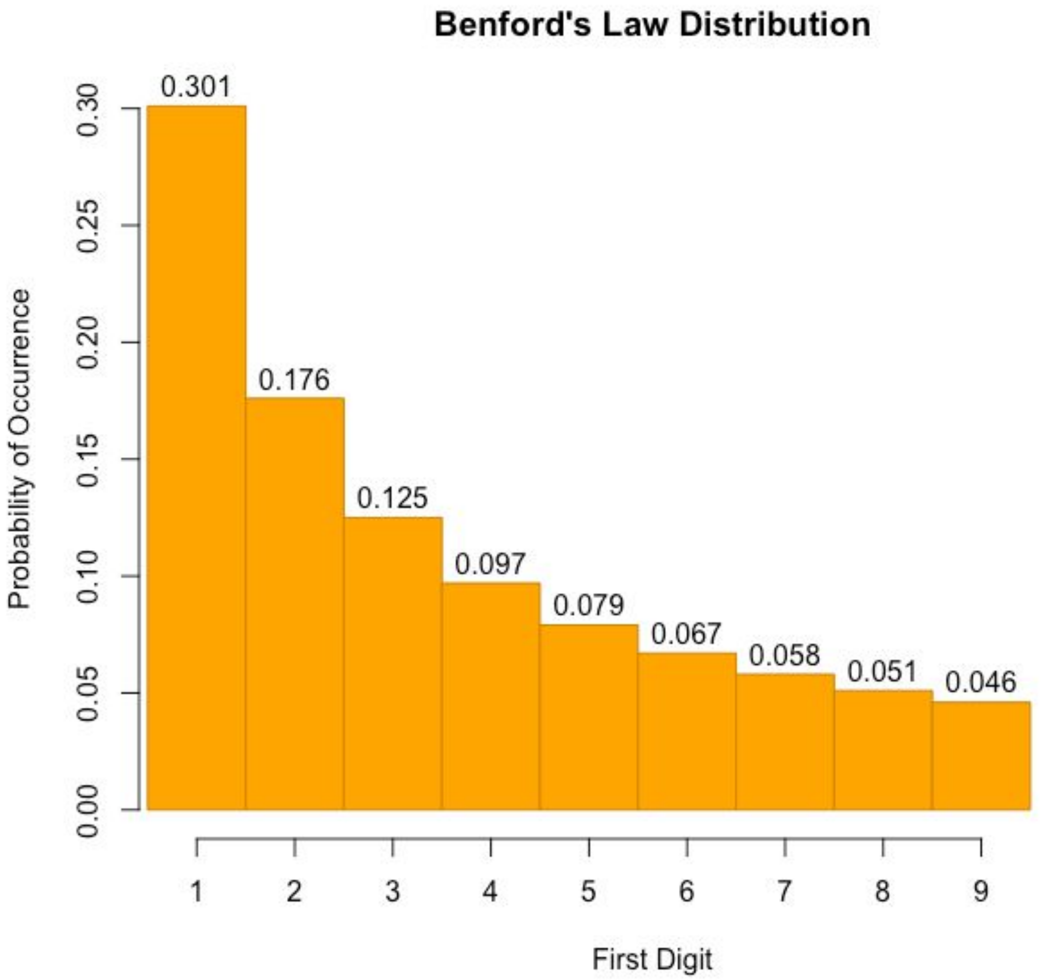
Benford’s Law

Tanli Su
Sophie Wang
Saunak Badam
Aaron Lee
Alexander (Sasha) Kokoshinskiy

A set of numbers is said to satisfy Benford's law if the first significant digit *D* of the numbers follows the following probability distribution:

$$Pr(D = d) = \log_{10}\left(1 + \frac{1}{d}\right), d \in \{1, ..., 9\}$$


Benford's Law Distribution
0.301
0.176
0.125
0.097
0.079
0.067
0.058
0.051
0.046
Probability of Occurrence
0.00
0.05
0.10
0.15
0.20
0.25
0.30
1
2
3
4
5
6
7
8
9
First Digit

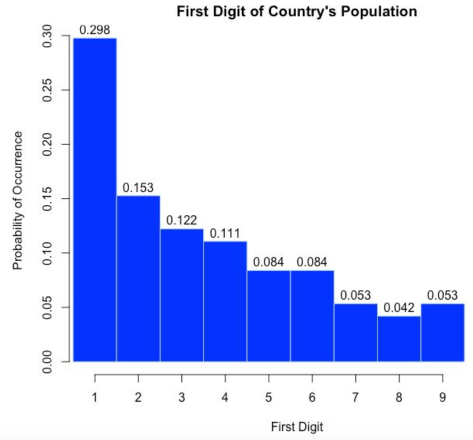
History of Benford's Law

- First discovered in 1881 by astronomer Simon Newcomb
  - "the law of probability of the occurrence of numbers is such that all mantissa of their logarithms is are equally likely"
- In 1938, physicist Frank Benford tested this on datasets in 20 different disciplines
  - Physical constants, molecular weights, numbers in Reader's Digest, etc.

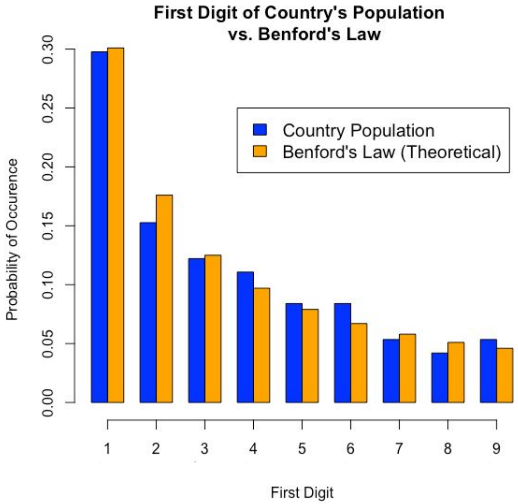
Title	1	2	3	4	5	6	7	8	9	Samples
Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
Newspaper items	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
Spec. Heat	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
H.P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
Mol. Wgt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
Atomic Wgt.	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
n^{-1}, $\sqrt{n}$	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5000
Design	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
Digest	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
X-Ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1458
Blackbody	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
$n^1, n^2, \ldots, n!$	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
Average	30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1011
Probable Error ($\pm$)	0.8	0.4	0.4	0.3	0.2	0.2	0.2	0.2	0.3	

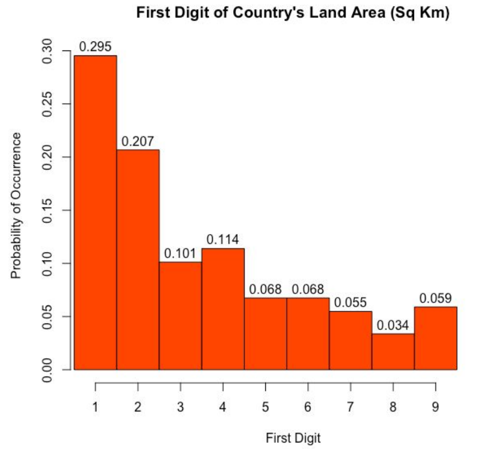
Distributions That Fit Benford’s Law

DATA: THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

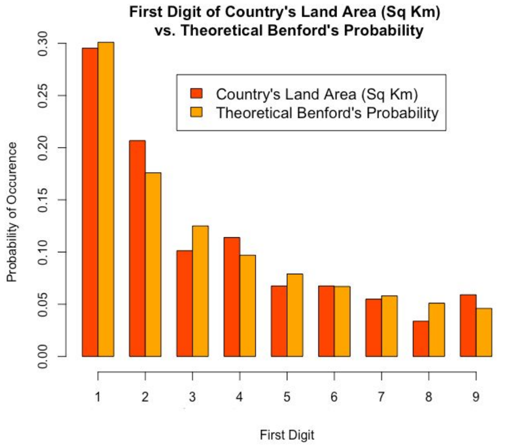


Population (per Country)





Land Area (per Country)



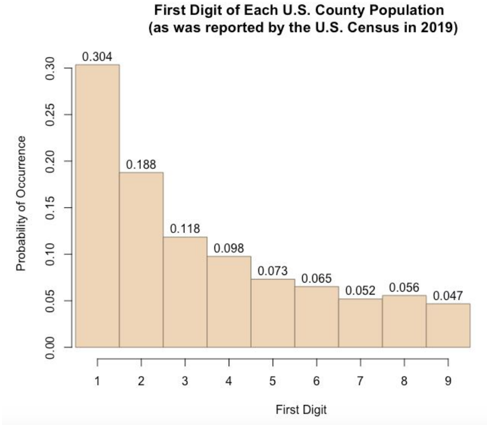
DATA: MONGABAY
NEWS & INSPIRATION FROM NATURE'S FRONTLINE

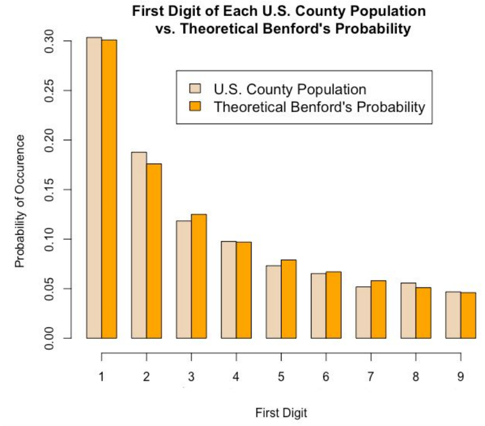
Screenshot of Excel Spreadsheets

H	I	J
Country Name	2018 Population	First Numbe
Aruba	105845	1
Afghanistan	37172386	3
Angola	30809762	3
Albania	2866376	2
Andorra	77006	7
Arab World	419790588	4
United Arab Emirates	9630959	9
Argentina	44494502	4
Armenia	2951776	2
American Samoa	55465	5
Antigua and Barbuda	96286	9
Australia	24982688	2
Austria	8840521	8
Azerbaijan	9939800	9
Burundi	11175378	1
Belgium	11433256	1
Benin	11485048	1
Burkina Faso	19751535	1
Bangladesh	161356039	1

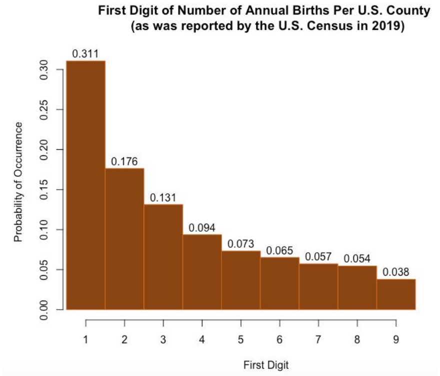
	A	B	C	D	E
1	Rank	Country	Capital City	Land Area (S	FirstDigit
2	1	Russia	Moscow	17075200	1
3	2	Canada	Ottawa	9976140	9
4	3	United State	Washington	9629091	9
5	4	China	Beijing	9596960	9
6	5	Brazil	Brasilia	8511965	8
7	6	Australia	Canberra	7686850	7
8	7	India	New Delhi	3287590	3
9	8	Argentina	Buenos Aires	2766890	2
10	9	Kazakhstan	Astana	2717300	2
11	10	Sudan	Khartoum	2505810	2
12	11	Algeria	Algiers	2381740	2
13	12	Congo (Dem.	Kinshasa	2345410	2
14	13	Greenland	Nuuk	2166086	2
15	14	Mexico	Mexico	1972550	1
16	15	Saudi Arabia	Riyadh	1960582	1
17	16	Indonesia	Jakarta	1919440	1
18	17	Libya	Tripoli	1759540	1
19	18	Iran	Tehran	1648000	1
20	19	Mongolia	Ulaanbaatar	1565000	1
21	20	Peru	Lima	1285220	1
22	21	Chad	N'Djamena	1284000	1

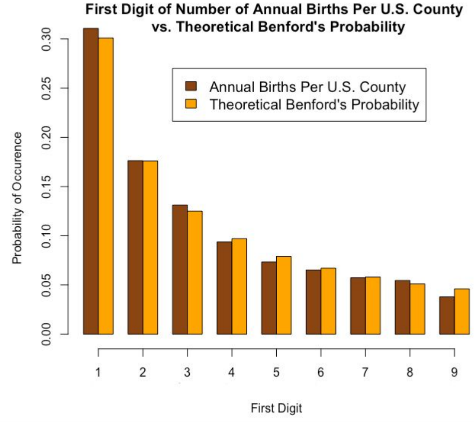
Population (per U.S.County)



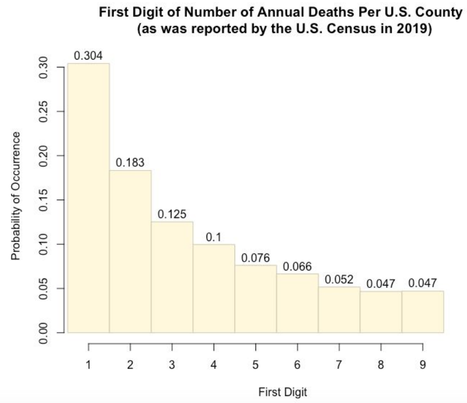


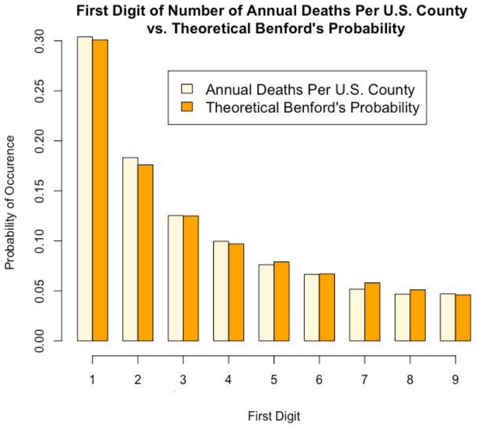
Annual Births (per U.S. County)





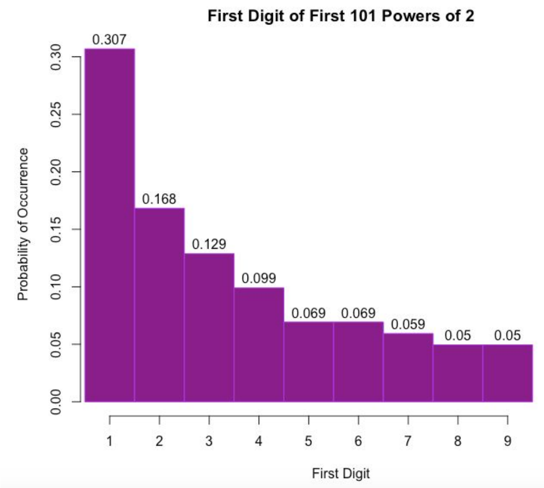
Annual Deaths (per U.S. County)

