Technische Universiteit Delft
Faculteit Elektrotechniek, Wiskunde en Informatica
Delft Institute of Applied Mathematics

Statistics of War Casualties

Verslag ten behoeve van het
Delft Institute of Applied Mathematics
als onderdeel ter verkrijging

van de graad van

BACHELOR OF SCIENCE
in
TECHNISCHE WISKUNDE

door

Peter Mokhov

Delft, Nederland

BSc verslag TECHNISCHE WISKUNDE

"Statistics of War Casualties"

Peter Mokhov

Technische Universiteit Delft

Begeleider

Dr. P. Cirillo

Overige commissieleden

Dr. D. Kurowicka

Dr. B. van den Dries

July, 2019

Delft

Abstract

In this thesis we statistically analyze violent conflicts. The main focus lies on the risk of occurrence of large wars. We collected data that provides the total amount of casualties, for every known war, in the time span 768CE - 2019. The distribution of the data suggests a presence of a long right tail. We have used different graphical tools to determine that the tail is Paretian and to locate the threshold value u for which the tail starts. Fitting the Generalized Pareto Distribution we have found that this tail starts from the 70% quantile which corresponds to a total number of casualties of 70000. We have found through the method of maximum likelihood a shape parameter of $\xi = 1.3$ and a scale parameter $\beta = 193397$. The threshold and these estimates provide us enough material to determine the tail risk using the survival function.
We have researched the inter-arrival times and we support the idea from earlier studies that the occurrence of wars follow a homogeneous Poisson process and that therefore no particular trend can be stated.

Contents

1. Introduction

In human history World War II has proven to be a significant event that seemingly changed the frequency of large scale wars after this war. The great global powers of our world did not wage war against each other after World War II which leads historians and scientists to believe that this period of stability is an unprecedented occurrence in history. This peace is also called the "Long Peace" referring to the period starting from the end of World War II until now. This long peace hypothesis is actively debated, whether consciously or not, among intellectuals/academics and among common people. People who debate the occurrence of this "extraordinary" armistice are usually split into two groups. One group is optimistic and claims that the world has learned from its barbaric ways in the past and now is in a period of understanding that we as human kind should respect and accept each others differences. The other group on the other hand counters this with data and statistics of war casualties. This includes quite a few studies such as the ones from Clauset [3] and Cirillo and Taleb [2]. They have used data of the number of casualties of wars and conclude that there is no clear pattern towards a "Long Peace".

Technology in war is also an interesting topic of discussion and could support both camps of the debate. For instance the development of nuclear weapons in modern warfare may favor the optimist side because, as all military historians seem to agree on, the weapons scare off the global superpowers from attacking each other and therefore they are forced in a state of armistice. On the other hand this lack of direct conflict has resulted in so called proxy wars all over the world and have led to further ongoing instability in these regions e.g. Middle East, Africa. Thus, it can be said that nuclear weapons trigger both peace and conflict in some way.

Since we will be working with data, it would be interesting to ask ourselves what we *a priori* can expect from it. Generally what we know is that World War II has been the deadliest conflict in human history and therefore if we would look at the total number of casualties it should correspond to the maximum number in the data. The earlier mentioned long peace would imply that we should not expect large scale wars in the data after the occurrence of World War II. Before this would there be any other war of such a scale that reaches the numbers of World War II? Perhaps World War I would be a close guess that we know also has a great number of casualties. What we do know for sure, or at least what we can agree on early in this paper, is that there are a few extremes in terms of the amount of total casualties in war. Intuitively it is clear that there are way more smaller scale conflicts than there are large interstate conflicts. In other words, the frequency of smaller wars is a lot higher than the frequency of these large events approximating the size of World War II. In this thesis we are interested and will research these large events and the risk of these events happening. The aim is to research the tail risk of violent conflicts and we aim to find a threshold value from which we can speak of "large" wars and thus were the tail of violent conflicts will start. We will apply methods from extreme value theory to determine the tail properties. To further motivate the reader the methods used in this thesis can be applied in other fields (finance, physics, biology etc.) because it turns out that the

tail distribution of war casualties follows a power-law distribution which is known to be found in many fields of research.
Next, the thesis will move on to the inter-arrival times of wars. This chapter of the thesis should answer the question: Is there a particular trend between the tail events? Earlier research has shown that there is no trend to be found towards a long term peace. We will check whether we could make the same statement given our data or perhaps it could be that there is a certain trend.

To give the reader a short overview the report will be structured as follows: firstly, the methods/tools used in this paper will be explained in the Methodology chapter. Secondly, the data on which we will apply the methods will be described and the way it is collected in the Data chapter. Afterwards we will provide the results of the methods which will be shown in the Results chapter. The thesis will end with a Conclusions chapter where we will discuss the results we have obtained and if they answer the questions we have raised in this chapter.

2. Methodology

This section is devoted to explaining the methods that are used to obtain the results. The first thing to do when we start working with clean data is to provide an overview of standard statistics to have a rough idea of the behavior of the data. Some lesser known statistics will be thoroughly explained and the results will contain well known plots such as histogram and box plots.

Next we will be looking at the tail of the data. In other words considering wars that have a high number of casualties. We will be investigating what kind of tail we are dealing with and search for a certain threshold above which the tail of the distribution starts. We will make use of the Pareto distribution and the method of maximum likelihood and it will therefore be explained thoroughly.

Lastly we will research the inter-arrival time between wars in our data. We will provide a theorem from probability that we are going to use for checking the Poisson process hypothesis from earlier research [3, 2].

2.1 Descriptive Statistics

To get an idea of how the data looks like it is always a good idea to first compute some straight forward statistics such as the mean, standard deviation, min/ max etc. to summarize the data. Some statistics that are less known or may need to be refreshed are explained below:

Q1 is defined as that value for which 25% of the data is below that value and 75% is greater than that value. Now the median which is actually Q2 is that value for which 50% of the data is below that value. Q3 is that value for which 75% of the data is below that value and 25% procent of the data is above that value. Lastly, the maximum value is simply the maximum value of the data.

The standard deviation, denoted by sd in the table, is a measure that is used to quantify the amount of variation or dispersion [5]. In other words it will tell how close the data will be to its mean value. A low sd means that the data tends to be close to the mean and a high sd means that the data is more spread out.

Another measure of dispersion is the skewness. It measures the asymmetry of the distribution. A negative skew means that the tail of the distribution on the left side of the mean is greater than on the right. A distribution with a zero skewness means that the distribution is symmetric (e.g. Normal). Now a positive skewness means that the distribution has a greater right tail then a left tail.

Last that is used in the descriptive statistics is the kurtosis. The kurtosis is a measure of the tailedness of the distribution [5]. In other words it measures whether the distribution has a lot of extremes or not. It is usually compared with the kurtosis of the normal distribution which has the value 3.

Next some standard plots will be produced. To start with a histogram will be made to visualize the distribution of the data. A histogram accurately depicts the empirical distribution of the data and gives a rough idea what the underlying probability density function looks like.

The other graph that is made is the boxplot. A boxplot graphically shows the quartiles, Q1, Q2, Q3 described earlier. In the Results chapter we will provide two boxplots. One will be made for the complete data and the other one will be made without the extremes. The extremes are defined as points that differ significantly from other observations but are not reporting errors. We will later be looking at the extremes separately because this thesis will be focusing on the tail of the data i.e. large wars.

2.2 Tail in Data

How would we know that the data we have contains a heavy tail? In this section, we will explain some graphical methods to investigate what kind of tail we may have in the data. The most general class that we have when we talk about extreme events are the heavy tails. However in practice they are not easy to work with. The Fat-tailed distributions on the other hand form the most interesting class for modeling extreme and rare events. They are asymptotically scale invariant and their tails mimic that of a Pareto distribution [7]. This leads us to some graphical tools to check whether our data has a Fat-tail. These graphical tools will be explained in the following paragraphs.

Mean Excess Function
The first method described for detecting the presence of a heavy tail is by using the *mean excess function plot* [2]. It is defined as follows:

Let X be a random variable with distribution F and right endpoint $x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}$. The function

$$e(u) = \mathbb{E}[X - u | X > u] = \frac{\int_u^\infty (t-u) dF(t)}{\int_u^\infty dF(t)}, \quad 0 < u < x_F, \tag{2.1}$$

is called mean excess function of X.

In practice, when data is provided, the empirical MEF of a sample $X_1, X_2, \ldots, X_n$ is computed as

$$e_n(u) = \frac{\sum_{i=1}^n (X_i - u)}{\sum_{i=1}^n \mathbf{1}_{\{X_i > u\}}}, \tag{2.2}$$

with u being the threshold value, which are values in $\{X_i : i = 1, \ldots, n\}$. To generate the plot we will be using the R function meplot from the evir package. Now if we spot an upward trend in the ME plot it would mean that the data shows heavy-tailed behaviour [8].

