Pregledni rad
Primljeno: 30. rujna 2017.
343.98:657

Milan Papić[*]
Nenad Vudrić[**]
Krešimir Jerin[***]

BENFORDOV ZAKON I NJEGOVA PRIMJENA U FORENZIČKOM RAČUNOVODSTVU

Sažetak

Benfordov zakon, poznat i kao zakon prve znamenke, izjava je o distribuciji frekventnosti vodećih znamenaka u mnogim kvantitativnim skupovima podataka iz stvarnog života. Intuitivno očekivanje je da distribucija vodećih znamenaka bude uniformna (odnosno da za svaku znamenku bude podjednaka vjerojatnost da se nađe na prvom mjestu), ali Benfordov zakon tvrdi kako vodeće znamenke prate logaritamsku distribuciju. U ovom radu dan je dokaz Benfordovog zakona te je prikazana njegova primjena u forenzičkom računovodstvu, jednom od mnogih područja njegove moguće primjene.

Ključne riječi: Benfordov zakon, forenzičko računovodstvo, revizija, zakon prve znamenke

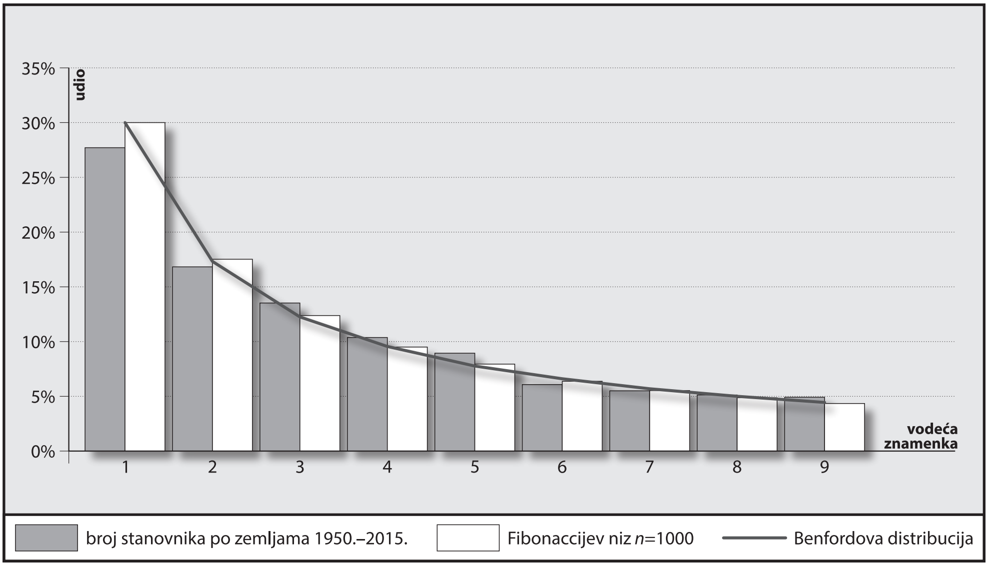
1. Uvod

Zamislimo scenarij u kojem posjedujemo skup od 10 000 brojčanih podataka te nas zanima koliko brojeva u tom skupu brojeva započinje s brojem 1. Intuitivno, ako prva znamenka može biti bilo koji broj između 1 i 9, za očekivati je da je odgovor oko 11 %, odnosno oko 1 100. Benfordov zakon[1] tvrdi da u prirodnim skupovima brojčanih podataka distribucija vodećih znamenaka nije uniformna, već je logaritamska sa znamenkom 1 na prvom mjestu u oko 30 %. Dapače, za svaku znamenku, ili skup znamenaka, postoji određena vjerojatnost pojavljivanja gdje veću vjerojatnost imaju niže znamenke.

[*] mr. sc. Milan Papić, Libertas međunarodno sveučilište, mpapic@libertas.hr
[**] Nenad Vudrić, Libertas međunarodno sveučilište, nvudric@yahoo.com
[***] Krešimir Jerin, bacc. oec., kresimir.jerin@gmail.com

[1] Prva osoba koja je primijetila da znamenke brojeva prate logaritamsku distribuciju je Simon Newcomb 1881. godine, no njegovom objavljenom članku nije posvećena velika pažnja. Pola stoljeća kasnije Frank Benford provodi poznato istraživanje nad bazama podataka od 20 229 „prirodnih" promatranja kojim je pokazana spomenuta logaritamska distribucija, koja time postaje poznata kao Benfordov zakon. Rezultati tog istraživanja su prikazani u poglavlju 2, tablica 1.

Grafikon 1: Benfordova distribucija te distribucije vodećih znamenaka Fibonaccijevog niza i broja stanovnika po zemljama



Izvor: https://data.worldbank.org/indicator/SP.POP.TOTL, 12. rujna 2017.

Informatizacijom poslovanja dolazi do popularizacije Benfordovog zakona te njegove sve veće primjene u raznim područjima. Jedno od glavnih svojstava Benfordove distribucije, koje je omogućilo njegovu primjenu u raznim područjima diljem svijeta, jest invarijantnost skalara, što znači da mjerne jedinice iskazanih brojčanih podataka ne utječu na Benfordov zakon. U ovom se radu proučavaju metode primjene Benfordovog zakona u forenzičkom računovodstvu te se proučavaju prednosti i nedostaci njegove primjene naprimjeru poduzeća prisutnih u Republici Hrvatskoj.

2. Benfordov zakon

Radi razumijevanja Benfordovog zakona potrebno je definirati pojam vodeće znamenke. *k* vodećih znamenki realnog broja *x* je definirano kao:

$$\lfloor |d| \rfloor, \text{ gdje je } x = d \cdot 10^i,\ 10^{k-1} \le d < 10^k,\ k \in N,\ i \in Z \setminus \{0\},\ \lfloor . \rfloor = \text{ najveće cijelo.}$$

Na primjer, 3 vodeće znamenke broja π su 314, a 2 vodeće znamenke broja −0,251 su 25.

Benfordov zakon tvrdi da u mnogim skupovima podataka (matematičke tablice, podaci iz stvarnog života kao i njihove kombinacije) vodeće znamenke nisu uniformno raspoređene već prate određenu logaritamsku distribuciju. Rezultati jednog od njegovih prvih istraživanja su dani u tablici 1.

Tablica 1. Benfordova analiza provedena 1938. godine s nazivom uzorka, učestalošću pojavljivanja prvih znamenaka i opsegom uzorka

Naziv	1	2	3	4	5	6	7	8	9	#N
Rijeke, površina	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
Stanovništvo	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
Konstante	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
Novine	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
Specifična toplina	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
Tlak	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
H. P. gubitak	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
Molekularna težina	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
Isušivanje	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
Atomska težina	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
$n-1$, $\sqrt{n}$	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5000
Dizajn	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
Reader's Digest	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
Cijene	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
Rendgenska voltaža	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
Statistika baseballa	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1458
Vodljivost	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
Adrese	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
$n^1, n^2, \ldots n!$	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
Stopa smrtnosti	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
Prosjek	**30.6**	**18.5**	**12.4**	**9.4**	**8.0**	**6.4**	**5.1**	**4.9**	**4.7**	
Vjerojatna pogreška	± 0.8	± 0.4	± 0.4	± 0.3	± 0.2	± 0.2	± 0.2	± 0.2	± 0.3	

Izvor: Nigrini (2012: 4).