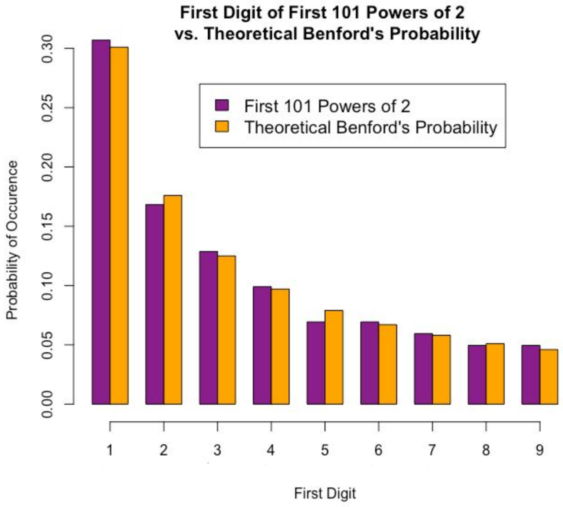




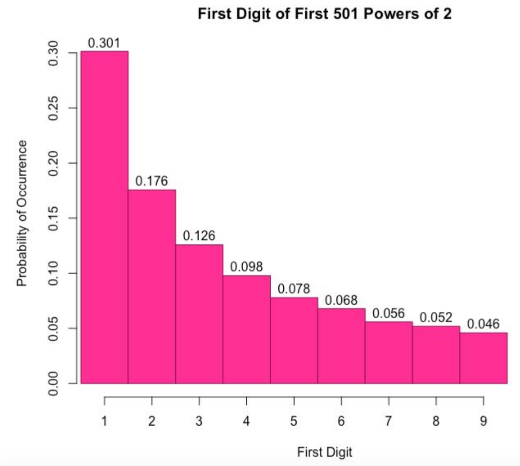
Pure Mathematical Sequences

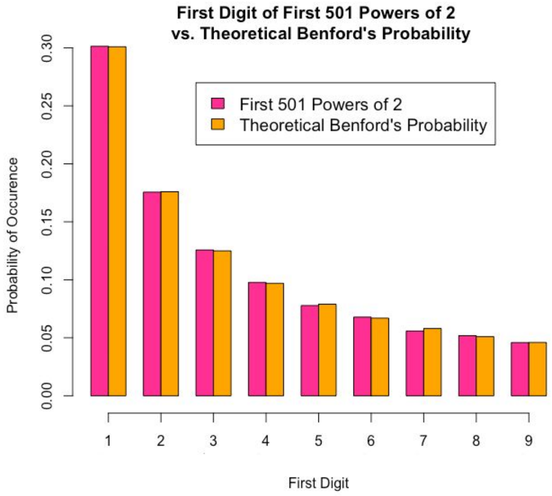
First 101 Powers of 2



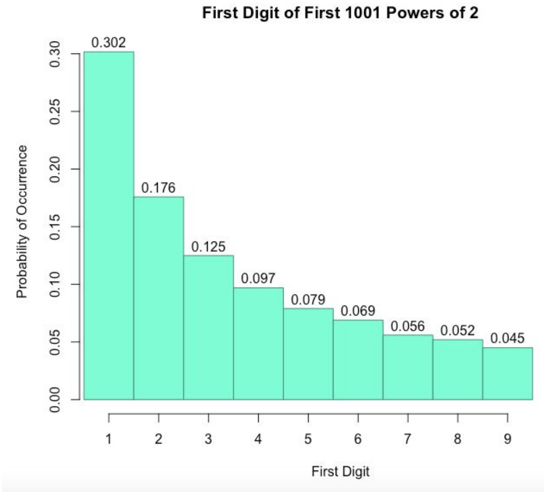


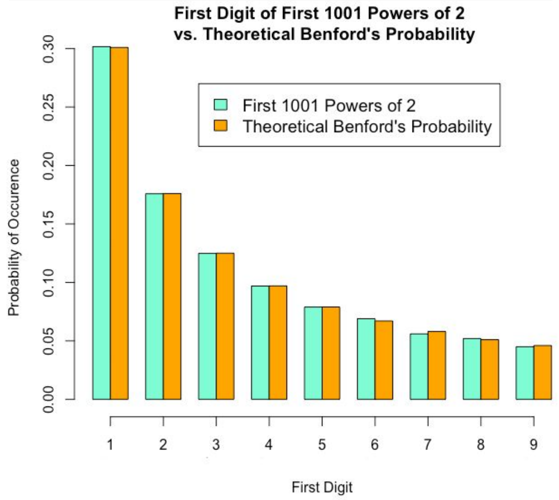
First 501 Powers of 2



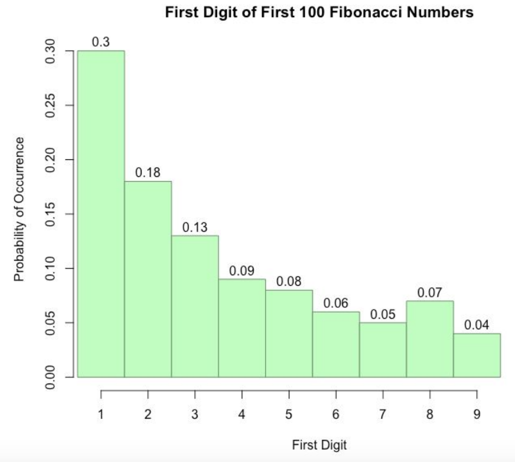


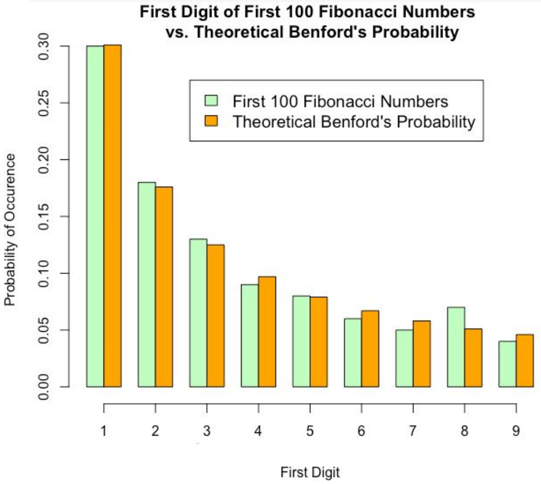
First 1001 Powers of 2



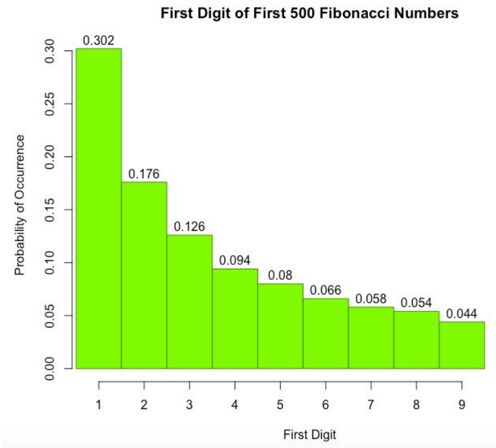


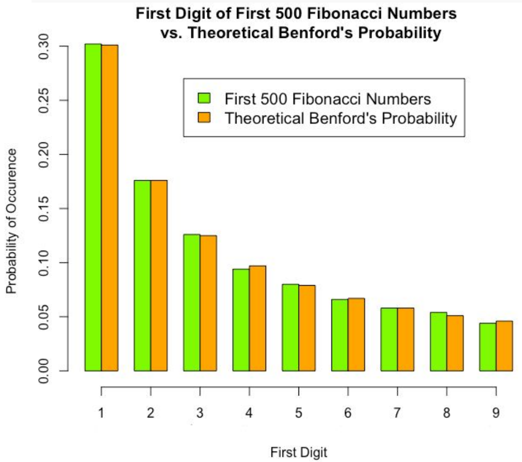
First 100 Fibonacci Numbers



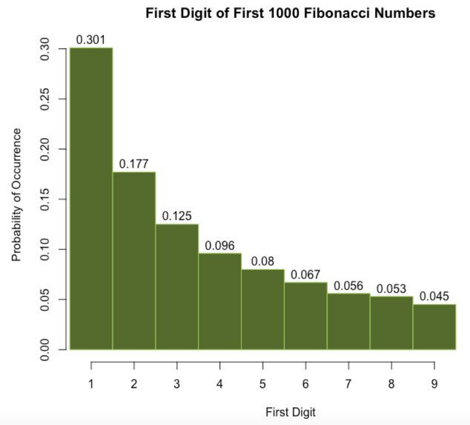


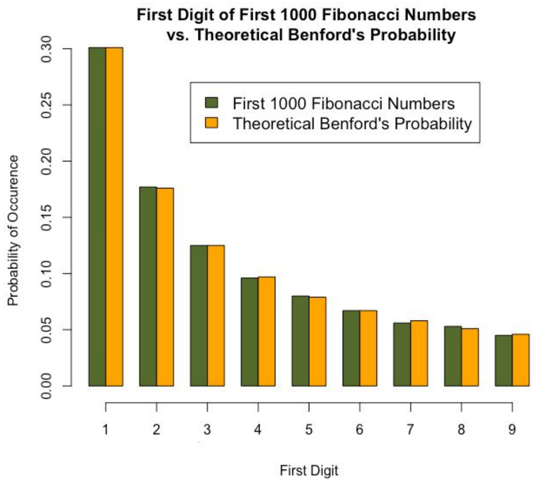
First 500 Fibonacci Numbers





First 1000 Fibonacci Numbers





Factorials

$$1! = 1$$

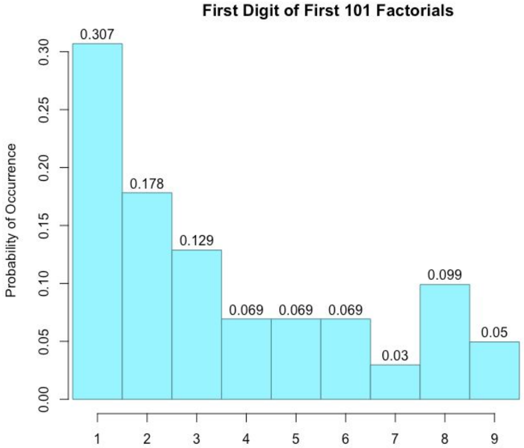
$$2! = 2(1) = 2$$

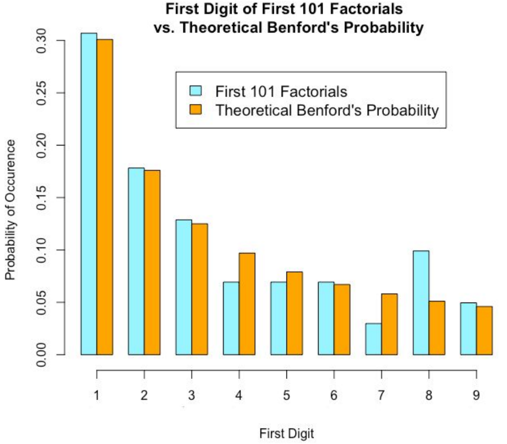
$$3! = 3(2)(1) = 6$$

$$4! = 4(3)(2)(1) = 24$$

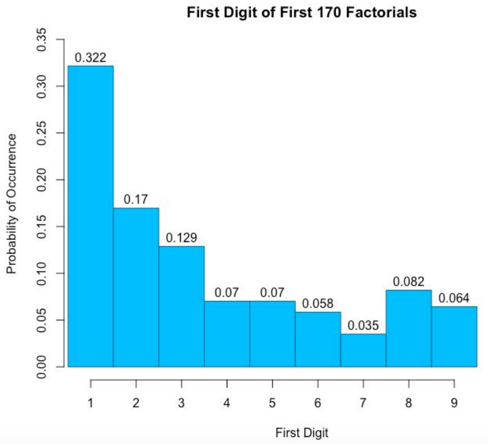
$$5! = 5(4)(3)(2)(1) = 120$$

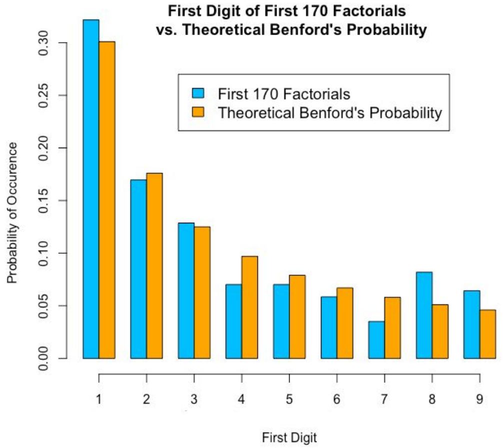
First 100 Factorials





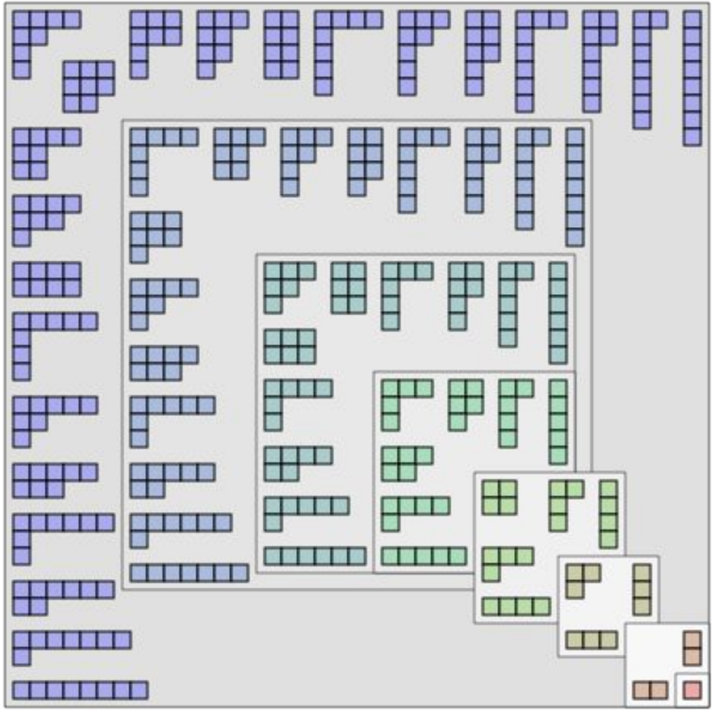
First 170 Factorials

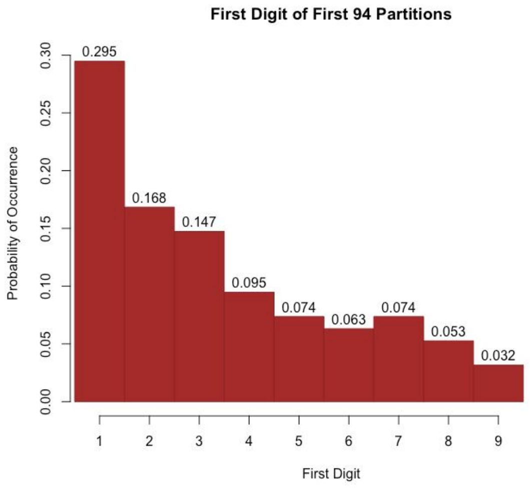


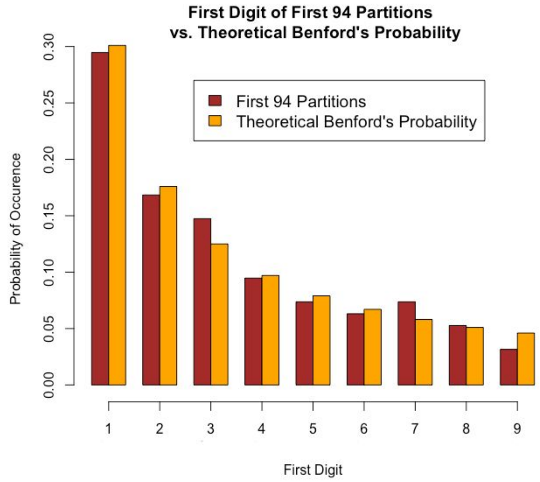


Partitions

There is 1 partition of 1,
There are 2 partitions of 2,
There are 3 partitions of 3,
There are 5 partitions of 4,
There are 7 partitions of 5,
There are 11 partitions of 6,
There are 15 partitions of 7,
There are 22 partitions of 8...



