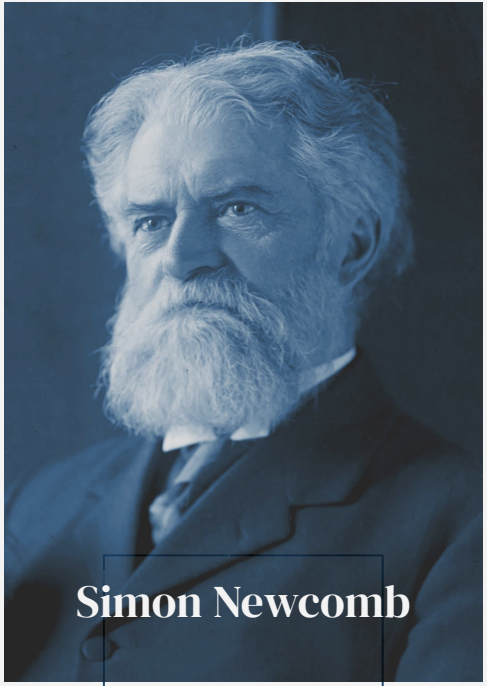
Benford's Law

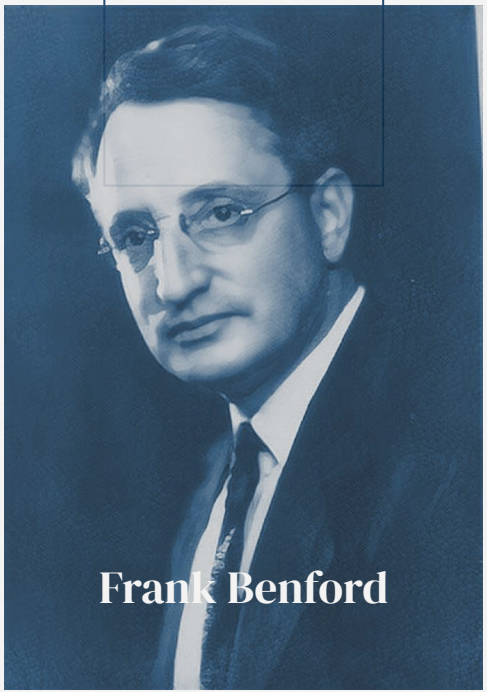
Carson Gampell, Gibran Erlangga, Shrey Khetrepal, Jessica Serrao

Table of Contents

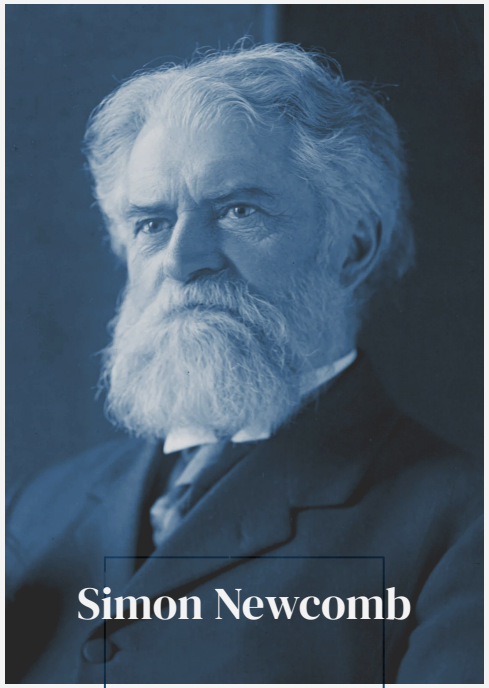
01 Problem Overview

The names you should know...





The pattern was first discovered by Newcomb back in 1880s.



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

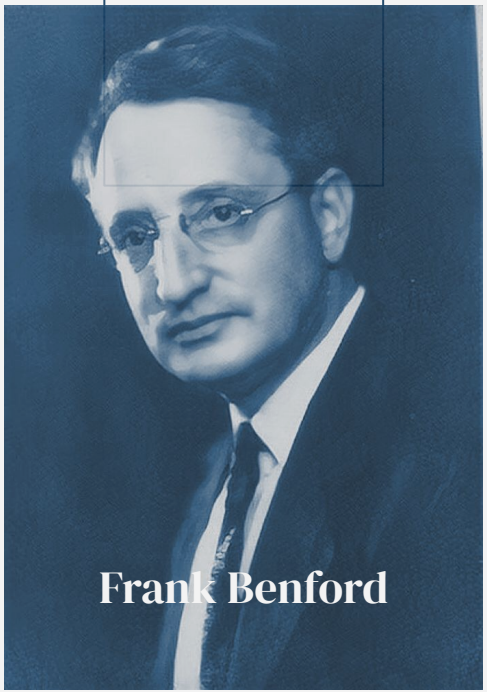
BY SIMON NEWCOMB.

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

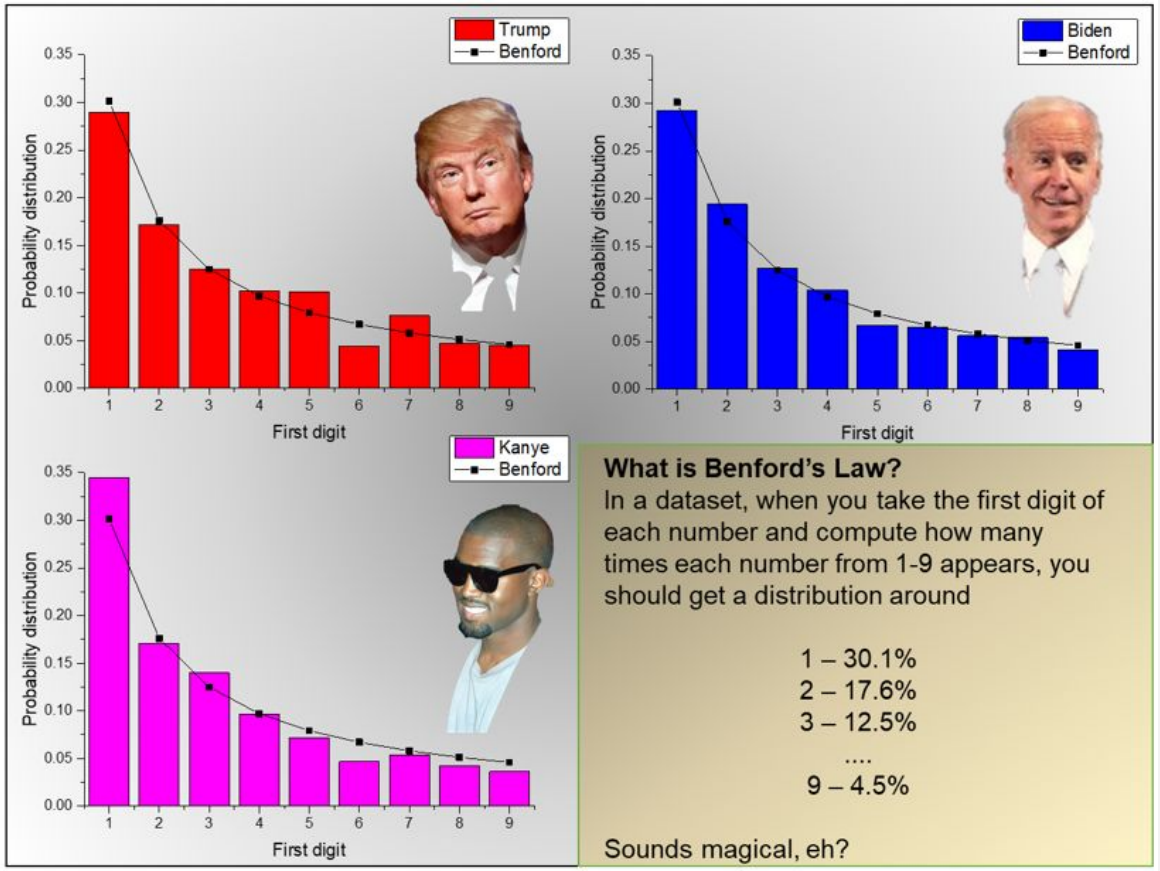
.. the pages of numbers whose leading digit was 1 were **more worn** than the pages of numbers whose leading digit was 9.

