THE STATISTICAL ESTIMATION OF EXTREME WAVES

by

NEIL GRANT MACKENZIE

B.Sc(Hons), University of Newcastle-upon-Tyne, 1973.

91

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

IN THE FACULTY OF GRADUATE STUDIES

in the Department

of

CIVIL ENGINEERING

We accept this thesis as conforming
to the required standard

The University of British Columbia

June, 1979

© Neil Grant MacKenzie, 1979

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the Head of my Department or by his representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Civil Engineering

The University of British Columbia
2075 Wesbrook Place
Vancouver, Canada
V6T 1W5

Date 2 August, 1979

DE-6 BP 75-511E

ABSTRACT

This thesis contains a review of existing statistical techniques for the prediction of extreme waves for coastal and offshore installation design. A description of the four most widely used probability distributions is given, together with a detailed discussion of the methods commonly used for the estimation of their parameters.

Although several of these techniques have been in use for several years, it has never been satisfactorily shown which are capable of yielding the most reliable predictions. The main purpose of this thesis is to suggest a practical method of solving this problem and achieving the best estimate.

The basic theory for the prediction of extreme values was described in detail by Gumbel (1958) who concentrated largely on the double exponential distribution which is named after him. An order to evaluate the quality of fit between this law and the data, Gumbel derived expressions which enabled one to plot confidence intervals to enclose the data. The method described in this thesis in partly an extension of Gumbel's work, and similar confidence interval methods are given for the remaining distributions, thus permitting direct comparisons to be drawn between their performances. The outcome of this is that the most reliable model of the data may be chosen, and hence the best prediction made.

The method also contains a curvature test which has been devised to facilitate computation and lead more directly to the end result. The particular form of the wave data, which is quite different from wind records, is also taken into consideration and a working definition of the sample tail is suggested.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

APPENDIX

LIST OF FIGURES

LIST OF FIGURES (continued) Page

LIST OF TABLES

ACKNOWLEDGEMENT

The author profits by this occasion to thank his supervisor, Dr. M. de St. Q. Isaacson, for the constant help and encouragement given. The typing and ingenious presentation of the text was the work of Ms. S. McLintock.

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

The technique of plotting collected wave data on probability paper in order to predict the probable magnitude of extreme values has a firm place in engineering design. Although its use is widespread, a general method of selecting the most suitable probability model has not been previously suggested. Considerable attention has been given in design to accurately predicting the effect of a selected design wave on a structure. However, the process for selecting a design wave still remains comparatively unreliable, and represents a weak link within the design process.

In engineering practice and current literature only four distributions are commonly used for this model. These are the LOGNORMAL, GUMBEL (Type I), FRETCHET (Type II), and WEIBULL (Type III) distributions. Each of these four distributions is actually a family of probability functions whose properties vary subtly as their parameters change in value. The first three are each defined by two parameters, conveniently called "shape" and "scale" parameters. The Type III distribution requires a third parameter for its definition. This is referred to as a "location" parameter and enables the Type III distribution to be used in two alternative forms. These are later designated the Type III_L and Type III_U distributions.

The success of the method described here is dependent upon a close empirical fit between the model and the data being achieved. Depending on the data used, there is often a tendency for one

distribution to be more suitable than the others, and hence able to give more reliable predictions. The systematic search for this distribution has received very little attention in the literature, and is the basic problem considered in this thesis.

Fisher and Tippet [1926] showed that there are only three asymptotic distributions, and that these describe the behaviour of the maximum value from any parent distribution as the sample size approaches infinity. Gumbel [1958] developed the Type I distribution to a considerable degree as a tool for flood prediction. As a result of his work, this distribution has, until recently, been the most widely used throughout the various applications of extremal statistics. Gumbel also popularised the method of moments, and to some extent the method of least-squares for estimating the parameters of the Type I distribution. The former method was generally adopted since it could be carried out by hand, whereas the latter required a computer program.

Thom [1954] developed the Type II distribution for wind analysis and suggested using the method of maximum likelihood for estimating the parameters. Both the Type I and Type II distributions have been adopted by meteorologists for the prediction of extreme wind speeds in the United States and elsewhere.

Jasper [1956] suggested the use of the lognormal distribution for describing the occurrence of significant wave heights. This has been commonly used by many authors including Draper [1963], though in recent years its popularity may have diminished slightly.

Following the successes with wind speed prediction, Thom [1971] went on to advocate the application of the Type II distribution to wave heights. He argued that this distribution was superior to the Type I since it had a lower bound of zero height. He plotted data taken from several Ocean Station Vessels based in the Pacific and Atlantic Oceans, achieving a good fit in some cases. The data from these vessels was mainly based on visual estimates, and the vessels themselves stationed in the deep waters of mid-ocean. However, it is not unreasonable to expect the distributions of waves in shallower, more restricted sites to be rather different from those described.

There are two alternative forms of the Type III distribution which are denoted by the Type III_L and Type III_U distributions respectively. The Type III_L (Weibull lower-bound) distribution of minima (see TABLE I) was used in combination with the method of moments by Gumbel [1954] for estimating the worst drought occurring in a river. This was a natural choice since it enabled the engineer to place a lower limit ε on the least flow ever possible in the river.

Bretschneider [1965] suggested the use of the Type III_L distribution for the short-term significant wave heights associated with a given storm. Hogben [1967] used data gathered in the mid-Atlantic to make a comparison between the Type III_L distribution and the lognormal distribution. He concluded that, although the latter gave a better fit for lesser wave heights, the Type III_L was superior for larger heights.

Battjes [1970], using instrument-recorded data from the mid-Atlantic and the Celtic Sea, found a strong departure in the data from the lognormal distribution for extreme wave heights. In this instance the lognormal distribution gave an over-estimation of wave height for a specified probability of exceedance. He found that the Type III_L distribution gave a superior fit when a small positive value was used for ε. Generally ε was less than one metre and represented the extent of background noise which was always found to be present.

St. Denis [1973] suggested that the Type III_U (Weibull upper-bound - see TABLE I) distribution should be used for the description of wave heights in situations where a physical upper limit in the height could be expected. A typical case might be that of shoaling, or of a distinctly limited fetch. No reports of its use have been found in the published literature, although its use has also been advocated by Borgman [1975]. This may be the result of difficulties surrounding parameter estimation. These are discussed in detail in CHAPTER 5 on the estimation of parameters.

Petrauskas and Aagaard [1971] described a computer method which enabled them to select the most suitable distribution for a data sample from eight chosen possibilities. These consisted of the Type I distribution together with seven Type III_L distributions, each with a different prescribed shape parameter. The process of parameter estimation then simplified to one of determining two rather than three parameters for each of the eight cases. This was then achieved by a direct least-squares approach. The resulting distribution was then plotted with "uncertainty intervals" to indicate the degree of error in prediction.

In a paper on rubble-mound breakwaters, Ouellet [1974], noted the wide variability in sample of wave data and the need for a consistent approach to predict from them. FIGURE 1, which is taken from his paper, shows five sets of data from different sources. It can be seen that in one case (Moffat Beach, Australia) the researcher did not fit a single straight line but used three straight sections. This implies that the sample was a mixture of data from three quite different lognormal populations. From the point of view of prediction this is quite undesirable, since only one-third of the sample could be used for long-term forecasts. Two other sets of data (Benghazi Harbour and Mangalore Harbour) develop pronounced curvature as the exceedance probability decreases. In CHAPTER 4, the role of this property is examined in detail.

In the next chapter the distributions are described in detail. Each subsequent chapter discusses a step in the derivation of a design wave, as indicated by their titles. The conclusion to this thesis describes a complete procedure and a worked example is provided for demonstration.

CHAPTER 2

THE DISTRIBUTIONS AND THEIR PROPERTIES

The distributions described here are the most commonly used for extreme wave prediction, and an outline of their properties is given.

2.1 The Lognormal Distribution

The Lognormal Distribution is derived by transforming a variable to its logarithm before applying the normal distribution. This results in a density which only exists for a positive variable, as shown in FIGURE 2. If Y is an $N(\mu,\sigma^2)$ variable, that is it possesses a Normal distribution with mean μ and variance σ^2, then $X = \exp(Y)$ is a lognormal variable with parameters μ and σ^2. The density is given in TABLE I with $\alpha \equiv \sigma$ and $\theta \equiv \mu$.

The popularity of this distribution amongst coastal engineers is largely due to its connection with the normal distribution, and its considerable flexibility rendered by the scale and shape parameters μ and σ^2 respectively. In other related fields, such as meteorology, the lognormal distribution is much less popular. It is possible that meteorologists feel justified in only using asymptotic distributions by the relative abundance of weather data.

Of the various distributions considered here, the lognormal is an exception in that it is not an asymptotic form. In other words, it does not limit its description to the tail of a parent distribution from which the body of data might be collected.

Lognormal Paper is constructed as follows:

a) The ordinate scale carries the Standard Normal Distribution critical points corresponding to the exceedance probability $Q(h)$. A critical point is the value of variate which defines the lower limit of area representing the exceedence probability $Q(h)$ under the density curve. The procedure has been described by Draper [1963].

b) The abscissa scale is simply the natural logarithm of the wave height.

True lognormal data will lie on a straight line whose slope and intercept will be determined by the two parameters α and θ^2.

2.2 Asymptotic Distributions of the Extreme Value

Generally the distribution of data occurring within two standard deviations of the mean value is well described by the parent distribution. However, in many cases (for example the NORMAL DISTRIBUTION) the areas within the tail, corresponding to comparatively rare events, are difficult to calculate with precision. This problem does not arise in practice since the distribution of the maximum value occurring within a sample from any parent family tends in distribution to one of the three asymptotic types as the sample size approaches infinity. The three types, namely Gumbel, Fretchet, and Weibull distributions, all have cumulative distribution functions

(TABLE 1) which may be evaluated by pocket-calculator instead of tables or a computer program.

In order to simplify the description of these distributions the following notation will be adopted:

α - shape parameter which determines basic shape of the density function.

θ - scale parameter controlling the density scale or spread along the variate axis.

ε - location parameter locating the position of the density function on the variate axis. In the special case of the Types III_U and III_L distributions, ε locates one end of the density function (see Figure 3).

2.3 Gumbel Distribution - Type I

The Type I distribution is the limiting form for maximum values taken from the exponential class of parent distributions, which include the Normal, Exponential and Gamma distributions. The probability function is defined in TABLE I, and the density function is sketched in FIGURE 2. This asymptotic distribution has become an accepted form for the prediction of extreme winds in the U.S.A. according to Simiu [1976]. It sets no upper limit on the intensity of the windspeed which may occur.

In common with the other two types, the GUMBEL PAPER has a linear ordinate scale given by

$$y = -\ln\{-\ln[1 - Q(h)]\} \qquad \ldots.(1)$$

where $Q(h)$ is the exceedance probability of a given wave height. The abscissa scale is simply the wave height, which need not be standardised before plotting.

2.4 Fretchet Distribution - Type II

The probability function of the Fretchet Distribution is defined in TABLE I and typical shapes of the density function are shown in FIGURE 2. The Type II asymptote is the limiting distribution of maximum values taken from the Cauchy class of parent distributions which are not commonly used in engineering because their means and variances do not always exist. The Cauchy class generally have densities which are functions of the reciprocal of the intensity (wave height). A useful property of the Type II asymptote is that its density decays more slowly than the other two asymptotes. This property has made the Type II distribution invaluable for the prediction of hurrican intensities in the U.S.A. (Simiu 1976).

Thom [1971] suggested that the Type II distribution is particularly suited to the description of wave heights. He argued that wave heights are bounded quantities since they cannot have negative values, and thus merit quite different treatment from temperatures and pressures, which are unbounded. The transformation from an unbounded variate to its extreme value is achieved by a translation, whereas a bounded variate is transformed by a change of scale. The Type II distribution may be considered as a Type I distribution in

which the variate has been transformed to its logarithm, thus giving it a change of scale.

Fretchet Paper has a linear ordinate scale given by

$$y = -\ln\{-\ln[1 - Q(h)]\}$$

as before. The linear abscissa scale is now given by

$$x = \ln\{h\} \qquad .....(2)$$

where h is the wave height.

2.5 Weibull Distribution - Type III

The Weibull probability functions are given in TABLE I, and typical density functions are shown in FIGURE 3. As already mentioned, two forms of the Weibull distribution are available. The upper bound distribution, Type III_U, is the distribution of maximum values taken from parent distributions with a finite tail length, such as the UNIFORM DISTRIBUTION. The lower bound distribution, Type III_L, is the distribution of minimum values from the same source. The Type III_L distribution has been used to a certain extent as an empirical tool for wave height prediction. The second form of Type III_U does not appear to have been widely applied to the problem considered here.

Type III_U paper has the same ordinate scale as Types I and II, which is given by

$$y = -\ln\{-\ln[1 - Q(h)]\}$$

Its abscissa scale is dependent upon one of the three parameters, and so varies from one set of data to another.

$$x = -\ln\{\varepsilon - h\} \qquad .....(3)$$

where ε is the maximum wave height ever possible and is finite.

Type III_L Paper has a different ordinate scale to that used for the other asymptotic distributions, and this is generally

$$y = +\ln\{-\ln Q(h)\} \qquad .....(4)$$

and the abscissa scale is again parameter dependent.

$$x = +\ln\{h - \varepsilon\} \qquad .....(5)$$

where ε is the smallest wave height possible and $\varepsilon \geq 0$.

CHAPTER 3

PROCEDURE FOR COLLECTING WAVE DATA

An ideal data source would consist of a continuous wave height recording over a period of several years. In practice, the collection of a continuous sample is not feasible. The accepted alternative is systematic intermittent sampling which consists of chart-recording over a short period of several minutes in each successive interval of a few hours. The recording period is often set at twenty minutes for a recording interval of three hours. Generally, the sea state changes slowly enough for this to be representative and a typical wave elevation recording is sketched in FIGURE 4.

Engineers concerned with the prediction of extreme winds generally use data which has been collected over a decade or more. This is usually available from nearby airports, which keep records of this length. It is rare for a coastal engineer to have an equally long record for wave heights at some locality. Wave heights are sensitive to physical influences such as fetch length and water depth. The engineer is often obliged to use recordings made at the location of interest in order to account for local effects, such as diffraction by coastline projections or refraction.

A further difficulty arises because the engineer rarely knows the location of the project some years before the design is started, and hence is often forced to work with a short record. A final problem arises because there are usually practical difficulties in operating wave recorders accurately over long periods of time since

they usually have to be rigged with a buoy and anchor. Experience shows that the chances of a malfunction or loss of an instrument deployed in this fashion is quite high.

In cases where it becomes necessary to initiate a local wave-recording program, the record will rarely cover more than a year or two, except perhaps in long term wave research projects. More often the time period studied will be one year. Although a shorter study period would be expected to introduce seasonal variations, there is reason to suggest that provided the winter months are covered in detail no serious information loss should occur (see SECTION 3.3). In this extreme case the parameter estimation should be carried out by a least-squares approach (SECTION 5.2 et seq.). If summer months are omitted, it is often relatively simple to confirm that no storms more severe than those measured during the winter occurred.

The result of a wave recording program would be a series of representative wave heights, one for each recording interval, e.g. the significant heights, or the maximum height measured. The method of converting a continuous record into a series of statistics is not described here and has been well documented by Tucker [1963].

The basic form for the presentation of wave data is the bivariate histogram or "scatter diagram". This consists of a table containing significant wave heights which are divided, by frequency of occurrence, into intervals of wave period. The total number of occurrances is equal to the number of wave records (including calms)

gathered during the study period, (e.g. one year). FIGURE 5 provides an example of this, and from such a diagram one may assemble a table of height classes and their frequencies by summing over the wave periods. The resulting data is then used for plotting.

Although the methods of prediction used for extreme wind speeds are very similar to those used for wave heights, there is one basic difference in approach which results from the much shorter wave period. Engineers concerned with the extreme wind speed usually have a sample containing one maximum speed for each year of the record. For the reasons just discussed, the wave prediction has often to be based on values occurring within a single year's recording. As one might expect then, the wave prediction must lack the degree of precision of a wind prediction and this is reflected in the width of the confidence intervals (see SECTION 7.8).

3.1 Forming The Sample

Sampling is a critical stage in any statistical analysis. Not only does it enable the statistician to reduce the vast universe of data into something which is both meaningful and manageable, but it also largely determines the shape of the questions which may be answered.

As far as a literature review could show, very few authors attempted any other method of forming the sample than the one used by Draper [1963] and summarised below:

a) Each short length of wave recording corresponding to a recording period is reduced to a significant height h_s and a zero crossing period T_Z. The

method proposed by Tucker [1961] is used. Each pair of statistics then applies to one recording interval of several hours.

b) To facilitate handling, a scatter diagram is prepared. This is a table of h_s against T_Z, both divided into classes, and each element is appropriately marked with a number of recording intervals, (see FIGURE 5).

c) Each wave height class (0 - 1.99 ft., 2 - 3.99 ft. etc.) is summed over all classes of T_Z to give marginal frequencies of height.

d) The probability that the significant wave height may exceed the lower limit of any class is then calculated as

$$Q(h) = \frac{\text{Number of height values} \geq h}{1 + \text{total number of height values}} \quad .....(6)$$

where h is the lower limit of each height class (0 ft., 2 ft. etc) and $Q(h)$ is [Prob $H > h$]

e) Paired values of h and $Q(h)$ may then be used to plot the lower limit of each class onto a probability paper (e.g. Type III_L paper).

This approach is commonly used in statistics and was developed to fit a distribution to the body of the sample. It is particularly useful when estimating by the method of moments. However, it is rather non-specific in the way it achieves a fit and does not always give the engineer the type of fit he requires.

Since the purpose of sampling wave heights is to arrive at reliable estimates of the rare occurance wave, it seems unreasonable to concentrate on achieving a good fit at the median of the sample. In fact the quality of fit for wave heights which are exceeded almost daily is quite irrelevant to the problem considered here. Clearly, a high quality fit in the vicinity of the tail of the parent distribution is required in order to predict events whose exceedence probabilities occupy this region.

It would be most helpful if the "tail" of the sample could be defined so that a distribution could be fitted directly to this portion of the data. In practical applications of statistics it is quite common to define the finite end of a tail as being a fixed number of standard deviations from the mean. The result is an empirical rule of the form: Lowest height within tail $\geq \bar{h} + a\, S_h$

where $\bar{h}$ is the sample mean height

S_h is the sample standard deviation

a is a constant

An alternative approach, which yields the constant 'a', is based on the method of calculating plotting positions and will be given in SECTION 3.3.

3.2 Determination of Plotting Positions

In order to plot the data on probability paper one must assign a fixed probability to each value in the sample. To do this the data is ordered according to height and the suffix m is used to denote its position or RANK. Thus $m = 1$ corresponds to the largest value and $m = n$ to the smallest of a sample containing n wave heights.

A formula which has gained wide acceptance for calculating the plotting position is:

$$Q(h_m) = 1 - P(h_m) = m/n+1 \qquad .....(7)$$

It has been shown by Gumbel [1958] that the expected probability for the mth ordered observation is given by $m/n+1$ and that this is independent of the distribution.

However, it has been demonstrated by Kimball [1960] and Gringorten [1963] that this formula tends to introduce a slight bias towards the distribution being estimated. Although it is possible to form unbiased forms for the distributions considered here, such expressions would vary according to the parameters. Since the parameters still have to be estimated, this would lead to either approximate forms or to an iterative procedure. The example given by Gringorten suggests that the bias introduced by the simple formula Eqn. 7 is small enough to be considered a second-order effect in comparison with those introduced by adopted different estimation methods or sampling procedures. The simple rule is invariably used for plotting and, for example, has been strongly recommended by

Borgman [1975]. It will be adopted throughout the present study and the effects of alternative formulae are not examined.

It should be noted that the rank value m is assigned to each individual wave height recorded but not directly to the height class limits. This may be seen in the worked example in CHAPTER 9. As a consequence of this the class limits in the example have the approximate ranks shown in TABLE III.

3.3. Definition of the Sample Tail

It has already been mentioned that for the prediction of extreme wave heights, a good fit in the tail of the data is of considerable importance, i.e. the distribution should give a good fit to the worst of the extreme measurements made. It is therefore convenient to define the sample tail. In SECTION 3.1 one simple method of defining the sample tail was mentioned. An alternative approach may be based on the fact that for all three asymptotic distributions a function of probability provides the ordinate scale for plotting. The ordinate of the mth wave height statistic is given by

$$y_m = -\ln\{-\ln[1 - Q(h_m)]\}$$

where $Q(h_m)$ is calculated as in SECTION 3.2.

A plot of this function against $Q(h_m)$ describes the distortion applied to obtain the scale along the ordinate and is shown in FIGURE 6(a). The gradient of the curve is given by

$$dy/dQ = \{(1-Q)\ \ell n(1-Q)\}^{-1} \qquad .....(8)$$

and this is plotted against $Q(h_m)$ in FIGURE 6(b).

As $Q(h_m)$ approaches the median value of 0.5, the gradient decreases and becomes almost constant for values less then, say, 0.1. The vertical distance between plotted points is controlled directly by this gradient and hence a lower limit for $Q(h_m)$ of 0.1 is chosen to locate the start of the tail.

The extent of the tail is then determined by the position of the sample wave height having a rank w

$$\text{where} \quad w = (n+1)/10 \qquad .....(9)$$

and n is the number of wave heights in record.

As a result of this procedure only 10% of the original sample is used for estimation. A proportion of the bulk of data, which is used in the fitting procedure should contain all measurements made during the summer months of lower storm activity. As a result, it no longer becomes necessary to rely on precise measurements during these periods of low storm activity. Thus gaps in the record for these months need not be serious, and often this may be confirmed by an inspection of local meteorological records. In the rare cases when a valuable piece of data is missed it may be possible to estimate the approximate number of recording intervals involved and their order within the sample. This has been suggested in connection with similar applications by Borgman [1961].

CHAPTER 4

INITIAL SELECTION OF PARAMETRIC FAMILY

To carry out a detailed analysis using each of the distributions in turn would be tedious. To base the final selection of a distribution solely on the width of the confidence intervals could be difficult and misleading, often resulting in solutions which appeared to fit well in the extreme tail but poorly elsewhere. In order to eliminate these procedures it is convenient to make use of a simple property which is shared to a differing degree by all distributions. Such a property is the curvature of a distribution when plotted on Type I paper.

Literature often shows noticeable curvature of data as it approaches the tail of the distribution. In many cases this curvature is detectible to some degree, e.g. FIGURE 1. Typical examples may be found in papers by Khanna and Andru [1974], and Ovellet [1974].

