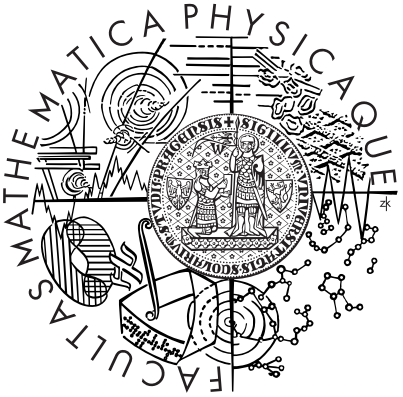
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

# BAKALÁŘSKÁ PRÁCE




Jiří Dvořák

## Benfordovo rozdělení

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Zdeněk Hlávka Ph.D.
Studijní program: matematika, obecná matematika

2008

Děkuji Mgr. Zdeňku Hlávkovi, Ph.D. za cenné rady v odborných i technických záležitostech a stejně tak všem ostatním, kteří svými připomínkami a komentáři přispěli k vylepšení této práce.

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

V Praze dne 14. 7. 2008 Jiří Dvořák

# Obsah

Název práce: Benfordovo rozdělení
Autor: Jiří Dvořák
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: Mgr. Zdeněk Hlávka, Ph.D.
e-mail vedoucího: Zdenek.Hlavka@mff.cuni.cz

Abstrakt: Cílem této práce je podat přehled o problematice nerovnoměrného rozdělení prvních platných číslic přirozeně se vyskytujících čísel. Tento fenomén je známý jako Benfordův zákon.

Tvrzení Benfordova zákona říká, že v přirozeně se vyskytujících číslech se budou na prvním platném místě nízké číslice vyskytovat častěji než číslice vysoké. Tohoto jevu lze využít například při odhalování účetních podvodů nebo při analýze zaokrouhlovacích chyb při rozsáhlých numerických výpočtech.

V práci bude dokázáno kritérium umožňující rozhodnout, zda se zkoumaný soubor čísel řídí Benfordovým zákonem. Dále budou prezentovány původní výsledky rozboru skutečných dat, ilustrující různé aspekty této problematiky.

Klíčová slova: Benfordův zákon, Benfordovo rozdělení, první platná číslice

Title: Benford distribution
Author: Jiří Dvořák
Department: Department of Probability and Mathematical Statistics
Supervisor: Mgr. Zdeněk Hlávka, Ph.D.
Supervisor's e-mail address: Zdenek.Hlavka@mff.cuni.cz

Abstract: The aim of this work is to review the subject of the logarithmic distribution of leading significant digits of naturally occurring numbers. This phenomenon is known as Benford's Law.

It states that the first significant digits of naturally occurring numbers are the lower ones more often than the high ones. This fact can be put to use in detecting fraud in accounting data or in probabilistic analysis of round-off errors in extensive numerical calculations.

A theorem which helps to decide whether the data in question follow Benford's Law will be proved and original results of analysis of real data that illustrate various aspects of Benford's Law will be presented.

Keywords: Benford's law, Benford's distribution, first significant digit

# Úvod

V mnoha souborech přirozeně se vyskytujících čísel se na prvním platném místě objevují nízké číslice výrazně častěji než číslice vysoké. Tento fenomén bývá označován jako Benfordův zákon a v české odborné literatuře mu zatím bylo věnováno minimum pozornosti.

Tato práce si klade za cíl seznámit čtenáře s problematikou Benfordova zákona, od čistě teoretických pokusů o vysvětlení tohoto fenoménu až po možnosti využití v praktických aplikacích.

V první kapitole jsou definovány základní pojmy používané v dalším textu a odvozeny některé vlastnosti rozdělení prvních (a dalších) platných číslic. Také jsou zde shrnuty nejzajímavější pokusy o „dokázání“ Benfordova zákona.

Následuje zavedení pojmů invariance vzhledem ke změně měřítka a invariance vzhledem ke změně základu číselné soustavy (kapitola 2). To umožní odvodit kritérium, které pomáhá rozhodnout, zda se zkoumaná množina čísel řídí Benfordovým zákonem.

V kapitole 3 jsou popsány metody zkoumání shody dat s Benfordovým zákonem. Některé z nich je možné využít k posouzení shody v případě náhodného výběru z nějakého pravděpodobnostního rozdělení bez znalosti konkrétní realizace výběru. Dále jsou zde uvedeny některé výsledky týkající se rozdělení mantisy v číselných posloupnostech.

Kapitola 4 shrnuje empirické výsledky F. Benforda a dalších autorů zabývajících se touto problematikou spolu s novými výsledky, na nichž jsou ilustrovány aspekty Benfordova zákona zmiňované v předchozích kapitolách. Grafické znázornění těchto výsledků je možné najít v příloze A.

Možné aplikace Benfordova zákona v různých oblastech od účetnictví až po analýzu zaokrouhlovacích chyb při numerických výpočtech jsou nastíněny v kapitole 5.

# Kapitola 1

# Benfordův zákon

Americký astronom a matematik Simon Newcomb v roce 1881 publikoval článek [1], v němž upozornil na skutečnost, že první stránky logaritmických tabulek jsou zřetelně opotřebovanější a špinavější než stránky na konci. Z toho usoudil, že uživatelé těchto tabulek (návštěvníci knihovny, vědci a studenti přírodovědných i společenských oborů) se při své práci častěji setkávají s čísly začínajícími číslicí 1 nebo 2, jejichž logaritmy jsou uvedeny v přední části tabulek, než s těmi, které začínají číslicí 8 nebo 9.

Na první pohled se zdá přirozené předpokládat, že první platná číslice čísel, s nimiž se lidé setkávají, bude se stejnou pravděpodobností jednička, dvojka i devítka. Newcombovo tvrzení je ale v rozporu s touto intuitivní představou.

Jako příklad nerovnoměrného rozdělení prvních číslic mohou sloužit údaje o počtu obyvatel v 6249 obcích v České republice k 1. 1. 2007 (podrobnosti v kapitole 4). Četnosti výskytu číslic 1, 2, ..., 9 na prvním platném místě jsou uvedeny v následující tabulce.

| **Číslice** | **Četnost výskytu** |
|---|---|
| 1 | 1820 |
| 2 | 1134 |
| 3 | 794 |
| 4 | 595 |
| 5 | 539 |
| 6 | 420 |
| 7 | 356 |
| 8 | 332 |
| 9 | 259 |

Tato čísla snad přesvědčí i skeptického čtenáře, že v některých souborech „přírodních dat“ se na prvním platném místě vyskytují nízké číslice častěji než vysoké.

Ve 30. letech 20. století si Frank Benford všiml stejného nerovnoměrného opotřebení stránek logaritmických tabulek a zřejmě bez znalosti [1] vydal v roce 1938 vlastní článek [2]. Díky tomu začal být fakt, že v mnoha případech nejsou první platné číslice rozděleny rovnoměrně, označován jako *Benfordův zákon.*

## 1.1 Benfordův zákon a Benfordovo rozdělení

Před vlastní formulací Benfordova zákona je vhodné zavést několik pojmů, které budou v následujícím textu používány.

**Definice 1.1** *Mantisa (při vyjádření čísel v desítkové soustavě) je funkce* $m : (0, \infty) \to [1, 10)$ *taková, že každé* $x \in (0, \infty)$ *se dá vyjádřit ve tvaru* $x = m(x) \cdot 10^n$ *pro nějaké* $n \in \mathbb{Z}$.

Takové číslo $m(x)$ je v intervalu $[1, 10)$ jediné a proto je definice korektní. Pokud bude v dalším textu nutné rozlišit, že jde o mantisu při vyjádření čísla v desítkové soustavě, bude použito označení $m^{(10)}(x)$. Podobně jako výše se dá definovat mantisa při vyjádření čísel v soustavě o jiném základu.

**Definice 1.2** *Mantisa (při vyjádření čísel v soustavě o základu b) je funkce* $m^{(b)} : (0, \infty) \to [1, b)$ *taková, že každé* $x \in (0, \infty)$ *se dá vyjádřit ve tvaru* $x = m^{(b)}(x) \cdot b^n$ *pro nějaké* $n \in \mathbb{Z}$.

**Definice 1.3** $D_1 : (0, \infty) \to \{1, 2, \ldots, 9\}$ *je funkce určující první platnou číslici argumentu při vyjádření v desítkové soustavě.* $D_k : (0, \infty) \to \{0, 1, \ldots, 9\}$ *je pro* $k = 2, 3, \ldots$ *funkce určující k-tou platnou číslici.*

Například tedy platí $D_1(\pi) = D_1(10\pi) = 3$, $D_2(\pi) = 1$, $m(100\pi) = 3.1415\ldots$ a podobně.

**Definice 1.4** *Náhodná veličina* $X$ *má Benfordovo rozdělení, pokud platí*

$$\mathbf{P}\,(X < t) = \log_{10} t,\ t \in [1, 10].$$

**Tvrzení 1.5 (Benfordův zákon)** *V některých přirozeně se vyskytujících souborech číselných údajů je rozdělení čísel takové, že jejich mantisy (při zápisu čísel v desítkové soustavě) mají Benfordovo rozdělení, tedy platí*

$$\mathbf{P}\,(m(x) < t) = \log_{10} t,\ t \in [1, 10]. \tag{1.1}$$

Toto tvrzení trpí jedním nedostatkem. Říká pouze, že v *některých* souborech numerických údajů je rozdělení mantis logaritmické (Benfordovo). Nedává žádný návod, jak předem rozhodnout, který soubor dat tuto vlastnost mít bude a který ne. Ve druhé kapitole však budou formulovány a dokázány některé postačující podmínky pro to, aby se mantisy čísel řídily Benfordovým rozdělením.

Simon Newcomb i Frank Benford dospěli k vyjádření, které odpovídá vzorci (1.1), každý však jinou cestou. Newcombovy úvahy v [1] jdou popsat takto: každé kladné reálné číslo $x$ lze zapsat ve tvaru $x = 10^s$ pro nějaké $s \in \mathbb{R}$. Protože celá část čísla $s$ neovlivní první platnou číslici (ani mantisu $m(x)$), stačí uvažovat pouze $s$ mod 1. Po krátké úvaze dochází k závěru, že v případě „v přírodě se vyskytujících čísel“ má $s$ mod 1 rovnoměrné rozdělení na intervalu $(0, 1)$.

Nechť tedy $S$ je náhodná veličina s rovnoměrným rozdělením na intervalu $(0, 1)$, která odpovídá výše uvedeným hodnotám $s$ mod 1. Pak náhodná veličina $Y = 10^S$ má podle věty o transformaci hustoty (např. Anděl [3], věta 3.5) hustotu

$$f_Y(u) = \frac{1}{u \ln 10}\ I_{(1,10)}(u),$$

kde $I_{(1,10)}$ značí indikátor intervalu $(1, 10)$ a ln přirozený logaritmus ($\log_{10}$ označuje dekadický logaritmus). Pro $y \in [1, 10]$ tedy platí

$$\mathbf{P}(Y < y) = \int_1^y \frac{1}{u \ln 10} du = \frac{1}{\ln 10}[\ln u]_{u=1}^{y} = \frac{1}{\ln 10} \ln y = \log_{10} y.$$

Náhodná veličina $Y$ má tedy Benfordovo rozdělení. Pro realizaci $Y = y$ platí $y = m(y)$, protože $Y$ nabývá pouze hodnot z intervalu $(1, 10)$. Dále je $y = m(y) = m(10^{s \bmod 1}) = m(10^s) = m(x)$. To znamená, že veličina $Y$ popisuje mantisu původně uvažovaného čísla $x$. Podle Newcomba se tedy množina „všech čísel vyskytujících se v přírodě“ řídí Benfordovým zákonem.

Na rozdíl od Newcomba Benford (v článku [2]) svá tvrzení založil na empirických pozorováních. Několik let shromažďoval číselné údaje z různých zdrojů a oborů, například plochy povodí 335 řek, měrné skupenské teplo 1389 chemických sloučenin, čísla vyskytující se na titulní stránce novin a další. Dohromady zpracoval více než 20 000 údajů a ukázal, že první číslice se opravdu nevyskytují všechny stejně často.

Když hledal jednoduchý vzorec, kterým by mohl popsat rozdělení prvních platných číslic ve svých datech, dospěl k výrazu

$$\mathbf{P}\left(D_1(x) = d\right) = \log_{10}\left(1 + \frac{1}{d}\right),\ d = 1, 2, \ldots, 9. \tag{1.2}$$

Ten je ovšem důsledkem vztahu (1.1), protože pro $d = 1, 2, \ldots, 9$ je

$$\mathbf{P}\left(D_1(x) = d\right) \ = \ \mathbf{P}\left(d \le m(x) < d + 1\right)$$

$$
\begin{aligned}
&= \mathbf{P}\left(m(x) < d+1\right) - \mathbf{P}\left(m(x) < d\right) \\
&= \log_{10}(d+1) - \log_{10} d = \log_{10}\left(\frac{d+1}{d}\right).
\end{aligned}
$$

Podobně se ukáže, že

$$
\mathbf{P}\left(D_2(x) = d\right) = \sum_{k=1}^{9} \log_{10}\left(1 + \frac{1}{10k+d}\right),\ d = 0, 1, \ldots, 9.
$$

To také odpovídá hodnotám, které udává Newcomb. Následující tabulka ukazuje, s jakou pravděpodobností se jednotlivé číslice objeví na prvním, resp. druhém platném místě (s přesností na 4 desetinná místa).

| **Číslice** | $p_1$ | $p_2$ |
|---|---|---|
| 0 | | 0.1197 |
| 1 | 0.3010 | 0.1139 |
| 2 | 0.1761 | 0.1088 |
| 3 | 0.1249 | 0.1043 |
| 4 | 0.0969 | 0.1003 |
| 5 | 0.0792 | 0.0967 |
| 6 | 0.0669 | 0.0934 |
| 7 | 0.0580 | 0.0904 |
| 8 | 0.0512 | 0.0876 |
| 9 | 0.0458 | 0.0850 |

Vztah (1.1) dokonce určuje *sdružené rozdělení* veličin $D_1, D_2, \ldots$ Pro všechna $k \in \mathbb{N},\ d_1 \in 1, 2, \ldots, 9,\ d_j \in 0, 1, \ldots, 9,\ j = 2, \ldots, k,$ platí

$$
\begin{aligned}
&\mathbf{P}(D_1 = d_1, \ldots, D_k = d_k) \\
&\quad = \mathbf{P}\left(\sum_{i=1}^{k} d_i \cdot 10^{1-i} \le m(x) < \sum_{i=1}^{k-1} d_i \cdot 10^{1-i} + (d_k+1)\cdot 10^{1-k}\right) \\
&\quad = \log_{10}\left(\sum_{i=1}^{k-1} d_i \cdot 10^{1-i} + (d_k+1)\cdot 10^{1-k}\right) - \log_{10}\left(\sum_{i=1}^{k} d_i \cdot 10^{1-i}\right) \\
&\quad = \log_{10}\left(\frac{\sum_{i=1}^{k-1} d_i \cdot 10^{1-i} + (d_k+1)\cdot 10^{1-k}}{\sum_{i=1}^{k} d_i \cdot 10^{1-i}}\right) \\
&\quad = \log_{10}\left(\frac{10^{1-k}\cdot\left(\sum_{i=1}^{k-1} d_i \cdot 10^{k-i} + d_k + 1\right)}{10^{1-k}\cdot\left(\sum_{i=1}^{k} d_i \cdot 10^{k-i}\right)}\right) \\
&\quad = \log_{10}\left(\frac{\sum_{i=1}^{k} d_i \cdot 10^{k-i} + 1}{\sum_{i=1}^{k} d_i \cdot 10^{k-i}}\right) = \log_{10}\left(1 + \frac{1}{\sum_{i=1}^{k} d_i \cdot 10^{k-i}}\right). \qquad (1.3)
\end{aligned}
$$

Z toho ovšem krátkým výpočtem vyplyne, že jednotlivé číslice na sobě *nejsou nezávislé*:

$$\mathbf{P}(D_1 = 1) \doteq 0.3010$$
$$\mathbf{P}(D_2 = 2) \doteq 0.1088$$
$$\mathbf{P}(D_1 = 1, D_2 = 2) = \log_{10}\left(1 + \frac{1}{12}\right) \doteq 0.0348$$
$$\mathbf{P}(D_2 = 2 \mid D_1 = 1) = \frac{\mathbf{P}(D_1 = 1, D_2 = 2)}{\mathbf{P}(D_1 = 1)} \doteq 0.1155$$

