

# Power-Law Size Distributions

Last updated: 2018/09/16, 11:25:58

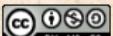
Principles of Complex Systems | @pocsvox  
CSYS/MATH 300, Fall, 2018

Prof. Peter Dodds | @peterdodds

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



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



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



# These slides are brought to you by:

Sealie & Lambie  
Productions



PoCS  
@pocsvox

Power-Law Size  
Distributions

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf  $\leftrightarrow$  CCDF

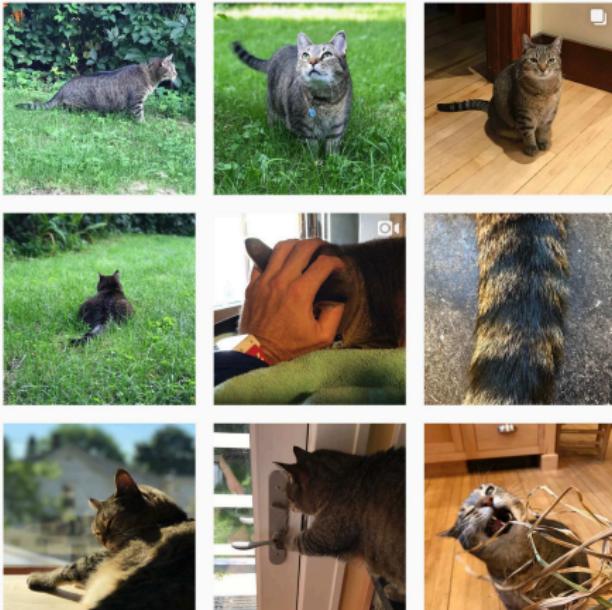
References

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



# These slides are also brought to you by:

## Special Guest Executive Producer



Our Intuition  
Definition  
Examples  
Wild vs. Mild  
CCDFs  
Zipf's law  
 $\text{Zipf} \leftrightarrow \text{CCDF}$   
References

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

On Instagram at [pratchett\\_the\\_cat](#)



# Outline

PoCS  
@pocsvox

Power-Law Size  
Distributions

Our Intuition

Our Intuition

Definition

Definition

Examples

Examples

Wild vs. Mild

Wild vs. Mild

CCDFs

CCDFs

Zipf's law

Zipf's law

Zipf  $\Leftrightarrow$  CCDF

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

References

References



# Two of the many things we struggle with cognitively:

## 1. Probability.

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

## 2. Logarithmic scales.

### On counting and logarithms:



- cube Listen to Radiolab's 2009 piece:  
"Numbers." ↗.
- cube Later: Benford's Law ↗.

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

Also to be enjoyed: the magnificence of the Dunning-Kruger effect ↗

Our Intuition  
Definition  
Examples  
Wild vs. Mild  
CCDFs  
Zipf's law  
Zipf  $\leftrightarrow$  CCDF  
References



# Homo probabilisticus?

## The set up:

- ⬢ A parent has two children.

Simple probability question:

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

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?

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



## 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?

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



# Let's test our collective intuition:

PoCS  
@pocsvox

Power-Law Size  
Distributions



Money  
≡  
Belief

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

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



## Our Intuition

Definition

Examples

Wild vs. Mild

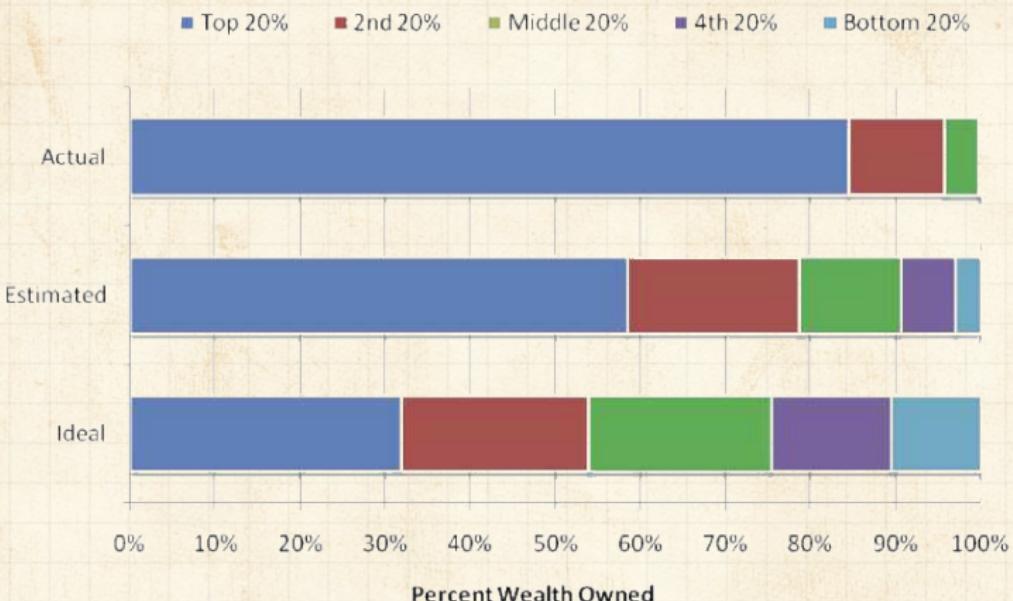
CCDFs

Zipf's law

Zipf  $\leftrightarrow$  CCDF

References

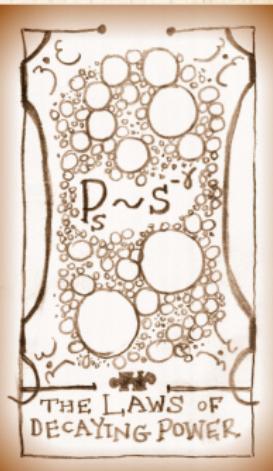
## Wealth distribution in the United States: [13]



