TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
MATEMAATILISE STATISTIKA INSTITUUT

Gea Pajula

Benfordi seadus

Bakalaureusetöö (6 EAP)

Juhendaja: Anne Selart, MSc

Tartu 2014

Benfordi seadus

Käesolevas bakalaureusetöös antakse ülevaade Benfordi seadusest, mille kohaselt algab arv numbriga 1 tõenäosusega $\log 2 \approx 0{,}301$, numbriga 2 tõenäosusega $\log(3/2) \approx 0{,}176$ ja nii monotoonselt kahanevalt kuni tõenäosuseni $\log(10/9) \approx 0{,}046$, et esimene number on 9. See seadus kehtib paljudes andmestikes, näiteks rahvaarvude, riikide pindalade, aktsiaturgude indeksite ja valimistulemuste korral. Antud töös uuritakse ka 2013. aasta kohaliku omavalitsuse volikogu valimiste tulemuste Benfordi seaduse järgmist.

Märksõnad: Benfordi seadus, matemaatiline statistika, tõenäosus, invariantsus, valimised, kohalikud omavalitsused.

Benford's Law

The purpose of this bachelor thesis was to give an overview of Benford's Law, which states that a number has leading significant digit 1 with probability $\log 2 \approx 0.301$, leading significant digit 2 with probability $\log(3/2) \approx 0.176$ and so on monotonically down to probability $\log(10/9) \approx 0.046$ for leading digit 9. This law has been found to apply to a wide variety of data sets, for example population numbers, areas of world states, stock market indices and election results. In this thesis is analysed the results of Estonian local government council election in 2013 conforming to this law.

Keywords: Benford's Law, mathematical statistics, probability, invariance, elections, local governments.

Sisukord

Sissejuhatus

Käesolevas bakalaureusetöös antakse ülevaade Benfordi seadusest, mille kohaselt algab arv numbriga 1 tõenäosusega $\log 2 \approx 0{,}301$ (kus $\log$ tähistab arvu kümnendlogaritmi), numbriga 2 tõenäosusega $\log(3/2) \approx 0{,}176$ ja nii monotoonselt kahanevalt kuni tõenäosuseni $\log(10/9) \approx 0{,}046$, et esimene number on 9. Seda seadust järgivad paljud arvuhulgad: Fibonacci arvud, riikide pindalad, aktsiaturgude indeksid, valimiste tulemused jne. Benfordi seadust saab kasutada ka pettuste avastamiseks: vähene kooskõla seadusega osutab suurele tõenäosusele, et andmeid on muudetud.

Bakalaureusetöö on jagatud viieks osaks. Esimeses peatükis on kirjeldatud Benfordi seaduse avastamise ja tõestamise lugu. Teises osas on uuritud Benfordi enda analüüsitud andmestike tulemusi, toodud näiteid Benfordi seadust järgivatest arvuhulkadest ning ka näited andmestikest, mille korral antud seadus ei kehti. Samuti vaadatakse Benfordi seaduse kohta aastas ilmunud artiklite arvude esinumbrite jaotust. Kolmandas peatükis on ära toodud seaduse matemaatiline definitsioon ja intuitiivne tõestus. Neljandas osas tutvustatakse Benfordi seaduse kehtimise kontrollimiseks kasutatavaid mantissi ja keskmise hälbe testi. Viiendas peatükis uuritakse 2013. aasta kohaliku omavalitsuse volikogu valimiste tulemuste Benfordi seaduse järgimist.

Benfordi seaduse kohta aastas ilmunud artiklite arvud on saadud internetileheküljelt *http://www.benfordonline.net/* ja valimiste tulemuste failid Vabariigi Valimiskomisjoni käest. Mõlemal juhul on soovitud arvude eraldamiseks kasutatud Pythonit (versioon 3.3.3), programmikood on toodud lisas 1. Edasine analüüs, mille programmikood on lisas 2, on tehtud rakendustarkvara R (versioon 3.0.2) abil, kasutatud on teeke BenfordTests (Joenssen, 2013) ja benford.analysis (Cinelli, 2014). Töö on kirjutatud tekstitöötlusprogrammiga Texmaker (versioon 4.1.1).

Autor tänab juhendajat Anne Selartit rohkete nõuannete, paranduste ning töö struktuuri puudutavate ideede eest.

1. Ajalugu

Järgnev Benfordi seaduse avastamise ja tõestamise lugu põhineb Matthews 1999. aasta artiklil.

Arvude algusnumbrite jaotuse avastamisel mängivad olulist rolli enne kalkulaatoreid kasutusel olnud logaritmitabelid, mille abil lihtsustati arvude korrutamist ja jagamist. Kiireim viis kahe arvu korrutamiseks oli leida nende logaritmid tabelist, need liita ning kasutada summa antilogaritmi korrutise saamiseks. (Nigrini, 2012, lk 2)

Ameerika astronoom Simon Newcomb märkas, et logaritmitabelite esimesed leheküljed kuluvad viimastest lehtedest kiiremini. Tundus, et numbritega 1 ja 2 algavaid arvusid kasutatakse arvutustes sagedamini kui numbritega 8 ja 9 hakkavaid. Selle tähelepaneku põhjal sõnastas Newcomb reegli: arvude esinemise tõenäosusjaotus on selline, et nende mantissid on võrdtõenäolised (Newcomb, 1881). Ta kasutas logaritmilist jaotust, et leida iga numbri esimesele ja teisele numbrikohale sattumise tõenäosus (vt Tabel 1.1). Samuti väitis ta, et kolmanda numbrikoha jaoks on tõenäosused peaaegu võrdsed ning neljanda ja järgnevate numbrikohtade korral on erinevused märkamatud.

Rohkem kui pool sajandit hiljem pani logaritmitabelite ebaühtlast kulumist tähele ka füüsik Frank Benford. Kogudes rohkem kui $20\ 000$ arvu erinevatest andmestikest (näiteks erisoojused, arvud ajalehe esikülgedel), sai ka Benford samad esimeste numbrite esinemistõenäosused, mis viitavad logaritmilisele jaotusele. Sarnaselt Newcombile ei andnud ka Benford head selgitust nähtuse tekkepõhjustest, kuid näited selle esinemisest olid piisavad, et antud seadust tema järgi nimetataks.

Alles aastakümneid hiljem 1960ndatel leiti selgitus, miks see seadus kehtib nii paljudel erinevatel andmetel. Nimelt, kui on olemas mingi üldine seadus, mis määrab ära nii jõgede pindalad, ainete soojusmahtuvused kui ka arvud firma maksuaruandes, siis see arvude jaotus ei tohiks olla mõjutatud sellest, mis ühikutes antud näide fikseeritakse: kas jõe pindala esitatakse meetri- või tollisüsteemis, kas soojusmahtuvus leitakse džau-

Tabel 1.1: Iga numbri esimesele ja teisele numbrikohale sattumise tõenäosus

Number	Esimene numbrikoht	Teine numbrikoht
0	...	0,1197
1	0,3010	0,1139
2	0,1761	0,1088
3	0,1249	0,1043
4	0,0969	0,1003
5	0,0792	0,0967
6	0,0669	0,0934
7	0,0580	0,0904
8	0,0512	0,0876
9	0,0458	0,0850

lides kelvini kohta või kilokalorites kelvini kohta või kas firma peab raamatupidamist eurodes või dollarites. Matemaatik Roger Pinkham tõestaski, et kui esinumbrid järgivad Benfordi seadust, siis andmete teisendamine ei muuda esinumbrite jaotust. Lisaks näitas ta seda, et Benfordi seadus on ainuke viis selliste omadustega numbrite jaotumiseks. Teisisõnu, iga skaala invariantsuse nõudega numbrite jaotus on paratamatult Benfordi seadus. Tänu Pinkhami tööle muutus seadus tõsiseltvõetavamaks ja hakati mõtlema võimalikest kasutusaladest.

Järgnevalt püüti selgusele jõuda, missuguste andmete korral on seaduse kehtivust oodata. Saadi aru, et arve peab olema piisavalt palju ning peab olema võimalik saada põhimõtteliselt igat väärtust, seega ei tohi esineda rangeid piire. Näiteks 10 erineva saia hinna jaoks ei ole mõtet oodata Benfordi seaduse kehtimist, sest valimimaht on liiga väike ning turumajanduse tõttu jäävad hinnad kitsasse fikseeritud vahemikku.

Benfordi seaduse tekkimise põhjuse avastas 1995. aastal Theodore Hill. Ta jõudis mõtteni, et see tuleneb erinevatest viisidest, kuidas erinevad mõõtmistulemused tekivad.

Lõppkokkuvõttes on kõik universumis mõõdetav erinevate protsesside tulemus: näiteks aatomi juhuslikest tõugetest või geneetikast tingitud. Hill sõnastas teoreemi, mille kohaselt järgivad arvud Benfordi seadust, kui need on pärit juhuslikest valimitest, mis on võetud juhuslikult valitud tõenäosusjaotustest (Hill, 1995). Andmestikud rahvaloendustulemustest kuni aktsiahindadeni on tingitud juhuslikus koosluses igasugustest jaotustest ning järgivad teoreemi kohaselt antud seadust. Benfordi enda uurimus ja paljud teised on näidanud, et see tõesti kehtib.