Navíc stojí za povšimnutí, že s rostoucím $k$ se rozdělení veličiny $D_k$ blíží rovnoměrnému rozdělení na množině $\{0, 1, \ldots, 9\}$. Díky vyjádření (1.3) se dá totiž psát

$$\mathbf{P}(D_k = d_k) = \sum_{d_1=1}^{9}\sum_{d_2=0}^{9}\cdots\sum_{d_{k-1}=0}^{9} \log_{10}\left(\frac{\sum_{i=1}^{k} d_i \cdot 10^{k-i} + 1}{\sum_{i=1}^{k} d_i \cdot 10^{k-i}}\right),$$

a dále pro $d_k = 0, 1, \ldots, 8$:

$$\begin{aligned}
&\mathbf{P}(D_k = d_k) - \mathbf{P}(D_k = d_k + 1) \\
&= \sum_{d_1=1}^{9}\sum_{d_2=0}^{9}\cdots\sum_{d_{k-1}=0}^{9}\left[\log_{10}\left(\frac{\sum_{i=1}^{k-1} d_i \cdot 10^{k-i} + d_k + 1}{\sum_{i=1}^{k-1} d_i \cdot 10^{k-i} + d_k}\right)\right. \\
&\quad \left. - \log_{10}\left(\frac{\sum_{i=1}^{k-1} d_i \cdot 10^{k-i} + d_k + 2}{\sum_{i=1}^{k-1} d_i \cdot 10^{k-i} + d_k + 1}\right)\right] \\
&= \sum_{d_1=1}^{9}\sum_{d_2=0}^{9}\cdots\sum_{d_{k-1}=0}^{9} \log_{10}\left(\frac{\frac{\sum_{i=1}^{k} d_i \cdot 10^{k-i} + 1}{\sum_{i=1}^{k} d_i \cdot 10^{k-i}}}{\frac{\sum_{i=1}^{k} d_i \cdot 10^{k-i} + 2}{\sum_{i=1}^{k} d_i \cdot 10^{k-i} + 1}}\right) \\
&= \sum_{d_1=1}^{9}\sum_{d_2=0}^{9}\cdots\sum_{d_{k-1}=0}^{9} \log_{10}\left(\frac{\left(\sum_{i=1}^{k} d_i \cdot 10^{k-i} + 1\right)^2}{(\sum_{i=1}^{k} d_i \cdot 10^{k-i})(\sum_{i=1}^{k} d_i \cdot 10^{k-i} + 2)}\right).
\end{aligned}$$

Pro odhad předchozího výrazu se hodí uvažovat funkci $f(x) = \frac{(x+1)^2}{x(x+2)}$. Platí, že $f(x) > 1$ a $f'(x) < 0$ pro kladná $x$. V posledním výrazu výše jsou v závorce právě hodnoty $f(\sum_{i=1}^{k} d_i \cdot 10^{k-i})$, ty jsou větší než 1 a proto $\mathbf{P}(D_k = d_k) - \mathbf{P}(D_k = d_k + 1) > 0$.

Protože funkce $f(x)$ je klesající a $\log_{10}(y)$ je rostoucí, je funkce $\log_{10}(f(x))$ klesající a v kontextu předchozího výpočtu nabývá svého maxima v bodě

$x = 10^{k-1}$, to je totiž nejmenší přípustná hodnota výrazu $\sum_{i=1}^{k} d_i \cdot 10^{k-i}$. To dává odhad

$$\begin{aligned}\mathbf{P}(D_k = d_k) - \mathbf{P}(D_k = d_k + 1) &\leq 9 \cdot 10^{k-2} \cdot \log_{10}\left(\frac{\left(10^{k-1}+1\right)^2}{10^{k-1}\cdot(10^{k-1}+2)}\right) \\ &= \frac{9}{10}\cdot 10^{k-1}\cdot\log_{10}\left(\frac{\left(1+\frac{1}{10^{k-1}}\right)^2}{1+\frac{2}{10^{k-1}}}\right) \\ &= \frac{9}{10}\cdot\log_{10}\left(\frac{\left(1+\frac{1}{10^{k-1}}\right)^{2\cdot 10^{k-1}}}{\left(1+\frac{2}{10^{k-1}}\right)^{10^{k-1}}}\right) \longrightarrow \frac{9}{10}\cdot\log_{10}\left(\frac{(e^1)^2}{e^2}\right), k\to\infty,\end{aligned}$$

a tedy platí

$$0 \leq \lim_{k\to\infty}\left(\mathbf{P}(D_k = d_k) - \mathbf{P}(D_k = d_k+1)\right) \leq \frac{9}{10}\cdot\log_{10}\frac{e^2}{e^2} = \frac{9}{10}\cdot\log_{10} 1 = 0.$$

Z toho plyne, že veličiny $D_k$ pro $k$ rostoucí nade všechny meze konvergují v distribuci k veličině s rovnoměrným rozdělením na množině $\{0, 1, \ldots, 9\}$.

## 1.2 Pokusy o vysvětlení

V průběhu let se objevilo mnoho pokusů o vysvětlení Benfordova zákona, většinou čistě matematické povahy. Některé z nich budou v následujících odstavcích nastíněny. Často se autor snažil ukázat, že množina reálných (případně přirozených) čísel splňuje (1.1) a z toho učinit závěr, že fenomén nerovnoměrného rozdělení první číslice je jednoduše vlastností používaného číselného systému.

Cílem je na $\mathbb{N}$, resp. na $\mathbb{R}$ zavést pravděpodobnostní míru, která bude určovat rozdělení první platné číslice (případně mantisy). Typicky je prvním krokem určit pravděpodobnost, že přirozené číslo $n$ má první platnou číslici 1, tj. patří do množiny $\{D_1 = 1\} = \{k \in \mathbb{N}, D_1(k) = 1\}$. Buď $\alpha(n)$ indikátor množiny $\{D_1 = 1\}$, tedy $\alpha(n) = 1$ pro $n \in \{D_1 = 1\}$ a $\alpha(n) = 0$ jinak. Pak by se zdálo přirozené definovat

$$\mathbf{P}(D_1 = 1) = \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\alpha(k).$$

Tato limita ale neexistuje, hodnoty výrazu oscilují mezi přibližně $\frac{1}{9}$ pro $n = 10^k,\ k \in \mathbb{N}$ a $\frac{5}{9}$ pro $n = 2\cdot 10^k,\ k \in \mathbb{N}$.

To vedlo k používání různých zobecněných sčítacích metod, které měly množině $\{D_1 = 1\}$ přiřadit „správnou“ pravděpodobnost $\log_{10} 2$. V tomto

ohledu je zajímavá práce Flehinger [4]. Používá iterace Cesàrovy sčítací metody:

$$\alpha_1(k) = \frac{1}{k}\sum_{i=1}^{k}\alpha(i),$$

$$\alpha_t(k) = \frac{1}{k}\sum_{i=1}^{k}\alpha_{t-1}(i)$$

a dokáže, že ačkoliv funkce $\alpha_t(k)$ jsou pořád oscilující, celý iterační proces konverguje ke kýžené hodnotě ve smyslu

$$\lim_{t\to\infty}\liminf_{k\to\infty}\alpha_t(k) = \lim_{t\to\infty}\limsup_{k\to\infty}\alpha_t(k) = \log_{10}2.$$

Žádný z těchto postupů nevede k uspokojivému výsledku i přesto, že uvažované sčítací metody jsou regulární (pro konvergentní řady dávají jako součet limitu částečných součtů) a množinám typu $\{D_1 = 1\}$ přiřazují požadovanou pravděpodobnost. Jak totiž píše Raimi [5], existuje mnoho regulárních sčítacích metod, které jim naopak přiřazují jiné pravděpodobnosti, a nelze *a priori* rozhodnout, která sčítací metoda je „správná“.

Ve spojitém případě (kdy byla místo přirozených čísel uvažována kladná reálná čísla) byly použity různé integrační metody, metody Fourierovy analýzy i teorie Banachových měr, žádný postup však nevyústil v zavedení požadované pravděpodobnosti na $\mathbb{R}^+$ ve smyslu $\sigma$-aditivní množinové funkce. Raimi [5] podává obsáhlý přehled výsledků v diskrétním i spojitém případě.

Dvě další hypotézy jsou často zmiňovány v souvislosti s Benfordovým zákonem: *invariance vzhledem ke změně základu* a *invariance vzhledem ke změně měřítka*. První z nich říká, že by se data, která se řídí Benfordovým zákonem, měla řídit jeho obdobou i při vyjádření čísel v soustavě o libovolném jiném základu, nejen v desítkové soustavě.

Invariance vzhledem ke změně měřítka je požadavek, aby to, zda se soubor číselných údajů Benfordovým zákonem řídí nebo ne, nezáviselo na tom, v jakých jednotkách jsou údaje vyjádřeny, zda v litrech, galonech, apod. Benfordův zákon tedy musí zůstat v platnosti, i pokud jsou uvažované údaje přenásobeny libovolnou kladnou konstantou. To dává podmínku, že $\mathbf{P}(X \in (0,1)) = \mathbf{P}(X \in (0,s))$ pro všechna $s \in (0,\infty)$, kde $X$ je náhodná veličina s rozdělením, které odpovídá rozdělení čísel ve zkoumaném souboru.

Například Pinkham [6] uvažuje abstraktní „soubor všech čísel“ ve smyslu hodnot vyhledávaných v logaritmických tabulkách a distribuční funkci $F$ určující jejich rozdělení. Z předpokladu invariance vzhledem ke změně měřítka a spojitosti $F$ (aby se žádné konkrétní číslo z uvažovaného souboru všech čísel vyhledávaných v logaritmických tabulkách neobjevovalo s kladnou pravděpodobností) pak ukáže, že rozdělení mantis je Benfordovo.

Problém ovšem je, že na $(0, \infty)$ neexistuje borelovská pravděpodobnostní míra invariantní vzhledem k měřítku. Pokud by $\mu$ byla borelovská pravděpodobnostní míra na $(0, \infty)$ splňující

$$\mu(0, 1) = \mu(0, s), \ \forall s \in (0, \infty),$$

vyjde podle věty o spojitosti míry a rostoucí posloupnosti $\mu-$měřitelných množin (např. Jarník [7], věta 23), že $\mu(0, s) = 0$ pro všechna $s \in (0, \infty)$. Stejná věta potom dává

$$\mu(0, \infty) = 0,$$

ale protože $\mu$ je pravděpodobnostní míra, musí být $\mu(0, \infty) = 1$, a to je spor. Způsob, jak zdůvodnit platnost Benfordova zákona, je třeba hledat jinde.

# Kapitola 2

# Náhodné výběry z náhodně vybíraných rozdělení

## 2.1 Úvod

Protože už Newcomb formuloval svá tvrzení jako pravděpodobnostní problém – „jaká je pravděpodobnost, že vybrané číslo bude mít na prvním platném místě číslici $d$“ – snažil se v devadesátých letech dvacátého století T. P. Hill zasadit Benfordův zákon do rámce moderní teorie pravděpodobnosti. V této kapitole bude předvedena část výsledků, kterých dosáhl, tak, jak je publikoval v článcích [8] a [9].

Výchozím bodem Hillových úvah je názor, že borelovská $\sigma$-algebra na $\mathbb{R}^+$ není pro zkoumání Benfordova zákona vhodná. Místo ní pracuje s tzv. *mantisovou $\sigma$-algebrou* (dokud nebude explicitně řečeno jinak, uvažuje se mantisa při vyjádření čísel v desítkové soustavě).

**Definice 2.1** *Mantisová $\sigma$-algebra $\mathcal{M}$ na $\mathbb{R}^+$ je $\sigma$-algebra generovaná funkcí mantisa: $x \mapsto m(x)$.*

Mantisová $\sigma$-algebra $\mathcal{M}$ je generována množinami typu

$$\bigcup_{n=-\infty}^{\infty} [a,b) \cdot 10^n,\ 1 \leq a < b < 10,$$

tedy je obsažena v borelovské $\sigma$-algebře na $\mathbb{R}^+$ a jde popsat takto:

$$S \in \mathcal{M} \Leftrightarrow S = \bigcup_{n=-\infty}^{\infty} B \cdot 10^n \text{ pro nějakou borelovskou } B \subseteq [1,10). \quad (2.1)$$

Každá $S \in \mathcal{M}$ má tedy neprázdný průnik se všemi intervaly typu $[10^k, 10^{k+1})$, $k \in \mathbb{Z}$.

**Definice 2.2** *Nechť $a > 0$ a $S \in \mathcal{M}$, pak symbolem $aS$ bude označena množina $\{as,\ s \in S\}$ a podobně $S^a$ bude značit množinu $\{s^a,\ s \in S\}$.*

**Tvrzení 2.3** *$\sigma$-algebra $\mathcal{M}$ má následující vlastnosti:*

*(i) každá neprázdná množina $S \in \mathcal{M}$ je neomezená s hromadnými body v 0 a $+\infty$,*

*(ii) $\mathcal{M}$ je uzavřená vzhledem k násobení kladnými reálnými čísly, tj. pro $s > 0$, $S \in \mathcal{M}$ je $sS \in \mathcal{M}$,*

*(iii) $\mathcal{M}$ je uzavřená vzhledem k odmocňování, tj. pro $m \in \mathbb{N}, S \in \mathcal{M}$ je $S^{1/m} \in \mathcal{M}$, ale ne vzhledem k umocňovnání,*

*(iv) $\mathcal{M}$ má vlastnost soběpodobnosti v tomto smyslu: je-li $S \in \mathcal{M}$, pak $10^m S = S$ pro každé $m \in \mathbb{Z}$.*

První vlastnost říká, že konečné intervaly jako $[1, 2)$ nejsou v $\mathcal{M}$ (tzn. nejdou popsat pomocí funkce mantisa), a tím pádem je odstraněn problém s neexistencí borelovské pravděpodobnostní míry invariantní vzhledem k měřítku, který byl uveden v části 1.2. Vlastnost (iii) si zaslouží podrobnější prozkoumání. Pro $S \in \mathcal{M}$ se může množina $S^{1/2}$ skládat ze dvou „částí“, a podobně pro další odmocniny. Například

$$S = \{D_1 = 1\} = \bigcup_{n=-\infty}^{\infty} [1, 2) \cdot 10^n,$$

$$S^{1/2} = \bigcup_{n=-\infty}^{\infty} [1, \sqrt{2}) \cdot 10^n \cup \bigcup_{n=-\infty}^{\infty} [\sqrt{10}, \sqrt{20}) \cdot 10^n \in \mathcal{M}, \text{ ale}$$

$$S^2 = \bigcup_{n=-\infty}^{\infty} [1, 4) \cdot 10^{2n} \notin \mathcal{M},$$

protože $S^2$ má prázdný průnik s intervaly typu $[10^{2k+1}, 10^{2k+2}), k \in \mathbb{Z}$ a nejde popsat pomocí funkce mantisa.

Vlastnost (ii) $\sigma$-algebry $\mathcal{M}$ umožňuje formalizovat hypotézu invariance vzhledem ke změně měřítka a podobně vlastnost (iii) je klíčem k hypotéze o invarianci vzhledem ke změně základu.

## 2.2 Invariance vzhledem ke změně měřítka a základu

Číselné údaje v uvažovaných souborech dat mohou být vyjádřeny v různých jednotkách. Například v tabulce délek britských řek (v mílích) znamená převod z mílí na kilometry vynásobení všech čísel koeficientem 1.6094. Takových převodů je ale libovolně mnoho, protože není žádné omezení na to, jaké jednotky se ta která skupina lidí rozhodne používat.

Hypotéza o invarianci vzhledem ke změně měřítka dává požadavek, aby se soubor údajů, který se řídí Benfordovým zákonem, tímto zákonem řídil i při použití libovolných jiných jednotek.

**Definice 2.4** *Pravděpodobnostní míra $P$ na měřitelném prostoru $(\mathbb{R}^+, \mathcal{M})$ je invariantní vzhledem ke změně měřítka, pokud $P(S) = P(sS)$ pro všechna $s > 0$ a každou $S \in \mathcal{M}$.*

Uvedená definice je korektní, protože množiny typu $sS$ jsou $\mathcal{M}-$měřitelné díky vlastnosti 2.3 (ii). Vlastnost invariance vzhledem ke změně měřítka dokonce charakterizuje Benfordův zákon (1.1).

