Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft Institute of Applied Mathematics

Heuristics of heavy-tailed distributions and the Obesity index

A thesis submitted to the
Delft Institute of Applied Mathematics
in partial fulfillment of the requirements

for the degree

MASTER OF SCIENCE
in
APPLIED MATHEMATICS

by

D. Nieboer

Delft, the Netherlands
January 2011

MSc THESIS APPLIED MATHEMATICS

"Heuristics of heavy-tailed distributions and the Obesity index"

D. Nieboer

Delft University of Technology

Daily supervisor

Prof.Dr.ir. R. M. Cooke

Responsible professor

Prof.Dr.ir. R. M. Cooke

Other thesis committee members

Dr. E.A. Cator

Dr. A.M. Hanea

Prof. Dr. Ir. J. Misiewicz

January 2011

Delft, the Netherlands

Contents

Chapter 1

Introduction

Many distributions that are found in practice are thin-tailed distributions. Heights and weights of animals for example are usually thin-tailed distributions, but there are also a lot of examples where heavy-tailed distributions can be found. The first example of heavy-tailed distributions was found in Mandelbrot [1963] where it was shown that the change in cotton prices was heavy-tailed. Since then many other examples of heavy-tailed distributions are found, among these are data file traffic on the internet (Crovella and Bestavros [1997]), returns on financial markets (Rachev [2003], Embrechts et al. [1997]) and magnitudes of earthquakes and floods (Latchman et al. [2008], Malamud and Turcotte [2006]). But what are heavy-tailed distributions and what is the difference with thin-tailed distributions, and how can one detect whether a distribution is heavy-tailed or not?

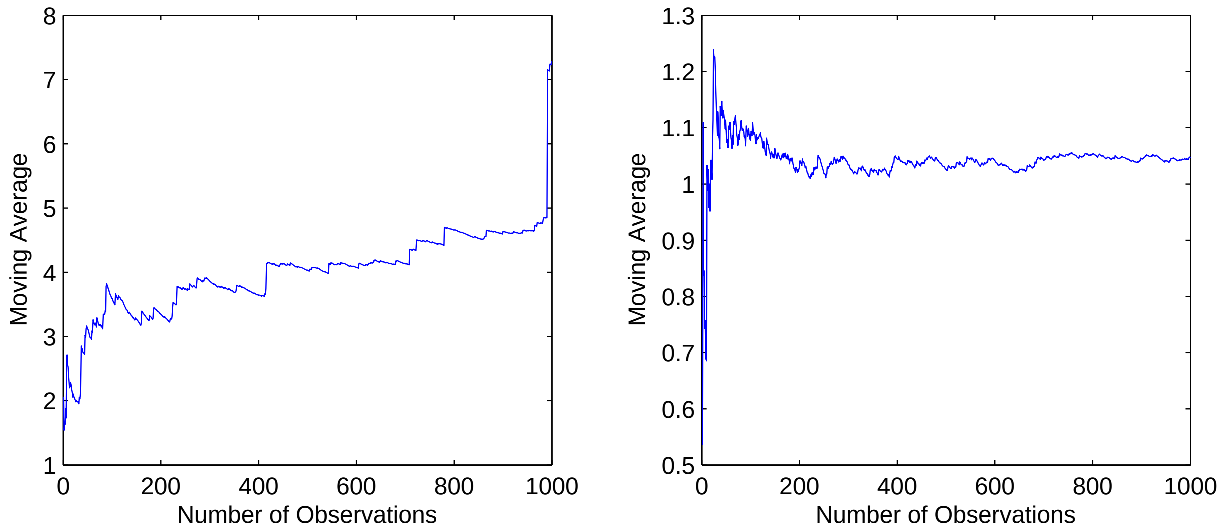
1.1 Heavy-tailed distributions

There are a few different definitions of heavy-tailedness of a distribution. These definitions all relate to the decay of the survivor function of a random variable. Two widely used classes of heavy-tailed distributions are the regularly varying distributions and subexponential distributions. We do not discuss the exact mathematical definition of these classes here, but these definitions can be found in Chapter 4. For now the following notions suffice, a distribution function is called regularly varying if the survivor function of this distribution looks like the survivor function of the Pareto(α) distribution for large values. The survivor function of the Pareto(α) distribution is given by

$$1 - F(x) = x^{-\alpha}$$

The parameter α is often called the tail index. A generalization of the class of regularly varying distributions is the class of subexponential distributions. A distribution function is called subexponential if the survivor function of the maximum of n of these random variables looks like the survivor function of the sum of n of these random variables. The regularly varying distribution functions are a strict subset of the subexponential distribution functions.

An important property of regularly varying distribution functions is that the m-th moment does not exist whenever $m \geq \alpha$. For a regularly varying distribution function the mean and variance can be infinite. This has a few important implications. First consider the sum of independent and identically distributed random variables that have a tail index $\alpha < 2$. This means that the variance of these random variables is infinite, and hence the central limit theorem does not hold for these random variables. Instead the generalized central limit theorem (Uchaikin and Zolotarev [1999]) holds. This theorem shows that if the tail index of a random variable is less than two, then the sum of these random variables have a stable distribution as a limiting distribution. These stable distributions are also regularly varying with the same tail index as the original random variable. When we consider a random variable that has a regularly varying distribution, with a tail index less than one, then the mean of this random variable is infinite. This limits the use of historical data to make an inference about the future. Consider for instance the moving average of such a random variable. If we consider a dataset with n observations, $X_1, X_2, X_3, \ldots$ then the k-th moving average is defined as the average of the first k observations. The moving average of a random variable, with a regularly varying distribution with tail index less than one, increases as we add more observations. From this one might conclude that there is a time trend in this plot, whilst actually one looks at a finite estimator of infinity that becomes more and more accurate. In Figures 1.1 (a)–(b) we see the moving average of respectively a



(a) Pareto(1) (b) Standard Exponential

Figure 1.1: Moving average of two datasets

Pareto(1) distribution and a Standard Exponential distribution. The samples we generated had a size of 1000. The mean of the Pareto(1) distribution is infinite whilst the mean of the Standard Exponential distributions is equal to one. As we can see the moving average of the Pareto(1) distribution shows an upward trend, whilst the moving average of the Standard Exponential distribution converges to the real mean of the Standard Exponential distribution.

1.1.1 Properties of heavy-tailed distributions

In this section we discuss different properties of heavy-tailed distributions and how these properties differ from the properties of thin-tailed distributions. One of the characteristics of heavy-tailed distributions is the fact that there are usually a few very large values compared to the other values of the dataset. In the insurance business this is called the Pareto law or the 20-80 rule-of-thumb. This rule states that 20% of the claims account for 80% of the total claim amount in an insurance portfolio. This can be elucidated by looking at the mean excess function. The mean excess function of a random variable X is defined in the following way.

$$e(u) = E\left[X - u | X > u\right]$$

The mean excess function gives the expected excess of a random variable over a certain threshold given that this random variable is larger than this threshold. It can be shown that for random variables with a subexponential distribution the mean excess function tends to infinity as u tends to infinity. This means that if we know that an observation from a subexponential distributions is above a very high threshold then we expect that this observation is much larger than this threshold. Another important fact of the mean excess function is that for random variables with a regularly varying distribution, with tail index $\alpha > 1$, the mean excess function is ultimately linear with slope $\frac{1}{\alpha-1}$.

Another way to measure the distance between two observations is by looking at the ratio between two observations. It turns out that for regularly varying distribution the ratio between the larger values in a dataset have a non-degenerate limiting distribution, and for distributions like the normal and exponential distribution this ratio tends to zero as we increase the number of observations. Furthermore we find that if we order a dataset from a Pareto distribution, then the ratio between two consecutive observations also has a Pareto distribution. In Table 1.1 we

Number of observations	Standard Normal Distribution	Pareto(1) Distribution
10	0.2343	$\frac{1}{2}$
50	0.0102	$\frac{1}{2}$
100	0.0020	$\frac{1}{2}$

Table 1.1: Probability that the next record value is twice as large as the previous record value for different size datasets

see the probability that the largest value in the dataset is twice as large as the second largest value for the standard normal distribution and the Pareto(1) distribution. The probability stays constant for the Pareto distribution, but it tends to zero for the standard normal distribution as the number of observations increases.

1.2 Extreme Value Theory

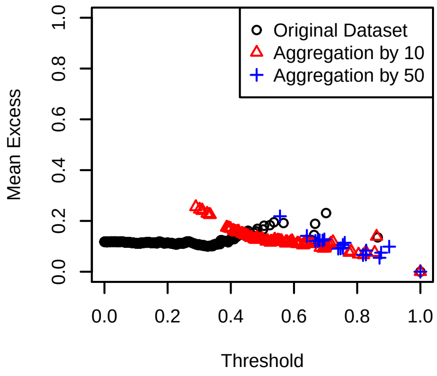
Another field where heavy-tailed distributions play an important role is in extreme value theory. Here one tries to make an inference about the limiting behaviour of the extreme values in a dataset. There are two main approaches in this field (Coles [2001]). In the block maxima method, one studies the asymptotic behaviour of the maximum of a dataset. It turns out that the only non-degenerate limiting distribution of the maximum of a distribution is the generalized extreme value distribution. The limiting distribution of the maximum of a regularly varying distribution is also regularly varying with the same tail index. One major drawback of this method is that only one observation in a block is used to make an inference about the limiting distribution of the maximum. In hydrology for example the typical size of a block is a year, so one only use one observation each year.

That is why this method is mostly superseded by the peaks-over-threshold method. Here one considers the behaviour of the distribution over a certain high threshold. It turns out that if the maxima have the generalized extreme value distribution as a limiting distribution then the exceedance over a high threshold is distributed according to the Generalized Pareto distribution. The exceedances of a regularly varying distribution are also regularly varying with the same tail index. But this method also has a drawback, since one of the assumptions made in the theory is the fact that the observations need to be independent. In many natural processes however there is a time dependence, for instance when one considers the height of the water in a river. There is a build up to a high water level and then the water level stays high for a while before it returns to the normal level. This means that if we use the peaks-over-threshold method that we see exceedances that are clustered and hence the observations are not independent. This problem is usually countered by using methods to identify these clusters and taking the largest value in this cluster as one observation.

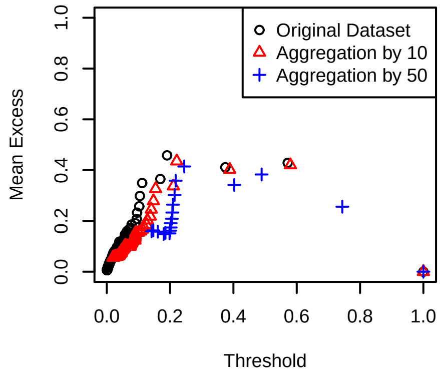
1.3 Self-similarity

The sums of regularly varying distributions with infinite variance converge to a stable distribution with the same tail index. This can be observed in the mean excess plot of a dataset from a regularly varying distribution. In the mean excess plot the empirical mean excess function of a dataset is plotted. Define the operation aggregating by k as dividing a dataset into groups of size k and summing each of these k values. If we consider a dataset of size n and compare the mean excess plot of this dataset with the mean excess plot of a dataset we obtained through aggregating the original dataset by k, then we find that both mean excess plots are very similar. Whilst for datasets from thin-tailed distributions both mean excess plots look very different. In order to compare the shapes of the mean excess plots we have standardized the data such that the largest value in the dataset we consider is equal to one. This does not change the shape of the mean excess plot, since we can easily see that $e(cu) = ce(u)$. In Figure 1.2 (a)–(d) we see the standardized mean excess plot of a sample from an exponential distribution, a Pareto(1) distribution, a Pareto(2) distribution and a Weibull distribution with shape parameter 0.5, the sample size is equal to 1000 for all samples. Together with the standardized mean excess plots of a dataset acquired through aggregating by 10 and 50. The Weibull distribution is a subexponential distribution whenever the shape parameter $\tau < 1$. Aggregating by k for the exponential distribution leads to a collapse of the slope of the standardized mean excess plot. For the Pareto(1) distribution aggregating the sample does not have a big effect on the mean excess plot, and for the Pareto(2) distribution we see that by taking large groups to sum the slope collapses. The same can also be observed for the dataset from a Weibull distribution. This means that whenever we encounter a regularly varying distribution we can see whether its tail index is less than two by comparing the mean excess plot of the original dataset with the mean excess plot of a aggregated dataset and checking whether the slope of the mean excess plot changed a lot. We can observe the same behaviour when we look at the mean excess plot of a dataset. Figures 1.3 (a)–(b) show the standardized mean excess plot for two datasets. The standardized mean excess plot in Figure 1.3a is based upon the Income- and Exposure-adjusted flood claims from the National Flood Insurance program in the United States from the years 1980 to 2006. From now on we refer to this dataset by NFIP. The second dataset we consider is taken from the National Crop Insurance Data. This dataset, maintained by the US Department of Agriculture, contains all crop insurance payments by county over the years 1979-2008. After closer inspection of the database only the years 1980-2008 for 2898 counties across the USA contain useful data entries. Our dataset contains all pooled values per county with claim sizes larger than \$ 1.000.000,-. The standardized mean excess plot of the NFIP database in Figure 1.3a seems to stay the same as we aggregate the dataset. This indicates that this dataset has infinite variance. The standardized mean excess plot of the National Crop Insurance data in Figure 1.3b changes a lot when taking random aggregations, which indicates a heavy-tailed distribution with finite variance.

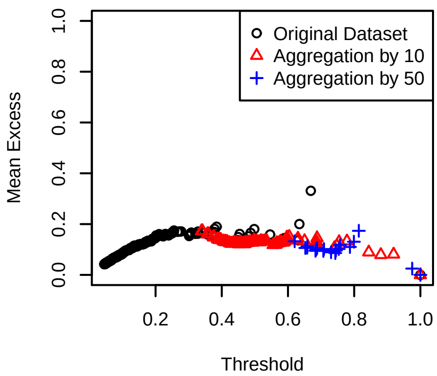
The self-similarity of heavy-tailed distributions was used in Crovella and Taqqu [1999] to construct an estimator for the tail index. In this paper he plotted the empirical survivor function



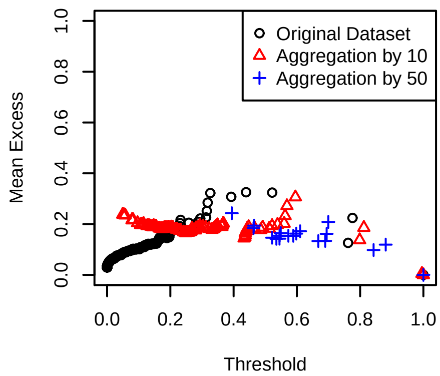
(a) Exponential Distribution



(b) Pareto $\alpha = 1$



(c) Pareto $\alpha = 2$



(d) Weibull distribution $\tau = 0.5$

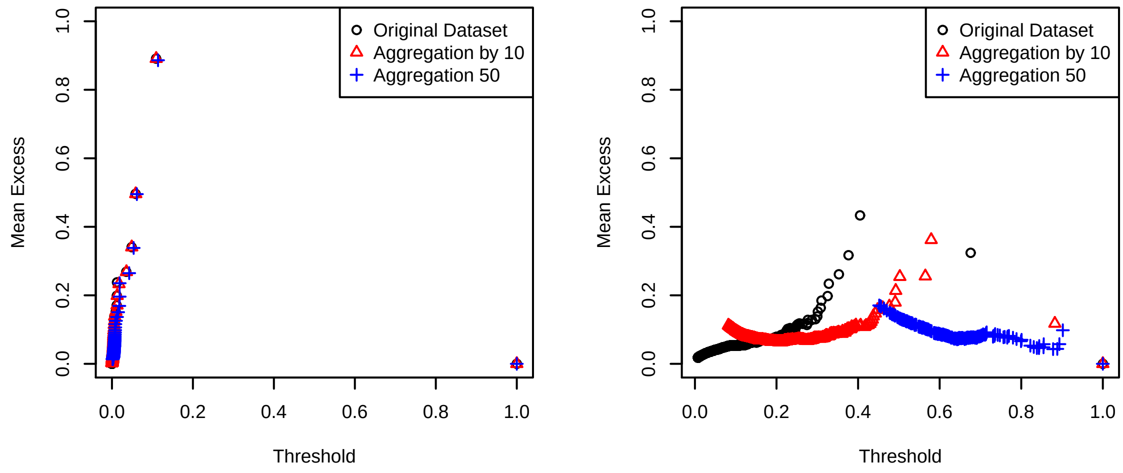
Figure 1.2: Standardized mean excess plots

on log-log axes of different levels of aggregating and searched for a proper region to fit a straight line through. This estimator did not perform too well when compared to the Hill estimator and could not discriminate between subexponential distributions and regularly varying distributions. The Hill estimator is a widely used estimator of the tail index and we define it in the next section.

1.4 Estimating the tail index

One of the most widely used classes of heavy-tailed distributions is the class of regularly varying distributions. In order to make an inference using these distributions one needs to be able to estimate the tail index of a dataset. In this section we review different methods to estimate the tail index from a dataset.