QQ-plot
The QQ-plot is another well known graphical method that can not be missed in this thesis. It is short for Quantile-Quantile plot which means it compares the quantiles of the empirical distribution with that of a theoretical distribution that would be interesting to compare it with. The empirical distribution is defined as follows. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables. The empirical distribution F_n will then be

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \leq x}. \tag{2.3}$$

While the theoretic distribution of a random variable X is the function

$$F_X(x) = \mathbb{P}(X \leq x). \tag{2.4}$$

The theoretic distribution that we will choose to detect heavy-tails is an exponential distribution because it is *the* distribution with thin tails and thus an appropriate distribution to compare the tails with. Now, according to [8] if we observe concativity in the QQ-plot it would be a sign of heavy tailed behavior.

Zipf Plot
There is also another and quite common way of verifying the presence of Paretianity in the tail. It is based on the survival function which is a function that provides the probability that a object of interest will survive beyond any specified time. In other words it is the complement of the cumulative distribution function 2.4

$$\overline{F}(x) = 1 - F(x) = \mathbb{P}(X > x). \tag{2.5}$$

Considering the Pareto I distribution 2.14 we will get the survival function

$$\overline{F}(x) = \left(\frac{x}{x_0}\right)^{-\alpha}, \quad 0 < x_0 \leq x. \tag{2.6}$$

If we now take the logs on both sides we will get

$$\log(\overline{F}(x)) = \alpha \log(x_0) - \alpha \log(x). \tag{2.7}$$

Setting $C = \alpha \log(x_0)$ we get $\log(\overline{F}(x)) = C - \alpha \log(x)$ thus a negative linear relationship between the logarithm of the survival function and the logarithm of x. When producing a Zipf plot for the data we do not need to expect a negative linear relationship in the whole plot. For example if there is a Paretian tail in the data the plot will show a negative linear relationship after some threshold value x. A Zipf plot is therefore also useful to heuristically identify this threshold value [6].

2.3 Moments

Since we suspect a distribution with a heavy tail it would be interesting to know the behavior of the moments in the sample. For instance a characteristic of a power law distribution is the non-existence of higher-order moments [2]. A graphical tool showing this behavior is called the *Maximum-to-Sum* plot. It is defined as follows:

Let $X_1, X_2, \ldots, X_n$ be nonnegative iid random variables. Define for any $p > 0$ the following quantities

$$M_n^p = \max\left(X_1^p, \ldots, X_n^p\right) \tag{2.8}$$

$$S_n^p = \sum_{i=1}^{n} X_i^p\,, \tag{2.9}$$

with M_n^p being the partial maximum and S_n^p the partial sum respectively. Now for $p = 1, 2, 3, \ldots,$ we have the following equivalence

$$\mathbb{E}\left[X^p\right] < \infty \iff R_n^p := \frac{M_n^p}{S_n^p} \xrightarrow{a.s.} 0 \tag{2.10}$$

as $n \to \infty$.

In other words will we know the behavior of the p-th moment when we check whether this ratio converges to zero or not. We will show plots for $p = 1, 2, 3, 4$ where the ratio R_n will be plotted against the number of observations n. So if we notice for a certain p that there is a clear convergence to zero we can conclude from equivalence 2.10 that the p-th moment must be finite.

2.4 Distribution Fit of Tail

The question we want to have answered is: "Is the tail that we have in our data Paretian?". We will test this by fitting a Pareto distribution to the tail of our data. With the tail of the data being all the values above some threshold u that can be found using methods discussed earlier such as Zipf plots or mean excess function plots. We are now interested in the *exceedance distribution* above this threshold u. More formally:

Let Z be a random variable with unknown distribution function G and right endpoint $z_G = \sup\{z \in \mathbb{R} : G(z) < 1\}$. The exceedance distribution function of Z above a given threshold u is defined as

$$G_u(z) = \mathbb{P}\left(Z \leq z | Z > u\right) = \frac{G(z) - G(u)}{1 - G(u)}, \quad \text{for } z \geq u. \tag{2.11}$$

For a suitable u that is large enough this exceedance distribution can be approximated by a Generalized Pareto distribution [6] and this is what we will do. We will firstly explain the Pareto distribution and its properties and next the method of maximum likelihood for the estimation of the parameters.

Pareto Distribution
Is a power-law distribution that is originally used to describe the distribution of wealth in a society but occurs in many other fields in physics and finance. The Italian Vilfredo Pareto discovered this distribution first by noticing that income is distributed by a power-law. Generally speaking a power law distribution is a function of the following form

$$g(x) = ax^{-k}. \tag{2.12}$$

Now a random variable X is said to follow a Pareto distribution if its density function f(x) is such that

$$f(x) = \frac{\alpha x_0^\alpha}{x^{\alpha+1}}\,. \tag{2.13}$$

with α being the shape parameter, which measures the heaviness of the right tail, and x_0 the scale parameter [6]. The scale parameter is basically speaking a lower bound on the possible values that a Pareto distributed random variable can take on.

The most important parameter of the Pareto distribution is the α shape parameter. As its name suggests it determines the shape and the behavior [6]. The smaller the value we take for this α the fatter the tail will become.

The CDF of the Pareto distribution is

$$F_X(x) = \mathbb{P}(X \leq x) = 1 - \left(\frac{x}{x_0}\right)^{-\alpha}, \quad 0 < x_0 \leq x. \tag{2.14}$$

This Pareto distribution is known as Pareto I because it is historically the earliest form that is developed by Vilfredo Pareto himself. Throughout the years generalizations have been proposed and the one that is widely used is the Generalized Pareto distribution (GPD) which is defined as follows:

$$GPD(x;\xi,\beta,u) = \begin{cases} 1 - \left(1 + \frac{\xi(x-u)}{\beta}\right)^{-\frac{1}{\xi}} & \xi \neq 0 \\ 1 - \exp\left(-\frac{x-u}{\beta}\right) & \xi = 0, \end{cases} \tag{2.15}$$

where $x \geq u$ for $\xi \geq 0$, $u \leq x \leq u - \beta/\xi$ for $\xi < 0$, u, $\xi \in \mathbb{R}$ and $\sigma > 0$ [6]. Note that if we set $\alpha = 1/\xi$ when $\xi > 0$ and $u = \beta/\xi$ we would get back the Pareto I distribution.

The tail risk that we are interested in is the probability of X taking on a value larger than the value of x. That is to say that we are interested in the survival function, Equation (2.5), from before. Thus what we want is

$$\mathbb{P}(X > x) = 1 - GPD(x;\xi,\beta,u). \tag{2.16}$$

An important property of a power law/ Paretian distribution, that has been mentioned before explaining the maxmum-to-sum plot, is the infiniteness of the higher order moments and is closely related to the ξ parameter. In fact the moment of order p of a Generalized Pareto distributed random variable only exists if and only if $\xi < \frac{1}{p}$. From this we can see that a small ξ implies an infinite p-th moment and will thus give us the information about the heaviness of the tail. The ξ parameter is therefore quite important for the research in this paper as we are looking at the tail risk of the war data.

In this thesis it is key to find the threshold value u in the data for which we can fit the GPD on. There are quite some different ways to find this value heuristically such as the zipf plot we have discussed earlier. A more systematic approach will be described later.

Maximum Likelihood

The way in which the parameters ξ and β are calculated is by the maximum likelihood estimation (MLE)[9]. This method obtains the parameter values that maximize the likelihood function. For a continuous probability distribution we have the following definition:

Let X be a random variable following an absolutely continuous probability distribution with density function f depending on parameter θ. Then the function

$$\mathcal{L}(\theta|x) = f_\theta(x), \tag{2.17}$$

considered as a function of θ is known as the likelihood function. Maximizing this function will give us, provided that it exists, the maximum likelihood estimate

$$\hat{\theta} \in \arg\max_{\theta \in \Theta} \{\mathcal{L}(\theta; x)\}. \tag{2.18}$$

Intuitively the maximum likelihood estimate is the one that is most consistent with the data that is provided. In other words it is the most likely parameter that is used in the distribution that produces the data.

Statistical Testing

The method of maximum likelihood estimation (MLE) provides us a way of obtaining suitable parameter estimates. Additionally, we would also like to know how well of an estimate we have gained. That is to say we need to check the significance of the estimator. We will do this by means of a hypothesis test with the null hypothesis being the case when the estimate is not significantly different from zero. Hence, if we let x_0 be our estimate, we will get the following t-statistic

$$\frac{x_0 - 0}{\mathrm{SE}(x_0)} = \frac{x_0}{\mathrm{SE}(x_0)} = t. \tag{2.19}$$

Now we would like to build a 95% confidence interval. A 95% confidence interval corresponds, under the normal distribution, to the critical values ± 1.96. If the t-statistic were to be outside of the interval we would reject the null. We can do this because the MLE estimates are asymptotically normal [14]. This implies that we would reject the null if the t-statistic is larger than 1.96 or smaller than -1.96 and rejecting the null means that the estimate is significantly different from zero.

