BACHELOR

A comparative study of methods to estimate and test power-law tails

Schoemaker, Marit T.A.

Award date:
2023

2WH40 Bachelor final project

A comparative study of methods to estimate and test power-law tails

Student
Marit Schoemaker
1485008

Supervisors
R.M. Castro
W.L.F. van der Hoorn

Eindhoven, July 6, 2023

Contents

1 Introduction

The population of cities, the distribution of wealth [17], the word frequency in a book [9], the frequencies of family names [22], the copies of a book sold [10] and the magnitude of earthquakes [11]. At first glance, it does not appear that these various collections of information have anything in common. But nothing could be further from the truth. In fact, these data are well modeled as samples from a power-law distribution. This phenomenon was first observed by the Italian Vilfredo Pareto in 1906 [17]. He noticed that approximately 80% of the land in Italy was owned by 20% of the Italian population. From this emerged the Pareto principle, also known as the 80/20 rule. The Pareto principle states that roughly 80% of the consequences follow from 20% of causes. Or, put another way, with 20% of our efforts, we achieve 80% of our results.

The Pareto principle is a specific form of a Pareto distribution. The Pareto distribution, in turn, is a specific form of the so-called power law distribution. The power law is a heavy tailed distribution, meaning that the distribution has heavier tails than the exponential distribution [7]. This implies that there is a large probability of obtaining extremely large values. The term 'power-law' is derived from this characteristic, as the probability of obtaining values larger than a fixed value follows a power relationship with that fixed value. In empirical phenomena, it is often observed that the power law applies only from a certain lower bound. In such cases, the distribution is said to have a power law tail.

In this thesis, we tackle the prevalent problem of testing if it is reasonable to model data as a sample from a power law distribution (in the tail). In practice, it is hard to be completely certain about this. All we can assert is that we have enough evidence not to reject the hypothesis that the data is drawn from a distribution with a power law (tail). In pursuit of this objective, several methods have been developed that determine whether data follows a power law (tail) as well as methods for estimating the parameters of power laws. However, as is commonly encountered with estimation and testing procedures, these methods are not flawless. For example, distributions that do not follow a power law distribution (in the tail) are often identified as power law distributions. In addition, parameter estimation can be completely wrong, leading to misinterpretations of the empirical data.

In this research, we will propose a new method for estimating and testing data that is believed to follow a power law distribution (in the tail). This method leverages the properties of the power law distribution, along with concepts of conditional distributions, transformations, and confidence intervals. Therefore, this new method is called the confidence interval method. To assess the efficacy of the new method, we will conduct a comprehensive comparative analysis against a well-known existing method. This existing method is referred to as the PLFit method and was developed by Clauset, Shalizi, and Newman [8]. Our aim is to present a more accurate approach for the estimation and testing procedure compared to the PLFit method.

We are going to put into practice what we have analyzed by applying our methods to a dataset of earthquakes in Limburg. Throughout the years, it has been suggested that the magnitudes of earthquakes, in general, adhere to a power law. Now, our objective is to determine if this pattern holds true for our specific dataset, and whether our methodologies yield similar conclusions.

To begin with, we will present several definitions that will be used throughout this research. This includes the formalization of the power law distribution, as well as the calibration and power of an hypothesis test. Next, we will provide a brief outline of the methods employed, state the hypothesis test, and construct a distribution with a power law tail. Once these definitions are familiar and a brief outline of the methods and hypothesis test is presented, we will explain the main problem of this research through a problem statement. Furthermore, we will formulate the main research questions that we seek to answer.
In order to compare the methods effectively, it is crucial to have a thorough understanding of how each method operates. Therefore, we provide a detailed description of the estimation and testing procedures for both the PLFit method and the confidence interval method. Once we have acquainted ourselves with the utilized methods, we analyze the performance of the estimators computed by both methods. We do this using a simulation study, which means we apply the estimation procedure to multiple datasets drawn from a power law distribution (in the tail). Following this,

we conduct a number of numerical experiments to evaluate the performance of the proposed testing methods.
After conducting the numerical results of the estimation and testing procedures, we will proceed to review the obtained findings. We will explain how these results can be interpreted and discuss their implications for the accuracy of the methods. Additionally, we will explore potential areas within the context of our research that could be interesting for further investigation. Furthermore, we will address any limitations encountered during the course of this study. Finally, we will conclude the report by presenting an overview of the study conducted and summarizing the key findings.

2 Mathematical formalism

In the previous section, we introduced the concept of Pareto and power law distributions and discussed their practical applications. We also emphasized the importance of accurate estimation and testing to extract meaningful insights from empirical data. In this section, we will present the definition of a power law distribution that will be used throughout this research. Additionally, we will state the hypothesis test. Furthermore, we will delve deeper into the previously introduced problem and address it in greater detail. We will define the main research questions that this report aims to address. Lastly, we construct a distribution with a power law tail that will be used for the analysis.

2.1 The power-law distribution

Many phenomena can be effectively modeled as samples from a distribution, and a significant number of these phenomena exhibit power law distributions. Power law distributions capture the intriguing fact that very large observations or extreme events are not terribly uncommon. A classic example illustrating this concept is the 80/20 rule, also known as the Pareto principle. According to this rule, approximately 80% of the effects come from about 20% of the causes. It highlights the unequal distribution of cause-and-effect relationships, with a small portion of causes leading to a significant portion of outcomes. In the context of power law distributions this rule exemplifies how a small number of instances or events within the 80% range can yield disproportionately large impacts or values. In other words, extreme values or significant effects are more prevalent than expected within the majority of outcomes governed by the power law distribution.

It is worth noting that unlike many other well-known distributions, a power-law does not possess a single fixed definition; rather, the properties defining a distribution as power-law can vary across different areas of research. By understanding the intricacies of power law behavior, we can gain valuable insights into diverse fields such as physics, biology, earth and planetary sciences, economics and finance, computer science, demography, and the social sciences.

We examine the population distribution of cities in the United States, using it as a real-world example that demonstrates characteristics resembling a power-law relation. The dataset we analyze includes cities with a population of 50,000 or more in 2020, obtained from the United States Census Bureau [19]. It is important to acknowledge that the power-law model is a mathematical abstraction that captures, albeit approximately, certain phenomena observed in the real world. While the model provides a useful framework for understanding and analyzing the size distribution of cities, it is essential not to confuse the mathematical representation with the complexity of actual reality.

The population distribution of U.S. cities often exhibits a heavy-tailed distribution, which means that there is a higher likelihood of extreme or rare occurrences compared to other distributions. In other words, most U.S. cities have a small population, while a few cities, like New York and Los Angeles, have much larger populations. This phenomenon is visually demonstrated in Figure 1, which displays the (empirical) complementary cumulative distribution function (ccdf) of the data. The ccdf provides information about the probability that a random variable exceeds a given threshold. In this case, Figure 1 shows the proportion of U.S. cities with populations of at least a certain size, highlighting the decreasing proportion of cities as the population threshold increases.

When data follows a power law distribution, it becomes particularly interesting to plot the ccdf on a logarithmic scale for both the x and y axes, as depicted in Figure 2. An intriguing observation is that the ccdf forms an almost straight line with a negative slope. This negative slope indicates that as the threshold value increases, the probability of observing values greater than the threshold decreases following a power-law pattern. This observation provides valuable insights into the nature and extent of power law behavior within the dataset.

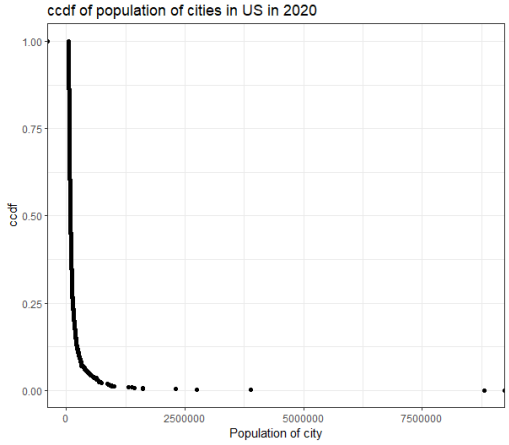


Figure 1: ccdf of the data

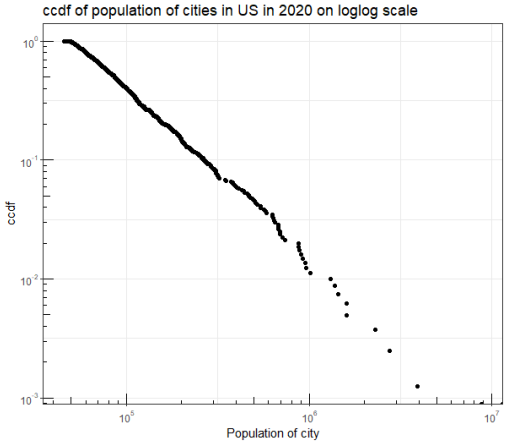


Figure 2: ccdf of the data on log-log scale

The power law is characterized by a mathematical equation in which the frequency or magnitude of an event is inversely proportional to its size or rank. A nonnegative random variable X is said to follow a power law distribution if

$$\mathbb{P}(X > x) \sim cx^{-\alpha+1}, \tag{1}$$

for $c > 0$ and $\alpha > 0$. The constant c is the normalization constant and α the constant parameter of the distribution called the scaling parameter. The "$\sim$" represents an asymptotic equivalence. It indicates that the probability $\mathbb{P}(X > x)$ is asymptotically equivalent to the term $cx^{-\alpha}$ as x grows large. This means that the ratio of the probability to the term approaches 1 as x becomes larger [15].

In the context of the power law distribution, we can further examine a specific form known as the Pareto distribution. It is named after the Italian economist Vilfredo Pareto, who first applied it to describe the distribution of wealth [17]. The Pareto distribution is supported on $[x_{min}, \infty)$ with $x_{min} > 0$ and the relationship in Eq. (1) holds with equality. The presence of the minimum value $x_{\min}$ in the Pareto distribution guarantees that no values can occur below $x_{\min}$, establishing it as the absolute lower bound. In other words, the probability of observing no value below $x_{\min}$ in the Pareto distribution is precisely one.

Therefore the ccdf of a Pareto distribution is given by,

$$\mathbb{P}(X > x) = \begin{cases} \left(\frac{x}{x_{min}}\right)^{-\alpha+1} & x \geq x_{min}, \\ 1 & x < x_{min}. \end{cases} \tag{2}$$

If X has a Pareto distribution, then in a log-log plot of $\mathbb{P}(X > x)$, the behavior is exactly linear for $x \geq x_{min}$. For this, we take the logarithm of Eq. (2),

$$\log(\mathbb{P}(X > x)) = (-\alpha + 1)(\log(x) - \log(x_{min})).$$

In conclusion, the power law is a more general term that can describe various distributions with a power-law relationship, but does not explicitly include the concept of a minimum value or cutoff, which is the case with the Pareto distribution.

In this study, we will focus on the Pareto distribution. In natural phenomena, it is often observed that the power law behavior only happens in the tail of the distribution, so only from a certain value onwards. That gives the following expression for the density function:

$$f(x) = \begin{cases} C\left(\frac{x}{x_{min}}\right)^{-\alpha} & x \geq x_{min}, \\ g(x) & 0 \leq x < x_{min}, \end{cases} \tag{3}$$

where $\alpha > 1$, $x_{min} > 0$ and $g(x) \geq 0$ uniquely define the distribution. Note that $g(x) : [0, x_{min}) \to [0, \infty)$ is such that $\int_0^{x_{min}} g(x)\, dx \leq 1$. Note that C is the normalization constant. The constant

C can be computed by solving $\int_0^{x_{min}} g(x)\ dx + \int_{x_{min}}^{\infty} C\left(\frac{x}{x_{min}}\right)^{-\alpha}\ dx = 1$. Provided $\alpha > 1$ and $g(x) = 0$ we find that $C = \frac{\alpha-1}{x_{min}}$. For modeling purpose, the scaling parameter is typically found within the range of $2 < \alpha < 3$, which is considered to be the most interesting range. However, it is worth noting that occasional deviations from this pattern may occur in certain cases [8]. In the case when $g(x) = 0$, $f(x)$ represents the probability density function of a pure Pareto distribution. It's worth noting that not all distributions satisfying Eq. (1), which is an asymptotic statement, can be represented by Eq. (3). The latter requires the tail of the distribution to strictly follow a pure Pareto distribution. In contrast, Eq. (1) only requires the condition that $\mathbb{P}(X > x)$ approaches a constant as x tends to infinity.

Remark: In practical scenarios, observing a precise manifestation of a Pareto distribution as described in Equation (2) is uncommon. To address situations where the exact power law behavior may not be observed, the class of regularly varying distributions is considered an alternative to Pareto distributions. This class provides a more realistic framework for modeling phenomena that exhibit power-law-like behavior but do not strictly conform to a strict Pareto distribution. The class of regularly varying distributions contains all distributions whose ccdf is given by,

$$\bar{F}(x) = l(x)x^{-\alpha+1},$$

where $l(x)$ is a slowly varying function that allows for a power law in a looser sense [20]. A slowly varying function is a mathematical concept used in the field of analysis to describe functions that change relatively slowly compared to another variable. More formally, a function $l(x)$ is said to be slowly varying if, for any positive constant t, the ratio $\frac{l(tx)}{l(x)}$ approaches 1 as x approaches infinity [13]. Note that this distribution also does not satisfy Eq. (1), the regularly varying distributions relax the strictness of the power-law behavior by introducing the slowly varying function $l(x)$.

For the purposes of this report, we define a distribution as having a power-law distribution if it exhibits the probability density function given by Equation (3).

2.2 Hypothesis test, calibration and power

In this thesis, our goal is to address the prevalent problem of determining whether it is reasonable to model data as a sample from a power law distribution. To accomplish this, we make the assumption that the data follow a specific underlying distribution denoted as F. Our objective is to investigate and assess whether this distribution can be characterized as a power law distribution. In order to achieve this, we employ a **hypothesis test**, which is a statistical procedure used to evaluate the validity of a claim based on sample data.

The purpose of our hypothesis test is to determine the plausibility of regarding the observed data as a sample from a power-law distribution, as defined in Eq. (3). We formulate the null and alternative hypotheses as follows:

$$H_0 : F \text{ is a power law for some } \alpha > 1, x_{min} > 0 \text{ and } g(x) \geq 0,$$
$$H_1 : F \text{ is not a power law.}$$

By conducting this hypothesis test and analyzing the statistical evidence obtained from the data, we aim to determine whether the underlying distribution can indeed be characterized as a power law distribution. To assess the significance of our findings, we rely on a statistical measure known as the p-value.

In hypothesis testing, the data serves as the input for the testing procedure, and the typical output is a p-value. A p-value is "the probability, computed assuming that H_0 is true, that the test statistic would take a value as extreme or more extreme than that actually observed." (Quote from Moore 2007, [8]). The p-value is a statistical measure that is used to determine the strength of evidence against a null hypothesis in a hypothesis test. It is commonly used in statistical hypothesis testing to assess the significance of the results obtained from an experiment or study.

Therefore we would like the p-value such that, assuming H_0 is true, not to reject H_0 that often, only with probability δ. This probability represents the type I error rate, which is the likelihood of incorrectly rejecting H_0 when it is actually true. Our objective is to ensure that the probability of rejecting a true null hypothesis is equal to δ. To achieve this, we reject the null hypothesis when the observed p-value $\leq \delta$. The only way this condition is satisfied for any δ value is when the p-value follows a standard uniform distribution.

Therefore, when formulating it formally, when the test statistic follows a continuous distribution under the null hypothesis, the resulting p-value should have a uniform distribution between 0 and 1. This means that for a **well-calibrated** hypothesis test, the p-value, which is a random variable, is uniformly distributed in the interval $[0, 1]$ [2].