4.1 Curvature Properties

When a Type II distribution is represented on any other pair of axes than those used to form a Fretchet plot, the result will be a curved line. This principle applies to all the distributions considered, and forms the basis of the curvature test suggested here.

The comparison is made by examining the curvature of each type of distribution when plotted on Gumbel Paper (Type I). A typical result is shown in FIGURE 7 and the actual degree of curvature for each distribution is dependent upon the parameters used. The curvature is defined as the second derivative

$$\text{CURVATURE} = d^2y/dx^2 \qquad \ldots\ldots(10)$$

where $y = -\ln\{-\ln[1 - Q(h)]$

$x = h.$

The resulting slope and curvature relationships are summarized in TABLE II and are briefly described below.

TYPE I - remains a straight line

TYPE II - develops strong negative curvature, and the tail decays more slowly than any of the other distributions.

TYPE III_U - develops the strongest positive curvature which enables it to achieve a finite limiting wave height, (i.e. one which has an exceedance probability of zero).

NOTE: As the parameter α approaches infinity both the Types II and III_U become straight lines, i.e. Type I.

TYPE III_L - may develop both negative or positive curvature depending on the size of α. (see TABLE II). In the special case of $\alpha = 1$ the line becomes straight. The relative flexibility of the tail of this distribution makes it useful for spanning the gap between Types II and III_U, in the ranges where their tails become inflexible.

The Type III_L distribution has a very wide range of curvature, and might often be an acceptable choice even without the curvature test.

LOGNORMAL - behaves in a similar fashion to Type II, though developing relatively mild curvatures.

The curvature relationships for Types I, II, III_U and III_L are relatively simple to derive and are given in Appendices A.5 - A.7. However, the lognormal distribution's behaviour is most easily demonstrated graphically, FIGURE 21. Although some overlap in curvature is expected, particularly between the lognormal and the Type III_L, both are retained as possibilities.

In order to make an initial choice of the distributions to be studied in detail, three groups may be used.

POSITIVE GROUP - Types III_U, III_L ($\alpha > 1$)

STRAIGHT GROUP - III_L ($\alpha \simeq 1$), LOGNORMAL, I.

NEGATIVE GROUP - LOGNORMAL, III_L ($\alpha < 1$), II.

The curvature test does not provide any method of selecting a distribution from within one of the groups just mentioned. However, this may be achieved by using the method of confidence intervals which is described in CHAPTER 6.

4.2 The Curvature Test

A simple procedure for selecting one of the three groups may now be used.

i) Plot the tail of the data onto a Type I Gumbel Paper with axes shown in FIGURE 7. The plotting procedure is that described in SECTION 3.2.

ii) The presence, type and degree of curvature is then assessed by eye and leads to a choice of one of the three groups.

Since the data is assembled for plotting by the method of SECTIONS 3.2 and 3.3, all points should occur within the parent distribution's tail, and hence one's decision would be based upon an 'overall' curvature for all of the plotted points.

CHAPTER 5

METHODS OF PARAMETER ESTIMATION

Each of the four distributions considered here is actually a family of different distributions which have widely different properties depending upon the parameter values. Having chosen one of the four distributions as a likely model, it remains to find the parameter values which fit that distribution to the data the closest.

In both the fields of wave and wind prediction, three methods of estimation have been adopted.

i) Method of Moments

ii) Method of Least Squares

iii) Method of Maximum Likelihood

Each of these three methods may give a different estimate of the parameters based on the same sample. It should be noted that all three methods provide a 'point estimate' i.e. a simple parameter value for each sample. Strictly speaking then, the estimates are themselves random variables, though they are treated as if they are stationary. It is mentioned in passing that the method of fitting a line by eye can give comparable results to those obtained by the method of least squares.

A brief description of each method and its application to the distributions is now given. The estimated values of parameters are indicated by a hat "$\wedge$", and parameter notation is as used in TABLE 1.

5.1 Method of Moments

The method of moments operates by successively approximating the shape of the model distribution to that of the sample histogram. This is achieved by equating the first k moments to give one equation for each of the k parameters required. The first three or four moments tend to exert the strongest influence on the shape of a distribution and so the procedure often leads to a reasonable model. One disadvantage of this method is that it uses all the collected data and does not emphasise the role played by the distribution tail.

5.1.1 Gumbel Distribution - Type I

The Type I distribution has two parameters (see TABLE I). The derivation of the moments and their properties are given in Appendix A.1. The two equations resulting from this procedure are

$$\overline{H} = \hat{\varepsilon} + \gamma\,\hat{\theta} \qquad .....(11)$$

and

$$\overline{H^2} - (\overline{H})^2 = \frac{\pi^2}{6}\,\hat{\theta}^2 \qquad .....(12)$$

where

$$\overline{H} = \frac{1}{n}\,\Sigma_1^{\,n}\,h_m\ ;\quad \overline{H^2} = \frac{1}{n}\,\Sigma_1^{\,n}\,h_m^2 \qquad .....(13)$$

and γ is Euler's Number (0.57722)

$\wedge$ denotes an estimated parameter.

Hence estimated values of the two parameters are:

$$\hat{\varepsilon} = \overline{H} - 0.4501\sqrt{\overline{H^2} - (\overline{H})^2} \qquad .....(14)$$

and

$$\hat{\theta} = 0.7797\sqrt{\overline{H^2} - (\overline{H})^2} \qquad .....(15)$$

Both $\hat{\varepsilon}$ and $\hat{\theta}$ are random variables whose values depend upon the particular random sample used for their estimation. The method of moments estimator for ε has a variance which is only 5% larger than that obtainable by the more complicated method of maximum likelihood. However, this method gives an estimator for $\hat{\theta}$ with a variance which is 80% greater than that obtainable by the method of maximum likelihood, and thus the resulting value of $\hat{\theta}$ tends to be unreliable.

It should be noted that the moment estimators for Type I are relatively straight forward to use since this distribution has the same shape for all parameter values.

5.1.2 Fretchet Distribution - Type II

The Type II distribution again has two parameters. However, one of these, α, controls the basic shape of the distribution and this parameter must be estimated first. A commonly used method for determining the shape is to equate the skewness of the sample to that of the model. The skewness of a distribution is defined by:

$$\sqrt{\beta} = \mu_3/\mu_2^{3/2} \qquad \text{.....(16)}$$

where μ_2 and μ_3 are the second and third central moments. The skewness of the Type II distribution is shown in FIGURE 8 as a function of α. It can be seen that in the region $\alpha > 5$ the skewness becomes increasingly insensitive to α and is liable to provide an inefficient estimation.

The estimation of the two parameters is achieved by the following method:

i) The sample skewness is calculated as

$$\sqrt{\beta_s} = [\overline{H^3} - 3\overline{H}\,\overline{H^2} + 2(\overline{H})^3][\overline{H^2} - (\overline{H})^2]^{-3/2} \qquad .....(17)$$

where

$$\overline{H^3} = \frac{1}{n} \sum_{1}^{n} h_m^3 \qquad .....(18)$$

and $\overline{H}$, $\overline{H^2}$ are defined in Eqn. 13.

ii) The estimated value of the shape parameter $\hat{\alpha}$ is obtained from FIGURE 8.

iii) The scale parameter is then calculated

$$\hat{\theta} = \overline{H} / \Gamma\{1 - 1/\hat{\alpha}\} \qquad .....(19)$$

where $\Gamma\{\}$ is the gamma function and is available in tabulated form. e.g. Abramowitz and Stegun [1970].

5.1.3 Weibull Distribution - Type III

The Weibull distribution requires the estimation of three parameters α, ε and θ. Since α controls the basic shape of the density function this parameter takes precedence in the estimation process. The skewness, which is the ratio of the second and third central moments, Eqn. 16 is found to provide satisfactory estimations, as for the Type II distribution, for $\alpha < 20$, but for higher values it becomes asymptotic and loses its sensitivity to α. This presents no real problem since the range of sensitivity has been found to be more than adequate for this study. The skewness is shown plotted against α in FIGURE 9.

Since there are three parameters to estimate, the method of moments requires equations involving the first three moments. These may be central moments or moments about the origin (or a combination of the two types). The derivation of the Weibull moments is given in Appendix A.3.

The method then reduces to the following steps:

i) Calculate the sample skewness from Eqns. 13, 17 and 18.

ii) The estimated value of the shape parameter $\hat{\alpha}$ is obtained from FIGURE 9.

It should be noted that for the Type III_U distribution the sign of the skewness should be changed before executing ii). This adjustment is not required for the Type III_L distribution.

iii) Solve for $\hat{\theta}$ from

$$\hat{\theta} = \{[\overline{H}^2 - (\overline{H})^2][\Gamma(1 + 2/\hat{\alpha}) - \Gamma^2(1 + 1/\hat{\alpha})]^{-1}\}^{\frac{1}{2}} \qquad .....(20)$$

iv) Solve for $\hat{\varepsilon}$

$$\hat{\varepsilon} = \overline{H} + \hat{\theta}\,\Gamma(1 + 1/\hat{\alpha}) \qquad .....(21)$$

5.1.4 <u>Lognormal Distribution</u>

The lognormal distribution has two parameters μ and σ. In the transformation from a normal to a lognormal distribution their properties:

μ changes from a location parameter to a scale parameter θ, while

σ changes from a scale parameter to the shape parameter α.

The normal density has only one standardised shape, however the lognormal can assume a variety of shapes depending upon the value of the shape parameter α.

The two moment estimators are given by

$$\alpha = [\ln\{\overline{H^2}\} - 2\ln\{\overline{H}\}]^{\frac{1}{2}} \qquad .....(22)$$

and

$$\theta = 2\ln\{\overline{H}\} - \tfrac{1}{2}\ln\{\overline{H^2}\} \qquad .....(23)$$

The use of these is straightforward, but they can be expected to perform poorly if the density is skewed too highly, i.e. best results will occur when the sample histogram of $\ln(H)$ is nearly symmetrical about the mean value, Bury [1975; p.281].

5.2 Method of Least-Squares

It was recommended in SECTION 3.3 that priority be given to fitting a line to the data occuring within the tail rather than using the entire sample. Of the procedures discussed in this chapter, the method of least-squares is the only one which may be used effectively with a portion of the sample and forms an important part of the approach recommended in this thesis.

Since all the types of probability paper described here give a straight-line plot for their family of distributions, it is feasible to use the linear version of the method of least-squares when fitting a line to the data. The method in its basic form is directly

applicable to the lognormal, Type I and Type II distributions and is described below. However, a modification is required for the Type III distribution since its abscissa scale is dependent upon the parameter ε, which is to be estimated.

Two Parameter Estimation

The least-squares method for determining the line of best fit to a group of data points is well known. In FIGURE 10 a series of data points are to be fitted by a line of slope a and intercept b. The vertical distance from a data point to the line is given by

$$r = |y_i - a\,x_i - b| \qquad .....(24)$$

the sum of the squared distances is

$$q = \sum_{1}^{N} (y_i - a\,x_i - b)^2 \qquad .....(25)$$

the method of least squares selects values of a and b which minimize q.

$$\frac{\partial q}{\partial a} = -2 \sum_{1}^{N} (y_i - a\,x_i - b)\,x_i = 0 \qquad .....(26)$$

$$\frac{\partial q}{\partial b} = -2 \sum_{1}^{N} (y_i - a\,x_i - b) = 0 \qquad .....(27)$$

where i is the point index

N is the number of points.

Whence

$$a = \left[N \sum^{N} x_i y_i - \sum^{N} x_i \sum^{N} y_i\right]\left[N \sum^{N} x_i^2 - \left(\sum^{N} x_i\right)^2\right]^{-1} \quad \ldots(28)$$

and

$$b = \left[\sum^{N} y_i \sum^{N} x_i^2 - \sum^{N} x_i \sum^{N} x_i y_i\right]\left[N \sum^{N} x_i^2 - \left(\sum^{N} x_i\right)^2\right]^{-1} \quad \ldots(29)$$

Thus the slope and the intercept may be calculated directly from a table of wave heights and their frequencies of occurrence.

5.2.1 Gumbel Distribution - Type I

The Type I distribution uses the general method described earlier. Type I paper uses an ordinate scale of

$$y = - \ln \{- \ln [1 - Q(h)]\} \quad \ldots\ldots(30)$$

and an abscissa scale of

$$x = h$$

The data plotting positions are calculated by the method in SECTION 3.2 and the least-squares estimates are:

$$\hat{\theta} = {}^{1}/_{a} \quad \ldots\ldots(31)$$

$$\hat{\varepsilon} = {}^{-b}/_{a} \quad \ldots\ldots(32)$$

where a and b are the slope and intercept.

5.2.2 Fretchet Distribution - Type II

The basic method is applied to the Type II distribution, which has the same ordinate scale as Type I (Eqn. 30), but with an abscissa scale:

$$x = \ln h$$

The treatment is the same as used for Type I, and yields the following estimations

$$\hat{\alpha} = a \qquad \text{.....(33)}$$

$$\hat{\theta} = \exp\{-b/a\} \qquad \text{.....(34)}$$

where a and b are the slope and intercept respectively.

5.2.3 Lognormal Distribution

Lognormal paper has, as its ordinate scale, critical points of probability from the standard Normal distribution (SECTION 2.1).

of $$P(Z) = \int_{-\infty}^{Z} \exp^{-t^2/2} \, dt \qquad \text{.....(35)}$$

then the ordinate scale is given by

$$y = Z$$

The abscissa scale is given by

$$x = \ln H$$

P(Z) is tabulated as NORMAL PROBABILITIES and is available in any statistical text.

Application of the basic method to data plotted according to SECTION 3.2 gives the estimators:

$$\hat{\alpha} = 1/a \qquad \text{.....(36)}$$

$$\hat{\theta} = -b/a \qquad \text{.....(37)}$$

5.2.4 Weibull Distribution - Type III

The three parameter Weibull distribution has the same ordinate scale as the Types I and II distributions, but has an abscissa scale which is itself dependent upon one of the parameters ε to be estimated. See FIGURE 15 and SECTION 2.0. Hence, without prior knowledge of ε the data cannot even be plotted. An iterative least squares procedure must be adopted to overcome this problem.

Using the least squares criterion, three equations must be analyzed. This is achieved by the basic method of minimizing the sum of the squared errors as described previously.

For the Type III_U distribution the three resulting equations now become

$$a = \left[N\sum^{N} x_i y_i - \sum^{N} x_i \sum^{N} y_i\right]\left[N\sum^{N} x_i^2 - \left(\sum^{N} x_i\right)^2\right]^{-1} \qquad \text{...(38)}$$

$$b = \left[\sum^{N} y_i \sum^{N} x_i^2 - \sum^{N} x_i \sum^{N} x_i y_i\right]\left[N\sum^{N} x_i - \left(\sum^{N} x_i^2\right)\right]^{-1} \qquad \text{...(39)}$$

$$r = a\sum^{N} {}^{y_i}/_{(\hat{\varepsilon}-H_i)} - ab\sum^{N} {}^{1}/_{(\hat{\varepsilon}-H_i)} + a^2\sum^{N} {}^{\ell n(\hat{\varepsilon}-H_i)}/_{(\varepsilon-H_i)} \qquad \text{...(40)}$$

where $r = \frac{\partial q}{\partial \varepsilon}$, $x_i = -\ell n(\hat{\varepsilon}-H_i)$, and $y_i = -\ell n\{-\ell n[1-Q(H_i)]\}$

Similarly, three equations may be assembled for the Type III_L distribution. The procedure for the Type III_U now becomes:

i) Select an initial value ε_o. This may be the largest measured wave height $H_{i=1}$

ii) Calculate a and b using Eqns. 38 and 39.

iii) Calculate r from Eqn. 40 and check for solution when $r = 0$

iv) Increase ε by an increment $\Delta\hat{\varepsilon}$, $\hat{\varepsilon}_1 = \varepsilon_o + \Delta\hat{\varepsilon}$, and repeat procedure until the value of ε which gives $r = 0$ is found.

The least-squares estimation of the Type III parameters is achieved by computer program, and convergence of r as ε is increased is shown for typical wave data in FIGURE 11.

5.3 Method of Maximum Liklihood

The method of maximum likelihood estimators (MLEs) attempts to provide estimated parameters which would give the data sample the highest probability of being observed in its particular form. The relative sizes of each of the data values play a fundamental part in the method, though their order is not important, and for this reason the method is unsuited for fitting a distribution specifically to the sample tail. The random sample is considered to consist of a series of independent observations from the same distribution. The probability of the intersection of these events is then the product of their individual probabilities.

The Likelihood function is defined

$$L = \prod_{m=1}^{n} p_\theta (h_m) \qquad \text{.....(41)}$$

where p_θ is the density of the parametric family, e.g. Type I, h_m are the individual wave heights, and n is the total number of wave heights.

The method of maximum likelihood then selects values for each parameter which maximise the likelihood function. Since most of the common density functions have an exponential form, this procedure is simplified by minimising the logarithm of the likelihood function.

MLE's have an important advantage over all other estimators in that they can yield unbiased estimators with minimum variance. This results in a comparatively efficient use of the data and estimates which are more likely to be close to their true values. Against this

quality lies the consideration that for two or more parameters they are troublesome to solve, requiring lengthy iterative manipulation. Furthermore the method uses the entire sample of wave heights and thus is unsuitable to studies directed specifically at the distribution tail. As a result the MLE's are rarely used in offshore engineering since it is usually felt that their drawbacks outweigh the principal advantage.

For completeness, a short description of the Maximum Likelihood procedure follows for each distribution.

5.3.1 Gumbel Distribution - Type I

The likelihood function is given by

$$L(h;\theta,\varepsilon) = \theta^{-n} \exp\left[- \sum_{1}^{n} \frac{h_m-\varepsilon}{\theta} - \sum_{1}^{n} \exp\left\{-\frac{h_m-\varepsilon}{\theta}\right\}\right] \quad \ldots.(42)$$

and by setting

$$\frac{\partial}{\partial\varepsilon} \ln L = 0$$

$$\frac{\partial}{\partial\theta} \ln L = 0$$

two equations in $\hat{\varepsilon}$ and $\hat{\theta}$ can be obtained

$$\frac{1}{n} \sum_{1}^{n} \exp\left\{-\frac{h_m-\varepsilon}{\hat{\theta}}\right\} = 1 \quad \ldots..(43)$$

and

$$\hat{\theta} + \hat{\varepsilon} + \frac{1}{n} \sum_{1}^{A} h_m \exp\left\{-\frac{h_m - \varepsilon}{\hat{\theta}}\right\} - \frac{\hat{\varepsilon}}{n} \sum_{1}^{n} \exp\left\{-\frac{h_m - \hat{\varepsilon}}{\hat{\theta}}\right\} = \overline{H} \qquad \ldots\ldots(44)$$

It can be seen that these involve an iterative solution for $\hat{\varepsilon}$ and $\hat{\hat{\theta}}$.

5.3.2 Fretchet Distribution - Type II

The likelihood function is given by

$$L(h;\alpha,\theta) = \alpha^n \exp\left\{-\sum_{1}^{n}\left(\frac{h_m}{\theta}\right)^{-\alpha}\right\}\prod_{1}^{n}\left(\frac{h_m}{\theta}\right)^{-(\alpha+1)} \qquad \ldots\ldots(45)$$

and the two ML equations in α and β are

$$n/\hat{\theta} - \hat{\theta}^{(\hat{\alpha}-1)} \sum_{1}^{n} h_m^{-\hat{\alpha}} = 0 \qquad \ldots\ldots(46)$$

and

$$n/\hat{\alpha} - \sum_{1}^{n} \ln h_m + \left[n \sum_{1}^{n} h_m^{-\hat{\alpha}} \ln h_m\right]\left[\sum_{1}^{n} h_m^{-\hat{\alpha}}\right]^{-1} = 0 \qquad \ldots\ldots(47)$$

where n is the total number of wave heights.

A more detailed account of the distribution is given by Thom [1954]. Equations 46 and 47 must be solved simultaneously by computer or a graphical method. A closed form for the maximum likelihood estimator does not exist.

5.3.3 Weibull Distribution - Type III_U

The likelihood function is given by

$$L(h;\varepsilon,\alpha,\theta) = \left(\frac{\alpha}{\theta^{\alpha}}\right)^{n} \exp\left\{-\frac{1}{\theta^{\alpha}}\sum \lambda_m^{\alpha}\right\}\prod_{1}^{n} \lambda_m^{\alpha-1} \qquad \ldots(48)$$

where $\lambda_m = \varepsilon - h_m$

the three equations which result from maximizing the likelihood function are

$$\hat{\theta}^{\hat{\alpha}} \frac{\hat{\alpha}-1}{\hat{\alpha}} = \left[\sum_{1}^{n} \lambda_m^{\alpha-1}\right]\left[\sum_{1}^{n} \lambda_m^{-1}\right]^{-1} \qquad \ldots\ldots(49)$$

$$\hat{\theta} = \left[\frac{1}{n}\sum_{1}^{n} \lambda_m^{\hat{\alpha}}\right]^{1/\hat{\alpha}} \qquad \ldots\ldots(50)$$

and

$$\frac{n}{\hat{\alpha}} - n\,\ell n\,\hat{\theta} + \sum_{1}^{n} \ell n\,\lambda_m - \sum_{1}^{n}\left(\frac{\lambda m}{\hat{\theta}}\right)^{\hat{\alpha}} \ell n\left(\frac{\lambda m}{\hat{\theta}}\right) = 0 \qquad \ldots\ldots(51)$$

Equations 49, 50 and 51 must be solved simultaneously for $\hat{\theta}$, $\hat{\alpha}$ and $\hat{\varepsilon}$, and again a closed form maximum likelihood estimator does not exist. The treatment of the Type III_L distribution is similar but not given here.

5.3.4 Lognormal Distribution

As a result of its direct connection with the Normal distribution, the lognormal case is comparatively straighforward. The estimators are:

$$\hat{\varepsilon} = \frac{1}{n} \sum_{1}^{n} \ln h_m \qquad .....(52)$$

and

$$\hat{\theta}^2 = \frac{1}{n-1} \sum_{1}^{n} (\ln h_m - \hat{\varepsilon})^2 \qquad .....(53)$$

In view of the simplicity of the M.L. estimators for the lognormal distribution and because of their desirable properties of unbiasedness and minimum variance, these should be used in place of the method of moments (which have a lower efficiency), and may be considered a good alternative to least-squares.

CHAPTER 6

TESTS OF FIT BETWEEN THE DISTRIBUTION AND THE DATA

Basically there are two methods of testing fit between the distribution and the data:

i) by hypothesis testing

ii) by confidence intervals

These two methods operate in much the same way since both use a pivotal quantity and have a level of significance or confidence. The confidence interval approach has been almost universally accepted for this type of work since it can be plotted in a form which may be readily appreciated by the engineer.