**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.

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

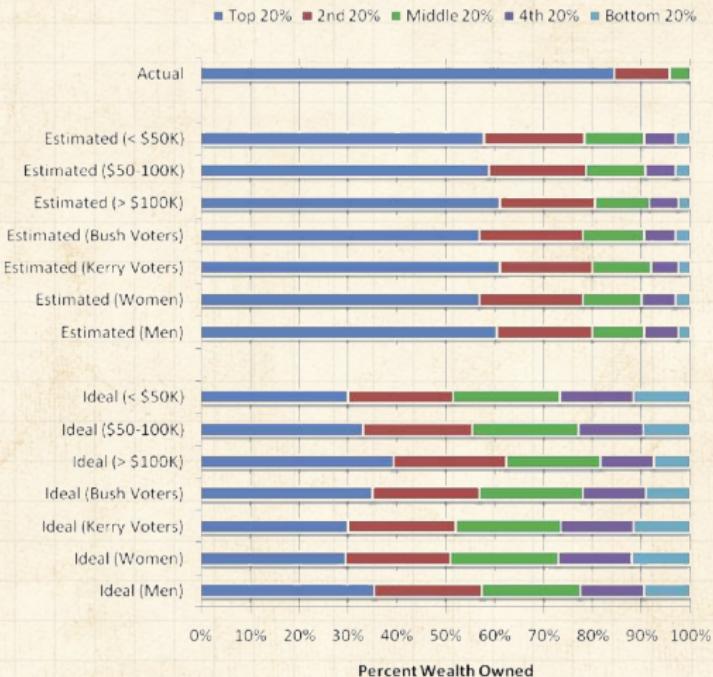
“Building a better America—One wealth quintile at a time”  
Norton and Ariely, 2011. [13]



# Wealth distribution in the United States: [13]

PoCS  
@pocsvox

## Power-Law Size Distributions



**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

References



A highly watched video based on this research is

# The Boggoracle Speaks:

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf  $\leftrightarrow$  CCDF

References



# The Boggoracle Speaks:

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf  $\leftrightarrow$  CCDF

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  $\gamma > 1$ .

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf  $\leftrightarrow$  CCDF

References

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

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

- ➌ We use base 10 because we are **good people**.



# 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 ↗, Weibull distributions ↗, ...

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf’s law

Zipf  $\leftrightarrow$  CCDF

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.



# Word frequency:

Brown Corpus ↗ ( $\sim 10^6$  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 ⇔ CCDF  
References



Jonathan Harris's Wordcount: ↗

## A word frequency distribution explorer:



PoCS  
@pocsvox

## Power-Law Size Distributions

## Our Intuition

### Definition

## Examples

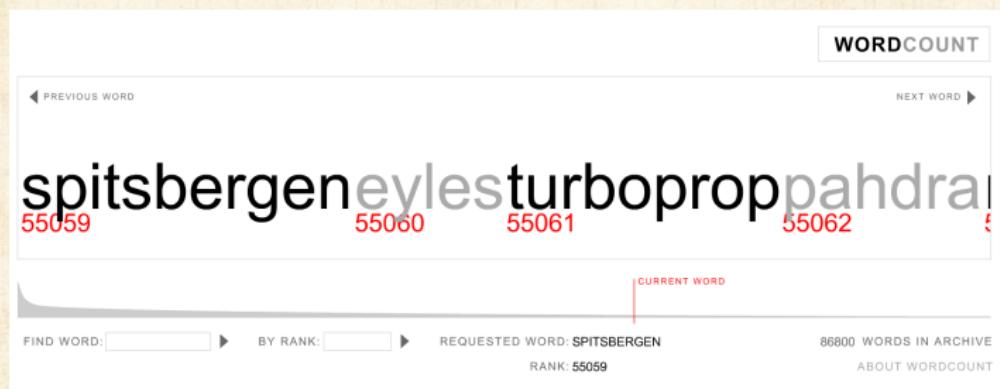
### Wild vs. Mild

CCDEs

### Zipf's law

Zipf ⇔ CCDF

## References



The cover of the book 'Thing Explainer: Complicated Stuff in Simple Words' by Randall Munroe. The title is at the top in large red letters, with 'by RANDALL MUNROE' below it. Below the title is a black and white illustration of a complex machine or system, possibly a nuclear reactor, with various pipes, valves, and components labeled with simple words like 'FUEL' and 'COOLANT'.

