Treball final de grau

GRAU DE MATEMÀTIQUES

Facultat de Matemàtiques
Universitat de Barcelona

Llei de Benford

Autor: Wei Huang

Director:	**Dr. Josep Fortiana Gregori**
Realitzat a:	**Facultat de matemàtiques (Departament de matemàtiques i informàtica)**

Barcelona, **27 de juny de 2016**

Abstract

This work is about *Benford's Law* (also know as first digit law) that asserts that, in some situations, the fraction of numbers that start with the digit d is not the intuitively –and yet reasonable– $1/9$ but the remarkable $\log_{10}(1 + d^{-1})$. We also study, in a generalized way, the behaviour of the others digits and we will see how certains sequences (Fibonacci's numbers, powers, etc) follows almost perfectly the values predicted by the law. Finally we will discuss daily situations that also follows the Benford's Law (lists populations, payments, etc).

Resum

Aquest treball consisteix en estudiar la *llei de Benford* (també coneguda com a llei del primer dígit) que ens assegura que, en determinades situacions, la proporció de nombres que comencen amb el dígit d, no és el valor intuïtiu –i raonable– $1/9$ sinó que segueix un valor més remarcable: $\log_{10}(1 + d^{-1})$. També estudiarem, de manera més generalitzada, el comportament del successius dígits i observarem com algunes successions (nombres de Fibonacci, potències, etc) s'ajusten perfectament als valors predits per la llei. Finalment, farem una anàlisi de certes situacions de la vida quotidiana on les dades segueixen la llei de Benford (dades del cens, factures, etc).

Agraïments

M'agradaria expressar el meu sincer agraïment per a totes aquelles persones que m'han ajudat en la realització d'aquest treball. Especialment vull agrair al Sr. Berenguer Sabadell per ajudar-me a revisar el treball i donar-me molts suggeriments. També vull donar moltes gràcies al Dr. Josep Fortiana per tutoritzar el meu projecte, des de guiar-me en la forma de fer la investigació del meu treball fins a com fer els programes i aprendre a redactar un treball d'aquest nivell. Un agraïment de tot cor per la comprenssió i suports morals rebuts de la meva família i amics. Gràcies a ells he tingut forces per continuar i acabar-lo, malgrat les dificultats que han aparegut al llarg del projecte.

Índex

1 Introducció

El projecte

La llei de Benford bàsicament diu que els dígits de determinades llistes de dades no es distribuexien d'una manera uniforme, com normalment s'esperaria. A banda de la llei del primer dígit, que és la més coneguda, els altres dígits també presenten un comportament especial. A mesura que la posició dels dígits avança cap a la dreta, els dígits sí que tendeixen a distribuir-se de manera uniforme.

Introduïré els conceptes bàsics i les propietats més importants, la majoría trets del llibre de Berger i Hill [1]. La llei de Benford és l'única llei que té la particularitat que la distribució dels dígits és invariant per canvi d'escala i per canvi de base. És a dir, si una llista de dades segueix la llei, després de multiplicar tot el conjunt per un factor, la nova llista generada continua seguint la llei. No és un fet aïllat, sinó que molts fenòmens naturals i econòmics segueixen aquesta llei. En aquest treball poso un exemple de la població d'Espanya de l'any 2015 i veurem que segueix la llei. I uns exemples aplicats a la detecció dels fraus empresarials. Atès que no tinc disponibilitat per treure dades corporatives i empresarials, he posat un dels exemples més típics del llibre de Nigrini [2], amb els tests de bondat d'ajust que he anat citant al treball. És també cert que no totes les llistes de dades segueixen la llei, com ara, per exemple, el números d'identitat, codis postals o, un dels casos més curiosos, el nombres primers[1], etc.

L'anàlisi de les dades l'he fet amb l'ajut del programa `R`, que té dos paquets de funcions especialitzades en l'estudi de la llei de Benford: el paquet `BenfordTests` i el paquet `benford.analysis` . Aquests paquets contenen funcions que permeten calcular els dígits d'un conjunt de dades, representar-los gràficament i comparar les dades reals amb les que hauríem d'obtenir segons llei de Benford. El paquet `BenfordTests` es concentra més en els diferents tests per valorar la conformitat, com el test de la χ^2, test de distància euclidiana, etc. El paquet `benford.analysis` té implementades unes bases de dades extretes dels exemples del llibre de Nigrini [2]. Als annexos del treball hi he afegit els codis de programació en `R` utilitzats.

[1] Ja ho va dir Leonhard Euler (1707-1783)

> *"Els matemàtics han intentat, en va fins avui, descobrir algun ordre en la successió dels nombres primers. I tenim raons per creure aquest és un misteri on la ment humana no hi podrà penetrar mai."*

Estructura de la Memòria

Aquest memòria s'estructura de la manera següent:

- Comencem fent una introducció del tema, contextualitzant-lo històricament i definim els conceptes bàsics per poder entendre els fonaments de la llei.
- En una segona part estudiem les propietats més importants que caracteritzen la llei de Benford.
- A la tercera part analitzem els test estadístics que permeten plantejar la conformitat, o no, de si un conjunt de dades s'ajusta a la llei de Benford. En aquesta part del treball no només analitzem la llei del primer dígit, sinó que fem una anàlisi de la llei més generalitzada (test de sumació, test de mantissa, etc).
- A continuació es presenten un seguit d'exemples i contraexemples per visualitzar millor els aspectes teòrics estudiats als apartats anteriors i apliquem tots els conceptes apresos per analitzar unes situacions de la vida quotidiana.
- Finalment exposem unes conclusions sobre la validesa i les limitacions de la llei i plantegem possibles ampliacions del treball.

En aquesta memòria utilitzarem les notacions següents:

1. $\log(x)$ per expressar $\log_{10}(x)$, $\ln(x)$ per expressar el logaritme neperià, o de base e, i $\log_b(x)$ per expressar el logaritme de base b qualsevol ($b > 0$).
2. La funció $t \to \lfloor t \rfloor$ és la funció part entera i representa l'enter immediatament inferior a t.
3. La funció $\langle t \rangle = t - \lfloor t \rfloor$ representa la part fraccionària de t.

2 Context històric

Els nombres formen una part important de la nostra vida. Des que ens aixequem fins que anem a dormir, la nostra vida està envoltada de números: rellotges, números de telèfon, números de les cases del carrer, els preus, les temperatures, els codis, etc.

És possible que aquests nombres segueixen un patró? Podríem usar aquests patrons per determinar si una dada és real o ha estat manipulada? Les respostes a les preguntes anteriors comencen amb Simon Newcomb.

El fenomen de certes regularitats en taules numèriques va ser documentat per primera vegada per l'astrònom i matemàtic Simon Newcomb (1835-1909) a l'any 1881 en un article a l'*American Journal of Mathematics*. Newcomb va observar que les primeres pàgines de les taules de logaritmes estaven més gastades que les darreres. Es pot deduir que això es deu a la presència més freqüent de dígits petits. En aquella època sense calculadora, la gent usava taules de logaritmes molt sovint per transformar els productes i els quocients en sumes i restes de logaritmes per després fer l'exponencial al resultat obtingut i així obtenir el producte, o el quocient, dels dos números originals[2].

L'any 1938, el físic Frank Benford va observar exactament el mateix fenomen sense conèixer prèviament l'article de Newcomb. Va proposar la mateixa llei que havia obtingut Newcomb i, a més a més, va recollir 20 col·leccions de dades de diferents àmbits per comprovar la llei, reunint un total de 20 229 registres, com ara l'àrea de rius, els pesos de elements atòmics, els nombres que apareixen en un número de la revista *Reader's Digest*, etc (veure la taula 1). La probabilitat mitjana del primer dígit de aquestes col·leccions s'apropen bastant bé a la llei del primer dígit. L'article de Benford va tenir més ressò que no pas l'article de Newcomb i així va ser com aquesta llei logarítmica es va conèixer a partir d'aleshores la *llei de Benford*, o *llei del primer dígit*.

Frank Albert Benford va néixer l'any 1883 a Johnstown (Pennsiylvania, USA). Estudià enginyeria electrònica i física. Després de graduar-se de la llicenciatura d'enginyeria elèctrica a la Universitat de Michigan l'any 1910, va treballar per a la *General Electric Company* a Schenectady (New York, USA) fins que es jubilà el 1948. Benford és conegut per haver inventat, el 1937 un instrument per mesurar l'índex de refracció del vidre. Com a expert en mesures òptiques, va divulgar 109 articles en els camps de l'òptica i les matemàtiques i va obtenir 20 patents d'instruments òptics. Va morir a Schenectady el 1948.

La Llei de Benford és una distribució de probabilitat que descriu el comportament

[2] Com molt bé va dir Pierre-Simon de Laplace (1749-1827)

"L'aparició dels logaritmes va doblar automàticament la vida dels astrònoms"

El procés és el següent:

$$x \cdot y = b^{(\log_b xy)} = b^{\log_b x + \log_b y}$$

Taula 1: La taula del primer dígit estudiada originalment per Benford.

Description	Count	1	2	3	4	5	6	7	8	9
Rivers, Area	335	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1
Population	3259	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2
Constants	104	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6
Newspapers	100	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0
Spec. Heat	1389	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1
Pressure	703	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7
H. P. Lost	690	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6
Mol. Wgt.	1800	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2
Drainage	159	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9
Atomic Wgt.	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	560	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6
Digest	308	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2
Cost Data	741	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1
X-Ray Volts	707	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8
Am. League	1458	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0
Black Body	1165	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4
Addresses	312	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0
n1, n2,..., n!	900	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5
Death Rate	418	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1
Average	1011	30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7
Probable Error		±0.8	±0.4	±0.4	±0.3	±0.2	±0.2	±0.2	±0.2	±0.3

dels dígits en molts conjunts de dades que es troben a la vida real. Els dígits no estan uniformement distribuïts, com hom podria esperar. Aproximadament, el 30% de les dades que poden aparèixer en una taula tenen com a primer dígit un 1, un tant per cent més elevat de l'esperat 11.1% si tots els dígits tinguessin la mateixa probabilitat d'aparèixer com a primer dígit. Les freqüències d'aparició com a primer dígit d'un nombre són decreixents fins a arribar al 4.6% de probabilitat que té el dígit 9. De la llei de Benford se'n deriva que com més gran sigui un dígit, la probabilitat de que aquest dígit sigui el primer dígit d'un determinat nombre, es va fent més petita. Tanmateix hem d'anar amb compte perquè no totes les taules de dades numèriques segueixen la llei de Benford (com, per exemple, els codis postals)

Un ampli ventall de taules s'ajusten admirablement bé a la llei de Benford, com ara: poblacions, longituds de rius, àrees de llacs, constants matemàtiques i físiques, resultats electorals, factures i un llarg etcètera. La llei de Benford, no explica només la distribució del primer dígit. També explica la distribució dels segon dígit, del tercer i de tots els següents, així com la combinació dels dos primers dígits, etc.

Si comptem de l'1 fins al 9, cada xifra surt amb la mateixa probabilitat (1/9=11.1%). Però del 10 al 19 només tenim com a primera xifra l'1, ja deixa de ser uniformement distribuït i no ho tornarà a ser fins arribar al 99. De fet, en una successió $1, 2, 3, \ldots, N$, el primer dígit no es distribueix uniformement excepte si el valor màxim és de la forma $10^n - 1$. Per exemple, els mesos de l'any. Un terç dels mesos (numèrics) tenen com a primer dígit un 1. O el dia del mes, on clarament els dígits 1 i 2 apareixen més freqüent com a primera xifra respecte els altres dígits. Finalment, una altra propietat sorprenent de la llei és la invariança respecte a canvis d'escala. Si la grandària de l'univers fos el doble que l'actual, totes les quantitats físiques que ara comencen per 1, passarien a començar per 2 o per 3. Aquelles que comencen per 2, passarien a fer-ho per 4 o per 5. I així successivament. I totes aquelles quantitats que ara comencen per 5, 6, 7, 8 i 9 passarien a començar per 1!

Actualment utilitzem la llei de Benford en molts àmbits. Fins i tot per detectar situacions fraudulentes. Per exemple, els experts comptables poden detectar errors en les factures emeses si les quantitats considerades no segueixen el patró que de manera natural haurien de fer segons la llei de Benford.

Aquest treball de final de grau estudia els aspectes matemàtics més rellevants de la llei de Benford i presenta diverses situacions on es segueix, o no, aquesta llei.

3 Definicions bàsiques

3.1 Els dígits significatius i el significand

Per estudiar una llei del comportament dels primers dígits cal que introduir primer com definim els dígits d'un nombre. El primer dígit és el dígit diferent de 0 més a l'esquerra en un nombre, per exemple el primer dígit de 13.43 és 1, el primer dígit de 0.34 és 3. Per tant, el primer dígit hi ha 9 possibilitats, de l'1 al 9. A partir del segon dígit ja hi ha 10 possibilitats: $0, 1, \ldots, 9$. En tota aquesta secció tractem els nombres negatius a través dels seu valor absolut, i donem definicions bàsiques per tot el que segueix, i ens basem, sobretot, en l'article de Berger i Hill ([1]).

Definició 3.1.1. *Per tot real no nul x, existeixen $k \in \mathbb{Z}$ i $j \in \{1, 2 \cdots, 9\}$ únics tals que*

$$10^k j \leq |x| < 10^k (j+1).$$

Diem que el primer dígit significatiu de x és $D_1(x) = j$.
Per $m \geq 2$, $m \in \mathbb{N}$, existeixen $k \in \mathbb{Z}$ i $j \in \{0, 1, 2 \cdots, 9\}$, únics tals que

$$10^k \left(\sum_{i=1}^{m-1} D_i(x) 10^{m-i} + j \right) \leq |x| < 10^k \left(\sum_{i=1}^{m-1} D_i(x) 10^{m-i} + j + 1 \right)$$

Diem que el m-èsim dígit significatiu de x és $D_m(x) = j$. Per conveniència, $D_m(0) := 0$ $\forall m$.

Exemple 3.1.1.

1. $x = 23$, $10^1 \cdot 2 \leq |23| < 10^1(2+1) \Rightarrow 20 \leq 23 < 30$.
2. $D_1(\pi) = D_1(10\pi) = D_1(-\pi) = 3$, $D_2(\sqrt{3}) = 7$, $D_3(100) = 0$, $D_4(5.678) = 8$, $D_2(\frac{1}{2}) = 5$, $D_2(3) = 0$.

Definició 3.1.2. *La funció significand (decimal) $S : \mathbb{R} \to [1, 10)$ es defineix com: Si $x \neq 0$ aleshores $S(x) = t$, on $t \in [1, 10)$ és l'únic real que compleix $|x| = 10^k t$ per un únic $k \in \mathbb{Z}$. Per conveniència, $S(0) := 0$.*

El significand de x permet conèixer totes les xifres significatives de x en base 10, però no diu res sobre l'ordre de magnitud de x.

Observació 3.1.1.

1. $\forall x \in \mathbb{R}$, $S(10^k x) = S(x)$ $\forall k \in \mathbb{Z}$.
2. $S(S(x)) = S(x)$.
3. $S(x) = 10^{\log|x| - \lfloor \log|x| \rfloor}$ $\forall x \neq 0$.

Exemple 3.1.2.

1. $S(0.314) = S(-3.14) = S(3.14 \cdot 10^k) = 3.14$, on $k \in \mathbb{Z}$.

2. $S(314) = 10^{\log|314| - \lfloor \log|314| \rfloor} = 10^{\log|314|-2} = \frac{314}{100} = 3.14$.

Proposició 3.1.1. *Per tot $x \in \mathbb{R}$:*

1. $S(x) = \sum_{m\in\mathbb{N}} 10^{1-m} D_m(x)$
2. $D_m(x) = \lfloor 10^{m-1} S(x) \rfloor - 10 \lfloor 10^{m-2} S(x) \rfloor$, $\forall m \in \mathbb{N}$

Demostració. Veure [1] p.7. □

Exemple 3.1.3.

1. $S(\pi) = D_1(\pi) + 10^{-1}(\pi) + 10^{-2}(\pi) + \ldots = 3.14159\ldots = \pi$
2. $D_2(\sqrt{2}) = \left\lfloor 10^{2-1} S(\sqrt{2}) \right\rfloor - 10 \left\lfloor 10^{2-2} S(\sqrt{2}) \right\rfloor = 12 - 10 = 2$

Definició 3.1.3 (Nigrini(2012),p.10)**.** *La* mantissa *d'un nombre real positiu x és la part fraccionària del logaritme d'aquest nombre.*

$$\text{Mantissa}(x) = \langle \log x \rangle \in [0, 1)$$

La definició de la mantissa en el llibre de Berger-Hill([1]) és $\log S$. Observem que $\log S(x) = \text{Mantissa}(x)$.

Exemple 3.1.4.

1. $\text{Mantissa}(1) = \langle \log 1 \rangle = 0$
2. $\text{Mantissa}(10) = \text{Mantissa}(10^k) = 0$
3. $\text{Mantissa}(50) = \langle \log 50 \rangle = 0.6989$
4. $\text{Mantissa}(50) = \log S(50) = \log(5) = 0.6989$

3.2 La σ-àlgebra significand

Per formular d'una manera matemàticament precisa, les definicions de la secció anterior necessitem definir d'una manera més rigurosa certs conceptes de la teoria de probabilitats. Comencem amb la definició següent.

