Univerzitet u Beogradu
Elektrotehnički fakultet

Benfordov zakon i primene

Master rad

Mentor
Doc. Dr Bojana Mihailović

Kandidat
Jovana Savić 2020/3423

Beograd, Septembar 2021.

Sadržaj

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 $\log_{10}(N+1) - \log_{10}(N)$. 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 Benford dolazi do istog otkrića, primetivši istu stvar u logaritamskim tablicama. On je prikupio preko 20.000 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 [2] je formulisao isti zakon do kog je došao i Simon Newcomb, koji je danas poznat kao Benfordov zakon. Ovim je Benfordov zakon postao primer interesantnog Stiglerovog zakona koji tvrdi da većina otkrića nije dobila naziv po onome ko je prvi do njih došao[1].

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 [5, 3, 6, 16, 7, 15, 4].

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 validnost podataka [33]. Prvi radovi u kojima se zvanično koristi Benfordov zakon u analizi finansijskih podataka se pojavljuju 1992. godine, autori su Busta i Sundheim. Godine, 1992. Nigrini u svom doktorskom radu koristi Benfordov zakon za detekciju utaje poreza [31]. Nigrini je vodeći autor iz oblasti primene Benfordovog zakona u forenzici finansijskih podataka i autor preko 30 radova i knjige [30] iz ove oblasti. Danas se, u nekim državama, dokazi bazirani na Benfordovom zakonu priznaju na sudu [11].

U glavi 2 dajemo definicije osnovnih pojmova. U odeljku 2.1 definišemo prostor verovatnoće u okvirima teorije mere. U odeljku 2.2 dajemo precizne definicije vodećih cifara i značajnog dela broja (mantise). Na osnovu ovoga, u odeljku 2.3 definišemo sigma polje značajnog dela broja. Potom u odeljku 2.4 dajemo definicije Benfordovih objekata - nizova, funkcija, slučajnih promenljivih i mere verovatnoće. U odeljku 2.5 uspostavljamo vezu između uniformne raspodele po modulu 1 i Benfordove osobine. Na osnovu ove veze izvodimo brojne osobine i kroz primere pokazujemo da su mnogi poznati matematički objekti poput Fibonačijevog niza Benfordovi.

U glavi 3 razmatramo sve ono što je bitno kada pokušavamo da odgovorimo na pitanje da li od nekih podataka ima smisla očekivati poštovanje Benfordovog zakona. Sa tim u vezi u odeljcima 3.1, 3.2 i 3.3 pokazujemo da je Benfordova osobina invarijanta kada se podaci skaliraju,

[1] Interesantno je da je Stigler smatrao da je sociolog Robert K. Merton otkrio Stiglerov zakon.

da za neke promene brojevnog sistema takođe ostaje nepromenjena i da je zbir značajnih delova podataka koji imaju vodeću cifru konstantan. U odeljku 3.4 pokazujemo koji stohastički procesi dovode do pojave Benfordove osobine. Izlaganje teorije završavamo nekim otvorenim pitanjima koja se prirodno javljaju kao posledica predstavljene teorije.

Glava 4 se bavi problemom testiranja Benfordove raspodele onda kada se ona smatra nultom hipotezom. U odeljku 4.1 su predstavljene statistike koje se najčešće koriste. U odeljku 4.2 su razmatrani problemi koji postoje sa ovako korišćenim statistikama, da bi u odeljku 4.3 bila predložena relativna entropija kao mera usklađenosti sa Benfordovim zakonom. Argumenti koji se iznose u ovoj glavi su ilustrovani raznim simulacijama na računaru.

U glavi 5 se pokazuje primena Benfordovog zakona. U odeljku 5.1 su dati primeri u kojima je Benfordov zakon korišćen u forenzici finansijskih podataka, gde je možda najinteresantiji primer iz oblasti elektrotehnike onaj u kome je problem krađe struje rešen uz pomoć Benfordovog zakona. Konačno, u odeljku 5.2 dajemo kritičku analizu primene Benfordovog zakona koja je u poslednjih godinu i po dana postala veoma popularna, a radi se o detekciji anomalija u podacima Covid-19 pandemije.

U zaključku su sistematizovani rezultati teze i date smernice za dalja istraživanja.

Glava 2

Definicija osnovnih pojmova

U ovom poglavlju su predstavljene teorijske osnove Benfordovog zakona. U nastavku će biti pokazano da je Benfordov zakon vodeće cifre samo specijalan slučaj opštije pojave - logaritamske raspodele značajnog dela broja. Prvo ćemo precizno definisati prostor verovatnoće i značajni deo broja. Nakon toga definišemo sigma polje značajnog dela broja koje će biti osnova za dalju analizu Benfordovih osobina. Potom definišemo Benfordove nizove, funkcije, slučajne promenljive i meru verovatnoće. Konačno, u poslednjem odeljku uspostavljamo vezu između Benfordove osobine i uniformne raspodele po modulu 1. Ovde su, takođe, predstavljena razna svojstva uniformne raspodele po modulu 1 na osnovu kojih je moguće proveriti postojanje Benfordove osobine. Njihova primena je ilustrovana kroz primere u kojima je pokazano da mnogi matematički objekti poput geometrijskih nizova, eksponencijalnih funkcija i ostalih zapravo Benfordovi, i samim tim je delimično objašnjena česta pojava Benfordovog zakona u praksi.

2.1 Osnovni koncepti teorije verovatnoće

Osnovni koncept standardne teorije verovatnoće je prostor verovatnoće $(\Omega, \mathcal{A}, \mathbb{P})$, gde je Ω neprazan skup (skup događaja), $\mathcal{A}$ je sigma polje na Ω (čije elemente nazivamo događajima) i $\mathbb{P}$ je mera verovatnoće na $(\Omega, \mathcal{A})$. U nastavku ovog odeljka ćemo definisati ove pojmove i ukratko objasniti zašto se teorija verovatnoće definiše upravo na ovim osnovama.

Prvi pojam koji definišemo je sigma polje. Sa $\mathcal{P}(\Omega)$ ćemo označiti partitivni skup skupa Ω, odnosno skup svih podskupova skupa Ω.

Definicija 2.1. Neka je Ω neprazan skup. Familija $\mathcal{A}$ ($\mathcal{A} \subset \mathcal{P}(\Omega)$) je sigma polje na Ω ako su ispunjenjeni uslovi:

1. $\Omega \in \mathcal{A}$;
2. Ako $A_1, A_2, \cdots \in \mathcal{A}$, tada i $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$;
3. Ako $A, B \in \mathcal{A}$, tada i $A \setminus B \in \mathcal{A}$.

$\Diamond$

Iz definicije sigma polja proizilazi da je ono zatvoreno u odnosu na sve skupovne operacije koje se izvode prebrojivo mnogo puta u proizvoljnom poretku (dokaz se može naći u [27], teorema 1.8). Jedan od najznačajnijih primera sigma polja je Borelovo sigma polje čiju definiciju dajemo u nastavku.

Definicija 2.2. Najmanje sigma polje koje sadrži sve otvorene intervale $(a, b) \subset \mathbb{R}$ naziva se Borelovim sigma poljem. $\Diamond$

U kontekstu Benfordovog zakona će nam biti značajna sigma polja na $[0,1)$ i $[1,10)$, i to najmanja sigma polja, odnosno Borelova sigma polja. Ako je $C \subset \mathbb{R}$, sa $\mathcal{B}(C)$ označavamo sigma polje $C \cap \mathcal{B} := \{C \cap B : B \in \mathcal{B}\}$. Borelovo sigma polje na $[a,b)$ ćemo označavati sa $\mathcal{B}[a,b)$.

Sada možemo da definišemo pojam mere verovatnoće na $(\Omega, \mathcal{A})$. Ovaj pojam definišemo preko Kolmogorovih aksioma verovatnoće.

Definicija 2.3. Mera verovatnoće na $(\Omega, \mathcal{A})$ je funkcija $\mathbb{P} : \mathcal{A} \to [0,1]$ takva da važi

1. $\mathbb{P}(\emptyset) = 0$;
2. $\mathbb{P}(\Omega) = 1$;
3. $\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$ ako su skupovi $A_n \in \mathcal{A}$ disjunktni.

Napomena. Uređen par $(\Omega, \mathcal{A})$ je merljiv prostor. Ukoliko bismo verovatnoću definisali na podskupovima skupa događaja Ω kao funkciju koja zadovoljava prethodno navedene aksiome, umesto na sigma polju skupa Ω, kao što je to urađeno u definiciji 2.3, tada bi postojali događaji kojima se ne bi mogla pripisati verovatnoća. Vitalijev skup je primer takvog događaja kada je skup događaja skup $\mathbb{R}$. ◊

Uvešćemo još i pojam Lebegove mere na $([a,b), \mathcal{B}[a,b))$ kao

$$\lambda_{a,b}[c,d] := \frac{d-c}{b-a} \quad \forall [c,d] \subset [a,b) \tag{2.1}$$

koji predstavlja meru uniformne raspodele, pri tome podrazumevano je $a < b$.

2.2 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.... Zbog toga je potrebno uvesti precizne definicije koje dajemo u nastavku.

Definicija 2.4. 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 $m \in N$. ◊

Primer 2.1. Na osnovu date definicije je $D_1(2021) = 2$, $D_2(\pi) = 1$, $D_1(1.99\ldots) = 2$.

Definicija 2.5. Za realan broj $x \in \mathbb{R}^+$, značajni deo broja (mantisa) u dekadnom brojevnom sistemu, 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:

$$S(x) = 10^{\log|x| - \lfloor \log|x| \rfloor}, \quad \forall x \neq 0 \tag{2.2}$$

U nastavku ćemo tamo gde se pojavljuje logaritam smatrati da argument nije nula. Iz definicija 2.4 i 2.5 direktno sledi osobina koja uspostavlja vezu između značajnog dela broja i njegovih cifara.

Osobina 2.1. Za svaki realan broj x važi:

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

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

$\Diamond$

Navedena osobina pokazuje da su vodeće cifre broja i njegov značajni deo ekvivalentni pojmovi.

2.3 Sigma polje značajnog dela broja

U odeljku 2.1 je data definicija sigma polja formiranog od nekog skupa Ω. Međutim, sigma polje možemo da generišemo i korišćenjem funkcije.

Definicija 2.6. Neka je data funkcija $f : \Omega \to \mathbb{R}$, tada je sigma polje generisano funkcijom f, u oznaci $\sigma(f)$, dato sa

$$\sigma(f) = \left\{ f^{-1}(B) \,:\, B \in \mathcal{B} \right\}. \tag{2.3}$$

$\Diamond$

U kontekstu teorije verovatnoće, funkcija $f : \Omega \to \mathbb{R}$, gde je $\sigma(f) \subset \mathcal{B}$ predstavlja slučajnu promenljivu. Ako je $\mathbb{P}$ mera verovatnoće na $(\Omega, \mathcal{B})$, funkciju raspodele slučajne promenljive X, u oznaci P_X definišemo kao meru verovatnoće na $(\mathbb{R}, \mathcal{B})$ izrazom:

$$P_X\left((-\infty, t]\right) = \mathbb{P}\left(X \leq t\right), \quad \forall t \in \mathbb{R}. \tag{2.4}$$

Odnosno, ukoliko slučajnu promenljivu definišemo kao funkciju f, njenu funkciju raspodele definišemo kao meru verovatnoće njoj odgovarajućeg intervala

$$P_X = X_* \mathbb{P} = \mathbb{P}\left(f^{-1}(B)\right), \quad \forall B \in \mathcal{B}. \tag{2.5}$$

Na osnovu izloženog možemo da formiramo definiciju sigma polja generisanog značajnim delom, odnosno mantisom, nekog realnog broja.

Definicija 2.7. Sigma polje značajnog dela broja $\mathcal{S}$ je sigma polje na $\mathbb{R}^+$ generisano funkcijom S koja vraća značajni deo broja, odnosno $\mathcal{S} = \mathbb{R}^+ \cap \sigma(S)$. $\Diamond$

Sigma polje značajnog dela broja $\mathcal{S}$ je familija svih događaja $A \subset \mathbb{R}^+$ koji se mogu u potpunosti opisati svojim značajnim delom broja (videti [3], lemu 2.8).

Primer 2.2. Skup svih brojeva čiji je značajni deo racionalan broj, $A_1 = \{x > 0 : S(x) \in \mathbb{Q}\}$ pripada $\mathcal{S}$. Skup svih brojeva čija je vodeća cifra manja od 3, $A_2 = \{x > 0 : D_0(x) < 3\}$ i skup svih brojeva čija druga cifra nije 2, $A_3 = \{x > 0 : D_1(x) \neq 2\}$ takođe pripadaju $\mathcal{S}$. Sa druge strane, skup brojeva $A_4 = \{x : x \in [1, 2]\}$ ne pripada $\mathcal{S}$. Razlog je to što funkcija S ne razdvaja intervale $[1, 2]$ i $[10, 20]$ i slične.

U nastavku navodimo teoremu koja će nam pomoći da precizno definišemo oblik intervala koji pripadaju $\mathcal{S}$. Pritom ćemo za svako $t \in \mathbb{R}$, svako $n \in \mathbb{N}$ i svaki skup $C \subset \mathbb{R}$ sa tC označavati izraz $\{tc : c \in C\}$, a sa $C^{1/n}$ označavati $\{t > 0 : t^n \in C\}$.

Teorema 2.1. Za svako $A \in \mathcal{S}$ važi

$$A = \bigcup_{k \in \mathbb{Z}} 10^k S(A) \tag{2.6}$$

gde je $S(A) = \{S(x) : x \in A\} \subset [1, 10)$. Dodatno, važi i

$$\mathcal{S} = \mathbb{R}^+ \cap \sigma(D_1, D_2, D_3, \dots) = \left\{ \bigcup_{k \in \mathbb{Z}} 10^k B : B \in \mathcal{B}[1, 10) \right\} \tag{2.7}$$

Dokaz. Pun dokaz ove teoreme se može naći u [3], teorema 2.9. Dokaz izraza (2.6) se bazira na tome da na osnovu definicije 2.5 sledi $S\left(10^k x\right) = S(x)$ za svako $k \in \mathbb{Z}$. Izraz (2.7) dokazujemo korišćenjem osobine 2.1. □

Sada kada smo definisali oblik elemenata sigma polja $\mathcal{S}$, možemo da izvedemo neke bitne osobine, koje dajemo u lemi u nastavku.

Lema 2.1. Za sigma polje značajnog dela broja $\mathcal{S}$ važe sledeće osobine :

(i) $\mathcal{S}$ je samosličan za množenje stepenima broja 10, odnosno važi da je $10^k A = A$ za svako $A \in \mathcal{S}$ i $k \in \mathbb{Z}$.

(ii) $\mathcal{S}$ je zatvoren za operaciju množenja pozitivnim skalarima, odnosno važi $\alpha A \in \mathcal{S}$ za svako $A \in \mathcal{S}$ in $\alpha > 0$.

(iii) $\mathcal{S}$ je zatvoren za operaciju korenovanja, odnosno, važi $A^{1/n} \in S$ za svako $A \in \mathcal{S}$ i $n \in \mathbb{N}$.

Dokaz. (i) Ova sobina direktno sledi iz (2.6) jer je $S(10^k A) = S(A)$.

(ii) Na osnovu teoreme 2.1 imamo da za svako $A \in \mathcal{S}$ postoji $B \in \mathcal{B}[1, 10)$ takvo da je $A = \cup_{k \in \mathbb{Z}} 10^k B$. Na osnovu (i) možemo bez gubitka opštosti pretpostaviti da je $1 < \alpha < 10$. Tada je

$$\alpha A = \bigcup_{k \in \mathbb{Z}} 10^k \alpha B = \bigcup_{k \in \mathbb{Z}} 10^k \left((\alpha B \cap [\alpha, 10)) \cup \left(\frac{\alpha}{10} B \cap [1, \alpha) \right) \right) = \bigcup_{k \in \mathbb{Z}} 10^k C$$

gde $C \in \mathcal{B}[1, 10)$. Prema tome, $\alpha A \in \mathcal{S}$.

(iii) $\mathcal{B}[1, 10)$ možemo da definišemo kao $\mathcal{B}[1, 10) = \sigma(\{[1, 10^s] : 0 < s < 1\})$, odnosno sigma polje generisano intervalima oblika $[1, 10^s]$ gde $0 < s < 1$. Na osnovu teoreme 2.1 će biti $A = \cup_{k \in \mathbb{Z}} 10^k [1, 10^s]$ gde $0 < s < 1$. Tada je

$$A^{1/n} = \bigcup_{k \in \mathbb{Z}} 10^{k/n} \left[1, 10^{s/n}\right] = \bigcup_{k \in \mathbb{Z}} 10^k \bigcup_{j=0}^{n-1} \left[10^{j/n}, 10^{(j+s)/n}\right] = \bigcup_{k \in \mathbb{Z}} 10^k C$$

gde $C \in \mathcal{B}[1, 10)$. Prema tome, $A^{1/n} \in \mathcal{S}$.

□

Iz zatvorenosti $\mathcal{S}$ za množenje vidimo da, ukoliko definišemo meru verovatnoće na $(\mathbb{R}^+, \mathcal{S})$ tada će biti definisana i mera verovatnoće događaja koji su rezultat skaliranja.

Konačno, moguće je uspostaviti izomorfizam mere između $(\mathbb{R}^+, \mathcal{S})$ i mere verovatnoće na $([0, 1), \mathcal{B}[0, 1))$.

Lema 2.2. Funkcija $l : \mathbb{R}^+ \to [0,1)$ definisana sa $l(x) = \log S(x)$ uspostavlja *jedan-jedan* i *na* preslikavanje (izomorfizam mera) između mere verovatnoće na $(\mathbb{R}^+, \mathcal{S})$ i $([0,1), \mathcal{B}[0,1)))$.

Dokaz. Dokaz ove leme se može naći u [3], lema 2.16. □

Značaj ove leme je u tome što uspostavlja vezu $l_*\mathbb{B} = \lambda_{0,1}$, gde je $\mathbb{B}$ Benfordova mera verovatnoće, koja će biti definisana u nastavku. Ova činjenica će biti osnov za izvođenje mnogih osobina.

2.4 Definicije Benfordovih objekata

Ispostavlja se da Benfordov zakon važi za mnoge matematičke objekte kao što su mnogi poznati nizovi, funkcije, i slučajne promenljive. U ovom odeljku dajemo precizne definicije Benfodovih objekata, odnosno, objekata sa Benfordovom osobinom.

Definicija 2.8. Niz realnih brojeva (x_n) je Benfordov niz ako za svako $t \in [1,10)$ važi

$$\lim_{N\to\infty} \frac{\#\{1 \le n \le N : S(x_n) \le t\}}{N} = \log t \tag{2.8}$$

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

$$\lim_{N\to\infty} \frac{\#\{1 \le n \le N : D_j(x_n) = d_j, j = \{1,2,\ldots m\}\}}{N} = \log\left(1 + \left(\sum_{j=1}^{m} 10^{m-j} d_j\right)^{-1}\right). \tag{2.9}$$

◊

Možemo da primetimo da je ovo mnogo opštija definicija na osnovu koje je Benfordov zakon vodeće cifre specijalan slučaj u kome posmatramo situaciju $m = 1$ u alternativnoj verziji definicije.

Napomena. Ako vodeće cifre prate Benfordov zakon one nisu nezavisne. Tako, na primer, imamo $P(D_1 = 1, D_2 = 2) = \log\left(1 + \frac{1}{12}\right) = 0.035$, dok je $P(D_1 = 1) = \log(2) = 0.301$ i $P(D_2 = 2) = \sum_{1\le d_1\le 9} \log\left(1 + \frac{1}{10d_1+2}\right) = 0.109$. Prema tome, $P(D_1 = 1) \cdot P(D_2 = 2) = 0.032$, odnosno $P(D_1 = 1) \cdot P(D_2 = 2) \neq P(D_1 = 1, D_2 = 2)$. ◊

Benfordov zakon se takođe često pojavljuje kada radimo sa funkcijama koje su rešenja nekih diferencijalnih jednačina. Ovo je motivacija da definišemo Benfordove funkcije.

Definicija 2.9. Borelova[1] funkcija $f : [0,+\infty) \to \mathbb{R}$ je Benfordova funkcija ako za svako $t \in [1,10)$ važi

$$\lim_{T\to\infty} \frac{\lambda(\{\tau \in [0,T) : S(f(\tau)) \le t\})}{T} = \log t. \tag{2.10}$$

[1]Funkcija $f : \mathbb{R} \to \mathbb{R}$ je Borelova ako je $f^{-1}(I) \in \mathcal{B}$ za svako $I \subset \mathbb{R}$. Praktično, svaka funkcija sa kojom se srećemo u kontekstu Benfordovog zakona jeste Borelova funkcija.

Alternativno, funkcija je Benfordova funkcija ako 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 $j \geq 2$, važi

$$\lim_{T\to\infty} \frac{\lambda\left(\{\tau \in [0,T) : D_j\left(x_n\right) = d_j, j = \{1,2,\ldots m\}\}\right)}{T} = \log\left(1 + \left(\sum_{j=1}^{m} 10^{m-j} d_j\right)^{-1}\right). \tag{2.11}$$

$\Diamond$

Definicija 2.10. Slučajna promenljiva X na prostoru verovatnoće $(\Omega, \mathcal{A}, \mathbb{P})$ je Benfordova ako je

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

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

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

$\Diamond$

Iz definicije se lako dobija da je funkcija gustine verovatnoće značajnog dela Benfordove slučajne promenljive $\frac{1}{x\ln(10)}$.

