CHAPTER 3

Statistical Methods and Error Handling

3.1 INTRODUCTION

This chapter provides a review of some of the basic statistical concepts and terminology used in processing data. We need this information if we are to deal properly with the specific techniques used to edit and analyze oceanographic data. Our review is intended to establish a common level of understanding by the readers, not to provide a summary of all available procedures.

In the past, all collected data were processed and reduced by hand so that the individual scientist had an opportunity to become personally familiar with each data value. During this manual reduction of data, the investigator took into account important information regarding the particular instrument used and was able to determine which data were "better" in the sense that they had been collected and processed correctly. Within the limits of the observing systems, an accurate description of the data could be achieved without the application of statistical procedures. Individual intuition and familiarity with shipboard procedures took precedence in this type of data processing and analyses were made on comparatively few data. In such investigations, the question of statistical reliability was seldom raised and it was assumed that individual data points were correct.

For the most part, the advent of the computer and electronic data collection methods has meant that a knowledge of statistical methods has become essential to any reliable interpretation of results. Circumstances still exist, however, for which physical oceanographers still assign considerable weight to the quality of individual measurements. This is certainly true of water sample data such as dissolved oxygen, nutrients, and chemical tracers collected from bottle casts. In these cases, the established methods of data reduction, including familiarity with the data and knowledge of previous work in a particular region, still produce valuable descriptions of oceanic features and phenomena with a spatial resolution not possible with statistical techniques. However, for those more accustomed to having data collected and/or delivered on high density storage media such as magnetic tape, CD-ROM, or floppy disk, statistical methods are essential to determining the value of the data and to decide how much of it can be considered useful for the intended analysis. This statistical approach arises from the fundamental complexity of the ocean, a

multivariate system with many degrees of freedom in which nonlinear dynamics and sampling limitations make it difficult to separate scales of variability.

A fundamental problem with a statistical approach to data reduction is the fact that the ocean is not a stationary environment in which we can make repeated measurements. By "stationary" we mean a physical system whose statistical properties remain unchanged with time. In order to make sense of our observations, we are forced to make some rather strong assumptions about our data and the processes we are trying to investigate. Basic to these assumptions is the concept of randomness and the consequent laws of probability. Since each oceanographic measurement can be considered a superposition of the desired signal plus unwanted noise (due to measurement errors and unresolved geophysical variability), the assumption of random behavior often is applied to both the signal and the noise. We must consider not only the statistical character of the signal and noise contributions individually but also the fact that the signal and the noise can interact with each other. Only through the application of the concept of probability can we make the assumptions required to reduce this complex set of variables to a workable subset. Our brief summary of statistics will emphasize concepts pertinent to the analysis of random variables such as probability density functions and statistical moments (mean, variance, etc.). A brief glossary of statistical terms can be found in Appendix B.

3.2 SAMPLE DISTRIBUTIONS

Fundamental to any form of data analysis is the realization that we are usually working with a limited set (or sample) of random events drawn from a much larger population. We use our sample to make estimates of the true statistical properties of the population. Historically, studies in physical oceanography were dependent on too few data points to allow for statistical inference and individual samples were considered representative of the true ocean. Often, an estimate of the population distribution is made from the sample set by using the relative frequency distribution, or histogram, of the measured data points. There is no fixed rule on how such a histogram is constructed in terms of ideal bin interval or number of bins. Generally, the more data there are, the greater the number of bins used in the histogram. Bins should be selected so that the majority of the measurements do not fall on the bin boundaries. Since the area of a histogram bin is proportional to the fraction of the total number of measurements in that interval, it represents the probability that an individual sample value will lie within that interval (Figure 3.1).

The most basic descriptive parameter for any set of measurements is the sample mean. The mean is generally taken over the duration of a time series (time average) or over an ensemble of measurements (ensemble mean) collected under similar conditions (Table 3.1). If the sample has N data values, $x_1, x_2, \ldots, x_N$, the sample mean is calculated as

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{3.2.1}$$

The sample mean is an unbiased estimate of the true population mean, μ. Here, an

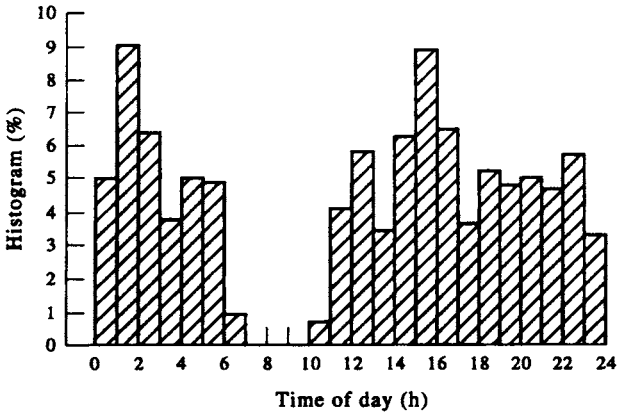

Figure 3.1. Histogram giving the percentage occurrences for the times of satellite position fixes during a 24-h day. Data are for satellite-tracked surface drifter #4851 deployed in the northeast Pacific Ocean from 10 December 1992 to 28 February 1993. During this 90-day period, the satellite receiver on the drifter was in the continuous receive mode.

Table 3.1. Statistical values for the data set $x = \{x_i, i = 1, \ldots, 9\} = \{-3, -1, 0, 2, 5, 7, 11, 12, 12\}$

Mean $\bar{x}$	Variance s'^2	Variance s^2	Standard deviation, s	Range	Median	Mode
5.00	30.22	34.00	5.83	15	5	12

"unbiased" estimator is one for which the expected value, $E[x]$, of the estimator is equal to the parameter being estimated. In this case, $E[x] = \mu$ for which $\bar{x}$ is an unbiased estimator. The sample mean locates the center of mass of the data distribution such that

$$\sum_{i=1}^{N} (x_i - \bar{x}) = 0$$

that is, the sample mean splits the data so that there is an equal weighting of negative and positive values of the fluctuation, $x' = x_i - \bar{x}$, about the mean value, $\bar{x}$. The weighted sample mean is the general case of (3.2.1) and is defined as

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} f_i x_i \tag{3.2.2}$$

where f_i/N is the relative frequency of occurrence of the ith value for the particular experiment or observational data set. In (3.2.1), $f_i = 1$ for all i.

The sample mean values give us the center of mass of a data distribution but not its width. To determine how the data are spread about the mean, we need a measure of the sample variability or *variance*. For the data used in (3.2.1), the *sample variance* is the average of the square of the sample deviations from the sample mean, expressed as

$$s'^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{3.2.3}$$

The *sample standard deviation* $s' = \sqrt{s'^2}$ the positive square root of (3.2.3), is a measure of the typical difference of a data value from the mean value of all the data points. In general, these differ from the corresponding true *population variance*, σ^2, and the *population standard deviation*, σ. As defined by (3.2.3), the sample variance is a biased estimate of the true population variance. An unbiased estimator of the population variance is obtained from

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{3.2.4a}$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{N}(x_i)^2 - \frac{1}{N}\left(\sum_{i=1}^{N}x_i\right)^2\right] \tag{3.2.4b}$$

where the denominator $N - 1$ expresses the fact that we need at least two values to define a sample variance and standard deviation, s. The use of the estimators s versus s' is often a matter of debate among oceanographers, although it should be noted that the difference between the two values decreases as the sample size increases. Only for relatively small samples ($N < 30$) is the difference significant. Because s' has a smaller mean square error than s and is an unbiased estimator when the population mean is known *a priori*, we recommend the use of (3.2.4). However, a word of caution: if your hypothesis depends on the difference between s and s', then you have ventured onto shaky statistical ground supported by questionable data. We further note that the expanded relation (3.2.4b) is a more efficient computational formulation than (3.2.4a) in that it allows one to obtain s^2 from a single pass through the data. If the sample mean must be calculated first, two passes through the same data set are required rather than one, which is computationally less efficient when dealing with large data sets.

Other statistical values of importance are the range, mode, and median of a data distribution (Table 3.1). The *range* is the spread or absolute difference between the end-point values of the data set while the *mode* is the value of the distribution that occurs most often. For example, the data sequence 2, 4, 4, 6, 4, 7 has a range of $|2 - 7| = 5$ and a mode of 4. The *median* is the middle value in a set of numbers arranged according to magnitude (the data sequence $-1, 0, 2, 3, 5, 6, 7$ has a median of 3). If there is an even number of data points, the median value is chosen mid-way between the two candidates for the central value. Two other measures, *skewness* (the third moment of the distribution and degree of asymmetry of the data about the mean) and *kurtosis* (a nondimensional number measuring the flatness or peakedness of a distribution) are less used in oceanography.

As we discuss more thoroughly later in this chapter, the shapes of many sample distributions can be approximated by a *normal* (also called a *bell* or *Gaussian*) distribution. A convenient aspect of a normal population distribution is that we can apply the following empirical "rule of thumb" to the data:

$\mu \pm \sigma$ spans approximately 68% of the measurements;
$\mu \pm 2\sigma$ spans approximately 95% of the measurements;
$\mu \pm 3\sigma$ spans most (99%) of the measurements.

The percentages are represented by the areas under the normal distribution curve spanned by each of the limits (Figure 3.2). We emphasize that the above limits apply only to normal distributions of random variables.

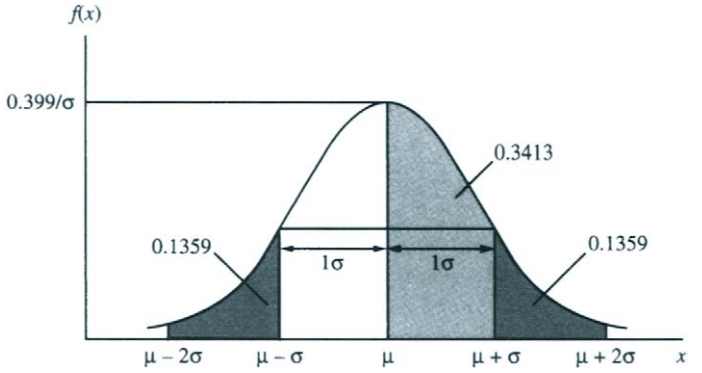

Figure 3.2. Normal distribution f(x) for mean μ and standard deviation σ of the random variable X. (From Harnett and Murphy, 1975.)

3.3 PROBABILITY

Most data collected by oceanographers are made up of samples taken from a larger unknown population. If we view these samples as random events of a statistical process, then we are faced with an element of uncertainty: "What are the chances that a certain event occurred or will occur based on our sample?" or "How likely is it that a given sample is truly representative of a certain population distribution? " (The last question might be asked of political pollsters who use small sample sizes to make sweeping statements about the opinions of the populace as a whole.) We need to find the best procedures for inferring the population distribution from the sample distribution and to have measures that specify the goodness of the inference. Probability theory provides the foundation for this type of analysis. In effect, it enables us to find a value between 0 and 1 which tells us just how likely is a particular event or sequence of events. A probability is a proportional measure of the occurrence of an event. If the event has a probability of zero, then it is impossible; if it has a probability of unity, then it is certain to occur. Probability theory as we know it today was initiated by Pascal and Fermat in the seventeenth century through their interest in games of chance. In the eighteenth century, Gauss and Laplace extended the theory to social sciences and actuarial mathematics. Well-known names like R. A. Fisher, J. Neyman, and E. S. Pearson are associated with the proliferation of statistical techniques developed in the twentieth century.

The *probability mass function*, $P(x)$, gives the relative frequency of occurrence of each possible value of a discrete random variable, X. Put another way, the function specifies the point probabilities $P(x_i) = P(X = x_i)$ and assumes nonzero values only at points $X = x_i$, $i = 1, 2, \ldots$. One of the most common examples of a probability mass function is the sum of the dots obtained from the roll of a pair of dice (Table 3.2). According to probability theory, the dice player is most likely to roll a 7 (highest probability mass function) and least likely to roll a 2 or 12 (lowest probability mass

Table 3.2. The discrete probability mass function and cumulative probability functions for the sum of the dots (variable X) obtained by tossing a pair of dice

Sum of dots (X)	Frequency of occurrence	Relative frequency	Probability mass function, $P(x)$	Cumulative probability function $F(x) = P(X \leq x)$
2	1	1/36	$P(x = 2) = 1/36$	$F(2) = P(X \leq 2) = 1/36$
3	2	2/36	$P(x = 3) = 2/36$	$F(3) = P(X \leq 3) = 3/36$
4	3	3/36	$P(x = 4) = 3/36$	$F(4) = P(X \leq 4) = 6/36$
5	4	4/36	$P(x = 5) = 4/36$	$F(5) = P(X \leq 5) = 10/36$
6	5	5/36	$P(x = 6) = 5/36$	$F(6) = P(X \leq 6) = 15/36$
7	6	6/36	$P(x = 7) = 6/36$	$F(7) = P(X \leq 7) = 21/36$
8	5	5/36	$P(x = 8) = 5/36$	$F(8) = P(X \leq 8) = 26/36$
9	4	4/36	$P(x = 9) = 4/36$	$F(9) = P(X \leq 9) = 30/36$
10	3	3/36	$P(x = 10) = 3/36$	$F(10) = P(X \leq 10) = 33/36$
11	2	2/36	$P(x = 11) = 2/36$	$F11) = P(X \leq 11) = 35/36$
12	1	1/36	$P(x = 12) = 1/36$	$F(12) = P(X \leq 12) = 1$
SUM	36	1.00		1.00

function). The dice example reveals two of the fundamental properties of all discrete probability functions: (1) $0 \leq P(X = x)$; and (2) $\sum P(x) = 1$, where the summation is over all possible values of x. The counterpart to $P(x)$ for the case of a continuous random variable X is the *probability density function* (abbreviated, PDF), $f(x)$, which we discuss more fully later in the chapter. For the continuous case, the above fundamental properties become: (1) $0 \leq f(x)$; and (2) $\int f(x)\,dx = 1$ where the integration is over all x in the range $(-\infty, \infty)$.

To further illustrate the concept of probability, consider N independent trials, each of which has the same probability of "success" p and probability of "failure" $q = 1 - p$. The probability of success or failure is unity; $p + q = 1$. Such trials involve binomial distributions for which the outcomes can be only one of two events: for example, a tossed coin will produce a head or a tail; an XBT will work or it won't work. If X represents the number of successes that occur in the N trials, then X is said to be a discrete random variable having parameters (N, p). The term "Bernoulli trial" is sometimes used for X. The probability mass function which gives the relative frequency of occurrence of each value of the random variable X having parameters (N, p) is the binomial distribution

$$p(x) = \binom{N}{x} p^x (1-p)^{N-x}, \quad x = 0, 1, \ldots, N \tag{3.3.1a}$$

where the expression

$$\binom{N}{x} = \binom{N}{N-x} = {}_N C_x \equiv N!/[(N-x)!x!] \tag{3.3.1b}$$

is the number of different *combinations* of groups of x objects that can be chosen from a total set of N objects without regard to order. The number of different combinations of x objects is always fewer than the number of *permutations*, ${}_N P_x$, of x objects $[{}_N P_x \equiv N!/(N-x)!]$. In the case of permutations, different ordering of the same objects counts for a different permutation (i.e. *ab* is different than *ba*). As an example,

the number of possible different batting orders (permutations) a coach can create among the first four hitters on a nine-person baseball team is $9!/(9-4)! = 9!/5! = 3024$. In contrast, the number of different groups of ball-players a coach can put in the first four lead-off batting positions without regard to batting order is $9!/[(9-4)!4!] = 9!/5!4! = 126$. The numbers

$$\binom{N}{x}$$

often are called *binomial coefficients* since they appear as coefficients in the expansion of the binomial expression $(a+b)^N$ given by the binomial theorem:

$$(a+b)^N = \sum_{k=0}^{N} \binom{N}{k} a^k b^{N-k} \tag{3.3.2}$$

The summed probability mass function

$$P(a \le x \le b) = \sum_{a}^{b} P(x)$$

for variable X over a specified range of values (a, b) can be demonstrated by a simple oceanographic example. Suppose there is a probability $1-p$ that a current meter will fail when moored in the ocean and that the failure is independent from current meter to current meter. Assume that a particular string of meters will successfully measure the expected flow structure if at least 50% of the meters on the string remain operative. For example, a two-instrument string used to measure the barotropic flow will be successful if one current meter remains operative while a four-instrument string used to resolve the baroclinic flow will be successful if at least two meters remain operative. We then ask: "For what values of p is a four-meter array preferable to a two-meter array?" Since each current meter is assumed to fail or function independently of the other meters, it follows that the number of functioning current meters is a binomial random variable. The probability that a four-meter mooring is successful is then

$$\begin{aligned} P(2 \le x \le 4) &= \sum_{k=2}^{4} \binom{4}{k} p^k (1-p)^{4-k} \\ &= \binom{4}{2} p^2 (1-p)^2 + \binom{4}{3} p^3 (1-p)^1 + \binom{4}{4} p^4 (1-p)^0 \\ &= 6p^2(1-p)^2 + 4p^3(1-p)^1 + p^4 \end{aligned}$$

Similarly, the probability that a two-meter array is successful is

$$\begin{aligned} P(1 \le x \le 2) &= \sum_{k=1}^{2} \binom{2}{k} p^k (1-p)^{2-k} \\ &= 2p(1-p) + p^2 \end{aligned}$$

From these two relations, we find that the four-meter string is more likely to succeed when

$$6p^2(1-p)^2 + 4p^3(1-p)^1 + p^4 \geq 2p(1-p) + p^2$$

or, after some factoring and simplification, when

$$(p-1)^2 + (3p-2) \geq 0$$

for which we find $3p - 2 \geq 0$, or $p \geq 2/3$. When compared to the two-meter array, the four-meter array is more likely to do its intended job when the probability, p, that the instrument works is $p \geq 2/3$. The two-meter array is more likely to succeed when $p \leq 2/3$.

As the previous example illustrates, we often make the fundamental assumption that each sample in our set of observations is an independent realization drawn from a random distribution. Individual events in this distribution cannot be predicted with certainty but their relative frequency of occurrence, for a long series of repeated trials (samples), is often remarkably stable. We further remark that the binomial distribution is only one type of probability density function. Other distribution functions will be discussed later in the chapter.

3.3.1 Cumulative probability functions

The probability mass function yields the probability of a specific event or probability of a range of events. From this function we can derive the *cumulative probability function*, $F(x)$—also called the cumulative distribution function, cumulative mass function, and probability distribution function—defined as that fraction of the total number of possible outcomes X (a random variable) which are less than a specific value x (a number). Thus, the distribution function is the probability that $X \leq x$, or

$$F(x) = P(X \leq x)$$

$$= \sum_{\text{all } X \leq x} P(x), \quad -\infty < x < \infty \text{ (discrete random variable, } X) \tag{3.3.3a}$$

$$= \int_{-\infty}^{x} f(x)\,dx \text{ (continuous random variable, } X) \tag{3.3.3b}$$

The discrete cumulative distribution function for tossing a pair of fair dice (Table 3.2) is plotted in Figure 3.3. Since the probabilities P and f are limited to the range 0 and 1, we have $F(-\infty) = 0$ and $F(\infty) = 1$. In addition, the distribution function $F(x)$ is a nondecreasing function of x, such that $F(x_1) \leq F(x_2)$ for $x_1 < x_2$, where $F(x)$ is continuous from the right (Table 3.2).

It follows that, for the case of a continuous function, the derivative of the distribution function F with respect to the sample parameter, x

$$f(x) = \frac{dF(x)}{dx} \tag{3.3.4}$$

recovers the probability density function (PDF), f. As noted earlier, the PDF has the property that its integral over all values is unity

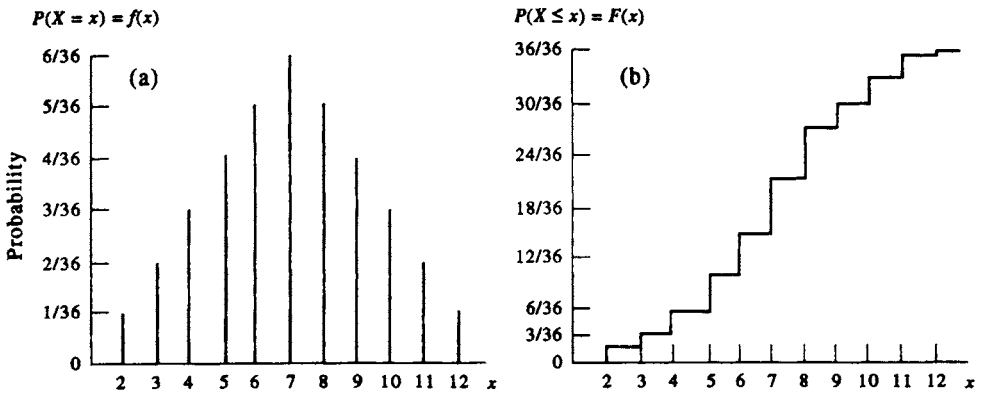

Figure 3.3. The discrete mass function P(x) and cumulative distribution function F(x) from tossing a pair of dice (see Table 3.2). (From Harnett and Murphy, 1975.)

$$\int_{-\infty}^{\infty} f(x)\,dx = F(\infty) - F(-\infty) = 1$$

In the limit $dx \to 0$, the fraction of outcomes for which x lies in the interval $x < x' < x + dx$ is equal to $f(x')\,dx$, the probability for this interval. The random variables being considered here are continuous so that the PDF can be defined by (3.3.4). Variables with distribution functions that contain discontinuities, such as the steps in Figure 3.3, are considered discrete variables. A random variable is considered discrete if it assumes only a countable number of values. In most oceanographic sampling, measurements can take on an infinity of values along a given scale and the measurements are best considered as continuous random variables. The function $F(x)$ for a continuous random variable X is itself continuous and appears as a smooth curve. Similarly, the PDF for a continuous random variable X is continuous and can be used to evaluate the probability that X falls within some interval $[a, b]$ as

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx \tag{3.3.5}$$

3.4 MOMENTS AND EXPECTED VALUES

The discussion in the previous section allows us to determine the probability of a single event or experiment, or describe the probability of a set of outcomes for a specific random variable. However, our discussion is not concise enough to describe fully the probability distributions of our data sets. The situation is similar to section 3.2 in which we started with a set of observed values. In addition to presenting the individual values, we seek properties of the data such as the sample mean and variance to help us characterize the structure of our observations. In the case of probability

distributions, we speak not of *observed* mean and variance but of the *expected* mean and variance obtained from an infinite number of realizations of the random variable under consideration.

Before discussing some common PDFs, we need to review the computation of the parameters used to describe these functions. These parameters are, in general, called "moments" by analogy to mechanical systems where moments describe the distribution of forces relative to some reference point. The statistical concept of degrees-of-freedom is also inherited from the terminology of physical–mechanical systems where the number of degrees-of-freedom specifies the motion possible within the physical constraints of the mechanical system and its distribution of forces. As noted earlier, the population mean, μ, and standard deviation, σ, define the first and second moments which describe the center and spread of the probability function. In general, these parameters do not uniquely define the PDF since many different PDFs can have the same mean and standard deviation. However, in the case of the Gaussian distribution, the PDF is completely described by μ and σ. In defining moments we must be careful to distinguish between moments taken about the origin and moments taken about the mean (central moments).

When discussing moments it is useful to introduce the concept of expected value. This concept is analogous to the notion of weighted functions. For a discrete random variable, X, with a probability function $P(x)$ (the discrete analogue to the continuous PDF), the expected value of X is written as $E[X]$ and is equivalent to the arithmetic mean, μ, of the probability distribution. In particular, we can write the expected value for a discrete PDF as

$$E[x] = \sum_{i=1}^{N} x_i P(x_i) = \mu \tag{3.4.1}$$

where μ is the population mean introduced in Section 3.2. The probability function $P(x)$ serves as a weighting function similar to the function f_i/N in equation (3.2.2). The difference is that f_i/N is the relative frequency for a single set of experimental samples whereas $P(x)$ is the expected relative frequency for an infinite number of samples from repeated trials of the experiment. The expected value, $E[X]$, for the sample which includes X, is the sample mean, $\bar{x}$. Similarly, the variance of the random variable X is the expected value of $(X - \mu)^2$, or

$$V[X] = E[(X - \mu)^2] = \sum_{i=1}^{N} (x - \mu)^2 P(x_i) = \sigma^2 \tag{3.4.2}$$

In the case of a continuous random variable, X, with PDF $f(x)$, the expected value is

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx \tag{3.4.3}$$

while for any function $g(X)$ with a PDF $f(x)$, the expected value can be written as

$$E[X] = \int_{-\infty}^{\infty} g(x) f(x)\, dx \quad \text{(continuous variable)} \tag{3.4.4a}$$

$$= \sum_{i=1}^{N} g(x_i)P(x_i) \quad \text{(discrete case)} \tag{3.4.4b}$$

Some useful properties of expected values for random variables are:

(1) For c = constant; $E[c] = c$, $V[c] = 0$;
(2) $E[cg(X)] = cE[g(X)]$, $V[cg(X)] = c^2V[g(X)]$;
(3) $E[g_1(X) \pm g_2(X) \pm \ldots] = E[g_1(X)] \pm E[g_2(X) \pm \ldots]$;
(4) $V[g(X)] = E[(g(X) - \mu)^2] = E[g(X)^2] - \mu^2$, (variance about the mean);
(5) $E[g_1g_2] = E[g_1]E[g_2]$;
(6) $V[g_1 \pm g_2] = V[g_1] + V[g_2] \pm 2C[g_1, g_2]$.

Property (6) introduces the *covariance function* of two variables, C, defined as

$$C[g_1, g_2] = E[g_1g_2] - E[g_1]E[g_2] \tag{3.4.5}$$

where $C = 0$ when g_1 and g_2 are independent random variances. Using properties (1) to (3), we find that $E[Y]$ for the linear relation $Y = a + bX$ can be expanded to

$$E[Y] = E[a + bX] = a + bE[X]$$

while from (1) and (6) we find

$$V[Y] = V[a + bX] = b^2V[X]$$

At this point, we remark that averages, expressed as expected values, $E[X]$, apply to ensemble averages of many (read, infinite) repeated samples. This means that each sample is considered to be drawn from an infinite ensemble of identical statistical processes varying under exactly the same conditions. In practice, we do not have repeated samples taken under identical conditions but rather time (or space) records. In using time or space averages as representative of ensemble averages, we are assuming that our records are *ergodic*. This implies that averages over an infinite ensemble can be replaced by an average over a single, infinitely long time series. An ergodic process is not to be confused with a stationary process for which the PDF of $X(t)$ is independent of time. In reality, time/space series can be considered stationary if major shifts in the statistical characteristics of the series occur over intervals that are long compared to the averaging interval so that the space/time records remain homogeneous (exhibit the same general behavior) throughout the selected averaging interval. A data record that is quiescent during the first half of the record and then exhibits large irregular oscillations during the second half of the record is not stationary.

3.4.1 Unbiased estimators and moments

As we stated earlier, $\bar{x}$ and s^2 defined by (3.2.2) and (3.2.4) are unbiased estimators of the true population mean, μ, and variance, σ^2. That is, the expected values of $\bar{x}$ and $(x - \bar{x})^2$ are equal to μ and σ^2, respectively. To illustrate the nature of the expected value, we will first prove that $E(\bar{x}) = \mu$. We write the expected value as the normalized sum of all $\bar{x}$ values

$$E[\bar{x}] = E\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum_{i=1}^{N} E[x_i] = \frac{1}{N}\sum_{i=1}^{N} \mu = \mu$$

as required. Next, we demonstrate that $E[s^2] = \sigma^2$. We again use the appropriate definitions and write

$$\begin{aligned} E[s^2] &= E\left[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2\right] \\ &= E\left[\frac{1}{N-1}\left\{\sum_{i=1}^{N}\left[(x_i - \mu)^2 - N(\bar{x} - \mu)^2\right]\right\}\right] \\ &= \frac{1}{N-1}\left\{\sum_{i=1}^{N} E\left[(x_i - \mu)^2\right] - NE\left[(\bar{x} - \mu)^2\right]\right\} \\ &= \frac{1}{N-1}\left\{\sum_{i=1}^{N}(\sigma^2) - N\frac{\sigma^2}{N}\right\} = \frac{\sigma^2}{N-1}(N-1) = \sigma^2 \end{aligned}$$

where we have used the relations $x_i - \bar{x} = (x_i - \mu) - (\bar{x} - \mu)$, $E[(x_i - \mu)^2] = V[x_i] = \sigma^2$ (the variance of an individual trial) and $E[(\bar{x} - \mu)^2] = V[\bar{x}] = \sigma^2/N$ (the variance of the sample mean relative to the population mean). The last expression derives from the central limit theorem discussed in Section 3.6.

Returning to our discussion of statistical moments, we define the ith moment of the random variable X, taken about the origin, as

$$E[X^i] = \mu_i \tag{3.4.6}$$

Thus, the first moment about the origin (i=1) is the population mean, $\mu = \mu_1$. Similarly, we can define the ith moment of X taken about the mean (called the ith central moment of X) as

$$E[(X - \mu)^i] = \mu_i \tag{3.4.7}$$

The population variance, σ^2, is the second (i = 2) central moment, μ_2.

3.4.2 Moment generating functions

Up to this point, we have computed the various characteristics of the random variable X using the probability functions directly. Now, suppose we look for a "generating" function that enables us to find all of the expected properties of the variable X using just this one function. For a discrete or continuous random variable X we define a *moment generating function* as $m(t) = E[e^{tX}]$ for the real variable, t. The moment generating function $m(t)$ serves two purposes. First, if we can find $E[e^{tX}]$, we can find any of the moments of X; second, if $m(t)$ exists it is unique and can be used to establish that both random variables have the same probability distributions. In other words, it is not possible for random variables with different probability distributions to have the same moment generating functions. Likewise, if the moment generating functions for two random variables are the same, then both variables must have the same

probability distribution. For a single-valued function X with a probability function, $P(X = x_k), k = 1, 2, \ldots$ in the discrete case, and $f(x)$ in the continuous case, the moment generating function, $m(t)$, is

$$m(t) = E[e^{tX}] = \sum_{i=1}^{\infty} e^{tx} iP(x_i) \tag{3.4.7a}$$

$$= \int_{-\infty}^{\infty} e^{tx} f(x)\, dx \tag{3.4.7b}$$

The advantages of the moment generating function become more apparent if we expand e^{tx} in the usual way to get

$$e^{tX} = 1 + tX + (tX)^2/2! + \ldots + (tX)^n/n! + \ldots$$

and apply this to $m(t)$ so that

$$\begin{aligned} m(t) = E[e^{tX}] &= E[1 + tX + (tX)^2/2! + \ldots + (tX)^n/n! + \ldots] \\ &= 1 + tE[X] + t^2E[X^2]/2! + \ldots + t^nE[X^n]/n! + \ldots \end{aligned} \tag{3.4.8}$$

Taking the derivatives of (3.4.8), we find

$$m'(t) = E[X] + tE[X^2] + \ldots + t^{n-1}E[X^n]/(n-1)! + \ldots \tag{3.4.9a}$$

$$m''(t) = E[X^2] + tE[X^3] + \ldots + t^{n-2}E[X^n]/(n-2)! + \ldots \tag{3.4.9b}$$

and so on (here, $m' \equiv dm/dt$). Setting $t = 0$ in (3.4.9) and continuing in the same way, we obtain

$$m'(0) = E[X]; m''(0) = E[X^2]; \ldots; m^{(n)}(0) = E[X^n] \tag{3.4.10}$$

In other words, we can easily obtain all the moments of the generating function $m(t)$ from the derivatives evaluated at $t = 0$. Specifically, we note that

$$E[X] = m'(0) \tag{3.4.11a}$$

$$V[X] = E[X^2] - (E[X])^2 = m''(0) - [m'(0)]^2 \tag{3.4.11b}$$

As a first example, suppose that the discrete variable X is binomially distributed with parameters N and p as in (3.3.1a). Then

$$\begin{aligned} m(t) &= \sum_{k=0}^{N} e^{tk} \binom{N}{k} p^k (1-p)^{N-k} \\ &= \sum_{k=0}^{N} \binom{N}{k} (p\, e^t)^k (1-p)^{N-k} = [p\, e^t + (1-p)]^N \end{aligned}$$

where we have used the binomial expansion

$$\binom{N}{k}$$

from (3.3.2). Taking the derivatives of this function and evaluating the results at $t = 0$, as per (3.4.11), yields the mean and variance for the binomial probability function,

$$E[X] = m'(0) = Np$$
$$V[X] = m''(0) - \{m'(0)\}^2 = Np(1-p) = Npq$$

As a further example, consider the density of a continuous random variable x given by

$$f(x) = \alpha^2 x\, e^{-\alpha x}, \quad \text{if } x > 0$$
$$= 0, \quad \text{otherwise}$$

Using (3.4.7b), we first write the moment generating function $m(t)$ as

$$m(t) = \int_{-\infty}^{\infty} e^{tx} f(x)dx = \alpha^2 \int_{o}^{\infty} x\, e^{-(\alpha - t)x}\, dx$$

For $\alpha - t > 0$, and hence $t < \alpha$

$$m(t) = \alpha^2 \int_0^{\infty} x\, e^{-(\alpha-t)x} dx$$
$$= \alpha^2 \left[\frac{\alpha e^{-(\alpha-t)x}}{(\alpha-t)} \Big|_0^{\infty} + \int_0^{\infty} \frac{e^{-(\alpha-t)x}}{(\alpha-t)} dx \right] = \alpha^2 \frac{[-e^{-(\alpha-t)}]}{(\alpha-t)^2} \Big|_0^{\infty} = \frac{\alpha^2}{(\alpha-t)^2}, \quad \text{for } t < \alpha$$

For $t \geq \alpha$, $m(t)$ is not defined. Using (3.4.11), we find

$$E[X] = m'(t=0) = \mu = \frac{2\alpha^2}{(\alpha-t)^3}\Big|_{t=0} = \frac{2}{\alpha}$$

Similarly, we find the second moment $V[X] = m''$ as

$$V[X] = m''(t=0) - \mu^2 = \frac{6\alpha^2}{(\alpha-t)^4}\Big|_{t=0} - \mu^2 = 6/\alpha^2 - 4/\alpha^2 = 2/\alpha^2$$

Several properties of moment generating functions (MGFs) are worth mentioning since they may be used to simplify more complicated functions. These are: (1) if the random variable X has a moment generating function $m(t)$, then the MGF of the random variable $Y = aX + b$ is $m(t) = e^{bt} m(at)$; (2) if X and Y are random variables with respective MGFs $m(t; X)$ and $m(t: Y)$, and if $m(t: X) = m(t: Y)$ for all t, then X and Y have the same probability distribution; (3) If X_k, $k = 1, \ldots, n$, are independent random variables with MGFs defined by $m(t: X_k)$, then the MGF of the random variable $Y = X_1 + X_2 + \ldots + X_n$ is given by the product, $m(t: Y) = m(t: X_1) m(t: X_2) \ldots m(t: X_n)$.

For convenience, the probability density functions, means, variances, and moment generating functions for several common continuous variables are presented in Appendix C. Moments allow us to describe the data in terms of their PDFs. Comparisons between moments from two random variables will establish whether or not they have the same PDF.

3.5 COMMON PROBABILITY DENSITY FUNCTIONS

The purpose of this section is to provide examples of three common PDFs. The first is the uniform PDF given by

$$\begin{aligned} f(x) &= \frac{1}{x_2 - x_1}, \quad x_1 \leq x \leq x_2 \\ &= 0, \quad \text{otherwise} \end{aligned} \tag{3.5.1}$$

(Figure 3.4) which is the intended PDF of random numbers generated by most computers and handheld calculators. The function is usually scaled between 0 and 1. The cumulative density function $F(x)$ given by (3.3.3b) has the form

$$\begin{aligned} F(x) &= 0, \quad x < x_1 \\ &= \frac{x - x_1}{x_2 - x_1}, \quad x_1 \leq x \leq x_2 \\ &= 1, \quad x \geq x_2 \end{aligned}$$

while the mean and standard deviation of (3.5.1) are given by $\mu = (x_2 + x_1)/2$ and $\sigma = (x_2 - x_1)/2\sqrt{3}$.

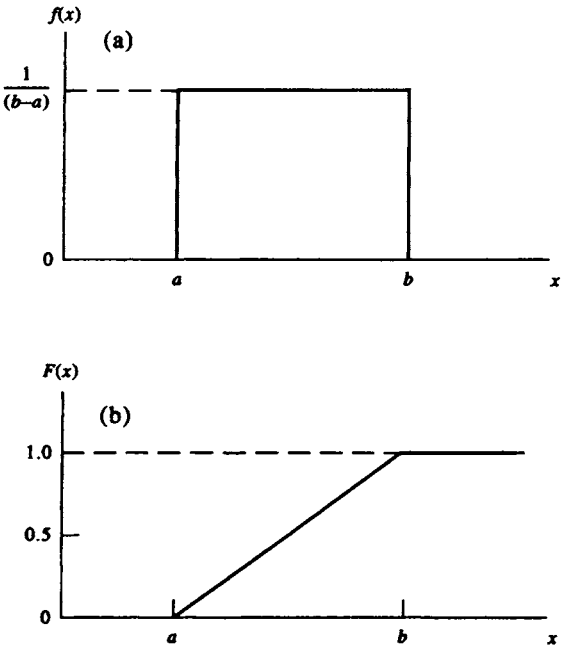

Figure 3.4. Uniform probability density distribution functions. (a) The probability density function, f(x); and (b) the corresponding cumulative probability distribution function, F(x). (From Bendat and Piersol, 1986.)

Perhaps the most familiar and widely used PDF is the normal (or Gaussian) density function:

$$f(x) = \frac{e^{\left[-(x-\mu)^2/2\sigma^2\right]}}{\sigma\sqrt{(2\pi)}}, \quad \sigma > 0, -\infty < \mu < \infty, -\infty < x < \infty \tag{3.5.2}$$

where the parameter σ represents the standard deviation (or spread) of the random variable X about its mean value μ (Figure 3.2). For convenience, (3.5.2) is often written in shorthand notion as $N(\mu, \sigma^2)$. The height of the density function at $x = \mu$ is $0.399/\sigma$. The cumulative probability distribution of a normally distributed random variable, X, lying in the interval a to b is given by the integral (3.3.5)

$$P(a \leq X \leq b) = \int_a^b \frac{e^{\left[-(x-\mu)^2/2\sigma^2\right]}}{\sigma\sqrt{(2\pi)}}\, dx \tag{3.5.3}$$

which is the area under the normal curve between a and b. Since a closed form of this integral does not exist, it must be evaluated by approximate methods, often involving the use of tables of areas. We have included a table of curve areas in Appendix D (Table D.1). The normal distribution is symmetric with respect to μ so that areas need to be tabulated only on one side of the mean. For example, $P(\mu \leq x \leq \mu + 1\sigma) = 0.3413$ so by symmetry $P(\mu - 1\sigma \leq x \leq \mu + 1\sigma) = 2(0.3413) = 0.6826$. The latter is the value used in the rule of thumb estimates for the range of the standard deviation, σ. For the normal distribution, the tabulated values represent the area between the mean and a point z, where z is the distance from the mean measured in standard deviations. This leads to the familiar transform for a normal random variable X given by

$$Z = \frac{X - \mu}{\sigma} \tag{3.5.4}$$

called the standardized normal variable. The variable Z gives the distances of points measured from the mean of the normal random variable in terms of the standard deviation of the normal random variable, X (Figure 3.5). The standard normal variable Z is normally distributed with a mean of zero (0) and a standard deviation of unity (1). Thus, if X is described by the function $N(\mu, \sigma^2)$, then Z is described by the function $N(0, 1)$.

