Chapter I

INTRODUCTION

The general framework

The purpose of STATISTICAL ANALYSIS of EXTREMES is to deal with, analyse and predict (or forecast) aspects of natural phenomena that correspond to the largest or smallest values of sampled data, or to over - or underpassing some level that sometimes lead to natural disasters. Floods, large fire claims, heavy rains, gusts of wind, and large waves, are examples of maxima or largest values of samples, as well as droughts, breaking strength of materials, failure of equipment or apparatus, low temperatures, etc. are examples of minima or smallest values of samples, of extremes of samples or extreme order statistics; many times not only the first but the second, third, etc. extremes are relevant.

Those problems can be analysed using suitable or available samples — according to the manageability of technique, error allowance for models, convenience and simplicity, etc — as univariate samples, multivariate samples or excursion(s) of random sequences or of stochastic processes (if time is continuous). Evidently, from the philosophical point of view, as a rule, natural phenomena should be considered as multivariate random sequences or stochastic processes (according to the discreteness or the continuity of the time set). This procedure, however sound it may be, by its complication would not allow for a (practical) description of the observations (the past and/or present) but also to predict about the next observations (the future) to legitimate sound inference and/or decision design.

In fact, in data analysis, we can sometimes proceed through a sequence of models, adapted to the case. Case studies, as in the last chapter and in the statistical chapters of the book, may be of help to practitioners.

We have not tried to prove all results, to keep the book to a manageable size (though not necessarily completely balanced) and to allow for easier use for data analysis and design; essential references are given for the proved (sometimes in a new or different way) or unproved results.

Basically, the applications will be based on the use of asymptotic distributions as approximations (better or worse) to the description of extremes, useful for forecasting: so obtaining asymptotic distributions, as well some reference to the speed of convergence, plays an essential role in the book.

If the approximation is good we will, with good efficiency, estimate or predict "extreme" quantiles (for large or small probabilities) for the design of structures, evaluation of extreme conditions, forecasting catastrophes, etc..

As a rule, the behaviour of sums and of extremes (maxima and minima) of samples are different and not strongly related, although there are some connexion as regards convergence of averages and of extremes. But, as a general rule, in the i.i.d. case, characteristic functions for sums vs. distribution functions for maxima (or survival functions for minima), the linear combination $aX + bY$ vs. the maximal combination $\max (X + a, Y + b)$, seem to play a similar or dual role; but the duality almost stops here. The Central Limit Theorem leads to the normal distribution (at least when the variance exists) or more generally to the infinitely divisible distributions with a central role for the normal, but the Extremal Limit Theorem leads to three asymptotic distributions (which can be integrated in one form, with a shape parameter) where the Gumbel distribution plays a central role; not only are both distributions indefinitely differentiable but, also, sums of Gumbel random variables are asymptotically normal as well as maxima or minima of normal random variables have asymptotically the Gumbel distribution. It does not seem that we can extend the analogies and this "experimental" digression will play only a vague role — behind the scenes — in the book. But see the Annex I, of Part I, for more details that extend to bivariate distributions.

Let us consider one example to clarify the use of the theory of statistical extremes. Suppose that we are considering the levels (or discharges) of a river at some fixed point and that we are interested in the yearly floods or droughts at that point. As the flood (drought) is the maximum (minimum) of the discharges we would be led to consider the exact statistical behaviour of the maxima (minima). The difficulties of dealing with them and the fact the yearly maximum considered is obtained from a large number of (daily) observations suggest using the asymptotic theory of maxima, whose distribution is known except for two or three parameters. The validity of this approximate substitution will be considered and referred to. A set of 60 or more years of observations in general gives sufficient data to estimate those unknown parameters with adequate accuracy and allows subsequent use of the asymptotic distribution function of the maxima or minima.

For bivariate extremes the situation is similar, although more complicated. Suppose that we are measuring the discharges of a river at two different points every day, say, at noon. It is evident that the measurements are positively correlated, the correlation being stronger and stronger as the points are closer and closer. From this series of daily measurements (pairs) we can obtain a sample sequence of pairs of maxima (the floods). We will see later how to deal with those samples of (yearly) pairs of maxima, considered as random pairs.

As one year is a relatively large sample, we will use the large sample (asymptotic) distribution of pairs of maxima as a substitute for its very complex and unknown actual distribution, but the representation of the large sample distribution of pairs of maxima has some unknown parameters and a dependence function.

