Učitel matematiky

Luděk Spíchal
Benfordův zákon

Učitel matematiky, Vol. 28 (2020), No. 3, 131–149

Persistent URL: http://dml.cz/dmlcz/148641

BENFORDŮV ZÁKON

Luděk Spíchal

Úvod

V roce 1881 popsal Simon Newcomb[1] zajímavé zjištění. Logaritmické tabulky, tehdy a ještě dlouho poté používané pro výpočty, které dnes běžně provádíme kalkulačkou, byly nejvíce ohmatané na stránkách popisujících čísla začínající číslicí 1 (Bellos, 2016). Zjištění zůstalo nepovšimnuté až do roku 1938, kdy jej nezávisle znovu objevil Frank Benford[2], jehož jméno dnes zákon popisující různou distribuci prvních číslic nese.

Cílem článku je uvést popis základních principů Benfordova zákona, porovnat četnosti prvních (popř. druhých) číslic v číslech tvořících datové soubory získané z veřejně dostupných statistik s teoretickými četnostmi udávanými Benfordovým zákonem, vhodným statistickým testem ověřit shodu empirických a teoretických hodnot datových souborů a nabídnout příklad možného využití popisovaného zákona.

Benfordův zákon

Benfordův zákon (*first-digits law, first-digit phenomenon*) vychází z empirických pozorování, ze kterých vyplývá, že v mnoha přirozeně se vyskytujících souborech číselných dat nemají první číslice stejné zastoupení, ale řídí se určitým typem logaritmické distribuce (Berger & Hill, 2011a). V datových souborech s náhodnou

[1] Newcombe, S. (1835–1909) byl kanadsko-americký astronom a matematik.

[2] Benford, F. (1883–1948) byl americký elektroinženýr a fyzik pracující v laboratořích firmy General Electric.

distribucí čísel (včetně výsledků početních operací) je pravděpodobnost (relativní četnost) výskytu menších číslic na první pozici větší než číslic větších (Kruger, 2017).

Newcomb odhadl, že pravděpodobnost výskytu platné číslice na první pozici je:

$$P(d_1) = \log_{10}\left(1 + \frac{1}{d_1}\right),$$

kde $d_1 = 1, \ldots, 9$.

Zákon tedy říká, že pravděpodobnost výskytu číslice 1 na první pozici je $\log_{10} 2 \cong 0{,}301$, pravděpodobnost výskytu číslice 2 na první pozici je $\log_{10}(3/2) \cong 0{,}176$, pravděpodobnost výskytu číslice 3 na první pozici je $\log_{10}(4/3) \cong 0{,}124$ atd., až k číslici 9, kde $\log_{10}(10/9) \cong 0{,}046$. V souborech obsahujících alespoň stovky čísel se tak vyskytují na první pozici číslice s relativní četností uvedenou v tabulce 1 a obrázku 1.

Tab. 1: Očekávané relativní četnosti číslic podle Benfordova zákona (Nigrini, 1996)

Číslice	**1. pozice**	**2. pozice**	**3. pozice**	**4. pozice**
0		0,119 68	0,101 78	0,100 18
1	0,301 03	0,113 89	0,101 38	0,100 14
2	0,176 09	0,198 82	0,100 97	0,100 10
3	0,124 94	0,104 33	0,100 57	0,100 06
4	0,096 91	0,100 31	0,100 18	0,100 02
5	0,079 18	0,096 68	0,099 79	0,099 98
6	0,066 95	0,093 37	0,099 40	0,099 94
7	0,057 99	0,090 35	0,099 02	0,099 90
8	0,051 15	0,087 57	0,098 64	0,099 86
9	0,045 76	0,085 00	0,098 27	0,099 82

Benfordův zákon je možné formulovat v obecnějším tvaru popisujícím pravděpodobnost výskytu druhé číslice (Berger & Hill, 2011; Hindls & Hronová, 2015):

$$P(d_2) = \sum_{k=1}^{9} \log_{10}\left(1 + \frac{1}{10k + d_2}\right), \quad d_2 = 0, \ldots, 9,$$

případně další platné číslice:

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

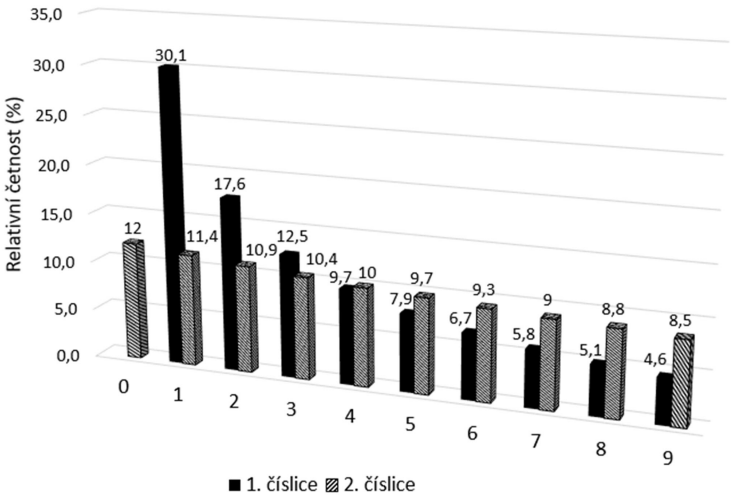
$d_k = 0, \ldots, 9$.

