Zipf’s law and friends

Hera He

January 8, 2016
Outline

1. Power law
2. Zipf’s law
3. Heap’s law
4. Benford’s law
Normal distribution

Figure: Left: histogram of heights in centimetres of American males. Data from the National Health Examination Survey, 1959-1962 (US Department of Health and Human Services). Right: histogram of speeds in miles per hour of cars on UK motorways. Data from Transport Statistics 2003 (UK Department for Transport).
Power law: first example

**Figure**: Left: histogram of the populations of all US cities with population of 10,000 or more. Right: another histogram of the same data, but plotted on logarithmic scales. The approximate straight-line form of the histogram in the right panel implies that the distribution follows a power law. Data from the 2000 US Census.
Power law: Definition

\[
\ln p(x) = -\alpha \ln x + c
\]
\[\iff p(x) = Cx^{-\alpha}\]

- Distributions of the above form are said to follow a power law.
- The constant \(\alpha\) is called the **exponent** of the power law.
- Other names: Patero distribution (continuous), Zipf’s law (discrete).
Measuring power law

How to detect or identify power-law behaviour?

First thought: make a simple histogram and plot it on log scale.

*Poor way to proceed!*
Measuring power law

We simulate $10^6$ random numbers from a power-law probability with exponent $\alpha = -2.5$. (plot a) Simple histogram plotted on log-scale produces noisy fluctuation in the tail. (plot b)

Figure: (a) Histogram of the set of 1 million random numbers described in the text, which have a power-law distribution with exponent $\alpha = 2.5$. (b) The same histogram on logarithmic scales. Notice how noisy the results get in the tail towards the right-hand side of the panel. This happens because the number of samples in the bins becomes small and statistical fluctuations are therefore large as a fraction of sample number.
Measuring power law

A better plot with logarithmic binning: each bin is a fixed multiplier wider than the one before. e.g. use bins of size 0.1, 0.2, 0.4, 0.8 etc. (break points (1, 1.1), (1.1, 1.3), (1.3, 1.7), (1.7, 2.5) etc.)
Measuring power law

A better plot with logarithmic binning: each bin is a fixed multiplier wider than the one before. e.g. use bins of size 0.1, 0.2, 0.4, 0.8 etc. (break points (1, 1.1), (1.1, 1.3), (1.3, 1.7), (1.7, 2.5) etc.)

Figure: (c) A histogram constructed using logarithmic binning.
Measuring power law

A superior plotting method: Plot ccdf(tail probability) \( P(X > x) \) against data values.

Or equivalently, use rank/frequency plot.

\[
P(X > x) = \int_{x}^{\infty} p(t)dt = \frac{C}{\alpha - 1}x^{-(\alpha-1)}
\]
Measuring power law

A superior plotting method: Plot ccdf(tail probability) $P(X > x)$ against data values.
Or equivalently, use rank/frequency plot.

$$P(X > x) = \int_x^\infty p(t)dt = \frac{C}{\alpha - 1}x^{-(\alpha-1)}$$

**Figure**: (d) A cumulative histogram or rank/frequency plot of the same data. The cumulative distribution also follows a power law, but with an exponent of $\alpha - 1 = 1.5$. 
Estimation of the exponent

How to estimate the exponent $\alpha$ from samples?

- least squares fit is not reliable. e.g. least squares fit to plot (b) gives $\hat{\alpha}_{LS} = 2.26 \pm 0.02$.
- formula obtained as MLE:

$$\alpha = 1 + n \left[ \sum_{i=1}^{n} \ln \frac{x_i}{x_{\text{min}}} \right]$$

Error can be derived by bootstrap/jackknife.

MLE estimate of simulated data $\hat{\alpha}_{MLE} = 2.500 \pm 0.002$
Examples of power laws
Real-world distributions typically follow power law only after some minimum value $x_{\text{min}}$. 

One often hears a quantity “has a power law tail”. 

A judgement is required to determine the value $x_{\text{min}}$. One way is to perform a scan over all values of $x_{\text{min}}$. 

