Julian Espinoza Martinez

Benford's Law

What is Benford's law?

Benford's law is the finding that the first digits of the numbers found in series of records of the most varied sources do not display a uniform distribution. They are arranged in such a way that the digit "1" has the highest frequency, followed by "2", then "3". And so in a successively decreasing manner down to "9".

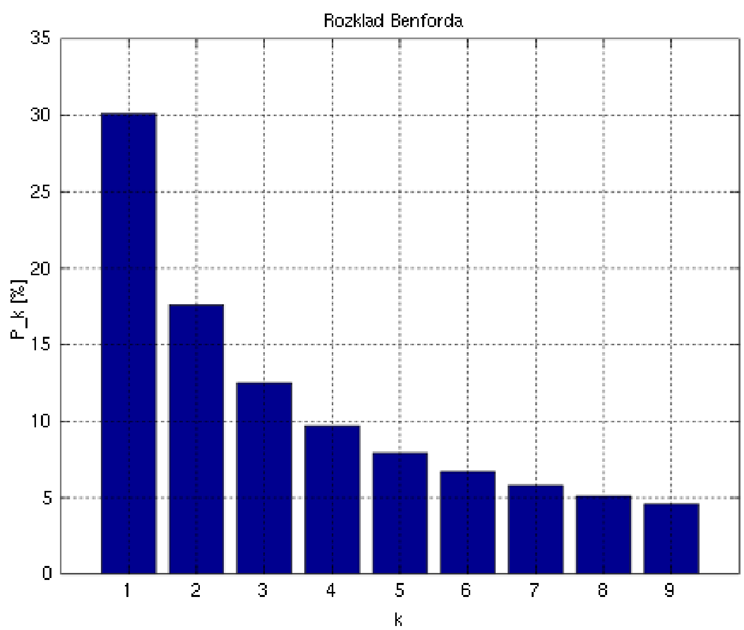

Image Credit: Wikipedia, https://en.wikipedia.org/wiki/Benford%27s_law#/media/File:Rozklad_benforda.svg

In general, it has been seen a series of numerical records follow Benford's Law when they:

- Represent magnitudes of events.
- Do not have pre-established minimum or maximum limits.
- Are not made up of numbers used as identifiers.
- Have a mean which is less than the median, and the data is not concentrated around the mean.

Credit card transaction data application:

Benford's law can be used to find potential fraud in credit card transactions. It could be the case that a cardholder or a merchant is fabricating fraudulent transactions. Someone making up transactions usually doesn't know about Benford's law, so the transaction amounts are random numbers distributed evenly. We can look at the amount distributions for each cardholder and merchant to see if the amount distributions substantially violate Benford's law.

Application on credit card transaction dataset:

For this dataset, we want to quantify how different is the first digit distribution from Benford's law distribution. The very first thing we want to do is remove all the transactions from FedEx since we see that they violate Benford's law and they are not unusual). Then, we will split this analysis into two, one where we will group the dataset by card number (Cardnum) and the other where we will group the dataset by merchant number (Merchnum).

The natural way to do conduct this analysis would be to bin the transaction amounts into 9 bins, one for each possible first digit, but this may be too many bins for this dataset, given that we may not have enough records for some entities (merchant or account). To accommodate for this we will just divide the grouped dataset into two bins, a low bin, and a high bin. We should have about 52.37 / 47.7 = 1.096 ratio of digits (3 through 9 for the high bin) to (1 and 2 for the low bin).

After grouping the dataset by merchant number or card number we will divide the grouped dataset into two bins to quantify how different is the first digit distribution from Benford's law distribution. For each group i.e. for each card number or merchant number depending on which analysis you are working on we will count the number of transactions there are and will store it in a new column called n. Second, for each group, we will count the number of first digits beginning with either 1 or 2 and store this count in a new column called n_low. Then we will create a new column called n_high that is computed by subtracting n_low from n. If either n_low or n_high is zero we will set it to 1 to avoid dividing by zero. The next step is to calculate a new column called R using the ratio of digits calculated previously, n_low, and n_high.

$$R = \frac{1.096 \times n_{\text{low}}}{n_{\text{high}}}$$

After we calculate R we need need to calculate the reciprocal of R and store it as a new field to use it to find the measure of unusualness U.

$$U = max(R,\ 1/R)$$

To be careful about statistics we will use a better measure of unusualness called smoothed U*

$$U^* = 1 + \left(\frac{U-1}{1+\exp^{-t}}\right)$$

To do this we will need to create another field called t :

$$t = (n - n_{mid}) / c$$

With smoothing parameters $c = 3$and $n_{mid = 15}$as suggested by Professor Coggeshall.

After doing all of this we can sort the card numbers and merchant numbers by their U* score in a descending manner to analyze for potential fraud.