Ensuring a well-calibrated test is important because it ensures the validity of hypothesis testing. If a test is poorly calibrated, it may lead to incorrect inferences and flawed decision-making. Therefore, it is essential to assess the calibration of the test by examining the distribution of p-values under the null hypothesis.

In hypothesis testing, the null hypothesis is typically assumed to be true, and the general goal is to gather evidence that either supports or contradicts it. A smaller p-value indicates stronger evidence against the null hypothesis, suggesting that the observed data is unlikely to occur by chance alone if the null hypothesis were true.

Now, when considering the **power** of a test, we are interested in its ability to detect a true effect when it exists. If the alternative hypothesis is true, and there is indeed a meaningful difference or effect in the observed data, we would like the test to have high power, meaning it has a high probability of correctly rejecting the null hypothesis. This means that the test is sensitive enough to detect the true effect, even in the presence of random variability in the data.

2.3 Problem statement

In this research, our main objective is to tackle the prevalent problem of determining whether it is reasonable to model data as a sample from a power law distribution. To achieve this, we specifically focus on comparing and evaluating various methods for parameter estimation and hypothesis testing specifically for statistical distributions with a power law (in the tail). Under the assumption that the underlying distribution of the data follows a power law, our research focuses on exploring various estimation methods that accurately estimate the parameters α and x_{min}, that govern the power law behavior. Additionally, we address the challenge of conducting hypothesis tests to validate this power law hypothesis.

The research aims to investigate and compare two methods, the PLFit method and the confidence interval method, for estimating the parameters α and x_{min} in power law distributions. It seeks to determine which method performs better in accurately estimating these parameters and capturing the tail behavior of the distribution. Additionally, the research aims to compare three methods, the PLFit method, confidence interval method and a combined method, for hypothesis testing in power law distributions to identify the most effective approach for validating the power law hypothesis. By conducting a thorough analysis and comparison of these methods, the research intends to provide insights into the most reliable techniques for parameter estimation and hypothesis testing in power law distributions.

This leads to the main research questions of this report:

1. Which of the two methods provides more accurate estimates of the parameters α and x_{min}?
2. Among the three methods, which one yields more reliable results in hypothesis testing?

Note: The data under consideration is assumed to be real-valued. This assumption allows for a continuous range of possible values, enabling us to perform various statistical analyses and order the data. Once we have collected samples, we assume that they are almost surely unique in the case of real-valued data. This means that the probability of encountering duplicate values is 0. However, it should be noted that in the case of discrete data, duplicates may exist. Although duplicates can arise in the case of discrete data, we acknowledge that handling ties or duplicated values is beyond the scope of this thesis. The focus of this research is primarily on the ordering and analysis of the data, rather than explicitly addressing methods for breaking ties or handling duplicates.

2.4 Power law distributions used for analysis

While it would provide a comprehensive analysis to explore all possible power law distributions, for the purpose of this research, we focus on two specific cases to assess the performance of the estimators and evaluate the calibration of the hypothesis test. One is a pure Pareto distribution, meaning that in Equation (3) $g(x) = 0$. The other is a distribution with a power law tail, so $g(x) \neq 0$. We chose to use a piecewise probability distribution to create the power law tail. For values of x that are less than x_{min}, a uniform distribution is used. On the other hand, for values of x that are greater than or equal to x_{min}, a Pareto distribution is used. These distributions will be used in all three methods with the same parameters so that the methods are analyzed under the same conditions.

The probability density function of the piecewise probability distribution is denoted by $h(x)$ and given by,

$$h(x) = \begin{cases} c_1 \frac{1}{x_{min}} & x < x_{min}, \\ c_2 x^{-\alpha} & x \geq x_{min}. \end{cases} \tag{4}$$

Here, c_1 and c_2 are the normalization constants. Their purpose is to ensure that the probability distribution function $h(x)$ satisfies the properties of a valid probability distribution, namely, that the total probability over the entire range of the random variable x is equal to 1.

At the point $x = x_{min}$ we want the function values of the left and right limits of x_{min} to be equal. By ensuring that the left and right limits are equal, we create a continuous transition at

that point, making it more challenging to precisely determine the location of x_{min}. Paradoxically, this challenge is beneficial because we want our estimation methods to remain effective in accurately estimating x_{min}. Then the following relation between c_1 and c_2 is obtained, $c_2 = c_1 x_{min}^{\alpha-1}$. Substituting this in Equation (4) and computing the integral of $h(x)$ being equal to 1 gives,

$$\begin{aligned}
\int_0^{\infty} h(x)\, dx &= 1, \\
\int_0^{x_{min}} c_1 \frac{1}{x_{min}}\, dx + \int_{x_{min}}^{\infty} c_1 x_{min}^{\alpha-1} x^{-\alpha}\, dx &= 1, \\
c_1 + c_1 x_{min}^{\alpha-1} \left[\frac{1}{-\alpha+1} x^{-\alpha+1} \right]_{x_{min}}^{\infty} &= 1, \\
c_1 - c_1 \frac{1}{\alpha+1} &= 1, \\
c_1 &= \frac{\alpha-1}{\alpha}, \\
c_2 &= \frac{\alpha-1}{\alpha} x_{min}^{\alpha-1}.
\end{aligned}$$

It is important to note that while the support of the power law distribution theoretically extends to infinity, in practice, the observable domain may be limited by the nature of the data or constraints of the phenomenon being studied. In Figure 3, we present the theoretical distribution on a log-log scale. The domain considered for this visualization is $[0, 100000]$, showcasing the behavior of the piecewise probability distribution within this specific domain. The parameters used in the distribution are $\alpha = 2.3$ and $x_{min} = 8$. The vertical dotted line indicates the separation between the uniform and Pareto distributions at x_{min}. Furthermore, the horizontal dotted line represents $\mathbb{P}(X > 8) = 1/\alpha$, which is derived from the complementary cumulative distribution function (ccdf) using Equation (4).

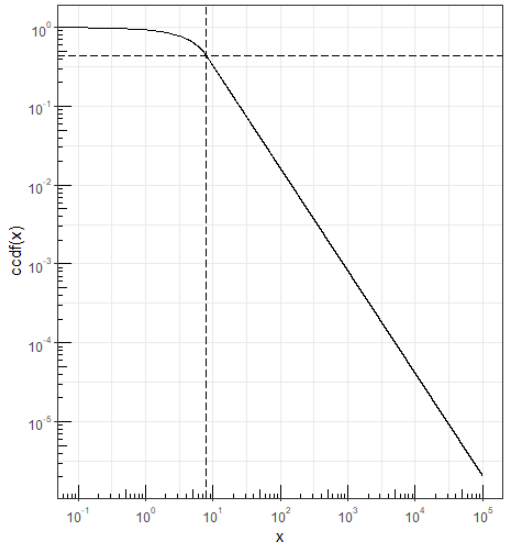


Figure 3: The ccdf of a piecewise uniform and Pareto distribution on log-log scale

3 The PLFit method

In the previous section, the definition of the power-law that is used in this research was discussed. In addition, as also discussed earlier, one of the goals of this research is to define a new method for estimating the scaling parameter and lower bound for the power law and a new goodness-of-fit test with more power than the testing procedure of the PLFit method for datasets with little data.

To be able to compare this new method with the PLFit method, it is important to first develop a better understanding of the PLFit method. Therefore, in this section a detailed explanation will be given for the estimation method and testing procedure of the PLFit method.

3.1 The estimation method

The estimation method is used when the empirical data is from a power-law, but the scaling parameter α and the lower bound x_{min} as defined in Equation (3) are unknown. First, the estimation of α will be explained after which the estimation of x_{min} will be described. The reason for following this order is that the estimation of α depends on the estimator of x_{min}. The method used for the estimation of x_{min} will also be used in the testing procedure, which will be discussed in Section 3.2.

The method leverages the concepts of conditional distributions and transformations to convert the problem of estimating x_{min} into a statistical estimation task that can be more easily addressed. In this section, we make the assumption that $\{X_i\}_{i=1}^n$ represents an independent and identically distributed sample drawn from a distribution with a power law tail, denoted by $F(x)$. Under this assumption, we investigate the behavior of a specific subset within $\{X_i\}_{i=1}^n$ by conditioning on the event that each X_i is greater than or equal to a specified threshold $\tau \in [0, \infty)$. The specific subset, which is conditioned on X_i being greater than or equal to τ, an independent and identically distributed sample from $F_\tau(x) = \mathbb{P}(X \leq x \mid X > \tau)$. In Section 4 we will see that this claim is reasonable, and for now we assume it.

3.1.1 Estimating the scaling parameter α

Often, in practical scenarios, both α and x_{min} are unknown. As already mentioned, the estimation of α depends on the estimator of x_{min}. We define $\hat{\alpha}(x)$ as the estimator of α based on all datapoints larger or equal than x, denoted as a function of x. To describe the overall procedure it is helpful to explain how one can estimate α when an estimate $\hat{x}_{min}$ is given. Therefore the estimator we will derive in this section is denoted by $\hat{\alpha}(\hat{x}_{min})$.

The method of choice for fitting parametrized models such as power-law distributions to observed data is the method of maximum likelihood. The maximum likelihood estimator is consistent under certain regularity conditions, meaning that the MLE converges in probability to the true value [21]. These conditions ensure that the pdf is sufficiently smooth and well-behaved for the purpose of maximum likelihood estimation.

Given a dataset $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ that is sampled from a power-law distribution as defined in Eq. (3), the goal is to find the value of α for the power-law distribution that is most likely to have generated the data using the given value for $\hat{x}_{min}$.

Let $k(z)$ be the function that denotes the number of observations in the dataset $\mathbf{x}$ with a value of at least z, where z is a variable. Therefore, $k(z)$ is given by,

$$k(z) = \sum_{i=1}^{n} \mathbb{1}\{x_i \geq z\}.$$

In this context, we are going to construct an estimator for α based only on datapoints with values larger than or equal to $\hat{x}_{min}$, k is set to

$$k = k(\hat{x}_{min}) = \sum_{i=1}^{n} \mathbb{1}\{x_i \geq \hat{x}_{min}\}, \tag{5}$$

where we start with a given value for $\hat{x}_{min}$, that is needed for the estimation of α.

We will use the order statistics sorted in decreasing order for the estimation of α. The order statistics are denoted by $x_{(i)}$. A dataset is sorted in decreasing order when $x_{(i)} \geq x_{(i+1)}$ for all $i \in \{1, \ldots, n-1\}$. Thus $x_{(1)}$ is the largest value and $x_{(n)}$ the smallest value in the dataset.

The MLE of α based on the restricted data (where $x \geq \hat{x}_{min}$) and assuming that $\hat{x}_{min} \geq x_{min}$ is the value of α that maximizes the likelihood function. The likelihood function is given by,

$$\mathcal{L}_k(\alpha) = \prod_{i=1}^{k} f(x_{(i)}) = \prod_{i=1}^{k} \frac{\alpha - 1}{\hat{x}_{min}} \left(\frac{x_{(i)}}{\hat{x}_{min}} \right)^{-\alpha}.$$

This is only the likelihood function if $\hat{x}_{min} \geq x_{min}$ and relies on the claim of the conditional distribution made in Section 3.1.

In practice, calculations are often done using the logarithm of the likelihood function, which is called the log-likelihood function, denoted by ℓ_k. The log-likelihood function can be used as it has the maximum at the same place. So maximizing the log-likelihood gives the same answer as maximizing the likelihood. Then the following can be obtained,

$$\ell_k(\alpha) = \log(\mathcal{L}_k(\alpha)) = k \log(\alpha - 1) - k \log(\hat{x}_{min}) - \alpha \sum_{i=1}^{k} \log \left(\frac{x_{(i)}}{\hat{x}_{min}} \right).$$

The maximum of this function can be found when the derivative with respect to α is equal to 0,

$$\frac{\partial \ell}{\partial \alpha} = \frac{k}{\alpha - 1} - \sum_{i=1}^{k} \log \left(\frac{x_{(i)}}{\hat{x}_{min}} \right) = 0.$$

Therefore setting $\frac{\partial \ell}{\partial \alpha} = 0$ and solving for α give us the following MLE for the scaling parameter,

$$\hat{\alpha}(\hat{x}_{min}) = 1 + k \left(\sum_{i=1}^{k} \log \left(\frac{x_{(i)}}{\hat{x}_{min}} \right) \right)^{-1}, \tag{6}$$

The MLE of α is denoted by $\hat{\alpha}(x)$, the hatted symbols denote estimators derived from data. This MLE estimator for α is known as the Hill estimator [6]. Now it is clear why the estimation method of α depends on the estimator of x_{min}.

Since the dataset follows a power law distribution for $x \geq x_{min}$. There are k observations in this dataset that are greater than or equal to $\hat{x}_{min}$. Which means that in the decreasing ordered statistics $\hat{x}_{min} = x_{(k)}$ and that gives the following formula for $\hat{\alpha}$,

$$\hat{\alpha}(x_{(k)}) = 1 + k \left(\sum_{i=1}^{k} \log \left(\frac{x_{(i)}}{x_{(k)}} \right) \right)^{-1}. \tag{7}$$

3.1.2 Estimating the lower bound x_{min}

When x_{min} is unknown, it must be estimated before the estimate for scaling parameter α can be calculated, as explained earlier. Therefore, it will first be necessary to determine from which point in the data the power law model is valid. In order to determine an accurate value for α, there must also be an accurate method for determining an estimate for x_{min}. Choosing an estimate for x_{min} that is too low can result in a biased estimate of the scaling parameter because it tries to fit a power-law model to non-power-law data. Conversely, selecting an estimate for x_{min} that is too high means that legitimate data points with $x_i < \hat{x}_{min}$ are discarded, which leads to increased statistical error on the scaling parameter [8]. The discarded data points may contain valuable information about the tail behavior of the distribution. When $\hat{x}_{min}$ is too high, the remaining dataset can be relatively small. As a result, the estimation of α can be heavily influenced by the specific values and characteristics of the observed data points. As a consequence, this can lead

to biased estimates of α that do not accurately represent the true underlying parameter of the observed data.

The general idea of the method is that $\hat{x}_{min}$ will be chosen such that the probability distribution of the observed data resembles the best-fit power-law as much as possible. The measure used to determine the distance between the two probability distributions is the Kolmogorov–Smirnov statistic, also known as the KS statistic. The statistic is defined as the maximum value of the absolute difference between two cumulative distribution functions [18]. This maximum value is denoted by $D(F(x), G(x))$ and is given by,

$$D(F(x), G(x)) = \max_{x \in \mathbb{R}} |F(x) - G(x)|,$$

where $F(x)$ and $G(x)$ are both cumulative distribution functions.

This statistic is often used with the cumulative distribution function of the observed data and the cumulative distribution function of the fitted model, and thereby see how well the fitted model resembles the observed data. An example of this is shown in Figure 4. The blue line is an empirical CDF of a sample with $n = 30$ sampled from a normal distribution with mean 0.5 and standard deviation 1.1. The red line is the CDF of this normal distribution and the black arrow is the KS statistic. It can be seen that for that value of x the distance between the two probability functions is maximized.

