Chapter 11

Statistical Analysis of Extreme Waves

11.1 Introduction

11.1.1 *Data for Extreme Wave Analysis*

(A) *Preparation of sample*

The first step in designing a maritime structure is the selection of design waves. In most cases, storm wave heights which would be exceeded once in a given period of years, say 100 years, are chosen on the basis of statistical analysis of extreme events. Needs for the analysis of extremes arise in many branches of sciences and engineering. Flood discharges for floodplain protection, hurricane winds for suspension bridge designs, and storm surge heights for coastal defense works are well-known examples in civil engineering. In this chapter, statistical techniques frequently used in extreme wave analysis are introduced and discussed.

Depending on the method of selecting a set of wave data (which is called a *sample* in statistics), there are three different approaches. One method tries to utilize the whole data of wave heights observed visually or instrumentally during a number of years. The data are analyzed in a form of cumulative distribution to be fitted to some distribution function. Once a best-fitting distribution function is found, the design wave height is estimated by extrapolating the distribution function to the level of probability which corresponds to a given period of years being considered in design process. This method is called the *total sample method*. Some people call it the *initial distribution method* or the *cumulative distribution function method*.

The other two methods use only the maxima of wave heights in time series data. The *annual maxima method* picks up the largest significant wave height in each year, whereas the *peaks-over-threshold method* takes the peak heights of storm waves over a certain threshold value.

The three methods have their own proponents. The choice to make among these methods is somewhat subjective. One important requisite for a statistical sample is *independency*. It means that individual data in a sample must be statistically independent of each other; in other words, the correlation coefficient between successive data should nearly be zero. Another important requisite is *homogeneity*. Individual data in a sample must have a common parent distribution, all belonging to a single group of data, which is called the *population*. Storm waves during the monsoon season and waves during the off-monsoon season would exhibit some difference in their cumulative distributions, and thus they would belong to different populations. A population of waves generated by tropical cyclones would probably be different from that by extratropical cyclones.

The total sample method is not recommended according to the above requisites. Ocean waves have a tendency of being persistent for many hours. The correlation coefficient between wave heights, 24 hours apart, has been found to have a high value of 0.3 to 0.5;[1] thus independency is not satisfied for data sets of the total sample method. Furthermore, the group of small wave heights is likely to constitute a population different from that of the group of large heights. Van Vledder *et al.*[2] reports a case that the total sample method predicts a 100-year wave height 10% larger than that estimated by the peaks-over-threshold method, probably owing to the influence of low wave height data. Thus, no further discussion will be given to the total sample method.

The annual maxima method and peaks-over-threshold method both satisfy the requisite of independency. The annual maxima method is widely used in the analysis of extreme flood discharges and other data of environmental loads. However, existing data bases of storm waves in various countries rarely cover a period of more than 20 years. Such a short record length of extreme wave data brings forth a problem of low reliability in statistical sense; a small sample size induces a wide range of confidence interval. The peaks-over-threshold method (henceforth abbreviated as POT) can have a relatively large number of data in a sample, and thus have a smaller range of confidence interval. Therefore, the discussion hereinafter is mainly focused on POT. Nevertheless, the techniques of extreme data analysis for the annual maxima method are identical with

those of POT, and the following descriptions are also applicable for the use of the annual maxima method.

(B) *Parameters of sample of extreme data*

Two parameters are important in describing the nature of a sample of extreme data. One is the *mean rate* of the extreme events. The mean rate denoted by λ is defined with the number of events N_T during the period of K years as

$$\lambda = \frac{N_T}{K}. \tag{11.1}$$

In the annual maxima method, one data is taken from each year so that $\lambda = 1$. In POT, the mean rate may vary from a few to several dozens depending on the threshold value which defines the extreme events. The number of years K need not be an integer but can have a decimal.

