O čem pojednává Benfordův zákon

JAROSLAV SEIBERT, JAROMÍR ZAHRÁDKA

Fakulta ekonomicko-správní Univerzity Pardubice

Každý dobrý učitel matematiky se stále snaží hledat vhodné příležitosti, které mohou ukázat matematickou teorii v poněkud netradičních, ale přitom navenek atraktivnějších situacích. Navíc je jistě účelné, když se jedná o užití matematických poznatků v problematice, která je poněkud vzdálená od tradičních oblastí aplikací matematiky. Jednomu méně známému matematickému poznatku je věnovaný tento náš příspěvek.

Uvažujme libovolnou množinu číselných dat, která vyjadřují hodnotu jisté přirozeně vymezené veličiny z reálného světa. Přitom nezáleží na tom, zda se jedná o data geografická, ekonomická (ceny zboží, fakturované částky, platby pojištění), tabulky fyzikálních veličin, hodnoty některých funkčních závislostí mezi veličinami na jistém diskrétním definičním oboru a podobně. Jedním z důležitých předpokladů je, aby byl rozsah souboru alespoň v řádu stovek nebo ještě lépe tisíců. Určuje se potom četnost výskytu jednotlivých číslic na prvním platném místě. V našem případě tedy první číslice různé od nuly v zápisu číselného údaje v desítkové soustavě. Pro jednoduchost budeme v dalším textu používat pouze pojmenování „první číslice“. Zdá se zřejmé, že by pravděpodobnost $P(d)$ výskytu číslice $d = 1,2,...9$ jako první číslice měla podléhat rovnoměrnému rozložení, tedy konkrétně $P(d) = \frac{1}{9} = 0,11...$. Ve skutečnosti se v souborech přirozených dat nejčastěji objevuje jako první číslice jednička a četnost výskytu dalších číslic postupně klesá v pořadí od 2 až k 9. Jak se můžeme v mnoha časopiseckých pramenech přesvědčit, je z uváděných hodnot zřejmé, že v těchto souborech zhruba 30% čísel začíná jedničkou a čím vyšší je první číslice, tím je menší pravděpodobnost, že se objevuje na začátku zápisu čísel.

Úvod k problematice zákona první číslice

Proveďme nejprve jednoduchý experiment se členy posloupností $a_n = n\sqrt{n}$, $b_n = n\,|\sin n|$, a členy tří rekurentně zadaných posloupností. Jde o Fibonacciovu posloupnost danou předpisem $F_{n+2} = F_{n+1} + F_n$, kde $F_1 = F_2 = 1$, Lucasovu posloupnost $L_{n+2} = L_{n+1} + L_n$, $L_1 = 1$, $L_2 = 3$ a Pellovu posloupnost $P_{n+2} = 2P_{n+1} + P_n$, $P_1 = 1, P_2 = 2$. Funkční vyjádření členů těchto posloupností (v závislosti na n) lze najít například řešením jejich rekurentních předpisů jako lineárních diferenčních rovnic 2. řádu s danými počátečními členy. Pro Fibonacciovu posloupnost pak platí tzv. Binetův vzorec $F_n = \frac{\alpha^n - \beta^n}{\alpha - \beta}$, kde $\alpha = \frac{1+\sqrt{5}}{2}$, $\beta = \frac{1-\sqrt{5}}{2}$. Podobně lze odvodit, že pro Lucasovu posloupnost platí $L_n = \alpha^n + \beta^n$ a pro Pellovu posloupnost $P_n = \frac{\gamma^n - \delta^n}{\gamma - \delta}$, kde $\gamma = 1+\sqrt{2}$, $\delta = 1-\sqrt{2}$.

Použitím programu MATLAB jsme určili hodnoty dostatečného počtu členů zmíněných posloupností. V tab.1 jsou pak uvedeny procentuálně vyjádřené relativní četnosti výskytu jednotlivých číslic jako prvních číslic členů těchto posloupností. Počet testovaných členů zkoumaných posloupností byl ovlivněn možnostmi zobrazení čísel v použitém softwaru.