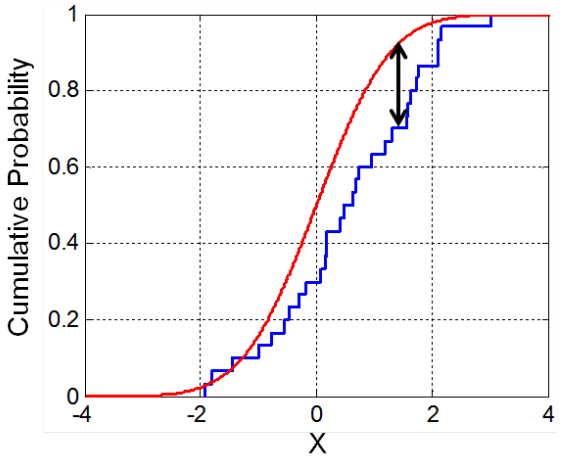


Figure 4: Visualization of the KS statistic

In the context of estimating x_{min}, the KS statistic is also used that way. Consider a dataset, denoted as $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ which has been sampled from a power-law distribution as defined in Eq. (3). The dataset is assumed to be ordered in decreasing order, denoted by $x_{(1)} \geq x_{(2)} \geq \ldots \geq x_{(n)}$. Each value $x_{(i)}$ in the ordered dataset is considered a potential candidate for $\hat{x}_{min}$. For each candidate $x_{(i)}$, we compute $\hat{\alpha}(x_{(i)})$, which represents the estimate of α based on all the data points larger than or equal to $x_{(i)}$. Once we have the estimate $\hat{\alpha}(x_{(i)})$, we can compare the conditional distribution of the tail using this estimated scaling parameter with the empirical distribution of the observed data points that are larger than or equal to $x_{(i)}$. The conditional distribution considers the values above $x_{(i)}$ and is denoted by $S_{x_{(i)}}(x)$. The empirical distribution is denoted by $P_{x_{(i)}}(x)$.

The functions of $S_{x_{(i)}}(x)$ and $P_{x_{(i)}}(x)$, where $x \geq x_{(i)}$ and $k = k(x_{(i)})$, are given by,

$$S_{x_{(i)}}(x) = \frac{1}{k} \sum_{j=1}^{k} \mathbb{1}\{x_{(j)} \leq x\}, \quad P_{x_{(i)}}(x) = 1 - \left(\frac{x}{x_{(i)}}\right)^{1-\hat{\alpha}(x_{(i)})}.$$

To compare the two distributions, we use the KS statistic. By varying the candidate values of $x_{(i)}$ and calculating the KS statistic for each, we can identify the value of $\hat{x}_{min}$ that minimizes

the KS statistic. This corresponds to the candidate $x_{(i)}$ that provides the best match between the conditional distribution of the tail, based on the estimated scaling parameter $\hat{\alpha}(x_{(i)})$, and the empirical distribution of the observed data points above $x_{(i)}$.

Then the KS statistic is calculated for $x \geq x_{(i)}$ in the following way,

$$D_{x_{(i)}}(S_{x_{(i)}}(x), P_{x_{(i)}}(x)) = \max_{x \geq x_{(i)}} |S_{x_{(i)}}(x) - P_{x_{(i)}}(x)|. \tag{8}$$

Now, since this procedure is done for every $x_{(i)}$ in the dataset, n KS statistics are obtained. The $x_{(i)}$ for which the KS statistic is minimized is chosen as $\hat{x}_{min}$, and $\hat{k}$ is defined to be the number of observations with value at least $\hat{x}_{min}$, i.e.

$$\hat{x}_{min} = \arg\min_{x_{(i)}} \left(D_{x_{(i)}}\right), \quad \hat{k} = k(\hat{x}_{min}). \tag{9}$$

Since for this value of $x_{(i)}$ the probability distribution of the observed data resembles the best-fit power-law the most.

3.2 The testing procedure

Now that the estimation method is explained, we can continue with the testing procedure. Recall that we will test whether the observed data follows a power law in the tail for some α and x_{min}. A goodness-of-fit method will be used to test this. The method that will be used is the so-called Bootstrap method. This is a method based on generating data samples from the assumed distribution of the observed data and compare it in some way with the real observed data. A pseudo-code for the Bootstrap method is given in Algorithm 1 and will be further explained below.

Algorithm 1: A Bootstrap approach for GoF testing power-law

Input : Data $\{x_1, \ldots, x_n\}$; $B \in \mathbb{N}$, number of Bootstrap samples

Compute $\hat{\alpha}$ and $\hat{x}_{min}$ for the observed data using PLFit method

Compute $D = \max_{x \geq \hat{x}_{min}} |S_{\hat{x}_{min}}(x) - P_{\hat{x}_{min}}(x)|$

for $b \leftarrow 1$ **to** B **do**

  Generate $\mathbf{Y}^{(b)} = (Y_1^{(b)}, \ldots Y_n^{(b)})$, the bootstrap samples, detailed method explained below

  Compute $\hat{\alpha}^{(b)}$ and $\hat{x}_{min}^{(b)}$

  Compute $D^{(b)} = \max_{x \geq \hat{x}_{min}^{(b)}} |\frac{1}{k} \sum_{i=1}^{k} \mathbb{1}\{Y_i^{(b)} \leq x\} - P_{\hat{x}_{min}^{(b)}}(x)|$

end

Output: Bootstrap p-value computed as $\frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\{D^{(b)} > D\}$

As can be seen in Algorithm 1, the first step is to fit the empirical data to the power-law model using the methods explained earlier in this section. The KS statistic will be calculated using the fitted model. Then, a large number of bootstrap samples will be drawn from the model that is fitted to the data. Generating these bootstrap samples needs some more explanation. A random variable Y is created that follows the fitted power-law for values above x_{min} and resembles the empirical distribution for values below x_{min}. This means Y will follow the following distribution,

$$Y = \begin{cases} \text{Pareto}(\hat{\alpha}, \hat{x}_{min}) & \text{with probability } \frac{k}{n}, \\ \text{Uniform}(x_{(k+1)}, \ldots, x_{(n)}) & \text{with probability } 1 - \frac{k}{n}, \end{cases}$$

where $x_{(i)}$ is still the order statistics sorted in decreasing order and $k = k(\hat{x}_{min})$ the number of observations where $x_i \geq \hat{x}_{min}$. When sampling an element from Y, with probability $\frac{k}{n}$ the element is sampled from the pure power-law, so the sampled value will be larger or equal to $\hat{x}_{min}$. And with probability $1 - \frac{k}{n}$, the element is sampled from the empirical distribution of the observed data, which is not assumed to be a power-law. In this case, the sampled value will be smaller than $\hat{x}_{min}$.

We sample B bootstrap samples with length n from distribution Y. For each bootstrap sample, $\hat{\alpha}^{(b)}$ and $\hat{x}_{min}^{(b)}$ will be computed using the PLFit method. And the KS statistic for this fitted power-law model will be computed. By estimating the parameters for each bootstrap sample assuming the null hypothesis is true, we are creating a null distribution of the test statistic. The null distribution represents what we would expect to observe if the null hypothesis were true.

The p-value can be computed as follows,

$$p\text{-value} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\{D^{(b)} > D\}.$$

In words, the p-value is defined to be the fraction of bootstrap KS distances that are larger than the empirical KS distance.

When the p-value is large, it means that with low probability, a bootstrap sample generated from the fitted distribution has a KS distance closer to that distribution than the KS distance of the original dataset. This suggests that there is a high probability that the original data, which is assumed to have a power-law distribution in the tail, comes from the fitted distribution. Any differences between the fitted distribution and the original data can be attributed to the random

sampling process, indicating that the observed data is likely to exhibit a power-law distribution in the tail.

Conversely, when the p-value is small, it indicates that with high probability, a bootstrap sample has a KS distance closer to the fitted distribution than the KS distance of the original dataset. In this case, it is not likely that the data comes from a distribution with a power-law tail, as there is a high probability that the bootstrap data is closer to a power-law distribution. The small p-value provides strong evidence against the null hypothesis, suggesting that the original data does not exhibit a power-law distribution in the tail as hypothesized.

According to Clauset et al., when p-value ≤ 0.1, we decide that the tail of the distribution does not follow a power-law distribution [8]. So the null hypothesis is rejected. However, the fixed-significance testing at level 0.05 is chosen for the purpose of comparing tests. It is important to note that the determination of when the p-value is small enough to conclude that the tail of the distribution does not follow a power law depends on the context.

4 Confidence interval method

The well-known PLFit method, discussed in Section 3, is used for fitting the correct power law to empirical distributions. In this section, a new method called the confidence interval method is proposed, which differs from the PLFit method in how it estimates x_{min}. The name of the method, "confidence interval", will become clear as we explain how the lower bound x_{min} is estimated.

Similar to the PLFit method, the estimation method of the confidence interval method assumes that the data follows a power law, but the values of α and x_{min} are unknown. Once the estimation method is explained, we will also discuss the testing procedure, which determines whether the observed data follows a power law.

4.1 Estimating the x_{min} and α

As already mentioned this method differs from the PLFit method in estimating x_{min}. However, this method leverages on the same conditional distribution argument as introduced Section in 3. In this section we will argue why the argument is reasonable.

The core idea of the method is to introduce a parameter, denoted as τ, which represents a potential candidate for x_{min}. By varying τ over a range of values, the method constructs a set of confidence intervals for the power law exponent α, which characterizes the shape of the distribution. When $\mathbf{X} = (X_1, \ldots, X_n)$ are i.i.d. random variables from a power law (with parameters α, x_{min} and g in Eq. (3)), and $\tau \geq x_{min}$ the distribution of the data restricted to the the points for which $X_i \geq \tau$ is i.i.d. from a Pareto distribution, for which we know how to construct good confidence statements for α. Therefore, for any $\tau \geq x_{min}$ the resulting confidence intervals are likely to all contain the true value α. However, when $\tau < x_{min}$ there is a mismatch, and this will often present itself as incompatible confidence intervals. The proposed method leverages on this idea.

Suppose that $\mathbf{X} = (X_1, \ldots, X_n)$ are i.i.d. random variables from a distribution with a power law tail, denoted by $F(x)$, with probability density function $f(x)$ as defined in Section 2.1:

$$f(x) = \begin{cases} \frac{\alpha-1}{x_{min}} \left(\frac{x}{x_{min}}\right)^{-\alpha} & x \geq x_{min}, \\ g(x) & 0 \leq x < x_{min}. \end{cases}$$

Take $\tau \in [0, \infty)$ and define $\kappa_\tau = \sum_{i=1}^{n} \mathbb{1}\{X_i \geq \tau\}$ as the number of random variables greater than or equal to τ, which makes κ_τ a random variable as well. Consider the unique strictly increasing map $\pi^* : \{1, \ldots, \kappa_\tau\} \to \{1, \ldots, n\}$ such that:

- We define $S_{\pi^*} = \{\pi^*(i) : i \in \{1, \ldots, \kappa_\tau\}\}$, where
- $\forall j \in S_{\pi^*} : X_j \geq \tau, \forall j \notin S_{\pi^*} : X_j < \tau$.

Note that for a given k the possible maps π are in one-to-one correspondence with the subsets of $\{1, \ldots, n\}$ with cardinality k, and there are exactly $\binom{n}{k}$ such sets. Let Γ_k denote all the possible maps π satisfying the conditions above (so that $|\Gamma_k| = \binom{n}{k}$). For now we assume that the value for κ_τ is given and equal to k.

Define $\mathbf{Y}_\tau = (Y_1, \ldots Y_k) = (X_{\pi^*(1)}, \ldots, X_{\pi^*(k)})$.

The set Γ_k provides a unique mapping between $\mathbf{Y}_\tau$ and $\mathbf{X}$ so that $\mathbf{Y}_\tau$ contains only random variables from $\mathbf{X}$ that are greater than or equal to τ. To be able to derive calculations with random sample $\mathbf{Y}_\tau$ we need to have that $(Y_1, \ldots, Y_k)$ conditioned on $\kappa_\tau = k$ are independent and identically distributed samples from $F_\tau(x)$, where $F_\tau(x)$ is the conditional distribution of $F(x)$ given $X \geq \tau$. I.e. $F_\tau(x) = \mathbb{P}(X \leq x \mid X \geq \tau)$.

Therefore, the first claim to be proven is:

Claim 1: $Y_1, \ldots, Y_k$ conditioned on $\kappa_\tau = k$ are independent and identically distributed samples from $F_\tau(x)$.
Proof Claim 1:
We need to show that Y_i, given $\kappa_\tau = k$, is sampled from $F_\tau(x)$ and that the joint distribution factors into the product of the individual conditional distributions. Let $y_1, y_2, \ldots, y_k$ be specific values. Then, the joint distribution of $Y_1, Y_2, \ldots, Y_k$ conditioned on $\kappa_\tau = k$ can be written as:

$$\begin{aligned}
\mathbb{P}(Y_1 \leq y_1, \ldots, Y_k \leq y_k \mid \kappa_\tau = k) &= \frac{\mathbb{P}(X_{\pi^*(1)} \leq y_1, \ldots, X_{\pi^*(k)} \leq y_k, \kappa_\tau = k)}{\mathbb{P}(\kappa_\tau = k)} \\
&= \frac{\mathbb{P}(\exists \pi \in \Gamma_k : X_{\pi(1)} \leq y_1, \ldots, X_{\pi(k)} \leq y_k, \{\forall j \in S_\pi : X_j \geq \tau, \forall j \notin S_\pi : X_j < \tau\})}{\mathbb{P}(\exists \pi \in \Gamma_k : \{\forall j \in S_\pi : X_j \geq \tau, \forall j \notin S_\pi : X_j < \tau\})} \\
&= \frac{\sum_{\pi \in \Gamma_k} \mathbb{P}(X_{\pi(1)} \leq y_1, X_{\pi(1)} \geq \tau, \ldots, X_{\pi(k)} \leq y_k, X_{\pi(k)} \geq \tau, \{\forall j \notin S_\pi : X_j < \tau\})}{\sum_{\pi \in \Gamma_k} \mathbb{P}(\forall j \in S_\pi : X_j \geq \tau, \forall j \notin S_\pi : X_j < \tau)} \\
&= \frac{\sum_{\pi \in \Gamma_k} \prod_{i=1}^{k} \mathbb{P}(X \leq y_i, X \geq \tau) \mathbb{P}^{n-k}(X < \tau)}{\sum_{\pi \in \Gamma_k} \prod_{i=1}^{k} \mathbb{P}(X \geq \tau) \mathbb{P}^{n-k}(X < \tau)} \\
&= \ldots = \prod_{i=1}^{k} \mathbb{P}(X \leq y_i \mid X \geq \tau).
\end{aligned}$$

Now we have shown that $Y_1, \ldots, Y_k$ conditioned on $\kappa_\tau = k$ are independent and identically distributed samples from $F_\tau(x)$.

Claim 2: If $\tau \geq x_{min}$ then $F_\tau(x)$ has density $f_\tau(x) = \frac{\alpha-1}{\tau}(\frac{x}{\tau})^{-\alpha} \mathbb{1}\{x \geq \tau\}$
Proof Claim 2:
The conditional distribution function $F_\tau(x)$, for $x \geq \tau$, is given by:

$$F_\tau(x) = \mathbb{P}(X \leq x \mid X \geq \tau) = \frac{\mathbb{P}(X \leq x, X \geq \tau)}{\mathbb{P}(X \geq \tau)} = \frac{F(x) - F(\tau)}{\mathbb{P}(X \geq \tau)}.$$

For any $x \geq \tau$, the conditional density function $f_\tau(x)$ is given by:

$$f_\tau(x) = \frac{f(x)}{\mathbb{P}(X \geq \tau)},$$

where $f(x)$ is the original density function of a distribution with a power law tail. And we find that for $\tau \geq x_{min}$,

$$f_\tau(x) = \frac{f(x)}{\mathbb{P}(X \geq \tau)} = \frac{\frac{\alpha-1}{x_{\min}} \left(\frac{x}{x_{\min}}\right)^{-\alpha}}{\left(\frac{\tau}{x_{\min}}\right)^{1-\alpha}} = \frac{\alpha - 1}{\tau} \left(\frac{x}{\tau}\right)^{-\alpha} \mathbb{1}\{x \geq \tau\}. \tag{10}$$

Hence, Claim 2 is proven: If $\tau \geq x_{min}$ then $F_\tau(x)$ has density $f_\tau(x)$ derived in equation (10).

