NOTEBOOK

Extreme value distributions

Emma Eastoe introduces a family of distributions that help statisticians model the extreme highs and lows of the natural world, and much more besides

What are extreme value distributions?

Extreme value distributions form the foundation of extreme value theory, a branch of statistics used to model the unusually high, or low, values in a data set. Examples of the application of extreme value analysis abound within both the environmental sciences and finance. For example, ocean engineers are interested in the largest wave heights in order to decide what safety constraints need to be put in place to protect structures such as oil rigs against the most devastating storms. In flood risk management, hydrologists need to produce maps of the highest river levels that can be expected within a given time period. Other areas of application include climate change, wind speeds, air pollution levels, temperatures and precipitation.

There are three types – or shapes – of extreme value distribution, which are collectively represented by a single formula known as the generalised extreme value (GEV) distribution. The shape determines the rate of decay of the upper tail of the distribution. This rate of decay tells us (i) how quickly the largest values of the distribution get larger, and (ii) whether they can get infinitely large (non-negative shape), or if there is an upper bound on how large they can be (negative shape). In addition to the shape parameter, the GEV distribution is usually also described by location and scale parameters.

Figure 1(a) shows a plot of the density functions of three GEV distributions, each of which has a different shape parameter: −0.2, 0 and 0.2. The difference between these distributions is more clearly seen when examining large quantiles as in Figure 1(b). From this plot there is a clear difference in the rate of decay of the three distributions; in particular, the quantiles of the distribution with the negative shape parameter tend to a finite upper bound (−5 in this case).

The GEV distribution is not chosen by chance as an appropriate model for the largest values of a data set. In fact its usefulness arises as a result of it having a property known as "max-stability". That is, the maximum of any (fixed) number of variables with common GEV distribution will also have a GEV distribution.

Emma Eastoe is a lecturer in statistics at Lancaster University

Who discovered it?

Fisher and Tippett derived the three basic shapes represented by the GEV distribution; these are known as the Fréchet (positive shape), Gumbel (zero shape) and negative Weibull (negative shape).[1] These three types were joined into a single family of distributions by von Mises[2] and Jenkinson.[3] The statistical advantage of using the generalised form of the distribution is that the user does not need to decide in advance which of the three types to use as their model; instead the shape parameter can be estimated as a model parameter.

When should it be used?