Pravděpodobnost výskytu číslic na jednotlivých pozicích se postupně vyrovnává a od páté číslice se blíží rovnoměrnému rozdělení, tj. pravděpodobnost výskytu každé z číslic $0, \ldots, 9$ je cca 10 %.

Uveďme několik příkladů oblastí, kde byla prokázána frekvence počátečních číslic podle Benfordova zákona. Patří mezi ně např. čísla tvořící Fibonacciho a Lucasovu posloupnost (Berger, 2011a), počáteční číslice fyzikálních konstant (Burke & Kincanon, 1991), emise skleníkových plynů, síla a hloubka zemětřesení (Sambridge, Tkalčić, & Jackson, 2010), kontrola věrohodnosti údajů z klinických studií (Beer, 2009), daňové příjmy (Nigrini, 1996), ceny akcií na burze (Pietronero at al., 2001). Uvedený výčet není zdaleka úplný, počet článků zabývajících se Benfordovým zákonem je v posledních letech poměrně značný.

Na druhou stranu v řadě doložených případů (množina přirozených čísel, prvočísla) četnosti prvních číslic neodpovídají Benfordovu zákonu (Berger & Hill, 2011b). Rovněž v některých dalších experimentálních číselných souborech se platnost zákonu nepotvrdila, např. Ausloos (2015).[3]

[3]Studie se zaměřila na data narození dětí v rodinách s různou náboženskou příslušností. Distribuce dat narození neprokázala shodu s Benfordovým zákonem.



Obr. 1: Relativní četnost číslic na první, resp. druhé pozici podle Benfordova zákona

Nezávislost (invariance) ke změně měřítka

T. Hill[4] v souvislosti s Benfordovým zákonem prohlásil, že pokud existuje nějaký univerzální zákon, který řídí rozdělení číslic, může to být pouze tento zákon (Bellos, 2016). Ukázal, že zákon představuje jediné rozdělení, které není závislé na měřítku (Hill, 1995).

Nezávislost na měřítku můžeme demonstrovat na příkladu čísel tvořících Lucasovu[5] posloupnost (L_n). Rekurentní vzorec posloupnosti je:

$$L_n = L_{n-1} + L_{n-2},\ L_1 = 2,\ L_2 = 1.$$

[4]Hill, T. (nar. 1943) je americký matematik zabývající se teorií pravděpodobnosti, zejména Benfordovým zákonem.

[5]Lucas, F. É. A. (1842–1891) byl francouzský matematik.

Lucasovu posloupnost tedy tvoří posloupnost čísel

$$2, 1, 3, 4, 7, 11, 18, \ldots$$

V tabulce 2 je uvedena relativní četnost prvních číslic v L_n, $8L_n$ a $20L_n$. Snadno zde zjistíme, že nezávisle na měřítku jsou relativní četnosti prvních číslic blízké Benfordovu zákonu.

Tab. 2: Relativní četnost prvních číslic v L_n, $8L_n$ a $20L_n$

Číslice	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	$\boldsymbol{\chi^2}$
B. zákon	30,1	17,6	12,5	9,7	7,9	6,7	5,8	5,1	4,6	
L_n	31	17	14	10	8	5	7	4	4	**1,23**
$8L_n$	29	18	13	9	8	7	4	7	5	**1,44**
$20L_n$	28	18	13	10	7	8	6	5	5	**0,58**

Chí-kvadrát testem dobré shody nezamítáme v žádném z uvedených příkladů nulovou hypotézu shody s relativní četností číslic podle Benfordova zákona. Použití vyššího počtu členů posloupnosti by dále zpřesnilo shodu.

Nezávislost na měřítku lze ukázat také např. na relativní četnosti prvních číslic v souborech stejných peněžních hodnot uvedených ovšem v různých národních měnách (Pietronero at al., 2001). Obdobně bychom zjistili, že převod plošných výměr mezi jednotkou míle a kilometr nemění četnosti prvních číslic.[6]

Obce v ČR

Vhodnou statistikou k ověření Benfordova zákona jsou údaje o počtu občanů žijících v obcích v České republice. Statistika je každoročně aktualizovaná k 1. lednu a dostupná na stránkách Ministerstva vnitra ČR.[7]

[6] Většina číselných souborů, které distribucí prvních číslic vyhovují Benfordovu zákonu je nezávislá rovněž ke změně základu ($b \geq 2$). Rovnici pro první číslici lze tedy zapsat ve tvaru: $P(d_1) = \log_b \left(1 + \frac{1}{d_1}\right) = \log_b(d_1+1) - \log_b d_1$.

[7] Zdrojem dat je MV ČR, dostupné online: `http://www.mvcr.cz/clanek/statistiky-pocty-obyvatel-v-obcich.aspx`.

Vzhledem k velkému rozsahu dat se přirozeně nabízí otázka, zda distribuce obyvatelstva odpovídá Benfordovu zákonu. Současně může být zajímavou možností testovat nejen první, ale i druhé pořadí číslic. V tabulce 3 jsou uvedeny počty obcí, jejichž počet obyvatel začíná (resp. má na druhé pozici) určitou číslicí (stav k 1. lednu 2018).

