Power-Law Distributions and Binned Empirical data

by

Yogesh S. Virkar

B.E., University of Mumbai, 2010

M.S., University of Colorado, Boulder, 2012

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Masters of Science

Department of Computer Science

2012

This thesis entitled:
Power-Law Distributions and Binned Empirical data
written by Yogesh S. Virkar
has been approved for the Department of Computer Science

Aaron Clauset

Michael Mozer

Vanja Dukic

Date __________________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Virkar, Yogesh S. (M.S., Computer Science)

Power-Law Distributions and Binned Empirical data

Thesis directed by Professor Aaron Clauset

Many man-made and natural phenomenon, including the intensity of earthquakes, population of cities, and sizes of wars, are believed to follow power-law distributions, and the detection of these patterns has significant consequences for our understanding of the underlying mechanisms. However, the large fluctuations in the tail of these distributions makes it difficult to provide clear evidence for or against the power-law hypothesis, particularly when the empirical data have been binned. Clauset, Shalizi and Newman recently provided a statistically principled framework for identifying and testing power-law distributions in continuous or discrete valued data, based on maximum-likelihood fitting, goodness-of-fit test based on the Kolmogorov-Smirnov (KS) statistic and likelihood ratios for model comparison. We adapt these techniques to the less common but important case of binned empirical data. We evaluate the effectiveness of our techniques on synthetic data with known structure and apply them to ten real-world data sets with heavy-tailed patterns.

Acknowledgements

Its my pleasure to thank the many people who made this thesis possible.

Firstly, I would like to thank my advisor, Professor Aaron Clauset. Without his guidance, advice and support, this project would have been impossible. I also thank him for being patient with me and making mathematics a fun subject.

I would also like to thank my committee, Professor Michael Mozer and Professor Vanja Dukic for taking time from their busy schedules and being a part of my thesis committee. Their suggestions and comments helped improve this work. I also thank Professor Elizabeth Bradley for her comments on this work.

I am grateful to my friends, my colleagues and my lab group for their invaluable comments and observations which improved my defense presentation.

Lastly, and most importantly, I would like to thank my grandmother, Lata Virkar and my parents, Shrikant Virkar and Manda Virkar, who raised me, taught me, loved me and supported me, and my brother Siddhesh Virkar, who has given me invaluable advice throughout my life. To them I dedicate this thesis.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

Power-law distributions occur in a wide range of natural and man-made phenomena such as physics, chemistry, biology, computer science and economics [12, 15, 20, 9]. Power-laws are of scientific interest due to their interesting mathematical properties [22] and their identification can indicate unusual underlying processes such as self organized criticality [2] or highly optimized tolerance [15], or non-trivial sampling effects [18].

A power-law distribution is characterized by its heavy-tailed nature which unlike the Normal distribution places considerable weight on extremely large values. In this sense, power laws are not well characterized by a typical or average value. For instance, the 2000 U.S. Census indicates that the average population of a city, town or village in the United States is 8226, but this value gives no indication that a significant fraction of the U.S. population lives in cities like New York and Los Angeles, whose populations are roughly three orders of magnitude larger than the average.

Mathematically, when the probability of measuring a quantity of interest x varies inversely as some power of x, the quantity x is said to follow a power-law distribution,

$$\begin{aligned} \mathrm{p}(x) &\propto x^{-\alpha} \quad \text{, i.e.,} \\ \mathrm{p}(x) &= Cx^{-\alpha} \qquad (1.1) \end{aligned}$$

where x is the quantity of interest, and $p(x)$ is the probability density function of x. C denotes the normalization constant. The constant α is known as the exponent or scaling parameter whose typical value lies between $2 < \alpha < 3$. In practice, power-law behavior is not seen for all values of

x. Usually this behavior holds true for values above certain lower bound, say $x_{\min}$, in which case the **tail** of the distribution is said to follow a power law. If we consider continuous valued data, we can derive the probability density for the power law as follows (details are given in Chapter 2),

$$\mathrm{p}(x) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-\alpha} \tag{1.2}$$

A power law has many interesting mathematical properties most of which are due to its heavy tailed nature or right-skewness. We can derive the k^{th} moment for a power-law distribution with lower-bound $x_{\min}$ as,

$$\begin{aligned} \langle x^k \rangle &= \int_{x_{\min}}^{\infty} x^k p(x)\,\mathrm{d}x \\ &= (\alpha - 1) {x_{\min}}^{(\alpha-1)} \int_{x_{\min}}^{\infty} x^{(k-\alpha)}\,\mathrm{d}x \\ &= {x_{\min}}^{k} \frac{\alpha - 1}{\alpha - 1 - k} \qquad (\alpha - 1) > k \end{aligned} \tag{1.3}$$

This implies that for $1 < \alpha < 2$, the first moment, i.e, the mean of the distribution, is infinite and so are all higher moments. With $2 < \alpha < 3$, the first moment (mean) is finite but the second moment (variance) and all higher moments become infinite. An infinite moment implies that with increasing sample size, the estimate of that moment increases.

Another property of power laws that makes them very interesting, is scale-invariance. If we multiply x by a constant factor, say f, such that the probability density for the power law changes from $p(x)$ to $p(fx)$, then

$$\begin{aligned} \mathrm{p}(fx) &= \frac{\alpha - 1}{x_{\min}} \left(\frac{fx}{x_{\min}} \right)^{-\alpha} \\ &= f^{-\alpha} \left[\frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-\alpha} \right] \\ &= f^{-\alpha} \mathrm{p}(x) \\ \mathrm{p}(fx) &\propto \mathrm{p}(x) \end{aligned} \tag{1.4}$$

It can be shown that amongst all the heavy-tailed distributions, a power law is the only distribution with this property. Taking logarithms of both sides of Equation (1.1) yields,

$$\ln \mathrm{p}(x) = \ln C - \alpha \ln x$$

implying that a true power law distribution appears as a straight line on log-log axes. If we change scale from $x \to fx$,

$$\begin{aligned}\ln \mathrm{p}(x) &= \ln C - \alpha \ln fx \\ &= \ln C - \alpha \ln x - \alpha \ln f\end{aligned}$$

we simply shift the power law up or down along the y-axis by a constant, on a logarithmic scale.

For scientific interpretations, scale invariance implies that large and small events are not qualitatively different. Thus, identifying a power-law pattern can provide important clues about a system's underlying mechanisms, and can facilitate statistical extrapolations about the likelihood of very large events [6]. Reliably identifying such patterns becomes extremely important as there are large fluctuations in the distribution's upper tail, where we wish to have the most accuracy. Also, a straight line behavior on a doubly logarithmic axes, is necessary but not sufficient condition for a power law hypothesis. For these reasons, it is difficult to distinguish a power law from other heavy-tailed distributions such as stretched-exponential or the log-normal, as such heavy-tailed distributions also appear fairly straight on a log-log axes for certain parameter settings and small sample sizes.

Recently, Clauset, Shalizi and Newman [5] introduced a statistically principled framework for fitting and testing the power-law hypothesis with empirical data. Their approach combines maximum-likelihood techniques for fitting a power-law distribution to the empirical data's upper tail, a distance-based method [19] for automatically identifying the point at which the power-law behavior begins [8], a goodness-of-fit test based on the Kolmogorov-Smirnov (KS) statistic for testing the plausibility, and a likelihood ratio test for comparing the power-law model to alternative heavy-tailed distributions [23]. These techniques can only be applied to continuous or discrete-valued empirical data.

The goal of this thesis is to adapt their framework to the important case of binned empirical data, i.e, when instead of observing raw values, we see only counts of values within a set of given ranges. The reasons for binning the data could be diverse. For example, measurements may be done roughly because direct measurement is impossible or impractical. Alternatively, the empirical data may have been recovered from an existing histogram, in which values were binned for ease of presentation, and the original values lost, or the data could come to us already binned for other reasons. In any such case, we would like to be able to make strong statistical inferences despite the binning, and this requires specialized tools.

In broad outline, the adapted framework proposed for analysis of binned empirical data can be summarized as follows-

(1) Estimate the parameters $x_{\min}$ and α of the power-law model using the methods described in Chapter 3.

(2) Calculate the goodness-of-fit between the data and the fitted model using the methods described in Chapter 4. If the resulting p-value is greater than 0.1, the power law is a plausible hypothesis for the data, otherwise it is rejected.

(3) Compare the power law to alternative heavy-tailed distributions via a likelihood ratio test, as described in Chapter 5. For each alternative, if the normalized likelihood ratio statistic is significantly away from zero, then its sign indicates whether or not the alternative is favored over the power-law model.

The effectiveness of this framework is tested for correctness using synthetic data with known structure and gives accurate results for sufficiently large sample sizes. These methods make no assumptions about the type of binning scheme used, and can thus be applied to linear, logarithmic or arbitrary bins. However, binning itself induces a loss of information that increases the uncertainty in parameter estimates and decreases the statistical power of hypothesis tests and model comparisons. To quantify the size of this information loss, quantitative measures are derived and their implications for the collection of binned data are discussed.

To demonstrate the utility of these methods, we apply them to analyze the structure of ten real-world data sets, all of which exhibit heavy-tailed, possibly power-law behavior. We can conclude that several cannot reasonably be considered to follow power laws.

A further motivation for using specialized tools to handle such data is given in Appendix A. In this section, an attempt is made to make binned data look like raw-valued data such that we can leverage existing tools for analysis. In this way we adapt data to suite methods instead of adapting methods to suite data. However this is shown not to work, further supporting the need for using our adapted framework for binned data.

Chapter 2

Binned Power-Law Distributions

Before adapting the methods given by Clauset et.al. in [5] to the case of binned data, we need to define what we mean by a power-law distribution with binned data.

Conventionally, a power-law distributed quantity can be either continuous or discrete. For continuous values, the probability density function for a power-law distribution can be defined as

$$\begin{aligned} p(x)dx &= \Pr(x \leq X < x + dx) \\ &= Cx^{-\alpha}dx \end{aligned} \tag{2.1}$$

where X is the observed value and C is the normalization constant. Clearly this density diverges as $x \to 0$ and so Equation (2.1) cannot hold for all $x \geq 0$; there must be some lower-bound $x_{\min}$ to the power-law behavior. In this case, so long as $\alpha > 1$, it is easy to calculate the normalizing constant by using definition of probability density, i.e,

$$\int_{x_{\min}}^{\infty} Cx^{-\alpha} = 1$$

Solving the above equation gives $C = \dfrac{\alpha - 1}{{x_{\min}}^{1-\alpha}}$ and the complete probability density function can be given as,

$$\Pr(x) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-\alpha} \tag{2.2}$$

When real continuous or integer discrete values are binned, we see only counts of values within a set of given ranges and this changes the nature of the data. After binning, data are composed of two pieces: the counts of values H and the given ranges or bin boundaries B. For concreteness, let k be the number of bins. The bin boundaries are then,

$$B = (b_1, b_2, ..., b_k, b_{k+1}) \tag{2.3}$$

where the i^{th} bin includes values $x \in [b_i, b_{i+1})$. Also, if any bin boundary, say b_i, denotes a lower bin boundary for a particular bin, then it also denotes upper bin boundary for the previous bin, except when b_i is the first or the last bin boundary.

$$H = (h_1, h_2, ..., h_k) \tag{2.4}$$

that is, h_i counts the number of raw observations $x \in [b_i, b_{i+1})$.

To write down the probability density function for such data, we need the probability that an observation falls in a particular bin. This is given by the fraction of the full distribution in the i^{th} bin:

$$\begin{aligned} \Pr(b_i < x < b_{i+1}) &= \int_{b_i}^{b_{i+1}} p(x)\, dx \\ &= {x_{\min}}^{\alpha-1}[{b_i}^{(1-\alpha)} - {b_{i+1}}^{(1-\alpha)}] \end{aligned} \tag{2.5}$$

Equation (2.5) shows that for binned power-law distributions, there are three parameters: $\alpha, x_{\min}$ and the bin boundaries $b_i, b_{i+1} \in B$. Here, we assume that the binning scheme B is given to us, as otherwise we would have access to the raw data and we could apply directly the methods of Clauset et.al.

Given B, fitting the power-law model, requires estimating α and $x_{\min}$. However, the binning process destroys information about the value $x_{\min}$, and so our estimate $\hat{x}_{\min}$ is restricted to the bin boundaries, i.e., $\hat{x}_{\min} \in B$. For this reason, we make the notational choice of $b_{\min}$, that denotes the smallest bin boundary for which the power-law behavior holds, and $\hat{b}_{\min}$ denotes its estimate such that $\hat{b}_{\min} \in B$.

