Histogram in log-log scale

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

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

Applying logarithms to both sides we get

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

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

Thus, we expect a linear relationship between $\log P(k)$ and $\log(k)$, i.e., an approximate linear dependency when we plot the data distribution in log-log scale. However, to obtain good results, we cannot make the fit over all data but over a histogram, in such a way that the bins are of equal size in logarithmic scale.

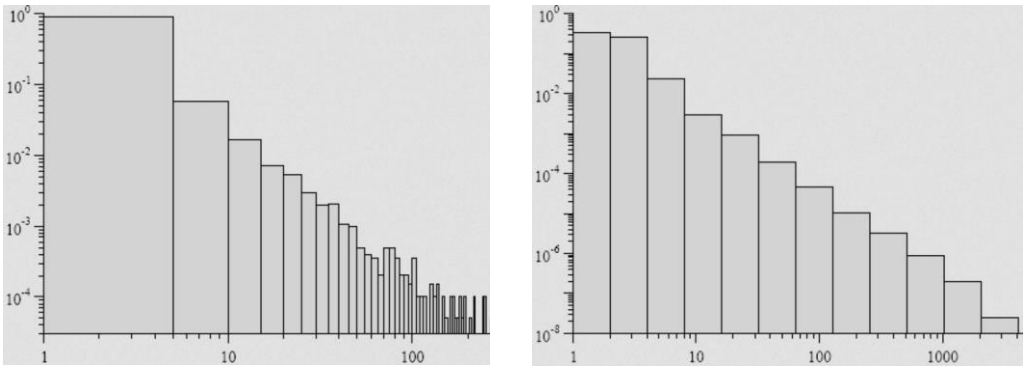

Fig 1. Wrong (left) and correct (right) histograms in log-log scale.

The procedure to generate a correct histogram of the probability density (PDF) in log-log scale 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 $[\log(k_{\min}), \log(k_{\max})]$ in equal size bins, e.g., 10 bins, to get de values $x_0 = \log(k_{\min}), x_1 x_2 \dots, x_{10} = \log(k_{\max} + 1)$
4. Count how many elements k_i have their $\log(k_i)$ in each bin $[x_0, x_1)$, $[x_1, x_2), [x_2, x_3), \dots, [x_9, x_{10})$
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 $[x_b, x_{b+1})$

Similarly, if we are interested in the complementary cumulative distribution function (CCDF), it 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 $[x_0, x_1)$, $[x_1, x_2), \dots, [x_9, x_{10})$ are $p_1, p_2, \dots, p_{10}$, then the CCDF has values:

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

For more information on histograms and the binning process, check:

- http://www.mkivela.com/binning_tutorial.html