé straně rovnosti (2.11) a sčítáním přes všechna přirozená čísla n dostáváme

$$U(s) - 1 = A(s)U(s),$$

což je ekvivalentní s (2.10). □

Poznámka 3. Tvrzení 2 jsme dokázali při definici iterace úspěchů druhého druhu. Tvrzení ovšem zůstává v platnosti i pro definici iterace úspěchů prvního druhu. První výskyt iterace délky k je v obou případech definován ekvivalentně. Pravděpodobnost jevu $[T_k = n]$ závisí pouze na tomto prvním výskytu, a proto zůstává vytvořující funkce pravděpodobnosti veličiny T_k stejná bez ohledu na uvažovanou definici iterace. Přitom definice iterace druhého druhu nám pomohla výrazně zjednodušit techniku důkazu.

V dalším textu již opět uvažujeme definici iterace prvního druhu.

2.3 Meze a aproximace

Ačkoliv známe explicitní vyjádření distribuční funkce a rozdělení veličiny T_k, pro velké hodnoty n a k je výpočet přesných hodnot $\mathsf{P}(T_k = n)$ výpočetně náročný a vzhledem k tomu je pro praxi užitečné zabývat se aproximacemi, jejichž výpočet je jednodušší. Definujme nejprve pomocnou funkci $G(n) := 1 - F(n)$. Zřejmě platí

$$G(n) = \mathsf{P}(T_k > n) = \mathsf{P}(L_n < k).$$

Z (2.6) okamžitě dostáváme

$$\mathsf{P}(T_k = n) = qp^k G(n-k-1), \quad n \geq k+2. \tag{2.12}$$

Vidíme tedy, že nám k aproximaci $\mathsf{P}(T_k = n)$ stačí hledat vhodný odhad hodnoty $G(n)$. Z (2.1) plyne, že se můžeme omezit na odhady $G(n)$ pro $n \geq k$. Pro

$$Z_i := \prod_{j=i}^{i+k-1} X_j,\ i = 1, 2, \ldots, n-k+1,$$

můžeme psát

$$F(n) = \mathsf{P}\left(\bigcup_{i=1}^{n-k+1} [Z_i = 1]\right) \leq \sum_{i=1}^{n-k+1} \mathsf{P}(Z_i = 1) = (n-k+1)p^k$$

neboli

$$G(n) \geq 1 - (n-k+1)p^k, \tag{2.13}$$

neboť $Z_i = 1$ tehdy a jen tehdy, když se v posloupnosti $\{X_n\}$ vyskytne iterace úspěchů délky k počínající veličinou X_i. Naopak máme

$$\begin{aligned} G(n) &= \mathsf{P}(Z_i = 0,\ i = 1, 2, \ldots, n-k+1) \\ &\leq \mathsf{P}(Z_1 = 0, Z_{k+1} = 0, Z_{2k+1} = 0, \ldots, Z_{rk+1} = 0), \end{aligned}$$

kde $r := \left\lfloor \frac{n-k}{k} \right\rfloor$. Náhodné veličiny $Z_1, Z_{k+1}, Z_{2k+1}, \ldots, Z_{rk+1}$ jsou nezávislé, a tedy platí

$$G(n) \leq \prod_{j=0}^{r} \mathsf{P}(Z_{jk+1} = 0) = (1-p^k)^{\lfloor n/k \rfloor}. \tag{2.14}$$

Dosazením (2.13) a (2.14) do (2.12) získáváme

$$qp^k[1-(n-2k)p^k] \leq \mathsf{P}(T_k = n) \leq qp^k(1-p^k)^{\lfloor \frac{n-k-1}{k} \rfloor},\ n \geq k+2.$$

Nyní odvodíme Fellerovu aproximaci pravděpodobnosti $\mathsf{P}(T_k = n)$. K tomu budeme potřebovat dvě pomocná tvrzení.

Lemma 1. Nechť $P(s) = \sum p_k s^k = \frac{U(s)}{V(s)}$ je vytvořující funkce, kde $U(s)$ a $V(s)$ jsou polynomy bez společných kořenů, a kde s_1 je jednoduchý kořen polynomu $V(s)$ takový, že je v absolutní hodnotě menší než všechny ostatní kořeny. Pak $p_n \sim a_1 s_1^{-(n+1)}$, kde $a_1 = \frac{-U(s_1)}{V'(s_1)}$.

Důkaz. Uvažujme nejprve situaci, kde stupeň U je menší než stupeň V a $m = \deg(V)$. Nejprve předpokládejme, že polynom V má právě m různých kořenů $s_1, s_2, \ldots, s_m$, tedy

$$V(s) = (s-s_1)(s-s_2)\cdots(s-s_n). \tag{2.15}$$

Potom můžeme pomocí rozkladu na parciální zlomky psát

$$P(s) = \frac{a_1}{s_1 - s} + \frac{a_2}{s_2 - s} + \cdots + \frac{a_m}{s_m - s}, \tag{2.16}$$

kde $a_1, a_2, \ldots, a_m$ jsou konstanty. Vynásobením rovnosti (2.16) výrazem $s_1 - s$ dostáváme

$$(s_1 - s)P(s) = a_1 + \frac{(s_1 - s)a_2}{s_2 - s} + \cdots + \frac{(s_1 - s)a_m}{s_m - s},$$

což se blíží k a_1 pro $s \to s_1$. Z $P(s) = \frac{U(s)}{V(s)}$ a (2.15) plyne

$$(s_1 - s)P(s) = \frac{-U(s)}{(s - s_2)(s - s_3)\cdots(s - s_m)}. \tag{2.17}$$

Pro $s \to s_1$ se čitatel na pravé straně rovnosti (2.17) blíží k $-U(s)$ a jmenovatel k $(s_1 - s_2)(s_1 - s_3)\cdots(s_1 - s_m)$, což je právě $V'(s_1)$. Zcela obdobně můžeme postupovat i pro $a_2, a_3, \ldots, a_m$ a vidíme, že

$$a_k = \frac{-U(s_k)}{V'(s_k)}.$$

Zřejmě

$$\frac{1}{s_k - s} = \frac{1}{s_k} \cdot \frac{1}{1 - s/s_k}.$$

Pro $|s| < |s_k|$ můžeme rozvinout $\frac{1}{1-s/s_k}$ do geometrické řady neboli

$$\frac{1}{1 - s/s_k} = 1 + \frac{s}{s_k} + \left(\frac{s}{s_k}\right)^2 + \left(\frac{s}{s_k}\right)^3 + \cdots \tag{2.18}$$

Vzhledem k (2.18) je

$$\left(\frac{a_k}{s_k - s}\right)^{(n)}_{s=0} = \frac{a_k n!}{s_k^n + 1}.$$

Z rovnosti $p_n = \frac{P^{(n)}(0)}{n!}$ již pak okamžitě plyne

$$p_n = \frac{a_1}{s_1^{n+1}} + \frac{a_2}{s_2^{n+1}} + \cdots + \frac{a_m}{s_m^{n+1}}. \tag{2.19}$$

Kořen s_1 je v absolutní hodnotě menší než všechny ostatní kořeny, tedy $|s_1| < |s_k|$, $k = 2, 3, \ldots, m$. Díky tomu je jmenovatel prvního zlomku v (2.19) nejmenší. Se zvětšujícím se n váha zlomku se jmenovatelem s_1^{n+1} v (2.19) roste oproti ostatním zlomkům. Jinými slovy $p_n \sim a_1 s_1^{-(n+1)}$. Během dokazování jsme zavedli dodatečné požadavky na stupeň polynomu U a násobnosti kořenů polynomu V. K dokončení důkazu zbývá ukázat, že tvrzení platí i bez těchto předpokladů. Buď nejprve $\deg(U) = m + r$, $r \geq 0$. Dělením polynomů dostáváme

$$P(s) = \frac{U_1(s)}{V(s)} + U_2(s),$$

kde U_1 je polynom stupně menšího než m a U_2 je polynom stupně r. Přitom vzhledem k tomu, že p_n lze vyjádřit jako $p_n = \frac{P^{(n)}(0)}{n!}$, má polynom U_2 vliv pouze na hodnoty $p_1, p_2, \ldots, p_{r+1}$ a na $\frac{U_1(s)}{V(s)}$ lze aplikovat předchozí postup. Nyní nechť existuje r-násobný kořen s_k polynomu $V(s)$. Vytvořující funkci $P(s)$ lze opět rozložit na parciální zlomky, přičemž vyjádření (2.16) bude obsahovat dalších $r - 1$ členů ve tvaru $b_l/(s - s_k)^l$, $l = 2, 3, \ldots, r$. V (2.19) se toto projeví opět dalšími $r - 1$ členy, tentokrát ve tvaru $b_l c_l s_k^{(n+l)}$, kde b_l a c_l, $l = 2, 3, \ldots, r$ jsou konstanty. Pokud je však s_1 jednoduchý kořen, nebudou mít tyto další členy vliv na asymptotické vyjádření p_n. □

Lemma 2. Mějme polynom $W(s) = 1 - qs[1 + ps + \ldots + (ps)^{k-1}]$, kde $0 < p < 1$, $q = 1 - p$. Pak existuje právě jeden kladný kořen x polynomu $W(s)$. Navíc $x > 1$.

Důkaz. Protože koeficienty p a q jsou kladné, je na kladné poloose $W(s)$ ostře klesající a jakožto polynom i spojitá funkce na celém $\mathbb{R}$. Přitom $W(0) = 1$ a zřejmě W není funkce zdola omezená. Z toho již jasně plyne, že existuje právě jedno $x > 0$, že $W(x) = 0$. Pro $y = 1$ je $W(y)$ rovno

$$W(y) = 1 - q(1 + p + p^2 + \cdots + p^k - 1) = 1 - q \cdot \frac{1 - p^k}{1 - p}.$$

Protože $0 < p < 1$, je $\frac{1-p^k}{1-p} < 1$, a tedy i $q \cdot \frac{1-p^k}{1-p} < 1$. Vzhledem k tomu 1 není kořenem polynomu $W(s)$ a z výše zmíněné monotonie plyne, že nutně $|x| > 1$. □

Nyní již máme dostatek prostředků k odvození Fellerovy aproximace geometrického rozdělení řádu k.

Tvrzení 3. Nechť T_k je náhodná veličina s geometrickým rozdělením řádu k. Pak

$$G(n) \sim \frac{1 - px}{(k + 1 - kx)q} \cdot \frac{1}{x^{n+1}}, \tag{2.20}$$

kde x je jediný kladný kořen polynomu $W(s) = 1 - qs[1 + ps + \cdots + (ps)^{k-1}]$.

