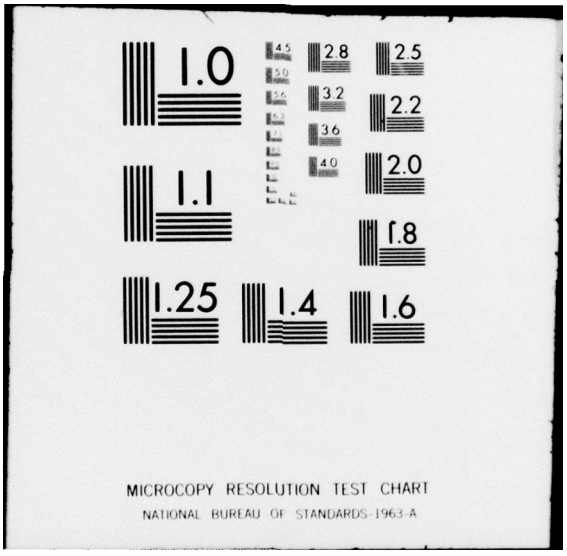
AD-A059 021 TEMPLE UNIV PHILADELPHIA PA DEPT OF MATHEMATICS F/G 12/1
EXTREME VALUE THEORY IN APPLIED PROBABILITY.(U)
JUL 78 J GALAMBOS
AFOSR-78-3504
UNCLASSIFIED AFOSR-TR-78-1138 NL

| OF |
AD
A059021

END
DATE
FILMED
11-78
DDC


1.0
1.1
1.25
1.4
1.6
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.8
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

LEVEL II

AD A059021

DDC FILE COPY

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: AFOSR-TR- 78-1193	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): EXTREME VALUE THEORY IN APPLIED PROBABILITY	5. TYPE OF REPORT & PERIOD COVERED: Interim
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Janos Galambos	8. CONTRACT OR GRANT NUMBER(s): AFOSR 78-3504
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Temple University, Department of Mathematics, Philadelphia, PA 19122	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Office of Scientific Research/NM, Bolling AFB, Washington, DC 20332	12. REPORT DATE: July 1978
	13. NUMBER OF PAGES: 11
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

DDC RECEIVED SEP 26 1978 D

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Conference on Stochastic Processes and their Applications, 6-10 July 1978, Canberra, Australia

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Several problems of applied probability lead to the investigation of extremes of a sequence of random variables (floods, strength of materials, failure models, maximal queue length, air pollution data, etc.). It is unrealistic in most cases to assume that the random variables in question are independent and identically distributed. However, there are several dependent models when the distribution of the k-th largest is well approximated by the distribution of the k-th largest of independent, and

DD FORM 1 JAN 73 1473

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Abstract

some times independent and identically distributed random variables (possibly changing the fixed size of the sequence in question to a random one). See the book below, particularly Sections 3.1, 3.12.6.1 and 6.2, in which both the possibility of such approximations and the limit laws when such approximation fails are discussed.

We shall describe several models with emphasis on the possibility of approximations by independent random variables. In addition, we shall analyze the following embarrassing question of an applied scientist:

in a sample of size n = 1,000, say, is the 10-th largest an extreme in the sense that k=10 is fixed as n increases (if n = 1,000 is not large enough, take $n = 10^6$), or should one consider the 10-th largest as the $\sqrt[3]{n}$ - th largest? Since the limiting distributions are different in the two cases, the decision on the actual case to be applied is significant.

Through this analysis, we shall point out several important "rules" for applying asymptotic extreme value theory.

AFOSR-TR-78-1138

(16) (19)

(9) Interim rept.

(6) Extreme Value Theory in Applied Probability

by

(10) Janos / Galambos

Department of Mathematics
Temple University
Philadelphia, Pa. 19122.

(11) Jul 78 (12) 12 p.

(15) AFOSR-78-3504

(16) 2304 (17) A5

Preliminary Report. Presented at the Conference on Stochastic Processes and their Applications, 6-10 July, 1978, Canberra, Australia.