Chapter 3

Fitting Power-Law to Binned Empirical data

One of the main goals of this thesis is to show correct fitting of the power-law form to the empirical data that are binned. Traditional approaches for fitting a power-law model, include fitting regression lines to the PDF or the complementary CDF of the empirical data on a log scale. If we take logarithm on both sides of Equation (2.1), we get

$$\ln p(x) = \ln C - \alpha \ln x \tag{3.1}$$

which means that visually a power law looks like a straight line on a log-scale. Hence least-squares linear regression could be used to recover the slope of this straight line and we could claim this slope to be the estimate of scaling parameter α. However, one of the most important assumptions of linear regression is Gaussian noise at every value of independent variable, in this case x. On a log scale this assumption does not hold, which is why linear regression does poorly and should be avoided. (See [5] for detailed analysis.) This section deals with adapting the MLE (method of maximum likelihood) for the case of binned empirical data. The correctness of these adapted methods is tested using synthetically generated binned data and comparison to the linear regression techniques is made.

3.1 Estimating the scaling parameter

Let us consider the estimation of the scaling parameter α for binned data sets. Estimating α correctly requires, as we will see, a value for the lower bound $b_{\min}$ of power-law behavior in the data. We will assume for now that this lower-bound, i.e, $b_{\min}$ is known. In cases where it is not known, we can estimate it from the data using the methods given in section 3.3.

To fit parameterized models such as the power law to the observed data we use the method of maximum likelihood, which has provable accuracy for parameter estimates in the limit of large sample sizes [24]. Assuming that our data are drawn from a distribution that follows a power law exactly above the bin boundary $b_{\min}$, we can derive the maximum likelihood estimator (MLE) of the scaling parameter for the binned case. Details of the derivation are given in Appendix B.

The MLE for the binned case can be calculated by numerically maximizing the following equation,

$$\mathcal{L} = n(\alpha - 1)\ln(b_{\min}) + \sum_{i=\min}^{k} h_i \ln[{b_i}^{(1-\alpha)} - {b_{i+1}}^{(1-\alpha)}] \tag{3.2}$$

where h_i is the count of values in bin i, i.e, $h_i \in H$, and the bin boundaries $b_i, b_{i+1} \in B$ bound the i^{th} bin. The symbol n denotes the summation of counts of values of bins above $b_{\min}$, i.e, $n = \sum\limits_{i=min}^{k} h_i$. We use the symbol N to denote the total sample size, i.e, $N = \sum\limits_{i=1}^{k} h_i$.

For logarithmic binning scheme, in which the bin boundaries are in powers of some constant c, i.e, $B = (1, c, c^2, c^3, ...)$, we can obtain an analytic solution for α by solving $\partial\mathcal{L}/\partial\alpha = 0$, and we get estimate of alpha as,

$$\hat{\alpha} = 1 + \log_c \left[1 - \frac{n\ln(c)}{\left(n - \sum\limits_{i=\min}^{k} ih_i\right)\ln(c) + n\ln(b_{\min})} \right] \tag{3.3}$$

It can be shown that the estimator $\hat{\alpha}$ is asymptotically consistent,, i.e, in the limit of large sample size $n \to \infty$, $\hat{\alpha} \to \alpha$. (See Appendix B and Figure B.1). The standard deviation in estimation of $\hat{\alpha}$ is given as,

$$\sigma = \frac{(c - c^{\alpha})}{n\sqrt{c^{1+\alpha}}[\ln(c)]} \tag{3.4}$$

Note that the above value for σ becomes positively biased with very small values of n (for $c = 2$, $n \lesssim 50$).

The c used in the logarithmic binning scheme can have a lot of impact on the accuracy of our results. Intuitively, a higher value of c, i.e., greater bin widths, would imply more information loss and hence lesser accuracy for $\hat{\alpha}$. It can be shown that, if one uses a binning scheme with $c = 10$, instead of a binning scheme with $c = 2$, then one would need to almost triple the sample size n, in order to have the same accuracy of $\hat{\alpha}$ (See Appendix D.1). Thus we need to make the bin boundaries as tight as possible, so that the information loss is minimal.

3.2 Performance of scaling parameter

To demonstrate that the estimator found by numerically maximizing Equation (3.2) gives us the correct parameter estimate for $\hat{\alpha}$, experiments are done on synthetically generated power-law data. This implies that we know that the true underlying model is a power law, and its true parameter value is α, before hand. In practical situations however we would not know if the underlying model that produced the data was in fact a power law. MLE described in Equation (3.2) just gives the best fit to the power-law form, but does not tell us if the power law is infact a good model to fit to the data. These important questions are addressed in Chapters 4 and 5.

Transformation method described in [5], is one of the ways to generate random deviates following the power-law form. Using this method, two sets of continuous power-law data following Equation (2.2) were generated, both with $x_{\min} = 1$ and $n = 10000$. The first set of random deviates were binned using the bin boundaries $B = (1, 11, 21, 31, 41, 51, ...)$. Thus bin width was constant and equal to 10. The second set of random deviates were binned using $B = (1, 2, 4, 8, 16, 32, 64, ...)$. Thus the bin boundaries are in powers of 2. We then fit the power-law form using the binned version of MLE. Figure 3.1, shows the plot of estimated alpha, i.e, $\hat{\alpha}$ as a function of the true underlying α. Our binned MLE approach, gives high accuracy in parameter estimation for both binning schemes.

Comparison of binned MLE is done with the ordinary least-squares (OLS) linear regression. Fitting regression line to PDF using OLS is shown in Figure 3.1a. For the logarithmic binning strategy, the slope is consistently underestimated. For the constant-width binning scheme, the

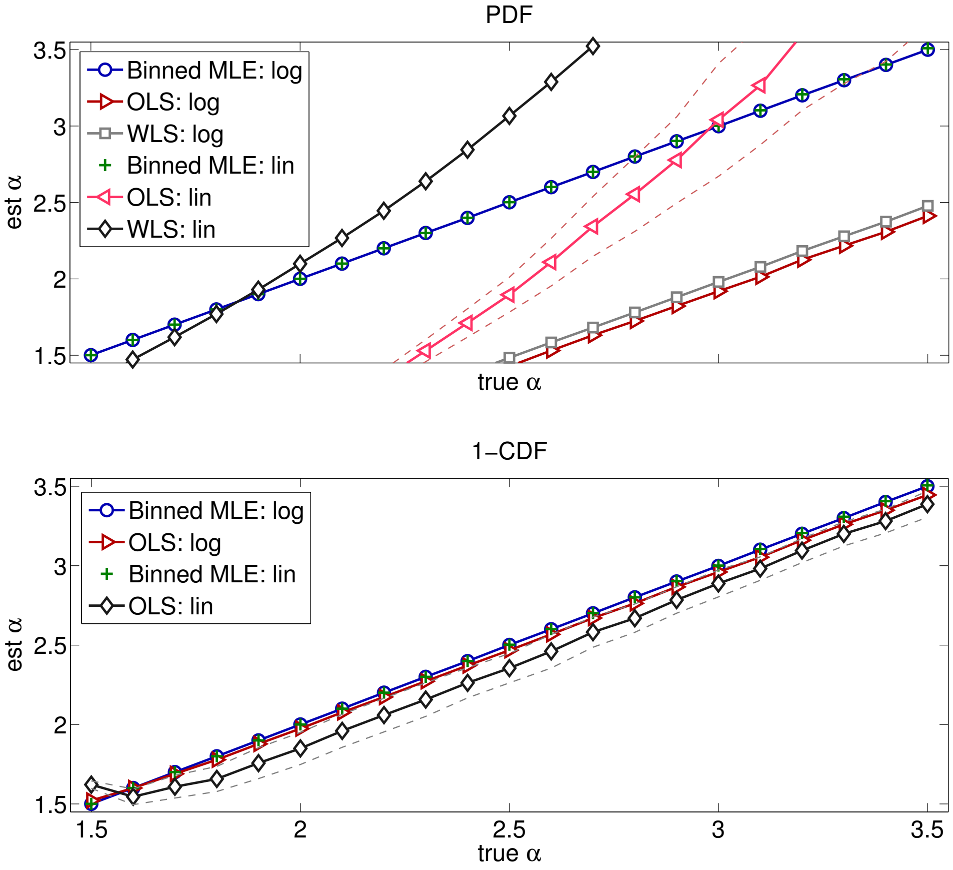


Figure 3.1: Comparison of binned version of MLE (denoted by Binned MLE) and linear regression methods (denoted by OLS for ordinary least squares and WLS for weighted least squares). Comparison is also made on different binning schemes such as logarithmic (log) and constant width (const-width). The variance is denoted by dashed lines wherever significant. (a) For OLS and WLS fits are made to pdf. (b) For OLS fits are made to 1-cdf.

accuracy for linear regression is very poor as the estimates for α go way off and there is some variance in estimation shown by the dashed lines. This is because with constant-width binning, the tail of the distribution is extremely noisy. To improve the accuracy of the linear regression approach, we also use weighted least-squares method (WLS). However, this improves accuracy only slightly for logarithmic binning and very little for constant-width bins.

As a second test, we compare MLE to regressions on the complementary CDF (i.e., 1-CDF) (Figure 3.1b). Though this approach greatly improves accuracy, with logarithmic binning, the slope is consistently and slightly underestimated for high values of α, while with constant-width binning the slope is consistently underestimated with some variance in estimation.

To further improve results using linear regression, fits are made to complementary CDF (i.e., 1-CDF) as shown in (Figure 3.1b). Though this approach greatly improves accuracy, with logarithmic binning, slope gets slightly underestimated for high values of true α, while with constant-width binning the slope is consistently underestimated with some variance in estimation.

With numerically maximizing approach used for the binned MLE, we can expect the binned MLE method to be slightly slower than the linear regression techniques. In practice however, if logarithmic binning scheme is used, we can obtain an analytic solution given by Equation (3.3), in which case binned MLE is equally fast, with better accuracy. For these reasons, our binned MLE version is recommended over linear regression techniques for estimating α.

3.3 Estimating lower bound on power-law behavior

In practice, very few empirical phenomenon follow a power-law distribution for all values of x. Instead, the power law holds above some lower threshold, which we call $b_{\min}$. Thus x scales above $b_{\min}$ and this region is called the tail of the distribution. To fit a power-law model, we need to identify $b_{\min}$ and then discard any data below $b_{\min}$. It is no surprise that if we cannot identify the correct lower-bound or threshold on power-law behavior then the estimation of scaling parameter, i.e, α would be incorrect. Choosing a low value of $b_{\min}$ results in estimate of α that is biased as we try to fit a power law to non power-law data. Overestimating $b_{\min}$ would result in throwing away legitimate data and increasing parameter uncertainty, and finite size effects might bias our estimate of α. Thus we need an estimate of $b_{\min}$ which we denote as $\hat{b}_{\min}$, to be fairly accurate.

A common way of choosing a lower bound on power-law behavior is by estimating $\hat{b}_{\min}$ visually as a point beyond which the PDF or the CCDF i.e. complementary cumulative distribution function becomes roughly straight on a log-log plot. However such eyeballing the exact location of $b_{\min}$ would be subjective. Another way is to plot $\hat{\alpha}$ (or a related quantity) as a function of $\hat{b}_{\min}$ and identify a point beyond which $\hat{\alpha}$ appears relatively stable or constant. This is called the Hill-plot [11]. With this scheme we can see that $\hat{\alpha}$ deviates rapidly from its true value, i.e, α, if $\hat{b}_{\min}$ is estimated below its true value $b_{\min}$. Above the true $b_{\min}$, we would get fairly stable $\hat{\alpha}$ and this would be the range over which we pick our estimate $\hat{b}_{\min}$. The potential problem in using Hill-plot is to identify this stable region where estimate of $\hat{\alpha}$ is constant, as this becomes subjective.

The algorithm presented by Clauset et.al. for finding $x_{\min}$ for non-binned data is an objective and robust approach to determine $x_{\min}$, and can be adapted to the binned case to determine $b_{\min}$.

The adapted algorithm works as follows-

(1) Choose a value of $\hat{b}_{\min} \in B'$ where $B' = (b_1, b_2, b_3, ..., b_{k-1})$ for empirical data set with k bins.

(2) Fit the power-law model, i.e., estimate α using methods described in Section 3.2

(3) Compare the CDFs of the fitted model and the data using Kolmogorov-Smirnov or KS statistic. Store the computed KS-statistic value.

