NBSIR 74-602

# Efficient Methods of Extreme-Value Methodology

Julius Lieblein

Technical Analysis Division
Institute for Applied Technology
National Bureau of Standards
Washington, D. C. 20234

October 1974

Final Report

U.S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

NBSIR 74-602

# EFFICIENT METHODS OF EXTREME-VALUE METHODOLOGY

Julius Lieblein

Technical Analysis Division
Institute for Applied Technology
National Bureau of Standards
Washington, D. C. 20234

October 1974

Final Report

U.S. DEPARTMENT OF COMMERCE, Frederick B. Dent, Secretary
NATIONAL BUREAU OF STANDARDS, Richard W. Roberts, Director

Contents



Tables

Efficient Methods of Extreme-Value Methodology

Julius Lieblein

This report presents the essentials of modern efficient methods of estimating the two parameters of a Type I extreme-value distribution. These methods are an essential phase of the analysis of data that follow such a distribution and occur in the study of high winds, earthquakes, traffic peaks, extreme shocks and extreme quantities and phenomena generally. Methods are given that are appropriate to the quantity of data available—highly efficient methods for smaller samples and nearly as efficient methods for large or very large samples. Necessary tables are provided. The methods are illustrated by examples and summarized as a ready guide for analysts and for computer programming. The report outlines further work necessary to cover other aspects of extreme-value analysis, including other distribution types that occur in failure phenomena such as consumer product failure, fatigue failure, etc.

Key words: Distribution of largest values; efficient estimators; extreme values; linear unbiased estimators; statistics; Type I distribution.

## 1. Introduction and Purpose

An increasing number of applications involve analysis of what have come to be known as "extreme values". These follow a statistical distribution that is quite different from that which governs ordinary data considered to come from a normal or Gaussian distribution. Analysis of extreme-value data requires estimation of the parameters of the extreme-value distribution that gives rise to such data.

It is the purpose of this report to present the most improved version of the essentials of such methodology, and make it available to those carrying out extreme-value analysis or developing computer programs for such purposes. Not all aspects of extreme-value analysis are presented in this report, only those concerned with estimation of parameters described in Section 2; neither is much theory given. More detailed treatment and theoretical development may be found in the sources indicated herein. Good general surveys that include extensive lists of references and cover various approaches will be found in [3] and [9].

## 2. Best Linear Unbiased Estimator, Sample Size $\leq$ 16

### a. Type I Extreme-Value Distribution and Its Parameters

A set of data $x_1, x_2, \ldots, x_n$ is said to follow a Type I extreme-value distribution" if the set is an independent random sample from a population represented by the cumulative distribution function (c.d.f.)

$$\text{Prob}\,\{X \leq x\} = e^{-e^{-(x-u)/b}}, \quad -\infty < x < \infty,\ -\infty < u < \infty,\ 0 < b < \infty \qquad (1)$$

in which u represents the parameter of location and b the parameter of scale.

The designation "extreme-value distribution" is based on the following. If $y_1, y_2, \ldots, y_p$ are a random sample of data from, say, a normal distribution, then the average value of such a sample also has a normal distribution, whereas the extreme—the smallest or largest—has quite a different distribution, which of course depends on the amount of data, p. However, for increasing p, such distribution approaches one of three types of limiting forms called "asymptotic distribution of extreme values", or simply, "extreme-value distribution". The type occurring most frequently in NBS applications appears to be Type I for largest values, whose c.d.f. is given above, and to which attention will be limited in this report. (See Addendum for discussion of amount of basic data, p, and sample size, n).

### b. Estimation of Parameters

Fitting an extreme-value distribution to a given set of data $x_1$, $x_2$, ..., $x_n$, or what is the same thing, estimating the parameters of the distribution, means performing calculations on these data to obtain a numerical value (estimate) for u and for b. Many different types of functions, called "estimators", have been utilized for this purpose, with widely differing statistical properties.

It is generally agreed that an estimator function, $f(x_1, \ldots, x_n)$, should have the following two desirable properties:

(1) Unbiased. On the average, the estimator function, f, which is a random variable with its own distribution, should equal the parameter being estimated, say u, and similarly for a function, g, used to estimate b, i.e.,

$$E(f) = u, \qquad E(g) = b,$$

"E" denoting mathematical expection. Such estimators are said to be "unbiased".

(2) Minimum variance. Generally, as more and more data are taken from the extreme-value distribution, the values of the estimator function tend to be more and more concentrated about a constant value. If this constant value is the parameter being estimated,

then the estimates thus tend to become more and more accurate with increasing sample size. This property of concentration or dispersion is represented by the variance of the estimator function. For a fixed amount of data, n, different estimator functions will have different concentrations, and it is evidently desirable to have an estimator with smallest variance among all unbiased estimator.

Best linear unbiased estimator (BLUE).* The above two desired properties of estimators may be incorporated in the statement that the "best" estimator of a parameter is one which has minimum variance among the class of unbiased estimators.

Generally, a lower (non-zero) bound exists, called the "Cramér-Rao lower bound" (C.R.B.), for the variance of unbiased estimators of a parameter based on a given sample size, n (see [1]). This lower bound could thus be considered to be a standard against which to compare the variance of any proposed unbiased estimator, f. Their ratio thus cannot exceed unity, and is generally less.

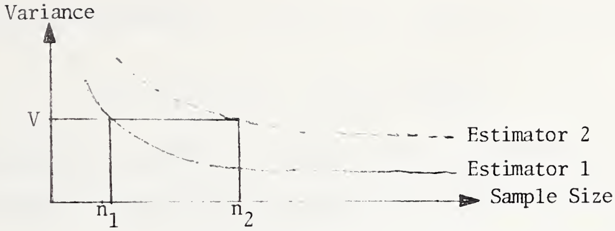
This ratio,

$$\text{(C.R.B.)}/(\text{variance of estimator } f) = (Q/n)/V_{(n)}(f) = \text{Eff}(f) \qquad (2)$$

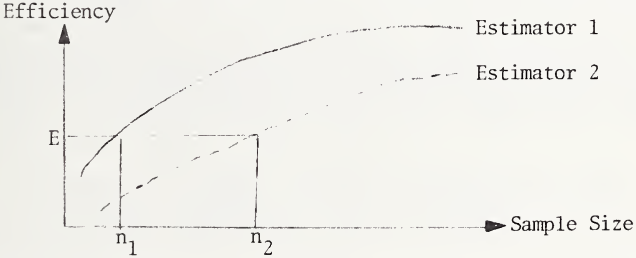
is designated the (Cramér-Rao) "efficiency" of the estimator f for estimating the designated parameter, where Q is independent of n.

