Appendix A

Probability and Statistics

A.1 INTRODUCTION

A *random variable* X is a real-valued function associated with the (chance) outcome of an experiment. That is, to each possible outcome of the experiment, we associate a real number x. The collection of numbers x_i is called the domain of the random variable X. A number x is referred to as a possible value or realization or sample value of the random variable X. Generally for this appendix, uppercase letters will represent random variables and lowercase letters will represent numerical values or realizations of the random variable.

Example 1:
A coin is flipped n times, let X denote the number of times heads appears. The domain of X is the set of numbers

$$\{x = 0, 1, \ldots, n\}.$$

Example 2:
A point on the surface of the Earth is chosen at random. Let X and Y denote its latitude and longitude. The domain of the random variable X and Y is the pair of numbers

$$\{X, Y, -\pi/2 \leq X \leq \pi/2,\ 0 \leq Y \leq 2\pi\}.$$

These two examples indicate the two types of random variables, discrete and continuous. The discrete random variable X has as its domain or possible outcomes the finite number of real numbers $x_1, x_2 \ldots x_n$ (see Example 1). For a continuous random variable, X has as its domain or outcome an interval on the real line, and to each of these intervals a probability may be assigned. The probability at a point

is zero. In Example 2, there are an infinite number of real numbers between 0 and 2π; hence, the probability of obtaining any one of them is zero. Consequently, we must speak of the probability of X over an interval.

A.2 AXIOMS OF PROBABILITY

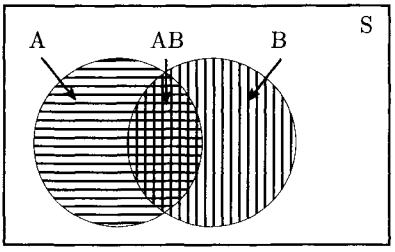
To define a probability associated with the possible values of a random variable, we introduce the fundamental axioms of probability. Let S represent the set of all possible outcomes of a trial or experiment. S is called the sample space of the experiment. Let A be any subset of points of the set S (i.e., a collection of one or more possible trials). For example, in one toss of a die the certain event S would be $s = 1, 2, 3, 4, 5, 6$, and a subset, A, may be $a = 3$.

Write p for a function that assigns to each event A a real number $p(A)$, called the probability of A; that is, $p(A)$ is the numerical probability of the event A occurring. For the previous example, $p(A) = 1/6$.

In modern probability theory, the following minimal set of axioms is generally met (Helstrom, 1984)

1. $p(A) \geq 0$
2. $p(S) = 1$
3. $p(A + B) = p(A) + p(B)$; A and B are any pair of mutually exclusive events.

Here, $p(A + B)$ indicates the probability of A or B or both occurring. For axiom 3, A and B cannot both occur since the events are mutually exclusive. If the events are not mutually exclusive, $p(A + B) = p(A) + p(B) - p(AB)$, where $p(AB)$ is the probability that both events A and B occur. Note that for mutually exclusive events $p(AB) = 0$ (by definition, both events cannot occur simultaneously). A *Venn diagram* (Freeman, 1963) can be used to demonstrate some of the concepts associated with probability. Consider an event A, then we define the complement of A as the event that A does not happen and denote it as $\overline{A}$. Thus A and $\overline{A}$ are said to be mutually exclusive—if A occurs, $\overline{A}$ cannot occur.



If A and B are any two events in a sample space S, we define AB to be the intersection of events A and B and assign to it all points in the space that belongs to both A and B; that is, the occurrence of AB implies that both events A and B have occurred. The intersection of A and B (meaning that both events have occurred) is sometimes denoted as $A \cap B = AB$. We define the union of A and B (written as $A \cup B = A + B$) as all points in the space for which A or B or both have occurred. For the *Venn diagram* shown,

S	:	all points in the rectangle
A	:	points in horizontal hatch region
B	:	points in vertical hatch region
AB	:	points in cross hatch region
$A + B$	:	all points in hatched area

Example:
Throw 1 red and 1 green die. What is probability that at least one "1" will appear? Let

$A =$	1 on red die,
$\overline{A} =$	2–6 on red die,
$B =$	1 on green die,
$\overline{B} =$	2–6 on green die.

The probability of at least one "1" occurring is the union of events A and B, $p(A + B)$. The probability of A or B or both occurring is seen from the Venn diagram to be

$$p(A + B) = p(A) + p(B) - p(AB). \tag{A.2.1}$$

Note that $p(A) + p(B)$ includes $2 \times p(AB)$. Now,

$$
\begin{aligned}
p(A) &= 1/6 \\
p(\overline{A}) &= 5/6 \\
p(B) &= 1/6 \\
p(\overline{B}) &= 5/6 \\
p(AB) = p(A)p(B) &= 1/36 .
\end{aligned}
$$

Hence,

$$p(A+B) = 1/6 + 1/6 - 1/36 = 11/36 .$$

Note that the events AB, $A\overline{B}$, $\overline{A}B$, $\overline{AB}$ are exhaustive (i.e., one must occur), and they are exclusive (i.e., only one can occur). Hence,

$$
\begin{aligned}
p(AB,\ A\overline{B},\ \overline{A}B,\ \overline{A}\,\overline{B}) &= p(AB) + p(A\overline{B}) + p(\overline{A}B) + p(\overline{A}\,\overline{B}) \\
&= 1/36 + 5/36 + 5/36 + 25/36 = 1 .
\end{aligned}
$$

Also, the probability of at least one "1" occurring can be expressed as

$$
\begin{aligned}
p(AB) + p(A\overline{B}) + p(\overline{A}B) &= p(A)p(B) + p(A)p(\overline{B}) + p(\overline{A})p(B) \\
&= (1/6)(1/6) + (1/6)(5/6) + (1/6)(5/6) \\
&= 11/36 .
\end{aligned}
$$

Furthermore, the probability of two "1"s occurring is given by events A and B both occurring,

$$p(AB)^* = p(A)\,p(B) = 1/36 .$$

Carrying on with this example, let

$A =$	1 on red die	$D =$	1 on green die
$B =$	2 on red die	$E =$	2 on green die
$C =$	3, 4, 5, 6 on red die	$F =$	3, 4, 5, 6 on green die

where $p(A) = p(B) = 1/6$ and $p(C) = 4/6$.

