math.e

Hrvatski matematički elektronički časopis

Benfordov zakon

nizovi znamenaka slučajna varijabla statistika

Bojan Basrak i Ivan Varga

PMF-Matematički odsjek, Bijenička 30, 10000 Zagre
bojan.basrak@math.h

1 Uvod

Prije široke dostupnosti računala i ručnih kalkulatora, znanstvenici su se često oslanjali na tzv logaritamske tablice. Zahvaljujući njima mnogi su se izračuni mogli pojednostaviti ili barem približn provesti. Posebno su bile važne u astronomiji. Tako je, američki astronom Simon Newcomb, jo 1881. godine primjetio da su početne stranice u logaritamskim tablicama istrošenije od ostali stranica. Kako tablice sadrže logaritme decimalnih brojeva poredanih po prvoj značajnoj znamenk Newcomb je naslutio da prva značajna znamenka stvarnih podataka nije jednoliko distribuirana Njegova opažanja ga na kraju dovode do pretpostavke da je vjerojatnost pojave znamenke d ka prve znamenke nekog od podataka, jednaka

$$\log_{10}(1+d) - \log_{10}(d),$$

za sve $d \in \{1, 2, \ldots, 9\}$. Taj isti fenomen primjećuje i fizičar Frank Benford 1938. godine. On g detaljnije istražuje i testira na različitim skupovima podataka, kao što su površine rijeka, veličin stanovništva, fizikalne konstante itd., pa se zbog toga otkrivanje ovog zakona pripisuj upravo Benfordu.

U praksi prikupljene numeričke podatke, mi danas matematički modeliramo slučajnim varijablama. Ako slučajnu varijablu označimo sa X, a njenu prvu značajnu znamenku sa $D_1(X)$, ove oznake možemo iskoristiti da iskažemo Benfordov zakon. Benford je jednostavno pretpostavio da će vjerojatnost pojavljivanja značajne znamenke d zadovoljavati

$$P(D_1(X) = d) = \log_{10}\left(1 + \frac{1}{d}\right), \tag{1}$$

za sve $d \in \{1, 2, \ldots, 9\}$. Upravo kako je naslutio i Newcomb. Za razdiobe za koje vrijedi ov pretpostavka, kažemo da zadovoljavaju {\em Benfordov zakon za prvu značajnu znamenku}.

Lako se uvjeriti da ovaj zakon ipak ne vrijedi za mnoge teorijske i često korištene razdiobe. Ako X npr. uniformno izaberemo iz intervala $(0, 1)$, i prva značajna znamenka imat će jednaku vjerojatnos da poprimi vrijednosti od 1 do 9. Ni za najvažniju razdiobu u statistici Benfordov zakon ne vrijedi. Naime, ako je X normalna (ili Gaussova) slučajna varijabla, može se pokazati da (1) ne vrijedi. Unatoč tome Newcombova i Benfordova slutnja potvrđene su empirijski na mnogim skupovima podataka.

U nastavku ćemo detaljnije prikazati Benfordov zakon, kao i neka teorijska opravdanja za njegovo pojavljivanje koje su matematičari (predvođeni T. Hillom) pronašli u zadnjih nekoliko desetljeća.

2 Benfordovo svojstvo

Pokazuje se da Benfordov zakon možemo iskazati i preciznije. Takav precizniji zakon određuj

razdiobu i za sve ostale značajne znamenke slučajno odabranog broja iz dane razdiobe.

Za svaki realan broj x različit od nule, prvu značajnu znamenku, u oznaci $D_1(x)$, formalno definiramo kao jedinstven broj $j \in \{1, 2, \ldots, 9\}$ za koji vrijedi

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

za neki $k \in \mathbb{Z}$. Jasno je da su brojevi k i j s tim svojstvom jedinstveni. Korisno je definirati i tzv *signifikant* (ili mantisu) realnog broja. Za $x \neq 0$, signifikant je jedinstven broj $S(x)$ iz interval $[1, 10)$ za koji vrijedi \linebreak[4] $|x| = 10^k S(x)$ za neki $k \in \mathbb{Z}$. Funkciju koja svakom realnor broju x pridružuje njegov *signifikant*

$$x \mapsto S(x),$$

nazivamo signifikantna funkcija. Pri tom za $x = 0$, definiramo $S(0) := 0$.