"Thing Explainer: Complicated Stuff in Simple Words" [a ↗](#)  
by Randall Munroe (2015). [11]

## Our Intuition

### Definition

## Examples

## Wild vs. Mild

CCDFs

## Zipf's law

Zipf  $\leftrightarrow$  CCDF

## References



# BOAT THAT GOES UNDER THE SEA

We've always had boats that go under the sea, but in the last few hundred years, we've learned to make ones that come back up.

**WORLD-ENDING BOAT**  
The boat shown here carries up to two dozen city-burning war machines. People have added up the power used during the Second World War—all the machines that blew up, all the guns that fired, and all the cities that burned. It's a lot of fire power. Each of these boats carries several times that amount.

**SPECIAL SEA WORDS**  
Most of the time, if you call a really big boat a "boat," people who know a lot about boats will get mad at you. But boats that go under the sea are really called "vessels."

#### S HEAVY METAL POWER MACHINE

These boats are powered by heavy metal, just like some power buildings. This means they can stay hidden for a long time without running out of power. Any time heavy metal is used for power, people worry about something going wrong. Of course, given what these boats are built for, people worry even more about the idea of one of these working right.

used those boats to shoot at  
make holes in them, or stick  
them that blew up.

**BREATHING STICK**  
This brings fresh air into the boat, but the boat can also make its own air by breaking water into the parts it's made of. This takes a lot of power, but the boat is powered by heavy metal, so it.

just like  
I stay  
out of power.  
People worry  
about what  
I'm more

ter, we found a new use for these boats: keeping our city-burning machines hidden, safe, and ready to use if there's a war.

**MIRROR LOOKERS**  
When the boat is hiding under the sea, it can come near the surface and use these sticks with mirrors in them to let the people outside see out of the water.

Light can't go far under

**OTHER BOATS THAT GO UNDER THE SEA**  
These are some other boats, drawn to show how big they are next to the medium-sized boat above.

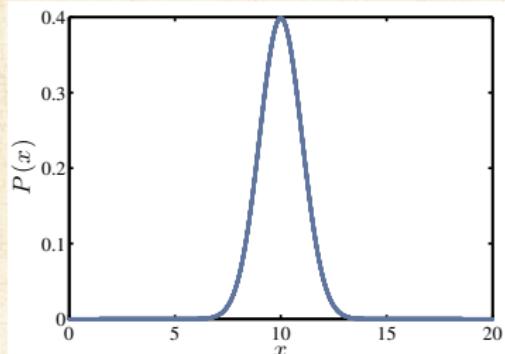


# The statistics of surprise—words:

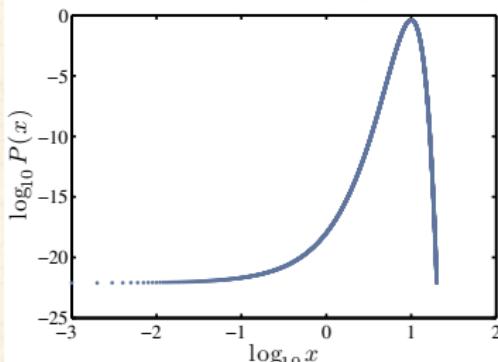
First—a Gaussian example:

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

linear:



log-log



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

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf  $\leftrightarrow$  CCDF

References



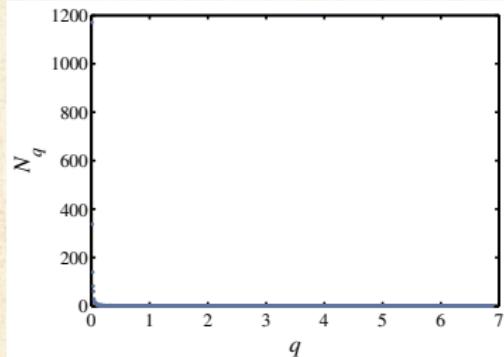
**Activity:** Sketch  $P(x) \sim x^{-1}$  for  $x = 1$  to  $x = 10^7$ .



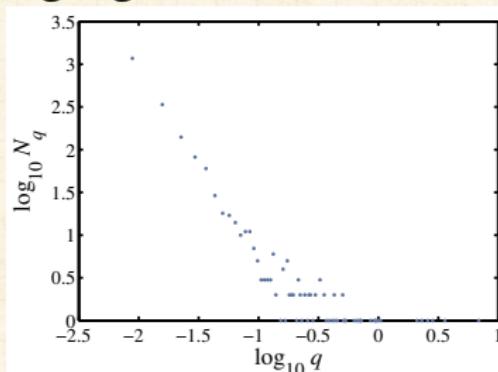
# The statistics of surprise—words:

Raw ‘probability’ (binned) for Brown Corpus:

linear:



log-log



- ❖  $q_w$  = normalized frequency of occurrence of word  $w$  (%).
- ❖  $N_q$  = number of distinct words that have a normalized frequency of occurrence  $q$ .
- ❖ e.g.,  $q_{\text{the}} \simeq 6.9\%$ ,  $N_{q_{\text{the}}} = 1$ .



