A Benford-törvény
avagy meghamisították-e az adatainkat?

8. előadás

Melyik adathalmaz hamis?

Állam/terület	valódi vagy hamis terület (km^2)	
Afganisztán	645807	796467
Albánia	28748	9943
Algéria	2381741	3168262
Amerikai Szamoa	197	301
Andorra	464	577
Anguilla	96	82
Antigua	442	949
Argentina	2777409	4021545
Aruba	193	367
Ausztrália	7682557	6563132
Ausztria	83858	64154
Azerbajdzsán	86530	71661
Bahamák	13962	9125
Bahrein	694	755
Banglades	142615	347722
Barbados	431	818
Belgium	30518	47123
Belize	22965	20648
Benin	112620	97768
…		

Miért foglalkozzunk az adatok minőségével?

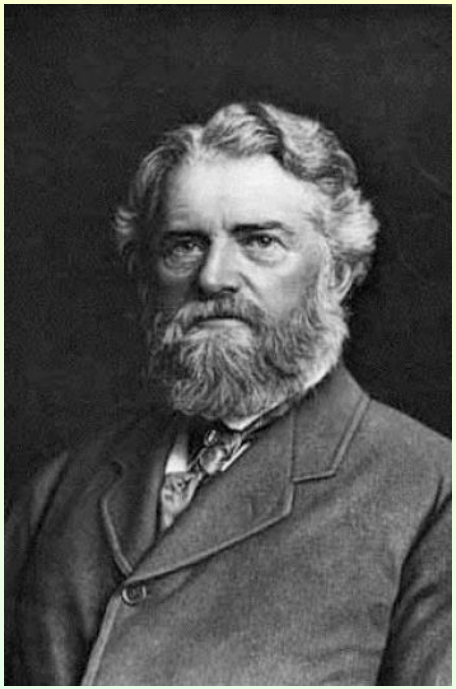
- a nagytömegű adatok mennyiségével fordított arányban csökken az adatok minősége (feldolgozási, programhibák)
- a kitalált, hamis adatok ‘gyártása’ egyre nagyobb probléma, még tudományos körökben is
    - Darsee-eset (1981, Harvard, orvoskutató)
    - Bruening-eset (1979-83 között a mentálisan hátramaradott emberek pszichofarmakológiájával foglalkozó cikkek 34%-át ő írta)

Mi a Benford-törvény?

menjünk vissza 1881-be

1881: Simon Newcomb

1835 – 1909



- Korának leghíresebb amerikai csillagásza
- Matematika és csillagászat professzora a Johns Hopkins Egyetemen
- Michaelson-nal együtt megmérte a fény sebességét
- 1881-ben észrevette, hogy a logaritmus táblázatok eleje elhasználódottabb a végüknél
- Arra következtetett, hogy az 1, 2, 3-mal kezdődő számokat gyakrabban keresik ki, mint a 7, 8, 9-cel kezdődőket
- Feltette, hogy az első számjegyek előfordulásának valószínűsége $P(d) = \log_{10}(1 + 1/d)$, ahol $d = 1, 2, 3, 4, 5, 6, 7, 8, 9$, és $\sum P(d) = 1$
- Eredményét publikálta: Amer J Math 4, 1881 (pp 39-40)

Newcomb (1881) cikke
Amer J Math 4, pp 39-40

Note on the Frequency of Use of the Different Digits in Natural Numbers.

By Simon Newcomb.

That the ten digits do not occur with equal frequency must be evident to any one making much use of logarithmic tables, and noticing how much faster the first pages wear out than the last ones. The first significant figure is oftener 1 than any other digit, and the frequency diminishes up to 9. The question naturally arises whether the reverse would be true of logarithms. That is, in a table of anti-logarithms, would the last part be more used than the first, or would every part be used equally? The law of frequency in the one case may be deduced from that in the other. The question we have to consider is, what is the probability that if a natural number be taken at random its first significant digit will be n, its second n', etc.

1938: Frank Benford

1883 – 1948

- fizikus a General Electric-nél
- ő is észrevette, hogy a logaritmus táblázatok eleje koszosabb a végüknél

- Arra következtetett, hogy az 1, 2, 3-mal kezdődő számokat gyakrabban keresik ki, mint a 7, 8, 9-cel kezdődőket
- Feltette, hogy az első számjegyek előfordulásának valószínűsége $P(d) = \log_{10}(1 + 1/d)$, ahol $d = 1, 2, 3, 4, 5, 6, 7, 8, 9$, és $\sum P(d) = 1$
- Megvizsgált különböző adathalmazokat:
  - 335 folyó területe, 3259 település lakosságszáma, a természetes számok hatványai, kémiai elemek mol-tömegei, fizikai állandók, stb…

F. Benford (1938) cikke

Proc. Amer. Phil. Soc. 78, 551-572

THE LAW OF ANOMALOUS NUMBERS

FRANK BENFORD

Physicist, Research Laboratory, General Electric Company,
Schenectady, New York

(Introduced by Irving Langmuir)

(*Read April 22, 1937*)

ABSTRACT

It has been observed that the first pages of a table of common logarithms show more wear than do the last pages, indicating that more used numbers begin with the digit 1 than with the digit 9. A compilation of some 20,000 first digits taken from widely divergent sources shows that there is a logarithmic distribution of first digits when the numbers are composed of four or more digits. An analysis of the numbers from different sources shows that the numbers taken from unrelated subjects, such as a group of newspaper items, show a much better agreement with a logarithmic distribution than do numbers from mathematical tabulations or other formal data. There is here the peculiar fact that numbers that individually are without relationship are, when considered in large groups, in good agreement with a distribution law—hence the name " Anomalous Numbers."

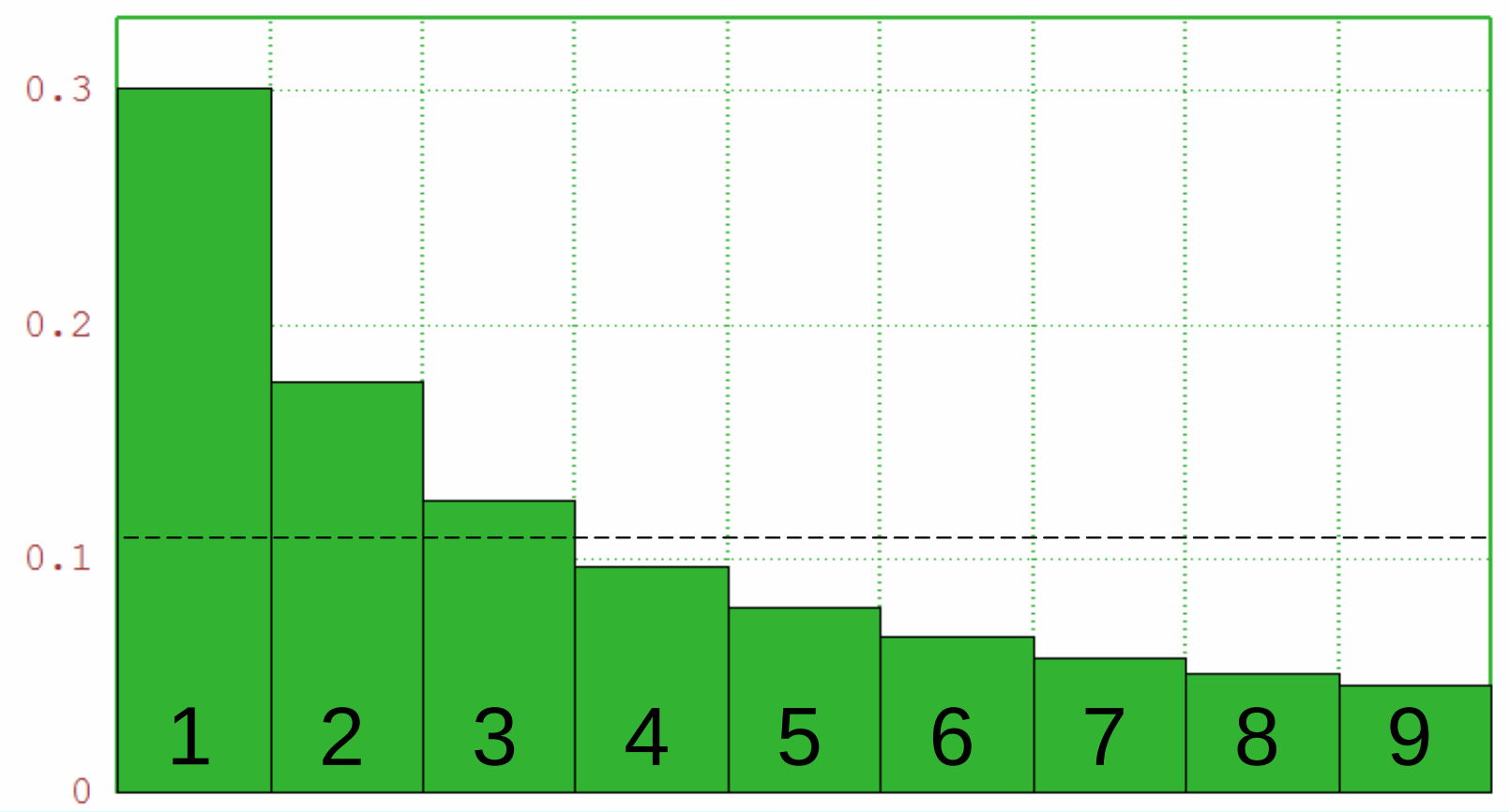
A Benford-törvény

- Nagyon sok számhalmazban a számok első értékes számjegyeinek eloszlása ezt a törvényt követi:

$$P(d) = \log_{10}\left(1+\frac{1}{d}\right), \quad \text{ahol} \quad d = 1, 2, ..., 9$$

d	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
***P* (%)**	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6

