

L'1 el més acaparador

Inés López, Pau Campuzano i Carla Ruiz

Índex

- ▶ Introducció
- ▶ Objectiu
- ▶ Hipòtesi
- ▶ Marc teòric
- ▶ Recerca d'informació
- ▶ Com ho farem?
- ▶ Mida de la mostra
- ▶ Tria de la mostra
- ▶ Comptem les xifres
- ▶ Loteries: euromillones
- ▶ Loteries: primitiva
- ▶ Loteries: joker (primitiva)
- ▶ Natalitat
- ▶ Eleccions als EEUU
- ▶ Eleccions a Veneçuela
- ▶ Dades d'en Barcenás

Introducció

Avui dia, només sentim a parlar de dades, de “big data”, d'ordinadors molt potents que treballen amb quantitats immenses de dades. Dades per estadístiques, dades per desxifrar el genoma humà, dades de resultats electorals, dades obertes, dades fiscals, dades de tot tipus i molt diverses.

Ens hem preguntat si aquestes dades segueixen alguna norma o algun patró.

Objectiu

- ▶ El principal objectiu del treball és posar a prova la Llei de Bendford. Analitzar diferents grups de dades per a comprovar que és el que diu aquesta llei.



Hipòtesi

1. La Llei de Benford es complirà al conjunt de dades/xifres de cada municipi
2. La Llei de Benford es complirà per el conjunt de global de les xifres dels municipis
3. Les dades de loteries no seguiran la Llei de Benford.
4. Les dades de les loteries seguiran les Lleis de Probabilitat.
5. Les dades dels resultats electorals dels Estats Units no seguiran la Llei de Benford.
6. Les dades dels resultats electorals de Veneçuela per a la presidència no seguiran la Llei de Benford

Hipòtesi

7. Les dades sobre les eleccions alemanyes l'any 1933, en què el partit Neonazi, amb Hitler al capdavant no compliran la Llei de Benford
8. La natalitat serà una dada que al igual de la població seguirà la Llei de Benford.
9. Les dades sobre la comptabilitat d'en Barcenas, no compliran la Llei de Benford.
10. Altres dades d'Idescat seguiran la Llei de Benford

Marc teòric: descobriment de la llei

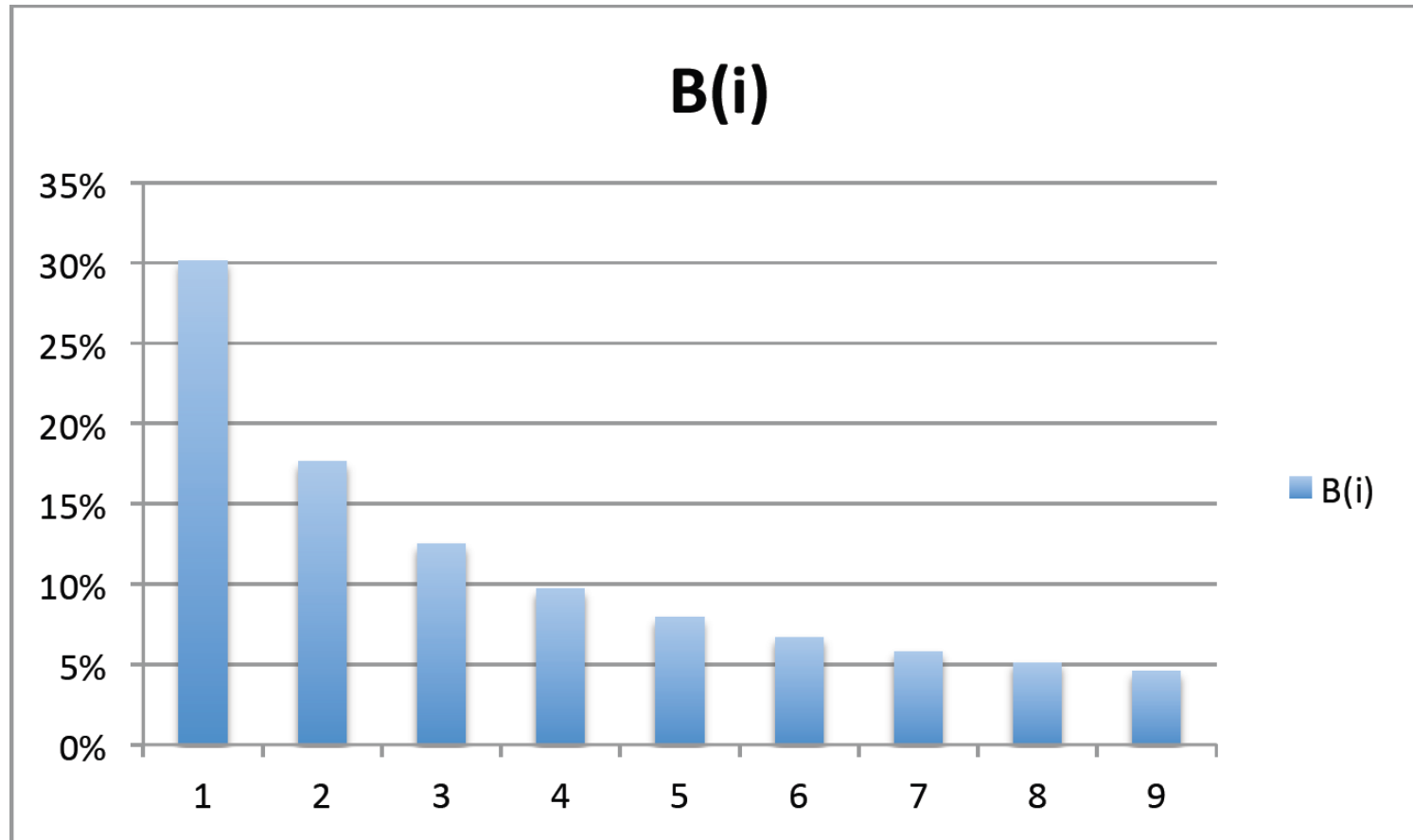
- ▶ La llei de Benford tracta dels nombres que comencen per 1, els quals apareixeran de forma més freqüent que els números començats per altres dígit. La distribució és asimètrica, es a dir, la freqüència estimada pels números començats per 1 és del 30%, per dos del 17%, per 3 el 12% i va disminuint.
- ▶ La primera persona que ho va descobrir va ser Simon Newcombe. Aquest descobriment va quedar en l'oblit fins que va arribar Frank Benford.

Marc teòric: funcionament de la llei

- ▶ La llei de Benford és una llei que tracta l'aparició del primer dígit d'un nombre, de l'1 al 9. es fa partint d'una distribució asimètrica.
- ▶ La llei de Benford és invariable a canvis d'escala i de base.
- ▶ $p(d) = \log((d+1)/d) = \log(d+1) - \log(d)$ $d \in \{1, 2, \dots, 9\}$

Dígit	1	2	3	4	5	6	7	8	9
P(d)	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Marc teòric: Gràfica Llei de Benford



Marc teòric: Detecció de frau

- ▶ Si tractem de falsificar la declaració de la renda, ens haurem d'inventar qualsevol dada, al fer això la tendència a utilitzar números a meitat d'escala és molt alta (5, 6, 7,...) i no a començar per l'1.



Recerca d'informació

- ▶ Per posar a prova la llei de Benford, el que hem fet és buscar dades que no siguin aleatòries i dades que sí ho siguin.
- ▶ Les dades que hem utilitzat a la recerca les hem tret de l'Idescat i d'altres webs (diaris, web oficials d'altres països...).



