Cambridge University Press
978-1-108-47241-8 — Statistical Methods for Climate Scientists
Timothy DelSole , Michael Tippett
Excerpt
More Information

1

Basic Concepts in Probability and Statistics

> Probability theory is nothing more than common sense reduced to calculation.
>
> *Pierre Simon Laplace*

This chapter reviews some essential concepts of probability and statistics, including the following:

- line plots, histograms, scatter plots
- mean, median, quantiles, variance
- random variables
- probability density function
- expectation of a random variable
- covariance and correlation
- independence
- the normal distribution (also known as the Gaussian distribution)
- the chi-squared distribution.

These concepts provide the foundation for the statistical methods discussed in the rest of this book.

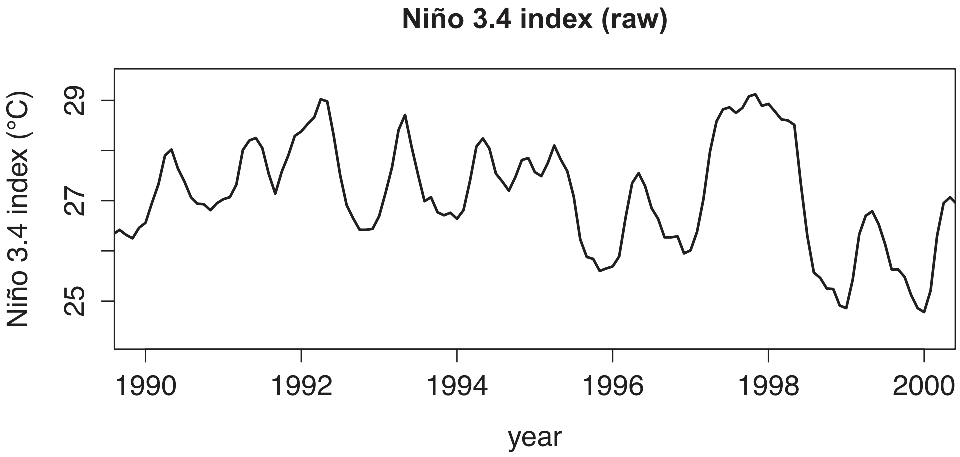


Figure 1.1 A time series of the monthly Niño 3.4 index over the period 1990–2000.

1.1 Graphical Description of Data

Scientific knowledge is based on observations. However, a mere list of observational facts rarely advances science. Instead, the data need to be organized in ways that help the scientist interpret the data in a scientific interpret the data in a scientific framework and formulate new hypotheses that can be checked in independent data or experiments. To illustrate ways of describing the main characteristics of a data set, consider a specific observable quantity: the area-average sea surface temperature in the equatorial Pacific in the region 170°W − 120°W and 5°S − 5°N. This quantity is called the Niño 3.4 index and is an indicator of seasonal climate variations. The monthly average value of this index over a period of 50 or more years is readily available from various data portals. What are some ways of describing such a data set?

Data taken sequentially in time are known as *time series*. A natural way to visualize time series is to plot them as a function of time. A *time series plot* of Niño 3.4 is shown in Figure 1.1. The figure reveals that peaks and valleys occur at nearly periodic intervals, reflecting the annual cycle for this region. The figure also reveals that the time series is “smooth” – the value at one time is close to the value at neighboring times. Such time series are said to be *serially correlated* or *autocorrelated* and will be studied in Chapter 5. Another feature is that the minimum values generally decreased from 1993 to 2000, suggesting a possible long-term change. Methods for quantifying long-term changes in time series will be discussed in Chapters 8 and 9. Note how much has been learned simply by plotting the time series.

Another way to visualize data is by a *histogram*.

Definition 1.1 (Histogram) *A histogram is a plot obtained by partitioning the range of data into intervals, often equal-sized, called bins, and then plotting a*

