

Power-Law Size Distributions

Principles of Complex Systems | @pocsvox

CSYS/MATH 300, Fall, 2015 | #FallPoCS2015

Prof. Peter Dodds | @peterdodds

Dept. of Mathematics & Statistics | Vermont Complex Systems Center
Vermont Advanced Computing Core | University of Vermont



Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References

$$P(x) \sim x^{-\delta}$$



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



These slides are brought to you by:

PoCS | @pocsvox

Power-Law Size
Distributions

Sealie & Lambie Productions

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References

$$P(x) \sim x^{-\delta}$$

Outline

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References

PoCS | @pocsvox

Power-Law Size
Distributions

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

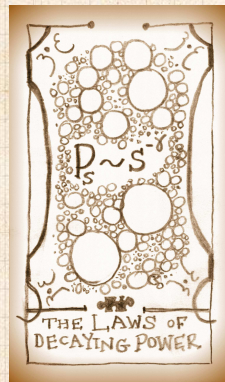
References

$$P(x) \sim x^{-\delta}$$

The deal:

PoCS | @pocsvox

Power-Law Size
Distributions



Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References

$$P(x) \sim x^{-\delta}$$

Two of the many things we struggle with cognitively:

1. Probability.

- ▶ Ex. The Monty Hall Problem.[↗](#)
- ▶ Ex. Daughter/Son born on Tuesday.[↗](#)
(see next two slides; Wikipedia entry here [↗](#).)

2. Logarithmic scales.

On counting and logarithms:



- ▶ Listen to Radiolab's 2009 piece: "Numbers."[↗](#).
- ▶ Later: Benford's Law[↗](#).

Also to be enjoyed: the magnificence of the Dunning-Kruger effect[↗](#)



Homo probabilisticus?

The set up:

- ▶ A parent has two children.

Simple probability question:

- ▶ What is the probability that both children are girls?
- ▶ 1/4...

The next set up:

- ▶ A parent has two children.
- ▶ We know one of them is a girl.

The next probabilistic poser:

- ▶ What is the probability that both children are girls?
- ▶ 1/3...

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



Try this one:

- ▶ A parent has two children.
- ▶ We know one of them is a girl **born on a Tuesday**.

Simple question #3:

- ▶ What is the probability that both children are girls?
- ▶ ?

Last:

- ▶ A parent has two children.
- ▶ We know one of them is a girl **born on December 31**.

And ...

- ▶ What is the probability that both children are girls?
- ▶ ?



Let's test our collective intuition:

PoCS | @pocsvox

Power-Law Size
Distributions



Money
≡
Belief

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References

Two questions about wealth distribution in the United States:

1. Please estimate the percentage of all wealth owned by individuals when grouped into quintiles.
2. Please estimate what you believe each quintile should own, ideally.
3. Extremes: 100, 0, 0, 0, 0 and 20, 20, 20, 20, 20



Wealth distribution in the United States:^[11]

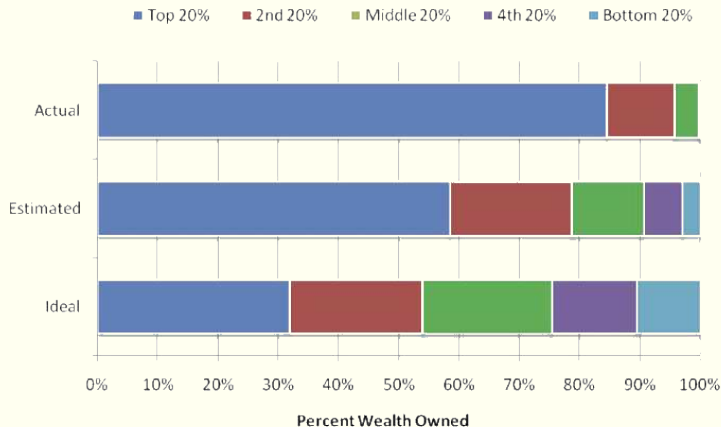


Fig. 2. The actual United States wealth distribution plotted against the estimated and ideal distributions across all respondents. Because of their small percentage share of total wealth, both the “4th 20%” value (0.2%) and the “Bottom 20%” value (0.1%) are not visible in the “Actual” distribution.