Com ho farem?

▶ **MUNICIPIS**

- ▶ El que hem fet és fer una tria a l'atzar dels pobles de Catalunya.
- ▶ Seguidament, mirarem les dades si compleixen la llei de Benford

Com ho farem?

- ▶ A Catalunya hi ha 948 pobles, repartits en 42 comarques
- ▶ Com no podem analitzar totes les dades perquè son moltes i necessitaríem molt de temps, hem fet servir una mostra dels municipis per veure que passaria.

Mida de la mostra

- ▶ La calculadora per a calcular la mostra ens diu que la nostra seria de 105 pobles. Amb un nivell de confiança del 95% i un interval de confiança de 9.

Determine Sample Size

Confidence Level: 95% 99%

Confidence Interval:

Population:

Sample size needed:

Tria de la mostra

- ▶ Per a triar la mostra, el que farem serà llistar tots els pobles en un excel i fer un “aleatorio entre” d’aquesta manera tots els municipis tenen la mateixa probabilitat de ser triats. ($948/105 = 8.8$)



Tria de la mostra

- ▶ Fem aleatorio entre 1 i 948, ens ha sortit 209. El primer poble és Catellgalí.
- ▶ Tornarem a fer “aleatorio entre 1-9”. Ens ha sortit: 7.
- ▶ Cada 7 municipis el triarem, així fins a obtenir els 105 municipis que necessitem. (al final hem agafat 106 pobles).

Pobles triats

209	Castellgalí	580	Prades	391	Llimiana	762	Santa Perpètua de Mogoda
216	Castellserà	587	Preixana	398	Lloret de Mar	769	Saus, Camallera i Llampaiés
223	Cellera de Ter, la	594	Puigdàlber	405	Maials	776	Senterada
230	Cervià de les Garrigues	601	Quar, la	412	Margalef	783	Seu d'Urgell, la
237	Coll de Nargó	608	Regencós	419	Masies de Roda, les	790	Soleràs, el
244	Conesa	615	Ribera d'Ondara	426	Masroig, el	797	Subirats
251	Cornellà de Llobregat	622	Ripollet	433	Mediona	804	Talarn
258	Cubells	629	Riudoms	440	Moià	811	Tarroja de Segarra
265	Escala, l'	636	Roquetes	447	Molsosa, la	818	Tiana
272	Esponellà	643	Rupit i Pruit	454	Montclar	825	Tordera
279	Esterrí d'Àneu	650	Salomó	461	Montmajor	832	Torre de l'Espanyol, la
286	Figaró-Montmany	657	Sant Andreu de Llavaneres	468	Montornès del Vallès	839	Torrelavit
293	Flix	664	Sant Carles de la Ràpita	475	Morell, el	846	Torroella de Montgrí
300	Fonollosa	671	Sant Esteve de la Sarga	482	Navàs	853	Ullà
307	Forès	678	Sant Feliu de Pallerols	489	Òdena	860	Vacarisses
314	Gaià	685	Sant Guim de la Plana	496	Olius	867	Vallbona de les Monges
321	Garrigàs	692	Sant Jaume dels Domenys	503	Omells de na Gaia, els	874	Vallirana
328	Gimenells i el Pla de la Font	699	Sant Julià de Cerdanyola	510	Os de Balaguer	881	Vandellòs i l'Hospitalet de l'Infant
335	Gombrèn	706	Sant Llorenç de Morunys	517	Palau d'Anglesola, el	888	Vespella de Gaià
342	Granyanella	713	Sant Martí Sarroca	524	Palma d'Ebre, la	895	Vilablareix
349	Guiamets, els	720	Sant Pau de Segúries	531	Parlavà	902	Viladrau
356	Horta de Sant Joan	727	Sant Pol de Mar	538	Perafita	909	Vilaller
363	Ivars d'Urgell	734	Sant Sadurní d'Osormort	545	Pineda de Mar	916	Vilanant
370	Juià	741	Santa Cecília de Voltregà	552	Planes d'Hostoles, les	923	Vilanova de Sau
377	Lladorre	748	Santa Eulàlia de Riuprimer	559	Pobla de Mafumet, la	930	Vila-sacra
384	Llardecans	755	Santa Maria de Besora	566	Pont de Bar, el	937	Vilella Alta, la
				573	Ponts	944	Vinaixa

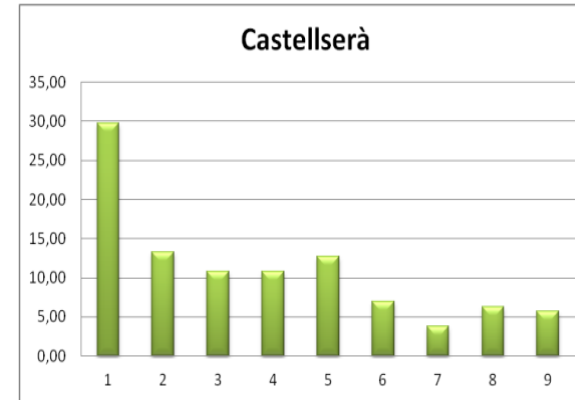
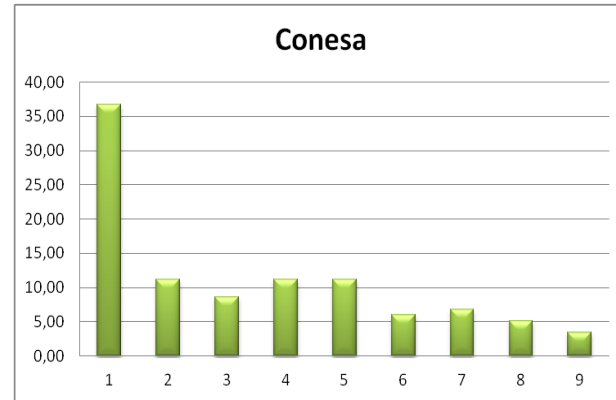
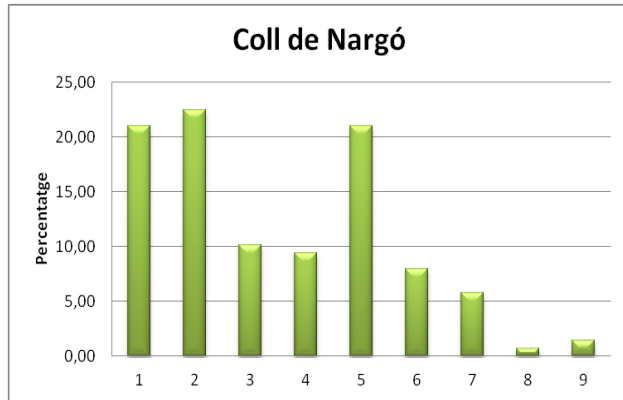
Com comptem les xifres

- ▶ Ens baixarem totes les dades de cada municipi de la pàgina d'Idescat.
- ▶ Separarem els nombres en columnes per a poder comptar les xifres i en farem la gràfica en percentatges.
- ▶ Per a fer això ens hem descarregat els 106 excels d'Idescat, hem comptat les primeres xifres i les hem ajuntat en un excel