The sample of pairs of maxima obtained in 50, 60, ..., 100 years of observations can then be used to solve the statistical decision problems concerning the unknown elements of the representation, specifically those that are most important in engineering applications: estimation of the parameters and testing hypotheses about them, testing for the independence of the maxima, and estimation of the dependence function.

We will not refer specially to the behaviour of minima (droughts, for example) because it is similar to that of maxima; we will recall the technique of conversion of results concerning maxima into results concerning minima.

The stochastic model

Let $\{X_1,..., X_n\}$ be a set of multidimensional random vectors with the joint distribution function $F_n(x_1,..., x_n) = \text{Prob}\{X_1 \le x_1,..., X_n \le x_n\}$ and with survival function $S_n(x_1, ..., x_n) = \text{Prob}\{X_1 > x_1, ..., X_n > x_n\}$; the relations between $F_n(x_1, ..., x_n)$ and $S_n(x_1, ..., x_n)$ are known in the continuity points of the margins being extended afterwards by right continuity for F_n and S_n.

We have $\text{Prob}\{\max_1^n \{X_i\} \le x\} = F_n(x, ..., x) = F_n(x)$ and

$\text{Prob}\{\min_1^n \{X_i\} > x\} = S_n(x, ..., x) = S_n(x)$; in any point x we see that

$$\text{Prob}\{\max_1^n X_i > x\} = 1 - F_n(x, \dots, x) = 1 - F_n(x) \qquad \text{and}$$

$$\text{Prob}\{\min_1^n X_i > x\} = 1 - S_n(x, \dots, x) = 1 - S_n(x).$$

Evidently as $\max_1^{n+1} X_i \geq \max_1^n X_i$ we have $F_n(x) \geq F_{n+1}(x)$, and as

$\min_1^{n+1} X_i \leq \min_1^n X_i$ we obtain $S_{n+1}(x) \leq S_n(x)$.

Note that if $h(.)$ is an increasing continuous function then $\max_1^n \{h(X_i)\} = h(\max_1^n X_i)$ and

$\min_1^n \{h(X_i)\} = h(\min_1^n X_i)$; if $h(.)$ is a decreasing continuous function we obtain $\max_1^n \{h(X_i)\}$

$= h(\min_1^n X_i)$ and $\min_1^n \{h(X_i)\} = h(\max_1^n X_i)$.

When we take $h(u) = -u$ we obtain the essential equalities

$$\max_1^n \{X_i\} = -\min_1^n \{-X_i\} \quad \text{and} \quad \min_1^n \{X_i\} = -\max_1^n \{-X_i\} .$$

These relations translate maxima results into minima results and vice-versa. In this book we will almost always deal with maxima results, the conversion to minima results following immediate; the exceptions will be, essentially, the study of the Weibull distribution for minima and the study of the dependence function with (standard) exponential margins in bivariate extremes.

Notice that these simple results in $h(.)$ — a very general transformation — may after be helpful in dealing with data, by a convenient transformation.

Let us apply the results above to the i.i.d. case; we have $F_n(x) = F_1^n(x) = (1 - S_1(x))^n$ and $S_n(x) = S_1^n(x) = (1 - F_1(x))^n$.

Order statistics and exceedances

Let $(X_1, \dots, X_n)$ be an i.i.d. univariate sample where $F(x)$ is the distribution function of the X_i. Let us order the sample as $X'_{1,n} \leq X'_{2,n} \leq \dots \leq X'_{n,n}$ (*ascending order statistics*) or $X''_{1,n}$

$\geq X''_{2,n} \geq \ldots \geq X''_{n,n}$ (*descending order statistics*) ; note that $X'_{k,n} = X''_{n+1-k,n}$ and that $X'_{1,n} = X''_{n,n}$ is the minimum of the sample, $X'_{n,n} = X''_{1,n}$ is the maximum of the sample, $X'_{2,n} = X''_{n-1,n}$ and $X'_{n-1,n} = X''_{2,n}$ are second minimum and maximum (second extremes), $X'_{k,n} = X''_{n+1-k,n}$ and $X'_{n+1-k,n} = X''_{k,n}$ are the k-th minimum and maximum. When possible and clear we will use the simple notations X'_k and X''_k.