Until very recently, the coastal/ocean engineering literature indicated that confidence intervals could only be applied to the Type I distribution derived by Gumbel [1958] and in the form summarized by St. Denis [1969]. The discussion here will show that this is certainly not the case, and that they may be generated for any of the four distributions, and that Gumbel's form was only an approximation to the exact derivation.

The intervals described here are not presented in the standard parametric intervals form most commonly used by statisticians, and which sets probabilistic limits to the estimated values of the parameters. The confidence intervals used for this work set probabilistic limits on the range of each of the data values given the estimated distribution. This permits an engineer to review the results and form a conclusion on the closeness with which a model fits the data.

A summary of the derivation of confidence intervals is presented in the next section. More detailed discussions may be found in accounts by Kendall [1947] and Borgman [1959].

6.1 Derivation of Confidence Intervals

Starting with a data sample $H_1, H_2 \ldots H_n$ which has a continuous parent distribution $P(h)$, we arrange the data in order of magnitude:

$H_{(1)}, H_{(2)} \ldots H_{(n)}$ so that $H_{(1)}$ is the largest and $H_{(n)}$ is the smallest.

Each data value $H_{(m)}$ is then assumed to behave as an independent random variable which has an identical parent distribution $P(h)$.

The general probability density of the mth statistic $H_{(m)}$ of a sample containing n values is the probability that

$$h - \frac{dh}{2} < H_{(m)} \leq h + \frac{dh}{2}$$

In order to achieve this, $m-1$ values must fall above $h + \frac{dh}{2}$, the mth fall within $h \pm \frac{dh}{2}$, and the rest below $h - \frac{dh}{2}$.

Hence $$f_m(h)\, dh \propto P(h)^{n-m} . \; p(h)\, dh . \; 1-P(h)^{m-1}$$

where $p(h)$ is the parent density function

$f_m(h)$ is the density function of the mth statistic

Since there is still some ambiguity over which values go above and below h, we choose (n-m) values below h, then one value at h, leaving the remaining (m-1) to fall above h.

$$f_m(h)\ dh = m \binom{n}{n-m} P(h)^{n-m} [1 - P(h)]^{m-1} p(h)\ dh \quad .....(54)$$

$$m \binom{n}{n-m} = \frac{m\ \Gamma(n+1)}{\Gamma(n-m+1)\Gamma(m+1)}$$

$$= {}^{1}/_{B(n-m+1,m)}$$

where B( ) is the Beta function.

The cumulative probability of the mth largest variable from n values is

$$F_m(h) = \text{Prob}\ [H_{(m)} \leq h]$$

and

$$F_m(h) = \int_{-\infty}^{\infty} f_m(h)\ dh$$

$$Fm(h) = \frac{1}{B(n-m+1,m)} \int_{-\infty}^{h} P(h)^{n-m} [1-P(h)]^{m-1} p(h)\ dh \quad ....(55)$$

let $w = P(h)$

$dw = p(h)\ dh$

$$F_m(h) = \frac{1}{B(n-m+1,m)} \int_0^{P(h)} w^{n-m} \, [1-w]^{m-1} \, dw$$

$$= \frac{B_P(n-m+1,m)}{B(n-m+1,m)} = I_P \, (n-m+1,m) \qquad \ldots\ldots(56)$$

where $I_P(\,)$ is the incomplete beta function which can be expressed in terms of the binomial expansion

$$I_P(m,n-m+1) = \sum_{j=m}^{n} \binom{n}{j} P^{n-j} \, [1-P]^j$$

and $$I_P(a,b) = 1 - I_{1-P}(b,a)$$

$$F_m(h) = 1 - \sum_{j=m}^{n} \binom{n}{j} P(h)^{n-j} \, [1-P(h)]^j \qquad \ldots\ldots(57)$$

Equation 57, then, expresses the probability function of the mth data point in terms of the parent distribution $P(h)$ which governs the occurrence of all wave heights. At this point it is clear that the form or type of the parent distribution has not been specified, and that the method is equally applicable to data from any of the four distributions considered.

6.2 Asymptotic Distribution of the mth Statistic

If the number of wave heights n is increased without limit while both m and $P(h)$ are held stationary, the function $F_m(h)$ tends to zero. This is very inconvenient and it would be desirable to have a stable non-zero limiting function which could be used for large samples.

This is achieved by replacing $P(h)$ by the parameter

$$w_n(h) = n\,[1-P(h)] \qquad \text{.....(58)}$$

and tabulating $w_n(h)$ instead of $P(h)$. In this case the incomplete beta function is replaced by the incomplete gamma function. Tables of w_n were prepared by Borgman [1959] and are summarized in FIGURES 24 to 28 which are plots of $\ell n\, w_n = \ell n\,\{n[1-P(h)]\}$ against the sample size n for several values of $F_m(h)$.

6.3 Approximate Distribution of the mth Statistic

Gumbel [1958] described an approximate form for the standard error of the mth statistic. It was assumed that m/n was approximately ½, so that it was strictly valid only for statistics taken from the centre of the ordered sample. He showed that Eqn. 54 could be expressed as a power series of which all but the first few terms could be neglected. This simplied to the density of a normal distribution.

6.4 Limitations

It should be noted that the assumptions made in its derivation now render the Gumbel version quite inaccurate in the vicinity $m \simeq 1$, (FIGURE 12). Since this region is of prime interest here, especially when using the least-squares method, the more direct method of solving Eqn. 57 is to be preferred.

In addition to a variation between the true interval and the standard deviation indicated by the normal distribution approximation, there is a very noticeable development of skewness (see FIGURE 12) in the distribution of the mth value as m approaches unity. This is quite important when using the confidence interval lines, since they develop a strong bias towards the right of the fitted line, indicating that if other sets of samples were used, the values would tend to fall more often to the right of the fitted line as m decreased and have greater heights than indicated by it.

A restriction on the use of the mth value distribution is that it is not defined in the region $m < 1$, and hence cannot be directly extended to predicted values. Gumbel effectively suggested that predicted maximum values would each retain the same interval size as the statistic at $m = 1$, but this practice does not seem to be generally accepted.

6.5 Method of Determining Critical Points

For a given sample of n wave heights and a statistic ordered position m the distribution of the mth value is described in terms of the original parent distribution. Since our sample consists of extreme values, the parent distribution may be chosen from one of the four distributions used here, e.g. Type III. If the confidence intervals are defined in terms of a probability that a given height will not be exceeded by the mth statistic then Eqn. 57 can be solved for the critical value h. However a more useful form can be reached by solving for the value of the parent distribution $P(h)$ which satisfies $F_m(h) = \phi$, the confidence probability. The value of $P(h)$ given by this may then be applied to any parent distribution to get a critical value of h. This approach results in tabulated values of $P(h)$ for given n, m and γ. Comprehensive tables for $m = 1(1)5$ have been compiled by Borgman [1959].

6.6 Procedure for Plotting Confidence Intervals

A plot of the data is prepared on the paper of the selected distribution by following the sample preparation and plotting procedure given in CHAPTER 3. A best-fit line is then located by one of the techniques in CHAPTER 5. The method for plotting the confidence intervals is as follows:

i) A confidence probability level γ e.g. 0.25 is selected. Each value of m is then treated separately and is used to obtain a pair of

values of $P(h)$ from FIGURES 24 to 28. That is, for a given sample size n, m and setting $F_m(h) = \phi$ for both $\phi = (1-\gamma)/2$ and $(1+\gamma)/2$ respectively, as shown in FIGURE 17, the appropriate FIGURE (24 to 28) is used to determine two values of $P(h)$.

ii) Each value of $P(h)$ is used as an ordinate position (see FIGURE 13) and, when projected onto the best-fit line will given two limiting values of height for the mth statistic.

iii) These may then be plotted on either side of the mth plotting position and a faired line drawn through equivalent limits for the remaining points.

6.7 Examples of Confidence Intervals

The sample described by St. Denis [1969] is shown plotted in FIGURES 13 to 16 with 25% confidence intervals. It may be seen that this level of confidence is sufficient to contain all points in the Type I, Type II and lognormal plots. However, the Type III_U confidence intervals are too narrow to contain the second and third highest points. This indicates that the Type III_U distribution is the least suitable model for this data.

CHAPTER 7

METHODS OF PREDICTION

The processes of selecting the most suitable distribution together with estimates of the parameters have been described in detail. This chapter will discuss methods of using the best-fit line to predict the "extreme wave". In order to describe this wave one requires a representative height, and for this either the significant or the maximum wave height may be used. The calculation of these two values is outlined in SECTIONS 7.1 and 7.3.

The methods described in previous chapters have been directed entirely towards predicting wave heights. Although extreme value analysis have not been applied to wave periods in this thesis, the period of the extreme wave may still be calculated from the limiting steepness as described in SECTION 7.2. The predicted values of wave height and period apply over the same recording interval that was used for data collection. Once an extreme wave height has been obtained it may be desirable to set a pair of probabilistic limits on its value to reflect the size of the sample and the quality of the estimating process. A discussion of such limits is given in SECTION 7.4.

In SECTION 7.5, the encounter probability and return period are discussed in detail. The encounter probability quantifies the risk of a wave with a given return period occurring within the lifetime. Additionally, another value might be used to estimate

the number of smaller waves, occuring within the same period, and which might hinder operation or promote fatigue. By this process the designer would be able to consider a "limit-state" and a "service condition".

7.1 Expected Significant Height

Once the extreme value plot has been drawn the next stage is usually to estimate the so-called 50 or 100 year design wave height. That is the wave height, defined in the same way as the recorded heights (e.g. the significant height over a recording interval), which would only be exceeded on average once during a period of 50 or 100 years respectively. This time interval is called the RETURN PERIOD or RECURRENCE INTERVAL, and is usually selected on the basis of a given structure lifetime.

A non-dimensional EXPECTED WAITING TIME, R is defined as the average number of trials between exceedances of a given height h. Each trial amounts to one recording interval and hence the waiting time is given by

$$\text{EXPECTED WAITING TIME} = \frac{\text{RETURN PERIOD}}{\text{RECORDING INTERVAL}} = R \qquad \ldots(59)$$

Let W be the waiting time, i.e. the random number of observations preceding and including the first exceedance of a given height h. W then has a GEOMETRIC DISTRIBUTION and if

$$P = P_r\,[H \leq h]$$

then the expected value of W is

$$R = \sum_{w=1}^{\infty} w\, P^{w-1}(1 - P) \qquad \ldots.(60)$$

which yields

$$\text{EXPECTED WAITING TIME} = {}^{1}/_{(1 - P)} \qquad \ldots.(61)$$

Thus for a return period T_r of 50 years, and using data which was recorded at 3 hourly intervals, the expected waiting time R would be 146,000. For a given data record and a return period one may thus calculate R and hence $Pr[H \leq h]$ from Eqns. 59 and 61. This will correspond to a value of height h on the probability plot. Since the data consisted of a series of significant heights, this predicted height will represent the significant height of a corresponding recording interval and with the required return period.

7.2 Extreme Wave Period

There are three common approaches to estimating the extreme wave period. The first is to repeat the entire procedure using wave periods instead of wave heights. The marginal frequencies are obtained by summing the number of wave occurrences in each period class. By using the same return period as the heights, one may obtain a predicted value of the 50 year zero-crossing period T_z for a future wave record with the same recording interval. Draper [1963] has suggested that this value of T_z may be used with the predicted height. This suggestion is based on the fact that there is a noticeable correlation between the two variables in the scatter diagram.

The second method of obtaining a representation wave period involves the use of a one-parameter wave spectrum such as the Pierson-Moskowitz spectrum. The spectrum is calculated for the predicted value of significant height and the value of frequency locating the spectrum peak is used to obtain the T_p. Again, the value corresponds to the same recording interval as does the data.

The third method, which is the simplest to use, involves using the predicted wave height to set a lower limit on the wave period. By assuming a Pierson-Moskowitz spectrum, Battjes [1970] has shown that, for deep water and intermediate depths, the wave steepness defined as $2\pi H_s/gT^2$ is limited by

$$\frac{2\pi H_s}{gT_L^2} \leq 1/16 \qquad \ldots.(62)$$

where g is the gravitational constant. Thus a lower limit of period T_L for a given significant height may be set as:

$$T_L = \{32\pi H_s/g\}^{\frac{1}{2}} \qquad \ldots.(63)$$

The method then involves trying different combinations of the period with the predicted height to find the worst effect on the structure.

7.3 Maximum Wave Height

Once the extreme values of significant wave height and mean zero-crossing period have been established, the maximum wave height may be calculated. Longuet-Higgins and Cartwright [1956] showed that for a wave spectrum of arbitrary shape the expected maximum individual wave height could be expressed as

$$\frac{\text{Expected Maximum Height}}{\text{Significant Height}} \simeq [\tfrac{1}{2}\, \ell n\, (t/T_Z)]^{\frac{1}{2}} \qquad \ldots.(64)$$

where t is the recording interval

and T_Z is the mean zero crossing period.

This relationship is valid provided t/T_Z is large, i.e. the recording interval contains a large number of waves. The sea-state is assumed to be stationary throughout this period. In the preceding section 7.2, it was shown that a lower limit would be placed on the period attached to the predicted value of the significant height.

The procedure for calculating the expected maximum wave height is:

i) Using the predicted value of the significant height H_S and the limiting steepness, calculate the minimum period T_L from Eqn. 62:

$$T_L = \{32\pi\, H_S/g\}^{\frac{1}{2}} \qquad \ldots.(63)$$

where g is the gravitational constant.

ii) Calculate the expected maximum wave height from Eqn. 64 using T_L from Eqn. 63. to T_Z.

It is quite sufficient to use the minimum period here since Eqn. 64 is insensitive to variations in t/T_Z. For example, a 10% error in the central period of 7 seconds over a recording interval of 3 hours would result in a height error of less than 0.5%.

7.4 Confidence Intervals for Prediction

Gumbel [1958] suggested that the confidence intervals described in the last chapter could be extended, beyond the region containing data, for prediction. The method he suggested was to draw a pair of lines parallel to the fitted line and passing through the interval offset points of the highest data point. Although the concept of using intervals to indicate error in prediction was very attractive, this method has not generally been adopted according to Chow [1964]. It was, however, restated by St. Denis [1969] in a paper devoted to wave prediction.

As has already been discussed, there is always a degree of variability involved in parameter estimation. An estimate is a function of a random sample and hence is itself a random variable. The estimator's variability is not lessened by the fact that the various methods of estimation often yield slightly different results. Hence, the Weibull distribution has three possible sources of error once it has been estimated. In view of the difficulties described, it is not surprising that the problem was left untouched until quite recently. Thoman, Bain and Antle [1969] prepared confidence interval tables for the parameters of the two-parameter Weibull distribution. This special case occurs when a Type III_L distribution is used with epsilon set equal to zero. The tables were made by using Monte Carlo Simulation to generate a series of random samples from the Type III_L distribution, and thence deriving an empirical distribution for each parameter. Confidence intervals were then taken from these empirical

distributions by a similar process to that used in the last section. This approach was used by Petrauskas and Aagaard [1971] to produce "uncertainty intervals" for prediction. The two limits calculated for each of the parameters involved resulted in a pair of straight lines, each having a slope and intercept which were different from the least-fit line. An example of the uncertainty intervals is shown in FIGURE 23. The intervals generated by this method were found to diverge from the least-fit line as the variate increased instead of remaining parallel to it, as originally suggested by Gumbel.

7.5 Encounter Probability and Waiting Time

It is accepted practice to refer to a design wave of given height by its RETURN PERIOD at a specific location. Thus a 25-year wave means that waves as large as the design wave or larger, occur on average once in each 25-year period. It is evident that in fact several such waves could possibly occur within the same 25-year period. Borgman [1963] has given a description of the distribution of the waiting time between events and of the probability of encounter.

The concept of return period enables one to represent the continuous time dimension as a series of discrete integer multiples of the original recording interval, or of a conveniently related quantity such as one year. Thus time can be used as a discrete random variable which has two possible states, which reflect whether or not the design wave has been exceeded within the associated time period.

The probability of an exceedance is given by

$$p = 1 - P(h)$$

and $q = 1-p$ represents the probability that a recording interval t will contain a wave higher than h.

If T_W is the waiting time until the first exceedance of h occurs, then it has a geometric distribution and

$$P_r \{T_W \leq \tau.t\} = 1 - q^{\tau} \qquad \ldots.(65)$$

where τ is a dimensionless integer multiple from Eqns. 59 and 61, the expected waiting time is

$$T_{r/t} = {}^{1}/_{[1-P(h)]} \qquad \ldots.(66)$$

thus from Eqns. 65 and 66

$$P_r \{\frac{T_W}{t} \leq \tau\} = 1 - (1 - {}^{t}/_{T_r})^{\tau} \qquad \ldots.(67)$$

This is the probability distribution of waiting time, and it is independent of the timing of the previous exceedance. If $(\tau.t)$ represents the lifetime of the structure then Eqn. 67 may be used to calculate the probability of it experiencing a wave with a given return period. FIGURE 22 has been constructed from Eqn. 67 for the special case where $t = 1$ year. When $T^2/t^2\tau >> 1$ (i.e. generally when $T/t >> 1$) a suitable approximation to Eqn. 67 is given by

$$P_r \{\frac{T_W}{t} \leq \tau\} \simeq 1 - \exp\{-\frac{\tau.t}{T}\} \qquad \ldots.(67a)$$

CHAPTER 8

CONCLUSIONS - A RECOMMENDED PROCEDURE

A recommended procedure for the prediction of extreme waves is now given below:

a) The data is taken from an intermittent record over a period of at least one year. Usually this will consist of a series of short continuous records (10 - 20 minute duration) which have been started at intervals of 3 - 10 hours. For reasons given in SECTION 3.3 it is possible to use data which has some of the results from 'low activity' months missing provided it can be shown that their approximate ranking positions fall outside the tail as defined in SECTION 3.3.

b) Each continuous record is reduced by the method described in SECTION 3.1 to a pair of single values, i.e. the significant height H_s and the zero crossing period T_z.

c) A scatter diagram with both H_s and T_z divided into a number of equal classes is then prepared. The number of records falling into each joint interval should be marked as shown in FIGURE 5.

d) The marginal height frequencies are fixed by summing over the period T_z for each class of H_s.

e) The plotting positions of each class lower limit are calculated according to the method of SECTION 3.1.

f) Each class lower limit is plotted on Type I paper and the curvature test is applied as in SECTION 4.2. This will result in a choice of one of three groups of distribution which are described in SECTION 4.1.

g) The tail of the sample is isolated for further use by the method given in SECTION 3.3. If this yields less than five points, one may return to step (c) and further subdivide the height classes in the scatter diagram. If this is not feasible one may have to accept a more general fit and include some lower limit classes.

h) The class lower limits of the tail are then plotted onto each paper of the distribution group selected in step (f). A straight line is fitted by one of the methods described in CHAPTER 5.

i) For each distribution of the group, the confidence intervals corresponding to the plotted points are drawn. Initially a confidence level of 60% may be used and narrower bands drawn until points start to fall outside the limits. On this basis one may select one distribution which gives the best fit to the data.

j) Predicted values of wave height and period may now be made by the methods of CHAPTER 7, and encounter probabilities assigned where appropriate.

CHAPTER 9

A WORKED EXAMPLE

The data used for this worked example was collected by the Department of Public Works of Canada at TINIER POINT, NEW BRUNSWICK. It covers a period of one year and originally appeared in a paper by Khanna and Andru [1974]. The example is analysed using the same steps as summarized in CHAPTER 8.

Steps a) to c) were originally carried out by the collecting agency and the starting point was the scatter diagram shown in FIGURE 5. The marginal frequencies of significant height H_S and zero crossing period T_Z are also shown [step d)].

Steps e) and f) were carried out to give the results shown in FIGURE 18. A curved line, which was fitted to the points by eye for convenience, has a strong positive curvature. On the basis of the method discussed in SECTION 4.1 this permitted the analysis to be confined to two possible models:

- Type III_U distribution
- Type III_L distribution

Step g) was carried out according to SECTION 3.3 as follows:

total number of wave heights on record

$N = 2245$

longest rank within tail $w = \frac{(n+1)}{10} = 224$

By adding the marginal frequencies of each height class the position of the class containing a value with a rank of 224 was located. This

resulted in the nine points which were plotted in FIGURE 18. The extent of the tail is indicated and the remaining points have been included to show their behaviour.

The result of step h), using the method of least squares, is shown in FIGURES 19 and 20. The confidence intervals have also been fitted according to step i). The calculation of confidence band positions is given in TABLE V. Since FIGURES 24 to 28 only cover the first five ranked positions [m=1 to 5] they cannot be directly applied to higher ranked values. This is overcome by introducing an approximate method which may be justified by the fact that the confidence intervals only serve as a test of comparison, and hence do not require the rigorous approach of parameter estimation. The method is given as follows:

i) The rank m of the smallest value occurring in each class is used and shown in column 2.

ii) The mean frequency is calculated as

$$\overline{Q} = \frac{m}{n+1}$$

iii) When the rank becomes greater than five, m is set equal to 5 and a new sample size n' is chosen to give an approximate frequency (column 4) which is close to the value in column 3.

$$n' = \frac{5}{\overline{Q}_m} - 1$$

The resulting values of n' are shown in column 5.

iv) An initial confidence level of $\gamma = 0.60$ was chosen. The procedure of SECTION 6.6 was then used to plot the confidence intervals with the following modifications

- n' replaces the true sample size
- FIGURE 28 is used for $m \geq 5$.

The final plots together with confidence intervals are shown in FIGURES 19 and 20.

Since the 60% confidence bands did not enclose all the points in FIGURE 19, intervals for 80% confidence were plotted. It was found that 60% confidence was sufficient for FIGURE 20, and that the highest pair of points could be enclosed by a narrower band of 40% confidence. Thus the Type III_U distribution with $\varepsilon = 15$ feet was the most suitable model for this data.

The prediction of the 100 year design wave is made according to the methods given in CHAPTER 7 for a RECORDING INTERVAL of 3 hours, and a RETURN PERIOD of 100 years, which results in an EXPECTED WAITING TIME of 292,000. The probability of non-exceedance is calculated from Eqn. 64 as $P(h) = 0.99999657$, which corresponds to $y = 12.584$ on the ordinate scale of FIGURE 20. This yields a 100 year SIGNIFICANT HEIGHT OF 14.47 feet.

The data used for this worked example was analysed by Khanna and Andru [1974]. Their estimate for the 100 year significant height varied between 20 and 30 feet. The lowest value was taken from a

Type III_L plot and the largest from a lognormal plot. The value of significant height for the 100 year wave suggested by this worked example was less than 15 feet. The large difference is attributable to the limiting effect of the Type III_U distribution and to the fitted line now lying to the left of the $m=1$ point.

