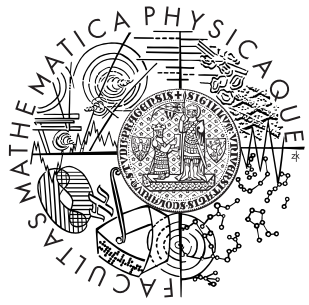
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Daniel Petrík

Narozeninový problém a jeho modifikace

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Ing. Marek Omelka, Ph.D.

Studijní program: Matematika obor: Finanční matematika

2009

Rád bych poděkoval Ing. Marku Omelkovi, Ph.D. za vedení mé práce, a také svým rodičům za podporu nejen ve studiu.

Prohlašuji, že jsem svou bakalářskou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 29. května 2009 Daniel Petrík

Obsah

Název práce: Narozeninový problém a jeho modifikace
Autor: Daniel Petrík
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: Ing. Marek Omelka, Ph.D.
e-mail vedoucího: omelka@karlin.mff.cuni.cz

Abstrakt: V předložené práci studujeme jeden z klasických příkladů ze základů teorie pravděpodobnosti. Zabýváme se způsoby stanovení pravděpodobnosti, že ve skupině n náhodně vybraných lidí existují alespoň dva, kteří se narodili ve stejný den (ne nutně ve stejný rok). Nejprve problém zkoumáme za platnosti zjednodušujícího předpokladu o rovnoměrném diskrétním rozdělení pravděpodobnosti narození v daném dnu, dojdeme k přesnému analytickému vyjádření hledané pravděpodobnosti a uvedeme i jeho některé vhodné aproximace. Dále tento předpoklad uvolníme a uvážíme reálnou situaci, že tato pravděpodobnost zcela stejná není. Do zkoumání zahrneme i reálná data z Českého statistického úřadu o četnosti porodů v jednotlivých měsících v ČR Analytické výsledky porovnáváme a zjišťujeme, na kolik je zjednodušující předpoklad poškozující. Také se budeme zabývat zcela obecným problémem pravděpodobnosti kolize, při umisťovaní n předmětů do k přihrádek, jehož konkrétním případem je narozeninový problém při volbě $k = 365$. Součástí práce jsou i počítačové simulace, jejichž výsledky porovnáváme s analyticky zjištěnými.

Klíčová slova: pravděpodobnost, narození, kolize, rovnoměrné a nerovnoměrné rozdělení

Title: Birthday problem and its modifications
Author: Daniel Petrík
Department: Department of Probability and Mathematical Statistics
Supervisor: Ing. Marek Omelka, Ph.D.
Supervisor's e-mail address: omelka@karlin.mff.cuni.cz

Abstract: In the present work we study one of the classical problems from the basics of the probability theory. We deal with the ways of determination of the probability that in the group of n radomly selected people there are at least two of them who were born on the same day (not necessarily in the same year). At first we explore the problem with a simplyfing assump-

tion of a discrete uniform distribution of birth on the same day, we reach an exact analytic formula of wanted probability and we mention some convenient approximations of this result. Furthermore we leave this assumption and consider the real situation where the probabilities aren't completely the same. To our investigation we also include some real facts and numbers from the Czech Statistical Office concerning the frecuency of birth in different months in the Czech Republic. We compare the analytic results and find out the negative impact of the simplyfing assumption of uniformity on the credibility of results. We also study a generalization of Birthday problem which is the task to compute the probability of collision when we are placing n objects to k ($= 365$) boxes (each box is only for one object; the so called collision means that there are two objects in the same box). There are also some computer simulations as a part of the work, whose results are confronted with the analytically determined.

Keywords: probability, birth, collision, uniform and non-uniform distribution

Úvod

Pravděpodobnost se kromě výlučně teoretických problémů často přímo dotýká i našeho každodenního života. S oblibou se uveřejňují nejrůznější statisticky určené šance na úmrtí na určitou nemoc, na dopravní nehodu v určitém dopravním prostředku apod. Jiným typem problému je například zkoumání pravděpodobnosti výhry n-té ceny v loterii, již je možno určit teoreticky jen na základě kombinatorického počítání. K této skupině problémů můžeme beze sporu přiřadit i narozeninový problém, tedy zkoumání jevu, že ve skupině o určitém počtu lidí existují alespoň dvě osoby mající narozeniny ve stejný den (vyžadujeme přitom shodu ve dni a měsíci, nikoli v roce). Nicméně v obecném případě se ukáže, že ačkoli jsme schopni napsat formuli pro danou pravděpodobnost, ani nejvýkonnější počítač ji nezvládne přesně vyčíslit ani za 1000 let, proto se buď tato pravděpodobnost aproximuje, nebo se přibližně určuje statisticky, tak jako se vyhodnocuje úmrtnost na určitou nemoc nebo nehodovost v určitých dopravních prostředcích.

I když narozeninový problém patří ke klasickým příkladům teorie pravděpodobnosti a jistě se nad ním mohli pozastavovat lidé už před několika staletími, o jeho první publikování se zasloužil až rakouský matematik a fyzik Richard von Mises (1883-1953). Ve svém článku *Über Aufteilungs und Besetzungs-Wahrscheinlichkeiten* (O rozdělení pravděpodobnosti v přihrádkovém problému) si v roce 1939 klade otázku: „Kolik lidí musí být v místnosti, aby pravděpodobnost toho, že alespoň dvě osoby mají narozeniny ve stejný den, byla přinejmenším 50%, pokud ignorujeme přestupné roky?" Tento článek byl v roce 1964 přeložen do angličtiny a vytištěn v knize *Selected Papers of Richard von Mises*, což zvýšilo popularitu a povědomí o narozeninovém problému a podnítilo další prohlubování znalostí v této oblasti napříč anglosaským světem.

V následujících letech se zájem matematiků ubíral cestou nejrůznějšího zobecňování klasického narozeninového problému. Již v 60. letech matema-

tikové jako E. H. McKinney, M. Klamkin a D. Newman začali uvažovat nad tím, jaký dopad má na úvahy zjednodušující předpoklad, se kterým Mises tiše počítal, a sice, že rozdělení pravděpodobnosti narození v jednolitvých dnech roku je rovnoměrné. Jedním z nejdůležitějších výsledků se v tomto směru záhy stala *narozeninová nerovnost*, o níž pojednáváme v podkapitole (3.3) a která v důsledku říká, že shoda narozenin při nerovnoměrném rozdělením pravděpodobnosti narození během roku nastává s vyšší pravděpodobností než při rozdělení rovnoměrném. Formální důkaz této nerovnosti podal David Bloom v roce 1973. S rozvojem informatiky se ukázalo, že narozeninový problém ve svém zobecnění na přihrádkový problém má i jisté aplikace v kryptografii při hashování, čemuž se věnujeme ve druhé kapitole.

Teoretické závěry, které učiníme, lze neformálně předvést i na několika spíše kuriozních případech z reality: Například jedenáctý a dvacátý devátý prezident USA se narodili 2. listopadu, ze 73 oceněných herců v hlavní roli na amerických Oscarech se 6 dvojic z nich narodilo ve stejný den, v případě hereček se jedná o 3 dvojice ze 67 oceněných a v kategorii režisérů, 5 dvojic ze 61 oceněných. Z 52 premiérů Velké Británie se 2 dvojice z nich narodily ve stejný den a mezi australskými premiéry tato shoda nastává až s 24 premiérem. Jak je tedy vidět, nemusíme zkoumat narozeninový problém jen ze synchronní perspektivy, ale můžeme tento jev analyzovat i napříč stoletími. Podstatné však je, aby byl výběr osob dostatečně náhodný. Ve třetí kapitole uvidíme, že různorodé ročníkové složení takové skupiny osob je velmi žádoucí.

Na závěr zmiňme, že v současnosti se teorie skrytá za narozeninovým (přihrádkovým) problémem uplatňuje vedle kryptografie také v epidemiologických registrech či v ekologii při určování velikosti populací.

Kapitola 1

Narozeninový problém s předpokladem rovnoměrnosti

1.1 Předpoklady rovnoměrnosti rozdělení

Chceme-li stanovit hodnotu pravděpodobnosti toho, že ve skupině n náhodně vybraných osob existují alespoň dva lidé mající narozeniny ve stejný den, bez potřeby obrovského množství vstupních dat (jakým by byly např. reálné pravděpodobnosti narození v jednotlivých dnech v roce zjištěné statisticky na určitém území v určitém období), je vhodné problém řešit s předpokladem rovnoměrného rozdělení pravděpodobností narození během roku, tj. že pravděpodobnost narození v libovolném dni v roce (29. únor vyjímaje) je $\frac{1}{365}$.

K předpokladu rovnoměrnosti rozdělení pravděpodobnosti narození v jednotlivých dnech během roku ještě přidáme následující dva předpoklady:

1. rok má vždy 365 dní (tj. neexistuje datum 29. únor[1], které by vykazovalo podstatně menší pravděpodobnost narození než ostatní dny v roce).

2. nerodí se dvojčata, trojčata či vícerčata (vícerčata porušují předpoklad nezávislosti pro jev, že dva lidé mají narozeniny ve stejný den, jenž budeme využívat při odvozování vzorců).

[1]Dle gregoriánského kalendáře je přestupný každý rok dělitelný 4 s výjimkou let dělitelných 100, které současně nejdou dělitelné 400 (tj. během každných 400 let je právě $25 + 24 + 24 + 24 = 97$ přestupných.

Zřejmě vliv obou výše uvedených předpokladů na přesnost výsledků (o vlivu předpokladu rovnoměrnosti budeme pojednávat ve třetí a čtvrté kapitole) je jen velmi malý:

1. pravděpobnost toho, že se člověk narodí v kterýkoli jiný den než 29. únor, je $\frac{146000}{146097} = 99,93\%$, tedy 29. únor není relevantní pro zkoumaný jev, navíc se zahrnutím 29. února do úvah by se pravděpodobnost narození v libovolném dnu v roce (kromě 29. února) zmenšila jen o asi $1,819 \cdot 10^{-6}$ z původních $\frac{1}{365}$ na $\frac{400}{146097}$.

2. v roce 2005 připadlo v ČR pouze 1939 vícečetných porodů na celkem 100 546 porodů[2], přijmeme-li příslušnou relativní četnost vícečetného porodu 1,93% jako odhad jeho pravděpodobnosti, potom v předložené práci vlastně zkoumáme pravděpodobnost existence shody narozenin v nějaké skupině lidí za podmínky neexistence vícečetného porodu (98,07%), tzn. nahrazením jevu prosté existence shody narozenin v dané skupině lidí příslušným jevem podmíněným se dopouštíme jen zanedbatelné nepřesnosti.

1.2 Analytický výpočet pravděpodobnosti

Pro určení pravděpodobnosti $p(n)$ jevu, že mezi n lidmi existují alespoň dvě osoby, které mají narozeniny ve stejný den, bude nejsnazší nejprve určit pravděpodobnost $q(n)$ příslušného doplňkového jevu, tj. že mezi n lidmi mají každé dvě náhodně vybrané osoby narozeniny v různý den. Je-li $n \geq 365$, potom s využitím Dirichletova principu snadno nahlédneme, že mezi n lidmi jistě existují aspoň dvě osoby mající narozeniny ve stejný den, tudíž $p(n) = 1$ a $q(n) = 0$. Pro $n \in \{1, 2, .., 365\}$ stačí použít jednoduchou úvahu a kombinatorický princip součinu: Totiž druhá osoba nemůže mít narozeniny ve stejný den jako první, čili připadá na ni 364 možných dnů, a tedy pravděpodobnost, že má druhá osoba narozeniny v jiný den než osoba první je $\frac{364}{365}$. Dále třetí osoba nemůže mít narozeniny ve stejný den jako má první či druhá osoba, neboli připadá na ni 363 možných dnů, a tedy pravděpodobnost, že má třetí osoba narozeniny v jiný den než osoba první nebo druhá je $\frac{363}{365}$ atd.

Obecně lze říci, že pravděpodobnost toho, že při $2 \leq k < n$ má k-tá osoba narozeniny v jiné dny než první až $(k-1)$-ní osoba, je $\frac{365-k+1}{365}$. Jelikož

[2]dle údajů Českého statistického úřadu [1]

takovou vlastnost vyžadujeme pro všechna přípustná k, a to nezávisle na sobě, dostáváme vztah (1.1) pro pravděpodobnost $q(n)$:

$$q(n) = \prod_{k=2}^{n} \frac{365-k+1}{365} = \prod_{k=1}^{n} \frac{365-k+1}{365} = \frac{365 \cdot 364 \cdot \ldots \cdot (365-n+1)}{365^n} =$$

$$= \frac{365!}{365^n (365-n)!} \tag{1.1}$$

Jelikož jsou oba diskutované jevy (vyčíslené pravděpodobnostmi $p(n)$ a $q(n)$) komplementární, musí platit, že $p(n) + q(n) = 1$. Úpravou obdržíme hledané vyjádření (1.2) pro pravděpodobnost $p(n)$:

$$p(n) = 1 - q(n) = 1 - \frac{365!}{365^n (365-n)!} \tag{1.2}$$

1.3 Narozeninový paradox

Spočítáme-li pro zajímavost několik hodnot $p(n)$ (tabulka 1.1) a vykreslíme-li graf závislosti pravděpodobnosti $p(n)$ na počtu osob n (obrázek 1.2), dospějeme k tzv. *narozeninovému paradoxu*. Totiž fakt, že ve skupině 23 náhodně vybraných lidí platí $p(23) > 0,5$ je na první pohled paradoxní. Většina laických úsudků počítá s tím, že by tato pravděpodobnost měla být výrazně nižší, nebo, že by pro dosažení alespoň 50%-ní pravděpodobnosti bylo nutné uvažovat mnohem početnější skupinu lidí. Tento zjevný nesoulad pramení z intuice a zkušeností, že např. na základní či střední škole nebo v pracovním kolektivu se nikdy neobjevila žádná osoba mající narozeniny ve stejný den jako my. Jenže předešlá úvaha není ekvivalentní s původně formulovaným problémem; původní problém totiž zkoumá jev, že nastane jakákoli shoda ve dni narození, kdežto problém právě diskutovaný zkoumá jev, že nastane konkrétní shoda jedné pevně zvolené osoby s jinou.

Označme $\tilde{p}(n)$ pravděpodobnost, že ve skupině n náhodně vybraných lidí existuje člověk mající narozeniny ve stejný den jako jiná ze skupiny pevně určená osoba, potom s vyjmutím pevně zvolené osoby z úvah má každý ze zbývajících $n-1$ lidí (a to nezávisle na sobě) pravděpodobnost $\frac{364}{365}$, že se narodil v jiný den než pevně určená osoba. Jelikož hledáme pravděpodobnost komplementárního jevu, pravděpodobnost $\tilde{p}(n)$ musí vyhovovat vzorci (1.3).

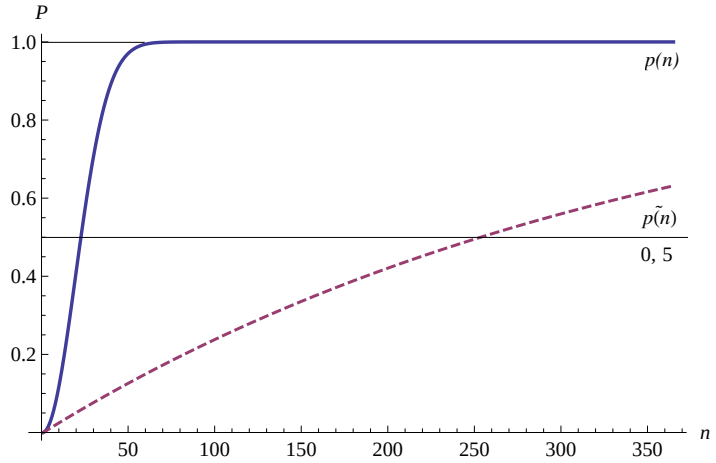
$$\tilde{p}(n) = 1 - \left(\frac{364}{365}\right)^{n-1} \tag{1.3}$$

Pro porovnání obou problémů i v tomto případě numericky spočteme některé hodnoty $\tilde{p}(n)$ (tabulka 1.1) a vykreslíme graf (obrázek 1.1).