Our third continuous PDF is the gamma density function which applies to random variables which are always nonnegative thus producing distributions that are skewed to the right. The gamma PDF is given by

$$\begin{aligned} f(x) &= \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)}\ \alpha, \beta > 0;\ 0 \leq x \leq \infty \\ &= 0, \text{elsewhere} \end{aligned} \tag{3.5.5}$$

where α and β are parameters of the distribution and $\Gamma(\alpha)$ is the gamma function

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}e^{-x}dx \tag{3.5.6}$$

For any integer n

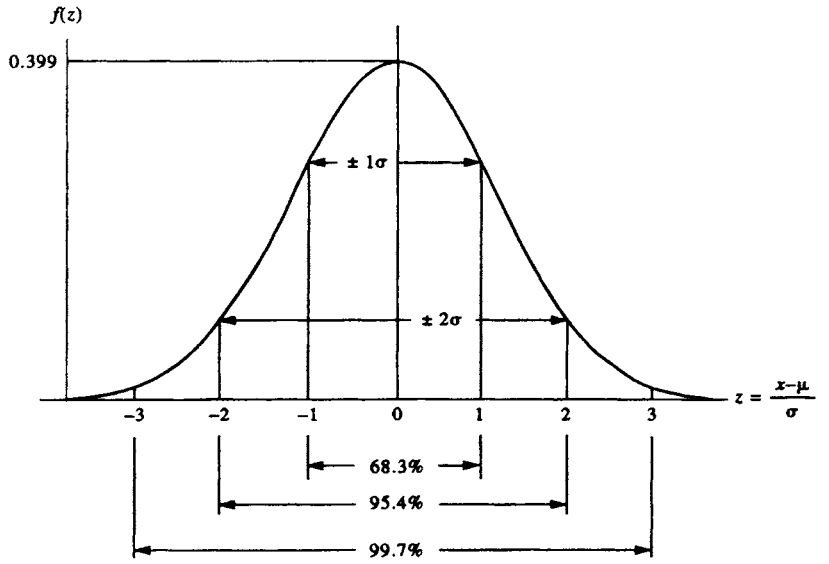

Figure 3.5. Distribution f(z) for the standardized normal random variable, $Z = (X - \mu)/\sigma$ (cf. Figure 3.2). (From Harnett and Murphy, 1975.)

$$\Gamma(n) = (n - 1) \tag{3.5.7}$$

while for a continuous variable α

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1), \text{ for } \alpha \geq 1 \tag{3.5.8}$$

where $\Gamma(1) = 1$. Plots of the gamma PDF for $\beta = 1$ and three values of the parameter α are presented in Figure 3.6. Since it is again impossible to define a closed form of the integral of the PDF in (3.5.5), tables are used to evaluate probabilities from the PDF.

One particularly important gamma density function has a PDF with $\alpha = \nu/2$ and $\beta = 2$. This is the *chi-square random distribution* (written as χ^2_ν and pronounced "ki square") with ν degrees of freedom (Appendix D, Table D.2). The chi-square distribution gets its name from the fact that it involves the square of normally distributed random variables, as we will explain shortly. Up to this point, we have dealt with a

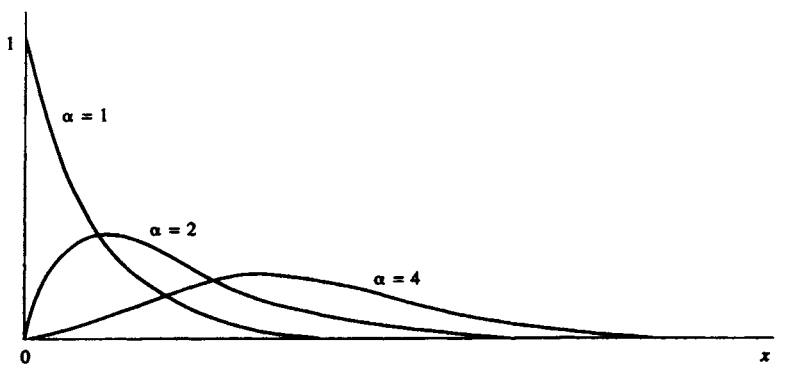

Figure 3.6. Plots of the gamma function for various values of the parameter α ($\beta = 1$).

single random variable X and its standard normalized equivalent, $Z = (X - \mu)/\sigma$. We now wish to investigate the combined properties of more than one standardized independent normal variable. For example, we might want to investigate the distributions of temperature differences between reversing thermometers and a CTD thermistor for a suite of CTD versus reversing thermometer intercomparisons taken at the same location. Each cast is considered to produce a temperature difference distribution x_k with a mean μ_k and a variance σ_k^2. The set of standardized independent normal variables Z_k formed from the casts is assumed to yield ν independent standardized normal variables $Z_1, Z_2, \ldots, Z_\nu$. The new random variable formed from the sum of the squares of the variables $Z_1, Z_2, \ldots, Z_\nu$ is the chi-square variable χ_ν^2 where

$$\chi_\nu^2 = Z_1^2 + Z_2^2 + \ldots + Z_\nu^2 \tag{3.5.9}$$

has ν degrees of freedom. For the case of our temperature comparison, this represents the square of the deviations for each cast about the mean. Properties of the distribution are

$$\text{Mean} = E[\chi_\nu^2] = \nu \tag{3.5.10a}$$

$$\text{Variance} = E[(\chi_\nu^2 - \nu)^2] = 2\nu \tag{3.5.10b}$$

We will make considerable use of the function χ_ν^2 in our discussion concerning confidence intervals for spectral estimates.

It bears repeating that probability density functions are really just models for real populations whose distributions we do not know. In many applications, it is not important that our PDF be a precise description of the true population since we are mainly concerned with the statistics of the distributions as provided by the probability statements from the model. It is not, in general, a simple problem to select the right PDF for a given data set. Two suggestions are worth mentioning: (1) Use available theoretical considerations regarding the process that generated the data; and (2) use the data sample to compute a frequency histogram and select the PDF that best fits the histogram. Once the PDF is selected, it can be used to compute statistical estimates of the true population parameters.

We also keep in mind that our statistics are computed from, and thus are functions of, other random variables and are, therefore, themselves random variables. For example, consider sample variables $X_1, X_2, \ldots, X_N$ from a normal population with mean μ and variance σ^2, then

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{3.5.11}$$

is normally distributed with mean μ and variance σ^2/N. From this it follows that

$$Z = \frac{\overline{X} - \mu}{\sigma_x} = \frac{\overline{X} - \mu}{\sigma/\sqrt{N}} = \sqrt{N}\frac{\overline{X} - \mu}{\sigma} \tag{3.5.12}$$

has a standard normal distribution $N(0, 1)$ with zero mean and variance of unity. Using the same sample, $X_1, X_2, \ldots, X_N$, we find that

$$\frac{1}{\sigma^2}\sum_{i=1}^{N}(X_i-\overline{X})^2=\frac{(N-1)s^2}{\sigma^2}=\chi_\nu^2 \tag{3.5.13}$$

has a chi-square distribution (χ_ν^2) with $\nu=(N-1)$ degrees of freedom. (Only $N-1$ degrees of freedom are available since the estimator requires use of the mean which reduces the degrees of freedom by one.) Here, the sample standard deviation, s, is an unbiased estimate of σ. We also can use $(X-\overline{X})/(s/\sqrt{N})$ as an estimate of the standard normal statistic, $(X-\mu)/(\sigma/\sqrt{N})$. The continuous sample statistic $(X-\overline{X})/(s/\sqrt{N})$ has a PDF known as the *Student's t-distribution* (Appendix D, Table D.3) with $(N-1)$ degrees of freedom. The name derives from an Irish statistician named W. S. Gossett who was one of the first to work on the statistic. Because his employer would not allow employees to publish their research, Gossett published his results under the name "Student" in 1908. Mathematically, the random variable t is defined as a standardized normal variable divided by the square root of an independently distributed chi-square variable divided by its degrees of freedom; viz. $t=z/\sqrt{(\chi^2/\nu)}$, where z is the standard normal distribution. Thus, one can safely use the normal distribution for samples $\nu>30$, but for smaller samples one must use the t-distribution. In other words, the normal distribution gives a good approximation to the t-distribution only for $\nu>30$.

The above relations for statistics computed from a normal population are important for two reasons:

(a) often, the data or the measurement errors can be assumed to have population distributions with normal probability density functions;
(b) one is working with averages that themselves are normally distributed regardless of the probability density function of the original data. This statement is a version of the well-known *central limit theorem.*

3.6 CENTRAL LIMIT THEOREM

Let X_1, X_2, ... , X_i, ... be a sequence of independent random variables with $E[X_i]=\mu_i$ and $V[X_i]=\sigma_i^2$. Define a new random variable $X=X_1+X_2+\ldots+X_N$. Then, as N becomes large, the standard normalized variable

$$Z_N=\frac{\left(X-\sum_{i=1}^{N}\mu_i\right)}{\left(\sum_{i=1}^{N}\sigma_i^2\right)^{1/2}} \tag{3.6.1}$$

takes on a normal distribution regardless of the distribution of the original population variable from which the sample was drawn. The fact that the X_i values may have any kind of distribution, and yet the sum X may be approximated by a normally distributed random variable, is the basic reason for the importance of the normal distribution in probability theory. For example, X might represent the summation of fresh water added to an estuary from a large number of rivers and streams, each with its own particular form of variability. In this case, the sum of the rivers and stream input would result in a normal distribution of the input of fresh water. Alternatively, the variable X, representing the success or failure of an AXBT launch, may be

represented as the sum of the following independent binomially-distributed random variables (a variable that can only take on one of two possible values)

$$X_i = 1 \text{ if the } i\text{th cast is a success}$$
$$= 0 \text{ if the } i\text{th cast is a failure}$$

with $X = X_1 + X_2 + \ldots + X_N$. For this random variable, $E[X] = Np$ and $V[X] = Np(1-p)$. For large N, it can be shown that the variable $(X - E[X])/\sqrt{V[X]}$ closely resembles the normal distribution, $N(0, 1)$.

A special form of the central limit theorem may be stated as: the distribution of mean values calculated from a suite of random samples X_i ($X_{i,1}, X_{i,2}, \ldots$) taken from a discrete or continuous population having the same mean μ and variance σ^2 approaches the normal distribution with mean μ and variance σ^2/N as N goes to infinity. Consequently, the distribution of the arithmetic mean

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{3.6.2}$$

is asymptotically normal with mean μ and variance σ^2/N when N is large. Ideally, we would like $N \to \infty$ but, for practical purposes, $N \geq 30$ will generally ensure that the population of X is normally distributed. When N is small, the shape of the sample distribution will depend mainly on the shape of the parent population. However, as N becomes larger, the shape of the sampling distribution becomes increasingly more like that of a normal distribution no matter what the shape of the parent population (Figure 3.7). In many instances, the normality assumption for the sampling distribution for $\overline{X}$ is reasonably accurate for $N > 4$ and quite accurate for $N > 10$ (Bendat and Piersol, 1986).

The central limit theorem has important implications for we often deal with average values in time or space. For example, current meter systems average over some time interval, allowing us to invoke the central limit theorem and assume normal statistics for the resulting data values. Similarly, data from high-resolution CTD systems are generally vertically averaged (or averaged over some set of cycles in time), thus approaching a normal PDF for the data averages, via the central limit theorem. An added benefit of this theorem is that the variance of the averages is reduced by the factor N, the number of samples averaged. The theorem essentially states that individual terms in the sum contribute a negligible amount to the variation of the sum, and that it is not likely that any one value makes a large contribution to the sum. Errors of measurements certainly have this characteristic. The final error is the sum of many small contributions none of which contributes very much to the total error. Note that the sample standard error is an unbiased estimate (again in the sense that the expected value is equal to the population parameter being estimated) even though the component sample standard deviation is not.

To further illustrate the use of the central limit theorem, consider a set of independent measurements of a process whose probability distribution is unknown. Through previous experimentation, the distribution of this process was estimated to have a mean of 7 and a variance of 120. If $\bar{x}$ denotes the mean of the sample measurements, we want to find the number of measurements, N, required to a give a probability

$$P(4 \leq \bar{x} \leq 10) = 0.866$$

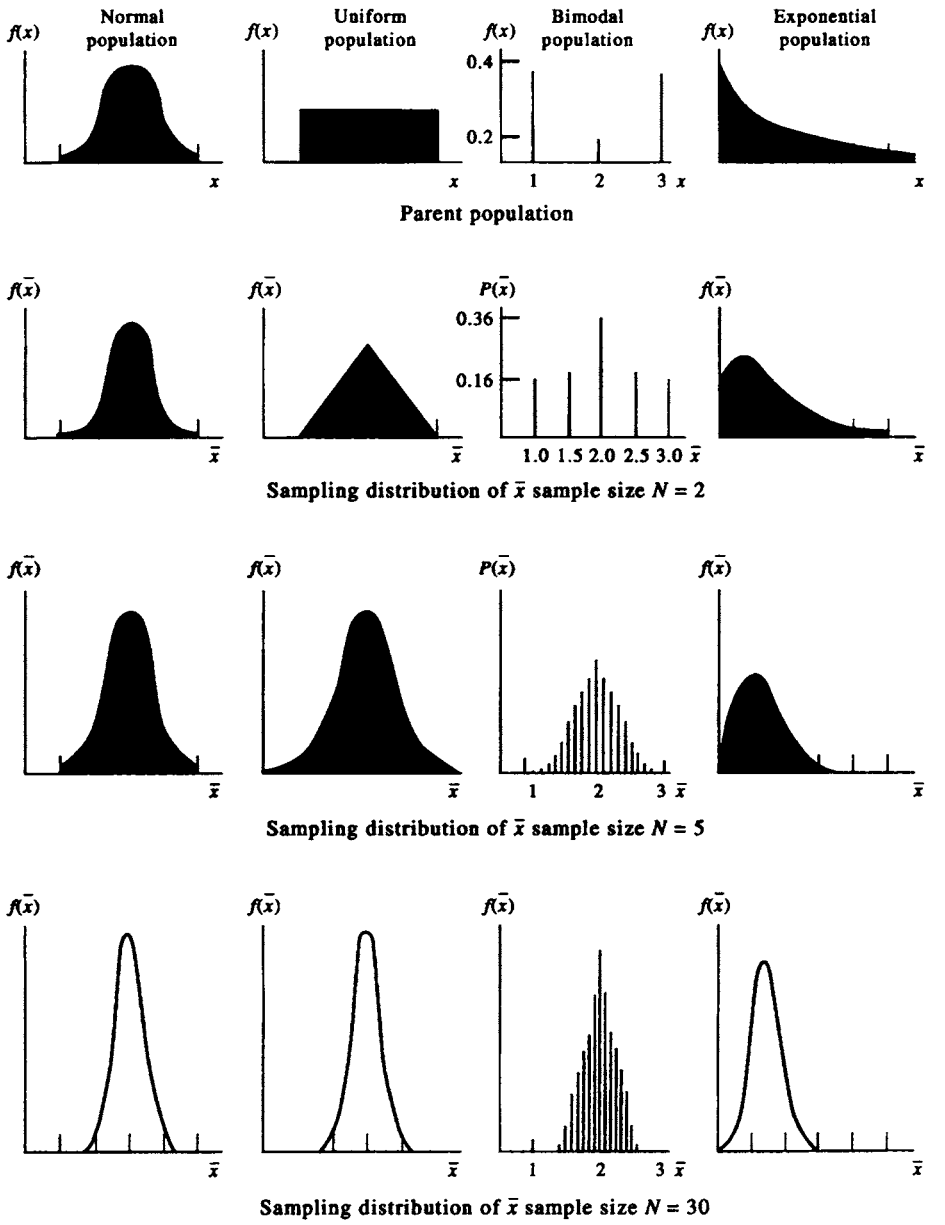

Figure 3.7. Sampling distribution of the mean $\bar{x}$ for different types of population distributions for increasing sample size, $N = 2$, 5, and 30. The shape of the sampling distribution becomes increasingly more like that of a normal distribution regardless of the shape of the parent population.

where 4 and 10 are the chosen problem limits. Here, we use the central limit theorem to argue that, while we don't know the exact distribution of our variable, we do know that means are normally distributed. Using the standard variable, $z = (x - \mu)/(\sigma/\sqrt{N})$, substituting $\bar{x}$ for x, and using the fact that $\sigma = \sqrt{120} = 2\sqrt{30}$, we can then write our probability function as

$$
\begin{aligned}
P(4 \le \bar{x} \le 10) &= P\left[\frac{(4-\mu)\sqrt{N}}{\sigma} < z < \frac{(10-\mu)\sqrt{N}}{\sigma}\right] \\
&= P\left[\frac{(4-7)\sqrt{N}}{2\sqrt{30}} < z < \frac{(10-7)\sqrt{N}}{2\sqrt{30}}\right] \\
&= P\left[\frac{-3\sqrt{N}}{2\sqrt{30}} < z < \frac{3\sqrt{N}}{2\sqrt{30}}\right] \\
&= 2P\left[z < \frac{3\sqrt{N}}{2\sqrt{30}}\right] - 1 = 0.866
\end{aligned}
$$

from which we find

$$
P\left[z < \frac{3\sqrt{N}}{2\sqrt{30}}\right] = 0.933
$$

Assuming that we are dealing with a normal distribution, we can look up the value 0.933 in a table to find the value of the integrand to which this cumulative probability corresponds. In this case, $3/2\sqrt{(N/30)} = 1.5$, so that $N = 30$.

3.7 ESTIMATION

In most oceanographic applications, the population parameters are unknown and must be estimated from a sample. Faced with this estimation problem, the objective of statistical analysis is twofold: To present criteria that allow us to determine how well a given sample represents the population parameter; and to provide methods for estimating these parameters. An *estimator* is a random variable used to provide *estimates* of population parameters. "Good" estimators are those that satisfy a number of important criteria: (1) Have average values that equal the parameter being estimated (*unbiasedness* property); (2) have relatively small variance (*efficiency* property); and (3) approach asymptotically the value of the population parameter as the sample size increases (*consistency* property). We have already introduced the concept of estimator bias in discussing variance and standard deviation. Formally, an estimate $\hat{\theta}$ of a parameter θ (here, the hat symbol ($\hat{}$) indicates an estimate), is an unbiased estimate provided that $E[\hat{\theta}] = \theta$; otherwise, it is a biased estimate with a bias $B = E[\hat{\theta}] - \theta$. An unbiased estimator is any estimate whose average value over all possible random samples is equal to the population parameter being estimated. An example of an unbiased estimator is the mean of the noise in an acoustic current meter record created by turbulent fluctuations in the velocity of sound speed in water; an example of a biased estimator is the linear slope of a sea-level record in the presence of a long-term trend (a slow change in average value). Other examples of unbiased estimators are $\bar{x}$ for $\hat{\theta}$, μ for $E[\hat{\theta}]$, and σ^2/N for $\sigma^2_{\hat{\theta}}$. The mean square error of our estimate $\hat{\theta}$ is

$$
E[(\hat{\theta} - \theta)^2] = V[\hat{\theta}] + B^2 \tag{3.7.1}
$$

The most efficient estimator (property 2) is the estimator with the smallest mean square error. Since it is possible to obtain more than one unbiased estimator for the

same target parameter, θ, we define the efficiency of an estimator as the ratio of the variances of the two estimators. For example, if we have two unbiased estimates $\hat{\theta}_1$ and $\hat{\theta}_2$, we can compute the relative efficiency of these two estimates as

$$\text{efficiency} = V\left[\hat{\theta}_2\right] / V\left[\hat{\theta}_1\right] \tag{3.7.2}$$

where $V[\hat{\theta}_1]$ and $V[\hat{\theta}_2]$ are the variances of the estimators. A low value of the ratio would suggest that $V[\hat{\theta}_2]$ is more efficient while a high value would indicate that $V[\hat{\theta}_1]$ is more efficient. As an example, consider the efficiency of two familiar estimators of the mean of a normal distribution. In particular, let $\hat{\theta}_1$ be the median value and $\hat{\theta}_2$ be the sample mean. The variance of the sample median is $V[\hat{\theta}_1] = (1.2533\sigma)^2/N$ while the sample mean has a variance $V[\hat{\theta}_2] = \sigma^2/N$. Thus, the efficiency is

$$\begin{aligned} \text{efficiency} &= V[\hat{\theta}_2]/V[\hat{\theta}_1] \\ &= (\sigma^2/N)/(1.2533^2\sigma^2/N) \\ &= 0.6366 \end{aligned}$$

Therefore, the variability of the sample mean is 63% of the variability of the sample median, which indicates that the sample *mean* is a more efficient estimator than the sample *median*.

As a second example, consider the sample variances s'^2 and s^2 given by (3.2.3) and (3.2.4), respectively. The efficiency of these two sample variances is the ratio of s'^2 to s^2

$$\frac{1/N \sum_{i=1}^{N} (x_i - \bar{x})^2}{1/(N-1) \sum_{i=1}^{N} (x_i - \bar{x})^2} = \frac{N-1}{N} < 1$$

which indicates that s'^2 is a more efficient statistic than s^2.

We can view the difference $\hat{\theta} - \theta$ as the distance between the population "target" value and our estimate. Since this difference is also a random variable, we can ask probability-related questions, such as: "What is the probability

$$P(-b < (\hat{\theta} - \theta) < b)$$

for some range $(-b, b)$?" It is common practice to express b as some multiple of the sample standard deviation of θ (e.g. $b = k\sigma_\theta$, $k > 1$). A widely used value is $k = 2$, corresponding to two standard deviations. Here, we can apply an important result known as *Tshebysheff's theorem* which states that for any random variable Y, for $k > 0$:

$$P(|Y - \mu| < k\sigma) \geq 1 - \frac{1}{k^2} \tag{3.7.3a}$$

or

$$P(|Y - \mu| \geq k\sigma) \leq \frac{1}{k^2} \tag{3.7.3b}$$

where $\mu = E[\hat{Y}]$ and $\sigma^2 = V[\hat{Y}]$. Applying this to the problem at hand, we find that for $k = 2$, $P(|\hat{\theta} - \theta| < 2\sigma_\theta) \geq 1 - 1/(2)^2 = 0.75$. Therefore, most random variables occurring in nature can be found within two standard deviations ($\pm 2\sigma$) of the mean

with a probability of 0.75. Note that the probability statement (3.7.3a) indicates that the probability is greater than or equal to the value of $1 - 1/k^2$ for any type of distribution. We can, therefore, expect somewhat more than 75% of the values to lie with the range $(-2\sigma, 2\sigma)$. In fact, this is generally a conservative estimate. If we assume that oceanographic measurements are typically normally distributed, we find $P(|Y - \mu| < 2\sigma) = 0.95$, so that 95% of the observations lie within $\pm 2\sigma$. This is an important conclusion in terms of editing methods which use criteria designed to select erroneous values from data samples based on probabilities.

3.8 CONFIDENCE INTERVALS

An important application of interval estimates for probability distribution functions is the formulation of *confidence intervals* for parameter estimates. These intervals define the degree of certainty that a given quantity θ will fall between specified lower and upper bounds θ_L, θ_U, respectively, of the parameter estimates. The confidence interval (θ_L, θ_U) associated with a particular confidence statement is usually written as

$$P(\theta_L < \theta < \theta_U) = 1 - \alpha, \quad 0 < \alpha < 1 \tag{3.8.1}$$

where α is called the *level of significance* (or confidence coefficient) for the confidence statement and $(1 - \alpha)100$ is the percent significance level for the variable θ. (The terms confidence coefficient, significance level, confidence level and confidence are commonly used interchangeably). A typical value for α is 0.05, which means that 95% of the cumulative area under the probability curve (3.8.1) is contained between the points θ_L and θ_U (Figure 3.8). For both symmetric and nonsymmetric probability distributions, each of the two points cuts off $\alpha/2$ of the total area under the distribution curve, leaving a total area under the curve of $1 - \alpha$; θ_L cuts off the left-hand part of the distribution tail and θ_U cuts off the right-hand part of the tail.

If θ_L, θ_U are derived from the true value of the variable θ (such as the population mean, μ), then the probability interval is fixed. However, where we are using sample estimates (for example, the mean, $\overline{X}$) to determine the variable value, θ, the probability interval will vary from sample to sample because of changes in the sample mean and standard deviation. Thus, we must inquire about the probability that the true value of θ will fall within the intervals generated by each of the given sample estimates. The statement $P(\theta_L < \theta < \theta_U)$ does not mean that the population variable θ has a probability of $P = 1 - \alpha$ of falling in the sample interval (θ_L, θ_U), in the sense that θ was behaving like a sample. The population variable is a fixed quantity. Once

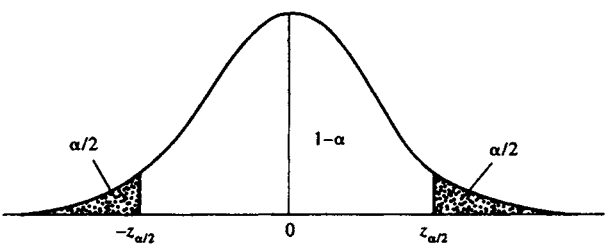

Figure 3.8. Location of the limits $(\theta_L, \theta_U) = (-z_{\alpha/2}, +z_{\alpha/2})$ for a normal probability distribution. For $\alpha = 0.05$, the cumulative area $1 - \alpha$ corresponds to the 95% interval for the distribution.

the interval is picked, the population variable θ is either in the interval or it isn't (probability 1 or 0). For the sample data, the interval may shift depending on the mean and variance of the particular sample we select from the population. We should, therefore, interpret (3.8.1) to mean that there is a probability P that the specified random sample interval (θ_L, θ_U) contains the true population variable θ a total of $(1 - \alpha)$ 100% of the time. That is, $(1 - \alpha)$ is the fraction of the time that the true variable value θ is contained by the sample interval (θ_L, θ_U).

In general, we need a quantity, called a *pivotal quantity*, that is a function of the sample estimator $\hat{\theta}$ and the unknown variable θ, where θ is the only unknown. The pivotal quantity must have a PDF that does not depend on θ. For large samples $(N \geq 30)$ of unbiased point estimators, the standard normal distribution $Z = (\hat{\theta} - \theta)/\sigma_{\hat{\theta}}$ is a pivotal quantity. In fact, it is common to express the confidence interval in terms of Z. For example, consider the statistic $\hat{\theta}$ with $E[\hat{\theta}] = \theta$ and $V[\hat{\theta}] = \sigma_{\hat{\theta}}^2$; find the $100(1 - \alpha)\%$ confidence interval. To do this, we first define

$$P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha \tag{3.8.2}$$

and then use the above relation $Z = (\hat{\theta} - \theta)/\sigma_{\hat{\theta}}$ to get

$$P(\hat{\theta} - Z_{\alpha/2}\sigma_{\hat{\theta}} < \theta < \hat{\theta} + Z_{\alpha/2}\sigma_{\hat{\theta}}) = 1 - \alpha \tag{3.8.3}$$

This formula can be used for large samples to compute the confidence interval for θ once α is selected. Again, the significance level, $1 - \alpha$, refers to the probability that the population parameter θ will be bracketed by the given confidence interval. The meaning of these limits is shown graphically in Figure 3.8 for a normal population. We remark that if the population standard deviation σ is known it should be used in (3.8.3) so that $\sigma_{\hat{\theta}} = \sigma$; if not, the sample standard deviation s can be used with little loss in accuracy if the sample size is sufficiently large (i.e. $N > 30$).

The three types of confidence intervals commonly used in oceanography are listed below. Specific usage depends on whether we are interested in the mean or the variance of the quantity being estimated.

3.8.1 Confidence interval for μ (σ known)

When the population standard deviation, σ, is known and the parent population is normal (or $N > 30$), the $100(1 - \alpha)$ percent confidence interval for the population mean is given by the symmetrical distribution for the standardized normal distribution, z

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}} \tag{3.8.4}$$

As an example, we wish to find the 95% confidence interval ($\alpha = 0.05$) for μ given the sample mean $\bar{x}$ and known normally distributed population variance, σ^2. Suppose that a thermister installed at the entrance to the ship's engine cooling water intake samples every second for $N = 20$ s and yields a mean ensemble temperature $\bar{x} = 12.7°C$ for the particular burst. Further, suppose that the water is isothermal and that the only source of variability is instrument noise, which we know from previous calibration in the lab has a known noise level $\sigma = 0.5°C$. Since we want the 95% confidence interval, the appropriate values of z for the normal distribution are $z_{\alpha/2} = 1.96$ and $-z_{\alpha/2} = -1.96$ (Appendix D, Table D.1). Substituting these values into (3.8.4) along

with $N = 20$, $\sigma = 0.5°C$, and $\bar{x} = 12.7°C$ we find

$$[12.7 - (1.96)0.5/\sqrt{20}]°C < \mu < [12.7 + (1.96)0.5/\sqrt{20}]°C$$

$$12.48°C < \mu < 12.92°C$$

Based on our 20 data values, there is a 95% probability that the true mean temperature of the water will be bracketed by the interval (12.48°C, 12.92°C) derived from the random interval

$$(\bar{x} - z_{\alpha/2}\sigma/\sqrt{N}, \bar{x} + z_{\alpha/2}\sigma/\sqrt{N})$$

3.8.2 Confidence interval for μ (σ unknown)

In most real circumstances, σ is not known and we must resort to the use of the sample standard deviation, s. Similarly, for small samples ($N < 30$), we cannot use the above technique but must introduce a formalism that works for any sample size and distribution, as long as the departures from normality are not excessive. Under these conditions, we resort to the variable $t = (\bar{x} - \mu)/(s/\sqrt{N})$, which has a Student's t-distribution with $\nu = (N - 1)$ degrees of freedom. Derivation of the $100(1 - \alpha)\%$ confidence interval follows the same procedure used for the symmetrically distributed normal distribution, except that we must modify the limits. In this case

$$P\left[-t_{\alpha/2,\nu} < (\bar{x} - \mu) \Big/ \frac{s}{\sqrt{N}} < t_{\alpha/2,\nu}\right] = 1 - \alpha \tag{3.8.5}$$

This is easily arranged to give the $100(1 - \alpha)\%$ confidence interval for μ

$$\bar{x} - t_{\alpha/2,\nu}\frac{s}{\sqrt{N}} < \mu < \bar{x} + t_{\alpha/2,\nu}\frac{s}{\sqrt{N}} \tag{3.8.6}$$

Note the similarity between (3.8.6) and the form (3.8.3) obtained for μ when σ is known. We return to our previous example of ship injection temperature and this time assume that $s = 0.5°C$ is a measured quantity obtained by subtracting the mean value $\bar{x} = 12.7°C$ from the series of 20 measurements. Turning to Appendix D (Table D.3) for the cumulative t-distribution, we look for values of $F(t)$ under the column for the 95% confidence interval ($\alpha = 0.05$) for which $F(t) = 1 - \alpha/2 = 0.975$. Using the fact that $\nu = (N - 1) = 19$, we find $t_{\alpha/2,\nu} = t_{0.025,19} = 2.093$. Substituting these values into (3.8.6), yields

$$[12.7 - 2.093(0.5/\sqrt{20})]°C < \mu < [12.7 + 2.093(0.5/\sqrt{20})]°C$$

$$12.47°C < \mu < 12.93°C$$

There is a 95% chance that the interval (12.47°C, 12.93°C) will bracket the true mean temperature. Because of the large sample size, this result is only slightly different than the result obtained for the normal distribution in the previous example when σ was known *a priori*.

3.8.3 Confidence interval for σ^2

Under certain circumstances, we are more interested in the confidence interval for the signal variance than the signal mean. For example, to determine the reliability of a spectral peak in a spectral density distribution (or spectrum), we need to know the confidence intervals for the population variance, σ^2, based on our sample variance, s^2. To do this, we seek a new pivotal quantity. Recall from (3.5.13) that for N samples of a variable x_i from a normal population, the expression

$$\frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{(N-1)/s^2}{\sigma^2} \tag{3.8.7}$$

is a χ^2 variable with $(N-1)$ degrees of freedom. Using this as a pivotal quantity, we can find upper and lower bounds χ_U^2 and χ_L^2 for which

$$P\left[\chi_L^2 < \frac{N-1}{\sigma^2/s^2} < \chi_U^2\right] = 1 - \alpha \tag{3.8.8}$$

or, upon rearranging terms,

$$P\left[\frac{(N-1)s^2}{\chi_L^2} < \sigma^2 < \frac{(N-1)s^2}{\chi_U^2}\right] = 1 - \alpha \tag{3.8.9}$$

Note that χ^2 is a skewed function (Figure 3.9), which means that the upper and lower bounds in (3.8.9) are asymmetric; the point $1 - \alpha/2$ rather than $-\alpha/2$ determines the point that cuts off $\alpha/2$ of the area at the lower end of the chi-square distribution.

From expression (3.8.9) we obtain the well-known $100(1-\alpha)\%$ confidence interval for the variance σ^2 when sampled from a normal population

$$\frac{(N-1)s^2}{\chi_{\alpha/2,\nu}^2} < \sigma^2 < \frac{(N-1)s^2}{\chi_{1-\alpha/2,\nu}^2} \tag{3.8.10}$$

where the subscripts $\alpha/2$ and $1-\alpha/2$ characterize the endpoint values of the confidence interval and $\nu = (N-1)$ gives the degrees of freedom of the chi-square distribution. The larger value of $\chi^2(=\chi_{\alpha/2,\nu}^2)$ appears in the denominator of the lower endpoint for σ^2 while the smaller value of $\chi^2(=\chi_{1-\alpha/2,\nu}^2)$ is in the denominator of the

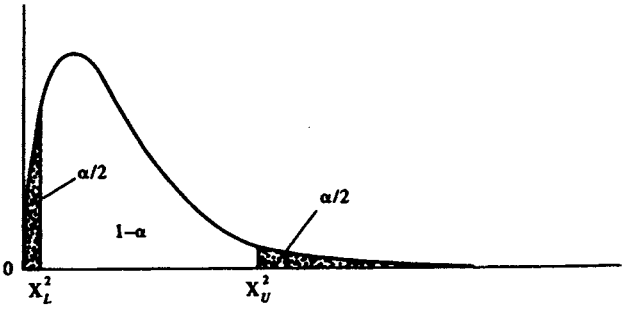

Figure 3.9. Location of the limts $(\theta_L, \theta_U) = (\chi_L^2, \chi_U^2)$ for a chi-square probability distribution. For $\alpha = 0.05$, the cumulative area $1 - \alpha$ corresponds to the 95% interval for the distribution. χ^2 is a skewed function so that the upper and lower bounds are asymmetric.

upper endpoint for σ^2. As an example, suppose that we have $\nu = 9$ in our spectral estimate of the eastward component of current velocity and that the background variance of our spectra near a distinct spectral peak is $s^2 = 10\ \mathrm{cm}^2/\mathrm{s}^2$. What is the 95% confidence interval for the variance? How big would the peak have to be to stand out statistically from the background level? (Details on spectral estimation can be found in Chapter 5.) In this case, $\alpha/2 = 0.025$ and $1 - \alpha/2 = 0.975$. Turning to the cumulative distribution $F(\chi^2)$ for 9 degrees of freedom in Appendix D (Table D.2), we find that $\chi^2_9 = 2.70$ for a cumulative integral $F(\alpha/2 = 0.025)$ and that $\chi^2_9 = 19.02$ for a cumulative integral $F(1 - \alpha/2 = 0.975)$. Thus, $P(2.70 < \chi^2_{\nu=9} < 19.02) = 1 - \alpha = 0.95$. Substituting $N - 1 = 9, s^2 = 10\,\mathrm{cm}^2/\mathrm{s}^2$, $\chi^2_{\alpha/2,\nu} = 19.02$ for the value that cuts off $\alpha/2$ of the upper end area under the curve and $\chi^2_{1-\alpha/2,\nu} = 2.70$ for the value that cuts off $1 - \alpha/2$ of the lower end area of the curve, (3.8.10) yields

$$9(10\ \mathrm{cm}^2/\mathrm{s}^2)/19.02 < \sigma^2 < 9(10\ \mathrm{cm}^2/\mathrm{s}^2)/2.70$$
$$4.7\,\mathrm{cm}^2/\mathrm{s}^2 < \sigma^2 < 33.3\,\mathrm{cm}^2/\mathrm{s}^2$$

Thus, the true background variance will lie between 4.7 and 33.3 $\mathrm{cm}^2/\mathrm{s}^2$. If a spectral peak has a greater range than these limits then it represents a statistically significant departure from background energy levels.

In most instances, spectral densities are presented in terms of the log of the spectral density function versus frequency or log-frequency (see Chapter 5). Dividing through by s^2 in (3.8.10) and taking the log, yields

$$\log(N-1) - \log\left(\chi^2_{\alpha/2.\nu}\right) < \log(\sigma^2/s^2) < \log(N-1) - \log\left(\chi^2_{1-\alpha/2.\nu}\right)$$

or, upon subtracting $\log(N-1)$ and rearranging the inequality

$$\log\left(\chi^2_{1-\alpha/2,\nu}\right) < \log(N-1) - \log(\sigma^2/s^2) < \log\left(\chi^2_{\alpha/2.\nu}\right)$$

The range R of the variance is then

$$R = \log\left(\chi^2_{\alpha/2,\nu}\right) - \log\left(\chi^2_{1-\alpha/2,\nu}\right) \tag{3.8.11}$$

while the pivot point p_o of the interval is

$$p_o = \log(N-1) - \log(\sigma^2/s^2) \tag{3.8.12}$$

If we assume that the measured background value of s^2 is a good approximation to σ^2, so that $\sigma^2/s^2 = 1$, then $p_o = \log(N-1)$. The ranges between the maximum value and p_o, and the minimal value and p_o, are $\log(\chi^2_{\alpha/2,\nu}) - p_o$ and $p_o - \log(\chi^2_{1-\alpha/2.\nu})$, respectively. Returning to our previous example for the 95% confidence interval, we find that

$$\log(2.70) < \log(9) < \log(19.02), \quad 0.43 < 0.95 < 1.28$$

giving a range $R = 0.848$ with the pivot point at $p_o = 0.95$.

3.8.4 Goodness-of-fit test

When the set of outcomes for an experiment is limited to two outcomes (such as success or failure, on or off, and so on), the appropriate test statistic for the

distribution is the binomial variable. However, when more than two outcomes are possible, the preferred statistic is the chi-square variable. In addition to providing confidence intervals for spectral estimates and other measurement parameters, the chi-square variable is used to test how closely the observed frequency distribution of a given parameter corresponds to the expected frequency distribution for the parameter. The expected frequencies represent the average number of values expected to fall in each frequency interval based on some theoretical probability distribution, such as a normal distribution. The observed frequency distribution represents a sample of values drawn from some probability distribution. What we want to know is whether the observed and expected frequencies are similar enough for us to conclude that they are drawn from the same probability function (the "null hypothesis"). The test for this similarity using the chi-square variable is called a "goodness-of-fit" test.

Consider a sample of N observations from a random variable X with observed probability density function $p_o(X)$. Let the N observations be grouped into K intervals (or categories) called *class intervals*, whose graphical distribution forms a frequency histogram (Bendat and Piersol, 1986). The actual number of observed values that fall within the ith class interval is denoted by f_i, and is called the *observed frequency* in the ith class. The number of observed values that we would expect to fall within the ith class interval if the observations really followed the theoretical probability distribution, $p(X)$, is denoted F_i, and is called the *expected frequency* in the ith class interval. The difference between the observed frequency and the expected frequency for each class interval is given by $f_i - F_i$. The total discrepancy for all class intervals between the expected and observed distributions is measured by the sample statistic

$$\mathbf{X}^2 = \sum_{i=1}^{K} \frac{(f_i - F_i)^2}{F_i} \tag{3.8.13}$$

where division by F_i transforms the sum of the squares into the chi-square-type variable, X^2.

