Chapter Three

Random Variables and Probability Distributions

3.1 Introduction

An event is defined as the possible outcome of an experiment. In engineering applications, the outcomes are usually associated with quantitative measures, such as the time-to-failure of a product, or qualitative measures, such as whether a product is safe or risky. When considering the continuous quantitative measurements, we use a quantity X, which varies in a certainty range, including $-\infty < X < \infty$, to denote a random event measurement. The variable X is also called a *continuous random variable*. If X can take on only limited values, it is called a *discrete random variable*. We will discuss only continuous random variables herein.

The following symbol convention is used throughout this course. An uppercase letter denotes a random variable; a lowercase letter denotes an observation (or a realization) of a random variable, or a deterministic variable; and a bold letter denotes a vector. For instance, X stands for a random variable; x denotes a realization of X. $\mathbf{X}$ stands for a vector of random variables, and $\mathbf{x}$ stands for a vector of realizations of $\mathbf{X}$ or a vector of deterministic quantities.

Next, we will introduce how to use a cumulative distribution function or probability density function to fully describe a random variable X.

3.2 Cumulative Distribution Function and Probability Density Function

For a physical quantity, the possible outcomes are usually within a range of measured or observed values. For example, if the nominal value of the length of a shaft is 100 mm, and its manufacturing tolerance is 0.1 mm, the actual length will be within the range of 100 ± 0.1 mm. When the length is measured, its actual values may vary from 99.90 mm to 100.10 mm. 100 sample measurements of the length are given in Table 3.1. As shown in the table, within the range from 99.90 mm to 100.10 mm, curtain values occur more frequently than others. The values around the nominal length 100 mm occur with a higher chance than the values near both endpoints.

If we divide the range [99.90, 100.10] into several equal segments and plot the number of values of the length that reside the segments, we will have a bar-like graph (see Fig. 3.1). This type of graph is called a *histogram*. It shows the frequency of the values that occur in different segments.

Table 3.1 100 Measurements of the Beam Length

99.90	99.90	99.93	99.94	99.95	99.95	99.95	99.95	99.95	99.96
99.96	99.96	99.96	99.96	99.96	99.96	99.96	99.96	99.96	99.97
99.97	99.97	99.97	99.97	99.97	99.97	99.97	99.97	99.97	99.98
99.98	99.98	99.98	99.98	99.98	99.98	99.99	99.99	99.99	99.99
99.99	99.99	99.99	100.00	100.00	100.00	100.00	100.00	100.00	100.00
100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.01
100.01	100.01	100.01	100.01	100.01	100.01	100.01	100.01	100.01	100.01
100.01	100.02	100.02	100.02	100.02	100.02	100.02	100.02	100.02	100.02
100.02	100.02	100.03	100.03	100.03	100.03	100.03	100.03	100.03	100.04
100.04	100.04	100.04	100.04	100.04	100.05	100.05	100.05	100.06	100.08

From the histogram, we see that it is more likely that the values of the length are around the nominal value 100 mm.