Another parameter is related with the process of censoring. When a wave hindcasting project is undertaken for the purpose of collecting samples of extreme wave heights, there is a possibility that medium to minor storms have not been detected on weather maps and waves generated by these storms are dropped from the list of data. Thus, it is often recommended to employ the data of large storm waves only, by omitting the data of low wave heights. This is an example of censoring process. Another example is the treatment of downtime of wave measurement system. If the maximum wave during the downtime is known to be below a certain moderate value by information from some other sources, the period of downtime can be included in the effective duration of measurement period K, provided that other measured data below that value be omitted from the extreme wave analysis. In these examples, the existence of minor data should be taken into account in extreme analysis so as not to distort the shape of distribution function. For this purpose, the following parameter called the *censoring parameter* denoted by ν is introduced:

$$\nu = \frac{N}{N_T}, \tag{11.2}$$

where N refers to the number of data taken in the analysis and N_T the total number of storm events which would have occurred during the period of extreme wave analysis; N_T need not be accurate, but its approximate estimate suffices.

11.1.2 *Distribution Functions for Extreme Waves*

In the extreme data analysis, many theoretical distribution functions are employed for fitting to samples. In theoretical statistics, a data of extremes refers to the maximum or minimum among a sample of independent data. When extreme analysis is applied for a sample of such extreme data, it is known that three types of theoretical functions should fit such samples, depending on the population distribution of initial data.[3] However, the data of extreme wave heights collected by POT are different from the extreme data of theoretical statistics. POT data carry no meaning of maxima of samples, but they are initial data defined as the peak heights of storm waves. Therefore, there is no theoretical ground to recommend any distribution function *a priori* to samples collected by POT.

Current consensus among people practicing extreme wave analysis is such as to apply various distribution functions to a sample and to select a best-fitting one as the most probable distribution of the population. The candidate functions often employed in extreme wave analysis are listed in the following, but there are several other distribution functions favored by statisticians. The cumulative distribution is denoted by $F(x)$ and the probability density function by $f(x)$, where x stands for the extreme variate (i.e., wave height).

1) Fisher–Tippett type I (abbreviated as FT-I) or Gumbel distribution:

$$F(x) = \exp\left[-\exp\left(-\frac{x-B}{A}\right)\right] \quad : \quad -\infty < x < \infty\,. \tag{11.3}$$

2) Fisher–Tippett type II (abbreviated as FT-II) or Frechét distribution:

$$F(x) = \exp\left[-\left(1+\frac{x-B}{kA}\right)^{-k}\right] \quad : \quad B - kA \le x < \infty\,. \tag{11.4}$$

3) Weibull distribution:

$$F(x) = 1 - \exp\left[-\left(\frac{x-B}{A}\right)^{k}\right] \quad : \quad B \le x < \infty\,. \tag{11.5}$$

4) Lognormal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}Ax}\exp\left[-\frac{(\ln x - B)^2}{2A^2}\right] \quad : \quad 0 < x < \infty\,. \tag{11.6}$$

These functions have two or three parameters. The parameter A is called the *scale parameter* because it governs the linear scale of x. The parameter B is called the *location parameter* because it fixes the location of the axis of x. The parameter k is called the *shape parameter* because it determines the functional shape of distribution. The parameter k has no dimension, but the parameters A and B have the same units with x except for the lognormal distribution. The notations for these parameters are not universal; the readers are advised to check the notations when they refer to various literatures.

The expression for the FT-II distribution has been so chosen that it does converge to the FT-I function at the limit of $k \to \infty$. Figure 11.1 exhibits the probability density of the FT-II distribution with the shape parameter $k = 2.5$, 3.3, 5 and 10 together with that of the FT-I distribution, which is designated with $k = \infty$. The abscissa of Fig. 11.1 is a dimensionless variate of $y = (x - B)/A$, which is called the *reduced variate*. As the value of the shape

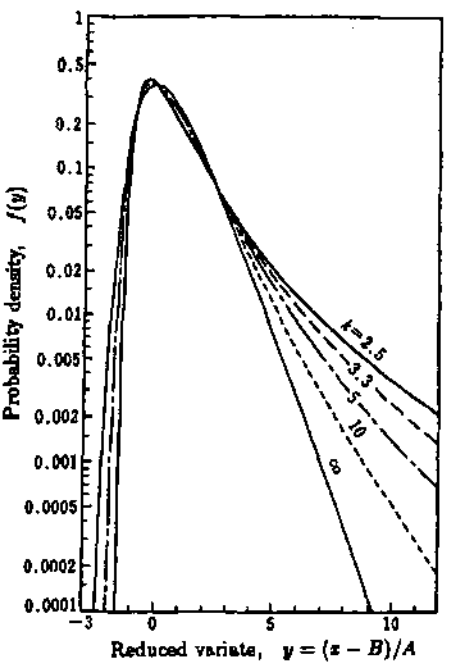

Fig. 11.1. Probability densities of FT-I ($k = \infty$) and FT-II ($k = 2.5$ to 10) distributions.[10]

