A Non-mathematical Introduction to Statistics of Extremes

by Erik Brodin

Erik Brodin
ebrodin@math.chalmers.se

Extreme events do happen – we know that – but what is the probability? As the extreme events seldom occur they are in general too difficult to handle with standard statistical methods. However, by the use of extreme value theory one has a methodology which is explicitly made for dealing with extreme events.

In this introduction we describe the theory of extremes from a statistically point of view. We give examples of the most used statistical methods for analysing data. Finally, we perform a case study where we investigate the extreme behaviour of fire insurance claims.

Introduction

This short introduction aims at describing extreme value theory and its applications. However, the language is statistics and the grammar is mathematics but this will be tried to be held at a minimum. The introduction is by no means complete and the interested reader is referred to Coles (2001) and Embrechts et al. (1997). The book by Coles (2001) is from a statistical point of view while the book by Embrechts et al. (1997) is more theoretical but still focusing on applications such as insurance and finance.

A common problem in many areas of science is to answer the question how likely it is that a certain event occurs. Also, one is often more interested in events that has a low probability to occur, so called extreme events, as these events tends to influence an environment more than a likely event. Natural examples are storm damages like the storm Gudrun over southern Sweden in 2005 and stock crashes as the Black Monday in 1929. However, these events are in general too difficult to handle with standard statistical tools and therefore demand special methodology. To see this, we give the following example:

> *Consider the constructions of the dikes in Holland. These dikes are vital for protection against flooding. Here one is interested in building the dikes higher than the 10 000 year return wave, i.e. the highest wave during 10 000 years. How high should the dikes be?*

Erik Brodin is a Ph. D student in Mathematical Statistics at Chalmers University of Technology. His research area is risk modelling with application to insurance and finance. In 2005 he received a grant from the foundation "Willis Corroons insamlingsstiftelse till Lennart Elmlunds minne".

A natural way to answer this question is to use statistical methods to estimate the height of the highest wave in 10 000 years. But how should one do this when one only has measurements from a couple of hundred years? If one instead, for some reason, chose to build the dikes higher than the 100 year return wave, it would be straight forward to use the data to estimate the height. However, to estimate the probability of an event that is more extreme than any that has already been observed demands different methodology. Such methodology is extreme value theory.

Extreme value theory is an applied and theoretical science which has been developed rapidly during the last 50 years but is by no means uncontroversial. Using mathematical results based on extreme value theory (under suitable assumptions) one can extrapolate observed data to answer questions about extreme events. Naturally, this is easy to criticize as extrapolation is by nature unreliable. However, extreme value theory has a mathematical foundation and no other credible alternative has been proposed. As Professor Richard Smith said:

> *There is always going to be an element of doubt, as one is extrapolating into areas one doesn't know about. But what extreme value theory is doing is making the best use of whatever you have about extreme phenomena.*

Also, all statistical work using extreme value theory is based on model assumptions which of course almost never are exactly as the complex real world. Hence, one must, as always, proceed with caution.

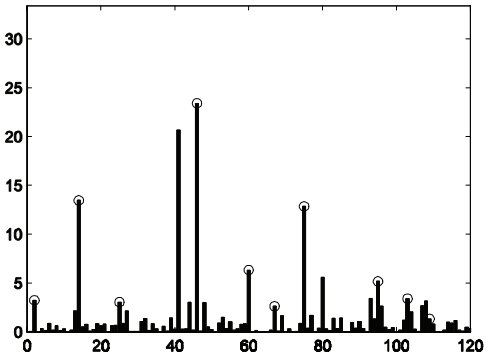
Basic Concept

When using extreme value theory in statistics there is basically two different approaches. The first is called the *Blocks method* and the second the *Peaks over Threshold method*.

Blocks method

The Blocks method deals with maximum observation during a given time period. For instance, the highest yearly observation of rain. Statistically one takes advantage of the fact that a maximum observation of a time period, given that the time period is sufficient long, starts to behave as maxima observations of longer time periods. An example of the Blocks method can be seen in Figure 1.

Figure 1: 10 years of randomly generated monthly observations. Yearly maximums marked with circles.