Once $x_{\text{min}}$ is determined, the usual MLE estimate for $\alpha$ can be used. 

R package: poweRlaw
MLE Estimation: proof

\[ p(x) = Cx^{-\alpha} = \frac{\alpha - 1}{x_{\min}} \left( \frac{x}{x_{\min}} \right)^{-\alpha} \]

\[ \mathcal{L} = \ln \prod_{i=1}^{n} p(x_i) \]

\[ = \ln \prod_{i=1}^{n} \frac{\alpha - 1}{x_{\min}} \left( \frac{x_i}{x_{\min}} \right)^{-\alpha} \]

\[ = n \ln(\alpha - 1) - n \ln x_{\min} - \alpha \sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}} \] (1)

Setting \( \frac{\partial \mathcal{L}}{\partial \alpha} = 0 \), we have

\[ \alpha = 1 + n \left[ \sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}} \right] \]
Zipf’s law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table.

Let $N$ be the number of elements, $k$ be their rank, $s$ be the value of the exponent characterizing the distribution. Zipf’s law predicts that out of a population of $N$ elements, the frequency of elements of rank $k$, $f(k; s, N)$, is:

$$f(k; s, N) = \frac{[\text{constant}]}{k^s}$$

Pareto distribution and Zipf’s law differ from each other in the way the C.D.F. is plotted. Unlike Pareto, Zipf’s made the rank on $x$-axis and frequency on $y$-axis.
Zipf’s law: Zipf’s plot

Figure: Zipf CDF for $N = 10$. The horizontal axis is the index $k$. (Note that the function is only defined at integer values of $k$. The connecting lines do not indicate continuity.)
Zipf’s law: Estimation

Let the p.m.f of $X$ with a Zipf’s law be

$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{\text{min}})}, \quad x \geq x_{\text{min}}$$

where $\zeta(\alpha, x_{\text{min}}) = \sum_{n=0}^{\infty} (n + x_{\text{min}})^{-\alpha}$

MLE esitmator numerically maximizes

$$\mathcal{L} = -n \ln \zeta(\alpha, x_{\text{min}}) - \alpha \sum_{i=1}^{n} \ln x_i$$

Details see Reference [3].
A generalization of Zipf’s law is the Zipf-Mandelbrot law, proposed by Benoît Mandelbrot, whose frequencies are:

$$f(k; N, q, s) = \frac{[\text{constant}]}{(k + q)^s}$$
Zipf’s law: Explanation 1, optimization model

- Mandelbrot experiment: design a language over an alphabet of size $d$ to optimize information per character.
  - Probability of $j$th most frequently used word is $p_j$
  - Length of $j$th most frequently used word is $c_j \sim \log_d j$

- Average information per word (entropy): $H = -\sum p_j \log p_j$

- Average characters per word: $C = \sum p_j c_j$

Optimization leads to Zipf’s law.
Zipf’s law: Explanation 2, monkey typing randomly

- Miller (psychologist, 1957) suggests following: monkeys type randomly at a keyboard.
  - Hit each of \( n \) characters with probability \( p \).
  - Hit space with probability \( 1 - np > 0 \).
  - A word is a sequence of characters separated by a space.

Resulting distribution of word frequencies follows a Zipf’s law.

Conclusion: Mandelbrot’s “optimization” not required for languages to have power law.
Consider a dynamic Web graph.
- pages join one at a time.
- Each page has one outlink.

Let $X_j(t)$ be the number of pages with degree $j$ at time $t$.

New page links:
- With probability $\alpha$, link to a random page.
- With probability $1 - \alpha$, link to a page chosen proportionally to indegree.

The resulting node degree follows a power law.
Heap’s law

A typical Heaps-law plot. The x-axis represents the text size, and the y-axis represents the number of distinct vocabulary elements present in the text. Compare the values of the two axes.
Heap’s law

\( V_R: \) number of distinct words in an instant text of size \( n \).