The concepts of variance and efficiency have important implications for sample size. The variances of two (unbiased) estimators may be portrayed as in the following sketch, showing the tendency of the variance to diminish with increasing sample size, for two estimators of the same parameter.

*Plural is also "BLUE", not "BLUE's".




Estimator 2 is the less efficient estimator, having greater variance, at any sample size, than Estimator 1. It is apparent that a given variance V (or a given accuracy) requires a smaller sample size, $n_1$, with Estimator 1 than with Estimator 2, for which the sample size required is $n_2 > n_1$. The corresponding situation for efficiency of the same two estimators is indicated in the next sketch, showing its increasing tendency with increasing sample size. Estimator 2 is less efficient




Estimator 1 at a given sample size, n, and requires an increase in sample size from $n_1$ to $n_2$ in order to reach the same efficiency, E. Thus the variance and its related efficiency have a bearing on sample size, and the most economical estimator to use is the one with smallest variance, or, what is the same thing, with greatest efficiency.

In the case of sampling from the extreme-value distribution, an optimally efficient estimator (i.e., E = 1) has not been found for either n or b. However, recent investigations have shown how to obtain estimators, based on "order statistics", which have very high efficiencies.

Henceforth, the sample values will be assumed arranged in ascending order:

$$x_1 \leq x_2 \leq \cdots \leq x_n \tag{3}$$

The x's are then called "order statistics". The desired estimators of the extreme-value parameters then take the form of a linear function of the order statistics:

$$\hat{u}_{(n)} = \sum_{x=1}^{n} a_i x_i, \qquad \hat{b}_{(n)} = \sum_{i=1}^{n} b_i x_i \tag{4}$$

The BLUE are then obtained by finding numerical coefficients, $a_i$, $b_i$, for which each of the two linear functions of the order statistics is an unbiased estimator of u (or b) having minimum variance. These values, presented in Table 1 for $n \leq 16$, were obtained from [11]. Their efficiencies are given in Table 1a based on [8] which shows that relatively little improvement is obtainable by increasing sample size much beyond 10, at which efficiency, for u, has reached 98%, and for u is 85%.

3. Good Linear Unbiased Estimator, Sample Size Exceeding 16

The list of coefficients for increasing sample size beyond 16 would involve increasingly cumbersome tables to use. Instead, a method has been developed in [6] that produces the coefficients of an estimator, for any sample size, that is considered to be the best obtainable on the basis of the known coefficients of a best estimator (BLUE) for a smaller sample size. This method is as follows.

Let the known coefficients for a sample of m < n observations be denoted by $a_t$ and $b_t$, giving the following BLUE for estimating u and b from a sample of size n:

$$\hat{u}_{(m)} = \sum_{t=1}^{m} a_t x_t, \qquad \hat{b}_{(m)} = \sum_{t=1}^{m} b_t x_t \tag{5}$$

Then coefficients $a_i'$ and $b_i'$ for good estimators $\hat{u}'_{(m)}$ and $\hat{b}'_{(m)}$ to take the place of the a's and b's in (4) are given by (see [7]).

$$a_i' = \sum_{t=1}^{m} a_t \cdot (t/i)p(n,m,i,t), \tag{6a}$$

$$i = 1,2,\ldots,n$$

$$b_i' = \sum_{t=1}^{m} b_t \cdot (t/i)p(n,m,i,t), \tag{6b}$$

where

$$p(n,m,i,t) = \binom{i}{t}\binom{n-i}{m-t}\Big/\binom{n}{m}$$

is the hypergeometric probability function tabulated in [5].

A small example, for n = 6, m = 4, will illustrate how these calculations are made. A worksheet format is provided in Table 2, which may serve as a guide for computer programming in the case of large samples. Starting with the BLUE for sample size m = 4,

$$\hat{u}_{(4)} = \sum_{t=1}^{4} a_t x_t, \qquad \hat{b}_{(4)} = \sum_{t=1}^{4} b_t x_t,$$

with the $a_t$ and $b_t$ shown in the first two columns of the worksheet, the calculations represented in equations (6a) and (6b) are carried out and produce the "good" estimators

$$\hat{u}'_{(6)} = \sum_{i=1}^{6} a_i' x_i, \qquad \hat{b}'_{(6)} = \sum_{i=1}^{6} b_i' x_i$$

with coefficients (rounded to 5D) as follows, compared to those of the BLUE taken from Table 1 for n = 6:

| $i$ | $a'_i$ | $a_i$ (BLUE) | $b'_i$ | $b_i$ (BLUE) |
|---|---|---|---|---|
| 1 | 0.34067 | 0.35545 | -0.37241 | -0.45927 |
| 2 | .24184 | .22549 | - .11460 | - .03599 |
| 3 | .17038 | .16562 | .04190 | .07320 |
| 4 | .11902 | .12105 | .12333 | .12672 |
| 5 | .08051 | .08352 | .15591 | .14953 |
| 6 | .04759 | .04887 | .16586 | .14581 |

The variances and efficiencies for these estimators are given by

| Estimator | (Variance)/$b^2$ = V (from [7]) | (Cramer-Rao Lower Bound)/$b^2$ = B (see [6]) | Efficiency B/V |
|---|---|---|---|
| $\hat{u}_{(6)}$ | 0.19123 | 0.18478 | .96627 |
| $\hat{u}$ (BLUE) | .19117 | .18478 | .96657 |
| $\hat{b}_{(6)}$ | .13417 | .10132 | .75516 |
| $\hat{b}$ (BLUE) | .13196 | .10132 | .76781 |

According to reference [7], p. 548, the variance of this "expected value estimator" for samples of size n (with regard to either of the parameters) will not exceed (m/n) times the variance of the BLUE for samples of size m, where m is the size of subgroup. It follows that the efficiency will not be less than that of the BLUE for samples of size m, since (see [2]),

$$\mathrm{Eff}_{(n)} = \frac{Q/n}{V_{(n)}} \geq \frac{Q/n}{(m/n)\, V_{(m)}} = \frac{Q/m}{V_{(m)}} = \mathrm{Eff}_{(m)}.$$

Thus, the efficiency pertaining to subgroup size, m, might be taken as a rough measure of the efficiency associated with this method. Table 1a, which gives efficiencies for different sample sizes, then shows that little improvement is to be gained by using subgroup sizes greater than 10, a fact already noted above with regard to total sample size, n.