Threshold

We have a method to fit a Pareto distribution to the data using the Method of Maximum likelihood to estimate the parameters and also a way of testing the significance of the estimates. We would now like to know what the threshold value u is for which above the data is Pareto distributed. If this tail, the data that is above the threshold, is Pareto distributed then we say that it is Paretian or in other words Fat tailed. The method we use to search for this threshold is as follows:

We first start of by calculating the 99% quantile of the data. This value will be our candidate threshold value. Next step is to fit a generalized pareto distribution on the data exceeding this value. Since computing maximum likelihood estimates by hand is rather tedious and can get complex we will use the function gpd in the R package evir. Now we will check if the estimates we get are significant by looking at the standard errors that are provided with this function. Suppose everything is alright and we get significant estimators. We then lower the quantile, say to 98% and check if it is again significant. We repeat the process until we get to a quantile value for which the gpd function gives us estimates that are not significant. We can then say that this value (or the one above) is our threshold value which we want to know. From this value on the tail starts and it will be a Paretian/Fat tail.

Robustness

Lastly we would like to check if the shape estimate ξ we computed, relating to the threshold

value we found earlier, is robust with respect to the quality and reliability of the data we have. As will be explained in the next chapter the data we have at our disposal is far from perfect and we may actually miss some observations. That is to say we are dealing with non-precise data, accurately defined in the book [4]. This is possible because some wars may not be well studied/documented or may be missed completely due to no sources being available, e.g. wars in the Americas hundreds of years ago.
The way we will check the robustness of the estimate is by resampling methods, used also in paper [2]. These methods have proven to be effective in dealing with non-precise observations and missing data as we will be dealing with in this thesis. We will make use of two well known methods in statistics:

- First method to check the robustness will be through *jackknifing*. The jackknife method was first developed by the British statistician Maurice Quenouille in the 20th century. What it basically does it computes an estimate of a statistic from different jackknife samples which are samples taken from the original data but have fewer observations than the original. In other words we "cut away" some observations from the original sample and we compute an estimate. The estimate of the shape parameter is computed with the before mentioned `gpd` function that uses MLE. In our case we will remove up to 20% of the data and for each estimate we computed we will be storing it into a vector. Next, a histogram will be made from this vector and if the estimate were to be robust we would expect that the estimates center around the ξ value, which we have earlier computed.
- The second method is through *bootstrapping*. The bootstrap method is inspired by the jackknife and is developed by the American Bradley Efron. It is similar with the jackknife in the way that it produces different samples, but instead of removing observations it creates samples from the original with replacement. We will use 100k of these samples and make a histogram to check whether the estimates center around the ξ value.

2.5 Inter-arrival Times

Maybe the most interesting part of researching statistics in war data is the inter-arrival times between wars. This can give us insight in whether there is statistically speaking a pattern in the occurrence of wars. In other words we are interested if there is a trend towards less frequent onsets of wars or if it remains the same pattern.

What we first would like to know is whether there is any dependence in events in the data and if the data is generated by the same distribution. In other words we want to know if the data is iid or not. This is quite an important step because it implies what kind of analysis we need to do such as time series analysis if there is significant autocorrelation in the data.

Record Plot
A way of checking the iid nature of the data is by means of records. In statistics a *record* is defined in the book [10] as follows:

Let $X_1, X_2, \ldots$ be a sequence of random variables. A *record* occurs if $X_n > M_{n-1} = \max(X_1, \ldots, X_{n-1})$. By definition we take X_1 as a record.

In the definition of record we remark that the number of records is dependent on the order of the data. Because let us suppose that the X_2 is a global maximum in the data set. Then we

would have only one record which won't tell us much. We would rather like to know that if we were to take a random sample of our data what the most likely amount of records would be. Therefore we will take a large amount of samples of our data and compute the number of records for each sample. Then for each sample the computed number of records will be stored in a vector and the average of this vector will give us the most likely number of records. The record plot will then be plotted with on the y-axis the record and on the x-axis the observation n where the record is located. This plot is the first record plot that is made.

The second plot involves the R plot function from the evir package called records. Here again the records are counted and the observations at which they occur are recorded [8]. It is compared with the expected behaviour for iid data and 95% confidence bounds are made. If the records fall within this bound we could therefore say that the data is most likely iid.

Poisson Process

Earlier research has shown that the inter-arrival times of wars follow a homogeneous Poisson process [3, 2, 11]. We will show that the data that is collected in this thesis supports this idea.

Firstly what is a Poisson process? Formally it is defined as follows: Let $(N_t)_{t\geq 0}$ be a random process with the following properties [12]:

(1) $N_0 = 0$

(2) *independence*: if $0 \leq s < t$, then the number of events which arrive during the time interval $(s, t]$ is independent of the arrivals of events prior to time s. In other words we say that $(N_t)_{t\geq 0}$ has *independent increments*.

(3) the number of events in any interval of length t is a Poisson random variable with parameter λt.

To empirically check whether the "events" in the data are generated by a Poisson process the following theorem will be helpful:

Theorem 2.5.1. *Let $(N_t)_{t\geq 0}$ be a Poisson process and let $X_1, X_2, X_3, \ldots$ be the inter-arrival times between successive arrivals of events. The random variables $X_1, X_2, X_3, \ldots$ are then independent and exponentially distributed.*

A proof of this theorem can be found in [13]. Now we will apply Theorem 2.5.1 to our data. In other words we will calculate the inter-arrival times between wars, check for their independence and check whether they are exponentially distributed. Since this is a necessity for a Poisson process we will know that the inter-arrival times of wars follow a (homogeneous) Poisson process.

To achieve this in practice we will first calculate a vector in R with each entry an inter-arrival time. An inter-arrival time is calculated by taking the difference between the starting year of a successive war. Next we will first check if the inter-arrival times are exponentially distributed. As we did before we will use a QQ-plot with theoretic distribution the exponential distribution. If the points in the vector lay roughly on a diagonal line we could then say that the inter-arrival times are most likely exponentially distributed.

Last thing is to check the independence of the inter-arrival times. We will check this by using again the record development plots explained before. To remind the reader, a slow logarithmic growth in a record development plot indicates that data is iid.

3. Data

The very first thing that is needed, to do the statistical analysis described in Chapter 2, is data. This chapter will explain the data that we will be using throughout this thesis. Some natural questions that arise concerning war data are for example the following:

- How to quantify war and the severity of the war?
- How to compare wars?
- What sources for the data is available?

Important is to first establish that in this research a "war" will be defined as war in the broadest sense of the definition. This means for example that also civil wars, regimes (e.g. Stalin's regime) and revolutions are also included in the data.

There could be a lot of ways to quantify the severity of the wars. For instance you could count direct military victims as a way to measure how large a war is. Or perhaps the economic consequences of a war might also be a good measure on the impact of a war on the population. Ideally, if this data were easily available it would be indeed quite interesting way of quantifying a war. Unfortunately this is not the case and we need to look for another way to measure a war. So what can we do? Given that a war has some "side effects" such as disease, hunger and a rise in criminality, most historical records of wars, especially the large ones, denote the total number of casualties and thus, also the ones that are not caused by direct conflict. This seems not ideal and perhaps it is indeed not. However, since this is available, it is quite likely the closest we can get into doing proper statistics on a topic that is not as concrete as we would like it to be. Therefore, in this thesis we will utilize the total number of casualties as a measure of the severity of a war.

3.1 Data Collection

The data that contains the total number of casualties of each war starting from roughly the 16th century will be used throughout this thesis and is thankfully available on Matthew White's website Necrometrics [1]. Per known war the website lists a number of sources (historians, journalists or governments) with the total amount of casualties during that war.

From this website the data has been stored into Microsoft Excel. The standard format that has been used for collection is the following per war:

- name_of_war
- begin_year_war
- end_year_war
- header_number
- century
- min_estimate
- max_estimate
- median_estimate
- mean_estimate
- notes

Most of the column names are somewhat self-explanatory. Every war contains a total number of casualties. Since we have multiple sources often claiming different numbers we need to find the "most likely" estimate of the number of casualties during a war. Intuitively it makes sense to take the most central value of the published numbers. In this way the party that perhaps overestimates the number of casualties compensates more or less for the party that underestimates.