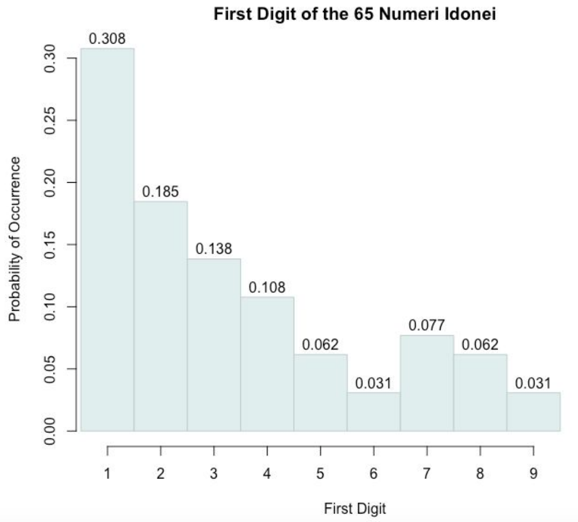


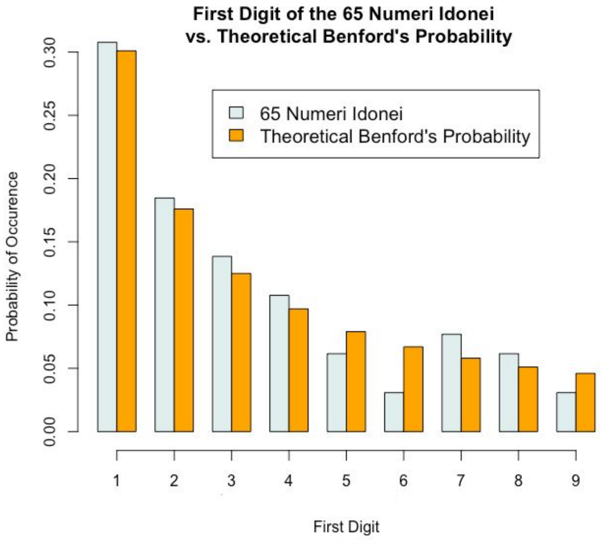


https://www.mathpages.com/home/kmath058.htm

Numeri Idonei

Euler found sixty-five integers, which he called "numeri idonei", that could be used to prove the primality of certain numbers.





Other Bases

Generalization of Benford Law

How we traditionally look at Benford's Law

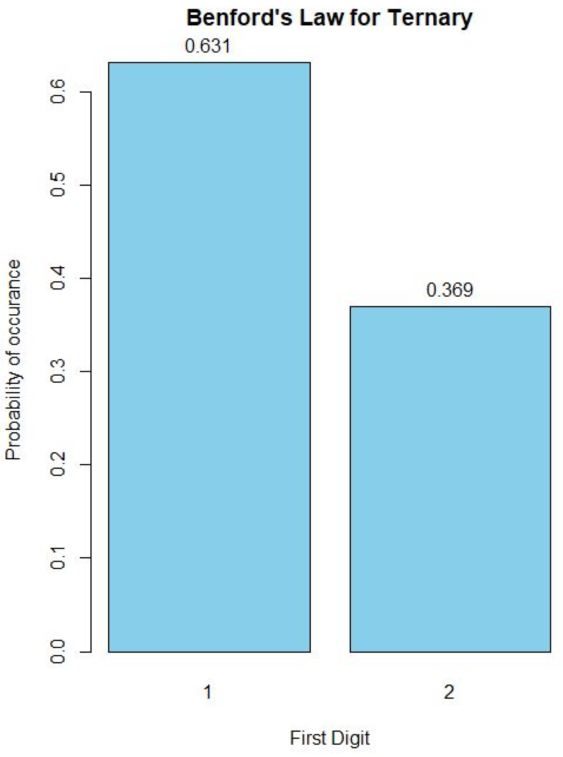
$$P(d) = \log_{10}(d+1) - \log_{10}(d) = \log_{10}\left(\frac{d+1}{d}\right) = \log_{10}\left(1 + \frac{1}{d}\right)$$

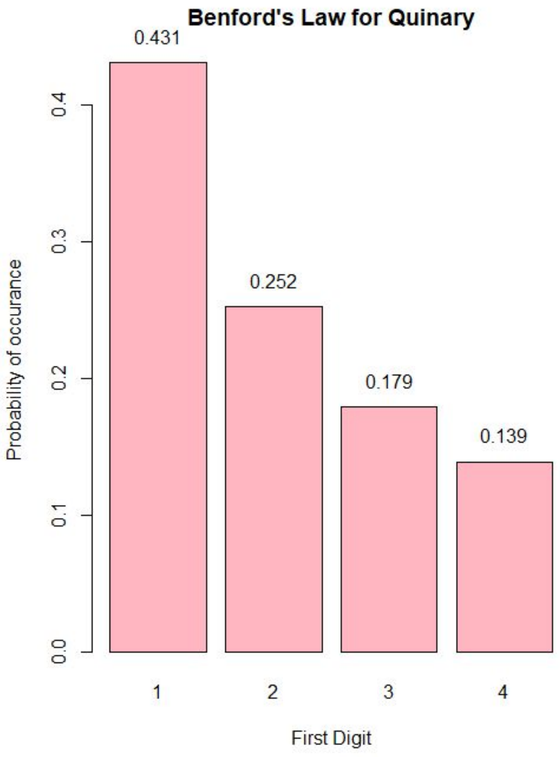
How this can be generalized for other bases

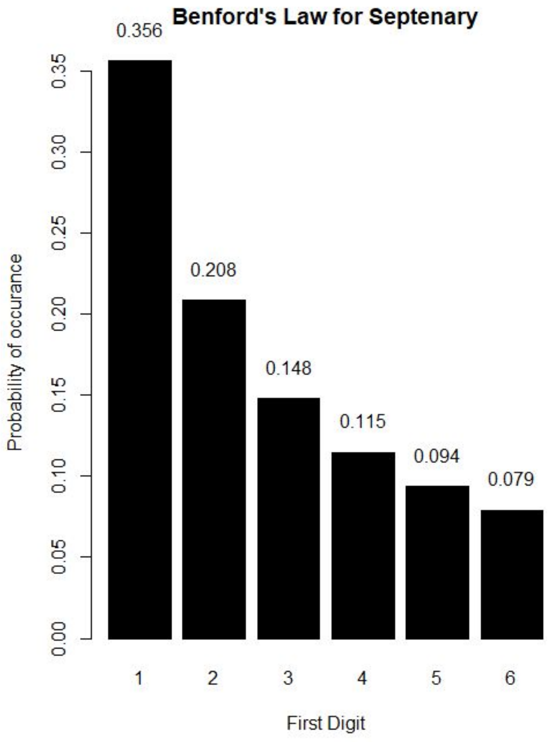
$$P(d) = \log_b(d+1) - \log_b(d) = \log_b\left(1 + \tfrac{1}{d}\right).$$

Some popular examples coming up

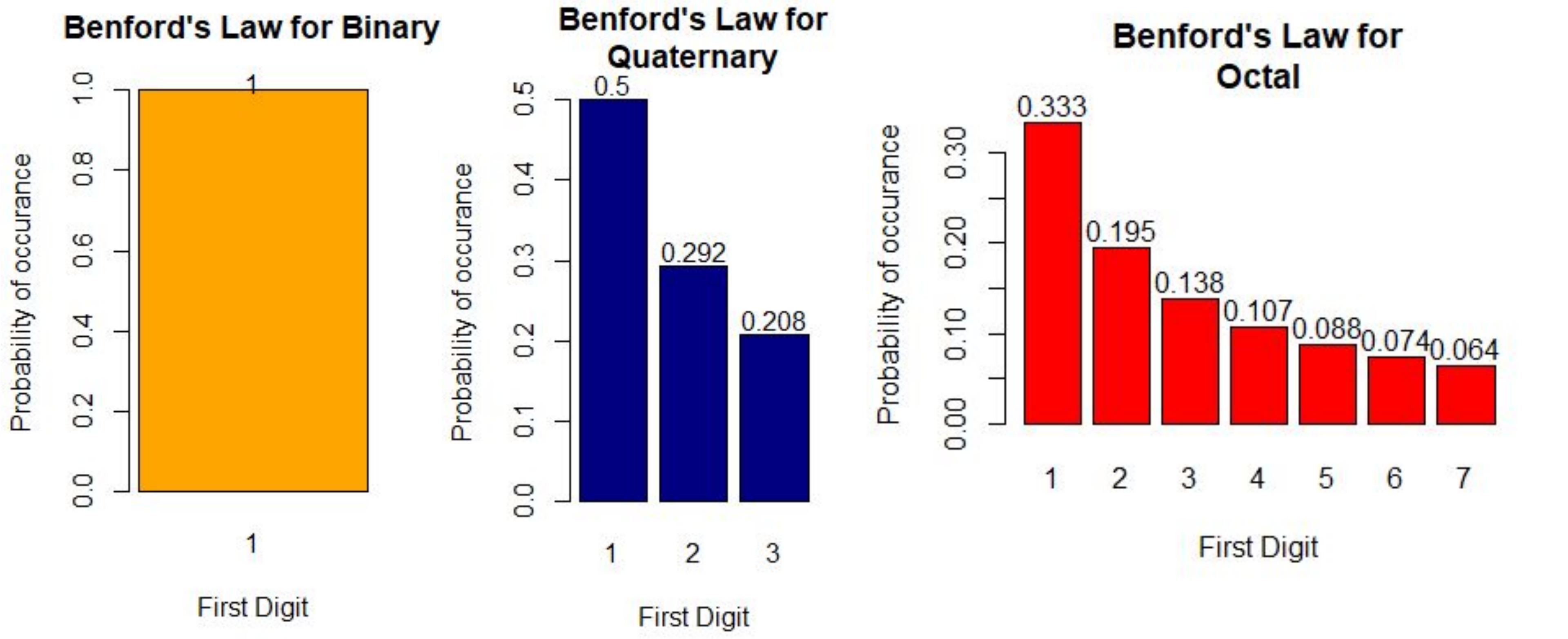
Benford’s Law in Interesting Bases







Bases Computers Work in



Powers of 2 and leading digit of 1

$$P(d) = \log_b(d+1) - \log_b(d) = \log_b\left(1 + \frac{1}{d}\right).$$

$\log_2(2) = 1$, because $2^1 = 2$; $\log_{2^1}(2) = x$, $2 = (2^1)^x$, $2 = 2^x$, $x = 1$

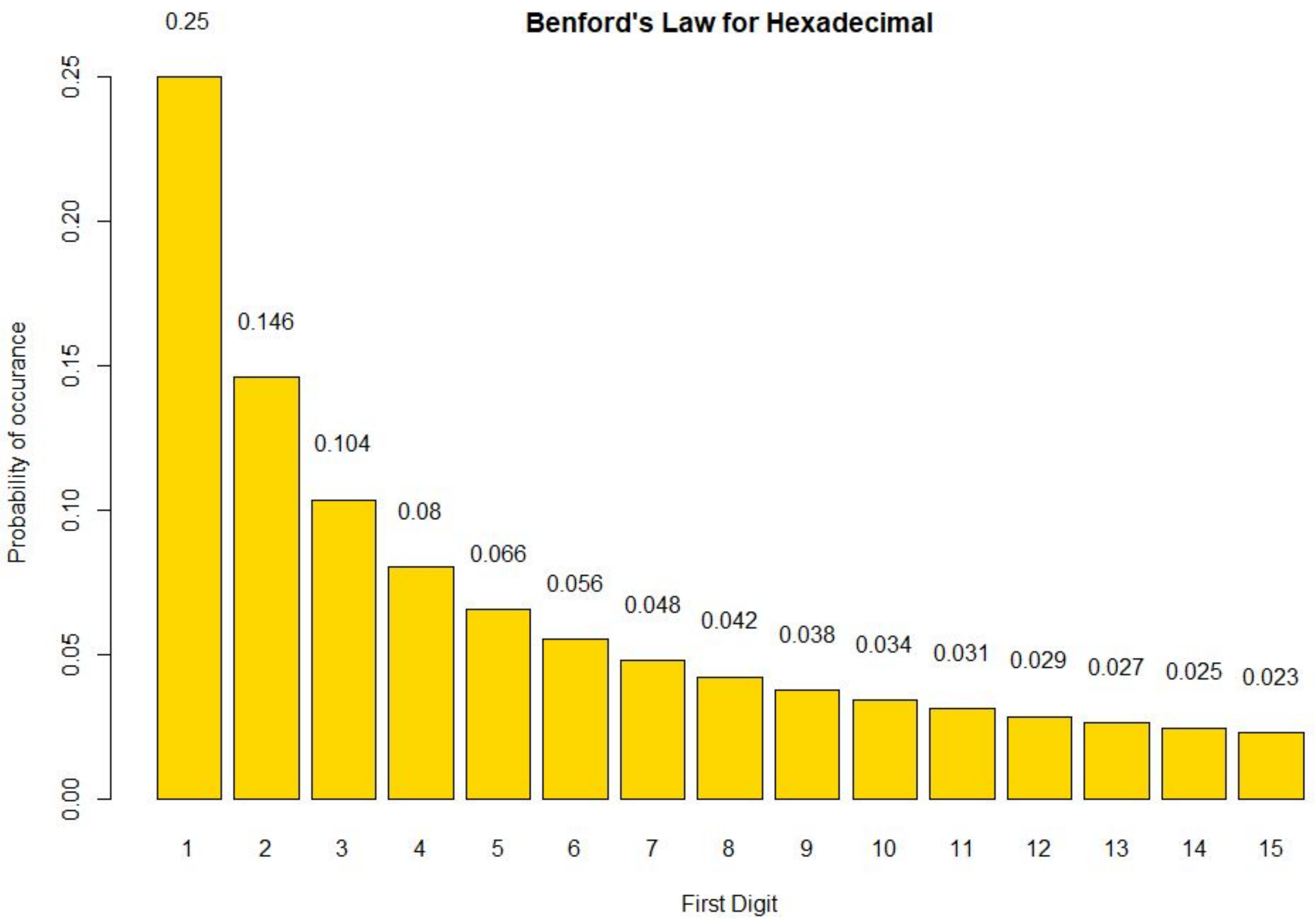
$\log_4(2) = ½$, because $4^{(½)} = 2$; $\log_{2^2}(2) = x$, $2 = (2^2)^x$, $2 = 2^{(2*x)}$, $x = ½$