The minimum period is determined from Eqn. 64 as $T_L = 6.7$ seconds, and hence the EXPECTED MAXIMUM HEIGHT is given by Eqn. 63 as 27.9 feet.

CHAPTER 10

FUTURE WORK

The procedure which has been described is primarily concerned with predicting a wave height of given return period. In cases where the dynamic response of a structure to waves is of importance, one must consider the distribution of wave periods. In order to calculate the combined effect of height and period variation, it becomes necessary to introduce a long-term bivariate distribution. This describes the joint probability of a given height and period occurring in combination. To a limited extent this problem has been examined by Battjes [SECTION 1.0] who used a discrete approach. The bivariate distribution developed could be continuous and its marginal distributions of height and period may be quite different, e.g. a Type I for periods with a Type III_U for significant height. Thus, instead of fitting a straight line one would use a "surface of best fit". One advantage of such an approach would be that a designer could take the fundamental frequencies of the structure into account when predicting design values. An approach would be to predict a wave of given return period, e.g. fifty year wave, given that the period of concern lay between limits, e.g. 6 to 8 seconds.

From SECTION 7.4 it can be seen that there is still no direct approach for obtaining confidence bands for predicted values. It would be most useful to the engineer if a method based on tables could be prepared for office use.

BIBLIOGRAPHY

Abramowitz, M. and Stegun, I. *Handbook of Mathematical Functions,* Dover, New York, 1970.

Battjes, J.A. "Long-term wave height distribution at seven stations around the British Isles," *National Institute of Oceanography,* England, Internal Report No. A.44, July, 1970.

Borgman, L.E. "The frequency distribution for the mth largest of n values," *M.S. Thesis, University of Houston,* Houston, Texas 1959.

Borgman, L.E. "The frequency distribution of near extremes," *Journal of Geophysical Research,* Vol. 66, No. 10, pages 3295-3307, October, 1961.

Borgman, L.E. "Risk Criteria," *Journal of the Waterways and Harbours Division,* American Society of Civil Engineers, Vol. 89, No. WW3. August, 1963.

Borgman, L.E. "Extremal Statistics in Ocean Engineering," *Proceedings of Civil Engineering in the Oceans,* University of Delaware, 1975.

Bretschneider, C.L. "Generation of wave by winds, state-of-the-art," *National Engineering Science Company,* Report SN-134-6, January, 1965.

Bury, K.V. *Statistical Models in Applied Science,* John Wiley and Sons, 1975.

Cartwright, D.E. and Longuet-Higgins, M.S. "The Statistical distribution of the maxima of a random function," *Philosophical Transactions Royal Society of London,* A247, pages 22-48, 1958.

Chow, V.T. *Handbook of Applied Hydrology: a compendium of water resources technology,* McGraw-Hill, New York, 1964.

Draper, L. "Derivation of a design wave from instrumental records of sea states," *Proceedings, Institution of Civil Engineers,* London, Vol. 26, pages 291-304.

Fisher, R.A. and Tippet, L.H.C. "Limiting forms of the frequency distrubution of the largest or smallest member of a sample," *Proceedings, Cambridge Philosophical Society,* Vol. 24, 1926.

Gringorten, I.I. "Envelopes for ordered observations applied to meterological extremes," *Journal of Geophysical Research,* Vol. 63, No. 3, pages 815-826, February 1963

Gumbel, E.J. "Statistical theory of droughts," *Journal of Hydraulics Division,* American Society of Civil Engineers, Vol. 80, May 1954.

Gumbel, E.J. *Statistics of Extremes,* Columbia University Press, 1958.

Hogben, N. "A companion of log-normal and Weibull functions for fitting long-term wave height distributions in the North Atlantic," National Physics Laboratory, England. Ship Division, T.M.190, October 1967.

Jasper, N.H. "Statistical distribution patterns of ocean waves and of wave-induced ship stresses and motions with engineering applications," *S.N.A.M.E., New York Meeting,* Preprint No. 6, 1956.

Kendall, M.G. *The Advanced Theory of Statistics,* Vol. 1, Charles Griffin and Co., London, 1947.

Khanna, J. and Andru, P. "Lifetime wave height curve for Saint John Deep, Canada," *Proceedings, International Symposium, Ocean Wave Measurement and Analysis,* Vol. 1, pages 301-319, ASCE, New Orleans, pages 301-319, September 1974.

Kimball, B.F. "On the choice of plotting positions on probability paper," *Journal of the American Statistical Association,* Vol. 55, pages 546-560 1960.

Ouellet, Y. "On the need of wave data for the design of rubble mound breakwaters," *Proceedings, International Symposium on Ocean Wave Measurement and Analysis,* Vol. 1, pages 500-522, ASCE, New Orleans, September 1974.

Petrauskas, C. and Aagaard, P. "Extrapolation of historical storm data for estimating design-wave heights," *Journal of the Society of Petroleum Engineers,* Vol. 11, pages 23-37, March 1971.

Simiu, E. and Filliben, J.J. "Probability distributions of extreme wind speeds," *Journal of the Structural Division,* American Society of Civil Engineers, Vol. 102, No. ST9, pages 1861-1877, September 1976.

St. Denis, M. "Determination of Extreme Waves", *Topics in Ocean Engineering,* Vol. 1, C.L. Bretschneider (ed.) pages 37-41, Gulf Publishing Co., Texas, 1969.

St. Denis, M. "Some cautions on the employment of the spectral technique to describe the waves of the sea and the response thereto of oceanic systems," *Offshore Technology Conference,* Houston, Paper OTC 1819, May 1973.

Thom, H.C.S. "Frequency of maximum wind-speeds" *Journal of the Structural Division,* American Society of Civil Engineers, Vol. 80, November 1954.

Thom, H.C.S. "Asymptotic extreme-value distributions of wave heights in the open ocean," *Journal of Marine Research,* Vol. 29, pages 19-27, 1971.

Thoman, D.R., Bain, L.J. and Antle, C.E. "Inference parameters of the Weibull distribution," *Technometrics,* Vol. 11(3), pages 445-460, 1969.

Tucker, M.J. "Analysis of records of sea waves," *Proceedings, Institution of Civil Engineers,* London, Vol. 26, pages 305-316, 1963.

TABLES

DISTRIBUTION	RANGE	PROBABILITY FUNCTION P(h)	EXPECTED VALUE	VARIANCE
LOGNORMAL	$0<H<\infty$ $-\infty<\theta<\infty$ $0<\alpha<\infty$	$\frac{1}{\sqrt{2\pi}}\int_0^h \frac{1}{\alpha h} \exp\{-\frac{1}{2}\left(\frac{\ell nh-\theta}{\alpha}\right)^2\}dh$	$\exp\{\theta + \frac{\alpha^2}{2}\}$	$\exp\{2\theta+\alpha^2\}[\exp(\alpha^2)-1]$
TYPE I	$-\infty<H<\infty$ $-\infty<\varepsilon<\infty$ $0<\theta<\infty$	$\exp\{-\exp[-\left(\frac{H-\varepsilon}{\theta}\right)]\}$	$\varepsilon + 0.57722\ \theta$	$1.64493\ \theta^2$
TYPE II	$0<\alpha<\infty$ $0<\theta<\infty$ $0<H<\infty$	$\exp\{-\left(\frac{H}{\theta}\right)^{-\alpha}\}$	$\theta\ \Gamma(1-\frac{1}{\alpha})$	$\theta^2\{\Gamma(1-\frac{2}{\alpha})-\Gamma^2(1-\frac{1}{\alpha})\}$
TYPE III_U UPPERBOUND	$-\infty<H<\varepsilon$ $0<\alpha<\infty$ $0<\theta<\infty$	$\exp\{-\left(\frac{\varepsilon-H}{\theta}\right)^{\alpha}\}$	$\varepsilon-\theta\ \Gamma(1+\frac{1}{\alpha})$	$\theta^2\{\Gamma(1+\frac{2}{\alpha})-\Gamma^2(1+\frac{1}{\alpha})\}$
TYPE III_L LOWER BOUND	$\varepsilon<H<\infty$ $0<\alpha<\infty$ $0<\theta<\infty$	$1-\exp\{-\left(\frac{H-\varepsilon}{\theta}\right)^{\alpha}\}$	$\varepsilon+\theta\ \Gamma(1+\frac{1}{\alpha})$	$\theta^2\{\Gamma(1+\frac{2}{\alpha})-\Gamma^2(1+\frac{1}{\alpha})\}$

TABLE 1 PROBABILITY DISTRIBUTIONS AND THEIR PROPERTIES

DISTRIBUTION	SLOPE	TAIL CURVATURE
Lognormal	positive	negative
Type I	$1/\theta$	straight line
Type II	$+\alpha/H$	negative curve $-\alpha/H^2 \leq 0$
Type III_L	$\frac{\alpha}{\theta}\left(\frac{H-\varepsilon}{\theta}\right)^{\alpha-1}$	$\frac{\alpha(\alpha-1)}{\theta^{\alpha}}(H-\varepsilon)^{\alpha-1}$ $\alpha<1$ negative $\alpha=1$ straight $\alpha>1$ positive
Type III_U	$+\alpha/(\varepsilon-H)$	positive curve $\alpha/(\varepsilon-x)^2 \geq 0$

TABLE II CURVATURE PROPERTIES OF THE DISTRIBUTIONS

1/ALPHA	SHAPE FACTOR	ALPHA
0.0125	1.2161	80.00
0.0250	1.2970	40.00
0.0375	1.3827	26.67
0.0500	1.4739	20.00
0.0625	1.5713	16.00
0.0750	1.6757	13.33
0.0875	1.7883	11.43
0.1000	1.9103	10.00
0.1125	2.0433	8.89
0.1250	2.1893	8.00
0.1375	2.3505	7.27
0.1500	2.5302	6.67
0.1625	2.7324	6.15
0.1750	2.9621	5.71
0.1875	3.2265	5.33
0.2000	3.5351	5.00
0.2125	3.9015	4.71
0.2250	4.3456	4.44
0.2375	4.8974	4.21
0.2500	5.6051	4.00
0.2625	6.5509	3.81
0.2750	7.8872	3.64
0.2875	9.9327	3.48
0.3000	13.4835	3.33
0.3125	21.2472	3.20
0.3250	52.1732	3.08
0.3375	-102.1849	2.96
0.3500	-24.9316	2.86
0.3625	-13.8496	2.76
0.3750	-9.3816	2.67
0.3875	-6.9457	2.58
0.4000	-5.3957	2.50
0.4125	-4.3085	2.42
0.4250	-3.4908	2.35
0.4375	-2.8404	2.29
0.4500	-2.2965	2.22
0.4625	-1.8180	2.16
0.4750	-1.3691	2.11
0.4875	-0.8994	2.05
0.5000	-0.0018	2.00

TABLE III Shape Factors for the FRETCHET Distribution

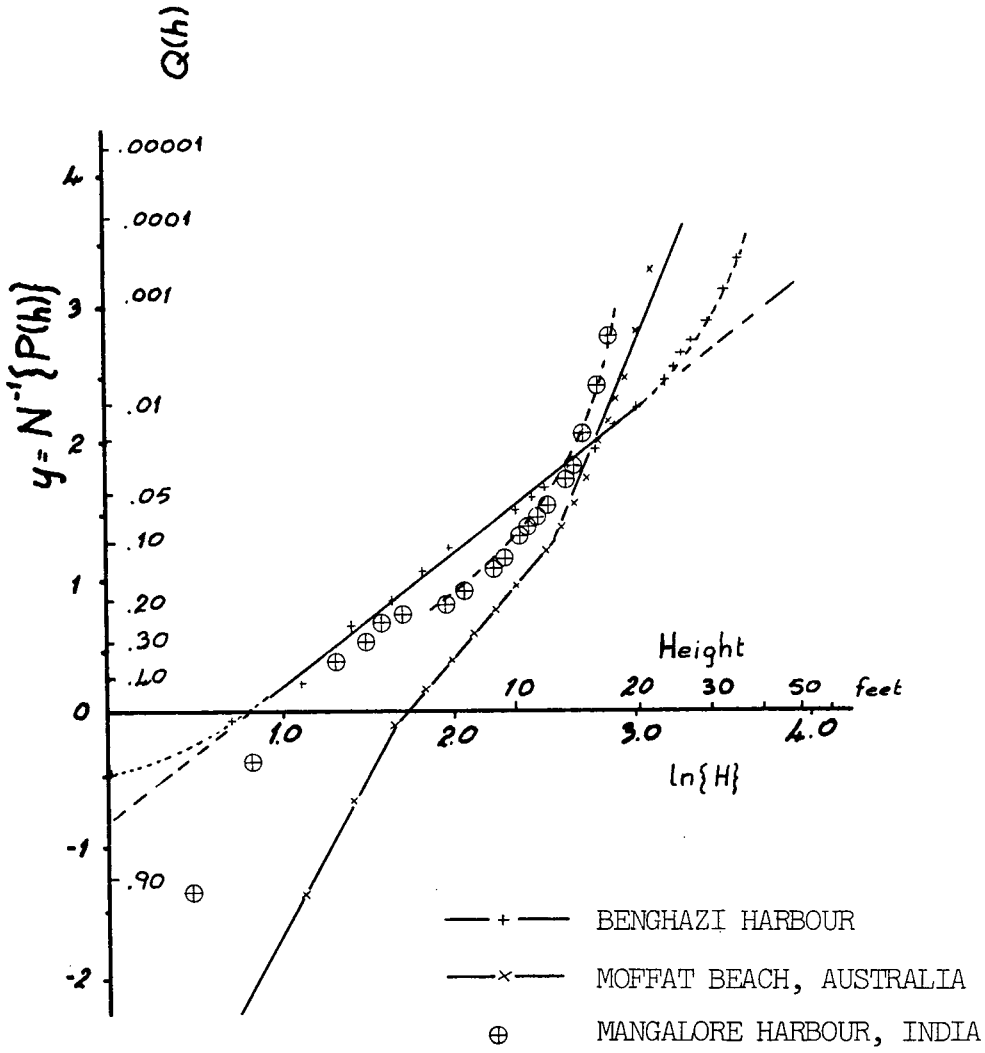
1/ALPHA	SHAPE FACTOR	ALPHA
0.05	-0.8680	20.000
0.10	-0.6376	10.000
0.15	-0.4357	6.667
0.20	-0.2541	5.000
0.25	-0.0872	4.000
0.30	0.0687	3.333
0.35	0.2167	2.857
0.40	0.3586	2.500
0.45	0.4963	2.222
0.50	0.6311	2.000
0.55	0.7640	1.818
0.60	0.8960	1.667
0.65	1.0279	1.538
0.70	1.1604	1.429
0.75	1.2941	1.333
0.80	1.4295	1.250
0.85	1.5674	1.176
0.90	1.7080	1.111
0.95	1.8521	1.053
1.00	2.0000	1.000
1.05	2.1523	0.952
1.10	2.3093	0.909
1.15	2.4718	0.870
1.20	2.6400	0.833
1.25	2.8146	0.800
1.30	2.9961	0.769
1.35	3.1851	0.741
1.40	3.3820	0.714
1.45	3.5875	0.690
1.50	3.8023	0.667
1.55	4.0269	0.645
1.60	4.2621	0.625
1.65	4.5086	0.606
1.70	4.7671	0.588
1.75	5.0385	0.571
1.80	5.3235	0.556
1.85	5.6230	0.541
1.90	5.9381	0.526
1.95	6.2697	0.513
2.00	6.6188	0.500

TABLE IV Shape Factors for the WEIBULL Distribution

	m	$\overline{Q}_m$	$\overline{Q}^1_m$	n^1		
1 DATA POINT	2 APPROX RANK	3 TRUE MEAN FREQ.	4 APPROX MEAN FREQ.	5 EQUIV. SAMPLE SIZE	6 60% CONFIDENCE	
					y LOWER	Y UPPER
HIGHEST 1	1	1/2245	-	2245	7.24	9.22
2	2	2/2245	-	2245	6.62	7.91
3	8	8/2245	5/14000	1400	5.34	6.11
4	15	15/2245	5/750	750	4.71	5.49
5	24	24/2245	5/450	450	4.20	4.98
6	38	38/2245	5/245	295	3.77	4.55
7	64	64/2245	5/175	175	3.25	4.02
8	93	93/2245	5/120	120	2.87	3.64
9	133	133/2245	5/85	85	2.51	3.29

TABLE V: ESTIMATION OF CONFIDENCE INTERVALS FOR TYPE III_U PLOT

FIGURES



NOTE: $N^{-1}\{P(h)\}$ is the value of variate which corresponds to an area equal to $P(h)$ under the Standard Normal Distribution density curve.

FIGURE 1 Typical Examples of Data on Lognormal Paper

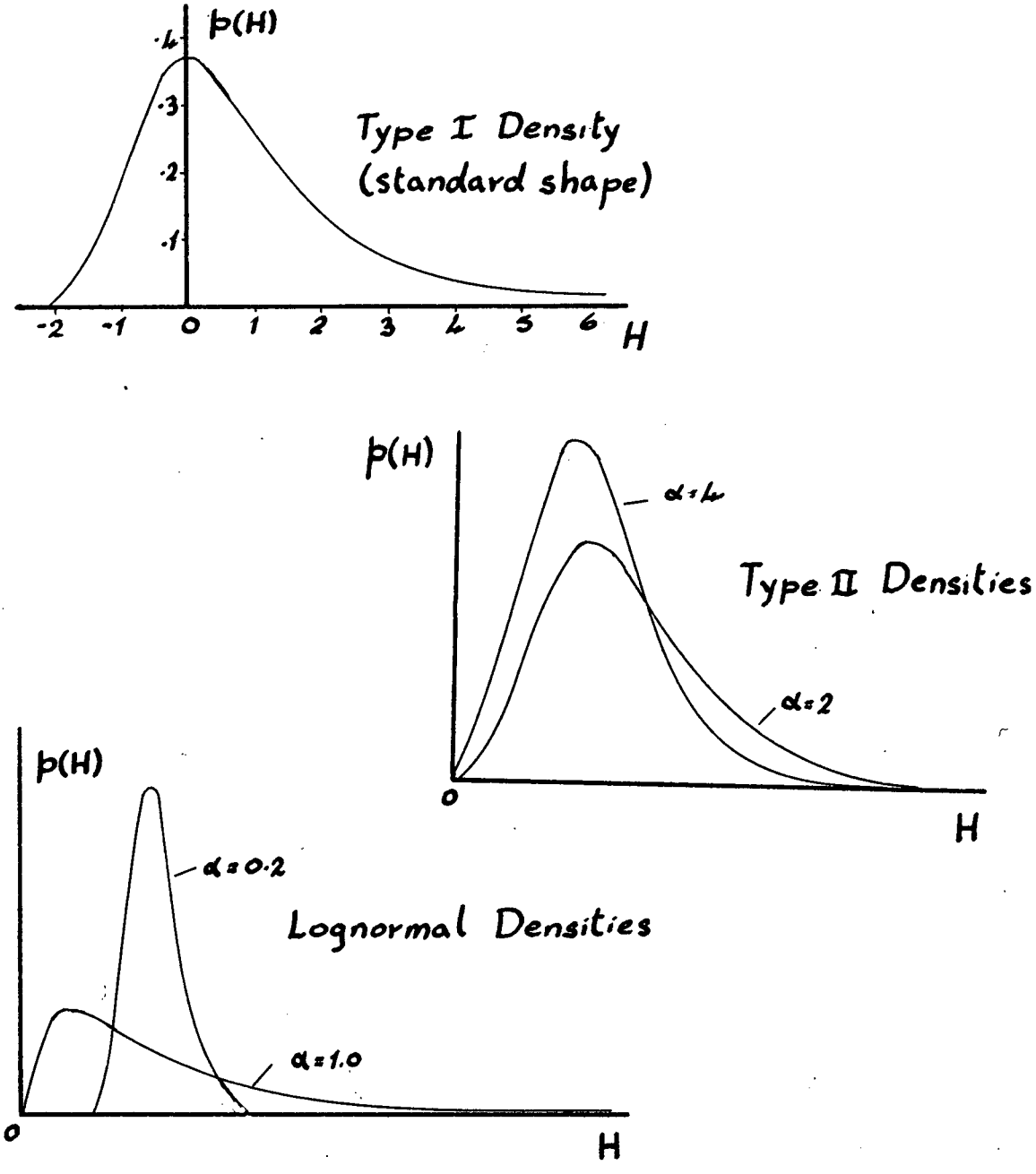
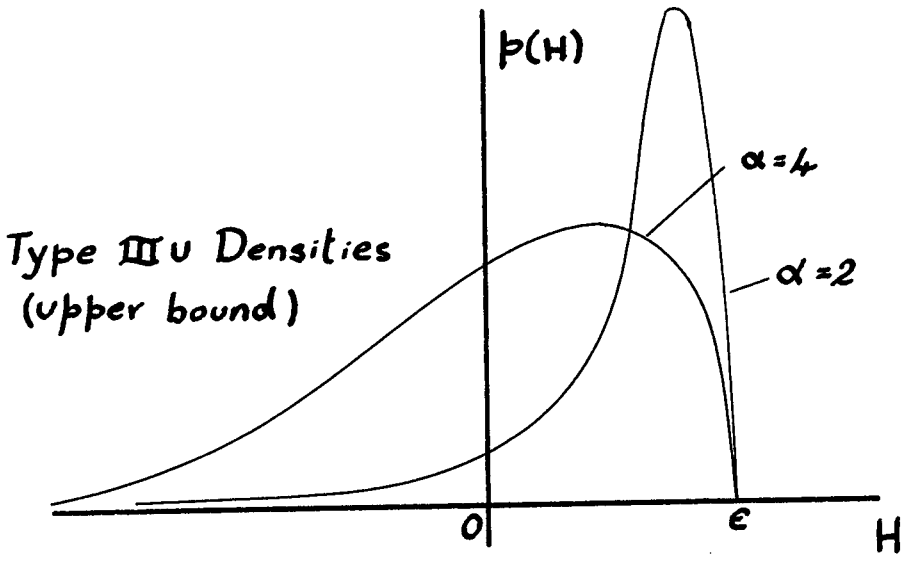