Primer 2.3 (Benfordove slučajne promenljive)**.**

(i) Mera verovatnoće P_k sa gustinom verovatnoće $f_k(x) = 1/\left(x \ln 10\right)$ na $\left[10^k, 10^{k+1}\right)$ je Benfordova za svako $k \in \mathbb{Z}$. Nije teško videti da su mere verovatnoće sa funkcijom gustine verovatnoće oblika $\sum_{k\in\mathbb{Z}} q_k P_k$ gde $0 \leq q \leq 1$ i $\sum_{k\in\mathbb{Z}} q_k = 1$ takođe Benfordove.

(ii) Promenljiva $U \sim \mathcal{U}\left(0, 1\right)$ očigledno nije Benfordova promenljiva. Ovo se lako proverava, $P\left(S\left(X\right) \leq 2\right) = \frac{1}{9} < \log 2$. U radu [4] je data tačna granica odstupanja uniformne slučajne promenljive na bilo kom opsegu od Benfordovog zakona.

(iii) Promenljiva $X = 10^U$ je Benfordova promenljiva. Na osnovu formule (2.2) 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 značajnog dela broja za Benfordovu slučajnu promenljivu koja je generisana na ovaj način.

(iv) Benfordovu slučajnu promenljivu možemo da generišemo i korišćenjem metode odbacivanja (videti [27], odeljak 14.3.2). U tom slučaju, promenljivu generišemo na sledeći način:

(a) Generišemo slučajnu promenljivu $Y \sim \mathcal{U}\left(1, 10\right)$.

(b) Generišemo slučajnu promenljivu $U \sim \mathcal{U}\left(0, 1\right)$.

(c) Ako je $U \leq \frac{f(Y)}{cg(Y)}$ uzimamo Y, u suprotnom vraćamo se na korak 1.

Funkcija $f(x)$ je funkcija gustine verovatnoće Benfordove promenljive, odnosno $f(x) = 1/\left(x \ln x\right)$. Funkcija $g(x)$ je funkcija gustine verovatnoće uniformne raspodele na $(1, 10)$, odnosno $g(x) = 1/9$. Konstantu c biramo tako da za svako x iz domena važi $cg(x) \geq f(x)$. Obzirom da je prosečan broj eksperimenata potreban da dobijemo jednu promenljivu jednak c, najbolje je uzeti što manju vrednost koja zadovoljava ovaj uslov, na primer $c = 5$.

U prethodnim primerima su prikazana dva načina za generisanje Benfordove slučajne promenljive, ali oba načina generišu slučajnu promenljivu koja je u opsegu $[1, 10)$. Postavlja se pitanje kako generisati Benfordovu slučajnu promenljivu koja zauzima drugi opseg vrednosti. U nastavku ovog rada će biti izložene neke osobine Benfordovih promenljivih koje će dati odgovor na ovo pitanje.

Definicija 2.11. Benfordova raspodela $\mathbb{B}$ je mera verovatnoće na $(\mathbb{R}^+, \mathcal{S})$ sa osobinom

$$\mathbb{B}(S \leq t) = \mathbb{B}\left(\bigcup_{k\in\mathbb{Z}} 10^k [1, t]\right) = \log t, \quad \forall t \in [1, 10) \tag{2.12}$$

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

$$\mathbb{B}(D_j(X) = d_j, 1 \leq j \leq m) = \log\left(1 + \left(\sum_{j=1}^{m} 10^{m-j} d_j\right)^{-1}\right). \tag{2.13}$$

$\Diamond$

Sa ovako definisanom Benfordovom raspodelom možemo da se vratimo na lemu 2.2 i pokažemo da zaista važi $l_*\mathbb{B} = \lambda_{0,1}$. Neka je $B = [0, t] \in \mathcal{B}[0, 1)$, odnosno $t < 1$. Tada je

$$\begin{aligned} l_*\mathbb{B} &= \mathbb{B}\left(l^{-1}([0, t])\right) = \mathbb{B}\left(S^{-1}\left[1, 10^t\right]\right) \\ &= \mathbb{B}\left(\bigcup_{k\in\mathbb{Z}} 10^k \left[1, 10^t\right]\right) = \log 10^t = t. \end{aligned} \tag{2.14}$$

2.5 Karakterizacija na osnovu uniformne raspodele

U ovom odeljku ćemo uspostaviti vezu između Benfordove raspodele i uniformne raspodele po modulu 1, skraćeno *u.r.* mod 1. Ova veza je osnov za izvođenje najbitnijih osobina Benfordovih objekata. Osim toga, ovo je jedan od glavnih alata koji koristimo kada želimo da proverimo da li neki matematički objekat ima Benfordovu osobinu.

U nastavku ćemo sa $\langle t \rangle$ označavati deo realnog broja desno od decimalne tačke, to jest $\langle t \rangle = t - \lfloor t \rfloor$.

Definicija 2.12. Niz realnih brojeva (x_n) ima *u.r.* mod 1 ako

$$\lim_{N\to\infty} \frac{\#\{1 \leq n \leq N : \langle x_n \rangle \leq s\}}{N} = s, \quad s \in [0, 1). \tag{2.15}$$

Funkcija (sa Borelovom merom) $f : [0, +\infty) \to \mathbb{R}$ je *u.r.* mod 1 ako

$$\lim_{T\to\infty} \frac{\lambda\{\tau \in [0, T) : \langle f(\tau) \rangle \leq s\}}{T} = s, \quad \forall s \in [0, 1). \tag{2.16}$$

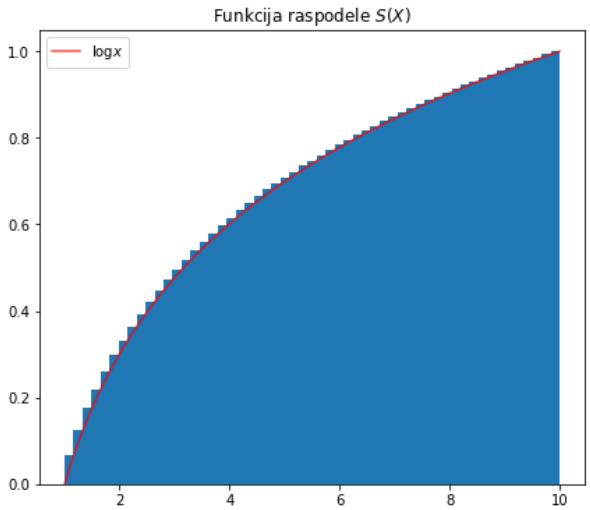
Slučajna promenljiva X na prostoru verovatnoće $(\Omega, \mathcal{A}, \mathbb{P})$ je *u.r.* mod 1 ako

$$\mathbb{P}(\langle X \rangle \leq s) = s, \quad \forall s \in [0, 1). \tag{2.17}$$

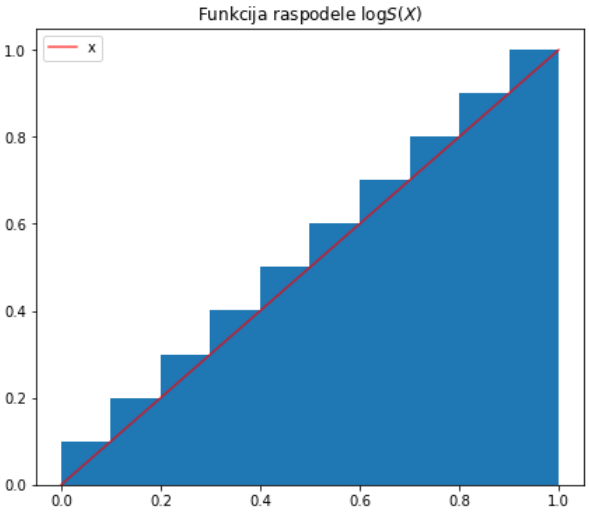
Mera verovatnoće P na $(\mathbb{R}, \mathcal{B})$ je *u.r.* mod 1 ako

$$P(\{x : \langle x \rangle \leq s\}) = P\left(\bigcup_{k\in\mathbb{Z}} [k, k + s]\right) = s, \quad \forall s \in [0, 1). \tag{2.18}$$

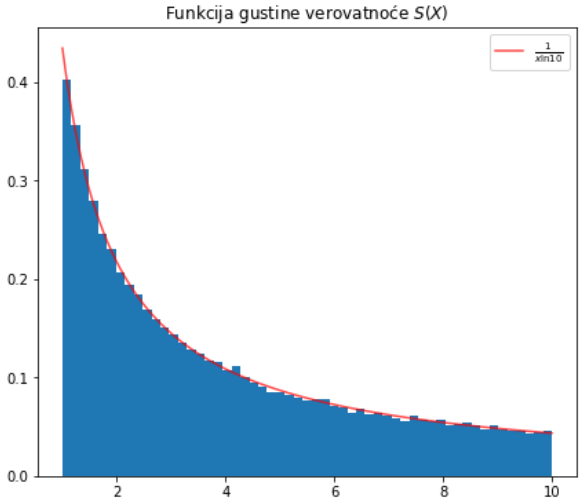
$\Diamond$



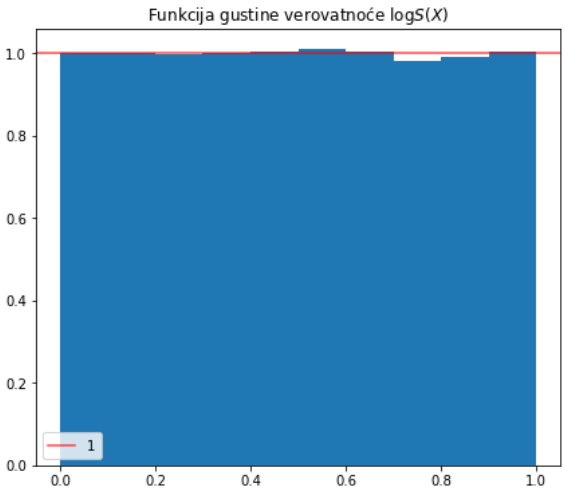
(a) Funkcija raspodele $S(X)$



(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

Teorema koju navodimo u nastavku je jedan od glavnih alata za analizu Benfordovog zakona.

Teorema 2.2 (Karakterizacija na osnovu 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. Neka je X slučajna promenljiva, i bez gubitka opštosti, pretpostavimo da je $\mathbb{P}(X = 0) = 0$. Tada, za svako $s \in [0, 1)$,

$$\begin{aligned}\mathbb{P}(\langle \log |X| \rangle \leq s) &= \mathbb{P}\left(\log |X| \in \bigcup_{k\in\mathbb{Z}} [k, k+s]\right) \\ &= \mathbb{P}\left(|X| \in \bigcup_{k\in\mathbb{Z}} \left[10^k, 10^{k+s}\right]\right) = \mathbb{P}\left(S(X) \leq 10^s\right)\end{aligned}$$

Na osnovu definicije 2.10 X je Benfordova promenljiva ako i samo ako je $\mathbb{P}\left(S(X) \leq 10^s\right) = s$, što zajedno za definicijom 2.12 dovodi do zaključka da je to ako i samo je je $\log |X|$ *u.r.* mod 1. Slično dokazujemo za nizove, funkcije i raspodele verovatnoće. □

U nastavku navodimo lemu koja daje značajne osobine objekata sa *u.r.* mod 1. Ova lema sa prethodnom teoremom daje osnov za izvođenje Benfordovih osobina.

Lema 2.3 (Definicije *u.r.* mod 1)**.**

(i) Niz (x_n) je *u.r* mod 1 ako i samo ako je niz $(kx_n + b)$ *u.r* mod 1 za svako $k \in \mathbb{Z}\setminus\{0\}$ i svako $b \in \mathbb{R}$. Takođe, (x_n) je *u.r* mod 1 ako i samo ako je (y_n) *u.r* mod 1 kad god je $\lim_{n\to\infty} |y_n - x_n| = 0$.

(ii) Funkcija f je *u.r* mod 1 ako i samo ako je $t \rightarrowtail kf(t) + b$ *u.r* mod 1 za svaki ceo broj k različit od nule i svako $b \in \mathbb{R}$.

(iii) Slučajna promenljiva X je *u.r* mod 1 ako i samo ako je $kX + b$ *u.r* mod 1 za svaki ceo broj k različit od nule i svako $b \in \mathbb{R}$.

(iv) Nz (x_n) je *u.r.* mod 1 ako i samo ako je $(x_n + \alpha \log n)$ *u.r.* mod 1. za svako $\alpha \in \mathbb{R}$.

Dokaz. Videti lemu 4.3, [3]. □

Iz teoreme 2.2 i leme 2.3 direktno slede sledeće osobine.

Osobina 2.2 (Definicije Benfordove osobine)**.**

(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) Funkcija $f : [0, +\infty) \to \mathbb{R}$ je Benfordova ako i samo ako je q/f Benfordova funkcija.

(iii) Funkcija $f : [0, +\infty) \to \mathbb{R}$ je Benfordova ako i samo ako je svako $\alpha \in \mathbb{R}$ i svako $k \in \mathbb{Z}$, gde je $\alpha k \neq 0$, $\alpha f(x)^k$ Benfordova funkcija.

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

(v) Slučajna promenljiva X je Benfordova ako i samo ako je $\alpha \in \mathbb{R}$ i svako $k \in \mathbb{Z}$, gde je $\alpha k \neq 0$, αX^k Benfordova.

U narednoj osobini ćemo dati potrebne uslove za *u.r.* mod 1.

Osobina 2.3. Neka je (x_n) niz realnih brojeva.

(i) Ako je $\lim_{n\to\infty}(x_{n+1} - x_n) = \theta$, gde je θ iracionalno, tada je (x_n) *u.r.* mod 1.

(ii) Ako je (x_n) periodično, odnosno važi $x_{n+p} = x_n$ za nako $p \in \mathbb{N}$ i svako $n \in \mathbb{N}$, tada je niz $(n\theta + x_n)$ *u.r.* mod 1 ako i samo ako je θ iracionalan broj.

(iii) Ako je niz (x_n) *u.r.* mod 1 i neopadajući, tada je $(x_n/\log n)$ neograničen niz.

U nastavku dajemo dve teoreme koje se koriste u teoriji *u.r.* mod 1 i daju način da proverimo da li neki niz ima *u.r.* mod 1.

Teorema 2.3 (Weylov kriterijum)**.** Niz (x_n) je *u.r.* mod 1 ako i samo ako je za svako $k \in \mathbb{N}$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} e^{2i\pi k x_n} = 0. \tag{2.19}$$

$\Diamond$

Teorema 2.4 (Van der Corputova procena za trigonometrijske sume)**.** Neka su $a, b \in \mathbb{Z}$, $a < b$, i neka je funkcija f dvostruko diferencijabilna na $[a, b]$ gde je $f''(x) \geq \rho > 0$ ili $f''(x) \leq -\rho < 0$ za $x \in [a, b]$. Tada je

$$\left| \sum_{n=a}^{b} e^{2i\pi f(n)} \right| \leq \left(\left| f'(a) - f'(b) \right| + 2 \right) \left(\frac{4}{\rho} + 3 \right) \tag{2.20}$$

$\Diamond$

Na osnovu izloženih osobina o uniformnoj raspodeli po modulu 1 dolazimo do veoma značajne teoreme koja na jednostavan način definiše Benfordove nizove.

Teorema 2.5. 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.

Dokaz. Na osnovu teoreme 2.2 niz $(\alpha^n a + \beta^n b)$ je Benfordov ako i samo ako je $\log|\alpha^n a + \beta^n b|$ *u.r.* mod 1.

Možemo da pišemo

$$\lim_{n\to\infty} \log|\alpha^n a + \beta^n b| - \log|\alpha^n a| = \lim_{n\to\infty} \log\left|1 + \frac{\beta^n b}{\alpha^n a}\right| = 0 \tag{2.21}$$

jer je $\alpha \neq 0$ i $|\alpha| > |\beta|$, pa je $\lim_{n\to\infty} \frac{\beta^n b}{\alpha^n a} = 0$. Odavde vidimo da je $(\log|\alpha^n a + \beta^n b|)$ *u.r.* mod 1 ako i samo ako je $(\log|\alpha^n a|) = (\log|a| + n\log|\alpha|)$ *u.r.* mod 1.

Na osnovu svojstva (i) 2.3 dati niz će biti *u.r.* mod 1 ako je $\log|\alpha|$ iracionalno. Ukoliko je $\log\alpha$ racionalan broj, tada $\langle \log|a| + n\log|\alpha| \rangle$ ima konačan broj različitih vrednosti, pa samim tim ne može da bude *u.r.* mod 1, čime je dokazana implikacija i u drugom smeru. $\square$

U nastavku dajemo teoremu koja definiše osobine slučajnih promenljivih sa *u.r.* mod 1.

Teorema 2.6. Neka su X i Y slučajne promenljive. Tada važi:

(i) Ako je X *u.r.* mod 1 i X i Y nezavisne, tada je $X + Y$ takođe *u.r.* mod 1.

(ii) Ako $\langle X \rangle$ i $\langle X + \alpha \rangle$ imaju istu raspodelu za neko iracionalno α tada je X *u.r.* mod 1.

(iii) Ako je (X_n) niz nezavisnih identično raspodeljenih slučajnih promenljivih i za X_1 važi $\mathbb{P}(X_1 \in C) < 1$ za svaki prebrojivi skup $C \subset \mathbb{R}$, tada je

$$\lim_{n\to\infty} \mathbb{P}\left(\langle \sum_{j=1}^{n} X_j \rangle \leq s\right) = s \quad \forall 0 \leq s < 1 \tag{2.22}$$

odnosno, $\langle \sum_{j=1}^{n} X_j \rangle$ konvergira u raspodeli ka $U(0,1)$ kada $n \to \infty$.

Dokaz. Dokaz ove teoreme se može naći u [3] (teorema 4.13). □

Sada na osnovu navedenih osobina i teorema možemo da ispitamo gde se sve pojavljuje Benfordova osobina. U primerima u nastavku ćemo predstaviti neke poznate matematičke objekte koji su Benfordovi i ilustrovati primenu do sada izložene teorije.

Primer 2.4 (Benfordovi nizovi)**.**

(i) 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.

(ii) Niz (2^n) nije Benfordov u brojevnom sistemu sa osnovom 8 (i uopšte sistemu koji je stepen broja 2). Sa druge strane, niz (10^n) jeste Benfordov u oktalnom brojevnom sistemu jer je $\log_8 10$ iracionalan broj.

(iii) 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 2.5 vidimo da je Fibonačijev niz Benfordov jer je $\varphi > 1$ i $\log \varphi$ iracionalan broj.

(iv) 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 $\lim_{n\to\infty} \frac{p_n}{n \log n} = 0$. Na osnovu svojstva 2.3 (iii) potreban uslov da niz (x_n) bude *u.r.* mod 1 je da je niz $(x_n \log n)$ neograničen. 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.

(v) Niz $(n!)$ je takođe Benfordov. Da bismo ovo dokazali iskoristićemo Stirlingovu formulu

$$\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n} = 1. \tag{2.23}$$

Na osnovu teoreme 2.2 i leme 2.3 (i), potreban i dovoljan uslov da dati niz bude Benfordov je da niz $\log\left(\sqrt{2\pi n}\left(\frac{n}{e}\right)^n\right)$ bude *u.r.* mod 1. Ovaj niz možemo da napišemo kao $\frac{1}{2}\log(2\pi) + \frac{1}{2}\log n + n\log\left(\frac{n}{e}\right)$. Na osnovu leme 2.3, stavke (i) možemo da izbacimo prvi član jer je konstantan. Na osnovu osobine 2.3 (iv), izbacujemo i drugi član i sada ostaje da pokažemo da je $n\log\left(\frac{n}{e}\right)$ *u.r.* mod 1.

Da bismo ovo dokazali iskoristićemo Van der Corputovu procenu i Weylov kriterijum. Imamo:

$$\begin{aligned} f'(n) &= \frac{k \ln n}{\ln 10} \\ f''(n) &= \frac{k}{n \ln 10} \end{aligned} \tag{2.24}$$

gde je $f(n) = kn \log \frac{n}{e}$. Uzećemo $\rho = \sqrt{\frac{|k|}{n \ln 10}}$. Imamo

$$\begin{aligned} \frac{1}{N} \left| \sum_{i=1}^{N} e^{2i\pi kn \log(n/e)} \right| &\leq \frac{1}{N} \left(|k| \frac{\ln N}{\ln 10} + 2 \right) \left(4 \sqrt{\frac{N \ln 10}{|k|}} + 3 \right) \\ &= \mathcal{O}\left(\frac{\ln N}{\sqrt{N}} \right). \end{aligned} \tag{2.25}$$

Odavde je na osnovu Weylov kriterijuma niz $n \log \frac{n}{e}$ *u.r.* mod 1, čime je dokazana polazna pretpostavka.

(vi) Niz (n^n) je takođe Benfordov. I u ovom slučaju se tvrđenje dokazuje korišćenjem teoreme 2.2 i pokazivanjem da je niz $n \log n$ *u.r.* mod 1. U ovom slučaju možemo direktno da primenimo Van der Corputovu procenu i Weylov kriterijum. Dokaz je jako sličan dokazu koji je dat u prethodnom primeru.