$\log_8(2) = ⅓$, because $8^{(⅓)} = 2$; $\log_{2^3}(2) = x$, $2 = (2^3)^x$, $2 = 2^{(3*x)}$, $x = ⅓$

We see a pattern

$\log_{2^c}(2) = x$, $2 = (2^c)^x$, $2 = 2^{(c*x)}$, $x = 1/c$

Base 16

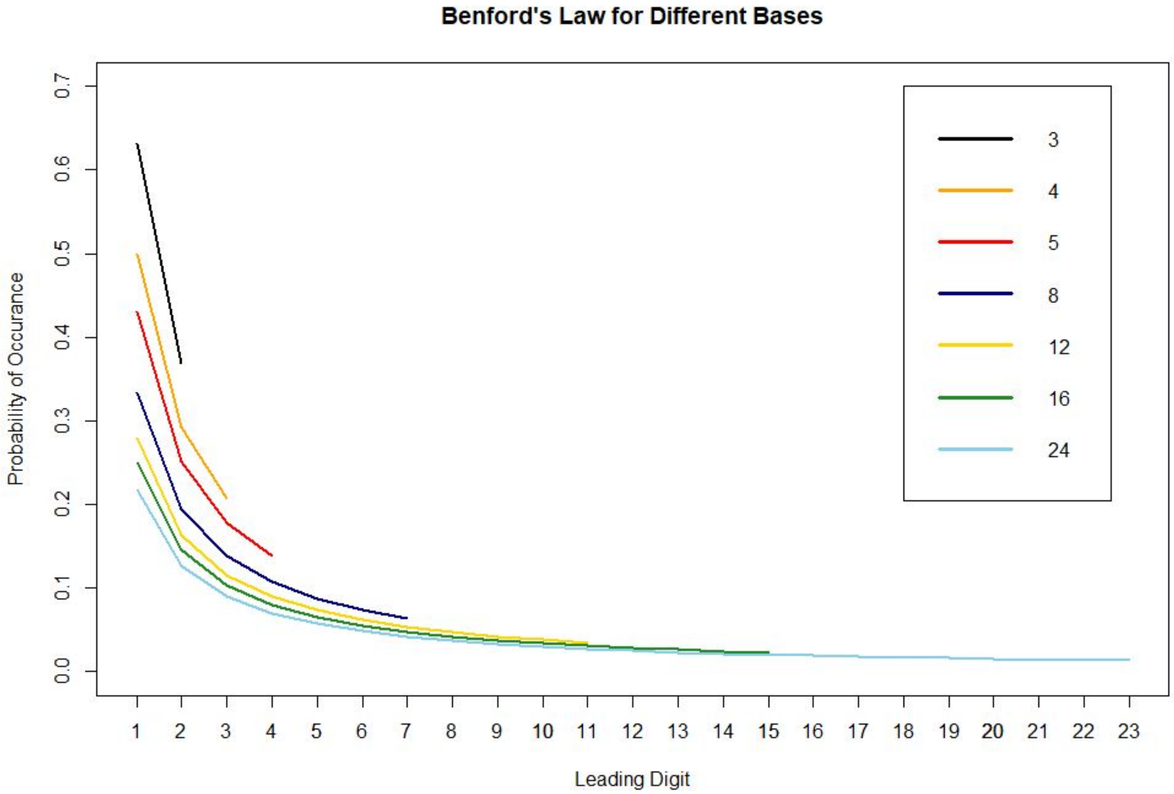

Benford's Law for Hexadecimal
Probability of occurance
0.00
0.05
0.10
0.15
0.20
0.25
0.25
0.146
0.104
0.08
0.066
0.056
0.048
0.042
0.038
0.034
0.031
0.029
0.027
0.025
0.023
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
First Digit

$$P(d) = \log_b(d+1) - \log_b(d) = \log_b\left(1 + \frac{1}{d}\right).$$

Interesting Points

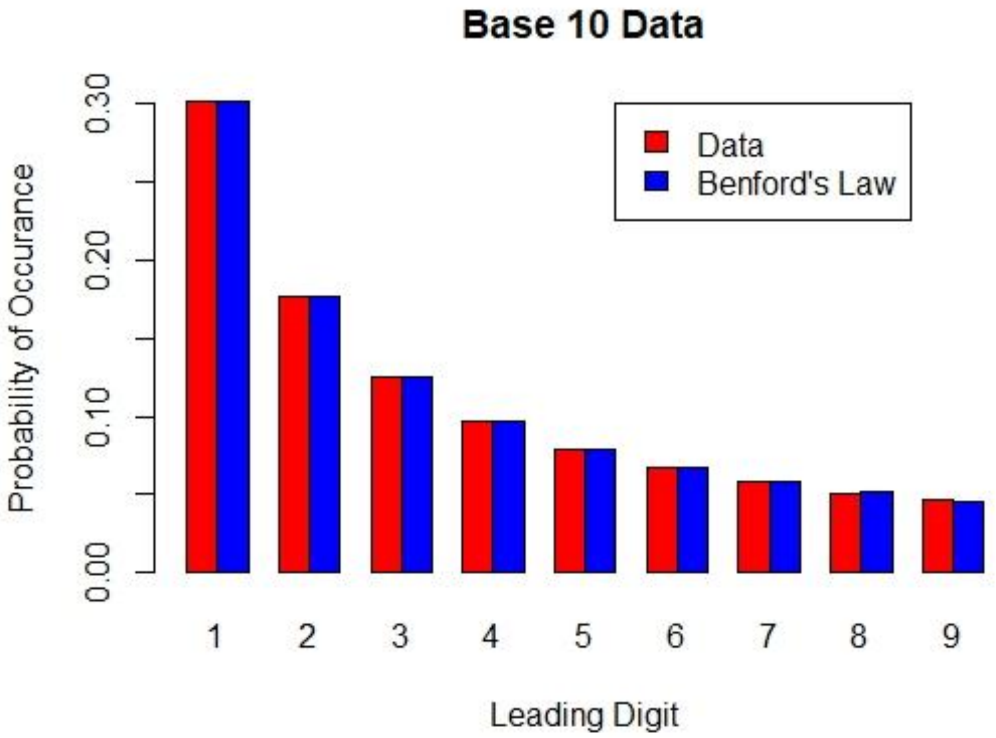
1. Leading digit of 1 will always have highest probability of occuring (maximize last equation)
2. As base increases, probability of leading digit 1 decreases
3. As base approaches infinity, Pr(leading digit = 1) approaches 0
4. [remember Pr(leading digit = 1 | Base = 2^c) = 1/c]
5. [rewritten as Pr(leading digit = 1 | Base = c) = 1/$\log_2$(c)]
6. The curve will flatten out as base approaches infinity

Visualization for many bases in a single graph



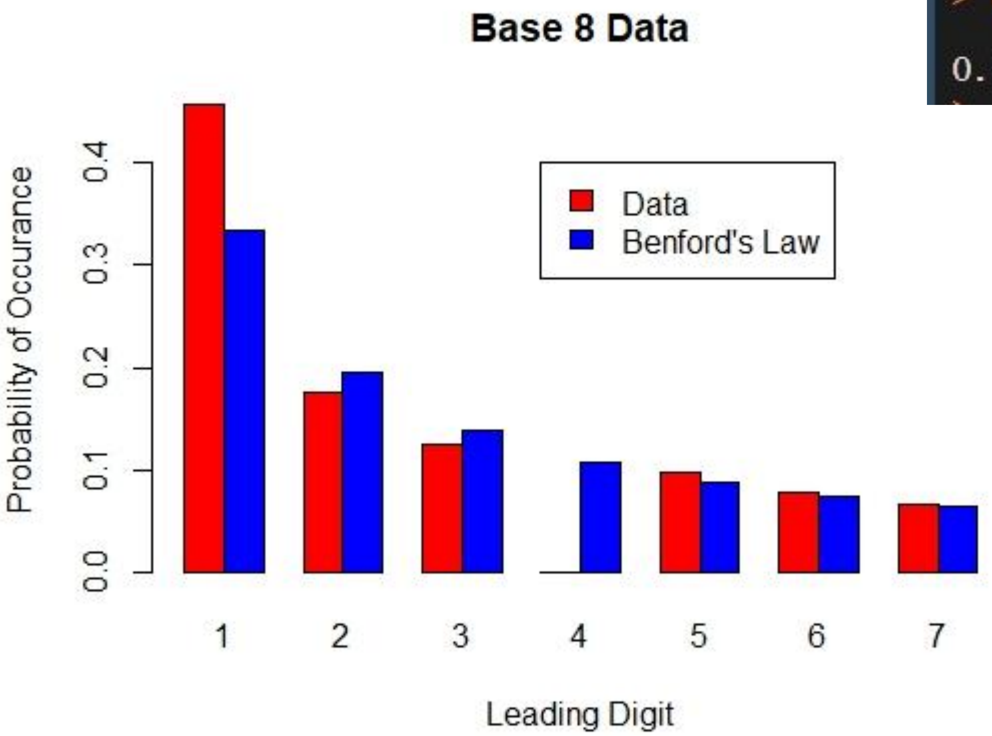
Catching Human Made Data

```
> base10
    1     2     3     4     5     6     7     8     9 
0.301 0.176 0.125 0.097 0.079 0.067 0.058 0.051 0.046 
```




```
> base10 %>% table
.
 10  20  30  40  50  60  70  80  90 
301 176 125  97  79  67  58  51  46 
> 
```

Look at it in Base 8




```
> base8
    1     2     3     4     5     6     7
0.456 0.176 0.125 0.000 0.097 0.079 0.067
```

Base 10	Base 8
10	12
20	24
30	36
40	50
50	62
60	74
70	106
80	120
90	132

Use in Letters

Letters

Benford's Law and First Letter of Word

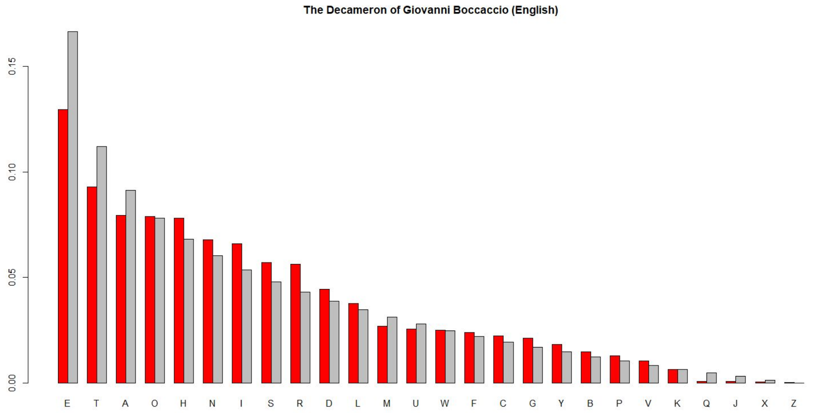
Article *in* Physica A: Statistical Mechanics and its Applications · December 2017

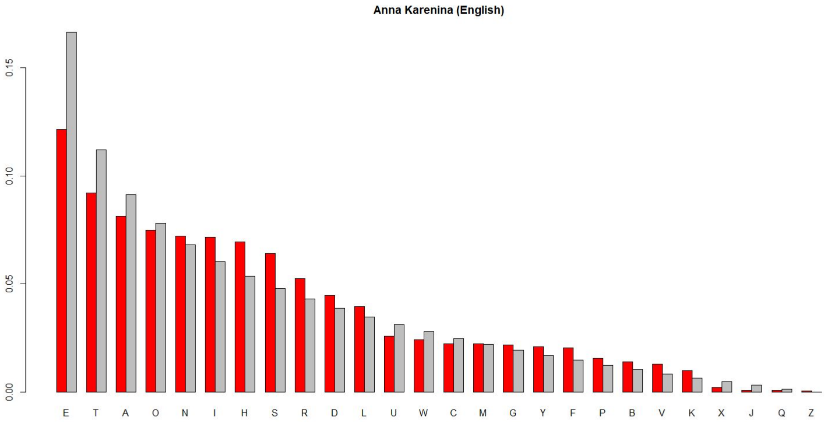
DOI: 10.1016/J.physa.2018.08.133

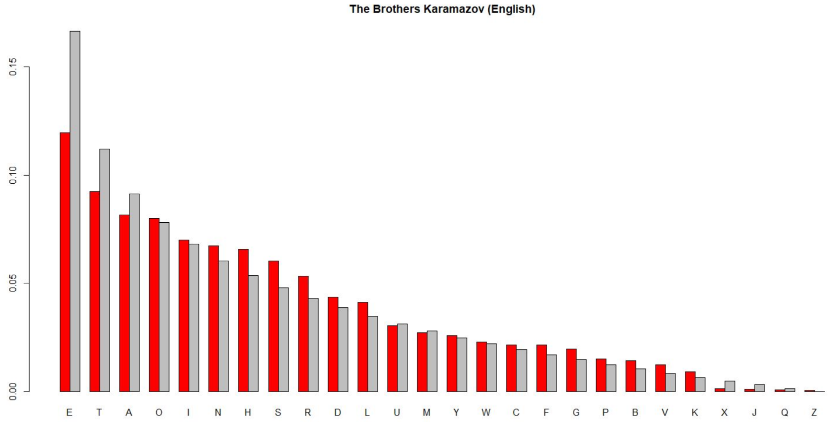
$$p_i = \frac{X - (X-1)\log_X(X-1) - i\log_X i + (i-1)\log_X(i-1)}{X(X-1)\log_X\left(\frac{X}{X-1}\right)}. \tag{2}$$

