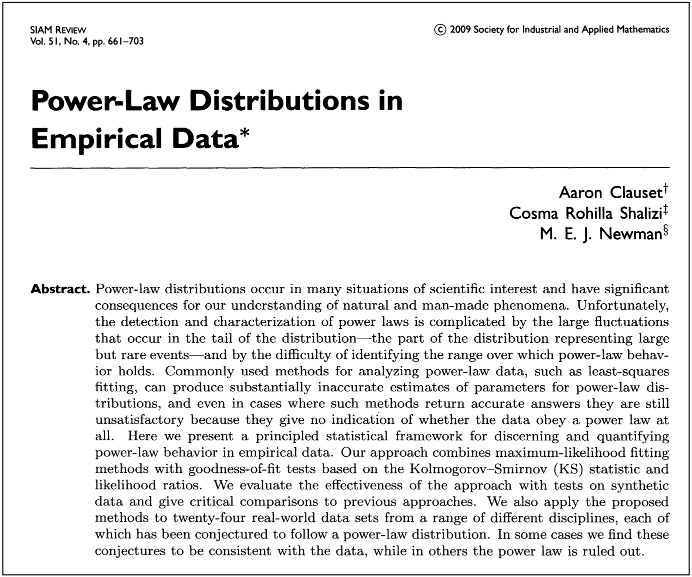
Power-Law Distributions in Empirical Data

Article for Advanced Methods in Applied Statistics

Christian Anker Rosiek

8th March 2018



SIAM Review
Vol. 51, No. 4, pp. 661–703

Power-Law Distributions in Empirical Data*

Aaron Clauset[†]
Cosma Rohilla Shalizi[‡]
M. E. J. Newman[§]

Abstract. Power-law distributions occur in many situations of scientific interest and have significant consequences for our understanding of natural and man-made phenomena. Unfortunately, the detection and characterization of power laws is complicated by the large fluctuations that occur in the tail of the distribution—the part of the distribution representing large but rare events—and by the difficulty of identifying the range over which power-law behavior holds. Commonly used methods for analyzing power-law data, such as least-squares fitting, can produce substantially inaccurate estimates of parameters for power-law distributions, and even in cases where such methods return accurate answers they are still unsatisfactory because they give no indication of whether the data obey a power law at all. Here we present a principled statistical framework for discerning and quantifying power-law behavior in empirical data. Our approach combines maximum-likelihood fitting methods with goodness-of-fit tests based on the Kolmogorov–Smirnov (KS) statistic and likelihood ratios. We evaluate the effectiveness of the approach with tests on synthetic data and give critical comparisons to previous approaches. We also apply the proposed methods to twenty-four real-world data sets from a range of different disciplines, each of which has been conjectured to follow a power-law distribution. In some cases we find these conjectures to be consistent with the data, while in others the power law is ruled out.

doi:10.1137/070710111

`http://tuvalu.santafe.edu/~aaronc/powerlaws/`

Power law distributions

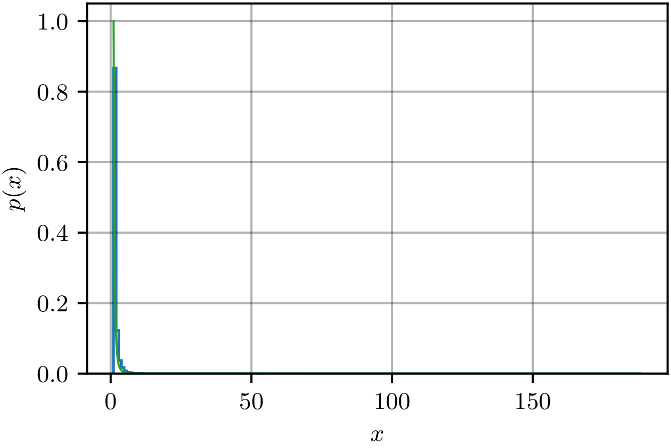
Continuous distribution

$$p(x) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-\alpha} \qquad (1)$$

Discrete distribution

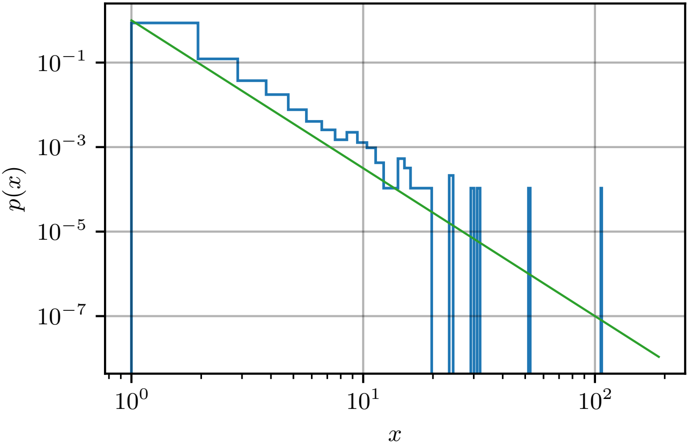
$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{\min})} \qquad (2)$$