“Building a better America—One wealth quintile at a time”
Norton and Ariely, 2011.^[11]



Wealth distribution in the United States:^[11]

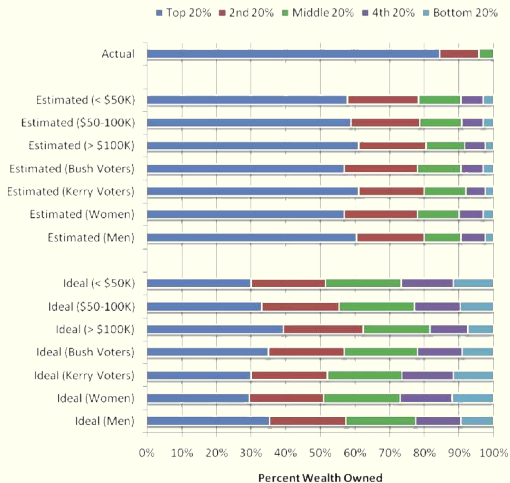


Fig. 3. The actual United States wealth distribution plotted against the estimated and ideal distributions of respondents of different income levels, political affiliations, and genders. Because of their small percentage share of total wealth, both the “4th 20%” value (0.2%) and the “Bottom 20%” value (0.1%) are not visible in the “Actual” distribution.

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



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

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

where $0 < x_{\min} < x < x_{\max}$ and $\gamma > 1$.

- ▶ x_{\min} = lower cutoff, x_{\max} = upper cutoff
- ▶ Negative linear relationship in log-log space:

$$\log_{10} P(x) = \log_{10} c - \gamma \log_{10} x$$

- ▶ We use base 10 because we are good people.
- ▶ power-law decays in probability:
The Statistics of Surprise.

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



Size distributions:

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

$$P(x) \sim c x^{-\gamma} \text{ for } x \text{ large.}$$

- ▶ Still use term 'power-law size distribution.'
- ▶ Other terms:
 - ▶ **Fat-tailed** distributions.
 - ▶ **Heavy-tailed** distributions.

Beware:

- ▶ Inverse power laws aren't the only ones:
lognormals[!\[\]\(13dd0e1ab3baa23f7c1ed52b3eec2756_img.jpg\)](#), Weibull distributions[!\[\]\(5ed985c65f50e5350eeeb77f03c2e095_img.jpg\)](#), ...

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



Size distributions:

Many systems have discrete sizes k :

- ▶ Word frequency
- ▶ Node degree in networks: # friends, # hyperlinks, etc.
- ▶ # citations for articles, court decisions, etc.

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

where $k_{\min} \leq k \leq k_{\max}$

- ▶ Obvious fail for $k = 0$.
- ▶ Again, typically a description of distribution's tail.

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



The statistics of surprise—words:

PoCS | @pocsvox

Power-Law Size
Distributions

Brown Corpus ↗ (~ 10^6 words):

rank	word	% q
1.	the	6.8872
2.	of	3.5839
3.	and	2.8401
4.	to	2.5744
5.	a	2.2996
6.	in	2.1010
7.	that	1.0428
8.	is	0.9943
9.	was	0.9661
10.	he	0.9392
11.	for	0.9340
12.	it	0.8623
13.	with	0.7176
14.	as	0.7137
15.	his	0.6886

rank	word	% q
1945.	apply	0.0055
1946.	vital	0.0055
1947.	September	0.0055
1948.	review	0.0055
1949.	wage	0.0055
1950.	motor	0.0055
1951.	fifteen	0.0055
1952.	regarded	0.0055
1953.	draw	0.0055
1954.	wheel	0.0055
1955.	organized	0.0055
1956.	vision	0.0055
1957.	wild	0.0055
1958.	Palmer	0.0055
1959.	intensity	0.0055

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



Jonathan Harris's Wordcount:

PoCS | @pocsvox

Power-Law Size
Distributions

A word frequency distribution explorer:



Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf ⇔ CCDF

Appendix

References



The statistics of surprise—words:

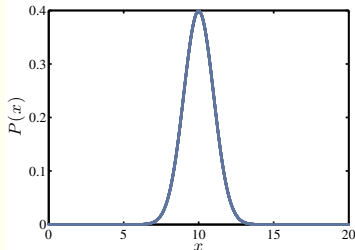
PoCS | @pocsvox

Power-Law Size
Distributions

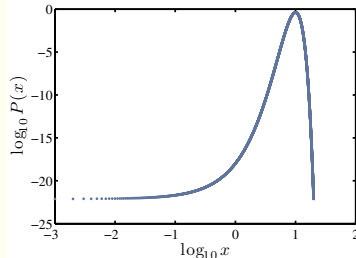
First—a Gaussian example:

$$P(x)dx = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} dx$$

linear:



log-log



mean $\mu = 10$, variance $\sigma^2 = 1$.

► **Activity:** Sketch $P(x) \sim x^{-1}$ for $x = 1$ to $x = 10^7$.

