JU UNIVERZITET U TUZLI
PRIRODNO-MATEMATIČKI FAKULTET
Odsjek za matematiku

Zoran Jasak

BENFORDOV ZAKON
I
REINFORCEMENT UČENJE

MAGISTARSKI RAD

Tuzla, Decembar 2010. godine

Rad je izrađen u 5 primjeraka
UDK Narodne i univerzitetske biblioteke „Derviš Sušić“ u Tuzli : 519.21:004.85
Mentor rada : Naser Prljača
Rad ima 122 stranice, 27 dijagrama, 8 slika i 18 tabela

Text is made in 5 copies
UDK Public and University library „Derviš Sušić“ in Tuzla : 519.21:004.85
Mentor : Naser Prljača
Text has 122 pages, 27 diagrams, 8 pictures and i 18 tables

Sadržaj

Sažetak

Benfordov zakon izražava vjerovatnoću pojave cifara na vodećim pozicijama brojeva. Nazvan je po Franku Albertu Benfordu (1938) koji je formulisao matematički model za ovu vjerovatnoću. Isto pravilo prije njega je uočio Simon Newcomb (1899). Ovaj zakon je u velikoj mjeri promijenio ustaljenu pretpostavku da se sve cifre sa istom vjerovatnoćom javljaju na svim pozicijama broja.

Cilj rada je pokazati mogućnosti, prednosti i nedostatke korištenja Benfordovog zakona u metodama reinforcement učenja. Posebno, cilj je analizirati uticaj izbora veličina formiranih na osnovu Benfordovog zakona a koje su definisane u patentnoj prijavi Fletcher Lu i Efrim Boritza, u kojoj su formulisali Adaptivnu Benfordovu metodu. Ova veličina se koristi kao odziv za pojedino stanje u metodi reinforcementu učenja. U ovom tekstu se predlaže korištenje odziva koji se dobija kao količnik veličina za tri i dvije vodeće cifre.

Ključne riječi : Benfordov zakon, reinforcement učenje, intervali povjerenja, nivo kontaminacije, Adaptivna Benfordova metoda, Q učenje, Sarsa, akcija, stanje, količnik.

Abstract

Benford's law describes probability for leding digits of number. It's named by Frank Albert Benford (1938) who stated mathematical model of this probability. Before him, the same observation is made by Simon Newcomb. This law has changed usual preasumption of equal probability of each digit on each position in number.

Main goal of text is to present posibilities, advantages and disadvantages od Benford's in reinforcement learning methods. Particulary, goal is to analyze influence of values made by Benford's law and defined in patent application by Fletcher Lu and Efrim Boritz, in which they formulated Adaptive Benford's method. This value is used as reward for statest in reinforcement learning. In this text using quotients of values for three and two leading digits is proposed.

Key words : Benford's law, reinforcement learning, confidence intervals, contamination level, Adaptive Benford's method, Q learning, Sarsa, action, state, quotient.

1 Uvod

1.1 Istorijat

Godine 1881. astronom Simon Newcomb je primijetio da su logaritamske tablice, koje su u to vrijeme korištene za numerička izračunavanja, u prvom dijelu očito bile više uprljane i korištene od ostalih. Smatrao je da je to stoga što se brojevi koji počinju sa 1 ili 2 koriste mnogo češće od ostalih.

"*Da se svih deset cifara ne pojavljuje sa jednakom frekvencijom mora biti jasno svakome ko koristi logaritamske tablice i ko primijeti koliko brže se početne stranice habaju od onih na kraju. Cifra je* 1 *je mnogo češća prva cifra dok se frekvencije ostalih cifara do* 9 *smanjuju*".

Bez objašnjavanja ili primjera zaključio je da frekvencije zadovoljavaju zakon koji se formuliše izrazom $P\{D = d\} = \log_{10}(1 + 1/d)$. Ova zapažanja, koja je objavio u matematičkom časopisu American Journal of Mathematics (Vol. 4., No. 1, p. 39-40) u članku pod nazivom "Note on the frequency of use of the different digits in natural numbers", bila su potpuno ignorisana skoro 60 godina.

Godine 1938. Frank Benford, fizičar u General Electric Research Laboratories, došao je do istog saznanja kao i Newcomb ali je nastavio sa razradom ideje [1]. Napravio je analizu frekvencije cifara po pozicijama na skupu od 20.229 slučajeva iz 20 različitih izvora, od novinskih stranica do kućnih adresa. Veličine uzoraka su bile od 91 za atomske težine do 5000 za pojmove iz matematičkih priručnika. Uočio je da se kao prva cifra najčešće pojavljuje 1, zatim 2 i tako dalje. Za razliku od Newcomba, Benford je, koristeći relativne frekvencije, formulisao matematički model i po njemu se odnosi koje je odredio zovu Benfordov zakon. Ovaj zakon se pojavljuje pod raznim nazivima i to *Zakon prve cifre*, *Analiza cifara*, *Newcomb-Benfordov fenomen* ili slično. U literaturi se do 80-ih godina koristio termin *Logaritamski zakon.*

Neki aspekti iz sažetka njegovog teksta su [1] :
Analiza brojeva iz raznih izvora pokazuje da brojevi uzeti od nepovezanih objekata kao što su grupe novinskih članaka, pokazuju mnogo bolju saglasnost sa logaritamskom distribucijom nego brojevi iz matematičkih tabela ili drugih formalnih skupova podataka. Postoji čudna činjenica da su brojevi, koji pojedinačno nisu u vezi kada se posmatraju u velikim grupama u velikoj saglasnosti sa zakonom distribucije, odakle i naziv Anomalični brojevi.
***
Dalja analiza podataka pokazuje strogu tendenciju grupa numeričkih podataka da prate geometrijske nizove. Ako su nizovi sastavljeni od brojeva koji imaju tri ili više cifara prve cifre formiraju logaritamske nizove. Ako brojevi sadrže samo jednu cifru još uvijek vrijedi geometrijska veza ali ne vrijedi strogi logaritamski odnos.
***
Izraz takođe daje frekvenciju cifara na drugoj, trećoj,... poziciji višecifrenih brojeva i pokazano je da isti zakon vrijedi za recipročne vrijednosti.

U tekstovima se veoma rijetko pominje zapažanje da isti zakon vrijedi i za recipročne vrijednosti !

Uloženo je puno napora da se pruži adekvatno objašenje ovog fenomena [2]. Bez obzira na kontraprimjere koji se mogu naći, neki autori tvrde da je ovaj zakon inherentan rezultat brojnog sistema odnosno da je u njegovoj osnovi. Postoje brojne ilustracije ovakvog stava. Prvi primjer je tablica množenja brojeva od 1 do 9. Udio cifre 1 na prvoj poziciji svih proizvoda je $18/81 = 2/9$. Ovo je manje od 30% što bi se moglo očekivati prema Benfordovom zakonu ali je ipak duplo više od 1/9 koliko se može intuitivno očekivati u smislu očekivanja da se svaka cifra na svakoj poziciji može pojaviti sa jednakom vjerovatnoćom. Ako se ova analogija prenese na neprekidnu okolinu moguće je pokazati da frekvencije prvih cifara proizvoda n realnih brojeva iz intervala $[1, 10)$ zadovoljavaju Benfordov zakon ako se n (broj umnožaka) neograničeno povećava.

Jedna od ilustracija iz realnog života su kamate. Da bi svota od 100 NJ narasla na 200 NJ uz stopu od 10% treba proći nešto više od 7 godina. Suština ove kalkulacije je da se prva cifra promijeni sa 1 na 2. Da bi suma narasla sa 200 na 300 NJ potrebno je novih 4 godine. Očigledno je da će najmanje vremena trebati da svota naraste sa 900 na 1000 NJ. Da bi svota narasla sa 1000 na 2000 treba novih 7 godina itd. Ako se uzme bilo koja druga početna svota prva cifra svih stanja u odabranom momentu mora približno zadovoljavati Benfordov zakon. Može se reći da zakon vrijedi za bilo koju geometrijski niz oblika $a \cdot q^n$ osim slučaja kada je q stepen od 10.

Mark Nigrini i drugi autori su formulisali kriterije koji podaci mogu biti predmet analize putem ovog zakona [3, 5, 6]. Pokazano je da ovaj zakon vrijedi za numeričke podatke, sa ili bez decimala, koji opisuju prirodni proces nad kojim se ne postavljaju ograničenja u pogledu minimalnih ili maksimalnih vrijednosti, koji po svojoj prirodi imaju raspon vrijednosti u najmanje dva reda veličine baze i koji nisu strukturirani. Primjer procesa koji mogu generisati ovakve podatke su cijene vrijednosnih papira na berzama, finansijske transakcije svih vrsta, kartično poslovanje, telekomunikacioni sistemi, računarski sistemi, procesi koji se opisuju rekurentnim izrazima, euklidske distance u pojedinim vrstama problema, prirodni procesi rasta populacija biljaka i životinja i bezbroj drugih primjera u prirodi, medicini i tehnici, uključujući brojne primjere iz teorijske matematike. Predmet analize ne mogu biti kategorijski podaci ali mogu njihove frekvencije. Mogući primjeri primjene su navedne u posebnom poglavlju ovog teksta.

Benfordov zakon ne ulazi u prirodu izvora numeričke veličine koja je predmet analize. Naprimjer, veličina $128,9456$ može označavati iznos, dužinu, distancu, vrijednost berzanskog indeksa, temperaturu, električni napon, količnik ili može imati neko drugo značenje. Mada sa stanovišta Benfordovog zakona brojevi $12.894,56$ i $1,289456$ imaju iste početne značajne cifre očigledno je da one, u kontekstu određenog problema imaju bitno drugačije značenje. Sa stanovišta analize putem ovog zakona izvor numeričkog podatka igra ulogu u momentu interpretacije rezultata analize. Drugim riječima, postupak testiranja je potpuno nezavisan od izvora samog podatka odnosno njegove prirode i značenja. Ova karakteristika ovom zakonu daje dimenziju objektivnosti koja nedostaje mnogim drugim metodama i koja je osnovni razlog sve većeg zanimanja. Benfordov zakon je u međunarodnoj praksi priznat kao vjerodostojan revizorski metod. Ugrađen je u gotovo sve specijalističke revizorske programske alate.

1.2 Mašinsko učenje

Moderna nauka i inžinjering su bazirani na korištenju principijelnih modela kako bi se opisali fizički, biološki i socijalni sistemi [4]. Takav pristup počinje sa osnovnim naučnim modelom kao što su Njutnovi zakoni kretanja ili Maxwell-ove jednadžbe elektromagnetizma a zatim se nadograđuje njihovom različitom primjenom u mehaničkom ili električnom inžinjerstvu. U tom pristupu se koriste eksperimentalni podaci kako bi se verifikovali modeli i napravila procjena nekih parametara za koje je direktno mjerenje teško ili čak nemoguće. Međutim, u mnogim domenima principi koji stoje iza procesa su nepoznati ili su sistemi koji su predmet istraživanja previše kompleksni kako bi bili matematički formalizirani. Rastuće korištenje računara je dovelo do generisanja ogromnih količina podataka. U odsustvu primarnih modela takvi stalno raspoloživi podaci mogu biti korišteni kako bi se izveli modeli putem procjene korisnih relacija (tj. nepoznate ulazne - izlazne zavisnosti). Stoga je prisutan pomak od klasičnog modeliranja i analiza baziranih na primarnim modelima ka razvoju modela i odgovarajućih analiza direktno iz podataka.

Postepeno se navikavamo na činjenicu da na našim računarima i mrežama postoje ogromne količine podataka. Vladine agencije, naučne institucije i poslovno okruženje su sve svoje resurse posvetili prikupljanju i pohrani podataka. U stvarnosti, samo mali dio tih podataka će biti ikada iskorišten u svrhu analiza jer su, u mnogim slučajevima, smještajni kapaciteti naprosto preveliki za upravljanje ili su strukture podataka previše kompleksne same po sebi kako bi bile efektivno analizirane. Primarni razlog za ovo je što se osnovni napor kreiranja skupa podataka često fokusira na pitanja kao što je efikasnost pohrane. Ovaj napor ne uključuje plan kako će podaci eventualno biti korišteni ili analizirani.

Potreba da se razumiju veliki, kompleksni i informacijama bogati skupovi podataka postoji u skoro svim poljima poslovanja, nauke i inžinjerstva. U poslovnom svijetu korporativni i korisnički podaci se počinju prepoznavati kao strateška imovina. Sposobnost da se izvuče korisno znanje sakriveno u tim podacima i djelovati na osnovu tog znanja dobija rastuću važnost u današnjem konkurentskom svijetu.

Postoje dva osnovna tipa metoda induktivnog učenja poznata pod nazivom nadzirano i nenadzirano učenje [4].

Nadzirano učenje se koristi kako bi se procijenila nepoznata zavisnost iz poznatih ulazno izlaznih primjera. Klasifikacija i regresija su poslovi podržani ovim tipom induktivnog učenja. Nadzirano učenje pretpostavlja postojanje funkcije učenja ili druge eksterne metode procjene predloženog modela. Pojam 'nadzirani' označava da su izlazne vrijednosti u uzorcima za treniranje poznati (tj. služe kao 'učitelji').

U šemi nenadziranog učenja uzimaju se samo uzorci sa ulaznim vrijednostima i ne postoji pretpostavka o izlazu tokom procesa učenja. Nenadzirano učenje eliminiše 'učitelja' i zahtijeva da onaj ko uči sam formira i evaluira model. Cilj nenadziranog učenja je da otkrije 'prirodnu' strukturu u ulaznim podacima. U biološkim sistemima percepcija je zadatak koji se uči putem nenadziranih tehnika. S obzirom da se analiza na osnovu Benfordovog zakona oslanja na matematički model, u smislu datog tumačenja, taj zakon spada u metode nenadziranog učenja. Ovaj matematički model je izveden najprije empirijski, na primjerima iz svakodnevnog života, a zatim su matematičari i drugi analitičari isti zakon izveli i teorijski.

Treći tip mašinskog učenja je reinforcement učenje. To je učenje šta uraditi - kako mapirati situaciju u akcije - tako da se maksimizira numerička vrijednost signala odziva (reward) [9]. Onome koji uči (learner) ne govori se koje akcije treba preduzeti, kako je slučaj u većini formi mašinskog učenja već umjesto toga mora otkriti putem pokušaja koje akcije daju najveći odziv. U najinteresantnijim i najizazovnijim slučajevima akcije mogu uticati ne samo na neposredni odziv već i na sljedeću situaciju i, putem toga, na naredne odzive. Ove dvije karakteristike - metoda pokušaja i greške - i odgođeni odziv - dva su najvažnija i najistaknutija svojstva metode reinforcement učenja.

1.3 Benfordov zakon i metode mašinskog učenja

Evolucija primjene Benfordovog zakona je bila spora i duga [2]. Poznati ekonomist Hal R. Varian je u pismu čitalaca (1972) izdavaču sugerisao da se ovaj zakon može koristiti kako bi se testirala prihvatljivost ekonomskih modela. Njegovo rezonovanje je bilo sljedeće :

Na kraju krajeva, Benfordov zakon je samo zanimljiv empirijski, gotovo numerološki fenomen; šta bi on trebao imati sa ekonomskim prognozama ? Međutim, mora se primijetiti da ako su ulazni podaci zadovoljavali Benfordov zakon a izlazni podaci nisu to bi moralo stvoriti osjećaj traćenja vremena zbog nesigurnosti korištenja takvog izlaza" (Varian [1972], p. 65).

Ova kratka sugestija nije privukla previše pažnje. Do kraja 1980-ih ovaj zakon je bio najvećim dijelom predmet matematičke teorije, bez posebnih praktičnih primjena. Stvarne primjene Benfordovog zakona datiraju iz druge polovine 80-ih. Carslaw [1988] je koristio ovaj zakon u dijelu koji se odnosi na drugu cifru kako bi pokazao da kompanije na Novom Zelandu sistematski zaokružuju podatke o prihodima. Cifra 0 se na drugoj poziciji javljala sa frekvencijom koja je bila mnogo veća u odnosu na očekivanja dok se cifra 9 pojavljivala mnogo manje od očekivanja. Studija je u SAD ponovljena i proširena i dala je suprotne rezultate u smislu da su detektovana zaokruživanja na manje iznose u izvještajima o gubicima. Dodatno, zarade po dionicama (Earnings Per Share) u SAD-u su u najvećem dijelu bili multiplikatori od 5 dok su se brojevi završavali cifrom 9 mnogo manje nego što bi se to moglo očekivati.

Osnovna priroda Benfordovog zakona je takva da daje jednu vrstu generalne ocjene cijelog skupa odnosno svih instanci jednog atributa nekog skupa. Ovo je aspekt koji je, u velikoj mjeri, bio povod za pisanje ovog teksta. Posebni segmenti koji su obrađeni u ovom tekstu su rad sa dijelom nedostajućih vrijednosti i mašinsko učenje. Rad sa djelimično dostupnim podacima je jedan od praktičnih problema primjene Benfordovog zakona. Kad je u pitanju mašinsko učenje, Benfordov zakon je našao primjenu u kalkulaciji parametara funkcija reinforcement učenja i vrijednosti na osnovu kojih se kreiraju politike ponašanja. Praktična (realna) primjena metode reinforcement learninga je dokazana mnogim primjerima a posebno, ponovo, u detekciji anomalija, sa trendom stalnog rasta praktičnih primjena. Svaki od željenih ciljeva rada obrazložen je u nastavku.

- Frekvencija grupa cifara računata prema Benfordovom zakonu je osnov za proračun stepena anomalije uzorka

Theodor Hill je napravio eksperiment u kojem je zamolio 742 studenta da zamisle šestocifrene slučajne brojeve i zapišu ih na komad papira. Skupio je odgovore i našao da ovi brojevi ne slijede Benfordov zakon [6]. Busta i Weinberg su, u tekstu u kojem su istraživali način klasifikacije podataka na osnovu Benfordovog zakona korištenjem neuronskih mreža, pošli od pretpostavke da bilo kakvi namješteni skupovi slijede Hillovu distribuciju s obzirom na pretpostavljeni sličan kognitivni proces [7]. Ista tzv. Hillova distribucija, korištena je i u drugom tekstu koji se bavi korištenjem Benfordovog zakona u neuronskim mrežama. U oba ova teksta korišten je termin 'nivo kontaminacije' (contamination level). Osim činjenice da nije empirijski i teorijski podržana, bitan nedostatak izbora tzv. Hillove distribucije je u činjenici da je kontaminacija simulirana odnosno unijeta u sam skup za koji se pretpostavlja da slijedi Benfordov zakon. S obzirom da nije moguće naći njenu egzaktnu formulaciju pretpostavka je da se radi o uzorku iz uniformne raspodjele.

Sa stanovišta praktične primjene, podatak o stepenu anomalije može biti bitan u analizama koje imaju za cilj eliminaciju anomaličnih podataka, njihovu detekciju kao osnovni cilj ili izbor prikladne tehnike analize. U smislu prve primjene može se navesti eliminacija ponavljajućih vrijednosti, eliminacija vrijednosti iznad nekog praga i slično. Detekcija prevara je najočigledniji primjer primjene u smislu detekcije anomalija kao osnovnog cilja.

U ovom tekstu se predlaže pristup u kojem se mogući stepen anomalije identifikuje na osnovu samog skupa, bez korištenja simuliranih distribucija ili eksternih mehanizama. Granice intervala povjerenja daju mogućnost procjene teorijskih gornjih i donjih granica frekvencija cifara po pozicijama. Zbog diskretne prirode ovog zakona, na osnovu teorijskih granica intervala povjerenja moguće je dati procjenu broja slučajeva odstupanja koja su značajna u smislu da izlaze iz procijenjenih intervala odnosno udjela ovih slučajeva u cjelokupnom uzorku. Taj dio uzorka se može nazvati stepen anomalije. S obzirom na prirodu zakona na kojem je zasnovan, može se ukazati na mogući obim uzorka koji izlazi van granica ali ne i na pojedinačne elemente uzorka.

- Izbor donjih i gornjih granica uzorka ima bitan uticaj na rezultat analize cifara putem Adaptivne Benfordove metode

Jedan od zahtjeva za primjenu Benfordovog zakona je da ne postoji ograničenje u smislu minimuma ili maksimuma [5]. Kada analiza obuhvata dio podataka to ne mora neophodno značiti da podaci ne slijede Benfordovu distribuciju; to jedino znači da skup nije kompletan. Unatoč tome, još uvijek je poželjno primijeniti Benfordov zakon za analizu cifara ako je moguće kako bi se, osim činjenice da podaci nedostaju, uočile anomalije. U Septembru 2008. godine patentirana je Adaptivna Benfordova metoda putem koje se vrši analiza cifara na djelimičnom odnosno maksimalno raspoloživom skupu. Metoda je prezentirana na način da se simuliraju slučajevi ograničenja donje i gornje granice i uticaj takvih ograničenja na bitne parametre koji se koriste za analizu cifara. Posebno, bitno je ustanoviti da li i koje bitne razlike u analizama postoje ako se kao vještačka ograničenja biraju veličine 10^k, gdje je $k \in \mathbb{Z}$ i kad se kao granice biraju proizvoljni brojevi iz intervala $\left[10^k, 10^{k+1}\right]$, $k \in \mathbb{Z}$.

U jednom broju tekstova vezanih za korištenje Adaptivne Benfordove metode u detekciji prevara Fletcher Lu je formulisao veličinu

$$BE\left(i\right)=\frac{f_{1i}}{b_{1i}}+\frac{f_{2i}}{b_{2i}}+\frac{f_{3i}}{b_{3i}} \qquad (1.1)$$

U ovom izrazu f_{ji} predstavlja uzoračku a b_{ji} teorijsku frekvenciju grupa cifara dužine j za stanje (slog) i. Ako se formira veličina $BE\left(2\right)=f_1/b_1+f_2/b_2$ tada je moguće računati količnik

$$BK32=\frac{BE\left(3\right)}{BE\left(2\right)}=1+\frac{f_3/b_3}{f_1/b_1+f_2/b_2} \qquad (1.2)$$

Ovaj količnik mjeri uticaj treće cifre u frekvencijama. Po istoj analogiji moguće je mjeriti uticaj četvrte, pete,... cifre ali se u praksi sa njima ne radi, osim možda u nekim područjima u kojima je bitna preciznost u vidu većeg broja decimala.

U uzorcima koji odražavaju regularne procese u kojima se ne postavljaju smetnje u smislu odbarane donje ili iznad odabrane gornje granice, ovaj količnik se poklapa sa zapisima za koje je $BE\left(3\right)$ najveći. Simulacija donjih i gornjih granica pokazuje da neslaganje počinje ako se donje granice uzimaju iznad veličina za koje frekvencije prvih cifara prelaze prag značajnosti odnosno bitno narušavaju Benfordov zakon. Drugim riječima, neslaganje veličina $BE\left(3\right)$ i $BK32$ daje mogućnost detekcije u skupu, bilo u smislu nedostajućih (namjerno ili nenamjerno izostavljenih / zanemarenih) podataka ili u smislu neregularnosti procesa koji se opisuje tim podacima. Česta je praksa da se u analizama ove vrste odbacuju neke male veličine (npr. iznosi manji od 10 NJ). Onaj ko to radi mora znati da li takav postupak može imati bitnog uticaja na rezultate analiza.

Neke metode data mininga se bave problemom nedostajućih vrijednosti, bilo da se radi o potrebi njihove tačne ili približne procjene ili načina da se tehnike prilagode činjenici da vrijednosti ne postoje i da nije moguća njihova rekonstrukcija bez uticaja na rezultete analiza. Benfordov zakon ne može biti korišten za identifikaciju odnosno procjenu nedostajućh pojedinačnih vrijednosti po nekom numeričkom atributu.

- Izbor kriterija Benfordovog koeficijenta ili Benfordovog količinika za početni atribut u algoritmu reinforcement učenja ima uticaj na politiku pretraživanja.

Reinforcement učenje je učenje šta uraditi - kako mapirati situaciju u akcije - tako da se maksimizira numerička vrijednost signala odziva (reward) [9]. Onome koji uči (learner) ne govori se koje akcije treba preduzeti, kako je slučaj u većini formi mašinskog učenja već umjesto toga mora otkriti putem pokušaja koje akcije daju najveći odziv. U najinteresantnijim i najizazovnijim slučajevima akcije mogu uticati ne samo na neposredni odziv već i na sljedeću situaciju i, putem toga, na naredne odzive. Ove dvije karakteristike, metoda pokušaja i greške i odgođeni odziv, dva su najvažnija i najistaknutija svojstva metode reinforcement učenja. Nadzirano učenje samo po sebi nije adekvatno za učenje iz interakcija. U interaktivnim problemima je često nepraktično pribavljati primjere željenog ponašanja koje je bilo ispravno ili je reprezentativno za sve situacije u kojima agent mora djelovati. Na nepoznatom terenu, gdje bi se moglo očekivati najuspješnije učenje, agent mora biti sposoban učiti iz sopstvenog iskustva. Reinforcement učenje se definiše ne samo karakterizacijom metoda učenja već i karakterizacijom problema. Za

bilo koji metod koji je prikladan za rješavanje problema na ovaj način smatra se da je reinforcement metod učenja.

Nezaobilazni element svake metode reinforcement učenja su odzivi, numeričke veličine koje karakterišu pojedino stanje. Benfordov zakon se po svojoj prirodi veže za uzorke velikog obima i određen tip numeričkih veličina. Primjer takvih uzoraka su npr. bankarske transakcije. Reinforcement učenje je na ovakvim uzorcima moguće koristiti na način da su stanja pojedine stavke a akcije atributi pojedinog zapisa (sloga).

Kako je rečeno, Fletcher Lu je, u okviru svoje patentne prijave Adaptivne Benfordove metode, predložio veličinu $BE\,(3)$ datu sa (1.1) kao parametar za primjenu Benfordovog zakona u metodama reinforcement učenja na način da startno stanje bude jedno iz skupa stanja za koje je ova veličina najveća. U ovom tekstu se predlaže da se, alternativno, startno stanje odabire po osnovu količnika $BK32$, koji definisan izrazom (1.2) najveći. Ako se gornja i donja granica mijenjaju, ovaj količnik ukazuje na mogući drugačiji, u pravilu širi, skup zapisa od onoga na koji ukazuje veličina $BE\,(3)$ u kompletnom uzorku i daje slogove koji ne moraju u svim slučajevima odgovarati onima za koje je $BE\,(3)$ najveći. Odstupanje je posebno značajno ako se povećava donja granica uzorka (simulirajući pri tom činjenicu da u uzorku nedostaje određena kategorija podataka). Ako se ovaj količnik koristi kao kriterij za početak tada se mijenja i politika reinforcement učenja koja nije ista kao za slučaj veličina $BE(3)$.

Osim najveće vrijednosti za $BE(3)$ kriterij izbora može biti odabrani raspon ovih vrijednosti, čime se dobijaju nešto drugačije politike. Po istoj analogiji se kao kriterij starta mogu uzeti rasponi vrijednosti $BK32$. Normalizacija po metodi Nigrinija daje mogućnost da se koriste svi slogovi. Odbacivanje slogova ispod nekih granica može imati odlučujući uticaj na pretraživanje, što baca novo svjetlo na cijeli problem (anomalije se mogu odnositi i na male vrijednosti).

Na istim uzorcima se primjenjuje metoda Q-učenja, sa dvije varijante izbora početnih kriterija. Za jednu će kao kriterij biti korištena veličina $BE\,(3)$ a za drugu veličina $BK32$. Cilj je pokazati da ovakav izbor ima bitan uticaj na rezultujuću politiku.

1.4 Struktura rada

Tekst rada je podijeljen u četiri dijela.

U prvom dijelu je prezentiran matematički aspekt Benfordovog zakona odnosno različiti pristupi koji su doveli do istog zakona. Na ovaj način je ujedno ilustrovana bogata teorijska podloga i velike teorijske mogućnosti ovog zakona.

U drugom dijelu je pregled testova koji se koriste u analizama skupova, sa njihovim osnovnim karakteristikama. Prezentirani su neki testovi i odgovarajući teorijski okvir.

U trećem dijelu je pregled primjera primjene ovog zakona. Iz navedenog je moguće vidjeti da gotovo nema naučnog ili praktičnog područja u kojem nije moguća primjena Benfordovog zakona.

U četvrtom dijelu su prezentirani rezultati kojima se postavljene hipoteze potvrđuju. U prvom primjeru je da prikaz kalkulacije stepena anomalije. Pod tim terminom se podrazumijeva dio uzorka koji izlazi iz intervala povjerenja. Posebno, vrši se poređenje kalkulacija stepena anomalije na osnovu klasičnog i adaptivnog Benfordovog zakona na istom uzorku. U drugom primjeru je analiza uticaja izbora granica na primjenom Adaptivne Benfordove metode. Uzimaju se razni slučajevi donjih i gornjih granica i posmatra kretanje nekih parametara opšteg tipa (MAD) te veličina BE i BK. U trećem primjeru je data razrada korištenja Benfordovog zakona za reinforcement učenje. Provedeni su testovi kojima se pokazuje uticaj izbora (opsega) veličina $BE\,(3)$ i $BK32$ na rezultujuće politike.

Metode su prezentirane na podacima iz stvarnih sistema ili su simulirani odnosno generisani alatima koji su opisani ili su njihova mješavina. U svim analizama je korišten programski paket MS Excell 2007 sa makroima koji su urađeni alatom VBA.

1.5 Zaključak

Svakim danom sve je više radova u kojem se pokazuju raznovrsne i nove mogućnosti Benfordovog zakona. S obzirom na njegovu prirodu, fokus primjene je većinom usmjeren na anomalije što je u uskoj vezi sa detekcijom prevara. Sa stanovišta metoda data mininga, ovo je postupak detekcije anomalija (outliera). Sa druge strane, pokazuje se da ovaj zakon može biti korišten i u metodama za koje se na prvi pogled ne čine kao prikladne za ovaj zakon. Primjer su mašinsko učenje i neuronske mreže.

S obzirom na prirodu podataka koji se koriste, ovaj zakon može naći svoje mjesto u svim metodama data mininga koje koriste distance (distance based), kao što su problem trgovačkog putnika, klastering i slično.

Priroda Benfordovog zakona daje mogućnost njegovog korištenja u analizama podataka koji imaju vremensku dimenziju. Ovo je vidljivo iz radova u kojima je fokus na finansijskim podacima koji nastaju kao dio poslovanja u vremenu. Drugi primjer su cijene dionica, procesi rasta i slično. Ovo ukazuje na mogućnost njegovog korištenja u analizama vremenskih serija.

U nekim metodama analize podataka kao što su mašinsko učenje, neuronske mreže ili klasifikacija, prirodno je koristiti veličine izvedene na osnovu Benfordovih svojstava. Primjer je veličina reward u metodi reinforcement učenja.

Najveća prednost ovog zakona je jednostavnost implementacije. Za najveći dio analiza kao alat se može koristiti Excell tabela sa odgovarajućim internim i/ili korisničkim funkcijama. Za veće obime podataka, a posebno za neuronske mreže i druge složene metode, potrebni su drugi prikladni alati. Zbog ove činjenice Benfordov zakon je već zauzeo mjesto u nizu programskih paketa koji su raspoloživi na tržištu za potrebe revizora i drugih specijalista, posebno onih koji se bave prevarama jer je za njih najvažnija mogućnost egzaktne provjere na osnovu nezavisnih mjerila koje je moguće uvijek ponoviti i interpretirati.

2 Matematička osnova Benfordovog zakona

2.1 Uvod

U ovom poglavlju je dat pregled važnijih metoda izvođenja i formulacije Benfordovog zakona. Svaka od metoda odražava specifičan pristup, nastao u praktičnom i teorijskom radu, koji dovodi do zajedničkog zaključka što potvrđuje njegovu izuzetno široku prisutnost.

Za bilo koju bazu $B > 1$ svaki pozitivan broj $x \in \mathbb{R}$ se može jedinstveno zapisati kao $x = M_B(x) \cdot B^k$, gdje je $k \in \mathbb{Z}$ a $M_B(x) \subset [1, B)$. Broj $M_B(x)$ se zove *mantisa* [11].

Definicija 2.1. [2]. Za svako $B > 1$ *mantisa funkcija* je funkcija $M_B : \mathbb{R}^+ \to [1, B)$ takva da je $M_B(x) = x \cdot B^{-[\log_B x]}$ za bilo koje $x \in \mathbb{R}^+$, gdje je $[r]$ označava cijeli dio broja r, najveći cijeli broj koji nije veći od r. Broj $M_B(x)$ se zove *mantisa* za x.

Najopštiji oblik Benfordovog zakona se može napisati u obliku [2]

$$P\{M_B(x) \leq t\} = \log_B(t), \qquad \forall x \in \mathbb{R}^+$$

Korištenjem ove jednakosti lako se izvodi poznati oblik Benfordovog zakona za prvu cifru

$$P\{D_1 = d\} = \log_B\left(1 + \frac{1}{d}\right) \tag{2.1}$$

kao i zakon za bilo koju poziciju n :

$$P\left\{D_n^{(B)} = d_n\right\} = \sum_{k=B^{n-2}}^{B^{(n-1)}-1} \log_B\left(1 + \frac{1}{k \cdot B + d_n}\right)$$

Na tabeli 2.1 su vjerovatnoće proračunate ovom formulom za prvu, drugu, treću i četvrtu cifru.

Na grafikonu P.2.1. u prilogu dat je grafički prikaz ovih frekvencija. Može se primijetiti da su vjerovatnoće za cifru 4 veoma bliske u gotovo svim slučajevima. Zbog gotovo uniformne distribucije za četvrtu i sve naredne cifre, u praktičnim primjenama je uobičajeno da se analize vrše na najviše tri prve pozicije.

	Pozicije			
Cifre	1	2	3	4
0		0,11968	0,10178	0,10018
1	0,30103	0,11389	0,10138	0,10014
2	0,17609	0,10882	0,10097	0,10010
3	0,12494	0,10433	0,10057	0,10006
4	0,09691	0,10031	0,10018	0,10002
5	0,07918	0,09668	0,09979	0,09998
6	0,06695	0,09337	0,09940	0,09994
7	0,05799	0,09035	0,09902	0,09990
8	0,05115	0,08757	0,09864	0,09986
9	0,04576	0,08500	0,09827	0,09982

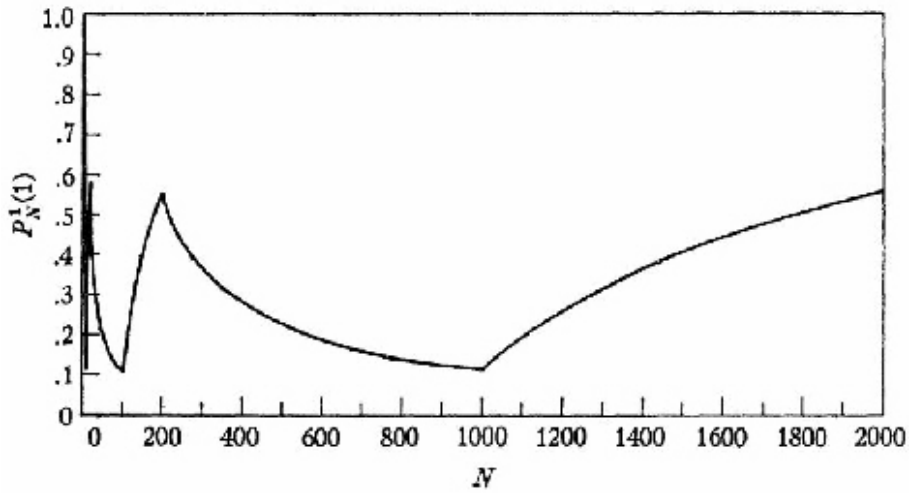
Table 1: Tabela 2.1. Teorijske frekvencije prema Benfordovom zakonu za cife na prve cetiri pozicije

2.2 Klasični pristupi izvođenju Benfordovog zakona

Flehinger. Flehinger [15] je analizirala distribuciju prve značajne cifre slučajno odabranog broja u $\mathbb{N}\setminus\{0\}$. Tražen je odgovor na heurističko pitanje : "Koji dio cijelih pozitivnih brojeva ima početnu cifru manju ili jednaku A ?" ili "Kolika je vjerovatnoća da slučajno odabran cijeli broj ima početnu cifru koja je manja ili jednaka A ?". Prema Benfordovom zakonu, ovo se dešava sa vjerovatnoćom

$$P_{Benf}\left\{[1, i+1)\right\} = \log_{10}(1+i)$$

Kao prvi korak, razmatraju se cijeli brojevi koji su manji ili jednaki konačnom broju N i neka je $P_N^1(A)$ dio tog podskupa sastavljen od brojeva čija je prva cifra manja ili jednaka A. Ako postoji $\lim_{N\to\infty} P_N^1(A)$ on će biti zadovoljavajući odgovor na postavljeno pitanje. Međutim, kako N raste $P_N^1(A)$ oscilira između $A/9$ (za $N = 10^j$, cijelo j) i približno $10 \cdot A/(9(A+1))$ (za $N = (A+1)\cdot 10^j$). Varijacija za $P_N^1(A)$ je data na slici 2.1. Stoga je $P_N^1(A)$ sporo divergentan niz koji oscilira u sve dužim intervalima. Odgovor se traži u generaliziranom ili Banachovom limesu ovog niza.



Slika 2.1. Prikaz spore divergencije niza $P_N^1(A)$ (izvor : Bret Flehinger, On the Probability That a Random Integer Has Initial Digit A, American Mathematical Monthly, 73:1056-1061, 1966.)

Ako se pokuša naći Cesaro limit ovog niza tj. forma kumulativnih prosjeka

$$P_N^2(A) = \frac{1}{N}\sum_{M=1}^{N} P_M^1(A)$$

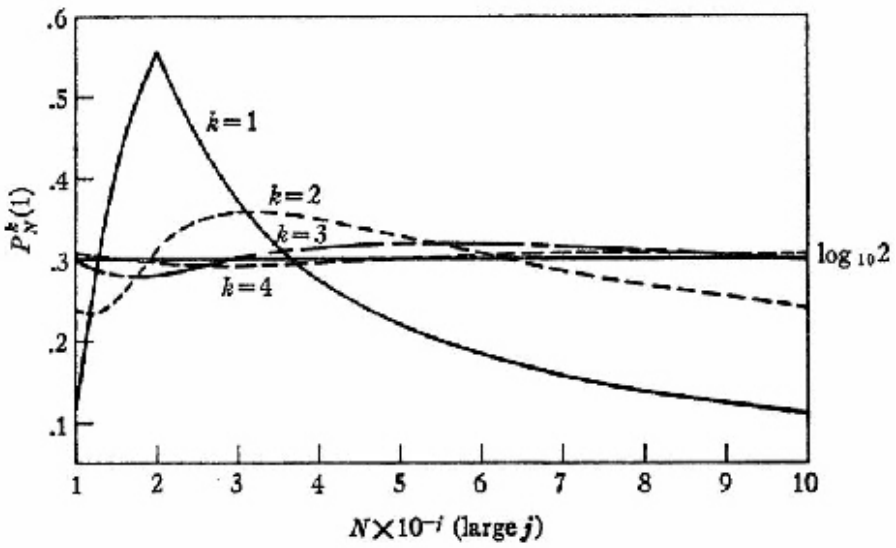
ponovo se generiše divergentni niz ali koji oscilira između užih granica. Ako se naprave sukcesivni kumulativni prosjeci (Hoelderove sume)

$$P_N^k(A) = \frac{1}{N}\sum_{M=1}^{N} P_M^{k-1}(A)$$

generiše se niz za sve konačne k. Kako k raste limiti unutar kojih nizovi osciliraju za veliko N postaju sve bliži. Ovo ponašanje je ilustrovano na slici 2.2. U nastavku teksta dokazuje se da vrijedi

$$\lim_{k\to\infty}\lim_{N\to\infty}\inf P_N^k(A) = \lim_{k\to\infty}\lim_{N\to\infty}\sup P_N^k(A) = \log_{10}(A+1)$$

Drugim riječima, pronađen je regularni granični proces koji vodi ka probabilističkoj mjeri na skupu cijelih brojeva sa inicijalnim ciframa koje su manje ili jednake A. Ova mjera je saglasna sa logaritamskim zakonom.



Slika 2.2. Prikaz brzine divergencije izraza $P_N^k(A)$ za različite vrijednost k (izvor : Bret Flehinger, On the Probability That a Random Integer Has Initial Digit A, American Mathematical Monthly, 73:1056-1061, 1966.)

Flehinger je predložila iterativnu proceduru. Za $k \geq 1$ predlaže

$$P_n^k(i) = \frac{1}{n}\sum_{m=1}^{n} P_n^{k-1}(i)$$

Korištenjem Cesaro prosjeka Flehinger je dokazala da amplituda oscilacija funkcija $P_n^k(i)$ opada kako ovaj proces konvergira ka Benfordovom zakonu na sljedeći način.

Teorem 2.1. (Flehinger)

$$\lim_k \liminf_n P_n^k(i) = \lim_k \limsup_n P_n^k(i) = \log_{10}(1+i)$$

Pokazuje se da $P_n^k(i)$ odgovara procesu od k koraka formiranja lanca Markova. Svojstvo Markova je jedno od ključnih svojstava za metodu reinforcement učenja.

U svom tekstu [15] Flehinger je koristila termin 'logaritamski zakon'. Može se pokazati da je Flehinger metod jači, u smislu da je ekvivalentan ili primjenljiv na mnogo više serija, od bilo kojih drugih iteracija putem Cesaro metoda i da je ekvivalentan bilo kojoj matričnoj metodi kadgod se koristi [2]. Postoji i neprekidna verzija ovog pristupa sa šemama integracije ili Furijeovom analizom. Jedan od zanimljivijih rezultata je tzv. Stiglerov zakon prema kojem se vjerovatnoće vodećih cifara razlikuju od onih po Benfordovom zakonu ali je tendencija ista.

Pristup Miller i Takloo-Bighash. Miller i Takloo-Bighash u formulaciji Benfordovog zakona polaze od niza brojeva [14]. Niz pozitivnih brojeva $\{x_n\}$ je Benfordovski (baza B) ako je vjerovatnoća da će prva cifra za x_n po bazi B biti d jednaka $\log_B(1+1/d)$. Preciznije

$$\lim_{N\to\infty} \frac{\#\{n \le N : D_1(x_n|B) = d\}}{N} = \log_B\left(1+\frac{1}{d}\right)$$

Pritom je $d \in \{1,2,...,B-1\}$ a $\#$ je oznaka za broj uočenih slučajeva. Ovo je distribucija vjerovatnoća jer se mora desiti jedan od $B-1$ događaja a ukupna vjerovatnoća je

$$\sum_{d=1}^{B-1} \log_B\left(1+\tfrac{1}{d}\right) = \log_B \prod_{d=1}^{B-1}\left(1+\tfrac{1}{d}\right) = \log_B \prod_{d=1}^{B-1} \frac{d+1}{d} = \log_B B = 1$$

2.3 Statistička formulacija Benfordovog zakona

Ako se sa D_1, D_2, ... označe funkcije značajnih cifara (npr. $D_1 = (0.314) = 3$, $D_2 = (0.314) = 1$, $D_3 = (0.314) = 4$) opšti zakon ima formu:

$$P\{D_1 = d_1, ...D_k = d_k\} = \log_{10}\left[1 + \left(\sum_{i=1}^{k} d_i \cdot 10^{k-i}\right)^{-1}\right] \qquad (2.2)$$

Uobičajeni pristup formulaciji Benfordovog zakona počinje sa 'ako se izabere slučajan broj vjerovatnoća da njegova prva značajna cifra bude d je $\log_{10}(1+1/d)$' [2]. Glavni nedostatak ove fomulacija je u činjenici da ne postoji prirodan metod da 'se broj slučajno izabere' iz skupa svih pozitivnih (realnih ili prirodnih) brojeva. Obično se počinje sa tim da se (2.2) potvrdi za pozitivne brojeve iz $\mathbb{N}$ počev od prototipskog skupa

$$\{D_1 = 1\} = \{1, 10, 11, 12, 13, 14, ..., 19, 100, 101, ...\}$$

tj. za skup pozitivnih brojeva koji počinju cifrom 1. U ovom pristupu problem je u tome da $\{D_1 = 1\}$ nema prirodnu gustinu u skupu cijelih brojeva tj.

$$\lim_{n\to\infty} \frac{1}{n} \{D_1 = 1\} \cap \{1, 2, ..., n\}$$

ne postoji, za razliku od skupa parnih ili prostih brojeva koji imaju prirodne gustine $1/2$ i 0 respektivno. Skup brojeva koji počinju cifrom d nema prirodnu gustinu u okviru bilo cijelih bilo realnih brojeva, za razliku od npr. parnih ili neparnih brojeva [16]. Nijedan od (brojnih) pristupa nije rezultirao definicijom (prebrojivo aditivne) vjerovatnoće, što je problem isti kao temeljni problem 'slučajnog izbora cijelog broja'. U tom smislu, radovi Newcomba i Benforda imaju glavni nedostatak u smislu da nemaju prikladnu definiciju prostora vjerovatnoća.

Zahtjev da se zakon značajnih cifara stavi u adekvatan prebrojivo aditivan prostor vjerovatnoće je, u suštini, dosta lagan. S obzirom da je zaključak zakona (2.2) jednostavno tvrdnja o funkcijama značajnih cifara (slučajnih varijabli) D_1, D_2,... neka je uzorački prostor $\mathbb{R}^+$, skup pozitivnih realnih brojeva, i neka je σ-algebra događaja jednostavno σ-polje generisano sa $\{D_1, D_2...\}$ (ili, što je ekvivalentno, generisano funkcijom $x \mapsto mantisa(x)$).

Definicija 2.2. [2]. (*sigma-algebra po bazi* B). Za svaki $E \in \mathcal{B}([1, B))$ neka je $\mathcal{M}_B(E) = \bigcup_{n\in\mathbb{Z}} B^n E$ i neka je $\mathcal{M}_B = \{\mathcal{M}_B(E) : E \in \mathcal{B}([1, B))\}$. Algabera $\mathcal{M}_B$ će se zvati sigma algebra po bazi B.

U ovoj definiciji $\mathcal{B}(A)$ označava Borelove skupove na A. Lako se pokazuje da ova σ-algebra σ-potpolje Borela i da u slučaju baze $B = 10$ vrijedi:

$$S \in \mathcal{M} \Leftrightarrow S = \bigcup_{n=-\infty}^{\infty} E \cdot 10^n, \qquad Borelov \;\; E \subseteq [1, 10) \qquad (2.3)$$

Iako veoma jednostavna, σ-algebra $\mathcal{M}$ ima neka interesantna svojstva.

1. Svaki neprazan skup u $\mathcal{M}$ je beskonačan sa tačkama nagomilavanja u tačkama 0 i $+\infty$;
2. $\mathcal{M}$ je zatvoren u odnosu na skalarno množenje ($s > 0, S \in \mathcal{M} \Rightarrow sS \in \mathcal{M}$);

3. $\mathcal{M}$ je zatvoren u odnosu na cjelobrojne korijene ($m \in \mathbb{N}, S \in \mathcal{M} \Rightarrow S^{1/m} \in \mathcal{M}$) ali ne i u odnosu na stepene;

4. $\mathcal{M}$ je sličan sam sebi u smislu da ako je $S \in \mathcal{M} \Rightarrow 10^m \cdot S \in \mathcal{M}$ za svaki cijeli m (gdje aS i S^a označava skupove $\{as : s \in S\}$ i $\{s^a : s \in S\}$ respektivno)

Svojstvo (1) znači da konačni intervali kao što je $[1, 2)$ nisu u $\mathcal{M}$ (tj. ne mogu se izraziti putem samo jedne značajne cifre; značajne cifre same po sebi ne mogu praviti razliku npr. između brojeva 2 i 20) i stoga nestaju kontradikcije povezane sa svojstvom mjerne invarijantnosti. Svojstva (1), (2) i (4) slijede lako iz (2.3) ali (3) zaslužuje bliže ispitivanje. Kvadratni korijen skupa u $\mathcal{M}$ se može sastojati od dva dijela a slično vrijedi i za korijene višeg reda. Naprimjer,

$$S = \{D_1 = 1\} = \bigcup_{n=-\infty}^{\infty} [1, 2) \cdot 10^n$$

$$S^{1/2} = \bigcup_{n=-\infty}^{\infty} \left[1, \sqrt{2}\right) \cdot 10^n \cup \bigcup_{n=-\infty}^{\infty} \left[\sqrt{10}, \sqrt{20}\right) \cdot 10^n \in \mathcal{M}$$

$$S^2 = \bigcup_{n=-\infty}^{\infty} [1, 4) \cdot 10^{2n} \notin \mathcal{M}$$

s obzirom da su razlike (gaps) u tom slučaju prevelike i ne mogu biti napisane pomoću $\{D_1, D_2, ...\}$. Svojstvo (2) je ključ hipoteze o mjernoj invarijantnost a svojstvo (4) je ključ hipoteze o baznoj invarijantnosti.

2.4 Entropijski princip formulisanja Benfordovog zakona

U tekstu [21] je dat entropijski pristup definisanju Benfordovog zakona. Posmatra se model kutija i loptica na sljedeći način :
A) Cifra n je ekvivalentna 'kutiji' koja sadrži n loptica koje nisu u interakciji
B) N sekvenca ovih 'kutija' je ekvivalentna broju ili numeričkom fajlu
C) Sve moguće kombinacije kutija i loptica, za dati broj loptica, imaju jednaku vjerovatnoću

Posljednja pretpostavka je definicija ekvilibrija i slučajnosti u statističkoj fizici. U informacionoj teoriji ovo znači da nad fajlom postoji Shannonov limit. Broj je napisan kao kombinacija cifara u bazi B. Broj sa N cifara u bazi B se može opisati kao skup od N kutija od kojih svaka sadrži n kuglica, gdje n može biti bilo koji cijeli broj od 1 do $B-1$. Ukupan broj loptica označava se sa P. Nepristrasna distibucija loptica u kutijama znači jednaku vjerovatnoću da svaka loptica bude u bilo kojoj kutiji. Pokazuje se da je ova pretpostavka ekvivalentna pretpostavci C) i povlači Benfordov zakon.

Naprimjer, ako se uzme baza $B = 4$ tada su moguće četiri cifre $(0, 1, 2, 3)$. Naveća vrijednost trocifrenog broja u ovoj bazi je 333, što znači da je u kutijama 9 loptica. Pritom je moguća samo jedna ovakva kombinacija. U slučaju da se raspoređuje 3 loptice u 3 kutije postoji više kombinacija $(300, 030, 003, 210, 201, 120, 102, 012, 021$ i $111)$. Cifra

1 se javlja 9 puta, cifra 2 se javlja 6 puta, cifra 3 se javlja 3 puta. Pritom je nevažno što je odnos broja cifara $9:6:3$, $\rho(1)=0.5$, $\rho(2)=0.33\dot{3}$, $\rho(3)=0.166\dot{6}$, nezavisan od N što je i razlog zašto 0 nije uključena u Benfordov zakon. Kako se vidi u primjeru, svaka kutija ima jednaku vjerovatnoću da u njoj bude 1, 2 ili 3 kuglice kao i u drugim kutijama. Međutim, vjerovatnoća da je u kutiji 3 kuglice manja je od vjerovatnoće da u njoj budu 2 kuglice a najveća je vjerovatnoća da u kutiji bude jedna kuglica. Razlog za ovo je što se vjerovatnoća da kutija ima n loptica smanjuje kako se n povećava, manje cifre imaju veću vjerovatnoću. Formalna kalkulacija distribucije loptica u kutijama urađena je računanjem svih mogućih deset konfiguracija (u opštem slučaju $\frac{(N+P-1)!}{(N-1)!\cdot P!}$), pri čemu im se daje jednaka vjetovatnoća, a zatim se broje frekvencije cifara, bez obzira na lokaciju.