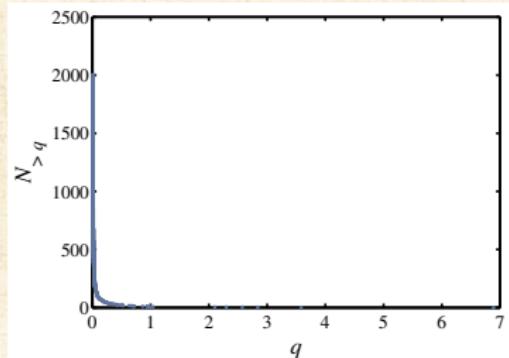
# The statistics of surprise—words:

PoCS  
@pocsvox

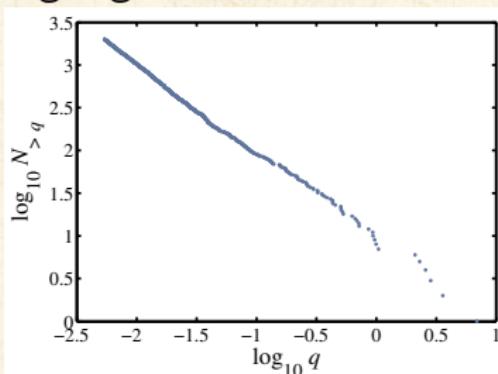
Power-Law Size  
Distributions

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

References



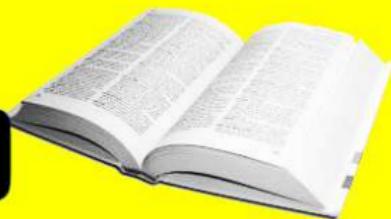
# My, what big words you have ...

PoCS  
@pocsvox

Power-Law Size  
Distributions

# Test your vocab

*How many words  
do you know?*

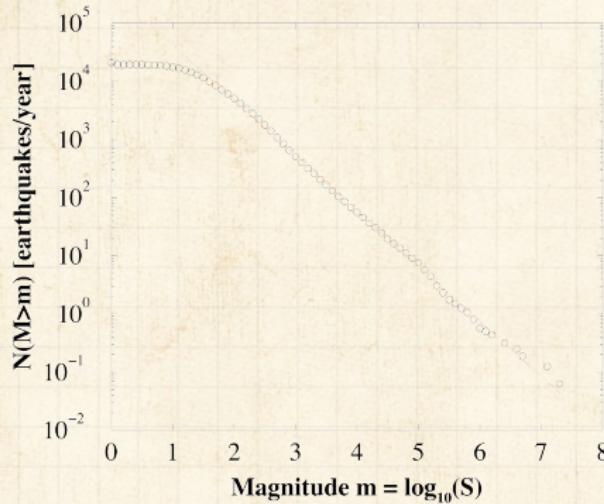


- ➊ Test ↗ capitalizes on word frequency following a heavily skewed frequency distribution with a decaying power-law tail.
- ➋ This Man Can Pronounce Every Word in the Dictionary ↗ (story here ↗)
- ➌ Best of Dr. Bailly ↗



# The statistics of surprise:

## Gutenberg-Richter law ↗



Log-log plot



Base 10



Slope = -1

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

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf  $\leftrightarrow$  CCDF

References



- From both the very awkwardly similar Christensen et al. and Bak et al.:  
“Unified scaling law for earthquakes” [4, 1]



Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf  $\leftrightarrow$  CCDF

References

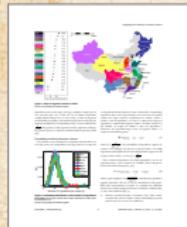
# The statistics of surprise:

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, ...'

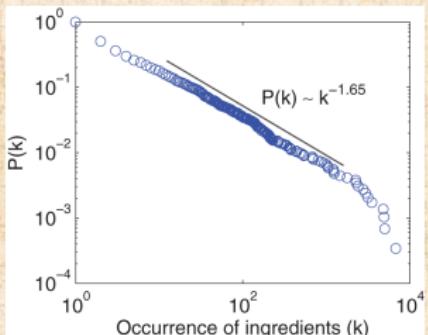




## "Geography and similarity of regional cuisines in China" ↗

Zhu et al.,

PLoS ONE, 8, e79161, 2013. [18]



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

References



Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

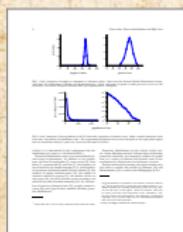
Zipf  $\leftrightarrow$  CCDF

References



## "On a class of skew distribution functions" ↗

Herbert A. Simon,  
Biometrika, **42**, 425–440, 1955. [15]



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

M. E. J. Newman,  
Contemporary Physics, **46**, 323–351,  
2005. [12]



## "Power-law distributions in empirical data" ↗

Clauset, Shalizi, and Newman,  
SIAM Review, **51**, 661–703, 2009. [5]



Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

$\text{Zipf} \Leftrightarrow \text{CCDF}$

References

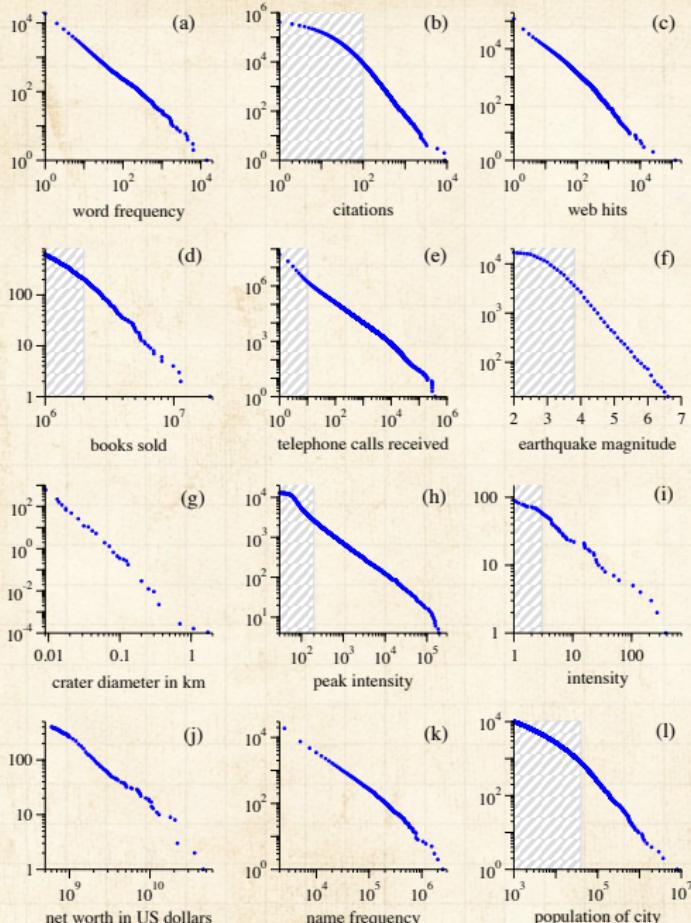


FIG. 4 Cumulative distributions or “rank/frequency plots” of twelve quantities reputed to follow power laws. The distributions were computed as described in Appendix A. Data in the shaded regions were excluded from the calculations of the exponents in Table I. Source references for the data are given in the text. (a) Numbers of occurrences of words in the novel *Moby Dick* by Herman 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 California between January 1910 and May 1992. Magnitude is proportional to the logarithm of the 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:

PoCS  
@pocsvox

Power-Law Size  
Distributions

## Some examples:

- Earthquake magnitude (Gutenberg-Richter law [9, 1]):  $P(M) \propto M^{-2}$
- # war deaths: [14]  $P(d) \propto d^{-1.8}$
- Sizes of forest fires [8]
- Sizes of cities: [15]  $P(n) \propto n^{-2.1}$
- # links to and from websites [2]
- Note: Exponents range in error



# Size distributions:

## More examples:

- .getBlockIcon() # citations to papers: [6, 7]  $P(k) \propto k^{-3}$ .
- .getBlockIcon() Individual wealth (maybe):  $P(W) \propto W^{-2}$ .
- getBlockIcon() Distributions of tree trunk diameters:  $P(d) \propto d^{-2}$ .
- getBlockIcon() The gravitational force at a random point in the universe: [10]  $P(F) \propto F^{-5/2}$ . (See the Holtsmark distribution ↗ and stable distributions ↗.)
- getBlockIcon() Diameter of moon craters: [12]  $P(d) \propto d^{-3}$ .
- getBlockIcon() Word frequency: [15] e.g.,  $P(k) \propto k^{-2.2}$  (variable).
- getBlockIcon() # religious adherents in cults: [5]  $P(k) \propto k^{-1.8 \pm 0.1}$ .
- getBlockIcon() # sightings of birds per species (North American Breeding Bird Survey for 2003): [5]  
 $P(k) \propto k^{-2.1 \pm 0.1}$ .
- getBlockIcon() # species per genus: [17, 15, 5]  $P(k) \propto k^{-2.4 \pm 0.2}$ .

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf ⇔ CCDF

References



# Table 3 from Clauset, Shalizi, and Newman [5]:

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



We'll explore various exponent measurement techniques in assignments.

# power-law size distributions

PoCS  
@poocsvox

Power-Law Size  
Distributions

## Gaussians versus power-law size distributions:

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

Our Intuition  
Definition  
Examples  
Wild vs. Mild  
CCDFs  
Zipf's law  
Zipf  $\leftrightarrow$  CCDF  
References

### THE BLACK SWAN



The Impact of the  
HIGHLY IMPROBABLE

Nassim Nicholas Taleb

- .getBlockIcon() See "The Black Swan" by Nassim Taleb.<sup>[16]</sup>
- .getBlockIcon() Terrible if successful framing:  
Black swans are not that surprising ...

