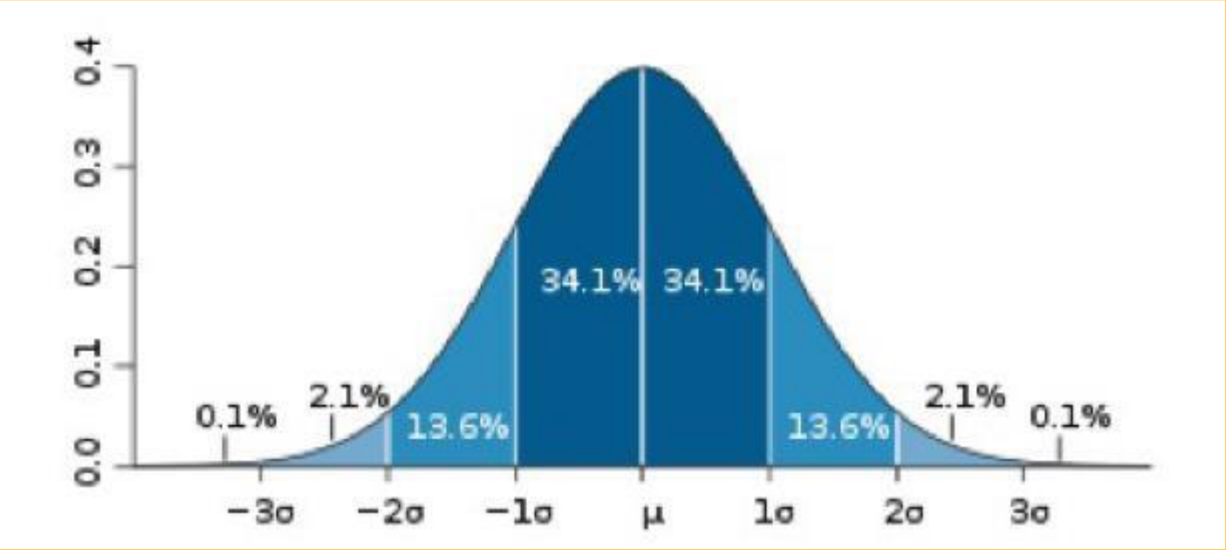
Universidade de São Paulo

FFCLRP: Lei de Potência

Prof$_a$: Fernando FF

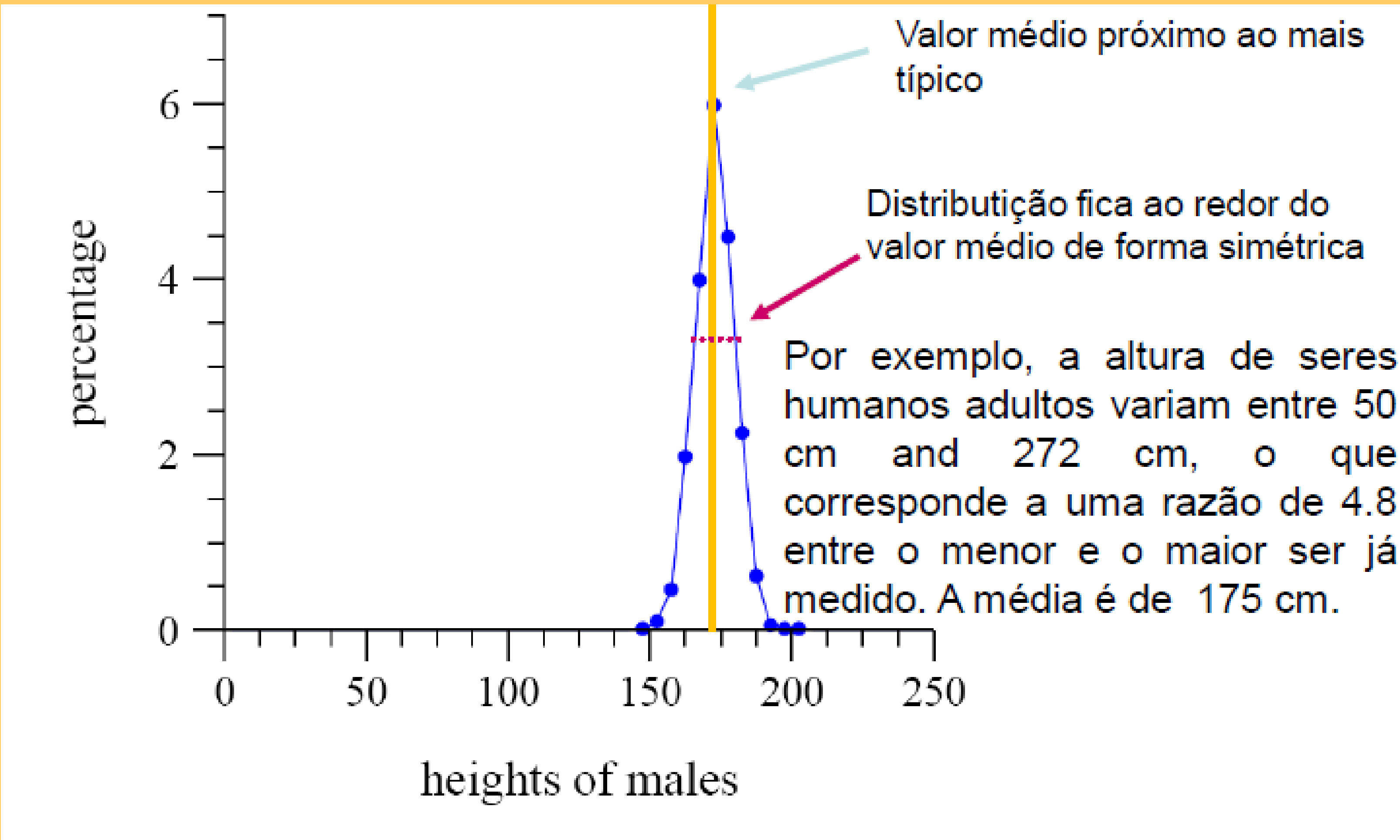
Ribeirão Pretro
Junho - 2020

Por que a base das ciências aplicadas é a distribuição normal?