(4) Repeat above steps for the candidate $\hat{b}_{\min}$ values, i.e, for all the elements in set B except the last two boundaries, i.e, till b_{k-1}.

(5) Choose $\hat{b}_{\min}$ that minimizes the KS statistic

If $P(x)$ is the CDF for the binned version of the power-law model and $S(x)$ is the CDF of

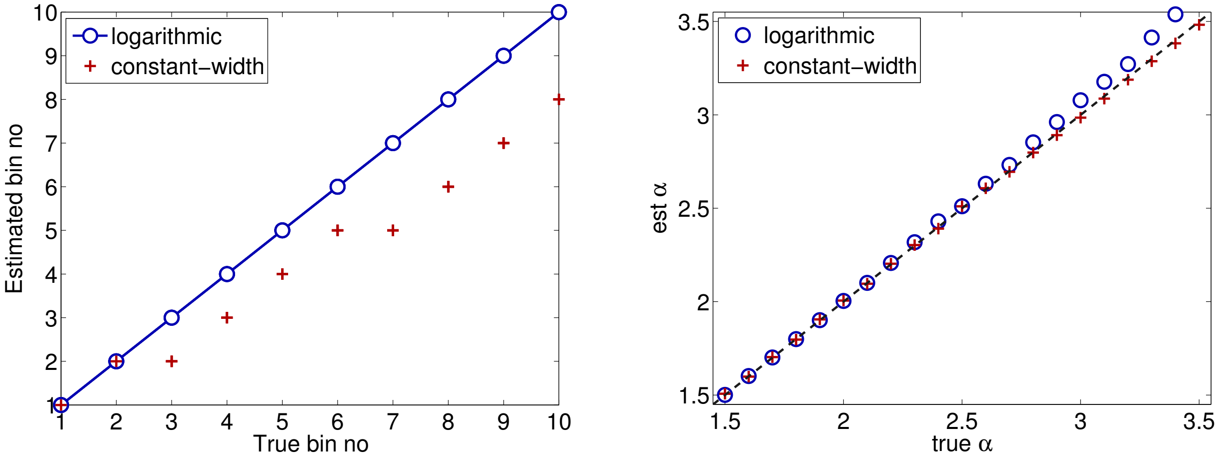


Figure 3.2: (a)True Bin number vs Estimated Bin number (b) True α vs Estimated α for true bin number 10. Results are for logarithmic binning strategy as well as constant-width binning.

the binned data, we choose $\hat{b}_{\min}$ that minimizes D given below

$$D = \max_{x \geq \hat{b}_{\min}} |S(x) - P(x)| \tag{3.5}$$

As mentioned before, choosing a low value of $b_{\min}$ results in fitting a power law to non-power-law data resulting in biased α. The correct value of $b_{\min}$ lies in this trade-off between bias due to fitting wrong model and variance due to small sample size. Since KS-statistic compares CDFs of fitted models and sees how far the model is from data, it captures this trade-off very well. When our model tries to fit data that does not follow the model, KS-statistic gives high value as we would expect. When sample size is small, KS-statistic would still give a high value, as large variance would cause the CDFs of our model and the data to be fairly apart. Thus this method is highly sensitive below and above true $b_{\min}$, which is why we expect this method to work well.

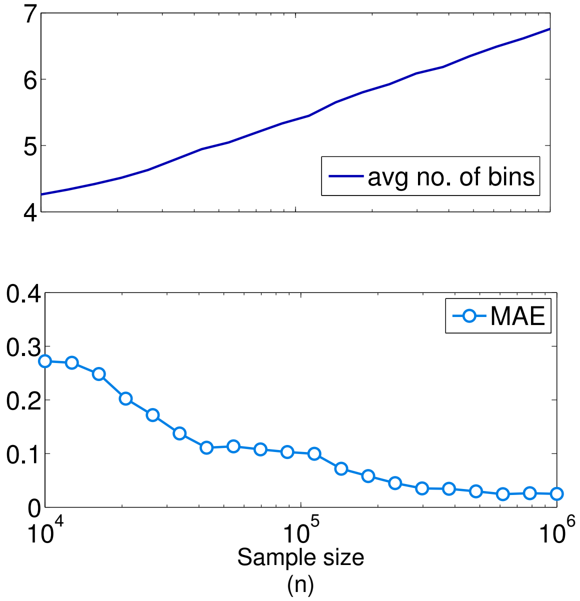


Figure 3.3: (a)Average number of bins after $b_{\min}$ as a function of sample size b (b) Mean absolute error as a function of n. Results are for logarithmic binning strategy, with bin boundaries in powers of 2 and for fixed underlying $b_{\min} = 512$ and $\alpha = 3.5$.

3.4 Testing Estimation of Lower Bound

To check the accuracy of the approach for binned data, we generated synthetic data drawn from a distribution of the form given below and bin it using some binning scheme,

$$p(x) = \begin{cases} Ce^{-\alpha(x/b_{\min}-1)}, & \text{for } x < b_{\min} \\ C(x/b_{\min})^{-\alpha}, & \text{for } x \geq b_{\min} \end{cases} \tag{3.6}$$

Random deviates following the above distribution were generated using $\alpha = 2.5$ and $N = 10,000$. This raw-vaued or non-binned data was then binned according to two different binning schemes. First with logarithmic binning scheme, where bin boundaries were in powers of 2, thus $B = (1, 2, 4, 8, 16, 32, 64, ...)$ and then with constant-width binning, where the bin-width was kept constant at 50 making $B = (1, 51, 101, 151, 201, 251, ...)$. For experiments with both the binning schemes we assume the true underlying $b_{\min}$ as one of the bin boundaries denoted by set B. Since the data has a continuous slope at $b_{\min}$, this makes estimating $b_{\min}$ difficult. With binned data, the task of estimating $b_{\min}$ is effectively choosing the right bin. Hence in checking accuracy of the approach to find $b_{\min}$, we are interested only in finding if the bin number for the true and the estimated $b_{\min}$ is the same. Bin number i denotes the bin from b_i to b_{i+1}.

In the first experiment we determine $\hat{b}_{\min}$ using the algorithm given in the previous section. Since we are interested in finding the correct bin number, the Figure 3.2a gives the plot of true versus estimated bin numbers for both the binning schemes. With the logarithmic binning strategy, estimates of the bin number are highly accurate for every assumed true $b_{\min}$ value. However, with constant-width binning scheme, for high values of true $b_{\min}$, the bin number gets underestimated.

There are some subtleties that we need to consider. Firstly the bin numbered i is different for constant-width binning and logarithmic binning. For instance the third bin for constant-width binning is $b_3 = 101$, while for logarithmic binning this value is between bins b_7 and b_8 as shown by set B for both cases. Also since bin widths for logarithmic binning increase in powers of some constant, higher the true value of $b_{\min}$, greater is the bias for the bins above $b_{\min}$ and lesser the variance. However with constant-width binning the bin widths remains the same throughout, and hence with higher $b_{\min}$, greater variance is observed for bins above $b_{\min}$. This results in underestimating $\hat{b}_{\min}$ value.

Since with the linear binning scheme the $b_{\min}$ value gets underestimated, another experiment was made varying $b_{\min}$, but instead of calculating $\hat{b}_{\min}$, the estimate of slope $\hat{\alpha}$ was calculated. For true underlying $b_{\min} = b_{10}$, i.e, for the 10^{th} bin we see that the estimated bin number, is highly underestimated. If we prove that inspite of this, we can correctly determine the slope, then it would prove that the methods are robust even with linearly spaced bin boundaries.

For constant-width binning, Figure 3.2b shows that for true bin number 10 estimated $\hat{\alpha}$ follows true α very closely. This means that inspite of underestimation of the bin number seen with constant-width binning, our methods calculate the scaling parameter α correctly, which is really the most important piece of power-law model. One could estimate α correctly if one chooses large $b_{\min}$ and throws out more data than required. However Figure 3.2a shows that $b_{\min}$ is actually getting a little underestimated in our case. The effect of the above two results is that we do not throw away more data points than required, while still calculating the slope correctly.

An interesting result is obtained for logarithmic binning. From Figure 3.2b, it seems that estimate $\hat{\alpha}$ is overestimated for high α value. If our fitting procedure has less number of bins to work with, slope might get overestimated. With high underlying $b_{\min}$ and α, this is exactly what occurs. Thus logarithmic binning should be treated with caution. If used for representational purposes, one must consider the sample size and hence the number of bins on average after $b_{\min}$ so that calulation of slope is done correctly.

Figure 3.3 shows the dependence of number of bins on average above $\hat{b}_{\min}$ on sample size n, for a fixed underlying $b_{\min} = 512$ and fixed $\alpha = 3.5$, when logarithmic binning scheme is used. As expected, the average number of bins, go on increasing, as the sample size n increases. The second part of the figure shows that with increasing average number of bins, the mean-absolute error $MAE(\hat{\alpha}) = \langle|\alpha - \hat{\alpha}|\rangle$ goes on decreasing and tends to 0. Figure 3.2a proves that we can estimate $b_{\min}$ correctly for the logarithmic binning, while Figure 3.3 shows that, given sufficient sample size, the slope can be estimated correctly as well.

Chapter 4

Testing the Power-Law Hypothesis

In the previous section, we answered the question of fitting power-law form to binned data. However, regardless of the true underlying distribution of the data, we can always fit our hypothesized model, i.e, the power-law distribution. Thus we need some way to test our hypothesis and see if the power law is a plausible fit to the data. The importance of testing power-law hypothesis has been discussed in [13]. Since, there are numerous models that yield power laws (see [12], [15]), and since more than one of these models can be used to explain an observed power law, validation becomes a crucial part of power law research. In studying empirical data, one strives to understand the underlying process that generated the data. Claiming a power law would imply mechanisms such as preferential attachment, self-organized criticality, random walks etc., and hence carefully testing such an hypothesis is important to avoid wrong inferences or predictions.

Since qualitative appraisals of the data, such as visualization are typically misleading and can lead to false conclusions, we need a quantitative way of testing the power-law hypothesis. This can be illustrated with an example (depicted in Figure 4.1a), which considers three heavy tailed distributions and compares them visually on log-log axes. Three synthetic data sets ($n = 100$) drawn from a power-law distribution with $\alpha = 2.5$, an exponential distribution with $\lambda = 0.125$ and a log-normal distribution with $\mu = 0.3$, $\sigma = 2$ are shown. Lower bound in each case was $b_{\min} = 16$ and the data was binned with bin boundaries as, $B = (16, 32, 64, 128, 256, ...)$, i.e, logarithmic binning with boundaries in powers of 2. In the figure, the x-axis gives the bin number and y-axis gives the count of elements in each bin. Upon cursory judgement one might conclude all three binned datasets to follow the power-law form, as all appear fairly straight on log-log axes, albeit with different slopes.

Given a set of models, in our case, heavy tailed distributions, if the task is to validate fit of one of the conjectured models, the task becomes quite challenging. This is typically because one needs to distinguish between fit of the conjectured model from its close relatives and fit of the conjectured model from deviations caused by sampling process. To answer the first part we use the likelihood-ratio tests as decried in Chapter 5, while the latter part is answered by the goodness-of-fit tests.

4.1 Goodness-of-fit tests

A goodness-of-fit test describes how well our hypothesized model, i.e, the power law, fits the observed or empirical data. Generally the outcome of such a test is a p-value that quantifies the plausibility of the null hypothesis, which in our case, is that the data follows a power-law distribution. If this p-value is below some level of significance, one can reject the null hypothesis. However, a high p-value cannot be used to accept the null hypothesis. A high p-value could mean that the null hypothesis is plausible explanation of the data, or that our test has low statistical power.

The basic approach is summarized below,

(1) Fit the power-law model $(\hat{b}_{\text{min}}, \hat{\alpha})$ to the binned empirical data using methods described in Chapter 3.

(2) Measure how far the fitted model is from the data using some distance measure d^*. We use the Kolmogorov-Smirnov statistic[1] given by Equation (3.5) to calculate d^*.

(3) Using semi-parametric bootstrap, generate m resamples of original binned empirical data, such that the resamples follow fitted power-law above $\hat{b}_{\text{min}}$, but follow the binned empirical

[1] Since in our case, data is binned, the pdf exists and we could potentially use the Pearson's χ^2 statistic as a distance measure. However due to high central tendency and variance associated with the statistic, it is not a good choice and hence not used here. See [16] for detailed analysis.

data below $\hat{b}_{\min}$.

