An Introduction to Extreme Value Theory

Michael P. Mersic

February 24, 2007

1 What is Extreme Value Theory?

What happened on February 1, 1953? There was a severe storm in the Netherlands [2]. This caused several dykes to fail. The flooding killed over 1800 people. After this natural disaster, the Dutch government formed a Delta Committee. The purpose of the Delta-Committee was, among other things, to determine how high the sea dykes need to be to protect against a 1 in 10,000 year storm. The obvious problem with this is that there is only one hundred or so years of historical data.

Traditionally, statistics looks at the mean values of a data set. The Central Limit Theorem states that sample means will be normally distributed around the population mean [4]. But, for many problems, the mean doesn't matter. What matters is the max event or the min event. Events like:

- The strength of a 1 in 10,000 year storm.
- The probability the stock market will crash tomorrow.
- The premium to charge for insurance over a large threshold.

There are two questions that I will ask in this paper. The first is, "What are the Extreme Value distributions?" This is the Extremal Limit problem [1]. The other question is, "To what distributions does EVT apply?" This is the Domain of Attraction problem [1].

Before answering these questions I will discuss some preliminaries, including Quantile-Quantile plots, and a more formal formulation of the Extremal Limit Problem and Domain of Attraction problem.

2 Comparing distributions with QQ-Plots

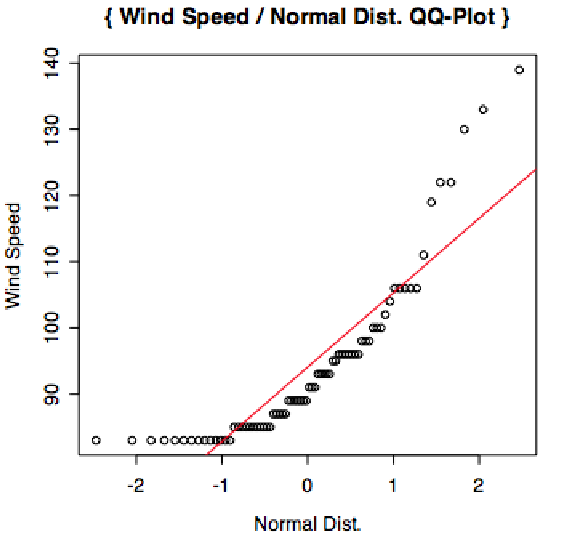
Before getting into EVT, I want to describe a method of comparing distributions called a Quantile-Quantile Plot, or QQ-Plot. A QQ-Plot allows us to quickly see if two sets of data seem to come from the same distribution [1]. It is important to note that it is not possible to prove that two sets of data came from the same distribution, but with a QQ-Plot we can see if it is a likely hypothesis.

Lets take a set of independent and identically distributed (iid) data $X_1, X_2, .., X_n$. The data comes from the distribution: $F(x) = P(X > x)$. How can we tell if this distribution is like the normal distribution? One way is to take the Quantiles of the two distributions and plot them. The Quantile is essentially the inverse of $F(x)$,

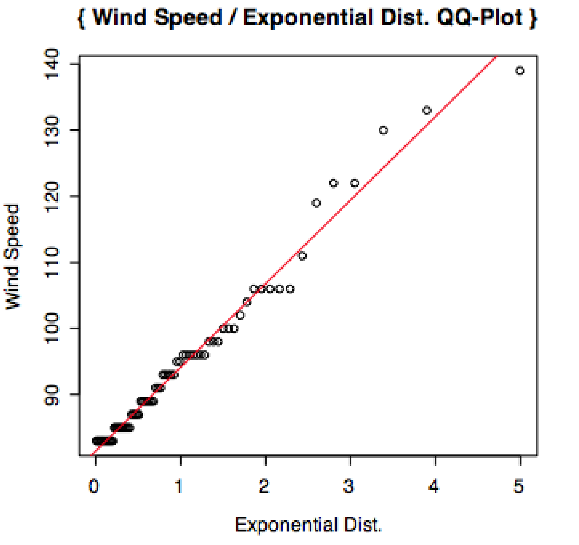
$Q(p) = inf\{x : F(x) \geq p\}$.