Escalas Típicas

Muitas coisas que os cientistas medem tem um tamanho típico ou escala definida

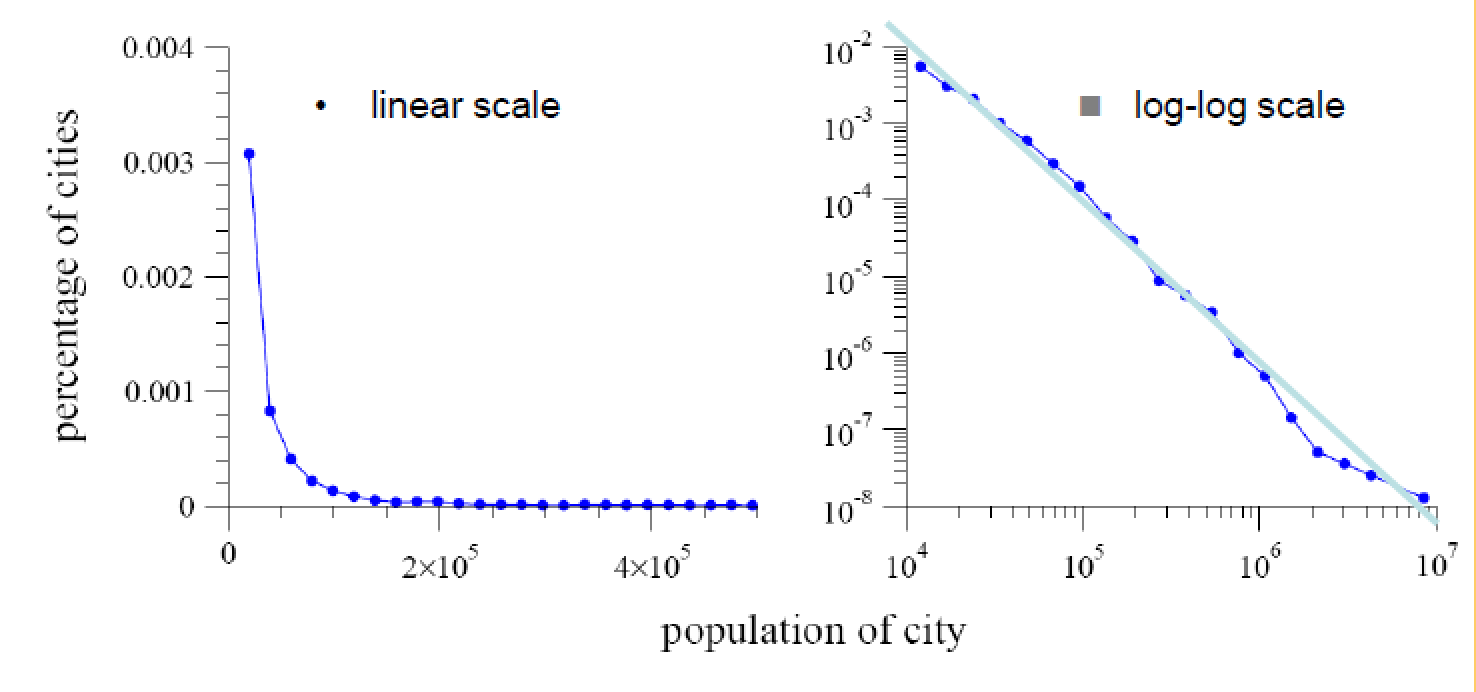


Distribuições Leis de Potência

$$P(x) = C\, x^{-\alpha}$$

Distribuições Leis de Potência

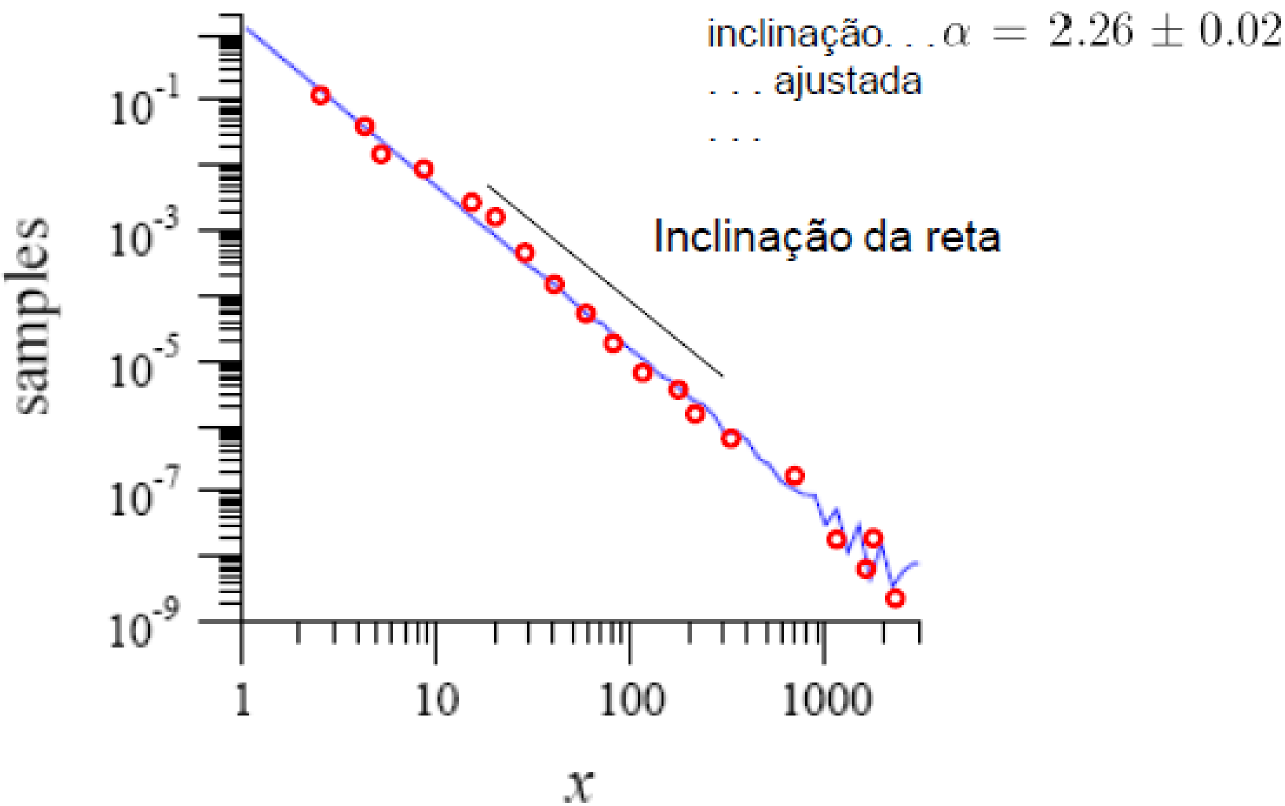
$P(x) = C\,x^{-\alpha}$



Alta assimetria
Linha reda no log-log plot
Ausência de Escala

Log-log plot

$P(x) = C\, x^{-\alpha}$



<u>NO typical value or a typical scale</u> (all sizes, all scales).

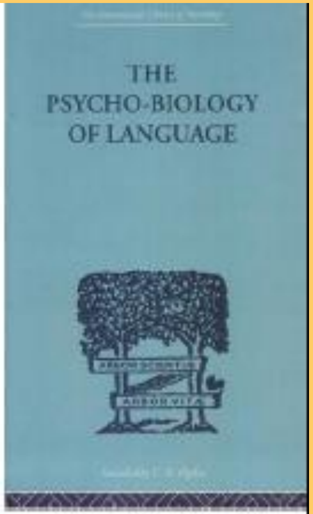
power law exponent α

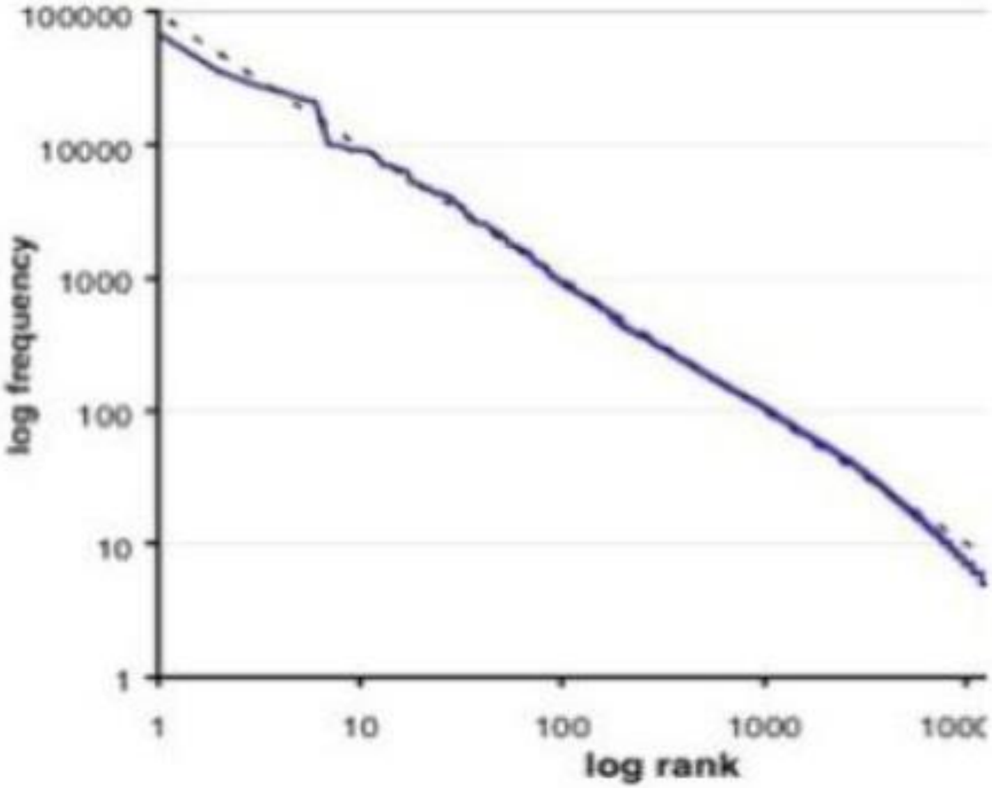
$$P(x) = C\,x^{-\alpha}$$

normalization constant (probabilities over all x must sum to 1)

- **Lei de Zipf** $f(x) = x^{-1}$

- *A segunda palavra no raking (x) tem a metade da probabilidade de ocorrência que a primeira.*





• Lei de Pareto

- The Italian economist Vilfredo Pareto was interested in the distribution of income.