Frank Benford rediscovered it 50 years later as the **Law of Anomalous Numbers.**

Title	1	2	3	4	5	6	7	8	9	Count
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
Newspapers	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
Black Body	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
$n, n^2, \ldots, n!$	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
Benford's Law	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6	



Votes numbers for Trump, Biden, and West follow Benford's Law.



source: Reddit

Application of Benford's Law

02

Current Research utilizing Benford's Law

Open Access Article

Application of Benford's Law on Cryptocurrencies

by Jernej Vičič [1,2,*,†] and Aleksandar Tošić [1,3,†]

1 Faculty of Mathematics Natural Sciences and Information Technologies, University of Primorska, 6000 Koper, Slovenia
2 Research Centre of the Slovenian Academy of Sciences and Arts, The Fran Ramovš Institute, 1000 Ljubljana, Slovenia
3 InnoRenew CoE, 6310 Izola, Slovenia
* Author to whom correspondence should be addressed.
† These authors contributed equally to this work.

Academic Editor: Jani Merikivi

J. Theor. Appl. Electron. Commer. Res. **2022**, 17(1), 313-326; https://doi.org/10.3390/jtaer17010016

Received: 7 November 2021 / Revised: 7 February 2022 / Accepted: 8 February 2022 / Published: 25 February 2022

(This article belongs to the Special Issue Blockchain Commerce Ecosystem)

View Full-Text | Download PDF | Browse Figures | Review Reports | Version Export

Abstract

The manuscript presents a study of the possibility of use of Benford's law conformity test, a we
fraud discovery, on a new domain: the discovery of anomalies (possibly fraudulent behaviour)
transactions. Blockchain-based currencies or cryptocurrencies have become a global phenor
disruptive technology, and a new investment vehicle. However, due to their decentralized na
presented regulators with difficulties in finding a balance between nurturing innovation, and
concerns about illicit activity have forced regulators to seek new ways of detecting, analyzin
blockchain transactions. Extensive research on machine learning, and transaction graph a
track suspicious behaviour. However, having a macro view of a public ledger is equally imp
grained analysis. Benford's law, the law of first digit, has been extensively used as a tool t
other use cases exist). The basic motivation that drove our research presented in this pa

Cornell University

arXiv > cs > arXiv:2203.13352

Computer Science > Computation and Language

[Submitted on 24 Mar 2022]

Does human speech follow Benford's Law?

Leo Hsu, Visar Berisha

Researchers have observed that the frequencies of leading digits in many man-made and naturally occurring datasets follow a logarithmic curve, with digits that start with the number 1 accounting for ~ 30% of all numbers in the dataset and digits that start with the number 9 accounting for ~ 5% of all numbers in the dataset. This phenomenon, known as Benford's Law, is highly repeatable and appears in lists of numbers from electricity bills, stock prices, tax returns, house prices, death rates, lengths of rivers, and naturally occurring images. In this paper we demonstrate that human speech spectra also follow Benford's Law. We use this observation to motivate a new set of features that can be efficiently extracted from speech and demonstrate that these features can be used to classify between human speech and synthetic speech.

Subjects: Comput...

...ing (cs.LG)

Information Sciences

Volume 582, January 2022, Pages 369-381

ELSEVIER

Feature selection using Benford's law to support detection of malicious social media bots

Innocent Mbona, Jan H.P. Eloff

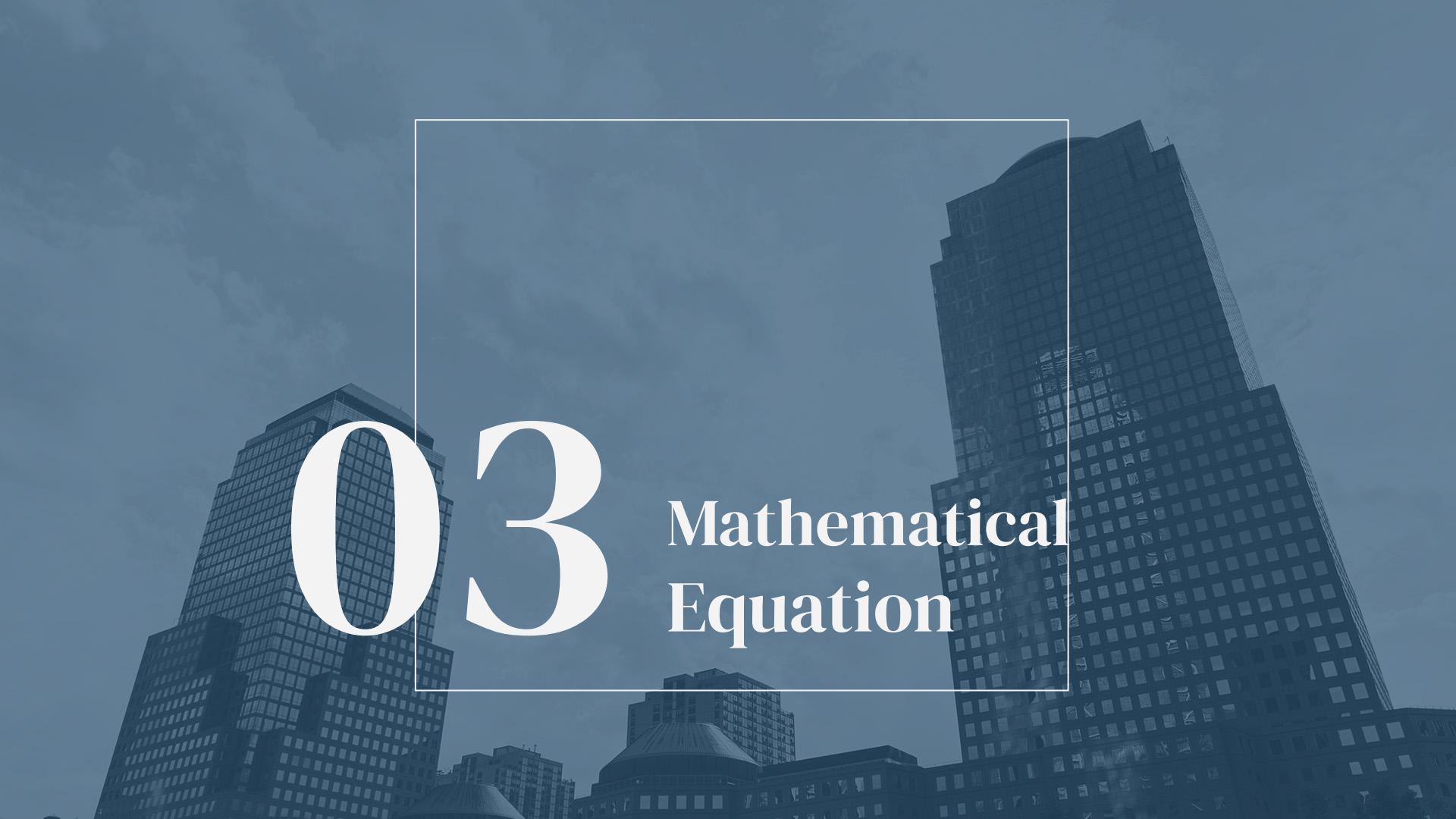
Show more

+ Add to Mendeley | Share | Cite

https://doi.org/10.1016/j.ins.2021.09.038 Get rights and content

Abstract

The increased amount of high-dimensional imbalanced data in online social networks challenges existing feature selection methods. Although feature selection methods such as principal component analysis (PCA) are effective for solving high-dimensional imbalanced data problems, they can be computationally expensive. Hence, an effortless approach for identifying meaningful features that are indicative of anomalous behaviour between humans and malicious bots is presented herein. The most recent Twitter dataset that encompasses the behaviour