Definició 3.2.1. *Sigui Ω un conjunt no buit, i sigui $\mathcal{A}$ una col·lecció de subconjunts de Ω, diem que $\mathcal{A}$ és una σ-àlgebra si es compleix:*

1. *El conjunt buit $\emptyset \in \mathcal{A}$*
2. *$A \in \mathcal{A} \Rightarrow A^c := \{\omega \in \Omega : \omega \notin A\} \in \mathcal{A}$*
3. *$A_n \in \mathcal{A}\ \forall n \in \mathbb{N} \Rightarrow \bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$.*

Definició 3.2.2. *La σ-àlgebra significand $\mathcal{S}$ és la σ-àlgebra en $\mathbb{R}$ generada pel la funció significand S, és a dir, $\mathcal{S} = \mathbb{R}^+ \cap \sigma(S)$.*

La σ-àlgebra significand $\mathcal{S}$ compleix els resultats següents:

Teorema 3.2.1. *Sigui $A \in \mathcal{S}$, aleshores*

$$A = \bigcup_{k\in\mathbb{Z}} 10^k S(A)$$

on $S(A) = \{S(x) : x \in A\} \subset [1, 10)$. A més a més,

$$\mathcal{S} = \mathbb{R}^+ \cap \sigma(D_1, D_2, D_3, \ldots) = \left\{ \bigcup_{k\in\mathbb{Z}} 10^k B : B \in \mathcal{B}[1, 10) \right\}$$

Demostració. Veure [1] p.11. □

Lema 3.2.1. *Sigui $\mathcal{S}$ una σ-àlgebra significand. Aleshores:*

1. *$10^k A = A$ per tot $A \in \mathcal{S}$ i $k \in \mathbb{Z}$.*
2. *$\alpha A \in \mathcal{S}$ per tot $A \in \mathcal{S}$ i $\alpha > 0$.*
3. *$A^{\frac{1}{n}} \in \mathcal{S}$ per tot $A \in \mathcal{S}$ i $n \in \mathbb{N}$.*

Demostració. Veure [1] p.13. □

Exemple 3.2.1.

1. Un conjunt de nombres positius:

   $$A = \left\{10^k, k \in \mathbb{Z}\right\} = \{\ldots, 0.01, 0.1, 1, 10, 100, \ldots\}$$

   pertany a $\mathcal{S}$ atès que:

   $$A = \{x > 0 : D_1(x) = 1, D_m = 0 \quad \forall m \geq 2\} \quad \text{o} \quad A = \bigcup_{k\in\mathbb{Z}} \{1\}$$

   on $\{1\} \in \mathcal{B}[1, 10)$.

2. Un conjunt de nombres positius amb el primer dígit significatiu 2:

   $$A = \{x > 0 : D_1(x) = 2\} = \{x > 0 : 2 \leq S(x) < 3\} = \bigcup_{k\in\mathbb{Z}} 10^k [2, 3)$$

   aleshores:

   $$3A = \{x > 0 : D_1(x) \in \{6, 9\}\} = \{x > 0 : 6 \leq S(x) < 9\} = \bigcup_{k\in\mathbb{Z}} 10^k [6, 9) \in \mathcal{S}$$

   $$A^{1/2} = \left\{x > 0 : S(x) \in \left[\sqrt{2}, \sqrt{3}\right) \cup \left[\sqrt{20}, \sqrt{30}\right)\right\} \in \mathcal{S}$$

   però

   $$A^2 = \bigcup_{k\in\mathbb{Z}} 10^{2k}[4, 9) \notin \mathcal{S}$$

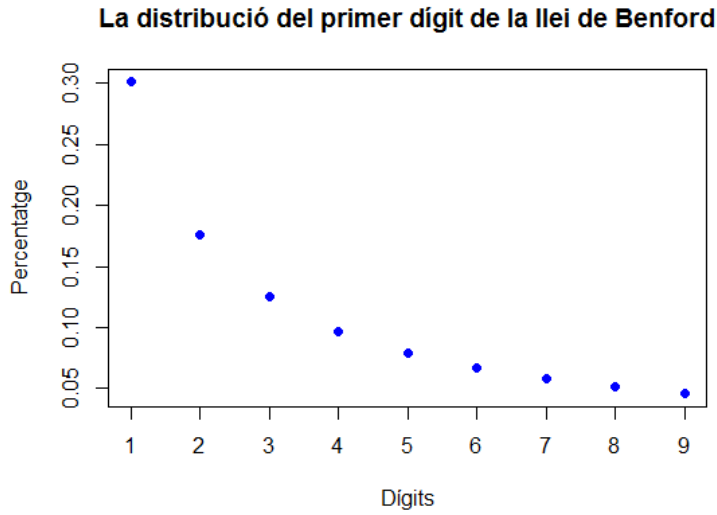
   ja que $[40, 90)$ no està inclòs a A^2.

3.3 Llei de Benford

Passem ara a definir formalment el que es coneix com la llei de Benford, que explica de maneara precisa la distribució del significand.

Definició 3.3.1. *La llei de Benford o llei del primer dígit queda definida per la funció de massa de probabilitat:*

$$\text{Prob}(d_1) = \log\left(1 + d_1^{-1}\right), \qquad d_1 \in \{1, 2, \ldots, 9\}\,.$$



Observació 3.3.1. $\text{Prob}(d_1 \leq d) = \sum_{d_1=1}^{d} \log\left(1 + d_1^{-1}\right) = \log(d+1)$, $d \in \{1, 2, \ldots, 9\}$.

La llei de Benford no descriu només la distribució del primer dígit significatiu, passem a veure la llei de Benford general per a tots els dígits significatius.

Definició 3.3.2. *Més en general, la llei general dels dígits significatius(en base 10) és, per tot $m \in \mathbb{N}$, per $d_1 \in \{1, 2, \ldots, 9\}$, i per $d_j \in \{0, 1, 2, \cdots, 9\}$, $j = 2, \ldots, m$, la probabilitat de que els dígits significatius són $d_1, d_2, \ldots, d_j$ és:*

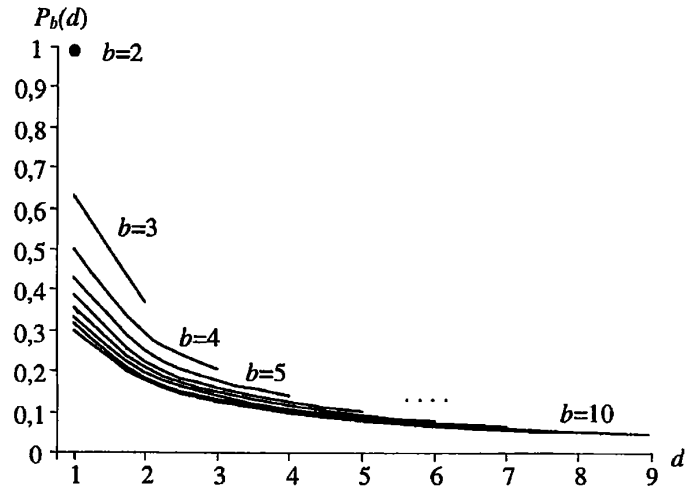
$$Prob\left((D_1, D_2, \ldots, D_m) = (d_1, d_2, \ldots, d_m)\right) = \log\left(1 + \left(\sum_{j=1}^{m} 10^{m-j} d_j\right)^{-1}\right)$$

La llei de Benford es vàlid per qualsevol base.

Definició 3.3.3. *Es pot estendre més, la llei general dels dígits significatius(en base b), per tot $m \in \mathbb{N}$, per $d_1 \in \{1, 2, \ldots, b-1\}$, i per $d_j \in \{0, 1, 2, \cdots, b-1\}$, $j = 2, \ldots, m$, la probabilitat de que els dígits significatius són $d_1, d_2, \ldots, d_j$ és:*

$$Prob\left(\left(D_1^{(b)}, D_2^{(b)}, \ldots, D_m^{(b)}\right) = (d_1, d_2, \ldots, d_m)\right) = \log_b\left(1 + \left(\sum_{j=1}^{m} b^{m-j} d_j\right)^{-1}\right)$$

Figura 1: La probabilitat del primer dígit en diferents bases



A la figura 1 estudiant el valor de probabilitat del primer dígit per a diverses bases s'observa clarament el decreixement a mesura que augmenta d.

Lema 3.3.1. *La definició 3.3.1 es pot generalitzar d'una manera contínua per la distribució del significand:*

$$\text{Prob}(S \leq t) = \log(t), \qquad \textit{on } t \in [1, 10).$$

Demostració. Primer mirem el cas quan t només té un dígit significatiu d_1:

$$\text{Prob}(S \leq d_1) = \begin{cases} 0 & \text{si } d_1 \leq 1 (\text{Prob}(S = 1) = 0), \\ \text{Prob}(D_1 \leq d_1 - 1) = \log(d_1) & \text{si } 1 < d_1 \leq 10. \end{cases}$$

Ara suposem que t té $m > 1$ dígit, és a dir, $t = \sum_{j=1}^{m} 10^{m-j} d_j$, on $d_1 > 1$ i $d_j > 0$,

$j = 2, \ldots, m$.

$$\begin{aligned}
\text{Prob}(S \le t) &= \text{Prob}(D_1 \le d_1 - 1) + \text{Prob}(D_1 = d_1, D_2 \le d_2 - 1) + \cdots \\
&\quad + \text{Prob}(D_1 = d_1, D_2 = d_2, \ldots, D_{m-1} = d_{m-1}, D_m \le d_m - 1) \\
&= \text{Prob}(D_1 \le d_1 - 1) + \sum_{d_2=0}^{d_2-1} \text{Prob}(D_1 = d_1, D_2 = d_2) + \cdots \\
&\quad + \sum_{d_m=0}^{d_m-1} \text{Prob}(D_1 = d_1, D_2 = d_2 - 1, \ldots, D_m = d_m) \\
&= \log(d_1) + \sum_{d_2=0}^{d_2-1} \log\left(1 + \frac{1}{10d_1 + d_2}\right) + \ldots + \sum_{d_m=0}^{d_m-1} \log\left(1 + \left(\sum_{j=1}^{m-1} 10^{m-j} d_j + d_m\right)^{-1}\right) \\
&= \log\left(\frac{\sum_{j=1}^{m} 10^{m-j} d_j}{10^m - 1}\right) \\
&= \log\left(\sum_{j=1}^{m} 10^{-j-1} d_j\right) = \log(t)
\end{aligned}$$

□

Observem que la proporció de nombres x tals que $S(x) \in [a, b)$, on $1 \le a < b \le 10$ és:

$$\text{Prob}(S(x) \in [a, b)) = \log b - \log a.$$

En particular, la probabilitat de d_2 com a segon dígit significatiu és:

$$\text{Prob}(D_2 = d_2) = \sum_{j=1}^{9} \log\left(1 + \frac{1}{10j + 1}\right), \qquad d_2 \in \{0, 1, 2, \ldots, 9\}$$

La probabilitat de $d_1 d_2$ com els dos primers dígits significatius és:

$$\text{Prob}((D_1, D_2) = (d_1, d_2)) = \log\left(1 + \frac{1}{d_1 + 10d_2}\right)$$

on $d_1 \in \{1, \ldots, 9\}$ i $d_2 \in \{0, 1, 2, \ldots, 9\}$.

La llei contínua en [1,10) que discretitza la llei del primer dígit és: $\frac{\mathbb{1}_{[1,10)}}{t \ln 10}$. En efecte,

$$\text{Prob}(D_1 = d_1) = \int_{d_1}^{d_1+1} \frac{1}{x \ln 10} dx = \frac{1}{\ln 10} \ln\left(\frac{d_1 + 1}{d_1}\right) = \log\left(1 + \frac{1}{d_1}\right)$$

on $d_1 \in \{1, 2, \ldots, 9\}$.
L'esperança de D_m és:

$$\text{E}(D_m) = \sum_{d_j=1}^{9} d_j \ \text{Prob}(D_m = d_j)$$

La variància de D_m és:

$$\mathrm{Var}(D_m) = \sum_{d_j=1}^{9} d_j^2 \ \mathrm{Prob}(D_m = d_j) - E(D_m)^2$$

.

Taula 2: Esperança i variància de D_j per j=1 a 7

k	E(D_j)	Var(D_j)
1	3.44023696712	6.0565126313757
2	4.18738970693	8.2537786232732
3	4.46776565097	8.2500943647286
4	4.49677537552	8.2500009523513
5	4.49967753636	8.2500000095245
6	4.49996775363	8.2500000000953
7	4.49999677536	8.2500000000014
8	4.49999967754	8.2500000000080
9	4.50000000000	8.2500000000000

Taula 3: Probabilitat dels D_j dígits, $j = 1, 2 \ldots, 9$

d_j	P(d_1)%	P(d_2)%	P(d_3)%	P(d_4)%	$\cdots$	P(d_8)%	P(d_9)%
0		11.968	10.178	10.018	$\cdots$	10.000002	10
1	30.103	11.389	10.138	10.014	$\cdots$	10.000001	10
2	17.609	10.882	10.097	10.010	$\cdots$	10.000001	10
3	12.494	10.433	10.057	10.006	$\cdots$	10.000001	10
4	9.691	10.031	10.018	10.002	$\cdots$	10.000000	10
5	7.918	9.668	9.979	9.998	$\cdots$	10.000000	10
6	6.695	9.337	9.940	9.994	$\cdots$	9.999999	10
7	5.799	9.035	9.902	9.990	$\cdots$	9.999999	10
8	5.115	8.757	9.864	9.986	$\cdots$	9.999999	10
9	4.575	8.500	9.827	9.982	$\cdots$	9.999998	10

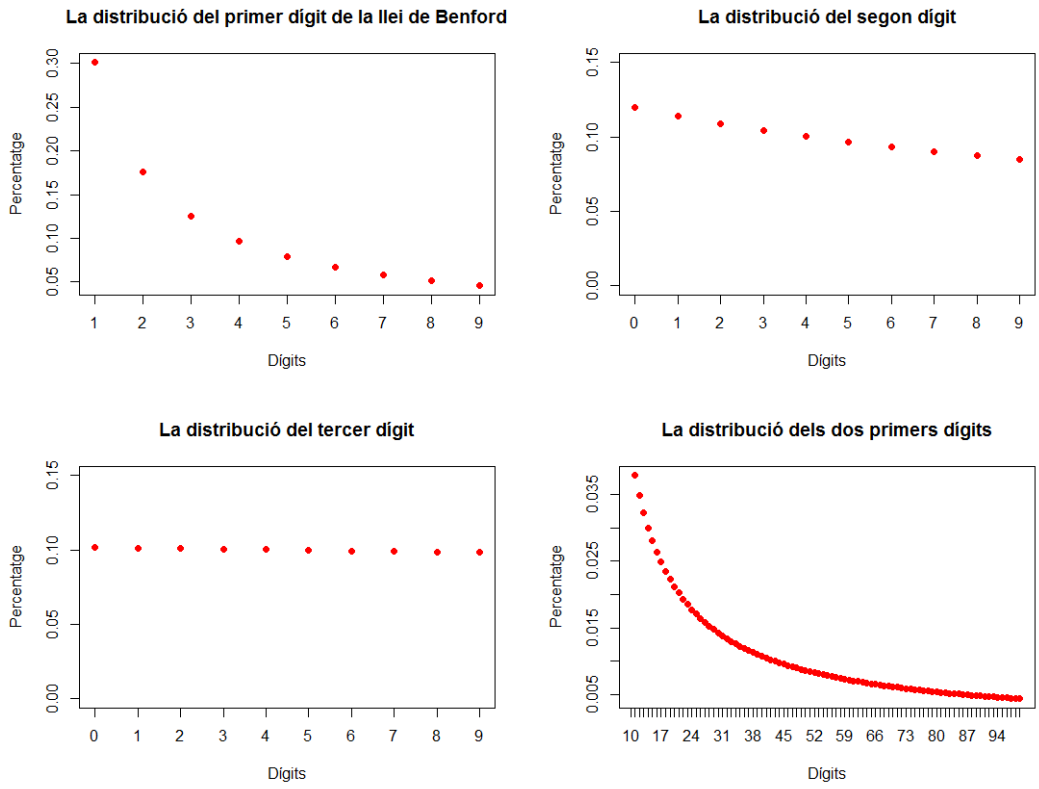
La taula 3 mostra que els dígits tendeixen a ser uniformement distribuïts a mesura que incrementem la posició del dígit. També es pot observar als gràfics de la figura 2 que cada cop és menys esbiaixats respecte la posició l'anterior, i més lineal. A partir del quart dígit, la diferència ja és quasi no apreciable. La distribució dels dos primers dígits es decreixen de 0.0378 a 0.00436.