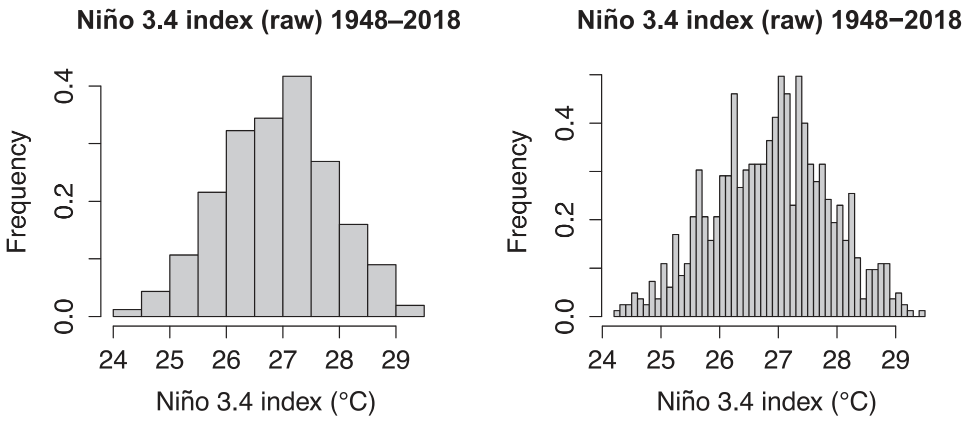


Figure 1.2 Histograms of the monthly mean Niño 3.4 index over the period 1948–2017. The two histograms show the same data, but the left histogram uses a wider bin size than the right.

rectangle over each bin such that the area of each rectangle equals the empirical frequency with which samples fall into the bin. The total area of the rectangles equals one. (Sometimes, histograms may be defined such that the total area of the rectangles equals the total number of samples, in which case the area of each rectangle equals the number of samples that fall into that bin.)

Histograms of the Niño 3.4 index for different bin sizes are shown in Figure 1.2. The figure shows that this index varied between 24°C and 29.5°C over the period 1948–2017. Also, values around 27° occur more frequently than values around 25° or 29°. However, the shape of the histogram is sensitive to bin size (e.g., compare Figures 1.2a and b); hence, the conclusions one draws from a histogram can be sensitive to bin size. There exist guidelines for choosing the bin size, e.g., Sturges' rule and the Freedman–Diaconis rule, but we will not discuss these. They often are implemented automatically in standard statistical software.

The *scatterplot* provides a way to visualize the relation between *two* variables. If X and Y are two time series over the same time steps, then each point on the scatterplot shows the point $(X(t), Y(t))$ for each value of t. Some examples of scatterplots are illustrated in Figure 1.3. Scatterplots can reveal distinctive relations between X and Y. For instance, Figure 1.3a shows a tendency for large values of X to occur at the same time as large values of Y. Such a tendency can be used to *predict* one variable based on knowledge of the other. For instance, if X were known to be at the upper extreme value, then it is very likely that Y also will be at its upper extreme. Figure 1.3b shows a similar tendency, except that the relation is weaker, and therefore a prediction of one variable based on the other would have more uncertainty. Figure 1.3c does not immediately reveal a relation between the two variables. Figure 1.3d shows that X and Y tend to be *negatively* related to each other,