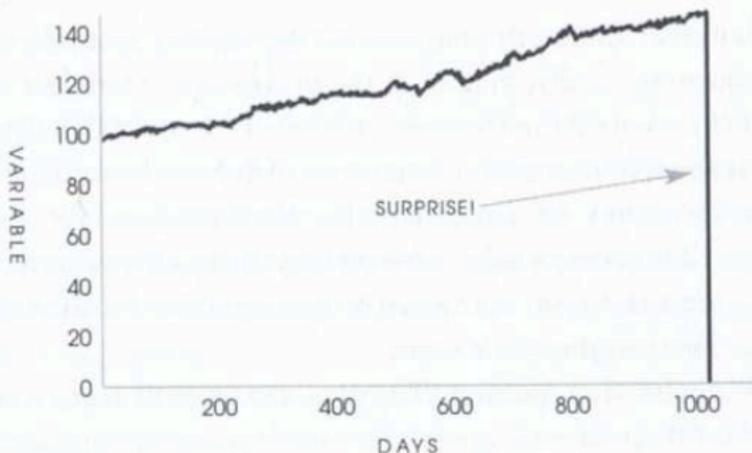


# Turkeys ...

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.



From "The Black Swan"<sup>[16]</sup>



## 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



# Size distributions:



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

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

PoCS  
@pocsvox  
Power-Law Size Distributions

Our Intuition  
Definition  
Examples  
Wild vs. Mild  
CCDFs  
Zipf's law  
Zipf  $\leftrightarrow$  CCDF  
References



# Devilish power-law size distribution details:

## 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

References



# And in general ...

PoCS  
@poocsvox

Power-Law Size  
Distributions

## Moments:

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

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf  $\leftrightarrow$  CCDF

References

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'
- If  $\gamma > 3$ , distribution is less terrifying and may be easily confused with other kinds of distributions.



Insert question from assignment 3 ↗



# Moments

PoCS  
@pocsvox

Power-Law Size  
Distributions

Standard deviation is a mathematical convenience:

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

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

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

$\langle |x - \langle x \rangle| \rangle$  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

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

References



Insert question from assignment 2 ↗



## Complementary Cumulative Distribution Function:

CCDF:



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

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf  $\leftrightarrow$  CCDF

References



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



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



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



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



## 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.

Our Intuition

Definition

Examples

Wild vs. Mild

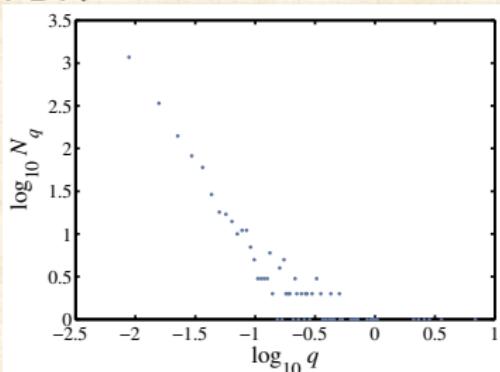
CCDFs

Zipf's law

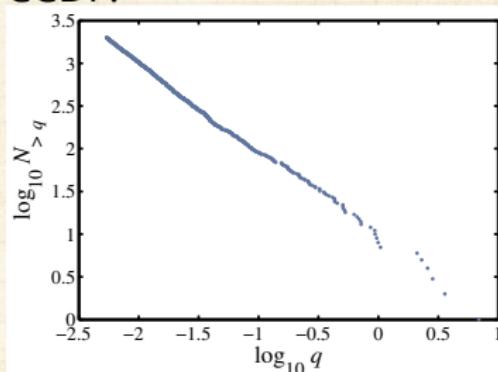
Zipf  $\leftrightarrow$  CCDF

References

## PDF:



## CCDF:



## Complementary Cumulative Distribution Function:

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf  $\leftrightarrow$  CCDF

References

 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.

# The Boggoracle Speaks:

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf  $\leftrightarrow$  CCDF

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 ↗:

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf  $\leftrightarrow$  CCDF

References



- ⬢ We'll study Zipf's law in depth ...



# 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

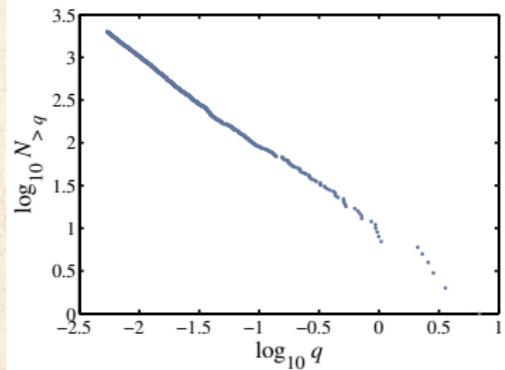
References



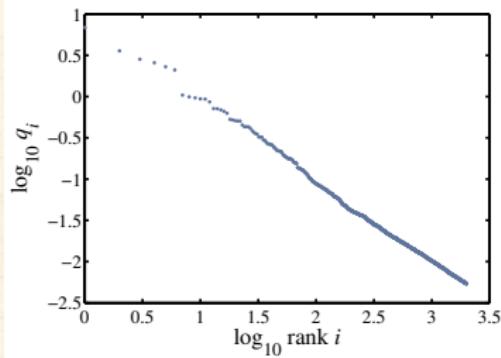
# Size distributions:

Brown Corpus (1,015,945 words):

CCDF:



Zipf:



Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf  $\leftrightarrow$  CCDF

References

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

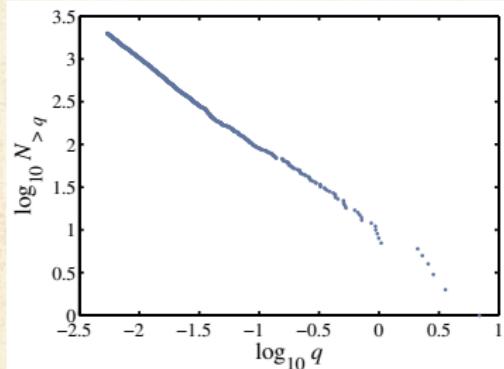
...