The number of degrees of freedom, ν, for the variable X^2 is equal to K minus the number of different independent linear restrictions imposed on the observations. As discussed by Bendat and Piersol (1986), one degree of freedom is lost through the restriction that, if $K - 1$ class intervals are determined, the Kth class interval follows automatically. If the expected theoretical density function is normally distributed then the mean and variance must be computed to allow comparison of the observed and expected distributions. This results in the loss of two additional degrees of freedom. Consequently, if the chi-square goodness-of-fit test is used to test for normality of the data, the true number of degrees of freedom for X^2 is $\nu = K - 3$.

Formula (3.8.13) measures the goodness-of-fit between f_i and F_i as follows: when the fit is good (that is, f_i and F_i are generally close), then the numerator of (3.8.13) will be small and the hence the value of X^2 will be low. On the other hand, if f_i and F_i are not close, the numerator of (3.8.13) will be comparatively large and the value of X^2 will be large. Thus, the critical region for the test statistic X^2 will always be in the upper tail of the chi-square function because we wish to reject the null hypothesis whenever the difference between f_i and F_i is large. More specifically, the region of acceptance of the null hypothesis (see Section 3.14) is

$$X^2 \leq \chi^2_{\alpha;\nu} \tag{3.8.14}$$

where the value of $\chi^2_{\alpha;\nu}$ is available from Appendix D (Table D.2). If the sample value X^2 is greater than $\chi^2_{\alpha;\nu}$, the hypothesis that $p(X) = p_o(X)$ is rejected at the level of significance. If X^2 is less than or equal to $\chi^2_{\alpha;\nu}$, the hypothesis is accepted at the α level of significance (i.e. there is a $100\alpha\%$ chance that we are wrong in accepting the null hypothesis that our data are drawn from a normal distribution). For example, suppose our analysis involves 15 class intervals and that the fit between the 15 estimates of f_i and F_i (where F_i is normally distributed) yields $X^2 = 23.1$. From tables for the cumulative chi-square distribution, $F(X) = p(x > \chi^2_{\alpha;\nu})$, we find that $p(X^2 > 21.03) = 0.05$ for $\nu = K - 3 = 12$ degrees of freedom. Thus, at the $\alpha = 0.05$ level of significance (95% certainty level) we cannot accept the null hypothesis that the observed values came from the same distribution as the expected values.

Chi-square tests for normality are typically performed using a constant interval width. Unless one is dealing with a uniform distribution, this will yield different expected frequency distributions from one class interval to the next. Bendat and Piersol recommend a class interval width of $\Delta x \approx 0.4s$, where s is the standard deviation of the sample data. A further requirement is that the expected frequencies in all class intervals be sufficiently large that X^2 in (3.8.13) is an acceptable approximation to $\chi^2_{\alpha;\nu}$. A common recommendation is that $F_i > 3$ in all class intervals. When testing for normality, where the expected frequencies diminish on the tails of the distribution, this requirement is attained by letting the first and last intervals extend to $-\infty$ to $+\infty$, respectively, so that $F_1, F_K > 3$.

As an example of a goodness-of-fit test, we consider a sample of $N = 200$ surface gravity wave heights measured every 0.78 s by a Datawell waverider buoy deployed off the west coast of Canada during the winter of 1993–1994 (Table 3.8.1). The wave record spans a period of 2.59 min and corresponds to a time of extreme (5 m high) storm-generated waves. According to one school of thought (e.g. Phillips *et al.*, 1993),

Table 3.8.1. Wave heights (mm) during a period of anomalously high waves as measured by a Datawell waverider buoy deployed in 30 m depth on the inner continental shelf of Vancouver Island, British Columbia. The original N = 200 time-series values have been rank ordered. The upper bounds of the K-class intervals have been underlined. (Courtesy, Diane Masson)

4636	4840	4901	4950	4980	5014	5034	5060	5095	5130
4698	4842	4904	4954	4986	5014	5037	5066	5095	5135
4702	4848	4907	4955	4991	5015	5037	5066	5096	5135
4731	4854	4907	4956	4994	5017	5038	5069	5102	5145
4743	4856	4908	4956	4996	5020	5039	5069	5103	5155
4745	4867	4914	4956	4996	5020	5040	5071	5104	5157
4747	4867	4916	4959	4996	5021	5040	5072	5104	5164
4749	4870	4917	4960	4997	5023	5044	5073	5104	5165
4773	4870	4923	4961	4998	5024	5045	5074	5106	5166
4785	4874	4925	4963	5003	5025	5045	5074	5110	5171
4793	4876	4934	4964	5006	5025	5047	5074	5111	5175
4814	4877	4935	4964	5006	5025	5048	5078	5115	5176
4817	4883	4937	4966	5006	5025	5050	5079	5119	5177
4818	4885	4939	4966	5006	5028	5051	5080	5119	5181
4823	4886	4940	4970	5006	5029	5052	5081	5120	5196
4824	4892	4941	4971	5010	5029	5053	5086	5121	5198
4828	4896	4942	4972	5011	5029	5057	5089	5122	5201
4829	4897	4942	4974	5011	5030	5058	5091	5123	5210
4830	4898	4943	4977	5012	5031	5059	5093	5125	5252
4840	4899	4944	4979	5012	5032	5059	5094	5127	5299

extreme wave events in the ocean are part of a Gaussian process and the occurrence of maximum wave heights is related in a linear manner to the statistical distribution of the surrounding wave field. If this is true, then the heights of high-wave events relative to the background seas should follow a normal frequency distribution. To test this at the $\alpha = 0.05$ significance level, $K = 10$ class intervals for the observed wave heights were fitted to a Gaussian probability distribution. The steps in the goodness-of-fit test are as follows:

(1) Specify the class interval width Δx and list the upper limit of the standardized values, z_α, of the normal distribution that correspond to these values (as in Table 3.8.2). Commonly Δx is assumed to span 0.4 standard deviations, s, such that $\Delta x \approx 0.4s$; here we use $\Delta x \approx 0.5s$. For $\Delta x = 0.4s$, the values of z_α we want are (... , −2.4, −2.0, ... , 2.0, 2.4, ...) while for $\Delta x = 0.5s$, the values are (... , −2.5, −2.0, ..., 2.0, 2.5, ...).

(2) Determine the finite upper and lower bounds for z_α from the requirement that $F_i > 3$. Since $F_i = NP_i$ (where $N = 200$ and P_i is the normal probability distribution for the ith interval), we require $P > 3/N = 0.015$. From tables of the standardized normal density function, we find that $P > 0.015$ implies a lower bound $z_\alpha = -2.0$, and an upper bound $z_\alpha = +2.0$. Note that for a larger sample, say $N = 2000$, we have $P > 3/2000 = 0.0015$ and the bounds become ± 2.8 for the interval $\Delta x = 0.4s$ and ± 2.5 for the interval $\Delta x = 0.5s$.

(3) Calculate the expected upper limit, $x = sz_\alpha + \bar{x}$ (mm), for the class intervals and mark this limit on the data table (Table 3.8.1). For each upper bound, z_α, in Table 3.8.2, find the corresponding probability density value. Note that these values apply to intervals so, for example, $P(-2.0 < x < -1.5) = 0.0668 - 0.0228 = 0.044$; $P(2.0 < x < \infty) = 0.0228$.

(4) Using the value of P, find the expected frequency values $F_i = NP_i$. The observed frequency f_i is found from Table 3.8.1 by counting the actual number of wave heights lying between the marks made in step 3. Complete the table and calculate X^2. Compare to $\chi^2_{\alpha;\nu}$.

Table 3.8.2. Calculation table for goodness-of-fit test for the data in Table 3.8.1. The number of intervals has been determined using an interval width $\Delta x = 0.5s$, *with* z_α *in units of 0.5 and requiring that* $F_i > 3$. $N = 200$, $\bar{x}$ *(mean)* $= 4997.6$ *mm,* s *(standard deviation)* $= 115.1$ *mm, and* ν *(degrees of freedom)* $= K - 3 = 7$

Class interval	Upper limit of data interval z_α	$x = sz_\alpha + \bar{x}$	P_i	$F_i = NP_i$	f_i	$F_i - f_i$	$\frac{(F_i - f_i)^2}{F_i}$
1	−2.0	4767.4	0.0228	4.6	8	3.4	2.51
2	−1.5	4825.0	0.0440	8.8	8	0.8	0.07
3	−1.0	4882.5	0.0919	18.4	16	2.4	0.31
4	−0.5	4940.1	0.1498	30.0	23	7.0	1.63
5	0	4997.6	0.1915	38.3	33	5.3	0.73
6	0.5	5055.2	0.1915	38.3	48	9.7	2.46
7	1.0	5112.7	0.1498	30.0	35	5.0	0.83
8	1.5	5170.2	0.0919	18.4	18	0.4	0.01
9	2.0	5227.8	0.0440	8.8	9	0.2	0.00
10	∞	∞	0.0228	4.6	2	2.6	1.47
Totals			1.0000	200	200		10.02

In the above example, $X^2 = 10.02$ and there are $\nu = 7$ degrees of freedom. From Table A.3, we find $P(X^2 > \chi^2_{\alpha;\nu}) = P(X^2 > \chi^2_{0.05;7}) = 14.07$. Thus, at the $\alpha = 0.05$ level of significance (95% significance level), we can accept the null hypothesis that the large wave heights measured by the waverider buoy had a Gaussian (normal) distribution in time and space.

3.9 SELECTING THE SAMPLE SIZE

It is not possible to determine the required sample size N for a given confidence interval until a measure of the data variability, the population standard deviation, σ, is known. This is because the variability of $\overline{X}$ depends on the variability of X. Since we do not usually know *a priori* the population standard deviation (the value for the true population), we use the best estimate available, the sample standard deviation, s. We also need to know the frequency content of the data variable so that we can ensure that the N values we use in our calculations are statistically independent samples. As a simple example, consider a normally distributed, continuous random variable, Y, with the units of meters. We wish to find the average of the sample and want it to be accurate to within ± 5 m. Since we know that approximately 95% of the sample means will lie within $\pm 2\sigma_Y$ of the true mean μ, we require that $2\sigma_Y = 5$ m. Using the central limit theorem for the mean, we can estimate σ_Y by

$$\hat{\sigma}_Y = \frac{\sigma}{\sqrt{N}}$$

so that $2\sigma/\sqrt{N} = 5$, or $N = 4\sigma^2/25$ (assuming that the N observations are statistically independent). If σ is known, we can easily find N.

When we don't know σ, we are forced to use an estimate from an earlier sample within the range of measurements. If we know the sample range, we can apply the empirical rule for normal distributions that the range is approximately 4σ and take one-fourth the range as our estimate of σ. Suppose our range in the above example is 84 m. Then, $\sigma = 21$ m and

$$N = 4\sigma^2/25 = (4)(21\text{ m})^2/(25\text{ m}^2) = 70.56 \approx 71$$

This means that, for a sample of $N = 71$ statistically independent values, we would be 95% sure (probability = 0.95) that our estimate of the mean value would lie within $\pm 2\sigma_Y = \pm 5$ m of the true mean.

One method for selecting the sample size for relatively large samples is based on Chebyshev's theorem known as the "weak law of large numbers". Let $f(x)$ be a density function with mean μ and variance σ^2, and let $\overline{x}_N$ be the sample mean of a random sample of size N from $f(x)$. Let ε and δ be any two specified numbers satisfying $\varepsilon > 0$ and $0 < \delta < 1$. If N is any integer greater than $(\sigma^2/\varepsilon^2)\delta$ then

$$P[-\varepsilon < \overline{x}_N - \mu < \varepsilon] \geq 1 - \delta \tag{3.9.1}$$

To show the validity of condition (3.9.1), we use Thebyshev's inequality

$$P[g(x) \geq k] \geq \frac{E[g(x)]}{k} \tag{3.9.2}$$

for every $k > 0$, random variable x, and nonnegative function $g(x)$. An equivalent

formula is

$$P[g(x) < k] \geq 1 - \frac{E[g(x)]}{k} \tag{3.9.3}$$

Let $g(x) = (\bar{x}_N - \mu < \varepsilon)^2$ and $k = \varepsilon^2$, then

$$\begin{aligned} P[-\varepsilon < \bar{x}_N - \mu < \varepsilon] &= P[|\bar{x}_N - \mu| < \varepsilon] \\ &= P\left[|\bar{x}_N - \mu|^2 < \varepsilon^2\right] \geq 1 - \frac{E\left[(\bar{x}_N - \mu)^2\right]}{\varepsilon^2} \\ &= 1 - \frac{\sigma^2}{N\varepsilon^2} \geq 1 - \delta \end{aligned} \tag{3.9.4}$$

For $\delta > \sigma^2/N\varepsilon^2$ or $N > \sigma^2/\delta\varepsilon^2$, the latter expression becomes

$$P[|\bar{x}_N - \mu| < \varepsilon] \geq 1 - \delta \tag{3.9.5}$$

We illustrate the use of the above relations by considering a distribution with an unknown mean and variance $\sigma^2 = 1$. How large a sample must be taken in order that the probability will be at least 0.95 that the sample mean, $\bar{x}_N$, will lie within 0.5 of the true population mean? Given are: $\sigma^2 = 1$ and $\varepsilon = 0.5$. Rearranging the inequality (3.9.5)

$$\delta \geq 1 - P[|\bar{x}_N - \mu| < 0.5] = 1 - 0.95 = 0.05$$

Substituting into the relation $N > (\sigma^2/\delta\varepsilon^2) = 1/(0.05)(0.5)^2$ tells us that $N \geq 80$ independent samples.

3.10 CONFIDENCE INTERVALS FOR ALTIMETER BIAS ESTIMATES

As an example of how to estimate confidence limits and sample size, consider an oceanographic altimetric satellite where the altimeter is to be calibrated by repeated passes over a spot on the earth where surface-based measurements provide a precise, independent measure of the sea surface elevation. A typical reference site is an offshore oil platform having sea-level gauges and a location system, such as the multisatellite global positioning system (GPS). For the TOPEX/POSEIDON satellite one reference site was an oil platform in the Santa Barbara channel off southern California (Christensen *et al.*, 1994). Each pass over the reference site provides a measurement of the satellite altimeter bias which is used to compute an average bias after repeated calibration observations. This bias is just the difference between the height measured by the altimeter and the height measured independently by the *in situ* measurements at the reference site. If we assume that our measurement errors are normally distributed with a mean of zero, then the uncertainty of the true mean bias, σ_b, is

$$\sigma_b = z s_b / \sqrt{N}$$

where z is the standard normal distribution, s_b is the estimated standard deviation of the measurements, and N is the number of measurements (i.e. the number of calibration passes over the reference site).

Suppose we are required to know the true mean of the altimeter bias to within 2 cm, and that we estimate the uncertainty of the individual measurements to be 3 cm. We then ask: "What is the number of independent measurements required to give a bias of 2 cm at the 90%, 95%, and 99% confidence intervals?" Using the above formulation for the standard error we find

$$N = (z_{\alpha/2} s_b / \sigma_b)^2 \tag{3.10.1}$$

from which we can compute the required sample size. As before, the parameter α refers to the chosen significance level. Now $\sigma_b = 2$ cm (required) and $s_b = 3$ cm (estimated), so that we can use the standard normal table for $z_{\alpha/2} = N(0, 1)$ in the appendix to obtain the values shown in Table 3.10.1. If we require the true mean to be 1.5 cm instead of 2.0 cm, the values in Table 3.10.1 become those in Table 3.10.2.

Finally, suppose the satellite is in a 10-day repeat orbit so that we can only collect a reference measurement every 10 days at our ground site; we are given 240 days to collect reference observations. What confidence intervals can be achieved for both of the above cases if we assume that only 50% of the calibration measurements are successful and that the 10-day observations are statistically independent? We can, in general, write the confidence intervals as

$$P(-c < z < c) = \alpha, \quad \text{and } P(z < c) = (\alpha + 1)/2$$

Now, since we have only one calibration measurement every 10 days for 50% of 240 days we have

$$\begin{aligned} c &= (0.5)(240 \text{ days})\,(1 \text{ measurement}/10 \text{ days}) \\ &= 12 \text{ measurements/year} \end{aligned}$$

Referring to the above tables, we see that for the first case (Table 3.10.1), where the mean bias was required to be 2.0 cm we can achieve the 95% interval; for the case where the mean bias is restricted to 1.5 cm (Table 3.10.2), only the 90% confidence interval is possible.

Table 3.10.1. Calculation of the number of satellite altimeter observations required to attain a given level of confidence in elevation using the relation (3.10.1) for $\sigma_b = 2$ cm and $s_b = 3$ cm

Confidence level (α)	Standard normal value (z_α)	Exact number of observations (N)	Actual number of observations
90%	1.645	6.089	7
95%	1.960	8.644	9
99%	2.576	14.931	15

Table 3.10.2. Calculation of the number of satellite altimeter observations needed for a given level of confidence in sea level elevation using the equation (3.10.1) for $\sigma_b = 1.5$ cm and $s_b = 3$ cm

Confidence level (α)	Standard normal value (z_α)	Exact number of observations (N)	Actual number of observations
90%	1.645	10.82	11
95%	1.960	15.37	16
99%	2.576	26.54	27

3.11 ESTIMATION METHODS

Now that we have introduced methods to calculate confidence intervals for our estimates of μ and σ^2, we need procedures to estimate these quantities themselves. There are many different methods we could use but space does not allow us to discuss them all. We first introduce a very general technique, known as minimum variance unbiased estimation (MVUE), and then later discuss a popular method called the maximum likelihood method which leads to MVUE estimators. We will also discuss one of the oldest methods for finding point estimates, the method of moments.

Before introducing the MVUE procedure, we need to define two terms: *sufficiency* and *likelihood*. Let $x_1, x_2, \ldots, x_N$ be a random sample from a probability distribution with an unknown statistical parameter, θ (mean, variance, etc.). The statistic $U = g(x_1, x_2, \ldots, x_N)$ is said to be sufficient for θ if the conditional distribution $x_1, x_2, \ldots, x_N$, given U, does not depend on θ. In other words, once U is known, no other combination of $x_1, x_2, \ldots, x_N$ provides additional information about θ. This tells us how to check if our statistic is sufficient but does not tell how to compute the statistic.

To define likelihood, let $y_1, y_2, \ldots, y_N$ be sample observations of random variables Y_1, $Y_2, \ldots, Y_N$. For continuous variables, the likelihood $L(y_1, y_2, \ldots, y_N)$ is the joint probability density $f(y_1, y_2, \ldots, y_N)$ evaluated at the observations, y_i. Assuming that the Y_i are statistically independent

$$L(y_1, y_2, \ldots, y_N) = f(y_1, y_2, \ldots, y_N) = f(y_1)f(y_2)\ldots f(y_N) \tag{3.11.1}$$

where $f(y_i)$, $i = 1, 2, \ldots, N$, is the probability density function (PDF) for the random variable Y_i.

As an oceanographic example, consider a record of daily-average current velocities obtained using a single current meter moored for a period of one month ($N = 30$ days). Show that the monthly mean velocity, V, is a sufficient statistic for the population mean if the variance is known (in this case, estimated from the range of current values). Since the daily velocities are average values of shorter-term current velocity measurements (e.g. 30 min values), we can invoke the central limit theorem to conclude that the daily velocities are normally distributed. Hence the probability density function can be written as

$$f(V) = \frac{1}{\sigma(2\pi)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(V-\mu)^2\right]$$

We can write the likelihood L of our sample as

$$\begin{aligned} L = f(V_1, V_2, \ldots, V_{30}) &= f(V_1)f(V_2)\ldots f(V_{30}) \\ &= \frac{1}{\sigma(2\pi)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(V_1-\mu)^2\right] \\ &\quad \times \frac{1}{\sigma(2\pi)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(V_2-\mu)^2\right] \\ &\quad \ldots \frac{1}{\sigma(2\pi)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(V_{30}-\mu)^2\right] \\ &= \frac{1}{[\sigma(2\pi)]^{15}} \exp\left[-\frac{1}{2^{30}\sigma^{60}} \sum_{i=1}^{30} (V_i-\mu)^2\right] \end{aligned}$$

Because σ is known from our range of current velocities then L is a function of V and μ only. Hence, V is a sufficient statistic for μ the population mean.

3.11.1 Minimum variance unbiased estimation

For random variables $Y_1, Y_2, \ldots, Y_N$, with probability density function, $f(y)$, and unknown parameter θ, let one set of sample observations be $(x_1, x_2, \ldots, x_N)$ and another be $(y_1, y_2, \ldots, y_N)$. The ratio of the likelihoods of these two sets of observations can be written as

$$\frac{L(x_1, x_2, \ldots, x_N)}{L(y_1, y_2, \ldots, y_N)} \tag{3.11.2}$$

In general, this ratio will not be a function of θ if, and only if, there is a function $g(x_1, x_2, \ldots, x_N)$ such that $g(x_1, x_2, \ldots, x_N) = g(y_1, y_2, \ldots, y_N)$ for all choices of x and y. If such a function can be found, it is the minimum sufficient statistic for θ. Any unbiased estimator that is a function of a minimal sufficient statistic will be an MVUE; this means that it will possess the smallest possible variance among the unbiased estimators.

We illustrate what we mean with an example. Let $x_1, x_2, \ldots, x_n$ be a random sample from a normal population with the unknown parameters μ and σ^2. We want to find the MVUE of μ and σ^2. Writing the likelihood ratio we have

$$\frac{L(x_1, x_2, \ldots, x_n)}{L(y_1, y_2, \ldots, y_n)} = \frac{f(x_1, x_2, \ldots, x_n)}{f(y_1, y_2, \ldots, y_n)}$$

$$\begin{aligned}
&= \frac{\dfrac{1}{\sigma\sqrt{2\pi}} \exp\left[-\dfrac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2\right]}{\dfrac{1}{\sigma\sqrt{2\pi}} \exp\left[-\dfrac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mu)^2\right]} \\
&= \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}(x_i - \mu)^2 - \sum_{i=1}^{N}(y_i - \mu)^2\right]\right\} \\
&= \exp\left\{-\frac{1}{2\sigma^2}\left[\left(\sum_{i=1}^{N}x_i^2 - \sum_{i=1}^{N}y_i^2\right) - 2\mu\left(\sum_{i=1}^{N}x_i - \sum_{i=1}^{N}y_i\right)\right]\right\}
\end{aligned} \tag{3.11.3}$$

For this ratio to be independent of μ, we must have

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i \tag{3.11.4}$$

for the ratio to be independent of σ^2, requires both (3.11.4) as well as

$$\sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} y_i^2 \tag{3.11.5}$$

Thus, both $\sum x_i$ and $\sum x_i^2$ are minimum sufficient statistics for μ and σ^2. Since $\bar{x}$ is an unbiased estimate of μ

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N-1}\left(\sum -N\bar{x}^2\right) \tag{3.11.6}$$

is an unbiased estimate of σ^2. Since both $\bar{x}$ and s^2 are functions of the minimal sufficient statistics

$$\sum_{i=1}^{N} x_i \quad \text{and} \quad \sum_{i=1}^{N} x_i^2$$

as expressed by (3.11.4) and (3.11.5), they also are MVUEs for μ and σ^2.

3.11.2 Method of moments

As suggested earlier, the method of moments is one of the oldest methods for parameter estimation. It is simple and straightforward to apply. Recall that the kth moment of a random variable Y, taken about the origin, is

$$\mu_k' = E[Y^k] \tag{3.11.7}$$

and the corresponding sample moment is

$$m_k' = \frac{1}{N}\sum_{i=1}^{N} y_i^k \tag{3.11.8}$$

The method of moments is based on the assumption that sample moments should provide good estimates of the corresponding population moments (i.e. m_k' is a good estimate of μ_k'). Thus we choose our estimates as those parameter values which are solutions of the equations $\mu_k' = m_k'$, $k = 1, 2, \ldots, r$ where r is the number of parameters.

We again illustrate with an example. A random sample $y_1, y_2, \ldots, y_N$ is selected from a population with a uniform PDF over the interval $(0, \theta)$, where θ is unknown. We use the method of moments to estimate θ. The first moment of the population is $\mu_1' = \mu = \theta/2$ (see Appendix C). The corresponding first sample moment is

$$m' = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}$$

If we equate the moments and solve for θ

$$\hat{\theta}/2 = \bar{y}, \text{ or } \hat{\theta} = 2\bar{y}$$

Thus, θ has a moment estimate of $2\bar{y}$.

We remark, that while the method of moments is straightforward to apply, the resulting estimates are not minimal sufficient statistics. In addition, these estimates may not even be unbiased. The primary advantage of this procedure is that it often yields results when others do not.

3.11.3 Maximum likelihood

The procedure introduced earlier to compute the MVUE is complicated by the fact that one must find some function of the minimal sufficient statistic that gives the sought-after target parameter. Finding this function is, in general, a matter of trial and error. We then introduced the method of moments which, while it is easy to apply, yields estimates which may not be optimal. A more sophisticated procedure, the maximum likelihood method, often leads to the MVUE.

The formal statement of this method is quite simple. Choose as estimates those parameter values that maximize the likelihood $L(y_1, y_2, \ldots, y_N)$. A simple example using discrete variables helps to illustrate the logic in the maximum likelihood method. Assume we have a bag containing three marbles. The marbles can be black or white. We randomly sample two of the three and find that they are both black. What is the best estimate of the total number of black marbles in the bag? If there are actually two black and one white in the bag, the probability of sampling two black marbles is

$$\frac{\binom{2}{2}\binom{1}{0}}{\binom{3}{2}} = 1/3$$

where, as in Section 3.3, the binomial expression is

$$\binom{N}{r} = {}_NP_r/r! = N!/[r!(N-r)!] \tag{3.11.9}$$

and ${}_NP_r$ is the number of permutations of N discrete variables sampled r at a time. In the above expression

$$\binom{2}{2}$$

indicates the first sample of two marbles, with both being black. The next term is the remaining unsampled marble (hence the 0 in the denominator) if it were white. Now if there are three black marbles in the bag the probability of sampling two blacks is

$$\frac{\binom{3}{2}\binom{0}{0}}{\binom{3}{2}} = 1$$

On this basis, it seems reasonable to choose three as the estimate of the number of black marbles in the bag in order to maximize the probability of the observed sample.

A more complex example can be used to illustrate the application of this method to our estimates of the mean, μ, and variance, σ^2, for a normal population. Again, let y_1, $y_2, \ldots, y_N$ be a random sample from a normal population with parameters μ and σ^2. We want to find the maximum likelihood estimators of μ and σ^2. Note we used this same example for our discussion of the method of moments. To find the maximum likelihood, we need to write the joint PDF of the independent observations $y_1, y_2, \ldots, y_N$

$$
\begin{aligned}
L = f(y_1, y_2, \ldots, y_N) &= f(y_1) f(y_2) \ldots f(y_N) \\
&= \left\{ \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left[\frac{-(y_1 - \mu)^2}{2\sigma^2} \right] \right\} \\
&\quad \times \left\{ \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left[\frac{-(y_2 - \mu)^2}{2\sigma^2} \right] \right\} \\
&\quad \ldots \left\{ \frac{1}{\sigma\sqrt{(2\pi)}} \exp\left[\frac{-(y_N - \mu)^2}{2\sigma^2} \right] \right\} \\
&= \left\{ \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left[- \sum_{i=1}^{N} \frac{-(y_i - \mu)^2}{2\sigma^2} \right] \right\}
\end{aligned}
\tag{3.11.10}
$$

We simplify this expression by taking $\log_N(L)$, which we then differentiate to find the maximum. Specifically

$$
\log_N(L) = -\frac{N}{2} \log_N(\sigma^2) - \frac{N}{2} \log_N(2\pi) - \sum_{i=1}^{N} \frac{(y_i - \mu)^2}{2\sigma^2} \tag{3.11.11}
$$

Taking derivatives of (3.11.11) with respect to μ and σ^2, we find

$$
\frac{\mathrm{d}[\log_N(L)]}{\mathrm{d}\mu} = \sum_{i=1}^{N} \frac{(y_i - \mu)}{\sigma^2} \tag{3.11.12a}
$$

$$
\frac{\mathrm{d}[\log_N(L)]}{\mathrm{d}\sigma^2} = -\frac{N}{\sigma^2} + \sum_{i=1}^{N} \frac{(y_i - \mu)}{2\sigma^4} \tag{3.11.12b}
$$

Setting (3.11.12a, b) to zero and solving yields the required estimates of μ and σ^2

$$
\sum_{i=1}^{N} \frac{(y_i - s)}{\sigma^2} = 0 \quad \text{or} \quad \mu = \frac{1}{N} \sum_{i=1}^{N} y_i = \bar{y} \tag{3.11.13}
$$

Substituting $\bar{y}$ into (3.11.12b)

$$
-N/\hat{\sigma}^2 + \sum_{i=1}^{N} (y_i - \bar{y})^2 / \hat{\sigma}^4 = 0 \tag{3.11.14a}
$$

or

$$
\hat{\sigma}^2 = \sum_{i=1}^{N} (y_i - \bar{y})^2 / N = s'^2 \tag{3.11.14b}
$$

Thus, $\bar{y}$ and s'^2 are the maximum likelihood estimators of μ and σ^2. Although, $\bar{y}$ is an unbiased estimate of μ, s'^2 is not unbiased for σ^2, as noted at the beginning of the chapter. However, s'^2 can easily be adjusted to the unbiased estimator

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2 \tag{3.11.15}$$

Since the Maximum Likelihood Method has widespread application, we present another simple example to illustrate its use. Let $y_1, y_2, \ldots, y_N$, be a random sample taken from a uniform distribution with $f(y_i) = 1/\theta = \text{constant}$, $0 \leq y_i \leq \theta$, and $i = 1, 2, \ldots, N$. We want to find the maximum likelihood estimate of θ. Again, we write the likelihood, L, as the joint probability function

$$\begin{aligned} L &= f(y_1, y_2, \ldots, y_N) = f(y_1)f(y_2)\ldots f(y_N) = (1/\theta)(1/\theta)\ldots(1/\theta) \\ &= (1/\theta)^N \end{aligned} \tag{3.11.16}$$

In this case, L is a monotonically decreasing function of θ and nowhere is $dL/d\theta = 0$. Instead, L increases monotonically as θ decreases and must be greater than or equal to the largest sample value, y_N. L is, therefore, not an unbiased estimate of θ. It can be adjusted to

$$\theta = \frac{(N+1)}{N} y_N \tag{3.11.17}$$

which is unbiased. We note that if any statistic U can be shown to be a sufficient statistic for estimating θ then the maximum likelihood estimator is always some function of U. If this maximum likelihood estimate can be found, and then adjusted to be unbiased, the result will generally be an MVUE.

To demonstrate the application of the maximum likelihood approach, assume that a random sample of size N is selected from the normal distribution (equation (3.5.2)) with μ and σ^2 as the mean and variance for each x_i (where we assume that the x_i values are independent). We ask: If $\bar{\theta} = (\theta_1, \theta_2) = (\mu, \sigma^2)$ is the parameter space for the probability density function $f(x_1, x_2, \ldots, x_N)$, then what is the likelihood function? Also, find the maximum likelihood estimator $\hat{\theta}_1$ of θ_1 which maximizes the likelihood function and find the maximum likelihood estimator $\hat{\theta}_2$ which maximizes the likelihood function θ_1. We first write the PDF as

$$\begin{aligned} f(\bar{x}, \bar{\theta}) &= (\tfrac{1}{2}\pi\sigma^2)^{N/2} = \exp\left[\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2\right] \\ &= \prod_{i=1}^{N}\left\{\frac{1}{\sqrt{(2\pi\sigma)}}\exp\left[-(x_i - \mu)^2/\sigma^2\right]\right\} = L(\bar{x}, \bar{\theta}) \end{aligned}$$

which is the likelihood function written in terms of the product, Π, of the exponential. Taking the natural log of the above expression with respect to our estimated parameter, θ_1, and setting it equal to zero to find the maximum, we get

$$\ln(L) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$$

where $\sigma > 0$ and $-\infty < \mu < \infty$. The derivative of this function with respect to θ_1 (which is μ) is

$$\frac{\partial L}{\partial \mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)(-2) = \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i - \mu) = 0$$

so that our estimate of μ is

$$\hat{\mu} = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Furthermore, the maximum likelihood estimator of θ_2 (which is σ^2) is given by

$$\frac{\partial L}{\partial \sigma^2} = -\frac{N}{2\sigma^2} - \frac{(-1)}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{2\sigma^2}\left[\frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2 - N\right] = 0$$

which yields the estimator

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$$

For a normally distributed oceanographic data set, we can readily obtain maximum likelihood estimates of the mean and variance of the data. However, the real value of this technique is for variables that are not normally distributed. For example, if we examine spectral energy computed from current velocities, the spectral values have a chi-square distribution rather than a normal distribution. If we follow the maximum likelihood procedure, we find that the spectral values have a mean of ν, the number of degrees of freedom, and a variance of 2ν. These are the maximum likelihood estimators for the mean and variance. This example can be used as a pattern for applying the maximum likelihood method to a particular sample. In particular, we first determine the appropriate PDF for the sample values. We then find the joint likelihood function, take the natural logs and then differentiate with respect to the parameter of interest. Setting this derivative equal to zero to find the maximum subsequently yields the value of the parameter being sought.

3.12 LINEAR ESTIMATION (REGRESSION)

Linear regression is one of a number of statistical procedures that fall under the general heading of linear estimation. Since linear regression is widely treated in the literature and is available in many software packages, our primary purpose here is to establish a common vocabulary for all readers. In our previous discussion and examples, we assumed that the random variables $Y_1, Y_2, \ldots, Y_N$ were independent (in a probabilistic sense) and identically distributed, which implies that $E[Y_i] = \mu$ is a constant. Often this is not the case and the expected value $E[Y_i]$ of the variable is a function of some other parameter. We now consider the values y of a random variable, Y, called the dependent variable, whose values are a function of one or more *nonrandom* variables $x_1, x_2, \ldots, x_N$, called independent variables (in a mathematical, rather than probabilistic sense).

If we model our random variable as

$$y = E[y] + \varepsilon = b_0 + b_1 x + \varepsilon \tag{3.12.1}$$

we invariably find that the points y are scattered about the regression line $E[y] = b_0 + b_1 x$. The random variable ε in the right-hand term of (3.12.1) gives the departure from linearity and has a specific PDF with a mean value $\mu_\varepsilon = 0$. In other words, we can think of y as having a deterministic part, $E[y]$, and a random part, ε, that is randomly distributed about the regression line. By definition, simple linear regression is limited to finding the coefficients b_0 and b_1. If N independent variables $(x_1, x_2, \ldots, x_N)$ are involved in the variability of each value y, we must deal with *multiple linear regression*. In this case, (3.12.1) becomes

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_N x_N + \varepsilon \tag{3.12.2}$$

3.12.1 Method of least squares

One of the most powerful techniques for fitting a dependent model parameter y to independent (observed input) variables x_i $(i = 1, 2, \ldots, N)$ is the *method of least squares*. We apply the method in terms of linear estimation and will later readdress the topic in terms of more general statistical models. (Note: by "linear" we mean linear in the parameters $b_0, b_1, \ldots, b_N$. Thus, $y = b_0 + b_1 x_1^2 + \varepsilon$ is linear but $y = b_0 + \sin(b_1 x_1) + \varepsilon$ is not.) We begin with the simplest case, that of fitting a straight line to a set of points using the "best" coefficients b_0, b_1 (Figure 3.10). In a sense, the least squares procedure does what we do by eye—it minimizes the vertical deviations (residuals) of data points from the fitted line. Let

$$y_i = \hat{y}_i + \varepsilon_i \tag{3.12.3}$$

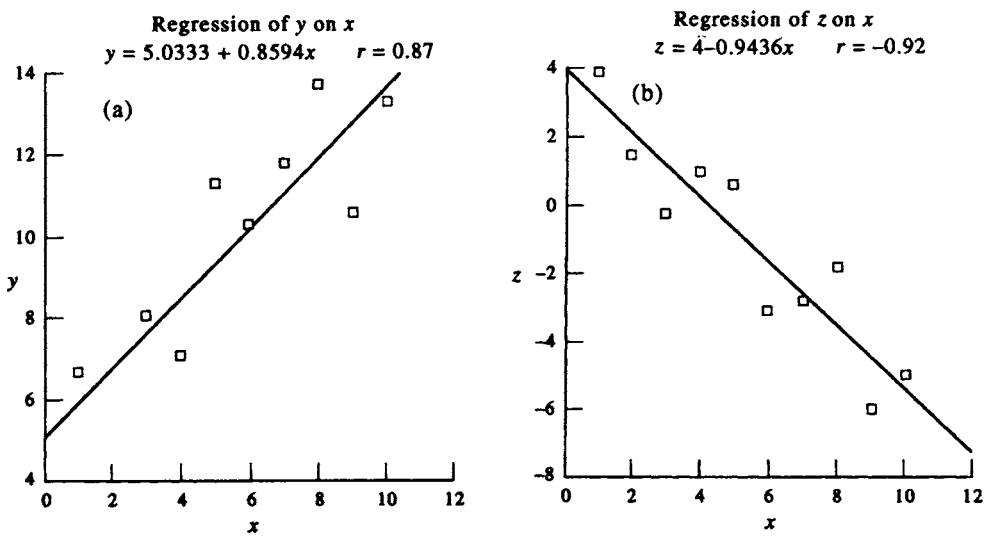

Figure 3.10. Straight line (linear regression) fits to the sets of points in Table 3.12.1 using the "best" coefficients b_0, b_1. (a) Regression of y on x, for which $(b_0, b_1) = (5.0333, 0.8594)$; and (b) regression of z on x, for which $(b_0, b_1) = (4.0, -0.9436)$. r is the correlation coefficient.

where

$$\hat{y}_i = b_0 + b_1 x_i \tag{3.12.4}$$

is our estimator for the deterministic portion of the data and ε is the residual or error. To find the coefficients b_0, b_1 we need to minimize the sum of the squared errors (SSE) where SSE is the total variance that is not explained (accounted for) by our linear regression model given by (3.12.3) and (3.12.4)

$$\text{SSE} = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} [y_i - (b_0 + b_1 x_i)]^2 \tag{3.12.5a}$$

$$= \text{SST} - \text{SSR} \tag{3.12.5b}$$

in which

$$\text{SST} = \sum_{i=1}^{N} (y_i - \bar{y})^2 \quad \text{and} \quad \text{SSR} = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2 \tag{3.12.5c}$$

Here, SST (sum of squares total) is the variance in the data and SSR (sum of squares regression) is the amount of variance explained by our regression model. Minimization amounts to finding those coefficients that minimize the unexplained variance (SSE). Taking the partial derivatives of (3.12.5a) with respect to b_0 and b_1 and setting the resultant values equal to zero, the minimization conditions are

$$\frac{\partial \text{SSE}}{\partial b_0} = 0; \quad \frac{\partial \text{SSE}}{\partial b_1} = 0 \tag{3.12.6}$$

Substituting (3.12.5a) into (3.12.6), we have for b_0

$$\begin{aligned} \frac{\partial \text{SSE}}{\partial b_0} &= \frac{\partial}{\partial b_0} \left\{ \sum_{i=1}^{N} [y_i - (b_0 + b_1 x_i)]^2 \right\} = -2 \sum_{i=1}^{N} [y_i - (b_0 + b_1 x_i)] \\ &= -2 \left(\sum_{i=1}^{N} y_i - N b_0 - b_1 \sum_{i=1}^{N} x_i \right) = 0 \end{aligned} \tag{3.12.6a}$$

Now for b_1

$$\begin{aligned} \frac{\partial \text{SSE}}{\partial b_1} &= \frac{\partial}{\partial b_1} \left\{ \sum_{i=1}^{N} [y_i - (b_0 + b_1 x_i)]^2 \right\} = -2 \sum_{i=1}^{N} [y_i - (b_0 + b_1 x_i] \\ &= -2 \left(\sum_{i=1}^{N} x_i y_i - b_0 \sum_{i=1}^{N} x_i - b_1 \sum_{i=1}^{N} x_i^2 \right) = 0 \end{aligned} \tag{3.12.6b}$$

Once the mean values of y and x are calculated, these least squares equations can be solved simultaneously to find an estimate of the coefficient b_1 (the slope of the regression line); this is then used to obtain an estimate of the second coefficient, b_0 (the intercept of the regression line). In particular

$$\hat{b}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{\left[N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i\right]}{\left[N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2\right]} \tag{3.12.7a}$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x} \tag{3.12.7b}$$

Several features of the regression values are worth noting. First, if we substitute the intercept $b_0 = \bar{y} - b_1\bar{x}$ into the line $\hat{y} = b_0 + b_1 x$, we obtain

$$\hat{y} = \bar{y} + b_1(x - \bar{x})$$

As a result, whenever $x = \bar{x}$, we have $\hat{y} = \bar{y}$. This means: (1) That the regression line always passes through the point $(\bar{x}, \bar{y})$; and (2) that because the operation $\partial \mathrm{SSE}/\partial b_0 = 0$ minimizes the error $\sum \varepsilon_i = 0$, the regression line not only goes through the point of averages $(\bar{x}, \bar{y})$ but it also splits the scatter of the observed points so that the positive residuals (where the regression line passes below the true point) always cancel exactly the negative residuals (where the line passed above the true point). The sample regression line is therefore an unbiased estimate of the population regression line.