Power-law histogram (continuous distribution)



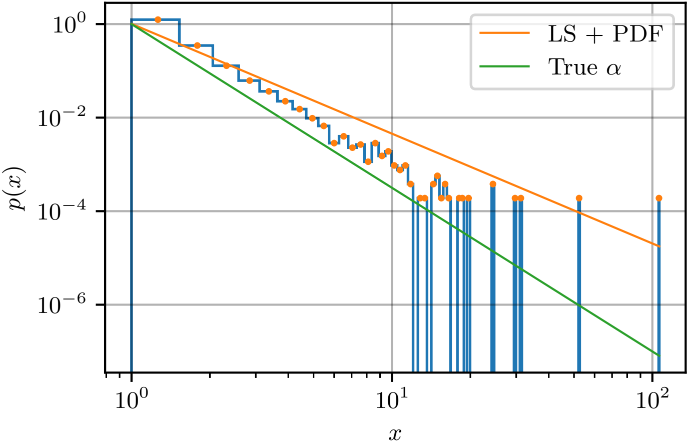
$$n = 10\,000\,, \quad \alpha = 3.5\,, \quad x_{\min} = 1\,.$$

Power-law histogram (continuous distribution)



$$n = 10\,000\,, \quad \alpha = 3.5\,, \quad x_{\min} = 1\,.$$

Linear least squares fit



$$n = 10\,000\,, \quad \alpha = 3.5\,, \quad x_{\min} = 1 \quad \rightarrow \quad \hat{\alpha}_{\mathsf{LS}} = 3.34(10)\,.$$

Maximum likelihood parameter estimation

Continuous distribution

$$\hat{\alpha}_{\textsf{MLE}} = 1 + n \left(\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}} \right)^{-1} . \tag{3}$$

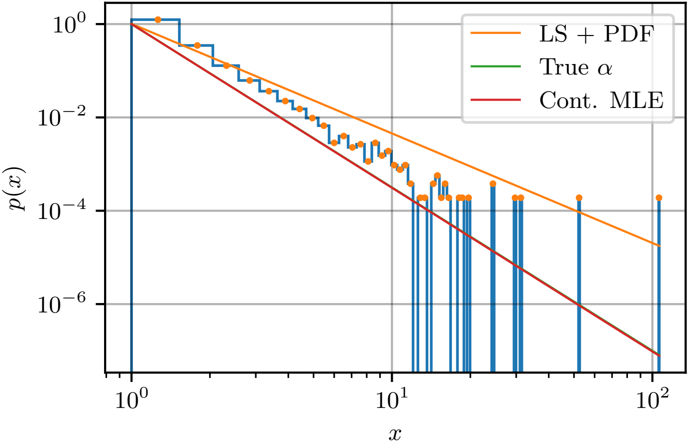
Discrete distribution

$$\hat{\alpha}_{\textsf{MLE}} = \arg\max_{\alpha} \mathcal{L} \tag{4a}$$

with

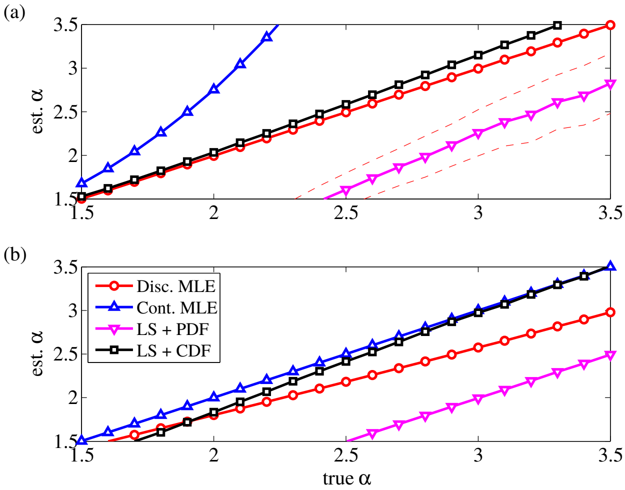
$$\mathcal{L} = -n \ln \zeta(\alpha, x_{\min}) - \alpha \sum_{i=1}^{n} \ln x_i \, . \tag{4b}$$

Maximum likelihood parameter estimation



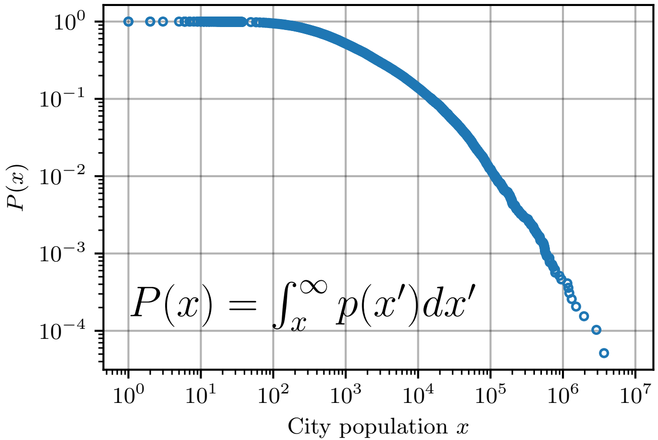
$$n = 10\,000\,, \quad \alpha = 3.5\,, \quad x_{\min} = 1 \quad \rightarrow \quad \hat{\alpha}_{\mathsf{MLE}} = 3.51(2)\,.$$