## 4. Estimators for Very Large Samples

The preceding methods become less practical for very large samples, say 50 or 100 or more. Fortunately, it has been found that it is not necessary to use all the sample values in order to obtain good estimates. In fact, as few as four suitably selected observations from a large sample of observations can yield almost as efficient estimates as using all n. Theoretical development and detailed accounts are presented in [4]. Some results from this paper are summarized in Table 3.

The n sample values are first arranged in order of increasing size

$$x_1 \leq x_2 \leq \cdots \leq x_n. \tag{7}$$

The idea is to find a small number, k, of the sample values ("quantiles")

$$x_{n_1}, x_{n_2}, \ldots, x_{n_k}, \tag{8}$$

which "space" the full set of n values in a manner analagous to the way a median spaces values corresponding to the fraction, $\lambda = .50$, with half of the observations above the median and the same fraction below—the median is then the "50-percent quantile" of the sample. The "optimum spacing" sought will then yield fractions

$$o < \lambda_1 < \lambda_2 < \ldots < \lambda_k < 1$$

with the $n_i$ in (8) being given by

$$n_i = \{n\lambda_i\}$$

where the brackets denote the next larger integer, e.g., $\{4.3\} = 5$, $\{16\} = 17$.

The $\lambda_i$ (for each k) for producing good estimates given in Table 3 were found in [4] by minimizing expressions for variances simultaneously for estimators of the parameters u and b by means of asymptotic normal theory results in large samples, under the constraint of unbiasedness. This procedure also produced the other quantities in Table 3. The $a_i^*$ and $b_i^*$ are the coefficients that give the good estimators for u and b based on the k selected observations, k = 2, 3, 4:

$$\hat{u}^* = \sum_{i=1}^{k} a_i^* x_{n_i}$$

$$\hat{b}^* = \sum_{i=1}^{k} b_i^* x_{n_i}$$

Asymptotic efficiency is the ratio of asymptotic variance to the Cramér-Rao lower bound. This variance, while not obtained from the exact variances calculated for the indicated order statistics of the sample of n (which would require very extensive calculation) is based on normal theory which becomes more and more exact as n increases indefinitely, and gives surprisingly close results even for reasonable-size samples. Thus, asymptotic efficiency is a usable measure of goodness to compare with exact efficiency for those sample sizes where the latter is available. The comparison is made in Table 3 with BLUE for sample size 16,* where the ratio is obtained to the exact values given in Table 1a. The relative efficiency of the optimum-spaced estimators thus obtained for k = 4 selected observations, is more than 90 percent as good for the estimator $\hat{u}^*$, and almost 85 percent as good for the estimator $\hat{b}^*$, as the efficiency of the best linear unbiased estimator obtainable for n = 16. These high ratios justify calling these starred values good estimators, although not best for the sample size n. What is very useful about them is that, for a given value of k, both the same spacings, $\lambda_i$, and the same coefficients, $a^*_i$

*The largest size shown in Table 1.

and $b_i^*$, can be used for <u>any</u> (large) n, and do not have to be recalculated for each different n. Thus, a very compact table such as Table 3 can serve for many cases. As an example of the use of Table 3 in forming estimators, suppose n = 100. Then, for the k = 4 selected ordered observations, $x_{n_1}$, $n_{n_2}$, $x_{n_3}$, $x_{n_4}$, the spacing, using the $\lambda_i$, is that given by $n_1 = \lceil .03(100) \rceil = 4$, $n_2 = \lceil .25(100) \rceil = 26$, $n_3 = \lceil .63(100) \rceil = 64$, and $n_4 = \lceil .90(100) \rceil = 91$ (the cut brackets denoting the next larger integer), so that the estimators are

$$\hat{u}^* = .1893\ x_4 + .4566\ x_{26} + .2772\ x_{64} + .0768\ x_{91}$$

$$\hat{b}^* = -.3127\ x_4 - .1123\ x_{26} + .2607\ x_{64} + .1643\ x_{91}$$

These four selected observations thus yield estimators that are almost as good as the best available from a sample of 16 observations, namely (from Table 3) 91% as efficient as the BLUE for u, and 84% as efficient as the BLUE for b.

## 5. Examples

The first example illustrates calculation of a BLUE using, as data, artificial random numbers from a known extreme-value distribution. The second example applies the procedure for large samples to maximum wind data to illustrate the calculation of good linear unbiased estimators.

<u>Ex. 1.</u> The simulated data, representing a sample of n = 8 observations from the extreme-value distribution with u = 4, b = 1, are as follows:

| i | Unordered | Ordered, $x_i$ | Coefficients to Obtain BLUE $a_i$ | $b_i$ |
|---|---|---|---|---|
| 1 | 5.41 | 3.62 | 0.273535 | -0.394187 |
| 2 | 3.70 | 3.62 | .189428 | - .075767 |
| 3 | 3.97 | 3.70 | .150200 | .011124 |
| 4 | 4.39 | 3.82 | .121174 | .058928 |
| 5 | 4.66 | 3.97 | .097142 | .087162 |
| 6 | 3.62 | 4.39 | .075904 | .102728 |
| 7 | 3.82 | 4.66 | .056132 | .108074 |
| 8 | 3.62 | 5.41 | .036485 | .101936 |
| Check Sum | 33.19 | 33.19 | 1.000000 | - .000002 |

The coefficients $a_i$ and $b_i$ are taken from Table 1 for n = 8. Before calculating the estimates there are a few useful points to be noted. Check sums should be obtained for every column of values. The first two columns should give the same sum because reordering does not change the total. The coefficients $a_i$ should add as closely as possible to unity and the coefficients $b_i$ should add to zero. It is shown by theory that these are the conditions for the estimator to be unbiased. These checks are satisfied above. When they are not, one should look for an error in transcription or some other arithmetical or "human" cause.