The probability that a "1" or a "2" occurs on either or both die (call this event G) is given by

$$
\begin{aligned}
p(G) = p(AD) &+ p(AE) + p(AF) + p(BD) \\
&+ p(BE) + p(BF) + p(DC) + p(EC) \\
= 1/36 &+ 1/36 + 4/36 + 1/36 + 1/36 \\
&+ 4/36 + 4/36 + 4/36 \\
= 20/36 .&
\end{aligned}
$$

*Events A and B are independent (see definition in A.3).

Note that $p(G)$ also can be described by

$$\begin{aligned} p(G) &= p(A) + p(B) + p(CD) + p(CE) \\ &= 1/6 + 1/6 + 4/36 + 4/36 = 20/36. \end{aligned}$$

If events A and B occur, it does not matter what happens on the green die, but if event C occurs, then either event D or E must occur. As stated earlier, if two events A and B are mutually exclusive, or disjoint, then the intersection, $A \cap B$, has no elements in common and $p(AB) = 0$; that is, the probability that both A and B occur is zero. Hence,

$$p(A + B) = p(A) + p(B). \tag{A.2.2}$$

For example, if we throw one die and let

$$\begin{aligned} A = &\quad \text{a "1" appears} \\ B = &\quad \text{a "2" appears.} \end{aligned}$$

Then

$$p(A + B) = p(A) + p(B) = 2/6.$$

A.3 CONDITIONAL PROBABILITY

In wide classes of problems in the real world, some events of interest are those whose occurrence is conditional on the occurrence of other events. Hence, we introduce *conditional probability*, $p(A/B)$. By this we mean the probability that event A occurs, given that event B has occurred. It is given by

$$p(A/B) = \frac{p(AB)}{p(B)}. \tag{A.3.1}$$

We say that two events A and B are independent if

$$p(A/B) = p(A) \text{ and } p(B/A) = p(B); \tag{A.3.2}$$

that is,

$$p(AB) = p(A)p(B). \tag{A.3.3}$$

A.4 PROBABILITY DENSITY AND DISTRIBUTION FUNCTIONS

For a continuous random variable, we assume that all events of practical interest will be represented by intervals on the real line and to each of these intervals

a probability may be assigned. Recall that the probability at a point is zero. We define a *probability density function*, $f(x)$, which represents the probability of X assuming a value somewhere in the interval $(x,\ x + dx)$. We define probability over the interval $x, x + dx$ in terms of area,

$$p(x \leq X \leq x + dx) = f(x)dx\,. \tag{A.4.1}$$

For the continuous random variable, axioms 1 and 2 become

1. $f\,(x) \geq 0$ (A.4.2)

2. $\displaystyle\int_{-\infty}^{\infty} f\,(x)\,dx = 1\,.$ (A.4.3)

The third axiom becomes

3. $p\,(a \leq X \leq c) = \int_a^c f\,(x)\,dx$ (A.4.4)

which for $a < b < c$

$$\begin{aligned} p\,(a \leq X \leq c) &= \int_a^b f\,(x)\,dx + \int_b^c f\,(x)\,dx \\ &= p\,(a \leq X \leq b) + p\,(b \leq X \leq c)\,. \end{aligned} \tag{A.4.5}$$

Note that for the continuous random variable we need not distinguish between $a \leq X \leq c$, $a \leq X < c$, $a < X \leq c$, and $a < X < c$, since the probability is the same for each (i.e., the probability at a point is zero).

Out of interest in the event $X \leq x$, we introduce $F\,(x)$, the *distribution function* of the continuous random variable X, and define it by

$$F\,(x) = p\,(X \leq x) = \int_{-\infty}^{x} f\,(t)\,dt. \tag{A.4.6}$$

It follows that

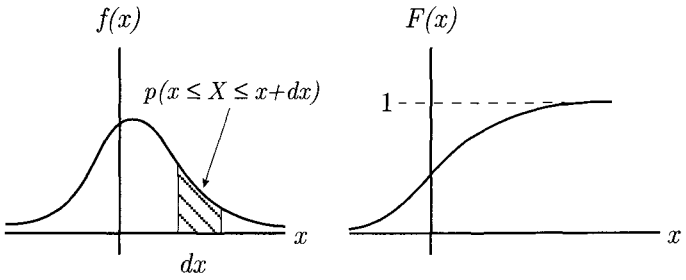
$$F(-\infty) = 0 \text{ and } F(\infty) = 1\,.$$

From elementary calculus, at points of continuity of F

$$\frac{d\,F(x)}{dx} = f\,(x) \tag{A.4.7}$$

which relates distribution and density functions for continuous random variables.

Consider the following sketch of the distribution and density function of a continuous random variable X.



Density function of a continuous random variable

Distribution function of a continuous random variable

From the definition of the density and distribution functions we have

$$p\,(a \leq X \leq b) = \int_a^b f(x)\,dx = F\,(b) - F\,(a)\,. \tag{A.4.8}$$

From axioms 1 and 2, we find

$$0 \leq F(x) \leq 1 \tag{A.4.9}$$

and $F(x)$ is monotonically increasing.

A.5 EXPECTED VALUES

We will now discuss the characteristics of probability distributions. A probability distribution is a generic term used to describe the probability behavior of a random variable. We shall now define certain useful parameters associated with probability distributions for continuous random variables. The *expected value* or the *mean* of X is written $E\,(X)$, and is defined by

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,dx\,. \tag{A.5.1}$$

Now consider a second random variable, where

$$Y = g\,(X)$$

then

$$E\,[g(X)] = \int_{-\infty}^{\infty} g(x)\,f(x)\,dx \tag{A.5.2}$$

which defines the expected value of a function of a random variable.