Iako nas prije svega zanima Benfordovo svojstvo za slučajne varijable, isto svojstvo mogu imati i nizovi. Označimo sa $\#A$ kardinalitet proizvoljnog skupa A. Niz realnih brojeva (x_n) je *Benfordov niz*, ako

$$\lim_{N\to\infty} \frac{\#\{1 \leq n \leq N : S(x_n) < t\}}{N} = \log_{10} t \quad \text{za sve } t \in [1, 10). \tag{2}$$

Benfordovo svojstvo dakle, precizira razdiobu signifikanta takvog niza. Samim tim, uočimo da (2 određuje razdiobu prve, ali i bilo koje druge značajne znamenke u nizu. Posebno npr.

$$\begin{aligned}
&\lim_{N\to\infty} \frac{\#\{1 \leq n \leq N : D_1(x_n) = d_1\}}{N} \\
&= \lim_{N\to\infty} \frac{\#\{1 \leq n \leq N : d_1 \leq S(x_n) < d_1 + 1\}}{N} \\
&= \lim_{N\to\infty} \frac{\#\{1 \leq n \leq N : S(x_n) < d_1 + 1\}}{N} \\
&\quad - \lim_{N\to\infty} \frac{\#\{1 \leq n \leq N : S(x_n) < d_1\}}{N} \\
&= \log(d_1 + 1) - \log(d_1) = \log\left(1 + \frac{1}{d_1}\right),
\end{aligned}$$

za sve $d_1 \in \{1, 2, \ldots, 9\}$. Slično, ako sa $D_2(x)$ označimo drugu značajnu znamenku realnog broj x, a niz (x_n) je Benfordov, tada

$$\lim_{N\to\infty} \frac{\#\{1 \leq n \leq N : D_1(x_n) = d_1 \text{ i } D_2(x_n) = d_2\}}{N} = \log\left(1 + (10d_1 + d_2)^{-1}\right),$$

za sve $d_1 \in \{1, 2, \ldots, 9\}$ i $d_2 \in \{0, 1, \ldots, 9\}$.

Poznato je npr. da je niz potencija $2^n, n \in \mathbb{N}$, Benfordov. Isto vrijedi i za niz faktorijela ili Fibonaccijev niz. S druge strane niz prirodnih odn. prostih brojeva nema ovo svojstvo.

Ako pak slučajna varijabla X zadovoljava

$$P(S(X) < t) = \log_{10} t, \tag{3}$$

za sve $t \in [1, 10)$, kažemo da X (odn. njena razdioba) posjeduje Benfordovo svojstvo. Za sv ovakve X, kao direktnu posljedicu dobivamo Benfordov zakon za prvu značajnu znamenku. Naime iz (3) slijedi

$$\begin{aligned}
P(D_1(X) = d) &= P(d \leq S(X) < d + 1) \\
&= P(S(X) < d + 1) - P(S(X) < d)
\end{aligned}$$

$$= \log_{10}(d+1) - \log_{10}(d)$$

za sve $d \in \{1, 2, \ldots, 9\}$, što dokazuje tvrdnju.

3 Povezana svojstva

Interesantno je da je Benfordovo svojstvo nizova usko povezano sa tzv. svojstvom uniformnos modulo 1. Ovo potonje svojstvo je vrlo značajno i podrobno proučavano u teoriji brojeva npr. označimo sa $\{x\}$ tzv. razlomljeni dio realnog broja x. Preciznije, $\{x\} = x - \lfloor x \rfloor$, gdje je $\lfloor x$ oznaka za najveći cijeli broj manji ili jednak x. Tako, npr. $\{2.71\} = 0.71$ i $\{-2.71\} = 0.29$. Za ni (x_n) kažemo da je uniformo distribuiran modulo 1, ako vrijedi

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

Analogno, slučajna varijabla X (odn. njena razdioba) uniformno je distribuirana modulo 1, ako

$$\mathbb{P}(\{X\} < s) = s, \quad \text{za sve } \ s \in [0,1).$$

Vezu između ovih svojstava objašnjava idući teorem, koji odmah daje i jedan recept za praktičn provjeru Benfordovog svojstva. Dokaz teorema se može pronaći u Hill [1].