Using the result of this claim the estimator of α can be derived. The method used is the maximum likelihood estimation, which is also used in the PLFit method explained in Section 3. Again we define $\hat{\alpha}(x)$ as the estimator based on all datapoints larger or equal than x. The following estimator for α is obtained, where $k = k(\tau)$,

$$\hat{\alpha}(\tau) = 1 + k \left(\sum_{i=1}^{k} \log \left(\frac{Y_i}{\tau} \right) \right)^{-1}. \tag{11}$$

Now that the claims are proven and the estimator for α is derived, we can start estimating the lower bound x_{min}. As already mentioned, τ represents a potential candidate for x_{min}. We consider values for τ in the interval of $[0, \infty)$. As we deal with a finite dataset, this results in setting τ equal to every value X_i in $\mathbf{X}$. The rest of the estimation method we will work with a given value for k.

Note that k depends on τ, so by setting τ equal to every X_i in $\mathbf{X}$ we actually consider k from 1 to n.

The method begins by transforming the random variable Y_i into a new variable Z_i using a logarithmic transformation, $Z_i = \log(\frac{Y_i}{\tau})$. By doing this, the distribution will be transformed into an exponential distribution which is easier to find the distribution of α.

The exponential distribution can be obtained in the following way. Since random variable Y_i is said to follow a power law for $x \geq \tau$, it has the following complementary cumulative distribution function,

$$\mathbb{P}(Y_i > x) = \left(\frac{x}{\tau}\right)^{1-\alpha}, \tag{12}$$

this can be found by integrating the density function $f_\tau(x)$ or by elaborating the equation of $F_\tau(x)$. For the complementary cumulative distribution function of the random variable Z_i, the following can be obtained,

$$\begin{aligned}\mathbb{P}(Z_i > x) &= \mathbb{P}\left(\log\left(\frac{Y_i}{\tau}\right) > x\right) \\ &= \mathbb{P}(Y_i > \tau e^x) \quad \text{, substituting this in (12) gives} \\ &= e^{(1-\alpha)x} = e^{-(\alpha-1)x},\end{aligned}$$

and this is indeed the ccdf of an exponential distribution with rate parameter $\alpha - 1$.

The method relies on the fact that the sum of k independent exponential random variables with rate parameter $\alpha - 1$ follows a Gamma distribution with parameters k and $\alpha - 1$.

$$\sum_{i=1}^{k} \log\left(\frac{Y_i}{\tau}\right) \sim \text{Gamma}(k, \alpha - 1). \tag{13}$$

This is a known result in probability theory [5]. The estimator $\hat{\alpha}(\tau)$ (see Eq. (11)) can be rewritten such that $\sum_{i=1}^{k} \log\left(\frac{Y_i}{\tau}\right) = \frac{k}{\hat{\alpha}(\tau)-1}$. By using this relationship, the method constructs a pivotal quantity, which is a function of the data and the unknown parameter α, and can be obtained by using the scaling property of a Gamma distribution.

$$(\alpha - 1)\frac{k}{\hat{\alpha}(\tau) - 1} \sim \text{Gamma}(k, 1).$$

To obtain a confidence interval for α, the method calculates confidence intervals for each value of k from 1 to n. The confidence intervals are constructed based on the quantile function of the Gamma distribution, denoted by $Q(a, p)$. The quantile function, also known as the inverse cumulative distribution function (CDF), provides the value corresponding to a given probability within a distribution. Therefore, by utilizing the quantile function $Q(a, p)$ with shape parameter a and rate parameter 1, we can determine the value within the Gamma distribution that corresponds to the given probability p. The quantile function $Q(a, p)$ is expressed as follows:

$$Q(a, p) = \begin{cases} -\infty, & \text{if } p = 0, \\ \gamma^{-1}(a, \Gamma(a) \cdot p), & \text{if } p > 0, \end{cases} \tag{14}$$

where $\gamma^{-1}(s, y)$ is the inverse of the lower incomplete gamma function $\gamma(s, x) = \int_0^x t^{s-1}\, e^{-t}\, dt$ and where $\Gamma(a) = \int_0^\infty t^{a-1}\, e^{-t}\, dt$ is the gamma function.

When constructing the confidence interval for α using a Gamma distribution with shape parameter $k = k(\tau)$, rate parameter 1 and a significance level δ, we can describe the process as follows:

When constructing a confidence interval for a parameter, we use both an upper and a lower quantile function. The upper quantile function, denoted by $Q_U^{k,\delta} = Q(k, 1 - \delta/2)$, represents the critical value for the upper bound of the interval. It ensures that only a small fraction of the distribution's probability lies above this value. The lower quantile function, denoted by $Q_L^{k,\delta} = Q(k, \delta/2)$,

represents the critical value for the lower bound of the interval. It ensures that only a small fraction of the probability lies below this value.

The values $1-\delta/2$ and $\delta/2$ are used because they correspond to the desired confidence level. The confidence level indicates the success rate of the method in capturing the true parameter. To distribute the remaining probability mass δ evenly in both tails of the distribution, we divide δ by 2. This allocation results in $\delta/2$ for the lower tail and $1-\delta/2$ for the upper tail, allowing us to construct a two-sided confidence interval with the desired confidence level of $1-\delta$.

The following can be obtained,

$$\mathbb{P}\left(Q_L^{k,\delta} \leq (\alpha-1)\frac{k}{\hat{\alpha}(\tau)-1} \leq Q_U^{k,\delta}\right) = 1-\delta.$$

By rewriting this expression, the confidence interval for α can be obtained,

$$\mathbb{P}\left(1+Q_L^{k,\delta}\cdot\frac{\hat{\alpha}(\tau)-1}{k} \leq \alpha \leq 1+Q_U^{k,\delta}\cdot\frac{\hat{\alpha}(\tau)-1}{k}\right) = 1-\delta. \tag{15}$$

The idea of the method is that confidence interval for α is computed for every τ as lower bound such that n confidence intervals are obtained. We want them to be simultaneously valid to be able to compare them. The concept is as follows, suppose we have a distribution with a power law tail.

From some minimum threshold x_{min} onwards, the parameter α will be roughly the same for all τ's greater than x_{min}. Since from then on the power law applies, and as can be easily seen on log-log scale, the scaling parameter α is constant (i.e. the "slope" of the straight line on log-log scale is constant). In Figure 5, an empirical distribution with a power law tail is shown on logarithmic scale. One can clearly see that from the dotted line, representing from when the power law applies, the data represents a straight line with a constant "slope". The first step is to set τ equal to every value $X_i \in \mathbf{X}$ and calculate the corresponding estimator for τ, namely $\hat{\alpha}(\tau)$. After that the corresponding confidence interval for $\alpha(\tau)$ is calculated, note this is the confidence interval for the true unknown value of α with τ as lower bound and can be computed using Eq. (15). This is also done for the empirical data shown in Figure 5. The result is shown in Figure 6. It can be seen that from x_{min} onwards the estimator $\hat{\alpha}$ is roughly the same for all values of τ. With the exception of the largest values for τ, due to the few data points that remain then.

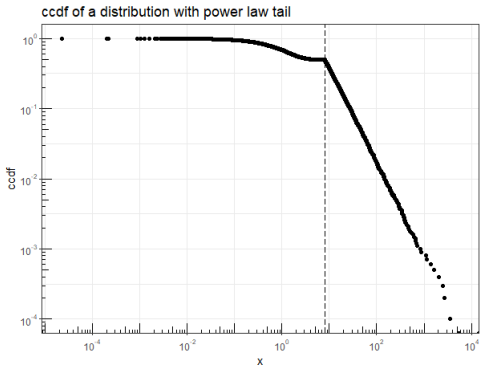


Figure 5: Example of an empirical distribution with a power law tail; ccdf where half of data is an exponential distribution with $\lambda = 1$, and the other half is a Pareto distribution with $\alpha = 2.3$ and $x_{min} = 8$

When considering values of τ smaller than x_{min}, it's important to understand that the power law tail of the distribution is no longer applicable. As a result, the estimators $\hat{\alpha}(\tau)$ for these smaller τ values may exhibit different behavior compared to the estimators of α calculated for $\tau \geq x_{min}$. The reason why the estimators $\hat{\alpha}(\tau)$ may exhibit different behavior for smaller τ values is because the underlying distribution in that range is different from the power law tail distribution. The estimators of α are derived based on the assumption of a power law distribution for $x \geq x_{min}$. The maximum likelihood estimation method is used to estimate α by fitting the power law model

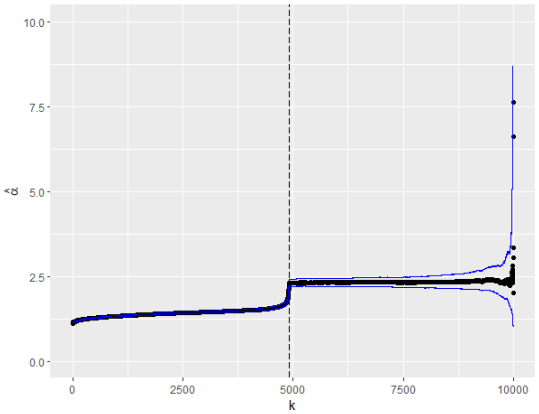


Figure 6: Estimator $\hat{\alpha}(\tau)$ with the corresponding confidence interval; the x axis shows the index of each value in the dataset sorted in increasing order, so k ranges from 1 to $n = 10000$; the dotted line is the index of x_{min} in the data; the black dots are the $\hat{\alpha}(\tau)$ for each τ in the dataset; the blue lines are the lower and upper confidence intervals

to the data. However, for values of τ smaller than x_{min}, there are values considered in the data for which the power law tail assumption no longer holds, and a different distribution may govern the data.

Due to the different distribution in that range, the behavior of the MLE estimators of α can vary. The estimators may not align with the confidence intervals calculated for $\tau \geq x_{min}$, as those intervals are based on the assumption of a power law.

Therefore, it is important to identify the minimum τ value at which the estimators $\hat{\alpha}(\tau)$ fall within the confidence intervals. This threshold marks the point where the power law tail assumption breaks down, and deviations in the behavior of the estimators occur. By determining this threshold, we can establish the range of values where the power law assumption holds, providing reliable estimation of x_{min}.

The rational above is now formalized:
Let us define the following confidence interval for α:

$$[L_{\tau,\delta}, U_{\tau,\delta}] = \left[1 + Q_L^{k,\delta} \cdot \frac{\hat{\alpha}(\tau) - 1}{k}, 1 + Q_U^{k,\delta} \cdot \frac{\hat{\alpha}(\tau) - 1}{k}\right]. \tag{16}$$

We can construct a confidence interval of this form for any $\tau > 0$ and $\delta > 0$. Please note here that $k = k(\tau)$, so k is dependent on τ. As already mentioned before, for every $\tau \in [0, \infty)$ we want to compute the confidence interval for α according to Eq. (16). Our aim is for all of these confidence intervals to hold true simultaneously, to be able to compare them.
Now we consider every $x \in \mathbf{X}$ as the value of τ. Then for all confidence intervals to be simultaneously true we want to obtain the following,

$$\mathbb{P}(\forall x \in \mathbf{X} : \alpha \in [L_{x,\delta}, U_{x,\delta}]) \geq 1 - \delta_0.$$

This can be rewritten as follows,

$$\begin{aligned}
\mathbb{P}(\forall x \in \mathbf{X} : \alpha \in [L_{x,\delta}, U_{x,\delta}]) &= 1 - \mathbb{P}(\exists x \in \mathbf{X} : \alpha \notin [L_{x,\delta}, U_{x,\delta}]) \\
&= 1 - \mathbb{P}(\bigcup_{x \in \mathbf{X}} \{\alpha \notin [L_{x,\delta}, U_{x,\delta}]\}) \\
&\geq 1 - \sum_{x \in \mathbf{X}} \mathbb{P}(\alpha \notin [L_{x,\delta}, U_{x,\delta}]) \\
&\geq 1 - \sum_{x \in \mathbf{X}} \frac{\delta_0}{n}.
\end{aligned}$$

This implies that, $[L_{x,\delta}, U_{x,\delta}]$ is the $\frac{\delta_0}{n}$ confidence interval of α based on the datapoints that are larger or equal than x. The interval suggests we are $(1 - \frac{\delta_0}{n})100\%$ confident that the true unknown parameter lies within this interval. It provides a plausible range of values for the unknown parameter based on the sample data and the chosen confidence interval. However this is a very crude bound due to the union, therefore other deltas will be considered as well.

To find the minimum x for which $\hat{\alpha}(x)$ falls within all the following confidence intervals, a set S is constructed. This will be defined as the set of all $x \in \mathbf{X}$ such that $\hat{\alpha}(x)$ is in all the following confidence intervals,

$$S = \{x \in \mathbf{X} : \forall \eta \geq x, \hat{\alpha}(x) \in [L_{\eta,\delta}, U_{\eta,\delta}]\}.$$

Then $\hat{x}_{min}$ is defined as the minimum of the set S, so

$$\hat{x}_{min} = \min_{x \in \mathbf{X}}\{S\} = \min_{x \in \mathbf{X}}\{x \in \mathbf{X} : \forall \eta \geq x, \hat{\alpha}(x) \in [L_{\eta,\delta}, U_{\eta,\delta}]\}. \tag{17}$$

Having determined the estimator for x_{min}, we can now calculate the estimator for the scaling parameter, denoted as $\hat{\alpha}(\hat{x}_{min})$, using Eq. (11), where $k = k(\hat{x}_{min})$.

Note: The largest value from the dataset is disregarded as potential $\hat{x}_{min}$ since otherwise $\hat{\alpha}(1) = 1 + 1/\sum_{i=1}^{1} \log(\frac{\tau}{\tau})$. Which is division by zero and therefore equal to infinity.

4.2 Testing procedure

The approach for the testing procedure is similar to the testing procedure of the PLFit method. The purpose of the test is to find out whether the observed data follows a power law. The Bootstrap method is used again. Meaning that data is generated from the assumed distribution of the observed data and compared in some way with the real observed data. In the PLFit method the comparison is made using the KS statistic. In this method another test statistic will be used for the comparison. A pseudo-code for the Bootstrap method is given in Algorithm 2 and will be further explained below.

Algorithm 2: A Bootstrap approach for GoF testing power-law

Input : Data $\mathbf{x} = \{x_1, \ldots, x_n\}$; $B \in \mathbb{N}$, number of Bootstrap samples

Compute $\hat{\alpha}$ and $\hat{x}_{min}$ for the observed data using confidence method
Compute $T = \max(\mathbf{x})$
for $b \leftarrow 1$ **to** B **do**
    Generate $\mathbf{Y}^{(b)} = (Y_1^{(b)}, \ldots Y_n^{(b)})$, the bootstrap samples with $\hat{\alpha}$ and $\hat{x}_{min}$
    Compute $T^{(b)} = \max(\mathbf{Y}^{(b)})$
end

Output: Bootstrap p-value computed as $\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{T^{(b)} \leq T\}$

As can be seen in Algorithm 2, the first step is to fit the empirical data to the power-law model using the methods explained in Section 4. The test statistic in this case does not need the fitted model. The test statistic used in this procedure is the maximum value of the observed or generated data. The reasoning behind this statistic will be explained further in a moment. But first, we continue with the explanation of the Algorithm. A large number of bootstrap samples will be drawn from the model that is fitted to the data. Generating these bootstrap samples follows the same method as in the PLFit method, see Algorithm 1.

