CHAPTER 1

INTRODUCTION

1.1 Preface

Nowadays, statistics admittedly holds an important place in all the fields of our lives. Almost everything is quantified and, most often, averaged. Indeed, 'averaging' is the statistical notion most easily understood and most widely used. On the other hand, not few are the cases where the central tendency, as this is captured by an average measure, is not suitable to 'statistically' describe the situations and their impacts.

Two important dates in the history of risk management are February 1, 1953 and January 28, 1986. During the night of February 1, 1953 at various locations the sea-dykes in the Netherlands collapsed during a severe storm, causing major flooding in large parts of coastal Holland and killing over 1800 people. The second date corresponds to the explosion of the space shuttle Challenger. According to Dalal et al. (1989), a probable cause was the insufficient functioning of the so-called O-rings due to the exceptionally low temperature the night before launching. In both of these cases, an extremal event caused a protective system to breakdown. One could argue that statistical concepts, such as averages, also 'broke-down', since not only they offer no help but they can also be misleading, if used. In such cases, it is the examination of extremes that provides us with insight of the situations.

Natural or man-made disasters, crashes on the stock market or other extremal events form part of society. The systematic study of extremes may be useful in contributing towards a scientific explanation of these. As will be made clearer in the sequel, analysis of extreme values is an aspect of statistical science that has much to offer to many fields of human activity.

1.2 Genesis and Historical Development

The extreme value theory is a blend of a variety of applications concerning natural phenomena such as rainfall, floods, wind gusts, air pollution and corrosion and sophisticated mathematical results on point processes and regular varying functions. So, engineers and hydrologists on the one hand and theoretical probabilists on the other, were the first to be interested in the development of extreme value theory. It is only recently that extreme value theory attracted mainstream statisticians. Indeed, the founders of probability and statistical theory (Laplace, Pascal, Fermat, Gauss, et al.) were too occupied with the general behaviour of statistical masses to be interested in rare extreme values.

Historically, work on extreme value problems may be dated back to as early as 1709 when N. Bernoulli discussed the mean largest distance from the origin when n points lie at random on a straight line of length t (Johnson et al., 1995). A century later Fourier stated that, in the Gaussian case, the probability of a deviation being more than three times the square root of two standard deviations from the mean is about 1 in 50,000, and consequently could be omitted (Kinnison, 1985). This seems to be the origin of the common, though erroneous, statistical 'rule' that plus or minus three standard deviations from the mean can be regarded as the maximum range of valid sample values from a Gaussian distribution.

The first to investigate extreme value statistics were early astronomers who were faced with the problem of utilizing or rejecting 'suspect' observations that appeared to differ greatly from the rest of a data-set. Still, systematic study and exploration of extreme value theory started in Germany in 1922. At that time a paper by Bortkiewicz (1922) appeared which dealt with the distribution of the range of random samples from the Gaussian distribution. The contribution of Botkiewicz is that he was the one to introduce the concept of '*distribution of largest values*'. A year later another German, von Mises, introduced the concept of 'expected value of the largest member of a sample of observations from the Gaussian distribution' (Mises, 1923). Essentially, he initiated the study of the asymptotic distribution of extreme values in samples from Gaussian distribution. At the same time, Dodd (1923) studie0d largest values from distributions other than the normal.

Still, the fathers of extreme value theory are Tippet and Fisher. Indeed, a major first step occurred in 1925, when Tippet presented tables of the largest values and corresponding probabilities for various sample sizes from a Gaussian distribution, as well as the mean range of such samples (Tippet, 1925). The first paper where asymptotic distributions of largest values (from a class of individual distributions) were considered appeared in 1927 by Frechet (1927). A year later, Fisher and Tippet (1928) published the paper that is now considered the foundation of the asymptotic theory of extreme value distributions. Independently, they found Frechet's asymptotic distribution and constructed two others. These three distributions have been found adequate to describe the extreme value distributions of all statistical distributions. We will explore further this result in subsequent chapter. Moreover, they showed the extremely slow convergence of the distribution of the largest value from Gaussian samples toward asymptote, which has been the main reason for the difficulties encountered by prior investigators. Indeed, the use of the Gaussian distribution as starting point has hampered the development of the theory, because none of the fundamental extreme value theorems is related in a simple way to the Gaussian distribution.

Some simple and useful sufficient conditions for the weak convergence of the largest order statistic to each of the three types of limit distributions were given by von Mises (1936). A few years later Gnedenko (1943) provided a rigorous foundation for the extreme value theory and necessary and sufficient conditions for the weak convergence of the extreme order statistics.

