Univerzitet u Beogradu
Elektrotehnički fakultet

Matematička statistika

Benfordov zakon

Jovana Savić 2020/3423

mentor
Prof. dr Milan Merkle

Glava 1

Uvod

Godine 1881. astronom *Simon Newcomb* je primetio da su stranice u tablici logaritama koje predstavljaju brojeve koji počinju jedinicom mnogo više pohabane nego ostale. On je tada formulisao zakon verovatnoće prve cifre koji je tvrdio da je verovatnoća da vodeća cifra broja bude N jednak $\log(N+1)-\log(N)$ [1]. U tom trenutku je zakon bio samo formulisan, ali nije postojalo ni objašnjenje, kao ni podaci koji dokazuju da isti važi.

Nekih pedeset godina kasnije, fizičar *Frank Benfrod* dolazi do istog otkrića, primetivši istu stvar u logaritamskim tablicama. On je prikupio preko 20,229 brojeva koji su predstavljali razne matematičke i fizičke konstante, površine reka, koji su se pojavljivali u novinama i slično. Posmatrajući vodeće cifre ovih brojeva, u radu [1] je formulisao isti zakon do kog je došao i *Simon Newcomb*, koji je danas poznat kao Benfordov zakon.

Danas ovaj zakon ima i svoju mnogo precizniju formulaciju koja se bazira na značajnom delu broja, odnosno, na svim ciframa, u kojoj je Benfordov zakon poseban slučaj koji se odnosi na vodeću cifru. Na osnovu ovoga su i dokazane razne osobine koje pomažu da malo bolje razumemo zašto se ovakva raspodela vodećih cifara toliko često pojavljuje i gde možemo da je očekujemo. Matematičari *Arno Berger* i *Ted Hill* su dali veliki doprinos razvijanju i sistematizaciji teorije Benfordovog zakona [2, 3, 4, 5].

Benfordov zakon je jako zanimljiva pojava, ali od 1972. godine postaje i nešto što ima veliku primenu. Te godine je *Hal Varian* dao predlog da se ovaj zakon iskoristi da se proveri da li su podaci validni [6]. Njegova pretpostavka se bazirala na tome da ljudi nemaju osećaja za ovaj zakon kada manipulišu podacima i imaju tendenciju da ih menjaju tako da ravnomerno raspoređuju vodeće cifre. Ispostavilo se da je njegova pretpostavka opravdana, što zbog loše procene od strane ljudi koji nameštaju podatke, što zbog samih osobina Benfordovih promeljivih koje ćemo videti u ovom seminarkom radu. Danas se i u sudskim sporovima priznaju dokazi bazirani na Benfordovom zakonu. Kada je u pitanju primena Benfordovog zakona upućujemo na knjigu [7] u kojoj se mogu naći različiti primeri primene Benfordovog zakona u finansijskoj forenzici.

U glavi 2 su definisani osnovni pojmovi poput značajnih cifara i značajnog dela broja, kao i Benfordovih nizova i slučajnih promenljivih.

U glavi 3 su definisana svojstva i date teoreme koje opisuju Benfordove

[1] U ovom seminarskom radu ćemo sa log označavati isključivo dekadni logaritam.

slučajne promenljive. Prikazano je da Benfordov zakon važi i nakon skaliranja i promene brojevnog sistema.

U glavi 4 je ukratko definisano koji to deterministički i stohastički procesi dovode do pojave Benfordovog zakona, ali treba imati na umu da on i dalje nije potpuno objašnjena pojava [8].

U glavi 6 je diskutovana provera validnosti Covid-19 podataka primenom Benfordovog zakona.

Glava 2

Definicije osnovnih pojmova

2.1 Značajne cifre i značajni deo broja

Benfordov zakon govori o raspodeli cifara najveće težine, pa je u skladu sa tim neophodno prvo definisati ovaj pojam. Neformalno, cifra najveće težine nekog decimalnog broja je prva cifra različita od nule koja se pojavljuje u njegovom decimalnom zapisu. Tako, na primer, cifra najveće težine brojeva 2021 i 0.2021 je 2. Ipak, ovakva neformalna definicija ostavlja prostora za interpretaciju kada su u pitanju brojevi kao što je 1.99....

U ovom seminarskom radu ćemo sa log označavati dekadni logaritam. Takođe, koristićemo $\log 0 := 0$.

Definicija 2.1.1. Za svaki realan broj x različit od nule, prva cifra, odnosno cifra najveće težine ili vodeća cifra, u oznaci $D_1(x)$, je ceo broj $j \in \{1, 2, \ldots, 9\}$ koji zadovoljava uslov

$$10^k j \leq |x| < 10^k(j+1)$$

gde je k jedinstven ceo broj. Za svako $m \geq 2$, $m \in N$, m-ta cifra realnog broja x, u oznaci $D_m(x)$ se definiše kao ceo broj $j \in \{0, 1, \ldots, 9\}$ koji zadovoljava uslov

$$10^k \left(\sum_{i=1}^{m-1} D_i(x) 10^{m-i} + j \right) \leq |x| < 10^k \left(\sum_{i=1}^{m-1} D_i(x) 10^{m-i} + j + 1 \right)$$

gde je k jedinstven ceo broj.

Kada je u pitanju nula, imamo da je $D_m(0) = 0$ za svako $m \in N$.

Definicija 2.1.2. Za realan broj $x \in \mathbb{R}^+$, značajni deo broja, u oznaci $S(x)$, je dat kao $S(x) = t$, gde je t jedinstveni broj iz $[1, 10)$ za koji važi $x = 10^k t$ za neko jedinstveno $k \in \mathbb{Z}$. Ukoliko je x negativan broj imamo $S(x) = S(-x)$. Definišemo i $S(0) = 0$.

Značajni deo broja možemo da definišemo i eksplicitno:

$$S(x) = 10^{\log|x| - \lfloor \log|x| \rfloor}, \quad \forall x \neq 0 \tag{2.1}$$

Primer 2.1.1. Na osnovu datih definicija imamo $D_2(2021) = 0$, $S(\pi) = \pi$ i $D_1(1.99...) = 2$.

Iz definicija 2.1.1 i 2.1.2 direktno sledi svojstvo koje uspostavlja vezu između značajnog dela broja i njegovih cifara.