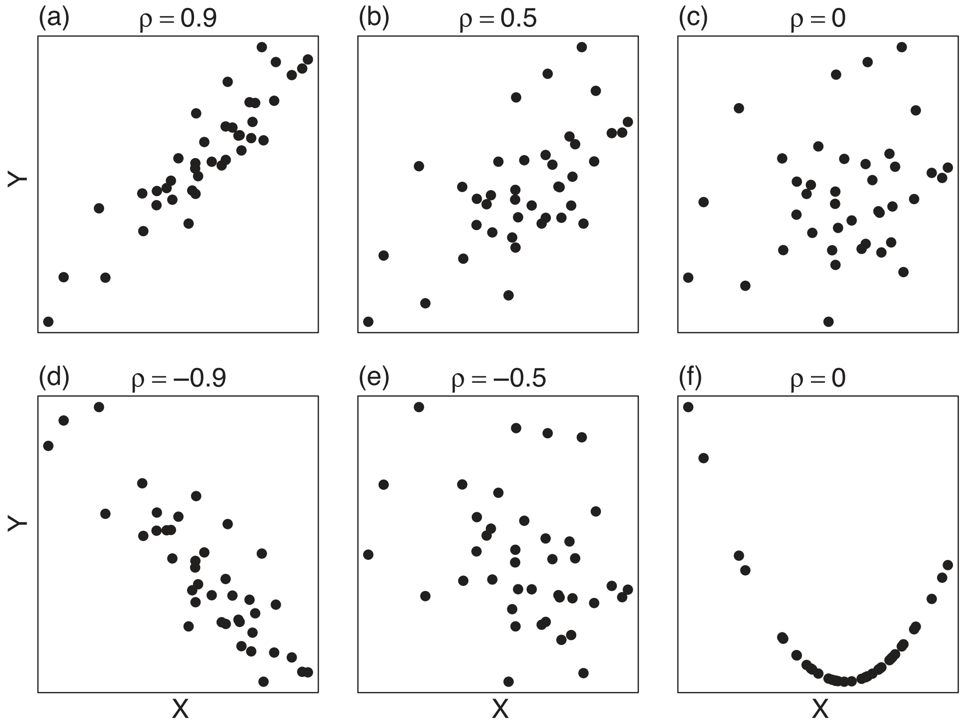


Figure 1.3 Scatterplots of X versus Y for various types of relation. The correlation coefficient ρ, given in the title of each panel, measures the degree of linear relation between X and Y. The data were generated using the model discussed in Example 1.7, except for data in the bottom right panel, which was generated by the model $Y = X^2$, where X is drawn from a standardized Gaussian.

when one goes up, the other goes down. Methods for quantifying these relations are discussed in Section 1.7.

1.2 Measures of Central Value: Mean, Median, and Mode

Visual plots are informative, but ultimately data must be described *quantitatively*. A basic descriptor of a set of numbers is their *central value*. For instance, the central value could be identified with the *most frequent* value, called the *mode*. The mode could be estimated by the location of the peak of a histogram, although this definition would depend on bin size. Also, for the Niño 3.4 time series, each value occurs *only once*, so there is no “most frequent value.” Other measures of central value are the *mean* and *median*. When these quantities are computed from data, the qualifier *sample* is used to emphasize its dependence on data.

Definition 1.2 (Sample Mean) *The sample mean (or average) of N numbers $X_1, \ldots, X_N$ is denoted $\hat{\mu}_X$ and equals the sum of the numbers divided by N*

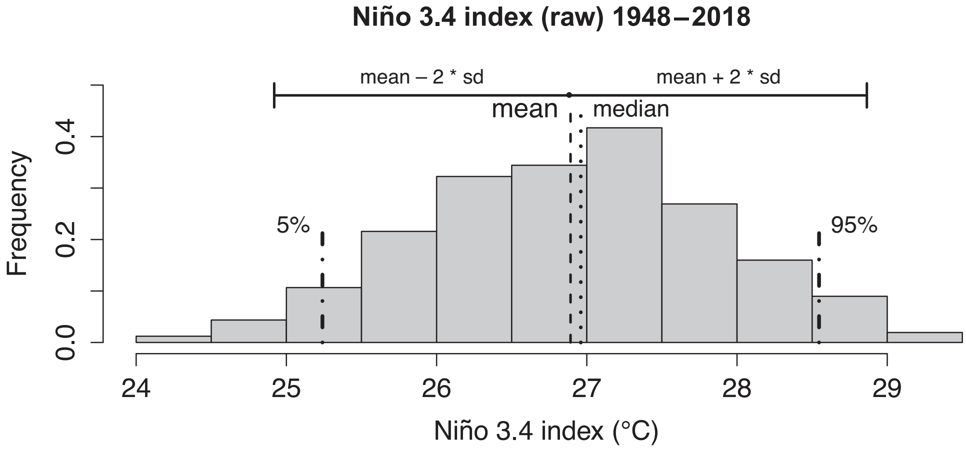


Figure 1.4 Histogram of the monthly mean (raw) Niño 3.4 index over the period 1948–2017, as in Figure 1.2, but with measures of central value and dispersion superimposed. The mean and median are indicated by dashed and dotted vertical lines, respectively. The dash-dotted lines indicate the 5th and 95th percentiles. The horizontal "error bar" at the top indicates the mean plus or minus two standard deviations. The empirical mode is between 27°C and 27.5°C.

$$\hat{\mu}_X = \frac{X_1 + X_2 + \cdots + X_N}{N} = \frac{1}{N}\sum_{n=1}^{N} X_n. \tag{1.1}$$

The mean of the Niño 3.4 index is indicated in Figure 1.4 by the dashed vertical line. The mean is always bounded by the largest and smallest elements.

Another measure of central value is the median.

Definition 1.3 (Sample Median) *The sample median of N numbers $X_1, \ldots, X_N$ is the middle value when the data are arranged from smallest to largest. If N is odd, the median is the unique middle value. If N is even, then two middle values exist and the median is defined to be their average.*

The median effectively divides the data into two equal halves: 50% of the data lie above the median, and 50% of the data lie below the median. The median of the Niño 3.4 index is shown by the dotted vertical line in Figure 1.4 and is close to the mean. In general, the mean and median are equal for symmetrically distributed data, but differ for asymmetrical distributions, as the following two examples illustrate.