El coeficient de correlació entre D_i i D_j és:

$$\rho_{D_i D_j} = \frac{\mathrm{Cov}(D_i, D_j)}{\sqrt{\mathrm{Var}(D_i)\mathrm{Var}(D_j)}}, \qquad \text{per} \quad 0 < i < j$$

A la taula 4 podem observar que la dependència entre dígits significatius disminueix a mesura que la distància $(j - i)$ augmenta.

Figura 2: Gràfiques de les distribucions de la posició dels dígits (en percentatges).



Taula 4: Coeficient de correlació entre D_i i D_j

i, j	1	2	3	4
1	0.0560563	0.0059126	0.0005916	0.0000591
2		0.0020566	0.0002059	0.0000205
3			0.0000228	0.0000022
4				0.0000002

Exemple 3.3.1.

1. $\text{Prob}(D_1 = 1) = \log 2 = 0.3010, \text{Prob}(D_1 = 9) = \log 9 = 0.046$
2. $\text{Prob}(D_1 = 1, D_2 = 5, D_3 = 7) = \log\left(1 + \frac{1}{157}\right) = 0.0028$
3. $\text{Prob}(D_1 = \text{Parell}) = \text{Prob}(D_1 \in \{2, 4, 6, 8\}) = 0.391.$
4. $\text{Prob}(D_1 = \text{Senar}) = \text{Prob}(D_1 \in \{1, 3, 5, 7, 9\}) = 0.609.$
5. $\text{Prob}(D_2 = 1 | D_1 = 1) = \dfrac{\log 12 - \log 11}{\log 2} = 0.1255 > \text{Prob}(D_2 = 1) = 0.1138$

Observació 3.3.2.

1. $\displaystyle\sum_{d_1=1}^{9} \log\left(1+\frac{1}{d_1}\right) = \log\left(\prod_{d_1=1}^{9}\left(\frac{1+d_1}{d_1}\right)\right) = 1$

2. $\text{Prob}(D_1 = 2) + \text{Prob}(D_1 = 3) = \log 2 = \text{Prob}(D_1 = 1)$.

3. $\text{Prob}(D_1 = 2) = \text{Prob}(D_1 = 4) + \text{Prob}(D_1 = 5)$.

4. $\text{Prob}(D_1 = 3) = \text{Prob}(D_1 = 6) + \text{Prob}(D_1 = 7)$.

5. $\text{Prob}(D_1 = 4) = \text{Prob}(D_1 = 8) + \text{Prob}(D_1 = 9)$.

L'observació 1 afirma que la suma de les probabilitats del primer dígit és 1, com s'ha d'esperar. L'observació 2 diu la probabilitat del primer dígit sigui 1 és igual a la suma de la probabilitat del primer dígit sigui 2 i 3. Podem pensar que després de multiplicar els nombres que tenen el primer dígit 1 per 2, el primer dígit serà 2 o 3. La idea de la resta és igual.

Una successió (x_n) és Benford si, es tria un nombre a l'atzar dels primers N elements de (x_n), la probabilitat que el primer dígit d'aquest nombre sigui d_1 s'aproxima a $\log(1+d_1^{-1})$ a mesura que $N \to \infty$, per tot $d_1 \in \{1, 2, \ldots, 9\}$, i de manera similar per als altres dígits signicatius.

Definició 3.3.4. *Una successió de nombres reals (x_n) és una successió de Benford (abreujament és Benford) si*

$$\lim_{N\to\infty} \frac{\#\{1 \le n \le N : S(x_n) \le t\}}{N} = \log t \quad \forall t \in [1, 10).$$

O equivalentment, per tot $m \in \mathbb{N}$, $d_1 \in \{1, 2, \ldots, 9\}$ i $d_j \in \{0, 1, 2, \cdots, 9\}$, $j \ge 2$,

$$\lim_{N\to\infty} \frac{\#\{1 \le n \le N : D_j(x_n) = d_j, \forall j = 1, 2, \ldots, m\}}{N} = \log\left(1 + \left(\sum_{j=1}^{m} 10^{m-j} d_j\right)^{-1}\right),$$

Una funció $f : [0, +\infty) \to \mathbb{R}$ és Benford si, es trien τ a l'atzar a l'interval $[0, T)$, la probabilitat que el primer dígit de $f(\tau)$ sigui d_1 s'aproxima a $\log(1 + d_1^{-1})$ a mesura que $T \to \infty$, per tot $d_1 \in \{1, 2, \ldots, 9\}$, i de manera similar per als altres dígits signicatius.

Definició 3.3.5. *Una funció $f : [0, +\infty) \to \mathbb{R}$ és Benford si:*

$$\lim_{T\to\infty} \frac{\lambda\left(\{\tau \in [0, T) : S(f(\tau)) \le t\}\right)}{N} = \log t \quad \forall t \in [1, 10).$$

O equivalentment, per tot $m \in \mathbb{N}$, $d_1 \in \{1, 2, \ldots, 9\}$ i $d_j \in \{0, 1, 2, \cdots, 9\}$, $j \ge 2$,

$$\lim_{T\to\infty} \frac{\lambda\left(\{\tau \in [0, T) : D_j(f(\tau)) = d_j, \forall j = 1, 2, \ldots, m\}\right)}{N} = \log\left(1 + \left(\sum_{j=1}^{m} 10^{m-j} d_j\right)^{-1}\right).$$

Definició 3.3.6. *Una mesura de probabilitat Borel en $\mathbb{R}$ és Benford si:*

$$P(\{x \in \mathbb{R} : S(x) \leq t\}) = P(S^{-1}(\{0\} \cup [1,t])) = \log t \quad \forall t \in [1,10).$$

Una variable aleatòria X en un espai de probabilitat $(\Omega, \mathcal{A}, \mathbb{P})$ és Benford si P_X és Benford, és a dir, si

$$\mathbb{P}(S(X) \leq t) = P_X(\{x \in \mathbb{R} : S(x) \leq t\}) = \log t \quad \forall t \in [1,10).$$

o equivalentment, per a tot $m \in \mathbb{N}$, $d_1 \in \{1,2,\ldots,9\}$ i $d_j \in \{0,1,2,\cdots,9\}$, $j \geq 2$,

$$\mathbb{P}(D_j(X) = d_j, j = 1,2,\ldots,m) = \log\left(1 + \left(\sum_{j=1}^{m} 10^{m-j} d_j\right)^{-1}\right)$$

Definició 3.3.7. *La distribució de Benford $\mathbb{B}$ és l'única mesura de probabilitat en $(\mathbb{R}^+, \mathcal{S})$ que compleix:*

$$\mathbb{B}(S \leq t) = \mathbb{B}\left(\bigcup_{k \in \mathbb{Z}} 10^k [1,t]\right) = \log t \quad \forall t \in [1,10),$$

o equivalentment, per a tot $m \in \mathbb{N}$, $d_1 \in \{1,2,\ldots,9\}$ i $d_j \in \{0,1,2,\cdots,9\}$, $j \geq 2$,

$$\mathbb{B}(D_j = d_j, j = 1,2,\ldots,m) = \log\left(1 + \left(\sum_{j=1}^{m} 10^{m-j} d_j\right)^{-1}\right)$$

La probabilitat $\mathbb{B}(\{1\})$ no és defineix simplement, perquè el conjunt 1 no es pot expressar com termes dels dígits significatius.

4 Caracteritzacions de la llei de Benford

En aquesta secció resumim quatre caracteritzacions de la llei de Benford en el context de successió, funcions, distribucions i variables aleatòries, en general sense les demostracions, per les quals ens referim a *Berger-Hill* (2011) i (2015).

4.1 Caracterització a través de la distribució uniforme

Definició 4.1.1. *Una successió de nombres reals (x_n) és uniformement distribuïda mòdul 1(u.d mod 1), si*

$$\lim_{N \to \infty} \frac{\#\{1 \leq n \leq N : \langle x_n \rangle \leq s\}}{N} = s \quad \forall s \in [0,1).$$

Una funció (mesurable Borel) $f : [0,+\infty) \to \mathbb{R}$ és u.d mod 1 si:

$$\lim_{T \to \infty} \frac{\lambda(\{\tau \in [0,T) : \langle f(\tau) \rangle \leq s\})}{N} = s \quad \forall s \in [0,1).$$

Una variable aleatòria X en un espai de probabilitat $(\Omega, \mathcal{A}, \mathbb{P})$ és u.d mod 1 si

$$\mathbb{P}(\langle X \rangle \leq s) = s \quad \forall s \in [0,1).$$

Una mesura de probabilitat P en $(\mathbb{R}, \mathcal{B})$ és u.d mod 1 si

$$P(\{x : \langle x \rangle \leq s\}) = P\left(\bigcup_{k\in\mathbb{Z}} [k, k+s]\right) = s \quad \forall s \in [0,1).$$

Teorema 4.1.1. *(Caracterització a través de la distribució uniforme) Una successió de nombres reals (una funció mesurable Borel, una variable aleatòria, una mesura de probabilitat Borel, respectivament) és Benford si i només si la successió dels logaritmes decimals dels valors absoluts dels seus termes es distribueix uniformement mòdul 1. És a dir, la mantissa es distribueix uniformement.*

Demostració. Sigui X una variable aleatòria, suposant $\mathbb{P}(X = 0) = 0$, aleshores per tot $s \in [0,1)$,

$$\mathbb{P}(\langle \log|X| \rangle \leq s) = \mathbb{P}(\log|X| \in \bigcup_{k\in\mathbb{Z}} [k, k+s]) = \mathbb{P}(|X| \in \bigcup_{k\in\mathbb{Z}} [10^k, 10^{k+s}]) = \mathbb{P}(S(X) \leq 10^s).$$

X és Benford si i només si $\mathbb{P}(S(x) \leq 10^s) = \log 10^s = s$ per tot $s \in [0,1)$, és a dir, si i només si $\log|X|$ és u.d mod 1. □

Lema 4.1.1. *1. Una successió (x_n) és u.d.mod 1 si i només si la seqüència $(kx_n + b)$ és u.d.mod 1 per tot $k \in \mathbb{Z} \setminus \{0\}$ i $b \in \mathbb{R}$. A més a més, (x_n) és u.d mod 1 si i només si (y_n) és u.d mod 1 quan $\lim_{n\to\infty} |y_n - x_n| = 0$.*

2. La funció f és u.d mod 1 si i només si $t \mapsto kf(t) + b$ és u.d mod 1 per a tot enter no nul k i $b \in \mathbb{R}$.

3. La variable aleatòria X és u.d mod 1 si i només si $kX + b$ és u.d mod 1 per a tot enter no nul k i $b \in \mathbb{R}$.

Corol·lari 4.1.1. *1. Una successió (x_n) és Benford si i només si (αx_n^k) és Benford per tot $\alpha \in \mathbb{R}$ i $k \in \mathbb{Z}$ amb $\alpha k \neq 0$.*

2. Una funció $f : [0, +\infty) \to \mathbb{R}$ és Benford si i només si $1/f$ és Benford.

3. Una variable aleatòria X és Benford si i només si $1/X$ és Benford.

Proposició 4.1.1. *Sigui (x_n) una successió de nombres reals:*

1. Si $\lim_{n\to\infty}(x_{n+1} - x_n) = \theta$ per algú θ irracional, aleshores (x_n) és u.d mod 1.

2. Si (x_n) és periòdic, és a dir $x_{n+p} = x_n$ per algú $p \in \mathbb{N}$ i per tot n, aleshores $(n\theta + x_n)$ és u.d mod 1 si i només si θ és irracional.

3. La seqüència (x_n) és n.d mod 1 si i només si $(x_n + \alpha \log_n)$ és u.d mod 1 per tot $\alpha \in \mathbb{R}$.

4. Si (x_n) és u.d mod 1 i no decreixent, aleshores $(x_n/\log n)$ no és fitat.

Teorema 4.1.2. *Sigui a, b, α, β nombres reals amb $\alpha \neq 0$ i $|\alpha| > |\beta|$, llavors $(\alpha^n a + \beta^n b)$ és Benford si i només si $\log|\alpha|$ és irracional.*

Demostració. Atès que $a \neq 0$ i $|\alpha| > |\beta|$, $\lim_{n\to\infty} \frac{\beta^n b}{\alpha^n a} = 0$, aleshores:

$$\log\left|\alpha^b a + \beta^n b\right| - \log|\alpha^n a| = \log\left|1 + \frac{\beta^n b}{\alpha^n a}\right| \to 0,$$

$\log\left|\alpha^b a + \beta^n b\right|$ és u.d.mod.1 si i només si $\log|\alpha^n a| = (\log|a| + n\log|\alpha|)$ ho és. Segons la proposició 4.1.1(i), això compleix, si $\lim_{n\to\infty}(\log|a| + (n+1)\log|\alpha| - \log|a| - n\log|\alpha|) = \log|\alpha|$ és irracional. Per tant, apliquem el teorema 4.1.1, obtenim $(\alpha^n a + \beta^n b)$ és Benford. □

Teorema 4.1.3. *Sigui X, Y variables aleatòries, aleshores:*

1. *Si X és u.d mod 1 i Y és independent de X, aleshores $X + Y$ és u.d mod 1.*
2. *Si $\langle X\rangle$ i $\langle X + \alpha\rangle$ tenen la matexia distribució per algú α irracional, aleshores X és u.d mod 1.*
3. *Si X_n és una seqüència de variables aleatòria i.i.d no és purament atòmic ($\mathbb{P}(X_1 \in C) < 1$ per tot conjunt numerable $C \subset \mathbb{R}$), aleshores:*

$$\lim_{n\to\infty} \mathbb{P}\left(\left\langle \sum_{j=1}^{n} X_j \right\rangle \leq s\right) = s \quad \textit{per tot} \quad 0 \leq s \leq 1,$$

Això és, $\left\langle \sum_{j=1}^{n} X_j \right\rangle \to U(0,1), n \to \infty$.

Proposició 4.1.2. *Sigui Y una variable aleatòria de Benford i X una variable aleatòria qualsevol amb la densitat contínua, X i Y són independents, aleshores XY és Benford.*

Proposició 4.1.3. *Sigui Y una variable aleatòria (no nul·la) de Benford i X una variable aleatòria (no nul·la) qualsevol amb la densitat contínua, aleshores X/Y i Y/X són Benfords.*

Teorema 4.1.4. *Sigui (X_n) una successió de variables aleatòries i.i.d no purament atòmic, és a dir, $\mathbb{P}(X_1 \in C) < 1$ per tot conjunt numerable $C \in \mathbb{R}$, aleshores:*

1. *$\left(\prod_{j=1}^{n} X_j\right)$ convergeix en distribució a la llei de Benford.*
2. *Amb la probabilitat 1, $\left(\prod_{j=1}^{n} X_j\right)$ és Benford.*

4.2 Caracterització a través de la invariància per canvi d'escala

Una propietat raonable a esperar de qualsevol llei que regeixi la distribució de primers (o successius) dígits és que sigui invariant front a un canvi d'escala. És a dir, si les àrees del rius segueixen la llei de Benford, hauria de ser irrellevant si les unitats expressen en kilòmetres o milles. Pinkham (1961) va proposar la demostració del fet que la llei de Benford és l'única distribució que és invariant per canvi d'escala. Tot i que aquesta conclusió de la invariància per canvi d'escala és intuïtivament correcta, la demostració conté un error. La suposició que hi ha una mesura de probabilitat invariant per canvi d'escala en $\mathbb{R}^+$ no és correcte(un tal mesura no pot ser finita), no existeix cap mesura de probabilitat compleix això. L'única variable aleatòria que és invariant per canvi d'escala és la variable aleatòria degenerada igual a la constant 0, és a dir, $\mathbb{P}(X = 0) = 1$. Però sí que poden tenir els dígits significatius invariants per canvi d'escala.

Definició 4.2.1. *Sigui $\mathcal{A} \supset \mathcal{S}$ una σ-àlgebra en $\mathbb{R}^+$, una mesura de probabilitat P en $(\mathbb{R}^+, \mathcal{A})$ té dígits significatius invariants per canvi d'escala si compleix:*

$$P(\alpha A) = P(A) \quad \forall \alpha > 0 \quad i \quad A \in \mathcal{S}$$

o equivalentment, per tot $m \in \mathbb{N}$, $d_1 \in \{1, 2, \ldots, 9\}$ i $d_j \in \{0, 1, 2, \cdots, 9\}$, $j \geq 2$,

$$P(\{x : D_j(\alpha x) = d_j, \forall j = 1, 2, \ldots, m\}) = P(\{x : D_j(x) = d_j, \forall j = 1, 2, \ldots, m\})$$

on $\alpha > 0$.