Tabulka 1.1: Některé numerické hodnoty $p(n)$ a $\tilde{p}(n)$

n	$p(n)$	$\tilde{p}(n)$
10	$0,1169$	$0,0244$
20	$0,4114$	$0,0508$
23	$0,5073$	$0,0586$
30	$0,7063$	$0,0765$
40	$0,8912$	$0,1015$
50	$0,9704$	$0,1258$
60	$0,9941$	$0,1494$
70	$0,9992$	$0,1725$
80	$0,99991$	$0,1949$
90	$0,999994$	$0,2166$
100	$0,9999997$	$0,2378$
200	$1-1,61\cdot 10^{-30}$	$0,4207$
254	$1-6,52\cdot 10^{-54}$	$0,5005$
300	$1-6,25\cdot 10^{-82}$	$0,5597$
350	$1-3,03\cdot 10^{-131}$	$0,6161$
366	1	$0,6326$

Obrázek 1.1: Graf závislosti $p(n)$ a $\tilde{p}(n)$ na n



Dle numerických hodnot z tabulky 1.1 i z grafu na obrázku 1.1 lze jednoznačně usuzovat, že pravděpodobnost $p(n)$ konverguje k 1 mnohem rychleji (dokonce této hodnoty i dosáhne pro $n \geq 365$) než pravděpodobnost $\tilde{p}(n)$, která této hodnoty nikdy nedosáhne. Rychlost konvergence $p(n)$ k 1 je natolik velká, že už ve skupině 57 náhodně vybraných osob existují s pravděpodobností převyšující 99% aspoň dva jedinci mající narozeniny ve stejný den. Je zřejmé, že aby v určité skupině náhodně vybraných osob existovali aspoň dva mající narozeniny ve stejný den s pravděpodobnostní aspoň 0,5, stačí, aby tato skupina obsahovala aspoň 23 lidí. Na druhou stranu, na to, aby ve skupině existoval člověk mající narozeniny v ten samý den s pevně určenou osobou z dané skupiny, musela by tato skupina obsahovat alespoň 254 osob. Tento zjevný nepoměr můžeme považovat za konrétní příklad narozeninového paradoxu.

1.4 Aproximace pravděpodobnosti

Vzhledem k výpočetní náročnosti vzorců (1.2) a (1.3), která pro vyšší n umožňuje přesné vyčíslení uvažovaných pravděpodobností snad jen s použitím výpočetní techniky, je vhodné např. ke stanovení pravděpodobnosti $p(n)$ pro daný počet osob n používat aproximací. Dále uvedeme několik běžných aproximací užívaných v těchto souvislostech, které poskytují dostatečně přesné výsledky. Numerické hodnoty a grafické porovnání zmiňovaných aproximací s přesnou hodnotou pravděpodobnosti i navzájem provedem v tabulce (1.2) a obrázku (1.2) na závěr této podkapitoly.

a) Exponenciální aproximace 1. řádu

Využívá první členy rozvoje exponenciely v Taylorovu řadu; tedy z rozvoje $e^x = \sum_{k=1}^{\infty} \frac{x^k}{k!} = 1 + x + o(x)$, kde $o(x)$ je funkce taková, že $\lim_{x \to 0} \frac{o(x)}{x} = 0$, lze prohlásit, že $e^x \approx 1 + x$ pro x blízká 0. Použijeme-li vztah (1.1) a dosadíme-li přibližnou hodnotu e^x za $1+x$, dojdeme k aproximaci (1.4) pro pravděpodobnost $q(n)$:

$$q(n) = \prod_{k=1}^{n} \left(1 - \frac{k-1}{365}\right) \approx \prod_{k=1}^{n} e^{-\frac{k-1}{365}} = e^{-\sum_{k=1}^{n} \frac{k-1}{365}} = e^{-\frac{n(n-1)}{2 \cdot 365}} \equiv q_{\exp}(n) \quad (1.4)$$

Z komplementarity příslušných jevů vyplývá aproximace (1.5) pro pravděpodobnost $p(n)$:

$$p(n) = 1 - q(n) \approx 1 - q_{\exp}(n) \approx 1 - e^{-\frac{n(n-1)}{2 \cdot 365}} \equiv p_{\exp_1}(n) \tag{1.5}$$

Numerickými hodnotami (tabulka 1.2) i graficky (obrázek 1.2) lze demonstrovat vysokou přesnost rovněž pro hrubší aproximaci (1.6) odvozenou z (1.5):

$$p_{\exp_2}(n) \equiv 1 - e^{-\frac{n^2}{2 \cdot 365}} \tag{1.6}$$

Hodnoty aproximace $p_{\exp_1}(n)$ mají zajímavou vlastnost, a to že jsou dolním odhadem hodnot $p(n)$. Snadno nahlédneme, že pro každé $n \in \{1, 2, \ldots, 366\}$ platí nerovnost $p_{\exp_1}(n) < p(n)$. Stačí si uvědomit, že pro $x \geq 0$ platí nerovnost (1.7):

$$1 - x < e^{-x} = \sum_{i=0}^{\infty} (-1)^i \frac{x^i}{i!} \tag{1.7}$$

Zřejmě nerovnost (1.7) platí triviálně pro každé $x > 1$ (neboť na levé straně vychází záporné číslo a na pravé straně kladné). Pokud $0 \leq x \leq 1$, je nerovnost (1.7) ekvivalentní s:

$$0 < \sum_{i=2}^{\infty} (-1)^i \frac{x^i}{i!} = \sum_{j=1}^{\infty} \left(\frac{x^{2j}}{(2j)!} - \frac{x^{2j+1}}{(2j+1)!} \right) \Leftarrow (\forall j \in \mathbb{N}) \left(\frac{x^{2j}}{(2j)!} > \frac{x^{2j+1}}{(2j+1)!} \right)$$

Nerovnost z pravé strany implikace lze ekvivalentně upravit na tvar $2j+1 > x$ pro $\forall j \in \mathbb{N}$, což je vzhledem k předpokládaným hodnotám $x \in [0; 1]$ platná nerovnost, tudíž i (1.7) je platná nerovnost.

Přímým důsledkem nerovnosti (1.7) je nerovnost (1.8):

$$\prod_{k=1}^{n} \left(1 - \frac{k-1}{365} \right) < \prod_{k=1}^{n} e^{-\frac{k-1}{365}} \tag{1.8}$$

Z nerovnosti (1.8) snadnou úpravou dostáváme, že:

$$1 - \prod_{k=1}^{n} e^{-\frac{k-1}{365}} < 1 - \prod_{k=1}^{n} \left(1 - \frac{k-1}{365} \right) \Leftrightarrow p_{\exp_1}(n) < p(n)$$

b) Kombinatorická aproximace

Jedná se o aproximaci, jež vychází z jednoduchého pozorování: Vybereme-li náhodně dvojici z n osob, pak jev, že tyto dvě osoby mají narozeniny v různé dny nastává s pravděpodobností $\frac{364}{365}$. Jelikož dvojici osob z n lidí můžeme vybrat celkem $\binom{n}{2}$ způsoby a pro jednotlivé dvojice předpokládáme nezávislost jevu, že tyto dvě osoby mají narozeniny v různý den, potom dle pravidla násobení by měla být pravděpodobnost toho, že žádná dvojice nebude mít narozeniny ve stejný den $\left(\frac{364}{365}\right)^{\binom{n}{2}}$. Jev komplementární (tj., že ve skupině n náhodně vybraných lidí existují aspoň dva mající narozeniny ve stejný den) musí mít pravděpodobnost (1.9):

$$1 - \left(\frac{364}{365}\right)^{\binom{n}{2}} = 1 - \left(\frac{364}{365}\right)^{\frac{n(n-1)}{2}} \equiv p_{\text{comb}}(n) \tag{1.9}$$

Právě popsaná aproximace budí dojem, že by se nemělo jednat o aproximaci, ale přímo o přesnou hodnotu pravděpodobnosti $p(n)$, leč není tomu tak. Problém se nachází v předpokladu nezávislosti pro jevy, že jednotlivé dvojice mají narozeniny v různý den. Stačí si uvědomit, že máme-li osoby A, B, C a označíme-li $A \neq B$ jev, že osoba A má narozeniny v jiný den než osoba B, pak evidentně $P(A \neq B, B \neq C, C \neq A) = \frac{364 \cdot 363}{365^2}$ není rovno $P(A \neq B) \cdot P(B \neq C) \cdot P(C \neq A) = \frac{364}{365} \cdot \frac{364}{365} \cdot \frac{364}{365}$, což je spor s nezávislostí. Jak však uvidíme, je tento zjednodušující předpoklad příčinou jen velmi malé chyby, a tedy jedná se o velmi dobrou aproximaci.

c) Poissonova aproximace

Definujme náhodnou veličinu X_n jako počet všech možný dvojic osob majících narozeniny ve stejný den ze skupiny n náhodně vybraných lidí. Náhodná veličina X_n má binomické rozdělení s parametry $\binom{n}{2}$ (tj. nabývá hodnot z množiny $\left\{0, 1, \ldots, \binom{n}{2}\right\}$) a $\frac{1}{365}$ (pravděpodobnost toho, že dva náhodně vybraní lidé mají narozeniny ve stejný den). Poissonova aproximace vychází z faktu, že limitním rozdělením $\text{Bi}(n, p)$ pro $n \to \infty$ je Poissonovo rozdělení s parametrem $\lambda \to np$ pro np konečné. Tedy v našem případě pro dostatečně velkou hodnotu $\binom{n}{2}$ nahradíme rozdělení $\text{Bi}\left(\binom{n}{2}, \frac{1}{365}\right)$ Poissonovým rozdělením s parametrem $\lambda = \frac{\binom{n}{2}}{365} = \frac{n(n-1)}{2 \cdot 365}$, čili $P(X_n = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ pro $\forall k \in \left\{0, 1, \ldots, \binom{n}{2}\right\}$. Úpravami snadno nahlédneme, že:

$$P(X_n > 0) = 1 - P(X_n = 0) = 1 - e^{-\lambda} = 1 - e^{-\frac{n(n-1)}{2 \cdot 365}} \equiv p_{\text{Po}}(n) = p_{\exp_1}(n)$$

Zřejmě Poissonova aproximace je identická s exponenciální aproximací 1. řádu uvedenou v bodě a).

Na závěr této podkapitoly se budeme věnovat srovnání kvality jednotlivých aproximací na základě jejich numerických hodnot ve vybraných bodech (tabulka (1.2)) i z grafického průběhu chyb v závislosti na počtu lidí n (obrázek (1.2)). V tabulce můžeme ihned pozorovat, že všechny hodnoty $p(n)$, $p_{\exp_1}(n)$, $p_{\exp_2}(n)$ a $p_{\text{comb}}(n)$ rychle konvergují k 1. Jelikož pravděpodobnost $p(n)$ pro $n = 366$ mění svou definiční formuli[3], hodnoty 1 je tedy přímo dosaženo. U aproximací $p_{\exp_1}(n)$, $p_{\exp_2}(n)$ a $p_{\text{comb}}(n)$ jsme jejich vzorec pro $n = 366$ neupravovali, neboť uvedené formule jsou definované pro všechna přirozená $n > 1$. Neboli aproximace $p_{\exp_1}(n)$, $p_{\exp_2}(n)$ a $p_{\text{comb}}(n)$ hodnoty 1 dosahují pouze limitně v ∞ (nicméně odchylka $p_{\exp_1}(366)$, $p_{\exp_2}(366)$ a $p_{\text{comb}}(366)$ od 1 nastává až na osmdesátém desetinném místě, čili je zanedbatelně malá.

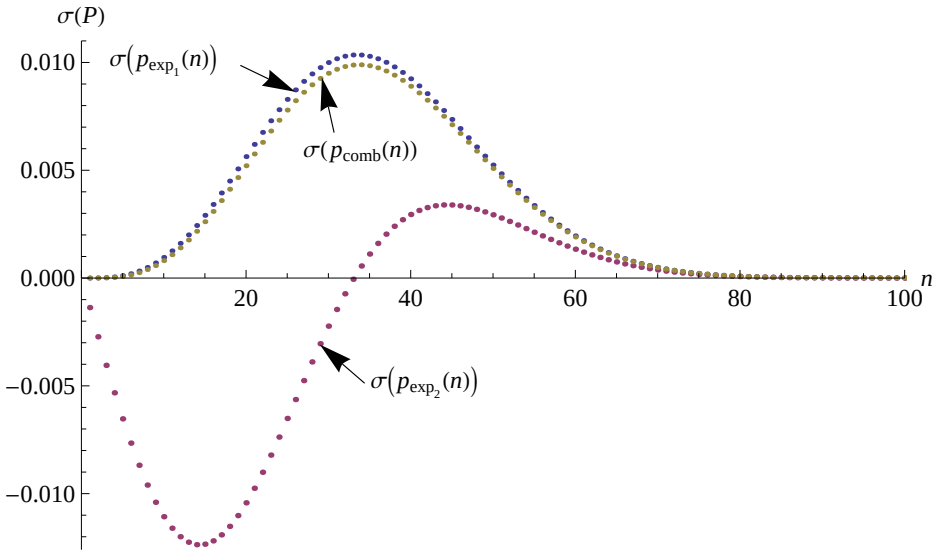
Tabulka 1.2: Srovnání numerických hodnot $p(n)$, $p_{\exp_1}(n)$, $p_{\exp_2}(n)$ a $p_{\text{comb}}(n)$

n	$p(n)$	$p_{\exp_1}(n)$	$p_{\exp_2}(n)$	$p_{\text{comb}}(n)$
10	$0,1169$	$0,1160$	$0,1280$	$0,1161$
20	$0,4114$	$0,4058$	$0,4217$	$0,4062$
23	$0,5073$	$0,5000$	$0,5155$	$0,5005$
30	$0,7063$	$0,6963$	$0,7085$	$0,6968$
40	$0,8912$	$0,8820$	$0,8883$	$0,8823$
50	$0,9704$	$0,9651$	$0,9674$	$0,9653$
60	$0,9941$	$0,9922$	$0,9928$	$0,9922$
70	$0,9992$	$0,9987$	$0,9988$	$0,9987$
80	$0,99991$	$0,9998$	$0,9999$	$0,9998$
90	$0,999994$	$0,99998$	$0,99998$	$0,99998$
100	$0,9999997$	$0,999999$	$0,999999$	$0,999999$
200	$1-1,61\cdot10^{-30}$	$1-2,10\cdot10^{-24}$	$1-1,60\cdot10^{-24}$	$1-1,95\cdot10^{-24}$
250	$1-2,29\cdot10^{-51}$	$1-9,25\cdot10^{-38}$	$1-6,57\cdot10^{-38}$	$1-8,23\cdot10^{-38}$
300	$1-6,25\cdot10^{-82}$	$1-4,32\cdot10^{-54}$	$1-2,86\cdot10^{-54}$	$1-3,65\cdot10^{-54}$
350	$1-3,03\cdot10^{-131}$	$1-2,14\cdot10^{-73}$	$1-1,32\cdot10^{-73}$	$1-1,70\cdot10^{-73}$
366	1	$1-3,34\cdot10^{-80}$	$1-2,02\cdot10^{-80}$	$1-2,60\cdot10^{-80}$

[3]neboť vzorce (1.1) a (1.2) nepřipouští dosazovat za n čísla větší jak 365, považujeme-li faktoriál ze záporného čísla za nedefinovaný

Nechť funkce σ je odchylka tak, že $\sigma(x) = p(n) - x$ pro $x \in \langle 0; 1 \rangle$, potom graf na obrázku (1.2) znázorňuje průběh funkcí $\sigma\left(p_{\exp_1}(n)\right)$, $\sigma\left(p_{\exp_2}(n)\right)$ a $\sigma\left(p_{\text{comb}}(n)\right)$, tedy odchylek (chyb) jednotlivých aproximací od skutečné hodnoty pravděpodobnosti $p(n)$ pro $n \in \{2, 3, \ldots, 100\}$. Nicméně pro $n \geq 66$ jsou hodnoty všech tří zobrazovaných odchylek menší než 0,001 a v grafu jsou vzájemně nerozlišitelné, navíc nerozlišitelné od osy n. Lze však usuzovat vzhledem ke konvergenci[4] všech tří křivek k ose n, že všechny aproximace jsou velmi přesné pro $n \in \{66, 67, \ldots, 365\}$. Pokud $n \in \{2, 3, \ldots, 65\}$, obrázek napovídá, která ze všech tří aproximací je pro dané konkrétní n nejpřesnější. Grafy $\sigma\left(p_{\exp_1}(n)\right)$, resp. $\sigma\left(p_{\text{comb}}(n)\right)$ mají velmi podobný průběh, jejich body leží na jednovrcholových křivkách s kladnými hodnotami, které svého maxima (tj. maximální kladné odchylky od skutečné pravděpodobnosti $p(n)$) dosahují pro $n = 33$ s funkční hodnotou 0,01035, resp. pro $n = 34$ s funkční hodnotou 0,0099. Tedy i v nejhorším možném případě se jedná o aproximace s chybou nejhůře okolo 0,01. Graf odchylky $\sigma\left(p_{\exp_2}(n)\right)$ leží na dvouvrcholové křivce s kladnými i zápornými hodnotami, která má dva lokální extrémy: pro $n = 14$ s funkční hodnotou -0,0124 (maximální možnou zápornou odchylku) a pro $n = 44$ s funkční hodnotou 0,0034 (maximální možná kladná odchylka), čili i tato aproximace je v nejhorším možném případě poměrně dobrá.

