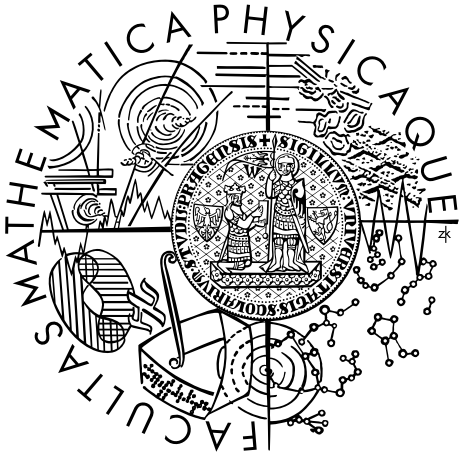
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Adam Láf

Diskrétní skenovací statistika

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Zbyněk Pawlas, Ph.D.
Studijní program: Matematika
Studijní obor: Obecná matematika

Praha 2016

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

Název práce: Diskrétní skenovací statistika

Autor: Adam Láf

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Zbyněk Pawlas, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Pro náhodný výběr z celočíselného rozdělení je zavedena diskrétní skenovací statistika jako maximum z klouzavých součtů daného počtu po sobě jdoucích náhodných veličin. Tato práce seznámí čtenáře s několika postupy pro odhad rozdělení diskrétní skenovací statistiky, přičemž uvedené aproximace následně zhodnotí na konkrétních případech. Zaměří se převážně na náhodné výběry z alternativního rozdělení, pro které bude uveden i návod pro výpočet přesných výsledků. Zmíněny budou i souvislosti s narozeninovým problémem a s odhadováním největšího počtu po sobě jdoucích úspěchů v řadě bernoulliovských pokusů.

Klíčová slova: diskrétní skenovací statistika, aproximace, pravděpodobnostní rozdělení, narozeninový problém

Title: Discrete scan statistics

Author: Adam Láf

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Zbyněk Pawlas, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The discrete scan statistic is defined as the maximum of moving sums of a given number of consecutive observations in a sequence of i.i.d. integer valued random variables. This thesis introduces various ways to approximate the distribution of the discrete scan statistic. These approximations are evaluated based on enumerations in specific cases. The main focus is on random variables with Bernoulli distribution, the only case where exact results for the distribution of the discrete scan statistic are available. Some connections with well-known problems as the birthday problem and the longest success run in a sequence of Bernoulli trials are also discussed.

Keywords: discrete scan statistics, approximation, probability distribution, birthday problem

Rád bych poděkoval vedoucímu této práce doc. RNDr. Zbyňku Pawlasovi, Ph.D. za vstřícnost, ochotu a veškeré připomínky, které pomohly k jejímu vypracování.

Obsah

Úvod

Představme si situaci, kdy nám ve 100 hodech mincí padlo jen 8 panen. To nevypadá na spravedlivou minci. Ale 5 z těch panen se nalézalo v 10 hodech v řadě. Je to hodně, nebo málo? Změnil se během hodů styl házení, nebo hraje roli nějaký jiný vnější faktor? A kolik nejvíce panen v řadě můžeme očekávat, aniž bychom začali pochybovat o férovosti dané mince? Pomocí diskrétní skenovací statistiky můžeme najít odpovědi.

Mějme pozorování na úsečce (časovém úseku) či jen v konečně mnoha seřazených bodech. Skenovací statistika je maximum součtu pozorování v libovolném časovém intervalu o pevně dané délce. Její základní použití je v testování nulové hypotézy o rovnoměrnosti pozorování vůči jejich shlukující alternativě. Využití nalézá v nepřeberném množství oborů, od analýzy struktur DNA a bílkovin přes meteorologii, sociologii, archeologii, telekominukace, teorii spolehlivosti a kontroly kvality až k práci s minovými poli, radarem a dál.

Tato práce se zabývá diskrétním případem, tj. naše pozorování nejsou rozložená na přímce, ale v diskrétních po sobě jdoucích bodech. Seznámíme čtenáře s různými aproximacemi rozdělení skenovací statistiky, naznačíme jejich odvození a provedeme jejich porovnání. Hlavní pozornost věnujeme pozorováním s alternativním rozdělením, což je jediný případ, pro který lze nalézt odvozené i přesné výsledky.

V kapitole 1 uvedeme přesnou definici diskrétní skenovací statistiky společně s dalšími poznatky potřebnými pro její aproximace.

Kapitola 2 se zabývá nepodmíněným případem, tj. situacemi, kdy předpokládáme rozdělení náhodných pozorování a odhadujeme pravděpodobnosti pro hodnoty, kterých skenovací statistika může nabývat.

Ve 3. kapitole prozkoumáme podmíněný případ, ve kterém pravděpodobnosti pro hodnoty skenovací statistiky podmiňujeme součtem všech naměřených pozorování. Tato oblast se využívá zejména pro alternativní rozdělení, kde lze situaci interpretovat jako hledání rozdělení nejvyššího počtu úspěchů v okně o dané délce v řadě pokusů se známým počtem úspěchů.

Nakonec v kapitole 4 vyčíslíme dříve uvedené aproximace, zhodnotíme jejich přesnost a uvedeme některé spojitosti se známými příklady, např. zobecněním narozeninového problému.

1. Základní pojmy a vzorce

1.1 Základní definice

Definice 1. *Nechť $X_1, X_2, \ldots, X_N$ jsou nezávislé stejně rozdělené nezáporné celočíselné náhodné veličiny. Pak pro přirozené číslo m, $2 \leq m \leq N$, definujeme* diskrétní skenovací statistiku *jako*

$$S_m = \max\{X_i + \cdots + X_{i+m-1}; 1 \leq i \leq N - m + 1\}. \tag{1.1}$$

Nezápornost není tak omezující, jestliže jsou X_i omezené zdola (popř. shora), neboť pakliže je nejmenší možná hodnota X_i rovna $-c$, $c \in \mathbb{N}$, můžeme pro všechna pozorování místo X_i uvažovat $X_i + c$.

Diskrétní skenovací statistika se užívá pro testování nulové hypotézy, dle které mají všechna pozorování stejné rozdělení, proti alternativě, že hypotéza neplatí. Pro stejně rozdělené nezávislé náhodné veličiny zamítáme nulovou hypotézu tehdy, když $S_m \geq k$, kde k je určeno na základě zvolené pravděpodobnosti pro chybu prvního druhu. To znamená, že je třeba znát rozdělení S_m, tj. pro celé číslo $k \geq 2$ nás zajímá

$$\mathsf{P}(S_m \geq k) = \mathsf{P}(k; m, N). \tag{1.2}$$

V podmíněném případě pak zkoumáme rozdělení S_m podmíněné znalostí součtu všech pozorování:

$$\mathsf{P}\left(S_m \geq k \middle| \sum_{i=1}^{N} X_i = a\right) = \mathsf{P}(k; m, N, a). \tag{1.3}$$

Dále se nám ještě bude hodit pro celé číslo j, $m \leq j$, zavést

$$Q_j = 1 - \mathsf{P}(k; m, j), \tag{1.4}$$

$$Q_j^*(a) = 1 - \mathsf{P}(k; m, j, a). \tag{1.5}$$

Přesné výsledky pro výrazy 1.2 a 1.3 lze najít pouze pro případ s alternativním rozdělením, a i pro něj je lze v rozumném čase počítat jen pro velmi omezenou množinu parametrů. Proto je zapotřebí počitatelných aproximací, které budou uvedeny v kapitolách 2 a 3. Tyto aproximace však stále využívají přesných výsledků pro speciální parametry (zpravidla Q_j či $Q_j^*(a)$ pro $j \in \{m, 2m, 3m\}$). Glaz a Balakrishnan (1999) i Glaz, Naus a Wallenstein (2001) získávají pro jiná než alternativní rozdělení potřebná Q_j či $Q_j^*(a)$ pomocí simulací.

Ještě by bylo dobré zmínit, že naše definice 1.1 uvažuje tzv. *lineární* případ, přičemž rozlišujeme ještě *kruhový* případ, kde za poslední pozorování zařadíme první a skenovací okno projede až do své startovní pozice. Takovým případem se zde zabývat nebudeme, avšak Glaz a Balakrishnan (1999, str. 10) říkají, že na něj lze rozšířit metody z lineárního případu. Rovněž je možné definovat skenovací statistiku i pro více rozměrů, kde se však situace značně komplikuje a není to náplní této práce (přičemž už pro 3 dimenze v podstatě ani žádné výsledky odvozené nejsou (Glaz a Balakrishnan, 1999, str. 12)).

