UDK: 343.53:658.14/.17 DOI: 10.7251/AP2106032J

BENFORDOV ZAKON I NJEGOVA PRIMJENA

Zoran Jasak[1]

Apstrakt

Predmet ovog rada je Benfordov zakon, odnosno statistička karakteristika velikih skupova numeričkih podataka. Cilj je čitaoce upoznati sa osnovnim svojstvima i prednostima ovog zakona, s obzirom na dosta nizak nivo znanja o njemu i mogućnostima njegove praktične i teorijske primjene na ovim prostorima. Benfordov zakon je logaritamski zakon koji opisuje dinamiku vodećih cifara u velikim skupovima numeričkih podataka. Detektovan je prije 140 godina, ali je ekspanzija njegove primjene nastupila razvojem računarske tehnike, posebno personalnih računara. Gotovo da nema područja ljudske djelatnosti i nauke u kojem nije detektovana mogućnost njegove primjene. Postao je jedan od ključnih alata za detekciju finansijskih prevara, kao i anomalija bilo koje vrste.

Ključne riječi: Benfordov zakon, signifikand, anomalije, prevare, statistički testovi.

Uvod

Postoji naizgled čudno pravilo da se u velikim skupovima numeričkih podataka vodeće cifre pojavljuju po tačnom pravilu. Ovo je uočio Simon Newcomb u svom članku na dvije strane (Newcomb, 1881). Ustvrdio je da se na vodećoj poziciji (prva slijeva) najčešće pojavljuje 1, zatim 2 itd. Zaključak je uslijedio na osnovu zapažanja da su logaritamske tablice pohabanije na početku nego na kraju, što mu je ukazivalo na naviku korišćenja brojeva.

Istu činjenicu je detektovao Frank Albert Benford (Benford, 1938), po kojem je zakon dobio ime. Benford je bio inžinjer u General electric kompaniji. Svoj zaključak je izveo na osnovu uzorka koji je obuhvatio preko 22.000 podataka iz raznih izvora (novinski članci, ulični brojevi, dužine rijeka, površine jezera, berzanski indeksi i slično). Za razliku od Simona Newcomba, dao je matematičku formulaciju koja je danas poznata u formi:

$$P[D = d] = \log_{10}\left(1 + \frac{1}{d}\right)$$

[1] Zoran Jasak, PhD, docent, Univerzitet „Bijeljina“ Pavlovića put bb, 76300 Bijeljina, Republika Srpska, BiH, E-mail: jasak_z@bih.net.ba

Ovdje $P[D = d]$ označava vjerovatnoću da je $d \in \{1, \ldots, 9\}$ vodeća cifra u skupu numeričkih podataka. Ove vjerovatnoće je izračunao i Newcomb, ali nije prezentirao formulu. Matematičkim metodama ova formula je proširena na cifre koje nisu vodeće te na grupe cifara na bilo kojoj poziciji.

Osim naziva Benfordov zakon, u literaturi se mogu naći nazivi Newcomb-Benford fenomen, Zakon prve cifre i slično.

Teorijske osnove i uslovi korišćenja

Zakon je teorijski dokazan od strane matematičara i statističara jer su ovaj problem shvatili kao izazov. Od samih početaka su poznati kriterijumi pod kojima se može vršiti bilo kakva analiza po osnovu ovog zakona.

Glavni slučajevi u kojima Benfordov zakon **ne vrijedi** su:

- Strukturirani brojevi kao što su telefonski brojevi, registarske oznake, brojevi računa, serijski (inventarni) brojevi, itd.;
- Uzorci malog obima. Smatra se da je potrebno minimalno 1000 brojeva za bilo kakvu analizu;
- Veličine istog tipa mjerene različitim mjernim jedinicama (npr. iznosi u različitim valutama, težine izražene u gramima i kilogramima, ...);
- Slučajni brojevi;
- Brojevi unutar jednog reda veličina (npr. visine odraslih osoba);

Teorijski razvoj traje i danas, zbog enormne dinamike rasta njegove primjene u velikom broju djelatnosti.

Jasno je pokazano da prirodni procesi slijede ovaj zakon. Biolozi su davno primijetili čvrstu vezu između prirode i Fibonačijevih brojeva. Radi podsjećanja, to je niz u kojem se svaki broj dobije kao suma prethodna dva: 1, 1, 2, 3, 5, 8, 13, 21, ... Primjer je broj latica u cvjetovima. Ono što čini vezu sa ovim zakonom je rekurzivna (samodefinišuća) priroda ovog niza. Kad su u pitanju prirodni procesi, Benfordov zakon vrijedi u svim slučajevima gdje se proces može opisati kao rekurzivni proces, a time i Fibonačijeve brojeve. U medicini je poznata veza ovog zakona i tzv. Weber-Fechner-ovog zakona koji opisuje reakciju na vanjski stimulans (svjetlost, zvuk, tonus mišića i slično).