Theorem 1. *Slučajna varijabla je Benfordova ako i samo ako je logaritam po bazi deset njen apsolutne vrijednosti uniformno distribuiran modulo 1.*

Analogni teorem vrijedi i za nizove realnih brojeva. Jasno je da stvarni podaci Benfordovo svojstv mogu imati tek približno. Ipak, u mnogim primjenama razumno je očekivati da podaci (barer približno) zadrže Benfordovo svojstvo i nakon promjene skale. Ako se npr. radi o novčanim iznosim Benfordovo svojstvo bismo mogli očekivati i nakon promjene valute. Slično, Benfordovo svojstvo z duljine rijeka očekivali bismo da vrijedi neovisno o tome da li te duljine izražavamo u miljama i kilometrima. Izuzetno je zanimljivo da invarijatnost na množenje skalarom daje alternativn karakterizaciju Benfordovog svojstva.

Theorem 2. *Za svaku slučajnu varijablu X, za koju je $\mathbb{P}(X = 0) = 0$, sljedeće s tvrdnje ekvivalentne:*

- *[(i)] X je Benfordova.*
- *[(ii)] Postoji znamenka $d \in \{1, 2, \ldots, 9\}$ tako da*

$$\mathbb{P}(D_1(\alpha X) = d) = \mathbb{P}(D_1(X) = d) \quad \text{za svaki } \alpha > 0,$$

gdje je $\mathbb{P}(D_1(X) = d) = \log(1 + d^{-1})$.

Dokaz teorema se može vidjeti u Hill [1].

Na sličan način možemo karakterizirati Benfordovo svojstvo i za realne nizove. Navedimo tek da je signifikantna funkcija niza realnih brojeva (x_n) invarijantna na množenje skalarom ako za svaki $\alpha > 0$ i $t \in [1, 10)$ vrijedi,

$$\lim_{N\to\infty} \frac{\#\{1 \le n \le N : S(\alpha x_n) < t\}}{N} = \lim_{N\to\infty} \frac{\#\{1 \le n \le N : S(x_n) < t\}}{N}. \qquad (4)$$

Naglasimo još da postoje i razna druga interesantna svojstva Benfordovih razdioba, koja dijelor nadilaze ambicije ovog pregleda, za detalje pogledajte npr. [1].

4 Primjeri

Fibonaccijevi brojevi F_n, $n = 0, 1, 2, \ldots$, predstavljaju jedan od najzanimljivijih nizova matematici. Ovaj niz izazivao je fascinaciju još u staroj Indiji, a svojstva mu pročavaju matematičar

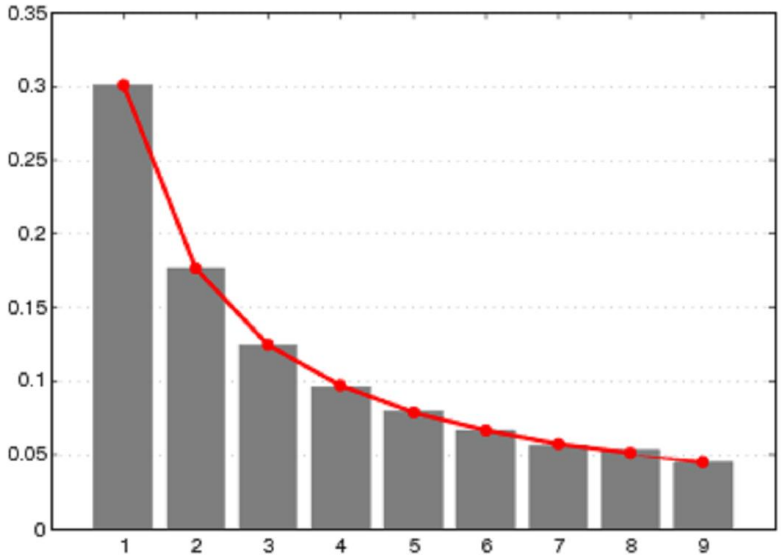
i danas. Brojevi F_n zadovoljavaju jednostavnu rekurziju