A.6 EXAMPLES AND DISCUSSION OF EXPECTATION

There are particular functions, $g(X)$, whose expectations describe important characteristics of the probability distribution of X. We will now consider several particular cases.

$E(X)$ is sometimes written as λ_1 or λ and is called the arithmetic mean of X. Geometrically, λ_1 is one of a number of possible devices for locating the "center" or centroid of the probability distribution with respect to the origin. The kth moment of X about the origin is

$$E[X^k] = \lambda_k = \int_{-\infty}^{\infty} x^k f(x)\, dx\,. \tag{A.6.1}$$

We also may speak of the kth moment of X about the mean λ_1. In this case, we define

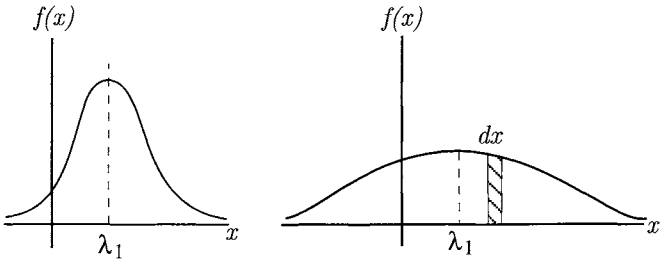
$$\mu_k \equiv E(X - \lambda_1)^k = \int_{-\infty}^{\infty} (x - \lambda_1)^k f(x) dx\,. \tag{A.6.2}$$

Note that $\mu_1 = 0$. For $k = 2$

$$\mu_2 = E(X - \lambda_1)^2 = \int_{-\infty}^{\infty} (x - \lambda_1)^2 f(x) dx\,. \tag{A.6.3}$$

This is usually denoted as σ^2 or $\sigma^2(X)$ and called the *variance* of X, the variance of the probability distribution of X, or the second moment of X about the mean. Note that μ_2 is always greater than zero unless $p(X = \lambda_1) = 1$. In this case $\mu_2 = 0$. The positive square root of the variance, σ, is called the standard deviation of X. It is one measure of the dispersion of the distribution about its mean value.

We interpret the density function as a mass distribution, the first moment about the origin becomes the center of gravity of the distribution, and the variance is the moment of inertia about an axis through the mean. For example, consider the following density functions:



Both density functions a and b have their mean value indicated by λ_1. Note that the variance of b obviously will be much larger than that of a, since $(x - \lambda_1)^2$ generally will be larger than that of a for an increment of area dx.

A few useful results follow readily from the definition of the expected value (see Eq. (A.5.1)) and the fact that it is a linear operator

$$E(a + bX) = a + bE(X) \tag{A.6.4}$$

where a and b are constants. Also,

$$\mu_1 = \int_{-\infty}^{\infty} (x - \lambda_1)\, f(x)dx = \lambda_1 - \lambda_1 = 0 \tag{A.6.5}$$

$$\begin{aligned}\mu_2 &= \int_{-\infty}^{\infty} (x - \lambda_1)^2 f(x)\, dx \\ &= \int_{-\infty}^{\infty} (x^2 - 2x\lambda_1 + \lambda_1^2) f(x)dx \\ &= \lambda_2 - 2\lambda_1^2 + \lambda_1^2 \\ &= \lambda_2 - \lambda_1^2 .\end{aligned} \tag{A.6.6}$$

The higher order moments are of theoretical importance in any distribution, but they do not have a simple geometric or physical interpretation as do λ_1 and μ_2.

All the information that can be known about the random variable X is contained in the probability density function. In guidance and estimation applications, we are most concerned with the first two moments, namely the mean and variance. Since the probability density function will change with time, the prediction of the future values for the state of the dynamic system can be obtained by propagating the joint density function (see Section A.9) forward in time and using

it to calculate the mean and variance. The equations for propagating the mean and variance of a random vector X are discussed in Section 4.8 of the text.

A.7 MOMENT GENERATING FUNCTIONS

Consider the particular case of

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx \tag{A.7.1}$$

for which

$$g(X) = e^{\theta x}$$

where θ is a dummy variable. Since

$$e^{\theta x} = 1 + \theta x + \frac{(\theta x)^2}{2!} + \cdots \frac{(\theta x)^n}{n!} + \cdots \tag{A.7.2}$$

substituting Eq. (A.7.2) into Eq. (A.7.1) results in

$$E(e^{\theta x}) = \lambda_0 + \theta\lambda_1 + \frac{\theta^2\lambda_2}{2!} + \ldots \frac{\theta^n\lambda_n}{n!} + \cdots \tag{A.7.3}$$

Thus $E(e^{\theta x})$ may be said to generate the moments $\lambda_0, \lambda_1 \ldots \lambda_n$ of the random variable X. It is called the *moment generating function* of X and is written $M_X(\theta)$. Note that

$$\frac{\partial^k M_X(\theta)}{\partial\theta^k}\Big|_{\theta=0} = \lambda_k. \tag{A.7.4}$$

Accepting the fact that the moment generating function for the function $h(X)$ is given by

$$M_{h(X)}(\theta) = \int_{-\infty}^{\infty} e^{\theta h(X)} f(x)dx, \tag{A.7.5}$$

let $h(X) = X - \lambda_1$, then

$$M_{(X-\lambda_1)}(\theta) = e^{-\theta\lambda_1} M_X(\theta) \tag{A.7.6}$$

which relates moments about the origin to moments about the mean,

$$\mu_k = \frac{\partial^k M_{(X-\lambda_1)}(\theta)}{\partial\theta^k}\Big|_{\theta=0}. \tag{A.7.7}$$

From Eqs. (A.7.3) and (A.7.6)

$$M_{(X-\lambda_1)}(\theta) = e^{-\theta\lambda_1}\left(\lambda_0 + \theta\lambda_1 + \frac{\theta^2\lambda_2}{2!} + \cdots \frac{\theta^n\lambda_n}{n!}\cdots\right)$$

and for example,

$$\begin{aligned}\mu_2 &= \left.\frac{\partial^2 M_{(X-\lambda_1)}(\theta)}{\partial\theta^2}\right|_{\theta=0} \\ &= \left.\lambda_1^2 e^{-\theta\lambda_1}(\lambda_0 + \theta\lambda_1) - \lambda_1 e^{-\theta\lambda_1}\lambda_1 - \lambda_1 e^{-\theta\lambda_1}\lambda_1 + e^{-\theta\lambda_1}\lambda_2\right|_{\theta=0} \\ &= \lambda_2 - \lambda_1^2, \quad \text{recall that} \quad \lambda_0 = 1\,.\end{aligned}$$

This is identical to the result in Eq. (A.6.6).