Iz tablice 1 se vidi da, suprotno intuitivnom očekivanju, prve vodeće znamenke nisu uniformno raspoređene, nego vjerojatnost pojavljivanja većih prvih vodećih znamenaka opada. Pogotovo je zanimljivo što je to svojstvo primijećeno kod svih uzoraka koji su potpuno međusobno neovisni.

2.1. Benfordov zakon o vjerojatnosti vodeće znamenke

Benfordov zakon tvrdi da u mnogim skupovima podataka vodeće znamenke prate određenu logaritamsku distribuciju. Najpoznatija verzija Benfordovog zakona je vezana za distribuciju prvih vodećih znamenki, koja tvrdi (Berger i Hill, 2011):

$$P(D) = \log\left(1 + \frac{1}{D}\right),\ D \in \{1, 2, 3, \ldots, 9\} \tag{1}$$

Na primjer, vjerojatnost pojavljivanja broja 3 kao vodeće znamenke je:

$$P(3) = \log\left(1 + \frac{1}{3}\right) = 0{,}1249 = 12{,}49\%$$

Dostupno je više dokaza za izraz (1), no većini je zajednička primjena svojstava funkcije gustoće te invarijantnosti skalara. Definicija funkcije gustoće je sljedeća (Benšić i Šuvak, 2014: 61):

Slučajna varijabla X na vjerojatnosnom prostoru (Ω, *F*, *P*) je neprekidna slučajna varijabla ako postoji funkcija $f: \Re \to \left[0, +\infty\right\rangle$ takva da je:

$$P\left(X \leq \alpha\right) = \int_{-\infty}^{\alpha} f(x)dx, \forall \alpha \in \Re$$

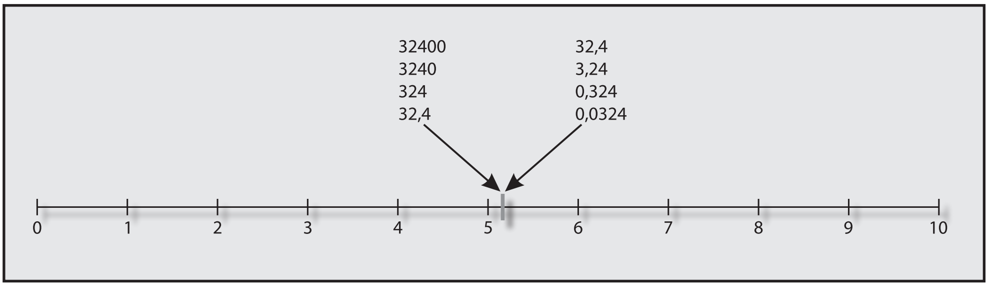
Funkcija f naziva se funkcija gustoće slučajne varijable X. Bitna svojstva funkcije gustoće neprekidne slučajne varijable:

- nenegativnost: $f(x) \geq 0,\ \forall \alpha \in \Re$
- normiranost: $\int_{-\infty}^{+\infty} f(x)dx = 1$

Prvi korak dokaza je definiranje neprekidne slučajne varijable X, koja će u našem slučaju biti mantisa logaritma.[2] Svaki realan broj ima jedinstven zapis u obliku znanstvene notacije $a \cdot 10^k$, $a \in \left[1{,}10\right\rangle$, $k \in Z$. Mantisa logaritma ne ovisi o vrijednosti 10^k nego samo o a jer $\log(a \cdot 10^k) = \log(a) + k$.

X je definiran kao $\Re^+ \to \left[1{,}10\right\rangle$ gdje realni brojevi poprimaju vrijednosti mantisa (slika 1).

Slika 1: Preslikavanje realnih brojeva u mantisu logaritma



Vjerojatnost *P*(*D*) mora odgovarati površini ispod funkcije gustoće na intervalu $\left[D, D+1\right\rangle$

$$P(D) = \int_{D}^{D+1} f(x)dx$$

2 Mantisa logaritma jest decimalni dio logaritma, npr. log 2 = 0,30103… pa je mantisa 30103…

Koristi se svojstvo invarijantnosti na skalar, odnosno skalar β ne utječe na iznos vjerojatnosti

$$P\left(D\right)=\int_{\beta D}^{\beta\left(D+1\right)}f(x)dx$$

Deriviranjem se dobiva funkcijska jednadžba

$$\frac{\partial P\left(D\right)}{\partial\beta}=0$$
$$\left(D+1\right)f\left(\beta\left(D+1\right)\right)-Df\left(\beta D\right)=0$$

Supstitucija D s 1, 2, … se redom dobivaju jednadžbe

$$2f\left(2\beta\right)-f\left(\beta\right)=0$$
$$3f\left(3\beta\right)-2f\left(2\beta\right)=0$$
$$\ldots$$
$$f\left(\beta\right)=cf\left(c\beta\right)$$

Supstitucijom $c\to\frac{1}{\beta}$ dolazi se do izraza

$$f\left(\beta\right)=\frac{f\left(1\right)}{\beta}$$

odnosno funkcija gustoće je oblika

$$f\left(x\right)=\frac{c}{x}$$

Parametar c se dobiva iz uvjeta

$$\int_{1}^{10}f(x)dx=1$$
$$c\left(\ln 10-\ln 1\right)=1$$
$$c=\frac{1}{\ln 10}$$

Te (kako je očito $f(x)\geq 0$ za svaki $x>0$) funkcija gustoće glasi:

$$f(x)=\frac{1}{x\ln(10)} \tag{2}$$

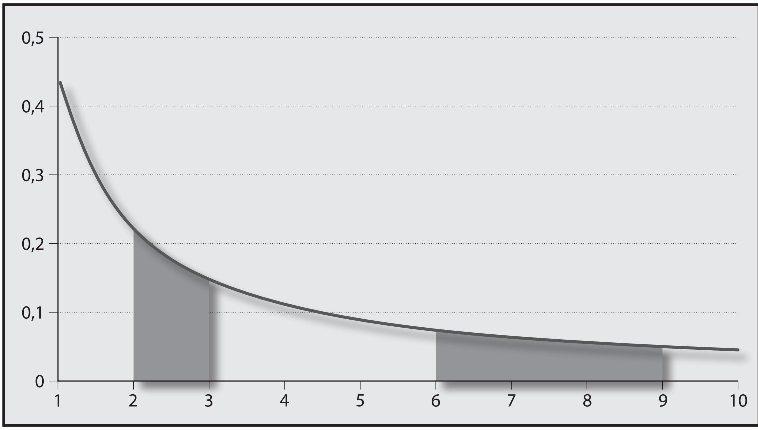
Uvrštavanjem dobivene funkcije gustoće u početnu jednadžbu dobivamo:

$$P(D)=\int_{D}^{D+1}\frac{1}{x\ln 10}dx=\log(D+1)-\log(D)=\log\left(1+\frac{1}{D}\right)$$

Q.E.D.