03 Mathematical Equation

What is Benford's Law ?

observation about the leading digits of numbers found in real-world, natural data sets

Intuitively, the leading digits of these numbers would be uniformly distributed so that each of the digits from 1 to 9 is equally likely to appear

often in natural data sets 1 occurs more frequently than 2, 2 more frequently than 3, and so on

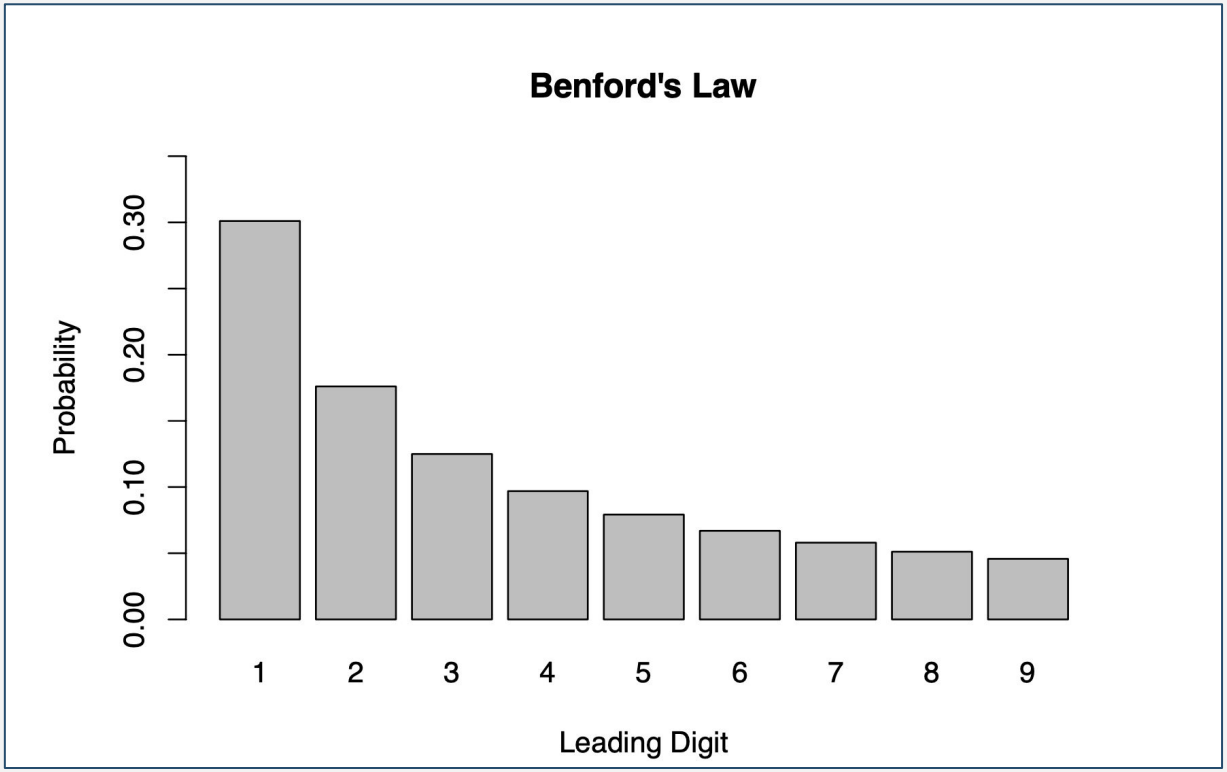
Benford's Law gives a prediction of the frequency of leading digits, using base-10 log, which decreases as digits increase from 1 to 9. A set of numbers is satisfied by Benford's Law if the leading digit D occurs with probability

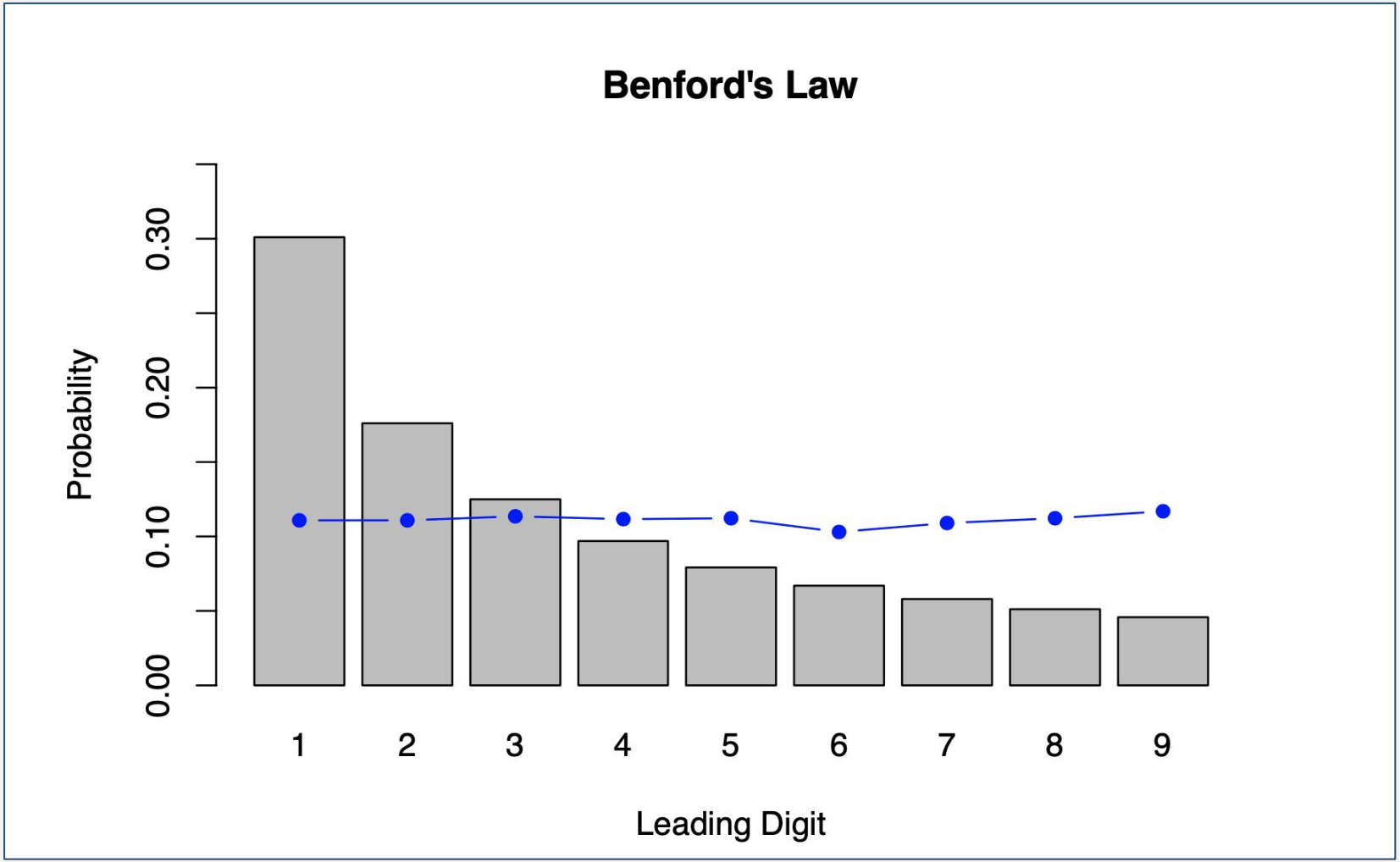
$$P(D = d) = log_{10}(1 + 1/d), \qquad d \in 1, 2, ..., 8, 9$$

```
Digits = c(1,2,3,4,5,6,7,8,9)
Probabilities = log10(1+1/1:9)
df <- data.frame(Digits, Probabilities)
print(df)
```

```
##   Digits Probabilities
## 1      1    0.30103000
## 2      2    0.17609126
## 3      3    0.12493874
## 4      4    0.09691001
## 5      5    0.07918125
## 6      6    0.06694679
## 7      7    0.05799195
## 8      8    0.05115252
## 9      9    0.04575749
```

```r
Benfords <- function(d) log10(1 + (1/d))
blawplot <- barplot(Benfords(Digits), names.arg = Digits, xlab = "Leading Digit",
        ylab = "Probability", main = "Benford's Law", ylim = c(0, .35))
```