Distribucija P loptica u N kutija je ekvilibrij u klasičnom problemu termodinamike. U statističkoj mehanici ekvilibrij se definiše kao statističko stanje u kojem sve moguće konfiguracije imaju jednaku vjerovatnoću. Ekvilibrij funkcije distibucije $\rho(n)$ (količnik kutija u kojima je n loptica) računa se na način da se traži maksimalna entropija, što znači jednaku vjerovatnoću za sve konfiguracije (mikrostanja).

2.5 Generalizacije Benfordovog zakona

Benfordova generalizacija. Jedan oblik generalizacije je formulisao Benford [19]. Njemu je bilo logično da su relativne frekvencije ili vjerovatnoće P za prve značajne cifre drugačije ako ne postoje druge (naredne) cifre nego u slučaju da se one jednostavno zanemare (truncate) u kalkulacijama frekvencija. Benford je ovaj rezultat zvao Opšta jednadžba zakona anomalnih brojeva a dat je izrazima

$$P_1^r = \frac{1}{N}\left[\log_e \frac{10\cdot(2\cdot 10^{r-1}-1)}{10^r-1}+\frac{8}{10^r}\right]$$
$$P^r{}_{\substack{a\\ a\neq 1}} = \frac{1}{N}\left[\log_e \frac{(a+1)\cdot 10^{r-1}-1}{a\cdot 10^r-1}-\frac{1}{10^r}\right]$$

Ovdje je r dozvoljeni broj cifara. Jednadžba se aproksimira sa $P(dd)=\log_{10}(1+1/dd)$ za veći poredak od r što omogućava korištenje jednostavnije formule u tipičnim slučajevima. Za niže vrijednosti za r koji predstavljaju brojeve koji su zaokruženi (rounded) ili u kojima je ostatak zanemaren (truncated) i svedeni na jednu cifru Benford je proračunao već poznate teorijske frekvencije.

Stiglerov pristup. Benford je sugerisao da zakon važi kada podaci dolaze iz mješavine uniformnih distribucija za koje je izvjesnije da imaju relativno male gornje granice [65]. Raimi (1976) je postulirao da je Benfordova šema mješavine uniformnih distribucija proizvoljna i aproksimativna jer implicira da razni drugi zakoni takođe mogu biti kreirani miješanjem raznih distribucija što nekoga može navesti da se pita zašto bi ova mješavina bila relevantna kako bi opisala distribuciju prvih značajnih cifara. George Stigler, dobitnik Nobelove nagrade za ekonomiju (1982), je tvrdio da je specifična mješavina uniformnih distribucija sa neuniformno distribuiranim maksimalnim vrijednostima, u najmanju ruku, nekonzistentna. Ova opservacija je natjerala Stiglera (1945) da predloži

alternativnu distribuciju prvih značajnih cifara koja je bila manje asimetrična za manje cifre i bila je izvedena bez korištenja ovakvih pretpostavki.

Stigler je (1945) provjerio Newcomb-Benfordov fenomen prve značajne cifre i predložio da prosječna relativna frekvencija vodeće značajne cifre d bude :

$$F_d = \frac{d \ln(d) - (d+1)\ln(d+1) + \left(1 + \frac{10}{9}\ln(10)\right)}{9}$$

Do ovog zaključka je došao najprije pretpostavljajući da je za najveću stavku u statističkoj tabeli podjednako izvjesno da počinje sa $d = 1, 2, ..., 9$ i da su ostale stavke u tabeli slučajno izabrane iz uniformne distribucije brojeva manjih od najveće stavke. Definisanjem r-tog ciklusa brojeva kao intervala $[10^r; 10^{r+1})$ za neki realni broj r Stigler je našao distribuciju prvih značajnih cifara za najveće stavke u ciklusu brojeva iz tabele i zatim formiranjem prosjeka vjerovatnoća za sve najveće stavke. S obzirom da su stavke tabele iz uniformne distribucije, bilo koja cifra d bi se na kraju $(r-1)$-og ciklusa, koji se ponovio $(10^r - 1)/9$ puta kao prva značajna cifra od $10^r - 1$ brojeva, trebala pojavljivati približno $10^r/9$ i 10^r respektivno. Naprimjer, na kraju prvog ciklusa tj. $[10, 100)$ cifra 2 se kao prva značajna cifra pojavljuje $(10^2 - 1)/9 = 11$ puta od $10^2 - 1 = 99$ brojeva, uključujući i one iz svih prethodnih ciklusa. Nakon $(r-1)$-og ciklusa d se ne pojavljuje kao prva cifra u sljedećih $(d-1) \cdot 10^r$ brojeva tj. cifra 2 se ne javlja kao prva značajna cifra u intervalu $[10^2, 10^2 + (2-1) \cdot 10^2) = [100, 200)$. Slijedeći tu logiku vidi se da vrijedi

$$p_i = \frac{d_i \cdot \ln d_i - (d_i + 1) \cdot \ln(d_i + 1) + m}{9}$$

gdje je m definisano sa

$$m = \frac{\sum_{i=1}^{9} \left(i^2 \ln d_i - d_i (d_i + 1) \ln(d_i + 1)\right)}{9 - \sum_{i=1}^{9} d_i}$$

Frekvencije dobijene Stiglerovom formulom prezentirane su u tabeli 2.2. Frekvencije iz Benfordovog zakona su date radi poređenja.

Iako se relativne frekvencije razlikuju, skupovi frekvencija su slični po obrascu monotonog opadanja. S obzirom da logaritamska distribucija prvih značajnih cifara ne vrijedi generalno za sve skupove podataka Stiglerov i Benfordov zakon mogu biti smatrani kao članovi porodice monotono opadajućih distribucija prvih značajnih cifara. Stigler tvrdi da razlika između njegove alternative i Benfordovog zakona potiče od skrivenih pretpostavki koje je napravio Benford o relativnim frekvencijama najvećih cifara u statističkim tabelama. Benford je pretpostavio da se mali brojevi sa odgovarajućim malim prvim značajnim ciframa pojavljuju češće kao granice za statističke tabele. Posebno, za datu mješavinu

PZC	Stiglerov zakon	Benfordov zakon
1	0.241	0.301
2	0.183	0.176
3	0.146	0.125
4	0.117	0.097
5	0.095	0.079
6	0.077	0.067
7	0.061	0.058
8	0.047	0.051
9	0.034	0.046

Table 2: Tabela 2.2 Uporedni pregled Benfordove i Stiglerove distribucije (izvor : Joanne Lee, Wendy K. Tam Cho, George G. Judge, Stigler's approach to recovering the distribution of first significant digits in natural data sets, Statistics and Probability Letters 80 (2010) 82 88, journal homepage: www.elsevier.com/locate/stapro)

uniformnih distribucija $U[0, b)$ pretpostavlja se da je gornja granica b proporcionalna sa $1/b$. Stigler smatra da je ta pretpostavka nepotrebna u izvođenju logaritamskog pravila s obzirom da niti je proširila obim zakona niti je doprinijela teorijskoj osnovi modeliranja distribucije prvih značajnih cifara. Za razliku od toga, Stiglerova pretpostavka je da je za najveće stavke u tabeli jednako izvjesno da počinju sa $d = 1, 2, ..., 9$ (Stigler, 1945).

Ovaj pristup omogućava da se riješe brojne dileme o prirodi veze između Benfordovog i Zipfovog zakona prema kojima se, u nekim slučajevima, izvodi pogrešan zaključak da je Benfordov zakon specijalni slučaj Zipfovog zakona.

2.6 Test drugog reda

Test drugog reda je analiza frekvencije cifara razlika između sortiranih (rangiranih) vrijednosti u skupu podataka [13]. Frekvencije cifara razlika aproksimiraju frekvencije Benfordovog zakona za većinu distribucija originalnih podataka. Testovi drugog reda generišu manje lažnih pozitivnih incidencija i mogu detektovati zaokruživanja podataka, podatke generisane linearnom regresijom, podatke generisane inverznim funkcijama poznatih distribucija i netačno sortiranje. Ovi uslovi ne bi mogli biti detektovani korištenjem tradicionalnih analitičkih procedura.

Test drugog reda je motivisan sa dva razloga. Prvo, već od poznatog Benfordovog teksta mnogi istraživači su pokazali da objedinjavanje podataka iz više izvora vodi ka Benfordovskom ponašanju. Drugo, mnoge standardne distribucije vjerovatnoća bliske su Benfordovskom ponašanju. Istražuje se distribucija cifara za razlike dvije susjedne vrijednosti jedne varijable složene rastućim redom. Za bilo koje $\delta < 1$ analizira se najviše N^{δ} uzastopnih razlika u skupu obima N. Rezultujuća distribucija vodećih cifara veoma slabo zavisi od distribucije unutar podataka i bliska je Benfordovom zakonu. Važno je pitanje da li sve razlike vode ka Benfordovom zakonu. Inspirisano je navedenim zapažanjem i dovodi do novih testova integriteta podataka, koji su lagani za primjenu i uspješno detektuju probleme u nekim skupovima podataka, što je bitno sa stanovišta praktične primjene.

Test drugog reda je motivisan željom da se detektuju anomalije u razlikama između elemenata skupa. Frekvencija prvih cifara razlika koja značajnije odstupa od predviđenog modela može ukazivati na neregularnosti / anomalije (npr. pokušaj prevaranta da lažira podatke na način da unosi veličine u pravilnim razmacima) ali i na određenu karakteristiku posmatranog sistema za koju se može pokazati kao veoma bitna.

2.7 Modeliranje skupova saglasnih sa Benfordovim zakonom

Jedno od čestih praktičnih potreba i pitanja je generisanje skupova brojeva koji su saglasni sa Benfordovim zakonom. Drugim riječima, potreban je odgovor na pitanje : da li je moguće nekom funkcijom ili kombinacijom više funkcija generisati skup traženog obima koji je, uz prihvatljiv nivo značajnosti, saglasan sa Benfordovim zakonom ? Odgovor je potvrdan a u nastavku je dat pregled nekih metoda.

Metoda Nigrinija. Nigrini i Miller [23] predlažu korištenje geometrijskih nizova za vještačko generisanje Benfordovih skupova i daju im naziv Sintetički Benfordovi nizovi. Takva sekvenca $(S_n)_{n\in\mathbb{N}}$ na realnom intervalu $[a, b]$ sa N elemenata generiše se izrazom $S_n = a \cdot r^{n-1}$ gdje je

$$r = 10^{(\log b - \log a)/(N-1)} = 10^{\frac{1}{N-1}\log\left(\frac{b}{a}\right)}$$

Gornja granica b dostiže se jedino za dovoljno veliko N. Frekvencije prvih cifara u geometrijskom nizu će slijediti Benfordov zakon ako su zadovoljena dva uslova. Prvi uslov je da obim uzorka, N, bude dovoljno velik. Zahtjev da obim uzorka bude 'dovoljno velik' može biti pomalo nejasan zbog toga što čak i savršeni geometrijski niz ne može perfektno slijediti Benfordov zakon. Naprimjer, očekivane proporcije da prve dvije cifre budu od 90 do 99 su u opsegu od 0.0044 do 0.0048. S obzirom da broj pojavljivanja mora biti cijeli broj to znači da će se stvarne frekvencije (bilo 4 ili 5) u uzorku obima npr. 10.000 prenijeti u proporcije od 0.004 ili 0.005. Sa porastom broja članova geometrijskog niza raste mogućnost da stvarne proporcije budu bliže onima koje se očekuju u skladu sa Benfordovim zakonom. Drugi uslov je da razlika $\log b - \log a$ bude cijeli pozitivan broj. Geometrijski niz mora obuhvatiti cijeli broj redova veličina odnosno stepena baze, kako bi se mogle pojaviti sve cifre sa očekivanim frekvencijama.

Ovaj pristup je modeliran u Excell-u za $N = 10.000$. Nakon generisanja podaci su sortirani u rastućem redoslijedu. Grafikon upućuje na podatke u kojima ima više malih nego velikih brojeva. Oko 80% elemenata uzorka je manje od 10.000 odnosno oko 10% raspona svih vrijednosti iz uzorka. Ovo je u skladu sa jednim od zahtjeva za Benfordovske skupove. Mjera asimetrije je oko $2,75$, što je dosta skromno.

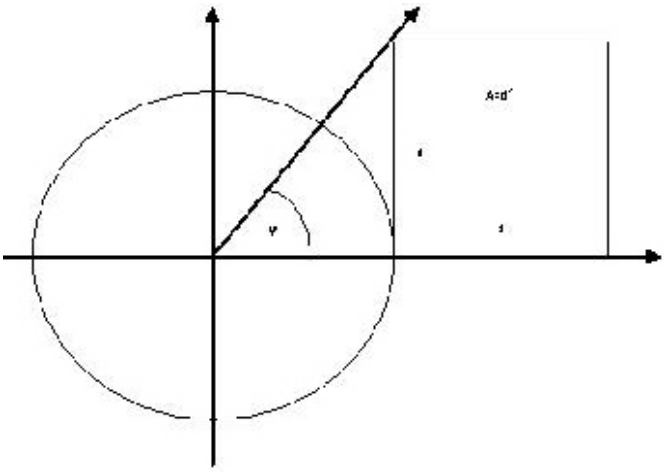
Eksponencijalna funkcija. Kako bi se generisali skupovi saglasni sa Benfordovim zakonom dovoljno je uzimati brojeve uniformno sa logaritamske skale [10]. Naredna teorema pokazuje da su takvi podaci saglasni sa Benfordovim zakonom.

Teorema 2.7. Slučajna varijabla $10^{\mathrm{rand}(k)}$ slijedi Benfordov zakon pod pretpostavkom da je k pozitivan cijeli broj i funkcija rand (k) vraća slučajne brojeve iz intervala $[0, k)$.

U Excell-u je ovaj pristup modeliran sa parametrom $k = 1$. Korištena je funkcija rand (x) u paketu Excell kojom su generisani slučajni brojevi iz intervala $(0, 1)$ a stepenovanjem se dobijaju brojevi iz intervala $(1, 10)$. Nakon generisanja uzorka obima $N = 10.000$ podaci su sortirani u rastućem redoslijedu.

Ako se y osa prikaže na logaritamskoj skali dobija se prava linija. Koeficijent asimetrije za slučajeve koji su analizirani po ovom metodu je bio oko $0,75$, što je dosta nisko. Uzrok je u činjenici da se zbog brzog rasta eksponencijalne funkcije ne dobija dovoljan broj malih vrijednosti, što je karakteristika nužna za Benfordovske skupove.

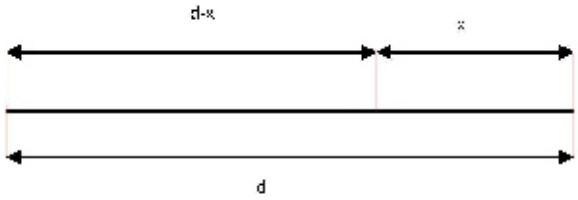
Geometrijska metoda. El-Muhsi teorema daje dva načina izvođenja Benfordovog skupa geometrijskim konstrukcijama [24]. Prema prvom metodu, u krugu jediničnog poluprečnika na slučajan način se bira ugao $\varphi \in (0, 2k\pi)$ i računa $h = \tan(\varphi)$. Kvadrat površine $A = h^2 = (\tan(\varphi))^2$ je broj koji se uzima kao element skupa. Ilustracija je data na slici 2.3.



Slika 2.3. Geometrijska metoda generisanja Benforodvskih skupova putem funkcije $y = (\tan(\varphi))^2$
(izvor : http://reocities.com/CapeCanaveral/hangar/4577/elmuhsi.htm)

Parametar k označava broj punih krugova odnosno numerički interval za zadati broj krugova. Ako se umjesto $r = 1$ uzme neka druga vrijednost veličina A se računa kao $A = (r \cdot \tan(\varphi))^2$. Umjesto kvadrata može se koristiti bilo koja druga figura upisana u kvadrat ili trodimenzionalni ekvivalent (kocka) odnosno tijelo upisano u kocku. U oba slučaja rezultujući skup će, zbog osobine mjerne invarijantnosti, slijediti Benfordov zakon.

Drugi metod se sastoji u tome da se uzme duž fiksne dužine d i bira slučajan broj $x \in (0, d)$. Kao generator Benfordovog skupa uzima se funkcija $r = \dfrac{d}{x} - 1$ odnosno $r = \dfrac{d - x}{x}$.



Slika 2.4. Geometrijska metoda generisanja Benfordovskih skupova korištenjem funkcije $y = \frac{d}{x} - 1$
(izvor : http://reocities.com/CapeCanaveral/hangar/4577/elmuhsi.htm)

Vrijednost d se može odnositi na interval dužine d tj. na bilo koji interval $[a, a+d]$. Metoda očigledno podrazumijeva poznat opseg d u koji padaju generisani podaci. Ovo, na izvjestan način, može biti i nedostatak s obzirom na jedan od osnovnih zahtjeva da skup ne smije imati nametnutih ograničenja.

Na istom izvoru (izvor : http://reocities.com/CapeCanaveral/hangar/4577/elmuhsi.htm) su dostupni testni podaci odnosno makroi u Excell-u za provjeru oba metoda na skupovima od 10.000 i 60.000 elemenata. Za potrebe ovog teksta napravljene su probe obje navedene metode, na način da generišu grafikone za prvu, drugu i posljednju cifru.

Prva metoda je simulirana sa parametrima $r = 1$, $k = 3$ što znači da je $\varphi \in [0, 18.85]$. Uzorak obima $N = 10.000$ je, nakon generisanja, sortiran u rastućem redoslijedu. Grafikon koji se dobije ovom metodom karakterističan je za podatke koji slijede Benfordov zakon; najveći dio uzorka ima vrijednosti u dijelu grafikona koji ima bitno sporiji rast u odnosu na dijelove u kojima su ekstremne vijednosti. Ovo osigurava da se najveći dio članova skupa nalazi u opsegu od bar dva reda veličina. Druga karakteristika je da je broj ekstremno malih ili velikih brojeva bitno manji u odnosu na ostale podatke. Na ovaj način se postiže visok stepen asimetrije, koji je u ovom slučaju dosta visok i iznosi $99,86$.

Druga metoda je modelirana sa parametrom $d = 100.000$. Uzorak obima $N = 10.000$ je, nakon generisanja, sortiran u rastućem redoslijedu. Mjera asimetrije za ove parametre je $50,20$ što je dosta visoko.

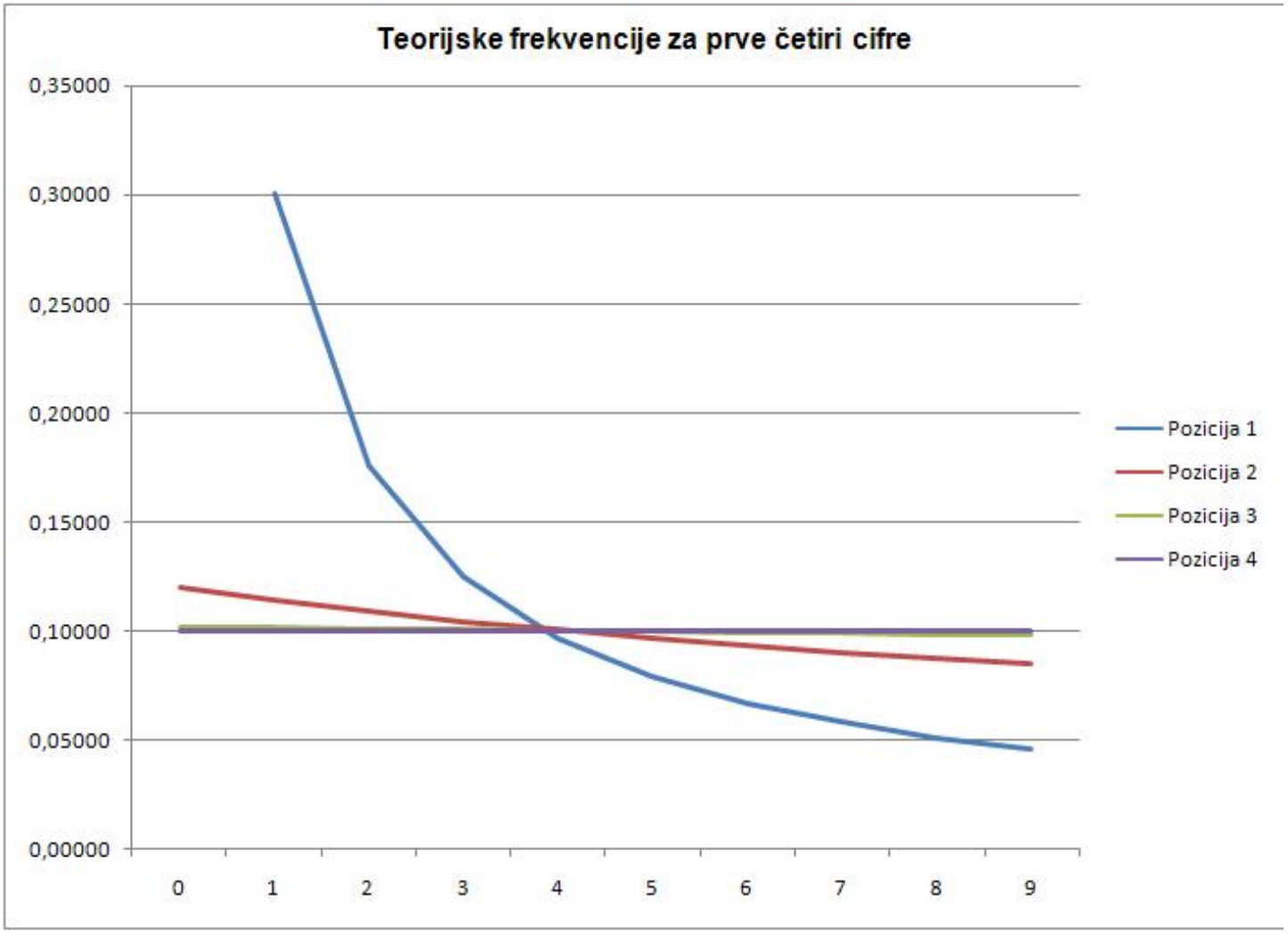
Na osnovu izloženog se može reći da metode generisanja po El-Muhsi teoremi daju uzorke bolje karakteristike asimetrije a samim tim i skupove koji imaju bolju saglasnost sa Benfordovim zakonom.

2.8 Zaključak

Benfordov zakon je izveden korištenjem brojnih i različitih metoda. Svaki od njih odražava pristup na osnovu specifičnih tipova teorijskih i praktičnih problema. Ovo ujedno odražava i veliku prisutnost ovog zakona u gotovo svim segmentima teorijskog i praktičnog rada.

Metode generisanja podataka koji su saglasni sa Benfordovim zakonom predstavljaju dragocjen alat prilikom provođenja testova i simulacija.

2.9 Prilog



Grafikon P.2.1. Teorijske frekvencije cifara na prve četiri pozicije. Grafikon urađen na osnovu tabele 2.1. Potrebno je primijetiti da su teorijske frekvencije za cifru 4 gotovo jednake u svim slučajevima

3 Testiranje Benfordovog zakona

Testiranje je postupak poređenja teorijskih i uzoračkih veličina kako bi se utvrdio stepen saglasnosti uzorka sa traženim zakonom odnosno pravilom. U velikom broju slučajeva, testiranje saglasnosti sa Benfordovim zakonom ima naglašen kontekst detekcije anomalija što se u praksi najčešće dovodi u vezu sa detekcijama prevara [25]. Bez obzira na to, prilikom izbora metoda testiranja bitni su odgovori na dva pitanja.

Prvo je pitanje prikladnosti testova. Neki postojeći testovi su suviše konzervativni pa se moraju izvesti kritične vrijednosti koje testovima daju veću snagu i evaluiraju vrijednosti za male uzorke. Često se koriste mjere saglasnosti (measures of fit) kao 'pravilo palca' kako bi se provjerila saglasnost sa Benfordovim zakonom. U nastavku je data interpretacija takvih mjera i kritične vrijednosti za testiranje hipoteza. Drugo pitanje je primjena testova na podacima koji inherentno ne zadovoljavaju Benfordov zakon. Odbacivanje testova na osnovu Benfordovog zakona na tim podacima neće pomoći u otkrivanju prevare ili greške.

U provođenju ovih testova se polazi od nulte hipoteze :

H_0 : Značajne cifre odabranog skupa imaju distribuciju u skladu sa Benfordovim zakonom

Testiranja se provode računanjem uzoračkih statistika koje se porede sa kritičnim vrijednostima za odbarani nivo značajnosti. Uobičajeno je da nivo značajnosti testa bude 5%.

3.1 Statistički testovi

U ovom dijelu je dat pregled jednog broja testova koji se koriste za ispitivanje saglasnosti uzorka sa Benfordovim zakonom.

3.1.1 Srednja apsolutna devijacija

Nigrini je (2000) predložio korištenje srednje aposlutne devijacije (MAD - Mean Absolute Deviation) koja se računa na sljedeći način [27] :

$$MAD = \frac{1}{9}\sum_{d=1}^{9} \left| \overline{P}\left\{D_1 = d\right\} - P\left\{D_1 = d\right\} \right|$$

Za slučaj testa prve dvije cifre MAD se računa na sljedeći način :

$$MAD = \frac{1}{90} \sum_{d_1 d_2 = 10}^{99} \left| \overline{P} \left\{ D_1 D_2 = d_1 d_2 \right\} - P \left\{ D_1 D_2 = d_1 d_2 \right\} \right|$$

Umjesto MAD neki autori koriste srednju kvadratnu grešku (MSE) ali to ne rješava najozbiljniji nedostatak ovog testa : ne postoje objektivno utvrđene kritične vrijednosti [2]. Koliko je poznato, trenutno postoje dva, donekle subjektivna, metoda da se utvrde kritične vrijednosti. Prvi i najčešće korišteni metod je predložio Nigrini. Njegove kritične vrijednosti su zasnovane na iskustvu praktičnih testova i stoga se moraju koristiti sa oprezom. Tabela 3.1. ilustruje kritične vrijednosti za tri tipa testova.

Tip testa / odluke	Prva cifra	Druga cifra	Prve dvije cifre
Bliska saglasnost	< 0.004	< 0.008	< 0.006
Prihvatljiva saglasnost	$0.004 - 0.008$	$0.008 - 0.012$	$0.006 - 0.012$
Marginalna saglasnost	$0.008 - 0.012$	$0.012 - 0.016$	$0.012 - 0.018$
Nesaglasnost	> 0.012	> 0.016	> 0.018

Table 3: Tabela 3.1. Kriticne vrijednosti za MA (izvor : Tamas Lolbert, Digital analysis : Theory and applications in auditing, Hungarian Statistical Review, Special number 10)

Na tabeli je moguće uočiti da su kritične vrijednosti za prve dvije cifre bliske prosjeku kritičnih vrijednosti za prvu i drugu cifru. Drugi mogući pristup je da se kritične vrijednosti dobiju Monte-Karlo simulacijom (Posch, 2004). U tom slučaju najprije se generiše Benfordov skup koji se zatim postepeno kontaminira odgovarajućim ne-Benfordovskim slučajnim brojevima. Evaluacija očekivanog MAD u svim fazama daje grubu procjenu u kojoj mjeri originalni podaci smiju biti kontaminirani. Ovaj pristup se mora koristiti sa oprezom s obzirom da rezultati zavise od načina na koji se generišu ne-Benfordovski brojevi. Dodatno, što je veći stepen kontaminacije veća je varijansa simuliranog MAD.

3.1.2 Pearsonov hi-kvadrat test

Pearsonov χ^2 test je, skupa sa testom Kolmogorov-Smirnov (D_N) i Kuiper (V_N), prirodan kandidat za testiranje saglasnosti sa Benfordovim zakonom [32]. U osnovnom obliku, χ^2 statistika se računa na sljedeći način :

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Ovdje su O_i i E_i uzoračke (observed) i očekivane (expected) frekvencije a k broj klasa (grupa) na koje je uzorak podijeljen. Broj klasa u ovom slučaju je broj cifara za koje se pravi analiza (9 za prvu cifru i 10 za drugu i ostale cifre jer se računanje obavlja za pojedine pozicije). Ako se radi sa uzorkom obima N tada je $\overline{P}\left\{D_i = d\right\} = O_i/N$ uzoračka relativna frekvencija cifre na d na poziciji i a $P\left\{D_i = d\right\}$ vjerovatnoća prema Benfordovom zakonu. U skladu sa tim je $E_i = N \cdot P\left\{D_i = d\right\}$ odnosno $P\left\{D_i = d\right\} = E_i/N$ i tada vrijedi [28] :

$$T_1 = N \cdot \sum_{d=1}^{9} \frac{\left(\overline{P}\left\{D_1 = d\right\} - P\left\{D_1 = d\right\}\right)^2}{P\left\{D_1 = d\right\}},$$
$$T_k = N \cdot \sum_{d=0}^{9} \frac{\left(\overline{P}\left\{D_k = d\right\} - P\left\{D_k = d\right\}\right)^2}{P\left\{D_k = d\right\}}, \qquad k = 2, 3, ...$$

Pod nultom hipotezom da je distribucija značajnih cifara u skladu sa Benfordovim zakonom ovo su statistike sa 8 (9) stepena slobode.

Kao kvadratna mjera, ova statistika je osjetljiva na obrazac odstupanja od Benfordovog zakona. Sa fiksiranim nivoom značajnosti test je veoma osjetljiv na povećanje obima uzorka (N). Iz tog razloga nulta hipoteza o saglasnosti distribucije cifara sa Benfordovim zakonom može biti odbačena kako se, povećanjem uzorka, vjerovatnoća greške Tipa II (β) približava nuli.

3.1.3 z-test

Za svaku pojedinačnu cifru može se računati standardna devijacija [30] na sljedeći način:

$$s_i = \sqrt{\frac{P_d\left(1 - P_d\right)}{N}}$$

gdje je P_d očekivana vjerovatnoća pojavljivanja cifre d a N obim uzorka. Da bi se ispitala frekvencija svake pojedine cifre odnosno da li se neka cifra pojavljuje više / manje nego što predviđa Benfordov zakon, može se koristiti z-statistika, koja se računa na sljedeći način [28] :

$$z = \frac{|P_o - P_d| - \frac{1}{2N}}{s_i} = \frac{|P_o - P_d| - \frac{1}{2N}}{\sqrt{\frac{P_d\left(1 - P_d\right)}{N}}} = \sqrt{N}\frac{|P_o - P_d| - \frac{1}{2N}}{\sqrt{P_d\left(1 - P_d\right)}}$$

Ovdje je P_o uzoračka (observed) relativna frekvencija, P_d očekivana (expected) vrijednost relativne frekvencije za cifru d koja se očekuje prema Benfordovom zakonu a N obim uzorka. Član $1/\left(2N\right)$ je korektivni faktor i koristi se samo ako je manji od člana pod znakom apsolutne vrijednosti. U tekstu [28] ova statistika je data u obliku bez korektivnog faktora :

$$T_d = \sqrt{N}\frac{P_o - P_d}{\sqrt{P_d \cdot \left(1 - P_d\right)}} = \frac{P_o - P_d}{\sqrt{\frac{P_d \cdot \left(1 - P_d\right)}{N}}} = \sqrt{N}\frac{P_o - P_d}{\sqrt{P_d \cdot \left(1 - P_d\right)}}$$

Na osnovu ovoga se izvode gornja i donja granica intervala povjerenja za frekvencije pojedine cifre [29] :

$$DG = P_d - \varphi_{\frac{1+\alpha}{2}} \cdot \sqrt{\frac{P_d \cdot (1 - P_d)}{N}} - \frac{1}{2N}$$
$$GG = P_d + \varphi_{\frac{1+\alpha}{2}} \cdot \sqrt{\frac{P_d \cdot (1 - P_d)}{N}} + \frac{1}{2N}$$

U ovim izrazima je $\varphi_{\frac{1+\alpha}{2}}$ kvantil za odabrani nivo povjerenja. Za nivo značajnosti od 5% ova kvantil je 1.96 a 10% nivo povjerenja on je 1.64. Tako, interval povjerenja za najčešće korišteni nivo povjerenja od 5% je

$$\left[P_d - 1.96 \cdot \sqrt{\frac{P_d \cdot (1 - P_d)}{N}} - \frac{1}{2N}, P_d + 1.96 \cdot \sqrt{\frac{P_d \cdot (1 - P_d)}{N}} + \frac{1}{2N}\right]$$

Ako uzoračka frekvencija prelazi donju ili gornju granicu odstupanje se kvalifikuje statistički značajno. z-skor je distanca od uzoračke srednje vrijednosti izražena u jedinicama standardne devijacije.

Posmatrano u kontekstu detekcije prevara odnosno traženja neregularnosti, ovdje se mogu pojaviti dva problema, jedan intuitivni a drugi statistički [30]. Prvo, intuitivno gledano, ako npr. u uzorku postoji samo nekoliko prevarantskih transakcija značajna razlika neće biti detektovana čak i ako je ukupna suma veoma velika. Drugo, statistički gledano, ako je predmet analize veliki broj numeričkih veličina (npr. veliki broj transakcija na jednom računu) to će davati malu proporciju nekonzistentnih brojeva kako bi se detektovala značajna razlika u odnosu na očekivanu nego kada bi račun imao mnogo manje transakcija. Ovo je razlog zašto veliki broj programskih paketa koji uključuju test na Benfordov zakon traži da se testira cjelokupna populacija (npr. ukupan promet na računu) umjesto da se uzima uzorak.

3.1.4 Karakterizacija Benfordovog zakona putem invarijantnosti suma

Važno svojstvo skupova koji slijede Benfordov zakon je invarijantnost suma. Nigrini je u tekstu [31] dao zapažanje :

Sume svih elemenata u Benfordovskom skupu sa vodećom cifrom x za sve $x \in \{1, 2, ..., 9\}$ su jednake. Kao posljedica, suma svih elemenata sa vodećom cifrom x je $1/9$ sume svih elemenata (Nigrini, 1992, 71).

Osnovna teorema na osnovu koje se razrađuje invarijantnost suma glasi : *Distribucija je invarijantna u smislu sume ako i samo ako je Benfordovska* [22]

Mada je Nigrini dao objašnjenje, na osnovu teorije brojeva, Allaart je dao probabilistički dokaz [22]. Kako bi se dobila prikladna formulacija bitne su tri stvari. Prvo, u smislu ovog svojstva, predmet sumiranja su mantise a ne brojevi sami po sebi. Drugo, riječ *konstantna* u definiciji koju je dao Nigrini treba biti zamijenjena sa *konstanta u očekivanju.* Razlog je u tome da za bilo koji konačni uzorak iz Benfordove distribucije sume skoro sigurno nisu konstantne. Uslov striktne jednakosti suma je pretežak. Može se pokazati da u slučaju Benfordove distribucije, podrazumijevajući nezavisne stavke, devet suma ima različite

momente drugog reda. Momenti prvog reda su u potpunosti saglasni sa ovom tvrdnjom. Treće, kako bi se uspostavila jedinstvenost neophodno je uzeti u obzir druge i treće cifre itd.

Ako se ovo ima u vidu, invarijantnost suma (sum-invariance) se može definisati na sljedeći način :

Distribucija je invarijantna u smislu sume ako je za bilo koji prirodni broj k očekivana suma mantisa svih stavki, počev od fiksne k-torke vodećih cifara, jednaka sa onima za bilo koju drugu k-torku.

Pod pojmom k-torka podrazumijeva se vodećih k cifara. Ovdje neće biti prezentiran teorijski aparat iza ovakve tvrdnje. Očigledna je praktična vrijednost ovog svojstva, koje zaslužuje da mu se posveti veća teorijska i praktična pažnja. Bitan element potreban za praktičnu primjenu ovog kriterija je izvođenje kritičnih vrijednosti.

3.2 Neke druge mjere iz podataka

3.2.1 Faktor izobličenja

U tekstovima *A taxpayer compliance application of Benford's law* (Journal of the American Taxation Association 18(1), 72-91) i *The Detection of Income Tax Evasion Through an Analysis of Digital Frequencies* (PhD thesis, University of Cincinnati, OH, USA, 1992) Nigrini razvija jednostavnu mjeru smjera i veličine izobličenja u skupu podataka koji slijedi Benfordov zakon pod uslovima nepristrasnog izvještavanja [33]. Nigrinijeva mjera se naziva Model Faktora Izobličenja (Distortion Factor Model - DFM) i zavisi od dvije pretpostavke. Prvo, bilo kakva manipulacija podacima ne mijenja red veličina vrijednosti sa kojima se radi. Pretpostavka je bazirana na psihološkom iskustvu da ljudi koriste red veličina kao referentne tačke, da su oni koji manipuliraju podacima oprezni od te tendencije i da stoga prilikom manipulacije izbjegavaju promjene koje daju upadljive promjene redova veličina. Drugo, model podrazumijeva da je relativna magnituda podataka nakon manipulacije slična magnitudi izvornih podataka (drugim riječima, prosječan stepen manipulacije jednak je u cijelom opsegu veličina). Ova pretpostavka je konzistentna sa činjenicom da manipulator izabire da mijenja podatke na način da je nivo značajnosti promjena sličan unutar redova veličina.

Model faktora izobličenja (MFI) podrazumijeva da skup podataka koji je predmet manipulacije slijedi Benfordov zakon i pokriva opseg $[10, 100)$. Ako skup pokriva veći opseg tada su podaci zbijeni (colapsed) ili razvučeni (expanded) na pretpostavljeni opseg micanjem decimalne tačke (zareza) putem transformacije :

$$X_{\text{col}} = \frac{10 \cdot X}{10^{int(\log_{10} X)}} \qquad (3.1)$$

gdje je X početna (sirova) vrijednost, X_{col} odgovarajuća sažeta (ili razvučena) vrijednost, *int* funkcija 'cijeli dio' koja uklanja decimalna mjesta desno od decimalne tačke. S obzirom na pretpostavke :

1. za podatke se smatra da su nepristrasni i slijede Benfordov zakon
2. Benfordova distribucija je mjerno invarijantna
3. sve manipulacije podacima su proporcionalne u odnosu na red veličina

zbijanje (collapsing) odnosno razvlačenje (expanding) podataka ne izobličuje bilo kakvu procentualnu manipulaciju koja je možda prisutna u podacima. Brojevi sa manje od dvije značajne cifre nakon zbijanja / razvlačenja brišu se iz skupa. Probom je moguće utvrditi da korištenje funkcija zaokruživanja (ROUND()) nije isto kao korištenje funkcije cijeli dio (INT()).

Model faktora izobličenja poredi srednju vrijednost izvedenog skupa i srednju vrijednost skupa nepristrasnih podataka koji sadrže isti broj opservacija unutar istog opsega i slijede Benfordov zakon. Stvarna srednja vrijednost, AM (Actual Mean) podataka je

$$AM = \frac{1}{N} \sum X_{\text{col}} \tag{3.2}$$

gdje je N broj opservacija. Nigrini [33] pokazuje da očekivana srednja vrijednost, EM (Expected Mean), nepristrasnog skupa podataka sa N opservacija na intervalu $[10, 100)$ iznosi :

$$EM = \frac{90}{N \cdot (10^{1/N} - 1)} \tag{3.3}$$

Faktor izobličenja, DF(Distortion Factor), se računa kao :

$$DF = \frac{100 \cdot (AM - EM)}{EM} \tag{3.4}$$

DF daje prosječan procenat manipulacije podacima. Nigrini pokazuje da je očekivana vrijednost za DF 0 i da je standardna devijacija za DF, STD(DF) jednaka :

$$STD\,(DF) = \frac{\left[11 \cdot N \cdot \left(10^{1/N} - 1\right)\right] - \left[9 \cdot \left(10^{1/N} + 1\right)\right]}{9 \cdot N \cdot (10^{1/N} + 1)} \tag{3.5}$$

S obzirom da je AM očekivanje za N slučajnih varijabli, na osnovu centralne granične teoreme, distribucija za DF se približava normalnoj distribuciji sa očekivanjem 0 i varijansom $[STD\,(DF)]^2$ kako se N povećava. Kao rezultat, z-statistika može biti računata za DF za relativno veliko N.

3.2.2 Normalizacija

Primjer skupa koji se dobije kada se primijeni transformacija data jednadžbom (3.1) dat je na grafikonu P.3.1 u prilogu. Vidljivo je da ova transfomacija vrši translaciju podskupova podataka iz intervala $[10^n, 10^{n+1})$ gdje je n u opsegu od minimalne do maksimalne vrijednosti veličine $int\,(\log_{10} X)$.

Jednadžba (3.5) se može napisati u opštem obliku :

$$X_{\text{col}} = \frac{10^{k-1} \cdot X}{10^{int(\log_{10} X)}} \qquad (3.6)$$

pri čemu je k odabrani broj značajnih cifara. Bitno je primijetiti da transformacija daje brojeve isključivo iz jednog reda veličina koji se odabire sa k. Ako je $k = 2$ dobija se izraz (3.1). Ako je $k = 3$ rezultat su brojevi iz intervala $[100, 1000)$ itd. Ova transformacija može biti korištena za normalizaciju putem decimalnog skaliranja. Osnova za ovo je svojstvo mjerne invarijantnosti Benfordovih skupova. Ako je S skup numeričkih veličina koji je predmet analize i ako je

$$m = \min_{X \in S} int\,(\log_{10} X)$$

postupak normalizacije bi se provodio putem transformacije

$$X_{nor} = \frac{X}{10^m}$$

Ako je m negativan to znači da je najmanji broj manji od 1. Rezultat noramlizacije su numeričke veličine u kojima je decimalni zarez pomaknut za $|m|$ pozicija udesno tako da najmanji podatak ima prvu značajnu cifru različitu od nule ispred decimalnog zareza. Ako je m pozitivan to znači da najmanji broj ima $m + 1$ značajnih cifara. Rezultat normalizacije su veličine u kojima je decimalni zarez pomaknut za m pozicija ulijevo. Drugim riječima, ova normalizacija osigurava da je najmanji podatak iz intervala $[1, 10)$. Korištenje funkcija $\lfloor \cdot \rfloor$ (floor) ili $\lceil \cdot \rceil$ (ceiling) umjesto funkcije int() ne bi osiguralo ovaj rezultat.

Za razliku od transformacije koju predlaže Nigrini, ovaj postupak normalizacije ne mijenja početni međusobni sortni poredak bilo koje dvije stavke u uzorku. Iz tog razloga je ovaj metod pogodan u situacijama kada treba osigurati da i najmanji broj ima potreban broj značajnih cifara ispred zareza. Naprimjer, u programskom paketu Excell korištenje funkcije LEFT() može dati rezultat koji nije iskoristiv. Tako, funkcija LEFT(12,396;3) daje rezultat "12," a ne "123".

Za Benfordove skupove ne vrijedi svojstvo invarijantnosti u odnosu na transformaciju translacije za konstantnu vrijednost. Drugim riječima, oduzimanje / dodavanje iste vrijednosti narušava Benfordovska svojstva.

3.2.3 Regresija

Nigrini je predložio korištenje linearne regresije u cilju testiranja sličnosti između dva grafa provjerom korelacije između njihovih dijelova [31]. Veliki stepen korelacije povlači sličnost oblika dva grafa u analizi cifara. Linearna regresija se izražava relacijom

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$$

gdje je Y_i vrijednost varijable odgovora u i-tom pokušaju, X_i poznata konstanta odnosno nezavisna varijabla u i-tom pokušaju, β_0 i β_1 parametri, a ε_i slučajna greška. Pritom je očekivanje $E(\varepsilon_i) = 0$, varijansa $\sigma^2 = \{\varepsilon_i\}$, a veličine ε_i i ε_j su nezavisne za $i \neq j$ tj. kovarijansa je jednaka nuli : $\sigma\{\varepsilon_i, \varepsilon_j\} = 0$. Potpuno slaganje znači da je $\beta_0 = 0$ i $\beta_1 = 1$.

Predmet regresije su teorijske frekvencije u odnosu na uzoračke frekvencije. Ako su predmet analize prva, druga ili treća cifra tada broj opservacija nije veći od 10. Stoga je ovaj test prikladan za testove prve dvije ili prve tri cifre [2]. Prvi korak u ovom postupku je da se formira scaterplot tako da se na jednoj osi stavlja teorijska a na drugoj uzoračka frekvencija. Ako male (velike) frekvencije iz jednog skupa odgovaraju malim (velikim) vrijednostima iz drugog skupa tada se može reći da su dva grafa slična.

U prilogu P.3.2 je primjer jednog scaterplot grafa za test prve dvije cifre. Korištenjem programskog alata Excell dobijen je regresioni model $y = 1,1947 \cdot x - 0,0022$. Vidljiv je povećani stepen grupisanja za male vrijednosti što nije slučaj za veće vrijednosti. Ovakav izgled može ukazivati na značajne razlike dva uzorka.

3.2.4 Odzivi na osnovu Benfordovog zakona

Benfordov zakon daje mogućnost definisanja veličina koje se mogu koristiti za metode reinforcement učenja. Fletcher Lu i Efrim Boritz [5] su u patentnoj prijavi Adaptivne Benfordove metode predložili metod računanja odziva (reward) koji se dodjeljuje pojedinoj numeričkoj veličini ili grupi numeričkih veličina u posmatranom skupu podataka.

Odziv (reward) za grupu stanja. Ako uzorak sadrži više numeričkih kolona koje zadovoljavaju uslove za analizu putem Benfordovog zakona predlaže se računanje veličine

$$R(s) = \sum_{cv} \sum_{\text{seq}} \frac{|P_{ocek} - P_{uzor}|}{P_{ocek}}$$

U ovom izrazu cv predstavlja numeričke kolone uzorka; veličina seq označava vodeće sekvence dužine tri; veličina P_{ocek} očekivanu (teorijsku) relativnu frekvenciju prve tri cifre a P_{uzor} uzoračku relativnu frekvenciju prve tri cifre. Prema tekstu patenta, ovo daje zbirnu vrijednost za nekoliko numeričkih atributa. Ovakva veličina je pogodna za mnoge metode data mininga kao što su mjere sličnosti, klastering itd.

Odziv za jedno stanje. U tekstu u kojem obrađuje primjer detekcije prevara korištenjem metoda reinforcement učenja [63, 64] Fletcehr Lu predlaže korištenje veličine :

$$BE\,(i) = \frac{f_{1i}}{b_{1i}} + \frac{f_{2i}}{b_{2i}} + \frac{f_{3i}}{b_{3i}}$$

U ovom izrazu f_{ji} predstavlja uzoračku a b_{ji} teorijsku frekvenciju grupa cifara dužine j za stanje (slog) i. Ovaj veličina se može računati za bilo koji numerički atribut koji odgovara kriterijima Benfordovog zakona bilo da se koristi osnovna odnosno adaptivna metoda računanja. Radi jednostavnosti, u nastavku će ova veličina, iz razloga praktične prirode, biti obilježena sa

$$BE\,(3) = \frac{f_1}{b_1} + \frac{f_2}{b_2} + \frac{f_3}{b_3}$$

Ovim se želi naglasiti da se kalkulacija pravi za najviše tri prve cifre. Pritom se ova veličina računa za svako pojedino stanje odnosno slog u uzorku. Mada je, po analogiji, moguće je praviti kalkulaciju za više od tri prve cifre, u praksi to nije je uobičajeno. Uticaj strukture podataka na ovu veličinu i korištenje ove veličine u metodama reinforcement učenja je osnovni predmet ovog teksta.

3.3 Postupak testiranja

3.3.1 Raspoloživi testovi

Nigrini i Mittermaier (1997) daju pregled šest testova u domenu analize cifara [29]. To su :

1. Test prve cifre,
2. Test druge / treće / četvrte cifre,
3. Test prve dvije cifre,
4. Duplikacija brojeva,
5. Zaokruživanje,
6. Posljednje dvije cifre

U istim tekstovima su date njihove karakteristike sa stanovišta (kontinuirane) revizije, što je logično ako se ima u vidu da se ovi autori bave u prvom redu računovodstvenim analizama [29, 31, 2].

Test prve cifre je inicijalni test razložnosti kojem nije namjena izbor uzorka. Visok stepen saglasnosti sa Benfordovim zakonom, sa generalnog stanovišta, signalizira da su podaci prošli test razložnosti. Ovo je, generalno gledano, prvi test koji se provodi tokom analize cifara. Sa stanovišta operativne potrebe, testovi prve i druge cifre možda i nisu potrebni jer su sve informacije informacije iz testova prve i druge cifre sadržane u testu prve dvije cifre.

Test druge / treće / četvrte cifre se koristi kao dodatni test razložnosti. Ovi testovi, takođe, nisu namijenjeni za izbor uzorka.

Test prve dvije cifre se predlaže kao relevantan sa stanovišta kontinuirane revizije. Ovaj test je balans između između toga da ne bude prevelikog nivoa (kao test prve i druge cifre) i ne previše fokusiran (kao test prve tri cifre). Ovaj test može biti korišten za izbor uzorka ali više na indirektan način. Kombinacije grupa cifara (odnosno stavke predstavljene ovim grupama vodećih cifara) koje imaju frekvenciju značajno različitu od predviđene idealni su kandidati za pažljivije istraživanje. S obzirom da su stvarni skupovi podataka samo aproksimativno Benfordovski, testovi prve dvije / tri / četiri cifre često rezultiraju lažnim alarmima koje mogu značiti moguće uzaludne napore revizora. Stoga ovi testovi imaju samo indirektnu ulogu u izboru uzorka u odnosu na direktne metode. Oni indiciraju koje grupe su previše korištene i gdje se druge analitičke / suštinske procedure kontrole trebaju usmjeriti. Test prve i druge cifre su preliminarni testovi.

Test dupliciranja brojeva rangira pojedine brojeve u skladu sa frekvencijom pojavljivanja. Postoji uska povezanost između ovog testa i testa prve dvije cifre. Naprimjer, izboj (spike) na grafu testa prve dvije cifre za neku vrijednost, npr. 50, može značiti da postoji veliki broj brojeva koji počinju sa 50 (50, 500, 5000, 505,...). Ovo sugeriše da se ovaj test koristi samo ako se na grafikonu uoče spikeovi koji se zatim porede sa stavkama iz uzorka. Test dupliciranja je prirodno proširenje analize prve dvije cifre [2]. Smatra se da je povećana frekvencija nekih vodećih cifara uzrokovana dupliciranjima koja signaliziraju neefikasnost, greške u radu, (ne)namjerne greške u izvještavanju i vrhunske prevare. Primjer neefikasnosti u računovodstvu je procedura evidentiranja pojedinačnih nabavki koje se ponavljaju umjesto korištenja grupnih računa. Pogrešno izještavanje je npr. sistematsko zaokruživanje iznosa a prevara je ako neko u kompaniji izdaje fiktivne narudžbe za usluge koje se ne izvršavaju. Posebna pažnja bi se trebala posvetiti brojevima koji su malo ispod psiholoških pragova internih veličina autorizacije jer su one najviše svojstvene pogrešnom izvještavanju i prevarama. Prilikom provođenja testa na dupliciranje brojeva može se računati tzv. Number Frequency Factor (NFF), za odabrane skupove.

Test na zaokružene brojeve se koristi tamo gdje procjena nije prihvatljiva ni na koji način. Pretjerana frekvencija brojeva koji su multiplikanti za 5, 10, 25, 100, 1000 i slično mogu signalizirati neprihvatljiv nemar ili čak moguću prevaru. Testovi na zaokružene brojeve su posebno prikladni u analizama pranja novca.