El municipi en xifres -> Bages	
Castellgalí (080615)	
El municipi en xifres En Excel	
Població	Treball
Donants	Relació amb l'activitat econòmica
Sexe	Altres ocupacions sobre ocupació resident
Tipus d'edat	Atenció a l'1.1. sector industrial de la finca
Grups etnics: Homes	Atenció al règim general de la L.
Grups etnics: Dones	Atenció al règim d'autònoms de la L.
Lloc de naixement	Atenció al règim gen. per jubilats del centre
Nacionalitat	Atenció de cotització L.
Nacionalitat: Homes	Persones de 1.1. sobre existència del pensionista
Nacionalitat: Dones	Altres ocupats per sector
Tipus d'ocupació	Altres ocupats per sector
Població ocupada a l'estranger	
Mobilitat obligada per subvenció	
Religiositat	Cultura i esports
Religiosos	Equipament del catòlic
Madrinats	Biblioteques
Madrinats: Homes	Esports residents
Madrinats: Dones	
Immigració estrangera	Eleccions
Eleccions (percentatge de la població)	Eleccions municipals
	Eleccions al Parlament de Catalunya
	Eleccions al Congrés dels Diputats
	Eleccions al Parlament Europeu
Habitatges i llocs	
Tipus d'habitatge	Medi ambient
Edifici de tipus d'ús d'habitatge	Residus municipals
Superfície útil dels habitatges	Residus industrials
Macromagnituds i impostos	
Impost sobre la renda de les persones físiques	
Impost de béns immobles urbans	
Impost sobre el patrimoni de les persones físiques	
Sectors econòmics	
Superfície agrícola irrigada	
Tipus d'irrigació	
Tipus de terres	
Tipus de ramaderia	
País de origen	
Construcció d'habitatges	
Abitament turístic	
País de vehicles	

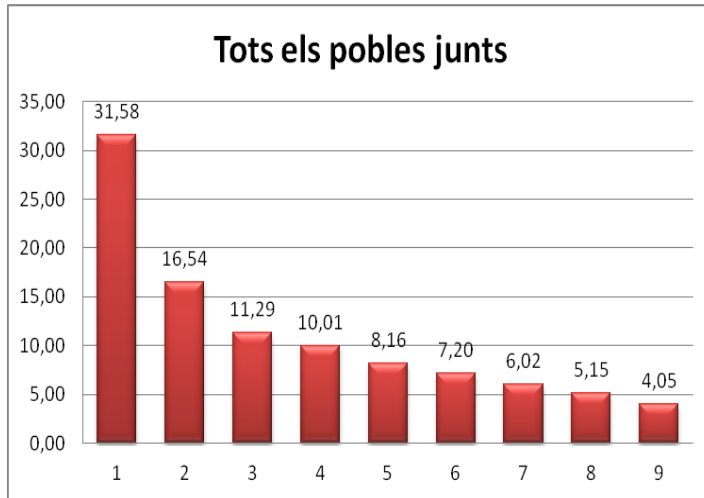


Gràfiques d'alguns pobles

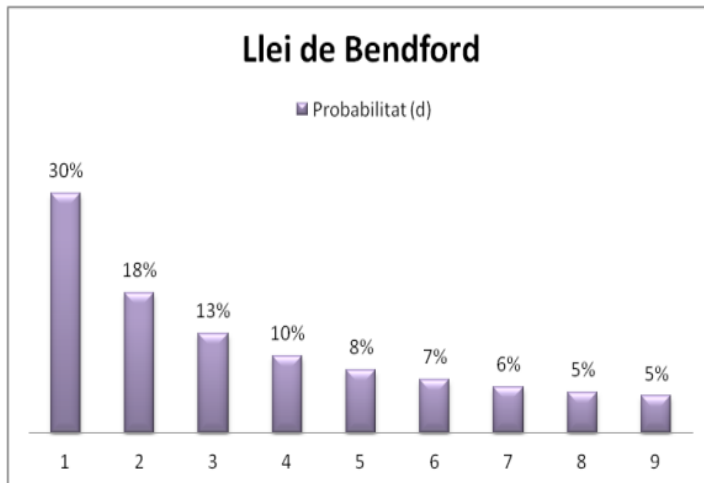


- ▶ No sembla que segueixin la Llei de Benford, en totes les seves xifres, sinó que només passa en l'1 i a més tampoc s'acosta al 30%.

Gràfiques de l'acumulat de pobles



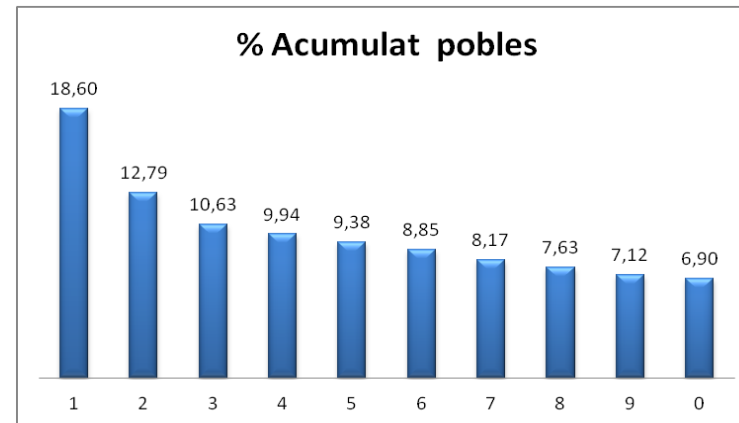
Xifres	Totals	%
1	4920	31,58
2	2577	16,54
3	1759	11,29
4	1560	10,01
5	1272	8,16
6	1122	7,20
7	938	6,02
8	802	5,15
9	631	4,05
Total	15581	



- Si mirem l'acumulat de tots els pobles resulta que sí compleixen la Llei de Benford

Gràfiques de l'acumulat de pobles

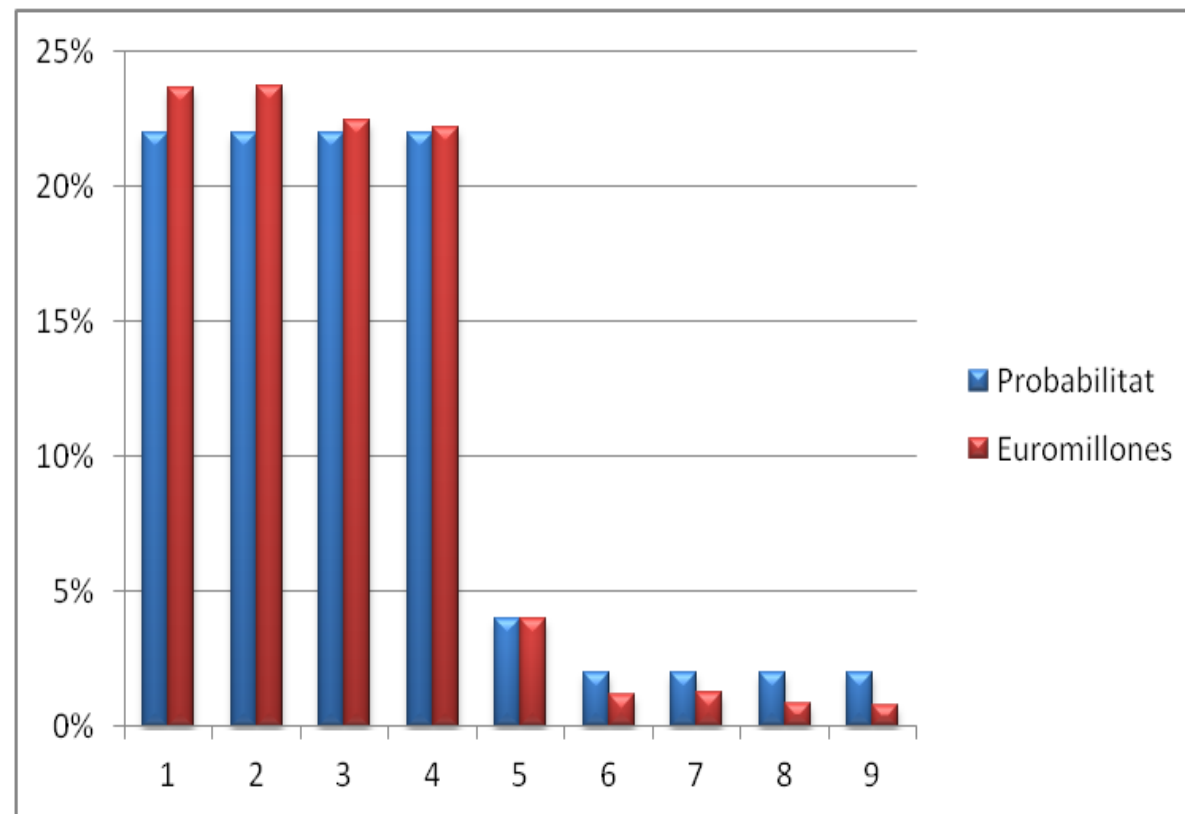
- ▶ Si amb les dades dels municipis, enlloc de comptar la primera xifra, les comptem totes inclosos els zeros obtenim aquesta gràfica:



- ▶ Observem que fent-ho d'aquesta manera no segueixen la Llei de Benford perquè els percentatges són lleugerament diferents de les probabilitats de Benford
- ▶ Però segueixen un ordre decreixen de percentatges com la llei de Benford. **CURIOSITAT**

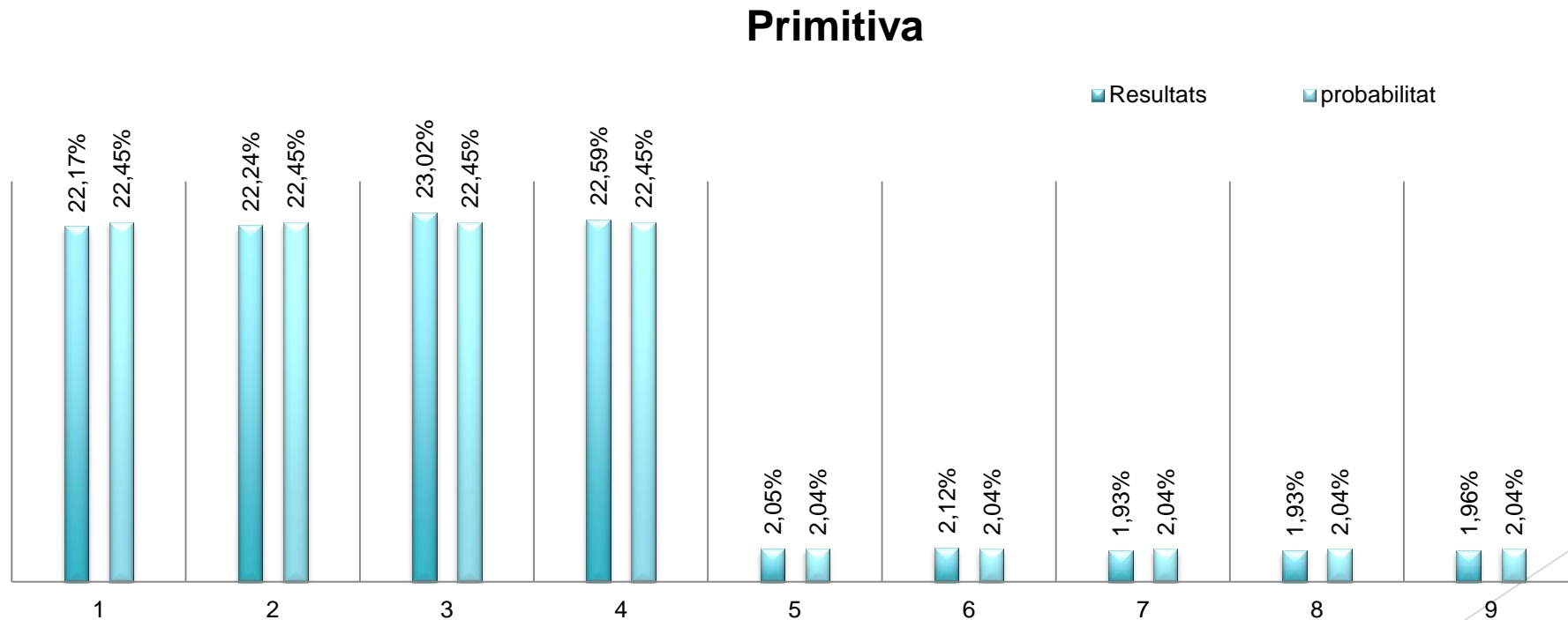
Loteries: euromillones

- ▶ Els percentatges de probabilitat s'acosten molt, per tant, podem afirmar que no ens fan trampa amb els resultats.