There will be drawn B number of bootstrap samples of length n from Y. For every sample $\mathbf{Y}^{(b)}$ the maximum of the sample is computed and assigned to the test statistic.

The p-value can be computed as follows,

$$p\text{-value} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{T^{(b)} \leq T\}.$$

In words, the p-value is defined to be the fraction of bootstrap maximum values that are smaller than the empirical maximum value.

The power law distribution is a heavy tailed distribution. When data is simulated from a power law the maximum of the data is expected to be large. Therefore, a low p-value indicates that the observed maximum value is relatively small compared to what would be obtained under the assumption of a power law distribution. Conversely, a high p-value suggests that the maximum values obtained from the bootstrap samples are consistent with what would be expected if the observed data were indeed drawn from a power law distribution.

In other words, if the observed data truly follows a power law distribution, we would expect the maximum values obtained from the bootstrap samples to be similar to or larger than the maximum value of the observed data. If the maximum value of the observed data is relatively high compared to the maximum values from the bootstrap samples, it suggests that the observed data is not exhibiting extreme behavior that is inconsistent with a power law distribution.

When p-value ≤ 0.05, we decide that the tail of the distribution does not follow a power-law distribution. So the null hypothesis is rejected. The fixed-significance testing at level 0.05 is chosen for the purpose of comparing tests. It is important to note that the determination of when the p-value is small enough to conclude that the tail of the distribution does not follow a power law depends on the context.

5 Combination of the methods

The third and last method that is used for estimation of α and x_{min} and testing for a power law tail is a combined method of the two methods described in the previous sections. The PLFit estimation method will be utilized to estimate the scaling parameter and the lower bound. The estimators will be employed to fit the data to the best fit power law model. Subsequently, the testing procedure will be applied using the max test statistic. A pseudo-code of this testing procedure is given in Algorithm 3. As can be seen, the only difference with the confidence interval method is the way α and x_{min} are estimated to generate the bootstrap samples.

There are several reasons why this could be a good idea for a method. A disadvantage of the PLFit method is that it uses the KS statistic for both estimation and testing, it can make the results vulnerable to the assumptions and limitations of this particular statistic. Furthermore, the estimation provided by PLFit seems to be more accurate than that of the confidence interval method. This is because in the confidence interval method, there is a slack in the intersection of confidence intervals for α, which leads to a negative bias. Moreover, from a computational perspective, this method offers advantages. Unlike the PLFit testing procedure, it it doesn't have to recompute the estimators for every bootstrap sample.

Algorithm 3: A Bootstrap approach for GoF testing power-law

Input : Data $\mathbf{x} = \{x_1, \ldots, x_n\}$; $B \in \mathbb{N}$, number of Bootstrap samples

Compute $\hat{\alpha}$ and $\hat{x}_{min}$ for the observed data using PLFit method

Compute $T = \max(\mathbf{x})$

for $b \leftarrow 1$ **to** B **do**

  Generate $\mathbf{Y}^{(b)} = (Y_1^{(b)}, \ldots Y_n^{(b)})$, the bootstrap samples with $\hat{\alpha}$ and $\hat{x}_{min}$

  Compute $T^{(b)} = \max(\mathbf{Y}^{(b)})$

end

Output: Bootstrap p-value computed as $\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{T^{(b)} \leq T\}$

6 Performance of the estimators

In this section we will discuss the performance of the estimators for both the PLFit method and the confidence interval method. The general approach to evaluating their performance is the same for both methods. To gauge their performance, we employed two distributions and estimated the values of α and x_{min}. The first distribution is the pure Pareto distribution, and the second distribution is a combination of the piecewise uniform and Pareto distribution, see Eq. (4).

To generate the samples for the analysis, we set $\alpha = 2.3$ and $x_{min} = 8$ for the Pareto distribution. We collected 500 samples, each consisting of 5000 data points, from both distributions. Subsequently, we calculated the estimators for each sample and visualized the results using histograms. By examining these histograms, we can gain valuable insights into the performance and accuracy of the estimators obtained through the methods. In the histograms, the vertical dotted line indicates the true value of the parameter.

We also calculated the mean, variance and mean squared error (MSE) of the estimates obtained. The mean of the estimates provides an indication of the central tendency or average value of the estimates. If the mean of the estimates is close to the true values of α and x_{min} used to generate the samples, it suggests that the PLFit method is providing reasonably accurate estimates. A large deviation from the true values could indicate a bias or systematic error in the estimation process.

The variance of the estimates measures the spread or variability of the estimates around the mean. A smaller variance indicates that the estimates are relatively consistent and close to each other, while a larger variance suggests more variability and potential instability in the estimation process. Ideally, we would like the estimates to have a low variance, indicating robustness and stability in the PLFit method.

The MSE provides an overall measure of the accuracy of the estimates by considering both bias and variability. It quantifies the average squared difference between the estimates and the true values of α and x_{min}. A smaller MSE indicates lower overall error and better accuracy of the estimators.

6.1 PLFit method

For the pure Pareto, α is well estimated, as one can see, the mean is 2.30e+0 and variance and MSE are close to 0, see Table 1. In the histogram in Figure 7 it can be seen that the spread of the estimate is close around the true value. The estimation of x_{min} on the other hand is slightly higher, see Figure 8, but part of the reason for that is that the minimum value in the dataset is 8, so x_{min} cannot be estimated lower than 8.

For Uniform and Pareto distribution, $\hat{\alpha}$ is again very well estimated, with a mean of 2.30e+0, see Table 1. The dispersion is slightly larger than in pure Pareto, see Figure 9. For x_{min}, the mean, which is 8.62e+0, is larger than the true value but lower than in pure Pareto. This is the case because the dataset includes values smaller than 8. However, it doesn't necessarily mean that the method recognizes the power law tail in general. That will be investigated in Section 7.

The large spread of $\hat{x}_{min}$ is striking for both distributions. Outliers that are much larger than the majority of the data can distort the estimation process. Since the PLFit method utilizes the tail region of the distribution to estimate $\hat{x}_{min}$, the presence of extreme outliers can skew the estimation towards higher values.

	Pure Pareto		Uniform Pareto	
	$\hat{\alpha}$	$\hat{x}_{min}$	$\hat{\alpha}$	$\hat{x}_{min}$
mean	2.30e+0	9.75e+0	2.30e+0	8.62e+0
variance	6.83e−4	9.54e+0	1.47e−3	5.93e+0
MSE	6.82e−4	1.26e+1	1.47e−3	6.31e+0

Table 1: Table of mean, variance and MSE of the estimators of PLFit

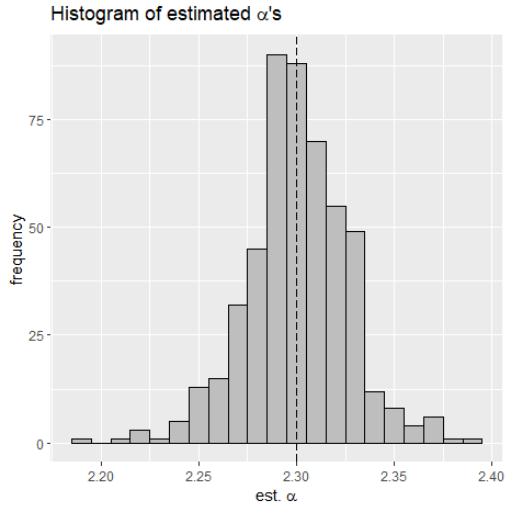


Figure 7: Histogram of $\hat{\alpha}$'s with Pure Pareto

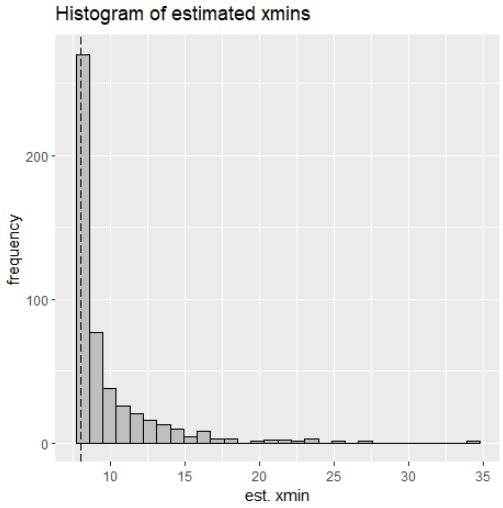


Figure 8: Histogram of $\hat{x}_{min}$'s with Pure Pareto

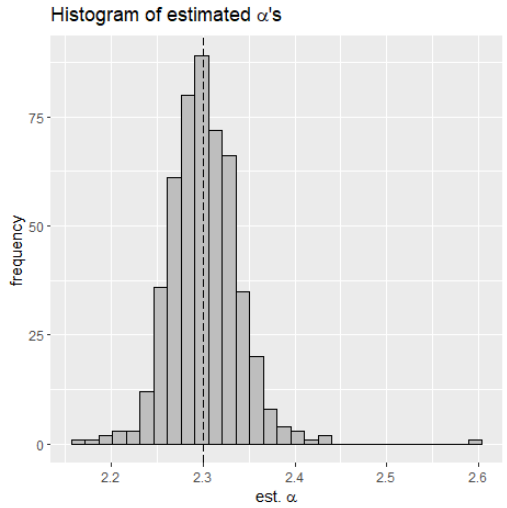


Figure 9: Histogram of $\hat{\alpha}$'s with Uniform and Pareto

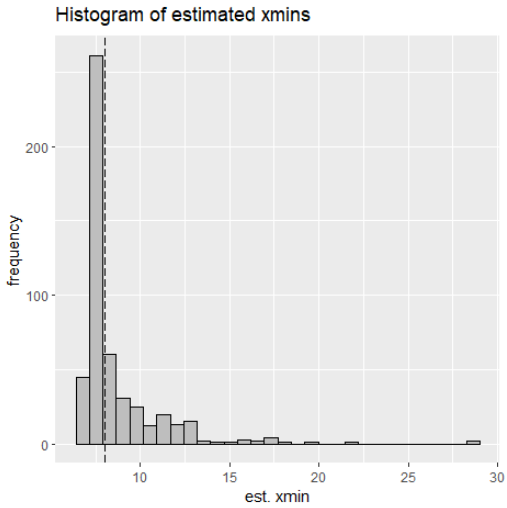


Figure 10: Histogram of $\hat{x}_{min}$'s with Uniform and Pareto

6.2 Confidence interval method

As we already saw in Section 4.1 we have the significance level δ for every $\hat{\alpha}$ that is computed. The choice of δ has an effect on the value of x_{min} that will be estimated, and therefore also on the estimation of α. We will see in this section what the influence of δ is on the performance of the estimators.

Several δ's were used for the estimation. These are from small to large, $1/n^2$, $0.05/n$, $1/n$ and $0.05/\log(n)$. We will not go too deeply into the choices of these δ's in this report. But in short, here is the reasoning. To begin with $\delta = 0.05/n$, this comes from the calculation for all confidence intervals to be simultaneously true. This choice will ensure the intervals are simultaneously valid at level δ_0, but is based on a crude union bound. Then $\delta = 0.05/\log(n)$, for estimating α we're summing independent, identically distributed random variables, then what we might expect (roughly, related to the law of the iterated logarithm) is that $\delta = 0.05/\log(n)$ would suffice. The δ's $1/n$ and $1/n^2$ come from the MSE of $\hat{\alpha}$, where in the calculation is conditioned on the confidence intervals being true and wrong for the true α.

6.2.1 Pure Pareto

To begin with the performance of the Pure Pareto. Figure 11 shows the histogram of the experiment using $\delta = 1/n^2$ for $\hat{\alpha}$. Similarly, Figure 12 illustrates the histogram for $\hat{x}_{min}$ in the same experiment. The histograms of $\delta = 0.05/\log(n)$, $\delta = 1/n$ and $\delta = 0.05/n$ can be seen in Appendix A.1.1.

One can observe that for δ equal to $1/n$, $0.05/n$ and $(1/n)^2$, the mean of $\hat{\alpha}$ is around 2.3. For $0.05/n$ and $(1/n)^2$, the mean of $\hat{x}_{min}$ is around 8. Thus, for these deltas, the estimates for the pure Pareto distribution are quite accurate. While for $0.05/\log(n)$ there are quite a few outliers. With a mean equal to 1.13e+1 and a variance of 8.67e+3, see Table 2, it is indeed not centered around the true x_{min}. This also makes the estimates for $\hat{\alpha}$ far too high, since $\hat{x}_{min}$ determines what the value of $\hat{\alpha}$ will be.

This is because for the larger deltas, the confidence intervals for α become smaller. Which makes the interval not wide enough to get over hickups at the largest values of the dataset. We will discuss in Section 6.2.3 a specific example of the influence of delta on the estimation of x_{min} and hence on the estimation of α.

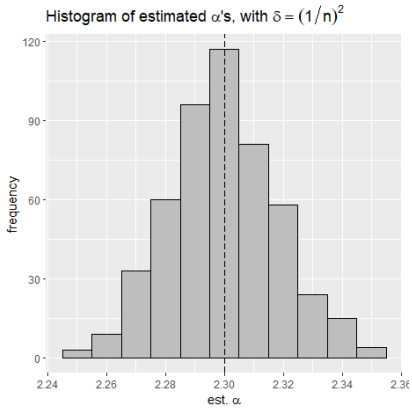


Figure 11: Histogram of $\hat{\alpha}$'s, with $\delta = 1/n^2$

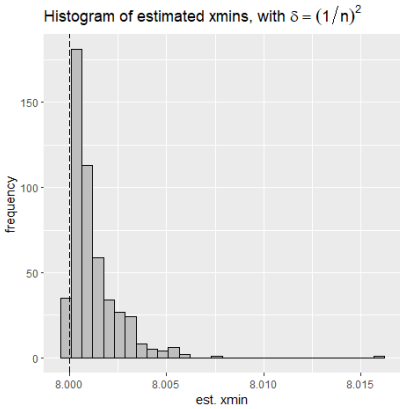


Figure 12: Histogram of $\hat{x}_{min}$'s, with $\delta = 1/n^2$

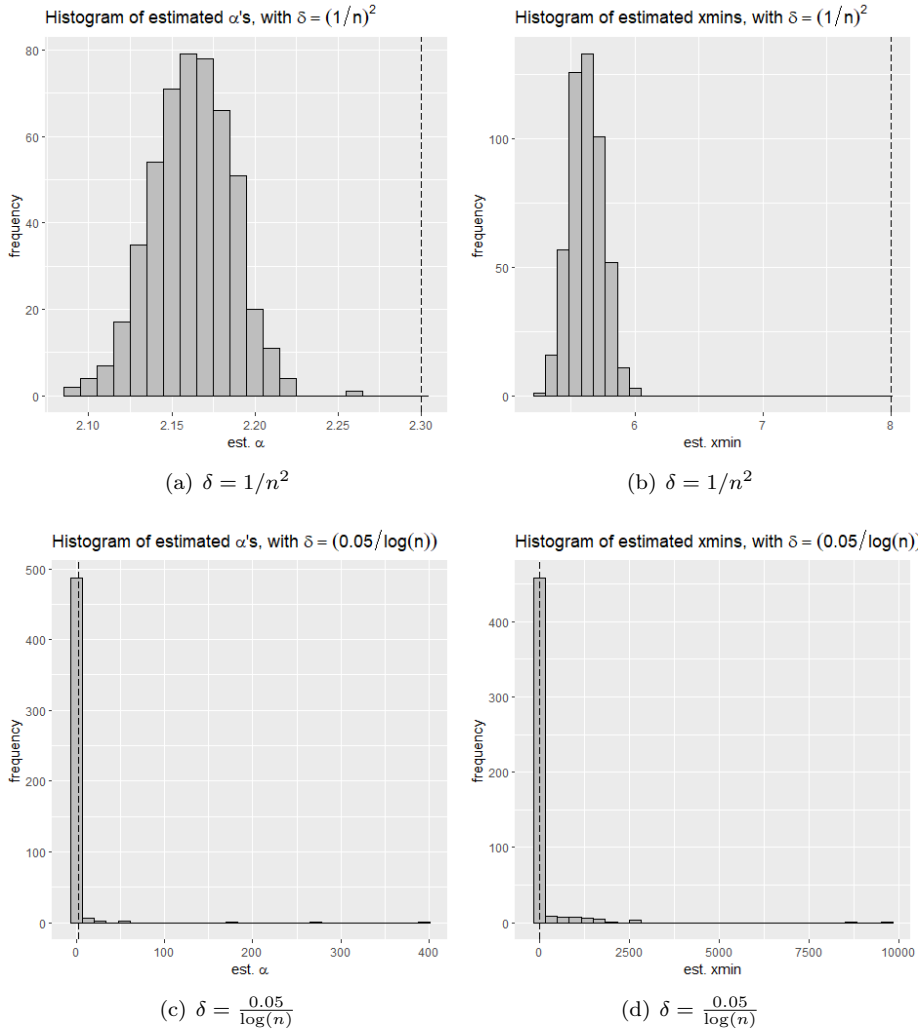
	mean		variance		MSE	
δ	$\hat{\alpha}$	$\hat{x}_{min}$	$\hat{\alpha}$	$\hat{x}_{min}$	$\hat{\alpha}$	$\hat{x}_{min}$
$0.05/\log(n)$	1.13e+1	2.86e+2	8.67e+3	1.45e+6	8.73e+3	1.52e+6
$1/n$	2.31e+0	1.75e+1	2.24e−2	2.17e+4	2.24e−2	2.17e+4
$0.05/n$	2.30e+0	8.00e+0	3.05e−4	1.44e−6	3.06e−4	2.91e−6
$(1/n)^2$	2.30e+0	8.00e+0	3.39e−4	1.83e−6	3.38e−4	3.28e−6