Test posljednje dvije cifre je fokusirana verzija testa na zaokružene brojeve i relevantan je kada se pojavi sumnja na 'sistematsko izmišljanje / namještanje' brojeva. Pristrasnost prema manjim ciframa se smanjuje pri kretanju udesno po pozicijama brojeva. Iako rezultat ima asimptotsko značenje simulacije pokazuju da je razložno pretpostaviti da su u stvarnim podacima posljednje dvije cifre uniformno distribuirane sa relativnom frekvencijom 1/100 pod uslovom da broj ima barem pet značajnih cifara.

3.3.2 Karakteristike testova

Bilo kakvo odstupanje od pretpostavljene distribucije cifara može se pripisati jednom od sljedeća dva faktora [2] :

- Greška u postupku uzimanja uzorka
- Manipulacija podacima, prevara, greške, neefikasnost

Kako bi se napravila razlika između dva moguća uzroka odstupanja koristi se nekoliko klasičnih i ne-klasičnih testova. Najvažniji od tih testova su :

- Vizualno ispitivanje
- Srednje apsolutno odstupanje (MAD - Mean Absolute Deviation)
- χ^2-test
- z-test
- Kolmogorov-Smirnov test
- Test sumiranja

Vizuelno ispitivanje je važan prvi korak u analizi cifara kao i u mnogim drugim primjenama statistike. Stavljanje vrijednosti na dijagram može dati brzu smjernicu za naredni korak. Iz tog razloga analitički alati moraju podržavati vizualno ispitivanje.

Srednje apsolutno odstupanje (MAD) mjeri ukupno odstupanje uzorka od pretpostavljene teorijske distribucije. Osnovni nedostatak ovog testa je nepostojanje egzaktno utvrđenih kritičnih vrijednosti. Analitičar ne može reći da je za test prve dvije cifre npr. vrijednost 0.0006 prihvatljiva a da je npr. 0.0186 neprihvatljiva. Sve što se može zaključiti jeste da je prva situacija bolja procjena od druge. Drugo, MAD može sakriti značajne probleme. Ako se pretpostavi situacija u kojoj je devijacija za prvih 88 cifara prihvatljiva (razlika skoro nula) a da su dva odstupanja plus 4.5% i minus 4.5% respektvno MAD bi trebao iznositi oko 0.1%. Ako se pretpostavi situacija u kojoj su sva apsolutna odstupanja oko 0.1% MAD bi u tom slučaju iznosio oko 0.1%. Obje situacije daju isti MAD ali sa stanovišta analize prva situacija skreće pažnju mnogo više od druge. Nigrini je predložio skalu kritičnih vrijednosti koja je donekle subjektivna i stoga se mora koristiti sa oprezom.

Pearsonov hi-kvadrat test je standardni metod u statistici kojim se testira u kojoj mjeri se uzoračka populacija poklapa sa teorijskom distribucijom. Prije korištenja u obzir se mora uzeti sljedeće :

- Ovaj test je samo aproksimacija pa je za male uzorke preporučljivo uzimati tačne (npr. multinomialne) testove
- Hi-kvadrat distribucija zahtijeva nezavisne sabirke

Testiranje na Benfordovu distribuciju obavlja se na velikim skupovima podataka (1.000 a čak i više od 10.000 stavki) skoro u svim slučajevima, tako da prvi uslov nije poseban problem. U nekim primjenama je najupitniji drugi uslov s obzirom da podaci u nekim skupovima nisu nezavisni[1].

[1] Varijansa cifara po pozicijama ukazuje na to da cifre u Benfordovskim skupovima nisu nezavisne

Test z-statistike je najbolje koristiti ako se želi utvrditi koja kategorija u uzorku ima prekomjernu frekvenciju u odnosu na teorijske vrijednosti [34]. Dok hi-kvadrat evaluira skup u cjelini, ovo je parcijalni test za pojedine kategorije, u ovom slučaju pojedine cifre. Ovo je dvostrani test sa kritičnim vrijednostima iz standardne normalne distribucije. Kalkulacija ovog testa koristi obim podataka (N) i ako je uzorak veći i ako je sve ostalo jednako, izračunata z-statistika će biti veća. Za velike uzorke z-statistika indicira značajne razlike, mada je sa praktičnog gledišta ta razlika često nebitna. Za prve dvije cifre trebalo bi računati 90 z-statistika i nema formule koja bi kombinovala 90 podataka kako bi se izvukao zaključak o podacima u cjelini.

Kolmogorov-Smirnoff test ima nekoliko važnih ograničenja :

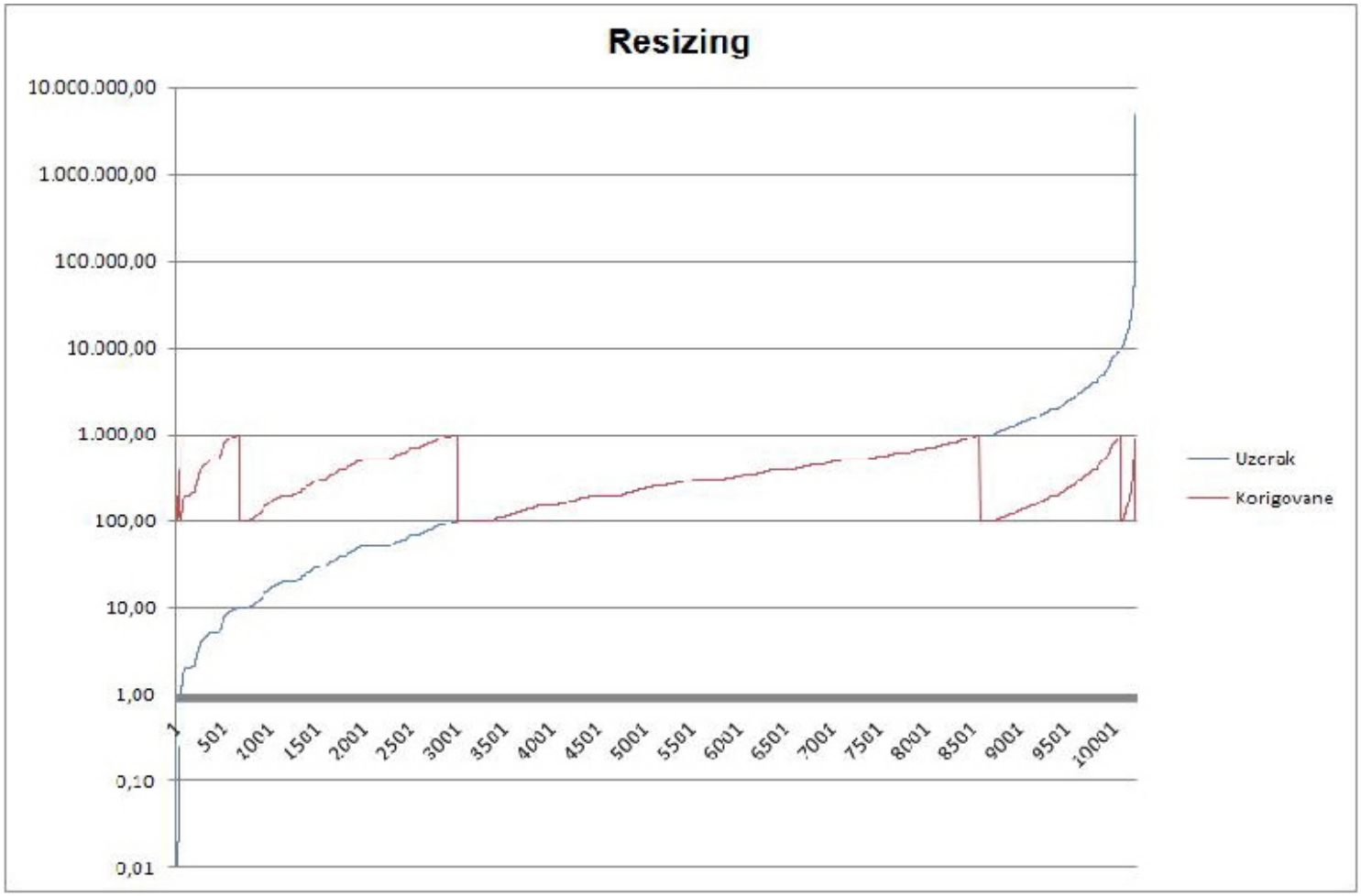
- Primjenjiv je samo na neprekidne distribucije
- Teži da bude osjetljiviji oko centra distribucije nego na njenim krajevima
- Distribucija mora biti u potpunosti definisana, što je možda najozbiljnije ograničenje. To znači da ako su parametri lokacija, skale i oblika procijenjeni na osnovu podataka kritični region ovog testa nije više validan

Test sumacije je alternativni pristup koji je razvio Nigrini (1992). Matematičku osnovu je dao Allaart koji je pokazao da je invarijantnost suma ekskluzivno svojstvo Benfordove distribucije [2]. Glavni cilj korištenja ovog testa je detekcija grešaka u magnitudama podataka putem sumiranja vrijednosti sa istim vodećim ciframa. Ovaj test čuva od malog broja velikih brojeva koji bi ostali nedetektovani testom prve ili druge cifre. Nigrini je naveo istinit primjer iz Wall Street Journal u kojem je čovjek, koji je očekivao da plati porez u visini od 513 USD, dobio obavještenje da duguje $300.000.007,57$ USD [31]. Isto se desilo sa još 3.000 ljudi. Zvanično objašnjenje je bilo da je u pitanju 'judska greška u programiranju'. Ovakva greška ne bi mogla biti detektovana korištenjem standardnih testova analize s obzirom da 3.000 grešaka nije značajno na nivou cijele populacije SAD. Ovoliko brojeva koji imaju 30 kao prve dvije cifre ne bi napravilo izboj (spike) na grafikonu. Međutim, suma ovih brojeva je ekstremno velika i kao takva izaziva pažnju. Ako je npr. 412.00 NJ najčešća stavka u isplatama plaća tada greška pomjeranja zareza dva mjesta udesno dovodi do broja 41.200 NJ a ovo ne bi moglo biti detektovano grafikonom za prve dvije cifre.

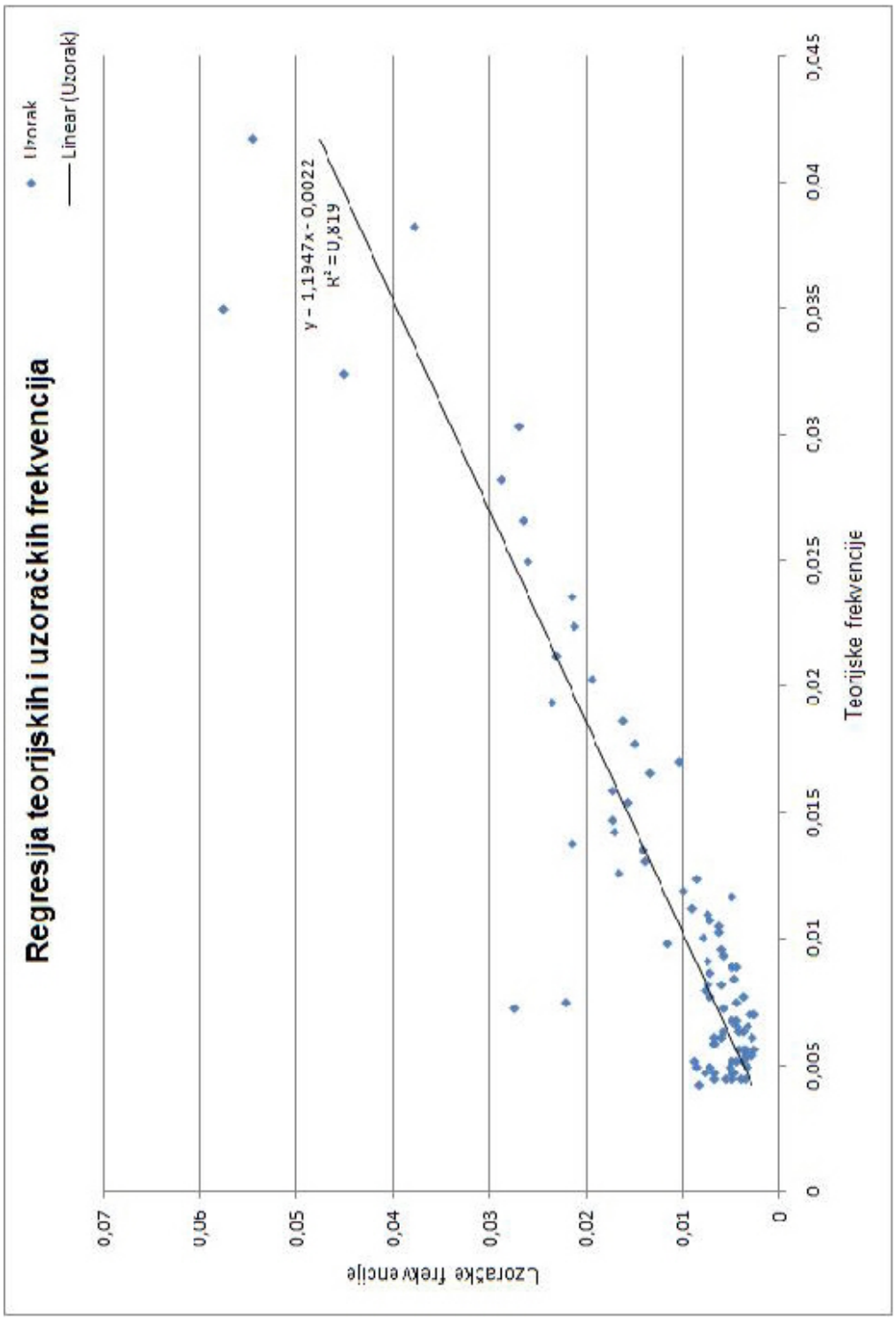
3.4 Zaključak

U ovom poglavlju je dat pregled testova saglasnosti sa Benfordovim zakonom, zajedno sa njihovim osnovnim karakteristikama. Izbor testa u velikoj mjeri zavisi od konteksta podataka i analize koja se provodi.

3.5 Prilog



Prilog 3.1. Uticaj transformacije 3.1 na uzorak. Vidljiv je efekat vertikalne translacije uzrokovan pomjeranjem decimalnog zareza



Prilog 3.2. Primjer scaterplot grafa za test prve dvije cifre. Vidljivo je grupisanje za manje vrijednosti uzoračkih i teorijskih frekvencija

4 Kada (ne) koristiti Benfordov zakon

4.1 Opšti uslovi primjene Benfordovog zakona

Opšte uslove koje trebaju zadovoljavati podaci koji su predmet analize putem ovog zakona formulisali su Mark Nigrini i Linda Mittermeier [3,29] a mogu se naći i u drugim tekstovima [npr. 30, 36, 37] a mogu se formulisati na sljedeći način :

1. Podaci moraju opisivati sličan fenomen odnosno isti atribut
2. Ne smije se postavljati limit u obliku minimalnih i/ili maksimalnih vrijednosti
3. Podaci ne bi trebali biti generisani putem prethodno definisanih ili dodijeljenih vrijednosti odnosno trebali bi imati slučajnu prirodu
4. Podaci bi trebali sadržavati više malih nego velikih vrijednosti
5. Podaci trebaju biti iskazani istim mjernim jedinicama
6. Podaci trebaju obuhvatiti barem dva reda veličina

Prvi uslov upućuje na podatke koji imaju istu prirodu odnosno isti izvor. Primjer su finansijske transakcije, rezultati raznih mjerenja, dužine, količine ili drugo svojstvo koje se izražava numeričkom veličinom.

Drugi uslov upućuje na to da se ovaj zakon ne može posmatrati na skupu podataka za koje je unaprijed poznato da su generisani sa ograničenjima. Ovaj uslov je bitan jer dopušta da ekstremne budu prirodan dio podataka koji su predmet analize. Drugim riječima, minimalne ili maksimalne vrijednosti ne moraju nužno biti anomalične.

Isti uslov mora važiti i za skup podataka koji je izdvojen iz nekog skupa. Naprimjer, iz skupa godišnjih transakcija se ne mogu izolovati isplate sa bankomata i testirati ih na Benfordov zakon jer su to vrijednosti za koje postoje tačno definisana ograničenja i koje su zaokružene. Sa druge strane, može biti zanimljivo da li skup iz kojeg se izbaci neki podskup podataka i dalje slijedi Benfordov zakon odnosno da li i u kojoj mjeri podskup podataka koji se izbacuje ima uticaj na usklađenost sa Benfordovim zakonom. Ovaj metod u obliku izbacivanja iznosa manjih od nekog praga Nigrini je u svojim analizama [3,29] koristio kao legitiman. Bitan pomak u pogledu ovog uslova napravljen je putem tzv. Adaptivne Benfordove metode koju su patentirali Fletcher Lu i Efrim Boriz [5].

Treći uslov upućuje na to da za analizu nisu pogodni podaci koji imaju utvrđenu fiksnu strukturu ili se formiraju prema unaprijed poznatim pravilima. Primjeri su serijski brojevi, telefonski brojevi, matični brojevi građana i firmi, brojevi socijalnog osiguranja, poreski brojevi, registracije automobila, brojevi knjigovodstvenih računa, analitičke partije i slično.

Četvrti uslov se interpretira na način da srednja vrijednost podataka treba biti manja od medijane i da skup treba imati pozitivnu asimetriju (skewness) [30]. Što je veći količnik srednje vrijednosti podijeljen medijanom skup je prikladniji za Benfordovu analizu. Drugo objašnjenje je da uzorak treba biti unimodalan s tim da modalna vrijednost nije nula.

Radi podsjećanja, mod skupa je vrijednost koja se najviše pojavljuje. Unimodalnost znači da ne postoji više vrijednosti koje imaju visoku frekvenciju. Provjera ovih uslova bi, u suštini, trebala biti prvi korak u pristupu analizi.

Peti uslov upućuje na to da prije testa treba provjeriti da li su podaci iskazani istim mjernim jedinicama kao što su valute, dužinske ili težinske mjere, na istoj skali temperature i slično.

Šesti uslov upućuje na uslov da podaci trebaju biti iz intervala (B^m, B^n) gdje je $n-m \in \mathbb{Z}$ [23]. Formalno matematički, ovaj uslov je potvrđen formulacijom koju su dali Leemis, Schmeiser i Evans [43] :
Neka je $W \sim U(a,b)$ *gdje su* a *i* b *realni brojevi za koje vrijedi* $a < b$. *Ako interval* $\left(10^a, 10^b\right)$ *obuhvata cijeli broj redova veličina tada prve cifre slučajne varijable* $T = 10^W$ *slijede Benfordov zakon tačno.*

Značenje navedene tvrdnje je da je to distribucija vjerovatnoća prvih cifara svih mogućih vrijednosti varijable T koje formiraju Benfordov skup. T je slučajna varijabla i samo jedan broj može ne biti 'Benfordov'. Dakle, ako je razlika $\log b - \log a$ cijeli broj i ako su logaritmi ekvidistribuirani tada brojevi nastali eksponenciranjem slijede Benfordov zakon.

Svi navedeni uslovi imaju svoje formalne matematičke formulacije. Dio tih formulacija predstavljen je u dijelu u kojem su opisani postupci izvođenja Benfordovog zakona.

4.2 Primjeri korištenja Benfordovog zakona

Primjeri u nastavku su ilustracije veoma velikog broja naučnih područja i ljudskih djelatnosti u kojima su pokazane i dokazane praktične vrijednosti analize cifara putem Benfordovog zakona.

4.2.1 Matematika

Dinamički sistemi. U tekstu [38] je pokazano da mnogi dinamički sistemi generišu podatke koji zadovoljavaju Benfordov zakon, uključujući mnoge stepene, eksponencijalne i racionalne funkcije, linearno dominantne sisteme, autonomne i neautonomne dinamičke sisteme. Ovo je samo dokaz stalno rastuće porodice sistema za koje se vjeruje da slijede Benfordov zakon kao što su fizičke konstante, cijene dionica na berzama, porezi, sume i proizvodi slučajnih varijabli, faktorijeli, Fibonačijevi brojevi i mnogi drugi. Analiziraju se nizovi oblika $x_{n+1} = T_n(x_n)$, gdje je (T_n) sekvenca mapa realne ose na samu sebe. Osnova analize je direktna korespondencija između Benfordovskih sekvenci i uniformne distribucije $\bmod 1$ u obliku sljedeće propozicije :

Propozicija 4.1. Sekvenca $(x_n)_{n\in\mathbb{N}_0}$ realnih brojeva je Benfordovska ako i samo ako je $(\log_b |x_n|)_{n\in\mathbb{N}_0}$ uniformno distribuiran $\bmod 1$.

Kongruencija $\bmod 1$ se definiše na sljedeći način.

Definicija 4.1. $u \equiv v \bmod 1$ ako i samo su mantise za B^u i B^v jednake, po odabranoj bazi B.

Jedna od posljedica je da Fibonačijevi brojevi zadovoljavaju Benfordov zakon.

Diferentne jednadžbe. Povezanost rekurentnih relacija i Benfordovog zakona detaljno je elaborirana u [14]. Rješenja diferentne jednadžbe

$$a_{n+k} = c_1 \cdot a_{n-k+1} + c_2 \cdot a_{n-k+2} + ... + c_k \cdot a_n$$

se uzimaju u obliku

$$a_n = u_1 \cdot \lambda_1^n + u_2 \cdot \lambda_2^n + ... + u_k \cdot \lambda_k^n$$

pri čemu se uzima $|\lambda_1| \geq |\lambda_2| \geq ... \geq |\lambda_k|$. Naredna teorema, koja ovdje neće biti dokazivana, formuliše uslove pod kojima značajne cifre rješenja a_n zadovoljavaju Benfordov zakon.

Teorema 4.1. *Neka je a_n rješenje diferentne jednadžbe reda k sa različitim realnim korijenima. Pretpostavimo da je $|\lambda_1| \neq 1$ gdje je $|\lambda_1|$ korijen koji ima najveću apsolutnu vrijednost. Dalje, pretpostavimo da su inicijalni uslovi takvi da koeficijent za λ_1 nije nula. Ako vrijedi $\log_b |\lambda_1| \notin Q$ tada je niz a_n Benfordovski.*

U tekstu [66] je pokazana duboka veza Benfordovog zakona i konačnih lanaca Markova. Sekvenca realnih brojeva $\{x_n\}$ je Benfordovska ako su njene značajne cifre odnosno decimalni dio prezentacije članova $\{x_n\}$ u obliku pokretnog zareza, distribuirane logaritamski. Slično, diskretni nereducibilni i neperiodični lanac Markova sa konačnim brojem stanja sa matricom vjerovatnoća prelaza P i matricom ograničenja P^* je Benfordovski ako je svaka komponenta sekvenci matrica $(P^n - P^*)$ i $(P^{n+1} - P^n)$ Benfordovska ili eventualno nula. Korištenjem alata kojima je utvrđeno Benfordovsko svojstvo Njutnove metode i konačno dimenzionalnih linearnih mapa, putem klasičnih teorija uniformne distribucije mod 1 i Perron-Frobenius, u ovom tekstu je izveden dovoljan uslov ('nerezonansa') koji garantuje da su ili P ili lanac Markova povezan sa njom Benfordovski. Ovaj uslov se koristi kako bi se pokazalo da skoro svi lanci Markova imaju Benfordovsko svojstvo u smislu da ako su vjerovatnoće prelaza odabrane nezavisno i neprekidno tada rezultujući lanac Markova ima Benfordovsko svojstvo sa vjerovatnoćom 1.

Jedna od činjenica na koju se oslanja dokaz je da ako je $\{x_n\}$ Benfordovski tada za $\forall \alpha \in \mathbb{R} \wedge k \in \mathbb{Z} : \alpha k \neq 0$ vrijedi da je niz $\left\{\alpha x_n^k\right\}$ takođe Benfordovski. Druga činjenica, usko povezana sa ovim je da ako su α, β, a, b realni brojevi za koje je $|\alpha| > |\beta|$ i $a \neq 0$ tada je $(a\alpha^n + b\beta^n)$ Benfordovski ako i samo ako je $\log |\alpha|$ iraconalan.

U tekstu je naznačena praktična vrijednost u problemima koji se pojavljuju u računskim operacijama u pokretnom zarezu gdje se javljaju problemi zaokruživanja (round-off), prekoračenja (overflow) i potkoračenja (undreflow). Ako se ima u vidu da je svojstvo Markova jedan od ključnih preduslova za korištenje metoda reinforcement učenja ova veza sa Benfordovim zakonom je dokaz duboke veze ovih procesa i distribucije cifara odnosno veličina.

Colatzova konjuktura. Sljedeći primjer rekurentne relacije je Colatzova konjuktura, poznata kao $3x + 1$ problem. Ako se uzme cijeli broj x_i naredni se izračunava na sljedeći način :

$$x_{i+1} = \begin{cases} 3x_i + 1 & x_i \text{ neparan, } x_i \neq 1 \\ \dfrac{x_i}{2} & x_i \text{ paran} \end{cases}$$

Analiza ovog problema u [40] polazi od toga da ako je x neparan prirodni broj tada je $3x+1$ paran broj pa se može naći prirodni broj k takav da $2^k \parallel (3x+1)$ i da je $y = \dfrac{3x_i+1}{2^k}$ neparan. Na taj način je definisana mapa $M : x \longmapsto y$. Vrijednost k iz definicije se zove k-vrijednost za x. Ovdje je y neparno i relativno prosto sa 3 tako da je prirodan domen za M skup $\not\geq$ brojeva relativno prostih sa 3 i 2. Neka je $\Pi = 6\cdot\mathbb{N}+E$ gdje je $E = \{1,5\}$ skup mogućih klasa kongruencije modulo 6. Kompjuterski je potvrđeno da za svaki prirodan broj $0 < x < 2^{60}$ proces dovodi do fiksne tačke 1.

Riemanova zeta-funkcija. U tekstu [40] je pokazano da i L-funkcije generišu Benfordov skup. Posmatra se Riemanova zeta funkcija

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{(p\ prost)} \left(1 - \frac{1}{p^s}\right)^{-1}$$

Inicijalno definisana za $\mathrm{Re}\,(s) > 1$ funkcija $\zeta(s)$ ima meromorfno proširenje na $\mathbb{C}$. Generalno gledano, predmet analize je L-funkcija

$$L(s,f) = \sum_{n=1}^{\infty} \frac{a_f(s)}{n^s} = \prod_{(p\ prost)} \prod_{j=1}^{d} \left(1 - \frac{\alpha_{f,d}(p)}{p^s}\right)^{-1}$$

gdje koeficijenti $\alpha_f(n)$ imaju aritmetičko značenje. Poznati primjeri uključuju Dirichlet L-funkcije (gdje je $\alpha_f(n) = \chi(n)$ za Dirichlet simbol χ) i L-funkcije eliptičkih krivih (gdje se $\alpha_f(p)$ odnosi na broj tačaka na eliptičkoj krivoj $\mathrm{mod}\, p$). Istražuju se prve značajne cifre vrijednosti L-funkcije. Početna tačka analize vrijednosti duž kritične linije $s = \dfrac{1}{2} + it$ je lognormalni zakon :

$$\lim_{T\to\infty} \frac{\mu\left(\left\{0 \leq t \leq T : \log\left|\zeta\left(\frac{1}{2}+it\right)\right| \leq y\sqrt{\frac{1}{2}\log\log T}\right\}\right)}{T} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-u^2/2} du$$

Iz tog razloga, gustine vrijednosti za $\log\left|\zeta\left(\frac{1}{2}+it\right)\right|$ za $t \in [0,T]$ su dobro aproksimirane sa očekivanjem 0 i standardnom devijacijom $\psi_T = \sqrt{\frac{1}{2}\log\log T} + \mathcal{O}(\log\log T)$

Prosti brojevi. U tekstu [41] Bartolo Luque i Lucas Lacasa su pokazali da Benfordov zakon važi za prve cifre prostih brojeva.

Operacije sa prirodnim brojevima. U tekstu [20] je dat nešto širi pregled frekvencije značajnih cifara za neke tipove operacija sa prirodnim brojevima kao što su kvadrati,

treći stepeni, kvadratni korijeni i slično. Statističkim χ^2 testom je pokazan visok stepen saglasnosti sa Benfordovim zakonom za Bellove brojeve, Catallan brojeve, Numeri ideoni, Fibonačijeve brojeve i Partition brojeve. S obzirom da se radi o rekurzivnim definicijama ovo u velikoj mjeri potvrđuje hipotezu da rekurzivni izrazi generišu brojeve po Benfordovom zakonu. Treba zapaziti da na malim uzorcima ($n < 10000$) prostih brojeva saglasnost nije potpuna.

Njutnova metoda. Arno Berger i Theodore P. Hill su pokazali da iteracije Njutnovom metodom zadovoljavaju Benfordov zakon [42]. Centralnu ulogu u tekstu ima sljedeća teorema.

Teorema 4.2. Neka je $f : R \to I$ realna funkcija za koju vrijedi $f(x^*) = 0$ i koja nije linearna.

1. Ako je x^* jednostruki korijen za f tada su $(x_n - x^*)$ i $(x_{n+1} - x_n)$ Benfordovski (po bazi b) za skoro sve x_0 u okolini od x^* i za sve b iz skupa $\mathbb{N}\setminus\{1\}$

2. Ako je x^* dvostruki korijen za f tada isti zaključak vrijedi za sve x_0 različite od x^* u nekoj okolini od x^* ako nije $b = 2^j$ za neko $j \in \mathbb{N}$

3. Ako je x^* korijen reda barem 3 tada isti zaključak vrijedi za sve x_0 različite od x^* u nekoj okolini od x^* i za sve $b \in \mathbb{N}\setminus\{1\}$

Ovo su samo neki od brojnih primjera mogućnosti primjene Benfordovog zakona u matematici. Poseban interes izazivaju rekurentne jednadžbe s obzirom da se pomoću njih mogu modelirati procesi koji su po svojoj prirodi pogodni za analizu putem ovog zakona. Primjer se može naći u tekstu [44], u kojem se obrađuju rekurentne i geometrijske serije, kao i u [45] koji se bavi vezom između diferencijalnih jednadžbi i ovog zakona. U tekstu [49] je data detaljnija analiza veze jednodimenzionih dinamičkih sistema i Benfordovog zakona.

4.2.2 Ekonomija

Najpoznatiji primjer primjene Benfordovog zakona u ekonomiji je analiza poreskih prijava koju je napravio Mark Nigrini i koja je označila početak njegovog korištenja u postupcima revizije. U nekoliko tekstova [3, 29, 46 su neki od njih] opisan je metod detekcije prevara. Osnovna premisa detekcije prevara je da neslaganje frekvencije prvih cifara sa Benfordovim zakonom može ukazivati na moguću neregularnost. Neslaganje sa Benfordovim zakonom ne mora automatski značiti da se radi o prevari već samo o mogućnosti koju treba ispitati. Ova metoda je brzo prihvaćena a od strane nekih regulatronih i nadzornih organa priznata kao validan revizorski alat, što je podržano i sve većim brojem standardnih programskih paketa koje koriste revizori. Detekcija prevara je jedna od najrašireniji primjena Benfordovog zakona. Sa analize poreskih prijava ovaj metod je brzo proširen na detekciju kartičarskih prevara i drugih oblika prevara podržanih elektronskim poslovanjem, s obzirom na njihov rastući broj, posebno ako se imaju u vidu neposredne štete koje treba identifikovati i/ili spriječiti.

U tekstu [47] je dat prikaz primjene Benfordovog zakona za analizu makroekonomskih podataka. Uzeti su različiti makroekonomski podaci 80 država svijeta. Države su podijeljene u šest grupa. Analiza pokazuje da odstupanja od Benfordovog zakona ne moraju

biti povezana sa pitanjima kvaliteta podataka već mogu biti rezultat naglašenih ekonomskih fluktuacija i strukturalnih nedostataka u raspoloživim podacima. Stoga, nesaglasnost sa Benfordovim zakonom ne bi trebala biti interpretirana kao znak lošeg kvaliteta makroekonomskih podataka. Bitno je obratiti pažnju u ovoj formulaciji da se ovaj zakon koristi za strukturalnu analizu podataka. Drugim riječima, dobijamo generalnu, široku sliku skupa u kojoj Benfordov zakon upućuje na moguće strukturalne nedostatke podataka. Druga bitna karakteristika ove analize je metod segmentiranje podataka, razdvajanje na ekonomski opravdane cjeline.

Primjena u ekonomiji se ne zaustavlja na prevarama. Ona je moguća u analizama investicionih programa, knjigovodstvenih izvještaja, prometa i brojnim drugim primjerima.

4.2.3 Kompjuteri

Peter Schatte je u tekstu "On mantissa distributions in computing and Benford's law" [J. Inform. Process. Cybernet 24 (1988), 443-455], na osnovu Benfordovske analize, utvrdio da je kompjuterski dizajn koji minimizira potreban kompjuterski smještajni prostor onaj zasnovan na bazi 8. Ovo je imalo bitan uticaj na razvoj kompjuterske tehnike u informatičkom dobu [51 i drugi]. Istraživači su počeli istraživati korištenje logaritamskih kompjutera kako bi našli načini ubrzanja kalkulacija.

Benfordov zakon u kompjuterskoj tehnici se može koristiti za analize veličine datoteka u folderima. Za potrebe te analize [48] pravi se lista naredbom :

```
dir /S /A-D /-C c:\ | sort > files.txt
```

Ovdje je files.txt naziv datoteke u koju se usmjerava rezultat naredbe. Pojava većeg broja datoteka čija se veličina iskazuje istim prvim značajnim ciframa, npr. veliki broj datoteka veličine 2048, može ukazivati na neki oblik anomalije. Također, postoji mogućnost da predmet analize budu veličine elektronskih žurnala (logova) koji se formiraju (u radu relacionih baza, operativnih sistema i slično) [46], vrijeme trajanja raznih procesa u multitasking i multiuser (višekorisničkim) okruženjima, numerički literali u izvornim programima i slično.

4.2.4 Ostale primjene

Weber-Fechnerov zakon. Brojne studije su posvećene vezi Benfordovog zakona i istraživanja koje je provodio Ernst Heinrich Weber [50]. On je formulisao zakon reakcije kod ljudi između stimulansa i reakcije (odgovora). Povećanjem stimulansa povećava se i odgovor i to po logaritamskoj stopi. Naprimjer, ako se stimulans poveća za faktor 2 reakcija se povećava za $\log_{10} 2$. Ako se stimulans poveća za faktor 3 reakcija se povećava za faktor $\log_{10} 3$. Prema Weber-Fechnerovom zakonu, ako uzmemo dovoljno veliki slučajan uzorak podataka o reakcijama naći ćemo da će 30.10% njih imati nivo stimulansa koji počinje cifrom 1, što je važno u kontekstu Benfordovog zakona. Prva istraživanja ovog tipa su vršena na način da su korišteni tegovi u rukama ispitanika kojima su bile vezane oči. Isti princip se može primijeniti na skoro sve vrste stimulansa kao što su svjetlo, zvuk, temperatura, doze lijekova i slično.

Demografija. U tekstu [52] je dat prikaz mogućnosti primjene Benfordovog zakona na analizu dinamike rasta populacija. Osnova su podaci o broju stanovnika 198 država 1997. godine. Zajedno sa veličinom populacije vršena je analogna analiza gustine naseljenosti. Brzina rasta populacija se daje izrazom

$$\frac{P(T)}{P(0)} = e^{rT}$$

gdje je $P(u)$ veličina populacije u momentu u a r stopa rasta. Ako se želi znati koliko treba vremenskih jedinica da se populacija udvostruči tada je

$$\frac{P(T_2)}{P(0)} = e^{rT_r} = 2 \Rightarrow rT_2 = \ln\left[\frac{P(T_2)}{P(T)}\right] = \ln 2$$
$$\Rightarrow T_2 = \frac{\ln 2}{r} = \frac{0.693\,15}{r} \approx \frac{0.70}{r}$$

Odavdje se dobija da za rast populacije od 100 do 200 uz stopu rasta od 2% treba 35 jedinica. Generalno gledano, broj vremenskih perioda potreban za rast populacije od $d \cdot 100$ do $(d+1) \cdot 100$ je dat izrazom

$$T = \left(\frac{d+1}{d}\right)/r = \frac{1}{M}\log_{10}\left(\frac{d+1}{d}\right)/r = \frac{1}{M}F_D/r$$

gdje je :
- M : konstanta
- $\ln(x) = \frac{1}{M}\log_{10}(x)$
- F_D : frekvencija prve značajne cifre d

Obrada fotografija. U [53] je pokazana mogućnost korištenja Benfordovog zakona za analizu digitalnih fotografija, posebno u smislu mogućnosti detekcije pokušaja da se one koriste za prenošenje skrivenih poruka i dokumenata. Pokazano je da magnituda gradijenta slike zadovoljava ovaj zakon i pruža mogućnosti korištenja kao što su entropija i kodiranje. U tekstu je posebno pokazano da distribucija značajnih cifara u blok-DCT koeficijentima zadovoljavaju Benfordov zakon i da PEG koeficijenti slijede Benfordov zakon u slučaju kada su komprimirani JPEG tehnikom samo jednom. Zakon nije zadovoljen ako se izvrši dvostruka kompresiija. Važan dio ovog teksta je prijedlog parametarski modela za formulaciju Benfordovog zakona

$$p(x) = N \cdot \log_{10}\left(1 + \frac{1}{s + x^q}\right), \qquad x = 1, 2, ..., 9$$

gdje je N faktor normalizacije koji čini da $p(x)$ bude distribucija vjerovatnoće, a s i q su parametri modela, kako bi se precizno opisala distribucija za različite vrste slika i različite faktore kompresije. Ova metoda je patentirana [58].

Obrada izbornih rezultata. Uspješnost primjene Benfordovog zakona u analizi obrade rezultata glasanja najavio je Walter R. Mebane Jr. koji je napravio više radova na ovu temu. Metod opisan u tekstu [56] ne zahtijeva da postoje kovarijacije u smislu pretpostavki kome su glasovi upućeni. Metod je baziran na distribuciji cifara nakon prebrajanja glasova tako da su brojevi glasova sami po sebi dovoljni. S obzirom da je baziran na maloj količini informacija metod sam po sebi ne može dijagnosticirati da li anomalija nužno upućuje na prevaru ili bilo koji oblik neregularnosti. Međutim, neki obrasci mogu ukazati na prevare. U tom smislu, metod se najbolje razumije kao indikator na koja mjesta treba usmjeriti pažnju.

Metoda je primijenjena na nivou lokacije i na nivou glasačke mašine[2]. Prvo pitanje je zašto treba očekivati mogućnost primjene Benfordovog zakona na podatke o glasanju ? Mada su neki predložili korištenje Benfordovog zakona za drugu cifru kao metod testiranja glasačkih prevara, jedan broj nadzornika glasanja se tome protivio. Autor je pokazao da bliži fokus na postupak dobijanja svakog glasa sugeriše statistički model koji veoma često producira frekvencije koje slijede Benfordov zakon. Još je važnija činjenica da uzoračke frekvencije broja glasova ne slijede Benfordov zakon. Situacija često prisutna u broju glasova je činjenica da podaci ne slijede precizno Benfordov zakon ali to važi za proces koji strogo koristi distribuciju druge cifre. U tekstu se koristi skraćenica 2BL za naziv ove distribucije. Ovdje neće biti prezentirani detalji metodologije ali je primjer ilustrativan u pogledu segmenta a posebno u smislu da je osnova metodologije bila distribucija druge a ne prve ili prve dvije cifre.

Stilometrija je analiza lingvističkih stilova i navika pisanja pojedinaca [57]. U pozadini je pretpostavka da svaki autor ima različite navike pisanja koji se ogledaju u elementima kao što je skup riječi, kompleksnost rečenica i frazeologija. Razlika među autorima može biti posljedica elemenata kao što su žanr ili sadržaj, autorsko iskustvo i kompetencija, komunikacione sposobnosti i očekivanja ciljane populacije. Stilometrija nastoji definisati svojstva autorovog stila i definiše statističke metode neophodne da se izmjere svojstva sličnosti između dva ili više tekstualnih izvora. Poseban fokus studija je korištenje stilometrije u identifikaciji autorstva e-mail poruka, kao moguća mjera borbe protiv spama i prevara. Jedan od načina za ovo je napraviti statistiku elemenata kao što su broj rečenica, broj riječi u rečenicama, broj riječi koje počinju velikim ili malim slovima, broj razmaka, broj znakova interpunkcije, broj nekih riječi i slično. Istraživanje opisano u [57] obuhvata 55 ovih svojstava a provedeno je na način da se statistika pravi za svakog autora a rezultati normaliziraju na interval $[0, 1]$. Zatim se provodi proces učenja u kojem se podaci klasificiraju korištenjem klastering metode k-najbližih susjeda (k-nearest neighbours) i Euklidskih distanci.

Pristup analizama tekstova putem Benfordovog zakona je usmjeren na analizu nekih specifičnih elemenata kao što je frekvencija riječi ili slova. Prvi primjer te primjene je analiza Biblije [54]. Uzeto je prvih 5 knjiga (Genesis, Exodus, Leviticus, Brojevi i Deuteronomija). U engleskoj verziji teksta utvrđena je frekvencija riječi : one, two, three, four, five, six, seven, eight, nine, ten, eleven, twelve, thirteen, fifteen, twenty, thirty, forty,

[2]Radi se o glasačkim mašinama putem kojih glasači u SAD imaju mogućnost glasanja

fifty, hundred, thousand, first, second, third, fifth, ninth, twelfth. Prebrojano je ukupno 1847 ovih slučajeva. Ispitivana je distribucija za svaku knjigu zasebno, što odgovara konceptu stratifikacije, i za sve knjige zajedno. Analiza je pokazala da distribucija ovih riječi po knjigama nije jednaka.

Drugi primjer analize je analiza Kur'ana [55]. Posmatran je broj stihova po surama, kojih je ukupno 114. Predmet analize su bile prve cifre brojeva kojima se izražava broj stihova po surama. Dobijena je tabela 4.1.

Prva cifra	1	2	3	4	5	6	7	8	9
Broj sura	30	17	12	11	14	7	8	10	5

Table 4: Tabela 4.1. Incidencija znacajnih cifara broja stihova po surama (izvor : Abdul Majid Motahari, Benford's Law and the Quran, http://www.submission.org/miracle/benford.html)

Drugim riječima, postoji 30 sura za koju broj stihova počinje značajnom ciform 1. Analogno se interpretiraju i ostali rezultati. Grafikon, koji se može napraviti na osnovu ove distribucije pokazuje veoma visok stepen saglasnosti sa Benfordovim zakonom.

Korištenje prvog slova je jedna od bitnih lingvističkih karakteristika. Benfordov zakon je razbio uvjerenje da se svaka cifra na svakoj poziciji riječi javlja sa jednakom vjerovatnoćom. Kad je u pitanju jezik, lako je pretpostaviti da šanse pojave pojedinog slova na prvoj poziciji nisu jednake. Dovoljno je uzeti riječnik i ustanoviti koji broj korijenskih riječi počinje nekim slovom. Izvjesno je da veoma mali broj riječi u engleskom jeziku počinje npr. slovima X, Y ili Z. Frekvencija slova na prvoj poziciji ne treba se miješati sa opštom frekvencijom slova na nivou jezika, koja je važan element u kriptoanalizi.

Kao osnova za analizu teksta u smislu frekvencije prvih slova riječi, po analogiji korištenja prvih cifara u brojevima, od strane autora ovog rada uzet je roman "Derviš i smrt" autora Meše Selimovića. Autor je na prvim pozicijama koristio 25 od raspoloživih 30 slova. Dobijene frekvencije su poredane opadajućim redom i proračunate teorijske frekvencije po bazi $B = 26$. Pritom je redni broj slova u tako dobijenom sortnom redoslijedu uziman kao 'cifra' u bazi $B = 26$. Uporedni dijagram ovako dobijenih teorijskih i uzoračkih frekvencija, dat na grafu 4.1, pokazao je primjetne razlike u korištenju nekih slova. Naravno, odstupanja od teorijskih frekvencija u ovom slučaju ne ukazuju na prevare već na jedno od jezičkih svojstava i navika autora.

Baza $B = 26$ je uzeta zbog nedostupnosti teorijskih frekvencija za prva slova, po uzoru na frekvencije vodećih cifara po bazi 10. Frekvencije se u tom slučaju računaju po formuli za drugu bazu:

$$P\{D = d/B\} = \log_B\left(1 + \frac{1}{d}\right) = \frac{\log_{10}(1 + 1/d)}{\log_{10} B}$$

Drugi mogući pristup je da se na neki način formiraju teorijske frekvencije. Mada se na prvi pogled ovaj posao čini težak, teorijske frekvencije je moguće dobiti prebrajanjem riječi po pojedinim slovima u odabranom riječniku. Na ovaj način se izbjegava pristrasnost i dobijaju frekvencije koje mogu poslužiti kao adekvatna empirijska osnova.

Ovakav pristup donosi bitnu prednost jer je izbor riječnika, u suštini, izbor odgovarajuće terminologije tzv. podjezika jer je poznato da nije moguće sve karakateristike podjezika prenijeti na korijenski jezik. Analiza ovog tipa se može praviti i za slogove.

4.2.5 Data mining

Data mining tehnike se koriste kako bi se iz velikih količina podataka dobile informacije koje nisu vidljive na prvi pogled, bilo da se radi o pravilima ponašanja, potrebi da se izvrši klasifikacija ili slično. Za te potrebe su razvijene brojne tehnike analiza koje se stalno dopunjuju i usavršavaju. Nova znanja i informacije se traže iz skupova za koja ne postoje unaprijed poznata pravila,

Benfordov zakon daje globalne ocjene i numeričke karakteristike skupa podataka. Iz tog razloga ga neki analitičari posmatraju kao jednu od tehnika data mininga. Međutim, ovakva kvalifikacija nije u potpunosti prihvatljiva. Ipak, karakteristika globalne analize je prirodan motiv za istraživanje mogućnosti primjene Benfordovog zakona u metodatama data mininga.

Mogući aspekti istraživanja primjene Benfordovog zakona u metodatama data mininga su :

- Da li i u kojoj mjeri saglasnost skupa sa Benfordovim zakonom ima uticaja na efikasnost, brzinu i stepen optimalnosti dobijenih rješenja
- Da li i na koji način veličine dobijene na osnovu Benfordovog zakona mogu biti korištene kao parametri u postojećim metodama i tehnikama data mininga

U smislu ovih aspekata mogući načini korištenja ovog zakona u metodama data mininga su :

- **Preprocesiranje**. Na početku svake analize numeričkih podataka uobičajeno je da se traže statistička svojstva skupova kao što je prosjek, standardna devijacija, asimetrija, mod, medijana i slično. Ove veličine analitičaru mogu biti jedan od pokazatelja globalnih karakteristika odnosno globalnu sliku skupa u skladu sa kojima može donijeti odluku o metodama koje treba koristiti. Za razliku od standardnih statističkih pokazatelja, veličine dobijene na osnovu Benfordovog zakona su dio specijalističkih alata
- **Anomalije (outliers)**. Ovo je najočiglednija praktična primjena, proistekla iz činjenice da Benfordov zakon stipulira distribuciju koja je nezavisna od prirode izvora numeričke veličine
- **Mašinsko učenje**. Jedan mogući način primjene u mašinskom učenju prezentiran je u ovom tekstu. Primjenu je moguće proširiti i na druge tipove problema mašinskog učenja
- **Problemi distanci**. Prema nekim istraživanjima, struktura distanci ima uticaj na brzinu i efikasnost algoritama u kojima se koriste. Kao paradigma može poslužiti problem trgovačkog putnika u kojem je potrebno naći optimalnu putanju obilaska skupa tačaka. U tekstu [10] se tvrdi da su problemi trgovačkog putnika brže rješivi što je veći stepen saglasnosti mjernih brojeva distanica sa Benfordovim zakonom

- **Neuronske mreže**. Dostupni izvori pokazuju da postoji interes da se putem neuronskih mreža vrši klasifikacija na osnovu parametara izvedenih iz Benfordove distribucije
- **Vremenske serije**. Benfordov zakon se može primijeniti na probleme koje imaju vremensku dimenziju, što je vidljivo iz činjenice da rekurentne serije, kao što su Fibonačijevi brojevi, prirodni procesi rasta (npr. populacija) i slični prirodni fenomeni slijede Benfordov zakon

Slučajevi u kojima Benfordov zakon nije moguće koristiti su :

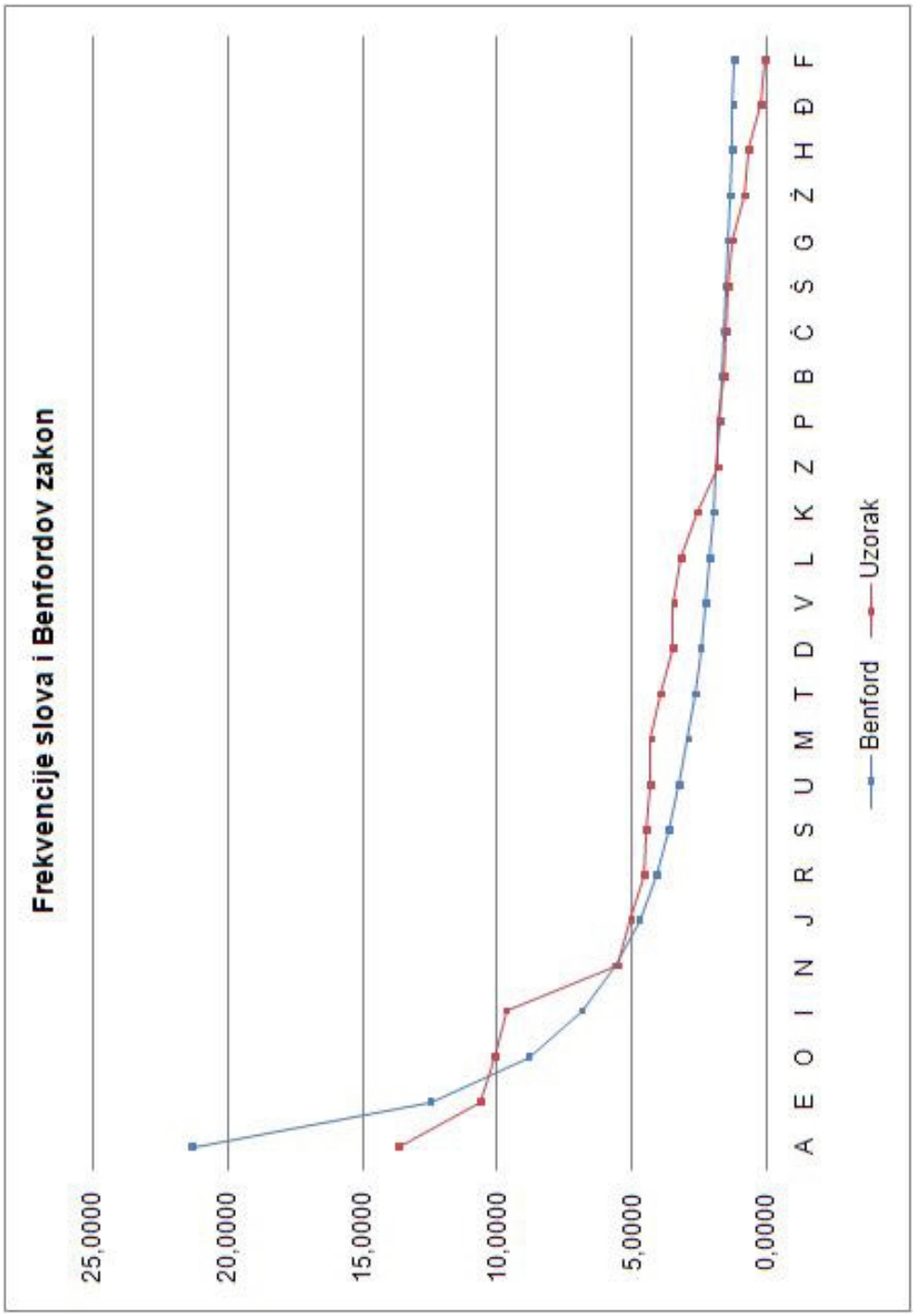
- Identifikacija ili procjena nedostajućih vrijednosti numeričkog atributa. Ako uzorak ima više dimenzija i ako po bilo kojoj dimenziji postoje nedostajući podaci njihova procjena ne može biti urađena putem Benfordovog zakona
- Analiza cifara ako su podaci normalizovani postupcima koji uključuju oduzimanje i sabiranje. To znači da prije analize putem ovog zakona ne bi trebalo raditi npr. minimaksnu normalizaciju ili normalizaciju putem standardne devijacije
- Analiza cifara u uslovima kada je dio elemenata uzorka predmet množenja / dijeljenja istim brojem. Benfordova distribucija je mjerno invarijantna pod uslovom da je cijeli skup predmet množenja / dijeljenja jednim istim brojem. Iz tog razloga npr. finansijski podaci moraju biti iskazani u samo jednoj novčanoj jedinici, podaci o mjerenju u samo jednom mjernom sistemu i slično

4.3 Zaključak

Gotovo svakodnevno se može naći novi primjer ljudske djelatnosti, prirodnih i drugih pojava u kojima Benfordov zakon može naći svoju primjenu. U tekstu [16] autor je pokazao mogućnost primjene Benfordovog zakona za analizu genske strukture i molekularnoj biologiji. Takođe, predmet analize mogu biti veličine ptičijih i životinjskih grupa (jata), veličine ledenih bregova, broj nastradalih u katastrofama (zemljotres, poplava, požar, ...), veličine grupa koje ulaze u restorane, veličine transakcija na berzama, veličine procesa na računarima, veličine naftnih i uljnih mrlja, veličine zahtjeva za osiguranja i slično.