Svojstvo 2.1.1. Za svaki realan broj x važi:

(i) $S(x) = \sum_{m\in\mathbb{N}} 10^{1-m} D_m(x)$;

(ii) $D_m(x) = \lfloor 10^{m-1} S(x) \rfloor - 10\lfloor 10^{m-2} S(x) \rfloor, \quad \forall m \in \mathbb{N}.$

2.2 Benfordovi nizovi i slučajne promenljive

U ovom seminarskom radu ćemo posmatrati Benfordove nizove i slučajne promenljive. Osim njih, moguće je definisati i Benfordove funkcije, Benfordovu meru verovatnoće i Benfordovu raspodelu. Definicije ovih pojmova se mogu naći u [3].

Definicija 2.2.1. Niz realnih brojeva (x_n) je Benfordov niz ako za svako $t \in [1, 10)$ važi

$$\lim_{N\to\infty} \frac{\#\{1 \le n \le N : S(x_n) \le t\}}{N} = \log t \tag{2.2}$$

gde smo sa # označili broj elemenata skupa.

Alternativno, niz je Benfordov niz ukoliko za svako za svako $m \in \mathbb{N}$, za svako $d_1 \in \{1, 2, \ldots, 9\}$ i za svako $d_j \in \{0, 1, \ldots, 9\}$ za koje je $j \ge 2$, važi

$$\lim_{N\to\infty} \frac{\#\{1 \le n \le N : D_j(x_n) = d_j, j = \{1, 2, \ldots m\}\}}{N} = \log\left(1 + \left(\sum_{j=1}^{m} 10^{m-j} d_j\right)^{-1}\right). \tag{2.3}$$

Primer 2.2.1. Ukoliko koristimo alternativnu definiciju Benfordovog niza, uvođenjem $m = 1$ dobijamo poznati zakon prve cifre:

$$\lim_{N\to\infty} \frac{\#\{1 \le n \le N : D_1(x_n) = d\}}{N} = \log\left(1 + \frac{1}{d}\right) \tag{2.4}$$

za svako $d \in \{1, 2, \ldots, 9\}$.

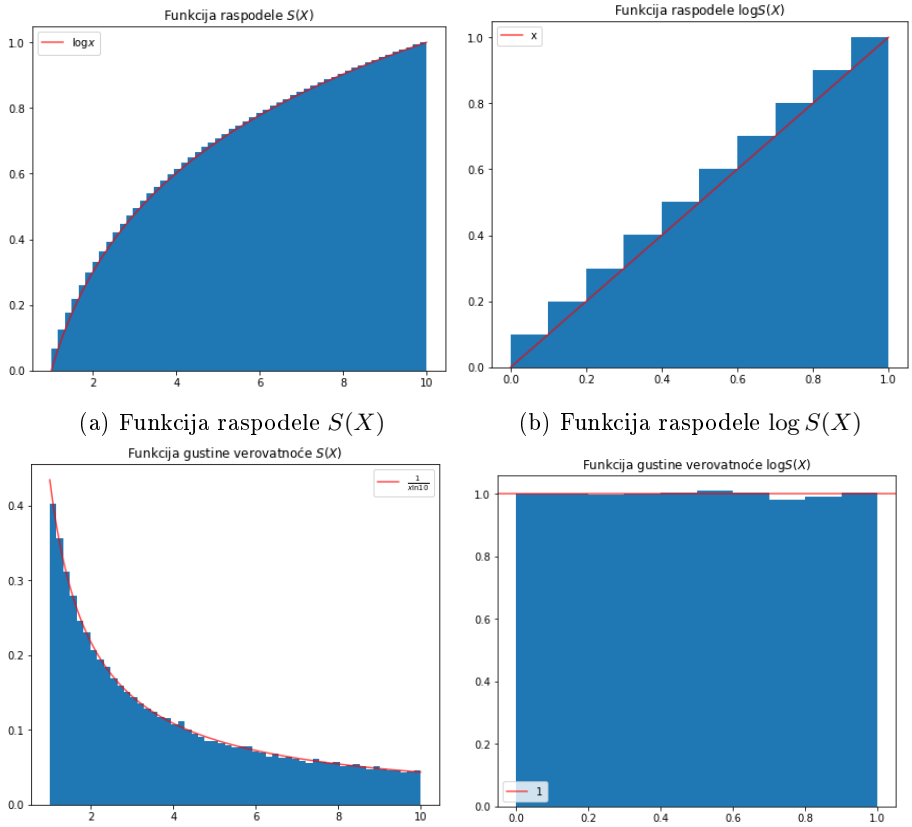
Definicija 2.2.2. Slučajna promenljiva X je Benfordova ako je

$$\mathbb{P}(S(X) \le t) = P_X(\{x \in \mathbb{R} : S(x) \le t\}) = \log t, \quad \forall t \in [1, 10)$$

odnosno, ako za svako $m \in \mathbb{N}$ svako $d_1 \in \{1, 2, \ldots, 9\}$ i svako $d_j \in \{0, 1, \ldots, 9\}$, $j \ge 2$ važi

$$\mathbb{P}(D_j(X) = d_j, 1 \le j \le m) = \log\left(1 + \left(\sum_{j=1}^{m} 10^{m-j} d_j\right)^{-1}\right)$$

Primer 2.2.2. Neka je U slučajna promenljiva sa uniformnom raspodelom na $[0, 1]$.



(a) Funkcija raspodele $S(X)$ (b) Funkcija raspodele $\log S(X)$

(c) Funkcija gustine verovatnoće $S(X)$ (d) Funkcija gustine verovatnoće $\log S(X)$

Slika 2.1: Funkcije gustine verovatnoće i funkcije raspodele značajnog dela broja i njegovog logaritma Benfordove slučajne promenljive

(i) Promenljiva U očigledno nije Benfordova promenljiva. Ovo se lako proverava, $P\left(S\left(X\right) \leq 2\right) = \frac{1}{9} < \log 2$.

(ii) Promenljiva $X = 10^U$ je Benfordova promenljiva. Na osnovu formule (2.1) dobijamo da je $S(X) = X$. Odavde nalazimo

$$P\left(S\left(X\right) \leq t\right) = P\left(X \leq t\right) = P\left(10^U \leq t\right) = P\left(U \leq \log t\right) = \log t$$

gde je $t \in [1, 10)$. Ovaj primer ujedno pokazuje kako se može generisati slučajna promenljiva sa Benfordovom raspodelom. Na slici 2.1 su prikazane funkcije gustine verovatnoće i funkcije raspodele značajnog dela broja Benfordove slučajne promenljive koja je generisana na ovaj način.

Glava 3

Svojstva Benfordovih nizova i slučajnih promenljivih

U ovom odeljku iznosimo najbitnija svojstva koja karakterišu Benfordove slučajne promeljive. Ispostavlja se da je logaritam značajnog dela broja Benfordove slučajne promenljive slučajna promenljiva sa uniformnom raspodelom po modulu 1. Ova raspodela je dosta proučavana, i ova veza je jedan od glavnih alata koji se koriste za izvođenje osobina Benfordovih objekata.

