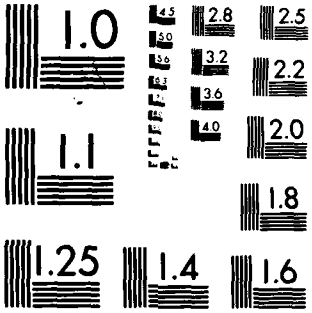
AD-A082 007 SACLANT ASW RESEARCH CENTRE LA SPEZIA (ITALY) F/G 12/1
THE DISTRIBUTION OF EXTREME VALUES OF SAMPLES DRAWN FROM A GAUS--ETC(U)
MAY 79 M J DAINTITH
UNCLASSIFIED SACLANTCEN-SM-124 NL

1 of 1

END
DATE
FILMED
4 80
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

SACLANTCEN Memorandum
SM - 124

SACLANT ASW
RESEARCH CENTRE
MEMORANDUM

THE DISTRIBUTION OF EXTREME VALUES OF SAMPLES DRAWN FROM A GAUSSIAN POPULATION

by

MAURICE J. DAINTITH

15 MAY 1979

NORTH
ATLANTIC
TREATY
ORGANIZATION

LA SPEZIA, ITALY

This document is released to a NATO Government at the direction of the SACLANTCEN subject to the following conditions:

1. The recipient NATO Government agrees to use its best endeavours to ensure that the information herein *disclosed*, whether or not it bears a security classification, is not dealt with in any manner (a) contrary to the intent of the provisions of the Charter of the Centre, or (b) prejudicial to the rights of the owner thereof to obtain patent, copyright, or other like *statutory* protection therefor.

2. If the technical information was originally released to the Centre by a NATO Government subject to restrictions clearly marked on this document the *recipient NATO Government* agrees to use its best endeavours to abide by the terms of the restrictions so imposed by the releasing Government.

Published by

INITIAL DISTRIBUTION

MINISTRIES OF DEFENCE	Copies
MOD Belgium	2
DND Canada	10
CHOD Denmark	8
MOD France	8
MOD Germany	15
MOD Greece	11
MOD Italy	10
MOD Netherlands	12
CHOD Norway	10
MOD Portugal	5
MOD Turkey	5
MOD U.K.	16
SECDEF U.S.	61

NATO AUTHORITIES	
Defence Planning Committee	3
NAMILCOM	2
SACLANT	10
SACLANTREPEUR	1
CINCWESTLANT/COMOCEANLANT	1
COMIBERLANT	1
CINCEASTLANT	1
COMSUBACLANT	1
COMMAIREASTLANT	1
SACEUR	2
CINCNORTH	1
CINCSOUTH	1
COMNAVSOUTH	1
COMSTRIKFORSOUTH	1
COMEDCENT	1
COMMARAIRMED	1
COMTWOATAF	1
CINCHAN	1

SCNR FOR SACLANTCEN	Copies
SCNR Belgium	1
SCNR Canada	1
SCNR Denmark	1
SCNR Germany	1
SCNR Greece	1
SCNR Italy	1
SCNR Netherlands	1
SCNR Norway	1
SCNR Portugal	1
SCNR Turkey	1
SCNR U.K.	1
SCNR U.S.	2
SECGEN Rep.	1
NAMILCOM Rep.	1
French Delegate	1

NATIONAL LIAISON OFFICERS	
NLO Denmark	1
NLO Germany	1
NLO Italy	1
NLO U.K.	1
NLO U.S.	1

NLR TO SACLANT	
NLR Belgium	1
NLR Canada	1
NLR Germany	1
NLR Greece	1
NLR Italy	1
NLR Norway	1
NLR Portugal	1
NLR Turkey	1

Total initial distribution	232
SACLANTCEN Library	10
Stock	37
Total number of copies	280

(14) SACLANTCEN [redacted]-SM-124

NORTH ATLANTIC TREATY ORGANIZATION

SACLANT ASW Research Centre
Viale San Bartolomeo 400, I-19026 San Bartolomeo (SP), Italy.

tel: national 0187 503540
international + 39 187 503540

telex: 271148 SACENT I

(9) Memorandum rept.,

(6)

THE DISTRIBUTION OF EXTREME VALUES OF SAMPLES DRAWN FROM A GAUSSIAN POPULATION.

by

(10) Maurice J. Daintith

(11) 15 May 1979

(12) 22

This memorandum has be n prepared within the SACLANTCEN Systems Research Division, as part of Project 09(75).

T. Mack
Division Chief

312950

TABLE OF CONTENTS

List of Figures

THE DISTRIBUTION OF EXTREME VALUES OF SAMPLES DRAWN FROM A GAUSSIAN POPULATION

by

Maurice J. Daintith

ABSTRACT

For some applications it is necessary to know the probability distribution functions of the maximum value (either absolute or else without regard to sign) occurring in large samples drawn from a parent stochastic population. Exact expressions are derived, and in addition useful approximations and limiting forms are presented; these also bring out clearly some interesting properties of the distributions.