1.2 Přesné výsledky pro alternativní rozdělení

Mějme $X_1, X_2, \ldots, X_N$ posloupnost pokusů, kde každý může skončit úspěchem ($X_i = 1$) či neúspěchem ($X_i = 0$). Předpokládejme, že jsme napozorovali a úspěchů a $N - a$ neúspěchů, přičemž všechna pořadí a úspěchů a $N - a$ neúspěchů jsou stejně pravděpodobná. Zajímá nás $\mathsf{P}(k; m, N, a)$. Saperstein (1972) uvádí odvození přesného vzorce pro $\mathsf{P}(k; m, N, a)$, pokud $k > a/2$. Podmínka $k > a/2$ je velmi omezující, avšak Glaz a kol. (2001, Theorem 12.1[1]) uvádějí následující větu (včetně důkazu) pro $\mathsf{P}(k; m, N, a)$, pokud $N = ml$, $2 \leq l \in \mathbb{N}$:

Věta 1. *Nechť $N, m, l, k, a \in \mathbb{N}$, $N/m = l \geq 2$, $2 \leq k \leq a \leq N$, $k \leq m \leq N$. Dále Θ_k označuje množinu všech uspořádaných l-tic $(n_1, n_2, \ldots, n_l)$, pro které platí $k > n_i \in \mathbb{N}_0$ pro všechna $i \in \{1, 2, \ldots, l\}$ a zároveň $\sum_{i=1}^{l} n_i = a$. Pak:*

$$1 - \mathsf{P}(k; m, N, a) = \frac{(m!)^l}{\binom{N}{a}} \sum \det D_{\mathbf{n}}, \tag{1.6}$$

kde suma je přes množinu Θ_k a $D_{\mathbf{n}}$ značí $l \times l$ matici s prvky

$$d_{rs} = \frac{1}{c_{rs}!(m - c_{rs})!}, \quad r = 1, 2, \ldots, l, \quad s = 1, 2, \ldots, l,$$

kde

$$c_{rs} = \begin{cases} (s - r)k + n_r - \sum_{i=r}^{s-1} n_i, & r < s, \\ (s - r)k + \sum_{i=s}^{r} n_i, & r \geq s, \end{cases}$$

přičemž $d_{rs} = 0$ pro $c_{rs} < 0$ nebo $m - c_{rs} < 0$.

K této větě patří i následující důsledek (Glaz a kol., 2001, Corollary 12.1):

Důsledek. Nechť $k > a/2$ a pro celé číslo l platí $N/m = l \geq 2$. Potom

$$\mathsf{P}(k; m, N, a) = \frac{2\sum_{s=k}^{a} \binom{m}{s}\binom{N-m}{a-s}}{\binom{N}{a}} + (lk - a - 1)\frac{\binom{m}{k}\binom{N-m}{a-k}}{\binom{N}{a}}.$$

Důsledek dává lépe počitatelný vzorec než Saperstein (1972), ačkoli musí platit $N = ml$. Nicméně pro m znatelně menší než N není podmínka $N = ml$ až tak omezující a v našich výpočtech v kapitole 4 se jinými případy zpravidla zabývat nebudeme. Tím máme návod na počítání $Q_j^*(a)$ pro alternativní rozdělení, kde $j = ml$, $2 \leq l \in \mathbb{N}$, ačkoli výpočet vzorce 1.6 bývá i pro nevelké parametry časově náročný.

Pro nepodmíněný případ (tj. $X_1, X_2, \ldots, X_N$ náhodný výběr z alternativního rozdělení s parametrem $p \in (0, 1)$) pak můžeme vhodným způsobem zprůměrovat výsledky pro podmíněné případy, což přesně provádí Naus (1982) pro nalezení Q_{2m} a Q_{3m}. Jeho zjednodušené vzorce tu nebudou uvedeny, nicméně zde provádíme onen výpočet Q_{rm}, $r \in \mathbb{N}$, pomocí podmíněných případů:

$$1 - \mathsf{P}(k; m, rm) = 1 - \sum_{a=k}^{rm} \mathsf{P}(k; m, rm, a)\binom{rm}{a} p^a (1 - p)^{rm-a}.$$

[1] Upozorněme na přepis ve znění v knize, místo $+\sum_{i=1}^{s-1} n_i$ má být $-\sum_{i=r}^{s-1} n_i$.

Nakonec k aproximacím uvedeným v kapitole 2 je potřeba ještě znát vzorec pro Q_m, který ze znalosti, že $\sum_{i=1}^{m} X_i$ má binomické rozdělení, určíme ihned jako

$$Q_m = \sum_{i=0}^{k-1} \binom{m}{i} p^i (1-p)^{m-i},$$

a dále vzorec pro Q_{m+1}, který v užité literatuře nenacházíme, ale lze ho snadno odvodit:

$$\begin{aligned}
Q_{m+1} &= 1 - \mathsf{P}(k; m, m+1) = 1 - \mathsf{P}\left(\sum_{i=1}^{m} X_i \geq k \vee \sum_{i=2}^{m+1} X_i \geq k\right) = \\
&= 1 - \mathsf{P}\left(\sum_{i=1}^{m} X_i \geq k\right) - \mathsf{P}\left(\sum_{i=1}^{m} X_i < k \wedge \sum_{i=2}^{m+1} X_i \geq k\right) = \\
&= Q_m - \mathsf{P}\left(X_1 = 0 \wedge \sum_{i=2}^{m} X_i = k-1 \wedge X_{m+1} = 1\right) = \\
&= Q_m - (1-p)\binom{m-1}{k-1} p^{k-1}(1-p)^{m-1-k+1} p = \\
&= Q_m - \binom{m-1}{k-1} p^k (1-p)^{m-k+1}.
\end{aligned}$$

2. Nepodmíněný případ

Mějme $X_1, \ldots, X_N$ nezávislé stejně rozdělené nezáporné celočíselné náhodné veličiny. Hledáme rozdělení diskrétní skenovací statistiky S_m z definice 1.1, tj. snažíme se najít $\mathsf{P}(k; m, N) = \mathsf{P}(S_m \geq k)$ (definice 1.2). V této kapitole ukážeme aproximace, jež prezentují Glaz a Balakrishnan (1999), a v kapitole 4 pak provedeme jejich porovnání na náhodných výběrech z určitých rozdělení.

Připomeňme zavedení Q_j jako $Q_j = 1 - \mathsf{P}(k; m, j)$ (definice 1.4) a dále uvažujme v této kapitole všechny proměnné jako celočíselné, nebude-li řečeno jinak.

2.1 Aproximace součinového typu

Uvažujme zjednodušený případ $N = lm$, $l \geq 2$. Pro $1 \leq i \leq l-1$ označme jevy B_i následovně:

$$B_i = \bigcap_{j=1}^{m+1} \left(X_{(i-1)m+j} + \cdots + X_{im+j-1} \leq k-1 \right). \tag{2.1}$$

Jev $(S_m \leq k-1)$ je roven jevu $\bigcap_{i=1}^{l-1} B_i$, z čehož můžeme získat následující aproximaci, pokud $l \geq 4$:

$$\begin{aligned}
\mathsf{P}(S_m \leq k-1) = \mathsf{P}\left(\bigcap_{i=1}^{l-1} B_i\right) &= \mathsf{P}(B_1)\,\mathsf{P}\left(\bigcap_{i=2}^{l-1} B_i \middle| B_1\right) = \\
&= \mathsf{P}(B_1) \prod_{i=2}^{l-1} \mathsf{P}\left(B_i \middle| \bigcap_{j=1}^{i-1} B_j\right) \approx \mathsf{P}(B_1) \prod_{i=2}^{l-1} \mathsf{P}(B_i | B_{i-1}) = \\
&= \mathsf{P}(B_1 \cap B_2) \prod_{i=2}^{l-2} \frac{\mathsf{P}(B_i \cap B_{i+1})}{\mathsf{P}(B_i)}.
\end{aligned}$$

Dále snadno z definice B_i vidíme, že pro $1 \leq i \leq l-2$ platí $\mathsf{P}(B_i) = Q_{2m}$ a $\mathsf{P}(B_i \cap B_{i+1}) = Q_{3m}$. Dostáváme předpis, který lze úspěšně užít při $l \geq 2$:

$$\mathsf{P}(S_m \geq k) \approx 1 - Q_{3m} \left(\frac{Q_{3m}}{Q_{2m}}\right)^{l-3}. \tag{2.2}$$

Lze se domnívat, že kdybychom v úpravách zanedbali v podmínce méně jevů, aproximace by byla přesnější, ale bylo by třeba znát Q_{rm} pro $r > 3$. Proto budeme ještě využívat jen o stupeň lepší aproximaci (za předpokladu $l \geq 3$):

$$\mathsf{P}(S_m \geq k) \approx 1 - Q_{4m} \left(\frac{Q_{4m}}{Q_{3m}}\right)^{l-4}. \tag{2.3}$$

Pakliže neplatí $N = ml$, nýbrž $N = ml - v$, $1 \leq v \leq m-1$, je doporučena následující úprava aproximace 2.2:

$$\mathsf{P}(S_m \geq k) \approx 1 - Q_{3m-v} \left(\frac{Q_{3m}}{Q_{2m}}\right)^{l-3}.$$

Vzhledem k obtížím spojeným s určením Q_{3m-v} a podobným výsledkům, jako má odhad 2.2, nebude poslední aproximace dále zkoumána.

2.2 Poissonovské aproximace

Součet n nezávislých náhodných veličin s alternativním rozdělením o stejném parametru $p \in (0,1)$ má binomické rozdělení s parametry n a p. Platí, že pro n jdoucí do nekonečna a p jdoucí k nule takovým způsobem, že np běží k nějaké reálné konstantě $\lambda > 0$, se binomické rozdělení blíží Poissonovu rozdělení s parametrem λ. Z toho vychází základní myšlenka poissonovských aproximací, přestože náhodné veličiny, se kterými se pracuje, často nesplňují podmínky nezávislosti či stejného rozdělení. Trochu podrobněji tuto myšlenku rozvádí Glaz a kol. (2001, podkapitola 7.6), nás budou ale více zajímat dva postupy vedoucí k aproximaci $\mathsf{P}(S_m \geq k)$, jež uvádí Glaz a Balakrishnan (1999).