Example 1.1 (The Sample Median and Mean for N Odd) ***Question:*** *What is the mean and median of the following data?*

$$2 \quad 8 \quad 5 \quad 9 \quad 3. \tag{1.2}$$

Answer: *To compute the median, first order the data:*

$$2 \quad 3 \quad 5 \quad 8 \quad 9. \tag{1.3}$$

The middle value is 5, hence the median is 5. The mean is

$$\frac{2+3+5+8+9}{5} = 5.4. \tag{1.4}$$

Example 1.2 (The Sample Median and Mean for N Even) ***Question:*** *What is the mean and median of the following data?*

$$2 \quad 8 \quad 5 \quad 9 \quad 3 \quad 100. \tag{1.5}$$

Answer: *To compute the median, first order the data:*

$$2 \quad 3 \quad 5 \quad 8 \quad 9 \quad 100. \tag{1.6}$$

The two middle values are 5 and 8, hence the median is their average, namely 6.5. In contrast, the mean is 21.2, which differs considerably from the median (contrary to example 1.1). Note that if the value of 100 were changed to some higher value X, the median would remain at 6.5 regardless of the value X, but the mean would increase with X. This example shows that the mean is sensitive to extreme values in a data set, whereas the median is not.

1.3 Measures of Variation: Percentile Ranges and Variance

Two data sets can have similar central values but differ by how they *vary* about the central value. Two common measures of variation are *quantile ranges* and *variance*. Sample quantiles are points that divide the sample into equal parts. Common quantiles have special names. For instance, *terciles* divide the sample into three equal parts; *quartiles* divide a sample into four equal parts. One of the most common quantiles is the *percentile*.

Definition 1.4 (Sample Percentiles) *A (sample) percentile is indicated by a number p, such that after the data are ordered from smallest to largest, at least $p \cdot 100\%$ of the data are at or below this value, and at least $100(1-p)\%$ are at or above this value. The resulting value is said to be the $100p$-th percentile (e.g., the 90th percentile corresponds to $p = 0.9$).*

The median is a special case of a percentile: It is the 50th percentile (i.e., $p = 0.5$). The above definition states merely that *at least* $p \cdot 100\%$ of the data lies below the $100p$'th percentile, hence the sample percentile is not unique. There are several definitions of sample quantiles; for instance, Hyndman and Fan (1996) discuss nine different algorithms for computing sample quantiles. The differences between these sample quantiles have no practical importance for large N and will not be of concern in this book. Mathematical software packages such as Matlab, R, and Python have built-in functions for computing quantiles.

The *percentile range* is the interval between two specified percentile points. For instance, the 5–95% range includes all values between the 5th and 95th precentiles. This percentile range is a measure of variation in the sense that it specifies an interval in which a random number from the population will fall 90% of the time. The 5th and 95th percentiles of the Niño 3.4 index are indicated in Figure 1.4 by the two dash-dot lines.

Another measure of variation is the variance.

Definition 1.5 (Sample Variance) *The sample variance of N numbers $X_1, \ldots, X_N$ is denoted $\hat{\sigma}_X^2$ and defined as*

$$\hat{\sigma}_X^2 = \frac{1}{N-1}\sum_{n=1}^{N}(X_n - \hat{\mu}_X)^2, \tag{1.7}$$

where $\hat{\mu}_X$ is the sample mean of the data, defined in (1.1).

The reader ought to be curious why the sum in (1.7) is divided by $N-1$, whereas the sum for the mean (1.1) was divided by N. The reason for this will be discussed in Section 1.10 (e.g., see discussion after Theorem 1.4). Based on its similarity to the definition of the mean, the variance is approximately the average squared difference from the sample mean.

Definition 1.6 (Standard Deviation) *The standard deviation is the (positive) square root of the variance:*

$$\hat{\sigma}_X = \sqrt{\hat{\sigma}_X^2}. \tag{1.8}$$

The standard deviation has the same units as X.