INTRODUCTION

If samples of size n are drawn from a gaussian population, two extreme values can be identified within each sample.

Two types of extremum may be defined. The first is the true maximum value. This would be the appropriate parameter for use in, say, designing a sieve. The second type is the maximum deviation from the mean. This would be natural in investigating noise peaks. This memorandum deals mainly with the second of these distributions.

The purpose of this memorandum is to investigate the statistical properties of these extremes, present numerical values, and derive some approximate expressions for practical use.

1 NOTATION, FUNDAMENTAL EQUATIONS, AND BASIC EXPRESSIONS

1.1 Notation

The notation adopted is that used by Abramovitz and Stegun [1]. Throughout, the parent population will be taken as having zero mean and unit standard deviation.

1.2 Fundamental equations

The gaussian probatility density function is

$$Z(x) = \exp(-\tfrac{1}{2}z^2)/\sqrt{2\pi} \qquad \text{(Eq. 1)}$$

and cumulative probability functions are

$$P(z) = \int_{-\infty}^{z} Z\,dz \qquad \text{(Eq. 2)}$$

$$Q(z) = 1 - P(z) = \int_{z}^{\infty} Z\,dz \qquad \text{(Eq. 3)}$$

$$A(z) = 2P(z)-1 = 1-2Q(z) = 2\int_{0}^{z} Z\,dz\ , \qquad \text{(Eq. 4)}$$

(z positive)

so that

$$Z = \partial P/\partial z = -\partial Q/\partial z = \tfrac{1}{2}\partial A/\partial z \qquad \text{(Eq. 5)}$$

Further useful relationships are

$$\partial Z/\partial z = -zZ \qquad \text{(Eq. 6)}$$

and

$$P(-z) = Q(z) = 1-P(z)\quad . \qquad \text{(Eq. 7)}$$

1.3 Basic expressions

Consider a sample of n. $C_n(x)$, the cumulative probability function of x, where x is the greatest value of $|z|$ in the sample, can be immediately written down. $C_n(x)$ is the probability that all members of the sample lie between ± x. For each member of the sample, the probability is $P(x) - P(-x)$, which from Eq. 7 is $2P(x)-1$, and from Eq. 4 is $A(x)$. Therefore for a sample of n

$$C_n(x) = A^n(x) \quad . \qquad \text{(E} \quad \text{(Eq. 8)}$$

The probability density function $p_n(x) = \partial C_n(x)/\partial x$, which, using Eq. 5, becomes

$$p_n(x) = 2nZ(x)\,A^{n-1}(x) \quad . \qquad \text{(Eq. 9)}$$

The maximum likelihood value of x is found by setting $\partial p_n(x)/\partial x = 0$. Differentiating Eq. 9, and using Eq. 6, yields

$$2(n-1)Z(x_1) = x_1 A(x_1) \quad , \qquad \text{(Eq. 10)}$$

which is a transcendental equation yielding x_1 as a function of n.

The mean value $\bar{x}$, is, from Eq. 9,

$$\bar{x} = 2n \int_0^\infty xZ(x)\, A^{n-1}(x)\, dx \qquad \text{(Eq. 11)}$$

and higher moments may be similarly defined.

2 EXACT NUMERICAL RESULTS

2.1 Confidence limits

Given a confidence limit C , from Eq. 8

$$A(x_C) = C^{1/n} \qquad \text{(Eq. 12)}$$

and from published tables of $P(x)$ $[\frac{1}{2}(1+A(x))]$, or by using an appropriate approximation, [e.g. *[1]* Ch. 26 para 2.22], x_C may be readily determined.

Figure 1 gives plots of x_C as a function of n over the range $n = 10$ to $n = 10^6$, for values of C of 0.99, 0.95, 0.5, and 0.05. The value for C = 0.5 is, by definition, the median, x_m.

Two features of these curves are noteworthy. In the first place, for all of them x is a very slowly varying function of n. For example, the median value increases from 1.83 to only 4.97 over a five decade range of n.

Secondly the 50% value does not lie midway between the 5% and 95% values, indicating that the distribution is markedly skew. The extent of the skewness can be appreciated from Fig. 2, which shows plots of the probability density function (Eq. 9) for values of n of 10^2, 10^4 and 10^6. Equation 9 shows immediately why this should be so. The factor Z in that expression is multiplied by A^{n-1}. Although A itself is very close to unity over the whole range of interest, when it is raised to the high power of $n-1$ it varies very rapidly, depressing the probability for the lower values of x.