**Věta 2.5** *Pravděpodobnostní míra $P$ na $(\mathbb{R}^+, \mathcal{M})$ je invariantní vzhledem ke změně měřítka právě tehdy, když*

$$P\left(\bigcup_{n=-\infty}^{\infty} [1,t)\cdot 10^n\right) = \log_{10} t \quad \textit{pro všechna } t \in [1,10). \tag{2.2}$$

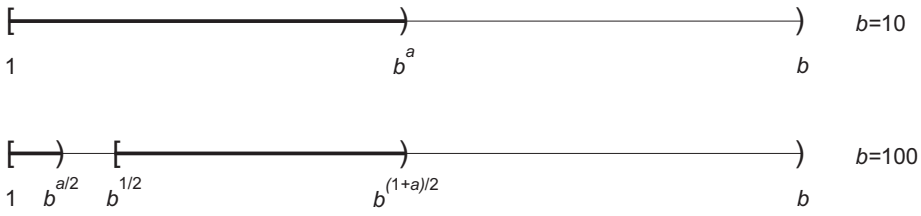
Důkaz je možné nalézt v [9].

Zdá se rozumné požadovat, aby zákon popisující rozdělení prvních platných číslic, má-li mít univerzální platnost, fungoval stejně i po přepsání do číselné soustavy o jiném základu než 10 (v soustavě o základu $b$ se bude pracovat s analogií Benfordova rozdělení, kde se $\log_{10}$ nahradí $\log_b$). Právě to je hypotéza o invarianci vzhledem ke změně základu.

Jako motivaci před zavedením pojmu pravděpodobnostní míry invariantní vzhledem ke změně základu lze uvažovat množinu $S \in \mathcal{M}$ čísel, která mají při zápisu v desítkové soustavě mantisu menší než 5. Na následujících řádcích bude $m^{(10)}$ značit funkci mantisa při zápisu čísel v soustavě o základu 10 a $m^{(100)}$ funkci mantisa při zápisu v soustavě o základu 100. Je tedy

$$S = \{1 \le m^{(10)} < 5\} = \{1 \le m^{(100)} < 5\} \cup \{10 \le m^{(100)} < 50\}.$$

Obrázek 2.1 graficky znázorňuje (při vyjádření v číselné soustavě o základu $b$, $b = 10$ a 100) množinu $S$, respektive její průnik s intervalem $[1, b)$.




Obrázek 2.1: K invarianci vzhledem ke změně základu ($a = \log_{10} 5$).

Dá se říct (za předpokladu, že nezáleží na základu použité číselné soustavy), že Benfordův zákon se neptá, jak velké jsou dílky na ose, ale pouze jakou část kladné reálné poloosy množina $S$ zabírá.

**Definice 2.6** *Pravděpodobnostní míra $P$ na měřitelném prostoru $(\mathbb{R}^+, \mathcal{M})$ je invariantní vzhledem ke změně základu, pokud $P(S) = P\left(S^{1/n}\right)$ pro všechna přirozená $n$ a každou $S \in \mathcal{M}$.*

Měřitelnost množin $S^{1/n}$ je zaručena vlastností 2.3 (iii). Podle vyjádření (2.1) nemá množina

$$S_1 = \{m^{(10)} = 1\} = \{\ldots, 0.1, 1, 10, 100, \ldots\} = \bigcup_{n=-\infty}^{\infty} \{1\} \cdot 10^n \in \mathcal{M}$$

žádné neprázdné $\mathcal{M}$-měřitelné podmnožiny, proto je Diracova míra $\delta_1$ množiny $S_1$ dobře definována (pokud $S \in \mathcal{M}$, pak $\delta_1(S) = 1$ pro $S_1 \subseteq S$ a $\delta_1(S) = 0$ jinak).

Nechť $P_L$ je pravděpodobnostní míra definovaná vzorcem (2.2), tedy

$$P_L\left(\bigcup_{n=-\infty}^{\infty} [1, t) \cdot 10^n\right) = \log_{10} t \quad \text{pro všechna } t \in [1, 10),$$

potom úplnou charakterizaci pravděpodobnostních měr invariantních vzhledem ke změně základu dává následující věta.

**Věta 2.7** *Pravděpodobnostní míra $P$ na $(\mathbb{R}^+, \mathcal{M})$ je invariantní vzhledem ke změně základu právě tehdy, když*

$$P = qP_L + (1 - q)\delta_1 \quad \textit{pro nějaké } q \in [0, 1].$$

Důkaz této věty je k dispozici v článku [9].

Věty 2.5 a 2.7 dohromady říkají, že invariance vzhledem ke změně měřítka implikuje invarianci vzhledem ke změně základu, ale ne naopak (například míra $\delta_1$ je invariantní vzhledem ke změně základu, ale ne měřítka).

## 2.3 Náhodné výběry z náhodně vybíraných rozdělení

Výše uvedená tvrzení jsou po matematické stránce správná, ale jak vysvětlí výskyt dat řídících se Benfordovým zákonem ve skutečném světě? V případě údajů sesbíraných Frankem Benfordem vykazovaly některé skupiny dobrou shodu s Benfordovým zákonem, a některé (jako třeba odmocniny přirozených čísel) v podstatě žádnou. Nejblíže Benfordovu zákonu byl ale souhrn všech 20 229 údajů z dvaceti různých skupin.

Spíše než to, že by všechny údaje pocházely ze stejného rozdělení „všech konstant světa“ (Pinkham [6]), je vhodné předpokládat, že údaje pocházejí z různých rozdělení. Frank Benford v článku [2] uvádí, že se skutečně snažil „sesbírat údaje z co nejvíce různých oblastí“.

Tady přichází ke slovu *náhodné pravděpodobnostní míry*. Reálná borelovská náhodná pravděpodobnostní míra $\mathbf{M}$ je náhodné zobrazení na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathbf{P})$, jehož hodnotami jsou borelovské pravděpodobnostní míry na $\mathbb{R}$ a které je regulární v tom smyslu, že pro každou borelovskou množinu $B \subset \mathbb{R}$ je $\mathbf{M}(B)$ náhodná veličina (pro podrobnější informace o náhodných pravděpodobnostních mírách viz Kallenberg [10]).

**Definice 2.8** *Očekávané rozdělení náhodné pravděpodobnostní míry* $\mathbf{M}$ *je pravděpodobnostní míra* $E\mathbf{M}$ *na* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *taková, že*

$$(E\mathbf{M})(B) = E[\mathbf{M}(B)] \quad \forall B \in \mathcal{B}(\mathbb{R}),$$

*kde* $\mathcal{B}(\mathbb{R})$ *značí* $\sigma$*-algebru borelovských množin na* $\mathbb{R}$.

Například pokud je $\mathbf{M}$ náhodná pravděpodobnostní míra, která s pravděpodobností $\frac{1}{2}$ určuje rovnoměrné rozdělení na $[0, 1]$ a jinak určuje exponenciální rozdělení se střední hodnotou 1, je očekávané rozdělení $E\mathbf{M}$ spojité rozdělení s hustotou

$$f(x) = \frac{1 + e^{-x}}{2} \text{ pro } 0 \leq x \leq 1, \quad f(x) = \frac{e^{-x}}{2} \text{ pro } x > 1.$$

Následující definice formalizuje myšlenku shromažďování údajů pocházejících z různých rozdělení. Dá se chápat takto: náhodně se vybere rozdělení a z něj se napozoruje výběr o rozsahu $k$, potom se náhodně vybere další rozdělení a z něj se napozoruje druhý výběr o rozsahu $k$, a tak dále.

**Definice 2.9** *Nechť* $\mathbf{M}$ *je náhodná pravděpodobnostní míra,* $\mathbf{M}_1, \mathbf{M}_2, \ldots$ *je posloupnost nezávislých stejně rozdělených náhodných pravděpodobnostních měr se stejným rozdělením jako* $\mathbf{M}$ *a* $k \in \mathbb{N}$*, potom posloupnost* $\mathbf{M}$*-náhodných k-výběrů je posloupnost náhodných veličin* $X_1, X_2, \ldots$ *na pravděpodobnostním prostoru* $(\Omega, \mathcal{A}, \mathbf{P})$ *taková, že pro každé* $j = 1, 2, \ldots$ *platí:*

*(i) při daném* $\mathbf{M}_j = P$ *jsou náhodné veličiny* $X_{(j-1)k+1}, \ldots, X_{jk}$ *nezávislé a stejně rozdělené s rozdělením* $P$*,*

*(ii)* $X_{(j-1)k+1}, \ldots, X_{jk}$ *jsou nezávislé na* $\{\mathbf{M}_i, X_{(i-1)k+1}, \ldots, X_{ik}\}$ *pro všechna* $i \neq j$*.*

Tyto posloupnosti mají zvláštní strukturu, jak ukáže následující lemma. Toto lemma je formulováno a dokázáno v poněkud silnější podobě než v článku [8].

**Lemma 2.10** *Nechť* $X_1, X_2, \ldots$ *je posloupnost* $\mathbf{M}$*-náhodných* $k$*-výběrů pro nějakou náhodnou pravděpodobnostní míru* $\mathbf{M}$ *a* $k \in \mathbb{N}$*, potom*

*(i)* $\{X_n\}$ *jsou stejně rozdělené s rozdělením* $E\mathbf{M}$*, ale obecně nejsou nezávislé,*

*(ii) při daných* $\{\mathbf{M}_1, \mathbf{M}_2, \ldots\}$ *jsou* $\{X_n\}$ *nezávislé, ale obecně nejsou stejně rozdělené.*

*Důkaz.* (i) Při vyjádření pomocí podmíněné pravděpodobnosti a podmíněné střední hodnoty platí pro všechny borelovské $B \subset \mathbb{R}$:

$$\mathbf{P}\left(X_j \in B \mid \mathbf{M}_j\right) = \mathbf{M}_j(B) \text{ skoro jistě,}$$

$$\mathbf{P}\left(X_j \in B\right) = E\left[\mathbf{P}\left(X_j \in B \mid \mathbf{M}_j\right)\right] = E\left[\mathbf{M}_j(B)\right] = E\left[\mathbf{M}(B)\right],$$

přičemž poslední rovnost plyne z toho, že $\mathbf{M}_j$ má stejné rozdělení jako $\mathbf{M}$. Proto všechna $X_j$ mají stejné rozdělení.

Druhá část tvrzení (i) plyne z toho, že výběry z neznámého rozdělení o něm mohou poskytovat určité informace, jak bude ukázáno na příkladě.

(ii) Nezávislost veličin $\{X_n\}$ plyne z definice 2.9, druhá část tvrzení pak z toho, že $X_{ik}$ má jiné rozdělení než $X_{jk}$, kdykoliv $\mathbf{M}_i \neq \mathbf{M}_j$.

□

Následující příklad (převzatý z [8]) ukáže, že posloupnost $\mathbf{M}$-náhodných $k$-výběrů $\{X_n\}$ obecně není stacionární ani markovská, netvoří martingal, $\{X_n\}$ nejsou nezávislé a nelze zaměňovat jejich pořadí.

Nechť $\mathbf{M}$ je náhodná pravděpodobnostní míra, která je s pravděpodobností $\frac{1}{2}$ Diracova míra $\delta(1)$ v bodě 1 a jinak je $\frac{\delta(1)+\delta(2)}{2}$, a $k = 3$.

Potom

$$\mathbf{P}\left((X_1, X_2, X_3) = (1,1,1)\right) = \frac{9}{16},$$

$$\mathbf{P}\left((X_2, X_3, X_4) = (1,1,1)\right) = \frac{15}{32},$$

a tedy posloupnost $\{X_n\}$ není stacionární.

Protože $\mathbf{P}(X_3 = 1|X_1 = X_2 = 1) = \frac{9}{10} \neq \frac{5}{6} = \mathbf{P}(X_3 = 1|X_2 = 1)$, netvoří posloupnost $\{X_n\}$ Markovův řetězec.

Je $E(X_2|X_1 = 2) = \frac{3}{2}$ a proto posloupnost $\{X_n\}$ není martingal.

Navíc $\mathbf{P}(X_2 = 2) = \frac{1}{4}$, ale $\mathbf{P}(X_2 = 2 \mid X_1 = 2) = \frac{1}{2}$, proto $X_1$ a $X_2$ nejsou nezávislé.

Dále platí

$$\mathbf{P}\left((X_1, X_2, X_3, X_4) = (1, 1, 1, 2)\right) = \frac{9}{64},$$
$$\mathbf{P}\left((X_1, X_2, X_3, X_4) = (2, 1, 1, 1)\right) = \frac{3}{64},$$

a tedy v posloupnosti $\{X_n\}$ nelze zaměňovat pořadí jednotlivých veličin.

Další lemma popíše asymptotické chování posloupnosti $\mathbf{M}$-náhodných $k$-výběrů.

**Lemma 2.11** *Nechť $\mathbf{M}$ je náhodná pravděpodobnostní míra a $\{X_n\}$ je posloupnost $\mathbf{M}$-náhodných $k$-výběrů pro nějaké $k \in \mathbb{N}$. Potom pro všechny borelovské množiny $B \subset \mathbb{R}$ platí*

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} I_{\{X_i \in B\}}}{n} = E\left[\mathbf{M}(B)\right] \textit{ skoro jistě},$$

*kde $I_{\{X_i \in B\}}$ je indikátor množiny $\{X_i \in B\}$.*

*Důkaz.* Nechť $B$ a $j \in \mathbb{N}$ jsou daná a nechť

$$Y_j = \sum_{m=1}^{k} I_{\{X_{(j-1)k+m} \in B\}}.$$

Pak platí, pokud limity existují,

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} I_{\{X_i \in B\}}}{n} = \lim_{m\to\infty} \frac{\sum_{j=1}^{m} Y_j}{km}. \tag{2.3}$$

Podle bodu (i) v definici 2.9 mají veličiny $Y_j$ při daném $\mathbf{M}_j$ binomické rozdělení s parametry $k$ a $E[\mathbf{M}_j(B)]$, a tedy pro všechna $j$

$$EY_j = E[E[Y_j \mid \mathbf{M}_j]] = E[kE[\mathbf{M}_j(B)]] = kE[\mathbf{M}_j(B)] = kE[\mathbf{M}(B)],$$

přičemž poslední rovnost plyne z toho, že $\mathbf{M}_j$ a $\mathbf{M}$ mají podle definice 2.9 stejné rozdělení.

Podle bodu (ii) v definici 2.9 jsou $Y_j$ nezávislé. Pro každé $j$ nabývá $Y_j$ pouze hodnot 1,2,...,$k$, rozptyl $Y_j$ se tedy dá pro všechna $j$ omezit stejnou konstantou a proto

$$\sum_{j=1}^{\infty} \frac{varY_j}{j^2} < \infty.$$

Silný zákon velkých čísel pro nestejně rozdělené náhodné veličiny (např. Dupač, Hušková [11], věta 4.6) potom dává

$$\lim_{m\to\infty} \frac{\sum_{j=1}^{m} Y_j}{m} = kE[\mathbf{M}(B)] \text{ skoro jistě.}$$

Platí tedy

$$\lim_{m\to\infty} \frac{\sum_{j=1}^{m} Y_j}{km} = E[\mathbf{M}(B)] \text{ skoro jistě}$$

a tvrzení lemmatu bezprostředně plyne z vyjádření (2.3).

□

Předpoklad, že z rozdělení $\mathbf{M}_j$ se bere vždy výběr o rozsahu $k$, není nezbytný. Tvrzení platí i tehdy, když se z $\mathbf{M}_j$ bere výběr o rozsahu $K_j$, kde $\{K_j\}$ jsou nezávislé náhodné veličiny s hodnotami v $\mathbb{N}$, které se dají omezit stejnou konstantou a jsou také nezávislé na všech ostatních veličinách $X_i$ a $\mathbf{M}_i$. Důkaz je v tomto případě o něco techničtější.

## 2.4 Statistické odvození

Téměř vše je nyní připraveno k formulaci limitní věty o chování prvních platných číslic. Tato věta říká zhruba toto: pokud jsou náhodně vybírána pravděpodobnostní rozdělení a z nich se potom berou náhodné výběry tak, že celý proces nezvýhodňuje nějakou volbu jednotek nebo základu číselné soustavy, pak rozdělení mantis v celém souboru bude konvergovat k Benfordovu rozdělení. Tím tato věta pomáhá vysvětlit nebo předpovídat shodu souborů číselných údajů s Benfordovým zákonem.