Samast, 1995ndast aastast alates kasvas ühes aastas Benfordi seaduse kohta avaldatud artiklite arv kiirelt: kui seni jäi aastas ilmunud artiklite arv enamasti alla kümne, siis edasi toimus järkjärguline kasv kuni 2009. aastani, kui ilmus kokku 64 artiklit. Viimase 5 aasta jooksul on avaldatud artiklite arv olnud kahanevas tendentsis. (*http://www.benfordonline.net/*)

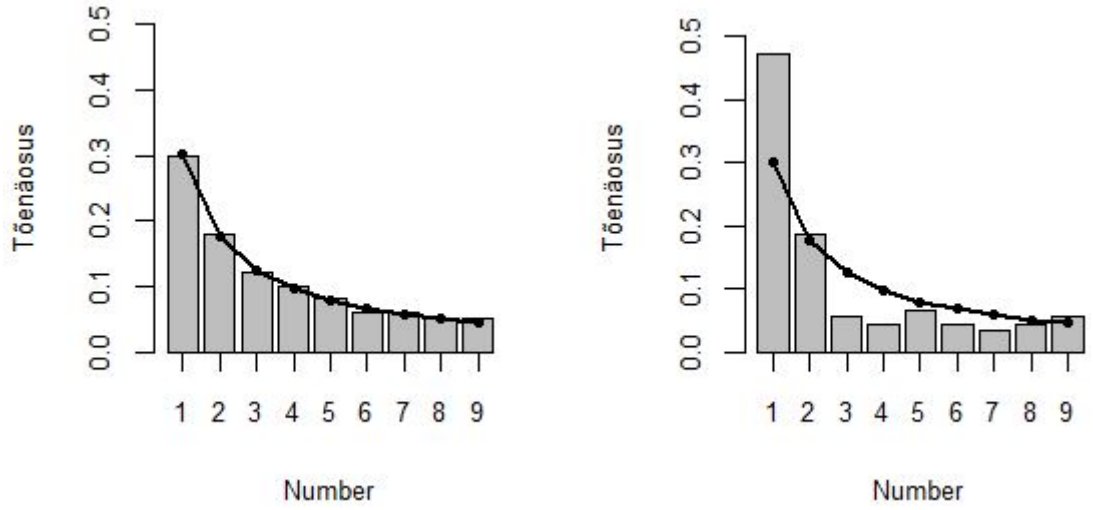
2. Näited andmetest

Frank Benford (Benford, 1938) analüüsis 20 229 arvu esimesi numbrikohti 20 erinevast andmestikust: näiteks füüsika konstandid, arvud ajalehe esikülgedel, erisoojused, molekulmassid, aatommassid, arvud ühes ajakirjas, majanumbrid aadressides ja suremuskordajad. Saadud esinumbrite osakaalud ning kõikide andmestike pealt leitud keskmised osakaalud, kus on parandatud Nigrini (2012) poolt viidatud kaks arvutusviga, on esitatud lisas 3 tabelis 1. Leitud tõenäosuste järgi sai Benford aru, et esinumbrite jaotus on logaritmiline ning kooskõla antud seadusega on parem juhuslikult tekkinud andmestikel.

Benfordi andmestikest on logaritmilise jaotusega kõige paremas kooskõlas ajalehtede esikülgedelt leitud arvud (ei arvestata kuupäevi ja sõnadega välja kirjutatud numbreid), mille korral on kõikide esimeste numbrite osakaalud enam-vähem võrdsed oodatavate tõenäosustega (vt joonis 2.1 a). Samuti vastavad Benfordi seadusele esimesed 342 majanumbrit ühes *American Men of Science* väljaandes ning kõik arvud peale kuupäevade ja leheküljenumbrite ajakirja *Readers' Digest* väljaandes.

Suurimad kõrvalekalded oodatud tõenäosustest on tihedalt seotud andmestikel: erisoojused, füüsika konstandid ning molekuli- ja aatomimassid. Aatommasside esimeste numbrite osakaalude suurimad erinevused oodatavatest tõenäosustest on numbrite 1, 3 ja 4 korral (vt joonis 2.1 b).

Kuigi kõik Benfordi uuritud andmestikud ei järgi antud seadust, siis andmestikke koos vaadates esimeste numbrite jaoks antud seadus kehtib. Saadud tulemuse: kui võtta juhuslikest jaotustest juhuslikke valimeid, siis saadud andmestik peaks olema Benfordi seadust järgiv, tõestas Hill (1995).

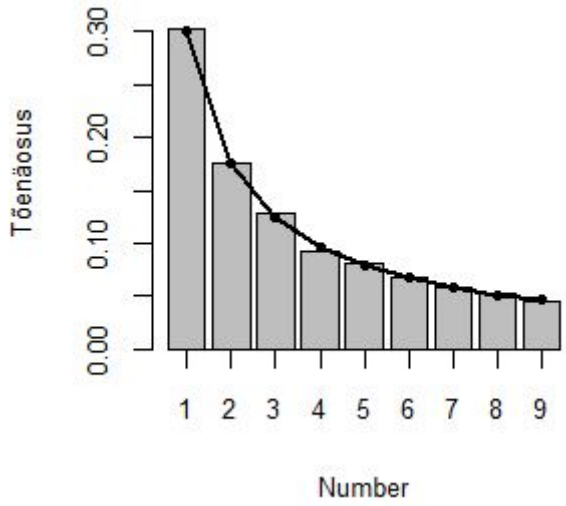


Joonis 2.1: Benfordi uuritud andmestike (a) parim kooskõla ajalehtede esikülgedel olnud arvude ja (b) halvim kooskõla aatommasside esimeste numbritega, mille vastavad osakaalud on toodud tulpdiagrammina ja oodatavad tõenäosused joondiagrammina.

Benfordi seadust järgivad täpselt mitmed arvujadad: faktoriaalid, number kahe astmed ja Fibonacci arvud (Berger, Hill ja 2011). Fibonacci arvud on suure hajuvusega, näiteks 400. arv katab 84 numbrikohta. Vaadates 400 esimest Fibonacci arvu, siis esinumbrite osakaalu ja oodatava tõenäosuse vahel kahe esimese komakoha seas erinevusi ei ole (vt joonis 2.2).

Andmeid testides on leitud, et antud seadust järgivad ka näiteks riikide pindalad ja rahvaarvud (Fewster, 2009), järvede pindalad, jõgede pikkused, aktsiaturgude indeksid ja failide suurused arvutis (Shao ja Ma, 2010). Ka raamatupidamis- ja finantsdokumentide, keskkonna raportite ja valimistulemuste korral peaks Benfordi seadus kehtima (Alexander, 2009). Kuna paljud andmestikud peaksid Benfordi seadust järgima, saab seda rakendada pettuste avastamiseks. Näiteks kasutatakse seda kahtlaste maksudeklaratsioonide leidmiseks, valimiste tulemustes pettuste tuvastamiseks ja muudetud digitaalpiltide välja selgitamiseks (Berger ja Hill, 2011).

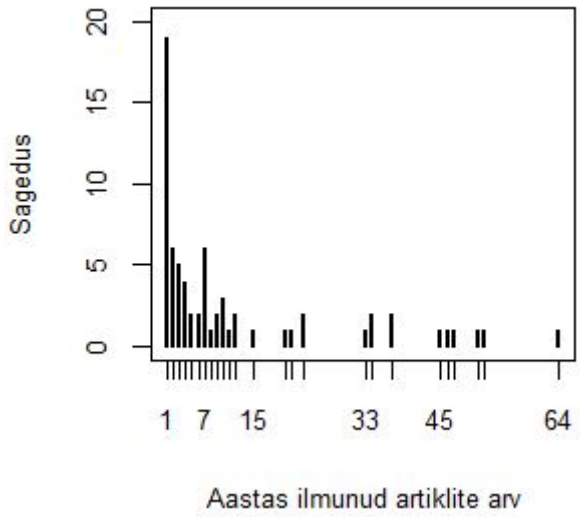
Kui andmed ei ole kooskõlas antud seadusega, tuleb neid lähemalt uurida ning on



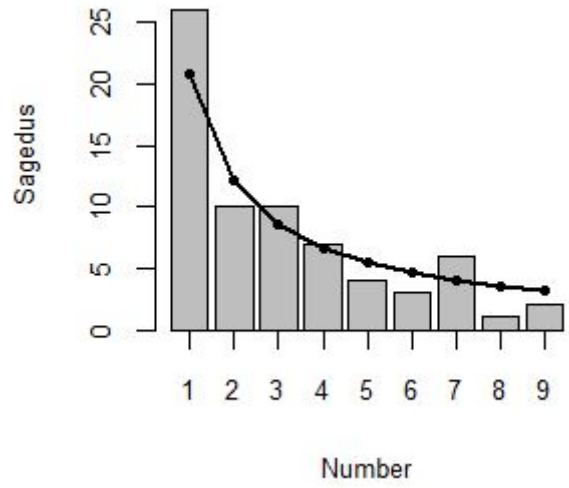
Joonis 2.2: Esimese 400 Fibonacci arvu esimeste numbrite osakaalud on toodud tulpdiagrammina ja oodatavad tõenäosused joondiagrammina.

suur tõenäosus, et arve on muudetud, kuid võib ka juhtuda, et need andmed üldse ei peagi Benfordi seadust järgima (Nigrini, 2012). Näiteks ei ole seaduse kehtimist oodata loterii ja telefoninumbrite (Shao, Ma, 2010), pangakonto ja lennunumbrite jaoks (Nigrini, 2012) ning samuti ei järgi Benfordi seadust algarvud (Berger ja Hill, 2011).

Järgnevalt uurime, kas Benfordi seadust võiks järgida ka antud seaduse kohta aastas ilmunud artiklite arvud. Veebilehe *http://www.benfordonline.net/list/chronological* andmetel on Benfordi seaduse kohta kokku ilmunud 799 artiklit, esimene 1881. aastal, järgmine 1912. aastal ning viimane käesoleval ehk 2014. aastal. Neist 65 aastal ei ilmunud ühtegi artiklit, väärtuseid 0 Benfordi seaduse testimisel ei arvestata. Üks artikkel ilmus 19 aastal, suurim aastas ilmunud artiklite arv on 64 aastal 2009 (vt joonis 2.3). Keskmine aastas ilmunud artiklite arv on 6 ja mediaan 1. Aastas ilmunud artiklite arvude esimeste numbrite seas on numbriga 1, 3 ja 7 algavaid oodatust rohkem ning numbriga 2, 5 ja 6 hakkavaid oodatust vähem (vt joonis 2.4). Jooniselt näeme, et esinumbrite osakaalud ei ole täpselt Benfordi seaduse jaoks oodatavad, kuid on tendents antud jaotuse suunas. Arvatavasti kooskõla paraneb aastate jooksul.