One of the simplest methods is by plotting the empirical survivor function on log-log axes and fitting a straight line above a certain threshold. The slope of this line is then used to



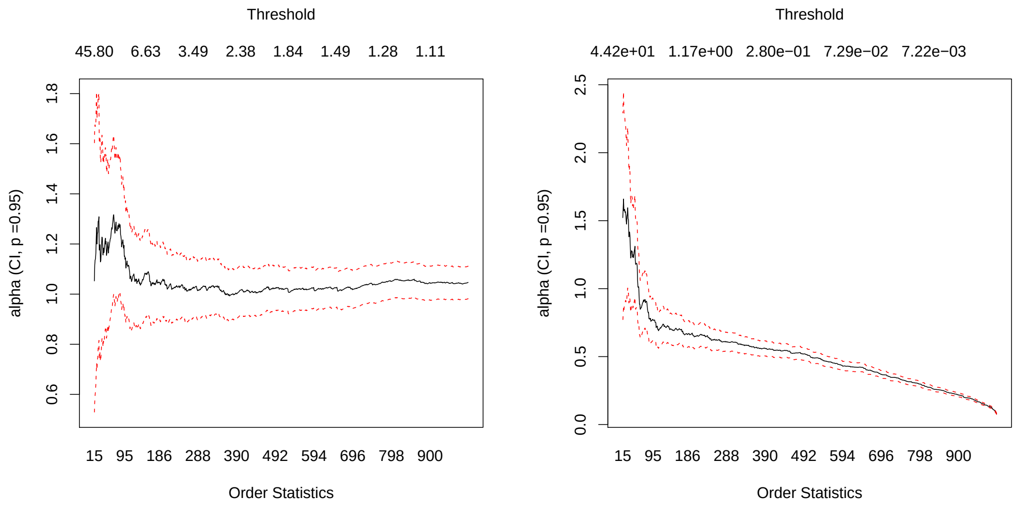
(a) NFIP (b) National Crop Insurance

Figure 1.3: Standardized mean excess plots of two datasets

estimate the tail index. A drawback of this method is that it is not clear how to choose the threshold. Another widely used estimator of the tail index is the Hill estimator. This estimator was proposed in Hill [1975] and is given by

$$\mathcal{H}_{k,n} = \frac{1}{k}\sum_{i=0}^{k-1}\left(\log(X_{n-i,n}) - \log(X_{n-k,n})\right),$$

where $X_{i,n}$ are such that $X_{1,n} \leq ... \leq X_{n,n}$. The tail index is estimated by $\frac{1}{\mathcal{H}_{k,n}}$. The idea behind this method is that if a random variable has a Pareto distribution then the log of this random variable has an exponential distribution with parameter equal to the tail index. The Hill estimator is then an estimator of the parameter of this exponential distribution. The Hill estimator has also a few drawbacks. First of all the Hill estimator depends on the value of k, and it is not clear which value of k needs to be chosen for the best estimate. A useful heuristic here is that k is usually less than $0.1 \cdot n$. There exist methods that choose k by minimizing the asymptotic mean squared error of the Hill estimator. Another drawback of the Hill estimator is that it works very well for Pareto distributed data, but for other regularly varying distribution functions the Hill estimator becomes less effective. To illustrate this we have drawn two different samples, one from the Pareto(1) distribution and one from a Burr distribution with parameters such that the tail index of this Burr distribution is equal to one. In Figure 1.4 (a)–(b) we see the Hill estimator for the two datasets together with the 95%-confidence bounds of the estimate. As we can see from Figure 1.4a, the Hill estimator gives a good estimate of the tail index, but from Figure 1.4b it is not clear that the tail index is equal to one. Finally note that in Figure 1.4 the Hill estimate is plotted against the different values in the dataset and the largest value of the dataset is plotted on the left of the x-axis. In Beirlant et al. [2005] various improvements of the Hill estimator are given, but these improvements require extra assumptions on the distribution of the dataset.



(a) Pareto distributed data (b) Burr distributed data

Figure 1.4: Hill estimator for samples of a Pareto and Burr distribution with tail index 1.

Distribution	Obesity index
Uniform	0.5
Exponential	0.75
Pareto(1)	$\pi^2 - 9$

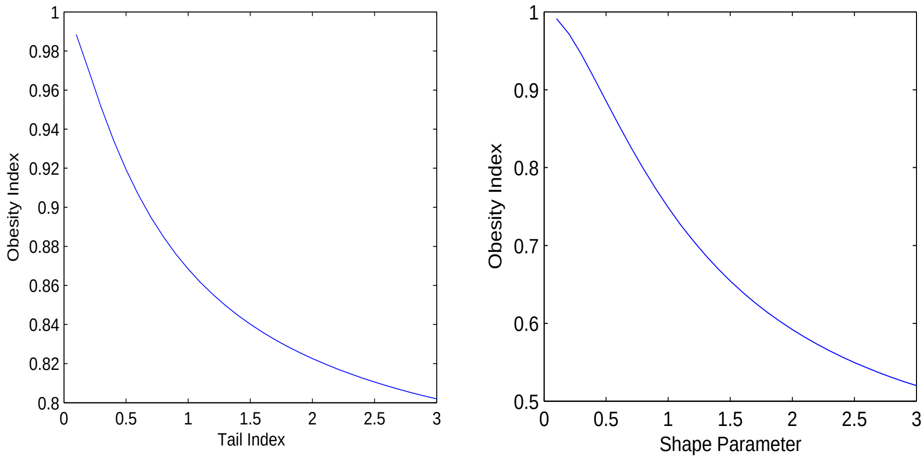
Table 1.2: Obesity index for a number of distributions

1.5 The Obesity Index

We have discussed two different classes of heavy-tailed distributions, the regularly varying distributions and subexponential distributions. We saw that the tail index of a regularly varying distributions could be used to characterize the heavy-tailedness of these distributions. The drawback of using the tail index as a characterization of the heavy-tailedness of a distribution is the fact that it is difficult to estimate the tail index from a dataset. This is due to the fact that the tail index is a parameter that can only be observed at infinity and not directly from a dataset. For the subexponential distributions there is not an index that measures the heavy-tailedness of a subexponential distribution. In this thesis we will search for in index that measures the heavy-tailedness of a distribution using an index that does not refer to the limiting behavior of a distribution. In this thesis we propose to use the following estimator as a measure of heavy-tailedness

$$\mathrm{Ob}(X) = P\left(X_1 + X_4 > X_2 + X_3 | X_1 \leq X_2 \leq X_3 \leq X_4\right), \qquad X_i \sim X, i = 1,2,3,4.$$

In table 1.2 the value of the Obesity index is given for a number of different distributions. In Figure 1.5 we see the Obesity index for the Pareto distribution, with tail index α, and for the Weibull distribution with shape parameter τ. From Figures 1.5a and 1.5b we find that the Obesity index follows the notion that if a Pareto distribution has a small tail index then the distribution is heavy-tailed. The same holds for the Weibull distribution, if $\tau < 1$ then the Weibull is a subexponential distribution and is considered heavy-tailed. The Obesity index increases as τ decreases. The question arises if we consider two random variables X_1 and X_2, which have the same regularly varying distribution function but with different tail indexes, α_1



(a) Pareto Distribution (b) Weibull Distribution

Figure 1.5: Obesity index for different distributions.

and α_2 respectively and for which $\alpha_1 < \alpha_2$ is the Obesity index of X_1 larger than the Obesity index of X_2. Numerical approximation of two Burr distributed random variables indicate that this is not the case. Consider X_1, a Burr distributed random variable with parameters $c = 1$ and $k = 2$, and a Burr distributed random variable with parameters $c = 3.9$ and $k = 0.5$. The tail index of X_1 is equal to 2 and the tail index of X_2 is equal to 1.95. But numerical approximation indicate that the Obesity index of X_1 is approximately equal to 0.8237 and the Obesity index of X_2 is approximately equal to 0.7463.

Another issue of the Obesity index is that for any symmetrical random variable the Obesity index is equal to $\frac{1}{2}$. This means the Obesity index of the normal distribution and the Cauchy distribution both are equal to $\frac{1}{2}$, but the Cauchy distribution is a regularly varying distribution with tail index 1 and the normal distribution is considered a thin-tailed distribution. This limits the use of the Obesity index to positive random variables, which is the case in many applications like insurance.

1.6 Outline of the thesis

In Chapter 2 and 3 we discuss different properties of order statistics from a distribution and some results from the theory of records. These results shall be used in Chapter 5 to derive different properties of the index we propose. And in Chapter 4 we discuss some different definitions of heavy-tailed distributions and the relationship between these definitions.

Chapter 2

Order Statistics

In this chapter we discuss some properties of order statistic. We use these properties later on to derive properties of the Obesity index. Most of these properties can be found in David [1981] or Nezvorov [2001]. We only consider order statistics from an i.i.d. sequence of continuous random variables. Now suppose we have a sequence of n continuous random variables $X_1, ..., X_n$ which are independent and identically distributed. If we order this sequence in ascending order we obtain the order statistics

$$X_{1,n} \leq ... \leq X_{n,n}.$$

2.1 Distribution of order statistics

In this section we derive the marginal and joint distribution of an order statistic. The distribution function of the r-th order statistic $X_{r,n}$, from a sample of a random variable X with distribution function F, is given by

$$\begin{aligned} F_{r,n}(x) &= P\left(X_{r,n} \leq x\right) \\ &= P\left(\text{at least } r \text{ of the } X_i \text{ are less than or equal to } x\right) \\ &= \sum_{m=r}^{n} P\left(\text{ exactly } m \text{ variables among } X_1, ..., X_n \ \leq \ x\right) \\ &= \sum_{m=r}^{n} \binom{n}{m} F(x)^m \left(1 - F(x)\right)^{n-m} \end{aligned}$$

Using the following relationship

$$\sum_{m=k}^{n} \binom{n}{m} y^m (1-y)^{n-m} = \int_0^y \frac{n!}{(k-1)!(n-k)!} t^{k-1} (1-t)^{n-k} dt, \qquad 0 \leq y \leq 1,$$

we get the following result

$$F_{r,n}(x) = I_{F(x)}\left(r, n-r+1\right), \tag{2.1}$$

where $I_x(p, q)$ is the regularized incomplete beta function which is given by

$$I_x\left(p, q\right) = \frac{1}{B\left(p, q\right)} \int_0^x t^{p-1} (1-t)^{q-1} dt,$$

and $B(p, q)$ is the beta function which is given by

$$B(p, q) = \int_0^1 t^{p-1} (1-t)^{q-1} dt.$$

Now assume that the random variable X_i has a probability density function $f(x) = \frac{d}{dx}F(x)$. Denote the density function of $X_{r,n}$ with $f_{r,n}$. Using (2.1) we get the following result.

$$\begin{aligned} f_{r,n}(x) &= \frac{1}{B(r,n-r+1)} \frac{d}{dx} \int_0^{F(x)} t^{r-1}(1-t)^{n-r}dt, \\ &= \frac{1}{B(r,n-r+1)} F(x)^{r-1} (1-F(x))^{n-r} f(x) \end{aligned} \tag{2.2}$$

The joint density of $X_{k(1),n}, ..., X_{k(r),n}$, here $\{k(1), ..., k(r)\}$ is a subset of the numbers $1, 2, 3, ..., n$, and $k(0) = 0$, $k(r+1) = n+1$ and finally $1 \leq r \leq n$, is given by

$$\begin{aligned} f_{k(1),...,k(n);n}(x_1, ..., x_r) &= \frac{n!}{\prod_{s=1}^{r+1}(k(s)-k(s-1)-1)!}, \\ &\prod_{s=1}^{r+1} (F(x_s) - F(x_{s-1}))^{k(s)-k(s-1)-1} \prod_{s=1}^{r} f(x_s), \end{aligned} \tag{2.3}$$

where $-\infty = x_0 < x_1 < ... < x_r < x_{r+1} = \infty$. We prove this for $r = 2$ and assume for simplicity that f is continuous at the points x_1 and x_2 under consideration. Consider the following probability

$$P(\delta, \Delta) = P\left(x_1 \leq X_{k(1),n} < x_1 + \delta < x_2 \leq X_{k(2),n} < x_2 + \Delta\right).$$

We show that as $\delta \to 0$ and $\Delta \to 0$ the following limit holds.

$$f(x_1, x_2) = \lim \frac{P(\delta, \Delta)}{\delta\Delta}$$

Now define the following events

$$\begin{aligned} A &= \{x_1 \leq X_{k(1),n} < x_1 + \delta < x_2 \leq X_{k(2),n} < x_2 + \Delta \text{ and the intervals} \\ &\quad [x_1, x_1+\delta) \text{ and } [x_2, x_2+\Delta) \text{ each contain exactly one order statistic}\}, \\ B &= \{x_1 \leq X_{k(1),n} < x_1 + \delta < x_2 \leq X_{k(2),n} < x_2 + \Delta \text{ and} \\ &\quad [x_1, x_1+\delta) \cup [x_2, x_2+\Delta) \text{ contains at least three order statistics}\}. \end{aligned}$$

We have that $P(\delta, \Delta) = P(A) + P(B)$. Also define the following events

$$\begin{aligned} C &= \{\text{at least two out of } n \text{ variables } X_1, ..., X_n \text{ fall into } [x_1, x_1+\delta)\} \\ D &= \{\text{at least two out of } n \text{ variables } X_1, ..., X_n \text{ fall into } [x_2, x_2+\Delta)\}. \end{aligned}$$

Now we have that $P(B) \leq P(C) + P(D)$. We find that

$$\begin{aligned} P(C) &= \sum_{k=2}^{n} \binom{n}{k} (F(x_1+\delta) - F(x_1))^k (1 - F(x_1+\delta) + F(x_1))^{n-k} \\ &\leq (F(x_1+\delta) - F(x_1))^2 \sum_{k=2}^{n} \binom{n}{k} \\ &\leq 2^n (F(x_1+\delta) - F(x_1))^2 \\ &= O(\delta^2), \qquad \delta \to 0, \end{aligned}$$

and similarly we obtain that

$$\begin{aligned} P(D) &= \sum_{k=2}^{n} \binom{n}{k} \left(F(x_2+\Delta) - F(x_2)\right)^k \left(1 - F(x_2+\Delta) + F(x_2)\right)^{n-k} \\ &\leq \left(F(x_2+\Delta)\right) \sum_{k=2}^{n} \binom{n}{k} \\ &\leq 2^n \left(F(x_2+\Delta) - F(x_2)\right)^2 \\ &= O\left(\Delta^2\right), \qquad \Delta \to 0. \end{aligned}$$

And so we obtain

$$\lim \frac{P(\delta,\Delta) - P(A)}{\delta\Delta} = 0 \qquad \text{as } \delta \to 0,\ \Delta \to 0.$$

So now it remains to see that

$$\begin{aligned} P(A) &= \frac{n!}{(k(1)-1)!(k(2)-k(1)-1)!(n-k(2))!} F(x_1)^{k(1)-1} \left(F(x_1+\delta) - F(x_1)\right) \\ &\quad \left(F(x_2) - F(x_1+\delta)\right)^{k(2)-k(1)-1} \left(F(x_2+\Delta) - F(x_2)\right) \left(1 - F(x_2)\right)^{n-k(2)}. \end{aligned}$$

From this equality we see that the limit exists and that

$$\begin{aligned} f(x_1,x_2) &= \frac{n!}{(k(1)-1)!(k(2)-k(1)-1)!(n-k(2))!} F(x_1)^{k(1)-1} \left(F(x_2) - F(x_1)\right)^{k(2)-k(1)-1} \\ &\quad \left(1 - F(x_2)\right)^{n-k(2)} f(x_1) f(x_2), \end{aligned}$$

which is the same as the joint distribution we wrote down earlier. Note that we only have found the right limit of $f(x_1+0, x_2+0)$, but since f is continuous we can obtain the other limits $f(x_1+0, x_2-0)$, $f(x_1-0, x_2+0)$ and $f(x_1-0, x_2-0)$ in a similar way.

Also note that when $r = n$ in (2.3) we get the joint density of all order statistics and that this joint density is given by

$$f_{1,\dots,n;n}(x_1,\dots,x_n) = \begin{cases} n! \prod_{s=1}^{n} f(x_s) & \text{if, } -\infty < x_1 < \dots < x_n < \infty \\ 0, & \text{otherwise.} \end{cases} \tag{2.4}$$

2.2 Conditional distribution

When we pass from the original random variables $X_1, \dots, X_n$ to the order statistics, we lose independence among these variables. Now suppose we have a sequence of n order statistics $X_{1,n}, \dots, X_{n,n}$, and let $1 < k < n$. In this section we derive the distribution of an order statistic $X_{k+1,n}$ given the previous order statistic $X_k = x_k, \dots, X_1 = x_1$. Let the density of this conditional random variable be denoted by $f(u|x_1, .., x_k)$. We show that this density coincides with the distribution of $X_{k+1,n}$ given that $X_{k,n} = x_k$, denoted by $f(u|x_k)$

$$\begin{aligned} f(u|x_1,...,x_k) &= \frac{f_{1,...,k+1;n}(x_1,...,x_k,u)}{f_{1,...,k;n}(x_1,...,x_k)} \\ &= \frac{\frac{n!}{(n-k-1)!}\left[1-F(u)\right]^{n-k-1}\prod_{s=1}^{k} f(x_s)f(u)}{\frac{n!}{(n-k)!}\left[1-F(x_k)\right]^{n-k}\prod_{s=1}^{k} f(x_s)} \\ &= \frac{\frac{n!}{(k-1)!(n-k-1)!}\left[1-F(u)\right]^{n-k-1}F(x_k)^{k-1}f(x_k)f(u)}{\frac{n!}{(k-1)!(n-k)!}\left[1-F(x_k)\right]^{n-k}F(x_k)^{k-1}f(x_k)} \\ &= \frac{f_{k,k+1;n}(x_k,u)}{f_{k,n}(x_k)} = f(u|x_k). \end{aligned}$$

From this we see that the order statistics form a Markov chain. The following theorem is an important theorem when one tries to find the distribution of functions of order statistics.