Research supported by a Grant (#78-3504) from AFOSR to Temple University.

404 207

A b s t r a c t

Several problems of applied probability lead to the investigation of extremes of a sequence of random variables (floods, strength of materials, failure models, maximal queue length, air pollution data, etc.). It is unrealistic in most cases to assume that the random variables in question are independent and identically distributed. However, there are several dependent models when the distribution of the k-th largest is well approximated by the distribution of the k-th largest of independent, and some times independent and identically distributed random variables (possibly changing the fixed size of the sequence in question to a random one). See the book below, particularly Sections 3.1, 3.12, 6.1 and 6.2, in which both the possibility of such approximations and the limit laws when such approximation fails are discussed.

We shall describe several models with emphasis on the possibility of approximations by independent random variables. In addition, we shall analyze the following embarrassing question of an applied scientist: in a sample of size $n = 1{,}000$, say, is the 10-th largest an extreme in the sense that $k=10$ is fixed as n increases (if $n = 1{,}000$ is not large enough, take $n = 10^6$), or should one consider the 10-th largest as the $\sqrt[3]{n}$ - th largest ? Since the limiting distributions are different in the two cases, the decision on the actual case to be applied is significant.

Through this analysis, we shall point out several important "rules" for applying asymptotic extreme value theory.

Reference

Galambos, Janos (1978). The Asymptotic Theory of Extreme Order Statistics. John Wiley & Sons, Inc., New York.

1. Stochastic Model Building

1.1. General comments. One of the most crucial part of an applied scientist's decision is the adoptation of a stochastic model for the random quantities $X_1, X_2, \ldots, X_n$ he is to deal with. Whether these random quantities are produced by nature (floods, winds, etc.), or by a manufacturing procedure (life lengths of components of an equipment), or the statistician collects them by random sampling, it is very rare that a scientific reasoning would lead to a unique dependence structure and a well defined family of distributions for the X's. Consequently, the basic underlying dependence and distributional assumptions are subjective to a large extent. The question thus arises whether substantially different conclusions can be reached by two scientists, assuming that both of them work under "reasonable assumptions"? Here "reasonable" means that the assumptions are compatible with general practice in the scientific literature.

For an outsider, the fact alone that this question arises should be shocking who trusted us that scientific decisions ought to be unquestionable. Unfortunately, not just the question arises but the answer to it is the real disappointment: the subjectivity of model building does influence the otherwise uniquely determined scientific procedure of decision making. Therefore, the choice of the model has to be based on more careful studies than a routine acceptance of independence and a family of distributions (normality or other popular ones). One cannot expect to build a theory of model building with a general appeal, since the practical problem to be settled should be a major consideration in adopting a model. Evidently, one has to be more accurate when human life is involved such as effects of food additives, drugs and medical treatment or safety of equipments. Financial and legal considerations also increase the demand for more accurate models. But whether newer approaches are more accurate is very doubtful. The almost daily

rejection of food additives on the base of moving away from the classical assumption of normality to a new underlying distribution does not correct the error of the model by which these same additives were accepted earlier as safe. Neither did the model used for the evaluation of air pollution data become more accurate by changing the assumption concerning the distribution of pollutant concentration first from normal to lognormal and more recently to other families, mainly mixtures. One general rule can, however, be set up: decisions of major importance should be reached after the data has been evaluated in several alternate models.

Examples in subsequent sections will support all claims of this introduction.

1.2. Stochastic models with extremes. In the book, Galambos (1978), the present author described and developed several dependent models, when the extremes govern the laws of interest. The aim of the present paper is to summarize some of these available models and to point to an important direction of future development.

Let us start with a list of applied fields where the solution is in terms of the maximum Z_n or minimum W_n of the basic random quantities X_j, $1 \leq j \leq n$, associated with the actual practical problem.