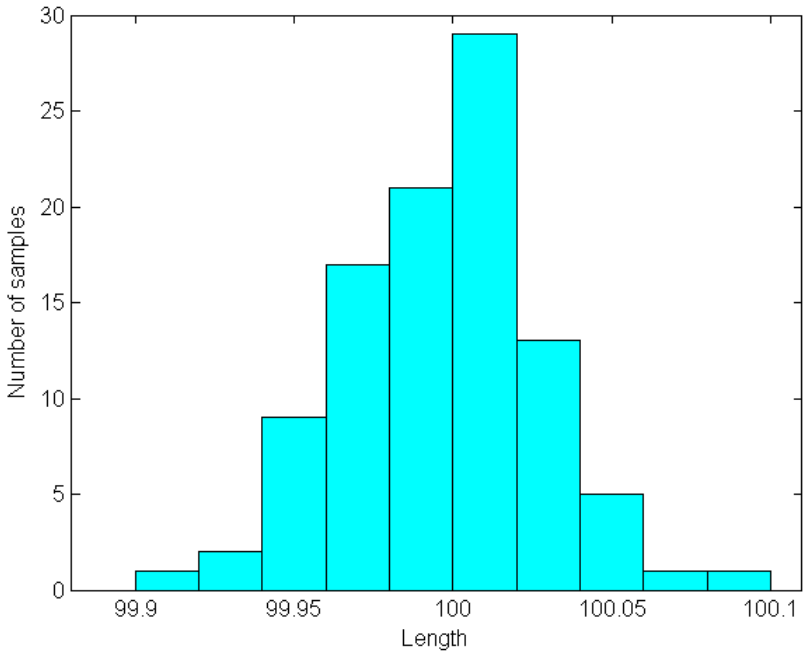

Figure 3.1 Histogram of the Length

If we plot the number of samples (measurements) divided by the total number of measurements, we obtain a variant of the histogram. As shown in Fig. 3.2, the vertical axis represents the number of measurements within each segment divided by the total number of measurements (100). Obviously, Fig. 3.2 is a scaled version of Fig. 3.1.

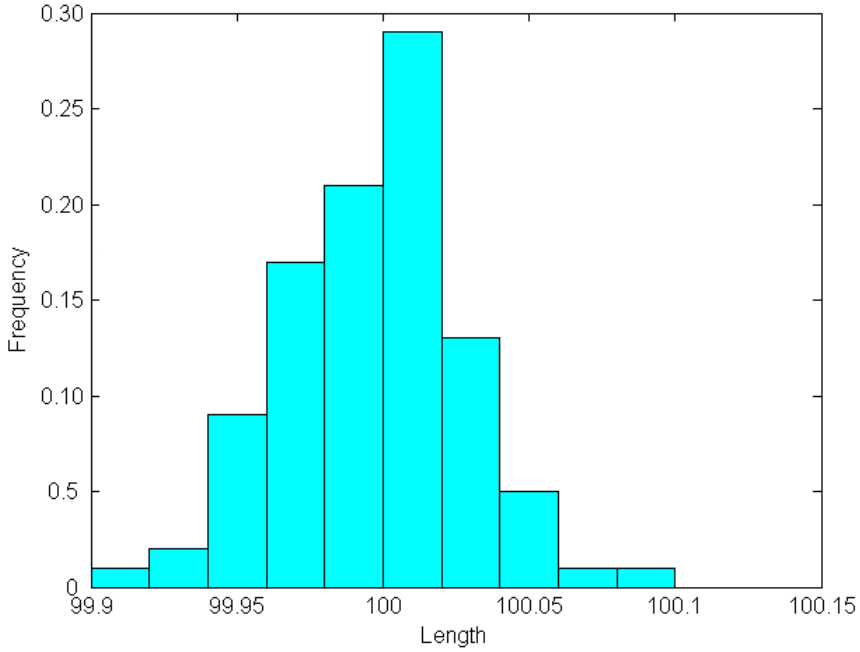

Figure 3.2 Histogram of the Length

If we have more samples and use more intervals to divide the range of the length, the bars in Fig. 3.2 will approach a smooth curve as shown in Fig. 3.3. This curve is called a probability density function (pdf).