Tab. 1: Relativní četnost prvních číslic u uvedeného počtu členů vybraných posloupností

První číslice	1	2	3	4	5	6	7	8	9	Počet členů
a_n	16,13	13,53	12,08	11,10	10,38	9,81	9,36	8,97	8,65	10^6
b_n	20,24	16,52	13,87	12,01	10,39	8,94	7,58	6,15	4,29	10^6
F_n	30,27	17,71	12,63	9,20	8,20	6,43	5,56	5,56	4,44	1476
L_n	30,44	17,15	12,85	9,89	7,80	6,35	6,19	4,90	4,42	1474
P_n	30,45	18,25	11,81	9,24	8,29	6,22	6,19	4,68	4,87	806

Jednoduchým porovnáním uvedených četností je vidět, že výskyt prvních číslic klesá od 1 k 9. Navíc je zřejmé, že větší shodu pozorujeme u členů rekurentně zadaných posloupností, kde relativní četnosti pro jednotlivé číslice vykazují pouze velmi malé odchylky. Je proto možné, že svoji roli v těchto případech hraje typ funkčního vyjádření členů posloupnosti, konkrétně závislost exponenciální. Z tohoto hlediska budeme k celému problému přistupovat v dalším textu.

Historické poznámky

Pozorovanou zákonitost poprvé zveřejnil v roce 1881 kanadský matematik a astronom S.Newcombe [6] v časopise *The American Journal of Mathematics*. Jeho tvrzení vycházelo z pozorování, že v logaritmických tabulkách v technické knihovně jsou evidentně nejvíce ohmatané stránky s čísly začínajícími jedničkou. Jeho pouze dvoustránkový text si však nezískal žádnou pozornost a upadl v zapomnění. Autor neuvedl žádnou analýzu konkrétních souborů dat, zato se pokusil o určité matematické zdůvodnění tohoto výsledku.

Ještě několik desetiletí tak jeho tvrzení nebyla věnována pozornost. Až v roce 1938 tento z určitého hlediska přírodní jev znovu objevil fyzik F.Benford [1]. Ten se však celým problémem zajímal mnohem systematičtěji. Prozkoumal přes 20 000 číselných údajů ve 20 různých souborech dat, jako byly například délky 335 řek, měrné tepelné kapacity 1389 chemických sloučenin, statistiky v americké baseballové lize, čísla uvedená v článcích na titulních stránkách novin apod. Řada autorů kriticky nahlížela na Benfordův výběr zkoumaných souborů dat s poukazem na účelovost tohoto výběru s ohledem na očekávaný výsledek. Přes veškeré výhrady se však ukázalo, že platnost zákona první číslice je velmi častý jev, a i proto se v současnosti používá pro tuto zákonitost pojmenování Benfordův zákon.

Je zajímavé, že zřejmě první významnější zmínkou o tomto zákoně v češtině je krátký článek P.Kantorka z roku 1998 v časopise Vesmír [5]. P.Kantorek je známý český

karikaturista, který vystudoval fyziku na Masarykově univerzitě, v roce 1968 odešel do Kanady a v současné době je profesorem fyziky na Univerzitě v Torontu. Ve svém článku konstatuje *„Nejde o žádný matematický trik, ale o skutečný přírodní zákon, jímž se řídí soubory jakýchkoliv přirozených dat, bez ohledu na jejich podstatu nebo fyzikální jednotky. Jedinou podmínkou je, že data musí být v minimálním rozsahu tří logaritmických intervalů.“*

Především v posledních dvaceti letech se v zahraničních časopisech objevilo mnoho příspěvků různé odborné kvality, které se věnují tomuto „first digit phenomenon“. V současnosti webové stránky *Benford Online Bibliography* [2] obsahují více než 600 odkazů na tyto příspěvky. I když autoři tohoto článku nemohli projít všechny tyto odkazy, je zřejmé, že okolo Bernfordova zákona je stále mnoho nedořešených otázek, jak z hlediska čistě matematického, tak především z hlediska možností jeho praktického využití v různých vědních oborech a oblastech lidské činnosti