The resulting numerical estimates of the parameters u and b are, respectively,

$$\hat{u} = \sum_{i=1}^{8} a_i x_i = 3.8724, \qquad \hat{b} = \sum_{i=1}^{8} b_i x_i = 0.4171.$$

If small calculations such as these are done on the modern hand-held or desk calculators that give sums of multipliers simultaneously with the sums of products then the above check sums will be found a very useful protection against error. As a result of this procedure it is estimated that the sample of 8 values comes from the statistical distribution or model (eq. (1))

$$\text{Prob}\{X \leq x\} = e^{-e^{-(x-3.8724)/.4171}}$$

From this one can estimate the probability that a given value of x will not be exceeded, or, conversely, find what x-value will be exceeded a small fraction, such as 1 percent, of the time. Detailed treatment of such matters, including variability of resulting estimates, involve aspects beyond the scope of the present report.

Ex. 2. The second example uses a set of maximum annual wind speeds from Chattanooga, Tenn., for the 21-year period 1944 through 1964. This example illustrates the simplified procedure for "very large" samples given previously in Section 4. Although the sample size n = 21 may not be considered "very large", it is adequate for illustrative purposes.

The 21 unordered values are (miles/hour):

| | | | |
|---|---|---|---|
| 53 | 62 | 49 | 59 |
| 40 | 45 | 50 | 45 |
| 49 | 63 | 57 | 45 |
| 53 | 63 | 57 | 42 |
| 41 | 67 | 52 | 41 |
| | | | 54 |
| | | Check sum = | 1087 |

Placed in ascending order they are the quantities $x_1, x_2, \ldots, x_{21}$; respectively:

| | | | |
|---|---|---|---|
| 40 | 45 | 52 | 57 |
| 41 | 45 | 53 | 59 |
| 41 | 49 | 53 | 62 |
| 42 | 49 | 54 | 63 |
| 45 | 50 | 57 | 63 |
| | | | 67 |
| | | Check sum = | 1087 |

Table 3 allows us to make good estimates by using only 2, 3, or 4 selected values out of the 21. Thus, for 2 values, the fractional distances to go in the ordered set of 21 is $\lambda_1 = .09$, $\lambda_2 = .73$, giving $n_1 = \{.09(21)\} = \{1.89\} = 2$; and $n_2 = \{.73(21)\} = \{15.33\} = 16$. Similarly, for k = 4 selected values, we would have $n_1 = \{.03(21)\} = 1$; $n_2 = \{.25(21)\} = 6$; $n_3 = \{.63(21)\} = 14$; and $n_4 = \{.90(21)\} = 19$. The calculations, based on the values from Table 3, are

$\underline{k = 2}$

| | | $a_i^*$ | $b_i^*$ |
|---|---|---|---|
| $x_2$ | = 41 | 0.5673 | -0.4837 |
| $x_{16}$ | = 57 | .4327 | .4837 |
| Ck. sum | 98 | 1.0000 | 0.0000 |

$$\hat{u}^* = \sum_{i=1}^{2} a_i^* x_{n_i} = 47.9232 \qquad \hat{b}^* = \sum_{i=1}^{2} b_i^* x_{n_i} = 7.7392$$

$\underline{k = 4}$

| | | $a_i^*$ | $b_i^*$ |
|---|---|---|---|
| $x_1$ | = 40 | 0.1893 | -0.3127 |
| $x_6$ | = 45 | .4566 | - .1123 |
| $x_{14}$ | = 54 | .2772 | .2607 |
| $x_{19}$ | = 63 | .0768 | .1643 |
| Ck. sum | 202 | .9999 | 0.0000 |

$$\hat{u}^* = \sum_{i=1}^{4} a_i^* x_{n_i} = 47.9262 \qquad \hat{b}^* = \sum_{i=1}^{4} b_i^* x_{n_i} = 6.8672$$

## 6. Summary of Instructions for Fitting a Type I Extreme-Value Distribution

If we have a set of n maximum values that follow the "double exponential" distribution, eq. (1), then the procedure for fitting the distribution, i.e., estimating the parameters u, b, is to first arrange the n values in ascending order

$$x_1 \leq x_2 \leq \ldots \leq x_n,$$

and then proceed as follows, according to the magnitude of sample size n:

a. $n \leq 16$

Form the product sums

$$\hat{u} = \sum_{i=1}^{n} a_i x_i, \qquad \hat{b} = \sum_{i=1}^{n} b_i x_i,$$

using the coefficients in Table 1. These give the numerical values of the best linear unbiased estimators (BLUE) of the parameters u, b. Further details were given in Section 2b.

b. $16 < n <$ about 50

For this range of n, take subgroup size m = 10 and proceed as follows:

(1) Compute the coefficients

$$a_i' = \sum_{t=1}^{n} [a_t \cdot (t/i)p(n,m,i,t)]$$

$$i = 1,2,\ldots,n,$$

$$b_i' = \sum_{t=1}^{n} [b_t \cdot (t/i)p(n,m,i,t)]$$

where the $a_t$'s and $b_t$'s (m in number) are the coefficients of the BLUE for sample size m, given in Table 1, and

$$p(n,m,i,t) = \frac{\binom{i}{t}\binom{n-i}{n-t}}{\binom{n}{m}}$$

is the hypergeometric probability function tabulated in [5].

(2) From the above coefficients, compute the estimators

$$\hat{u}'_{(n)} = \sum_{i=1}^{n} a_i' x_i, \qquad \hat{b}'_{(n)} = \sum_{i=1}^{n} b_i' x_i.$$

These give the linear unbiased estimators that are the best available on the basis of the known BLUE coefficients for samples of size m. Further details were discussed in Section 3.

c. $n >$ about 50

This is the range of large to very large sample sizes. The estimators are obtained by using only a small number, k, of selected values out of the n, and the procedure is as follows:

(1) Take $k = 4$ and obtain $\lambda_i$ and coefficients $a_i^*$ and $b_i^*$ from Table 3, $i = 1, 2, \ldots, k$.

(2) Compute $n_1, n_2, n_3, n_4$ with $n_i = [n\lambda_i]$, the next larger integer to $n\lambda_i$.