Tab. 3: Obce v ČR

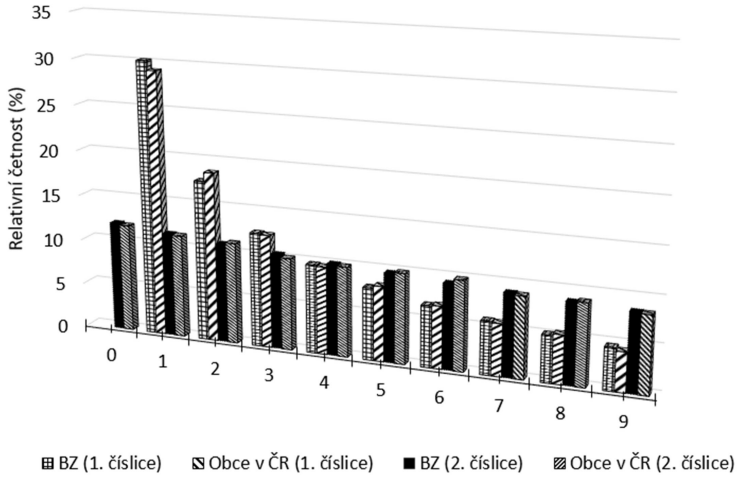
Číslice	**1. číslice**		**2. číslice**	
	Absolutní četnost	**Relativní četnost**	**Absolutní četnost**	**Relativní četnost**
0			740	11,83
1	1819	**29,07**	706	**11,28**
2	1167	**18,65**	694	**11,09**
3	775	**12,39**	631	**10,08**
4	601	**9,61**	616	**9,84**
5	513	**8,20**	616	**9,84**
6	427	**6,82**	614	**9,81**
7	356	**5,69**	555	**8,87**
8	332	**5,31**	558	**8,92**
9	267	**4,27**	527	**8,42**
Celkem	**6257**	**100,0**	**6257**	**100,0**

Grafické znázornění podobnosti empirických a teoretických hodnot (obr. 2) doplníme Pearsonovým χ^2 (chí-kvadrát) testem dobré shody (Ausloos, Herteliu, & Ileanu, 2015; Holčík & Komenda, 2015).[8]

[8]Předpokládejme, že náhodná veličina X nabývá konečného počtu hodnot $d_1, \ldots, d_m$, s pravděpodobnostmi $p_1, \ldots, p_m$, kde $\sum_{i=1}^{m} p_i = 1$. Požadovaná shoda nastává v případě, že se počet pozorování v jednotlivých variantách (pozorované četnosti $N_{i,o}$, $n = \sum_{i=1}^{m} N_i$) bude blížit hodnotě očekávaných četností $N_{i,e} = np_i$. Pokud má náhodná veličina X požadované rozdělení pravděpodobnosti, má statistika χ^2 chí-kvadrát rozdělení s $m-1$ stupni volnosti $X^2 = \sum_{i=1}^{m} \frac{(N_{i,o}-np_i)^2}{np_i} = \chi^2_{(m-1)}$. Nulovou hypotézu ($H_0$) o shodě

Dosazením empirických a teoretických hodnot získáme realizaci testové charakteristiky pro první číslici ve tvaru:

$$X^2_{d_1} = \sum_{i=1}^{9} \frac{(N_{i,d_1} - np_i)^2}{np_i} = 9{,}26.$$



Obr. 2: Porovnání empirických četností (tab. 3) s Benfordovým zákonem (BZ)

Srovnáme-li zjištěnou hodnotu testové charakteristiky s kvantilem příslušným hladině významnosti $\alpha = 0{,}05$ (Kruger & Yadavalli, 2017)

$$X^2_{d_1} \doteq 9{,}26 \leq \chi^2_8 = 15{,}51,$$

pak *nezamítáme* nulovou hypotézu shody distribuce prvních číslic s Benfordovým zákonem.

V případě druhé číslice je:

$$X^2_{d_2} = \sum_{i=0}^{9} \frac{(N_{i,d_2} - np_i)^2}{np_i} = 3{,}36.$$

rozdělení veličiny X s předpokládaným teoretickým (Benfordovým) rozdělením zamítáme na hladině významnosti α, když realizace testové statistiky překročí příslušný kvantil chí-kvadrát rozdělení, tedy když $X^2 \geq \chi^2_{(m-1)}(1-\alpha)$.

Srovnáme-li zjištěnou hodnotu testové charakteristiky s kvantilem příslušným hladině významnosti $\alpha = 0{,}05$ (Kruger & Yadavalli, 2017)

$$X^2_{d_2} \doteq 3,36 \leq \chi^2_9 = 16{,}91,$$

pak *nezamítáme* nulovou hypotézu shody distribuce druhých číslic s Benfordovým zákonem.

Na závěr můžeme tedy konstatovat, že relativní četnosti výskytu první a druhých číslic jsou v dobré shodě s Benfordovým zákonem.[9]

Benford – ano, či ne?

Benford ve svém původním článku sledoval distribuci prvních číslic v rozmanitých textech (Benford, 1938). Následující příklad vychází z textu, který se nachází na stránkách České lesnické akademie (ČLA) v Trutnově a týká se historie školy.[10] Text je zajímavý výskytem velkého počtu číselných údajů zaznamenávajících letopočty značně ovlivňující relativní četnost prvních číslic (tab. 4).