\( K \) and \( \beta \) are free parameters.

\[ V_R(n) = Kn^\beta \]

- can be derived from Zipf’s law (asymptotically equivalent to Zipf’s law under mild assumptions)
- implies diminishing returns in terms of discovery of the full vocabulary from which the distinct terms are drawn.
- is generalizable. E.g. objects - people, type - country
True or fake?

Table 1. One of the columns gives the land area of political states and territories in km$^2$. The other column contains faked data, generated with a random number generator.

<table>
<thead>
<tr>
<th>State/Territory</th>
<th>Real or Faked Area (km$^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>645,807</td>
</tr>
<tr>
<td>Albania</td>
<td>28,748</td>
</tr>
<tr>
<td>Algeria</td>
<td>2,381,741</td>
</tr>
<tr>
<td>American Samoa</td>
<td>197</td>
</tr>
<tr>
<td>Andorra</td>
<td>464</td>
</tr>
<tr>
<td>Anguilla</td>
<td>96</td>
</tr>
<tr>
<td>Antigua and Barbuda</td>
<td>442</td>
</tr>
<tr>
<td>Argentina</td>
<td>2,777,409</td>
</tr>
<tr>
<td>Armenia</td>
<td>29,743</td>
</tr>
<tr>
<td>Aruba</td>
<td>193</td>
</tr>
<tr>
<td>Australia</td>
<td>7,682,557</td>
</tr>
<tr>
<td>Austria</td>
<td>83,858</td>
</tr>
<tr>
<td>Azerbaijan</td>
<td>86,530</td>
</tr>
<tr>
<td>Bahamas</td>
<td>13,962</td>
</tr>
<tr>
<td>Bahrain</td>
<td>694</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>142,615</td>
</tr>
<tr>
<td>Barbados</td>
<td>431</td>
</tr>
<tr>
<td>Belgium</td>
<td>30,518</td>
</tr>
<tr>
<td>Belize</td>
<td>22,965</td>
</tr>
<tr>
<td>Benin</td>
<td>112,620</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Benford’s law (first digit law)

Imagine a large dataset, say something like a list of every country and its population.

<table>
<thead>
<tr>
<th>COUNTRY</th>
<th>POPULATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>29,117,000</td>
</tr>
<tr>
<td>Albania</td>
<td>3,195,000</td>
</tr>
<tr>
<td>Algeria</td>
<td>35,423,000</td>
</tr>
<tr>
<td>Andorra</td>
<td>84,082</td>
</tr>
<tr>
<td>Angola</td>
<td>18,993,000</td>
</tr>
</tbody>
</table>

Chances are, the leading digit will be a 1 more often than a 2. And 2s would probably occur more often than 3s, and so on. **This odd phenomenon is Benford’s law.**
Benford’s law: mathematical statement

A set of numbers is said to satisfy Benford’s law if the leading digit \(d\, (d \in \{1, ..., 9\})\) occurs with probability

\[
P(d) = \log_{10}(d + 1) - \log_{10}(d) = \log_{10}\frac{d + 1}{d}
\]

\[\begin{array}{|c|c|c|}
\hline
d & P(d) & \text{Relative size of } P(d) \\
\hline
1 & 30.1\% & \text{large} \\
2 & 17.6\% & \text{moderate} \\
3 & 12.5\% & \text{not very} \\
4 & 9.7\% & \text{not very} \\
5 & 7.9\% & \text{not very} \\
6 & 6.7\% & \text{not very} \\
7 & 5.8\% & \text{not very} \\
8 & 5.1\% & \text{not very} \\
9 & 4.6\% & \text{not very} \\
\hline
\end{array}
\]

\textbf{Figure:} The distribution of first digits depicted by Benford’s law.
Benford’s law: Basic Mechanism

\[ P(d) = \log_{10}(d + 1) - \log_{10}(d) = \log_{10} \frac{d + 1}{d} \]

- Benford’s law is expected if the mantissa (fractional part) of the logarithms of the numbers (but not the numbers themselves) are uniformly and randomly distributed.
- It tends to apply most accurately to data that are distributed uniformly across many orders of magnitude.
Benford’s law: Possible Explanations

Examples:

- **bacteria size:** Outcomes of exponential growth processes.
  (e.g. $1000 \times 2^n$ on day $n$)
  - exponentially growing quantity moves on a log-scale at a constant rate.