Here we have 120 randomly generated observations which can be interpreted as monthly observations during 10 years. Marked with circles are the yearly maxima. To answer the question about the dike height, with support in extreme value theory, one models the highest 10 000 year wave by considering the observed highest 1 year waves. Of course, it is better to consider a longer time period but this is a trade-off, as always in statistics, with the number of observed maximums. For instance, it is better to have 10 yearly maxima than one 10-year maximum. To be able to do the hopefully correct thing there exist several graphical tools. More on this topic can be found in the references.

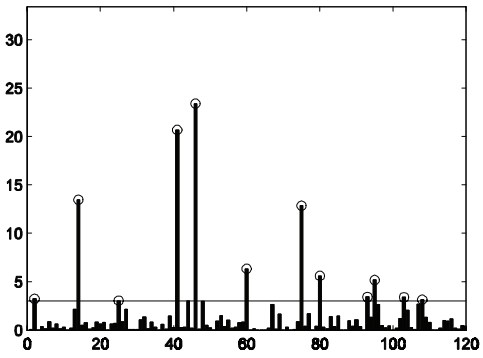
It is easy to see a disadvantage with this method. In any year there may have occurred an event that is more extreme then the maximum in other years and hence important information is not used. An advantage with this

method is that it is natural to store data as maximum.

Peaks over threshold method

The Peaks over threshold method does not consider blocks maxima. Instead all observations larger than a given threshold are used in a statistical analysis given that they are. This is illustrated in Figure 2 where the same observations as in Figure 1 are used. Extreme value theory then gives that data behave similarly over the threshold if the threshold is sufficiently high. This implies that one usually has more observations than with the Blocks method. However, one can show that the methods are the same, see Embrechts et al. (1997), but this is not the topic of this paper. One critical and difficult question is how to select the threshold, which in a way resembles choosing the time period of one block in the Blocks method. This is a delicate task and graphical tools are used. We will present some of them in the next section but we refer to the references for more in-depth material.

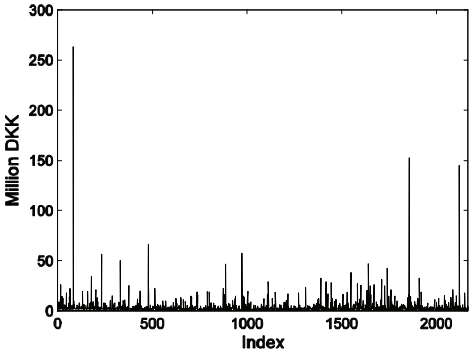
Figure 2: 10 years of randomly generated monthly observations. Values over the threshold marked with circles.



Danish fire insurance claims

Describing statistical tools is much easier when using an example. Therefore, in the following section we will perform a statistical extreme value analysis of the Danish fire insurance claim data using the peak over threshold method. This analysis is by far not complete but the purpose is to give a glimpse of the techniques used. The data consists of 2167 observations of large insurance claims between 1980 and 1990 (1985 prices) in Denmark. We will try to answer the question: "What is the loss level for which only 0.1 % of the losses would be larger?" In statistical language this is called the 0.999-quantile. This is an important question which is of importance in, for example, risk management or the reinsurance industry, which daily struggles with how to insure an extreme event.

Figure 3: Danish fire insurance claims, 1980-1990.



In Figure 3 we have depicted the data. The mean of the claims is 3.4 million DKK and the standard deviation is 8.5 million DKK. As we have several extreme values, one as large as 263 million DKK, the standard statistical methods do not describe the data well. Also, there is about 6 % of the events larger 10 million DKK but they sum up to about 40 % of the total loss amount.