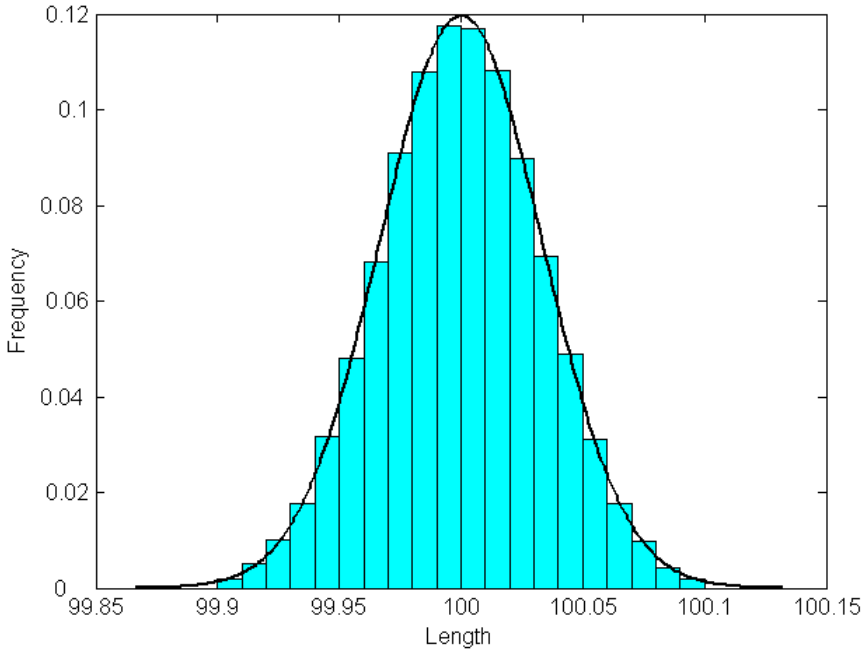

Figure 3.3 Histogram of the Length with More Samples

The pdf captures the chance property of a random variable as shown in Fig. 3.4 and fully describes a random variable. $f(x)$ is used the denote a probability density function of random variable X, where x is a realization (a specific value) of X. The significance of the pdf is that $f(x)dx$ is the probability that the random variable X is in the interval $[x, x+dx]$ (see Fig. 3.4), written as

$$\mathrm{P}(x \le X \le x+dx) = f(x)dx \tag{3.1}$$

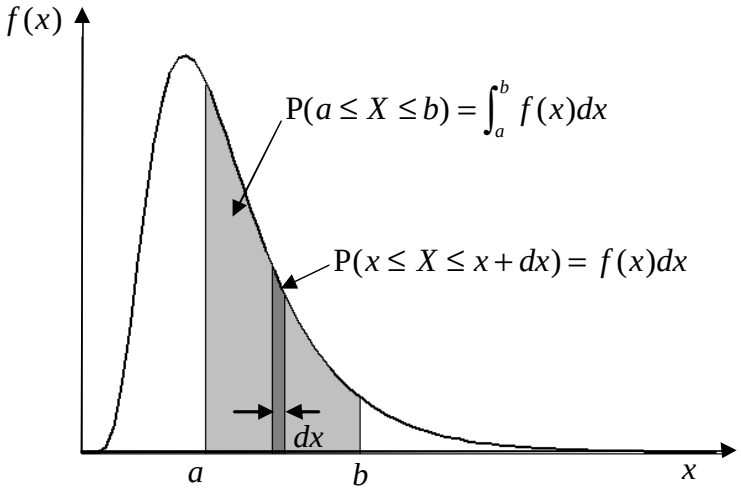

Figure 3.4 Probability Density Function

We can also determine the probability of X over a finite interval $[a,b]$ as

$$\mathrm{P}(a \le X \le b) = \int_a^b f(x)dx \tag{3.2}$$

which is the area underneath the curve of $f(x)$ from $x=a$ to $x=b$ (see Fig. 3.4).

A pdf must be nonnegative, i.e.

$$f(x) \ge 0 \tag{3.3}$$

and satisfies the following condition

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \tag{3.4}$$

Eq. 3.4 indicates that the area underneath the pdf curve is 1. In other words, the probability of X taking all possible values is equal to 1.0.

In addition to pdf, the cumulative distribution function (cdf) is also commonly used. It is defined as the probability that the random variable X is less than or equal to a constant x, namely,

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(x)dx \tag{3.5}$$

As shown in Fig. 3.5, the cdf $F(x)$ is the area underneath the pdf curve in the range of $(-\infty, x]$.