Svojstva dobivene funkcije gustoće (2) kao i invarijantnost skalara su prikazane na grafikonu 2. Funkcija gustoće je negativne na području $\langle -\infty, 0\rangle$, ali kako je domena $\left[1,10\right\rangle$ svojstvo nenegativnosti je zadovoljeno. Svojstvo normiranost je pokazano zbrojem vjerojatnosti svih mogućih slučajeva gdje je suma 1. Kao primjer invarijantnosti skalara je uzeto područje domene $\left[2,3\right\rangle$ (odnosno svi realni brojevi kojima je prva vodeća znamenka 2) i skalar $\beta=3$, odnosno $P(2)=P(6)+P(7)+P(8)$.

Grafikon 2: Osnovna svojstva funkcije gustoće te prikaz invarijantnosti skalara



d_i	$P(d_i)$
1	0,3010
2	0,1761
3	0,1249
4	0,0969
5	0,0792
6	0,0669
7	0,0580
8	0,0512
9	0,0458
∑	1

2.2. Benfordov zakon o vjerojatnosti *k* vodećih znamenaka

Benfordov zakon se može proširiti na prvih k vodećih znamenki koristeći već dobivenu funkciju gustoće (2). Neka je $k \in N$ tada je distribucija prvih k vodećih znamenki d_1, d_2, ..., d_k jednaka (Jamain, 2001: 14):

$$P(d_1,d_2,\ldots,d_k)=\log\left(1+\frac{1}{\left(\sum_{i=1}^{k}10^{k-i}d_i\right)}\right) \tag{3}$$

gdje je $d_1 \in \{1, 2, ..., 9\}$, a svi ostali $d_1 \in \{0, 1, 2, ..., 9\}$.

Najjednostavniji dokaz izraza (3) je korištenje funkcije gustoće (2). Definira se preslikavanje $M_k : N^k \rightarrow \left[1,10\right\rangle$ tako da:

$$(d_1,d_2,\ldots,d_k)\rightarrow\sum_{i=1}^{k}10^{-(i-1)}d_i$$

gdje je $k \in N$, $d_1 \in \{1, 2, ..., 9\}$, $d_{2, 3, ..., k} \in \{0, 1, 2, ..., 9\}$. Drugim riječima, skup k prirodnih brojeva se preslikava u realan broj oblika $d_1, d_2, d_3, ..., d_k$.

$$P\left(d_1, d_2, \ldots, d_k\right) = P\left(\sum_{i=1}^{k} 10^{-(i-1)} d_i \le M_k < \sum_{i=1}^{k-1} 10^{-(i-1)} d_i + \left(1 + d_k\right) 10^{-k+1}\right)$$

$$= \int_{\sum_{i=1}^{k} 10^{-(i-1)} d_i}^{\sum_{i=1}^{k-1} 10^{-(i-1)} d_i + (1+d_k) 10^{-k+1}} \frac{1}{x \ln 10} dx$$

$$= \log\left(\sum_{i=1}^{k-1} 10^{-(i-1)} d_i + \left(1 + d_k\right) 10^{-k+1}\right) - \log\left(\sum_{i=1}^{k} 10^{-(i-1)} d_i\right)$$

$$= \log\left(\frac{d_1 + d_2 10^{-1} + d_3 10^{-3} + \ldots + \left(1 + d_k\right) 10^{-k+1}}{d_1 + d_2 10^{-1} + d_3 10^{-3} + \ldots + d_k 10^{-k+1}}\right)$$

$$= \log\left(1 + \frac{1}{\left(\sum_{i=1}^{k} 10^{k-i} d_i\right)}\right) \qquad \text{Q.E.D.}$$

Na primjer, vjerojatnost da u nekom skupu podataka vodeće 3 znamenke budu 324 iznosi:

$$P\left(3, 2, 4\right) = \log\left(1 + \frac{1}{10^2 \cdot 3 + 10^1 \cdot 2 + 10^0 \cdot 4}\right) = \log\left(1 + \frac{1}{324}\right) = 0{,}13\%$$

Pomoću izraza (3) se može izračunati vjerojatnost pojavljivanja bilo kojih znamenki, a ne samo prvih k. Na primjer, vjerojatnost da je druga znamenka broj 2 iznosi:

$$\sum_{i=1}^{9} P\left(i, 2\right) = \sum_{i=1}^{9} \log\left(1 + \frac{1}{10i + 2}\right) = 10{,}88\%$$

Benfordov zakon se može proširiti i na druge baze preko formule gustoće

$$f(x) = \frac{1}{y \log b}$$

gdje je baza $b > 1$. Izraz za vjerojatnost prve vodeće znamenke u nekoj bazi onda glasi (http://www.math.uah.edu/stat/special/Benford.html, The University of Alabama in Huntsville, Mathemacial Sciences, 27. kolovoza 2017):

$$P_b\left(d\right) = \log_b\left(1 + \frac{1}{d}\right) \tag{4}$$

a za prvih k vodećih znamenki

$$P_b\left(d_1, d_2, \ldots, d_k\right) = \log_b\left(1 + \frac{1}{\left(\sum_{i=1}^{k} b^{k-1} d_i\right)}\right) \tag{5}$$

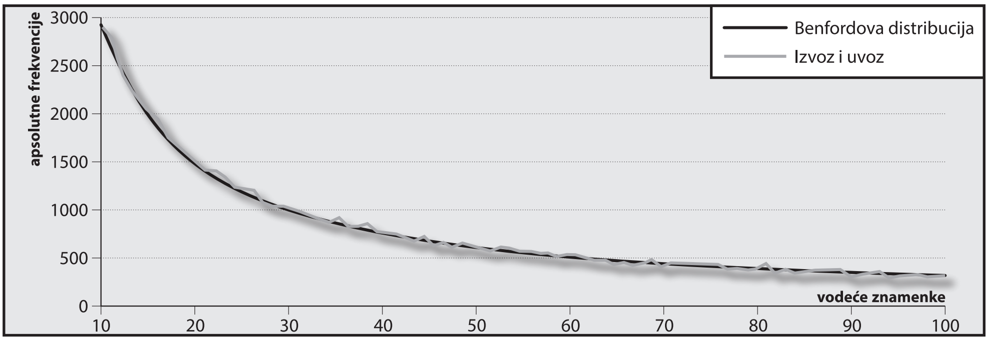
U sljedećim poglavljima fokus je na primjeni Benfordovog zakona, a kako je dekadski sustav daleko najkorišteniji, izrazi 4 i 5 neće biti primjenjivani u ostatku ovog rada.