Joonis 2.3: Benfordi seaduse kohta aastas ilmunud artiklite sagedused.



Joonis 2.4: Benfordi seaduse kohta aastas ilmunud artiklite arvude esimeste numbrite sagedused tulpdiagrammina ja oodatavad sagedused joondiagrammine.

3. Benfordi seaduse teooria

Selles peatükis esitame Benfordi seaduse definitsiooni ja intuitiivse tõestuse, mis põhineb eeldusel, et uuritavate arvude logaritmid jaotuvad normaaljaotuse kohaselt. Definitsiooni alapeatükis on tuginetud allikale Hill (1995) ja tõestuse aluseks on Fewster (2009).

3.1. Definitsioon

Olgu $X > 0$ juhuslik suurus jaotusfunktsiooniga F. See juhuslik suurus on avaldatav ühesel kujul tüve R ja astendaja n abil (arvusüsteemis alusel 10) järgnevalt

$$X = R \cdot 10^n, \text{ kus } 1 \leq R < 10, R \in \mathbf{R} \text{ ja } n \in \mathbf{Z}. \tag{3.1.1}$$

Näiteks $314 = 3{,}14 \cdot 10^2$ ja $0{,}0314 = 3{,}14 \cdot 10^{-2}$. Tähistame $D_1, D_2, \ldots$ abil (alusel 10) numbrikoha funktsioone. Näiteks $D_1(0{,}0314) = 3, D_2(0{,}0314) = 1, D_3(0{,}0314 = 4)$.

Benfordi seaduse kohaselt algab arv numbriga 1 tõenäosusega $\log 2 \approx 0{,}301$, numbriga 2 tõenäosusega $\log(3/2) \approx 0{,}176$ ja nii monotoonselt kahanevalt kuni tõenäosuseni $\log(10/9) \approx 0{,}046$, et esimene number on 9. Seega esimese numbri jaoks saame

$$P(D_1 = d) = \log[(d+1)/d] = \log(1 + d^{-1}),\ d = 1, 2, \ldots, 9.$$

Teise numbri korral kehtib

$$P(D_2 = d) = \sum_{k=1}^{9} \log[1 + (10k + d)^{-1}],\ d = 0, 1, 2, \ldots, 9.$$

Näiteks tõenäosuseks, et teine number on 1, saame $P(D_2 = 1) = \sum_{k=1}^{9} \log[1 + (10k + 1)^{-1}] = \log(1+(11)^{-1}) + \log(1+(21)^{-1}) + \log(1+(31)^{-1}) + \log(1+(41)^{-1}) + \log(1+(51)^{-1}) + \log(1+(61)^{-1}) + \log(1+(71)^{-1}) + \log(1+(81)^{-1}) + \log(1+(91)^{-1}) = 0{,}1139$.

Numbrikohtade jaotuse saab esitada ka üldisemalt pideval skaalal, kui vaatame arvu tüve jaotust:

$$P(R \leq t) = \log t,\ t \in [1, 10).$$

Sellest saab tüvenumbrite ühisjaotuse

$$P(D_1 = d_1, \ldots, D_k = d_k) =$$

$$P(d_1{+}d_2 10^{-1}{+}d_3 10^{-2}{+}\cdots{+}d_k 10^{-(k-1)} \le R < d_1{+}d_2 10^{-1}{+}d_3 10^{-2}{+}\cdots{+}(d_k{+}1)10^{-(k-1)}) =$$

$$\log(d_1{+}d_2 10^{-1}{+}d_3 10^{-2}{+}\cdots{+}(d_k{+}1)10^{-(k-1)}){-}\log(d_1{+}d_2 10^{-1}{+}d_3 10^{-2}{+}\cdots{+}d_k 10^{-(k-1)}) =$$

$$\log \frac{d_1 + d_2 10^{-1} + d_3 10^{-2} + \cdots + (d_k + 1)10^{-(k-1)}}{d_1 + d_2 10^{-1} + d_3 10^{-2} + \cdots + d_k 10^{-(k-1)}} =$$

$$\log(1 + \frac{10^{-(k-1)}}{d_1 + d_2 10^{-1} + d_3 10^{-2} + \cdots + d_k 10^{-(k-1)}}) =$$

$$\log(1 + \frac{10^{-(k-1)} \cdot 10^{k-1}}{d_1 10^{k-1} + d_2 10^{k-2} + d_3 10^{k-3} + \cdots + d_k 10^0}) =$$

$$\log[1 + (\sum_{i=1}^{k} d_i \cdot 10^{k-i})^{-1}],$$

$$d_1 \in 1, 2, \ldots, 9,\ d_j \in 0, 1, \ldots, 9,\ j = 2, \ldots, k.$$

Näiteks $P(D_1 = 3, D_2 = 1, D_3 = 4) = \log[1 + (3 \cdot 10^2 + 1 \cdot 10^1 + 4 \cdot 10^0)^{-1}] = \log(1 + 314^{-1}) = 0{,}0014$.

Tüvenumbrid on üksteisest sõltuvad. Näiteks kahe tüvenumbri korral $P(D_1 = 3, D_2 = 1) = \log(1 + 31^{-1}) = 0{,}0138$, kuid $P(D_1 = 3)P(D_2 = 1) = \log(1 + 3^{-1}) \cdot 0{,}1139 = 0{,}1249 \cdot 0{,}1139 = 0{,}0142$.

3.2. Tõestus

Juhuslik suurus $X > 0$ on avaldatav seose 3.1.1 järgi kujul $X = R \cdot 10^n$. Arvu X esimene number on sama, mis tüvel R. Näiteks kui X algab numbriga 1, siis võib ta olla poollõikudes $X \in [1; 2), X \in [10; 20), X \in [100; 200)$ jne. Kuid tüve R korral vaatleme ühte poollõiku: arvu X esimene number on 1, kui $1 \le R < 2$. Seda seost logaritmides saame

$$\log 1 \le \log R < \log 2 \ \text{ ehk } \ 0 \le \log R < 0{,}301. \tag{3.2.1}$$

Logaritmides seose 3.1.1 mõlemat poolt, saame

$$\log X = \log(R \cdot 10^n) = \log R + \log 10^n = \log R + n,$$

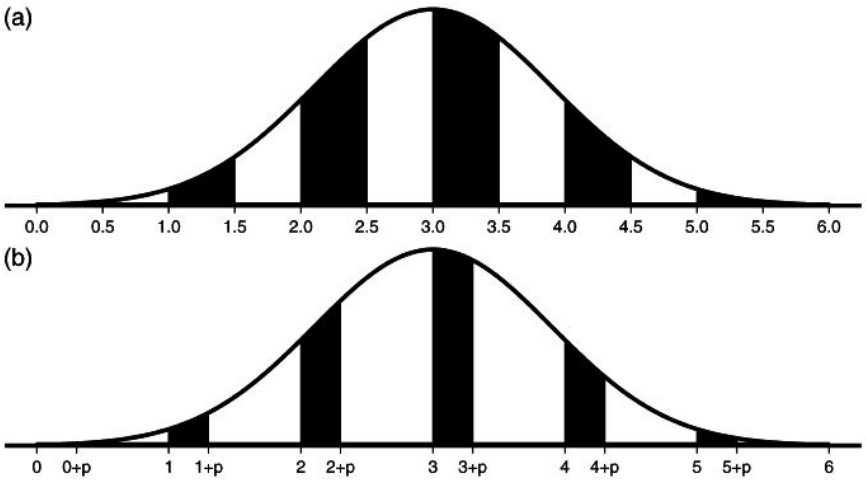
millest $\log R = \log X - n$ on arvu X logaritmi murdosa ehk mantiss. Seose 3.2.1 järgi saame, et numbriga 1 algamiseks peab mantiss olema poollõigus $[0; 0{,}301)$. Asendades $\log R$ seoses 3.2.1, saame $0 \leq \log X - n < 0{,}301$, millest

$$n \leq \log X < n + 0{,}301. \tag{3.2.2}$$

Näiteks $X = 124 = 1{,}24 \cdot 10^2$ korral $\log X = 2{,}093$ ja $2 \leq 2{,}093 < 2{,}301$ ning esimene number on 1. Samas näiteks $X = 76 = 7{,}6 \cdot 10^1$ puhul $\log X = 1{,}881 \notin [1; 1{,}301)$, kuid $X = 76$ ei alga ka numbriga 1.

Selleks, et jõuda X numbriga 1 algamise tõenäosuseni, vaatame kahemõõtmelist kübarakujulist pinda, mis on ühtlaselt kaetud võrdsete laiustega mustade ja valgete triipudega (Joonis 3.2.1 a). Sel juhul umbes pool kübarast on musta värvi. Üldisemalt, kui mustad triibud katavad kübarakuju servast osakaalu p, siis umbes p osa kujundist on musta värvi. Umbkaudne hinnang paraneb triipude arvu suurenedes: palju peenikesi triipe arvestab kübarakuju asümmeetriat väiksemast arvust laiematest triipudest paremini. Keskmiselt, liigutades triipe juhuslikult vasakule või paremale, katab triibutatud ala kübarakuju kogupindalast osa p.

Vastavalt seosele 3.2.2 asuvad numbriga 1 algavad X väärtused ühtlaste vahedega triipudena $\log X$ skaalal, kus iga triibu laius on $0{,}301$. Olgu $\log X$ tihedusel kuju, mis on toodud joonisel 3.2.1 b. Arvud X esimese numbriga 1 vastavad triipudele, millest igaüks katab poollõigu $[n; n + 0{,}301)$ mingi täisarvu n korral. Iga täisarvu n jaoks on üks triip $\log X$ tihedusfunktsioonis, seega triipude koguarv on täisarvude arv $\log X$ skaalal. Arvu X numbriga 1 algamise tõenäosus on sama, mis tõenäosus, et $\log X$ asub mustal triibul. Mustad triibud katavad $p = 0{,}301$ osa servast, seega umbes $p = 0{,}301$ pindalast on must. Tihedusfunktsiooni korral pindala vastab tõenäosusele. Seega arvu X numbriga 1 algamise tõenäosus on umbes $p = 0{,}301$.