(4) Fit the power-law model ($\hat{y}_{\min}$, $\hat{\beta}$) to these m resamples. Here β represents true underlying parameter value of the resampled data, i.e, $\beta = \hat{\alpha}$ and y represents its true underlying lower bound, i.e, $y = \hat{b}_{\min}$.

(5) Measure how far these fitted distributions are from the original fitted model using the same distance measure, i.e, the KS statistic, and store these distances as d.

(6) Then calculate a p-value as a fraction of the distances in the previous step that are larger than the distance measure calculated in step 2, i.e., $p = \Pr(d \geq d^*)$.

(7) If the p-value is less than some level of significance, reject the power-law hypothesis.

Step 3, requires more explanation. Let us assume we have the sum of frequencies in the histogram of empirical data as N. If out of those N elements, n elements lie above $\hat{b}_{\min}$, then with probability n/N, we generate a power-law random deviate with lower-bound $\hat{b}_{\min}$ and scaling parameter $\hat{\alpha}$. With probability $1 - n/N$, we sample one element following the binned empirical distribution below $\hat{b}_{\min}$. After generating N data points this way, we bin them using the same empirical binning scheme B. Thus we generate m synthetic binned data sets as required in step 3, such that the m resamples follow a power law above $\hat{b}_{\min}$, but have the same non power-law part as that of the binned empirical distribution below $\hat{b}_{\min}$.

Note that while in step 2 we measure fluctuations of the fitted model ($\hat{b}_{\min}$, $\hat{\alpha}$) to the empirical data, in step 5, we measure fluctuations of fits to the resamples, i.e, ($\hat{y}_{\min}$, $\hat{\beta}$) to the fitted model, i.e, ($\hat{b}_{\min}$, $\hat{\alpha}$) and not to the original empirical data. This is extremely important to get an unbiased p-value. If fits to the resamples are compared to the binned empirical data instead, significant fraction of the distances d would be greater than d^*, even though the model ($\hat{b}_{\min}$, $\hat{\alpha}$) is a poor fit, thus giving a high p-value.

For rejecting the power-law hypothesis, we choose a very conservative 10% level of significance. This means that if the p-value calculated in step 6 is, $p < 0.1$, then we reject the power-law hypothesis. In hypothesis testing incorrectly rejecting a true null hypothesis is called as a Type I error, while incorrectly accepting a false null hypothesis is called as a Type II error. Threshold that we have selected is conservative compared to say 5% level of significance, in the sense that we would try to reduce Type II error while perhaps allowing the occasional Type I error.

Note that, $p > 0.1$, does not allow us to accept a power law hypothesis as noted earlier. One of the reasons is that for relatively low values of N, empirical data might follow the power-law form closely leading to a high p-value, even when the underlying process is non power law. Other reason for not favoring the power-law hypothesis is that some other heavy-tailed distribution might fit the empirical data better. Thus in this case we need to investigate further, in order to make a strong case for power-law hypothesis. This is the topic for Chapter 5.

4.2 Performance of Goodness-of-fit tests

To demonstrate that we can correctly distinguish the power-law behavior from the non-power-law behavior even on binning the data, we consider data drawn from three distinct underlying distribution such as a power law, a log normal and an exponential. The parameter settings for the three distributions is as given in caption for Figure 4.1a. All of the three data sets were binned using logarithmic binning strategy with bin boundaries as powers of 2, i.e, $B = (1, 2, 4, 8, 16, ...)$.

In the Figure 4.1b, we show the average p-value for the data sets drawn from these three distributions, as a function of the sample size n. The average p-value for underlying power-law distribution is close to 0.5 for all n values. With log-normal and exponential cases, as n increases, the p-value falls below the rule-of-thumb threshold of 0.1 and thus the power-law hypothesis would be rejected in these cases. Thus the results obtained in [5] hold even on binning.

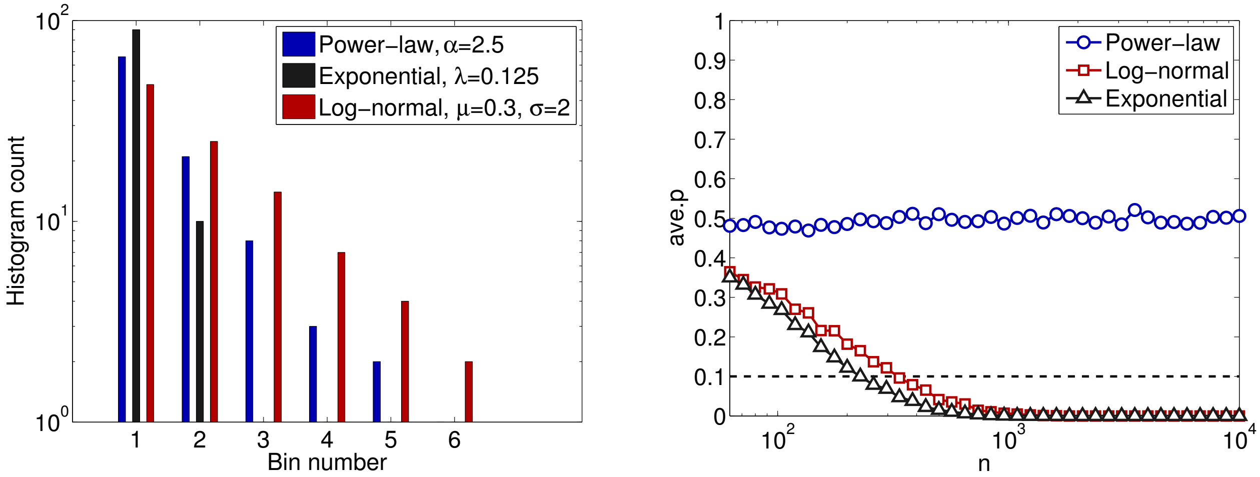


Figure 4.1: (a) The binned PDFs of three samples ($n = 100$) drawn by binning continuous distributions in a logarithmic binning scheme: a power law with $\alpha = 2.5$, an exponential with $\lambda = 0.125$ and a log-normal with $\mu = 0.3$ and $\sigma = 2$, all with $b_{\min} = 16$ (b)The average p-value for the maximum likelihood power-law model for samples from the given three distributions, as a function of the number of observations n. As n increases, only the p-value for power-law distributed data remains above our rule-of-thumb threshold $p = 0.1$, with the others falling off toward zero, indicating that p does correctly identify the true power-law behavior in this case.

Chapter 5

Alternative Distributions

We have the empirical data at hand with limited knowledge about it and our principal goal is to find a model that best describes our data. This way, we can make qualitative judgements about forecasting based on our model. Thus our aim is not just to test empirical data for the power-law hypothesis, but to find the hypothesis that best describes the data. The point is that even if the power law is a plausible fit to the data, alternative distributions such as the log normal or stretched exponential might fit the data equally well. Eliminating these alternatives can strengthen the case for power law.

Note that, in this case, our judgment will be always limited to the number of models we hypothesize, and there could be a better model which we did not consider to begin with. Thus, this is essentially a tough problem, and prior knowledge about the data at hand is required to decide what constitutes a good hypothesis. If reasonable models are not considered to begin with, we can never accurately forecast and hence choosing a set of reasonable models is critical in making qualitative judgements based on our model.

In our case, if one considers power-law distribution as a plausible hypothesis based on prior knowledge about the empirical data, then its highly probable that the empirical distribution is heavy tailed. This means that good alternative explanations of the empirical data could be other heavy-tailed distributions and thus exponential distribution, stretched exponential distribution or log-normal distributions would constitute as good alternative hypotheses.

To rule out such competing distributions we could use goodness-of-fit tests to generate a p-

value after fitting each of the competing distributions to the data as explained in previous sections. Thus the case for the power-law hypothesis would be strengthened if we get significant p-values for competing distributions, i.e, $p < 0.1$.

5.1 Direct Comparison of Models

The method mentioned above to compare different competing distributions, involves adapting fitting procedure of Chapter 3 and hypothesis testing described in Chapter 4 to all candidate distributions, which is a major effort. As noted earlier, if we are interested in knowing what process produced the observed data, we are only interested in knowing which of the hypothesized distributions is a better fit. In such cases, we can directly compare the different distributions using methods such as likelihood ratio tests, cross validation [21], minimum description length [10], etc. In this section, we use the method of likelihood ratio test.

The likelihood of a model given our data is the probability of the data given the model. The likelihood ratio R, is the ratio of likelihoods under different competing models. More specifically, the likelihood ratio compares the null hypothesis (i.e., the power law) to competing hypotheses (exponential, log-normal, etc.). In practice, for convenience, we take the log of R and the log-likelihood ratio is denoted by $\mathcal{R}$. The sign of $\mathcal{R}$ indicates the favored hypothesis. We get $\mathcal{R} > 0$ when the null hypothesis is a better fit, $\mathcal{R} < 0$ when the alternative hypothesis is a better fit and $\mathcal{R} = 0$ in case of a tie. Note that, a tie, implies that we have insufficient data to discriminate between the null hypothesis and the alternative hypothesis.

We calculate the likelihood ratio as,

$$\mathcal{R} = \ln\left(\frac{\mathcal{L}_A(x_i|\hat{\theta}_A)}{\mathcal{L}_B(x_i|\hat{\theta}_B)}\right) \tag{5.1}$$

Here, $\mathcal{L}_A$ denotes the likelihood for the model A, i.e, the power-law model and the $\mathcal{L}_B$ denotes the likelihood for one of the competing models given above. We take logarithm of the ratio of likelihoods, since its mathematically easier to deal with.

Since our data are independent random variables, the log-likelihood ratio or $\mathcal{R}$, is also a random variable and is subject to statistical fluctuations. Thus if the expected value of $\mathcal{R}$ is close to zero, its observed sign could be a result of chance and cannot be trusted. In this case, we need to measure the size of fluctuations associated with $\mathcal{R}$, i.e, its standard deviation, σ. For this we use a method proposed by Vuong [23]. This method gives us a p-value that gives us an estimate of the probability that we measured a given value of $\mathcal{R}$ when its true value is close to zero, and thus a large p-value (say, $p > 0.1$) indicates that we cannot trust the sign of $\mathcal{R}$. If it is small ($p < 0.1$), then the observed sign is reliable and we can trust $\mathcal{R}$ to decide which is the better fit. The technicalities involved in likelihood ratio tests are described in Appendix C.

5.2 Performance of log-likelihood tests

To test the sanity of this likelihood ratio test for binned data, we made two tests on synthetic data. In the first test, data was drawn from a power-law distribution with $\alpha = 2.5$ and $x_{\min} = 1$ and in the second test the data was drawn from a log-normal distribution with $\mu = 0.3$ and $\sigma = 2$. Binning scheme used in both cases was logarithmic in powers of 2. For both experiments normalized log-likelihood ratio, i.e., $n^{-1/2}\mathcal{R}/\sigma$ was found as a function of number of elements n. Normalizing the log-likelihood ratio, allows us to compare the different $\mathcal{R}$s found for different data sets. Figure 5.1a shows the results for the first test. In this case, as the number of elements n increases, normalized log-likelihood ratio becomes more and more positive, thus supporting the power-law hypothesis. In the second test, with underlying log-normal distribution, however, the normalized log-likelihood ratio becomes more and more negative for increasing n as shown in Figure 5.1b, thus rejecting the power-law hypothesis.