Pro první postup označme při $1 \leq j \leq N - m + 1$ náhodnou veličinu

$$I_j = \begin{cases} 1, & \text{pokud } X_j + \cdots + X_{j+m-1} \geq k, \\ 0, & \text{jinak.} \end{cases} \tag{2.4}$$

Tudíž

$$\mathsf{P}(S_m \leq k-1) = \mathsf{P}\left(\sum_{j=1}^{N-m+1} I_j = 0\right).$$

Nyní tedy budeme předpokládat, že rozdělení $\sum_{j=1}^{N-m+1} I_j$ je blízké Poissonovu rozdělení s parametrem λ, který určíme z rovnosti středních hodnot, tj.

$$\lambda = \mathsf{E}\left(\sum_{j=1}^{N-m+1} I_j\right) = \sum_{j=1}^{N-m+1} \mathsf{E}\,(I_j) = (N-m+1)(1-Q_m).$$

Pro Poissonovo rozdělení platí, že pravděpodobnost nastání nuly je $e^{-\lambda}$, z čehož získáváme aproximaci

$$\mathsf{P}(S_m \geq k) \approx 1 - e^{-\lambda}. \tag{2.5}$$

Druhý postup využívá jevy B_i z definice 2.1 (tedy chceme $N = ml$, $l \geq 2$). Zaveďme pro $1 \leq i \leq l-1$ náhodnou veličinu

$$H_i = \begin{cases} 1, & \text{pokud } B_i \text{ nenastal,} \\ 0, & \text{jinak.} \end{cases} \tag{2.6}$$

Opět

$$\mathsf{P}(S_m \leq k-1) = \mathsf{P}\left(\sum_{i=1}^{l-1} H_i = 0\right).$$

Obdobně jako výše i zde odhadujeme rozdělení $\sum_{i=1}^{l-1} H_i$ pomocí Poissonova rozdělení se střední hodnotou

$$\lambda_2 = \mathsf{E}\left(\sum_{j=1}^{l-1} H_i\right) = (l-1)(1-Q_{2m})$$

a získaná aproximace je dána předpisem

$$\mathsf{P}(S_m \geq k) \approx 1 - e^{-\lambda_2}. \tag{2.7}$$

Avšak vzhledem k definici náhodných veličin I_j a H_i není překvapivé, že jevy $(I_j = 1)$ a $(H_i = 1)$ mají sklon se shlukovat. Glaz a Balakrishnan (1999, str. 31) dávají návod, jak toto zohlednit a dané aproximace vylepšit. Postup budeme ilustrovat na aproximaci 2.7:

Nechť $1 \leq i \leq l-1$ a upravme náhodnou veličinu H_i následovně (B_0 definujme jako celý prostor):

$$H_i^* = \begin{cases} 1, & \text{pokud } B_i^c \cap B_{i-1} \text{ nastal,} \\ 0, & \text{jinak.} \end{cases}$$

Tím jsme předešli přílišnému shlukování jevů $(H_i^* = 1)$, ale stále platí

$$\mathsf{P}(S_m \leq k-1) = \mathsf{P}\left(\sum_{i=1}^{l-1} H_i^* = 0\right).$$

Spočítáme střední hodnotu λ_2^*:

$$\lambda_2^* = \mathsf{E}\left(\sum_{i=1}^{l-1} H_i^*\right) = \sum_{i=1}^{l-1} \mathsf{E}\left(H_i^*\right) = \mathsf{P}(H_1^* = 1) + \sum_{i=2}^{l-1} \mathsf{P}\left(H_i^* = 1\right),$$

kde

$$\mathsf{P}(H_1^* = 1) = \mathsf{P}\left[\bigcup_{j=1}^{m+1}\left(\sum_{t=j}^{j+m-1} X_t \geq k\right)\right] = \mathsf{P}(k; m, 2m) = 1 - Q_{2m}$$

a (pokud platí $l \geq 3$)

$$\begin{aligned}
\mathsf{P}(H_2^* = 1) &= \mathsf{P}\left[\bigcup_{j=1}^{m+1}\left(\sum_{t=j+m}^{j+2m-1} X_t \geq k\right) \cap \bigcap_{j=1}^{m+1}\left(\sum_{t=j}^{j+m-1} X_t \leq k-1\right)\right] = \\
&= \mathsf{P}\left[\bigcup_{j=1}^{2m+1}\left(\sum_{t=j}^{j+m-1} X_t \geq k\right) \cap \bigcap_{j=1}^{m+1}\left(\sum_{t=j}^{j+m-1} X_t \leq k-1\right)\right] = \\
&= \mathsf{P}\left[\bigcup_{j=1}^{2m+1}\left(\sum_{t=j}^{j+m-1} X_t \geq k\right) \setminus \bigcup_{j=1}^{m+1}\left(\sum_{t=j}^{j+m-1} X_t \geq k\right)\right] = \\
&= \mathsf{P}\left[\bigcup_{j=1}^{2m+1}\left(\sum_{t=j}^{j+m-1} X_t \geq k\right)\right] - \mathsf{P}\left[\bigcup_{j=1}^{m+1}\left(\sum_{t=j}^{j+m-1} X_t \geq k\right)\right] = \\
&= \mathsf{P}(k; m, 3m) - \mathsf{P}(k; m, 2m) = Q_{2m} - Q_{3m}.
\end{aligned}$$

Celkem dostáváme

$$\lambda_2^* = 1 - Q_{2m} + (l-2)(Q_{2m} - Q_{3m})$$

a naše upravená aproximace má tvar

$$\mathsf{P}(S_m \geq k) \approx 1 - e^{-\lambda_2^*}. \tag{2.8}$$

2.3 Složené poissonovské aproximace

Nechť $Y_1, Y_2, \ldots$ je posloupnost nezávislých stejně rozdělených nezáporných celočíselných náhodných veličin. Dále nechť L je náhodná veličina s Poissonovým rozdělením nezávislá s $\{Y_i\}_{i=1}^{\infty}$. Pak $\sum_{i=1}^{L} Y_i$ má složené Poissonovo rozdělení a je skoro jistě rovna $\sum_{j=1}^{\infty} jM_j$, kde M_j jsou určité nezávislé náhodné veličiny s Poissonovým rozdělením se středními hodnotami λ_j (Glaz a kol., 2001, str. 111). Pokud se zajímáme o pravděpodobnost jevu $(\sum_{i=1}^{L} Y_i = 0)$, tj. $(\sum_{j=1}^{\infty} jM_j = 0)$, pak platí:

$$\mathsf{P}\left(\sum_{j=1}^{\infty} jM_j = 0\right) = \mathsf{P}(M_j = 0 \ \ \forall j \in \mathbb{N}) = \prod_{j=1}^{\infty} \mathsf{P}(M_j = 0) = \exp\left(-\sum_{j=1}^{\infty} \lambda_j\right).$$

Odtud získáváme jakýsi základ pro složené poissonovské aproximace, ačkoli, stejně jako u poissonovských, nejsou mnohé předpoklady obvykle splněny. V praxi si dále vystačíme jen s několika prvními λ_j, které též musíme odhadovat. Uvádíme zde dvě aproximace rozdělení S_m, které Glaz a Balakrishnan (1999) označují jako dosahující dobrých výsledků.

První z aproximací pracuje s náhodnými veličinami I_j z definice 2.4, kde se snaží $\sum_{j=1}^{N-m+1} I_j$ odhadnout pomocí $\sum_{j=1}^{N-m+1} jM_j$, M_j obdobně jako výše. Při zavedení

$$p = \frac{\mathsf{P}(I_1 = 1 \wedge I_2 = 1)}{\mathsf{P}(I_1 = 1)} = \frac{1 - 2Q_m + Q_{m+1}}{1 - Q_m}$$

pak spočítáme λ_j ze vzorců

$$\begin{aligned}
\lambda_j &= (N - m + 1)(1 - Q_m)(1 - p)^2 p^{j-1}, && j = 1, \ldots, m-1, \\
\lambda_j &= \frac{(N - m + 1)(1 - Q_m)}{j} \left[2(1 - p)p^{j-1} + (2m - j - 2)(1 - p)^2 p^{j-1}\right], \\
& && j = m, \ldots, 2m-2, \\
\lambda_{2m-1} &= \frac{(N - m + 1)(1 - Q_m)p^{2m-2}}{2m - 1}.
\end{aligned}$$

Výsledná aproximace je dána vztahem

$$\mathsf{P}(S_m \geq k) \approx 1 - \exp\left(-\sum_{j=1}^{2m-1} \lambda_j\right). \tag{2.9}$$