The annual flood on a river is measured as the maximum Z_n of the daily highest water levels X_j, $j \geq 1$. The X's here are dependent and closer the observations are in time, the stronger the dependence is. Within a season, and at some locations throughout the year, the X's can be assumed to be identically distributed. In order to plan for building a dam on a river at a given location, an accurate description of Z_n is required. If two, or if several dams are to be built on the same river, then the X's have to be measured at those selected locations and our interest is the multivariate distribution of the corresponding Z's.

Notice that a choice of a mathematical model for the description of the annual flood is equivalent to accepting a concept

for a weakening dependence as time passes and a choice for the common distribution of the X's in a season or for the year as the case may be. Because other applied fields lead to this same mathematical problem, let us go ahead with the list mentioned earlier.

The random strength of a sheet of certain material under stress can be expressed as a minimum W_n as follows. Let us hypothetically subdivide the sheet into n parts. If X_j is the strength of the j-th part in this subdivision, then the so called weakest link principle tells us that the strength of the original sheet is indeed W_n of these X_j. Again, those X_j which represent the strength of distant parts have weaker dependence than neighbouring ones and in fact some kind of almost independence is valid as distance increases indefinitely. Here, unit is proportional to the reciprocal of the square root of n (make the hypothetical subdivision of a rectangle in such a way that both edges are divided into m equal parts and set $n = m^2$). If the division is done into equal parts, then the X's can again be assumed to be identically distributed.

The time to the first failure of a complicated equipment is best approached by first grouping components into so called cuts or pathes. By such an approach, an arbitrary equipment becomes equivalent to a parallel or series system. However, while parallel systems are installed for safety considerations and thus they are expected to function independently, the above reduction procedure can lead to very complicated dependence structures. The actual grouping procedure is as follows. A path is a set of components whose functioning insures the functioning of the equipment. A minimal path is a path whose number of elements cannot be reduced without violating the defining property of a path. On the other hand, a cut is a set of components whose failure causes the equipment to fail. Finally, a minimal cut is a cut with a minimal number of elements. Now if X_j represents the random time to the first failure of the j-th minimal path of the equipment, then Z_n is

the time to the first failure of the equipment. This same random waiting time until the equipment fails can also be expressed as the minimum W_s of the random times upto the failures of minimal cuts. Since either n or s is large for a complicated equipment, its failure is well described by an asymptotic model. However, since different path sets may contain several common components, the assumption of independence of the X_j is unjustified even in an approximate sense. Evidently, a similar remark applies to cuts as well.

Here we cannot speak of close and distant neighbours of path sets and yet the dependence structures of the three problems listed can be described by a unified model. What will be essentially different in the failure model when compared with the models for strength or flood that a single distribution cannot be used for all path sets. We should rather seek a mathematical solution when each path set is permitted to have different failure distribution. Strangely indeed, this generality leads to the mathematical conclusion that engineers applied all along: any distribution with monotonic hazard rate is an asymptotic failure distribution.

The unified dependence model for the three listed problems is given by a set of random variables $X_1, X_2, \ldots, X_n$ which form a so called E_n^*-sequence. Since it is accurately defined on pp.176-177 of the author's book referred to earlier, we describe it here in vague terms only. The requirement is that a set E_n^* of exceptional pairs (X_i, X_j) can be found with which the following three basic properties hold for the original sequence X_j, $1 \leq j \leq n$:

(i) the events $\{X_{j_t} \geq x\}$ are asymptotically independent as x becomes "large", whenever no pairs of the subscripts j_t are exceptional;

(ii) if exactly one pair, (i_m, i_t) say, is exceptional among the subscripts of X_{i_s}, $1 \leq s \leq k$, then the

probability of the intersection of the events $\{X_{i_s} \geq x\}$ is proportional to the product of $P(X_{i_m} \geq x, X_{i_t} \geq x)$ and the univariate tails $P(X_{i_s} \geq x)$, $s \neq m,t$;

and

(iii) the number of elements in the exceptional set E_n^* is of smaller magnitude than n^2, which is the number of all pairs of the subscripts of the X_j .

