Benfordova statistična inkvizicija

Borja Slamič, Eva Šantić Zadravec, Rok Jurinčič
Mentor: Nina Štempelj

Povzetek

Benfordov zakon opisuje lastnost nekaterih družin podatkov, da se nižje vodilne števke podatkov iz teh družin pojavljajo pogosteje kot višje. Zakon smo izpeljali in predstavili tudi njegovo uporabo in primere pojavljanja v vsakdanjem življenju. Zakon smo preoblikovali, da lahko poleg verjetnosti pojavljanja posameznih vodilnih števk izračunamo tudi verjetnost pojavljanja vseh naslednjih števk. Izpeljali smo ga tudi za druge številske sisteme.

1 Zgodovina Benfordovega zakona

Leta 1881 je Simon Newcomb opazil, da so prve strani njegove knjižice logaritemskih tabel bolj obrabljene kot tiste na koncu. Sklepal je, da največkrat računa s števili, katerih vodilna števka je 1 in da s števili z višjimi vodilnimi števkami redkeje računa. Ker niso imeli kalkulatorjev so namreč produkte velikih števil računali tako, da so logaritme teh števil sešteli in v knjižicah s predhodno izračunanimi vrednostmi logaritmov poiskali število, katerega logaritem je dobljena vsota, saj velja

$$\log a + \log b = \log(a \cdot b).$$

Objavil je članek s svojo teorijo zakaj naj bi do tega prihajalo. Predvideval je, da so podatki s katerimi je delal porazdeljeni logaritemsko. Zato je predlagal, da se verjetnost, da je števka k vodilna izračuna po formuli

$$P(d = k) = \log(k + 1) - \log(k).$$

Leta 1938 je enako opazil Frank Benford, ki mu je uspelo tudi dokazati enačbo in jo preveriti na večih različnih družinah podatkov.

2 Izpeljava Benfordovega zakona

Naj bo I interval, na katerem se nahajajo podatki - verjetnost, da se podatek nahaja na njem je 1. V našem primeru bo to interval $(0, \infty)$. Množica J je množica, za katero računamo verjetnost, da se podatek pojavi na njej.

Če bi bili podatki enakomeno porazdeljeni po celotnem intervalu I, se verjetnost, da so podatki v množici J izračuna tako, da delež elementov na intervalu I, ki so hkrati elementi množice J, primerjamo s številom elementov na celotnem intervalu I. Zanimajo nas vodilne števke in opazimo lahko, da se vzorec ponavlja za vsak velikostni red, torej vsak interval $L = [10^k, 10^{k+1}]$, pri čemer je $k \in \mathbb{N}$. Zato verjetnost, da je podatek element množice J izračunamo tako, da dolžino intervala $M = J \cap L = [10^k, 2 \cdot 10^k)$ primerjamo z dolžino intervala L.

Verjetnost, da se podatek nahaja na podintervalu ni odvisna samo od dolžine intervala, ampak tudi od gostote podatkov na njem.

Če za števila velja Benfordov zakon, podatki ne morejo biti porazdeljeni enakomerno, sicer bi bila verjetnost, da se podatki nahajajo na množici J enaka $\frac{1}{9}$. Torej bi morali poiskati gostoto podatkov in jo upoštevati pri izračunu verjetnosti.

Ko je Benford podatke pretvarjal iz ene enote v drugo je ugotovil, da se vzorec razporeditve vodilnih števk ohranja, vendar pa se interval $[a, b]$, na katerem so števila ki jih opazujemo, razširi (če je koeficient $|k| > 1$) oziroma skrči (če je koeficient $|k| < 1$).

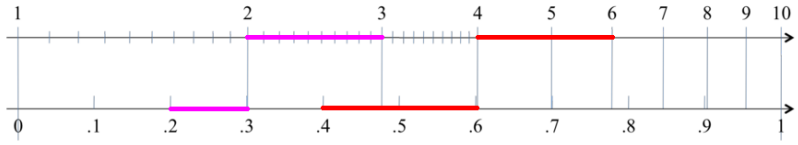
Na običajni skali lahko množenje podatkov s faktorjem k predstavimo tako, da krajišči pomnožimo s k in tako dobimo razširjen interval $[ka, kb]$ z enako količino a drugačno gostoto podatkov.Če želimo zagotoviti, da se bo gostota podatkov ohranjala, potrebujemo skalo, na kateri bo veljalo:

$$d(a, b) = d(ka, kb).$$

Opazimo da temu pogoju ustreza logaritemska skala, saj velja:

$$\log ka - \log kb = \log \frac{ka}{kb} = \log \frac{a}{b}$$

$$\log a - \log b = \log \frac{a}{b}.$$



Slika 1: Interval $[2, 3]$ (magenta) pomnožimo s $k = 2$ in dobimo interval $[4, 6]$ (rdeč). Intervala sta prikazana na navadni in logaritemski skali.