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



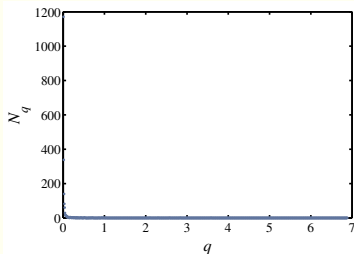
The statistics of surprise—words:

PoCS | @pocsvox

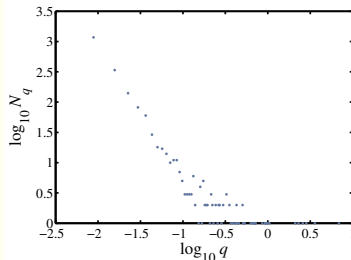
Power-Law Size
Distributions

Raw 'probability' (binned) for Brown Corpus:

linear:



log-log



q_w = frequency of occurrence of word q expressed as a percentage.

N_q = number of distinct words that have a frequency of occurrence q .

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



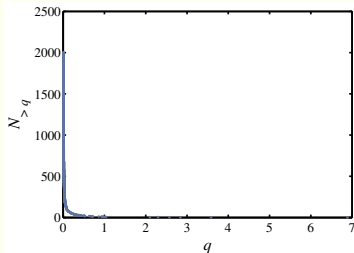
The statistics of surprise—words:

PoCS | @pocsvox

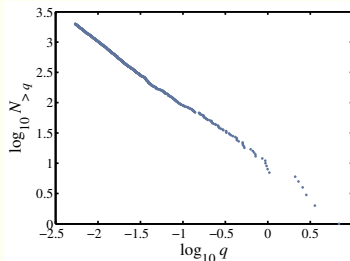
Power-Law Size
Distributions

Complementary Cumulative Probability Distribution $N_{>q}$:

linear:



log-log



► Also known as the 'Exceedance Probability.'

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



My, what big words you have...


PoCS | @pocsvox

Power-Law Size
Distributions

**Test
your
vocab**

*How many words
do you know?*



- Test  capitalizes on word frequency following a heavily skewed frequency distribution with a decaying power-law tail.

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

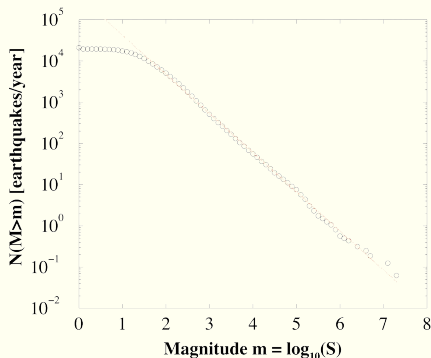
Appendix

References



The statistics of surprise:

Gutenberg-Richter law



- ▶ Log-log plot
- ▶ Base 10
- ▶ Slope = -1

$$N(M > m) \propto m^{-1}$$

- ▶ From **both** the very awkwardly similar Christensen et al. and Bak et al.:
“Unified scaling law for earthquakes” [3, 1]

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



The statistics of surprise:

From: "Quake Moves Japan Closer to U.S. and Alters Earth's Spin"  by Kenneth Chang, March 13, 2011, NYT:

'What is perhaps most surprising about the Japan earthquake is how misleading history can be. In the past 300 years, no earthquake nearly that large—nothing larger than magnitude eight—had struck in the Japan subduction zone. That, in turn, led to assumptions about how large a tsunami might strike the coast.'

"It did them a giant disservice," said Dr. Stein of the geological survey. That is not the first time that the earthquake potential of a fault has been underestimated. Most geophysicists did not think the Sumatra fault could generate a magnitude 9.1 earthquake, ...'

PoCS | @pocsvox

Power-Law Size
Distributions

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References

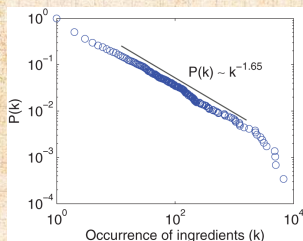




"Geography and Similarity of Regional Cuisines in China" ↗

Zhu et al.,

PLoS ONE, **8**, e79161, 2013. [16]



- ▶ Fraction of ingredients that appear in at least k recipes.
- ▶ Oops in notation: $P(k)$ is the Complementary Cumulative Distribution $P_{\geq}(k)$

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References





"On a class of skew distribution functions" ↗

Herbert A. Simon,
Biometrika, **42**, 425–440, 1955. ^[13]