It is important to note that blindly following the sign of $\mathcal{R}$ can result in misclassification of data. If we consider p-value associated with $\mathcal{R}$ as mentioned before, we can reduce this misclassification. We see that with underlying power-law distribution, we can start trusting the sign of $\mathcal{R}$ after $n = 10^4$, while with underlying log-normal distribution, we can trust $\mathcal{R}$ even with sample size of about $n = 10^2$. This experiment further supports the need to rigorously test power-law hypothesis as it implies that that the log-normal distribution approximates a power law pretty well

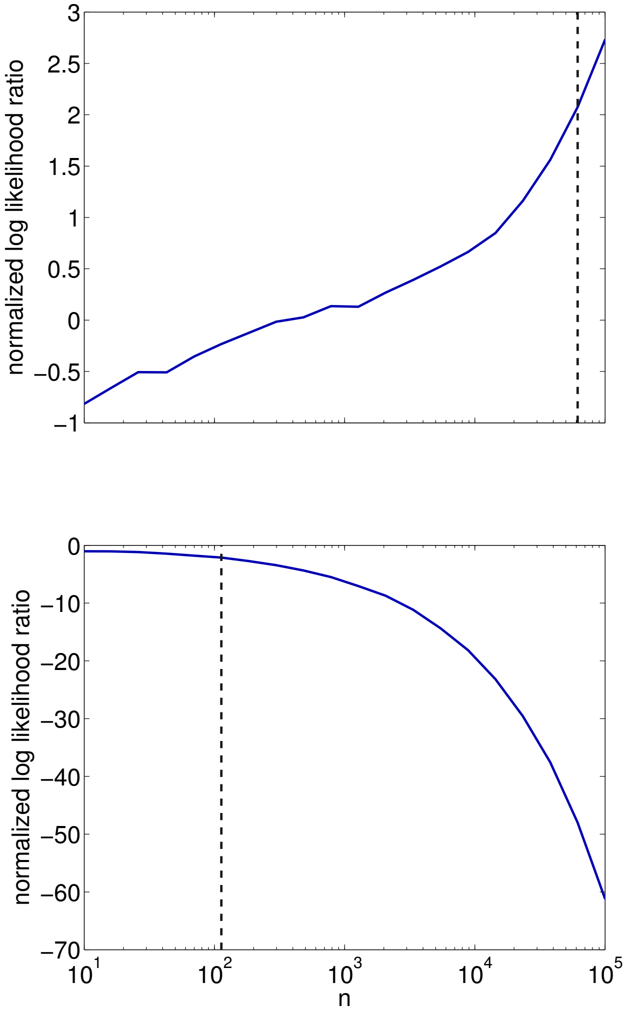


Figure 5.1: Behavior of normalized log-likelihood ratio $n^{-1/2}\mathcal{R}/\sigma$, for synthetic data sets drawn from (a) power-law distribution and (b) log-normal distribution. For both the cases, the synthetic data was binned using the logarithmic binning scheme, with bin boundaries in powers of 2. The dotted line indicates the value of n after which p-value becomes significant, i.e., $p < 0.1$, and we can trust the sign of $\mathcal{R}$.

even for large sample sizes.

Chapter 6

Applications to Empirical Data

In this chapter, we apply the methods described in this thesis, to some real-world binned data sets, some of which have been conjectured to follow the power law. As we will see, our results indicate that for none of the chosen data sets, the power-law hypothesis is strongly favored.

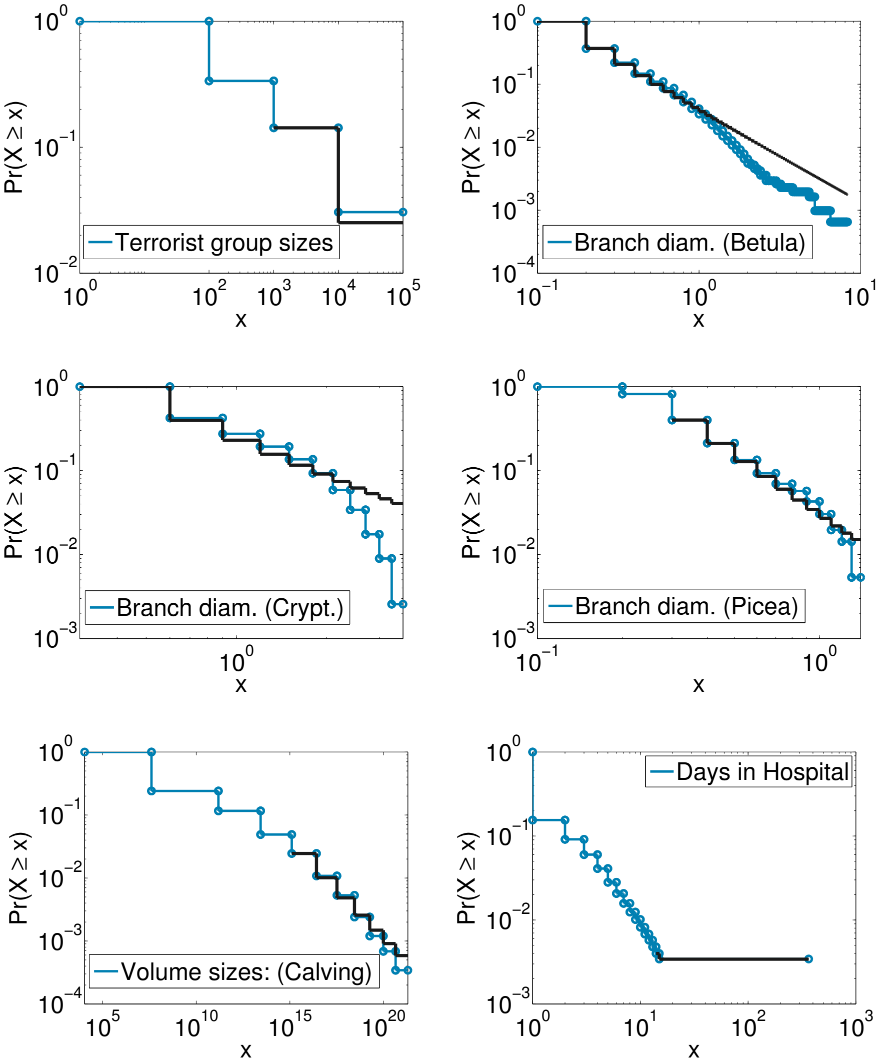
(1) Terrorist group sizes: The frequency-distribution of sizes of terrorist groups, i.e, number of personnel in terrorist organizations [1]. In all 393 groups are recorded. Bin boundaries are given by $B = (1, 100, 1000, 10000, 100000)$.

(2) Branch diameter (Species: Betula, Cryptomeria, Picea): The number of branches as a function of branch diameter [26]. For Betula species, the number of branches were 3096 while for Cryptomeria and Picea number of branches measured were 3897 and 560 respectively. For Betula and Picea measurements were made in 10mm intervals while for Cryptomeria measurements were made in intervals of 30mm.

(3) Volume sizes of Calving events: The frequency distribution of volumes for calving events [4]. The data consists of 5837 calving events in all, with 11 unique sizes. Size of calving event s, is related to volume as,

$$V = 12.6(s)^3.87$$

Using this equation we found volume in m^3 and binned the data. Binning scheme used was logarithmic in powers of 10 such that each of 11 unique volume sizes lie in the middle of a set of bin boundaries.


10^0
10^{-1}
10^{-2}
Pr(X ≥ x)
Terrorist group sizes
10^0
10^2
10^3
10^4
10^5
x
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
Pr(X ≥ x)
Branch diam. (Betula)
10^{-1}
10^0
10^1
x
10^0
10^{-1}
10^{-2}
10^{-3}
Pr(X ≥ x)
Branch diam. (Crypt.)
10^0
x
10^0
10^{-1}
10^{-2}
10^{-3}
Pr(X ≥ x)
Branch diam. (Picea)
10^{-1}
10^0
x
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
Pr(X ≥ x)
Volume sizes: (Calving)
10^5
10^{10}
10^{15}
10^{20}
x
10^0
10^{-1}
10^{-2}
10^{-3}
Pr(X ≥ x)
Days in Hospital
10^0
10^1
10^2
10^3
x

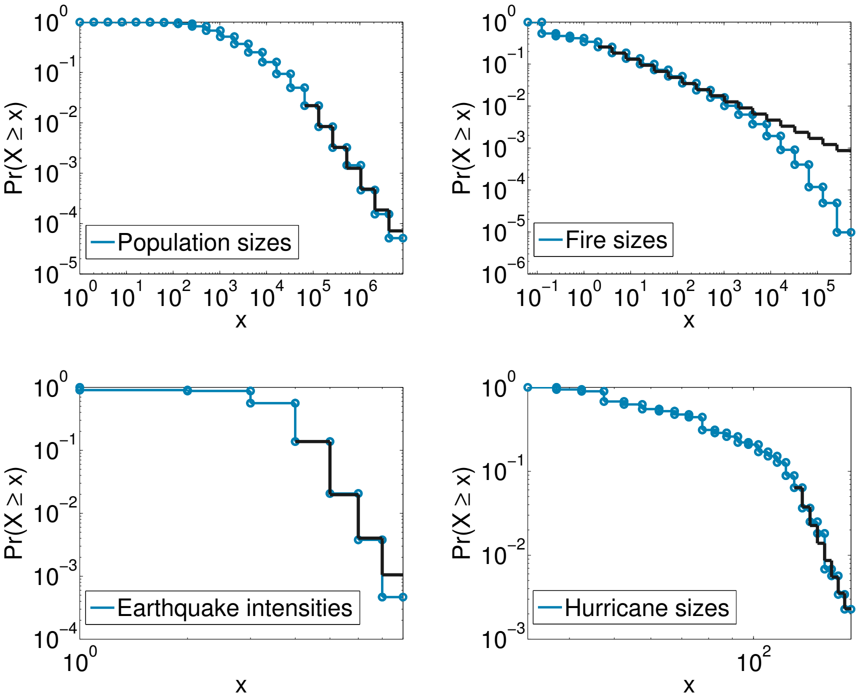


Figure 6.0: The CCDFs P(x) for the five empirical data sets and their maximum likelihood power-law fits.

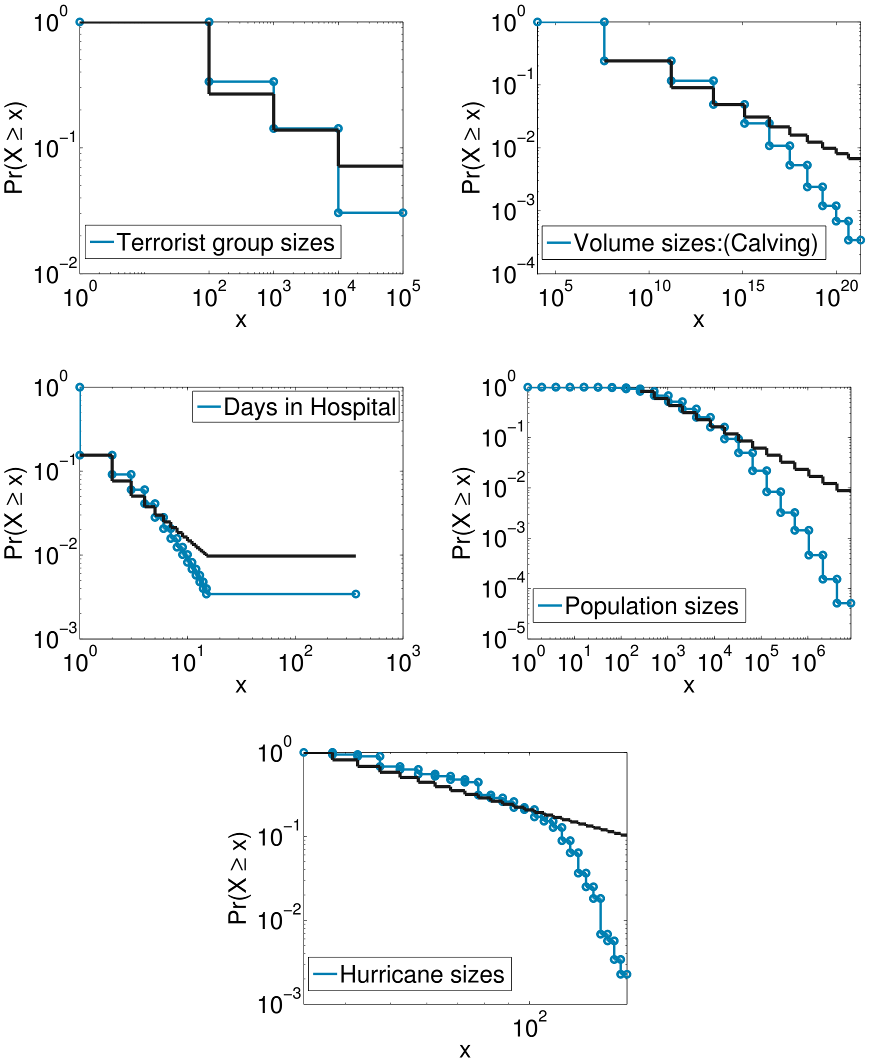


Figure 6.1: The CCDFs P(x) for empirical data sets with $b_{\min} = b_1$ for terrorist group sizes and hurricane sizes data sets, $b_{\min} = b_2$ for days in hospital and volume sizes data sets and $b_{\min} = b_9$ for population sizes data.