A.8 SOME IMPORTANT CONTINUOUS DISTRIBUTIONS

A.8.1 UNIFORM OR RECTANGULAR DISTRIBUTION

If X has equal probability over the range $a \le X \le b$, that is, every value of X in this range is equally likely to occur, we say that X is *uniformly distributed.* Its density function is

$$\begin{aligned}f(x) &= \frac{1}{b-a}, \quad a \le x \le b \\ &= 0 \qquad \text{elsewhere}\end{aligned} \tag{A.8.1}$$

$$F(x) = \int_{-\infty}^{a} f(x)dx + \int_{a}^{x} f(t)dt = \int_{a}^{x} \frac{dt}{b-a}$$

$$\begin{aligned}F(x) &= 0 \qquad x < a \\ &= \frac{x-a}{b-a} \qquad a \le x \le b \\ &= 1 \qquad x > b\,.\end{aligned}$$

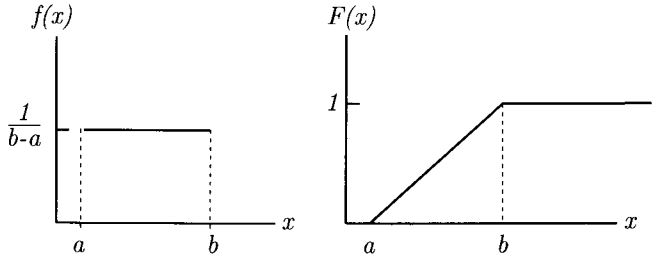
The first two moments (mean and variance) are

$$\begin{aligned}E(X) &= \int_{a}^{b} xf(x)dx = \int_{a}^{b} \frac{xdx}{b-a} \\ &= \left[\frac{x^2}{2(b-a)}\right]_a^b = \frac{a+b}{2}\end{aligned} \tag{A.8.2}$$

$$\sigma^2(X) = \int_a^b [x - E(X)]^2 f(x)dx$$

$$= \int_a^b \left(x - \frac{a+b}{2}\right)^2 \frac{dx}{b-a} = \frac{(b-a)^2}{12}. \tag{A.8.3}$$

Graphically the uniform density function is



The uniform density function is hardly a likely description of the probability behavior of many physical systems. Its importance lies in its utility in statistical theory. Any continuous probability distribution can be converted first into a uniform distribution, then into a given continuous distribution. This often facilitates the study of the properties of the distribution, which themselves are somewhat intractable.

A.8.2 The Gaussian or Normal Distribution

One of the most important distributions in probability theory is the one for which

$$f(x) = \frac{1}{\sqrt{2\pi}b} \exp\left[-\frac{1}{2}\left(\frac{x-a}{b}\right)^2\right] \quad \begin{matrix} -\infty < x < \infty \\ b > 0 \end{matrix}. \tag{A.8.4}$$

The moment generating function for this distribution is

$$M_X(\theta) = E[e^{\theta x}] = \int_{-\infty}^{\infty} e^{\theta x} f(x)\, dx$$

$$M_X(\theta) = \int_{-\infty}^{\infty} \frac{e^{\theta x}}{\sqrt{2\pi}\, b} \exp\left[-\frac{1}{2}\left(\frac{x-a}{b}\right)^2\right] dx$$

$$= \exp\left[\frac{\theta^2 b^2}{2} + a\theta\right]. \tag{A.8.5}$$

For details of the derivation of Eq. (A.8.5), see Freeman (1963).

From Eq. (A.8.5), we see that the mean of the normal distribution is

$$\frac{\partial M_X(\theta)}{\partial \theta}\Big|_{\theta=0} = a$$

hence

$$\lambda_1 = a\,.$$

From Eq. (A.7.6), we have

$$\begin{aligned} M_{X-\lambda_1}(\theta) &= e^{-a\theta} M_X(\theta) \\ &= e^{\theta^2 b^2/2} \end{aligned} \tag{A.8.6}$$

and from Eq. (A.8.6), the variance is

$$\frac{\partial^2 M_{X-\lambda_1}(\theta)}{\partial \theta^2}\Big|_{\theta=0} = b^2 = \sigma^2$$

and $\sigma = b$. Hence, Eq. (A.8.4) may be written

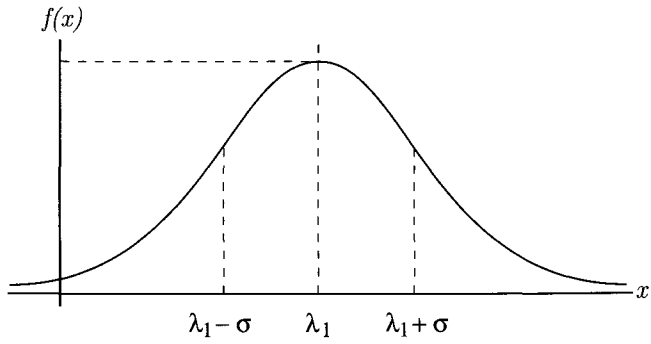
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\lambda_1}{\sigma}\right)^2\right] \quad -\infty < x < \infty\,. \tag{A.8.7}$$

Using Eq. (A.8.6), it may be shown that

$$\mu_{2k+1} = 0,$$

that is, the odd moments of a normal random variable about its mean are zero. In fact the odd moments of any symmetric distribution about its mean are zero provided that they exist.