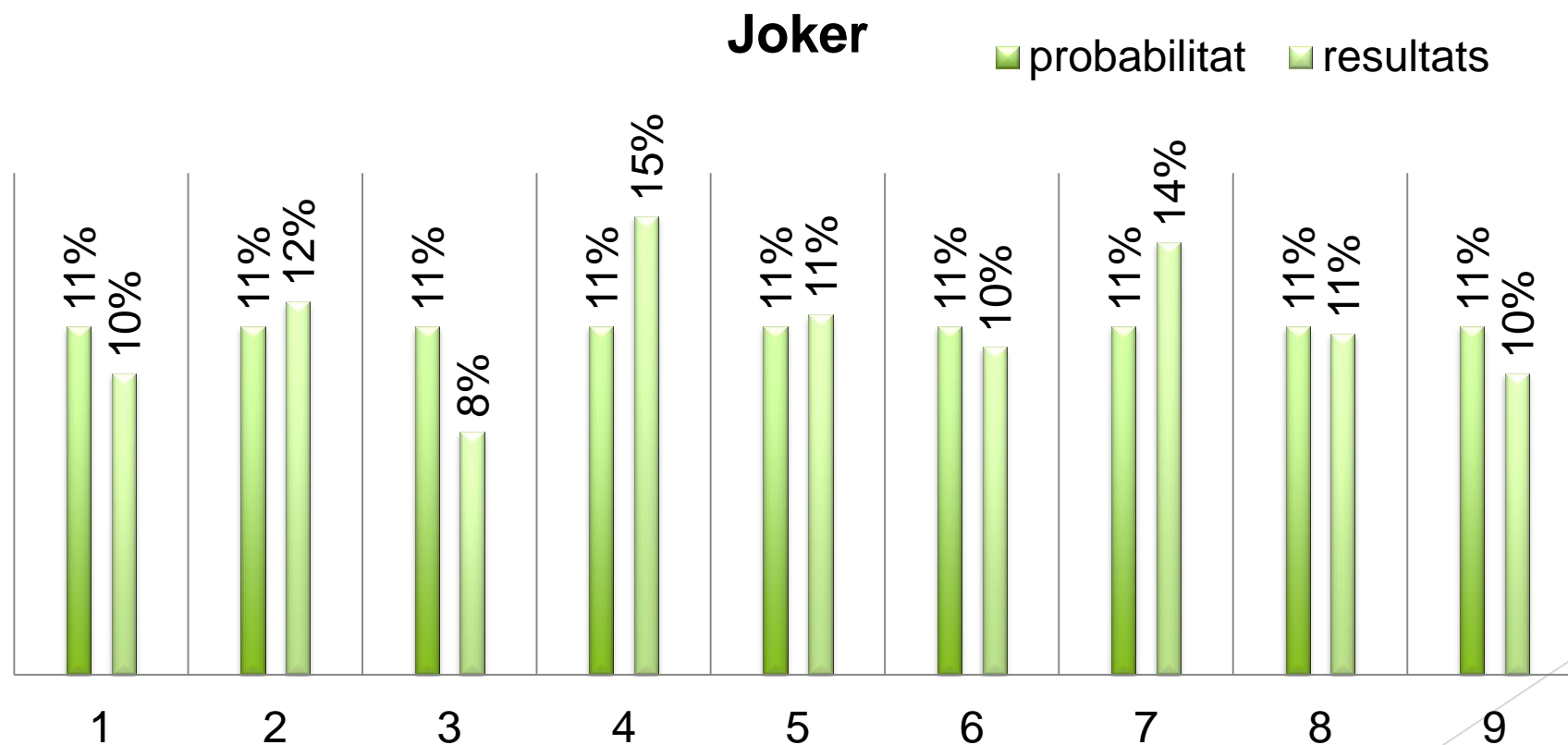
loteries: primitiva

- ▶ Al igual que la primitiva, la gràfica és molt semblant, també veiem que els percentatges s'apropen molt a la probabilitat.



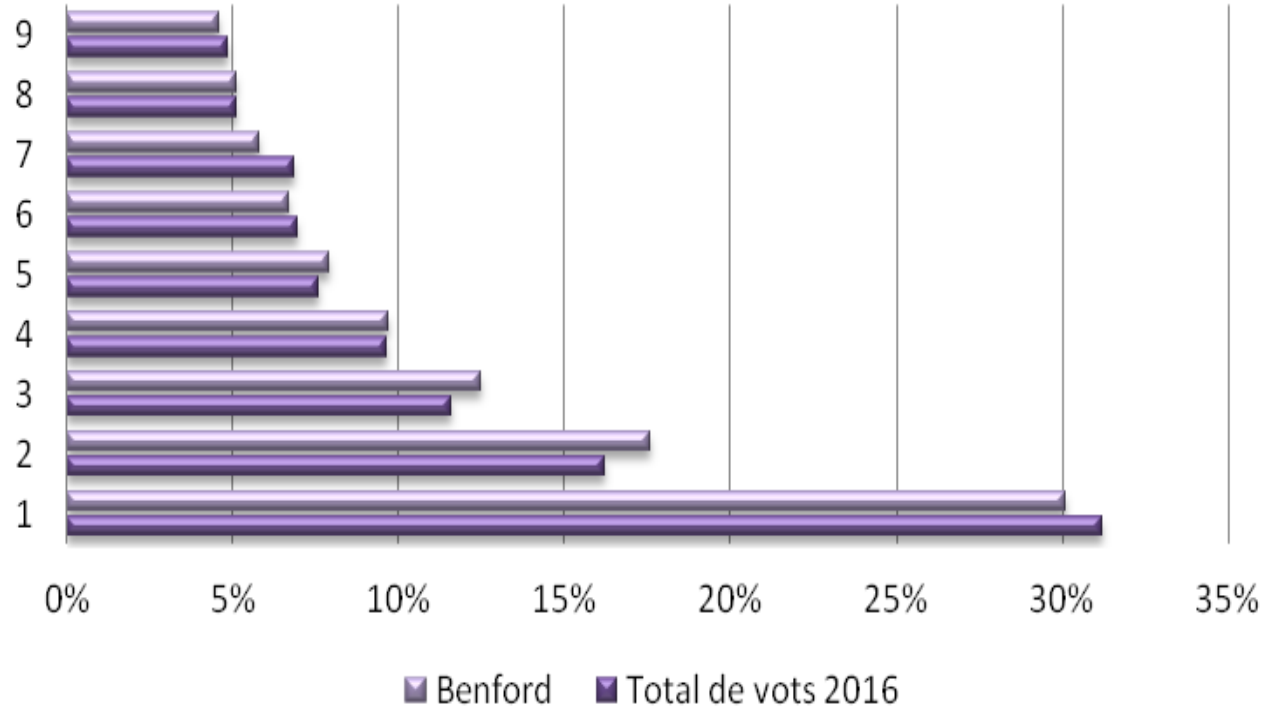
Loteries: joker(primitiva)

- ▶ En aquesta gràfica de la primitiva podem veure una gran irregularitat.
- ▶ No segueix la llei de Benford, segueix la Llei de probabilitat.