To summarize the linear regression procedure, we note that:

(1) For each selected x (independent variable) there is a distribution of y from which the sample (dependent variable) is drawn at random.
(2) The population of y corresponding to a selected x has a mean μ that lies on the straight line $\mu = b_0 + b_1 x$, where b_0 and b_1 are regression parameters.
(3) In each population, the standard deviation of y about its mean, $b_0 + b_1 x$, has the same value ($s_{xy} = s_\varepsilon, y = b_0 + b_1 x + \varepsilon$). Note that ε is a random variable drawn from a normal population with $\mu = 0$ and $s = s_{xy}$.

Table 3.12.1. Values for dependent variables y_i, z_i as functions of x_i. The estimated values $\hat{y}$ and $\hat{z}$ are derived from the linear regression analysis. Formulae at the bottom of the table are the total sum of squares (SST), sum of squares for the regression (SSR) and the sum of squares of the errors (SSE) to be derived in our regression analysis for $N = 10$

x_i	y_i	$\hat{y}_i$	z_i	$\hat{y}_i$
1.0	6.7	5.9	3.9	3.1
2.0	4.7	6.8	1.5	2.1
3.0	8.1	7.6	−0.2	1.2
4.0	7.1	8.5	1.0	0.2
5.0	11.3	9.4	0.6	−0.7
6.0	10.5	10.2	−3.1	−1.7
7.0	11.8	11.1	−2.8	−2.6
8.0	13.7	11.9	−1.8	−3.6
9.0	10.6	12.8	−6.0	−4.5
10.0	13.3	13.7	−5.0	−5.4

SST(y) = 80.64; SSR(y) = 61.11; SSE(y) = 19.53
SST(z) = 86.39; SSR(z) = 73.46; SSE(z) = 12.93

Thus, y is the sum of a random part ε and a fixed part x; the fixed part determines the mean values of the y population samples, with one distribution of y for each x that we pick. The mean values of y lie on the straight line, $\mu = b_0 + b_1 x$, which is the population regression line. The regression parameter b_0 is the y mean for $x = 0$ and b_1 is the slope of the regression line. The random part, ε, is independent of x and y. To compute the regression parameters, we need values of $N, \bar{x}, \bar{y}, \sum x^2, \sum y^2$, and $\sum xy$. Earlier, we discussed the computational shortcuts to compute $\sum x^2$ and $\sum y^2$ without first computing the means of x and y. The same can be accomplished for xy using,

$$\sum (x - \bar{x})(y - \bar{y}) = \sum xy - \sum x \sum y / N$$

As examples of linear regression, consider the data sets in Table 3.12.1 for dependent variables y_i and z_i which are both functions of the same independent variable x_i (for example, y_i could be the eastward and z_i the northward component of velocity as functions of time x_i). We will compute the regression coefficients b_0, b_1 plus the sample variance s^2 and percent of explained variance (100 SSR/SST) for each data set.

To estimate the regression parameters, we must first compute the means of the three series

$$\bar{x} = 5.50; \quad \bar{y} = 9.78; \quad \bar{z} = -1.19$$

We then use the means to calculate the sums in (3.12.6)

$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 82.50; \quad \sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y}) = 71.00;$$

$$\sum_{i=1}^{10} (x_i - \bar{x})(z_i - \bar{z}) = -77.85$$

$$SST(y) = \sum_{i=1}^{10} (y_i - \bar{y})^2 = 80.64$$

$$SST(z) = \sum_{i=1}^{10} (z_i - \bar{z})^2 = 86.36$$

For the regression of y on x ($\hat{y} = b_0 + b_1 x$), we find

$$b_0 = 5.05; b_1 = 0.861; s^2 = 2.44$$
$$100 \cdot SSR(y)/SST(y) = (100)61.11/80.64 = 75.8\%$$

while for the regression of z on x ($\hat{z} = b_0 + b_1 x$), we have

$$b_0 = 4.00; \quad b_1 = -0.94; \quad s^2 = 1.62$$
$$100 \cdot SSR(z)/SST(z) = (100)73.46/86.36 = 85.0\%$$

The ratio SSR/SST (variance explained/total variance) is a meaure of the goodness of fit of the regression curves called the *correlation of determination*, r^2. If the regression line fits perfectly all the sample values, all residuals would be zero. In turn, SSE = 0 and SSR/SST = r^2 = 1. As the fit becomes increasingly less representative of the data points, r^2 decreases towards a possible minimum of zero.

3.12.2 Standard error of the estimate

The measure of the absolute magnitude of the goodness of fit is the standard error of the estimate, s_ε, defined as

$$s_\varepsilon = [\mathrm{SSE}/(N-2)]^{1/2} = \left[\frac{1}{N-2}\sum_{i=1}^{N}(y-\hat{y})^2\right]^{1/2} \tag{3.12.8}$$

The number of degrees of freedom, $N-2$, for s_ε is based on the fact that two parameters, b_0 and b_1 are needed for any linear regression estimate. If s_ε is from a normal distribution and has a mean of zero, then, in analogy with our discussion of the standard deviation of values about their mean, approximately 68.3% of the observations will fall within $\pm 1 s_\varepsilon$ units of the regression line, 95.4% will fall within $\pm 2 s_\varepsilon$ units of the line and 99.7% will fall within $\pm 3 s_\varepsilon$ units of this line. For the examples of Table 3.12.1

$$\text{Variable } y\text{: } s_\varepsilon = [\mathrm{SSE}(y)/(N-2)]^{1/2} = (19.53/8)^{1/2} = 1.56$$
$$\text{Variable } z\text{: } s_\varepsilon = [\mathrm{SSE}(z)/(N-2)]^{1/2} = (12.93/8)^{1/2} = 1.27$$

As a result, the $\pm 2 s_\varepsilon$ ranges are $\pm 2(1.56)$ and $\pm 2(1.27)$, respectively.

We next turn to our estimate of the slope, b_1. Recalling that $b_1 = \sum x / \sum x^2$, we find

$$s_{b1}^2 = \frac{s_{xy}^2}{\sum x^2} = \frac{s_\varepsilon^2}{\sum x^2} \tag{3.12.9}$$

where $s_{b1}{}^2$ is the sample variance for our estimate, $\hat{b}_1$, for the slope of the regression line. For small samples ($N < 30$), we can write the 95% confidence interval as

$$\hat{b}_1 - t_{0.05} s_{b1} \leq b_1 \leq \hat{b}_1 + t_{0.05} s_{b1} \tag{3.12.10}$$

Turning to the regression line itself, we wish to say something about the standard deviation about $\bar{y}$ (i.e. the regression line). First we note that $\hat{\varepsilon}$ has variance σ_{xy}^2/N and $\hat{b}_1$ has variance $\sigma_{xy}^2/\sum x^2$. Since the errors, ε, are assumed independent, the variance of the sums is the sum of the variances

$$\sigma_y^2 = \sigma_{xy}^2\left[1/(N-2) + x^2/\sum x^2\right] \tag{3.12.11}$$

which leads to the standard error given above. These confidence limits would appear as hyperbolae in regression diagrams such as Figure 3.10. The hyperbolae are the confidence belts for the different significance levels. Note the increasing hazard of making predictions for values of x far removed from the mean value $\bar{x}$. Since the lines indicate that y must be within the confidence belt, higher signifance levels have narrower belts. Thus, estimates of $\bar{y}$ get worse as we move away from $\bar{x}, \bar{y}$. Remember that these confidence belts are for the regression line itself and not for the individual points. Hence, if repeated samples of y_i are taken of the same size and the same fixed value of x, the 95% of the confidence intervals, constructed for the mean value of y and x, will contain the true value of the mean of y and x. If only one prediction is made of x, then the probability that the calculated interval of this will contain the true value is 95%.

3.12.3 Multivariate regression

To extend the regression procedure to multivariate regression, we must formulate our linear estimation model in matrix terms. Suppose our model is of the form

$$Y = b_0 + b_1X_1 + b_2X_2 + \ \ldots\ + b_kX_k + \varepsilon \tag{3.12.12}$$

and that we make N independent (probabilistic) observations $y_1, y_2, \ldots, y_N$ of Y. This means that we can write

$$y_i = b_0 + b_1x_{i1} + b_2x_{i2} + \ \ldots\ + b_kx_{ik} + \varepsilon_i \tag{3.12.13}$$

where x_{ij} is the jth independent variable for the ith observation. Writing this in matrix form we have

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ \ldots \\ y_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_0 & x_{11} & \ldots & x_{1k} \\ x_0 & x_{21} & \ldots & x_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ x_0 & x_{N1} & \ldots & x_{Nk} \end{pmatrix}$$
$$\mathbf{B} = \begin{pmatrix} b_0 \\ b_1 \\ \ldots \\ \ldots \\ b_k \end{pmatrix}, \quad \mathbf{E} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \ldots \\ \varepsilon_N \end{pmatrix} \tag{3.12.14}$$

where the boldface letters indicate matrices. Using (3.12.14), we can represent the N equations relating y_i to the independent variable x_{ij} as

$$\mathbf{Y} = \mathbf{B} \cdot \mathbf{X} + \mathbf{E} \tag{3.12.15}$$

If we restrict our analysis to the first two coefficients, (3.12.15) reduces to the simple straight line fit model (3.12.4). In this case, the matrices for N observations become

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ \ldots \\ y_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_0 & \ldots & x_{1N} \\ x_0 & \ldots & x_{2N} \\ x_0 & \ldots & \ldots \\ x_0 & \ldots & \ldots \\ x_0 & \ldots & x_{NN} \end{pmatrix}$$
$$\mathbf{B} = \begin{pmatrix} b_o \\ b_1 \end{pmatrix}, \quad \mathbf{E} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \ldots \\ \varepsilon_N \end{pmatrix} \tag{3.12.16}$$

Using these N observations in (3.12.15), our least squares equations are

$$\begin{aligned} Nb_0 + b_1 \sum_{i=1}^{N} x_i &= \sum_{i=1}^{N} y_i \\ b_0 \sum_{i=1}^{N} x_i + b_1 \sum_{i=1}^{N} x_i^2 &= \sum_{i=1}^{N} x_i y_i \end{aligned} \tag{3.12.17}$$

which we can solve for b_0 and b_1. We can generalize the procedure further by realizing that for $x_0 = 1$ we have

$$\mathbf{x}' \cdot \mathbf{x} = \begin{pmatrix} 1 \ldots & \ldots & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots \\ x_1 & \ldots & \ldots & x_N \end{pmatrix} \begin{pmatrix} 1 & x_i \\ \ldots & \ldots \\ 1 & x_N \end{pmatrix} = \begin{pmatrix} N & \sum x_i \\ \ldots & \ldots \\ \sum x_i & \sum x_i^2 \end{pmatrix} \tag{3.12.18}$$

where $\mathbf{X}'$ is the transpose of the matrix $\mathbf{X}$ and, the sums are from 1 to N, and

$$\mathbf{X}' \cdot \mathbf{Y} = \begin{pmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \end{pmatrix} \tag{3.12.19}$$

Our least squares equations can then be expressed as

$$(\mathbf{X}'\mathbf{X}) \cdot \mathbf{B} = \mathbf{X}' \cdot \mathbf{Y} \tag{3.12.20}$$

where

$$B = \begin{pmatrix} b_0 \\ \\ b_1 \end{pmatrix} \tag{3.12.21}$$

Solving the above equations for $\mathbf{B}$, we obtain

$$\mathbf{B} = (\mathbf{X}' \cdot \mathbf{X})^{-1} \mathbf{X}' \cdot \mathbf{Y} \tag{3.12.22}$$

3.12.4 A computational example of matrix regression

Since linear regression is widely used in oceanography, we will illustrate its use by a simple example. Suppose we want to fit a line to the data pairs consisting of the independent variable x_i and the dependent variable y_i given in Table 3.12.2.

From these we find

$$\sum_{i=1}^{N} x_i = 0, \quad \sum_{i=1}^{N} y_i = 5, \quad \sum_{i=1}^{N} x_i y_i = 7, \quad \sum_{i=1}^{N} x_i^2 = 10$$

Substituting into equation (3.12.14), we have

Table 3.12.2. Data values used in least squares linear fit of a two-coefficient regression model, $y_i = F(x_i)$

Data		Solution values	
x_i	y_i	$(x_i)(y_i)$	x_i^2
−2	0	0	4
−1	0	0	1
0	1	0	0
1	1	1	1
2	3	6	4

$$b_1 = \frac{\left[N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i\right]}{\left[N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2\right]} = \frac{[(5)(7) - (0)(5)]}{[(5)(10) - 10^2]} = 0.7$$

$$b_0 = \bar{y} - b_1\bar{x} = 5/5 - (0.7)(0) = 1$$

This same problem can be put in matrix form

$$\mathbf{Y} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 3 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}$$

$$\mathbf{X}' \cdot \mathbf{X} = \begin{pmatrix} 5 & 0 \\ 0 & 10 \end{pmatrix}, \quad \mathbf{X}' \cdot \mathbf{Y} = \begin{pmatrix} 5 \\ 7 \end{pmatrix}, \quad (\mathbf{X}' \cdot \mathbf{X})^{-1} = \begin{pmatrix} 1/5 & 0 \\ 0 & 1/10 \end{pmatrix}$$

$$\mathbf{B} = (\mathbf{X}' \cdot \mathbf{X})^{-1}\mathbf{X}' \cdot \mathbf{Y} = \begin{pmatrix} 1/5 & 0 \\ 0 & 1/10 \end{pmatrix} \begin{pmatrix} 5 \\ 7 \end{pmatrix} \begin{pmatrix} 1 \\ 0.7 \end{pmatrix}$$

so that by (3.12.21), $b_0 = 1$ and $b_1 = 0.7$.

An important property of the simple straight line least-square estimators we have just derived is that b_0 and b_1 are unbiased estimates of their true parameter values. We have assumed that $E[\varepsilon] = 0$ and that $V[\varepsilon] = \sigma^2$; thus the error variance is independent of x and $V[Y] = V[\varepsilon] = \sigma^2$. Since σ^2 is usually unknown we estimate it using the sample variance (3.2.4) given by

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N} (y_i - \bar{y})^2 \tag{3.12.23}$$

However, if we use the output values, $\hat{y}_i$, from the least squares, to estimate $\varepsilon_i(Y) = y_i - \hat{y}_i$, we must write (3.12.23) as

$$s^2 = \frac{1}{N-2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \frac{1}{N-2}\text{SSE} \tag{3.12.24}$$

where SSE, given by (3.12.23), represents the sum of the squares of the errors and the $N-2$ corresponds to the fact that two parameters, b_0 and b_1, are needed in the model.

In matrix notation we can write the SSE as

$$\mathbf{SSE} = \mathbf{Y}' \cdot \mathbf{Y} - (\mathbf{B}' \cdot \mathbf{X}') \cdot \mathbf{Y} \tag{3.12.25}$$

Using this with our previous numerical example we write (3.12.25) as

$$\begin{pmatrix} 0 & 0 & 1 & 1 & 3 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 3 \end{pmatrix} - \begin{pmatrix} 1 & 0.7 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 3 \end{pmatrix}$$

$$= 11 - \begin{pmatrix} 1 & 0.7 \end{pmatrix} \begin{pmatrix} 5 \\ 7 \end{pmatrix} = 11 - 9.9 = 1.1$$

Since $s^2 = \text{SSE}/(N-2)$, we have $s^2 = 1.1/(3) = 0.367$ as our estimator of σ^2.

3.12.5 Polynomial curve fitting with least squares

The use of least-squares fitting is not limited to the straight line regression model discussed thus far. In general, we can write our linear model as any polynomial of the form

$$Y = b_0 + b_1 x + b_2 x^2 + \ldots + b_N x^N + \varepsilon \tag{3.12.26}$$

The procedure is the same as with the straight line case except that now the **X** matrix has $N+1$ columns. Thus, our least-squares fit will have $N+1$ linear equations with $N+1$ unknowns, $b_0, b_1, \ldots, b_N$. These equations are called the *normal equations*.

3.12.6 Relationship between least-squares and maximum likelihood

As discussed earlier, the maximum likelihood estimator is one that maximizes the likelihood of sampling a given parameter. In general, if we have a sample x_i from a population with the PDF $f(X_i, \theta)$, where θ is the parameter of interest, the maximum likelihood estimator $L(\theta)$ is the product of the individual independent probabilities

$$L(\theta) = f(x_1, \theta) f(x_2, \theta) \ldots f(x_N, \theta) \tag{3.12.27}$$

If the errors all come from a normal distribution, this becomes from equation (3.11.10)

$$L(\theta) = \frac{\exp\left[-\sum_{i=1}^{N}(x_i - \theta)^2/2\sigma^2\right]}{\sigma^N(2\pi)^{N/2}} \tag{3.12.28}$$

When this is maximized, it leads to the least-squares estimate

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x}$$

In other words, the least-squares estimate of the mean of θ can be derived from a normal distribution using the maximum likelihood criterion. This value is found to be the average of the independent variable x.

3.13 RELATIONSHIP BETWEEN REGRESSION AND CORRELATION

The subject of correlation will be considered in more detail when we examine time-series analysis methods. Our intension, here, is simply to introduce the concept in general statistical terms and relate it to the simple regression model just discussed. As with regression, correlation relates two variables but unlike regression it is measured without estimation of the population regression line.

The *correlation coefficient*, r, is a way of determining how well two (or more) variables co-vary in time or space. For two random variables x $(x_1, x_2, \ldots, x_N)$ and y $(y_1, y_2, \ldots, y_N)$ the correlation coefficient can be written

$$\begin{aligned} r &= \frac{1}{N-1}\sum_{i=1}^{N}\frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \\ &= C_{xy}/s_x s_y \end{aligned} \tag{3.13.1a}$$

where

$$C_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) \tag{3.13.1b}$$

is the *covariance* of x and y, and s_x and s_y are the standard deviations for the two data records as defined by equation (3.2.4). We note two important properties of r:

(1) r is a dimensionless quantity since the units of the numerator and the denominator are the same;
(2) the value of r lies between -1 and $+1$ since it is normalized by the product of the standard deviations of both variables.

For $r = \pm 1$, the data points (x_i, y_i) cluster along a straight line and the samples are said to have a perfect correlation (+ for "in-phase" fluctuations and minus (−) for 180° "out-of-phase" fluctuations). For $r \approx 0$, the points are scattered randomly on the graph and there is little or no relationship between the variables. The variables x_i, y_i in

equation (3.13.1a, b) could be samples from two different, independent random variables or they could represent the independent (input) and dependent (output) variables of an estimation model. Alternatively, they could be samples from the same variable. Known as an *auto-correlation*, the later is usually computed for increasing lag or shifts in the starting value for one of the time series. A lag of "m" means that the first m values of one of the series, say the x series, are removed so that x_{m+1} becomes the new x_1 and so on.

Some authors prefer to use r^2 (the coefficient of determination discussed in Section 3.12.1 in the context of straight-line regression) rather than r (the correlation coefficient) since the squared value can be used to construct a significance level for r^2 in terms of a hypothesis test that the true correlation squared is zero. Writing

$$C_{xy}^2/(s_x s_y)^2 = \text{SSR/SST} = r^2 \tag{3.13.2}$$

we see that r^2 = variance explained/total variance, as stated earlier. A value $r = 0.75$ means that a linear regression of y on x explains $r^2 = 56.25\%$ of the total sample variance. Our approach is to use r to get the sign of the correlation and r^2 to estimate the joint variances.

3.13.1 The effects of random errors on correlation

Before discussing the relationship between r and our simple regression model, it is important to realize that sampling errors in x_i and y_i can only cause r to decrease. This can be shown by writing our two variables as a combination of true values (α_i, β_i) and random errors (δ_i, ε_i). In particular

$$\begin{aligned} x_i &= \alpha_i + \delta_i \\ y_i &= \beta_i + \varepsilon_i \end{aligned} \tag{3.13.3}$$

Using equations (3.13.2) and (3.13.3), we can write the correlation between x_i and y_i as

$$r_{xy} = \frac{s_\alpha s_\beta r_{\alpha\beta} + s_\beta s_\delta r_{\beta\delta} + s_\alpha s_\varepsilon r_{\alpha\varepsilon} + s_\delta s_\varepsilon r_{\delta\varepsilon}}{s_x s_y} \tag{3.13.4}$$

where for convenience we have dropped the index i. Since the random errors δ and ε are assumed to be independent of each other and of the variables α and, β we know that

$$r_{\beta\delta} = r_{\alpha\varepsilon} = r_{\delta\varepsilon} = 0$$

so that (3.13.4) becomes

$$r_{xy} = \frac{s_\alpha s_\beta}{s_x s_y} r_{\alpha\beta} \tag{3.13.5}$$

This result means that the ratio between the product of the true standard deviations (s_α, s_β) to the product of the measured variable (s_x, s_y) determines the magnitude of the computed correlation coefficient (r_{xy}) relative to the true value ($r_{\alpha\beta}$).

To determine (3.13.5), we expand the variances of x and y as

$$\left(s_x^2, s_y^2\right) = \frac{1}{N-1}\sum_{i=1}^{N}\left[(x_i-\bar{x})^2, (y_i-\bar{y})^2\right]$$

where, as usual, $\bar{x}, \bar{y}$ are the average values for samples x_i, y_i respectively. Expanding the numerator into its component terms through (3.13.3), and using the fact that the errors are independent of one another, and of x and y, yields

$$\sum_{i=1}^{N}(x_i-\bar{x})^2 = \sum_{i=1}^{N}\left[(\alpha_i-\bar{\alpha})^2+\delta_i^2\right]$$

$$\sum_{i=1}^{N}(y_i-\bar{y})^2 = \sum_{i=1}^{N}\left[(\beta_i-\bar{\beta})^2+\varepsilon_i^2\right]$$

Dividing through by $(N-1)$ and using the definitions for standard devation, we find

$$s_x^2 = s_\alpha^2 + \frac{\sum_{i=1}^{N}\delta_i^2}{N-1}; \quad s_y^2 = s_\beta^2 + \frac{\sum_{i=1}^{N}\varepsilon_i^2}{N-1} \tag{3.13.6}$$

Since the second terms in each of the above expressions can never be negative ($N > 1$), the observed variances s_x^2 and s_y^2 are always greater than the corresponding true variances. Applying this result to equation (3.13.5), we see that the calculated correlation, r_{xy}, derived from the observations is always smaller than the true correlation, $r_{\alpha\beta}$. Because of random errors, the correlation coefficient computed from the observations will be smaller than (or, at best, equal to) the true correlation coefficient.

3.13.2 The maximum likelihood correlation estimator

Returning to the relationship between correlation and regression, we note the maximum likelihood estimator of the correlation coefficient is, by (3.13.1a)

$$r = \left[\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})\right] \Big/ \left[\sum_{i=1}^{N}(x_i-\bar{x})^2\sum_{i=1}^{N}(y_i-\bar{y})^2\right]^{1/2} \tag{3.13.7}$$

for a bivariate normal population (x_i, y_i). We can expand this using (3.13.1b) to get

$$r = \frac{\left[N\sum_{i=1}^{N}x_iy_i - \sum_{i=1}^{N}x_i\sum_{i=1}^{N}y_i\right]}{\left\{\left[N\sum_{i=1}^{N}x_i^2 - \left(\sum_{i=1}^{N}x_i\right)^2\right]\left[N\sum_{i=1}^{N}y_i^2 - \left(\sum_{i=1}^{N}y_i\right)^2\right]\right\}^{1/2}} \tag{3.13.8}$$

Note that the numerator in equation (3.13.8) is similar to the numerator of the estimator for b_1 in equation (3.12.7a). For the case where the regression line passes through the origin in (3.12.7b), we have $b_1 = 0$ and our model is

$$\hat{y}_i = \hat{b}_1 x_i$$

and we can rewrite (3.12.7a) as

$$\hat{b}_1 = \frac{\left[\sum_{i=1}^{N} x_i y_i\right]\left[\sum_{i=1}^{N} y_i^2\right]^{1/2}}{\left[\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i^2\right]^{1/2}\left[\sum_{i=1}^{N} x_i^2\right]^{1/2}} \tag{3.13.9}$$
$$= r s_y / s_x; \text{ or, } r = \hat{b}_1 s_x / s_y$$

Thus, r can be computed from $\hat{b}_1$ and vice versa if the standard deviations of the sample values x and y are known. Also, using the relationship between $\hat{b}_1$ and r we can write the variance of the parameter estimate in equation (3.13.9) as

$$s^2 = \frac{1}{N-2}\sum_{i=1}^{N}(y_i - \bar{y})^2 = \frac{1}{N-2}\text{SSE} \tag{3.13.10}$$

We can use this result to better understand the relationship between correlation and regression by writing the ratio of the regression variance in equation (3.13.10) to the sample variance in y alone; for large N, this becomes

$$\frac{s^2}{s_y^2} = \frac{(N-1)(1-r^2)}{N-2} \approx (1-r^2) \tag{3.13.11}$$

Thus, for N large, r^2 is that portion of the variance of y that can be attributed to its regression on x while $(1-r^2)$ is that portion of y's variance that is independent of x.

Earlier it was noted that a computationally efficient way to calculate the variance was to use equation (3.2.4b) which required only a single pass through the data sample. A similar saving can be gained in computing the covariance by expanding the product

$$\sum_{I=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{N}(x_i y_i) - \frac{\left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{N} \tag{3.13.12}$$

3.13.3 Correlation and regression: cause and effect

A point worth stressing is that a high correlation coefficient or a "good" fit of a regression curve $y = y(x)$ to a set of observations x, does not imply that x is "causing" y. Nor does it imply that x will provide a good predictor for y in the future. For example, the number of sockeye salmon returning to the Fraser River of British Columbia each fall from the North Pacific Ocean is often highly correlated with the mean fall surface water temperature at Amphitrite Point on the southwest coast of Vancouver Island. No one believes that the fish are responding directly to the temperature, but rather that temperature is a proxy variable for the real factor (or combination of factors) influencing the homeward migration of the fish. Of course, we are not saying that one should not draw inferences or conclusions from correlation or regression analysis, but only that caution is advised when seeking cause-and-effect relationships between variables. We further remark that there is little point in drawing any type of line through the data unless the scatter about the line is

appreciably less than the overall spread of the observations. There is a tendency to fit trend lines to data with large variability and scatter even if a trend is not justified on statistical grounds. If $|r| < 0.5$, it hardly seems reasonable to fit a line for predictive purposes.

There is another important aspect of regression–correlation analysis that is worth stressing: Although the value of the correlation coefficient or coefficient of determination does *not* depend on which variable (x or y) is designated as the independent variable and which is designated as the dependent variable, this distinction *is* very important when it comes to regression analysis. The regression coefficients a, b for the conditional distribution of y given x ($y = a_1 + b_1x$) are different than those for the conditional distribution of x given y ($x = a_2 + b_2y$). In general, $a_1 \neq -a_2/b_2$ and $b_1 \neq 1/b_2$ so that the regression lines are different. In the first case, we are solving for the line shown in Figure 3.11(a), while in the second case we are solving for the line in Figure 3.11(b).

As an example, consider the broken lines in Figure 3.11(c) which show the two different linear regression lines for the regression of the observed cross-channel sea-level differences $y = \Delta\eta_c$, as measured by coastal tide gauges, and the calculated cross-channel sea-level difference $x = \Delta\eta_m$ obtained using concurrent current meter data from cross-channel moorings. The term $\Delta\eta_c$ is simply the difference in the mean sea-level from one side of the 25-km-wide channel to the other, while $\Delta\eta_m$ is calculated from the current meter records assuming that the time-averaged along-channel flow is in geostrophic balance (Labrecque *et al.*, 1994). The dotted line is the regression $\Delta\eta_c = a_1 + b_1\Delta\eta_m$ while the dashed line is the regression $\Delta\eta_m = a_2 + b_2\Delta\eta_c$, with $b_1 \neq b_2$. The correlation coefficient $r = 0.69$ is the same for the two regressions. The solid line in Figure 3.11(c) is the so-called *neutral* regression line for the two parameters (Garrett and Petrie, 1981) and might seem the line of choice since it is not obvious which parameter should be the independent parameter and which should be the dependent parameter. Neutral regression is equivalent to minimizing the sum of the square distances from the regression line (Figure 3.11d).

In fisheries research, neutral regression is known as *geometeric mean functional regression* (GMFR) and is commonly used to relate fish body proportions when there is no clear basis to select dependent and independent variables (Sprent and Dolby, 1980). For two variables with zero means, the slope estimator, b, is given by the square roots of the variance ratios

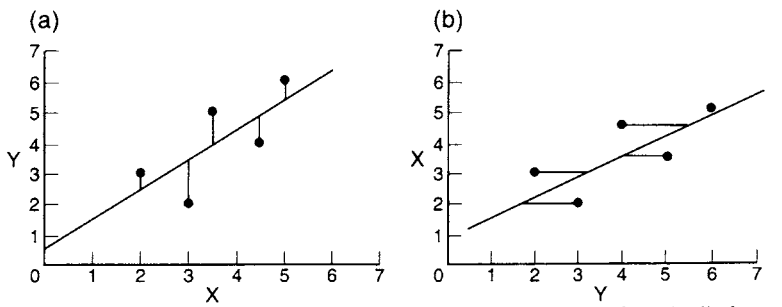

Figure 3.11. Straight line regressions (a) y on x, and (b) x on y showing the "direction" along which the variance is minimized.

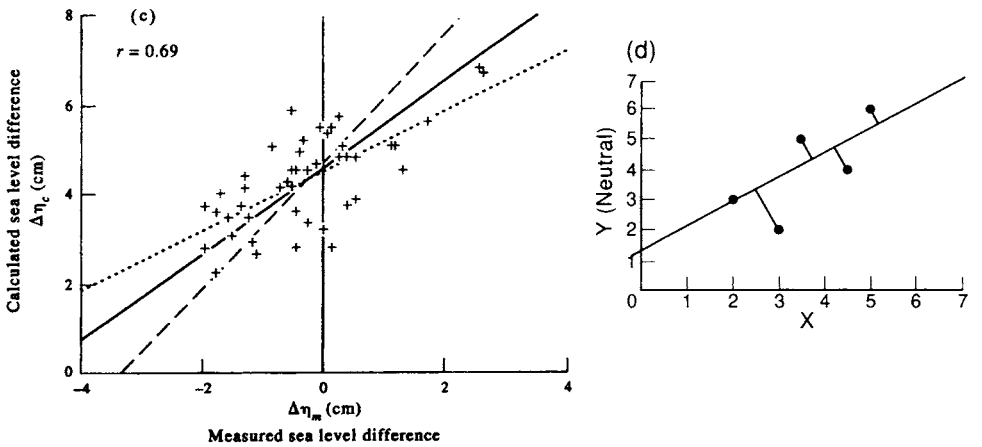

Figure 3.11. (c) Scatter plot of $\Delta\eta_c$ versus $\Delta\eta_m$ for a cross-section of the 22-km-wide Juan de Fuca Strait separating Vancouver Island from Washington State. Plots give the regression of the observed cross-channel sea level differences $y = \Delta\eta_c$, as measured by coastal tide gauges, and the calculated cross-channel sea level difference, $x = \Delta\eta_m$, obtained using concurrent current meter data from cross-channel moorings. The solid line gives the bisector regression fit to the data (slope and 95% confidence level = 0.96 ± 0.37); the dotted line (slope = 0.66 ± 0.14) and the dashed line (slope = 1.40 ± 0.32) are the standard slopes for $\Delta\eta_c$ versus $\Delta\eta_m$ and $\Delta\eta_m$ versus $\Delta\eta_c$, respectively. r = 0.69. (From Labrecque et al., 1994.) (d) The "direction" along which the variance for the data points in (a) and (b) is minimized.

$$b_{yx} = \text{sgn}\,(s_{xy})\frac{\sum_{i=1}^{N}(y_i-\bar{y})^2}{\left[\sum_{i=1}^{N}(x_i-\bar{x})^2\right]^{1/2}}; \quad \text{regression } \hat{y}_i = \hat{b}_{yx}x_i$$

$$b_{xy} = \text{sgn}\,(s_{xy})\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{\left[\sum_{i=1}^{N}(y_i-\bar{y})^2\right]^{1/2}}; \quad \text{regression } \hat{x}_i = \hat{b}_{xy}y_i \tag{3.13.13}$$

where sgn (s_{xy}) is the sign of the covariance function $s_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$ and $b_{yx} = 1/b_{xy}$, as required. Note that the slope b_{yx} lies midway between the slopes b_1 and b_2

$$b_1 = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{N}(x_i-\bar{x})^2}; \quad \text{regression line } \hat{y}_i = a_1 + \hat{b}_1 x_i$$

$$b_2 = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{N}(y_i-\bar{y})^2}; \quad \text{regression line } \hat{x}_i = a_2 + \hat{b}_2 y_i \tag{3.13.14}$$

given by (3.12.7a) for standard regression analyses (Figure 3.11a). The GMFR is then the geometric mean slope of the least-squares regression coefficient for the regression

slope of y on x and the regression of x on y; $b_{yx} = [b_1(b_2)^{-1}]^{1/2}$. Since the slope from the GMFR is simply a ratio of variances, it is "transparent" to the determination of correlation coefficients or coefficients of determination. It is these correlations, not the slope of the line, that test the strength of the linear relationship between the two variables. Moreover, none of the standard linear regression models reduces to the GMFR slope estimate except under unlikely circumstances. According to Sprent and Dolby (1980), *ad hoc* use of the GMFR is not recommended when there are errors in both variables. The GMFR model, though appealing, rests on shaky statistical ground and its use remains controversial.

3.14 HYPOTHESIS TESTING

Statistical inference takes one of two forms. Either we make estimates of population variables, as we have done thus far, or we test hypotheses about the implications of these variables. Statistical inference in which we choose between two conflicting hypotheses about the value of a particular population variable is known as *hypothesis testing*.

Hypothesis testing follows scientific methodology from whose nomenclature the terms are borrowed. The investigator forms a "hypothesis," collects some sample data and uses a statistical construct to either reject or accept the original hypothesis. The basic elements of a statistical test are: (1) the *null hypothesis*, H_o (the hypothesis to be tested); (2) the alternate hypothesis, H_a; (3) the test statistic to be used; and (4) the region of rejection of the hypothesis. The active components of a statistical test are the test statistic and the associated rejection region, with the latter specifying the values of the test statistic for which the null hypothesis is rejected. We emphasize the point that "pure" hypothesis testing originated from early work in which the null hypothesis corresponded to an idea or theory about a population variable that the scientist hoped *would be rejected*. "Null" in this case means incorrect and invalid so that we could call it the "invalid hypothesis". In other words, the null hypotheis specified those values of the population variable which it was thought did *not* represent the true value of the variable. This is a form of negative thinking and is the reason that many of us would rather think in terms of the *alternate hypothesis* in which we specify those values of the variable that we hope will hold true (the "valid" hypothesis). Regardless of which hypothesis is chosen, it is important to remember that the true population value under consideration must either lie in the test set covered by H_o or in the set covered by H_a. There are no other choices.

We restrict consideration of hypothesis testing to large samples ($N > 30$). In hypothesis testing, two types of errors are possible. In a type-1 error, the null hypothesis H_o is rejected when it is true. The probability of this type of error is denoted by α. Type-2 errors occur when H_o is accepted when it is false (H_a is true). The probability of type-2 errors is written as β. In Table 13.14.1, the probability P (accept $H_o | H_o$ is true) $= 1 - \alpha$ corresponds to the $100(1 - \alpha)\%$ confidence interval. Alternatively, the probability P(reject $H_o | H_o$ is false) $= 1 - \beta$ is the power of the statistical test since it indicates the ability of the test to determine when the null hypothesis is false and H_o should be rejected.

Table 13.14.1. The four possible decision outcomes in hypothesis testing and the probability of each decision outcome in a test hypothesis

		Possible situation	
		H_o is true	H_o is false
Action	Accept H_o	Correct decision; confidence level $1 - \alpha$	Incorrect decision; (Type-2 error); β
	Reject H_o	Incorrect decision (Type-1 error); α	Correct decision; power of the test $1 - \beta$
	Sum	1.00	1.00

For a parameter θ based on a random sample $x_1, \ldots, x_N$, we want to test various values of θ using the estimate $\hat{\theta}$ as a test statistic. This estimator is assumed to have an approximately normal sampling distribution. For a specified value of $\hat{\theta}\,(= \theta_o)$, we want to test the hypothesis, H_o, that $\hat{\theta} = \theta_o$ (written H_o: $\theta = \theta_o$) with the alternate hypothesis, H_a, that $\hat{\theta} > \theta_o$ (written H_a: $\theta > \theta_o$). An efficient test statistic for our assumed normal distribution is the standard normal Z defined as

$$Z = \frac{(\hat{\theta} - \theta)}{\hat{\sigma}_{\hat{\theta}}} \tag{3.14.1}$$

where $\hat{\sigma}_{\hat{\theta}}$ is the standard deviation of the approximately normal sampling distribution of $\hat{\theta}$, which can then be computed from the sample. For this test statistic, the null hypothesis (H_o: $\theta = \theta_o$) is rejected for $Z > Z_\alpha$ where α is the probability of a type-1 error. Graphically, this rejection region is depicted as the shaded portion in Figure 3.12(a), which is called an "upper-tail" test. Similarly, a "lower-tail" test would have the shaded rejection region starting at $-Z_\alpha$ and corresponds to $Z < -Z_\alpha$ and $\theta < \theta_o$ (Figure 3.12b). A two-tailed test (Figure 3.12c) is one for which the null hypothesis rejection region is $|Z| > Z_{\alpha/2}$ and $\theta \neq \theta_o$. The decision of which test alternative to use should be based on the form of the alternate hypothesis. If one is interested in parameter values greater than θ_o, an upper-tail test is used; for values less than θ_o, a lower-tail test is appropriate. If one is interested in any change from θ_o, it is best to use a two-tailed test. The following is an example for which a two-tailed test is appropriate.