FIGURE 2 Typical Density Curves for the Types I, II and Lognormal Distributions



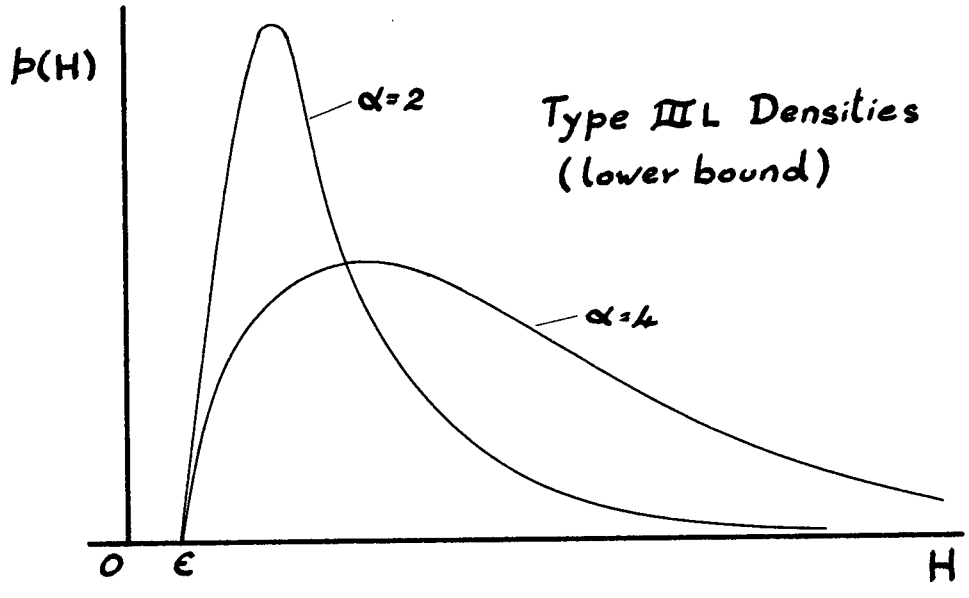
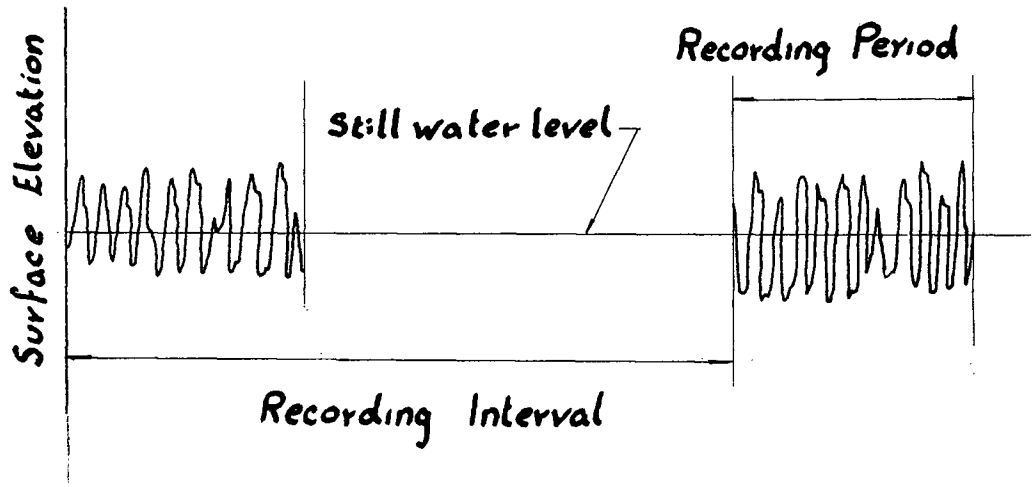