3. Analiza odstupanja od Benfordove distribucije

Distribucija znamenaka neke distribucije stvarnih podataka neće nikada u potpunosti biti jednaka Benfordovoj distribuciji. Kako stvarne baze podataka uvijek imaju konačno mnogo podataka u sebi, izračun udjela pojavljivanja znamenaka rezultira racionalnim brojem, dok izrazi 2 i 3 za izračun Benfordove distribucije rezultiraju iracionalnim brojevima. U ovom poglavlju obrađena su tri[3] najčešća načina mjerenja odstupanja neke distribucije znamenaka od Benfordove distribucije. Ako odstupanje bude signifikantno, to je znak da se u skupu podataka javljaju abnormalna ponavljanja znamenaka i/ili neke druge anomalije.

Pokazni primjer na kojem će biti izračunata odstupanja distribucija znamenaka je izvoz i uvoz Republike Hrvatske u razdoblju od 2010. do 2016. godine, prvenstveno zbog lake dostupnosti, točnosti i veličine spomenute baze podataka. Promatrana je distribucija prvih dviju znamenaka, odnosno $P(d_1, d_2)$ gdje je $d_1 \in \{1, 2, ..., 9\}$ i $d_1 \in \{0, 1, ..., 9\}$ (grafikon 3).

Grafikon 3: Distribucija prve dvije vodeće znamenke izvoza i uvoza RH po poglavljima Carinske tarife / Kombinirane nomenklature i zemljama namjene / podrijetla od 2010. do 2016. godine u odnosu na Benfordovu distribuciju (N=70.835)



Izvor: https://www.dzs.hr/, 19. rujna 2017.

3 Sva tri testa su temeljena na promatranju pojavljivanja znamenaka, dok postoje i testovi bazirani na pretpostavci uniformne distribucije mantise logaritama.

3.1. Preduvjeti testiranja Benfordovog zakona

Prije analize distribucije znamenaka skupa podataka s Benfordovom distribucijom treba utvrditi da se za promatrani skup podataka zbilja očekuje da odgovara Benfordovoj distribuciji. Na primjer, skup podataka koji čine visine svih ljudi na Zemlji mjerenih u centimetrima neće odgovarati Benfordovoj distribuciji, iako bi u tom slučaju broj podataka bio slučajan (prirodan) i veći od 7 milijardi. Razlog je očit, visina gotovo svih ljudi započinje znamenkom 1 (visina između 100 cm i 199 cm), a gotovo nijedna osoba nema prvu znamenku 3, 4 ili 5 (npr. za prvu znamenku 4 osoba bi trebala biti visoka između 40 cm i 49 cm ili između 400 cm i 499 cm). U nastavku je popis „nematematičkih" smjernica koje trebaju biti zadovoljene da skup podataka prati Benfordovu distribuciju (Nigrini, 2011: 97).

- Podaci predstavljaju kvantificirane činjenice ili događaje.
- Ne smije postojati zadani raspon podataka, odnosno ne smije biti postavljen minimum i maksimum. Iznimka je 0, koja može biti minimum ako se radi o podacima koji poprimaju samo pozitivne vrijednosti.
- Podaci ne bi trebali predstavljati identifikacijske brojeve ili oznake, na primjer brojevi telefona, žiro-računi, registracije automobila, anketni odgovori u Likertovoj skali i slično. Drugim riječima, podaci ne bi trebali biti numeričke oznake događaja, osoba i stvari korištenih umjesto riječi.
- Distribucija podataka trebala bi biti pozitivno asimetrična, odnosno median bi trebao biti manji od aritmetičke sredine, što znači da prevladavaju niže vrijednosti. Nadalje, podaci ne bi trebali biti grupirani oko aritmetičke sredine. Za primjer možemo uzeti već spomenute visine ljudi koje su usko grupirane oko aritmetičke sredine.
- Preporučeni minimalni opseg statističkog skupa je 1000 podataka. Ako je manje podataka, analiza bi trebala dopustiti veća odstupanja od Benfordove distribucije.

3.2. Hi-kvadrat test

Hi-kvadrat je baziran isključivo na analizi apsolutnih frekvencija, a odgovora na pitanje koliko neke frekvencije dobivene istraživanjem odstupaju od očekivanih frekvencija (u našem slučaju Benfordove distribucije) (Papić, 2014: 235). Izračun je dan izrazom:

$$\text{hi-kvadrat} = \sum_{i=1}^{n} \frac{\left(f_0 - f_t\right)^2}{f_t}$$

gdje f_0 predstavlja opažene apsolutne frekvencije, f_t teoretske apsolutne frekvencije i n predstavlja broj razreda. Dobiveni rezultat izraza 6 se uspoređuje s tablicom graničnih vrijednosti hi-kvadrat testa (lako dostupne na internetu ili u statističkim knjigama, npr. https://www.medcalc.org/manual/chi-square-table.php, 19. rujna 2017) ovisno o željenoj vjerojatnosti i stupnju slobode k koji je jednak broju razreda $n-1$.

U primjeru spomenutom na početku poglavlja $n=90$ jer je broj razreda jednak mogućim kombinacijama prve dvije znamenke, što je kardinalni broj skupa {10, 11, 12, ..., 98, 99}. Stupanj slobode iznosi 89, a granična vrijednost za vjerojatnost od 5% u tom slučaju iznosi 112,02. Rezultat izraza 6 je 103,26 (što je ispod granične vrijednosti) stoga distribucija prvih dviju znamenaka izvoza i uvoza Republike Hrvatske u razdoblju od 2010. do 2016. godine značajno ne odstupa (uz rizik od 5%) od Benfordove distribucije.

3.3. Kolmogorov-Smirnovljev test

Kolmogorov-Smirnovljev test (dalje u tekstu K-S) baziran je na kumulativnoj funkciji gustoće, odnosno on uspoređuje funkcije gustoće dviju distribucija. Najveće opaženo odstupanje se uspoređuje s graničnom vrijednošću koje je (za velike uzorke) dano izrazom:

$$\text{K-S} = \frac{\sqrt{-\frac{1}{2}\ln\left(\frac{\alpha}{2}\right)}}{\sqrt{n}} \qquad (7)$$

gdje je α stupanj sigurnosti, a n veličina uzorka.

U primjeru izvoza i uvoza RH uz stupanj sigurnosti od 5% i $n=70.835$ izraz 7 rezultira graničnom vrijednošću 0,0051. Izračun kumulativnih vrijednosti funkcija gustoće promatranih dviju distribucija jest zbroj udjela pojavljivanja znamenaka, za Benfordovu distribuciju glasi:

$$\sum_{d\in S_k}\log\left(1+\frac{1}{d}\right) \qquad (8)$$

gdje je $S=\{10, 11, 12, ..., 99\}$, a S_k oznaka za prvih $k\in\{1, 2, ..., 90\}$ elemenata skupa S.