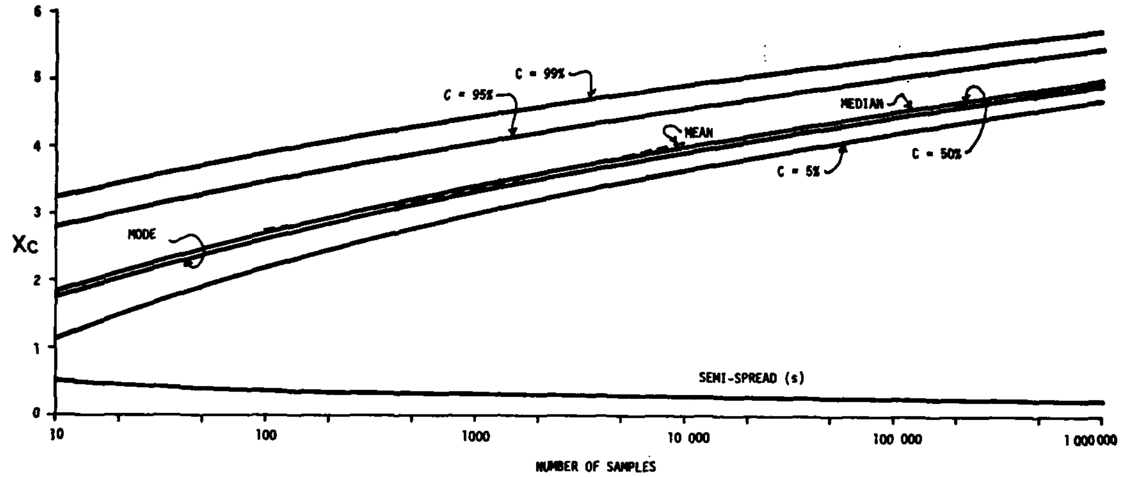


FIG. 1 CONFIDENCE LIMITS, MEAN, MODE AND SEMI SPREAD

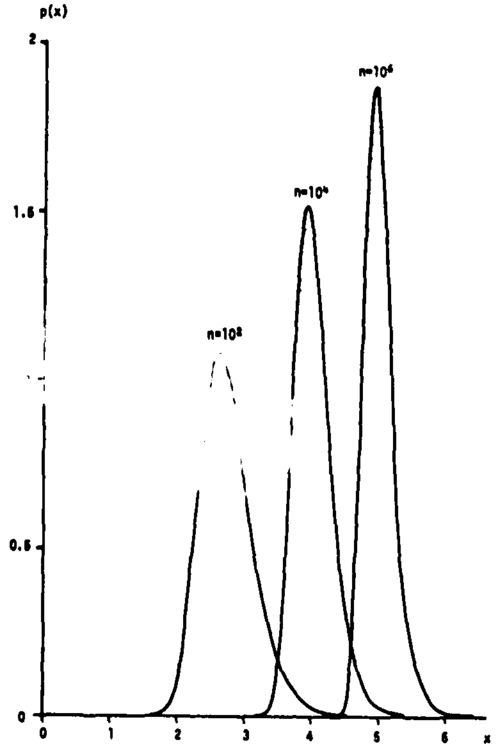


FIG. 2 PROBABILITY DENSITY FUNCTION

2.2 Maximum likelihood

Equation 10 may be written as

$$\exp(-\tfrac{1}{2}x_1^2) = x_1 A(x_1) \sqrt{\pi/2}/(n-1) \qquad \text{(Eq. 13)}$$

and in this form is readily solved iteratively, since the exponential term varies much more rapidly than the terms on the right-hand side. [Note that $A(x_1)$ is here only to the first power, and differs significantly from unity only for the smaller values of n; the lowest value, for $n = 10$, is 0.92. It is therefore justifiable to use a reasonable approximation for $A(x_1)$; a suitable one is given in *[1]* Ch. 26, para 2.16.]

The resulting values for the maximum likelihood value — the <u>mode</u> — are plotted in Fig. 1. They run nearly parallel to, and not far from, the median (50% confidence limit). This suggests that these central values, despite the non-gaussian distribution, are useful parameters.

2.3 Mean

Equation 11, for the mean value, cannot be evaluated analytically. Numerical integration would be a possibility, but it seemed preferable to search for an adequate approximation that would give some feeling for the way that $\bar{x}$ varies with n. An approximation sufficiently good over a part of the range ($n = 10^2$ to 10^4) is described in Ch. 3. The results are shown as the dotted line in Fig. 1. There is a reasonable coincidence with the median curve, once again supporting the usefulness of the central values.

2.4 Variance

In principle, the variance would be obtained by evaluating the mean square value of $x[=\bar{x}^2]$, and the mean value $\bar{x}$, and calculating the variance as

$$\overline{x^2} - \bar{x}^2$$

However, neither quantity can be calculated exactly, and since the difference turns out to be small compared with the two mean values, the approximation used for calculating the mean is not adequate. In addition, even in using the same type of approximation to calculate $\bar{x}^2$, the result appears as a small difference between two large quantities. For this reason, and since anyway the distribution is so skew, this approach was abandoned.

Instead, I have used a quantity which I refer to as the *semi-spread*, s , defined as follows.

Using Eq. 12, compute x for the confidence limits of $C = 0.84134$ and 0.15866. Half the difference between these x-values is defined as the semi-spread, s. If the distribution had been gaussian, s would in fact have been the standard deviation. In any event, it defines a range within which 68% of the x-values lie.