The distribution function of $X'_{k,n} = X''_{n+1-k,n}$ is

$$F_{k,n}(x) = \text{Prob}\,\{X'_{k,n} \leq x\} = \text{Prob}\,\{k \text{ or more } X_i \leq x\} =$$

$$= \sum_{j=k}^{n} \binom{n}{j} F(x)^j (1 - F(x))^{n-j} ,$$

by the binomial distribution.

If $F(x)$ has a probability density $f(x) = F'(x)$, the probability density of $X'_{k,n} = X''_{n+1-k'}$ is

$$F'_{k,n}(x) = F(x)^{k-1} (1 - F(x))^{n-k} f(x) = k\binom{n}{k} F(x)^{k-1} (1 - F(x))^{n-k} .$$

Analogously the distribution function of $(X'_{k,n}, X'_{m,n}) = (X''_{n+1-k,n}, X''_{n+1-m,n})$, can be written with $1 \leq k < m \leq n$; its probability density is immediately seen to be, for $X'_{k,n} = x$, $X'_{m,n} = y$,

$= 0$ if $y \leq x$

$$= \frac{n!}{(k-1)!(m-k-1)!(n-m)!} F^{k-1}(x) (F(y) - F(x))^{m-k-1} (1 - F(y))^{n-m} f(x) f(y) \text{ if } y \geq x.$$

From these distribution functions and their multivariate extensions — where concomitants can intervene, but which will not be dealt with in this book — we can obtain the asymptotic

behaviour of the sample quantiles which are *central* order statistics, i.e., in a sequence of sample the k_n-th ascending order statistic $X'_{k_n,n}$ is such that $k_n/n \to p$ with $0 < p < 1$.

Let us now suppose that we are thinking of a second sample of m random variables with the same distribution function F(x). The probability that *exactly* $J = j$ of the random variables of the second sample are smaller or equal to $X'_{k,n} = X''_{n+1-k,n} = x$ is evidently $\binom{m}{j} F(x)^j (1 - F(x))^{m-j}$

and, thus, the probability — *before* the first sample — that exactly j out of m observations are smaller or equal to the k-th ascending order statistic $X'_{k,n}$ of the first sample is

$$P(j,m|k,n) = \int_{-\infty}^{+\infty} \binom{m}{j} F^j(x) (1 - F(x))^{m-j} \; k\binom{n}{k} F^{k-1}(x) (1 - F(x))^{n-k} \; f(x)\, dx$$

$$= k \binom{m}{j} \binom{n}{k} \int_{-\infty}^{+\infty} F^{j+k-1}(x) (1 - F(x))^{m-j+n-k} \; f(x)\, dx$$

$$= \frac{k \binom{n}{k} \binom{m}{j}}{(k + j) \binom{n+m}{k+j}}$$

with $$\sum_{j=0}^{m} \frac{k \binom{n}{k} \binom{m}{j}}{(k + j) \binom{n+m}{k+j}} = 1 .$$

Note the symmetry $P(j,m|k,n) = P(m - j, m \mid n + 1 - k , n)$ and that this last probability is that exactly j observations of the second sample will be larger than or equal to $X''_k = X'_{n+1-k'}$.

These situations, for the second sample, are called *exceedances* : the underpassing of $X'_{k,n}$ or the overpassing of $X''_{k,n}$.

Let us now compute the mean value and variance of the random number of exceedances J (below $X'_{k,n}$ or above $X''_{k,n}$). We have

$$\mu(m|k,n) = M(J) = \sum_{j=0}^{m} j \, P(j,m|k,n) = \frac{m\,k}{n+1}$$

and

$$\sigma^2 (m|k,n) = M(J^2) - M^2(J) = M(J(J-1)) + M(J) - M^2(J)$$

$$= \frac{m(m+n+1)}{(n+1)^2 (n+2)} \; k\,(n+1-k)\,.$$

The variance is smaller for $k = 1$ or $k = n$ and the ratio of the variances $\sigma^2 (m|1,n) = \sigma^2 (m|n,n)$ to the variance at $k = \frac{n+1}{2}$ (the median if n is odd) is $\frac{4n}{(n+1)^2} \sim 4 / n$, which shows that, for exceedances, extremes are more reliable than central values, contrary to the usual belief.