Torej lahko izračunamo verjetnost, da imajo podatki, ki se ravnajo po Benfordovem zakonu, vodilno števko l tako, da na logaritemski skali dolžino intervala $[l \cdot 10^k, (l + 1) \cdot 10^k]$ delimo z dolžino intervala $[10^k, 10^{k+1}]$. Iz tega, da imajo na logaritemski skali poljubni intervali L dolžino 1 sledi, da je verjetnost da ima podatek vodilno števko l enaka dolžini intervala $[l, l + 1]$. Torej lahko verjetnost računamo kot

$$P(d = l) = \log(l + 1) - \log(l).$$

Poračunane verjetnosti so prikazane v tabeli 1.

1	2	3	4	5	6	7	8	9
30,1%	17,6%	12,5%	9,7%	7,9%	6,7%	5,8%	5,1%	4,9%

Tabela 1: Verjetnost pojavljanja za vse možne vodilne števke.

3 Benfordov zakon za drugo števko

Podobno, kot se med prehodom med redi velikosti ohranja verjetnost pojavljanja vodilne števke, velja to tudi za drugo števko. Intervale L je treba v tem primeru dodatno razdeliti na intervale s števili z enakimi drugimi števkami.

Ko hočemo opazovati verjetnost pojavljanja druge števke d moramo torej sešteti dolžine intervalov, na katerih se d pojavlja kot druga števka. To se

v našem primeru na intervalu L ponovi devetkrat. Poglejmo si primer za verjetnost pojavljanja dvojke kot druge števke.

$$\begin{aligned} P(d_2 = 2) =& (\log 13 - \log 12) + (\log 23 - \log 22) + \\ & (\log 33 - \log 32) + (\log 43 - \log 42) + \\ & (\log 53 - \log 52) + (\log 63 - \log 62) + \\ & (\log 73 - \log 72) + (\log 83 - \log 82) + \\ & (\log 93 - \log 92) \approx 0,1088 \end{aligned}$$

Benfordov zakon se da posplošiti tudi za poljubno števko. V tabeli 2 je prikazana verjetnost pojavljanja števk na drugem mestu.

0	1	2	3	4	5	6	7	8	9
12,0%	11,4%	10,9%	10,4%	10,0%	9,7%	9,3%	9,0%	8,8%	8,3%

Tabela 2: verjetnost pojavljanja števk na drugem mestu

Že v primeru verjetnosti pojavljanja števk na drugem mestu so razlike med verjetnostmi pojavljanja posameznih števk opazno manjše kot za vodilno števko. Za naslednje števke je porazdelitev vedno bolj enakomerna.

4 Benfordov zakon za druge številske sisteme

Pri pretvarjanju med številskimi sistemi se Benfordov zakon ohranja. Da bi to pokazali, moramo logaritem spremeniti na osnovo številskega sistema.

Število a pretvorimo iz desetiškega sistema v sistem z novo osnovo b po naslednjem postopku:

$$a = b \cdot a_1 + o_1$$
$$a_1 = b \cdot a_2 + o_2$$
$$\vdots$$
$$a_{n-1} = b \cdot a_n + o_{n-1}$$
$$a_n = b \cdot 0 + o_n$$

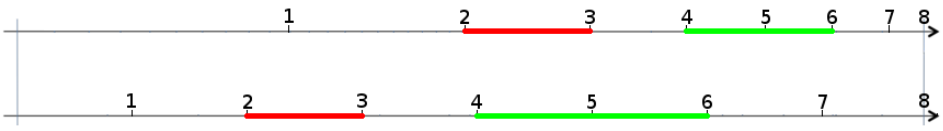
o_i = ostanek

$$o_n \cdot b^{n-1} + o_{n-1} \cdot b^{n-2} ... + o_2 \cdot b^1 + o_1,$$

pri čemer so $o_i \in \mathbb{N}$ in $o_i < b$ za vsak $i \in \mathbb{N}$.