(3) Compute the large-sample estimators

$$\hat{u}^* = \sum_{i=1}^{k} a_i^* x_{n_i}, \qquad \hat{b}^* = \sum_{i=1}^{k} b_i^* x_{n_i}.$$

These estimators are unbiased and have remarkably high efficiency considering that only a very small portion of the observations are used. Further details were discussed in Section 4.

## 7. Further Work

This report has presented a preliminary brief account of modern developments in extreme-value methodology that can yield optimum or good estimators of the parameters of the frequently occurring Type I distribution of largest values.

Current methodology concerns many additional aspects of extreme-value analysis which would be useful to the applied scientist. The possibilities for further work include the following:

a. Selection of extreme-value model. There are three types of extreme-value distributions, including the Type I considered in this report. Modern methodology includes

(1) quick graphical aids for selecting the most appropriate type; there are a variety of such methods that apply in different situations;

(2) computer methods for handling many sets of data. One study involved analysis of 38 sets of data on maximum wind speed with sample size varying from 15 to 53. Over half of the sets were found to follow the Type I distribution. Further work will compare these methods and determine the best ones to be recommended.

b. Fitting the selected distribution. The work of this report can be extended to include the other two types. They each involve a third parameter and require more elaborate methods of estimation. It will be essential to develop computer programming for comparing the various methods and testing their sensitivity to simplified approaches. For example, there is a whole class of estimators, called best linear invariant estimators (BLIE), which are biased, but have efficiencies exceeding those of the optimum unbiased estimators discussed in this report. Such estimators have found widespread use in reliability studies of many kinds [10]. This work will involve largest-value data but will be modified to include smallest-value data treatment in order to accommodate the Type III, or Weibull, distribution of smallest values that is also extremely useful in failure phenomena and reliability studies.

c. Statistical inference from fitted distribution. Specification of the distribution complete with the fitted parameters implies considerable information on future events, provided the model continues to hold (and the parameters do not change). Thus the probability of exceeding any given large value is determinable, which is of importance in study of high intensity winds, earthquakes, floods, and natural catastrophes. Other applications are extreme phenomena or peaks in general such as traffic peaks, as well as breaking strength and fatigue failure of materials, consumer product failure, etc. Conversely, values can be determined which will be surpassed with given probability, or risk, which is of evident value in these fields.

Table 1. Coefficients of Best Linear Unbiased Estimators (BLUE) for Type I Extreme-Value Distribution

| n | i | $a_i$ | $b_i$ |
|---|---|---|---|
| 2 | 1 | 0.916 373 | -0.721 348 |
| | 2 | .083 627 | .721 348 |
| 3 | 1 | .656 320 | - .630 541 |
| | 2 | .255 714 | .255 816 |
| | 3 | .087 966 | .374 725 |
| 4 | 1 | .510 998 | - .558 619 |
| | 2 | .263 943 | .085 903 |
| | 3 | .153 680 | .223 919 |
| | 4 | .071 380 | .248 797 |
| 5 | 1 | .418 934 | - .503 127 |
| | 2 | .246 282 | .006 534 |
| | 3 | .167 609 | .130 455 |
| | 4 | .108 824 | .181 656 |
| | 5 | .058 350 | .184 483 |
| 6 | 1 | .355 450 | - .459 273 |
| | 2 | .225 488 | - .035 992 |
| | 3 | .165 620 | .073 199 |
| | 4 | .121 054 | .126 724 |
| | 5 | .083 522 | .149 534 |
| | 6 | .048 867 | .145 807 |
| 7 | 1 | .309 008 | - .423 700 |
| | 2 | .206 260 | - .060 698 |
| | 3 | .158 590 | .036 192 |
| | 4 | .123 223 | .087 339 |
| | 5 | .093 747 | .114 868 |
| | 6 | .067 331 | .125 859 |
| | 7 | .041 841 | .120 141 |
| 8 | 1 | .273 535 | - .394 187 |
| | 2 | .189 428 | - .075 767 |
| | 3 | .150 200 | .011 124 |
| | 4 | .121 174 | .058 928 |
| | 5 | .097 142 | .087 162 |
| | 6 | .075 904 | .102 728 |
| | 7 | .056 132 | .108 074 |
| | 8 | .036 485 | .101 936 |
| 9 | 1 | 0.245 539 | -0.369 242 |
| | 2 | .174 882 | - .085 203 |
| | 3 | .141 789 | - .006 486 |
| | 4 | .117 357 | .037 977 |
| | 5 | .097 218 | .065 574 |
| | 6 | .079 569 | .082 654 |
| | 7 | .063 400 | .091 965 |
| | 8 | .047 957 | .094 369 |
| | 9 | .032 291 | .088 391 |
| 10 | 1 | .222 867 | - .347 830 |
| | 2 | .162 308 | - .091 158 |
| | 3 | .133 845 | - .019 210 |
| | 4 | .112 868 | .022 179 |
| | 5 | .095 636 | .048 671 |
| | 6 | .080 618 | .066 064 |
| | 7 | .066 988 | .077 021 |
| | 8 | .054 193 | .082 771 |
| | 9 | .041 748 | .083 552 |
| | 10 | .028 929 | .077 940 |
| 11 | 1 | .204 123 | - .329 210 |
| | 2 | .151 384 | - .094 869 |
| | 3 | .126 522 | - .028 604 |
| | 4 | .108 226 | .010 032 |
| | 5 | .093 234 | .035 284 |
| | 6 | .080 222 | .052 464 |
| | 7 | .068 485 | .064 071 |
| | 8 | .057 578 | .071 381 |
| | 9 | .047 159 | .074 977 |
| | 10 | .036 886 | .074 830 |
| | 11 | .026 180 | .069 644 |
| 12 | 1 | .188 361 | - .312 840 |
| | 2 | .141 833 | - .097 086 |
| | 3 | .119 838 | - .035 655 |
| | 4 | .103 673 | .000 534 |
| | 5 | .090 455 | .024 548 |
| | 6 | .079 018 | .041 278 |
| | 7 | .068 747 | .053 053 |
| | 8 | .059 266 | .061 112 |
| | 9 | .050 303 | .066 122 |
| | 10 | .041 628 | .068 357 |
| | 11 | .032 984 | .067 671 |
| | 12 | .023 894 | .062 906 |