Benford's Law
Probability
0.00
0.10
0.20
0.30
1
2
3
4
5
6
7
8
9
Leading Digit

these probabilities sum to 1

$$\sum_{d=1}^{9} log_{10}(1+\frac{1}{d}) = \sum_{d=1}^{9} log_{10}(\frac{d+1}{d})$$

$$= \sum_{d=1}^{9} log_{10}(d+1) - log_{10}(d)$$

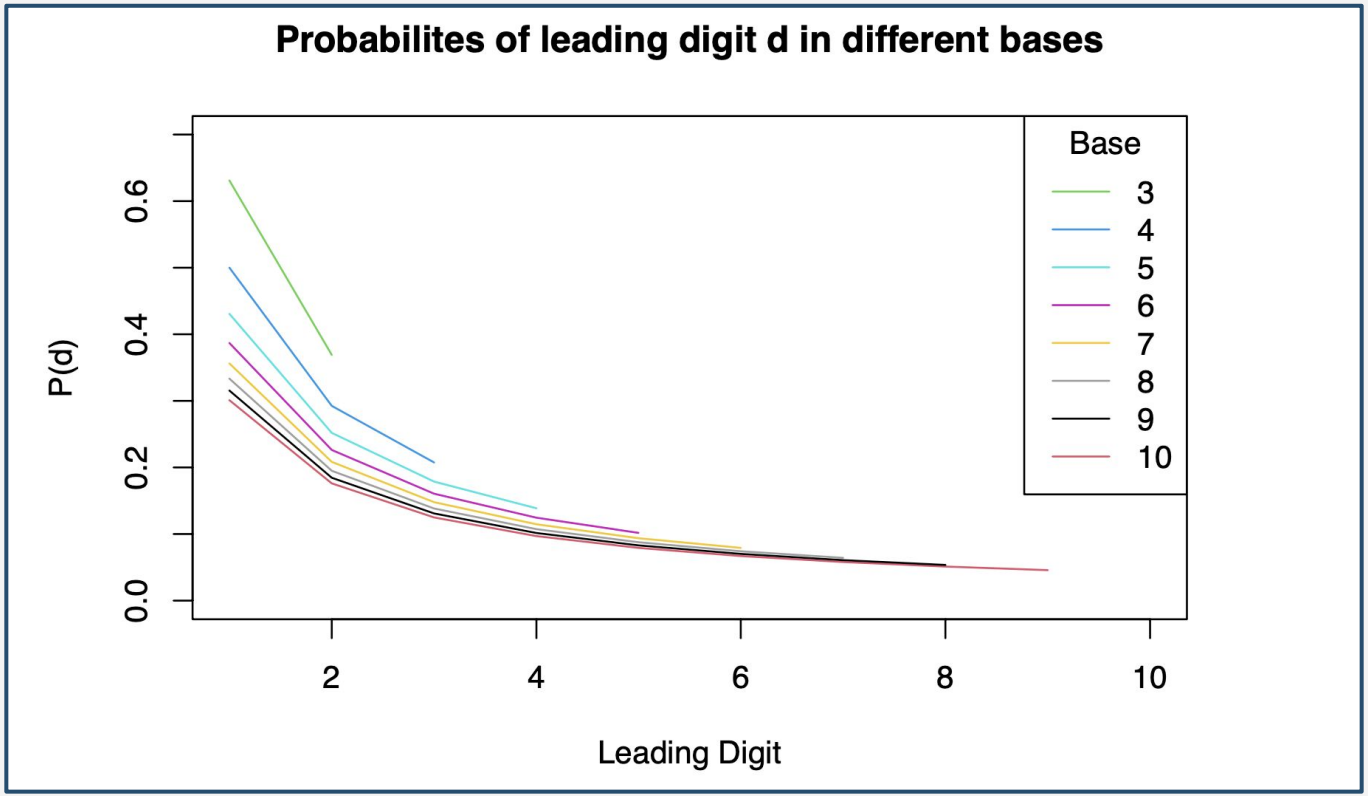
$$= log_{10}(10) - log_{10}(1)$$

$$= 1$$

Properties of Benford's Law

Benford's Law is applicable to other bases as well, taking the form

$$P(D = d) = log_b(1 + \frac{1}{d}) \quad \text{for other bases b} \geq 2$$



Properties of Benford's Law

Data the follows Benford's Law is also invariant to scaling. If a dataset is "Benford" then the set obtained by multiplying all the numbers in the original set by a fixed constant will also be "Benford."

Proof:

Let's set a positive integer d and constant c. For example, let's set c = 2 and d = 7.

If x is some number in the set, the probability that 2x starts with d is equal to the probability that x starts with the digits 5d, 5d+1, 5d+2, 5d+3 or 5d+4. In my example, since d = 7, then 35, 36, 37, 38 and 39 will start with 7 when multiplied by 2. If d = 12, then 60, 61, 62, 63 and 64 would start with 12 when multiplied by 2. Now, we can prove it for any d when c = 2.

Proof:

$$log_{10}(1+\frac{1}{5d})+log_{10}(1+\frac{1}{5d+1})+log_{10}(1+\frac{1}{5d+2})+log_{10}(1+\frac{1}{5d+3})+log_{10}(1+\frac{1}{5d+4})$$

$$=log_{10}(\frac{5d+5}{5})$$

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

Notes

The sequence of prime numbers follows Benford's Law.

The sequence of Fibonacci numbers follows Benford's Law.

The sequence of factorials also follows Benford's Law.

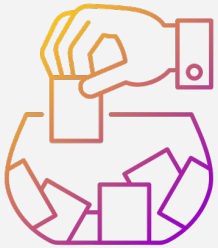
xn is Benford if and only if log(x) ∈ / Q