Let us now consider, more generally, the random time T of the *s-th* exceedance (the underpassing of $X'_{k,n}$ or the overpassing of $X''_{k,n}$) in a (possibly) infinite second sample (m not fixed as before). Thus the random order T in which the exceedance happens does not relate to a fixed event — as in the (classical) return period situation — but to a random event, defined previously to the observation of the first sample of n random variables. Consequently we will not need the knowledge the distribution of $X'_{k,n}$ or of $X''_{k,n}$, as was the case in the return period situation, because the set is defined previously to the observations : briefly speaking, we have an inverse binomial with an underlying random probability. The approach will, naturally, be non-parametric as for exceedances.

For the *s-th* underpassing of $X'_{k,n}$ at the $T = j$ observation we have

$$P(j|s,k,n) = \text{Prob}\,\{T = j|s,k,n) = \int_{-\infty}^{+\infty} [\,\binom{j-1}{s-1} F^s(x)\;(1 - F(x))^{j-s}\,] \;x$$

$$x\,[\,n\binom{n}{k}\;F(x)^{k-1}\,(1 - F(x))^{n-k}\,f(x)\,]\,dx,$$

where the meaning of the two brackets is clear from the previous study.

It is immediate that

$$P(j|s,k,n) = \frac{k\;\binom{n}{k}\,\binom{j-1}{s-1}}{(k+s)\,\binom{n+j}{k+s}}$$

with $\sum_{j=s}^{+\infty} P(j|s,k,n) = 1$ obviously. This probability is, clearly, also the probability that $X''_{k,n}$ is overpassed at the j-th observation of the second sample the s-th time.

The mean value and variance of T are

$$\mu\,(s|k,n) = M(T) = \sum_{j=s}^{+\infty} j\;\; P(j|s,k,n) = \frac{n\;\; s}{k\;\; -\;\; 1}$$

and

$$\sigma^2\,(s|k,n) = M(T^2) - M^2(T) = M(T(T+1)) - M(T) - M^2(T) =$$

$$= \frac{ns(n\;\; -\;\; k\;\; +\;\; 1)\;\; (k\;\; +\;\; s\;\; -\;\; 1)}{(k-1)^2\;\; (k-2)}\;\; .$$

For $k = 1$ we have $\mu(s|1,n) = +\infty$ which is easily understandable and for $k = n$ we get $\mu(s|1,n) = \frac{n\;\; s}{n\;\; -\;\; 1}$, results that could be obtained directly. Also for $k = 1$ and $k = 2$ we have $\sigma^2\,(s|1,n) = +\infty$ (as expected) and $\sigma^2\,(s|2,n) = +\infty$.

It is convenient to compare the two different approaches of this section: exceedances and the random return period, both for the second sample, previously to the knowledge of the first sample, that, allowing for the use of non-parametric methods, overcomes the necessity of knowledge of the underlying continuous F(x). In the first case (exceedances) we study the number of underpassings or overpassings of some order statistics of the first sample in a finite number of observations as the second sample has an infinite size. For the random return period we accept, admit, or consider an infinite second sample (thus allowing for infinite i.i.d. observations) and we are dealing with a random stopping time. The analogy with the direct and inverse binomial is obvious.

A note on the asymptotic behaviour of the sample quantiles

As it is well known the usual estimator of the unique theoretical quantile χ_p $(0 < p < 1)$ is the order statistic $X'_{[np]+1}$ called the *empirical* or *sample quantile* χ^*_p which we will denote also by $Q_n(p)$; note that $X'_1 \le Q_n(p) \le X'_n$ if $(0 < p < 1)$. In the case of existence of a

density $f(x) = F'(x)$ we know that $\chi_p^* \xrightarrow{P} \chi_p$ and $\sqrt{n}\ f(\chi_p) \dfrac{\chi_p^* - \chi_p}{\sqrt{p(1-p)}}$ is asymptotically standard normal if $0 < f(\chi_p) < \infty$; the joint distribution of the set of quantiles $(\chi^*_{p_1}, \ldots, \chi^*_{p_n}) =$ $(Q_n(p_1) \ldots, Q_n(p_k))$, is such that $\{\sqrt{n}\ f(\chi_{p_i}) \dfrac{\chi^*_{p_i} - \chi_{p_i}}{\sqrt{p_i(1-p_i)}}\}$ is asymptotically multinormal with standard margins and correlation coefficients, for $p_i < p_j$, $\rho_{ij} = \sqrt{\dfrac{p_i}{p_j} \dfrac{1-p_j}{1-p_i}}$ (>0); for details see Cramér (1946).