Důkaz. Abychom mohli použít metodu rozkladu na parciální zlomky na vytvořující funkci pravděpodobnosti, musíme ověřit, že x je pouze jednonásobný kořen a že je nejmenší v absolutní hodnotě ze všech kořenů polynomu. Z algebry víme, že kořen polynomu násobnosti n je současně kořenem jeho $(n-1)$-ní derivace. Ovšem

$$W'(s) = -q\{[1 + ps + \cdots + (ps)^{k-1}] + [ps + 2p^2s^2 + \cdots + (k-1)p^{k-1}s^{k-1}]\},$$

což je ale zjevně záporné pro všechna $s > 0$. Z trojúhelníkové nerovnosti pak okamžitě plyne, že pro každé $s \in \mathbb{C}$, $|s| \leq x$ je

$$|qs[1 + ps + \cdots + (ps)^{k-1}]| \leq qx[1 + px + \cdots + (px)^{k-1}] = 1.$$

Přitom rovnost nastává právě když $s = x$. Jinak řečeno, x je v absolutní hodnotě nejmenší ze všech kořenů polynomu $W(s)$. Nyní můžeme použít lemma 1 na vytvořující funkci pravděpodobnosti geometrického rozdělení řádu k. Dle tvrzení 2 máme $s_1 = x$, $U(s) = (ps)^k(1 - ps)$ a $V(s) = 1 - s + qp^k s^{k+1}$. Derivace polynomu V je $V'(s) = (-1)[1 - (k+1)qp^k s^k]$. Tato volba polynomů U a V je korektní i přesto, že mají společný kořen p^{-1}. Platí totiž

$$V'(s) = [(1 - ps) \cdot W(s)]' = -p \cdot W(s) + (1 - ps) \cdot W'(s)$$

a

$$V'(x) = (1 - ps) \cdot W'(x),$$

neboť $V(s) = (1 - ps) \cdot W(s)$ a $W(x) = 0$. Půjde tedy o ekvivalentní vyjádření jako při volbě $U(s) = (ps)^k$ a $V(s) = W(s)$. Dostáváme tedy

$$P(T_k = n) = a_n \sim \frac{p^k x^k (1 - px)}{1 - (k+1)qp^k x^k} \cdot \frac{1}{x^{n+1}}.$$

To můžeme dále upravit a s využitím rovnosti $V(x) = 0$ postupně píšeme

$$\begin{aligned}
\frac{p^k x^k (1-px)}{1-(k+1)qp^k x^k} &= \frac{p^k x^k (1-px)}{1-(k+1)qp^k x^k} \cdot \frac{x-1}{x-1} \\
&= \frac{p^k x^k (1-px)(x-1)}{x - qp^k x^{k+1} - 1 - kqp^k x^{k+1} + (k+1)qp^k x^k} \\
&= \frac{p^k x^k (1-px)(x-1)}{-\underbrace{(1-x+qp^k x^{k+1})}_{V(x)=0} + p^k x^k (k+1-kx)q} \\
&= \frac{(1-px)(x-1)}{(k+1-kx)q}.
\end{aligned}$$

Odvodili jsme tedy, že

$$a_n \sim \frac{(1-px)(x-1)}{(k+1-kx)q} \cdot \frac{1}{x^{n+1}},$$

kde a_n je pravděpodobnost prvního výskytu iterace délky k v n-tém pokusu. Přitom $G(n) = P(T_k > n) = a_{n+1} + a_{n+2} + \cdots$, takže

$$G(n) \sim \sum_{l=n+1}^{\infty} \frac{(1-px)(x-1)}{(k+1-kx)q} \cdot \frac{1}{x^{l+1}},$$

a protože dle lemmatu 2 je $|x| > 1$, můžeme psát

$$\begin{aligned}
\sum_{l=n+1}^{\infty} \frac{(1-px)(x-1)}{(k+1-kx)q} \cdot \frac{1}{x^{l+1}} &= \frac{(1-px)(x-1)}{(k+1-kx)q} \sum_{l=n+1}^{\infty} \frac{1}{x^{l+1}} \\
&= \frac{(1-px)(x-1)}{(k+1-kx)q} \cdot \frac{1}{(x-1)x^{n+1}} \\
&= \frac{1-px}{(k+1-kx)q} \cdot \frac{1}{x^{n+1}}.
\end{aligned}$$

Tím je tvrzení věty dokázáno. □

Poznámka 4. Ve speciálním případě předchozí věty, kdy $x = p^{-1}$, je čitatel i jmenovatel v aproximaci roven nule. Toto je způsobeno právě volbou U a V z prvního vyjádření vytvořující funkce pravděpodobnosti z věty 2, kde mají čitatel a jmenovatel společný kořen p^{-1}. Ovšem zkrátíme-li zlomek v aproximaci výrazem $(1-px)$, dojdeme k výsledku odpovídajícímu Tvrzení 3.

Pro praxi je rovněž užitečné znát velikost maximální chyby, jaké se můžeme odhadem dopustit. Nyní odvodíme maximální absolutní chybu, které se dopustíme při nahrazení přesné hodnoty $G(n)$ Fellerovou aproximací.

Tvrzení 4. Buď T_k náhodná veličina s geometrickým rozdělením řádu k. Nechť $(k+1)q \neq 1$. Potom platí

$$\left| G(n) - \frac{1-px}{(k+1-kx)q} \cdot \frac{1}{x^{n+1}} \right| \leq \frac{2(k-1)p^{n+1}}{(k+1+kp^{-1})q}, \tag{2.21}$$

kde x je jediný kladný kořen polynomu $W(s) = 1 - qs[1 + ps + \cdots + (ps)^{k-1}]$.

Důkaz. Z prvního vyjádření vytvořující funkce pravděpodobnosti ve větě 2 je zřejmé, že každý kořen jmenovatele splňuje rovnost

$$s = 1 + qp^k s^{k+1}, \tag{2.22}$$

přičemž (2.22) obsahuje navíc ještě kořen $s = p^{-1}$. Uvažujme nyní funkci $f(s) := 1 + qp^k s^{k+1}$. Snadno se přesvědčíme, že funkce f je na kladné poloose konvexní a ostře rostoucí. Protože jediné dva reálné kladné kořeny polynomu $V(s) = 1 - s + qp^k s^{k+1}$ jsou x a p^{-1}, protne se na kladné poloose graf funkce f s grafem identické funkce $h(s) := s$ právě v těchto dvou bodech. Jejich vzájemnou polohu určíme snadno z derivace funkce f. Platí totiž $f'(p^{-1}) = (k+1)q$. Funkce f je na kladné poloose konvexní a monotónní a $f(0) = 1$. V bodě p^{-1} se protíná graf f s grafem h (případně se dotýkají). Přitom mohou nastat právě tři případy: $(k+1)q > 1$, $(k+1)q < 1$, nebo $(k+1)q = 1$. Vyšetříme každý zvlášť. Nechť tedy nejprve

$$(k+1)q > 1. \tag{2.23}$$

Protože pro všechna reálná s je $h'(s) = 1$, roste v tomto případě funkce f v bodě p^{-1} rychleji než identita. Jinými slovy graf funkce f protne graf h v bodě p^{-1} zdola. Z toho a z výše zmíněných vlastností funkce f už jasně plyne, že $x < p^{-1}$. Mezi těmito dvěma body leží graf funkce f pod grafem h, přesněji pro všechna reálná s taková, že $x < s < p^{-1}$, je $f(s) < s$. Z Cauchyho-Schwarzovy nerovnosti dostáváme pro všechna komplexní čísla s nerovnost $|f(s)| \leq f(|s|)$ a celkem tedy pro všechna komplexní s splňující $x < |s| < p^{-1}$ platí

$$|f(s)| \leq f(|s|) < |s|. \tag{2.24}$$

Z důkazu předchozího tvrzení víme, že x je v absolutní hodnotě nejmenším kořenem polynomu W. Nyní už z (2.24) okamžitě vidíme, že pro všechny kořeny s_k polynomu W takové, že $s_k \neq x$ platí

$$|s| > p^{-1}, \tag{2.25}$$

neboť p^{-1} není díky (2.23) kořenem polynomu W. Protože potřebujeme zjistit velikost, jak velký přírůstek k funkci $G(n)$ tvoří kořeny, jež jsme aproximací zanedbali, bude nás zajímat jejich násobnost. S využitím výše uvedeného ji již snadno odvodíme. Předně derivací obou stran rovnosti (2.22) získáváme vztah

$$1 = (k+1)qp^k s^k, \tag{2.26}$$

přitom každý kořen polynomu W násobnosti větší než 1 musí splňovat rovnost (2.26). Pro s komplexní je $s^k = |s|^k(\cos(k\phi) + i\sin(k\phi))$, kde $\phi = \arg(z)$. Pokud má pro s platit (2.26), musí být $k\phi = 2n\pi$ pro nějaké přirozené číslo n. Pak ovšem $s^k = |s|^k$. Z (2.23) a (2.25) je již snadno vidět, že neexistuje žádné komplexní s

takové, které by současně bylo kořenem polynomu W a splňovalo rovnost (2.26). Jinak řečeno, všechny kořeny polynomu W jsou jednonásobné. W má tedy kromě x ještě $k-1$ různých kořenů $s_1, s_2, \ldots, s_{r-1}$ a jejich příspěvek C_r k hodnotě $G(n)$ je stejného tvaru jako příspěvek kořene x čili

$$C_r = \frac{1 - ps_r}{k + 1 - ks_r} \cdot \frac{1}{qs_r^{n+1}}. \tag{2.27}$$

Uvažujme reálné číslo t splňující $t > p^{-1} > k^{-1}(k+1)$. Potom můžeme psát

$$\left| \frac{pte^{i\phi} - 1}{kte^{i\phi} - (k+1)} \right| \leq \frac{pt + 1}{kt + k + 1}, \tag{2.28}$$

kde $\phi \in [-\pi, \pi]$. Dosazením komplexního čísla s, jehož vzdálenost od počátku činí právě t, do levé strany předchozí rovnice, dostáváme stejný horní odhad i pro s. S využitím (2.28) a (2.25) získáme odhad pro C_r

$$|C_r| < \frac{2p^{n+1}}{(k + 1 + kp^{-1})q}.$$

V aproximaci jsme zanedbali právě $k-1$ takových kořenů, a proto maximální možná chyba, které jsme se tímto mohli dopustit, činí

$$\frac{2(k+1)p^{n+1}}{(k + 1 + kp^{-1})q}.$$

K dokončení důkazu musíme ještě vyšetřit zbylé dva případy. Buď nyní

$$(k+1)q < 1. \tag{2.29}$$

Postupujeme obdobně jako v případě (2.23), přičemž tentokrát se graf funkce f protne s grafem h shora neboli $p^{-1} < x$. Nerovnosti (2.24) platí v tomto případě pro všechna komplexní s taková, že $p^{-1} < |s| < x$. Opět x je v absolutní hodnotě nejmenší kořen polynomu W, takže pro každý kořen s polynomu W takový, že $s \neq x$ platí $|s| > x$. Z (2.29) plyne, že graf funkce f protne graf h v bodě p^{-1} shora.V bodě x pak nutně graf f protne graf identické funkce h zdola, takže $f'(x) > 1$. Analogicky jako v předchozím případě dostáváme, že polynom W má k jednoduchých kořenů $x, s_1, s_2, \ldots, s_{k-1}$. Nerovnost (2.28) platí v tomto případě pro $t > x > k^{-1}(k+1) > p^{-1}$. Zbytek důkazu je již zcela totožný jako v předchozím případě. Nechť konečně

$$(k+1)q = 1.$$

Nyní máme $x = p^{-1}$. Platí, že $f'(x) = f'(p^{-1}) = 1$, ostatní kořeny jsou v absolutní hodnotě ostře větší než x, tedy opět nesplňují (2.26) a jsou jednoduché. (2.28) platí pro $t > p^1 \geq k^{-1}(k+1)$ a zbytek důkazu je analogický předchozím případům. □

K tomu, abychom mohli použít Fellerovu aproximaci, potřebujeme znát hodnotu kořene x. Explicitní vzorec pro jeho výpočet bývá až na několik speciálních případů (například při $(k+1)q = 1$ okamžitě vidíme, že $x = p^{-1}$) neznámý a potřebujeme ho získat jinak. V případě, že $(k+1)q > 1$, můžeme použít Fellerův

postup aproximace hodnoty kořene x. Podobně jako v důkazu tvrzení 4 uvažujme funkci f definovanou předpisem $f(s) := 1 + qp^k s^{k+1}$. Zkonstruujeme posloupnost $\{x_n\}_{n=0}^{\infty}$, kde $x_0 := 1$ a $x_{l+1} := f(x_l)$ pro všechna přirozená čísla l. Tato posloupnost je ostře rostoucí a s rostoucím l se blíží k menšímu ze dvou kladných kořenů rovnice $f(s) = s$. Protože předpokládáme, že $(k+1)q > 1$, je tento kořen právě hledané x. V případě, že platí nerovnost $(k+1)q < 1$, není možné tento postup použít a musíme třeba hodnotu x získat jinak. V tomto případě je ale buďto k velmi malé, nebo je naopak malá hodnota q. Pro každé q přitom existuje k_0 takové, že pro všechna$k \geq k_0$ již bude platit $(k+1)q > 1$.