Eleccions dels EEUU

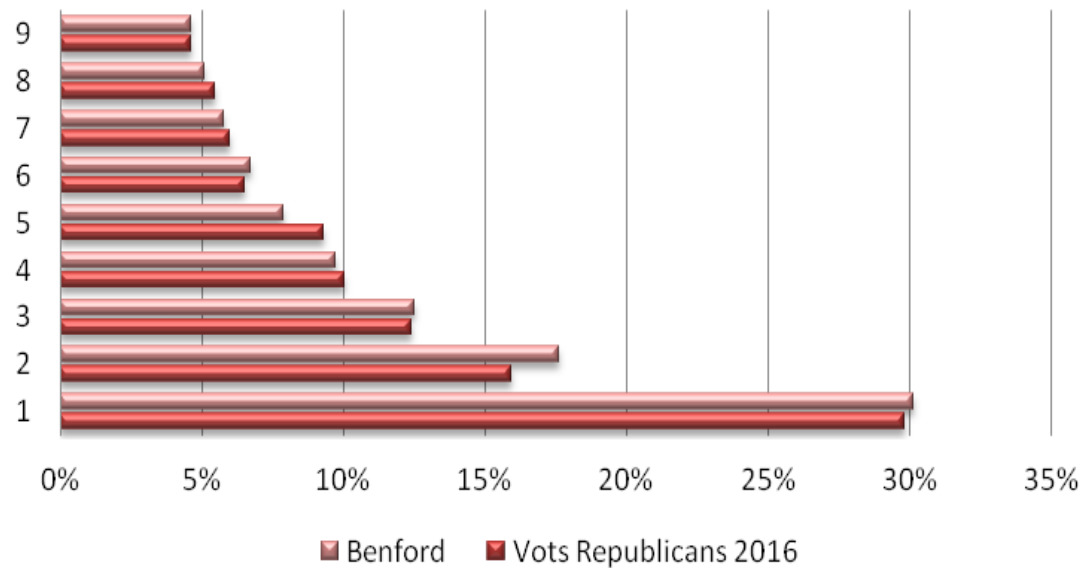
Total vots 2016



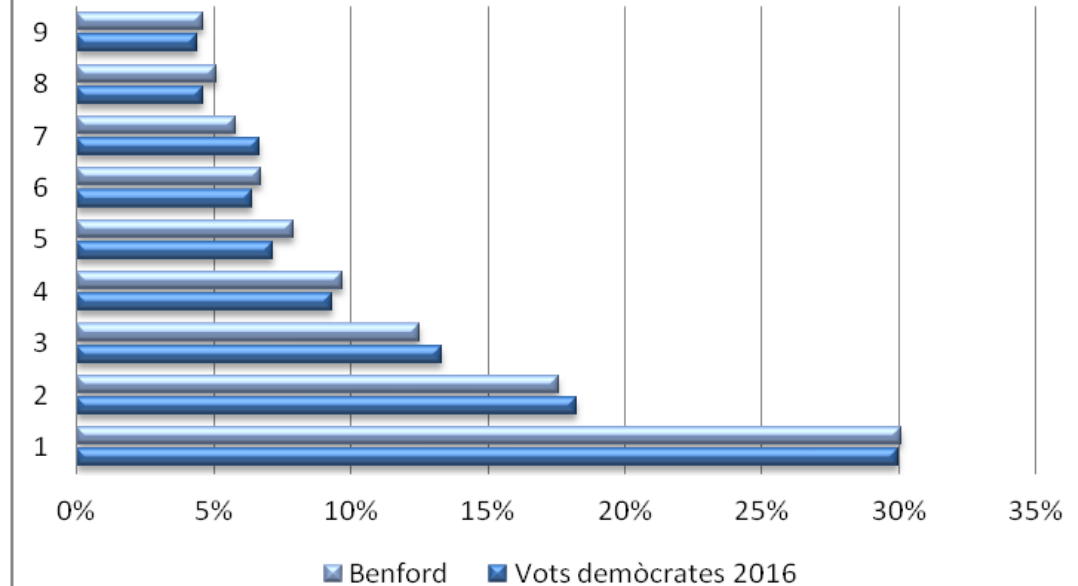
- ▶ Hem fet el recompte de les primeres xifres dels vots obtinguts pel partit demòcrata, pel partit republicà i pel total de vots en cadascun dels estats, repartits en pobles/ciutats