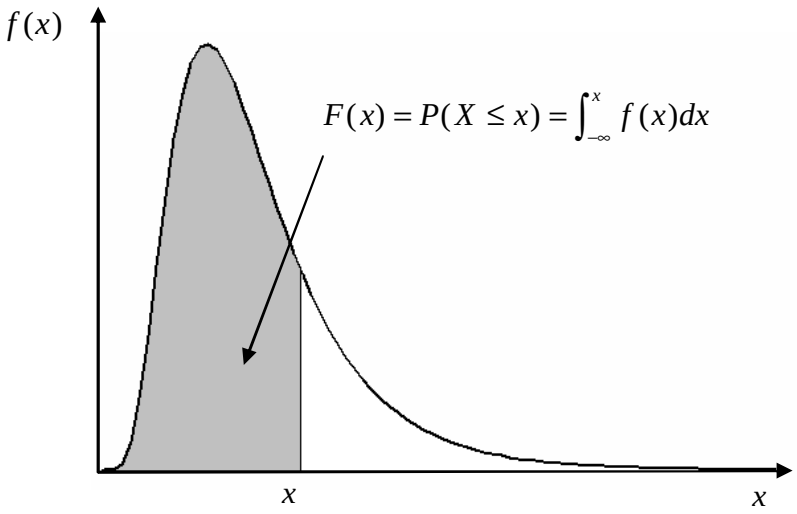

Figure 3.5 Probability Density Function

Note that since $f(x) \ge 0$ and the integral of $f(x)$ is normalized to unity, $F_X(x)$ possesses the following features:

- $F(x)$ is a nondecreasing function of x and $F_X(x) \ge 0$
- $F(-\infty) = 0$
- $F(+\infty) = 1$

Fig. 3.6 shows the cdf which corresponds to the pdf depicted in Fig. 3.4.

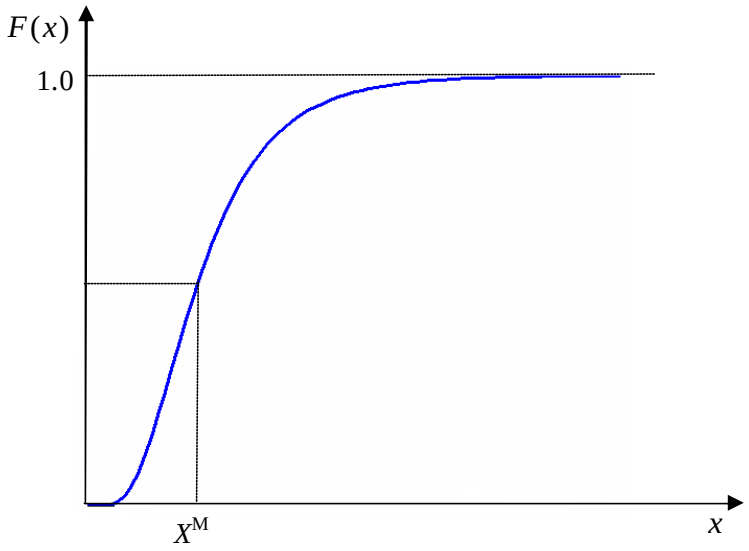

Figure 3.6 Cumulative Probability Function

Eq. 3.5 gives the mathematical relationship between the pdf and cdf. Conversely, the pdf can be derived from the cdf with the following equation

$$f(x) = \frac{d[F(x)]}{dx} \tag{3.6}$$

3.3 Population and Sample

The distribution we discussed above is referred to as *population distribution*. By definition, a population is any entire collection of objects from which we may collect data. It is the entire group in which we are interested, and about which we wish to describe or draw conclusions. If we use the concept of the set discussed in Section 2.3, the population can be viewed as a universal set. We use the pdf and cdf given above to describe a population distribution.

Because a population is too large to study in its entirety, usually a group of units selected from the population is used to draw conclusions about the population, such as the distribution shape and location. This group of units selected from the population is called a *sample* of that population. The sample should be representative of the general population. This is often best achieved by random sampling.