Obrázek 1.2: Srovnání chyb aproximací $p_{\exp_1}(n)$, $p_{\exp_2}(n)$ a $p_{\text{comb}}(n)$



[4] příslušné limity jsou vyšetřovány v následující kapitole jako (2.5), (2.6) a (2.7)

1.5 Aproximace počtu lidí při známé pravděpodobnosti

I v případě určování nejmenší možné velikosti skupiny lidí n takové, aby v této skupině byla pravděpodobnost aspoň jako předem dané p, je pro vyšší n dle vzorce dle vzorce (1.2) výpočetně náročné. Navíc tím, že hledáme jako výsledek určité přirozené číslo, není nutné takové číslo jako řešení rovnice získané ze vzorce (1.2) zjišťovat s přesností na několik desetinných míst. Pro takové účely je velmi vhodné využívat aproximace. Dvě z možných aproximací uvedeme v této podkapitole.

Budeme-li postupně upravovat vztah (1.5) z aproximace $p_{\exp_1}(n)$, resp. $p_{\exp_2}(n)$ a), dostaneme kvadratickou rovnici (1.10), resp. (1.11):

$$1 - e^{-\frac{n(n-1)}{2\cdot 365}} = p_{\exp_1}(n)$$

$$1 - p_{\exp_1}(n) = e^{-\frac{n(n-1)}{2\cdot 365}}$$

$$\ln\left[1 - p_{\exp_1}(n)\right] = -\frac{n(n-1)}{2\cdot 365}$$

$$n^2 - n + 2\cdot 365\cdot \ln\left[1 - p_{\exp_1}(n)\right] = 0 \tag{1.10}$$

$$1 - e^{-\frac{n^2}{2\cdot 365}} = p_{\exp_2}(n)$$

$$1 - p_{\exp_2}(n) = e^{-\frac{n^2}{2\cdot 365}}$$

$$\ln\left[1 - p_{\exp_2}(n)\right] = -\frac{n^2}{2\cdot 365}$$

$$n^2 + 2\cdot 365\cdot \ln\left[1 - p_{\exp_2}(n)\right] = 0 \tag{1.11}$$

Kvadratická rovnice (1.10) má kořeny:

$$n_{1,2} = \frac{1}{2} \pm \sqrt{\frac{1}{4} - 2\cdot 365\cdot \ln\left[1 - p_{\exp_1}(n)\right]}$$

Dále nahradíme $p_{\exp_1}(n)$ hodnotou p a provedeme úvahu o přípustnosti jednotlivých řešení kvadratické rovnice (1.10). Vezmeme-li v předešlém vztahu variantu s minusem a uvědomíme-li si, že na n klademe požadavek celočíselnosti, vycházelo by nám $n \approx 0$, tudíž přípustným řešením bude pouze varianta s plusem. Hodnota aproximace počtu lidí n nutných k dosažení úrovně zadané hodnoty pravděpodobnosti p je tedy dána vzorcem:

$$n \approx \frac{1}{2} + \sqrt{\frac{1}{4} - 2 \cdot 365 \cdot \ln(1-p)} \tag{1.12}$$

Obdobnou úvahou jako v předešlém odstavci získáme hrubší aproximaci počtu lidí n nutných k dosažení úrovně zadané hodnoty pravděpodobnosti p:

$$n \approx \sqrt{-2 \cdot 365 \cdot \ln(1-p)} \tag{1.13}$$

Jaká je přesnost uvedených aproximací a jak je s nimi třeba zacházet? Proveďme následující jednoduchý test na hodnoty získané aproximací (1.12): Dosazujme do vzorce (1.12) postupně hodnoty $p(2), p(3), \ldots, p(100)$ a sledujme, jak se bude vyvíjet hodnota $n^*\left(p(n)\right) = \left\lfloor \frac{1}{2} + \sqrt{\frac{1}{4} - 2 \cdot 365 \cdot \ln\left[1-p(n)\right]} \right\rfloor$ (neboli aproximované n zaokrouhlené směrem dolů, neboť není možné rozumně interpretovat např. jednu desetinu člověka). Při numerickém výpočtu v softwaru *Mathematica* pro $n \in \{2, 3, \ldots, 46\}$ vždy dostáváme $n^*\left(p(n)\right) = n$, poté však $n^*\left(p(47)\right) = 48$. Dále pro $n \in \{47, 48, \ldots, 64\}$ vždy platí $n^*\left(p(n)\right) = n + 1$, nicméně $n^*\left(p(65)\right) = 67$, atd.

Trend, kdy se postupně zvyšuje hodnota aproximovaného n^* oproti skutečné hodnotě n, stále pokračuje a každé navýšení o jednotku je čím dál tím častější. Čím je toto způsobeno? Totiž diference $p(n+1) - p(n)$ se zmenšují rychleji, než klesá funkce $\sigma\left(p_{\exp_1}(n)\right)$, což má za následek, že v určitých hodnotách n bude platit $p(n) = p_{\exp_1}(n) + \sigma\left(p_{\exp_1}(n)\right) \doteq p_{\exp_1}(n+k)$, kde přirozené číslo k bude splňovat $n^*(p(n)) = n + k$. Dva konkrétní příklady tohoto problému jsme uvedli v předešlém odstavci, stačí si totiž povšimnout, že $p_{\exp_1}(47) = 0,9483$, $\sigma\left(p_{\exp_1}(47)\right) = 0,0065$, $p_{\exp_1}(48) = 0,9545$ a $p_{\exp_1}(47) + \sigma\left(p_{\exp_1}(47)\right) \doteq 0,9548 \doteq 0,9549 \doteq p_{\exp_1}(48)$, rovněž také $p_{\exp_1}(65) = 0,9966$, $\sigma\left(p_{\exp_1}(65)\right) = 0,0010$, $p_{\exp_1}(67) = 0,9977$ a $p_{\exp_1}(65) + {} + \sigma\left(p_{\exp_1}(65)\right) \doteq 0,9976 \doteq 0,9977 \doteq p_{\exp_1}(67)$. Pro další ilustraci tohoto uvádíme tabulku, která srovnává diference a odchylky pro n od 90 do 99 (tabulka (1.3)):

Tabulka 1.3: Srovnání diferencí $p(n)$ a odchylek $\sigma\left(p_{\exp_1}(n)\right)$

n	$p(n)$	$p(n+1)-p(n)$	$\sigma\left(p_{\exp_1}(n)\right)$	$n^*\left(p(n)\right)$
90	$0,999994$	$1,52 \cdot 10^{-6}$	$0,000011$	94
91	$0,999995$	$1,16 \cdot 10^{-6}$	$8,78 \cdot 10^{-6}$	95
92	$0,999997$	$8,77 \cdot 10^{-7}$	$6,98 \cdot 10^{-6}$	96
93	$0,999997$	$6,63 \cdot 10^{-7}$	$5,52 \cdot 10^{-6}$	97
94	$0,999998$	$4,99 \cdot 10^{-7}$	$4,36 \cdot 10^{-6}$	98
95	$0,999999$	$3,75 \cdot 10^{-7}$	$3,43 \cdot 10^{-6}$	99
96	$0,999999$	$2,80 \cdot 10^{-7}$	$2,69 \cdot 10^{-6}$	100
97	$0,999999$	$2,09 \cdot 10^{-7}$	$2,10 \cdot 10^{-6}$	101
98	$0,999999$	$1,55 \cdot 10^{-7}$	$1,63 \cdot 10^{-6}$	102
99	$0,9999996$	$1,14 \cdot 10^{-7}$	$1,27 \cdot 10^{-6}$	104

Závěrem lze říci, že aproximace (1.12) dává přesné hodnoty pro pravděpodobnosti menší jak 0,95. Chceme-li aproximovat tímto způsobem pro některou z hodnot velmi blízkou 1, musíme se spokojit s tím, že nám aproximace dá orientační výsledek, od nějž bude třeba až několik jednotek odečíst, abychom se dostali k hledané hodnotě n. Je však třeba docenit, že pokud bychom neměli k dispozici aproximaci (1.12), museli bychom při známé hodnotě pravděpodobnosti $p(n)$ (ne příliš blízko u 1) dostávat příslušné n ze vztahu (1.2), tj. řešením rovnice $p(n) = 1-\frac{365!}{365^n(365-n)!}$ v neznámé n, což analyticky při obecném n nelze provést a numericky můžeme dostat pouze lepší či horší aproximace.

Kapitola 2

Přihrádkový problém s předpokladem rovnoměrnosti

Narozeninový problém lze zobecňovat v několika směrech. Jednou cestou zobecnění problému je uvolnit předpoklad o 365 dnech v roce a hledat pravděpodobnost toho, že při umisťování n předmětů do m přihrádek v některé přihrádce budou alespoň dva předměty, pak evidentně při volbě $m = 365$ dostaneme narozeninový problém studovaný v první kapitole. Jinou cestou zobecnění je uvolnit předpoklad o stejné pravděpodobnosti narození v libovolném dnu v roce, totiž můžeme například každému měsíci, týdnu případně dnu (v tom nejobecnějším případě) přiřadit jinou (skutečnou) pravděpodobnost narození. V této kapitole se budeme zabývat prvním typem zobecnění, v kapitolách následujících prozkoumáme druhou variantu zobecnění.

2.1 Pravděpodobnost shody v obecném přihrádkovém problému

Provedeme-li úvahu založenou na Dirichletově principu analogicky s podkapitolou 1.2, v níž všude nahradíme 365 obecným m a místo $p(n)$, resp. $q(n)$ budeme psát $p(m, n)$, resp. $q(m, n)$, bude platit, že pro $n > m$, je $q(m, n) = 0$ a $p(m, n) = 1$ a v případě, že $2 \leq n \leq m$, dostaneme:

$$q(m,n) = \prod_{k=2}^{n} \frac{m-k+1}{m} = \prod_{k=1}^{n} \frac{m-k+1}{m} = \frac{m!}{m^n(m-n)!} \qquad (2.1)$$

$$p(m,n) = 1 - q(m,n) = 1 - \frac{m!}{m^n(m-n)!} \tag{2.2}$$

2.2 Exponenciální aproximace

Ze stejných důvodů jako v předešlé kapitole, i v přihrádkovém problému je možné a za určitých okolností i výhodné uvažovat aproximaci pravděpodobnosti $p(m,n)$ pravděpodobností $p_{\exp}(m,n) = 1 - e^{-\frac{n(n-1)}{2m}}$. Viděli jsme, že pro $m = 365$ je exponenciální aproximace velmi přesná a vhodná pro určení minimálního počtu lidí, který by uspokojil požadovanou pravděpodobnost shody v narozeninovém problému. Jak je to s kvalitou exponenciální aproximace pro různé hodnoty n a m?

V podkapitole 1.4 jsme ukazáli, že $p_{\exp_1}(n) < p(n)$, analogicky nahlédneme, že i $p_{\exp}(m,n) < p(m,n)$ pro $n \in \{2,3,\ldots,m\}$. Označme:

$$A(m,n) \equiv p(m,n) - p_{\exp}(m,n) = e^{-\frac{n(n-1)}{2m}} - \frac{m!}{m^n(m-n)!} = e^{-\frac{n(n-1)}{2m}} - \frac{\Gamma(m+1)}{m^n\Gamma(m-n+1)}$$

Předpokládejme, že jsme dodefinovali funkci A předpisem pomocí gamma funkce[1] na všechna reálná čísla splňující $2 \leq n \leq m$, potom takto definovaná funkce je zřejmě spojitá[2]. Chceme-li určit kvalitu dané aproximace, musíme být schopni vyšetřit průbeh funkce $A(m,n)$, což je v obecné rovině velmi náročný problém. Derivace $\frac{\partial A(m,n)}{\partial m}$ a $\frac{\partial A(m,n)}{\partial n}$ vedou na funkce, jejichž kořeny ve vší obecnosti nebudeme schopni vyjádřit. Spočítáme první parciální derivace $A(m,n)$ dle m i n:

$$\frac{\partial A(m,n)}{\partial m} = \frac{(n-1)n}{2m^2} e^{-\frac{(n-1)n}{2m}} + \frac{\Gamma(m)\left(mH_{m-n} - mH_m + n\right)}{m^n\Gamma(m-n+1)} \tag{2.3}$$

$$\frac{\partial A(m,n)}{\partial n} = \frac{(1-2n)}{2m} e^{-\frac{(n-1)n}{2m}} + \frac{\Gamma(m+1)\left(-H_{m-n} + \ln(m) + \gamma\right)}{m^n\Gamma(m-n+1)}, \tag{2.4}$$

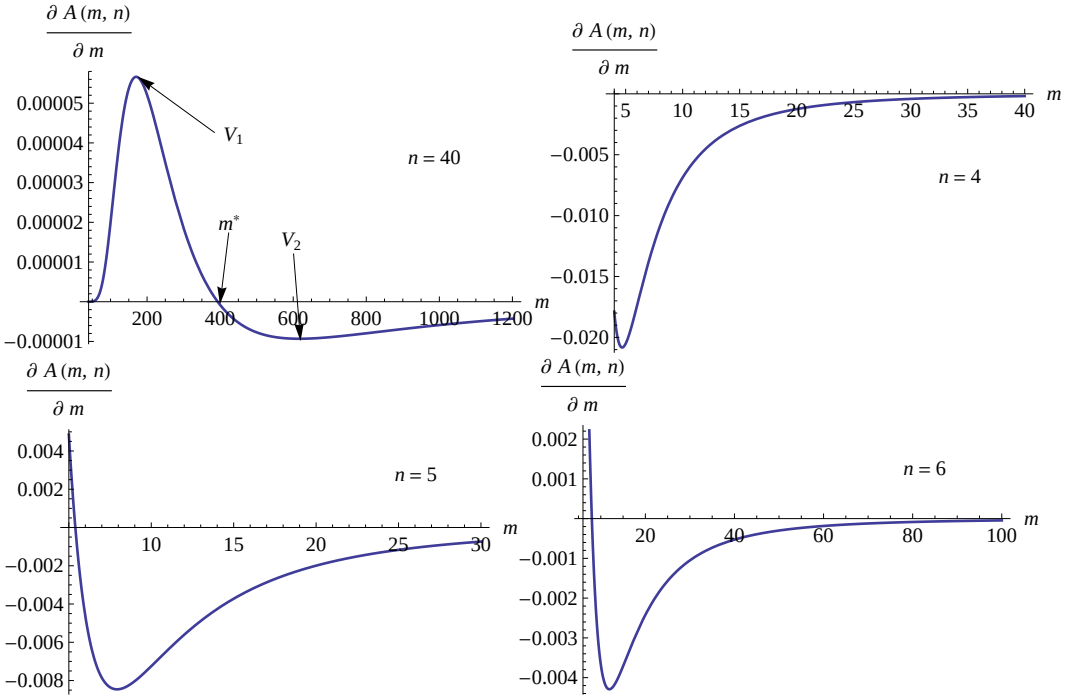
[1] Jak známo, faktoriál lze přepsat pomocí gamma funkce jako $\Gamma(k) = (k-1)!$ pro všechna přirozená k (definiční obor gamma funkce si navíc s pomocí komplexní analýzy rozšíříme na všechna nezáporná reálná čísla).