Rast populacije bilo kog tipa ima pravilnu strukturu svojstvenu kategoriji populacije (ljudi, bakterije, biljke, ...). Ako se napravi spisak svih mjesta u nekoj državi sa brojem stanovnika, u uslovima odsustva manipulacije vodeće cifre tih brojeva slijede Benfordov zakon.

Praktična primjena Benfordovog zakona

Poznavanje pravila je osnova na kojoj se razvijaju metode detekcije odstupanja. Priroda i mjera tih odstupanja u velikim skupovima određuju se statističkim testovima. Razvoj računarstva, posebno personalnih računara početkom 80-ih godina donio je prekretnicu u kojoj je ovaj zakon iz teorijske sfere prešao u područje praktične primjene.

Prvi poznati primjer primjene ovog zakona je bila analiza poreskih prijava Bill Clintona, koju je uradio Mark Nigrini (Nigrini, 2011). Sam rezultat analize nije ukazivao na manipulacije značajnijeg obima. Ovaj slučaj je pokazao neslućene mogućnosti ovog zakona i uslijedila je ekspanzija koja još nije zaustavljena. Logično je da su prvi korisnici bile finansijske institucije (banke, berze, kartične kuće, ...). U godinama koje su uslijedile ovim zakonom su detektovane neke od najvećih korporativnih prevara (Enron, Parmalat, Societe General i duge).

Prostor teksta ne dozvoljava da se navedu svi slučajevi praktičnog korišćenja ovog zakona. Na web stranici https://www.benfordonline.net/ su dostupni gotovo svi do sada objavljeni tekstovi na ovu temu, uključujući i neke tekstove autora ovog teksta. Tekstovi su dostupni hronološki, od 1881. godine do danas, alfabetski po naslovima i po autorima.

Metode analize

Značajnija odstupanja od poznatog pravila u skupovima podataka detektuju se metodama statističkog testiranja. Zbog prirode ovog zakona, koji mjeri frekvencije, koriste se neparametarski testovi. U početku je korišten gotovo isključivo hi-kvadrat test. Vremenom su razvijeni novi testovi ili verzije postojećih testova, kao što su Hosmer-Lemeshow test, Wilcoxon, G-test, ... (Jasak, 2015; Jasak, 2016).

Testiranje se provodi na neki od sljedećih načina:

- Test vodećih cifara. Najjednostavniji je i služi kao osnova za dobijanje opšte informacije o uzorku;
- Test prve dvije cifre. Isti kao prethodni, ali obuhvata prve dvije cifre slijeva. Daje precizniju informaciju o izvoru eventualnog neslaganja. Ovaj i prethodni test su dali dobre rezultate u detekciji pokušaja pranja novca; prevarant puno koristi iznose bliske gornjoj granici iznad koje je obavezna provjera svih aspekata transakcije;
- Test prve tri cifre. Rijetko se provodi, osim u slučaju izuzetno velikih uzoraka;
- Test cifara ili grupa cifara na pozicijama unutar broja. Provodi se u situacijama kad se smatra da vodeće cifre nisu dobar izvor informacija. Primjer su birački spiskovi: broj glasača ne prelazi 1.000 po biračkom mjestu

- Zaokruživanje. Testiraju se zaokruživanja na posljednja dva cijela mjesta ispred zareza, što može ukazivati na sistematsko zaokruživanje iznosa;
- Test drugog reda. Uzorak se poreda rastućim redoslijedom, nađu se razlike dva uzastopna broja i provode prethodni testovi. Uvećana frekvencija neke od cifara ili grupe cifara može ukazivati na obrazac sistematskog zasijavanja iznosima u nekim pravilnim intervalima (1.625,00; 2.125,00; 2.625,00; ...). Onaj koji to radi računa da analitičar nema adekvatan mehanizam da to otkrije.

Iz pregleda je vidljivo da se Benfordov zakon ne bavi prirodom i izvorom brojeva koje su predmet analize. Detekcija odstupanja analitičaru je putokaz gdje da traži uzroke i razloge. Na tom koraku može doći do izražaja priroda podataka.