U druhé aproximace požadujeme $N = ml$, $l \geq 4$. Má zdánlivě jednoduchý následující předpis:

$$\mathsf{P}(S_m \geq k) \approx 1 - \exp\left(-\sum_{j=1}^{3} \lambda_j^*\right), \tag{2.10}$$

kde pro $j = 1, 2, 3$ obdržíme λ_j^* jako

$$\lambda_j^* = \frac{1}{j} \mathsf{P}(B_1^c)[2\pi_{1,j} + (l - 3)\pi_{2,j}],$$

přičemž jev B_1 je z definice 2.1 (tedy platí $\mathsf{P}(B_1^c) = 1 - Q_{2m}$) a

$$\pi_{1,j} = \mathsf{P}(H_2 = j - 1 | H_1 = 1),$$
$$\pi_{2,j} = \mathsf{P}(H_1 + H_3 = j - 1 | H_2 = 1),$$

kde H_i jsou náhodné veličiny z definice 2.6. Glaz a Balakrishnan (1999, str. 35[1]) dávají hodnoty pro $\pi_{1,j}$ a $\pi_{2,j}$ pomocí Q_{2m}, Q_{3m} a Q_{4m}, po pár úpravách lze získat snadno počitatelný vzorec

$$\sum_{j=1}^{3} \lambda_j^* = 1 - Q_{3m} + \frac{(l-3)(1 - Q_{3m} - Q_{4m} + Q_{2m}^2)}{3}.$$

2.4 Střední hodnota a rozptyl

Ještě se nabízí otázka, jakou hodnotu skenovací statistiky bychom vlastně mohli očekávat. Proto pro S_m ještě vypočteme dvě základní charakteristiky náhodné veličiny: střední hodnotu a rozptyl. Střední hodnotu určíme snadno jako

$$\begin{aligned} \mathsf{E}\,(S_m) &= \sum_{k=1}^{m} k\,\mathsf{P}(S_m = k) = \sum_{k=1}^{m} k\,\mathsf{P}(S_m \geq k) - \sum_{k=1}^{m-1} k\,\mathsf{P}(S_m \geq k+1) = \\ &= \sum_{k=1}^{m} k\,\mathsf{P}(S_m \geq k) - \sum_{k=2}^{m} (k-1)\,\mathsf{P}(S_m \geq k) = \sum_{k=1}^{m} \mathsf{P}(S_m \geq k). \end{aligned} \tag{2.11}$$

K odvození rozptylu nejprve podobně spočtěme $\mathsf{E}\,(S_m^2)$:

$$\begin{aligned} \mathsf{E}\,(S_m^2) &= \sum_{k=1}^{m} k^2\,\mathsf{P}(S_m = k) = \sum_{k=1}^{m} k^2\,\mathsf{P}(S_m \geq k) - \sum_{k=1}^{m-1} k^2\,\mathsf{P}(S_m \geq k+1) = \\ &= \sum_{k=1}^{m} k^2\,\mathsf{P}(S_m \geq k) - \sum_{k=2}^{m} (k-1)^2\,\mathsf{P}(S_m \geq k) = \sum_{k=1}^{m} (2k-1)\,\mathsf{P}(S_m \geq k). \end{aligned}$$

Vyjde nám rovnost

$$\begin{aligned} \mathrm{var}(S_m) &= \mathsf{E}\,(S_m^2) - \mathsf{E}^{\,2}(S_m) = \sum_{k=1}^{m} (2k-1)\,\mathsf{P}(S_m \geq k) - \mathsf{E}^{\,2}(S_m) = \\ &= 2\sum_{k=1}^{m} k\,\mathsf{P}(S_m \geq k) - \mathsf{E}\,(S_m)[1 + \mathsf{E}\,(S_m)]. \end{aligned} \tag{2.12}$$

Vidíme, že obě charakteristiky dostaneme ihned ze znalosti $\mathsf{P}(S_m \geq k)$, tudíž je dokážeme snadno odhadnout za pomoci již uvedených aproximací.

[1] Ve vzorci pro $\pi_{2,3}$ je však v knize zjevný přepis, výraz $-3q_{2m}^2$ má být pouze $-3q_{2m}$.

3. Podmíněný případ

Uvažujme náhodný výběr $X_1, X_2, \ldots, X_N$ z alternativního rozdělení. Dále nechť víme, že z oněch N náhodných veličin jich a nabývá hodnoty 1 (úspěch) a $N - a$ hodnoty 0 (neúspěch). Za těchto předpokladů platí, že každá permutace našich a úspěchů a $N - a$ neúspěchů je stejně pravděpodobná. Zajímáme se o $\mathsf{P}(S_m \geq k | \sum_{i=1}^{N} X_i = a) = \mathsf{P}(k; m, N, a)$ (viz definice 1.3 a 1.1).

Naše aproximace budou vycházet ze stejných postupů jako pro nepodmíněný případ. Avšak $Q_j^*(a)$ z definice 1.5 není dostačující, neboť my bychom potřebovali znát pravděpodobnosti, že na úseku délky j nenabývá S_m hodnoty k či více za předpokladu, že a je součet všech N pozorování (nikoli pouze pozorování na onom úseku délky j). Ale můžeme spočítat, jaká je pravděpodobnost, že právě i úspěchů z a se nachází na souvislém úseku délky j, dále vzniklý výsledek přenásobit $Q_j^*(i)$ a sečíst přes všechna smysluplná i, čímž pro $r = 2, 3, 4$ dostaneme:

$$q_{rm}(a) = \mathsf{P}(S_m \leq k-1 \text{ na úseku délky } rm \text{ při } a \text{ úspěších v } N \text{ pokusech}) = \\ = \sum_{i=0}^{\min(rk-r,a)} Q_{rm}^*(i) \frac{\binom{rm}{i}\binom{N-rm}{a-i}}{\binom{N}{a}},$$

pokud $rm \leq N$. Tedy díky větě 1 dokážeme $q_{rm}(a)$ vyčíslit. Pro jiná rozdělení bychom $q_{rm}(a)$ mohli zadefinovat na základě stejné myšlenky, ačkoli k přesnému výpočtu by nám chyběl návod na spočtení příslušných $Q_{rm}^*(i)$.

V této kapitole představíme aproximace, jež uvádí Glaz a Balakrishnan (1999) a Glaz a kol. (2001) pro podmíněný případ za předpokladů uvedených výše (tj. alternativního rozdělení), a v následující kapitole je právě na náhodných výběrech z alternativního rozdělení zhodnotíme. Jelikož ale tyto aproximace vychází z obdobných postupů jako v minulé kapitole, lze je (při patřičně upraveném předpisu pro $q_{rm}(a)$) úspěšně použít i pro $X_1, X_2, \ldots, X_N$ nezávislé stejně rozdělené nezáporné celočíselné náhodné veličiny.

Není-li řečeno jinak, opět v rámci kapitoly uvažujme všechny proměnné jako celočíselné.

3.1 Aproximace součinového typu

Nechť $N = lm$, $l \geq 2$. Pak pro $1 \leq i \leq l-1$ definujeme jevy C_i:

$$C_i = \bigcap_{j=1}^{m+1} \left(X_{(i-1)m+j} + \cdots + X_{im+j-1} \leq k-1 \right). \tag{3.1}$$

Jedná se o jevy podobné jevům B_i z definice 2.1 s tím rozdílem, že v tomto případě $X_1, X_2, \ldots, X_N$ dosahují a jedniček a $N-a$ nul. Stále však jev $(S_m \leq k-1)$ je roven jevu $\bigcap_{i=1}^{l-1} C_i$, tedy analogickým způsobem jako při aproximaci 2.2 dostáváme pro dostatečně velká l:

$$1 - \mathsf{P}(k; m, N, a) \approx \mathsf{P}(C_1 \cap C_2) \prod_{i=2}^{l-2} \frac{\mathsf{P}(C_i \cap C_{i+1})}{\mathsf{P}(C_i)},$$

případně při odhadnutí $\mathsf{P}(C_i|\bigcap_{j=1}^{i-1} C_j)$ pomocí $\mathsf{P}(C_i|C_{i-1}\cap C_{i-2})$ dostaneme

$$1-\mathsf{P}(k;m,N,a)\approx \mathsf{P}(C_1\cap C_2\cap C_3)\prod_{i=2}^{l-3}\frac{\mathsf{P}(C_i\cap C_{i+1}\cap C_{i+2})}{\mathsf{P}(C_i\cap C_{i+1})}.$$

Opět snadno z definice C_i vidíme, že pro $1\le i\le l-3$ platí $\mathsf{P}(C_i)=q_{2m}(a)$, $\mathsf{P}(C_i\cap C_{i+1})=q_{3m}(a)$ a $\mathsf{P}(C_i\cap C_{i+1}\cap C_{i+2})=q_{4m}(a)$. Celkem pro $l\ge 3$ (příp. $l\ge 4$, neboť při zavedení $q_{rm}(a)$ jsme chtěli $rm\le N$) získáváme

$$\mathsf{P}(k;m,N,a)\approx 1-q_{3m}(a)\left(\frac{q_{3m}(a)}{q_{2m}(a)}\right)^{l-3}, \tag{3.2}$$

$$\mathsf{P}(k;m,N,a)\approx 1-q_{4m}(a)\left(\frac{q_{4m}(a)}{q_{3m}(a)}\right)^{l-4}. \tag{3.3}$$

3.2 Poissonovské aproximace

Mějme jevy $C_1,\dots,C_{l-1}$ z definice 3.1 (stále uvažujeme $N=ml$, $l\ge 2$) a pro $1\le i\le l-1$ zaveďme náhodnou veličinu

$$D_i=\begin{cases}1, & \text{pokud } C_i \text{ nenastal},\\ 0, & \text{jinak.}\end{cases} \tag{3.4}$$

