Estimation of the exponent of a power-law distribution

If we have a data set $\{k_i, i=1,\ldots,n\}$ which follows a power-law distribution of exponent γ, then

$$P(k) = C\, k^{-\gamma}$$

Applying logarithms to both sides we get

$$\log(P(k)) = -\gamma \log(k) + \log(C)$$

Thus, for the estimation of the exponent γ we just need to make a linear regression of $\log(P(k))$ as a function of $\log(k)$, and take the slope of the regression line with sign changed. However, to obtain good results, we cannot make the fit over all data but over a histogram.

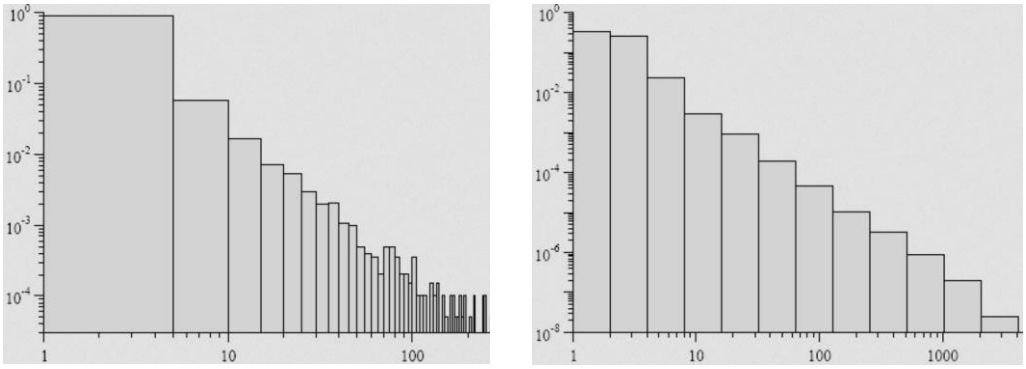

Fig 1. Wrong (left) and correct (right) histograms in log-log scale.

In doing so, the probability of the bins is not exactly proportional to $k^{-\gamma}$ but to $k^{-\gamma+1}$, see:

- E. P. White, B. J. Enquist and J. L. Green: "On estimating the exponent of power-law frequency distributions", Ecology 89(4) (2008) 905-912.

Therefore, the final procedure is the following:

1. Find $k_{min} = \min(k)$ and $k_{max} = \max(k)$
2. Calculate the logarithm of k_i for all the data elements
3. Divide the interval $[\log(k_{min}), \log(k_{max}+1)]$ in equal size bins, e.g. 10 bins, to get de values $x_0 = \log(k_{min})$, $x_1, x_2, \ldots, x_{10} = \log(k_{max}+1)$
4. Count how many elements k_i have their $\log(k_i)$ in each bin $[x_0, x_1)$, $[x_1, x_2)$, $[x_2, x_3)$,…, $[x_9, x_{10})$
5. Dividing the number of elements in each bin by the total number of elements n we get estimations for the probabilities p_b of bin $[x_{b-1}, x_b)$
6. Make the linear regression of pairs $(x_b, \log(p_b))$, to obtain the regression line of equation $y = m\,x + b$; finally, the estimation of the exponent is $\gamma = -m + 1$

Another alternative for the estimation of exponent γ consists in doing exactly the same but for the complementary cumulative distribution function (CCDF) instead of the probability density function (PDF). CCDF is calculated from the PDF just by summing up the probabilities of all the bins to the right of the bin you are considering (this one included in the sum). For example, if the probabilities (PDF) of the bins [x_0, x_1), [x_1, x_2),…, [x_9, x_{10}) are p_1, p_2, …, p_{10}, then the CCDF has values:

- [x_0, x_1) → $c_1 = p_1 + p_2 + \ldots + p_{10} = 1$
- [x_1, x_2) → $c_2 = p_2 + \ldots + p_{10}$
- …
- [x_8, x_9) → $c_9 = p_9 + p_{10}$
- [x_9, x_{10}) → $c_{10} = p_{10}$

Once we have obtained the CCDF, we make the linear regression with the values of the pairs (x_b, log(c_b)), to obtain the regression line of equation $y = m' x + b'$, and the new estimation of the exponent is $\gamma = -m' + 1$.

The quality of these two estimations of the exponents improves with the size of the data sample.

A completely different approach is derived by finding the estimator of maximum likelihood. You can find full details in the paper:

- A. Clauset, C. R. Shalizi and M. E. J. Newman: "Power-law distributions in empirical data", SIAM Rev. 51(4) (2009) 661-703.

Summarizing, if the values of k_i are discrete (integer numbers), the estimation is approximated using Eq. (3.7):

$$\gamma = 1 + n\left(\sum_{i=1}^{n} \ln \frac{k_i}{k_{min} - \frac{1}{2}}\right)^{-1}$$