Přítomnost letopočtů značně ovlivňuje distribuci prvních číslic, zejména vzhledem k číslici 1. Takový soubor neobsahuje náhodnou distribuci prvních číslic, tj. *zamítáme* nulovou hypotézu shody s Benfordovým zákonem.

Po odstranění letopočtů se charakter souboru výrazně změní. Zaznamenané absolutní a relativní četnosti výskytu číslic na první pozici bez letopočtů jsou v tabulce 4. Testem dobré shody nulovou hypotézu *nezamítáme*, distribuce prvních číslic může odpovídat Benfordovu zákonu (obr. 3).

[9] V biologii, ekonomii a inženýrských disciplínách se obvykle používá hladina významnosti $\alpha = 0{,}05$, tj. data jsou s pravděpodobností 95 % uvnitř mezí. Pro první číslici je na hladině významnosti $\alpha = 0{,}05$ kvantil $\chi^2_8 = 15{,}51$, pro druhou číslici je kvantil $\chi^2_9 = 16{,}91$.

[10] F. Zuman: Lesnická škola v Zákupech – dříve v Bělé – v prvních čtyřech letech (Z kroniky Umlaufovy), dostupné online: `https://www.clatrutnov.cz/index.php/cs/skola/historie/32-historie-trutnov`.

Tab. 4: Historie ČLA

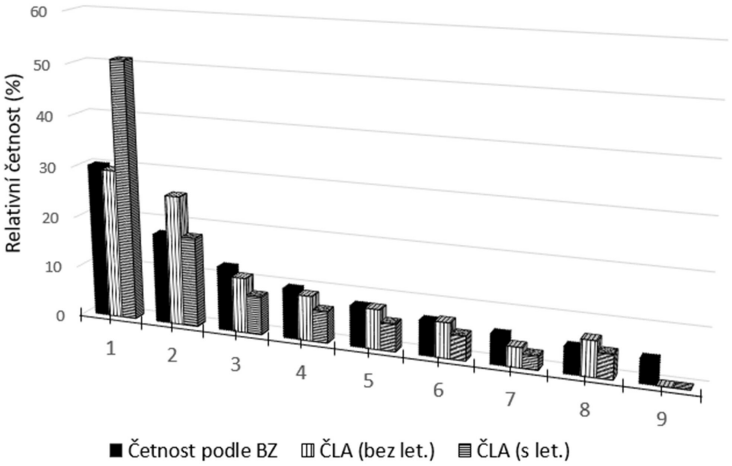
Číslice	1. číslice (s letopočty)		1. číslice (bez letopočtů)	
	Absolutní četnost	Relativní četnost	Absolutní četnost	Relativní četnost
1	75	**51,0**	30	**29,4**
2	26	**17,6**	26	**25,5**
3	11	**7,5**	11	**10,8**
4	9	**6,2**	9	**8,8**
5	8	**5,4**	8	**7,8**
6	7	**4,8**	7	**6,9**
7	4	**2,7**	4	**3,9**
8	7	**4,8**	7	**6,9**
9	0	**0**	0	**0**
Celkem	**147**	**100,0**	**102**	**100,0**
χ^2		**37,4**		**9,88**

Jak rozumět Benfordovu zákonu?

Přehled oblastí výskytu Benfordova zákonu uvedený v literatuře je opravdu působivý. Desítky, možná stovky popsaných příkladů distribuce prvních číslic na jedné straně poukazují na ukotvení zákonu v rozmanitých textech a číselných souborech, na druhou stranu přímo nenabízí klíč k vysvětlení podstaty tohoto fenoménu. Rigorózní vysvětlení, které je ovšem poměrně složité, vypracoval v 90. letech 20. století T. Hill (Hill, 1995).[11] Pokusy o intuitivní vysvětlení obvykle vycházejí z nezávislosti (invariance) vůči měřítku (viz výše) a základu logaritmu. Vychází z předpokladu, že hledaný univerzální zákon by neměl záviset na jednotce, ve které probíhá měření, nebo číselné soustavě, ve které měření probíhá.

[11] Hill, T. (nar. 1943) je americký matematik zabývající se teorií pravděpodobnosti, zejména Benfordovým zákonem. Více např.: `https://en.wikipedia.org/wiki/Ted_Hill_(mathematician)`.

Zjednodušené vysvětlení vyžadující základní znalost logaritmů a grafu funkce rozdělení pravděpodobnosti lze nalézt v článku Fewster (2009), ze kterého je převzata základní idea.[12] Cílem této sekce je upozornit zejména na obvyklé vlastnosti číselných souborů, které odpovídají Benfordovu zákonu.



Obr. 3: Porovnání empirické četnosti (tab. 4) s Benfordovým zákonem (první číslice)

Již v základních kurzech matematiky se zmiňuje možnost vyjádřit každé kladné reálné číslo ve tvaru

$$X = a \cdot 10^n,$$

kde $1 \leq a < 10$, $n \in \mathbb{Z}$. Pokud předchozí rovnici logaritmujeme

$$\log_{10} X = \log_{10}(a \cdot 10^n),$$

pak je

$$\log_{10} X = \log_{10}(a) + n.$$

[12]Hustota pravděpodobnosti je funkce, jejíž hodnotu pro libovolný zvolený prvek z množiny hodnot náhodné proměnné můžeme vyjádřit jako relativní četnost hodnoty tohoto prvku v rámci celé množiny možných hodnot.