Given a set of data, $X_1, X_2, .., X_n$, from $F(x)$, is the data normally distributed? To answer this question with a QQ-Plot we will first find the probability points, $p_1 = F(X_1), p_2 = F(X_2), ..., p_n = F(X_n)$. Then find the Quantiles of each probability point from the normal distribution $Y_1 = Q_N(p_1), Y_2 = Q_N(p_2), ..., Y_n = Q_N(p_n)$. Finally, plot each point $(Y_1, X_1), (Y_2, X_2), ..., (Y_n, X_n)$. If there is a relationship in the data, the points will fall close to the 45 degree line in the plot.

As an example I will plot the daily maximum wind speed, for days with maximum wind speed above 82-km/hr for Zeventem Belgium. This example is taken from [1]. In the next two graphs the Wind Speed data is plotted against the Normal Distribution and against the Exponential Distribution. First, the Normal Distribution with a least squares regression line:



Here the plot shows a very poor fit for the Wind Speed data against the Normal Distribution. Now the Exponential Distribution with a least squares regression line:



This is a much better fit.

The linearity of the data in these graphs can be measured by the correlation coefficient, r_Q. The correlation coefficient is bounded, $0 \leq r_Q \leq 1$. If the data is perfectly linear, then $r_Q = 1$. In the above examples, r_Q for Wind Speed against the Normal Distribution is 0.8969 and r_Q for Wind Speed against the Exponential Distribution is 0.9912. As expected, r_Q for the Exponential Distribution is much closer to 1 then r_Q for the Normal Distribution.

3 Formulation of the Extremal Value Problem

3.1 Maximum and Minimum are the same problem

It is easy to show that the maximum and the minimum distribution problems are the same problem [3]. Consider the maximum value of a random sample of iid variables, $X_{n,n} =$

$max\{X_1, X_2, ..., X_n\}$. The minimum value is, $X_{1,n} = -max\{-X_1, -X_2, ..., -X_n\}$.

3.2 The limiting distribution of $F^n(x)$ is degenerate

Now I will show that $F^n(x)$ is degenerate and needs to be normalized [3]. Let F be the underlying distribution. Let it's right end point (maximum value) be x^*. Then:

$$max(X_1, X_2, ..., X_n) \rightarrow x^* \text{as } n \rightarrow \infty.$$

and

$$\begin{aligned} &P(max(X_1, ..., X_n) \leq x) \\ &= P(X_1 \leq x, X_2 \leq x, ..., X_n \leq x) \\ &= F^n(x) \end{aligned}$$

so if $x < x^*$ then

$$F^n(x) \rightarrow 0 \text{ as } n \rightarrow \infty$$

and if $x \geq x^*$ then

$$F^n(x) \rightarrow 1 \text{ as } n \rightarrow \infty$$

Therefore the limiting distribution $\lim_{n\rightarrow\infty} F^n(x)$ is degenerate.

To deal with this we need to normalize $F^n(x)$. Suppose there exists a sequence of constants $a_n > 0$ and b_n real such that:

$$P(\frac{max(X_1, X_2, ..., X_n) - b_n}{a_n} \leq x) = F^n(a_n x + b_n)$$

then

$$\lim_{n\rightarrow\infty} F^n(a_n x + b_n) = G(x)$$

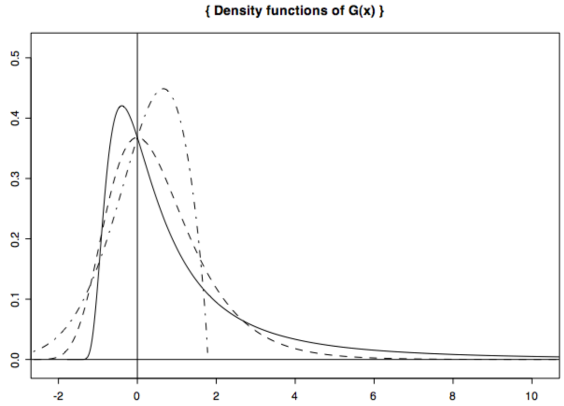
Finding the limiting distributions $G(x)$ is called the Extremal Limit Problem. Finding the $F(x)$ that have sequences of constants as described above leading to $G(x)$ is called the Domain of Attraction Problem.