**Definice 2.12** *Posloupnost náhodných veličin $X_1, X_2, \ldots$ má hodnoty mantisy nezávislé na měřítku, pokud*

$$\frac{|\sum_{i=1}^{n} I_{\{X_i\in S\}} - \sum_{i=1}^{n} I_{\{X_i\in sS\}}|}{n} \xrightarrow{n\to\infty} 0 \text{ skoro jistě}$$

*pro všechna $s > 0$ a všechna $S \in \mathcal{M}$, a má hodnoty mantisy nezávislé na číselném základu, pokud*

$$\frac{|\sum_{i=1}^{n} I_{\{X_i\in S\}} - \sum_{i=1}^{n} I_{\{X_i\in S^{1/m}\}}|}{n} \xrightarrow{n\to\infty} 0 \text{ skoro jistě}$$

*pro všechna $m \in \mathbb{N}$ a všechna $S \in \mathcal{M}$.*

Jako příklad poslouží posloupnosti $\{X_n\}$, $\{Y_n\}$ a $\{Z_n\}$, kde $X_n \equiv 1$, $Y_n \equiv 2$, $Z_n = 2^n$. Pak $\{X_n\}$ má hodnoty mantisy nezávislé na číselném základu, ale ne na měřítku, $\{Y_n\}$ nemá hodnoty mantisy nezávislé ani na měřítku, ani na číselném základu, a $\{Z_n\}$ má podle [8] hodnoty mantisy nezávislé na obojím.

Pěkný příklad „ze života“ je následující: náhodně se vybere evropská společnost vyrábějící nápoje a zaznamená se objem $k$ jejích náhodně vybraných produktů, potom se vybere druhá společnost atd. V tomto případě budou pravděpodobně hodnoty silně vázány na konkrétní jednotku, litr, a proto zaznamenaná posloupnost čísel bude mít hodnoty mantisy *závislé* na měřítku. Převod čísel na jiné jednotky, například pinty nebo galony, povede nejspíše k výrazně odlišnému rozdělení mantis.

Pokud je však náhodně vybírán druh savce žijící v Evropě a zjišťován objem $k$ náhodně vybraných zástupců tohoto druhu, bude tento proces pravděpodobně méně záviset na výběru jednotek.

Podobně lze náhodně vybírat města a u $k$ náhodně vybraných obyvatel zapisovat počet prstů na rukou – takto vzniklá posloupnost bude mít zajisté hodnoty mantisy výrazně závislé na číselném základu 10. Na druhou stranu pokud v těchto městech bude zkoumán počet listů $k$ náhodně vybraných stromů, bude mít tato posloupnost hodnoty mantisy mnohem méně závislé na číselném základu.

Protože mantisová $\sigma$-algebra $\mathcal{M}$ je obsažena v borelovské $\sigma$-algebře na $\mathbb{R}$, indukuje každá borelovská pravděpodobnostní míra na $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ jednoznačně určenou pravděpodobnostní míru na $(\mathbb{R}^+, \mathcal{M})$ ve smyslu restrikce míry na menší systém množin a tudíž je následující definice korektní.

**Definice 2.13** *Náhodná pravděpodobnostní míra* $\mathbf{M}$ *je nevychýlená měřítkem, pokud její očekávané rozdělení* $E\mathbf{M}$ *je invariantní vzhledem ke změně měřítka na* $(\mathbb{R}^+, \mathcal{M})$*, a* $\mathbf{M}$ *je nevychýlená základem, pokud její očekávané rozdělení* $E\mathbf{M}$ *je invariantní vzhledem ke změně základu na* $(\mathbb{R}^+, \mathcal{M})$.

Klíčovým bodem této definice je fakt, že nevyžaduje, aby konkrétní realizace $\mathbf{M}$ byly invariantní vzhledem ke změně měřítka nebo základu (často se dokonce stává, že žádná realizace $\mathbf{M}$ nemá požadovanou vlastnost), ale pouze aby celý proces sbírání údajů *v průměru* nezvýhodňoval nějakou volbu jednotek nebo základu číselné soustavy.

**Definice 2.14** *Pravděpodobnostní míra* $\mathbf{P}$ *nemá atomy na měřitelném prostoru* $(\mathrm{S}, \mathcal{S})$*, pokud pravděpodobnostní prostor* $(\mathrm{S}, \mathcal{S}, \mathbf{P})$ *nemá atomy, neboli v* $\sigma$*-algebře* $\mathcal{S}$ *není žádná množina* $A$, $\mathbf{P}(A) > 0$ *taková, že pro každou* $B \in \mathcal{S}$, $B \subseteq A$, $\mathbf{P}(B) < \mathbf{P}(A)$ *je už nutně* $\mathbf{P}(B) = 0$.

V následujícím tvrzení bude $\mathbf{M}(t)$ označovat náhodnou veličinu $\mathbf{M}(D_t)$, kde $D_t = \bigcup_{n=-\infty}^{\infty}[1,t)\cdot 10^n$ je množina kladných čísel, jejichž mantisa je v intervalu $[1,t)$.

**Věta 2.15** *Nechť* $\mathbf{M}$ *je náhodná pravděpodobnostní míra na* $(\mathbb{R}^+,\mathcal{M})$. *Následující tvrzení jsou ekvivalentní:*

*(i)* $\mathbf{M}$ *je nevychýlená měřítkem,*

*(ii)* $\mathbf{M}$ *je nevychýlená základem a* $E\mathbf{M}$ *nemá atomy na* $(\mathbb{R}^+,\mathcal{M})$,

*(iii)* $E[\mathbf{M}(t)] = \log_{10} t$ *pro všechna* $t \in [1,10)$,

*(iv) každý* $\mathbf{M}$*-náhodný k-výběr má hodnoty mantisy nezávislé na měřítku,*

*(v)* $E\mathbf{M}$ *nemá atomy na* $(\mathbb{R}^+,\mathcal{M})$ *a každý* $\mathbf{M}$*-náhodný k-výběr má hodnoty mantisy nezávislé na číselném základu,*

*(vi) každý* $\mathbf{M}$*-náhodný k-výběr* $X_1, X_2, \ldots$ *splňuje*

$$\frac{\sum_{i=1}^{n} I_{\{m(X_i)\in[1,t)\}}}{n} \xrightarrow{n\to\infty} \log_{10} t \text{ s.j. pro všechna } t \in [1,10).$$

*Důkaz.* „(i) ⇔ (iii)“ Nechť nejprve $\mathbf{M}$ je nevychýlená měřítkem. To podle definice 2.13 znamená, že $E\mathbf{M}$ je invariantní vzhledem ke změně měřítka na $(\mathbb{R}^+,\mathcal{M})$, a podle definice 2.8 a věty 2.5 platí

$$E[\mathbf{M}(t)] = (E\mathbf{M})(D_t) = \log_{10} t \text{ pro všechna } t \in [1,10).$$

Stejnou úvahou se dokáže i opačná implikace.

„(ii) ⇔ (iii)“ To, že $E\mathbf{M}$ je invariantní vzhledem ke změně základu ($\mathbf{M}$ je nevychýlená základem), spolu s faktem, že $E\mathbf{M}$ nemá atomy na $(\mathbb{R}^+,\mathcal{M})$, je podle věty 2.7 ekvivalentní tomu, že $E\mathbf{M} = P_L$, kde $P_L$ je pravděpodobnostní míra definovaná před větou 2.7. Ta splňuje $P_L(D_t) = \log_{10} t$ pro všechna $t \in [1,10)$.

„(iii) ⇔ (iv)“ Podle lemmatu 2.11:

$$A_n := \frac{\sum_{i=1}^{n} I_{\{X_i\in S\}}}{n} \to E[\mathbf{M}(S)] \text{ skoro jistě,}$$

$$B_n := \frac{\sum_{i=1}^{n} I_{\{X_i\in sS\}}}{n} \to E[\mathbf{M}(sS)] \text{ skoro jistě,}$$

proto $|A_n - B_n| \to 0$ skoro jistě právě tehdy, když $E\mathbf{M}(S) = E\mathbf{M}(sS)$. To je podle definice 2.4 a věty 2.5 ekvivalentní vlastnosti (iii).

„(iii) $\Leftrightarrow$ (v)“ Stejně jako výše, užitím lemmatu 2.11 vyjde, že vlastnost (iv) je ekvivalentní tomu, že $E\mathbf{M} = P_L$. To je ale ekvivalentní (iii).

„(iii) $\Leftrightarrow$ (vi)“ Lemma 2.11 dává

$$\frac{\sum_{i=1}^{n} I_{\{m(X_i)\in[1,t)\}}}{n} \to E[\mathbf{M}(t)] \text{ skoro jistě, } t \in [1, 10)$$

a z toho plyne požadovaná ekvivalence.

□

Věta 2.15 pomáhá vysvětlit, proč se některé soubory číselných údajů řídí Benfordovým zákonem lépe než jiné. Například na čísla vypsaná Frankem Benfordem z titulních stran místních novin (viz [2]) se dá dívat tak, že pocházejí z různých rozdělení, která spolu navzájem nesouvisí a tedy dohromady nezvýhodňují jednu volbu jednotek před ostatními. V kontextu uvedené věty to znamená, že rozdělení mantis v tomto výběru se bude asymptoticky blížit Benfordovu rozdělení.

Pokud naopak rozdělení, z něhož zkoumané údaje pochází, preferuje konkrétní volbu jednotek (jako je tomu v případě už zmíněných společností vyrábějících nápoje), *nebude* se rozdělení mantis blížit Benfordovu. Přínos věty 2.15 spočívá právě v tom, že poskytuje kritérium pro rozhodování, zda se zkoumaný soubor bude řídit Benfordovým zákonem či nebude.

# Kapitola 3

# Shoda s Benfordovým zákonem

## 3.1 Určování shody

Při analýze rozdělení mantisy v souboru číselných údajů se nabízí použití testů dobré shody (např. Pearsonova $\chi^2$ testu) k ověření, zda rozdělení výskytu prvních platných číslic odpovídá multinomickému rozdělení s parametry určenými Benfordovým zákonem. Použití tohoto testu bude podrobněji rozebráno a předvedeno v kapitole 4.

Určitou informaci o rozdělení mantisy lze ovšem získat i bez konkrétního souboru údajů. Například u náhodné veličiny se známým rozdělením nebo u náhodného výběru z nějakého pravděpodobnostního rozdělení lze předem (bez znalosti konkrétní realizace uvažované veličiny nebo výběru) buď přímo popsat rozdělení mantisy nebo určit, do jaké míry se toto rozdělení shoduje s Benfordovým.

Podobně u některých číselných posloupností definovaných určitou vlastností (ne výčtem prvků) je možné popsat asymptotické chování rozdělení mantisy bez znalosti konkrétních číselných hodnot prvků posloupnosti.

Předně je třeba zjistit, zda vůbec existují rozdělení, která se Benfordovým zákonem řídí *přesně.* Za příklad poslouží náhodná veličina s hustotou

$$f(u) = \frac{1}{u \ln 10}\ I_{(1,10)}(u),$$

uvedená v kapitole 1. Tu lze ještě mírně zobecnit do tvaru

$$g_{a,b}(u) = \frac{1}{u(b-a) \ln 10}\ I_{(10^a,10^b)}(u),$$

kde $a, b \in \mathbb{Z},\ a < b$. Nechť náhodná veličina $X$ má hustotu $g_{a,b}$. Přímý výpočet potom dává pro $t \in [1, 10]$ :

$$\mathbf{P}(m(X) < t) \ = \ \sum_{n=-\infty}^{\infty} \mathbf{P}(X \in [10^n, t \cdot 10^n))$$

$$
\begin{aligned}
&= \sum_{n=-\infty}^{\infty} \int_{10^n}^{t\cdot 10^n} \frac{1}{u(b-a)\ln 10}\; I_{(10^a,10^b)}(u)\; du \\
&= \sum_{n=a}^{b-1} \int_{10^n}^{t\cdot 10^n} \frac{1}{u(b-a)\ln 10}\; du \\
&= \frac{1}{(b-a)\ln 10} \sum_{n=a}^{b-1} [\ln u]_{u=10^n}^{t\cdot 10^n} \\
&= \frac{1}{(b-a)\ln 10} \sum_{n=a}^{b-1} (\ln t + n\ln 10 - n\ln 10) \\
&= \frac{(b-a)\ln t}{(b-a)\ln 10} = \log_{10} t.
\end{aligned}
$$

To znamená, že rozdělení mantisy náhodné veličiny $X$ je Benfordovo a $X$ se přesně řídí Benfordovým zákonem.

Další zobecnění vede k hustotě typu

$$
g(u) = \frac{1}{u\ln 10 \cdot \sum_{i=1}^{n}(b_i - a_i)}\; I_R(u),\;\; R = \bigcup_{i=1}^{n} (10^{a_i}, 10^{b_i}),
$$

přičemž $n \in \mathbb{N},\;\; a_i, b_i \in \mathbb{Z},\;\; i = 1, \ldots, n,\;\; a_1 < b_1 \le a_2 < \ldots \le a_n < b_n$.

Podobný výpočet jako výše ukáže, že i v tomto případě je rozdělení mantisy náhodné veličiny s takovou hustotou Benfordovo. Také libovolná konvexní kombinace hustot tohoto typu je opět hustota nějaké náhodné veličiny, která se přesně řídí Benfordovým zákonem.

Naproti tomu náhodná veličina $Y$ s rovnoměrným rozdělením na intervalu $[0, 1]$ má následující rozdělení mantisy:

$$
\begin{aligned}
\mathbf{P}(m(Y) < t) &= \sum_{n=-\infty}^{\infty} \mathbf{P}(Y \in [10^n, t\cdot 10^n)) = \sum_{n=-\infty}^{-1} (t\cdot 10^n - 10^n) \\
&= (t-1)\sum_{n=1}^{\infty} 10^{-n} = (t-1)\frac{\frac{1}{10}}{1-\frac{1}{10}} = \frac{t-1}{9},\;\; t \in [1, 10].
\end{aligned}
$$

Rozdělení veličiny $m(Y)$ je tedy rovnoměrné a ne Benfordovo.

Goudsmit a Furry [12] tvrdí, že fenomén nerovnoměrného výskytu prvních číslic „je pouhým důsledkem našeho způsobu zápisu čísel“ a tedy nezávisí na tom, z jakého rozdělení zkoumaná data pocházejí. To je v rozporu s pozorováním učiněným výše (že rovnoměrné rozdělení na $[0, 1]$ dává rovnoměrné rozdělení mantisy), přesto je zajímavé podívat se na jejich úvahy blíže.

Vychází z představy rozsáhlého souboru číselných údajů, kladných nebo braných bez ohledu na znaménko, a uvažují hustotu $f$, která popisuje jejich rozdělení. Pak platí

$$\int_0^\infty f(x)dx = 1.$$

Při zápisu čísel v desítkové soustavě má rozdělení mantisy uvažovaných čísel hustotu

$$h(p) = \sum_{m=-\infty}^{\infty} f(p \cdot 10^m) \cdot 10^m,\ p \in (1, 10), \tag{3.1}$$

což se dá ukázat pomocí zobecněné věty o transformaci hustoty (např. Anděl [3], věta 3.7). Následně se uvedená suma aproximuje pomocí integrálu:

$$h(p) = \sum_{m=-\infty}^{\infty} f(p \cdot 10^m) \cdot 10^m \approx \int_{-\infty}^{\infty} f(p \cdot 10^m) \cdot 10^m dm. \tag{3.2}$$

Tento integrál lze vypočítat pomocí substituce $x = p \cdot 10^m$ a vyjde

$$h(p) \approx \frac{\int_0^\infty f(x)dx}{p \cdot \ln 10} = \frac{1}{p \cdot \ln 10},\ p \in (1, 10).$$

To je ale hustota Benfordova rozdělení.

Goudsmit a Furry z toho vyvozují závěr, že nerovnoměrné rozdělení prvních platných číslic „nesouvisí s povahou číselných údajů ani s jejich rozdělením.“ Jsou si však vědomi toho, že „je nutné zodpovědět otázku, jak přesně integrál v (3.2) aproximuje uvažovanou sumu.“ Přesnost aproximace ovšem závisí na vlastnostech hustoty $f(x)$.

Tato problematika je podrobněji rozebrána v navazujícím článku Furry, Hurwitz [13]. Nechť pomocná funkce $\Psi$ je definována takto:

$$\Psi(q) = \sum_{m=-\infty}^{\infty} f(10^{m+q}) \cdot 10^{m+q} \cdot \ln 10,\ q \in \mathbb{R}. \tag{3.3}$$

Potom hustota rozdělení mantisy $h(p)$ uvedená v (3.1) se dá vyjádřit ve tvaru

$$h(p) = \frac{1}{p \cdot \ln 10} \cdot \Psi(\log_{10} p),\ p \in (1, 10).$$

Zřejmě platí $\Psi(q + 1) = \Psi(q)$, stačí tedy zkoumat $\Psi$ pouze na intervalu $[0, 1]$. Pokud by funkce $\Psi$ byla identicky rovna jedné, byla by hustota rozdělení mantisy $h(p)$ přesně hustotou Benfordova rozdělení.