The quantity s is plotted in Fig. 1. It decreases with n , although very slowly, demonstrating the central tendency in the distribution (as evident in Fig. 2). The small value of s compared with, say, x_1 or x_m shows that these latter values are reasonably well-defined, and therefore of potential value.

Some confirmation that the semi-spread s is a meaningful value in defining the precision of the distribution was obtained by finding the ratio of the range between the 5% and 95% values of x to the whole spread $2s$. Over the whole range of n from 10 to 10^6, this ratio was nearly constant (1.66 to 1.70), indicating that the shapes of the curves were sensibly the same. (It is of interest that the value of the ratio for a truly gaussian distribution is 1.64, not markedly different from the figures quoted above.)

3 APPROXIMATIONS

The aim of finding simple approximat[illegible] he statistical quantities previously calculated is partly to pr[illegible] quick method of calculation, and partly to give an easy understanding [illegible] he way in which the parameters vary with n.

3.1 Median

From Eq. 12, the median value $(c = \frac{1}{2})$ is determined by finding x_m from

$$A(x_m) = (\tfrac{1}{2})^{1/n} \qquad \text{(Eq. 14)}$$

Now over the range considered, A is not much less than unity, so that $Q(x_m) = \frac{1}{2}(1-A)$ is a small quantity.

Hence, since $A = 1-2Q$, $\ell n\, A \sim -2Q$, and, from Eq. 14,

$$-2Q \sim \frac{1}{n}\,\ell n 2, \qquad \text{(Eq. 15)}$$

or

$$Q(x_m) \sim \ell n 2/2n$$

The problem now is to find x_m from Eq. 15. In paras 2.22 or 2.23 of Ch. 26 of *[1]* it is shown that, with considerable accuracy, x_m is a function of $\sqrt{\ell n(1/Q^2)}$, so that, from Eq. 15, x_m^2 is a function of $\ell n(n)$. Plotting x_m^2 against $\ell n\ n$ produced a nearly straight line almost passing through the origin, suggesting that a crude approximation would be $x_m \propto \sqrt{\ell n\ n}$. An obvious attempt to find a better approximation would be to try a modified power law, viz: $x_m \propto (\ell n\ n)^\beta$, or, taking logarithms, $\ell n\ x_m = \propto + \beta \ell n \ell n\ n$, a linear relationship. A plot of $\ell n\ x_m$ against $\ell n \ell n\ n$ produced a straight line to within the limits of graph paper accuracy.

Accordingly, a linear regression curve was fitted to $\ell n\ x_m$ as a function of $\ell n \ell n n$ (using values of n of 10, 20, 50, 100, 200, ... 10^6). The result was

$$\ell n\ x_m = 0.1397 + 0.5593\ \ell n \ell n\ n \qquad \text{(Eq. 16)}$$

with the very high coefficient of determination of 0.99995.

Equation 16 yields

$$x_m = 1.1499(\ell n\ n)^{0.5593} \qquad \text{(Eq. 17)}$$

Equation 17 predicts the value of x_m over the whole range of n to an accuracy of better than 0.3% (i.e. of at most two units in the second decimal place).

It will be noted that this differs only slightly from the square root 'first guess'.

3.2 Mode

Since the modal curve so closely parallels the median (Fig. 1) it is natural to try the same type of approximation. The straight line regression fit yielded

$$\ell n\ x_1 = 0.0732 + 0.5795\ \ell n \ell n\ n$$

with a coefficient of determination of 0.99986, so that

$$x_1 = 1.0760(\ell n\ n)^{0.5795} \qquad \text{(Eq. 18)}$$

The errors from Eq. 18 are at most about 0.7% (about three units in the second decimal place).

3.3 Semi-spread s

The semi-spread is defined as $\frac{1}{2}(x_b - x_a)$, where

$$A(x_a) = C_a^{1/n} , \quad A(x_b) = C_b^{1/n}$$

$C_a = 0.15866$, $C_b = 0.84134$. Writing $A = 1-2Q$ and approximating, as above,

$$Q(x_a) \sim (\ell n C_a)/n, \quad Q(x_b) \sim (\ell n C_b)/n.$$

Now try the very crude approximation

$$x^2 \sim -2\ell n Q \quad ,$$

whence

$$x_b^2 - x_a^2 \sim 2\ \ell n(Q_a/Q_b) \sim 2\ \ell n[(\ell n C_a)/\ell n C_b] \ = \ 4.73 \quad .$$

But $x_b - x_a = 2s$, and $x_b + x_a \sim 2x_m$, whence $sx_m \sim 1.18$.

This relationship is not too badly astray, since, using the calculated values, it turns out that sx_m varies only from 0.92 to 1.13 over the whole range of n, while s itself varies by a factor of 2.2.