Under these conditions, the univariate and bivariate marginals of the X_j determine whether the maxima, when normalized, have a limiting distribution and when their do, what is the actual form of this limiting distribution. There is a simple criterion available, and which is of wide applicability to applied questions, which guarantees that the limiting distribution of the normalized maximum for an E_n^*-sequence exists and this limiting distribution is the same as if the X_j were completely independent. However, when this condition fails, the limiting distribution may still exist, but the corresponding theory is not well developed for such case.

This unified approach through E_n^*-sequences covers several well developed dependence concepts such as m-dependence and variants of mixing. As was emphasized earlier, the basic assumptions of a model are mainly a matter of belief when applied in a concrete situation; the concept of an E_n^*-sequence is developed exactly with this dilemma in mind. While the special case of a mixing concept, say, requires the validity of a weakening dependence for certain sequences of sigma fields as the major assumption, our only assumption that cannot be checked is (ii). Namely, one can construct the set E_n^* in such a way that (i) and (iii) be automatically satisfied. This construction of E_n^* for the three problems mentioned is self evident. For the flood and strength models, E_n^* will contain any pairs which represent "close neighbours", while, for the failure model, (i,j) belongs to E_n^* if X_i and X_j represent life length of such path sets which have common elements.

Another approach to failure models is possible through exchangeable variables. It is quite a surprising result that a kind of averaging is possible in such a sensitive field as extreme value theory (see Sections 3.1 and 3.2 of the book by Galambos (1978)). We do not discuss this approach here, but we wish to draw attention to it because of its varied possibility for further investigations.

We conclude this section by mentioning two approaches to specific problems of extreme value theory which bear interesting mathematical facts. One is the approach by C.C. Heyde (1971) to queue length, where dependent variables are "transformed" to independent ones (but with a random number of elements). The other one is the classical approach by H.E. Daniels (1945) to strength of bundles of threads, in which the opposite is done: originally independent variables are "transformed" to dependent ones. Interestingly, these dependent variables can also be represented as a function of independent random variables of random size, namely, a quantile with random index. However, no attempt has been made so far to develop the theory of strength of bundles on this latter line.

Further important applications are touched upon in the next section where we discuss the effects of the choice of a population distribution.

2. The choice of a population distribution

We speak of a population distribution when a random phenomenon can be described by a sequence of identically distributed random variables $X_1, X_2, \ldots, X_n$. This is the case of floods, strengths (both sheets and bundles), certain failure models (e.g. parallel and series systems), permitted level of air pollution concentration (the US Standard is set by the second largest observation) and effects of food additives or drugs on persons belonging to a group of identical physical conditions. Notice that independence is not emphasized. However, in the analysis that follows it will be

on occasion easier to stress a point by assuming that an approximation by independent variables is possible.

Let us put down some notations. The basic random variables are denoted by $X_1, X_2, \ldots, X_n$. They are assumed to be identically distributed and we use $F(x)$ for their common distribution function. We put Z_n and W_n for the maximum and the minimum, respectively, of the X's, whose distribution functions are $H_n(x)$ and $L_n(x)$, respectively. Finally, the empirical distribution function of the X's is denoted by $F_n(x)$. In the frequency approach to probability, $F_n(x)$ has to be a good approximation to $F(x)$. But can we rely on any good approximation to $F(x)$ if our interest is $H_n(x)$? As we shall see, no classical approximation is acceptable and thus specific techniques are needed in extreme value theory. Of course, if we know $F(x)$ accurately through a so called characterization theorem (for questions of normality, see the book by Kagan, Linnik and Rao (1973) and for other distributions, the one by Galambos and Kotz (1978)), then no statistical method is needed.