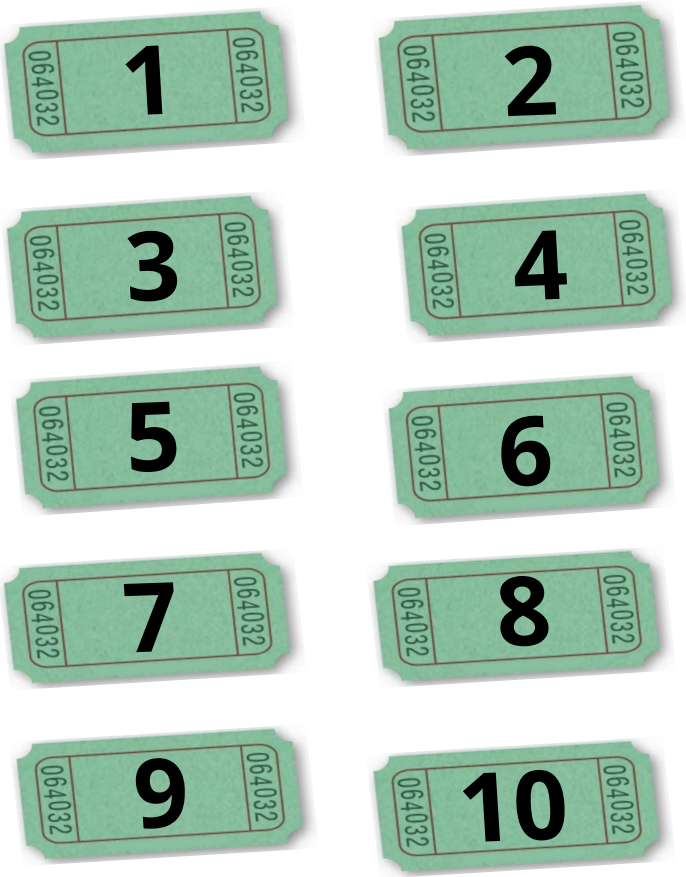
Intuition Behind the Law

04

Raffle

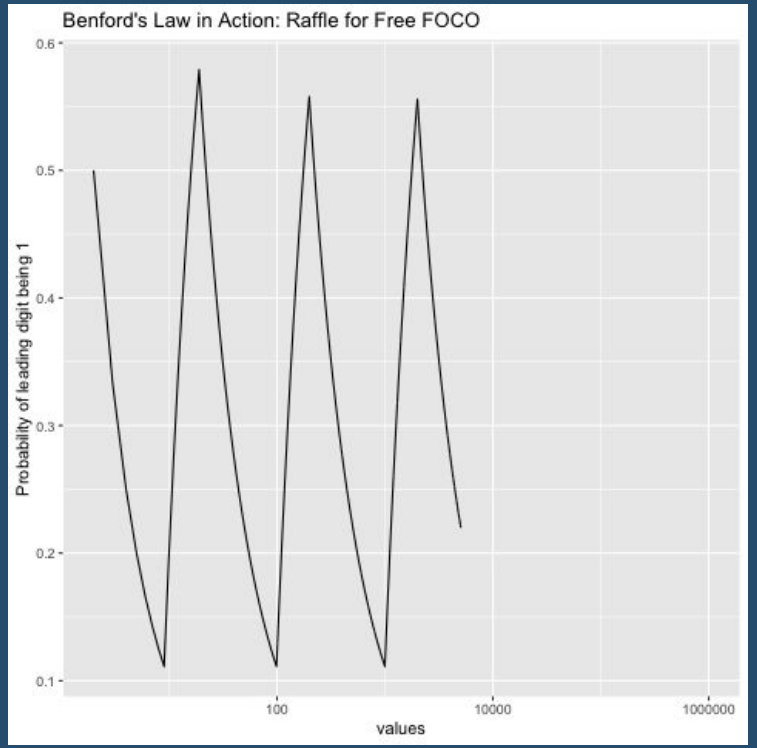
You enter a raffle to win free meals at FOCO for fall term!

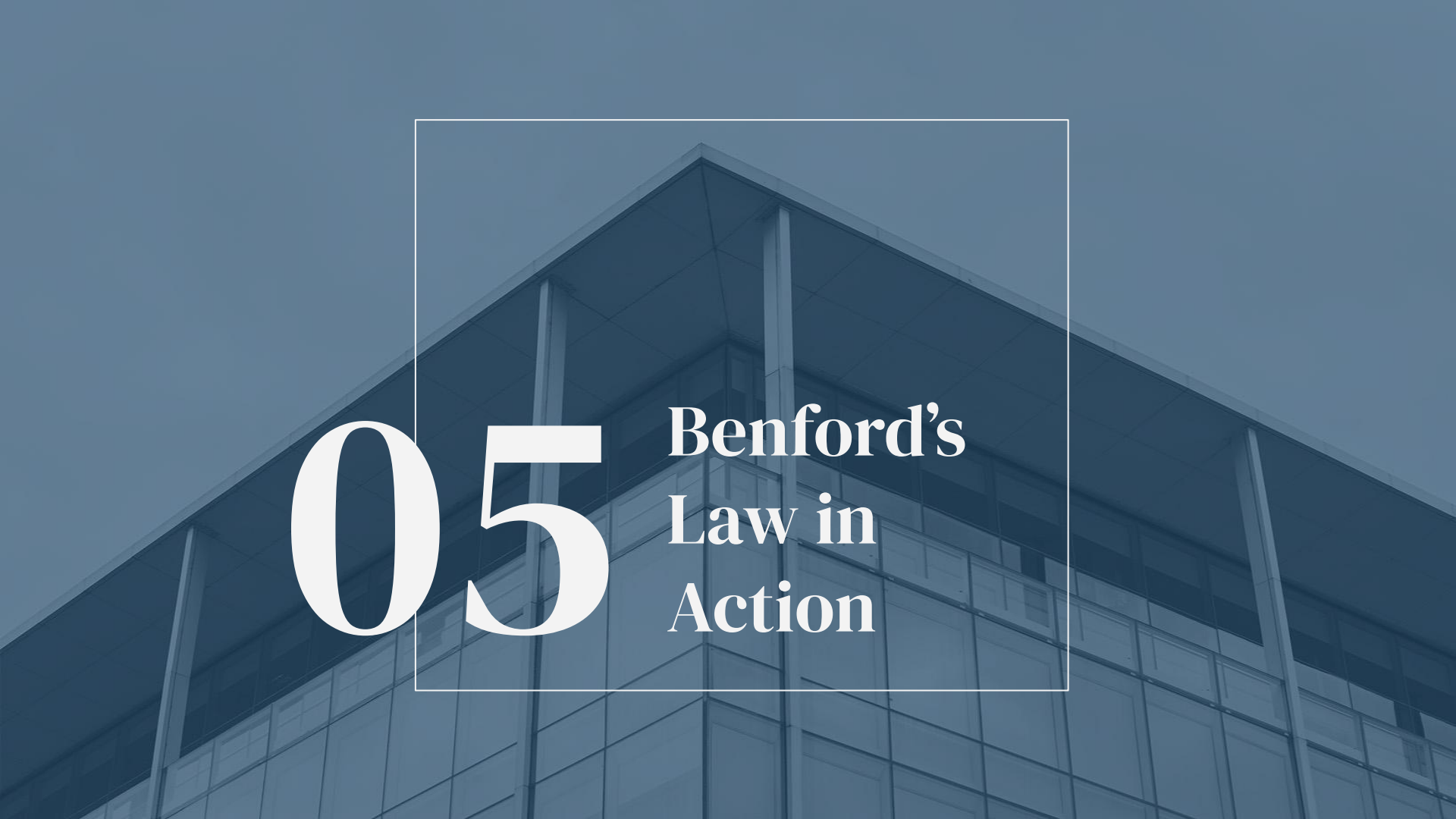




Number of tickets	Probability of leading digit being 1
2	50%
3	33% ↓
4	25% ↓
5	20% ↓
6	17% ↓
7	14% ↓
8	12.5% ↓
9	11% ↓
10	20% ↑

Benford's Law Raffle Simulation





05 Benford's Law in Action

Datasets

Forest Fires

The fires dataset was downloaded from Kaggle and describes forest fires that occurred in Turkey from 1988 to 2020. The data can be found here.

Air Pollution

The air pollution dataset was downloaded from Kaggle and describes the death rates due to air pollution in several countries. The data can be found here.

Soccer

The soccer data was downloaded from Github and describes various European soccer leagues and the associated stats for each player for the season. The data can be found here.