Uvažovaná funkce $f$ je nezáporná (jde o hustotu nějakého rozdělení) a požadavek

$$1 = \sum_{m=-\infty}^{\infty} f(10^{m+q}) \cdot 10^{m+q} \cdot \ln 10, \;\; q \in [0,1]$$

na ni klade tyto podmínky:

- na množině, na níž je $f > 0$, je tvaru $f(x) = c \cdot \frac{1}{x}$, kde $c$ je reálná konstanta (aby se odstranil vliv faktoru $10^{m+q}$);
- $f$ je kladná pouze na intervalech typu $(10^a, 10^b)$, $a, b \in \mathbb{Z}$ (jinak by $\Psi(q)$ nebyla konstantní);
- množina, na níž je $f > 0$, je omezená (kvůli konvergenci sumy).

To vede k funkcím typu $g_{a,b}$, případně jejich zobecněním, jak byla uvedena na začátku této kapitoly. Pro libovolnou jinou hustotu je porušena podmínka $\Psi(q) = 1, \; q \in [0,1]$, a rozdělení mantisy nebude (přesně) Benfordovo.

V tomto smyslu jsou tedy rozdělení s hustotami typu $g_{a,b}$ jediná, jejichž mantisa má *přesně* Benfordovo rozdělení.

Míru shody s Benfordovým rozdělením Furry a Hurwitz v [13] měří pomocí hodnot $\max|\Psi - 1|$ a pro některá běžná rozdělení uvádějí tyto hodnoty, získané numerickým výpočtem podle vzorce (3.3). Jejich výsledky jsou shrnuty v následující tabulce.

| Hustota | $\max\|\Psi - 1\|$ |
|---|---|
| $\frac{2}{\sqrt{2\pi\alpha^2}}\; e^{-\frac{x^2}{2\alpha^2}}$ | 0.33 |
| $\frac{2\alpha}{\pi(\alpha^2+x^2)}$ | 0.0557 |
| $\frac{4\alpha}{\pi^2}\frac{\ln(\frac{x}{\alpha})}{x^2-\alpha^2}$ | 0.00152 |
| $\frac{1}{\alpha}\; e^{-\frac{x}{\alpha}}$ | 0.115 |
| $\frac{\alpha}{(\alpha+x)^2}$ | 0.0065 |

Je vidět, že rozdíly v míře shody s Benfordovým rozdělením mohou být velmi výrazné.

## 3.2 Směsi rozdělení

Shodu s Benfordovým zákonem jde zlepšit vytvořením *směsi rozdělení*, kdy se získá nové rozdělení „smícháním“ různých rozdělení v určitém poměru v následujícím smyslu (definice je převzata z [14]).

**Definice 3.1** *Nechť $f(x,\lambda),\ x \in (0,\infty)$ je pro každé $\lambda \in \Lambda \neq \emptyset$ hustota nějakého rozdělení (buď stejného typu s parametrem $\lambda$, např. $Exp(\lambda)$, případně různého typu podle hodnoty parametru $\lambda$), a nechť $g(\lambda)$ je nezáporná měřitelná funkce taková, že*

$$\int_\Lambda g(\lambda)d\lambda = 1,$$

*pak smíšenou hustotou je funkce $r(x)$ definovaná jako*

$$r(x) = \int_\Lambda g(\lambda)f(x,\lambda)d\lambda,\ x \in (0,\infty).$$

Interval $(0,\infty)$, na němž jsou hustoty $f(x,\lambda)$ a $r(x)$ definovány, je použit pouze proto, že se v této kapitole pracuje s rozděleními čísel braných bez ohledu na znaménko a bez horní meze. V úplně obecné definici by byl tento interval nahrazen obecnou množinou $M \subset \mathbb{R}$, na níž jsou hustoty definovány.

Furry a Hurwitz [13] dokazují, že pokud $f$ je hustota popisující uvažované rozdělení čísel (bez ohledu na znaménko), pak smíšená hustota

$$f^*(x) = \int_0^\infty \frac{1}{\lambda} f\left(\frac{x}{\lambda}\right) f(\lambda)d\lambda$$

se více blíží hustotě Benfordova rozdělení než původní hustota $f$.

Iteracemi tohoto procesu se vytvářejí hustoty $f^{*n}$, pro něž platí (viz [13]):

$$\lim_{n\to\infty} |\Psi_n(q) - 1| = 0 \text{ stejnoměrně v } q,$$

kde $\Psi_n$ je funkce odpovídající $f^{*n}$ podle vzorce (3.3). Tímto postupem lze tedy získávat hustoty, které se libovolně přesně blíží hustotě Benfordova rozdělení.

## 3.3 Silně a slabě benfordovské posloupnosti

Některé číselné posloupnosti mají vlastnost, že se rozdělení mantisy jejich prvků blíží Benfordovu rozdělení, pokud se bere v úvahu čím dál tím delší úsek těchto posloupností. V takovém případě se dají označit jako *benfordovské*. Tuto problematiku přehledně shrnuje např. Raimi [5].

**Definice 3.2** *Posloupnost $\{b_n\}$ reálných čísel z intervalu $[0,1)$ se nazývá rovnoměrně distribuovaná na $[0,1)$, pokud pro každý interval $[a,b) \subset [0,1)$ platí*

$$\lim_{k\to\infty} \frac{1}{k} \sum_{n=1}^{k} \beta(n) = b - a,$$

*kde $\beta(n) = 1$, pokud $b_n \in [a,b)$, a $\beta(n) = 0$ jinak.*

Stejně jako v části 2.4 bude dále symbolem $D_p$ $(1 \leq p < 10)$ označena množina kladných reálných čísel, která mají hodnotu mantisy menší než $p$, tedy

$$D_p = \{x \in \mathbb{R}^+, m(x) < p\} = \bigcup_{n=-\infty}^{\infty} [10^n, p \cdot 10^n).$$

**Definice 3.3** *Nechť $\{a_n\}$ je posloupnost kladných reálných čísel a nechť $b_n = (\log_{10} a_n) \bmod 1$. Pokud je $\{b_n\}$ rovnoměrně distribuovaná na $[0, 1)$, nazývá se $\{a_n\}$ silně benfordovská posloupnost.*

Silně benfordovská posloupnost $\{a_n\}$ se řídí Benfordovým zákonem v tom smyslu, že

$$\lim_{k\to\infty} \frac{1}{k} \sum_{n=1}^{k} \alpha_p(n) = \log_{10} p, \ p \in [1, 10),$$

kde $\alpha_p(n) = 1$, pokud $a_n \in D_p$, a $\alpha_p(n) = 0$ jinak.

Například posloupnost přirozených čísel $\mathbb{N}$ není silně benfordovská posloupnost, protože uvedená limita neexistuje (viz kapitola 1), ale geometrické posloupnosti $\{ar^n\}$ jsou, pokud $r$ není racionální mocninou čísla 10. Pokud $a_n = ar^n$, je

$$b_n = (\log_{10} a_n) \bmod 1 = (\log_{10} a + n \log_{10} r) \bmod 1.$$

Přitom platí, že aritmetické posloupnosti s iracionální diferencí jsou (bráno mod 1) rovnoměrně distribuované na intervalu $[0, 1)$ - Raimi [5] uvádí odkaz na Hardy, Wright [15]. Právě požadavek, aby diference v posloupnosti $\{b_n\}$ byla iracionální, tedy $\log_{10} r \notin \mathbb{Q}$, vede k tomu, že posloupnost $\{ar^n\}$ je silně benfordovská, právě když $r$ není racionální mocninou 10.

Geometrické posloupnosti ale nejsou jedinými silně benfordovskými posloupnostmi. Tuto vlastnost mají i tzv. asymptoticky geometrické posloupnosti.

**Definice 3.4** *Posloupnost $\{a_n\}$ se nazývá asymptoticky geometrická, pokud existuje geometrická posloupnost $\{ar^n\}$ taková, že*

$$\lim_{n\to\infty} \frac{a_n}{ar^n} = 1.$$

V tomto případě $\log_{10} a_n - n \log_{10} r$ konverguje a $\log_{10} a_n \bmod 1$ je stejně rovnoměrně distribuovaná jako $n \log_{10} r$. Jinými slovy, pokud $\log_{10} r \notin \mathbb{Q}$, je asymptoticky geometrická posloupnost silně benfordovská.

Jako příklad asymptoticky geometrické posloupnosti Raimi [5] uvádí Fibonacciho posloupnost. V jejím případě je limitním poměrem $r$ z definice

3.4 zlatý řez, $r = \frac{1+\sqrt{5}}{2}$. Navíc $\log_{10} \frac{1+\sqrt{5}}{2}$ je iracionální číslo, a proto je Fibonacciho posloupnost silně benfordovská.

Dalším příkladem silně benfordovské posloupnosti je $\{n!\}$, jak ukazuje Diaconis [16].

Posloupnosti, které nejsou rovnoměrně distribuované na $[0, 1)$ proto, že limita v definici 3.2 neexistuje, stále mohou být (mod 1) v jistém smyslu rovnoměrně rozprostřené. Jiná sčítací metoda, než která byla použita v definici 3.2, vede k mírně upravené definici rovnoměrné distribuovanosti (viz Raimi [5]).

**Definice 3.5** *Posloupnost $\{b_n\}$ reálných čísel z intervalu $[0, 1)$ se nazývá L-rovnoměrně distribuovaná na $[0, 1)$, pokud pro každý interval $[a, b) \subset [0, 1)$ platí*

$$\lim_{k\to\infty} \frac{1}{\ln k} \sum_{n=1}^{k} \frac{\beta(n)}{n} = b - a,$$

*kde $\beta(n) = 1$, pokud $b_n \in [a, b)$, a $\beta(n) = 0$ jinak.*

**Definice 3.6** *Nechť $\{a_n\}$ je posloupnost kladných reálných čísel a nechť $b_n = (\log_{10} a_n) \bmod 1$. Pokud je $\{b_n\}$ L-rovnoměrně distribuovaná na $[0, 1)$, nazývá se $\{a_n\}$ slabě benfordovská posloupnost.*

Raimi [5] uvádí (spolu s odkazy na patřičnou literaturu), že posloupnost $\{a_n\}$ je slabě benfordovská, pokud $a_n = n$, $a_n = \sqrt{n}$, $a_n = n$-té prvočíslo nebo $a_n = Q(n)$ pro libovolný polynom $Q$.

To, že číselná posloupnost je slabě benfordovská, ještě neznamená, že rozdělení mantisy jejích prvků se asymptoticky blíží Benfordovu. Znamená to, že podíl prvků, jejichž mantisa je v intervalu $[1, t)$, $t \in [1, 10]$, se při použití určité sumační (vlastně průměrovací) metody blíží hodnotě $\log_{10} t$. Stejná situace se objevila už v kapitole 1 při zkoumání rozdělení mantisy v posloupnosti přirozených čísel (viz Flehinger [4]).

Rozdíl mezi silně a slabě benfordovskými posloupnostmi je tedy tento: v dostatečně dlouhém úseku silně benfordovské posloupnosti se bude vyskytovat přibližně 30% čísel začínajících číslicí 1, 18% začínajících číslicí 2 atd., v případě slabě benfordovské posloupnosti to očekávat nejde. Přesto je rozdělení mantisy ve slabě benfordovské posloupnosti při použití průměrovací techniky odpovídající metodě z definice 3.5 podobné Benfordovu rozdělení.

# Kapitola 4

# Empirická pozorování

## 4.1 Zpracování údajů

Benfordův zákon u některých souborů číselných údajů určuje konkrétní rozdělení mantisy a tím i rozdělení prvních platných číslic. Při rozhodování, zda se konkrétní (napozorovaná) data řídí Benfordovým zákonem, se velmi dobře uplatní Pearsonův $\chi^2$ test dobré shody.

Tento test se obecně používá k testování hypotézy, že vektor četností výskytu určitých jevů má multinomické rozdělení s předpokládanými parametry. Má-li náhodný vektor $\mathbf{X} = (X_1, \ldots, X_k)'$ multinomické rozdělení $M(n; p_1, \ldots, p_k)$, pak Pearsonova statistika

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \tag{4.1}$$

má při $n \to \infty$ asymptoticky rozdělení $\chi^2_{k-1}$ (viz např. Anděl [3], věta 12.5). Tento výraz jde ještě upravit do podoby

$$\chi^2 = \sum_{i=1}^{k} \frac{X_i^2}{np_i} - n,$$

která je vhodnější při praktických výpočtech. Protože má tato veličina rozdělení $\chi^2_{k-1}$ pouze asymptoticky, je nutné mít dostatečně velký rozsah zkoumaného souboru. Pro použití testů založených na veličině $\chi^2$ se obvykle požaduje, aby $np_i \geq 5$ pro každé $i = 1, \ldots, k$.

V případě zkoumání, zda se napozorovaná data řídí Benfordovým zákonem, bude vektor $\mathbf{X} = (X_1, \ldots, X_9)'$ představovat četnosti výskytu číslic $1, 2, \ldots, 9$ na prvním platném místě a $n$ celkový počet zpracovávaných údajů. Takto definovaný vektor $\mathbf{X}$ má multinomické rozdělení $M(n; q_1, \ldots, q_9)$ s neznámými parametry $q_i$, které představují pravděpodobnost výskytu číslice $i$ na prvním platném místě.

Nechť $p_1, \ldots, p_9$ jsou pravděpodobnosti předpovídané Benfordovým zákonem (podle vzorce (1.2)), tedy

$$p_i = \log_{10}\left(1 + \frac{1}{i}\right), \ i = 1, \ldots, 9.$$

Testovaná hypotéza potom bude taková, že tyto pravděpodobnosti odpovídají pravděpodobnostem $q_i$, neboli

$$H_0 : q_i = p_i, \ i = 1, \ldots, 9.$$

Za platnosti hypotézy $H_0$ má veličina $\chi^2$ definovaná vztahem (4.1) asymptoticky rozdělení $\chi^2_8$. Na hladině $\alpha = 0.05$ bude hypotéza $H_0$ zamítnuta, pokud hodnota $\chi^2$ přesáhne 95%-kvantil $\chi^2$-rozdělení o osmi stupních volnosti.

Výše popsaný test je označován jako Pearsonův $\chi^2$ test dobré shody a v této kapitole bude používán k testování hypotézy, že rozdělení prvních platných číslic zkoumaných údajů odpovídá rozdělení předpovězenému Benfordovým zákonem. Zamítnutí této hypotézy pak umožní vyvodit závěr, že rozdělení mantis *není* Benfordovo, při nezamítnutí hypotézy ale opačný závěr učinit nelze.

## 4.2 Benfordova data

V článku [2] předkládá Benford výsledky své několikaleté práce, během níž sbíral údaje „z tolika různých oblastí, kolik čas a energie dovolily.“ Dohromady zpracoval více než 20 000 údajů z 20 různých oblastí, což lze ve třicátých letech 20. století považovat za úctyhodný výkon (všechny výpočty se totiž prováděly ručně nebo na mechanických kalkulátorech).

Benford hledal rozdělení prvních platných číslic v „přirozeně se vyskytujících“ číselných údajích a uvádí pouze relativní četnosti výskytu prvních číslic ve všech 20 zkoumaných kategoriích spolu s údajem, kolik údajů bylo do jednotlivých kategorií zahrnuto. Pro každou číslici tak Benford získal 20 různých relativních četností (pro každou kategorii jednu hodnotu) a z nich vypočítal aritmetický průměr.

Na těchto průměrných hodnotách pak Benfordovi bylo nápadné, že se blíží hodnotám $p_i = \log_{10}\left(1 + \frac{1}{i}\right), \ i = 1, \ldots, 9$, a dále se ve svém článku pokusil zdůvodnit, proč právě tyto hodnoty jsou „ty správné.“

Kdyby Benford nepracoval s průměrnými relativními četnostmi, ale místo toho by zpracoval všechny údaje najednou v jednom velkém souboru, získal by odlišné relativní četnosti. Ty se sice více liší od hodnot $p_i$, přesto jsou jim stále dost blízké.