Uveďme ještě některé další aproximace, které jsou opět převzaty z knihy Balakrishnan, Koutras [3], přičemž autoři se odkazují dále na literaturu uvedenou dále u jednotlivých výsledků. Funkci $G(n)$ lze dobře aproximovat výrazem $\exp\{-(n-k+1)p^k\}$, přičemž pro maximální chybu tohoto odhadu platí

$$|G(n) - \exp\{-(n-k+1)p^k\}| \leq (2k-1)p^k + 2(k-1)p. \tag{2.30}$$

Odvození této aproximace založené na Chen-Steinově metodě (viz Chen [8]) je k nalezení v článku Chryssaphinou, Papastavridis [9]. Podobným způsobem dokazují Barbour, Holst a Janson [10] jiné vyjádření maximální chyby při stejné aproximaci, a to sice

$$|G(n) - \exp\{-(n-k+1)p^k\}| \leq \frac{2p}{q}(1-p^{k-1}) - (2k-3)p^k. \tag{2.31}$$

Ve stejné knize ještě autoři opět s pomocí Chen-Steinovy metody ukazují, že $G(n)$ lze lépe aproximovat výrazem $\exp\{-(n-k+1)qp^k\}$. Pro maximální možnou chybu, které se můžeme touto aproximací dopustit platí, že

$$|G(n) - \exp\{-(n-k+1)qp^k\}| \leq (2kq+1)p^k. \tag{2.32}$$

Jak uvádějí Balakrishnan a Koutras [3], aproximace uvedená v (2.32) zaručuje přesnost řádu p^k, kdežto aproximace v (2.30) a (2.31) je přesnosti řádu pouze p. Nakonec uveďme ještě další vyjádření horní a dolní meze funkce $G(n)$. Platí

$$(1-p^k)^{n-k+1} \leq G(n) \leq (1-qp^k)^{n-k+1}. \tag{2.33}$$

Různá odvození těchto mezí lze najít například v Chao, Fu[11], Chiang, Niu [12] nebo Papastavridis, Koutras [13]. Kombinací (2.12) s (2.21), (2.30), (2.31), (2.32) a (2.33) dostáváme horní a dolní meze pro $\mathsf{P}(T_k = n)$.

Fellerova aproximace dává dolní mez LB_F a horní mez UB_F tvaru

$$\begin{aligned} LB_F &= qp^k \left[\frac{1-px}{(k+1-kx)q} \cdot \frac{1}{x^{n-k}} - \frac{2(k-1)p^{n-k}}{(k+1+kp^{-1})q} \right], \\ UB_F &= qp^k \left[\frac{1-px}{(k+1-kx)q} \cdot \frac{1}{x^{n-k}} + \frac{2(k-1)p^{n-k}}{(k+1+kp^{-1})q} \right]. \end{aligned} \tag{2.34}$$

Použitím (2.30) dostáváme první horní a dolní Chen-Steinovu mez $UB_{CS}^{(1)}$, respektive $LB_{CS}^{(1)}$, kde

$$\begin{aligned} LB_{CS}^{(1)} &= qp^k\{\exp[-(n-2k)p^k] - (2k-1)p^k - 2(k-1)p\}, \\ UB_{CS}^{(1)} &= qp^k\{\exp[-(n-2k)p^k] + (2k-1)p^k + 2(k-1)p\}. \end{aligned}$$

Z (2.31) máme druhé Chen-Steinovy meze $UB_{CS}^{(2)}$ a $LB_{CS}^{(2)}$, přičemž

$$\begin{aligned} LB_{CS}^{(2)} &= qp^k\{\exp[-(n-2k)p^k] - \frac{2p}{q}(1-p^{k-1}) - (2k-3)p^k\}, \\ UB_{CS}^{(2)} &= qp^k\{\exp[-(n-2k)p^k] + \frac{2p}{q}(1-p^{k-1}) - (2k-3)p^k\}. \end{aligned}$$

Z (2.32) pak plynou třetí Chen-Steinovy meze $UB_{CS}^{(3)}$ a $LB_{CS}^{(3)}$. Platí

$$\begin{aligned} LB_{CS}^{(3)} &= qp^k\{\exp[-(n-2k)qp^k] - (2kq+1)p^k\}, \\ UB_{CS}^{(3)} &= qp^k\{\exp[-(n-2k)qp^k] + (2kq+1)p^k\}. \end{aligned}$$

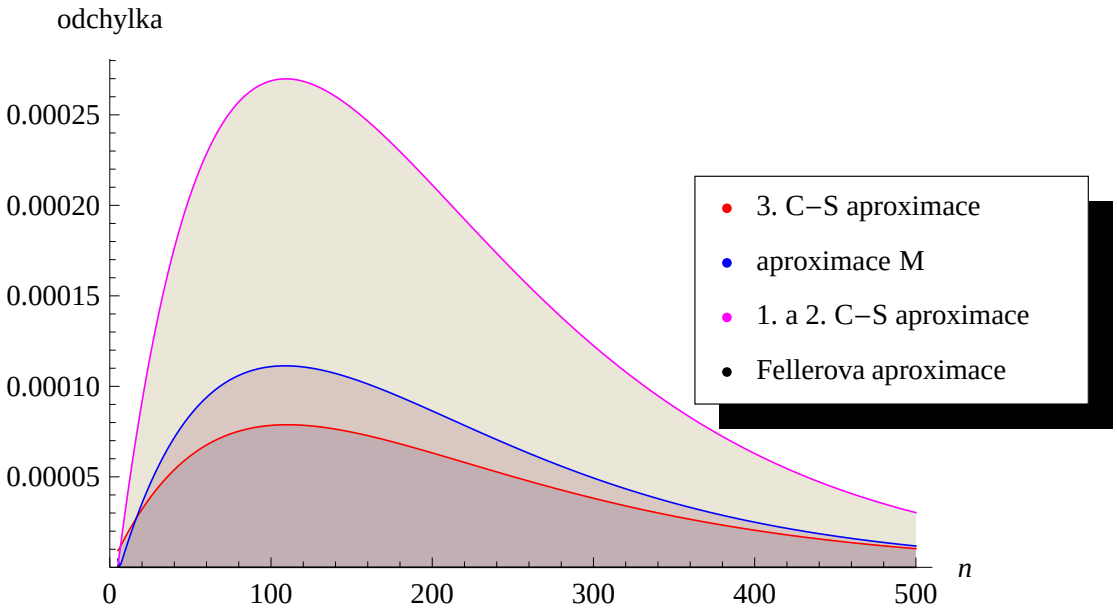
Konečně použitím (2.12) na (2.33) získáme horní mez UB a dolní mez LB, kde

$$\begin{aligned} LB &= qp^k(1-p^k)^{n-2k}, \\ UB &= qp^k(1-qp^k)^{n-2k}. \end{aligned}$$

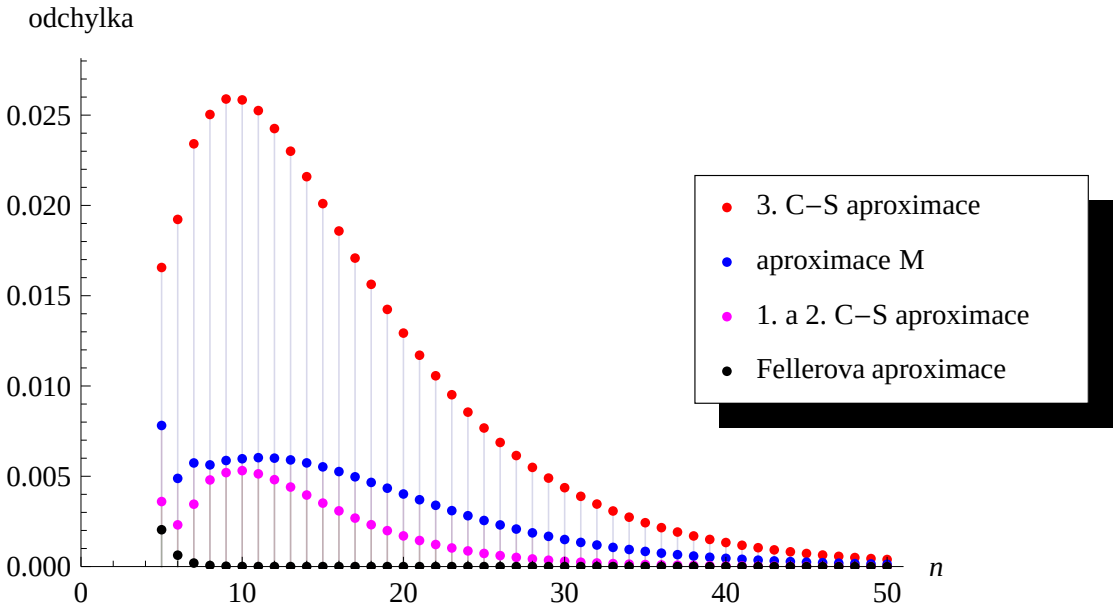
Poznámka 5. Fellerova a Chen-Steinovy meze mohou při určitých parametech být záporné, nebo být větší než 1. V takovém případě dosadíme za hodnotu meze nulu, respektive jedničku.

Srovnejme nyní numericky jednotlivé aproximace. Jako první Chen-Steinovu aproximaci označme výraz $[UB_{CS}^{(1)} + LB_{CS}^{(1)}]/2$, druhá Chen-Steinova aproximace nechť je $[UB_{CS}^{(2)}+LB_{CS}^{(2)}]/2$ a třetí $[UB_{CS}^{(3)}+LB_{CS}^{(3)}]/2$. Definujme aproximaci M jako $M := (LB + UB)/2$. Je okamžitě vidět, že první a druhá Chen-Steinova aproximace jsou totožné. Obrázky 2.1–2.6 ukazují absolutní odchylku jednotlivých aproximací od přesné hodnoty $\mathsf{P}(T_k = n)$, vypočtené pomocí vzorce z tvrzení 1. Konkrétní hodnoty p a k jsou uvedeny v popisu u každého grafu. Červeně je vyznačena odchylka první respektive druhé Chen-Steinovy aproximace, modře odchylka třetí Chen-Steinovy aproximace, purpurové hodnoty představují odchylku mezní aproximace a černě jsou označeny odchylky Fellerovy aproximace. Hodnoty jsou zobrazené pouze pro $n \geq 2k + 1$, neboť pro nižší hodnoty nemá aproximace vzhledem k (2.1) valný smysl.