Jestliže, např. pro číslo a platí, že $1 \leq a < 2$, pak po logaritmování je

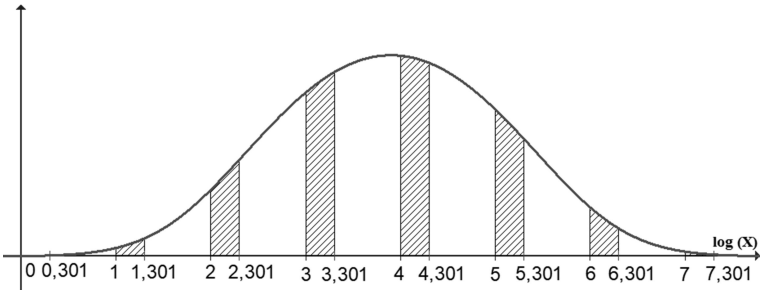
$$0 \leq \log_{10} a < 0{,}301,$$

nebo pro libovolné $n \in \mathbb{Z}$

$$n \leq \log_{10}(a) + n < 0{,}301 + n. \tag{1}$$

Poslední nerovnost ukazuje, že nezávisle na řádu čísla X je na jeho první pozici číslice 1, právě když dekadický logaritmus čísla a náleží intervalu $(n, n + 0{,}301)$, kde $n \in \mathbb{Z}$.

Uvažujme nyní soubor náhodných dat, o kterém budeme předpokládat, že se řídí nějakým rozdělením pravděpodobností. Z toho souboru náhodně vybereme číslo X, které vyjádříme jako $\log_{10} X$, a znázorníme na číselné ose. Pokud uvažované číslo X začíná 1, pak $\log_{10} X$ musí náležet nějakému intervalu $(n, n + 0{,}301)$, kde $n \in \mathbb{Z}$ (obr. 4). Pokud je dále křivka v obrázku 4 grafem hustoty pravděpodobnosti funkce $\log_{10} X$, pak pravděpodobnost, že číslo X začíná číslicí 1, je pravděpodobností, že $\log_{10} X$ leží v některém z vyšrafovaných pruhů. Součet ploch jednotlivých pruhů v grafu hustoty je pak hledanou pravděpodobností, že číslo X začíná číslicí 1.



Obr. 4: Graf hustoty pravděpodobnosti

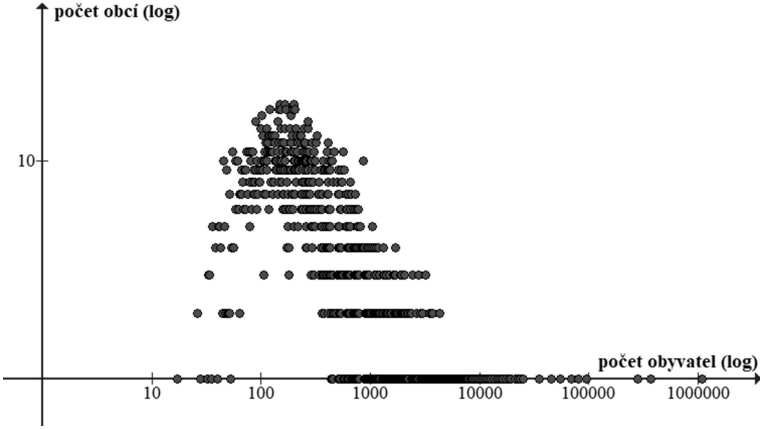
Fewster (2009) doplňuje výše uvedené vysvětlení konstatováním, že data obvykle odpovídají tím lépe relativní četnosti popisované Benfordovým zákonem,

- čím více v logaritmickém měřítku tvoří symetrickou křivku,

- čím více číselných řádů zahrnují (v grafu hustoty pravděpodobnosti se objeví více pruhů).

Scott & Fasli (2001) uvádí, že data mohou vyhovovat logaritmické distribuci popisované Benfordovým zákonem, pokud:

- je tvoří pouze kladné hodnoty,
- má graf (v logaritmickém měřítku) pouze jeden vrchol a vpravo od průměru se vyskytují odlehlejší hodnoty než vlevo (logaritmicko-normální rozdělení s pravým chvostem),
- není medián více než polovinou aritmetického průměru.



Obr. 5: Distribuce obyvatel v obcích ČR (log-log měřítko)

Uvedené závěry můžeme srovnat s distribucí obyvatel v obcích ČR (obr. 5), kde jsou hodnoty rozloženy do šesti číselných řádů a grafickým znázorněním je křivka protažená směrem doprava. Mediánová obec má 429 obyvatel, průměrná obec pak 1 644 obyvatel, tj. medián je méně než polovinou aritmetického průměru.

V řadě případů je tak možné posoudit data podle vlastností měřených hodnot. Tak, např. při měření tělesné výšky, budou určitě všechny hodnoty kladné a graf bude mít pouze jeden vrchol, nicméně medián nebude nejvýše polovinou průměru, hodnoty mediánu a průměru naopak budou velmi blízké. Pro distribuci tělesných výšek nelze tedy očekávat shodu s Benfordovým zákonem.