There are two options of calculating this middle value. One will be taking the average (mean) between the minimum estimate and the maximum estimate. The other is calculating the median of the sample of estimates given by the different sources. The median is the value separating the higher half from the lower half of a data sample [5].

The header_number refers to the number that will be Matthew White's estimate of the Median number [1]. This number is not always provided on the website and is included in the data for the sake of completeness and preventing mistakes when transferring it in the excel file. For the research done in this thesis the median number has been calculated again for every war listed on the website and these calculated medians will be used throughout the paper along with the calculated mean values for every war.

The data that is collected will be shared on the website [15] from the publisher MDPI. It will be made available at the end of this thesis.

Time Span

We will now comment on the time span of the data we use in this thesis. The earliest war present in the data provided by Necrometrics is a war that goes by the name of "Charlemagne" during 768-814CE. It is the name of a king (a.k.a. Charles the Great) in the 8th century who united Western and Central Europe after the fall of the Roman Western Empire three centuries earlier. During his leadership he was the first to campaign against the Saxons which has started the subsequent Saxon Wars. These wars are very well known today in the modern era through the Legend of King Arthur, who is said to have led the defence of Britain against Saxon invaders. It has been claimed that stories of Charlemagne have been a significant inspiration for the stories of King Arthur.

Important is to note that in a few cases we do not literally take the official time span of a war. A good example is the Korean conflict between North Korea and South Korea. Officially they are still (1945-2019) at conflict, however most violence took place between 1950 and 1953 and therefore in the data, we will consider only this time period where most casualties fell. The same goes for the ones that are still ongoing during this thesis. In these cases we needed to do a bit of research on which wars are being actively fought and which ones are in a state of armistice for a long period. Another example, and also one that has been quite hard to estimate, is the conflict in Colombia (1964-2019). In addition to non-constant fighting there are also multiple parties involved such as the FARC, Colombian government and far-right paramilitaries. This results in many sources giving different numbers over different time spans. Therefore some extra reading was involved to get the most reliable estimate in this case.

The latest wars in the data are the wars with the latest starting year and two of them are still being waged now during the writing of this thesis. We have three of such wars all starting from 2014 and they are respectively the following

- Ukrain vs. Russian separatists
- Israel vs. Hamas
- War on ISIS

The latest updates we can give at the time of conclusion of the thesis is that the War on ISIS is officially over and ended on the 23rd of March 2019. The other two wars are still officially ongoing due to persistent instability in the regions as the reader may know.

Now, it could be said that the exact time span that our data covers would be 768 till 2019. However we should emphasize that especially between 768 and the 16th century there are quite some missing wars and imprecisions due to the fact that wars have not been well documented then as it is now. In the next section we will discuss the problems and imprecisions further.

3.2 Data Problems

There are several problems in war data that need to be mentioned. It is best summarized in the following quote:

"Accounts of war casualties are often anecdotal, spreading via citations, and based on vague estimates, without anyone's ability to verify the assessments using period sources." (Cirillo & Taleb, 2016, p. 31)

Indeed the numbers that are provided can be quite dubious and it cannot be verified in any way. However there is the fact that when wars result in a great number of victims these wars tend to be studied more by historians. This is why we can be more confident about the numbers presented in large wars and this is also a nice motivation to research the tail of the data.

Imprecisions
Important is to note that in this thesis the total number of casualties will be counted during the period of war involving all the parties related to the war. This means that also the casualties due to disease, hunger and executions are included in the estimates of the data. Hence the estimates in the data would not only include battle deaths. In the Methodology chapter it has been mentioned that the data we have is not precise and that some wars could not be included because there are no sources available. However we are quite confident that the missing conflicts will not be occurring in the tail of the data simply because very large wars will not remain unnoticed by historians. We will show in the Results chapter through re-sampling methods that the estimates we get using our data are indeed robust against missing observations and imprecisions.

We give a short example of an imprecision in an estimation that is politically coloured. Let us consider Stalin's regime, that has an estimated median value of 25 million casualties. The numbers that resulted in this estimate come from sources that are on the left and right wing of the political spectrum [1]. These sources estimate a total amount of 20 million and 30 million casualties respectively. This is of course a huge difference in numbers and accounts perhaps as the most extreme variability in the sources of a single data entry.

Lastly and quite importantly we have the following problem. Namely that with 11 entries in the data the median is not available. This is due to the fact that for some wars only one source is given that provides a range instead of a concrete number of casualties. For example a war that is said to have 1000 to 2000 casualties coming from a single source will have no entry in the median_estimate column but will only have an entry in the mean_estimate column. In the analysis of the data these 11 special cases will have a median value equal to the mean value to solve this issue.

4. Results

The results from the methods applied on the data that we described earlier will be presented in this chapter.

4.1 Descriptive Statistics

After having collected the data the descriptive statistics will be given for the median_estimate and the mean_estimate columns. These statistics can be found in table 4.1:

Description	Median	Mean
min	42	42
Q1	2287	2500
median	14000	15000
mean	600900	596500
Q3	100100	124400
max	5e+07	5.55e+07
sd	3589946	3595663
skewness	10.08947	11.02434
kurtosis	113.4153	138.0859

Table 4.1: Descriptive Statistics Median/Mean.

The table starts of with the minimum value of the median_estimate and the mean_estimate columns respectively which is recorded in 1967 in Greece. Next the quantile values are given and the maximum value of 50 million from WWII. The sd in both the mean and the median columns seem to be relatively high which would indicate that the distribution is highly dispersed.

Looking at the shape of the distribution of the data the table shows a skewness of around 10. This indicates that the distribution is remarkably right tailed compared with the symmetric distribution of a normal. Also the kurtosis measured from the data of total casualties is well above a 100 which indicates that it has a lot more extremes then that of a normal.

From these descriptive statistics the conclusion can be made that the distribution of the data clearly takes the shape of a long or fat right-tailed one.

4.1.1 Graphical Representation

In this section the data will be described visually through some well known plots described earlier. To first get an idea of how the data is distributed we will provide a histogram in the following figure.

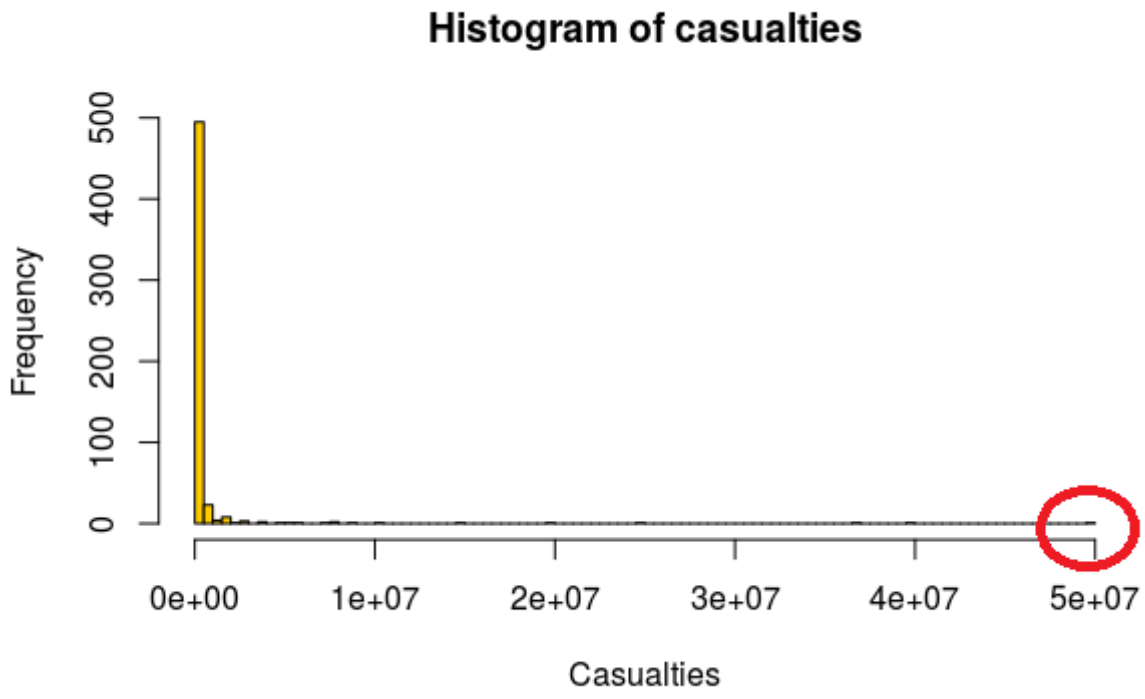
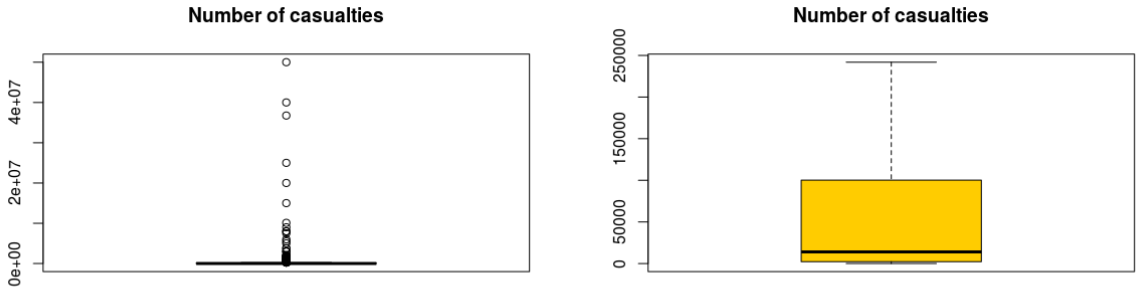


Figure 4.1: Histogram of the median war casualties in armed conflicts from 768CE to 2019.

