Power Law Size Distributions
Principles of Complex Systems
Course CSYS/MATH 300, Fall, 2009

Prof. Peter Dodds
Dept. of Mathematics & Statistics
Center for Complex Systems :: Vermont Advanced Computing Center
University of Vermont

Licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License.

Overview
Introduction
Examples
Zipf’s law
Wild vs. Mild
CCDFs

The Don

Extreme deviations in test cricket

Don Bradman’s batting average = 166% next best.

Size distributions

The sizes of many systems’ elements appear to obey an inverse power-law size distribution:

\[ P(\text{size} = x) \sim c x^{-\gamma} \]

where \( x_{\min} < x < x_{\max} \)

and \( \gamma > 1 \)

- Typically, \( 2 < \gamma < 3 \).
- \( x_{\min} \) = lower cutoff
- \( x_{\max} \) = upper cutoff
Power Law Size Distributions

Overview
Introduction
Examples
Zipf’s law
Wild vs. Mild
CCDFs
References

Size distributions

− Usually, only the tail of the distribution obeys a power law:

\[ P(x) \sim c x^{-\gamma} \text{ as } x \to \infty. \]

− Still use term ‘power law distribution’

Size distributions

Many systems have discrete sizes \( k \):

− Word frequency
− Node degree (as we have seen): # hyperlinks, etc.
− number of citations for articles, court decisions, etc.

\[ P(k) \sim c k^{-\gamma} \]

where \( k_{\text{min}} \leq k \leq k_{\text{max}} \)

Power law size distributions are sometimes called Pareto distributions \(^\square\) after Italian scholar Vilfredo Pareto.

− Pareto noted wealth in Italy was distributed unevenly (80–20 rule).
− Term used especially by economists

Size distributions

− Negative linear relationship in log-log space:

\[ \log P(x) = \log c - \gamma \log x \]
**Size distributions**

**Examples:**
- Earthquake magnitude (Gutenberg Richter law): $P(M) \propto M^{-3}$
- Number of war deaths: $P(d) \propto d^{-1.8}$
- Sizes of forest fires
- Sizes of cities: $P(n) \propto n^{-2.1}$
- Number of links to and from websites

(Note: Exponents range in error; see M.E.J. Newman arxiv.org/cond-mat/0412004v3 (Ξ))

**Power-law distributions are..**
- often called ‘heavy-tailed’
- or said to have ‘fat tails’

**Important!:**
- Inverse power laws aren’t the only ones:
  - lognormals, stretched exponentials, ...

**Zipfian rank-frequency plots**

**George Kingsley Zipf:**
- We'll study Zipf’s law in depth...
Zipfian rank-frequency plots

Zipf’s way:

- \( s_r \) = the size of the \( r \)th ranked object.
- \( r = 1 \) corresponds to the largest size.
- \( s_1 \) could be the frequency of occurrence of the most common word in a text.
- Zipf’s observation:

\[
S_r \propto r^{-\alpha}
\]

Power law distributions

Gaussians versus power-law distributions:

- Example: Height versus wealth.
- Mild versus Wild (Mandelbrot)
- Mediocristan versus Extremistan
  (See “The Black Swan” by Nassim Taleb \(^1\))

Turkeys...

From “The Black Swan” \(^1\)

Taleb’s table \(^1\)

Mediocristan/Extremistan

- Most typical member is mediocre/Most typical is either giant or tiny
- Winners get a small segment/Winner take almost all effects
- When you observe for a while, you know what’s going on/It takes a very long time to figure out what’s going on
- Prediction is easy/Prediction is hard
- History crawls/History makes jumps
- Tyranny of the collective/Tyranny of the accidental
Complementary Cumulative Distribution Function (CCDF):

\[ P_\geq(x) = P(x' \geq x) = 1 - P(x' < x) \]
\[ = \int_{x'=x}^{\infty} P(x') \, dx' \]
\[ \propto \int_{x'=x}^{\infty} (x')^{-\gamma} \, dx' \]
\[ = \frac{1}{-\gamma + 1} (x')^{-\gamma+1} \bigg|_{x'=x}^{\infty} \]
\[ \propto x^{-\gamma+1} \]

Complementary Cumulative Distribution Function:

\[ P_\geq(x) \propto x^{-\gamma+1} \]

- Use when tail of \( P \) follows a power law.
- Increases exponent by one.
- Useful in cleaning up data.

Size distributions

Brown Corpus (1,015,945 words):

**CCDF:**

\[ \log_{10} N_n > n \]

**Zipf:**

\[ \log_{10} n_i \]

- The, of, and, to, a, ... = ‘objects’
- ‘Size’ = word frequency
- **Beep:** CCDF and Zipf plots are related...
Size distributions

Brown Corpus (1,015,945 words):

CCDF:

Zipf (axes flipped):

▶ The, of, and, to, a, ... = ‘objects’
▶ ‘Size’ = word frequency
▶ Beep: CCDF and Zipf plots are related...

Details on the lack of scale:

Let’s find the mean:

\[ \langle x \rangle = \int_{x=x_{\text{min}}}^{x_{\text{max}}} x P(x) \, dx \]

\[ = c \int_{x=x_{\text{min}}}^{x_{\text{max}}} xx^{-\gamma} \, dx \]

\[ = \frac{c}{2-\gamma} \left( x_{\text{max}}^{2-\gamma} - x_{\text{min}}^{2-\gamma} \right). \]

The mean:

\[ \langle x \rangle \sim \frac{c}{2-\gamma} \left( x_{\text{max}}^{2-\gamma} - x_{\text{min}}^{2-\gamma} \right). \]

▶ Mean blows up with upper cutoff if \( \gamma < 2 \).
▶ Mean depends on lower cutoff if \( \gamma > 2 \).
▶ \( \gamma < 2 \): Typical sample is large.
▶ \( \gamma > 2 \): Typical sample is small.
And in general...

Moments:
- All moments depend only on cutoffs.
- No internal scale dominates (even matters).
- Compare to a Gaussian, exponential, etc.

Moments

Standard deviation is a mathematical convenience!:
- Variance is nice analytically...
- Another measure of distribution width:
  Mean average deviation (MAD) =
  \[ \langle |x - \langle x \rangle| \rangle \]
- MAD is unpleasant analytically...

Moments

For many real size distributions:
\[ 2 < \gamma < 3 \]
- mean is finite (depends on lower cutoff)
- \[ \sigma^2 \] = variance is ‘infinite’ (depends on upper cutoff)
- Width of distribution is ‘infinite’

How sample sizes grow...

Given \( P(x) \sim cx^{-\gamma} \):
- We can show that after \( n \) samples, we expect the largest sample to be
  \[ x_1 \gtrsim n^{1/(\gamma-1)} \]
- Sampling from a ‘mild’ distribution gives a much slower growth with \( n \).
- e.g., for \( P(x) = \lambda e^{-\lambda x} \), we find
  \[ x_1 \gtrsim \frac{1}{\lambda} \ln n. \]
References

N. N. Taleb.
*The Black Swan.*

G. K. Zipf.
*Human Behaviour and the Principle of Least-Effort.*
Addison-Wesley, Cambridge, MA, 1949.