First let us look at the meaning of the empirical distribution function. By definition, $F_n(x) = 1$ for all $x > Z_n$. The meaning of this fact is that the jump of $F_n(x)$ at Z_n represents the increament $1 - F(x)$ of the population distribution for an x close to Z_n but the actual variation of the tail $1 - F(x)$ is not approximated by $F_n(x)$ for x's exceeding Z_n. But since the tail $1 - F(x)$ alone determines $H_n(x)$ for large values of n, whenever the basic random variables are almost independent, we simply cannot expect $F_n(x)$ to be a guide in approximating $H_n(x)$. Now we just have to observe that both the chi-squared test for goodness of fit and the more frequently used fitting by a specific probability paper are based on $F_n(x)$ and we can thus conclude that classical methods are inapplicable to approximate $F(x)$ in extreme value theory. In order to be more specific, let us look at a numerical example.

Assume that a random quantity X is associated with the operation of an equipment (or the human body). The equipment fails

if X exceeds 3300 and its operation is safe in the range 2000 to 3200. We made n = 500 observations on X and all were in the safety range. In fact, $W_{500} = 2342$ and $Z_{500} = 3071$. Furthermore, the X's were varying around the mean 2698 with standard deviation 104, and the reduced values (X-2698)/104 gave a very good fit on a standard normal probability paper. Is it now justified to use a normal table to compute the safety of 100,000 equipments, say ? If a failure means only a financial loss, then not much problem is involved: one computes the expected number of losses, includes them in the price and gives a warranty against loss. However, if a single failure means disaster (loss of life or permanent damage to humans), then we have to determine the population distribution F accurately enough that the value 100,000(1 - F(y)) with y =(3300-2698)/104=5.788 be negligible. Since our observations do not give any information on F(y) beyond y = (3071-2698)/104 = 3.587, a fitting based on these observations is **insufficient** for our purposes. What we actually want is a good approximation to $H_s(x)$ with s = 100,000. A separate paper is devoted to this problem; see Galambos (1978a).

3. The accuracy of an approximate model

The discussion in the preceding section already shows that accuracy is more significant in extreme value theory than in most classical statistical problems. Unfortunately, very few speed of convergence estimates are available. There is actually only one result on this line which first appeared in the author's book (Section 2.10). That result can also be extended to E_n^*-sequences without much effort. We do not reproduce those results here. Rather we would describe another problem that also leads to a problem of accuracy.

In asymptotic theory one distinguishes between upper extremes and large order statistics depending whether the rank of an order statistic is fixed or varies with the sample size. More precisely, let $X_{k:n}$ be the k-th largest among n observations. If k is fixed as n increases indefinitely, then $X_{k:n}$ is called the

k-th upper extreme. On the other hand, if k varies with n but k/n converges to one as n goes to infinity, then we refer to $X_{k:n}$ as a large order statistic. Now, in several applied fields one faces the problem of evaluating the behaviour of $X_{k:n}$ where k is small and n is large. Which case applies if asymptotic theory is to be the basic tool ? For example, if $k = 10$ and $n = 1{,}000$, is $X_{k:n}$ an upper extreme or a large order statistic in the above sense ? The distinction is significant because in the case of extreme, a gamma type distribution is obtained for the limiting case while with varying k (in our case $\sqrt[3]{n}$), the limiting distribution is related to the normal distribution. Since the parameters are small here, a normal approximation is not justified for the gamma. The answer is not an easy one and a general answer cannot be given. When one decides to use one case or the other, the accuracy of the approximation has to be estimated. This can only be done if the population distribution is known. If not, then a method similar to the one described for the maximum in Galambos (1978a) should be developed. No result seems to be available on this line.

4. References

Daniels, H.E. (1945). The statistical theory of the strength of bundles of threads. Proc. Royal Soc. A 183, 405-435.

Galambos, J. (1978). The asymptotic theory of extreme order statistics. John Wiley and Sons, New York.

Galambos, J. (1978a). Statistical aspects of extreme value theory. Preliminary Report.

Galambos, J. and S. Kotz (1978). Characterizations of probability distributions. Lecture Notes Series, Springer Verlag, Heidelb

Heyde, C.C. (1971). On the growth of the maximum queue length in a stable queue. Operations Res. 19, 447-452.

Kagan, A.M., Linnik, Yu. V. and C.R. Rao (1973). Characterization problems in mathematical statistics. John Wiley and Sons, New York.