In extreme value analysis, the distribution can be used as a model

Main image: lightkite/Bigstock.com

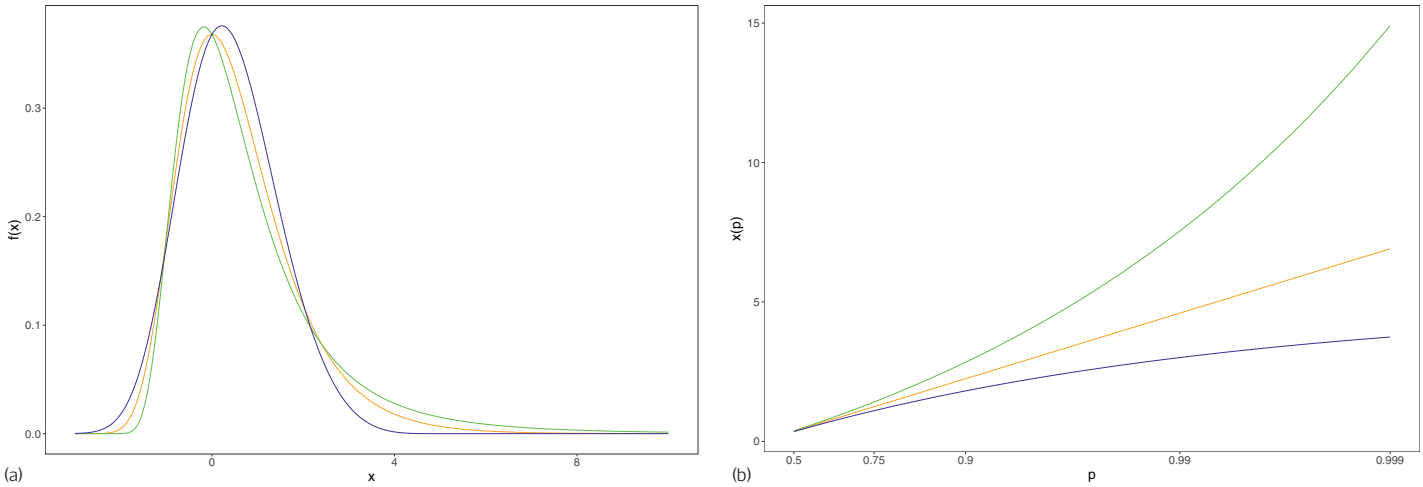

FIGURE 1 (a) Density and (b) quantile plots for the standard GEV distribution with shape parameter −0.2 (blue line), 0 (orange line) and 0.2 (green line)

for any extreme data set comprised of what are called "block maxima" – these are found by dividing the data set of interest into equal, non-overlapping segments, and noting the maximum (or minimum) data-point in each – for example, the minimum temperature each year, or the annual maximum wave height. When fitting the model, it is important to ensure that you have a large enough sample of maxima – 30 years or more would be ideal. Care must also be taken that the series of observations over which the maxima are taken is sufficiently long: for example, modelling the annual maximum of 365 daily observations would be sensible, but modelling the daily maximum of hourly observations would not.

As with all statistical models, the model fit should be assessed. Once this is done, the model is often used to make predictions of the block maxima return levels. For example, we might calculate the 500-year return level of the annual maximum wave height to give an estimate of the wave height exceeded, on average, once every 500 years. The choice of return period, in this case 500 years, is usually defined in advance of the analysis by the end-user, for example the ocean engineer or hydrologist.

When should it not be used?

It is generally only advisable to use the GEV distribution when an extreme value analysis is required and interest is in the block maxima. Often, interest is not so much in the block maxima – the annual maximum wave height, say – but in the extreme behaviour of the underlying process – three-hourly wave heights, for example. In these circumstances, it is much more efficient to model all extreme events rather than taking block maxima and discarding the rest of the data set. An appropriate distribution to model the sizes of all extreme events would be the generalised Pareto distribution (to be described later in the *Notebook* series).

Keep in mind...

The GEV distribution is a good model for the unusually high, or low, values of a data set, when the data set takes the form of block maxima. It is commonly used in environmental and financial applications. It is a very flexible distribution and can accommodate numerous data sets which have very different extreme behaviour. In particular, it is appropriate both for data sets that have no finite bound on the largest values that could be seen and for those that do, and can be used to estimate return levels in both cases.

However, we have made an implicit assumption that the phenomena being modelled are stationary and homogeneous, and this is not often the case. The effects of climate change, for example, can result in many physical processes having annual maxima/minima which vary over time: an example would be the annual minimum sea ice extent in the Arctic, which is gradually declining. In addition, processes such as surface-level air pollution, which are often modelled spatially, are unlikely to have annual maxima/minima which are the same across different locations. For example, much air pollution comes from sources such as roads or industry which are found in urban rather than rural areas. Care must always be taken to account for differences such as these, often by appropriate modelling of the GEV parameters using regression or random effects models, among others. ■

A note about *Notebook*

This article is the second in a new regular series on statistical distributions. Which do you want to read about next? Send suggestions – or a submission of your own – to significance@rss.org.uk

References

1. Fisher, R. A. and Tippett, L. H. C. (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, **24**, 180–190.

2. von Mises, R. (1964) La distribution de la plus grande de *n* valeurs. In Ph. Frank, S. Goldstein, M. Kac, W. Prager, G. Szegö and G. Birkhoff (eds), *Selected Papers of Richard von Mises: Volume II. Probability and Statistics, General* (pp. 271–294). Providence, RI: American Mathematical Society.

3. Jenkinson, A. F. (1955) The frequency distribution of the annual maximum (or minimum) values of meteorological events. *Quarterly Journal of the Royal Meteorological Society*, **81**, 158–172.