Figure 4.1 shows a histogram made from the column median_estimate. The red circle represents the datapoint of World War II which is known to be the deadliest conflict in human history. As we have just seen in the descriptive statistics it can clearly be seen that the data is (very) positively skewed and has a long right tail which in the next section will be properly detected by graphical tools.

Another simple but clear way of representing the data is by means of a boxplot. This is shown in the following figure



(a) With extremes

(b) Without extremes

Figure 4.2: Boxplots of the total number of casualties.

The histogram in Figure 4.1 suggests that there is a presence of a long right tail and therefore it is to be expected to have quite a few extremes. The boxplot in Figure 4.2a shows a squeezed boxplot with a significant amount of extremes above which would confirm that there is quite a long tail present in the data. By contrast Figure 4.2b shows the boxplot without the extremes and gives a better picture.

An interesting question that could be asked and perhaps give us a direction to answer would be: Given the data about the number of casualties, what would be considered a 'large' war? Clearly, there is not a concrete answer but Figure 4.2 gives us a few options. We could for instance definitely say that the extremes are considered to be 'extraordinary big' wars which according to Figure 4.2b would mean any war with greater than 241872 amount of casualties. A normally large war would then be any war past the median value of 14000 or past the Q3 value 100199. Later we will find a better way to find a certain threshold from which our tail i.e. the largest wars start.

4.2 Tail in Data

Now that we have seen what the distribution looks like through histograms and boxplots, we will use other methods to graphically detect the presence of a right fat tail.

We will start off with the earlier explained *mean excess plot*. We will use the R function meplot from the evir package and we will apply it on the data. We get the following plot:

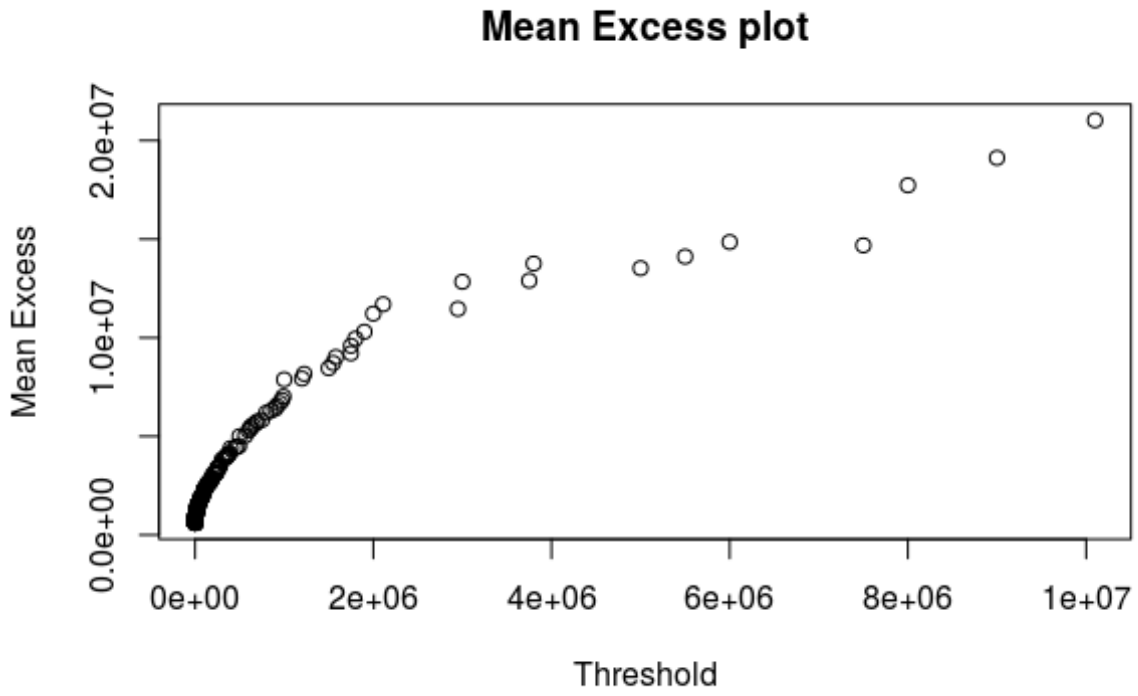


Figure 4.3: Mean Excess plot showing an upward trend indicating heavy-tailed behavior.

In Figure 4.3 we can clearly see an upward trend. If there is an upward trend in a ME plot it means that the data shows heavy-tailed behaviour [8]. This corresponds with what we have seen in the descriptive statistics.

Now when the empirical distribution gives a good approximation of the theoretical distribution the points in the plot should be close to the diagonal. In the following figure we can see the QQ-plot applied to the data with as theoretic distribution the exponential. Since the exponential distribution is *the* distribution with thin tails it is the natural benchmark for studying the tails.

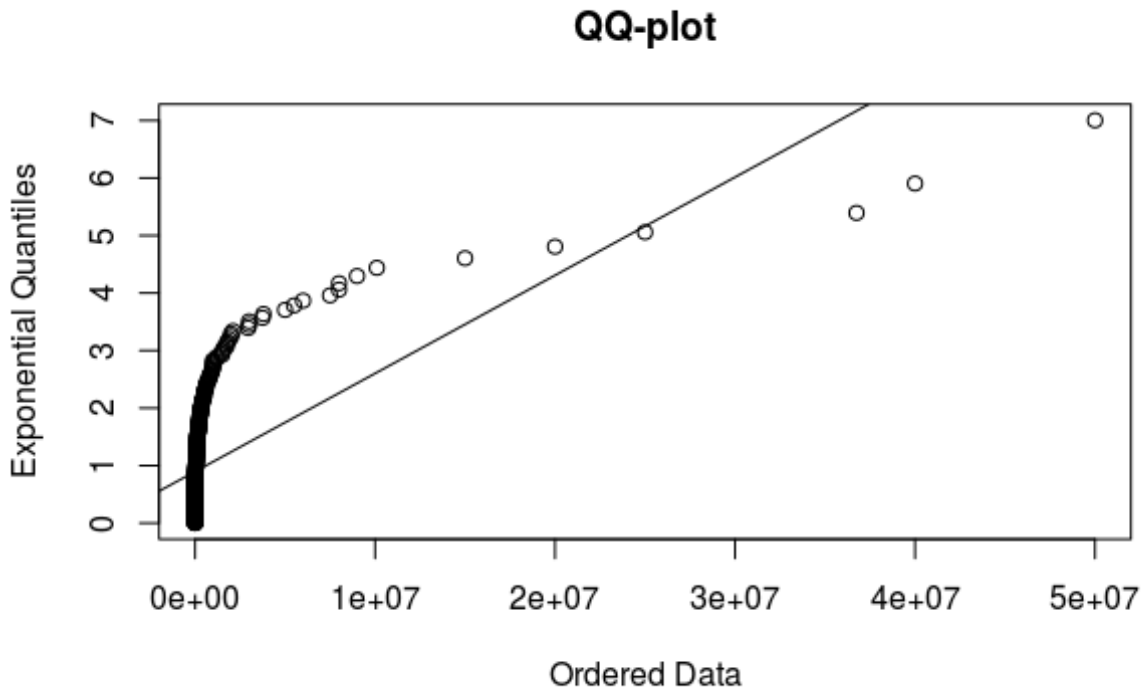


Figure 4.4: Concave departure from straight line indicating heavy-tailed behavior.

Now in Figure 4.4 we see that the points do not occur on the straight line. Rather it departs from the straight line in a concave way. This sort of shape is also a sign of heavy-tailed behavior in the data [8].

Zipf plot

We will have a look at our last graphical tool to detect heavy tails, namely the Zipf plot. Remember from the Methodology chapter that it plots the logarithm of the empirical survival function against the logs of the ordered values of x. The next figure will show the Zipf plot applied to our data.