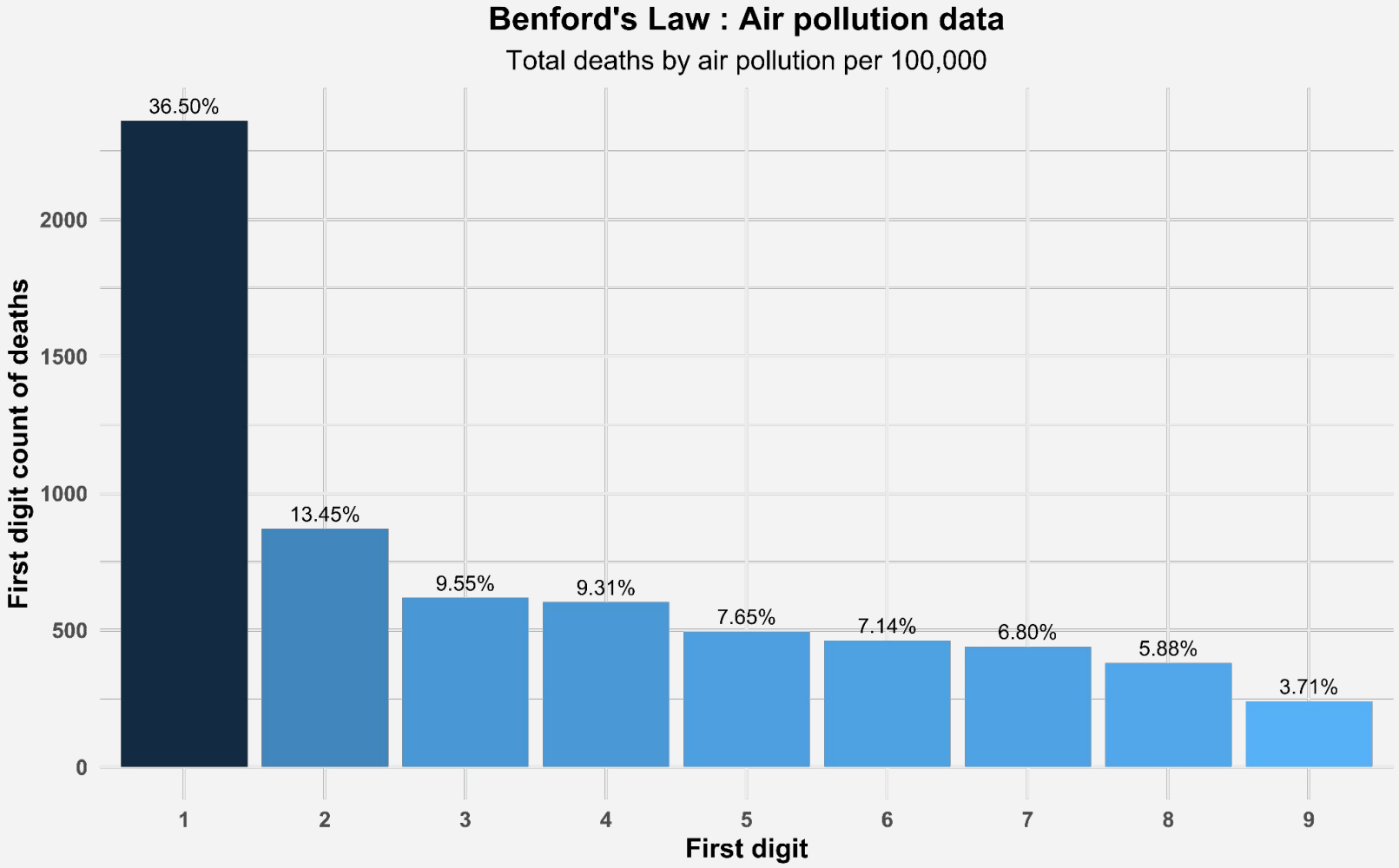
Financial Transactions

The financial transactions dataset was downloaded from Kaggle and describes synthetic data created by a simulator called PaySim that synthesizes data from a private dataset but further injects malicious behaviour to simulate fraudulent transactions. The data can be found here.

Air Pollution Data

No. of observations: 6498

Country	Year	Total deaths per 100,000	Indoor air pollution deaths per 100,000	Outdoor particulate matter death per 100,000	Outdoor ozone pollution deaths per 100,000
Japan	1990	24.97	0.37	22.66	2.34
Japan	1991	25.57	0.33	21.66	2.32


Benford's Law : Air pollution data
Total deaths by air pollution per 100,000
First digit count of deaths
0
500
1000
1500
2000
36.50%
13.45%
9.55%
9.31%
7.65%
7.14%
6.80%
5.88%
3.71%
1
2
3
4
5
6
7
8
9
First digit

Air Pollution Data

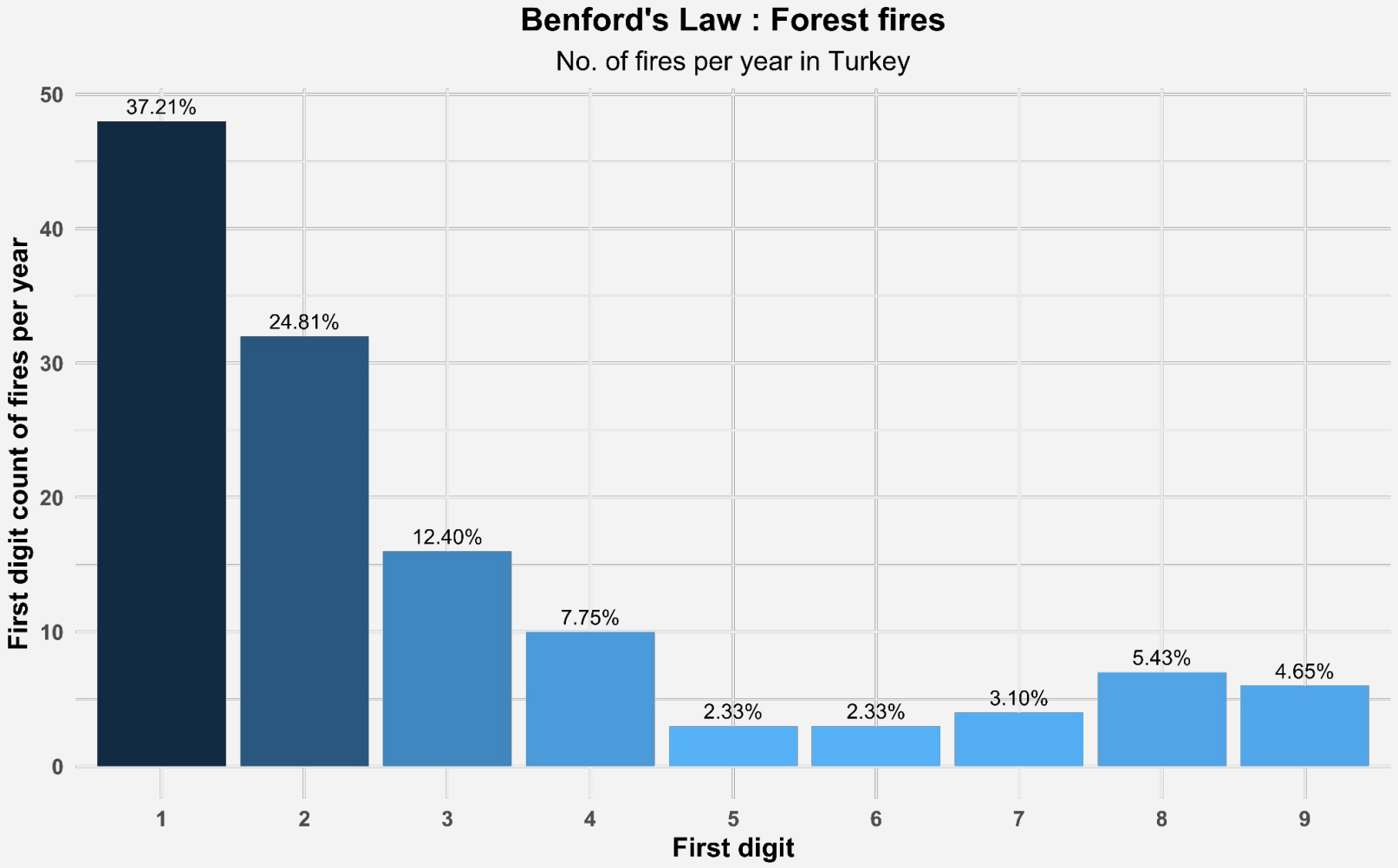
Digit	Actual	Expected
1	0.36	0.30
2	0.14	0.18
3	0.09	0.13
4	0.09	0.09
5	0.07	0.07
6	0.07	0.06
7	0.06	0.05
8	0.05	0.05
9	0.03	0.04