A Benford-törvény

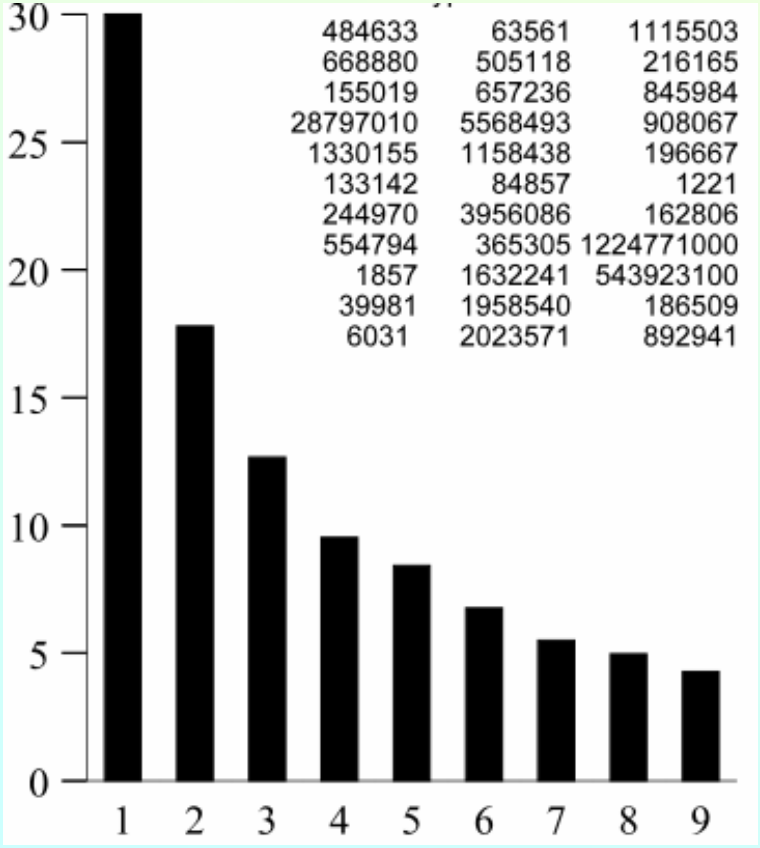


A Benford által vizsgált adatok

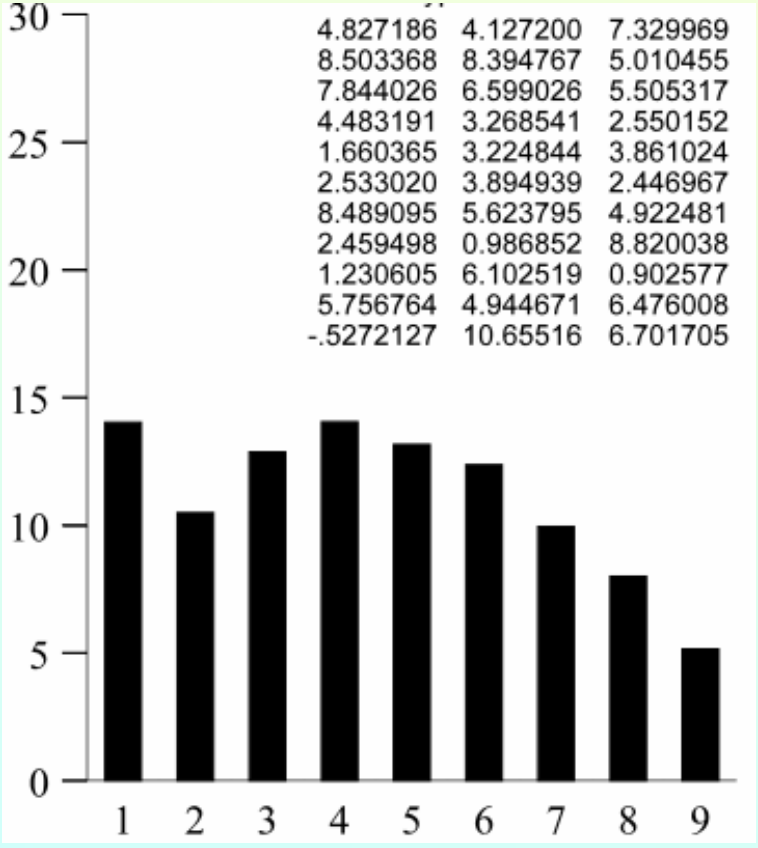
oszlop	név	1	2	3	4	5	6	7	8	9	minta
A	folyók, terület	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
B	népesség	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
C	állandók	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
D	újságok	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
E	fajhő	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
F	nyomásveszteség	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
G	teljesítményveszteség	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
H	moláris tömeg	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
I	csatornahálózat	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
J	atomtömeg	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
K	n^{-1}, $\sqrt{n}$	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5000
L	tervezés	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
M	Reader's Digest	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
N	költségadatok	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
O	röntgensugár feszültségek	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
P	Amerikai Liga	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1458
Q	feketetest	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
R	címek	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
S	n^1, n^2, ... n!	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
T	halálozási arány	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
	átlag	30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1011
	valószínű hiba	±0.8	±0.4	±0.4	±0.3	±0.2	±0.2	±0.2	±0.3		

Benford vagy nem Benford?

Smith (1997)



jövedelemadó, Benford



véletlen számok,
nem Benford

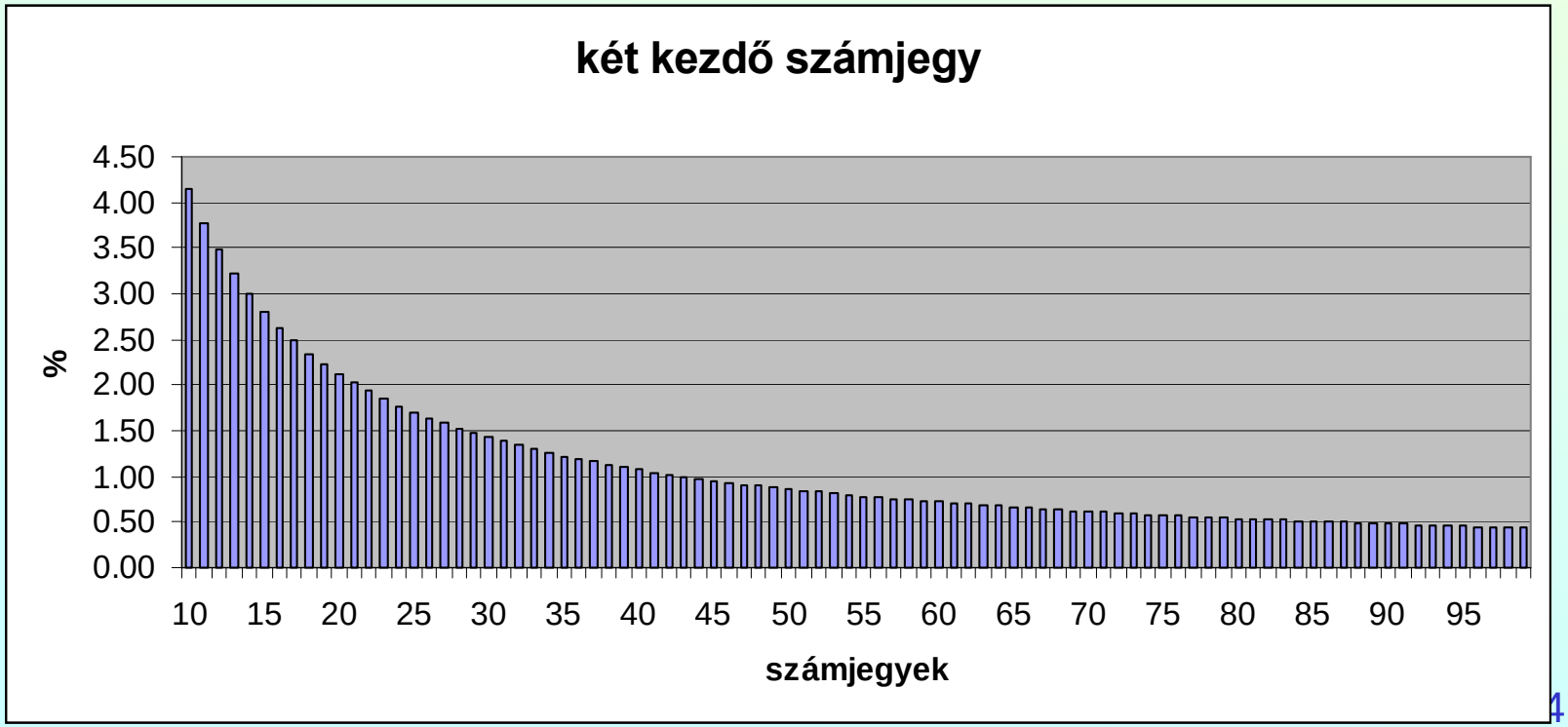
Tulajdonságok

- Ha az eloszlás Benford, akkor bármilyen számmal szorozva/osztva is Benford marad (**skála invariáns**)

- Ha az eloszlás Benford, akkor egy másik számrendszerben is Benford marad (**számrendszer invariáns**)

Első két számjegyre vonatkozó Benford-törvény

$$P(d) = \log_{10}\left(1 + \frac{1}{d_1 d_2}\right), \quad \text{ahol} \quad d_1 d_2 = 10, 11, ..., 99$$



Benford-törvényből várt számjegy gyakoriságok különböző helyiértékeken

számjegy	1. hely	2. hely	3. hely	4. hely
0		0.11968	0.10178	0.10018
1	0.30103	0.11389	0.10138	0.10014
2	0.17609	0.19882	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	0.09350	0.09902	0.09990
8	0.05115	0.08757	0.09864	0.09986
9	0.04576	0.08500	0.09827	0.09982

Mi a Benford-törvény magyarázata?

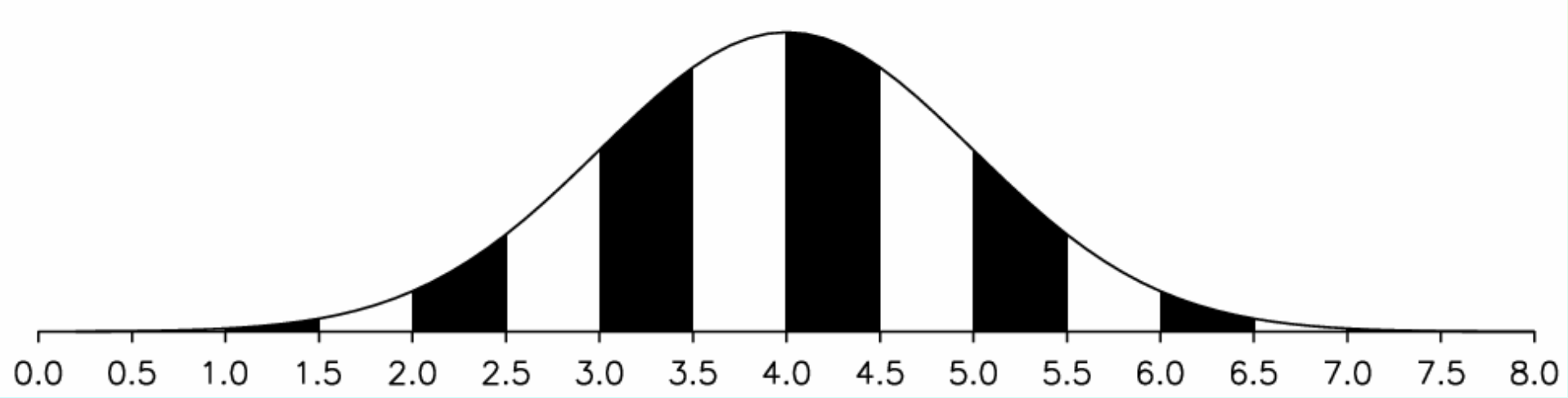
- a követőinek szinte kultusza van
- a természet valamilyen misztikus vagy paranormális jellemzője??
- Benford: „az ember egyesével számol: 1,2,3,4,…, a Természet így számol: e^0, e^1, e^2, e^3…”
- a Természetben van egy univerzális számeloszlás, függetlenül attól, hogyan vizsgáljuk
- stb…

- ezek a „magyarázatok” **mind** rossz irányba mennek

- a Benford-törvénynek **egyszerű** és **logikus** magyarázata van, ami mentes minden misztikától (Fewster, 2009)