The obvious suggestion is therefore to try an expression of the form $s = \alpha\ (\ell n\ n)^{\beta}$, in the same way that x_m was fitted. As might be expected, the approximation is worst for small n, and it appeared more satisfactory to omit the value $n = 10$, starting from $n = 50$. The result, with a coefficient of determination of 0.9989, is:

$$s = 0.787\ (\ell n\ n)^{-0.469}$$

in which the errors are $< 1\%$ for $n > 50$, becoming 2% for $n = 20$, and 5.5% for $n = 10$.

3.4 Mean

The evaluation of Eq. 11, repeated here for convenience, presents some difficulties

$$\bar{x} = 2n \int_0^\infty xZ(x)\ A^{n-1}(x)dx \quad . \qquad \text{(Eq. 11)}$$

Straightforwardly, one would endeavour to find an approximation for $A(x)$ that would allow Eq. 11 to be integrated. However, A is raised to a very high power for large n, and the errors in A^{n-1} would be intolerable. However, Eq. 11 may be transformed into a more malleable form as follows. Noting that $xZ = -\partial Z/\partial x$ [Eq. 6], Eq. 11 may be integrated by parts, yielding

$$\bar{x} = -2nZ\, A^{n-1} \int_0^\infty + 2n(n-1) \int_0^\infty ZA^{n-2} \frac{dA}{dx}\, dx$$

The first term vanishes at both limits, and in the integral it is legitimate to use A as the variable of integration (since A is a monotonic function of x).

Thus $\bar{x} = 2n(n-1) \int_0^1 Z\, A^{n-2}\, dA$ (Eq. 19)

Now in Eq. 19 if one can find an approximation for Z, as a function of A , that is correct to a few percent over the region of importance (which from Fig. 2 is confined to a restricted range of values), $\bar{x}$ can be determined to about the same accuracy (since Z is raised to the first power only).

The problem is, then, to find an approximation to Z as a function of A. Not only must it be of adequate accuracy, but it must also be of a form permitting Eq. 19 to be integrated. For example, an approximation can be found for x in terms of $\ell n\, Q$ i.e. in terms of $\ell n(1-A)$, and since $Z = 1/\sqrt{2\pi}\, \exp(-\tfrac{1}{2}x^2)$, this will give Z as a somewhat complicated exponential function of $\ell n(1-A)$. But this does not lead to an integrable form.

A possible solution is suggested as follows.

Para 2.4 of Ch. 26 in *[1]* gives an upper bound (which is also quite a good approximation) to P(x) , viz:

$$P(x) = \tfrac{1}{2}[1 + (1-\exp(-2x^2/\pi))^{\frac{1}{2}}]$$

or

$$A^2(x) = 1 - \exp(-2x^2/\pi)$$

so that, if the inversion were permissible,

$$Z = (1/\sqrt{2\pi})\, \exp(-\tfrac{1}{2}x^2)$$

$$\sim (1/\sqrt{2\pi})\, [1-A^2]^{\pi/4}$$

However, this is a poor approximation. Over the range $2 < x < 4$, errors of as much as 13% in Z are given by this expression. It does suggest however that, at least over a limited range, one may approximate may approximate z in the form $Z \sim \alpha[1-A^2]^{\beta}$, where α and β are constants. Taking logarithms, this implies that x^2 is a linear function of $\ell n(1-A^2)$. It was found, in practice, that, over the range $2 < x < 4$,

$$-x^2/2 \sim 0.2532 + 0.9164\ \ell n(1-A^2),$$

or

$$Z = 0.5139(1-A^2)^{0.9164} \quad .$$

with errors $<3\%$ over the whole range.

Applying this approximation to Eq. 19 yields

$$\bar{x} \sim 2n(n-1) \times 0.5139 \int_0^1 (1-A^2)^{0.9164} A^{n-2}\, dA \quad .$$

Writing $A = t^{½}$

$$\bar{x} = 0.5139n(n-1) \int_0^1 (1-t)^{\beta} t^{(n-3)/2} dt, \qquad \text{(Eq. 20)}$$

where $\beta = 0.9164$.

But the integral is a known beta function

$$\bar{x} = 0.5139\ n(n-1)\ B(1+\beta, \frac{n-1}{2}) \quad ,$$

where

$$B(1+\beta,(n-1)/2) = \Gamma(1+\beta)\ \Gamma(n-1)/2 \big/ \Gamma(n+1+2\beta)/2 \qquad \text{(Eq. 21)}$$

Expression 21 can be considerably simplified, using Stirling's approximation for the two gamma functions involving n.