The theoretical developments at the 1920s and mid 1930s were followed in the late 1930s and 1940s by a number of publications concerning applications of extreme value statistics. Gumbel was the first to study the application of extreme value theory. His first application was to old age, the consideration of the largest duration of life. In the sequel he showed that the statistical distribution of floods could be understood by the use of extreme value theory (Gumbel, 1941). 'Extreme value' procedures have also been applied extensively to other meteorological phenomena (such as rainfall analysis), to stress and breaking strength of structural materials and to the statistical problem of outlying observations.

The applications mentioned above all refer to the early development of statistical analysis of extremes from a theoretical as well as practical point of view. Gumbel's book of 1958 (Gumbel, 1958) contains a very extensive bibliography of the developed literature up to that point of time. Of course, since then many more refinements of the original ideas and further theoretical developments and fields of applications have emerged. Some of these recent developments will be further discussed in the chapters to follow. Still, while this extensive literature serves as a testimony to the validity and applicability of the extreme value distributions and processes, it also reflects the lack of co-ordination between researchers and the inevitable duplication of results appearing in a wide range of publications.

1.3 Fields of Application

The reader should have already gained an idea of the diversity of fields where extreme-value analysis can be applied. In the sequel, we give only a short review of the most important areas where extreme-value theory has already been successfully implemented.

- ***Hydrology – Environmental Data***

As we have mentioned hydrologists were of the first to use extreme-value theory in practice. Here, the ultimate interest is the estimation of the T-year flood discharge, which is the level once exceeded on the average in a period of T years. Under standard conditions, the T-year level is a high-quantile of the distribution of discharges. Thus, one is primarily interested in a quantity determined by the upper tail of the distribution. Since, usually the time span T is larger than the observation period, some additional assumptions on the underlying distribution of data have to be made. If the statistical inference is based on annual maxima of discharges, then hydrologists favoured model is the extreme-value model. Alternatively, if the inference is based on a partial duration series, which is the series of exceedances over a certain high threshold, the standard model for the flood magnitudes is the generalized Pareto model.

There is a large literature of extreme-value analyses applied to hydrological data. Hosking et al. (1985) and Hosking and Wallis (1987) apply their proposed estimation method to river Nidd data (to 35 annual maxima floods of the river Nidd, at Hunsingore, Yorkshire, England). Davidon and Smith (1990) apply the generalized Pareto distribution to more detailed data of the same river, taking into account both seasonality and serial dependence of data. Dekkers and de Haan (1989) are concerned with the high tide water levels in one of the islands at the Dutch coast.

The increasing need to exploit coast areas combined with the concern about the greenhouse effect has resulted in a demand for the height of sea defences to be estimated accurately, and to be such that the risk of the sea-dyke being exceeded is small and pre-specified. So, more elaborate techniques have recently been developed. Tawn (1992) performed extreme-value analysis to hourly sea levels by taking into account the fact that the series of observations is not a stationary sequence (due to astronomical tidal component). Barão and Tawn (1999) utilize bivariate extreme value distributions to model data of sea-levels at two UK east coast sites, while de Haan and de Ronder (1998) model wind and sea data of Netherlands using bivariate extreme value d.f.

Extreme low sea levels are of independent interest in applications to shipping and harbour developments and for the design of nuclear power station cooling water intakes. With simple adaptations most methods can be applied to sea-level minima to solve such problems.

Another related issue is that of rainfalls. The design of large-scale hydrological structures requires estimates to be made of the extremal behaviour of the rainfall process within a designated catchment region. It is common to simulate extreme events (rainfalls) and then to access the consequent effect on hydrological models of reservoirs, river flood networks and drainage systems. Coles and Tawn (1996) exploit extreme value characterizations to develop an explicit model for extremes of spatially aggregated rainfall over fixed durations for a heterogeneous spatial rainfall process.

Furthermore, in ecology, higher concentration of certain ecological quantities, like concentration of ozone, acid rain or SO_2 in the air are of great interest due to their negative response on humans and generally, on the biological system. For example, Smith (1989) performs extreme value analysis in ground-level ozone data, taking into account phenomena common in environmental time series, such as seasonality and clustering of extremes. Similar is the subject dealt with in Küchenhoff and Thamerus (1996).