Theorem 2.2.1. *Let $X_{1,n} \leq ... \leq X_{n,n}$ be order statistics corresponding to a continuous distribution function F. Then for any $1 < k < n$ the random vectors*

$$X^{(1)} = (X_{1,n},...,X_{k-1,n}) \text{ and } X^{(2)} = (X_{k+1,n},...,X_{n,n})$$

are conditionally independent given any fixed value of the order statistic $X_{k,n}$. Furthermore, the conditional distribution of the vector $X^{(1)}$ given that $X_{k,n} = u$ coincides with the unconditional distribution of order statistics $Y_{1,k-1},...,Y_{k-1,k-1}$ corresponding to i.i.d. random variables $Y_1,...,Y_{k-1}$ with distribution function

$$F^{(u)}(x) = \frac{F(x)}{F(u)} \qquad x < u.$$

Similarly, the conditional distribution of the vector $X^{(2)}$ given $X_{k,n} = u$ coincides with the unconditional distribution of order statistics $W_{1,n-k},...,W_{n-k;n-k}$ related to the distribution function

$$F_{(u)}(x) = \frac{F(x)-F(u)}{1-F(u)} \qquad x > u.$$

Proof. To simplify the proof we assume that the underlying random variables $X_1,...,X_n$ have density f. The conditional density is given by

$$\begin{aligned} f(x_1,...,x_{k-1},x_{k+1},...,x_n|X_{k,n}=u) &= \frac{f_{1,...,n;n}(x_1,...,x_{k-1},x_{k+1,n},...,x_n)}{f_{k;n}(u)} \\ &= \left[(k-1)!\prod_{s=1}^{k-1}\frac{f(x_s)}{F(u)}\right]\left[(n-k)!\prod_{r=k+1}^{n}\frac{f(x_r)}{1-F(u)}\right]. \end{aligned}$$

As we can see the first part of the conditional density is the joint density of the order statistics from a sample size $k-1$ where the random variables have a density $\frac{f(x)}{F(u)}$ for $x < u$. The second part in the density is the joint density of the order statistics from a sample of size $n-k$ where the random variables have a distribution $\frac{F(x)-F(u)}{1-F(u)}$ for $x > u$. □

2.3 Representations for order statistics

As we pointed out before one of the drawbacks of using the order statistics is losing the independence property among the random variables. But if we consider order statistics from the exponential distribution or the uniform distribution there are a few useful properties of the order statistics that can be used when looking at linear combinations of the order statistics.

Theorem 2.3.1. *Let $X_{1,n} \leq ... \leq X_{n,n}$, $n = 1, 2, ...,$ be order statistics related to independent and identically distributed random variables with distribution function F, and let*

$$U_{1,n} \leq ... \leq U_{n,n},$$

be order statistics related to a sample from the uniform distribution on $[0, 1]$. Then for any $n = 1, 2, ...$ the vectors $(F(X_{1,n}), ..., F(X_{n,n}))$ and $(U_{1,n}, ..., U_{n,n})$ are equally distributed.

Theorem 2.3.2. *Consider exponential order statistics*

$$Z_{1,n} \leq ... \leq Z_{n,n},$$

related to a sequence of independent and identically distributed random variables $Z_1, Z_2, ...$ with distribution function

$$H(x) = \max\left(0, 1 - e^{-x}\right).$$

Then for any $n = 1, 2, ...$ we have

$$(Z_{1,n}, ..., Z_{n,n}) \stackrel{d}{=} \left(\frac{v_1}{n}, \frac{v_1}{n} + \frac{v_2}{n-1}, ..., \frac{v_1}{n} + ... + v_n\right), \tag{2.5}$$

where $v_1, v_2, ...$ is a sequence of independent and identically distributed random variables with distribution function $H(x)$.

Proof. In order to prove Theorem 2.3.2 it suffices to show that the densities of both vectors in (2.5) are equal. Putting

$$f(x) = \begin{cases} e^{-x}, & \text{if } x > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{2.6}$$

and by substituting equation (2.6) into the joint density of the n order statistics given by

$$f_{1,2,...,n;n}(x_1, ..., x_n) = \begin{cases} n! \prod_{i=1}^{n} f(x_i), & x_1 < ... < x_n, \\ 0, & \text{otherwise,} \end{cases}$$

we find that the joint density of the vector on the LHS of equation (2.5) is given by

$$f_{1,2,...,n;n}(x_1, ..., x_n) = \begin{cases} n! \exp\{-\sum_{s=1}^{n} x_s\}, & \text{if } 0 < x_1 < ... < x_n < \infty, \\ 0, & \text{otherwise.} \end{cases} \tag{2.7}$$

The joint density of n i.i.d. standard exponential random variables $v_1, ..., v_n$ is given by

$$g(y_1, ..., y_n) = \begin{cases} \exp\{-\sum_{s=1}^{n} y_s\}, & \text{if } y_1 > 0, ..., y_n > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2.8}$$

The linear change of variables

$$(v_1, ..., v_n) = (\frac{y_1}{n}, \frac{y_1}{n} + \frac{y_2}{n-1}, \frac{y_1}{n} + \frac{y_2}{n-1} + \frac{y_3}{n-2}, ..., \frac{y_1}{n} + ... + y_n)$$

with Jacobian $\frac{1}{n!}$ which corresponds to the passage to random variables

$$V_1 = \frac{v_1}{n}, V_2 = \frac{v_1}{n} + \frac{v_2}{n-1}, ..., V_n = \frac{v_1}{n} + ... + v_n,$$

has the property that

$$v_1 + v_2 + ... + v_n = y_1 + ... + y_n$$

and maps the domain $\{y_s > 0 s = 1, ..., n\}$ into the domain $\{0 < v_1 < v_2 < ... < v_n < \infty\}$. Equation (2.8) implies that $V_1, ..., V_n$ have the joint density

$$f(v_1, ..., v_n) = \begin{cases} n! \exp\left\{-\sum_{s=1}^{n} v_s\right\}, & \text{if } 0 < v_1 < ... < v_n, \\ 0, & \text{otherwise.} \end{cases} \tag{2.9}$$

Comparing equation (2.7) with equation (2.9) we find that both vectors in (2.5) have the same density and this proves the theorem. □

Using Theorem 2.3.2 it is possible to find the distribution of any linear combination of order statistics from an exponential distribution, since we can express this linear combination as a sum of independent exponential distributed random variables.

Theorem 2.3.3. *Let $U_{1,n} \leq ... \leq U_{n,n}$, $n = 1, 2, ...$ be order statistics from an uniform sample. Then for any $n = 1, 2, ...$*

$$(U_{1,n}, ..., U_{n,n}) \stackrel{d}{=} \left(\frac{S_1}{S_{n+1}}, ..., \frac{S_n}{S_{n+1}}\right),$$

where

$$S_m = v_1 + ... + v_m, \qquad m = 1, 2, ...,$$

and where $v_1, ..., v_m$ are independent standard exponential random variables.

2.4 Functions of order statistics

In this section we discuss different techniques that can be used to obtain the distribution of different functions of order statistics.

2.4.1 Partial sums

Using Theorem 2.2.1 we can obtain the distribution of sums of consecutive order statistics, $\sum_{i=r+1}^{s-1} X_{i,n}$. The distribution of the order statistics $X_{r+1,n}, ..., X_{s-1,n}$ given that $X_{r,n} = y$ and $X_{s,n} = z$ coincides with the unconditional distribution of order statistics $V_{1,n}, ..., V_{s-r-1}$ corresponding to an i.i.d. sequence $V_1, ..., V_{s-r-1}$ where the distribution function of V_i is given by

$$V_{y,z}(x) = \frac{F(x) - F(y)}{F(z) - F(y)}, \qquad y < x < z. \tag{2.10}$$

From Theorem 2.2.1 we can write the distribution function of the partial sum in the following way

$$\begin{aligned} P\left(X_{r+1} + ... + X_{s-1} < x\right) &= \int_{-\infty<y<z<\infty} P\left(X_{r+1} + ... + X_{s-1} < x | X_{r,n} = y, X_{s,n} = z\right) f_{r,s;n}(y,z) dydz \\ &= \int_{-\infty<y<z<\infty} V_{y,z}^{(s-r-1)*}(x) f_{r,s;n}(y,z) dydz, \end{aligned}$$

where $V_{y,z}^{(s-r-1)*}(x)$ denotes the $s - r - 1$-th convolution of the distribution given by (2.10).

2.4.2 Ratio between order statistics

Now we look at the distribution of the ratio between two order statistics.

Theorem 2.4.1. *For $r < s$ and $0 \leq x \leq 1$*

$$P\left(\frac{X_{r,n}}{X_{s,n}} \leq x\right) = \frac{1}{B(s, n-s+1)} \int_0^1 I_{Q_x(t)}(r, s-r) t^{s-1}(1-t)^{n-s} dt, \tag{2.11}$$

where

$$Q_x(t) = \frac{F\left(xF^{-1}(t)\right)}{t}.$$

Proof.

$$\begin{aligned}
P\left(\frac{X_{r,n}}{X_{s,n}} \leq x\right) &= \int_{-\infty}^{\infty} P\left(\frac{y}{X_{s,n}} \leq x | X_{r,n} = y\right) f_{X_{r,n}}(y) dy, \\
&= \int_{-\infty}^{\infty} P\left(X_{s,n} > \frac{y}{x} | X_{r,n} = y\right) f_{X_{r,n}}(y) dy, \\
&= \int_{-\infty}^{\infty} \int_{\frac{y}{x}}^{\infty} f_{X_{s,n}|X_{r,n}=y}(z) dz f_{X_{r,n}}(y) dy, \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{zx} f_{X_{r,n}}(y) f_{X_{s,n}|X_{r,n}=y}(z) dy dz, \\
&= C \int_{-\infty}^{\infty} \int_{-\infty}^{zx} F(y)^{r-1} \left[1 - F(y)\right]^{n-r} f(y) \\
&\quad \frac{\left[F(z) - F(y)\right]^{s-r-1} \left[1 - F(z)\right]^{n-s} f(z)}{\left[1 - F(y)\right]^{n-r}} dy dz,
\end{aligned}$$

where $C = \frac{1}{B(r,n-r+1)B(s-r,n-s+1)}$. We apply the transformation $t = F(z)$ from which we get the following

$$P\left(\frac{X_{r,n}}{X_{s,n}} \leq x\right) = C \int_0^1 \int_{-\infty}^{xF^{-1}(t)} F(y)^{r-1} f(y) \left[t - F(y)\right]^{s-r-1} dy \,(1-t)^{n-s} \,dt.$$

Next we use the transformation $\frac{F(y)}{t} = u$.

$$\begin{aligned}
P\left(\frac{X_{r,n}}{X_{s,n}} \leq x\right) &= C \int_0^1 \int_0^{\frac{F(xF^{-1}(t))}{t}} t^{r-1} u^{r-1} \left(t - tu\right)^{s-r-1} t du \,(1-t)^{n-s} \,dt, \\
&= C \int_0^1 \int_0^{\frac{F(xF^{-1}(t))}{t}} u^{r-1} \left(1-u\right)^{s-r-1} du t^{s-1} \left(1-t\right)^{n-s} dt
\end{aligned}$$

We can rewrite the constant C in the following way

$$\begin{aligned}
C &= \frac{1}{B(r, n-r+1)B(s-r, n-s+1)} \\
&= \frac{n!}{(r-1)!(n-r)!} \frac{(n-r)!}{(s-r-1)!(n-s)!} \\
&= \frac{1}{(s-r-1)!(r-1)!} \frac{n!}{(n-s)!} \\
&= \frac{(s-1)!}{(s-r-1)!(r-1)!} \frac{n!}{(n-s)!(s-1)!} \\
&= \frac{1}{B(r, s-r)B(s, n-s+1)}.
\end{aligned}$$

If we substitute this in our integral, and define $Q_x(t) = \frac{F(xF^{-1}(t))}{t}$, we get the following

$$\begin{aligned} P\left(\frac{X_{r,n}}{X_{s,n}} \le x\right) &= \frac{1}{B(s,n-s+1)} \int_0^1 \frac{\int_0^{Q_x(t)} u^{r-1}(1-u)^{s-r-1}\, du}{B(s,s-r)} t^{s-1}(1-t)^{n-s}\, dt \\ &= \frac{1}{B(s,n-s+1)} \int_0^1 I_{Q_x(t)}(r,s-r) t^{s-1}(1-t)^{n-s}\, dt. \end{aligned}$$

□

In this chapter we looked at the distribution of order statistics and derived different properties of order statistics. We use these properties in chapter 5 to derive properties of the obesity index and the distribution of the ratio between order statistics. In the next chapter we review the theory of records which we use in Chapter 5.

Chapter 3

Records

In this chapter we discuss the theory of records, in Chapter 5 we use these results to define an estimator and derive some properties of this estimator. This theory is closely related to the field of order statistics. We give a short summary of the main results. For a more detailed discussion see Arnold et al. [1998] or Nezvorov [2001], where most of the results we present here can be found.

3.1 Standard record value processes

Let $X_1, X_2, ...$ be an infinite sequence of independent and identically distributed random variables. Denote the cumulative distribution function of these random variables by F and assume it is continuous. An observation is called an upper record value if its value exceeds all previous observations. So X_j is an upper record if $X_j > X_i$ for all $i < j$. We are also interested in the times at which the record values occur. For convenience assume that we observe X_j at time j. The record time sequence $\{T_n, n \geq 0\}$ is defined as

$$T_0 = 1 \text{ with probability } 1$$

and for $n \geq 1$,

$$T_n = \min\left\{j : X_j > X_{T_{n-1}}\right\}.$$

The record value sequence $\{R_n\}$ is then defined by

$$R_n = X_{T_n}, \qquad n = 0, 1, 2, ...$$

The number of records observed at time n is called the record counting process $\{N_n, n \geq 1\}$ where

$$N_n = \{\text{number of records among } X_1, ..., X_n\}.$$

We have that $N_1 = 1$ since X_1 is always a record.