$b_{\min}$	Quantity	Sample size N	$\hat{\alpha}$	$\hat{b}_{\min}$	Elements in tail n	p-value
Estimated	Terrorist group sizes	393	1.753 (0.105)	1000	56	0.13
	Branch diam.(Betula) (mm)	3096	2.437 (0.025)	0.1	3096	**0.02**
	Branch diam.(Crypt.) (mm)	3897	2.337 (0.020)	0.3	3897	**0.00**
	Branch diam.(Picea) (mm)	560	3.236 (0.130)	0.3	224	0.79
	Volume sizes (m^3)	5837	1.292 (0.022)	$(1.26) \times 10^{15}$	143	0.45
	Days in Hospital	76038	3.239 (0.266)	14	303	0.40
	Hurricane sizes (m/s)	879	14.200 (1.689)	122.5	56	0.36
	Earthquake intensities (M_{L})	19302	9.71 (0.185)	4	2659	**0.10**
	Population sizes	19447	2.380 (0.068)	65536	426	0.72
	Fire sizes $(acres)$	203785	1.482 (0.002)	2	52004	**0.00**
Fixed	Terrorist group sizes	393	1.286 (0.011)	1	393	**0.00**
	Volume sizes (m^3)	5837	1.120 (0.002)	$(4.1) \times 10^7$	1405	**0.00**
	Days in Hospital	76038	2.020 (0.007)	1	11769	**0.00**
	Hurricane sizes (m/s)	879	2.437 (0.027)	32.5	879	**0.00**
	Population sizes	19447	1.468 (0.003)	256	16015	**0.00**

Table 6.1: Results of parameter estimation for all empirical data sets. In first set of experiments, $b_{\min}$ is estimated while in second set of experiments it is kept fixed; $b_{\min} = b_1$ for terrorist group sizes and hurricane sizes data sets, $b_{\min} = b_2$ for days in hospital and volume sizes data sets and $b_{\min} = b_9$ for population sizes data.

$b_{\min}$	Quantity	Power-law	Log-normal		Exponential		Stretched exp.		Power law + cut-off		Support for power-law
		p	$\mathcal{R}$	p	$\mathcal{R}$	p	$\mathcal{R}$	p	$\mathcal{R}$	p	
Estimated	Terr. grp sizes	0.13	-2.011	**0.04**	3.91	**0.00**	-1.934	**0.05**	-2.57	0.11	weak
	diam.(Bet)	0.02	-1.768	**0.08**	7.194	**0.00**	-1.745	**0.08**	-7.503	**0.01**	weak
	diam.(Crpt.)	0.00	-9.71	**0.00**	1.99	**0.05**	-9.478	**0.00**	-123.76	**0.00**	weak
	diam.(Pic.)	0.79	-1.342	0.18	0.741	0.46	-1.41	0.16	-2.689	**0.10**	with cut-off
	Volume sizes	0.45	-1.501	0.13	10.612	**0.00**	-1.164	0.25	-0.233	0.63	moderate
	Days in Hosp.	0.40	-0.978	0.33	-1.018	0.31	-1.012	0.31	-0.231	0.63	weak
	Hurr. sizes	0.36	-0.352	0.73	6.17	**0.00**	-0.715	0.48	-0.298	0.59	weak
	Quake int.	0.10	-2.81	**0.01**	19.01	**0.00**	-2.165	**0.03**	-5.538	**0.02**	with cut-off
	Pop. sizes	0.72	-0.069	0.95	16.25	**0.00**	-0.081	0.94	-0.229	0.63	moderate
	Fire sizes	0.00	-16.03	**0.00**	9.26	**0.00**	-16.42	**0.00**	-410.014	**0.00**	with cut-off
Fixed	Terr. grp sizes	0.00	-4.32	**0.00**	4.59	**0.00**	-4.47	**0.00**	-26.26	**0.00**	none
	Volume sizes	0.00	-8.53	**0.00**	36.25	**0.00**	-8.777	**0.00**	-36.99	**0.00**	with cut-off
	Days in Hosp.	0.00	-18.37	**0.00**	-1.86	**0.06**	-18.69	**0.00**	-602.86	**0.00**	none
	Hurr sizes	0.00	-13.26	**0.00**	-20.712	**0.00**	-13.78	**0.00**	-117.07	**0.00**	with cut-off
	Pop. sizes	0.00	-29.276	**0.00**	11.66	**0.00**	-29.57	**0.00**	-917.03	**0.00**	with cut-off

Table 6.2: Comparison to other heavy-talied distributions. From left to right, the distributions are log-normal, Exponential, Stretched Exponential and Power-law with cutoff. Also, for each comparison, likelihood ratio and its associated p-value is calculated. In first set of experiments, $b_{\min}$ is estimated while in second set of experiments it is kept fixed; $b_{\min} = b_1$ for terrorist group sizes and hurricane sizes data sets, $b_{\min} = b_2$ for days in hospital and volume sizes data sets and $b_{\min} = b_9$ for population sizes data. Significant p-value results are in bold face.

(4) Days spent in Hospital: Distribution of number of days spent in hospital by patients in one year [14]. This data is provided by Health Heritage Provider Network. The total number of records is 76038, with bin boundaries given as $B = (0, 1, 2, 3, ..., 13, 14, 15, 365)$.

(5) Population sizes: The human populations of US cities in 2000 census.

(6) Fire sizes: The sizes in acres of wildfires occurring on U.S. federal land between 1986 and 1996 [15].

(7) Earthquake intensities: The intensities of earthquakes occurring in California between 1910 and 1992, measured as the maximum amplitude of motion during the quake [15].

(8) Hurricane sizes: Intensity of tropical storms and hurricanes in US from 1949 through 2010 [3]. Intensity is measured in wind speeds (unit: knots). For each of the 879 storms/hurricanes, we consider its maximum recorded wind speed as intensity measure. The reported wind speeds are such that if actual value of wind speed is 47.4 it is reported as 45, while if actual value is 47.6 it is reported as 50. This results in a binning scheme with constant-width of 5, such that $B = (32.5, 37.5, 42.5, 47.5, 52.5, ...)$. In this way each reported value lies exactly in middle of any bin.

Data-sets (5), (6) and (7) are taken from [5]. These are not originally binned. We binned the earthquake intensities data is binned using constant-width binning as $B = (0, 1, 2, 3, ..., 8)$. Both population sizes and fire sizes data sets were binned using logarithmic binning in powers of 2. Recovering the results obtained in [5] for these data, is a good check on our adapted methods.

Figure 6.0 and Figure 6.1 show the power-law fits to all the empirical data sets. Table 6.1 gives the parameter estimates for the same. Depending on whether $b_{\min}$ is estimated using procedure of Section 3.3, or whether $b_{\min}$ is fixed to a lower bin boundary, we classified our results in two sets.

In first set of experiments, we find that two of the branch diameter data sets, namely, species Betula and Cryptomeria and the earthquakes intensities and the fire sizes data sets, are ruled out as a power law, due to significant p-value, i.e, $p < 0.1$. The figures also show considerable curvature for these data. We find that the remaining data sets, namely, terrorist group sizes, branch diameter for Picea species, volume sizes of calving events, days spent in hospital, hurricane sizes and population sizes cannot be ruled out as a power law. To strengthen the case for power law, we can use likelihood ratio-tests to compare power-law fits to these data to fits of other heavy-tailed distributions such as log-normal, exponential, etc. Table 6.2 summarizes the results for these comparisons. For terrorist group sizes data, the log normal or the stretched exponential were found to be better explanations, while for branch diameter (Picea species) data, power law with exponential cut-off does better. These are indicated by negative sign of log-likelihood ratio and a significant p-value. Comparisons for hurricane sizes, population sizes and volume sizes data sets showed non-significant p-values in all cases. Notice from Table 6.1 that n, which is the number of elements after estimated $\hat{b}_{\min}$ or number of elements in the tail of the distribution, is low for these data sets and thus our analysis had low statistical power.

For this reason, the second set of tests, used fixed low value for $b_{\min}$, so that our statistical framework has more data to work with. This is in general a good practice since if we care about the endogenous process that generated our data, we should allow other plausible models to fit to as much data as possible. This way we avoid erroneously seeing power laws.

For all the data sets tried this way, we got $p = 0$ for the power-law hypothesis. Also power law with cut-off seemed to be better explanation than the power law in all the cases. This seems to imply that there could be a characteristic scaling region implying the possibility of large events, but extremely large events might be exponentially rare. Also, it implies that there is considerable curvature in the empirical distribution for these data sets.

It is interesting to note that for the days spent in hospital data, the stretched exponential did well. This data showed the number of days spent by patients in hospital in one year. If we consider that the more a person stays in the hospital, the more is the chance of staying since her condition must be bad. Also it is well known that hazard function used in survival analysis is derived from stretched exponential function by adopting an age-dependent exponent and so our intuitions and results agree. For hurricane data, power law with cut-off is favored. This means that there is some scaling region for wind speeds, but extremely high wind speeds are exponentially rare.

Chapter 7

Conclusions

The principal goal of this thesis was to adapt the framework given in [5] successfully to binned data sets with varied binning schemes. As seen from Chapters 3-5, on applying the modified framework to synthetically generated binned data, we get the expected results, which proves correctness of the methods. We can try to leverage existing analytical tools by sampling given binned data so that we adapt data to suite methods. However this does not work (see Appendix A) and we need to adapt methods to data instead of adapting data to suite existing methods, which was the primary motivation of this work.

Using our binned version of MLE, we get accurate estimate of the scaling parameter α. Also MLE has the nice properties of asymptotic consistency and normally distributed errors. From Appendix D.1, we see that, when we consider binning some existing data with raw values, we must strive to make the bin boundaries as tight as possible, so that there is least information loss. This is an important take-home message from this thesis.

We also proved in Appendix D.2, that even if the binning scheme is fairly arbitrary under the one assumption of non-overlapping bin boundaries, the accuracy of the results does not vary much and we get near correct value for the scaling parameter. In this appendix, we use the data set, frequency of occurrence of unique words in novel Moby-Dick [15], as our test data. On binning this Moby-Dick data set with varied binning schemes, we see that the slope estimation is fairly accurate and follows the results obtained in [5]. Though, this demonstrated the robustness of our methods, in practice, arbitrary binning schemes should be avoided, so that we obtain most accurate results.

On applying our tools to empirical data sets belonging to varied fields, we saw that on initial analysis, where we let our estimation procedure calculate $b_{\min}$, some of the data sets such as, branch diameter for species Betula and Cryptomeria, earthquake intensities and fire sizes, were found not to follow the power-law hypothesis. While other data sets could not be ruled out, the statistical power of these tests was low as our tools had less data to work with. This motivated further investigation. When $b_{\min}$ was fixed to some lower value such that our tools have more data to analyze, we found that the power law did not hold for any data set. Though in some cases such as the hurricane sizes data or the population sizes data where power law with exponential cut-off was favored, for the remaining data sets, stretched exponential or log normal seemed to provide better fits.

There are many complexities involved when we consider binned data sets. As mentioned before, binning would typically arise from uncertainty in estimation, and this means we would have some distribution around our empirical values. This in turn implies, that bin boundaries could potentially cross over, leading to more complex data. More specifically, the binned data in this case would have overlapping bin boundaries. Our framework does not handle such data and this is a potential avenue for future research.

Bibliography

[1] Victor Asal and R. Karl Rethemeyer. The nature of the beast: Organizational structures and the lethality of terrorist attacks. The Journal of Politics, 70(2):437–449, 2008.

[2] Per Bak, Chao Tang, and Kurt Wiesenfeld. Self-organized criticality. Physical Review A, 38(1):381–384, 1987.

[3] National Hurricane Center. http://www.nhc.noaa.gov/pastall.shtml, 2011 (Access: 2012).

[4] A. Chapuis and T. Tetzlaff. What do the distributions of calving-event sizes and intervals say about the stability of tidewater glaciers? Journal of Glaciology, 2011.

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

[6] Aaron Clauset and Ryan Woodard. Estimating the historical and future probabilities of large terrorist events. Preprint, 2012.

[7] H. Cramér. A contribution to the theory of statistical estimation. Skand. Ȧktuaries Tidskrift, 29:458–463, 1946.

[8] Holger Drees and Edgar Kaufmann. Selecting the optimal sample fraction in univariate extreme value estimation. Stochastic Processes and their Applications, 75:149–172, 1998.

[9] Xavier Gabaix. Power laws in economics and finance. Annual Review of Economics, 1:255–293, 2009.

[10] P. D. Grünwald. The Minimum Length Description Principle. MIT Press, Cambridge, MA, 2007.