Odhalte falšovaná data

Vše, co jsme až dosud zmínili, bychom mohli považovat za zajímavou kuriozitu bez praktického využití. Nicméně nerovnoměrné zastoupení číslic na první pozici umožňuje analyzovat různé texty a hledat takové, které tomuto rozdělení četnosti neodpovídají. Pokud vyloučíme situace, kdy z nějakého konkrétního důvodu určité číslice (zejména jiné než malé) převažují, pak texty porušující uvedené rozdělení četnosti mohou vykazovat známky manipulace s daty.

Hill (1998) v tomto případě nabízí jednoduchou variantu testu (vhodného např. do hodin základů pravděpodobností) ukazujícího, že lidé při falšování dat obvykle nevolí číslice zcela náhodně. Doporučuje rozdělit studenty ve třídě na dvě poloviny, z nichž jedna bude např. 200krát házet mincí a zapisovat výsledky, zatímco druhá polovina pořadí výsledků zapíše „náhodně“. Dodává, že lidé jen zřídka v takovém případě „náhodně“ zapíší delší sekvence tvořené pouze jedním z možných výsledků, které se však v takto dlouhé sekvenci pokusů objevují s vysokou pravděpodobností.[13]

Bellos (2016) uvádí několik konkrétních případů použití Benfordova zákonu při odhalování manipulovaných dat. V jednom takovém případu S. de Marchi a J. T. Hamilton z Dukeovy univerzity prokázali falšování údajů o emisích kyseliny dusičné a olova továrnou v Severní Karolíně (Marchi & Hamilton, 2006). V jiném

[13] Zvažme jako model situaci, kdy hledáme pravděpodobnost, že v průběhu 200 hodů mincí se objeví sekvence alespoň 8 lícových stran mince. Výpočet uvedené pravděpodobnosti není zcela snadný, využít lze aproximaci uvedenou v knize Feller, W. (1960). *An Introduction to Probability Theory and Its Applications*, str. 325, dostupné online na `https://archive.org/details/in.ernet.dli.2015.124388/page/n347`. Nechť q_n je pravděpodobnost, že se při n hodech mincí neobjeví žádná posloupnost r po sobě jdoucích lícových stran mince. Pak $q_n \approx \frac{1-px}{(r+1-rx)q} \cdot \frac{1}{x^{n+1}}$, kde p je pravděpodobnost, že padne lícová strana mince ($q = 1 - p$), a x je nejmenší kladný kořen rovnice $1 - x + qp^r x^{r+1} = 0$. Jestliže $r = 8$, $p = q = 0{,}5$, pak rovnice $1 - x + 0{,}5 \cdot 0{,}5^8 x^9 = 0$ má nejmenší kladný kořen (např. Wolfram Alpha) $x \approx 1{,}001\,99$. Pro hledanou pravděpodobnost tedy platí $P \approx 1 - q_{200} \approx 1 - 0{,}68 \approx 0{,}32$. Pokud bychom snížili požadavek na délku sekvence, např. na 6 po sobě jdoucích lícových stran, pak odhadovaná pravděpodobnost vzroste na cca 80 %.

případě W. Mebane z Michiganské univerzity ukázal, že prezidentská volba v Íránu v roce 2009 byla zřejmě zmanipulovaná, neboť počty hlasů odevzdané pro úřadujícího prezidenta se neshodovaly s Benfordovým zákonem. Vysvětlením by podle Mebaneho mohlo být umělé přidání hlasů (Mebane, 2010).[14] Diekmann (2007) při odhalování nesrovnalostí v účetních datech doporučuje zaměřit pozornost spíše na druhé a další číslice v pořadí, neboť četnosti prvních číslic často vykazují charakter odpovídající Benfordovu zákonu.

Tab. 5: Počet obyvatel v obcích 3. typu v Praze

Obec 3. typu	Počet obyvatel (P1)	Počet obyvatel (P2)	Obec 3. typu	Počet obyvatel (P1)	Počet obyvatel (P2)
Praha 1	24 320	29 411	Praha 12	60 035	67 093
Praha 2	39 170	42 005	Praha 13	56 663	51 434
Praha 3	62 553	63 664	Praha 14	42 560	47 321
Praha 4	122 176	82 369	Praha 15	43 534	53 776
Praha 5	76 381	71 113	Praha 16	22 703	25 539
Praha 6	102 356	89 225	Praha 17	27 515	21 597
Praha 7	38 922	52 056	Praha 18	25 851	35 778
Praha 8	104 173	114 678	Praha 19	12 697	11 704
Praha 9	47 612	44 334	Praha 20	14 126	8 236
Praha 10	97 177	100 297	Praha 21	16 828	7 675
Praha 11	76 651	96 592	Praha 22	14 873	12 979

Zkusme tedy na závěr nabídnout možnost využití Benfordova zákona při posouzení dvou datových souborů, kde jeden ze souborů obsahuje úmyslně pozměněná data. Jako příklad vezmeme distribuci obyvatel v Praze, kterou z hlediska správního členění

[14] Autor v článku ukazuje, že v případě volebních podvodů je vhodné se zaměřit na frekvenci druhých číslic, frekvence prvních číslic nemusí být průkazná (zejména v případě menšího počtu hlasů odevzdaných v jednotlivých volebních místnostech).

tvoří 22 obcí 3. typu s počty obyvatel uvedenými v jednom ze sloupců v tabulce 5.[15] Který ze sloupců obsahuje správné údaje? Řešení naleznete v příloze A.