(vii) Na osnovu osobine 2.3, stavke (iii) i teoreme 2.2 imamo da je potreban uslov da niz (x_n) bude Benfordov da niz $(\log x_n / \log n)$ bude neograničen. Na osnovu ovoga lako pokazujemo da sledeći nizovi nisu Benfordovi:

$$(n^a) : \frac{\log n^a}{\log n} \to a, \quad (an) : \frac{\log an}{\log n} \to 1, \quad (\log_b n) : \frac{\log \log_b n}{\log n} \to 0 \tag{2.26}$$

Primer 2.5 (Benfordove funkcije)**.**

(i) Funkcija $f(t) = at + b$, gde su $a, b \in \mathbb{R}$ očigledno nije *u.r.* mod 1 ako je $a = 0$. Za $a > 0$ imamo da $\langle a\tau + b \rangle \leq s$ važi na intervalima oblika $\tau \in \left[\frac{k-b}{a}, \frac{k-b+s}{a} \right]$ gde je $k \in \mathbb{Z}$. Svaki od ovih intervala je iste širine s/a. Za $T > 0$ broj ovih intervala je u granicama $\lfloor aT \rfloor \pm 2$. Prema tome,

$$\frac{s}{a} \left(\lfloor aT \rfloor - 2 \right) \leq \lambda \left(\{ \tau \in [0, T) : \langle a\tau + b \rangle \leq s \} \right) \leq \frac{s}{a} \left(\lfloor aT \rfloor + 2 \right). \tag{2.27}$$

Ako podelimo gornju nejednakost sa T imamo:

$$\lim_{T \to \infty} \frac{s}{aT} \left(\lfloor aT \rfloor \pm 2 \right) = s. \tag{2.28}$$

Prema tome data funkcija je *u.r.* mod 1. Dakle, linearna funkcija nije Benfordova. Na osnovu osobine 2.2 nalazimo da ni funkcija $f(t) = q/(at + b)$ takođe nije Benfordova.

(ii) Na osnovu prethodnog primera i teoreme 2.2 nalazimo da je funkcija $f(t) = e^{\alpha t}$ Benfordova za $\alpha \neq 0$.

Glava 3

Benfordova osobina kao nulta hipoteza

U ovoj glavi razmatramo gde i kada možemo da očekujemo pojavu Benfordovog zakona. S tim u vezi, prvo ćemo da definišemo neka svojstva koja karakterišu Benfordove objekte.

3.1 Benfordova osobina i skaliranje

Intuitivno, ako neki zakon važi za podatke koje srećemo u svakodnevnom životu, očekujemo da će isti ne zavisi od toga koje smo jedinice mere koristili. Međutim, 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 skaliranja. U nastavku pokazujemo da je potreban i dovoljan uslov da ovo važi da slučajna promenljiva bude Benfordova.

Definicija 3.1. Neka je $\mathcal{A} \supset \mathcal{S}$ sigma polje na $\mathbb{R}^+$. Mera verovatnoće P na $(\mathbb{R}^+, \mathcal{A})$ je invarijanta skaliranja ako za svako $\alpha > 0$ i svako $A \in \mathcal{S}$ važi

$$P(\alpha A) = P(A) \tag{3.1}$$

odnosno, ako 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 $j \geq 2$, za svako $\alpha > 0$ važi

$$P\left(\{x : D_j(\alpha x) = d_j, 1 \leq j \leq m\}\right) = P\left(\{x : D_j(x) = d_j, 1 \leq j \leq m\}\right).$$

$\Diamond$

Sledeća teorema govori o tome da je Benfordova raspodela jedina raspodela na sigma polju značajnog dela broja koja je invarijanta skaliranja.

Teorema 3.1. Mera verovatnoće P na $(\mathbb{R}^+, \mathcal{A})$ gde je $\mathcal{A} \in \mathcal{S}$ je invarijanta skaliranja ako i samo ako je $P(A) = \mathbb{B}(A)$ za svako $A \in \mathcal{S}$, odnosno, ako i samo ako je P Benfordova mera verovatnoće.

Dokaz. Neka je P neka mera verovatnoće na $(\mathbb{R}^+, \mathcal{A})$. Sa P_0 označimo njenu restrikciju na $(\mathbb{R}^+, \mathcal{S})$. Neka je $Q := l_* P_0 = P_0\left(l^{-1}(B)\right))$ za svako $B \in \mathcal{B}$, gde je l funkcija iz leme 2.2. Na osnovu ovoga Q je mera verovatnoće na $([0,1), \mathcal{B}[0,1))$. Prema tome, za svako $A \in \mathcal{S}$ postoji $B \in \mathcal{B}$ za koje je $l(A) = B$, odnosno $A = l^{-1}(B)$.

Na osnovu jednačine (2.2) funkciju l možemo da napišemo i kao $l(x) = \langle \log |x| \rangle$. Za $l(A) = B$ će biti $l(\alpha A) = \langle \log \alpha + \log A \rangle = \langle t + B \rangle$ gde $t \in \mathbb{R}$ jer je $\alpha > 0$.

Na osnovu ovoga, formula

$$P_0(\alpha A) = P_0(A), \quad \forall\, \alpha > 0, A \in \mathcal{A} \tag{3.2}$$

je ekvivalentna formuli

$$Q\left(\langle t + B\rangle\right) = Q\left(B\right), \quad \forall\, t \in \mathbb{R}, B \in \mathcal{B}. \tag{3.3}$$

Neka je X slučajna promenljiva. Jednačina (3.3) tvrdi da je za svako $t \in \mathbb{R}$ raspodela $\langle X\rangle$ i $\langle X + t\rangle$ ista. Iz stavke (i) teoreme 2.6 imamo da je ovo tačno ako je X *u.r.* mod 1. Iz stavke (ii) iz iste teoreme imamo da jednačina implicira da je X *u.r.* mod 1. Prema tome, jednakost (3.3) je tačna ako i samo ako je $Q = \lambda_{0,1}$.

Konačno, na osnovu (2.14) zaključujemo da je $P_0 = \mathbb{B}$. □

Napomena. Osobina invarijante skaliranja za Benfordove nizove, funkcije i slučajne promenljive je data u osobini 2.2.

Primer 3.1. U [32] se predlaže test koji proverava da li podaci prate Benfordov zakon tako što se proverava da li je frekvencija pojavljivanja jedinice kao vodeće cifre jednaka $\log 2$ i nakon skaliranja. Napomenimo da nije dovoljno podatke skalirati jednim brojem, jer je moguće da raspodela vodećih cifara ostane ista. Zbog toga se ovaj test bazira na uzastopnom množenju nekom konstantom veliki broj puta.

U primerima 2.3 je pokazano kako se može generisati Benfordova slučajna promenljiva. U oba primera je slučajna promenljiva bila u opsegu $[1, 10)$. Sada kada smo pokazali da će raspodela cifara ostati ista, jasno je da je moguće generisati slučajnu promenljivu i u proizvoljnom opsegu oblika $[\alpha, 10\alpha)$. Ovo zapažanje možemo formalno da definišemo kao osobinu koju dajemo u nastavku.

Osobina 3.1. Fiksirajmo $b > a > 0$. Tada je:

(i) Ako je $b < 10a$, ne postoji Benfordova slučajna promenljiva X takva da važi $X \in [a, b]$;

(ii) Ako je $b = 10a$, tada postoji tačno Benfordova slučajna promenljiva u opsegu $[a, b]$;

(iii) Ako je $b > 10a > 0$, tada za svako $c \in (0, 0.1b - a)$, postoji slučajna promenljiva $X_c = (a + c)\, 10^U$ koja ima opseg koji je u $[a, b]$ i Benfordova je. Takođe, za svako $c \in (0, b - 10a)$, postoji slučajna promenljiva $Y_c = U\left[a + c, 10a + c\right]$ koja zauzima opseg koji je sadržan u $[a, b]$ i nije Benfordova.

Dokaz. Osobina (i) je očigledno. Jasno je da u tom slučaju postoje vodeće cifre koje se uopšte ne pojavljuju, pa slučajna promenljiva koja uzima vrednosti iz tog opsega ne može biti Benfordova.

Osobina (ii) sledi iz osobine skaliranja i teoreme 2.2.

Osobina (iii) je posledica činjenice da je $X = 10^U$ Benfordova slučajna promenljiva (primer 2.3). Kako je Benfordova osobina invarijanta za skaliranje, i slučajne promenljive oblika $(a + c)\, 10^U$ jesu Benfordove. Drugi deo osobine (iii) je posledica toga što nijedna slučajna promenljiva sa uniformnom raspodelom nije Benfordova, kao što je pokazano u primeru 2.3. □

Prema tome, ako je opseg slučajne promenljive manji od jednog reda veličine, ne možemo očekivati pojavu Benfordovog zakona. Sa druge strane, ako je opseg makar malo veći od jednog reda veličine, postoji beskonačno mnogo slučajnih promenljivih iz tog opsega koje su Benfordove.

Dakle, da bi slučajna promenljiva imala Benfordovu raspodelu nije potrebno da zauzima opseg od nekoliko reda veličina. Često se može naći ideja da je neophodno da podaci zauzimaju veliki opseg da bi se mogla očekivati pojava Benfordovog zakona. Kao što se vidi iz datog svojstva i osobine invarijantnosti skaliranja, ovo nije tačno. Ova greška i njena propagacija u literaturi i

popularnoj nauci je detaljno diskutovana u radu [17]. U primerima u odeljku 2.5 smo videli da su mnogi geometrijski nizovi i eksponencijalne funkcije zapravo Benfordovi, i kao takvi, delimično uzrok pojave Benfordovog zakona u praksi. Ovo je verovatno jedan od glavnih razloga za to što se javila ideja o tome da podaci moraju da zauzimaju veliki opseg.

3.2 Benfordova osobina u različitim brojevnim sistemima

Pre nego što je teorija o Benfordovom zakonu bila razrađena, bilo je argumenata da je pojava Benfordovog zakona posledica toga što koristimo brojevni sistem sa osnovom 10. U ovom odeljku ćemo analizirati Benfordovu osobinu u različitim brojevnim sistemima.

Rad [3] daje sledeću definiciju mere verovatnoće značajnog dela broja koja je invarijanta za promenu brojevnog sistema.

Definicija 3.2. Neka je $\mathcal{A} \supset \mathcal{S}$ sigma polje na $\mathbb{R}^+$. Mera verovatnoće P na $(\mathbb{R}^+, \mathcal{A})$ je invarijanta za promenu brojevnog sistema ako za svako $\mathcal{A} \in \mathcal{S}$, i svako $n \in \mathbb{N}$ važi $P(A) = P\left(A^{1/n}\right)$. $\Diamond$

U nastavku ćemo probati da objasnimo motivaciju iza ovakve definicije. Nije teško videti da se teorema 2.1 i lema 2.1 lako proširuju i na druge brojevne sisteme. U sigma polju značajnog dela broja sa osnovom b, događaji su oblika $A = \bigcup_{k\in\mathbb{Z}} b^k B$, gde je $B \in \mathcal{B}[1, b)$. Odnosno, $A = \bigcup_{k\in\mathbb{Z}} b^k [b^x, b^y)$, gde su $x, y \in [0, 1]$ i podrazumevamo $x < y$.

Tada će događaj $A^{1/n}$ biti

$$A^{1/n} = \bigcup_{k\in\mathbb{Z}} b^k \left(\bigcup_{j=0}^{n-1} b^{j/n} \left[b^{x/n}, b^{y/n} \right) \right) \tag{3.4}$$

Ako uvedemo smenu $b_{1/n} = b^{1/n}$ dobijamo:

$$A^{1/n} = \bigcup_{k\in\mathbb{Z}} b^k \left(\bigcup_{j=0}^{n-1} b_{1/n}^j \left[b_{1/n}^x, b_{1/n}^y \right) \right) = \bigcup_{k\in\mathbb{Z}} b_{1/n}^k \left[b_{1/n}^x, b_{1/n}^y \right) = A_{1/n}. \tag{3.5}$$

Odavde vidimo da je događaj A u brojevnom sistemu sa osnovom b ekvivalentan događaju $A^{1/n}$ u brojevnom sistemu sa osnovom b^n. Prema tome, ako želimo meru verovatnoće koja je ista i kada promenimo brojevni sistem, prirodno je da za bar ove događaje očekujemo da ta mera bude ista.

Benfordov unuk Frank Benford je, na primer, zamerio definiciju invarijante brojevnog sistema koja je data iznad, odnosno, smatra da korišćenje naziva invarijanta brojevnog sistema, obzirom da se radi samo o brojevnim sistemima koji su stepeni originalnog brojevnog sistema nije opravdano [1].

Teorema 3.2. Mera verovatnoće P na $(\mathbb{R}^+, \mathcal{A})$, gde je $\mathcal{A} \supset \mathcal{S}$, je invarijanta u odnosu na brojevni sistem kada su u pitanju vodeće cifre ako i samo ako je za neko $q \in [0, 1]$

$$P(A) = q\delta_1(A) + (1 - q)\,\mathbb{B}(A) \tag{3.6}$$

za svako $A \in \mathcal{S}$. Sa δ_1 smo označili Dirakovu meru verovatnoće koncentrisanu u tački 1, definisanu sa:

$$\delta_a = \begin{cases} 1, & a \in A \\ 0, & \text{inače.} \end{cases} \tag{3.7}$$

Dokaz. Kompletan dokaz ove teoreme se može naći u [3], teorema 4.30. Ovde ćemo ukratko pokazati da je Benfordova mera verovatnoće invarijanta u odnosu na brojevni sistem.

Za $A = \bigcup_{k\in\mathbb{Z}} 10^k\,[1, 10^s]$, $0 < s < 1$, na osnovu definicije 2.11 imamo $\mathbb{B}(A) = s$. Sa druge strane je:

$$\begin{aligned}\mathbb{B}\left(A^{1/n}\right) &= \bigcup_{k\in\mathbb{Z}} 10^k \bigcup_{j=0}^{n-1} \left[10^{j/n}, 10^{(j+s)/n}\right] = \sum_{j=0}^{n-1} \mathbb{B}\left(\bigcup_{k\in\mathbb{Z}} \left[10^{j/n}, 10^{(j+s)/n}\right]\right) \\ &= \sum_{j=0}^{n-1} \left(\log 10^{(j+s)/n} - \log 10^{j/n}\right) = \sum_{j=0}^{n-1} \left((j+s)/n - j/n\right) = s = \mathbb{B}\left(A\right)\end{aligned} \tag{3.8}$$

□

Iz ove teoreme sledi da je jedina neprekidna mera verovatnoće koja je invarijanta za brojevni sistem upravo Benfordova mera verovatnoće.

Napomena. Iz teoreme 3.2 sledi da osobina invarijante skaliranja implicira osobinu invarijante brojevnog sistema. Obrnuto ne važi jer Dirakova mera verovatnoće nije invarijanta u odnosu na skaliranje. ◊

U skladu sa prethodno navedenom zamerkom u vezi sa definicijom invarijante promene brojevnog sistema navodimo jedan primer koji ilustruje problem promene brojevnog sistema i Benfordove osobine.

Primer 3.2. Videli smo da je slučajna promenljiva $X = 10^U$ Benfordova u dekadnom brojevnom sistemu. Međutim, u brojevnom sistemu sa osnovom 8 to nije slučaj. Vidimo da je $\log_8 10^U = U \ln 10/\ln 8 = 1.107U$, i ova slučajna promenljiva očigledno nije *u.r.* mod 1 (mada jeste jako bliska tome). Slično, promenljiva 8^U jeste Benfordova u oktalnom brojevnom sistemu, ali ne i u sistemu sa osnovom 10, jer je $\log 8^U = U \log 8 = 0.903U$.

U radovima [34] i [1] je veoma detaljno analizirano šta se dešava sa Benfordovom osobinom u različitim brojevnim sistemima. Jasno je da, ukoliko je neka slučajna promenljiva Benfordova u brojevnom sistemu sa osnovom b, kraće b-Benfordova, to ne mora biti slučaj u i kada koristimo drugu osnovu. Međutim, ispostavlja se da u ovakvim situacijama dobijamo promenljive koje su veoma bliske Benfordovoj. Ovaj rezultat je predstavljen u radu [34], teoremi 30, koju navodimo u nastavku.

Teorema 3.3. Neka je $X = \beta^{U[c,d]}$ slučajna promenljiva, gde je $1 < \beta$ i $c < d$. Tada je

$$\{b \in (1,\infty) : X \text{ je b-Benfordova } \} = \left\{\beta^{\frac{d-c}{n}} : n \in \mathbb{N}\right\} \tag{3.9}$$

i za svako $b > 1$ je

$$\sup_{A\in\mathcal{B}[0,1)} \left|P\left(\log_b X \bmod 1 \in A\right) - \lambda_{0,1}\left(A\right)\right| \leq \frac{1}{(d-c)\log_b \beta} \tag{3.10}$$

Dokaz. Videti [34], teoremu 30. □

Ova teorema je verovatno još jedan od izvora već pomenute česte greške koja tvrdi da podaci moraju zauzimati veliki opseg, odnosno red veličina, da bi bili Benfordovi. Iz (3.10) vidimo da je supremum greške obrnuto proporcionalan širini uniformne raspodele i veličini osnove. Prema tome, ako imamo podatke koji su ovog oblika, gde stepen biramo slučajno iz uniformne raspodele, nije teško doći u situaciju da isti deluju kao da imaju Benfordovu osobinu.

Sličnu situaciju imamo i kada smo u domenu deterministčkih procesa. Na primer, niz $\left(10^{n/100}\right)$ nije Benfordov na osnovu teoreme 2.5, međutim odstupanje od Benfordove raspodele je jako malo.

Ono što je najproblematičnije u ovim primerima je to što su ove promenljive oblika geometrijskih nizova, za koje uglavnom očekujemo pojavu Benfordove osobine, i to dodatno komplikuje testiranje.

3.3 Benfordova osobina i sumiranje značajnih delova

Nigrini je primetio i delimično dokazao, da, ako podaci prate Benfordov zakon, tada je suma značajnih delova koji počinju jedinicom jednaka sumi značajnih delova koji počinju dvojkom itd.

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

$$S_{d_1,\ldots,d_m}(x) := \begin{cases} S(x), & \text{ako } (D_1(x),\ldots,D_m(x)) = (d_1,\ldots,d_m) \\ 0, & \text{inače} \end{cases} \tag{3.11}$$

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

Definicija 3.3. Niz realnih brojeva (x_n) je invarijanta u odnosu na sumiranje značajnih delova ako za svako $m \in \mathbb{N}$ limes

$$\lim_{N\to\infty} \frac{\sum_{n=1}^{N} S_{d_1,\ldots,d_m}(x_n)}{N} \tag{3.12}$$

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

Definicija 3.4. Slučajna promenljiva X je invarijanta u odnosu na sumiranje značajnih delova 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. Slučajna promenljiva X sa $\mathbb{P}(X=0)=0$ je invarijanta u odnosu na sumiranje značajnih delova ako i samo ako je Benfordova.

Dokaz. Kompletan dokaz teoreme se može naći u [3], teorema 4.37. U nastavku pokazujemo da je Benfordova slučajna promenljiva zaista invarijanta u odnosu na sumiranje značajnih delova.

Ako je X Benfordova slučajna promenljiva, važi:

$$\mathbb{E}S_{d_1,\ldots,d_m}(X) = \int_{d_1+10^{-1}d_2+\cdots+10^{1-m}d_m}^{d_1+10^{-1}d_2+\cdots+10^{1-m}(d_m+1)} t\frac{1}{t\ln 10}\mathrm{d}t = \frac{10^{1-m}}{\ln 10}. \tag{3.13}$$

Prema tome, $\mathbb{E}S_{d_1,\ldots,d_m}(X)$ ne zavisi od $d_1,\ldots,d_m$. □

Primer 3.3. U [30] je ova osobina iskorišćena kao jedna vrsta testa za proveru da li podaci poštuju Benfordov zakon. Test se bazira na tome da saberemo značajne delove koji počinju istom cifrom i uporedimo sume.

3.4 Benfordova osobina i stohastički procesi

U odeljku 2.5 smo videli dosta primera Benfordovih nizova i funkcija, primetivši da se mnogi od njih pojavljuju u svakodnevnom životu. U ovom odeljku razmatramo pojavu Benfordove osobine u u okviru stohastičkih procesa. Ovde ujedno i završavamo teorijsku analizu Benfordovog zakona tako što navodimo dve teoreme koje su analogne centralnoj graničnoj teoremi.

Prva teorema govori kada kombinacija uzoraka iz slučajno izabranih raspodela teži Benfordovoj raspodeli. Druga teorema govori o tome da kombinacija uniformnih raspodela teži Benfordovoj, ukoliko je širina tih uniformnih raspodela Benfordova slučajna promenljiva.

Definicija 3.5. Beskonačan niz slučajnih promenljivih $(X_1, X_2, \dots)$ konvergira u raspodeli ka Benfordovom zakonu ako

$$\lim_{n\to\infty} P\left(S\left(X_n\right) \leq t\right) = \log t \quad \forall\, t \in [1, 10)\,. \tag{3.14}$$

Ovaj niz je Benfordov za verovatnoćom 1 ako

$$P\left((X_n) \text{je Benfordov niz }\right) = 1 \tag{3.15}$$

$\Diamond$

Teorema 3.5. Ako je X neprekidna promenljiva, tada (X^n) konvergira u raspodeli ka Benfordovom zakonu i Benfordova je sa verovatnoćom 1.