[11] B. M. Hill. A simple general approach to inference about the tail of a distribution. Ann. Statist., 3(3):1163–1174, 1975.

[12] M. Mitzenmacher. A brief history of generative models of power law and lognormal distributions. Internet Math., 1:226–251, 2004.

[13] M .Mitzenmacher. The future of power law research. Internet Math., 2:525–534, 2006.

[14] Heritage Provider Network. http://www.heritagehealthprize.com/c/hhp/data, 2011 (Access: 2012).

[15] M. E. J. Newman. Power laws, pareto distributions and zipfs's law. Contemp. Phy., 46:323–351, 2005.

[16] P. T. Nicholls. Estimation of zipf parameters. J. Am. Soc. Information Sci., 40:443–445, 1989.

[17] C. R. Rao. Minimum variance and the estimation of several parameters. Proc. Cambridge Phil. Soc., 43:280–283, 1946.

[18] W. J. Reed and B. D. Hughes. From gene families and genera to income and internet file sizes: Why power laws are so common in nature. Physical Review E, 66:067103, 2002.

[19] Rolf-Dieter Reiss and Michael Thomas. Statistical Analysis of Extreme Values: with Applications to Insurance, Finance, Hydrology and Other Fields. Birkhäuser, Basel, Switzerland, 2007.

[20] D. Sornette. Critical Phenomena in Natural Sciences. Springer, Berlin, 2nd edition, 2006.

[21] M. Stone. Cross-validatory choice and assessment of statistical. J. Roy. Statist. Soc. Ser. B, 36:111–133, 1974.

[22] Michael P. H. Stumpf and Mason A. Porter. Critical truths about power laws. Science, 335:665–666, 2012.

[23] Q. H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica, 57:307–333, 1989.

[24] L. Wasserman. All of Statistics: A Concise Course in Statistical Inference. Springer-Verlag, Berlin, 2003.

[25] L. Wasserman. All of non-parametric statistics. Springer, 2006.

[26] Geoffrey B. West, Brian J. Enquist, and James H. Brown. A general quantitative theory of forest structure and dynamics. Proceedings of the National Academy of Sciences of the United States of America, 106(17):7040–7045, 2009.

[27] S. S. Wilks. The large sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Statist., 9:60–62, 1938.

Appendix A

Using continuous approximation of binned data

One of the ways, we can quantify and validate the power-law behavior for binned data is to leverage existing methods in [5] by making given binned data look like continuous data. We could assume power-law distribution with slope $\alpha \in R$ within the bins, where R denotes the range over which we assume values for α. We could then draw continuous data from the power-law model with assumed α and apply fitting procedure described in [5]. We could then make estimate $\hat{\alpha}$ for each assumed α and average over all the estimates to make the final estimate as $\langle\hat{\alpha}\rangle$. This way we adapt data so that existing methods can be used.

However there are potential problems with scheme given above. While we get estimates of α which are close to its true value, the estimator becomes biased by the range R over which we expect true α to lie. We typically get a good estimate if true α lies in the middle of R.

We made an experiment to test this theory. We generated 10 synthetic binned data sets following a power law with true α ranging from 1.5 to 3.5 in steps of 0.1. Logarithmic binning scheme was used with bin boundaries in powers of 2. We then applied our binned MLE approach to make fits to the 10 data sets. We drew continuous data by assuming a power-law distribution within the bins and assuming $\alpha \in R$ where in the first case, $R = [1.5\ 3.5]$, in steps of 0.1 and in the second case $R = [1.5\ 4.5]$, also in steps of 0.1. We then applied continuous MLE to get estimates of α.

Figure A.1 shows the results. Binned MLE does well as estimate $\hat{\alpha}$ follows true α very closely. For continuous MLE, while the estimate $\hat{\alpha}$ is close to the true underlying value, $\hat{\alpha}$ is biased by R,

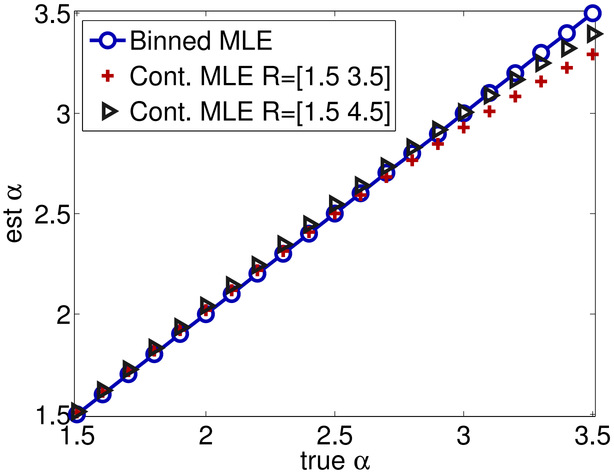


Figure A.1: Comparison of Binned MLE and Continuous MLE applied to binned data set. For Cont. MLE, we assume $R = [1.5\ 3.5]$ in steps of 0.1 and then $R = [1.5\ 4.5]$ in steps of 0.1.

and hence using continuous MLE on continuous approximation of binned data should be avoided to get unbiased estimated of α. Even though this implies adapting methods to data, Figure A.1 proves that this could be well worth the effort.

Another problem is that since we apply fitting procedure to continuous data is that we need to do the procedure for all values of $\alpha \in R$ and then average over the estimates. This could get computationally expensive and hence should be avoided.

Appendix B

Maximum Likelihood Estimator for Binned Power Law

In case of binned data, Chapter 2 defines the binned power-law distribution given by Equation (2.5) as,

$$\Pr(b_i < x < b_{i+1}) = {b_{\min}}^{\alpha-1} \left[{b_i}^{(1-\alpha)} - {b_{i+1}}^{(1-\alpha)} \right]$$

where α is the scaling parameter and $b_{\min}$ is the minimum value at which the power-law behavior holds. Given a data set containing N observations such that h_i observations fall in the i^{th} bin and the data follows a power law above lower-bound $b_{\min}$, we would like to know the value of α for the power-law model that is most likely to have generated our data. The probability that the data were drawn from the model is proportional to,

$$\Pr(b_i < x < b_{i+1}|\alpha) = \prod_{i=1}^{k} \left({b_{\min}}^{\alpha-1} \left[{b_i}^{(1-\alpha)} - {b_{i+1}}^{(1-\alpha)} \right] \right)^{h_i} \tag{B.1}$$

This probability is called the likelihood of data given the model. The data are most likely to have been generated by the model with scaling parameter α that maximizes this function. For ease of

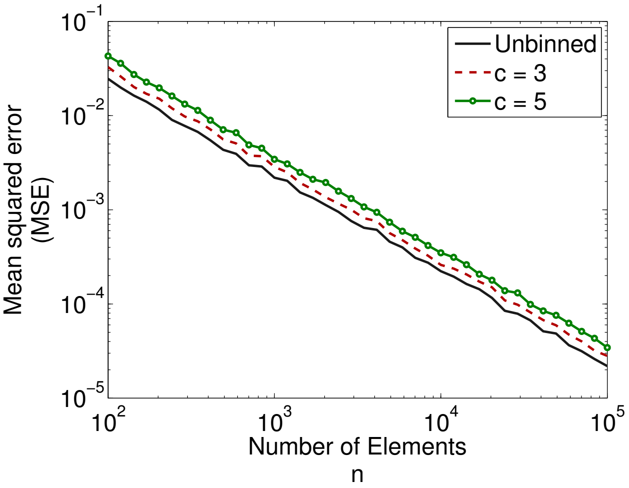


Figure B.1: Plot demonstrating asymptotic consistency for logarithmic binning with different number of bins. Comparison to continuous MLE on unbinned data is also shown.

use we normally work with logarithm $\mathcal{L}$ of the likelihood, which has its maximum at the same place:

$$\begin{aligned}
\mathcal{L} &= \ln \Pr(b_i < x < b_{i+1}|\alpha) \\
&= \ln \left[\prod_{i=\min}^{k} \left({b_{\min}}^{\alpha-1} \left[{b_i}^{(1-\alpha)} - {b_{i+1}}^{(1-\alpha)} \right] \right)^{h_i} \right] \\
&= \sum_{i=\min}^{k} \left[(\alpha-1)\ln(b_{\min}) h_i + h_i \ln \left[{b_i}^{(1-\alpha)} - {b_{i+1}}^{(1-\alpha)} \right] \right] \\
&= (\alpha-1)\ln(b_{\min}) \sum_{i=\min}^{k} h_i + \sum_{i=1}^{k} h_i \ln \left[{b_i}^{(1-\alpha)} - {b_{i+1}}^{(1-\alpha)} \right] \\
&= n(\alpha-1)\ln(b_{\min}) + \sum_{i=\min}^{k} h_i \ln \left[{b_i}^{(1-\alpha)} - {b_{i+1}}^{(1-\alpha)} \right] \qquad \text{(B.2)}
\end{aligned}$$

We could obtain an analytic solution for α by setting $\partial\mathcal{L}/\partial\alpha = 0$. However since the above equation is complex, we could solve for α by numerical maximization of the above equation.

However if we consider the logarithmic binning scheme where $b_i = c^{i-1}$, i.e, the bin boundaries are in powers of some constant c, then we can derive an analytic expression for α. In this case if we have k bins in the empirical data then the bin boundaries are given as $B = (1, c^1, c^2, c^3, ...c^k)$. Thus with this scheme, Equation (B.2) becomes,

$$\begin{aligned}
\mathcal{L} =& n(\alpha-1)\ln(b_{\min}) + \sum_{i=\min}^{k} h_i \ln \left[\left(c^{i-1}\right)^{(1-\alpha)} - \left(c^{i}\right)^{(1-\alpha)} \right] \\
=& n(\alpha-1)\ln(b_{\min}) + \sum_{i=\min}^{k} h_i \ln \left[c^{(i-1)(1-\alpha)} - c^{(i)(1-\alpha)} \right] \\
=& n(\alpha-1)\ln(x_{\min}) + \sum_{i=\min}^{k} h_i \left[\ln \left(c^{i(1-\alpha)} \right) + \ln \left(c^{(\alpha-1)} - 1 \right) \right] \\
=& n(\alpha-1)\ln(x_{\min}) + (1-\alpha)\ln(c) \sum_{i=1}^{k} i h_i + \ln \left[c^{(\alpha-1)} - 1 \right] \sum_{i=\min}^{k} h_i \\
=& n(\alpha-1)\ln(b_{\min}) + n \ln \left[c^{(\alpha-1)} - 1 \right] + (1-\alpha)\ln(c) \sum_{i=\min}^{k} i h_i \qquad \text{(B.3)}
\end{aligned}$$

We can get an analytic solution for α in this case using equation $\partial\mathcal{L}/\partial\alpha = 0$ to find the maximum. On solving for α we get,

$$\hat{\alpha} = 1 + \log_c \left[1 - \frac{n \ln(c)}{\left(n - \sum_{i=\min}^{k} i h_i \right) \ln(c) + n \ln(b_{\min})} \right] \tag{B.4}$$

To show that the estimator $\hat{\alpha}$ is asymptotically consistent, we generated one synthetic data set following continuous power-law distribution and two synthetic binned data sets following power-law distribution with $\alpha = 2.5$ and with lower-bound $b_{\min} = 1$. Amongst the binned data sets, the first was binned with $c = 3$, i.e, bin boundaries were $B = (1, 3, 9, 27, ...)$. The second was binned using $B = (1, 5, 125, 625, ...)$. Using continuous MLE given in [5] we made an estimate $\hat{\alpha}$ for continuous data set and using Equation (B.4) we made an estimate $\hat{\alpha}$ for both the binned data sets. We then calculated the mean-squared error as $MSE = \langle(\alpha - \hat{\alpha})^2\rangle$, where angular brackets imply averaging, which was done for 1000 iterations.

Figure B.1 shows the results for MSE plotted as a function of sample size n. We can see that the continuous MLE in [5] as well as our estimator denoted by Equation (B.4) are asymptotically consistent. On comparing to continuous estimator, we see that MSE is more for binned case. This is because we have lost information due to binning and hence we find MSE shifted above for binned data sets. Also note that the more the bin-width, the more is the loss of information and hence the more is the shift. Hence, MSE line for $c = 5$ lies above the one for $c = 3$. A detailed treatment of this aspect of binning can be found in Appendix D.1.