- Pareto's law is expressed in terms of the cumulative distribution
  - the probability that a person earns X or more

$$P[X > x] \sim x^{-k}$$

- Here we recognize k as just α -1, where α is the power-law exponent

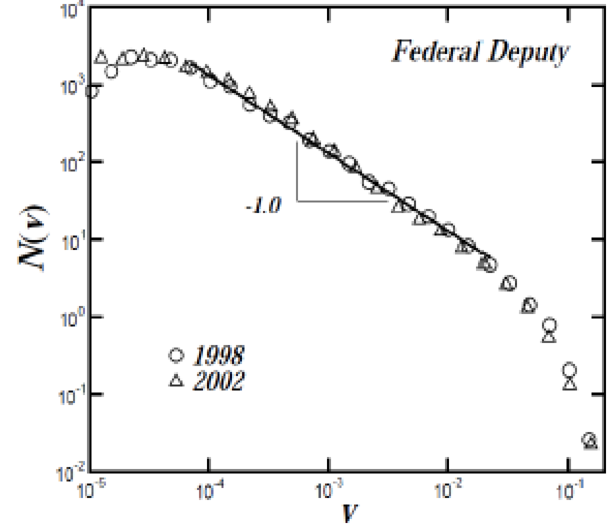
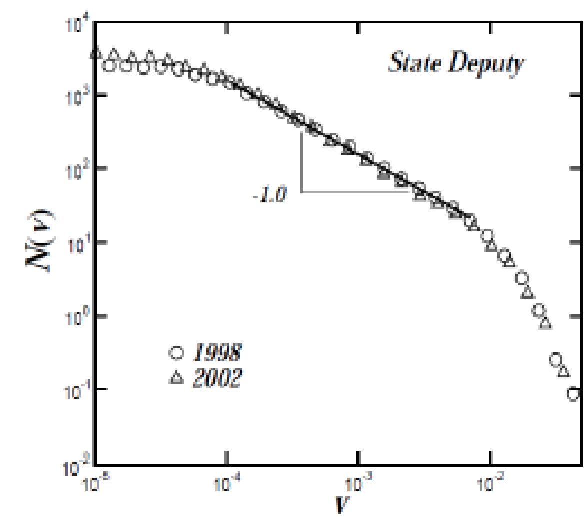


FIG. 2: (a) Double logarithmic plot of the voting distribution for state deputies in 1998 (cir
and 2002 (triangles). The solid lines are the least-squares fits to the data in the scaling regi
The numbers indicate the scaling exponent . (b) The same as in (a) but for federal deputies.

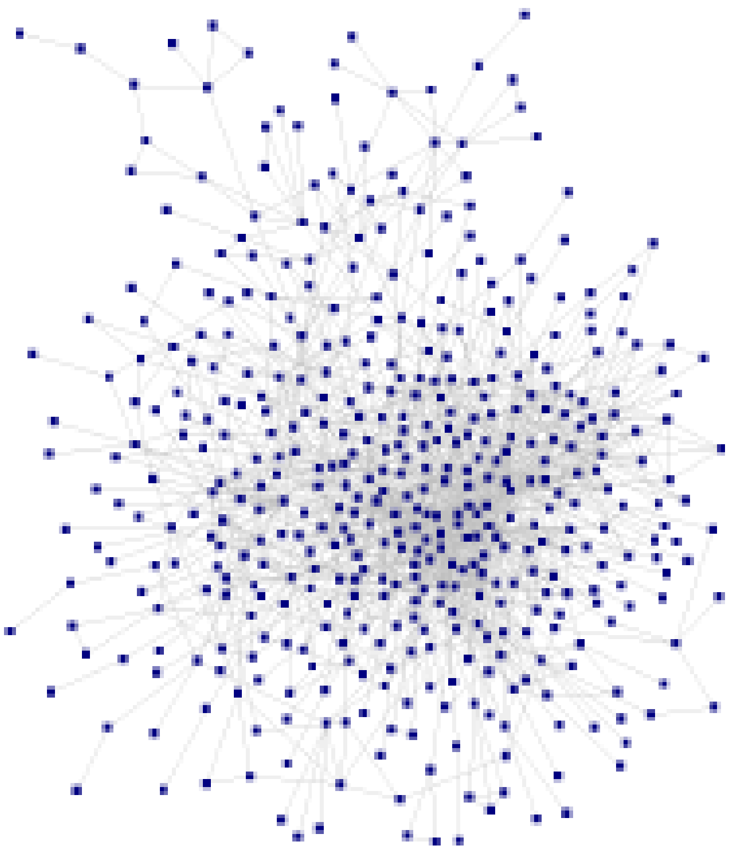


Brazilian elections: voting for a scaling democracy

R. N. Costa Filho*, M. P. Almeida, J. E. Moreira, J. S. Andrade Jr.
Departamento de Física, Universidade Federal do Ceará, Caixa Postal 6030

Scientific Collaboration Network

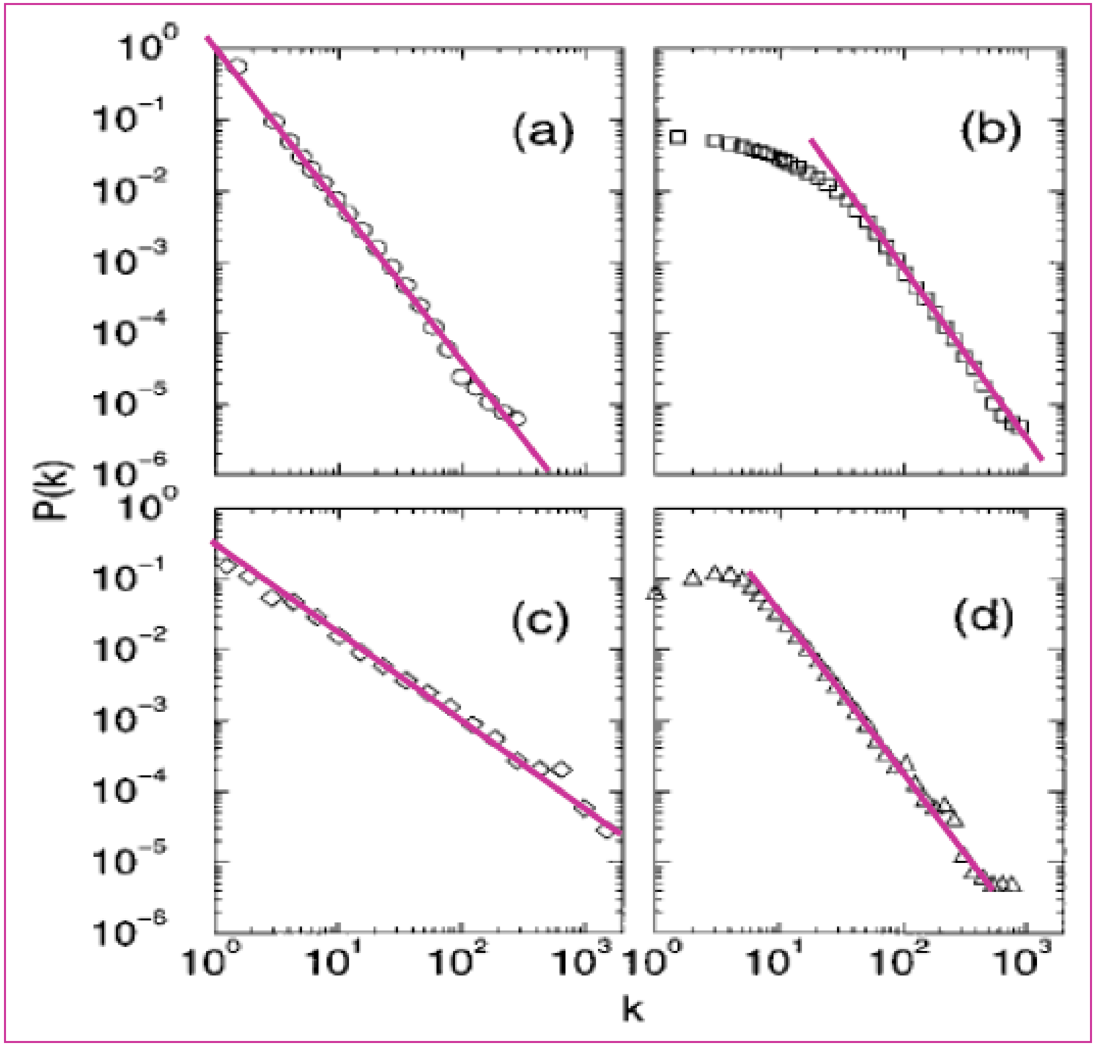
- 400,000 nodes, authors in *Mathematical Reviews* database
- An edge between two authors if they have a joint paper
- Just 676,000 edges