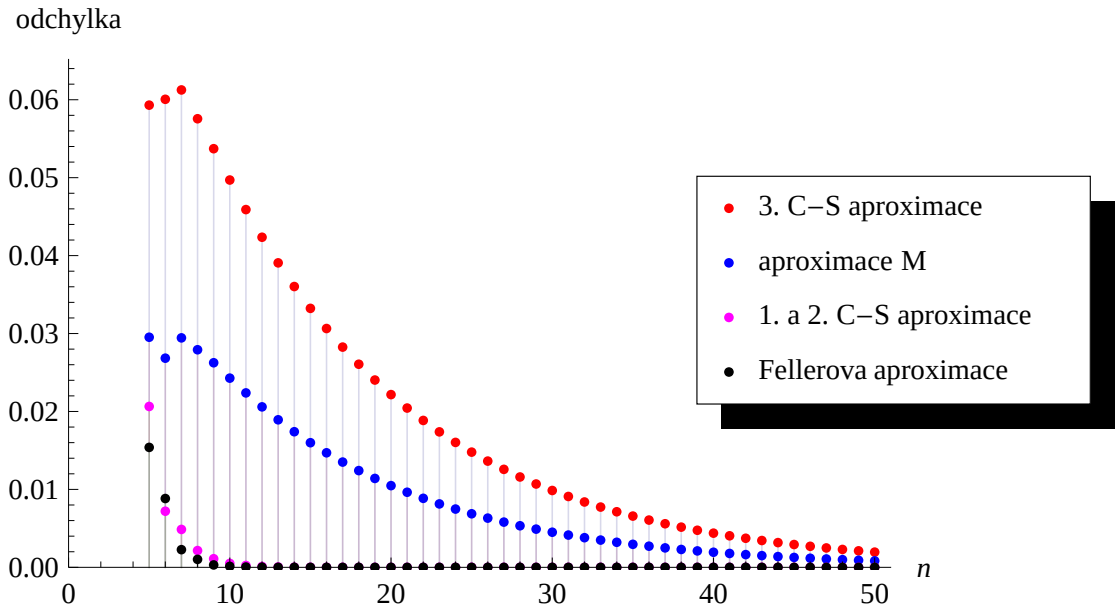
Poznámka 6. Veškeré výpočty byly prováděny a všechny grafy konstruovány v programu Wolfram Mathematica 7. V příloze je uveden příslušný zdrojový kód.