Závěr

Článek popsal zajímavý fenomén rozdělení číslic na vybraných pozicích označovaný jako Benfordův zákon. Platnost zákona jsme ukázali na příkladu distribuce obyvatelstva v obcích ČR. Zde jsme velmi dobrou shodu s Benfordovým zákonem získali jak pro první, tak pro druhé číslice v pořadí.

Zákon v posledních desetiletích nachází zajímavá využití, např. při odhalování účetních podvodů. Lidé snažící se falšovat např. účetní knihy mají tendenci čísla upravovat tak, aby začínala rovnoměrně všemi číslicemi. To, co na první pohled může vypadat jako „chytré podvádění", však naopak člověka znalého Benfordova zákona okamžitě upozorní na možnost podvodu. Samozřejmě jde pouze o podezření (indicii), které musí být následně podložené prokázáním podvodu. Durtschi et al. (2004) zmiňují použití Benfordova zákona při odhalování zpronevěr v účetních záznamech pomocí softwaru umožňujícího testovat tyto záznamy s ohledem na frekvenci prvních číslic. Zásadním problémem pro analytiky a auditory je v tomto případě volba vhodného vzorku jak s ohledem na rozsah dat, tak s ohledem na časový rámec. V účetních datech se v průběhu roku mohou objevit odchylky od očekávaného rozdělení, které neznamenají pokus o podvod. Benfordův zákon tak funguje jako síto k vyhledávání podezřelých dat, která budou dále prověřována.

Z pohledu českého čtenáře může být rovněž zajímavé použití Benfordova zákonu při posouzení práce zakladatele genetiky J. G. Mendela vzhledem k frekvenci prvních číslic, které uvádí Kruger & Yadavalli (2017).[16] Vychází z názoru anglického statistika

[15] Zdrojem dat je MV ČR, dostupné z `http://www.mvcr.cz/clanek/statistiky-pocty-obyvatel-v-obcich.aspx`.

[16] Mendel, J. G. (1822–1884) byl augustiniánský mnich a později opat kláštera na Starém Brně. Ve svých přírodovědných pozorováních se zaměřil na sledování kříženců hrachu. Zmíněnou práci uveřejnil roku 1866 pod názvem „Pokusy s rostlinnými hybridy".

R. Fishera, který ve své práci z r. 1936 nezpochybnil samotné Mendelovy závěry, řekl však, že „data jsou příliš dobrá, než aby byla pravdivá“.[17] Analýza dat provedená Krugerem poukazuje na zřetelné odchylky ve frekvenci některých číslic od frekvence dané Benfordovým zákonem. To může naznačovat jak možnou selekci dat ve prospěch takových, které podporovaly Mendelem navržený model křížení, tak fakt, že Mendel pokusy odchylující se od uvažovaného modelu doplňoval dalšími, aby dosáhl shody. Na Mendelovu obhajobu je na druhou stranu vhodné dodat, že exaktní statistika tehdy jako obor fakticky neexistovala a Mendel byl jedním z prvních přírodovědců, kteří v biologii aplikovali matematické metody.[18]

Problematiku Benfordova zákona lze využít k doplnění a zpestření výuky jak v oblasti logaritmů a jejich využití, tak v oblasti základů pravděpodobností a statistiky (rozdělení četnosti, práce s datovými soubory, grafy). Pro bližší seznámení s Benfordovým zákonem mohou určitě dobře posloužit zdroje uvedené na závěr tohoto článku (převážně v angličtině). V omezené míře lze nalézt informace o Benfordově zákonu i v česky psané literatuře. Bellos (2016) věnuje problematice Benfordova zákonu a dalších souvisejících zákonitostí celou kapitolu. Dále lze zmínit článek Seiberta & Zahrádky (2016) v časopise *Matematika – fyzika – informatika*, který obsahuje jinou variantu přístupného vysvětlení platnosti Benfordova zákonu.

Přes vše, co jsme zmínili, je při aplikování Benfordovým zákonem jistě užitečné postupovat s rozmyslem, nikoliv se slepou vírou v platnost. Ponechání prostoru pro zkušenost a intuici může zabránit chybnému použití, např. při odhalování podvodů a plagiátů.

Na závěr dodejme, že ačkoliv řada aspektů souvisejících s Benfordovým zákonem stojí na celkem pevných základech (např. Hill, 1995), nebyl dosud sjednocen přístup, který by současně spojil výskyt Benfordova zákona v tak vzdálených oblastech jako je

[17] Fisher, R. A. (1890–1962) byl anglický statistik, evoluční biolog a genetik. V anglicky psané literatuře je spor uváděn pod názvem „Fisher-Mendel controversy“.

[18] Více např. `https://cs.wikipedia.org/wiki/Gregor_Mendel`.

teorie čísel, dynamické systémy, statistika a reálná data (Berger, 2011b) a současně nabídl dostatečně intuitivní vysvětlení podstaty tohoto fenoménu.