Dokaz. Da bismo dokazali da (X^n) konvergira u raspodeli ka Benfordovom zakonu koristimo teoremu koju navodimo u nastavku ([3], teorema 4.17). Ako X ima gustinu, tada je

$$\lim_{n\to\infty} \mathbb{P}\left(\langle nX\rangle \leq s\right) = s \tag{3.16}$$

za svako $0 \leq s < 1$. Na osnovu ove teoreme je

$$\mathbb{P}\left(S\left(X\right) \leq t\right) = \mathbb{P}\left(\langle \log|X^n|\rangle \leq \log t\right) = \mathbb{P}\left(\langle n \log|X|\rangle \leq \log t\right) \to \log t \tag{3.17}$$

kada $n \to \infty$.

Što se tiče drugog dela tvrđenja, na osnovu teoreme 2.5 dati niz će biti Benfordov ako je $\log|X|$ iracionalno. Poznato je da je verovatnoća ovog događaja nula. $\square$

Teorema 3.6. Neka su X i Y dve nezavisne promenljive sa $\mathbb{P}\left(XY = 0\right) = 0$. Tada važi:

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

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

Dokaz. Dokaz ove teoreme se može naći u [3], teorema 6.3. $\square$

Primetimo da ova teorema kaže da se Benfordova osobina propagira dalje kada množimo slučajne promenljive. Prema tome, ako je neka pojava posledica proizvoda slučajnih promenljivih, dovoljno je da jedna od njih bude Benfordova da bi i posmatrana pojava bila Benfordova.

Teorema 3.7. 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. Videti teoremu 6.6. u [3]. $\square$

Konačno, daćemo teoremu koja je analogna centralnoj graničnoj teoremi. Da bi smo definisali ovu teoremu, potrebno je prvo definisati slučajnu meru verovatnoće i pojam nepomerene mere verovatnoće u odnosu na skaliranje i promenu brojevnog sistema[1].

Definicija 3.6. Slučajna mera verovatnoće $\mathbb{P}$ je slučajna promenljiva čije su vrednosti mere verovatnoće na $\mathbb{R}$. $\Diamond$

[1]Koristimo definiciju invarijante promene brojevnog sistema koju smo dali u 3.2

Primer 3.4. Primer slučajne mere verovatnoće $\mathbb{P}$ može da bude mera verovatnoće iz skupa $\{\mathcal{U}[0,1], \mathcal{N}(0,1)\}$, i te vrednosti uzima sa verovatnoćama $\{1/3, 2/3\}$. Odnosno, sa verovatnoćom $1/3$ generišemo slučajnu promenljivu sa uniformnom raspodelom, to jest $\mathbb{P}$ je mera uniformne raspodele, a sa verovatnoćom $2/3$ slučajnu promenljivu sa normalnom raspodelom, odnosno $\mathbb{P}$ je mera normalne raspodele.

Definicija 3.7. 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. Slučajna 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 brojevnog sistema. ◊

Primer 3.5. Za slučajnu meru verovatnoće iz prethodnog primera, prosečna mera verovatnoće je funkcija raspodele slučajne promenljive X sa funkcijom gustine verovatnoće $\frac{1}{3}+\frac{2}{3}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)$ za $0 < x < 1$ i $\frac{2}{3}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)$ za $x > 1$.

Primetimo da je jedina mera verovatnoće koja je istovremeno nepomerena u odnosu na brojevni sistem i skaliranje upravo Benfrodova mera verovatnoće. Prema tome, slučajna mera verovatnoće je nepomerena istovremeno u odnosu na skaliranje i promenu brojevnog sistema ako je njena prosečna mera verovatnoće Benfordova. U teoremi koju dajemo u nastavku će biti precizno formulisano, da ako je prosečna mera verovatnoće slučajne mere verovatnoće Benfordova, tada uzorak formiran od slučajnih promenljivih sa tom slučajnom merom verovatnoće konvergira ka Benfordovom zakonu.

Teorema 3.8. Neka je $\mathbb{P}$ slučajna mera verovatnoće takva da $\mathbb{P}(S \in \{0,1\}) = 0$ sa verovatnoćom 1. Neka je $P_1, P_2, \ldots$ niz nezavisnih identičnih slučajnih mera verovatnoće iz $\mathbb{P}$. Fiksirajmo prirodan broj m i neka $X_1, X_2, \ldots, X_m$ bude slučajni uzorak veličine m iz P_1, neka je $X_{m+1}, \ldots, X_{2m}$ 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 $X_1, X_2, \ldots, X_m, X_{m+1}, \ldots$ konvergira ka Benfordovom zakonu sa verovatnoćom jedan, to jest

$$P\left(\lim_{N\to\infty}\frac{\#\{1 \le n \le N \ : \ S(X_n) \le t\}}{N} = \log t \ \forall \ t \in [q, 10)\right) = 1 \tag{3.18}$$

Dokaz. Videti teoremu 6.20 u [3]. □

Na kraju navodimo teoremu koja je predstavljena u radu [19] (teorema 2.1) koja povezuje kombinaciju uniformnih raspodela i Benfordovu osobinu.

Teorema 3.9. Benfordova mera verovatnoće je jedina mera verovatnoće μ na $[1, 10)$ takva da, ako $S(Y)$ ima meru verovatnoće μ, istu meru ima i $S(X)$, gde je $X \sim \mathcal{U}[0, Y]$.

U ovom odeljku završavamo sa izlaganjem teorije Benfordovog zakona. Na kraju navodimo neka otvorena pitanja koja se prirodno javljaju nakon ove analize.

(i) Da li je teorema 3.9 samo specijalan slučaj teoreme 3.8?

(ii) U teoremi 3.8 je navedeno da uzorci teže Benfordovoj raspodeli ako je prosečna mera verovatnoće iz koje dolaze Benfordova. Koje to slučajne mere verovatnoće zadovoljavaju ove osobine i, još bitnije, kakav je odnos između njih i realnih mera koje srećemo u praksi?

(iii) Osim Benfordovih slučajnih promenljivih postoje brojne slučajne promenljive koje su približno Benfordove, i one će biti analizirane detaljno u narednoj glavi. Da li je moguće izvedenu teoriju generalizovati i na mere verovatnoće koje su sa nekom zadovoljavajućom tačnošću Benfordove? Kako se onda one uklapaju u teoremu 3.8? Kako se u sve to uklapa izabrani brojevni sistem?

(iv) Da li je moguće konstruisati mere verovatnoće čije je odstupanje od Benfordove mere kontinualno? Odnosno, da li je moguće konstruisati prostor mera verovatnoća u kome se možemo kretati i približavati, ili se udaljavati od Benfordove mere? Da li bi ovakav prostor mogao da objedini i prave Benfordove slučajne promenljive sa približno Benfordovim?Ako bi ovakav prostor mogao da se konstruiše, kako bi izgledala generalizacija teoreme 3.8?

Glava 4

Testiranje Benfordovog zakona

U prethodnoj glavi smo izneli teorijske rezultate koji bi trebalo da nam pomognu da shvatimo gde možemo da očekujemo pojavu Benfordovog zakona. Nažalost, iako smo u prethodnom odeljku pokazali kako se može dobiti Benfordova osobina, mogli smo i da naslutimo neke probleme.

Tako, na primer, odeljku 3.4 su date teoreme koje govore o tome da se Benfordova osobina može javiti kada kombinujemo podatke iz različitih raspodela ili uniformnih raspodela. Međutim, primetimo da ove teoreme imaju dosta ograničenja, čija pretpostavka, u opštem slučaju, može biti previše stroga da bi se olako pretpostavila u praksi. U teoremi 3.3 smo videli da mnoge slučajne promenljive koje nisu Benfordove u bazi b ipak mogu da budu jako bliske Benfordovim. U nastavku ćemo pokazati da mnoge raspodele poput normalne i eksponencijalne nisu Benfordove, ali je njihovo odstupanje od Benfordove prilično malo. Dakle, odgovor na pitanje da li bi podaci uopšte trebalo da prate Benfordovu raspodelu je dosta komplikovaniji nego što se čini na prvi pogled. U ovoj glavi ćemo predstaviti najčešće korišćene statistike, razmatrati probleme koji postoje u vezi sa njima i predložiti neka rešenja.

4.1 Najčešće korišćene statistike

Statistike koje se najčešće koriste za testiranje hipoteze o Benfordovoj raspodeli su Pirsonova statistika ili χ^2 test, Kolmogorov-Smirnov test, Kuiperov test i m i d statistike.

4.1.1 Pirsonov χ^2 test

Pirsonov test je jedan od najčešće korišćenih testova uopšte. Ovaj test je jako zgodan, pogotovu kada su podaci diskretne prirode. U nastavku dajemo teoremu na osnovu koje definišemo Pirsonov test [27].

Teorema 4.1. 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

$$\chi^2 = \sum_{j=0}^{r} \frac{(O_j - np_j)^2}{np_j} \tag{4.1}$$

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 4.1.

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

$$r = 1 + 3.3 \log n. \tag{4.2}$$

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 $n = 266$.

4.1.2 Kolmogorov-Smirnov test

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

Teorema 4.2. Neka je n veličina uzorka, $F = F_0$ nulta hipoteza (gde je F_0 neprekidna funkcija), a F_n empirijska raspodela. Tada statistika

$$\lambda = \sqrt{N} \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)| \tag{4.3}$$

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. Ovaj test se primenjuje i kada su funkcije raspodele diskretne (iako se teorema oslanja na to da je F_0 neprekidna). U radu [29] je diskutovana primena Kolmogorov-Smirnov testa na diskrenim funkcijama raspodele i pokazano je da se u tom slučaju mogu koristiti i manje stroge vrednosti statistika. 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 4.1, gde su statistike za diskretan slučaj posebno označene.

4.1.3 Kuiperov test

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

$$V_n = \left(D_n^+ + D_n^-\right)\left(\sqrt{n} + 0.155 + \frac{0.24}{\sqrt{n}}\right) \tag{4.4}$$

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 [29]. U tabeli 4.1 su date značajne vrednosti ove statistike.

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

4.1.4 m-statistika i d-statistika

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

Definicija 4.1. U kontekstu Benfordovog zakona, m-statistiku definišemo kao:

$$m_n = \sqrt{n} \max_{d \in \{1,\ldots,9\}} |\Pr(D_0(X) = d) - \log(1 + 1/d)| \tag{4.5}$$

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

Definicija 4.2. U kontekstu Benforfovog zakona, d-statistiku definišemo kao:

$$d_n = \sqrt{n}\left(\sum_{d\in\{1,\ldots,9\}} \left(\Pr\left(D_0\left(X\right)=d\right) - \log\left(1+1/d\right)\right)^2\right)^{1/2} \tag{4.6}$$

$\Diamond$

U ovom slučaju se pojavljivanje svake cifre smatra 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 4.1 su date brojne vrednosti za najčešće korišćene nivoe značajnosti.

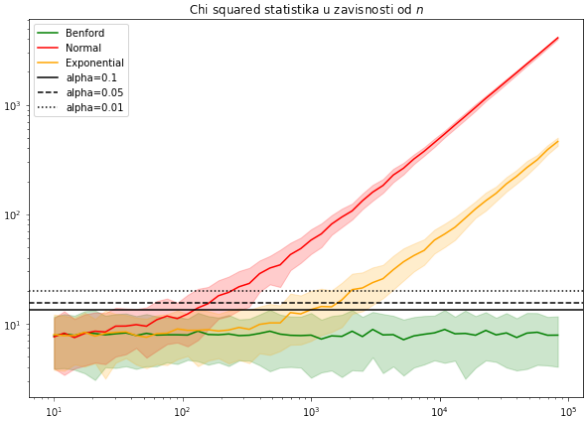
4.1.5 Značajne vrednosti statistika

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

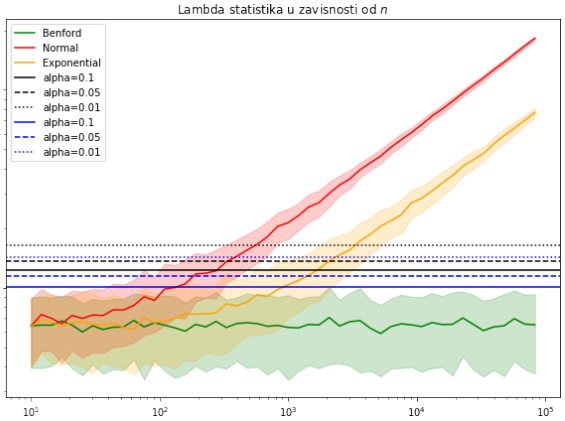
Da bi se stekla dodatna predstava o tome kako se različite statistike ponašaju u zavisnosti od veličine uzorka u odnosu na različite nivoe značajnosti generisano je po 10^5 uzoraka Benfordovih, eksponencijalnih i normalno raspodeljenih slučajnih promenljivih za različite vrednosti n i rezultati su prikazani na slici 4.1. Normalna i eksponencijalna raspodela su izabrane jer se radi o raspodelama koje su jako bliske Benfordovoj, što se vidi na slikama.

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

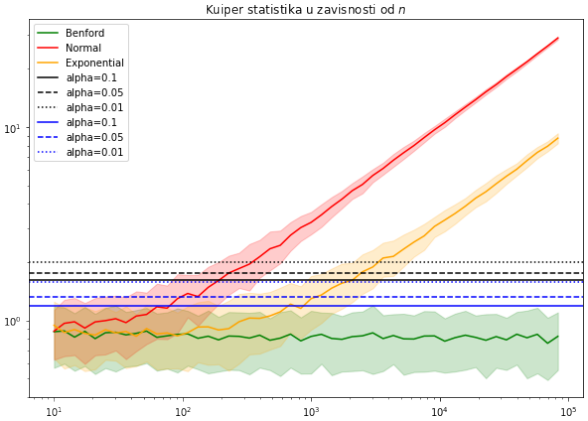
Tabela 4.1: Značajne vrednosti statistika u kontekstu testiranja Benfordovog zakona



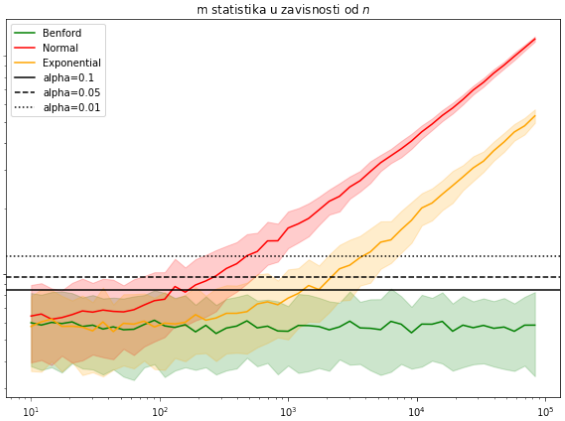
(a) χ^2 statistika



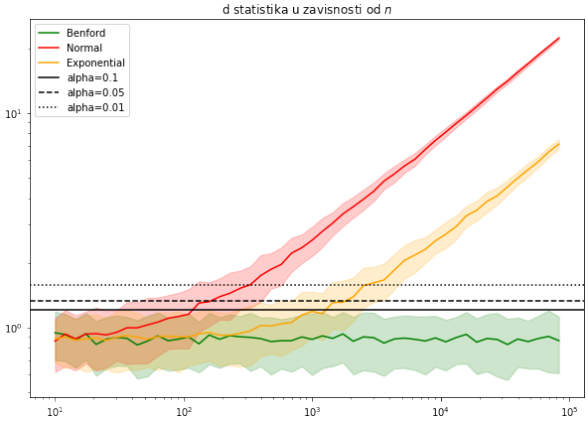
(b) λ statistika



(c) V_n statistika



(d) m_n statistika



(e) d_n statistika

Slika 4.1: Zavisnost različitih statistika od veličine uzorka. Punom linijom je prikazana prosečna vrednost, a svetlijom bojom je označena standardna devijacija. Linijama su označene značajne vrednosti.

4.2 Problemi sa najčešće korišćenim statistikama

Testovi predstavljeni u prethodnom odeljku nisu inače problematični i redovno se koriste, štaviše daju dobre rezultate u testiranju hipoteza. Osim toga, imaju dobru teorijsku podlogu. Međutim, kada primenjujemo ove testove u kontekstu Benfordovog zakona nailazimo na dosta problema. U nastavku ćemo probati da izložimo ove probleme i objasnimo njihov uzrok, pa na osnovu toga predložiti drugačiji način za testiranje Benfordovog zakona.

4.2.1 Približno Benfordove slučajne promenljive

Mnoge slučajne promenljive koje dolaze iz poznatih raspodela mogu biti približno Benfordove u smislu da, pogotovo ako koristimo mali broj odbiraka, njihov značajni deo ima raspodelu koja jako liči na Benfordovu. Ova sličnost je još više izražena ako izaberemo da posmatramo samo vodeće cifre.

Na slici 4.1 smo mogli da vidimo da se predstavljene statistike suštinski malo razlikuju. Stoga ćemo u nastavku probleme ilustrovati kroz korišćenje χ^2 statistike. Na slikama 4.2 i 4.3 su prikazani rezultati simulacije u kojoj je korišćeno 1000 odbiraka iz razlučitih raspodela koji daju raspodelu značajnog dela broja koja je veoma bliska Benfordovoj. Na slici 4.4 je pokazano da čak i kada koristimo 100 000 odbiraka, nije teško dobiti rezultate gde je odstupanje od Benfordovog zakona malo. Ove promenljive su uzete iz raspodela koje su približno Benfordove i potom skalirane. Dakle, ne možemo ih potpuno razlikovati ni na osnovu osobine skaliranja jer i tada ostaju približno Benfordove.

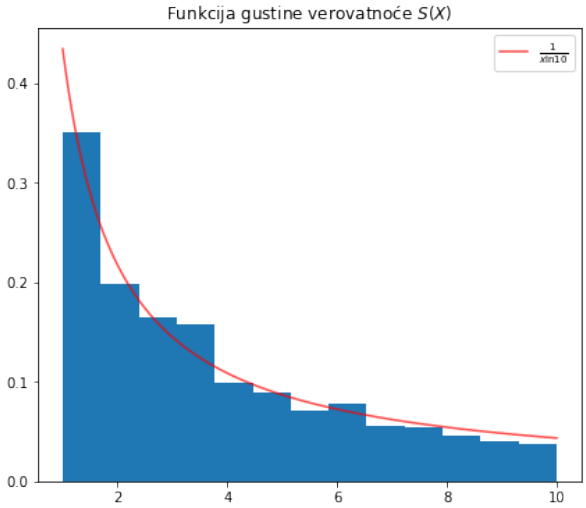
U radu [24] se kao dva osnovna zahteva za Benfordovu osobinu navode opseg podataka i oblik histograma podataka. Kao uslov za opseg podataka se navodi da je potrebno da on bude reda veličine bar 3. U osobini 3.1 smo videli i dokazali da je potreban uslov da podaci budu Benfordovi da zauzimaju opseg od bar jednog reda veličine, i da je tvrđenje da podaci moraju zauzimati veliki opseg veličina netačno, bar iz čisto teorijske perspektive. Uslov o obliku histograma govori o tome da se Benfordov zakon može očekivati tamo gde je matematičko očekivanje veće od medijane, odnosno, tamo gde je histogram ima pozitivni pomeraj. Ove heuristike se mogu naći u većini radova koji se bave primenom Benfordovog zakona i prilično su ustaljene. Ovakve heuristike dovode upravo do približno Benfordovih slučajnih promenljivih. Primetimo da ovo nije posledica šuma u podacima, obzirom da radimo sa značajnim delovima brojeva, šum u ovom kontekstu ima veoma mali uticaj.

Teorija Benfordovog zakona i primena su se prilično nezavisno razvijale. Ovo je verovatno glavni razlog što, kada malo detaljnije proučimo Benfordov zakon nailazimo na to da teorija i praksa ustvari, pod pojmom Benfordove raspodele, ne smatraju potpuno iste stvari. Ovo samo po sebi nije problem, niti je pojava koja se ovde prvi put pojavljuje. Međutim, problem nastaje onda kada pokušavamo da primenimo rezultate iz teorije u praksi, kao što je testiranje hipoteza.

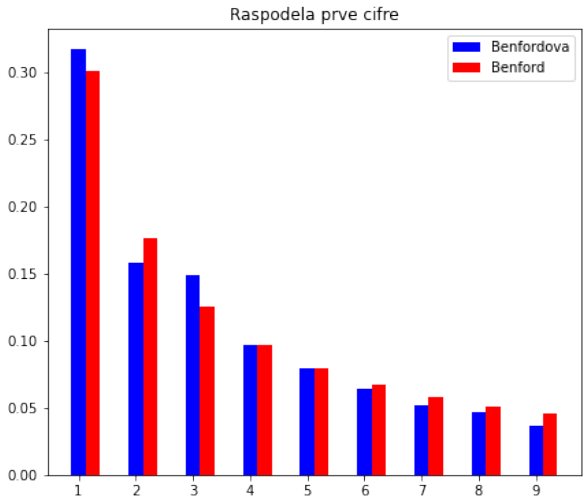
U radovima poput [18, 29] je, na primer, pomenuto kako je χ^2 statistika previše stroga, pogotovo kada broj odbiraka raste, i da odbacuje hipotezu o Benfordovom zakonu i onda kada podaci vizuelno odgovaraju istoj. Ovo je još jedno tvrđenje u vezi sa Benfordovim zakonom koje je problematično. Ne radi se o tome da χ^2 statistika ima neke mane koje se jedino primećuju u testiranju Benfordovog zakona, već je problem u tome što mi u praksi srećemo gotovo isključivo približno Benfordove promenljive ili uopšte ne ispunjavamo pretpostavke koje postoje u izvođenju, a primenjujemo rigorozno testiranje hipoteza. Dakle, nije problem u alatu koji koristimo, već to što primenjujemo alat koji nije odgovarajući problemu sa kojim se susrećemo.