Anem a veure que la mesura de probabilitat de Benford $\mathbb{B}$ en $(\mathbb{R}^+, \mathcal{S})$ té dígits significatius invariants per canvi d'escala. Sigui $A = \bigcup_{k\in\mathbb{Z}} 10^k[a, b]$ amb $1 \leq a < b < 10$, un $\alpha > 0$ qualsevol real,

$$\alpha A = \bigcup_{k\in\mathbb{Z}} 10^{k+\log\alpha}[a, b] = \bigcup_{k\in\mathbb{Z}} 10^{k+\langle\log\alpha\rangle}[a, b] = \bigcup_{k\in\mathbb{Z}} 10^k B,$$

on B és:

$$B = \begin{cases} \left[10^{\langle\log\alpha\rangle}a, 10^{\langle\log\alpha\rangle}b\right] & \text{si } 0 \leq \langle\log\alpha\rangle < 1 - \log b, \\ \left[1, 10^{\langle\log\alpha\rangle-1}b\right] \cup \left[10^{\langle\log\alpha\rangle}a, 10\right] & \text{si } 1 - \log b \leq \langle\log\alpha\rangle < 1 - \log a, \\ \left[10^{\langle\log\alpha\rangle-1}a, 10^{\langle\log\alpha\rangle-1}b\right] & \text{si } 1 - \log a \leq \langle\log\alpha\rangle < 1. \end{cases}$$

i :

$$\mathbb{B}(\alpha A) = \begin{cases} \log 10^{\langle\log\alpha\rangle}b - \log 10^{\langle\log\alpha\rangle}a \\ \log 10^{\langle\log\alpha\rangle-1}b + 1 - \log 10^{\langle\log\alpha\rangle}a \\ \log 10^{\langle\log\alpha\rangle-1}b - \log 10^{\langle\log\alpha\rangle-1}a \end{cases}$$

$= \log b - \log a = \mathbb{B}(A)$, per tant $\mathbb{B}$ té els dígits invariants per canvi d'escala.

Teorema 4.2.1. *Una mesura de probabilitat en $(\mathbb{R}^+, \mathcal{A})$ té dígits significatius invariants per canvi d'escala si i només si $P(A) = \mathbb{B}(A)$, $\forall A \in \mathcal{S}$, és a dir, si i només si P és Benford.*

La densitat d'un conjunt $A \subset \mathbb{N}$ és $\rho \in [0,1]$ si $\lim_{N\to\infty} \#\{1 \le n \le N : n \in A\}/N$ existeix i serà igual que ρ.

Teorema 4.2.2. *1. Sigui (x_n) una successió de nombres reals, $\{n : x_n \neq 0\} = \{n_1 < n_2 < \ldots\}$, diem (x_n) té dígits significatius invariants per canvi d'escala si i només si existeix una funció de densitat per $\{n : x_n \neq 0\}$ i o bé $\rho(\{x : x_n = 0\}) = 1$, o bé $(x_{n_j})_{j\in\mathbb{N}}$ és Benford. En particular, si $\rho(\{x : x_n = 0\}) = 0$ aleshores (x_n) té dígits significatius invariants per canvi d'escala si i només si és Benford.*

2. Una funció(Borel mesurable) $f : [0, +\infty) \to \mathbb{R}$ amb $\lambda(\{t \ge 0 : f(t) = 0\}) < +\infty$ té dígits significatius invariants per canvi d'escala si i només si és Benford. A més a més, una funció f és Benford si αf és Benford per tot $\alpha \neq 0$.

Teorema 4.2.3. *Sigui X una variable aleatòria que compleix $P(X = 0) = 0$, aleshores són equivalents:*

1. *X és Bendord.*
2. *Existeix un nombre $d \in 1, 2, \ldots, 9$ tal que*

$$\mathbb{P}(D_1(\alpha X) = d) = \mathbb{P}(D_1(X) = d) \quad \forall \alpha > 0$$

En particular, (2) implica que $\mathbb{P}(D_1(X) = d) = \log(1 + d^{-1})$.

4.3 Caracterització a través de la invariància per canvi de base

La invariància per canvi de base és una hipòtesi més subtil que també compleix la llei de Benford. La idea principal és que la llei s'ha de continuar complint quan escrivim els nombres en altres bases de numeració, diferent de $b = 10$.

Per demostrar aquest fet, primer veurem que la invariància per canvi de base es manté quan la nova base és una potència de la base original. En aquest sentit, aquesta condició necessària sembla més feble que la condició d'invariància per canvi a qualsevol base. Però més endavant veurem (Teorema 4.3.1) que aquesta condició també és suficient.

De la mateixa manera que l'única variable aleatòria que és invariant per canvi d'escala és la variable aleatòria degenerada igual a la constant 0, l'única variable aleatòria complint la invariància per canvi de base és $\mathbb{P}(X = 1) = 1$. Però, sí que existeixen altres variables aleatòries que tenen els dígits significatius invariants per canvi de la base.

Abans de donar la definició, mirem un exemple.

Exemple 4.3.1. Sigui

$$A = \left\{x > 0 : 1 \le D_1(x)^{(10)} < 3\right\} = \{x > 0 : S(x) \in [1, 3)\}$$

un conjunt de nombres reals té els primers dígits entre 1 i 3 en base 10. Aleshores

$$A^{1/2} = \left\{ x > 0 : S(x) \in [1, \sqrt{3}) \cup [\sqrt{10}, \sqrt{30}) \right\}$$

Considerem ara la funció significand en base 100, S_{100}. És a dir, per a $x \neq 0$, $S_{100}(x)$ és l'únic real entre [1,100) tal que $|x| = 100^k S_{100}(x)$ per un algun (necessàriament) únic $k \in \mathbb{Z}$. Aleshores

$$A = \left\{ x > 0, S_{100}(x) \in [1, 3) \cup \left[\sqrt{10}, \sqrt{30} \right) \right\}.$$

Tenim:

$$\left\{ x > 0 : S_b(x) \in [1, b^{a/2}) \cup [b^{1/2}, b^{(1+a)/2}) \right\} = \begin{cases} A^{1/2} & \text{si } b = 10, \\ A & \text{si } b = 100. \end{cases}$$

on $a = \log 3$. Per tant, una mesura de probabilitat P té els dígits significatius invariants si $P(A)$ i $P(A^{1/2})$ són iguals. Així $P(A) = P(A^{1/n})$ s'ha de complir per tot n. ($A^{1/n} \in \mathcal{S}$, per tot $A \in \mathcal{S}$ i $n \in \mathbb{N}$ (Lemma 3.2.1 (3))).

Passem ara a definir què vol dir que una mesura de probabilitat té els dígits significatius invariants per canvi de base.

Definició 4.3.1. *Sigui una σ-àlgebra $\mathcal{A} \supset \mathcal{S}$ en $\mathbb{R}^+$. Una mesura de probabilitat P en $(\mathbb{R}^+, \mathcal{A})$ té els dígits significatius invariants per canvi de base si $P(A) = P(A^{1/n})$ per tot $A \in \mathcal{S}$ i $n \in \mathbb{N}$.*

Anem a veure que la mesura de probabilitat $\mathbb{B}$ té els dígits significatius invariants per canvi de base. En efecte, per qualsevol $0 \leq s < 1$, sigui

$$A = \{x > 0 : S_{10}(x) \in [1, 10^s)\} = \bigcup_{k \in \mathbb{Z}} 10^k [1, 10^s) \in \mathcal{S}$$

Aleshores:

$$A^{1/n} = \bigcup_{k \in \mathbb{Z}} 10^k \bigcup_{j=0}^{n-1} [10^{j/n}, 10^{(j+s)/n})$$

Aleshores:

$$\mathbb{B}(A^{1/n}) = \sum_{j=0}^{n-1} \left(\log 10^{(j+s)/n} - \log 10^{j/n} \right) = \sum_{j=0}^{n-1} \left(\frac{j+s}{n} - \frac{j}{n} \right) = s = \mathbb{B}(A)$$

per tant $\mathbb{B}$ té els dígits significatius invariants per canvi de base.

Observació 4.3.1. Sigui $\delta_1(A) = 1$ si $1 \in A$, i $\delta_1(A) = 0$ si $1 \neq A$, observem que la mesura de probabilitat δ_1 té els dígits significatius invariants per canvi de base ja que $1 \in A$ si i només si $1 \in A^{1/n}$.

El teorema següent ens caracteritza les mesures de probabilitat que tenen els dígits significatius invariants per canvi de base.

Teorema 4.3.1. *(Caracterització a través de la invariància de base) Una mesura de probabilitat P en $(\mathbb{R}^+, \mathcal{A})$ amb $\mathcal{A} \supset \mathcal{S}$ té els dígtis significatius invariants per canvi de base si, i només si, per algún $q \in [0,1]$ es compleix:*

$$P(A) = q\delta_1(A) + (1-q)\mathbb{B}(A) \quad \forall A \in \mathcal{S}$$

D'aquesta manera, la contant 1 té la probabilitat positiu q.

Corol·lari 4.3.1. *Si una mesura de probabilitat en $\mathbb{R}^+$ té els dígits significatius invariant per canvi d'escala aleshores té els dígits significatius invariants per canvi de base.*

Observem que δ_1 té els dígits invariants per canvi de base, però no per canvi d'escala.

4.4 Caracterització a través de la invariància de suma

Nigrini va observar que en un conjunt de nombres reals aproxima molt bé a la llei de Benford, la suma dels significands dels nombres que tenen el primer dígit 1 és molt similar a la suma dels significands dels nombres que tenen el primer dígit 2, i la suma dels significands dels nombres començen amb altres primers dígits també se sembla molt. És evident que hi ha més nombres començen amb 1 que 2, i més nombres començen amb 2 que 3, així successivament. De la manera similar, les sumes dels significands dels nombres que comenin amb els mateixos dos primers dígtis és semblants, independentment quin sigui els dos primers dígits.

Definició 4.4.1. *Una successió de nombres reals té dígits significatius invariants per la suma si per tot $m \in \mathbb{Z}$,*

$$\lim_{N\to\infty} \frac{\sum_{n=1}^{N} S_{d_1,\ldots,d_m}(x_n)}{N}$$

existeix i és independent de $d_1, \ldots, d_m$.

En particular, si una successió de reals (x_n) té el primer dígit significatiu invariants per la suma si existeix c tal que:

$$\lim_{N\to\infty} \frac{\sum_{n=1}^{N} S_{d_1}(x_n)}{N} = c, \quad \forall d_1 = 1, 2, \ldots, 9.$$

Definició 4.4.2. *Una variable aleatòria té dígits significatius invariants per la suma si per tot $m \in \mathbb{Z}$, $\mathbb{E}S_{d_1,\ldots,d_m}(X)$ existeix i és independent de $d_1, \ldots, d_m$.*

Teorema 4.4.1. *(Caracterització a través de la invariància de suma) Una variable aleatòria X amb $\mathbb{P}(X = 0) = 0$ té els dígits significatius invariants per la suma si i només si és Benford.*

5 Els tests de la llei de Benford

En aquesta secció anem a veure diferents tests per comprovar la conformitat d'un conjunt de dades a la llei de Benford. A part de veure si segueix la llei del primer dígit, podem analitzar també els dígits següents.

5.1 Test dels dos primers dígits o test de primer ordre

Segons la fórmula de la definició 3.3.2 , la probabilitat de d_1d_2 com els dos primers dígit significatiu és:

$$\text{Prob}(D_1D_2 = d_1d_2) = \log\left(1 + \frac{1}{d_1d_2}\right), \qquad d_1d_2 \in \{10, 11, \ldots, 99\}.$$

Una successió de nombres reals (x_n) satisfà el test dels dos primers dígits si la freqüència relativa observada de cada dos primers dígits d_1d_2 és acceptablement pròxima a la probabilitat teòrica. El test dels primers dos dígits és més preferida que el test del primer dígit i el test del segon dígit per separat, perquè dóna més informació, excepte quan la mostra és relativament petita. Observem que un conjunt de nombre només tenen un dígit, podría seguir la llei del primer dígit, però no pas dels dos primers dígits.

5.2 Test de sumació

Sigui una successió de nombres reals (x_n), segons el teorema 4.4.1, si la suma dels significands de tots els nombre que comencen amb el(s) mateix(os) dígit(s) són acceptablement pròxima, satisfà la hipòtesi nul·la del test de sumació.
Cas del primer dígit, la suma és igual a:

$$\sum_{D_1(x_i)=d_1} x_i = \frac{\sum_1^N S(x_i)}{9}, \qquad \forall d_1 \in \{1, 2 \ldots, 9\}$$

Cas dels dos primers dígits, la suma és igual a:

$$\sum_{D_1D_2(x_i)=d_1d_2} x_i = \frac{\sum_1^N S(x_i)}{90}, \qquad \forall d_1 \in \{1, 2 \ldots, 9\}, d_2 \in \{1, 2 \ldots, 9\}$$

El test es pot generalitzar a tots els dígits (la suma dels tres primers dígits, la suma dels quatre primers dígits), si acceptem la hipòtesi nul·la del test de sumació per tots els dígits, la successió és Benford.

5.3 Test de segon ordre

Sigui (x_n) una successió de nombres reals, la ordenem en ordre creixent i s'obté (y_n). Fem la diferència entre les observacions adjacents $(z_j = y_{i+1} - y_i, i = 1, \ldots, n-1)$, obté una nova successió amb una observació menys. La hipòtesi nul·la és la successió (z_{n-1}) segueix la llei de Benford.

5.4 Test dels dos últims dígits

A la taula 2, hem observat que el m-ésim dígit tendeix a estar uniformement distribuït a mesura que augmentant m, indenpendent de quin valor sigui d_m. Des del tercer dígit en endavant, la probabilitat de tenir 1 o 2, o ..., 9 com dígit és quasi igual. Els últims dos dígits poden ser: $00, 01, \ldots, 99$, per tant la probabilitat esperada de dos últims dígits és constant igual a 1/100=0.01. El test de dos últims dígits té sentit si els nombres tenen més de 4 dígits, en el cas els nombres només tenen 2 dígits, els dos últims dígits són els dos primers dígits, la probabilitat no és constant, i no té sentit realitzar el test dels dos últims dígits.

5.5 Test de mantissa

Sigui una successió de nombres reals (x_n), la ordenem en ordre creixent, segons el Teorema 4.1.1, si els logaritmes dels significands dels nombre estan uniformement distribuïts en modùl 1, accepta la hipòtesi nul·la de que la dada és una successió de Benford.

5.6 Test khi-quadrat

La fórmula de la prova de khi-quadrat per a comprovar la bondat d'ajust de la mostra és:

$$\chi^2 = N * \sum_{d=10^{k-1}}^{10^k-1} \frac{(\text{Prob}^o(d) - \text{Prob}^e(d))^2}{\text{Prob}^e(d)}$$

on N és nombre d'observacions, Prob^o és la probabilitat observada del conjunt i Prob^e és la probabilitat teòrica segons la llei de Benford. En particular, el khi-quadrat del primer dígit:

$$\chi^2 = N * \sum_{d_1=1}^{9} \frac{(\text{Prob}^o(D_1 = d_1) - \log(1 + \frac{1}{d_1}))^2}{\log(1 + \frac{1}{d_1})} \quad o \quad \chi^2 = \sum_{d_1=1}^{9} \frac{(n_j - N_j)^2}{N_j}$$

on $N_j = N * \text{Prob}(D_1 = d_j)$, n_j és el nombre de vegades que el conjunt té d_j com el primer dígit significatiu . Els valors crítics per a 8 grau de llibertat en un nivell de significació de l'1% i 5% són respectivament 15.51 i 20.09.

5.7 Test de desviació mitjana absoluta

La desviació mitjana absoluta(MAD) calcula la mitjana absoluta de la diferència entre la probabilitat observada d'una successió real x_n i la probabilitat teòrica. Els MAD del primer dígit i dels dos primers dígits són:

$$\text{MAD} = \frac{\sum_{d_1=1}^{9} |\text{Prob}^o(D_1 = d_1) - \log(1 + \frac{1}{d_1})|}{9}$$

$$\text{i,} \quad \text{MAD} = \frac{\sum_{(d_1,d_2)=(1,0)}^{(9,9)} |\text{Prob}^o((D_1, D_2) = (d_1, d_2)) - \log(1 + \frac{1}{d_1+10d_2})|}{90}$$

Observem que el test no té en compte el nombre d'observació. Les dades amb el MAD més baix s'ajusten millor a la llei de Benford.