Joonis 3.2.1: (a) Kübarakuju vahelduvate sama laiade mustade ja valgete triipudega on ligikaudu poolenisti must ja poolenisti valge. (b) Kui mustad triibud katavad osakaalu p kübarakujutise servast, siis ligikaudu osa p kübara pindalast on musta värvi. Antud juhul $p = 0{,}301$ ning mustad triibud vastavad numbriga 1 algavatele arvudele X, kui kübar kujutab $\log X$ tihedusfunktsiooni.

Sarnaselt "X algab 1-ga" juhu selgitusele, saame seose üldkuju. Kirjutame $X = R \cdot 10^n, \text{ kus } 1 \leq R < 10, R \in \mathbf{R} \text{ ja } n \in \mathbf{Z}$. Arvu X esimene number on $d \in 1, 2, \ldots, 9$ parajasti siis, kui $d \leq R < d+1$. Viimase seosega samaväärselt kehtib $\log d \leq \log R < \log(d+1)$. Kui $\log X$ jaotus on sarnane normaaljaotusele, siis on ühtlase vahedega triipe $[n + \log d; n + \log(d+1))$ palju ning servast kaetud osa $\log(d+1) - \log d$ on võrdne triipude poolt kaetud pindalaga. Seega saame

$$P(\log d \leq \log R < \log(d+1)) = \log(d+1) - \log d = \log[(d+1)/d], \qquad (3.2.3)$$

mis ongi Benfordi seadusele vastav tõenäosus: $P(D_1(X) = d) = \log[(d+1)/d]$.

Artiklis Fewster (2009) on sama mütsiskeemi abil proovitud selgitada ka seda, millal andmed Benfordi seadust täitma peaksid. Nimelt, mida suurem on triipude arv, seda lähemale arvu 1-ga algamise tõenäosusele $0{,}301$ on osakaalu oodata, sest ühe triibu poolt põhjustatud lokaalne asümmeetria kübarakujus on paremini tasakaalustatud teiste trii-

pude poolt. Seepärast, mida suurem on triipude arv, seda paremini peaksid andmed Benfordi seadust järgima. Triipude arvu ei saa suurendada triipe üksteisele lähemale tuues, sest need peavad asuma poollõikudel $[0; 0{,}301), [1; 1{,}301), [2; 2{,}301)$ jne. Ainuke võimalus triipude arvu suurendamiseks on kübarakuju laiemaks muutmine. See tähendab $\log X$ jaotus peaks katma suurema ulatuse ehk X jaotus peaks hõlmama rohkem suurusjärke. Kui X saab väärtuseid $1 - 10^6$, siis $\log X$ katab kuus täisarvu, esitledes kuut triipu joonisel 3.2.1 b. See on Fewster (2009) järgi enamasti piisav, et andmed järgiksid Benfordi seadust.

Kuigi Fewster (2009) toodud tõestuse (mis sisuliselt eeldab $\log X$ normaaljaotust ehk X lognormaaljaotust) korral tõesti X hajuvuse suurendamine ehk mütsikuju venitamine parandab klappi Benfordi seadusega, ei pruugi see teistsuguse X jaotuse korral kehtida. Näiteks $X \sim U(1; 10^6)$, mis samuti katab kuus suurusjärku (on suure hajuvusega), korral esinumbrite jaotus ei järgi Benfordi seadust ja hajuvuse suurendamine ei muuda seda. Seega Fewster (2009) toodud tingimust ei saa kasutada andmete Benfordi seaduse vastavuse kontrolliks.

Valemi 3.2.3 kuju vaadates saame ka üldisema tingimuse juhusliku suuruse Benfordi seaduse vastavusele, sest $P(\log d \leq \log R < \log(d+1)) = \log(d+1) - \log d$, kui $\log R$ ehk X logaritmi murdosa on ühtlasest jaotusest lõigul $[0; 1]$ Seega positiivne juhuslik suurus X järgib Benfordi seadust parajasti siis, kui $\log R \sim U(0; 1)$ (Berger ja Hill, 2011). Kuid see tingimus ei anna aimu, mis andmestikud peaksid Benfordi seadust järgima ning see pole siiski tingimus X enda jaotuse kohta, mida lihtsalt mudelite põhjal kontrollida.

4. Benfordi seaduse kontrollimise testid

Benfordi seaduse testimiseks on mitmeid teste, antud peatükis on toodud mantissi ja keskmise hälbe test, mille kirjelduste aluseks on Nigrini (2012, ptk 7). Mantissi test on suure valimimahu korral väga tundlik ja talub ainult väikseid kõrvalekaldeid Benfordi seadusest. See eest keskmise hälbe test ei kasuta arvutustes valimi suurust ning tänu sellele ei teki probleeme suuri andmestikke analüüsides. Kuid sellel testil puuduvad objektiivsed väärtused, millega saadud statistikut otsuse tegemiseks võrrelda.

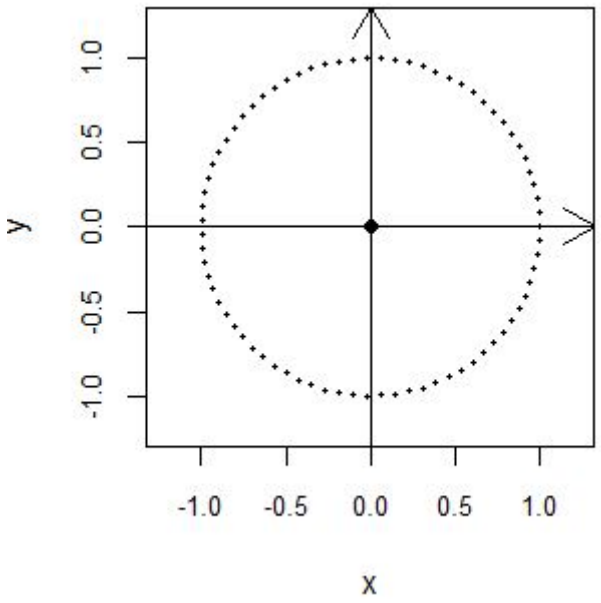
4.1. Mantissi test

Mantissi testis kujutatakse arvude mantissid ühikringil. Benfordi seaduse kehtimise kontrollimiseks leitakse raskuskeskme kaugus ühikringi keskpunktist.

Olgu N juhuslikku suurust, millele vastavad mantissid on $M_1, M_2, \ldots, M_N$. Mantissi testi jaoks kujutatakse need ühikringil, mille keskpunkt on $(0; 0)$. Koordinaadid arvutatakse järgnevalt: $x_i = \cos(2\pi M_i)$ ja $y_i = \sin(2\pi M_i)$, kus $i = 1, \ldots, N$. Benfordi seadust järgivate arvude mantissid asuvad ringjoonel ühtlaselt ja ringi raskuskese asub keskpunktis $(0; 0)$ (vt Joonis 4.1.1).

Punktis $(1; 0) = (\cos 0; \sin 0)$ asub mantiss 0, mis saadakse $\log 1$ tulemusel. Seega arvu tüvi $r = 1$ ning arvud on näiteks $0{,}1$, 1, 10, 100. Punktis $(0; 1)$ on mantiss $0{,}25$, millele vastavad arvud $0{,}17783$, $1{,}7783$, $17{,}783$ jne. Arvude $0{,}31623$, $3{,}1623$ ja $31{,}623$ mantiss on $0{,}5$, mis asub punktis $(-1; 0)$. Punktis $(0; -1)$ on mantiss $0{,}75$, mis saadakse arvudest $0{,}56234$, $5{,}6234$, $56{,}234$ jne.

Ühikringi ehk raadiusega 1 ringi ümbermõõt on $2\pi = 6{,}2832$. Ühendades numbriga 1 algavatele arvudele vastavad punktid, tekib kaar pikkusega $1{,}8914$ (vt Joonis 4.1.2). Ringi ümbermõõdu ja Benfordi seadust järgiva arvu numbriga 1 algamise tõenäosuse $0{,}30103$ korrutis on samuti $1{,}8914$. Korrutades ringi ümbermõõdu ja numbriga 5 algamise tõenäosuse $0{,}07918$ on tulemuseks $0{,}4975$, mis võrdub numbriga 5 hakkavatele



Joonis 4.1.1: Benfordi seadust järgivate arvude mantissid jaotuvad ühikringjoonel ühtlaselt ja ringi raskuskese asub keskpunktis.

arvudele vastavate punktide ühendamisel tekkinud kaare pikkusega.

