

# **EXTREME VALUE STATISTICS**

## **by Dougal Goodman and James Orr**

### **Beyond the Limits of Experience**

When Roger Bannister ran the first “sub” four-minute mile in 1954, he showed how athletes excel through ever-greater feats of speed and endurance. By following a pacemaker in the early laps and with benefits of improved training techniques, he was able to run faster than anyone had done before. However, others had the same benefits and had failed to break the magic barrier. Was this a truly exceptional performance?

Most of Holland lies below mean sea-level, with some 20% of its land area reclaimed from the sea. Continued economic prosperity and safety rest on building adequate flood defences, to ensure that flooding is a rare, or practically impossible, event. Accurate tidal records are only available for some 120 years. What height of defences will provide protection up to a 1000-year return period (roughly a 0.1% annual probability)? The answer must be accurate, as high floodwalls are expensive and unsightly.

These two problems are linked by a mathematical theory. Extreme value statistics is the study of rare events that lie beyond common experience. Extreme value statistics can be applied to extremes in many fields, including nature, engineering, sport and economics.

### **A Break from “The Norm”**

For many years, actuarial science has enjoyed increasing influence in the pricing and reserving of general insurance business. Much of this success is built on the Central Limit Theorem (CLT), which shows how long-term average values (expected values) behave for large numbers of observations from a random variable.

Actuaries have helped companies to fine-tune their pricing through multivariate analysis, which shows how underwriting factors (e.g. the age of a driver, the make or colour of a car) affect long-term average values. However, a new challenge and opportunity exists for actuaries, to gain influence in a new territory – where hurricanes, earthquakes and stock market “crashes” roam.

Companies in all industries are concerned about large losses that might reduce profits or threaten solvency. The banking world is using “Value at Risk” techniques to determine conditional expected values and answer questions like “How much will we lose, if we have a big loss?” Moreover, insurers are increasingly relying on excess of loss reinsurance, which provides protection specifically against exceptional losses.

However, how can you “fine-tune” a catastrophe? To sweeten the challenge, there isn’t much data; by definition, rare events do not appear very often. What if the physical processes underlying catastrophes are fundamentally different from other, less extreme, events? Extrapolating out of the observed data must be dangerous? It is, but the conceptual and theoretical structures of Extreme Value Theory (EVT) allow us to extract the maximum value from the data that we do have.

# **EXTREME VALUE STATISTICS**

## **by Dougal Goodman and James Orr**

### **The Distribution of “Maxima”**

Much in the same way that CLT helps us to estimate the distribution of average values, EVT tells us that the distribution of observed maxima (each from a large number of observations of a random variable) will approximate to the Generalised Extreme Value (GEV) distribution. For example, annual maximum sea-level data can be analysed with this model.

The GEV has three parameters, for location, scale and “shape”. The shape parameter plays a critical role in determining whether the tail of the distribution is finite (when the parameter is negative) or infinite (when it is non-negative). Indeed, the GEV encompasses three classes of distribution: Gumbel (zero shape parameter), Frechet (positive), and Weibull (negative), which may be familiar to some readers. Proving that the best estimate of the shape parameter for a set of data is negative can provide strong evidence for a limiting endpoint in a distribution.

However, where the evidence suggests otherwise (that infinite values can occur) a “reality check” is always worthwhile, to consider the underlying physical processes. The maximum possible value may be limited by physical or financial factors, such as the largest possible wave height or insurance claim from the total destruction of a property. If this is not reflected in the data, then even the most sophisticated mathematical methods will benefit from some practical input.

The GEV has been used extensively in assessing flood defence requirements, but it can only make use of annual maximum data. Other data, near to the maximum value, must be relevant in fitting the tail of a statistical distribution? Fortunately, statistical techniques that use more of the extreme data have been developed. These tools use data above a threshold.

### **The Distribution of “Exceedance Points”**

It can be shown that, for a sufficiently large threshold, data above that point will follow the Generalised Pareto Distribution (GPD). Again, the GPD has three parameters, including a shape parameter, scale and the “threshold” value itself. With GPD, it is possible to use far more of the sea-level data to assess flood defence requirements, rather than simply annual maxima. This greatly improves the credibility and accuracy of the results.