Najznačajnija svojstva Benfordovih promenljivih su invarijantnost u odnosu na skaliranje i promenu brojevnog sistema. Naime, ukoliko slučajna promenljiva prati Benfordov zakon, tada će isti važiti i kada se ona pomnoži proizvoljnom konstantom i kada se promeni brojevni sistem. Ovo ujedno eliminiše mogućnost da je pojava Benfordovog zakona posledica korišćenja dekadnog brojevnog sistema ili određenih jedinica mere.

3.1 Karakterizacija na osnovu uniformne raspodele

Najpre definišemo uniformnu raspodelu po modulu 1.

Definicija 3.1.1. Niz realnih brojeva (x_n) ima uniformnu raspodelu po modulu 1, u nastavku *u.r.* mod 1 ako

$$\lim_{N\to\infty} \frac{\#\{1 \le n \le N : \langle x_n \rangle \le s\}}{N} = s, \quad s \in [0,1).$$

Slučajna promenljiva X je *u.r.* mod 1 ako

$$\mathbb{P}(\langle X \rangle \le s) = s, \quad \forall s \in [0,1).$$

Teorema koju dajemo u nastavku uspostavlja vezu između uniformne raspodele po modulu 1 i Benfordove raspodele.

Teorema 3.1.1. *(Karakterizacija uniformne raspodele)* Niz realnih brojeva, funkcija sa Borelovom merom, slučajna promenljiva i mera verovatnoće su Benfordovi ako i samo ako njihov logaritam (sa osnovom 10) apsolutne vrednosti ima uniformnu raspodelu po modulu 1.

Dokaz. Videti [3], teoremu 4.2. □

U nastavku navodimo svojstva koja nam pomažu da odredimo koji nizovi i slučajne promenljive jesu Benfordovi.

Svojstva koja dajemo u nastavku direktno slede iz teoreme 3.1.1 i leme 4.3. iz rada [3].

Svojstvo 3.1.1. (i) Niz (x_n) je Benfordov ako i samo ako za svako $\alpha \in \mathbb{R}$ i svako $k \in \mathbb{Z}$, gde je $\alpha k \neq 0$, niz $\left(\alpha x_n^k\right)$ Benfordov.

(ii) Slučajna promenljiva X je Benfordova ako i samo ako je $1/X$ Benfordova.

U radu [3] je data teorema 4.2. sa dokazom koja obezbeđuje način za proveru da li je neki niz Benfordov. Radi potpunosti navodimo istu u nastavku.

Teorema 3.1.2. Ako su a, b, α, β realni brojevi gde je $\alpha \neq 0$ i $|\alpha| > |\beta|$ tada je $(\alpha^n a + \beta^n b)$ Benfordov niz ako i samo ako je $\log|\alpha|$ iracionalan broj.

U nastavku navodimo nekoliko primera u kojima ilustrujemo kako na osnovu teoreme 3.1.2 i navednih svojstava možemo da proverimo da li su određeni nizovi Benfordovi.

Primer 3.1.1. Niz (2^n) je Benfordov jer je $\log 2$ iracionalan broj. Sa druge strane, niz (10^n) nije Benfordov, jer je $\log 10$ racionalan broj. Na osnovu ovog primera vidimo da su mnogi geometrijski nizovi zapravo Benfordovi.

Primer 3.1.2. Kako je u prethodnom primeru pokazano da je niz (2^n) Benfordov, na osnovu svojstva 3.1.1 vidimo da su i nizovi $(\pi \cdot 4^n)$, $(3 \cdot 2^{-n})$ i slični, takođe Benfordovi nizovi.

Primer 3.1.3. Posmatrajmo niz prostih brojeva (p_n). Ovaj niz je Benfordov ako je $(\log p_n)$ *u.r.* mod 1. Na osnovu teoreme o prostim brojevima je $\lim_{n\to\infty} \frac{p_n}{n \log n} = 0$. Potreban uslov da niz (x_n) bude *u.r.* mod 1 je da je niz $(x_n \log n)$ neograničen (videti propoziciju 4.8. (iv) u [3]). U ovom slučaju je

$$\lim_{n\to\infty} \frac{\log p_n}{\log n} = \lim_{n\to\infty} \frac{\log(n \log n)}{\log n} = 1.$$

Odavde vidimo da niz prostih brojeva nije Benfordov niz.

Primer 3.1.4. Posmatrajmo niz Fibonačijevih brojeva (F_n) definisanih sa $F_{n+2} = F_{n+1} + F_n$, gde je $F_1 = F_2 = 1$. Iskoristićemo poznati rezultat

$$F_n = \frac{1}{\sqrt{5}}(\varphi^n - (-\varphi)^{-n})$$

gde je $\varphi = (1 + \sqrt{5})/2$. Direktnom primenom teoreme 3.1.2 vidimo da je Fibonačijev niz Benfordov jer je $\varphi > 1$ i $\log\varphi$ iracionalan broj.

3.2 Benfordova osobina kao invarijanta za skaliranje

Intuitivno, ako neki zakon važi za podatke koje srećemo u svakodnevnom životu, očekujemo da će isti važiti bez obzira na to koje jedinice koristimo. Ne

postoji pozitivna slučajna promenljiva čija je raspodela invarijanta skaliranja, odnosno, ne postoji slučajna promenljiva X takva da aX ima istu raspodelu za sve vrednosti skalara $a > 0$ (dokaz se može naći u [3]). Međtim, moguće je da raspodela njenog značajnog dela broja bude invarijanta za skaliranje. U nastavku pokazujemo da je potreban i dovoljan uslov da ovo važi da slučajna promenljiva bude Benfordova.

Definicija 3.2.1. Slučajna promenljiva X sa $P(X = 0) = 0$ ima cifre, odnosno značajni deo, čija je raspodela invarijanta za skaliranje ako su raspodele $S(aX)$ i $S(X)$ identične za svako $a \in \mathbb{R}$.

Teorema 3.2.1. Slučajna promenljiva X sa $P(X = 0) = 0$ ima cifre, odnosno značajni deo, čija je raspodela invarijanta skaliranja ako i samo ako je Benfordova.

Dokaz. Videti [3], teoremu 4.20. □

Sledeća teorema daje mogućnost formiranja neformalnog testa kojim bi se proverilo da li je slučajna promenljiva Benfordova.