Stejně jako v nepodmíněném případě (aproximace 2.7) platí

$$1-\mathsf{P}(k;m,N,a)=\mathsf{P}\left(\sum_{i=1}^{l-1}D_i=0\right)$$

a poissonovská aproximace má předpis

$$\mathsf{P}(k;m,N,a)\approx 1-e^{-\lambda}, \tag{3.5}$$

kde λ získáme z rovnosti

$$\lambda=\mathsf{E}\left(\sum_{j=1}^{l-1}D_i\right)=(l-1)[1-q_{2m}(a)].$$

Nyní znovu můžeme aproximaci vylepšit postupem, jímž jsme získali aproximaci 2.8. Pro $1\le i\le l-1$ upravme náhodnou veličinu D_i takto (C_0 definujme jako celý prostor):

$$D_i^*=\begin{cases}1, & \text{pokud } C_i^c\cap C_{i-1} \text{ nastal},\\ 0, & \text{jinak.}\end{cases}$$

Platí

$$1-\mathsf{P}(k;m,N,a)=\mathsf{P}\left(\sum_{i=1}^{l-1}D_i^*=0\right),$$

střední hodnotu λ^* lze (při $l\ge 3$) spočíst jako

$$\lambda^*=\mathsf{E}\left(\sum_{i=1}^{l-1}D_i^*\right)=1-q_{2m}(a)+(l-2)[q_{2m}(a)-q_{3m}(a)]$$

a vylepšená aproximace má tvar

$$\mathsf{P}(k;m,N,a)\approx 1-e^{-\lambda^*}. \tag{3.6}$$

3.3 Složené poissonovské aproximace

V této části uvedeme aproximace 2.9 a 2.10 upravené na podmíněný případ. Pro první z nich nejprve označme

$$p(a) = \frac{1 - 2q_m(a) + q_{m+1}(a)}{1 - q_m(a)},$$

kde $q_m(a)$ spočteme jako

$$q_m(a) = \sum_{i=0}^{\min(k-1,a)} \frac{\binom{m}{i}\binom{N-m}{a-i}}{\binom{N}{a}}$$

a pro $q_{m+1}(a)$ užijeme následující vzorec (Glaz a Balakrishnan, 1999, str. 39):

$$q_{m+1}(a) = \sum_{k_1=0}^{1}\sum_{k_3=0}^{1}\sum_{k_2=0}^{\min(k-1-k_1,k-1-k_3)} \frac{\binom{1}{k_1}\binom{m-1}{k_2}\binom{1}{k_3}\binom{N-m-1}{a-k_1-k_2-k_3}}{\binom{N}{a}}.$$

Potom spočítáme λ_j ze vzorců

$$\begin{aligned}
\lambda_j &= (N-m+1)[1-q_m(a)][1-p(a)]^2 p^{j-1}(a), && j = 1, \ldots, m-1,\\
\lambda_j &= \frac{(N-m+1)[1-q_m(a)]}{j} \times \\
&\quad \times \left\{2[1-p(a)]p^{j-1}(a) + (2m-j-2)[1-p(a)]^2 p^{j-1}(a)\right\}, \\
&&& j = m, \ldots, 2m-2,\\
\lambda_{2m-1} &= \frac{(N-m+1)[1-q_m(a)]p^{2m-2}(a)}{2m-1}.
\end{aligned}$$

Dostaneme výslednou aproximaci

$$\mathsf{P}(k; m, N, a) \approx 1 - \exp\left(-\sum_{j=1}^{2m-1} \lambda_j\right). \tag{3.7}$$

Druhou aproximaci lze užít, když platí $N = lm$, $l \geq 4$. Je dána předpisem

$$\mathsf{P}(k; m, N, a) \approx 1 - \exp\left(-\sum_{j=1}^{3} \lambda_j^*\right), \tag{3.8}$$

kde pro $j = 1, 2, 3$ obdržíme λ_j^* jako

$$\lambda_j^* = \frac{1}{j}\,\mathsf{P}(C_1^c)[2\pi_{1,j}^* + (l-3)\pi_{2,j}^*],$$

přičemž jev C_1 je z definice 3.1 (tedy platí $\mathsf{P}(C_1^c) = 1 - q_{2m}(a)$) a

$$\begin{aligned}
\pi_{1,j}^* &= \mathsf{P}(D_2 = j-1 | D_1 = 1),\\
\pi_{2,j}^* &= \mathsf{P}(D_1 + D_3 = j-1 | D_2 = 1),
\end{aligned}$$

kde D_i jsou náhodné veličiny z definice 3.4. Avšak nepotřebujeme určit každé λ_j^* zvlášť, postačí nám, že

$$\sum_{j=1}^{3} \lambda_j^* = 1 - q_{3m}(a) + \frac{(l-3)[1 - q_{3m}(a) - q_{4m}(a) + q_{2m}^2(a)]}{3}.$$

3.4 Střední hodnota a rozptyl

Střední hodnotu $\mathsf{E}\left(S_m \middle| \sum_{i=1}^{N} X_i = a\right)$ (dále jen $\mathsf{E}\,(S_m; a)$) a stejně tak i rozptyl $\mathrm{var}\left(S_m \middle| \sum_{i=1}^{N} X_i = a\right)$ (označme $\mathrm{var}(S_m; a)$) získáme ze vzorců 2.11 a 2.12 nahrazením pravděpodobností $\mathsf{P}(S_m \geq k)$ pomocí $\mathsf{P}\left(S_m \geq k \middle| \sum_{i=1}^{N} X_i = a\right)$ (neboli $\mathsf{P}(k; m, N, a)$). Tedy

$$\mathsf{E}\,(S_m; a) = \sum_{k=1}^{m} \mathsf{P}(k; m, N, a),$$

$$\mathrm{var}(S_m; a) = 2\sum_{k=1}^{m} k\,\mathsf{P}(k; m, N, a) - \mathsf{E}\,(S_m; a)[1 + \mathsf{E}\,(S_m; a)].$$

4. Příklady a porovnání

4.1 Srovnání aproximací

V minulých kapitolách jsme uvedli několik aproximací pro výpočet diskrétní skenovací statistiky a pro případ výběru z alternativního rozdělení dali i návod na spočtení všeho potřebného k jejich vyčíslení. Nyní je na základě vypočítaných hodnot pro množství případů budeme porovnávat. Pro výpočty byl užit programovací jazyk R (R Core Team, 2015, verze 3.1.3), konkrétněji vývojové prostředí RStudio (RStudio Team, 2015, verze 0.98.1103). Pomineme-li přesné výsledky pro alternativní rozdělení, pak výpočetní schopnosti užitého programu jsou zcela dostačující, přičemž časová náročnost výpočtu vzorce 1.6 roste se zvyšujícími se parametry příliš rychle na to, abychom věřili, že jinými cestami bychom se dostali znatelně dále.

V tabulce 4.3 představujeme pro nepodmíněný případ alternativního rozdělení přesné výsledky jak pro $\mathsf{P}(S_m \geq k)$ (definice 1.2), tak pro všechny aproximace zkoumané v kapitole 2. Vidíme, že aproximace součinového typu (2.2, 2.3) jsou téměř totožné s přesnými výsledky, přičemž dle očekávání aproximace 2.3 překonává aproximaci 2.2. Dokonce má v dané tabulce pouze jeden odlišně zobrazený výsledek od přesné hodnoty. To je způsobeno zaokrouhlováním, odlišnosti se začínají mnohem více projevovat až od pátého či šestého desetinného místa. Rovněž si ale uvědomme, že pro případ $N = 100$ a $m = 25$ nám předpis aproximace 2.3 skutečně dává přesné řešení. Dále si lze všimnout, že první uvedená poissonovská aproximace (2.5) je sice dle svého vzorce nejsnadněji spočitatelná, ale její výsledky jsou zcela neuspokojivé. Druhá poissonovská aproximace (2.7) je o poznání přesnější a její upravená podoba (2.8) je většinou skutečným výsledkům ještě blíže, avšak celkově poissonovské aproximace nemohou v kvalitě odhadu soupeřit s aproximacemi součinového typu. Nakonec aproximace využívající složené Poissonovo rozdělení (2.9, 2.10) bychom hodnotili jako dostatečné, ale s prostorem pro zlepšení.

V podmíněném případě alternativního rozdělení (tabulka 4.4) pozorujeme podobné úkazy jako v nepodmíněném. Stále považujeme za nejlepší aproximace součinového typu (3.2, 3.3), ačkoli rozdíl od přesných výsledků je již viditelný. Rovněž si můžeme všimnout, že poissonovská aproximace 3.5 se vůči ostatním aproximacím zhoršuje, když jsou odhadované pravděpodobnosti malé, naopak její upravená verze 3.6 a aproximace využívající složené Poissonovo rozdělení 3.8 mají sklony se od větších k menším hodnotám odhadované pravděpodobnosti zlepšovat.

Tabulky 4.5 a 4.6 pak i pro větší N pouze podporují naše závěry, že aproximace součinového typu dosahují celkově nejlepších výsledků, ačkoli v tabulce 4.6 se mnohdy ukazuje jako přesnější složená poissonovská aproximace 3.7. Dobře si stojí i poissonovské aproximace zohledňující shlukování určitých jevů na zkoumaném náhodném výběru (2.8, 3.6) a zbylé složené poissonovské aproximace, žádná z nich nad ostatními ale nijak zvlášť nevyniká. V těchto tabulkách porovnáváme aproximace s odhadem $\mathsf{P}(S_m \geq k)$, resp. $\mathsf{P}(k; m, N, a)$, získaným nasimulováním a následným prozkoumáním více než 10 000 náhodných výběrů z alternativního rozdělení pro každou kombinaci parametrů N a p, resp. N a a, jež jsou v tabulkách uvedeny. Samotné aproximace pak stále využívají přesných vzorců.

