

# Learning about Extremes

---

## KEYWORDS:

Teaching;  
Probability plot;  
Simulation

Stuart G Coles

Department of Mathematics, University of Nottingham

## Summary

For many physical processes it is extreme levels which are of greatest concern. This article gives a practical introduction to the problems and models involved.

---

## ◆INTRODUCTION◆

---

CONSIDER a random sample  $X_1, \dots, X_n$  taken from a population with common distribution function  $F$ . What is the distribution of the sample mean,  $\bar{X}$ . No problem. Any A-level student will know that provided  $n$  is sufficiently large, and under very broad regularity conditions,  $\bar{X}$  is well-approximated by a Normal distribution, regardless of the population distribution  $F$ . But how about the distribution of  $X_{\max} = \max(X_1, \dots, X_n)$ , the maximum value within the sample? Can anything be said about this distribution if  $F$  is unknown? There are a number of reasons why the investigation of this problem is a useful exercise for students undertaking a first course in statistics. Firstly, the problem can be approached in a purely practical way. This involves the use of a variety of statistical skills: simulation, exploratory data analysis and probability plots, for example. Furthermore, the distribution of  $X_{\max}$  is required in many practical situations, particularly in relation to the environment. For instance, sea-defences must be designed to withstand the most extreme sea-level likely to be experienced at that location. Taking the  $X_i$  to be hourly observations of sea-level, and  $n$  to be the number of observations in a year, design criteria for sea-walls are generally based on an estimate of the distribution of  $X_{\max}$ , the annual maximum sea-level. Other applications include the study of extreme temperatures, high pollution levels and heavy rainfall. In each case it is the behaviour of the most extreme value over a year, say, which is of most concern.

## ◆EXACT DISTRIBUTION OF $X_{\max}$ ◆

---

It is easy, in principle, to write down an expression for the distribution of  $X_{\max}$  in terms of the popula-

tion distribution function  $F$ :

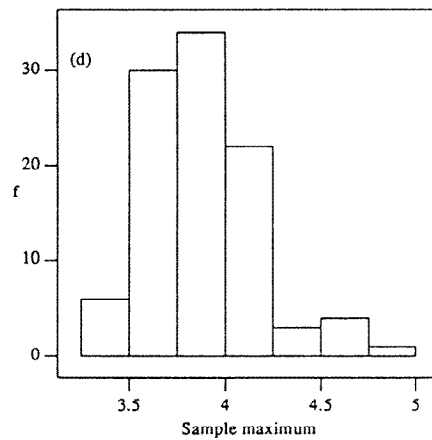
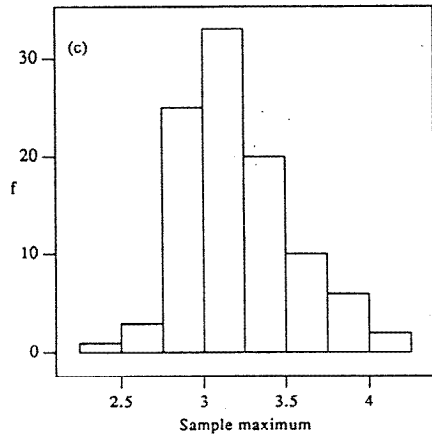
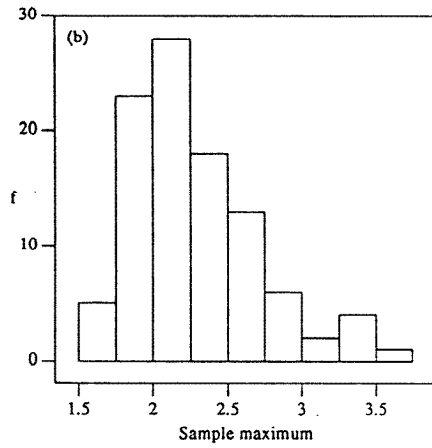
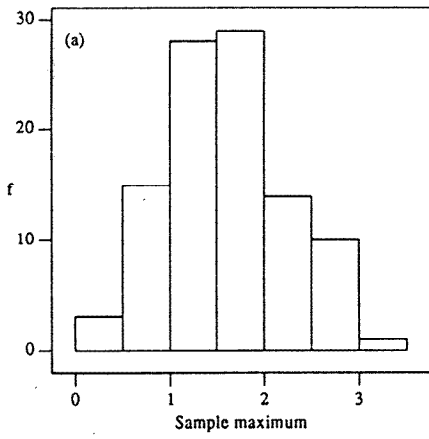
$$\begin{aligned} Pr(X_{\max} \leq x) &= Pr(X_1 \leq x, \dots, X_n \leq x) \\ &= Pr(X_1 \leq x) \dots Pr(X_n \leq x) \\ &= \{F(x)\}^n \end{aligned} \quad (1)$$