parameter k decreases, the distribution of FT-II becomes broader with longer tails. A broad distribution predicts a very large 100-year wave height, when compared with a narrow distribution.

Figure 11.2 shows the probability density of the Weibull distribution with the shape parameter $k = 0.75$, 1.0, 1.4 and 2.0. The case with $k = 1$ is the exponential distribution, which is included in the Weibull distribution. With the decrease in the k value, the distribution becomes broader. The lognormal distribution demonstrates behaviors similar to the Weibull distribution with $k = 2$; the two distributions are often fitted to a sample of extreme wave heights with almost the same degree of goodness of fit.[4] Thus, a fitting of lognormal distribution can be replaced to that of the Weibull distribution with $k = 2$, and explanations on the applications of the lognormal distribution for extreme waves data are deleted hereinafter.

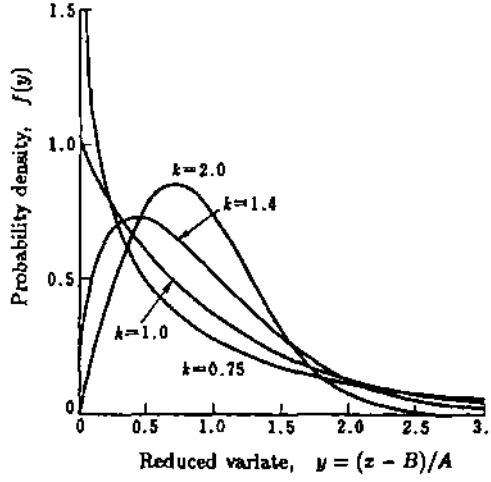

Fig. 11.2. Probability density of Weibull distribution ($k = 0.75$ to 2.0).[4]

The characteristics of the above four distributions are listed in Table 11.1 for their modes, means and standard deviations. When the shape parameter k of the FT-II distribution is less than 2, the distribution becomes too broad and its standard deviation cannot be defined. When the shape parameter k of the Weibull distribution is less than 1, the probability densities diverge as the variate x approaches B and the mode cannot be defined.

Table 11.1. Characteristics of distribution functions for extreme analysis.

Distribution	Mode	Mean	Standard Deviation
FT-II	$B+kA\left[\left(\frac{k}{1+k}\right)^{1/k}-1\right]$	$B+kA\left[\Gamma\left(1-\frac{1}{k}\right)-1\right]$	$kA\left[\Gamma\left(1-\frac{2}{k}\right)-\Gamma^2\left(1-\frac{1}{k}\right)\right]^{1/2}$
FT-I	B	$B+A\gamma$	$\frac{\pi}{\sqrt{6}}A$
Weibull	$B+A\left(1-\frac{1}{k}\right)^{1/k}$	$B+A\Gamma\left(1+\frac{1}{k}\right)$	$A\left[\Gamma\left(1+\frac{2}{k}\right)-\Gamma^2\left(1+\frac{1}{k}\right)\right]^{1/k}$
Lognormal	$\exp(B-A^2)$	$\exp\left(B+\frac{A^2}{2}\right)$	$\exp\left(B+\frac{A^2}{2}\right)(\exp A^2-1)^{1/2}$

Note: $\Gamma(\ )$ is the gamma function and γ is Euler's constant ($= 0.5772\ldots$).

11.1.3 *Return Period and Return Value*

Extreme statistics have the objective of estimating an expected value of extreme event which would occur once in a long period of time. For this purpose, the concepts of *return period* and *return value* are introduced. The return period is defined as the average duration of time during which extreme events exceeding a certain threshold value would occur once. The return value is the threshold value which defines a given return period. A 100-year wave height is a return value which would be exceeded once in every 100 years on the average.