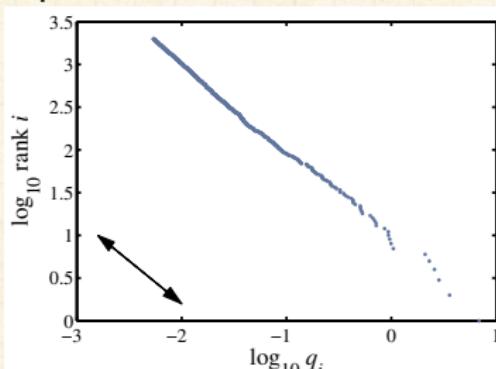
# Size distributions:

Brown Corpus (1,015,945 words):

CCDF:



Zipf:



Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf  $\leftrightarrow$  CCDF

References

- 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

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:

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

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





## "Zipf's Law in the Popularity Distribution of Chess Openings" ↗

Blasius and Tönjes,  
Phys. Rev. Lett., **103**, 218701, 2009. [3]

- ⬢ Examined all games of varying game depth  $d$  in a set of chess databases.
- ⬢  $n$  = popularity = how many times a specific game path appears in databases.
- ⬢  $S(n; d)$  = number of depth  $d$  games with popularity  $n$ .
- ⬢ Show "the frequencies of opening moves are distributed according to a power law with an exponent that increases linearly with the game depth, whereas the pooled distribution of all opening weights follows Zipf's law with universal exponent."
- ⬢ Propose hierarchical fragmentation model that produces self-similar game trees.

PoCS  
@pocts vox  
Power-Law Size Distributions

Our Intuition  
Definition  
Examples  
Wild vs. Mild  
CCDFs  
Zipf's law  
Zipf  $\leftrightarrow$  CCDF  
References



Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf  $\leftrightarrow$  CCDF

References

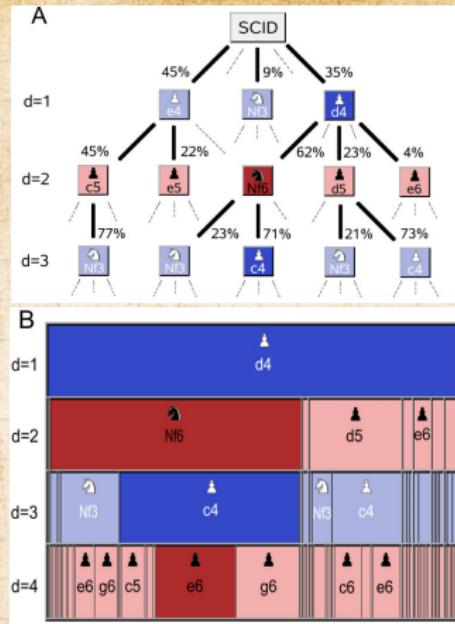


FIG. 1 (color online). (a) Schematic representation of the weighted game tree of chess based on the SCIDBASE [6] for the first three half moves. Each node indicates a state of the game. Possible game continuations are shown as solid lines together with the branching ratios  $r_d$ . Dotted lines symbolize other game continuations, which are not shown. (b) Alternative representation emphasizing the successive segmentation of the set of games, here indicated for games following a 1.d4 opening until the fourth half move  $d = 4$ . Each node  $\sigma$  is represented by a box of a size proportional to its frequency  $n_\sigma$ . In the subsequent half move these games split into subsets (indicated vertically below) according to the possible game continuations. Highlighted in (a) and (b) is a popular opening sequence 1.d4 Nf6 2.c4 e6 (Indian defense).

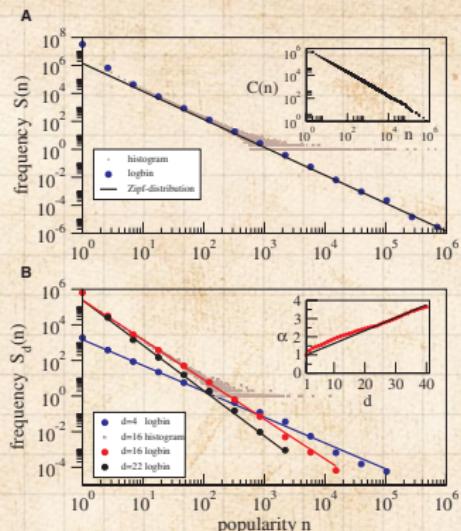
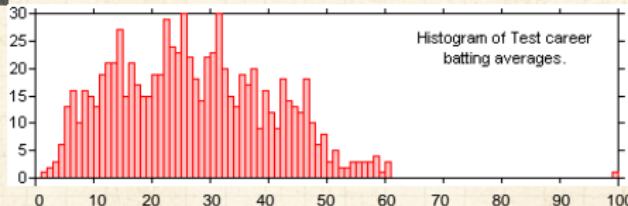