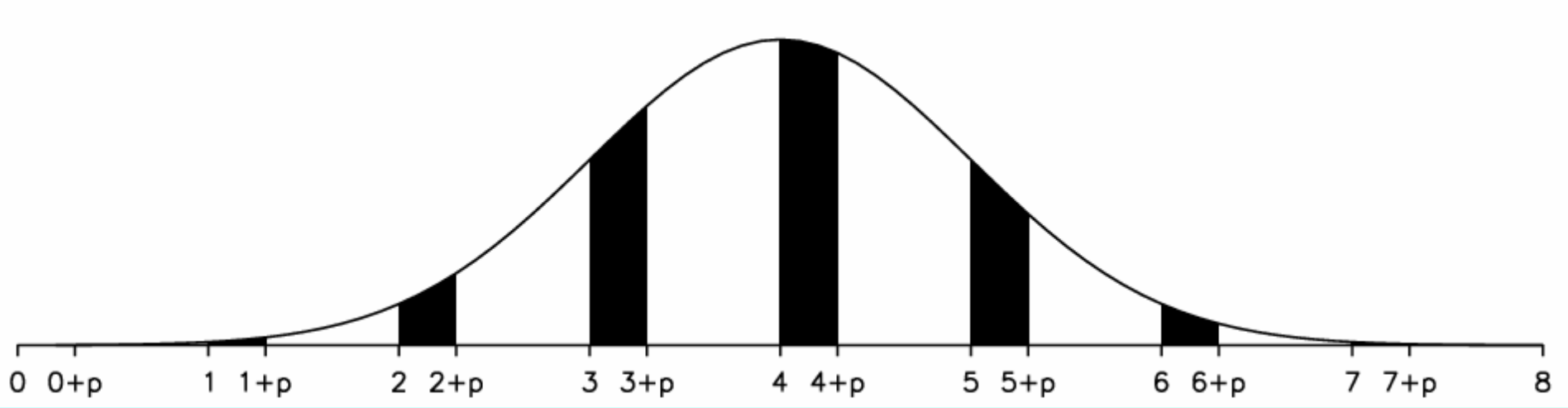
Valószínűség sűrűség függvény

- ha egy „kalapot” (= vsz. sűrűség függvény) egyenletesen becsíkozunk, nagyjából a fele lesz fekete



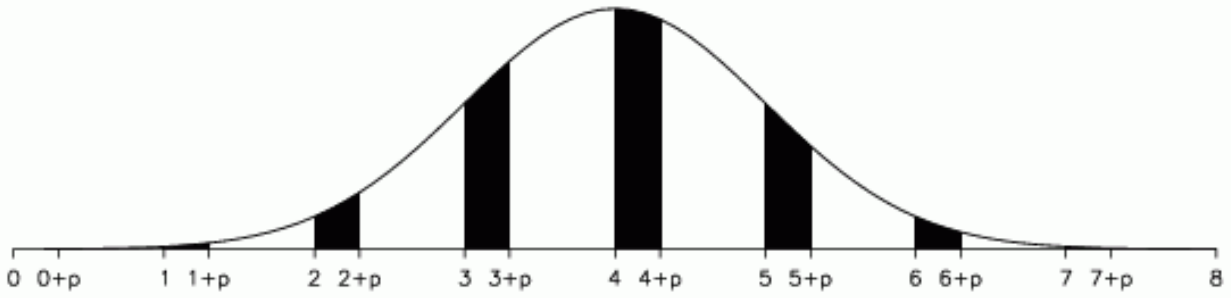
Valószínűség sűrűség függvény

- ha a *p*-ed részét csíkozzuk be, a terület is körülbelül a *p*-ed részére változik



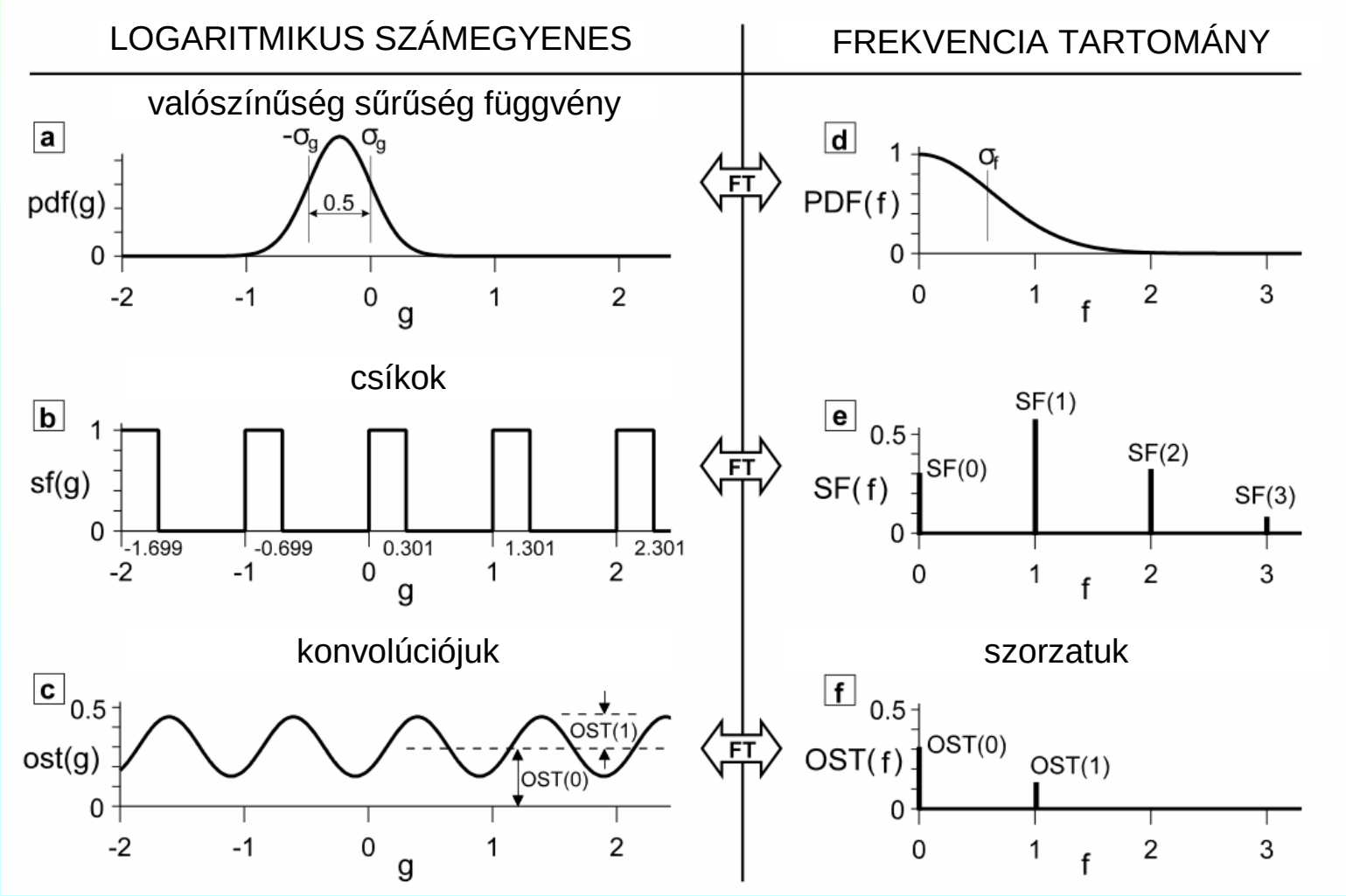
Valószínűség sűrűség függvény

- ha eltoljuk a csíkokat, átlagosan ezek a területnek körülbelül a *p*-ed részét fedik le



Konvolúció!

- a csíkok eltolása szabatosabban is megfogalmazható konvolúcióként és megoldható a frekvencia tartományban (Smith, 1997):



Első számjegy és logaritmus

- bármely pozitív *X* egész számnak az első számjegye pontosan akkor 1, ha $\log_{10}(X)$ értéke *n* és *n* + 0.301 közé esik valamilyen *n* egész számra ($\log_{10}2=0.301$)
- ha *X* egy valószínűségi változó, akkor a „kalap” a **$\log_{10}(X)$ valószínűség sűrűség függvény**e
- az **1**-el kezdődő *X* számok azok a **csíkok**, amelyek *n* és *n* + 0.301 közé esnek valamilyen *n* egész számra

Kapcsolat a Benford-törvénnyel

- a csíkok a „kalap” kb. 0.301-ed részét töltik ki, vagyis az *X* teljes valószínűségének kb. 0.301-ed részét kapjuk meg (a görbe alatti terület 1)

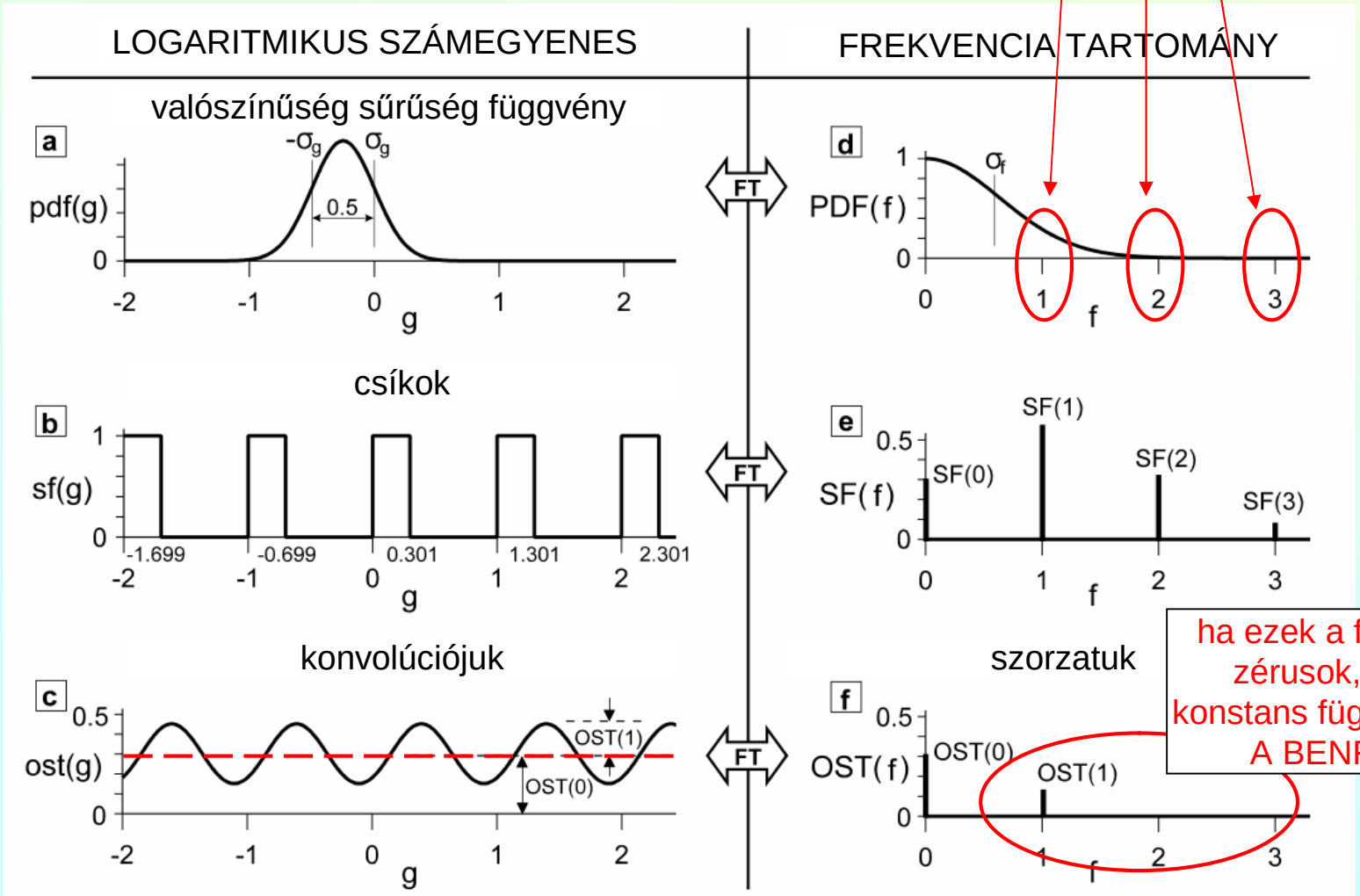
- az 1-el kezdődő *X* számok valószínűsége tehát 0.301 lesz, ahogy a Benford-törvény kimondja

Mikor kapunk Benford-eloszlást?