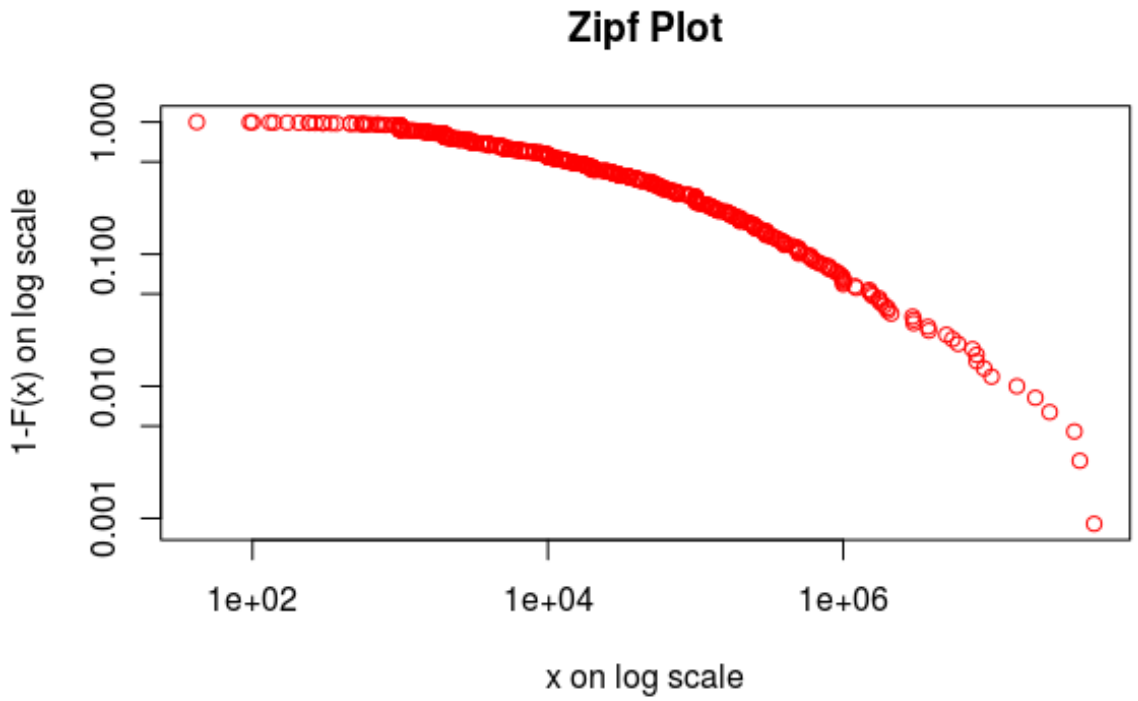


Figure 4.5: Zipf plot of data showing somewhat linear decline past some threshold value u.

From this plot we can observe that the graph declines in a somewhat negative linear way on the right hand side of the plot after some threshold value u. This would indicate that the data follows a power law and that above this threshold Paretianity holds [6].

4.3 Moments

Next we come to the essential property of a Pareto distribution. Namely the non-existence of moments and we will show the behavior of the moments in the maximum-to-sum plots that has been explained previously. The following figure shows the plots that have been made.

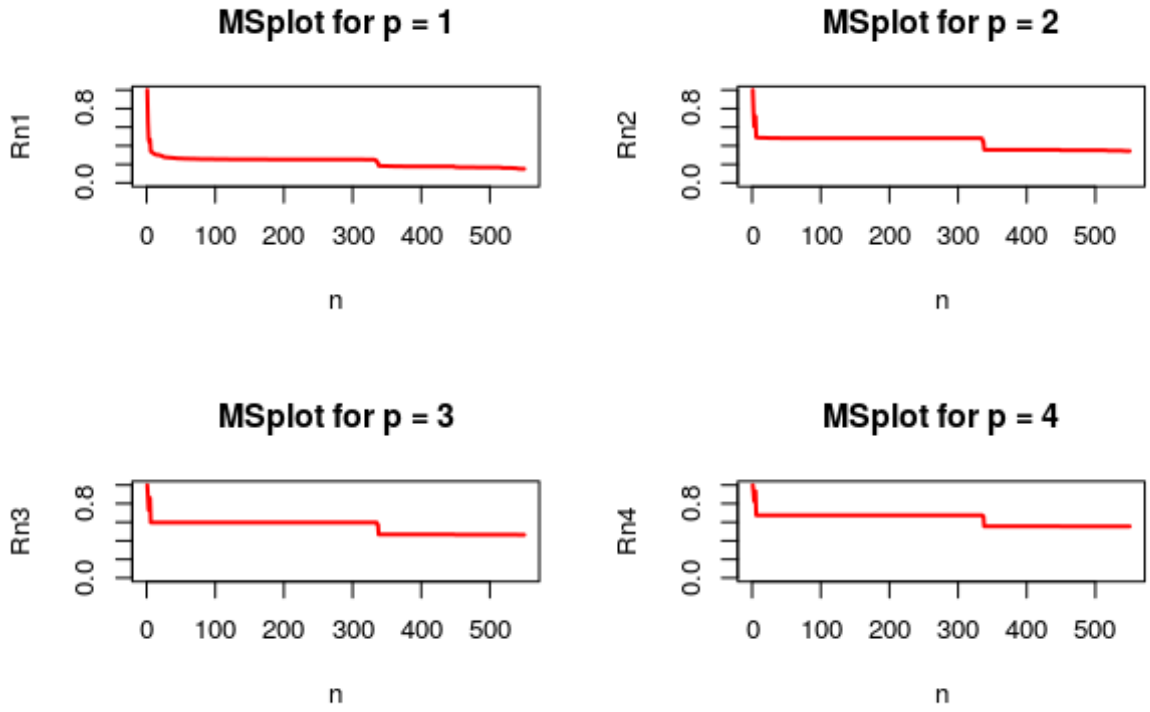


Figure 4.6: Maximum to sum plot showing the behavior of the four moments resulting in the ratio R_n not being convergent.

In Figure 4.6 we can clearly see that the ratio R_n does not converge to 0 for $p = 2, 3, 4$. From Equivalence (2.10) we can conclude that these p-th moments do not exist. This in turn would suggest that the right tail is such that the moments do not exist. This validates the suggestion that the tail is Paretian due to its property of non-existent moments.

4.4 Distribution Fit of Tail

The previous section gives us an idea of what the distribution looks like: it is a distribution with most of its mass at the beginning and has a long right tail. Now we will be having a closer look at this tail. Already with the QQ-plot in Figure 4.4 in previous section we found that the tail of the distribution shows some heavy tailed behavior.

In this section the main goal is to fit a Pareto distribution to the tail of the data. The (Generalized) Pareto distribution is thoroughly explained previously in the Methodology chapter. There, we said that a distribution like the Pareto had a few parameters namely the scale and the shape parameter. Since our goal is to model the tail of the distribution we are especially interested in the shape parameter because as it is mentioned earlier it is the one that decides the heaviness of the tail. In order to determine an appropriate shape parameter we first need to find the threshold value for which the tail starts. Previous plots as the mean excess Figure 4.3, the zipf plot Figure 4.5 and maximum to sum Figure 4.6 all indicate roughly where the threshold value

may be. Especially from the behavior of the moments in the maximum to sum plot we can see that it makes a slight jump between observations $n = 300$ and $n = 400$. From this moment on the tail takes over the behavior of the moments. Now in the sorted data observations $n = 300$ and $n = 400$ correspond to a quantile of approximately 57% and 75% respectively. Since our candidate threshold is somewhere in between we have fitted the GPD distribution on data from different quantile values. We proceed as explained in the Methodology chapter and come up with the following results.

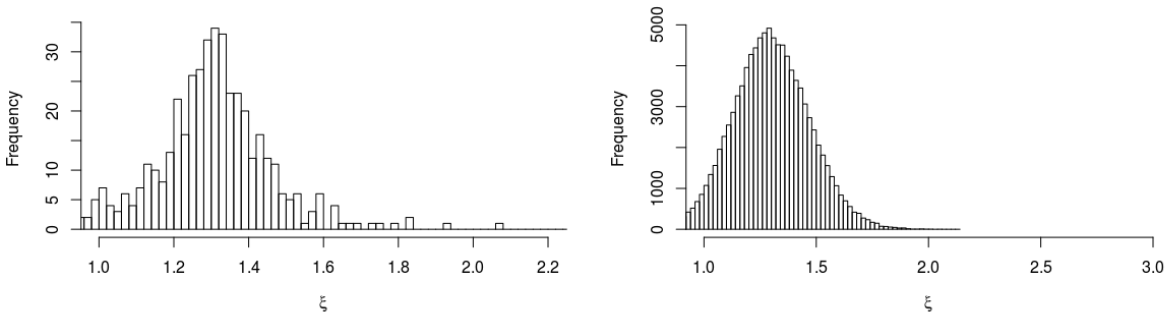
Quantile	ξ	SE
95%	1.043870	0.4709813
90%	1.117487	0.2857727
85%	1.336222	0.2689498
80%	1.300995	0.2225280
75%	1.268552	0.1932948
70%	1.310759	0.1810389
65%	1.358282	0.1747441
60%	1.345364	0.1590392
57%	1.389182	0.1571257

Table 4.2: Maximum likelihood estimates of the tail parameter ξ of the Generalised Pareto tail for different thresholds, expressed in terms of quantiles.