3.2 Distribution of record values

Let the record increment process be defined by

$$J_n = R_n - R_{n-1}, \qquad n > 1,$$

with $J_0 = R_0$. It can easily be shown that if we consider the record increment process from a sequence of i.i.d. standard exponential random variables then all the J_n are independent and

J_n has a standard exponential distribution. Using the record increment process we are able to derive the distribution of the n-th record from a sequence of i.i.d. standard exponential distributed random variables.

$$\begin{aligned} P\left(R_n < x\right) &= P\left(R_n - R_{n-1} + R_{n-1} - R_{n-2} + R_{n-2} - ... + R_1 - R_0 + R_0 < x\right) \\ &= P\left(J_n + J_{n-1} + ... + J_0 < x\right) \end{aligned}$$

Since $\sum_{i=0}^{n} J + i$ is the sum of $n+1$ standard exponential distributed random variables we find that the record values from a sequence of standard exponential distributed random variables has the gamma distribution with parameters $n+1$ and 1.

$$R_n \sim \text{Gamma}(n+1, 1),\ n = 0, 1, 2, ...$$

If a random variable X has a Gamma(n, λ) distribution then it has the following density function

$$f_X(x) = \begin{cases} \frac{\lambda(\lambda x)^{n-1} e^{-\lambda x}}{\Gamma(n)}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

We can use the result above to find the distribution of the n-th record corresponding to a sequence $\{X_i\}$ of i.i.d. random variables with continuous distribution function F. If X has distribution function F then

$$H(X) \equiv -\log(1 - F(X))$$

has a standard exponential distribution function. We also have that $X \stackrel{d}{=} F^{-1}(1 - e^{-X^*})$ where X^* is a standard exponential random variable. Since X is a monotone function of X^* we can express the n-th record of the sequence $\{X_j\}$ as a simple function of the n-th record of the sequence $\{X^*\}$. This can be done in the following way

$$R_n \stackrel{d}{=} F^{-1}(1 - e^{-R_n^*}), \qquad n = 0, 1, 2, ...$$

Using the following expression of the distribution of the n-th record from a standard exponential sequence

$$P\left(R_n^* > r^*\right) = e^{-r^*} \sum_{k=0}^{n} \frac{(r^*)^k}{k!}, \qquad r^* > 0,$$

the survival function of the record from an arbitrary sequence of i.i.d. random variables with distribution function F is given by

$$P\left(R_n > r\right) = [1 - F(r)] \sum_{k=0}^{n} \frac{-\left[\log(1 - F(r))\right]^k}{k!}.$$

3.3 Record times and related statistics

The definition of the record time sequence $\{T_n, n \geq 0\}$ was given by

$$T_0 = 1,\ \text{with probability } 1,$$

and for $n \geq 1$

$$T_n = \min\{j : X_j > X_{T_{n-1}}\}.$$

In order to find the distribution of the first n non-trivial record times $T_1, T_2, ..., T_n$ we first look at the sequence of record time indicator random variables. These are defined in the following way

$$I_1 = 1 \text{ with probability } 1,$$

and for $n > 1$

$$I_n = \mathbb{1}_{\{X_n > \max\{X_1,...,X_{n-1}\}\}}$$

So $I_n = 1$ if and only if X_n is a record value. We assume that the distribution function F, of the random variables we consider, is continuous. It is easily verified that the random variables I_n have a Bernoulli distribution with parameter $\frac{1}{n}$ and are independent of each other. The joint distribution for the first m record times can be obtained using the record indicators. For integers $1 < n_1 < ... < n_m$ we have that

$$\begin{aligned} P\left(T_1 = n_1, ..., T_m = n_m\right) &= P\left(I_2 = 0, ..., I_{n_1-1} = 0, I_{n_1} = 1, I_{n_1+1} = 0, ..., I_{n_m} = 0\right) \\ &= \left[(n_1 - 1)(n_2 - 1)...(n_m - 1)n_m\right]^{-1}. \end{aligned}$$

In order to find the marginal distribution of T_k we first review some properties of the record counting process $\{N_n, n \geq 1\}$ defined by

$$\begin{aligned} N_n &= \{\text{number of records among } X_1, ...X_n\} \\ &= \sum_{j=1}^{n} I_j. \end{aligned}$$

Since the record indicators are independent we can immediately write down the mean and the variance for N_n.

$$\begin{aligned} E\left[N_n\right] &= \sum_{j=1}^{n} \frac{1}{j}, \\ \text{Var}(N_n) &= \sum_{j=1}^{n} \frac{1}{j}\left(1 - \frac{1}{j}\right). \end{aligned}$$

We can obtain the exact distribution of N_n using the probability generating function. We have the following result.

$$\begin{aligned} E\left[s^{N_n}\right] &= \prod_{j=1}^{n} E\left[s^{I_j}\right] \\ &= \prod_{j=1}^{n}\left(1 + \frac{s-1}{j}\right) \end{aligned}$$

From this we find that

$$P(N_n = k) = \frac{S_n^k}{n!}$$

where S_n^k is a Stirling number of the first kind. The Stirling numbers of the first kind are given by the coefficients in the following expansion .

$$(x)_n = \sum_{k=0}^{n} S_n^k x^k,$$

where $(x)_n = x(x-1)(x-2)...(x-n+1)$. The record counting process N_n follows the central limit theorem.

$$\frac{N_n - \log(n)}{\sqrt{\log(n)}} \xrightarrow{d} N(0,1)$$

We can use the information about the record counting process to obtain the distribution of T_k. Note that the events $\{T_k = n\}$ and $\{N_n = k+1, N_{n-1} = k\}$ are equivalent. From this we get that

$$\begin{aligned} P\left(T_k = n\right) &= P\left(N_n = k+1, N_{n-1} = k\right) \\ &= P\left(I_n = 1, N_{n-1} = k\right) \\ &= \frac{1}{n}\frac{S_{n-1}^k}{(n-1)!} \\ &= \frac{S_{n-1}^k}{n!}. \end{aligned}$$

We also have asymptotic log-normality for T_k.

$$\frac{\log(T_k) - k}{\sqrt{k}} \xrightarrow{d} N(0,1)$$

3.4 k-records

There are two different sequences that are called k-record values in the literature. We discuss both definitions here. First define the sequence of initial ranks ρ_n given by

$$\rho_n = \#\{j : j \le n \text{ and } X_n \le X_j\}, \qquad n \ge 1.$$

We call X_n a Type 1 k-record value if $\rho_n = k$, when $n \ge k$. Denote the sequence that is generated through this process by $\left\{R_n^{(k)}\right\}$. The Type 2 k-record sequence is defined in the following way, let $T_{0(k)} = k$, $R_{0(k)} = X_{n-k+1,k}$ and

$$T_{n(k)} = \min\left\{j : j > T_{(n-1)(k)}, X_j > X_{T_{(n-1)(k)}-k+1, T_{(n-1)(k)}}\right\},$$

and define $R_{n(k)} = X_{T_{n(k)-k+1}}$ as the n-th k-record. Here a k record is established whenever $\rho_n \ge k$. Although the corresponding X_n does not need to be a Type 2 k-record, unless $k = 1$, but the observation eventually becomes a Type 2 k-record value. The sequence $\left\{R_{n(k)}, n \ge 0\right\}$ from a distribution F is identical in distribution to a record sequence $\{R_n, n \ge 0\}$ from the distribution function $F_{1,k}(x) = 1-(1-F(x))^k$. So all the distributional properties of the record values and record counting statistics do extend to the corresponding k-record sequences.

The difference between the Type 1 and Type 2 k-records can also be explained in the following way. We only observe a new Type 1 k-record whenever an observation is exactly the k-th largest seen yet. Whilst we also observe a new Type 2 k-record whenever we observe a new value that is larger than the previous k-th largest yet.

Chapter 4

Heavy-tailed distribtutions

In this chapter we discuss a number of classes of heavy-tailed distributions and look into the relationships between these classes. We also discuss different properties of these classes and how these are related to the properties of heavy-tailed distributions that we discussed in Chapter 1.

4.0.1 Regularly varying distribution functions

An important class of heavy-tailed distributions is the class of regularly varying distribution functions. A distribution function is called regular varying at infinity with index $-\alpha$ if the following limit holds.

$$\lim_{x\to\infty} \frac{\overline{F}(tx)}{\overline{F}(x)} = t^{-\alpha},$$

where $\overline{F}(x) = 1 - F(x)$. The parameter α is sometimes referred to as the tail index.

Regularly varying functions

In this section we discuss some results from the theory of regularly varying function. For a more detailed discussion about the theory of regularly variation we refer to Bingham et al. [1987].

Definition 4.0.1. *A positive measurable function h on $(0,\infty)$ is regularly varying at infinity with index $\alpha \in \mathbb{R}$ if the following limit holds*

$$\lim_{x\to\infty} \frac{h(tx)}{h(x)} = t^{\alpha}, \qquad t > 0. \tag{4.1}$$

We write $h(x) \in \mathcal{R}_\alpha$. If $\alpha = 0$ we call the function slowly varying at infinity.

Instead of writing that $h(x)$ is a regularly varying function at infinity with index α we simply call the function $h(x)$ regularly varying. If $h(x) \in \mathcal{R}_\alpha$ then we can rewrite the function $h(x)$ in the following way

$$h(x) = x^{\alpha} L(x),$$

where $L(x)$ is a slowly varying function. The following theorem, which is called Karamata's theorem, is an important tool when looking at the behaviour of regularly varying functions.

Theorem 4.0.1. *Let $L \in \mathcal{R}_0$ be locally bounded in $[x_0,\infty)$ for some $x_0 \geq 0$. Then*

- *for $\alpha > -1$,*

$$\int_{x_0}^{x} t^{\alpha} L(t)dt \sim (\alpha+1)^{-1} x^{\alpha+1} L(x), \qquad x \to \infty,$$

- *for* $\alpha < -1$

$$\int_x^\infty t^\alpha L(t)dt \sim -(\alpha+1)^{-1}x^{\alpha+1}L(x), \qquad x \to \infty.$$

Regular variation for distribution functions

The class of regularly varying distributions is an important class of heavy-tailed distributions. This class is closed under convolutions as can be found in Applebaum [2005], where the result was attributed to G. Samorodnitsky.

Theorem 4.0.2. *If* X *and* Y *are independent real-valued random variables with* $\overline{F}_X \in \mathcal{R}_{-\alpha}$ *and* $\overline{F}_Y \in \mathcal{R}_{-\beta}$*, where* $\alpha, \beta > 0$*, then* $\overline{F}_{X+Y} \in \mathcal{R}_\rho$*, where* $\rho = \min\{\alpha, \beta\}$*.*

The same theorem, but with the assumption that $\alpha = \beta$ can be found in Feller [1971].

Proposition 4.0.3. *If* F_1 *and* F_2 *are two distribution functions such that as* $x \to \infty$

$$1 - F_i(x) = x^{-\alpha}L_i(x) \tag{4.2}$$

with L_i *slowly varying, then the convolution* $G = F_1 * F_2$ *has a regularly varying tail such that*

$$1 - G(x) \sim x^{-\alpha}\left(L_1(x) + L_2(x)\right). \tag{4.3}$$

From Proposition 4.0.3 we obtain the following result using induction on n.

Corollary 4.0.1. *If* $\overline{F}(x) = x^{-\alpha}L(x)$ *for* $\alpha \geq 0$ *and* $L \in \mathcal{R}_0$*, then for all* $n \geq 1$*,*

$$\overline{F^{n*}}(x) \sim n\overline{F}(x), \qquad x \to \infty.$$

Now consider an i.i.d. sample $X_1, ..., X_n$ with common distribution function F, and denote the partial sum by $S_n = X_1 + ... + X_n$ and the maximum by $M_n = \max\{X_1, ..., X_n\}$. Then for all $n \geq 2$ we find that

$$\begin{aligned} P(S_n > x) &= \overline{F^{n*}}(x) \\ P(M_n > x) &= \overline{F^n}(x) \\ &= \overline{F}(x)\sum_{k=0}^{n-1} F^k(x) \\ &= n\overline{F}(x), \qquad x \to \infty. \end{aligned}$$

From this we find that we can rewrite corollary 4.0.1 in the following way. If $\overline{F} \in \mathcal{R}_{-\alpha}$ with $\alpha \geq 0$, then we have that

$$P(S_n > x) \sim P(M_n > x) \qquad \text{as } x \to \infty.$$

This means that the tail of the distribution of the sum is determined by the tail of the maximum. In table 4.1 a number of distribution functions from the class of regularly varying distributions.

Distribution	$\overline{F}(x)$ or $f(x)$	Index of regular variation
Pareto	$\overline{F}(x) = x^{-\alpha}$	$-\alpha$
Burr	$\overline{F}(x) = \left(\frac{1}{x^{\tau}+1}\right)^{\alpha}$	$-\tau\alpha$
Log-Gamma	$f(x) = \frac{\alpha^{\beta}}{\Gamma(\beta)} (\ln(x))^{\beta-1} x^{-\alpha-1}$	$-\alpha$

Table 4.1: Regularly varying distribution functions

4.0.2 Subexponential distribution functions

A generalization of the class of regularly varying distributions is the class of subexponential distributions. In this section we discuss several properties of distributions with a subexponential tails.

Definition 4.0.2. *A distribution function F with support $(0, \infty)$ is a subexponential distribution, if for all $n \geq 2$,*

$$\lim_{x \to \infty} \frac{\overline{F^{n*}}(x)}{\overline{F}(x)} = n. \tag{4.4}$$

If F is a subexponential distribution we denote this by $F \in \mathcal{S}$.

From equation (4.4) we find the same intuitive characterisation of subexponentiality as we did for the regularly varying distributions. So for subexponential distributions we also find that the following characteristic holds

$$P(S_n > x) \sim P(M_n > x) \qquad \text{as } x \to \infty.$$

In order to check if a distribution function is a subexponential distribution we do not need to check equation (4.4) for all $n \geq 2$. Instead we can use Lemma 4.0.4 which gives a sufficient condition for subexponentiality.

Lemma 4.0.4. *If the following equation holds*

$$\limsup_{x \to \infty} \frac{\overline{F^{2*}}(x)}{\overline{F}(x)} = 2,$$

then $F \in \mathcal{S}$.

Lemma 4.0.5 gives a few important properties of subexponential distributions, these properties come from Embrechts et al. [1997].

Lemma 4.0.5. *1. If $F \in \mathcal{S}$, then uniformly in compact y-sets of $(0, \infty)$,*

$$\lim_{x \to \infty} \frac{\overline{F}(x-y)}{\overline{F}(x)} = 1. \tag{4.5}$$

2. If (4.5) holds then, for all $\varepsilon > 0$,

$$e^{\varepsilon x}\overline{F}(x) \to \infty, \qquad x \to \infty$$

3. If $F \in \mathcal{S}$ then, given $\varepsilon > 0$, there exists a finite constant K such that for all $n \geq 2$,

$$\frac{\overline{F^{n*}}(x)}{\overline{F}(x)} \leq K(1+\varepsilon)^n, \qquad x \geq 0. \tag{4.6}$$

In Table 4.2 a number of distributions are given that are subexponential. Unlike the class of regularly varying distributions the class of subexponential distributions is not closed under convolutions, a counterexample was provided in Leslie [1989].

Distribution	Tail $\overline{F}$ or density f	Parameters
Lognormal	$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{\frac{-(\ln(x)-\mu)^2}{2\sigma^2}}$	$\mu \in \mathbb{R},\ \sigma > 0$
Benktander-type-I	$\overline{F}(x) = \left(1 + 2\frac{\beta}{\alpha}\ln(x)\right) e^{-\beta(\ln(x))^2-(\alpha+1)\ln(x)}$	$\alpha,\ \kappa > 0$
Benktander-type-II	$\overline{F}(x) = e^{\frac{\alpha}{\beta}} x^{-(1-\beta)} e^{-\alpha\frac{x^\beta}{\beta}}$	$\alpha > 0,\ 0 < \beta < 1$
Weibull	$\overline{F}(x) = e^{-cx^\tau}$	$c > 0,\ 0 < \tau < 1$

Table 4.2: Distributions with subexponential tails.

4.0.3 Related classes of heavy-tailed distributions

In this section we first give two more classes of heavy-tailed distributions, after this we discuss the relationships between these classes. The first class we give is the class of dominatedly varying distribution functions denoted by $\mathcal{D}$

$$\mathcal{D} = \left\{ F \text{ d.f. on } (0,\infty) : \limsup_{x\to\infty} \frac{\overline{F}\left(\frac{x}{2}\right)}{\overline{F}(x)} < \infty \right\}$$

The final class of distribution functions we define is the class of long-tailed distributions, denoted by $\mathcal{L}$, which is defined in the following way

$$\mathcal{L} = \left\{ F \text{ d.f. on } (0,\infty) : \lim_{x\to\infty} \frac{\overline{F}(x-y)}{\overline{F}(x)} = 1 \text{ for all } y > 0 \right\}$$

The two classes of distribution functions we already discussed are the regularly varying distribution functions ($\mathcal{R}$) and the subexponential distribution functions ($\mathcal{S}$).

$$\begin{aligned} \mathcal{R} &= \left\{ F \text{ d.f. on } (0,\infty) : \overline{F} \in \mathcal{R}_{-\alpha} \text{ for some } \alpha \geq 0 \right\}, \\ \mathcal{S} &= \left\{ F \text{ d.f. on } (0,\infty) : \lim_{x\to\infty} \frac{\overline{F^{n*}}(x)}{\overline{F}(x)} = n \right\}. \end{aligned}$$

For these classes we have the following relationships

1. $\mathcal{R} \subset \mathcal{S} \subset \mathcal{L}$ and $\mathcal{R} \subset \mathcal{D}$,
2. $\mathcal{L} \cap \mathcal{D} \subset \mathcal{S}$,
3. $\mathcal{D} \nsubseteq \mathcal{S}$ and $\mathcal{S} \nsubseteq \mathcal{D}$.

4.1 Mean excess function

A tool that is widely used to detect whether a dataset shows heavy-tailed behaviour is the mean excess function. This is due to the fact that if a distribution function is subexponential the mean excess function tends to infinity. Whilst for the exponential distribution the mean excess function is a constant and for the normal distribution the mean excess function tends to zero. The mean excess function of a random variable X with finite expectation is defined in the following way.

Definition 4.1.1. *Let X be a random variable with right endpoint x_F and $E[X] < \infty$, then*

$$e(u) = E[X - u | X > u], \qquad 0 \leq u \leq x_F,$$

is called the mean excess function of X.

Distribution	Mean excess function
Exponential	$\frac{1}{\lambda}$
Weibull	$\frac{x^{1-\tau}}{\beta\tau}$
Log-Normal	$\frac{\sigma^2 x}{\ln(x)-\mu}(1+o(1))$
Pareto	$\frac{\kappa+u}{\alpha-1}, \quad \alpha > 1$
Burr	$\frac{u}{\alpha\tau-1}(1+o(1)), \quad \alpha\tau > 1$
Loggamma	$\frac{u}{\alpha-1}(1+o(1))$

Table 4.3: Mean excess functions of distributions

In insurance $e(u)$ is called the mean excess loss function. Here $e(u)$ can be interpreted as the expected claim size over some threshold u. In reliability theory or in the medical field $e(u)$ is often called the mean residual life function. In data analysis one uses the empirical counterpart of the mean excess function which is given by

$$\widehat{e}_n(u) = \frac{\sum_{i=1}^n X_{i,n} \mathbb{1}_{X_{i,n}>u}}{\sum_{i=1}^n \mathbb{1}_{X_{i,n}>u}} - u.$$

The empirical version is usually plotted against the values $u = x_{i,n}$ for $k = 1, ..., n-1$.