To continue our analysis we have to find the threshold for Peaks over threshold analysis. This will be discussed later. The threshold will be denoted u. Claims which are larger than u will be assumed to behave statistically equally. Now, mathematics of extreme value theory gives, if X is an insurance claim:

$$P[X > u + x \mid X > u] \approx \left(1 + \frac{\gamma x}{\sigma}\right)^{-1/\gamma} \quad (1)$$

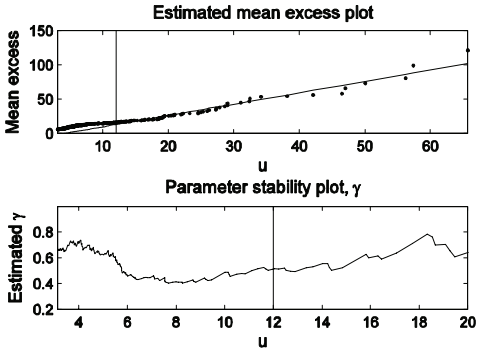
where $\sigma > 0$ and $\gamma \in \mathbf{R}$ are called parameters. This expression is in plain English "the probability that X is larger than u+x given that X is larger than u". The approximation becomes an equality if one increases the threshold. The right hand expression in the formula is the survival function of the generalized Pareto distribution. In the Peaks over threshold method we can use the excess values to estimate the unknown parameters γ and σ. Of course, for different data or extreme phenomena one has different values of u, γ and σ, but still the generalized Pareto distribution. With the estimated parameters we can calculate properties of large values such as quantiles and the probability that a certain event occurs.

We will present two graphical tools for finding the threshold u: Mean-Excess plot and parameter stability plot.

- ***Mean Excess plot:*** By properties of the generalized Pareto distribution one has that if the distribution is valid over a given threshold u_0 then it is also valid over a threshold $u > u_0$. Straight forward calculations give that the expected excess over a given threshold is linear in u. In other words, if we plot the estimated mean excess over a given threshold we should choose u when it behaves linearly (like a straight line).
- ***Parameter stability plot:*** By properties of the generalized Pareto distribution one has that the parameter γ from Equation 1 should be constant in u over a threshold u_0 if the generalized Pareto distribution is valid. Hence, we plot the estimated against u and look for a stable (flat) region. Also, there is a parameter stability plot with a version of σ.

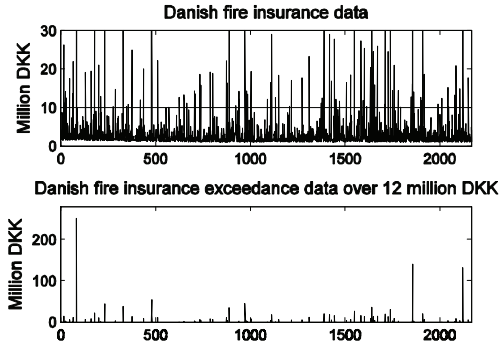
We once again point to the references for a more in-depth description. In Figure 4 we have the mean excess plot and parameter stability plot of $\hat{\gamma}$ for the Danish fire insurance claim data. The hat on top of γ denotes that the parameter is estimated. The estimation method is maximum likelihood.

Figure 4: Top plot: Mean excess plot for danish fire insurance claims. Straight line fitted over threshold 12 million DKK. Bottom plot: Parameter stability plot for estimated ?



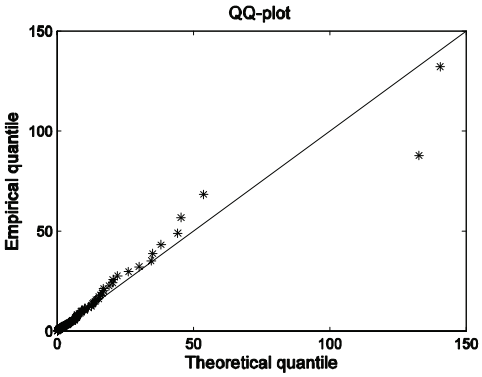
We interpret Figure 4 in the following way: In the bottom plot one can see two stable (flat) regions. The first one is for u=(6,10) and the second is for u=(11,14). However, if we chose the first region then we could not draw a straight line through the data due to the "bump" for u<*10* in the top plot. Hence, as a combination of the mean excess plot and the parameter stability plot the analysis gives us u=12, which has been marked in Figure 4 leading to 85 excesses as a reasonable choice. The excess data can be found in Figure 5. Observe that selecting the threshold is not an exact science and could easily lead to different choices if different people conduct the analysis. A sound strategy is to continue the analysis with several different thresholds and see how the main questions, in this case the 0.999-quantile, is affected by the threshold choice.