Raskuskeskme kauguse ruut keskpunktist on $Z^2 = X^2 + Y^2$, kus

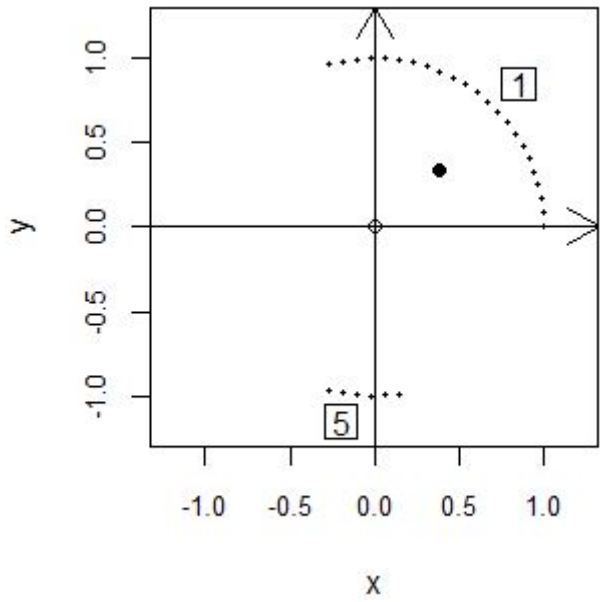
$$X = \frac{\sum_{i=1}^{N} \cos(2\pi M_i)}{N} \text{ ja } Y = \frac{\sum_{i=1}^{N} \sin(2\pi M_i)}{N}. \tag{5.1.1}$$

Benfordi seadust täpselt järgivate arvude korral asub ringi raskuskese keskpunktis $(0; 0)$. Mida rohkem kalduvad arvud Benfordi seadusest ehk asetsevad ringjoonel vähem ühtlasemalt, seda kaugemal asub raskuskese keskpunktist. Näiteks vaadates ainult numbriga 1 ja 5 algavaid arve, ei järgi need Benfordi seadust ning ka raskuskese ei asu keskpunktis (vt Joonis 4.1.2).

Juhul kui uuritavad arvud järgivad Benfordi seadust, siis nende mantissid on ühtlasest jaotusest ja kehtib järgmine:

$$E(\cos(2\pi M_i)) = E(\sin(2\pi M_i)) = 0 \text{ ning } D(\cos(2\pi M_i)) = D(\sin(2\pi M_i)) = 0{,}5.$$

Siis suuruste 5.1.1 korral kehtib $EX = EY = 0$ ja $DX = DY = 1/(2N)$. Tsent-



Joonis 4.1.2: Numbriga 1 ja 5 algavate arvude asetsemine ühikringil ja nende raskuskese.

raalse piirteoreemi järgi $X \dot{\sim} N(0; 1/(2N))$ ja $Y \dot{\sim} N(0; 1/(2N))$. Kuna kahe sõltumatu $N(0; \sigma^2)$ jaotusega juhusliku suuruse ruutude summa ruutjuur on Rayleigh jaotusega parameetriga σ (Forbes jt, 2011, lk 173), siis $Z = \sqrt{X^2 + Y^2}$ ehk raskuskeskme kaugus ühikringi keskpunktist, mis on mantissi testi teststatistik, on Rayleigh jaotusega, mille parameetriks on $\sqrt{1/(2N)}$. Rayleigh jaotuse jaotusfunktsioon on

$$F_Z(z) = 1 - e^{-z^2/(2(1/\sqrt{2N})^2)} = 1 - e^{-z^2 N}$$

(Forbes jt, 2011, lk 173). Selle põhjal saame olulisuse tõenäosuseks

$$p = P(Z > z|H_0) = 1 - P(Z \leq z|H_0) = e^{-z^2 N}. \tag{4.1.2}$$

Mantissi test on suure valimimahu korral väga tundlik, selle vähendamiseks saab Nigrini (2012) kohaselt kasutada olulisuse tõenäosuse arvutamisel valemis 4.1.2 N asemel $\sqrt{N}$ või $\sqrt[3]{N}$.

4.2. Keskmise hälbe test

Keskmise hälbe testi jaoks leitakse statistik järgmiselt

$$\text{Keskmine hälve} = \frac{\sum_{d=1}^{K} |E_d - O_d|}{K},$$

kus E_d tähistab numbriga d algavate arvude oodatavat tõenäosust ja O_d tähistab numbriga d algavate vaadeldud arvude osakaalu. Esimese numbrikoha kontrollimisel $d = 1, 2, \ldots, 9$, seega kokku on vaatluse all esimesi numbreid $K = 9$.

Antud testi jaoks puuduvad objektiivsed otsustuspiirid. Tabelis 4.2.1 on toodud Nigrini (2012) soovituslikud keskmise hälbe otsustuspiirid, mille määramise aluseks olid erinevad reaalsete andmestike tulemused.

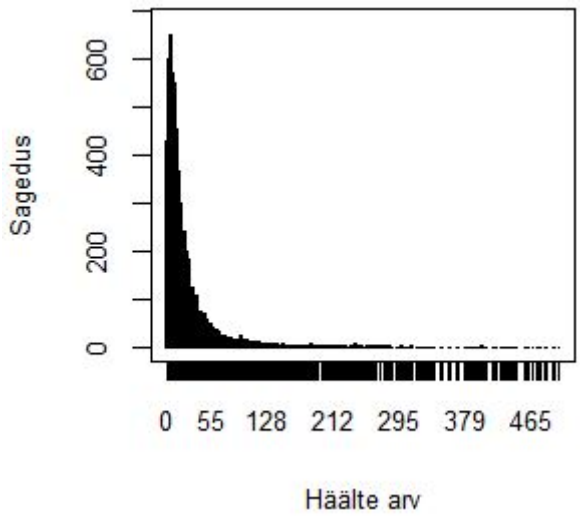
Tabel 4.2.1: Soovituslikud keskmise hälbe otsustuspiirid esimese numbrikoha korral.

Keskmine hälve	Otsus
$< 0{,}006$	Lähedane vastavus
$0{,}006 - 0{,}012$	Aktsepteeritav vastavus
$0{,}012 - 0{,}015$	Piiripealselt aktsepteeritav vastavus
$> 0{,}015$	Mittevastavus

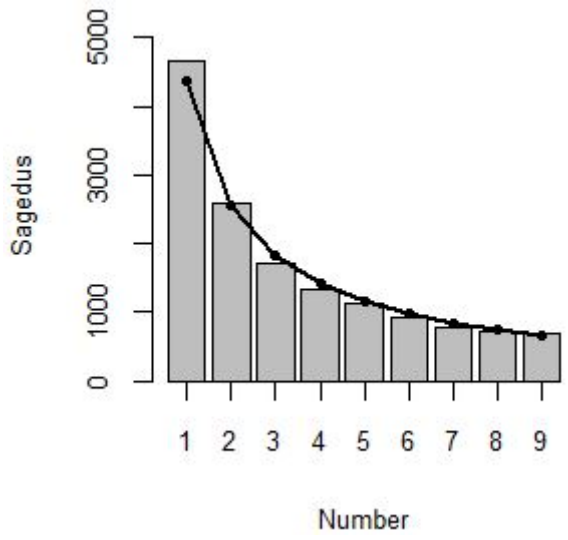
5. Kohaliku omavalitsuse volikogu valimiste analüüs

Kuna on arvatud, et valimistulemused peaksid järgima Benfordi seadust (Pericchi ja Torres, 2012), siis vaatame selles peatükis, kas 2013. aasta kohaliku omavalitsuse volikogu valimiste tulemused järgivad Benfordi seadust.

Valimistel oli 14 784 kandidaati, uurimiseks kasutatakse iga kandidaadi poolt häältetanute arvu. Kokku on andmestikus 625 334 kehtivat häält. Tulemuste jaotus on väga ebasümmeetriline, kuni 500 häält saadud tulemuste sagedused on toodud joonisel 5.1, sellest suuremaid tulemusi on 120. Kandidaatidest kõige rohkem hääli kogunud kandideerija sai 39 979 häält, suuruselt järgmine tulemus on 8 017. Väikseim saadud häälte arv on 0, mida esines 243 korda. Kuna arvul 0 pole esimest numbrit $1, \ldots, 9$ seas, siis Benfordi seaduse testimisel suurust 0 ei arvestata. Häälte arvu keskmine on 42,3 ja mediaan 14, seega väiksemaid tulemusi esineb rohkem. Valimiste tulemuste esimeste numbrite sagedused ja Benfordi seaduse jaoks oodatavad sagedused on kujutatud joonisel 5.2, millelt on silma järgi näha andmete väga head kooskõla antud seadusega.



Joonis 5.1: Kohaliku omavalitsuse volikogu valimistel kuni 500 häält saadud tulemuste sagedused.



Joonis 5.2: Valimiste tulemuste esimeste numbrite sagedused tulpdiagrammina ja Benfordi seaduse jaoks oodatavad sagedused joondiagrammina.