Ovaj zakon je veoma dobra numerička tehnika koja sve više nalazi mjesta u analizama velikih skupova podataka. Dugo vremena je posmatran kao zanimljiv numerički kuriozitet bez posebnih praktičnih vrijednosti. Ipak, sve više se uviđaju mogućnosti njegove praktične primjene. Tome u prilog ide činjenica da je na stranici www.patents.com u 2008. i 2009. godini prijavljeno devet patenata koji se direktno odnose na Benfordov zakon u raznim segmenatima nauke i prakse. Neki od tih patenata su izvor za ovaj tekst.

4.4 Prilog



Grafikon 4.1. Frekvencije prvih slova u romanu Derviš i smrt

5 Stepen anomalije

5.1 Uvod

Theodor Hill je napravio eksperiment u kojem je zamolio 742 studenta matematike da zamisle šestocifrene slučajne brojeve i zapišu ih na komad papira. Skupio je odgovore i ustanovio da ovi brojevi ne slijede Benfordov zakon [6]. Busta i Weinberg su, u tekstu u kojem su obrađivali način klasifikacije podataka korištenjem neuronskih mreža na osnovu Benfordovog zakona, pošli od pretpostavke da bilo kakvi namješteni skupovi slijede Hillovu distribuciju s obzirom na pretpostavljeni sličan kognitivni proces [7]. Ista tzv. Hillova distribucija, korištena je i drugom tekstu koji se bavi korištenjem Benfordovog zakona u neuronskim mrežama. U oba ova teksta korišten je termin 'nivo kontaminacije' (contamination level).

Osim činjenice da nije empirijski i teorijski podržana, bitan nedostatak izbora Hillove distribucije je u činjenici da je kontaminacija simulirana odnosno unijeta u sam skup. Drugim riječima, uzorak je 'zasijan' simuliranim podacima. S obzirom da nije moguće naći egzaktnu formulaciju tzv. Hillove distribucije pretpostavka je da se radi o uzorku iz uniformne raspodjele. Pored odsustva teorijske podržanosti, ovaj pristup ne uzima u obzir mogući stepen anomalije samog uzorka odnosno polazi se od pretpostavke da uzorak obavezno slijedi Benfordovu distribuciju. Ovo je moguće za uzorke koji su simulirani ali je ovakav pristup upitan kada se radi sa stvarnim uzorcima.

Značajnost odstupanja uzoračkih frekvencija se procjenjuje testovima u kojima se koriste Intervali povjerenja. Kvalifikacija značajnosti odstupanja ne daje odgovor na pitanje koliki dio uzorka je van intervala povjerenja. S obzirom da se odstupanja koja su unutar intervala povjerenja uzimaju kao statistički prihvatljiva, dio van intervala povjerenja tu značajnost narušava.

Povećana frekvencija po jednoj cifri znači da je bar na jednoj od ostalih cifara ta frekvencija manja. Ako tako povećana frekvencija prelazi gornju granicu intervala povjerenja teoretski se može desiti da manje frekvencije po ostalim ciframa budu unutar intervala povjerenja, posebno kada je u pitanju relativno mali obim uzorka. Uzrok je u činjenici da u izračunu intervala povjerenja učestvuje standardna devijacija koja zavisi od obima uzorka.

Frekvencija koja prelazi gornju granicu, odnosno razlika između frekvencija uzorka i teorijske frekvencije koja odgovara gornjoj granici, očito predstavlja ono što se može zvati nivo anomalije, dio uzorka koji utiče na statističku značajnost odstupanja od intervala povjerenja. Isto vrijedi za slučaj da je uzoračka frekvencija manja od donje granice. Dio koji u tom slučaju nedostaje je očito sadržan u ostalom dijelu uzorka na način da su neke od frekvencija povećane.

Ako se ova odstupanja sumiraju po ciframa odnosno grupama cifara, npr. kada su u pitanju prve dvije cifre, može se proračunati stepen anomalije. Hipoteza od koje se polazi je :

H_0: Stepen anomalije se može proračunati na osnovu uzorka korištenjem Benfordovog zakona

Neka je :

- N : Obim uzorka
- $\overline{F}\{D_i = d\}$: uzoračka frekvencija za cifru d na poziciji i
- s_i : standardna devijacija za cifru na poziciji i
- $F_L\{D_i = d\} = F\{D_i = d\} - 1.96 \cdot N \cdot s_i$: teorijske frekvencije za donju granicu intervala povjerenja
- $F_U\{D_i = d\} = F\{D_i = d\} + 1.96 \cdot N \cdot s_i$: teorijske frekvencije za gornju granicu povjerenja

Ukupan obim frekvencija koje su van intervala povjerenja dat je sumom
$CL = \frac{1}{N}\sum_{i=1}^{9} F_{D_i}$
gdje je

$$F_{D_i} = \begin{cases} F\{D_i = d\} - F_U\{D_i = d\} & F_U\{D_i = d\} < F\{D_i = d\} \\ 0 & F_L\{D_i = d\} \leq F\{D_i = d\} \leq F_U\{D_i = d\} \\ F_L\{D_i = d\} - F\{D_i = d\} & F\{D_i = d\} < F_L\{D_i = d\} \end{cases} \quad (5.1)$$

Ovaj podatak je moguće uzimati kao stepen anomalije jer predstavlja dio uzorka koji izlazi iz okvira datog intervalom povjerenja. Veličina 1.96 se uzima kao kvantil za 95% nivo povjerenja z-testa. Za 90% nivo povjerenja ovaj kvantil je1.64. Korištenje ovog kvantila bi dalo nešto uži interval povjerenja što može biti korisno za uzorke manjeg obima.

5.2 Algoritam

Koraci u kojima se izračunava stepen anomalije su :

1. Računanje uzoračkih frekvencija
2. Računanje donje i gornje granice intervala povjerenja za svaku cifru
3. Računanje teorijskih frekvencija za donju i gornju granicu
4. Računanje razlika između teorijskih frekvencija donjih odnosno gornjih granica povjerenja za svaku cifru
5. Sumiranje razlika iz prethodnog koraka
6. Računanje stepena anomalije kao količnika suma iz prethodnog koraka i obima uzorka

5.3 Eksperiment

U cilju demonstracije ove metode uzet je uzorak obima $N = 10.190$ stavki. Korištenjem Excell programskog paketa napravljena je simulacija ovog algoritma. Rezultat je prikazan u tabelama 5.1 i 5.2.

Značenja kolona su sljedeća :

Cifre	DG	BZ	GG	UZ	sd(d)
1	0, 2923	0, 3010	0, 3097	0, 2776	0, 00444
2	0, 1685	0, 1761	0, 1836	0, 1860	0, 00385
3	0, 1184	0, 1249	0, 1314	0, 1304	0, 00334
4	0, 0909	0, 0969	0, 1028	0, 1064	0, 00305
5	0, 0729	0, 0792	0, 0854	0, 1186	0, 00320
6	0, 0626	0, 0669	0, 0712	0, 0527	0, 00221
7	0, 0538	0, 0580	0, 0620	0, 0467	0, 00209
8	0, 0474	0, 0512	0, 0548	0, 0385	0, 00191
9	0, 0418	0, 0458	0, 0497	0, 0431	0, 00201

Table 5: Tabela 5.1. Rezultati simulacije algoritma. Date su relativne frekvencije

Cifre	DG_Fr	UZ_Fr	GG_Fr	Razlike_Fr	Procenti
1	2978, 392	2829	3155, 600	149, 392	1, 466
2	1716, 889	1895	1870, 851	24, 149	0, 237
3	1205, 995	1329	1339, 256	0, 000	0, 000
4	926, 011	1084	1048, 016	35, 984	0, 353
5	742, 377	1209	870, 337	338, 663	3, 323
6	637, 481	537	725, 894	100, 481	0, 986
7	548, 686	476	632, 189	72, 686	0, 713
8	482, 692	392	558, 796	90, 692	0, 890
9	425, 597	439	505, 941	0, 000	0, 000
Sume	9664, 120	10190	10706, 880	812, 048	7, 970

Table 6: Tabela 5.2. Kalkulacija za stepen kontaminacije. Date su uzoracke i teorijske frekvencije

- DG : Donja granica relativnih frekvencija odosno intervala povjerenja
- BZ : Relativne frekvencije u skladu sa Benfordovim zakonom
- GG : Gornja granica relativnih frekvencija odosno intervala povjerenja
- UZ : Uzoračke relativne frekvencije
- sd(d) : Standardna devijacija za cifru d
- DG_Fr : Donja granica frekvencija : $DG\_Fr = N \cdot DG$
- UZ_Fr : Uzoračke frekvencije
- GG_Fr : Gornja granica frekvencija : $GG\_Fr = N \cdot GG$
- Razlike_Fr : Razlike proračunate prema formuli (4.2.1)
- Procenti : Procenti odstupanja po pozicijama : $Procenti = \ Razlike\_Fr/N$

Dijagram 5.1 je ilustracija frekvencija i intervala povjerenja za prvu cifru u odbaranom uzorku. Ista kalkulacija je moguća za drugu ili treću cifru odnosno prve dvije ili tri cifre. Prema ovoj kalkulaciji, ukupno 812 stavki odnosno 7.97% ovog uzorka je van dobijenog intervala povjerenja. Ovaj podatak se može uzeti kao stepen anomalije uzorka. Rezultat

se može tumačiti na način da je iz uzorka moguće izdvojiti 812 stavki koje su uzrok da uzorak ne slijedi Benfordov zakon. Drugim riječima, ako se tokom analize utvrdi broj koji je bitno drugačiji od ovog broja postoji osnova za preispitivanje postupka analize. S obzirom da ovaj zakon nije orijentisan na pojedinačne stavke već daje generalnu ocjenu uzorka, test ne ukazuje na pojedinačne stavke. Njih treba tražiti drugim metodama.

Ista kalkulacija se može uraditi kada se provede test drugog reda. Radi podsjećanja, ovaj test se izvodi na način da se podaci sortiraju rastućim / opadajućim redoslijedom, izračunaju razlike između dvije uzastopne stavke i analize provode na ovako izračunatim razlikama.

Test drugog reda i kalkulacija na istom uzroku daje rezultate na tabelama 5.3 i 5.4. Testom drugog reda uzorak od $N = 10.190$ stavki sveden je na uzorak obima $N = 1.145$ stavki.

Cifre	DG	BZ	GG	UZ	st(d)
1	0, 2724	0, 3010	0, 3297	0, 4279	0, 0146
2	0, 1531	0, 1761	0, 1991	0, 1956	0, 0117
3	0, 1066	0, 1249	0, 1433	0, 1127	0, 0093
4	0, 0809	0, 0969	0, 1130	0, 0838	0, 0082
5	0, 0658	0, 0792	0, 0926	0, 0568	0, 0068
6	0, 0549	0, 0669	0, 0790	0, 0454	0, 0062
7	0, 0480	0, 0580	0, 0680	0, 0306	0, 0051
8	0, 0419	0, 0512	0, 0604	0, 0262	0, 0047
9	0, 0375	0, 0458	0, 0541	0, 0210	0, 0042

Table 7: Tabela 5.3. Relativne frekvencije u testu drugog reda za uzorak iz tabele 5.1

Cifre	DG_Fr	UZ_Fr	GG_Fr	Razlike_Fr	Procenti
1	311, 8644	490	377, 4943	0, 0000	0, 000
2	175, 3153	224	227, 9337	3, 9337	0, 344
3	122, 0850	129	164, 0247	35, 0247	3, 059
4	92, 5806	96	129, 3433	33, 3433	2, 912
5	75, 3156	65	106, 0095	41, 0095	3, 582
6	62, 8450	52	90, 4632	38, 4632	3, 359
7	54, 9839	35	77, 8177	42, 8177	3, 740
8	47, 9758	30	69, 1634	39, 1634	3, 420
9	42, 8915	24	61, 8932	37, 8932	3, 309
Sume	985, 8571	1145	1.304, 1429	271, 6486	23, 720

Table 8: Tabela 5.4 Kalkulacija teorijskih i uzorackih frekvencija za stepen kontaminacije

Dijagram 5.2 je ilustracija ovih frekvencija i intervala povjerenja za test drugog reda. Porast anomalije na preko 23% ukazuje na nepravilnosti čija priroda treba biti istražena u kontekstu njihovog nastanka i/ili njihove funkcije.

Korištenjem Excell-a modelirana su dva ekstremna slučaja koja se mogu desiti i na kojima će biti dato obrazloženje za ovu vrstu kalkulacije. U prvom slučaju frekvencija samo jedne cifre je iznad gornje granice intervala povjerenja. Kalkulacija je u tabelama 5.5 i 5.6.

Cifre	DG	BZ	GG	UZ	StDev
1	0, 2761	0, 3010	0, 3259	0, 3536	0, 0127
2	0, 1566	0, 1761	0, 1956	0, 1676	0, 0099
3	0, 1082	0, 1249	0, 1417	0, 1167	0, 0085
4	0, 0821	0, 0969	0, 1118	0, 0891	0, 0076
5	0, 0656	0, 0792	0, 0927	0, 0728	0, 0069
6	0, 0545	0, 0669	0, 0794	0, 0608	0, 0064
7	0, 0464	0, 0580	0, 0696	0, 0523	0, 0059
8	0, 0402	0, 0512	0, 0621	0, 0460	0, 0056
9	0, 0354	0, 0458	0, 0561	0, 0410	0, 0053

Table 9: Tabela 5.5. Relativne frekvencije za slucaj frekvencije iznad intervala povjerenja po jednoj cifri

Cifre	DG_Fr	Uz_Fr	GG_Fr	Razlike_Fr	Procenti
1	390, 4202	500	460, 8926	39, 1074	2, 766
2	221, 4638	237	276, 5223	0, 0000	0, 000
3	153, 0012	165	200, 3255	0, 0000	0, 000
4	116, 0329	126	158, 0286	0, 0000	0, 000
5	92, 8086	103	131, 1159	0, 0000	0, 000
6	77, 0479	86	112, 2776	0, 0000	0, 000
7	65, 5872	74	98, 4141	0, 0000	0, 000
8	56, 8951	65	87, 7642	0, 0000	0, 000
9	50, 0835	58	79, 3187	0, 0000	0, 000
Sume	1.223, 3405	1414	1.604, 6595	39, 1074	2, 766

Table 10: Tabela 5.6. Teorijske i uzoracke frekvencije za slucaj frekvencije iznad intervala povjerenja po jednoj cifri

Dijagram 5.3 je ilustracija ovih frekvencija i intervala povjerenja. Jedina frekvencija koja je van odnosno iznad gornje granice intervala povjerenja je ona za cifru 1. Prema kalkulaciji, ukupno je 39 ovakvih stavki, što čini 2.76% uzorka. Frekvencije ostalih cifara su ispod teorijskih.

U drugom slučaju, frekvencije samo jedne cifre su ispod donje granice intervala povjerenja. Kalkulacija je u tabelama 5.7 i 5.8.

Dijagram 5.4 je ilustracija ovih frekvencija i intervala povjerenja. Jedina frekvencija koja je van odnosno ispod donje granice intervala povjerenja je ona za cifru 2. Prema kalkulaciji, ukupno je 44 ovakvih stavki, što čini 3.14% uzorka. Smanjena frekvencija za cifru 2 je rezultat odstupanja frekvencija ostalih cifara od teorijskih. Iako unutar granica intervala povjerenja, ova odstupanja imaju bitan uticaj na frekvenciju cifre 2. Drugim riječima, uzorak bi trebao biti detaljnije istražen u smislu razloga za ovakvu distribuciju frekvencija a istraga bi trebala pokazati da je ovo minimalan broj stavki koje uzrokuju ovakva odstupanja.

Cifre	DG	BZ	GG	UZ	StDev
1	0, 2772	0, 3010	0, 3248	0, 2970	0, 0122
2	0, 1587	0, 1761	0, 1935	0, 1273	0, 0089
3	0, 1070	0, 1249	0, 1429	0, 1379	0, 0092
4	0, 0807	0, 0969	0, 1131	0, 1082	0, 0083
5	0, 0644	0, 0792	0, 0940	0, 0884	0, 0075
6	0, 0530	0, 0669	0, 0809	0, 0778	0, 0071
7	0, 0462	0, 0580	0, 0697	0, 0537	0, 0060
8	0, 0388	0, 0512	0, 0635	0, 0601	0, 0063
9	0, 0345	0, 0458	0, 0571	0, 0495	0, 0058

Table 11: Tabela 5.7. Relativne frekvencije za slucaj kada su frekvencije po jednom atributu ispod inveravala povjerenja po jednoj cifri

Cifre	DG_Fr	UZ_Fr	GG_Fr	Raz_Fr	Procenti
1	391, 9782	420	459, 3346	0, 0000	0, 000
2	224, 4276	180	273, 5585	44, 4276	3, 142
3	151, 2507	195	202, 0761	0, 0000	0, 000
4	114, 1361	153	159, 9254	0, 0000	0, 000
5	91, 0398	125	132, 8847	0, 0000	0, 000
6	74, 9219	110	114, 4036	0, 0000	0, 000
7	65, 3793	76	98, 6220	0, 0000	0, 000
8	54, 8109	85	89, 8484	0, 0000	0, 000
9	48, 7136	70	80, 6886	0, 0000	0, 000
Sume	1.216, 6580	1414	1.611, 3420	44, 4276	3, 142

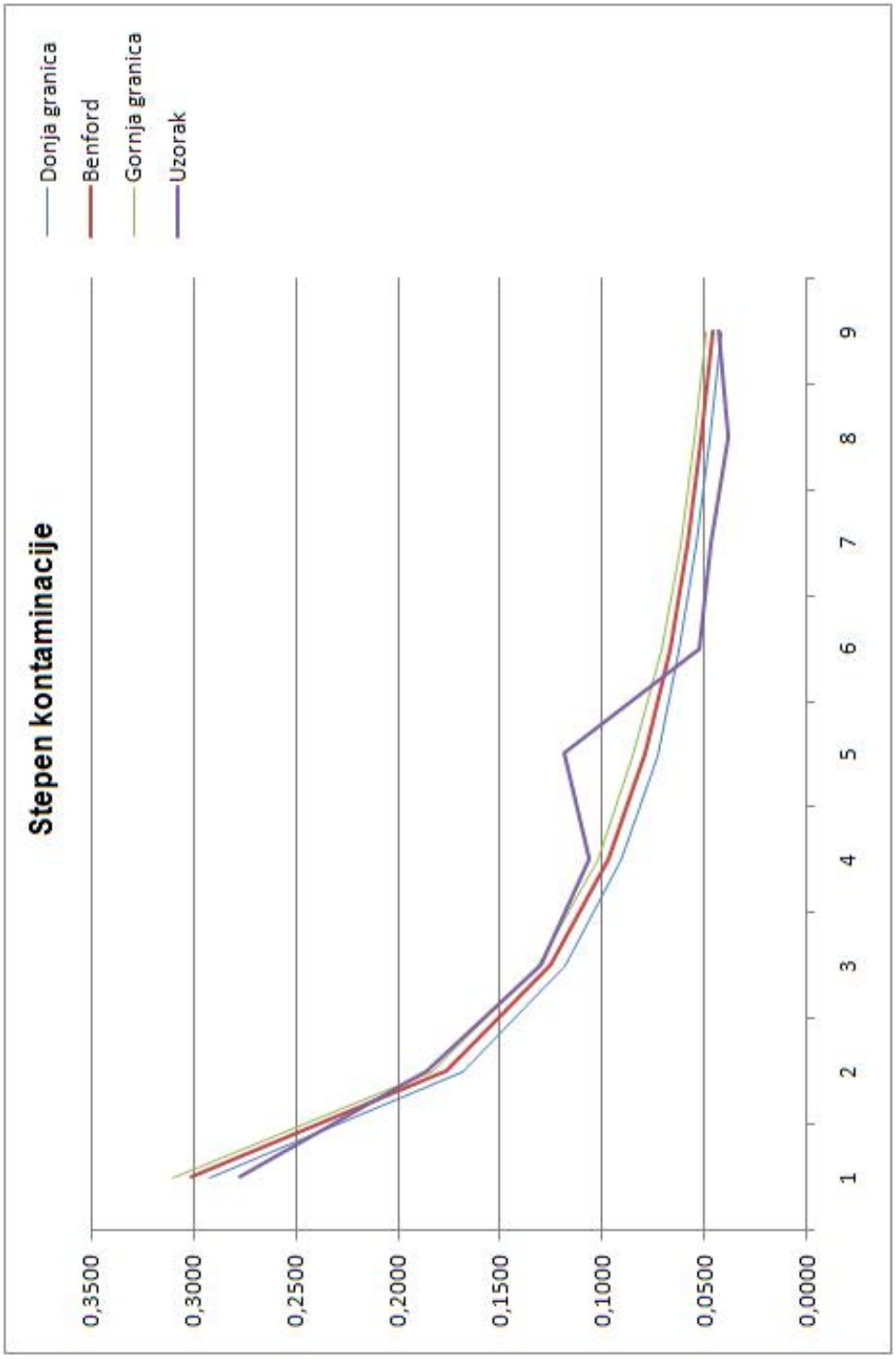
Table 12: Tabela 5.8. Teorijske i uzoracke frekvencije za slucaj kada su frekvencije po jednoj cifri ispod intervala povjerenja

5.4 Diskusija

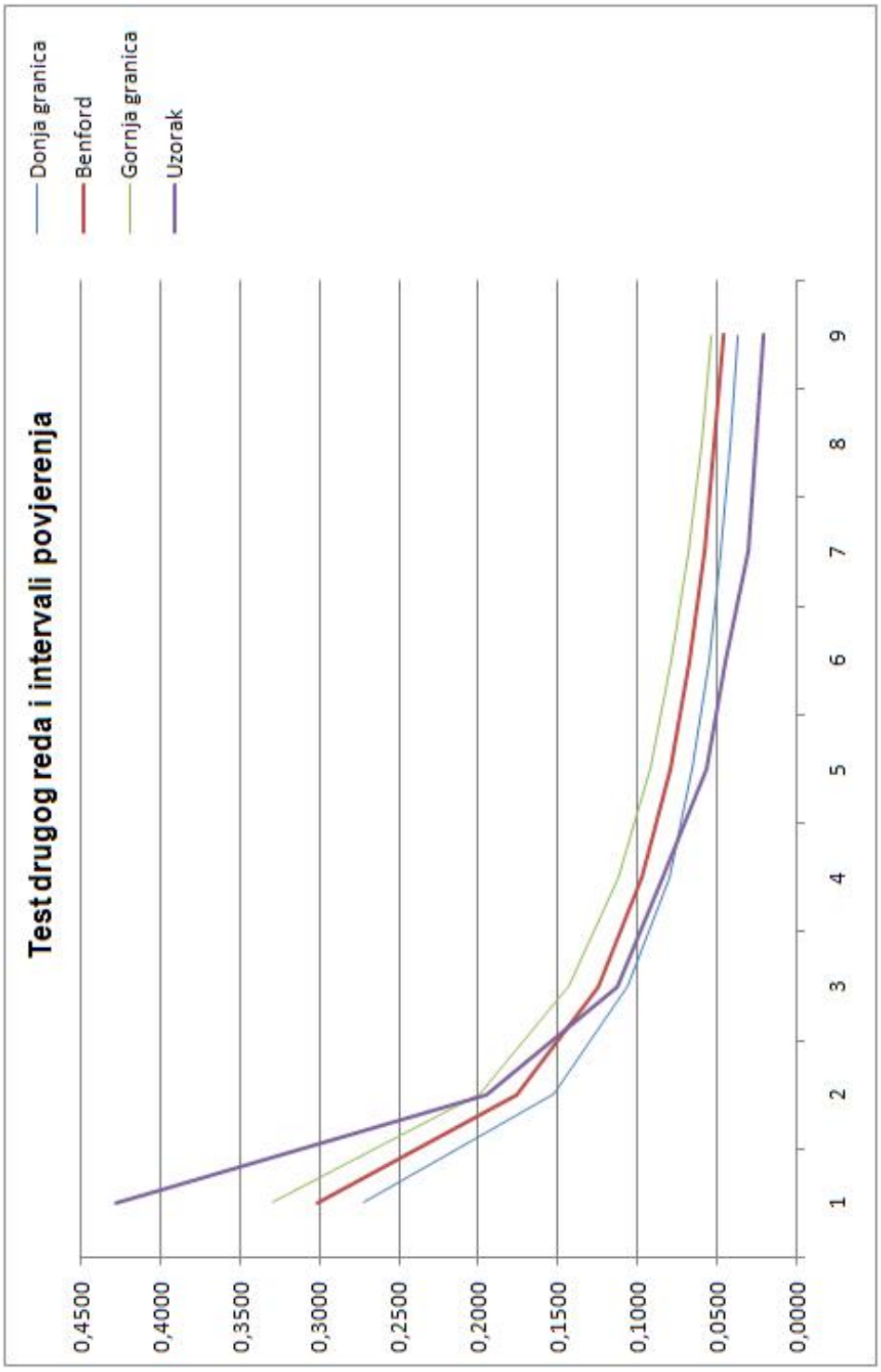
U ovom poglavlju je pokazano da je procjena stepena kontaminacije moguća na osnovu datog uzorka. Razlog leži u diskretnoj prirodi ovog zakona u kojem se kalkulacije provode na konačnom broju cifara odnosno grupa cifara. Očigledno je da ova kalkulacija ne može biti provedena ako se ne raspolaže podatkom o obimu uzorka.

Sa praktičnog stanovišta, ovo je korisno za analitičara koji na ovaj način može praviti barem okvirne procjene mogućeg obima odstupanja od osnovnog skupa.

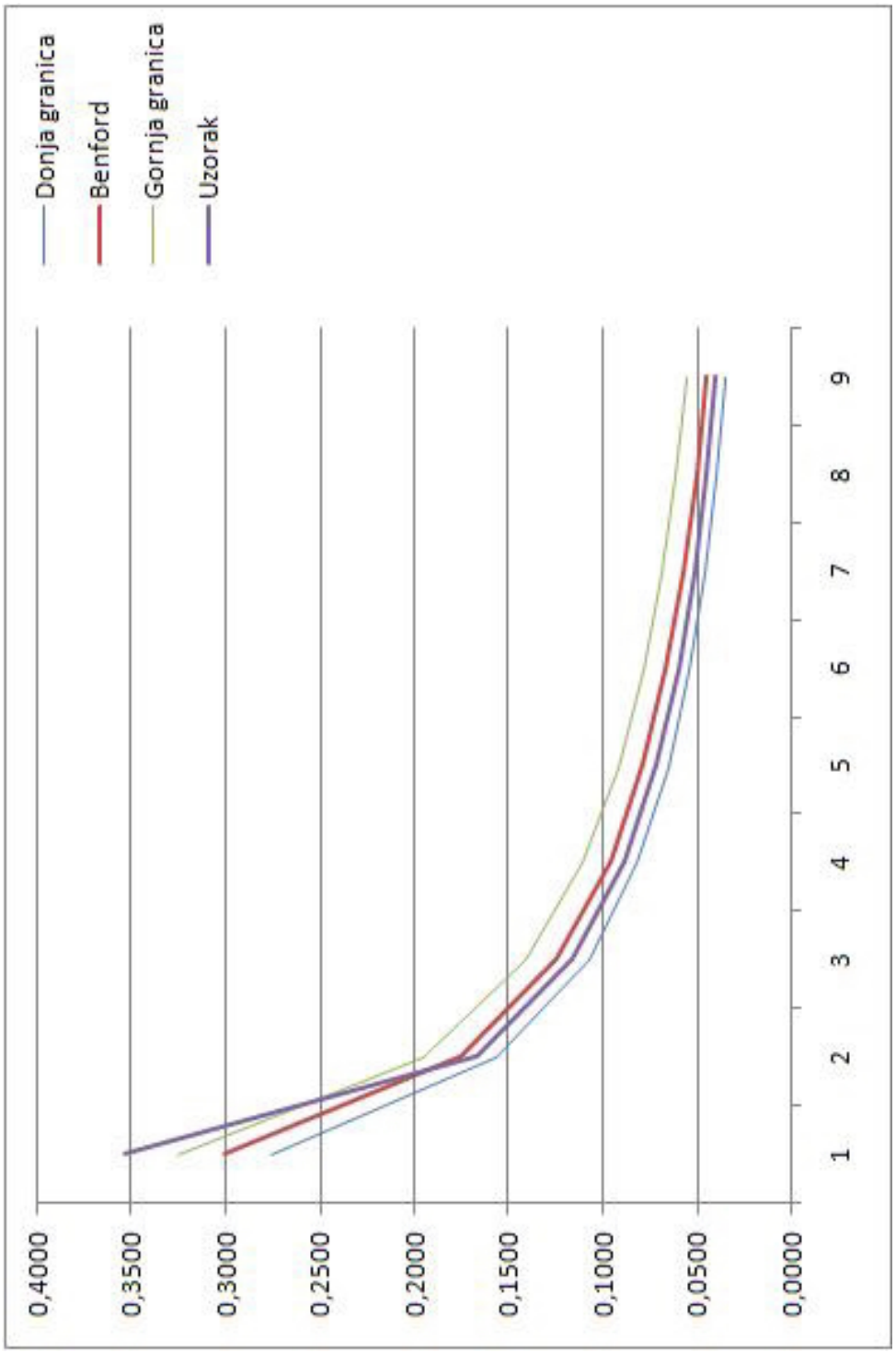
5.5 Prilog



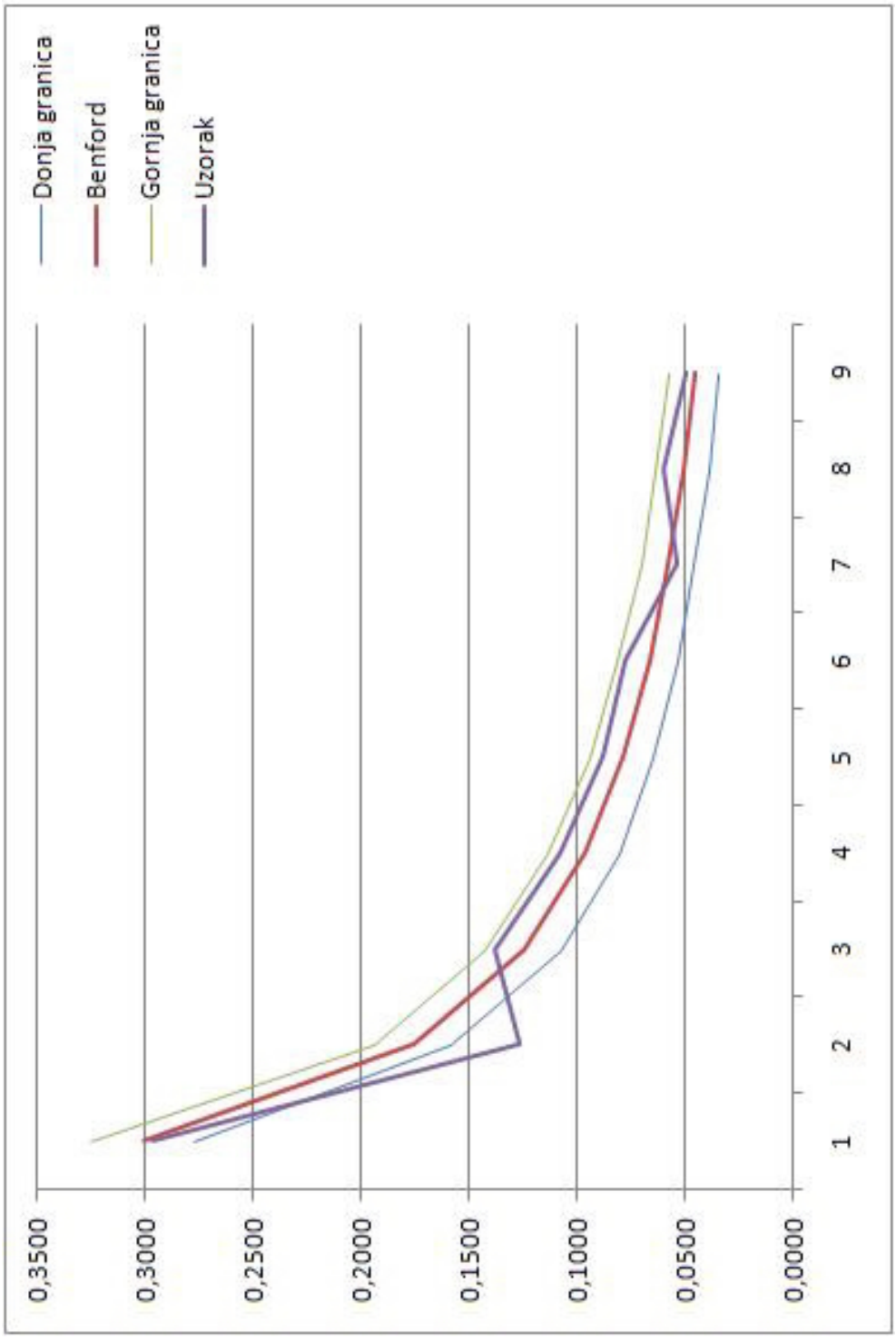
Slika 5.1. Frekvencije i intervali povjerenja za prvu cifru odabranog uzorka



Slika 5.2. Grafikon nakon testa drugog reda. Primjetno je neslaganje uzorka sa intervalima povjerenja i veliki porast kontaminacije



Slika 5.3. Relativne frekvencije za slučaj kada su frekvencije po jednom atributu iznad intervala povjerenja



Slika 5.4. Relativne frekvencije za slučaj kada su frekvencije po jednom atributu ispod intervala povjerenja

6 Adaptivna Benfordova metoda

6.1 Uvod

Benfordov zakon specificira distribuciju vjerovatnoća vodećih cifara na nivou cijelog skupa [5]. Jedan od zahtjeva za njegovu primjenu je da se ne postavljaju ograničenja u pogledu minimuma ili maksimuma. Kada su podaci obuhvaćeni samo djelimično to ne mora neophodno značiti da podaci ne slijede Benfordovu distribuciju. Umjesto toga, to jedino znači da predmet analize nije kompletan skup. Kad su u pitanju finansijski podaci ovo je pojava koja se može desiti npr. kada je u pitanju samo jedan mjesec ili period, pokušaj skrivanja velikih transakcija ili malih često ponavljanih iznosa koji samom svojom frekvencijom svraćaju pozornost i slično. U aktivnostima pranja novca prevaranti često koriste iznose neposredno ispod zakonski utvrđenog praga. Stoga nekompletnost u smislu da je odbačen dio podataka iznad ili ispod nekog praga koji analitičaru, u pravilu, nije poznat, predstavlja poseban problem u analizi cifara putem ovog zakona. Unatoč tome, još uvijek je poželjno primijeniti Benfordov zakon za analizu cifara kako bi se našle anomalije osim činjenice da podaci nedostaju.

Problem sa tradicionalnim Benfordovim zakonom u radu sa nekompletnim podacima je da frekvencije nekih cifara postaju naglašene prilikom računanja vjerovatnoća. Naprimjer, Benfordov zakon tvrdi da bi se cifra 4 u skupu podataka trebala pojavljivati sa vjerovatnoćom $\log_{10}(1 + 1/4) = 0.096\,91$. Ako se pretpostavi da se u kompletnom skupu od 100 opservacija cifra 4 na prvom mjestu pojavljuje 10 puta to približno aproksimira Benfordovu vjerovatnoću. Međutim, ako je skup podataka nekompletan i ima samo 50 evidentiranih slogova a cifra 4 se pojavila također 10 puta dobija se $10/50 = 0.20$, što bitno povećava vjerovatnoću cifara koje su evidentirane više puta zbog cifara koje nedostaju i nisu uključene u računanje vjerovatnoće.

Efrim Boritz i Fletcher Lu su patentirali metodu, poznatu kao Adaptivna Benfordova metoda, koja omogućava analizu putem Benfordovog zakona i na uzorcima za koje je, na osnovu strukture i drugih faktora, poznato da nedostaju ne zadovoljavaju kriterije za Benfordov zakon [5]. Ona usklađuje distribuciju frekvencija cifara kako bi se uzele u obzir bilo kakve nedostajuće granice i producirale vrijednosti pragova za različite rangove cifara. Zatim se ove vrijednosti koriste kako bi se analizirali testni podaci. Na taj način se mogu dobiti cifre ili grupe cifara koje izlaze iz dobijenog okvira vrijednosti.

Hipoteza od koje se polazi je :

H_0 : Izbor donjih i gornjih granica uzorka ima uticaj na rezultat analize cifara putem Adaptivne Benfordove metode

6.2 Algoritam Adaptivne Benfordove metode

Ako postoji uvjerenje da evidentirani podaci slijede Benfordovu distribuciju, da je ona graničena i ako nedostaju podaci ispod ili iznad nekog praga to saznanje se može koristiti za procjenu saglasnosti sa Benfordovom distribucijom [5]. Prvo, neka je :

- s : vodeća sekvenca dužine i. Sekvenca je u ovom slučaju niz vodećih cifara u dužini od jedne ili više vodećih cifara. U praksi je uobičajeno da se uzima najviše tri vodeće cifre

- $\overline{F}\{S_i = s\}$: uzoračka frekvencija sekvence s dužine i u posmatranom uzorku
- $P\{S_i = s\}$: Benfordova vjerovatnoća sekvence S_i, gdje je S_i bilo koja sekvenca dužine i (odnosno S_i je slučajna varijabla)
- $F\{S_i = s\}$: frekvencija sekvence s dužine i u posmatranom uzorku koja odgovara Benfordovoj vjerovatnoći

Ako je potrebno da se uzoračka frekvencija vodeće sekvence s poredi sa Benfordovom vjerovatnoćom računa se relativna frekvencija u funkciji uzoračke vjerovatnoće kao količnik

$$\frac{\overline{F}\{S_i = s\}}{\sum_{S_i} \overline{F}\{S_i = s\}} \simeq P\{S_i = s\} \qquad (6.1)$$

Imenilac je sumiran za sve sekvence cifara iste dužine i kao i sekvenca s. Neka je

$$C_i = \sum_{S_i} \overline{F}\{S_i = s\}$$

Jednadžba (6.1) sada može biti rearanžirana u sljedeći oblik

$$\frac{\overline{F}\{S_i = s\}}{P\{S_i = s\}} \simeq \sum_{S_i} \overline{F}\{S_i = s\} = C_i \qquad (6.2)$$

Iz jednadžbe (6.2) se vidi da je veličina C_i, u suštini, konstantni faktor skale za sve sekvence iste dužine i. Ako u podacima nedostaju podaci zbog prisustva pragova veličina C_i se može računati korištenjem uzoračkih frekvencija s obzirom da bi ove sekvence koje su stvarno prisutne trebale slijediti vjerovatnoće prema Benfordovom zakonu.

Da bi se mogla napraviti najbolja procjena nedostajućih podataka računa se srednja vrijednost (prosjek) svih mogućih C_i za date sekvence dužine i. Naprimjer, C_2 bi bio prosječni faktor skaliranja za sve frekvencije na prve dvije pozicije. Stoga, neka je

$$C_i = \frac{\overline{F}\{S_i = s\}/P\{S_i = s\}}{\#\,|sekvence\_dužine\_i|} = \frac{1}{\#\,|sekvence\_dužine\_i|} \cdot \sum_{S_i} \frac{\overline{F}\{S_i = s\}}{P\{S_i = s\}}$$

Faktor skaliranja C_i se koristi za 'popunjavanje' nedostajućih podataka Benfordove distribucije. Imenitelj predstavlja broj svih sekvenci dužine i koji se koristi u kalkulacijama. Veličina C_i se koristi za množenje Benfordovske vjerovatnoće za sekvence cifara dužine i što se zatim koristi za poređenje sa uzoračkim frekvencijama iz podataka.

Glavni koraci ovog algoritma su :

1. Računanje vrijednosti konstanti C_i za različite dužine sekvenci cifara

2. Računanje vještačke (izvedene) frekvencije za različite dužine sekvenci cifara

Vještačke (izvedene) Benfordovske frekvencije računaju se na sljedeći način

$$F\{S_i = s\} = C_i \times \log_{10}\left(1 + \frac{1}{d}\right) \qquad (6.3)$$

Stvarne frekvencije se simuliraju množenjem umjesto dijeljenjem sa sumom uzoračkih instanci koje bi trebale dati vjerovatnoće. Na ovaj način se izbjegava efekt bubrenja (inflating), pojave velikih frekvencija neke sekvence zbog toga što neke druge nedostaju.

Metoda je u patentnoj prijavi demonstrirana na uzorku poreskih prijava i na uzorku podataka zdravstvenog osiguranja. Uzorak poreskih podataka je bio predmet analize od strane Mark Nigrinija pa je napravljeno njihovo poređenje. Pritom su podaci uzimani iz raznih raspona koji su vidljivi u tabeli 6.1.

Rasponi vrijednosti	**Obim**	**Klasični Benford**	**Adaptivni Benford**
Cijeli uzorak	3141	96,2	94,9
100.000-1.100.000	431	85,0	89,4
200.000-1.200.000	220	85,6	96,9
300.000-1.300.000	144	49,5	94,4
400.000-1.400.000	106	30,8	91,7
500.000-1.500.000	84	32,0	87,2
600.000-1.600.000	65	34,0	84,0
700.000-1.700.000	48	33,8	82,4
800.000-1.800.000	36	19,2	82,7
900.000-1.900.000	24	11,8	73,5

Table 13: Tabela 6.1. Uporedna tabela analize frekvencija klasicnom i Adaptivnom Benfordovom metodom. Izvor : Efrim Boritz, Fletcher Lu, Method of Data Analysis, patentna prijava US 2008/0208946 A1, 28.08.2008

U ovoj tabeli kolona označena kao 'Klasični Benford' predstavlja procenat saglasnosti uzorka sa teorijskim frekvencijama koje se računaju na klasični način. Kolona označena kao 'Adaptivni Benford' označava procenat saglasnosti uzorka sa teorijskim frekvencijama prema adaptivnoj metodi. Ovakav izbor raspona je napravljen kako bi se demonstrirali efekti redukcije vodećih cifara. Tako, u rasponu 100.000 - 1.100.000 moguće su sve vodeće cifre od 1 do 9, dok je u rasponu 900.000 - 1.900.000 jedine moguće vodeće cifre su 9 i 1. Na ovoj tabeli se može uočiti korištenje Adaptivne Benfordove metode za male uzorke.

U nastavku je opšti opis algoritma adaptivne Benfordove metode. U algoritmu su objedinjeni svi navedeni koraci.

Neka je S posmatrani skup testnih podataka
Neka je U_i gornja granica broja standardnih devijacija
FOR sekvence $d = 1, 2, 3, ..., GG$
    $f_{d,uzorak}$ =broj koliko se puta sekvenca d pojavljuje u skupu S

FOR $i = 1...gg\_dužine\_sekvenci$

Neka je n_i broj vodećih cifara dužine i koji se u skupu S javljaju bar jednom

Za sve sekvence D_i dužine i računati:

$$\widetilde{C}_i = \frac{1}{n_i} \sum_{D_i} \frac{f_{D_i,uzorak}}{\log_{10}\left(1 + \frac{1}{D_i}\right)}$$

Za svaku sekvencu d dužine i računati:

$$f_{d,očekivanje} = \widetilde{C}_i \times \log_{10}\left(1 + \frac{1}{d}\right)$$

Za sve sekvence D_i dužine i računati:

$$\widetilde{\sigma}_i^2 = \frac{1}{n_i} \sum_{D_i} \left(f_{D_i,uzorak} - f_{d,očekivanje}\right)^2$$

Za svaku sekvencu d dužine i računati:

ako je $U_i < \dfrac{f_{D_i,uzorak} - f_{d,očekivanje}}{\sigma_i}$ tada je d anomalična

6.3 Kriteriji anomaličnosti

Broj standardnih devijacija daje gornju i donju granicu devijacija od vještačkih Benfordovskih frekvencija. Te granice se koriste da se utvrdi da li su odstupanja podataka iz uzorka dovoljna i potencijalno indikativna u smislu anomaličnosti. Nakon što se adaptivnom metodom napravi procjena uzoračkih frekvencija, anomaličnost podataka se provjerava u narednim koracima :

1. Računanje standardnih devijacija za svaku dužinu sekvenci. Računaju se devijacije za prvu cifru, prve dvije cifre i prve tri cifre
2. Procjena intervala povjerenja
3. Karakterizacija podataka koji odstupaju odnosno prelaze granicu broja standardnih devijacija od izvedenih Benfordovskih sekvenci

Broj standardnih devijacija, u oznaci u_i, u koraku 3. određuje interval povjerenja i uzima se tako da je $u_i \in [-2, +2]$.

Ako se sa u_i označi broj standardnih devijacija tada je

$$\begin{aligned} F\{S_i = s\} - u_i \cdot \sigma_i < \overline{F}\{S_i = s\} < F\{S_i = s\} + u_i \cdot \sigma_i \\ \Rightarrow \frac{\overline{F}\{S_i = s\} - F\{S_i = s\}}{\sigma_i} < u_i \\ -u_i < \frac{\overline{F}\{S_i = s\} - F\{S_i = s\}}{\sigma_i} \end{aligned}$$

Kriterij da sekvenca ima prihvatljivu frekvenciju je uslov

$$-u_i < \frac{\overline{F}\{S_i = s\} - F\{S_i = s\}}{\sigma_i} < u_i$$

Ovaj uslov se može napisati i na sljedeći način

$$\frac{\left|\overline{F}\left\{S_i = s\right\} - F\left\{S_i = s\right\}\right|}{\sigma_i} < u_i$$

Drugim riječima, ako vrijedi jedna od relacija

$$\frac{\overline{F}\left\{S_i = s\right\} - F\left\{S_i = s\right\}}{\sigma_i} > u_i$$
$$\frac{\overline{F}\left\{S_i = s\right\} - F\left\{S_i = s\right\}}{\sigma_i} < -u_i$$

sekvenca se kvalifikuje kao anomalična. Treba obratiti pažnju da se razlika u brojiocu računa tako da se teorijska frekvencija oduzima od uzoračke. Varijansa se, na osnovu uzoračkih podataka, računa na sljedeći način:

$$\overline{\sigma}_i^2 = \frac{1}{N_i}\sum_{D_i}\left(\overline{F}\left\{S_i = s\right\} - F\left\{S_i = s\right\}\right)^2 \qquad (6.4)$$
$$\sigma_i = \sqrt{\overline{\sigma}_i^2}$$

Ovdje je N_i ukupan broj svih sekvenci dužine i čije frekvencije u uzorku nisu nula.

6.4 Uzoračke veličine

Adaptivna Benfordova metoda daje mogućnost da se izračunavaju standardne veličine izvedene iz uzorka. Jedan od njih je srednja apsolutna devijacija (MAD). MAD se za prvu poziciju računa na sljedeći način:

$$MAD_1 = \frac{1}{9}\sum_{d=1}^{9}\left|\overline{P}\left\{D_1 = d\right\} - P\left\{D_1 = d\right\}\right|$$

a za prve dvije pozicije na sljedeći način :

$$MAD_2 = \frac{1}{90}\sum_{d_1 d_2=10}^{99}\left|\overline{P}\left\{D_1 D_2 = d_1 d_2\right\} - P\left\{D_1 D_2 = d_1 d_2\right\}\right|$$

Ove formule pretpostavljaju uzorke u kojima se mogu pojaviti sve cifre na prvoj odnosno prve dvije pozicije. Adaptivna Benfordova metoda uklanja ovaj nedostatak pa se u kalkulacijama može koristiti stvarni broj sekvenci cifara na prvoj, prvoj dvije i prve tri pozicije koje su se pojavile u uzorku. Drugim riječima, u gornje dvije formule umjesto imenilaca 9 i 90 može se uzimati stvarni broj cifara odnosno grupa cifara koje se javljaju. Ovo je korišteno u eksperimentima u kojima je vršeno poređenje dvije metode.

6.5 Eksperiment

Provedeni su eksperimenti kojima je ispitivan uticaj izbora donjih i gornjih granica na rezultate analiza putem Adaptivne Benfordove metode. Hipoteza koja se provjerava je :

H_0:Izbor donjih i gornjih granica uzorka ima uticaj na rezultat analize cifara putem Adaptivne Benfordove metode

Eksperimenti su provedeni na dva uzorka. Prvi uzorak je spisak datoteka sa laptopa autora teksta a drugi je primjer transakcija. Testovi su provedeni na način da se uzimaju razne vrijednosti gornjih i donjih granica i za svaki od tih slučajeva računaju vrijednosti odabranih globalnih parametara. Kalkulacije su izvedene u programskom paketu MS Excell.

6.5.1 Datoteke sa laptopa

Spisak datoteka je dobijen naredbom

```
dir /S /A-D /-C c:\ | sort > lista.txt
```

Kao rezultat navedene naredbe, za potrebe prvog uzorka, dobijen je spisak od 107.566 stavki. Veličine datoteka, izražene bajtovima, su bile u rasponu od 0 do 2.146.750.464. Nakon odbacivanja vrijednosti manjih od 100 uzorak je sveden na $N = 104.337$ stavke. Rasponi analiziranih frekvencija su u tabeli 6.2.