- ha több a csík, a területek jobban kiegyenlítődnek, így a csíkok összterülete jobban közelít 0.301-hez: az eloszlás jobban „Benford” lesz
- mivel a csíkok távolsága adott, szélesebb „kalap” esetén lesz több csík
- $\log_{10}(X)$ eloszlásának terjedelme nagyobb: ***X több nagyságrendet fog át***
  - pl. ha X 1-10^6 közötti, $\log_{10}(X)$ 6 csíkot tartalmaz
  - ez elég meggyőzően „Benford” eloszlást fog adni

Mikor kapunk Benford-eloszlást?

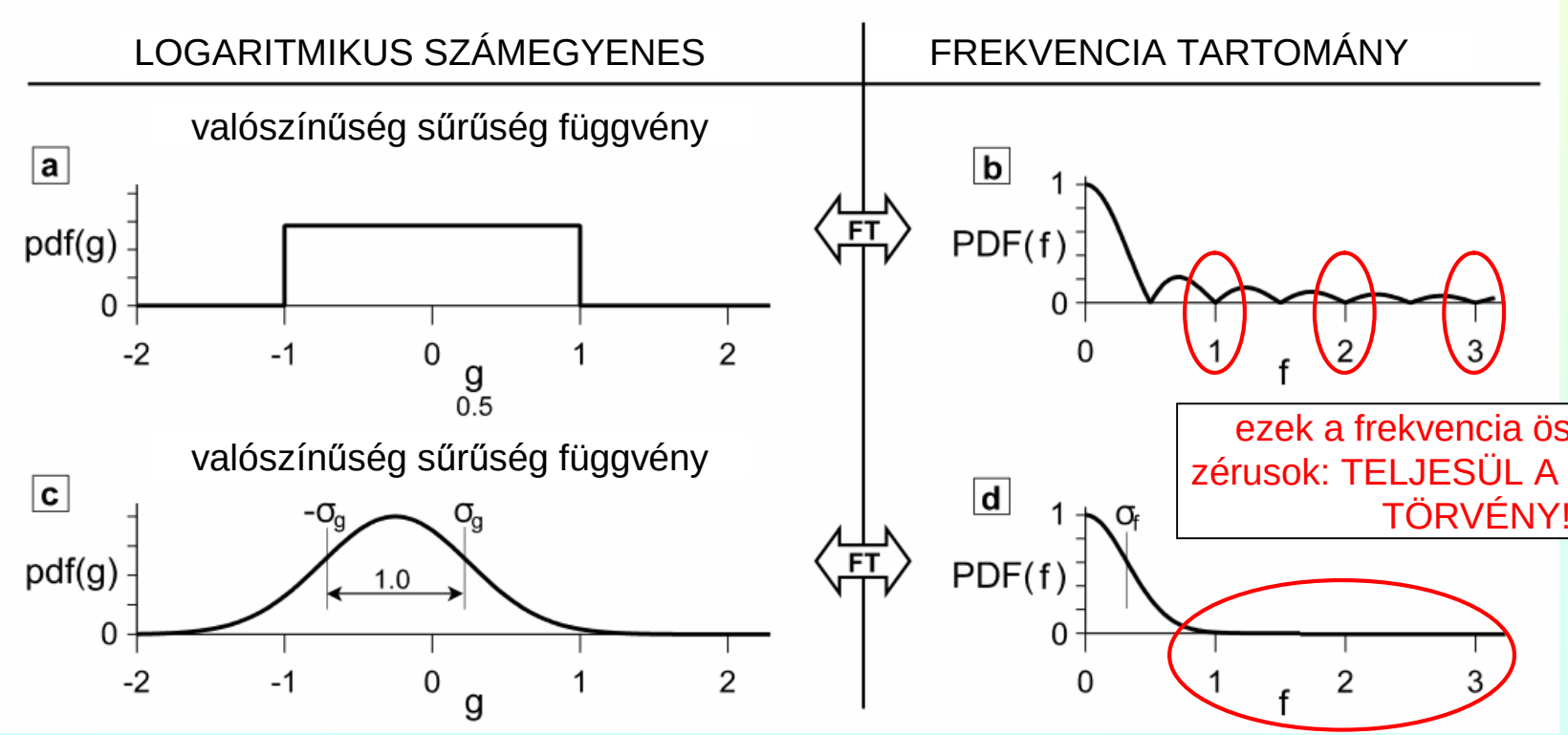
- akkor, ha a PDF(f) értéke zérus az egész értékű nemzérus *f* frekvenciákon (*f* = 1, 2, 3, ...):



ha ezek a frekvencia összetevők zérusok, akkor a konvolúció konstans függvény lesz: TELJESÜL A BENFORD-TÖRVÉNY!

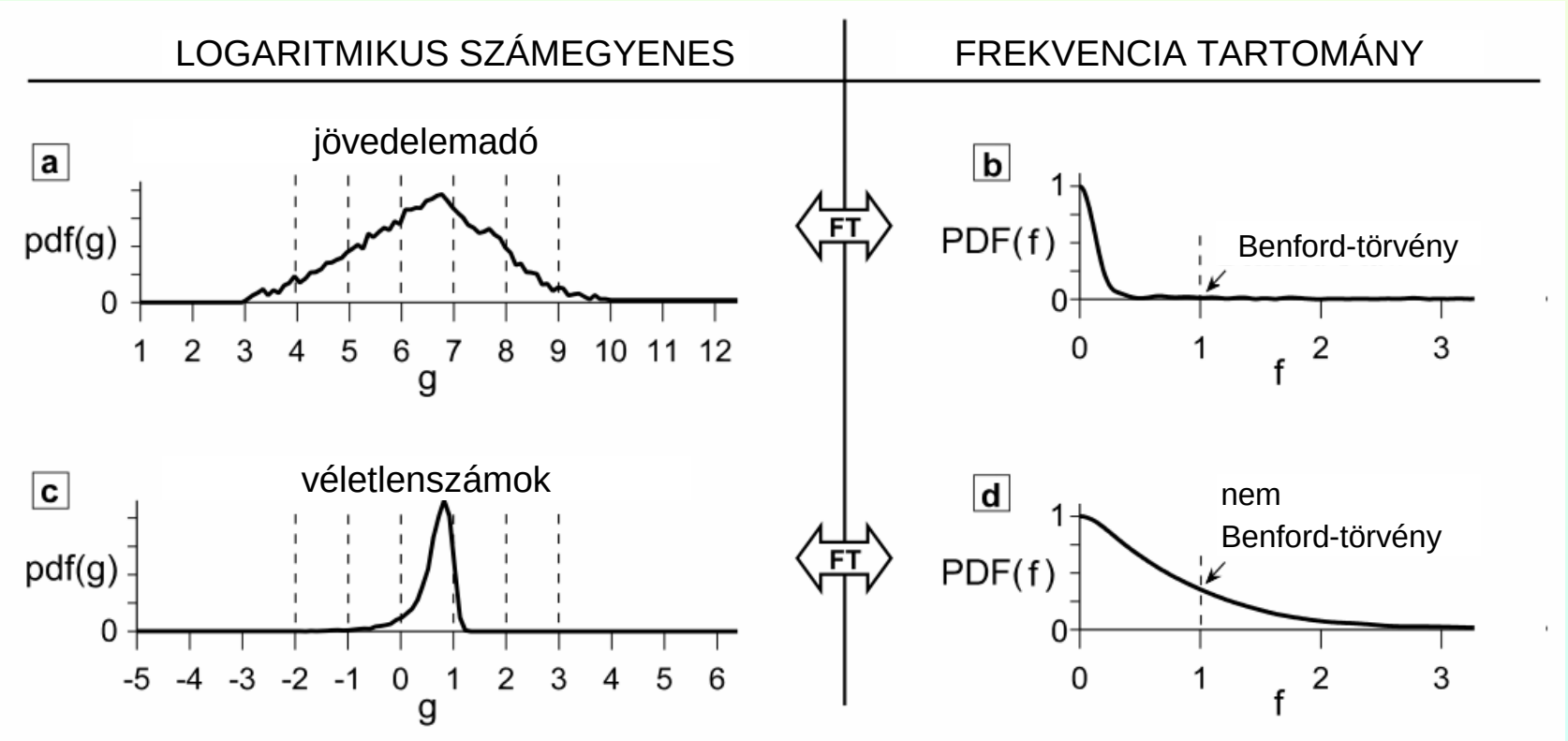
Mikor kapunk Benford-eloszlást?

- kétféle lehetőségünk van erre:



Bevezetőben említett példák

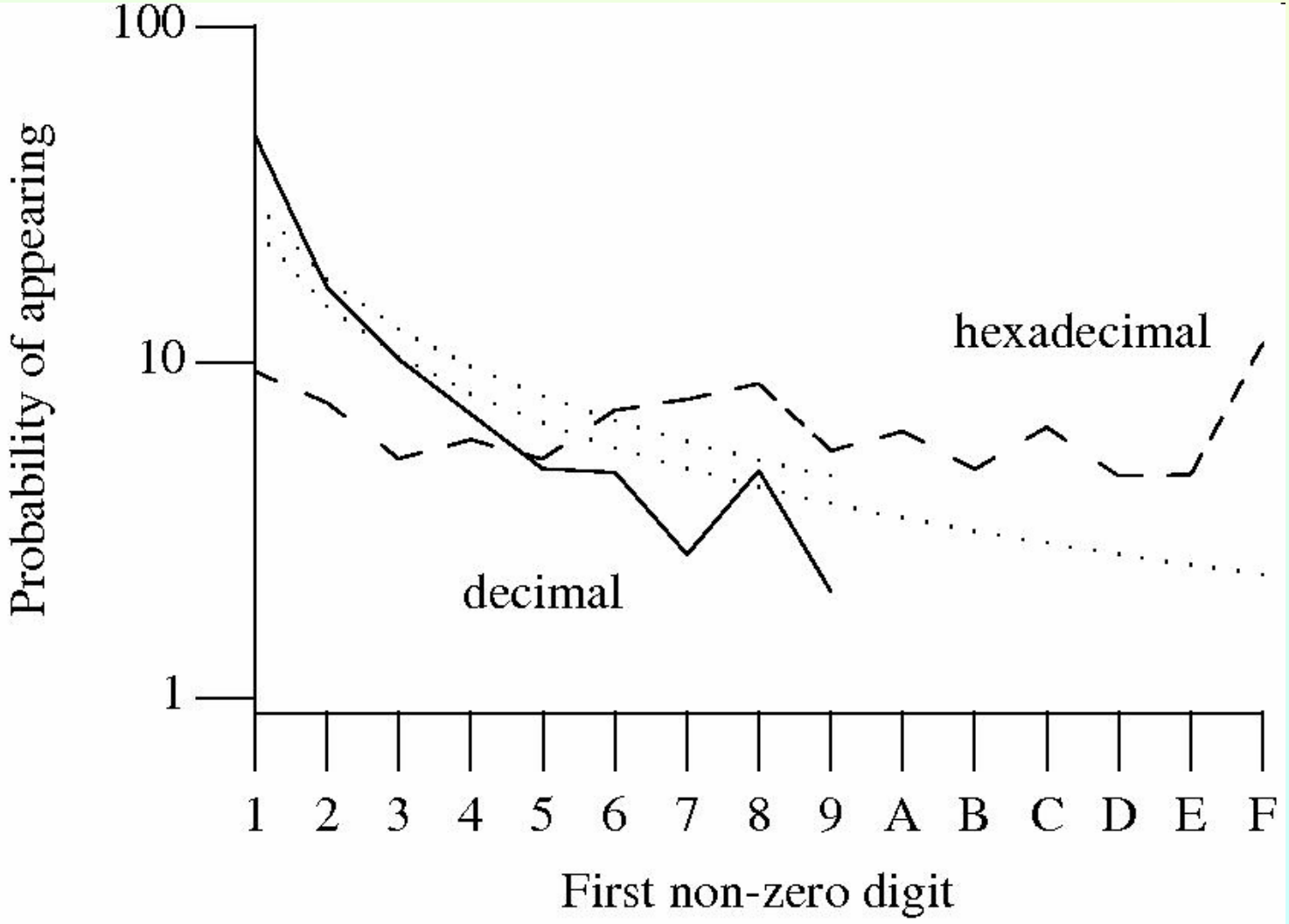
- Benford és nem Benford eloszlások:



Mi van a skála és számrendszer invarianciával?

- ha **szorozzuk/osztjuk** az adatokat, a $\log_{10}(X)$ eloszlása csak jobbra/balra **eltolódik**, alakja, terjedelme nem változik meg
- ha áttérünk **más számrendszer**re, megváltozik a csíkok távolsága
  - ha az alapszám **10-nél kisebb**, a csíkok **sűrűbbek**, jobban „Benford” lesz az eloszlás
  - ha az alapszám **10-nél nagyobb**, a csíkok **ritkábbak**, kevésbé „Benford” lesz az eloszlás

- C forráskódban található számok (Derek-Jones, 2008):



Mi van a többi számjeggyel?

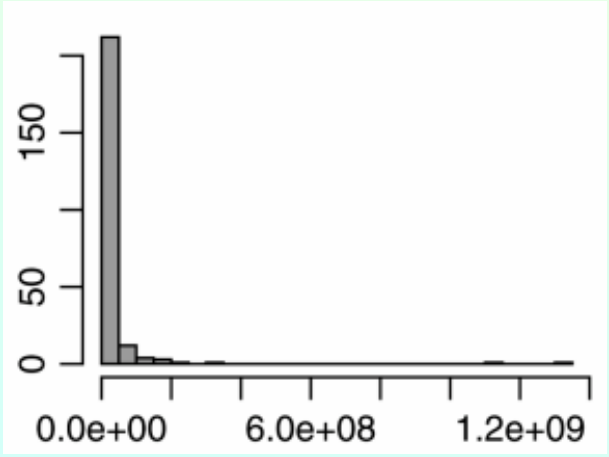
- a fenti gondolatmenet pontosan ugyanaz a 2-vel, 3-mal, … kezdődő számokra, csak a csíkok nem n és $n + \log_{10}2$ közé, hanem

  $n + \log_{10}d$ és $n + \log_{10}(d+1)$

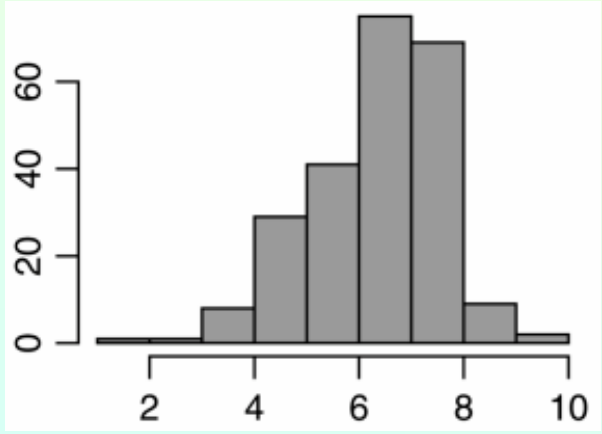
  közé fognak esni ($\log_{10}1 = 0$)
- az intervallum hossza pedig $\log_{10}(d+1) - \log_{10}d = \log_{10}(1 + 1/d)$

Példák eloszlásokra

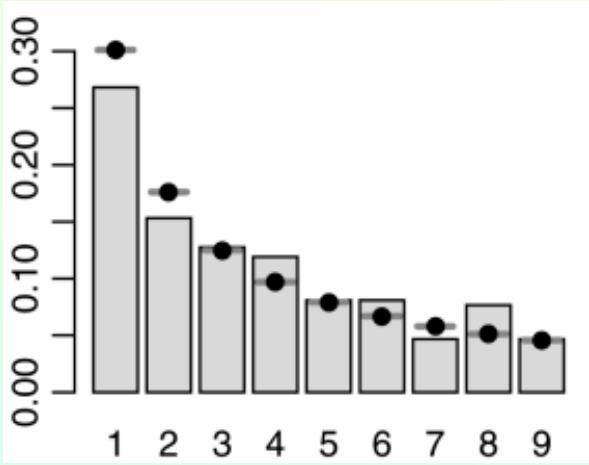
a világ államainak a népessége



népesség



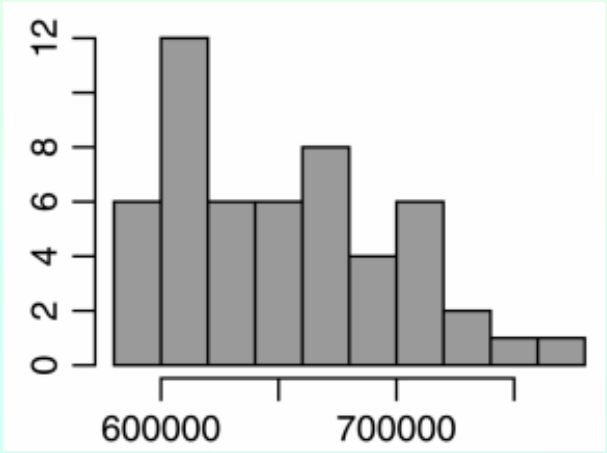
log10 (népesség)



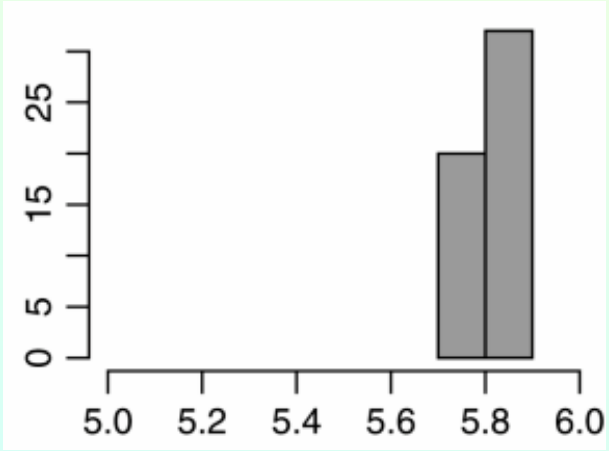
első számjegyek

Példák eloszlásokra

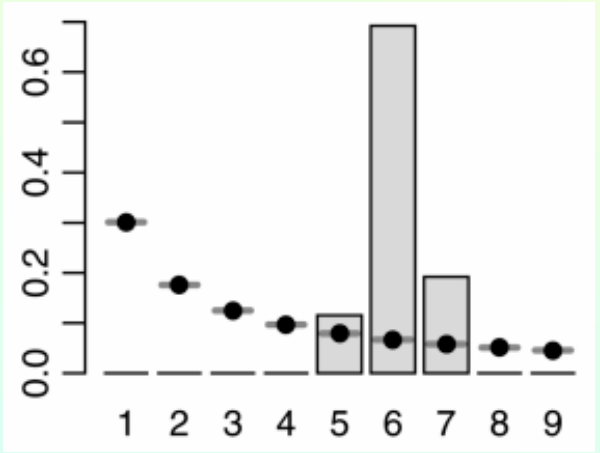
Kalifornia választókerületei



népesség



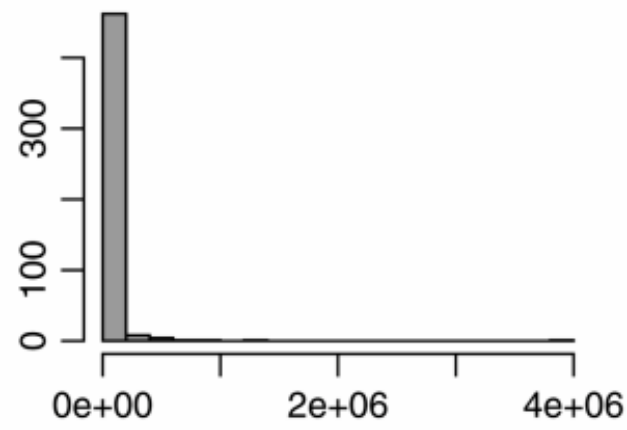
log10 (népesség)