[2] Vzhledem k požadavku $2 \leq n \leq m$ jsou totiž vyloučeny případné body nespojitosti, které by existovaly při definici A na celém $\mathbb{R}^2$.

kde $H_k = \sum_{i=1}^{k} \frac{1}{i}$ je příslušným harmonickým číslem (pro nějž rovněž připouštíme existenci zobecnění na nezáporných reálných číslech (viz. např. [8]) a γ je tzv. *Euler-Mascheroniho konstanta* s přibližnou hodnotou 0,5772156649.

Určení průběhu funkce $A(m, n)$ čistě analyticky je netriviální, pro klíčové nulové body a intervaly, kde je funkce $\frac{\partial A(m,n)}{\partial m}$ vzhledem k proměnné m a $\frac{\partial A(m,n)}{\partial n}$ vzhledem k n kladná či záporná, se nám nepodaří nalézt rozumné obecné explicitní vyjádření. Provedeme však určitá pozorování na základě grafů těchto funkcí pro přípustné hodnoty n a m, které splňují $2 \leq n \leq m$.

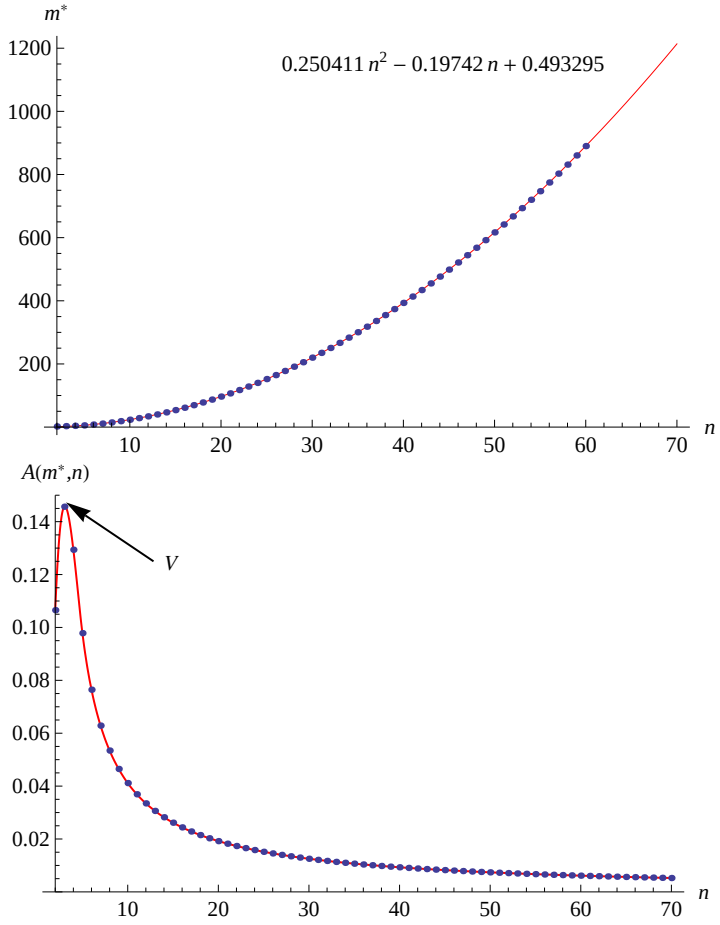
Obrázek 2.1: Průběh $\frac{\partial A(m,n)}{\partial m}$ pro různá n



Na prvním grafu na obrázku (2.1) vidíme průběh $\frac{\partial A(m,n)}{\partial m}$ typický pro většinu přirozených n (kromě malých hodnot). Bude se jednat o dvouvrcholovou křivku, pro kterou při $m = n$ hodnota $\frac{\partial A(m,n)}{\partial m}$ zprava konverguje k 0, dále funkce roste až do jistého vrcholu V_1 (lokální a zároveň i globální maximum), z nějž klesá, protíná osu m v jistém bodě m^* a následně klesá až do určitého vrcholu V_2, z nějž pak dále roste a asymptoticky se přibližuje k ose

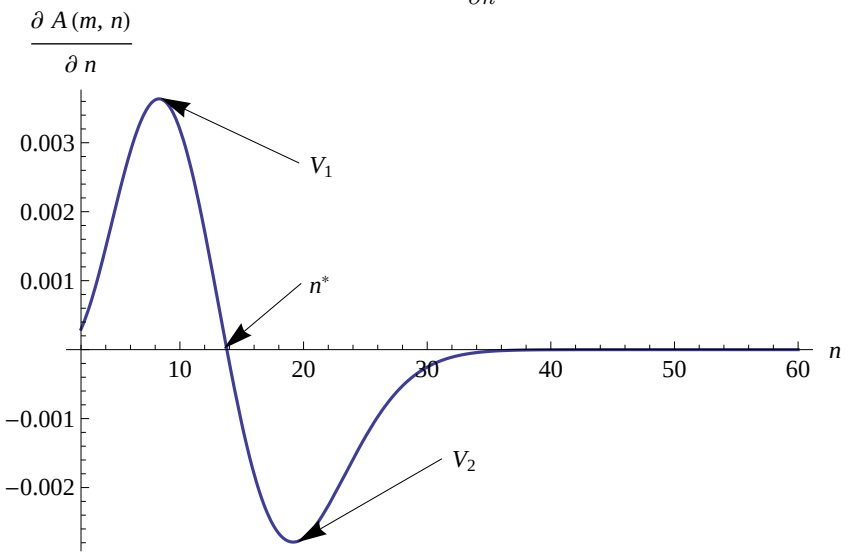
m. Na zbylých grafech vidíme odlišné situace pro malá n. Pro $n = 4$ neexistuje nulový bod m^*, pro $n = 5, 6$ již tento bod existuje, nicméně na grafech se neobjevuje vrchol V_1 (bude patrný právě s rostoucím n). Při vyšším n již zůstanou výše popsané vlastnosti křivky zachovány, s tím, že různé hodnoty (m, n) budou způsobovat různé souřadnice obou vrcholů i nulového bodu, jehož souřadnice vyšetřujeme v dalším odstavci. Zřejmě nyní $\frac{\partial A(m,n)}{\partial m} > 0$ na $(2, m^*)$, a tedy funkce $A(m, n)$ je rostoucí vzhledem k m, dále $\frac{\partial A(m,n)}{\partial m} < 0$ na (m^*, ∞), tudíž funkce $A(m, n)$ je klesající vzhledem k m a bod m^* je bodem lokálního maxima. V případě, že neexistuje bod m^*, je funkce $A(m, n)$ klesající vzhledem k m v celém uvažovaném definičním oboru.

Obrázek 2.2: Hodnoty m^* a $A(m^*, n)$ pro různá n



První graf na obrázku (2.2) ukazuje vývoj hodnot m^* (neboli hodnot maximalizujících $A(m,n)$ pro dané n). Navíc jsme tyto diskrétní hodnoty metodou nejmenších čtverců s využitím softwaru *Mathematica* proložili parabolou, jejíž rovnice je uvedena na obrázku a která nám může posloužit pro přibližné určení m^* při vyšším n[3]. Následující graf příslušné maximální hodnoty $A(m^*,n)$ (tedy maximální možné nepřesnosti exponenciální aproximace) vyčísluje. Všimněme si, že body grafu $A(m^*,n)$ leží na jednovrcholové křivce s vrcholem V (je maximem), přitom funkce $A(m^*,n)$ do vrcholu (v okolí bodu $n=3$ s funkční hodnotou 0,1457) roste a dále pak klesá a asymptoticky se přimyká k ose n (tento fakt později zdůvodníme). Je třeba si ještě uvědomit, že uvedeným postupem nemusí být m^* celé číslo, tedy maximalizujeme-li $A(m,n)$ při pevném n na oboru přirozených čísel, bude maxima dosaženo v okolí bodu m^*, tj. v hodnotě $\lfloor m^* \rfloor$ nebo $\lceil m^* \rceil$. V každém případě pro všechna přípustná přirozená m bude $A(m,n) \leq A(m^*,n)$. Konečně $A(m^*,70) = 0,0052$, tedy pro libovolné přípustné přirozené m je nepřesnost exponenciální aproximace při $n=70$ v nejhorším případě nejvýše 0,0052 a pro vyšší n se tato chyba dále snižuje vzhledem ke konvergenci $A(m^*,n)$ k 0.

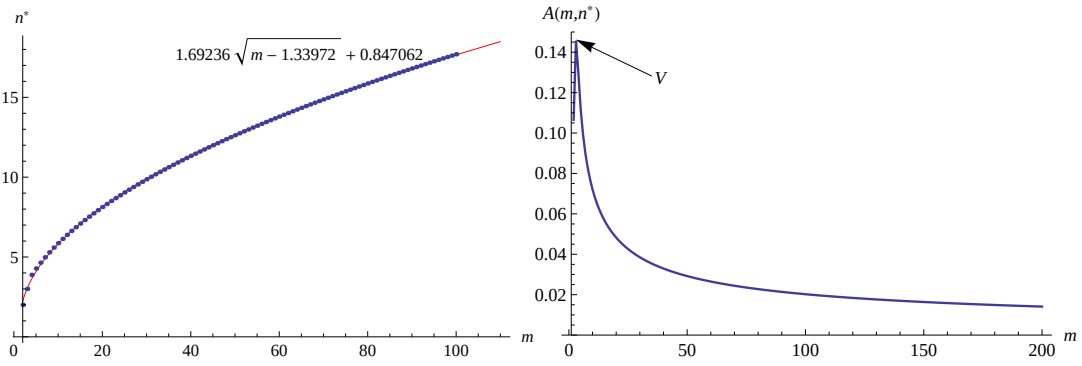
Obrázek 2.3: Průběh $\frac{\partial A(m,n)}{\partial n}$ pro $m = 60$



[3]Pokud budeme prokládat body m^* pro $n \in \{2,3,\ldots,1000\}$, dostaneme rovnici proložení $0.250004n^2 - 0.168906n + 0.172561$, čili vidíme, že koeficient (nejvýznamnějšího) kvadratického členu s rostoucím n bude nejspíš konvergovat k 0,25, absolutní i lineární člen se budou dále přibližovat k 0 a jejich význam bude mnohem menší.

Na obrázku (2.3) můžeme sledovat průběh $\frac{\partial A(m,n)}{\partial n}$ typický pro všechna přirozená $m > 2$. Bude se jednat o dvouvrcholovou křivku, pro kterou při $n = 2$ hodnota $\frac{\partial A(m,n)}{\partial n}$ zprava konverguje k 0, dále funkce roste až do jistého vrcholu V_1 (lokální a zároveň globální maximum), z nějž klesá, protíná osu n v jistém bodě n^* a dále klesá až do určitého vrcholu V_2, ze kterého následně roste a asymptoticky se přibližuje k ose n. Různé hodnoty (n, m) budou způsobovat různé souřadnice obou vrcholů i nulového bodu, které vyšetříme v následujícím odstavci. Konečně platí $\frac{\partial A(m,n)}{\partial n} > 0$ na $(2, n^*)$, a tedy funkce $A(m, n)$ je rostoucí vzhledem k n, dále $\frac{\partial A(m,n)}{\partial n} < 0$ na $(n^*, m\rangle$, tudíž funkce $A(m, n)$ je klesající vzhledem k n a bod n^* je bodem lokálního maxima.

Obrázek 2.4: Hodnoty n^* a $A(m, n^*)$ pro různá m



První graf na obrázku (2.4) ukazuje vývoj hodnot n^* (neboli hodnot maximalizujících $A(m, n)$ pro dané m). Opět jsme tyto diskrétní hodnoty metodou nejmenších čtverců za pomoci softwaru *Mathematica* proložili křivkou inverzní k jedné větvi paraboly, jejíž rovnice je uvedena na obrázku a která nám může posloužit pro přibližné určení n^* při vyšším m[4]. Následující grafy tyto maximální hodnoty $A(m, n^*)$ (tedy maximální možné nepřesnosti exponenciální aproximace) vyčíslují. Můžeme si povšimnout, že body grafu $A(m, n^*)$ leží na jednovrcholové křivce s vrcholem V (je maximem), přitom funkce $A(m, n^*)$ do vrcholu (v okolí bodu $m = 3$ s funkční hodnotou 0,1457) roste a dále pak klesá a asymptoticky se přimyká k ose m, ale pomaleji než se

[4]Pokud budeme prokládat body n^* pro $m \in \{2, 3, \ldots, 1000\}$, dostaneme rovnici proložení $1,72994\sqrt{m - 0,511614} + 0,447477$, čili vidíme, že nejvýznamnější koeficient před odmocninou má tendenci pozvolna růst s rostoucím n, naopak vliv absolutního členu se postupně eliminuje.

přimykala $A(m^*, n)$ k ose n. Je třeba si ještě uvědomit, že uvedeným postupem nemusí být n^* celé číslo, neboli při maximalizaci $A(m, n)$ pro pevné m na oboru přirozených čísel bude maxima dosaženo v okolí bodu n^*, tj. v hodnotě $\lfloor n^* \rfloor$ nebo $\lceil n^* \rceil$. Každopádně pro všechna přípustná přirozená n bude $A(m, n) \leq A(m, n^*)$. Konečně $A(1000, n^*) = 0,0062$, tedy pro libovolné přípustné přirozené m je nepřesnost exponenciální aproximace při $m = 1000$ v nejhorším případě nejvýše 0,0052 a pro vyšší m se tato chyba dále snižuje. Porovnáme-li tento výsledek se situací (2.2), vidíme, že pro zajištění srovnatelně vysoké garantované maximální chyby aproximace musíme vzít m výrazně vyšší než stačilo brát $n = 70$. Proto při volbách nižších hodnot m přesnost exponenciální aproximace klíčově závisí na hodnotě n, kdy při nešikovné kombinaci (m, n) se může nepřesnost pohybovat až v řádu desetin, což je nezanedbatelná hodnota.

Jak jsme si mohli povšimnout v předcházející analýze funkce $A(m, n)$, problematické jsou malé hodnoty m a n, pro něž se funkce nechová monotónně. Nicméně jak dále uvídíme z jejích asymptotických vlastností, od určitých hodnot $m_0, n_0 \in \mathbb{N}$ funkce $A(m, n)$ asymptoticky klesá k 0, což zaručí žádoucí přesnost dané exponenciální aproximace. Nuže určeme $\lim\limits_{m\to\infty} A(m, n)$, $\lim\limits_{n\to m} A(m, n)$ a $\lim\limits_{n\to m, m\to\infty} A(m, n)$:

$$\lim_{m\to\infty} A(m,n) = \lim_{m\to\infty} e^{-\frac{n(n-1)}{2m}} - \lim_{m\to\infty} \prod_{k=1}^{n} \left(1 - \frac{k-1}{m}\right) = e^0 - 1^n = 0 \quad (2.5)$$

$$\lim_{n\to m} A(m,n) = \lim_{n\to m} e^{-\frac{n(n-1)}{2m}} - \lim_{n\to m} \frac{m!}{m^n(m-n)!} = e^{-\frac{m-1}{2}} - \frac{m!}{m^m} \quad (2.6)$$

$$\lim_{m\to\infty} \left(e^{-\frac{m-1}{2}} - \frac{m!}{m^m}\right) = e^{-\infty} - \lim_{m\to\infty} \frac{\sqrt{2\pi m}\left(\frac{m}{e}\right)^m (1 + o(m))}{m^m} = 0 - \lim_{m\to\infty} \frac{\sqrt{2\pi m}}{e^m} = 0 \quad (2.7)$$

Dodejme, že ve vzorci (2.7) využíváme tzv. *Stirlingův rozvoj* pro faktoriál $m!$.