Srovnání nabízí tabulka 4.1. Ve sloupci „Průměrná relativní četnost“ jsou rel. četnosti vypočtené jako aritmetický průměr z rel. četností v jednotlivých skupinách dat, ve sloupci „Skutečná relativní četnost“ je hodnota odpovídající podílu čísel začínajících danou číslicí v celém souboru. Hodnoty jsou zaokrouhlené na tři desetinná místa.

| **Číslice** | **Průměrná r.č.** | **Skutečná r.č.** | $p_i = \log_{10}\left(1+\frac{1}{i}\right)$ |
|---|---|---|---|
| 1 | 0.306 | 0.289 | 0.301 |
| 2 | 0.185 | 0.195 | 0.176 |
| 3 | 0.124 | 0.127 | 0.125 |
| 4 | 0.094 | 0.091 | 0.097 |
| 5 | 0.080 | 0.075 | 0.079 |
| 6 | 0.064 | 0.064 | 0.067 |
| 7 | 0.051 | 0.054 | 0.058 |
| 8 | 0.049 | 0.055 | 0.051 |
| 9 | 0.047 | 0.051 | 0.046 |

Tabulka 4.1: Relativní četnosti v Benfordových datech.

Zvláštní je, že v jednotlivých skupinách v Benfordově tabulce v článku [2] se relativní četnosti nasčítají přesně na 1. Díky zaokrouhlování relativních četností na tři desetinná místa by mohlo dojít k tomu, že se v některé skupině nenasčítají přesně na 1, ale v případě Benfordem předložených dat dává součet 1 všech dvacet skupin.

Upozornili na to Diaconis a Freedman v článku [17], v němž se zabývají právě otázkou chování součtu zaokrouhlených procentuálních údajů. Při použití jejich modelů vyjde pravděpodobnost, že se relativní četnosti v každé z dvaceti skupin nasčítají přesně na 1, „astronomicky malá“ (zhruba $1 : 2^{20}$).

Vyvozují z toho, že se Benfordovy četnosti neřídí žádným z jejich modelů, a to vede k podezření, že Benford s částí dat manipuloval, aby docílil lepší shody s hodnotami $\log_{10}\left(1+\frac{1}{i}\right)$. Například v prvním řádku Benfordovy tabulky (viz [2]) je podíl čísel začínajících na 7 uveden jako 5.5% z celkového počtu 335 údajů. Snadný výpočet ukazuje, že počet čísel začínajících na 7 musel být 18 nebo 19, ale $\frac{18}{335}$ se zaokrouhlí na 5.4% a $\frac{19}{335}$ na 5.7% procenta. Žádná skutečná hodnota tedy nemohla vést k údaji 5.5%.

Ony chybně zaokrouhlené hodnoty jsou skutečně blíž požadovaným číslům než hodnoty zaokrouhlené správně, ale i neupravená data se poměrně dobře shodují s předpokládanými četnostmi.

Pearsonův $\chi^2$ test provedený na jednotlivé dílčí soubory vede na hladině 5% v osmi případech k zamítnutí hypotézy $H_0$, že se rozdělení prvních platných číslic v těchto souborech řídí Benfordovým zákonem, ve dvanácti případech test tuto hypotézu nezamítl.

Všech osm skupin, u nichž došlo k zamítnutí, se nachází mezi posledními devíti v seznamu skupin, seřazených podle součtu absolutních odchylek od očekávaných relativních četností, předloženém v [2]. Výsledky $\chi^2$ testu jsou tedy konzistentní s hodnocením shody u jednotlivých skupin, jak je podává Benford.

$\chi^2$ test provedený na celý soubor, s nímž Benford pracoval, vede k zamítnutí hypotézy $H_0$ (při tak velkém rozsahu souboru je síla testu obrovská, interval spolehlivosti se při pevné hladině významnosti zužuje s rostoucím počtem pozorování).

## 4.3 Další publikované výsledky

Od vydání Benfordova článku [2] se další autoři snažili ukázat, které soubory číselných údajů se řídí Benfordovým zákonem.

Například rozsáhlé soubory čísel obsažených v účetních záznamech nebo daňových přiznáních firem a jednotlivců se obvykle Benfordovým zákonem řídí velmi přesně (Nigrini [18]). To odpovídá výsledkům uvedeným v kapitole 2, protože účetní záznamy lze považovat za směs údajů pocházejících z různých rozdělení.

Ley [19] zjistil, že posloupnost jednodenních výnosů na indexu Dow-Jones Industrial Average (DJIA) a Standard and Poor's (S&P) je v dobré shodě s Benfordovým zákonem. Jde o řádově desetitisíce údajů z let 1900–1993 (DJIA), resp. 1926–1993 (S&P) a relativní četnosti prvních platných číslic jsou uvedeny v tabulce 4.2.

| **Číslice** | **DJIA** | **S&P** | **Benfordův zákon** |
|---|---|---|---|
| 1 | 0.2894 | 0.2917 | 0.3010 |
| 2 | 0.1678 | 0.1696 | 0.1761 |
| 3 | 0.1238 | 0.1342 | 0.1249 |
| 4 | 0.0999 | 0.0987 | 0.0969 |
| 5 | 0.0848 | 0.0776 | 0.0792 |
| 6 | 0.0723 | 0.0713 | 0.0669 |
| 7 | 0.0615 | 0.0560 | 0.0580 |
| 8 | 0.0532 | 0.0536 | 0.0512 |
| 9 | 0.0472 | 0.0473 | 0.0458 |

Tabulka 4.2: DJIA a S&P (Ley [19]).

Po provedení $\chi^2$ testu Ley zamítá hypotézu, že se rozdělení prvních platných číslic řídí Benfordovým zákonem, ale poukazuje na to, že zamítnutí je způsobeno velkou silou testu danou velkým počtem pozorování. Kdyby se

vzala v úvahu pouze data z posledních deseti let (1983–1993), uvádí, k zamítnutí nulové hypotézy by nedošlo.

Varian [20] popisuje, jak se koncem 60. let zabýval vývojem počítačových modelů předpovídajících ekonomický vývoj v San Franciské zátoce do roku 1990. Výstupy srovnával s Benfordovým zákonem na základě úvahy, že když se vstupní data modelu shodují s Benfordovým zákonem, měly by se jím řídit i výstupy, pokud je model „rozumný."

Po rozboru výstupů svého modelu dospěl Varian k tomu, že „data jsou v poměrně dobré shodě s Benfordovým zákonem." Konkrétně nejlepší shodu vykazovala ta část modelu, jejíž vstupní data pocházela ze sčítání lidu - to je podle Variana obecně považováno za přesnější než jiné zdroje údajů.

Benfordovým zákonem se také poměrně přesně řídí i údaje o odběru elektřiny jednotlivými spotřebiteli na Šalamounových ostrovech, jak uvádí Raimi [5]. Jeho data pochází z října 1969 a zajímavý je jejich původ.

Údaje nezískával autor sám, ale už zpracované mu byly zaslány ředitelem energetické společnosti ve městě Honiara na Šalamounových ostrovech, který byl zaujat problematikou Benfordova zákona poté, co si přečetl populárně laděný Raimiho článek [21] v časopisu Scientific American.

Příklady podobných výsledků lze nalézt nejen v ekonomii, ale třeba i ve fyzice. Burke a Kincanon [22] se rozhodli prozkoumat, zda se fyzikální konstanty řídí Benfordovým zákonem.

Vzali konstanty uvedené na vnitřním přebalu úvodní vysokoškolské učebnice fyziky a spočítali četnosti výskytu prvních platných číslic při vyjádření čísel v jednotkách SI a v britských jednotkách. Výsledky jsou v tabulce 4.3.

| **Číslice** | **Jednotky SI** | | **Britské jednotky** | |
|---|---|---|---|---|
| 1 | 8 | (40 %) | 7 | (35 %) |
| 2 | 2 | (10 %) | 2 | (10 %) |
| 3 | 1 | (5 %) | 3 | (15 %) |
| 4 | 0 | (0 %) | 1 | (5 %) |
| 5 | 2 | (10 %) | 1 | (5 %) |
| 6 | 3 | (15 %) | 3 | (15 %) |
| 7 | 0 | (0 %) | 0 | (0 %) |
| 8 | 2 | (10 %) | 1 | (5 %) |
| 9 | 2 | (10 %) | 2 | (10 %) |

Tabulka 4.3: Fyzikální konstanty (Burke a Kincanon [22]).

Pozoruhodné je, že při tak malém rozsahu souboru (pouze 20 údajů) je vidět vůbec nějaká shoda s Benfordovým zákonem. Při vyjádření čísel v soustavě SI a v britských jednotkách má jednička zdaleka nejvyšší četnost

výskytu na prvním platném místě. Ostatní hodnoty už není možné odlišit od statistického šumu.

Buck, Merchant a Perez [23] vyvinuli metodu, jak počítat poločas rozpadu nestabilních nuklidů podléhajících alfa rozpadu a při porovnávání vypočtených hodnot s experimentálně zjištěnými údaji si všimli, že rozdělení prvních platných číslic je nerovnoměrné.

Je známo 477 nuklidů podléhajících alfa rozpadu a jejich poločasy rozpadu pokrývají interval od zhruba $10^{-6}$ sekundy až po $10^{15}$ let. Tato data, nashromážděná v průběhu celého 20. století, poskytují výbornou příležitost zkoumat, za jakých situací se Benfordův zákon uplatňuje.

Při podrobnějším rozboru Buck a kol. zjistili, že četnosti výskytu jednotlivých číslic na prvním platném místě vypočtených i naměřených hodnot se poměrně dobře shoduje s četnostmi předpovězenými Benfordovým zákonem.

U vypočtených hodnot zkoumali také rozdělení druhých platných číslic a i zde data vykázala vysokou míru shody s Benfordovým zákonem (naměřené hodnoty jsou často zaznamenány pouze s přesností na jednu platnou cifru a výsledky získané rozborem druhých platných číslic by nebyly směrodatné).

Na druhou stranu ne v každé sadě přirozeně se vyskytujících údajů se rozdělení prvních platných číslic řídí Benfordovým zákonem. Raimi [5] uvádí pěkný příklad: telefonní seznam. U čísel na stránce vytržené z místního telefonního seznamu bude na prvním platném místě výrazně převažovat jedna číslice, ostatní budou zastoupeny minimálně.

## 4.4 Původní výsledky

V této části budou předložena autorem vybraná a zpracovaná data z různých oblastí, která mohou ilustrovat aspekty Benfordova zákona, jež byly zmíněny v předchozích kapitolách.

Údaje jsou rozděleny do devíti skupin podle toho, z jaké oblasti pocházejí, a skupiny jsou označeny názvem, který naznačuje povahu dat. Například data ve skupině „Populace“ tvoří počty obyvatel v obcích České republiky, skupina „Státy“ obsahuje údaje o rozlohách jednotlivých států světa atd.

Skupiny budou rozebrány jednotlivě, bude upřesněna povaha a původ dat a následovat bude krátký komentář a výsledek $\chi^2$ testu hypotézy $H_0$, že rozdělení prvních platných číslic se řídí Benfordovým zákonem (na hladině $\alpha = 5\%$). Hypotéza $H_0$ tedy bude zamítnuta, pokud je dosažená p-hodnota menší než 0.05. Grafy znázorňující odchylky od předpokládaného rozdělení jsou k dispozici v příloze A.

Přehled výsledků podává tabulka 4.4 (zjištěné četnosti prvních platných číslic v jednotlivých skupinách) a tabulka 4.5 (relativní četnosti a p-hodnoty dosažené při provedení $\chi^2$ testu dobré shody, pro srovnání jsou uvedeny i relativní četnosti předpovězené Benfordovým zákonem).

| **Skupina** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | $\sum$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Populace | 1820 | 1134 | 794 | 595 | 539 | 420 | 356 | 332 | 259 | 6249 |
| Státy | 50 | 38 | 22 | 22 | 13 | 13 | 12 | 7 | 12 | 189 |
| Nehody | 296 | 171 | 129 | 123 | 99 | 73 | 55 | 50 | 39 | 1035 |
| Soubory | 5814 | 3427 | 2583 | 2099 | 1600 | 1611 | 1509 | 1253 | 853 | 20749 |
| Web | 238 | 119 | 75 | 43 | 48 | 28 | 28 | 14 | 12 | 605 |
| Léky | 343 | 190 | 113 | 82 | 101 | 90 | 71 | 77 | 79 | 1146 |
| $2^n(100)$ | 30 | 17 | 13 | 10 | 7 | 7 | 6 | 5 | 5 | 100 |
| $2^n(1000)$ | 301 | 176 | 125 | 97 | 79 | 69 | 56 | 52 | 45 | 1000 |
| Fibonacci | 301 | 177 | 125 | 96 | 80 | 67 | 56 | 53 | 45 | 1000 |

Tabulka 4.4: Četnosti výskytu prvních platných číslic.

| **Skupina** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **p-hod.** |
|---|---|---|---|---|---|---|---|---|---|---|
| Benford | 30.1 | 17.6 | 12.5 | 9.7 | 7.9 | 6.7 | 5.8 | 5.1 | 4.6 | — |
| Populace | 29.1 | 18.1 | 12.7 | 9.5 | 8.6 | 6.7 | 5.7 | 5.3 | 4.1 | 0.2313 |
| Státy | 26.5 | 20.1 | 11.6 | 11.6 | 6.9 | 6.9 | 6.3 | 3.7 | 6.3 | 0.7832 |
| Nehody | 28.6 | 16.5 | 12.5 | 11.9 | 9.6 | 7.1 | 5.3 | 4.8 | 3.8 | 0.1337 |
| Soubory | 28.0 | 16.5 | 12.4 | 10.1 | 7.7 | 7.8 | 7.3 | 6.0 | 4.1 | <2.2e-16 |
| Web | 39.3 | 19.7 | 12.4 | 7.1 | 7.9 | 4.6 | 4.6 | 2.3 | 2.0 | 2.15e-07 |
| Léky | 29.9 | 16.6 | 9.9 | 7.2 | 8.8 | 7.9 | 6.2 | 6.7 | 6.9 | 8.73e-06 |
| $2^n(100)$ | 30.0 | 17.0 | 13.0 | 10.0 | 7.0 | 7.0 | 6.0 | 5.0 | 5.0 | 1 |
| $2^n(1000)$ | 30.1 | 17.6 | 12.5 | 9.7 | 7.9 | 6.9 | 5.6 | 5.2 | 4.5 | 1 |
| Fibonacci | 30.1 | 17.7 | 12.5 | 9.6 | 8.0 | 6.7 | 5.6 | 5.3 | 4.5 | 1 |

Tabulka 4.5: Relativní četnosti (v procentech) a dosažené p-hodnoty.

## Populace

**Zdroj dat:** Český statistický úřad ([24]).
**Výsledek testu:** Dosažená p-hodnota 0.2313 znamená, že není zamítnuta nulová hypotéza, že rozdělení prvních platných číslic je multinomické s parametry odpovídajícími Benfordovu zákonu.
**Komentář:** Jde o příklad uvedený v úvodu. Data tvoří počty obyvatel v 6249 obcích České republiky k 1. lednu 2007. U takto velkého souboru údajů jde o pozoruhodně dobrou shodu s předpokládaným rozdělením.

## Státy

**Zdroj dat:** Bateman, Egan: *Encyklopedie Zeměpis světa* ([25]).
**Výsledek testu:** Dosažená p-hodnota 0.7832 znamená, že nulová hypotéza není zamítnuta.
**Komentář:** Data tvoří rozlohy 189 nezávislých států uvedených v [25]. Výsledky uvedené v tabulkách 4.4 a 4.5 odpovídají rozlohám vyjádřeným v kilometrech čtverečních.

Tato data se hodí ke zkoumání invariance vzhledem k měřítku, je totiž možné použít různé jednotky plochy. Při vyjádření například ve čtverečních mílích (1 čtvereční míle je přibližně 2.59 $km^2$) vyjdou tyto hodnoty:

| **Skupina** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **p-hod.** |
|---|---|---|---|---|---|---|---|---|---|---|
| abs. čet. | 65 | 27 | 28 | 22 | 6 | 11 | 11 | 7 | 12 | — |
| r. čet. | 34.4 | 14.3 | 14.8 | 11.6 | 3.2 | 5.8 | 5.8 | 3.7 | 6.3 | 0.1740 |

Tabulka 4.6: Rozlohy států vyjádřené ve čtverečních mílích.

Po přepočítání hodnot na čtvereční míle zůstalo rozdělení prvních platných číslic nerovnoměrné a $\chi^2$ test nezamítá nulovou hypotézu. Tento soubor údajů tedy vykazuje do určité míry vlastnost invariance vzhledem k měřítku, při vyjádření hodnot v jiných jednotkách by rozdělení prvních platných číslic zůstalo podobné tomu, jaké předpovídá Benfordův zákon.