Teorema 3.2.2. Slučajna promenljiva X sa $P(X = 0) = 0$ je Benfordova ako i samo ako je za neko $d \in \{1, 2, \ldots, 9\}$

$$P(D_1(aX) = d) = P(D_1(X) = d) \tag{3.1}$$

za svako $a \in \mathbb{R}^+$. Dodatno, (3.1) imlicira da je $P(D_1(X) = d) = \log\left(1 + d^{-1}\right)$.

Dokaz. Videti [3], teoremu 4.25. □

3.3 Benfordova osobina kao invarijanta za promenu brojevnog sistema

Benfordov zakon važi bez obzira na to koji brojevni sistem koristimo.

Definicija 3.3.1. Slučajna promenljiva X sa $P(X = 0) = 0$ ima cifre, odnosno značajni deo broja, čija je raspodela invarijanta za promenu brojevnog sistema ako su raspodele $S(X)$ i $S(X^n)$ identične za svako $n \in \mathbb{N}$.

Lema 3.3.1. Neprekidna slučajna promenljiva je Benfordova ako i samo ako ima cifre, odnosno značajni deo broja čija je raspodela invarijanta za promenu brojevnog sistema.

Dokaz. Ova teorema je direktna posledica teoreme 4.30 iz [3] ukoliko se ograničimo na neprekidne promenljive. □

Iz leme 3.3.1 i teoreme 3.2.1 direktno sledi sledeće svojstvo.

Svojstvo 3.3.1. Ako slučajna promenljiva ima ima cifre, odnosno značajni deo broja, čija je raspodela invarijanta za promenu brojevnog sistema tada je raspodela njenih cifara, odnosno značajnog dela broja invarijanta za skaliranje.

3.4 Benfordova osobina kao invarijanta sumiranja

Nijedan konačan skup podataka ne može da ima Benfordovu raspodelu u teorijskom smislu jer je i broj vrednosti značajnog dela broja konačan. Ono što je primećeno jeste da ako sumiramo značajne delove brojeva koji počinju jedinicom, potom značajne delove brojeva koji počinju dvojkom, itd, ove sume su približno jednake. Ukoliko je raspodela cifara logaritamska, ovakav rezultat je očekivan. Ova osobina se naziva invarijantom sumiranja.

U teorijskom razmatranju naš skup podataka je beskonačan, pa samim tim i opisane sume. Da bismo mogli da ih poredimo i svedemo na slučaj sa konačnim skupom podataka, uvodimo značajni deo broja za brojeve sa konačnim decimalnim zapisom kao

$$S_{d_1,\ldots,d_m}(x) := \begin{cases} S(x), & \text{ako } (D_1(x),\ldots,D_m(x)) = (d_1,\ldots,d_m) \\ 0, & \text{inače} \end{cases} \tag{3.2}$$

za svako $m \in \mathbb{N}$, svako $d_1 \in \{1,2,\ldots,9\}$ i $d_2 \in \{0,1,\ldots,9\}$ gde je $j \geq 2$.

Definicija 3.4.1. Niz realnih brojeva (x_n) je invarijanta za sumiranje ako za svako $m \in \mathbb{N}$ limes

$$\lim_{N\to\infty} \frac{\sum_{n=1}^{N} S_{d_1,\ldots,d_m}(x_n)}{N} \tag{3.3}$$

postoji i ne zavisi od $d_1,\ldots,d_n$.

Definicija 3.4.2. Slučajna promenljiva X je invarijanta za sumiranje ako, za svako $m \in \mathbb{N}$, vrednost $\mathbb{E}S_{d_1,\ldots,d_m}(X)$ ne zavisi od $d_1,\ldots,d_m$.

Teorema 3.4.1. Slučajna promenljiva X sa $\mathbb{P}(X=0)=0$ je invarijanta sumiranja ako i samo ako je Benfordova.

Dokaz. Videti [3], teoremu 4.37. □

Glava 4

Gde možemo da očekujemo Benfordov zakon?

U ovom poglavlju ćemo razmatrati gde se možemo da očekujemo Benfordove nizove i slučajne promenljive i na taj način probati da bar delimično objasnimo pojavu Benfordovog zakona.

4.1 Benfordov zakon i deterministički procesi

Mnogi poznati nizovi su Benfordovi. Na osnovu teoreme 3.1.2 vidimo da je skoro svaki geometrijski niz Benfordov.U primeru 3.1.4 smo videli da je i Fibonačijev niz Benfordov niz. U radu [9] je dokazano da je i niz faktorijela $(n!)$ takođe Benfordov. U radu [10] je dokazano da su nizovi poput $(\prod_{k=1}^{n} k!)$ i $\left(\prod_{k=1}^{n} k^k\right)$ Benfordovi. Mnogi procesi u prirodi imaju eksponencijalni rast i time dovode do pojave Benfordovog zakona. Ovo je verovatno i jedan od glavnih razloga što se često, pogrešno, pretpostavlja da podaci moraju imati veliki raspon da bi pratili Benfordov zakon [4]. Ovo, kao što smo mogli da vidimo na primeru slučajne promenljive 10^U ne mora da bude slučaj.

U primeru 3.1.3 je pokazano da niz prostih brojeva nije Benfordov, međutim, niz Mersenovih brojeva $(2^{p_n} - 1)$ jeste Benfordov [11].

Sa druge strane, očigledno je da niz prirodnih brojeva (n) nije Benfordov. Na osnovu svojstva 3.1.1 vidimo da nijedan niz oblika $\left(\alpha n^k\right)$, gde je $\alpha \in \mathbb{R}$ i $k \in \mathbb{Z}$. Prema tome, aritmetički nizovi nisu Benfordovi.

4.2 Benfordov zakon i stohastički procesi

U ovom odeljku ćemo pokazati u kojim se stohastičkim procesima javlja Benfordov zakon. Proizvodi nezavisnih promenljivih i kombinacije slučajnih promenljivih iz različitih raspodela imaju Benfordove osobine.

Definicija 4.2.1. Beskonačan niz slučajnih promenljivih $(X_1, X_2, \dots)$ konvergira u raspodeli ka Benfordovom zakonu ako

$$\lim_{n\to\infty} P\left(S\left(X_n\right) \leq t\right) = \log t \quad \forall\, t \in [1, 10)\,. \tag{4.1}$$

Ovaj niz je Benfordov za verovatnoćom 1 ako

$$P\left((X_n)\,\text{je Benfordov niz }\right) = 1 \tag{4.2}$$

Teorema 4.2.1. Ako je X neprekidna promenljiva, tada (X^n) konvergira u raspodeli ka Benfordovom zakonu i Benfordova je sa verovatnoćom 1.