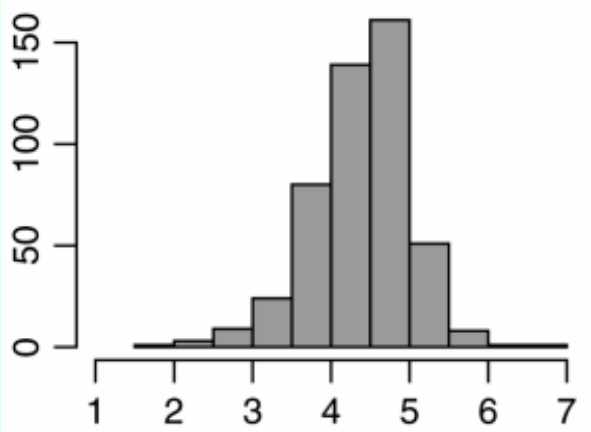
első számjegyek

Példák eloszlásokra

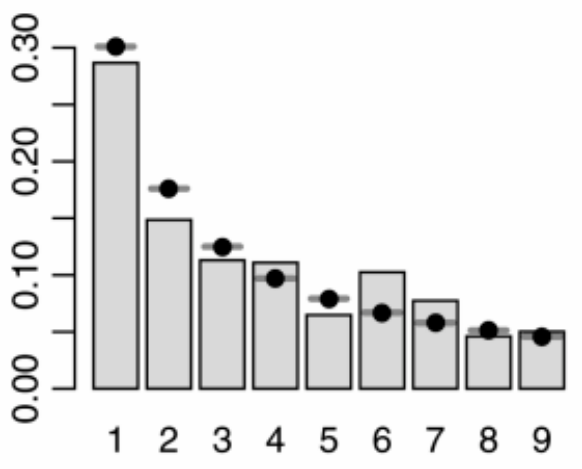
Kalifornia városai



népesség



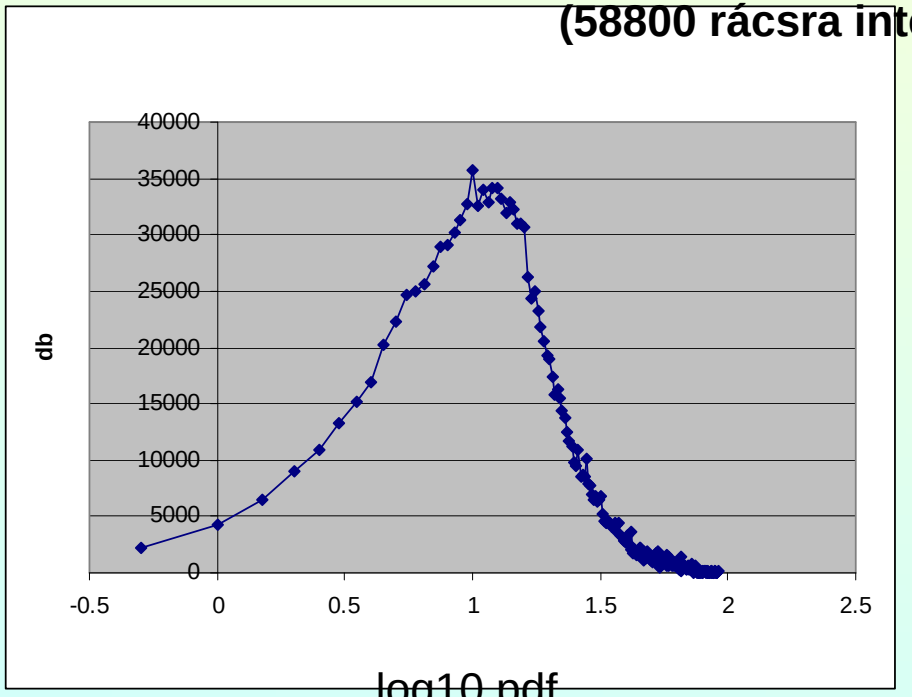
log10 (népesség)



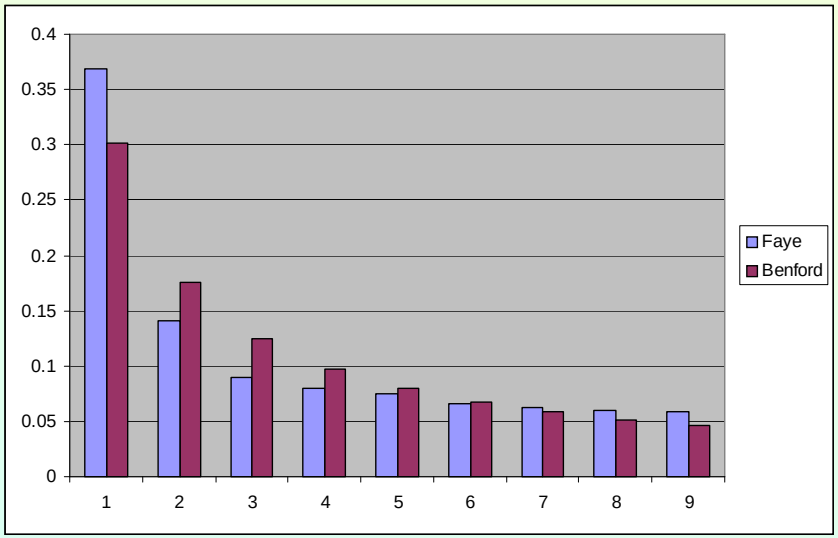
első számjegyek

Példák eloszlásokra

Magyarországi Faye nehézségi rendellenességek (58800 rácsra interpolált adat, ELGI)

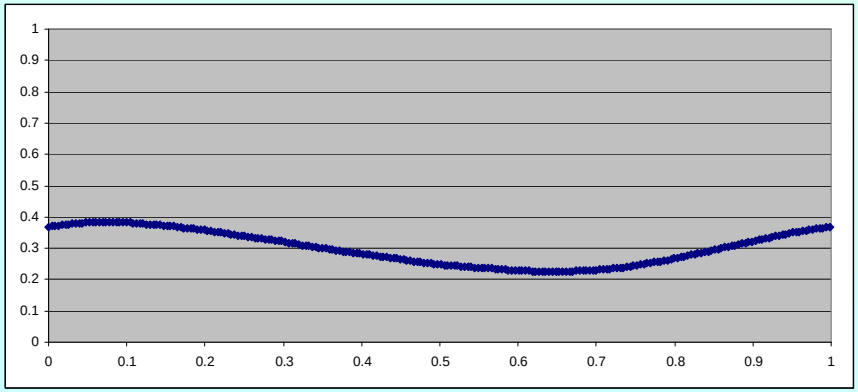


log10 pdf



első számjegyek

1-es számjegy konvolúció eredménye



Mikor kapunk Benford-eloszlást?

- Minél **több nagyságrend** az adataink terjedelme, annál inkább Benford-eloszlást kapunk (ezen az átskálázás nem változtat!)
- A $\log_{10}(X)$ valószínűség sűrűség függvénynek ésszerűen „**simá**nak” kell lennie

Mikor kapunk Benford-eloszlást?

- Kivételes esetben, ha a $\log_{10}(X)$ valószínűség sűrűség függvénye **konstans**, akkor nem lényeges követelmény az adatok terjedelme

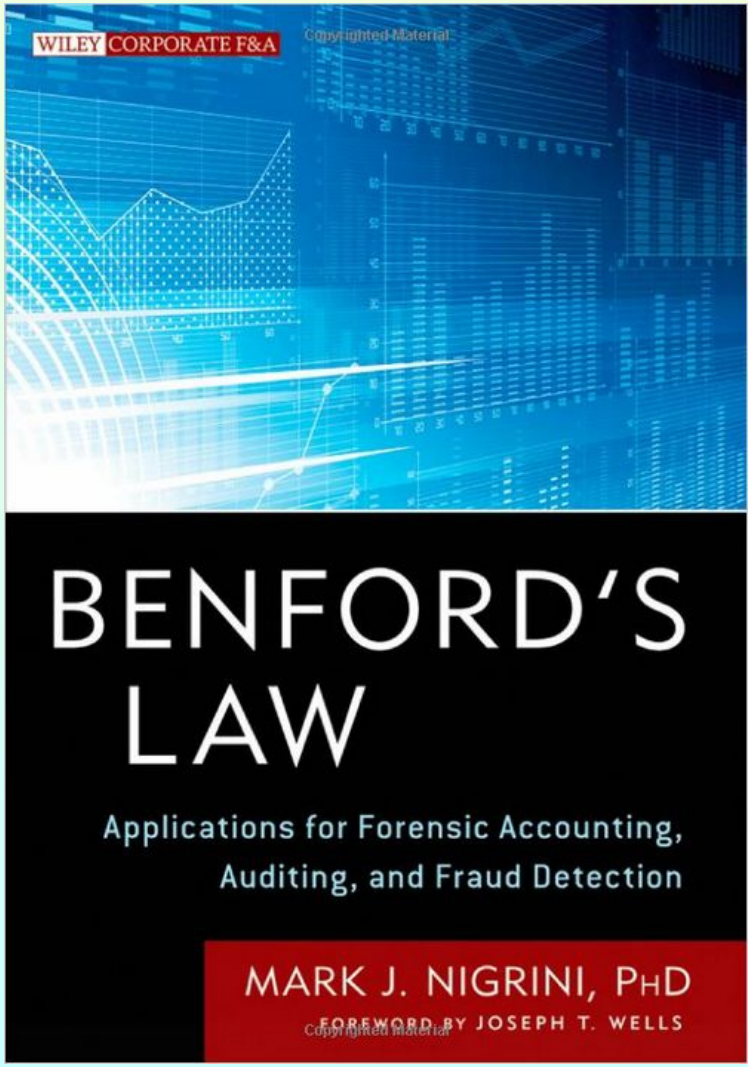
- A fenti feltételek (terjededelem, simaság) sok eloszlás függvényre igazak, ezért **gyakran kapunk Benford-eloszlást**

Most már tudjuk, melyik adathalmaz hamis?

Állam/terület	valódi vagy hamis terület (km^2)	
Afganisztán	645807	796467
Albánia	28748	9943
Algéria	2381741	3168262
Amerikai Szamoa	197	301
Andorra	464	577
Anguilla	96	82
Antigua	442	949
Argentina	2777409	4021545
Aruba	193	367
Ausztrália	7682557	6563132
Ausztria	83858	64154
Azerbajdzsán	86530	71661
Bahamák	13962	9125
Bahrein	694	755
Banglades	142615	347722
Barbados	431	818
Belgium	30518	47123
Belize	22965	20648
Benin	112620	97768
…		

Néhány alkalmazás adatok feldolgozására

- **ellenőrzés**: **adó-** illetve **könyvelési csalás**ok lebuktatására (Nigrini), hamisított interjúk, kérdőívek felderítésére statisztikai adatfelvétel esetén
- **választási csalások** kiderítésére (Irán, 2009)
- a processzorok lebegőpontos számításokhoz használt inputjainak eloszlása a Benford törvényt követi – ezt figyelembevéve megnőhet a számítási sebesség
- földrengés beérkezésének detektálása
- stb...