4.2.2 Benfordove promenljive iz različitih izvora

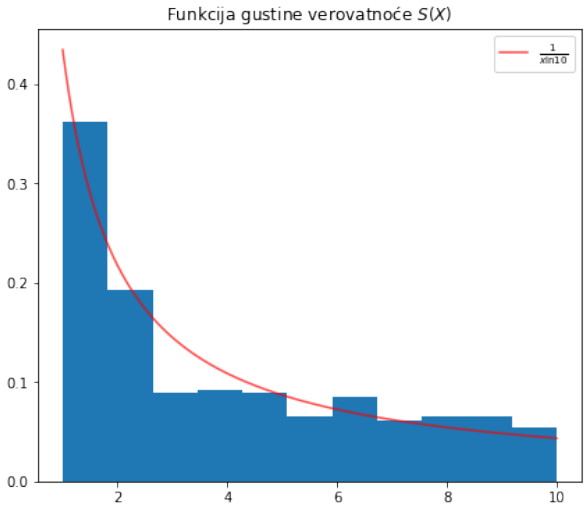
Ono što Benfordove slučajne promenljive razlikuje od standarnih slučajnih promenljivih je činjenica da verovatnoću definišemo na sigma algebri značajnog dela broja, a promenljive



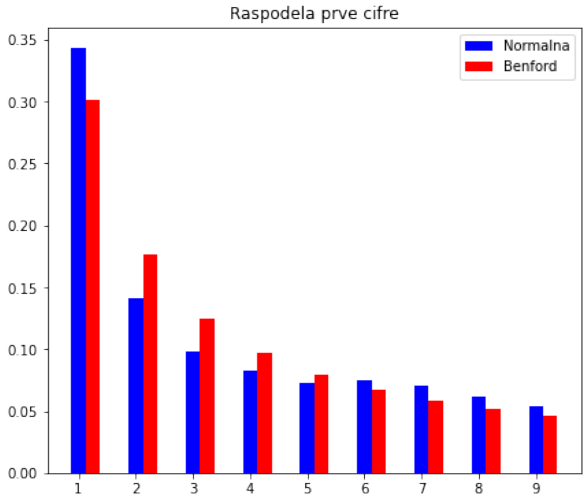
(a) Fgv $S(X)$, $X \sim$ Benford



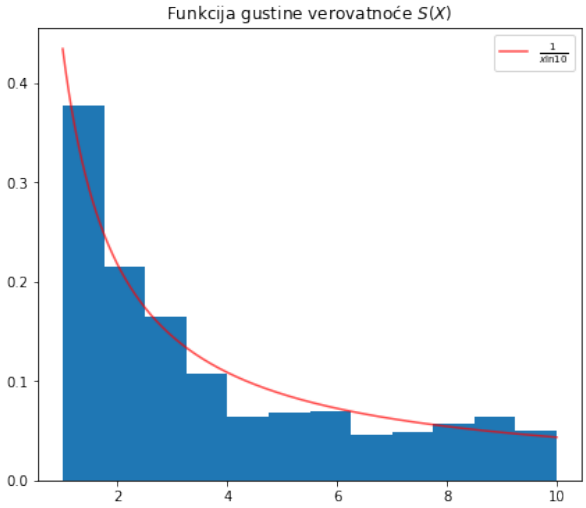
(b) Raspodela vodećih cifara X, $X \sim$ Benford



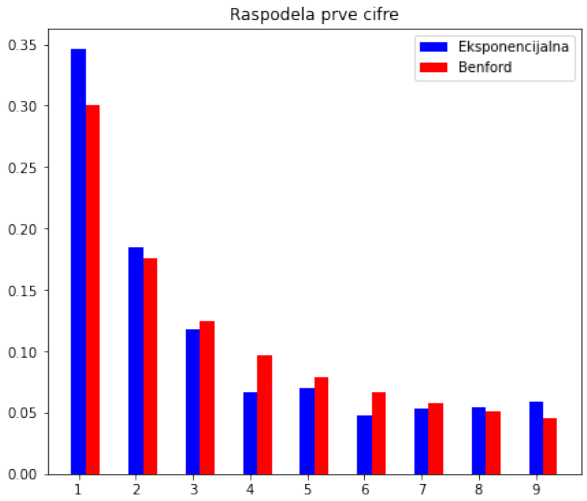
(c) Fgv $S(X)$, $X \sim \mathcal{N}(0,1)$



(d) Raspodela vodećih cifara X, $X \sim \mathcal{N}(0,1)$

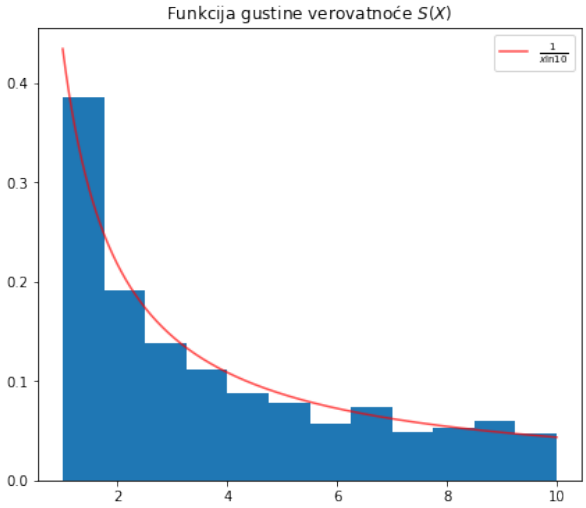


(e) Fgv $S(X)$, $X \sim \text{Exp}(1)$

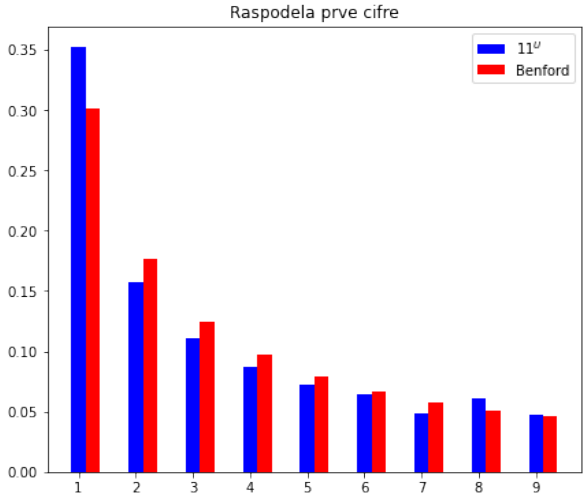


(f) Raspodela vodećih cifara X, $X \sim \text{Exp}(1)$

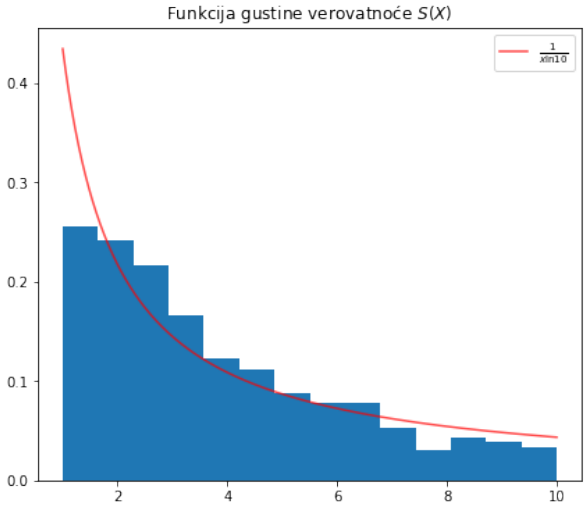
Slika 4.2: Funkcija gustine verovatnoće značajnog dela i raspodele vodećih cifara za Benfordovu slučajnu promenljivu, slučajnu promenljivu iz normalne raspodele i slučajnu promenljivu iz eksponencijalne raspodele. U simulaciji je korišćeno 1000 odbiraka.



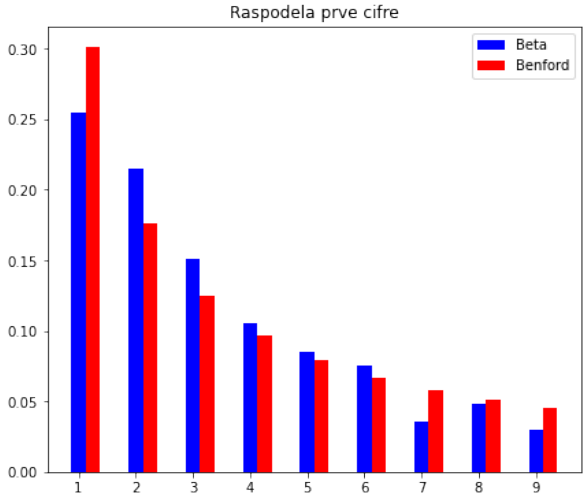
(a) Fgv $S(X)$, $X = 11^U$



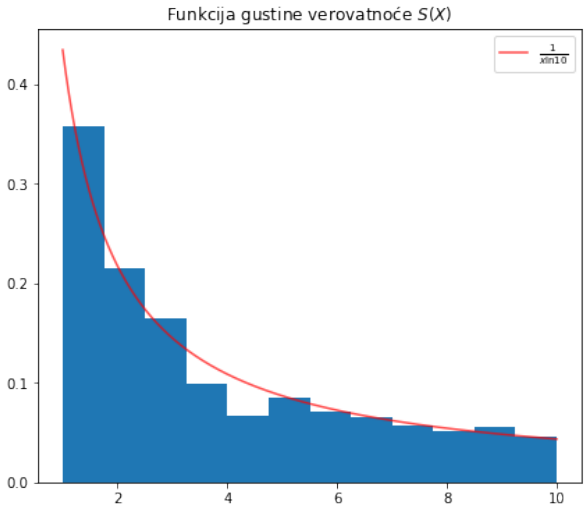
(b) Raspodela vodećih cifara X, $X = 11^U$



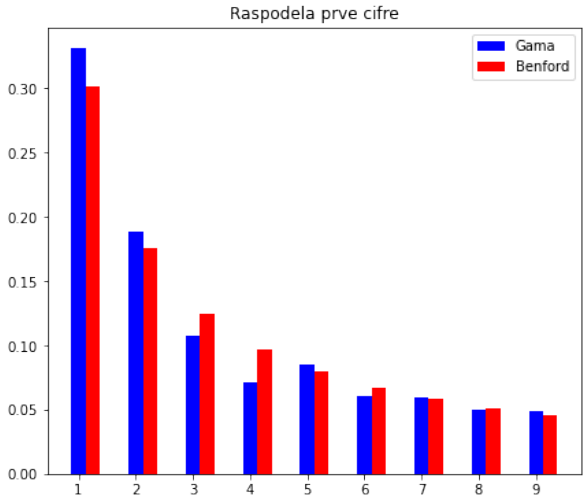
(c) Fgv $S(X)$, $X \sim B(1,3)$



(d) Raspodela vodećih cifara X, $X \sim B(1,3)$

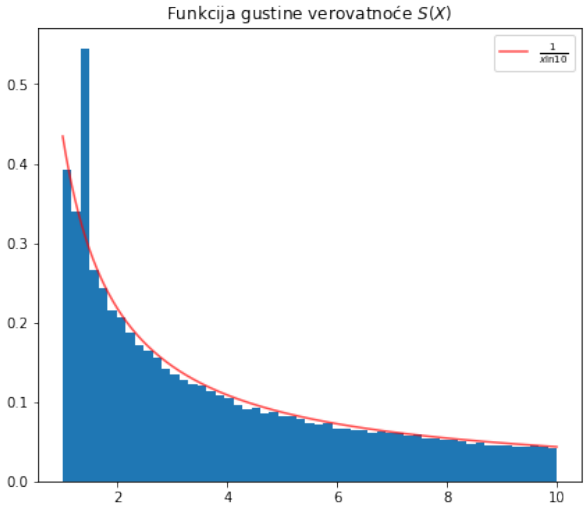


(e) Fgv $S(X)$, $X \sim \Gamma(1,1)$

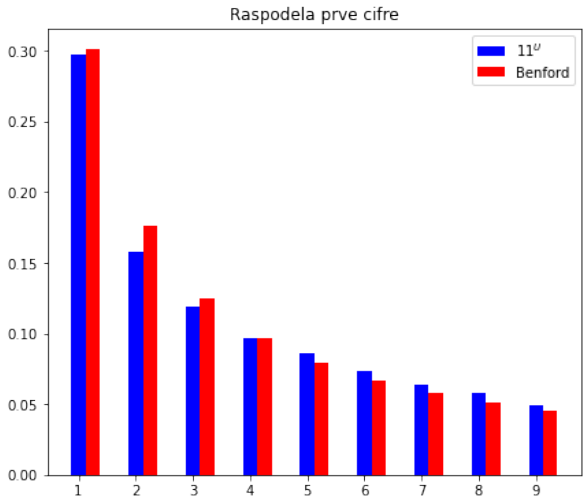


(f) Raspodela vodećih cifara X, $X \sim \Gamma(1,1)$

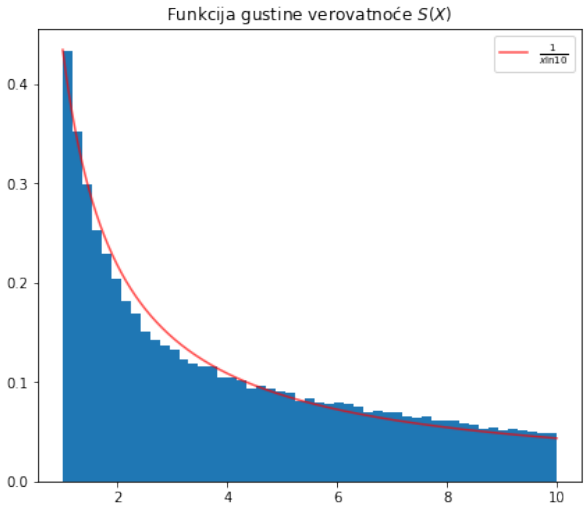
Slika 4.3: Funkcija gustine verovatnoće značajnog dela i raspodele vodećih cifara za slučajne promenljive $X = 11^U$, $X \sim B(1,3)$ i $X \sim \Gamma(1,1)$. U simulaciji je korišćeno 1000 odbiraka.



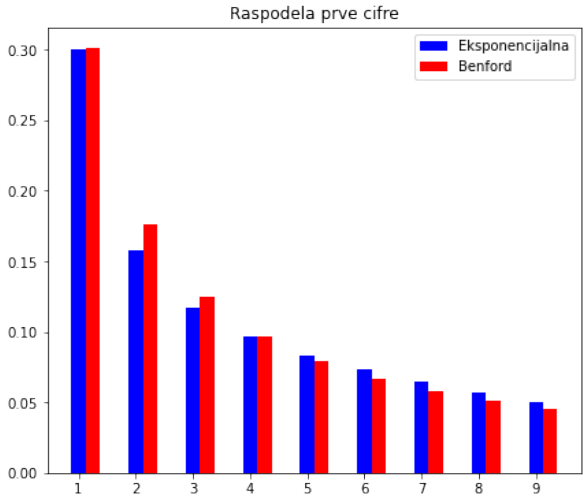
(a) Fgv $S(X)$, $X = 11^U$



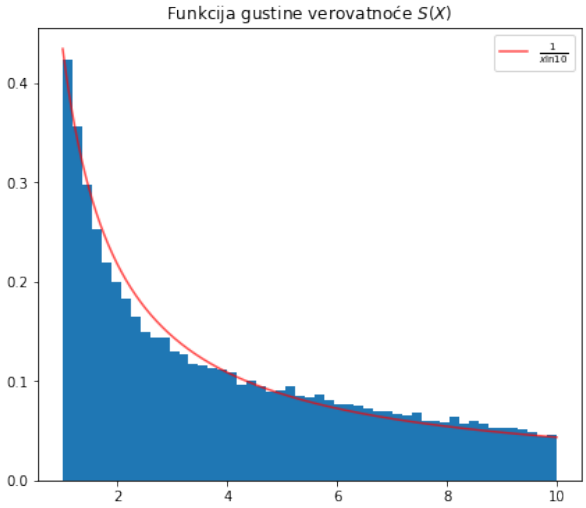
(b) Raspodela vodećih cifara X, $X = 11^U$



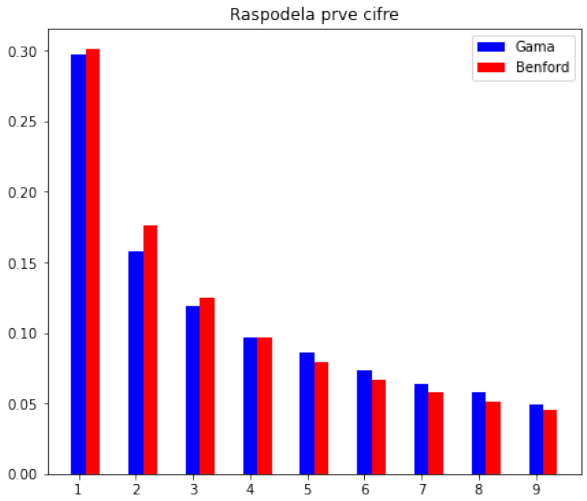
(c) Fgv $S(X)$, $X \sim \text{Exp}(1)$



(d) Raspodela vodećih cifara X, $X \sim \text{Exp}(1)$

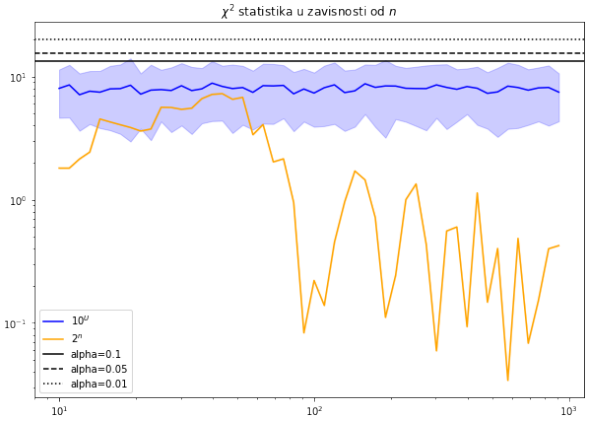


(e) Fgv $S(X)$, $X \sim \Gamma(1,1)$

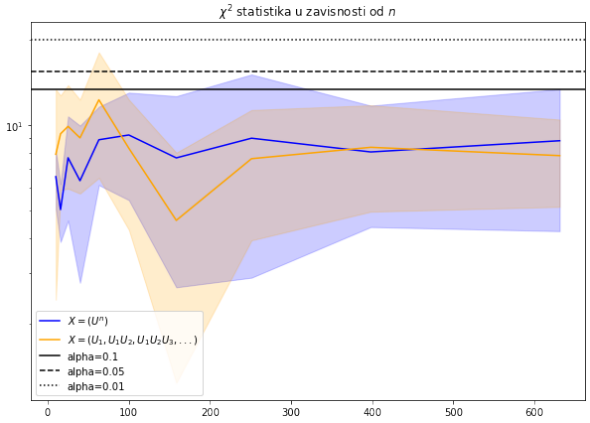


(f) Raspodela vodećih cifara X, $X \sim \Gamma(1,1)$

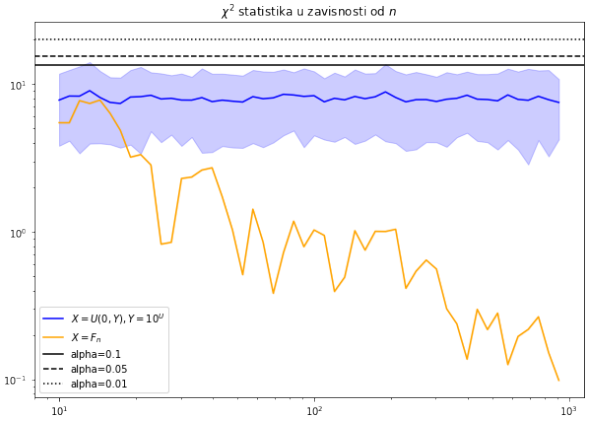
Slika 4.4: Funkcija gustine verovatnoće značajnog dela i raspodele vodećih cifara za slučajne promenljive $X = 11^U$, $X \sim \text{Exp}(1)$ i $X \sim \Gamma(1,1)$. Promenljive su skalirane brojem iz $\mathcal{U}[1,10]$. U simulaciji je korišćeno 100 000 odbiraka.



