Financial Data

The following presumably are "educational expenses" that I reported to the math department for reimbursement.

39.20	33.05	99.71	19.70	655.19	16.37	2.28	4.95
14.00	32.40	4.86	1.03	21.72	1.72	5.76	420.37
8.41	21.26	873.80	460.86	24.81	41.58	2.55	14.10
2.46	2.46	3.76	252.05	2.33	25.66	1.22	337.41
1.94	145.52	15.55	3.50	126.50	182.88	6.45	6.98

Total = \$3,935.48

98.48	7.96	64.3	17.36	9.54	2.43	426.93	925.51
4.96	98.55	8.25	4.34	634.87	937.75	3.22	79.75
48.79	317.80	3.69	494.86	8.73	88.31	23.47	6.44
31.68	1.04	39.00	2.48	44.989	96.62	36.96	99.57
5.73	34.29	52.66	8.40	925.30	1.23	518.56	24.17

Total = \$6,238.56

Financial Data

Q: Which of the two data sets is real, and which is fake?

39.20	33.05	99.71	19.70	655.19	16.37	2.28	4.95
14.00	32.40	4.86	1.03	21.72	1.72	5.76	420.37
8.41	21.26	873.80	460.86	24.81	41.58	2.55	14.10
2.46	2.46	3.76	252.05	2.33	25.66	1.22	337.41
1.94	145.52	15.55	3.50	126.50	182.88	6.45	6.98

Total = \$3,935.48

98.48	7.96	64.3	17.36	9.54	2.43	426.93	925.51
4.96	98.55	8.25	4.34	634.87	937.75	3.22	79.75
48.79	317.80	3.69	494.86	8.73	88.31	23.47	6.44
31.68	1.04	39.00	2.48	44.989	96.62	36.96	99.57
5.73	34.29	52.66	8.40	925.30	1.23	518.56	24.17

Total = \$6,238.56

Financial Data

We can try to look at various statistics for each data set:

Statistic	Data Set 1	Data Set 2
Mean	98.39	155.96
Median	15.96	35.62
Mode	N/A	N/A
St. Dev.	192.61	272.24
Range	872.77	936.71
Min	1.03	1.04
Max	873.80	937.75
Total	3,935.48	6,238.56

There is nothing here to imply that something sketchy is going on with either of the data sets!

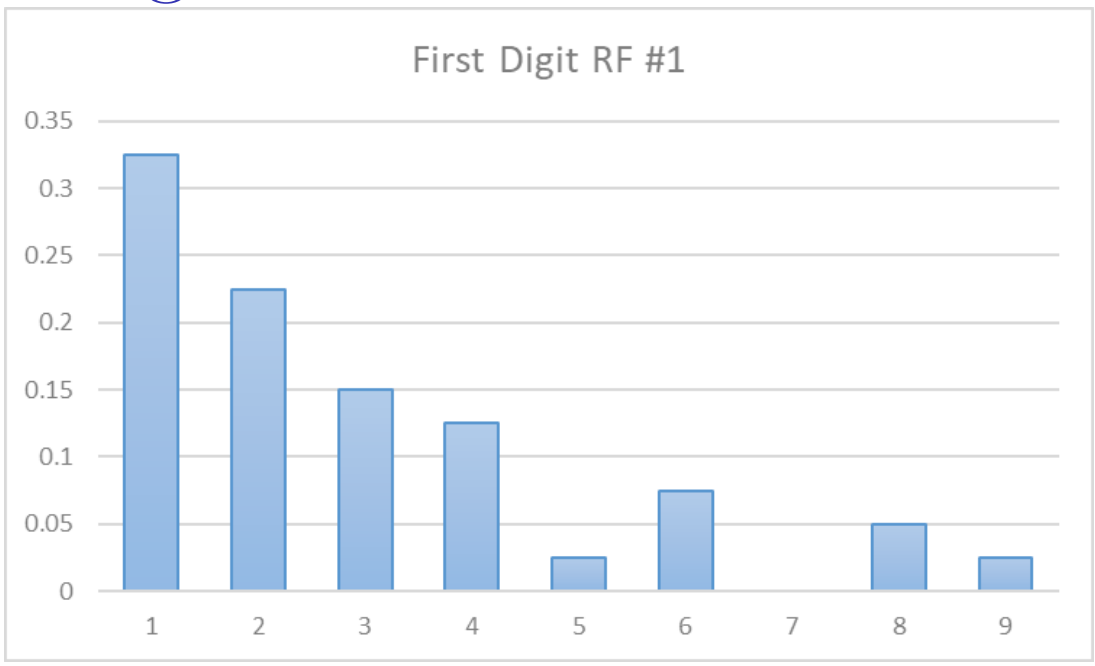
First Digit Statistics

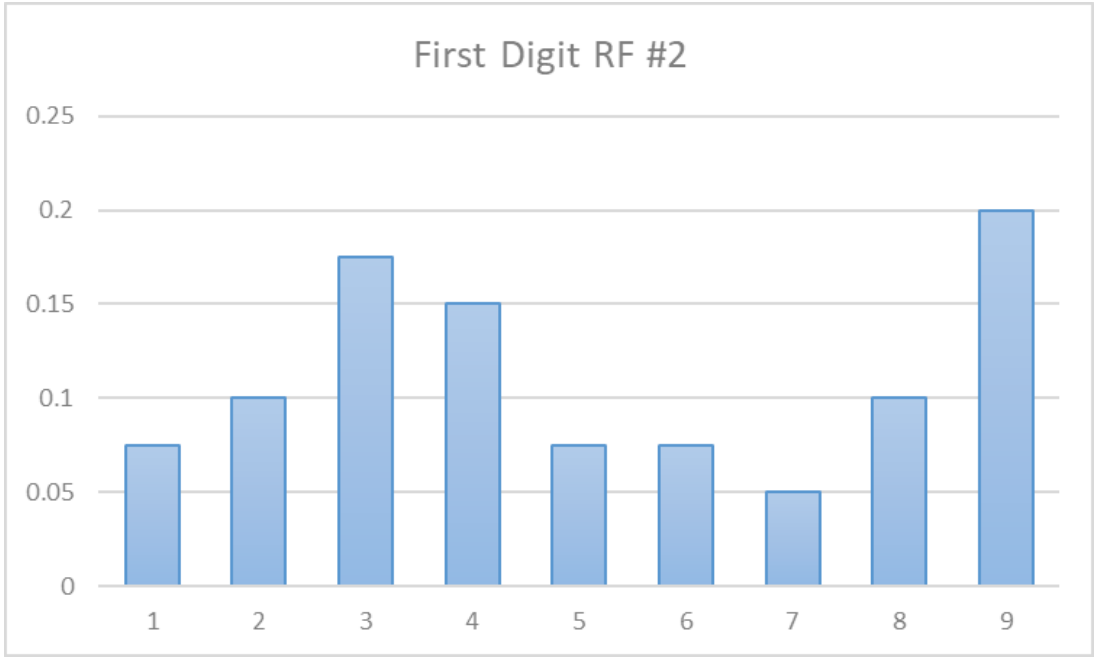
As a last resort, we can try to look at the numbers themselves. We could analyze the digits to see if there is randomness in them, or if that randomness is absent.