Top 40 Cardnum (potential fraud based on Benford's law):

Cardnum	n_low	n_high	R	1/R	U	n	t	U*
5142253356	61	5	13.37	0.07	13.37	66	17	13.37
5142299705	25	3	9.13	0.11	9.13	28	4.33	9.03
5142197563	15	134	0.12	8.15	8.15	149	44.67	8.15
5142194617	5	33	0.17	6.02	6.02	38	7.67	6.02
5142288241	1	13	0.08	11.86	11.86	14	-0.33	5.53
5142239140	16	3	5.85	0.17	5.85	19	1.33	4.83
5142144931	6	30	0.22	4.56	4.56	36	7	4.56
5142192606	13	2	7.12	0.14	7.12	15	0	4.06
5142204384	199	54	4.04	0.25	4.04	253	79.33	4.04
5142284940	21	6	3.84	0.26	3.84	27	4	3.78
5142189113	6	24	0.27	3.65	3.65	30	5	3.63
5142225308	4	17	0.26	3.88	3.88	21	2	3.53
5142116864	58	18	3.53	0.28	3.53	76	20.33	3.53
5142293257	2	13	0.17	5.93	5.93	15	0	3.47
5142173286	2	13	0.17	5.93	5.93	15	0	3.47
5142246929	79	25	3.46	0.29	3.46	104	29.67	3.46
5142224699	7	25	0.31	3.26	3.26	32	5.67	3.25
5142847398	10	35	0.31	3.19	3.19	45	10	3.19
5142273608	6	21	0.31	3.19	3.19	27	4	3.15
5142147267	22	76	0.32	3.15	3.15	98	27.67	3.15
5142224769	15	5	3.29	0.3	3.29	20	1.67	2.92
5142242241	16	51	0.34	2.91	2.91	67	17.33	2.91
5142260984	265	101	2.88	0.35	2.88	366	117	2.88
5142113192	2	12	0.18	5.47	5.47	14	-0.33	2.87
5142191416	18	7	2.82	0.35	2.82	25	3.33	2.76
5142308889	11	2	6.03	0.17	6.03	13	-0.67	2.71
5142194228	11	2	6.03	0.17	6.03	13	-0.67	2.71
5142212038	12	3	4.38	0.23	4.38	15	0	2.69
5142195887	12	3	4.38	0.23	4.38	15	0	2.69
5142225184	27	11	2.69	0.37	2.69	38	7.67	2.69
5142257356	142	58	2.68	0.37	2.68	200	61.67	2.68
5142216493	14	5	3.07	0.33	3.07	19	1.33	2.64
5142239106	8	23	0.38	2.62	2.62	31	5.33	2.62
5142144593	4	14	0.31	3.19	3.19	18	1	2.6
5142126842	38	16	2.6	0.38	2.6	54	13	2.6
5142117315	7	20	0.38	2.61	2.61	27	4	2.58
5142218798	21	9	2.56	0.39	2.56	30	5	2.55
5142180432	58	25	2.54	0.39	2.54	83	22.67	2.54

5142264155	27	12	2.47	0.41	2.47	39	8	2.47
5142294614	5	15	0.37	2.74	2.74	20	1.67	2.46

Top 40 Merchnum (potential fraud based on Benford's law):

Merchnum	n_low	n_high	R	1/R	U	n	t	U*
991808369338	1	181	0.01	165.15	165.15	182	55.67	165.15
8078200641472	59	1	64.66	0.02	64.66	60	15	64.66
308904389335	1	53	0.02	48.36	48.36	54	13	48.36
3523000628102	34	1	37.26	0.03	37.26	35	6.67	37.2
808998385332	1	36	0.03	32.85	32.85	37	7.33	32.83
55158027	27	1	29.59	0.03	29.59	28	4.33	29.22
8916500620062	1	31	0.04	28.28	28.28	32	5.67	28.19
3910694900001	25	1	27.4	0.04	27.4	26	3.67	26.74
881145544	24	1	26.3	0.04	26.3	25	3.33	25.43
8889817332	24	1	26.3	0.04	26.3	25	3.33	25.43
5600900060992	1	27	0.04	24.64	24.64	28	4.33	24.33
6844000608436	23	1	25.21	0.04	25.21	24	3	24.06
92891948003	1	24	0.05	21.9	21.9	25	3.33	21.18
5803301245621	21	1	23.02	0.04	23.02	22	2.33	21.07
3433000017263	53	3	19.36	0.05	19.36	56	13.67	19.36
467615916337	1	22	0.05	20.07	20.07	23	2.67	18.83
817004638227	19	1	20.82	0.05	20.82	20	1.67	17.67
2376700063599	30	2	16.44	0.06	16.44	32	5.67	16.39
993620816222	1	19	0.06	17.34	17.34	20	1.67	14.74
993620810220	5	76	0.07	13.87	13.87	81	22	13.87
465614140337	1	18	0.06	16.42	16.42	19	1.33	13.21
8999000079657	1	18	0.06	16.42	16.42	19	1.33	13.21
8317600900099	24	2	13.15	0.08	13.15	26	3.67	12.85
5000006000095	253	23	12.06	0.08	12.06	276	87	12.06
5186264200136	1	17	0.06	15.51	15.51	18	1	11.61
9420966064460	1	17	0.06	15.51	15.51	18	1	11.61
600000201284	4	50	0.09	11.41	11.41	54	13	11.41
5600000060302	1	16	0.07	14.6	14.6	17	0.67	9.99
7080606900600	1	16	0.07	14.6	14.6	17	0.67	9.99
6070095870009	26	3	9.5	0.11	9.5	29	4.67	9.42
999960264339	3	28	0.12	8.52	8.52	31	5.33	8.48
555400670006	1	15	0.07	13.69	13.69	16	0.33	8.39
881894855	1	15	0.07	13.69	13.69	16	0.33	8.39
1960400470068	23	3	8.4	0.12	8.4	26	3.67	8.22
993620559229	5	43	0.13	7.85	7.85	48	11	7.85
604901367333	1	14	0.08	12.77	12.77	15	0	6.89
8100544800098	1	14	0.08	12.77	12.77	15	0	6.89
2586000448258	1	14	0.08	12.77	12.77	15	0	6.89

6000330043193	13	1	14.25	0.07	14.25	14	-0.33	6.53
2644006060269	13	1	14.25	0.07	14.25	14	-0.33	6.53