Table 1. Coefficients of Best Linear Unbiased Estimators (BLUE) for Type I Extreme-Value Distribution (continued)

| n | i | $a_i$ | $b_i$ |
|---|---|---|---|
| 13 | 1 | 0.174 916 | - .298 313 |
| | 2 | .133 422 | - .098 284 |
| | 3 | .113 759 | - .041 013 |
| | 4 | .099 323 | - .006 997 |
| | 5 | .087 540 | .015 836 |
| | 6 | .077 368 | .032 014 |
| | 7 | .068 264 | .043 710 |
| | 8 | .059 900 | .052 101 |
| | 9 | .052 047 | .057 862 |
| | 10 | .044 528 | .061 355 |
| | 11 | .037 177 | .062 699 |
| | 12 | .029 790 | .061 699 |
| | 13 | .021 965 | .057 330 |
| 14 | 1 | .163 309 | - .285 316 |
| | 2 | .125 966 | - .098 775 |
| | 3 | .108 230 | - .045 120 |
| | 4 | .095 223 | - .013 039 |
| | 5 | .084 619 | .008 690 |
| | 6 | .075 484 | .024 282 |
| | 7 | .067 331 | .035 768 |
| | 8 | .059 866 | .044 262 |
| | 9 | .052 891 | .050 418 |
| | 10 | .046 260 | .054 624 |
| | 11 | .039 847 | .057 083 |
| | 12 | .033 526 | .057 829 |
| | 13 | .027 131 | .056 652 |
| | 14 | .020 317 | .052 642 |

| n | i | $a_i$ | $b_i$ |
|---|---|---|---|
| 15 | 1 | 0.153 184 | -0.273 606 |
| | 2 | .119 314 | - .098 768 |
| | 3 | .103 196 | - .048 285 |
| | 4 | .091 384 | - .017 934 |
| | 5 | .081 767 | .002 773 |
| | 6 | .073 495 | .017 779 |
| | 7 | .066 128 | .028 988 |
| | 8 | .059 401 | .037 452 |
| | 9 | .053 140 | .043 798 |
| | 10 | .047 217 | .048 415 |
| | 11 | .041 529 | .051 534 |
| | 12 | .035 984 | .053 267 |
| | 13 | .030 484 | .053 603 |
| | 14 | .024 887 | .052 334 |
| | 15 | .018 894 | .048 648 |
| 16 | 1 | .144 271 | - .262 990 |
| | 2 | .113 346 | - .098 406 |
| | 3 | .098 600 | - .050 731 |
| | 4 | .087 801 | - .021 933 |
| | 5 | .079 021 | - .002 167 |
| | 6 | .071 476 | .012 270 |
| | 7 | .064 771 | .023 168 |
| | 8 | .058 660 | .031 528 |
| | 9 | .052 989 | .037 939 |
| | 10 | .047 646 | .042 787 |
| | 11 | .042 539 | .046 308 |
| | 12 | .037 597 | .048 646 |
| | 13 | .032 748 | .049 860 |
| | 14 | .027 911 | .049 912 |
| | 15 | .022 969 | .048 602 |
| | 16 | .017 653 | .045 207 |

Table 1a. Efficiency of BLUE Listed in Table 1

| n | Cramér-Rao Efficiency for Estimators of $\hat{u}$ | Cramér-Rao Efficiency for Estimators of $\hat{b}$ |
|---|---|---|
| 2 | 0.84047 | 0.42700 |
| 3 | .91732 | .58786 |
| 4 | .94448 | .67463 |
| 5 | .95824 | .72960 |
| 6 | .96654 | .76782 |
| 7 | .94315 | .79606 |
| 8 | .97605 | .81785 |
| 9 | .97903 | .83519 |
| 10 | .98135 | .84938 |
| 11 | .98312 | .86119 |
| 12 | .98474 | .87121 |
| 13 | .98600 | .87982 |
| 14 | .98706 | .88725 |
| 15 | .98800 | .89381 |
| 16 | .98880 | .89961 |

Table 2. Worksheet for calculating coefficients $a_i'$, $b_i'$, $i = 1,2,\ldots,n$, for good extreme-value estimators based on coefficients $a_t$, $b_t$, $t = 1,2,\ldots,m<n$ of BLUE for smaller sample size, $n = 6$, $m = 4$:

$$a_i' = \sum_{t=1}^{4} a_t\, r(i,t), \quad b_i' = \sum_{t=1}^{4} b_t\, r(i,t), \text{ with } r(i,t) = (t/i)\, p(6,4,i,t)$$

| t | $a_t$ (From Table 1, for n = 4) | $b_t$ | t/i | p(6,4,i,t) = p | (t/i) p | t/i | p(6,4,i,t) | (t/i) p |
|---|---|---|---|---|---|---|---|---|
| | | | | i = 1 | | | i = 2 | |
| 1 | .510998 | -.558619 | 1 | .666667 | .666667 | .5 | .533333 | .266667 |
| 2 | .263943 | .085903 | 2 | 0 | 0 | 1.0 | .400000 | .400000 |
| 3 | .153680 | .223919 | 3 | 0 | 0 | 1.5 | 0 | 0 |
| 4 | .071380 | .248797 | 4 | 0 | 0 | 2.0 | 0 | 0 |
| Ck. Sum | 1.000001 | 0.000000 | 10 | .666667 | .666667 | 5.0 | .933333 | .666667 |
| | | | | $a_1' = .340666$, | $b_1' = -.372413$ | | $a_2' = .241844$, | $b_2' = -.114604$ |
| | | | | i = 3 | | | i = 4 | |
| 1 | (values above) | | .333333 | .2 | .066667 | .25 | 0 | 0 |
| 2 | | | .666667 | .6 | .400000 | .50 | .400000 | .200000 |
| 3 | | | 1.000000 | .2 | .200000 | .75 | .533333 | .400000 |
| 4 | | | 1.333333 | 0 | 0 | 1.00 | .066667 | .066667 |
| Ck. Sum | | | 3.333333 | 1.0 | .666667 | 2.50 | 1.000000 | .666667 |
| | | | | $a_3' = .170380$, | $b_3' = .041904$ | | $a_4' = .119019$, | $b_4' = .123335$ |
| | | | | i = 5 | | | i = 6 | |
| 1 | (values above) | | .20 | 0 | 0 | .166667 | 0 | 0 |
| 2 | | | .40 | 0 | 0 | .333333 | 0 | 0 |
| 3 | | | .60 | .666667 | .400000 | .500000 | 0 | 0 |
| 4 | | | .80 | .333333 | .266666 | .666667 | 1.000000 | .666667 |
| Ck. Sum | | | 2.00 | 1.000000 | .666666 | 1.666667 | 1.000000 | .666667 |
| | | | | $a_5' = .080507$, | $b_5' = .155913$ | | $a_6' = .047587$, | $b_6' = .165865$ |