The table below shows the **first digit** frequency and relative frequency of each data set.

Digit	Freq #1	RF #1	Freq #2	RF #2
1	13	0.325	3	0.075
2	9	0.225	4	0.10
3	6	0.15	7	0.175
4	5	0.125	6	0.15
5	1	0.025	3	0.075
6	3	0.075	3	0.075
7	0	0	2	0.05
8	2	0.05	4	0.10
9	1	0.025	8	0.20
Total	40	1	40	1

First Digit Statistics





Benford's First Digit Law

It turns out that the second data set is very likely to be FAKE! This is because of **Benford's Law**, which loosely states that the **leading digit** of many naturally occuring numbers is **more likely to be small** (1,2,3) than large (7,8,9).

Numerically, the distribution of the first digit being equal to d (for $1 \leq d \leq 9$) is given by

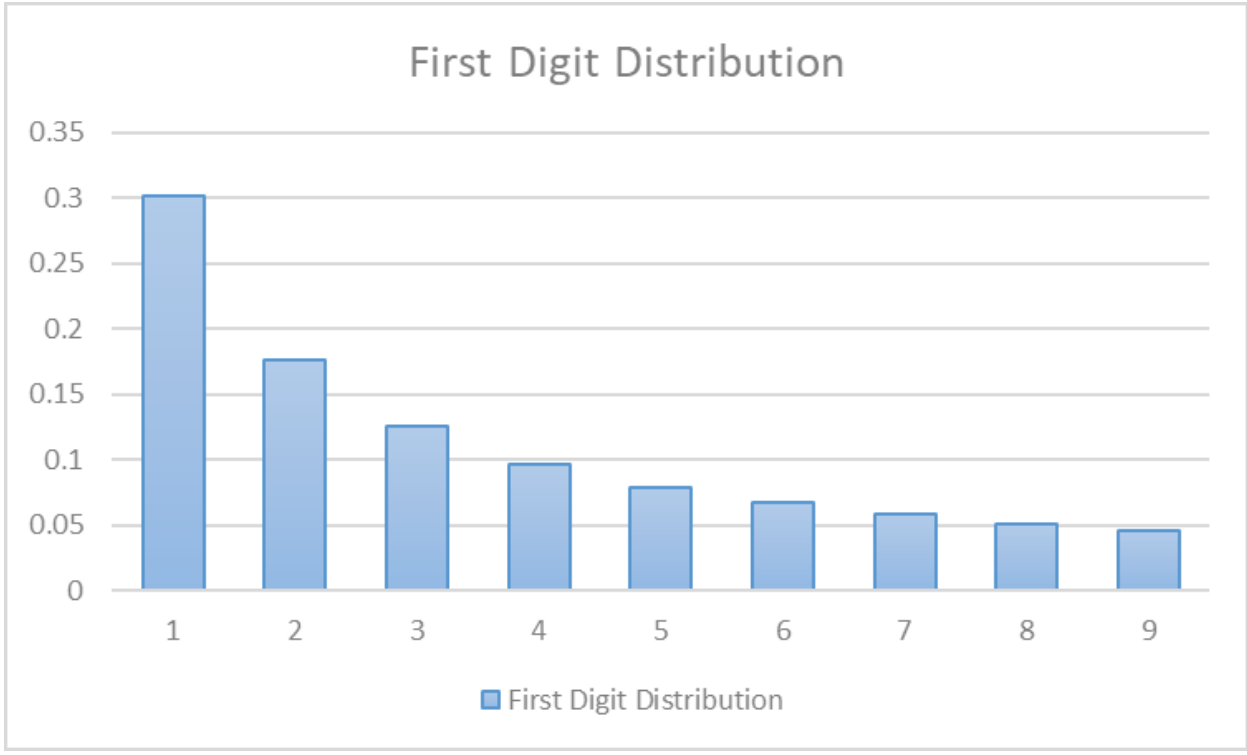
$$\boxed{P(d) = \log_{10}\left(\frac{1+d}{d}\right) = \log_{10}\left(1+\frac{1}{d}\right)}$$

This distribution is used in **fraud detection**.

Benford's First Digit Law

The table below shows the actual probability values predicted by Benford's Law:

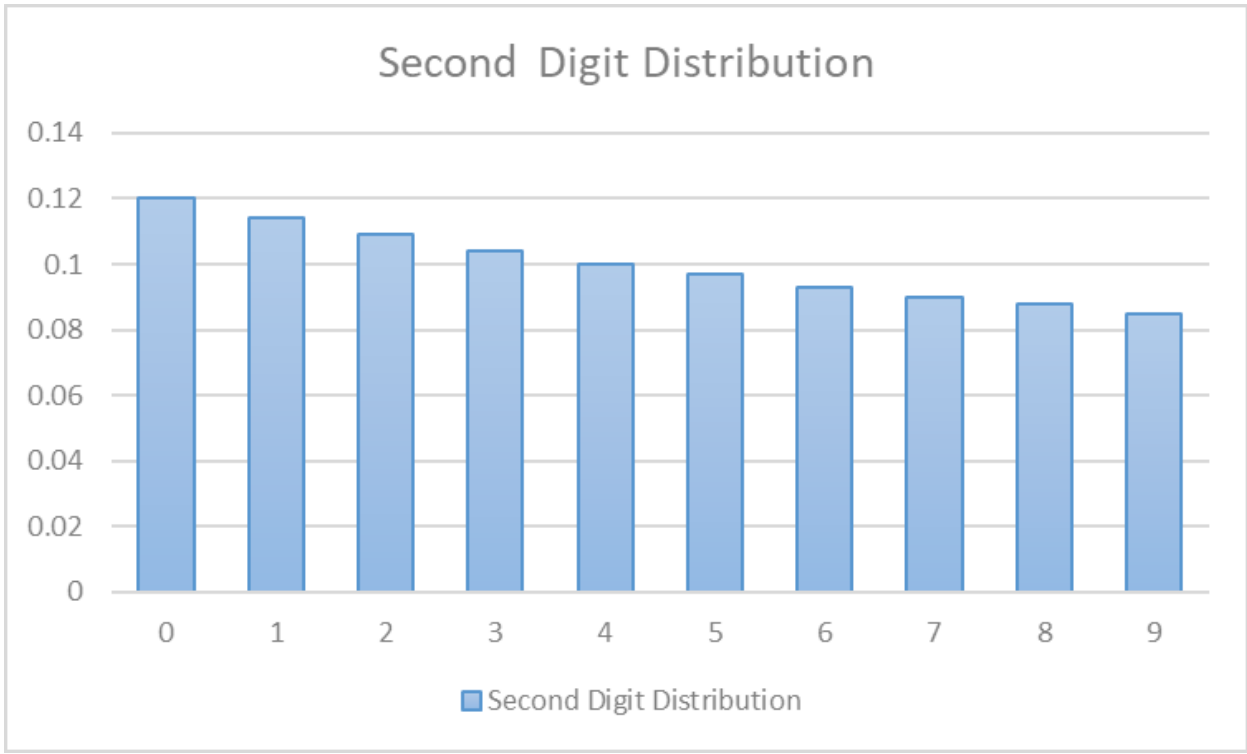
d	P(d)
1	0.301
2	0.176
3	0.125
4	0.097
5	0.079
6	0.067
7	0.058
8	0.051
9	0.046



Second Digit Distribution

It is possible to extend Benford's Law beyond the first digit for a more thorough analysis.

d	P(d)
0	0.12
1	0.114
2	0.109
3	0.104
4	0.1
5	0.097
6	0.093
7	0.09
8	0.088
9	0.085



The First Three Digits

As we analyze further digits, anything from the third digit and on can be considered to be approximately uniform (each digit is equally likely to appear).

d	$P(D_1 = d)$	$P(D_2 = d)$	$P(D_3 = d)$
0	——	0.120	0.102
1	0.301	0.114	0.101
2	0.176	0.109	0.101
3	0.125	0.104	0.101
4	0.097	0.100	0.100
5	0.079	0.097	0.100
6	0.067	0.093	0.099
7	0.058	0.090	0.099
8	0.051	0.088	0.099
9	0.046	0.085	0.098

History

Benford's Law was mentioned/observed by astronomer-mathematician Simon Newcomb in 1881, but was named after Frank Benford, who stated it in 1938.

Newcomb noticed logarithmic table pages and the end of his books were unevenly worn. The first pages wore out faster than the last!

Things That Satisfy Benford's Law

Benford's law appears in many number sets, such as:

- physical constants
- physical measurements
- scientific calculations
- financial/accounting data
- populations

Things That Don't Satisfy Benford's Law

The following data sets are bad fits to Benford's law:

- airline passenger counts
- telephone numbers
- small data sets with 500 or less transactions

Any time data is **restricted by min/max**, it will generally disobey Benford's law (randomly chosen integers 1 and 50).

Sequentially assigned numbers are also bad fits (insurance policy numbers, invoice numbers, account numbers, etc...).

Financial Data

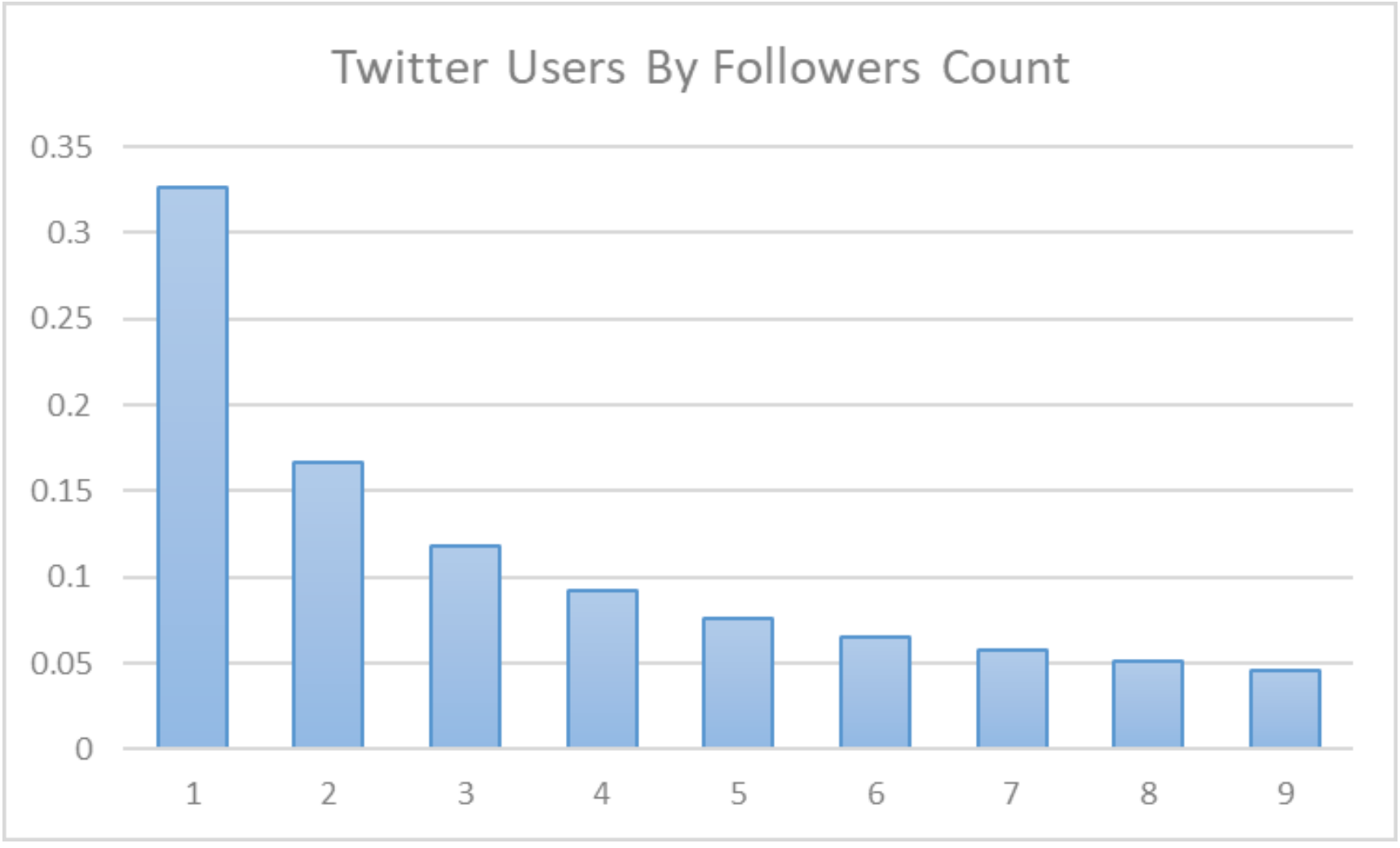
Examples of financial data that follow Benford's law:

- credit card transactions (Homework!)
- loan data
- customer balances
- stock prices
- inventory prices
- customer refunds

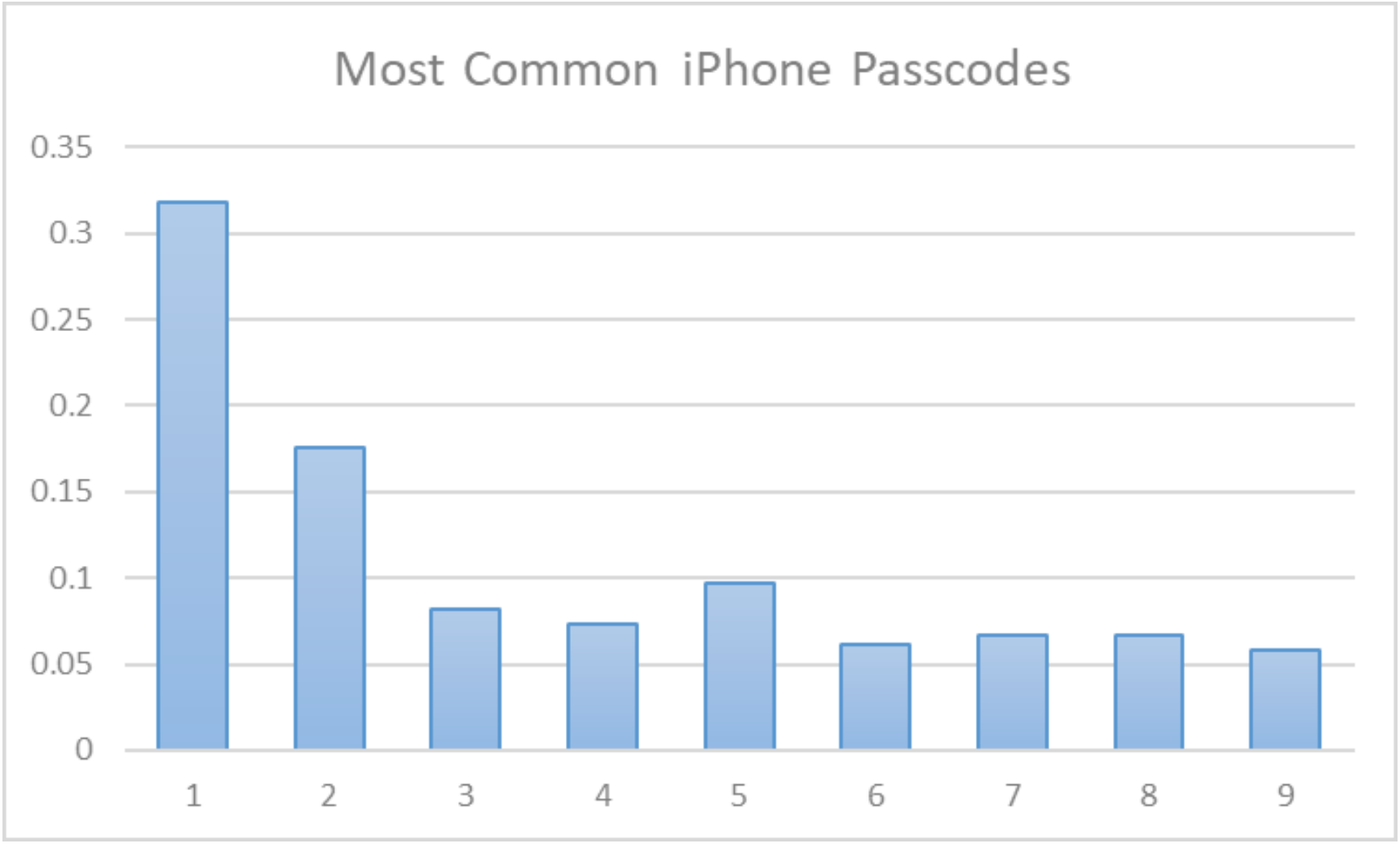
Manipulated/fraudulent data will not follow BL.

BL can also be used to spot duplicate entries in databases (such as double payments made by mistake), so it is very useful to companies!