For example, to understand the population of the length of the aforementioned shaft, 100 samples were collected randomly as shown in Table 3.1. These samples can be used to study the population of the length by using statistical tools such as the histogram drawn in Fig. 3.1.

3.4 Moments

Even though a cdf or pdf can fully describe a random variable *X*, neither of them may be straightforward enough for a direct interpretation. For convenience, we frequently use other additional parameters which can be derived from the cdf or pdf. The most important parameters are the moments, including

- *mean*, which is the first moment about the origin
- *variance*, which is the second moment about the mean
- *skewness*, which is the third moment about the mean

The k-*th* moment about the origin is given by

$$M_k^{'} = \int_{-\infty}^{+\infty} x^k f_X(x) dx \tag{3.7}$$

The k-*th* moment about the mean m_X is given by

$$M_k = \int_{-\infty}^{+\infty} (x - m_X)^k f_X(x) \tag{3.8}$$

The mean m_X is defined below.

3.4.1 Mean

The *mean value*, also known as the *expected value*, or *population mean*, is defined as the first moment measured about the origin

$$m_X = \int_{-\infty}^{+\infty} x f_X(x) dx \tag{3.9}$$

If there are *n* observations (samples) of the random variable *X*, $(x_1, x_1, \cdots, x_n)$, the average of the samples (sample mean) is calculated by

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3.10}$$

As the sample size *n* increases, the sample mean $\overline{X}$ will approach the population mean (the expected value) m_X. Therefore, the expected value m_X is the long-run average of random variable *X*. We can use a sample mean to estimate a population mean.

The 100 samples of the shaft length in Table 3.1 were drawn from a population with its mean $m_X = 100$ mm. The sample mean of the length is calculated by

$$\overline{X} = \frac{1}{100}\sum_{i=1}^{100} x_i = 99.96 \tag{3.11}$$

In this case, it is seen that the sample mean is close to the population mean.

3.4.2 Variance

The variance is the second moment about the mean. It is an indication of how the individual measurements scatter around its mean. The population variance is defined as

$$s^2 = \int_{-\infty}^{+\infty} (x - m_X)^2 f_X(x) dx \tag{3.12}$$

When *n* observations $(x_1, x_1, \cdots, x_n)$ are available, the sample variance can be calculated by

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{X})^2 \tag{3.13}$$

The value of the variance given by the above equation is biased. When the number of samples *n* approaches infinity, the estimate will not converge to the population variance s^2. The unbiased sample variance is then used and is given by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{X})^2 \tag{3.14}$$

The sample variance in the above equation will approach the population mean when the sample size *n* increases.

The use of a variance as a descriptor is not obvious due to its unit, which is the square of the unit of the random variable. It is not the same as the unit of either the random variable or its mean. Therefore, the square root of the variance is usually used and is called the standard deviation with the following formulation.

$$s = \sqrt{\int_{-\infty}^{+\infty} (x - m_X)^2 f_X(x) dx} \tag{3.15}$$

Similarly, the sample standard deviation is calculated by

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2} \tag{3.16}$$

Using the 100 samples in Table 3.1, we can calculate the sample variable and standard deviation of the shaft length. The results are $S^2 = 0.0324\ \text{mm}^2$ and $S = 0.18\ \text{mm}$. These two values can be used as the estimates of the population variance s^2 and standard deviation s, respectively.

The standard deviation is a measure of how a distribution spreads out; it is used to characterize the dispersion among the measures in a given population. Suppose that two shafts have the same mean value of the length $m_X = 100\ mm$. But their standard deviations of length are different: $s_1 = 0.0034$ and $s_2 = 0.0068$. Since the first shaft has a smaller standard deviation, its length is distributed more narrowly than the second shaft (see Fig. 3.7). Because of this, with the same other conditions, the variation of the performance (such as stress and deflection which are functions of the length) of the first shafts will be smaller than that of the second shaft. In this sense, we may say that the first shaft has higher quality (or is more robust) than the second shaft. The example shows that the standard deviation is an important indicator of quality or robustness.