Correlation : 0.96

Forest Fires Data

No. of observations: 129

Country	Year	Hectares	Number of fires	Status (Total/Negligence/ Unknown)
Turkey	1988	185	483	Total
Turkey	1989	174	385	Negligence


Benford's Law : Forest fires
No. of fires per year in Turkey
First digit count of fires per year
50
40
30
20
10
0
37.21%
24.81%
12.40%
7.75%
2.33%
2.33%
3.10%
5.43%
4.65%
1
2
3
4
5
6
7
8
9
First digit

Forest Fires Data

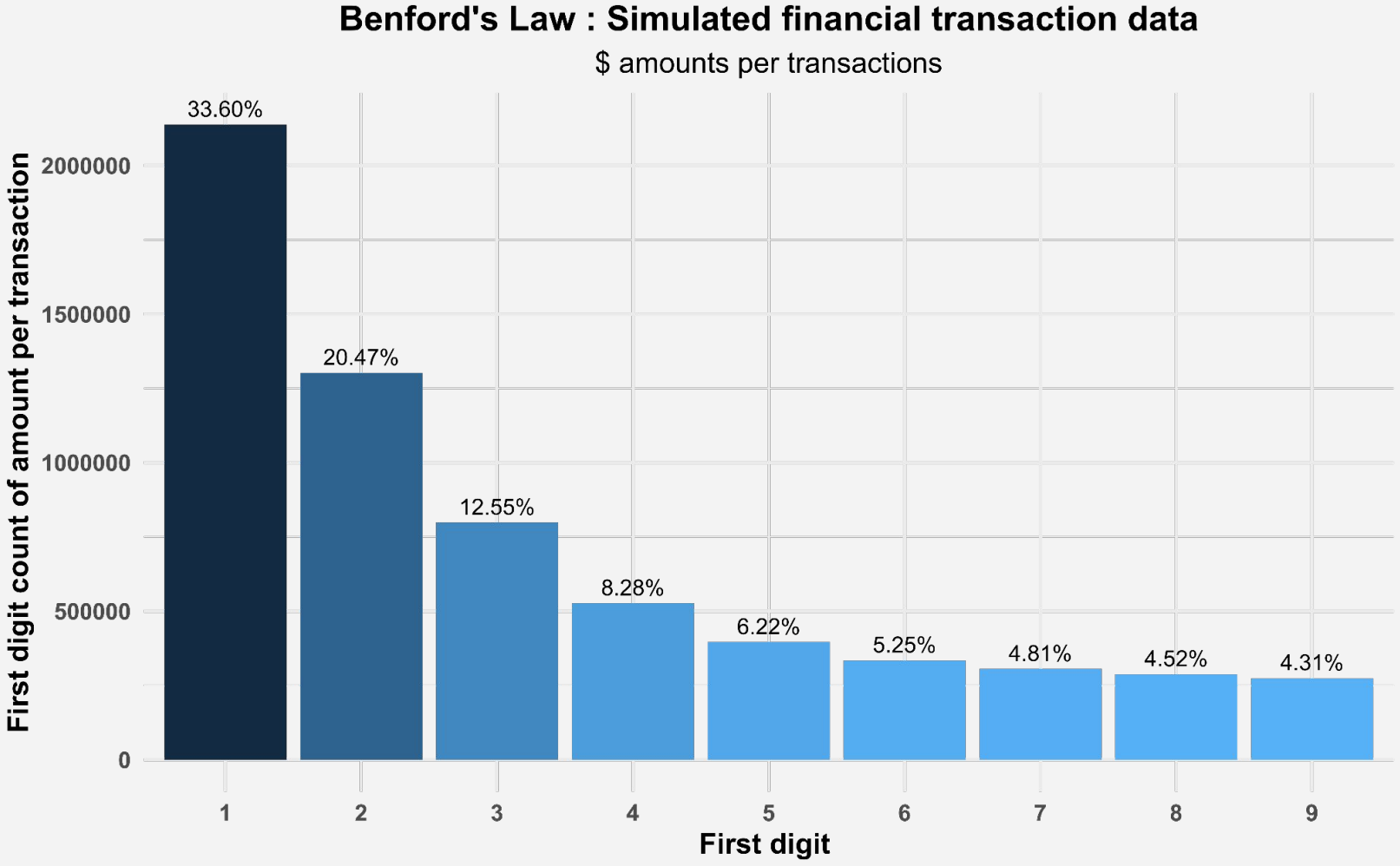
Digit	Actual	Expected
1	0.37	0.30
2	0.24	0.18
3	0.12	0.13
4	0.07	0.09
5	0.02	0.07
6	0.02	0.06
7	0.03	0.05
8	0.04	0.05
9	0.04	0.04

Correlation : 0.70

Simulated Financial Data

No. of observations: 6,352,488

Type (Payment, Debit, Cash Out)	Amount ($)	Origin Name	Origin old balance	Origin new balance	Is Fraud (0/1)
Debit	180	AHWIEN	500	320	0
Debit	110	AHWIDF	410	300	0


Benford's Law : Simulated financial transaction data
$ amounts per transactions
First digit count of amount per transaction
2000000
1500000
1000000
500000
0
33.60%
20.47%
12.55%
8.28%
6.22%
5.25%
4.81%
4.52%
4.31%
1
2
3
4
5
6
7
8
9
First digit

Simulated Financial data

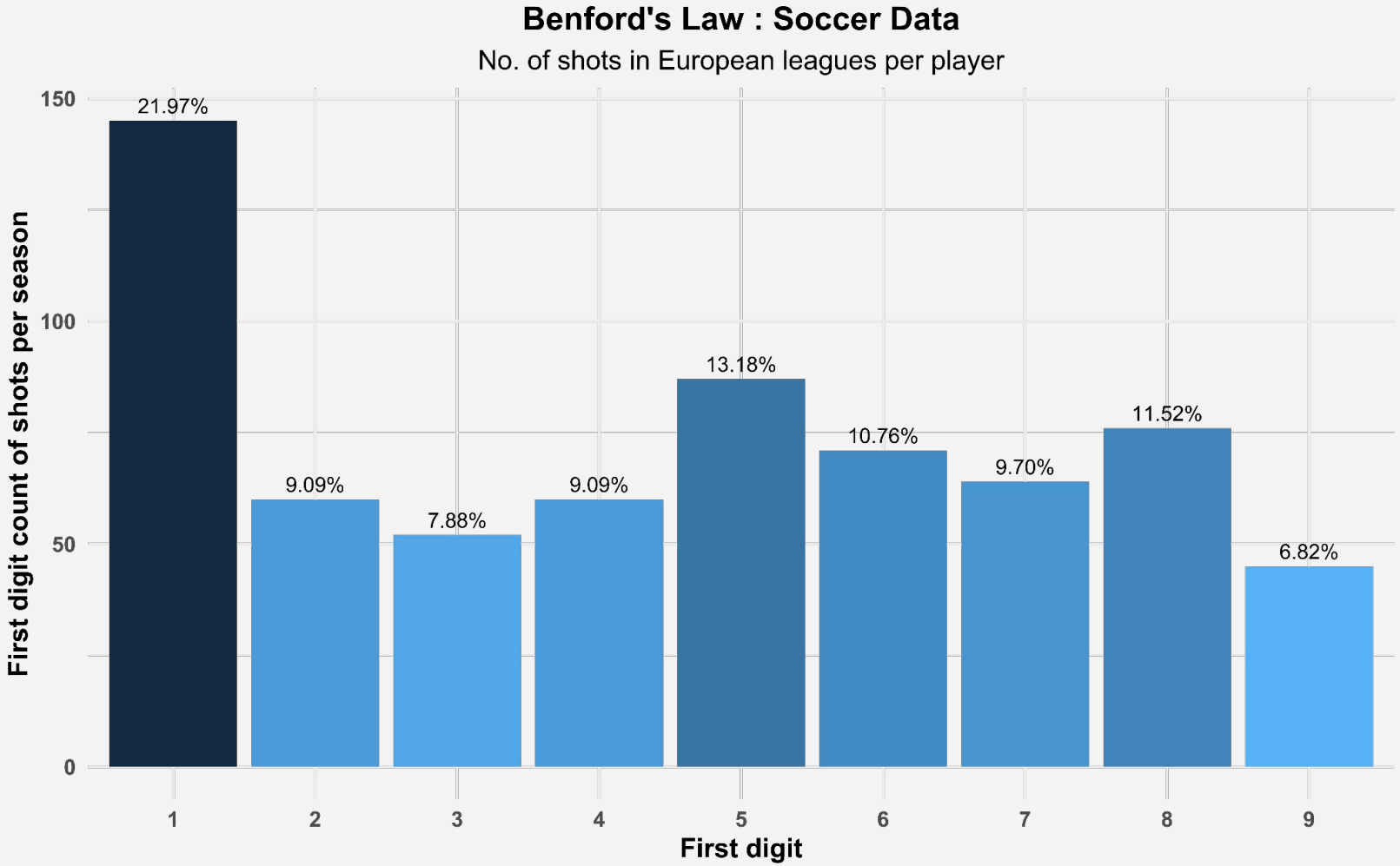
Digit	Actual	Expected
1	0.33	0.30
2	0.20	0.18
3	0.13	0.13
4	0.08	0.09
5	0.06	0.07
6	0.05	0.06
7	0.04	0.05
8	0.04	0.05
9	0.04	0.04