- **stock price:** Multiplicative fluctuations
  - logarithm of the stock price is undergoing a random walk

  The leading digits of data satisfying Zipf’s law with $s = 1$ satisfies Benford’s law.

Two major steps:

- scale invariance (Pinkham, 1961)
- mixture result (Ted Hill, 1995)
Benford’s law: original table

Benford (1938) “collected data from as many fields as possible and to include a wide variety of types”.

<table>
<thead>
<tr>
<th>Group</th>
<th>Title</th>
<th>First Digit</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Rivers, Area</td>
<td>31.0</td>
<td>12.4</td>
</tr>
<tr>
<td>B</td>
<td>Population</td>
<td>33.9</td>
<td>20.4</td>
</tr>
<tr>
<td>C</td>
<td>Constants</td>
<td>41.3</td>
<td>14.4</td>
</tr>
<tr>
<td>D</td>
<td>Newspapers</td>
<td>30.0</td>
<td>18.0</td>
</tr>
<tr>
<td>E</td>
<td>Spec. Heat</td>
<td>24.0</td>
<td>18.4</td>
</tr>
<tr>
<td>F</td>
<td>Pressure</td>
<td>29.6</td>
<td>18.3</td>
</tr>
<tr>
<td>G</td>
<td>H.P. Lost</td>
<td>30.0</td>
<td>18.4</td>
</tr>
<tr>
<td>H</td>
<td>Mol. Wgt.</td>
<td>26.7</td>
<td>25.2</td>
</tr>
<tr>
<td>I</td>
<td>Drainage</td>
<td>27.1</td>
<td>23.9</td>
</tr>
<tr>
<td>J</td>
<td>Atomic Wgt.</td>
<td>47.2</td>
<td>18.7</td>
</tr>
<tr>
<td>K</td>
<td>$n^{-1}$, $\sqrt{n}$, ...</td>
<td>25.7</td>
<td>20.3</td>
</tr>
<tr>
<td>L</td>
<td>Design</td>
<td>26.8</td>
<td>14.8</td>
</tr>
<tr>
<td>M</td>
<td>Digest</td>
<td>33.4</td>
<td>18.5</td>
</tr>
<tr>
<td>N</td>
<td>Cost Data</td>
<td>32.4</td>
<td>18.8</td>
</tr>
<tr>
<td>O</td>
<td>X-Ray Volts</td>
<td>27.9</td>
<td>17.5</td>
</tr>
<tr>
<td>P</td>
<td>Am. League</td>
<td>32.7</td>
<td>17.6</td>
</tr>
<tr>
<td>Q</td>
<td>Black Body</td>
<td>31.0</td>
<td>17.3</td>
</tr>
<tr>
<td>R</td>
<td>Addresses</td>
<td>28.9</td>
<td>19.2</td>
</tr>
<tr>
<td>S</td>
<td>$n^1$, $n^2$· · · $n$!</td>
<td>25.3</td>
<td>16.0</td>
</tr>
<tr>
<td>T</td>
<td>Death Rate</td>
<td>27.0</td>
<td>18.6</td>
</tr>
<tr>
<td></td>
<td><strong>Average</strong></td>
<td><strong>30.6</strong></td>
<td><strong>18.5</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Probable Error</strong></td>
<td>±0.8</td>
<td>±0.4</td>
</tr>
</tbody>
</table>
Benford’s law: Mixture result by Ted Hill, 1995

If one repeatedly ”randomly” chooses a probability distribution (with some regularity conditions) and then randomly chooses numbers according to that distribution, the resulting list of numbers will obey Benford’s Law.

References

1. Wikipedia (Zipf’s law, Heap’s law, Benford’s law)