The return period denoted by R is derived from the distribution function as follows. For the sake of simplicity, the case of annual maxima method is discussed first. The distribution function $F(x)$ is assumed to be known. The probability that the extreme variate x does not exceed a given value x_u in one year is $F(x_u)$ by definition. Suppose that the event of $x \geq x_u$ occurred in one year, the variate x did not exceed x_u during the other $n-1$ years, and it did exceed x_u in the nth year. Because the probability of nonexceedance for $n-1$ years is given by $F^{n-1}(x_u)$ and that of exceedance in one year is $1-F(x_u)$, the probability of the above event is expressed as

$$P_n = F^{n-1}(x_u)[1-F(x_u)]\,. \tag{11.7}$$

The above event may occur in the first year of $n=1$ but may not occur until $n=\infty$. The expected value of n is the return period by definition, and it is calculated as below.

$$R = E[n] = \sum_{n=1}^{\infty} nP_n = [1-F(x_u)]\sum_{n=1}^{\infty} nF^{n-1}(x_u) = \frac{1}{1-F(x_u)}\,. \tag{11.8}$$

The return value corresponding to the return period R is denoted by x_R. It is obtained with the inverse function of the cumulative distribution as

$$x_R = F^{-1}\left(1 - \frac{1}{R}\right) . \tag{11.9}$$

In the case of POT with the mean rate λ, each year is divided into time segments of $1/\lambda$ year by assuming that each time segment has the same probability of extreme events (seasonal variation of events is neglected). Then, the return period and the return value can be given by the following formulas:

$$R = \frac{1}{\lambda[1 - F(x_u)]} , \tag{11.10}$$

$$x_R = F^{-1}\left(1 - \frac{1}{\lambda R}\right) . \tag{11.11}$$

11.2 Estimation of Best-Fitting Distribution Function

11.2.1 *Selection of Fitting Method*

There are several methods of fitting a theoretical distribution function to a sample of extreme data and estimating the parameter values. They are

(i) graphical fitting method,
(ii) least squares method,
(iii) method of moments,
(iv) maximum likelihood method,
(v) others.

The graphical fitting method and the least squares method require a rearrangement of data in a given sample in the descending order by placing the largest data at the position of number one. The rearranged data are given the probability of nonexceedance according to their order number m and the sample size N (number of data in a sample). In the graphical fitting method, a special probability paper is devised in such a way that a particular distribution would be represented as a straight line on that paper. The extreme data are plotted on the probability paper of chosen distribution, and a straight line which best fits to the data is drawn by visual judgment. The least squares method eliminates the process of visual judgment and can make objective comparison of the goodness of fit. The assignment of respective probabilities

of nonexceedance to individual data is made with the so-called *plotting position formula* to be discussed later.

The method of moments calculates the mean and standard deviation of the sample and equates the results to the characteristics of distribution listed in Table 11.1. The parameters A and B of the FT-I and lognormal distributions can be estimated through this process. The FT-II and Weibull distributions require the information of the skewness of sample so as to estimate the shape parameter k and two other parameters. The method of moments and the graphical fitting method were the favorite ones in old days as described in the classical book by Gumbel[3]; in those days, computing capacity of analysts was quite limited and the least squares method could not be carried out for a large sample. The method of moments has a demerit of having a negative bias in the estimate of scale parameter A, unless a proper adjustment of sample size is made. It is because the standard deviation gradually decreases from the value listed in Table 11.1 as the sample size decreases; this will be discussed in Sec. 11.3.1. The method of moments also does not have a theoretical or empirical formula for estimating confidence interval of return value except for the FT-I distribution (see Sec. 11.3.3(C)); a numerical simulation by using the parameters obtained is required for estimation of confidence interval.

The maximum likelihood method is an iterative numerical scheme to find the parameter values which maximize the likelihood function defined as

$$L(x_1, \ldots, x_N; A, B, k) = \prod_{m=1}^{N} f(x_m; A, B, k), \tag{11.12}$$

where $x_1, \ldots, x_N$ represent the data values and f is the probability density function. The maximum likelihood method is favored by statisticians in recent years, because its characteristics can be examined mathematically. However the theory is not easy to understand and the algorithm of numerical scheme is rather complicated.