Dokaz. Videti teoremu 6.1. [3]. □

Teorema 4.2.2. Neka su X i Y dve nezavisne promenljive sa $\mathbb{P}(XY = 0) = 0$. Tada važi:

(i) Ako je X Benfordova, i XY je takođe Benfordova promenljiva.

(ii) Ako $S(X)$ i $S(XY)$ imaju istu raspodelu, tada je ili $\log S(Y)$ racionalan broj sa verovatnoćom 1, ili je X Benfordova promenljiva.

Dokaz. Videti teoremu 6.3. u [3]. □

Teorema 4.2.3. Ako su X_1, X_2, X_3, ... nezavisne identično raspodeljene neprekidne slučajne promenljive, tada niz $(X_1, X_1X_2, X_1X_2X_3, \dots)$ konvergira u raspodeli ka Benfordovom zakonu i Benfordov je sa verovatnoćom 1.

Dokaz. Vidteti teoremu 6.6. u [3]. □

Konačno, daćemo teoremu koja je analogna centralnoj graničnoj teoremi. Da bi smo definisali ovu teoremu, prvo ćemo uvesti dve definicije.

Definicija 4.2.2. Slučajna mera verovatnoće $\mathbb{P}$ je slučajna promenljiva čije su vrednosti mere verovatnoće na $\mathbb{R}$.

Definicija 4.2.3. Slučajna mera verovatnoće $\mathbb{P}$ je nepomerena u odnosu na skaliranje kada su u pitanju značajne cifre ako njena prosečna mera verovatnoće $P_{\mathbb{P}}$ ima značajne cifre koje su invarijanta za skaliranje. lučajna mera verovatnoće $\mathbb{P}$ je nepomerena u odnosu na promenu brojevnog sistema kada su u pitanju značajne cifre ako njena prosečna mera verovatnoće $P_{\mathbb{P}}$ ima značajne cifre koje su invarijanta za promenu brojevne baze.

Sada kada smo uveli ove definicije dolazimo do ključne teoreme koja objašnjava da ukoliko su uzeti odbirci čija je raspodela slučajna, raspodela njihove kombinacije teži Benfordovom zakonu.

Teorema 4.2.4. Neka je $\mathbb{P}$ slučajna mera verovatnoće takva da $\mathbb{P}(S \in \{0,1\}) = 0$ sa verovatnoćom 1. Neka je $P_1, P_2, \dots$ niz nzavisnih identičnih slučajnih mera verovatnoće iz $\mathbb{P}$. Fiksirajmo prirodan broj m i neka $X_1, X_2, \dots, X_m$ bude slučajni uzorak veličine m iz P_1, neka je $X_{m+1}, \dots, X_{2m}$ slučajan uzorak veličine m iz P_2, i tako dalje. Ako je $\mathbb{P}$ nepomerena u odnosu na skaliranje ili promenu brojevnog sistema, tada empirijska raspodela kombinovanog uzorka $X_1, X_2, \dots, X_m, X_{m+1}, \dots$ konvergira ka Benfordovom zakonu sa verovatnoćom jedan, to jest

$$P\left(\lim_{N\to\infty} \frac{\#\{1 \le n \le N \ : \ S(X_n) \le t\}}{N} = \log t \ \forall \ t \in [q, 10)\right) = 1 \tag{4.3}$$

Dokaz. Videti teoremu 6.20 u [3]. □

Glava 5

Testiranje Benfordovog zakona

Kao što je već rečeno, Benfordov zakon se može koristiti kao neka vrsta provere validnosti podataka. U tom slučaju, naravno, treba voditi računa o tome da podaci budu takvi da se može očekivati da poštuju Benfordov zakon. Ukoliko je poznato da podaci imaju Benfordovu raspodelu, tada će se određene vrste nepravilnosti koje su uglavnom posledica nekorektnosti podataka, ispoljiti tako što podaci više neće pratiti Benfordov zakon.

Jako je bitno napomenuti da dokaz nepoštovanja Benfordovog zakona ne može da bude dokaz da je došlo do neke manipulacije podacima, već samo upozorenje da je potrebno dodatno ispitati validnost podataka.

U nastavku ćemo dati opis testova koji se najčešće koriste za proveru da li podaci imaju poštuju Benfordov zakon.

5.1 Pirsonov χ^2 test

Pirsonov test je jedan od najčešće korišćenih testova uopšte. Ovaj test je jako zgodan kada su podaci diskretne prirode. U nastavku dajemo teoremu na osnovu koje definišemo Pirsonov test [12].

Teorema 5.1.1. Neka je (X_i) nezavisan uzorak veličine n. Neka je r broj različitih ishoda, i neka je p_j verovatnoća j-tog ishoda. Sa O_j označimo broj opservacija tipa j. Tada, statistika

$$\chi^2 = \sum_{j=0}^{r} \frac{(O_j - np_j)^2}{np_j} \tag{5.1}$$

ima asimptotski raspodelu $\chi^2\,(r-1)$.

Velike vrednosti ove statistike ukazuju na veliku razliku između očekivanih i empirijskih ishoda, pa su indikacija za odbacivanje nulte hipoteze. Ako je $\chi^2 > \varepsilon_{\alpha-1}$, gde je $\varepsilon_{\alpha-1}$ kvantil reda $\alpha - 1$ raspodele $\chi^2\,(r-1)$, tada originalnu hipotezu odbacujemo sa nivoom značajnosti α.

U kontekstu provere Benfordovog zakona, nulta hipoteza je uglavnom to da podaci imaju Benfordovu raspodelu. Ishodi su vodeće cifre, pa posmatramo raspodelu $\chi^2(8)$. Značajne vrednosti ove statistike su date u tabeli 5.1.

Postoji empirijsko pravilo, takozvano Sturgesovo pravilo, koje nam daje vezu između broja klasa, odnosno ishoda, r i broja opservacija n:

$$r = 1 + 3.3 \log n. \tag{5.2}$$

Pošto je u situaciji u kojoj primenjujem ovaj test broj ishoda broj različitih cifara, i on je fiksiran, ovo pravilo možemo da iskoristimo da definišemo okviran broj odbiraka koji nam je potreban da bismo primenili Pirsonov test. Zamenom vrednosti dobijamo $n = 266$.

5.2 Kolmogorov-Smirnov test

Kolmogorov-Smirnov test definišemo na osnovu teoreme koju dajemo u nastavku.

Teorema 5.2.1. Neka je n veličina uzorka, $F = F_0$ nulta hipoteza (gde je F_0 neprekidna funkcija), a F_n empirijska raspodela. Tada statistika

$$\lambda = \sqrt{N} \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)| \tag{5.3}$$

ima K raspodelu.