Jelikož aproximace pro nepodmíněný případ nebyly uvedeny jen za předpokladu alternativního rozdělení, ukažme, že obstojí i pro jiná. Zde však nejen samotné odhady $\mathsf{P}(S_m \geq k)$, ale i Q_j (z definice 1.4) potřebné pro výpočet aproximací musí být získány na základě vysokého počtu (v našem případě 100 000) simulovaných případů. Zkoumali jsme Poissonovo (tabulka 4.7) a binomické (tabulka 4.8) rozdělení. Ukazuje se, že i zde jsou aproximace součinového typu velmi přesné a neupravené poissonovské aproximace spíše selhávají.

4.2 Narozeninový problém

Známý příklad v teorii pravděpodobnosti je určit nejmenší počet lidí, pro který je nadpoloviční pravděpodobnost, že alespoň dva z nich mají narozeniny ve stejný den. Předpokládá se, že pravděpodobnost narozenin člověka je každý den stejná bez ohledu na narozeniny zbývajících lidí. Pro rok s 365 dny je hledaný počet lidí 23.

Zadání této úlohy lze zobecnit mnoha způsoby. Jedním z nich je zajímat se místo alespoň dvou narozenin v jeden den o alespoň dvoje narozeniny v průběhu m dní. Takovým případem se zabývá Naus (1968) a odvozuje potřebné pravděpodobnosti. Na rozdíl od původního problému je zde už třeba rozlišit, zda na sebe dny na konci roku navazují (tj. při standardním kalendáři uvažujeme 31. prosince a 1. ledna jako sousední dny) či ne. Vzhledem k práci s lineárním případem diskrétní skenovací statistiky v minulých kapitolách dále uvažujme, že ne.

Ale kde je ona souvislost se skenovací statistikou? Představme si situaci, kdy víme, že nenastaly dvoje narozeniny ve stejný den. Pak můžeme každému z N dnů přiřadit číslo odpovídající počtu narozenin v daný den a všechny permutace těchto a jedniček a $N - a$ nul jsou za našich předpokladů stejně pravděpodobné. V tom ale už vidíme posloupnost nezávislých stejně rozdělených náhodných veličin s alternativním rozdělením, jejichž součet je znám. Tedy pravděpodobnost, že nastaly alespoň dvoje narozeniny v $m \geq 2$ dnech ($m \leq N$), když každé narozeniny jsou v jiný den, je rovna $\mathsf{P}(2; m, N, a)$ (definice 1.3). Zapišme vše formálněji:

Mějme přirozená čísla N, a, m, přičemž N je z nich největší a představuje počet dní v roce, a pak udává počet uvažovaných narozenin. Označme pro m jev A_m jako nastání alespoň dvou narozenin v m po sobě jdoucích dnech. Potom zřejmě pro původní narozeninový problém chceme najít nejmenší a, že $\mathsf{P}(A_1) > 1/2$. To můžeme udělat za pomoci následujícího vzorce:

$$\mathsf{P}(A_1) = 1 - \mathsf{P}(A_1^c) = 1 - \frac{N}{N} \times \frac{N-1}{N} \times \cdots \times \frac{N-a+1}{N} = 1 - \frac{N!}{(N-a)!N^a}.$$

Při našem zobecnění pak usilujeme o $\mathsf{P}(A_m)$. Jelikož $A_m^c \subset A_1^c$, tak platí

$$\mathsf{P}(A_m) = 1 - \mathsf{P}(A_m^c \cap A_1^c) = 1 - \mathsf{P}(A_m^c | A_1^c)\, \mathsf{P}(A_1^c),$$

kde $\mathsf{P}(A_1^c)$ už umíme spočíst. Takto postupuje i Naus (1968) při hledání svých výsledků. Avšak pro nás jeho vzorce nejsou potřeba, nám stačí, že pro $m \geq 2$ jev $(A_m^c | A_1^c)$ nastává s pravděpodobností $1 - \mathsf{P}(2; m, N, a)$ (viz výše), tedy

$$\mathsf{P}(A_m) = 1 - [1 - \mathsf{P}(2; m, N, a)]\, \mathsf{P}(A_1^c).$$

Tabulka 4.1: Odhady nejmenšího počtu lidí, při kterém je nadpoloviční pravděpodobnost výskytu alespoň 2 narozenin v úseku m dní v roce o N dnech.

N	$m:$	1	2	3	4	5	6	7	14	30
100		13	8	6	6	5	5	5	4	3
365		23	14	11	10	9	8	7	6	4

Pozn: Při výpočtech byla použita aproximace 3.7.

Při vhodném m můžeme využít větu 1 a pro větší parametry můžeme použít právě naše aproximace z kapitoly 3. Pro zajímavost uvádíme odhady nejmenších a, aby $P(A_m) > 1/2$ pro vybrané hodnoty N a m (tabulka 4.1).

Celkově pak vidíme, že pro celé číslo $k \geq 2$ platí, že $\mathsf{P}(k; m, N, a)$ označuje pravděpodobnost, že v roce s N dny má alespoň k lidí narozeniny v m po sobě jdoucích dnech, pakliže žádné dvoje narozeniny nejsou ve stejný den. Bohužel zbavit se podmínky, že narozeniny jsou v různých dnech, není tak snadné, jak by se na první pohled mohlo zdát (předchozí postup selže, neboť méně než k narozenin v m dnech neimplikuje méně než dvoje narozeniny v jednom dnu).

4.3 Nejdelší série úspěchů

V úvodu této práce jsme uvažovali, zda ve 100 hodech mincí, kdy padlo jen 8 panen, není podezřelé mít 5 panen v 10 po sobě jdoucích hodech. Jedná se o podmíněný případ s alternativním rozdělením, můžeme snadno spočíst, že $\mathsf{P}(5; 10, 100, 8) = 0{,}00685$. Obvykle jevy, které mají pravděpodobnost pod 0,05, označujeme za velmi nepravděpodobné, takže ano, 5 či více panen je skutečně velmi hodně.

Také si můžeme položit otázku, kolik panen bychom při spravedlivé minci (tj. pravděpodobnost padnutí panny je 0,5 pro každý hod) mohli v úseku 10 hodů očekávat. Použijeme-li vzorec 2.11 a aproximaci 2.3, dostaneme $\mathsf{E}(S_{10}) \approx 8{,}12203$, tedy bychom očekávali zhruba 8 panen. Avšak mnohem přirozenější otázka je, jaký nejdelší úsek panen bychom očekávali v celých 100 hodech. Formulujme tento problém trochu obecněji:

Nechť máme N nezávislých pokusů, jež mohou skončit úspěchem s pravděpodobností $p \in (0, 1)$, či neúspěchem s pravděpodobností $1 - p$ (tzv. bernoulliovské pokusy). Nechť D označuje délku nejdelší série úspěchů. Jaké je rozdělení D?

Jako obvykle přiřaďme úspěchu hodnotu 1 a neúspěchu hodnotu 0, čímž získáme náhodný výběr z alternativního rozdělení. Jev $(D \geq k)$ je pak roven jevu $(S_k = k) \equiv (S_k \geq k)$. Takže rozdělení známe, platí:

$$\mathsf{P}(D \geq k) = \mathsf{P}(S_k \geq k)$$

a

$$\mathsf{P}(D = k) = \mathsf{P}(S_k \geq k) - \mathsf{P}(S_{k+1} \geq k + 1).$$

V příkladu s mincemi potom dostáváme

$$\mathsf{E}(D) = \sum_{k=1}^{N} k\,\mathsf{P}(D = k) = \sum_{k=1}^{N} \mathsf{P}(D \geq k) = \sum_{k=1}^{N} \mathsf{P}(S_k \geq k),$$

Tabulka 4.2: Odhady pravděpodobností jevu, že ve 100 bernoulliovských pokusech s parametrem p bude mít nejdelší série úspěchů délku alespoň k.

p	$k:$	3	4	5	6	7	10	15
1/6		,319	,061	,010	,002			
1/3		,931	,565	,236	,084	,029	,001	
1/2		1,000	,973	,810	,546	,318	,044	,001

Pozn: Při výpočtech byla použita aproximace 2.2.

což s využitím aproximace 2.2, kde za l dosazujeme N/m (přičemž $p = 0{,}5$, $N = 100$ a m je rovno právě zkoumanému k, tudíž ale porušujeme předpoklady $N/m \in \mathbb{N}$ a $N/m \geq 2$), dává přibližný výsledek 5,99142, tedy očekávaný největší počet po sobě padnoucích panen ve 100 hodech mincí je zhruba 6.

Některé odhady $\mathsf{P}(S_k \geq k)$ pro $N = 100$ získané též z aproximace 2.2 jsou pro ilustraci k nahlednutí v tabulce 4.2, přičemž můžeme vidět, že kdyby byly řádky tabulky kompletní (tj. od jedné do N), tak střední hodnota největšího počtu po sobě jdoucích pokusů s pravděpodobností úspěchu p je součet hodnot příslušného řádku.