A sample of extreme variates always has a statistical variability as will be discussed in Sec. 11.3.1. This sample variability causes a certain amount of deviation of the return value estimated on a sample from the population value. In consideration of such deviations, two criteria are employed in selection of the fitting method in the statistics of extremes. One is *unbiasedness* and the other is *efficiency*. The former refers to the condition that the return value should have no bias from the corresponding value of the population. Bias is evaluated

not for individual samples but for the ensemble of samples. The amount of bias is calculated by means of the Monte Carlo simulation technique when no theoretical evaluation is possible. Efficiency refers to the degree of deviation of estimated return values from the population. The smaller the deviation is, the larger the efficiency of a fitting method is.

According to the above two criteria, the method of moments cannot satisfy the unbiasedness. Its efficiency is said to be inferior to the maximum likelihood method and the least squares method. The maximum likelihood method tends to have a small amount of negative bias, but its efficiency seems to be the largest as exemplified in a comparative study with numerically simulated extreme data.[5] In this chapter, however, the least squares method for extreme analysis is discussed in detail because of its simplicity in algorithm and applications. The least squares method has often been accused of yielding a positive bias in the return value. However, it was caused by the use of inappropriate plotting position formula. With the use of plotting position formulas presented in the following, it has been verified that the least squares method satisfies the condition of unbiasedness.[4,6]

11.2.2 *Plotting Position Formulas*

A sample of data arranged in the ascending or descending order belongs to the category of *ordered statistics*. As the present chapter is concerned with the statistics of extremely large wave heights, the descending order is taken and the order number is expressed with m. The variate and its nonexceedance probability of the mth order are denoted with the subscript (m). The formula which assigns the probability to the ordered variate is the plotting position formula.

The best-known plotting position formula is the Weibull formula of the following:

$$\hat{F}_{(m)} = 1 - \frac{m}{N+1}. \tag{11.13}$$

Equation (11.13) is derived as the expected probability of the mth ordered variate in the population; i.e., $E[F(x_{(m)})]$. Gumbel (Ref. 11.3, Sec 1.2.7) advocated the use of this formula based on somewhat intuitive arguments. But the Weibull plotting position formula always produces a positive bias in the return value, amounting to several percent when the sample size is less than a few dozens.[4,6]

The unbiased plotting position formula varies depending on the distribution function applied. According to a numerical simulation study, the Gringorten formula[7] yields almost no bias when applied to the FT-I distribution. For the normal distribution, the Blom formula[8] brings forth little bias. For the Weibull distribution, Petruaskas and Aagaard[9] derived a formula in such a way that it gives the probability corresponding to the expected value of the mth ordered variate; i.e., $F\{E[x_{(m)}]\}$. When examined in a numerical simulation study, however, their formula has produced a small amount of negative bias. Then Goda[4,6] proposed a modified version of the Petruaskas and Aagaard formula. For the FT-II distribution, Goda and Onozawa[10] proposed an empirical formula based on another numerical simulation study. The latter two proposals are both based on the Monte Carlo simulations for the sample size ranging from 10 to 200, each size with 10,000 samples.

The unbiased plotting position formula can be expressed in the following general form:

$$\hat{F}_{(m)} = 1 - \frac{m - \alpha}{N_T + \beta}, \quad m = 1, 2, \ldots, N. \tag{11.14}$$

The values of constants α and β are given in Table 11.2. The above formula uses the total number N_T instead of the sample size N so that the formula can be applied for both censored and uncensored samples.

Table 11.2. Constants of unbiased plotting position formula.

Distribution	α	β	Authors
FT-II	$0.44 + 0.52/k$	$0.12 - 0.11/k$	Goda and Onozawa[10]
FT-I	0.44	0.12	Gringorten[7]
Weibull	$0.20 + 0.27/\sqrt{k}$	$0.20 + 0.23/\sqrt{k}$	Goda[4,6]
Normal	0.375	0.25	Blom[8]
Lognormal	0.375	0.25	Blom[8]

11.2.3 *Parameter Estimation by the Least Squares Method*

The first step in the parameter estimation is the selection of candidate distribution functions. As discussed in Sec. 11.1.2, the FT-I, FT-II and Weibull distributions are considered in this chapter as the candidates of the distribution of a population of extreme waves; the population distribution is called

the *parent distribution*. The scale, location, and shape parameters of these candidate functions are estimated for a given sample, and the goodness of fit to each function is compared for selection of the best-fitting distribution.