Figure 5: Top plot: Danish fire insurance claims with threshold 12 million DKK. For reasons of presentation values over 30 million DKK are truncated. Bottom plot: Excesses over the threshold.



Now, when we have selected the threshold u, it is straight forward to estimate the unknown parameters γ and σ from the exceed data, using a statistical program such as R, Splus or Matlab. Details on estimation can be found in the references. The estimated parameters are $\hat{\gamma} = 0.52$ and $\hat{\sigma} = 7.55$ million DKK with 95 % confidence interval (0.27,0.9) for γ respectively (5.22,10.66) for σ. The confidence intervals are calculated via profile likelihood which is described in Coles (2001). Also, with a qq-plot one can see how well the model fits. This is illustrated in Figure 6. A qq-plot compares the theoretical model, i.e. the generalized Pareto distribution, with the observed data. If the dots are close to the straight line then the model is acceptable which in this case it seems to be.

Figure 6: QQ-plot of fitted model of the Danish fire insurance claims.

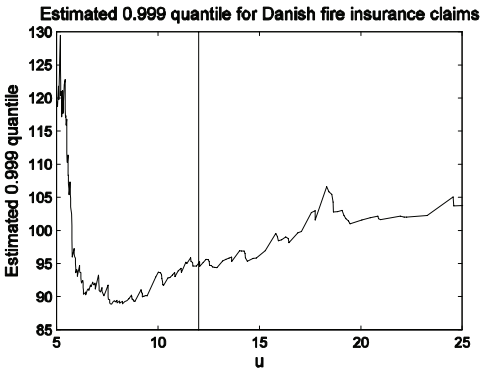


Now we are prepared to find the 0.999-quantile. Using straight forward calculations from Equation 1, noting that a p-quantile occurs with probability p, i.e. $P[X > x_p] = 1\text{-}p$, we end up with

$$x_p = u + \frac{\sigma}{\gamma}\left(\left(\frac{1-p}{P[X>u]}\right)^{-\gamma} - 1\right)$$

where x_p is the value of the p-quantile. We can estimate $P[X > u]$ as the proportion of data over the threshold u, 0.05 in our case. In Figure 7 we have depicted the 0.999-quantile as a function of the threshold.

Figure 7: Estimated 0.999 quantile of the Danish fire insurance claims.



The quantile seems stable in the neighbourhood of the selected threshold u=12 which in addition to the mean-excess plot and parameter stability plot indicates a good threshold selection. Using this threshold we conclude that 99.9 % of the claims is estimated to be smaller than $x_{0.999} = 95.6$ million DKK, with 95 % profile likelihood confidence interval (64,203) million DKK. The confidence interval is quite wide and could be lowered with a larger data set. This could be done by using a longer time interval than 1980-1990 or that insurance companies could share information about extreme events. However, the nature of extreme events is volatile and uncertainty is large. It is of course tempting to estimate higher quantiles and this is indeed possible but should be handled with extreme care. The estimated quantile could be used in risk management as a risk measure of the insurance portfolio and/or helping a insurance company selecting its reinsurance level.

Extensions

In the sections above the analysis has been conducted with observations which are independent and identically distributed in one dimension, i.e. the Danish fire insurance claims occur individually and behave in the same way. One can say that this is a naive way to model real phenomena. However, there exist well understood theory and statistical methodology for multivariate, time-dependent and non-stationary data but this lies outside the scope of this text. Examples of multivariate data are modelling of high wind speed and large amounts of rain which tend to be dependent and the empirical fact that in a stock crash the stocks seem to fall together. Time-dependent data are for instance stock data where a large absolute movement often is followed by another large absolute movement. Non-stationary data is for example data marked by seasonality such as temperature and of course phenomena changed due to a global environmental effect. For an interesting introduction and in-depth reading on the subject the references are a good starting point.

References

Coles, S. G. (2001) *An Introduction to Statistical Modeling of Extreme Values*. Springer, London.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997) *Modelling Extremal Events for Insurance and Finance.* Springer, Berlin.