Eleccions dels EEUU

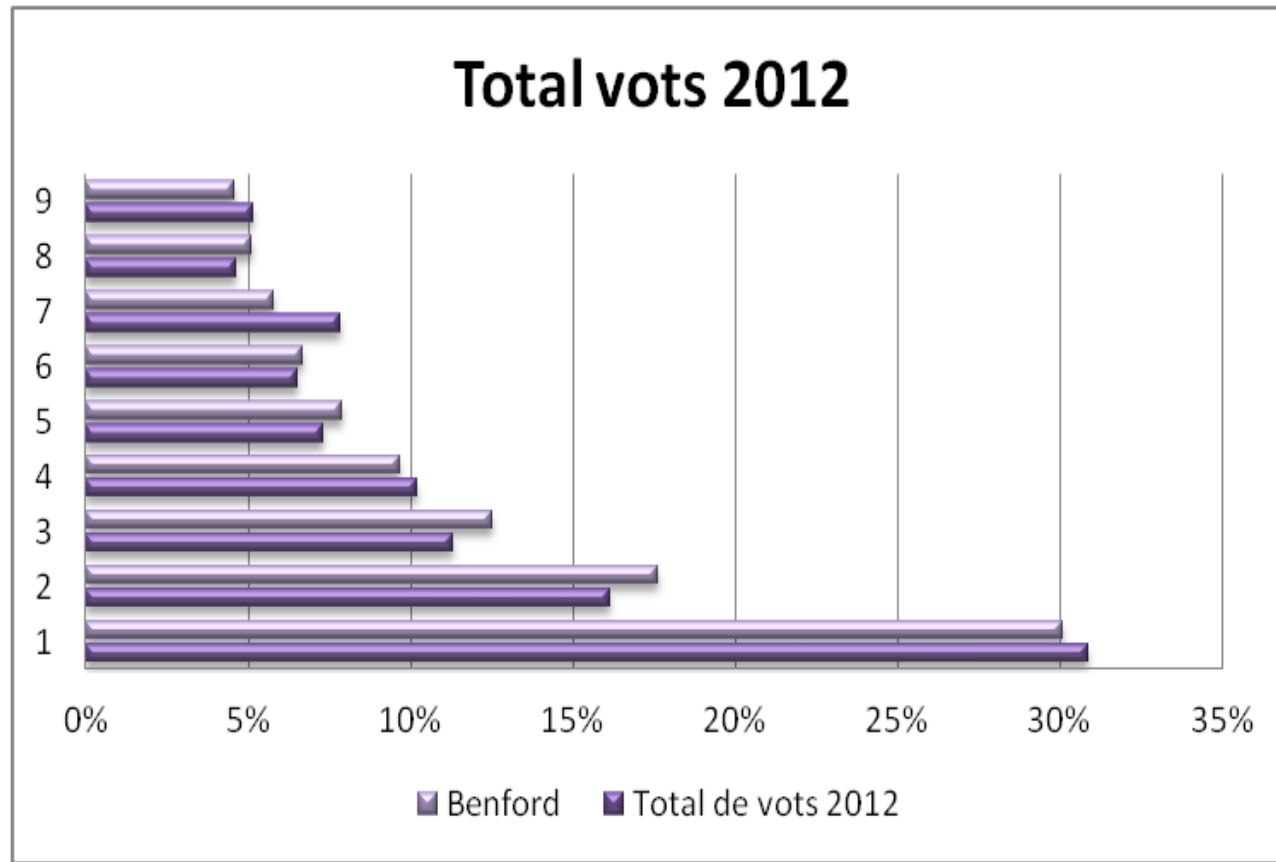
Vots Partit Republicà 2016



Vots Partit Demòcrata 2016



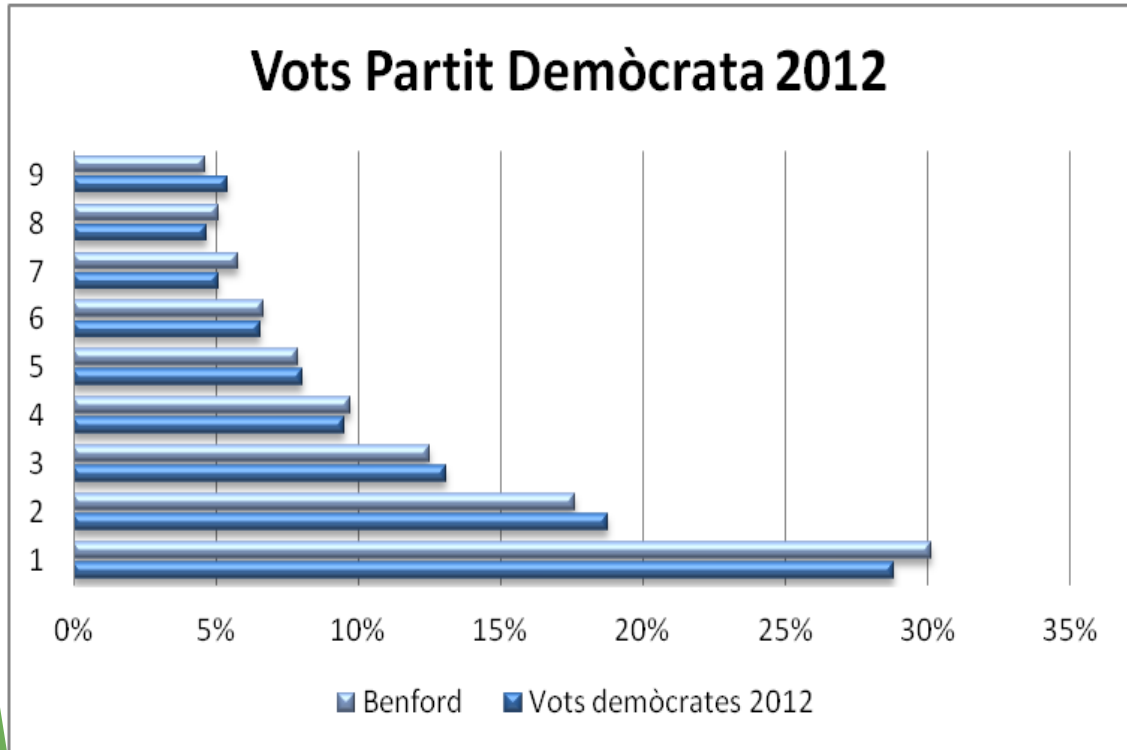
Eleccions dels EEUU



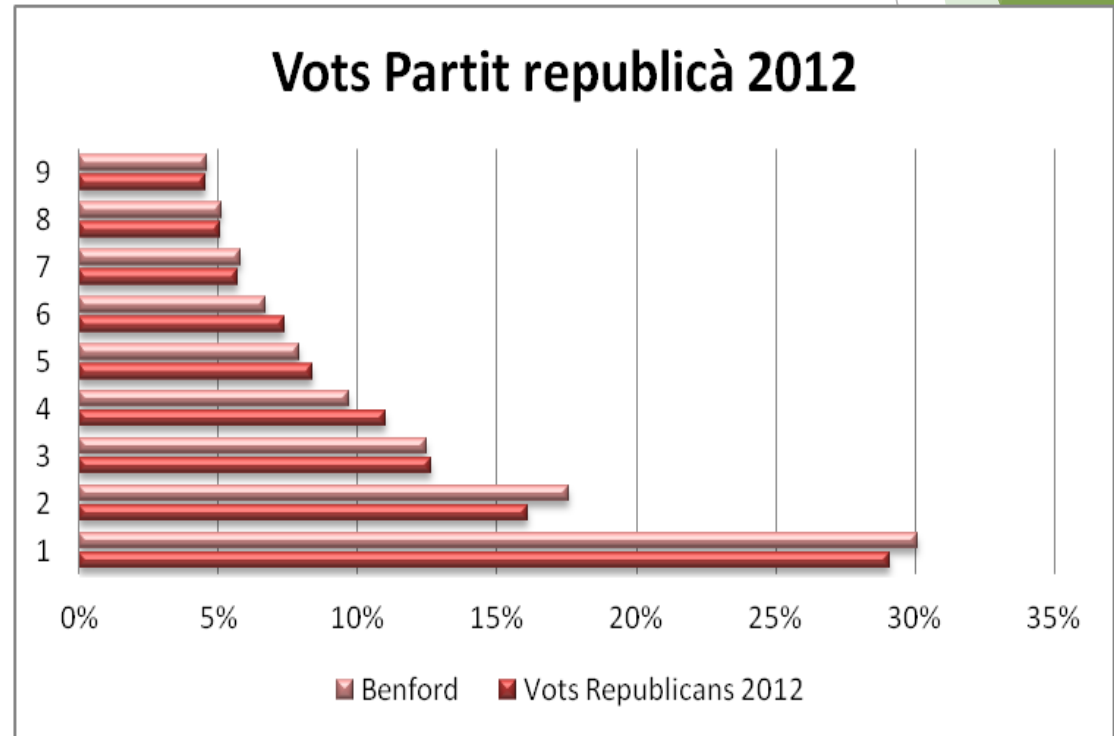
- ▶ Les primeres gràfiques segueixen la llei de Benford, per tant si s'ha fet trampa en qualsevol dels casos, s'han fet d'una manera poc notòria.

Eleccions dels EEUU

Vots Partit Demòcrata 2012

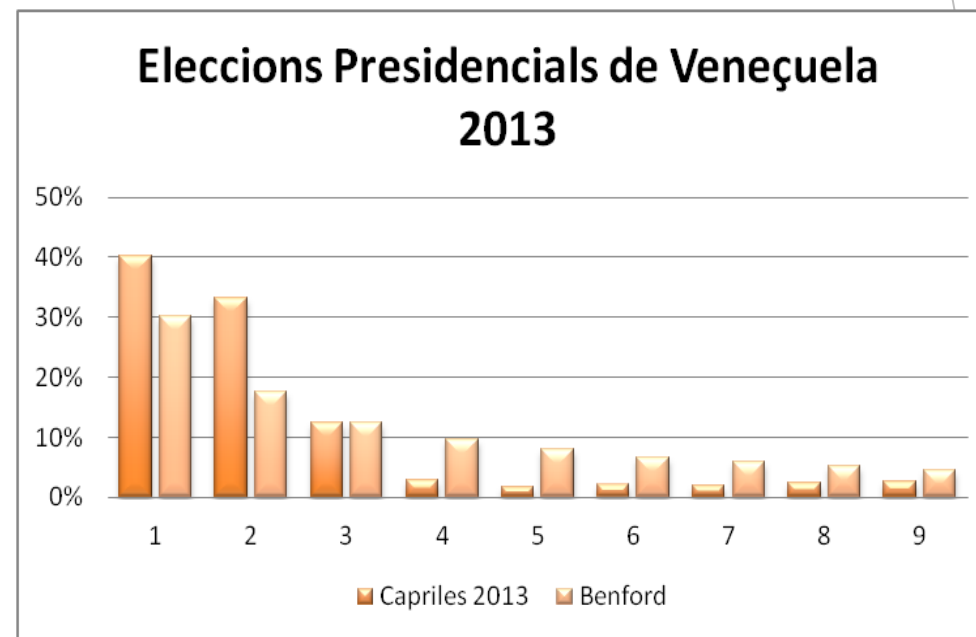
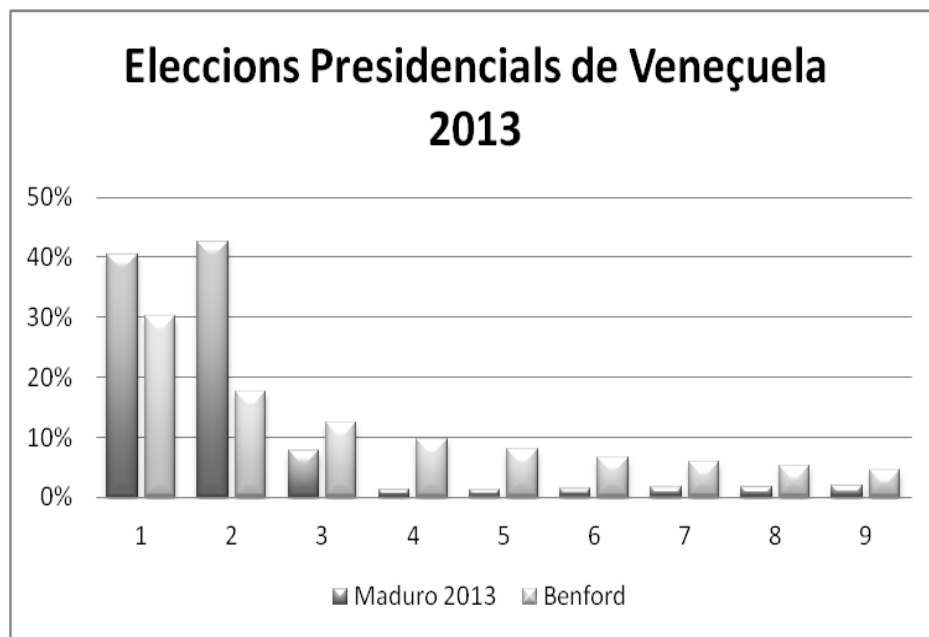