Konečně uveďme, že analogicky by bylo možné uvážit inverzní aproximaci pro počet lidí (předmětů) n, aby pravděpodobnost umístění alespoň dvojice předmětů do jedné přihrádky byla aspoň p, a to dosazením m na místo 365 v aproximacích (1.12) a (1.13) spočtených v podkapitole 1.4:

$$n \approx \frac{1}{2} + \sqrt{\frac{1}{4} - 2m \cdot \ln(1-p)} \tag{2.8}$$

$$n \approx \sqrt{-2m \cdot \ln(1-p)} \tag{2.9}$$

2.3 Význam v kryptografii a hashování

Chceme-li ukládat určitý (velký) objem dat do nějaké počítačové databáze a pak v ní i efektivně vyhledávat, sekvenční ukládání je díky své časové náročnosti na prohledávání nevyhovující a přímý přístup, při kterém známe přesný index záznamu, který takto nalezneme bez jakéhokoli vyhledávání, má zase vysoké paměťové nároky. V informatice je snaha snižovat časovou složitost i paměťové nároky všech procesů, proto je tento problém uvnitř počítače často řešen hashováním (rozptylováním), které je kompromisní variantou mezi sekvenčním a přímým přístupem.

Při hashování je každému vstupnímu datu na základě jistého předpisu (hashovací funkce) přiřazen index určující jeho pozici v databázi. Pokud následně máme za úkol takový záznam nalézt, stačí nám znát jeho index a použitou hashovací funkci. Platí, že objem dat je větší než kapacita databáze a souběžně (i jako důsledek), že hashovací funkce[5] není prostá, což vede k tzv. *kolizím*. Kolize nastává, když předpis hashovací funkce při zařazování n-tého záznamu ukáže na pozici, kde se již nějaký záznam nachází. Významným cílem při hashování je co možná nejvíce takovým kolizím zabránit, tj. maximálně snížit jejich pravděpodobnost[6], avšak určit pravděpodobnost takové kolize není ničím jiným než aplikovat poznatky získané o přihrádkovém problém v předešlé podkapitole.

Hashování se často používá v kryptografii jako jedna z možných forem zabezpečení. *Narozeninový paradox*, diskutovaný v podkapitole 1.3, má právě v kryptografii zásadní dopad: totiž tzv. *útoky hrubou silou* (tj. pokusy o prolomení hashovací funkce vyhledáváním dvojice kolidujících dat umístěných na stejnou pozici v databázi prostým zkoušením všech možností) jsou díky

[5]Měl by to být takový předpis, který bude data "dobře" rozptylovat, tj. bude mít co nejnižší pravděpodobnost kolize (že k jedné možné pozici v databázi přiřadí více záznamů), jednoduchými příklady takové hashovací funkce jsou např. funkce $h(n) \equiv n \bmod p$, kde p je nějaké vhodně zvolené prvočíslo vzhledem k velikosti databáze.

[6]U praktických hashovacích funkcích však nikdy nemůže být nulová, všechny uvažované hashovací funkce jsou přirozeně kolizní.

narozeninovému paradoxu snazší, než by se na první pohled mohlo zdát. Pro k-bitovou hashovací funkci (zřejmě umožňuje rozptylování až do 2^k různých pozic databáze) nastává kolize s nezanedbatelnou pravděpodobností přibližně 0,5 již v množině obsahující $2^{\frac{k}{2}}\sqrt{2 \cdot \ln 2} \approx 1,1774 \cdot 2^{\frac{k}{2}}$ dat (výsledek pro n získaný dosazením $p = 0,5$ a $m = 2^k$ do vztahu (2.9)), namísto očekávaných $\frac{1}{2} \cdot 2^k$ dat (výsledek plynoucí z intuitivní úvahy, že v polovině hashovaných dat musí být pravděpodobnost kolize přibližně 0,5). Tedy po prohledání $1,1774 \cdot 2^{\frac{k}{2}}$ dat již máme přibližně 50% šanci, že nalezneme kolidující dvojici, přitom číslo $2^{\frac{k}{2}}$ je vzhledem k bezpečnostním požadavkům pro malá k poměrně malé[7], což klade jistý nárok na délku hashovací funkce (tj. nejen na její kvalitu). Samotné hledání kolizních stavů je klíčové při autentizaci nebo autorizaci dat, které se používají v elektronickém podpisu atd.

Na závěr kapitoly zmíníme dva konkrétní příklady aplikace hashování: pro kontrolu zpráv odesílaných po síti a pro uchovávání hesel. Při posílání zpráv po sítí většinou dochází k tomu, že odesílaná zpráva vstoupí do jistého hashovacího algoritmu, jehož výsledkem je tvz. *message digest*, poté se vhodně zašifruje, připojí se k ní *message digest* a v této podobě se odešle adresátovi. Adresát zprávu pomocí sdíleného klíče dešifruje, provede hashování a výsledný *message digest* porovná s tím, který byl přiložen ke zprávě odesílatelem. Shodují-li se oba dva *message digest*, znamená to, že zpráva nebyla během přenosu nijak změněna. Hashování při uchovávání hesel má význam pro jejich ochranu před zneužitím, totiž z bezpečnostních důvodů není vhodné uchovávat hesla v jejich zdrojovém tvaru, proto se uchovává pouze výsledek, který nám poskytne hashovací funkce. Vzhledem k povaze hashovacího algoritmu (vhodně zvoleného) bude nesmírně složité z výsledku hashovací funkce odvodit původní heslo, nicméně pro operační systém bude naopak velice snadné podle výsledků hashování ověřit správnost hesla.

[7]Pro dosažení uspokojivé bezpečnosti dané hashovací funkce se obvykle volí $k \geq 256$.

Kapitola 3

Narozeninový problém v případě nerovnoměrného rozdělení

V této kapitole se budeme zabývat narozeninovým problémem v klasické podobě, v jaké jsme se s ním seznámili v první kapitole. Budeme dále využívat předpoklad o neexistenci přestupných let, ale uvolníme předpoklad o rovnoměrném rozdělení pravděpodobností narození v jednotlivých dnech během roku, což odpovídá skutečné situaci. Na kolik je předpoklad rovnoměrného rozdělení poškozující? Poskytuje dobrou aproximaci reálné situace? Jaký je vztah mezi pravděpodobností určenou v narozeninovém problému s a bez předpokladu rovnoměrnosti?

3.1 Příčiny nerovnoměrností v reálné distribuci porodů během roku

Pro zkoumání narozeninového problému bez předpokladu rovnoměrnosti je potřeba vzít v úvahu tzv. sezónnost porodů i různé cykličnosti, které je možné v distribuci porodů sledovat. Jsou totiž nejběžnějšími původci nerovnoměrností v rozdělení pravděpodobnosti narození v daný den během roku. Nerovnoměrnosti však vyvolávají i politická rozhodnutí (např. jedná-li se o zavádění či změny dávek pro matky, narozené děti apod.) i řada událostí nepřírodního charakteru.

> Četnost narození je v jednotlivých měsících a obdobích roku rozdílná. Rozložení počtu narozených dětí, resp. koncepcí během kalendářního roku souvisí s biologicko-klimatickými podmíněnostmi a životním

stylem obyvatelstva, proto se z dlouhodobého pohledu - nebere-li se v úvahu náhodné kolísání - mění jen málo a pozvolna. Kromě toho není sezónnost pro celkový charakter reprodukce nijak významná, v demografii je proto spíše okrajovým tématem.

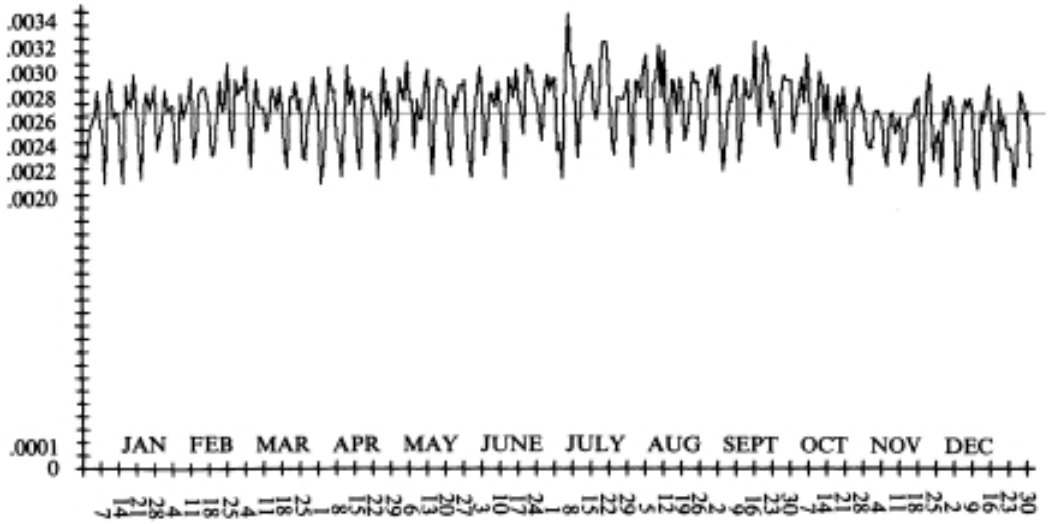
V České republice se pravidelně vyšší počet dětí rodí v jarních a letních měsících, zatímco na podzim a v první polovině zimy jsou počty narozených podprůměrné. V období 1993 - 2005 došlo k mírnému snížení měsíčních indexů[1] v první čtvrtině roku a naopak k nárůstu indexů v posledním čtvrtletí, čímž se těžiště porodnosti posunulo z období únor-červenec na duben-září[2]. Zároveň se však rozdíly v počtu narozených v jednotlivých měsících poněkud zmenšily. [1]

Český statistický úřad se však při evidování porodů během roku zabývá nejvýše měsíční sezónností porodů, a tedy neregistruje počty porodů pro jednotlivé dny v roce. Z důvodu absence potřebných dat pro ČR použijeme pro naše účely studii [2] Geoffrey C. Berresforda z Univerzity Long Island (USA), který srovnával pravděpodobnost narození $\frac{1}{365}$ za předpokladu rovnoměrného rozdělení s relativními četnostmi porodů pro každý den v roce 1977 vypočtenými na základě dat ze státu New York v témže roce. V této studii bylo zjištěno, že se relativní četnosti pro jednotlivé dny během roku pohybují od 0,002135 (pro neděli, 11. prosince 1977) do 0,003478 (pro středu, 6. července 1977), což představuje odchylku od teoretické pravděpodobnosti $\frac{1}{365} = 0,002740$ až o 27%.

[1] Jedná se o tzv. měsíční indexy porodnosti, které se získají následujícím postupem: Nejprve se výchozí absolutní měsíční počty narozených standardizují, tj. očistí se od rozdílné délky jednotlivých měsíců. Měsíční indexy jsou poté vypočteny jako poměr standardizovaného počtu narozených v daném měsíci k průměrnému měsíčnímu počtu narozených v daném roce. Hodnota indexu pro průměrný měsíc v roce je tedy rovna jedné.

[2] tyto závěry potvrzují i obrázky 3.1 a 3.3, které se zabývají sezónností porodů ve státě New York (USA)

Obrázek 3.1: Rozložení relativních četností porodů ve státě New York (USA) během roku 1977

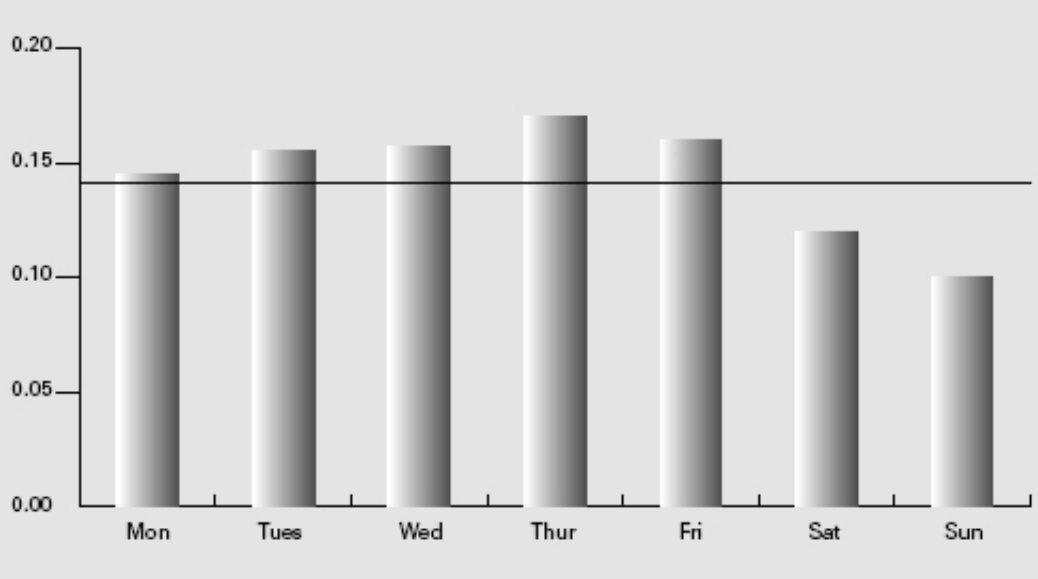


Na obrázku 3.1, jenž pochází z Berresfordovy studie, je vodorovnou linkou naznačena úroveň pravděpodobnosti za předpokladu rovnoměrného rozdělení porodů během roku. Obrázek jasně ukazuje značné kolísání skutečných relativních četností kolem teoretické pravděpodobnosti $\frac{1}{365}$ v celém sledovaném období a navíc napovídá, že v rozložení porodů bude hrát pravděpodobně roli jistý týdenní cyklus, jelikož trend výkyvů se periodicky opakuje přibližně každých 7 dní.

Joshua Gans a Andrew Leigh ve svém článku [3] hypotézu o existenci týdenního cyklu v porodech potvrzují a zkoumají příčiny tohoto jevu. Platí totiž, že o víkendu se rodí méně dětí než v pracovních dnech, nejvíce dětí se rodí ve čtvrtek, nejméně pak v neděli. Příčinou nerovnoměrného rozložení porodů během týdne je zřejmě možnost plánování si doby porodu (např. vyvoláním předčasného porodu apod.). Na nižším podílu porodů během víkendu se tak může podepsat i pracovní doba a s ní související ekonomické dopady pro chod porodnických zařízení. I personál těchto zařízení má standardní pracovní dobu od pondělí do pátku. O víkendu, kdy je provoz takových zařízení omezenější, navíc musejí zaměstnanci porodnic dostávat příplatky ke mzdě, což zvyšuje osobní náklady nemocnice. Obrázek 3.2 zachycuje rozložení porodů na jednotlivé dny v týdnu, byl sestaven na základě dat z Australského statistického úřadu z let 1975 až 2003. Dodejme, že vodorovná linka zachycuje úroveň teoretické pravděpodobnosti $\frac{1}{7}$ při rozdělení porodů rovnoměrně

mezi všech 7 dní.