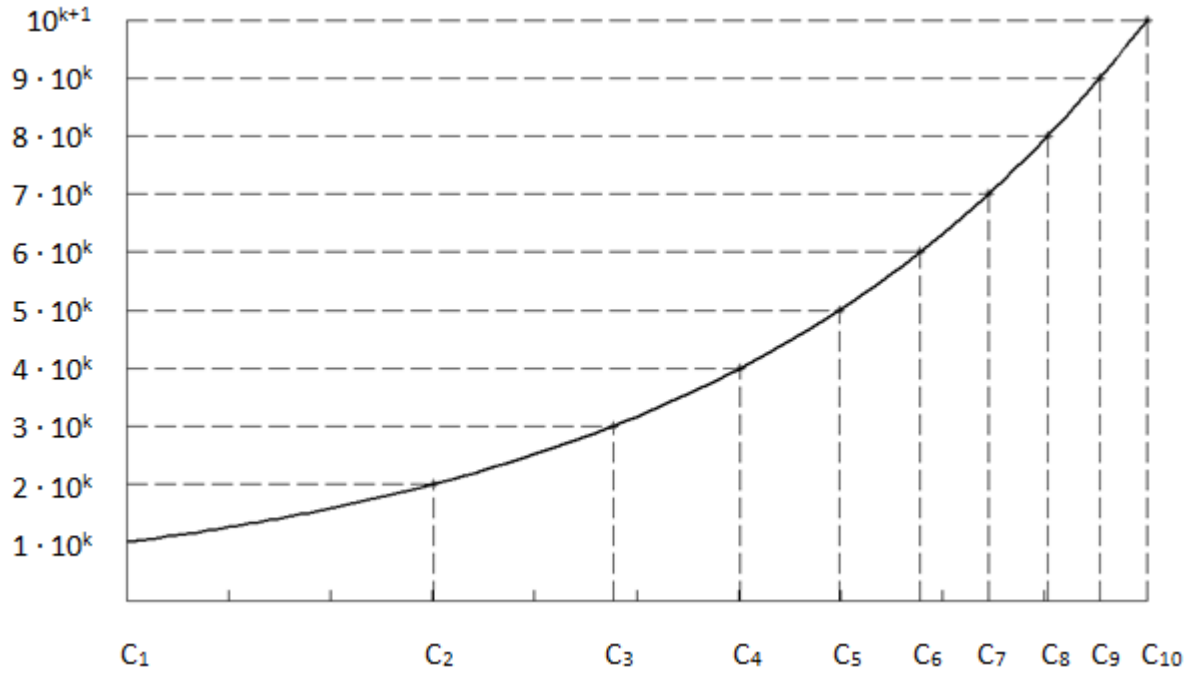
Matematické zdůvodnění a konkretizace zákona

Obraťme nyní pozornost k matematické stránce zákona. Je naprosto zřejmé, že zákon platit nebude v případě souborů dat, kde jednotlivá čísla jsou vytvářena náhodným výběrem číslic na jednotlivých pozicích, přičemž tyto číslice jsou vybírány se stejnou pravděpodobností, což představuje rovnoměrné rozložení pravděpodobností. Takovéto soubory dat nás proto zajímat nemohou,

Pro soubory dat vyjadřujících hodnotu jisté přirozeně vymezené veličiny z reálného světa lze najít v literatuře celou řadu různých důkazů, zdůvodnění či odvození zákona první číslice. Z pochopitelných důvodů většina z nich je založena na metodách statistických a poznatcích teorie pravděpodobnosti. Zájemcům můžeme doporučit například články T.P.Hilla [4], R.A.Raimiho [9], D.Cohena [3] nebo R.S.Pinkhama [7]. Pro naše potřeby však postačí následující odvození, jehož základní ideu jsme převzali z článku K.A.Rosse [10], otištěného v časopise *The American Mathematical Monthly*.

Uvažujme posloupnost danou funkčním předpisem $a_n = A\alpha^n$. Jako jednoduchý model pro tuto posloupnost můžeme použít peněžní hodnotu majetku firmy nebo nějaké výrobní či finanční společnosti, kde celočíselná proměnná n vyjadřuje čas v určitých jednotkách, například v letech. Číslo A udává peněžní hodnotu ve výchozím roce, kdy $n = 0$. Daný předpis předpokládá, že hodnota se mění exponenciálně, přičemž pro $\alpha > 1$ roste v čase a naopak při $0 < \alpha < 1$ klesá. Dále budeme pro zjednodušení předpokládat, že hodnota s časem roste. Protože vynásobením hodnot a_n mocninou čísla 10 se první číslice nemění, můžeme také předpokládat, že počáteční majetek $1 \le A < 10$ a základ exponenciální funkce vyhovuje podmínce $1 < \alpha < 10$. V obr. 1 je zobrazený průběh rostoucí exponenciální funkce $f(t) = A\alpha^t$ mezi funkčními hodnotami 10^k a 10^{k+1}. Z něho je jasně vidět, že čas, který odpovídá tomu, kdy první číslicí funkční hodnoty je 1 určuje interval $\langle c_1, c_2)$, první číslicí je 2 určuje interval $\langle c_2, c_3)$, atd. Proto 1 je tedy první číslicí mnohem déle, než když první číslicí je 9.