Suppose that daily averaged currents for some mooring locations are available for the same month from two different years (e.g. January 1984 and January 1985). We wish to test the hypothesis that the monthly means of the longshore component of the flow, V, for these two different years are the same. If the daily averages are computed from hourly observations, we invoke the central limit theorem and conclude that our sampling distributions are normally distributed. Taking each month as having 31 days, we satisfy the condition of a large sample ($N > 30$) and can use the procedure outlined above. Suppose we observe that for January 1984 the mean and standard deviation of the observed current is $V_{84} = 23 \pm 3$ cm/s while for January 1985 we find a monthly mean speed $V_{85} = 20 \pm 2$ cm/s (here, the standard deviations are obtained from the signal variances). We now wish to test the null hypothesis that the true (as opposed to our sampled) monthly mean current speeds were statistically the same for the two separate years. We use the two-tailed test to detect any deviations from

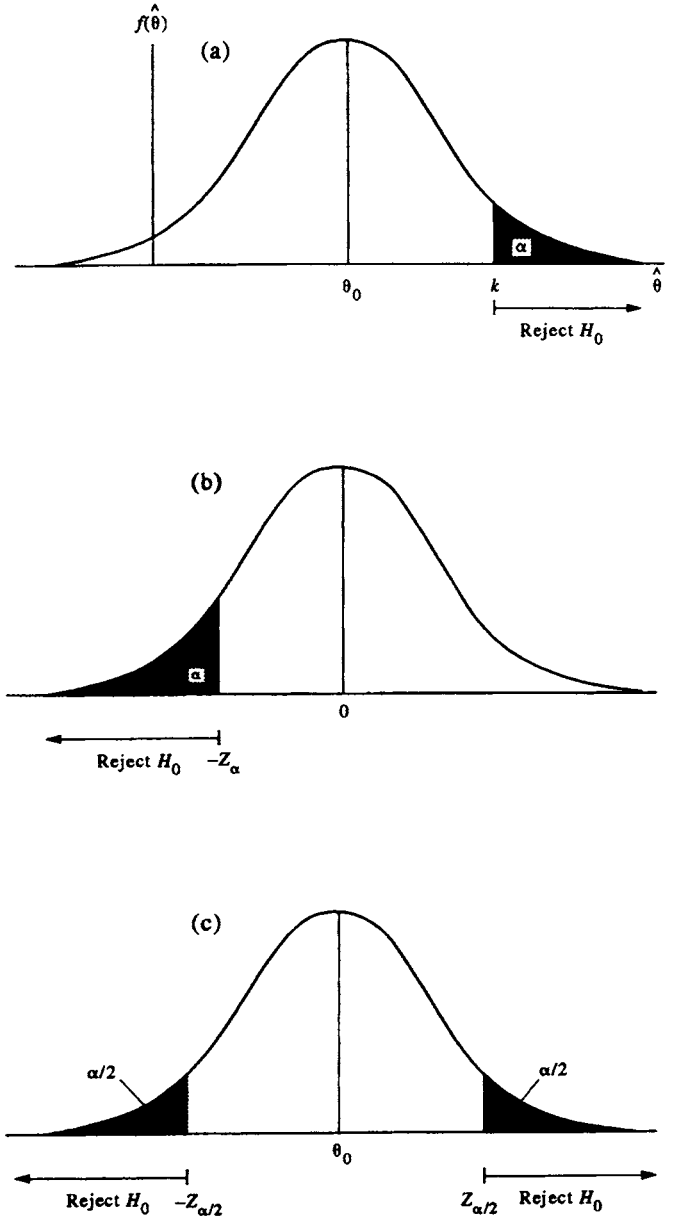

Figure 3.12. Large-sample rejection regions (shaded areas) for the null hypothesis $H_0 : \theta = \theta_0$, *for the normally distributed function* $f(\theta)$. *(a) Upper-tail test for* $H_o : \theta = \theta_o$, $H_a : \theta > \theta_o$; *(b) lower-tail test with the rejection region for* $H_o : \theta = \theta_o$, $H_a : \theta < \theta_o$; *and (c) two-tailed test for which the null hypothesis rejection region is* $|z| > z_{\alpha/2}$ *and* $H_a : \theta \neq \theta_o$.

equality. In this example, the *point estimator* used to detect any difference between the monthly mean records calculated from daily observed values is the sample mean difference, $\hat{\theta} - \theta_o = V_{84} - V_{85}$. Our test statistic (3.14.1) is

$$Z = \frac{(V_{84} - V_{85})}{\left[s_{84}^2/N_{84} + s_{85}^2/N_{85}\right]^{1/2}}$$

which yields

$$Z = \frac{(23 - 20)}{[9/31 + 4/31]^{1/2}} = 4.63$$

To determine if the above result falls in the rejection region, $Z > Z_\alpha$, we need to select the significance level α for type-1 errors. For the 95% significance level, $\alpha = 0.05$ and $\alpha/2 = 0.025$. From the standard normal table (Appendix D, Table D.1) $Z_{0.025} = 1.2$. Our test value $Z = 4.63$ is greater than 1.2 so that it falls within the rejection region, and we must reject the hypothesis that the monthly mean current speeds are the same for both years. In most oceanographic applications hypothesis testing is limited to the null hypothesis and thus type-1 errors are most appropriate. We will not consider here the implementation of type-2 errors which lead to the acceptance of an alternate hypothesis as described in Table 3.14.1.

Turning again to satellite altimetry for an example, we note that the altimeter height bias discussed earlier is one of the error sources that contributes to the overall error "budget" of altimetric height measurements. Suppose that we wish to know if the *overall* height error H_T is less than some specified amount, H_ε. We first set up the null hypothesis (H_o: $H_T < H_\varepsilon$) that the true mean is less than H_ε. At this point, we must also select a significance level for our test. A significance level of $1 - \alpha$ means that we do not want to make a mistake and reject the null hypothesis more than $\alpha(100)\%$ of the time. We begin by defining our hypothesis limit, H_T, as

$$H_T = H_\varepsilon + \frac{z_\alpha s_b}{\sqrt{N}} \tag{3.14.2}$$

where the standard normal distribution Z_α is given by equation (3.14.1) and s_b is the standard error (uncertainty) in our measurements. If the mean of our measurements is greater than H_T, then we reject H_o and conclude that the mean is greater than H_ε with a probability α of being wrong.

Suppose we set $H_\varepsilon = 13$ cm and consider $N = 9$ consecutive statistically independent satellite measurements in which each measurement is assumed to have an uncertainty of $s_b = 3$ cm. If the mean height error is 15 cm, do we accept or reject the null hypothesis for the probability level $\alpha = 0.10$? What about the cases for $\alpha = 0.05$ and $\alpha = 0.01$? Given our hypothesis limit $H_\varepsilon = 13$ cm and the fact that $N = 9$ and $s_b = 3$ cm, we can write equation (3.14.2) as $H_T = 13 + Z_\alpha$ cm. According to the results of

Table 3.14.2. Testing the null hypothesis that the overall bias error H_T of satellite altimetry data is less than 13 cm. Assumes normal error distributions

Significance level, α	Standard normal distribution, Z_α	Total error height, H_T	Decision
0.10	2.326	15.326 cm	reject H_o
0.05	1.645	14.645 cm	accept H_o
0.01	1.280	14.280 cm	accept H_o

Table 3.14.2, this means that we accept the null hypothesis that the overall error is less than 13 cm at the 5 and 10% probability levels but not at the 1% probability level (these are referred to as the 95%, 90%, and 99% significance levels, respectively).

3.14.1 Significance levels and confidence intervals for correlation

One useful application of null hypothesis testing is the development of significance levels for the correlation coefficient, r. If we take the null hypothesis as $r = r_o$, where r_o is some estimate of the correlation coefficient, we can determine the rejection region in terms of r at a chosen significance level α for different degrees of freedom $(N-2)$. A list of such values is given in Appendix E. In that table, the correlation coefficient r for the 95 and 99% significance levels (also called the 5 and 1% levels depending on whether or not one is judging a population parameter or testing a hypothesis) are presented as functions of the number of degrees of freedom.

For example, a sample of 20 pairs of (x, y) values with a correlation coefficient less than 0.444 and $N-2 = 18$ degrees of freedom would not be significantly different from zero at the 95% confidence level. It is interesting to note that, because of the close relationship between r and the regression coefficient b_1 of these pairs of values, we could have developed the table for r values using a test of the null hypothesis for b_1.

The procedure for finding confidence intervals for the correlation coefficient r is to first transform it into the standard normal variable Z_r as

$$Z_r = \frac{1}{2}[\ln(1+r) - \ln(1-r)] \tag{3.14.3}$$

which has the standard error

$$\sigma_z = \frac{1}{(N-3)^{1/2}} \tag{3.14.4}$$

independent of the value of the correlation. The appropriate confidence interval is then

$$Z_r - Z_{\alpha/2}\sigma_z < Z < Z_r + Z_{\alpha/2}\sigma_z \tag{3.14.5}$$

which can be transformed back into values of r using equation (3.14.3).

Before leaving the subject of correlations we want to stress that correlations are merely statistical constructs and, while we have some mathematical guidelines as to the statistical reliability of these values, we cannot replace common sense and physical insight with our statistical calculations. It is entirely possible that our statistics will deceive us if we do not apply them carefully. We again emphasize that a high correlation can reveal either a close relationship between two variables or their simultaneous dependence on a third variable. It is also possible that a high correlation may be due to complete coincidence and have no causal relationship behind it. A classic example (Snedecor and Cochran, 1967) is the high negative correlation (-0.98) between the annual birthrate in Great Britain and the annual production of pig iron in the United States for the years 1875–1920. This high correlation is statistically significant for the available $N-2 = 43$ degrees of freedom, but the likelihood of a direct relationship between these two variables, is very low.

3.14.2 Analysis of variance and the *F*-distribution

Most of the statistical tests we have presented to this point are designed to test for differences between two populations. In certain circumstances, we may wish to investigate the differences among three or more populations simultaneously rather than attempt the arduous task of examining all possible pairs. For example, we might want to compare the mean lifetimes of drifters sold by several different manufacturers to see if there is a difference in survivability for similar environmental conditions; or, we might want to look for significant differences among temperature or salinity data measured simultaneously during an intercomparison of several different commercially available CTDs. The *analysis of variance* (ANOVA) is a method for performing simultaneous tests on data sets drawn from different populations. In essence, ANOVA is a test between the amount of variation in the data that can be attributed to chance and that which can be attributed to specific causes and effects. If the amount of variability *between* samples is small relative to the variability *within* samples, then the null hypothesis H_o—that the variability occurred by chance—cannot be rejected. If, on the other hand, the ratio of these variations is sufficiently large, we can reject H_o. "Sufficiently large", in this case, is determined by the ratio of two continuous χ^2 probability distributions. This ratio is known as the *F-distribution*.

To examine this subject further, we need several definitions. Suppose we have samples from a total of $\mathcal{J}$ populations and that a given sample consists of N_j values. In ANOVA, the $\mathcal{J}$ samples are called $\mathcal{J}$ "treatments", a term that stems from early applications of the method to agricultural problems where soils were "treated" with different kinds of fertilizer and the statistical results compared. In the one-factor ANOVA model, the values y_{ij} for a particular treatment (input), x_j, differ from some common background value, μ, because of random effects; that is

$$y_{ij} = \mu + x_j + \varepsilon_{ij}; \quad \begin{matrix} j = 1, 2, ..., \mathcal{J} \\ i = 1, 2, ..., N_{\mathcal{J}} \end{matrix} \tag{3.14.6}$$

where the outcome y_{ij} is made up of a common (grand average) effect (μ), plus a treatment effect (x_j) and a random effect, ε_{ij}. The grand mean, μ, and the treatment effects, x_j, are assumed to be constants while the errors, ε_{ij}, are independent, normally distributed, variables with zero mean and a common variance, σ^2, for all populations. The null hypothesis for this one-factor model is that the treatments have zero effect. That is, H_o: $x_j = 0$ ($j = 1, 2, ..., \mathcal{J}$) or, equivalently, H_o: $\mu_1 = \mu_2 = ... = \mu_{\mathcal{J}}$ (i.e. there is no difference between the populations aside from that due to random errors). The alternative hypothesis is that some of the treatments have a nonzero effect. Note that "treatment" can refer to any basic parameter we wish to compare such as buoy design, power supply, or CTD manufacturer. To test the null hypotheses, we consider samples of size N_j from each of the $\mathcal{J}$ populations. For each of these samples, we calculate the mean value $\bar{y}_j$ ($j = 1, 2, ..., \mathcal{J}$). The grand mean for all the data is denoted as $\bar{y}$.

As an example, suppose we want to intercompare the temperature records from three types of CTDs placed in the same temperature bath under identical sampling conditions. Four countries take part in the intercomparison and each brings the same three types of CTD. The results of the test are reproduced in Table 3.14.3.

If H_o is true, $\mu_1 = \mu_2 = \mu_3$, and the measured differences between $\bar{y}_1, \bar{y}_2$, and $\bar{y}_3$ in Table 3.14.3 can be attributed to random processes.

Table 3.14.3. Temperatures in °C measured by three makes of CTD in the same calibration tank. Four instruments of each type are used in the test. The grand mean for the data from all three instruments is $\bar{y} = 15.002°C$

Measurement (i)	CTD Type 1 Sample $j = 1$	CTD Type 2 Sample $j = 2$	CTD Type 3 Sample $j = 3$
1	15.001	15.004	15.002
2	14.999	15.002	15.003
3	15.000	15.001	15.000
4	14.998	15.004	15.002
Mean $\bar{y}$ °C	15.000 $= \bar{y}_1$	15.003 $= \bar{y}_2$	15.002 $= \bar{y}_3$

The treatment effects for the CTD example are given by

$$x_1 = \bar{y}_1 - \bar{y} = -0.002°C$$
$$x_2 = \bar{y}_2 - \bar{y} = +0.001°C$$
$$x_3 = \bar{y}_3 - \bar{y} = 0.0°C$$

where $\bar{y} = (\bar{y}_1 + \bar{y}_2 + \bar{y}_3)/3$. The ANOVA test involves determining whether the estimated values of x_j are large enough to convince us that H_o is not true. Whenever H_o is true, we can expect that the variability between the $\mathcal{J}$ means is the same as the variability within each sample (the only source of variability is the random effects, ε_{ij}). However, if the treatment effects are not all zero, then the variability between samples should be larger than the variability within the samples.

The variation within the $\mathcal{J}$ samples is found by first summing the squared deviations of y_{ij} about the mean value $\bar{y}_j$ for each sample, namely

$$\sum_{i=1}^{N_\mathcal{J}} (\bar{y}_{ij} - \bar{y}_j)^2$$

If we then sum this variation over all $\mathcal{J}$ samples, we obtain the *sum of squares within* (SSW)

$$\text{Sum of squares within: SSW} = \sum_{j=1}^{\mathcal{J}} \sum_{i=1}^{N_j} (y_{ij} - \bar{y}_j)^2 \tag{3.14.7}$$

Note that the sample lengths, N_j, need not be the same since the summation for each sample uses only the mean for that particular sample. Next, we will need the amount of variation between the samples (SSB). This is obtained by taking the squared deviation of the mean of the $\mathcal{J}$th sample, $\bar{y}_j$, and the grand mean, $\bar{y}$. This deviation must then be weighted by the number of observations in the $\mathcal{J}$th sample. The overall sum is given by

$$\text{Sum of squares between: SSB} = \sum_{j=1}^{\mathcal{J}} N_j (\bar{y}_j - \bar{y})^2 \tag{3.14.8}$$

To compare the variability within samples to the variability between samples, we need to divide each sum by its respective number of degrees of freedom, just as we did with

other variance expressions such as s^2. For SSB, the degrees of freedom (DOF) $= \mathcal{J} - 1$ while for SSW

$$\text{DOF} = \left(\sum_{j=1}^{\mathcal{J}} N_j\right) - \mathcal{J}$$

The *mean square* values are then:

$$\text{Mean square between: MSB} = \frac{SSB}{\mathcal{J} - 1} \tag{3.14.9a}$$

$$\text{Mean square within: MSW} = \frac{SSW}{\left(\sum_{j=1}^{\mathcal{J}} N_j\right) - \mathcal{J}} \tag{3.14.9b}$$

In the above example, $\mathcal{J} - 1 = 2$ and $\sum_{j=1}^{\mathcal{J}} N_j - \mathcal{J} = 9$. The calculated values of MSB and MSW for our CTD example are given in Table 3.14.4. Specifically

$$\text{SSW} = \sum_{i=1}^{4} (y_{i1} - \bar{y}_1)^2 + \sum_{i=1}^{4} (y_{i2} - \bar{y}_2)^2 + \sum_{i=1}^{4} (y_{i3} - \bar{y}_3)^2$$

$$\text{SSB} = N_1(\bar{y}_1 - \bar{y})^2 + N_2(\bar{y}_2 - \bar{y})^2 + N_3(\bar{y}_3 - \bar{y})^2 + N_4(\bar{y}_4 - \bar{y})^2$$

where $N_j = 4$ ($j = 1, \ldots, 4$). To determine if the ratio of MSB to MSW is large enough to reject the null hypothesis, we use the F-distribution for $\mathcal{J} - 1$ and

$$\left(\sum_{j=1}^{\mathcal{J}} N_j\right) - \mathcal{J}$$

degrees of freedom.

Named after R. A. Fisher who first studied it in 1924, the F-distribution is defined in terms of the ratio of two independent χ^2 variables divided by their respective degrees of freedom. If X_1 is a χ^2 variable with ν_1 degrees of freedom and X_2 is another χ^2 variable with ν_2 degrees of freedom, then the random variable

$$F(\nu_1, \nu_2) = \frac{X_1/\nu_1}{X_2/\nu_2} \tag{3.14.10}$$

is a nonnegative chi-square variable with ν_1 degrees of freedom in the numerator and ν_2 degrees of freedom in the denominator. If $\mathcal{J} = 2$, in the CTD example above, the F-

Table 3.14.4. Calculated values of sum of squares and mean square values for the CTD temperature intercomparison. DOF = number of degrees of freedom

Type of variation	Sum of squares ($°C^2$)	DOF	Mean square ($°C^2$)
Between samples (type of CTD)	20×10^{-6}	2	10×10^{-6}
Within samples (all CTDs)	18×10^{-6}	9	2×10^{-6}
Total	38×10^{-6}	11	(ratio = 5.0)

test is equivalent to a one-sided t-test. There is no upper limit to F, which like the χ^2 distribution is skewed to the right. Tables are used to list the critical values of $P\,(F > F_\alpha)$ for selected degrees of freedom ν_1 and ν_2 for the two most commonly used significance levels, $\alpha = 0.05$ and $\alpha = 0.01$. In ANOVA, the values of SSB and SSW follow χ^2-distributions. Therefore, if we let $X_1 = \text{SSB}$ and $X_2 = \text{SSW}$, then

$$F\left(\mathcal{J}-1, \sum N_j - \mathcal{J}\right) = \frac{[\text{SSB}/(\mathcal{J}-1)]}{\text{SSW}(\sum N_j - \mathcal{J})} = \frac{\text{MSB}}{\text{MSW}} \tag{3.14.11}$$

When MSB is large relative to MSW, F will be large and we can justifiably reject the null hypothesis that the different CTDs (different treatment effects) measure the same temperature within the accuracy of the instruments. For our CTD intercomparison (Table 3.14.4), we have MSB/MSW = 5.0, ν_1=2 and $\nu_2 = 9$. Using the values for the F-distribution for 2 and 9 degrees of freedom from Appendix D, Table D.4a, we find $F_\alpha(2,9) = 4.26$ for $\alpha = 0.05$ (95% confidence level) and $F_\alpha(2,9) = 8.02$ for $\alpha = 0.01$ (99% confidence level). Since, $F = 5.0$ in our example, we conclude that a difference exists among the different makes of CTD at the 95% confidence level, but not at the 99% confidence level.

3.15 EFFECTIVE DEGREES OF FREEDOM

To this point, we have assumed that we are dealing with random variables and each of the N values in a given sample are statistically independent. For example, in calculating the unbiased standard deviation for N data points, we assume there are $N - 1$ degrees of freedom. (We use $N - 1$ rather than N since we need a minimum of two values to calculate the standard deviaton of a sample.) Similarly, in Sections 3.8 and 3.10, we specify confidence limits in terms of the number of samples rather than the "true" number of degrees of freedom of the sample. In reality, consecutive data values may not be independent. Contributions from low-frequency components and narrow band oscillations such as in inertial motions may lead to a high degree of correlation between values separated by large times or distances. The most common example of highly coherent narrow band signals are the tides and tidal currents which possess a strong temporal and spatial coherence. If we want our statistics to have any real meaning, we are forced to find the *effective number of degrees of freedom* using information on the coherence and autocorrelation of our data.

The effects of coherent nonrandom processes on data series lead us into the question of data redundancy in multivariate linear regression. Our general model is

$$\hat{y}(t_i) = \sum_{k=1}^{M} b_k x_k(t_i); \quad i = 1, \ldots, N \tag{3.15.1}$$

where the x_k represents M observed parameters or quantities at times t_i. The b_k are M linear-regression coefficients relating the independent variables $x_k(t_i)$ to the N model estimates, $\hat{y}(t_i)$. Here, the x_k observations can be measurements of different physical quantities or of the same quantity measured at different times or locations.

The estimate $\hat{y}$ differs from the true parameter by an error $\varepsilon_i = \hat{y}(t_i) - y(t_i) = \hat{y}_i - y_i$. Following our earlier discussion, we assume that this error is randomly distributed

and is therefore uncorrelated with the input data $x_k(t_i)$. To find the best estimate, we apply the method of least squares to minimize the mean square error

$$\overline{\varepsilon^2} = \sum_{i=1}^{M}\sum_{j=1}^{M} b_i b_j \overline{x_i x_j} - 2\sum_{j=1}^{M} b_j \overline{x_j y} + \overline{y^2} \tag{3.15.2}$$

In this case, the overbars represent ensemble averages. To assist us in our minimization, we invoke the Gauss–Markov theorem which says that the estimator, given by equation (3.15.1), with the smallest mean square error is that with coefficients

$$b_k = \sum_{j=1}^{M}\left[\left\{\overline{x_k x_j}\right\}^{-1}\overline{x_j y}\right] \tag{3.15.3}$$

where $\left\{\overline{x_k x_j}\right\}^{-1}$ is the i, j element of the inverse of the $M \times M$ cross covariance matrix of the input variables (note: $\left\{\overline{x_k x_j}\right\}^{-1} \neq 1/\overline{x_k x_j}$). This mean-square product matrix is always positive definite unless one of the input variables x_k can be expressed as an exact combination of the other input values. The presence of random measurement errors in all input data make this "degeneracy" highly unlikely. It should be noted however, that it is the partial correlation between inputs that increases the uncertainty in our estimator by lowering the degrees of freedom through a reduction in the independence of our input parameters.

We can write the minimum least-square error ε_o^2 as

$$\overline{\varepsilon_o^2} = \overline{y^2} - \sum_{i=1}^{M}\sum_{j=1}^{M} \overline{y x_j}\left\{\overline{x_k x_j}\right\}^{-1}\overline{y x_j} \tag{3.15.4}$$

At this point, we introduce a measure of the reliability of our estimate called the *skill* (S) of the model. This skill is defined as the fraction of the true parameter variance explained by our linear statistical estimator; thus

$$S = \left\{\overline{y^2}\right\}^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\overline{y x_j}\left\{\overline{x_k x_j}\right\}^{-1}\overline{y x_j}\right] \tag{3.15.5}$$

The skill value ranges from no skill ($S = 0$) to perfect skill ($S = 1$). We note that for the case ($M = 1$), S is the square of the correlation between x_1 and y.

The fundamental trade-off for any linear estimation model is that, while one wants to use as many independent input variables as possible to avoid interdependence among the estimates of the dependent variable, each new input contributes random measurement errors that degrade the overall estimate. As pointed out by Davis (1977) the best criterion for selecting the input data parameters is to use *a priori* theoretical considerations. If this is not possible, some effort should be made to select those inputs which contribute most to the estimation skill.

The conflicting requirements of limiting M (the observed parameters) and including all candidate input parameters is a dilemma. In considering this dilemma Chelton (1983) concludes that the only way to reduce the error limits on the estimated regression coefficients is to increase what are called the "effective degrees of freedom N^*." This can be done only by increasing the sample size of the input variable (i.e using a longer time series) or by high-passing the data to eliminate contributions from

unresolved, and generally coherent, low-frequency components. Since we are forced to deal with relatively short data records in which ensemble averages are replaced by sample averages over time or space, we need a procedure to evaluate N^*, the effective degrees of freedom.

In the case of real data, ensemble averages are generally replaced with sample averages over time or space so that the resultant values become estimates. Thus, the skill can be written as S given by (3.15.5). If we assume for a moment that the x_k input data are serially uncorrelated (i.e. we expand the data series into orthogonal functions), we can write the sample estimate of the skill as

$$\hat{S} = \sum_{i=1}^{M} \sum_{j=1}^{M} \frac{\overline{x_i x_j}^2}{\overline{x_j^2 y^2}}$$

Following Davis (1978) we can expand this skill estimate into a true skill plus an artificial skill

$$\hat{S} = S + S_A \tag{3.15.7}$$

The artificial skill, S_A, arises from errors in the estimates and can be calculated by evaluating the skill in equation (3.15.6) at a very long time (or space lag) where no real skill is expected. At this point, there is no true estimation skill and $\hat{S} = S_A$.

Davis (1976) derived an appropriate expression for the expected (mean) value of this artificial skill which relates it to the effective degrees of freedom N^*

$$\overline{S}_A = \sum_{k=1}^{M} (N_k^*)^{-1} \tag{3.15.8}$$

where N_k^* is the effective degrees of freedom associated with the sample estimate of the covariance between the output y and and input x_k of the model. Under the conditions that S (the true skill) is not large, that the record length N is long compared to the autocovariance scales of y and x, and that the N_k^* are the same for all N, we can write N^* as

$$N^* = \frac{N}{\left[\sum_{\tau=-\infty}^{\infty} C_{xx}(\tau) C_{yy}(\tau)\right] / \left[C_{xx}(0) C_{yy}(0)\right]} \tag{3.15.9a}$$

$$= \frac{N}{\left[\sum_{\tau=-\infty}^{\infty} \rho_{xx}(\tau) \rho_{yy}(\tau)\right]} \tag{3.15.9b}$$

where $\rho_{\zeta\zeta}(\tau) = C_{\zeta\zeta}(\tau)/C_{\zeta\zeta}(0) = C_{\zeta\zeta}(\tau)/s_\zeta^2$ is the normalized autocovariance function for any variable ζ (with variance s_ζ^2), and

$$\begin{aligned} C_{\zeta\zeta}(\tau) &= E[(\zeta(t_i) - \bar{\zeta})(\zeta(\tau + t_i) - \bar{\zeta})] \\ &= \frac{1}{N'} \sum_{i=1}^{N'} \{(\zeta(t_i) - \bar{\zeta})(\zeta(\tau + t_i) - \bar{\zeta})\} \end{aligned} \tag{3.15.10}$$

where N' is the number of data values used in the summation for the particular lag, τ. A more complete form of this expression was given by Chelton (1983) as

$$N^* = \frac{N}{\left[\sum_{\tau=-\infty}^{\infty} C_{xx}(\tau)C_{yy}(\tau) + C_{xy}(\tau)C_{yx}(\tau)\right] / [C_{xx}(0)C_{yy}(0)]} \tag{3.15.11a}$$

$$= \frac{N}{\left[\sum_{\tau=-\infty}^{\infty} \rho_{xx}(\tau)\rho_{yy}(\tau) + \rho_{xy}(\tau)\rho_{yx}(\tau)\right]} \tag{3.15.11b}$$

This expression now includes the cross-covariances between y and x [e.g. $C_{xy}(\tau)$ and $\rho_{xy}(\tau)$] and is not limited to cases where S is small.

In general, the true auto- and cross-covariances are not known and the computation of N^* requires the substitution of sample estimates over finite lags for the correlations in equation (3.15.11). The resulting effective degrees of freedom, N^*, can be used with standard tables to find the selected significance levels for $\hat{S}$. In the ideal case, when all input variables are neither cross- nor serially correlated (and therefore independent) the effective number of degrees of freedom is N, the sample size. In general, however input data series are serially correlated and $N^* \ll N$. The larger the time/space correlation scales in equation (3.15.11), the smaller the value of N^*. This means that it is the large scale, low-frequency components of the input data that lead to a decrease in the number of independent values in the data series.

The limitations of estimating regression characteristics for real data can be summarized as follows:

(1) Accurate statistical results require the use of the effective number of degrees of freedom, N^*, with N^* generally much less than the total number of observations N.
(2) The accuracy of the estimated regression coefficients increases as N^* increases.
(3) The accuracy of the regression coefficient decreases as the number of inputs M increases (measurement error is added).
(4) The accuracy increases as the model skill increases and decreases as the input parameters become more correlated.

The above considerations emphasize the need for careful selection of the input data and the careful evaluation of the characteristics of these data. As pointed out by Davis (1977), a fundamental part of this selection process is the determination of the space and time scales to be studied. The methods used to extract this fundamental scale information from the input data can range from cross-spectral analysis (see Chapter 5) to a filtering of the data using preselected windows. Performing this filtering in the time domain rather than the frequency domain is often less complicated. The filtering process has the goal of eliminating scales that are not expected to contribute to the true correlation but which will add artificial correlation due to instrument and sampling errors.

Once the space and/or time scales are determined, selected or set by filtering, the next step is the selection of the input series to use in the estimate. At this stage, the dilemma arises between limiting the effects of errors and at the same time including as many as possible uncorrelated input variables to increase the degrees of freedom. Davis (1977) recommends using dynamical considerations to make this selection and shows how the data required for proper statistical estimation are generally those required to make the dynamical system well posed. However, he also mentions that, in

general, the dynamics of most processes are not well enough understood and that specification data are not known with certainty. Nevertheless, some quantitative understanding of the physical system can serve as a useful guide to the selection of estimation data.

3.15.1 Trend estimates and the integral time scale

Most oceanographic variability arises through a combination of random and non-random processes. The presence of tidal and low-frequency components means that data points in time or space series are not independent of one another. The data that we collect are not truly random samples drawn from random populations. There is invariably a nonzero correlation between values in the series which must be taken into account when we tally-up the true number of independent samples or degrees of freedom we think we have in our system. This number is important when it comes to determining the confidence limits of linear regression slopes and parameter estimates.

As an example, consider the confidence limits on the slope of the least squares linear regression $\hat{y} = b_0 + b_1 x$ (where, again, $\hat{}$ denotes an estimator for the function y). From equation (3.8.6), the limits are

$$\pm(s_\varepsilon t_{\alpha/2,\nu})/[(N-1)s_x]^{1/2} \tag{3.15.12a}$$

or, in terms of the estimator β_1 for b_1

$$b_1 - \frac{(s_\varepsilon t_{\alpha/2,\nu})}{[(N-1)s_x]^{1/2}} < \beta_1 < b_1 + \frac{(s_\varepsilon t_{\alpha/2,\nu})}{[(N-1)s_x]^{1/2}} \tag{3.15.12b}$$

where $\nu = N - 2$ is the number of degrees of freedom for the Student's t-distribution at the $(1-\alpha)100\%$ confidence level, and the standard error of the estimate, s_ε, is given by

$$s_\varepsilon = \left[\frac{1}{N-2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right]^{1/2} = \left[\frac{1}{N-2}\mathrm{SSE}\right]^{1/2} \tag{3.15.13}$$

The standard deviation for the x variable, s_x, is given by

$$s_x = \left[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{1/2} \tag{3.15.14a}$$

or

$$\sqrt{N-1}\,s_x = \left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{1/2} \tag{3.15.14b}$$

The question is: what do we use for the number of degrees of freedom if the N samples in our series are not statistically independent? The reason we ask this question is that the characteristic amplitudes of the fluctuations s_ε and s_x are calculated using all N values in our data series when we really should be using some

sort of *effective* number of degrees of freedom $N^*(< N)$ which takes into account the degree of correlation that exists between data points (as discussed in the previous section).

Suppose we decide to err on the conservative side by agreeing to work with that value of N^* which makes the confidence limits $\pm(s_\varepsilon t_{\alpha/2.\nu})/[(N-1)s_x]^{1/2}$ as small as justifiably possible. This means that when we estimate the confidence limits for a regression slope for a given confidence coefficient, α, we know that we have probably been too cautious and that the confidence limits on the slope probably bracket those that we derive.

We begin by keeping s_ε as it is. If there are high frequency (possibly random) fluctuations superimposed on coherent low-frequency motions, retaining the high-frequency variability adds to the magnitude of s_ε. Had we low-pass filtered the data first and recomputed s_ε based on the true number of data points in our low-pass filtered record, we would expect s_ε to be somewhat smaller. By using s_ε as it is we are assuming that it is a fixed quantity no matter how we subsample or filter the data (s_ε = constant). We do the same with s_x but now replace $N-1$ with N^*-1, where $N^* < N$. This increases the magnitude of the confidence limits. All that remains is to assume that the number of degrees of freedom for the t-distribution are given by the effective number of degrees of freedom $\nu = N^* - 2$. This statistic has a larger value than for $\nu = N-2$ so that, again, we are overestimating the magnitude of the confidence interval. This confidence interval is then given by

$$\pm(s_\varepsilon t_{\alpha/2.\nu})/[(N^*-1)s_x]^{1/2} \tag{3.15.15a}$$

i.e.

$$b_1 - \frac{(s_\varepsilon t_{\alpha/2.\nu})}{[(N^*-1)s_x]^{1/2}} < \beta_1 < b_1 + \frac{(s_\varepsilon t_{\alpha/2.\nu})}{[(N^*-1)s_x]^{1/2}} \tag{3.15.15b}$$

with $\nu = N^* - 2$.

Our final task is define the effective number of degrees of freedom, N^*, based on a knowledge of the autocovariance function $C(\tau)$ (3.15.10) as a function of lag τ. To do this, we must first find the integral time scale T for the data record

$$T = \frac{1}{C(0)} \sum_{k=0}^{m-1} \frac{\Delta\tau}{2} [C(\tau_k + \Delta\tau) + C(\tau_k)] \text{(discrete case)} \tag{3.15.16a}$$

$$= \frac{1}{C(0)} \int_0^\infty C(\tau) \mathrm{d}\tau \text{ continuous case} \tag{3.15.16b}$$

where m is the number of lag values to be incorporated in the integral, $\Delta\tau$ is the time increment between data values and $\frac{1}{2}[C(\tau_k + \Delta\tau) + C(\tau_k)]$ is the mean value of C for the midpoint of the lag interval $(\tau_k, \tau_k + \Delta\tau)$. Once the integral time scale is known, the effective degrees of freedom are found by

$$N^* = \frac{N\Delta t}{T} \tag{3.15.17}$$

where Δt is the sampling increment and $N\Delta t$ is the total length (duration or distance) of the record. If, for example, $N = 120$, $\Delta t = 1$ h, and $T = 10$ h, then $N^* = 12$ ($\ll N$).

To find the autocovariance function, we let $\tau_k = k\Delta\tau$ be the kth lag ($k = 0, 1, \ldots$), then

$$C(\tau_k) = \frac{1}{N-1-k}\sum_{i=1}^{N-k}(y_i - \overline{y_i})(y_{i+k} - \overline{y_{i+k}});\ k = 0, \ldots, N_{\max} \tag{3.15.18a}$$

$$C(0) = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y_i})^2 = s_y^2 \tag{3.15.18b}$$

where $C(0)$ is the just the variance s_y^2 of the full data series. In both equations (3.15.18a) and (3.15.18b), the data start with the first value for $i = 1$; $N_{\max}$ is the maximum number of reasonable lag values (starting at zero lag and going to $\ll N/2$) that can be calculated before the summation becomes erratic. In theory, we would like $C(\tau) \to 0$ as $\tau \to N$. In reality, however, the data series will contain low-frequency components which will cause the autocovariance function to oscillate about zero or asymptote towards a nonzero value. It should also be obvious that the statistical significance of the summation becomes meaningless at large lag due to the fact that the statistic is based on fewer and fewer values as the lag becomes large. For example, at a lag $k = (N - 3)$ there are only four values that go into the summation and these are derived from neighboring points that are likely highly correlated.

We can picture the integral time scale using equation (3.15.16b). Writing

$$T \cdot C(0) = \int_{\text{all } \tau} C(\tau)\mathrm{d}\tau$$

we see that the area under the curve $C(\tau)$ has been equated to the rectangular region $T \cdot C(0)$ (Figure 13.13). In essence, we take a reasonable portion of the curve $C(\tau)$, obtain its area and divide the integral (sum) by its value at zero lag, $C(0)$. An example of the autocovariance function and the integral time scale derived from it are shown in Figure 13.14 for satellite-tracked drifter data in the North Pacific.

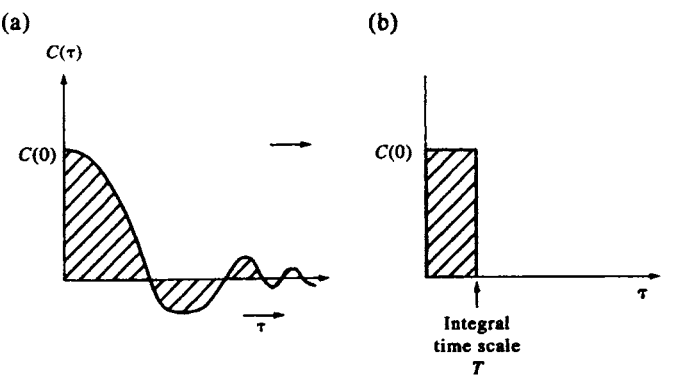

Figure 3.13. Definition of the integral time scale. The area under the curve $C(\tau)$ *versus* τ *in (a) is equated to the rectangular region TC(0) in (b). In practice, only a reasonable portion of the curve* $C(\tau)$ *is used to obtain the area in (a).*

3.16 EDITING AND DESPIKING TECHNIQUES: THE NATURE OF ERRORS

A major concern in processing oceanographic data is how to distinguish the true oceanic signal from measurement "errors" or other erroneous values. There are two very different types of measurement errors that can affect data. *Random errors*, usually equated with "noise", have random probability distributions and are generally small compared to the signal. Random errors are associated with inaccuracies in the measurement system or with real variability that is not resolved by the measurement system. The well-accepted statistical techniques for estimating the effects of such random errors are based largely on the statistics of a random population (see previous sections on statistics). Other errors which strongly influence data analysis are *accidental* errors. These errors are not representative of the true population and occur as a result of undetected instrument failures, misreading of scales, incorrect recording of data, and other human failings. In the following discussion, we will handle these two error types in reverse order since the large accidental errors must be removed first before techniques can be applied to treat the "statistical" (random) errors.