	Donja granica	**Gornja granica**
A	100	1.000
B	1.000	10.000
C	10.000	100.000
D	100.000	1.000.000
E	1000.000	10.000.000
F	10.000.000	100.000.000
G	100.000.000	1.000.000.000
H	1.000.000.000	10.000.000.000

Table 14: Tabela 6.2 Rasponi veliėina. Odabrani rasponi odrađavaju frekvencije prvih cifara

Ovako odabrani rasponi ujedno odražavaju i frekvencije vodećih cifara što će se pokazati kao bitna prednost u postupku analize. Frekvencije vrijednosti po rasponima iz tabele 6.2, proračunate Excell makroom, su u tabeli 6.3.
Ove frekvencije su prezentirane na grafikonu 6.1. Uočljiv je veliki stepen saglasnosti sa Benfordovim zakonom za grupe B , C i D. Izuzetak su veličine koje počinju cifrom 8 u grupi B. Ova činjenica je vidljiva i na grafikonu 6.2.

U tekstu u kojem obrađuje primjer detekcije prevara korištenjem metoda reinforcement učenja [63, 64] Fletcher Lu predlaže korištenje veličine :

	A	B	C	D	E	F	G	H
1	2.817	11.661	9.839	5.528	1.709	346	26	0
2	2.443	6.871	5.938	2.938	1.248	46	3	2
3	1.995	4.701	3.746	1.702	909	20	2	0
4	1.794	3.320	2.669	967	856	15	1	0
5	1.524	2.749	1.995	828	544	8	0	0
6	1.667	2.441	1.765	584	306	11	1	0
7	1.642	2.019	1.191	605	183	4	2	0
8	1.585	2.817	1.106	382	145	6	0	0
9	1.351	1.482	895	277	89	21	0	0
Ukupno	16.818	38.061	29.144	13.811	5.989	477	35	2

Table 15: Tabela 6.3. Frekvencije po rasponima i pocetnim ciframa

$$BE\,(i) = \frac{f_{1i}}{b_{1i}} + \frac{f_{2i}}{b_{2i}} + \frac{f_{3i}}{b_{3i}} \qquad (6.5)$$

U ovom izrazu f_{ji} predstavlja uzoračku a b_{ji} teorijsku frekvenciju grupa cifara dužine j za stanje (slog) i. Ovaj veličina se računa za svaki pojedini slog u uzorku, bilo da se koristi osnovna ili adaptivna metoda računanja. U nastavku će ova veličina, iz praktičnih razloga, biti obilježena sa

$$BE\,(3) = \frac{f_1}{b_1} + \frac{f_2}{b_2} + \frac{f_3}{b_3} \qquad (6.6)$$

Ako se po analogiji formira veličina

$$BE\,(2) = \frac{f_1}{b_1} + \frac{f_2}{b_2} \qquad (6.7)$$

tada je moguće računati količnik

$$BK32 = \frac{BR\,(3)}{BR\,(2)} = 1 + \frac{\dfrac{f_3}{b_3}}{\dfrac{f_1}{b_1} + \dfrac{f_2}{b_2}} \qquad (6.8)$$

Ovaj količnik mjeri uticaj treće cifre u frekvencijama. Ako podaci odražavaju regularan proces ili ako na njima nisu vršene intervencije u smislu odbacivanja vrijednosti iznad ili ispod odabranih granica, slogovi (odnosno stanja) za koje je $BE\,(3)$ najveći su isti slogovi za koje je $BK32$ najveći. U suprotnom ova saglasnost ne vrijedi. Ako se vrši odbacivanje tada je odstupanje posebno značajno ako se povećava donja granica. Ovo je zbog toga što Benfordovski skupovi, u pravilu, imaju veći broj malih veličina.

Kako bi se ove činjenice potvrdile provedeni su testovi tako da je pravljen izbor donjih i gornjih granica a zatim, putem odgovarajućeg makroa, računate veličine $BE\,(2)$ i $BK32$. Kao gornja granica je uzeta veličina 3.000.000, imajući u vidu da je iznad te granice oko $0,5\%$ cijelog uzorka. Za donje granice su uzimane veličine od 1.000 do 10.000 u koracima od 1.000 i iznos 8.500 te su računate veličine $BE\,(3)$ i $BK32$. U svim slučajevima kada je donja granica bila manja od 8.500 najveće vrijednosti za $BE\,(3)$ su odgovarale veličinama koje počinju sa 819 od kojih najveću frekvenciju ima veličina 8.192. U svim tim slučajevima najveća vrijednost za $BK32$ odgovara najvećim vrijednostima za $BE\,(3)$ iz uzorka.

Kada su kao donje granice uzimane vrijednosti 8.500 i veće najveća vrijednosti za $BE\,(3)$ nije odgovarala najvećoj vrijednosti količnika $BK32$. Uporedni pregled ove dvije veličine za razne vrijednosti donje granice dat je u tabeli 6.4.

	$BR\,(3)$	$BK32$
1.000	$35,036664$	$7,442831$
2.000	$37,114702$	$7,442884$
3.000	$39,460441$	$7,442710$
4.000	$42,018276$	$7,442566$
5.000	$44,662249$	$7,442756$
6.000	$47,706176$	$7,442831$
7.000	$51,386855$	$7,443100$
8.000	$55,474220$	$7,443527$
8.500	$8,125298$	$4,019510$
9.000	$8,388329$	$4,114099$
10.000	$9,161922$	$4,208542$

Table 16: Tabela 6.4. Velicine BE(3) i BK32 za razne vrijednosti donje granice. Uocljiva je nagla promjena izmedju velicina 8.000 i 8.500

Provedeni su dodatni testovi za donje pragove 8.100, 8.200, 8.300 i 8.400. Promjena je detektovana pri prelasku sa 8.100 na 8.200. Ovo ukazuje da granicu na kojoj se dešava ova promjena treba tražiti u grupi iznosa za koje frekvencija prve cifre značajno prelazi očekivanu teorijsku frekvenciju. Razlog za ovo je u trećem članu desne strane izraza (6.5) odnosno (6.6). Dodatnim testovima je ustanovljeno da je sa stanovišta ove analize kritična bila vrijednost 8.192 odnosno veličina 8K.

Uzimanje granice iznad veličine čije prve tri cifre imaju ekstremno visoku frekvenciju uzrokuje bitno smanjenje trećeg sabirka a samim tim i pripadajućeg zbira odnosno količnika. U ovom slučaju to se desilo pri prelasku veličine 8.192. Ilustracija ove činjenice je na grafikonu 6.3.

Ova činjenica je veoma bitna sa više aspekata. Prvi aspekt se odnosi na praktični pristup analizi. U slučaju ovakvog rezultata postoji osnova za sumnju u samu strukturu podataka. Prije toga se mora procijeniti postupak odbacivanja malih vrijednosti ako je takvo odbacivanje urađeno. Ovakav rezultat ukazuje da odbacivanje iznosa koji su veći od kritične vrijednosti to možda ne bi bio dobar potez. Drugi aspekt se odnosi na primjenu u problemima u kojima se veličine $BE\,(3)$ i $BK32$ ne koriste samo za detekciju anomalija već kao parametri analiza. Primjer je reinforcement učenje u kojem se ove veličine koriste kao kriterij izbora stanja odnosno akcije.

Bitna činjenica koja je ustanovljena i koja je vidljiva na grafikonu je da obje krive imaju trend rasta kada se donja granica povećava. Rast se nastavlja i nakon ove tačke do naredne kritične veličine. Dodatne analize mogu pokazati u kojoj mjeri ovaj uticaj zavisi od stepena značajnosti odstupanja od teorijskih vrijednosti za Benfordov zakon.

6.5.2 Slučajno odabrani uzorak

Uticaj izbora granica moguće je pokazati računanjem nekih globalnih veličina kao što su

- MAD : srednja apsolutna devijacija
- DF : Faktor distorzije

Predmet kalkulacija je mogao biti i jedan broj drugih veličina kao npr. totalna varijansa, očekivana cifra i slično. Ove veličine su odabrane jer je očigledno da sadrže dovoljan nivo globalne informacije o skupu. Način računanja ove dvije veličine dat je u poglavlju 2 ovog teksta.

Testovi su obavljeni na uzorku obima $N = 10.241$. Vrijednosti u uzorku su u rasponu od 0.01 do 500.000 i jedna stavka od 5.000.000. Ako se uzorak sortira u rastućem redoslijedu dobija se grafikon 6.4. u prilogu.

Svaka od izabranih veličina je data zasebno za prvu cifru, prve dvije i prve tri cifre. Pritom je MAD računat po klasičnoj i po adaptivnoj metodi. Iz praktičnih razloga, donji i gornji pragovi su uzimani u približno jednakim razmacima. Pritome najveća donja granica nije veća od najmanje gornje granice. Dobijeno je po deset vrijednosti za donji i gornji prag. Sa stanovišta ovog testa bitno je da li i pod kojim uslovima dolazi do promjene posmatranih veličina ako se mijenjaju donja i gornja granica.

Rezultati su predstavljeni u tabelama uz grafikone 6.5 do 6.12 u prilogu. Prvih šest tablica se odnosi na MAD za prvu, prve dvije i prve tri cifre, računat na standardan Benfordov način i po adaptivnoj metodi. Sedma tabela daje prikaz faktora distorzije za svaku kombinaciju pragova. Osma tabela daje procenat uzorka koji je bio predmet kalkulacija na svakom koraku.

6.6 Diskusija

Nakon provedenih testova se može zaključiti da promjene granica uzorka imaju uticaj na veličine kojima se opisuju neke globalne karakteristike i na veličine koje se izvode po osnovu Benfordovog zakona. Pritom, veći uticaj ima promjena donje granice. Ako se računaju veličine $BE(3)$ i $BK32$ promjene se dešavaju na vrijednostima za koje je frekvencija prvih cifara bitno veća od teorijske. Povećana frekvencija ima ima uticaj na treći sabirak u izrazu (6.1) odnosno (6.2). Ako uzorak ima osobinu da slogovi za koje najveće vrijednosti za $BE(3)$ nisu u svim slučajevima jednaki onima za koje je $BK32$ uzrok se može tražiti u činjenici da uzorak nije kompletan ili u prirodi odnosno izbora podataka.

Zaključak o zavisnosti rezultata Benfordove analize od granice donjeg praga isti je u slučaju kada se računaju neki globalni parametri (MAD, Faktor distorzije). Međutim, ove veličine ne ukazuju na pragove u kojima se dešavaju bitne promjene na način kako to čine veličine $BE(3)$ i $BK32$. Ovo je logično s obzirom da se MAD računa kao globalna veličina na nivou cijelog skupa dok se $BE(3)$ i $BK32$ računaju za svaki član tog skupa.

Na isti zaključak upućuju i grafikoni dobijeni na osnovu ovih tabela. Iz grafikona za prvi uzorak vidljiv je veoma mali broj elemenata iznad granice od 100.000 (u procentima u opsegu od $0,68\%$ do najviše $1,45\%$ obima uzorka), što ima uticaja na kalkulacije. Drugim riječima, uticaj gornje granice bi bio vidljiviji ako bi iza granice 100.000 obim uzorka bio najmanje $1,50\%$. Mogući uzrok je u činjenici da su 'velike' vrijednosti iz intervala koji obuhvata manji red veličina nego što je to slučaj sa 'malim' vrijednostima. U drugom uzorku iznad granice od 3.000.000 je manje od 1% uzorka ali je frekvencija u intervalima $[10^2, 10^3]$, $[10^3, 10^4]$ mnogo veća.

Uticaj na globalne pokazatelje i ukupnu analizu cifara je manji ako je razlika gornje i donje granice veća od dva reda stepena baze. Ovo spada u jedan od standardnih uslova za analize putem Benfordovog zakona. U drugom uzorku se to očituje tako što su anomaličnosti veće ako su donja i gornja granica unutar intervala $[10.000, 100.000]$ ili $[1.000, 10.000]$. U slučaju koji je bio analiziran ovo je posebno bilo vidljivo za faktor distorzije.

Kad je u pitanju veličina MAD, u slučaju kada granice obuhvataju barem dva reda veličina vrijedi da je $MAD_1 \geq MAD_2 \geq MAD_3$, gdje MAD_i označava MAD za sekvence dužine i. U slučajevima kada su gornja i donja granica iz opsega jednog reda veličina tada je $MAD_2 \leq MAD_3$. Ista činjenica je detektovana i korištenjem adaptivne Benfordove metode ali za donje granice koje su osjetno više u odnosu na analizu putem klasične metode. Pritom se detekcija putem adaptivne Benfordove metode ograničava na mnogo uži interval. Drugim riječima, klasična Benfordova metoda detektuje daleko veći broj anomaličnosti u odnosu na adaptivnu Benfordovu metodu pa u tom slučaju postoji rizik lažne detekcije. Detekcija pojave da je $MAD_2 \leq MAD_3$ daje osnovu za sumnju da nedostaje dio podataka, bilo ispod donje odnosno iznad neke gornje granice.

Na ovaj način je potvrđeno da izbor granica ima značajan uticaj na analize cifara putem Benfordove metode.

6.7 Zaključak

Adaptivna Benfordova metoda je napravila bitan pomak u mogućnosti primjene Benfordovog zakona jer se umanjuje uticaj ograničenja veličine skupa koji je predmet analize, posebno na slučajeve za koje se do sada smatralo da ne mogu biti predmet analize ovog tipa.

Prva karakteristika Adaptivne Benfordove metode je mogućnost analize na uzorcima malog obima. U patentnoj prijavi je dat pregled obima uzoraka koji su bili osnova za procjene po pojedinim intervalima. U najgorem slučaju uzorak je imao 24 elementa što je manje od 1% obima cjelokupnog uzorka. Naravno, obim uzorka diktira i metod analize. Na uzorcima ovakvog obima nije moguća npr. analiza frekvencija prve tri cifre.

Druga karakteristika Adaptivne Benfordove metode je veći stepen detekcije u odnosu na tradicionalnu Benfordovu metodu što je potvrđeno u provedenim eksperimentima. Stepen detekcije je povećan jer uzimanje prosječne vrijednosti na osnovu proračunatih frekvencija smanjuje efekat 'bubrenja' odnosno efekat da zbog nedostajućih vrijednosti pojedine teorijske frekvencije budu izrazito velike.

Treća bitna karakteristika Adaptivne Benfordove metode je mogućnost kalkulacije globalnih parametara s obzirom da Adaptivna metoda daje novu procjenu teorijskih frekvencija i intervala povjerenja.

Detekcija granice na kojoj se mijenja uticaj treće cifre pruža mogućnost kvalitetnije procjene eventualnih nedostataka kao i strukturnih ili sadržajnih anoamlija u poadcima.

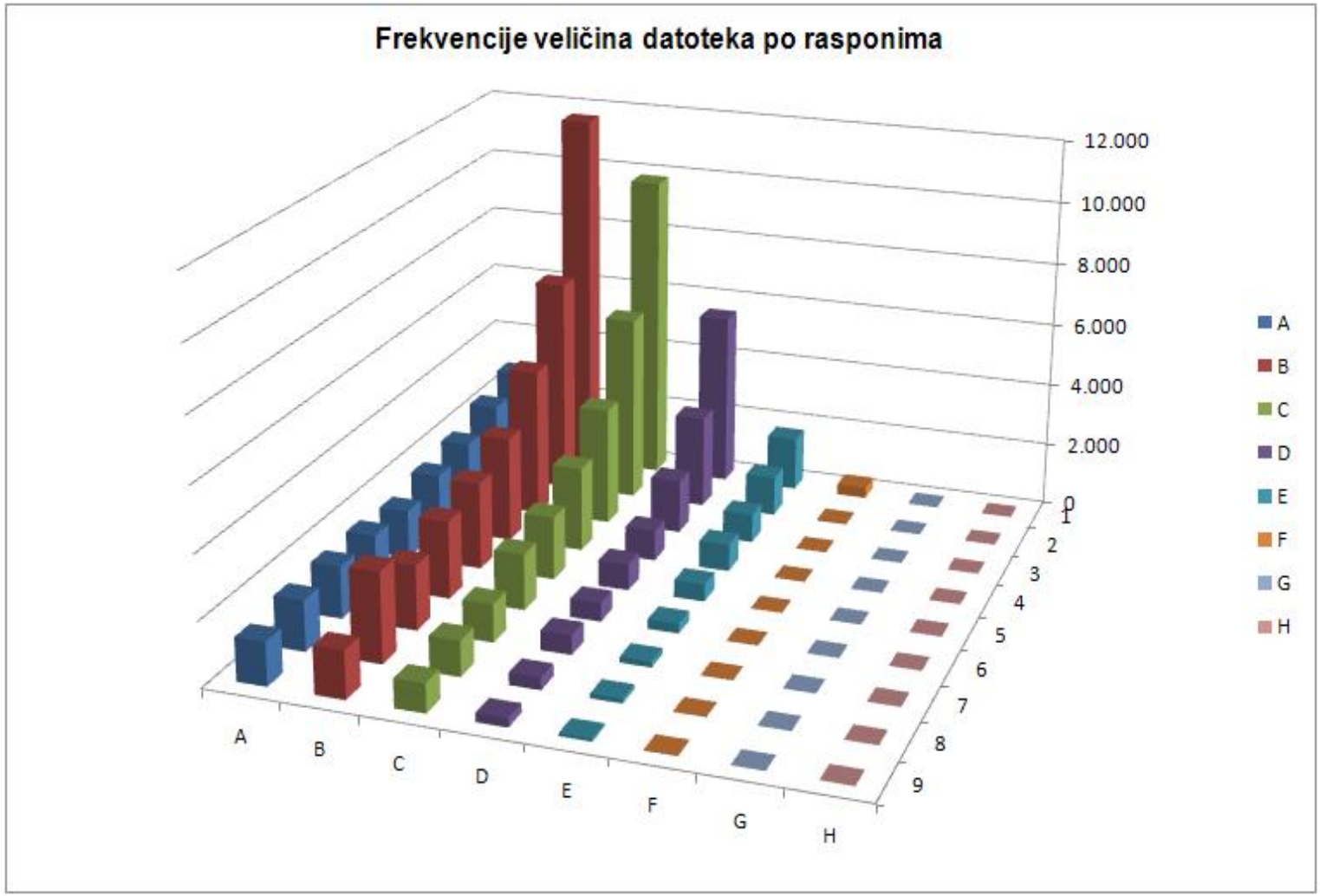
Na raspolaganju je više metoda kojima se može procijeniti da li je izvor anomaličnosti činjenica da nedostaje dio podataka ispod ili iznad nekih pragova :

1. Računanje veličina MAD_1, MAD_2 i MAD_3 i ispitivanje njihovog odnosa
2. Računanje veličina $BE\,(3)$ i $BK32$
3. Mijenjanje gornje granice uzorka u dovoljno malim koracima kako bi se ispitalo kako se mijenja faktor distorzije

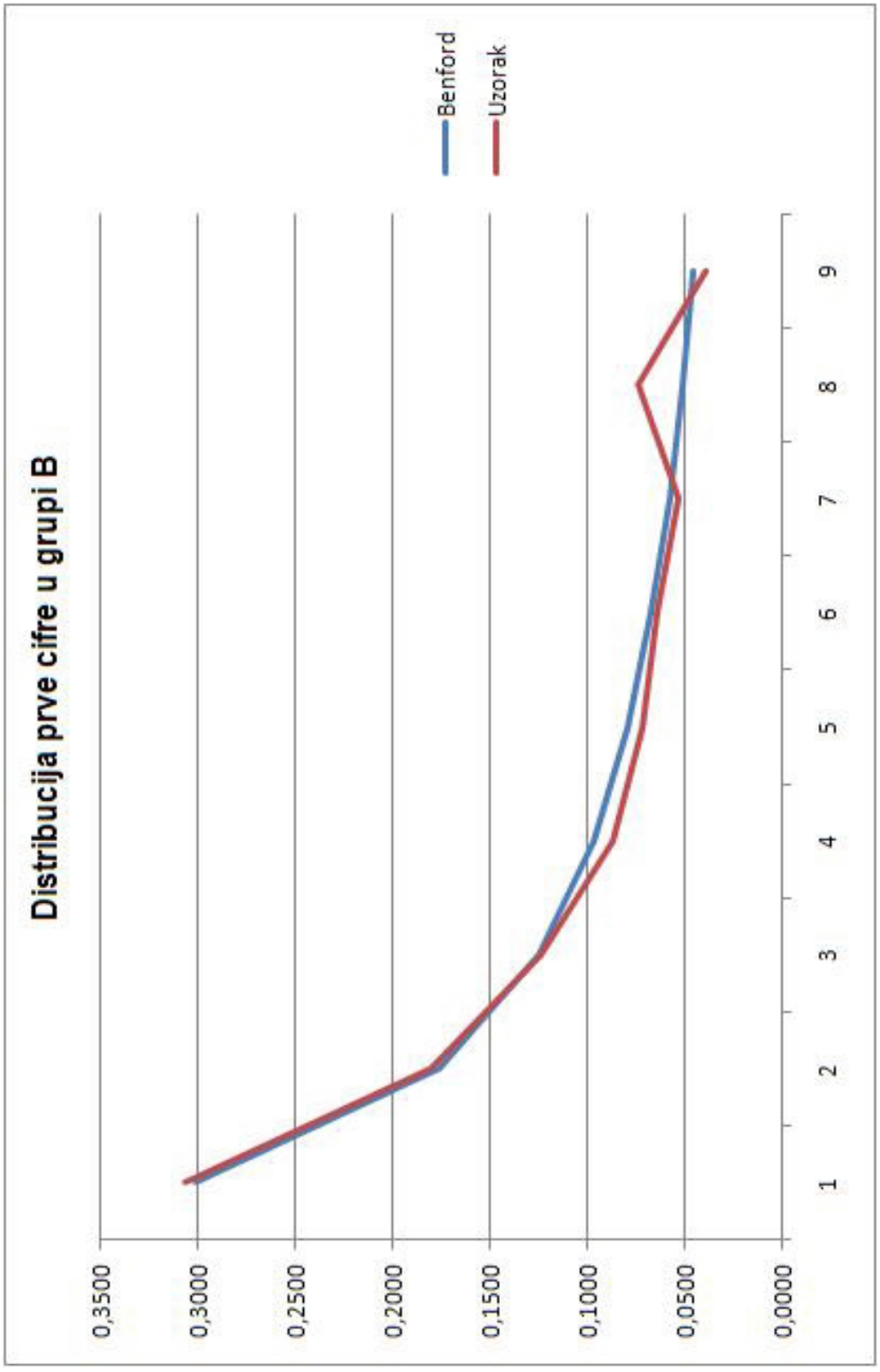
Pomoć u identifikaciji donje i gornje granice može biti grafička prezentacija uzorka sortiranog u rastućem redoslijedu.

U smislu praktične primjene, ako postoji potreba da se utvrdi adekvatnost uzorka prednost se može dati računanju vrijednosti $BE\,(3)$ i $BK32$. Razloge treba tražiti u bitno manjem naporu računanja. Druga prednost je u tome što je moguće prepoznati kritične vrijednosti na kojima dolazi do bitnih promjena veličina $BE\,(3)$ i $BK32$ i njihovih odnosa. Ako se računaju veličine MAD i faktor distorzijce ovakav zaključak se može izvesti indirektno, nakon kalkulacija koje mogu biti dugotrajne i složene.

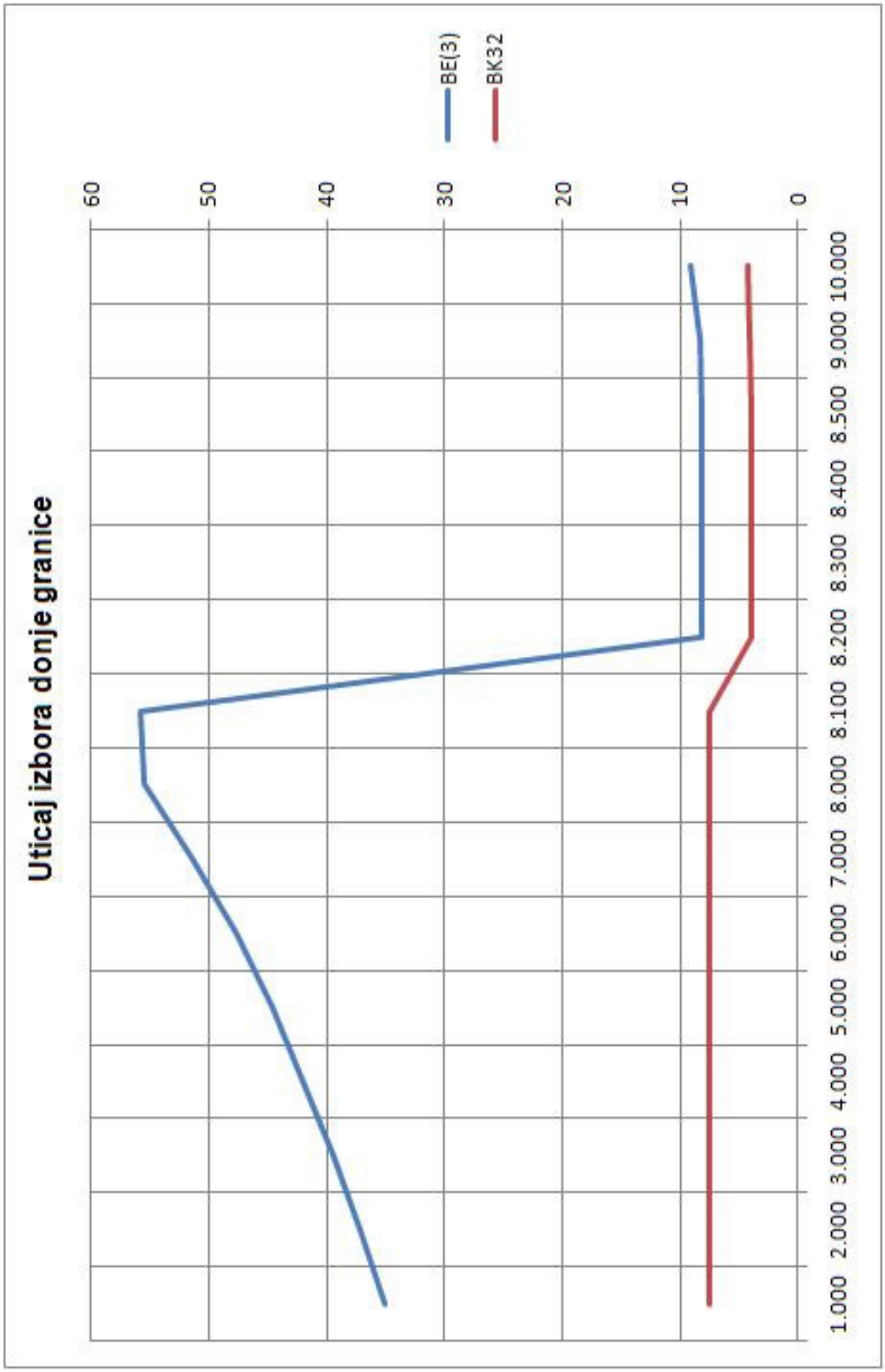
Važno je napomenuti da nijedna od ovih metoda ne daje mogućnost rekonstrukcije nedostajućih vrijednosti ispod donje granice već samo daje indikaciju anomaličnosti.



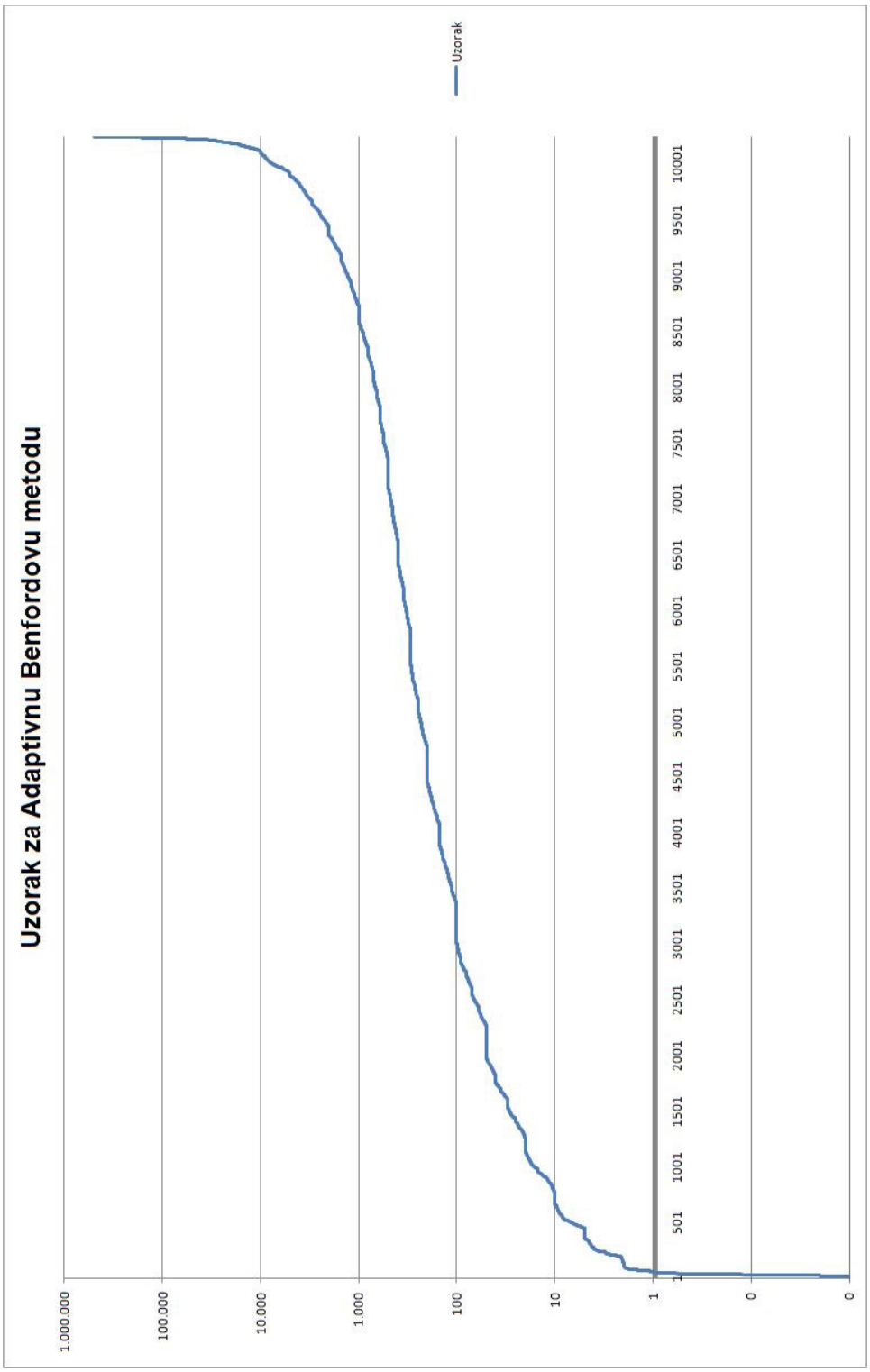
Grafikon 6.1. Frekvencije veličina datoteka po rasponima veličina. Za grupe B i C je uočljiv veliki stepen slaganja sa Benfordovim zakonom osim za vrijednosti iz grupe B koje počinju cifrom 8



Grafikon 6.2. Distribucija prve cifre za grupu B (1.000 - 10.000). Uočljiva je razlika u frekvencijama za veličine koje počinju cifrom 8



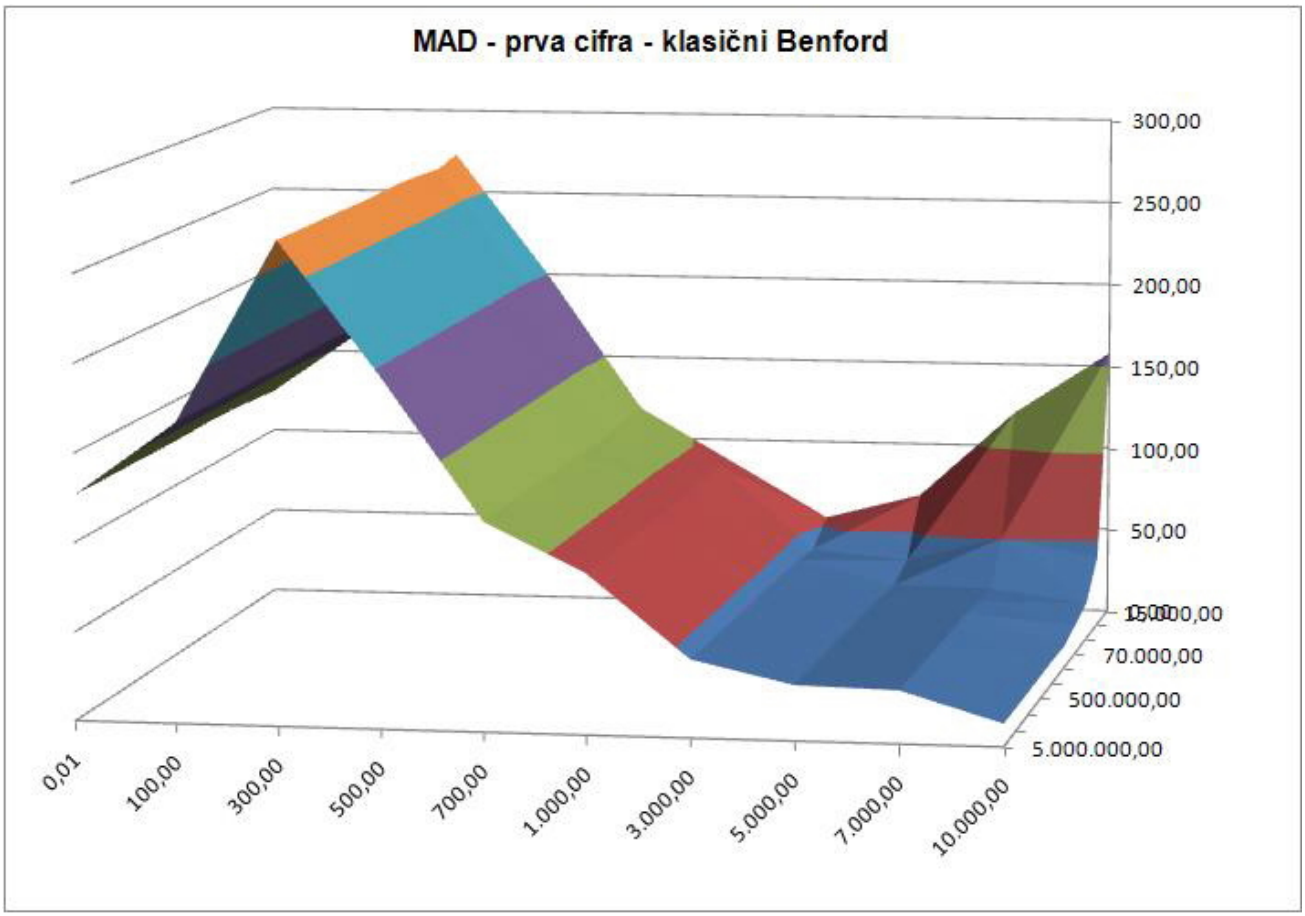
Grafikon 6.3. Uporedni prikaz veličina BE(3) i BK32. Uočljiv je pad obje veličine kada se pređe prag 8.192. Takođe, uočljiv je trend rasta obje veličine do momenta naglog pada nadalje



Grafikon 6.4. Grafički prikaz sortiranog uzorka. Na slici je vidljivo da je obim uzorka u rasponu (0;100) a posebno u rasponu (0;1.000) bitno veći od onog iz raspona (100.000;1.000.000), što je jedan od razloga za veću zavisnost adaptivne metode od vrijednosti donjeg praga

MAD - prva cifra - klasični Benford

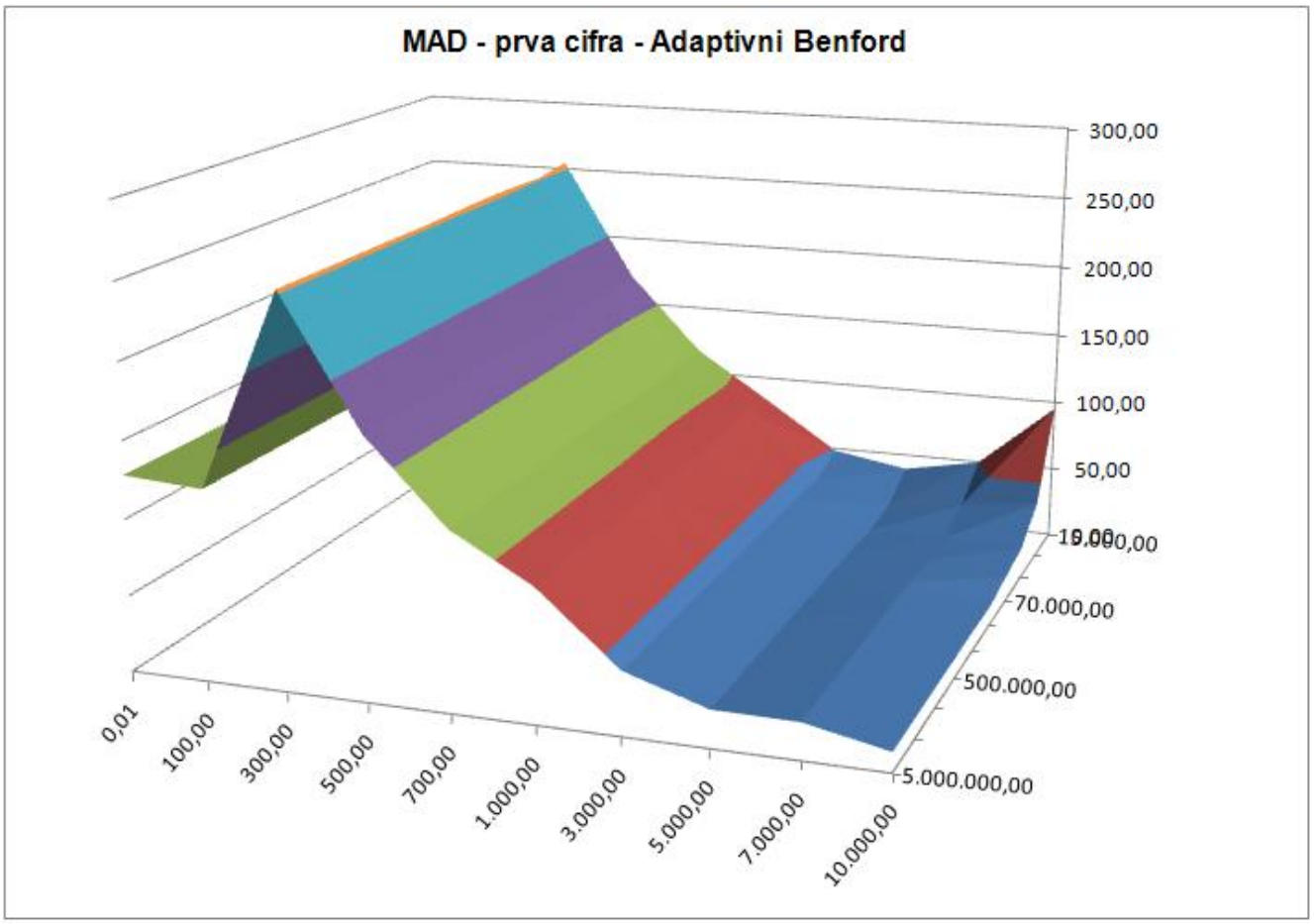
MAD - 1. cifra	15.000,00	30.000,00	50.000,00	70.000,00	100.000,00	300.000,00	500.000,00	700.000,00	1.000.000,00	5.000.000,00
0,01	124,31	125,45	126,61	126,99	126,37	126,05	126,28	126,28	126,28	126,40
100,00	164,18	168,13	168,58	168,45	167,61	168,04	168,09	168,09	168,09	168,11
300,00	272,92	268,96	270,38	270,93	270,56	270,07	270,37	270,37	270,37	270,52
500,00	199,24	195,08	195,00	195,96	196,28	195,84	196,22	196,22	196,22	196,40
700,00	118,31	117,58	116,58	116,32	117,08	117,44	117,24	117,24	117,24	117,14
1.000,00	85,86	90,90	90,66	89,93	89,48	89,99	89,71	89,71	89,71	89,56
3.000,00	53,32	41,41	42,76	43,27	43,71	43,21	43,49	43,49	43,49	43,64
5.000,00	68,38	36,18	28,98	29,98	30,79	30,36	30,75	30,75	30,75	30,94
7.000,00	120,25	51,23	28,45	28,80	30,02	29,60	29,62	29,62	29,62	29,63
10.000,00	157,89	40,68	21,25	16,81	12,27	12,91	12,76	12,76	12,76	12,87



Grafikon 6.5. Grafikon za veličinu MAD za prve cifre odabranih vrijednosti donjeg i gornjeg praga računat klasičnim Benfordovim metodom. Primjetan je veliki MAD za vrijednosti bliske pragovima

MAD - prva cifra - Adaptivni Benford

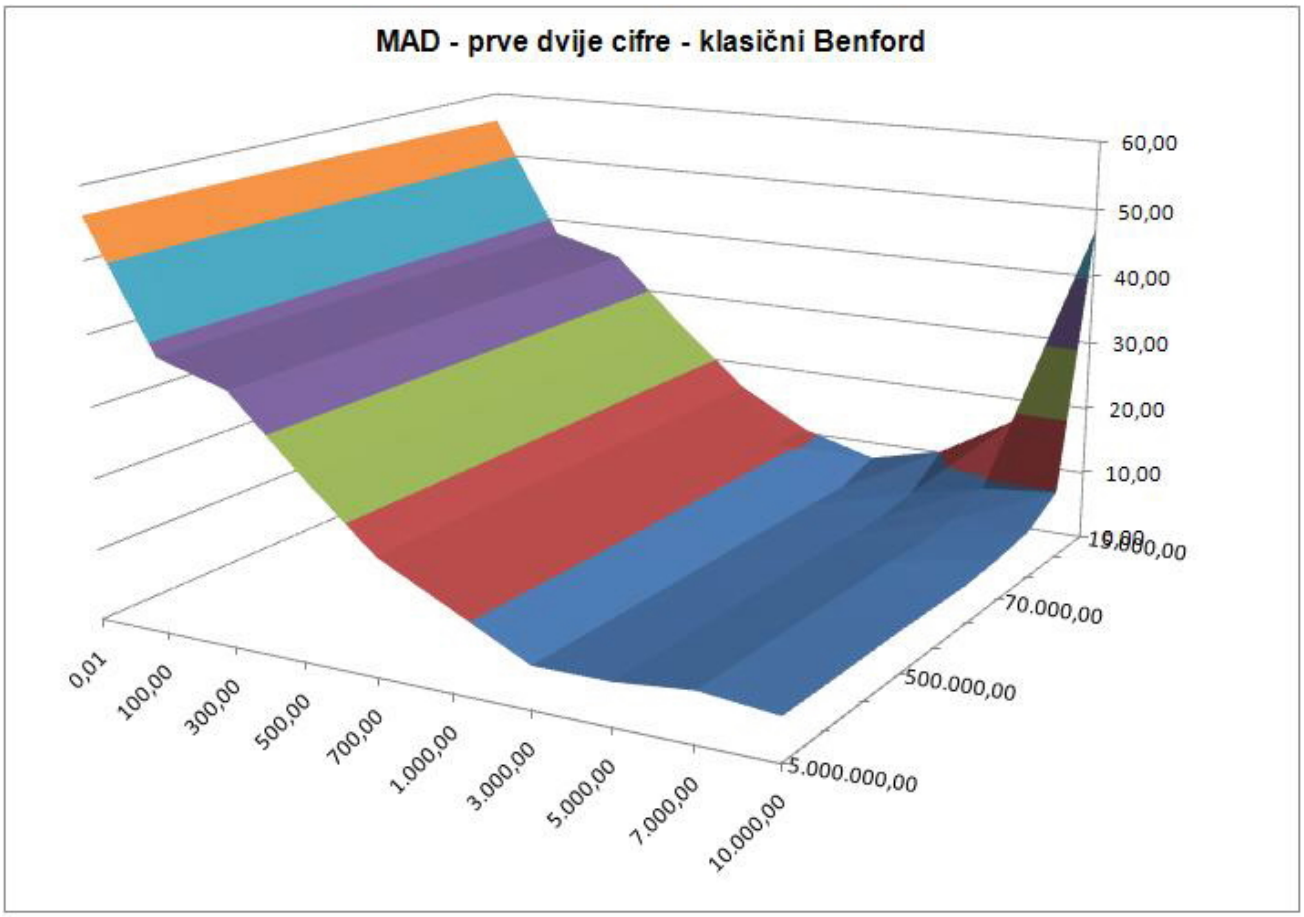
MAD - 1. cifra	15.000,00	30.000,00	50.000,00	70.000,00	100.000,00	300.000,00	500.000,00	700.000,00	1.000.000,00	5.000.000,00
0,01	125,83	127,09	128,26	128,61	127,98	127,67	127,90	127,90	127,90	128,02
100,00	121,63	124,14	124,64	124,66	124,07	124,31	124,41	124,41	124,41	124,46
300,00	254,62	251,34	252,74	253,22	252,73	252,32	252,61	252,61	252,61	252,75
500,00	170,96	167,55	167,33	168,10	168,59	168,26	168,57	168,57	168,57	168,72
700,00	116,88	116,38	115,37	115,08	115,81	116,19	115,99	115,99	115,99	115,89
1.000,00	82,91	87,95	86,89	86,36	86,17	86,75	86,54	86,54	86,54	86,43
3.000,00	48,23	39,06	40,40	40,85	41,20	40,76	41,03	41,03	41,03	41,16
5.000,00	39,31	25,64	22,57	23,40	23,97	23,71	24,05	24,05	24,05	24,22
7.000,00	49,51	32,26	23,05	22,73	23,55	23,98	23,75	23,75	23,75	23,64
10.000,00	95,06	38,82	21,38	17,02	12,55	13,20	12,97	12,97	12,97	12,85



Grafikon 6.6. Grafikon za veličinu MAD za odabrane vrijednosti donjeg i gornjeg praga računatu prema adaptivnoj metodi. Potrebno je primijetiti bitno manji MAD za bliske vrijednosti pragova u odnosu na obračun putem klasične Benfordove metode

MAD - prve dvije cifre - klasični Benford

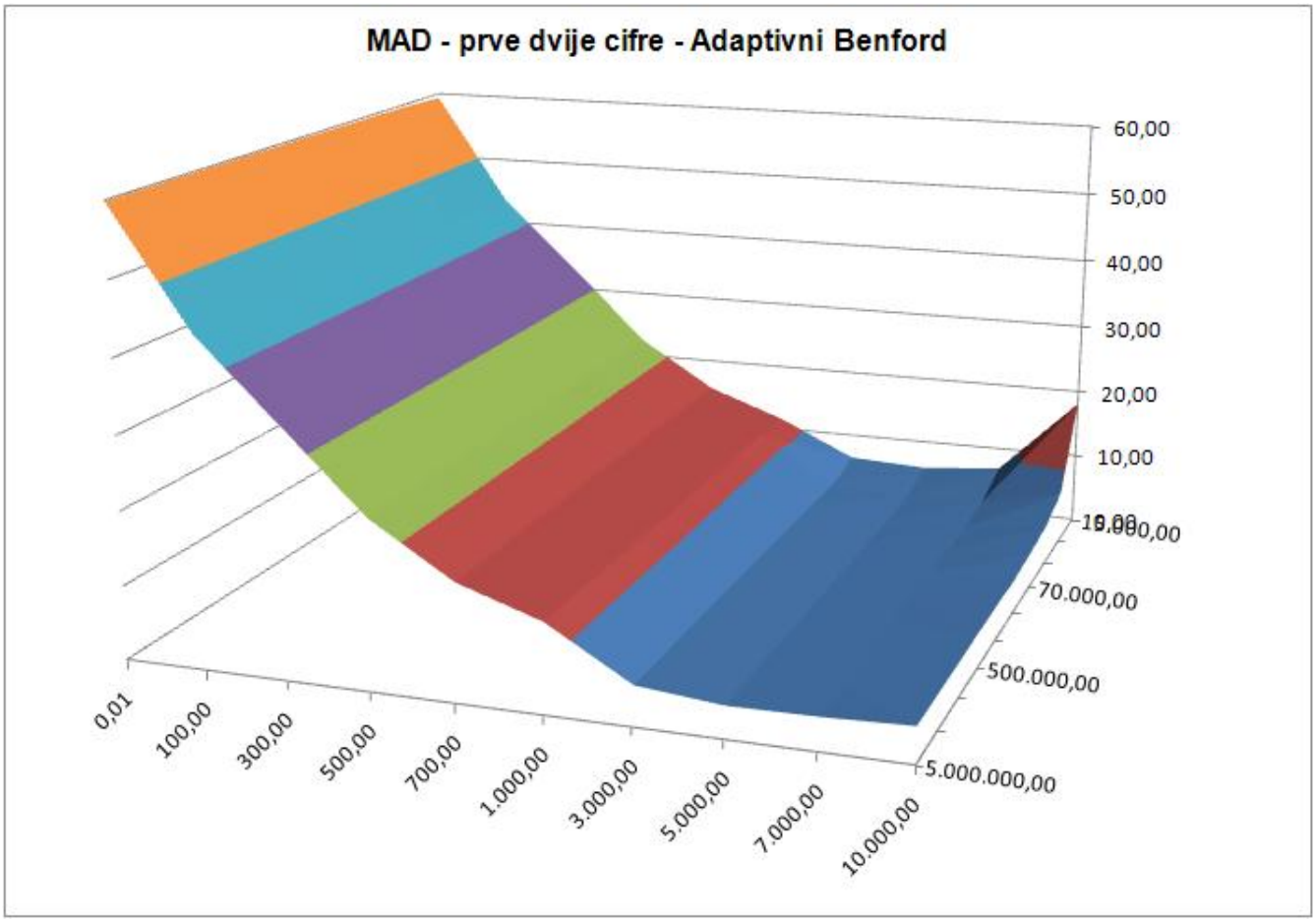
MAD - 2. cifra	15.000,00	30.000,00	50.000,00	70.000,00	100.000,00	300.000,00	500.000,00	700.000,00	1.000.000,00	5.000.000,00
0,01	55,61	55,71	55,82	55,85	55,94	55,94	55,98	55,98	55,98	56,00
100,00	37,81	37,93	38,03	38,06	38,13	38,13	38,17	38,17	38,17	38,19
300,00	35,10	34,90	35,04	35,10	35,17	35,12	35,16	35,16	35,16	35,17
500,00	25,23	25,17	25,16	25,23	25,31	25,29	25,33	25,33	25,33	25,35
700,00	16,15	16,12	16,12	16,14	16,25	16,23	16,26	16,26	16,26	16,28
1.000,00	10,51	10,91	10,98	10,98	11,03	11,09	11,11	11,11	11,11	11,12
3.000,00	7,53	5,57	5,74	5,81	5,90	5,86	5,89	5,89	5,89	5,91
5.000,00	10,17	6,33	5,54	5,70	5,89	5,85	5,90	5,90	5,90	5,93
7.000,00	16,50	8,80	7,04	6,66	6,96	6,91	6,95	6,95	6,95	6,99
10.000,00	46,90	9,90	6,79	6,01	5,54	5,70	5,88	5,88	5,88	5,97



Grafikon 6.7. Grafikon za veličinu MAD za prve dvije cifre odabranih vrijednosti donjeg i gornjeg praga računatog klasičnim Benfordovim metodom. Potrebno je primijetiti veliki MAD za bliske vrijednosti pragova