Obrázek 3.2: Rozložení relativních četností porodů v různých dnech v týdnu v Austrálii za odbobí let 1975-2003

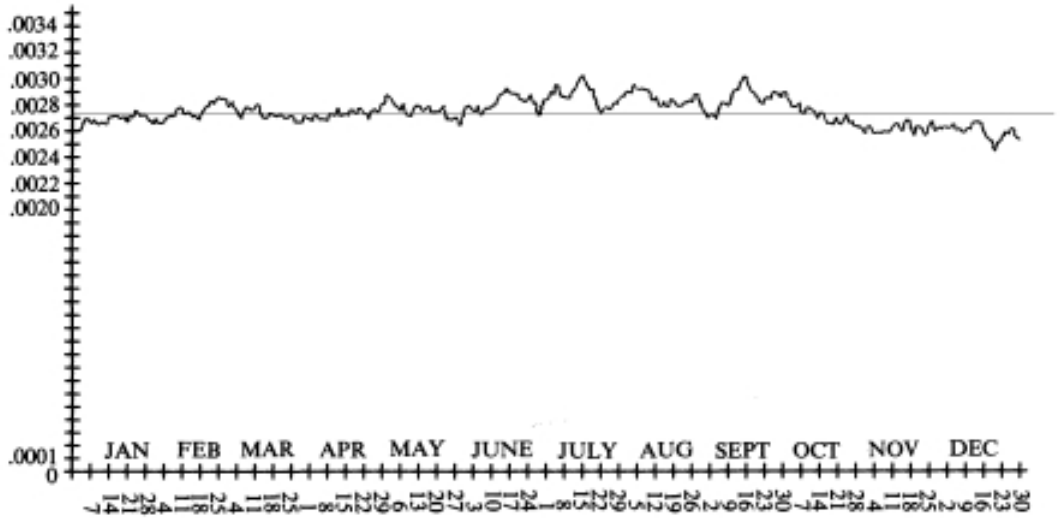


Zmiňovaná týdenní cykličnost v porodech má jistý dopad na rovnoměrnost distribuce porodů během jednoho roku, již bychom rádi předpokládali: Způsobuje tak významné kolísání, že bychom výsledky získané aproximací reálného problému problémem s předpokladem rovnoměrnosti z první kapitoly nemohli považovat za dostatečně spolehlivé. Ovšem je ve všech populacích týdenní cykličnost porodů relevantní? Snadno dovodíme, že nikoli. Předešlé úvahy narážejí na ten nedostatek, že jsme zkoumali data jen z jednoho jediného roku. Uvažujeme-li tedy narozeninový problém v populaci jednoho ročníku, opravdu se nelze příliš spolehlivě opírat o výsledky z první kapitoly. Navíc to vysvětluje, že pro případ školních tříd (žáci ze stejné třídy jsou obvykle jedním a tím samým populačním ročníkem), není vhodné kvůli týdenní cykličnosti porodů uvažovat zjednodušení tohoto problému na narozeninový problém s předpokladem rovnoměrnosti řešený v první kapitole.

Populace, ve kterých lze skutečný narozeninový problém dobře aproximovat problémem s předpokladem rovnoměrnosti, tedy musejí být složené z osob z různých ročníků (čím větší diverzita, tím je aproximace skutečné

situace spolehlivější)[3]. Proto Berresford ve své studii data upravil tím, že relativní četnost pro každý den nahradil průměrem relativní četnosti tohoto dne a šesti dnů po něm následujících, což evidentně eliminovalo vliv týdenní cykličnosti a zároveň to vypovídací hodnotu dat jako takových příliš nepokošdilo. Významným efektem této transformace bylo, že odchylky relativních četností nyní byly pouze do 10% od hodnoty teoretické pravděpodobnosti $\frac{1}{365}$, viz. obrázek 3.3:

Obrázek 3.3: Rozložení transformovaných relativních četností porodů (bez vlivu týdenní cykličnosti) ve státě New York (USA) během roku 1977



Evidentně za situace, kterou ukazuje obrázek 3.3, je aproximace narozeninového problému se skutečnou distribucí porodů problémem s předpokladem rovnoměrnosti dosatečně spolehlivá, neboť skutečná empirická distribuce porodů se již velmi blíží distribuční funkci rovnoměrného rozdělení.

[3]To je dáno tím, že počet dní v roce (365) je nesoudělný s počtem dní v týdnu (7), tím pádem i se zahrnutím přestupných let, se každých nejvýše 28 po sobě jdoucích let vlivy týdenní cykličnosti eliminují, neboť 1. leden (a samozřejmě i kterékoli jiné libovolně zvolené datum) pokaždé připadne na jiný den v týdnu, a to právě čtyřikrát na každý den v týdnu během 28 po sobě jdoucích let.

3.2 Analytický výpočet pravděpodobnosti

Předpokládejme, že jsme očíslovali všechny dny v roce čísly od 1 do 365. Nechť p_i značí pravděpodobnost toho, že se daná osoba narodila v i-tém dni a $i = 1, 2, \ldots, 365$. Uvažujme, že jsme pro dané n vybrali z množiny $\{1, 2, \ldots, 365\}$ podmožinu $\{i_1, i_2, \ldots, i_n\}$ splňující $i_1 < i_2 < \ldots < i_n$. Předpokládejme, že jsme si očíslovali n osob čísly od 1 do n. Zřejmě $p_{i_{\pi(1)}} p_{i_{\pi(2)}} \cdot \ldots \cdot p_{i_{\pi(n)}}$ je pravděpodobnost toho, že se j-tá osoba narodila právě v den $i_{\pi(j)}$ pro všechna $j = 1, 2, \ldots, n$, kde π je nějaká permutace prvků množiny $\{1, 2, \ldots, n\}$. Jaká je pravděpodobnost, označme ji $Q_{i_1,\ldots,i_n}(n)$, toho, že se každá z osob 1 až n narodila v právě jednom dni (nezáleží v jakém konkrétním) z množiny $\{i_1, i_2, \ldots, i_n\}$? Tuto pravděpodobnost evidentně dostaneme, sečteme-li dílčí pravděpodobnosti $p_{i_{\pi(1)}} p_{i_{\pi(2)}} \cdot \ldots \cdot p_{i_{\pi(n)}}$ přes všechny možné permutace π prvků množiny $\{1, 2, \ldots, n\}$. Jelikož permutace je vzájemně jednoznačné zobrazení, musí platit $p_{i_{\pi(1)}} p_{i_{\pi(2)}} \cdot \ldots \cdot p_{i_{\pi(n)}} = p_{i_1} p_{i_2} \cdot \ldots \cdot p_{i_n}$ pro libovolnou permutaci π. Konečně, jak známo, permutací n-prvkové množiny je $n!$, tudíž $Q_{i_1,\ldots,i_n}(n) = n!\, p_{i_1} p_{i_2} \cdot \ldots \cdot p_{i_n}$. Pak stačí vzít všechny možné n-prvkové množiny $\{i_1, i_2, \ldots, i_n\} \subset \{1, 2, \ldots, 365\}$ takové, že $i_1 < i_2 < \ldots < i_n$, to zřejmě při současném uvažování všech možných permutací indexů z každé takové množiny vyčerpává všechny možnosti, jak se n osob může narodit v n různých dnech během roku, aniž by existovala mezi nimi dvojice mající narozeniny ve stejný den. Sečteme-li (stejně jako ve vzorci (3.1)) přes všechny takové podmnožiny indexů $\{i_1, i_2, \ldots, i_n\}$ odpovídající pravděpodobnosti $Q_{i_1,\ldots,i_n}(n)$, obdržíme pravděpodobnost $Q(n)$ toho, že ve skupině n lidí neexistují dva lidé, kteří by se narodili ve stejný den.

$$Q(n) = n! \sum_{i_1 < \ldots < i_n} p_{i_1} \cdot \ldots \cdot p_{i_n} = n! \sum_{i_1 < \ldots < i_n} \prod_{j=1}^{n} p_{i_j} \tag{3.1}$$

Bohužel dle vzorce (3.1) by ve většině případů pravděpodobnost nespočítal v rozumném čase ani výkonnější počítač, neboť by musel sečíst $\binom{365}{n}$ členů. V tabulce (3.1) maximálních přirozených mocnin 10 menších jak $\binom{365}{n}$ vidíme, že, pro většinu hodnot n se jedná o příliš velké množství operací:

Tabulka 3.1: Představa o velikosti hodnot $\binom{365}{n}$

n	10	20	23	30	50	75	100	125	150	175	183
$\binom{365}{n}$	10^{19}	10^{32}	10^{36}	10^{43}	10^{62}	10^{79}	10^{91}	10^{100}	10^{105}	10^{108}	10^{108}

Budeme-li uvažovat, že jádro velmi výkonného procesoru je schopné uskutečnit až $3 \cdot 10^9$ operací za sekundu a nebudeme-li uvažovat omezení plynoucí z ukládání dat a mezivýsledků výpočtů, pak na vykonání $\binom{365}{10} \approx 10^{19}$ součtů bude třeba asi $3,33 \cdot 10^9$ sekund, což činí téměř 106 let. Pokud však vzorec (3.1) modifikujeme do méně konkrétní podoby, dojdeme k výpočetně schůdnější formuli (3.2). Potřebnost takového vztahu vyplývá i z požadavku zmenšit rozsah vstupních dat pro problém, neboli, aby uživatel nemusel pro výpočet znát 365 různých pravděpodobností narození ve všech dnech v roce[4].

Nechť máme k dispozici $m < 365$ různých pravděpodobností $p_1, \ldots, p_m$ tak, že pravděpodobnost p_i je pravděpodobností narození pro právě d_i různých dní, $\sum_{i=1}^{m} d_i = 365$ a $\sum_{i=1}^{m} d_i p_i = 1$, potom:

$$Q(n) = n! \sum_{n_1+\ldots+n_m=n} \prod_{i=1}^{m} \binom{d_i}{n_i} p_i^{n_i}, \tag{3.2}$$

kde sčítáme přes $\binom{m+n-1}{n}$ různých uspořádaných m-tic $(n_1, \ldots, n_m)$ nezáporných celých čísel takových, že jejich součet je n. Jak jsme ke vzorci (3.2) dospěli? n_i zřejmě označuje počet lidí z n osob, kteří se narodili v některém z dnů s pravděpodobností p_i, mluvme v takovém případě o i-té D-množině (dnů). Jelikož v i-té D-množině je právě d_i dní, existuje celkem $\binom{d_i}{n_i}$ možností[5], jak taková situace mohla nastat. Snadno nahlédneme, že $\prod_{i=1}^{m} \binom{d_i}{n_i} p_i^{n_i}$ vyjadřuje celkovou pravděpodobnost, že se pevně určených n_i osob ze zkoumané skupiny narodilo výhradně ve dnech i-té D-množiny. Vysčítáme-li tyto pravděpodobnosti přes všechna i, dostaneme pravděpodobnost toho, že při pevně stanovené příslušnosti každé osoby do D-množiny dle dnu svého narození, nebude mezi těmito lidmi existovat dvojice, která by se narodila ve stejný den. Uvažíme-li všechny možné permutace n osob, dostaneme celkovou pravděpodobnost $Q(n)$ bez pevně stanovené příslušnosti každé osoby do určité D-množiny.

Vztah (3.2) nám, mimojiné, umožňuje spočítat pravděpodobnost $Q(n)$, resp. $\breve{q}(n)$ (při značení dle první kapitoly) v narozeninovém problému s předpokladem rovnoměrnosti se zahrnutím existence 29. února (přestupných roků).

[4]Mnohdy ani taková data nejsou k dispozici, např. ČSÚ zveřejňuje jen měsíční relativní četnosti jako empirické odhady pravděpodobností pro jednotlivé měsíce.

[5]pokud $d_i < n_i$, považujeme $\binom{d_i}{n_i}$ za 0.

Vrátíme-li se k poznatkům z podkapitoly 1.1 a položíme-li $p_1 = \frac{97}{146097}$, $d_1 = 1$ (tedy p_1 bude pravděpodobnost narození 29. února a d_1 počet dní, které takovou pravděpodobnost mají) a $p_2 = \frac{400}{146097}$, $d_2 = 365$ (čili p_2 bude hodnota pravděpodobnosti narození v daný konkrétní den během roku kromě 29. února a d_2 celkový počet dní s touto stejnou hodnotou pravděpodobnosti), potom pro narozeninový problém s předpokladem rovnoměrnosti a přípustěním existence přestupných let obdržíme vztah (3.3):

$$\breve{q}(n) = n! \sum_{\{(n-k,k)|k\in\{0,\ldots,n\}\}} \binom{1}{n-k} \left(\frac{97}{146097}\right)^{n-k} \binom{365}{k} \left(\frac{400}{146097}\right)^{k} =$$

$$= n! \left[\binom{365}{n} \left(\frac{400}{146097}\right)^{n} + \binom{365}{n-1} \frac{97 \cdot 400^{n-1}}{146097^{n}}\right] =$$

$$= n! \binom{365}{n} \left(\frac{400}{146097}\right)^{n} \left[1 + \frac{97n}{400 \cdot (366-n)}\right] \tag{3.3}$$

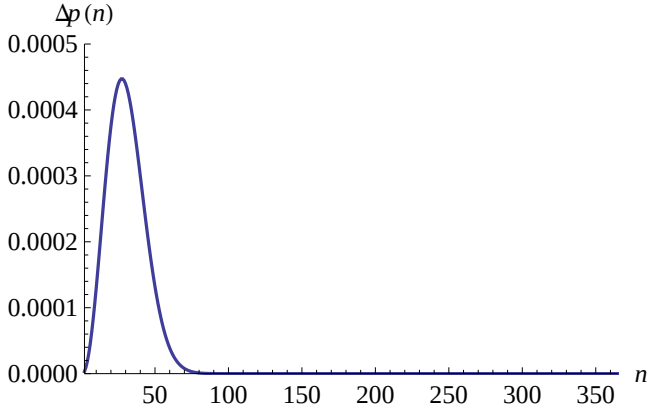
S využitím vztahu (3.3) porovnáme vybrané hodnoty $\breve{p}(n) = 1 - \breve{q}(n)$ s hodnotami $p(n)$ z problému z první kapitoly, kde jsme nepředpokládali existenci přestupných let. Intuitivně (zvýšil se počet možností narození) platí $\breve{p}(n) < p(n)$ pro všechna $n \in \{2, 3, \ldots, 365\}$. Bude-li nás zajímat výše $\Delta p(n) = p(n) - \breve{p}(n)$, můžeme sledovat tabulku (3.2) nebo graf na obrázku (3.4).

Jak je patrné z grafu, hodnoty $\Delta p(n)$ rychle rostou až do jistého maxima, které nastává v bodě $n = 27$ s funkční hodnotou přibližně 0,000447, z tohoto bodu dále prudce klesá a asymptoticky se přimyká k ose n. Je tedy patrné, že vynechání předpokladu existence přestupných let způsobí v nejhorším případě nepřesnost o velikosti nejvýše 0,000447, což je zanedbatelná hodnota.

Tabulka 3.2: Některé numerické hodnoty $\breve{p}(n)$ a $p(n)$

n	$\breve{p}(n)$	$p(n)$
10	$0,1168$	$0,1169$
20	$0,4111$	$0,4114$
23	$0,5069$	$0,5073$
30	$0,7059$	$0,7063$
40	$0,8909$	$0,8912$
50	$0,9702$	$0,9704$
60	$0,9941$	$0,9941$
70	$0,9992$	$0,9992$
80	$0,99991$	$0,99991$
90	$0,999994$	$0,999994$
100	$0,9999997$	$0,9999997$
200	$1-1,82\cdot 10^{-30}$	$1-1,61\cdot 10^{-30}$
300	$1-1,08\cdot 10^{-81}$	$1-6,25\cdot 10^{-82}$
350	$1-1,51\cdot 10^{-130}$	$1-3,03\cdot 10^{-131}$
366	$1-2,77\cdot 10^{-158}$	1

Obrázek 3.4: Průběh rozdílů $\Delta p(n) = p(n) - \breve{p}(n)$ v závislosti na n



Jiné a velmi užitečné použití vzorce (3.2) uvádíme v podkapitole 3.4, kde jej aplikujeme na konkrétní data z ČSÚ a dospíváme k závěru, že i přes významné omezení počtu vstupujících pravděpodobností ze 365 na 12, budou sice výpočetní omezení významně zmírněna (neboli řady výsledků se

dočkáme v dosažitelném čase), ale stále budeme schopni pravděpodobnost $Q(n)$ numericky vyčíslit jen pro několik nejmenších hodnot n.