4.1.1 Basic properties of the mean excess function

For positive random variables the mean excess function can be calculated using the following formula.

Proposition 4.1.1. *The mean excess function of a positive model $\overline{F}$ can be calculated using the following formula*

$$e(u) = \frac{\int_u^{x_F} \overline{F}(x)dx}{\overline{F}(u)}, \qquad 0 < u < x_F,$$

where x_F is the endpoint of the distribution function F.

The mean excess function determines the distribution uniquely.

Proposition 4.1.2. *Suppose we have a continuous distribution function F, then the following relationship holds,*

$$\overline{F}(x) = \frac{e(0)}{e(x)} \exp\left\{ -\int_0^a \frac{1}{e(u)} du \right\}.$$

In table 4.3 we see the first order approximations of the mean excess function for different distribution functions. As we pointed out at the beginning of this section the mean excess plot is a widely used tool to detect whether a distribution function is subexponential or not. This is due to the following two propositions.

Proposition 4.1.3. *If a positive random variable X has a regularly varying distribution function with a tail index $\alpha > 1$, then*

$$e(u) \sim \frac{u}{\alpha - 1}, \qquad \text{as } x \to \infty.$$

Proof. Since we consider a positive random variable we can use proposition 4.1.1 to find that

$$e(u) = \frac{\int_u^{\infty} \overline{F}(x)dx}{\overline{F}(u)} \tag{4.7}$$

Since $\overline{F} \in \mathcal{R}_{-\alpha}$ there exists a slowly varying function $l(x)$ such that

$$\overline{F}(x) = x^{-\alpha} l(x) \tag{4.8}$$

Using equations (4.7) and (4.8) we find that

$$\frac{\int_u^\infty \overline{F}(x)dx}{\overline{F}(u)} = \frac{\int_u^\infty x^{-\alpha} l(x)dx}{u^{-\alpha} l(x)}. \tag{4.9}$$

From theorem 4.0.1 we find that

$$\frac{\int_u^\infty x^{-\alpha} l(x)dx}{u^{-\alpha} l(x)} \sim \frac{u}{\alpha - 1}, \qquad u \to \infty$$

□

Proposition 4.1.4. *Assume that F is the distribution function of a positive continuous random variable X which is unbounded to the right and has a finite mean. If for all $y \in \mathbb{R}$*

$$\lim_{x \to \infty} \frac{\overline{F}(x-y)}{\overline{F}(x)} = e^{\gamma y}, \tag{4.10}$$

for some $\gamma \in [0, \infty]$, then

$$\lim_{u \to \infty} e(u) = \frac{1}{\gamma}.$$

Proof. Since we are considering a positive continuous random variable we know that we can calculate the mean excess function using proposition 4.1.1

$$e(u) = \frac{\int_u^\infty \overline{F}(x)dx}{\overline{F}(u)}$$

Taking the limit to infinity with respect to u gives us the following expression

$$\begin{aligned} \lim_{u \to \infty} \frac{\int_u^\infty \overline{F}(x)dx}{\overline{F}(u)} &= \lim_{u \to \infty} \int_0^\infty \frac{\overline{F}(u+x)}{\overline{F}(u)} dx \\ &= \int_0^\infty \lim_{u \to \infty} \lim_{u \to \infty} \frac{\overline{F}(u+x)}{\overline{F}(u)} dx \\ &= \int_0^\infty e^{-\gamma x} dx \\ &= \frac{1}{\gamma}. \end{aligned} \tag{4.11}$$

We know that for all $u > 0$ the following inequality holds $\overline{F}(x) > \overline{F}(x+u)$ and

$$E[X] = \int_0^\infty \overline{F}(x)dx < \infty.$$

This means that we can apply the dominated convergence theorem and can interchange the limit and the integral in equation (4.11. □

Note that if $F \in \mathcal{S}$, then equation (4.10) is satisfied with $\gamma = 0$. If a distribution function is a subexponential distribution function the mean excess function $e(u)$ tends to infinity as $u \to \infty$. If a distribution function is regularly varying with a tail index $\alpha > 1$ then we know that the mean excess function of this distribution is eventually linear with slope $\frac{1}{\alpha-1}$. One of the drawbacks of the mean excess function is that if we consider a regularly varying distribution function with a tail index $\alpha < 1$, then the mean excess function of this distribution function does not exist. But when we plot the empirical mean excess function the slope of this plot is still finite. Another drawback of the mean excess function is that the empirical mean excess plot is very sensitive to the largest few values in the dataset.

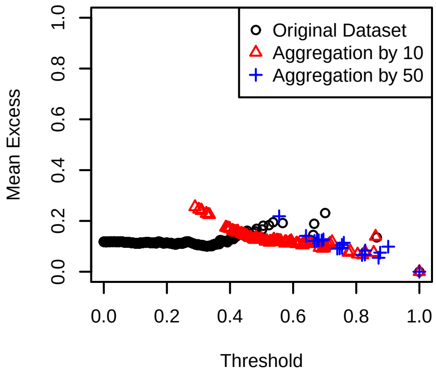
Chapter 5

Heuristics of heavy tailedness

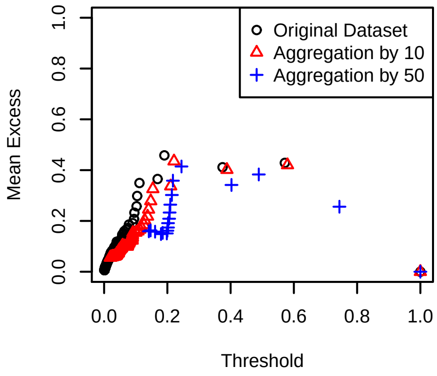
In this chapter we look for ways to use the self-similarity of the mean excess plot for heavy-tailed distributions to define a new measure of heavy-tailedness. First we look at the mean excess plot and see how the mean excess plot changes when we aggregate a dataset by k. From this we define two new measures, the first index is the ratio between the largest observation and the second-largest observation in a dataset and the second is the Obesity index which we define as the probability that the sum of the largest and the smallest observation in a dataset of four is larger than the sum of the other two observations.

5.1 Self-similarity

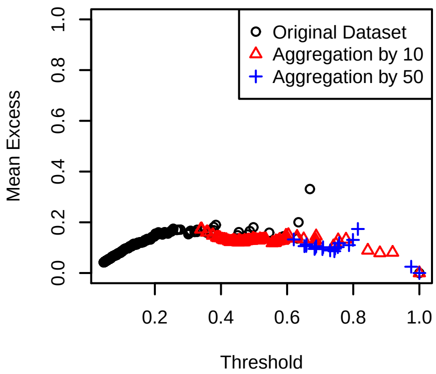
One of the heuristics we discussed in the Chapter 1 was the self-similarity of heavy-tailed distributions and how this could be seen in the mean excess plot of a distribution. Now consider a dataset of size n and create a new dataset by dividing the original dataset in to groups of size k and sum each of the k members of each group, then we obtain a new dataset. We call this operation aggregation by k. If we compare the mean excess plots of regularly varying distribution function with tail index $\alpha < 2$, then the mean excess plot of the original dataset and the dataset acquired through aggregating by k look very similar. For distributions with a finite variance the mean excess plot of the original sample and the aggregated sample look very different.



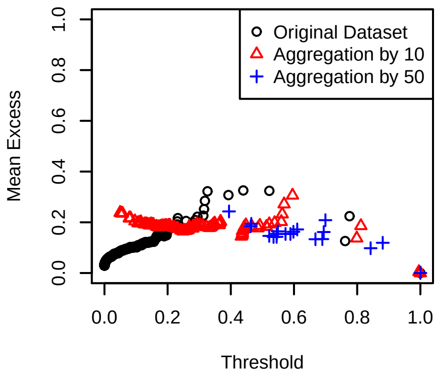
(a) Exponential Distribution



(b) Pareto $\alpha = 1$



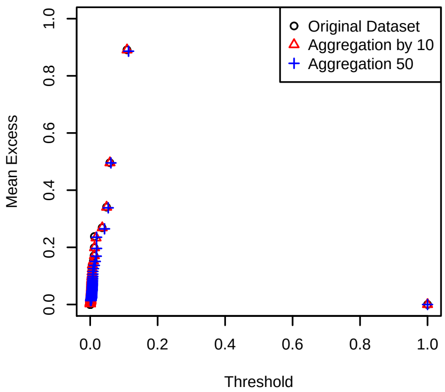
(c) Pareto $\alpha = 2$



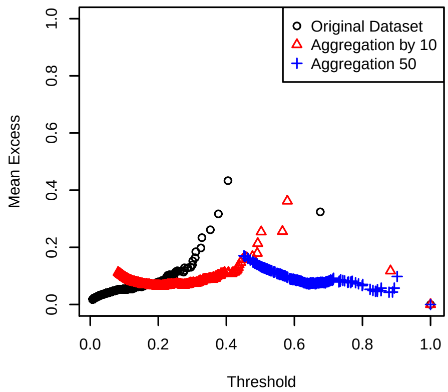
(d) Weibull distribution $\tau = 0.5$

Figure 5.1: Standardized mean excess plots

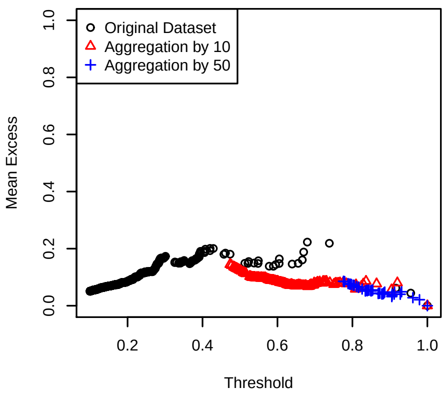
This can be explained through the generalized central limit theorem, which states that when we consider random variables with a regularly varying distribution with a tail index $\alpha < 2$ then the normalized sums of these random variables converge to a stable distribution with the same tail index. When we consider random variables with a finite variance then the normalized sums converge to a standard normal distribution for which the mean excess function tends to zero. In Figures 5.1 **??**–**??** we see the standardized mean excess plot of a number of simulated datasets of size 1000. As we can see the mean excess plots of the Exponential and Weibull distributed datasets quickly collapse under random aggregations, but the mean excess plot of the Pareto(2) distribution collapses slowly and the mean excess plot of the Pareto(1) does not change much when aggregated by 10 whilst aggregation by 50 leads to a shift in the mean excess plot but the slope stays approximately the same. Figures 5.2 (a)–(b) were also in the introduction these are the standardized mean excess plots of the NFIP database and the National Crop Insurance data. The standardized mean excess plot in Figure 5.2c is based upon a dataset that consists of the height of the bill that was written to a patient at the discharge date. Note that each of the mean excess plots in Figure 5.2 show some evidence of heavy-tailedness since each mean excess plot is increasing. But the NFIP dataset shows very heavy-tailed behaviour whilst the



(a) NFIP



(b) National Crop Insurance



(c) Hospital

Figure 5.2: Standardized mean excess plots of a few datasets

other datasets do not show very heavy-tailed behaviour since the mean excess plot collapses as we apply aggregations to the dataset. This indicates that the NFIP data has infinite variance and that the two other datasets hvae finite variance.

We have seen that when we look at the mean excess plot of a regularly varying distribution function with a tail index $\alpha < 2$ the standardized mean excess plot does not change much when we apply aggregation by k on the dataset. Consider only the largest value in the original dataset, denote this by M_n and the largest value in the dataset obtained through aggregation by k, denote this by $M_{n(k)}$. Then we know by definition that $M_n < M_{n(k)}$, but for regularly varying distributions with a small tail index the maximum of the aggregated dataset does not differ much from the original maximum. This indicates that M_n is a member of the group which produced $M_{n(k)}$. In general it is quite difficult to calculate the probability that the maximum of a dataset is contained in the group that produces $M_{n(k)}$. But we do know that, for positive random variables, whenever the largest observation in a dataset is at least k times as large as the second largest observation, then the group that contains M_n produces $M_{n(k)}$. So let us focus on the distribution of the ratio of the largest and the second largest value in a dataset.

5.1.1 Distribution of the ratio between order statistics

In Theorem 2.4.1 we derived the distribution of the ratio between two order statistics in the general case, this distribution was given by

$$P\left(\frac{X_{r,n}}{X_{s,n}} \leq x\right) = \frac{1}{B(s, n-s+1)} \int_0^1 I_{Q_x(t)}(r, s-r) t^{s-1}(1-t)^{n-s} dt, \qquad (r < s), \qquad (5.1)$$

where $B(x, y)$ is the beta function, $I_x(r, s)$ the incomplete beta function and $Q_x(t) = \frac{F(tF^{-1}(x)}{t}$. In our case we are interested in the case that $r = n-1$ and $s = n$ so the distribution function in equation (5.1) simplifies to

$$P\left(\frac{X_{n-1,n}}{X_{n,n}} \leq x\right) = n(n-1) \int_0^1 I_{Q_x(t)}(n-1, 1) t^{n-2} dt.$$

But as it turns out there is a much more simple form for the distribution function of the ratio of successive order statistics when we consider order statistics from a Pareto distribution.

Proposition 5.1.1. *When $X_{1,n}, ..., X_{n,n}$ are order statistics from a Pareto distribution then the ratio between two consecutive order statistics, $\frac{X_{i+1,n}}{X_{i,n}}$, also has a Pareto distribution with parameter $(n-i)\alpha$.*

Proof. The distribution function of $\frac{X_{i+1,n}}{X_{i,n}}$ can be found by conditionalizing on $X_{i,n}$ and using Theorem 2.2.1 to find the distribution of $X_{i+1,n}|X_{i,n} = x$.

$$\begin{aligned}
P\left(\frac{X_{i+1,n}}{X_{i,n}} > z\right) &= \int_1^\infty P\left(X_{i+1,n} > zx | X_{i+1,n} = x\right) f_{X_{i,n}}(x) dx \\
&= \int_1^\infty \left(\frac{1-F(zx)}{1-F(x)}\right)^{n-i} \frac{1}{B(i, n-i+1)} F(x)^{i-1} \left(1-F(x)\right)^{n-i} f(x) dx \\
&= \frac{1}{B(i, n-i+1)} \int_1^\infty (1-F(zx))^{n-i} F(x)^{i-1} f(x) dx \\
&= \frac{1}{B(i, n-i+1)} z^{-(n-i)\alpha} \int_1^\infty x^{-(n-i)\alpha} (1-x^{-\alpha})^{i-1} \alpha x^{-\alpha-1} dx \\
&= z^{-(n-i)\alpha} \frac{1}{B(i, n-i+1)} \int_0^1 u^{n-i} (1-u)^{i-1} du \qquad (u = x^{-\alpha}) \\
&= z^{-(n-i)\alpha} \frac{1}{B(i, n-i+1)} B(i, n-i+1) \\
&= z^{-(n-i)\alpha}.
\end{aligned}$$

□

One could wonder if the converse of proposition 5.1.1 also holds, i.e. if for some k and n the following ratio $\frac{X_{k+1,n}}{X_{k,n}}$ has a Pareto distribution, is the parent distribution of the order statistics a Pareto distribution. This is not the case which can be shown through the counter-example of Arnold [1983]. Let Z_1 and Z_2 be two independent $\Gamma(\frac{1}{2}, 1)$ random variables, and let $X = e^{Z_1 - Z_2}$. If one considers a sample of size 2, then we find that X_1 and X_2 are not Pareto distributed, but that the ratio $\frac{X_{2,2}}{X_{1,2}}$ does have a Pareto distribution.

Instead one needs to make additional assumptions like the ratio of two successive order statistics have a Pareto distribution for all n, as was shown in Rossberg [1972]. Here we will give a different proof of this result. The following lemma[1] is needed to proof the result.

[1] Result was found on 1 February 2010 at http://at.yorku.ca/cgi-bin/bbqa?forum=ask_an_analyst_2006; task=show_msg;msg=1091.0001

Lemma 5.1.2. *If $f(x)$ is a continuous function on $[0,1]$, and the following equation holds for all $n \geq 0$*

$$\int_0^1 f(x)x^n dx = 0, \tag{5.2}$$

then $f(x)$ is equal to zero.