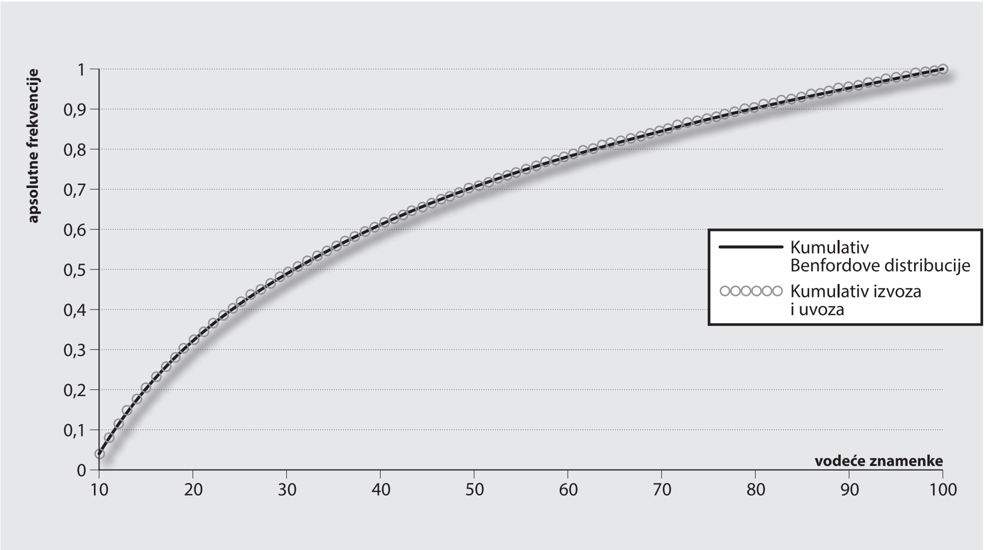
Za izračun kumulativnog zbroja relativnih frekvencija izvoza i uvoza RH se koristi izraz:

$$\sum_{d\in S_k}\frac{d}{n} \qquad (9)$$

U slučaju kad je $k=90$ (odnosno kad promatramo kumulativ udjela pojavljivanja svih znamenaka) izrazi 8 i 9 iznose 1 (što je i jedno od svojstava funkcije gustoće spoemutih u poglavlju 2). Najveća razlika izraza 8 i 9 iznosi 0,0039 (pri $k=17$) što je manje od granične vrijednosti te se može zaključiti, kao i kod hi-kvadrat testa, da distribucija prvih dviju znamenaka izvoza i uvoza Republike Hrvatske u razdoblju od 2010. do 2016. godine značajno ne odstupa (uz rizik od 5%) od Benfordove distribucije.

Kumulativ udjela pojavljivanja prvih dviju vodećih znamenki izvoza i uvoza RH po poglavljima Carinske tarife / Kombinirane nomenklaturo i zemljama namjene / podrijetla od 2010. do 2016. godine u odnosu na Benfordovou distribuciju (N = 70.835)

Grafikon 4: Kumulativ udjela pojavljivanja prve dvije vodeće znamenke izvoza i uvoza RH po poglavljima Carinske tarife / Kombinirane nomenklature i zemljama namjene / podrijetla od 2010. do 2016. godine u odnosu na Benfordovou distribuciju (N=70 .835)



Izvor: https://www.dzs.hr/, 19. rujna 2017.

3.4. Prosječno apsolutno odstupanje

Prosječno apsolutno odstupanje je test koji preporučuje M. J. Nigrini[4], koji za razliku od prethodna dva testa (hi-kvadrat i K-S) u izračun ne uzima veličinu uzorka. Dan je izrazom:

$$\text{prosječno apsolutno odstupanje} = \frac{\sum_{i=1}^{k} \left| r_o - r_t \right|}{k} \tag{10}$$

gdje je k broj razreda, a r_o i r_t predstavljaju udjele pojavljivanja promatranih razreda. Dobivena vrijednost se uspoređuje s kritičnim vrijednostima prikazanim u tablici 2.

[4] Mnogi autori ga smatraju začetnikom primjene Benfordovog zakona u forenzičkom računovodstvu

Tablica 2: Granične vrijednosti prosječnog apsolutnog odstupanja

Znamenke	Raspon	Zaključak
Prva znamenka	0,0000 - 0,0006	Vrlo malo odstupanje
	0,0006 - 0,0012	Prihvaćeno odstupanje
	0,0012 - 0,0015	Marginalno prihvaćeno odstupanje
	Iznad 0,0015	Veliko odstupanje
Druga znamenka	0,000 - 0,008	Vrlo malo odstupanje
	0,008 - 0,001	Prihvaćeno odstupanje
	0,0010 - 0,0012	Marginalno prihvaćeno odstupanje
	Iznad 0,0012	Veliko odstupanje
Prve dvije znamenke	0,0000 - 0,0012	Vrlo malo odstupanje
	0,0012 - 0,0018	Prihvaćeno odstupanje
	0,0018 - 0,0022	Marginalno prihvaćeno odstupanje
	Iznad 0,0022	Veliko odstupanje
Prve tri znamenke	0,00000 - 0,00036	Vrlo malo odstupanje
	0,00036 - 0,00044	Prihvaćeno odstupanje
	0,00044 - 0,00050	Marginalno prihvaćeno odstupanje
	Iznad 0,00050	Veliko odstupanje

Izvor: Nigrini (2012: 160).

U slučaju izvoza i uvoza RH, prosječno apsolutno odstupanje iznosi 0,000314 što znači da distribucija prvih dviju znamenaka izvoza i uvoza Republike Hrvatske u razdoblju od 2010. do 2016. godine vrlo malo odstupa od Benfordove distribucije.

4. Primjena Benfordovog zakona u forenzičkom računovodstvu

Benfordov zakon proučava se u raznim područjima poput: udari gromova (Manoochehrnia et al., 2010), potresi (Díaz et al., 2015), nuklearne znanosti (Wells et al., 2007), medicinske snimke (Sanches i Marques, 2006), aktivnosti Jehovinih svjedoka (Mir, 2014) i mnoga druga. Kao što je već rečeno, za većinu baza podataka iz stvarnog života (one koje zadovoljavaju „nematematičke" smjernice iz poglavlja 3) očekuje se da prate Benfordovu distribuciju. Ako odstupanje bude signifikantno, to je znak da se u skupu podataka javljaju abnormalna ponavljanja znamenaka i/ili neke druge anomalije. Analizom tih anomalija utvrđuje se uzrok njihovog nastajanja koji, među ostalim, može biti i namjerna manipulacija podacima. Na primjer, u nekom izvještaju o prodaji može se pojaviti povećan udio brojeva koji započinju na 19, 49 i 99, no umjesto zaključka da se radi o manipulaciji brojeva u izvještaju, najvjerojatniji je uzrok psihološko određivanje cijena pa zbilja većina prodanih artikala i ima te cijene.