One example of a large accidental error would be assigning an incorrect geographic location to an oceanographic measurement which then transfers the observations to a

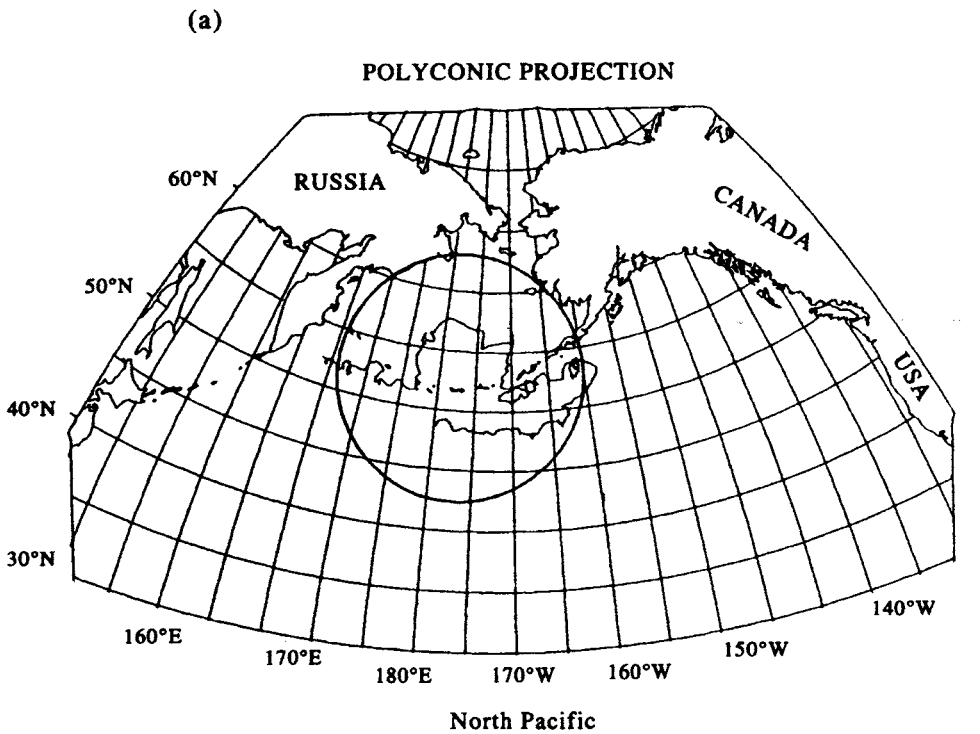

Figure 3.14. Autocovariance functions and corresponding integral time scales for zonal (u) and meridional (v) velocities of satellite-tracked drifter deployed to the south of the Aleutian Islands in the northeast Pacific (see insert) and covering the period 13 November 1991 to 30 July 1993 based on six-hourly sampling interval. (Courtesy of Adrian Dolling.)

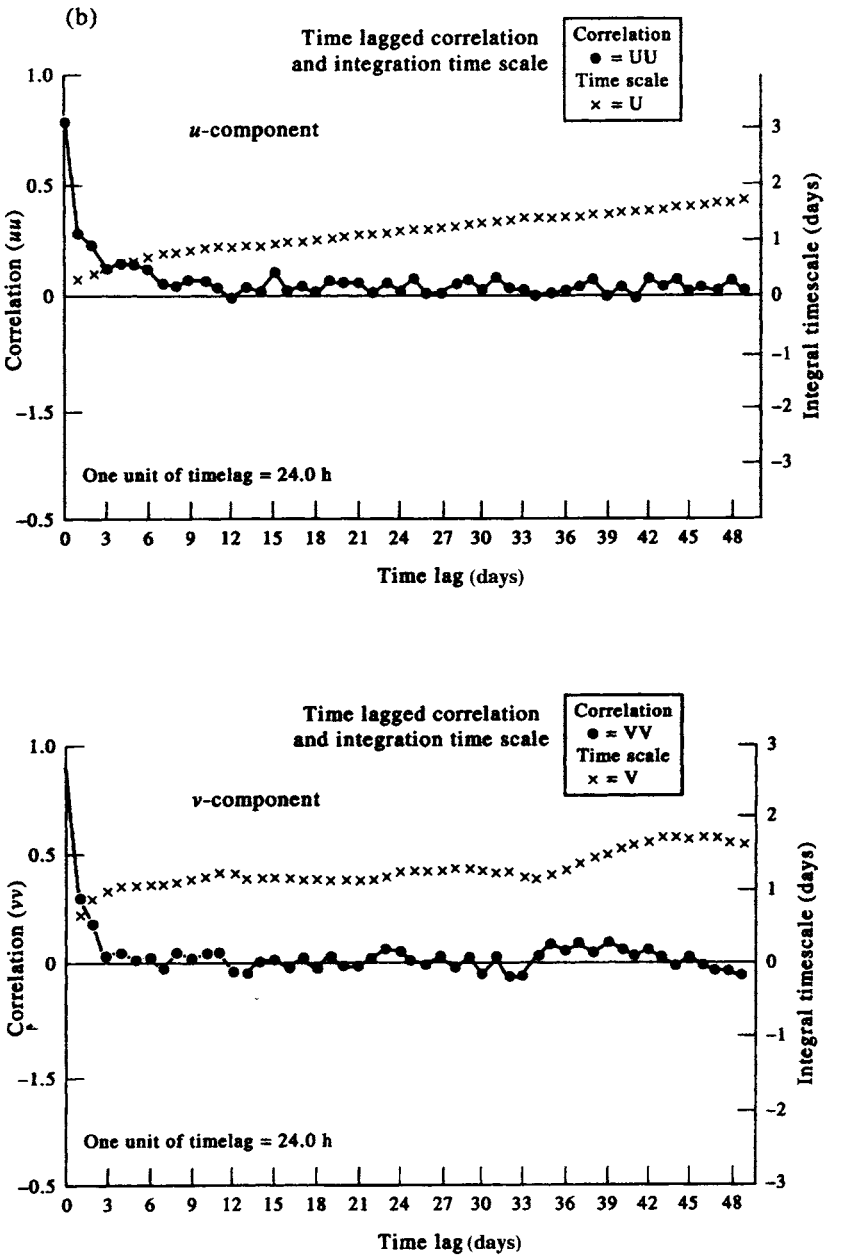

Figure 3.14. Autocovariance functions and corresponding integral time scales for zonal (u) and meridional (v) velocities of satellite-tracked drifter deployed to the south of the Aleutian Islands in the northeast Pacific (see insert) and covering the period 13 November 1991 to 30 July 1993 based on six-hourly sampling interval. (Courtesy of Adrian Dolling.)

region with which they have no direct relationship. Some of these errors, such as oceanographic stations on land, are easily detected, while others are less obvious. Another example of such errors would be biases in a group of measurements due to the application of incorrect sensor calibrations or undetected instrument malfunctions.

Again, the new data would be inconsistent with other existing measurements of the same phenomenon. Our goal is to remove or correct such errors in order to make the data set as self consistent as possible. If we know the history of the data, meaning the details of its collection and reduction, we may be in a better position to understand the sources of these errors. If we have received the data from another source, or are looking at archived data, we may not have available the necessary details on the "petigree" of the data and may have to come to some rather arbitrary decisions regarding its reliability. Considering the widespread use of computer-linked data banks, this is not a trivial problem. The question is how to ensure the necessary quality control yet ensure rapid dissemination and accessibility to data files.

3.16.1 Identifying and removing errors

There are two important axioms to follow when dealing with large erroneous values or "spikes":

(1) To identify the large errors, it is necessary to examine all of the data in visual form and to get a "feel" for the data;
(2) When large errors are encountered, it is usually best to eliminate them all together rather than try to "correct" them and incorporate them back into the data set.

Of course, care must be taken not to reject important data points just because they don't fit either the previous data structure or our preconceived notion of the process. A good example is the determination of heat transport in the South Atlantic. Bennett (1976) suggested that the oceanic heat transport in this ocean is directed toward the equator, contrary to the widely accepted notion that oceanic heat transports are generally poleward. Stommel (personal communication) noted that, in his tabulation of property fluxes for the South Atlantic, Wüst (1957) conspicuously left out the flux of heat while treating other less easily computed transports such as those of nutrients and oxygen. Through an exchange of letters with a former student of Wüst's, Stommel learned that the heat content calculation indeed showed that heat is transported equatorward. Wüst considered this to be the wrong direction and the results were not published along with the other flux values. The point of this story is to illustrate the way in which our prejudice can lead us to reject significant results. In such cases, there is no hard rule as to how this decision is made and a great deal of subjectivity will always be inherent in this level of data interpretion.

The need to examine all the data to detect errors presents a difficult task because of the large numbers of values and the difficulty of looking at unprocessed data. In this case, it is more important to think of ways in which we can present the data so as to ask and answer the questions regarding consistency of the measurements. A compact over-view of all the data is the best solution. This presentation may be as simple as a scatter diagram of the observations versus some independent variable, or a scatter diagram relating two concurrently measured parameters. While scatter diagrams cannot be used to to resolve visually individual points, they do reveal groupings of points which relate to the physical processes expressed by the data. As an example, consider a temperature–salinity scatter diagram (Figure 3.15) computed using a large number of hydrographic data collected from bottle casts. Here, the groups of dots labelled "*a*", "*b*" refer to different water masses present in the 5° square 35–40°N, 15–20°W where the data were collected. The data labelled "*c*" clearly represent a distinct

water mass since the points lie along a line divergent from the rest of the scatter values. If we look at other similar *TS* scatter plots, we recognize that this line is consistent with the *TS* relationship from a corresponding square at this same longitude but south of the equator. Thus, it is likely that the latitude recorded was incorrect and that these data are simply misplaced. We correct this by eliminating the points "*c*" from our square. However, we can't be sufficiently confident of our assumption to add the points to the other square even though the data coverage there is not very good.

Often it is not possible to develop a simple summary presentation of all the data. In the case of current meter data, a time-series presentation is the most appropriate way of looking at the data. As noted by Pillbury *et al.* (1974), error detection using this technique is very time consuming. They note that this procedure can be used successfully for speed, pressure, salinity, and temperature but not for direction, which varies widely. This is due to the fact that direction is limited to the range 0–360° and shows no extreme values. Because of the wrap-around (2π discontinuity) problem, in which $0° = 360°$ (or $-180° = +180°$), direction records tend to be very "spiky", especially in regions of strong tidal flow. A scatter diagram of speed versus direction can be used to detect systematic errors between the speed and direction sensors and to pinpoint those times when the current speed is below the threshold recording level of the instrument. This would be displayed by the direction readings at speeds below threshold and would be easier to identify on the scatter plot than in the individual time series. The only way around the problem with the direction channel is to transform the recorded time series of speed and direction (U, θ) to orthogonal components of velocity (u, v). In particular, separate plots of the east–west (u) and the north–south (v) velocity components (or alongshore and cross-shore components for data collected near the coast) quickly reveal any erroneous values in the data (Figure 3.16).

Pillsbury *et al.* (1974) report that, for Aanderaa RCM4 and RCM5 current meters, there are several sources of large errors. We will discuss these as typical of the errors inherent in moored current meter data since many of these instruments remain in use. One source of error is the current meter's encoder which might encounter a small electrical resistance. The probability of this occurring is considered small. A more likely error is due to nonuniformity in the 1/4-in magnetic tape which may have variations in the coating or carry residual magnetism. The tape transcriber is also a possible error source since it occasionally drops a bit. An error particular to the speed parameter where the speed is seen to abnormally increase, may be caused by nonuniformities in the speed potentiometer winding. A less frequent error type is that associated with clock and trigger malfunctions. Instances have been observed where a meter has cycled several times in rapid succession or conversely missed one or more cycles. These problems are addressed here under the section on timing errors. Direction errors are due to mechanical failures in the compass itself. In some cases the compass needle failed to contact the resistance ring around the compass while in others direction readings in one range all were recorded in a different range. Many of these compass problems were apparent in the raw data while others were only discovered later by looking at the direction histograms.

Other problems with Aanderaa RCM4/5 current meters have been noted over the years. These can be minimized if the following protocol is observed (assuming that the instrument is operational and calibrated):

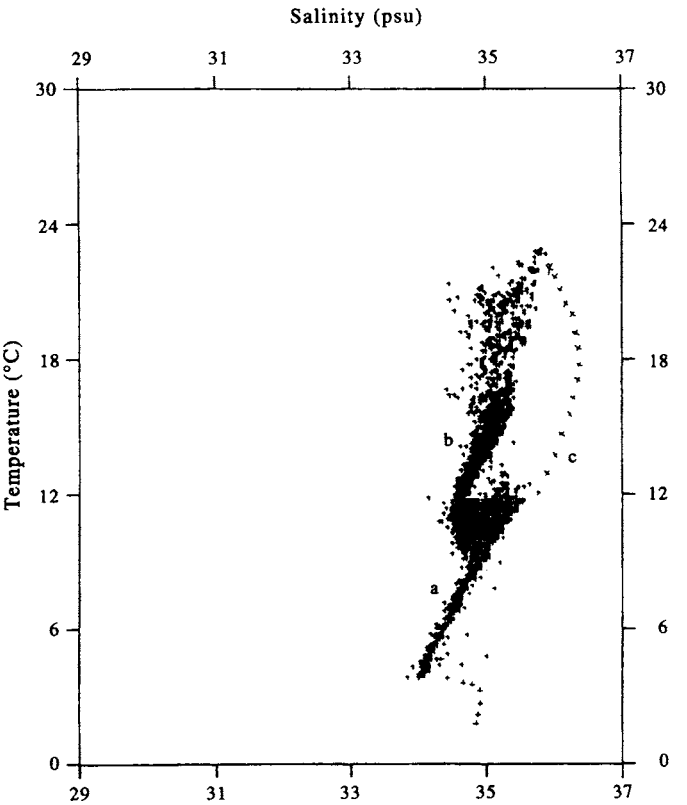

Figure 3.15. TS relationship computed using a large number of hydrographic data collected from bottle casts. Groups labeled "a", "b" refer to different water masses present in the 5° square (35–40°N, 15–20°W) where the data were collected. The data labeled "c" clearly represent a distinct water mass since the points lie along a line divergent from the rest of the scatter values.

(1) Use a new nonmagnetic battery and load test with a 100 ohm resistor to ensure that it meets the manufacturer's specification. Keep in mind that battery amp-hours decrease with decreasing water temperature.
(2) Do not overfill the supply spool with magnetic tape. Leave a 2 mm space so that the tape will not spill off the spool and jam the mechanical mechanism when the instrument is tilted or laid on its side.
(3) Check the tape take-up spool clearance between pinch-rollers spring, circlip, and frame. Spin spool by hand. Check for space between the feed spool and pressure sensor (if installed). Wrap 20 turns of leader on the take-up spool and check the clutch tension.
(4) Check that both spool nuts are in place and do not over-tighten. Do not over-tighten the nylon rotor pivot screw.
(5) Ensure that no ferrous metal screws are used near the compass. Replace these with stainless steel or brass. Also, do not use a ferrous bar to balance the direction vane—it may be close enough to cause the compass to "stick" and ruin the directional data.

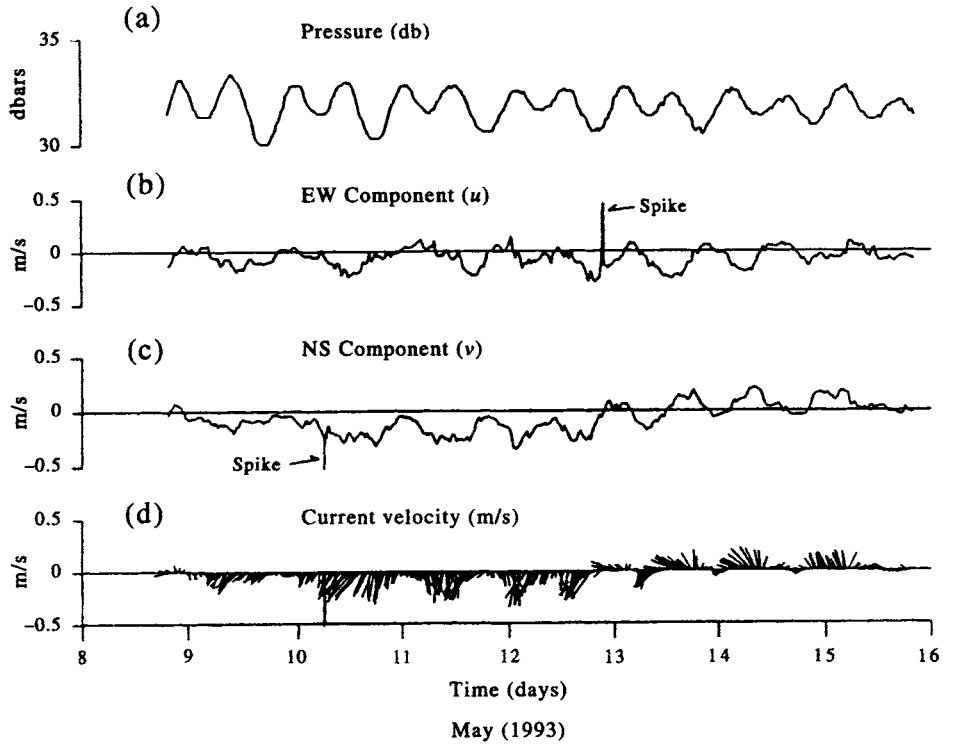

Figure 3.16. A plot of hourly data obtained from an Aanderra RCM4 current meter moored at 30 m depth data in 250 m of water near the entrance to Juan de Fuca Strait (48° 3.3'N, 125° 18.8'W) during the period 8–16 May 1993. (a) Ambient pressure (≈ instrument depth in meters); (b) East–west (u) component of velocity (m/s); (c) North–south (v) component of velocity (m/s); (d) Velocity stick vector (m/s). Erroneous current velocity values ("spikes") stand out in the (u, v) records. Flow consisted of moderate tidal currents superimposed on a surface estuarine outflow that weakened around May 13.

(6) Inspect the O-ring for cuts or nicks and do not trap loose wiring under the ring seat when closing the case. Leakage of small amounts of water to the bottom of the instrument case can cause electrical malfunctions when instrument tilts.
(7) Do not jam a spinning rotor with tissue paper or other material prior to deployment. It is better to shield rotor from wind while on deck. Too often the instrument is recovered with the material still jammed in place.
(8) It is essential to hand-record accurate times for the first and last data records. Make sure the time zone is recorded. Record the time the instrument enters the water on deployment and leaves the water on recovery. More problems can be linked to poor bookkeeping than any other cause.
(9) Spin the rotor in multiples of 24 times (or some multiple of four) to ensure that sampling interval and rotor counter switch (if applicable) are correct.

Another standard method for isolating large errors is to compute a histogram of the sample values. This amounts to completing step 1 in a goodness of fit calculation since a histogram is nothing more than a diagram showing the frequencies of occurrence of sample values. While this is a very straightforward procedure some care must be taken in selecting the parameter intervals, or bins, over which the sample frequencies are calculated. If the bins are too large, the histogram will not resolve the character of the

sample PDF and the effects of large error values will be suppressed by being grouped with more commonly occurring values. On the other hand, if the bins are too narrow, individual values take on more influence and the resulting distribution will not appear smooth. This makes it difficult to "see" the real shape of the distribution.

The use of a histogram in locating large errors is that it readily identifies the number of widely differing values that occur and shows whether these divergent values fit into the assumed PDF for the assumed variable. In other words, we can not only see how many values ("outliers") differ widely from the mean values, but also determine if the number of large values in the sample is consistent with the expected distribution of large values for the population. Thus, we have an added guideline for deciding whether the sample values should be retained or eliminated for subsequent analysis. Both PDFs and histograms use visual means of detecting large error values. It is possible to use more automated and objective techniques, such as eliminating all values that exceed a specified standard deviation (e.g. $\pm 3\sigma$). However, these approaches have the weakness that they must first consider all data points, including the extreme values, as valid in order to determine decision levels for selecting or rejecting data. Here, we could use an iterative process in which the values outside the accepted range are omitted from each subsequent recalculation of the mean and standard deviation, until the remaining data have near constant statistics with each new iteration. Large errors, which are usually easy to spot using visual editing techniques, should be removed before proceeding to a more objective step involving the detection of less obvious random deviations. An objective technique for identifying outlier values is to compute a function which selects extremes of the population such as the first derivative of the measured variable with respect to an independent parameter. An example would be a time series of temperature measured from a line in a satellite image. After the extreme gradients are identified in the first derivative calculation, there is still the question of how widely the extremes should be allowed to differ from the rest of the population and whether a value should be considered as an error value or as simply as a maximum (or minimum) of the process being observed.

In making such a decision, it is necessary to have an estimate of the variability of the process. As discussed above, the dispersion of the population distribution is best represented by the variance or the standard deviation. If we are dealing with a normal population, we know that the standard deviation specifies the spread of the distribution and that 66% of the population values lie within $\mu \pm \sigma$ while 95% of these values are in the interval $\mu \pm 2\sigma$. Beyond $\mu \pm 3\sigma$ lie only 0.26% of the total frequency of occurrence, leaving 99.74% within this interval. Thus, it is again a matter of probabilities and significance level; and we must choose at what level we will reject deviations from the mean as errors. If we choose to discard all measurements beyond 2σ, we will have retained 95% of the sample population as our new sample population for which we will repeat out estimation of the statistics. This suggests that we will make our statistical estimate twice; first to decide what data to retain, and second to make statistical inferences about the behavior exhibited by the revised sample data. It is customary to use a much coarser sub-sampling interval, or to use broadly smoothed data, to compute the initial sample standard deviations for the purposes of editing the data. For our *TS* curve example (Figure 3.15), we might initially have used a computational interval of 1 or 2°C to compute a standard deviation for the first-stage editing and then have used the newly defined sample population (original sample minus large deviations > 2°C) to recalculate the mean and standard deviation with a resolution of 0.1°C, closer to the measurement accuracy for reversing thermometers. In statistical analysis we should

not expect to exceed the inherent accuracy and resolution of our data. Modern computing facilities, and even pocket calculators, make it tempting to work with many decimal places despite the fact that higher place values are not at all representative of the ability of the instrument to make the oceanic measurement.

A form of two-step editing is used in the routine processing of CTD data which is typically sampled at $\approx$ 25 samples/s per channel ($\approx$ 25 Hz/channel). Since these instruments produce many more data than we are capable of examining, both smoothing and editing procedures are often built into the routine processing programs. The steps involved with processing calibrated CTD data at the Institute of Ocean Sciences are as follows:

(1) Write the data to a file for display on a computer screen using an interactive editing program written for the particular data set.
(2) Examine all data for a given set of parameters by displaying the data simultaneously on a computer monitor; as a consistency check, it is important to know if large errors in one paramater such as temperature, are associated with some real feature in another parameter, such as salinity.
(3) With the cursor, eliminate erroneous values collected near the ocean surface where the probe rises in and out of the water with the roll of the ship.
(4) Using the file in (3), calculate the pressure gradient versus depth for the data and eliminate those data values for which the depth is decreasing with time for a downcast and increasing with time for an upcast (wave action eliminator).
(5) Using the file of (4), produce a hardcopy plot of the entire profile plus an expanded version for the upper ocean (say 0 to 300 m depth).
(6) On the hardcopy, "flag" erroneous values and irregularities in all data channels.
(7) Use the interactive screen display to eliminate "bad" data identified in (5). If gaps between data points are small, linearly interpolate between adjacent values.
(8) Smooth the edited file by averaging values over a specified depth range. Typically, 1-m averaged files are generated for profile data and 1-s averaged files for time-series data.

Because of improved CTD technology in recent years, step (8) is often conducted first. This step is then eliminated or replaced with a larger averaging interval such as 5 m.

Fofonoff *et al.* (1974) used a 1/2-s average (15 scans) to smooth the measured pressure series. From this smoothed set, a 10th decibar pressure series was generated. Even with the smoothing, the pressure was oversampled, with roughly two observations for each pressure interval. The goal of this computation was to produce a uniform pressure series that could be used to generate profiles of T and S with depth. Processing routines could be added that first sorted out spurious extreme T and S values, based on a running mean standard deviation, and which ensured that the pressure series was monotonically increasing. This would correct for small variations in the depth of the probe due to ship motion or strong current shear. Also, in making these editing decisions we should always keep in mind the instrument characteristics and not discard data well within the noise level of the measurement system.

When editing newly collected data, we should always consider what is already known from similar, or related measurements in order to detect obvious errors. A typical example is the use of *TS* curves to evaluate the performance of sample bottles in a hydrocast. Since *TS* curves are known to remain relatively stationary for many areas, previously sampled *TS* curves for an area can be used to locate data points that may have been caused by the erroneous performance of a water sampler; these are generally due to inadvertant bottle

"trips" in which the sampler likely closed before or after the desired depth was achieved. Prior *TS* curves also have served as a means of interpolating a particular hydrocast or perhaps providing salinities to match measured temperatures. This approach is limited, however, to those areas and those parameters for which a sufficient number of existing observations are available to define the mean state and variability. In many areas, and for many parameters, information is too limited for existing data to be of any real use in evaluating the quality of new measurements. As a matter of curiosity, it would be interesting to determine the numbers of deep hydrocast data that were unknowingly collected at hydrothermal venting sites and discarded because they were "erroneous". Anomalously high temperatures would be difficult to justify if one did not know about hydrothermal circulation and associated buoyant plumes.

In contrast to large accidental errors, which lead to large offsets or systematic biases, random errors are generally small and normally distributed. These errors often are the result of inaccuracies in the instrumentation or data collection procedures and therefore represent the limit of our ability to measure the desired variable. Added to this is our inability to completely resolve the inherent variability in a particular parameter. This too may be a limitation of our instrument or of our sampling scheme. In either case, when we cannot directly measure a scale of oceanic variability that contributes to the alias of our measurement, the variability will form part of the uncertainty in the final calculated value.

The theory of random errors is well established (Scarborough, 1966). The fundamental approach is to treat the errors as random numbers with a normal PDF. Basic to this assumption is that positive and negative errors of the same size occur in about equal number and tend to cancel each other. This suggests that the appropriate way to treat data containing random errors is in terms of mean-square (MS) and root-mean-square (RMS) values. Another fundamental assumption is that the probability of an error occurring depends inversely on its magnitude; thus, small errors are more frequent than large ones. Following the first of these two assumptions, the PDF of the random errors might be written as

$$p(\varepsilon_x) = f(\varepsilon_x^2) \tag{3.16.1}$$

where p is the PDF of the errors ε_x. The second characteristic requires that the probability decreases with increasing ε_x so we can write for any real constant, k

$$p(\varepsilon_x) = C \exp(-k^2 \varepsilon_x^2) \tag{3.16.2}$$

Using the fact that the integral under the curve of any PDF is unity, we solve for C and get

$$p(\varepsilon_x) = \frac{k}{\sqrt{\pi}} \exp(-k^2 \varepsilon_x^2) \tag{3.16.3}$$

This expression is known as the probability equation or the error equation. A graph of the function gives the normal or Gaussian probability curve. The term k is a constant called the *index of precision* and sets both the amplitude and the width of the normal curve. As k increases, the normal curve becomes narrower and the errors get smaller, making the measurement more precise. (This description applies only for small random errors and not to systematic errors.)

3.16.2 Propagation of error

Suppose we have a quantity, F, which is calculated from a combination of a number (n) of independently observed variables. For example, F might be oceanic heat transport computed from independent velocity and temperature profiles, x. We can estimate the combined random error of F as the sum of squared errors of the individual variables provided that the errors are independent of the variables and that they are all normally distributed. As a simple example, let F be a linear combination of our measurement variables, x

$$F = a_1 x_1 + a_2 x_2 + \ldots + a_N x_N \tag{3.16.4}$$

where $a_1, \ldots, a_N$ are constants. The inverse of the squared error or *index of precision* (H) of F can be written

$$\frac{1}{H^2} = \frac{a_1^2}{h_1^2} + \frac{a_2^2}{h_2^2} + \ldots + \frac{a_N^2}{h_N^2} = \sum_{i=1}^{N} \frac{a_i^2}{h_i^2} \tag{3.16.5}$$

where h_i is the error for the ith measurement, x_i.

A more generalized formula for error calculations for arbitary F for which the contibuting variables are uncorrelated is

$$\begin{aligned} \frac{1}{H^2} &= \frac{(\partial F/\partial x_1)^2}{h_1^2} + \frac{(\partial F/\partial x_2)^2}{h_2^2} + \ldots + \frac{(\partial F/\partial x_N)^2}{h_N^2} \\ &= \sum_{i=1}^{N} \frac{(\partial F/\partial x_i)^2}{h_i^2} \end{aligned} \tag{3.16.6}$$

where partial derivatives $\partial F/\partial x_i$ are obtained from Taylor expansions of the function F in terms of the independent variables x_i. To convert this expression to one in terms of relative errors, we use the fact that

$$\frac{1}{h^2} = \frac{r_e^2}{\rho^2} \tag{3.16.7}$$

where r_e is the corresponding relative error and $\rho = 0.4769$ is a constant obtained from the error equation (3.16.3). Using this definition we can write our final error as

$$R_e = \left[(\partial F/\partial x_1)^2 r_1^2 + (\partial F/\partial x_2)^2 r_2^2 + \ldots + (\partial F/\partial x_N)^2 r_N^2\right]^{1/2} \tag{3.16.8}$$

In this form, R_e is really only the RMS error that describes the equivalent combined error in the equation of interest. This Taylor expansion of the contributing error terms is known as the *propagation of errors formula.* It is limited to small errors and uncorrelated independent variables. Since these principles apply only to small random errors, it is necessary to use some data editing procedure to remove any large errors or biases in the measurements before using this formula. By using a mean-square formulation, we take advantage of the fact that small random errors can be expected to often cancel each other resulting in a far smaller mean-square error than would result if the measurement errors were simply added regardless of sign to yield a maximum "worst case error". The primary application of equation (3.16.8) is in determining the

overall error in a quantity derived from a number of component variables all with measurement errors. This is a situation common to many oceanographic problems.

A more complicated propagation of error formula is needed if there is a nonzero correlation between the independent variables, x. In this case, we must also retain the covariance terms in any Taylor expansion of the small error terms. For example, the density ρ is a function of both temperature T and salinity S so that the errors (variances) in density σ_ρ^2 can be related to the measurement errors in temperature σ_T^2 and salinity σ_S^2 by

$$\sigma_\rho^2 = (\partial\rho/\partial T)^2\sigma_T^2 + (\partial\rho/\partial S)^2\sigma_S^2 + 2[(\partial\rho/\partial T)\cdot(\partial\rho/\partial S)]C(T,S) \qquad (3.16.9)$$

where $C(T, S)$ is the covariance between temperature and salinity fluctuations. Only when $C(T, S) = 0$ do we get the result (3.16.8). An example of a detailed error calculation is the measurement of flow through trawl nets towed at various angles through the water column is given in Burd and Thomson (1993).

3.16.3 Dealing with numbers: the statistics of roundoff

Since we must represent all measurements in discrete digital form, we are forced to deal with the consequences of numerical roundoff, or truncation. The problem results from the limitations of digital computing machines. For example, the irrational fraction 1/3 is represented in the computer as the decimal equivalent 0.3333 ... 3 with an obvious roundoff effect. This may not seem to be a problem for most applications since most computers carry a minimum of eight decimal places at single precision. The large number of arithmetic operations carried out in a problem lasting for only a few seconds of computer processing time can, however, lead to large errors in due to roundoff and truncation errors. The case of greatest concern is when two nearly identical numbers are subtracted, requiring proper representation to the smallest possible digit. Such differences can easily occur unknowingly in a complicated computational problem. Rather than discuss procedures for estimating this roundoff error, we will discuss the nature of the problem and emphasize the need to avoid roundoff.

General floating-point values (decimal numbers) in a computer follow closely the so-called "scientific notation" and are represented as a mantissa (to the right of the decimal point) and an exponent (the associated power of 10). For example, in a three-digit system, the number 64.282 would be represented as 0.643×10^2 where the roundoff is accomplished by adding five in the thousands' decimal place and then truncating after the third digit. This process of rounding off results in a slight bias because it always rounds up when there is a 5 in the least significant digit. A way to overcome this bias is to use the last digit retained to determine whether to round up or down when the next digit is exactly 5. This rule, which leads to the least possible error, is to round-up if the next to the last digit retained is odd and to round down when it is even. This procedure can be summarized as follows. When rounding a number to k decimals:

(1) if the $k + 1$ decimal is 0, 2, 4, 6, 8 then the k decimal is unchanged;
(2) if the $k + 1$ decimal is 1, 3, 5, 7, 9 then the k decimal is increased by 1.

This system of rounding-off will result in errors that are generally less than 0.5×10^{-k} and maximum roundoff errors of 0.6×10^{-k}. In most applications, the effect of this

roundoff bias is too small to justify the added numerical manipulation required to implement this even–odd roundoff scheme.

In computing systems, floating-point numbers are handled in a binary representation having 24 bits (wordlength is 32 bits but eight bits are used for the exponent) which results in seven significant decimal digits. Called *single precision*, this level of accuracy is adequate for many computations. For those problems with repeated calculations, and the subsequent high probability of differencing two nearly identical numbers, a *double-precision* representation is used which has 56 binary bits leading to 16 significant decimal digits. Roundoff, in the case of double precision, results in very small biases which can be completely ignored for most applications. Another approach to the problem of roundoff errors is to consider them to be random variables. In this way, statistical methods can be applied to better understand the effects of roundoff errors. Consider the roundoff of a single number x; for this number, all numbers occurring in the interval $x_o - 1/2 < x < x_o + 1/2$ (measured in units of the last digit) become that number. Thus, the roundoff has a uniform probability distribution in the last digit. We can write the corresponding probability density function $f(x)$ for x as

$$f(x) = \begin{cases} 1 & (x_o - 1/2,\ x_o + \tfrac{1}{2}) \\ 0, & \text{elsewhere} \end{cases} \tag{3.16.10}$$

and note that

$$\int_{-\infty}^{\infty} f(x)\,dx = 1 \tag{3.16.11}$$

The most common measures of a PDF are its first two moments, the mean and variance. The mean of $f(x)$ in equation (3.16.10) is x_o and the variance is

$$V[f(x)] = \sigma^2 = \int_{x_o-1/2}^{x_o+1/2} [x - x_o] f(x)\,dx = \int_{-1/2}^{+1/2} x'^2\,dx' = \frac{1}{12} \tag{3.16.12}$$

Experimental tests have verified the uniform distribution of roundoff in computer systems. In fact, computers generate random numbers by using the overflow value of the mantissa.

We can represent roundoff as an additive random error (ε) superimposed on the true variable (x). In this case, we can write the computer representation of our variable (which we assume is free from measurement and sampling errors) as $x + \varepsilon$. For a floating-point number system, it is better to use

$$x(1 + \varepsilon);\ |\varepsilon| < \tfrac{1}{2}(10^{-2}) \tag{3.16.13}$$

for the variable with roundoff error ε. This formulation has the effect of focusing attention on the consequences of roundoff for every application in which it appears. For example, the product

$$x_1(1 + \varepsilon_1)x_2(1 + \varepsilon_2) = x_1x_2(1 + \varepsilon_1 + \varepsilon_2 + \varepsilon_1\varepsilon_2) \tag{3.16.14}$$

demonstrates how roundoff propagates during multiplication. Generally, the product

$\varepsilon_1\varepsilon_2$ is sufficiently small to be ignored. However, in the above multiplication we must include the roundoff for this operation, whereby (3.16.14) becomes

$$\begin{aligned} x_3(1+\varepsilon_3) &= x_1x_2(1+\varepsilon_1+\varepsilon_2+\varepsilon) \\ |\varepsilon| &< \tfrac{1}{2}(10^{-2}); \quad \varepsilon_3 = \varepsilon_1+\varepsilon_2+\varepsilon \end{aligned} \tag{3.16.15}$$

Similar error propagation results are found for other arithmetical operations.

We can extend this to a generalized product

$$y_1(1+\varepsilon_1)y_2(1+\varepsilon_2)\ldots y_N(1+\varepsilon_N) \tag{3.16.16}$$

which becomes

$$y_1y_2\ldots y_N[1+(\varepsilon_1+\varepsilon_2+\ldots+\varepsilon_N)] \tag{3.16.17}$$

By the central limit theorem, the sum of n independent random numbers (the roundoff errors) approaches a normal distribution. The effect for the other operations is much the same; therefore, while individual roundoffs are from a uniform distribution, the result of many arithmetic roundoff operations tends toward a normal distribution. This also can be demonstrated experimentally.

As stated earlier, we will generally ignore roundoff as a source of error in the processing and analysis of oceanographic data. The above discussion has been presented here to make the reader aware of potential problems and provide some familiarity with the problems of using computing systems. In most data applications, the effects of roundoff error are small enough to be ignored. Only in the case of recursive calculations, where each computation depends on the previous one, do we anticipate large roundoff errors. This is usually a problem for numerical modelers who must deal with the repeated manipulation of computer-generated "data". In cases where roundoff errors are of some consequence, statistical methods can be used in which the errors can be treated as variables from a normal population.

3.16.4 Gauss–Markov theorem

The term *Gauss–Markov process* is often used to model certain kinds of random variability in oceanography. To understand the assumptions behind this process, consider the standard linear regression model, $y = \alpha + \beta x + \varepsilon$, developed in the previous sections. As before, α, β are regression coefficients, x is a deterministic variable and ε a random variable. According to the Gauss–Markov theorem, the estimators α, β found from least squares analysis are the *best linear unbiased estimators* for the model for the following conditions on ε:

(1) The random variable ε is independent of the independent variable, x;
(2) ε has a mean of zero; that is $\mathrm{E}[\varepsilon] = 0$;
(3) Errors ε_j and ε_k associated with any two points in the population are independent of one another; the covariance between any two errors is zero; $C[\varepsilon_j, \varepsilon_k] = 0, j \neq k$;
(4) ε has a finite variance $\sigma_\varepsilon^2 \neq 0$.

The estimators are *unbiased* since their expected value equals the population values (given 1 and 2) and they are *best* in that they are efficient (if 3 and 4 hold true), the variance of the least-squares estimators being smaller than any other linear unbiased estimator. A further assumption that is often made is that the errors, ε, are normally

distributed. In this case, the estimators of α, β, and μ using the least-squares requirements are identical to the estimators resulting from the use of maximum-likelihood estimation. This assumption, combined with the four previous assumptions, provide the rationale for the least-square procedure.

3.17 INTERPOLATION: FILLING THE DATA GAPS

Most analysis procedures used in the physical sciences are designed for comparatively long and densely sampled series with equally spaced measurements in time or space. The wealth of information on time-series analysis primarily applies to regularly spaced and abundant observations. There are two main reasons for this: (1) the mathematical necessity for long, equally-spaced data for the derivation of statistically reliable estimates from modern analytical techniques; and (2) the fact that most modern measurement systems both collect and store data in digital format. Spectral estimates, for example, improve with increased duration of the data series in the sense that one is able to cover an increasing range of the dominant frequency constitutents that make up the record. Digital sampling systems are considerably more economical than analog recording systems in that they cut down on storage space, power consumption and postprocessing effort.

3.17.1 Equally and unequally spaced data

Electronic systems now provide data at regularly spaced sampling increments. Unfortunately, such systems usually operate autonomously and any type of equipment failure generally leads to either data *gaps* or a premature termination of the record. The failure of electronic data logging systems is but one source of gappy records in physical oceanography. Because of their very nature, shipborne measurements are a source of gappy records. Oceanographic research vessels are expensive platforms to operate and must be used in an optimal fashion. As a consequence, it is often impossible to collect observations in time or space of sufficient regularity and spacing to resolve the phenomenon of interest. Efforts are usually made to space measurements as evenly as possible but, for a variety of reasons, station spacings are often considerably greater than desired. Weather conditions, as well as ship and equipment problems, almost invariably lead to unwanted gaps in the data set. Sometimes equipment failures are not detected until the data are examined in the laboratory. In addition, editing out errors produces unwanted gaps in the data record.