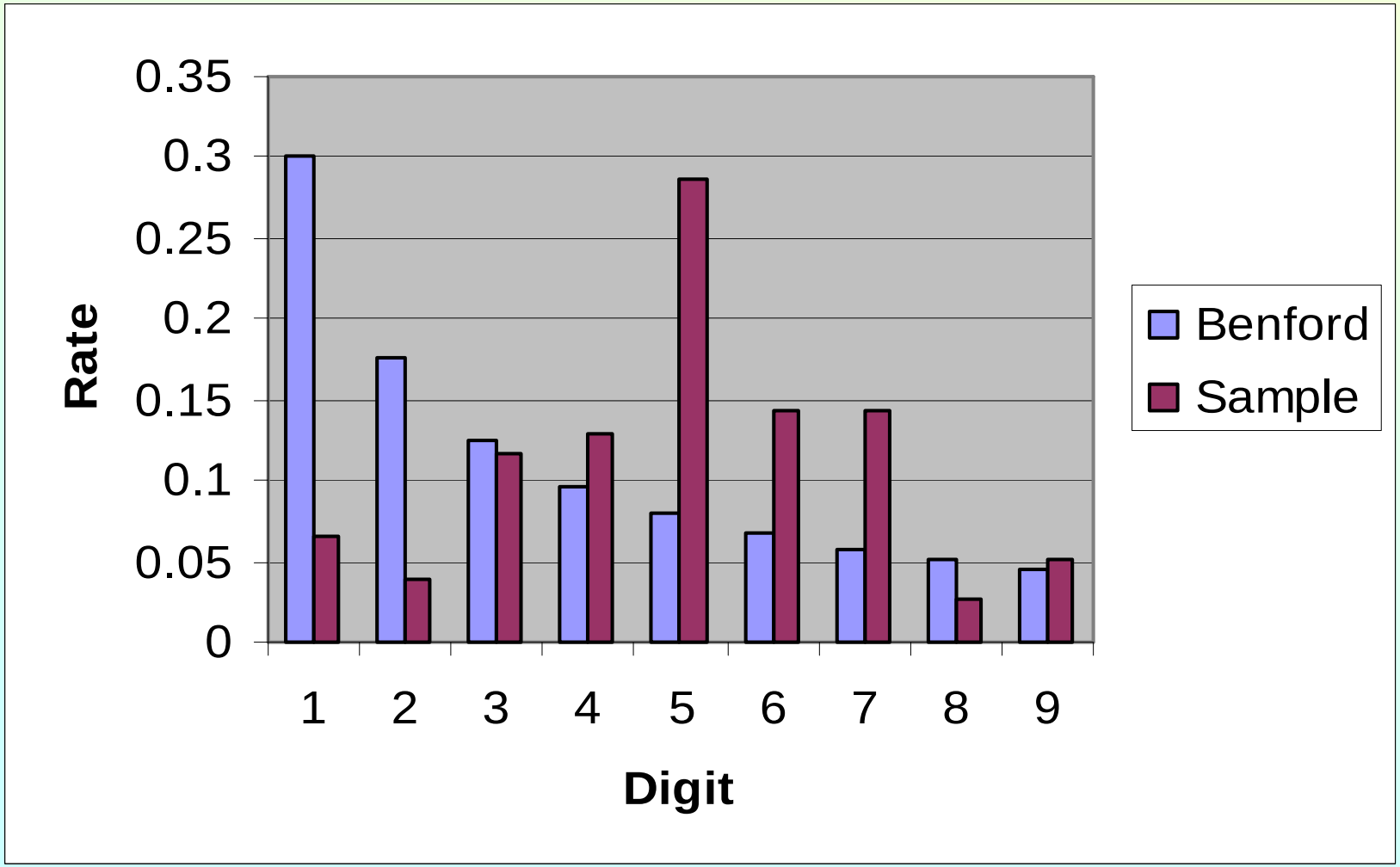
2012

Könyvelési csalás felderítése

Rose (2003) nyomán

- egy magán kisvállalkozás kibővítette egy áruházas családi vállalkozását négy áruházból álló üzletlánccá
- ki kellett engednie a kezéből a közvetlen irányítást bizonyos területeken
- aggódott a könyvelési hibák és csalás miatt
- Excelben elemezte az áruház kifizetéseit a Benford-törvény szerint

Első számjegy teszt



Első számjegy teszt

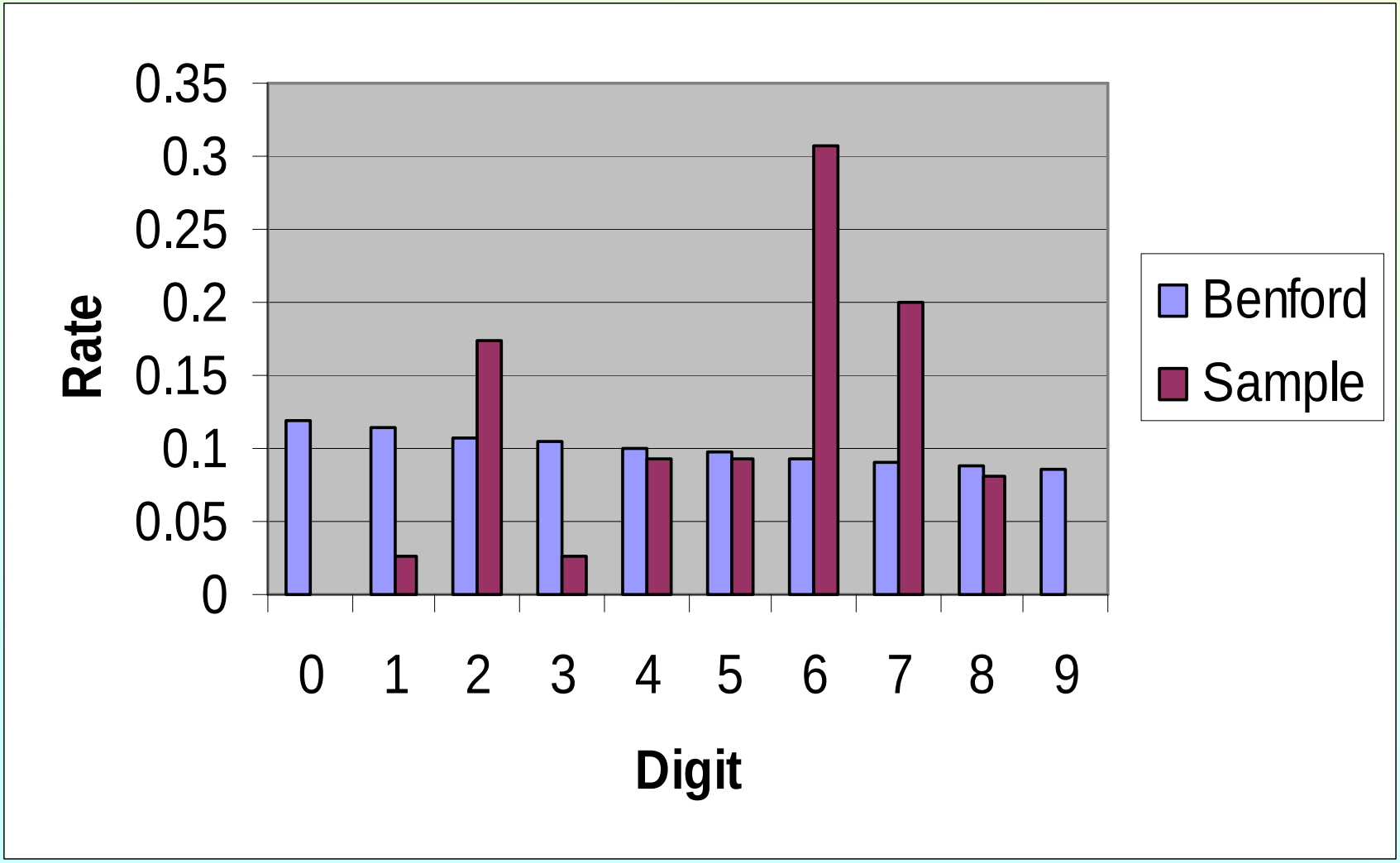
$$MAD = \sum_{i=1}^{9} \frac{\left| p^{i}_{empirikus} - p^{i}_{benford} \right|}{9}$$

alkalmazás / döntés	**Első számjegyek**	**Második számjegyek**	**Első két számjegy**
jó illeszkedés	< 0.004	< 0.008	< 0.006
elfogadható illeszkedés	0.004 – 0.008	0.008 – 0.012	0.006 – 0.012
Gyenge illeszkedés	0.008 – 0.012	0.012 – 0.016	0.012 – 0.018
nincs illeszkedés	> 0.012	> 0.016	> 0.018

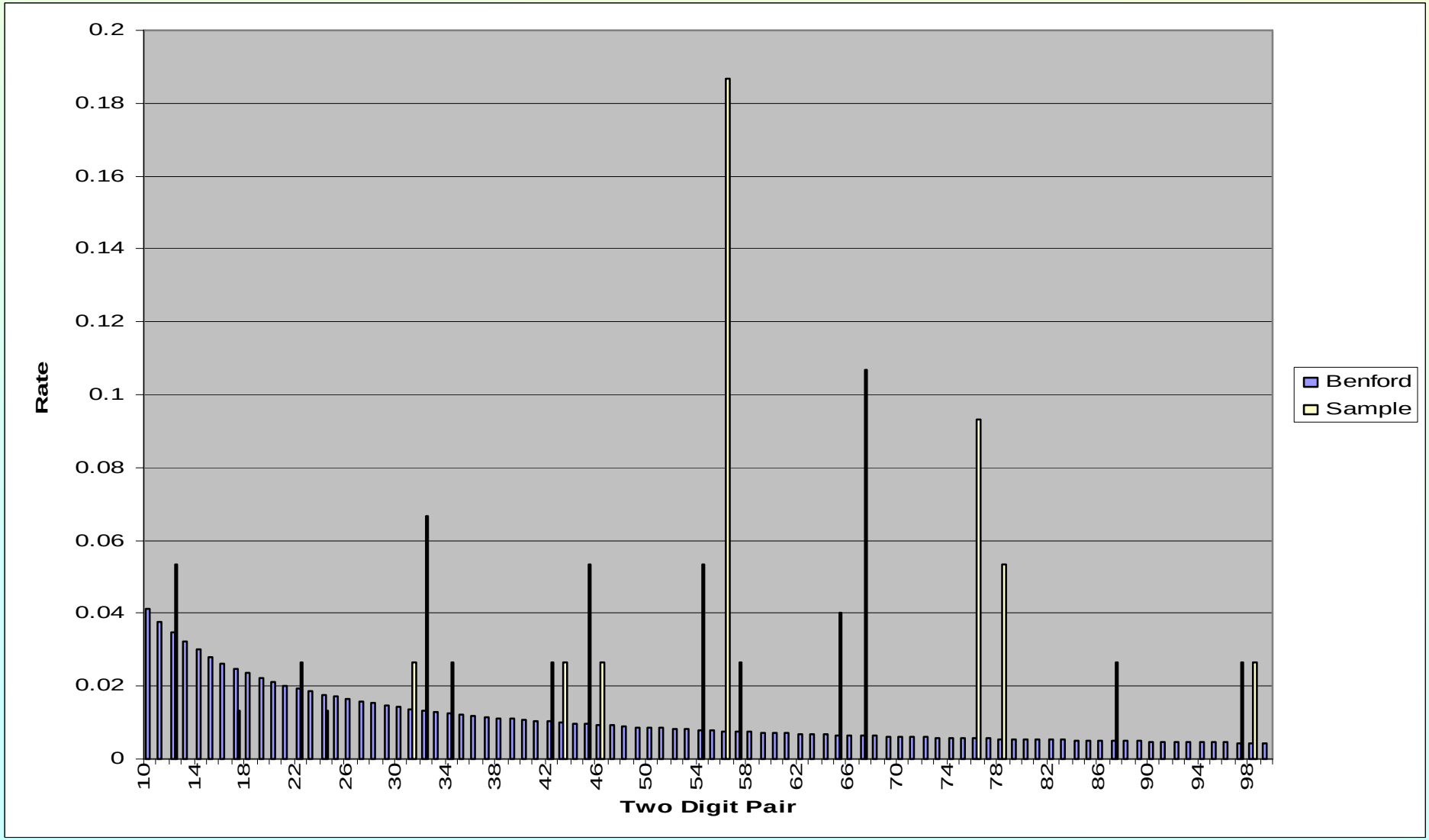
Forrás: Nigrini [2000]

6. Táblázat: A MAD teszt Nigrini által javasolt kritikus értékei

Második számjegy teszt



Első két számjegy teszt



Elemzés

- Első számjegy teszt
  - az 5, 6, 7-es számjegyek sokkal gyakoribbak a vártnál, viszont az 1-es sokkal kevésbé gyakori
- Második számjegy teszt
  - a 6, 7-es számjegyek ismét sokkal gyakoribbak, és a 0 egyáltalán nem fordult elő
- Első két számjegy teszt
  - az 56 és 67 a vártnál sokkal többször fordult elő
- A tulajdonos megkereste az 56-os és 67-es számjegyekkel kezdődő kifizetésekről
  - ismeretlen beszállítónak teljesített kifizetéseket talált
  - a további vizsgálat feltárta, hogy a beszállító nem létezik: a kifizetések magán számlára történtek

Egy 2011-es hír

Matematikusok igazolták a görög csalást

2011.10.25. 15:10 - Index |

Tudományos bizonyíték támasztja alá amit már sokan gyanítottak: Görögország éveken keresztül meghamisította költségvetését - legalábbis ezt állítják az ilmenaui műszaki egyetem matematikusai.