Taula 5: La taula de MAD

Digits	Range	Conclusion
First Digits	0.000 to 0.006	Close conformity
	0.006 to 0.012	Acceptable conformity
	0.012 to 0.015	Marginally acceptable conformity
	Above 0.015	Nonconformity
Second Digits	0.000 to 0.008	Close conformity
	0.008 to 0.010	Acceptable conformity
	0.010 to 0.012	Marginally acceptable conformity
	Above 0.012	Nonconformity
First-Two Digits	0.0000 to 0.0012	Close conformity
	0.0012 to 0.0018	Acceptable conformity
	0.0018 to 0.0022	Marginally acceptable conformity
	Above 0.0022	Nonconformity
First-Three Digits	0.00000 to 0.00036	Close conformity
	0.00036 to 0.00044	Acceptable conformity
	0.00044 to 0.00050	Marginally acceptable conformity
	Above 0.00050	Nonconformity

5.8 Test de desviació màxima

La desviació màxima del primer dígit és la diferència màxima en valor absolut de la probabilitat del primer dígit entre la successió observada i la llei del primer dígit:

$$d_{max} = \max_{\{1 \le d_1 \le 9\}} \left\{ \left| Prob^o(D_1 = d_1) - \log(1 + \frac{1}{d_1}) \right| \right\}.$$

La desviació màxima dels dos primers dígits és la diferència màxima en valor absolut de la probabilitat dels dos primers dígits entre la successió observada i la llei dels dos primers dígits:

$$d_{max} = \max_{\{10 \le d_1+10d_2 \le 99\}} \left\{ \left| Prob^o((D_1, D_2) = (d_1, d_2)) - \log(1 + \frac{1}{d_1 + 10d_2}) \right| \right\}.$$

5.9 Test d'arc de mantissa

Sigui (x_n) una successió de nombres reals, representem les mantisses dels nombres en un cercle de centre (0,0) i radi 1. Les coordenades del les mantisses són:

x-coordenada=$\cos(2\pi * (\log(x_i) \text{mod} 1))$, y-coordenada=$\sin(2\pi * (\log(x_i) \text{mod} 1))$

Coordenada del centre de gravetat:

x-coordenada=$\frac{\sum_{i=1}^{N} \cos(2\pi * (\log(x_i) \text{mod} 1))}{N}$, y-coordenada=$\frac{\sum_{i=1}^{N} \sin(2\pi * (\log(x_i) \text{mod} 1))}{N}$

La distància del centre de gravetat al (0,0):

$L^2 = (x - coordenada)^2 + (y - coordenada)^2$.

$p - value = 1 - e^{-L^2 \times N}$ amb la grau de llibertat 2.

Les mantisses ordenades d'una successió de Benford es distribueix uniformement al cercle de centre (0,0) amb radi 1. Si el centre de gravetat de les mantisses de la dada real és estadísticament indiferent de (0,0) (amb el nivell de significació 0.05), pot afirmar que la dada és u.d mod 1.

6 Exemples

6.1 Els enters positius

Primer anem a veure que $\log n$ no és u.d.mod 1. Per tot $s \in [0,1)$, tenim

$$\liminf_{N\to\infty} \frac{\#\{1 \le n \le N : \langle \log n \rangle \le s\}}{N} = \frac{10^s - 1}{9}$$

i

$$\limsup_{N\to\infty} \frac{\#\{1 \le n \le N : \langle \log n \rangle \le s\}}{N} = \frac{10(1 - 10^{-s})}{9}$$

El límit inferior i el límit superior no coincideixen, com el límit és diferent de s, segons la definició 4.1.1, $f(t) = \log n$ no és u.d.mod.1. Segons el teorema 4.1.1, podem concloure que la successió d'enters positius n no és Benford.

6.2 La Distribució uniforme contínua

Sigui $X \sim U(0,1)$ una distribució uniforme en $[0,1)$, aleshores per tot $1 \le t < 10$,

$$\mathbb{P}(S(X) \le t) = P_X\left(\bigcup_{k\in\mathbb{Z}} 10^k[1,t]\right) = \sum_{n\in\mathbb{N}} 10^{-n}(t-1) = \frac{t-1}{9} \neq \log t,$$

per tant $U(0,1)$ no és Benford.

6.3 Distribució exponencial Exp(1)

Sigui $X \sim Exp(1)$ una distribució exponencial amb paràmetre 1, observem

$$\begin{aligned}\mathbb{P}(D_1(X) = 1) = \mathbb{P}(X \in \bigcup_{k\in\mathbb{Z}} 10^k[1,2)) &= \sum_{k\in\mathbb{Z}}((1 - e^{-2\cdot 10^k}) - (1 - e^{-10^k})) \\ &= \sum_{k\in\mathbb{Z}}(-e^{-2\cdot 10^k} + e^{-10^k}) > \sum_{k=1}^{3}(-e^{-2\cdot 10^k} + e^{-10^k}) = 0.318\ldots > \log 2,\end{aligned}$$

com que $-e^{-2\cdot 10^k} + e^{-10^k}$ és positiu per tot $k \in \mathbb{Z}$, la penúltima desigualtat és sempre cert. Com que la probabilitat del primer dígit sigui 1 del la distribució exponencial és diferent de la llei del primer dígit, $exp(1)$ no és Benford.

Les simulacions de la taula 6 s'han obtingut amb set.seed(2016), observem que els valors de MAD són més gran que 0.015, per tant, indiquen no conformitat. Excepte el de Exp(1) amb mida 1000, té una conformitat lleugera. Els p-valors de χ^2 i d'arc de mantissa de aquestes 4 simulacions rebutjen la hipòtesi nul·la, podem dir que no segueixen la llei del primer dígit.

Taula 6: Bondat de l'ajust del primer dígit de simulacions de D.U i Exponencial

	N	MAD	d_{max}	Khi-quadrat	p-valor	L2	p-valor
U(0,1)	100	0.05	0.15	38.21	$6.88 \cdot 10^{-6}$	0.055	0.004
U(0,1)	1000	0.06	0.20	437.1	$< 2.2 \cdot 10^{-16}$	0.13	$< 2.2 \cdot 10^{-16}$
e^{-x}	100	0.026	0.07	9.56	0.30	0.02	0.15
e^{-x}	1000	0.012	0.047	18.21	0.020	0.007	0.0008

6.4 Funció lineal

Sigui la funció lineal $f(t) = at + b$, $a, b \in \mathbb{R}$. Primer anem a veure que $f(t) = \log|at + b|$ no és u.d.mod.1

$$\liminf_{T\to\infty} \frac{\lambda(\{\tau \in [0,T) : \langle \log|a\tau + b| \leq s\rangle\})}{T} = \frac{10^s - 1}{9}$$

i

$$\limsup_{T\to\infty} \frac{\lambda(\{\tau \in [0,T) : \langle \log|a\tau + b| \leq s\rangle\})}{T} = \frac{10(1 - 10^{-s})}{9}$$

El límit inferior i el límit superior no coincideixen, com el límit és diferent de s, segons la definició 4.1.1, $f(t) = \log|at + b|$ no és u.d.mod.1. Segons el teorema 4.1.1, podem concloure que $f(t) = at + b$ no és Benford.

6.5 $\alpha^n a + \beta^n b$, α, β, a, b nombres reals

Segons teorema 4.1.2, podem afirmar:

1. $2^n, 3^n, 5^n, 2^n + 3^n$ són Benfords, ja que $\log 2, \log 3, \log 5$ són irracionals.
2. $10^n, 0.1^n, \sqrt{10}$ no són Benfords, ja que $\log 10 = 1, \log 0.1 = -1, \log\sqrt{10} = \frac{1}{2}$ no són irracionls.
3. $0.1 \cdot 0.03^n + 0.03 \cdot 0.1^n$ no és Benford, ja que $0.1 > 0.03$ i $\log 0.1$ no és irracional.
4. $(0.2^n + (-0.2)^n)$ no és Benford, ja que quan n és senar, els termes s'anul·len.

La successió $r^n = 1.0002303^n \in [1, 10)$, $n = 1, 2 \cdots, 10000$ és una successió de nombres reals entre 1 i 10 amb les mantisses uniformement distribuïts a l'interval $[0, 1)$ (cada terme incrementa 0.0001 respecte l'anterior). Observem la taula 7, el primer dígit de la successió 2^n s'apropa més a la llei del primer dígit quan augmenta la mida de 100 a 1000, el valor de MAD que està a la taula 8 també ha disminuït, la desviació màxima s'ha canviat del dígit 5 al dígit 6. La combinació d'una successió de Benford i un de no també s'ajusten molt bé a la llei del primer dígit. La successió $r^n = 1.0002303^n \in [1, 10)$, $n = 1, 2 \cdots, 10000$ s'apropa molt bé com s'esperava (Teorema 4.1.1), té un MAD igual a $4.46e - 05$ és quasi 0, indica una conformitat perfecta. El test de χ^2 i d'arc de mantissa de tots aquestes successions no rebutjen la hipòtesi nul·la, podem afirmar que aquestes successions són Benfords.

Taula 7: El primer dígit de $a\alpha^n + b\beta^n$

d	P(1)%	P(2)%	P(3)%	P(4)%	P(5)%	P(6)%	P(7)%	P(8)%	P(9)%
2^{100}	30	17	13	10	7	7	6	5	5
2^{1000}	30.1	17.6	12.5	9.7	7.9	6.9	5.6	5.2	4.5
$2^{1000} + 7 \cdot 1^{1000}$	30.2	17.6	12.5	9.6	7.9	6.8	5.7	5.1	4.6
(1)	30.11	17.60	12.50	9.69	7.92	6.69	5.80	5.11	4.5
(2)	30.17	17.58	12.58	9.79	7.79	6.59	5.89	5.09	4.5
$10^2 \log(1 + d^{-1})$	30.10	17.60	12.49	9.691	7.918	6.694	5.799	5.115	4.575

on (1):$r^n = 1.0002303^n, N = 10000$ i (2):$ar^n = 3 \cdot 1.34^n, N = 1000$

Taula 8: Bondat de l'ajust de $a\alpha^n + b\beta^n$

	MAD	d_{max}	Khi-quadrat	p-valor	L2	p-valor
2^{100}	0.0039	0.009(d=5)	0.22	1	$1.53 \cdot 10^{-5}$	0.9985
2^{1000}	0.0007	0.002(d=6)	0.16	1	$1.35 \cdot 10^{-8}$	1
$2^{1000} + 7 \cdot 1^{1000}$	0.0005	0.001(d=6)	0.047	1	$4.85 \cdot 10^{-6}$	0.995
(1)	$4.46 \cdot 10^{-5}$	0.009(d=2)	0.002	1	$4.23 \cdot 10^{-9}$	1
(2)	0.0008	0.126(d=5)	0.008	1	$2.93 \cdot 10^{-6}$	0.997

6.6 Nombres de Fibonacci

Definició 6.6.1. *Els nombres de Fibonacci es defineix com:*

$$F(n) = \begin{cases} 0 & si\ n = 0, \\ 1 & si\ n = 1, \\ F(n-1) + F(n-2) & altrament. \end{cases}$$

La fórmula general dels nombres de Fibonacci és:

$$F(n) = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2} \right)^n - \frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2} \right)^n .$$

Com que $\lim_{n\to\infty} \frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2} \right)^n = 0$, $\lim_{n\to\infty} F(n) = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2} \right)^n$.

$\lim_{n\to\infty} \log F(n) = n \log \frac{1+\sqrt{5}}{2} + \log \frac{1}{\sqrt{5}}$, on $\log \frac{1+\sqrt{5}}{2}$ i $\log \frac{1}{\sqrt{5}}$ són constants, per tant el logaritme de $F(n)$ és uniforme distribuït, podem deduir que la successió de Fibonacci és Benford. O per teorema 4.1.1, com que $\log(\frac{1+\sqrt{5}}{2})$ és irracional, obtenim la mateixa conclusió.

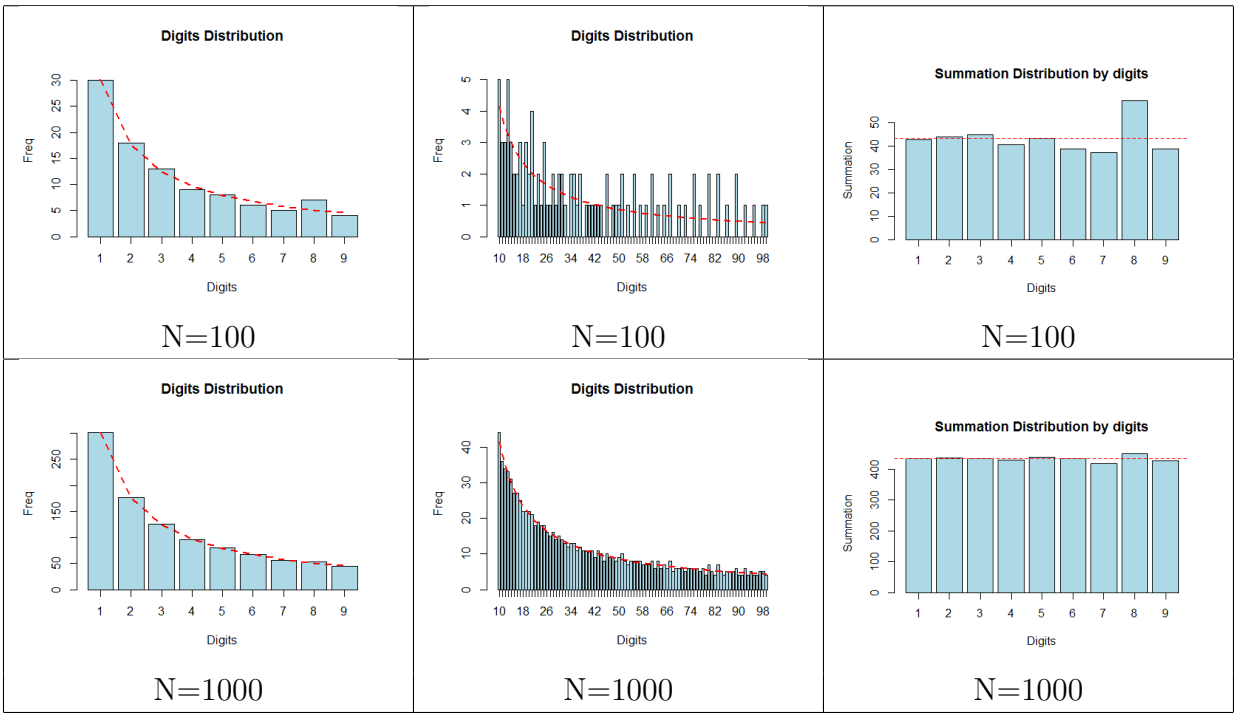
Taula 9: El primer dígit dels nombres de Fibonacci

d	P(1)%	P(2)%	P(3)%	P(4)%	P(5)%	P(6)%	P(7)%	P(8)%	P(9)%
$N = 10^2$	30	18	13	9	8	6	5	7	4
$N = 10^3$	30.1	17.7	12.5	9.6	8.0	6.7	5.6	5.3	4.5
$10^2 \log(1 + d^{-1})$	30.10	17.60	12.49	9.691	7.918	6.694	5.799	5.115	4.575

Taula 10: Bondat de l'ajust dels Nombres de Fibonacci

	MAD	$d_{max} * 10^2$	Khi-quadrat	p-valor	L2	p-valor
$N = 10^2$	0.0064	1.884(d=8)	1.0288	0.9981	$9.6481 \cdot 10^{-5}$	0.9904
$N = 10^3$	0.0008	0.199(d=7)	0.1695	1	$2.9671 \cdot 10^{-7}$	0.9997

Figura 3: Els gràfics de la successió de Fibonacci



A taula 9 veiem que a mesura augmentant n, la probabilitat del primer dígit conformen cada cop més a la llei del primer dígit. I a la taula 10 mostren la bondat d'ajust. Quan N=100, el valor MAD=0.0064 $<$ 0.012 indica la probabilitat del primer dígit de la successió té una conformitat acceptable, la diferència entre la probabilitat observada i teòrica del dígit 8 és més gran del tot: $1.88 \cdot 10^{-4}$. Els p-valors del test de χ^2 i test d'arc de mantissa són molt proper a 1, també impliquen que la successió s'ajusten molt bé a la llei del primer dígit. Quan N=1000, la conformitat ja és molt propera a la llei del primer dígit, la diferència màxima entre la probabilitat observada i teòrica és el dígit 7: $1.99 \cdot 10^{-3}$. Els p-valors del test

de Khi-quadrat i test d'arc de mantissa ara són encara més a prop a 1, podem condiderar el nombres de Fibonacci com una successió de Benford.

A la figura 2, observem que quan N=100, la distribució del primer dígit segueix bastant bé a la línea vermella discontínua (la llei del primer dígit) excepte el dígit 8 sobresurt una miqueta. A la distribució dels dos primers dígits no conforme gaire bé, té molts pics sobresurten a la línea vermella, vol dir la probabilitat observada és més gran que la probabiltat esperada de la llei. El gràfic de la suma és bastant uniforme, excepte la suma dels nombres tenen el primer dígit 8 és més gran de tot. Quan N=1000, tots 3 gràfics s'ajusten molt bé!

6.7 Els factorials

Per els primers 150 valors numèrics factorials, la prova de khi-quadrat no rebutja la distribució de la Benford, però, la desviació mitjana absoluta és massa gran, i indica no conformitat.

Taula 11: Percentage del primer dígit dels factorials de 1 a n

d	P(1)%	P(2)%	P(3)%	P(4)%	P(5)%	P(6)%	P(7)%	P(8)%	P(9)%
$N = 50$	24	22	16	6	8	12	2	10	0
$N = 100$	30	18	13	7	7	7	3	10	5
$N = 150$	32.67	16.67	13.33	6.00	7.33	6.67	2.00	8.67	6.67
$10^2 \log(1 + d^{-1})$	30.10	17.60	12.49	9.691	7.918	6.694	5.799	5.115	4.575

Taula 12: Bondat de l'ajust del primer dígit dels factorials

	MAD	$d_{max} * 10^2$	Khi-quadrat	p-valor	L2	p-valor
$N = 50$	0.040	6.1(d=1)	10.33	0.243	0.008	0.65
$N = 100$	0.014	4.88(d=8)	6.95	0.54	0.0035	0.71
$N = 150$	0.020	3.80(d=7)	11.53	0.17	0.0051	0.43

6.8 Exponencial

La funció $f(t) = at + b$ on $a, b \in \mathbb{R}$, és u.d mod 1 si i només si $a \neq 0$. Per tant, segons el teorema 4.1.1, la funció e^{at} és Benford si $a \neq 0$. ($\log f(t) = \frac{at}{\ln 10}$ és u.d mod 1).