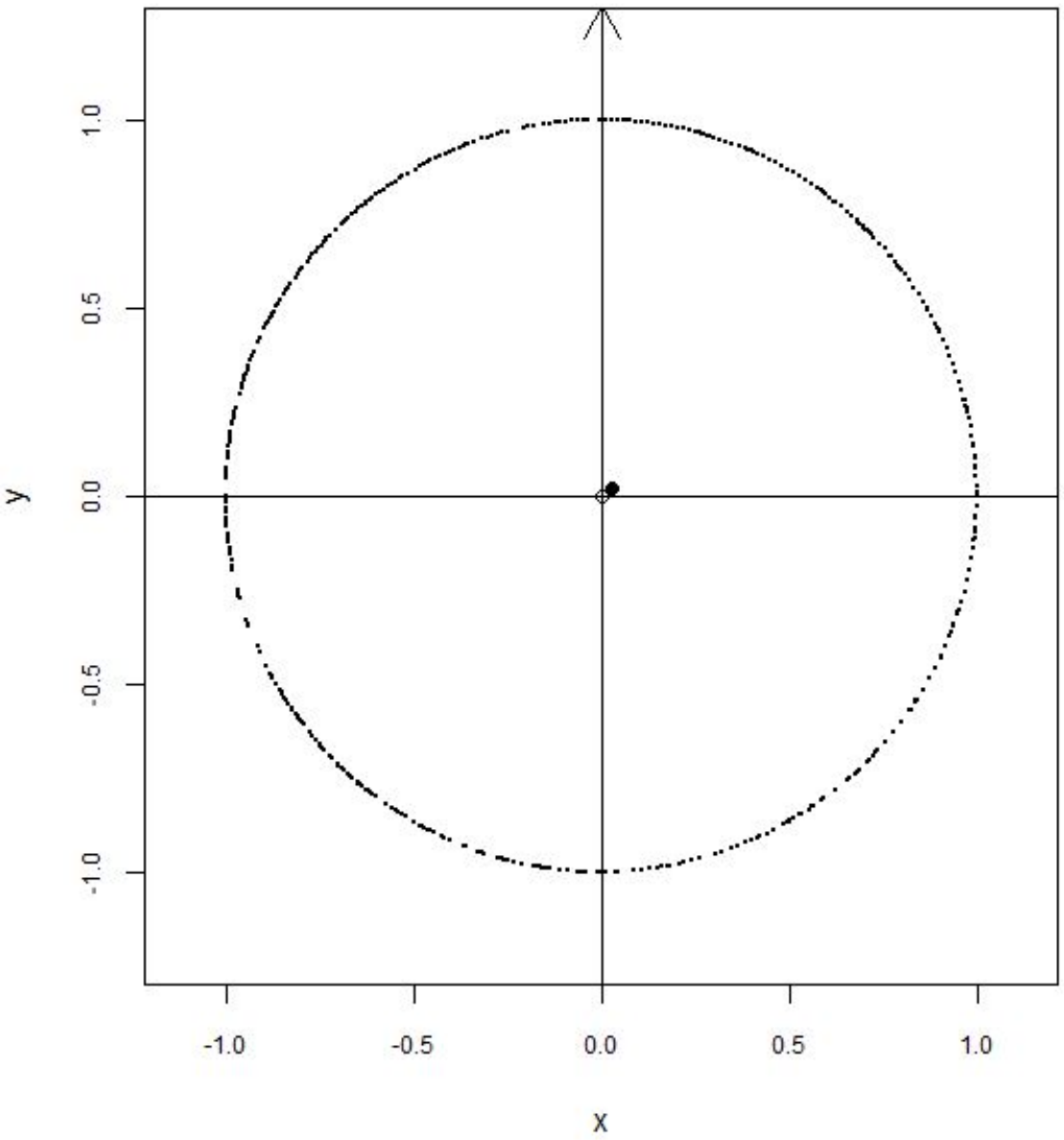
Kui kasutame Benfordi jaotuse sobivuse kontrolliks χ^2-testi, on testi vabadusastmete arvuks 8, statistiku väärtuseks saame $38{,}6$ ja olulisuse tõenäosuseks $0{,}000006$. See on väiksem olulisuse nivoost $0{,}05$, järelikult võtame vastu sisuka hüpoteesi, et andmed ei ole antud jaotusest. Seega valimiste tulemused ei järgi antud testi kohaselt Benfordi seadust. Kuid selle põhjuseks võib olla suure valimimahu korral χ^2-testi tundlikkus väikeste kõrvalekallete suhtes (Nigrini, 2012). Võttes 100 valimit andmetest, peaks nendest nullhüpoteesi kehtides olulisuse nivool $0{,}05$ jääma nullhüpoteesi juurde 95 tükki. Valimimahu 200 korral tulevad 100-st testist 86 kuni 100 selliseid, mis jäid nullhüpoteesi juurde. Seega kogu andmestik χ^2-testi järgi ei vasta Benfordi seadusele, kuid väiksema valimimahu korral jääb test antud andmete korral pigem nullhüpoteesi juurde.

Kontrollime nüüd kohalike valimiste tulemusi peatükis 4.1 tutvustatud mantissi testiga. Kandidaatide saadud häälte mantisside paiknemine ühikringil on toodud joonisel 5.3. Mantissi testi tulemuseks saame, et raskuskeskme kaugus keskpunktist on $0{,}0011$ ning testi olulisuse tõenäosus $0{,}0000002$, seega võtame vastu sisuka hüpoteesi, et and-

med ei järgi Benfordi seadust. Kuid ka mantissi test on suure valimimahu korral väga tundlik ja talub ainult väikseid kõrvalekaldeid Benfordi seadusest (Nigrini, 2012). Tundlikkuse vähendamiseks saab kasutada olulisuse tõenäosuse arvutamisel valemis 4.1.2 N asemel Nigrini (2012) soovitatud $\sqrt{N}$ või $\sqrt[3]{N}$. Valimiste tulemuste testimisel $\sqrt{N}$ kasutades, saame olulisuse tõenäosuseks $0{,}87$ ja $\sqrt[3]{N}$ kasutamisel $0{,}97$. Mõlemad olulisuse tõenäosused on suuremad olulisuse nivoost $0{,}05$, seega vähendatud tundlikkusega mantissi testi kohaselt kohaliku omavalitsuse volikogu valimiste tulemused järgivad Benfordi seadust.

Nii χ^2-test kui ka mantissi test võtavad arvesse valimi suuruse. Suure valimimahu korral peetakse parimaks lahenduseks kasutada peatükis 4.2 toodud keskmise hälbe testi, mis ei võta valimimahtu arvesse ning seega ei teki suure andmehulga tõttu probleeme (Nigrini, 2012). Kohalike valimiste tulemuste esimeste numbrikohtade keskmiseks hälbeks saame $0{,}005$. Tabelis 4.2.1 toodud otsustuspiiride järgi on tegemist lähedase vastavusega, seega tulemused vastavad Benfordi seadusele.

Seega saime, et suure valimimahu korral tundlike χ^2-testi ja mantissi testi korral valimiste tulemused ei järgi Benfordi seadust. Väiksemaid valimeid võttes jääb χ^2-test pigem nullhüpoteesi, et arvude esinumbrid on Benfordi jaotusest, juurde. Vähendatud tundlikkusega mantissi testi kohaselt andmed vastavad antud seadusele. Valimi suurust mitte arvestavat keskmise hälbe testi kasutades on 2013. aasta kohalike omavalitsuse volikogu valimiste häälte arvude esimesed numbrid lähedase vastavusega Benfordi seadusele.



Joonis 5.3: Kohaliku omavalituse volikogu valimiste tulemuste mantissid ühikringil.

Viited

Alexander, J.C., 2009. Remarks on the use of Benford's law. *Social Science Research Network*, [online] Kättesaadav: <http://ssrn.com/abstract=1505147> [Vaadatud 14. märts 2014].

Benford, F., 1938. The Law of Anomalous Numbers. *Proceedings of the American Philosophical Society*, [online] 78(4), lk 551-572. Kättesaadav: Tartu Ülikooli Raamatukogu URL <www.jstor.org/stable/984802> [Vaadatud 7. märts 2014].

Berger, A. ja Hill, T.P., 2011. Benford's Law Strikes Back: No Simple Explanation in Sight for Mathematical Gem. *The Mathematical Intelligencer*, [online] Kättesaadav: <http://www.math.ualberta.ca/~aberger/benford_bibliography/berger_hill_11a.pdf> [Vaadatud 2. aprill 2014].

Cinelli, C., 2014. benford.analysis: Benford Analysis for data validation and forensic analytics. R package version 0.1. <http://CRAN.R-project.org/package=benford.analysis>.

Fewster, R., 2009. A Simple Explanation of Benford's Law. *The American Statistician*, [online] Kättesaadav: <https://www.stat.auckland.ac.nz/~fewster/RFewster_Benford.pdf> [Vaadatud 7. märts 2014].

Forbes, C., Evans, M., Hastings, N. ja Peacock, B., 2011. *Statistical Distributions*. Hoboken, New Jersey: John Wiley & Sons.

Hill, T.P., 1995. A Statistical Derivation of the Significant-Digit Law. *Statistical Science*, [online] Kättesaadav: <http://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1042&context=rgp_rsr> [Vaadatud 7. märts 2014].

Hill, T.P. ja Rogers, E., 2013. Benford Online Bibliography. [online] Kättesaadav: <http://www.benfordonline.net/> [Vaadatud 2. mai 2014].

Joenssen, D.W., 2013. BenfordTests: Statistical Tests for Evaluating Conformity to Benford's Law. R package version 1.1.1. <http://CRAN.R-project.org/package=BenfordTests>.

Matthews, R., 1999. The power of one. *New Scientist*, [online] Kättesaadav: <http://people.math.gatech.edu/~hill/ARTICLES/The%20Power%20of%20One.pdf> [Vaadatud 2. aprill 2014].

Newcomb, S., 1881. Note on the Frequency of Use of the Different Digits in Natural Numbers. *American Journal of Mathematics*, [online] Kättesaadav: <http://www.jstor.org/stable/2369148> [Vaadatud 7. märts 2014].

Nigrini, M.J., 2012. *Applications for Forensic Accounting, Auditing, and Fraud Detection*. Hoboken, New Jersey: John Wiley & Sons.

Pericchi, L. ja Torres, D., 2012. Quick Anomaly Detection by the Newcomb?Benford Law, with Applications to Electoral Processes Data from the USA, Puerto Rico and Venezuela. *Statistical Science*, [online] Kättesaadav: <http://arxiv.org/pdf/1205.3290.pdf> [Vaadatud 2. mai 2014].

Shao, L. ja Ma, B.Q., 2010. The significant digit law in statistical physics. *Physica A*, [online] 389164), lk 3109-3116. Kättesaadav: Tartu Ülikooli Raamatukogu URL <http://www.sciencedirect.com/science/article/pii/S0378437110003596> [Vaadatud 25. aprill 2014].