The gap problem is even more severe when one is analyzing historical data or data collected from "platforms of opportunity." Historical data are a collection of many different sampling programs all of which had different goals and therefore very different sampling requirements. By its very nature, such collections of data will necessarily be irregularly spaced and variable in terms of accuracy and reliability. Further editing, dictated by the goals of the historical data analysis project, will add new gaps to the set of existing data series.

Monitoring stations, ships of opportunity, and satellite measurements frequently produce data series that are unevenly spaced. The geographic distribution of monitoring stations (e.g. Pacific island sea-level stations) is far from uniform in terms of the spacing between stations. Thus, while the data series collected at each

station, may themselves consist of evenly and densely spaced measurements in time, the space intervals between stations will be highly irregular. Open ocean buoys and current meter moorings also fit this classification of densely and evenly spaced temporal observations at widely and often irregularly spaced locations. Here again, any failure in the recording system, whether minor or catastrophic, will lead to gaps in the time-series record. Often these gaps are quite large since unplanned recovery efforts are required to correct the problem. Such a correction effort assumes the telemetering of data which is at present not widely done. Failures of on-board recording systems must wait until the scheduled servicing of the instrument which may then result in relatively large data gaps.

At the other end of the sampling spectrum satellite observing systems provide dense and evenly spaced measurements that are often very irregular in time. A familiar source of temporal gaps, in infrared image series, is cloud cover. Both occasional and persistent cloud cover can interrupt a sequence of images collected to study changes of sea surface temperature. The effects of cloud cover apply also to satellite remote sensing in the optical bands. In addition to the cloud-cover problem, there are often problems with the on-board satellite sensing systems or associated with the ground receiving station that lead to gaps in time series of image data. Microwave sensing of the surface is not as sensitive to cloud attenuation but it is subject to sensor and ground-recording failure problems.

Platforms of opportunity (usually merchant ships) produce uniquely irregular sets of measurements. Most merchant ships repeat the same course with minor adjustments for local weather conditions and season. A seasonal shift in course is generally seen at higher latitudes to take advantage of great circle routes during times of better weather. A return to lower latitudes is seen in winter data as the ships avoid problems with strong storms. Added to the seasonal track changes is the nature of the daily sampling procedure. Usually the ship takes measurements at some specified time interval which, due to variations in ship track, ship speed and weather conditions, may be at very different positions from sailing to sailing. Thus, the merchant ship data will be irregular in both space and time. Systems that operate continuously from ships of opportunity (e.g. injection SST) overcome this problem. These continuous measurements, however, are still subject to variations in ship track.

The net result of all these measurement problems is that oceanographers are often faced with short records of unequally spaced data. Even if the records are long they are often gappy in time or space. It is, therefore, necessary to interpolate these data to produce series of evenly spaced measurements. While some analysis procedures, such as least-squares harmonic analysis, apply directly to uneven or gappy data, it is more often the case that irregularly spaced data are interpolated to yield evenly spaced, regular data. These interpolated records can then be analyzed with familiar methods of time-series analysis.

Interpolation also may be required with evenly spaced data if the subject dynamics apply to smaller space/time scales than are resolved by the measurements. Thus, the data points that are interpolated produce another set of regularly spaced points with a finer resolution. Many interpolation procedures have been developed that only apply to evenly spaced data.

3.17.2 Interpolation methods

Interpolation techniques are needed for both irregularly spaced and evenly spaced data series. Before deciding which interpolation method is most effective, we need to consider the particular application. A series of appropriate questions regarding the selection of the best interpolation procedures are:

(1) What samples (original data series, derivatives, etc.) should we use?
(2) What class of interpolation function (linear, higher-order polynomial, cubic-spline, etc.) best satisifies the dynamical restrictions of the analysis?
(3) What mathematical criteria (exact data-point matching, least-squares fit, continuity of slopes, etc.) do we use to derive the interpolated values?
(4) Where do we apply these criteria?

Answers to these questions serve as guides to the selection of a unique interpolation procedure.

3.17.2.1 Linear interpolation

The type of interpolation scheme to be employed depends on how many data points we want our interpolation curve (polynomial) to pass through. Increasing the number of points we want our curve to fit, increases the order of the polynomial we need to do the fitting. The most straightforward and widely used interpolation procedure is that of *linear interpolation*. This consists of fitting a straight line between two data points and choosing interpolated values at the appropriate positions along that line. For a data series $y(x)$, this linear procedure can be written as

$$\begin{aligned} y(x) &= y(a) + \frac{x-a}{b-a}[y(b) - y(a)] \\ &= \frac{(b-x)y(a) + (x-a)y(b)}{b-a} \end{aligned} \tag{3.17.1}$$

where $x_{\text{start}} = a$ and $x_{\text{end}} = b$ are the times (positions) of the data collection at the start and end of the sampling increment being interploated, and x represents the corresponding time (position) of the desired interpolated value within the interval $[a, b]$. This is the customary procedure for interpolating between values in most tables. The same formula can be applied to *extrapolation* (extending the data beyond the domain of the observations) where the point x lies beyond the interval $[a, b]$. Equation (3.17.1) is a special case of the Lagrange polynomial interpolation formula discussed in the next section.

3.17.2.2 Polynomial interpolation

If we wish to interpolate between more than two points simultaneously, we need to use higher-order polynomials than the first-order polynomal (straight line) used in the previous section. For example, through three points we can find an unique polynomial of degree 2 (a quadratic); through four points, an unique polynomial of degree 3 (a cubic), and so on. The two methods described below are computationally robust in the sense that they yield reasonable results at most points. Polynomial interpolation techniques such as Vandermonde's method and Newton's method are awkward to program and suffer from problems with roundoff error.

3.17.2.2.1 Lagrange's method

The Lagrange polynomial interpolation formula is a method for finding an interpolating polynomial $y(x)$ of degree N which passes through all of the available data points (x_i, y_i); $i = 1, 2, \ldots, N + 1$. The general form for this polynomial, of which linear interploation is a special case, is given as

$$y(x) = a_o + a_1 x + a_2 x^2 + \ldots + a_N x^N = \sum_{k=0}^{N} a_k x^k$$

$$= \sum_{i=1}^{N+1} \left[y_i \left(\prod_{\substack{k=1 \\ k \neq i}}^{N+1} \frac{x - x_k}{x_i - x_k} \right) \right] \tag{3.17.2}$$

where Π is the product function. Note that in the product function, the ith term—corresponding to the particular data point, x_i, in the denominator—is not included when calculating the product for the term involving x_i. Even though k ranges from 1 to $N + 1$, Π uses only N terms and the final polynomial is of order N, as required.

The goal of the Lagrange interpolation method is to find an Nth degree polynomial which is constrained to pass through the original $N + 1$ data points and which yields a "reasonable" interpolated value for any position x located anywhere between the original data points. To see that the polynomial passes through the original data points, note that the ith product function, Π_i, defined for the data point x_i in the denominator is constructed in such a way that $\Pi_i(x_j; x_i) = \delta_{ij}$ whenever $x = x_j$ is one of the data values (δ_{ij} is the Kronecker delta function). This means that $\Pi_i(x_j; x_i) = 0$ for all x_j except for the specific value $x = x_i$ found in the original data series which matches the term in the denominator. In that case, $\Pi_i(x_j; x_i) = 1$ and $y_i \Pi_i(x_j; x_i) = y_j$

The general polynomial we seek is constructed as a sum of the product functions in equation (3.17.2) which can be expanded to give

$$y(x) = \sum_{i=1}^{N+1} y_i [Q_i(x)/Q_i(x_i)] \tag{3.17.3}$$

in which

$$Q_i(x) = (x - x_1)(x - x_2) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_{N+1}) \tag{3.17.4}$$

is the product of all the factors except the ith one. For any x, (3.17.3) can be expanded to give the interpolating polynomial

$$y(x) = y_1 \frac{(x - x_2)(x - x_3) \ldots (x - x_{N+1})}{(x_1 - x_2)(x_1 - x_3) \ldots (x_1 - x_{N+1})} + y_2 \frac{(x - x_1)(x - x_3) \ldots (x - x_{N+1})}{(x_2 - x_1)(x_2 - x_3) \ldots (x_2 - x_{N+1})}$$
$$+ \ldots + y_N \frac{(x - x_1)(x - x_2) \ldots (x - x_N)}{(x_{N+1} - x_1)(x_{N+1} - x_3) \ldots (x_{N+1} - x_N)} \tag{3.17.5}$$

Note that, for the original data points, $x = x_i$, the polynomial yields the correct output value $y(x_i) = y_i$, as required.

In the Lagrange interpolation method, the calculation is based on all the known data values. If the user wants to add new data to the series, the whole calculation must be repeated from the start. Although the above formula can be applied directly, programing improvements exist that should be taken into account (Press *et al.*, 1992). Use of Neville's algorithm for contructing the interpolating polynomial is more efficient and allows for an estimate of the errors resulting from the curve fit.

As an example of this interpolation method, consider four points (x_i, y_i), $i = 1, \ldots, 4$ given as (0, 2), (1, 2), (2, 0) and (3, 0) through which we wish to fit a (cubic) polynomial. Substituting these values into equation (3.17.5), we obtain

$$y(x) = 2\frac{(x-1)(x-2)(x-3)}{(0-1)(0-2)(0-3)} + 2\frac{(x-0)(x-2)(x-3)}{(1-0)(1-2)(1-3)} + 0 + 0$$
$$= \frac{2}{3}x^3 - 3x^2 + \frac{7}{3}x + 2$$

The resulting third-order curve is plotted in Figure 3.17.

3.17.2.3 Spline interpolation

In recent years, the method that has received the widest general acceptance is the spline interpolation method. Splines, unlike other polynomial interpolations such as the Lagrange polynomial interpolation formula, apply to a series of segments of the data record rather than the entire data series. This leads to the obvious question to ask in selecting the proper interpolation procedure: Do we want a single, high-order polynomial for the interpolation over the entire domain, or would it be better to use a sequence of lower-order polynomials for short segments and sum them over the domain of interest? This integration is inherently a smoothing operation but one must be careful of discontinuities, or sharp corners, where the segments join together. Spline functions are designed to overcome such discontinuities, at least for the lower-order derivatives. It is because discontinuities are allowed in higher-order derivatives that splines are so effective locally. Constraints placed on the interpolated series in one region have only very small effects on regions far removed. As a result, splines are more effective at fitting nonanalytic distributions characteristic of real data. The term "spline" derives from the flexible drafting tool used by naval architects to draw piecewise continuous curves.

Splines have other favorable properties such as good convergence, highly accurate derivative approximation, and good stability in the presence of roundoff errors. Splines represent a middle ground between a purely analytical description and numerical finite difference methods which break the domain into the smallest possible intervals. The piecewise approximation philosophy represented by splines has given rise to finite element numerical methods.

With spline interpolation, we approximate the interpolation function $y(x)$ over the interval $[a, b]$ by dividing the interval into subregions with the requirement that there be continuity of the function at the joints. We can define a spline function, $y(x)$, of degree N with values at the joints

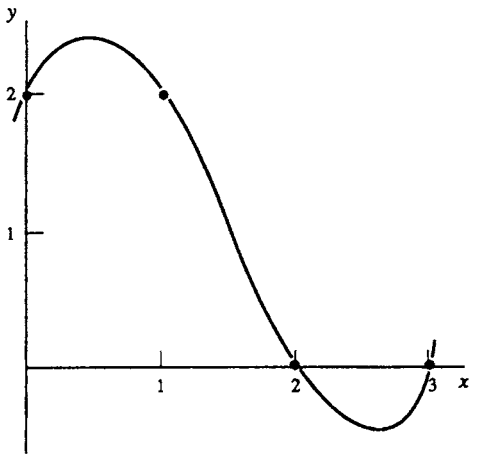

Figure 3.17. Use of Lagrange's method to fit a third-order (cubic) polynomial through the data points (x_i, y_i) *given by (0, 2), (1, 2), (2, 0), and (3, 0).*

$$a = u_0 \leq u_1 \leq u_2 \ldots \leq u_N = b \tag{3.17.6}$$

and having the properties:

(1) In each interval $u_{i-1} \leq x \leq u_i\,(i = 1, m)$, the function $y(x)$ is a polynomial of degree not greater than N.
(2) At each interior joint, $y(x)$ and its first $N - 1$ derivatives are continuous.

The spline function in widest use is the cubic spline ($N = 3$). To give the reader familiarity with the spline interpolation technique, we will develop the cubic spline equations and work through a simple example. Consider a data series with elements (x_i, y_i), $i = 1, \ldots, N$. Since we are working with a cubic spline interpolation, the first two dervivatives $y'(x)$ and $y''(x)$ of the interpolation function, $y(x)$, can be defined for each of the points x_i while the third derivative $y'''(x)$ will be a constant for all x. Here, the prime symbol denotes differentiation with respect to the independent variable x. We write the spline function in the form

$$y(x) = f_i(x);\ x_i \leq x \leq x_{i+1},\ i = 1, \ldots, N - 1 \tag{3.17.7}$$

and specify the following conditions at the junctions of the segments:

(1) Continuity of the spline function:

$$\begin{aligned} f_i(x_i) &= y(x_i) = y_i,\ i = 1, 2, \ldots, N - 1; \\ f_{i-1}(x_i) &= y(x_i) = y_i,\ i = 2, 3, \ldots, N; \end{aligned} \tag{3.17.8a}$$

(2) continuity of the slope:

$$f'_{i-1}(x_i) = f'_i(x_i),\ i = 1, 2, \ldots, N - 1; \tag{3.17.8b}$$

(3) continuity of second derivative:

$$f''_{i-1}(x_i) = f''_i(x_i),\ i = 1, 2, \ldots, N - 1; \tag{3.17.8c}$$

Since $y'''(x) = \text{constant}$, $y''(x)$ must be linear, so that

$$\begin{aligned} f_i''(x_i) &= y_i'' \frac{(x_{i+1} - x)}{x_{i+1} - x_i} \\ &= y_{i+1}'' \frac{(x - x_i)}{x_{i+1} - x_i} \end{aligned} \tag{3.17.9}$$

Integrating twice and selecting integration constants to satisfy the conditions (3.17.8a, b) on $f_i(x_i)$ and $f_{i-1}(x_i)$ gives

$$\begin{aligned} f_i(x) = {} & y_i \frac{(x_{i+1} - x)}{(x_{i+1} - x_i)} + y_{i+1} \frac{(x - x_i)}{(x_{i+1} - x_i)} \\ & - \frac{(x_{i+1} - x_i)^2}{6} y_i'' \left\{ \frac{(x_{i+1} - x)}{(x_{i+1} - x_i)} - \left[\frac{(x_{i+1} - x)}{(x_{i+1} - x_i)} \right]^3 \right\} \\ & - \frac{(x_{i+1} - x_i)^2}{6} y_{i+1}'' \left\{ \frac{(x - x_i)}{(x_{i+1} - x_i)} - \left[\frac{(x - x_i)}{(x_{i+1} - x_i)} \right]^3 \right\} \end{aligned} \tag{3.17.10}$$

which uniquely satisfies the continuity condition for the second derivative but not, in general, for the first derivative (slope). To ensure continuity of the slope at the seams, we expand (3.17.9) by differentiation to get

$$f_i'(x_i) = \frac{(y_{i+1} - y_i)}{x_{i+1} - x_i} - \frac{(x_{i+1} - x_i)}{6} (2y_i'' + y_{i+1}'') \tag{3.17.11a}$$

$$f_{i-1}'(x_i) = \frac{(y_i - y_{i-1})}{x_{i+1} - x_i} - \frac{(x_{i+1} - x_i)}{6} (y_{i-1}'' + 2y_i'') \tag{3.17.11b}$$

We then set (3.17.11a) and (3.17.11b) equal in order to satisfy slope continuity (3.17.8b), whereby

$$\begin{aligned} & (x_i - x_{i-1}) y_{i-1}'' + 2[(x_{i+1} - x_{i-1})] y_i'' + (x_{i+1} - x_i) y_{i+1}'' \\ & = 6 \frac{(y_{i+1} - y_i)}{x_{i+1} - x_i} - \frac{(y_i - y_{i-1})}{x_i - x_{i-1}}, \quad i = 2, \ldots, N - 1 \end{aligned} \tag{3.17.12}$$

which must be satisfied at $N - 2$ points by the N unknown quantities, y_i''. We require two more conditions on the y_i'' which we get by specifying conditions at the end points x_1 and x_N of the data sequence. After specifying these end values, we have $N - 2$ unknowns which we find by solving the $N - 2$ equations. There are two main ways of specifying the end points: (1) we set one or both of the second derivates, y_1'' and y_N'' at the end points to be zero (this is termed the *natural cubic spline*) so that the interpolating function has zero curvature at one or both boundaries; or (2) we set either y_1'' and y_N'' to values derived from equation (3.17.11) in order that the first derivatives of the interpolating function, y_i', take on specified values at one or both of the termination boundaries.

As a general example, we consider the spline solution for six evenly spaced points with the data interval $h = x_{i+1} - x_i$ and function d_i defined in terms of y_i as

$$d_i = \frac{(y_{i+1} - 2y_i + y_{i-1})}{2h^2} \tag{3.17.13}$$

We can write the equations (3.17.10) for these six equally spaced points in matrix form as

$$\begin{pmatrix} 4 & 1 & 1 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 4 \end{pmatrix} \begin{pmatrix} y_2'' \\ y_3'' \\ y_4'' \\ y_5'' \end{pmatrix} = \begin{pmatrix} 12d_2 - y_1''/h \\ 12d_3 \\ 12d_4 \\ 12d_5 - y_6''/h \end{pmatrix} \tag{3.17.14}$$

If we want to specify y_i' rather than y_i'', we need an equation relating both. If the end conditions are not known, the simplest choice is $y_1'' = 0$ (the *natural spline* noted above). Another, and smoother choice (in the sense of less inflection or curvature at the interpolated point) is $y_1'' = 0.05y_2''$. Although spline interpolation is a global, rather than a local, curve (altering a y_i'' or an end condition affects the overall spline), the dominant diagonal terms in equation (3.17.14) cause the effects to rapidly decrease as the distance from the altered point increases.

We should point out the method of splines offers no advantage over polynomial interpolation when applied to either the approximation of well-behaved mathematical functions or to curve fitting when the experimental data are dense. "Dense" means that the number of data points in a subregion is more than an order of magnitude larger than the number of inflection points in the fitted curve and that there are no abrupt changes in the second derivative. The advantage of splines is their inherent smoothness when dealing with sparse data.

As a numerical example of spline fitting, we consider the six-point fitting of the points represented in equation (3.17.14) for the 11 data points in Table 3.17.1. Using a general polynomial fit yields the curve in Figure 3.18. Here, all but one of the first six points lie on a straight line. Due to this single point, the polynomial curve oscillates with an amplitude that does not decrease. In contrast, the spline amplitude (Figure 3.19) for the same 11 values reduces each cycle by a factor of 3.

Often the first or second derivatives of the interpolated function are important. In Figure 3.18, we see that fitting a polynomial to sparse data can result in large, unrealistic changes in the second derivatives. The spline fit to the same points (Figure 3.19) using the end-point conditions $y_1'' = y_N'' = 0$ demonstrates the smoothness of the spline interpolation. In essence, the spline method sacrifices higher-order continuity to achieve second derivative smoothness.

Spline interpolation is generally accomplished by computer routines that operate on the dataset in question. Computer routines solve for the spline functions by solving the equation

$$\sum_{i=1}^{N} [(g(x_i) - y_i)/\delta y_i]^2 = S \tag{3.17.15}$$

where $g(x_i)$ is composed of cubic parabolas

$$g(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \tag{3.17.16}$$

for the interval $x_i \leq x \leq x_i + 1$. The terms δy_i are positive numbers that control the amount of smoothing at each point; the larger δy_i is the more closely the spline fits at each data point. A good choice of δy_i is the standard deviation of the data values.

The S term also controls smoothing, resulting in more smoothing when S increases. As S gets smaller, smoothing decreases and the splines fit the data points more closely.

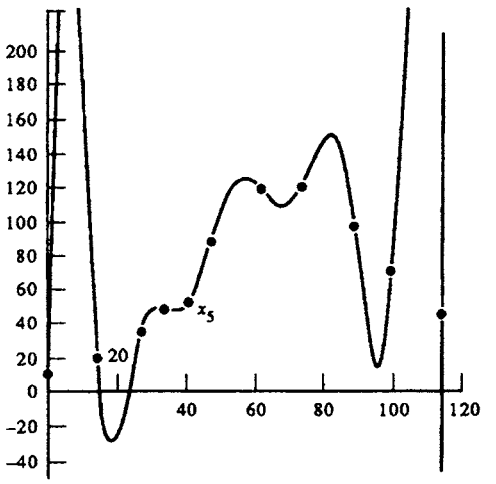

Figure 3.18. A general six-point polynomial fit to the data values in Table 3.17.1. Due to a single point, the polynomial curve oscillates with a amplitude that does not decrease with x.

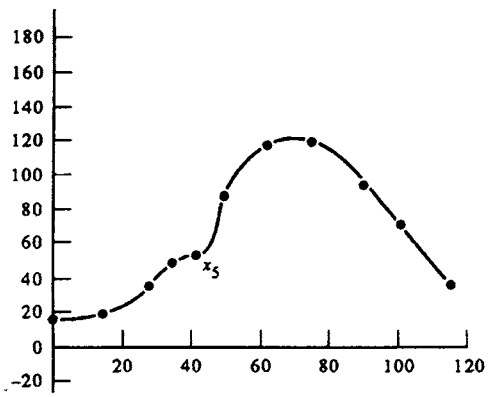

Figure 3.19. Cubic spline fit to the data values in Table 3.17.1. Amplitude of each cycle is reduced by a factor of three compared to Figure 3.18.

When $S = 0$ the data points are fitted exactly by the interpolating spline functions. A recommended value of S is $N/2$, where N is the number of data points. An even smoother interpolation can be achieved using splines under tension. Tension is introduced into the spline procedure to eliminate extraneous inflection points. An iterative procedure is usually used to select the best level for the tension parameter.

3.17.3 Interpolating gappy records: practical examples

Gaps or "holes" occur frequently in geophysical data series. Gaps in a stationary time series are, of course, analogous to gaps in a homogeneous spatial distribution. Small gaps are of little concern and linear interpolation is recommended for filling the gaps. If the gaps are large (of the size of the integral time or space scale), it is generally better to work with the existing short data segments than to "make up" data by pushing interpolation schemes beyond their accepted limitations. For the gray area between these two extremes, one wants to know how large the data loss can be and still

Table 3.17.1. Data pairs (x_i, y_i) used for interpolation schemes in Figures 3.18 and 3.19

i	x_i	y_i
1	0	16
2	14	19
3	27	36
4	33	48
5	41	53
6	48	90
7	62	119
8	74	120
9	89	96
10	99	71
11	114	36

permit reasonable use of standard interpolation techniques and processing methods. The problem of gappy data in oceanography was addressed by Thompson (1971) who suggested that a random sampling of data points might be an optimally efficient approach. Further insight into the problem of missing data can found in Davis and Regier (1977) and Bretherton and McWilliams (1980). In this section, we present two examples of how to deal with gappy data. One is a straightforward analysis by Sturges (1983) who used monthly tide gauge data to investigate what happens to spectral estimates when one punches holes in the data set. The other is a practical guide to the interpolation of satellite-tracked Lagrangian drifter data with its inherently irregular time steps.

3.17.3.1 Interpolating gappy records for time-series analysis

Sturges (1983) used a Monte Carlo technique to poke holes at random in a known time series of monthly mean sea-level. The original record had a "red" spectrum which fell off as f^{-3} at high frequencies and contained a single major spectral peak at a period of 12 months. A total of 120 months of data were used in the analysis. The idea was to reconstruct the gappy series using a cubic spline interpolation method and see how closely the spectrum from the interpolated time series resembled that of the original time series. Data loss was limited to less than 30% of the record length and, for any individual experiment, the holes were all the same length. However, different hole lengths were used in successive runs. The only stipulation was that the length of the data segment before the next gap be at least as long as the gap itself. The program was not allowed to eliminate the first and last data points.

Cross-spectra were computed between the original time series and the interpolated gappy series. For a specified hole size, holes were generated randomly in the data series, the cross-spectra computed and the entire process repeated 1000 times. The magnitudes of the resulting cross-spectra provided estimates of how much power was lost or gained during the interpolation while the corresponding phases was interpreted as the error introduced by the interpolation process (Figure 3.20). Several important conclusions arise from Sturges' analysis:

(1) Gaps have a more adverse effect on weak spectral components (spectral peaks) than on strong ones embedded in the same background spectrum.
(2) The phase can be estimated to roughly 10° at the 90% confidence level for data losses of over 30% for a strong spectral signal; the requirement is that the gaps are

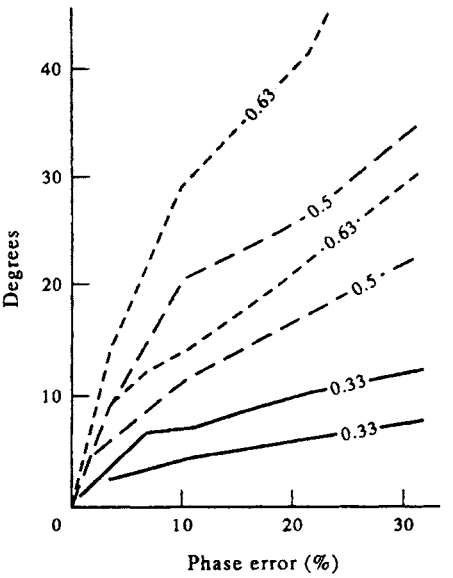

Figure 3.20. Absolute phase errors (°) expressed as a function of percent (%) data lost between the original sea level time series and the series with random holes filled in with a cubic spline fit. On each line, the ratio Δ/T *is shown, where* Δ *is the length of the gap and T is the spectral period of interest; the value 0.5 means that the holes were four units (months) long and the period 8 units long. Results are shown for the 90 and 99% confidence limits (lower and upper lines for each case). (From Sturges, 1983.)*

kept to about 1/3 of the period of the signal being examined. If the gaps are 1/2 of the period, the data loss can still be about 20%.

(3) Although correlation functions can be computed for gappy data, it is much more difficult to compute the cross-correlation function for these data.

According to Sturges' analysis, the adverse effects of gaps depends on the length of gaps relative to the length of data set and on the magnitudes of the dominant spectral components in the signal.

3.17.3.2 Interpolating satellite-tracked positional data

The analysis of positional (latitude, longitude) time series collected through the Service Argos satellite-tracking system illustrates some of the problems that may arise with standard interpolation procedures. Because the times that polar orbiting satellites pass over an oceanic region change through the day and because drifters move relative to the orbits of the satellite, the times between satellite fixes are irregular. At mid-latitudes, times between locational fixes can range from less than an hour to as long as 10 h. Typical average times between fixes are around 2–3 h (Thomson *et al.* 1997). The challenge is to generate regularly spaced time series of latitude (x) and longitude (y) from which one can derive regularly spaced time series of drifter zonal velocity ($u = \Delta x/\Delta t$) and meridional velocity ($v = \Delta y/\Delta t$). This challenge is especially problematic where a "duty cycle" has been programmed into the drifter transmitter to reduce the number (and cost) of transmissions to the passing NOAA satellites. A commonly used duty cycle, consisting of one day continuous transmission followed by two days of no transmission, results in large data gaps that

make calculation of mean currents difficult in regions having strong currents in the inertial and tidal frequency bands. The duty cycle of 8 h continuous transmission followed by 16 h of silence is superior for mid-latitude regions with strong inertial or tidal frequency variability.

Because of strong inertial motions in the upper layer of the open ocean and strong tidal motions over continental margins, sampling intervals of 3–4 h, or less, are preferable. A typical time step of 6 h used in many analyses of satellite-tracked drifters is inadequate to resolve inertial motions except in regions equatorward of 30° latitude where the inertial period $T = 1/f_{\text{inertial}}$ exceeds 24 h. (At 50° latitude, $T \approx$ 16.5 h; see *Coriolis frequency*.) To generate time series at a reasonably short time step, say 3 h, we need to interpolate between irregularly spaced data points. To do this, we use a cubic spline interpolation for each of the positional records. After the correct start and end times for the oceanic portion of the record have been determined, the first step in the process is to remove any erroneous points from the "raw" data by calculating speeds over adjacent time steps, t_i; e.g. $u_i = (x_{i+1} - x_i)/(t_{i+1} - t_i)$. One then omits any unrealistic velocity values that exceed some threshold value (say 5 m/s). This "edited" record needs to undergo further editing by averaging successive data positions for which the time step Δt is less than an hour. The reason for this is quite simple: Because positional accuracies Δx and Δy are about 350 m roughly 63% of the time, velocity errors are roughly $\Delta x/\Delta t > 0.1$ m/s when $\Delta t < 1$ h. Such error values are comparable to mean ocean currents and need to be eliminated from the records. Drifters located using GPS transmitters have smaller position errors and better velocity resolution. The time series also need to be examined for drogue-on, drogue-off. If a reliable strain sensor is built into the drogue system, it can be used to determine if and when the drogue fell off. Otherwise, one needs to calculate the speed-squared from the raw data and look for sudden major "jumps" in speed that signal loss of the drogue (Figure 3.21). We recommend this approach for all modern-day drifters since strain gauge sensors appear to be unreliable. At the time this book was being written, drogue loss and not battery or transmitter failure, was the primary cause of drifter "failure" in the open ocean.

Provided there are more than about six accurate satellite fixes per day, the edited positional records can be interpolated to regularly spaced 3-h time series using a cubic spline interpolation algorithm. In general, the spline curve will be well behaved and the fit will resemble the kind of curve one would draw through the data by eye. Inertial and tidal loops in the trajectory will be fairly well resolved. Spurious results will occur where data gaps are too large to properly condition the spline interpolation algorithm. Assuming that the spline interpolation of positions looks reasonable, the next step is to calculate the velocity components from the rate of change of position. It is tempting to equate the coefficient for the linear term in the cubic spline interpolation to the "instantaneous" velocity at any location along the drifter trajectory. That would be a mistake. Although trajectories can look quite smooth, curvatures can be large and resulting velocities unrealistic. In fact, use of the spline coefficients to calculate instantaneous velocity components leads to an increase in the kinetic energy of the motions. The reader can verify this by artifically generating a continuous time series of position consisting of a linear trend and time varying inertial motions. The artifical position record is then decimated to 3-hourly values and a cubic spline interpolation scheme applied. Using instantaneous velocity values at the 3-hourly time steps derived from the interpolation, one finds that the kinetic energy in most frequency bands is increased relative to the original record. The recommended procedure is to calculate

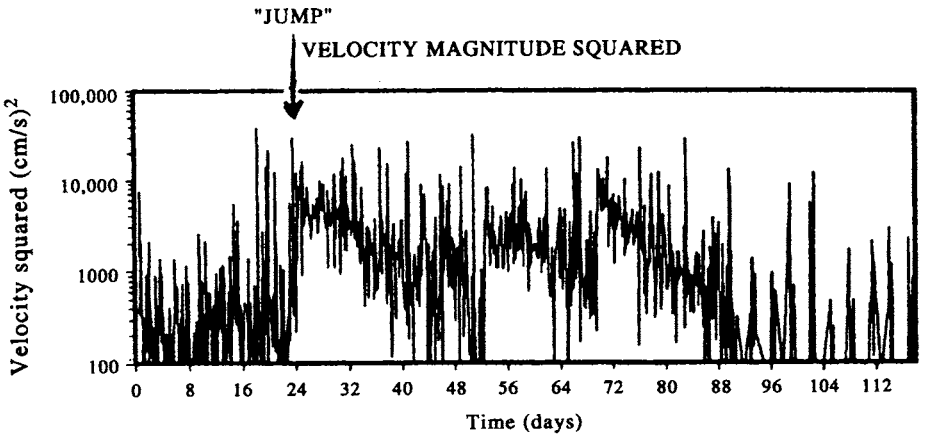

Figure 3.21. Sudden "jumps" in the speed-squared values from edited satellite-tracked drifter velocity data collected in the North Pacific near 50°N between 133–142°W longitude during the period 4 September to 30 December 1990. The "jump" indicates rapid acceleration of the drifter following probable loss of its drogue.

the two horizontal velocity components (u, v) from the central differences between three consecutive values of the 3-hourly positional data. From first differences, the velocity components at each point "i" are then: $u_i = (x_{i+1} - x_i)/(t_{i+1} - t_i)$ and $v_i = (y_{i+1} - y_i)/(t_{i+1} - t_i)$ for simple two-point differences or for the recommended centered values, $u_i = (x_{i+1} - x_{i-1})/(t_{i+1} - t_{i-1})$ and $v_i = (y_{i+1} - y_{i-1})/(t_{i+1} - t_{i-1})$. In summary, for those oceanic regions subject to pronounced inertial and tidal frequency motions, we have recommended the use of cubic spline interpolation to generate 2–4-hourly time series for position but simple linear interpolation of positional data to generate the corresponding time series for velocity. The interpolation requires more than 6–8 satellite fixes through the day to be successful.

Trajectories with data gaps that are long relative to the local inertial period require special consideration. For gaps associated with a transmitter duty cycle of 8 h "on" followed by 16 h "off", we can obtain accurate daily mean positional values by least-squares fitting a time-varying continuous function to successive segments of the irregular data and then averaging the resulting function over successive 24-h periods. This filtering processes is as follows (see Bograd *et al.*, 1999):

(1) Use least squares to fit a specified function, $\xi(t)$, to several (N) successive 8-h days of zonal (or meridional) trajectory data. The general model has the form $\xi(t) = a + bt + ct^2 + dt^3 + a_1 \sin(2\pi ft + \phi_1) + a_2 \sin(2\pi\omega_2 t + \phi_2)$ where a, b, c, d, a_1, ϕ_1, a_2 and ϕ_2 are the unknown coefficients, f is the local Coriolis frequency and ω_2 the semidiurnal frequency (0.081 cph). The phases ϕ_1, ϕ_2 for the two frequencies will vary from segment to segment. We suggest that four to five days (N = 4 or 5) of data be used for each segment fit. Shorter segments will have too few data for an accurate least-squares fit; longer segments will result in too much smoothing of the intermittant inertial and tidal motions;

(2) Repeat the least-squares operation for each segment of length N days, shifting forward in time by one day after each set of coefficients is determined. This yields one estimate for the first day $\xi_1 = \xi(t = t_1)$, two estimates for the second day, ξ_2, three estimates for the third day and four estimates for all other days until near the end of

the record when the number of estimates again falls to unity for the last record. Average all the values in each daily segment for each of the multiple curves $\xi_i(t)$ ($i=1$, ... , up to N) to get the average daily latitude $\xi_x(t)$ and longitude $\xi_y(t)$;

(3) The pairs of coefficients a_1, ϕ_1 and a_2, ϕ_2 can be used to give rough reconstructions of the inertial and semidiurnal tidal motions, respectively. Expect the phases to fluctuate considerably from segment to segment due to natural variability in the phases of the motions and from contamination by adjacent frequency bands.

For the duty cycle consisting of one day "on" followed by two days "off", the model is less useful (except at equatorial latitudes) and requires a much longer data segment (say 12 days instead of four) for each least-squares analysis.

3.17.3.3 Interpolation records from nearby stations

Provided that the spatial scales of the processes being examined are large compared to the separation between sampling sites, short gaps in the time series at one location can be filled using an identical type of time series from a nearby location. For example, missing hourly tide heights at one coastal tide gauge station can be filled using hourly tide heights from an adjacent station further along the coast. To do this, we first use coincident data segments to determine the relative amplitudes and phases of the time series at the two locations. A simple cross-correlation analysis can be used to determine the peak time lag between the series while the relative amplitudes can be obtained from the ratio of the standard deviations of the two series. Gaps in one time series (series 1) are then filled by applying the appropriate time lag and amplitude factor to the uninterupted data series (series 2). A more sophisticated approach would be to first obtain the complex transfer function $H_{12}(\omega) = |H_{12}(\omega)|\exp[i\phi_{12}(\omega)]$ as a function of frequency ω for the two coincident time series. The missing time series values at site 1 could then be filled using the amplitudes $|H_{12}(\omega)|$ and phase differences $\phi_{12}(\omega)$ of the transfer function applied to the uninterupted data series.

3.18 COVARIANCE AND THE COVARIANCE MATRIX

Covariance, like variance, is a measure of variability. For two variables, the covariance is a measure of the joint variation about a common mean. When extended to a multivariate population, the relevant statistic is the covariance matrix. As we shall see, it is equivalent to what will be introduced later as the "mean product matrix." The covariance and covariance matrix are the fundamental concepts behind the spatial analysis techniques discussed in the next chapter.

3.18.1 Covariance and structure functions

The covariance $C(Y_1, Y_2)$, also written as $\text{cov}[Y_1, Y_2]$, between variables Y_1, Y_2 is

$$C(Y_1, Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)] \tag{3.18.1}$$

where $\mu_1 = E[Y_1]$ and $\mu_2 = E[Y_2]$. A positive covariance indicates that Y_2 and Y_1 increase and decrease together while a negative covariance has Y_2 decreasing as Y_1

increases, and vice versa. We can expand equation (3.18.1) into a more convenient computational form

$$C(Y_1, Y_2) = E[Y_1 Y_2] - E[Y_1]E[Y_2] \tag{3.18.2}$$

Note, that if Y_1, Y_2 are independent random variables, then $C[Y_1, Y_2] = 0$.