U području forenzičkog računovodstva preporučuje se (Nigrini, 2012: 87) analiza distribucija prve znamenke i prvih dviju znamenaka. Nema pravila koji izvještaji se analiziraju (dokle god zadovoljavaju „nematematičke" smjernice iz poglavlja 3), nego bi se analitičar trebao voditi vlastitom prosudbom. Na primjer, ako postoji sumnja da je neko poduzeće manipuliralo brojevima u financijskim izvještajima 2014. i 2015. godine, ne bi se trebali uzeti financijski izvještaji od 2000. godine nego recimo od 2013. do 2016. godine. Kao što je prethodno spomenuto, analizom znamenaka identificiraju se anomalije, a ne manipulacije, tako da je nužna uporaba i drugih forenzičkih metoda kako bi se identificirala moguća manipulacija. Glavne prednosti primjene Benfordovog zakona u forenzičkom računovodstvu su:

- izračun je jednostavan
- otkrivaju sestavke u kojima je došlo do moguće manipulacije na koje revizor treba obratiti pozornost
- vrlo je teško manipulirati brojevima u izvještajima, a da znamenke i dalje prate Benfordov zakon
- revizor ima slobodu da prema vlastitoj prosudbi odabere izvještaje i razdoblje za potrebe analize distribucije znamenaka.

4.1. Primjena Benfordov zakona na poduzećima izlistanim na Zagrebačkoj burzi

Kako autor nema pristup povlaštenim informacijama o poslovanjima poduzeća (analitičke bruto bilance, izvještaji o prodaju, izvještaji o uvozu i izvozu i sl.), primjena Benfordovog zakona ograničena je na javno objavljene financijske izvještaje. Na portalu Financijske agencije objavljuju se godišnji financijski izvještaji (uglavnom bilanca i račun dobiti i gubitka), no za potrebe analize znamenaka je potrebno minimalno 1000 podataka (za analizu prvih dviju znamenki još i više) za što bi bilo potrebno promatrati razdoblje duže od 10 godina. Zbog tog razloga su promatrana poduzeća (dionička društva) izlistana na Zagrebačkoj burzi (dalje u tekstu: ZSE) jer ona imaju obvezu kvartalne javne objave bilance, računa dobiti i gubitka i novčanog toka.

U ovome poglavlju analizirana je distribucija prve znamenke i prvih dviju znamenki u financijskim izvještajima 12 nasumično odabranih poduzeća[5] izlistanih na ZSE. Konkretno, analizirani su nekonsolidirana bilanca, račun dobiti i gubitka i izvještaj o novčanom tijeku za razdoblje od 2011. do 2016. godine. Provedene su sve tri metode analize odstupanja: hi-kvadrat test (u tablici 3: χ^2), Kolmogorov-Smirnovljev

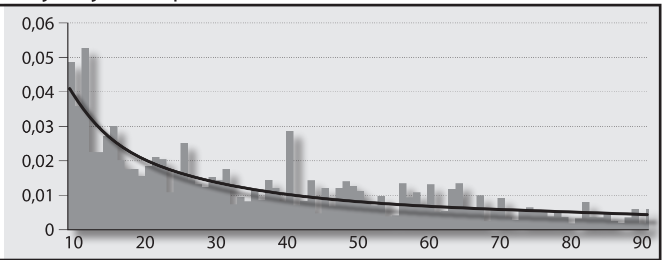
[5] Imena poduzeća su zamijenjena abecednim slovima jer cilj ovog rada nije identifikacija poduzeća s anomalijama.

test (u tablici 3: K-S) i prosječno apsolutno odstupanje (u tablici 3: MAD). U tablici 3 su prikazani rezultati analize, tamno je označeno signifikantno odstupanje (uz p = 5%), a sivo da nema signifikantnog odstupanja. Nadalje, na grafikonima je prikazana distribucija udjela prvih dviju znamenaka Benfordove distribucije (crna linija) i znamenaka u promatranim financijskim izvještajima (sivi stupci). Sa #N je označen ukupan broj podataka po poduzeću.

Tablica 3: Rezultati analize distribucije prve znamenke i prve dvije znamenke u financijskih izvještajima 12 poduzeća izlistanih na ZSE

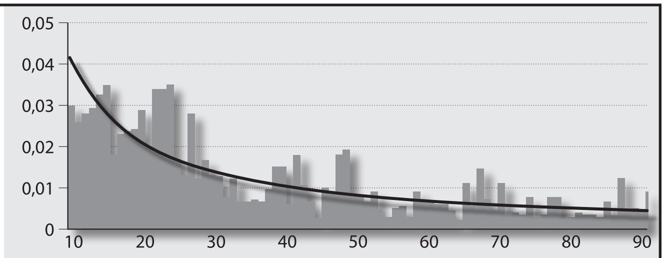
A d.d.

	d_1	d_1d_2
χ^2	41,92	399,9
K - S	0,0273	0,0311
MAD	0,0099	0,0031
#N	2791	



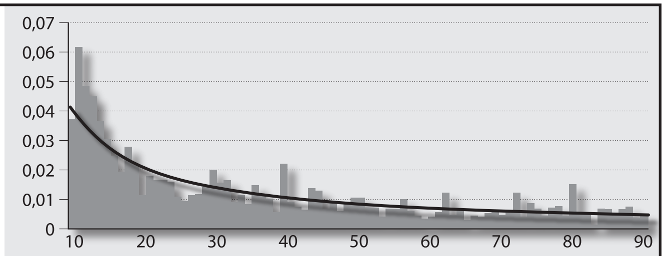
B d.d.

	d_1	d_1d_2
χ^2	79,84	366,7
K - S	0,0325	0,0335
MAD	0,0193	0,0036
#N	1968	



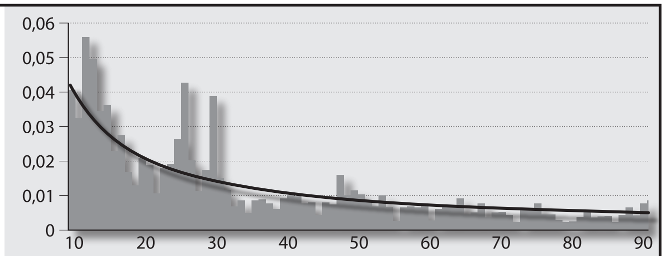
C d.d.

	d_1	d_1d_2
χ^2	80,00	384,2
K - S	0,0528	0,0556
MAD	0,0159	0,0031
#N	2942	



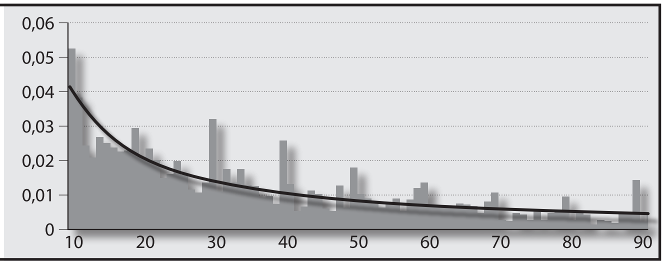
D d.d.