FIGURE 3 Typical Density Curves for the Type III Distribution



NOTE: The wave height is measured from trough to crest

FIGURE 4 Typical Wave Elevation Recording

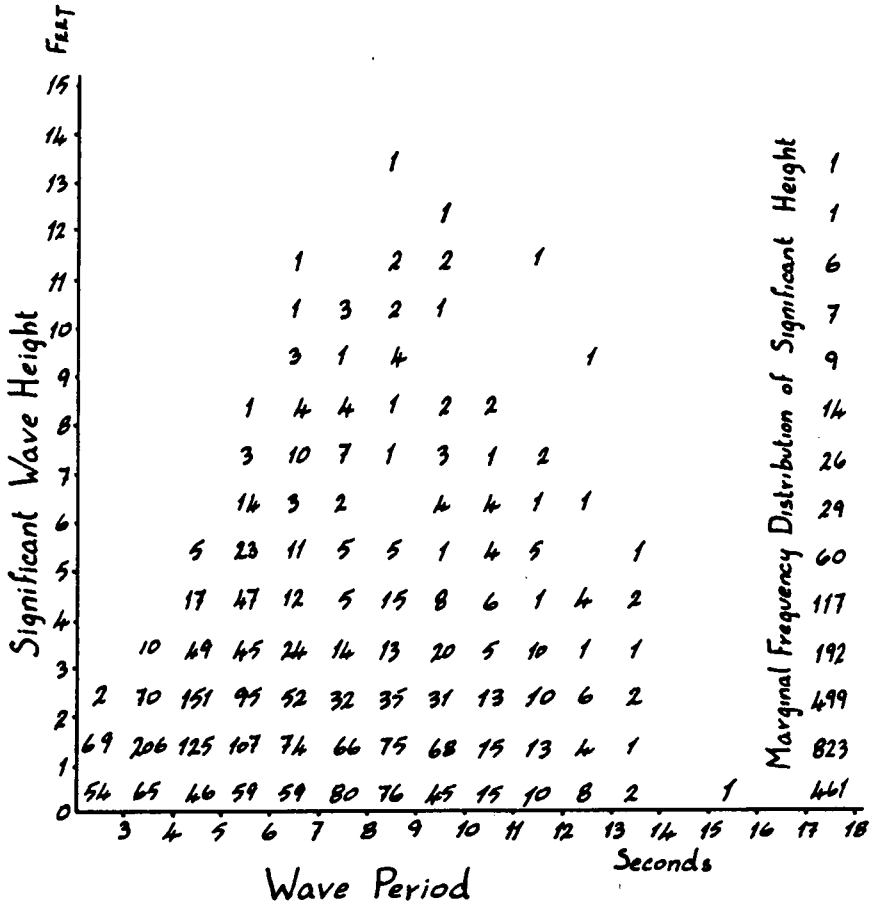


FIGURE 5 A Bivariate Scatter Diagram

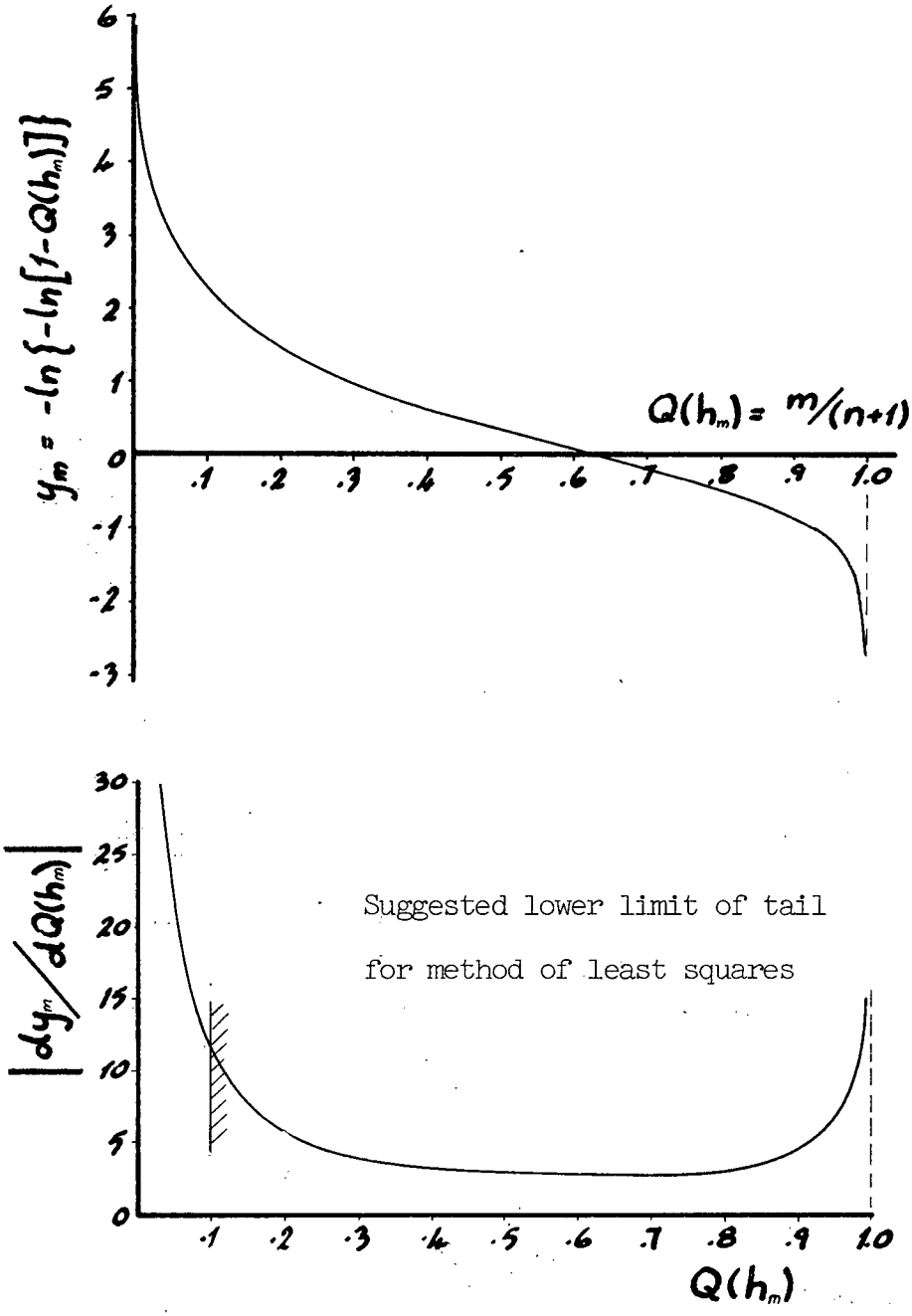


FIGURE 6 The Definition of Sample Tail

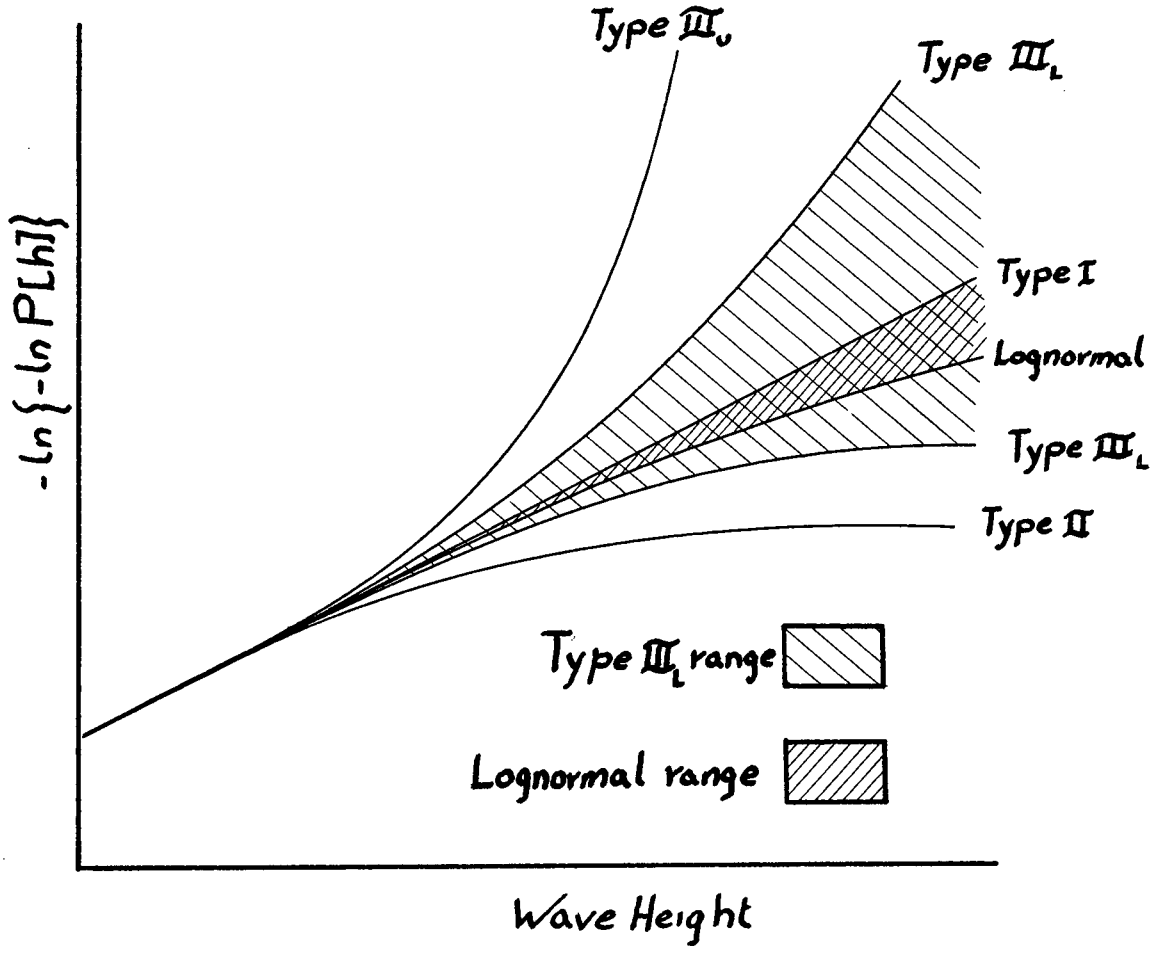


FIGURE 7 Comparison of Tail Curvatures on Gumbel paper

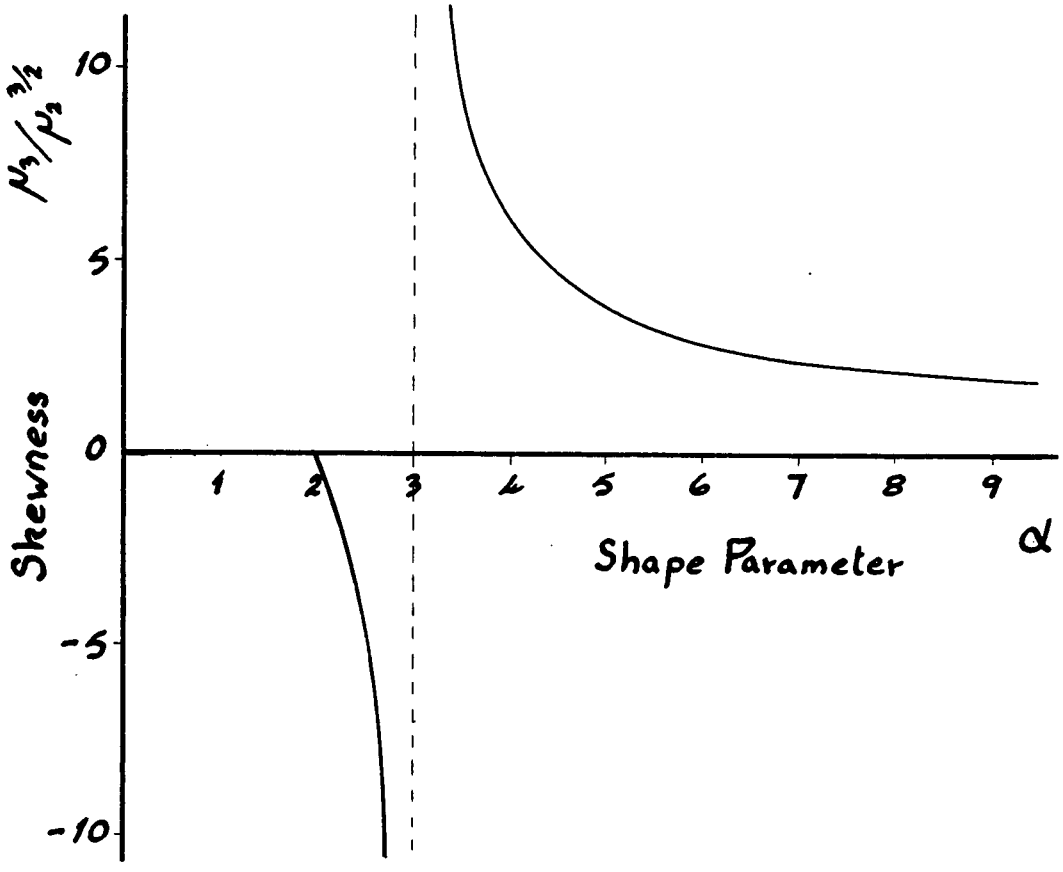


FIGURE 8 Skewness of the Fretchet Distribution

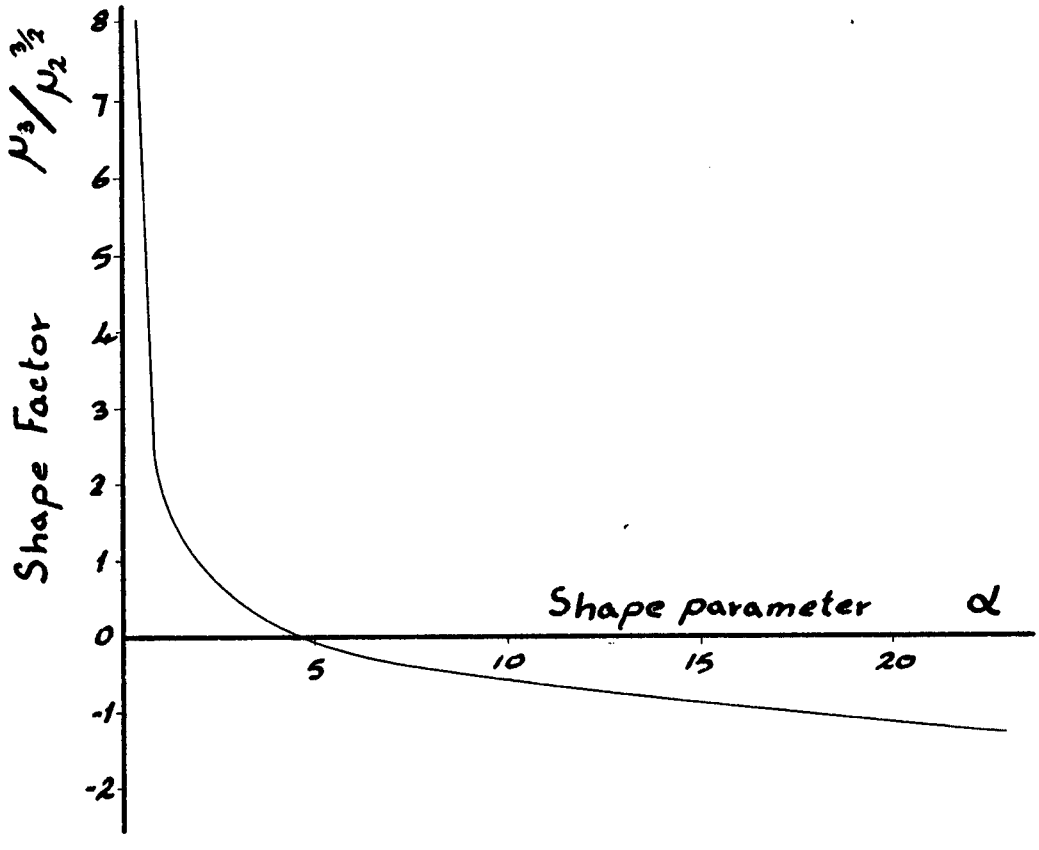


FIGURE 9 Shape Factor of the Weibull Distribution

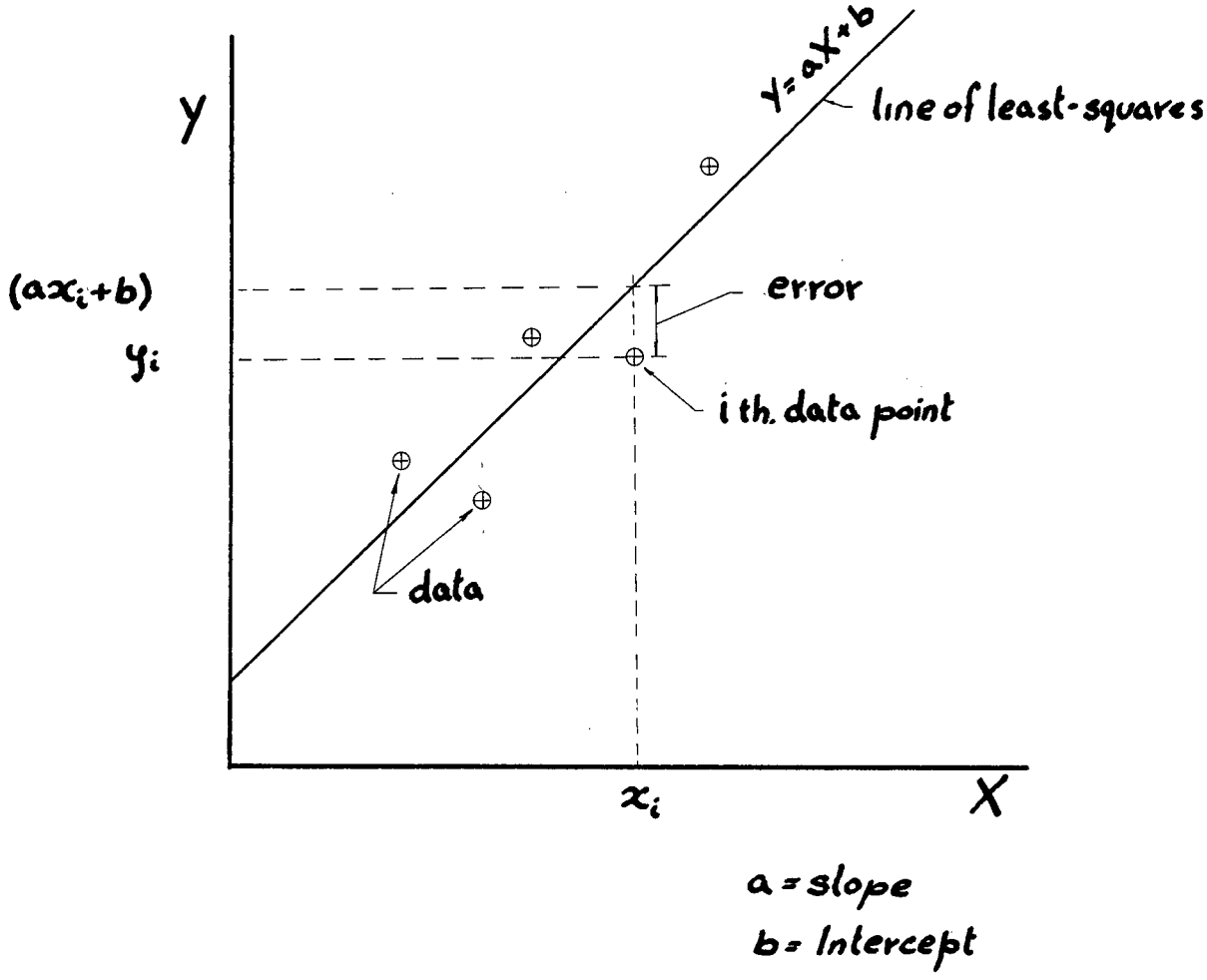


FIGURE 10 Method of Least Squares

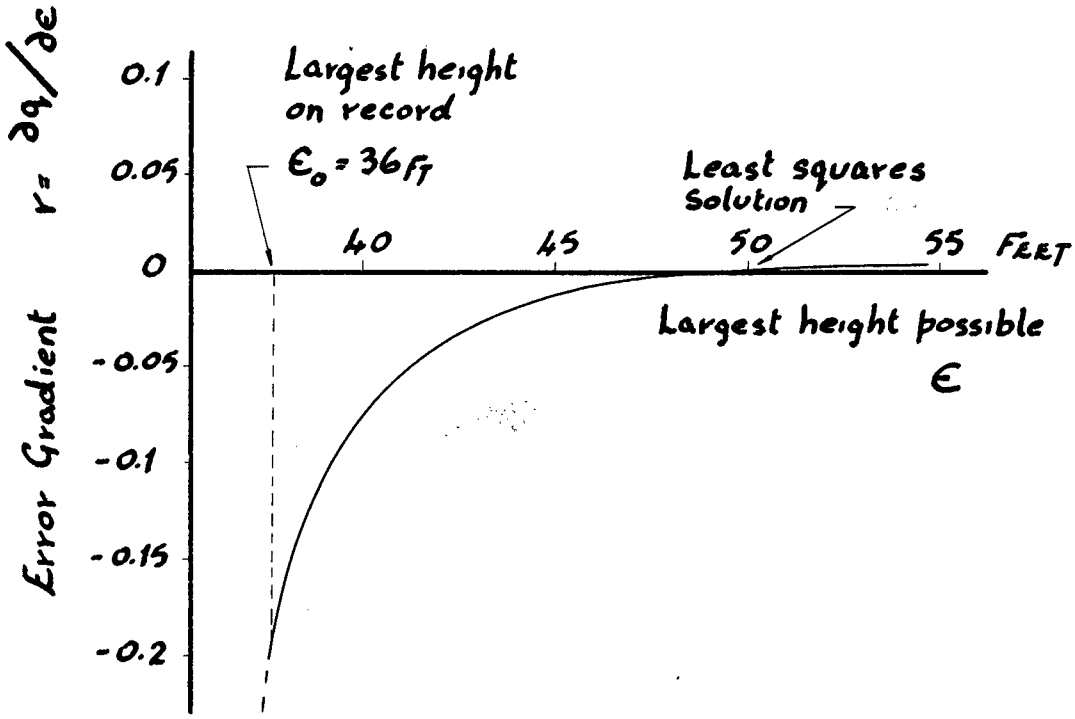


FIGURE 11 Convergence of Least Squares Procedure for Type III_U Distribution

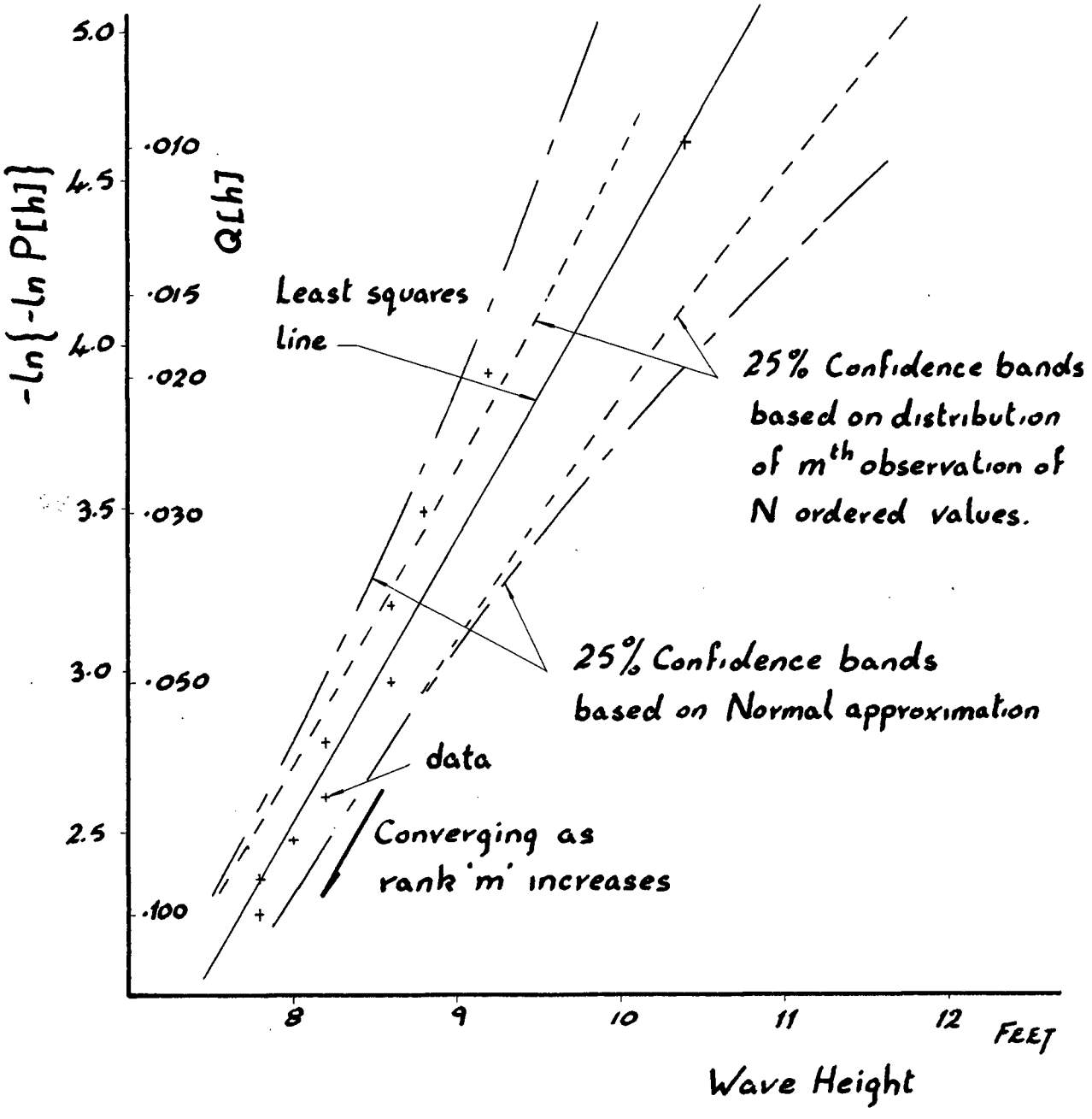


FIGURE 12 Comparison Between the Approximate and Exact Confidence Intervals

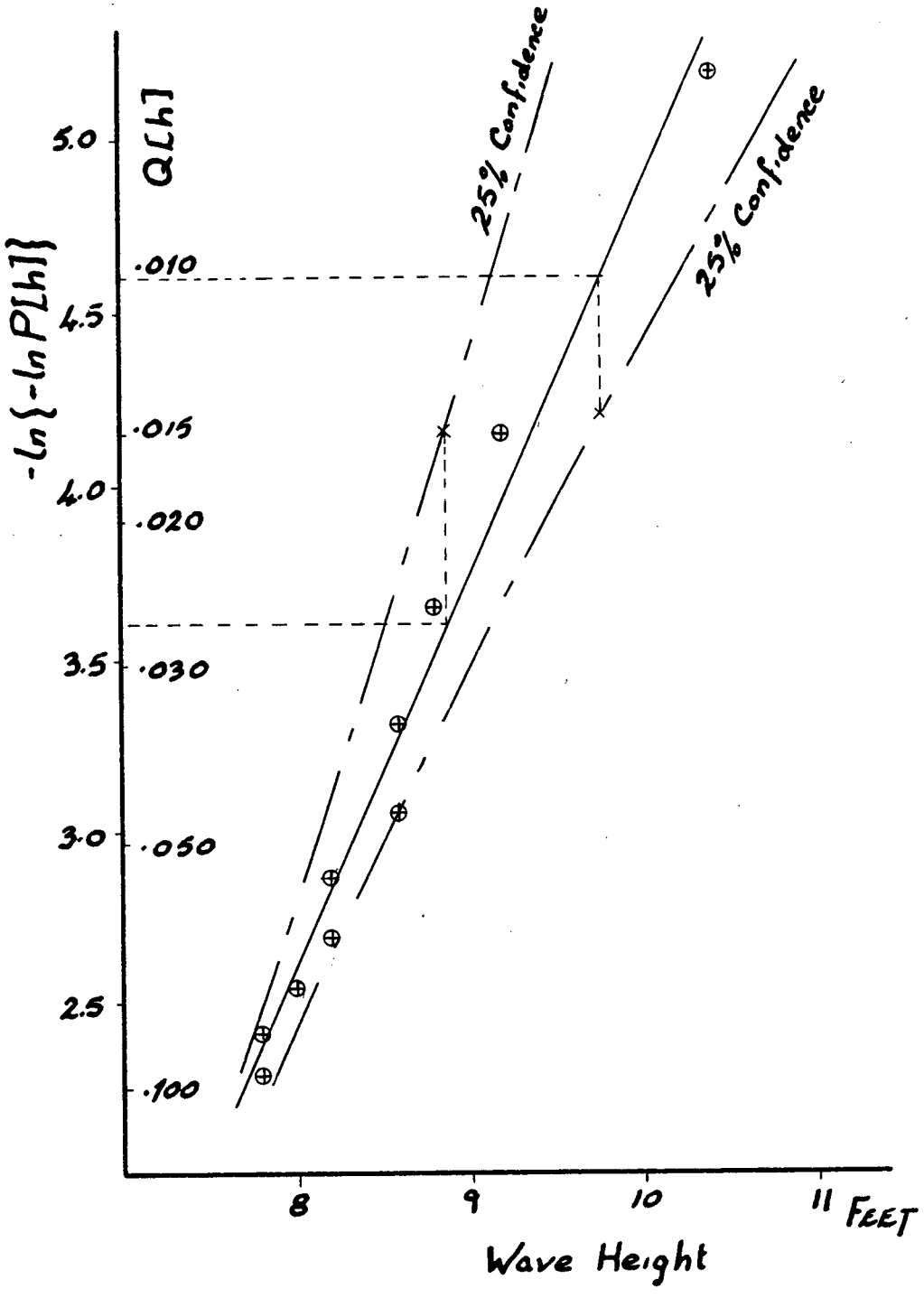
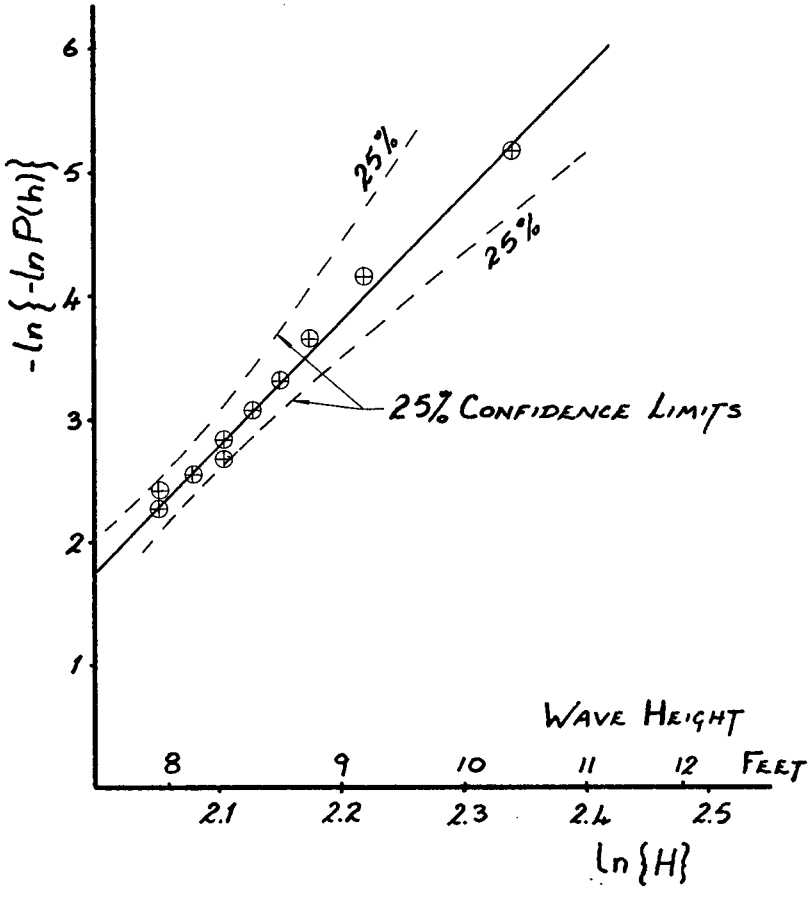
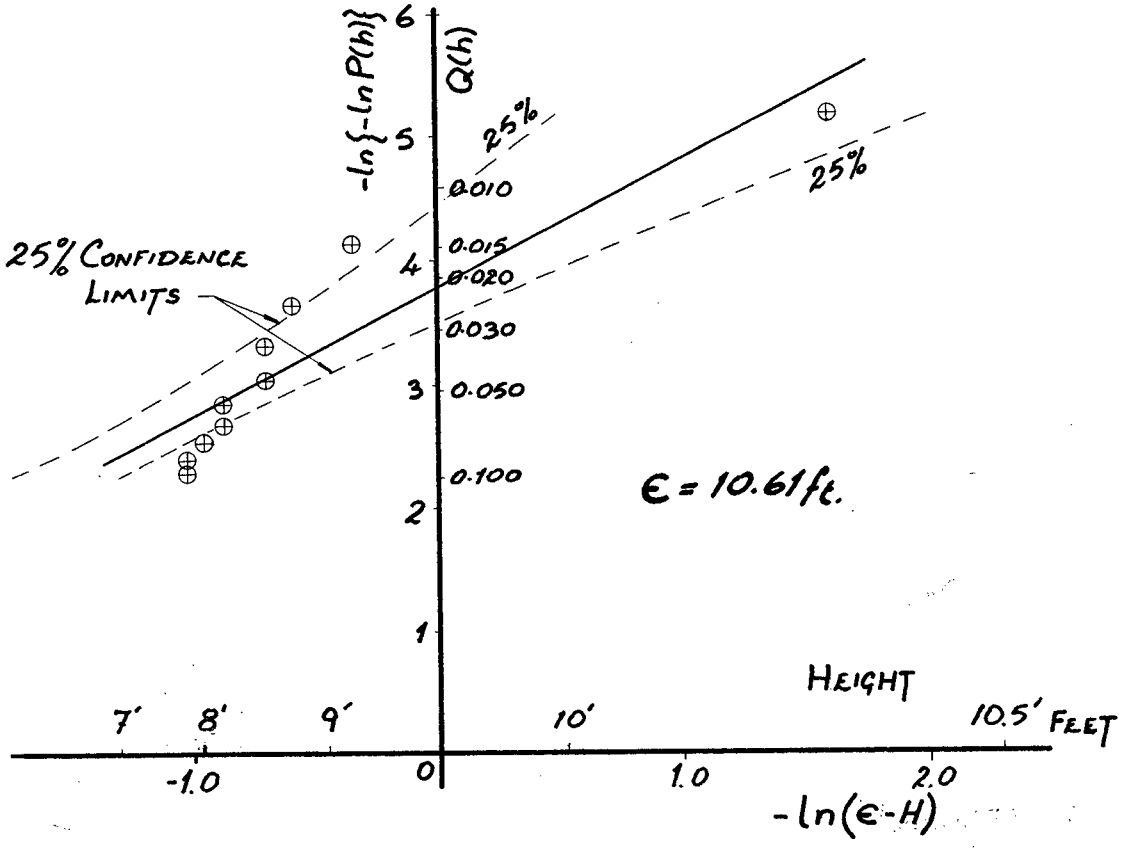


FIGURE 13 Confidence Intervals on a Type I Plot



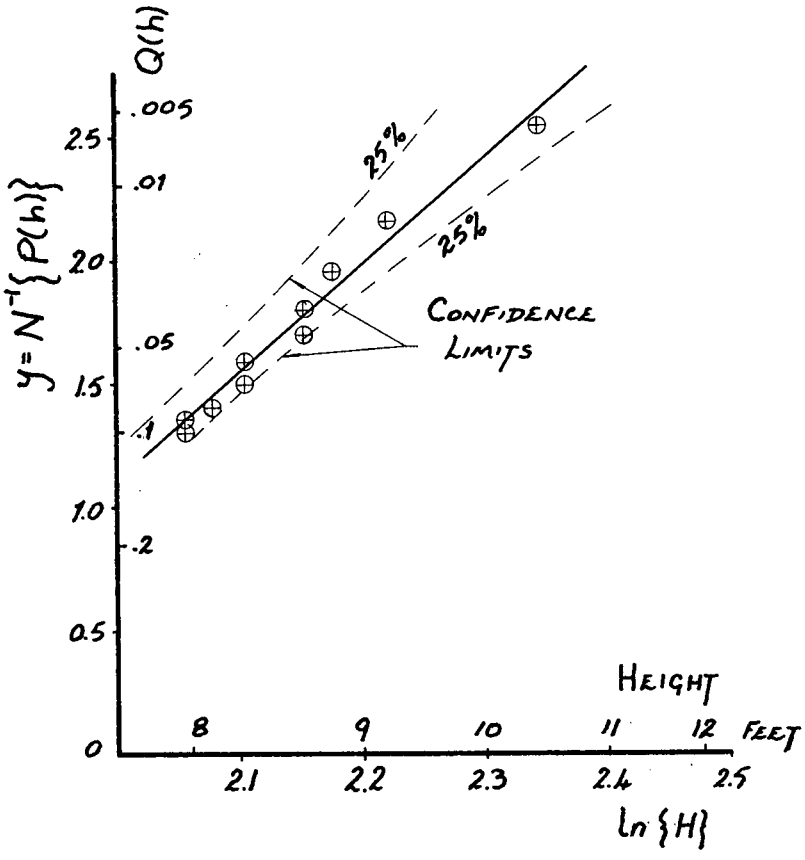
Note: data from St. Denis [1969]

FIGURE 14 Confidence Intervals on a Type II plot



NOTE: Data from St. Denis [1969]

FIGURE 15 Confidence Intervals on a Type III Plot



NOTE: i) $N^{-1}\{P(h)\}$ is the value of variate which corresponds to an area equal to $P(h)$ under the Standard Normal Distribution density curve.

ii) Data from St. Denis [1969]

FIGURE 16 Confidence intervals on Lognormal Plot

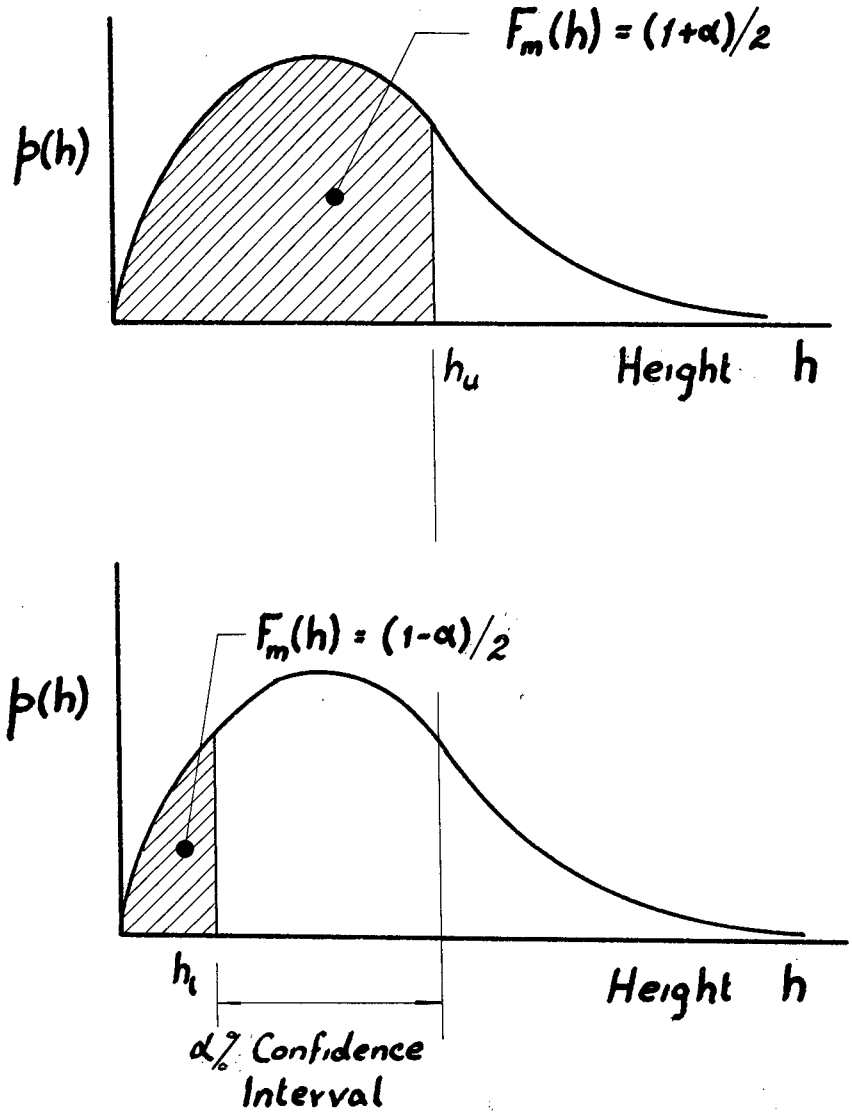


FIGURE 17 Determination of Confidence Interval from Typical Distribution Density of the mth Observation.