- ***Insurance***

Estimating loss severity distributions (i.e. distributions of individual claim sizes) from historical data is an important actuarial activity in insurance. In the context of re-insurance, where we are required to choose or price a high-excess layer, we are specifically interested in estimating the tails of loss severity distributions. In this situation it is essential to find a good statistical model for the largest observed historical losses; it is less important that the model explains smaller losses. In fact, a model chosen for its overall fit to all historical losses may not provide a particularly good fit to the large losses. Such a model may not be suitable for pricing a high-excess layer. It is obvious that extreme-value theory is the most appropriate tool for this job, either by using extreme value distribution to model large claims or generalized Pareto distribution to model exceedances over a high threshold.

The applicability of extreme-value theory to insurance is discussed by Beirlant et al. (1994), Mikosch (1997), McNeil (1997), McNeil and Saladin (1997) with application to Danish data on large fire insurance losses, Rootzen and Tajvidi (1997) with application to Swedish windstorm insurance claims.

- ***Finance – Risk Management***

Finance and, even more general, risk management are areas where only recently extreme-value theory has gained ground. Insurance and financial data can both be investigated from the viewpoint of risk analysis. Therefore, the insight gained from insurance data can also be helpful for the understanding of financial risks.

Mainly due to the increase in volume and complexity of financial instruments traded, risk management has become a key issue in any financial institution or corporation of some importance. Globally accepted rules are put into place aimed at monitoring and managing the full diversity of risk. Extreme event risk is present in all areas of risk management. Whether we are concerned with credit, market or insurance risk, one of the greatest challenges to the risk manager is to implement risk management models which allow for rare but damaging events, and permit the measurement of their consequences. In market risk, we might be concerned with the day-to-day determination of the Value-at-Risk (VaR) for the losses we incur on a trading book due to adverse market movements. In credit or operational risk management our goal might be the determination of the risk capital we require as a cushion against irregular losses from credit downgrading and defaults or unforeseen operational problems. No discussion has perhaps been more heated that the one on VaR. The biggest problem with VaR is the main assumption in the conventional models, i.e. that portfolio returns are normally distributed.

In summary the main points of risk management are the followings

- Risk management is interested in estimating tail probabilities and quantiles of profit-loss distributions, and indeed of general financial data
- Extremes do matter
- We want to have methods for estimating conditional probabilities concerning tail-events: “Given that we incur a loss beyond VaR, how far do we expect the excess to go?”
- Financial data show fat tails.

Extreme-Value Theory is a subject whose motivations match the four points highlighted above. It has a very important role to play in some of the more technical discussions related to risk management issues. The usefulness of extreme-value theory to risk management is stressed by Danielsson and de Vries (1997), McNeil (1998 and 1999), Embrechts et al. (1998, 1999), Embrechts (1999).

- ***Teletraffic Engineering***

Classical queuing and network stochastic models contain simplifying assumptions guaranteeing the Markov property and insuring analytical tractability. Frequently, inter-arrival and service times are assumed to be i.i.d. and typically underlying distributions are derived from operations on exponential distributions. At a minimum, underlying distributions are usually assumed nice enough that moments are finite. Increasing instrumentation of teletraffic networks has made possible the acquisition of large amounts of data. Analysis of this data is disturbing since there is strong evidence that the classical queuing assumption of thin tails and independence are inappropriate for these data. Such phenomena as file lengths, CPU time to complete a job, call holding times, inter-arrival times between packets in a network and length of on/off cycles appear to be generated by distributions which have heavy tails. Resnick and Stărică (1995), Kratz and Resnick (1996), and Resnick (1997a) deal with such kind of data.

Other areas where extreme-value analysis has found application are engineering strength of materials (Harter, 1978 provides a detailed literature on this subject), earthquake size distribution (see, e.g., Kagan, 1997), athletic records (Strand and Boes, 1998, Barão and Tawn, 1999), city-sizes, corrosion analysis, exploitation of diamond deposits, demography, geology and meteorology among others.

1.4 Overview

In the second chapter we provide laws, theorems and propositions that constitute the theoretical background of extreme values. In chapter 3, some fully parametric estimation methods are described. Another family of estimation methods is presented in chapter 4, the family which includes the well-known extreme-value index estimators Hill, moment, and so on. These are the so-called semi-parametric estimation methods, on which special emphasis is put. In chapter 5 some recently suggested methods for improving the performance of extreme-value index estimators are described. Their performance is evaluated via simulation. In chapter 6, extreme-value analysis on teletraffic data is performed. Chapter 7 is the final chapter of this thesis, where the main findings are summarized and suggestions for further research are made.