„Meghamisított számoknál a számjegyek eloszlása eltérést mutat a Benford-képlethez képest" – állítják az egyetem kutatói, ami kreatív könyvvitelre enged következtetni.

A matematikusok az összes uniós tagország 1999 és 2009 közötti adatait alapul véve végezték el számításaikat és a legnagyobb eltérést a Benford-képlettől Görögország esetében tapasztalták. Minden egyes ország esetében 156 adatot vizsgáltak meg, beleértve az adósságállományt, a beruházásokat és a költségvetési kiadásokat is.

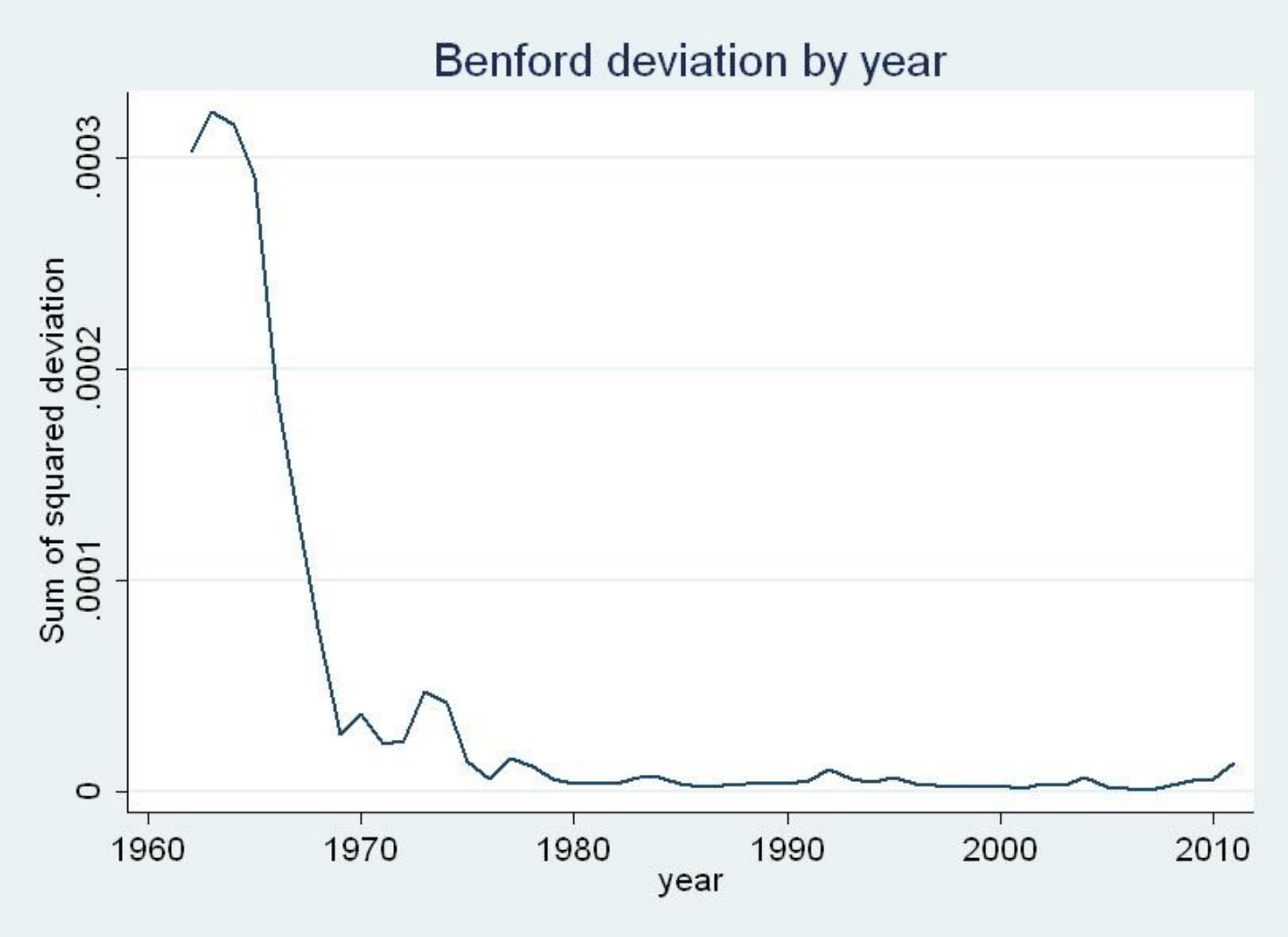
A matematikusok egyúttal arra a meglepő megállapításra jutottak, hogy Belgium adatai alig valamivel bizonyultak jobbnak Görögországénál, ami szerintük megérne egy alaposabb vizsgálódást is.

Compustat pénzügyi adatbázis

- 20000 cég könyvelési adatainak eltérése a Benford-törvénytől
- egyre kevésbé tükrözik a valóságot... ☹

Compustat pénzügyi adatbázis 2.

- A szerző azóta javította a grafikont, mert a zérus adatok is szerepeltek a korábbi statisztikában... ☺



Földrengések detektálása Benford-törvény alapján (Sambridge et al. 2010)

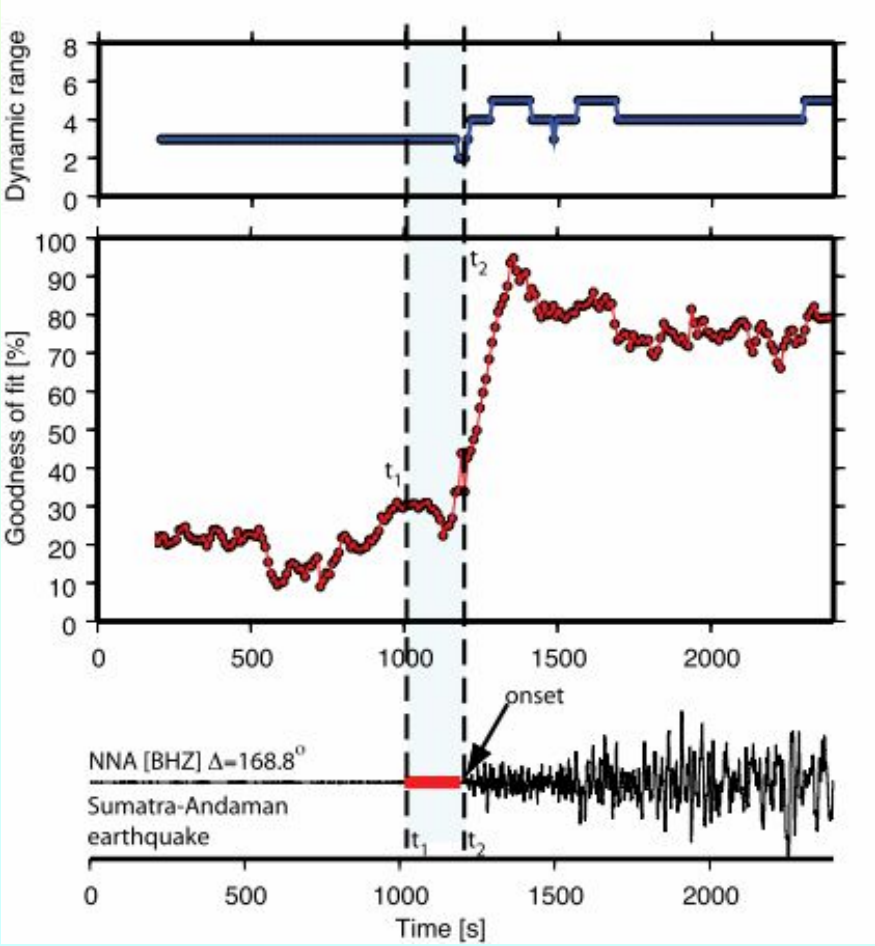
- detektálható-e egy földrengés csupán a szeizmikus idősor adatok első számjegyei alapján?
- felszín elmozdulások a 2004-es Szumátra-Andaman földrengés kapcsán, Peru állomás
- 200 másodperces mozgó ablak alapján illesztési jellemzőt számítottak:

n_D: mért, P_D: elméleti gyakoriság, n: adatok száma

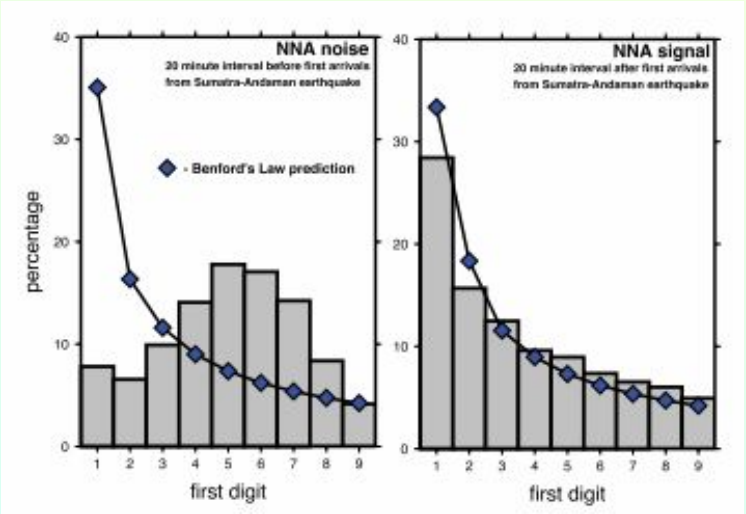
$$\phi = \left[1 - \left(\sum_{D=1}^{9} \frac{(n_D - nP_D)^2}{nP_D} \right)^{1/2} \right] \times 100$$

Eredmény

Φ



első számjegy eloszlása



rengés előtt rengés közben

Hivatkozások

- Derek-Jones (2008). Benford’s law and numeric literals in source code, The Shape of Code, online
- Fewster, RM (2009). A simple Explanation of Benford's Law. American Statistician 63(1), 20-25.
- Lolbert Tamás (2008). Statisztikai eljárások alkalmazása az ellenőrzésben, különös tekintettel a pénzügyi ellenőrzésre. PhD értekezés, Budapesti Corvinus Egyetem.
- Nigrini, M.J.(2012). Benford’s Law: Applications for forensic accounting, auditing, and fraud detection, Wiley, 2012
- Rose, AM and Rose, JM (2003). Turn Excel into a financial sleuth: an easy-to-use digital analysis tool can red-flag irregularities. Journal of Accountancy 196(2), 58-60.
- Sambridge, M, Tkalčić, H and Jackson, A (2010). Benford's law in the Natural Sciences. Geophysical Research Letters
- Smith, SW (1997). Explaining Benford's Law. Chapter 34 in: The Scientist and Engineer's Guide to Digital Signal Processing.