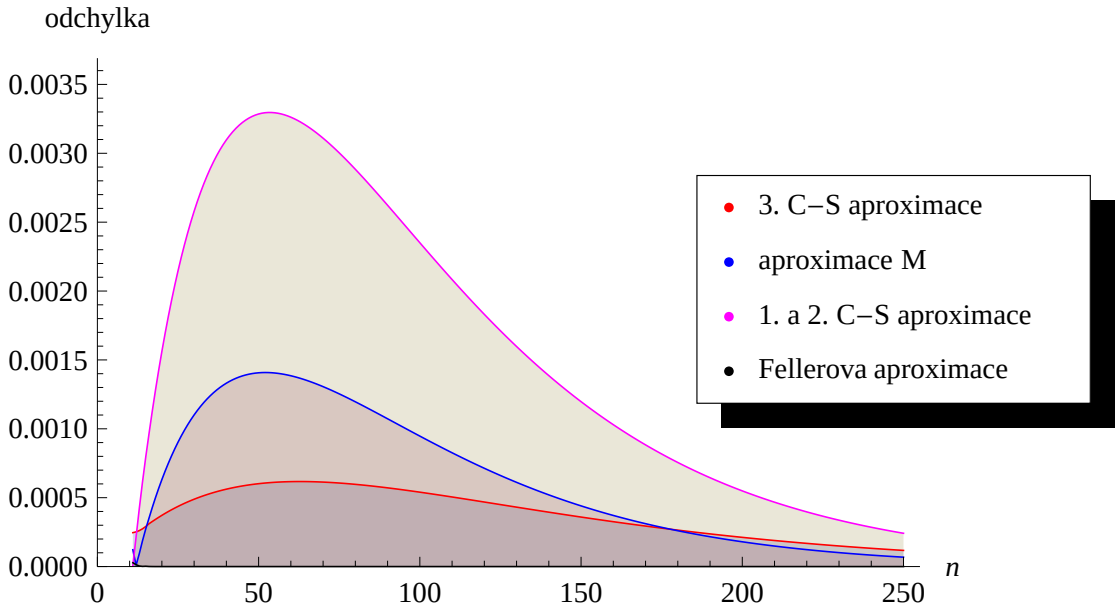
Obrázek 2.1: $p = 0.1$, $k = 2$



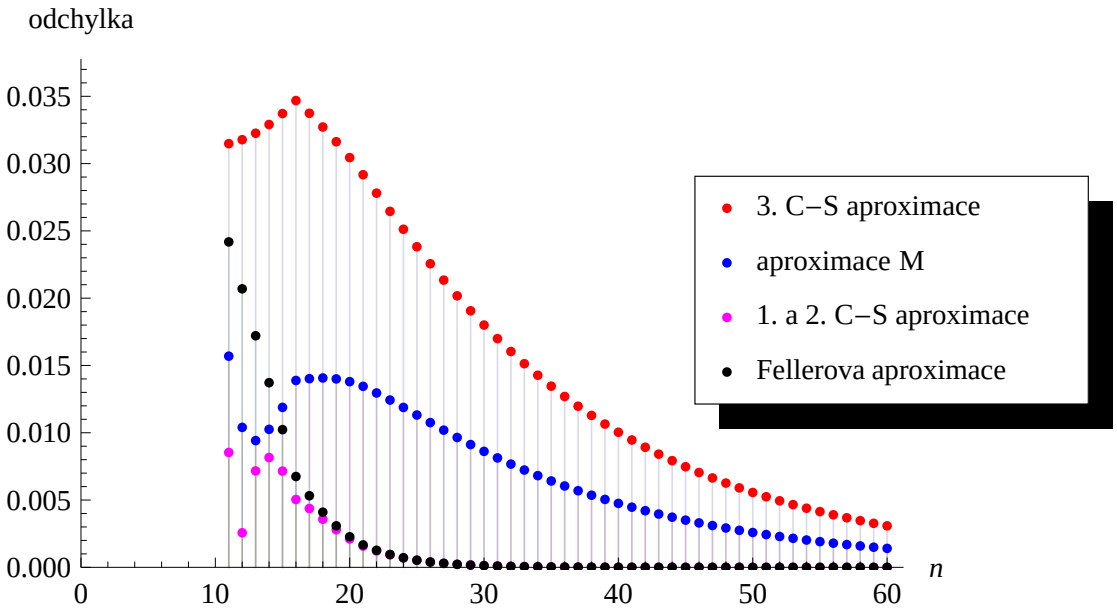
Obrázek 2.2: $p = 0.5$, $k = 2$



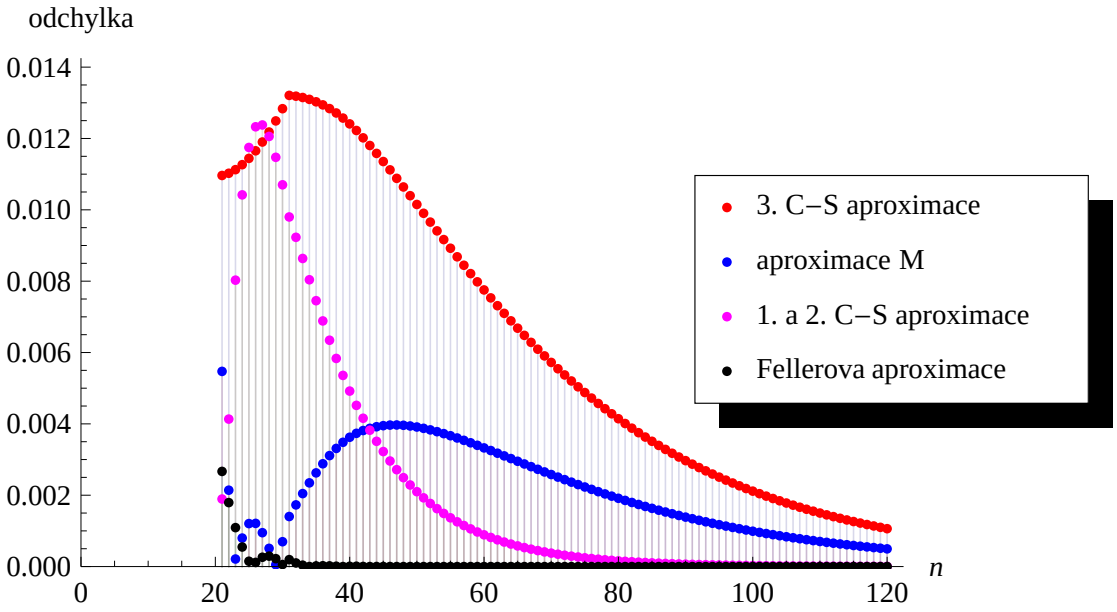
Obrázek 2.3: $p = 0.9$, $k = 2$



Obrázek 2.4: $p = 0.5$, $k = 5$



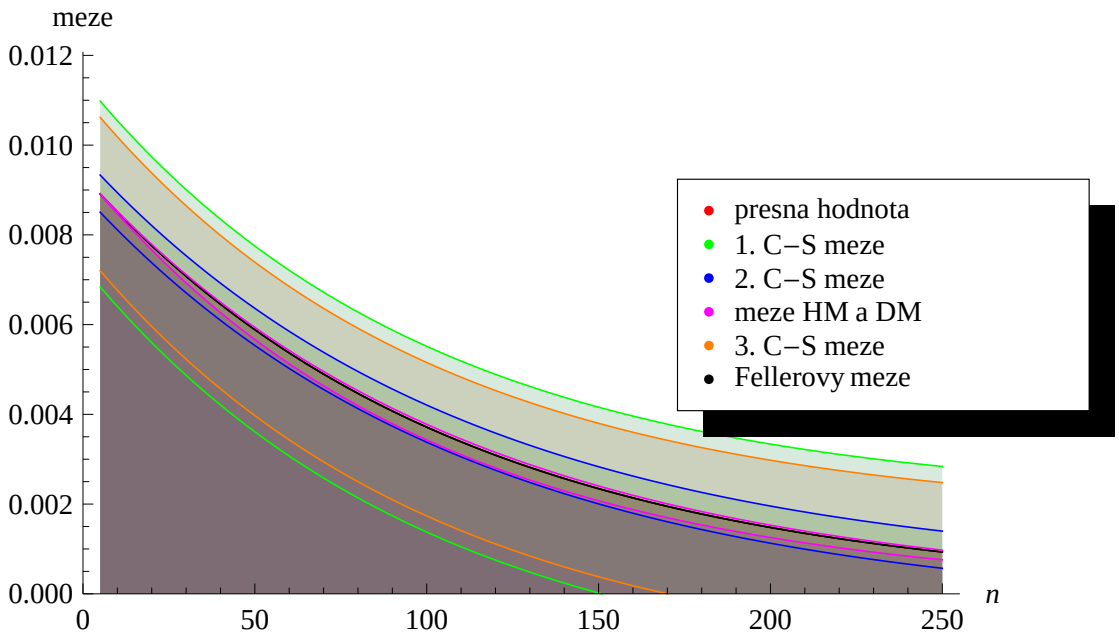
Obrázek 2.5: $p = 0.9$, $k = 5$



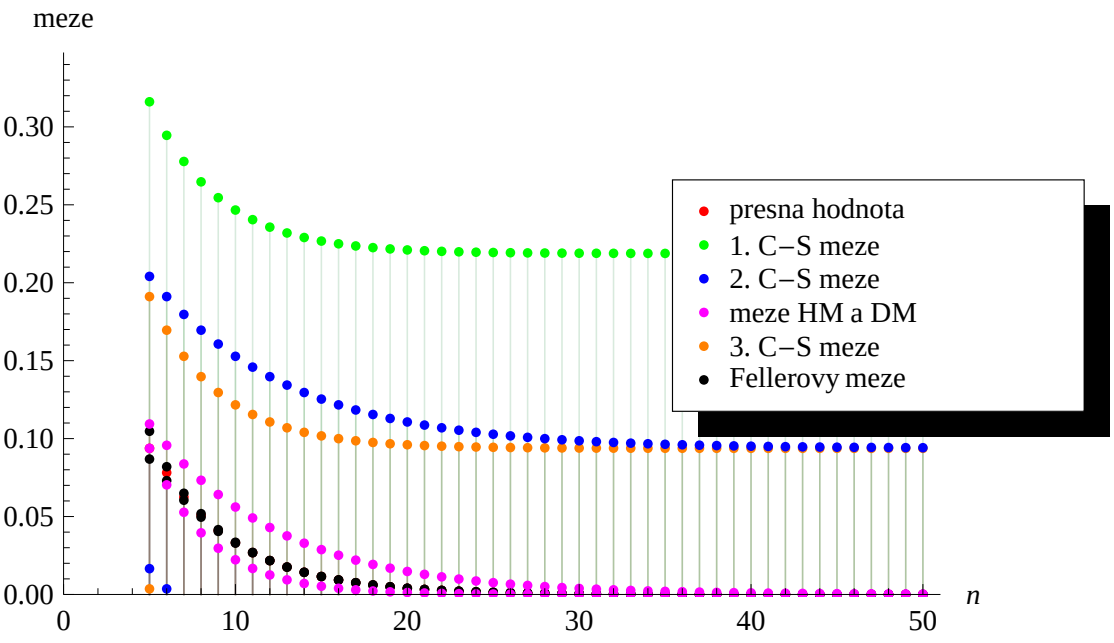
Obrázek 2.6: $p = 0.9$, $k = 10$

Zřejmě Fellerova aproximace je ze všech uvedených vždy nejpřesnější a konverguje k přesné hodnotě $\mathsf{P}(T_k = n)$ velmi rychle, lze ji proto efektivně použít i při nižších hodnotách n. Bohužel její výpočet je ze všech nejnáročnější, jelikož musíme hledat kořen x, který je třeba ve většině případů aproximovat výše uvedeným způsobem. Jak je vidět z obrázků, je-li výskyt iterace „vzácný“ jev čili je-li p^k „malé“, ostatní aproximace dávají také poměrně přesný odhad $\mathsf{P}(T_k = n)$ i pro nízká n, přičemž jejich konvergence k přesné hodnotě není nutně vždy monotónní. Ve zmíněném případě, tedy pro malé p^k, se jeví jako nejvhodnější (neuvažujeme-li aproximaci Fellerovu) Chen-Steinovy aproximace, pokud je naopak výskyt iterace délky k častý jev, pak nám aproximace M dává přesnější hodnoty.

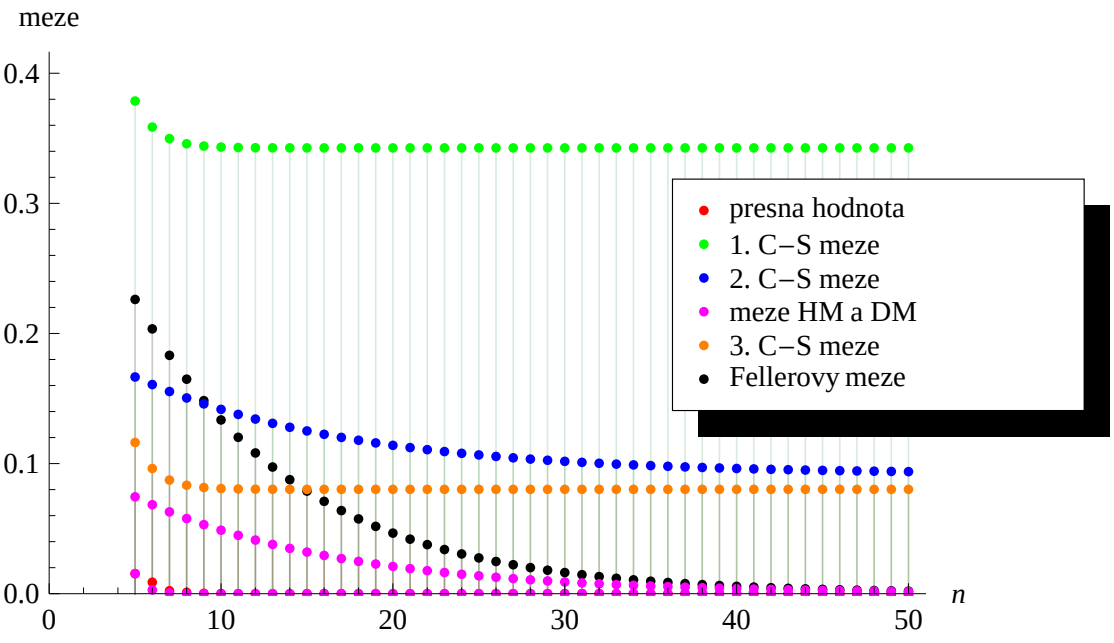
Nyní se zaměřme na meze příslušné jednotlivým aproximacím a opět je numericky srovnáme. Obrázky 2.7–2.13 zobrazují přesnou hodnotu $\mathsf{P}(T_k = n)$ (červeně) a hodnoty horních a dolních mezí první, druhé a třetí Chen-Steinovy aproximace (po řadě obě meze zeleně, obě modře, obě oranžově), mezí LB a UB (obě purpurově) a horní a dolní Fellerovy meze (obě vyznačeny černou barvou). V duchu Poznámky 5 není v některých případech vyznačena dolní Chen-Steinova nebo Fellerova mez, neboť nabývají pro dané p a k záporných hodnot.