"Power laws, Pareto distributions and Zipf's law" ↗

M. E. J. Newman,
Contemporary Physics, **46**, 323–351,
2005. ^[10]



"Power-law distributions in empirical data" ↗

Clauset, Shalizi, and Newman,
SIAM Review, **51**, 661–703, 2009. ^[4]

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



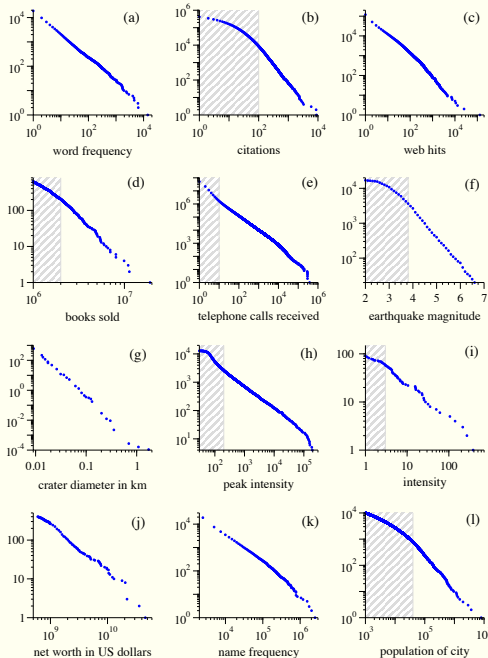



FIG. 4 Cumulative distributions or "rank/frequency plots" of twelve quantities reputed to follow power laws. The distributions were computed as described in Appendix A. Data in the shaded regions were excluded from the calculations of the exponents in Table I. Source references for the data are given in the text. (a) Numbers of occurrences of words in the novel *Moby Dick* by Hermann Melville. (b) Numbers of citations to scientific papers published in 1981, from time of publication until June 1997. (c) Numbers of hits on web sites by 60,000 users of the America Online Internet service for the day of 1 December 1997. (d) Numbers of copies of bestselling books sold in the US between 1895 and 1965. (e) Number of calls received by AT&T telephone customers in the US for a single day. (f) Magnitude of earthquakes in California between January 1910 and May 1992. Magnitude is proportional to the logarithm of the maximum amplitude of the earthquake, and hence the distribution obeys a power law even though the horizontal axis is linear. (g) Diameter of craters on the moon. Vertical axis is measured per square kilometre. (h) Peak gamma-ray intensity of solar flares in counts per second, measured from Earth orbit between February 1980 and November 1989. (i) Intensity of wars from 1816 to 1980, measured as battle deaths per 10,000 of the population of the participating countries. (j) Aggregate net worth in dollars of the richest individuals in the US in October 2003. (k) Frequency of occurrence of family names in the US in the year 1990. (l) Populations of US cities in the year 2000.

Size distributions:

Some examples:

- ▶ Earthquake magnitude (Gutenberg-Richter law 
 - ▶ Note: Exponents range in error

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF



Appendix

References



Size distributions:

More examples:

- ▶ # citations to papers: ^[5, 6] $P(k) \propto k^{-3}$.
- ▶ Individual wealth (maybe): $P(W) \propto W^{-2}$.
- ▶ Distributions of tree trunk diameters: $P(d) \propto d^{-2}$.
- ▶ The gravitational force at a random point in the universe: ^[9] $P(F) \propto F^{-5/2}$. (See the Holtmark distribution  and stable distributions .)
- ▶ Diameter of moon craters: ^[10] $P(d) \propto d^{-3}$.
- ▶ Word frequency: ^[13] e.g., $P(k) \propto k^{-2.2}$ (variable).
- ▶ # religious adherents in cults: ^[4] $P(k) \propto k^{-1.8 \pm 0.1}$.
- ▶ # sightings of birds per species (North American Breeding Bird Survey for 2003): ^[4]
 $P(k) \propto k^{-2.1 \pm 0.1}$.
- ▶ # species per genus: ^[15, 13, 4] $P(k) \propto k^{-2.4 \pm 0.2}$.

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



Table 3 from Clauset, Shalizi, and Newman^[4]:

Basic parameters of the data sets described in section 6, along with their power-law fits and the corresponding p -values (statistically significant values are denoted in **bold**).