The least squares method can yield the best estimate of two parameters in a single operation. As the FT-I distribution has two parameters of A and B, it can be analyzed by the least squares method directly. The FT-II and Weibull distributions have three parameters however, and thus they have to be modified into a form of two parameter functions. In this chapter, the shape parameter k is fixed at one of the following values for these distributions:

$$\left.\begin{array}{l}\text{FT-II distribution} \quad : k = 2.5,\ 3.33,\ 5.0 \text{ and } 10.0,\\ \text{Weibull distribution} : k = 0.75,\ 1.0,\ 1.4 \text{ and } 2.0.\end{array}\right\} \tag{11.15}$$

Once the shape parameter is fixed, each distribution becomes an independent candidate function and is competed with other functions for best fitting. Thus, a proposal is hereby made to employ nine cumulative distributions (one FT-I, four FT-II's, and four Weibull's) as the candidate distributions.

The main reason for fixing the shape parameter is the difficulty in predicting the true parent distribution from a sample of small size, say a few dozen to one hundred. Goda[4,6] has demonstrated this difficulty by a Monte Carlo simulation study. The statistical variability of these distributions and confidence intervals of parameter estimates and return values have also been analyzed for the above nine distributions, and the results are presented in a form of tables and empirical formulas.

The second step in the parameter estimation is the preparation of the order statistics $x_{(m)}$ of extreme data in the descending order and the assignment of the nonexceedance probability $\hat{F}_{(m)}$ by Eq. (11.14). Then the reduced variate $y_{(m)}$ for the mth ordered data is calculated by the following equation:

$$\left.\begin{array}{ll}\text{FT-I distribution} & : y_{(m)} = -\ln[-\ln \hat{F}_{(m)}]\,,\\ \text{FT-II distribution} & : y_{(m)} = k[(-\ln \hat{F}_{(m)})^{-1/k} - 1]\,,\\ \text{Weibull distribution} & : y_{(m)} = [-\ln(1 - \hat{F}_{(m)})]^{1/k}\,.\end{array}\right\} \tag{11.16}$$

The third step is the application of the least squares method for the parameters $\hat{A}$ and $\hat{B}$ in the following equation:

$$x_{(m)} = \hat{B} + \hat{A} y_{(m)}\,. \tag{11.17}$$

The correlation coefficient r between $x_{(m)}$ and $y_{(m)}$ must be estimated together with $\hat{A}$ and $\hat{B}$. Any numerical algorithm for the least squares method suffices for solving Eq. (11.17). However, attention is called for the expression of Eq. (11.17) which is different from the conventional form of $y = a + bx$.

As an explanatory example, the Kodiak data of hindcasted storm waves[2] is analyzed below. Wave hindcasting was carried out by the Coastal Engineering Research Center of the US Army[11] for the North-Eastern Pacific Ocean. The data was retrieved from a grid point off Alaska, located at 57°50′N and 148°78′W. The data set consists of all peak storm waves with the significant height exceeding 6 m, which were generated by 78 storms during a period of 20 years. Table 11.3 lists the Kodiak data set in chronological sequences. As the exact number of storm events in this period was not scrutinized, the data set is treated bere as an uncensored sample. The Kodiak data is one of the two extreme wave data sets which were jointly analyzed by a working group on extreme statistics of the Section of Maritime Hydraulics of the International Association of Hydraulic Research (IAHR), as reported by van Vledder *et al.*[2]

Table 11.3. Peak significant wave heights of Kodiak data set ($K = 20$ years, $N = N_T = 78$, $\lambda = 3.9$, $\nu = 1$).