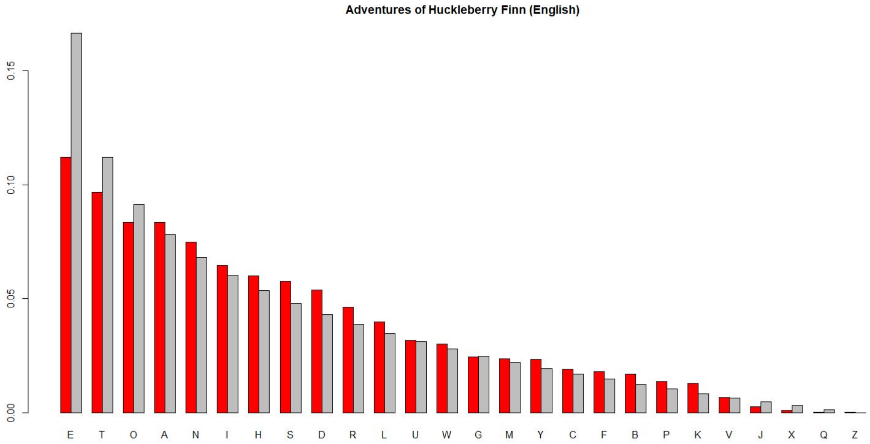
First Letter Law - similar idea to Benford’s law, however it is simply an observation

Use in books

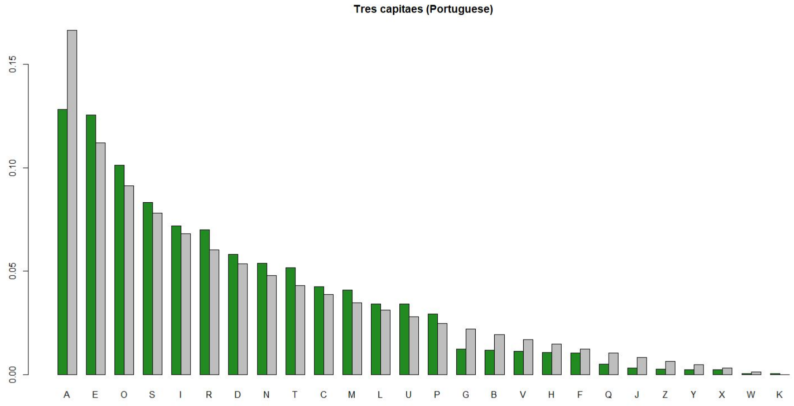

The Decameron of Giovanni Boccaccio (English)
0.00
0.05
0.10
0.15
E T A O H N I S R D L M U W F C G Y B P V K Q J X Z

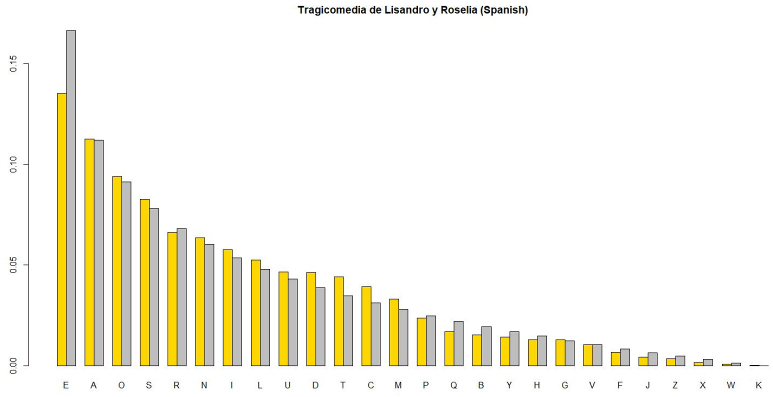

Anna Karenina (English)
0.00
0.05
0.10
0.15
E T A O N I H S R D L U W C M G Y F P B V K X J Q Z

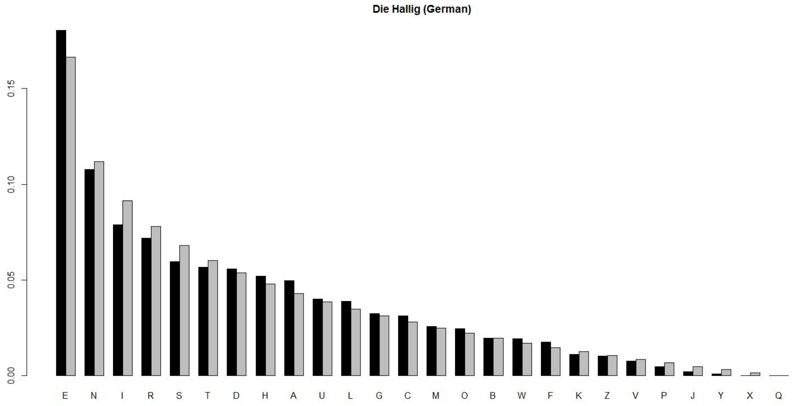

The Brothers Karamazov (English)
0.00
0.05
0.10
0.15
E T A O I N H S R D L U M Y W C F G P B V K X J Q Z


Adventures of Huckleberry Finn (English)
0.00
0.05
0.10
0.15
E T O A N I H S D R L U W G M Y C F B P K V J X Q Z

Use in books in other languages







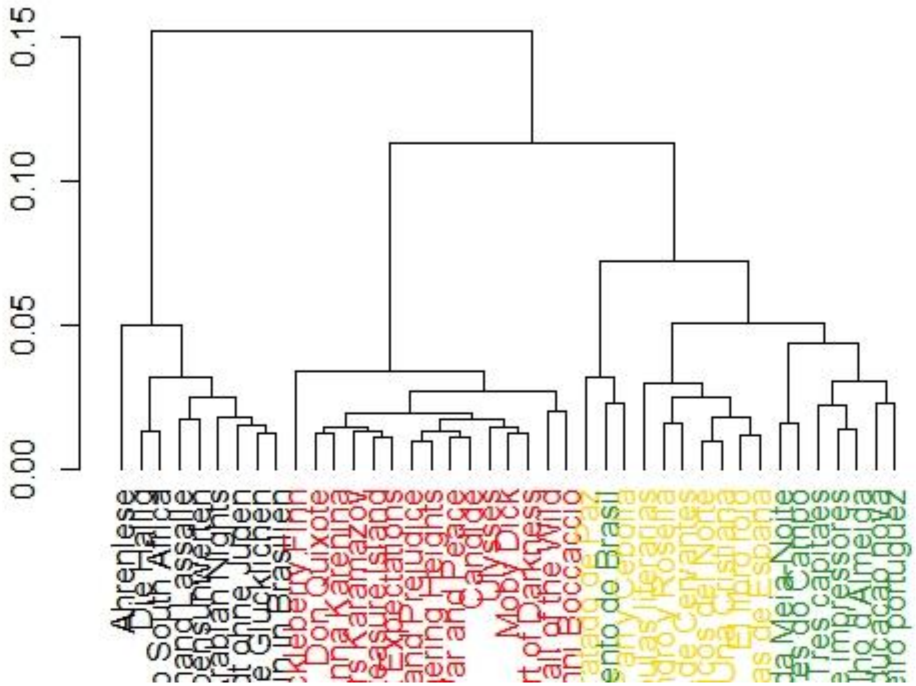
Differentiating language by letter frequency

Black - German

Red - English

Yellow - Spanish

Green - Portuguese



Proof

Prove that the mixture distribution of distributions that obey Benford’s law also obeys Benford’s law.

in other words,

Let $X_1, ..., X_n$ be random variables such that for all X_i, the distribution of the first digits is given by

$$Pr(D_{X_i} = d) = \log_{10}\left(1 + \frac{1}{d}\right), \qquad d \in \{1, ..., 9\}$$

Let $p_1, ..., p_n$ be probabilities that sum to 1.

We want to prove that if $Y = \sum p_i X_i$, (or equivalently, we pick distribution X_i with probability p_i), then the distribution of the first digits of Y is also given by

$$Pr(D_Y = d) = \log_{10}\left(1 + \frac{1}{d}\right), \qquad d \in \{1, ..., 9\}$$

Proof

From Theorem 2.84 in the textbook, the distribution of the first digits of a random variable with cdf $F(x)$ is given by

$$Pr(D = d) = \sum_{k=-\infty}^{\infty} \left(F(10^k \cdot (d+1)) - F(10^k \cdot d)\right)$$

In addition, since a mixture of random variables translates to a linear combination of the cdfs, we have

$$F_Y(y) = \sum p_i \cdot F_{X_i}(y)$$

Then we can write the distribution of the first digits of Y as

$$Pr(D_Y = d) = \sum_{k=-\infty}^{\infty} \left(\sum p_i \cdot F_{X_i}(10^k \cdot (d+1)) - \sum p_i \cdot F_{X_i}(10^k \cdot d)\right)$$

Proof

$$Pr(D_Y = d) = \sum_{k=-\infty}^{\infty} \left(\sum p_i \cdot F_{X_i}(10^k \cdot (d+1)) - \sum p_i \cdot F_{X_i}(10^k \cdot d) \right)$$

which simplifies to

$$Pr(D_Y = d) = \sum_{k=-\infty}^{\infty} \left(\sum p_i \cdot (F_{X_i}(10^k \cdot (d+1)) - F_{X_i}(10^k \cdot d)) \right)$$

$$= \sum p_i \cdot \left(\sum_{k=-\infty}^{\infty} (F_{X_i}(10^k \cdot (d+1)) - F_{X_i}(10^k \cdot d)) \right)$$

$$= \sum p_i \cdot Pr(D_{X_i} = d)$$

$$= \sum p_i \cdot \log_{10}\left(1 + \frac{1}{d}\right)$$

Proof

$$= \sum p_i \cdot \log_{10}\left(1 + \frac{1}{d}\right)$$

$$= \sum \log_{10}\left(1 + \frac{1}{d}\right)^{p_i}$$

$$= \log_{10}\left(\left(1 + \frac{1}{d}\right)^{p_1} \cdot \ldots \cdot \left(1 + \frac{1}{d}\right)^{p_n}\right)$$

$$= \log_{10}\left(1 + \frac{1}{d}\right)^{p_1 + \ldots + p_n}$$

$$= \log_{10}\left(1 + \frac{1}{d}\right)$$

Thus, we have

$$Pr(D_Y = d) = \log_{10}\left(1 + \frac{1}{d}\right), \qquad d \in \{1, \ldots, 9\}$$

Thus, a mixture distribution of distributions that obey Benford’s law also obeys Benford’s law.

Demonstrate Using Data

Mixture Distribution using various mathematical sequences

Let Y be a mixture distribution where:

- $X_1, \dots, X_4$ are distributed as:
  - powers of 2
  - factorials
  - Fibonacci numbers
  - Bell numbers

  (these are all distributions that obey Benford's law)

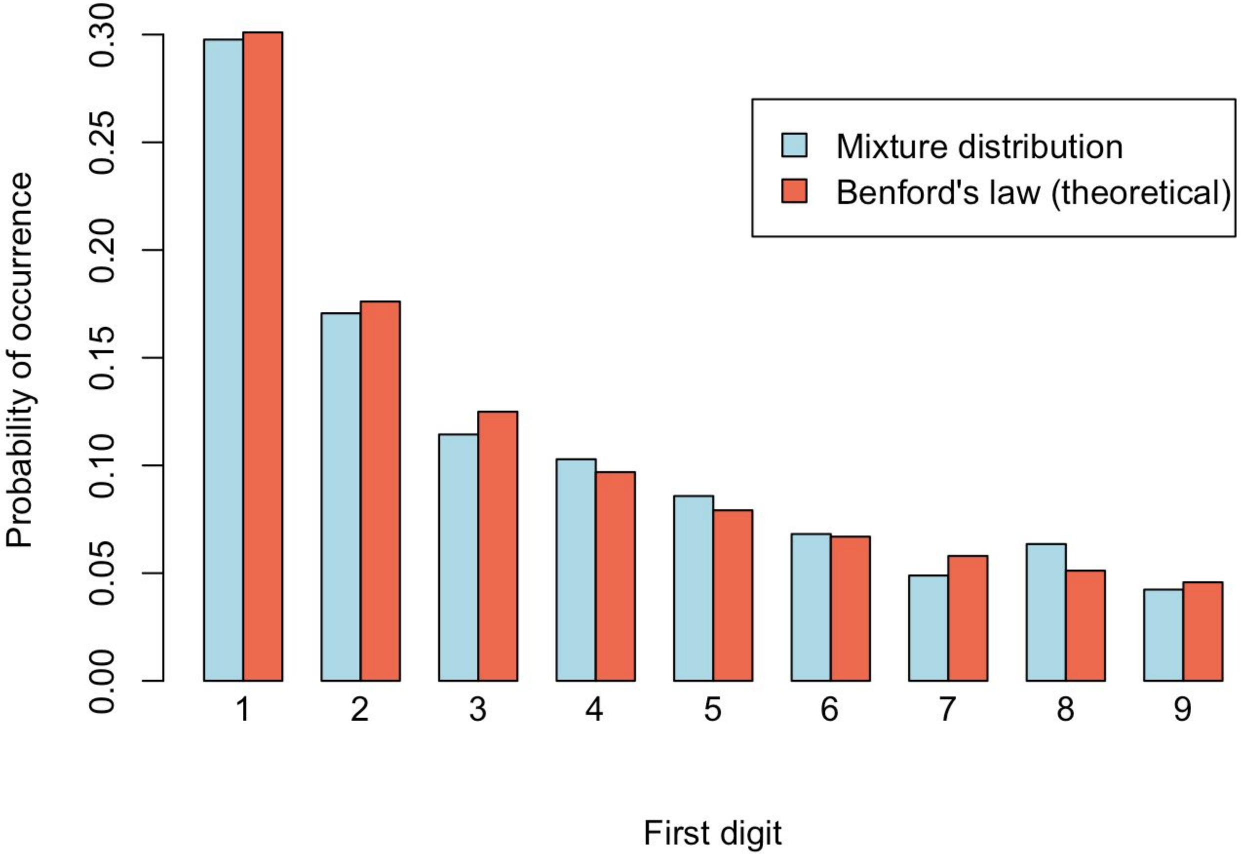
- $p_1, \dots, p_4 = 0.363, 0.017, 0.274, 0.345$
  - randomly generated
  - sum to 1