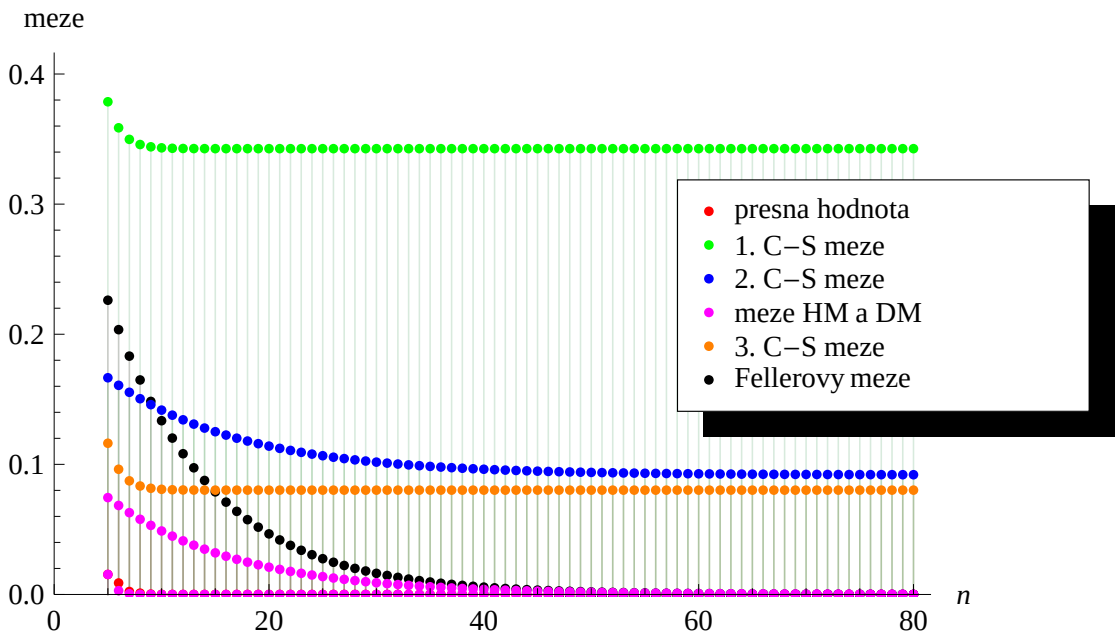
Obrázek 2.7: $p = 0.1$, $k = 2$



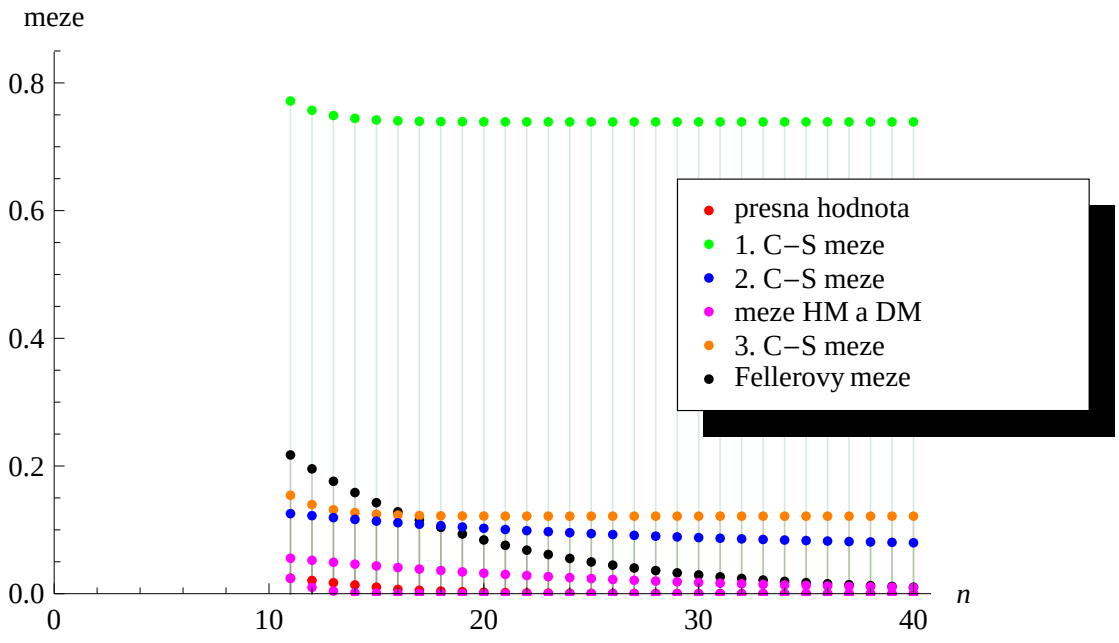
Obrázek 2.8: $p = 0.5$, $k = 2$



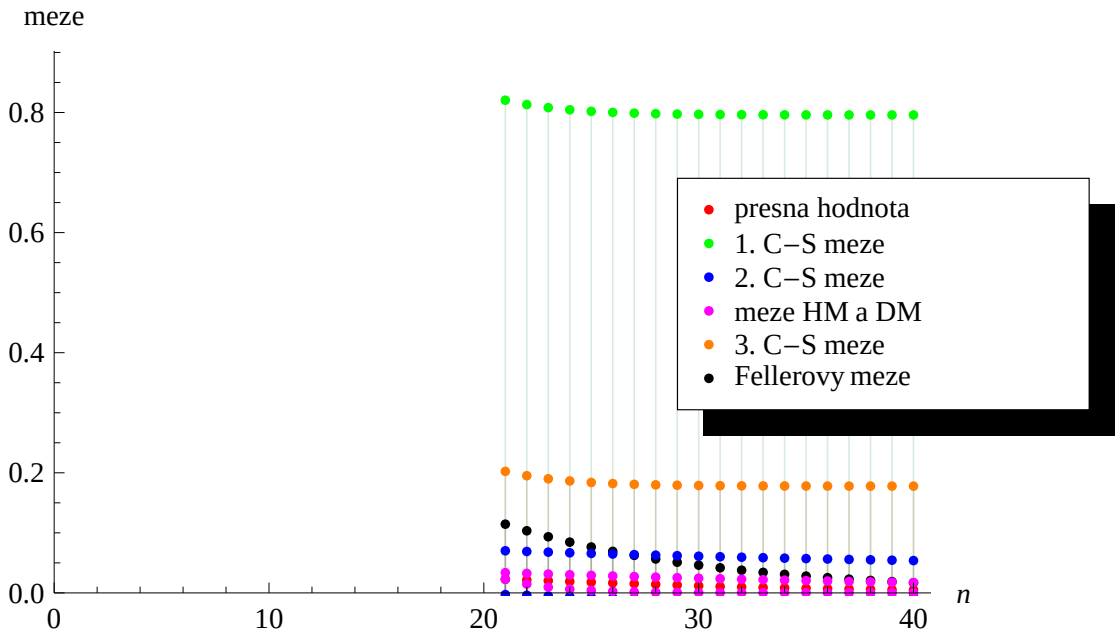
Obrázek 2.9: $p = 0.9$, $k = 2$



Obrázek 2.10: $p = 0.5$, $k = 5$



Obrázek 2.11: $p = 0.9$, $k = 5$



Obrázek 2.12: $p = 0.9$, $k = 10$

V tomto se jeví jako nejvhodnější meze LB a UB. Fellerovy meze rovněž dávají velmi dobré výsledky, pro některé hodnoty p a k jsou dokonce o něco přesnější než meze LB a UB, jejich nevýhodou ale zůstává již výše zmíněná náročnost výpočtu. Z Chen-Steinových mezí se ukazuje třetí varianta jako ta, jejíž intervaly jsou nejužší.

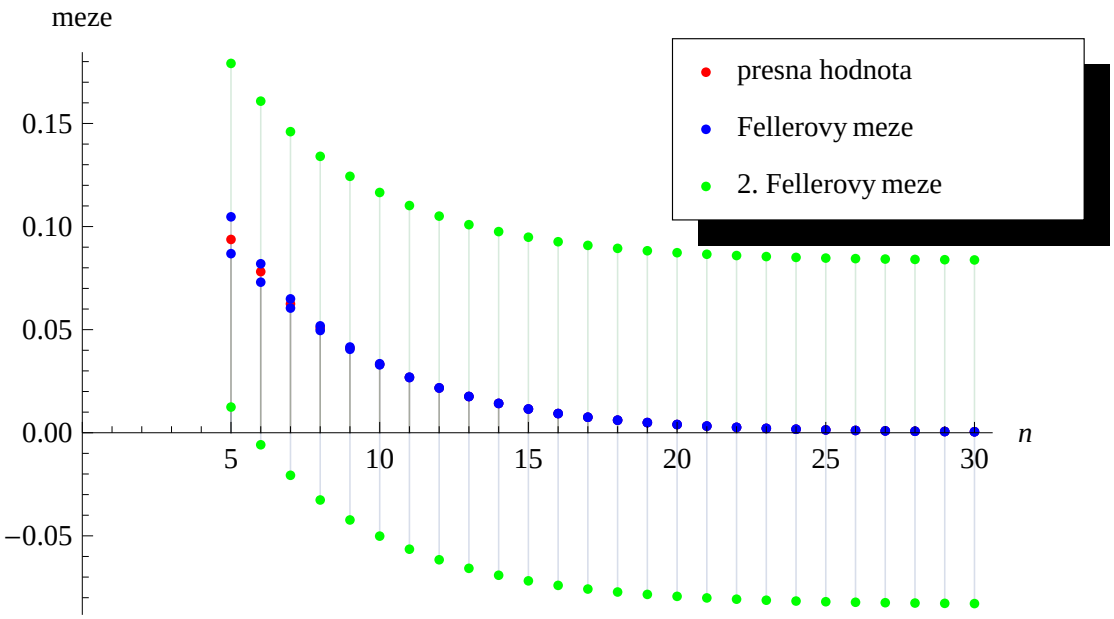
Poznámka 7. V knize Feller [2] na straně 326 je jako maximální možná odchylka Fellerovy aproximace od $G(n)$ uveden výraz

$$\frac{2(k+1)p}{qp(k+1)}.$$

Dostáváme pak druhé horní a dolní Fellerovy meze UB_{F2} a LB_{F2} tvaru

$$\begin{aligned} LB_{F2} &= qp^k \left[\frac{1-px}{(k+1-kx)q} \cdot \frac{1}{x^{n-k}} - \frac{2(k+1)p}{qp(k+1)} \right], \\ UB_{F2} &= qp^k \left[\frac{1-px}{(k+1-kx)q} \cdot \frac{1}{x^{n-k}} + \frac{2(k+1)p}{qp(k+1)} \right]. \end{aligned} \tag{2.35}$$

Tato aproximace je sice platná, ale dává zbytečně široké intervaly pro hodnotu $\mathsf{P}(T_k = n)$. Srovnejme si meze (2.35) s mezemi (2.34). Modře jsou na obrázku 2.13 vyznačeny meze (2.34) a zeleně meze (2.35), červená barva opět značí přesnou hodnotu $\mathsf{P}(T_k = n)$.



Obrázek 2.13: $p = 0.5$, $k = 2$

2.4 Podmíněné rozdělení

Na konec této kapitoly se ještě podívejme jak vypadá podmíněné rozdělení veličin T_k a L_n. Výsledky, které zde získáme, vzápětí použijeme ve třetí kapitole k odvození testu nezávislosti založeném na délce nejdelšího řetězce úspěchů.

Mějme nyní konečnou posloupnost n Bernoulliových pokusů, přičemž tentokrát předpokládáme, že počet úspěchů a neúspěchů je nám znám. Označme počet úspěchů jako S_n a počet neúspěchů jako y. Zřejmě platí, že $S_n := n - y$ a $0 \leq y \leq n$. Situaci si můžeme představit snadno na urnovém modelu, kde y neúspěchů představuje $y + 1$ urn, do kterých umístíme $n - y$ míčků, tedy přesně tolik, kolik se v posloupnosti vyskytlo úspěchů. Předpokládejme, že $L_n < k$, tedy všechny iterace úspěchů jsou délky nejvýše $k - 1$. V našem urnovém modelu to bude znamenat, že v každé urně je nejvýše $k - 1$ míčků. Nechť $N(a, b, c)$ představuje počet možností, jak lze rozdělit a identických míčků do b různých urn za předpokladu, že do každé urny se vejde nanejvýš c míčků. Dle knihy Balakrishnan, Koutras [3] lze $N(a, b, c)$ vyjádřit jako

$$N(a, b, c) = \sum_{j=0}^{\lfloor \frac{a}{c+1} \rfloor} (-1)^j \binom{b}{j} \binom{a - j(c+1) + b - 1}{b - 1}. \tag{2.36}$$

Autoři se odkazují na knihy Riordan [14] a Johnson a Kotz [15]. S použitím (2.36) můžeme psát

$$\mathsf{P}(L_n < k, S_n = n - k) = \frac{N(n - y, y + 1, k - 1)}{2^n}.$$

Protože zřejmě

$$\mathsf{P}(S_n = n - y) = \frac{\binom{n}{y}}{2^n},$$

dostáváme

$$\begin{aligned}\mathsf{P}(L_n < k | S_n = n-y) &= \frac{\mathsf{P}(L_n < k, S_n = n-k)}{\mathsf{P}(S_n = n-y)} = \frac{N(n-y, y+1, k-1)}{\binom{n}{y}} \\ &= \binom{n}{y}^{-1} \sum_{j=0}^{\lfloor \frac{n-y}{k} \rfloor} (-1)^j \binom{y+1}{j} \binom{n-jk}{y}.\end{aligned}$$

S pomocí vztahu (2.4) dostáme podmíněnou distribuční funkci veličiny T_k jako

$$\begin{aligned}\mathsf{P}(T_k \leq n | S_n = n-y) &= 1 - \mathsf{P}(L_n < k | S_n = n-y) \\ &= \frac{\binom{n}{y} - \binom{n}{y} - \sum_{j=1}^{\lfloor \frac{n-y}{k} \rfloor} (-1)^j \binom{y+1}{j} \binom{n-jk}{y}}{\binom{n}{y}} \\ &= \binom{n}{y}^{-1} \sum_{j=1}^{\lfloor \frac{n-y}{k} \rfloor} (-1)^{j+1} \binom{y+1}{j} \binom{n-jk}{y}. \quad (2.37)\end{aligned}$$

Podmíněnou pravděpodobnost $\mathsf{P}(T_k = n | S_n = n-y)$ dostaneme snadnou úvahou z našeho urnového modelu. Má-li k prvnímu výskytu iterace úspěchů délky k dojít v n-tém pokusu, znamená to, že v poslední urně se nachází právě k míčků, takže nám k rozdělení do zbylých y uren zbývá $n-y-k$ míčků, a protože před n-tým pokusem se iterace úspěchů délky k nikdy nevyskytla, v každé z těchto zbylých uren může být míčků nanejvýš $k-1$. Z této úvahy již snadno odvodíme podmíněné rozdělení veličiny T_k jako

$$\begin{aligned}\mathsf{P}(T_k = n | S_n = n-y) &= \frac{\mathsf{P}(T_k = n, S_n = n-k)}{\mathsf{P}(S_n = n-y)} = \frac{N(n-y-k, y, k-1)}{\binom{n}{y}} \\ &= \binom{n}{y}^{-1} \sum_{j=0}^{\lfloor \frac{n-y-k}{k} \rfloor} (-1)^j \binom{y}{j} \binom{n-j(k+1)-1}{y-1}\end{aligned}$$

pro $n \geq y+k$. Pro $n < y+k$ je tato pravděpodobnost očividně nulová.

3. Testy nezávislosti

Protože množství různých (a často velmi zajímavých) aplikací iterací v poslopnosti Bernoulliových pokusů vysoce převyšuje rozsah této práce, zaměříme se pouze na použití iterací při testování hypotézy nezávislosti náhodných veličin. Je nutno podotknout, že i toto téma je velice rozsáhlé a podrobně se budeme zabývat pouze vybranými metodami. Mnoho dalších zajímavých aplikací včetně odkazů na literaturu, která se jimi zabývá podrobněji, je k nalezení v knize Balakrishnan, Koutras [3].