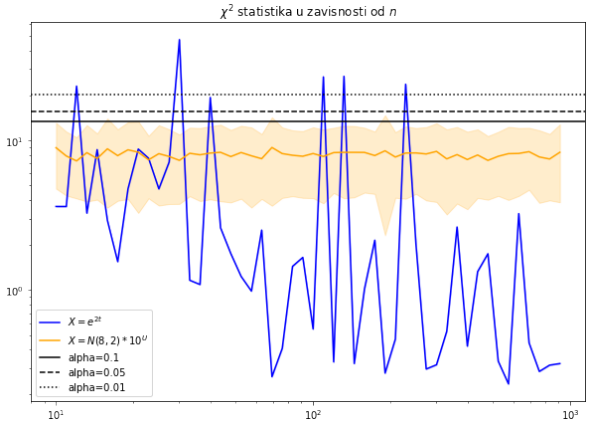
(a) Benfordova slučajna promenljiva i geometrijski niz 2^n



(b) Benfordova slučajna promenljiva i niz $(U_1, U_1U_2, \dots)$



(c) Benfordova slučajna promenljiva i Fibonačijev niz



(d) Benfordova slučajna promenljiva i eksponencijalna funkcija

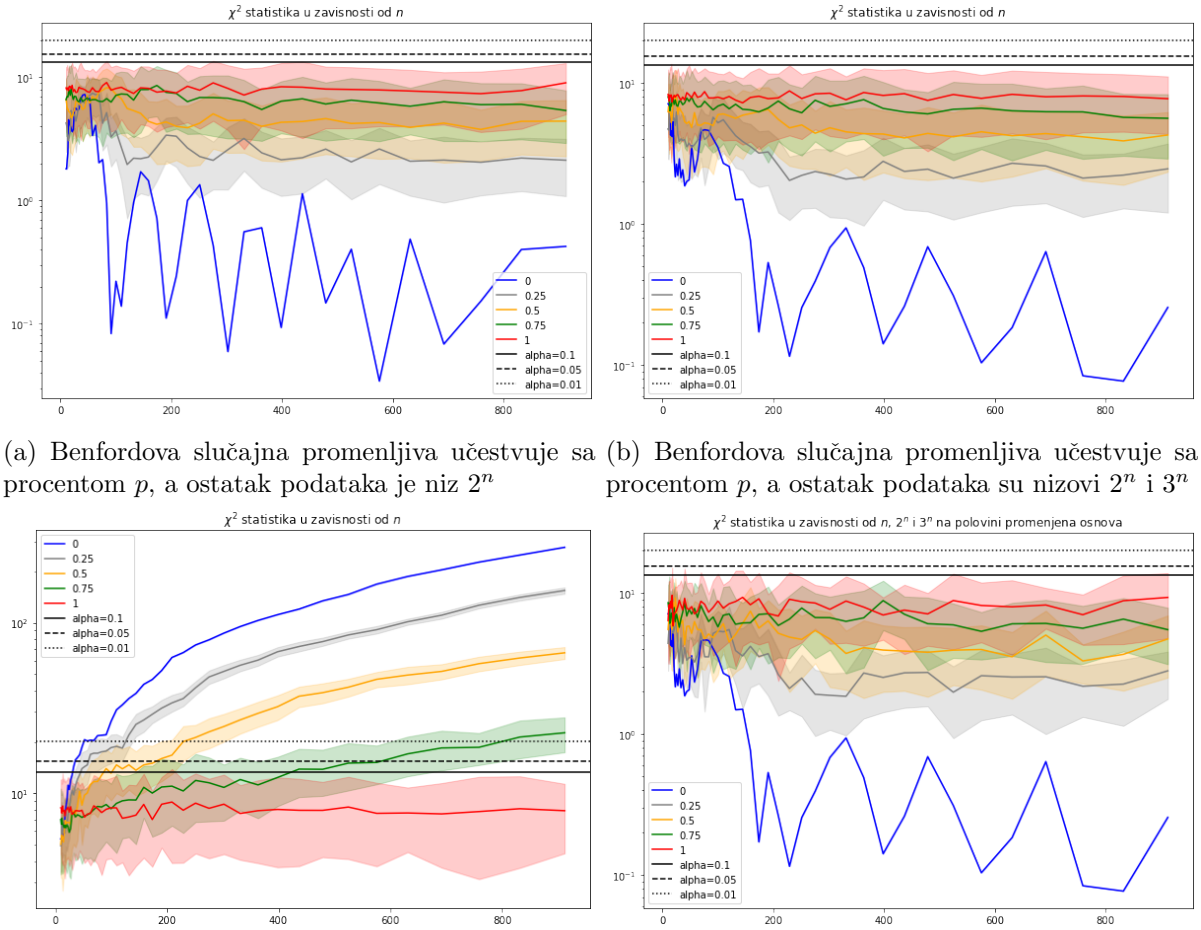
Slika 4.5: Zavisnost χ^2 statistike od broja odbiraka za Benfordove promenljive iz različitih izvora

generišemo iz skupa $\mathbb{R}$. Zbog toga je moguće imati Benfordove slučajne promenljive koje dolaze iz različitih izvora i u skladu sa tim se i drugačije ponašaju. Osim toga, postoje Benfordove promenljive koje su posledica determinističkih procesa, kao što je na primer, Fibonačijev niz. Iako sa teorijske tačke gledišta vrlo lako možemo da razdvojimo determinističke od stohastičkih procesa, u praksi nije jednostavno razlikovati ove pojave i često ih je teško razdvojiti u podacima.

Na slici 4.5 je prikazano kako se ponaša χ^2 statistika kada povećavamo broj odbiraka. Korišćeni su različiti načini za generisanje Benfordovih slučajnih promenljivih – slučajna promenljiva $X = 10^U$ za koju je u primeru 2.3 (i) pokazano da je Benfordova, $X = (U^n)$ na osnovu teoreme 3.5, $X = (U_1, U_1U_2, U_1U_2U_3, \dots)$ na osnovu teoreme 3.7, $X = U[0, Y]$, gde je $Y = 10^U$ kao što je pokazano u teoremi 3.6 i slučajna promenljiva $X = Y10^U$ gde je $Y \sim \mathcal{N}(8, 2)$ na osnovu osobine (i) teoreme 3.6. Za ove slučajne promenljive su prikazane srednje vrednosti statistike i standardne devijacije koje su označene svetlijom bojom.

Osim slučajnih promenljivih prikazane su i vrednosti χ^2 statistike za nizove 2^n i F_n, kao i za funkciju $f(t) = e^t$. Opseg koji prikazujemo na grafiku je posledica ograničene tačnosti i maksimalnih vrednosti sa kojima možemo da radimo na računaru (obzirom da ove promenljive jako brzo rastu).

Ono što možemo da vidimo na graficima je da geometrijski i Fibonačijev niz jako brzo dostižu veoma malu vrednost statistike. Eksponencijalna funkcija takođe ima manje odstupanje, ali kao što vidimo na slici, ono dosta osciluje, za razliku od nizova gde se greška prilično sigurno



(a) Benfordova slučajna promenljiva učestvuje sa procentom p, a ostatak podataka je niz 2^n

(b) Benfordova slučajna promenljiva učestvuje sa procentom p, a ostatak podataka su nizovi 2^n i 3^n

(c) Benfordova slučajna promenljiva učestvuje sa procentom p, a ostatak podataka su nazmenični nizovi 2^n i 3^n

(d) Benfordova slučajna promenljiva učestvuje sa procentom p, a ostatak podataka je geometrijski niz gde je na polovini promenjena osnova sa 2 na 3

Slika 4.6: Zavisnost χ^2 statistike od broja odbiraka

smanjuje.

Dati rezultat nije čudan. Mnogo manja vrednost χ^2 statistike je posledica toga što ove promenljive nisu slučajne. Iz čisto teorijske perspektive, problema nema i rezultati su potpuno očekivani. Međutim, iz perspektive prakse gde je potrebno testirati Benfordov zakon, i samim tim izabrati značajne vrednosti statistika, ova činjenica uvodi dodatne probleme.

Na slici 4.6 je prikazano šta se dešava sa χ^2 statistikom kada formiramo skup podataka u kome sa određenim procentom učestvuje Benfordova slučajna promenljiva, a ostatak čini geometrijski niz 2^n. Sa slike se vidi da smanjenje determinizma u podacima dovodi to porasta ove statistike, što je i očekivani rezultat. Na slici takođe možemo da primetimo da se χ^2 statistika smanjuje kada uvodimo Benfordovu slučajnu promenljivu u podatke koji nisu Benfordovi. Ovo je veoma interesantna pojava koja nije tako česta kada radimo sa drugim raspodelama, jer je u tom slučaju prosto potencijalni domen funkcije gustine verovatnoće mnogo veći. Konačno, na ovoj slici možemo da vidimo da χ^2 statistika uopšte nije problematična, štaviše jako je precizna i uspeva da razlikuje Benfrodovu raspodelu od raspodela koje su joj veoma slične.

4.2.3 Konceptualni problemi sa najčešće korišćenim statistikama

Neki od konceptualnih problema koji postoje su već objašnjeni u prethodnim odeljcima, gde je glavni problem prilično slobodno pretpostavljanje Benfordovog zakona kao nulte hipoteze

onda kada podaci nisu teorijski Benfordovi, već samo približno, a potom primena rigoroznih statističkih testova za testiranje hipoteze. Potrebno je uskladiti ova dva – ili primeniti veću slobodu u pri testiranju, tako da i podaci koji vizuelno deluju kao Benfrodovi prolaze testove, ili prihvatiti da je pojava Bendfordovog zakona u teorijskom smislu retka i odbacivati većinu hipoteza. Obzirom na samu istoriju Benfordovog zakona, ali i na široku primenu, mnogo bolja opcija jeste uvesti testove koji nisu tako strogi i koji će davati rezultat koji je u skladu sa vizuelnim, odnosno uskladiti teoriju sa utvrđenom primenom, umesto odbacivanja utvrđene primene radi usklađivanja sa teorijom.

Drugi konceptualni problem, koji smo takođe predstavili u prethodnom odeljku, je sama činjenica da primenjujemo testiranje hipoteza na podacima koji su možda potpuno ili delimično deterministički.

Naredni problem je u tome što su sve statistike koje smo naveli do sada baziraju na centralnoj graničnoj teoremi i ideji da su uzorci nezavisni i dolaze iz nekog većeg skupa koji ima raspodelu koja je naša nulta hipoteza. Upravo zato one postaju sve strože kako broj odbiraka raste – što je više odbiraka to će njihova raspodela ličiti na onu iz koje su uzeti.

Za razliku od centralne granične teoreme, teorema 3.8 koja je njoj analogna, ima mnogo veća ograničenja. Naime, podaci iz različitih raspodela imaju Benfordovu meru verovatnoće ako su te raspodele u proseku Benfordove. Ova teorema se navodi kao objašenjenje za to što podaci dobijeni iz različitih raspodela često poštuju Benfordov zakon, ali ovakvo objašnjenje nije potpuno. Ovde se prirodno postavlja pitanje koje su to slučajne mere verovatnoće koje su Benfordove i koliko često se one pojavljuju u praksi.

Konačno, uzorci koje imamo kada želimo da primenimo Benfordov zakon su retko nezavisni, samim tim pretpostavke na osnovu kojih izvodimo statistike nisu opravdane.

4.3 Relativna entropija kao mera usaglašenosti sa Benfordovim zakonom

Obirom na predstavljene probleme u vezi sa često korišćenim statistikama, predlažemo korišćenje relativne entropije kao meru za procenu usaglašenosti sa Benfordovim zakonom. U radu [24] je predstavljena slična ideja. Sličnu metriku koja takođe ne zavisi od broja odbiraka - srednje apsolutno odstupanje (*eng.* MAD) daje Nigrini u knjizi [30].

Relativna entropija, ili Kullback–Leibler divergencija je mera odstupanja jedne funkcije gustine verovatnoće od referentne funkcije gustine verovatnoće, i ova mera ne zavisi od broja odbiraka i daje rezultate koji su u skladu sa vizuelnim očekivanjima. Za razliku od pomenutih metrika, ova je pomnožena verovatnoćom koja se očekuje, pa se samim tim daje veća težina odstupanjima na onim ciframa koje su i verovatnije. U nastavku dajemo definiciju relativne entropije.

Definicija 4.3. Za zakone raspodele P i Q definisane na domenu D, relativna entropija Q u odnosnu na referentni zakon raspodele P je data izrazom

$$D_{KL}\left(P||Q\right) = \sum_{x \in D} P(x) \log\left(\frac{P\left(x\right)}{Q\left(x\right)}\right). \tag{4.7}$$

$\Diamond$

U definiciji je moguće koristiti logaritam sa različitim osnovama, ali obzirom da smo u okviru Benfordovog zakona, prirodno je uzeti dekadni logaritam. Uvek će važiti $D_{KL}\left(P||Q\right) \geq 0$ (videti Gibsovu nejednakost), pri čemu je relativna entropija veoma bliska nuli onda kada su P i Q gotovo istovetne. U kontekstu Bajesovskog rezonovanja, relativna entropija se može shvatiti kao

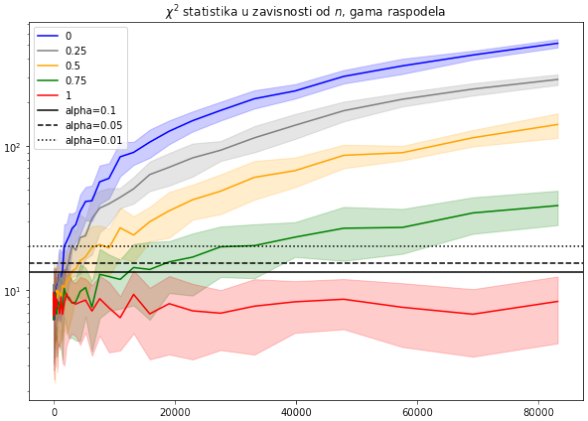
količina informacija koja je izgubljena time što se zakon raspodele Q koristi za aproksimaciju zakona raspodele P. Zbog toga je izabrano da Q bude Benfordova raspodela, a P stvarna raspodela u eksperimentima.

Na slikama 4.7, 4.8, 4.9 je prikazana zavisnost relativne entropije od broja odbiraka. Ono što možemo da primetimo je da mnoge približno Benfordove slučajne promenljive imaju tendenciju da uđu u zasićenje, odnosno, imaju konstantno odstupanje od Benfordovog zakona. Ovo odstupanje je uglavnom veoma malo. Na slici 4.10 su prikazane raspodele vodećih cifara i vrednosti relativne entropije. Vidimo da je relativna entropija reda veličine između 10^{-3} i 10^{-2} granica do koje podaci vizuelno i dalje odgovaraju Benfordovom zakonu. Primetimo da korišćenjem neke od ustaljenih statistika hipotezu o Benfordovom zakonu u ovim slučajevima treba odbaciti.

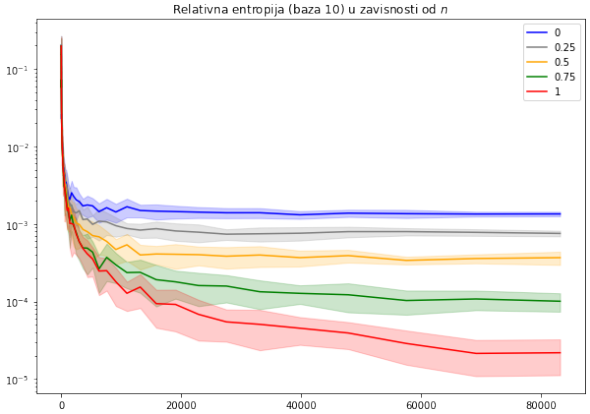
Osim što relativna entropija mnogo bolje ilustruje ono što intuitivno očekujemo, već za 200-300 odbiraka jasno razlikujemo one promenljive koje su bliske Benfordovom zakonu od onih koje nisu.

Zbog svega ovoga, relativna entropija, ili MAD deluju kao bolje metrike nego neke od korišćenih statistika koje se često navode reda radi, dok se zaključak uglavom bazira na vizuelnoj proceni koja nije potpuno objektivna. U nekim radovima se nailazi na rezonovanje u kome se usaglašenost sa Benfordovim zakonom vizuelno procenjuje, a onda se na osnovu ove procene prihvataju ili odbijaju dobijene statistike. Relativna entropija je objektivna mera odstupanja od Benfordove raspodele, a daje rezulte koji su u skladu sa intuitivnom i vizuelnom procenom. U isto vreme izbegavamo sve one probleme koji postoje kada primenjujemo testiranje hipoteza, a postavljanje nulte hipoteze i njeno testiranje radimo u na istom nivou tačnosti.

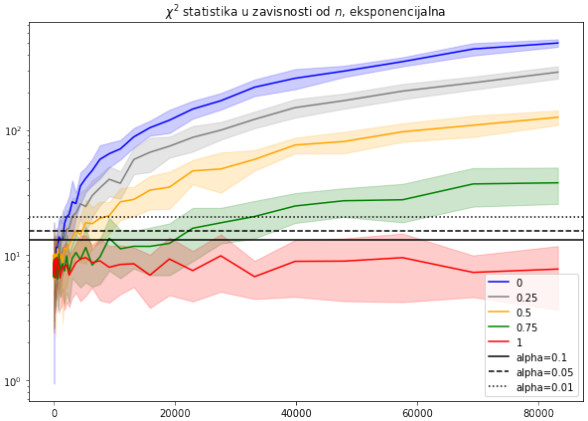
U ovoj glavi smo videli da postoje mnoge promenljive koje su približno Benfordove u pravom smislu te reči - imaju raspodelu značajnog dela koja liči na Benfordovu čak i nakon skaliranja. Videli smo da relativna entropija ovakvih promenljivih ima tendenciju da bude ograničena, pa bi upravo ona mogla da se iskoristi za definisanje istih. Osim toga, videli smo da dodavanje Benfordovih slučajnih promenljivih u skupove podataka ima tendenciju da raspodelu približi Benfordovoj. Ovo tendencija, zajedno sa množenjem bi mogla da se iskoristi da se nađe odgovor na pitanje kada to slučajna mera verovatnoće u proseku teži Benfordovoj. Takođe ukazuje da da postoje skupovi podataka koji se mogu približavati ili udaljavati od Benfordove mere verovatnoće.



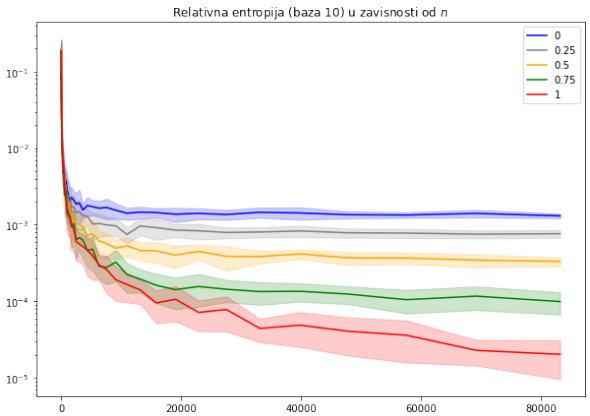
(a) Zavisnost χ^2 od broja odbiraka gde Benfordova slučajna promenljiva učestvuje sa procentom p, a ostatak podataka je iz $\Gamma(1,1)$ raspodele



(b) Zavisnost $D_{KL}(P||Q)$ od broja odbiraka gde Benfordova slučajna promenljiva učestvuje sa procentom p, a ostatak podataka je iz $\Gamma(1,1)$ raspodele

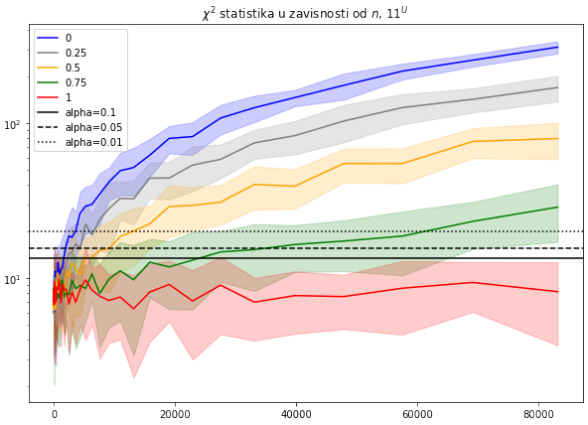


(c) Zavisnost χ^2 od broja odbiraka gde Benfordova slučajna promenljiva učestvuje sa procentom p, a ostatak podataka je iz $\text{Exp}(1)$ raspodele

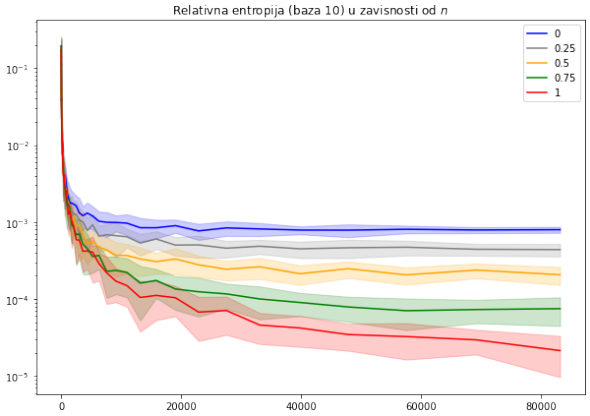


(d) Zavisnost $D_{KL}(P||Q)$ od broja odbiraka gde Benfordova slučajna promenljiva učestvuje sa procentom p, a ostatak podataka je iz $\text{Exp}(1)$ raspodele

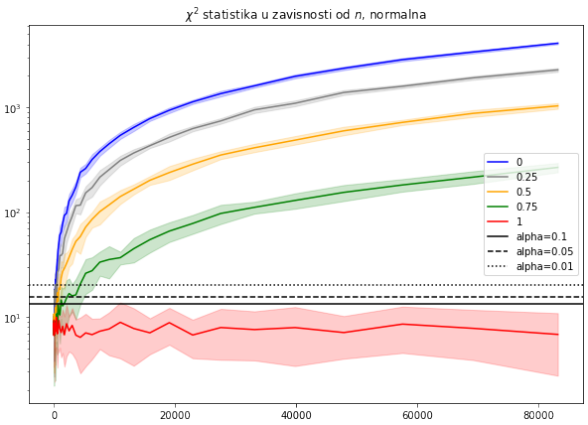
Slika 4.7: Zavisnost χ^2 statistike i relativne entropije od broja odbiraka