Obr. 1: Graf růstu majetku firmy



Nyní můžeme dokončit odůvodnění funkčního vyjádření zákona první číslice. Uvažujme libovolnou nenulovou číslici d, pak její umístění jako první číslice odpovídá intervalu $\langle c_d, c_{d+1})$. Chceme určit délku tohoto intervalu v závislosti na hodnotě d. Podle obr. 1 platí zřejmě

$$f(c_d) = A\alpha^{c_d} = d \cdot 10^k$$

$$f(c_{d+1}) = A\alpha^{c_{d+1}} = (d+1) \cdot 10^k$$

a po zlogaritmování (při základu 10)

$$\log A + c_d \log \alpha = \log d + k$$

$$\log A + c_{d+1} \log \alpha = \log(d+1) + k\,.$$

Odečtením posledních rovností a jejich jednoduchou úpravou dostáváme

$$\log(d+1) - \log d = (c_{d+1} - c_d) \cdot \log \alpha$$

neboli

$$c_{d+1} - c_d = \frac{1}{\log \alpha}\left[\log(d+1) - \log d\right],$$

kde $\frac{1}{\log\alpha}$ je konstanta nezávislá na k. Navíc čas, kdy křivka roste od hodnoty 10^k do je přesně

$$c_{10} - c_1 = \frac{1}{\log\alpha}\left[\log 10 - \log 1\right] = \frac{1}{\log\alpha}.$$

Jinými slovy, v tomto „jednotkovém" časovém rozpětí, je poměrný časový úsek, že první číslice hodnot $A\alpha^t$ je právě d, roven

$$P(d) = \log(d+1) - \log d = \log\frac{d+1}{d} = \log\left(1 + \frac{1}{d}\right).$$

Je zřejmé, že tato úvaha platí pro libovolnou číselnou hodnotu k.

Jestliže se vrátíme k modelové situaci s majetkem firmy, a tedy k diskrétním hodnotám $f(n) = A\alpha^n$ pro nezáporná celá čísla n, jsou příslušné hodnoty $[n, f(n)]$ rozmístěné podél křivky v obr. 1. Můžeme tedy očekávat, že poměrná část těchto bodů v intervalu $\langle c_d, c_{d+1})$ by měla být také určena hodnotou $\log\frac{d+1}{d}$.

Získali jsme tedy funkční vyjádření zákona první číslice neboli Benfordova zákona, v běžně používaném tvaru. Příslušné pravděpodobnosti pro jednotlivé číslice jsou v tab 2. Náhodné veličiny, jejichž rozdělení odpovídá uvedeným hodnotám, se někdy označují jako veličiny podléhající Benfordovu rozdělení.

Tab. 2: Pravděpodobnosti podle Benfordova zákona

d	1	2	3	4	5	6	7	8	9
$\log\frac{d+1}{d}$	0,3010	0,1761	0,1249	0,0969	0,0792	0,0670	0,0580	0,0512	0,0458

Ověření platnosti zákona na vybraných souborech

V zahraniční literatuře lze najít ověřování zákona první číslice na mnoha datových souborech přírodního (přirozeného) charakteru. Předchozí odvození ukazuje, že zákon jistě platí, jestliže se jedná o hodnoty náhodné veličiny, které se mění podle jistého exponenciálního růstu nebo klesání. Řada autorů se pokoušela stanovit přesnější charakteristiku číselných souborů, pro které zákon bude platit.

Jestliže bychom shrnuli alespoň nejdůležitější požadavky na dané soubory, můžeme je formulovat v těchto čtyřech bodech:

1. Číselná data musí být dána ve stejných měrných jednotkách (počet obyvatel, délková či plošná míra, peněžní měna apod.).
2. Dat musí být dostatečné množství, většinou se požaduje řádově tisíce dat, jako minimální se uvádí 500 hodnot.