Ve statistické teorii i praxi často předpokládáme nezávislost námi zkoumaných veličin. Proto je důležité mít k dispozici nástroj, který nám pomůže s ověřením nezávislosti nějakého výběru náhodných veličin. Jedním ze způsobů, jak testovat hypotézu nezávislosti, jsou testy založené na iteracích Bernoulliových pokusů. Nemusíme se přitom omezovat na náhodné veličiny s alternativním rozdělením. Nabývají-li naše náhodné veličiny reálných hodnot, stanovíme určitou hranici, většinou medián výběru, a všechny hodnoty, které jsou větší než tato hranice, označíme jako úspěch. Naopak menší hodnoty označíme jako neúspěch. Hodnoty, které jsou mediánu rovny, vynecháme a o jejich počet snížíme rozsah výběru. V případě, že náhodné veličiny nabývají nějakých kvalitativních hodnot (například muž/žena), stanovíme v závislosti na konkrétním případě nějaké kritérium, podle kterého opět převedeme výběr na posloupnost Bernoulliových pokusů. Podrobně se podíváme na dva testy vycházející z iterací v posloupnosti Bernoulliových pokusů. Jde o test na základě délky nejdelší iterace úspěchů a test na základě celkového počtu iterací.

Předpokládejme, že máme k dispozici právě jednu realizaci náhodných veličin $X_1, \ldots, X_n$. V úvodu této podkapitoly jsme popsali způsob, jakým naši realizaci náhodných veličin $X_1, \ldots, X_n$ upravíme na realizaci posloupnosti Bernoulliových pokusů $X_1^*, \ldots, X_n^*$. Jako H_0 označíme hypotézu, že veličiny $X_1^*, \ldots, X_n^*$ jsou nezávislé, a tedy jsou nezávislé i veličiny $X_1, \ldots, X_n$. Chceme testovat H_0 proti alternativní hypotéze, že mezi veličinami $X_1^*, \ldots, X_n^*$ existuje závislost.

3.1 Test založený na délce nejdelší iterace úspěchů

K odvození testu založeném na délce nejdelší iterace úspěchů potřebujeme najít pravděpodobnost, že nejdelší iterace v posloupnosti n Bernoulliových pokusů je délky k, za podmínky, že z n pokusů je jich právě y neúspěšných. Tuto pravděpodobnost získáme přímo ze vzorce (2.37) a ze vztahu (2.5). Můžeme psát

$$\begin{aligned}
\mathsf{P}(L_n &= k|S_n = n-y) = \mathsf{P}(T_k \le n) - \mathsf{P}(T_{k+1} \le n) \\
&= \binom{n}{y}^{-1} \sum_{j=1}^{\lfloor \frac{n-y}{k} \rfloor} (-1)^{j+1} \binom{y+1}{j} \binom{n-jk}{y} \\
&\quad - \binom{n}{y}^{-1} \sum_{j=1}^{\lfloor \frac{n-y}{k+1} \rfloor} (-1)^{j+1} \binom{y+1}{j} \binom{n-j(k+1)}{y}. \qquad (3.1)
\end{aligned}$$

Nechť je počet úspěchů v $X_1^*, \ldots, X_n^*$ roven $S_n = n - y$, kde y je počet neúspěchů. V případě platnosti hypotézy H_0 se bude délka nejdelší iterace úspěchů při daném S_n řídit rozdělením (3.1). Hypotézu nezávislosti zamítneme, jestliže skutečná délka nejdelší iterace úspěchů je buď příliš velká, nebo naopak příliš malá. Nechť l_1 je nejmenší číslo splňující nerovnost $\mathsf{P}(L_n \leq l_1 | S_n = n - y) \leq \frac{\alpha}{2}$ a l_2 největší takové číslo, jež splňuje $\mathsf{P}(L_n \geq l_1 | S_n = n - y) \leq \frac{\alpha}{2}$. Hypotézu H_0 tedy zamítáme v případě, že pro skutečnou délku nejdelší iterace úspěchů L_n^* platí, že $L_n^* \leq l_1$ nebo $L_n^* \geq l_2$. Přitom skutečná hladina testu je nejvýše rovna α.

3.2 Test založený na celkovém počtu iterací

V tomto případě nás nebudou zajímat délky iterací, ale jejich počet, a tentokrát bereme v úvahu i iterace neúspěchů. Bude-li celkový počet iterací příliš nízký, nebo naopak příliš vysoký, zamítneme hypotézu nezávislosti. K odvození testu potřebujeme znát rozdělení veličiny R, která představuje celkový počet iterací v posloupnosti n (nezávislých) Bernoulliových pokusů. Dále označme R_1 jako počet iterací úspěchů a R_2 jako počet iterací neúspěchů. Zřejmě platí $R = R_1 + R_2$. Počet úspěchů v tomto případě označme jako n_1 a počet neúspěchů jako n_2. Opět je zjevně $n = n_1 + n_2$. Budeme postupovat obdobně jako Gibbons [16]. Nejdříve určíme sdružené rozdělení veličin R_1 a R_2 a z něj pak rozdělení jejich součtu.

Poznámka 8. Pro $a < b$ budeme definovat $\binom{a}{b}$ jako nulu.

Tvrzení 5. Sdružená pravděpodobnost veličin R_1 a R_2 je

$$\mathsf{P}(R_1 = r_1, R_2 = r_2) = \frac{c \cdot \binom{n_1-1}{r_1-1}\binom{n_2-1}{r_2-1}}{\binom{n_1+n_2}{n_1}},$$

kde $r_1 = 1, 2, \ldots, n$, $r_2 = 1, 2, \ldots, n_2$, $c = 2$ pro $r_1 = r_2$ a $c = 1$ pro $r_1 = r_2 \pm 1$. Při jiných hodnotách r_1 a r_2 je tato pravděpodobnost nulová.

Důkaz. Situaci si zde (podobně jako při odvození podmíněného rozdělení v předchozí kapitole) představíme opět na urnovém modelu. Máme tolik bílých míčků, kolik je v posloupnosti úspěchů,tedy n_1 a tolik černých míčků, kolik je neúspěchů (n_2). Dohromady máme $r = r_1 + r_2$ uren, do kterých míčky rozmisťujeme. Má-li být počet iterací úspěchů roven r_1, znamená to, že musíme n_1 bílých míčků (úspěchů) rozmístit do r_1 uren, přičemž v každé urně má být alespoň jeden míček. Takto lze míčky umístit právě $\binom{n_1-1}{r_1-1}$ způsoby. Obdobná úvaha platí pro počet iterací neúspěchů. Začíná-li posloupnost úspěchem (v první urně je bílý míček), existuje tedy právě $\binom{n_1-1}{r_1-1}\binom{n_2-1}{r_2-1}$ způsobů, jak uspořádat úspěchy a neúspěchy tak, aby bylo v posloupnosti právě r_1 iterací úspěchů (r_1 uren obsahujících pouze bílé míčky) a r_2 iterací neúspěchů (r_2 uren obsahujících pouze černé míčky). Začíná-li posloupnost neúspěchem (v první urně je černý míček), je úvaha zcela analogická. Protože iteraci úspěchů musí nutně následovat iterace neúspěchů a naopak, urny s bílými a černými míčky se pravidelně střídají. Vzhledem k tomu, že $r = r_1 + r_2$, musí tedy nutně platit buď $r_1 = r_2 \pm 1$, nebo $r_1 = r_2$, a tedy v ostatních případech je pravděpodobnost $\mathsf{P}(R_1 = r_1, R_2 = R_2)$ zřejmě nulová. Pokud $r_1 = r_2 + 1$ respektive $r_1 = r_2 - 1$, začíná posloupnost iterací úspěchů respektive neúspěchů. V případě, že $r_1 = r_2$, může posloupnost začínat jak úspěchem, tak neúspěchem, a proto v tomto případě je počet možných uspořádání dvojnásobný ($c = 2$). Tím je tvrzení dokázáno. □

Tvrzení 6. Rozdělení pravděpodobnosti veličiny $R = R_1 + R_2$ je dáno vzorcem

$$\mathsf{P}(R=r) = \begin{cases} \frac{2\cdot\binom{n_1-1}{r/2-1}\binom{n_2-1}{r/2-1}}{\binom{n_1+n_2}{n_1}} & \text{pro } r \text{ sudé,} \\ \frac{\binom{n_1-1}{(r-1)/2}\binom{n_2-1}{(r-3)/2}+\binom{n_1-1}{(r-3)/2}\binom{n_2-1}{(r-1)/2}}{\binom{n_1+n_2}{n_1}} & \text{pro } r \text{ liché,} \end{cases}$$

je-li $r = 2, 3, \ldots, n_1 + n_2$. Pro jiné hodnoty r je $\mathsf{P}(R = r) = 0$.

Důkaz. Je-li počet všech iterací r sudé číslo, musí existovat stejně iterací úspěchů jako neúspěchů čili $r_1 = r_2 = r/2$. Vzhledem k tomu dostáváme z tvrzení 5 pravděpodobnost $\mathsf{P}(R = r)$ pro r sudé jako

$$\mathsf{P}(R=r) = \mathsf{P}(R_1 = r/2, R_2 = r/2) = \frac{2\cdot\binom{n_1-1}{r/2-1}\binom{n_2-1}{r/2-1}}{\binom{n_1+n_2}{n_1}}.$$

V případě, že r je liché, máme $r_1 = \pm r_2$, takže buď $r_1 = (r+1)/2$ a $r_2 = (r-1)/2$, nebo $r_1 = (r-1)/2$ a $r_2 = (r+1)/2$. Opět s pomocí tvrzení 5 získáme pravděpodobnost $\mathsf{P}(R = r)$ pro sudé r. Můžeme psát

$$\begin{aligned} \mathsf{P}(R=r) &= \mathsf{P}(R_1 = (r+1)/2, R_2 = (r-1)/2) \\ &\quad + \mathsf{P}(R_1 = (r-1)/2, R_2 = (r+1)/2) \\ &= \frac{\binom{n_1-1}{(r-1)/2}\binom{n_2-1}{(r-3)/2} + \binom{n_1-1}{(r-3)/2}\binom{n_2-1}{(r-1)/2}}{\binom{n_1+n_2}{n_1}}. \end{aligned}$$

Tím je tvrzení věty dokázáno. □

Jestliže platí hypotéza nezávislosti H_0, bude se celkový počet iterací řídit rozdělením veličiny R. H_0 tedy zamítneme na hladině nejvýše rovné α, platí-li pro skutečný počet iterací R^*, že buď $R^* \leq l_1$, nebo $R^* \geq l_2$, kde l_1 je nejmenší číslo splňující nerovnost $\mathsf{P}(R \leq l_1) \leq \frac{\alpha}{2}$ a l_2 největší takové číslo, jež splňuje $\mathsf{P}(L_n \geq l_1) \leq \frac{\alpha}{2}$.

Předpokládejme, že n roste nade všechny meze tím způsobem, že $\lambda = \frac{n_1}{n}$ zůstává konstantní. Pak lze dokázat (viz Gibbons [16]), že veličina

$$Z = \frac{R - 2n\lambda(1-\lambda)}{2\sqrt{n}\lambda(1-\lambda)}$$

má asymptoticky rozdělení $N(0,1)$. Při velkých hodnotách n můžeme test založit na veličině Z a v tom případě zamítáme hypotézu nezávislosti na hladině α, platí-li

$$|Z^*| = \left|\frac{R^* - 2n\lambda(1-\lambda)}{2\sqrt{n}\lambda(1-\lambda)}\right| \geq u_{\alpha/2}.$$

Jak úvádí Gibbons [16], síla testu založeného na celkovém počtu iterací může být v závislosti na srovnávaných datech vyšší nebo naopak nižší než síla testu založeného na délce nejdelší iterace úspěchů. Oba testy používají pouze část dostupné informace, neboť délka nejdelší iterace neurčuje zcela délky ostatních iterací a jejich celkový počet a naopak z celkového počtu iterací neplynou jednoznačně jejich délky.