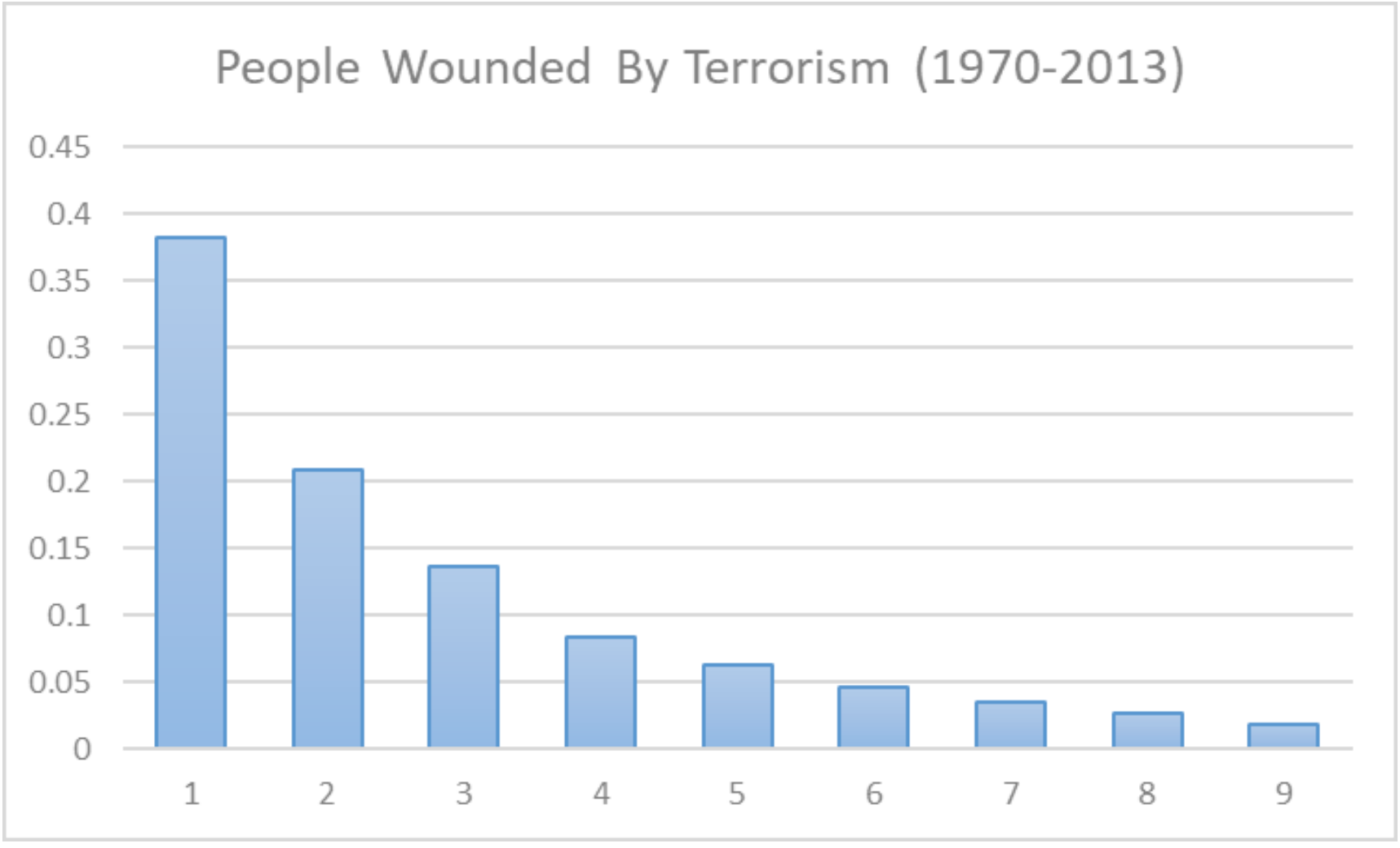
Examples


Twitter Users By Followers Count
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1
2
3
4
5
6
7
8
9