Among the different measures listed above, the ones that will be used most often in this book are the *mean* for central tendency, and the *variance* for variation. The main reason for this is that the mean and variance are algebraic combinations of the data (i.e., they involve summations and powers of the data); hence, they are easier to deal with theoretically compared to mode, median, and percentiles (which require ranking the data). Using the mean and variance, a standard description of variability is the mean value plus and minus one or two standard deviations. For the Niño 3.4 index shown in Figure 1.4, the mean plus or minus two standard deviations is indicated by the error bar at the top of the figure.

Selected Properties of the Sample Variance

If $\hat{\sigma}_X^2$ is the sample variance of $X_1, \ldots, X_N$ and k is a constant, then

- variance of k times each X_n: $\hat{\sigma}_{(kX)}^2 = k^2\hat{\sigma}_X^2$.
- variance of k plus each X_n: $\hat{\sigma}_{(X+k)}^2 = \hat{\sigma}_X^2$.

An identity that is occasionally useful is

$$\hat{\sigma}_X^2 = \left(\hat{\mu}_{(X^2)} - \hat{\mu}_X^2\right)\frac{N}{N-1}. \tag{1.9}$$

Numerically, computation of sample variance based on (1.7) requires *two* passes of the data: one to compute the mean, and a second to compute deviations from the

Cambridge University Press
978-1-108-47241-8 — Statistical Methods for Climate Scientists
Timothy DelSole , Michael Tippett
Excerpt
More Information

mean. With (1.9), the sample variance can be computed from one pass of the data, but requires tracking two quantities, namely the means of X and X^2. The sample variance is nonnegative, but in practice (1.9) can be (slightly) negative owing to numerical precision error.

1.4 Population versus a Sample

An observation is defined as the outcome of an experiment performed on nature. We will conceive of a theoretical collection of all possible outcomes, and then interpret an observation as a random draw from this theoretical collection. The theoretical collection of all possible observations is called the *population*, while a random draw from this collection is called a *sample* or *realization*. The goal of statistics is to make inferences or decisions about a population based on information derived from a sample.

In nature, population properties are never known with complete certainty. Knowledge of population properties is tantamount to knowledge of the "inner machinery" of the system. Except in idealized settings, we never know the inner workings of the system on which we experiment, and therefore we can never be sure about the population properties. Rather, we can only *infer* population properties based on the outcome of experiments. We might attempt to approximate the population probability of an event by measuring the relative frequency with which the event occurs in a large number of independent samples, but this approach meets fundamental difficulties with defining "large," "approximate," and "independent." These and other subtle problems can be avoided by *defining* probability in axiomatic terms, much like geometry is developed strictly from a set of axioms and rules of logic. This is the approach mathematicians have adopted. For the problem considered in this book, this axiomatic abstraction is not required. Therefore, we briefly review basic concepts in probability theory that are needed to get started. Most text books on statistics and probability cover these concepts in detail and can be consulted for further information.

1.5 Elements of Probability Theory

What is the probability of tossing a fair coin and getting heads? A typical 10-year-old child knows that the probability is 50%. However, that same 10-year-old child can become confused by an experiment where 6 out of 10 tosses are heads, since 6/10 is not 50%. The child eventually learns that "50% probability" refers to the idea that in a *long sequence* of coin tosses the relative frequency of heads *approaches* 50%. However, the relative frequency of heads in a small number of experiments can differ considerably from 50%. Asserting that heads occurs with 50% probability is tantamount to asserting knowledge of the “inner machinery” of nature. We refer to the “50% probability” as a *population property*, to distinguish it from the results of a particular experiment, e.g., “6 out of 10 tosses,” which is a *sample property*. Much confusion can be avoided by clearly distinguishing population and sample properties. In particular, it is a mistake to equate the relative frequency with which an event occurs in an experiment with the probability of the event in the population.

A *random variable* is a function that assigns a real number to each outcome of an experiment. If the outcome is numerical, such as the temperature reading from a thermometer, then the random variable often is the number itself. If the outcome is not numerical, then the role of the function is to assign a real number to each outcome. For example, the outcome of a coin toss is heads or tails, i.e., not a number, but a function may assign 1 to heads and 0 to tails, thereby producing a random variable whose only two values are 0 and 1. This is an example of a *discrete* random variable, whose possible values can be counted. In contrast, a random variable is said to be *continuous* if its values can be any of the infinitely many values in one or more line intervals.