$$F_n = F_{n-1} + F_{n-2}\,,$$

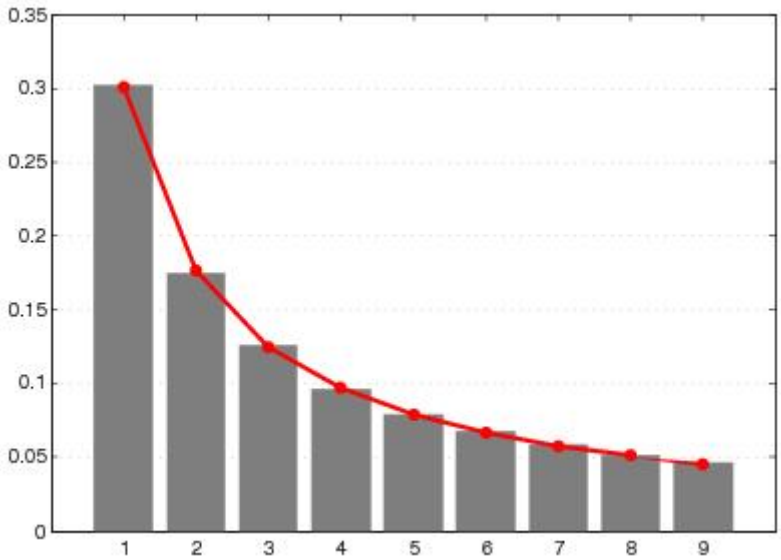
za sve brojeve $n \geq 2$, a pri tom je $F_0 = 0$ i $F_1 = 1$. Prisjetimo se inicijalni članovi niza su

$$0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, \ldots$$

Slika 1 donosi usporedbu razdiobe prve značajne znamenke za prvih 1000 članova niza (F_n) Benfordovom razdiobom. Na slici 2 promatramo razdiobu prve značajne znamenke niza (πF_n) Primjetimo da je i ovdje prisutna vrlo dobra podudarnost s Benfordovim zakonom, baš kako sm mogli očekivati na osnovu razmatranja iz prethodnog odjeljka.



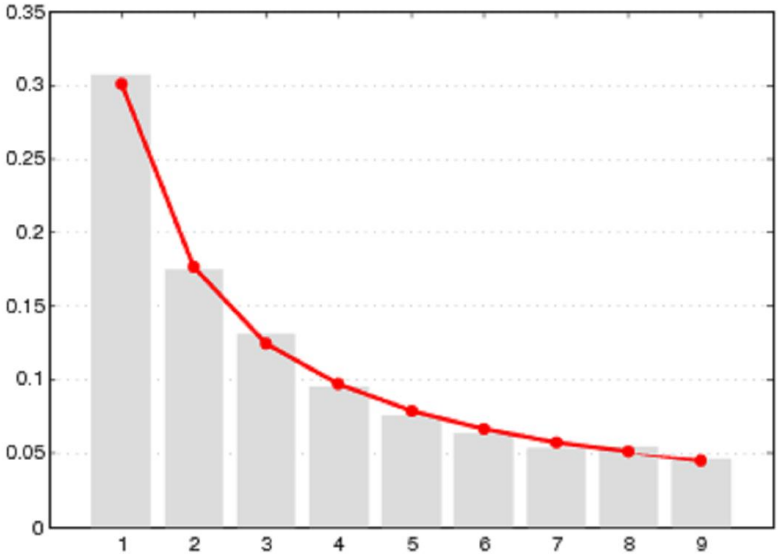
Slika 1: Usporedba Benfordove distribucije prve značajne znamenke (crvena linija) i distribucije prve značajne znamenke za prvih 1000 ($N = 1000$) članova Fibonaccijevog niza (sivi pravokutnici)



Slika 2: Usporedba Benfordove distribucije prve značajne znamenke (crvena linija) i distribucije prve značajne znamenke za prvih 1000 ($N = 1000$) članova niza (πF_n)(sivi pravokutnici).