Analiza frekvencija je dugo bila jedini mehanizam, sve do uvođenja pojma **signifikand**. Svaki realni ***x*** broj može se napisati u obliku:

$$x = S(x) \cdot 10^E$$

Ovo je tzv. inžinjerski zapis, koji odgovara ispisu u formi sa jednim cijelim mjestom i eksponentom. Primjer je broj:

$$x = 15.678{,}35 = 1{,}567835 \cdot 10^4$$

U datom primjeru je $S(x) = 1{,}567835$, signifikand, dobijen pomicanjem zareza 4 mjesta ulijevo. U velikim skupovima brojeva ova konverzija se obavlja transformacijom:

$$S(x) = \frac{x}{10^{\lfloor \log_{10} x \rfloor}}$$

Ako se žele dobiti signifikandi sa 2 cijela mjesta u brojilac se stavi $10 \cdot x$.

Mark Nigrini (Nigrini, 2011) je prvi ustvrdio da ako se svi brojevi iz uzorka transformišu u signifikande, a zatim saberu signifikandi koji imaju jednake vodeće cifre, te sume su očekivano jednake. Na ovaj način je dobijen novi način detekcije anomalija. Cifre mogu imati frekvencije koje su u skladu sa teorijskim modelom, ali sume koje se značajno razlikuju mogu ukazivati na korištenje ekstremnih vrijednosti.

Jedan od aspekata je primjena na podatke na vremenskoj skali. U nekim područjima rada je od bitne je važnosti da se u obzir uzme vremenski raspored podataka. Primjer su podaci o dnevnim varijacijama vlažnosti i temperature, koji se ne mogu tretirati kao skup slučajnih brojeva bez poretka. Finansijske transakcije se obavljaju računo određenim redom koji ne može i ne smije biti zanemaren. Za ove potrebe je razvijena posebna veličina, tzv. G-konstanta. Za normalnu raspodjelu ova konstanta je $\sqrt{2}/2$; za uniformu raspodjelu (odnosno za stvarno slučajne brojeve) ova konstanta

je $4/3$. Pripremljen je tekst ovog autora u kojem se izlaže formiranje i upotreba ove konstante za Benofordov zakon.

Zaključak

Benfordov zakon predstavlja trenutno najmoćniju metoda detekcije anomalija i analize velikih skupova numeričkih podataka. Kako je rečeno, odstupanje od pravila samo po sebi ne mora značiti i prevaru, ali sigurno ukazuje na situacije koje treba posebno istražiti sa stanovišta izvora i strukture podataka.

Poseban kvalitet ovog zakona, koje ga izdvaja od drugih, je njegova objektivnost i nepristrasnost, izražena formulom koja se jednostavno realizuje u računskom i proceduralnom smislu. Numerički podaci se uzimaju bez pretpostavki o njihovoj prirodi i izvoru, koja postaje važna nakon provedenog testiranja i drugih analiza. Ovo je razlog sve većeg zanimanja stručnjaka svih profila, koji traže načine njegove primjene. Sigurno je da još nisu u potpunosti sagledane sve mogućnosti njegove primjene.

Literatura

1. Newcomb, S. (1881): *Note on the Frequency of use of different digits in natural numbers*, American Journal of Mathematics. Vol. 4. No. 1/4. pp. 39-40
2. Benford, F.A. (1938): *The Law of Anomalous Numbers*, Proceedings of the American Philosophical Society. Vol. 78. No. 4 p. 551-572
3. Nigrini, M. (2011): *Forensic Analytics – Methods and Techniques for Forensic Accounting Investigations*, pp. 144–146. Wiley, Hoboken
4. Jasak, Z. (2015): *Benford's law and arithmetic sequences*, Journal of Mathematical Sciences: Advances and Applications. Volume 32. pages 1 - 16. ISSN 0974-5750
5. Jasak, Z. (2016): *Benford's Law and Hosmer-Lemeshow test*, Journal of Mathematical Sciences: Advances and Applications. Volume 41. Pages 57-73.
6. https://www.benfordonline.net/ (Pristupljeno 15.05.2021.)

BENFORD'S LAW AND ITS APPLICATION

Zoran Jasak[1]

Abstract

Subject of this paper is Benford's law, statistical characteristics of big data sets. Goal is to introduce users with basic properties and advantages of this law, having in mind relatively low level of knowledge about this law and of possibilities its practical and theoretical application on our region. Benford's law is logarithmic law which describes dynamics of leading digits in big data sets. It's detected 140 years ago, but expansion of practical use has started by development of computer technique, particularly of personal computers. There is almost no area of human activity and science in which possibility of it's application is not detected. Benford's law became one of key tools for detection of financial frauds, and anomalies of any kind as well.

Key words: Benford's law, significand, anomalies, frauds, statistical tests.

[1] Zoran Jasak, PhD, Assistant professor, Bijeljina University, Pavlovica put bb, 76300 Bijeljina, Republic of Srpska, BiH, E-mail: jasak_z@bih.net.ba