Neglecting terms of the order $1/n$ (and noting $\ell n(1-1/n) \sim -1/n$) ,

$$\ell n\ \Gamma(n-1)/2 \sim ½\ \ell n(2\pi) + (n/2-1)[\ell n(n-1) - \ell n\ 2] - n/2+½$$

$$\sim ½\ \ell n(2\pi) + (n/2-1)\ \ell n(n/2) -½ -n/2 +½$$

Similarly,

$$\ell n\ \Gamma(n+1+2\beta/2) \sim ½\ \ell n(2\pi)+(n/2+\beta\ )\ \ell n(n/2)+½\ (1+2\beta\ ) - n/2 -½-\beta.$$

whence $\ell n[\ \Gamma(n-1)/2)\ /\ \Gamma(n+1+2\beta)/2)]$

$$\sim -(1+\beta)\ell n(n/2) \quad ,$$

so that $B(1+\beta,(n-1)/2) \sim \Gamma(1+\beta)\ (n/2)^{-(1+\beta)}$.

and from Eq. 20

$$\bar{x} = 0.5139\ \Gamma(1+\beta)\ (n-1)\ n^{-\beta}\ 2^{\beta}\ .$$

$\Gamma 1+\beta = 0.9678$, so the final result is

$$\bar{x} = 1.877 \times\ (n-1)/n^{0.9164}$$

or, for large n , $\bar{x} \sim 1.877\ n^{0.0836}$

It is this equation that has been used over the range shown in Fig. 1 to calculate a rough value of $\bar{x}$. Errors of several units in the second decimal place would be expected.

4 THE ABSOLUTE MAXIMUM

If interest lies in the distribution of the maximum value of x, taking account of sign, a similar treatment to that for the modular values presents no extra problems. Evidently the cumulative distribution function would be $C(x) = P^n(x)$ (where x can now range between $+\infty$ and $-\infty$), and the distribution function would be, by differentiating, $p(x) = nP^{n-1}Z$.

However, it is possible to obviate the whole problem by the following considerations. Evidently, we are interested only in the range of values over which $p(x)$ differs significantly from zero, which we may plausibly expect will be the range over which $C(x)$ is not too far removed from 0.5. Furthermore, we would expect $P(x)$ and $A(x)$, for large n at any rate, to be not much smaller than unity; i.e. $Q(x)$ will be small.

Now since $P = 1-Q$, and $A = 1-2Q$, we may expand the two cumulative functions P^n and A^n by taking logarithms. Retaining two terms in the expansions

$$\ell n\ P \sim -\ Q+\tfrac{1}{2}Q^2$$

$$\ell n\ A \sim -2Q+2Q^2\ ,$$

whence

$$\ell n\ P^n \sim -\ nQ+\tfrac{1}{2}nQ^2$$

$$\ell n\ A^n \sim -2nQ+2nQ^2\ .$$

Substitute 2n in the expression for $\ell n\ P^n$

$$\ell n\ P^{2n} \sim -2nQ+nQ^2$$

and hence

$$\ell n[P^{2n}/A^n] = -nQ^2 \quad .$$

Now nQ is not too far removed from 0.5 and therefore at maximum is, say, unity;

so that $\ell n[P^{2n}/A^n]$ is of the order of $-Q$

or $P^{2n}/A^n \sim e^{-Q} \sim 1-Q$, and since Q is small $P^{2n} \sim A^n$.

The implication of this is that the distribution of absolute values is almost exactly the same as the distribution of modular values for a sample of half the size. In other words, the curves of Figs. 1 and 2 may be used to estimate absolute values if entered with n = half the number in the sample.

A few trials, in fact, showed that this holds true even down to n = 10 to an accuracy of at least three decimal places.

This is not an unexpected result, noting that, for each sample of n, selecting the greatest modulus is equivalent to selecting the largest value from either the positive values or the negative values, while the selection for the absolute maximum will involve generally only the positive values (the change, in a large sample, of the true maximum being negative can be neglected for all except very small values of x). Hence, for the absolute maximum we normally can ignore all the negative values, which will constitute about one-half of the total sample.

5 A GENERAL LIMITING FORM FOR LARGE SAMPLES

In the course of this small study, I observed a very interesting limiting form that holds for a large class of parent distributions. The observation is not, however, new, since it appears in *[1]* Ch. 26, para 1.30 as *"Extreme-Value (Fisher-Tippett Type 1, or doubly exponential)"*. The reference to the original work is not readily available, so that I do not know if the derivation I give below corresponds with that of Fisher and Tippett.

Consider a parent distribution Z(x), in which, initially, the only requirement is that Z approaches zero asymptotically as x tends to infinity. The cumulative probability P(x) is defined as usual, by

$$P = \int_{-\infty}^{x} Z\, dx \qquad \text{(Eq. 22)}$$

For the absolute maximum of a sample of n, the cumulative probability as before

$$C(x) = P^n(x) \qquad \text{(Eq. 23)}$$

and the probability distribution

$$p(x) = \partial C/\partial x \qquad \text{(Eq. 24)}$$

Define

$$y = \ell n\ C = n\ \ell n\ P \qquad \text{(Eq. 25)}$$

Since $C \rightarrow 1$, $y \rightarrow 0$ as $x \rightarrow \infty$

differentiating Eq. 25, and using Eqs. 22 and 24,

$$\partial y/\partial x = p/C = nZ/P \quad . \qquad \text{(Eq. 26)}$$

and, differentiating again

$$\begin{aligned} \partial^2 y/\partial x^2 &= (\partial p/\partial x)/C - p^2/C^2 . \\ &= n(\partial Z/\partial x)/P - nZ^2/P^2 \qquad \text{(Eq. 27)} \end{aligned}$$