For a two-dimensional isotropic velocity field, $u_i(y)$, the covariance tensor $C(r)$, also called the *structure function* from earlier studies of turbulence, takes the form

$$\begin{aligned} C_{ij}(r) &= \langle u_i(\mathbf{y}) u_j(\mathbf{y}+\mathbf{r}) \rangle \\ &= \sigma^2 \frac{[f(r) - g(r)] r_i r_j}{r^2 + g(r)\delta_{ij}} \end{aligned} \tag{3.18.3}$$

where $\langle \cdot \rangle$ denotes an ensemble average, $r \equiv |\mathbf{r}|$, $\mathbf{y} = (y_1, y_2)$ is the position vector, $f(r)$ and $g(r)$ are, respectively, the one-dimensional longitudinal and transverse correlation functions, and $\sigma^2 = \langle u_i(\mathbf{y})^2 \rangle$. The longitudinal and transverse correlation functions are

$$f(r) = \langle u_L(\mathbf{y}) u_L(\mathbf{y}+\mathbf{r}) \rangle \tag{3.18.4a}$$

$$g(r) = \langle u_P(\mathbf{y}) u_P(\mathbf{y}+\mathbf{r}) \rangle \tag{3.18.4b}$$

where $u_L(\mathbf{y})$ and $u_P(\mathbf{y})$ are the velocity fluctuations parallel and perpendicular to $\mathbf{r} = (r_1, r_2)$. The velocity fluctuations are normalized so that the correlations equal unity at $r = 0$. If the two-dimensional flow field is horizontally nondivergent, homogenous and isotropic, then $C_{ij}(r) = 0$ and

$$g(r) = \frac{\mathrm{d}}{\mathrm{d}r}[rf(r)] \tag{3.18.5}$$

Freeland *et al.* (1975) have used (3.18.5) to test for two-dimensional, nondivergent, homogenous, and isotropic low-frequency velocity structure in SOFAR float data collected in the North Atlantic. Stacey *et al.* (1988) used this relation to test for similar flow structure in the Strait of Georgia. Although close to the error limits in certain cases, the observed structure is generally consistent with horizontal, nondivergent, homogeneous and isotropic flow (Figure 3.22). The dotted lines in Figure 3.22 are the analytical functions

$$f(r) = (1 + br)\,\mathrm{e}^{-br} \tag{3.18.6a}$$

$$g(r) = (1 + br - b^2 r^2))\mathrm{e}^{-br} \tag{3.18.6b}$$

3.18.2 A computational example

If Y_1, Y_2 have a joint probability density function

$$f(y_1, y_2) = \begin{cases} 2y_1, & 0 \le y_1 \le 1; \quad 0 \le y_2 \le 1 \\ 0, & \text{elsewhere} \end{cases} \tag{3.18.7}$$

what is the covariance of Y_1, Y_2? We first write the expected value of Y_1, Y_2 as

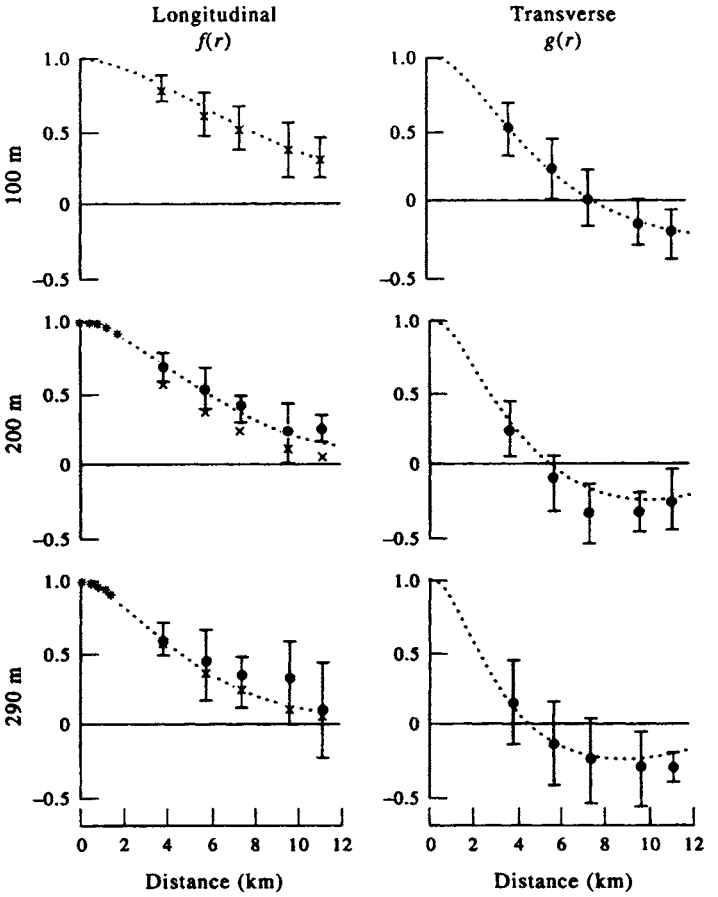

Figure 3.22. Longitudinal and transverse correlations at 100, 200, and 280/290 m depths. The dots ar measured average values and error bars are the standard deviations. The mean and trend were remove from each time series before calculation of the correlations. The crosses are predicted values of f(r calculated using (3.18.5) by drawing straight line segments between the average values of g(r) an integrating under the curve. (From Stacey et al., 1988.)

$$E[Y_1Y_2] = \int_0^1 \int_0^1 y_1y_2 f(y_1,y_2)dy_1dy_2 = \int_0^1 \int_0^1 y_1y_2(2y_1)dy_1dy_2$$

$$= \int_0^1 \frac{1}{3}y_2(2y_1^2)\,|_0^1 dy_2 = \int_0^1 \frac{2}{3}y_2dy_2 = \frac{2y_2^2}{3\,2}\Big|_0^1 = \frac{1}{3}$$

Recall that, for discrete variables

$$E[g(Y_1, ..., Y_k)] = \sum_{y_k} \cdots \sum_{y_1} g(y_1, ..., y_k)P_1(y_1, ..., y_k)$$

or for continuous variables

$$E[g(Y_1, ..., Y_k)] = \int_{y_k} ... \int_{y_1} g(y_1, ..., y_k) f(y_1, ..., y_k) dy_1 ... dy_k$$

For this example, we find $E[Y_1Y_2] = 1/3$. Now

$$E[Y_1] = \int_0^1 \int_0^1 y_1(2y_1) dy_1 dy_2 = \int_0^1 \frac{2}{3} y_1^3 \Big|_0^1 dy_2 = \frac{2}{3} y_2 \Big|_0^1 = \frac{2}{3}$$

and $E[Y_2] = 1/2$, so that $\text{cov}[Y_1Y_2] = E[Y_1Y_2] - \mu_1\mu_2 = 1/3 - (2/3)(1/2) = 0$. Therefore, Y_1 and Y_2 are independent. Of course, we could have anticipated this result since $f(y_1, y_2)$ in equation (3.18.7) is independent of y_2.

3.18.3 Multivariate distributions

In the case of multivariate distributions, the covariance becomes the *covariance matrix*. If we have n measurements (samples) of N variables (Y), we can describe this as n random variables having a joint N-dimensional probability density function (PDF)

$$f_{1,2,...,N}(Y_1, Y_2, ..., Y_N) \tag{3.18.8}$$

If the random variables, Y, are mutually independent, the joint PDF can be factored in the usual way as

$$f_{1,2,...,N}(Y_1, Y_2, ..., Y_N) = f_1(Y_1) f_2(Y_2) ... f_N(Y_N) \tag{3.18.9}$$

An important multivariate PDF is the multivariate normal PDF

$$f_Y(Y) = \frac{1}{2\pi^{N/2} |\mathbf{W}|^{1/2}} \exp\left[-\frac{1}{2} (Y - \mu)^T \mathbf{W}^{-1} (\mathbf{Y} - \mu) \right]$$

where $Y^T = (Y_1, Y_2, ..., Y_N)$, $\mu^T = (\mu_1, \mu_2, \ddot{O}, \mu_N)$, are the row vectors and $\mathbf{W}^{-1}$ is the inverse of the covariance matrix $\mathbf{W}$

$$\mathbf{W} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} & \sigma_1\sigma_3\rho_{13} & ... & \sigma_1\sigma_N\rho_{1N} \\ \sigma_2\sigma_1\rho_{12} & \sigma_2^2 & \sigma_2\sigma_3\rho_{23} & ... & \sigma_2\sigma_N\rho_{2N} \\ ... & ... & ... & ... & ... \\ \sigma_N\sigma_1\rho_{N1} & \sigma_2\sigma_N\rho_{2N} & ... & ... & \sigma_N^2 \end{pmatrix} \tag{3.18.10}$$

or

$$\mathbf{W} = \begin{pmatrix} V[Y_1] & C[Y_1Y_2] & ... & ... & C[Y_1Y_N] \\ C[Y_2Y_1] & V[Y_2] & ... & ... & C[Y_2Y_N] \\ ... & ... & ... & ... & ... \\ C[Y_NY_1] & C[Y_NY_2] & ... & ... & V[Y_N] \end{pmatrix} \tag{3.18.11}$$

Note that $C[Y_iY_j] = C[Y_jY_i]$ and therefore $\mathbf{W}$ is symmetric ($\mathbf{W} = \mathbf{W}^T$). In addition, $\mathbf{W}$ is positive semi-definite; that is, $|\mathbf{W}|$ and all its principal minors are nonnegative. Another way to show this is

$$V[\lambda^T Y] = E[\lambda^T (Y - \mu)(Y - \mu)^T \lambda] = \lambda^T |\mathbf{W}| \tag{3.18.12}$$

which will always be nonnegative for any λ.

3.19 THE BOOTSTRAP AND JACKKNIFE METHODS

Many data series in the natural sciences are nonreproducible and the researcher is left with only one set of observations with which to work. With only one realization of a series, it is impossible to compare it with a related series to determine if they are drawn from the same, or from different, populations. There are numerous oceanographic examples, including tsunami oscillations recorded by a coastal tide gauge, a single seasonal cycle of monthly mean currents at a mooring location, and a trend in long-term temperature data from a climate monitoring station. Marine biologists face similar limitations when analyzing groups of animal species caught in nets or bottom grab samples. The problem is that empirical observations are prone to error and any interpretation of an event must be devised based on statistical measures of the probability of the event. A fundamental measure for testing the validity of any property of a data set is its variance. Parametric statistical models have been developed which help the investigator decide the degree of faith to be placed in a given statistic. However, data and model are often nonlinear so that it is not usually possible to find an analytical expression for model variance in terms of the data variance.

The *parametric* statistical methods presented in the previous sections were institutionalized long before the time of modern digital computers when use of analytical expressions greatly simplified the laborious hand calculation of statistical properties. During the past few decades, *nonparametric* statistical methods have been developed to take advantage of the increasing computational efficiencies of computers. An advantage of the new methods is that they permit investigations of the statistical properties of a sample which do not conform to a specific analytical model. Equally importantly, they can be applied to small data sets while still providing a reliable estimation of confidence limits on the statistic of interest. "Bootstrapping" and "jackknifing" are two of the more commonly used methods that could not be used effectively until the invention of the digital computer. Both are resampling techniques in which artificial data sets are generated by selection of points from an original set of data. Specifically, we start with a single realization of an "experiment" and from that one set of experimental data we create a multitude of new artificial realizations of the experiment without having to repeat the observations. These realizations are then used to estimate the reliability of the particular statistic of interest, with the underlying assumption that the sample data are representative of the entire population.

In the bootstrapping method, random samples selected during the resampling process are replaced before each new sample is created. As a consequence, any data value can be drawn many times, or not at all. The name bootstrap arises from the expression "to lift oneself up by one's bootstraps". In jackknifing, artificial data sets are created by selectively and systematically removing samples from the original data set. The statistics of interest are recalculated for each resulting truncated data set and

the variability among the artificial samples used to describe the variability of these statistics. "Cross-validation" is an older technique. The idea is to split the data into two parts and set one part aside. Curves are fitted to the first part and then tested against values in the second part. Cross-validation consists of seeing how well the fitted curves predict the values in the portion of data set aside. The data can be randomly split in many ways and many times in order to obtain the needed statistical reliability. For additional information on this technique, the reader is referred to Efron and Gong (1983).

3.19.1 Bootstrap method

Introduced by Efron in 1977 (Diaconis and Efron, 1983), bootstrapping provides freedom from two limiting factors that have constrained statistical theory since its beginning: (1) the assumption of normal (Gaussian) data distributions; and (2) the focus on statistical measures whose theoretical properties can be analyzed mathematically. As with other nonparametric methods, bootstrapping is insensitive to assumptions made with respect to the statistical properties of the data and does not need an analytical expression for the connection between model and data statistical properties. Resampling techniques are based on the idea that we can repeat a particular experiment by constructing multiple data sets from the one measured data set. Application of the resampling procedure must be modeled on a testable hypothesis so that the resulting probability can be used to accept or reject the null hypothesis. The methods can be applied just as well to any statistic, simple or complicated. A *bootstrap sample* is a "copy" of the original data that may contain a certain value (datum, x_n) more than once, once, or not at all (i.e. the number of occurrences of x_n lies between 0 and N, where N is the number of independent data points). Introductions into the bootstrapping procedure can be found in Efron and Gong (1983), Diaconis and Efron (1983), and Tichelaar and Ruff (1989). Nemec and Brinkhurst (1988) apply the method to testing the statistical significance of biological species cluster analysis for which there are duplicate or triplicate samples for each location.

Suppose that we have N values of a scalar or vector variable, x_n $(1 \leq n \leq N)$, whose statistical properties we wish to investigate in relation to another variable. This could be a univariate variable such as sea-level height $x_n = \eta(t_n)$ at a single location over a period of N time steps, t_n, or the structure of the first mode empirical orthogonal function $\phi_1(x_n)$ as a function of location, x_n. Alternatively, we could be dealing with a bivariate variable (x_{1n}, x_{2n}) such as water temperature versus dissolved oxygen content from a series of vertical profiles. Results apply to any other set of measurements whose statistics we wish to determine. We may want to compare means and standard deviations (variances) of different records to see if they are significantly different. Alternatively, we might want to place confidence limits on the slope of a line derived using a standard least-squares fit to our bivariate data (x_{1n}, x_{2n}), or, determine how much confidence we can have in the coefficients we obtained from the least-squares fit of an annual cycle to a single set of 12 monthly mean current records from a mooring location. Note that if there is a high degree of correlation among the N data values, the N are not statistically independent samples and we are faced with the usual problem of dealing with an effective number of degrees of freedom N^* for the data set.

The procedure is to equate each of our N independent data points with a number produced by a random number generator. We can do this by assigning each of the data values to separate uniform-width bins lying along the line $(-1, +1)$, or (0, 1), depending on the random number generator being used. For N values, there will be N

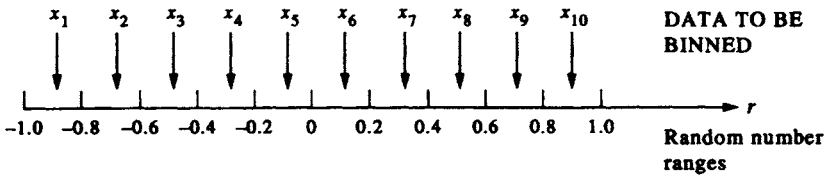

Figure 3.19.1. The assignment (binning) of observed data values x_n ($n = 1, \ldots, 10$) to 10 range values of the random number, r_k ($k = 1, \ldots, 10$). For each bootstrap sample of 10 values, 10 random numbers are selected and located according bin range. The datum values x_n assigned to each range are then used to form the bootstrap sample.

uniform-width bins on the line and each bin will be equated with one of the N data values (Figure 3.19.1). The bin width is $2/N$ if the line -1 to $+1$ is used. A random number generator such as a Monte Carlo scheme is used to randomly select sequences of N bins corresponding to the multiple bootstrap samples. Suppose that the random number generator picks a number, r, from the range $-1 \leq r \leq 1$. If this number falls into the range of bin k, corresponding to the range $[2(k-1)/N] - 1 \leq r_k \leq (2k/N) - 1$, for $k = 1, \ldots, N$, then the data value x_k assigned to bin k is taken to be one of the samples we need to make up our bootstrap data set. In Figure 3.19.1, there are 10 data values and 10 corresponding random number segments of length 0.2, with datum value x_1 assigned to the range -1.0 to -0.8, x_2 assigned to -0.8 to -0.6, and so on. Since bootstrapping works with replacement, it is quite possible to get the same bin several times, or not at all. The first N data values from our resampling constitute the first bootstrap sample. The process is then repeated again and again until hundreds or thousands of bootstrap samples have been generated. Diaconis and Efron (1983) discuss making a billion bootstrap samples. They also take another approach. Instead of generating one bootstrap sample at a time by equating bins along the real line $(-1, 1)$ with N samples, they generate all the needed multiple copies of all the N data values (say one million copies of each of the original data values or data points) and place them all in a rotating "lotto" bin. They then reach in and pull out all the requisite number of N-value bootstrap samples from the shuffled points, being careful to throw each data point back into the bin before selecting the next value. This requires some sort of label for each value in the bin based on a random selection process that can identify a data point that has been selected.

Although bootstrapping has yet to find widespread application in the marine sciences, there are several noteworthy examples in the literature. Enfield and Cid (1990) examined the stationarity of different groupings of El Niño recurrence rates based on the chronology of Quinn *et al.* (1987). For example, group 1 consisted of all strong (S) and very strong (VS) events for the period 1525–1983, while groups 4 and 5 consisted of S/VS events for times of high and low solar activity for this period. Groups 6–10 contained different samples of intensities for the modern period of 1803–1987. Maximum likelihood estimation was used to fit a two-parameter Weibull distribution $f(t)$ to each sample group,

$$f(t) = (\beta t^{\beta-1}/\tau^{\beta}) \exp[-(t/\tau)^{\beta}] \tag{3.19.1}$$

where β and τ are, respectively, the shape (peakedness) and time scale (RMS return interval) parameters, and t is the random variable for the return interval. For each group, only a single distribution could be fitted. To derive estimates of the mean and

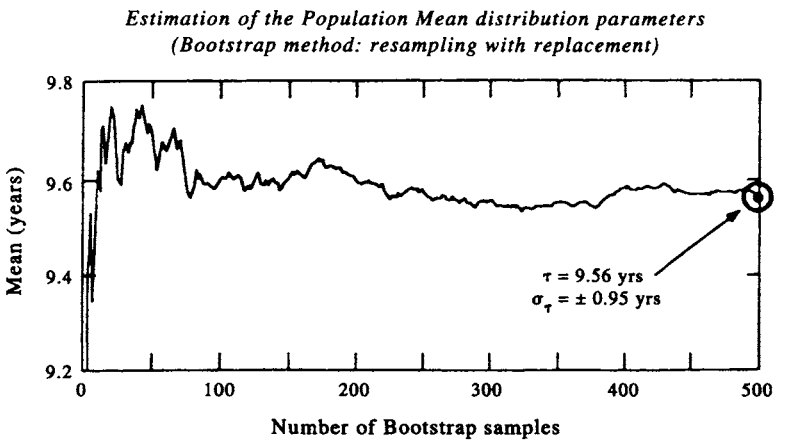

Figure 3.19.2. Estimation of the population mean distribution parameters (mean return time in years) using the bootstrap method for El Niño events taking place during times of low solar activity for the period 1525–1983. τ is the return time and σ_τ its standard deviation. (Enfield and Cid, 1990.)

standard deviations of the parameters for each group, 500 bootstrap samples were generated and the Weibull parameters obtained for each sample. As indicated by Figure 3.19.2, this number of samples provides good convergence to the mean value for the Weibull distribution fit for each group. The distribution of El Niño return events for bootstrap samples for all intensities for the "early modern" period 1803–1891 is shown in Figure 3.19.3 along with its corresponding Weibull distribution. Enfield and Luis use the resampling analysis to show that, for the groups associated with times of low solar activity and those associated with times of high solar activity, there is comparatively little overlap between the bootstrap-derived frequency histograms and mean return time scales, τ (years) (Figure 3.19.4). These results suggest that there is a statistical difference in the return times for the two groups and that return times are nonstationary.

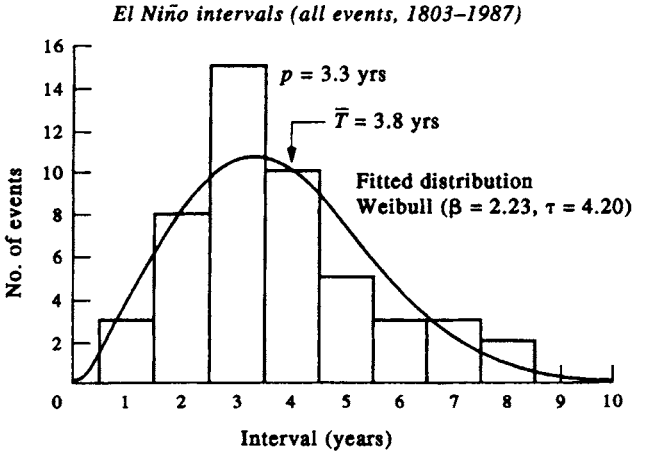

Figure 3.19.3. Histogram of El Niño return times for all events between 1803 and 1987 (group #7) derived using the bootstrapping resampling technique. The solid curve is the Weibull distribution fitted to the histogram. The modal and mean return intervals (3.3 and 3.8 years, respectively) are the derived from the MLE-estimated population parameters. (From Enfield and Cid, 1990.)

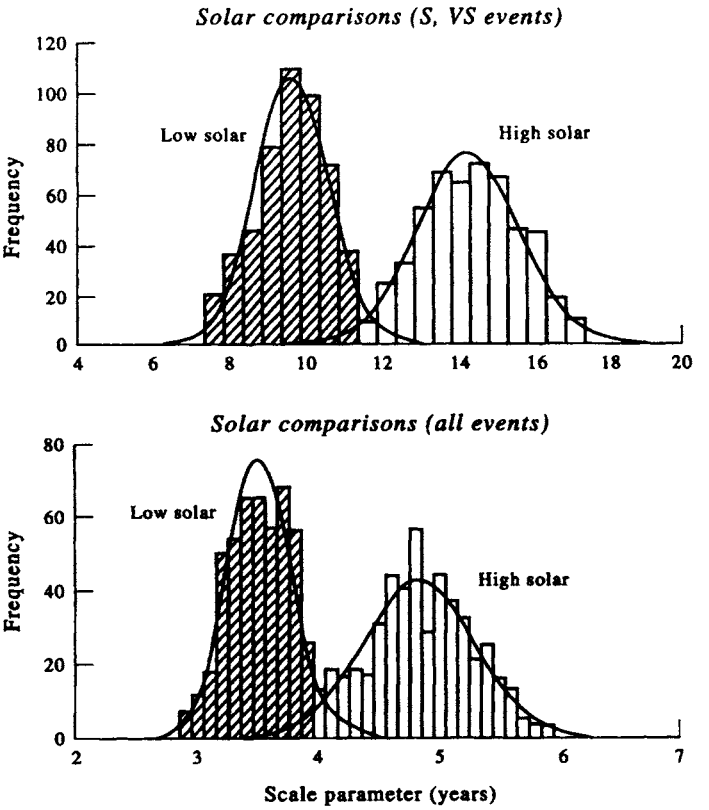

Figure 3.19.4. Histograms and fitted Weibull distributions obtained using the bootstrapping method. Plots show the occurrences of El Niño events for the times of low and high solar flare activity for (a) Strong and very strong El Niño events, only; (b) All El Niño events.

Much of the present evidence for possible global warming is based on Northern Hemispheric annual surface air temperature records over the past 100 years (Jones *et al.*, 1986; Hansen and Lebedeff, 1987; Gruza *et al.*, 1988). Interest in the reliability of the means and trends of these records (labeled H, J, and G) prompted Elsner and Tsonis (1991) to examine differences in means and trends of pairs of these records for the three global mean temperature curves. The data sets have been constructed using different averaging methods and different observational data bases. Data set H contains only observations from land stations whereas data set J uses both land and ship-based observations. Averages for set H are derived using equal-area boxes over the globe whereas data set G is constructed by visual inspection of anomalies from sea-level temperature analyses. The usual assumption is that these time series are representative of the same population, a result that appears to be supported by the statistically significant correlation $r > 0.79$ among the different curves. As pointed out by Elsner and Tsonis, however, the presence of trends in these data means that the linear cross-correlation coefficient may not be a reliable measure of the covariability of the records. Two questions can be addressed using the bootstrapping method: (1) are the three versions of the temperature records significantly different that we can say they are not drawn from the same population? (The null hypothesis is false.); and (2) are the trends in the three records sufficiently alike that they are measuring a true rise in global temperature?

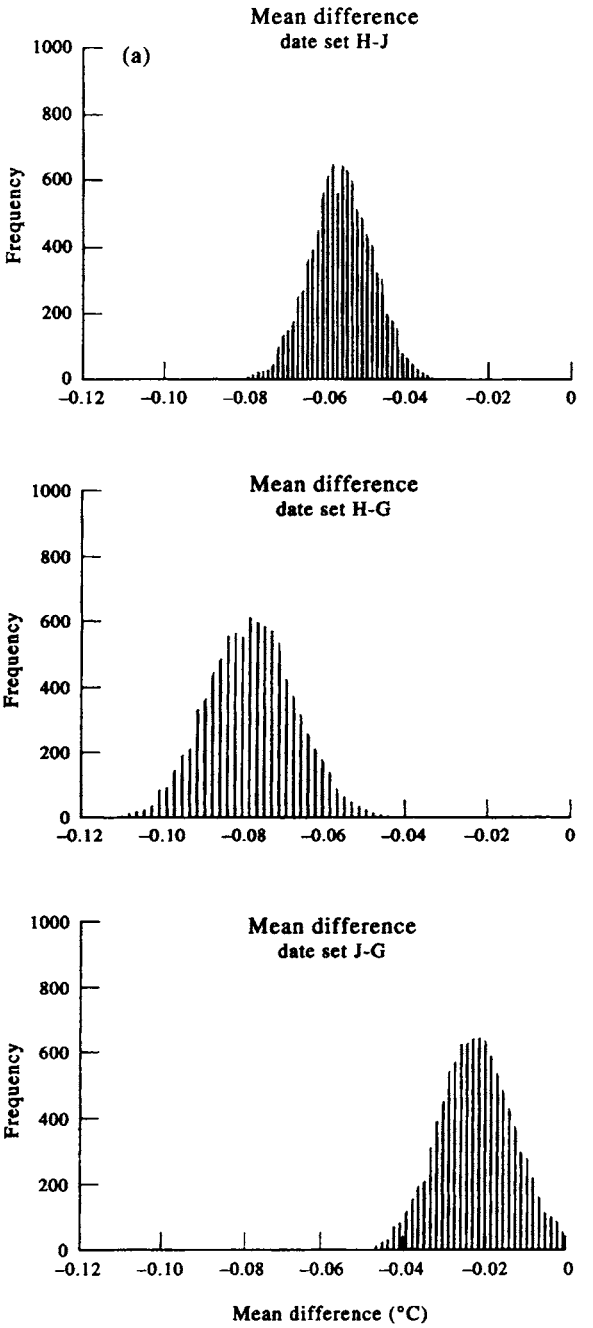

Figure 3.19.5. Bootstrap-generated histograms of global air temperature difference records obtained by subtracting the temperature records of Jones et al. (1986) (J), Hansen and Lebedeff (1987) (H), and Gruza et al. (1988) (G). (a) Frequency distributions of the mean differences plotted for 10^4 bootstrap samples. The x-axis (ordinate) gives the number of times the bootstrap mean fell into a given interval. All three distributions are located to the left of a zero mean difference. (From Elsner and Tsonis, 1991.)

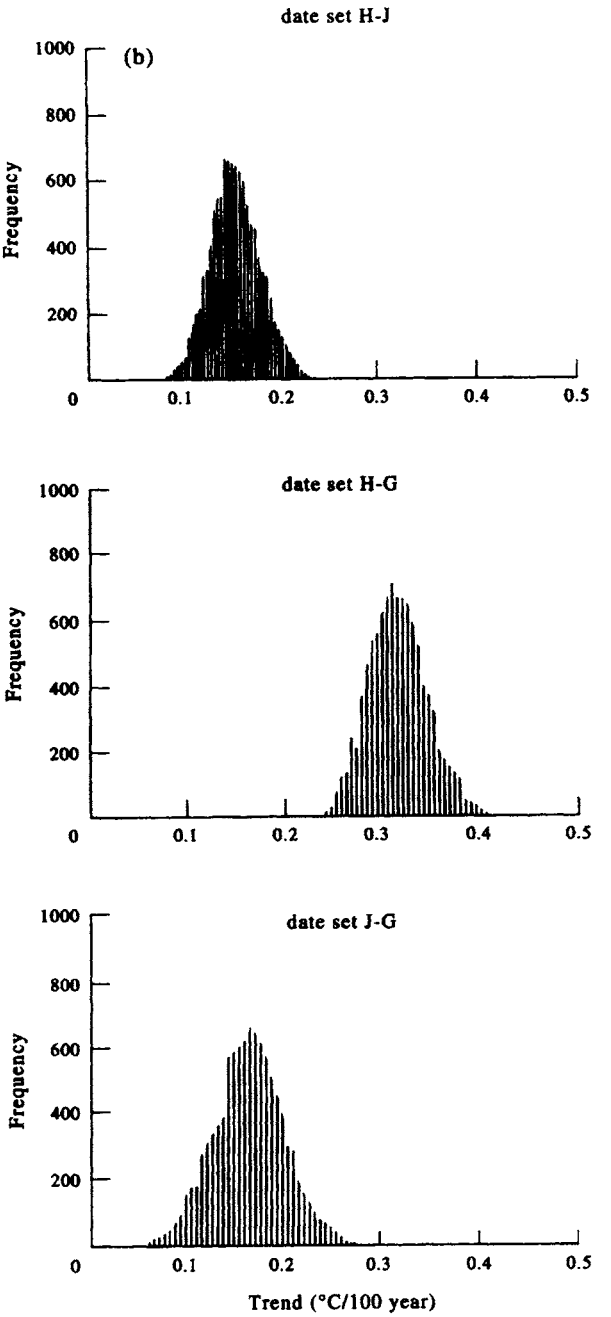

Figure 3.19.5. Bootstrap-generated histograms of global air temperature difference records obtained by subtracting the temperature records of Jones et al. (1986) (J), Hansen and Lebedeff (1987) (H), and Gruza et al. (1988) (G). (b) Same as (a), but for slope (trend) of the temperature difference curves. All three distributions are separated from zero indicating significant differences between long-term surface temperature trends given by each of the three data sets. (From Elsner and Tsonis, 1991.)

Because of the strong linear correlation in the records, the authors work with difference records. A difference record is constructed by subtracting the annual (mean removed) departure record of one data set from the annual departure record of another data set. Although not zero, the cross-correlation for the difference records is considerably less than those for the original departure records, showing that differencing is a form of high-pass filtering that effectively reduces biasing from the trends. The average difference for all 97 years of data used in the analyses (the difference record H–J relative to the years 1951–1980) is $-0.05°C$, indicating that the hemispheric temperatures of Jones *et al.* (1986) are slightly warmer than those of Hansen and Lebedeff (1987). Similar results were obtained for H–G and J–G. To see if these differences are statistically significant, 10,000 bootstrap samples of the difference records were generated. The results (Figure 13.9.5a) suggest that all three hemispheric temperature records exhibit significantly different nonzero means. The overlap in the distributions is quite minimal. The same process was then used to examine the trends in the difference records. For the H–J record, the trend is $+0.15°C$/century so that the trend of Hansen and Lebedeff is greater than that of Jones *et al.* As indicated in Figure 13.9.5(b), the long-term trends were distinct. On the basis of these results, the authors were forced to conclude that at least two of the data sets do not represent the true population (i.e. the truth). More generally, the results bring into question the confidence one can have that the long-term temperature trends obtained from these data are representative of trends over hemispheric or global scales.

Biological oceanographers often have difficult sampling problems that can be addressed by bootstrap methods. For example, the biologist may want to use cluster analyses of animal abundance for different locations to see if species distributions differ statistically from one sampling location (or time) to the next. Cluster analyses of ecological data use dendrograms—linkage rules which group samples according to the relative similarity of total species composition—to determine if the organisms in one group of samples have been drawn from the same or different statistical assemblages of those of another group of samples. Provided there are, at least, replicates for most samples, bootstrapping can be used to derive tests for statistical significance of similarity linkages in cluster analyses (Burd and Thomson, 1994). For further information on this aspect of bootstrapping, the reader is referred to Nemec and Brinkhurst (1988). Finally, in this section, we note that it is possible to vary the bootstrap size by selecting samples smaller than N, the original size of the data set, to compare various estimator distributions obtained from different sample sizes. This allows one to observe the effects of varying sample size on sample estimator distributions and statistical power.

3.19.2 Jackknife method

Several other methods are similar in concept to bootstrapping but differ significantly in detail. The idea, in each case, is to generate artificial data sets and assess the variability of a statistic from its variability over all the sets of artificial data. The methods differ in the way they generate the artificial data. Jackknifing differs from bootstrapping in that data points are not replaced prior to each resampling. This technique was first proposed by Maurice Quenouille in 1949 and developed by John Tukey in the 1950s. The name "jackknife" was used by Tukey to suggest an all-purpose statistical tool.

A jackknife resample is obtained by deleting a fixed number of data points (j) from the original set of N data points. For each resample, a different group of j values is removed so that each resample consists of a distinct collection of data values. In the "delete-j" jackknife sample, there will be $k = N - j$ samples in each new truncated data set. The total number of new artificial data that can be generated is

$$\binom{N}{j}$$

which the reader will recognize as ${}_NP_j = N!/(N-j)!$, the number of permutations of N objects taken j at a time. Consider the simple delete-1 jackknife. In this case, there are $N - 1$ samples per artificial data set and a total of ${}_NP_j = N$ new data sets that can be created by systematically removing one value at a time. As illustrated by Figure 3.19.6, an original data set of four data values will yield a total of four distinct delete-1 jackknife samples, each of size three (3), which can then be used to examine various statistics of the original data set. The sample average of the data derived by deleting the ith datum, denoted by the subscript (i), is

$$\bar{x}_{(i)} = \frac{N\bar{x} - x_i}{N-1} = \frac{1}{N-1}\sum_{j\neq i}^{N} x_j \tag{3.19.2}$$

where

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

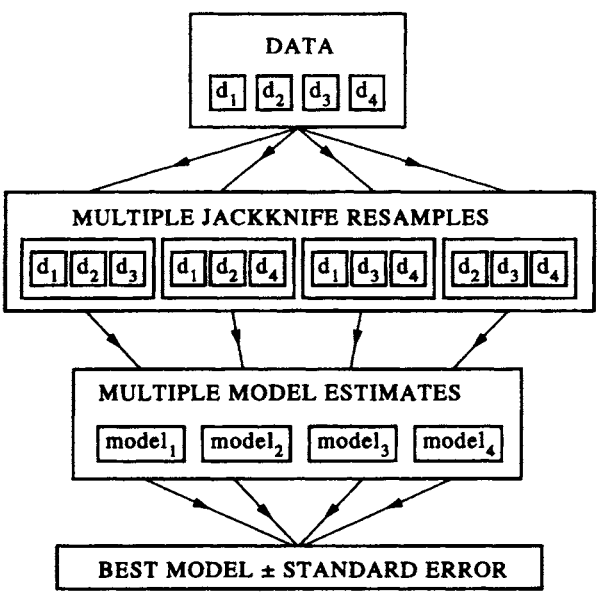

Figure 3.19.6. Schematic representation of the jackknife. The original data vector has four components (samples), labeled d_1 to d_4. The data are resampled by deleting a fixed number of components (here, one) from the original data to form multiple jackknife resamples (in case, four). Each resample defines a model estimate. The multiple model estimates are then combined to a best model and its standard deviation. (From Tichelaar and Ruff, 1989.)

is the mean found using all the original data. The average of the N jackknife averages, $\bar{x}_{(i)}$, is

$$\bar{x}^* = \frac{1}{N}\sum_{i=1}^{N} \bar{x}_{(i)} = \bar{x} \tag{3.19.3}$$

The last result, namely that the mean of all the jackknife samples is identical to the mean of the original data set, is easily obtained using equation (3.19.2). The estimator for the standard deviation, σ_j, of the delete-1 jackknife is

$$\sigma_j = \sum_{i=1}^{N} \left[\left(\bar{x}_{(i)} - \bar{x}^* \right)^2 \right]^{1/2} \tag{3.19.4a}$$

$$= \frac{1}{N-1} \sum_{i=1}^{N} \left[(x_i - \bar{x})^2 \right]^{1/2} \tag{3.19.4b}$$

where (3.19.4b) is the usual expression for the standard deviation of N data values. Our expression differs slightly from that of Efron and Gong (1983) who use a denominator of $1/[(N-1)N]$ instead of $1/(N-1)^2$ in their definition of variance. The advantage of (3.19.4a) is that it can be generalized for finding the standard deviation of any estimator θ that can be derived for the original data. In particular, if θ is a scalar, we simply replace $x_{(i)}$ with $\theta_{(i)}$ and x^* with θ^* where $\theta_{(i)}$ is an estimator for θ obtained for the data set with the ith value removed. Although the jackknife requires fewer calculations than the bootstrap, it is less flexible and at times less dependable (Efron and Gong, 1983). In general, there are N jackknife samples for the delete-1 jackknife as compared with

$${}_{2N-1}P_N = \binom{2N-1}{N}$$

bootstrap points.

Our example of jackknifing is from Tichelaar and Ruff (1989) who generated $N = 20$ unequally spaced data values y_i that follow the relation $y_i = cx_i + \varepsilon_i$ ($c = 1.5$, exactly), where ε_i is a noise component drawn from a "white" spectral distribution with a normalized standard deviation of 1.5 and mean of zero. The least squares estimator for the standard deviation of the slope is

$$\hat{\sigma} = \sum_{i=1}^{N} \left[(y_i - \hat{c}x_i)^2 \right] \Big/ \left[(N-1) \sum_{i=1}^{N} x_i^2 \right] \tag{3.19.5}$$

where $\hat{c} \sum y_i x_i / \sum x_i^2$ Two jackknife estimators were used: (1) The delete-1 jackknife, for which the artificial sample sizes are $N - 1 = 19$; and (2) The delete-half ($N/2$) jackknife for which the sample sizes are $N - N/2 = 10$. In both cases, the jackknife resamples had equal weighting in the analysis. For the delete-half jackknife, a Monte Carlo determination of 100 subsamples was used since the total samples ${}_{20}P_{10} = 20!/10!$ is very large. The results are presented in Figure 3.19.7. The last panel gives the corresponding result for the bootstrap estimate of the slope using 100 bootstrap samples. Results showed that the bootstrap standard error of the slope was slightly lower than those for both jackknifing estimates.

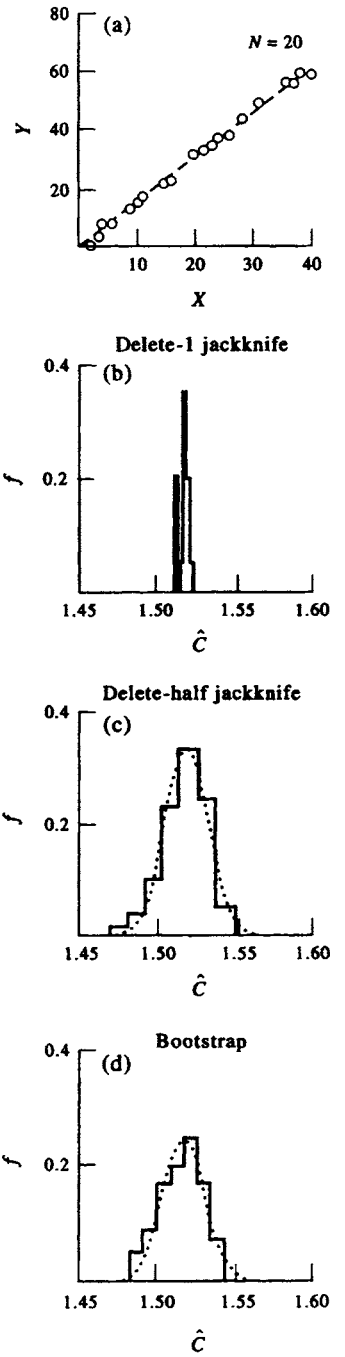

Figure 3.19.7. Use of the bootstrapping technique to estimate the reliability of a linear regression line. (a) A least-squares fit through the noisy data, for which the estimated slope $\hat{c} = 1.518 \pm 0.0138$ *(±1 standard error); (b) The normalized frequency of occurrence distribution, f, for the delete-1 jackknife which yields* $\hat{c} = 1.518 \pm 0.0136$*; (c) As in (b) but for the delete-half jackknife for which* $\hat{c} = 1.517 \pm 0.0141$*; (d) The corresponding bootstrapping estimate, for which* $\hat{c} = 1.517 \pm 0.0132$*. Note the scale difference between (b) and (c). The dashed line is the analytical distribution of* $\hat{c}$*. (From Tichelaar and Ruff, 1989.)*