Proof. Since equation (5.2) holds, we know that for any polynomial $p(x)$ the following equation holds

$$\int_0^1 f(x)p(x)dx.$$

From this we find that for any polynomial $p(x)$

$$\begin{aligned}\int_0^1 f(x)^2 dx &= \int_0^1 [f(x) - p(x)]\, f(x) + f(x)p(x)dx \\ &= \int_0^1 [f(x) - p(x)]\, f(x)dx.\end{aligned}$$

Since $f(x)$ is a continuous function on $[0,1]$ we find by the Weierstrass theorem that for any $\varepsilon > 0$ there exists a polynomial $P(x)$ such that

$$\sup_{x\in[0,1]} |f(x) - P(x)| < \varepsilon.$$

By the Min-Max theorem we also know that there exists a constant M such that $|f(x)| \leq M$ for all $0 \leq x \leq 1$. From this we find that for any $\varepsilon > 0$ there exists a polynomial $P(x)$ such that

$$\begin{aligned}|\int_0^1 f(x)^2 dx| &= |\int_0^1 [f(x) - P(x)]\, f(x)dx|, \\ &\leq \int_0^1 |f(x) - P(x)||f(x)|dx, \\ &\leq \varepsilon M.\end{aligned} \tag{5.3}$$

But since equation (5.3) holds for all $\varepsilon > 0$ we find that

$$\int_0^1 f(x)^2 dx = 0. \tag{5.4}$$

By assumption $f(x)$ is a continuous function and hence $f(x)^2 \geq 0$ is also a continuous function. The only thing we need to proof now is that $f(x)^2 = 0$ for all $x \in [0,1]$. Now assume that $f(x)^2$ is positive for some $\xi \in [0,1]$, then we know that there exists an $M > 0$ such that $f(c)^2 = M$ and $f(x) \leq M$ for all $x \in [0,1]$. Now let $A = \frac{1}{2}M$, now there exists an interval $[p,q]$ which contains c such that $f(x) > A$ for all $x \in [p,q]$. And since $f(x)^2 \geq 0$ for all x we find the following

$$\begin{aligned}\int_0^1 f(x)^2 &= \int_0^p f(x)^2 dx + \int_p^q f(x)^2 dx + \int_q^1 f(x)^2 dx \\ &\geq \int_p^q f(x)^2 dx \\ &\geq \int_p^q A dx \\ &= A(p-q) > 0.\end{aligned}$$

Which is a contradiction and hence we find that $f(x)^2 = 0$ for all $x \in [0,1]$ and hence $f(x) = 0$ for all $x \in [0,1]$. ☐

Theorem 5.1.3. *If for a positive continuous random variable, for which the distribution function has an inverse, the following equation holds for all $n \geq 2$*

$$P\left(\frac{X_{n,n}}{X_{n-1,n}} > x\right) = x^{-\alpha}, \tag{5.5}$$

for some $\alpha > 0$, then X has the following distribution function

$$F(x) = 1 - \left(\frac{\kappa}{x}\right)^{\alpha},$$

for $x > \kappa$.

Proof. From equation (5.5) we find that

$$n(n-1)\int_0^{\infty} (1 - F(xz))\, F(z)^{n-2} f(z)dz = x^{-\alpha}.$$

From this we find that

$$n(n-1)\int_0^{\infty} x^{\alpha} (1 - F(zx))\, F(z)^{n-2} f(z)dz = 1. \tag{5.6}$$

Since we know that $x > 1$ the following inequality holds $1 - F(xz) < 1 - F(z)$, and since

$$\int_0^{\infty} (1 - F(z))F(z)^{n-2} f(z)dz = 1,$$

since $(1 - F(z))F(z)^{n-2} f(z)$ is the density of the $n-1$-th order statistic from a sample of n. Now differentiate equation (5.6) with respect to x, we are able to interchange integration and differentiation since we can apply the dominated convergence theorem. From this we find that

$$\int_0^{\infty} \left(\alpha x^{-\alpha-1}(1 - F(xz) - x^{-\alpha} z f(zx)\right) F(z)^{n-2} f(z)dz = 0.$$

Using the substitution $u = F(x)$ and dividing by $x^{\alpha-1}$ we find that

$$\int_0^{1} \left(\alpha(1 - F(xF^{-1}(u)) - xF^{-1}(u) f(xF^{-1}(u))\right) u^{n-2} du. \tag{5.7}$$

Since equation (5.3) holds for all $n \geq 2$ we can apply lemma 5.1.2 and find that

$$\alpha(1 - F(xF^{-1}(u)) - xF^{-1}(u) f(xF^{-1}(u)) = 0.$$

We can rewrite this in the following way, by putting $t = xF^{-1}(u)$

$$\alpha\overline{F}(t) + t\frac{d}{dx}\overline{F}(t) = 0.$$

Solving this differential equation gives us

$$\overline{F}(t) = Ct^{-\alpha}$$

for some constant C. We know that $\overline{F}(t) \leq 1$, this means that this solution is only valid for $x > C^{-1/\alpha}$. We can also write the survivor function in the following way

$$\overline{F}(t) = \left(\frac{\kappa}{t}\right)^{\alpha}, \qquad x > \kappa$$

□

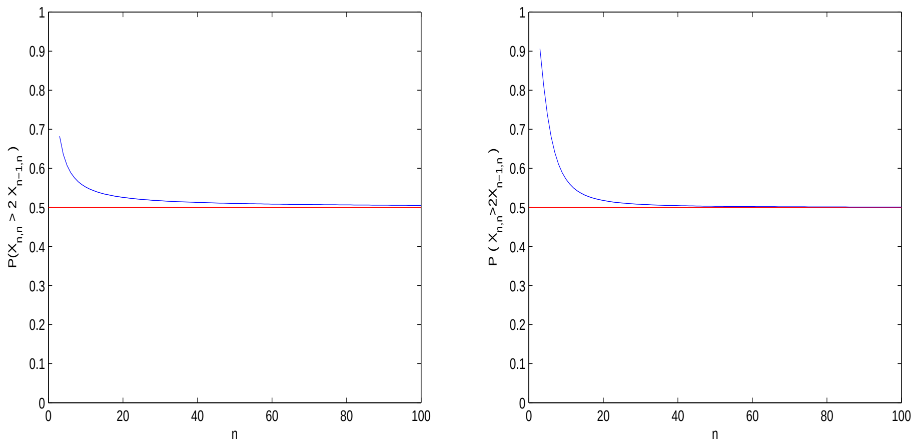
By conditionalizing on the second-largest order statistic we find that the distribution of the ratio between two upper order statistics can be obtained by evaluating the following integral.

$$P\left(\frac{X_{n,n}}{X_{n-1,n}} > z\right) = \int_{-\infty}^{\infty} (1 - F(zx))\, F(x)^{n-2} f(x)dx \tag{5.8}$$

We have found an analytical solution for the integral in equation (5.8) of a Weibull distribution. The distribution function of the ratio between the two upper order statistics from a Weibull distribution is given by

$$\begin{aligned}
P\left(\frac{X_{n,n}}{X_{n-1,n}} > x\right) &= n(n-1)\int_0^{\infty} (1 - F(zx))\, F(x)^{n-2} f(x)dx \\
&= n(n-1)\int_0^{\infty} \left(e^{-(\lambda x z)^{\tau}}\right)\left(1 - e^{-(\lambda z)^{\tau}}\right) \tau\lambda^{\tau} z^{\tau-1} e^{-(\lambda z)^{\tau}} dz \\
&= n(n-1)\int_0^{1} u^{x^{\tau}} (1-u)^{n-2} du \\
&= n(n-1)B(x^{\tau}+1, n-1).
\end{aligned}$$

In Figures 5.3 (a)–(b) we see the approximation of the probability in equation (5.8), with $z = 2$, for the Burr distribution with parameters $c = 1$ and $k = 1$ and the Cauchy distribution. This Burr distribution and the Cauchy distribution both have a tail index one. The probability that the largest order statistic is twice as large as the second largest order statistic seems to converge to a half. Which is exactly the probability that the largest order statistic from a Pareto(1) distribution is twice as large as the second largest order statistic. This indicates that if we consider the distribution of the ratio of the two largest order statistics then the distribution of this ratio converges to the distribution of the ratio between the two largest observation from a Pareto distribution with the same tail index. The following theorems show that the ratio between



(a) Burr Distribution, $c = 1$, $k = 1$ (b) Cauchy Distribution

Figure 5.3: $P(X_{n,n} > 2X_{n-1,n})$ for a few distributions

the upper order statistics have a non-degenerate limit if and only if the parent distribution of the order statistics is regularly varying. Before we present these theorems we repeat some results from the theory of regular variation. If the following limit exists for $z > 1$

$$\lim_{x\to\infty} \frac{\overline{F}(zx)}{\overline{F}(x)} = \beta(z) \in [0,1]\,, \quad z > 1$$

then the three following cases are possible.

1. if $\beta(z) = 0$ for all $z > 1$, then $\overline{F}$ is called a rapidly varying function,

2. if $\beta(z) = z^{-\alpha}$ for all $z > 1$, where $\alpha > 0$, then $\overline{F}$ is called a regularly varying distribution function,

3. if $\beta(z) = 1$ for all $z > 1$, then $\overline{F}$ is called a slowly varying function.

This is exactly the result that can be found in Balakrishnan and Stepanov [2007].

Theorem 5.1.4. *Let F be a distribution function such that $F(x) < 1$ for all x. If $1 - F$ is rapidly varying distribution function and $0 < l \leq k$, then*

$$\frac{X_{n-k+l}}{X_{n-k}} \xrightarrow{P} 1, \qquad (n \to \infty)$$

If $1 - F$ is regularly varying with index $-\alpha$ and $0 < l \leq k$, then

$$P\left(\frac{X_{n-k+l}}{X_{n-k}} > z\right) \to \sum_{i=0}^{l-1} \binom{k}{i} (1 - z^{-\alpha})^i z^{-\alpha(k-i)}, \qquad (z > 1)$$

If $1 - F$ is a slowly varying distribution function and $0 < l \leq k$, then

$$\frac{X_{n-k+l}}{X_{n-k}} \xrightarrow{P} \infty, \qquad (n \to \infty).$$

The converse of Theorem 5.1.4 is also true, as was shown in Smid and Stam [1975].

Theorem 5.1.5. *If for some $j \geq 1$, $z \in (0,1)$ and $\alpha \geq 0$,*

$$\lim_{n\to\infty} P\left(\frac{X_{n-j}}{X_{n-j+1}} < z\right) = z^{j\alpha} \tag{5.9}$$

then

$$\lim_{y\to\infty} \frac{1 - F\left(\frac{y}{z}\right)}{1 - F(y)} = z^{\alpha}$$

From this theorem we get the following corollary

Corollary 5.1.1. *If (5.9) holds for all $z \in (0,1)$, then $1 - F(x)$ is regularly varying of order $-\alpha$ as $x \to \infty$.*

Theorem 5.1.5 was generalized and extended in Bingham and Teugels [1979].

Theorem 5.1.6. *Let $s \in \{0,1,2,...\}$, $r \in \{1,2,...\}$ be fixed integers. Let F be concentrated on the positive half-line. If $\frac{X_{n-r-s,n}}{X_{n-s,n}}$ converges in distribution to a non-degenerate limit, then for some $\rho > 0$, $1 - F(x)$ varies regularly of order $-\rho$ as $x \to \infty$.*

5.2 The ratio as index

In the previous section we have shown that if the ratio between the two largest order statistics converges in distribution to some non-degenerate limit, then the parent distribution is regularly varying. This raises the question can we use this as a measure for heavy-tailedness of a distribution function. In this section we look at the following probability

$$P\left(\frac{X_{n,n}}{X_{n-1,n}} > k\right), \tag{5.10}$$

and how we can estimate this from a dataset. The most obvious way would be by choosing some n and bootstrapping datasets of size n from the dataset and checking whether $X_{n,n} > kX_{n-1,n}$, but this has the drawback that it is not obvious which value n should be. Now consider the following estimator. Suppose we have a dataset of size n, first consider the first two observations. Now observe the next value in the dataset, if the new observation is larger than the previous second largest value take $ntrials = ntrials + 1$ and if the largest value is larger than k times the second largest value in the dataset take $nsucces = nsucces + 1$. Repeat this until we have observed all values. The estimator of $P\left(\frac{X_{n,n}}{X_{n-1,n}} > k\right)$ is defined by $\frac{nsucces}{ntrials}$. We have not proven that this is a consistent estimator, but simulations show that for the Pareto distribution the estimator behaves as expected. Note that the $ntrials$ is the number of observed type 2 2-record values, the probability that a new observation is a type 2 2-record value is equal to

$$P\left(X_n > X_{n-2,n-1}\right) = \int_{-\infty}^{\infty} P\left(X > y\right) f_{n-2,n}(y)dy = \frac{2}{n}$$

This means that if we have a dataset of size n then the expected number of observed Type 2 2-records equals.

$$\sum_{j=3}^{n} \frac{2}{j} = 2\left(\sum_{i=1}^{n} \frac{1}{j} - 1.5\right) \approx 2(\log(n) + \gamma - 1.5).$$

Where γ is the Euler-Mascheroni constant and approximately equal to 0.5772. In figure 5.4 we

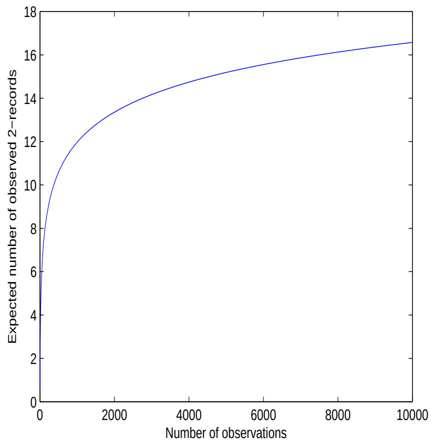
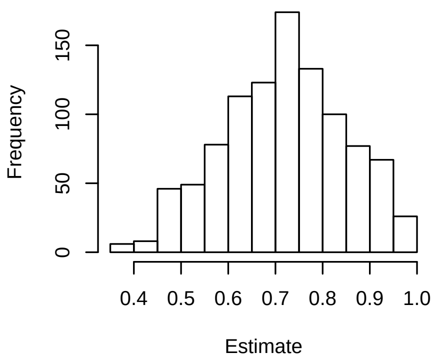
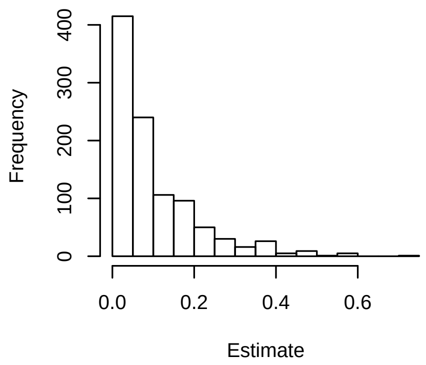


Figure 5.4: Expected number of observed 2-records

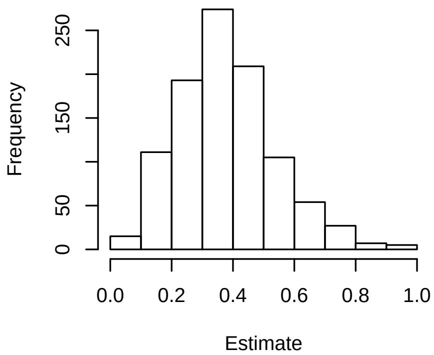
see the expected number of 2-records plotted against the size of the dataset. In a dataset of size 10000 we only expect to see 16 2-records. Since we do not observe a lot of 2-records we do not expect the estimator to be very accurate. We have used the estimate for $P\left(\frac{X_{n,n}}{X_{n-1,n}} > k\right)$



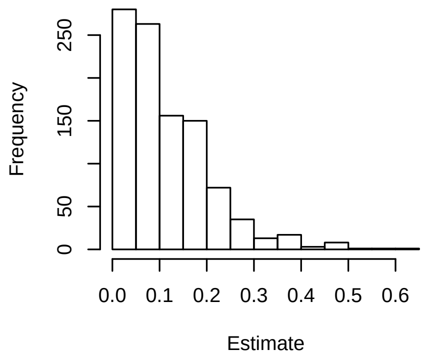
(a) Pareto $\alpha = 0.5$



(b) Pareto $\alpha = 3$



(c) Weibull $\tau = 0.5$



(d) Standard Exponential

Figure 5.5: Histograms of the estimate of $P\left(\frac{X_{n,n}}{X_{n-1,n}} > x\right)$

on a number of simulated datasets. All these simulated dataset were of size 1000. In Figures 5.5 (a)–(d) we see a histogram of the estimator we proposed, we calculated the estimator 1000 times by reordering the dataset and calculating the estimator for the reordered dataset. For the Pareto(0.5) distribution in Figure 5.5a we see that on average the estimator seems to be accurate but that the estimator ranges from as low as 0.4 to as high as 1. For a Weibull distribution with shape parameter $\tau = 0.5$ we see that the estimate of the probability is much larger than zero. This is due to the slow convergence of the probability to zero and the fact that we expect to see more 2-records early in a dataset. In Table 5.1 we have summarized the results of applying the estimators to a Pareto(0.5) distribution, a Pareto(3) distribution, a Weibull distribution with shape parameter $\tau = 0.5$ and a Standard Exponential distribution. We also applied these estimators to the NFIP data, the National Crop Insurance Data and the Hospital data. From Figure 5.6a we see that the estimate of the probability $P\left(\frac{X_{n,n}}{X_{n-1,n}} > 2\right)$ suggest more heavy-tailed behavior than the estimate of the probability of the National Crop Insurance Data and the Hospital data. Which is supported by looking at the mean excess plots of these datasets. We have bootstrapped the dataset by reordering the data in order to calculate more than one

Distribution	Expected Value	Mean Estimate
Pareto $\alpha = 0.5$	0.7071068	0.7234412
Pareto $\alpha = 3$	0.125	0.09511546
Weibull $\tau = 0.5$	0	0.3865759
Exponential	0	0.1104443

Table 5.1: Mean estimate of $P\left(\frac{X_{n,n}}{X_{n-1,n}} > 2\right)$

Dataset	Mean Estimate
NFIP	0.5857
National Crop Insurance	0.2190
Hospital	0.0882

Table 5.2: Mean estimate of $P\left(\frac{X_{n,n}}{X_{n-1,n}} > 2\right)$

realization of the estimator. Again we see that the estimator gives a nice result on average but that the individual values seem to be very spread out.