6.9 Nombres primers

Sigui (p_n) una successió de n nombrs primers, per teorema dels nombrs primers, sabem $p_n = O(n \log p_n)$, com que $p_n / \log n \approx n$ és fitat, quan $n \to \infty$. Per tant,

segons la proposició 4.1.1(4), els nombres primers no són Benford.

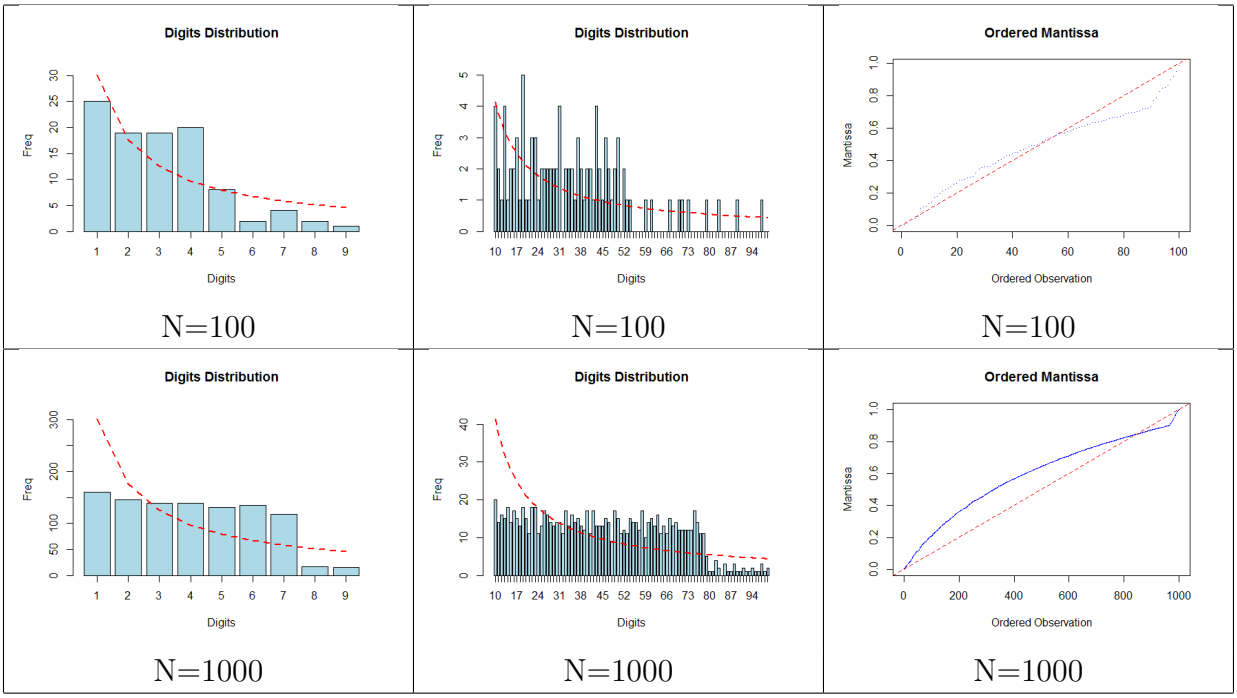
Taula 13: El primer dígit dels nombres primers

d	P(1)%	P(2)%	P(3)%	P(4)%	P(5)%	P(6)%	P(7)%	P(8)%	P(9)%
$N = 10^2$	25	19	19	20	8	2	4	2	1
$N = 10^3$	16.0	14.6	13.9	13.9	13.1	13.5	11.8	1.7	1.5
$10^2 \log(1 + d^{-1})$	30.10	17.60	12.49	9.69	7.91	6.69	5.79	5.11	4.57

Taula 14: Bondat de l'ajust del primer dígit dels nombres primers

	MAD	$d_{max} * 10^2$	Khi-quadrat	p-valor	L2	p-valor
$N = 10^2$	0.041	10.30(d=4)	23.872	0.002408	0.088054	0.0001499
$N = 10^3$	0.052	141.03(d=1)	299.74	$< 2.2 \cdot 10^{-16}$	0.093277	$< 2.2 \cdot 10^{-16}$

Figura 4: Els gràfics dels nombres primers



A la taula 13, observem quan N=100, les probabilitats del primer dígit és bastant diferent de la llei del primer dígit, a mesura augentant la mida, tampoc millora. Ho podem observar també a la figura 4, quan N=100, els primers 4 dígits tenen molt més freqüència que els altres, i quan N=1000, els primers 7 dígits tenen freqüència relativa més o menys iguals, i els dígits 8 i 9 tenen freqüència relativa molt més

menor, aquest comportament segueixen als dos primers dígits. Els gràfics de mantissa no són uniforment distribuïts. I a la taula 14, observem que els MAD són major que 0.015, indican una conformitat molt malament. Quan N=100, el dígit 4 comporta més diferent que els altres. Tant el test de Khi-quadrat com el test d'arc de mantissa ens diu el primer dígit dels nombres primers no s'ajusten gens. Quan augmenta la mida a 1000, es comporten encara pitjor.

6.10 Distribucions comuns

A banda de la distribució uniforme i exponencial, les distribucions comuns com normal, gamma, beta, binomial tampoc segueixen la llei de Benford.
He creat una successió alestòria de Benford de 1000 observacions amb set.seed(1992) i les altres amb set.seed(2016). Observem a la taula 15, la multiplicació de la successió de benford amb la distribció uniforme i exponencial sí que segueixen la llei, amb una conformitat bastant bona (el MAD són més petit que 0.006 i el p-valors propen a 1). La divisió de la distribució uniforme entre la distribució de Benford i la divisió de la distribució de Benford entre la distribució exponencial també tenen una conformitat acceptable. Els dos tests no rebutjen la hipòtesi nul·la, les resultats concorden amb les proposicions 4.1.2 i 4.1.3. Observem que la multplicació de les distribucions uniforme, normal i expoencial té una conformitat acceptable (MAD=0.006), i els p-valors són més gran que 0.05, no rebutjen la hipòtesi nul·la de que no és Benford (Teorema 4.1.4). La multiplicació de aquests 5 distribucions no Benfords té un MAD=0.013¡0.015, indica una conformitat marginalment acceptable, però tampoc rebutja la hipòtesi nul·la (p-valors més gran 0.05), podem dir que la distribució de la multiplicació de normal, uniforme, exponencial, gamma i beta és Benford.

Taula 15: Simulació de 1000 observacions

	MAD	d_{max}	Khi-quadrat	p-valor	L2	p-valor
rbenf	0.004	0.009	3.23	0.9	0.0001	0.84
rbenf*U(0,1)	0.006	0.017	7.29	0.50	0.0001	0.89
rbenf*e^{-1}	0.003	0.006	1.74	0.99	0.0003	0.70
U(0,1)/rbenf	0.008	0.013	7.65	0.47	0.0004	0.70
rbenf/e^{-1}	0.006	0.015	4.5	0.8	0.001	0.33
N(0,1)	0.02	0.06	230.9	$< 2.2 \cdot 10^{-16}$	0.025	$< 2.2 \cdot 10^{-16}$
N(0,1)*U(0,1)*e^{-1}	0.006	0.016	2.46	0.96	0.0006	0.71
Beta(1/2,1/2)	0.06	0.19	1023.2	$< 2.2 \cdot 10^{-16}$	0.09	$< 2.2 \cdot 10^{-16}$
Gamma(1.5,1.5)	0.016	0.04	235.3	$< 2.2 \cdot 10^{-16}$	0.012	$< 2.2 \cdot 10^{-16}$
(3)	0.013	0.034	9.07	0.33	0.004	0.14

on (3)=U(0,1)*N(0,1)*e^{-1}*beta(1/2,1/2)*Gamma(1.5,1.5)

7 Aplicacions

A la taula 1, hem vist que Benford va recollir una àmplia varietat de tipus de dades, tan diversa com li fou possible per comprovar la llei de Benford. Per exemple, els pesos atòmics dels elements, nombres d'habitants, etc. Després de fer-se famosa la llei, una gran quantitat d'evidència empírica ha aparegut, com les constants físiques, dades de comptabilitat, dades d'eleccions, etc. En aquesta secció anem a veure la llei de Benford no tant com una curiositat, sinó també la seva aplicació pràctica a través dels diferents tests basats en la llei.
Aplicant la llei de Benford a un determinat conjunt de dades, podem comprovar si compleixen la llei. Si, després de fer proves repetides amb dades similars veiem que, en general la compleixen, es pot inferir que qualsevol conjunts de dades d'aquest àmbit mantindrà el compliment de la llei. Si troben un cas concret d'un conjunt de dades d'aquesta classe que no la segueixi, podem sospitar que alguns valors estan alterats, degut a un defecte en la recollida o processat de les dades, o bé a un possible frau.

7.1 Dades del cens

La bona correlació existent entre les estadístiques poblacionals i la Llei de Benford significa que pot usar-se per a verificar models demogràfics. Les xifres de població de pobles i ciutats poden variar des de desenes o centenars a milers o milions, i els afecta un gran ventall de factors. És plausible, doncs, pensar que poden seguir la llei de Benford. He tret tres conjunt de dades a la pàgina web del Institut Nacional d'Estadísitca, la primera [3] és xifres oficials de població dels municipis d'Espanya de l'any 2015. A la pàgina hi ha 52 fitxers separats de la població dels municipis de cada província, he analitzat junts tots els municipis de totes les provincies d'Espanya, són 8119 observacions en total. A la mateixa pàgina, hi ha les dades de població de les provincies d'Espanya, el fitxer conté 52 observacions. I la última dada [4] es tracta de la població de cada país del món.

Taula 16: El primer dígit de la Població de l'any 2015

	P(1)%	P(2)%	P(3)%	P(4)%	P(5)%	P(6)%	P(7)%	P(8)%	P(9)%
E	30.73	18.03	12.43	8.84	8.12	6.42	5.91	4.90	4.62
E^3	30.40	17.76	12.53	9.15	8.12	6.69	5.43	5.47	4.46
E^4	30.98	17.93	11.77	9.23	7.54	6.87	5.79	5.25	4.64
$E * 2$	29.97	18.11	12.62	9.94	8.09	6.65	5.78	4.64	4.20
$E * 7$	30.52	17.19	12.56	9.21	7.81	6.49	6.43	5.30	4.48

[3] Al link: http://www.ine.es/dynt3/inebase/es/index.html?padre=517&dh=1

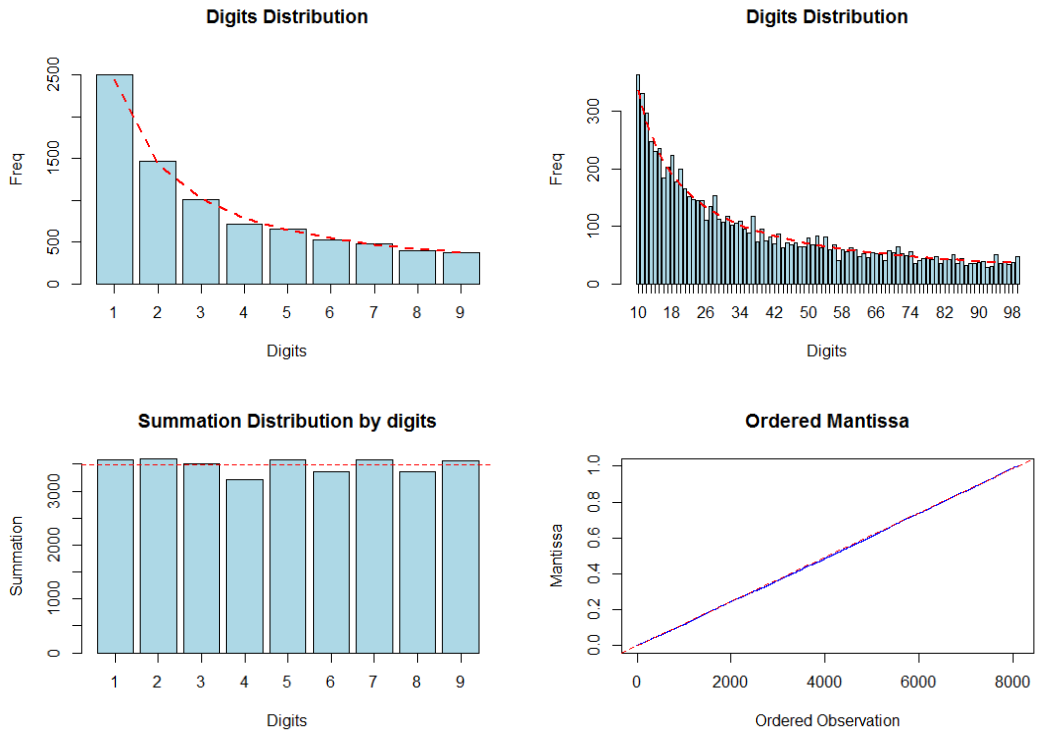
[4] Al link: http://www.ine.es/dynt3/inebase/index.htm?type=pcaxis&path=/t42/p02/&file=pcaxis on conté les informacions de demografia de dades internacionals.

Taula 17: Població de l'any 2015

dades	tamany	MAD	d_{max}	χ^2	p-valor	L2	p-valor
E	8119	0.003	0.0085	10.20	0.25	0.00015	0.289
E $*$ 2	8119	0.002	0.005	8.20	0.41	0.00015	0.29
E $*$ 7	8119	0.0029	0.006	10.07	0.26	0.00015	0.29
E^4	8119	0.002	0.005	7.31	0.50	9.51e-05	0.46
E^3	8119	0.0035	0.0087	9.95	0.29	0.0003	0.06
Les provincies d'Espanya	52	0.0525	0.115	16.88	0.031	0.115	0.0025
Els països del món	232	0.0456	0.203	53.98	$6.97 \cdot 10^{-9}$	0.125	$2.24 \cdot 10^{-13}$

on E és la dada dels municipis d'Espanya

Figura 5: Distribució del primer i del dos primers dígits de la població d'Espanya per municipis



A la taula 17, observem que el valor MAD la població d'Espanya per municipis és $0.003 < 0.006$, podem dir que les dades són molt pròxima a la llei del primer dígit, també ho podem veure al primer gràfic de la figura , la freqüència relativa del primer dígit i del dos primers dígits s'apropen molt a les probabilitats teòriques. El valor observat de khi-quadrat és $10.20 < 15.51$, acceptem la hipòtesi nul·la. I el gràfic 3 mostra que la suma és quasi constant per el primer dígit. Obsevem el gràfic 4 de la figura les mantisses ordenades observades són uniformement distribuïts tal

com explica el valor d'arc de mantissa, podem afirmar que el primer dígit d'Espanya per municipis segueixen la llei de Benford. Multiplico tots els nombres per 2 i 7, els conjunts nous continuan seguint la llei del primer dígit, afirma la caraterística de la invariància per canvi de l'escala. Elevem tot el conjunt per 3, o 4, el conjunt nou continuant seguir la llei del primer dígit, afirma la caraterística de la invariància per canvi de base. Però si analitzem la població de les provincies d'Espanya o de tots els països del món, els valors de MAD són superiors a 0.015, i els p-valors són més gran que 0.05, no conformen a la llei! Podem suposar que aquests dos casos les dades són agregades.

7.2 Detecció de frau

Al nostre voltant, existeixen diferents tipus de frau, com frau fiscal, frau electoral, frau bancari, frau financer, etc. És important trobar eines per detecctar-los i evidar-los. En l'àmbit del frau financer, pot passar frau en la nòmina, en les vendes, en els pagaments, en els xecs, o en l'evasió d'impostos, etc. No hi ha una tècnica de detecció que és el millor de tot, en qualsevol procés de detecció de fraus, les dades han de ser sotmès a diverses proves. La llei de Benford és una tècnica que podria aplicar al començament, a través d'anàlisi de freqüència dels dígits detecten irregularitats sense necessitat de revisar tots els nombres.

La gent a l'hora d'intentar falsificar números té tendència d'utilitzar massa sovint els nombres que comencen pels dígits com 5,6,7, i pocs que comencen per 1. Aquest fet contradiu la llei de Benford. Aquesta violació de la llei no implica necessàriament frau, però si constitueix un bon indici per justificar una inspecció més detallada. Per exemple, la Hisenda dels EUA va determinar que si una xifra comença per tres i apareix amb una probabilitat 40% en lloc del 12,5%, hi ha motius per investigar el frau fiscal.