3. Číselné hodnoty by měly být v rozpětí alespoň tří (logaritmických) řádů.
4. Číselné hodnoty by neměly být v příslušném rozpětí zdola ani shora nijak omezeny.

Na druhou stranu se podařilo například dokázat, že nezáleží na základu číselné soustavy, v níž jsou údaje vyjádřeny.

V české literatuře se objevilo jen několik seriózních pokusů ověřovat nebo přímo aplikovat v praxi Benfordův zákon [8], Rozhodli jsme se proto vyzkoušet platnost zákona na dvou datově odlišných souborech, které jsou veřejně přístupné na internetu. V mnoha zahraničních odkazech je citováno ověřování zákona na určitých demografických souborech, které se často konkrétně týkají počtu obyvatel obcí, měst nebo států. Využili jsme proto toho, že v roce 2011 proběhlo v České republice sčítání lidu a analyzovali jsme soubor tvořený počty obyvatel všech 602 měst v naší republice [11]. Pro současný článek jsme aktualizovali toto šetření a použili jsme datový soubor počtu obyvatel všech 6253 obcí v ČR ke dni 31.12.2013. Zdrojem byl materiál *ČSÚ a územně analytické podklady* [13]. Výsledky jsou uvedeny v tab.3, příslušné relativní četnosti vykazují ještě přesnější shodu s hodnotami získanými z Benfordova zákona, než v případě počtu obyvatel měst. Domníváme se, že lepší shoda je důsledkem zvětšení rozsahu zkoumaného souboru o jeden řád. Tento soubor splňuje všechny čtyři výše připomenuté podmínky na posuzovaný soubor dat. Největší počet obyvatel má samozřejmě hlavní město Praha, konkrétně 1 243 201, pouze 2 obyvatele má Březina, okres Vyškov, v tomto případě se jedná o vojenský újezd.

Jako druhý zkoumaný soubor jsme použili data ze zcela odlišné oblasti. *Státní ústav pro kontrolu léčiv* v souladu se zákonem zveřejňuje *Seznam cen a úhrad léčivých přípravků a potravin pro zvláštní lékařské účely hrazených ze zdravotního pojištění*. Seznam je každý měsíc aktualizován a uvádí úplný výčet léčivých přípravků a potravin pro zvláštní lékařské účely, o jejichž úhradě z veřejného zdravotního pojištění ústav rozhoduje [12]. Ceny (MFC) v tomto seznamu uvedené ke dni 5.4.2015 se pohybovaly přibližně v rozpětí od 10 do 100 000 Kč a celkový počet položek byl 8614. Příslušné absolutní a relativní četnosti jsou uvedeny v posledních dvou řádcích tab. 3. Z údajů v tabulce je zřejmé, že soubor cen v ceníku také velice dobře odpovídá Benfordovu rozdělení pro první číslice.

Obr. 2 zcela jasně dokumentuje vizuálně míru shody relativních četností výskytu jednotlivých číslic 1 až 9 jako prvních číslic položek v analyzovaných souborech s předpokládanou pravděpodobností výskytu podle Benfordova zákona. Při detailním statistickém zpracování analýz datových souborů se ovšem zpravidla postupuje tak, že se stanoví jako nulová hypotéza shoda s Benfordovým rozložením a vhodným kritériem se hypotéza testuje. Nejčastěji se používá test dobré shody χ-kvadrát.