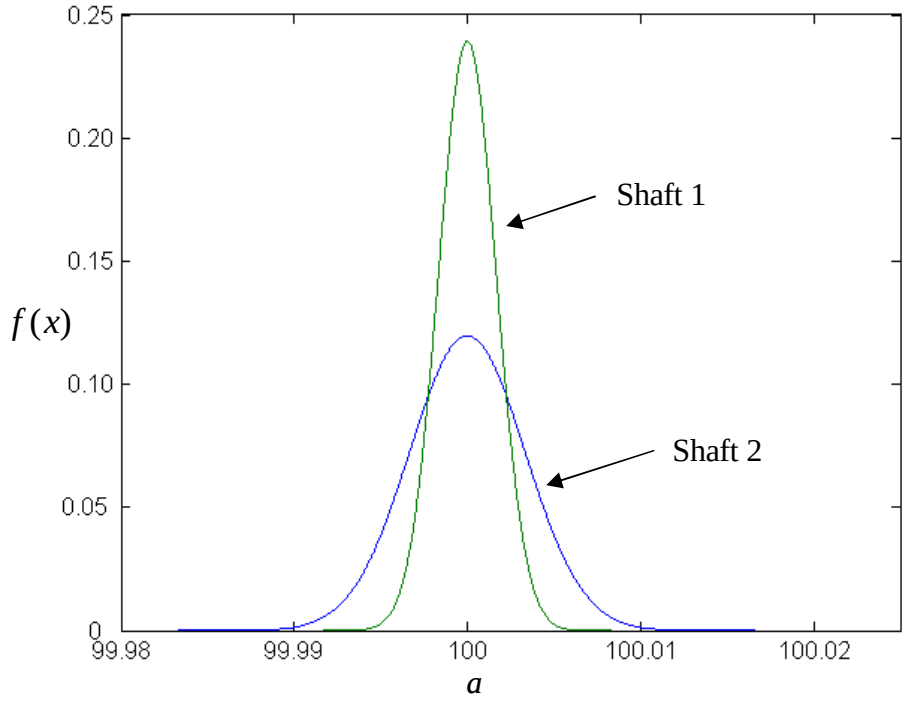

Figure 3.7 pdfs of Two Shafts

3.4.3 Skewness

The skewness is defined as the third moment about the mean with the following equation,

$$g_{0X} = \int_{-\infty}^{+\infty} (x - m_X)^3 f_X(x) dx \tag{3.18}$$

A nondimensional measurement of the skewness known as the *skewness coefficient* is defined as

$$g_X = \frac{g_{0X}}{s_X^3} \tag{3.19}$$

The skewness describes the degree of asymmetry of a distribution. A symmetric distribution has a skewness of zero, while an asymmetric distribution has a nonzero skewness. If more extreme tail of the distribution is to the right of the mean, the skewness is positive; if the more extreme tail is to the left of the mean, the skewness is negative. The skewness is illustrated in Fig. 3.8.