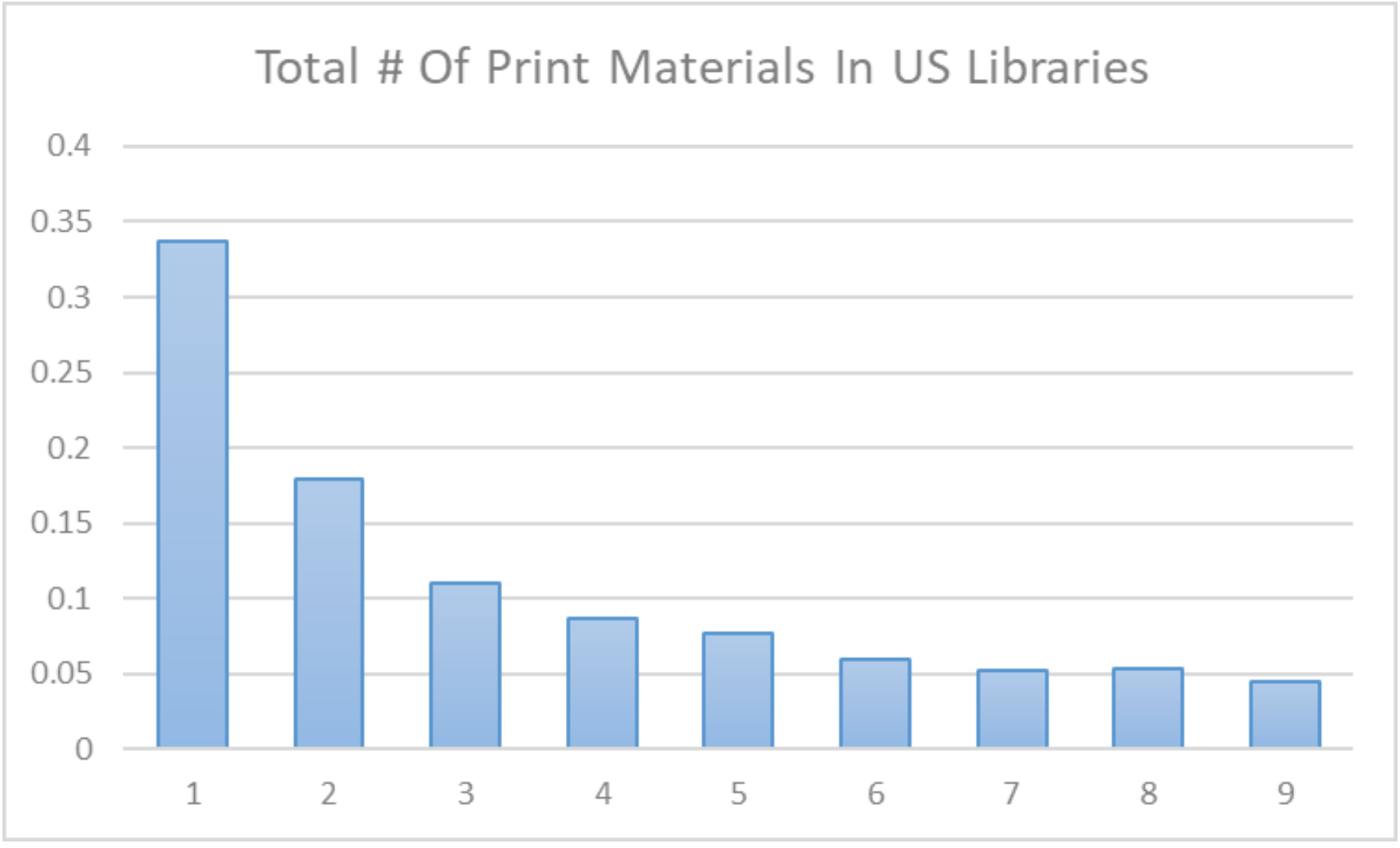
Examples


Most Common iPhone Passcodes
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1
2
3
4
5
6
7
8
9

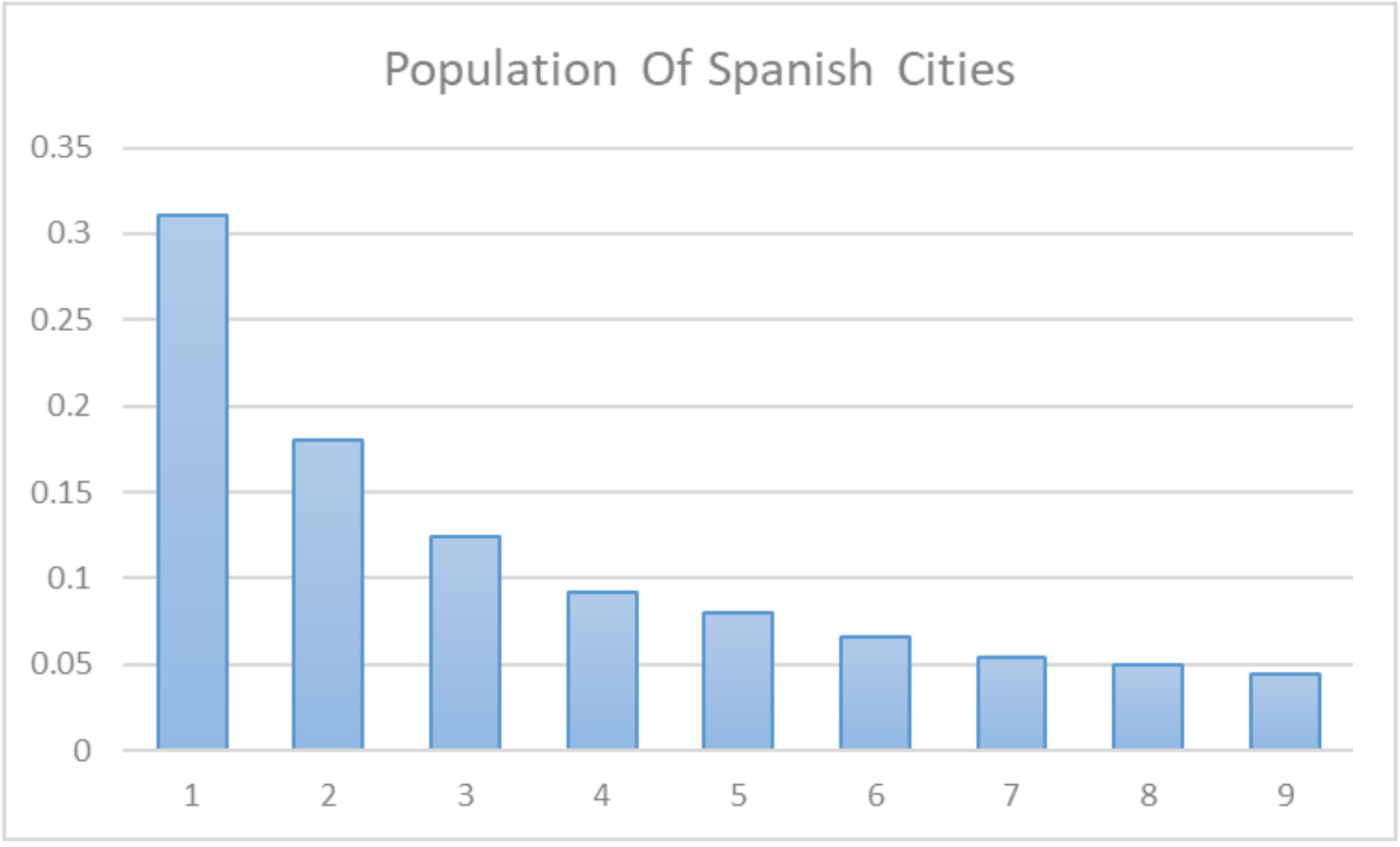
Examples


People Wounded By Terrorism (1970-2013)
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1
2
3
4
5
6
7
8
9