Quantity	n	$\langle x \rangle$	σ	x_{\max}	\hat{x}_{\min}	$\hat{\alpha}$	n_{tail}	p
count of word use	18 855	11.14	148.33	14 086	7 ± 2	1.95(2)	2958 ± 987	0.49
protein interaction degree	1846	2.34	3.05	56	5 ± 2	3.1(3)	204 ± 263	0.31
metabolic degree	1641	5.68	17.81	468	4 ± 1	2.8(1)	748 ± 136	0.00
Internet degree	22 688	5.63	37.83	2583	21 ± 9	2.12(9)	770 ± 1124	0.29
telephone calls received	51 360 423	3.88	179.09	375 746	120 ± 49	2.09(1)	$102 592 \pm 210 147$	0.63
intensity of wars	115	15.70	49.97	382	2.1 ± 3.5	1.7(2)	70 ± 14	0.20
terrorist attack severity	9101	4.35	31.58	2749	12 ± 4	2.4(2)	547 ± 1663	0.68
HTP size (kilobytes)	226 386	7.36	57.94	10 971	36.25 ± 22.74	2.48(5)	6794 ± 2232	0.00
species per genus	509	5.59	6.94	56	4 ± 2	2.4(2)	233 ± 138	0.10
bird species sightings	591	3384.36	10 952.34	138 705	6679 ± 2463	2.1(2)	66 ± 41	0.55
blackouts ($\times 10^3$)	211	253.87	610.31	7500	230 ± 90	2.3(3)	59 ± 35	0.62
sales of books ($\times 10^3$)	633	1986.67	1396.60	19 077	2400 ± 430	3.7(3)	139 ± 115	0.66
population of cities ($\times 10^3$)	19 447	9.00	77.83	8 009	52.46 ± 11.88	2.37(8)	580 ± 177	0.76
email address books size	4581	12.45	21.49	333	57 ± 21	3.5(6)	196 ± 449	0.16
forest fire size (acres)	203 785	0.90	20.99	4121	6324 ± 3487	2.2(3)	521 ± 6801	0.05
solar flare intensity	12 773	689.41	6520.59	231 300	323 ± 89	1.79(2)	1711 ± 384	1.00
quake intensity ($\times 10^3$)	19 302	24.54	563.83	63 096	0.794 ± 80.198	1.64(4)	$11 697 \pm 2159$	0.00
religious followers ($\times 10^6$)	103	27.36	136.64	1050	3.85 ± 1.60	1.8(1)	39 ± 26	0.42
freq. of surnames ($\times 10^3$)	2753	50.59	113.99	2502	111.92 ± 40.67	2.5(2)	239 ± 215	0.20
net worth (mil. USD)	400	2388.69	4 167.35	46 000	900 ± 364	2.3(1)	302 ± 77	0.00
citations to papers	415 229	16.17	44.02	8904	160 ± 35	3.16(6)	3455 ± 1859	0.20
papers authored	401 445	7.21	16.52	1416	133 ± 13	4.3(1)	988 ± 377	0.90
hits to web sites	119 724	9.83	392.52	129 641	2 ± 13	1.81(8)	$50 981 \pm 16 898$	0.00
links to web sites	241 428 853	9.15	106 871.65	1 199 466	3684 ± 151	2.336(9)	$28 986 \pm 1560$	0.00

- We'll explore various exponent measurement techniques in assignments.

power-law size distributions

PoCS | @pocsvox

Power-Law Size
Distributions

Gaussians versus power-law size distributions:

- ▶ Mediocristan versus Extremistan
- ▶ **Mild** versus **Wild** (Mandelbrot)
- ▶ Example: Height versus wealth.

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References

THE BLACK SWAN



The Impact of the
HIGHLY IMPROBABLE

Nassim Nicholas Taleb

- ▶ See “The Black Swan” by Nassim Taleb.^[14]

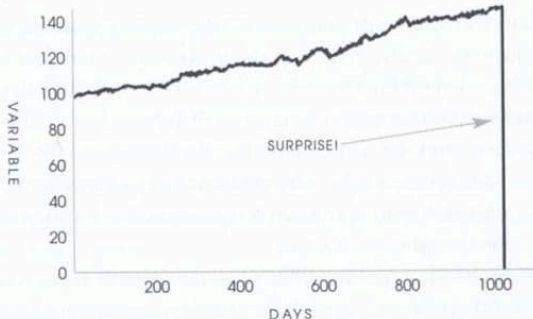


Turkeys...

PoCS | @pocsvox

Power-Law Size
Distributions

FIGURE 1: ONE THOUSAND AND ONE DAYS OF HISTORY



A turkey before and after Thanksgiving. The history of a process over a thousand days tells you nothing about what is to happen next. This naïve projection of the future from the past can be applied to anything.

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



From "The Black Swan" ^[14]

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 rare and accidental

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



Size distributions:

PoCS | @pocsvox

Power-Law Size
Distributions

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



Power-law size distributions are sometimes called Pareto distributions after Italian scholar Vilfredo Pareto.