Where the GPD applies, the average of data values in excess of the threshold (“exceedances”) will be a linear function of the threshold itself. This provides a useful diagnostic tool, in plotting “mean exceedance” against “threshold”. Using this, the appropriateness of the GPD can be assessed and a suitable threshold point can be estimated. By deciding where the linear relationship breaks down, it is possible to set the threshold and partition the data, to fit GPD to the tail and an empirical or other traditional distribution to the lower values.

Having fitted a statistical distribution to the data, we can estimate extreme points (such as the 0.1% percentile) and answer questions about conditional expectations (to answer “Value at Risk” questions). However, two major issues remain, in potential time dependencies in the data and the uncertainty in our estimates.

# EXTREME VALUE STATISTICS

by Dougal Goodman and James Orr

## Changing Times and Uncertainty

Developments over time, such as improved diet, training techniques and equipment, have drastically changed middle-distance running standards. Sub 4-minute miles are now commonplace and world-class times are some fifteen seconds faster! Clearly, performance data from previous decades is not relevant to assessing performance in the present time. “Non stationary” analytical techniques, that can identify and model time dependency like this, are also being supported and promoted by TSUNAMI, but are not the subject of this article.

The second issue, of measuring the uncertainty in our estimates, can be elegantly addressed through Bayesian statistics. Bayes’ Theorem tells us that:

$$P(B|A) = P(A|B) \times P(B) / P(A),$$

where  $P(A)$  is the probability of “A” and  $P(A|B)$  is the probability of “A given B”.

By substituting “Data” for “A” and “Parameters” for “B”, we can determine the (posterior) probability distribution of the parameters given the data, namely  $P(B|A)$ . Using the “likelihood”  $P(A|B)$  we can readily place credible intervals around the parameter values. Knowledge of  $P(B)$  and  $P(A)$  is not required to achieve this. These limits then allow us to assess the accuracy and decide on the credibility of the estimates that we obtain for high percentile points and conditional expectations.

## Who’s Who

EVT has been adopted and developed by a number of researchers. Principal exponents include Professor Richard Smith of the University of North Carolina, who developed the “exceedance over threshold” techniques, and Professor Paul Embrechts of ETH in Zurich, who has published extensively on financial modelling applications.

Paul Embrechts co-authored *Modelling Extremal Events for Insurance and Finance*, which provides a comprehensive guide to practitioners. Alexander McNeil, who works with Paul Embrechts, has also produced a number of routines for the “S-Plus” statistical package, which supports extreme value analysis. These routines are available, free of charge, at [www.math.ethz.ch/~mcneil/softward.html](http://www.math.ethz.ch/~mcneil/softward.html).

Dr Mark Dixon of City University has run a useful course for newcomers, in June of this year, and is planning a repeat programme in September. This course includes routines in the freeware “R” statistical package, for tuition in and evaluation of the methods. Further details can be obtained from Mark at [m.j.dixon@city.ac.uk](mailto:m.j.dixon@city.ac.uk).

The Research Committee of the Faculty and Institute of Actuaries recognised the importance of extreme value techniques some time ago. They have supported a programme of research into pricing excess of loss reinsurance contracts against catastrophic losses using EVT, by a researcher based in Georgia State University. A paper from this research is expected soon and should shortly be published in *The North American Actuarial Journal*.

**EXTREME VALUE STATISTICS**  
**by Dougal Goodman and James Orr**

**TSUNAMI**

TSUNAMI is committed to promoting extreme value statistical techniques, following its sponsorship of a six-month programme, last year, at the Isaac Newton Institute for Mathematical Sciences on non linear and non stationary statistics. Full references to related publications are on our Web Site at [www.nerc-bas.ac.uk/public/tsunami](http://www.nerc-bas.ac.uk/public/tsunami) under "Funded Projects - Extreme Value Statistics".

-/-

*References:*

*"Bayesian Risk Analysis" by R.L. Smith and D.J. Goodman. This paper discusses how the application of Bayesian methods using Monte Carlo Markov chains provides estimates of the distributions of parameters in the Generalised Pareto Distribution. To be published shortly.*

*"Modelling Extremal Events" by Paul Embrechts, Claudia Kluppelberg and Thomas Mikosch (1997). Published by Springer, ISBN 3-540-60931-8.*

*"Models for Exceedance over High Thresholds (with discussion)" by A.C. Davison and R.L. Smith. The Journal of the Royal Statistical Society, Series B, volume 52, pages 342-442.*

*"Statistics for Exceptional Athletics Records" by R.L. Smith (1997). Letter to the editor, Applied Statistics, volume 46, pages 123-127.*