Picture from orgnet.com

Redes Sociais

Albert and Barabasi (1999)

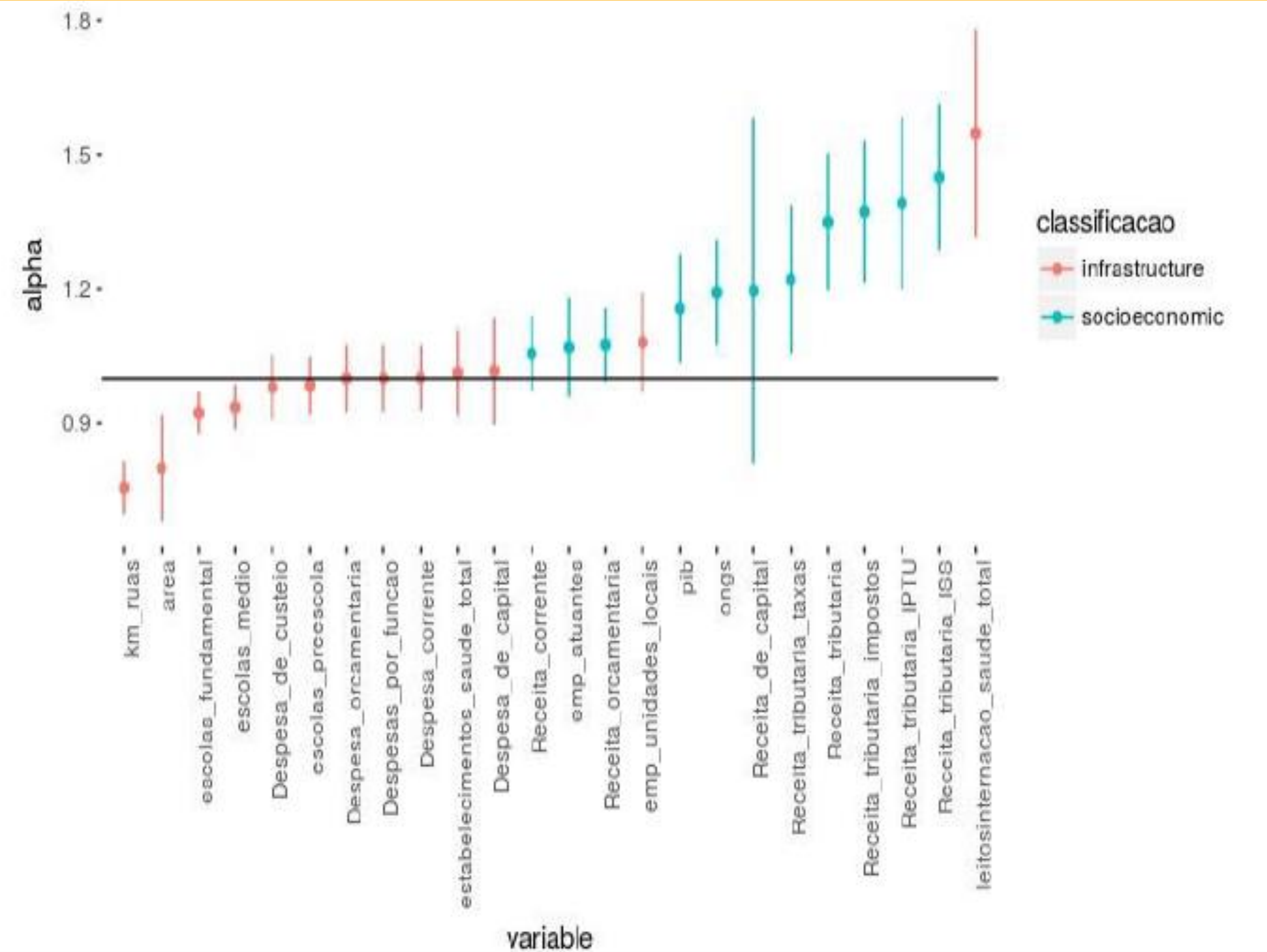


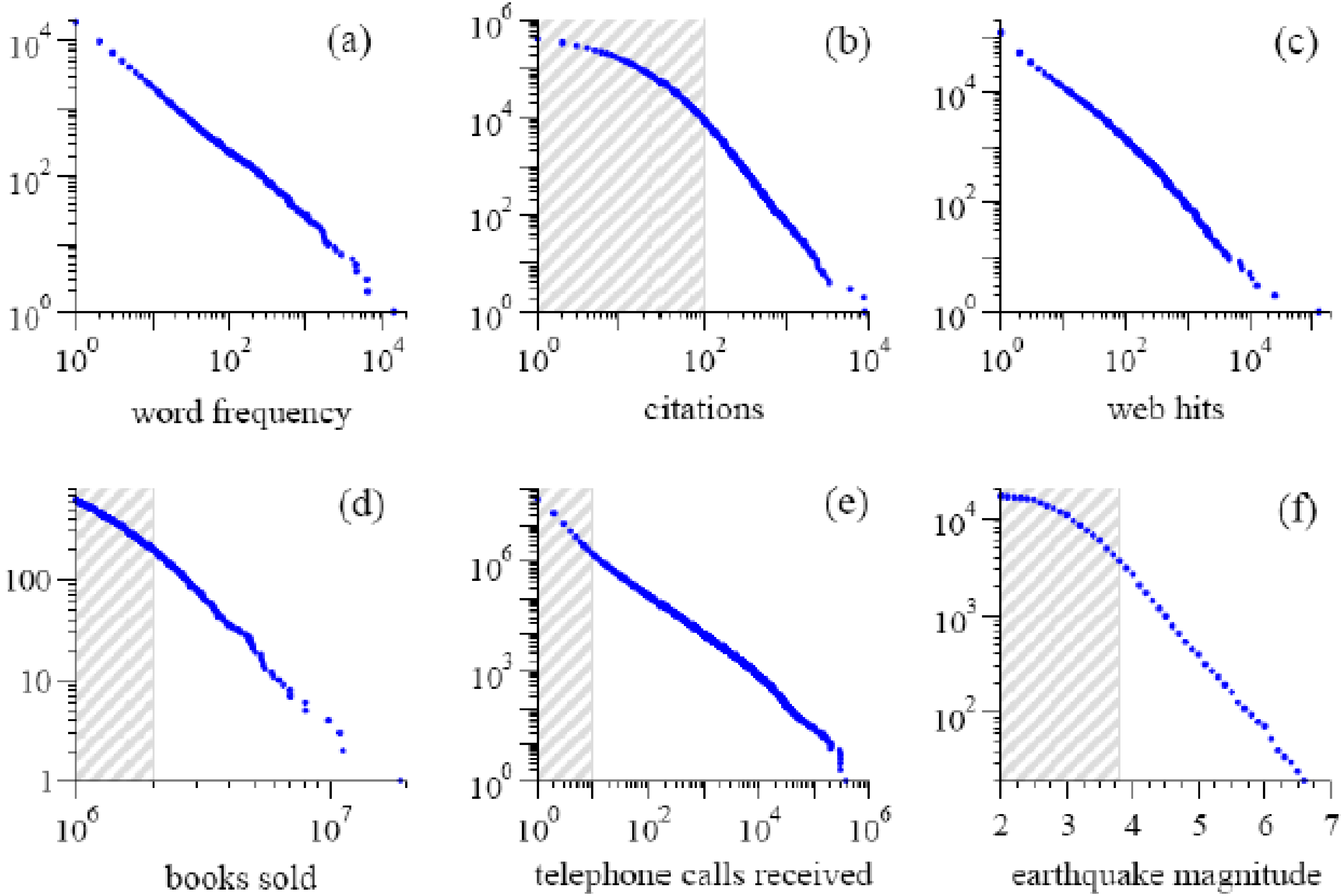
(b) co-starring in movies
(c) co-authorship of physicists
(d) co-authorship of neuroscientists