Ještě je dobré si uvědomit, že občas lze podobný postup nasadit, i když mají původní pozorování více než dva možné výsledky. Například pokud házíme obyčejnou šestistěnnou kostkou a ptáme se na pravděpodobnost, že v N hodech má nejdelší série po sobě padnoucích stejných čísel délku alespoň k, můžeme původní posloupnost přepsat a využít postup na nově získanou posloupnost, která nám bude značit, zda byl předchozí hod roven současnému:

2	5	4	4	4	6	2	1	6	5	1	1	4	3	5	5...
	0	0	1	1	0	0	0	0	0	0	1	0	0	0	1...

Zjevně pak hledáme $\mathsf{P}(S_{k-1} \geq k-1)$ při $N-1$ nezávislých pozorováních s alternativním rozdělením s parametrem $p = 1/6$.

Tabulka 4.3: Aproximace pro nepodmíněný případ alternativního rozdělení ve srovnání s přesnými výsledky při $N = 50$ (horní část) a $N = 100$ (dolní část).

m	p	k	$\mathsf{P}(S_m \geq k)$	2.2	2.3	2.5	2.7	2.8	2.9	2.10
5	,05	2	,3179	,3179	,3179	,6463	,4238	,2990	,3194	,3027
		3	,0288	,0288	,0288	,0519	,0379	,0287	,0292	,0287
	,10	2	,7304	,7307	,7304	,9764	,8337	,6375	,7155	,6594
		3	,1784	,1784	,1784	,3255	,2316	,1726	,1830	,1735
		4	,0146	,0146	,0146	,0209	,0182	,0145	,0148	,0145
10	,05	2	,4849	,4849	,4849	,9707	,5675	,4085	,4541	,4201
		3	,1143	,1143	,1143	,3760	,1435	,1098	,1159	,1103
		4	,0146	,0146	,0146	,0413	,0177	,0145	,0149	,0145
10	,05	2	,7476	,7478	,7476	,9996	,8483	,6504	,7390	,6739
		3	,2272	,2272	,2272	,6489	,2942	,2178	,2393	,2193
		4	,0311	,0311	,0311	,0893	,0394	,0310	,0328	,0310
		5	,0025	,0025	,0025	,0058	,0031	,0025	,0026	,0025
	,10	2	,9855	,9856	,9855	1,0000	,9912	,8457	,9673	,9069
		3	,7462	,7464	,7462	,9983	,8313	,6530	,7607	,6723
		4	,2816	,2816	,2816	,6879	,3503	,2676	,3069	,2696
		5	,0542	,0542	,0542	,1382	,0672	,0537	,0591	,0538
		6	,0063	,0063	,0063	,0133	,0076	,0063	,0067	,0063
	,20	4	,9193	,9196	,9193	1,0000	,9484	,7938	,9252	,8257
		5	,5706	,5707	,5706	,9494	,6555	,5158	,6186	,5241
		6	,1904	,1904	,1904	,4399	,2315	,1841	,2138	,1848
20	,05	2	,8714	,8717	,8714	1,0000	,8781	,6423	,7893	,6906
		3	,5010	,5011	,5009	,9978	,5691	,4210	,5022	,4318
		4	,1714	,1714	,1714	,7242	,2085	,1616	,1859	,1626
		5	,0385	,0385	,0385	,1882	,0464	,0380	,0423	,0380
		6	,0063	,0063	,0063	,0263	,0073	,0062	,0068	,0062
	,10	3	,9439	,9442	,9439	1,0000	,9225	,6803	,8865	,7437
		4	,7266	,7270	,7266	1,0000	,7584	,5671	,7336	,5911
		5	,3953	,3954	,3953	,9697	,4515	,3454	,4419	,3509
		6	,1500	,1500	,1500	,5981	,1783	,1426	,1764	,1432
		7	,0418	,0418	,0418	,1757	,0494	,0412	,0490	,0413
		8	,0090	,0090	,0090	,0331	,0104	,0090	,0103	,0090
	,20	5	,9764	,9766	,9764	1,0000	,9451	,6948	,9452	,7724
		6	,8725	,8730	,8725	1,0000	,8636	,6505	,8818	,6901
		7	,6404	,6408	,6404	,9991	,6765	,5161	,7104	,5317
		8	,3605	,3606	,3605	,9260	,4072	,3192	,4371	,3232
		9	,1542	,1542	,1542	,5545	,1797	,1464	,1938	,1470
25	,05	2	,8962	,8964	,8962	1,0000	,8583	,6204	,7671	,6624
		3	,5877	,5879	,5877	,9999	,6212	,4623	,5602	,4750
		4	,2525	,2525	,2525	,9250	,2924	,2276	,2683	,2296
		5	,0744	,0744	,0744	,4199	,0880	,0722	,0824	,0723
		6	,0162	,0162	,0162	,0881	,0188	,0161	,0179	,0161
		7	,0028	,0028	,0028	,0127	,0031	,0028	,0030	,0028

Tabulka 4.4: Aproximace pro podmíněný případ alternativního rozdělení ve srovnání s přesnými výsledky při $N = 50$ a $N = 100$.

N	m	a	k	$\mathsf{P}(k; m, N, a)$	3.2	3.3	3.5	3.6	3.7	3.8
50	5	5	2	,8687	,7415	,7638	,8161	,6514	,6944	,6740
			3	,1217	,1170	,1181	,1485	,1146	,1162	,1152
			4	,0038	,0038	,0038	,0047	,0038	,0038	,0038
		10	3	,7607	,6715	,6886	,7420	,5981	,6459	,6138
			4	,1223	,1182	,1192	,1451	,1158	,1188	,1163
			5	,0049	,0049	,0049	,0058	,0049	,0049	,0049
	10	5	2	,9991	,9512	,9788	,8972	,7097	,8132	,7531
			3	,4941	,4523	,4773	,4786	,3899	,4053	,4006
			4	,0637	,0630	,0634	,0734	,0617	,0615	,0619
			5	,0025	,0025	,0025	,0029	,0025	,0024	,0025
100	10	5	2	,8987	,7654	,7885	,8355	,6703	,7207	,6945
			3	,1665	,1578	,1598	,1988	,1534	,1573	,1546
			4	,0090	,0090	,0090	,0110	,0090	,0090	,0090
		10	3	,8522	,7471	,7665	,8080	,6583	,7321	,6779
			4	,2320	,2177	,2210	,2641	,2095	,2266	,2115
			5	,0263	,0261	,0261	,0314	,0260	,0271	,0260
			6	,0016	,0016	,0016	,0018	,0016	,0016	,0016
		15	3	,9996	,9771	,9835	,9786	,8570	,9566	,8948
			4	,7432	,6590	,6753	,7237	,5890	,6690	,6031
			5	,2025	,1923	,1946	,2302	,1859	,2057	,1873
			6	,0270	,0268	,0268	,0318	,0266	,0284	,0267
			7	,0021	,0021	,0021	,0024	,0021	,0021	,0021
		20	4	,9866	,9327	,9445	,9448	,8141	,9209	,8426
			5	,5901	,5322	,5441	,5977	,4857	,5584	,4946
			6	,1441	,1389	,1402	,1652	,1356	,1507	,1363
			7	,0188	,0187	,0187	,0219	,0186	,0199	,0186
			8	,0014	,0014	,0014	,0016	,0014	,0014	,0014
	20	5	2	,9994	,9527	,9798	,8989	,7102	,8173	,7543
			3	,5370	,4873	,5167	,5097	,4155	,4351	,4278
			4	,0828	,0816	,0824	,0947	,0795	,0791	,0798
			5	,0043	,0043	,0043	,0049	,0043	,0042	,0043
		10	4	,8653	,7798	,8254	,7487	,6107	,7183	,6342
			5	,3510	,3307	,3433	,3563	,2967	,3358	,3022
			6	,0752	,0742	,0748	,0844	,0724	,0785	,0727
			7	,0096	,0096	,0096	,0108	,0096	,0100	,0096
			8	,0007	,0007	,0007	,0008	,0007	,0007	,0007
		15	5	,9545	,8841	,9218	,8313	,6744	,8335	,7037
			6	,5680	,5228	,5496	,5321	,4417	,5350	,4532
			7	,1896	,1836	,1875	,2033	,1729	,2026	,1746
			8	,0415	,0412	,0414	,0464	,0407	,0456	,0407
			9	,0062	,0062	,0062	,0069	,0062	,0067	,0062
			10	,0006	,0006	,0006	,0007	,0006	,0007	,0006

Tabulka 4.5: Vybrané aproximace pro nepodmíněný případ alternativního rozdělení pro větší N ve srovnání s výsledky získanými z více než 10 000 simulací.

N	m	p	k	$\hat{\mathsf{P}}(S_m \geq k)$	2.3	2.7	2.8	2.9	2.10
500	10	,01	2	,3331	,3270	,4518	,3227	,3386	,3236
			3	,0149	,0159	,0213	,0159	,0162	,0159
			4	,0004	,0004	,0005	,0004	,0004	,0004
		,05	3	,7364	,7405	,8500	,7221	,7714	,7256
			4	,1580	,1541	,1967	,1532	,1649	,1533
			5	,0129	,0134	,0166	,0134	,0141	,0134
	20	,01	2	,5264	,5218	,6716	,5007	,5399	,5056
			3	,0600	,0638	,0858	,0635	,0667	,0635
			4	,0034	,0038	,0048	,0038	,0039	,0038
1000	10	,01	2	,5523	,5487	,7032	,5428	,5658	,5441
			3	,0345	,0317	,0426	,0317	,0325	,0317
			4	,0010	,0008	,0010	,0008	,0008	,0008
		,05	3	,9335	,9337	,9784	,9238	,9491	,9257
			4	,2818	,2861	,3575	,2846	,3049	,2848
			5	,0257	,0269	,0332	,0268	,0283	,0268
			6	,0017	,0015	,0017	,0015	,0015	,0015
	20	,01	2	,7775	,7742	,8971	,7532	,7947	,7583
			3	,1337	,1249	,1673	,1243	,1313	,1244
			4	,0086	,0077	,0098	,0077	,0080	,0077

Tabulka 4.6: Vybrané aproximace pro podmíněný případ alternativního rozdělení pro větší N ve srovnání s výsledky získanými z více než 10 000 simulací.