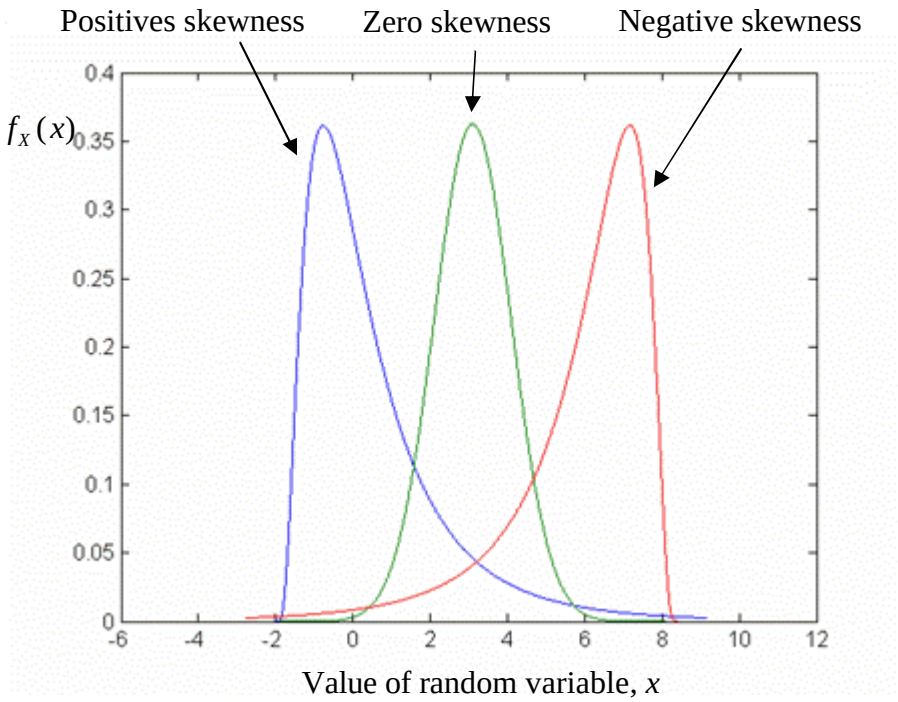

Figure 3.8 Skewness of Distributions

3.4.4 Median

The median of a population X_m is the point that divides the distribution of a random variable in half. Numerically, half of the measurements in a population will have values that are equal to or larger than the median, and the other half will have values that are equal to or smaller than the median.

If the cdf of a random variable is given, the median can be found by the fact that at the median, the cdf is equal to 0.5, i.e.

$$F_X(X_m) = 0.5 \tag{3.20}$$

The population mean is demonstrated in Fig. 3.6.

To find the *median* from a set of samples, we need first to arrange all the samples from lowest value to highest value and then pick the middle one(s). If there are an even number of samples, we take the average of the two middle values.

For example, there are two sets of samples, A = (3.2, 5, 2, 6.5, 7) and B = (3.2, 5, 2, 6.5, 7, 8). First we sort the samples as A = (2, 3.2, 5, 6.5, 7) and B = (2, 3.5, 5, 6.5, 7, 8). Then, we calculate the medians. The median of A is 5 and that of B is (5 + 6.5) /2 = 5.75.

3.4.5 Percentile Value

A percentile value X^a is a value below which the probability of the actual values of random variable X less than X^a is a, i.e.

$$P(X \le X^a) = F_X(X^a) = \int_{-\infty}^{X^a} f(x)dx = a \tag{3.21}$$

The percentile value is illustrated in Fig. 3.9. It is shown that the shaded area under the pdf curve is equal to a.

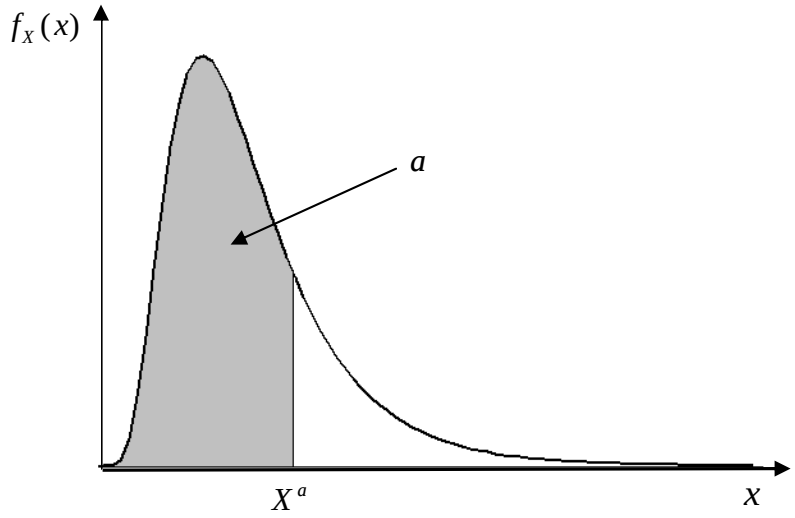

Figure 3.9 Percentile Value of a Distribution

3.5 Jointly Distributed Random Variables

When two or more random variables are being considered simultaneously, their joint behavior is determined by their joint probability distribution function. We will first discuss the situation of two random variables. The discussions can be easily extended to the general situation where more than two random variables are involved.