The normal distribution is depicted graphically here:



The density function has its maximum at $X = \lambda_1$, is symmetric about the line $X = \lambda_1$, and has two inflection points at $X = \lambda_1 \pm \sigma$ (i.e., points where $\frac{d^2 f(x)}{dx^2} = 0$), and the minimum occur at $\pm\infty$. Other interesting properties of the univariate normal distribution function are

$$p(\lambda_1 - \sigma \leq X \leq \lambda_1 + \sigma) = \int_{\lambda_1 - \sigma}^{\lambda_1 + \sigma} f(x)dx = .68268$$
$$p\left[(\lambda_1 - 2\sigma) \leq X \leq (\lambda_1 + 2\sigma)\right] = .95449$$
$$p\left[(\lambda_1 - 3\sigma) \leq X \leq (\lambda_1 + 3\sigma)\right] = .99730.$$

A.9 TWO RANDOM VARIABLES

The joint distribution function $F(x, y)$ of two random variables is defined as

$$F(x_0, y_0) = p\{X \leq x_0, Y \leq y_0\}. \tag{A.9.1}$$

It has the following properties

$$\begin{aligned} &0 \leq F(x, y) \leq 1 \qquad \text{for all } x, y \\ &F(-\infty, y) = F(x, -\infty) = 0, F[\infty, \infty] = 1 \\ &F(\infty, y_0) = p\{Y \leq y_0\}, \qquad F(x_0, \infty) = p\{X \leq x_0\} \end{aligned} \tag{A.9.2}$$

The concept of the density function now follows. Because $F(x, y)$ is continuous, a function $f(x, y)$ exists such that

$$F(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u, v)\, du\, dv \tag{A.9.3}$$

or the equivalent

$$f(x, y) = \frac{\partial F^2(x, y)}{\partial x \partial y}. \tag{A.9.4}$$

By definition,

$$p(a \leq X \leq b, c \leq Y \leq d) = \int_{c}^{d} \int_{a}^{b} f(x, y)\, dx\, dy \tag{A.9.5}$$

which is analogous to the relationship between the probability and the density function for the single random variable case. Eq. (A.9.5) may be written

$$p(x \leq X \leq x + dx, y \leq Y \leq y + dy) = f(x, y)dxdy. \tag{A.9.6}$$

In summary:

$$\begin{aligned} F(x, y) &\equiv \text{joint distribution function of } X, Y. \\ f(x, y) &\equiv \text{joint density function of } X, Y. \\ f(x, y)\, dx\, dy &\equiv \text{joint probability element of } X, Y. \end{aligned}$$

A.10 MARGINAL DISTRIBUTIONS

We often want to determine the probability behavior of one random variable, given the joint probability behavior of two. This is interpreted to mean

$$p(X \leq x, \text{no condition on } Y) = F(x, \infty). \tag{A.10.1}$$

For the continuous case

$$\begin{aligned} F(x, \infty) &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(u, v)\, dv\, du = \int_{-\infty}^{x} \left[\int_{-\infty}^{\infty} f(u, v)\, dv \right] du \\ &= \int_{-\infty}^{x} g(u)\, du. \end{aligned} \tag{A.10.2}$$

Hence,

$$g(x) = \int_{-\infty}^{\infty} f(x, y)\, dy \tag{A.10.3}$$

is called the marginal density function of X. Similarly,

$$F(\infty, y) = \int_{-\infty}^{y} \left[\int_{-\infty}^{\infty} f(u, v)\, du \right] dv = \int_{-\infty}^{y} h(v)\, dv$$

and

$$h(y) = \int_{-\infty}^{\infty} f(x, y) dx \tag{A.10.4}$$

is the marginal density function of Y. Hence, the marginal density function of a random variable is obtained from the joint density function by integrating over the unwanted variable.

A.11 INDEPENDENCE OF RANDOM VARIABLES

We have previously defined the independence of two events A and B by $p(A, B) = p(A)\,p(B)$. For the case of random variables X and Y, we say that they are independent if we can factor their joint density function into

$$f(x, y) = g(x)h(y) \tag{A.11.1}$$

where $g(x)$ and $h(y)$ are the marginal density functions of X and Y.

A.12 CONDITIONAL PROBABILITY

For the simple events A and B, we define $p(B/A)$ by

$$p(B/A) = \frac{p(AB)}{p(A)}\,. \tag{A.12.1}$$

In the spirit of Eq. (A.12.1), we wish to define a *conditional density function* for continuous random variables X and Y with density functions $f(x, y)$, $g(x)$, and $h(y)$. Accordingly,

$$g(x/y) = \frac{f(x,y)}{h(y)},\ h(y/x) = \frac{f(x,y)}{g(x)}\,. \tag{A.12.2}$$

As an immediate consequence of Eq. (A.12.2), we have

$$\begin{aligned} p(a \le X \le b/Y = y) &= \int_a^b g(x/y)\,dx = \frac{\int_a^b f(x,y)\,dx}{h(y)} \\ p(c \le Y \le d/X = x) &= \int_c^d h(y/x)\,dy = \frac{\int_c^d f(x,y)\,dy}{g(x)}\,. \end{aligned} \tag{A.12.3}$$

Note that in Eq. (A.12.3), we are talking about X and Y in the vicinity of the values x and y.