Poiščemo ostanek o_1 pri deljenju števila a z osnovo b. Ta ostanek bo zadnja števka pretvorjenega števila. Postopek ponovimo za koeficient a_1 in dobimo naslednjo števko o_2. Postopek ponavljamo, dokler je koeficient $a_i > b$.

Številke, ki smo jih zapisali v številskem sistemu b, predstavimo na številski osi. Da bi se izognili raztegovanju in spreminjanju gostote, smo, kot pri desetiškem sistemu, pretvorili skalo številske osi v logaritemsko. Tokrat mora imeti logaritem osnovo b, tako da je enaka osnovi številskega sistema, sicer dolžina intervala do b ne bi bila enaka 1.



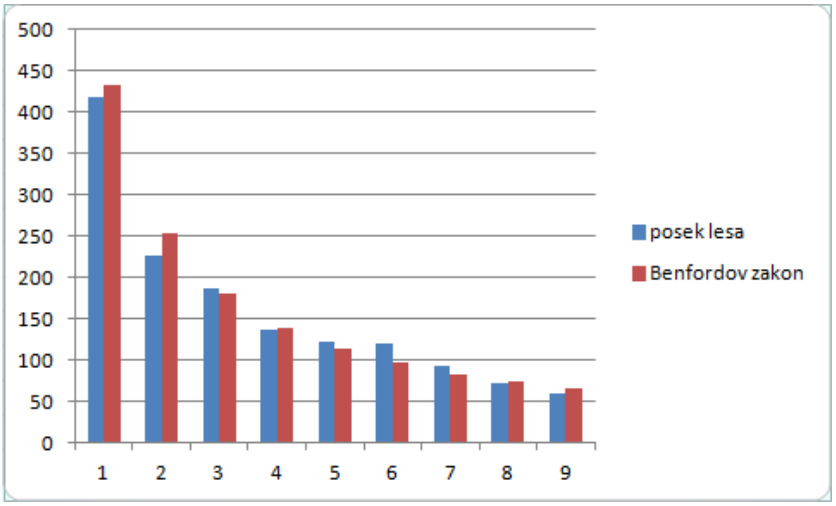
Slika 2: Interval $[2, 3]$ (rdeč) pomnožimo s $k = 2$ in dobimo interval $[4, 6]$ (zelen). Intervala sta prikazana na navadni in logaritemski skali v osmiškem sistemu.

Ker smo spremenili samo osnovo logaritma se pri logaritmu z osnovo b ohrani enakomerna porazdelitev podatkov .

5 Uporabnost Benfordovega zakona

Benfordov zakon danes s pridom izkoriščamo za odkrivanje prevar, pojavlja pa se tudi na mnogih drugih področjih. Finančne transakcije, število prebivalcev, dolžine rek in vrednosti matematičnih ter fizikalnih konstant so le nekatere izmed skupin podatkov, ki sledijo Benfordovemu zakonu. Kot že omenjeno pa se morajo podatki raztezati čez več redov velikosti.

Odločili smo se grafično prikazati primer, ko Benfordov zakon velja za podatke o poseku lesa na družinskih kmetijah za različne namene uporabe. Podatki so bili zbrani v Sloveniji leta 2000 in 2010 in so urejeni po kohezijskih regijah ter velikostnih razredih površine gozda. Modri stolpci na grafu označujejo dejanske podatke o pojavljanju vodilnih števk, medtem ko rdeči stolpci predstavljajo podatke, ki se popolnoma skladajo z Benfordovim zakonom.



Slika 3: prikaz primera Benfordovega zakona

Literatura

[1] Gozdarstvo Slovenija in kohezijske regije [online] [citirano 18. avgust 2017] `http://pxweb.stat.si/pxweb/Database/Kmetijstvo_2010/06_gozdarstvo/01_15P50_kohez_regije/01_15P50_kohez_regije.asp`

[2] Benford's law [online] [citirano 18. avgust 2017] `http://mathworld.wolfram.com/BenfordsLaw.html`

[3] Benford's Law - Wikipedia [online] [citirano 18. avgust 2017] `https://en.wikipedia.org/wiki/Benford%27s_law`