## Nehody

**Zdroj dat:** Ministerstvo vnitra České republiky ([26]).
**Výsledek testu:** Dosažená p-hodnota 0.1337 znamená, že nulová hypotéza není zamítnuta.
**Komentář:** V tomto případě jsou zkoumány odhady hmotných škod vzniklých v průběhu roku 2007 na jednotlivých kilometrech dálnic v České republice. Vyřazeny byly případy, kdy odhadnutá škoda byla nulová (na daném kilometru k žádné nehodě nedošlo), protože používané metody pracují pouze s kladnými čísly.

Také tento soubor je vhodný ke zkoumání invariance vzhledem k měřítku, má totiž smysl vyjadřovat odhady škod v jiných měnách, než je česká koruna (údaje v tabulkách 4.4 a 4.5 odpovídají právě vyjádření škod v Kč).

Při převodu zkoumaných odhadů na eura, resp. americké dolary podle kurzu České národní banky ze dne 11. 5. 2008 (1 EUR za 25.145 Kč, resp. 1 USD za 16.267 Kč) vyjdou jiné hodnoty (viz tabulka 4.7).

Z tabulky je vidět, že se jednotlivé četnosti změnily (a při vyjádření v amerických dolarech je dokonce na hladině 5% zamítána nulová hypotéza,

| Skupina | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | p-hod. |
|---|---|---|---|---|---|---|---|---|---|---|
| EUR (abs. č.) | 341 | 187 | 123 | 93 | 90 | 65 | 46 | 48 | 42 | — |
| EUR (r. č.) | 32.9 | 18.1 | 11.9 | 9.0 | 8.7 | 6.3 | 4.4 | 4.6 | 4.1 | 0.3304 |
| USD (abs. č.) | 276 | 199 | 161 | 92 | 70 | 75 | 57 | 53 | 52 | — |
| USD (r. č.) | 26.7 | 19.2 | 15.6 | 8.9 | 6.8 | 7.2 | 5.5 | 5.1 | 5.0 | 0.0315 |

Tabulka 4.7: Odhady škod vyjádřené v EUR a USD.

že první platné číslice mají předpokládané rozdělení). Přesto však zůstává rozdělení prvních platných číslic nerovnoměrné a podobné tomu, jaké předpovídá Benfordův zákon (viz grafy v příloze A). Vyjádření hodnot v ještě dalších jednotkách by vedlo k podobným výsledkům. Lze tedy říct, že tento soubor dat se vyznačuje invariancí vzhledem k měřítku.

## Soubory

**Zdroj dat:** Velikosti všech souborů na pevném disku autorova počítače zjištěné postupem uvedeným na stránce *www.nigrini.com* ([27]).
**Výsledek testu:** Dosažená p-hodnota je menší než $2.2 \cdot 10^{-16}$ a tedy nulová hypotéza je zamítnuta (test má velkou sílu díky vysokému počtu pozorování).
**Komentář:** Na uvedené stránce je zveřejněn návod, podle kterého si každý zájemce může sám provést analýzu velikostí souborů na svém pevném disku.

Přestože je nulová hypotéza zamítnuta, graf v příloze A ukazuje, že Benfordův zákon poskytuje dobrý model pro popis rozdělení prvních platných číslic ve zkoumaných datech.

Tato data lze využít k ilustraci invariance vzhledem ke změně základu číselné soustavy. Při vyjádření čísel v soustavě o základu $b$, $b = 2, 3, 4, \ldots$ přejde Benfordův zákon do odpovídající podoby

$$\mathbf{P}\left(m^{(b)}(X) < t\right) = \log_b t, \ t \in [1, b],$$

kde $m^{(b)}$ značí mantisu čísla vyjádřeného v soustavě o základu $b$. Relativní četnosti výskytu prvních platných číslic předpovídané Benfordovým zákonem jsou tedy

$$p_i = \log_b \left(1 + \frac{1}{i}\right), \ i = 1, \ldots, b-1.$$

Ve speciálním případě $b = 10$ se tyto výrazy shodují s těmi, které byly používány v předchozích kapitolách.

Degenerovaným případem je situace, kdy jsou čísla vyjádřena v dvojkové soustavě – pak je totiž první platnou číslicí libovolného kladného čísla

jednička. To ale souhlasí s tím, že

$$p_1 = \log_2\left(1 + \frac{1}{1}\right) = 1,$$

proto není třeba tento případ zkoumat zvlášť.

Zkoumané velikosti souborů lze převést například do osmičkové nebo šestnáctkové soustavy a poté zkoumat rozdělení první platné číslice (s odpovídajícími změnami při provádění $\chi^2$ testu).

Srovnání výsledků tohoto rozboru s předpokládaným rozdělením se lépe ukáže graficky než výčtem konkrétních četností (viz příloha A, grafy jsou označeny číslem odpovídajícím tomu, v jaké soustavě jsou zkoumaná čísla vyjádřena, např. graf „Soubory [8]“ se týká vyjádření v osmičkové soustavě).

Při pohledu na grafy v příloze A lze říci, že míra shody s teoretickým rozdělením zůstala zachována a soubor dat má tedy vlastnost invariance vzhledem ke změně základu číselné soustavy. Benfordův zákon tedy poskytuje dobrý model pro popis rozdělení prvních platných číslic u zkoumaných dat.

Provedení $\chi^2$ testu při vyjádření čísel v osmičkové i šestnáctkové soustavě přesto vede k zamítnutí nulové hypotézy (stejně jako v desítkové soustavě), že rozdělení prvních platných číslic je multinomické s parametry odpovídajícími Benfordovu zákonu upravenému do podoby pro uvažovanou číselnou soustavu.

## Web

**Zdroj dat:** *www.pocitadlo.cz* ([28]).
**Výsledek testu:** Dosažená p-hodnota $2.15 \cdot 10^{-7}$ znamená, že nulová hypotéza je zamítnuta.
**Komentář:** Server Pocitadlo.cz vede statistiky návštěvnosti stránek svých registrovaných uživatelů, data tvoří počty přístupů na jednotlivé stránky v kategorii „Cestování“ za den 4. 4. 2008.

Počet přístupů na každou stránku lze chápat jako náhodnou veličinu s Poissonovým rozdělením s neznámým parametrem $\lambda$, který je určen atraktivitou a dostupností stránky, a počty přístupů na různé stránky je možné považovat za nezávislé. Situace tedy odpovídá části 2.3 – „Náhodné výběry z náhodně vybíraných rozdělení.“

Přestože je zamítnuta nulová hypotéza, je vidět, že rozdělení prvních platných číslic je velmi nerovnoměrné a nižší číslice jsou výrazně preferovány. Toto rozdělení je tedy podobné tomu, jaké předpovídá Benfordův zákon.

## Léky

**Zdroj dat:** Všeobecná zdravotní pojišťovna České republiky ([29]).
**Výsledek testu:** Dosažená p-hodnota $8.73 \cdot 10^{-6}$ znamená, že nulová hypotéza je zamítnuta.
**Komentář:** Data tvoří orientační ceny volně prodejných léčivých přípravků uvedených v číselníku VZP ČR.

Stejně jako v minulém případě je rozdělení prvních platných číslic nerovnoměrné a nižší číslice se vyskytují výrazně častěji než ty vyšší. Přestože je nulová hypotéza zamítnuta, je rozdělení prvních platných číslic podobné tomu, jaké předpovídá Benfordův zákon.

## $2^n$(100), $2^n$(1000), Fibonacci

**Zdroj dat:** Data tvoří hodnoty posloupností $a_n = \{2^n\}_{n=1}^{100}$, $b_n = \{2^n\}_{n=1}^{1000}$ a $c_n = \{n - \textit{té Fibonacciho číslo}\}_{n=1}^{1000}$.
**Výsledek testu:** Ve všech třech případech je dosažená p-hodnota velmi blízká 1, proto nulová hypotéza není zamítnuta.
**Komentář:** Tyto soubory jsou tvořeny hodnotami geometrické posloupnosti $2^n$, resp. asymptoticky geometrické Fibonacciho posloupnosti (viz oddíl 3.3).

Skupiny $2^n(100)$ a $2^n(1000)$ se liší pouze počtem prvků, které byly analyzovány. V prvním případě to bylo prvních 100 prvků, v druhém případě prvních 1000.

Už prvních 100 prvků posloupnosti vykazuje pozoruhodně dobrou shodu s Benfordovým zákonem, při rozboru prvních platných číslic tisíce prvků je shoda ještě lepší.

Stejně tak prvních 1000 prvků Fibonacciho posloupnosti dává relativní četnosti výskytu prvních platných číslic, které se s předpokládanými shodují téměř přesně. Vysokou míru shody lze vysvětlit tím, že jde o asymptoticky geometrickou posloupnost (viz oddíl 3.3 nebo Raimi [5]).

# Kapitola 5

# Aplikace Benfordova zákona

V literatuře se objevuje několik návrhů na využití Benfordova zákona v praxi. Například Schatte [30] uvádí, že během dlouhých počítačových výpočtů v aritmetice s pohyblivou čárkou (tzv. *floating-point* aritmetika) mají mantisy zpracovávaných čísel přibližně Benfordovo rozdělení. Tuto informaci je možné využít při analýze zaokrouhlovacích chyb nebo při návrhu rychlejších algoritmů.

Knuth [31] předkládá určité výpočty, kterými zdůvodňuje předpoklad, že rozdělení mantisy ve vstupních datech jeho počítačových výpočtů je Benfordovo. Kromě jiného pak srovnává podle různých kritérií výhodnost používání dvojkové a šestnáctkové soustavy (s ohledem na předpokládané rozdělení mantis vstupních dat).

Z hlediska přesnosti výpočtů je vhodnější použití binární aritmetiky, horní odhad relativní chyby vzniklé zaokrouhlením je totiž poloviční než při použití šestnáctkové soustavy.

Naproti tomu výpočty prováděné v hexadecimální aritmetice probíhají o něco rychleji, protože není třeba tak často čísla normalizovat (posouvat desetinnou čárku). Cenou za rychlost je tedy nižší přesnost a naopak.

Jiná možnost využití Benfordova zákona je spíše psychologické povahy. Hill [32] uvádí příklad loterie provozované v americkém státě Massachussets: hráči sází na jedno čtyřciferné číslo, poté je taženo vítězné číslo a výhra se rozdělí mezi všechny hráče, kteří na toto číslo vsadili. Protože všechna čísla mohou být vítězná se stejnou pravděpodobností, je výhodné sázet na taková čísla, na která nikdo jiný nesází (aby se případná výhra dělila mezi co nejmenší počet výherců).

Pokud lidé sází na čísla, s nimiž se setkávají (v novinách, v práci, apod.) a pokud se taková čísla řídí Benfordovým zákonem, má pro hráče smysl vybírat čísla začínající vysokou číslicí. Idea je taková, že za uvedených předpokladů se lidé častěji setkávají s čísly jako 1989 nebo 2008, často na taková čísla sází a proto je výhodné sázet na čísla typu 9981 nebo 8002. Pravdě-

podobnost výhry se tím nezvýší, ale v případě vytažení takového čísla bude peněžitá výhra vyšší.

Poněkud užitečnější aplikace je ta, kterou navrhuje Varian [20]. Zabýval se vývojem matematických modelů předpovídajících ekonomický vývoj (viz oddíl 4.3) a fakt, že vstupní data jeho modelů se poměrně dobře řídí Benfordovým zákonem, ho přivedl k myšlence zkoumat rozdělení prvních platných číslic výstupních dat.

Pokud se toto ukáže jako výrazně odlišné od rozdělení, které předpovídá Benfordův zákon, navrhuje Varian takový model označit za „podezřelý“ a doporučuje vytvořit jiný model, který vykáže lepší shodu s Benfordovým zákonem.

Techniky založené na Benfordovu zákonu je možné využít i při odhalování nesrovnalostí v účetních záznamech, například daňových úniků nebo nesprávného proplácení faktur. Jednu takovou metodu předkládá Nigrini v článku [18].

Vychází z pozorování, že se soubor čísel z rozsáhlých účetních záznamů (zpracovaných bez chyb a neoprávněné manipulace s daty) velmi dobře řídí Benfordovým zákonem. Pokud prověřovaný soubor vykazuje velké odchylky od předpokládaného rozdělení mantis, je označen za podezřelý a Nigrini doporučuje podrobit dotyčnou společnost auditu, který podezření potvrdí nebo vyvrátí (může totiž jít o náhodnou odchylku, byť je její pravděpodobnost malá).

Za testovou statistiku bere Nigrini veličinu $DF$, kterou nazývá faktor zkreslení (anglicky *distortion factor*). Při počítání této statistiky je nejdříve nutné všechna čísla ze zkoumaného souboru případným posunutím desetinné čárky převést do intervalu $[10, 100)$ a vypočítat jejich průměr $AM$ (*actual mean*).

Za předpokladu, že se původní data řídí Benfordovým zákonem, je možné určit střední hodnotu tohoto průměru – $EM$ (*expected mean*). Nigrini [18] nabízí metodu výpočtu $EM$ a uvádí, že při velkém počtu záznamů $n$ tato hodnota závisí na $n$ pouze zanedbatelně a blíží se číslu 39.1.

Pokud by rozdělení mantis zkoumaných čísel bylo rovnoměrné, vyšlo by $EM = 55.0$. Hodnota 39.1 je tedy v souladu s tím, že podle Benfordova zákona jsou nižší mantisy preferovány před vysokými.

Nyní už je možné spočítat testovou statistiku $DF$ podle vzorce

$$DF = \frac{AM - EM}{EM}.$$

$DF$ určuje odchýlení skutečného průměru od očekávané hodnoty. Pokud se v původních datech vyskytují čísla s nízkými mantisami častěji, než předpovídá Benfordův zákon, bude $AM$ menší než $EM$ a hodnota $DF$ bude zá-

porná (a naopak). Díky tomu lze usoudit, zda případná manipulace s daty obnášela snižování nebo navyšování skutečných hodnot.

Například u položek odečitatelných z daní vzbudí manipulace směrem nahoru (navyšování čísel) větší podezření něž odchylka směrem dolů, která spíše vznikla náhodou.

Slabinou této metody je fakt, že není citlivá na *systematickou* manipulaci s daty. Pokud se původní údaje řídí Benfordovým zákonem, což by skutečná účetní data měla, pak jejich vynásobení stejným číslem shodu s Benfordovým zákonem nepokazí díky invarianci vzhledem k měřítku. Pokud jsou tedy údaje takto systematicky nadhodnocovány nebo podhodnocovány, faktor zkreslení to neodhalí.

Nigrini sice neuvádí žádné kritické hodnoty pro $DF$, ale předpokládá provádění této analýzy na mnoha souborech dat současně a následné prověření těch nejpodezřelejších (s nejvyšší absolutní hodnotou $DF$). To vystihuje například situaci finančního úřadu, který nemůže detailně prověřit údaje všech firem, ale tato metoda mu poskytuje možnost rozhodnout, které společnosti podrobit auditu.

Hill [8] uvádí, že díky softwaru, který test založený na faktoru zkreslení a další testy implementoval a na jehož vývoji se sám Nigrini podílel, bylo odhaleno sedm newyorských společností, které byly později obviněny z daňových podvodů. Úřady v Nizozemí projevily zájem tyto testy využít k odhalování daňových úniků a v USA se o jejich používání také jednalo.

Díky rostoucí dostupnosti digitálních dat a výkonnosti výpočetní techniky bude pokračovat trend k používání citlivějších statistických testů k odhalování podvodů a manipulace s daty. Aplikace Benfordova zákona je zřejmě pouze první krok.

# Závěr

V předcházejících kapitolách byl podán stručný přehled o fenoménu nerovnoměrného rozdělení prvních platných číslic známém jako Benfordův zákon.

Byla nastíněna historie Benfordova zákona od prvních zmínek a pokusů o vysvětlení až po některé nedávné výsledky a moderní aplikace, například při odhalování účetních podvodů. Také bylo odvozeno kritérium umožňující rozhodnout, zda se daný soubor číselných údajů Benfordovým zákonem řídí nebo ne.

Rozborem konkrétních dat, vybraných a zpracovaných autorem, potom byly ilustrovány různé vlastnosti Benfordova zákona, jako je invariance vzhledem ke změně měřítka a vzhledem ke změně základu číselné soustavy.

Dále by bylo zajímavé prostudovat nejmodernější teoretické výsledky v oblasti Benfordova zákona a také pokročilé aplikace, například citlivější testy účetních dat založené na podrobnějším rozboru číselných údajů (hlavně na zkoumání rozdělení prvních dvou, resp. tří platných číslic nebo naopak posledního dvojčíslí). Rozsah této práce už ale hlubší proniknutí do těchto oblastí nedovoluje.