Powers of 2	FirstDigit	Factorials	FirstDigit	Fibonacci	FirstDigit	Bell	FirstDigit
1	0	1	0	1	0	1	0
2	2	1	1	1	1	2	2
4	4	2	2	2	2	5	5
8	8	6	6	3	3	15	1
16	1	24	2	5	5	52	5
32	3	120	1	8	8	203	2
64	6	720	7	13	1	877	8
128	1	5040	5	21	2	4140	4
256	2	40320	4	34	3	21147	2
512	5	362880	3	55	5	115975	1
1024	1	3628800	3	89	8	678570	6
2048	2	39916800	3	144	1	4213597	4
4096	4	479001600	4	233	2	27644437	2
8192	8	6227020800	6	377	3	190899322	1
16384	1	8.7178E+10	8	610	6	1382958545	1
32768	3	1.31E+12	1	987	9	1.048E+10	1
65536	6	2.09E+13	2	1597	1	8.2865E+10	8
131072	1	3.56E+14	3	2584	2	6.82E+11	6
262144	2	6.40E+15	6	4181	4	5.83E+12	5
524288	5	1.22E+17	1	6765	6	5.17E+13	5
1048576	1	2.43E+18	2	10946	1	4.75E+14	4
2097152	2	5.11E+19	5	17711	1	4.51E+15	4
4194304	4	1.12E+21	1	28657	2	4.42E+16	4
8388608	8	2.59E+22	2	46368	4	4.46E+17	4
16777216	1	6.20E+23	6	75025	7	4.64E+18	4
33554432	3	1.55E+25	1	121393	1	4.96E+19	4
67108864	6	4.03E+26	4	196418	1	5.46E+20	5
134217728	1	1.09E+28	1	317811	3	6.16E+21	6
268435456	2	3.05E+29	3	514229	5	7.13E+22	7
536870912	5	8.84E+30	8	832040	8	8.47E+23	8
1073741824	1	2.65E+32	2	1346269	1	1.03E+25	1
2147483648	2	8.22E+33	8	2178309	2	1.28E+26	1
4294967296	4	2.63E+35	2	3524578	3	1.63E+27	1
8589934592	8	8.68E+36	8	5702887	5	2.12E+28	2
1.718E+10	1	2.95E+38	2	9227465	9	2.82E+29	2
3.436E+10	3	1.03E+40	1	14930352	1	3.82E+30	3
6.8719E+10	6	3.72E+41	3	24157817	2	5.29E+31	5
1.37E+11	1	1.38E+43	1	39088169	3	7.46E+32	7
2.75E+11	2	5.23E+44	5	63245986	6	1.07E+34	1
5.50E+11	5	2.04E+46	2	102334155	1	1.57E+35	1
1.10E+12	1	8.16E+47	8	165580141	1	2.35E+36	2
2.20E+12	2	3.35E+49	3	267914296	2	3.57E+37	3
4.40E+12	4	1.41E+51	1	433494437	4	5.53E+38	5
8.80E+12	8	6.04E+52	6	701408733	7	8.70E+39	8
1.76E+13	1	2.66E+54	2	1134903170	1	1.39E+41	1

- Used data containing the first 1000 numbers of each mathematical sequence for $X_1, \ldots, X_4$
- Generated 1 million simulated values of Y
  - For each value, picked distribution X_i with probability p_i and then picked a random value for X_i from the data
- Found the distribution of the first digits of Y

First digit distribution of Mixture Distribution vs. Benford's Law

Mixture distribution using powers of 2, factorials, Fibonacci numbers, and Bell numbers



Another Example

Mixture Distribution using births in each U.S. county in 2019

Let Y be a mixture distribution where:

- each X_i is distributed according to the total number of births in each county of a certain U.S. state in 2019
  - X_1 , … , X_{51} for 50 U.S. states + Washington D.C.
  - (population per county almost follows Benford's law)

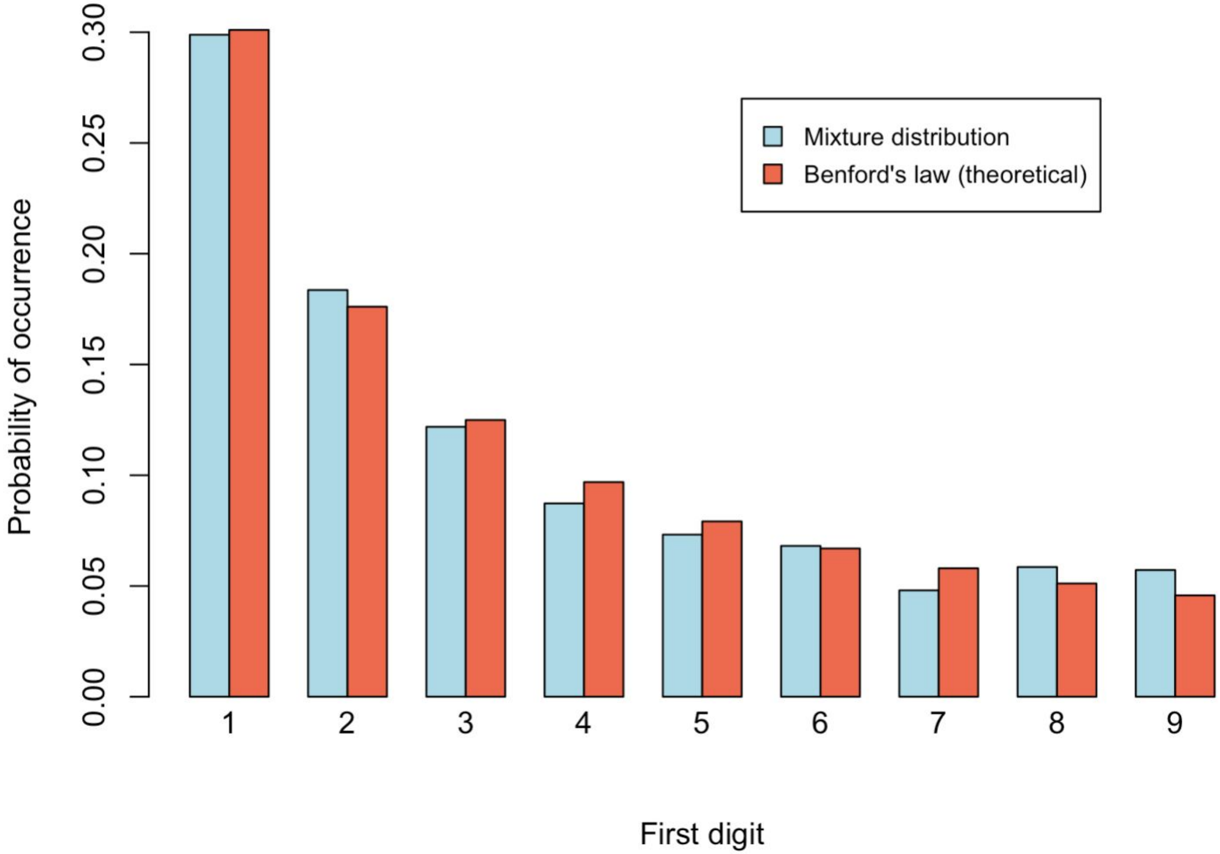
- each p_i = total births in state *i* in 2019 / total births in the U.S. in 2019
  - p_i represents the probability that a random birth occurred state *i*

STNAME	CTYNAME	BIRTHS2019	FirstDigit
Alabama	Alabama	57313	5
Alabama	Autauga County	624	6
Alabama	Baldwin County	2304	2
Alabama	Barbour County	256	2
Alabama	Bibb County	240	2
Alabama	Blount County	651	6
Alabama	Bullock County	109	1
Alabama	Butler County	213	2
Alabama	Calhoun County	1269	1
Alabama	Chambers County	354	3
Alabama	Cherokee County	222	2
Alabama	Chilton County	551	5
Alabama	Choctaw County	133	1
Alabama	Clarke County	266	2
Alabama	Clay County	143	1
Alabama	Cleburne County	187	1
Alabama	Coffee County	599	5
Alabama	Colbert County	630	6
Alabama	Conecuh County	123	1
Alabama	Coosa County	89	8
Alabama	Covington County	413	4
Alabama	Crenshaw County	137	1
Alabama	Cullman County	990	9
Alabama	Dale County	648	6
Alabama	Dallas County	438	4
Alabama	DeKalb County	795	7
Alabama	Elmore County	932	9
Alabama	Escambia County	420	4
Alabama	Etowah County	1175	1
Alabama	Fayette County	174	1
Alabama	Franklin County	432	4
Alabama	Geneva County	278	2
Alabama	Greene County	96	9
Alabama	Hale County	190	1
Alabama	Henry County	167	1
Alabama	Houston County	1304	1
Alabama	Jackson County	562	5
Alabama	Jefferson County	8422	8
Alabama	Lamar County	159	1
Alabama	Lauderdale County	876	8
Alabama	Lawrence County	349	3
Alabama	Lee County	1825	1
Alabama	Limestone County	1014	1
Alabama	Lowndes County	116	1
Alabama	Macon County	172	1
Alabama	Madison County	4242	4

- Used data for the number of births in each U.S. county in 2019, separated by state for $X_1, \ldots, X_{51}$
- Generated 1 million simulated values of Y
    - For each value, picked distribution X_i with probability p_i and then picked a random value for X_i from the data
- Found the distribution of the first digits of Y

First digit distribution of Mixture Distribution vs. Benford's Law

Mixture distribution using births in each U.S. county in 2019



Pearson Chi-Square Tests

Rationale for chi-square test

- Benford's law is a multinomial distribution with m frequencies
- The likelihood ratio test asymptotically follows a chi-square distribution which and is approximately equivalent to the Pearson/Wald chi-square test
- For one-sample tests against the theoretical Benford distribution we will use the **Pearson test**
- For two-sample tests we will use the **Wald test**

General case (one-sample Pearson test)

Ref: Section 7.9.1

Let X be a multinomial distribution with m frequencies. To test that X follows a known distribution, we test the hypotheses below:

$H_0 : p_1 = p_{10}, ..., p_m = p_{m0}$
$H_a : p_j \neq p_{j0}$ for at least one j $\leq m$

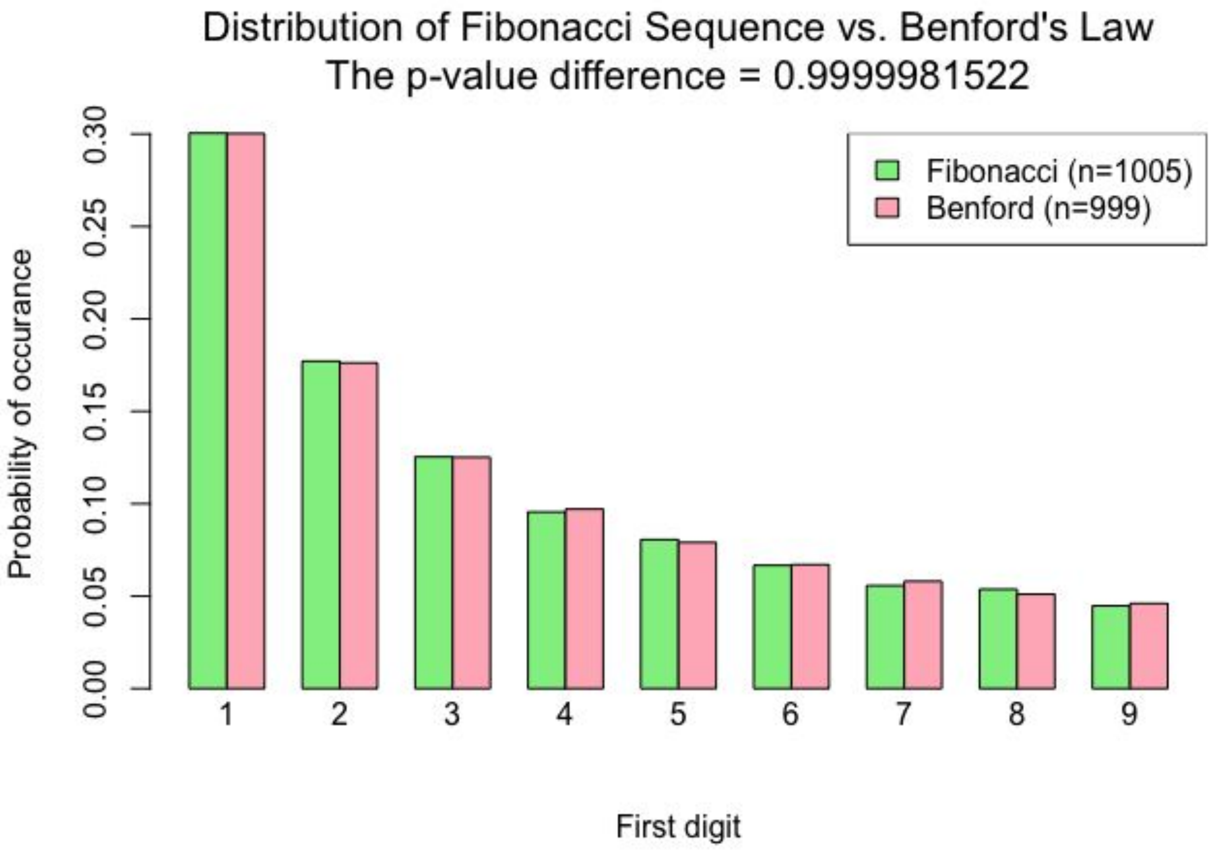
To test the null hypothesis, we conduct the Pearson chi-square test with $m - 1$ degrees of freedom:

$$n \sum_{j=1}^{m} \frac{(\widehat{p_j} - p_{j0})^2}{p_{j0}} \simeq \chi^2(m-1)$$