3.3 Extremal Limits

The class of Extreme Distributions are described by [1]:

$$\begin{aligned} G_\gamma(x) &= exp(-(1+\gamma x)^{\frac{-1}{\gamma}}) && (1) \\ &\quad \text{where } 1+\gamma x > 0 \\ G_\gamma(x) &= exp(-e^{-x}) && (2) \\ &\quad \text{where } \gamma = 0 \end{aligned}$$

When $\gamma = 0$, $G(x)_{\gamma=0}$ is called the Gumbel Distribution. It's density function is $f(x) = e^{-x}e^{-e^{x}}$ and it looks like this:



The solid line is with $\gamma = 0.56$. The dashed line is with $\gamma = 0$. The dash-dotted line is with $\gamma = -0.56$.

The range of $G(x)_\gamma$ depends on the value of γ:

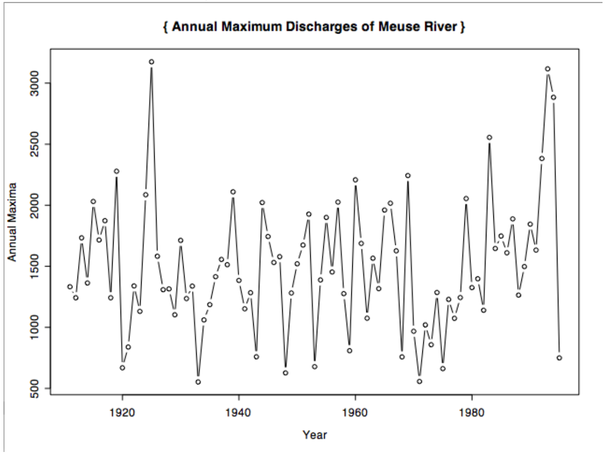
For $\gamma > 0$	the range is	$\frac{-1}{\gamma} < x < \infty$
For $\gamma < 0$	the range is	$-\infty < x < \frac{-1}{\gamma}$
For $\gamma = 0$	the range is	$-\infty < x < \infty$

Now, I'll finish with an example.

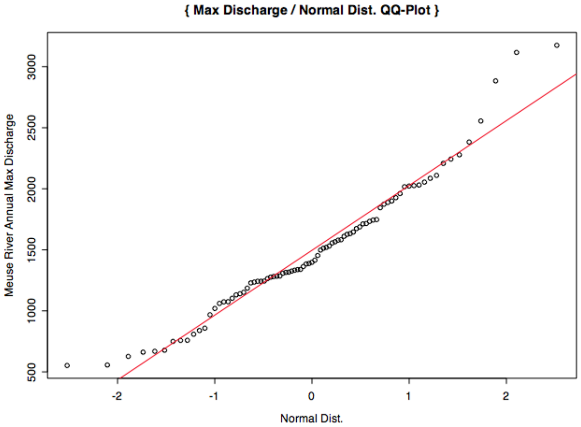
4 Meuse river example

The Meuse river runs from France through Belgium and the Netherlands into the North Sea. In this section I will look at the annual maximum discharge data. The data for this example, and the optimal value of γ was taken from [1]. First I will compare the empirical distribution with the Normal Distribution, then with the Gumbel Distribution. Finally, using the Generalized Extreme Value Distribution developed in the last section I will provide an optimal Distribution (by maximizing correlation) and then estimate a 1 in 100 year maximum discharge.