FIG. 2 (color online). (a) Histogram of weight frequencies  $S(n)$  of openings up to  $d = 40$  in the Scid database and with logarithmic binning. A straight line fit (not shown) yields an exponent of  $\alpha = 2.05$  with a goodness of fit  $R^2 > 0.9992$ . For comparison, the Zipf distribution Eq. (8) with  $\mu = 1$  is indicated as a solid line. Inset: number  $C(n) = \sum_{m=n+1}^N S(m)$  of openings with a popularity  $m > n$ .  $C(n)$  follows a power law with exponent  $\alpha = 1.04$  ( $R^2 = 0.994$ ). (b) Number  $S_d(n)$  of openings of depth  $d$  with a given popularity  $n$  for  $d = 4$ ,  $d = 16$ , and  $d = 22$ . Solid lines are regression lines to the logarithmically binned data ( $R^2 > 0.99$  for  $d < 35$ ). Inset: slope  $\alpha_d$  of the regression line as a function of  $d$  and the analytical estimation Eq. (6) using  $N = 1.4 \times 10^6$  and  $\beta = 0$  (solid line).



## Extreme deviations in test cricket:



- ➊ Don Bradman's batting average  = 166% next best.
- ➋ That's pretty solid.
- ➌ 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

References



# A good eye:

PoCS  
@pocsvox

Power-Law Size  
Distributions

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf  $\leftrightarrow$  CCDF

References



# References I

PoCS  
@poocsvox

Power-Law Size  
Distributions

- [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] B. Blasius and R. Tönjes.  
Zipf's law in the popularity distribution of chess  
openings.

Phys. Rev. Lett., 103:218701, 2009. pdf ↗

- [4] K. Christensen, L. Danon, T. Scanlon, and P. Bak.  
Unified scaling law for earthquakes.  
Proc. Natl. Acad. Sci., 99:2509–2513, 2002. pdf ↗

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf  $\leftrightarrow$  CCDF

References



# References II

PoCS  
@poocsvox

Power-Law Size  
Distributions

- [5] 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

- [6] D. J. de Solla Price.  
Networks of scientific papers.  
*Science*, 149:510–515, 1965. [pdf ↗](#)

Definition

- [7] D. J. de Solla Price.  
A general theory of bibliometric and other  
cumulative advantage processes.  
*J. Amer. Soc. Inform. Sci.*, 27:292–306, 1976. [pdf ↗](#)

Examples

- [8] P. Grassberger.  
Critical behaviour of the Drossel-Schwabl forest  
fire model.  
*New Journal of Physics*, 4:17.1–17.15, 2002. [pdf ↗](#)

Wild vs. Mild

CCDFs

Zipf's law

Zipf  $\leftrightarrow$  CCDF

References



# References III

PoCS  
@pocsvox

Power-Law Size  
Distributions

- [9] B. Gutenberg and C. F. Richter.  
Earthquake magnitude, intensity, energy, and acceleration.  
Bull. Seism. Soc. Am., 499:105–145, 1942. pdf ↗
- [10] J. Holtsmark.  
Über die verbreiterung von spektrallinien.  
Ann. Phys., 58:577–, 1919.
- [11] R. Munroe.  
Thing Explainer: Complicated Stuff in Simple Words.  
Houghton Mifflin Harcourt, 2015.
- [12] M. E. J. Newman.  
Power laws, Pareto distributions and Zipf's law.  
Contemporary Physics, 46:323–351, 2005. pdf ↗

Our Intuition  
Definition  
Examples  
Wild vs. Mild  
CCDFs  
Zipf's law  
Zipf  $\leftrightarrow$  CCDF  
References



# References IV

PoCS  
@poocsvox

Power-Law Size  
Distributions

- [13] 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 ↗

- [14] L. F. Richardson.

Variation of the frequency of fatal quarrels with magnitude.

J. Amer. Stat. Assoc., 43:523–546, 1949.

- [15] H. A. Simon.

On a class of skew distribution functions.

Biometrika, 42:425–440, 1955. pdf ↗

- [16] N. N. Taleb.

The Black Swan.

Random House, New York, 2007.

Our Intuition

Definition

Examples

Wild vs. Mild

CCDFs

Zipf's law

Zipf  $\leftrightarrow$  CCDF

References



# References V

PoCS  
@pocsvox

Power-Law Size  
Distributions

[17] G. U. Yule.

A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis, F.R.S.

Phil. Trans. B, 213:21–87, 1925. pdf ↗

[18] 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 ↗

[19] G. K. Zipf.

Human Behaviour and the Principle of Least-Effort.

Addison-Wesley, Cambridge, MA, 1949.