Table 1. Scaling exponents for urban indicators vs. city size

Y	β	95% CI	Adj-R^2	Observations	Country–year
New patents	1.27	[1.25,1.29]	0.72	331	U.S. 2001
Inventors	1.25	[1.22,1.27]	0.76	331	U.S. 2001
Private R&D employment	1.34	[1.29,1.39]	0.92	266	U.S. 2002
"Supercreative" employment	1.15	[1.11,1.18]	0.89	287	U.S. 2003
R&D establishments	1.19	[1.14,1.22]	0.77	287	U.S. 1997
R&D employment	1.26	[1.18,1.43]	0.93	295	China 2002
Total wages	1.12	[1.09,1.13]	0.96	361	U.S. 2002
Total bank deposits	1.08	[1.03,1.11]	0.91	267	U.S. 1996
GDP	1.15	[1.06,1.23]	0.96	295	China 2002
GDP	1.26	[1.09,1.46]	0.64	196	EU 1999–2003
GDP	1.13	[1.03,1.23]	0.94	37	Germany 2003
Total electrical consumption	1.07	[1.03,1.11]	0.88	392	Germany 2002
New AIDS cases	1.23	[1.18,1.29]	0.76	93	U.S. 2002–2003
Serious crimes	1.16	[1.11, 1.18]	0.89	287	U.S. 2003
Total housing	1.00	[0.99,1.01]	0.99	316	U.S. 1990
Total employment	1.01	[0.99,1.02]	0.98	331	U.S. 2001
Household electrical consumption	1.00	[0.94,1.06]	0.88	377	Germany 2002
Household electrical consumption	1.05	[0.89,1.22]	0.91	295	China 2002
Household water consumption	1.01	[0.89,1.11]	0.96	295	China 2002
Gasoline stations	0.77	[0.74,0.81]	0.93	318	U.S. 2001
Gasoline sales	0.79	[0.73,0.80]	0.94	318	U.S. 2001
Length of electrical cables	0.87	[0.82,0.92]	0.75	380	Germany 2002
Road surface	0.83	[0.74,0.92]	0.87	29	Germany 2002

Data sources are shown in *SI Text*. CI, confidence interval; Adj-R^2, adjusted R^2; GDP, gross domestic product

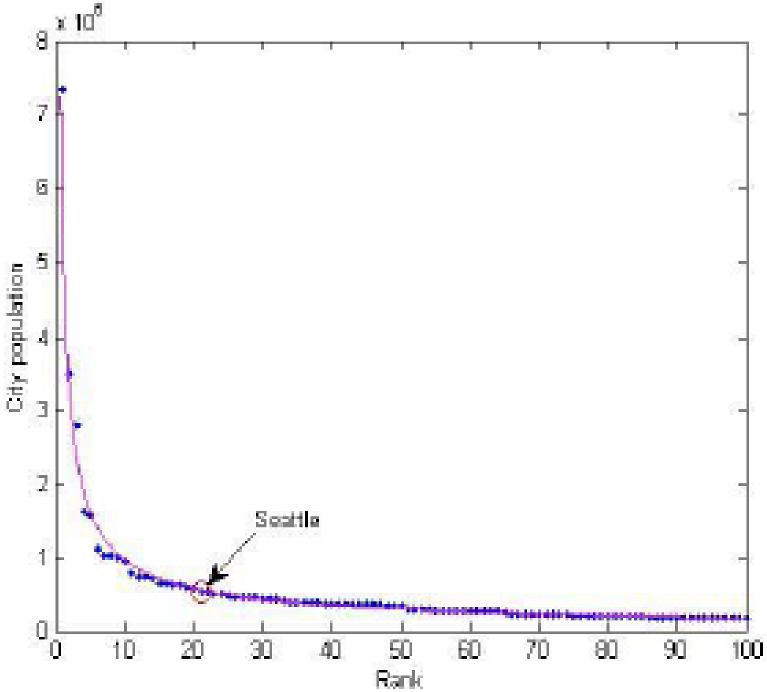

alpha
1.8
1.5
1.2
0.9
variable
km_ruas
area
escolas_fundamental
escolas_medio
Despesa_de_custeio
escolas_preescola
Despesa_orcamentaria
Despesas_por_funcao
Despesa_corrente
estabelecimentos_saude_total
Despesa_de_capital
Receita_corrente
emp_atuantes
Receita_orcamentaria
emp_unidades_locais
pib
ongs
Receita_de_capital
Receita_tributaria_taxas
Receita_tributaria
Receita_tributaria_impostos
Receita_tributaria_IPTU
Receita_tributaria_ISS
leitosinternacao_saude_total
classificacao
infrastructure
socioeconomic

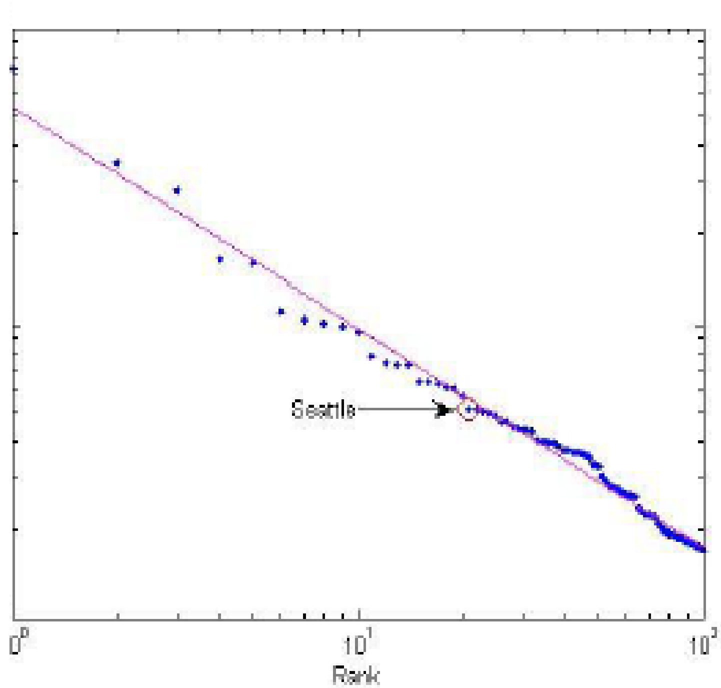


(a)
word frequency
(b)
citations
(c)
web hits
(d)
books sold
(e)
telephone calls received
(f)
earthquake magnitude