MAD - prve dvije cifre - Adaptivni Benford

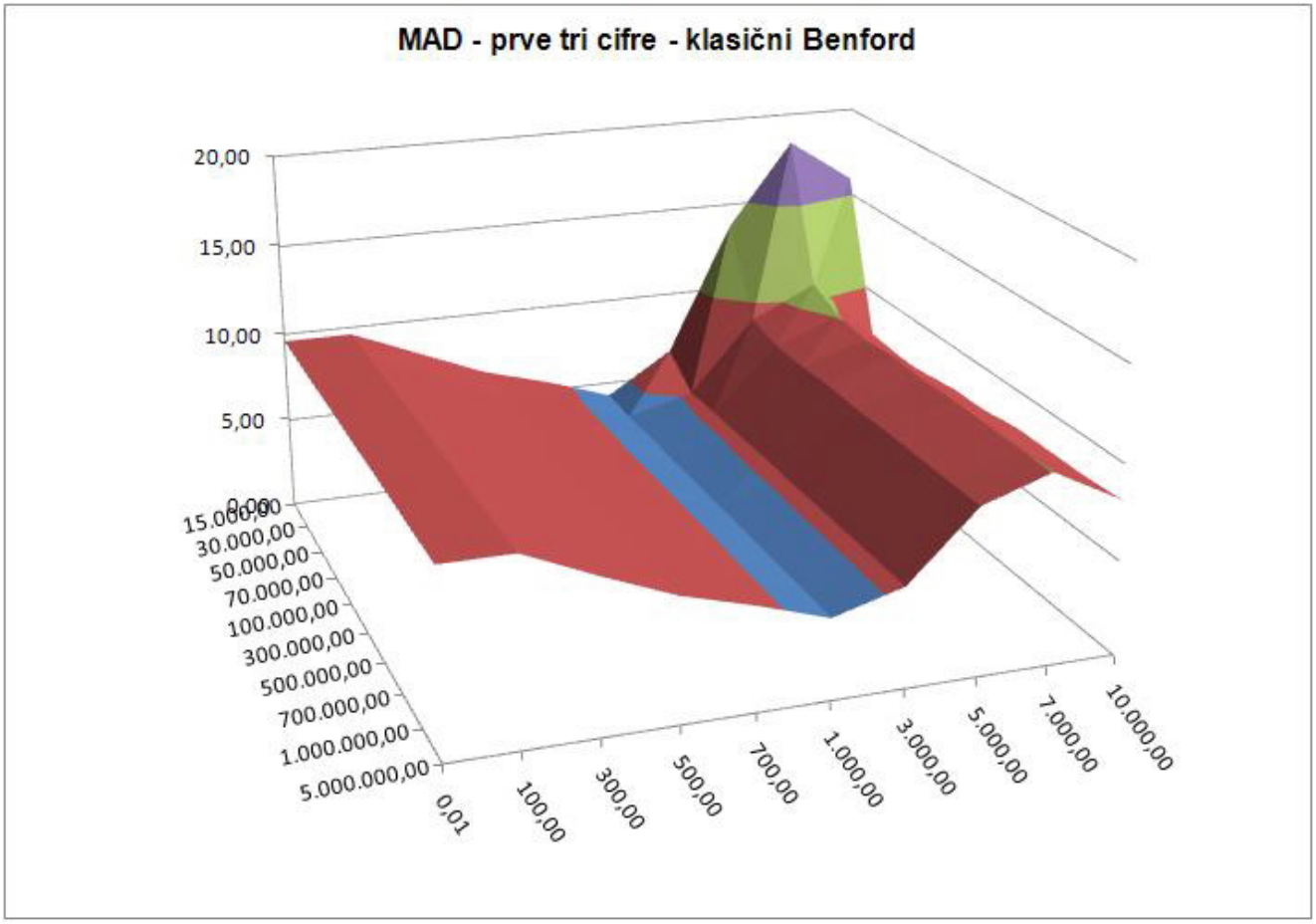
MAD - 2. cifra	15.000,00	30.000,00	50.000,00	70.000,00	100.000,00	300.000,00	500.000,00	700.000,00	1.000.000,00	5.000.000,00
0,01	59,33	59,58	59,69	59,71	59,77	59,79	59,82	59,82	59,82	59,84
100,00	43,33	43,59	43,71	43,73	43,79	43,81	43,85	43,85	43,85	43,87
300,00	33,17	33,07	33,22	33,27	33,31	33,28	33,31	33,31	33,31	33,32
500,00	22,48	22,49	22,47	22,55	22,63	22,62	22,65	22,65	22,65	22,67
700,00	15,56	15,58	15,60	15,61	15,70	15,68	15,71	15,71	15,71	15,72
1.000,00	11,58	11,85	11,89	11,87	11,93	11,96	11,99	11,99	11,99	12,01
3.000,00	6,36	5,06	5,21	5,28	5,37	5,32	5,35	5,35	5,35	5,37
5.000,00	5,83	4,31	3,98	4,07	4,18	4,22	4,25	4,25	4,25	4,27
7.000,00	6,83	4,91	4,27	4,14	4,28	4,33	4,38	4,38	4,38	4,41
10.000,00	18,04	7,24	5,45	4,90	4,53	4,69	4,77	4,77	4,77	4,81



Grafikon 6.8. Grafikon za veličinu MAD za prve dvije cifre za odabrane vrijednosti donjeg i gornjeg praga računatog Adaptivnim Benfordovim metodom. Potrebno je primijetiti bitno manji MAD za bliske vrijednosti pragova u odnosu na klasični Benfordov metod

MAD - prve tri cifre - klasični Benford

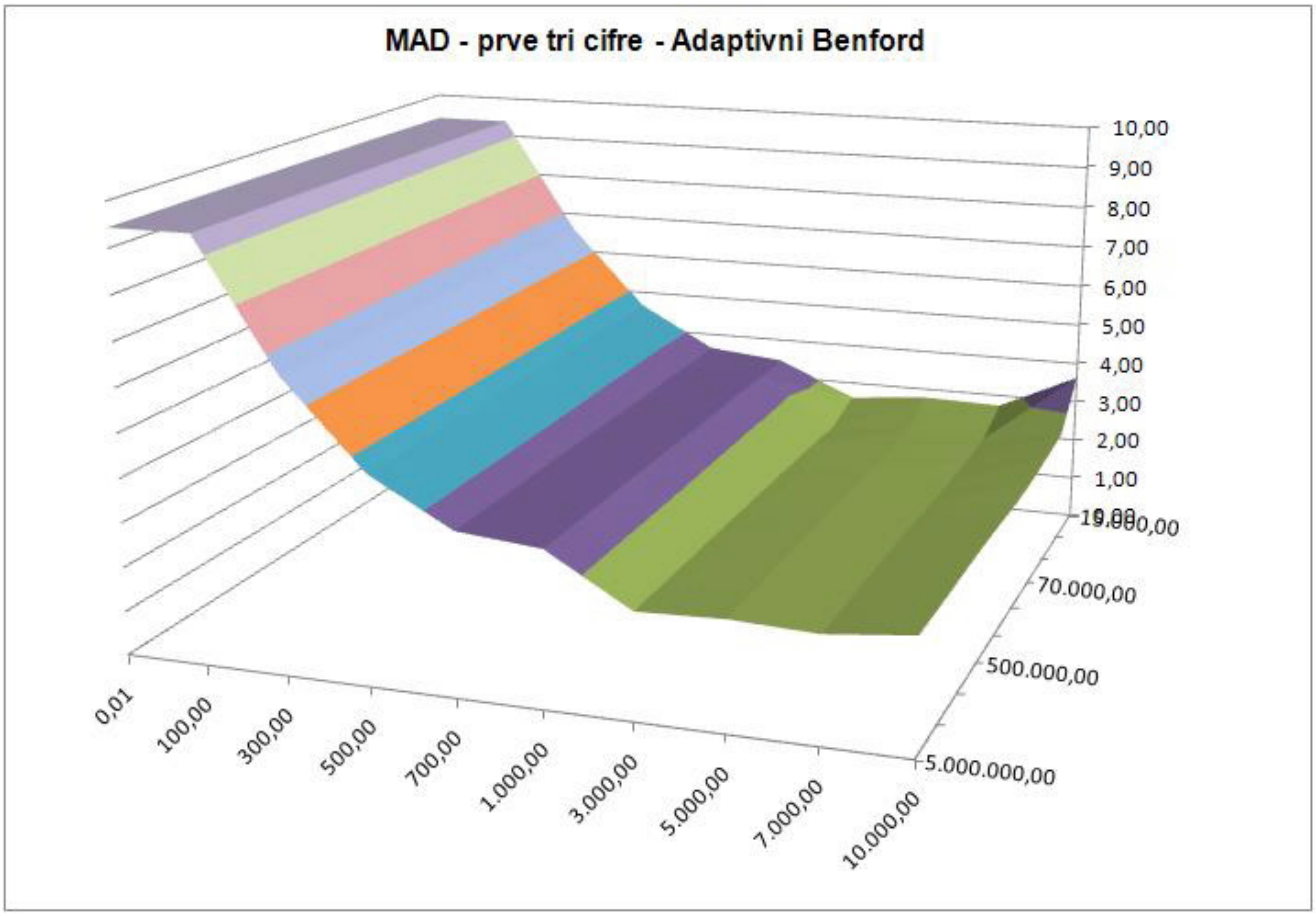
MAD - 3. cifra	15.000,00	30.000,00	50.000,00	70.000,00	100.000,00	300.000,00	500.000,00	700.000,00	1.000.000,00	5.000.000,00
0,01	9,60	9,59	9,61	9,62	9,63	9,64	9,64	9,64	9,64	9,64
100,00	9,60	9,59	9,61	9,62	9,63	9,64	9,64	9,64	9,64	9,64
300,00	8,07	7,95	7,97	7,98	8,00	8,00	8,00	8,00	8,00	8,01
500,00	6,54	6,43	6,45	6,46	6,48	6,49	6,49	6,49	6,49	6,50
700,00	5,50	5,40	5,42	5,39	5,42	5,43	5,44	5,44	5,44	5,44
1.000,00	4,24	4,18	4,21	4,19	4,21	4,22	4,24	4,24	4,24	4,24
3.000,00	6,54	5,16	5,25	5,23	5,28	5,25	5,28	5,28	5,28	5,29
5.000,00	13,72	9,15	8,70	8,67	8,79	8,73	8,79	8,79	8,79	8,82
7.000,00	18,33	10,85	10,06	9,79	9,98	9,90	9,99	9,99	9,99	10,03
10.000,00	15,89	7,50	7,31	7,30	7,66	7,71	7,99	7,99	7,99	8,14



Grafikon 6.9. Grafikon za veličinu MAD za prve tri cifre odabranih vrijednosti donjeg i gornjeg praga računatog klasičnim Benfordovim metodom. Izgled dijagrama bitno odudara od onog za prve dvije cifre. I dalje je vidljiva velika vrijednost za bliske vrijednosti pragova

MAD - prve tri cifre - Adaptivni Benford

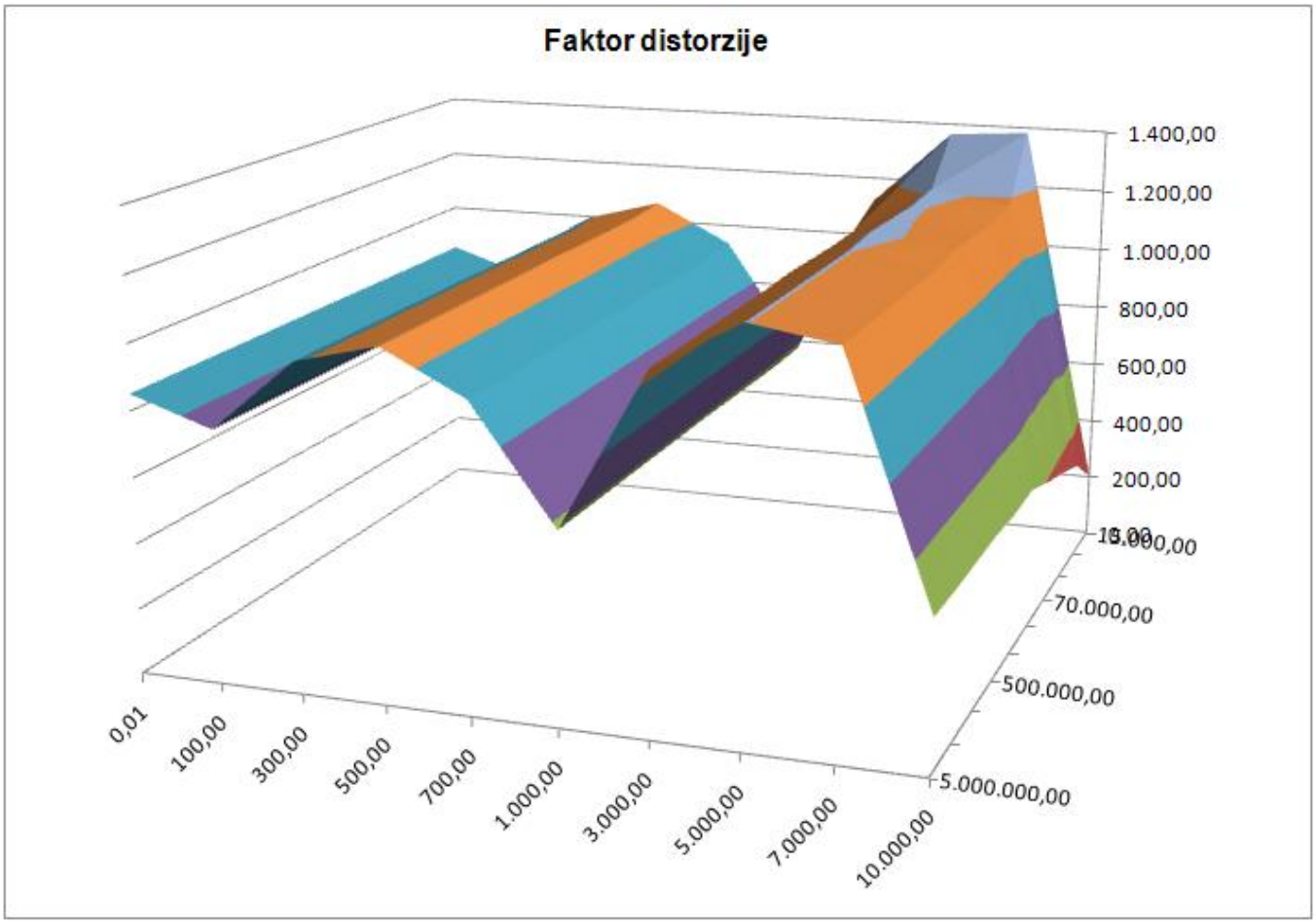
MAD - 3. cifra	15.000,00	30.000,00	50.000,00	70.000,00	100.000,00	300.000,00	500.000,00	700.000,00	1.000.000,00	5.000.000,00
0,01	9,39	9,40	9,41	9,42	9,43	9,43	9,44	9,44	9,44	9,44
100,00	9,39	9,40	9,41	9,42	9,43	9,43	9,44	9,44	9,44	9,44
300,00	6,58	6,53	6,55	6,56	6,57	6,57	6,58	6,58	6,58	6,58
500,00	4,65	4,62	4,63	4,64	4,65	4,66	4,66	4,66	4,66	4,66
700,00	3,63	3,61	3,62	3,61	3,63	3,63	3,64	3,64	3,64	3,64
1.000,00	3,44	3,44	3,46	3,44	3,45	3,46	3,47	3,47	3,47	3,47
3.000,00	2,61	2,29	2,33	2,34	2,35	2,36	2,36	2,36	2,36	2,37
5.000,00	2,78	2,39	2,36	2,37	2,40	2,42	2,43	2,43	2,43	2,44
7.000,00	2,73	2,34	2,29	2,27	2,31	2,33	2,35	2,35	2,35	2,36
10.000,00	3,63	2,64	2,55	2,48	2,47	2,52	2,55	2,55	2,55	2,57



Grafikon 6.10. Grafikon za veličinu MAD za prve tri cifre odabranih vrijednosti donjeg i gornjeg praga računatog Adaptivnim Benfordovim metodom. Potrebno je primijetiti bitno drugačiji oblik dijagrama u odnosu na dijagram dobijen korištenjem klasičnog Benfordovog metoda

Faktor distorzije

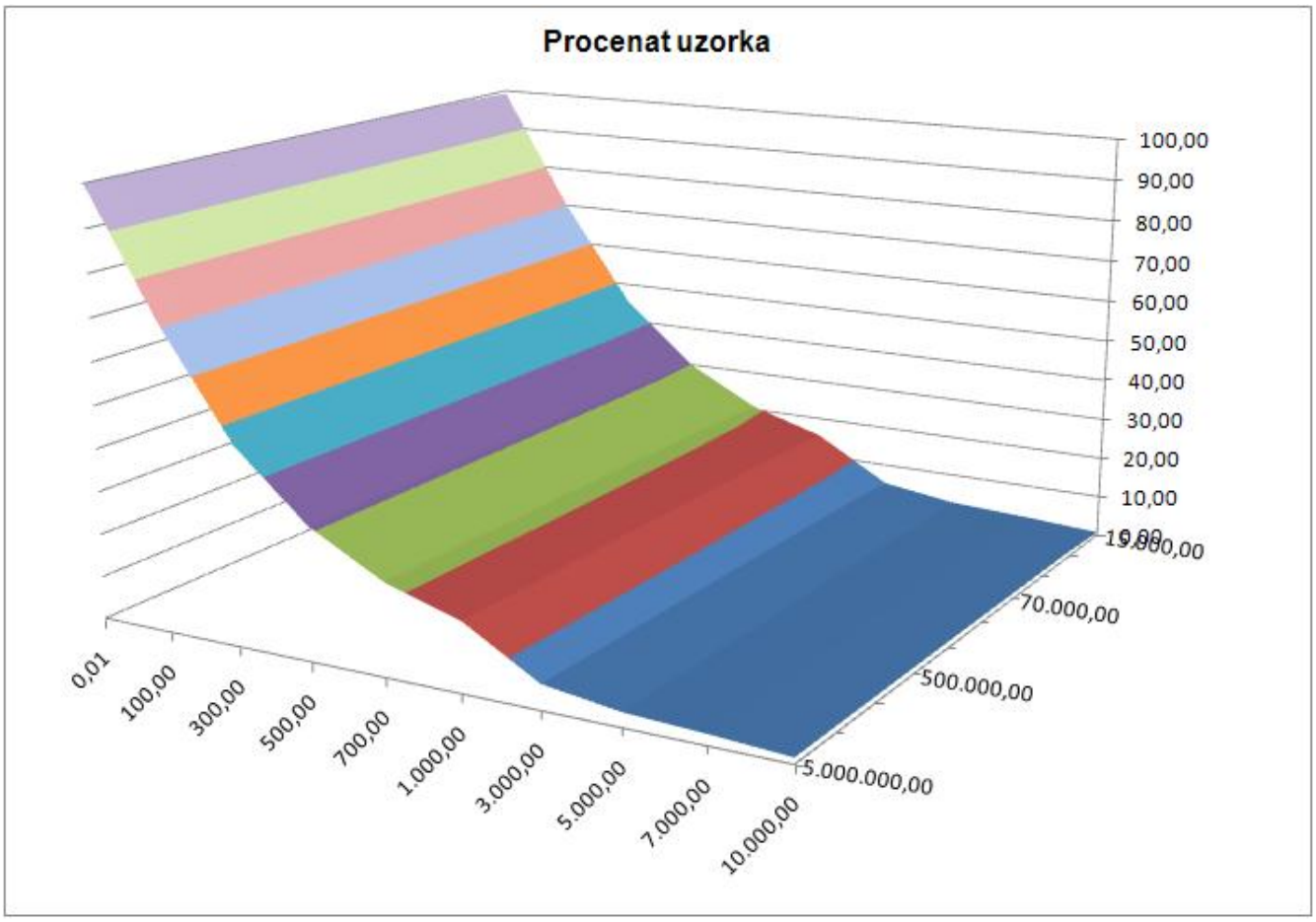
DF	15.000,00	30.000,00	50.000,00	70.000,00	100.000,00	300.000,00	500.000,00	700.000,00	1.000.000,00	5.000.000,00
0,01	852,88	850,85	850,89	851,15	851,51	851,27	851,34	851,34	851,34	851,37
100,00	769,77	767,49	767,67	768,09	768,67	768,39	768,50	768,50	768,50	768,56
300,00	995,45	989,48	989,28	989,69	990,30	989,63	989,71	989,71	989,71	989,75
500,00	1.064,26	1.054,18	1.053,68	1.054,19	1.054,97	1.053,88	1.053,96	1.053,96	1.053,96	1.054,00
700,00	935,55	924,07	923,94	924,97	926,46	925,18	925,41	925,41	925,41	925,52
1.000,00	574,79	570,78	572,76	575,21	578,45	577,80	578,53	578,53	578,53	578,89
3.000,00	1.120,34	1.060,07	1.057,23	1.059,95	1.064,08	1.058,21	1.058,62	1.058,62	1.058,62	1.058,82
5.000,00	1.365,63	1.217,30	1.207,42	1.209,95	1.214,53	1.202,10	1.201,97	1.201,97	1.201,97	1.201,90
7.000,00	1.382,62	1.176,81	1.165,17	1.169,43	1.176,48	1.160,74	1.160,89	1.160,89	1.160,89	1.160,96
10.000,00	202,68	308,41	354,33	392,74	439,15	436,57	446,75	446,75	446,75	451,74



Grafikon 6.11. Faktor distorzije za različite vrijednosti pragova. Bitno je primijetiti periodični izgled u odnosu na redove veličina

Obim uzorka u skladu sa izborom granica

Procenat	15.000,00	30.000,00	50.000,00	70.000,00	100.000,00	300.000,00	500.000,00	700.000,00	1.000.000,00	5.000.000,00
0,01	99,22	99,72	99,81	99,86	99,92	99,97	99,99	99,99	99,99	100,00
100,00	69,87	70,36	70,46	70,51	70,57	70,62	70,64	70,64	70,64	70,65
300,00	45,25	45,75	45,85	45,89	45,95	46,00	46,02	46,02	46,02	46,03
500,00	30,03	30,52	30,62	30,67	30,73	30,78	30,80	30,80	30,80	30,81
700,00	20,71	21,21	21,31	21,36	21,41	21,46	21,48	21,48	21,48	21,49
1.000,00	15,41	15,91	16,00	16,05	16,11	16,16	16,18	16,18	16,18	16,19
3.000,00	5,08	5,58	5,67	5,72	5,78	5,83	5,85	5,85	5,85	5,86
5.000,00	2,60	3,10	3,19	3,24	3,30	3,35	3,37	3,37	3,37	3,38
7.000,00	1,78	2,28	2,37	2,42	2,48	2,53	2,55	2,55	2,55	2,56
10.000,00	0,66	1,16	1,26	1,31	1,37	1,42	1,44	1,44	1,44	1,45



Grafikon 6.12. Procenat uzorka u skladu sa promjenama vrijednosti pragova

7 Reinforcement učenje

7.1 Uvod

Ideja učenja putem interakcije sa okruženjem vjerovatno je prvo što o čemu se razmišlja kada je u pitanju priroda učenja [8]. Dijete koje se igra, maše rukama ili gleda okolo nema eksplicitnog učitelja već direktnu senzomotornu vezu sa okolinom. Iskušavanje te veze producira obilje informacija o uzroku, efektu i posljedicama akcija i o tome šta uraditi kako bi se postigli ciljevi. Učenje iz interakcija je fundamentalna ideja u osnovi skoro svih teorija učenja i inteligencije.

Reinforcement learning je učenje šta uraditi, kako mapirati situaciju u akcije tako da se maksimizira numerička vrijednost signala odziva (reward). Onome koji uči ne govori se koje akcije treba preduzeti kao u većini formi mašinskog učenja. Umjesto toga, putem pokušaja on mora otkriti koje akcije daju najveći odziv. U najinteresantnijim i najizazovnijim slučajevima, akcije mogu uticati ne samo na neposredni odziv već i na narednu situaciju i tako na naredne odzive. Ove dvije karakteristike, metoda pokušaja i greške i odgođeni odziv, dva su najvažnija i najistaknutija svojstva metode reinforcement učenja.

Termin *reinforcement* se prevodi kao *ojačanje*, *armatura* i slično. U tom smislu bi se termin *reinforcement learning* mogao prevesti sa *učenje pojačavanja, učenje ubrzavanja* ili slično. Iz praktičnih razloga, u tekstu će biti zadržan termin *reinforcement učenje*.

Ključni termin u teoriji i praksi reinforcement učenja je *reward*. Prevedeno, ovaj termin označava *nagradu*, *naknadu* i slično. U cijelom tekstu se kao prikladna zamjena ovog prevoda koristi termin *odziv*. Ovakav izbor je napravljen jer je u tom terminu sadržano tehničko i praktično značenje termina *reward* u kontekstu reinforcement učenja odnosno interaktivnog odnosa agenta i okruženja.

Reinforcement učenje se razlikuje od nadziranog učenja, koje je predmet studija u velikom dijelu istraživanja metoda mašinskog učenja, prepoznavanja statističkih obrazaca i vještačkih neuronskih mreža [8]. Nadzirano učenje je učenje iz primjera na osnovu znanja od strane eksternog supervizora. Ovo je važna metoda učenja ali sama po sebi nije adekvatna za učenje iz interakcija. U interaktivnim problemima često je nepraktično pribavljati primjere željenog ponašanja koje je ispravno ili je reprezentativno za sve situacije u kojima agent mora djelovati. Na nepoznatom terenu, gdje bi se moglo očekivati najuspješnije učenje, agent mora biti sposoban učiti iz sopstvenog iskustva.

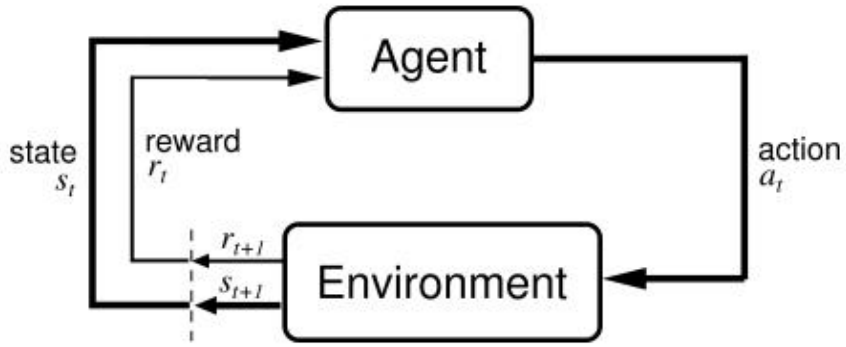
Jedan od izazova koji se javlja vezano za metodu reinforcement učenja, kojeg nema u drugim tipovima učenja, jeste balans između istraživanja (exploration) i korištenja (exploitation). Da bi pribavio veći odziv agent mora biti skloniji akcijama koje je probao u prošlosti a za koje je ustanovio da su efektivni u smislu odziva. Da bi ih otkrio mora birati akcije koje nije birao ranije. Agent mora eksploatisati ono što već zna kako bi dobio odziv ali, također, mora istraživati kako bi napravio bolju akciju u budućnosti. Dilema je u tome što ni istraživanje ni eksploatacija ne mogu biti ekskluzivno napravljeni bez propusta u zadatku. Agent mora probati više akcija i progresivno favorizirati one za koje ustanovi da su najbolje. Stohastički gledano, svaka akcija se mora napraviti više puta kako bi se dobila pouzdana procjena očekivanog odziva. Dilema istraživanje - korištenje predmet je intenzivnog studiranja.

Drugo ključno svojstvo metode reinforcement učenja je da eskplicitno uzima u obzir *cijeli* problem ciljno orijentisanog agenta koji je u interakciji sa nepoznatim okruženjem. Ovo je u suprotnosti sa pristupima koji posmatraju podprobleme bez naznake kako se oni uklapaju u veću sliku. Naprimjer, veći dio istraživanja mašinskog učenja se bavi nadziranim učenjem a da se pritom eksplicitino ne ukazuje kako će ta mogućnost na kraju biti iskorištena.

7.2 Elementi reinforcement učenja

Agent je onaj koji uči i donosi odluke [8]. Ono sa čim je agent u interakciji, podrazumijevajući sve van njega, zove se okruženje. Oni su u interakciji kontinuirano, agent odabire akcije a okruženje odgovara na njih i prezentira mu nove situacije. Okruženje povećava odzive, numeričke vrijednosti koje agent nastoji maksimizirati tokom vremena. Kompletna specifikacija okruženja definiše zadatak, jednu instancu problema reinforcement učenja.

Specifično, agent i okruženje su u interakciji na svakoj sekvenci diskretnih vremenskih koraka $t = 1, 2, 3, ...$ U svakom vremenskom momentu t agent dobija prezentacije stanja okruženja $s_t \in \mathcal{S}$, gdje je $\mathcal{S}$ skup mogućih stanja i na toj osnovi odabire akciju $a_t \in \mathcal{A}$, gdje je $\mathcal{A}$ skup akcija raspoloživih u stanju s_t. Jedan vremenski korak dalje, kao posljedica te akcije, agent dobija numerički odziv $r_{t+1} \in \mathfrak{R}$ i prelazi u novo stanje s_{t+1}. Slika 7.1 prikazuje ovu interakciju.



Slika 7.1. Interakcija agenta i okruženja (Izvor : Richard S. Sutton, Andrew G. Barto, Reinforcement Learning: An Introduction,)

Osim agenta i okruženja mogu se identifikovati četiri glavna elementa sistema reinforcement učenja : politika, funkcija odziva, funkcija vrijednosti i, opcionalno, model okruženja.

Politika definiše način agentovog ponašanja u datom vremenu. Grubo govoreći, politika je mapiranje uočenih stanja okruženja na akcije koje se trebaju preduzeti u tim stanjima. Označava se sa π_t gdje je $\pi_t(s, a)$ vjerovatnoća da se odabere akcija $a = a_t$ ako je $s = s_t$. U nekim slučajevima politika može biti jednostavna funkcija ili tabela dok u drugim može uključivati intenzivne kalkulacije i korištenje računara kao što je proces pretraživanja. Politika je suština agenta reinforcement učenja u smislu da je sama po sebi dovoljna da odredi ponašanje.

Funkcija odziva definiše cilj problema reinforcement učenja. Ona mapira svako uočeno stanje (ili par stanje - akcija) okruženja na jedan broj, odziv, koji indicira unutrašnju poželjnost stanja. Cilj agenta reinforcement učenja je maksimizirati ukupan odziv dugoročno. Funkcija odziva definiše šta su za agenta dobri i loši događaji. Odzivi su neposredna svojstva koja se mogu definisati za problem sa kojim se agent suočava. Ako akcija odabrana u skladu sa politikom rezultira niskim odzivom tada se politika može izmijeniti kako bi se odabrala neka druga akcija u takvoj situaciji u budućnosti. Generalno gledano, funkcije odziva mogu biti stohastičke.

Funkcija vrijednosti specificira šta je dobro u dugoročnom smislu, dok funkcija odziva indicira šta je dobro u neposrednom smislu. Grubo govoreći, vrijednost stanja je ukupan iznos povrata za kojeg agent može očekivati da bude akumuliran u budućnosti, počev od nekog stanja. Dok odziv određuje neposrednu, unutrašnju poželjnost stanja, vrijednosti stanja indiciraju dugotrajnu poželjnost stanja nakon uzimanja u obzir stanja koja se trebaju slijediti i odziva raspoloživih u tim stanjima. Naprimjer, stanje može uvijek davati odziv niskog nivoa ali još uvijek ima veliku vrijednost jer se pravilno slijedi od strane drugih stanja koja daju visoke odzive. Može se desiti i obratno.

Model okruženja je četvrti element nekih sistema reinforcement učenja. To je ono što simulira ponašanje okruženja. Naprimjer, za dato stanje i datu akciju model može predvidjeti rezultujuće naredno stanje i naredni odziv. Modeli se koriste za planiranje, pod kojim se podrazumijeva bilo koji oblik odlučivanja o smjeru akcija razmatranjem mogućh budućih situacija prije nego se one stvarno iskuse. Uključivanje modela i planiranja u sisteme reinforcement učenja je relativno novi razvoj. Postepeno postaje jasno da su metode reinforcement učenja u bliskoj vezi sa metodama dinamičkog programiranja koje koriste modele i koje su blisko povezane sa metodama planiranja stanje - prostor. Moderno reinforcement učenje proširuje spektar sa učenja niskog nivoa putem pokušaja i greške na učenje visokog nivoa, putem promišljenog planiranja.

7.3 Formalni okvir

7.3.1 Svojstvo Markova

U reinforcement učenju agent donosi odluke kao funkciju signala iz okruženja koji se zove stanje [8]. Pod pojmom *stanje* se smatra bilo koja informacija raspoloživa agentu. Pretpostavlja se da je stanje dato putem nekog sistema preprocesiranja koji je dio okruženja.

Za signal stanja koji uspijeva zadržati relevantne informacije kaže se da ima svojstvo Markova. Naprimjer, trenutna pozicija u igri šaha ima svojstvo Markova jer sumira sve važno o cijeloj sekvenci (redu poteza) koja je do nje dovela. Veći dio informacije o redu poteza je izgubljen ali sve bitno za nastavak igre je zadržano. Ovo se nekada naziva svojstvo 'nezavisnosti od putanje' jer sve bitno je u signalu stanja.

Kako bi se održao jednostavan matematički model polazi se od pretpostavke da postoji konačan broj stanja i vrijednosti odziva. Ovo omogućava rad u pojmovima suma i vjerovatnoća umjesto integrala i gustina vjerovatnoća ali se model može proširiti da uključi neprekidna stanja i odzive. U najopštijem slučaju, okruženje u momentu $t + 1$ odgovara na akciju u momentu t a to može zavisiti od svega što se desilo ranije. U ovom slučaju dinamika može biti definisana specificiranjem kompletne distribucije vjerovatnoća

$$P\{s_{t+1} = s', r_{t+1} = r | s_t, a_t, t_t, s_{t-1}, a_{t-1,}..., r_1, s_0, a_0\} \qquad (7.1)$$

za sve s' i r i sve moguće vrijednosti prethodnih događaja $s_t, a_t, t_t, s_{t-1}, a_{t-1,}..., r_1, s_0, a_0$. Sa s' je označeno bilo koje naredno stanje. Ako signal stanja ima svojstvo Markova tada odgovor okruženja u momentu $t+1$ zavisi samo od stanja i prezentacije akcija u momentu t. U tom slučaju dinamika okruženja može biti definisana samo sa

$$P\{s_{t+1} = s', r_{t+1} = r | s_t, a_t, t_t\} \qquad (7.2)$$

za sve s', r, s_t i a_t. Drugim riječima, signal stanja ima svojstvo Markova ako i samo ako je (7.1) jednako sa (7.2) za sve s', r i istorije $s_t, a_t, r_t, ..., r_1, s_0, a_0$. U tom slučaju se za okuženje i zadatak u cjelini takođe kaže da imaju svojstvo Markova.

Formalno gledano, proces odlučivanja markova (MDP - Markov Decision Process) se sastoji od sljedećih elemenata [62] :

- Skup stanja $\mathcal{S}$: prostor svih mogućih stanja. Može biti diskretan ($\mathcal{S} \in \mathbb{N}$), neprekidan ($\mathcal{S} \in \mathbb{R}^n$) ili njihova mješavina
- Prostor stanja $\mathcal{A}$: prostor svih akcija koje agent može odabrati. Može biti diskretan, neprekidan ili njihova mješavina
- Funkcija numeričkog odziva $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- Funkcija tranzicije stanja $f : \mathcal{S} \times \mathcal{A} \to S$
- Inicijalna distribucija stanja $d : \mathcal{S} \to [0, 1]$ nad prostorom stanja

Ako okruženje ima svojstvo Markova tada dinamika jednog koraka (7.2) omogućava da se predvidi sljedeće stanje i očekivani sljedeći odziv dat tekućim stanjem i akcijom. Iteracijom te jednadžbe mogu se predvidjeti sva buduća stanja i očekivani odzivi iz poznavanja samo tekućeg stanja kao što bi to bilo moguće na osnovu cjelokupno date istorije do tog momenta. Iz ovoga slijedi da stanja Markova daju najbolju moguću osnovu za izbor akcija. To znači da je najbolja politika izbora akcija kao funkcija stanja Markova dobra onoliko koliko i najbolja politika izbora akcija kao funkcija kompletne istorije.

Svojstvo Markova je važno u reinforcement učenju jer se podrazumijeva da su odluke i vrijednosti funkcije samo trenutnog stanja [8]. Kako bi one bile efektivne i informativne, prezentacija stanja mora biti informativna. Pretpostavka prezentacije stanja Markova nije jedinstvena za reinforcement učenje već je prisutna u većini ako ne i u svim drugim pristupima vještačke inteligencije. Najveći broj algoritama zahtijeva svojstvo Markova za njihovu dokazanu konvergenciju optimalne politike ali još uvijek rade korektno ako svojstvo Markova nije drastično prekršeno [62].

Zadatak reinforcement učenja koji zadovoljava svojstvo Markova zove se Proces odlučivanja Markova ili MDP (Markov Decision Process) [8]. Ako je prostor stanja konačan tada se on zove Konačni proces odlučivanja Markova i veoma je važan za teoriju reinforcement učenja.

Poseban proces odlučivanja Markova je definisan svojim stanjem i dinamikom jednog koraka. Za bilo koje dato stanje s i akciju a vjerovatnoća svakog narednog stanja s' je data sa

$$\mathcal{P}^a_{ss'} = P\{s_{t+1} = s' | s_t = s, a_t = a\}$$

Ove veličine se zovu *vjerovatnoće tranzicija*. Slično, za bilo koje stanje i akciju s i a zajedno sa narednim stanjem s' očekivana vrijednost narednog odziva je

$$\mathcal{R}^a_{ss'} = E\{r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\}$$

Ove veličine, $\mathcal{P}^a_{ss'}$ i $\mathcal{R}^a_{ss'}$, kompletno specificiraju najvažnije aspekte dinamike konačnog procesa odlučivanja Markova; izgubljena je samo informacija o distribuciji odziva oko očekivane vrijednosti.

7.3.2 Povrat

Agentov cilj je maksimizirati povrat koji dugoročno dobija. Ako je niz odziva nakon vremenskog momenta t označen sa $r_{t+1}, r_{t+2}, r_{t+3}, ...$ generalno gledano, traži se maksimiziranje očekivanog povrata R_t koji je definisan kao specifična funkcija sekvenci odziva. U najjednostavnijem slučaju povrat je suma odziva

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + ... + r_T \qquad (7.3)$$

gdje je T konačni vremenski korak. Ovaj pristup ima smisla u aplikacijama u kojima prirodno postoji konačni vremenski moment tj. gdje se interakcija agent - okruženje prirodno cijepa na podsekvence koje se zovu *epizode*, kao što je jedna igra, put kroz mrežu (maze) ili bilo koja vrsta ponavljajućih interakcija. Svaka epizoda se završava u posebnom stanju koje se zove terminalno stanje, nakon čega slijedi reset na standardno početno stanje ili na uzorak iz standardne distribucije startnih stanja. Zadaci ove vrste zovu se epizodni poslovi. U njima nekada treba razlikovati skup neterminalnih stanja, u oznaci $\mathcal{S}$, od skupa svih stanja plus terminalno stanje, u oznaci $\mathcal{S}^+$.

Dodatni koncept je diskontovanje. U skladu sa tim pristupom, agent nastoji izabrati akcije tako da se maksimizira suma diskontovanih odziva koje će dobijati u budućnosti. Posebno, odabire maksimizaciju očekivanog diskontovanog povrata

$$R_t = r_{t+1} + \gamma t_{t+2} + \gamma^2 r_{t+3} + ... = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \qquad (7.4)$$

gdje je γ parametar, $0 \leq \gamma \leq 1$, koji se zove *faktor diskontovanja*. On određuje trenutnu vrijednost budućih odziva : odziv dobijen u k vremenskih momenata vrijedan je γ^{k-1} puta od onoga što bi vrijedio da je dobijen trenutno. Ako je $\gamma < 1$ beskonačna suma ima konačnu vrijednost sve dok je niz odziva $\{r_k\}$ ograničen. Ako je $\gamma = 0$ agent je 'kratkovidan', zabrinut samo da maksimizira neposredne odzive; njegov cilj u tom slučaju je naučiti kako izabrati a_t kako bi se maksimizirao samo r_{t+1}. Kako se γ približava ka 1 cilj jače uzima u obzir buduće odzive; agent gleda u sve dalju budućnost (far-sighted).

7.3.3 Funkcije vrijednosti stanja

Skoro svi algoritmi reinforcement učenja su bazirani na procjeni funkcija vrijednosti - funkcija stanja (ili parova stanje - akcija) kojom se procjenjuje koliko je za agenta dobro da bude u datom stanju odnosno koliko je dobro da preduzme akciju u datom stanju. Napomena *koliko dobro* je definisana u pojmovima očekivanih odziva. Naravno, odzivi koje agent može očekivati u budućnosti zavise od akcija koje će preduzeti. U skladu sa tim, funkcije vrijednosti su definisane poštujući posebne politike.

Politika π je mapiranje iz svakog stanja $s \in \mathcal{S}$ i akcije $a \in \mathcal{A}(s)$ na vjerovatnoću $\pi(s,a)$ ako se preduzme akcija a u stanju s. Neformalno, vrijednost stanja s u politici π, u oznaci $V^{\pi}(s)$, je očekivani povrat sa početkom u stanju s kada se slijedi politika π. Za proces odlučivanja Markova veličina $V^{\pi}(s)$ se formalno definiše sa

$$V^{\pi}(s) = E_{\pi}\{R_t | s_t = s\} = E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\right\} \qquad (7.5)$$

gdje $E_{\pi}\{\cdot\}$ označava očekivanu vrijednost kada agent slijedi politiku π a t bilo koji vremenski moment. Vrijednost terminalnog stanja, ako postoji, je uvijek nula. Slično, vrijednost akcije a u stanju s pod politikom π, u oznaci $Q^{\pi}(s,a)$, je očekivani povrat koji počinje od s preduzimanjem akcije a kada se slijedi politika π.

$$Q^{\pi}(s,a) = E_{\pi}\{R_t | s_t = s, a_t = a\} = E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\right\} \qquad (7.6)$$

Veličina Q^{π} se zove funkcija vrijednosti akcije za politiku π.

Rekurzivne relacije su fundamentalno svojstvo funkcija vrijednosti koje se koriste u okviru reinforcement učenja i dinamičkog programiranja. Za bilo koju politiku π i stanje s vrijedi naredni uslov konzistencije između vrijednosti s i vrijednosti njegovih mogućih narednih stanja :

$$\begin{aligned}
V^{\pi}(s) &= E_{\pi}\{R_t | s_t = s\} \\
&= E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\right\} \\
&= E_{\pi}\left\{r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} | s_t = s\right\} \\
&= \sum_a \pi(s,a) \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} | s_{t+1} = s'\right\}\right] \\
&= \sum_a \pi(s,a) \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma V^{\pi}(s')\right] \qquad (7.7)
\end{aligned}$$

Ovdje se implicitno smatra da se akcije a preduzimaju iz skupa $\mathcal{A}(s)$, naredna stanja s' se uzimaju iz skupa $\mathcal{S}$ ili iz skupa $\mathcal{S}^+$ za slučaj epizodnog problema. Jednadžba (7.7) je Bellmanova jednadžba za V^{π}. Ona izražava odnos između vrijednosti tekućeg stanja

i vrijednosti njegovih narednih stanja. Ova jednadžba daje prosjek svih mogućnosti, uz ponderisanje svake od njih vjerovatnoćom pojavljivanja. To znači da vrijednost početnog stanja mora biti jednaka (diskontovanoj) vrijednosti očekivanog narednog stanja plus odzivi do tog momenta.

7.4 Temporal difference učenje

Postoje tri fundamentalne klase metoda za rješavanje problema reinforcement učenja : Dinamičko programiranje, Monte Carlo metode i Temporal difference učenje [8]. Svaka od ovih klasa metoda ima svoje prednosti i nedostatke. Metode dinamičkog programiranja su dobro dobro podržane matematički ali zahtijevaju tačan i potpun model okruženja. Monte Carlo metode ne zahtijevaju model i konceptualno su jednostavne ali nisu pogodne za inkrementalna izračunavanja korak-po-korak. Temporal difference metode ne zahtijevaju model, u potpunosti su inkrementalne ali su kompleksne za analizu. Ove metode se razlikuju u više aspekata po pitanju efikasnosti i konvergencije.

Termin *temporal difference* se prevodi kao t*emporalna razlika*. Iz praktičnih razloga, u tekstu je zadržan izvorni termin *temporal difference*.

Odzivi su primarni dok su vrijednosti, kao predikcije odziva, sekundarne [8]. Bez odziva ne može biti vrijednosti a jedina namjena procjena vrijednosti je dobiti veće odzive. Bez obzira na to, vrijednosti su ono što je najbitnije sa stanovišta donošenja i evaluacije odluka. Izbori akcija se prave na osnovu procjena vrijednosti. Traže se akcije koje vode ka stanjima najveće vrijednosti (ne najveće odzive) jer te akcije donose najveću veličinu odziva u dugoročnom smislu. Nažalost, mnogo je teže odrediti vrijednosti nego odzive. Odzivi su osnova koja se dobija direktno iz okruženja ali vrijednosti mogu biti procijenjene i ponovo procijenjene iz sekvenci opservacija koje agent pravi tokom svog životnog ciklusa.

Ustvari, najvažnija komponenta algoritama reinforcement učenja je metod efikasne procjene vrijednosti.

7.4.1 Teorijski osnov

Temporal difference učenje je kombinacija ideja metoda Monte Carlo i Dinamičkog programiranja [8]. Slično dinamičkom programiranju, TD metoda procjene ažurira na osnovu drugih procjena, ne čekajući na konačni izlaz; one su samopunjeće - bootstrap. Zajedničko za ove metode je njihova rekurentna priroda i u pravilu se sastoje od dva koraka. Prvi je problem predikcije ili evaluacije politike odnosno procjena funkcije vrijednosti za datu politiku. Drugi je problem upravljanja odnosno nalaženja optimalne politike.

I TD i Monte Carlo metode koriste iskustvo za rješenje problema predikcije. Za iskustvo koje slijedi politiku π obje metode ažuriraju procjene V za V^{π}. Ako je neterminalno stanje s_t posjećeno u momentu t tada obje metode ažuriraju svoje procjene $V(s_t)$ na osnovu onoga što se dešava nakon te posjete. Grubo govoreći, Monte Carlo metode čekaju sve dok nije poznat povrat nakon posjete a zatim taj povrat koriste kao cilj za $V(s_t)$. Jednostavan Monte Carlo metod pogodan za nestacionarno okruženje je

$$V(s_t) \leftarrow V(s_t) + \alpha [R_t - V(s_t)] \qquad (7.8)$$

gdje je R_t realizovani povrat u momentu t, α konstantni parametar koraka. Monte Carlo metode moraju čekati do kraja epizode kako bi se odredio inkrement za $V(s_t)$ (tek tada je R_t poznat). TD metode trebaju čekati samo sljedeći vremenski moment. U momentu $t+1$ automatski se formira cilj i kreira korisno ažuriranje koristeći uzoračku odziv r_{t+1} i procjenu $V(s_{t+1})$. Najjednostavnija TD metoda, poznata kao $TD(0)$ je

$$V(s_t) \leftarrow V(s_t) + \alpha [r_{t+1} + \gamma V(s_{t+1}) - V(s_t)] \qquad (7.9)$$

U suštini, cilj ažuriranja u Monte Carlo metodi je R_t dok je u TD cilj ažuriranja $r_{t+1} + \gamma V_t(s_{t+1})$. S obzirom da TD metoda svoje ažuriranje bazira na dijelu postojeće procjene kaže se da je to bootstraping metoda, kao DP. Poznato je da vrijedi

$$\begin{aligned} V^\pi(s) &= E_\pi \{R_t | s_t = s\} \qquad (7.10) \\ &= E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s \right\} \\ &= E_\pi \left\{ r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} | s_t = s \right\} \\ &= E_\pi \{ r_{t+1} + \gamma V^\pi(s_{t-1}) | s_t = s \} \qquad (7.11) \end{aligned}$$

Grubo govoreći, Monte Carlo metode kao cilj koriste procjenu za (7.10) a DP metode procjenu (7.11). TD metode formiraju uzorak očekivane vrijednosti u (7.11) i tekuću procjenu za V_t koriste umjesto stvarne V^π. Stoga, TD metode kombinuju Monte Carlo uzorkovanje sa samopunjenjem (bootstraping) za DP.

Prva prednost TD metoda je u tome što ne zahtijevaju model okruženja, njegovih odziva i distribucija vjerovatnoće narednih stanja. Druga prednost TD metoda je da se veoma prirodno implementiraju on-line, na potpuno inkrementalni način. Sa Monte Carlo metodama se mora čekati dok se epizoda ne završi, dok se sa TD metodama treba čekati samo jedan korak.

Za bilo koju fiksiranu politiku π dokazano je da ranije opisan TD algoritam konvergira ka V^π u pojmovima konstantnog parametra koraka ako je on dovoljno malen i sa vjerovatnoćom 1 ako parametar koraka opada u skladu sa uobičajenim uslovima stohastičke aproksimacije

$$\sum_{k=0}^{\infty} \alpha_k(a) = \infty \qquad \sum_{k=0}^{\infty} \alpha_k^2(a) < \infty$$

Ovdje je α_k parametar veličine koraka za vremenski moment k i akciju a.

7.4.2 SARSA on-policy učenje

SARSA je algoritam upravljanja koji generalizira TD učenje. Ažuriranje se ne obavlja samo na trajektoriji (s_t, r_t, s_{t+1}) već na $(s_t, a_t, r_t, s_{t+1}, a_{t+1})$ po čemu je metoda i dobila naziv (state - action - reward - state - action) [60]. Prvi korak je naučiti funkciju vrijednosti akcije umjesto funkcije vrijednosti stanja. On-policy metoda pokušava evaluirati ili unaprijediti politiku koja se koristi za donošenje odluka.