Ako je $\lambda > \varepsilon_{\alpha-1}$, gde je $\varepsilon_{\alpha-1}$ kvantil reda $\alpha - 1$ raspodele K, tada originalnu hipotezu odbacujemo sa nivoom značajnosti α.

I u ovom slučaju posmatramo vodeće cifre brojeva, koje su diskretna slučajna promenljiva. Vidimo da ovde nije zadovoljen uslov teoreme. U radu [13] je pokazano da su, kada primenimo ovaj test na diskretnim funkcijama raspodele, tipične značajne vrednosti ovog testa jako stroge. To jest, hipotezu možemo da odbacimo i na nižim vrednostima statistike u odnosu na slučaj sa neprekidnom funkcijom raspodele. Ove vrednosti su date u tabeli 5.1, gde su statistike za diskretan slučaj posebno označene.

5.3 Kuiperov test

Za uzorak veličine n, sa empirijskom funkcijom raspodele F_n i nultom hipotezom $F = F_0$, Kuiperov test koristi statistiku

$$V_n = \left(D_n^+ + D_n^-\right)\left(\sqrt{n} + 0.155 + \frac{0.24}{\sqrt{n}}\right) \tag{5.4}$$

gde je $D_n^+ = \sup_{x \in \mathbb{R}} F_n(x) - F_0(x)$, a $D_n^- = \sup_{x \in \mathbb{R}} F_0(x) - F_n(x)$.

Za određivanje ove statistike se koristi tabela [13]. U tabeli 5.1 su date značajne vrednosti ove statistike.

Ovaj test je jako dobar za slučaj u kome je nulta hipoteza uniformna raspodela.

5.4 m-statistika i d-statistika

Za proveru da li podaci prate Benfordov zakon se još koriste i m i d statistike.

Definicija 5.4.1. U kontekstu Benfordovog zakona, m-statistiku definišemo kao:

$$m_n = \sqrt{n} \max_{d \in \{1,\ldots,9\}} \left| \Pr\left(D_0\left(X\right) = d\right) - \log\left(1 + 1/d\right) \right| \tag{5.5}$$

gde smo sa $D_0(X)$ označili vodeću cifru slučajne promenljive X.

Definicija 5.4.2. U kontekstu Benforfovog zakona, d-statistiku definišemo kao:

$$d_n = \sqrt{n} \left(\sum_{d \in \{1,\ldots,9\}} \left(\Pr\left(D_0\left(X\right) = d\right) - \log\left(1 + 1/d\right) \right)^2 \right)^{1/2} \tag{5.6}$$

U ovom slučaju se pojavljivanje svake cifre može smatrati jednim Bernulijevim eksperimentom. Tako se svakoj slučajnoj promenljivoj X može pridružiti vektor u kome se na osnovu indikatorske funkcije definiše njena vodeća cifra. Tada će prema centralnoj graničnoj teoremi ove statistike konvergirati ka normalnoj raspodeli. Na osnovu ove ideje se formiraju značajne vrednosti statistika. U tabeli 5.1 su date brojne vrednosti za najčešće korišćene nivoe značajnosti.

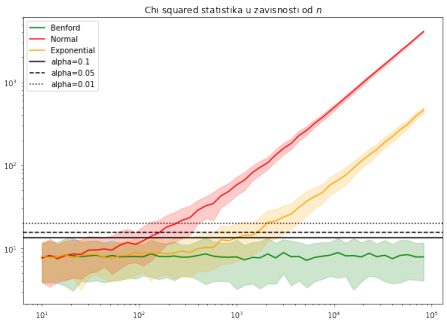
5.5 Značajne vrednosti statistika

U ovom odeljku predstavljamo tabelu u kojoj su sistematizovane značajne vrednosti svake statistike za tri najčešće korišćena nivoa značajnosti.

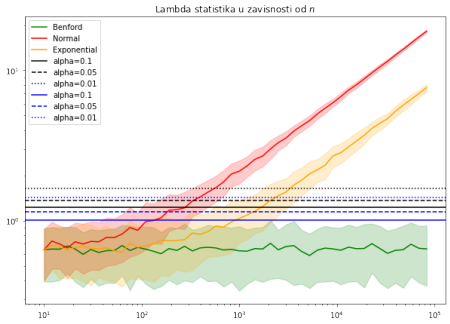
statistika	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
χ^2	13.362	15.507	20.090
λ	1.224 *1.012*	1.358 *1.148*	1.628 *1.420*
V_n	1.620 *1.191*	1.747 *1.321*	2.001 *1.579*
m_n	0.851	0.967	1.212
d_n	1.212	1.330	1.569

Tabela 5.1: Značajne vrednosti statistika u kontekstu testiranja Benfordovog zakona

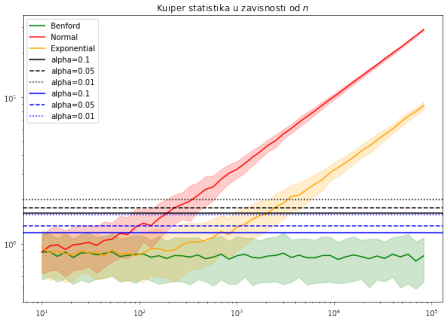
Da bi se stekla dodatna predstava o tome kako se različite statistike ponašaju u zavisnosti od veličine uzorka u odnosu na različite nivoe značajnosti generisano je po 100 uzoraka Benfordovih, eksponencijalnih i normalno raspodeljenih slučajnih promenljivih za različite vrednosti n i rezultati su prikazani na slici 5.1. Normalna i eksponencijalna raspodela su izabrane jer se radi o raspodelama koje su jako bliske Benfordovoj, što se vidi na slikama.



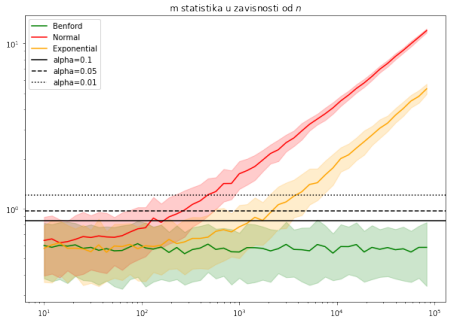
(a) χ^2 statistika



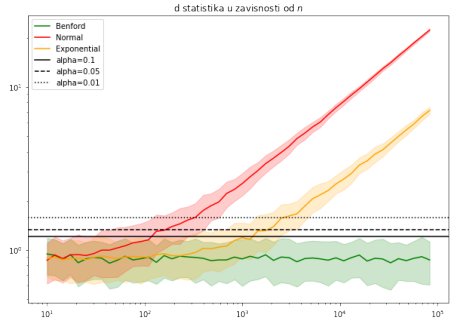
(b) λ statistika



(c) V_n statistika



(d) m_n statistika



(e) d_n statistika

Slika 5.1: Zavisnost različitih statistika od veličine uzorka. Punom linijom je prikazana prosečna vrednost, a svetlijom bojom je označena standardna devijacija. Linijama su označene značajne vrednosti.