Also,

$$\begin{aligned} p(a \le X \le b/c \le Y \le d) &= \frac{p(a \le X \le b, c \le Y \le d)}{p(c \le Y \le d)} \\ &= \frac{\int_a^b \int_c^d f(x, y)\,dy\,dx}{\int_c^d h(y)\,dy}\,. \end{aligned} \tag{A.12.4}$$

In the case of statistically independent variables, X and Y, the definition of statistical independence given by Eq. (A.11.1) leads to the following result from

Eq. (A.12.2)

$$g(x/y) = \frac{f(x, y)}{h(y)} = \frac{g(x)h(y)}{h(y)} = g(x)$$

$$h(y/x) = \frac{f(x, y)}{g(x)} = h(y). \tag{A.12.5}$$

A.13 EXPECTED VALUES OF BIVARIATE FUNCTIONS

Following the arguments for one random variable, we say that the mean or expected value $E[\phi(X, Y)]$ of an arbitrary function $\phi(X, Y)$ of two continuous random variables X and Y is given by

$$E[\phi(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(x, y) f(x, y) dx\, dy. \tag{A.13.1}$$

As with one random variable, the expected value of certain functions is of great importance in identifying characteristics of joint probability distributions. Such expectations include the following. Setting

$$\phi(X, Y) = X^l Y^m$$

gives

$$E[X^l Y^m] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^l y^m f(x, y) dx\, dy \equiv \lambda_{lm} \tag{A.13.2}$$

written λ_{lm}, the lm^{th} moment of X, Y about the origin. The lm^{th} moment about the mean is obtained by setting

$$\phi(X, Y) = [X - \lambda_{10}]^l [Y - \lambda_{01}]^m.$$

This results in

$$E\left\{[X - \lambda_{10}]^l [Y - \lambda_{01}]^m\right\}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [x - \lambda_{10}]^l [y - \lambda_{01}]^m f(x, y) dx\, dy \equiv \mu_{lm} \tag{A.13.3}$$

written μ_{lm}, the lm^{th} moment of X, Y about the respective moments λ_{10} and λ_{01}. Particular cases of μ_{lm} and λ_{lm} often used are

l	m	
0	0	$\lambda_{00} = 1$
1	0	$\lambda_{10} = E(X)$, the mean of X
0	1	$\lambda_{01} = E(Y)$, the mean of Y
0	0	$\mu_{00} = 1$
1	1	$\mu_{11} = E\{[X - E(X)](Y - E(Y))\}$, the covariance of X and Y
2	0	$\mu_{20} = \sigma^2(X)$, the variance of X
0	2	$\mu_{02} = \sigma^2(Y)$, the variance of Y.

Consider as an example the computation of

$$\begin{aligned}
\mu_{11} &= E(X - \lambda_{10})(Y - \lambda_{01}) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \lambda_{10})(y - \lambda_{01}) f(x, y) dx\, dy \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (xy - \lambda_{10} y - \lambda_{01} x + \lambda_{10}\lambda_{01}) f(x, y) dx\, dy \\
&= \lambda_{11} - 2\lambda_{10}\lambda_{01} + \lambda_{10}\lambda_{01} \\
&= \lambda_{11} - \lambda_{10}\lambda_{01}\,. \qquad \text{(A.13.4)}
\end{aligned}$$

The result is analogous to Eq. (A.6.6).

A.14 THE VARIANCE-COVARIANCE MATRIX

The symmetric matrix

$$\begin{aligned}
P &= E\left[\begin{bmatrix} X - E(X) \\ Y - E(Y) \end{bmatrix} \begin{bmatrix} X - E(X) & Y - E(Y) \end{bmatrix}\right] \\
&= E\begin{bmatrix} (X - E(X))^2 & (X - E(X))(Y - E(Y)) \\ (Y - E(Y)(X - E(X)) & (Y - E(Y))^2 \end{bmatrix} \\
P &= \begin{bmatrix} \sigma^2(X) & \mu_{11} \\ \mu_{11} & \sigma^2(Y) \end{bmatrix} \qquad \text{(A.14.1)}
\end{aligned}$$

is called the *variance-covariance matrix* of the random variables X and Y. As seen from Eq. (A.14.1), the diagonals contain the variances of X and Y, and the off diagonal terms contain the covariances.

The covariance of the random variables X and Y often is written in terms of the *correlation coefficient* between X and Y, ρ_{XY}. The correlation coefficient is defined as

$$\rho_{XY} \equiv \frac{E\{[X - E(X)]\,[Y - E(Y)]\}}{\{E[X - E(X)]^2\}^{1/2}\{E[Y - E(Y)]^2\}^{1/2}} = \frac{\mu_{11}}{\sigma(X)\sigma(Y)}\,. \tag{A.14.2}$$

The variance-covariance matrix for an n-dimensional random vector, X, can be written as

$$P = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & & \ddots & \vdots \\ \rho_{1n}\sigma_1\sigma_n & \rho_{2n}\sigma_2\sigma_n & \cdots & \sigma_n^2 \end{bmatrix} \tag{A.14.3}$$

where ρ_{ij} is a measure of the degree of linear correlation between X_i and X_j. It can also be written as

$$\rho_{ij} \equiv \frac{P_{ij}}{\sigma_i\sigma_j}, \quad i \neq j\,. \tag{A.14.4}$$

The correlation coefficient can be shown to be the covariance between the *standardized random variables*

$$U \equiv \tfrac{X - E(X)}{\sigma(X)}, V \equiv \tfrac{Y - E(Y)}{\sigma(Y)}\,. \tag{A.14.5}$$

That is,

$$\rho_{XY} = \mathrm{cov}\,(\mathrm{U},\,\mathrm{V}) = E[(U - E(U))(V - E(V))]\,.$$

This is easily demonstrated by noting that

$$E(U) = E(V) = 0\,. \tag{A.14.6}$$

Thus

$$E[(U - E(U))(V - E(V))] = E(UV)\,. \tag{A.14.7}$$

So

$$\mathrm{cov}\,(\mathrm{UV}) = E\,(UV) = E\left[\left(\frac{X - E(X)}{\sigma\,(X)}\right)\left(\frac{Y - E(Y)}{\sigma(Y)}\right)\right] = \frac{\mu_{11}}{\sigma(X)\,\sigma(Y)} \equiv \rho_{XY}\,. \tag{A.14.8}$$