3.5.1 Joint density and distribution functions

The joint cdf of two random variables X and Y is defined as

$$F_{X,Y}(x,y) = P(X \le x, Y \le y) \tag{3.22}$$

The joint cdf obeys following conditions:

- $F_{X,Y}(-\infty,-\infty) = 0$ (3.23)
- $F_{X,Y}(x,-\infty) = 0$ (3.24)
- $F_{X,Y}(-\infty, y) = 0$ (3.25)
- $F_{X,Y}(+\infty,+\infty) = 1$ (3.26)
- $F_{X,Y}(x,+\infty) = F_X(x)$ (3.27)
- $F_{X,Y}(+\infty, y) = F_Y(y)$ (3.28)
- $F_{X,Y} \ge 0$ (3.29)
- $F_{X,Y}$ is a non-decreasing function of X and Y.

The joint pdf is given by

$$f_{X,Y}(x,y) = \frac{\partial F_{X,Y}(x,y)}{\partial x \partial y} \tag{3.30}$$

If the joint pdf is given, the joint cdf can be calculated by

$$F_{X,Y}(x,y) = P(X \le x, Y \le y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(x,y)dxdy \tag{3.31}$$

3.5.2 Marginal density function

Knowing the joint pdf, we can obtain the individual pdf, called *marginal pdf*.

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy \tag{3.32}$$

and

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dx \tag{2.33}$$

3.5.3 Covariance and correlation

Similar to the variance of a single random variable, the *covariance* of two random variables X and Y, denoted as $\mathrm{Cov}(X,Y)$, is the second moment about their respective means m_X and m_Y.

$$\mathrm{Cov}(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(X - m_X)(Y - m_Y) f_{X,Y}(x,y)dx \tag{2.34}$$

The covariance of two random variables X and Y provides a measure of how the two random variables are linearly correlated, and it hence indicates the linear relationship between the two random variables. The derived dimensionless quantity, known as correlation coefficient, is usually used, which is given by

$$r_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{s_X s_Y} \tag{2.35}$$

Values of $r_{X,Y}$ range between -1 and +1.

- $r_{X,Y} = 0$, there is no linear relationship between X and Y.

- $0 < r_{X,Y} < 1$, there is a positive relationship between X and Y; Y increases as X increases.
- $-1 < r_{X,Y} < 0$, there is a negative relationship between X and Y; Y decreases as X increases.
- $r_{X,Y} = 1$, there is a perfect positive linear relationship between X and Y; Y linearly increases as X increases.
- $r_{X,Y} = -1$, there is a perfect negative linear relationship between X and Y; Y linearly decreases as X increases.

Appendix
MATLAB Statistics Toolbox

The MATLAB Statistics Toolbox is a collection of statistical tools built on the MATLAB numeric computing environment. The toolbox supports a wide range of common statistical tasks, such as random number generation, curve fitting, Design of Experiments, and statistical process control.

If a set of samples of a random variable exists, we can use the following functions to study the samples.

mean(X) – average or mean value

For a vector x, mean(x) is the mean value of the samples in x. For a matrix x, mean(x) returns a row vector containing the mean value of each column in x.

std(X) – standard deviation

For a vector x, std(x) returns the standard deviation. . For a matrix x, std(x) returns a row vector containing the standard deviation of each column in x.

skewness(X) – skewness coefficient

For a vector x, skewness(x) returns the sample skewness. . For a matrix x, skewness(x) returns a row vector containing the sample skewness of each column in x.

moment(X) – central moments of all orders

moment(x, order) returns the central moment of a vector x specified by the positive integer, order. For matrix, x, moment(x, order) returns central moments of the specified order for each column in x.