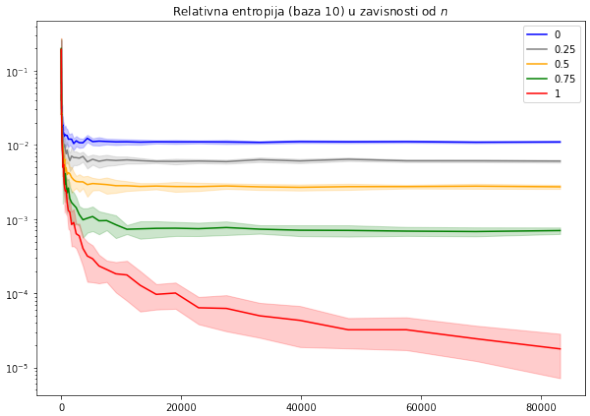
(a) Zavisnost χ^2 od broja odbiraka gde Benfordova slučajna promenljiva učestvuje sa procentom p, a ostatak podataka je slučajna promenljiva 11^U



(b) Zavisnost $D_{KL}(P||Q)$ od broja odbiraka gde Benfordova slučajna promenljiva učestvuje sa procentom p, a ostatak podataka je slučajna promenljiva 11^U

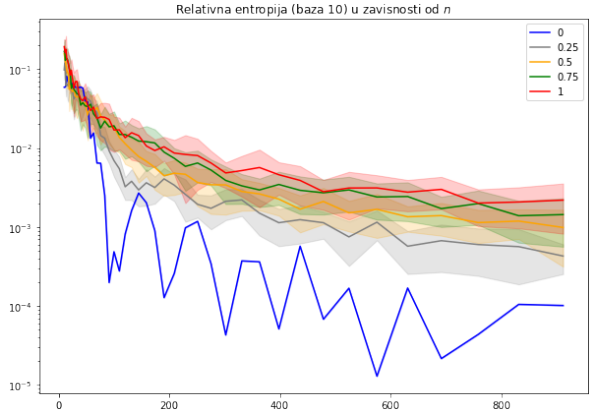


(c) Zavisnost χ^2 od broja odbiraka gde Benfordova slučajna promenljiva učestvuje sa procentom p, a ostatak podataka je iz $\mathcal{N}(0,1)$ raspodele

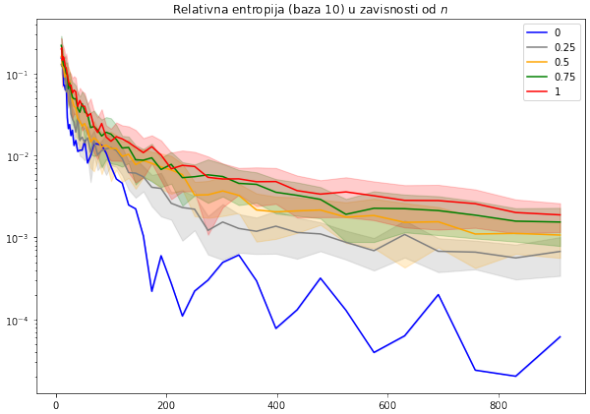


(d) Zavisnost $D_{KL}(P||Q)$ od broja odbiraka gde Benfordova slučajna promenljiva učestvuje sa procentom p, a ostatak podataka je iz $\mathcal{N}(0,1)$ raspodele

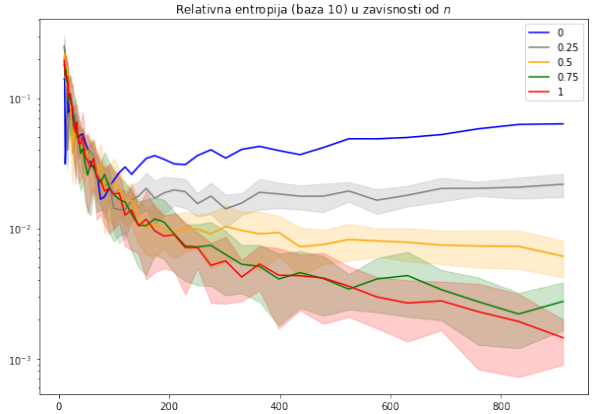
Slika 4.8: Zavisnost χ^2 statistike i relativne entropije od broja odbiraka



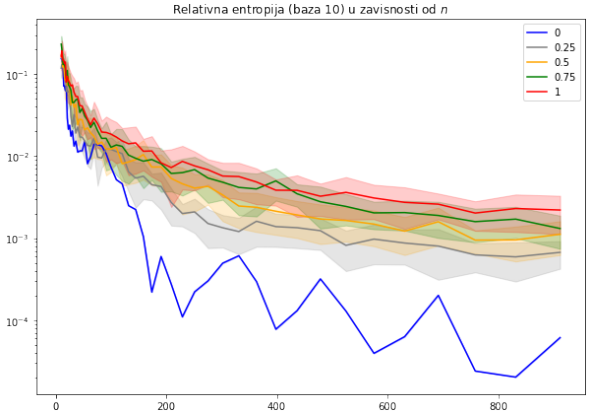
(a) Zavisnost $D_{KL}(P||Q)$ od broja odbiraka gde Benfordova slučajna promenljiva učestvuje sa procentom p, a ostatak podataka je niz 2^n



(b) Zavisnost $D_{KL}(P||Q)$ od broja odbiraka gde Benfordova slučajna promenljiva učestvuje sa procentom p, a ostatak podataka su nizovi 2^n i 3^n

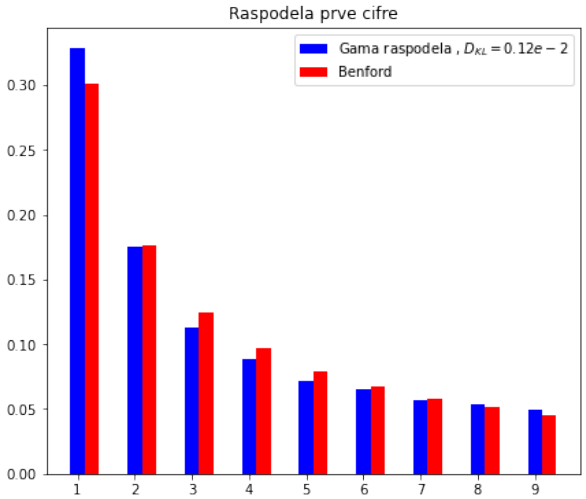


(c) Zavisnost $D_{KL}(P||Q)$ od broja odbiraka gde Benfordova slučajna promenljiva učestvuje sa procentom p, a ostatak podataka su nazmenični nizovi 2^n i 3^n

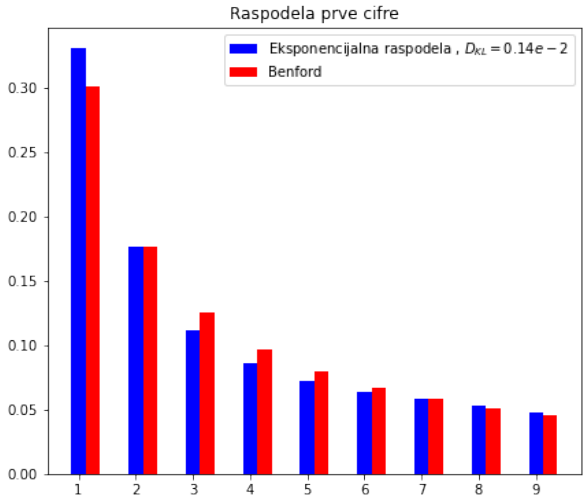


(d) Zavisnost $D_{KL}(P||Q)$ od broja odbiraka gde Benfordova slučajna promenljiva učestvuje sa procentom p, a ostatak podataka je geometrijski niz gde je na polovini promenjena osnova sa 2 na 3

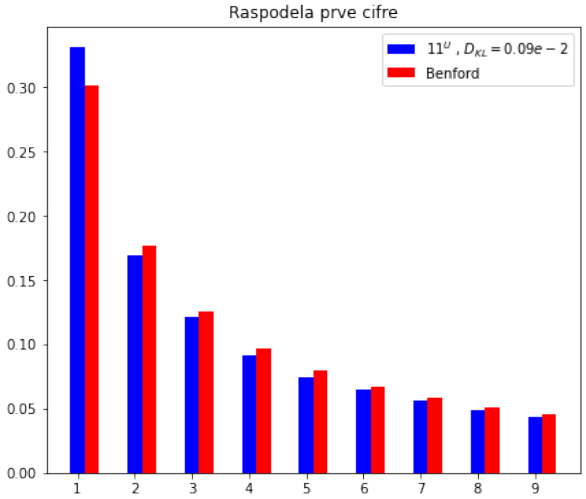
Slika 4.9: Zavisnost relativne entropije od broja odbiraka



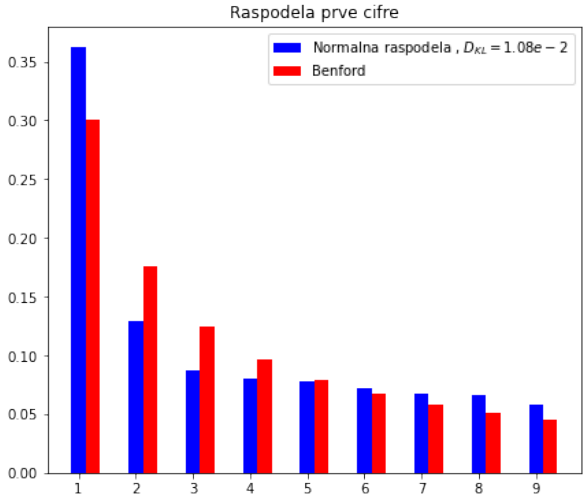
(a) Raspodela vodeće cifre za podatke iz $\Gamma(1,1)$ raspodele



(b) Raspodela vodeće cifre za podatke iz $\mathrm{Exp}(1)$ raspodele



(c) Raspodela vodeće cifre za slučajne promenljive oblika 11^U



(d) Raspodela vodeće cifre za podatke iz $\mathcal{N}(0,1)$ raspodele

Slika 4.10: Raspodele vodeće cifre i vrednost relativne entropije za približno Benfordove promenljive

Glava 5

Primena Benfordovog zakona u forenzici podataka

Postoje dve osnovne primene Benfordovog zakona. Jedna od najpoznatijih je primena Benfordovog zakona u forenzici podataka, i u ovom slučaju se Benfordov zakon primenjuje u raznim oblastima nauke. Manje poznata primena koristi Benfordov zakon u digitalnoj obradi signala i srodnim oblastima.

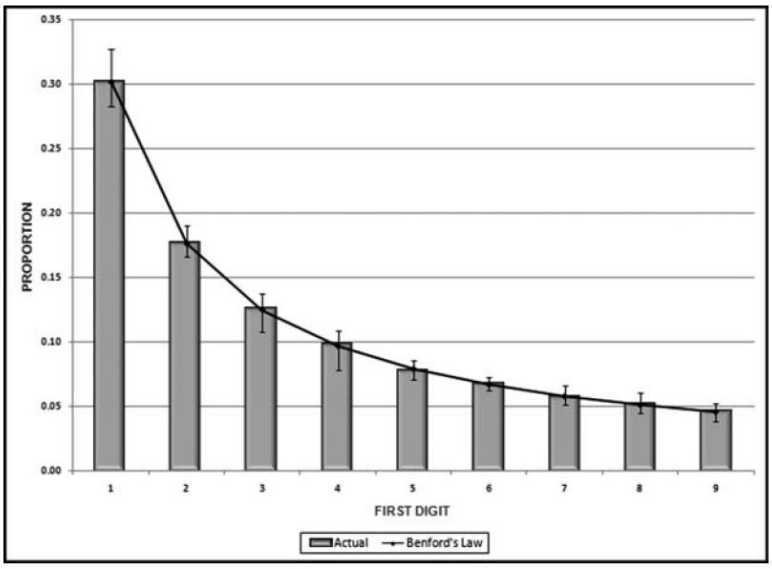
U ovom odeljku ćemo prikazati dve primene Benfordovog zakona u forenzici podataka – primena u detektovanju lažnih finansijskih izveštaja i detektovanju manipulacije brojem prijavljenih slučajeva zaraze Covidom 19.

5.1 Finansijski podaci

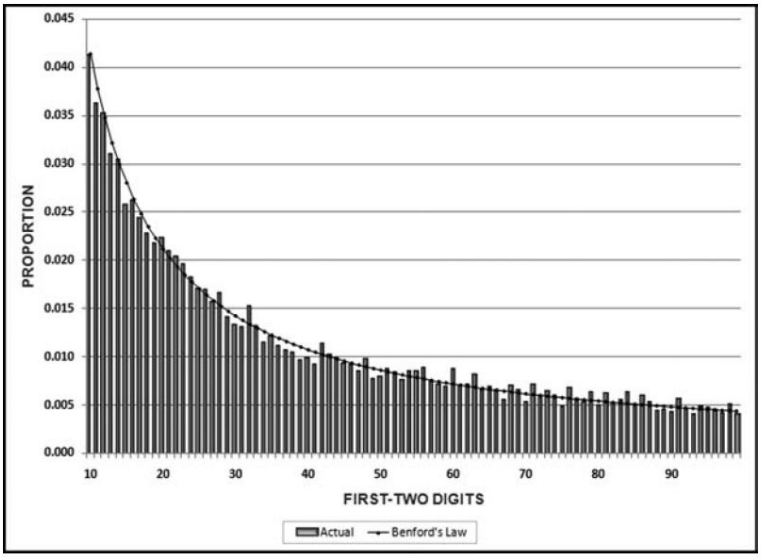
Prva primena Benfordovog zakona, i danas najčešća, jeste u analizi finansijskih izveštaja. Ova ideja se prvi put javlja 1972. godine kada je je Hal Varian predložio da se Benfordov zakon iskoristi za detekciju potencijalnih manipulacija u podacima socioekonomske prirode u radu [33]. Već 1975. se pojavljuju četiri rada iz psihologije koja se bave raspodelama vodećih cifara koje se dobijaju kada se od učesnika traži da generišu slučajne brojeve. Ovi radovi su potvrdili da ljudi imaju jako lošu procenu kada se od njih očekuje da izmisle slučajne brojeve i da imaju tendenciju da ravnomerno raspoređuju vodeće cifre, čime je predložena ideja o primeni Benfordovog zakona dobila svoje utemeljenje i sa strane psihologije.

Prvi radovi u kojima se zvanično koristi Benfordov zakon u analizi finansijskih podataka se pojavljuju 1992. godine, autori su Busta i Sundheim. U pitanju su dva rada – u prvom je pokazano da prijavljene vrednosti o povraćaju poreza u SAD-u prate Benfordov zakon, da bi u sledećem ovaj zaključak bio iskorišćen za detekciju anomalija u istim. Konačno, 1992. godine Nigrini u svom doktorskom radu koristi Benfordov zakon za detekciju utaje poreza [31]. Nigrini je vodeći autor iz oblasti primene Benfordovog zakona u forenzici finansijskih podataka i autor preko 30 radova i knjige [30] iz ove oblasti.

Ovakvi podaci po pravilu imaju jako mala odstupanja od Benfordove raspodele. Na slici 5.1 je prikazan tipičan primer raspodele vodećih cifara u podacima koji se javljaju u finansijskim izveštajima. U nastavku navodimo nekoliko primera koji su preuzeti iz [30].



(a) Raspodela vodeće cifre



(b) Raspodela prve dve cifre

Slika 5.1: Raspodela vodećih cifara sa označenim minimalnim i maksimalnim vrednostima dobijena analizom 25 skupova finansijskih podataka iz 2005. godine [30]

Pre nego što navedemo primere, napomenimo da u se u svima njima kao metrika koristilo srednje apsolutno odstupanje koje ne zavisi od broja odbiraka.

5.1.1 Promenjeni računi

Primena Benfordovog zakona je nekad prilično jednostavna. Na primer, u [30] se navodi primer prodavca koji je radio u osiguravajućoj kući koji je lažirao troškove hrane i puta. Politika firme je bila da im se dostave skenirani računi troškova, a da oni potom svojim zaposlenima daju novac za iste. Skenirani računi su bili jako loše rezolucije, pa je bilo lako promeniti brojeve tako da se to ne primeti. Njegova greška je bila u tome što se prevara svodila na to da se vodeća cifra jedinica promeni u sedmicu ili devetku. Kako je u pitanju bila velika osiguravajuća kuća računovođe su odlično poznavali alate poput Benfordovog zakona i koristili ga da provere prijavljene troškove. Naravno, ovako promenjeni troškovi su imali velika odstupanja i vrlo brzo

je detaljna analiza u kojoj se jasno videlo da su cene previsoke pokazala da se radilo o prevari.

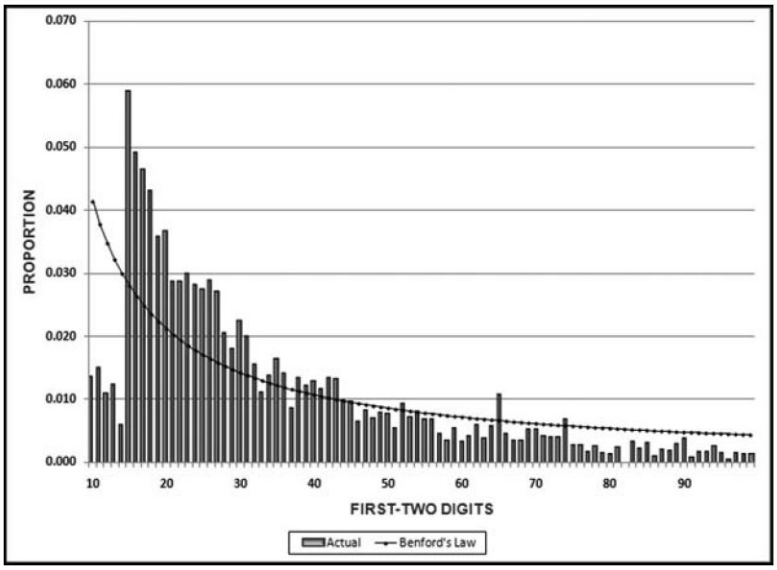
5.1.2 Troškovi zdravstvenog osiguranja

Još jedan zanimljiv primer primene Benfordovog zakona, dat u [30], je u vezi lažnih troškova zdravstevnog osiguranja. Radilo se kompaniji koja je imala lanac motela i koja je pružala zaposlenima zdravstveno osiguranje. Zaposleni su svoje račune za troškove lečenja slali glavnoj kancelariji i dobijali pokriće. U kompaniji nije postojala interna kontrola, ali je primećeno da su se ovi troškovi značajno povećali.

Supervizor zadužen za proveru ovih podataka je analizirao histogram prve dve cifre i rezultat je prikazan na slici 5.2. Dobijeni grafik izgleda veoma čudno, ali razlog za to je što je snaga računara u to vreme bila prilično ograničena i supervizor nije mogao da analizira ceo skup podataka. Zbog toga je on izbacio sva potraživanja koja su bila manja od 1.500 $, smatrajući da su izmišljena potraživanja verovatno veća od ove vrednosti. Ovo je razlog što u histogramu fale brojevi koji počinju ciframa od 10 do 14. Razlog što postoji znatno više odbiraka sa malim vodećim ciframa je posledica toga što podaci ne zauzimaju veliki opseg, a 1.500 $ je minimalna vrednost.

Ono što je bilo interesantno je usamljeni vrh za vodeće cifre 65. Potražovanja koja počinju sa 65 su detaljno analizirana i ispostavilo se da postoji 13 lažnih čekova za iznose od 6.500 $ do 6.599 $. Naime, jedna od supervizora je našla način da lažira izveštaje o manjim operacijama srca čija je cena bila u navedenom problematičnom opsegu. Interesantno je da je zaposlene koji su bili navodni pacijenti veoma pažljivo birala tako da izgleda kao da je potreba za ovim operacijama bila najveća upravo u onim motelima u kojima je i najveća prosečna starost zaposlenih.

Dodatnim ispitivanjima je otkriveno da je ukupna naneta šteta kompaniji bila milion dolara.