4. Závěr

Definovali jsme pojem iterace a odvodili explicitní vzorec pro geometrické rozdělení řádu k. Protože přesný výpočet tohoto rozdělení je pro velká k výpočetně náročný, odvodili jsme ještě Fellerovu aproximaci a určili její horní a dolní mez. Poté jsme provedli numerické srovnání Fellerovy a několika dalších aproximací a jejich mezí. Z výsledků vyplývá, že zdaleka nejpřesnější aproximace bývá právě Fellerova, bohužel její výpočet je současně ze všech srovnávaných aproximací nejnáročnější vzhledem k nutnosti hledat kořen x pomocí aproximujících iterací. Z ostatních aproximací bývají nejvhodnější Chen-Steinovy nebo aproximace, kterou jsme označili jako M, a to podle velikosti hodnoty p^k. Ze všech zkoumaných mezí nám v závislosti na parametrech p a k nejužší intervaly dávají Fellerovy meze nebo meze LB a UB. Jiná verze Fellerových mezí, která je uváděná v knize Feller [2], je přitom naopak ze všech nejméně přesná. Na konci druhé kapitoly jsme odvodili podmíněné rozdělení při daném počtu úspěchů a pomocí něj pak test hypotézy nezávislosti založený na délce nejdelší iterace úspěchů. Nakonec jsme získali ještě test nezávislosti založený na celkovém počtu iterací. Z časových důvodů už bohužel nebylo provedeno vlastní srovnání těchto testů a omezili jsme se na citaci výsledků z knihy Gibbons [16].

Seznam použité literatury

[1] RÉVÉSZ, P. *Strong theorems on coin tossing.* Proceedings of the International Congress of Mathematicians, 749–759, Helsinki, 1978

[2] FELLER, W. *An Introduction to Probability Theory and Its Applications.* Volume I, Third edition, John Wiley and Sons, New York, 1968

[3] BALAKRISHNAN, N., KOUTRAS M. V. *Runs and Scans with Applications.* John Wiley and Sons, New York, 2002

[4] PHILIPPOU, A. N., MUWAFI, A. A. *Waiting fo the k-th consecutive success and the fibonacci sequence of order k.* The Fibonacci Quarterly, 20, 28–32, 1982

[5] PHILIPPOU, A. N., MAKRI, F. S. *Longest success runs and Fibonacci-type polynomials.* The Fibonacci Quarterly, 23, 338–346, 1985

[6] UPPULURI, V. R. R, PATIL, S. A. *Waiting times and generalized Fibonacci sequences.* The Fibonacci Quarterly, 21, 242–249, 1983

[7] MUSELLI, M. *Simple expressions for success run distributions in Bernoulli trials.* Statistics and Probability Letters, 31, 121–128, 1996

[8] CHEN, L. H. Y. *Poisson approximation for dependent trials.* Annals of Probability, 3, 534–545, 1975

[9] CHRYSSAPHINOU, O., PAPASTAVRIDIS, S. *Limit distribution for a consecutive-k-out-of-n: F system.* Advances in Applied Probability, 22, 491–493, 1990

[10] BARBOUR, A. D., HOLST, L., JANSON S. *Poisson Approximations.* Oxford University Press, New York, 1992

[11] CHAO, M. T., FU, J. C. *The reliability of large series system under a Markovian structure.* Advances in Applied Probability, 23, 894–908, 1991

[12] CHAO, M. T., FU, J. C. *Reliability of consecutive-k-out-of-n: F system.* IEEE Transactions on Reliability, 30, 87–89, 1981

[13] PAPASTAVRIDIS, S. G., KOUTRAS, M. V. *Bounds for reliability of consecutive-k-within-m-out-of-n systems.* IEEE Transactions on Reliability, 42, 156–160, 1993

[14] RIORDAN, J. *An Introduction to Combinatorial Analysis.* John Wiley and Sons, New York, 1958

[15] JOHNSON, N. L., KOTZ, S. *Urn Models and their Applications.* John Wiley and Sons, New York, 1977

[16] GIBBONS J. D. *Nonparametric Statistical Inference.* Second edition, Marcel Dekker, New York, 1985

Funkce v programu Wolfram Mathematica

Zde jsou uvedeny funkce pro výpočet přesné hodnoty pravděpodobnosti výskytu první iterace délky k v posloupnosti n Bernoulliových pokusů, funkce pro výpočet aproximací této přesné hodnoty a jejich horních a dolních mezí a funkce pro tvorbu grafů obsažených v této práci.

```
<< PlotLegends‘
(*pro tvorbu legend grafů je nutné zavolat tento balík*)
p :=
q := 1-p
k :=

A = SparseArray[{{i_, 1} /; i < k + 1 -> q, {i_, j_} /;
i + j == 2 i + 1 -> p,
{k + 1, k + 1} -> 1}, {k + 1, k + 1}, 0];
(*matice pravděpodobnosti předchodu*)
l := UnitVector[k + 1, 1]
r := UnitVector[k + 1, k]
Exact[n_] := p*l.MatrixPower[A, IntegerPart[n] - 1].r
(*funkce pro výpočet přesné hodnoty pravděpodobnosti
prvního výskytu
iterace délky k po n krocích*)

CS12[n_] := q*p^k*(Exp[-(IntegerPart[n] - 2 k)*p^k])
(*první a druhá Chen-Steinova aproximace*)

CS3[n_] := q*p^k*(Exp[-(IntegerPart[n] - 2 k)*q*p^k])
(*třetí Chen-Steinova aproximace*)

ev[x_] := SetAccuracy[1 + q*p^k*x^(k + 1),1000]
(*pomocná funkce pro výpočet kořene z ve Fellerově aproximaci*)

z := SetAccuracy[Nest[ev, 1, 500], 10]
Feller[n_] := q*p^k*(1 - p*z)/((k + 1 - k*z)*q)*1/z^(n - k)
(*Fellerova aproximace*)

M[n_] := q*p^k*((1 - p^k)^(n - 2 k) + (1 - q*p^k)^(n - 2 k))/2
(*aproximace M*)

FEUB[n_] := Feller[n] + q*p^k*(2 (k - 1)*p^(n - k))
/(k + 1 + k*p^(-1))*q
FELB[n_] := Feller[n] - q*p^k*(2 (k - 1)*p^(n - k))
/(k + 1 + k*p^(-1))*q
(*horní a dolní Fellerovy meze*)
```

```
FEUB2[n_] :=  Feller[n] +  q*p^k*(2 (k - 1)*p)/(k*q (1 + p))
FELB2[n_] := Feller[n] - q*p^k*(2 (k - 1)*p)/(k*q (1 + p))
(*horní a dolní druhé Fellerovy meze*)

CS1UB[n_] := CS12[n] + q*p^k*((2 k - 1)*p^k + 2 (k - 1)*p)
CS1LB[n_] := CS12[n] - q*p^k*((2 k - 1)*p^k + 2 (k - 1)*p)
(*horní a dolní první Chen-Steinova mez*)

CS2UB[n_] := CS12[n] + q*p^k*((2 p)/
q*(1 - p^(k - 1)) - (2 k - 3)*p^k)
CS2LB[n_] := CS12[n] - q*p^k*((2 p)/
q*(1 - p^(k - 1)) - (2 k - 3)*p^k)
(*horní a dolní druhá Chen-Steinova mez*)

CS3UB[n_] := CS3[n] + q*p^k*((2 k*q + 1) p^k)
CS3LB[n_] := CS3[n] - q*p^k*((2 k*q + 1) p^k)
(*horní a dolní třetí Chen-Steinova mez*)

HM[n_] := q*p^k*(1 - q*p^k)^(n - 2 k)
DM[n_] := q*p^k*(1 - p^k)^(n - 2 k)
 (*horní a dolní meze HM a DM*)

ShowLegend[ DiscretePlot[{Abs[Exact[n] - CS3[n]],
Abs[Exact[n] - M[n]],
 Abs[Exact[n] - CS12[n]], Abs[Exact[n] - Feller[n]]},
{n, 2 k + 1, 120}, PlotRange -> {{0, Automatic},
 {0, 1.3*Max[Abs[Exact[2 k + 1] - CS3[2 k + 1]],
 Abs[Exact[2 k + 1] - M[2 k + 1]],
 Abs[Exact[2 k + 1] - Feller[2 k + 1]],
  *Abs[Exact[2 k + 1] - Abs[CS12[2 k + 1]]]]}},
  PlotStyle -> {Red, Blue, Magenta, Black},
  AxesLabel -> {n, odchylka}], {{{Graphics[{Red, Point[{0, 0}]}],
    "3. C-S aproximace"}, {Graphics[{Blue, Point[{0, 0}]}],
    "aproximace M"}, {Graphics[{Magenta, Point[{0, 0}]}],
    "1. a 2. C-S aproximace"}, {Graphics[{Black, Point[{0, 0}]}],
    "Fellerova aproximace"}}, LegendPosition -> {0.4, -0.2},
  LegendTextSpace -> 7}]
(*graf zaznamenávající absolutní odchylky aproximací od
přesné hodnoty pravděpodobnosti,
že k prvnímu výskytu iterace délky k dojde po n krocích *)

ShowLegend[
 DiscretePlot[{Exact[n], FEUB[n], FELB[n], CS1UB[n], CS1LB[n],
   CS2LB[n], CS2UB[n], CS3UB[n], CS3LB[n], DM[n], HM[n]},
{n, 2 k + 1,
    40}, PlotRange -> {{0, Automatic}, {0, 1.1*CS1UB[2 k + 1]}},
  PlotStyle -> {Red, Black, Black, Green, Green,
Orange, Orange, Blue,
```

```
     Blue, Magenta, Magenta},
  AxesLabel -> {n, meze}], {{{Graphics[{Red, Point[{0, 0}]}],
    " presna hodnota"}, {Graphics[{Green, Point[{0, 0}]}],
    " 1. C-S meze"}, {Graphics[{Blue, Point[{0, 0}]}], \!\(\*
TagBox[
StyleBox["\"\< 2. C-S meze\>\"",
ShowSpecialCharacters->False,
ShowStringCharacters->True,
NumberMarks->True],
FullForm]\)}, {Graphics[{Magenta, Point[{0, 0}]}],
    " meze HM a DM"}, {Graphics[{Orange, Point[{0, 0}]}],
    " 3. C-S meze"}, {Graphics[{Black, Point[{0, 0}]}],
    " Fellerovy meze"}}, LegendPosition -> {0.3, -0.18},
  LegendTextSpace -> 12, LegendBorderSpace -> 0.5}]
(*graf zaznamenávající přesnou hodnotu
 pravděpodobnosti, že k prvnímu
výskytu iterace délky k dojde po n krocích a hodnoty horních a
dolních mezích jednotlivých aproximací*)

ShowLegend[
 DiscretePlot[{Exact[n], FEUB[n], FELB[n], FEUB2[n], FELB2[n]},
{n, 2 k + 1, 30}, PlotRange -> {{0, Automatic}, Automatic},
  PlotStyle -> {Red, Blue, Blue, Green, Green},
  AxesLabel -> {n, meze}], {{{Graphics[{Red, Point[{0, 0}]}],
    "presna hodnota"}, {Graphics[{Blue, Point[{0, 0}]}],
    "Fellerovy meze"}, {Graphics[{Green, Point[{0, 0}]}],
    "2. Fellerovy meze"}}, LegendPosition -> {0.3, 0.15},
  LegendTextSpace -> 7}]
(*graf srovnávající první a druhé Fellerovy meze*)
```