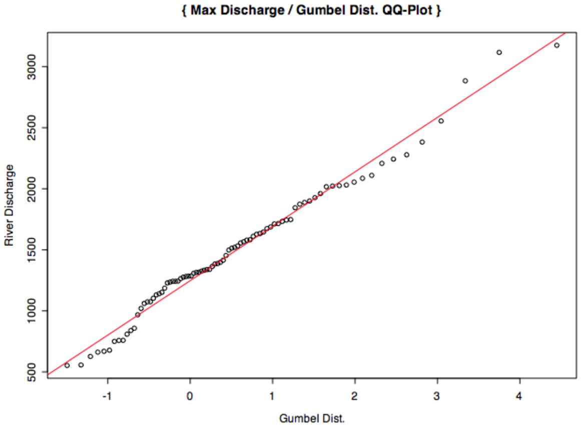
The data set is of yearly maximum discharges from 1911 to 1995.



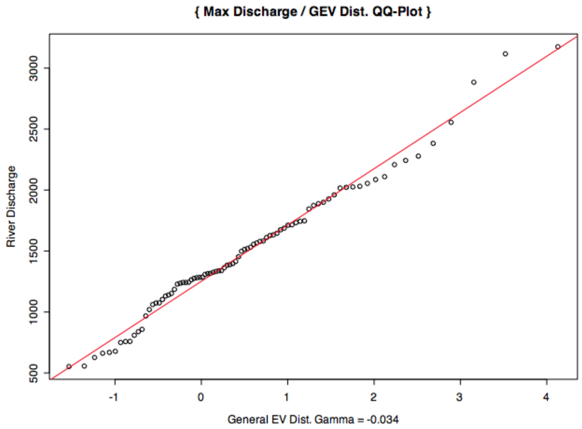
Here is a comparison of the empirical distribution with the Normal Distribution. Significantly, one should note that if the least-squares regression line is used as a model for the actual maximum distribution, the maximum discharge will be underestimated.



Now, a comparison of the empirical data with the Gumbel Distribution, that is $\gamma = 0$. Visually the least-squares regression line is a much better fit. It does not seem to underestimate the maximum values like the Normal Distribution does.



Finally if the Extremal Value Index, that is γ, is chosen to maximize the correlation between the empirical data and the Extreme Value Distribution, γ is -0.034.



To estimate the one in one hundred year maximum discharge, let $y = 100$. We want to find the value of the regression line where $x = Q(1 - \frac{1}{y})$. Or, more commonly, $x = U(y)$, where $U(y) = Q(1 - \frac{1}{y})$. The value $U(100)$ for the Normal, Gumbel, and $\gamma = -0.034$ Distributions is:

$$\begin{aligned} U(100)_N &= 2.3263 \\ U(100)_{\gamma=0} &= 4.60014 \\ U(100)_{\gamma=-0.034} &= 4.25845 \end{aligned}$$

Then inserting these $U(100)$ values into their respective regression lines, we get the following estimates for the one in one hundred year maximum discharge.

$$\begin{aligned} R_N(U(100)_N) &= 509(2.3263) \\ &\quad +1467 = 2651 \\ R_{\gamma=0}(U(100)_{\gamma=0}) &= 462(4.60014) \\ &\quad +1252 = 3377 \\ R_{\gamma=-0.034}(U(100)_{\gamma=-0.034}) &= 434(4.25845) \\ &\quad +1259 = 3107 \end{aligned}$$

As noted above, the Normal Distribution significantly underestimates the one in one hundred year maximum discharge. Although the Gumbel Distribution gives the highest estimate, it is important not to overestimate the maximum value. If the actual risk is overestimated then a higher cost is implied to compensate for that risk.

References

[1] Jan Beirlant, Yuri Goegebeur, Jozef Teugels, and Johan Segers. *Statistics of Extremes.* Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chicheser, West Sussex PO19 8SQ, England, 2004.

[2] Bassi F., Embrechts P., and Kafetzaki M. Risk management and quantile estimation. In R.J. Adler et al, editor, *A Practical Guide to Heavy Tails.* Birkhaeuser, Boston, 1998.

[3] Laurens de Haan and Ana Ferreira. *Extreme value theory : an introduction.* Springer Series in Operations Research and Financial Engineering. Springer, New York ; London, 2006 edition, 2006.

[4] R. Lyman Ott and Michael T. Longnecker. *A First Course in Statistical Methods.* Brooks/Cole, Belmont, CA, 2004.