Vots Partit republicà 2012



Veneçuela

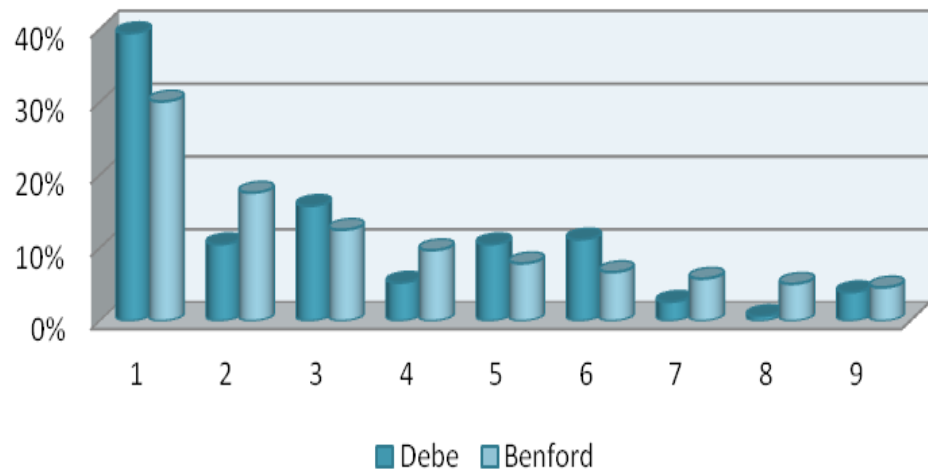
Com podem observar, en cap dels dos casos es compleix la Llei de Bendford. Per tant, sembla que s'hauria de fer una investigació més a fons sobre aquests resultats.



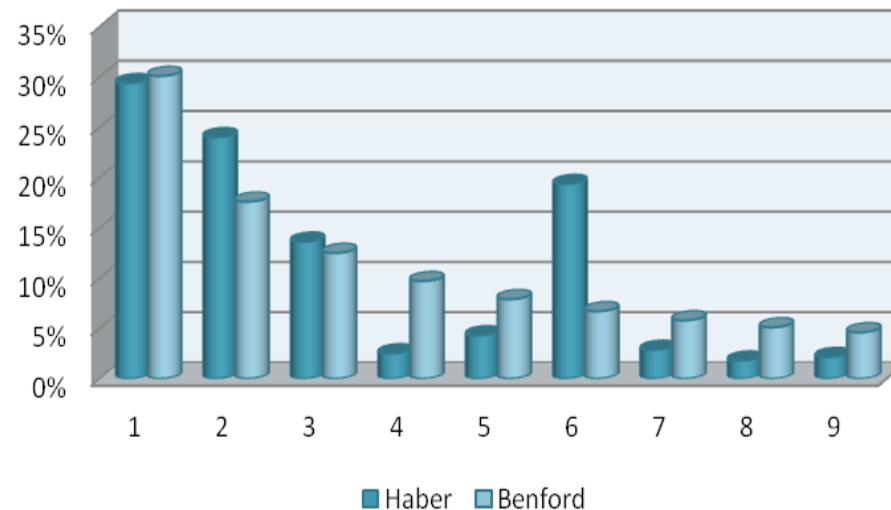
Barcenas

- ▶ Hem trobat un excel amb les dades d'en Barcenas, com podem veure no segueix la llei de Benford. En la gràfica del “Debe” la columna dels 1 està molt alta i la dels 2 molt baixa. En la gràfica del haver, sobresurt el 6 de forma molt estranya i en la del Saldo, les 4 primeres xifres són molt més elevades del que haurien de ser.

Debe Barcenas



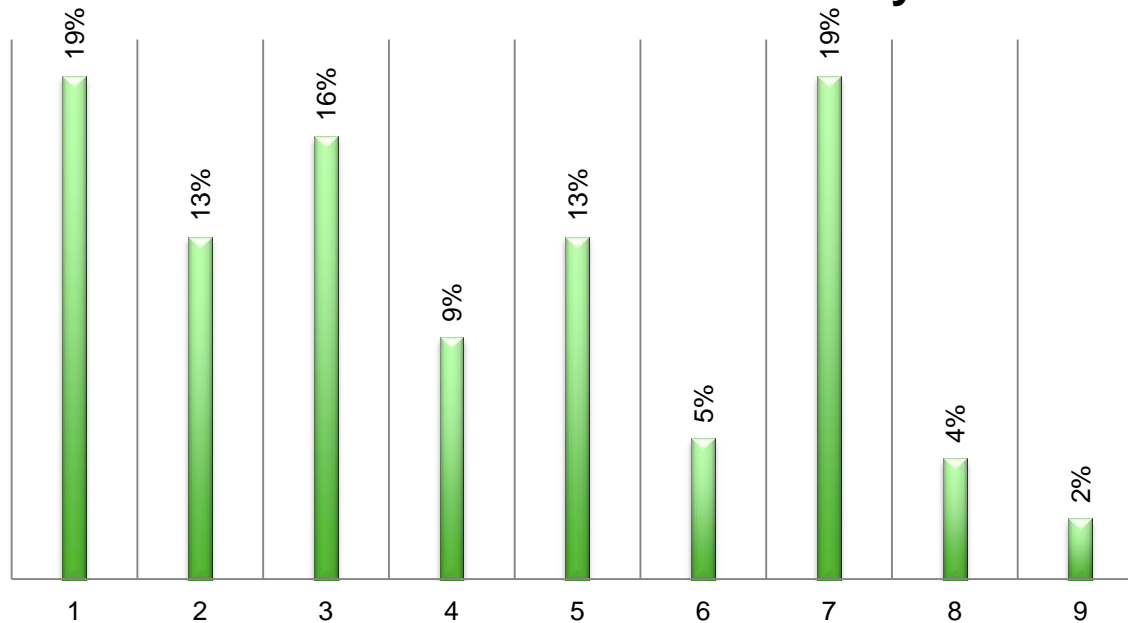
Haber Barcenas







Natalitat

- ▶ Quan és del tot estrany perquè si ens quedem amb la primera xifra, hi ha moltes més possibilitats de 1 i 2 que la resta de les xifres. (En un mes hi ha 13 xifres 1, 12 xifres 2, 2 o 3 xifres 3, i una per la resta de xifres).

dades de la natalitat a catalunya



Conclusions

-  1. La Llei de Benford es complirà al conjunt de dades/xifres de cada municipi
-  2. La Llei de Benford es complirà per el conjunt de global de les xifres dels municipis
-  3. Les dades de loteries no seguiran la Llei de Benford.
-  4. Les dades de les loteries seguiran les Lleis de Probabilitat.

Conclusions



5. Les dades dels resultats electorals dels Estats Units no seguiran la Llei de Benford.







6. Les dades dels resultats electorals de Veneçuela per a la presidència no seguiran la Llei de Benford



7. Les dades sobre les eleccions alemanyes l'any 1933, en què el partit Neonazi, amb Hitler al capdavant no compliran la Llei de Benford

Conclusions

-   8. La natalitat serà una dada que al igual de la població seguirà la Llei de Benford.
-  9. Les dades sobre la comptabilitat d'en Barcnas, no compliran la Llei de Benford.
-  10. Les dades de l'IPC seguiran la Llei de Benford

Moltes gràcies