The estimates of the shape parameter can be seen in Table 4.2 with their standard errors. An additional technicality that has been proven to be crucial is re-scaling the data. The numerical algorithm used in the R package evir turns out to have trouble with the large numbers that our data contains which resulted in the standard errors not being computed. Therefore we have divided the numbers by 1000 to get estimates that we can validate.

If we take a 5% significance level we can conclude that all the ξ estimates are significant. Now what can we say about these estimates? It can be seen that from the 57% quantile onwards the ξ estimate remains more or less stable around the value 1.3. The stability of this value is a result of a property of a power law distribution which is the scale invariance [10].
However for higher quantiles above 90% we see that it starts to deviate remarkably from 1.3. We will comment on these deviations for higher quantiles later in the Conlusions chapter. For now we know that the tail starts somewhere between 57% and 75% and that for fitting a GPD distribution from these quantiles on we get a ξ estimate of approximately 1.3. As our desired threshold value u we will take the 70% quantile which corresponds (coincidentally) to 70000 casualties in the data. Similarly the scale parameter β has been estimated for the 70% quantile and for which we obtained $\beta = 193397$ with an SE of 32577. We have now succeeded in estimating every parameter for the tail risk survival function 2.16 that we aimed to attain.

Lastly in this section we will check for the robustness of the ξ estimate. In Chapter 2 we have explained the resampling methods we will use. The results are the following plots



(a) Jackknife, removing up to 20% of observations (b) Bootstrap, using 100k bootstrap samples

Figure 4.7: Re-sampling methods for testing the robustness of the ξ estimate.

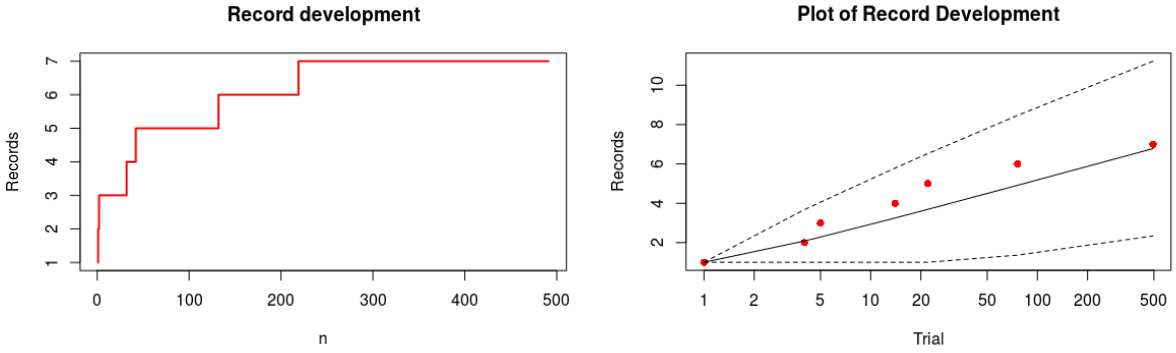
In both Figure 4.7a and more so in Figure 4.7b we see that the ξ estimates neatly center around the value 1.3 from Table 4.2. As expected from reasoning in Chapter 3 that large wars are not so much affected by imprecisions and also from an earlier paper [2] we can conclude that our ξ estimate is remarkably robust.

4.5 Inter-arrival Times

The inter arrival times between wars will be researched to investigate if there is a pattern or a trend in the data.

4.5.1 Record Development

We first need to remark that the data we have about wars do not in general have time dependence. This makes quite some sense if we for example consider a war in Europe and a war somewhere in Asia around the same time. Quite often it is the case that wars being waged at such a great distance apart do not have any relation with each other. Of course there are cases in which this can be true e.g. World War II. Generally speaking however from an epistemological we do not believe these data represent a proper time series. We will further show graphically that indeed the data has an iid nature through the before explained record plot.



(a) Record development

(b) Record plot

Figure 4.8: Record development plot with confidence intervals

Figure 4.8 shows the development of the records over the sample. As has been explained before we have taken multiple samples of the data and calculated the average amount of records which results in 7 records in total and plotting the records against the sample it results in figure 4.8a. Now the number of records of iid data grows very slowly and will graphically represent a logarithmic function [10]. Indeed the record plot in Figure 4.8a of our data does show a logarithmic like function. The plot therefore supports the idea that the data we work with is iid. The second plot Figure 4.8b is made with the R function records and builds confidence intervals. The plot shows that it falls comfortably within the bounds and thus the data is most likely to be iid. Therefore we can fortify our claim that there is no time dependence, hence no meaningful war generator over time.

4.5.2 Distribution and Independence

Here the inter-arrival times will be tested for independence and whether they are exponentially distributed because then the idea would be supported that large armed conflicts follow a homogeneous Poisson process. The QQ-plot can be seen in the following figure.

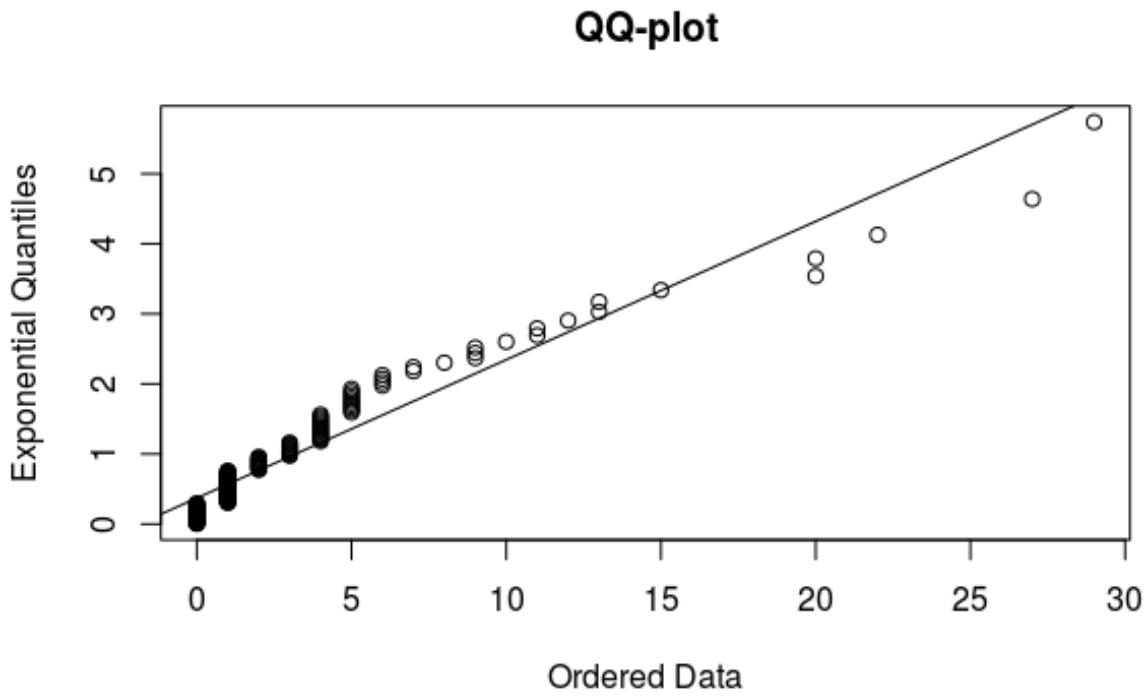
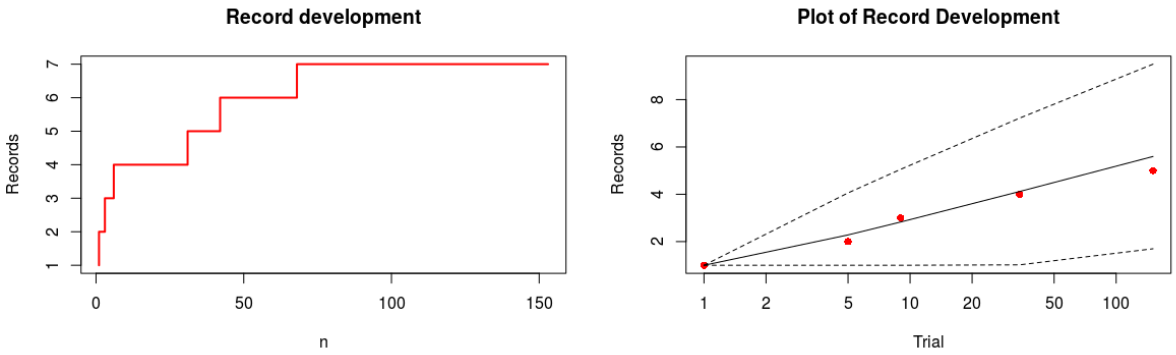


Figure 4.9: QQ-plot with theoretical exponential distribution

Figure 4.9 shows the data points being close to the diagonal line but still deviates a bit from it. The figure could support the hypothesis of the data coming from an exponential distribution. In the conclusion we will further discuss this plot.