Glava 6

Primena Benfordovog zakona

Benfordov zakon je jedan od najčešće korišćenih alata za proveru validnosti podataka. U knjizi [7] je detaljno objašnjena primena ovog zakona u analizi finansijskih podataka sa mnogo primera iz stvarnog života.

Od početka pandemije virusa Covid-19, pojavilo se dosta radova u kojima se koristi Benfordov zakon za analizu epidemioloških podataka, i u ovom seminarskom radu ćemo razmatrati primenu Benfordovog zakona u ovom kontekstu.

6.1 Covid-19 - prijavljeni slučajevi

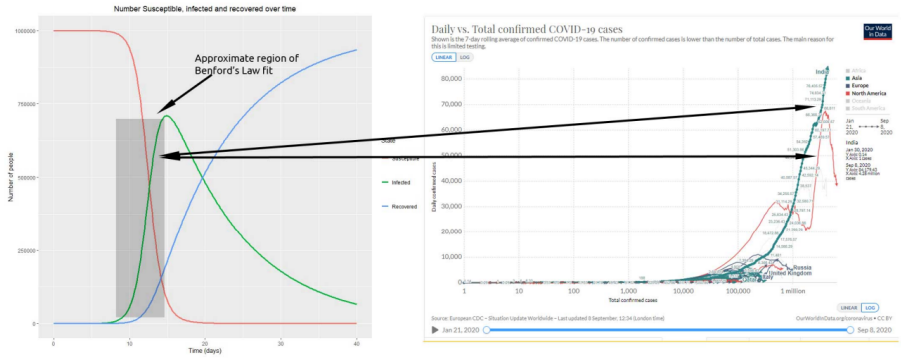
U radu [14] je pokazano da postoji osnova za korišćenje Benfordovog zakona za proveru prijavljenog broja slučaja zaraženih u toku epidemije. Ipak, treba imati na umu da Benfordov zakon možemo da očekujemo tamo gde broj zaraženih eksponencijalno raste. Onog trenutka kada počne uvođenje mera, prestaje eksponencijalni rast i samim tim više ne postoje osnove da nulta hipoteza bude ta da podaci imaju Benfordovu raspodelu. Ovo je prikazano na slici 6.1. U ovom radu je pokazano da broj zaraženih treba da raste 10% ili više, period u kome se dešava eksponencijalni rast treba da bude bar 50 dana i podaci treba da promene red veličine za bar 3. U radovima [14, 15] je predloženo da se odstupanje od Benfordovog zakona koristi kao mera za utvrđivanje uspešnosti suzbijanja epidemije Covid-19 virusa.

U radu [16] je posmatran prijavljeni broj slučaja virusa Covid-19 u Kini, Italiji i SAD-u [1]. Posmatrani su prijavljeni slučajevi pre i nakon uvođenja mera u različitim geografskim oblastima.

U ovom radu su korišćene χ^2, d_n, m_n i V_n statistike za testiranje toga da li podaci prate Benfordov zakon. U tabeli 6.1 su prikazane vrednosti ovih statistika dobijenih u radu [16] za prijavljeni broj zaraženih pre uvođenja mera.

U radu [16] piše da χ^2 statistika odbacuje nultu hipotezu, ali ovde nalazimo da to nije slučaj. Nijedna statiska ne odbacuje nultu hipotezu sa značajnošću boljom od $\alpha = 0.05$. Kada se uzme u obzir i veličina uzorka, podaci su zaista jako bliski Benfordovoj raspodeli i nema osnova da se odbaci nulta hipoteza. U [16] se takođe došlo do istog zaključka.

[1] Podaci se mogu naći na sajtu `https://datahub.io/core/covid-19#resource-countries-aggregated`



Slika 6.1: Primer SIR modela u kome je označena regija u kojoj se može očekivati Benfordov zakon [14].

Država	N	χ^2	d_n	m_n	V_n
Kina	581	16.04	1.16	0.79	0.33
Italija	359	5.00	0.65	0.29	0.64
SAD	1867	11.40	1.31	1.06	1.34

Tabela 6.1: Dobijene vrednosti statistika pre uvođenja mera [16]. Plavom bojom su označene vrednosti koje nisu dovoljne za odbacivanje hipoteze (smatrajući da značajnost mora da bude bar 0.1 za odbacivanje). Vrednosti na osnovu kojih odbacujemo nultu hipotezu sa nivoom značajnosti $\alpha = 0.05$ su označene braon bojom.

U slučaju SAD-a, ukoliko uzmemo strožu vrednost V_n statistike, date u tabeli 5.1, imamo tri statistike koje navode da je moguće odbaciti hipotezu sa nivoom značajnosti 0.05. Ipak, to što je odstupanje malo veće je najverovatnije posledica toga što je u ovom slučaju posmatran broj zaraženih od 29. februara 2020. do 30. juna 2020. godine. Podaci koji su korišćeni za Kinu i Italiju podrazumevaju datume do 16. marta 2021. i 16. aprila 2020. godine, respektivno. Tokom juna meseca su već uveliko uvođene različite mere u čitavom svetu i samim tim je najverovatnije da se tokom maja i juna usporava eksponencijalno širenje virusa, i da to dovodi do manjeg poklapanja sa Benfordovim zakonom.

Konačno, u kontekstu testiranja validnosti epidemioloških podataka je potrebno setiti se teoreme 3.2.1 koja kaže da, ako promenljiva ima Benfordovu raspodelu, imaće je i ako se skalira nekim brojem. Ovo je jako bitna osobina jer, ukoliko bi postojala manipulacija podacima, ona bi mogla da bude i u vidu skaliranja, a ne u vidu izmene pojedinačnih vrednosti. Pretpostavimo da nam je cilj da prikažemo manji broj registrovanih slučaja nego što zaista jeste. Dovoljno je taj broj podeliti istom konstantom svakog dana. Podaci će tada i dalje imati Benfordovu raspodelu, a prijavljeni broj će dati mnogo optimističniju sliku nego što to jeste slučaj. Broj testova koji se radi u jednom danu je jedan od načina da se manipuliše brojem zaraženih, koji ima efekat množenja konstantom. Čak i značajna promena u vidu broja testova (ukoliko nije česta) jeste nevidljiva za

ovakav test. Neka je broj prosečno urađenih testova u jednom danu N_1 do nekog trenutka, a potom N_2 u nastavku, podrazumevajući da smo u delu epidemije u kome broj zaraženih eksponencijalno raste. Tada će broj zaraženih virusom biti kombinacija dva geometrijska niza, i kao takav će i dalje pratiti Benfordov zakon [2]. Prema tome, samo posmatranje broja registrovanih slučajeva, bez razmatranja metodologije testiranja nije dovoljno.