Table 3. Optimum spacings ($\lambda_i$), coefficients, and ratios of asymptotic efficiencies (A.R.E.) of $\hat{u}^*$ and $\hat{b}^*$ to efficiencies of BLUE for n = 16, for k = 2, 3, or 4 selected observations $x_{n_1} \le x_{n_2} \le x_{n_3} \le x_{n_4}$, $n_i = \{n\lambda_i\}$

| k | | $x_{\{n\lambda_1\}}$ | $x_{\{n\lambda_2\}}$ | $x_{\{n\lambda_3\}}$ | $x_{\{n\lambda_4\}}$ | Ratio of A.R.E. to efficiency of BLUE for n = 16, for | $\hat{u}^* = \Sigma a_i^* x_{n_i}$ | $\hat{b}^* = \Sigma b_i^* x_{n_i}$ |
|---|---|---|---|---|---|---|---|---|
| | | | | | | A.R.E.: | .7335 | .5660 |
| 2 | $\lambda_i$ | .09 | .73 | | | Ratio: | .7418 | .6292 |
| | $a_i^*$ | .5673 | .4327 | | | | | |
| | $b_i^*$ | -.4837 | .4837 | | | | | |
| | | | | | | A.R.E.: | .8564 | .6709 |
| 3 | $\lambda_i$ | .06 | .45 | .85 | | Ratio: | .8661 | .7458 |
| | $a_i^*$ | .3550 | .5055 | .1395 | | | | |
| | $b_i^*$ | -.4456 | .1700 | .2756 | | | | |
| | | | | | | A.R.E.: | .9041 | .7589 |
| 4 | $\lambda_i$ | .03 | .25 | .63 | .90 | Ratio: | .9143 | .8436 |
| | $a_i^*$ | .1893 | .4566 | .2772 | .0768 | | | |
| | $b_i^*$ | -.3127 | -.1123 | .2607 | .1643 | | | |

References


1. Cramér, Harald, Mathematical Methods of Statistics, Princeton University Press, Princeton (1946), pp. 479-480.

2. Goldstein, N., "Random Numbers from the Extreme Value Distribution," Publications de l'Institut de l'Université de Paris, Vol. XII, fascicule 2 (1963) 137-158.

3. Harter, H. Leon, Order Statistics and Their Use in Testing and Estimation, Vol. 2, Aerospace Research Laboratories, Wash., D.C. (1969), 125-131.

4. Hassanein, K. M., "Estimation of the Parameters of the Extreme Value Distribution by Use of Two or Three Order Statistics," Biometrika, Vol. 56, 429-436 (1969).

5. Lieberman, G. and Owen, D., Tables of the Hypergeometric Probability Distribution, Stanford University Press, Stanford, California, (1961).

6. Lieblein, J., "A New Method of Analyzing Extreme-Value Data," NACA Tech. Note 3053, National Advisory Committee for Aeronautics, 1954, Table III(a).

7. McCool, John I., "The Construction of Good Linear Unbiased Estimates from the Best Linear Estimates for a Smaller Size," Technometrics, Vol. 7, 543-552 (1965).

8. Mann, N. R., "Results on Location and Scale Parameter Estimation with Application to the Extreme-Value Distribution," Report ARL 67-0023, Aerospace Research Laboratories, 1967, Table D.II.

9. Mann, N. R., "Point and Interval Estimation Procedures for the Two-Parameter Weibull and Extreme-Value Distributions," Technometrics, Vol. 10, 231-256 (1968).

10. Mann, N. R., Schafer, R. E. and Singpurwalla, N. D., Methods for the Statistical Analysis of Reliability and Life Data, John Wiley, New York (1974).

11. White, J. S., "Least-Squares Unbiased Censored Linear Estimation for the Log Weibull (Extreme Value) Distribution," J. Indust. Math. Soc., Vol. 14, 21-60 (1964).

## ADDENDUM

### Consideration of Sample Size

As pointed out in Section 2a above, there may be two sample sizes involved in an "extreme-value" situation, namely, (i) the amount of data, p, from which an extreme is taken; and (ii) the number of such extreme values, n, each extreme value in a sense representing a different set of data from the same population. It was also indicated that extreme-value analysis in theory depends upon an asymptotic situation, one where each amount of data, p, is "large". It would be useful to have some guide lines as to how much is large.

This has always been recognized to be one of the most difficult questions by workers in the field, from the pioneering efforts of Professor E. J. Gumbel [b] to the present day, and the question is still largely open.

As an illustration of how p and n may be considered, take Example 2 of the text above, dealing with maximum annual wind speeds. Records of wind speed are obtained by means of continuous-recording instruments throughout the year. Maximum values for five-minute periods are read and the largest of these is taken as the single maximum for the year. Thus the amount of data "in back of" the year's maximum is represented not by the 365 daily maxima, but by the much larger number of five-minute periods. Thus, p is at least several thousand, and there is little argument that this is a large enough amount of data. In the example, sample size n = 21, and this is the distinction between p and n. What may be considered additionally is whether the maxima of all the five-minute periods are from the same population, as this is a basis for the theoretical derivation of the extreme-value distribution. Here, discussion must be heuristic, in the absence of definitive studies. Long experience has made it seem likely that considerable departure from such a "stationarity" assumption can be tolerated without being detrimental to the application of extreme-value methods. It is as though the limiting process implicit in use of the extreme-value distribution operates to "smooth" out irregularities in the fundamental data, which may not even be accessible, and is not essential to application of the methods. At any rate, it is true that in the cited example, and in many other such cases analyzed by Simiu and Filliben [f] of the National Bureau of Standards, the maxima appear to follow the extreme-value distribution. A large number of other successful applications will be found in Professor Gumbel's definitive book [b], substantial portions of which have been updated by Mann, Schafer, and Singpurwalla [d].