# Literatura


[1] Newcomb, S.: *Note on the frequency of use of the different digits in natural numbers*, Amer. J. Math. **4** (1881) 39–40.

[2] Benford, F.: *The law of anomalous numbers*, Proc. Amer. Philos. Soc. **78** (1938) 551–572.

[3] Anděl, J.: *Základy matematické statistiky*, Matfyzpress, Praha, 2007.

[4] Flehinger, B. J.: *On the probability that a random integer has initial digit A*, Amer. Math. Monthly **73** (1966) 1056–1061.

[5] Raimi, R. A.: *The first digit problem*, Amer. Math. Monthly **83** (1976) 521–538.

[6] Pinkham, R. S.: *On the distribution of first significant digits*, Ann. Math. Statist. **32** (1961) 1223–1230.

[7] Jarník, V.: *Integrální počet (II)*, Academia, Praha, 1984.

[8] Hill, T. P.: *A statistical derivation of the significant-digit law*, Statist. Sci. **10** (1995) 354–363.

[9] Hill, T. P.: *Base-invariance implies Benford's law*, Proc. Amer. Math. Soc. **123** (1995) 887–895.

[10] Kallenberg, O.: *Random measures*, Academic Press, New York, 1983.

[11] Dupač, V., Hušková, M.: *Pravděpodobnost a matematická statistika*, Karolinum, Praha, 2005.

[12] Goudsmit, S. A., Furry, W. H.: *Significant figures of numbers in statistical tables*, Nature **154** (1944) 800–801.

[13] Furry, W. H., Hurwitz, H.: *Distribution of numbers and distribution of significant figures*, Nature **155** (1945) 52–53.

[14] http://en.wikipedia.org/wiki/Mixture_density (aktualizace 3. 7. 2008).

[15] Hardy, G. H., Wright, E. M.: *An introduction to the theory of numbers*, 4.vyd., Oxford Univ. Press, New York, 1960.

[16] Diaconis, P.: *The distribution of leading digits and uniform distribution mod 1*, Ann. Probab. **5** (1977) 72–81.

[17] Diaconis, P., Freedman, D.: *On rounding percentages*, J. Amer. Statist. Assoc. **74** (1979) 359–364.

[18] Nigrini, M.: *A taxpayer compliance application of Benford's law*, J. Amer. Taxation Assoc. **18** (1996) 72–91.

[19] Ley, E.: *On the peculiar distribution of the U.S. stock indexes' digits*, Amer. Statist. **50** (1996) 311–313.

[20] Varian, H.: *Benford's law*, Amer. Statist. **26** (1972) 65–66.

[21] Raimi, R. A.: *The peculiar distribution of first digits*, Sci. Amer. **221** (1969) 109–120.

[22] Burke, J., Kincanon, E.: *Benford's law and physical constants: the distribution of initial digits*, Amer. J. Phys. **59** (1991) 952.

[23] Buck, B., Merchant, A., Perez, M.: *An illustration of Benford's first digit law using alpha decay half lives*, Eur. J. Phys. **14** (1993) 59–63.

[24] Český statistický úřad: *Počet obyvatel v obcích České republiky k 1. 1. 2007* (http://www.czso.cz/csu/2007edicniplan.nsf/p/1301-07, aktualizace 16. 9. 2007).

[25] Bateman, G., Egan, V.: *Encyklopedie Zeměpis světa*, Columbus, Praha, 1999.

[26] Ministerstvo vnitra České republiky: *Přehled počtu nehod na dálnicích v České republice za rok 2007* (http://www.mvcr.cz/statistiky/doprava/2007/dalnice_2007.xls, aktualizace 27. 3. 2008).

[27] http://www.nigrini.com/images/file_sizes.html (aktualizace 26. 3. 2008).

[28] http://www.pocitadlo.cz/kstat.php?kat=3 (aktualizace 4. 4. 2008).

[29] Všeobecná zdravotní pojišťovna České republiky: *Volně prodejné léčivé přípravky přihlášené do číselníku VZP ČR k 1. 2. 2008* (http://www.vzp.cz/cms/internet/cz/Lekari/Ciselniky/Volne_prodejne, aktualizace 6. 4. 2008).

[30] Schatte, P.: *On mantissa distributions in computing and Benford's law*, J. Inform. Process. Cybernet. **24** (1988) 443–455.

[31] Knuth, D.: *The art of computer programming*, 2. díl, Addison-Wesley, New York, 1969.

[32] Hill, T. P.: *The significant-digit phenomenon*, Amer. Math. Monthly **102** (1995) 322–327.

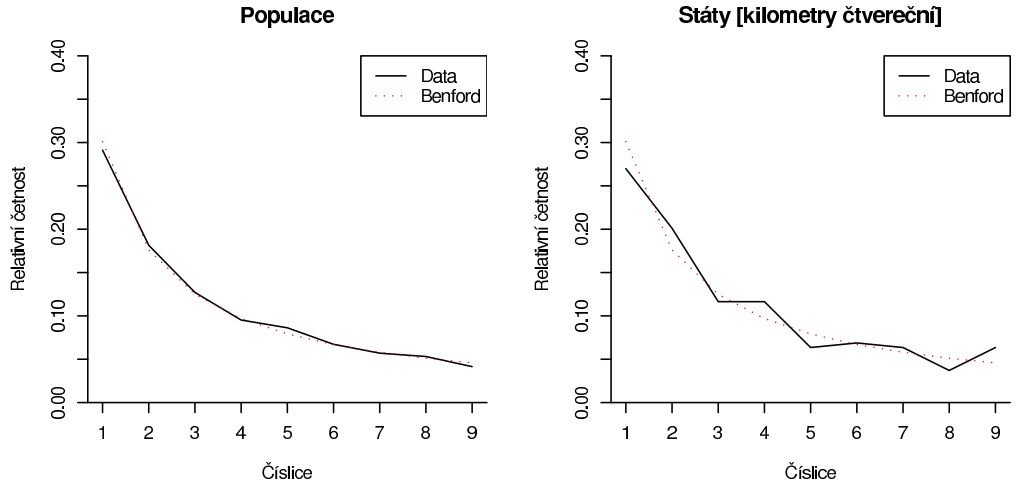
# Příloha A

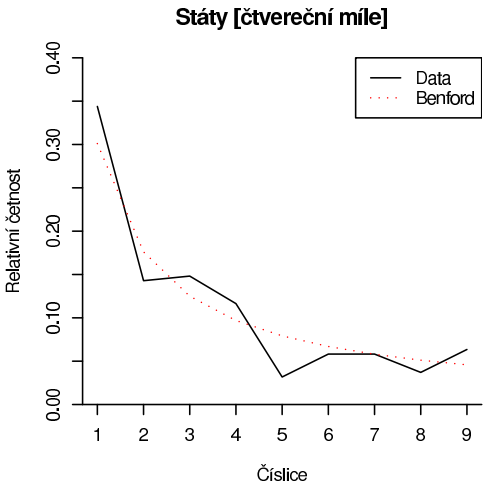
# Relativní četnosti výskytu prvních platných číslic

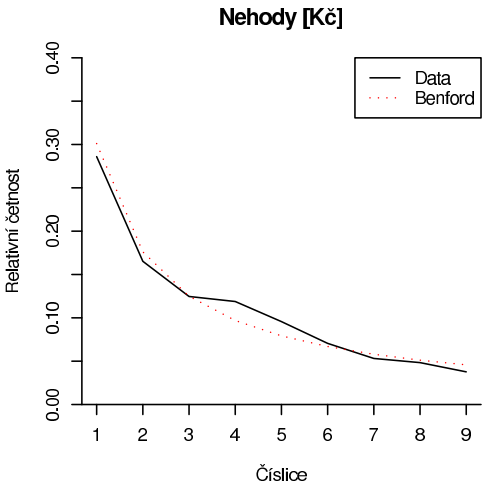
Následující grafy ukazují relativní četnosti výskytu prvních platných číslic v souborech dat popisovaných v části 4.4. Plnou čarou jsou znázorněny hodnoty zjištěné v konkrétním souboru, pro srovnání jsou v grafech uvedeny i hodnoty předpovídané Benfordovým zákonem. Ty jsou znázorněny červenou tečkovanou čarou.

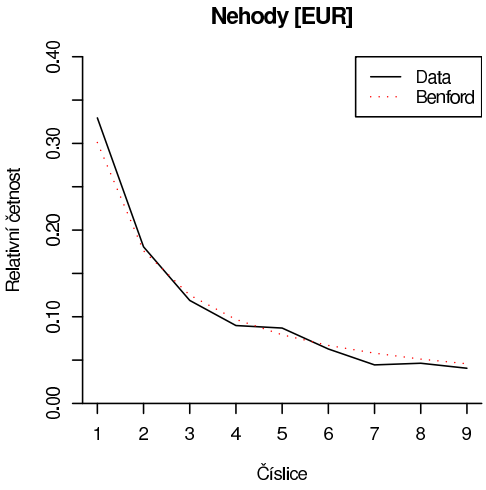
Linky v grafech vystupují pouze jako spojnice bodů odpovídajících zjištěným, resp. předpovídaným hodnotám a samy o sobě nemají žádný věcný význam. Jsou použity pouze ke zlepšení přehlednosti grafů a snadnějšímu vizuálnímu posouzení shody dat s Benfordovým zákonem.

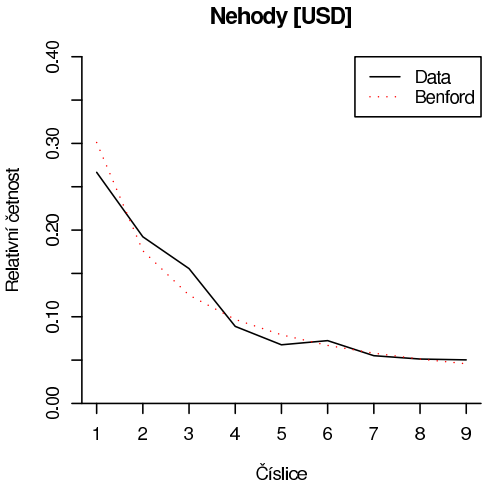
U většiny grafů jde o první platné číslice při vyjádření čísel v desítkové soustavě, kromě těch označených „Soubory [8]“ a „Soubory [16]“, tam jde o čísla vyjádřená v osmičkové, resp. šestnáctkové soustavě. V tom případě jsou pro srovnání znázorněny hodnoty předpovídané Benfordovým zákonem upraveným pro odpovídající číselnou soustavu (viz oddíl 4.4).

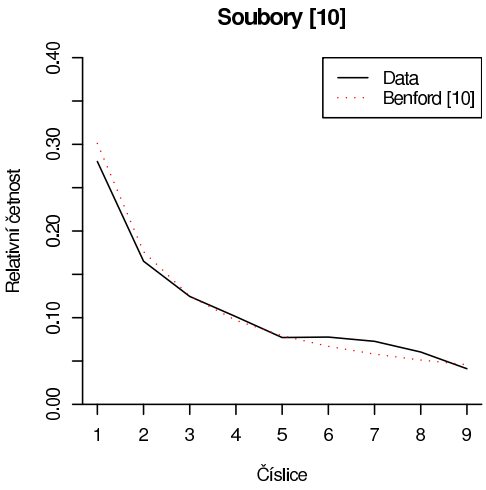



Státy [čtvereční míle]
Data
Benford
Relativní četnost
0.00
0.10
0.20
0.30
0.40
1
2
3
4
5
6
7
8
9
Číslice

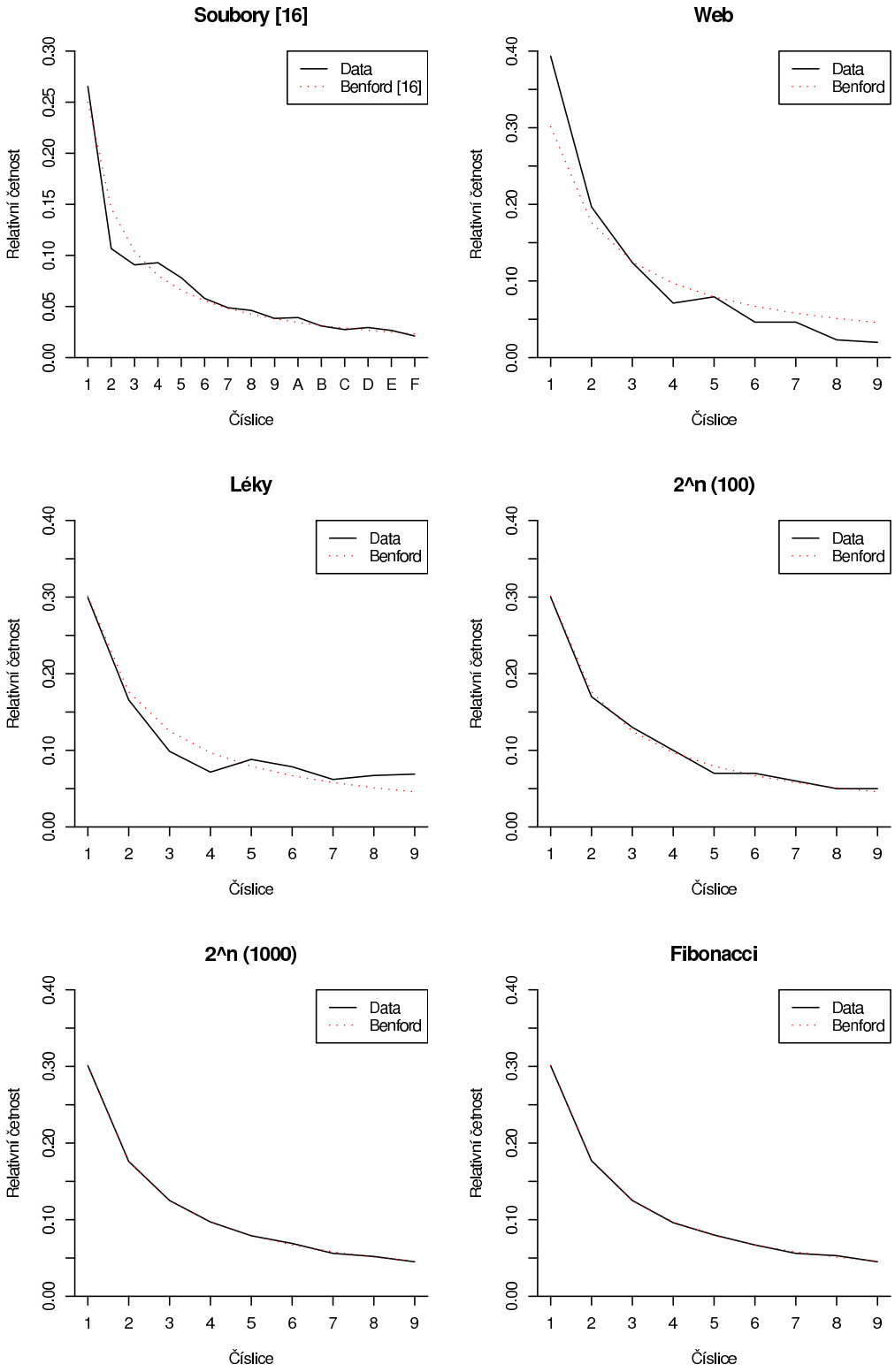

Nehody [Kč]
Data
Benford
Relativní četnost
0.00
0.10
0.20
0.30
0.40
1
2
3
4
5
6
7
8
9
Číslice


Nehody [EUR]
Data
Benford
Relativní četnost
0.00
0.10
0.20
0.30
0.40
1
2
3
4
5
6
7
8
9
Číslice


Nehody [USD]
Data
Benford
Relativní četnost
0.00
0.10
0.20
0.30
0.40
1
2
3
4
5
6
7
8
9
Číslice


Soubory [10]
Data
Benford [10]
Relativní četnost
0.00
0.10
0.20
0.30
0.40
1
2
3
4
5
6
7
8
9
Číslice


Soubory [8]
Data
Benford [8]
Relativní četnost
0.00
0.10
0.20
0.30
0.40
1
2
3
4
5
6
7
Číslice


Soubory [16]
Data
Benford [16]
Relativní četnost
Číslice
Web
Data
Benford
Relativní četnost
Číslice
Léky
Data
Benford
Relativní četnost
Číslice
2^n (100)
Data
Benford
Relativní četnost
Číslice
2^n (1000)
Data
Benford
Relativní četnost
Číslice
Fibonacci
Data
Benford
Relativní četnost
Číslice