L'economista americà Mark Nigrini[2] ha acumulat una gran col·lecció d'impostos dels EUA i les dades de comptabilitat dels ingressos i dels inventaris informats per l'IRS [5], en la majoria d'aquests casos la distribució de Benford és un excel·lent ajust (potser precisament perquè cada un és una barreja imparcial de les dades de diferents distribucions). Es postula que Benford és una distribució raonable a esperar dels dígits significatius de grans conjunts de dades de comptabilitat. Es podria esperar que la conformitat amb Benford no varia molt en el temps a causa de la teorema d'invariància d'escala, un conjunt de Benford seguirà sent un conjunt de Benford fins i tot amb els números que inflen any rere any. Fem notar que hi ha una discrepància entre la definició de Nigrini i el que fem servir la defnició 4.4.1, creiem que és un error. Per suplir-lo, he desenvolupat els codis nous. Més detall sobre aquesta discrepància es poden trobar a l'Annex, pàg 45.

[5]Internal Revenue Service

7.2.1 Uns exemples petits

Primer anem a veure uns exemples petits sense dades analitzades per tenir una idea com funciona la llei de Benford a l'hora de la detecció del frau. Són exemples trets del llibre Nigrini(2012) [2].

A la revisió de la despesa de cada empleat per la seva conformitat amb Benfrod podria usar la prova dels dos primers dígits. Un dels empleats es va destacar amb al voltant del 25 per cent dels nombres de la seva despesa que comencen amb 48. La proporció esperada per al 48 és una mica menys d'1 per cent. La revisió va mostrar que la despesa diària de 4.82 dòlars pel seu esmorzar.

Duanes i Rendes al Regne Unit van analitzar els nombres de negocis anual, van informar sobre les declaracions d'impostos d'autoavaluació dels contribuents britànics. Hi va haver un augment significatiu en 14 i una escassetat de nombres amb els dos primera dígits 15 i 16. Va adonar-se que els contribuents estaven apuntant els números de vendes que estaven per sota de 15000£. El control de 15000£ es va mantenir fins a 2007. En 2008, el límit es va augmentar a 64000£ per a la versió curta. El límit es va elevar a 67000£ en 2009 i novament va elevar 68.000£ en 2010. I el límit de 2011 va ser de 70000£. Seria interessant veure els gràfics dels dos primers dígits amb més salts al 62-63, 65-66, 66-67, 68-69 en aquests quatre anys respectivament.

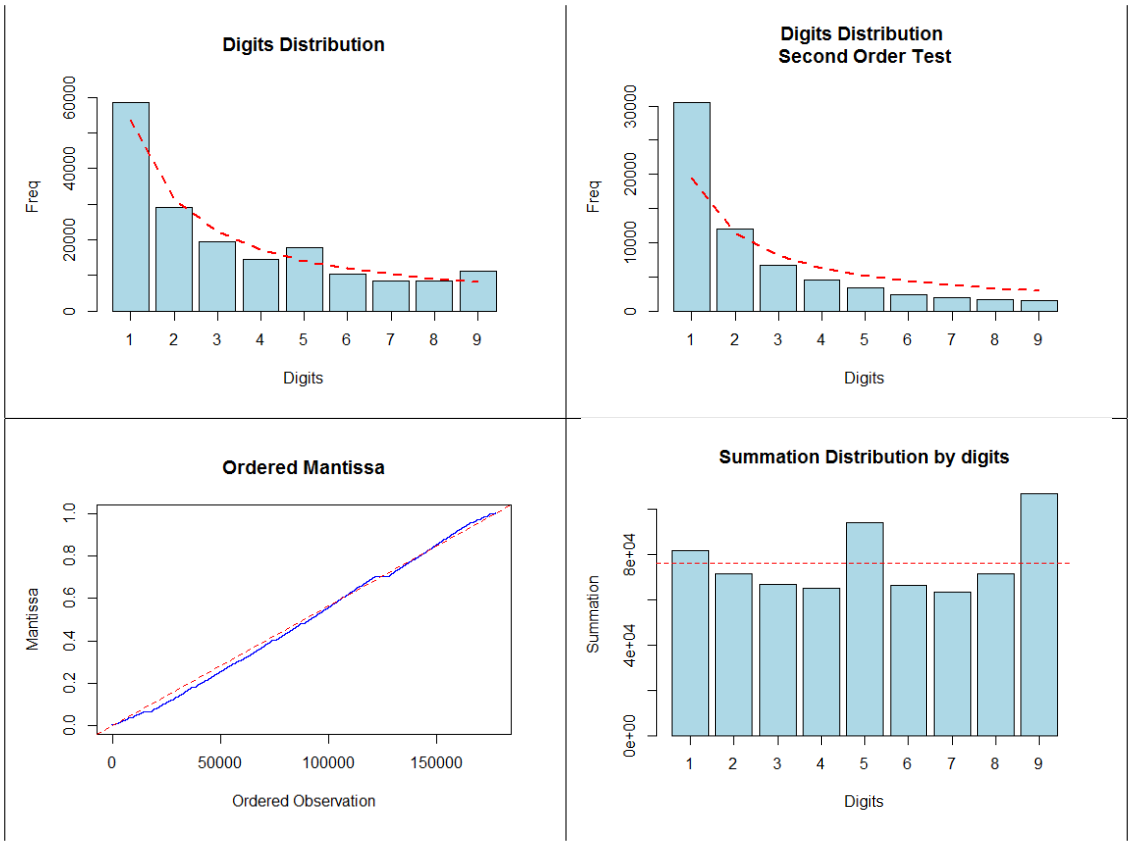
7.2.2 Els dígits de les factures empresarials

Les dades [6] procedien d'una empresa de servei públic de la Costa Oest de l'any 2010, són les factures dels béns i serveis rebuts de venedors. En total, n'hi ha 189.470 observacions, l'import de la factura varia de -71388$ a 26763476$. Analitzarem només els nombres positius més gran o igual que 10$, i al final es queda 177.763 observacions.

Observem la figura 6, al gràfic de la distribució del primer dígit, de simple vista se sembla la dada segueix bastant bé la llei, hi ha petits salts a 1, 5 i 9, però en realitat la conformitat és molt feble, amb un MAD=0.015. La suma de les freqüències relatives d'aquests 3 dígits és 49.2%, és poc pràctic estudiar totes aquestes factures. Al gràfic del dos primers dígits, observem que el salt més gran és al 50, els dígits 10, 11 també representa salts bastant grans. El salts al 98, 99 podrien ser interessant, són nombres sota el llindar 100. Hi ha salts petits al 90 i 92. La distribució dels dos primers dígits representa una tendència general, però no ho segueix, té MAD igual a $0.0024 > 0.0022$. El test del primer dígit i del segon dígits són molt agregats, el test dels dos primers dígits és bastant millor, inclou tota la informació dels altres dos, ja que podem observar els salts dels dos primers dígits és combinació del salt del primer dígit i del segon dígit.
El test del segon ordre conté 64.578 diferències no nuls. La diferència és 0 quan els

[6] La dada tret d'un exemple del Nigrini (2012), es pot trobar a la pàgina web http://www.nigrini.com/ForensicAnalytics.htm a l'apartat de *Benford's Law, Chapter 2-6, Corporate Payment Data*. O també es pot accedir a `R` amb comanda *data(corporate.payment)*.

Figura 6: Els gràfics del primer dígit de la factura

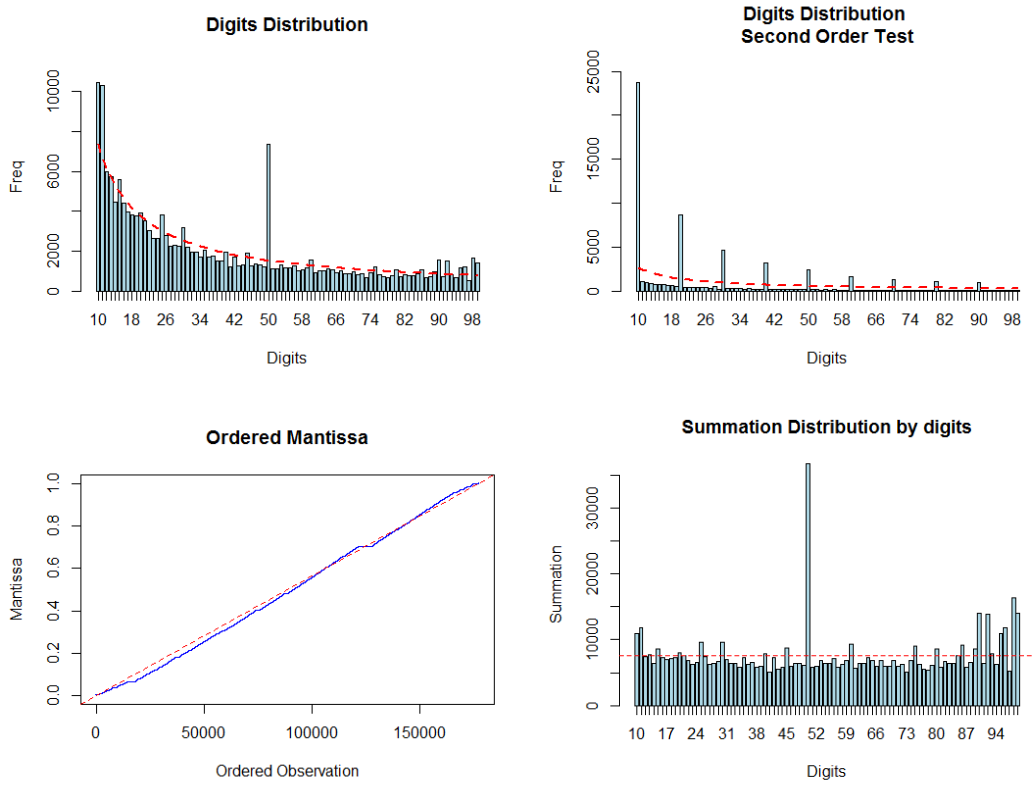


Taula 18: Taula de duplicació

Rank	Nombre	Duplicats	Rank	Nombre	Duplicats
1	50.00	6022	11	30.00	672
2	1153.35	2264	12	250.00	657
3	1083.45	1185	13	200.00	631
4	150.00	1056	14	100.00	624
5	988.35	1018	15	300.00	617
6	1159.35	976	16	45.00	602
7	25.00	956	17	1118.40	578
8	90.00	938	18	1318.35	565
9	928.45	907	19	964.40	559
10	994.35	729	20	1018.30	517

nombres successius de la llista ordenada de la dada són iguals, i els 0's no estan inclòs. El gràfic del segon ordre dels dos primer dígits es veu dos patrons, un amb els dos primers dígits divisibles per 10 (10, 20, $\cdots$, 90), i l'altre no. Hi ha 1319105 observacions estan entre 10 i 999.99, però només hi ha 99000 nombres entre 10 a 99999, vol dir 40% dels nombres a aquest interval estan repetits, els nombres estan

Figura 7: Els gràfics del dos primers dígits de la factura



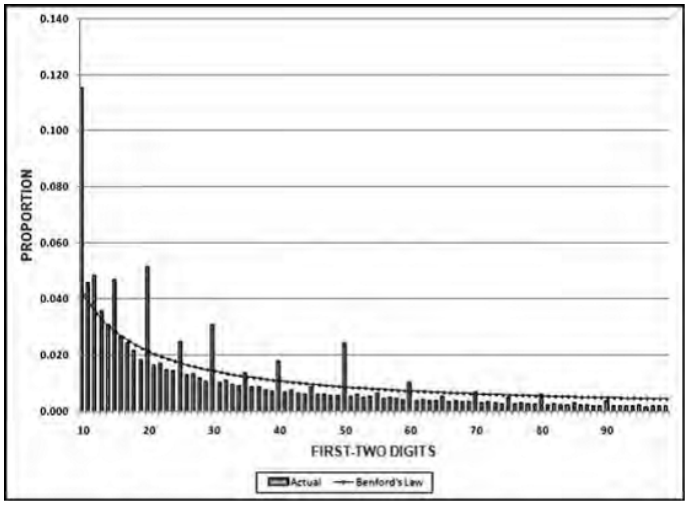
molt centrats a aquest interval. Podem deduir que la diferència dels nombre és relativament petit com 0.01, 0.02, n'hi ha 22374 amb la diferència 0.01, i 8145 amb la diferència 2, etc. La diferència 0.01 provoca el salt 10, i la diferència 20 provoca el salt 20, successivament.
Al test de duplicació, observem que l'import 50 repeteix 6022 vegades, provoca el salt als gràfics. La freqüència relativa és 0.033 supera la probabilitat esperada de 0.009. L'import 1153.35, 1159.35 i 1118.40 provoca el salt del dígit 11, en total hi ha 3818 factures, s'aproxima a la diferència de la probabilitat (0.02·177763=3555.26), vol dir la majoria de les factures que provoca el salt 11 estan aquí. Amb més recerca, mostra que l'import 1153.53 estan agrupats només a 84 dies diferents, necessita més investigació! Una altra trobada és 2263 de les transaccions de l'import 1153.35 es provenen de dos venedors, una revisió més profunda es necessita.

7.2.3 Els dígits dels xecs bancaris

El frau en comptes de xecs és un problema greu per als bancs. Aquests fraus poden ser en forma de signatures falsificades, falsificació de xecs i xecs alterats. La idea era que si els patrons de dígits dels xecs fraudulents diferien dels bons controls, la desigualtat podria ser utilitzat com un indicador addicional de frau. El primer pas va ser buscar en els patrons de dígits dels bons controls per veure si els patrons

van ser consistents en tots els seus centres de processament. Les dades consistien en 13.2 milions de xecs a partir dels 20 centres de processament de EUA. Com la major part de les pèrdues per frau de xecs eren més gran que 100$, seria un pas eficaç per eliminar tots els números de menys de 100$. La supressió dels controls de petites quantitats redueix la bona mostra de verificació a 6700000. Els dígits dels bons xecs es mostren a la Figura .

Figura 8: Els gràfics del dos primers dígits de xecs de bons controls

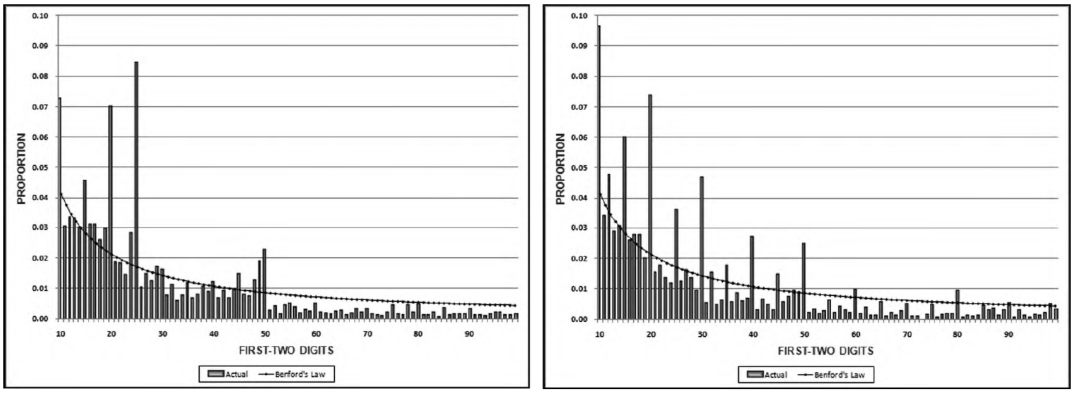


La figura que mostra que els dígits dels bons controls no s'ajusta a la Llei de Benford , amb un MAD de 0.004. Això no va ser un problema ja que l'objectiu era buscar diferències entre els bons i sobre els controls de frau. Els salts van ser en gran part causats per quantitats rodones (50, 100, 200...). Els dos nombres que s'utilitzen amb més freqüència van ser 100$ i 50$, això va ser consistent a totes les regions.

El següent pas era veure si els patrons dels dígits diferien en tots els centres de processament. Els patrons dels dígits per a cada centre es van calcular, i els gràfics es van comparar amb el patró de tot el país a la Figura . Es van calcular les desviacions (MAD) entre les gràfiques de centres de processament i la gràfica en tot el país. Va oscil·lar entre un mínim de 0.0003 i un màxim de 0.0012. Les majors desviacions dels dígits entre els gràfics regionals i el patró de tot el país van ser de 10 i 50.

Els dígits dels xecs fraudulents a la figura 8.8 es diferenciaven dels dígits de la bona xecs, i els patrons dels dígits dels controls de frau van ser diferents en les diferents regions(no demostrada). El següent pas va consistir en calcular la relació entre les proporcions de frau per a cada regió i les proporcions bones per cada dígit. Per exemple, el 25, la proporció real dels controls de frau era (aproximadament) de 0.084, i per als controls bons, la proporció de 25 era 0.025. Això dóna una proporció de 3.36 (0.084 dividit per 0.025), mostra que el 25 va ser molt afavorida

Figura 9: Els gràfics del dos primers dígits de xecs de dos regions



Regió 1 Regió 2

pels defraudadors. Per a la regió 1, els percentatges més alts van ser per 25, 49, 48 i 24, respectivament. Per a la regió 2, els percentatges més alts van ser el 98, 99, 48 i 45, respectivament. Les relacions ens diuen que els números de frau van ser més sovint relacionats amb el nombres de control de llindar. Els defraudadors estaven intentant entrar just sota un control real (potser una quantitat de diners en un caixer de supermercat requereix l'aprovació del supervisor) o una percepció de control. L'últim pas en el projecte va ser identificar els nombres reals associats als dos primers dígits com l'exemple anterioir.