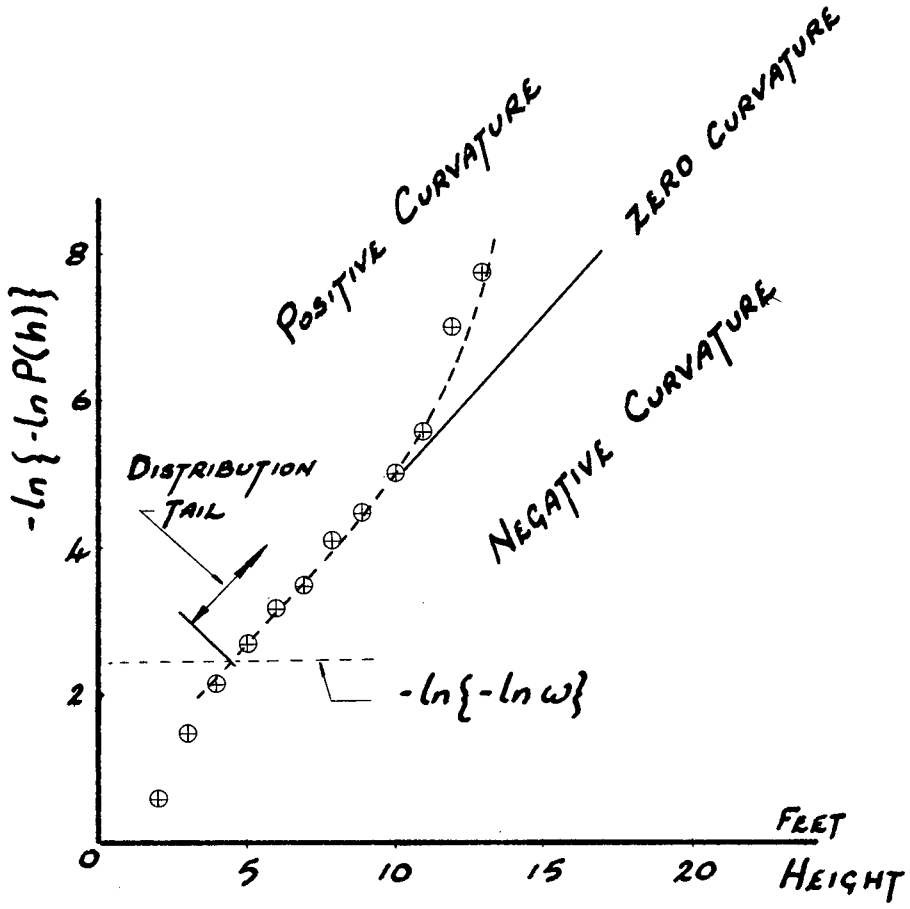


FIGURE 18 The Curvature Test

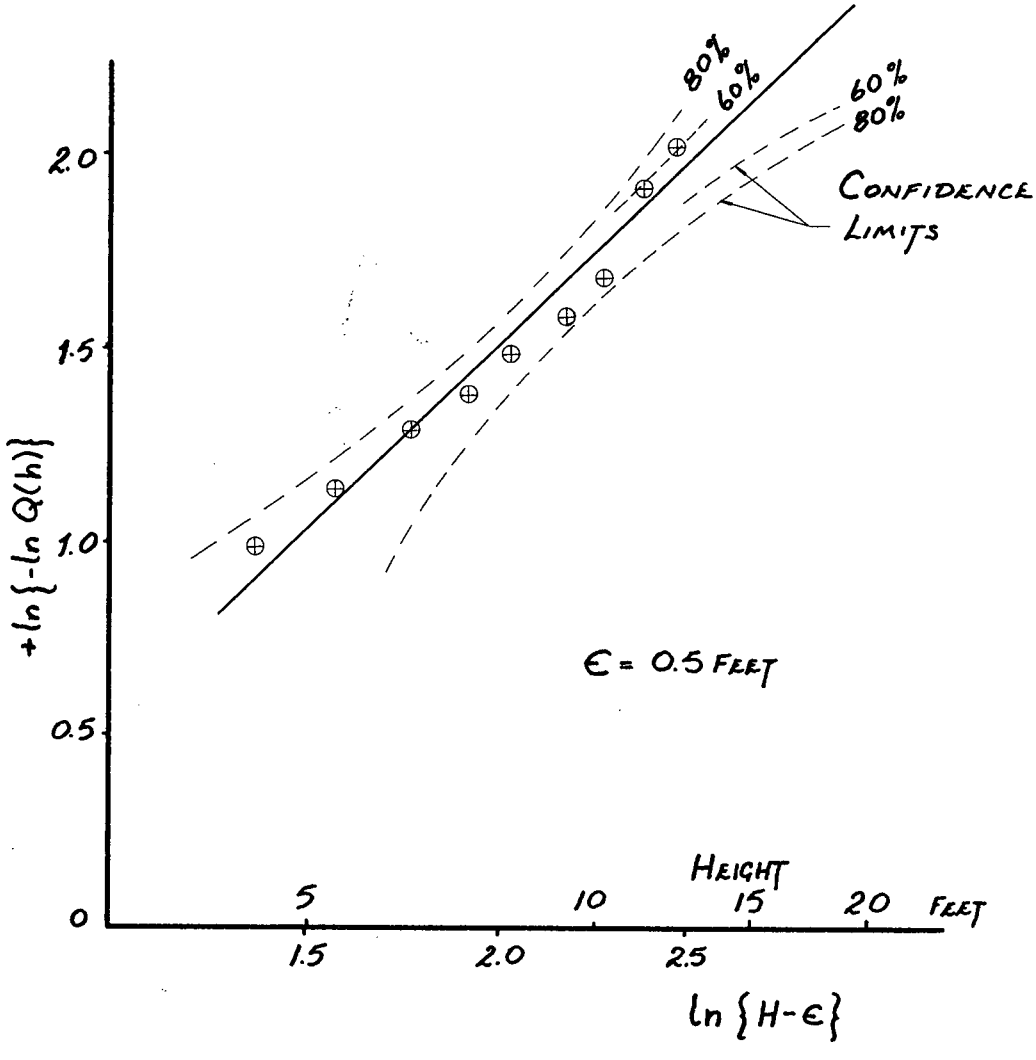


FIGURE 19 Type III_L Plot with Confidence Intervals

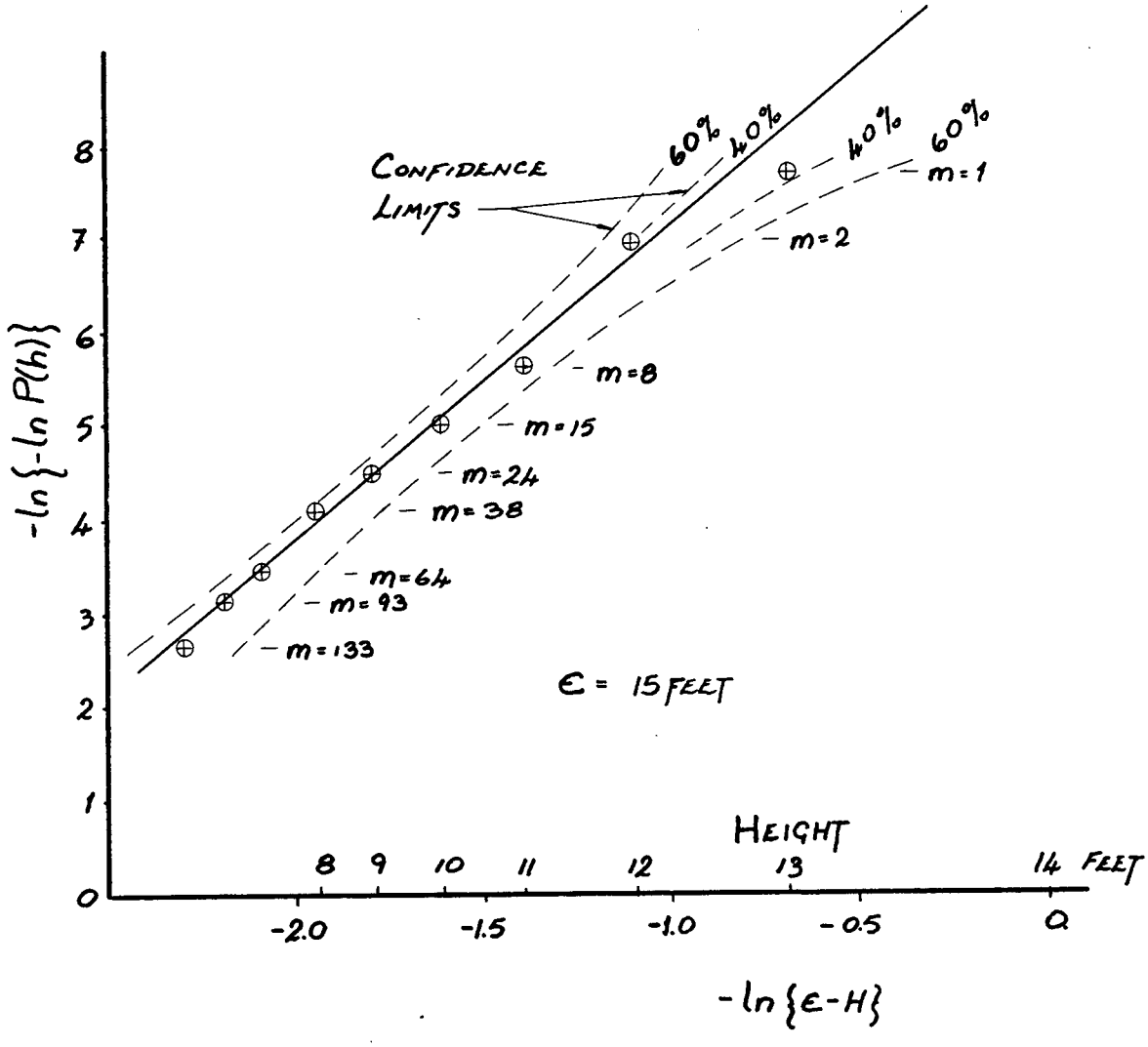
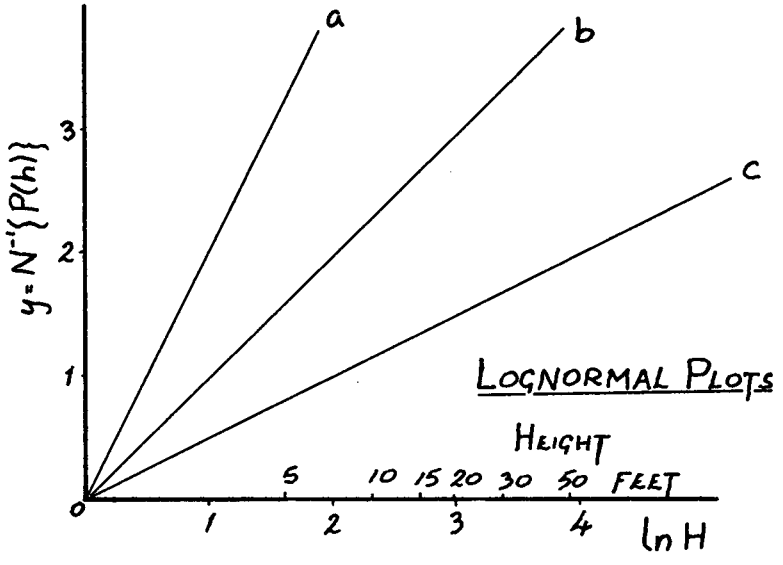
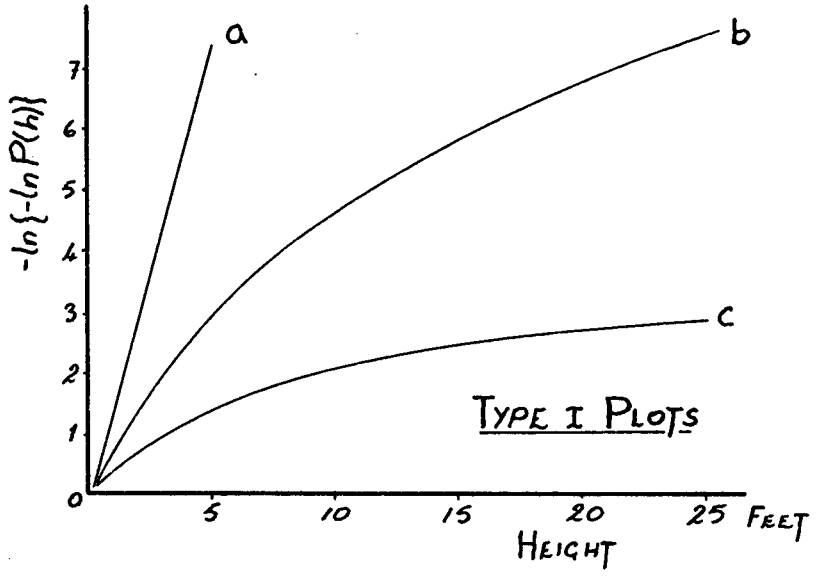


FIGURE 20 Type III_U Plot with Confidence Intervals





Note: $N^{-1}\{P(h)\}$ is the value of variate which corresponds to an area equal to $P(h)$ under the Standard Normal Distribution density curve

FIGURE 21 Curvature of the Lognormal Distribution on Type I Paper

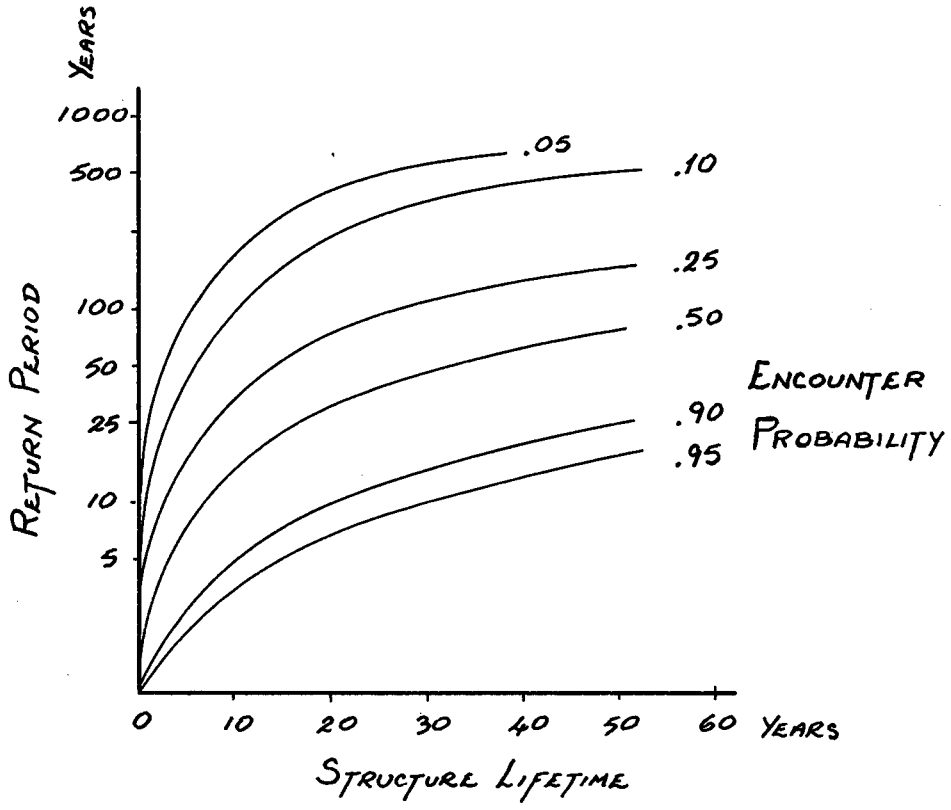


FIGURE 22 The Relationship Between Return Period and Encounter Probability

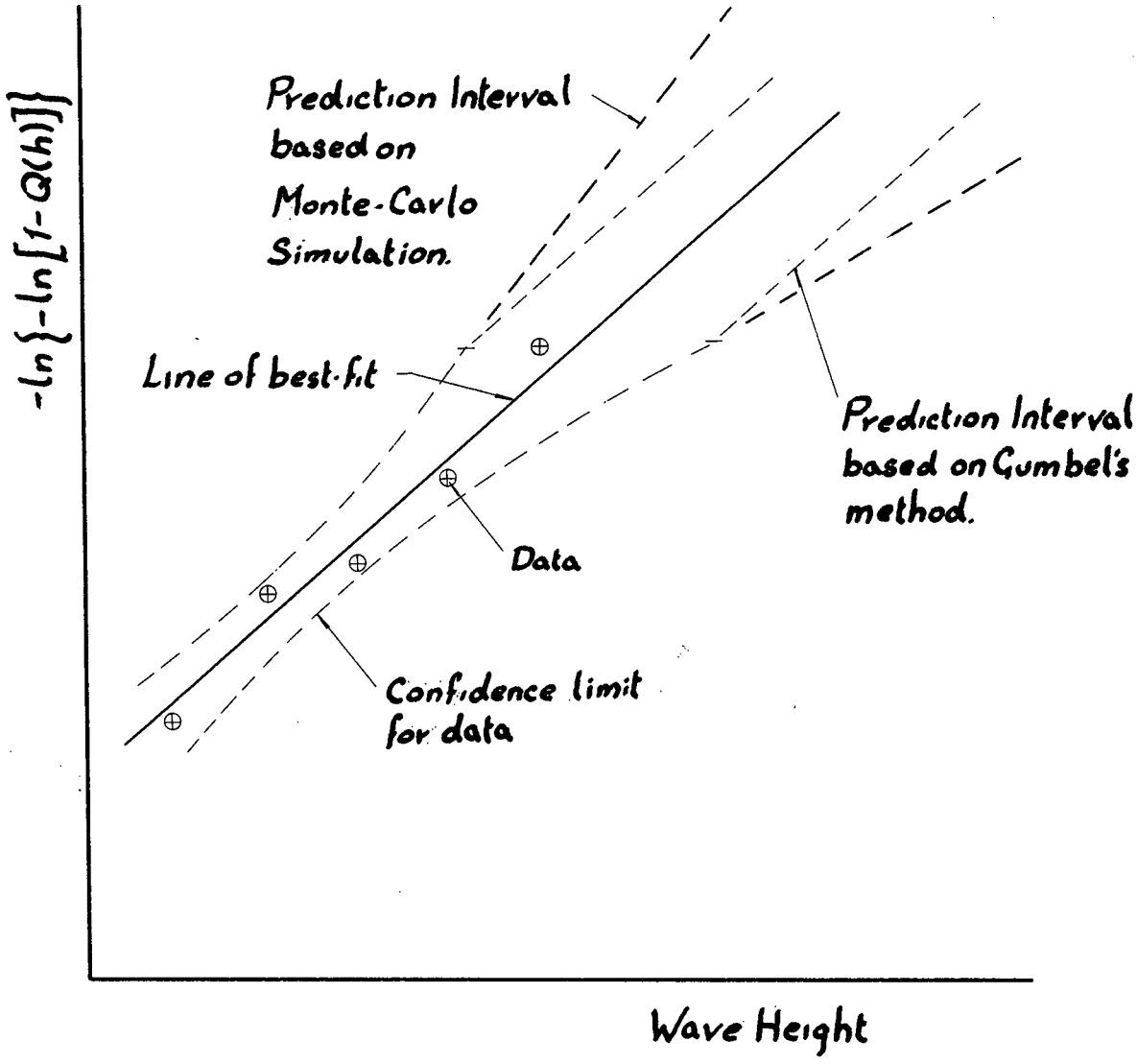


FIGURE 23 Typical Prediction Intervals

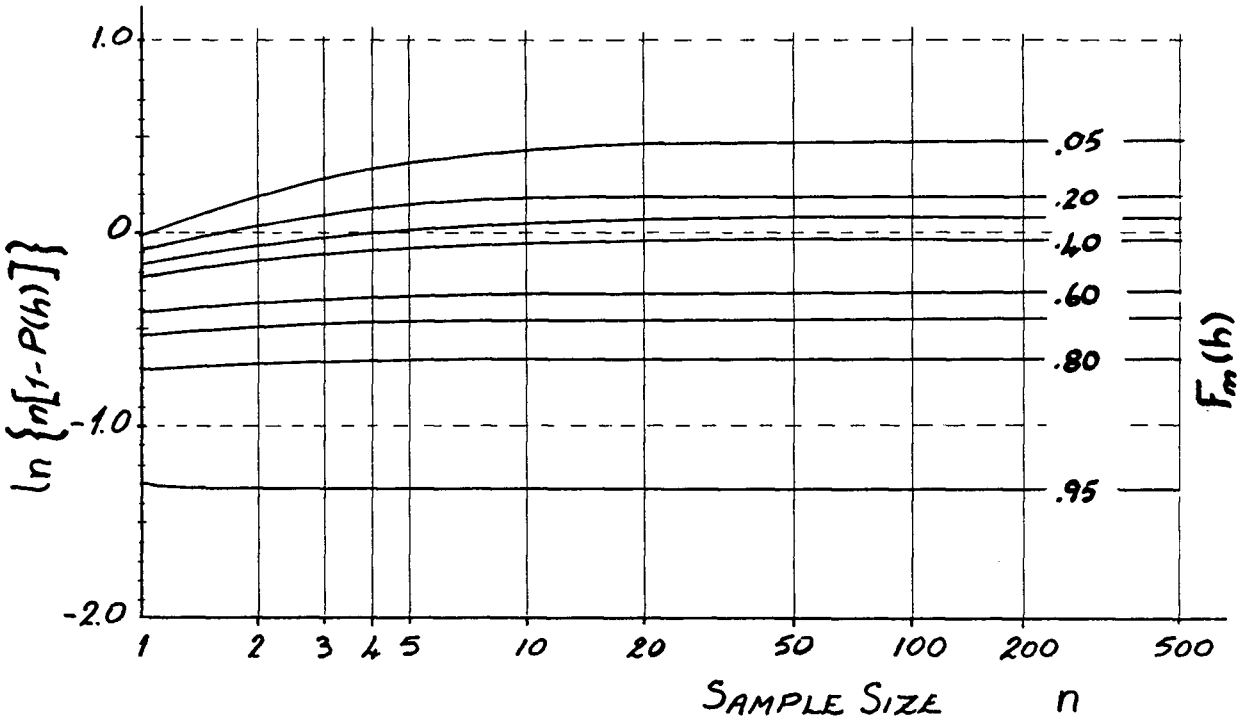


FIGURE 24 The Relationship Between P(h) and $F_m(h)$ for m = 1.