Parameter estimation comparison

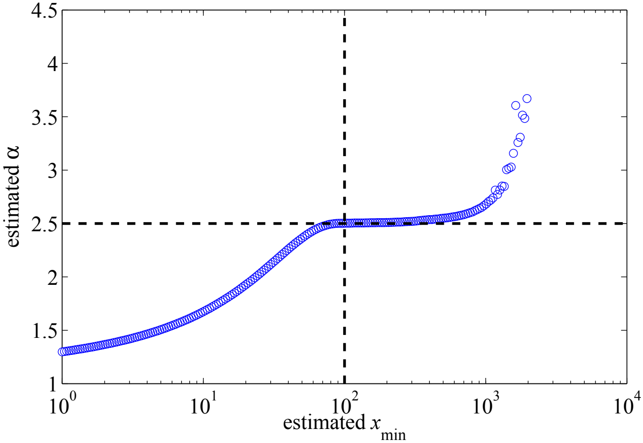


Article [1] Figure 3.2. Different α-estimators used with (a) discrete and (b) continuous power-laws.

US city population

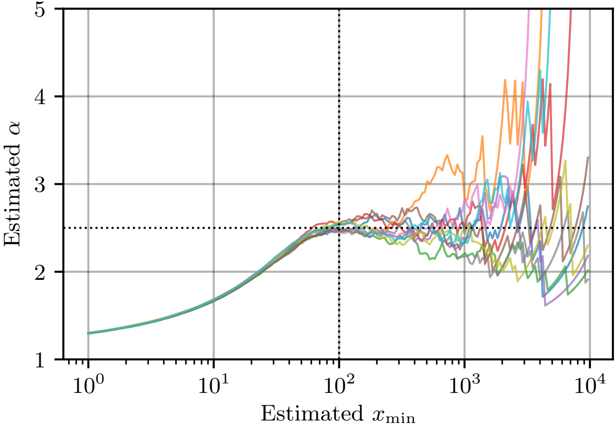


Estimating cut-off $x_{\min}$



Article [1] Figure 3.3. 5000 samples with $\alpha = 2.5$, $x_{\min} = 100$ averaged over 2500 trials.

Estimating cut-off $x_{\min}$



5000 samples with $\alpha = 2.5\,,\ x_{\min} = 100$. 10 individual trials.

Estimating cut-off $x_{\min}$

One method: Maximize similarity between measured data distribution and best-fit distribution. Similarity is here measured with

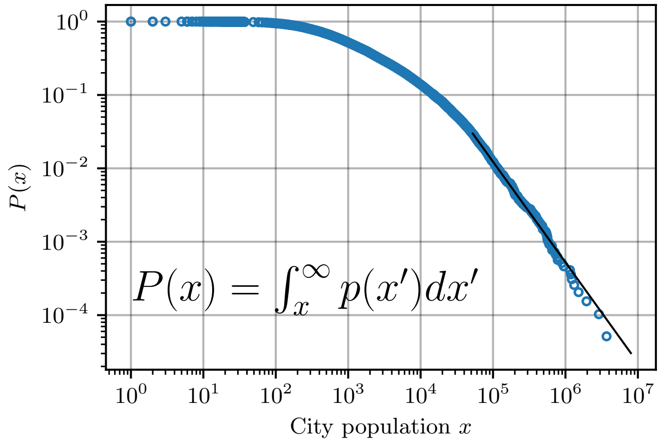
Kolmogorov-Smirnov test statistic:

$$D = \max_{x \geq x_{\min}} |S(x) - P(x)| \tag{5}$$

where $P(x)$ is measured data CDF and $S(x)$ is best-fit CDF.

Additionally, proposed Monte Carlo GOF: Sample a large number of artificial observations from distributions with the best-fit parameters. p-value is now the ratio of simulated samples that have worse D. (Note: Greater p-value is better.)

US city population



Rounding off

Not covered here:

- Model comparison using likelihood ratios.
- Application to real-world datasets.
- Appendices: Mathematical and computational details, e.g. MLE convergence, sampling from power-law distributions, etc.

Follow-up article [2]: *Power-law distributions in binned empirical data.*

References:

[1] Aaron Clauset, Cosma Rohilla Shalizi, and M.E.J. Newman. *Power-law distributions in empirical data.* SIAM Review **51**, 661–703 (2009).

[2] Y. Virkar, and A. Clauset. *Power-law distributions in binned empirical data.* The Annals of Applied Statistics **8**, 89–119 (2014).