- ▶ Pareto noted wealth in Italy was distributed unevenly (80–20 rule; misleading).
- ▶ Term used especially by practitioners of the Dismal Science.



Devilish power-law size distribution details:

PoCS | @pocsvox

Power-Law Size
Distributions

Exhibit A:

- ▶ Given $P(x) = cx^{-\gamma}$ with $0 < x_{\min} < x < x_{\max}$, the mean is ($\gamma \neq 2$):

$$\langle x \rangle = \frac{c}{2-\gamma} (x_{\max}^{2-\gamma} - x_{\min}^{2-\gamma}).$$

- ▶ 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.

Insert question from assignment 2 

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References




And in general...

Moments:

- ▶ All moments depend only on cutoffs.
- ▶ No internal scale that dominates/matters.
- ▶ Compare to a Gaussian, exponential, etc.

For many real size distributions: $2 < \gamma < 3$

- ▶ mean is finite (depends on lower cutoff)
- ▶ σ^2 = variance is 'infinite' (depends on upper cutoff)
- ▶ Width of distribution is 'infinite'
- ▶ If $\gamma > 3$, distribution is less terrifying and may be easily confused with other kinds of distributions.

Insert question from assignment 2 

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



Standard deviation is a mathematical convenience:


- ▶ Variance is nice analytically...
- ▶ Another measure of distribution width:

$$\text{Mean average deviation (MAD)} = \langle |x - \langle x \rangle| \rangle$$

- ▶ For a pure power law with $2 < \gamma < 3$:

$$\langle |x - \langle x \rangle| \rangle \text{ is finite.}$$

- ▶ But MAD is mildly unpleasant analytically...
- ▶ We still speak of infinite 'width' if $\gamma < 3$.

Insert question from assignment 2 

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



How sample sizes grow...

Given $P(x) \sim cx^{-\gamma}$:

- ▶ We can show that after n samples, we expect the largest sample to be

$$x_1 \gtrsim c' n^{1/(\gamma-1)}$$

- ▶ Sampling from a finite-variance distribution gives a much slower growth with n .
- ▶ e.g., for $P(x) = \lambda e^{-\lambda x}$, we find

$$x_1 \gtrsim \frac{1}{\lambda} \ln n.$$

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



Insert question from assignment 2 

CCDF:

$$P_{\geq}(x) = P(x' \geq x) = 1 - P(x' < x)$$

$$= \int_{x'=x}^{\infty} P(x') dx'$$

$$\propto \int_{x'=x}^{\infty} (x')^{-\gamma} dx'$$

$$= \frac{1}{-\gamma + 1} (x')^{-\gamma+1} \Big|_{x'=x}^{\infty}$$

$$\propto x^{-\gamma+1}$$

[Our Intuition](#)[Definition](#)[Examples](#)[Wild vs. Mild](#)[CCDFs](#)[Zipf's law](#)[Zipf \$\Leftrightarrow\$ CCDF](#)[Appendix](#)[References](#)

Complementary Cumulative Distribution Function:

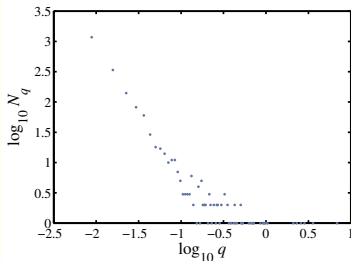
CCDF:



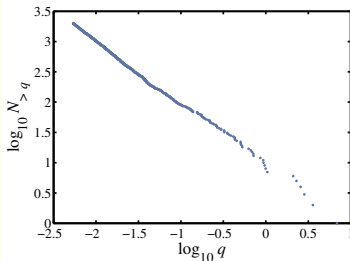
$$P_{\geq}(x) \propto x^{-\gamma+1}$$

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

PDF:



CCDF:



Complementary Cumulative Distribution Function:

- ▶ Same story for a discrete variable: $P(k) \sim ck^{-\gamma}$.



$$P_{\geq}(k) = P(k' \geq k)$$

$$= \sum_{k'=k}^{\infty} P(k)$$

$$\propto k^{-\gamma+1}$$

- ▶ Use integrals to approximate sums.

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References




Zipfian rank-frequency plots

PoCS | @pocsvox

Power-Law Size
Distributions

George Kingsley Zipf:

- ▶ Noted various rank distributions have power-law tails, often with exponent -1 (word frequency, city sizes...)
- ▶ Zipf's 1949 Magnum Opus :
- ▶ We'll study Zipf's law in depth...