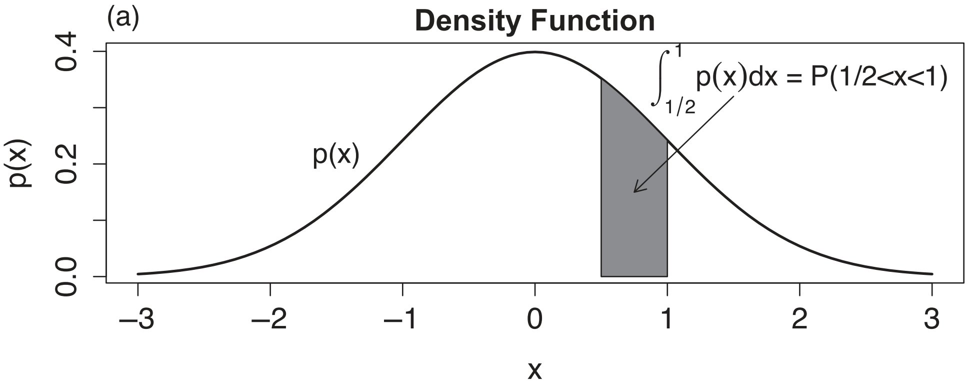
Sometimes a random variable needs to be distinguished from the value that it takes on. The standard notation is to denote a random variable by an uppercase letter, i.e. X, and denote the specific value of a random draw from the population by a lowercase letter, i.e. x. We will adopt this notation in this chapter. However, this notation will be adhered to only lightly, since later we will use uppercase letters to denote matrices and lowercase letters to denote vectors, a distinction that is more important in multivariate analysis.

If a variable is discrete, then it has a countable number of possible realizations $X_1, X_2, \ldots$. The corresponding probabilities are denoted $p_1, p_2, \ldots$ and called the *probability mass function*. If a random variable is continuous, then we consider a class of variables X such that the probability of $\{x_1 \leq X \leq x_2\}$, for all values of $x_1 \leq x_2$, can be expressed as

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p_X(x)dx, \tag{1.10}$$

where $p_X(x)$ is a nonnegative function called the *density function*. By this definition, the probability of X falling between x_1 and x_2 corresponds to the *area under the density function*. This area is illustrated in Figure 1.5a for a particular distribution. If an experiment always yields some real value of X, then that probability is 100% and it follows that

$$\int_{-\infty}^{\infty} p_X(x)dx = 1. \tag{1.11}$$



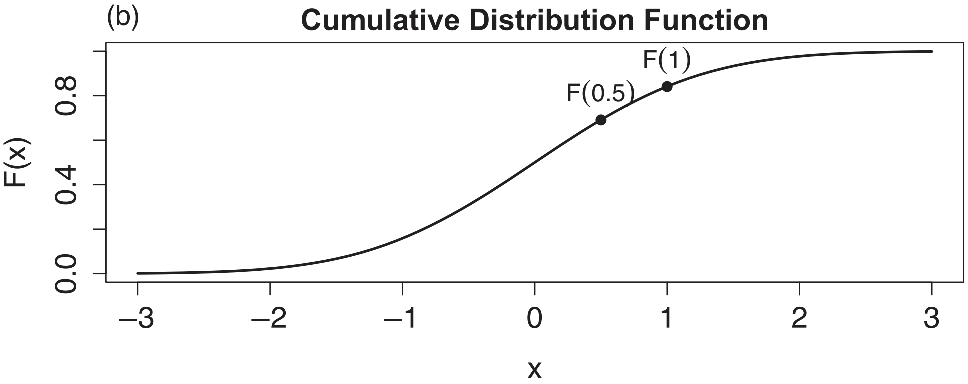


Figure 1.5 Schematic showing (a) a probability density function for X and the fact that the probability that X lies between 1/2 and 1 is given by the area under the density function $p(x)$, and (b) the corresponding cumulative distribution function $F(x)$ and the values at $x = 0.5$ and $x = 1$, the difference of which equals the area of the shaded region in (a).

The histogram provides an estimate of the density function, provided the histogram is expressed in terms of relative frequencies. Another function is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} p_X(u)du, \tag{1.12}$$

which is called the *cumulative distribution function* and illustrated in Figure 1.5b. The probability that X lies between x_1 and x_2 can be expressed equivalently as

$$P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1). \tag{1.13}$$

The above properties do not uniquely specify the density function $p_X(x)$, as there is more than one $p_X(x)$ that gives the same left-hand side of (1.10) (e.g., two density functions could differ at isolated points and still yield the same probability of the