Literatura

[1] Bellos, A. (2016). *Alex za zrcadlem. Jak se čísla odrážejí v životě a život v číslech.* Dokořán.

[2] Hill, T. P. (1995). A statistical derivation of the significant-digit law. *Statistical Science, 10*, 354–363.

[3] Hill, T. P. (1998). The First Digit Phenomenon: A century-old observation about an unexpected pattern in many numerical tables applies to the stock market, census statistics and accounting data. *American Scientist, 86*(4), 358–363.

[4] Berger, A., & Hill, T. P. (2011a). A basic theory of Benford's Law. *Probability Surveys, 8*, 1–126.

[5] Berger, A., & Hill, T. P. (2011b). Benford's Law strikes back: No simple explanation in sight for mathematical gem. *Springer Science, Business Media*, LLC, *33*(1), 85–91.

[6] Ausloos, M., Herteliu, C. & Ileanu, B. (2015). Breakdown of Benford's law for birth data. *Physica A, 419*, 736–745.

[7] Kruger, P. S., & Yadavalli, V. S. S. (2017). The power of one: The Benford's law. *South African Journal of Industrial Engineering, 28*(2), 1–13.

[8] Holčík, J., Komenda, M. (Eds.) et al. (2015). *Matematická biologie: e-learningová učebnice* [online]. Brno: Masarykova univerzita. `http://portal.matematickabiologie.cz/`

[9] Hindls, R., & Hronová, S. (2015). Benford's Law and possibilities for its use in governmental statistics. *Statistika, 95*(2), 54–64.

[10] Fewster, R. M. (2009). A simple explanation of Benford's Law. *The American Statistician, 63*(1), 26–32.

[11] Mir, T. A. (2012). The law of the leading digits and the world religions. *Physica A, 391*, 792–798.

[12] Pietronero, L., Tosatti, E., Tosatti, V., & Vespignani, A.

(2001). Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf. *Physica A, 293*, 297–304.

[13] Nigrini, M., J. (1996). Taxpayer compliance application of Benford's law. *Journal of the American Taxation Association, 18*(1), 72–92.

[14] Seiber, J., & Zahrádka, J. (2016). O čem pojednává Benfordův zákon. *Matematika – fyzika – informatika, 25*(2), 89–98. http://mfi.upol.cz/files/25/2502/mfi_2502_089_098.pdf

[15] Durtschi, C., Hillison, W., & Pacini, C. (2004). The effective use of Benford's Law to assist in detecting fraud in accounting data. *Journal of Forensic Accounting, V*, 17–34.

[16] Diekmann, A. (2007). Not the first digit! Using Benford's Law to detect fraudulent scientific data. *Journal of Applied Statistics, 34*(3), 321–329.

[17] Scott, P. D., & Fasli, M. (2001). Benford's Law: An empirical investigation and a novel explanation. CSM Technical Report 349. https://cswww.essex.ac.uk/technical-reports/2001/CSM-349.pdf

[18] Burke, J., & Kincanon, E. (1991). Benford law and physical constants – The distribution of initial digits. *American Journal of Physics, 59*(10), 952.

[19] Sambridge, M., Tkalčić, H., & Jackson, A. (2010). Benford's law in the natural sciences. *Geophysical Research Letters, 37*(22). https://doi.org/10.1029/2010GL044830

[20] Beer, T. W. (2009). Terminal digit preference: beware of Benford's law. *Journal of Clinical Pathology, 62*, 192.

[21] Benford, F. (1938). The law of anomalous numbers. *Proc. American Philosophical Society, 78*(4), 551–572.

[22] Marchi, S., & Hamilton, J. T. (2006). Assessing the accuracy of self-reported data: an evaluation of the toxics release inventory. *Journal of Risk and Uncertainty, 32*(1), 57–76.

[23] Mebane, W. R., Jr. (2010). Fraud in the 2009 presidential election in Iran? *Chance, 23*(1), 6–15.

Abstract

This article refers to the Benford's Law, also known as the first--digit law, which is one of the most mysterious law of nature. The article provides the basic characteristic of the law and a simple, intuitive explanation of why and when the law applies. The last part is focused on using the law in case of suspicion that the data are manipulated.

Luděk Spíchal
Masarykova univerzita v Brně
Ústav matematiky a statistiky
Kotlářská 267/2
611 37 Brno
Česká lesnická akademie Trutnov
Lesnická 9
541 11 Trutnov

Příloha A

Řešení je celkem nasnadě, určíme četnosti prvních číslic v obou souborech.

Číslo	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Celkem**
Očekávaná četnost	6,62	3,87	2,75	2,13	1,74	1,48	1,28	1,12	1,01	22
Varianta P1	7	4	2	3	1	2	2	0	1	22
Varianta P2	4	3	1	3	3	2	2	3	1	22

I přes velmi malý rozsah dat lze snadno porovnáním s očekávanou četností usoudit, že se pozměněné údaje o počtu obyvatel nacházejí pravděpodobněji ve variantě P2. Zajímavá je rovněž (přes velký rozsah dat) míra shody varianty P1 s distribucí určenou Benfordovým zákonem, kterou ovšem v tomto případě není vhodné pro malý rozsah dat ověřovat testem dobré shody.