This aspect of the investigation is a useful topic for discussion with students in itself.

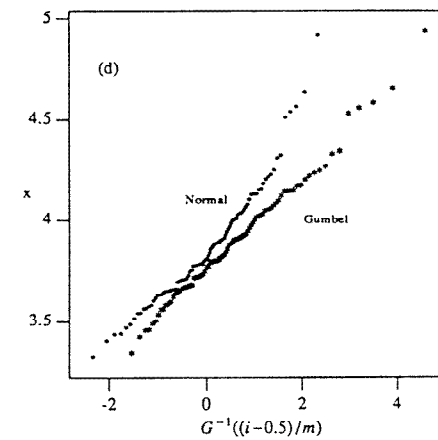
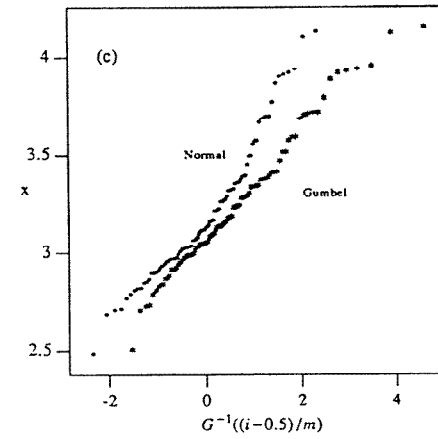
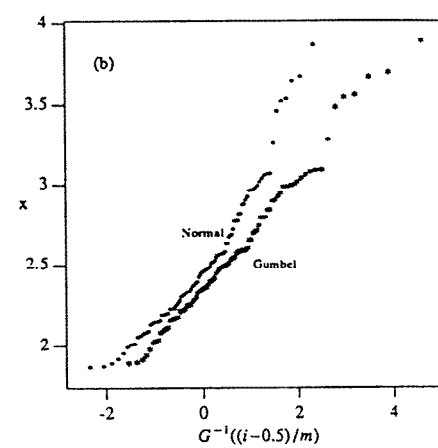
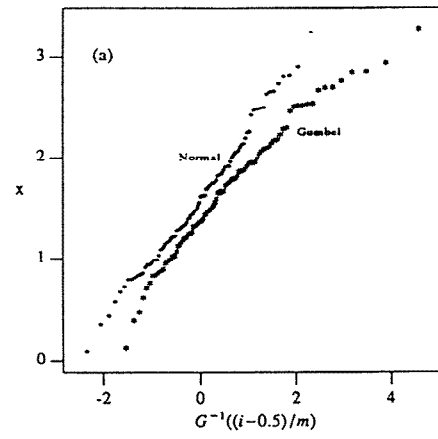
For illustration, Figure 1 gives histograms of  $m = 100$  values of  $x_{\max}$  for samples of various sizes simulated from a standard Normal distribution. For a sample size of  $n = 10$  there is little to suggest that the distribution of  $X_{\max}$  itself is not Normal. But with increasing  $n$  it becomes evident that  $X_{\max}$  is certainly not Normally distributed, but follows a distribution with substantial positive skew. This will hopefully dispel the myth that all continuous distributions in practice, especially those involving limits, are Normal.

Further evidence that the distribution of  $X_{\max}$  is not Normal is obtained by drawing a Normal probability plot. This can be done on specially scaled paper, or equivalently by the following method. If a sample  $x_1, \dots, x_n$  is sorted into ascending order, then a plot of  $x_i$  against  $G^{-1}\{(i-0.5)/m\}$ , where  $G$  is the standard Normal distribution function, should be approximately linear. The rationale behind this is that the empirical distribution function of the data should be comparable to the distribution function itself. Departures from linearity suggest the Normal model is wrong. Furthermore, if the linearity of the plot is acceptable then the intercept and slope of the fitted line give estimates of  $\mu$  and  $\sigma$ , the mean and standard deviation of the Normal distribution respectively.