Year	H_s (m)	Year	H_s (m)
1956	6.2	1966	7.3, 8.6, 7.4
1957	–	1967	7.1, 6.0, 6.3, 6.0, 6.7
1958	8.8, 6.6, 6.9, 7.8, 6.3	1968	6.6, 6.5, 6.9, 7.7, 8.2, 6.7, 7.4
1959	11.7, 7.2, 7.4	1969	6.4, 6.1, 7.1, 6.5, 8.5, 8.8, 9.1
1960	9.9, 8.9, 7.5, 7.0, 6.7	1970	8.0, 6.3, 9.1
1961	9.2, 6.2, 6.3	1971	6.6
1962	8.1, 6.3, 7.2, 6.3, 6.0	1972	6.7, 7.2, 10.2, 7.0, 10.1
1963	8.4, 6.8, 9.3, 6.7, 6.5, 7.2, 8.5	1973	7.8, 6.1, 6.3, 8.6, 7.1, 10.0
1964	6.9. 6.6, 9.4, 8.2	1974	8.0, 6.1, 8.4
1965	6.3, 7.6	1975	7.4, 8.2, 8.1

The Kodiak data set is rearranged in the descending order according to the magnitude of significant wave height, the nonexceeding probability is assigned, and the reduced variate is calculated for several candidate distributions. Then the least squares method is applied. Table 11.4 lists a part of the results of calculation. Graphical representation of the Kodiak data will be given in Sec. 11.3.3.

Table 11.4. Results of Kodiak data analysis by the least squares method (sample size: $N = 78$, mean: $\bar{x} = 7.501$ m, standard deviation: $\sigma_x = 1.214$ m).

m	$x_{(m)}$	FT-II ($k = 10$)		FT-I		Weibull ($k = 1.4$)		Weibull ($k = 2.0$)	
		$\hat{F}_{(m)}$	$y_{(m)}$	$\hat{F}_{(m)}$	$y_{(m)}$	$\hat{F}_{(m)}$	$y_{(m)}$	$\hat{F}_{(m)}$	$y_{(m)}$
1	11.7	0.9935	6.540	0.9928	4.934	0.9927	3.121	0.9922	2.204
2	10.2	0.9807	4.825	0.9800	3.903	0.9800	2.648	0.9795	1.971
3	10.1	0.9679	4.081	0.9672	3.402	0.9672	2.405	0.9667	1.845
4	10.0	0.9551	3.607	0.9544	3.065	0.9544	2.238	0.9539	1.754
5	9.9	0.9423	3.261	0.9416	2.811	0.9417	2.109	0.9412	1.683
6	9.4	0.9295	2.990	0.9288	2.606	0.9289	2.003	0.9284	1.624
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
74	6.1	0.0589	−0.989	0.0584	−1.044	0.0615	0.140	0.007	0.250
75	6.1	0.0461	−1.063	0.0456	−1.128	0.0488	0.118	0.0479	0.222
76	6.0	0.0333	−1.152	0.0328	−1.229	0.0360	0.094	0.0351	0.189
77	6.0	0.0205	−1.270	0.0200	−1.364	0.0233	0.069	0.0224	0.150
78	6.0	0.0077	−1.464	0.0072	−1.597	0.0105	0.039	0.0096	0.098
Parameters		$\hat{A} = 0.8292$ m $\hat{B} = 6.937$ m		$\hat{A} = 0.9567$ m $\hat{B} = 6.955$ m		$\hat{A} = 1.8621$ m $\hat{B} = 5.805$ m		$\hat{A} = 2.6228$ m $\hat{B} = 5.178$ m	
Correlation		$r = 0.98738$		$r = 0.99191$		$r = 0.99629$		$r = 0.98906$	

11.2.4 *Selection of Most Probable Parent Distribution*

(A) *Goodness of fit tests*

As mentioned earlier, the candidate distribution which best fits to the sample is selected as the most probable parent distribution. Nevertheless this does not exclude use of a single candidate distribution based on one's postulation about the parent distribution of storm wave heights. As the data base of extreme waves by observations and hindcasting is expanded, there will be many cases of reliable extreme wave analyses. Then it could be possible in the future to establish some parent distribution of extreme wave heights, which would vary from coast to coast.