Eliminating P between Eqs. 26 and 27 yields the two equations

$$\partial^2 y/\partial x^2 - (\partial \ell n\ Z/\partial x)(\partial y/\partial x) + (\partial y/\partial x)^2/n = 0 \qquad \text{(Eq. 28)}$$

and

$$(\partial p/\partial x)/C = (\partial y/\partial x)^2 + (\partial \ell n\ Z/\partial x)(\partial y/\partial x) + (\partial y/\partial x)^2/n \quad , \qquad \text{(Eq. 29)}$$

Now consider the maximum value of p (the modal value), and let this occur for $x = \alpha$ (this was denoted by x in previous sections). From Eq. 29 this will be given by

$$(\partial y/\partial x)\ (1-1/n) = -(\partial \ell n\ Z/\partial x) \quad . \qquad \text{(Eq. 30)}$$

Since Z decreases with x, by hypothesis, the right-hand side of Eq. 3 is positive, and will be written as $(1/\beta)$, where β is a positive quantity. Since we are considering very large n, the term $1/n$ on the left-hand side may be dropped, yielding

$$(\partial y/\partial x)_\alpha = -(\partial \ell n\ Z/\partial x)_\alpha = 1/\beta \quad . \qquad \text{(Eq. 31)}$$

Equation 28 may be written as

$$\partial^2y/\partial x^2 + (\partial y/\partial x)\ [-\partial \ell n\ Z/\partial x + (\partial y/\partial x)/n] = 0.$$

The limiting form of this equation for large n becomes

$$\partial^2y/\partial x^2 - (\partial y/\partial x)\ (\partial \ell n\ Z/\partial x) = 0 \qquad \text{(Eq. 32)}$$

Again, since $p(x)$ is significantly greater than zero over only a limited range, about $x = \alpha$, we may insert the central value of the slowly varying factor $(\partial \ell n\ Z/\partial x)$ in Eq. 32, i.e. from Eq. 31, the constant $-1/\beta$. Equation 32 thus becomes

$$\partial^2y/\partial x^2 + 1/\beta\ \partial y/\partial x = 0 \quad ,$$

which integrates immediately to

$$\partial y/\partial x = K \exp(-x/\beta) \quad .$$

To find K , put $x = \alpha$, then, using Eq. 31

$$1/\beta \ = \ K \exp(-\alpha/\beta),$$

yielding

$$\partial y/\partial x = 1/\beta \exp -[(x-\alpha)/\beta] \quad , \qquad \text{(Eq. 33)}$$

Equation 33 also integrates immediately to

$$y = - \exp -[(x-\alpha)/\beta] \quad , \qquad \text{(Eq. 34)}$$

there being no constant of integration since $y \to 0$ as $x \to \infty$.

Hence

$$C = e^y = \exp\left(-e^{-t}\right) \quad , \qquad \text{(Eq. 35)}$$

where $t = (x-\alpha)/\beta$.

But from Eq. 26

$$p = C\partial y/\partial x = (1/\beta) \exp(-t - e^{-t}) \qquad \text{(Eq. 36)}$$

Equation 36 is the expression given by Fisher & Tippett quoted in *[1]*.

From Eqs. 26 and 31, α is obtained from the transcendental equation

$$\ell n\ Z(\alpha)/\partial\alpha + n\ Z(\alpha)/P(\alpha) = 0 \qquad \text{(Eq. 37)}$$

and from Eq. 31

$$1/\beta = -(\partial \ell n\ Z(\alpha)/\partial\alpha) \qquad \text{(Eq. 38)}$$

If the distribution is symmetrical about the mean, the same type of procedure will yield the distribution of deviations from the mean, the results being

$$C = \exp(-e^{-t}) \qquad \text{(Eq. 39)}$$

$$p = (1/\beta)\ \exp(-t\ -e^{-t})\ , \qquad \text{(Eq. 40)}$$

where

$$t = (x-\alpha)/\beta \qquad \text{(Eq. 41)}$$

$$(\partial \ell n\ Z(\partial)/\partial\alpha) + 2n\ Z/(2P-1) = 0 \qquad \text{(Eq. 42)}$$

$$(1/\beta) = -(\partial \ell n\ Z(\alpha)/\partial\alpha) \qquad \text{(Eq. 43)}$$

With one exception, these are the same equations as for the absolute maximum. The exception is the value of $(\partial \ell n\ Z(\alpha)/\partial\alpha)$ (Eqs. 37 and 42; viz nZ/P and $2nZ/(2P-1)$ respectively. Since for very large n, α becomes very large, so that $P(\alpha)$ tends to unity, the first expression tends to nZ , and the second to $2nZ$.