Example: City Populations

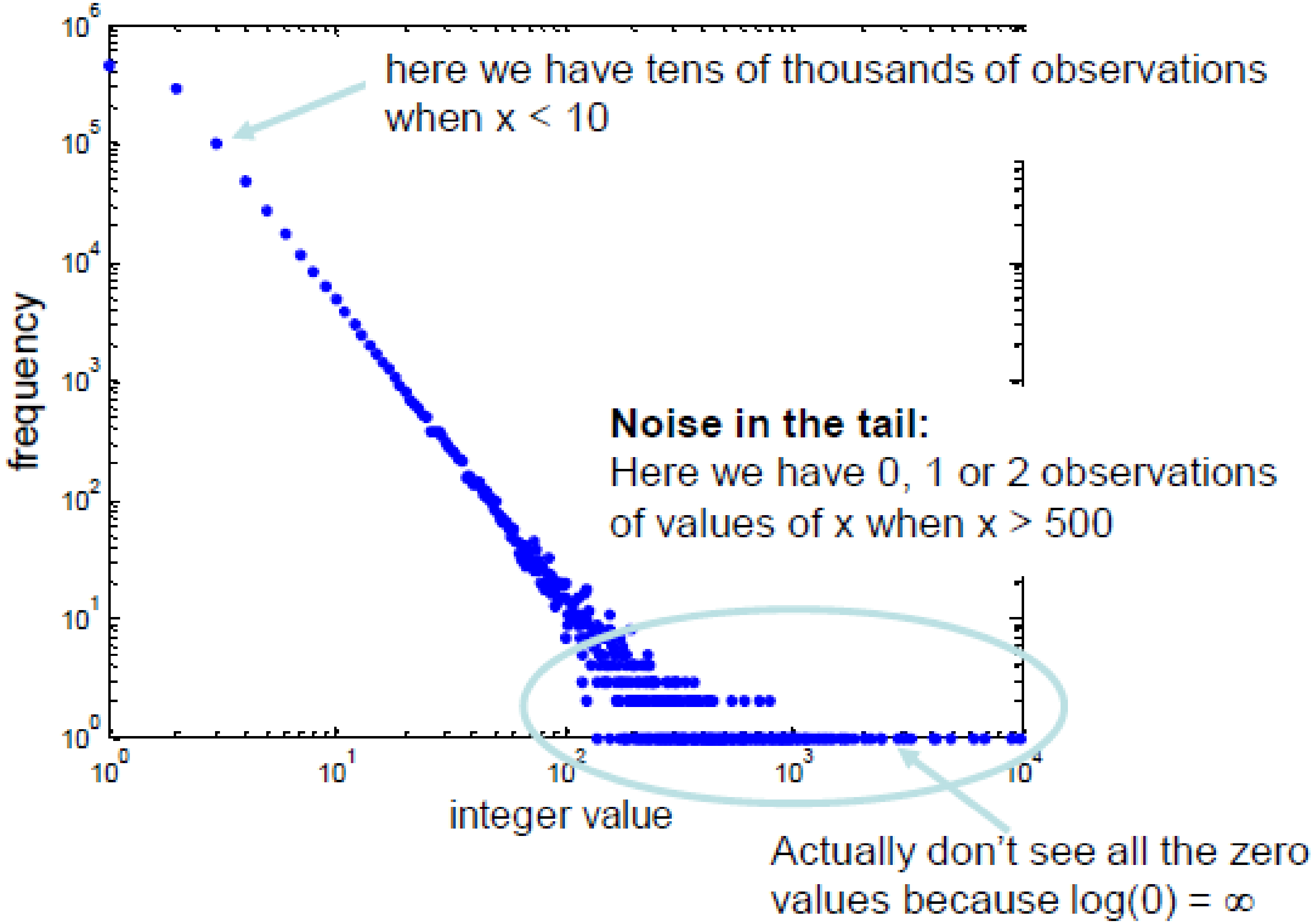
- Power law exponent: c = 0.74





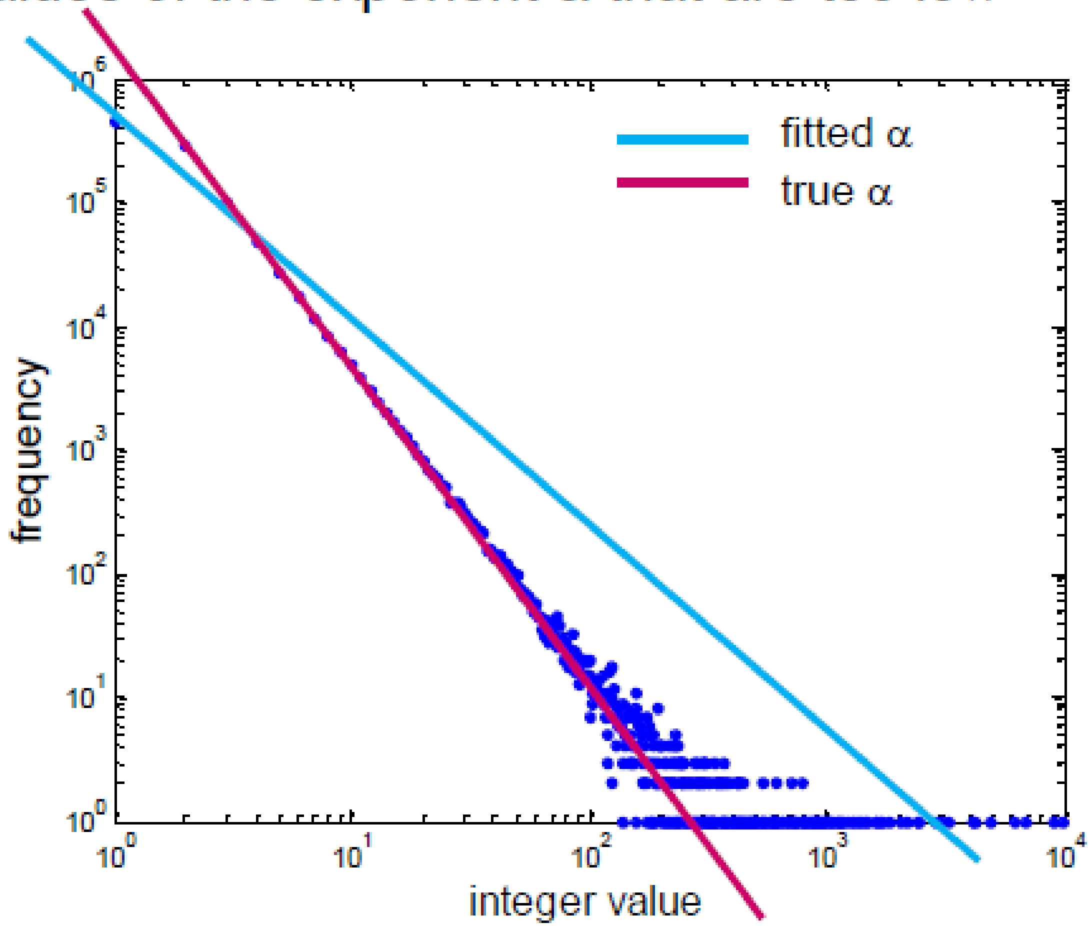
Log-log scale plot of straight binning of the data

- Same bins, but plotted on a log-log scale



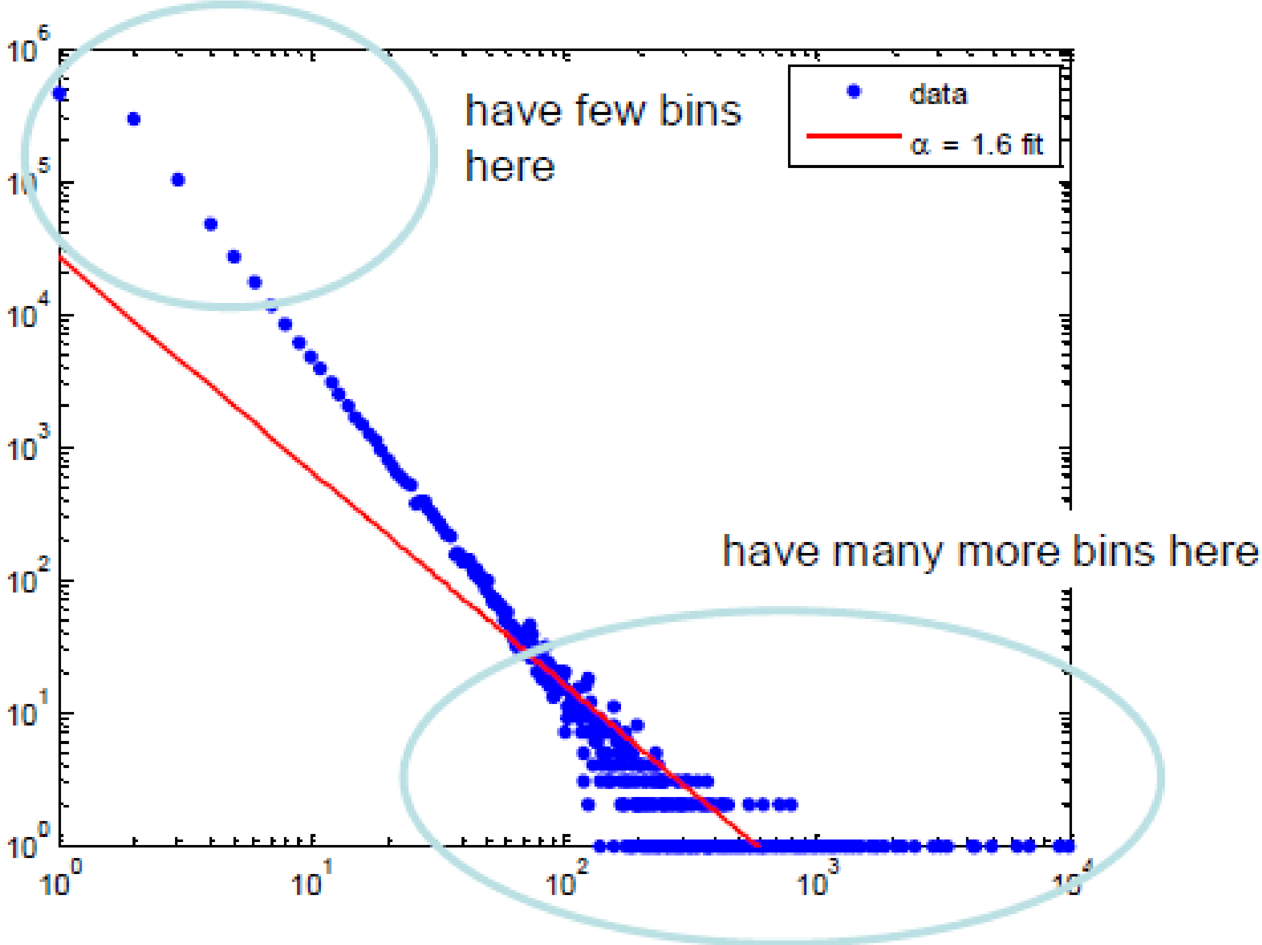
Log-log scale plot of straight binning of the data