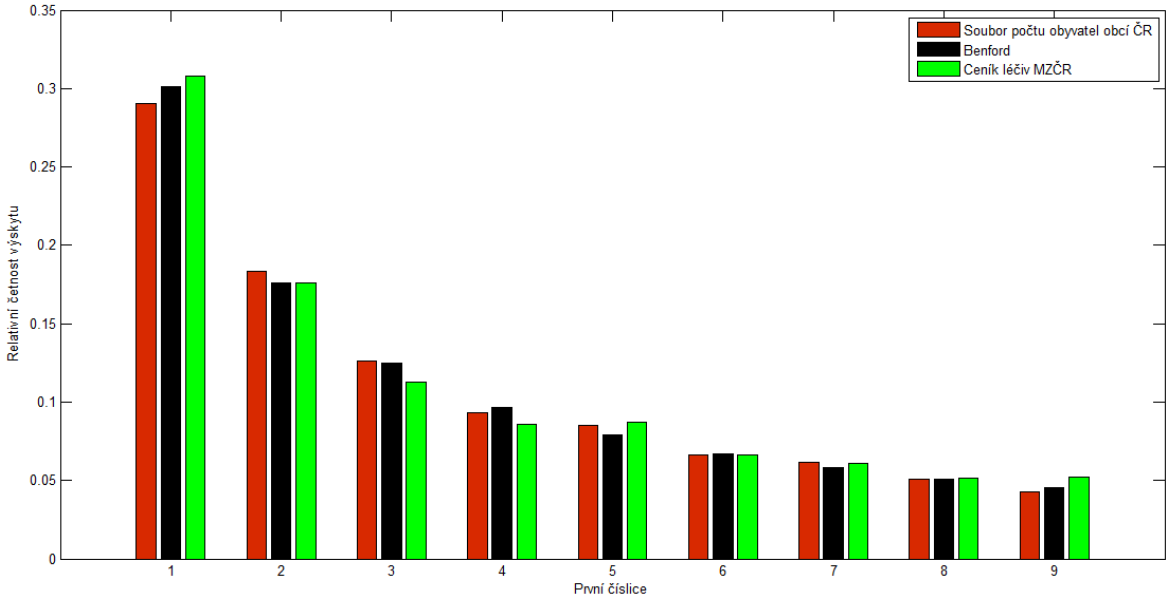
Tab.3: Absolutní a relativní četnosti (v %) výskytu první číslice d v daných souborech: A – počet obyvatel obcí ČR; B - ceník léčiv

d	1	2	3	4	5	6	7	8	9
A abs.	1815	1147	790	583	531	414	387	319	267
A rel.	29,03	18,34	12,63	9,32	8,49	6,62	6,29	5,10	4,27
B abs.	2655	1514	970	742	749	569	522	446	447
B rel.	30,82	17,58	11,26	8,61	8,70	6,61	6,06	5,18	5,19

Pro případnou analýzu uvedeného typu lze najít na internetu řadu datových souborů, které je možné takto zkoumat. Vhodným zdrojem jsou webové stránky Českého statistického úřadu nebo Eurostatu. Největším problémem je především zajistit dostatečný rozsah souboru a

dostatečné rozpětí dat v souboru. Domníváme se ale, že by takováto aktivita mohla žáky zaujmout, protože většina z nich s internetovými vyhledávači běžně pracuje.

Obr. 2: Srovnání relativních četností v uvedených souborech



Závěrečné shrnutí

Jak už bylo dříve uvedeno, obsahuje on-line databáze [2] více než 600 titulů pojednávajících o Benfordově zákoně a jeho možném využití v různých oborech. Je zřejmé, že se tato databáze bude dále rozrůstat. Pro matematiky je zde však celá řada nedořešených otázek. Několik z nich zformuloval K.A. Ross ve svém článku [10], který považujeme za stěžejní pro matematické chápání zákona.

V posledních letech se objevila řada studií, které se pokoušely aplikovat zákon první číslice při analýze datových souborů v ekonomické a finanční oblasti, kde se bezesporu jeví uplatnění tohoto zákona za nejefektivnější. Snažili jsme se proto zmapovat některé významnější pokusy o aplikaci Benfordova zákona v ekonomických vědách [11], včetně konkrétních výsledků hlavně v zahraničí. Cílem takových analýz bylo posoudit, zda číselné údaje v jednotlivých finančních výkazech a jiných ekonomických statistikách odpovídají Benfordovu rozdělení.

Z hlediska možného použití zákona první číslice můžeme také připomenout jednu situaci z politologie. Opakovaně se objevily pokusy detekovat s pomocí tohoto zákona falšování výsledku voleb v některých zemích, v případě zájmu může čtenář najít na internetu příslušné odkazy. Těžko lze posoudit, zda se jedná o skutečně detailní a fundovanou analýzu volebních výsledků. Zhodnocení je nutné nechat na odborníky v této oblasti.

Pro účinnější využití Benfordova zákona v praxi, je důležité rozšíření zákona na další platné číslice dat v souborech. Bez těchto úvah nejsou možnosti efektivního uplatnění zákona při formulování konkrétních závěrů příslušných analýz datových souborů plně fundované. Abychom lépe pochopili matematické okolnosti popisované zákonitosti, bylo by také třeba detailněji přihlédnout k odlišnostem v přístupu jednotlivých odborníků na problematiku zákona první číslice. To vše už je nad reálné možnosti tohoto příspěvku, který považujeme jen za základní vhled do „tajemství“ v našich zemích dosud málo připomínaného fenoménu.