7.3 Quan s'aplica la llei de Benford?

Benford no aportava cap indicació quan els conjunts de dades han de seguir la llei, i què tenen en comú entre ells. És evident que no tots els conjunts de dades satisfan la llei, i mai hi va haver una clara definició d'un experiment d'estadística general que podria predir quines taules faria, i el que no ho faria. Nigrini va proposar unes maneres per veure si un conjunt de nombres reals compleix la llei de Benford:

1. La mostra s'ha de presentar les mides dels fets o esdeveniments.

2. Pel conjunt que els nombres tenen almenys 4 dígits, serà més proper a la llei.

3. Els valors no estan restringida pel mínims o màxims, excepte un mínim de 0 per a les dades que només poden ser positius (resultats de les eleccions, nombre de població). O un mínim de 10 també es permet. L'ànalisi de llei de Benfrod en general no es veu influenciada per un valor mínim que és una potència sencera de 10.

4. Els nombres no són identificacions o etiquetes(número de seguretat, codi postal, número de telèfon).

5. Quan la mitjana és més gran que la mediana i el biaix és positiu. La distribució de la variable ha de ser lleugerament asimètrica positiva, és a dir, ha d'haver un major nombre de valors petits que grans. Per exemple, hi ha més pobles que ciutats grans, hi ha més companyes petites que les grans, la quantitat de rius petits és més gran que rius grans. Tenint en compte, els salaris de la mateixa professió normalment no compleix la llei de Benford, ja que la diferència entre ells no és molt rellevant. Però, els salaris que inclou diferents professions podria ser que seguir la llei.
   No conformitat de Benford podria ser indicadors de:

   - Un conjunt de dades incompletes.
   - La mostra no ser representatius de la població.
   - Arrodoniment excessiu de les dades.
   - Dades anormals o duplicats

8 Conclusions

8.1 Sumari

Un sumari de les propietats de la llei de Benford:

1. La probabilitat d'aparèixer com el primer dígit decreix de l'1 al 9.
2. A mesura que la posició dels dígits es mou cap a la dreta, tendeixen a ser uniformement distribuïts.
3. Els dígits significatius són dependents, la dependència disminueix a mesura que la separació entre ells augmenta.
4. Una succeesió de nombre reals segueix la llei de Benford si el seu logaritme és uniformement distribuït mòdul 1.(Teorema 4.1.1)
5. La distribució de Benford és la única distribució contínua que té els dígits significatius invariants per canvi d'escala.(Teorema 4.2.1)
6. La distribució de Benford és la única distribució que té els dígits significatius invariants per canvi de base. (Teorema 4.3.1)
7. La distribució de Benford és la única distribució que té els dígits invariants per la suma. (Teorema 4.4.1)
8. la distribució de Benford és invariant sota l'inversió i la multiplicació per qualsevol distribució.(Proposicions4.1.2 i 4.1.3)

8.2 Limitacions de l'aplicació de la llei

Una de les limitacions de la llei és que no distingeix els nombre que tenen el mateix significand, com 30 o 300, tots dos tenen el primer dígit 3, i els altres dígits 0.

Els resultats de l'aplicació de la llei de Benford no necessàriament indiquen que hi ha frau en la dada analitzada, simplement ens dóna un senyal d'alerta d'on posar més atenció. Els resultats s'han de valorar d'acord a les particularitats del cas. Considerar què tipus d'operacions són, en quines dates es realitzen, amb quina freqüència es realitzen i si estan d'acord a normes de la institució financera. Es complementen amb les altres tècniques estadístiques: anàlisi de regressió i correlació, anàlisi de dispersió, anàlisi de patrons i seqüències, anàlisi de faltants i duplicats, anàlisi històrica de tendències, etc.

La facilitat d'entendre i aplicar la llei també és un defecte, els defraudadors poden aprendre ràpidament, i prendre cures especials a l'hora de fer canvi de les entrades a la taula de comptes amb la finalitat de conformar la llei de Benford.

Hi ha altres situacions en les quals no compleixen la llei de Benford. Alguns casos particulars:

- Alçada dels habitants. La majoria tenen alçada entre 100 i 200 centímetres, en aquest rang només hi ha primer dígits un, i el segon dígit 4,5,6,7,8. Per tant, l'alçada no és Benford.
- El resultat de loteria que són totalment aleatoris. En general cap joc d'atzar compleix la llei de Benford.
- Els nombres són assignats o etiquetes que tenen caràcter identificats (número de seguretat, codi postal, número de telèfon, número de document identitat), no són generats naturalment.

8.3 Possibles ampliacions

Al meu treball, només he pogut explicar els conceptes bàsics, les caràcterístiques fonamentals i unes aplicacions de la llei Benford. Aquesta distribució logarítmica és molt més complex i fort que ho sembla aparentment. La llei de Benford pot ampliar als processos lineals multidimensionals i molt més. Una gran varietat de dades reals que segueixen la llei de Benford poden usar per detectar els canvis en el fenomens naturals, el fraus ecomòmics o les utilitzacions tecnològics, la seva aplicació queda molt per descobrir.

Referències

[1] Berger, Arno.; Hill, Theodore P.: A basic theory of Benford's Law, *Probability Surveys.*, Vol.8 (2011). pp 1-126.

[2] Nigrini, Mark J.: Benford's Law. Applications for Forensic Accounting, Auditing, and Fraud Detection, *John Wiley & Sons. Hoboken, New Jersey.*, 2012.

[3] Berger, Arno; Hill, Theodore P.:A introduction to Benford's Law, *Princeton University Press*, 26 de maig, 2015

[4] Elise, Janvresse; Thierry, de la Rue.: La llei de Benford, *Societat Catalana de Matemàtiques.*, Vol.24, núm.1, 2009. pp 5-12.

[5] Manual, Perera Domínguez; Juan, David Ayllón Burguillo.: El primer dígito significativo.

[6] Hill, Theodore P.: The Significant-Digit Phenomenon, *Amer. Math. Monthly.*, Vol.102, 322-327, 1995b.

[7] Hill, Theodore P.: The First Digit Phenomenon, *Amer. Sci.*, Vol.86, 358-363, 1998.

[8] Hill, Theodore P.: A statistical Derivation of the Significant-Digit Law, *Statistical Science.*, Vol. 10, No.4, 354-363, 1995.

[9] Li, ZhiPeng; Cong, Lin; Wang, Huajia.: Discussion on Benfrod's law and its application,
`arXiv:math/0408057v2 [math.ST]`, 4 Oct 2004.

[10] Adrian Jamain.: Benford's Law, *Imperial College of London Department of Mathematics; Ecole Nationale Superieure d'Informatique et de Mathematiques Appliquees de Grenoble.*, abril-Septembre, 2001.

[11] Wikipedia contributors.: Benford's law. *Wikipedia, The free Encyclopedia*, 27 de juny, 2016,
`https://en.wikipedia.org/w/index.php?title=Benford%27s_law&oldid=727201723`.

[12] Wikipedia contributors.: Frank Benford. *Wikipedia, The free Encyclopedia*, 6 de maig, 2016,
`https://en.wikipedia.org/w/index.php?title=Frank_Benford&oldid=718918647`.

[13] Nigrini, MarK j.:Data Analysis Technology for the Audit Community,
`http://www.nigrini.com/`.

Annex: Implementacions i codi

Com ja hem dit a la introducció, hi ha dos paquets de R sobre la llei de Benford: "BenfordTests" i "benford.analysis".

El paquet "BenfordTests" conté diverses proves estadístiques especialitzades i funcions de suport per determinar si les dades numèriques podrien ajustar-se a la llei de Benford. Totes les proves es poden realitzar usant més d'un dígit. Totes les proves simulen els valors específics requerits per a la inferència estadística, mentre que els p-valors per als tests estadístiques també es poden determinar fent servir les seves distribucions asimptòtiques. Al meu treball utilitzo especialment el test de khi-quadrat.

El paquet "benford.analysis" proporciona eines que fan més fàcil per validar les dades usant la Llei de Benford. El propòsit principal del paquet és identificar dades sospitosos que necessiten una verificació addicional.

Volem advertir que hem trobat el que sembla un error en un dels gràfics que produeix la funció plot(benford(data)): "Summation Distribution by digits". Nigrini dóna una definició aparentment és incorrecta, i en el paquet com que segueixen el llibre de Nigrini, fan gràfic respecte d'això. La definició del test de la suma que va posar Nigrini és diferent de la definició 4.4.1. Si una successió de nombres reals és Benford, hauría de complir les sumes dels significands dels nombres que tenen el mateix dígit són iguals en lloc de ser la suma dels nombres (un nombre gran influeix molt en el cas de la suma dels nombres). Si els nombres estan entre 0 i 10, sí que és indiferent la suma dels significands dels nombres o la suma dels nombres. Com que ens interessa és detectar si les dades segueixen Benford, es surti el test de sumació de la distribució de Benford sigui distribuïts uniforme, agafem la definició que tenim nostres. El codi per fer el càlcul és següent:

```
library(benford.analysis)
library(BenfordTests)
####Summation Distribution by digits####
##El primer dígit##
sum<-rep(0,9)
vec1<-vec
for(j in 1:9){
  for(i in 1:length(vec1)){
    vec1[i]<-vec1[i]/(10^(floor(log10(vec1[i]))))
      if(signifd(vec1[i])==j) {
          sum[j]<-sum[j]+vec1[i]
    }
  }
 }
sum1(vec)
barplot(sum,ylim=c(0,max(sum)),col="lightblue",main="
   Summation Distribution by digits", xlab="Digits", ylab
   ="Summation",names.arg = c(1:9),axis.lty=1)
```

```
abline(h=mean(sum),col="red", lty=2)

##El dos primers dígits##
sum2<-rep(0,90)
vec2<-vec
for(j in 10:99){
  for(i in 1:length(vec2)){
    vec2[i]<-vec2[i]/(10^(floor(log10(vec2[i]))))
    if(signifd(vec2[i],2)==j) {
     sum2[j-9]<-sum2[j-9]+vec2[i]
    }
  }
}
barplot(sum2,ylim=c(0,max(sum2)),col="lightblue",main="
   Summation Distribution by digits", xlab="Digits", ylab
   ="Summation",names.arg = c(10:99),axis.lty=1)
abline(h=mean(sum2),col="red",lty=2)

###Figura 2###
benford <- log10(1+1/(1:9))
plot(benford, col="red", main="La distribució del primer
   dígit de la llei de Benford",
xlab="Dígits", ylab="Percentatge", type="p", pch=19, xaxt
   ="n")
axis(side=1,at=c(1:9))

segondigit<-NULL
for(i in 0:9) {
  segondigit<-c(segondigit,p.this.digit.at.n(i,2))
}
plot(cbind(0:9,segondigit),col="red", main="La
   distribució del segon dígit", xlab="Dígits", ylab="
   Percentatge", xaxt="n",type="p", pch=19, ylim=c
   (0,0.15))
axis(side=1,at=c(0:9))

tercerdigit<-NULL
for(i in 0:9) {
  tercerdigit<-c(tercerdigit,p.this.digit.at.n(i,3))
}
plot(cbind(0:9,tercerdigit),col="red",main="La
   distribució del tercer dígit", xlab="Dígits", ylab="
   Percentatge", xaxt="n",type="p", pch=19, ylim=c
   (0,0.15))
axis(side=1,at=c(0:9))
```

```
primersdosdigits<-log10(1+1/(11:99))
plot(cbind(11:99,primersdosdigits),col="red", main="La
   distribució dels dos primers dígits", xlab="Dígits",
   ylab="Percentatge", xaxt="n",type="p", pch=19)
axis(side=1,at=c(10:99))

##Taula2:Esperança##
esp<-rep(0,9)
for(m in 1:9) {
   for(i in 1:9){
      esp[m]<-i*p.this.digit.at.n(i,m)+esp[m]
   }
}
print(esp,digits=12)

##Variància##
var<-rep(0,9)
for(m in 1:9) {
  for(i in 1:9){
    var[m]<-i^2*p.this.digit.at.n(i,m)+var[m]
  }
  var[m]<-var[m]-esp[m]^2
}
print(var,14)

###Taula 3###
benf=matrix(NA,10,9)
for(i in 1:9){
  for(j in 0:9){
    benf[j+1,i]<-p.this.digit.at.n(j,i)*100
  }
}


round(benford(vec,1)$bfd$data.dist,4)*100#calcular la
   freqüència relativa del primer dígits
benford(vec,1)$MAD
max(abs(benford(vec,1)$bfd$data.dist-benford(vec,1)
   $bfd$benford.dist))
benford(vec,1)#obté els valors de khi-quadrat, el seu p-
   valor, el test d'arc de mantissa.
Aquestes 4 comandes són utilitzats per les succession
   següent:
###Potència###
vec<-NULL
n<-100#n=100,1000
```

```
for(i in 1:n) {
  vec[i]<-2^i
}
vec<-NULL
n<-1000
for(i in 1:n) {
  vec[i]<-2^i+7*1^i
}
##(1)##
a<-1
r<-1.0002303
vec<-NULL
vec[1]=1
for(i in 2:10000){
  vec<-c(vec,a*r^i)
}
##(2)##
a<-3
r<-1.34
vec<-NULL
vec[1]=3
for(i in 1:1000){
  vec<-c(vec,a*r^i)
}

##fibonaci##
vec-NULL
n<-100#n=1000
vec<-numeric(n)
vec[1]<-1
vec[2]<-1
for(i in 3:n){
  vec[i]<-vec[i-1]+vec[i-2]
}
##Factorial##
vec<-NULL
n<-50#n=100,n=150
for(i in 1: n){
  vec[i]<-factorial(i)
}

##Exponencial##
vec<-NULL
a=1#a=0.5, a=0.7
b=0#b=2
n=100#n=500
```

```
for(i in 1:n){
  vec[i]=exp(a*i+b)
}

##nombre primers##
library(matlab)
i<-1
n=100#n=100
while(length(primes(i))<n) {i<-i+1}
vec<-primes(i)
vec

##Distribucions comuns##
library(stats)
set.seed(2015)
vec1<-rbenf(1000)
set.seed(2016)
vec2<-runif(100)
vec3<-runif(1000)
vec4<-rexp(100,1)
vec5<-rexp(1000,1)
vec6<-rnorm(1000,1)
vec7<-rbeta(1000, 1/2,1/2)
vec8<-rgamma(1000,1.5,1.5)

vec9<-vec1*vec3
vec10<-vec1*vec5
vec11<-vec3/vec1
vec12<-vec1/vec5
vec13<-vec6*vec3*vec5
vec14<-vec6*vec3*vec5*vec7*Vec8

####Municipis d'Espanya####
vec<-NULL
vec<-ts(read.table("espanyamunicipio.txt"))
plot(benford(vec,1))
plot(benford(vec,2))
round(benford(vec,1)$bfd$data.dist,4)*100
benford(vec,1)$MAD
max(abs(benford(vec,1)$bfd$data.dist-benford(vec,1)
   $bfd$benford.dist))
benford(vec,1)
benford(vec,1)$bfd$data.dist*100
plot(benford(vec,1),except = "none",multiple=FALSE)

####Les provincies d'Espanya####
```

```
vec<-NULL
vec<-ts(read.table("espanyaprovincia.txt"))
Els valors es calculen de la mateix manera que l'anterior
   .

#####Els països del món######
vec<-NULL
vec<-ts(read.table("mundo.txt"))
Els valors es calculen de la mateix manera que l'anterior
   .


###Factures###
vec0<-read.csv("Còpia de CorporatePaymentsData-1.csv",dec
   =",",header = TRUE, sep=";")
dim(vec0)#189470
min(vec0$Amount)#-71388
max(vec0$Amount)#26763476
vec<-vec0$Amount[vec0$Amount>=10]
length(vec)#177763
benford(vec,1)$bfd$data.dist[1]+benford(vec,1)$bfd$data.
   dist[5]+benford(vec,1)$bfd$data.dist[9]#49.2%
benford(vec,1)$MAD#0.01464187
benford(vec,2)$MAD#0.002429906
plot(benford(vec,1))
plot(benford(vec,1),except = "none",multiple=FALSE)
plot(benford(vec,2),except = "none",multiple=FALSE)
benford(vec,1)
vec2<-benford(vec,2)$s.o.data$second.order
length(vec2)#64578
for(i in 1:9){
  print(length(vec2[vec2==0.01*i]))
}
vecD<-length(vec[vec<=999.99])#139105 obs
duplicatesTable(benford(vec,1))[c(1:20),]
(benford(vec,1)$bfd$data.dist[1]-benford(vec,1)
   $bfd$benford.dist[1])*177763#4977.005
summary(vec0$Date[vec0$Amount==1153.35]) #84 dies
   differents
table(vec0$VendorNum[vec0$Amount==1153.35])#3 venedors, 1
    d'ells només amb 1 import 1153.35
```