[2]Ukupan broj slučajeva u različitim geografskim oblastima jeste kombinacija više geometrijskih nizova, i kao što je pokazano, podaci i dalje imaju Benfordovu raspodelu.

Glava 7

Zaključak

Benfordov zakon je veoma interesantan i iz čisto matematičke perspektive i iz perspektive primene. Na sajtu `https://www.benfordonline.net/` se mogu naći radovi koji su u vezi sa Benfordovim zakonom. Veliki broj radova koji se pojavio u poslednje vreme primenjuje Benfordov zakon u nekoj vrsti analize podataka Covid-19 pandemije. Veliki deo radova koristi Benfordov zakon za proveru validnosti podataka.

Glavni problem u ovakvoj primeni Benfordovog zakona jeste to što uvođenje mera dovodi do toga da podaci više ne prate Benfordov zakon i samim tim je ova primena ograničena samo na podatke koji su prikupljeni na početku pandemije. Sa druge strane, upravo ova osobina se može koristiti za analizu uspešnosti uvedenih mera.

Tako, na primer, u radu [17] je dat pregled vrednosti statistika χ^2, V_n i MAD za sve države i označeno je sa kojom značajnošću koja statistika odbacujuje hipotezu o Benfordovom zakonu. Uzeti su podaci počev od prvih registrovanih slučajeva u svakoj državi do 12. novembra 2020. godine. Situacije u kojima se odbacuje nulta hipoteza su interpretirane kao potencijalno netačno prijavljivanje registrovanih slučajeva. Kao što smo videli, ovo nije potpuno opravdana interpretacija, jer podaci uključuju period u kome su uveliko uvođene mere za suzbijanje pandemije.

Drugi problem u primeni Benfordovog zakona za proveru validnosti epidemioloških podataka je činjenica da se mnoge manipulacije mogu ispoljiti kao množenje konstantom i njih ne možemo da detekujemo na ovaj način. Interesantno je da ovakva vrsta manipulacije verovatno lakša za izvođenje i zahteva manju koordinaciju, obzirom da je moguće izvesti je ograničavanjem dostupnog broja testova i slično.

Ono u čemu je poređenje sa Benfordovim zakonom jako dobro jeste manipulacija podacima koja uključuje neku vrstu zaokruživanja tako da se dobije broj koji ima manji broj cifara. Poznato je da, iako su brojevi 990 i 1000 jako bliski, njihov psihološki uticaj se veoma razlikuje. Dakle, ukoliko neka vlada želi da prikaže situaciju boljom nego što jeste slučaj (da bi, na primer, opravdala sporo reagovanje), možemo da očekujemo odstupanje od Benfordovog zakona upravo na najvećim ciframa.

Primena Benfordovog zakona za proveru validnosti podataka zahteva oprez, pogotovo kada se proveravaju epidemiološki podaci. U ovom kontekstu, odstupanje ili neodstupanje od Benfordovog zakona se vrlo jednostavno proverava, ali

se teško interpretira. Ovo nisu ni potrebni, ni dovoljni uslovi za zaključivanje da li su podaci validni ili ne. Problem ispitivanja da li je broj prijavljenih slučajeva tačan nije lak i zavisi od ogromnog broja faktora koje treba uzeti u obzir. Uklapanje ili neuklapanje u Benfordov zakon se, eventualno, može iskoristiti kao dopunski argument onda kada postoje i drugi dokazi, nikako samostalno.

Literatura

[1] Frank Benford. The law of anomalous numbers. *Proceedings of the American philosophical society*, pages 551–572, 1938.

[2] Arno Berger and Theodore P. Hill. *An Introduction to Benford's Law*. Princeton University Press, 2015.

[3] Arno Berger and Theodore P. Hill. A basic theory of Benford's Law. *Probability Surveys*, 8(none):1 – 126, 2011.

[4] Arno Berger and Theodore P. Hill. The mathematics of benford's law – a primer, 2020.

[5] Arno Berger and Theodore P. Hill. What is ... benford's law? *Notices of the American Mathematical Society*, 64(2):132–134, February 2017.

[6] Hal R Varian. Benfords law. *American Statistician*, 26(3):65, 1972.

[7] M.J. Nigrini and J.T. Wells. *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*. Wiley Corporate F&A. Wiley, 2012.

[8] Arno Berger and Theodore P Hill. Benford's law strikes back: No simple explanation in sight for mathematical gem. *The Mathematical Intelligencer*, 33(1):85–91, 2011.

[9] Persi Diaconis et al. The distribution of leading digits and uniform distribution mod 1. *The Annals of Probability*, 5(1):72–81, 1977.

[10] Bruno Massé and Dominique Schneider. Fast growing sequences of numbers and the first digit phenomenon. *International Journal of Number Theory*, 11(03):705–719, 2015.

[11] Zhaodong Cai, Matthew Faust, AJ Hildebrand, Junxian Li, and Yuan Zhang. Leading digits of mersenne numbers. *Experimental Mathematics*, pages 1–17, 2019.

[12] Milan Merkle. *Verovatnoća i statistika*. Akademska misao, Beograd, 2010.

[13] John Morrow. Benford's law, families of distributions and a test basis. 2014.

[14] Nils Koesters, Andrena McMenemy, and Yohan Bélanger. Simulating epidemics with a sird model and testing with benford's law. *Pre-print) doi: https://www. newyorker. com/news/daily-comment/thecoronavirus-hits-brazil-hard-but-jair-bolsonaro-is-unrepentant*, 2020.

[15] Kang-Bok Lee, Sumin Han, and Yeasung Jeong. Covid-19, flattening the curve, and benford's law. *Physica A: Statistical Mechanics and its Applications*, 559:125090, 2020.

[16] R Isea. How valid are the reported cases of people infected with covid-19 in the world. *Int J Coronav*, 1(2):53–56, 2020.

[17] A Kilani and G P Georgiou. Countries with potential data misreport based on Benford's law. *Journal of Public Health*, 43(2):e295–e296, 01 2021.