Correlation : 0.995

Soccer Data

No. of observations: 660

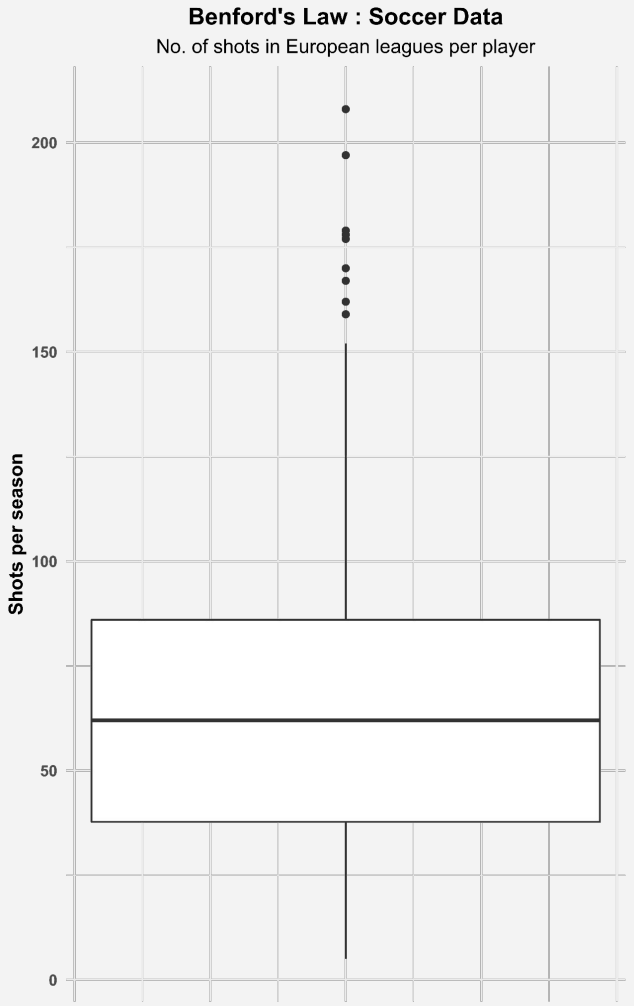
Country	League	Player Name	Matches Played	Total shots	Year
Italy	Serie A	Christiano Ronaldo	30	180	2018
Italy	Serie A	Christiano Ronaldo	38	152	2019


Benford's Law : Soccer Data
No. of shots in European leagues per player
First digit count of shots per season
150
100
50
0
21.97%
9.09%
7.88%
9.09%
13.18%
10.76%
9.70%
11.52%
6.82%
1
2
3
4
5
6
7
8
9
First digit

Soccer Data

Digit	Actual	Expected
1	0.22	0.30
2	0.09	0.18
3	0.08	0.13
4	0.09	0.09
5	0.13	0.07
6	0.10	0.06
7	0.10	0.05
8	0.11	0.05
9	0.07	0.04

Correlation : 0.21


Benford's Law : Soccer Data
No. of shots in European leagues per player
Shots per season
200
150
100
50
0

Strengths and Limitations

06

Strengths

Benford's Law works well with **large data** that **covers orders of magnitudes** in values. They should be **naturally occurring**, in other words they just exist and do not follow any synthetic processes.

The law can help to **uncover fictitious numbers** among random data because it detects manual intervention. Thus it is often used to detect electoral fraud, or financial fraud such as tax evasion or cooking the books.

Limitations

Benford's Law **will not work on small samples**. If you examine a bank account with 20 checks for the year, your sample isn't likely to follow the predicted patterns. You need larger samples to make the test work.

not all data can be tested; assigned numbers, such as telephone numbers, Social Security numbers, and account numbers generally can't be expected to follow Benford's Law.

Data to which Benford's Law does not apply...

- data that is not varied enough in range of values
  - atmospheric pressure
  - phone numbers
  - recorded times for the Olympic 1,500 m race
  - height of humans

Final Remarks

Citations

1. d'Aquin M. What does/doesn't follow Benford's law. Medium. Published February 20, 2022. Accessed May 28, 2022. https://towardsdatascience.com/what-does-doesnt-follow-benford-s-law-7d0b3c14afa5
2. Miller SJ. A Quick Introduction to Benford's Law. :16.
3. andrehk19. [OC] Votes numbers for Trump, Biden, and West follow Benford's Law. Benford's Law, or the first digit law, is consistently recognized as a valid method to assess data manipulation in accounting and financial fields. r/dataisbeautiful. Published November 5, 2020. Accessed May 28, 2022. www.reddit.com/r/dataisbeautiful/comments/jogujo/oc_votes_numbers_for_trump_biden_and_west_follow/
4. Vičič J, Tošić A. Application of Benford's Law on Cryptocurrencies. *J Theor Appl Electron Commer Res*. 2022;17(1):313-326. doi:10.3390/jtaer17010016
5. Feature selection using Benford's law to support detection of malicious social media bots - ScienceDirect. Accessed May 28, 2022. https://www.sciencedirect.com/science/article/pii/S0020025521009695
6. Hsu L, Berisha V. *Does Human Speech Follow Benford's Law?* arXiv; 2022. doi:10.48550/arXiv.2203.13352
7. Marchand C, Maahs D. Benford's Law and COVID-19 Data. CHANCE. Accessed May 28, 2022. https://chance.amstat.org/2021/04/benfords-law/
8. Can Benford's Law Detect Tax Fraud? Accessed May 28, 2022. https://www.forbes.com/sites/taxnotes/2021/08/19/can-benfords-law-detect-tax-fraud/?sh=5a531b854d70
9. Benford's Law | Brilliant Math & Science Wiki. Accessed May 31, 2022. https://brilliant.org/wiki/benfords-law/
10. RPubs - Benford's Law Graphed in R. Accessed May 31, 2022. https://rpubs.com/chiefmurph/205239

Thanks