Table 2: Table of mean, variance and MSE of the estimators Pure Pareto of CI method

6.2.2 Piecewise uniform & Pareto

In Figure 13 the histograms of the estimates of $\hat{\alpha}$ and $\hat{x}_{min}$ are shown for the piecewise uniform and Pareto distribution. In panel (a) and (b) the histograms of the estimates of $\hat{\alpha}$ and $\hat{x}_{min}$ are shown respectively, both for $\delta = 1/n^2$. What is noticeable compared to the pure Pareto is that the $\hat{x}_{min}$ is underestimated and therefore so is $\hat{\alpha}$. Because there is a uniform distribution before the Pareto tail. This also causes a larger dispersion of $\hat{\alpha}$ and $\hat{x}_{min}$ which can be seen in the variance and MSE, see Table 3.

In panel (c) and (d) the histograms of the estimates of $\hat{\alpha}$ and $\hat{x}_{min}$ are shown respectively, both for $\delta = 0.05/\log(n)$. When $\delta = 0.05/\log(n)$ the same story applies as for the pure Pareto, $\hat{x}_{min}$ is often much too large causing $\hat{\alpha}$ to be overestimated on average. The histograms of $\delta = 1/n$ and $\delta = 0.05/n$ can be seen in Appendix A.1.2.



(a) $\delta = 1/n^2$ (b) $\delta = 1/n^2$

(c) $\delta = \frac{0.05}{\log(n)}$ (d) $\delta = \frac{0.05}{\log(n)}$

Figure 13: Each panel shows the histograms for the estimated $\hat{\alpha}$'s and $\hat{x}_{min}$ for the piecewise uniform and Pareto distribution. Panel (a) and (b) show results when $\delta = 1/n^2$. Panel (b) and (c) show results when $\delta = 0.05/\log(n)$.

	mean		variance		MSE	
δ	$\hat{\alpha}$	$\hat{x}_{min}$	$\hat{\alpha}$	$\hat{x}_{min}$	$\hat{\alpha}$	$\hat{x}_{min}$
$0.05/\log(n)$	4.39e+0	1.35e+2	5.31e+2	4.63e+5	5.34e+2	4.78e+5
$1/n$	2.21e+0	9.37e+0	2.22e−3	3.09e+3	9.69e−3	3.08e+3
$0.05/n$	2.19e+0	5.92e+0	6.27e−4	2.32e−2	1.25e−2	4.35e+0
$(1/n)^2$	2.16e+0	5.62e+0	5.68e−4	1.66e−2	1.96e−2	5.69e+0

Table 3: Table of mean, variance and MSE of the estimators of Uniform & Pareto CI method

6.2.3 The influence of δ on $\hat{x}_{min}$

It is now clear that delta affects the estimation of x_{min}. In this section, we will examine its effect using a specific example. As shown in Figure 13d, an x_{min} is estimated to be about 10000 while the true value is equal to 8. The ccdf of the dataset for which this x_{min} is estimated is shown in Figure 14.

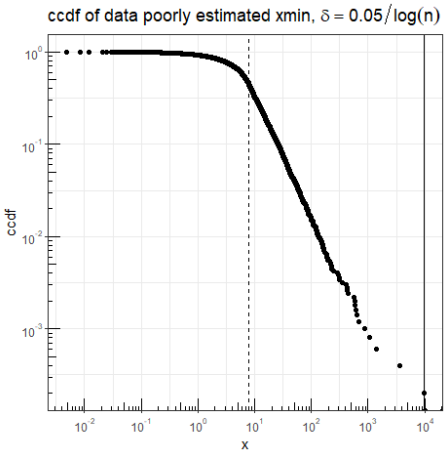


Figure 14: The ccdf of the dataset with poorly estimated x_{min}; the dotted line is the true $x_{min} = 8$ and the solid line is the estimated $\hat{x}_{min} = 9670.5$

The tail of the distribution is not a smooth linear line. It can be seen that the largest two values are much larger than the ones following them, namely 10069.9, 9670.5 and then 3631.0.

As explained in Section 4.1, the smallest value of x is chosen in which the confidence intervals of $\alpha(\eta)$ with $\eta \geq x$ overlap. This determines the value of $\hat{x}_{min}$, and subsequently, the corresponding $\hat{\alpha}(\tau)$ is determined. For example, let's consider the second-largest value as $\hat{x}_{min}$ (equal to 9670.5), and calculate the associated $\hat{\alpha}(\tau)$,

$$\hat{\alpha}(\tau) = 1 + 2\left(\log\left(\frac{10069.9}{9670.5}\right) + \log\left(\frac{9670.5}{9670.5}\right)\right)^{-1} \approx 50.4.$$

The confidence interval is relatively small due to $\delta = 0.05/\log(n)$, which means it is not wide enough to correct for this large $\hat{\alpha}(\tau)$ value. As a result, this $\hat{\alpha}(\tau)$ becomes the smallest $\hat{\alpha}(\tau)$ with an overlapping confidence interval. This can also be observed in the graph depicting the confidence intervals of α for each k and the corresponding estimate $\hat{\alpha}(\tau)$. In Figure 15 one can see that the first $\hat{\alpha}(\tau)$ is so large that one cannot even distinguish the confidence intervals of the subsequent α's from the estimates. In Figure 16, the same estimates for α are shown but now without estimate where the largest value is considered as $\hat{x}_{min}$. It can be seen that the estimates are estimated around 2.3 until it drops at some point (where the actual x_{min} is located).

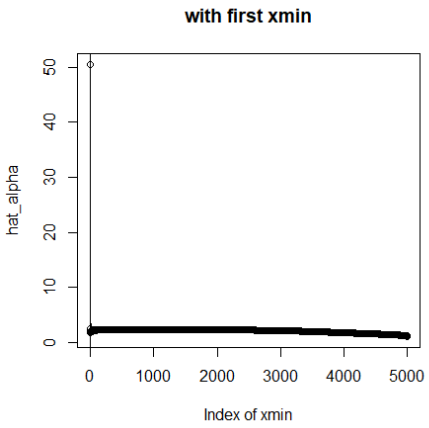


Figure 15: Estimates of $\hat{\alpha}(\tau)$ for each k and corresponding confidence interval for α

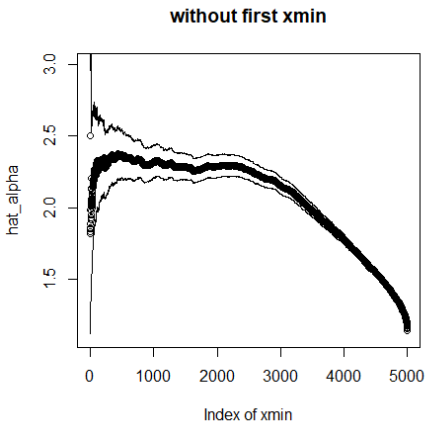


Figure 16: Estimates of $\hat{\alpha}(\tau)$ for each $k \in \{3, \ldots n\}$ and corresponding confidence interval for α

So there is an interaction between δ and $\hat{x}_{min}$. For small δ's, the estimator $\hat{x}_{min}$ will consistently exhibit a negative bias, since this is based on simultaneously valid confidence intervals. The magnitude of this bias increases as the confidence intervals become wider, allowing for more flexibility in the estimates of α. The width of the confidence intervals is inversely related to the value of δ, meaning that a smaller δ leads to a larger negative bias. This relationship can be observed in Tables 2 and 3.

However, it is important to note that excessively increasing δ can lead to a significant risk of the intervals no longer encompassing the true value of the power law exponent. Consequently, for extremely large values of δ, the bias may actually become positive and even considerably high. Which can be seen in Figure 17, here various δ are plotted against $\hat{x}_{min}$ for a piecewise uniform and Pareto distribution with $\alpha = 2.3$ and $x_{min} = 8$ and $n = 10000$. We started with $\delta = 1/n$ and continued till $\delta = 50/n$. It can be seen that indeed for larger values of δ the values for $\hat{x}_{min}$ become very high.

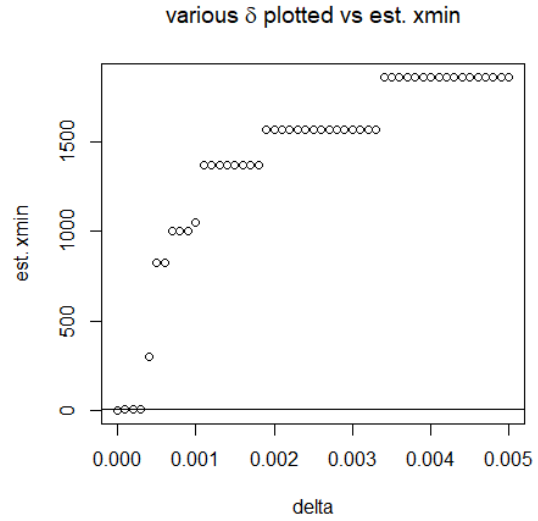


Figure 17: Plot of δ vs $\hat{x}_{min}$ for piecewise uniform and Pareto distribution with $n = 10000$

7 Numerical results

In this section a number of numerical experiments are conducted to evaluate (empirically) the performance of the proposed testing methods. As the methods are computationally demanding, the number of bootstrap samples used is $B = 200$ to compute the p-values. Also, the experiments are repeated $R = 500$ times.

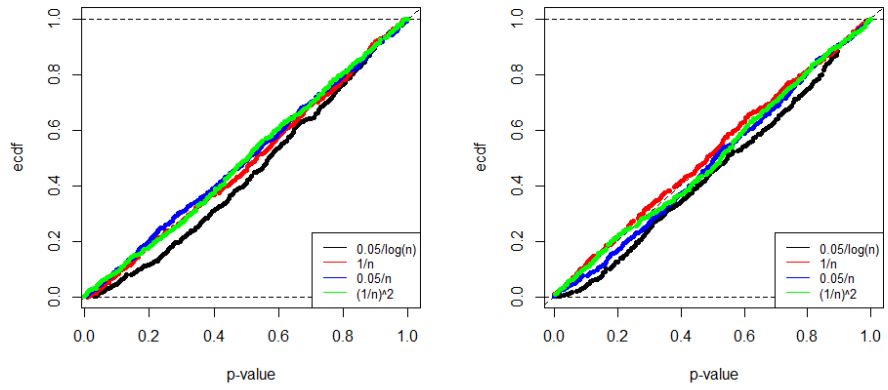
7.1 Calibration confidence interval method for various δ

In Figure 18 we investigate for which δ the testing procedure of the confidence interval method is well calibrated. Data was generated from a pure Pareto distribution and a piecewise uniform & Pareto distribution with $\alpha = 2.3$ and $x_{min} = 8$. In panel (a) the empirical distribution functions of the p-values corresponding to the four δ's are shown for the case $n = 1000$, and in panel (b) for $n = 10000$, where the data was generated from the pure Pareto distribution. Here we see the test seems to be well calibrated when a pure Pareto is used as the distribution under the null. However, in panel (b) and (c) the empirical distribution functions of the p-values are shown, where the data was generated from the piecewise uniform and Pareto distribution. Here we see that test is not well calibrated for any δ.

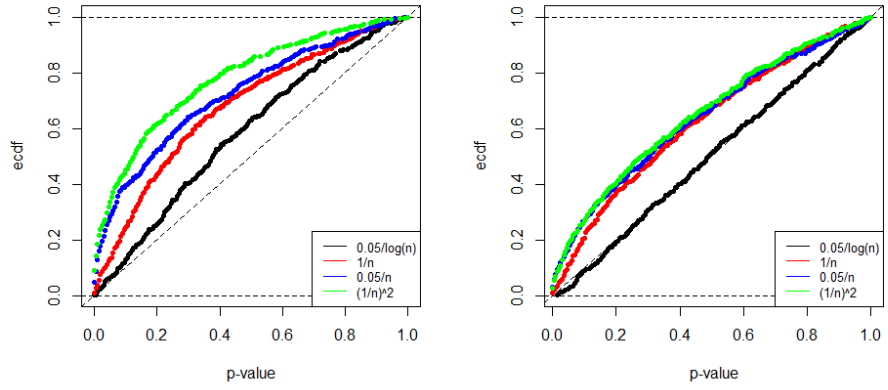
Except for $\delta = 0.05/\log(n)$ in the case where $n = 10000$, but as we already saw in Section 6.2, $\hat{x}_{min}$ is often estimated as the second-largest value of the data sample. Once the bootstrap samples are generated in the testing procedure, $k = 2$ so that with a probability of $\frac{2}{n}$ an element is sampled from the pure Pareto, and with a probability of $1 - \frac{2}{n}$ an element is sampled from the observed data. Such a high $\hat{x}_{min}$ causes the bootstrap samples to be very similar to the observed data, and therefore the maximum of the observed data is approximately equal to the maxima of the bootstrap samples, suggesting that the data would come from a power law.

On the other hand, when δ is small, the estimate $\hat{x}_{min}$ will have a big negative bias. This is what causes the problem with the calibration of the test. Therefore, the choice of δ in the estimation procedure is very important for the calibration of the test in the testing procedure.

The delta value chosen for the alternative hypothesis testing using the confidence interval method is given by $\delta = 1/n$. This choice is motivated by the fact that it results in the smallest mean squared error (MSE) for the estimated parameter $\hat{\alpha}$. Furthermore, when estimating the value of x_{min}, although there are a few outliers, $\delta = 1/n$ appears to produce the least biased results compared to other delta values under consideration. Additionally, this testing approach seems to be the most accurately calibrated.