3.3 Narozeninová nerovnost

Narozeninový problém s předpokladem rovnoměrnosti hraje úlohu mezního případu pro narozeninový problém se skutečným (nerovnoměrným) rozdělením pravděpodobností narození v určitém dnu. Ukážeme, že právě pro rovnoměrné rozdělení je pravděpodobnost toho, že ve skupině n lidí existují alespoň dva mající narozeniny ve stejný den, minimální. Právě uvedené tvrzení je důsledkem tzv. *narozeninové nerovnosti*:

Tvrzení: (*Narozeninová nerovnost*)

$$Q\left[(p_1, p_2, \ldots, p_{365}), n\right] \leq Q\left[\left(p_1 = \frac{1}{365}, p_2 = \frac{1}{365}, \ldots, p_{365} = \frac{1}{365}\right), n\right],$$

kde $Q\left[(p_1, p_2, \ldots, p_{365}), n\right]$ je pravděpodobnost toho, že ve skupině $n \geq 2$ lidí nemá ani jedna dvojice osob narozeniny ve stejný den, je-li pravděpodobnost narození pro den i dána hodnotou p_i, kde $i = 1, \ldots, 365$ a $\sum_{i=1}^{365} p_i = 1$.

Důkaz: Pokud na levé straně nerovnosti platí $p_1 = p_2 = \ldots = p_{365}$, v nerovnosti nastává rovnost a tedy tvrzení platí. Dále budeme předpokládat, že mezi pravděpodobnostmi p_1 až p_{365} existuje dvojice $p_i \neq p_j$. Bez újmy na obecnosti, nechť $p_1 \neq p_2$, jinak bychom mohli pravděpodobnosti jednoduše přečíslovat. Ukážeme, že nahradíme-li obě hodnoty p_1 a p_2 jejich aritmetickým průměrem $\frac{p_1+p_2}{2}$, potom hodnotu $Q\left[(p_1, p_2, \ldots, p_{365}), n\right]$ zvýšíme, neboli $Q\left[(p_1, p_2, \ldots, p_{365}), n\right] < Q\left[\left(\frac{p_1+p_2}{2}, \frac{p_1+p_2}{2}, p_3, \ldots, p_{365}\right), n\right]$. Následně vyjádření $Q\left[(p_1, p_2, \ldots, p_{365}), n\right] = Q(n)$ ze vzorce (3.1) rozdělíme na tři části: na členy obsahující jak p_1, tak i p_2, členy obsahující právě jednu z pravděpodobností p_1 a p_2 a konečně členy neobsahující žádnou z obou pravděpodobností:

$$Q\left[(p_1, p_2, \ldots, p_{365}), n\right] =$$

$$= n! \left(p_1 p_2 \sum_{2<i_1<\ldots<i_{n-2}} p_{i_1} \ldots p_{i_{n-2}} + \right.$$

$$+ (p_1 + p_2) \sum_{2<i_1<\ldots<i_{n-1}} p_{i_1} \ldots p_{i_{n-1}} +$$

$$\left. + \sum_{2<i_1<\ldots<i_n} p_{i_1} \ldots p_{i_n} \right) \tag{3.4}$$

Pokud nyní ve vzorci (3.4) nahradíme hodnoty p_1 a p_2 jejich společným průměrem, zřejmě tím hodnotu druhého ani třetího členu nezměníme. Nicméně zvýšíme tím hodnotu prvního členu, což vyplývá z AG-nerovnosti $\sqrt{p_1 p_2} \leq \frac{p_1+p_2}{2}$ s rovností právě, když $p_1 = p_2$. V našem případě je dle předpokladu $p_1 \neq p_2$, čili můžeme použít AG-nerovnost s ostrou nerovností. Jelikož při tomto nahrazení se první člen v (3.4) zvýší a ostatní zůstanou beze změny, skutečně jsme tím hodnotu $Q\left[(p_1, p_2, \ldots, p_{365}), n\right]$ zvýšili. Protože pravděpodobnost $Q\left[(p_1, p_2, \ldots, p_{365}), n\right]$ může být zvýšena, kdykoli mezi hodnotami p_i existuje dvojice $p_i \neq p_j$, a to nahrazením obou hodnot jejich aritmetickým průměrem, musí být pravděpodobnost $Q\left[(p_1, p_2, \ldots, p_{365}), n\right]$ maximální[6] právě, když všechna p_i mají totožnou hodnotu.■

Na závěr této podkapitoly si stačí jen uvědomit, že pravděpodobnost shody v narozeninovém problému je komplementární k pravděpodobnosti $Q\left[(p_1, p_2, \ldots, p_{365}), n\right]$, odtud evidentně dostáváme, že za předpokladu rovnoměrného rozdělení pravděpodobností narození v určitém dni je pravděpodobnost toho, že ve skupině n osob existují alespoň dvě osoby mající narozeniny ve stejný den, minimální.

[6] pokud takové maximum existuje, což je v našem případě zaručeno tím, že funkce Q je spojitou funkcí na kompaktu daném podmínkami $\sum_{i=1}^{365} p_i = 1$, $0 \leq p_i \leq 1$, pro $\forall i \in \{1, \ldots, 365\}$.

3.4 Odhad pravděpodobnosti neexistence shody na základě dat z ČSÚ

Nyní se budeme snažit určit odhad $\widehat{Q}(n)$ na základě relativních četností narození pro průměrný den v každém měsíci[7] za období let 1991-2005. Zmiňované relativní četnosti budeme považovat za empirické odhady $\hat{p}_1, ..., \hat{p}_{12}$. V tabulce (3.3) uvádíme přehledně poslední řádek z tabulky (P.1).

Tabulka 3.3: Odhady pro pravděpodobnosti narození v jednotlivých měsících přepočtené na jeden den

Měsíc	$f_{\text{měsíc}}$	Co odhaduje?
Leden	0,002659	$\hat{p}_1$
Únor	0,002768	$\hat{p}_2$
Březen	0,002855	$\hat{p}_3$
Duben	0,002933	$\hat{p}_4$
Květen	0,002927	$\hat{p}_5$
Červen	0,002925	$\hat{p}_6$
Červenec	0,002903	$\hat{p}_7$
Srpen	0,002756	$\hat{p}_8$
Září	0,002745	$\hat{p}_9$
Říjen	0,002514	$\hat{p}_{10}$
Listopad	0,002451	$\hat{p}_{11}$
Prosinec	0,002428	$\hat{p}_{12}$

Relativní četnost pro měsíc i byla určena jako:

$$\hat{p}_i = \frac{N_i}{\sum_{k=1}^{12} N_k} \cdot \frac{1}{\overline{D_i}},$$

kde N_i značí celkový počet živě narozených v měsíci i v celém sledovaném období (řádek „Celkem" v tabulce (P.1)), D_i značí celkový počet dní, které

[7]Nejlepší by samozřejmě bylo užívat různé relativní četnosti pro různé dny a nejen pro měsíce, nicméně i tak bychom měli obdržet věrnější odhad $P(n)$ za předpokladu reálné distribuce narození během roku, než aproximujeme-li pravděpodobností $p(n)$ z narozeninového problému s předpokladem rovnoměrného rozdělení z první kapitoly. Druhou překážkou je i fakt, že údaje o relativních četnostech pro jednolitvé dny pro oblast ČR jednoduše nemáme k dispozici.

náležely do měsíce i ve zkoumaném období (řádek „Dnů celkem" v tabulce (P.1)) a konečně $\overline{D_i} = \frac{D_i}{15}$ je průměrný počet dní, který připadal na 1 rok za celé sledované období (řádek „Průměrně dní" v tabulce (P.1))[8]. Dané relativní četnosti budeme aplikovat na vzorec (3.2), v němž dále budeme klást $m = 12$ a $(d_1, d_2, ..., d_{12}) = (31, 30, 28, 30, 31, 30, 31, 31, 30, 31, 30, 31)$ (tj. počty dní v jednotlivých kalendářních měsících v nepřestupném roce), čili budeme využívat následující formuli:

$$\widehat{Q}(n) = n! \sum_{n_1+...+n_{12}=n} \prod_{i=1}^{12} \binom{d_i}{n_i} \hat{p}_i^{n_i} \tag{3.5}$$

Nicméně již pro malé hodnoty n narážíme na značné výpočetní problémy, a to i při použití výpočetní techniky. Vzorec (3.5) zadáme k výpočtu programu *Mathematica* následující sérií příkazů:

```
<< Combinatorica‘
Clear [n, d, m, p, rozklad]
n := //je třeba doplnit nějaké přirozené číslo//
d := {31, 28, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31}
m := 12
p := {0.002659487, 0.00276795, 0.002855341, 0.002933096, 0.002927356,
0.002925005, 0.002902664, 0.002755975, 0.002744918, 0.002514381,
0.002450937, 0.002427543}
rozklad := Compositions[n, m]
delka := Length[rozklad]
Timing[n!*Sum[Apply[Times,Flatten[Binomial[d,rozklad[[i]]]]*
(p^rozklad[[i]])]],{i, delka}]]
```

V tabulce (3.4) uvádíme výsledky $\widehat{Q}(n)$ pro $n \in \{2, 3, 4, 5, 6, 7\}$ a zároveň je porovnáváme s hodnotou $q(n)$ z první kapitoly za předpokladu rovnoměrnosti. Můžeme přitom pozorovat, že odlišnosti nastávají v řádu tisícin, což je přijatelná nepřesnost, všimneme-li si, že doba přesného výpočtu exponenciálně roste:

[8]Evidentně pro všechny měsíce kromě února to bude skutečný počet dní v daném měsíci, tj. 30 nebo 31, v případě února dospějeme k desetinnému číslu 28,2, což je dáno zahrnutím tří přestupných roků 1992, 1996 a 2004.

Tabulka 3.4: Srovnání odhadů $\widehat{Q}(n)$ s teoretickými hodnotami $q(n)$

n	$\widehat{Q}(n)$	doba výpočtu	$q(n)$	$q(n) - \widehat{Q}(n)$
2	0,9961	1,212 s	0,9973	0,0011
3	0,9901	25,987 s	0,9918	0,0017
4	0,9814	6 min 10,813 s	0,9836	0,0022
5	0,9701	1 h 7 min 51,920 s	0,9729	0,0028
6	0,9562	8 h 59 min 25,341 s	0,9595	0,0034
7	0,9399	2 dny 6 h 46 min 44 s	0,9438	0,0039

Vzhledem k narozeninové nerovnosti jistě platí, že $q(n) > \widehat{Q}(n) > Q(n)$, neboli odhad $\widehat{Q}(n)$ pocházející z našeho pseudorovnoměrného modelu (kdy připouštíme skutečnou nerovnoměrnost mezi dny z různých měsíců, ale v rámci jednoho měsíce uvažujeme rovnoměrné rozdělení pravděpodobnosti narození pro jednotlivé dny tohoto měsíce) je přesnějším odhadem (horní mezí) skutečné pravděpodobnosti $Q(n)$.

Kapitola 4

Monte Carlo odhad při nerovnoměrné distribuci narození během roku

Jak jsme viděli v předešlé kapitole, ani výpočetní technika není dostačující na to, aby spočítala přesnou analytickou hodnotu pravděpodobnosti $Q(n)$, jejíž význam zůstává stejný jako v předešlé podkapitole. Nicméně pomocí počítačové simulace jsme schopni tento problém vyřešit v rozumně dlouhém čase i s uspokojivou přesností. Pro účely této kapitoly označme $\pi(n) = = 1 - Q(n)$, tedy pravděpodobnost, že v dané skupině n lidí existují alespoň dva mající narozeniny ve stejný den.

4.1 Reprezentace dat a realizace počítačové simulace

Dále nechť náhodná veličina $X(n)$ nabývá hodnoty 0, není-li v dané skupině o n osobách žádná shoda ve dni narozenin, a 1, existuje-li mezi těmito osobami dvojice mající narozeniny ve stejný den. Snadno nahlédneme, že takto definovaná náhodná veličina $X(n)$ má alternativní rozdělení s parametrem $\pi(n)$, neboli $X(n) \sim \mathrm{Alt}(\pi(n))$. Provedeme-li dostatečný počet pozorování veličiny $X(n)$, budeme moci sestrojit intervalový odhad $B(\boldsymbol{X(n)}) = = (C_L(\boldsymbol{X(n)}, C_R(\boldsymbol{X(n)})$ pro hledaný parametr $\pi(n)$.

Zvolme jako počet pozorování 10 000. Pomocí počítačové simulace v softwaru *Mathematica* vygenerujeme 10 000 n-tic náhodných čísel z diskrétního

rozdělení daného pravděpodobnostmi $\hat{p}_1, \ldots, \hat{p}_{12}$ na množině $\{1, 2, \ldots, 365\}$ tak, že $\hat{p}_i$ je pravděpodobností vygenerování pro každé přirozené číslo od $\sum_{j=1}^{i-1} d_i + 1$ do $\sum_{j=1}^{i} d_i$ pro $i \in \{1, 2, \ldots, 12\}$. Tyto n-tice vhodně reprezentují v jakých dnech mohou mít osoby z dané n-členné skupiny narozeniny s distribucí, která odpovídá skutečné distribuci dané relativními četnostmi porodů pro jednotlivé měsíce. Následující serie příkazů generuje 10 000 takových n-tic a navíc na výstupu vyhodnocuje, v kolika z nich se vyskytlo aspoň jedno číslo dvakrát (tj. nastala shoda u aspoň dvou osob ve dnu narození):

```
d := {31, 28, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31}
p := {0.002659487, 0.00276795, 0.002855341, 0.002933096, 0.002927356,
0.002925005, 0.002902664, 0.002755975, 0.002744918, 0.002514381,
0.002450937, 0.002427543}
n := //je třeba doplnit nějaké přirozené číslo//
pocetpozorovani := 10000
pozorovani := Table[RandomChoice[Flatten[Table[Table[p[[i]], {j, d[[i]]}],
{i, Length[d]}]] -> Range[365], {n}], {k, pocetpozorovani}]
shoda := Table[If[Length[Union[SeedRandom[1]; pozorovani[[l]]] < n, "A",
"N"],{l,pocetpozorovani}]
Count[shoda, "A"]
```

4.2 Popis a použití statistické metody intervalového odhadu

Konkrétní výsledky počítačových simulací pro jednotlivá n uvádíme v tabulce (4.1). Nejprve získáme údaj $\sum_{i=1}^{10000} X_i(n)$, tj. počet shod, který nastal při provedení 10 000 pozorování. Odtud snadno určíme aritmetický průměr $\hat{\pi}(n) \equiv \overline{X_{10000}(n)} = \frac{1}{10000} \sum_{i=1}^{10000} X_i(n)$, který je, jak známo, nestranným a konzistentním odhadem střední hodnoty $\mathrm{E}X_i(n) = \pi(n)$ [1] (viz. třetí sloupec tabulky). Pro srovnání $\hat{\pi}(n)$ s přesnou hodnotou z problému s předpokladem rovnoměrnosti uvádíme i hodnotu $p(n)$. Dále z rozptylu $\mathrm{Var}X(n)$ odhad-

[1] platí $X \sim \mathrm{Alt}(p) \Rightarrow \mathrm{E}X = p$

nutého předpisem[2]:

$$\frac{1}{10000}\sum_{i=1}^{10000}\left(X_i(n)-\overline{X_i(n)}\right)^2 = \overline{X_{10000}(n)}(1-\overline{X_{10000}(n)}) \equiv \widehat{\mathrm{Var}}X(n)$$

ihned obdržíme jako $\sqrt{\widehat{\mathrm{Var}}X(n)}$ odhad[3] směrodatné odchylky $\hat{\sigma}(X(n))$. V následujících sloupcích tabulky uvádíme hodnoty $C_L(\boldsymbol{X(n)})$, $C_R(\boldsymbol{X(n)})$ a jejich rozdíl (délku intervalu). Jedná se o levý a pravý krajní bod intervalového odhadu na 95%-ní hladině spolehlivosti, který získáme známým postupem popsaným v následujícím odstavci.