- Fitting a straight line to it via least squares regression will give values of the exponent α that are too low



What goes wrong with straightforward binning

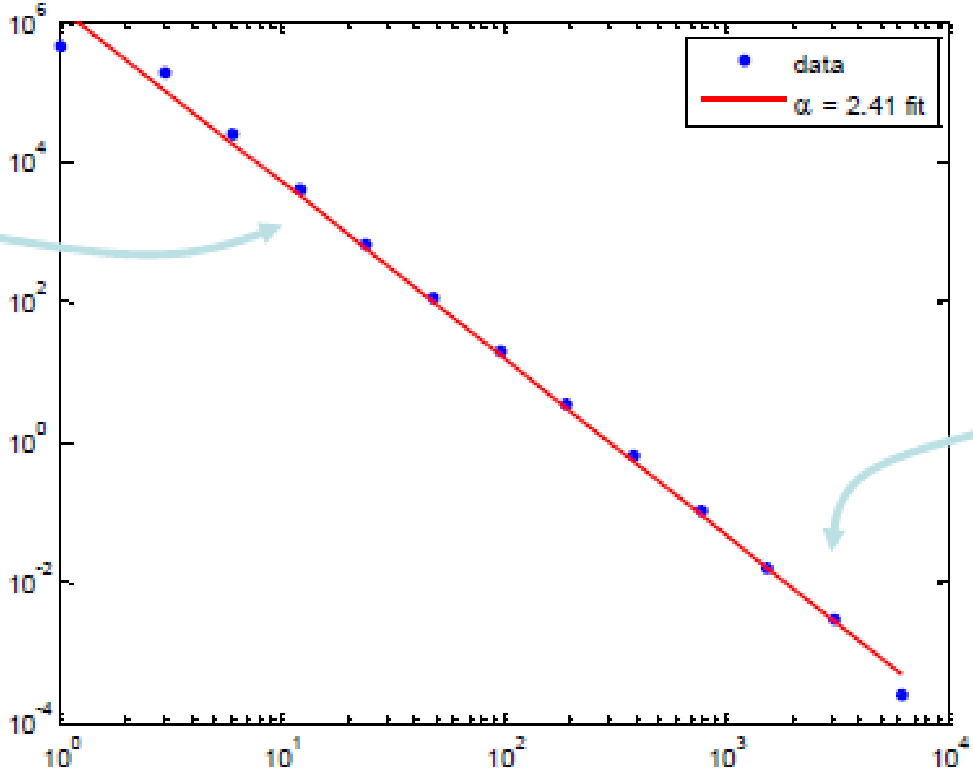
- Noise in the tail skews the regression result



First solution: logarithmic binning

- bin data into exponentially wider bins:
  - 1, 2, 4, 8, 16, 32, …
- normalize by the width of the bin

evenly spaced datapoints



less noise in the tail of the distribution

- disadvantage: binning smoothes out data but also loses information

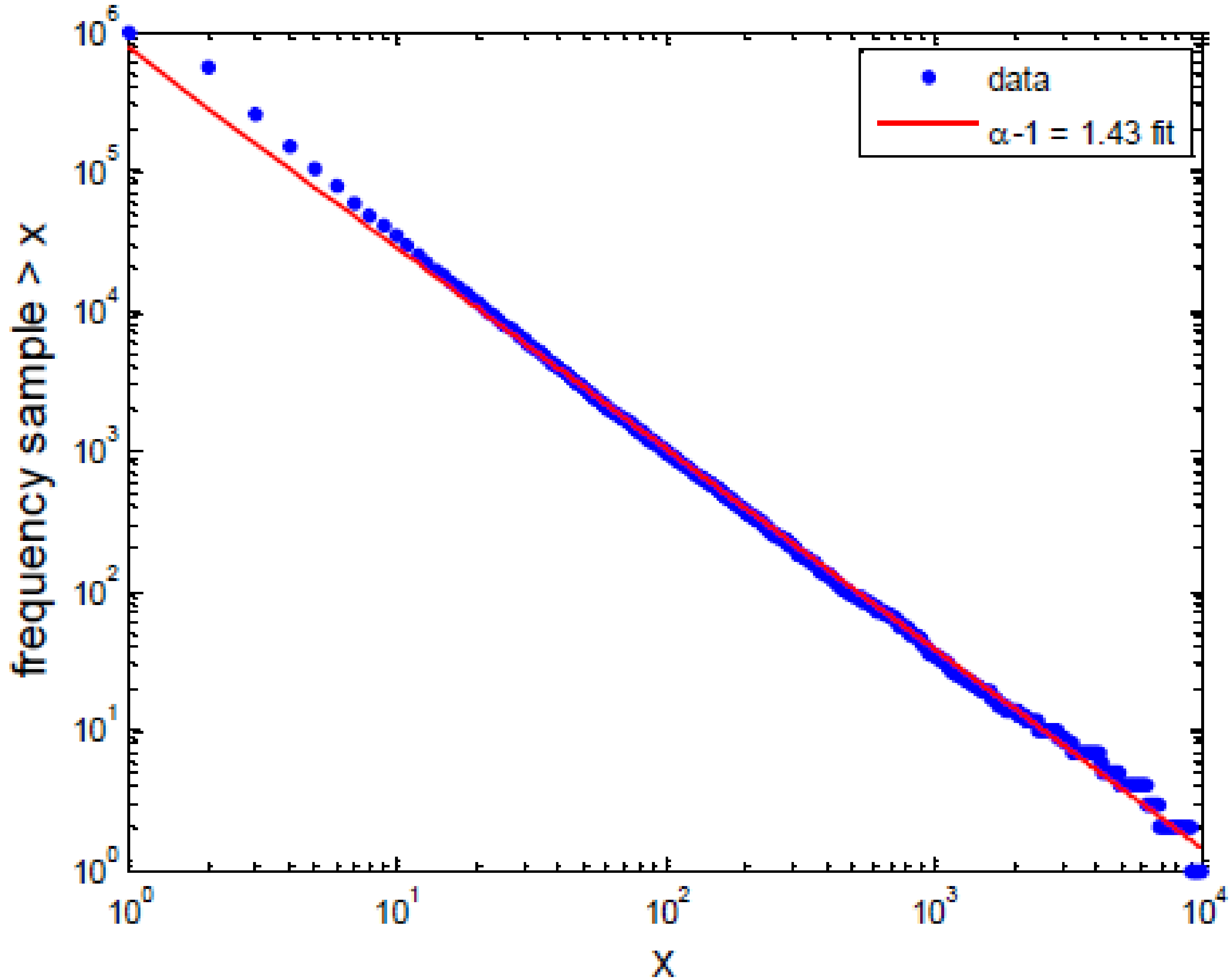
Distribuição Acumulada

$$\int cx^{-\alpha} = \frac{c}{1-\alpha} x^{-(\alpha-1)}$$

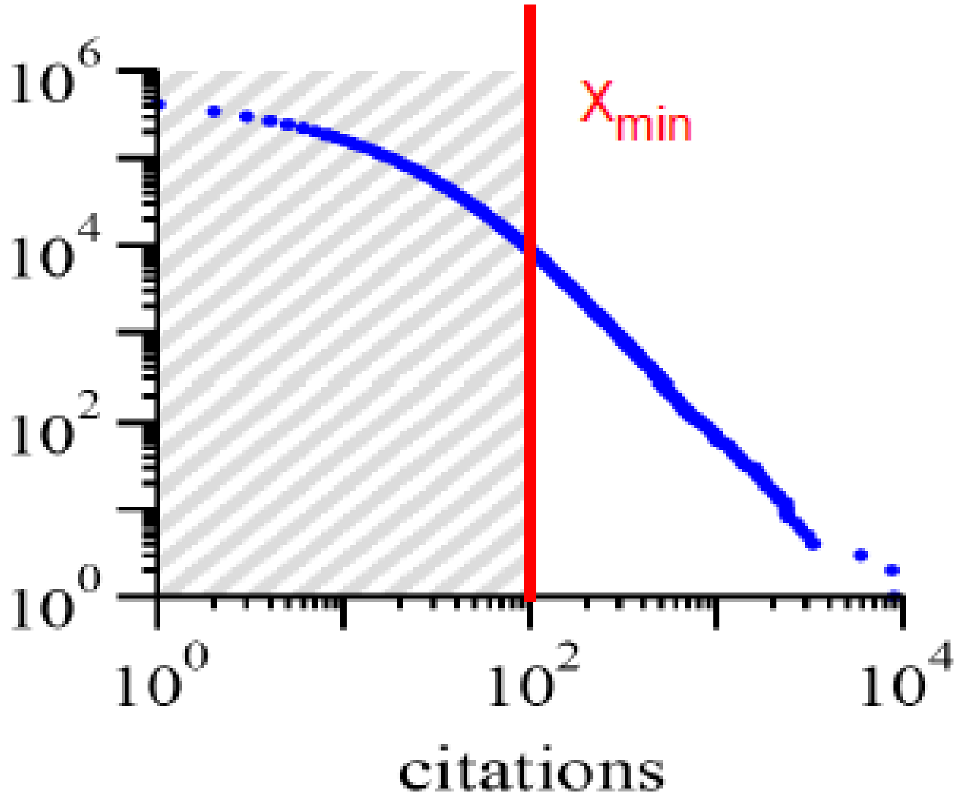
Note que o expoente agora é: $\alpha - 1$