N	m	a	k	$\hat{\mathsf{P}}(k; m, N, a)$	3.3	3.5	3.6	3.7	3.8
500	10	25	3	,7826	,7221	,8293	,7025	,7463	,7063
			4	,1382	,1314	,1668	,1306	,1390	,1307
			5	,0099	,0096	,0118	,0096	,0100	,0096
		50	4	,8600	,8119	,8911	,7902	,8449	,7940
			5	,2405	,2314	,2840	,2292	,2511	,2295
			6	,0288	,0270	,0327	,0270	,0288	,0270
	20	25	4	,6802	,6204	,7163	,5898	,6629	,5953
			5	,1706	,1623	,2002	,1599	,1801	,1603
			6	,0233	,0233	,0282	,0233	,0256	,0233
1000	10	25	3	,3399	,3263	,4188	,3240	,3394	,3244
			4	,0245	,0210	,0267	,0210	,0217	,0210
			5	,0005	,0007	,0008	,0007	,0007	,0007
		50	3	,9621	,9291	,9754	,9185	,9436	,9206
			4	,2718	,2667	,3331	,2652	,2828	,2654
			5	,0231	,0229	,0282	,0229	,0240	,0229
			6	,0014	,0011	,0013	,0011	,0012	,0011
	20	25	3	,8214	,7612	,8624	,7400	,7898	,7442
			4	,1787	,1699	,2155	,1687	,1827	,1689
			5	,0195	,0165	,0203	,0165	,0175	,0165

Tabulka 4.7: Aproximace pro nepodmíněný případ Poissonova rozdělení při $N = 100$ ve srovnání s výsledky získanými ze 100 000 simulací.

m	λ	k	$\hat{\mathsf{P}}(S_m \geq k)$	2.2	2.3	2.5	2.7	2.8	2.9	2.10
10	,05	2	,7536	,7574	,7518	,9997	,8524	,6585	,7230	,6805
		3	,2560	,2468	,2525	,7293	,3396	,2352	,2595	,2393
		4	,0459	,0424	,0489	,1526	,0601	,0421	,0481	,0443
		5	,0055	,0061	,0053	,0171	,0069	,0061	,0055	,0058
		6	,0006	,0006	,0013	,0018	,0003	,0006	,0009	,0008
	,10	2	,9838	,9837	,9833	1,0000	,9906	,8427	,9641	,9036
		3	,7656	,7719	,7663	,9994	,8532	,6723	,7520	,6925
		4	,3462	,3587	,3479	,8342	,4296	,3365	,3622	,3362
		5	,0971	,0939	,0920	,2968	,1268	,0923	,0699	,0919
		6	,0194	,0199	,0183	,0540	,0233	,0199	,0357	,0193
		7	,0031	,0026	,0030	,0100	,0043	,0026	,0011	,0027
20	,05	2	,8702	,8690	,8669	1,0000	,8755	,6413	,7804	,6883
		3	,5132	,5154	,5095	,9984	,5820	,4310	,5588	,4412
		4	,1910	,1915	,1898	,7824	,2353	,1792	,2075	,1801
		5	,0507	,0462	,0482	,2487	,0639	,0454	,0756	,0462
		6	,0101	,0087	,0100	,0574	,0118	,0087	,0016	,0091
		7	,0017	,0020	,0015	,0065	,0018	,0020	,0024	,0019

Tabulka 4.8: Vybrané aproximace pro nepodmíněný případ binomického rozdělení při $N = 100$ ve srovnání s výsledky získanými ze 100 000 simulací.

m	n	p	k	$\hat{\mathsf{P}}(S_m \geq k)$	2.2	2.3	2.7	2.8	2.9	2.10
10	5	,01	2	,7495	,7527	,7471	,8518	,6543	,7541	,6767
			3	,2500	,2574	,2464	,3301	,2454	,2827	,2440
			4	,0420	,0345	,0455	,0589	,0342	,0420	,0379
			5	,0052	,0040	,0065	,0065	,0040	,0073	,0048
		,05	5	,8826	,8807	,8872	,9252	,7601	,8774	,7906
			6	,5968	,5995	,5905	,6882	,5386	,6788	,5462
			7	,2826	,2869	,2818	,3507	,2726	,3428	,2729
			8	,1004	,1046	,0994	,1250	,1027	,0954	,1012
			9	,0283	,0320	,0272	,0312	,0318	,0600	,0302
			10	,0073	,0068	,0059	,0077	,0068	,0145	,0065
	10	,01	2	,9843	,9847	,9846	,9906	,8454	,9650	,9053
			3	,7654	,7667	,7710	,8520	,6676	,7321	,6907
			4	,3413	,3502	,3436	,4270	,3287	,3753	,3298
			5	,0929	,0842	,0924	,1224	,0828	,0831	,0858
			6	,0187	,0170	,0177	,0228	,0170	,0145	,0172
			7	,0029	,0021	,0039	,0031	,0021	,0027	,0027
20	5	,01	2	,8674	,8686	,8669	,8759	,6408	,7551	,6882
			3	,5091	,5063	,5087	,5761	,4244	,4873	,4363
			4	,1869	,1838	,1848	,2274	,1725	,2230	,1740
			5	,0472	,0459	,0488	,0573	,0452	,0187	,0462
			6	,0092	,0096	,0092	,0105	,0096	,0145	,0094
			7	,0014	,0017	,0011	,0015	,0017	,0048	,0015

Závěr

V práci jsme čtenáře seznámili s diskrétní skenovací statistikou a uvedli množství aproximací pro výpočet jejího rozdělení. Zabývali jsme se nepodmíněným i podmíněným případem, přičemž v obou jsme dané aproximace dělili do tří skupin: aproximací součinového typu, poissonovských aproximací a aproximací využívajících složené Poissonovo rozdělení. V případě aproximací součinového typu bylo předvedeno odvození, pro zbylé skupiny jsme nastínili myšlenky, ze kterých vycházejí. Rovněž jsme odvodili vzorce pro střední hodnotu a rozptyl diskrétní skenovací statistiky.

Na základě vyčíslení aproximací pro množství případů, převážně pro množství případů s alternativním rozdělením, jsme zhodnotili uvedené aproximace a uvedli zkoumané hodnoty čtenáři k nahlédnutí. Díky uvedení postupů pro výpočet přesných výsledků pro alternativní rozdělení by čtenář měl být schopen jen s využitím obsahu této práce příslušné hodnoty přepočíst.

Nejlépe jsme zhodnotili aproximace součinového typu, které pokaždé, i při binomickém či Poissonově rozdělení, dosahovaly velmi přesných výsledků. Nejjednodušší poissonovské aproximace se ukázaly být značně nevyhovující, avšak názorně jsme předvedli úpravu zohledňující závislost jevů, ze kterých tyto aproximace vycházejí, po které se již výsledky ukázaly jako obstojné. Aproximace využívající složené Poissonovo rozdělení dosahují proměnlivých výsledků, převážně si však v hodnocení stojí o něco lépe než i ty nejlepší poissonovské aproximace.

Dále byla uvedena spojitost s narozeninovým problémem, přesněji využití diskrétní skenovací statistiky při hledání pravděpodobnosti 2 či více narozenin ve stanoveném počtu po sobě jdoucích dní. Též jsme na příkladu předvedli odhad největšího počtu po sobě jdoucích úspěchů v řadě bernoulliovských pokusů.

Skenovací statistika je značně rozsáhlá oblast, přičemž i pro její diskrétní případ (oproti spojitému) se lze zabývat mnohými souvisejícími problémy, ať už prací ve více rozměrech či například zkoumáním více posloupností náhodných jevů najednou. Vzhledem k možným aplikacím v mnoha odvětvích, kde je třeba dávat pozor na neobvyklé shluky určitých jevů, považujeme případný hlubší výzkum této oblasti nejen za zajímavý, ale i přínosný.

Seznam použité literatury

Glaz, J. a Balakrishnan, N., editors (1999). *Scan Statistics and Applications*. Birkhäuser, Boston, MA. ISBN 0-8176-4041-X.

Glaz, J., Naus, J. a Wallenstein, S. (2001). *Scan Statistics*. Springer-Verlag, New York, NY. ISBN 0-387-98819-X.

Naus, J. I. (1968). An Extension of the Birthday Problem. *The American Statistician*, **22**(1), 27–29.

Naus, J. I. (1982). Approximations for Distributions of Scan Statistics. *Journal of the American Statistical Association*, **77**(377), 177–183.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`. Version 3.1.3.

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA. URL `http://www.rstudio.com/`. Version 0.98.1103.

Saperstein, B. (1972). The Generalized Birthday Problem. *Journal of the American Statistical Association*, **67**(338), 425–428.

Seznam tabulek