Za dati par stanje - akcija (s_t, a_t) SARSA simulira akciju a_t u stanju s_t kako bi se našao odziv r_t i stanje u tranziciji s_{t+1}. Algoritam zatim koristi tekuću optimalnu politiku baziranu na Q vrijednostima, kako bi generisao narednu akciju a_{t+1} (akciju odabire slučajno sa vjerovatnoćom ε). U ovoj tački SARSA ažurira $Q(s_t, a_t)$ na sljedeći način

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha_k [r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \tag{7.12}$$

Ažuriranje se vrši nakon svake tranzicije iz neterminalnog stanja s_t. Ako je stanje s_{t+1} terminalno tada se $Q(s_{t+1}, a_{t+1})$ uzima kao nula. Ovo pravilo koristi svaki element n-torke $(s_t, a_t, r_t, s_{t+1}, a_{t+1})$ čime se pravi tranzicija od jednog para stanje-akcija na drugi. Ovo pravilo ažuriranja je zasnovano na sljedećoj varijanti Bellmanove jednadžbe optimalnosti

$$Q^*(s_t, a_t) = R(s_t, a_t) + \gamma \mathbb{E}[Q^*(s_{t+1}, \pi^*(s_{t+1}))] \tag{7.13}$$

gdje je

$$\pi^*(s_t) \in \arg\max_a Q^*(s_{t+1}, a) \tag{7.14}$$

7.4.3 Q off-policy učenje

Jedno od najvažnijih napredaka u reinforcement učenju je razvoj off-policy TD kontrolnog algoritma poznatog kao Q-učenje (Watkins, 1989). U najprostijoj formi jednokoračno Q-učenje se definiše sa

$$Q^*(s_t, a_t) = R(s_t, a_t) + \gamma E\left[\max_a Q(s_{t+1}, a)\right] \tag{7.15}$$

Odgovarajuće pravilo ažuriranja, koje je osnova Q-učenja, je

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha_k \left[r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right] \tag{7.16}$$

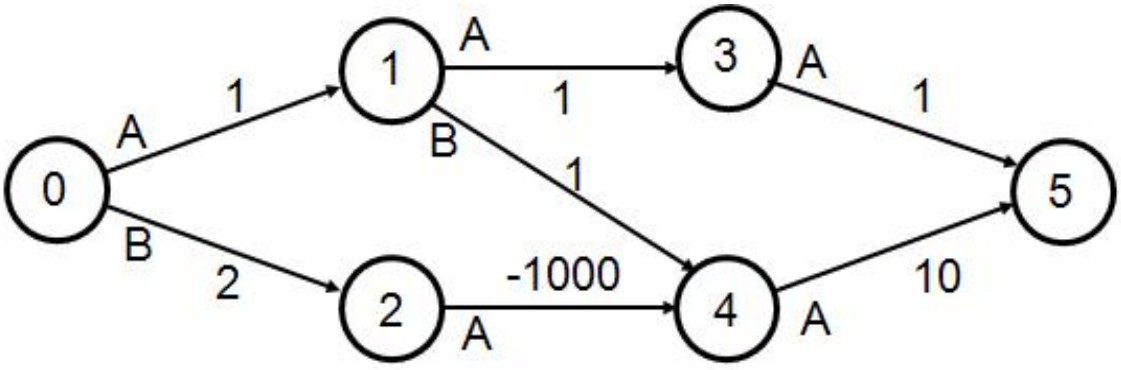
U off-policy metodama funkcije razdvojene su politike procjene i korištenja [8]. Politika koja se koristi za generisanje ponašanja, nazvana politika ponašanja, može biti nepovezana sa politikom koja se koristi za evaluacije i poboljšanja, nazvana politika procjene. Prednost ovakvog pristupa je da politika procjene može biti deterministička (npr. pohlepna) dok politika ponašanja može nastaviti da pravi uzorke svih mogućih akcija.

U ovom slučaju naučena funkcija vrijednosti akcija Q direktno aproksimira Q^*, optimalnu funkciju vrijednosti akcija, nezavisno od politike koja se slijedi. Ovo bitno pojednostavljuje analizu algoritma i omogućava brze konvergencije. Politika ima efekt u kojem se određuje koji su parovi stanja-akcija posjećeni i ažurirani. Međutim, sve što se zahtijeva za korektnu konvergenciju jeste da se ažuriranja nastave za sve parove. Agent ne može

procijeniti parove stanja - akcija koje nije posjetio. Ovo je minimalni zahtjev u smislu da bilo koji metod koji garantuje da će naći optimalno ponašanje u opštem slučaju mora ovo zahtijevati. Pod tom pretpostavkom i varijantom uobičajenih stohastičkih uslova aproksimacije za niz parametara veličine koraka za Q_t je pokazana konvergencija ka Q^* sa vjerovatnoćom 1.

Primjer. Na grafikonu 7.2 je ilustracija problema u kojem agent treba izabrati optimalnu putanju od stanja 0 do stanja 5 [61]. Sa A i B su označene moguće varijante kretanja u situacijama kada agent treba napraviti izbor. Moguće su tri politike

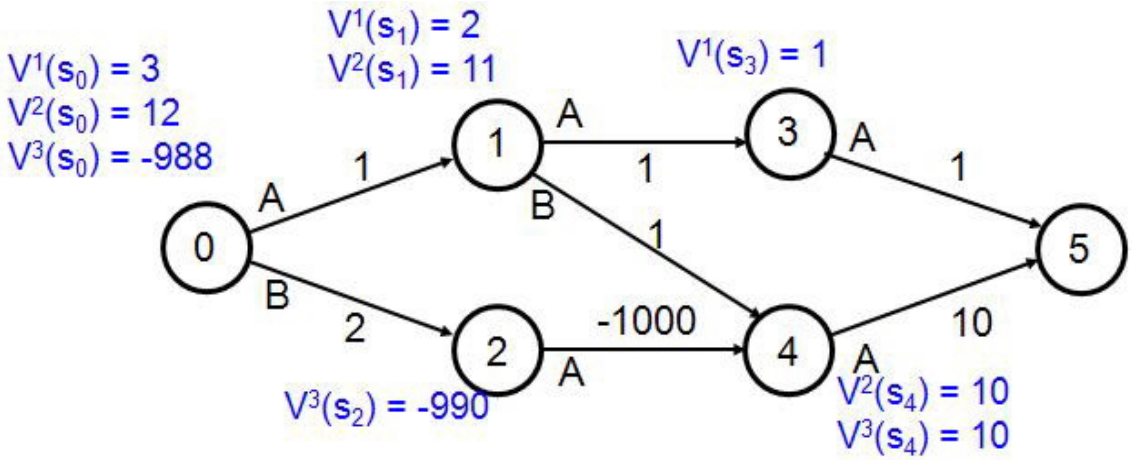
1. $0 \to 1 \to 3 \to 5$
2. $0 \to 1 \to 4 \to 5$
3. $0 \to 2 \to 4 \to 5$



Slika 7.2. Primjer mreže i mogućih politika na mreži (izvor : Bill Smart, Reinforcement Learning : A User's Guide, Department of Computer Science and Engineering, Washington University in St. Louis)

Vrijednosti povrata koje agent može prikupiti ako slijedi ove putanje su

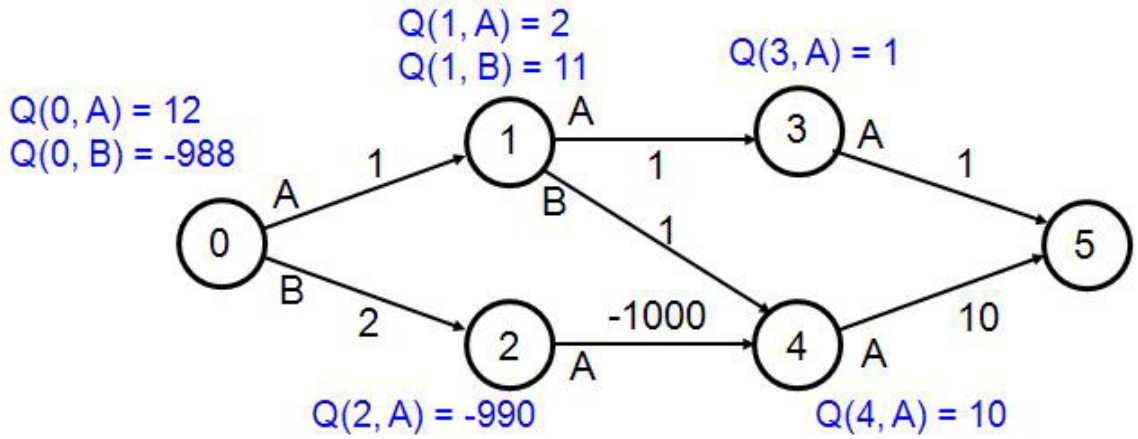
1. $0 \to 1 \to 3 \to 5 = 1 + 1 + 1 = 3$
2. $0 \to 1 \to 4 \to 5 = 1 + 1 + 10 = 12$
3. $0 \to 2 \to 4 \to 5 = 2 - 1000 + 10 = -988$



Slika 7.3. Primjer mreže i vrijednosti stanja u skladu sa odabranom politikom Pored svakog čvora su vrijednosti stanja za svaku moguću politiku ako se pođe iz tog stanja (izvor : Bill Smart, Reinforcement Learning : A User's Guide, Department of Computer Science and Engineering, Washington University in St. Louis)

Na slici 7.3 je ujedno moguće sagledati opštu šemu računanja vrijednosti. Tako, $V^1(s_0)$ označava vrijednost koja se dobije kada se startuje iz stanja s_0 i kad se slijedi politika 1 (putanja $0 \rightarrow 1 \rightarrow 3 \rightarrow 5$).

Ako se politika ne specificira tada se vrijednosti računaju na način koji je dat na slici 7.4. Vrijednosti stanja (u čvorovima) su najveće vrijednosti koje se mogu dobiti od tog stanja do kraja putanje.



Slika 7.4. Primjer mreže i vrijednosti akcija za pojedine politike (izvor : Bill Smart, Reinforcement Learning : A User's Guide, Department of Computer Science and Engineering, Washington University in St. Louis)

Osnovne karakteristike on-policy algoritama su [61]

- Konačna politika je zavisna od metode (politike) pretraživanja
- Generalno gledano, politika pretraživanja treba biti 'bliska' konačnoj politici
- Može se zaustaviti u lokalnom maksimumu

Osnovne karakteristike off-policy algoritama su

- Konačna politika je nezavisna od politike pretraživanja
- Mogu se koristiti proizvoljne politike pretraživanja
- Ne može se zaustaviti u lokalnom maksimumu

7.5 Benfordov zakon i reinforcement učenje

Priroda Benfordovog zakona fokus njegovog korištenja usmjerava na masovne skupove podataka, sa atributima koji zadovoljavaju odgovarajuće uslove. Tipični predstavnici takvih skupova su finansijske transakcije, berzanski indeksi, rezultati mjerenja nekog sistema i slično.

Osnovu za primjenu metoda reinforcement učenja u skupu podataka koji je predstavljen slogovima je koncept u kojem se slogovi posmatraju kao moguća stanja a kolone kao moguće akcije [5]. Da bi to bilo moguće svakom stanju treba biti dodijeljen odziv, vrijednost koja će ga karakterizirati. Benfordov zakon daje mogućnost definisanja takvih

veličina. S obzirom da ovaj zakon stipulira dinamiku značajnih cifara svaka veličina izvedena na osnovu njega nosi isto ili ekvivalentno značenje. U tom smislu, korištenje ovog zakona u metodama reinforcement učenja za rezultat, u pravilu, ima skup stanja koja slijede određeni obrazac (pattern) ponašanja posmatranog sistema.

Fletcher Lu i Efrim Boritz su u svojim radovima i patentnoj prijavi predložili dvije veličine na osnovu Benfordovog zakona, pogodne za skupove podataka koji se predstavljaju u obliku nizova zapisa (slogova) sa kolonama kao što su npr. finansijske transakcije.

Odziv (reward) za grupu stanja. Ako uzorak sadrži više numeričkih kolona koje zadovoljavaju uslove za analizu putem Benfordovog zakona, Fletcher Lu i Efrim Boritz [5] su u patentnoj prijavi predložili računanje veličine

$$R\left(s\right)=\sum_{cv}\sum_{\text{seq}}\frac{\left|P_{ocek}-P_{uzor}\right|}{P_{ocek}} \tag{7.17}$$

U ovom izrazu cv predstavlja kolone uzorka koje ulaze u računicu; veličina seq označava vodeće sekvence dužine tri; veličina P_{ocek} očekivanu (teorijsku) relativnu frekvenciju a P_{uzor} uzoračku relativnu frekvenciju prve tri cifre. Prema tekstu patenta, ovo daje zbirnu vrijednost za odabrani broj numeričkih atributa (cv). Izbor računanja za tri pozicije je prirodan ako se ima u vidu da su testovi prve tri cifre više fokusirani od testova za prve dvije cifre. Ovakva zbirna veličina je pogodna za mnoge metode data mininga kao što su mjere sličnosti, klastering itd. U metodama reinforcement učenja ova veličina se može koristiti za slučajeve kada se atributi iz bilo kog razloga ne mogu posmatrati odvojeno ili je takav pristup poželjan iz bilo kog razloga.

Odziv za jedno stanje. U tekstovima u kojima obrađuje primjer detekcije prevara korištenjem metoda reinforcement učenja [63, 64] Fletcher Lu predlaže korištenje veličine

$$BE\left(i\right)=\frac{f_{1i}}{b_{1i}}+\frac{f_{2i}}{b_{2i}}+\frac{f_{3i}}{b_{3i}} \tag{7.18}$$

U ovom izrazu f_{ji} predstavlja uzoračku a b_{ji} teorijsku frekvenciju grupa cifara dužine j za stanje (slog) i. Računa se za svako pojedino stanje predstavljeno kao numerički atribut, bilo da se koristi osnovna ili adaptivna metoda računanja. U nastavku će ova veličina, iz razloga praktične prirode, biti obilježena sa

$$BE\left(3\right)=\frac{f_1}{b_1}+\frac{f_2}{b_2}+\frac{f_3}{b_3} \tag{7.19}$$

Ovim se želi naglasiti kalkulacija za tri vodeće cifre. Po analogiji, moguće je praviti kalkulaciju više od tri vodeće cifre mada takva analiza. Ako se, po analogiji sa (7.19), formira veličina

$$BR\left(2\right)=\frac{f_1}{b_1}+\frac{f_2}{b_2} \tag{7.20}$$

moguće je računati količnik

$$BK32=\frac{BE\left(3\right)}{BE\left(2\right)}=1+\frac{f_3/b_3}{f_1/b_1+f_2/b_2} \tag{7.21}$$

Ovaj količnik mjeri uticaj treće cifre na način da je ovaj količnik veći za slogove u kojima frekvencija treće cifre bitno odstupa u smislu da je veća od teorijske. Svaki metod koji se oslanja na ovu činjenicu će imati mogućnost većeg fokusiranja na ovu grupu slogova. Ovakav pristup favorizira slogove sa povećanim frekvencijama na trećoj poziciji. Međutim, postoji interes da se za određene probleme u analizu uključe i slogovi za koje je frekvencija treće cifre bitno ispod teorijske. Kako bi se postigao ovaj efekat, Fletcher lu predlaže da se umjesto izraza (7.19) koristi

$$BE\,(3) = q_1 + q_2 + q_3 \qquad (7.22)$$

gdje je

$$q_i = \begin{cases} \dfrac{f_i}{b_i} & f_i > b_i \\ \dfrac{b_i}{f_i} & inače \end{cases}$$

U ovom tekstu će se koristiti isključivo izraz (7.19).

U skupu koji ne odražava regularni proces ili na kojem su pravljene transformacije kao što je odbacivanje stanja ispod donje ili iznad gornje odabrane granice, slogovi (stanja) za koje je $BE\,(3)$ najveći ne moraju biti isti oni za koje je $BK32$ najveći. Odstupanje je, u većini slučajeva, posebno značajno što je donja granica veća odnosno ako u uzorku, iz bilo kog razloga, nedostaje određena kategorija podataka. Ako je u pitanju odbacivanje vrijednosti ispod donje granice onda se to dešava zbog činjenice da se Benfordov zakon, u pravilu, koristi za skupove koji imaju veliki broj malih veličina. Ako se ovaj količnik koristi kao kriterij za izbor stanja tada se mijenja i politika reinforcement učenja; ona sigurno nije ista kao za slučaj veličina $BE(3)$. Ova činjenica je bila osnovna ideja eksperimenata.

Fletcher Lu sugeriše da učenje startuje u stanjima za koje je $BE\,(3)$ najveće. Eksperiment pokazuje da li se i u kojoj mjeri politike pretraživanja mijenjaju ako učenje startuje od stanja za koje je ova veličina iz odabranog opsega. Razlog za ovo je u činjenici da najveće vrijednosti mogu biti veoma bliske dok je istovremeno skup početnih stanja širi. Jedan od mogućih načina je da širina intervala bude izračunata putem standardnih devijacija.

Alternativa od interesa je da se kao početna uzmu ona stanja za koja je $BK32$ najveće. Dodatno, cilj je ispitati da li se i u kojoj mjeri politike pretraživanja mijenjaju ako učenje startuje u jednom od stanja za koje je ova veličina iz odabranog opsega vrijednosti, iz razloga koji su identični onima za $BE\,(3)$. Bitno je napomenuti da se ove varijante testiraju na istom skupu i sa istim parametrima kako bi se jasnije iskazale eventualne razlike u rezultatima. Jednake alternative se mogu primijeniti i na veličinu $R\,(s)$.

7.6 Eksperimenti

7.6.1 Hipoteza i priprema

Provedeni su eksperimenti kojima je ispitivan uticaj izbora početnih početnih stanja korištenjem veličina koje su izvedene na osnovu Benfodovog zakona na rezultujuću politiku reinforcement učenja. Hipoteza koja se provjerava je :

H_0:Izbor početnih uslova korištenjem veličina izvedenih na osnovu Benfordovog zakona ima uticaj na rezultujuću politiku reinforcement učenja

Eksperiment je proveden na više skupova podataka. Jedan uzorak je formiran na osnovu liste datoteka na laptopu. Ostali uzorci su primjeri finansijskih transakcija.

Eksperiment je u svim slučajevima obuhvatio sljedeće korake i aktivnosti :
I faza : Priprema

- Računanje veličina $BE(3)$ i $BK32$ za numerički atribut i izbor gornjih i donjih granica, u funkciji odziva za numerički atribut
- Računanje odziva za kategorijske atribute, kao odnos frekvencije pojedine vrijednosti i ukupnog broja stanja na nivou cijelog uzorka (relativna frekvencija)
- Računanje vjerovatnoća prelaza tako da se frekvencija svakog para stanja iz različitih kolona u jednom redu podijeli sa ukupnim brojem stanja (obimom uzorka).

II faza : Provođenje eksperimenta

- Demonstracija algoritma za opseg vrijednosti veličine $BE(3)$
- Demonstracija algoritma za opseg vrijednosti veličine $BK32$

U oba slučaja početni atribut je numerički atribut koji odgovara uslovima korištenja Benfordovog zakona.

III faza : Analiza Q vrijednosti putem Benfordovog zakona

IV faza : Diskusija rezultata
U okviru analize se daju tumačenja dobijenih rezultata i mogućih varijanti.

Donje i gornje granice uzorka su birane na način da slogovi koji imaju najveće vrijednosti za $BE(3)$ nisu isti oni za koje je $BK32$ najveći. Testiranja su provedena na način da su kao stanja uzimane prve tri pozicije, skupa sa odgovarajućim odzivima i to u dvije varijante. U jednoj varijanti je za odziv stanja uzimana veličina $BE(3)$ a u drugom $BK32$. Proširenje testa na prve četiri cifre omogućava da se provedu dodatna dva testa na istom uzorku.

U uzorcima koji su bili predmet testova datumi su rastavljeni na dan, dan u sedmici i mjesec. U slučaju finansijskih transakcija ovo može pružiti informaciju o obrascu ponašanja kao što je sklonost da se određeni tip ili visina transakcija provodi u određenom periodu. U slučaju datoteka na laptopu, uz dodatak podatka o vremenu, ovo može pružiti informaciju o obrascu određenih sistemskih i drugih funkcija (npr. sigurnosna ažuriranja) a posebno odstupanja od nekih poznatih obrazaca kao što su napadi virusa, neregularan rad softvera i slično. Kao kategorijski podaci može se uzeti bilo koji tip podatka koji je od interesa, npr. bankomati na kojima su obavljene transakcije i slično.

Eksperimenti su provedeni korištenjem programskog paketa Excell i samostalno kreiranih makroa koji su dizajnirani VBA alatima za ove potrebe. Excell je odabran jer daje brze rezultate, lako se upravlja parametrima, nije potreban poseban korisnički interfejs, nisu potrebne konverzije formata koje su potencijalni izvor rizika tačnosti, tako da je veća pažnja posvećena konceptu i algoritmu u cjelini.

7.6.2 Algoritam

Opšta šema algoritma koji je korišten za eksperimente je sljedeća :

- Inicijalizirati vrijednosti :
  - β : brzina učenja
  - γ : faktor diskonotovanja
  - ε : parametar pretraživanja
  - $nRun$: broj prolaza (obrada)
  - $nEpiz$: broj epizoda po jednoj obradi
- Za svaki od prolaza $i = 1, ..., nRun$ raditi
  - Ako je $i = 1$ Anulirati tabelu Q vrijednosti
  - Slučajno odabrati početnu akciju
  - U odabranoj akciji Izabrati vrijednost koja predstavlja početno stanje, po uslovu opsega frekvencija
  - Provoditi postupak po epizodama
- Za svaku iteraciju (epizodu) $j = 1, ..., nEpiz$ raditi
  - Ako je $j = 1$ tada

    Odabrati početni slog prema odabranom početnom stanju

    Inače

    Preuzeti početno stanje iz prethodne epizode
- Za odabrani slog raditi postupak izbora novog stanja i akcije
  - Generisati slučajan broj
  - Ako je slučajan broj manji od ε ili veći od $1 - \varepsilon$

    Narednu akciju izabrati na slučajan način

    Inače

    Provesti postupak za sve naredne kolone
- Za sve naredne kolone raditi
  - Odabrati akciju (kolonu) u koju nije vršen prelaz (raspoloživa akcija)
  - Pronaći najveću Q vrijednost za stanje u tekućoj koloni
  - U odabranoj koloni tražiti stanje s_{k+1} za koje je vjerovatnoća prelaza iz stanja s_k najveća
  - Izračunati Q vrijednost

$$Q(s,a) = (1-\beta)\cdot Q(s_k,a_k) + \beta\cdot\left(r_k + \gamma\cdot\max_{a'} Q(s_{k+1},a_{k+1})\right) \text{ odnosno}$$

$$Q(s,a) = Q(s_k,a_k) + \beta\cdot\left(r_k + \gamma\cdot\max_{a'} Q(s_{k+1},a_{k+1}) - Q(s_k,a_k)\right)$$

gdje je :

- – * $Q(s_k,a_k)$ Q vrijednost za stanje s_k (najveća Q vrijednost za akcije koje su raspoložive iz stanja s_k)
  * $Q(s_{k+1},a_{k+1})$ Q vrijednost za stanje s_{k+1} (najveća Q vrijednost za akcije koje su raspoložive iz stanja s_{k+1})
  * r_k : odziv za stanje s_k
  – Tekuću akciju uzeti kao narednu : $a_k \leftarrow a_{k-1}$

Na algoritmu je vidljivo da se Q vrijednosti resetuju nakon jednog prolaza koji se sastoji od odabranog broja iteracija. U makro je dodata mogućnost da se Q vrijednosti ne resetuju i eksperimenti su provedeni sa oba ova uslova. Jedan od motiva za pristup da se Q vrijednosti ne resetuju je mogućnost da se u prolazima koriste Q vrijednosti iz prethodnih koraka čime se može postići brža konvergencija. Ne postoji konflikt s obzirom da su matrice prelaza bazirane na parovima stanja iz različitih akcija (kolona). Kao generator slučajnih brojeva korištena je Excell funkcija rand(). Jednom odabrana akcija (kolona) ne može biti ponovo odabrana. U tom smislu, u algoritmu se koristi termin *raspoloživa stanja* odnosno ona stanja (kolone) koja nisu korištena u prethodnim koracima, čime se omogućava obavezan izlazak iz procesa izbora; uvijek postoji terminalno stanje mada ono nije poznato na početku. Ovaj proces je analogan sa postupcima analize podataka u kojima se obično polazi od jednog skupa atributa a zatim se dodatnim kriterijima po drugim atributima krug sužava na podatke koji su predmet posebne pažnje (npr. datumi za iznose koji počinju specifičnim sekvencama i slično).

Ključni moment primjene ovog zakona za reinforcement učenje je izbor skupa početnih stanja. U skladu sa metodologijom koju predlaže Fletcher Lu, postupak učenja počinje od numeričkog atributa i to od onih stanja (slogova) za koje je veličina $BE(3)$ najveća. S obzirom na pokazanu razliku u skupovima stanja koja odgovaraju ovim vrijednostima, u algoritam je ugrađena mogućnost odabira početnih stanja iz slogova za koje je odziv $BE(3)$ odnosno količnik $BK32$ iz odabranog opsega. Drugi oblik selekcije skupa početnih stanja je unos donje i gornje granice frekvencija. Ovim se postiže fokusiranje na različita početna stanja čime se, u suštini, mogu dobiti specifične politike. Svaki prolaz (skup iteracija) počinje odabirom novog početnog stanja. Početno stanje odabrano u prvoj iteraciji svakog prolaza se prenosi u narednu iteraciju.

Prva akcija se bira slučajnim izborom. Ostale akcije se biraju po principu ε pohlepnog pretraživanja tako da se slučajno generisani broj poredi sa zadatom veličinom ε. Ako je taj broj manji od ε ili veći od $1-\varepsilon$ tada se akcija izabire slučajnim izborom iz skupa preostalih akcija. Cilj ovakvog načina rada je da se postigne veća fluktuacija pri izboru akcije.

Makroi kreirani za potrebe ovog testa daju mogućnost eliminacije željenog atributa kako bi se ispitao uticaj i nivo promjena u rezultirajućoj politici.

Rezultat rada algoritma reinforcement učenja je politika. U ovom slučaju ona označava utvrđeni redoslijed selekcije kolona u postupku analize što se može posmatrati kao obrazac ponašanja.

Parametri kojima je ovaj algoritam određen su

- β : parametar brzine učenja
- γ : parametar diskontovanja
- ε : parametar pretraživanja;

Parametar $\varepsilon \in (0; 1)$ upravlja pretraživanjem na tako da se akcija odabire slučajno ako je slučajni broj manji od ε ili veći od $1 - \varepsilon$.. Ovaj način izbora akcija se naziva ε pohlepno pretraživanje. Izmjenom vrijednosti ovog parametra može se pratiti uticaj na politiku, vidljiv na grafikonima rezultujućih Q vrijednosti. Za potrebe ove analize parametar brzine učenja (β) je postavljen na $\beta = 0,8$, parametar diskontovanja (γ) je postavljen na $\gamma = 0,9$ a parametar pretraživanja (ε) je postavljen na $\varepsilon = 0,05$.

Svaki algoritam reinforcement učenja ima dva osnovna zadatka, istraživanje (exploration) i korištenje (exploitation). Balans između ova dva zadatka određuje kvalitet odabrane metode učenja. Algoritam koji je predmet ove analize, generalno gledano, ima faze :

- Istraživanje (exploration) : traženje početnog stanja, traženje stanja sa najvećom vjerovatnoćom prelaza
- Korištenje (exploitation) : izbor najbolje akcije iz odabranog stanja

Ova struktura odgovara algoritmu Q off-policy učenja. U ovom tipu algoritma politike pretraživanja i politike procjene su odvojene. Politika pretraživanja je sadržana u konceptu traženja početnog stanja (slučajan izbor akcije u prvom koraku, slučajan izbor stanja u prvom koraku i njegovo zadržavanje u svakom narednom koraku) i načinu traženja narednog stanja za odabranu akciju (traženje stanja sa najvećom vjerovatnoom prelaza). Politika procjene je sadržana u načinu traženja najbolje akcije (izbor po osnovu najveće Q vrijednosti, ažuriranje Q vrijednosti, način na koji se mijenja brzina učenja).

Testovi su provedeni na način da se posmatraju razlike u rezultujućim politikama ako se kao kriterij početnog stanja uzme veličina $BE\,(3)$ odnosno $BK32$. U nastavku su prezentirani rezultati kada su stanja uzimana kao prve tri cifre numeričke veličine. Isti komparativni rezultati su postignuti kada su korištene prve četiri cifre kao predstavnici stanja pa ti rezultati neće biti posebno obrazlagani. Rezultati su prezentirani u obliku grafikona na kojima su predstavljene Q vrijednosti. Svi testovi su provedeni u 10 obrada sa po 25 iteracija. Veći broj iteracija daje rezultate većeg stepena vjerodostojnosti. Vrijeme testiranja se bitno povećava ako se radi sa prve četiri cifre s obziom na povećanu frekvenciju parova stanja iz dva atributa. Vrijeme jedne iteracije varira o jedne minute, u testovima sa tri prve cifre, do pet minuta u ostalim testovima. S obzirom da su $BE\,(3)$ i $BK32$ jednake za prve tri i prve četiri cifre broja, efekat uzimanja četiri cifre je povećanje stepena fokusiranosti. Ovo može biti važno u praktičnim primjenama ovog algoritma. Parametar opsega frekvencija početnih stanja u svim slučajevima je postavljen tako da obuhvata minimalno pet mogućih vrijednosti. Vrijednosti dobijene algoritmom su pohranjene u posebnim Excell tabelama tako da je moguća njihova naknadna analiza.

U cilju analiza napravljeni su grafikoni dobijenih Q vrijednosti. Na x osi je broj iteracija a svaka obrada je predstavljena jednom linijom na grafikonu. Odabrano je predstavljanje Q vrijednosti putem linija kako bi se bolje uočile uporedne karakteristike svake obrade. Na grafikonima su posebno označene prva i posljednja obrada.

7.7 Uzorak 1 - datoteke

Prvi uzorak je skup podataka o datotekama na laptopu dobijen naredbom

```
dir /S /A-D /-C c:\ | sort > lista.txt
```

Kao rezultat navedene naredbe dobijen je spisak od 107.566 stavki. Na osnovu spisak je napravljen izbor atributa koji je na tabeli 7.1.

Rb	Naziv kolone	Instance	Napomene
1.	Velicina	900	Velična datoteke u bajtovima
2.	Ekstenzija	566	Ekstenzija od tri slova
3.	Dan	31	Dan iz datuma formiranja
4.	DanUSed	7	Dan u sedmici, u skladu sa realnim datumom
5.	Mjesec	12	Mjesec iz datuma formiranja
6.	Period	24	Period formiranja, na osnovu vremena formiranja datoteke, u koracima od 1 sat
7.	DuzinaNaziva	85	Dužina naziva datoteke

Table 17: Struktura atributa datoteke koja je predmet analize. Kolona Instance daje broj razlicitih stanja po datom atributu

Kolona označena sa Instance daje broj različitih stanja po pojedinom atributu. Kolona Period označava period u toku dana u intervalima od jedan sat kada je transakcija obavljena. Dobijen je uzorak koji u numeričkom atributu ima samo cijele brojeve.

Veličine datoteka, u atributu Velicina, su u rasponu od 0 do 2.146.750.464, izražene u bajtovima. Nakon odbacivanja vrijednosti manjih od 100 uzorak je sveden na $N = 104.337$ stavke. Razlog odbacivanja je u činjenici da su vrijednosti veličina $BE\,(3)$ i $BK32$ jednaki za dvocifrene brojeve. Kao donja granica veličine uzeta je vrijednost 8.200 a kao gornja granica vrijednost 3.000.000. Ovakav izbor je napravljen u skladu sa analizama veličina $BE\,(3)$ i $BK32$ koje su opisane u poglavlju 6 ovog teksta za ovaj uzorak. Radi podsjećanja, kritična numerička vrijednost je 8.192. Nakon ovoga uzorak je sveden na obim $N = 48.752$. Raspon vrijednosti za veličinu $BE\,(3)$ u ovako dobijenom uzorku je interval $[1,3907; 8,1499]$ a za veličinu $BK32$ je interval $[1,0909; 4,0199]$.

Struktura podataka po rasponima vrijednosti je predstavljena na grafikonu 6.1. Rasponi su dati po stepenima baze 10 (A: $100 - 1.000$, B: $1.000 - 10.000$, C: $10.000 - 100.000$, D: $100.000 - 1.000.000$, E: $1.000.000 - 10.000.000$, F: $10.000.000 - 100.000.000$, G:$100.000.000 - 1.000.000.000$, H:$1.000.000.000 - 10.000.000.000$). Ovo odražava i raspored frekvencija vodećih cifara.

Kao predstavnici stanja uzimane su tri vodeće cifre iznosa i provedene su dvije grupe testova. U jednoj grupi testova za odzive su uzimane veličine $BE\,(3)$ a u drugom $BK32$. Uslov za početna stanja je njihova frekvencija u rasponima od 300 do 380. Procjena je napravljena u skladu sa planom da se test provede u 10 prolaza sa po 25 iteracija. Broj početnih stanja sa ovakvim frekvencijama je dosta skroman tako da 10 prolaza

omogućava da svako stanje bude više puta odabrano kao početno kako bi se mogla praviti odgovarajuća poređenja.

Formirane su Q vrijednosti i na osnovu njih napravljeni grafikoni. Prva očigledna razlika između ove dvije metode je u rasponima Q vrijednosti. U prvoj metodi, kada se kao odziv koristi $BE\,(3)$, vrijednosti dostižu nivo od $161,7808$. U drugoj metodi, kada se kao odziv koristi $BK32$, vrijednosti dostižu nivo od 69.7421. Uzrok je u razlici raspona vrijednosti $BE\,(3)$ i $BK32$.

Od rezultujuće politike se može očekivati da pruži informacije o tome da li postoji i kakva je priroda veze između veličine datoteke, vremena i perioda njenog nastanka i drugih atributa. Drugim riječima, algoritam treba pružiti informaciju da li se određeni tip datoteka formira po nekom obrascu koji je prepoznatljiv.

Rezultati za Q vrijednosti su date na grafikonima. Iteracije za različita stanja su označena različitim bojama. Vrijednosti svakog prolaza za jedno početno stanje su označene različitom linijom. Q vrijednosti posljednjeg prolaza za svako početno stanje su označene duplom linijom.

Rezultati za Q vrijednosti prema prvoj metodi su na grafikonu 7.2. Iteracije su obuhvatile pet stanja (pet puta stanje 107, dva puta stanje 101, jednom stanje 102, jednom stanje 231, jednom stanje 122). Uočljivo je da Q vrijednosti za stanje 107 imaju osjetno veće vrijednosti u odnosu na ostala stanja. Najveća Q vrijednost, $161,780$, je dostignuta u tri različite iteracije. Putanje za ovu vrijednost su navedene kao Putanja1, Putanja2 i Putanja 3 respektivno u tabeli 7.2.

Rezultati za Q vrijednosti prema drugoj metodi su na grafikonu 7.3. Iteracije su obuhvatile 5 stanja (tri puta stanje 101, dva puta stanje 102, dva puta stanje 107, dva puta stanje 122, jednom stanje 231). Prva očigledna razlika u odnosu na prethodni grafikon je u broju početnih stanja pri istim početnim uslovima. Druga razlika je u odnosu vrijednosti koje su dobijene u pojedinim prolazima. Najveća Q vrijednost za stanje 107 je na tabeli 7.2 kao Putanja4. Ovo nisu najveće vrijednosti po ovoj metodi; one su dobijene za početno stanje 122 $(69,742)$ što ukazuje na bitnu razliku rezultujućih politika. Ovdje su te putanje date kako bi se ilustrovale razlike u putanjama. Varijacije u vrijednostima unutar jednog prolaza se mogu pripisati politici pretraživanja.

Najveće vrijednosti po drugoj metodi $(45,296)$ postignuti su za početno stanje 101 za dvije putanje, dodate kao kolone Putanja 5 i Putanja 6 u tabeli 7.2.

Prvi red u svakoj od grupa po tri reda odgovara oznaci akcije (kolone), drugi red daje stanje a treći red njegovu frekvenciju. Redoslijed akcija odgovara postupku filtriranja u radu sa Excell tablicama, analogno procesu selekcije tokom analize podataka.

Postupak filtriranja redoslijedom koji je dat u bilo kojoj od ove dvije tabele se može zaustaviti na bilo kojem od koraka. Tokom analize u ovom slučaju je obratiti pažnju na slučajeve kada su frekvencije na uzastopna dva koraka jednake ili veoma bliske. Ovo ukazuje na vezu među atributima za koje ovo važi.

Korak	**Putanja1**	**Putanja2**	**Putanja3**	**Putanja4**	**Putanja5**	**Putanja6**
Q	161, 780	161, 780	161, 780	25, 153	45, 296	45, 296
Poč. st.	**0**	**0**	**0**	**0**	**0**	**0**
	107	107	107	107	101	101
	313	313	313	313	307	307
Akcija1	**1**	**3**	**2**	**6**	**1**	**4**
	DLL	UT	14	12	EXE	MAJ
	45	61	35	57	56	84
Akcija2	**2**	**1**	**1**	**1**	**2**	**1**
	14	DLL	DLL	DLL	20	DLL
	12	7	12	6	1	2
Akcija3	**3**	**2**	**3**	**3**	**3**	**3**
	PO	4	PO	PO	SR	NE
	12	1	12	4	1	2
Akcija4	**4**	**4**	**4**	**4**	**4**	**2**
	APR	NOV	APR	APR	AVG	23
	12	1	12	4	1	2
Akcija5	**5**	**5**	**5**	**5**	**5**	**5**
	15	4	15	15	21	24
	12	1	12	4	1	2
Akcija6	**6**	**6**	**6**	**2**	**6**	**6**
	12	12	12	14	10	10
	4	1	4	4	1	1

Table 18: Tabela 7.2 Putanje dobijene algoritmom Q ucenja. Prvi red u grupama od po tri reda odgovara akcijama (kolonama), drugi red oznacava stanje a treci red frekvenciju tog stanja. Putanja1 i Putanja 2 odgovaraju rezultatima po prvoj a ostale pod drugoj metodi

7.8 Diskusija

Nakon provedenih testova može se reći da postoji bitna razlika u rezultujućim politikama ako se kao uslov početnog stanja koristi opseg količnika ($BK32$) u odnosu na vrijednosti za $BE\,(3)$ iz odabranog opsega. Parametar prema kojem se biraju početna stanja za koje je frekvencija u datom opsegu u svim slučajevima je postavljen tako da obuhvata minimalno pet mogućih vrijednosti. Korištenje veličine $BK32$ kao odziva za rezultat ima drugačiji, češće manji broj početnih stanja unutar istog broja prolaza. Istovremeno, nije se mijenjala politika ažuriranja Q vrijednosti. Ovo je veoma značajan rezultat jer pokazuje da se politike pretraživanja mogu bitno poboljšati i unaprijediti bez posebnog uticaja na politiku ažuriranja Q vrijednosti.

Kada je uslov početnog stanja opseg količnika ($BK32$) Q vrijednosti su manje u odnosu na one koje se dobiju ako se kao uslov koriste vrijednosti iz opsega veličina $BE\,(3)$. Ovo je važno sa stanovišta mogućnosti korištenja nekih politika pretraživanja kao što je SoftMax u kojoj je važan izbor parametra τ, u skladu sa nivoom tekućih Q vrijednosti i rezultatima detaljne analize prije korištenja.

Identične karakteristike rezultata su dobijene ako se početna stanja uzimaju sa četiri umjesto sa tri cifre. Takođe, rezultati provedeni sa drugim vrijednostima parametara β, γ i ε su u saglasnosti sa prethodno prezentiranim rezultatima. Varijacije vrijednosti

između uzastopnih iteracija mogu se pripisati postupku pretraživanja.

S obzirom na relativno mali broj provedenih iteracija, postupak se može nastaviti tako da se provede novi krug iteracija na istom uzorku i sa potpuno istim parametrima ali da se pritom zanemare optimalne putanje dobijene u ovom koraku, što je jednako njihovom fizičkom uklanjanju iz uzorka. Na taj način se dobijaju nove nešto manje optimalne politike. Postupak se može nastaviti do momenta dok se ne procijeni da se na taj način ne mogu dobiti kvalitativno nove informacije. Procjenu o broju ovakvih ponavljanja donosi isključivo analitičar. Algoritam koji je razvijen za ove potrebe je prilagođen i za ovaj način rada i uspješno testiran.

Na osnovu navedenog može se zaključiti da izbor odziva $BE\,(3)$ odnosno količnika $BK32$ kao kriterija izbora početnog stanja odnosno njihovo korištenje u funkciji odziva ima bitan uticaj na redoslijed akcija, na Q vrijednosti i time na rezultujuću politiku. Time je polazna hipoteza potvrđena.

7.9 Testiranje rezultata putem Benfordovog zakona

Q vrijednosti generisane za opisani uzorak su analizirane putem Benfordovog zakona. Predmet analize su bile vrijednosti po obje metode zasebno. Proveden je test prve cifre i druge cifre izračunatih Q vrijednosti, bez obzira što se radi o relativno malom uzorku.

Benfordova analiza za rezultate dobijene po prvoj metodi su na grafikonima 7.4 do 7.6. Grafikon 7.4 daje distribuciju Q vrijednosti po rasponima vrijednosti. Uočljive su velike frekvencije u rasponu od 55 do 105. Test prve cifre, dat na grafikonu 7.5, ukazuje na naglašene frekvencije veličina koje počinju sa 7, 8 i 9. Ovo je u skladu sa izgledom grafikona 7.4. na kojem je vidljivo da najveći dio frekvencija otpada na Q vrijednosti iz opsega 55 do 105. Test druge cifre, dat na grafikonu 7.6 ukazuje na povećanu frekvenciju cifara 1 i 2 na drugoj poziciji.

Benfordova analiza za rezultate dobijene po drugoj metodi su na grafikonima 7.7 do 7.9. Grafikon 7.7 daje distribuciju Q vrijednosti po rasponima vrijednosti. Uočljive su velike frekvencije u rasponu od 20 do 35. Test prve cifre, dat na grafikonu 7.8, ukazuje na naglašene frekvencije veličina koje počinju sa 1 i 2 što je i logično ako se u ima u vidu grafikon 7.7. na kojem je vidljivo da najveći dio frekvencija otpada na Q vrijednosti iz opsega 100 do 220. Test druge cifre, dat na grafikonu 7.9 ukazuje na povećanu frekvenciju cifre 4 na drugoj poziciji, što je ponovo u saglasnosti sa rezultatima na grafikonu 7.7.

7.10 Zaključak

Provedeni testovi ukazuju da izbor vrijednosti $BE\,(3)$ i $BK32$ u funkciji odziva za numeričke veličine ima bitan uticaj na rezultate reinforcement učenja.

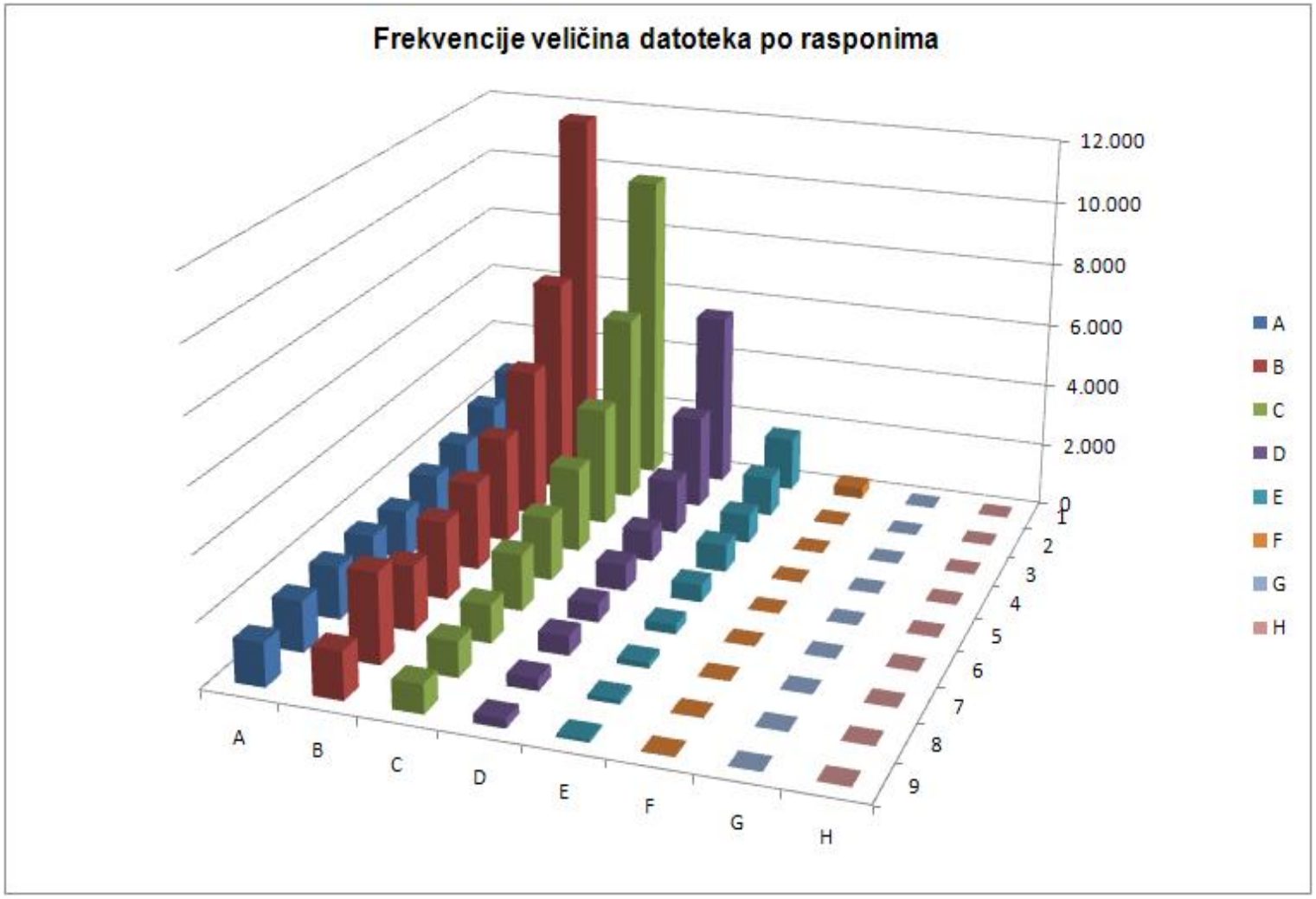
Najočitiji i najbitniji je uticaj na početna stanja. Skup početnih stanja dobijen po osnovu izbora u opsegu velične $BE\,(3)$ je u nekim skupovima podataka bitno drugačiji od skupa koji se dobije po uslovu izbora u opsegu veličine $BK32$. Polazak od različitih početnih stanja ima za posljedicu drugačiju strukturu mogućih narednih stanja odnosno

ukupne politike. Veći količnik $BK32$ ukazuje na povećanu frekvenciju treće cifre odnosno povećanu frekvenciju sličnih numeričkih veličina.

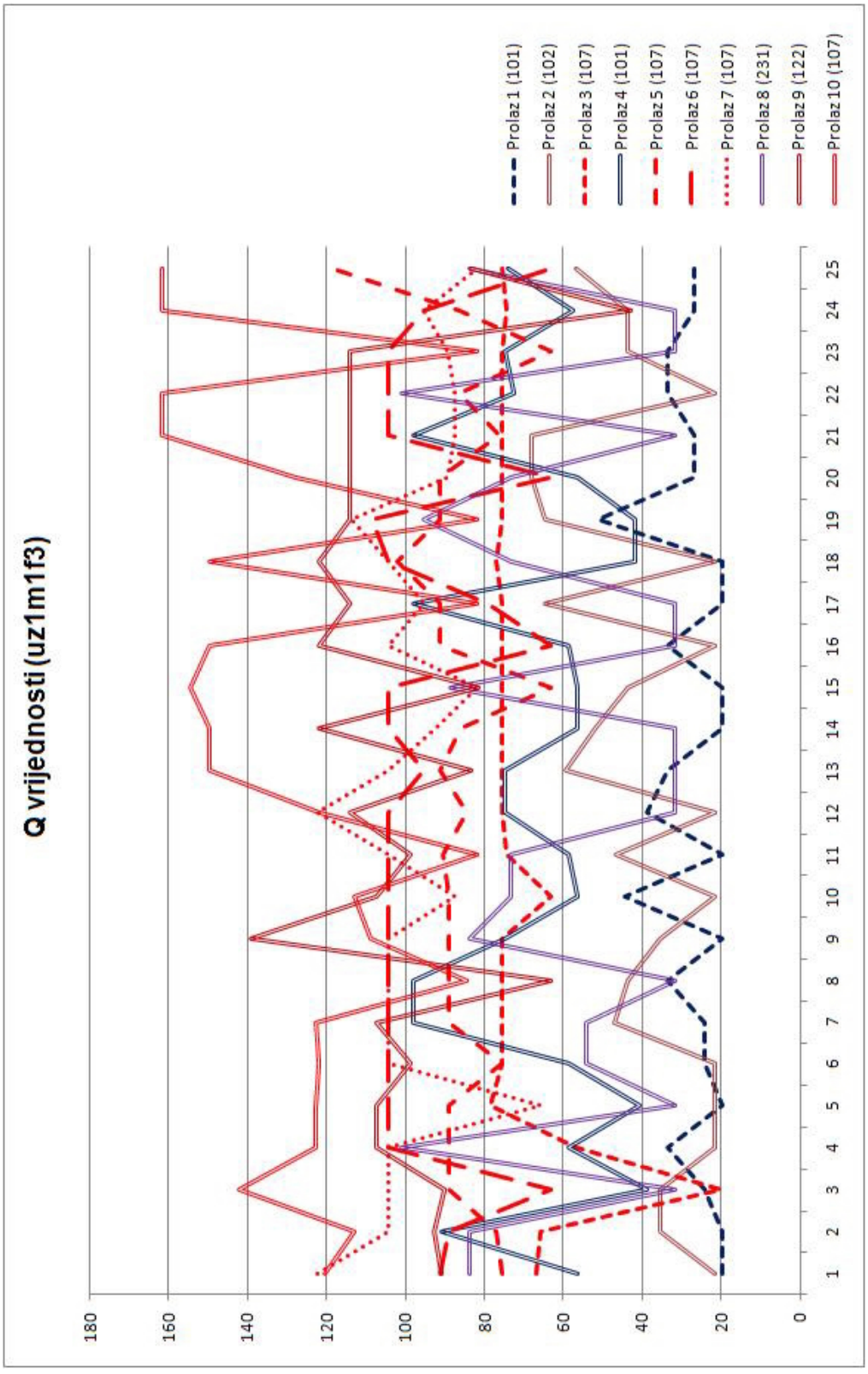
Korištenjem metoda reinforcement učenja dobija se mogućnost utvrđivanja konteksta u kojem se javljaju određene grupe podataka, zavisno od njihovog sadržaja a ne samo od redoslijeda. Ovo je izuzetno važno sa stanovišta praktične primjene. Treba naglasiti da rezultate Q učenja treba posmatrati kao oblik / obrazac ponašanja odnosno obrazac nastajanja i održavanja veza između sadržaja atributa. Tek analiza sa stanovišta izvora i okolnosti nastajanja i toka podataka može ukazati da li se radi o anomaličnosti. Pri tom pojam 'anomaličnost' treba posmatrati u nešto širem okviru u smislu da odstupanje može predstavljati legitiman mada specifičan način rada. Određeni tip transakcija u određenom periodu, npr. često dizanje gotovine na određenim uređajima u određeno vrijeme, nekad može upućivati samo na najobičnije navike klijenata. Nastajanje datoteka određenog tipa u određene dane umjesto na anomaličnost može upućivati na određena pravila kao što su ažuriranje antivirus zaštite, ažuriranje operativnog sistema i slično.

Varijacije vrijednosti unutar jednog ciklusa iteracija ukazuje na uticaj konteksta na dobijene Q vrijednosti. Jednaki redoslijed atributa ne mora u svakom slučaju značiti jednake Q vrijednosti. Manje Q vrijednosti za jednake redoslijede ukazuju na manje vjerovatne situacije u kontekstu reinforcement učenja.