5.3 The Obesity Index

In the previous section we tried to use the ratio between the largest observation and the second-largest observation as a measure of heavy-tailedness. There were a few drawbacks with this approach, when we tried to estimate the index we saw that the estimate was not very accurate. This was due to the fact that we tried to make an inference based upon the observations from a Type 2 2-record sequence and in this case we do net expect to see a lot of Type 2 2-records. In this section we try and define another index of heavy-tailedness. We tried to construct a measure of heavy-tailedness based upon the fact that if we apply aggregation by k on a dataset, the maximum of the aggregated dataset is usually the sum of the group that contained the maximum value. Now consider aggregation by 2 in a dataset of size 4 containing the observation X_1, X_2, X_3, X_4. Which are independent and identically distributed random variables. Assume that $X_1 < X_2 < X_3 < X_4$. By definition we have that $X_4+X_2 > X_3+X_1$ and $X_4+X_3 > X_2+X_1$, so the only interesting case arises whenever we sum X_4 with X_1. Now define the Obesity index by

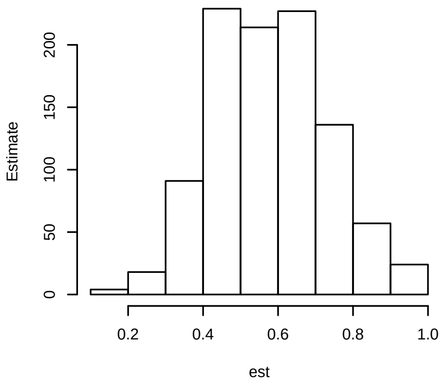
$$\text{Ob}(X) = P\left(X_4 + X_1 > X_2 + X_3 | X_1 \leq X_2 \leq X_3 \leq X_4\right), \qquad X_i \sim X.$$

We expect that for heavy-tailed distribution this probability is larger than for thin-tailed distributions. We can rewrite the inequality in the probability in equation (5.3) in the following way.

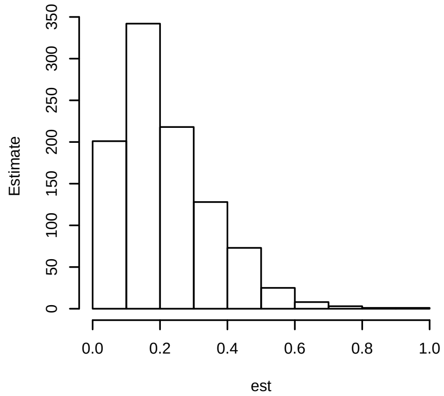
$$X_4 - X_3 > X_2 - X_1,$$

which was one of the heuristics of heavy-tailed distributions we discussed in Chapter 1, i.e. the fact that larger observations lie further apart than smaller observations. Note that the Obesity index is invariant under multiplication by a positive constant and translation, i.e. $\text{Ob}(aX+b) = \text{Ob}(X)$ for $a > 0$ and $b \in \mathbb{R}$. In the following propositions we calculate the Obesity index for a number of distributions. First note that whenever we consider a random variable X, for which $P(X = C) = 1$ where C is a constant, has an Obesity index equal to zero.

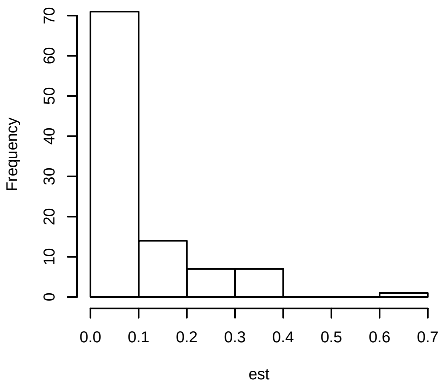
Proposition 5.3.1. *The obesity index of the uniform distribution is equal to $\frac{1}{2}$.*



(a) NFIP



(b) National Crop Insurance



(c) Hospital

Figure 5.6: Histograms of the estimate of $P\left(\frac{X_{n,n}}{X_{n-1,n}} > 2\right)$

Proof. The obesity index can be rewritten in the following way

$$P\left(X_4 - X_3 > X_2 - X_1 | X_1 < X_2 < X_3 < X_4\right) = P\left(X_{4,4} - X_{3,4} > X_{2,4} - X_{1,4}\right). \tag{5.11}$$

Using theorem 2.3.3 we can calculate the probability in equation (5.11). We get the following

$$P\left(X_{4,4} - X_{3,4} > X_{2,4} - X_{1,4}\right) = P\left(X > Y\right), \tag{5.12}$$

where X and Y are standard exponential random variables. Since the random variables X and Y in equation (5.12) are independent and identically distributed random variables this probability is equal to $\frac{1}{2}$. ☐

Proposition 5.3.2. *The obesity index of the exponential distribution is equal to* $\frac{3}{4}$.

Proof. Again we rewrite the obesity index in the following way

$$P\left(X_4 - X_3 > X_2 - X_1 | X_1 < X_2 < X_3 < X_4\right) = P\left(X_{4,4} - X_{3,4} > X_{2,4} - X_{1,4}\right). \tag{5.13}$$

Using theorem 2.3.2 we see that the obesity index of the exponential distribution is equal to

$$P\left(X_{4,4} - X_{3,4} > X_{2,4} - X_{1,4}\right) = P\left(X > \frac{Y}{3}\right), \tag{5.14}$$

where X and Y are independent standard exponential random variables. We can calculate the probability on the RHS in equation (5.14).

$$\begin{aligned} P\left(X > \frac{Y}{3}\right) &= \int_0^\infty P\left(X > y\right) f_{\frac{Y}{3}}(y)dy \\ &= \int_0^\infty e^{-y} 3e^{-3y} dy \\ &= \frac{3}{4}. \end{aligned}$$

□

Proposition 5.3.3. *If X is a symmetrical random variable with respect to zero, $X \stackrel{d}{=} -X$, then the obesity index is equal to $\frac{1}{2}$.*

Proof. If $X \stackrel{d}{=} -X$, then we have that $F_X(x) = 1 - F_X(-x)$, and $f_X(x) = f_X(-x)$. The joint density of $X_{3,4}$ and $X_{4,4}$ is now given by

$$\begin{aligned} f_{3,4;4}(x_3, x_4) &= \frac{24}{2} F(x_3)^2 f(x_3) f(x_4), \quad x_3 < x_4 \\ &= \frac{24}{2}(1 - F(-x_3))^2 f(-x_3) f(-x_4), \quad -x_4 < -x_3 \\ &= f_{1,2;4}(-x_4, -x_3) \end{aligned}$$

Which is equal to the joint density of $-X_{1,4}$ and $-X_{2,4}$, and from this we find that

$$X_{4,4} - X_{3,4} \stackrel{d}{=} X_{2,4} - X_{1,4}.$$

And hence the obesity index is equal to $\frac{1}{2}$. □

From proposition 5.3.3 we find that for a random variable with a symmetrical distribution with respect to some constant μ the Obesity index is equal to zero. This is the case because if X is symmetric with respect to μ, $X - \mu$ is symmetric with respect to zero. And we have that $\text{Ob}(X) = \text{Ob}(X - \mu)$. This means that the Obesity index of both the Cauchy and the Normal distribution have an Obesity index that is equal to $\frac{1}{2}$. But the Cauchy distribution has a regularly varying distribution function with tail index 1, and the Normal distribution is considered to be a thin-tailed distribution. This limits the use of the Obesity index to positive random variables, which is usually the case in applications like insurance and magnitudes of natural disasters.

Theorem 5.3.4. *The obesity index of a random variable X with distribution function F and density f can be calculated by evaluating the following integral,*

$$24 \int_{-\infty}^\infty \int_{x_1}^\infty \int_{x_2}^\infty \overline{F}(x_2 + x_3 - x_1) f(x_1) f(x_2) f(x_3) dx_3 dx_2 dx_1.$$

Proof. The obesity index can be rewritten in the following way

$$P\left(X_1 + X_4 > X_2 + X_3 | X_1 \leq X_2 \leq X_3 \leq X_4\right).$$

Recall that the joint density of all n order statistics out of a sample of size n is given by

$$f_{1,2,...,n;n}(x_1, x_2, ..., x_n) = \begin{cases} n! \prod_{i=1}^{n} f(x_i), & x_1 < x_2 < ... < x_n, \\ 0, & \text{otherwise,} \end{cases}$$

In order to calculate the obesity index we need to integrate this joint density over all numbers such that

$$x_1 + x_4 > x_2 + x_3, \text{ and } x_1 < x_2 < x_3 < x_4.$$

To calculate the obesity index we then need to evaluate the following integral.

$$\text{Ob}(X) = 24 \int_{-\infty}^{\infty} f(x_1) \int_{x_1}^{\infty} f(x_2) \int_{x_2}^{\infty} f(x_3) \int_{x_3+x_2-x_1}^{\infty} f(x_4) dx_4 dx_3 dx_2 dx_1$$

Now the innermost integral is the probability that the random variable X is larger than $x_3 + x_2 - x_1$ so we can simplify this expression to

$$\text{Ob}(X) = 24 \int_{-\infty}^{\infty} \int_{x_1}^{\infty} \int_{x_2}^{\infty} \overline{F}(x_2 + x_3 - x_1) f(x_1) f(x_2) f(x_3) dx_3 dx_2 dx_1.$$

□

Using Theorem 5.3.4 we are able to calculate the Obesity index whenever the parameter α is an integer. We have done this using Maple, in Table 5.3 the exact and approximate value of the Obesity index for a number of α are given. From Table 5.3 we can observe that the Obesity

α	Exact value	Approximate value
1	$\pi^2 - 9$	0.8696
2	$593 - 60\pi^2$	0.8237
3	$\frac{-124353}{5} + 2520\pi^2$	0.8031
4	$\frac{19150997}{21} - 92400\pi^2$	0.7912

Table 5.3: Obesity index of Pareto(α) distribution when α is an integer

index seems to increase as the tail index decreases. Which was exactly what we expected. Now we are going to discuss some properties for the Obesity index of a Pareto random variable. These properties are derived using the theory of majorization. We give a short review of the theory of majorization before we derive the properties of the Obesity index.

5.3.1 Theory of Majorization

The theory of majorization is used to give a mathematical meaning to the notion that the components of one vector are less spread out than the components of another vector.

Definition 5.3.1. *A vector* $\mathbf{y} \in \mathbb{R}^n$ *majorizes a vector* $\mathbf{x} \in \mathbb{R}^n$ *if*

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i,$$

and

$$\sum_{i=1}^{k} x_{[i]} \leq \sum_{i=1}^{k} y_{[i]}, \qquad k = 1, ..., n$$

where $x_{[i]}$ are the ordered elements of the vector $\mathbf{x}$ such that

$$x_{[1]} \geq ... \geq x_{[n]}.$$

We denote this by $\mathbf{x} \prec \mathbf{y}$.

There also exist functions which preserve the ordering we defined above. These functions are called Schur-convex functions.

Definition 5.3.2. *A function $\phi : \mathcal{A} \to \mathbb{R}$, where $\mathcal{A} \subset \mathbb{R}^n$, is called Schur-convex on $\mathcal{A}$ if*

$$\mathbf{x} \prec \mathbf{y} \text{ on } \mathcal{A} \Rightarrow \phi(x) \leq \phi(y)$$

The following proposition gives sufficient conditions for a function ϕ to be Schur-convex.

Proposition 5.3.5. *If $I \subset \mathbb{R}$ is an interval and $g : I \to \mathbb{R}$ is convex, then*

$$\phi(x) = \sum_{i=1}^{n} g(x_i),$$

is Schur-convex on I^n.

We prove two theorems about the inequality in the obesity index.

Theorem 5.3.6. *If $0 < x_1 < x_2 < x_3 < x_4$, then as $\alpha \to 0$ the following inequality holds*

$$\lim_{\alpha \to 0} x_1^{-1/\alpha} + x_4^{-1/\alpha} > \lim_{\alpha \to 0} x_2^{-1/\alpha} + x_3^{-1/\alpha}$$

Proof. From Hardy et al. [1934] we know that

$$\lim_{p \to \infty} \left(\sum_{i=1}^{n} x_i^p \right)^{\frac{1}{p}} = \max \{x_1, ..., x_n\}.$$

From this we get that

$$\lim_{\alpha \to 0} \frac{x_1^{-1/\alpha} + x_4^{-1/\alpha}}{\max \left\{ x_1^{-1}, x_4^{-1} \right\}} = 1.$$

So this means that as α tends to 0, the terms $x_1^{-1/\alpha} + x_4^{-1/\alpha}$ tends to $\max \left\{ x_1^{-1/\alpha}, x_4^{-1/\alpha} \right\}$, which by definition is equal to $x_1^{-1/\alpha}$. The same limit holds for $x_2^{-1/\alpha} + x_4^{-1/\alpha}$, where the maximum of these two by definition is equal to $x_2^{-1/\alpha}$. And by definition we have that $x_1^{-1/\alpha} > x_2^{-1/\alpha}$. □

Lemma 5.3.7. *If $1 < y_1 < y_2 < y_3$ and*

$$y_3 + 1 > y_2 + y_1$$

then for all $q > 1$

$$y_3^q + 1 > y_2^q + y_1^q$$

Proof. Note that $(y_1, y_2) \prec (y_2 + y_1 - 1, 1)$ and that the function $g : \mathbb{R} \to \mathbb{R}$ defined by

$$g(x) = x^q, \qquad q > 1,$$

is convex. Then the function

$$\phi(x_1, x_2) = \sum_{i=1}^{2} g(x_i),$$

is Schur-convex on $\mathbb{R}^2$ by Proposition 5.3.5. From this we find that

$$\begin{aligned}\phi(x_1, x_2) &= y_1^q + y_2^q \\ &\leq (y_2 + y_1 - 1)^q + 1 \\ &\leq y_3^q + 1.\end{aligned}$$

□

Theorem 5.3.8. *For all* $0 < x_1 < x_2 < x_3 < x_4$ *and* $\alpha > \beta > 0$, *if the following inequality holds*

$$x_4^{-1/\alpha} + x_1^{-1/\alpha} > x_2^{-1/\alpha} + x_3^{-1/\alpha},$$

then

$$x_4^{-1/\beta} + x_1^{-1/\beta} > x_2^{-1/\beta} + x_3^{-1/\beta}.$$

Proof. We know that the following inequality holds

$$x_4^{-1/\alpha} + x_1^{-1/\alpha} > x_2^{-1/\alpha} + x_3^{-1/\alpha}.$$

From this we get

$$1 + \left(\frac{x_4}{x_1}\right)^{\frac{1}{\alpha}} > \left(\frac{x_4}{x_2}\right)^{\frac{1}{\alpha}} + \left(\frac{x_4}{x_3}\right)^{\frac{1}{\alpha}}.$$

Now apply Lemma 5.3.7 with $q = \frac{\alpha}{\beta} > 1$. □

When X has a Pareto(α) distribution, then $X^{\alpha/\beta}$ has a Pareto(β) distribution. This means that if $\beta < \alpha$, then $\mathrm{Ob}(X) \leq \mathrm{Ob}\left(X^{\alpha/\beta}\right)$. And from Theorem 5.3.6 we know that

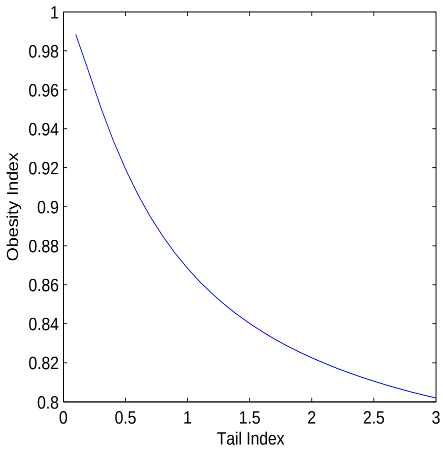
$$\lim_{\alpha \to 0} \mathrm{Ob}(X) = 1.$$

In fact we know that if we have two positive random variable X and X^a, with $a > 1$, then $\mathrm{Ob}(X) \leq \mathrm{Ob}(X^a)$. And $\lim_{a\to\infty} X^a = 1$. Using Theorem 5.3.4 we have approximated the obesity index of the Pareto distribution, the Weibull distribution, the Log-normal distribution, the Generalized Pareto distribution and the Generalized Extreme Value distribution. The generalized extreme value distribution is defined by

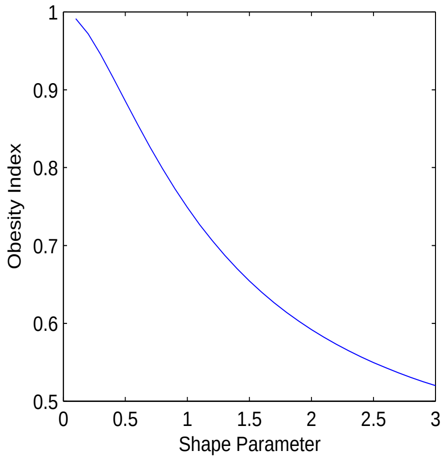
$$F(x; \mu, \sigma, \xi) = \exp\left\{ -\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi} \right\},$$

for $1 + \xi(x - \mu)/\sigma > 0$, the location parameter $\mu \in \mathbb{R}$, the scale parameter $\sigma > 0$ and shape parameter $\xi \in \mathbb{R}$. Where the case $\xi = 0$ the generalized extreme value distribution corresponds to the Gumbel distribution. The Generalized Pareto distribution is defined by