It was stated above that the basic data behind the maxima may not even be available, yet this need not impair the application of extreme-value methods. Many successful applications where "basic data" are not available occur in the fields of reliability and in failure phenomena of materials, products, structures, and systems in general.

This can be illustrated by the simplest case, tensile strength of materials. It is epitomized by the saying "A chain is no stronger than its weakest link", characterized as the "weakest link" hypothesis. The idea is that a bar of, e.g., steel is made up of many hypothetical small segments, with tensile strengths represented by a probability distribution. When the bar is subjected to tensile stress, its failure stress is determined by the strength of the weakest of its "segments", which acts as a "weakest link". This idea is widely attributed to Griffith, in his theory of flaws enunciated in 1920 [a]; and the first statistical treatment based on this, to Pierce in 1926 [e].[1/]

A practical problem involving tensile strength is, knowing the strenght of a bar of given length, can we predict the strength of a bar, say, twice as long? The answer is generally Yes, by considering the larger bar as though it were made up of twice as many "small segments" as the half-size bar, i.e., the "amount of data, p" is twice as much, even though the value of p remains unknown and hypothetical.

Thus, many practical problems are successfully tractable without knowing the amount of "basic data, p". This may be a reason that little attention has been given to this aspect of the methodology.

The methodology presented in this report is primarily intended for workers who are already applying extreme-value methods to their problems and may be using methods of estimation that may not be the most efficient and best available in the present state-of-the-art. These workers have found, through other methods, outside the scope of the present preliminary report, that the extreme-value distribution appears applicable to their problems. This would imply, again on a heuristic basis, that the amount of basic data, p, known or hypothetical, is adequate.

At present, only the most general guide lines can be given as to how small a value of p would be adequate. Thus, if it is known that the

---

[1/] However, the writer of the present report has found evidence to justify earlier priority, namely, to Chaplin in 1880 (see Lieblein [c]).

basic data (in amount p) come from a simple exponential distribution (a failure model used in many reliability studies [d]), then p can be as small as 5 or even less. However, for some other distributions of basic data, such as the normal (Gaussian), it is known that such small values of p are inadequate. On the other hand, it is felt that generally a value of p of several hundred is probably sufficient, while for much smaller values the situation is doubtful. But, as already indicated, this is a matter of conjecture and requires considerable further research. In any case, if the user of extreme-value methods finds such methods applicable in a given case, he generally need not be concerned about the value of p.

## References


a. Griffith, A. A., "The phenomena of rupture and flow in solids," Phil. Trans. A, Vol. 221, 163 (1920).

b. Gumbel, E. J., Statistics of Extremes, Columbia University Press, New York (1958).

c. Lieblein, J., "Two early papers on the relation between extreme values and tensile strength," Biometrika, Vol. 41, Parts 3 and 4, 559-560 (December 1954).

d. Mann, N. R., Schafer, R. E., and Singpurwella, N. D., Methods for the Statistical Analysis of Reliability and Life Data, John Wiley, New York (1974).

e. Peirce, F. T., "Tensile tests for cotton yarns. V. "The weakest link--theorems on the strength of long and of composite specimens," Journal Textile Inst. Trans., Vol. 17, 355 (1926).

f. Simiu, E. and Filliben, J. J., "Statistical Analysis of Extreme Winds" (Technical Note in preparation).


USCOMM-NBS-DC

NBS-114A (REV. 7-73)

| U.S. DEPT. OF COMM. BIBLIOGRAPHIC DATA SHEET | 1. PUBLICATION OR REPORT NO. NBSIR 74-602 | 2. Gov't Accession No. | 3. Recipient's Accession No. |
|---|---|---|---|
| 4. TITLE AND SUBTITLE<br>Efficient Methods of Extreme-Volue Methodology | | | 5. Publication Date<br>October 1974 |
| | | | 6. Performing Organization Code |
| 7. AUTHOR(S)<br>Julius Lieblein | | | 8. Performing Organ. Report No.<br>NBSIR 74-602 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>NATIONAL BUREAU OF STANDARDS<br>DEPARTMENT OF COMMERCE<br>WASHINGTON, D.C. 20234 | | | 10. Project/Task/Work Unit No. |
| | | | 11. Contract/Grant No. |
| 12. Sponsoring Organization Name and Complete Address (*Street, City, State, ZIP*)<br>Same as No. 9 | | | 13. Type of Report & Period Covered<br>Final |
| | | | 14. Sponsoring Agency Code |

15. SUPPLEMENTARY NOTES

16. ABSTRACT (*A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.*)

This report presents the essentials of modern efficient methods of estimating the two parameters of a Type I extreme-value distribution. These methods are an essential phase of the analysis of data that follow such a distribution and occur in the study of high winds, earthquakes, traffic peaks, extreme shocks and extreme quantities and phenomena generally. Methods are given that are appropriate to the quantity of data available—highly efficient methods for smaller samples and nearly as efficient methods for large or very large samples. Necessary tables are provided. The methods are illustrated by examples and summarized as a ready guide for analysts and for computer programming. The report outlines further work necessary to cover other aspects of extreme-value analysis, including other distribution types that occur in failure phenomena such as consumer product failure, fatigue failure, etc.

17. KEY WORDS (*six to twelve entries; alphabetical order; capitalize only the first letter of the first key word unless a proper name; separated by semicolons*)

Distribution of largest values; efficient estimators; extreme values; linear unbiased estimators; statistics; Type I distribution.

| 18. AVAILABILITY [X] Unlimited | 19. SECURITY CLASS (THIS REPORT) UNCLASSIFIED | 21. NO. OF PAGES 31 |
|---|---|---|
| [ ] For Official Distribution. Do Not Release to NTIS | | |
| [ ] Order From Sup. of Doc., U.S. Government Printing Office Washington, D.C. 20402, SD Cat. No. C13 | 20. SECURITY CLASS (THIS PAGE) UNCLASSIFIED | 22. Price |
| [ ] Order From National Technical Information Service (NTIS) Springfield, Virginia 22151 | | |

USCOMM-DC 29042-P74