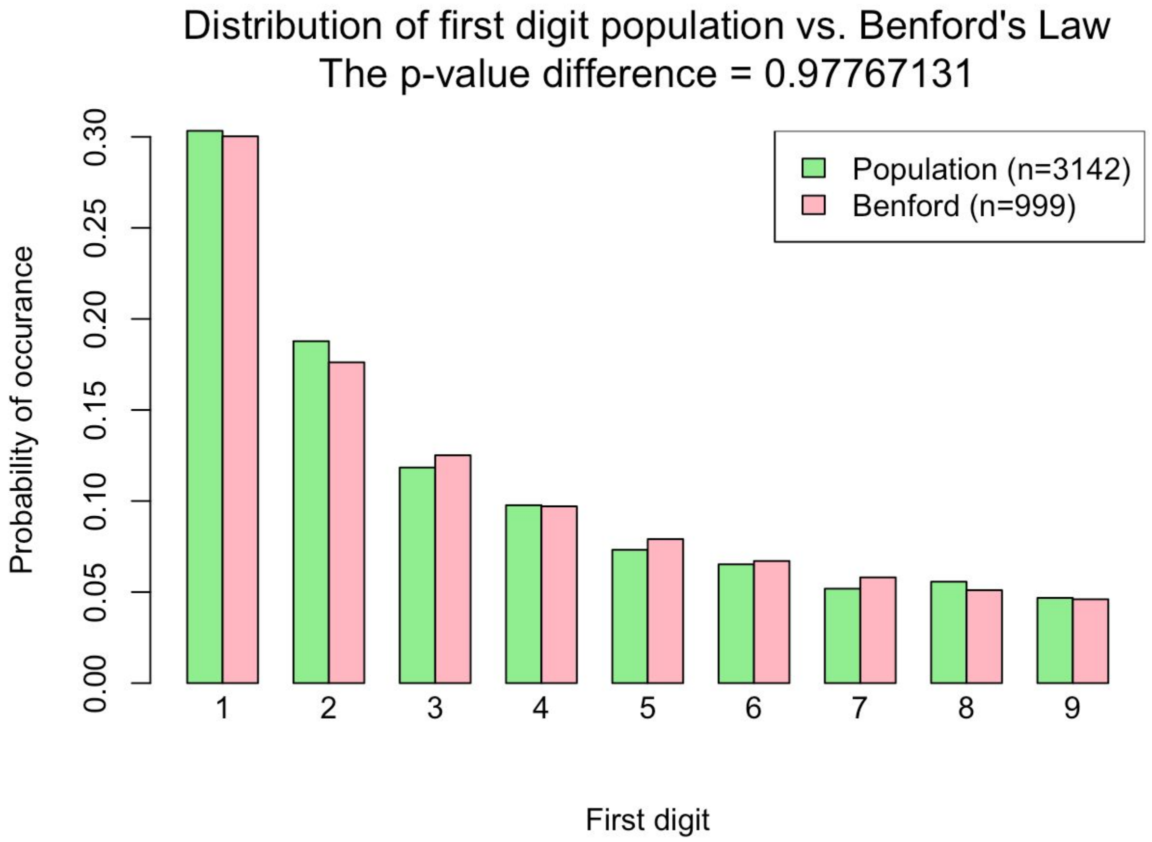
$\hat{p_j}$ is the MLE, defined as $\hat{p_j} = \frac{X_j}{n}$

Demonstrate Using Data
(distributions that follow)

Pearson Test for Fibonacci vs. Benford's Law

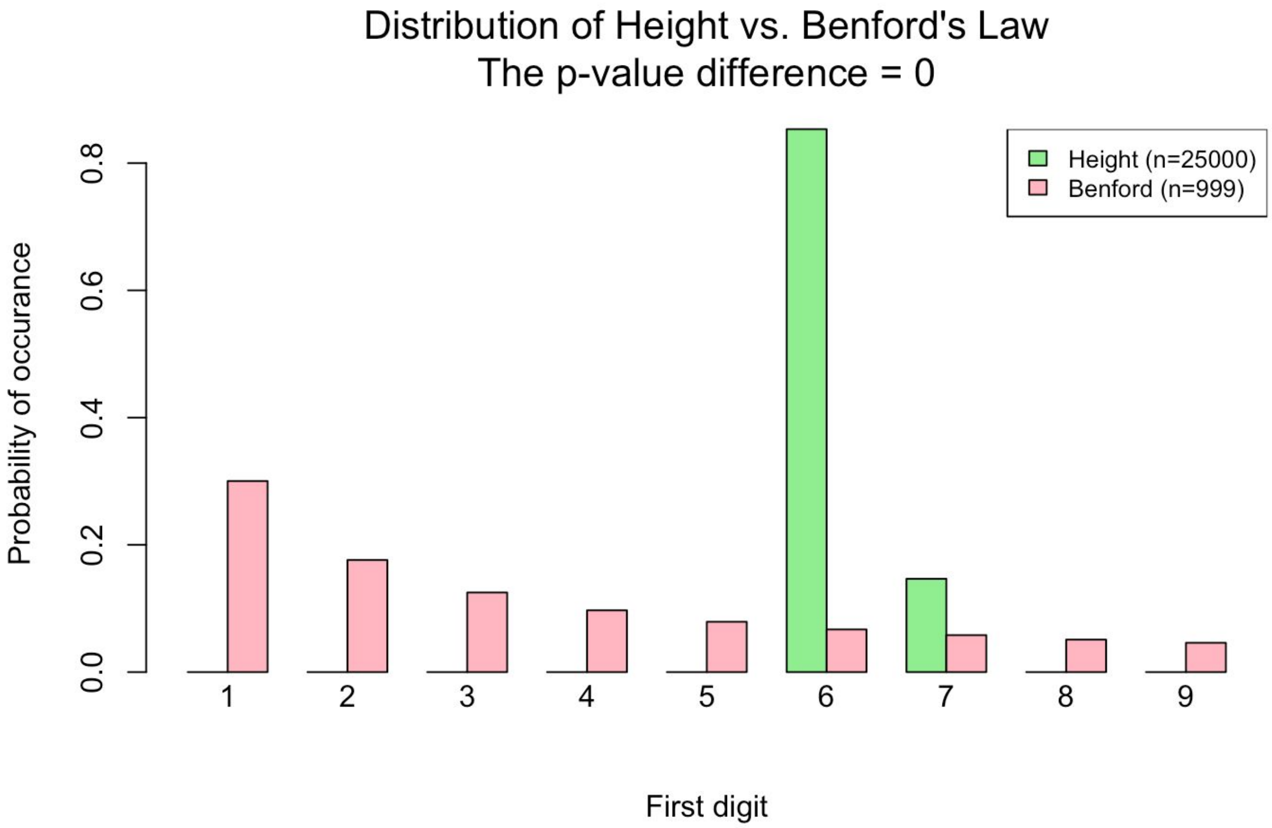


Pearson Test for population by county vs. Benford's Law

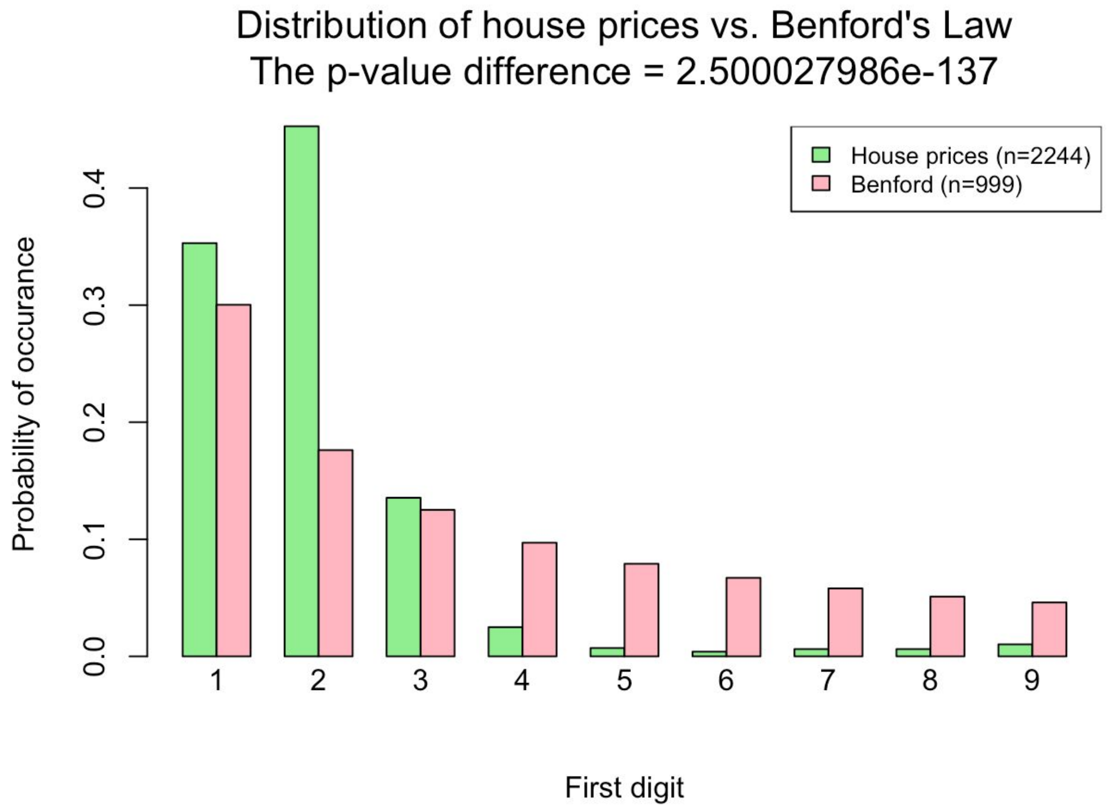


Demonstrate Using Data (distributions that don’t follow)

Pearson Test for normal data vs. Benford’s Law



Another normal distribution vs. Benford's Law



Themes for distributions that follow Benford’s Law

- Sufficient sample size
- Large span of number values
- 3+ orders of magnitude
- Non human-assigned numbers
- Right-skewed data
- Scale invariance

General case (two-sample Wald test)

Let X_j and Y_j be two frequency distributions with sample size n_X and n_Y:

$H_0 : p_{1X} = p_{1Y}, ..., p_{mX} = p_{mY}$
$H_a : p_{jX} \neq p_{jY}$ for at least one j $\leq m$

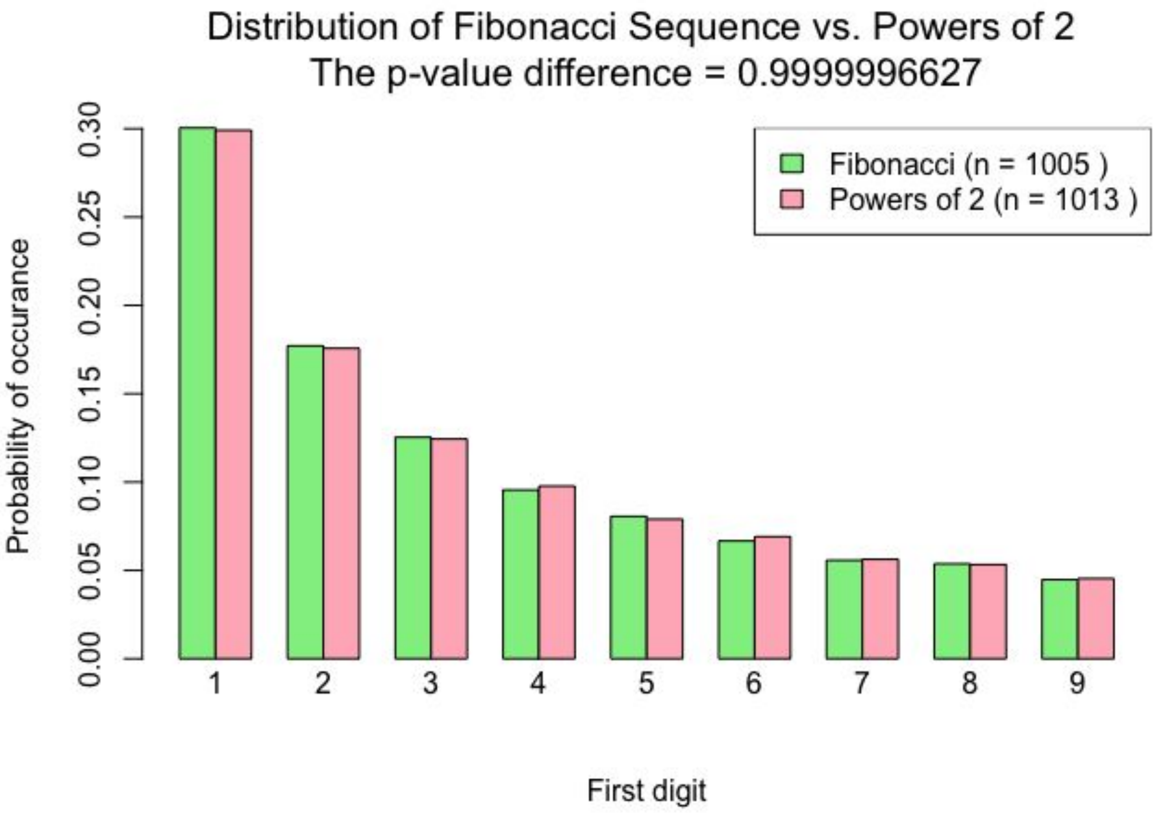
Let $\hat{p_j}$ represent the probability estimate under the null hypothesis that the probabilities within the two distributions are the same.