Fitting via regression to the cumulative distribution

- fitted exponent (2.43) much closer to actual (2.5)



- Distribution of citations to papers
- power law is evident only in the tail
  - $x_{min} > 1$



Source: MEJ Newman, 'Power laws, Pareto distributions and Zipf's law

1.1 Estimating α

For the moment, let's assume we already know $x_{\min}$. In that case, we can use the method of maximum likelihood to derive an MLE for α.

$$\begin{aligned}\ln \mathcal{L}(\{x_i\} \,|\, \alpha, x_{\min}) &= \ln\left[\prod_{i=1}^{n} \frac{\alpha-1}{x_{\min}}\left(\frac{x_i}{x_{\min}}\right)^{-\alpha}\right] \\ &= n\ln\left(\frac{\alpha-1}{x_{\min}}\right) - \alpha\sum_{i=1}^{n}\ln\left(\frac{x_i}{x_{\min}}\right) \quad .\end{aligned}$$

Now solving $\partial\mathcal{L}/\partial\alpha = 0$ for α yields

$$\hat{\alpha} = 1 + n \left/ \sum_{i=1}^{n}\ln\left(\frac{x_i}{x_{\min}}\right)\right. \quad , \tag{1}$$

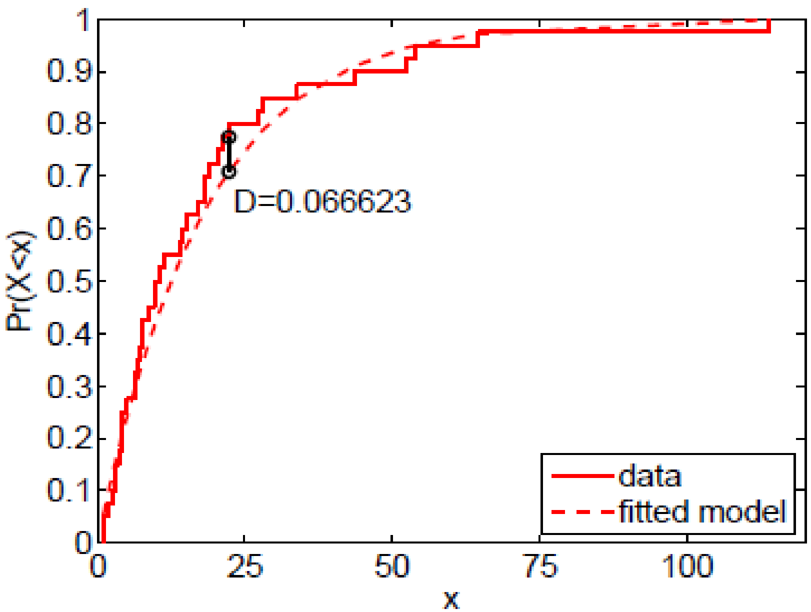
and the standard error estimate in the MLE can be shown to be $\hat{\sigma} = (\hat{\alpha}-1)/\sqrt{n}$.

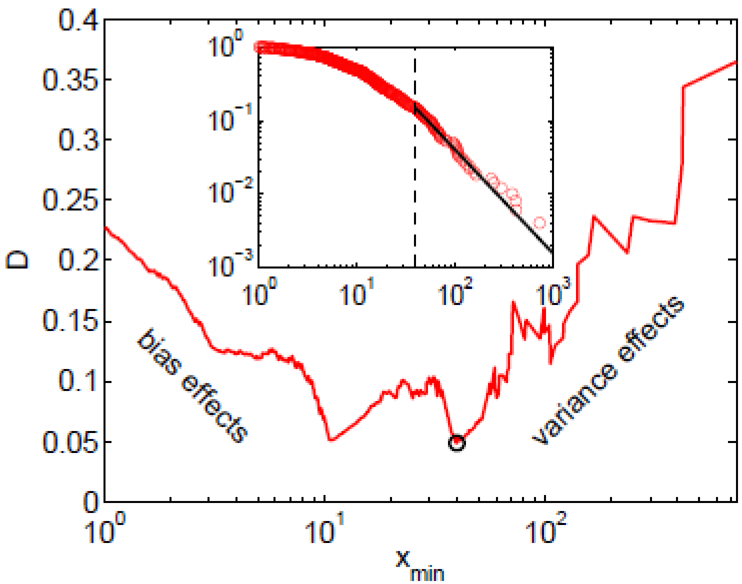
THE KOLMOGOROV-SMIRNOV STATISTIC

The KS statistic is simply the maximum distance between the CDFs of the data and the fitted model:

$$D = \max_{x \geq x_{\min}} \left| S(x) - P(x) \right|.$$

Here S(x) is the CDF of the data for the observations with value at least x_{min}, and P(x) is the CDF for the power-law model that best fits the data in the region $x \geq x_{min}$.





NORMALIZATION

The constant C is given by the normalization requirement that

$$1 = \int_{x_{\min}}^{\infty} p(x)dx = C\int_{x_{\min}}^{\infty} x^{-\alpha} dx = \frac{C}{1-\alpha}\left[x^{-\alpha+1}\right]_{x_{\min}}^{\infty}$$

This only makes sense if α > 1, since otherwise the right-hand side of the equation would diverge: power laws with exponents less than unity cannot be normalized and don't normally occur in nature. If α > 1 then $C = (\alpha-1)x_{\min}^{\alpha-1}$ and the correct normalized expression for the power law itself is

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

Some distributions follow a power law for part of their range but are cut off at high values of *x*. That is, above some value they deviate from the power law and fall off quickly towards zero. If this happens, then the distribution may be normalized no matter what the value of the exponent .

MOMENTS

The mean value of the power-law distributed quantity *x* is given by

$$\langle x \rangle = \int_{x_{\min}}^{\infty} x\,p(x)dx = C\int_{x_{\min}}^{\infty} x^{-\alpha+1}dx$$

Note that this expression becomes infinite if $\alpha \le 2$. Power laws with such low values of α have no finite mean: if we were to repeat our finite experiment many times and calculate the mean for each repetition, then the mean of those many means is itself also formally divergent, since it is simply equal to the mean we would calculate if all the repetitions were combined into one large experiment. This implies that, while the mean may take a relatively small value on any particular repetition of the experiment, it must occasionally take a huge value, in order that the overall mean diverge as the number of repetitions does. Thus there must be very large fluctuations in the value of the mean, and this is what the divergence really implies.

We can also calculate higher moments of the distribution *p(x)*.

$$\langle x^m \rangle = \frac{\alpha - 1}{\alpha - 1 - m} x_{\min}^m$$

REFERENCES

- Clauset, A., Rohilla Shalizi, C. And Newman, M.E.J. (2007). Power law distribution in empirical data, arXiv:0706.1062v1 [physics.data-an] (7 Jun 2007).
- Dutta & Horn (1981). Low-frequency fluctuations: 1/f noise, *Rev. Mod. Phys.*, 53, 497.
- Goldstein, M.L., Morris, S.A. And Yen, G.G. (2004). Problems with fitting to the power-law distribution, Eur. Phys.J.B, **41**, 255-258.
- Newman, M.E.J. (2005) Power laws, Pareto distributions and Zipf's law, Contemporary Physics, **46**, 323–351. / arXiv:cond-mat/0412004v3 [cond-mat.stat-mech] (29 May 2006).
- Pelletier, J.D. & Turcotte, D.L. (1999). Self affine time series: II. Applications and Models, in Advances in Geophysics, vol. 40, *Long-Range Persistence in Geophysical Time Series*. Academic Press, London.
- Sornette, D. (2000). Critical phenomena in Natural Sciences. Springer, Heidelberg.