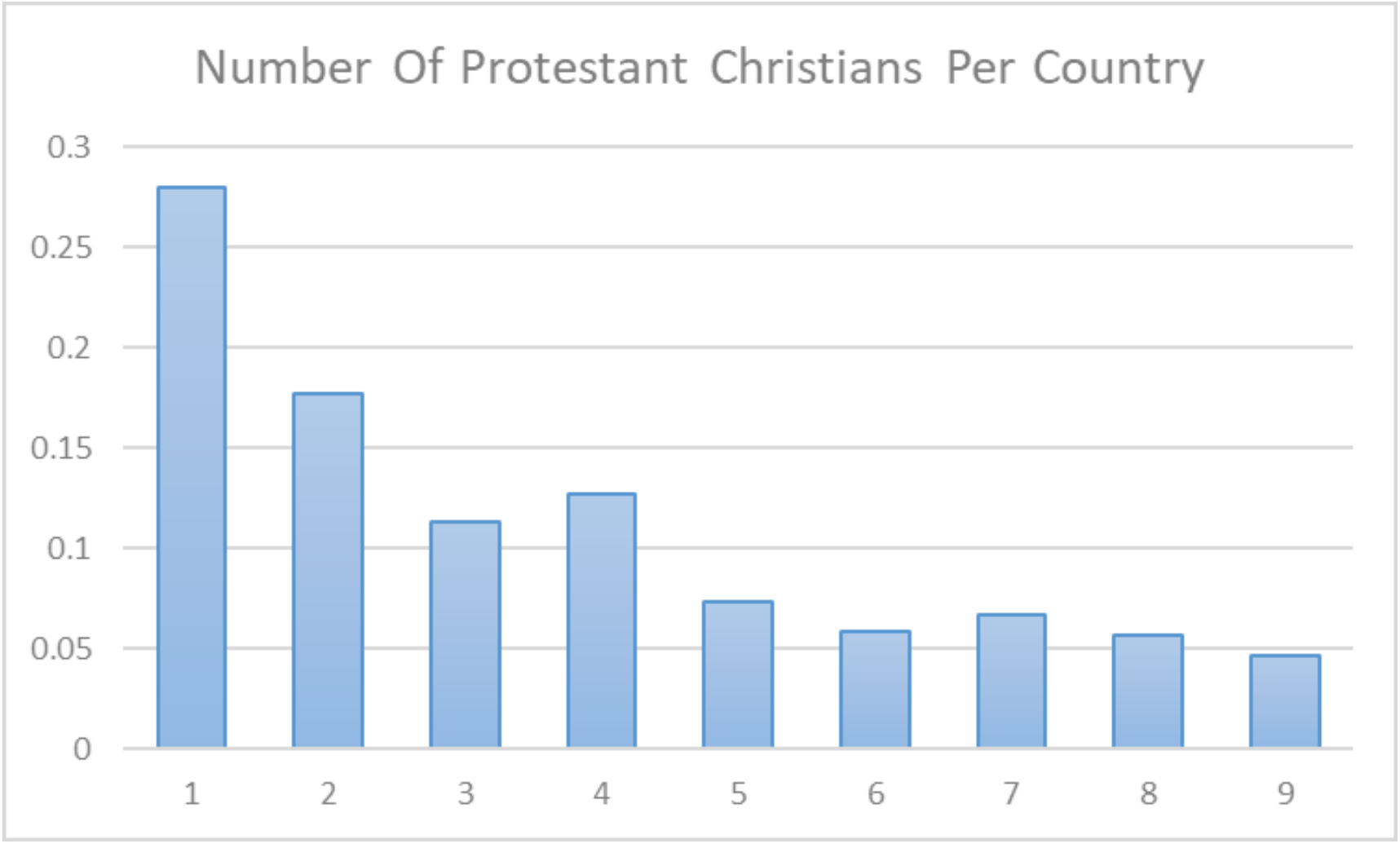
Examples


Total # Of Print Materials In US Libraries
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1
2
3
4
5
6
7
8
9