To check that the inter-arrival times are independent we provide again a record plot.



(a) Record development

(b) Record plot

Figure 4.10: Record development plot for the inter-arrival times

We observe again the slow logarithmic growth in figure 4.10 that indicates that we are dealing

with iid data [10]. Figure 4.10b shows that the records are quite close to the diagonal straight line that represents expected iid data. In conclusion, we can say that the inter-arrival times are quite likely iid.

Now that the inter-arrival times seem to be exponentially distributed and independent Theorem 2.5.1 tells us that the inter-arrival times follow a (homogeneous) Poisson process. Considering earlier research, this is as we have expected.

5. Conclusion

The goal of this thesis was to research the risk of large wars and their occurrence. Large wars are interesting to research due to their impact on human population, politics, economy and so on. The fact that extremely large wars that are researched the most by (independent) historians also gives us the right motivation to study these particular cases. We wanted to give the reader an idea of how to go about objectively researching a topic that is quite vague and perhaps controversial at the same time. The goal was to quantify wars as best as possible and compose objective statements about the risk and occurrence of large ones. More precisely we looked at the tail of the distribution, what kind of tail we are dealing with and from what number of casualties we can speak of a "tail" in the distribution. Graphical tools have given us a way of heuristically detecting a fat tail and gave us some degree of validation that we can fit a Pareto distribution to the tail in order to find the tail risk we are interested in. In this chapter we will comment and discuss the results that we have found and if they agree or disagree with the expectations we coined earlier and the research that has been done before.

We first started with the very fundamental thing we needed for this research, namely data. We expected to see a large number of relatively small scale conflicts and a small number of relatively large scale conflicts. The always useful histogram Figure 4.1 and the boxplots Figure 4.2 showed us immediately that indeed the frequency of "small" conflicts (below 14000 casualties) is way higher than for the larger ones. Here we have observed a tail and ask ourselves what kind of tail this is, as it represents the wars with the largest amount of casualties we have in the data. Here, the use of some other quite common graphical tools had come to our aid. The mean excess plot, the QQ-plot and the Zipf plot all have indicated in their own way that we are dealing with a tail that shows heavy tailed and even fat tailed behavior. In other words, that we are dealing with a Paretian tail.

A property of a Paretian distribution is the non-existence of higher order moments and the maximum-to-sum plot we made for moments $p = 1, 2, 3, 4$ told us exactly this. At this point, we became quite confident that we can model the tail through a (Generalized) Pareto distribution. In order to do this we needed to find a suitable threshold u that is used as a parameter in the distribution. The best we could do of finding this threshold value was by using the graphical tools we have used that detected heavy tailed behavior.

The Zipf plot for instance indicates a fat tail if it shows a negative linear relationship. The article [6] told us that the threshold u can be found at the point where the plot starts to become negative linear. If we look at the Zipf plot in Figure 4.5 that we created, it is not clear from which point on the plot becomes negative linear. It almost seems as if the plot becomes piecewise negative linear which resulted in not being immediately useful in finding the threshold value u.

What did turn out to be very useful in finding the value were the maximum to sum plots: Between observation 300 and 400 we spotted a little break and it indicated that the threshold

value could be somewhere in between. The next step was to start fitting the GPD to different quantiles between observation 300 and 400. Here, we noticed that the parameter estimate ξ was somewhat stable around the value 1.3. However, as can be seen in Table 4.2 for higher quantiles, the estimate changes. There are methods in dealing with these higher quantiles but they won't be covered in this thesis. Since the estimate ξ seems to be stable and significant between 57% and 75% we could take an arbitrary quantile within as our threshold u. For our results we went with the 70% quantile as threshold; of course, there may be better more precise ways of finding the value. However with the tools we used in this research it seems the best we could do. We thus have achieved to determine the tail risk as we have successfully fitted the GPD to the tail.

Furthermore we have tested the robustness of the $\xi = 1.3$ estimate. We used re-sampling methods and found that the estimate is robust against missing data and imprecisions. This was expected due to results from earlier research [2] and also from the fact that the data that represents large wars do not or carry very little missing data because they are well studied by historians.

The last part of this thesis involves the inter-arrival times between the wars. Considering previous research we wanted to prove or disprove the idea that the large conflicts follow a homogeneous Poisson process and we wanted to show this using theorem 2.5.1. We first checked if the inter-arrival times are exponentially distributed.

As we did before to check for heavy tailed behavior we made a QQ-plot with the theoretic distribution the exponential. The QQ-plot 4.9 shows that the discrete points are close to the diagonal line but not right on it. This could suggest that the distribution of the inter-arrival times deviates slightly from the exponential. We remark that in an earlier paper [2] the points came much closer to the straight diagonal line. At the same time however, the points do not deviate as much as for example in the earlier QQ-plot 4.4 that we have made suggesting that both ways of thought can be reasonable.

The last part was checking for independence. We already reasoned that the data could not possibly represent a proper time series, since there does not exist a war generating process. If we were however, for a moment, considered the hypothesis that there is time dependence, we have showed graphically that there is not through records. Thus, based on the record development plots, the idea that the inter-arrivals are iid seems to be indeed very likely. Now from being exponentially distributed and from independence, Theorem (2.5.1) follows and therefore the inter-arrival times are described by a Poisson process.

From deriving that the inter-arrival times follow a Poisson process we can state that there is no particular or special trend to be found in the onsets of wars. This indeed supports the idea of earlier research such as the ones from Cirillo and Taleb (2016) and Clauset (2018). Namely the idea that big wars, i.e. in our case wars that have over 70000 casualties, have neither decreased nor increased over time.

Recommendations

The data used in this thesis has been collected and explained in the Data chapter. The attentive reader may have noticed that the results throughout the thesis mainly have been produced using just two columns from the whole data set. Namely the median estimates and the column with the start year of wars have been used. More and different results could be obtained from using the other columns and we highly recommend the reader to try things out using the data set. The data set will be open for research on the publishing website MDPI [15] when this thesis is finished.

Bibliography

[1] Matthew White (2014).
URL https://necrometrics.com/

[2] Cirillo, P., Taleb, N.N. (2016). *On the statistical properties and tail risk of violent conflicts.* Physica A.

[3] Aaron Clauset. *Trends and fluctuations in the severity of interstate wars.* Science Advances, University of Colorado, 2018.

[4] Reinhard Viertl (1995). *Statistical Methods for Non-Precise Data* CRC Press, Inc.

[5] David J. Sheskin (2003) *Handbook of Parametric and Nonparametric Statistical Procedures Third Edition.* CRC Press.

[6] Pasquale Cirillo. *Are your data really Pareto distributed?.* Physica A, Delft University of Technology, 2013

[7] Pasquale Cirillo. *Not all tails are the same, A taxonomy* https://twitter.com/drcirillo, 2018

[8] Bernhard Pfaff and Alexander McNeil (2018). evir: Extreme Values in R. R package version 1.7-4. https://CRAN.R-project.org/package=evir

[9] Ronald Fischer. *Inverse Probability.* Mathematical Proceedings of the Cambridge Philosophical Society, Cambridge, 1930

[10] Embrechts, P., Kluppelberg, C., Mikosch, T. (1997) *Modelling Extremal Events*, Chapter 6, Springer-Verlag

[11] Hayes, B. (2002) *Statistics of deadly quarrels* Am. Sci. 90, 10-14,

[12] Sheldon M. Ross (1996). *Stochastic processes* Wiley. pp. 5960

[13] Grimmet, Welsh (2014). *Probability: An Introduction*

[14] John A. Rice (2007). *Mathematical Statistics and Data Analysis* Brooks/Cole

[15] MDPI (1996).
URL https://www.mdpi.com/journal/data