Benfordov zakon možemo ilustrirati i na konkretnim podacima. Promatrat ćemo razdiobu prv značajne znamenke na podacima o broju stanovnika naselja u Hrvatskoj. Podaci su preuzeti iz baz podataka za posljednji popis stanovništva iz 2011. godine, Državnog zavoda za statistiku Promatrani skup ima oko 6700 podataka. Raspon podataka je reda veličine 10^5. Naime, prem

popisu stanovnistva iz 2011. godine, Zagreb je imao 686568 stanovnika, dok je najmanje naselje Hrvatskoj – Špigelski Breg, imalo tek jednog stanovnika. Iz grafa na slici 3 vidimo da i ovaj sku podataka vrlo dobro slijedi Benfordov zakon.



Distribucija prve značajne znamenke za 2011. godinu, koja je prikazana svijetlo sivim pravokutnicima, dok je Benfordova distribucija prve značajne znamenke prikazana crvenom linijom.

5 Benfordov zakon u primjeni

Gotovo od samog otkrivanja Benfordovog zakona postojala su nastojanja da ga se iskoristi razotkrivanju raznih prevara. Istraživanja Marka Nigrinija pokazuje kako se Benfordov zakon mož koristiti kao indikator u financijskim prevarama, npr. analizirajući koliko dobro isplate, uplate, iznos osiguranja itd. slijede Benfordovu distribuciju (vidi Nigrini [2]). Osim za financijske podatke, zako se pokazuje koristan i u otkrivanju falsificiranja znanstvenih i makroekonomskih podataka. Tako j npr. Rauch [3] na ovoj osnovi doveo u sumnju makroekonomske podatke koje je Grčka slala prij ulaska u Europsku Uniju. Slični razlozi, nedavno su natjerali ANZ (Australia & New Zealand Bankin Group) da posumnja u kineske ekonomske podatke o godišnjoj bruto domaćoj proizvodnji (BDP), čemu su izvjestili i mnogi svjetski mediji. Naglasimo ipak, ako podaci ne odgovaraju Benfordovor zakonu, to ne mora značiti da se njima manipuliralo. Unatoč tome, Benfordov zakon se u Ameri katkad koristi kao službeni dokaz i u sudskoj praksi.

Osim zbog manipulacije podacima, u praksi, podaci neće slijediti Benfordov zakon ako su ograničeni tako da počinju samo određenim značajnim znamenkama, kao npr. podaci o visini, kvocijentu inteligencije, opsegu glave ili rasponu ruku. Primjetimo, ti podaci su tipično približno normalno distribuirani. Nadalje, ako skupovi podataka imaju raspon kroz samo 1 ili 2 reda veličine (npr. podac su između 1 i 100), Benfordov zakon isto tako tipično neće vrijediti. Slično, zakon nije primjenjiv za podatke na koje je postavljen maksimum ili minimum. Da smo npr. na skupu podataka o broju stanovnika promatrali samo naselja koja imaju izmedu 500 i 3000 stanovnika, podudarnost podataka s Benfordovim zakonom bila bi puno slabija.

Bibliografija

[1]

Berger, A. i T.P. Hill: *A basic theory of Benford's Law.* Probability Surveys, 8:1–126, 2011.

[2]

Nigrini, M.:*Benford's Law: Aplications for Forensic Accounting, Auditing and Fraud Detection*, svezak 586. Wiley, 2012.

[3]

Rauch, B., M. Goettsche, G. Braehler i S. Engel : *Fact and Fiction in EU-Governmental Economic Data.* German Economic Review, 12(3):243–255, 2011.

ISSN 1334-6083