Figure 2 gives Normal probability plots for each of the simulated sets of  $X_{\max}$  described above. The plot corresponding to  $n = 10$  seems reasonably linear, but for larger  $n$  there is a distinct kink in the plot, which also occurs lower down in the sample for larger sample sizes. This again illustrates that the distribution of  $X_{\max}$  does not tend to Normality as the sample



**Figure 1.** Histograms of maxima of simulated samples from standard Normal distribution. Sample sizes: (a) 10, (b) 100, (c) 1000, (d) 10,000



**Figure 2.** Probability plots of simulated sample maxima assuming Normal and Gumbel distributions. Sample sizes: (a) 10, (b) 100, (c) 1000, (d) 10,000

size increases. Indeed the departure from Normality becomes more marked with increasing  $n$ , corresponding to the distribution becoming more positively skew as the histograms suggested.

---

### ◆THE GUMBEL DISTRIBUTION ◆

---

The non-Normality of the distribution of  $X_{max}$  is not surprising. In fact, careful mathematical treatment of equation (1) suggests that for large  $n$  the distribution function of  $X_{max}$  should be approximately what is known as a Gumbel distribution. This has distribution function

$$G(x) = \exp\{-\exp(-(x-a)/b)\} \quad (2)$$

where  $a$  and  $b$  are location and scale parameters respectively, rather like  $\mu$  and  $\sigma$  for the Normal distribution. Do the simulated data support this argument? This can be assessed by the use of probability plots, also shown on Figure 2, where this time the ordered sample is plotted against  $G^{-1}(i/(m+1)) = -\log(-\log(i/(m+1)))$ , so that the standard Normal distribution function has been replaced with a standard Gumbel distribution function ( $a=0, b=1$  in equation (2)). If the Gumbel distribution is appropriate this plot will be approximately linear, with intercept and slope giving estimates of  $a$  and  $b$  respectively. In this case the linearity of the plot appears to improve with increasing sample size, confirming that the Gumbel distribution is a reasonable approximation to the distribution of  $X_{max}$  for large samples.

One point to note, however, is that although the Gumbel probability plots are reasonably linear, the largest values of  $x_{max}$  do seem to deviate from the line somewhat. This suggests that even for very large samples the Gumbel distribution may not provide very accurate probabilities for the most extreme events. This is a moot point, since it is precisely these events which are of greatest concern in practice.

---

### ◆FURTHER IDEAS ◆

---

The above procedure can easily be modified to investigate other aspects of extremes. Here are a few ideas.

1. Base simulation on a distribution other than the Normal distribution. Does the Gumbel distribution always seem to be a valid approximation to the distribution of  $X_{max}$ ? (A spectacular indication that the answer is no' can be obtained by taking  $F$  to be a Cauchy distribution, given by  $F(x) = 0.5 + (1/\pi)\tan^{-1}(x)$ ,  $-\pi/2 < x < \pi/2$ .

Try it. In fact, the Gumbel distribution is just a special case of a broader distributional family known as the Generalized Extreme Value distribution, which gives all possible limiting forms of  $X_{max}$ ;

2. Use the probability plots to estimate the Gumbel parameters  $a$  and  $b$ . How do these vary with  $n$ ?
3. Most variables in practice have a seasonal effect. Try adding on a seasonal component a sinusoidal term for example to the simulated variables before finding  $x_{max}$ . What effect does this have on the distribution of  $X_{max}$ ?
4. Modify the above procedure to investigate the distribution of  $X_{min}$  the minimum value of the sample. This problem also has practical motivation: minimum temperature, reservoir level etc.

Finally, the study of extremes is becoming increasingly important. There is, for example, current concern that the greenhouse effect may lead to a rise in mean sea-level. This may be serious, but if climatic change were also to substantially affect the behaviour of extreme sea-levels the consequences could be catastrophic. Statistics clearly has an important role in identifying any change in the behaviour of  $X_{max}$ : the techniques and models described above comprise the first step in this process.