Goodness of fit is measured by several tests. The Kolmogorov–Smirnov test, the Anderson–Darling test and the chi-square test are often used for this

purpose. When the parameter estimate is done by the least squares method, however, the degree of goodness of fit is simply represented with the value of correlation coefficient between the ordered data $x_{(m)}$ and its reduced variate $y_{(m)}$; the nearer the coefficient is toward 1, the better the fitting is. The Kodiak data set has been fitted to nine distributions including the four distributions listed in Table 11.4. Among the candidate functions, the Weibull distribution with $k = 1.4$ yields the correlation coefficient closest to 1, and is judged as the best-fitting one.

The degree of correlation coefficient being near to 1 depends on a candidate distribution. Samples from a distribution with a narrow range of spreading such as the Weibull with $k = 2$ tend to yield the correlation coefficient much closer to 1 compared with samples from a distribution with a broad spreading. To examine the statistical characteristics of correlation coefficient, its residue from 1 is defined here as $\Delta r = 1 - r$. The residue is a statistical variate, the value of which varies from sample to sample. Goda and Kobune[6] reported the results of another Monte Carlo simulation study on extreme statistics. Figure 11.3 shows the mean value of the residue of correlation coefficient of samples for several distributions. Simulation was done with 10,000 samples for each sample size ranging from 10 to 400 for respective distributions. As seen in Fig. 11.3, Δr_{mean} of the Weibull distribution with $k = 0.75$ is larger than

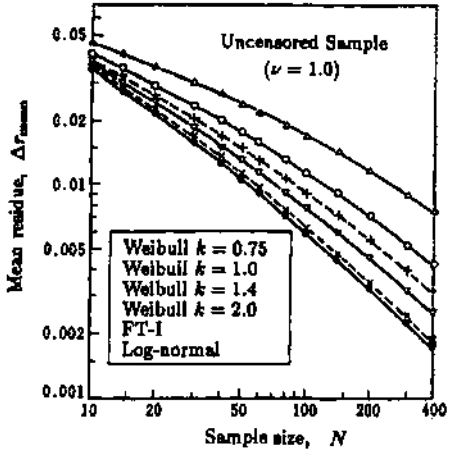

Fig. 11.3. Mean residue Δr_{mean} of various distributions.[6]

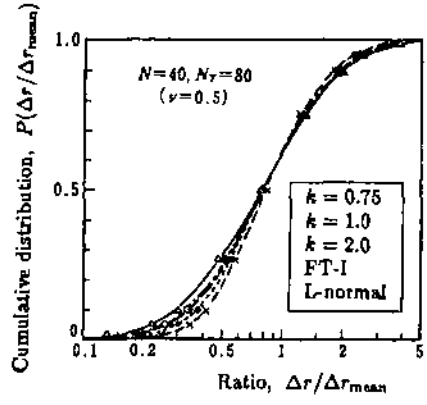

Fig. 11.4. Examples of cumulative distributions of the ratio $\Delta r/\Delta r_{mean}$.[6]

that with $k = 2$ at any sample size. Thus, a test of goodness of fit by means of the absolute value of correlation coefficient tends to yield unfavorable results against a broad distribution.

A remedy for the above bias is to use the ratio of the residue of a sample to the mean residue of a fitted distribution. Figure 11.4 shows examples of the cumulative distributions of the ratio $\Delta r/\Delta r_{\text{mean}}$; the sample size is $N = 40$ and the data are censored ones with the censoring parameter $\nu = 0.5$. When the residue of correlation coefficient is normalized with its mean value, differences between various distributions are greatly reduced and a fair comparison of goodness of fit becomes possible. Goda and Kobune[6] proposed to use the MIR (MInimum Ratio of residual correlation coefficient) criterion for judgment of best fitting; a distribution with the smallest ratio is a best fitting one. They derived an empirical formula for estimating the mean residue Δr_{mean} for a given distribution, sample size and censoring parameter from the data of simulation study. The formula is given as

$$\Delta r_{\text{mean}} = \exp[a + b \ln N + c (\ln N)^2] \,. \tag{11.18}$$

The coefficients a, b and c are formulated for various distributions as listed in Table 11.5. Relative error of Eq. (11.18) in predicting the mean residue value is less than $\pm 3\%$.