(a) $n = 1000$ (b) $n = 10000$



(c) $n = 1000$ (d) $n = 10000$

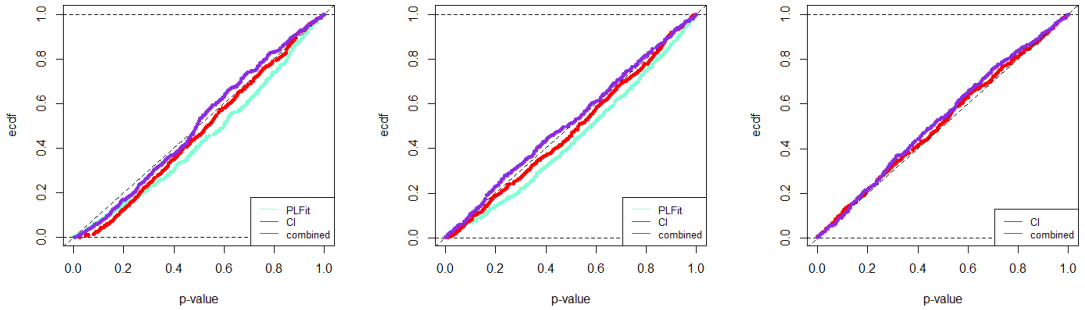
Figure 18: Each panel shows the empirical distribution function for the p-values of the tests. (based on $R = 500$ repetitions). Panel (a) and (b) show results under the pure Pareto distribution, for sample sizes $n = 1000, 10000$ respectively. Panel (b) and (c) show results under the piecewise uniform and Pareto distribution.

7.2 Calibration of the three testing methods

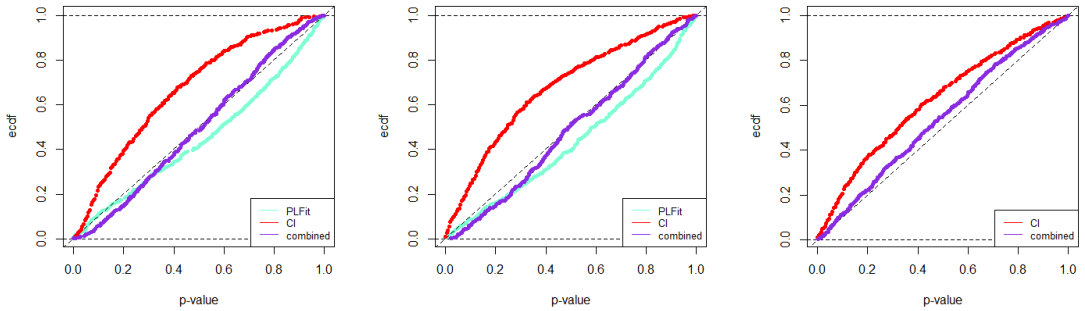
In Figure 19 we investigate if the three testing methods are well calibrated. In panel (a) - (c), the data was generated from a pure Pareto distribution and in panel (d) - (f), the data was generated from a piecewise uniform and Pareto distribution. In panel (a) and (b) the empirical distribution functions of the p-values corresponding to the three testing methods are shown for $n = 200, 1000$. Due to its high computational demands, the PLFit method is tested for $n = 200, 1000$. And in panel (c) we have the empirical distribution functions of the p-values corresponding to the confidence interval method and the combined method. It can be seen that under the pure Pareto, all three test methods are well calibrated. Where, as the sample size increases, the tests become better calibrated.

But problems arise again when the piecewise uniform and Pareto distribution is used as distribution under the null. In panel (d) - (f), it can be seen that the confidence interval method is not well calibrated. When comparing the PLFit method and the combined method, the combined method is better calibrated than the PLFit method. But both tests are reasonably well calibrated.

As already mention in Section 2.2, ensuring a well-calibrated test is important because it ensures the validity of hypothesis testing. When an hypothesis test is not well calibrated, it may lead to incorrect inferences and flawed decision-making. In such cases, the test may produce misleading results or incorrect conclusions about the hypotheses being tested.



(a) $n = 200$ pure Pareto (b) $n = 1000$ pure Pareto (c) $n = 10000$ pure Pareto



(d) $n = 200$ uniform & Pareto (e) $n = 1000$ uniform & Pareto (f) $n = 10000$ uniform & Pareto

Figure 19: Each panel shows the empirical distribution function for the p-values of the tests. (based on $R = 500$ repetitions). Panel (a)-(c) show results under the pure Pareto distribution, for sample sizes $n = 200, 1000, 10000$ respectively. Panel (d)-(f) denote results under the uniform and Pareto distribution.

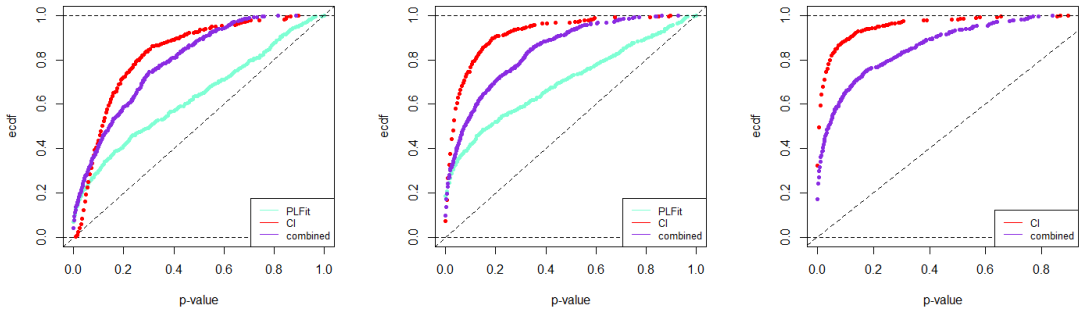
7.3 Statistical power of the three testing methods

In Figure 20 we investigate the power of the three testing methods. For investigating the power of the various tests, data was generated from an exponential distribution $\lambda = 0.125$ and a log-normal distribution with $\mu = 0.3$ and $\sigma = 2$.

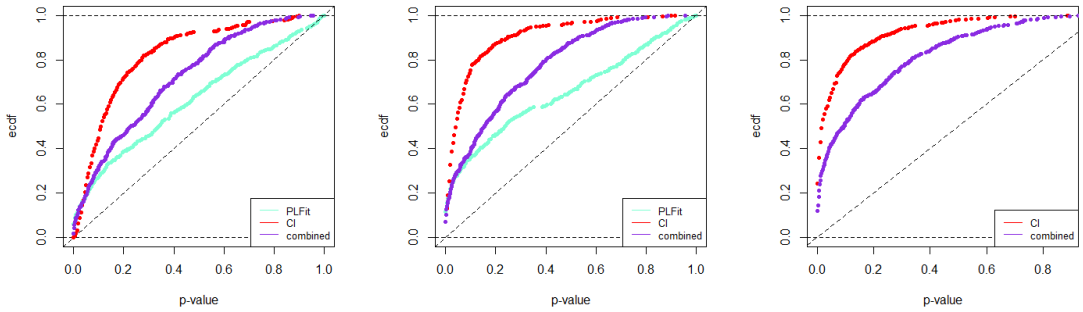
The power of a test is the probability of correctly rejecting the null hypothesis. To determine the power of the tests in our numerical results, we calculated the fraction of times where the data was recognised as data not coming from a power-law distribution. This calculation involved counting the number of times the p-value was less than or equal to the significance level of 0.05. We divided this count by the total number of p-values, which was 500 (as we performed 500 repetitions). The detailed results can be found in Table 4.

As we analyze the results, it becomes evident that the tests exhibit increasing power as n increases. In panel (a) - (c), where data was generated from an exponential distribution, it seems that the confidence interval method has the highest power. However, it is important to note that the reliability of these results is compromised due to the poor calibration of the confidence interval method. On the other hand, the PLFit method demonstrates the lowest power. Consequently, the combined method emerges as the most reliable and powerful approach.

Moving on to the log-normal distribution in panels (d)-(f), we observe a similar trend where the confidence interval method appears to have the highest power, albeit once again being unreliable. Interestingly, all three tests struggle to accurately identify the log-normal distribution as lacking a power law tail. Among them, the PLFit method encounters the greatest difficulty, followed by the combined method, and finally the confidence interval method, where the latter is still unreliable.



(a) $n = 200$ exponential (b) $n = 1000$ exponential (c) $n = 10000$ exponential



(d) $n = 200$ log-normal (e) $n = 1000$ log-normal (f) $n = 10000$ log-normal

Figure 20: Each panel shows the empirical distribution function for the p-values of the tests. (based on $R = 500$ repetitions). Panel (a)-(c) show results under the exponential distribution, for sample sizes $n = 200, 1000, 10000$ respectively. Panel (d)-(f) denote results under the log-normal distribution.

	PLFit method		Confidence interval method		Combined method	
	Log-normal	Exponential	Log-normal	Exponential	Log-normal	Exponential
$n = 200$	0.186	0.224	0.236	0.194	0.192	0.28
$n = 1000$	0.288	0.352	0.564	0.628	0.298	0.416
$n = 10000$	-	-	0.65	0.82	0.42	0.53

Table 4: Power of the three tests for the log-normal and exponential distribution

8 Case study

In this section, we will apply the methods we have analyzed to a dataset of earthquakes in Limburg. Over the years, there have been suggestions that earthquake magnitudes, in general, follow a power law distribution. Our aim now is to verify if this pattern remains consistent within our specific dataset and assess whether our methodologies lead to similar findings.

8.1 Magnitude and seismic moment

The seismic moment is a measure of the total energy released by an earthquake. It takes into account the area of the fault that slipped, the average amount of slip along the fault, and the rigidity of the rocks involved. The Richter magnitude scale, is a logarithmic scale used to quantify the magnitude or size of an earthquake. Furthermore, the Richter scale is often used in news media and in the KNMI data.

It's important to note that the Richter magnitude scale primarily provides an estimate of the amplitude of seismic waves, while the seismic moment provides a more comprehensive measure of the earthquake's energy release based on the fault area, slip, and rock rigidity.

The relationship between the magnitude m and seismic moment x of an earthquake, as proposed by Kagan [12], can be expressed by the equation:

$$m = \frac{2}{3}\log_{10}(x) - 6.$$

This relation will be used to convert between the magnitude and seismic moment.

8.2 Exploratory data analysis

The KNMI, which stands for "Koninklijk Nederlands Meteorologisch Instituut" or Royal Netherlands Meteorological Institute, has multiple seismic stations throughout the Netherlands. These stations are responsible for detecting and documenting earthquakes, as well as reporting the corresponding local magnitudes. All recorded earthquakes are listed and available through the KNMI database [1]. Each earthquake entry in the database includes information such as the date, location of the epicentre (latitude and longitude), and magnitude. In 1906, the first seismic stations in the Netherlands was installed in the Bilt.

The KNMI faced limitations in terms of the number and accuracy of seismic stations, which resulted in the inability to register all earthquakes since the installation of the first seismic station. This non-registration of earthquakes has led to incomplete datasets. However, there is a way to deal with this non-registration, called the Magnitude of Completeness (MOC), which establishes a minimum magnitude threshold above which all earthquakes within a specific region are reliably recorded. Unfortunately, the specific date when the KNMI started implementing this MOC value has not been reported.

In 1985, a significant earthquake highlighted the need for enhanced seismic monitoring capabilities, leading to improvements in the seismic network. For this research, we assume that starting from 1985 onwards, all earthquakes with a magnitude greater than or equal to 1.8 have been registered.

In this case study, our investigation centers specifically on the earthquakes registered in the province of Limburg. To achieve this, a dataset filtering procedure was implemented, focusing on the altitude and longitude criteria of Limburg. As a result, our filtered dataset exclusively comprises 127 earthquakes that occurred within the boundaries of Limburg with magnitudes larger than or equal to 1.8 and registered after 1985. The empirical ccdf of the seismic moments of these earthquakes that are registered can be seen in Figure 21. In Section 8.3 we will apply the estimation and testing methods discussed in this thesis on this dataset.

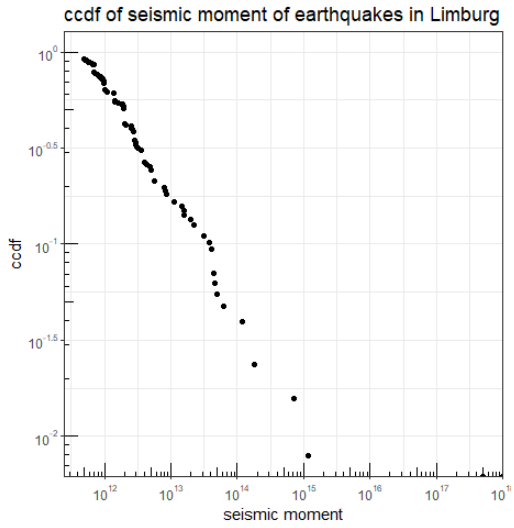


Figure 21: Empirical ccdf of the seismic moments of the earthquakes in Limburg

8.3 Applying the methods

We applied the two estimation methods on the data of earthquakes in Limburg and we obtained the following results for estimation procedure, see Table 5:

	PLFit method	Confidence interval method
$\hat{\alpha}$	1.56e+00	1.56e+00
$\hat{x}_{min}$	6.41e+11	5.01e+11

Table 5: Estimation methods applied on data of earthquakes in Limburg

It can be observed that both methods yield similar estimations for the scaling parameter α, assuming that the underlying distribution of the data follows a power law distribution. However, there is a discrepancy in the estimation of the lower bound x_{min} between the methods. When converted to magnitudes, the estimated $\hat{x}_{min}$ is approximately 1.87 for the PLFit method and around 1.8 for the confidence interval method. As already seen in Section 6 the estimation of the lower bound x_{min} differs in accuracy between the two methods, which could potentially explain the discrepancy. Additionally, it is worth noting that the sample size is very small, which can have a significant influence on the accuracy of the estimators. The accuracy of estimators in power law distribution heavily relies on the availability of sufficient data points in the tail of the distribution.

The assumption that the underlying distribution of the earthquake data in Limburg follows a power law distribution was tested using three different testing methods. The results obtained from these methods are as follows:

$$p\text{-value PLFit method} : 0.155$$
$$p\text{-value confidence interval method} : 0.965$$
$$p\text{-value combined method} : 0.935$$

The null hypothesis states that the underlying distribution of the data denoted by F follows a power law distribution. The alternative hypothesis states that F does not follow a power law. This means that H_1 suggests the presence of a different distribution that deviates from the power law.

Given the p-values obtained from the testing methods, all of which are larger than the significance level of 0.05, we do not have sufficient evidence to reject the null hypothesis. Therefore, we do not find significant support to conclude that the data does not follow a power law distribution.

It is important to note that the small sample size may limit the statistical power to detect significant deviations from the assumed power law distribution. Thus, the lack of sufficient evidence to reject the null hypothesis should be interpreted in the context of the sample size limitations.

Therefore, while the estimations and p-values provide insights into the data analysis, it is essential to be cautious in drawing definitive conclusions about the suitability of a power law distribution for the earthquake data in Limburg, given the impact of the small sample size on the accuracy of estimators and hypothesis testing outcomes.

9 Discussion

In Sections 6.1 and 6.2 we analyzed the performance of the estimators for the PLFit method and the confidence interval method respectively. The PLFit method is accurate in estimating α and x_{min} for the pure Pareto and piecewise uniform and Pareto distribution. The confidence interval method is not for the δ's we considered. For small δ's, $\hat{x}_{min}$ is negatively biased and for large δ's, on the contrary, the estimator is positively biased.

In Section 7.2 we investigated the calibration of the three testing methods. The confidence interval method is poorly calibrated. But the PLFit method and combined method are both reasonably well calibrated. This provides a solid foundation for evaluating the power of the test and its ability to detect alternative hypotheses or deviations from the null hypothesis.