$\widehat{p_j} = \frac{X_j+Y_j}{n_X+n_Y}$

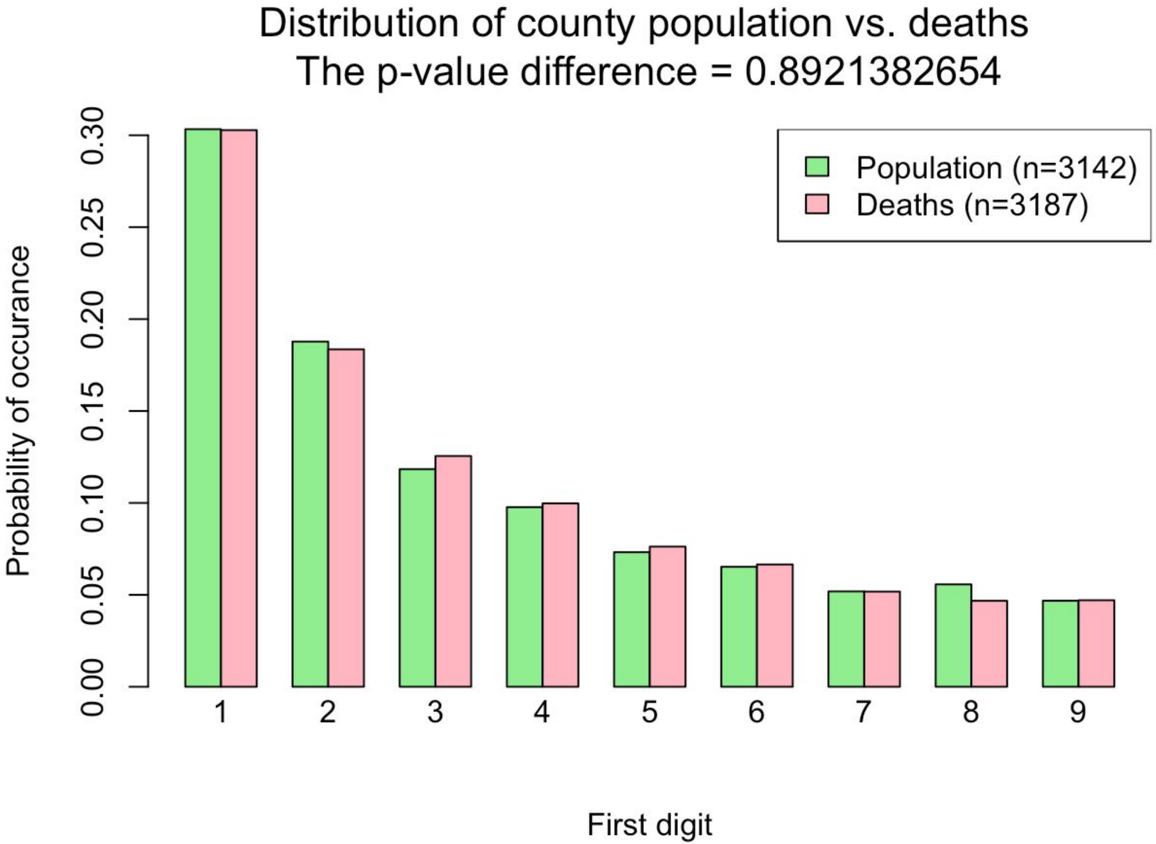
To test the null hypothesis, we conduct the Wald test which follows a chi-square distribution with degrees of freedom $m-1$:

$$\frac{1}{1/n_x + 1/n_y} = \sum_{j=1}^{m} \frac{(\widehat{p}_{jX} - \widehat{p}_{jY})^2}{\widehat{p_j}} \simeq \chi^2(m-1)$$

Two-sample Wald test for first-digits of two distributions following Benford's law



Two-sample Wald test for first digits of county population vs. deaths



Conditions for Conformance

A Word of Caution

Note the difference between the Benford-conforming digit distribution and the Benford-conforming random variable.

$$X = 0.101142\ldots$$

Only the red digit has the first digit distribution:

$$\Pr(D = d) = \log_{10}(1 + 1/d).$$

The r.v. itself can have a much less well-behaved distribution.

Leemis et al. (2000)

A Criterion for Benford-ness

A Benford-conforming random digit D has pmf (for d = 1, 2, …, 9)

$$\Pr(D = d) = \log_{10}(1 + 1/d).$$

and hence cdf

$$\begin{aligned} F_D(d) = \Pr(D \le d) &= \sum_{i=1}^{d} \log_{10}(1 + 1/i) \\ &= \sum_{i=1}^{d} \log_{10}\left(\frac{1+i}{i}\right) \\ &= \log_{10}\left(\frac{2}{1} \cdot \frac{3}{2} \cdot \cdots \cdot \frac{1+d}{d}\right) \\ &= \log_{10}(1 + d). \end{aligned}$$

Leemis et al. (2000)

A Criterion for Benford-ness

Recall from Math 40 that if we have some target cdf F_X, we can generate samples of this distribution by defining

$$X = F_X^{-1}(U) \text{ where } U \sim \mathcal{R}(0,1).$$

In this case, we compute the inverse of the previous cdf:

$$F_D^{-1}(x) = \lceil 10^x - 1 \rceil.$$

We can generate Benford-conforming first digits with

$$D = \lceil 10^U - 1 \rceil$$

or equivalently,

$$D = \lfloor 10^U \rfloor. \qquad (*)$$

Leemis et al. (2000)

A Criterion for Benford-ness

We can compute the first significant digit of any positive value X as follows:

$$M = \lfloor \log_{10}(X) \rfloor \qquad\qquad D = \lfloor X \cdot 10^{-M} \rfloor = \lfloor 10^{\log_{10} X - M} \rfloor.$$

Then D is the first significant digit.

Ex. $X = 365$. Then

$$X = 365 = \underbrace{3.65}_{X \cdot 10^{-M}} \times 10^{\overset{M}{2}}.$$

Leemis et al. (2000)

A Criterion for Benford-ness

D is Benford-conforming if

$$D = \lfloor 10^{\log_{10} X - M} \rfloor = \lfloor 10^{U} \rfloor$$

or in other words, if

$$Z = \log_{10} X - M \sim \mathcal{R}(0, 1).$$

Even more explicitly, we can check that

$$\begin{aligned} F_Z(z) &= \Pr(Z \le z) \\ &= \sum_{k=-\infty}^{\infty} \Pr(10^k \le X < 10^{k+1}) \cdot \Pr(\log_{10} X - k \le z \mid 10^k \le X < 10^{k+1}) \\ &= z \end{aligned}$$

holds.

Leemis et al. (2000)

Benford-ness Criterion in Practice

Ex. Consider W with pdf given by

$$f_W(x) = \begin{cases} x & 0 < x < 1 \\ 2 - x & 1 \le x < 2. \end{cases}$$

Define $X = 10^W$. Does X conform to Benford's Law?

Leemis et al. (2000)

Benford-ness Criterion in Practice

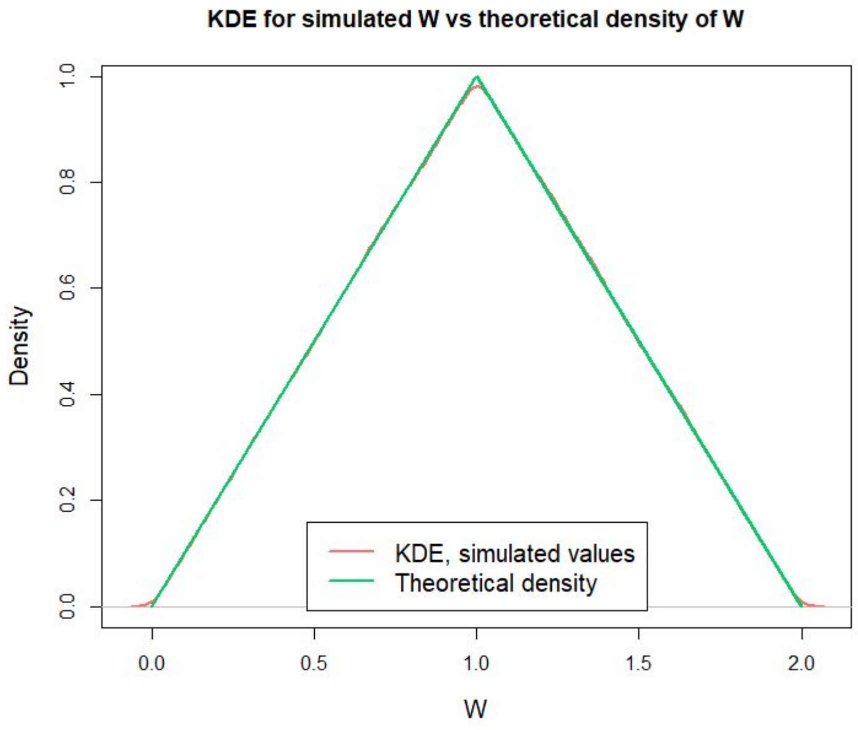
Solution. Apply the criterion:

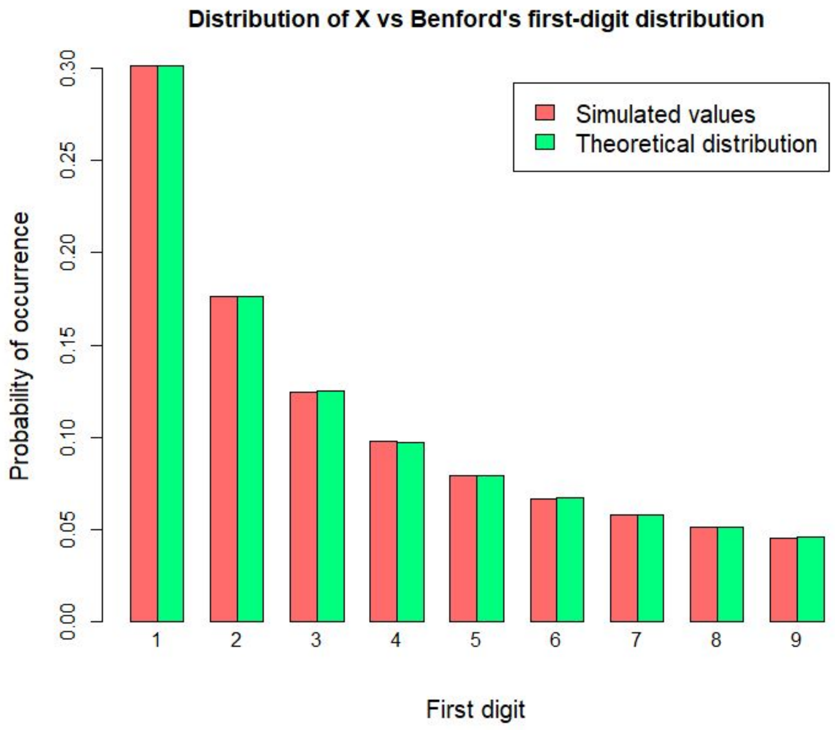
$$
\begin{aligned}
F_Z(z) &= \sum_{k=-\infty}^{\infty} \Pr(10^k \le X < 10^{k+1}) \cdot \Pr(W - k \le z \mid 10^k \le X < 10^{k+1}) \\
&= 0.5 \cdot \Pr(W \le z \mid W \le 1) + 0.5 \cdot \Pr(W \le z + 1 \mid 1 \le W < 2) \\
&= \frac{0.5 \int_0^z x\,dx}{0.5} + \frac{0.5 \int_1^{z+1} (2 - x)\,dx}{0.5} \\
&= \left[\frac{x^2}{2}\right]_0^z + \left[2x - \frac{x^2}{2}\right]_1^{z+1} \\
&= \frac{z^2}{2} + \frac{2z - z^2}{2} \\
&= z.
\end{aligned}
$$

So X should indeed conform to Benford's Law.

Leemis et al. (2000)

Benford-ness Criterion in Practice





Leemis et al. (2000)

Other Observations

A few more empirical observations the paper makes (without proof):

1. If W is distributed with a single extreme mode, Benford's law will be poorly fit

2. W has certain limiting distributions that conform well (e.g. W normally distributed with variance tending to infinity)

3. Many distributions' conformity is highly parameter-specific (certain parameters will conform very closely while others not at all)

Textbook, p. 142–143

Other Observations

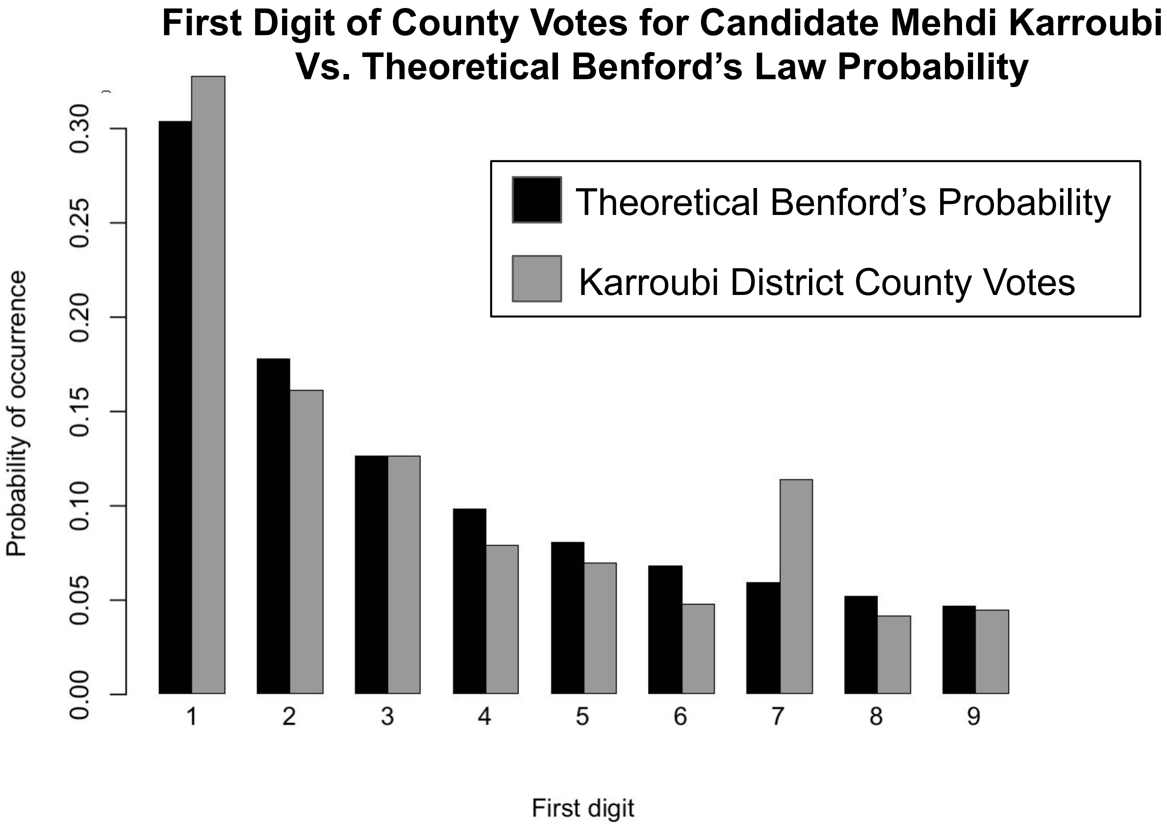
A couple related observations from the textbook:

1. If W is uniformly distributed, and the endpoints a and b are far apart, then X is an almost exact match for Benford's law.

2. If W follows the standard normal distribution, then X (note. X is lognormal) is an almost exact match for Benford's law.

Applications

2009 Iranian Presidential Election Results



"... there are significantly more vote totals for Karoubi beginning with digit "7" than would be expected by Benford's Law."

Other Applications

- Tax fraud
- Greece election
- People fabricating coefficients in academic papers

Thank you for your time!