Hence,

$$\mathrm{cov}\,(\mathrm{UV}) = \rho_{XY}\,. \tag{A.14.9}$$

Note also that

$$\sigma^2(U) = E[U - E(U)]^2 = E[U^2]$$
$$= \frac{E[X - E(X)]^2}{\sigma^2(X)} = \frac{\sigma^2(X)}{\sigma^2(X)} = 1. \tag{A.14.10}$$

Likewise

$$\sigma^2(V) = 1.$$

So

$$\sigma^2(U) = \sigma^2(V) = 1. \tag{A.14.11}$$

Note that the *standard deviation* is defined to be the positive square root of the variance. Consequently,

$$\rho_{UV} = \frac{\text{cov}(\text{UV})}{\sigma(U)\,\sigma(V)} = \text{cov}(\text{UV}) = \rho_{XY}. \tag{A.14.12}$$

A.15 PROPERTIES OF THE CORRELATION COEFFICIENT

It is first convenient to prove two elementary relationships for a function of two random variables. For a and b constant

$$E(a\,X + b\,Y) = aE(X) + bE(Y). \tag{A.15.1}$$

This follows from the linear property of the expectation operator. Next,

$$\sigma^2(aX + bY) \equiv E[(aX + bY) - E(aX + bY)]^2$$
$$= E[a(X - E(X)) + b(Y - E(Y))]^2$$
$$= a^2E[X - E(X)]^2 + 2ab\,E[(X - E(X))(Y - E(Y))] + b^2E[Y - E(Y)]^2$$
$$= a^2\sigma^2(X) + 2ab\,\mu_{11} + b^2\sigma^2(Y).$$

In terms of the correlation coefficient defined by Eq. (A.14.2), this result becomes

$$\sigma^2(aX + bY) = a^2\sigma^2(X) + 2ab\rho_{XY}\sigma(X)\sigma(Y) + b^2\sigma^2(Y). \tag{A.15.2}$$

Using Eqs. (A.15.1) and (A.15.2), we can demonstrate certain useful properties of ρ.

It will be convenient to use the standardized random variables defined by Eq. (A.14.5). From Eq. (A.15.2), we have that

$$\sigma^2(U + V) = \sigma^2(U) + 2\,\rho_{UV}\sigma(U)\sigma(V) + \sigma^2(V)$$
$$\sigma^2(U - V) = \sigma^2(U) - 2\,\rho_{UV}\sigma(U)\sigma(V) + \sigma^2(V)$$

but from Eq. (A.14.11)

$$\sigma^2(U) = \sigma^2(V) = 1$$

hence

$$\sigma^2(U \pm V) = 2(1 \pm \rho_{UV})\,. \tag{A.15.3}$$

Since by definition a variance is a nonnegative quantity

$$1 + \rho_{UV} \geq 0$$
$$1 - \rho_{UV} \geq 0\,.$$

Hence

$$-1 \leq \rho_{UV} \leq 1 \tag{A.15.4}$$

and from Eq. (A.14.12)

$$-1 \leq \rho_{XY} \leq 1\,. \tag{A.15.5}$$

It can be shown that when ρ_{XY} assumes its extreme values $+1$ or -1, the relationship between X and Y is perfectly linear. That is, all values of the random variable pair X, Y lie on a straight line of positive or negative slope.

From Eq. (A.15.3), if $\rho_{UV} = +1$

$$\sigma^2(U - V) = 0\,.$$

Hence, $(U - V)$ is a constant with all probability concentrated at that constant. Or in terms of X and Y

$$\frac{X - E(X)}{\sigma(X)} - \frac{Y - E(Y)}{\sigma(Y)} = \text{const}\,. \tag{A.15.6}$$

This is an equation of the form

$$Y = a + bX$$

where $b = \sigma(Y)/\sigma(X)$, a positive constant.

A similar expression may be written for $\rho_{UV} = -1$. In this case, b is a negative constant $(-\sigma(Y)/\sigma(X))$. Also, the converse holds. Suppose that

$$Y = a \pm bX\,.$$

Then it can easily be shown that $\mu_{11} = \pm b\sigma^2(X)$ and that $\sigma^2(Y) = b^2\,\sigma^2(X)$. Using the definition of ρ, we have $\rho_{XY} = \pm 1$.

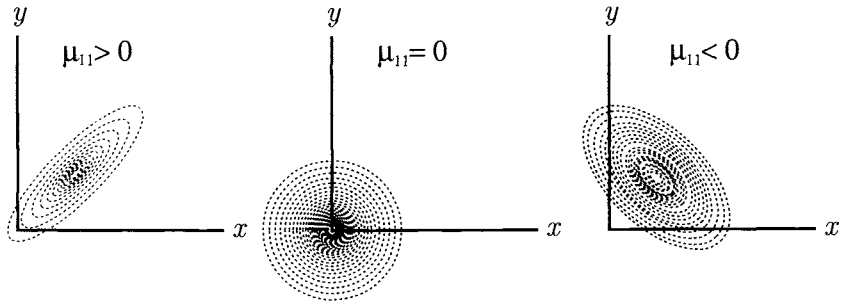
Nonlinear functional relationships between random variables do not necessarily result in the correlation coefficient assuming a value of 1.

A.16 PROPERTIES OF COVARIANCE AND CORRELATION

From the definition of μ_{11}

$$\mu_{11} = E[(X - E(X))(Y - E(Y))]$$

we see that if large values of the random variable X are found paired generally with large values of Y in the function $f(x, y)$, and if small values of X are paired with small values of Y, μ_{11} and hence ρ_{XY} will be positive. Also, if large values of X are paired with small values of Y in $f(x, y)$, then μ_{11} and hence ρ_{XY} will be negative. Finally, if some large and small values of X and Y are paired then $\mu_{11} \simeq 0$. Graphically assume that we sample a value of X and Y and plot the results. Three cases are possible:



An example of positive correlation would be a sampling of human height and weight. Note that if $\rho_{XY} = 1$, the variance-covariance matrix will be singular.