Lisad

Lisa 1. Andmete sisselugemise programmikood

```
# Leiame Benfordi seaduse kohta avaldatud artiklite arvu aastate kaupa

# Failis on lehelt http://www.benfordonline.net/list/chronological kopeeritud aastaarvud
f = open("benfordonline_aastad.txt", encoding="UTF-8")
sisu = f.read()
f.close()

# Sõnastik, mille võtmeteks on aastaarvud 1881-2014
artikleid = {}
for aasta in range(1881, 2015):
    artikleid[aasta] = 0

# Täidame sõnsatiku, väärtuseks on sel aastal ilmunud artilite arv
a = sisu.split()
for rida in a:
    artikleid[int(rida)] += 1
print(artikleid)

# Kirjutame aastad ja ilmunud artiklite arvud faili
f = open("artikleid.csv", mode="w")
# Tunnuste nimed
f.write("aasta, artikleid\n")

# Leiame, mis aastal ilmus kõige rohkem artikleid
max = 0
max_aastad = []

for aasta in artikleid:
  # Kirjutame aasta ja sellel aastal ilmunud artiklite arvu faili
    f.write(str(aasta) + ", "+ str(artikleid[aasta]) + "\n")
  # Kontrollime, kas see on hetke suurim aastas artikleid ilmunud arv
    if artikleid[aasta] > max:
        max = artikleid[aasta]
        max_aastad = [aasta]
    elif artikleid[aasta] == max:
        max_aastad.append(aasta)

print(max, max_aastad)


```

```
42 # Leiame 2013. aasta kohaliku omavalitsuse volikogu valimiste
43 # häälte arvud kandidaatide kaupa
44
45 from os import listdir
46 from os.path import join
47
48 def failist_lugemine(failinimi):
49     f = open(join(kaust,failinimi), encoding="UTF-8")
50     sisu = f.read()
51     f.close()
52     kandidaadid = sisu.split("<candidate>")
53     hääled = []
54     for i in range(1,len(kandidaadid)):
55         info = kandidaadid[i]
56         hääli_algus = info.index("<value>") + len("<value>")
57         hääli_lõpp = info.index("</value>")
58         hääli = info[hääli_algus : hääli_lõpp]
59         hääled.append(int(hääli))
60     #print(hääled)
61
62     # Kontrollimiseks:
63     # Kuna Tallinnas on ka info valimisringkondade kohta,
64     # siis faili struktuur on teistest erinev
65     if failinimi == "KOV_2013_ELECTION_RESULT_0784.xml":
66         ringkond_hääled = 218505
67         # häälte arv vaadatud http://kov2013.vvk.ee/detailed_0784.
               html
68     else:
69         ringkond = sisu.split("</voteDistributionByParties>")
70         ringkond_hääli_algus = ringkond[1].index("<value>") + len("<
               value>")
71         ringkond_hääli_lõpp = ringkond[1].index("</value>")
72         ringkond_hääled = ringkond[1][ringkond_hääli_algus:
               ringkond_hääli_lõpp]
73     if sum(hääled) != int(ringkond_hääled): # Kontroll
74         print("VIGA: õiget häälte tulemust ei saanud failis " + \
75               failinimi + ", hääli " + str(sum(hääled)) + \
76               ", ringkonnas hääli " + ringkond_hääled)
77     return hääled
78
79 kaust = "kov2013_results"
80 failid = listdir(kaust) # kaustas asuvad failid
81
82 # Failidest häälte arvude lugemine:
83 kõik_hääled = []
84 for failinimi in failid:
85     # "KOV_2013_ELECTION_RESULT" algav fail käib
86     # valimistulemuste kohta valdades ja linnades
87     if "KOV_2013_ELECTION_RESULT" in failinimi:
```

```
88          hääled = failist_lugemine(failinimi)
89          kõik_hääled += hääled
90
91  print(len(kõik_hääled)) # kandidaatide arv: 14784
92  print(sum(kõik_hääled)) # häälte arv: 625334
93  print(min(kõik_hääled)) # väikseim häälte arv: 0
94  print(max(kõik_hääled)) # suurim häälte arv: 39979
95
96  # Häälte arvude uude faili kirjutamine:
97  f = open("tulemused.csv", mode="w")
98  for hääl in kõik_hääled:
99      f.write(str(hääl) + "\n")
100 f.close()
```

Lisa 2. Andmete analüüsi ja testide jooniste programmikood

```
install.packages("BenfordTests")
library("BenfordTests")
citation("BenfordTests")
install.packages("benford.analysis")
library("benford.analysis")
citation("benford.analysis")

########## Peatükk 2 ##########
########## Näited andmetest ##########

########## Benfordi uuritud andmestike parim ja halvim kooskõla, joonis 2.1 ##########

e_tn = c(0.301, 0.176, 0.125, 0.097, 0.079, 0.067, 0.058, 0.051, 0.046) # eeldatavad tõenäosused
par(mfrow = c(1, 2))

# Arvud ajalehe esikülgedel
o_tn = c(0.300, 0.180, 0.120, 0.100, 0.080, 0.060, 0.060, 0.050, 0.050) # tegelikud tõenäosused
xmarks = barplot(o_tn, xlab = "Number", ylab = "Tõenäosus", ylim=c(0, 0.5), cex.lab=.8, cex.axis = .8)
axis(1, at=xmarks, labels = seq(1,9), cex.lab=.8, cex.axis = .8)
lines(xmarks, e_tn, lwd=2, col="black") #lty=2
points(x=xmarks, y = e_tn, pch=20)

# Aatommassid
o_tn = c(0.472, 0.187, 0.055, 0.044, 0.066, 0.044, 0.033, 0.044, 0.055) # tegelikud tõenäosused
xmarks = barplot(o_tn, xlab = "Number", ylab = "Tõenäosus", ylim=c(0, max(o_tn, e_tn) * 1.1), cex.lab=.8, cex.axis = .8)
axis(1, at=xmarks, labels = seq(1,9), cex.lab=.8, cex.axis = .8)
lines(xmarks, e_tn, lwd=2, col="black") #lty=2
points(x=xmarks, y = e_tn, pch=20)

########## Fibonacci arvud, joonis 2.2 ##########
fibonacci = function(n){
  fibvals = numeric(n)
  fibvals[1] = 1
  fibvals[2] = 1
  for (i in 3:n)
    fibvals[i] = fibvals[i-1] + fibvals[i-2]
  return (fibvals)
}

tulemused = fibonacci(400)

```

```
(bfd = benford(tulemused, number.of.digits = 1))
(e_tn = bfd$bfd$benford.dist) # eeldatavad tõenäosused
(o_tn = bfd$bfd$data.dist) # tegelikud tõenäosused
(round(e_tn - o_tn, 2))

par(mfrow = c(1, 1))
xmarks = barplot(o_tn, xlab = "Number", ylab = "Tõenäosus", ylim=c(0,
    max(o_tn, e_tn) * 1.1), cex.lab=.8, cex.axis = .8)
axis(1, at=xmarks, labels = seq(1,9), cex.lab=.8, cex.axis = .8)
lines(xmarks, e_tn, lwd=2, col="black") #lty=2
points(x=xmarks, y = e_tn, pch=20)


########## Benfordi seaduse kohta aastas ilmunud artiklid ##########
andmed = read.csv("C:\\Users\\G\\Dropbox\\6.semester\\Lõputöö\\
    artikleid.csv", header=TRUE, sep=",")
dim(andmed)
names(andmed)
tulemused = sort(andmed$artikleid*1.0)
length(tulemused) # 134 (1881-2014)
sum(tulemused) # 799
max(tulemused) # 64
min(tulemused) # 0
mean(tulemused) # 6
median(tulemused) # 1
table(tulemused)

########## Aastas ilmunud artiklite sagedused, joonis 2.3 ##########
#plot(table(tulemused), xlab="Aastas ilmunud artiklite arv", ylab = "
    Sagedus", cex.lab=.8, cex.axis = .8)
plot(table(tulemused[tulemused > 0]), xlab="Aastas ilmunud artiklite
    arv", ylim=c(0, 20), ylab = "Sagedus", cex.lab=.8, cex.axis = .8)

########## Esimeste numbrikohtade sagedused, joonis 2.4 ##########
# Kasutame teegi benford.anaysis funktsiooni benford
(bfd = benford(tulemused, number.of.digits = 1))

(e = bfd$bfd$benford.dist.freq) # eeldatavad sagedused
(o = bfd$bfd$data.dist.freq) # tegelikud sagedused

xmarks = barplot(o, xlab = "Number", ylab = "Sagedus", ylim=c(0, max(
    o, e) * 1.1), cex.lab=.8, cex.axis = .8)
axis(1, at=xmarks, labels = seq(1,9), cex.lab=.8, cex.axis = .8)
lines(xmarks, e, lwd=2, col="black") #lty=2
points(x=xmarks, y = e, pch=20)

# Erinevused
round(e-o,1)