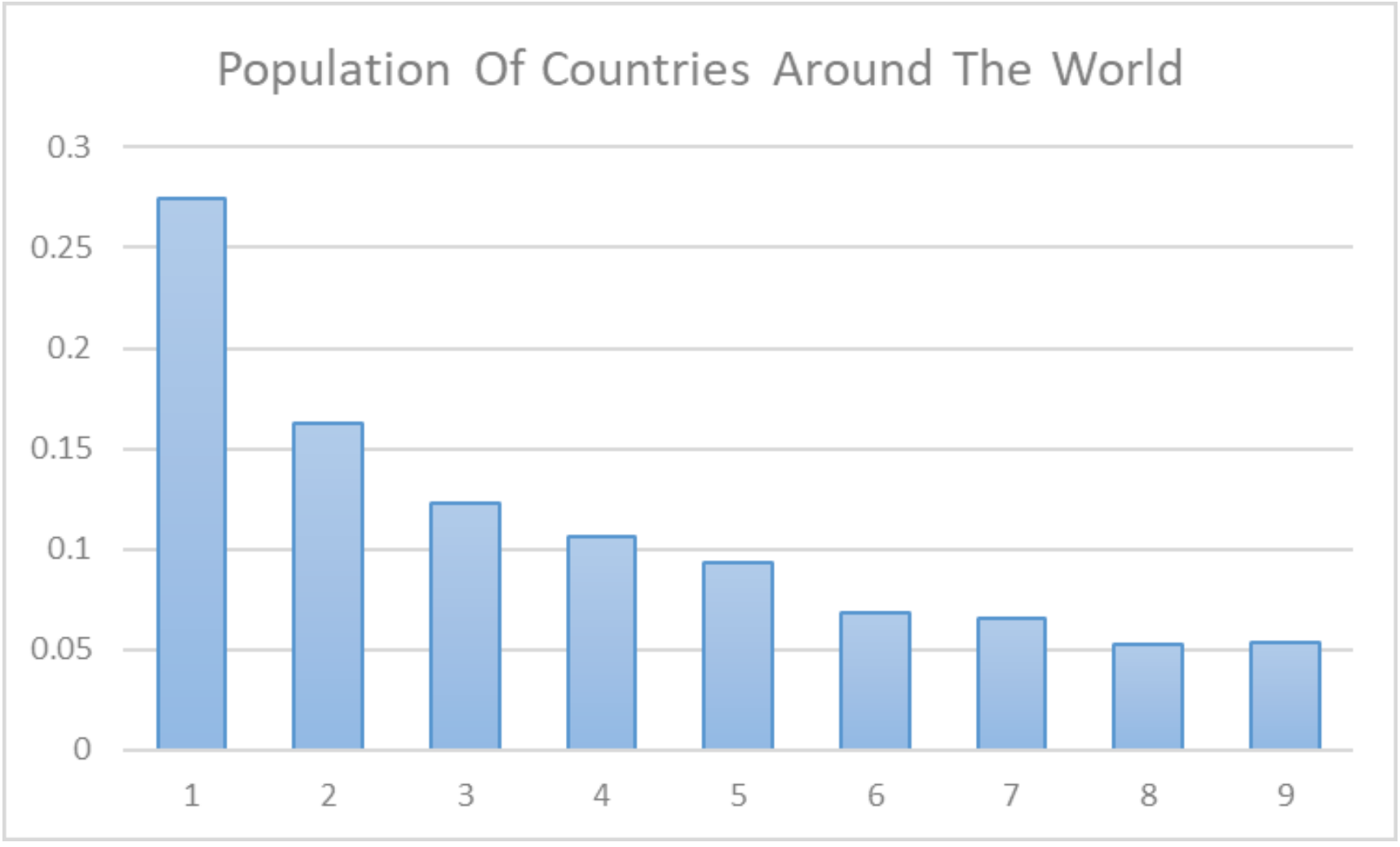
Examples


Population Of Spanish Cities
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1
2
3
4
5
6
7
8
9

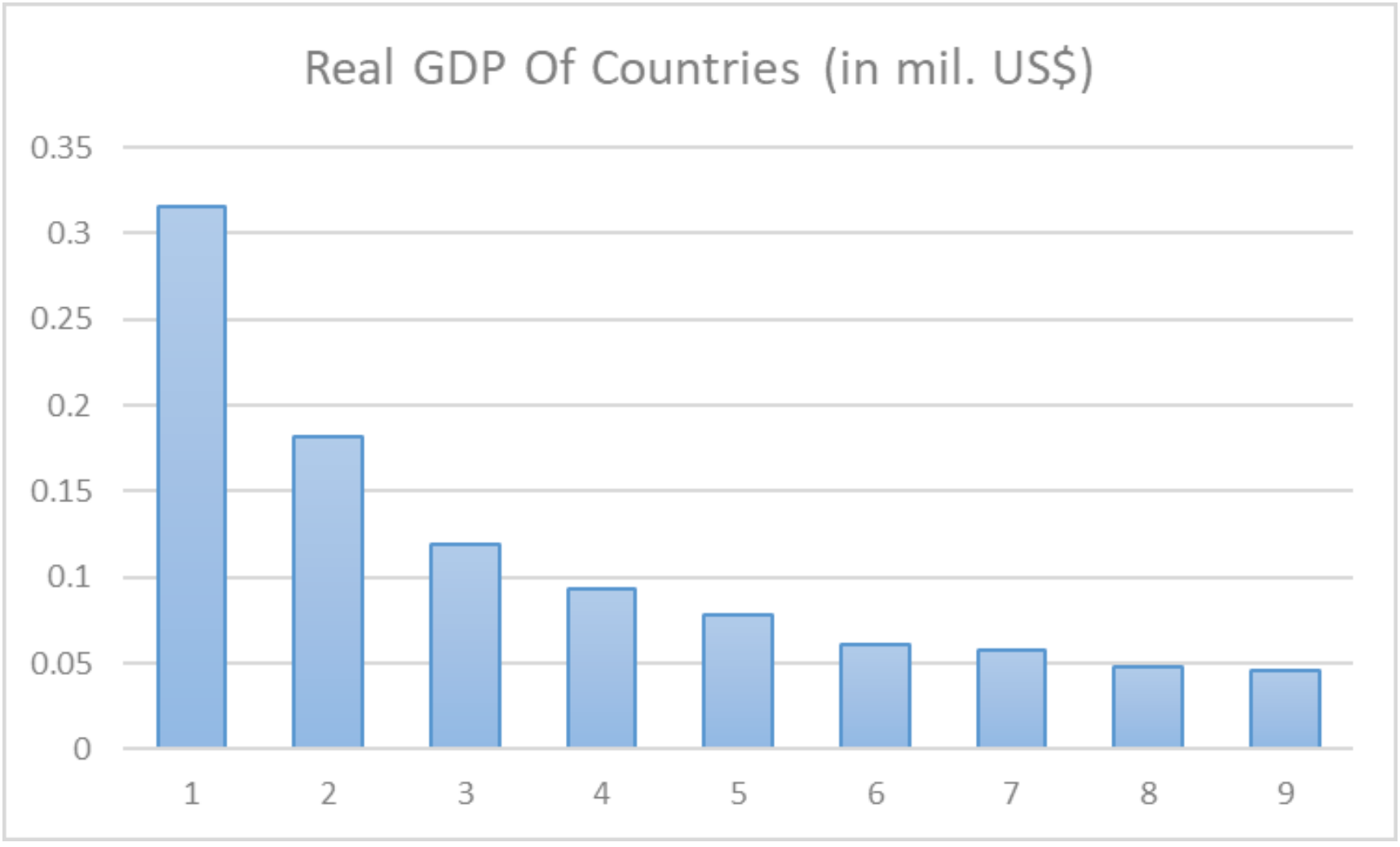
Examples


Number Of Protestant Christians Per Country
0.3
0.25
0.2
0.15
0.1
0.05
0
1
2
3
4
5
6
7
8
9

Examples


Population Of Countries Around The World
0.3
0.25
0.2
0.15
0.1
0.05
0
1
2
3
4
5
6
7
8
9

Examples



More examples at *www.testingbenfordslaw.com.*

Number Sequences

Certain number sequences also obey Benford's law, as shown in the table below.

d	F_n	2^n	3^n	4^n	5^n	32^n
1	301	301	300	304	302	305
2	177	176	177	177	176	170
3	125	125	123	121	125	130
4	96	97	98	100	96	100
5	80	79	79	77	80	75
6	67	69	66	69	66	70
7	56	56	59	58	59	51
8	53	52	52	50	50	49
9	45	45	46	44	46	50
Total	1000	1000	1000	1000	1000	1000