	d_1	d_1d_2
χ^2	49,44	513,9
K - S	0,0543	0,0799
MAD	0,014	0,0035
#N	2576	



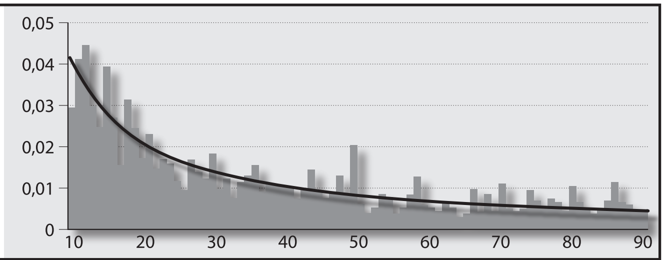
E d.d.

	d_1	d_1d_2
χ^2	43,14	460,6
K - S	0,0244	0,0244
MAD	0,0110	0,0029
#N	3108	



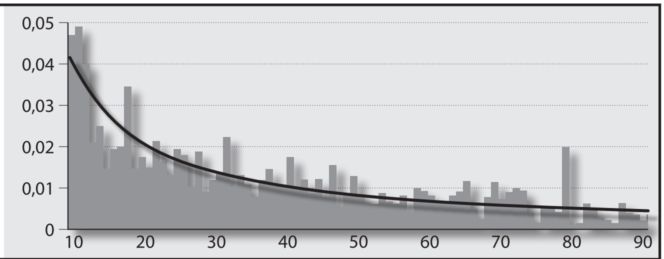
F d.d.

	d_1	d_1d_2
χ^2	34,75	301,1
K - S	0,0212	0,0253
MAD	0,0092	0,0027
#N	2937	



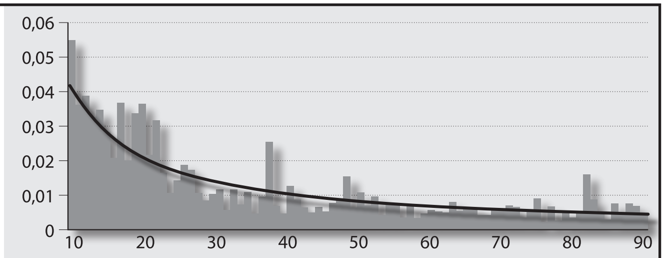
G d.d.

	d_1	d_1d_2
χ^2	24,22	411,7
K - S	0,0307	0,0333
MAD	0,0086	0,0032
#N	2715	



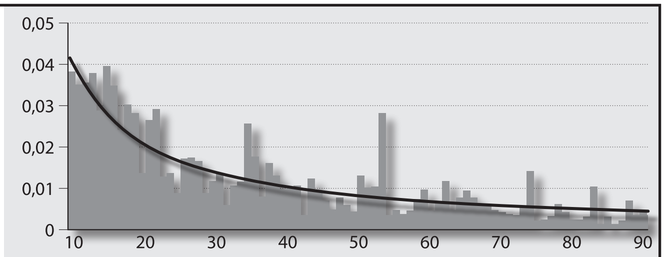
H d.d.

	d_1	d_1d_2
χ^2	52,83	413,9
K - S	0,0472	0,0661
MAD	0,0138	0,0032
#N	2737	



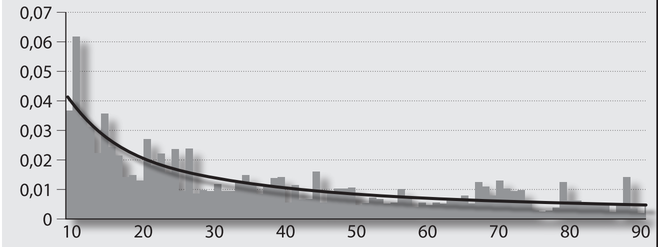
I d.d.

	d_1	d_1d_2
χ^2	51,17	546,4
K - S	0,0331	0,0403
MAD	0,0135	0,0034
#N	2873	



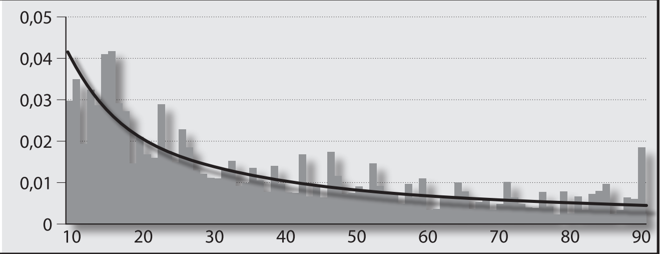
J d.d.

	d_1	d_1d_2
χ^2	32,94	411,1
K - S	0,0373	0,0449
MAD	0,0083	0,0033
#N	2825	



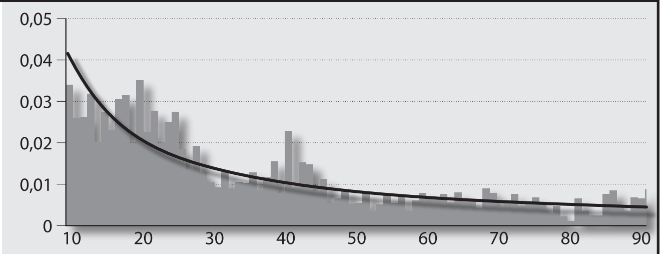
K d.d.

	d_1	d_1d_2
χ^2	12,61	412,7
K - S	0,0133	0,0318
MAD	0,0056	0,0030
#N	2841	



L d.d.

	d_1	d_1d_2
χ^2	86,94	309,2
K - S	0,0309	0,0420
MAD	0,0169	0,0029
#N	2887	



Izvor: autor, prema http://www.zse.hr/default.aspx?id=36774, 22. rujna 2017.

Granične vrijednosti hi-kvadrat testa su 15,51 za prvu znamenku i 112,02 za drugu znamenku, za Kolmogorov-Smirnovljev test su izračunate preko izraza (7) gdje je $\alpha = 0{,}05$, odnosno

$$\text{granična vrijednost K-S} = \frac{1{,}36}{\sqrt{n}} \tag{11}$$

gdje je n veličina uzorka, za prosječno apsolutno odstupanje su uzete granične vrijednosti prikazane u tablici 2. Veličina uzorka je u svim promatranjima veća od preporučenog minimuma koji iznosi 1000.

Provedenom analizom utvrđeno je da su odstupanja distribucije znamenaka (prve i prve dvije) od Benfordove distribucija signifikantna u svim promatranim poduzećima i u svim metodama, osim za F d.d. gdje odstupanje za prvu znamenku nije signifikantno po Kolmogorov-Smirnovljev testu i za K d.d. gdje odstupanje nije signifikantno za prvu znamenku po hi-kvadrat i Kolmogorov-Smirnovljev testu. Rezultati su u skladu s intuitivnim očekivanjem dobivenim promatrajući priložene grafove, gdje su velika odstupanja lako primjetna (ali se neovisno o tome vidi tendencija podataka da odgovaraju Benfordovoj distribuciji).