This implies that the distribution for the absolute maximum has exactly the same form as for the maximum deviation from the mean, provided that for the former, the tables and curves are entered with $(n/2)$ instead of n . This confirms the finding in a previous section.

For a gaussian distribution, $Z = (1/\sqrt{2\pi})\ \exp(-\frac{1}{2}x^2)$, so that, from Eq. 43, $1/\beta = \alpha$.

Unfortunately, this limiting expression is not very accurate, even for $n = 10^6$. It could hardly be expected to be so, since it is independent of the parent distribution (within limits).

There is, however, the advantage that the mean, median, and variance are readily computable.

From Ref. 1,

$$\text{Mean } \bar{x} = \alpha + \gamma\beta \quad \text{(where } \gamma = \text{Euler's constant} = 0.577....)$$

$$\text{variance } \sigma^2 = (\pi\beta)^2/6$$

In addition, by putting $C = \frac{1}{2}$ in Eqs. 35 or 39,

$$\text{median } x_m = \alpha -(\ell n\ \ell n\ 2)\beta = \alpha + 0.3665\ \beta\ .$$

From this, it appears that the mean and the median are very nearly equal, and both are greater than the mode. This compares well with the curves of Fig. 1.

Finally, for a gaussian distribution, where $\beta = 1/\alpha$

$$\sigma\alpha = \pi/\sqrt{6} = 1.28$$

This may be compared with the approximation

$s x_m = 1.18$, shown earlier.

SUMMARY AND DISCUSSION

The main results are summarized as follows:

(1) For samples of n, where n ranges from ten to one million, the following approximations may be used to estimate the statistical parameters of the maximum deviation (the modulus)

(a) The most likely value (the mode) is given by

$$x_1 \sim 1.076(\ell n\ n)^{0.58}$$

(b) The median value is

$$x_m \sim 1.15(\ell n\ n)^{0.56}$$

(c) The semi-spread (half the range, which includes 68% of values) is

$$s \sim 0.79(\ell n\ n)^{-0.47}$$

(d) The probability distribution function $p(x)$ can be described by a universal curve relating

$sp(x)$ to $(x-x_1)/s$.

(2) For the absolute maximum the formulae and curves for the modulus may be used, by entering with n equal to half the number in the sample.

(3) A limiting form as $n \to \infty$, which holds for a large class of parent distributions $Z(x)$, is that

$$C = e^{-t}$$

$$p = \exp[-t - e^{-t}]$$

where

$$t = (x-\alpha)/\beta.$$

α , the modal value of x , is given by

$$\partial \ell n\ Z(\alpha)/\partial\alpha + nZ = 0 \quad \text{(absolute maximum)}$$

$$\partial \ell n\ Z(\alpha)/\partial\alpha + 2nZ = 0 \quad \text{(maximum deviate)}$$

$$1/\beta = -\partial \ell n\ Z(\alpha)/\partial\alpha$$

For this distribution

the mean $\bar{x} = \alpha + \gamma\beta$ (γ = Euler's constant = 0.577....)

the median $x_m = \alpha - (\ell n\ \ell n\ 2)/\beta$

the variance $\sigma^2 = (\pi\beta)^2/6$.

This expression is not as accurate, up to $n = 10^6$, as the empirical expressions in this study.

In discussing these results it must be remembered that the aim is pragmatic — to find easily handleable expressions for the distributions, not elegant theoretical results.

The most striking fact is the accuracy with which parameters of the distributions may be estimated to a respectable accuracy, over a very large range of sample size, by the use of very simple expressions. The derivation of the limiting form sheds some light on why this may be expected.

The second feature is that, perhaps contrary to intuition, the distribution of extreme values turns out to be predictable within quite narrow limits. The implication justifies the common use of confidence limits, particularly for large samples.

A third feature is perhaps of interest. In general, statisticians search for 'efficient' statistics (in the Fisher sense). That is, they prefer to deal with statistics that give the maximum possible information about the parent distribution. This is, of course, especially true of investigations in which the aim is, by experiment, to determine what the parent distribution is.

However, the inverted problem, that of finding a predictable statistic, does not necessarily mean that the best choice of statistic is an efficient one. Since an inefficient statistic gives little information about the parent distribution it follows, conversely, that in many instances lack of knowledge of the parent distribution is of relatively minor importance. This can be of value in system design.

The limit of this approach is, naturally, to find the currently fashionable 'distribution-free' parameters. Here, in a noise background, for example, it has been shown that, for large samples, the distribution of the maxima is sensibly independent of the nature of the parent distribution, and will be scaled according to the variance. This implies that, in detecting a signal against noise, the effect of noise spokes hardly depends on whether the noise is gaussian or of some other distribution: it can be predicted for all practical purposes merely from a knowledge of the mean intensity — a somewhat unexpected result.

REFERENCES

1. ABRAMOWITZ, M. and STEGUN, I.A. Handbook of Mathematical Functions, Applied Mathematics Series. 55. Washington, D.C. National Bureau of Standards, 1964.