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



Zipfian rank-frequency plots

PoCS | @pocsvox

Power-Law Size
Distributions

Zipf's way:

- ▶ Given a collection of entities, rank them by size, largest to smallest.
- ▶ x_r = the size of the r th ranked entity.
- ▶ $r = 1$ corresponds to the largest size.
- ▶ Example: x_1 could be the frequency of occurrence of the most common word in a text.
- ▶ Zipf's observation:

$$x_r \propto r^{-\alpha}$$

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

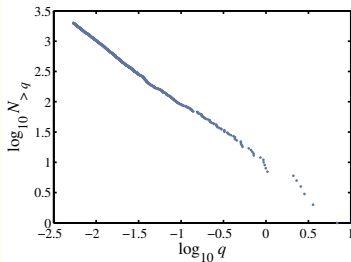
References



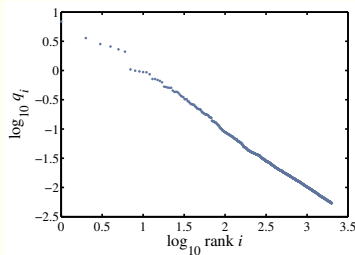
Size distributions:

Brown Corpus (1,015,945 words):

CCDF:



Zipf:



- ▶ The, of, and, to, a, ... = 'objects'
- ▶ 'Size' = word frequency
- ▶ **Beep:** (Important) CCDF and Zipf plots are related...

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

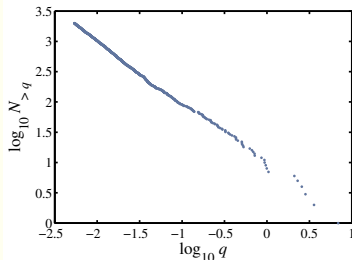
References



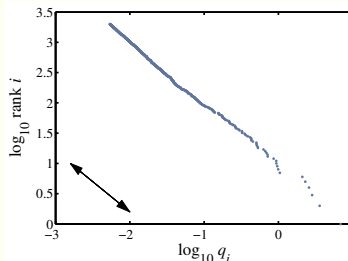
Size distributions:

Brown Corpus (1,015,945 words):

CCDF:



Zipf:



- ▶ The, of, and, to, a, ... = 'objects'
- ▶ 'Size' = word frequency
- ▶ **Beep:** (Important) CCDF and Zipf plots are related...

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



Observe:

- ▶ $NP_{\geq}(x)$ = the number of objects with size at least x
where N = total number of objects.
- ▶ If an object has size x_r , then $NP_{\geq}(x_r)$ is its rank r .
- ▶ So

$$x_r \propto r^{-\alpha} = (NP_{\geq}(x_r))^{-\alpha}$$

$$\propto x_r^{(-\gamma+1)(-\alpha)} \text{ since } P_{\geq}(x) \sim x^{-\gamma+1}.$$

We therefore have $1 = (-\gamma + 1)(-\alpha)$ or:

$$\alpha = \frac{1}{\gamma - 1}$$

- ▶ A rank distribution exponent of $\alpha = 1$ corresponds to a size distribution exponent $\gamma = 2$.

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

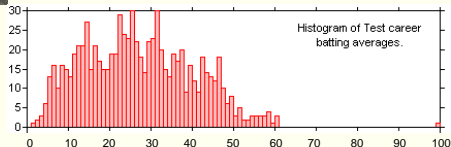
Zipf \Leftrightarrow CCDF

Appendix

References



Extreme deviations in test cricket:



- ▶ Don Bradman's batting average ↗
= 166% next best.
- ▶ That's pretty solid.
- ▶ Later in the course: Understanding success—is the Mona Lisa like Don Bradman?

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



A good eye:

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

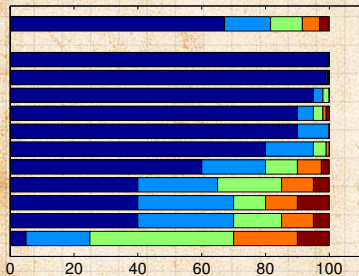
Zipf \Leftrightarrow CCDF

Appendix

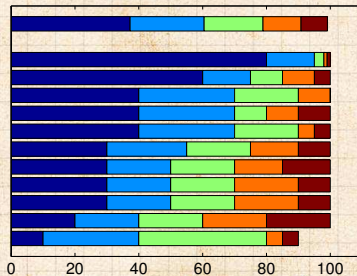
References



Actual:
Fall 2014:



Ideal:
Fall 2014:



Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

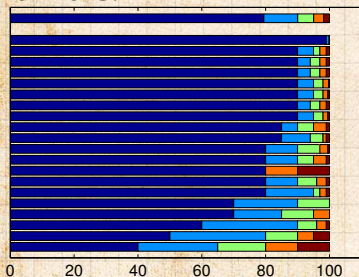
Zipf \Leftrightarrow CCDF

Appendix

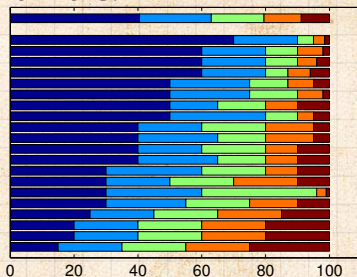
References



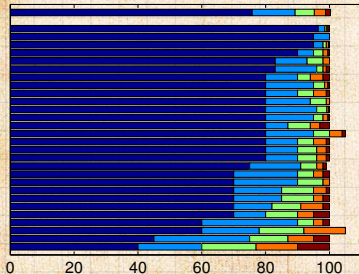
Actual:
Fall 2013:



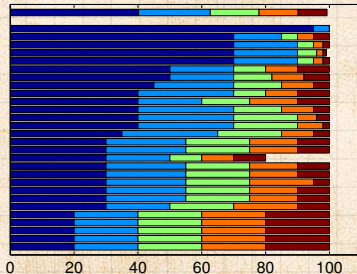
Ideal:
Fall 2013:



Spring 2013:



Spring 2013:



PoCS | @pocsvox

Power-Law Size
Distributions

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



References I

- [1] P. Bak, K. Christensen, L. Danon, and T. Scanlon.
Unified scaling law for earthquakes.
[Phys. Rev. Lett.](#), 88:178501, 2002. pdf ↗
- [2] A.-L. Barabási and R. Albert.
Emergence of scaling in random networks.
[Science](#), 286:509–511, 1999. pdf ↗
- [3] K. Christensen, L. Danon, T. Scanlon, and P. Bak.
Unified scaling law for earthquakes.
[Proc. Natl. Acad. Sci.](#), 99:2509–2513, 2002. pdf ↗
- [4] A. Clauset, C. R. Shalizi, and M. E. J. Newman.
Power-law distributions in empirical data.
[SIAM Review](#), 51:661–703, 2009. pdf ↗

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



References II

- [5] D. J. de Solla Price.
Networks of scientific papers.
[Science](#), 149:510–515, 1965. pdf ↗
- [6] D. J. de Solla Price.
A general theory of bibliometric and other
cumulative advantage processes.
[J. Amer. Soc. Inform. Sci.](#), 27:292–306, 1976.
- [7] P. Grassberger.
Critical behaviour of the Drossel-Schwabl forest
fire model.
[New Journal of Physics](#), 4:17.1–17.15, 2002. pdf ↗
- [8] B. Gutenberg and C. F. Richter.
Earthquake magnitude, intensity, energy, and
acceleration.
[Bull. Seism. Soc. Am.](#), 499:105–145, 1942. pdf ↗

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References



References III

- [9] J. Holtsmark.
Über die verbreiterung von spektrallinien.
[Ann. Phys.](#), 58:577–, 1919.
- [10] M. E. J. Newman.
Power laws, Pareto distributions and Zipf's law.
[Contemporary Physics](#), 46:323–351, 2005. pdf ↗
- [11] M. I. Norton and D. Ariely.
Building a better America—One wealth quintile at
a time.
[Perspectives on Psychological Science](#), 6:9–12,
2011. pdf ↗
- [12] L. F. Richardson.
Variation of the frequency of fatal quarrels with
magnitude.
[J. Amer. Stat. Assoc.](#), 43:523–546, 1949. pdf ↗

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law



Zipf \Leftrightarrow CCDF

Appendix

References



References IV

- [13] H. A. Simon.
On a class of skew distribution functions.
[Biometrika](#), 42:425–440, 1955. [pdf](#) 
- [14] N. N. Taleb.
[The Black Swan](#).
Random House, New York, 2007.
- [15] G. U. Yule.
A mathematical theory of evolution, based on the
conclusions of Dr J. C. Willis, F.R.S.
[Phil. Trans. B](#), 213:21–, 1924.
- [16] Y.-X. Zhu, J. Huang, Z.-K. Zhang, Q.-M. Zhang,
T. Zhou, and Y.-Y. Ahn.
Geography and similarity of regional cuisines in
china.
[PLoS ONE](#), 8:e79161, 2013. [pdf](#) 

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

[References](#)



Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf \Leftrightarrow CCDF

Appendix

References

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