Jako závěrečný výrok, který v jistém zjednodušení vyjadřuje cíl tohoto příspěvku, můžeme použít jemně upravenou citaci z článku P.Kantorka. *„Aplikace Benfordova zákona je dalekosáhlá. Máme-li větší soubor jakýchkoliv dat, můžeme poměrně jednoduchým statistickým rozborem lehce zjistit, jsou-li data skutečná (přírodní), nebo podezřelá. Pochopitelně ale platí základní pravidlo statistiky: Čím více dat, tím lepší souhlas s teoretickou křivkou. Zákon se bohužel nedá použít k zlepšení šance výhry ve Sportce, protože nahodile tažená čísla 1 až 9 mohou být seřazena v libovolném pořadí. V každém případě je to však zákon fascinující a je zajímavé, jak málo i renomovaných akademiků o něm ví.“*

Literatura

[1] *Benford, F.*: The law of anomalous numbers. Proc. Amer. Philos. Soc., vol. 78, 1938, 551-572.

[2] *Berger, A - Hill, T. P.:* Benford Online Bibliography. 2015. [cit. 2015-04-14]. Dostupné z http://www.benfordonline.net.

[3] *Cohen, D., I., A.:* An explanation of the first digit phenomenon. Journal of Combinatorial Theory, vol. 20, 1976, 367-370.

[4] *Hill, T. P.:* A statistical derivation of the significant.digit law. Statistical Science, vol. 10, no. 4., 1996, 354-365.

[5] *Kantorek, P.*: Benfordův zákon. Vesmír, roč. 77, 1998, s. 583.

[6] *Newcombe, S.:* Note on the frequency of use of the different digits in natural numbers. Americal Journal of Mathematics, vol.4, 1881, 39-40.

[7] *Pinkham, R., S.:* On the distribution of first significant digits. Annals of Mathematical Statistics, vol. 32, 1961. 1223-1230.

[8] *Plaček, M.:* Benfordův zákon: fakta a mýty. Bulletin komory certifikovaných účetních. 1/2013, Praha, 2013, 43-46.

[9] *Raimi, R., A.:* The first digit problem. The American Mathematical Monthly, vol. 83, 1976. 521-538.

[10] *Ross, K., A.:* Benford's law, a growth industry. The American Mathematical Monthly, . vol. 118, no. 8, 2011, 571-583..

[11] *Seibert, J - Zahrádka, J.:* Zákon první číslice a jeho aplikace, Scientific papers of the University of Pardubice. Series D, Faculty of Economics and Administration, roč.30, 1/2014, 75-83.

[12] Seznam cen a úhrad LP/PZLÚ k 2. 4. 2015. [cit. 2015-04-05]. Dostupné z http://www.sukl.cz/sukl/oprava-predchoziho-seznamu-cen-a-uhrad-lp-pzlu.

[13] ČSÚ a územně analytické podklady. [cit. 2015-04-05]. Dostupné z https://www.czso.cz/csu/czso/csu_a_uzemne_analyticke_podklady.

Kontaktní adresa

RNDr. Jaromír Zahrádka, Ph.D.
Univerzita Pardubice

Fakulta ekonomicko-správní
Ústav matematiky a kvantitativních metod
Studentská 95, 532 10 Pardubice, Česká republika
jaromir.zahradka@upce.cz
+420466036047

Doc.RNDr.Jaroslav Seibert,CSc.
Univerzita Pardubice
Fakulta ekonomicko-správní
Ústav matematiky a kvantitativních metod
Studentská 95, 532 10 Pardubice, Česká republika
jaroslav.seibert@upce.cz
+420466036016

Abstract: The paper deals with the Benford law which is also called as the first digit law. The history, empirical evidence and a simple explanation of its validity are reviewed. This law showed that the data in the real world have the property that the first digit 1 appears in 100 numbers about 30 times, the first digit 2 about 17 times, and so on the first digit 9 about 5 times. This law states that data sets from different fields leading digits tend to be distributed logarithmically. The Benford's distribution is verified for the set consists of the numbers of inhabitants of all towns in Czech Republic and for the set of prices of the medicinal drugs.