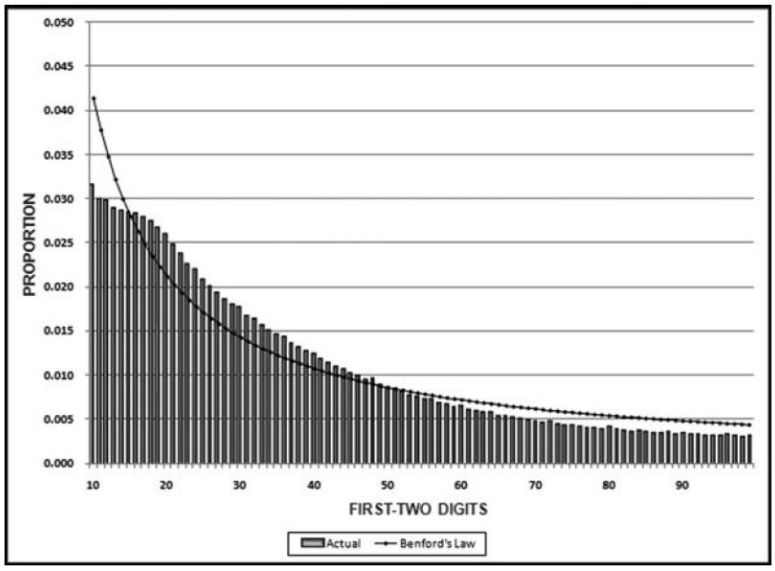


Slika 5.2: Prve dve cifre potraživanja za zdravstvenu zaštitu zaposlenih [30]

5.1.3 Krađa struje

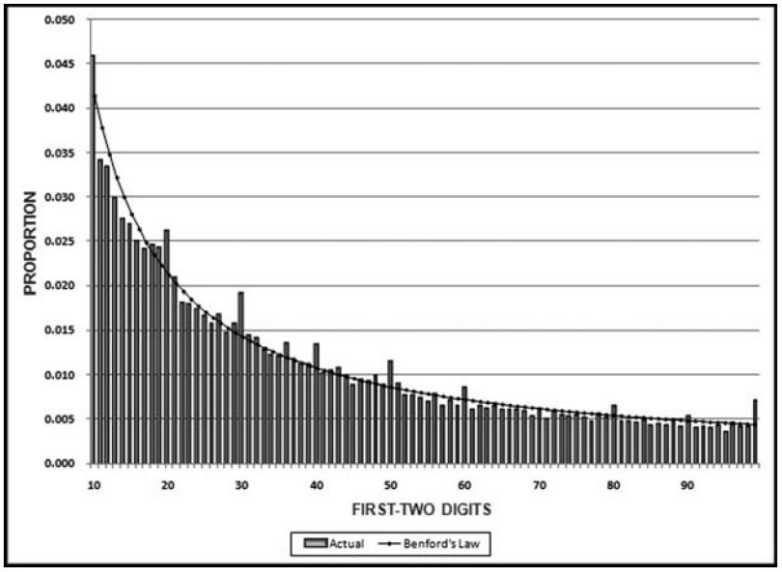
Elektroprivreda iz Južne Amerike je znala da postoji neslaganje između naplaćene i stvarno potrošene struje. Finansijski gubici su bili veoma veliki i zbog toga je započeta istraga. Prvi korak je bio da se proveri da li cifre naplaćenih kWh poštuju Benfordov zakon. Dobijeni rezultati su prikazani na slici 5.3. Sa slike se vidi da postoji problem sa podacima, ali ne postoje nikakve vrednosti koje odstupaju od susednih, pa nije jasno gde treba tražiti nepravilnosti. Ovo nije

čudno, naime, nas interesuje gde se dešava krađa struje, to jest slučajevi koji nisu naplaćeni koji se ne nalaze na ovom grafiku.



Slika 5.3: Vodeće cifre naplaćenih kWh [30]

Pošto ovakva analiza nije dala rezultate posmatrani su kWh kreditirani kupcima. Raspodela ovih cifara je prikazana na slici 5.4. Ono što je je bilo čudno u ovim podacima je činjenica da je u jednoj godini bilo 90.000 kredita, što je delovalo previše za sistem sa automatskom naplatom. Kreditirane vrednosti i naplaćene očigledno nemaju isti oblik histograma cifara, i to je takođe bilo veoma čudno. Konačno, kreditirane vrednosti su bile jako bliske Benfordovim, ali su imale usamljene vrhove koji ukazuju na dupliranje podataka ili neku drugu vrstu anomalije koja može biti prevara.



Slika 5.4: Vodeće cifre kreditiranih kWh [30]

Dalja analiza se fokusirala na kredite i merenja koji su pokazivali značajna smanjenja tokom godine. Jedna od stvari koje su bile analizirane u kreditima su bile količine sa vodećim ciframa 99 koje se pojavljuju češće nego što bi trebalo. Ove vrednosti su takođe u vezi sa ograničenjima

iznad kojih se povećava cena, pa mogu ukazivati na dogovor između radnika koji vrši naplatu i kupaca, koji bi obezbedio kupcu da ostane ispod granice. Analizom kredita koji počinju ciframa 99 je uočeno nekoliko kredita koji su jako blizu 1 milionu kWh i preko 200 vrednosti kredita koje su za jedan manje od 100.000 kWh.

Konačno, posmatrana je merena potrošnja, gde je fokus bio na onim vrednostima koje imaju veliku negativnu korelaciju sa prosečnim šablonima korišćenja, veliko relativno smanjenje potrošnje u između prve i druge polovine godine i velika smanjenja u broju naplaćenih kWh.

Ovakva analiza je dala listu od nekih 1.200 kupaca koji su bili sumnjivi. Konačan izveštaj je sadržao stotine kupaca koji su imali velike kredite, a u isto vreme nisu imali registrovanu potrošnju u drugoj polovini godine. Ova istraga je dovela do povraćaja od nekoliko miliona dolara i sprečila dalje gubitke.

U ovo odeljku smo videli kako se ispravno primenjuje Benfordov zakon. Prikazani primeri su bili dobri jer je, pre svega, data jaka osnova za pretpostavljanje Benfordovog zakona kao nulte hipoteze. Dalja primena je uvek, osim u trivijalnim slučajevima, podrazumevala proveru koja ne posmatra samo vodeće cifre već i značajni deo broja. Dodatno, u [31] se mogu naći još mnogi primeri u kojima su i poslednje cifre takođe korišćene za detekciju nepravilnosti. Osim toga, u ovim situacijama su manipulacije podacima takve da dovode do odstupanja od Benfordovog zakona - izbacivanje vrednosti, ponavljanje vrednosti, izmišljanje brojeva i slične. Primetimo još da je je mereno odstupanje od Benfordovog zakona, odnosno nije korišćeno testiranje hipoteza za koje smo već videli da je problematično. I konačno, nakon ovakve analize koja je gotovo uvek dodatno prilagođena konkretnom problemu, izvršeno je detaljno ispitivanje sumnjivih podataka. Benfordov zakon nigde nije iskorišćen kao dokaz samostalno, već isključivo kao nešto što bi moglo da ukaže na probleme ili smanji prostor pretrage.

5.2 Covid-19 pandemija

Od početka pandemije do trenutka pisanja ovog rada je prošlo nešto više od godinu i po dana, a na sajtu `www.benfordonline.net` se već može naći preko 40 radova koji u svom naslovu sadrže ključnu reč Covid. Veliki broj ovih radova koristi Benfordov zakon za analizu validnosti podataka, pri čemu se najčešće proverava broj registrovanih slučajeva. U nastavku ćemo na osnovu do sada iznetog probati da kritički sagledamo ovaj pristup.

U prethodnim odeljcima smo pokazali primere ispravne primene Benfordovog zakona i sistematizovali koji su to uslovi bili ispunjeni da bi primena bila valjana. Kada su u pitanju Covid-19 podaci, nailazimo na gotovo suprotne rezultate.

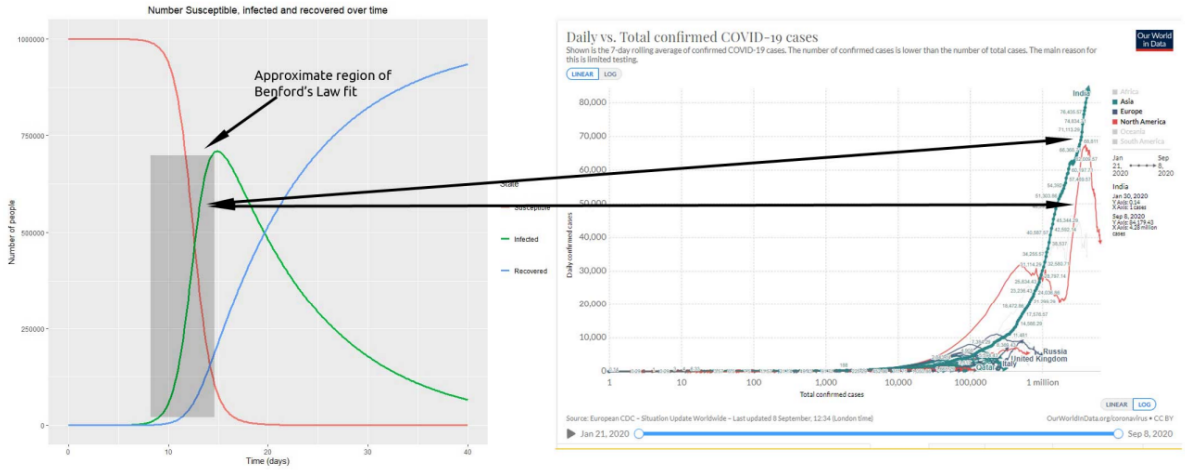
Prvi problem je uvođenje Benfordove raspodele kao nulte hipoteze. U radu [23] je pokazano da postoji osnova za korišćenje Benfordovog zakona za proveru prijavljenog broja slučaja zaraženih u toku epidemije. Dobijeni rezultat nije čudan jer se širenje virusa aproksimira geometrijskim nizovima ili eksponencijalnim funkcijama. Međutim, onda kada broj zaraženih prestane da raste eksponencijalno, više nema osnova za očekivanje Benfordovog zakona.

Poznato je da uvođenjem različitih 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 5.5. 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. Samo na osnovu činjenice da je potrebno imati eksponencijalno širenje opasnog virusa 50 dana, bez ikakvih mera koje bi to sprečile, očekivanje Benfordovog zakona postaje veoma problematično.

U radovima [23, 25] je predloženo da se odstupanje od Benfordovog zakona koristi kao mera za utvrđivanje uspešnosti suzbijanja Covid-19 pandemije.

U radu [18] je posmatran prijavljeni broj slučaja virusa Covid-19 u Kini, Italiji i SAD-u [1].

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



Slika 5.5: Primer SIR modela u kome je označena regija u kojoj se može očekivati Benfordov zakon [23].

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 5.1 su prikazane vrednosti ovih statistika dobijenih u radu [18] za prijavljeni broj zaraženih pre uvođenja mera.

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 5.1: Dobijene vrednosti statistika pre uvođenja mera [18]. 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 radu [18] piše da χ^2 statistika odbacuje nultu hipotezu, ali ovde nalazimo da to nije slučaj. Nijedna statistika 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 [18] se takođe došlo do istog zaključka.

U slučaju SAD-a, ukoliko uzmemo strožu vrednost V_n statistike, date u tabeli 4.1, 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, ono što je dodatno problematično u ovom radu jeste korišćenje podataka iz različitih geografskih oblasti. Poznato je da kada posmatramo podatke o broju stanovnika imamo pojavu Benfordovog zakona [30]. Ako umesto broja stanovnika uzimamo taj broj pomnožen konstantom, rezultat se neće promeniti. Prema tome, to što podaci i dalje prate Benfordov zakon

bi mogla biti samo indikacija da je procenat registrovanih slučajeva u proseku isti u različitim geografskim oblastima.

Primetimo, na kraju, da je posmatrana samo vodeća cifra. U primerima ispravne primene Benfordovog zakona smo videli da podaci mogu biti Benfordovi, i da postoje anomalije, koje se često vide tek onda kada se posmatraju i ostale cifre. Štaviše, u mnogim primerima je upravo veliko poklapanje sa Benfordovim zakonom dovelo do toga da se problematični podaci još jasnije vide.

Konačno, u kontekstu testiranja validnosti epidemioloških podataka je potrebno setiti se teoreme 3.1 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. Prema tome, samo posmatranje broja registrovanih slučajeva, bez razmatranja metodologije testiranja nije dovoljno.

Jedan od problema 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.

Tako, na primer, u radu [21] je dat pregled vrednosti statistika χ^2, V_n i MAD za sve države i označeno je sa kojom značajnošću koja statistika, odnosno metrika, 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. Osim što ovakav skok u zaključivanju nije opravdan, postoji problem i sa veličinom ovih podataka.

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. Ovakva analiza bi gotovo izvesno zahtevala proveru više vodećih cifara i ne bi lako mogla da se uoči posmatranjem samo prve cifre.

Primena Benfordovog zakona za proveru validnosti podataka zahteva oprez, pogotovo kada se proveravaju epidemiološki podaci. U kontekstu Covid-19 podataka se za primenu Benfordovog zakona može iskoristiti izreka *Kada je jedini alat koji imamo čekić, svaki problem izgleda kao ekser*.

Kada sagledamo ove analize vidimo da u samom korenu analiza postoji problem u smislu postavljanja nulte hipoteze. Onda i kada je nulta hipoteza opravdana (podaci iz različitih

geografskih oblasti ili gradova) analiza nije dovoljno detaljna. Konačno, kao što je već rečeno, Benfordov zakon se, pre svega koristi kao signal da nešto nije u redu, ne mogu se na osnovu toga izvoditi zaključci da postoje anomalije u podacima, a pogotovo je problematično napraviti još jedan skok i reći da su anomalije posledica prevare. Onda kada se i potvrdi da podaci prate Benfordov zakon, to nema isto značenje koje bi imalo u slučaju finansijskih izveštaja. Najlakša i najverovatnija manipulacija broja zaraženih se ispoljava u vidu skaliranja koje ovakav test ne može da otkrije.

Glava 6

Zaključak

U ovom radu je predstavljena teorija Benfordovog zakona i analizirana njegova primena u forenzici podataka. U glavi 2 je utemeljena teorija na osnovu koje su izvođene dodatne osobine Benfordovih objekata. Benfordov zakon je značajan i u teorijskom smislu. Kada radimo sa merama verovatnoće koje definišemo na skupu realnih brojeva, moguće je zanemariti sigma polja, što se često i radi u literaturi koja je fokusirana na primenu. Razlog je to što su nemerljivi događaji u ovim situacijama takvi da se ne mogu efektivno konstruisati. Kada definišemo meru verovatnoće na značajnom delu broja, problemi nemerljivih događaja prestaju da budu tako apstraktni i potreba za sigma poljima postaje mnogo konkretnija. Na osnovu ovako precizno definisanog prostora verovatnoće su u glavi 3 predstavljene značajne osobine Benfordovih objekata.

Doprinos teorijskog izlaganja je u tome što je ovde prvi put na srpskom jeziku izneta teorijska osnova Benfordovog zakona. Pored toga, otvorena su i mnoga pitanja koja su data na kraju ove celine. Od predstavljenih pitanja izdvajamo pitanje o tome da li je teorema o kombinaciji uniformnih raspodela (teorema 3.9) samo specijalan slučaj teoreme o kombinaciji raspodela koje dovode do pojave Benfordove raspodele (3.8) i pitanje o slučajnim merama verovatnoće koje su u proseku Benfordove i njihovom odnosu sa realnim merama koje srećemo u praksi.

U glavi 4 je detaljno analizirano testiranje Benfordovog zakona. U ovoj glavi je uspostavljena veza između rigorozne teorije i dosta manje rigorozne primene Benfordove raspodele koje su se, čini se, razvijale dosta odvojeno jedna od druge. Problemi koji postoje kada pokušavamo da uspostavimo vezu između ove dve strane su identifikovani i predložena su rešenja. Na osnovu dobijenih rezultata su date ideje i smernice koje bi mogle da pomognu da odgovorimo na pitanja postavljena u prethodnoj glavi.

U glavi 5 je pokazana metodologija primene Benfordovog zakona u forenzici podataka. Kao ilustracija dobre i ispravne primene su dati primeri iz Nigrinijeve knjige. Kao značajan primer za oblast elektrotehnike navodimo primer u kome je pokazano kako su korišćenjem Benfordovog zakona pronađeni oni koji su učestvovali u krađi struje. U istoj glavi je, takođe, ukazano i na neispravnu primenu Benfordovog zakona za analizu podataka Covid-19 pandemije i detaljno je objašnjeno šta u ovakvoj primeni nije bilo valjano.

Spisak simbola

$\log x$	Dekadni logaritam
$\ln x$	Prirodni logaritam
$D_1(x)$	Vodeća cifra broja x
$S(x)$	Značajni deo (mantisa) broja x u dekadnom brojevnom sistemu
$\mathbb{N}$	Skup prirodnih brojeva
$\mathbb{Z}$	Skup celih brojeva
$\mathbb{R}$	Skup realnih brojeva
$\mathbb{R}^+$	Skup pozitivnih realnih brojeva
Ω	Skup događaja
$\mathbb{P}$	Mera verovatnoće na $(\Omega, \mathcal{A})$
$(\Omega, \mathcal{A}, \mathbb{P})$	Prostor verovatnoće
$\mathcal{B}$	Borelovo sigma polje
$\mathcal{B}[a, b)$	Borelovo sigma polje na $[a, b)$
$\mathcal{P}(\Omega)$	Partitivni skup skupa Ω
$\lambda_{a,b}$	Lebegova mera na $([a, b), \mathcal{B}[a, b))$
$\mathcal{S}$	Sigma polje na značajnom delu broja
X	Slučajna promenljiva
$\lfloor x \rfloor$	Ceo deo broja x
$\langle x \rangle$	Razlomljeni deo broja
U	Slučajna promenljiva sa uniformnom raspodelom na $[0, 1]$
$\mathbb{B}$	Benfordova mera verovatnoće
$u.r.$ mod 1	Uniformna raspodela po modulu 1
δ_a	Dirakova mera verovatnoće koncentrisana u tački a
$D_{KL}(P\|\|Q)$	Relativna entropija P u odnosu na Q

Spisak slika

Spisak tabela

Literatura

[1] Frank Benford. Base dependence of Benford random variables. *Stats-MDP*, 4(3):578–594, 2021.

[2] Frank A. Benford. The law of law of anomalous numbers. *Proceeding of the American Philosophical Society*, 78(4):551–572, 1938.

[3] Arno Berger and Theodore P. Hill. A basic theory of Benford's law. *Probability Surveys*, (8):1–126, 2011.

[4] 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.

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

[6] Arno Berger and Theodore P. Hill. The Mathematics of Benford's Law: a primer. *Statistical methods and Applications*, 2020. ArXiv 1909.07527 (2020).

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

[8] Zhaodong Cai, Matthew Faust, Adolf J. Hildebrand, Junxianan Li, and Yuan Zhang. Leading digits of mersenne numbers. *Experimental Mathematics*, pages 1–17, 2019.

[9] Zhaodong Cai, Matthew Faust, Adolf J. Hildebrand, Junxianan Li, and Yuan Zhang. The surprising accuracy of Benford's law in mathematics. *Amer. Math. Monthly*, 127(3):217–237, 2020. ArXiv 1907.08894v.2, 2019.

[10] Amélia Sofia Carvalho Damiao da Silva. The application of Benford's law in detecting accounting fraud in the financial sector. Master's thesis, Lisboa School of Economics and Management, 2013.

[11] Dorrell D. Darrell and Gadawski A. Gregory. Counterterrorism: Conventional tools for unconventional warfare. *United States Atorneys' Bulletin*, March 2005.

[12] Joseph Deckert, Mikhail Myagkov, and Peter C. Ordeshook. Benford's law and the detection of election fraud. *Political Analysis*, 19(3):245–268, 2011.

[13] Dragan S. Djordjević. *Mera, integral i izvod.* Univerzitet u Nišu, Prirodno-matematički fakultet, 2014.

[14] William Goodman. The promises and pitfalls of Benford's law. *Significance*, 13(3):38–41, 2016.

[15] Theodore P. Hill. Base-invariance implies Benford's law. *Proceedings of the American Mathematical Society*, 123(3):887–895, 1995.

[16] Theodore P. Hill. A statistical derivation of the significant-digit law. *Statistical science*, 10(4):354–363, 1995.

[17] Theodore P. Hill. A widespread error in the use of Benford's law to detect election and other fraud, 2020. ArXiv:2011.13015.

[18] Raúl Isea. How valid are the reported cases of people infected with Covid-19 in the world. *Internantial Journal of Coronaviruses*, 1(2):53–56, 2020.

[19] Élise Janvresse and Thierry de la Rue. From uniform distributions to Benford's law. *Journal of Applied Probability*, 41(4):1203–1210, 2004.

[20] Walter R. Mebane Jr. Election forensics: Vote counts and Benford's law. In *Summer Meeting of the Political Methodology Society, UC-Davis*, volume 17, 2006.

[21] Ana Kilani and Gina P. Georgiu. Countries with potential data misreport based on Benford's law. *Journal of Public Health*, 43(2):e295–e296, 2021.

[22] Christoffer Koch and Ken Okamura. Benford's law and Covid-19 reporting. *Economics Letters*, 196(109573), 2020.

[23] Nils Koester, Andrena McMenemyand, and Yohan Bélanger. Simulating epidemics with a SIRD model and testing with Benford's law, 2020. Preprint. DOI: 10.13140/RG.2.2.15903.38566.

[24] Alex Ely Kossovsky. On the mistaken use of the chi-square test in Benford's law. *Statist-MDPI*, 4:419–453, 2021.

[25] Kang-Bok Lee, Sumin Han, and Yeasung Jeong. COVID-19, flattening the curve, and Benford's law. *Physica A*, 559(125090), 2020.

[26] Bruno Massé and Dominique Schneider. Fast growing sequences of numbers and the first digit phenomenon. *International Journal of Number Theory*, 11(3):705–719, 2015.

[27] Milan Merkle. *Verovatnoća i statistika za inženjere i studente tehnike*. Četvrto izmenjeno i dopunjeno izdanje, Akademska misao, Beograd, 2016.

[28] Steven J. Miller, editor. *Benford's Law*. Princeton University Press, 2015.

[29] John Morrow. Benford's law, families of distributions and a test basis. CEP Discussion paper 1291, Centre for Economic Performance, London School of Economics and political science, August 2014.

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

[31] Mark John Nigrini. *The detection of income tax evasion through an analysis of digital distributions*. University of Cincinnati, 1993.

[32] Steven W. Smith. *The Scientist and Engineer's Guide to Digital Signal Processing*. California Technical Pub., 2002.

[33] Hal R. Varian. Benford's law. *The American Statistician*, 26(3):65–66, 1972. (In Letters to the Editor.).

[34] Michał Ryszard Wójcik. Notes on scale-invariance and base-invariance for Benford's law, 2013.