S1: Fitting power-laws in empirical data with estimators that work for all exponents

Rudolf Hanel¹, Bernat Corominas-Murtra¹, Bo Liu¹, Stefan Thurner¹,²,³,⁴

¹ Section for Science of Complex Systems, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria
² Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA
³ IIASA, Schlossplatz 1, 2361 Laxenburg, Austria
⁴ Complexity Science Hub Vienna, Josefstadterstrasse 39, A-1090 Vienna, Austria

* Current Address: Section for Science of Complex Systems, CeMSIIS, Medical University of Vienna, Spitalgasse 23, Bauteil 86, A-1090, Vienna, Austria
* stefan.thurner@meduniwien.ac.at

APPENDIX A: Sampling from continuous sample spaces

If events \(x\) are drawn from a continuous sample space \(\Omega = [x_{\text{min}}, x_{\text{max}}]\), for instance the magnitude of earthquakes, then the ‘natural order’ of possible events is simply given by the magnitude \(x\) of the observation. Events \(x\) are drawn from a continuous power-law distribution

\[ p(z|\lambda, \Omega) = \frac{z^{-\lambda}}{Z} \]

with

\[ Z = Z_{\lambda}([x_{\text{min}}, x_{\text{max}}]) \]

(compare Eq. (main 3) first line).

To work with well defined probabilities we have to bin the data first. Probabilities to observe events within a particular bin depend on the margins of the \(W\) bins

\[ b = (b_0, b_1, \cdots, b_W), \]

with \(b_0 = x_{\text{min}}\) and \(b_W = x_{\text{max}}\). The histogram \(k = (k_1, \cdots, k_W)\) counts the number \(k_i\) of events \(x\) falling into the bin \(b_i > x \geq b_{i-1}\), and the probability of observing \(x\) in the \(i\)'th bin is given by

\[ p(i|\lambda, x) = \frac{b_i^{1-\lambda} - b_{i-1}^{1-\lambda}}{x_{\text{max}}^{1-\lambda} - x_{\text{min}}^{1-\lambda}}. \tag{1} \]

Binning events sampled from a continuous distribution may have practical reasons. For instance data may be collected from measurements with different physical resolution levels, so that binning should be performed at the lowest resolution of data points included in the collection of samples. We will not discuss the ML estimator for binned data in detail but only remark that for given bin margins \(b\) it is sufficient to insert \(p(i|\lambda, x)\) of Eq (1) into Eq. (main 7) with \(\theta = \{\lambda\}\), to derive the appropriate ML condition for binned data. An algorithm for binned data \texttt{r_plhistfit}, where we assume the bin margins \(b_i\) to be given, is found in [2].

We point out that if margins for binning have not been specified prior to the experiments, then specifying the optimal margins for binning the data becomes a parameter estimation problem in itself, i.e. the optimal margins \(b_i\) have to be estimated from the data as well. One major source of uncertainty in the estimates of \(\lambda\) from binned data is related to the uncertainty in choosing the upper and lower bounds \(x_{\text{min}}\) and \(x_{\text{max}}\) of the data, i.e. specifying the bounds of the underlying continuous sample space.

Binning becomes irrelevant for clean continuous data for the following reason. Suppose we fix the sample space \([x_{\text{min}}, x_{\text{max}}]\) and cut this domain into \(M\) bins of width...
\[ \Delta = (x_{\text{max}} - x_{\text{min}})/M. \] Since the data \( x = \{x_1, \ldots, x_N\} \) is drawn from a continuous sample space, the chance for two observations \( x_m \) and \( x_n \) to be exactly equal becomes zero for \( m \neq n \), if \( M \) has been chosen sufficiently large. Then each bin almost certainly contains either one sample \( x_n \) or none. The probability of observing \( x \) then is asymptotically (as \( \Delta \) approaches zero) given by

\[
P(x|\lambda) = \Delta^N \prod_{n=1}^{N} \left( \frac{x_n^{-\lambda}}{Z(x_{\text{min}}, x_{\text{max}})} \right).
\]

The parameter estimation problem of finding the optimal \( \lambda \) becomes possible to derive Bayesian estimators for \( x \) for instance, one asks how likely can the maximal value \( \text{max}(x) \) of the sampled data \( x = (x_1, \ldots, x_N) \) be found to be larger than some value \( y \). By deriving \( P(\text{max}(x) > y|\lambda, [x_{\text{min}}, x_{\text{max}}]) \) and \( P(\text{min}(x) < y|\lambda, [x_{\text{min}}, x_{\text{max}}]) \), as a consequence, it becomes possible to derive Bayesian estimators for \( x_{\text{min}} \) and \( x_{\text{max}} \).
References


2. http://www.complex-systems.meduniwien.ac.at/SI2016/r_plfit.m
   http://www.complex-systems.meduniwien.ac.at/SI2016/r_plhistfit.m
   http://www.complex-systems.meduniwien.ac.at/SI2016/r_randi.m
   http://www.complex-systems.meduniwien.ac.at/SI2016/r_plfit_calibrate.m
   http://www.complex-systems.meduniwien.ac.at/SI2016/r_plfit_calib_eval.m
   Alternatively, see also S4 File Appendix D for the code.