Předpokládejme nyní, že jsme učinili m pozorování veličiny $X(n)$, potom dle centrální limitní věty platí:[4]

$$\mathcal{L}\left(\sqrt{m}\,\frac{\overline{X_m(n)}-\mathrm{E}\ X_i(n)}{\sqrt{\mathrm{Var}X_m(n)}}\right) = \mathcal{L}\left(\sqrt{m}\,\frac{\overline{X_m(n)}-\pi(n)}{\sigma(X_m(n))}\right) \longrightarrow \mathrm{N}(0,1), m\to\infty \tag{4.1}$$

Chceme-li sestrojit (symetrický) intervalový odhad pro parametr $\pi(n)$ na hladině spolehlivosti $1-\alpha$ pro vhodné $\alpha\in(0;1)$, stačí využít, že platí $P[u_{\frac{\alpha}{2}} < \varphi(\boldsymbol{X(n)},\pi(n)) < u_{1-\frac{\alpha}{2}}] = 1-\alpha$, kde funkce $\varphi(\boldsymbol{X(n)},\pi(n))$ je pivotální statistikou se známým rozdělením a $u_{\frac{\alpha}{2}}$, resp. $u_{1-\frac{\alpha}{2}}$ jsou příslušné kvantily rozdělení statistiky $\varphi(\boldsymbol{X(n)},\pi(n))$. Aplikujme nyní předchozí teoretický rozbor na náš problém. Za pivotální statistiku zvolíme funkci φ danou přepisem $\varphi(\boldsymbol{X(n)},\pi(n)) = \sqrt{m}\frac{\overline{X_m(n)}-\pi(n)}{\sigma(X(n))}$, která má dle centrální limitní věty asymptoticky standardní normální rozdělení $\mathrm{N}(0,1)$, čili dostáváme:

$$P\left[\Phi^{-1}\left(\frac{\alpha}{2}\right) < \sqrt{m}\,\frac{\overline{X_m(n)}-\pi(n)}{\sigma(X_m(n))} < \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right] \to 1-\alpha, m\to\infty \tag{4.2}$$

Jak známo, Φ označuje distribuční funkci rozdělení $\mathrm{N}(0,1)$. Vzhledem k symetrii hustoty normálního rozdělení platí, že $\Phi^{-1}\left(\frac{\alpha}{2}\right) = -\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$, a tedy předešlý vztah je možno přepsat do ekvivalentní podoby:

[2] platí $X\sim\mathrm{Alt}(p)\Rightarrow \mathrm{Var}X = p(1-p)$

[3] konzistentní, nikoli však nestranný; nicméně jeho vychýlení o velikosti $\frac{1}{9999}\widehat{\mathrm{Var}}X(n)$ je prakticky zanedbatelné

[4] $\mathcal{L}(X)$ značí rozdělení náhodné veličiny X

$$P\left[-\Phi^{-1}\left(1-\frac{\alpha}{2}\right) < \sqrt{m}\,\frac{\overline{X_m(n)}-\pi(n)}{\sigma(X_m(n))} < \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right] \to 1-\alpha, m \to \infty \tag{4.3}$$

Osamostatněním parametru $\pi(n)$ v nerovnostech v (4.3) získáme klíčovou formuli (4.4), pomocí níž zkonstruujeme intervalový odhad parametru $\pi(n)$ v okolí jeho bodového odhadu $\overline{X_m(n)}$ pro $m \to \infty$:

$$P\left[\overline{X_m(n)} - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{\sigma(X_m(n))}{\sqrt{m}} < \pi(n) < \overline{X_m(n)} + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\frac{\sigma(X_m(n))}{\sqrt{m}}\right] \to 1-\alpha \tag{4.4}$$

Konečný vztah použitý při konstrukci intervalových odhadů v tabulce (4.2) získáme dosazením konkrétních číselných hodnot pro hladinu spolehlivosti testu $1-\alpha$, příslušný $\left(1-\frac{\alpha}{2}\right)$-tý kvantil standardního normálního rozdělení a počet pozorování m. Zvolme $\alpha = 0,05$, což je běžná hodnota při práci se statistickou spolehlivostí, potom $\Phi^{-1}\left(1-\frac{\alpha}{2}\right) = \Phi^{-1}(0,975) \doteq 1,96$ a v důsledku při $m = 10\,000$ pozorováních dostáváme vztah:

$$P\left[\hat{\pi}(n) - 0,0196 \cdot \hat{\sigma}(X(n)) < \pi(n) < \hat{\pi}(n) - 0,0196 \cdot \hat{\sigma}(X(n))\right] \to 0,95 \tag{4.5}$$

Tedy asymptotický intervalový odhad se spolehlivostí 95% pro parametr $\pi(n)$ je interval:

$$(C_L(\boldsymbol{X(n)}, C_R(\boldsymbol{X(n)}) = (\hat{\pi}(n) - 0,0196 \cdot \hat{\sigma}(X(n)); \hat{\pi}(n) + 0,0196 \cdot \hat{\sigma}(X(n)))$$

Ačkoli teoreticky má platit narozeninová nerovnost, výsledky simulací pro tento fakt nedopadly zcela přesvědčivě. Vzhledem k narozeninové bychom očekávali, že $\hat{\pi}(n) > p(n)$, avšak tato situace mezi zvolenými n nastala jen pro hodnoty 7, 16, 22, 24, 26 a 55. Čemu to lze přičíst? Předně, nezanedbatelné pravděpodobnosti $\hat{\pi}(n) \leq p(n)$, za níž stojí skutečnost, že počet pozorování zvolený jako 10 000 není dostatečný. Při 10 000 pozorováních se totiž hodnoty $|\hat{\pi}(n) - \pi(n)|$ nijak markantně neliší[5] od $\pi(n) - p(n)$, což je žádoucí. Kdybychom zvolili vyšší počet pozorování, snížili bychom pravděpodobnost platnosti $\hat{\pi}(n) \leq p(n)$ a zvýšili bychom přesnost takového intervalového odhadu.

[5]Odchylky pravděpodobností narození pro jednotlivé měsíce od pravděpodobnosti za předpokladu rovnoměrnosti nejsou tolik významné, aby po provedení 10 000 pozorování způsobovaly výraznější odlišnost hodnoty $\hat{\pi}(n)$ od $p(n)$, ať už v jednom anebo i ve druhém směru.

Tabulka 4.1: Intervalové odhady pro pravděpodobnost $\pi(n)$

n	$p(n)$	$\hat{\pi}(n)$	$C_L(\boldsymbol{X(n)})$	$C_R(\boldsymbol{X(n)})$	$C_R - C_L$
7	0,0562	0,0589	0,0543	0,0635	0,0092
15	0,2529	0,2477	0,2392	0,2562	0,0169
16	0,2836	0,2872	0,2783	0,2961	0,0177
17	0,3150	0,3134	0,3043	0,3225	0,0182
18	0,3469	0,3438	0,3345	0,3531	0,0186
19	0,3791	0,3736	0,3641	0,3831	0,0190
20	0,4114	0,4105	0,4009	0,4201	0,0193
21	0,4437	0,4432	0,4335	0,4529	0,0195
22	0,4757	0,4761	0,4663	0,4859	0,0196
23	0,5073	0,5034	0,4936	0,5132	0,0196
24	0,5383	0,5384	0,5286	0,5482	0,0195
25	0,5687	0,5657	0,5560	0,5754	0,0194
26	0,5982	0,6002	0,5906	0,6098	0,0192
27	0,6269	0,6255	0,6160	0,6350	0,0190
28	0,6545	0,6527	0,6434	0,6620	0,0187
29	0,6810	0,6708	0,6616	0,6800	0,0184
30	0,7063	0,7030	0,6940	0,7120	0,0179
35	0,8144	0,8116	0,8039	0,8193	0,0153
40	0,8912	0,8889	0,8827	0,8951	0,0123
45	0,9410	0,9409	0,9363	0,9455	0,0092
50	0,9704	0,9702	0,9669	0,9735	0,0067
55	0,9863	0,9865	0,9842	0,9888	0,0045
60	0,9941	0,9934	0,9918	0,9950	0,0032

V tabulce si lze rovněž povšimnout, že délka jednotlivých intervalových odhadů je přiměřená počtu pozorování (s roustoucím počtem pozorování se takto konstruované intervaly zkracují) a pohybuje se od 0,0032 do 0,0196. Ze vztahu (4.5) jednoduše plyne, proč je tento interval nejdelší pro hodnoty n okolo 23. Totiž $x(1-x)$ pro $x \in (0;1)$ je, jak známo, maximální, pokud $x = 0,5$, z toho důvodu je i $\hat{\sigma}(X(n)) = \sqrt{\hat{\pi}(n)(1-\hat{\pi}(n))}$ maximální pro hodnoty $\hat{\pi}(n)$ okolo 0,5. Naopak pro hodnoty $\hat{\pi}(n)$ u 0 nebo 1 je délka takového odhadu nejmenší.

V případě $n = 7$ jsme určili pravděpodobnost $p(n)$, resp. $P(n)$ všemi v práci použitými metodami. K jakým výsledkům jsme dospěli? Sledujme tabulku 4.2:

Tabulka 4.2: Srovnání odhadů $\widehat{Q}(n)$ s teoretickými hodnotami $q(n)$

n	$p(n) = 1 - q(n)$	$\widehat{P}(n) = 1 - \widehat{Q}(n)$	$\hat{\pi}(n)$	95% interval spolehlivosti
7	0,0562	0,0601	0,0589	(0,0543;0,0635)

Můžeme si povšimnout, že rozdíly mezi jednotlivými výsledky se pohybují pouze v řádu tisícin. Lze konstatovat, že zvolíme-li vhodně vysoký počet pozorování, můžeme obdržet na základě počítačových simulací v dosažitelném čase výsledky se srovnatelnou přeností, jako když problém s nerovnoměrnou distribucí aproximujeme problémem s rovnoměrnou distribucí narození během roku. Navíc tato metoda odhadu nám umožňuje libovolně upravovat čas výpočtu i přesnost takového odhadu (a to pouhou volbou počtu pozorování), na rozdíl od přesného výpočtu z podkapitoly 3.4, kde je třeba čekat několik dní i déle na výsledek s přesností až na velké množství desetinných míst, která prakticky nevyužijeme k žádnému dalšímu výpočtu ani interpretaci.

Přílohy

Tabulka P.1:

Měsíc	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
1991	11 456	10 670	11 907	11 770	11 914	11 344	11 197	10 713	10 546	9 671	9 066	9 240
1992	10 028	9 646	10 985	10 559	10 988	10 772	10 902	10 333	10 115	9 418	8 764	9 276
1993	9 885	9 632	10 790	10 472	10 933	10 819	11 100	10 613	10 170	9 274	8 707	8 630
1994	9 169	8 766	10 149	10 049	10 218	9 541	9 385	8 454	8 083	7 674	7 484	7 587
1995	8 155	7 560	8 875	8 179	8 561	8 411	8 551	8 165	7 708	7 643	7 017	7 176
1996	7 613	7 273	7 979	7 901	8 340	7 983	8 216	7 389	7 069	6 947	6 827	6 909
1997	7 488	6 883	7 854	8 208	8 580	7 814	8 367	7 712	7 521	7 037	6 372	6 698
1998	7 064	6 754	7 846	8 129	8 144	8 056	8 491	7 772	7 504	7 102	6 782	6 789
1999	7 187	6 907	7 875	7 892	8 007	7 927	8 007	7 759	7 648	6 816	6 408	7 038
2000	7 557	7 142	8 025	7 983	8 391	7 837	7 906	7 797	7 269	7 177	6 946	6 880
2001	7 574	6 798	7 878	7 948	8 207	7 870	8 088	7 889	7 396	7 301	6 950	6 816
2002	7 432	7 184	8 166	8 145	8 417	7 834	8 122	8 073	7 717	7 498	6 968	7 230
2003	7 538	6 992	7 999	7 862	8 279	8 021	8 849	8 279	8 015	7 678	6 816	7 357
2004	7 822	7 586	8 100	8 381	8 457	8 584	8 785	8 507	8 266	7 702	7 575	7 899
2005	8 004	7 581	8 676	8 838	9 023	9 139	9 343	9 015	8 801	8 271	7 885	7 635
Celkem	123973	117 375	133 103	132 317	136 460	131 952	135 309	128 471	123 828	117 209	110 566	113 161
Dnů celkem	465	423	465	450	465	450	465	465	450	465	450	465
Průměrně dní	31	28,2	31	30	31	30	31	31	30	31	30	31
$f_{\text{měsíc}}(\text{v } \frac{1}{1000})$	2,6595	2,7680	2,8553	2,9331	2,9274	2,9250	2,9027	2,7560	2,7449	2,5144	2,4509	2,4275

Pozn: Tabulka až na poslední tři řádky uvádí absolutní počty živě narozených v daném měsíci a daném roce, řádek „Celkem" tyto počty sčítá pro jednotlivé kalendářní měsíce, „Dnů celkem" je počet dní, které náležely danému měsíci za celé období 1991-2005, „Průměr dní" značí průměrný počet počet dní daného měsíce v celém sledovaném období, „$f_{\text{měsíc}}$" je relativní četnost živě narozených pro jeden konkrétní průměrný den v daném měsíci (tzn. v rámci jednoho kalendářního měsíce předpokládáme rovnoměrné rozdělení, tedy každý den bude mít tuto stejnou relativní četnost)

Literatura

[1] AUTOŘI ČSÚ: *Sezónnost a vícečetné porody.* In Porodnost a plodnost 2001 až 2005, Český statistický úřad, Praha, 2006. [online] [cit. 2009-01-04]. Dostupné z: <http://www.czso.cz/csu/2006edicniplan.nsf/t/360034F8EF/$File/400806a3.pdf>

[2] BERRESFORD, Geoffrey. C.: *The Uniformity Assumption in the Birthday Problem*, Mathematics Magazine, Vol. 53, No. 5, Mathematical Association of America, 1980, s. 286-288. [online] [cit. 2009-06-04]. Dostupné z: <http://www.jstor.org/stable/2689391>

[3] GANS, Joshua & LEIGH, Andrew: *Unusual days in births and deaths*, The Melbourne Review, Vol. 3 No. 1, Melbourne, 2007, s. 72-79. [online] [cit. 2009-06-04]. Dostupné z: <http://econrsss.anu.edu.au/~aleigh/pdf/UnusualDaysBirthsDeaths.pdf>

[4] KOULA, Petr: *Hashovaní (hashing).* [online] [cit. 2009-06-04]. Dostupné z: <http://koula.networld.cz/dsa/>

[5] MALÝ, Jan: *Jednocestné zabezpečení citlivých údajů v databázi*, Fakulta elektrotechniky a komunikačních technologií VUT v Brně, Brno, 2008. [online] [cit. 2009-10-04]. Dostupné z: <www.elektrorevue.cz/cz/clanky/informacni-techologie/0/jednocestne-zabezpeceni-citlivych-udaju-v-databazi/>

[6] WEISSTEIN, Eric W.: *Birthday Problem*, MathWorld - A Wolfram Web Resource. [online] [cit. 2009-06-04]. Dostupné z: <http://mathworld.wolfram.com/BirthdayProblem.html>

[7] WIKIPEDIA CONTRIBUTORS: *Birthday Problem*, Wikipedia, The Free Encyclopedia. [online] [cit. 2009-01-03]. Dostupné z:

<http://en.wikipedia.org/w/index.php?title=Birthday_problem
&oldid=288518822>

[8] WIKIPEDIA CONTRIBUTORS: *Harmonic number*, Wikipedia, The Free Encyclopedia. [online] [cit. 2009-05-05]. Dostupné z: <http://en.wikipedia.org/w/index.php?title=Harmonic_number
&oldid=275618303>

[9] SUCHAN, Martin: *Porovnání současných a nových hašovacích funkcí*, bakalářská práce, Matematicko-fyzikální fakulta UK v Praze, Praha, 2007.