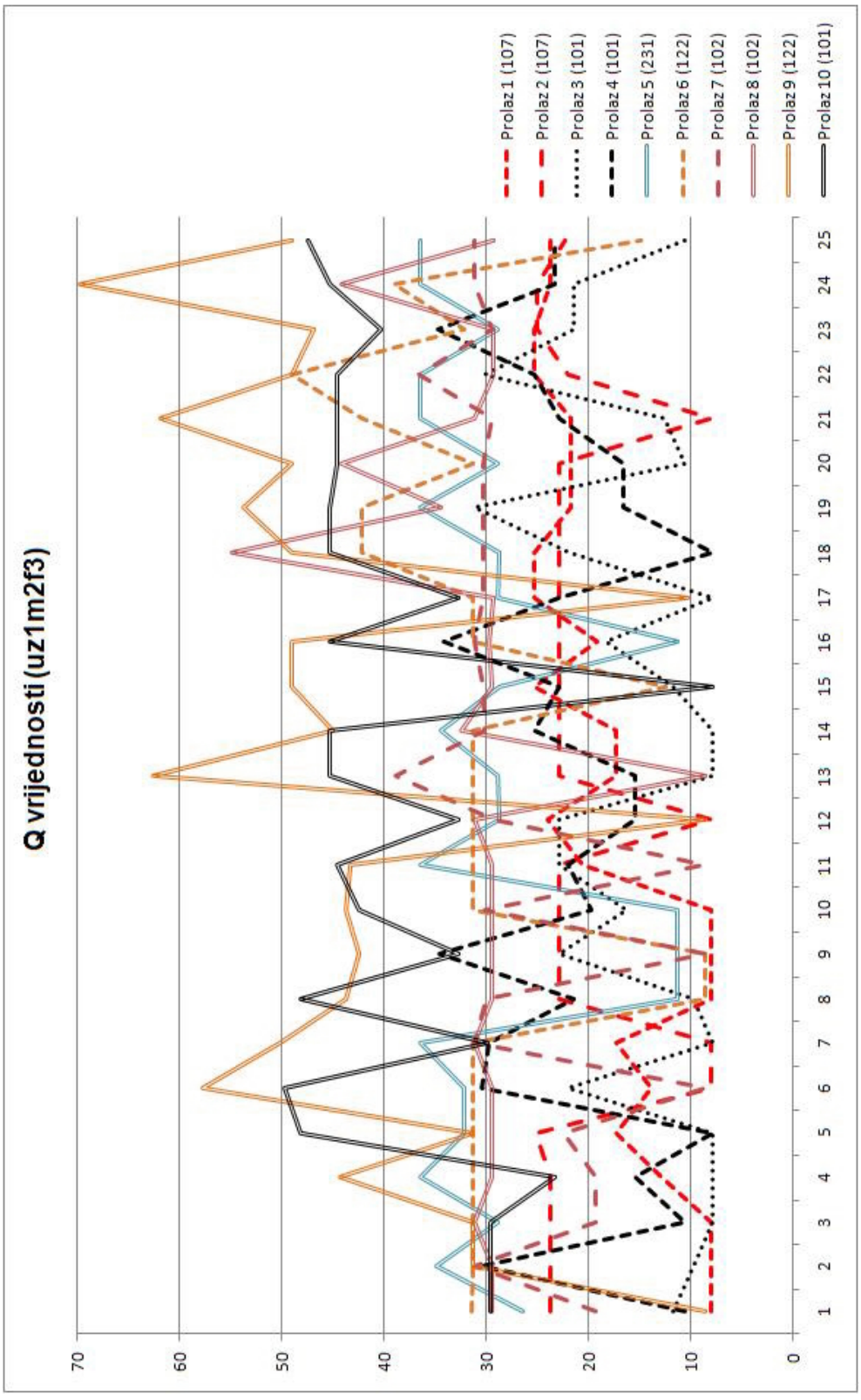
Na taj način je hipoteza dokazana.



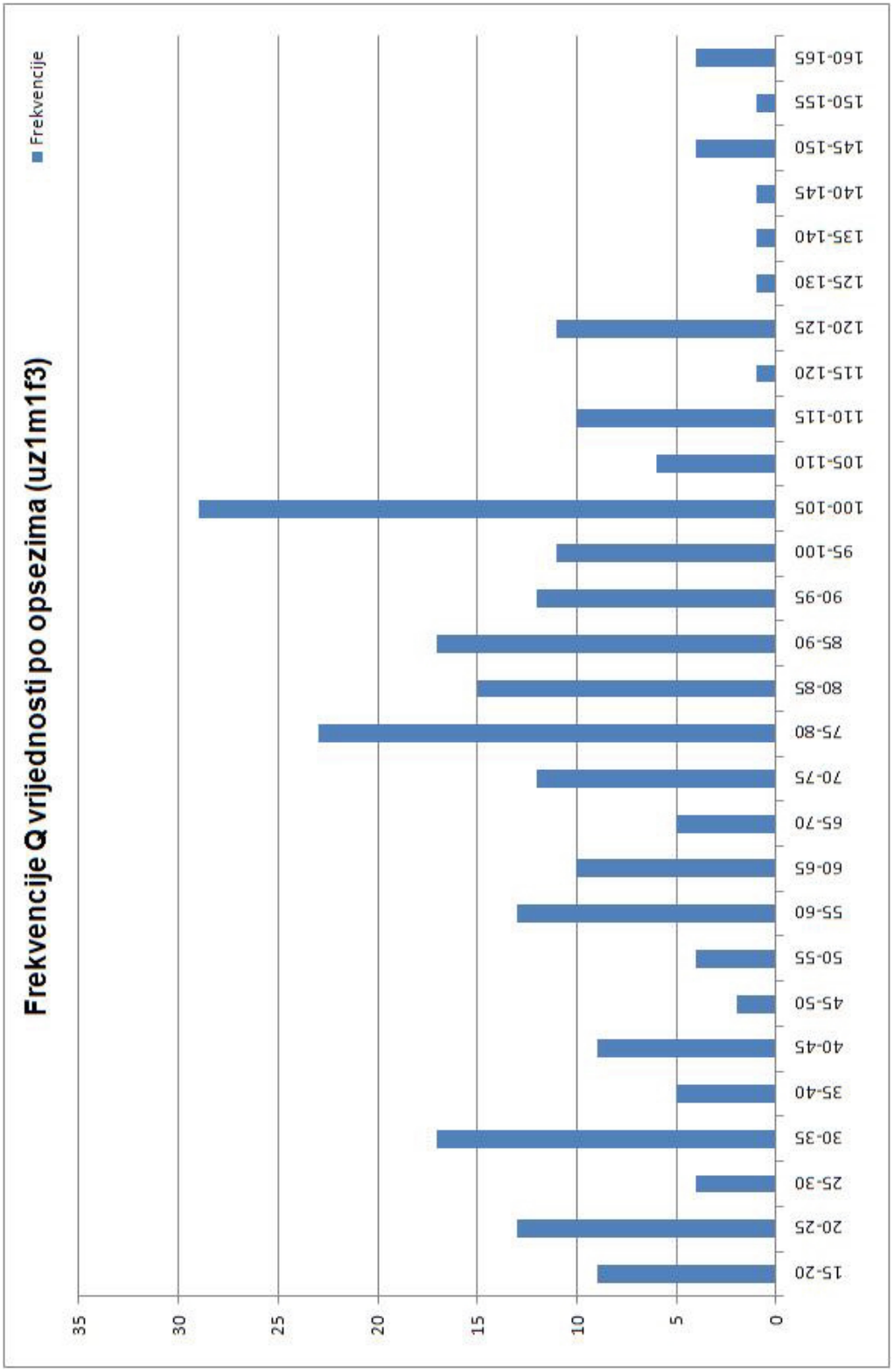
Grafikon 7.1. Frekvencije veličina datoteka po rasponima veličina. Za grupe B i C je uočljiv veliki stepen slaganja sa Benfordovim zakonom osim za vrijednosti iz grupe B koje počinju cifrom 8



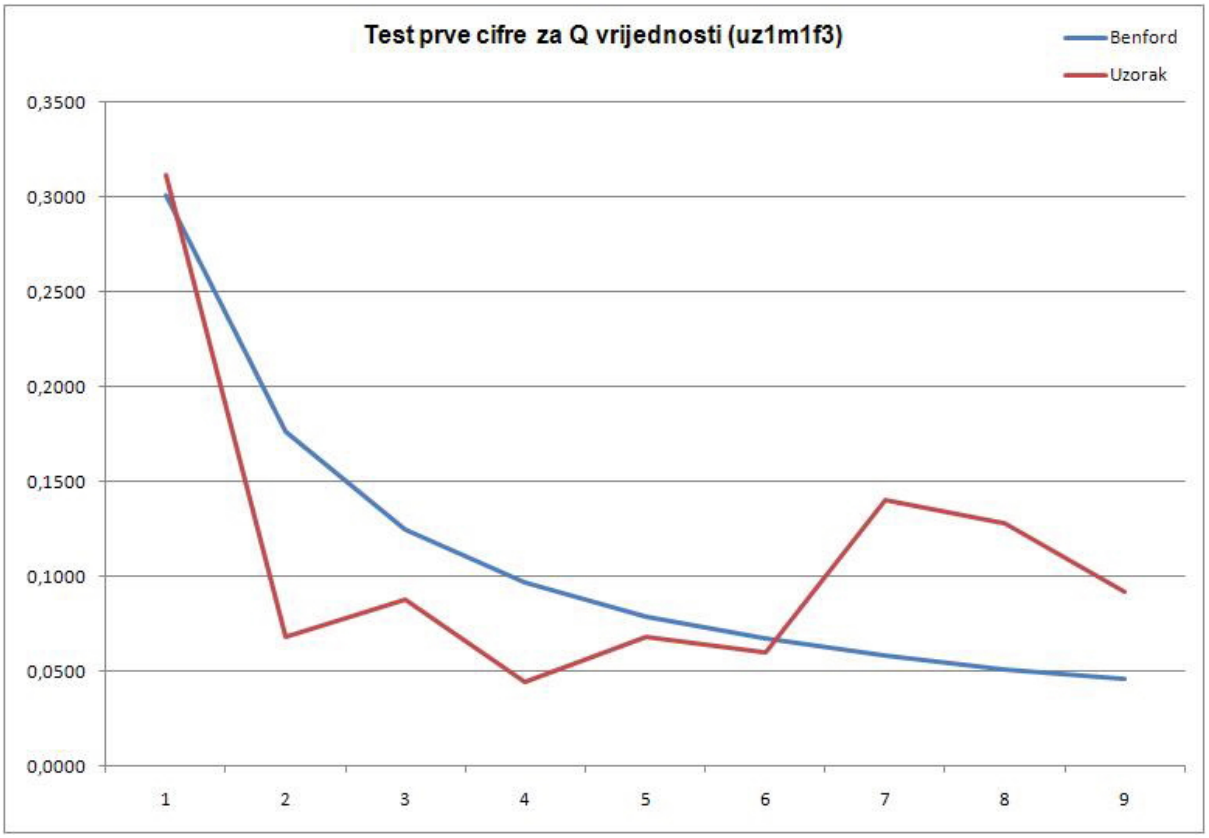
Grafikon 7.2. Q vrijednosti ako se kao odziv uzima vrijednost $BE\,(3)$. Različita stanja su označena različitim bojama. Posljednja iteracija za svako stanje je označena dvostrukom linijom.



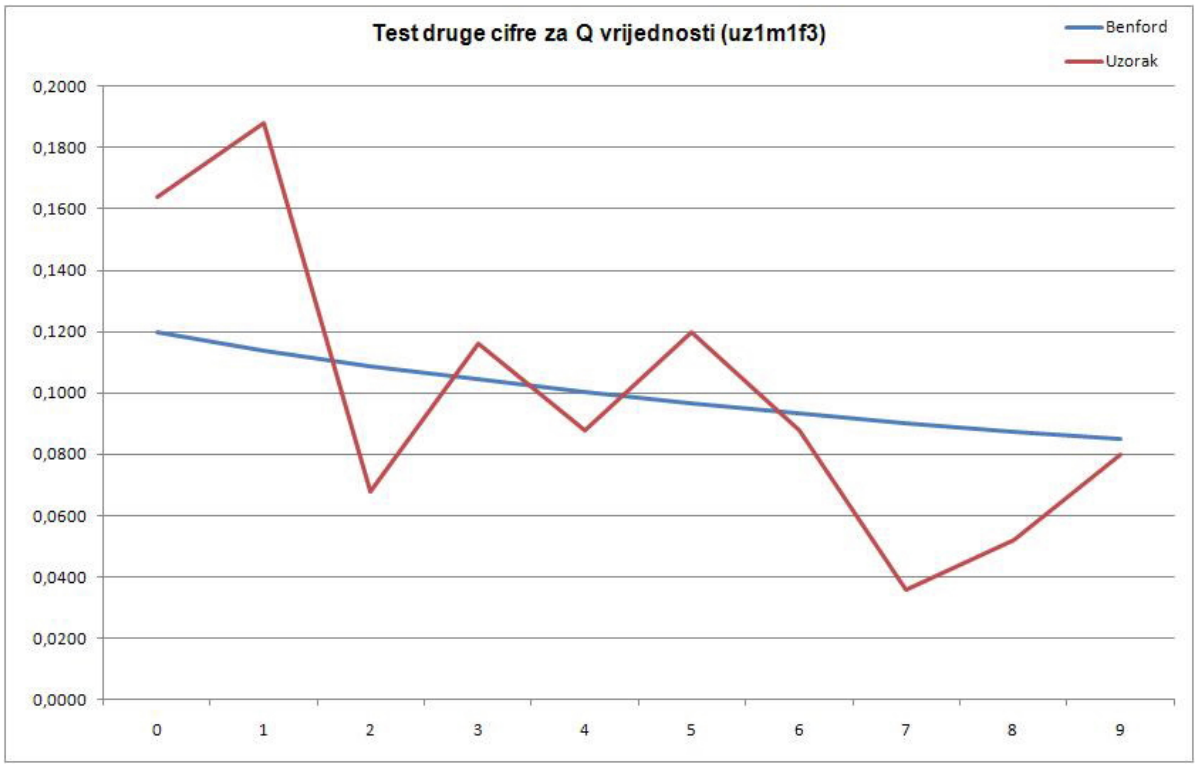
Grafikon 7.3. Q vrijednosti ako se kao odziv uzima vrijednost $BK32$. Različita stanja su označena različitim bojama. Posljednja iteracija za svako stanje je označena dvostrukom linijom



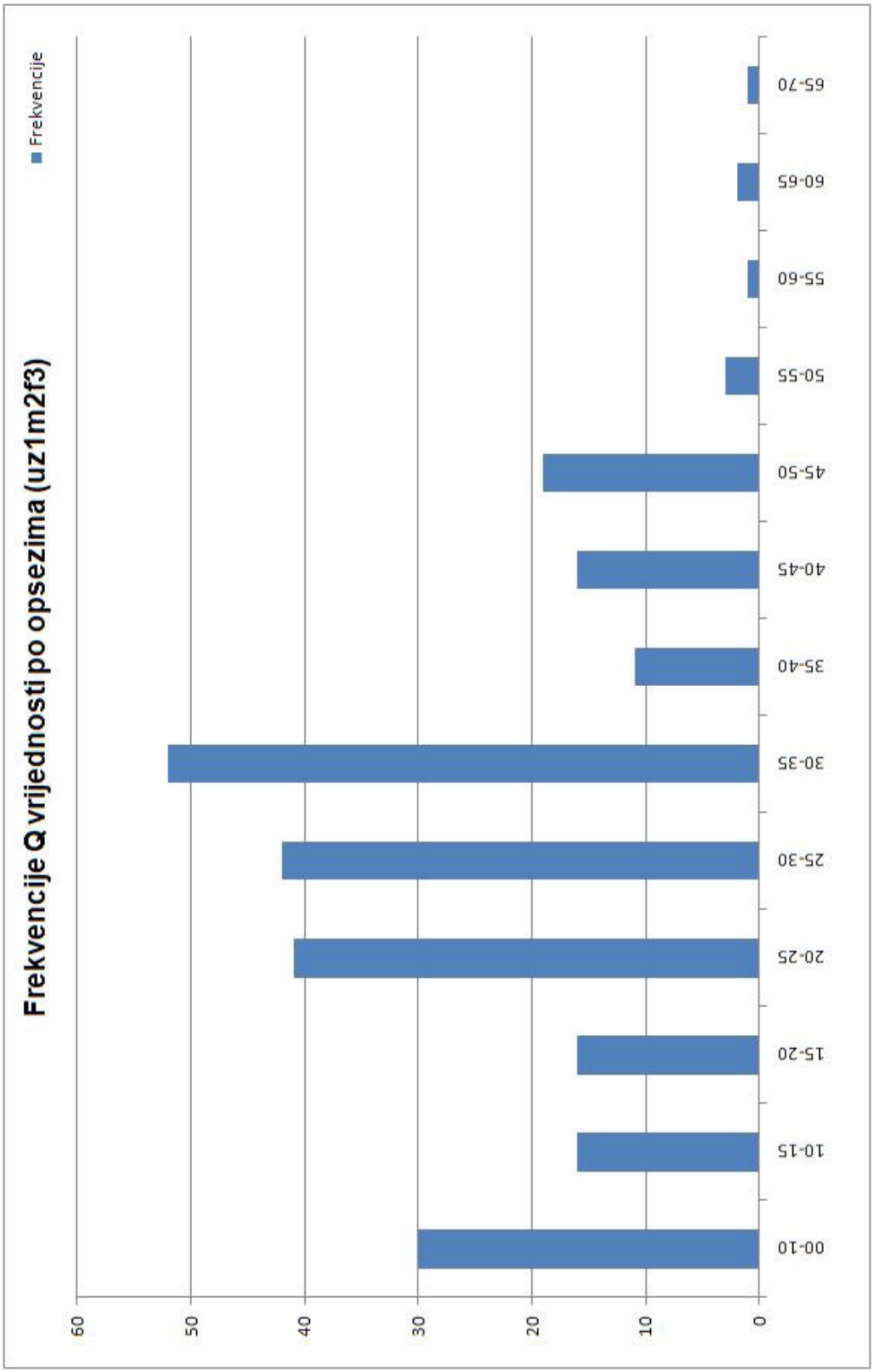
Grafikon 7.4. Distribucija frekvencija Q vrijednosti po ospezima (prva metoda). Uočljive su velike frekvencije u rasponu od 55 do 105



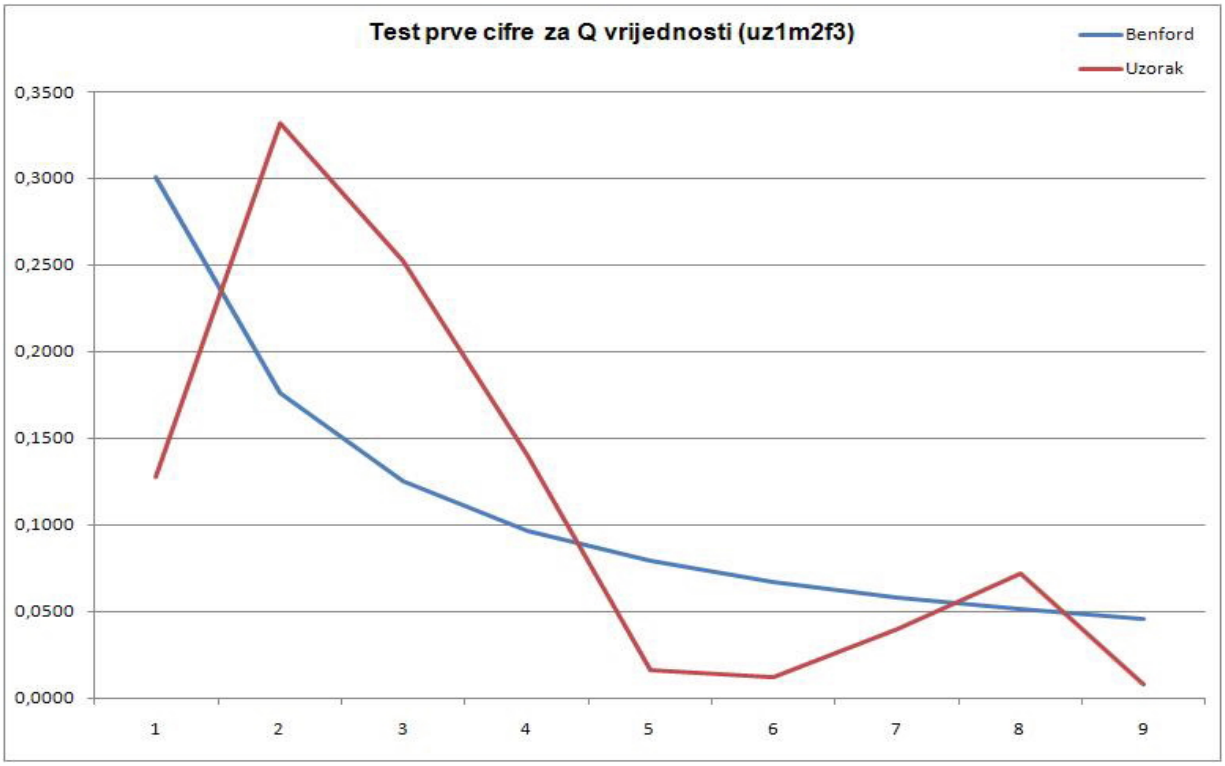
Grafikon 7.5. Frekvencije prvih cifara Q vrijednosti. Uočljiva je povećana frekvencija cifara 7, 8 i 9 na vodećim pozicijama



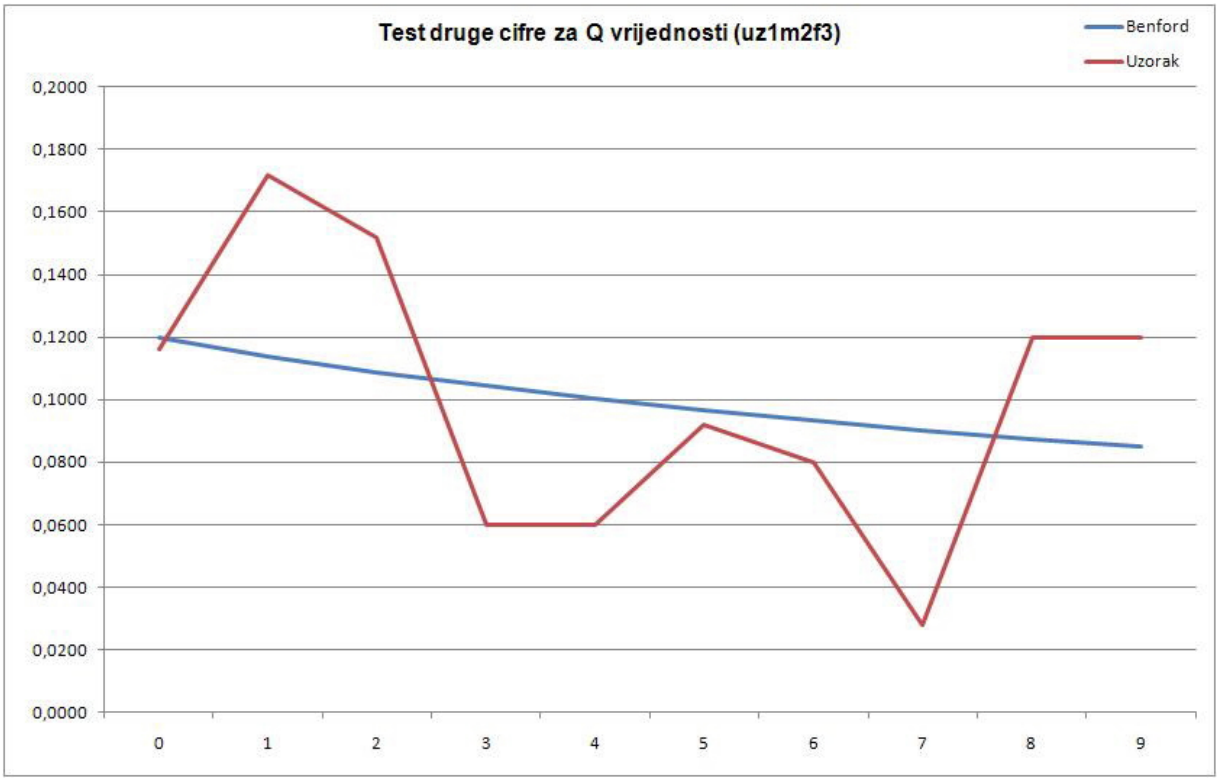
Grafikon 7.6. Frekvencije cifara Q vrijednosti na drugoj poziciji. Uočljiva je povećana frekvencija cifara 1 i 2 na drugoj poziciji



Grafikon 7.7. Distribucija frekvencija Q vrijednosti po ospezima. Uočljive su velike frekvencije u rasponu od 20 do 35



Grafikon 7.8. Test prve cifre Q vrijednosti. Naglašene su frekvencije cifara 2, 3 i 4 na prvoj poziciji. Frekvencija cifre 8 se može uzeti kao prihvatljiva



Grafikon 7.9. Test druge cifre na proračunatim Q vrijednostima. Uočljivo je odstupanje za cifru 4 i cifru 7

8 Zaključak

Benfordov zakon je logaritamski zakon distribucije prvih značajnih cifara. Pojava da se određene cifre na početnim pozicijama javljaju češće od ostalih uočena je još krajem XIX vijeka od strane astronoma Simona Newcomba. Primijetio je da se početne stranice logaritamskih tablica više koriste od ostalih. Iz toga je zaključio da su ljudi iz nekog razloga skloniji korištenju brojeva koji počinju manjim ciframa. Istu pojavu je uočio fizičar Frank Albert Benford (1938). Za razliku od Simona Newcomba, on je proveo eksperiment i predložio izraz kojim se izražava vjerovatnoća pojave odabrane cifre na nekoj od vodećih pozicija. Kasnije je ovaj zakon po njemu dobio ime. Eksperiment se sastojao u prikupljanju velikog broja numeričkih veličina iz raznih izvora i kalkulisanju relativnih frekvencija. Rezultati su pokazali iznenađujuće poklapanje sa predloženom distribucijom bez obzira na izvor podataka. Ovo ukazuje da je sa stanovišta Benfordovog zakona potpuno nevažno da li se radi o kupu novčanih transakcija, biološkom procesu rasta ili dinamike populacije, izbornim rezultatima, berzanskim transakcijama, radu računarskog sistema, telekomunikacijama, pisanom ili govornom jeziku ili čemu drugom. Drugim riječima, izraz ni na koji način ne ulazi u samu prirodu ili izvor numeričke veličine.

Objašnjenje ovog zakona je bio i ostao veliki teorijski izazov jer se čini da ovaj zakon izražava daleko dublje odnose nego što se to čini na prvi pogled. George Joseph Stigler (1911 - 1991), dobitnik Nobelove nagrade za ekonomiju (1982), iznio je ideju (1945) prema kojoj je Benfordov zakon specijalan slučaj familije monotono opadajućih distribucija vjerovatnoće. Primijećeno je da postoje skupovi za koje Benfordov zakon ne vrijedi. Pritom je značajno da čak i kada prve značajne cifre odstupaju od logaritamskog obrasca Benfordovog zakona čini se da relativne frekvencije još uvijek favoriziraju manje cifre i opadaju monotono na način srodan Benfordovom zakonu. Nakon njega su mnogi matematičari dali svoje teorijska tumačenja i dokaze koja su polazila od različitih premisa ali je rezultat uvijek bio isti. Poseban doprinos su dali g-đa Bret Flehinger (1965), Theodor P. Hill, Takloo i Bigash, Raimi, Miller i drugi.

Benfordov zakon je primjenljiv na velike skupove numeričkih vrijednosti koje zadovoljavaju određene uslove kao što su raspon vrijednosti u najmanje dva reda stepena baze, odsustvo vidljivih ograničenja, veći broj elemenata, odsustvo strukturiranosti pojedine numeričke veličine i slično. Provjeren je i eksperimentalno potvrđen na velikom broju uzoraka iz različitih izvora.

Jedna od bitnih karakteristika Benfordovog zakona su mjerna i bazna invarijantnost. Mjerna invarijantnost je osobina prema kojoj se zakon distribucije prvih cifara ne mijenja ako se sve veličine uzorka pomnože istim brojem. Ako su npr. finansijske veličine iskazane u jednoj valuti iste osobine distribucije se zadržavaju prilikom konverzije. Bazna invarijantnost je osobina prema kojoj se distribucija zadržava ako se veličine iskažu u drugim bazama.

Formulacija u obliku izraza koji povezuje poziciju cifre sa vjerovatnoćom njene pojave na toj poziciji je inicirala ideju da se uzoračke vrijednosti porede sa ovim zakonom i da se iz toga pokušaju izvući određeni zaključci. Prvi od njih se nameće sam po sebi : ako postoji zakon koji izražava vezu cifre i njene pozicije i ako se pokaže da distribucija na uzorku ne odgovara tom zakonu onda postoji osnova za sumnju u neki oblik anomalije unutar samog skupa. Naročiti zamah ovoj ideji dao je razvoj informatike, bolje rečeno personalnih

računara, i programskih okruženja u kojima je moguće provoditi brze kalkulacije i obradu sve većih količina podataka po sve manjoj cijeni. Jedan od prvih koji se u radu služio ovom idejom je Mark Nigrini koji je napravio analizu poreskih prijava. Nakon toga je naglo porastao broj autora koji su ovaj zakon koristili ili koriste u gotovo svim područjima nauke i života. Ovakva priroda zakona je njegovu primjenu primarno usmjerila na detekciju prevara kao jednog od logičnih zaključaka kada se u skupu detektuje anomalija. Zbog svoje objektivnosti i potpune nezavisnosti od izvora ili prirode podataka koji se analiziraju ovaj zakon je priznat kao legitiman revizorski metod. Rezultati analize putem ovog zakona na američkim sudovima se priznaju kao vjerodostojni i neporecivi.

Brojni su drugi primjeri mogućnosti primjene ovog zakona za raznolike segmente nauke i života. Pokazano je da rekurzivni izrazi, kao što su Fibonačijevi brojevi, iterativne odnosno rekurzivne numeričke metode i slično, produciraju brojeve koji slijede ovaj zakon.

Prilikom korištenja Benfordovog zakona u bilo kom smislu logično je postaviti pitanje o tome koliki dio skupa je anomaličan odnosno u kojoj mjeri se eventualna anomaličnost može procijeniti. To je jedno od prvih pitanja sa kojim se susreće analitičar u svom radu koji je zainteresovan da na osnovu ovakve procjene odredi strategiju i naredne korake. U tekstu je pokazano da je ovakvu procjenu moguće napraviti na osnovu teorijskih i uzoračkih frekvencija na način da se, na osnovu teorijske distribucije, formira interval povjerenja i zatim mjere odstupanja odnosno slučajevi koji izlaze iz tako formiranog intervala. Rezultat je okvirna procjena broja stavki koje se mogu posmatrati u smislu značajnog odstupanja od teorijske distribucije. Istovremeno, na osnovu prve cifre je moguće ukazati na skup stavki iz kojeg dolazi detektovano odstupanje odnosno skup koji je 'krivac' za stepen anomalije. Pritom je važno istaći da ovo ne znači mogućnost detekcije pojedinačnih stavki već se za to moraju uraditi druge vrste analiza.

U testiranjima putem Benfordovog zakona se polazi od pretpostavke da skup zadovoljava uslove koje je definisao Nigrini, posebno u smislu odsustva bilo kog vida donjih ili gornjih ograničenja. U praksi ovaj uslov nekada nije moguće osigurati. Ipak, analitičar i dalje želi provoditi testove putem Benfordovog zakona. Primjena klasičnog testa na Benfordov zakon bi dala procjenu o velikom odstupanju a time i sliku koja, u suštini, nije tačna. Kako bi prevazišli ovaj problem Efrim Boritz i Fletcher Lu su u svojoj patentnoj prijavi (2008) prezentirali ideju Adaptivne Benfordove metode, koja na osnovu raspoloživog skupa pravi transformaciju kako bi se i dalje moglo vršiti testiranje putem Benfordovog zakona. U tekstu je pokazano na koji način izbor donje i gornje granice uzorka ima uticaj na rezultate testiranja. Na odabranom uzorku je izvršena simulacija na način da se prave različiti izbori donjih granica i posmatraju odabrane veličine koje svojim promjenama mogu dati sliku uticaja. Kao veličina za ovu svrhu odabrano je srednje apsolutno odstupanje (Mean Absolute Deviation - MAD). Nakon testiranja je pokazano da izbor donjih i gornjih granica ima značajnog uticaja na veličinu MAD i na rezultate eventualnog testiranja. Nedostatak rada sa veličinom MAD je u činjenici da test ne ukazuje na eventualnu grupu stavki koja je mogući izvor detektovanog odstupanja.

Jedan od ciljeva testiranja uticaja izbora donjih i gornjih granica je simulacija realne situacije u kojoj analitičar prije početka svjesno odbacuje ekstremno male ili ekstremno velike vrijednosti. Npr. ako su predmet analize finansijski podaci jedan od najčešćih koraka je odbacivanje veličina manjih od odabranog praga, npr. manje od $10,00$ i/ili nekoliko ekstremnih veličina. Drugi cilj testiranja je bio provjeriti mogućnost da se na

osnovu dobijenih veličina napravi procjena o tome da li je skup bio i u kojoj mjeri predmet bilo kakve manipulacije. Iste ove analize su moguće i na skupovima iz kojih je izbačen dio podataka po principu 'prosijavanja'. Primjer je izbacivanje npr. svih transakcija na bankomatu, iznosa koji su zaokruženi, transakcija u određenom periodu i slično, iz regularnog skupa.

U patentnoj prijavi u kojoj su dali prijedlog Adaptivne Benfordove metode, Fletcher Lu i Efrim Boritz su predložili neke veličine koje se izračunavaju na osnovu Benfordovog zakona. Jedna od njih, koja je u ovom tekstu označena sa $BE\,(3)$, svakoj numeričkoj veličini dodjeljuje broj koji odražava uticaj cifara na pojedinim početnim pozicijama. Istu veličinu je moguće računati za dvije ili više cifara. Iz praktičnih razloga se zadržava rad sa najviše tri cifre. Računanje ove veličine za prve dvije cifre, $BE\,(2)$, daje mogućnost računanja količnika koji je u ovom tekstu označen sa $BK32 = BE\,(3)\,/BE\,(2)$. Testovi pokazuju da mijenjanje donjih i gornjih granica, čime se simulira nedostatak određenog skupa podataka, ima bitnog uticaja na veličinu $BK32$. Isti uticaj se može detektovati ako je uzorak bio predmet bilo kog oblika 'prosijavanja' podataka.

Posebno važan rezultat, koji je prezentiran u ovom tekstu, je u činjenici da je ustanovljeno da se najveće promjene ovog količnika dešavaju ako se kao donje granice uzmu veličine koje su iznad vrijednosti koje počinju ciframa čije uzoračke frekvencije značajno odstupaju od teorijskih. Uzrok je u strukturi računanja veličine $BE\,(3)$ u kojem se dešavaju značajne promjene u članu izraza kojim se izražava frekvencija treće cifre.

Ako skup u potpunosti slijedi Benfordov zakon tada slogovima za koje je veličina $BE\,(3)$ najveća odgovaraju slogovi za koje je veličina $BK32$ takođe najveća. Tokom testiranja je ustanovljeno da ako se uzmu drugačije donje granice ovaj odnos ne mora vrijediti i on se bitno narušava ako se kao donje granice, kao u prethodnom slučaju, uzmu veličine koje su iznad vrijednosti koje počinju ciframa čije uzoračke frekvencije značajno odstupaju od teorijskih. Drugim riječima, dobijen je kriterij putem kojeg se može zaključiti da li postoji mogućnost bilo kojeg oblika manipulacije u uzorku. Ovo je drugi bitan rezultat koji se nudi u ovom tekstu.

Priroda Benfordovog zakona primarni fokus njegove primjene stavlja u kontekst detekcije anomalija i sa tim povezanih primjena kao što su detekcije prevara. Generalno gledano, postupak se sastoj u tome da se porede teorijske i uzoračke relativne frekvencije u funkciji vjerovatnoća i na osnovu toga donose zaključci odnosno prave drugi testovi. Formiranje veličina $BE\,(3)$, $BE\,(2)$ i $BK32$, koje su izvedene na osnovu Benfordovog zakona, daje okvir za korištenje Benfordovog zakona u sasvim drugom kontekstu. U ovom tekstu je dat primjer njihovog korištenja u okviru metode reinforcement učenja.

Reinforcement learning je učenje šta uraditi, kako mapirati situaciju u akcije tako da se maksimizira numerička vrijednost signala odziva (reward). Onome koji uči ne govori se koje akcije treba preduzeti kao što je slučaj u većini formi mašinskog učenja. Umjesto toga, putem pokušaja on mora otkriti koje akcije daju najveći odziv. U najinteresantnijim i najizazovnijim slučajevima, akcije mogu uticati ne samo na neposredni odziv već i na narednu situaciju i tako na naredne odzive. Ove dvije karakteristike, metoda pokušaja i greške i odgođeni odziv, dva su najvažnija i najistaknutija svojstva metode reinforcement učenja. Reinforcement učenje se razlikuje od nadziranog učenja, prepoznavanja statističkih obrazaca i vještačkih neuronskih mreža. Nadzirano učenje je učenje iz primjera na

osnovu znanja od strane eksternog supervizora. Ovo je važna metoda učenja ali samo po sebi nije adekvatna za učenje iz interakcija. Na nepoznatom terenu, gdje bi se moglo očekivati najuspješnije učenje, agent mora biti sposoban učiti iz sopstvenog iskustva.

Ključni termin u teoriji i praksi reinforcement učenja je *reward*. Izvorno ovaj termin označava *nagradu*, *naknadu* i slično. U cijelom tekstu se kao prikladna zamjena ovog prevoda koristi termin *odziv*. Ovakav izbor je napravljen jer je u tom terminu sadržano tehničko i praktično značenje termina *reward* u kontekstu reinforcement učenja odnosno interaktivnog odnosa agenta i okruženja.

U tekstu je prezentiran primjer korištenja veličina $BE\,(3)$ i $BK32$ u funkciji odziva za početni numerički atribut. Relativne frekvencije pojedinih vrijednosti kategorijskih atributa, proračunate na nivou cijelog skupa, uzete su kao odziv za kategorijske atribute.

Cilj je bio potvrditi tezu da korištenje ovih veličina ima bitnog uticaja na rezultujuće politike. Osnova za tezu je u (prethodno potvrđenoj) činjenici da slogovi za koje je $BE\,(3)$ najveće ne moraju biti isti oni za koje je $BK32$ najveće, posebno ako su donje granice numeričkih vrijednosti veće od određenih kritičnih vrijednosti. Samim tim, izbor opsega ovih veličina diktira moguća početna stanja a time i politiku u cjelini.

Da bi se ova teza provjerila proveden je jedan broj testova. Osnova za testiranje su bile datoteke sa jednim numeričkim i šest kategorijskih atributa. Akcija u smislu metoda reinforcement učenja u ovom slučaju je bio 'izbor kolone'. Stanje u smislu metoda reinforcement učenja u ovom slučaju su pojedini slogovi datoteke. Svaki test je proveden u deset obrada sa po 25 iteracija. Početno stanje na prvom koraku svake iteracije je zadržavano u svakoj narednoj iteraciji. Na početku svake obrade odnosno grupe iteracija pravljen je novi izbor početnog stanja. Sve simulacije su provedene korištenjem Excell tabela odnosno makroa koji su za ove potrebe razvijeni potupno samostalno.

Testiranja su pokazala značajne razlike prije svega u dinamici konvergencije. Razlike su posebno vidljive u slučajevima kada unutar predviđenih deset obrada postoji više različitih početnih stanja. Ovo je važan rezultat jer pokazuje da algoritam ima aktivnu funkciju istraživanja (exploration) koja nakon određenog broja koraka daje prednost vrijednostima koja su se pokazala kao najbolja u smislu očekivanog povrata.

Rezultat reinforcement učenja je politika. U ovom slučaju politika je redoslijed atributa sa vrijednostima putem kojih je pravljen izbor akcija odnosno koje su vrijednosti koje su imale funkciju kriterija tokom algoritma. Ovakva politika odražava u velikoj mjeri analitički proces u kojem se, počev od numeričkog atributa, vrijednosti postepeno filtriraju i to određenim redoslijedom. Primjer je izbor dana datuma unutar iznosa koji počinju određenim nizom cifara, zatim izborom dana u sedmici za datume iz prethodnog koraka i tako dalje. Kao i za Benfordov zakon, rezultujuća politika ne mora nužno ukazivati na anomaliju, prevaru ili bilo koji drugi oblik devijacije. Ako su predmet analize datoteke na računaru politika daje sliku rada sistema u smislu sklonosti da se određene datoteke kreiraju pod određenim uslovima. U finansijskom poslovanju politika daje npr. obrazac ponašanja klijenta u smislu navika da određene transakcije obavlja po određenom obrascu. Kao i u slučaju Benfordovog zakona, posljednju riječ ima analitičar koji treba biti u mogućnosti da interpretira dobijene podatke.

Nakon svega, može se reći da Benfordov zakon daje velike mogućnosti u analitičkim i forenzičkim postupcima. Osnovni razlozi su jednostavna implementacija i potpuna

nezavisnosti od prirode samih podataka. Svakim danom je sve više primjera njegove primjene. Ovaj tekst je pokazao da primjena u reinforcement učenju otvara potpuno nove mogućnosti.

9 Literatura

References

[1] Frank Benford [1938], The Law of Anomalous Numbers, Proceedings of the American Philosophical Society, Vol. 78, No. 4, p. 551-572

[2] Tamas Lolbert, Digital analysis : Theory and applications in auditing, Hungarian Statistical Review, Special number 10

[3] Mark Nigrini, I've Got Your Number, Journal of Accountancy, May 1999, 187, 5; ABI/INFORM Global pg. 79 - 83

[4] Mehmed Kantardžić, Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons 2003

[5] Efrim Boritz, Fletcher Lu, Method of Data Analysis, patentna prijava US 2008/0208946 A1, 28.08.2008.

[6] Sukanto Bhattacharya, Dongming Xu, Kuldeep Kumar, An Artificial Neural Network-based Analytical Review Procedure for Detection of Manipulated Dataset

[7] Bruce Busta, Randy Weinberg, Using Benford's Law and Neural Networks as a Review Procedure, 14.01.1998

[8] Richard S. Sutton, Andrew G. Barto, Reinforcement Learning: An Introduction,

[9] John Lenz, Reinforcement Learning and the Temporal Difference Algorithm,

[10] Ian Gent, Toby Walsh, Benford's Law

[11] Peter N. Posch, A Survey on Sequences and Distribution Functions satisfying the First-Digit-Law, Department of Finance, University of Ulm, Germany, draft version : 25th October 2004

[12] Bojan Radman, Benfordov zakon, Hrvatski matematički elektronički list, Vol. 5, http://e.math.hr/benford/index.html

[13] Steven J. Miller, Mark J. Nigrini, Order Statistics And Benford's Law, Research article, Hindawi Publishing Corporation, International Journal of Mathematics and Mathematical Sciences, Volume 2008, Article ID 382948, 19 pages, doi:10.1155/2008/382948

[14] Steven J. Miller, Ramin Takloo-Bighash, An Invitation to Modern Number Theory, Princeton University press, Princeton and Oxford, ProbStat_Chaps8And9 June 7, 2007

[15] Bret Flehinger, On the Probability That a Random Integere Has Initial Digit A, American Mathematical Monthly, 73:1056-1061, 1966.

[16] Theodore P. Hill, A Statistical Derivation of the Significant-Digit Law, School of Mathematics and Center for Applied Probability Georgia Institute of Technology, Atlanta, March 20, 1996

[17] Elise Janvresse, Thierry de la Rue, From Uniform Distributions to Benford's Law, Universit e de Rouen, LMRS, UMR 6085 - CNRS

[18] Diaconis Persi, The Distribution of leading Digits and Uniform Distribution mod 1, The Annals of Probability, 1977, Vol. 5, No. 1, p. 72-81

[19] Dean Brooks, Naked-Eye Quantum Mechanics - Practical Applications of Benford's Law for Integer Quantities, Frequencies, The Journal of Size Law Applications, Special Paper #1, Ekaros Analytical Inc.

[20] Werner Hürlimann, Generalizing Benford's Law Using Power Laws : Application to Integer Sequences, Research Article, Hindawi Publishing Corporation International Journal of Mathematics and Mathematical Sciences Volume 2009, Article ID 970284, 10 pages, doi:10.1155/2009/970284

[21] Oded Kafri, Entropy Principle in Direct Derivation of Benford's Law, Varicom Communications

[22] Pieter C Allaart, An Invariant Sum Characterization of Benford s Law, Department of Mathematics and Computer Science De Boelelaan a HV Amsterdam The Netherlands

[23] Mark J. Nigrini, Steven J. Miller, Data Diagnostics Using Second Order Tests Of Benford's Law, June 1, 2006

[24] http://reocities.com/CapeCanaveral/hangar/4577/elmuhsi.htm

[25] John Morrow, Benford's Law, Families of Distributions And A Test Basis, January 22, 2009

[26] Li ZhiPeng, Cong Lin, Wang Huajia, Discussion On Benford's Law And Its Application, arXiv:math/0408057v2 [math.ST] 4 Oct 2004

[27] Bartolo Luque, Lucas Lacasa, The first digit frequencies of primes and Riemann zeta zeros tend to uniformity following a size-dependent generalized Benford's law, arXiv:0811.3302v1 [math.NT] 20 Nov 2008

[28] Stefan Günnel, Karl-Heinz Tödter, Does Benford's law hold in economic research and forecasting?, Discussion Paper, Series 1: Economic Studies, No 32/2007, Deutsche Bundesbank

[29] Mark J. Nigrini, Linda J. Mittermaier, The Use of Benford's Law As An Aid In Analytical Procedures, Auditing, A Journal of Practice & Theory, Vol. 16, No. 2, Fall 1997

[30] Cindy Durtschi, William Hillison, Carl Pacini, The Effective Use Of Benford's Law To Assist In Detecting Fraud In Accounting Data, Journal of Forensic Accounting, 1524-5586, Vol V(2004), pp. 17-34

[31] Mark Nigrini, Continuous Auditing, Ernst & Young Center for Auditing Research and Advanced Technology, University of Kansas, August 30, 2000

[32] Wendy K., Tam Cho, Brian J. Gaines, Breaking the (Benford) Law: Statistical Fraud Detection in Campaign Finance, The American Statistician, August 2007, Vol. 61, No. 3

[33] Christopter F. Dumas, John H. Devine, Detecting Evidence of NonCompliance In SelfReported Pollution Emissions Data An Application of BenforcTs Law, Selected Paper, American Agricultural Economics Association Annual Meeting Tampa, FL July 30-August 2, 2000

[34] www.wikipedia.com, z-test

[35] Mark Nigrini, Using Benford's Law To Detect Fraud, Chapter 5, NO. 02-5410

[36] A Guide to Benford's Law, A CaseWare IDEA Research Department document, June 24, 2003

[37] Allen Long, Olga Pavlova, Olga Pechinkina, Lillian Clark, First Significant Digit Law And Fraud Detection In Factoring Data, London South Bank University

[38] Fletcher Lu, J. Efrim Boritz, Detecting Fraud in Health Insurance Data: Learning to Model Incomplete Benford's Law Distributions, School of Computer Science, University of Waterloo

[39] David Groce, Benford's Law, Lyncean Group, March 23, 2005

[40] Alex V. Kontorovich, Steven J. Miller, Benford's Law, Values of L-functions And the 3x + 1 Problem, arXiv:math/0412003v2 [math.NT] 27 Jun 2005

[41] Bartolo Luque, Lucas Lacasa, The first-digit frequencies of prime numbers and Riemann zeta zeros, Proceedings of the Royal Society A, doi:10.1098/rspa.2009.0126, published online

[42] Arno Berger i Theodore P. Hill, Newton's Method Obeys Benford's Law, The American Monthly, August / September 2007; 114, 7; Research Library

[43] Lawrence M. Leemis, Bruce W. Schmeiser, Diane L. Evans, Survival Distributions Satisfying Benford's Law, The American Statistician, August 2000, Vol. 54, No. 3

[44] Steven J. Miller, Some Thoughts on Benford's Law, November 11, 2004

[45] Arno Berger, S. Sigmund, On the distribution of mantissae in nonautonomous difference equations, Journal of Difference Equations and Applications, Vol. 13, Nos. 8–9, August–September 2007, 829–845

[46] Zoran Jasak, Banjanovich-Mehmedovich, Detecting Anomalies by Benford's Law, ISSPIT 2008

[47] Jesus Gonzalez-Garcia, Gonzalo Pastor, Benford's Law and Macroeconomic Data Quality, IMF Working Paper, Statistic Department, January 2009

[48] http://www.nigrini.com/images/file_sizes.html

[49] Arno Berger, Leonid A. Bunimovich, Theodore P. Hill, One-Dimensional Dynamical Systems And Benford's Law, Transaction of the American Mathematical Society, Volume 357, Number 1, Pages 197-219, S 0002-9947(04)03455-5, Article electronically published on April 16, 2004

[50] Charley Tichenor, Bobby Davis, Why Benford's Law Works For Function Point Analysis

[51] Frank Albert, Benford's Law,

[52] Frederic Sandron, Do Populations Conform to the Law of Anomalous Numbers?, Institut National d'Etudes Démographiques, 2002/4-5 - Volume 57, ISSN 1634-2941 | pages 753 - 761

[53] Dongdong Fu, Yun Q. Shi, Wei Su, A generalized Benford's law for JPEG coefficients and its applications in image forensics; Dept. of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark; U.S. Army Communication-Electronics RD&E Center, Intelligence & Information Warfare Directorate, Fort Monmouth

[54] Denka Markova, Linda Njoh, Michael Lloyd, Ph.D., Benford's Law And The Bible

[55] Abdul Majid Motahari, Benford's Law and the Quran, http://www.submission.org/miracle/benford.html

[56] Walter R. Mebane, Election Forensics: Vote Counts and Benford's Law, July 17, 2006, Prepared for presentation at the 2006 Summer Meeting of the Political Methodology Society, UC-Davis, July 20 - 22

[57] K. Calix, M. Connors, D. Levy, H. Manzar, G. McCabe, S. Westcott, Stylometry for E-mail Author Identification and Authentication, Proceedings of CSIS Research Day, Pace University, May 2008

[58] Yun-Quing Shi, DongDong FU, Apparatus and Methods for Generalized Benford's Law Analysis Of DCT And JPG Ceofficients, Patentna prijava US 2008/0031535 A1, 07 Februar 2008.

[59] Engelbrecht A. P., Computational Intelligence : An Introduction, pogl. 6,

[60] Greenwald Amy, Introduction to Artifical Intelligence, Lectures 19 -22, March - May 2009, Brown University Providence, Rhode Island

[61] Bill Smart, Reinforcement Learning : A User's Guide, Department of Computer Science and Engineering, Washington University in St. Louis

[62] Gerhard Neumann, The Reinforcement Learning Toolbox, Reinforcement Learning for Optimal Control Tasks, Diplomski rad, University of Technology, Graz, Maj 2005

[63] Fletcher Lu, J. Efrim Boritz, Dominic Covvey, Adaptive Fraud Detection using Benford's Law, Canadian Institute of Chartered Accountants i University of Waterloo

[64] Fletcher Lu, Uncovering Fraud in Direct Marketing Data with a Fraud Auditing Case Builder, Department of Math and Computer Science, University of Maryland Eastern Shore

[65] Joanne Lee, Wendy K. Tam Cho, George G. Judge, Stigler's approach to recovering the distribution of first significant digits in natural data sets, Statistics and Probability Letters 80 (2010) 82 88, journal homepage: www.elsevier.com/locate/stapro

[66] Bahar Kaynar, Arno Berger, Theodore P. Hill, Ad Ridder, Finite-state Markov Chains Obey Benford's Law, Tinbergen Institute Discussion Paper, TI 2010-030/4, March 1, 2010