Appendix C

Likelihood Ratio Tests

Let p_1 and p_2 be the PDFs of two different candidate distributions. The likelihoods of given data under these distributions can be given as,

$$L_1 = \prod_{i=1}^{n} p_1(x_i), \qquad L_2 = \prod_{i=1}^{n} p_2(x_i) \tag{C.1}$$

If data points x_i, are binned into k bins using bin boundaries $B = (b_1, b_2, ..., b_k, b_{k+1})$, such that $b_j < x_i < b_{j+1}$, then we replace the probability density for x_i, i.e, $p(x_i)$, with the binned probability density, $p(b_j < x_i < b_{j+1})$. This is the only change we need to make for binned data. Taking ratio of the likelihoods given by Equation (C.1), we get,

$$R = \frac{L_1}{L_2} = \prod_{i=1}^{n} \frac{p_1(x_i)}{p_2(x_i)} \tag{C.2}$$

To get log-likelihood ratio, we simply take logarithm on both sides of above equation,

$$\mathcal{R} = \sum_{i=1}^{n} \left[\ln p_1(x_i) - \ln p_2(x_i)\right] = \sum_{i=1}^{n} \left[\ell_i^{(1)} - \ell_i^{(2)}\right] \tag{C.3}$$

where $\ell_i^{(j)} = \ln p_j(x_i)$ is the log-likelihood of a single data point x_i within distribution j.

Since, by hypothesis, the x_i, are assumed to be independent, the differences $\ell_i^{(1)} - \ell_i^{(2)}$ and hence, by the central limit theorem, their sum $\mathcal{R}$ beomes normally distributed in limit of large n, with expected variance $n\sigma^2$ where σ^2 is the variance of the data and is given as,

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \left[\left(\ell_i^{(1)} - \ell_i^{(2)}\right) - \left(\bar{\ell}_i^{(1)} - \bar{\ell}_i^{(2)}\right)\right]^2 \tag{C.4}$$

where

$$\bar{\ell}_i^{(1)} = \frac{1}{n}\sum_{i=1}^{n} \ell_i^{(1)}, \qquad \bar{\ell}_i^{(2)} = \frac{1}{n}\sum_{i=1}^{n} \ell_i^{(2)} \tag{C.5}$$

If the true expectation of the log-likelihood ratio is zero, the observed sign of $\mathcal{R}$ is purely a result of the fluctuations and cannot be trusted as an indicator of which model is preferred. The probability that the measured log-likelihood ratio has a magnitude as large as or larger than the observed value $|\mathcal{R}|$ can be quantified by a p-value,

$$\begin{aligned} p &= \frac{1}{\sqrt{2\pi n\sigma^2}}\left[\int_{-\infty}^{-|\mathcal{R}|} \mathrm{e}^{-t^2/2n\sigma^2}\,dt + \int_{|\mathcal{R}|}^{\infty} \mathrm{e}^{-t^2/2n\sigma^2}\,dt\right] \\ &= \operatorname{erfc}(|\mathcal{R}|/\sqrt{2n}\sigma) \end{aligned} \tag{C.6}$$

where erfc is the complementary error function, which is defined as,

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_{x}^{\infty} \mathrm{e}^{-t^2}\,dt \tag{C.7}$$

This p-value gives us an estimate of the probability that we measured a given value of $\mathcal{R}$ when its true value is close to zero, and thus a large p-value (say, $p > 0.1$) indicates that we cannot trust the sign of $\mathcal{R}$ and likelihood ratio test is inadequate to discriminate between the selected distributions. On the other hand, a low p-value suggests that the observed sign of $\mathcal{R}$ is not a chance of fluctuations and hence is an indicator of which model is a better fit to the data.

Since the distributions p_1 and p_2 that we are dealing with, are fixed by fitting to the same data that are the basis for the likelihood ratio test, there are correlations between the data and the log-likelihoods. Vuong [23] has shown that so long as p_1 and p_2 come from distinct, non nested families of distributions and the estimation is done by maximizing the likelihood within each family, the above results indeed hold.

For nested hypotheses such as power law and power law with exponential cut-off, if the true distribution lies in the smaller family, i.e., in this case the power law, then the best fits of both families converge to the true distribution as n becomes large. Thus $|\mathcal{R}| \to 0$ and so does σ. Consequently p-value given in Equation (C.6), tends to $0/0$ and its distribution does not obey the simple central limit theorem. A more careful analysis shows that $\mathcal{R}$ actually adopts a chi-squared distribution as n becomes large [27]. This result could be used to calculate the correct p-value. If this p-value is small, then the smaller family can be ruled out. If not, then we can say that the larger family may be needed to fit the data, but nothing can be said for certain.

Appendix D

Impact of binning

D.1 Bound on Mean-squared-error

In order to find a lower-bound on the mean-squared error, i.e, MSE, we use fisher information (see [25]) given as,

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln\mathcal{F}(X;\theta)\right] \tag{D.1}$$

Here the function $\mathcal{F}$ is the likelihood function and $\ln\mathcal{F}$ is given by B.3. Taking derivative wrt α,

$$\mathcal{I}(\alpha) = -E\left[\frac{\partial^2}{\partial\alpha^2}\mathcal{L}(X;\alpha)\right]$$

If we assume logarithmic binning scheme, we can use the log-likelihood function Equation (B.3). On solving the above double derivative we get fisher information for the logarithmic binning case to be,

$$\begin{aligned}\mathcal{I}(\alpha) &= -E\left[-\frac{nc^{1+\alpha}[\ln(c)]^2}{(c-c^{\alpha})^2}\right]\\ &= \frac{nc^{1+\alpha}[\ln(c)]^2}{(c-c^{\alpha})^2}\end{aligned} \tag{D.2}$$

For an unbiased estimator, the Cramér-Rao bound (see [7], [17]) shows that the inverse-fisher information serves as the lower-bound for the variance of an estimator,

$$\begin{aligned}Var(\hat{\alpha}) &\geq \frac{1}{n\mathcal{I}(\alpha)}\\ &\geq \frac{(c-c^{\alpha})^2}{n^2c^{1+\alpha}[\ln(c)]^2}\end{aligned} \tag{D.3}$$

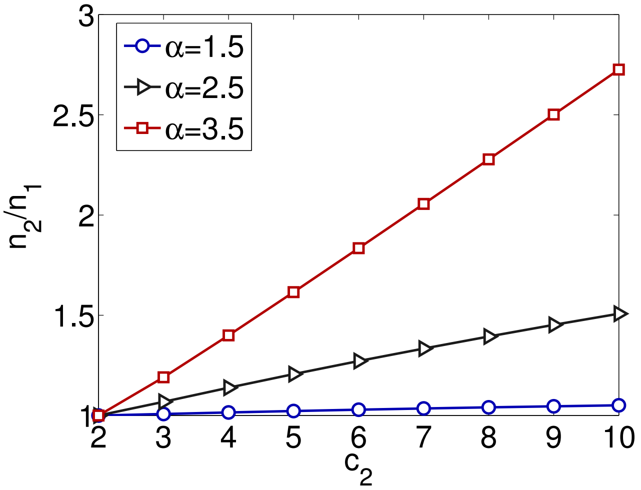


Figure D.1: Comparison of the number of elements required with two different binning strategies such that bound on MSE remains constant. Ratio of number of elements required n_2, in binning scheme c_2, to the number of elements n_1, used with binning scheme $c_1 = 2$, as a function of c_2. For $\alpha = 3.5$, we would require almost thrice as many elements if we used a logarithmic binning scheme in powers of $c_2 = 10$, instead of a logarithmic binning scheme in powers of $c_1 = 2$.

The standard error for an estimator is given as,

$$\sigma = \sqrt{Var(\hat{\alpha})} \tag{D.4}$$

Using Equation (D.3) and Equation (D.4), we can get a lower-bound on standard error as,

$$\sigma \geq \frac{(c - c^{\alpha})}{n\sqrt{c^{1+\alpha}}[\ln(c)]} \tag{D.5}$$

For sufficiently large n, we can consider the above equation to be an equality as,

$$\sigma = \frac{(c - c^{\alpha})}{n\sqrt{c^{1+\alpha}}[\ln(c)]} \tag{D.6}$$

However in the small n regime (for $c = 2$, $n \lesssim 50$), this equality won't hold, and σ becomes positively biased.

Further using the bias-variance decomposition (see [25]),

$$MSE(\hat{\alpha}) = Var(\hat{\alpha}) + [Bias(\alpha, \hat{\alpha})]^2 \tag{D.7}$$

and then combining D.3 and D.7, we get,

$$\begin{aligned} MSE(\hat{\alpha}) - [Bias(\alpha, \hat{\alpha})]^2 &\geq \frac{1}{n\mathcal{I}(\alpha)} \\ MSE(\hat{\alpha}) &\geq \frac{1}{n\mathcal{I}(\alpha)} + [Bias(\alpha, \hat{\alpha})]^2 \\ MSE(\hat{\alpha}) &\geq \frac{1}{n\mathcal{I}(\alpha)} \end{aligned} \tag{D.8}$$

Thus $\frac{1}{n\mathcal{I}(\alpha)}$ gives the bound on MSE as well.

For a given α, we can find a set of parameters (n_1, c_1) and (n_2, c_2) such that the lower-bound on MSE shown in Equation (D.8) for both parameter settings is equal. Assuming that we know n_1, c_1, c_2, we can find an equation for n_2 by using the following equation

$$\left|\frac{1}{n_1\mathcal{I}(\alpha)}\right|_{n_1,c_1} = \left|\frac{1}{n_2\mathcal{I}(\alpha)}\right|_{n_2,c_2}$$

Using above equation along with Equation (D.2), we get,

$$\frac{(c_1 - {c_1}^{\alpha})^2}{n_1^2 {c_1}^{1+\alpha}[\ln(c_1)]^2} = \frac{(c_2 - {c_2}^{\alpha})^2}{n_2^2 {c_2}^{1+\alpha}[\ln(c_2)]^2}$$

Binning Scheme	Estimate of α	Estimate of b_{min}
Unbinned [5]	1.95	7
Log Binning (1, 5, 81)	1.932	13.33
Log Binning (1, 5, 61)	1.962	3.83
Log Binning (1, 5, 41)	1.93	13.33
Log Binning (1, 5, 21)	1.985	10
Linear Binning (width=100)	1.96	1
Linear Binning (width=10)	1.963	1

Table D.1: Results for parameters α and b_{min} for different binning schemes.

On solving the above equation for n_2, we get,

$$n_2 = \frac{\ln(c_1){c_1}^{\frac{1+\alpha}{2}}(c_2 - {c_2}^{\alpha})n_1}{\ln(c_2){c_2}^{\frac{1+\alpha}{2}}(c_1 - {c_1}^{\alpha})} \tag{D.9}$$

Figure D.1 shows a plot of ratio of the number of elements of the two binning schemes, i.e, n_2/n_1 with binning schemes given by c_2 and c_1, respectively. In the figure, $c_1 = 2$ and the ratio is plotted as a function of c_2 for three distinct α values. The plot shows a linear increase in the ratio as c_2 increases, which is what we expect as the bin widths increase. For $\alpha = 3.5$, if we use logarithmic binning scheme with $c = 10$, we would require approximately more than double the number of elements as we would require with binning scheme of $c = 2$ to have the same statistical power. The point of this experiment is to show that we need to be extremely carefully when we bin our data, and we should always opt for a binning scheme with bin boundaries as tight as possible so that inference becomes more accurate.

D.2 Effect of different binning strategies

As mentioned in Chapter 1, we do not make any implicit assumptions about the spacing between the bin boundaries of the binned data. Thus it is possible to have a binning scheme with bin boundaries separated linearly or logarithmically. In this section, we experiment with different variations of such schemes and see if we get any significant deviations for the parameter estimates of the power-law form namely b_{min} and α.

In this experiment we apply different binning schemes to a continuous valued data set. Then we apply our fitting procedure as described in Chapter 3 and find the parameter estimates namely, $\hat{b}_{\min}$ and $\hat{\alpha}$. For this experiment we chose the frequency of occurrence of unique words in the novel Moby Dick [15]. This data set was analyzed in [5], and parameter estimates for the power-law model using continuous MLE were found to be, $x_{\min} = 13.33$ and $\alpha = 1.932$.

On applying different logarithmic and linear binning schemes, we get the results as given in table D.1. The logarithmic binning scheme used here is slightly different. In the binning scheme column for logarithmic binning, (a,b,n) implies that the bin boundaries start from a and go upto b in n logarithmic steps. We see that even on varying the bin spacing linearly or logarithmically, the estimates for $\hat{b}_{\min}$ and $\hat{\alpha}$ that are pretty closed to the unbinned results.