Promatrajući grafove revizor može uočiti neke zanimljive pojave koje možda ne bi inače uočio, npr. poduzeća A d.d. i L d.d. oba imaju veliku frekvenciju pojavljivanja brojeva koji započinju znamenkama 41, uglavnom zbog stavka *Rezerve za vlastite dionice* i *Vlastite dionice i udjeli*. Konkretno, kod poduzeća A d.d. spomenute stavke iznose 41 459 113 kuna za sve kvartale u razdoblju od 2011. do 2014., dok kod poduzeća L d.d. iznose 41 884 885 kuna za sve kvartale od 2011. godine do trećeg kvartala 2014. godine. Još jedna zanimljiva anomalija se može primijetiti kod poduzeća E d.d. koje ima vrlo izražene skokovima u deseticama 10, 30, 40, 50, 60, 70, 90.Za traženje uzroka tih anomalija su potrebni detaljniji izvještaji od onih javno objavljenih po grupnim kontima (oznake konta s dvjema znamenkama).

5. Zaključak

Protuinvitno ponašanje vodećih znamenaka, koje prati Benfordov zakon, matematički je dokazivo i ono se može proširiti na izračun vjerojatnosti pojavljivanja bilo kojih znamenaka u bilo kojoj bazi. Benfordova distribucija ima svojstvo invarijantnosti na skalar, što znači da ako neka baza podataka zadovoljava Benfordov zakon, onda će i nakon množenja sa skalarom ostati zadovoljen Benfordov zakon. Spomenuto svojstvo važno je jer to znači da za potrebe analize distribucije znamenaka nije bitno jesu li vrijednosti u kunama, eurima, stopama, metrima i sl., čime Benfordov zakon dobiva i univerzalnu primjenu svugdje u svijetu.

Problem koji se javlja pri analizi stvarnih baza podataka jest da su one konačne (dok se dokaz Benfordovog zakona temelji na beskonačnom skupu vrijednosti), zbog čega se koriste metode procjene odstupanja distribucije znamenaka od Benfordovog zakona. Te metode zahtijevaju veliki broj podataka što je i jedan od razloga da se primjena Benfordovog zakona tek počela razvijati krajem 20. st. kad su računala postala dostupna kao i razne baze podataka zahvaljujući informatizaciji poslovanja. Najčešće korištene metode su hi-kvadrat test, Kolmogorov-Smirnovljev test i prosječno apsolutno odstupanje.

Prije analize distribucije znamenaka, potrebno je utvrditi treba li promatrana baza podataka zbilja zadovoljavati Benfordov zakon koristeći nematematiče smjernice iskazane u poglavlju 3.1. (veličina uzorka barem 1000, pozitivna asimetričnost, ne postoji ugrađeni minimum ili maksimum, podaci su kvantificirani događaji ili činjenice, a ne oznake npr. OIB, JMBG i sl.). Ako odstupanje znamenaka od Benfordovog zakona bude signifikantno, to ukazuje da postoje abnormalna pojavljivanja znamenaka i/ili druge anomalije u promatranoj bazi podataka. Konkretno u forenzičkom računovodstvu, revizoru je to signal da je došlo do moguće manipulacije brojevima, što je i jedan od razloga sve veće primjene Benfordovog zakona u tom području. Nadalje, za razliku od mnogih drugih metoda manipulacije izvještajima, teško je „sakriti" manipulacije u financijskim izvještajima od analize distribucije znamenaka. Za područje forenzičkog računovodstva je preporučena analiza distribucije prve znamenke i prve dvije znamenke.

Analizom financijskih izvještaja 12 poduzeća izlistanih na Zagrebačkoj burzi utvrđeno je da sva bilježe odstupanja od Benfordove distribucije (ali je i primjetna tendencija znamenaka prema Benfordovoj distribuciji), no na tome je analiza završila zbog manjka pristupa drugim i/ili detaljnijim izvještajima. Ukazana odstupanja mogu biti od pomoći revizorima jer ukazuju na stavke u kojima su prisutne anomalije, a na revizorima jest da utvrde jesu li te anomalije uzrokovane manipulacijom. Mišljenje je autora da su za daljnja istraživanja primjene Benfordovog zakona u forenzičkom računovodstvu potrebne mnoge detaljne baze podataka o poslovanjima poduzeća koje, na žalost, spadaju u povlaštene informacije.

Literatura

1. Benšić, Mirta i Šuvak, Nenad. 2014. *Uvod u vjerojatnost i statistiku*. Osijek: Sveučilište J. J. Strossmayera.
2. Berger, Arno i Hill, Theodore P. 2011. A basic theory of Benford's Law. *Probability Surveys*, 8: 1–126.
3. Díaz Cusi, J. et al. 2015. On the Ability of the Benford's Law to Detect Earthquakes and Discriminate Seismic Signals. *Seismological Research Letters*, 86 (1): 192–201.
4. Jamain, Adrien. 2001. *Benford's law*. Imperial College of London.
5. Manoochehrnia, Pooyan et al. 2010. Benford's Law and Its Application to Lightning Data. *IEEE Transactions on Electromagnetic Compability*, 52 (4): 956–961.
6. Mir, Tariq Ahmad. 2014. The Benford law behavior of the religious activity data. *Physica A: Statistical Mechanics and its Applications,* 408: 1–9.
7. Nigrini, Mark. 2012. *Benford's Law: Application for Forencis Accounting. Auditing, and Fraud detection.* John Wiley and Sons.
8. Nigrini, Mark. 2011. *Forensic analytics: Methods and Techniques for Forensic Accounting Investigations*. John Wiley and Sons.
9. Papić, Milan. 2014. *Primijenjena statistika u MS Excelu*. Zagreb: Zoro.
10. Sanches, João M. i Marques, Jorge S. 2006. Image Reconstruction Using the Benford Law. *2006 International Conference on Image Processing:* 2029–2032.
11. Wells, Kevin et al. 2007. Quantifying the Partial Volume Effectin PET Using Benford's Law. *IEEE Transactions on Nuclear Science,* 54 (5): 1616–1625.

Internetski izvori

12. Državni zavod za statistiku. www.dzs.hr.
13. The University of Alabama in Huntsville, Mathemacial Sciences. www.math.uah.edu/stat/special/Benford.html.
14. Medalc. www.medcalc.org/manual/chi-square-table.php.
15. The World Bank. https://data.worldbank.org/indicator/SP.POP.TOTL, 12. rujna 2017.
16. Zagrebačka burza. www.zse.hr.

Benford's law and its application in forensic accounting

Abstract

Benford's law, also known as the first-digit law, is a statement about the frequency distribution of leading digits in many real-life sets of numerical data.It is intuitively expected that leading digit distribution is uniform (i.e. the probability for each digit to be a leading digit is equally), but Benford's Law states that leading digits follow logarithmic distribution. In this paper the proof of Benford's Law is shown, alongside with its application in forensic accounting, one of many possible fields of it's application.

Key words: Benford's Law, forensic accounting, revision, first-digit law