```

```
########## Peatükk 4 ##########
########## Bennfordi seaduse kontrollimise testid ##########

# Funktsioon, mis teeb täpselt Benfordi seadust järgivad arvud
# (Nigrini, 2012, lk 163-164)
exact_benford = function(n){
  i = seq(1, n)
  exact_vals = 10*(10**(4/n))**(i-1)
  return (exact_vals)
}

########## Benfordi seadust järgivad arvud ühikringil, joonis 4.1.1
    ##########

# Arvud:
vals = exact_benford(150)
benford(vals)
# Koordinaadid:
x = cos(2*pi*(log10(vals)%%1))
y = sin(2*pi*(log10(vals)%%1))
# Raskuskese koordinaadid:
(x_raskuskese = sum(x) / length(x))
(y_raskuskese = sum(y) / length(y))

# Kujutame joonisel
plot(0, 0, xlim = c(-1, 1), ylim = c(-1.2, 1.2), asp = 1, xlab = "x",
     ylab = "y", cex.lab=1, cex.axis = .8)
arrows(-1.33, 0, 1.33, 0) # x-telg
arrows(0, -1.3, 0, 1.3) # y-telg
points(x, y, pch=20, cex=.6) # arvud
points(x_raskuskese, y_raskuskese, pch=16) # raskuskese

########## Numbriga 1 ja 5 algavate arvude mantissid ühikringil,
    joonis 4.1.2 ##########

# Teegi BenfordTests funktsioon signifd leiab vaikimisi arvu esimese
    numbri

# Numbriga 1 algavad arvud ja neile vastavad koordinaadid:
yhega = vals[signifd(vals) == 1]
x_yhega = cos(2*pi*(log10(yhega)%%1))
y_yhega = sin(2*pi*(log10(yhega)%%1))

# Numbriga 5 algavad arvud ja neile vastavad koordinaadid:
viiega = vals[signifd(vals) == 5]
x_viiega = cos(2*pi*(log10(viiega)%%1))
y_viiega = sin(2*pi*(log10(viiega)%%1))

# Raskuskese koordinaadid:
```

```
(x_raskuskese = (sum(x_yhega) + sum(x_viiega)) / (length(x_yhega) +
    length(x_viiega)))
(y_raskuskese = (sum(y_yhega) + sum(y_viiega)) / (length(y_yhega) +
    length(y_viiega)))

# Kujutame joonisel ainult arvud, mis algavad numbriga 1 ja 5
plot(0, 0, xlim = c(-1, 1), ylim = c(-1.2, 1.2), asp = 1, xlab = "x",
     ylab = "y", cex.lab=1, cex.axis = .8)
arrows(-1.33, 0, 1.33, 0) # x-telg
arrows(0, -1.3, 0, 1.3) # y-telg
points(x_yhega, y_yhega, pch=20, cex=.6)
points(x_viiega, y_viiega, pch=20, cex=.6)
points(x_raskuskese, y_raskuskese, pch=16) # raskuskese

# Number 1 ruudu sees:
points(0.85, 0.85, pch=0, cex=2.3)
text(0.87, 0.85, "1")
# Number 5 ruudu sees:
points(-0.2, -1.15, pch=0, cex=2.3)
text(-0.19, -1.15, "5")

########## Peatükk 5 ##########
########## Kohalike omavalitsuste valimiste tulemuste analüüs
    ##########

########## Kohaliku omavalitsuse valimiste tulemused ##########
andmed = read.csv("C:\\Users\\G\\Dropbox\\6.semester\\Lõputöö\\
    tulemused.csv", header=FALSE)
dim(andmed)
names(andmed)
tulemused = (andmed$V1)*1.0
length(tulemused) # 14784
sum(tulemused) # 625334
max(tulemused) # 39979
min(tulemused) # 0
mean(tulemused) # 42.3
median(tulemused) # 14
table(tulemused)

########## Kuni 500 häält saadud tulemuste sagedused, joonis 5.1
    ##########
#plot(table(tulemused))
#plot(table(tulemused[tulemused != 39979]))
plot(table(tulemused[tulemused < 500]), xlab="Häälte arv", ylab = "
    Sagedus", ylim=c(0,675), cex.lab=.8, cex.axis = .8)
length(tulemused[tulemused > 500])

########## Esimeste numbrikohtade sagedused, joonis 5.2 ##########
# Kasutame teegi benford.anaysis funktsiooni benford
```

```
(bfd = benford(tulemused, number.of.digits = 1))

(e = bfd$bfd$benford.dist.freq) # eeldatavad sagedused
(o = bfd$bfd$data.dist.freq) # tegelikud sagedused

xmarks = barplot(o, xlab = "Number", ylab = "Sagedus", ylim=c(0, max(
    o, e) * 1.1), cex.lab=.8, cex.axis = .8)
axis(1, at=xmarks, labels = seq(1,9), cex.lab=.8, cex.axis = .8)
lines(xmarks, e, lwd=2, col="black") #lty=2
points(x=xmarks, y = e, pch=20)

########## Hii-ruut test ##########
# Hii-ruut statistik
sum(bfd$bfd$squared.diff)
sum((e-o)**2 / e)
# Olulisuse tõenäosus
(bfd$stats$chisq$p.value)

# Funktsioon, mis teeb andmetest "arv" valimit ja
# tagastab nendest nullhüpoteesi juurde jäänute arvu
nullhüpoteese = function(andmed, suurus, arv) {
  # Leiame p-väärtused
  p = numeric(arv)
  for (i in 1:arv) {
    valim = sample(andmed, size = suurus)
    bfd = benford(valim, number.of.digits=1)
    p[i] = bfd$stats$chisq$p.value
  }
  # Leiame nullhüpoteesi juurde jäänute arvu
  H0 = p[p > 0.05]
  #plot(sort(p)) # Ühtlane jaotus
  H0_arv = length(H0)

  return (H0_arv)
}

# Teeme 50 korda 100 valimit suurusega 200
H0 = numeric(50)
for (i in 1:50)
  H0[i] = nullhüpoteese(tulemused, 200, 100)
sort(H0) # 50-st 50 korral peaagu kõik 100 valimit võtsid vastu
    nullhüpoteesi

valim = sample(tulemused, size=200)
(bfd = benford(valim, number.of.digits = 1))

(e = bfd$bfd$benford.dist.freq) # eeldatavad sagedused
(o = bfd$bfd$data.dist.freq) # tegelikud sagedused
e_tn = bfd$bfd$benford.dist # eeldatavad tõenäosused
```

```
o_tn = bfd$bfd$data.dist # tegelikud tõenäosused
chisq.test(o, e_tn)
p=log10(2:10/1:9)
chisq.test(o, p=p)

########## Mantissi test ##########
L2 = 0.0011 #(L2 = bfd$stats$mantissa.arc.test$L2)
bfd$stats$mantissa.arc.test$p.value # p=0.0000001812

# Tundlikkuse vähendamiseks kasutame sqrt(N) ja N**(1/3)
exp(-L2*sqrt(length(tulemused))) # p=0.87
# Tundlikkuse vähendamiseks kasutame sqrt(N)
exp(-L2*(length(tulemused))**(1/3)) # p=0.97

########## Tulemuste mantissid ühikringil, joonis 5.3 ##########

# Jätame välja 0-d, sest sellel ei ole esimest numbrit (1-9) ja
# mantissi leidmisel ei saa leida logaritmi 0-st
vals = tulemused[tulemused != 0]

# Koordinaadid:
x = cos(2*pi*(log10(vals)%%1))
y = sin(2*pi*(log10(vals)%%1))
# Raskuskese koordinaadid:
(x_raskuskese = sum(x) / length(x))
(y_raskuskese = sum(y) / length(y))

# Kujutame joonisel
plot(0, 0, xlim = c(-1, 1), ylim = c(-1.2, 1.2), asp = 1, xlab = "x",
     ylab = "y", cex.lab=1, cex.axis = .75)
arrows(-1.33, 0, 1.33, 0) # x-telg
arrows(0, -1.3, 0, 1.3) # y-telg
points(x, y, pch=20, cex=.2) # arvud
points(x_raskuskese, y_raskuskese, pch=16) # raskuskese

########## Keskmise hälbe test ##########
(bfd$MAD) # 0.005193287
e_tn = bfd$bfd$benford.dist # eeldatavad tõenäosused
o_tn = bfd$bfd$data.dist # tegelikud tõenäosused
sum(abs(o_tn - e_tn))/length(o_tn)
```

Lisa 3. Benfordi analüüsitud andmestike tulemused

Tabel 1: Benfordi uuritud andmestike esimeste numbrikohtade osakaalud

Andmed	Arv	1	2	3	4	5	6	7	8	9
Jõgede pindalad	335	31,0	16,4	10,7	11,3	7,2	8,6	5,5	4,2	5,1
Rahvaarvud USAs	3259	33,9	20,4	14,2	8,1	7,2	6,2	4,1	3,7	2,2
Füüsika konstandid	104	41,3	14,4	4,8	8,6	10,6	5,8	1,0	2,9	10,6
Arvud ajalehe esikülgedel	100	30,0	18,0	12,0	10,0	8,0	6,0	6,0	5,0	5,0
Erisoojused	1389	24,0	18,4	16,2	14,6	10,6	4,1	3,2	4,8	4,1
Õhuvoolu rõhu kadumised	703	29,6	18,3	12,8	9,8	8,3	6,4	5,7	4,4	4,7
Õhuvoolu võimsuse kadumised	690	30,0	18,4	11,9	10,8	8,1	7,0	5,1	5,1	3,6
Molekulmassid.	1800	26,7	25,2	15,4	10,8	6,7	5,1	4,1	2,8	3,2
Jõgede äravoolud	159	27,1	23,9	13,8	12,6	8,2	5,0	5,0	2,5	1,9
Aatommassid	91	47,2	18,7	5,5	4,4	6,6	4,4	3,3	4,4	5,5
$n^{-1}, \sqrt{n}, \ldots$	5000	25,7	20,3	9,7	6,8	6,6	6,8	7,2	8,0	8,9
Design Data Generators	560	26,8	14,8	14,3	7,5	8,3	8,4	7,0	7,3	5,6
Arvud ajakirjas	308	33,4	18,5	12,4	7,5	7,1	6,5	5,5	4,9	4,2
Betooni hinnad	741	32,4	18,8	10,1	10,1	9,8	5,5	4,7	5,5	3,1
Röntgenkiirguse pinged	707	27,9	17,5	14,4	9,0	8,1	7,4	5,1	5,8	4,8
Pesapalli statistika	1458	32,7	17,6	12,6	9,8	7,4	6,4	4,9	5,6	3,0
Musta keha kiirgused	1165	31,0	17,3	14,1	8,7	6,6	7,0	5,2	4,7	5,4
Majanumber aadressis	342	28,9	19,2	12,6	8,8	8,5	6,4	5,6	5,0	5,0
$n^1, \ldots, n^8, n!$	900	25,3	16,0	12,0	10,0	8,5	8,8	6,8	7,1	5,5
Suremuskordajad	418	27,0	18,6	15,7	9,4	6,7	6,5	7,2	4,8	4,1
Keskmine	1101	30,6	18,5	12,3	9,4	8,0	6,4	5,1	4,9	4,8
Tõenäoline viga		±0,8	±0,4	±0,4	±0,3	±0,2	±0,2	±0,2	±0,2	±0,3
Benfordi seadus		30,1	17,6	12,5	9,7	7,9	6,7	5,8	5,1	4,6

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Gea Pajula,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose "Benfordi seadus", mille juhendaja on Anne Selart,

    1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
    1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace't'i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 05.05.2014