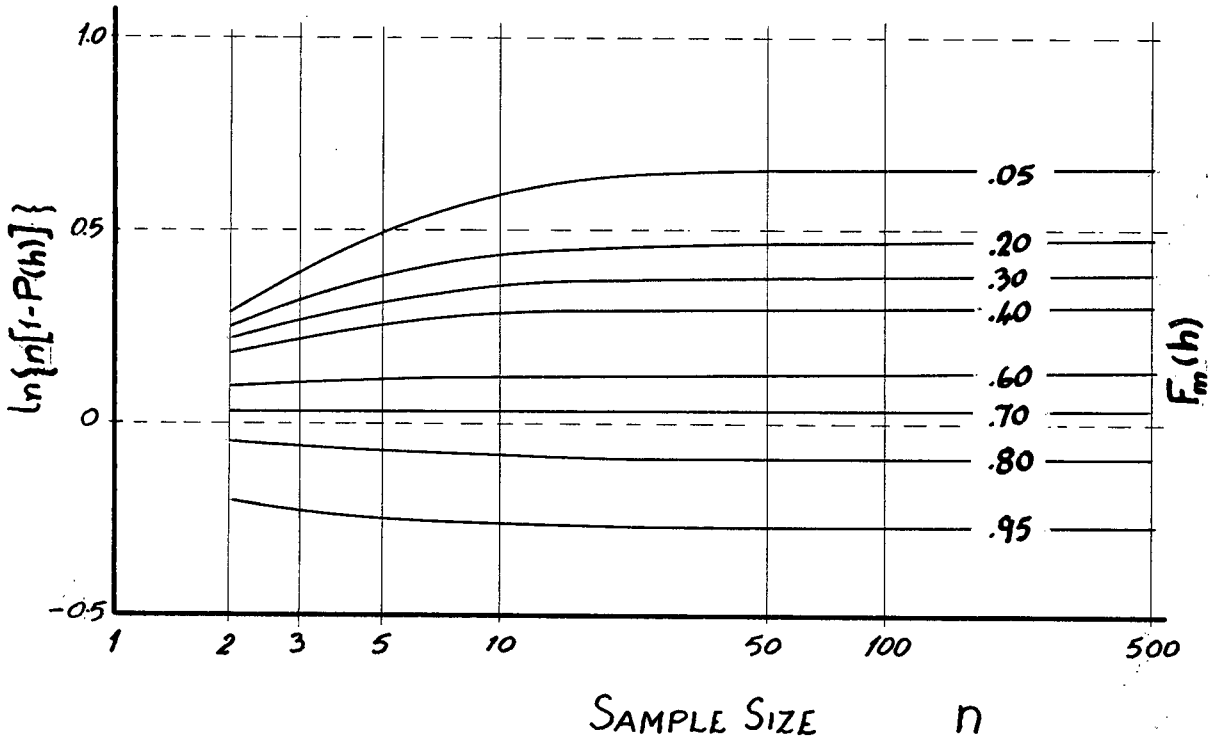


FIGURE 25 The Relationship Between P(h) and $F_m(h)$ for m = 2.

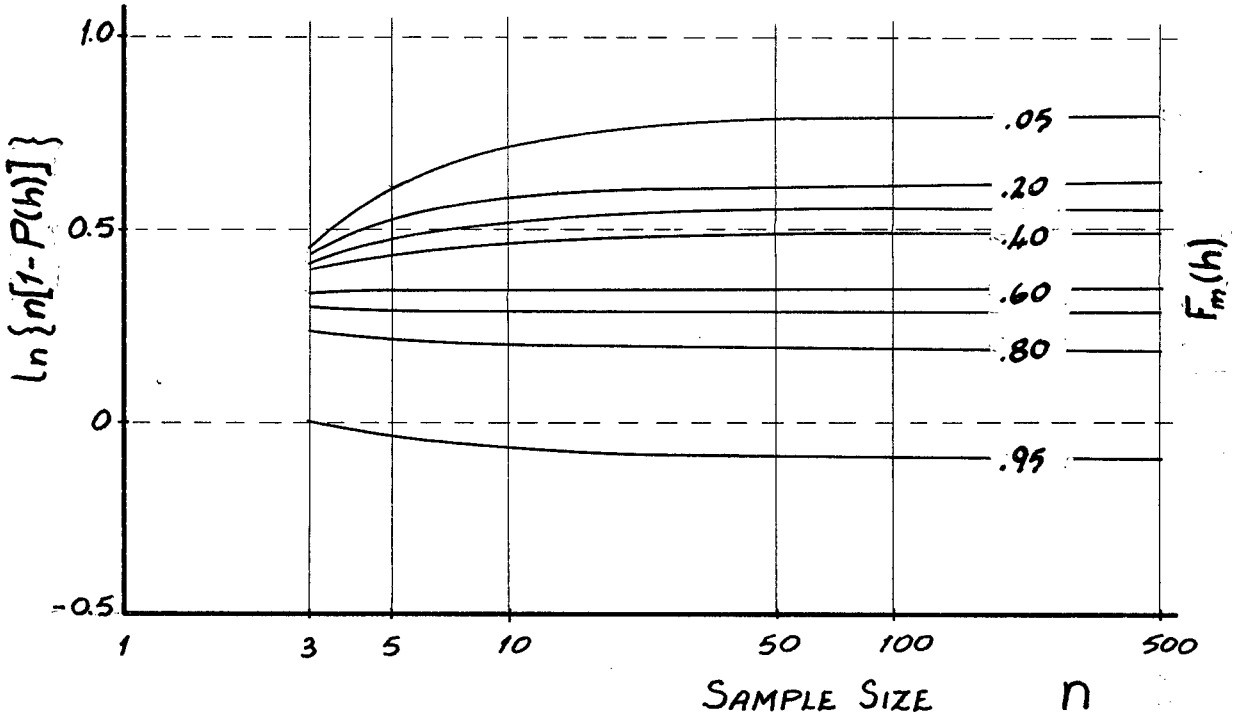


FIGURE 26 The Relationship Between P(h) and $F_m(h)$ for m = 3.

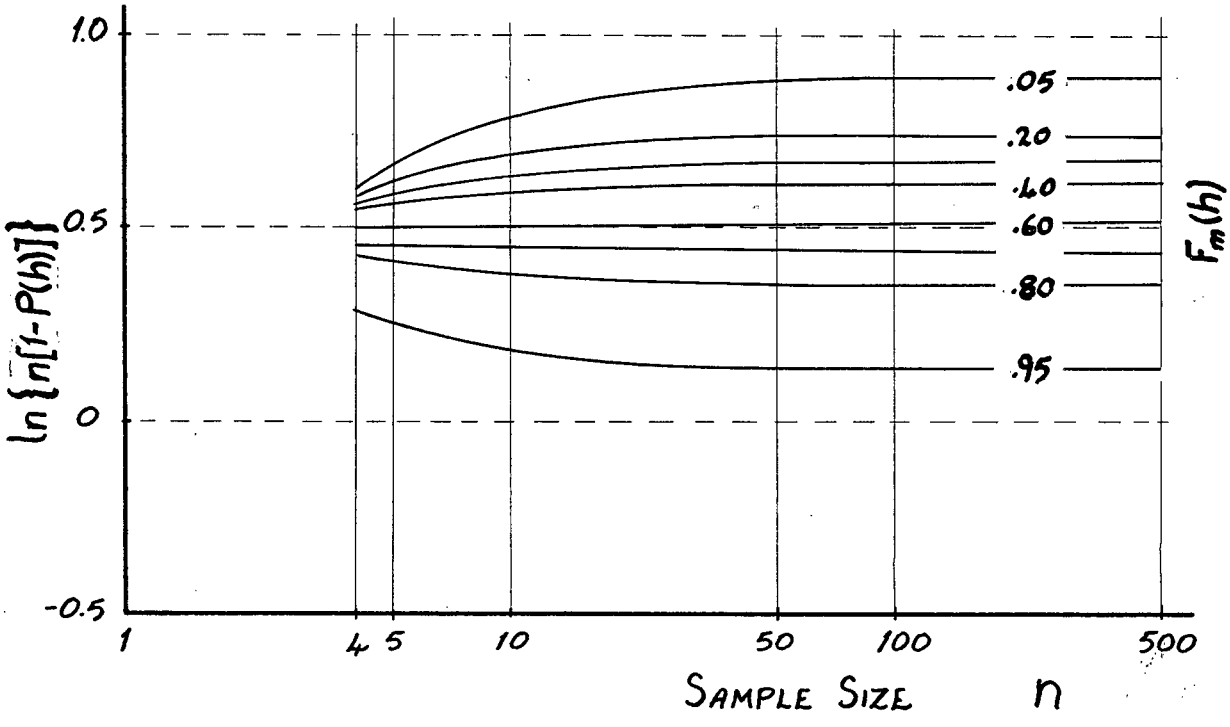


FIGURE 27 The Relationship Between P(h) and $F_m(h)$ for m = 4.

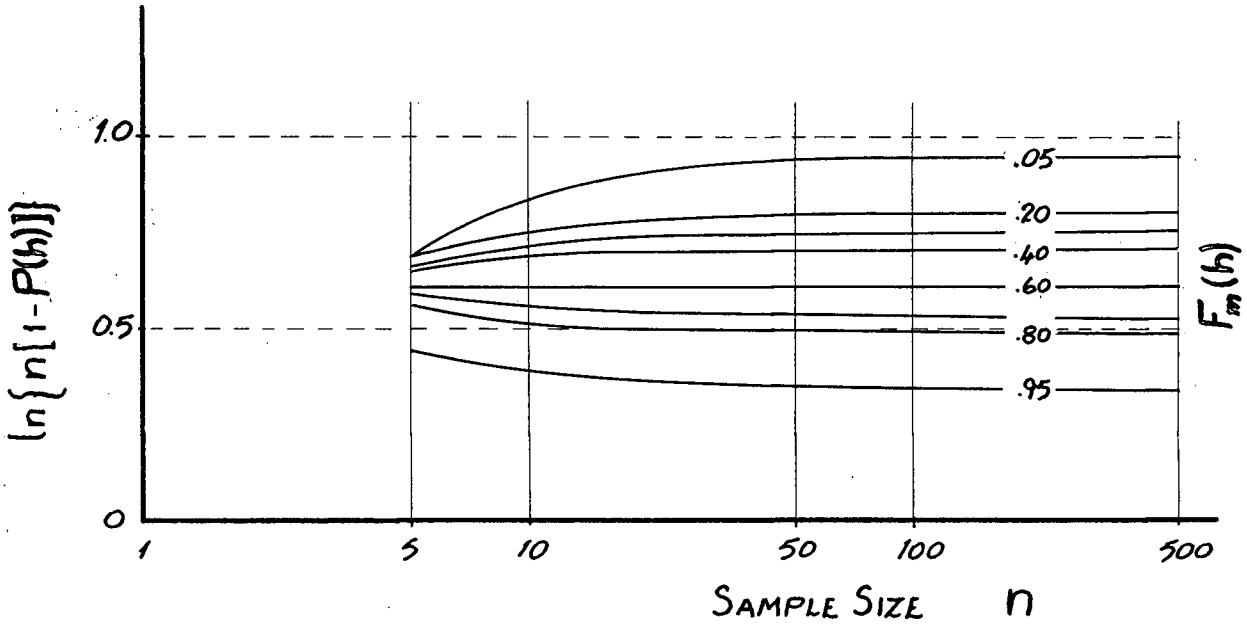


FIGURE 28 The Relationship Between P(h) and $F_m(h)$ for m = 5.

APPENDIX

A.1 Properties of the Type I Distribution

The cumulative probability function is given by

$$P(h) = \exp\{-\exp[-\frac{h-\varepsilon}{\theta}]\} \qquad \ldots.(68)$$

$$\text{let} \quad z = \frac{h-\varepsilon}{\theta} \qquad \ldots.(69)$$

where ε is a location parameter

and θ is a scale parameter

$$P(z) = \exp\{-\exp(-z)\} \qquad \ldots.(70)$$

the density function is then given by

$$p(z) = \frac{d}{dz} P(z) = \exp\{-z\} \exp\{-\exp(-z)\}$$

$$\text{Type I density} \quad p(z) = \exp\{-z - \exp[-z]\} \qquad \ldots.(71)$$

A moment generating function is defined as

$$M\{t\} = \int_{-\infty}^{\infty} \exp\{tz\}\, p(z)\, dz \qquad \ldots.(72)$$

where t is a dummy variable.

set $y = \exp(-z)$

$$M\{t\} = \int_{0}^{\infty} y^{(1-t)-1} \exp\{-y\}\, dy$$

$$= \Gamma(1-t) \qquad \ldots.(73)$$

which is the Gamma function with argument $(1-t)$.

The basic properties of the moment generating function [Bury 1975; page 44] give the kth moment about the origin

$$m_k(z) = \frac{d^k}{dk^t} M\{t\} \Big|_{t=0}$$

thus $$m_1(z) = \frac{d}{dt} \Gamma(1-t) \Big|_{t=o} = \gamma \qquad \ldots.(74)$$

where γ is Euler's Number (0.57722).

Similarly higher moments are found to be

$$m_2(z) = \gamma^2 + \frac{\pi^2}{6} = 1.97811$$

$$m_3(z) = 5.44487$$

$$m_4(z) = 23.56147$$

The first four central moments are given as

$$\mu_1(z) = 0$$

$$\mathrm{Var}(z) = \mu_2(z) = \frac{\pi^2}{6} = 1.64493$$

$$\mu_3(z) = 2.40411$$

$$\mu_4(z) = 14.61136$$

The first shape factor is constant and given by

$$\frac{\mu_3}{\mu_2^{3/2}} = 1.13955$$

Since the shape factor is constant, the Type I distribution is only capable of having one shape which is shown in FIGURE 2.

The estimation of parameters used in Eqn. 68 is achieved by using the properties of the parameterless version, Eqn. 70.

from Eqn. 71

$$E\{z\} = E\left\{\frac{h-\varepsilon}{\theta}\right\} = \frac{m_1(h)}{\theta} - \frac{\varepsilon}{\theta}$$

$$m_1(h) = \varepsilon + 0.57722\ \theta \qquad \ldots.(75)$$

Similarly

$$Var(z) = m_2(z) - m_1{}^2(z)$$

$$= \frac{\mu_2(h)}{\theta^2}$$

$$\mu_2(h) = \frac{\pi^2}{6}\ \theta^2 \qquad \ldots.(76)$$

Equations 75 and 76 form the basis of the method of moments described in SECTION 5.1.1.

A2 Properties of the Type II Distribution

The cumulative probability function is given by

$$P(h) = \exp\left[-\left(\frac{h}{\theta}\right)^{-\alpha}\right] \qquad \ldots.(77)$$

$$h \geq 0$$

$$\theta \geq 0$$

$$\alpha > 0$$

where α is a shape parameter

and θ is a scale parameter

The density is

$$p(h) = \frac{d}{dh} P(h) = \frac{\alpha}{\theta}\left(\frac{h}{\theta}\right)^{-(\alpha+1)} \exp\left[-\left(\frac{h}{\theta}\right)^{-\alpha}\right] \qquad \ldots.(78)$$

The kth moment is defined as

$$m_k(h) = \int_0^\infty h^k \, p(h) \, dh$$

setting $t = \left(\frac{h}{\theta}\right)^{-\alpha}$

$$\theta^{-\alpha} t = h^{-\alpha}$$

$$dt = \frac{\alpha}{\theta}\left(\frac{h}{\theta}\right)^{-(\alpha+1)} dh$$

$$m_k(t) = \int_0^\infty \theta^k \, t^{-k/\alpha} \, e^{-t} \, dt$$

$$m_k(t) = \theta^k \, \Gamma\{1 - (k/\alpha)\} \qquad \ldots.(79)$$

The function $\Gamma\{1-k/\alpha\}$ is discontinuous for all integer values of (k/α), and this distribution is only valid when $(k/\alpha) < 1$. [Gumbel 1958 pages 262-264]. This implies that when α is integer, the expectation or moment of order α will not exist. In addition, when $\alpha \leq 1$ the Type II distribution does not have a mean, and its variance cannot exist for $\alpha \leq 2$. Thus, this distribution must be used with considerable care.

A3 Properties of the Type III_U Distribution

The cumulative probability function is

$$P(h) = \exp\left\{-\left(\frac{\varepsilon-h}{\theta}\right)^{\alpha}\right\} \qquad \ldots.(80)$$

$$\text{for} \quad \varepsilon - h \geq 0$$

$$\theta > 0 \; ; \quad \alpha > 0$$

where ε is the highest wave ever possible

i.e. $P(\varepsilon) = 1.0$

and θ is a scale parameter.

The distribution may be simplified by setting

$$\lambda = (\varepsilon - h)$$

the density then becomes

$$P(\lambda) = \frac{d}{d\lambda} P(\lambda) = \frac{\alpha}{\theta}\left(\frac{\lambda}{\theta}\right)^{\alpha-1} \exp\left\{-\frac{\lambda}{\theta}\right\}^{\alpha} \qquad \ldots.(81)$$

again defining the kth moment by

$$M_k(\lambda) = \int_{-\infty}^{\infty} \lambda^k \, p(\lambda) \, d\lambda$$

$$= \int_{-\infty}^{\infty} \frac{\alpha}{\theta}\left(\frac{\lambda}{\theta}\right)^{\alpha-1} \exp\left\{-\left(\frac{\lambda}{\theta}\right)^{\alpha}\right\} \lambda^k \, d\lambda$$

setting $t = \left(\frac{\lambda}{\theta}\right)^{\alpha}$

gives the gamma integral

$$M_k(\lambda) = \int_{0}^{\infty} \theta^k \, t^{k/\alpha} \, e^{-t} \, dt$$

$$M_k(\lambda) = \theta^k \, \Gamma\{1 + (k/\alpha)\} \qquad \ldots.(82)$$

The size of the moment is independent of ε. Unlike the moments of the Type II distribution, the moments of the Type III_U distribution are continuous. Equation 82 forms the basis of the method of moments described in SECTION 5.1.3. The treatment for the Type III_L distribution is basically the same and is easily derived by the same method.

A4 Properties of the Lognormal Distribution

The lognormal model is obtained by using the logarithm of the height as a reduced variate, and applying the Normal model.

The reduced variate

$$y = \ell n(h)$$

and the NORMAL probability function

$$P(h) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^h \frac{1}{h} \exp\{-\frac{1}{2}\left(\frac{\ell nh - \mu}{\sigma}\right)^2\} dh$$

yield the Lognormal probability function

$$P(h) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^h \frac{1}{h} \exp\{-\frac{1}{2}\left(\frac{\ell nH - \mu}{\sigma}\right)^2\} dh \qquad \ldots.(83)$$

where μ is the scale parameter

and σ is a shape parameter

The density is given by

$$p_\ell(h) = \frac{1}{h\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2}\left(\frac{\ell nh - \mu}{\sigma}\right)^2\} \qquad \ldots.(84)$$

for $h > 0$

$\sigma > 0$; $-\infty < \mu < \infty$

The kth moment is given as

$$M_k(h) = \int_{-\infty}^{\infty} h^k p_\ell(h)\, dh$$

setting $y = \ell n(h)$

$$M_k(h) = \int_{-\infty}^{\infty} \exp\{ky\}\, p_n(y)\, dy \qquad \ldots.(85)$$

where $p_n(y)$ is the Normal density.

Equation 85 is the standard form of the moment generating function with argument k, giving

$$M_k(h) = \exp\{\mu k + \frac{k^2\sigma^2}{2}\}$$

The first moment is then

$$E\{h\} = m_1(h) = \exp\{\mu + \frac{\sigma^2}{2}\} \qquad \ldots.(86)$$

and the central moments are

$$\mu_1(h) = 0$$

$$\mu_2(h) = [\exp(\sigma^2) - 1]\exp(2\mu+\sigma^2) \qquad \ldots.(87)$$

$$\mu_3(h) = m_1^3(\gamma^6 - 3\gamma^4) \qquad \ldots.(88)$$

where γ is the coefficient of dispersion

$$\gamma = \mu_2^{\frac{1}{2}}/m_1 = [\exp(\sigma^2) - 1]^{\frac{1}{2}}$$

The first shape factor is given by

$$\phi = \gamma^3 + 3\gamma \qquad \ldots.(89)$$

$$> 0$$

indicating that all lognormal densities are skewed to the right. The simultaneous solution of Eqns. 86 and 87 leads to the method of moment estimators given in Eqns. 22 and 23 of SECTION 5.1.4.

A5 Curvature of the Type II Distribution

The ordinate scale of Type I paper is given by

$$y = -\ln\{-\ln P(h)\}$$

the Type II probability function is

$$P(h) = \exp\{-\left(\frac{h}{\theta}\right)^{-\alpha}\}$$

Thus the equation of a Type II line on Type I paper is

$$y = \alpha(\ln h - \ln \theta) \qquad \ldots.(90)$$

which has slope

$$\frac{dy}{dh} = \alpha/h$$

and curvature

$$\frac{d^2y}{dh^2} = \frac{-\alpha}{h^2} \leq 0$$

Hence the Type II distribution will have a negative curvature when plotted on Type I paper.

A6 Curvature of the Type III_U Distribution

The Type III_U probability function is given by

$$P(h) = \exp\left\{-\left[\frac{\varepsilon - h}{\theta}\right]^{\alpha}\right\}$$

Again using Eqn. 1 as the ordinate of the Type I paper.

The equation of a Type III_U line on Type I paper is

$$y = -\alpha \ln(\varepsilon - h) + \alpha \ln \theta \qquad \ldots.(91)$$

which has slope

$$\frac{dy}{dh} = \frac{+\alpha}{(\varepsilon - h)}$$

and curvature

$$\frac{d^2y}{dx^2} = \frac{\alpha}{(\varepsilon - h)^2} \geq 0$$

Thus, a Type III_U distribution will have positive curvature when plotted on Type I paper.

A7 Curvature of the Type III_L Distribution

The Type III probability function is

$$P(h) = 1 - \exp\{-\left(\frac{h-\varepsilon}{\theta}\right)^{\alpha}\}$$

as $P(h) \to 1$

$$\ell n\, P(h) = \ell n\,[1 - \exp\{-\left(\frac{h-\varepsilon}{\theta}\right)^{\alpha}\}]$$

$$\simeq -\exp\{-\left(\frac{h-\varepsilon}{\theta}\right)^{\alpha}\}$$

Using Eqn. 1 as the ordinate of the Type I paper, the equation of the Type III_L distribution becomes:

$$y \simeq \left(\frac{h-\varepsilon}{\theta}\right)^{\alpha} \qquad \ldots.(92)$$

which has slope

$$\frac{dy}{dh} = \frac{\alpha}{\theta}\left(\frac{h-\varepsilon}{\theta}\right)^{\alpha-1}$$

and curvature

$$\frac{d^2y}{dh^2} = \frac{\alpha}{\theta}\left(\frac{\alpha-1}{\theta}\right)\left(\frac{h-\varepsilon}{\theta}\right)^{\alpha-2}$$

since θ, α and $(h-\varepsilon)$ are positive

sign of $\frac{d^2y}{dh^2}$ = sign of $(\alpha-1)$

Thus the curvature of the Type III_U distribution becomes:

i) positive when $\alpha > 1$

ii) zero when $\alpha = 1$

iii) negative when $\alpha < 1$