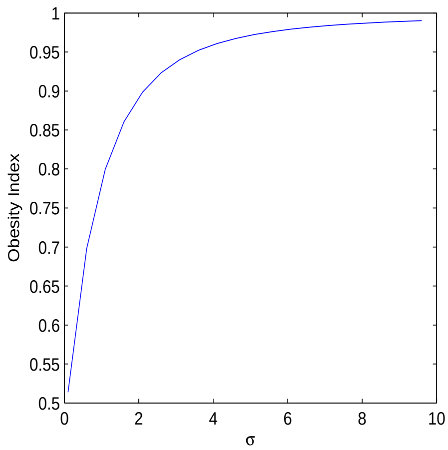
$$F(x; \mu, \sigma, \xi) = \begin{cases} 1 - \left(1 + \frac{\xi(x-\mu)}{\sigma}\right)^{-1/\xi} & \text{for } \xi \neq 0, \\ 1 - \exp\left\{-\frac{x-\mu}{\sigma}\right\} & \text{for } \xi = 0, \end{cases}$$



(a) Pareto Distribution



(b) Weibull Distribution



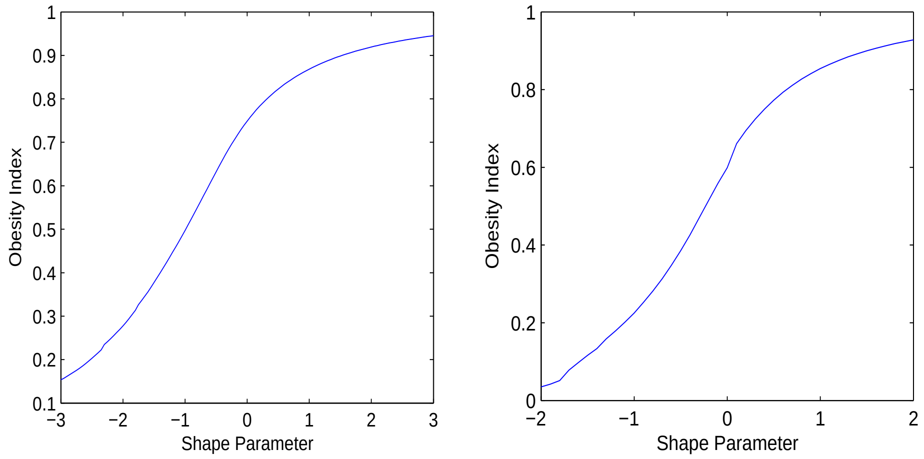
(c) Log-Normal distribution

Figure 5.7: Obesity index for different distributions.

for $x \geq \mu$ when $\xi \geq 0$, and $x \leq \mu - \frac{\sigma}{\xi}$ when $\xi < 0$, the location parameter $\mu \in \mathbb{R}$, the scale parameter $\sigma > 0$ and the shape parameter $\xi \in \mathbb{R}$. Whenever $\xi > 0$ the generalized extreme value distribution and the Generalized Pareto distribution are regularly varying distribution functions with tail index $\frac{1}{\xi}$. As we can see in Figures 5.7 and 5.8 the Obesity index of all the distributions we consider here behave nicely. This is due to the fact that if we consider a random variable X that has anyone of these distributions and consider X^a, then X^a also has the same distribution but with different parameters. In these figures we have plotted the Obesity index against the parameter that changes when considering X^a and that cannot be changed through adding a constant to X^a or by multiplying X^a with a constant. One can wonder if the Obesity index of a regularly varying distribution increases as the tail index of this distribution decreases. The following numerical approximation does indicate that the this is not the case in general.

If X has a Pareto distribution with parameter k, then the following random variable has a Burr distribution with parameters c and k

$$Y \stackrel{d}{=} (X-1)^{\frac{1}{c}}.$$



(a) Generalized Pareto distribution (b) Generalized Extreme Value distribution

Figure 5.8: Obesity index for different distributions.

This holds since when X has a Pareto(k) distribution then

$$P\left((X-1)^{\frac{1}{c}} > x\right) = P\left(X > x^c + 1\right) = \left(x^c+1\right)^{-k}.$$

From table 4.1 we know that the tail index of the Burr distribution is equal to ck. This means that if we consider a Burr distributed random variable with parameters k and $c = 1$, then the Obesity index of this Burr distributed random variable equals the Obesity index of a Pareto random variable with parameter k. From this we find that the Obesity index of a random variable X_1 with a Burr distribution with parameters $c = 1, k = 2$ is equal to $593 - 60\pi^2 \approx 0.8237$. If we now consider a random variable X_2 with a Burr distribution with parameters $c = 3.9$ and $k = 0.5$ and we approximate the Obesity index numerically we find that the Obesity index of this random variable is approximately equal to 0.7463, which is confirmed by simulations. So although the tail index of X_1 is larger than the tail index of X_2, we have that $\mathrm{Ob}(X_1) > \mathrm{Ob}(X_2)$. In Figure 5.9 the Obesity index of the Burr distribution is plotted for different values of c and k. these

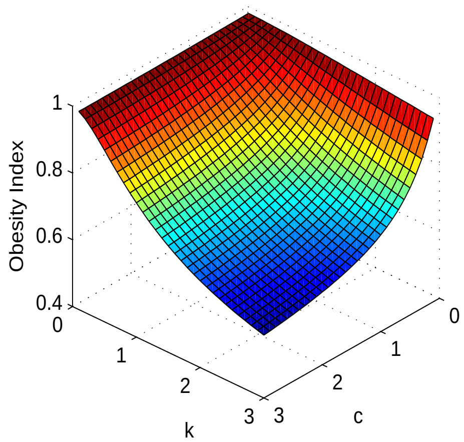


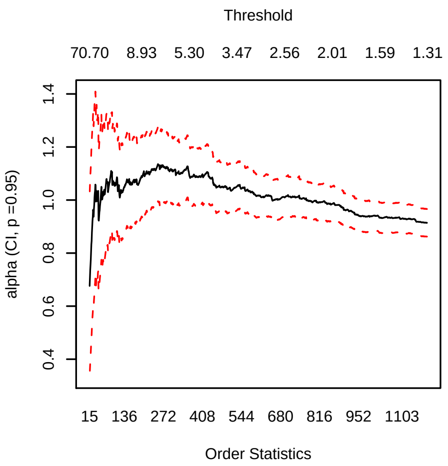
Figure 5.9: The Obesity index of the Burr distribution

Dataset	Obesity Index	Confidence Interval
Hospital Data	0.8	(0.7928,0.8072)
NFIP	0.876	(0.8700,0.8820)
National Crop Insurance	0.808	(0.8009,0.8151)

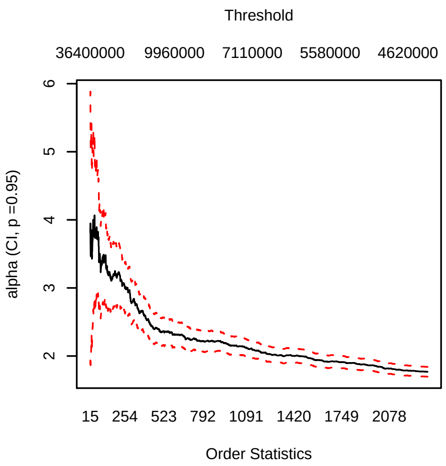
Table 5.4: Estimate of the Obesity index

5.3.2 The Obesity Index of a few Datasets

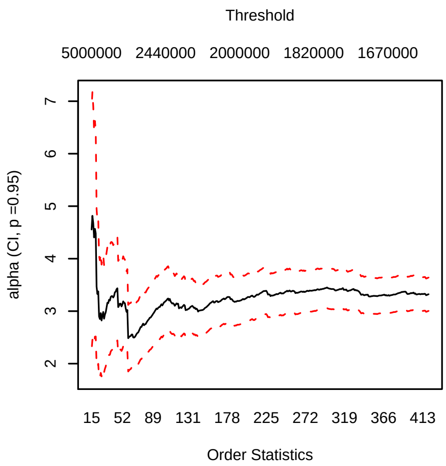
In this section we estimate the Obesity index of a number of datasets, and check whether the Obesity index and the estimate of the tail index are different. In Table 5.4 we see the estimate of the Obesity index based upon 250 bootstrapped values, and the 95%-confidence bounds of the estimate. The estimate was based upon 250 samples from the dataset. From Table 5.4 we get that the NFIP dataset is more heavy-tailed than the National Crop Insurance data and the Hospital data. These conclusions are supported when looking at the mean excess plots of these datasets, this could also be concluded from the Hill estimates of this datasets. In figure 5.10 we see the Hill estimates based upon the top 20% observations of each dataset. Note that the Hill plots in Figures 5.10a and 5.10c are quite stable, but that the Hill plot of the National Crop Insurance Data in Figure 5.10b is not.



(a) NFIP



(b) National Crop Insurance



(c) Hospital

Figure 5.10: Hill estimator of a number of datasets

The final dataset we consider is the G-econ database from Nordhaus et al. [2006]. This dataset consists of environmental and economical characteristics of cells of 1 degree latitude and 1 degree longitude of the earth. One of the entries is the average precipitation. If we look at the mean excess plot of this dataset, it is not clear whether we are dealing with a heavy-tailed distribution or not. In Figure 5.11 we see that the mean excess plot first decreases and after that shows increasing behavior. If we estimate the obesity index of this dataset we get an estimate of 0.728 with 95%-confidence bounds equal to $(0.6728, 0.7832)$. This estimate suggest that we are dealing with thin-tailed distribution. This conclusion is supported if we look at the exponential QQ-plot of the dataset which shows that the data follows a exponential distribution almost perfectly.

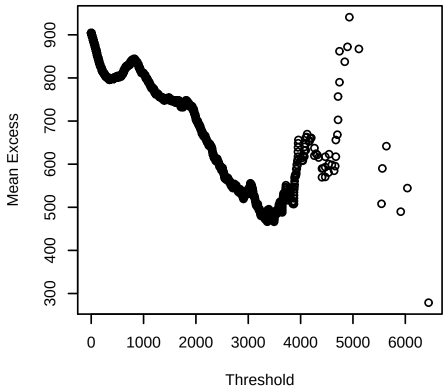


Figure 5.11: Mean excess plot average precipitation.

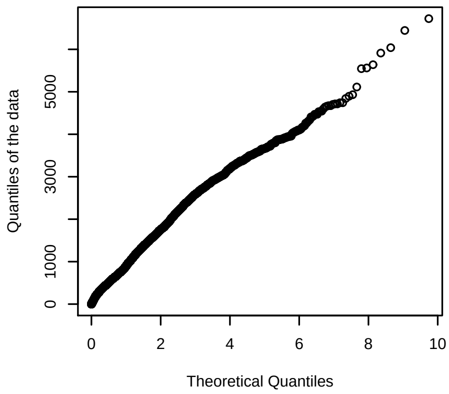


Figure 5.12: Exponential QQ-plot for the average precipitation.

Chapter 6

Conclusions and future research

In this thesis we have looked at two different candidates to characterize the heavy-tailedness of a dataset based upon the behavior of the mean excess plot under aggregations by k. We first looked at the ratio between the largest observation and the second largest observation. As it turned out this ratio has a non-degenerate limit if and only if the distribution is regularly varying. We tried to construct an estimator that would estimate this limiting probability, the problem with this estimator was that it was very inaccurate, since we based it on the observed Type 2 2-records in a dataset. And one does not observe a lot of k-records when considering a large dataset. For example the expected number of Type 2 2-records in a dataset of size 10000 is approximately equal to 16.58. The estimator was also biased when considering thin-tailed distributions, since the limiting probability was equal to zero but the estimate was usually much larger than zero. This was due to the fact that most 2-records will be observed early in the dataset, and then the probability that the largest observation is twice as large as the second largest observation is still quite large. This is why we looked for another characterization of heavy-tailedness of a dataset. We then defined the Obesity index of a random variable in the following way

$$\text{Ob}(X) = P\left(X_1 + X_4 > X_2 + X_3 | X_1 < X_2 < X_3 < X_4\right), \qquad X_i \sim X.$$

We saw that for a lot of regularly varying distribution functions if the tail index got smaller the Obesity index got larger. Unfortunately this was not the case in general. We found a counter-example for which the tail index of a regularly varying distribution function got smaller and the Obesity index also got smaller.

Consider the Burr distribution which has parameters c and k, and has a tail index ck. If the parameter c is equal to one the Obesity index of the Burr distribution will be equal to the Obesity index from the Pareto distribution with a parameter k. Using numerical approximation we have shown that the Obesity index of the first Burr distribution is likely to be larger than the Obesity index of the second Burr distribution.

When we estimated the Obesity index for a number of datasets we saw that the Obesity index and the Hill estimator both gave the same results on the heavy-tailedness of the dataset. Although the Hill estimator of the National Crop Insurance did not gave a clear estimate of the tail index of the distribution. We also saw that when the mean excess plot of a distribution does not give a clear indication whether a dataset is heavy-tailed or not the Obesity index could be used to get an indication of the heavy-tailedness of the distribution.

The Obesity index seems to be an useful index when characterizing the heavy-tailedness of a distribution. In the examples we considered we saw that the Obesity index and the Hill estimator both indicated the same amount of heavy-tailedness in a distribution. A drawback of the Obestiy index seems to be that for the Burr distribution the Obesity index does not indicate a heavy-tailed distribution although the tail index indicates that it is. Future research should investigate under which conditions the Obesity index increases as the tail index of a regularly varying distribution decreases.

Bibliography

D. Applebaum. Levy-type stochastic integrals with regurlary varying tails. *Stochastic Analysis and Applications*, 2005.

B. Arnold. *Pareto Distributions.* Fairland International Co-operative Publishing, 1983.

B. Arnold, N. Balakrishnan, and H. Nagaraja. *Records.* John Wiley Sons, 1998.

N. Balakrishnan and A. Stepanov. Asymptotic properties of the ratio of order statistics. *Statistics and Probability letters*, 78:301–310, 2007.

J. Beirlant, G. Dierckx, and A. Guillou. Estimation of the extreme-value index and generalized quantle plots. *Bernoulli*, 11:949–970, 2005.

N. Bingham and J. Teugels. Conditions implying domains of attraction. In *Proceedings of the Sixth conference on Probability Theory*, pages 23–24, 1979.

N. Bingham, C. Goldie, and J. Teugels. *Regular Variation.* Cambridge University Press, 1987.

S. Coles. *An Introduction to Statistical Modelling of Extreme Values.* Springer-Verlag, 2001.

M. Crovella and A. Bestavros. Self-similarity in world wide web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking*, 1997.

M. Crovella and M. Taqqu. Estimating the heavy tail index from scaling properties. *Methodology and Computing in Applied Probability*, 1:55–79, 1999.

H. A. David. *Order Statistics.* John Wiley Sons, 2nd edition edition, 1981.

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremeal Events for Insurance and Finance.* Springer, 1997.

W. Feller. *An Introduction to Probability Theory and Its Applications*, volume II. John Wiley Sons, 2nd edition edition, 1971.

G. Hardy, J. Littlewood, and G. Pólya. *Inequalities.* cambridge University Press, 1934.

B. Hill. A simple general approach to inference about the tail of a distribution. *Annals Of Statistics*, 3, 1975.

J. Latchman, F. Morgan, and W. Aspinall. Temporal changes in the cumulative piecewise gradient of a variant of the gutenbergrichter relationship, and the imminence of extreme events. *Earth-Science Reviews*, 87:94–112, 2008.

J. Leslie. On the non-closure under convolution of the subexponential family. *Journal of Applied Probability*, 1989.

B. Malamud and D. Turcotte. The applicability of power-law frequency statistics to floods. *Journal of Hydrology*, 322:168–180, 2006.

B. Mandelbrot. The variation of certain speculative prices. *The Journal of Business*, 36:394–419, 1963.

V. B. Nezvorov. *Records: Mathematical Theory*. American Mathematical Society, 2001.

W. Nordhaus, Q. Azam, D. Corderi, K. Hood, N. M. Victor, M. Mohammed, A. Miltner, and J. Weiss. The g-econ database on gridded output: Methods and data. *Yale university*, 2006.

S. Rachev. *Handbook of heavy-tailed distributions in finance*. Elsevier, 2003.

H. Rossberg. Characterization of the exponential and the pareto distributions by means of some properties of the distributions which the differences and quatients of order statistics are subject to. *Statistics*, 3:207–216, 1972.

B. Smid and A. Stam. Convergence in distribution of quotients of order statistics. *Stochastic Processes and their Applications*, 1975.

V. Uchaikin and V. Zolotarev. *Chance and stability: stable distributions and their applications*. VSP International Science Publishers, 1999.