In Section 7.3 the power of the three testing methods is investigated. Since the confidence interval method is not well calibtrated, the results in the power of the test can not be relied upon. That leaves the PLFit method and combined method. Since the PLFit method has a lower power for both the exponential and log-normal distribution, the combined method emerges as the most reliable and powerful method.

Moving forward, there are several avenues for future research that can build upon the findings of this study. Firstly, other test statistics than the maximum of the dataset can be investigated using the confidence interval method. Ideas for test statistics include, for example, using other quantiles of the dataset. Now the maximum value is used, which is the 100% quantile, but one can look at the 99% quantile or the 50% quantile, also known as the median of the data. Power law tails are characterized by a slow decay of probabilities, which implies that extreme values are relatively more likely to occur. By focusing on a quantile in the tail region, one can detect potential deviations from power law behavior. Another idea is considering the mean or median of the data that is larger or equal than $\hat{x}_{min}$. However, one should be cautious with this as under the alternative $\hat{x}_{min}$ will be lower so more data is in the tail, while for the bootstrap samples this will not be the case. And that will affect the median.

Secondly, as we already saw, the choice of δ is very important for accurate estimators of the power law distribution. Further research can be done on the effect of δ and determining an optimal value for δ that ensures an unbiased estimation of x_{min}.

Thirdly, the distributions that are used for analyzing the estimation and testing methods could be extended. In this study, only a piecewise uniform and Pareto distribution is used as a distribution with a power law tail. However, an intriguing piecewise distribution that could be further investigated within the context of the methods is a piecewise exponential and Pareto distribution. This distribution has the potential to provide new insights into the performance of the estimation and testing methods. In addition, other distributions can also be tested for the alternative hypothesis, such as a piecewise probability distribution of uniform and exponential, or uniform and log-normal, for example. By testing these distributions for goodness-of-fit, one can determine whether the tests indeed reject these distributions as inconsistent with the power law hypothesis.

Finally, a pure power law tail rarely occurs in nature. Therefore, the estimation and testing procedure can be extended to less strict Pareto's, such as a truncated Pareto or a tapered Pareto, or regularly varying distributions.

It is important to acknowledge that our research has certain limitations. For example, we tested our methods on a limited number of distributions. While there are a lot more variations of distributions that the methods could be tested for. The conclusions drawn in this research are based on the distributions we used. While attempts were made to make the distributions diverse enough, the current results may not accurately represent all scenarios. It is possible that the performance varies for specific distributions compared to what we have observed so far. Conducting thorough research on other distributions can help determine if this situation arises.

Furthermore, due to the inefficiency of the PLFit method's simulation, the processing time was quite long. Therefore, the PLFit method is tested for a maximum sample size of $n = 5000$ in estimation and $n = 1000$ in the testing procedure. This should be taken into account when considering the results.

Having examined the results obtained from the simulation study on the estimation and testing procedure, as well as offering recommendations for future investigations, we can now proceed to draw conclusions for this research.

10 Conclusion

In this report, we introduce the concept of power laws. Power laws play a crucial role in understanding the distribution of various phenomena in natural, social, and technological systems. Power laws are a family of probability distributions with heavy tails that exhibit remarkable characteristics, such as the occurrence of extreme events or outliers. We delved into the PLFit method, a well-known technique developed specifically for analyzing empirical data that is believed to follow a power-law distribution.

The problem tackled in our research is that the PLFit method has low power, meaning that the power-law hypothesis is often not rejected when it should have been. Chances are that there appears to be evidence for a certain model choice, but in reality, this model choice is not supported. The main purpose of this study was to introduce a new method called the confidence interval method and compare the performance of the two methods in estimation and testing.

The research process began by examining the PLFit method and the confidence interval method in order to get a better understanding of both methods. Following this, a simulation study was conducted to evaluate the performance of the estimators in both methods. Additionally, an analysis was performed on the effect of δ on the estimators, specifically for the confidence interval method. Lastly, a simulation study was conducted to assess the power of the different testing methods; here the combined method comes into play.

To conclude the research of this report, in the estimation part, we discovered the PLFit method is accurate for the distributions we tested for. The confidence interval estimation procedure is not very accurate now but can perform much better with the right choice for δ. In the testing for goodness-of-fit, we found out that the PLFit method has less power than the confidence interval method. However, the confidence interval method is unreliable due to the bad estimators and results in a poorly calibrated test. Therefore, the combined method seems to bring out the best in both methods; the estimation of the PLFit method has a positive effect on the testing procedure of the confidence interval method in calibration and power. Another overall advantage of the combined method is that it is much faster in computation, it does not use the same statistic for both estimation and testing, and the testing procedure is easier to compute.

To conclude the case study where we applied the methods to a dataset of earthquakes in Limburg, the estimation part yielded similar results for the scaling parameter but differed in the estimation of the lower bound. The hypothesis testing indicated that there is not enough evidence to reject the assumption of a power law distribution, although the small sample size may limit the reliability of the conclusions. Therefore, caution should be exercised when drawing definitive conclusions about the suitability of a power law distribution for the earthquake data in Limburg.

This report has provided an overview and comparison of three methods − the PLFit method, the confidence interval method, and a combination of the two − for analyzing empirical data following a power-law distribution (in the tail). Through careful analysis and simulation studies, we have gained valuable insights into the strengths and limitations of each method. This study is the basis for further investigation into the new estimation and testing procedures. By acquiring more knowledge about the behavior of estimation and testing procedures, researchers can make informed decisions and refine these methods to better suit the needs of various applications.

References

[1] KNMI - Data. URL http://rdsa.knmi.nl/fdsnws/event/1/query?format=text&nodata=404&eventtype=earthquake.

[2] F. Abramovich and Y. Ritov. *Statistical theory : a concise introduction*. CRC Press, Boca Raton, 2013.

[3] L. Adamic and B. Huberman. The nature of markets in the World Wide Web. *Q. J. Electron. Commerce*, 1:5–12, 2000.

[4] A. Bhattacharya, B. Chen, R. van der Hofstad, and B. Zwart. Consistency of the PLFit estimator for power-law data. 2 2020. URL http://arxiv.org/abs/2002.06870.

[5] J. K. Blitzstein and J. Hwang. *Introduction to Probability: Texts in Statistical Science*. CRC Press, 10 2015.

[6] Bruce M. Hill. Hill inference tail of distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975.

[7] M. C. Bryson. Heavy-Tailed Distributions: Properties and Tests. *Technometrics*, 16(1):61–68, 2 1974.

[8] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data, 2009. ISSN 00361445.

[9] J. Estoup. *Gammes Sténographiques*. Institut Sténographique, Paris, 4 edition, 2016.

[10] T. Fenner, M. Levene, and G. Loizou. Predicting the long tail of book sales: Unearthing the power-law exponent. *Physica A*, 389:2416–2421, 2010. doi: 10.1016/j.physa.2010.02.021. URL www.longtail.com.

[11] B. Gutenberg and C. F. Richter. Frequency of earthquakes in California. *Bulletin of the Seismological Society of America*, 34(4):185–188, 10 1944. ISSN 1943-3573. doi: 10.1785/BSSA0340040185.

[12] Y. Y. Kagan. Seismic moment distribution revisited: I. Statistical results. Technical report, 2002. URL https://academic.oup.com/gji/article/148/3/520/822773.

[13] J. Karamata. Sur un mode de croissance regulière des fonctions. *Mathematica (Cluj)*, pages 38–53, 1930.

[14] D. Leynaud, D. Jongmans, H. Teerlynck, and T. Camelbeeck. Seismic hazard assessment }in Belgium. *Geologica Belgica*, pages 67–86, 2001.

[15] M. Mitzenmacher. A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics*, 1(2):226–251, 2004.

[16] D. S. Moore. *Basic Practice of Statistics*. Freeman, New York, 4 edition, 2007.

[17] V. Pareto. *Cours d'Economie Politique*, volume 2. Librairie Droz, Geneva, 1897.

[18] W. H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical recipes in C : the art of scientific computing*. Cambridge University Press, Cambridge, UK, 2 edition, 1992. ISBN 0521431085.

[19] United States Census Bureau. City and Town Population Totals: 2020-2022, 2022. URL https://www.census.gov/data/tables/time-series/demo/popest/2020s-total-cities-and-towns.html.

[20] I. Voitalov, P. Van Der Hoorn, R. Van Der Hofstad, and D. Krioukov. Scale-free networks well done. *Physical Review Research*, 1(3), 10 2019. ISSN 26431564. doi: 10.1103/PhysRevResearch.1.033034.

[21] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, New York, 2 edition, 2004.

[22] H.-Y. Wu, C.-I. Chou, and J.-J. Tseng. A simple marriage model for the power-law behaviour in the frequency distributions of family names. *Computer Physics Communications*, 182: 201–204, 2011. doi: 10.1016/j.cpc.2010.07.051. URL `www.elsevier.com/locate/cpc`.

A Extra Figures

A.1 Performance estimators CI method

A.1.1 Pure Pareto

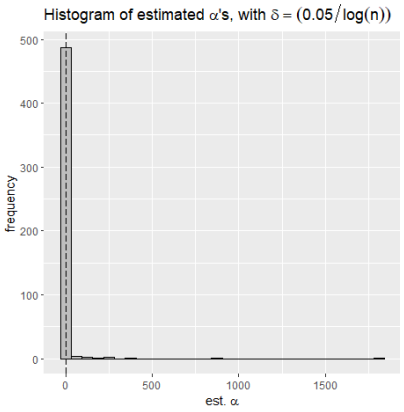


Figure B.1: Histogram of $\hat{\alpha}$'s, with $\delta = \frac{0.05}{\log(n)}$

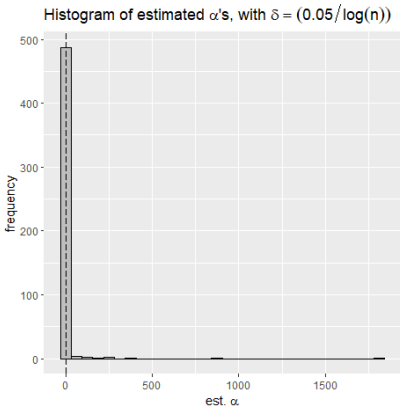


Figure B.2: Histogram of $\hat{x}_{min}$'s, with $\delta = \frac{0.05}{\log(n)}$

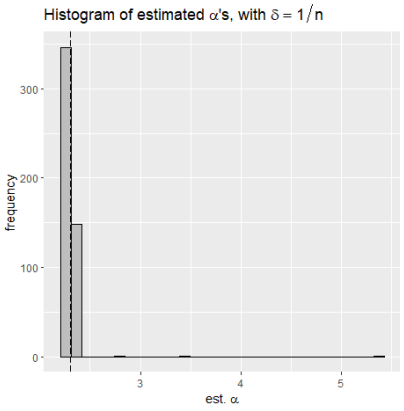


Figure B.3: Histogram of $\hat{\alpha}$'s, with $\delta = 1/n$

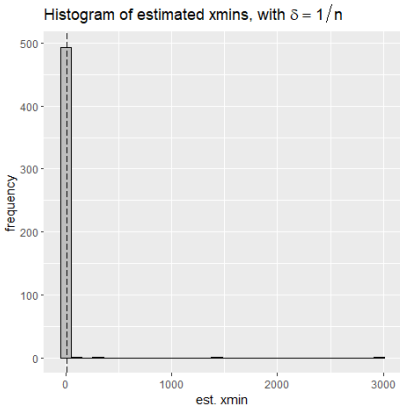


Figure B.4: Histogram of $\hat{x}_{min}$'s, with $\delta = 1/n$

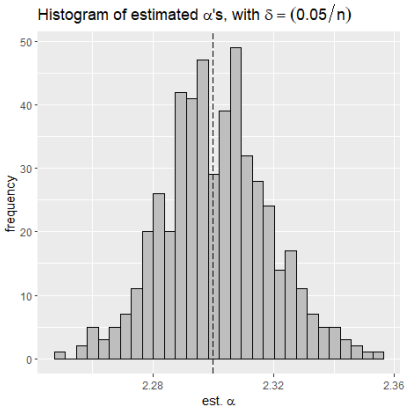


Figure B.5: Histogram of $\hat{\alpha}$'s, with $\delta = \frac{0.05}{n}$

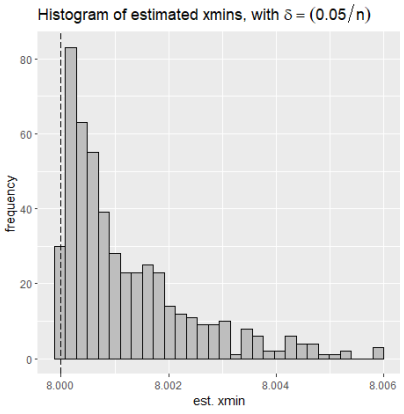


Figure B.6: Histogram of $\hat{x}_{min}$'s, with $\delta = \frac{0.05}{n}$

A.1.2 Uniform Pareto

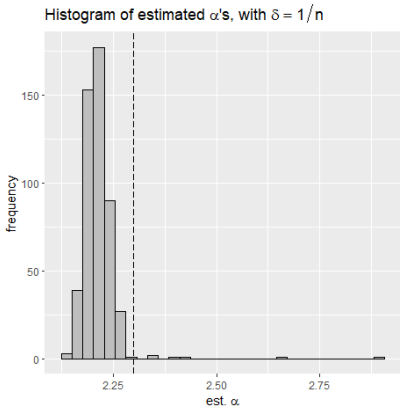


Figure B.7: Histogram of $\hat{\alpha}$'s, with $\delta = 1/n$

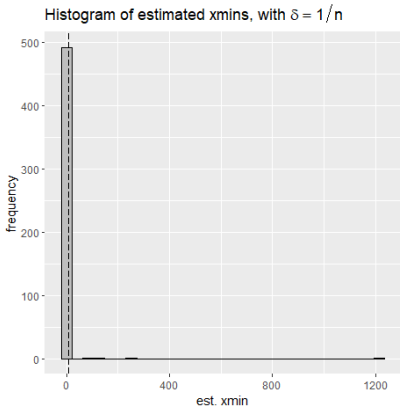


Figure B.8: Histogram of $\hat{x}_{min}$'s, with $\delta = 1/n$

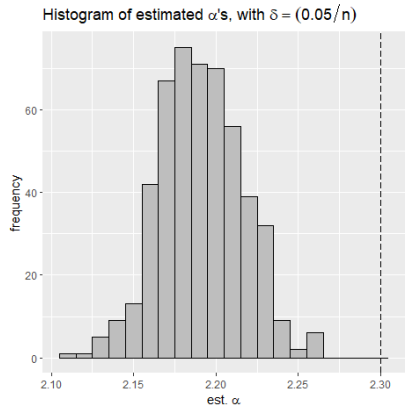


Figure B.9: Histogram of $\hat{\alpha}$'s, with $\delta = \frac{0.05}{n}$

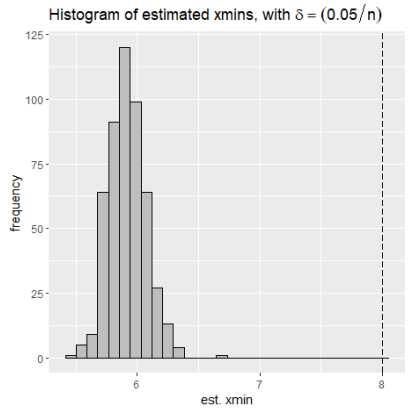


Figure B.10: Histogram of $\hat{x}_{min}$'s, with $\delta = \frac{0.05}{n}$