A.17 BIVARIATE NORMAL DISTRIBUTION

The *bivariate normal density function* is given by

$$f(x,y) = \frac{1}{2\pi\,\sigma(x)\,\sigma(y)\,\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left[\frac{x-\lambda_{10}}{\sigma(x)}\right]^2 - 2\rho\frac{[x-\lambda_{10}][y-\lambda_{01}]}{\sigma(x)\sigma(y)} + \left[\frac{y-\lambda_{01}}{\sigma(y)}\right]^2\right\}\right] \quad \begin{matrix} -\infty < x < \infty \\ -\infty < y < \infty \end{matrix} \tag{A.17.1}$$

which has five parameters λ_{10}, λ_{01}, $\sigma(x)$, $\sigma(y)$, and ρ. From Eq. (A.17.1) note that if $\rho = 0$, $f(x, y)$ may be factored into $f(x, y) = g(x)h(y)$. Hence, $\rho = 0$ is a sufficient condition for statistical independence of bivariate normal variables. This is not true for most density functions.

A.18 MARGINAL DISTRIBUTIONS

It can be shown that each of the random variables in Eq. (A.17.1) is normally distributed. By carrying out the integral,

$$g(x) = \int_{-\infty}^{\infty} f(x,y)dy. \tag{A.18.1}$$

It can be shown that

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma(x)} \exp\left[-\frac{1}{2}\left(\frac{x-\lambda_{10}}{\sigma(x)}\right)^2\right] \tag{A.18.2}$$

which is the normal density function of X. Similar results exist for the marginal distribution of Y. The converse is not true—if the marginal distributions $g(x)$ and $h(y)$ are normal, the joint density function $f(x,y)$ is not necessarily bivariate normal.

Now consider the conditional density function

$$h(y/x) = \frac{f(x,y)}{g(x)} \tag{A.18.3}$$

for the normal distribution. The numerator and denominator are given by Eqs. (A.17.1) and (A.18.2), respectively. Inserting these in Eq. (A.18.3) and simplifying, we obtain

$$h(y/x) = \frac{1}{\sigma(y)\sqrt{2\pi}\sqrt{1-\rho^2}} \times \exp -\frac{1}{2}\left[\frac{y - \{\lambda_{01} + [\rho\sigma(y)/\sigma(x)][x-\lambda_{10}]\}}{\sigma(y)\sqrt{1-\rho^2}}\right]^2. \tag{A.18.4}$$

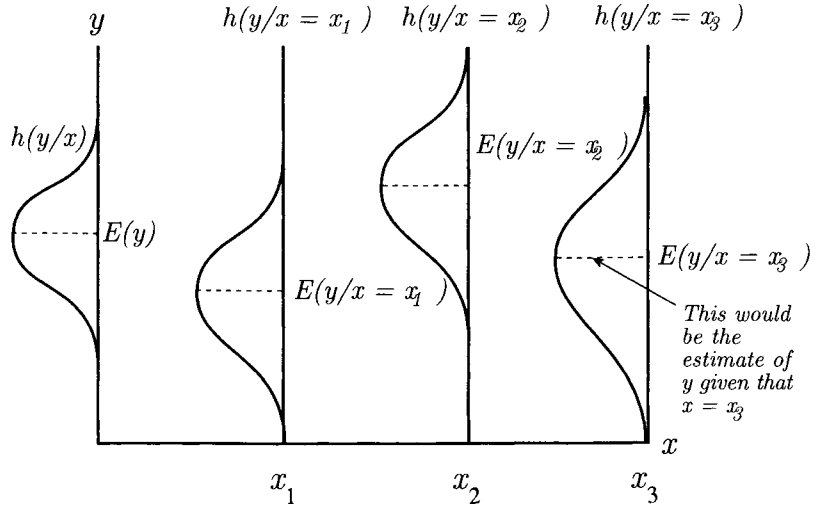
Hence, the conditional density function of Y is normal with conditional mean

$$E(Y/x) = \lambda_{01} + \rho\frac{\sigma(y)}{\sigma(x)}[x-\lambda_{10}]$$

and conditional standard deviation

$$\sigma(Y/x) = \sigma(y)\sqrt{1-\rho^2}.$$

Thus the conditional as well as the marginal distribution of the bivariate normal distribution are normal. A graphic example of the conditional density function follows.



A.19 THE MULTIVARIATE NORMAL DISTRIBUTION

For the multivariate case, consider a vector of random variables; for example,

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}.$$

The *multivariate normal density function* is given by

$$f(x_1, x_2, \ldots, x_p) = \frac{1}{(2\pi)^{p/2}|V|^{1/2}} e^{-1/2\,(\mathbf{X}-\mathbf{\Lambda})^T V^{-1}\,(\mathbf{X}-\mathbf{\Lambda})} \qquad -\infty < x_i < \infty \tag{A.19.1}$$

where

V is the $p \times p$ variance-covariance matrix of the vector $\mathbf{X}$

$|V|$ is the determinant of V

$\mathbf{\Lambda}$ is a $p \times 1$ vector of mean values of X.

The matrix V is defined as

$$V = E\{[\mathbf{X} - \mathbf{\Lambda}][\mathbf{X} - \mathbf{\Lambda}]^T\} \tag{A.19.2}$$

in terms of the correlation coefficient

$$\rho_{ij} = \frac{\mu_{ij}}{\sigma(x_i)\sigma(x_j)} \tag{A.19.3}$$

$$V = \begin{bmatrix} \sigma^2(x_1) & \rho_{12}\sigma(x_1)\sigma(x_2) & \rho_{13}\sigma(x_1)\sigma(x_3) \cdots \rho_{1p}\sigma(x_1)\sigma(x_p) \\ \rho_{12}\sigma(x_1)\sigma(x_2) & \sigma^2(x_2) & \rho_{23}\sigma(x_2)\sigma(x_3) \cdots \rho_{2p}\sigma(x_2)\sigma(x_p) \\ \vdots & & \\ \rho_{1p}\sigma(x_1)\sigma(x_p) & \rho_{2p}\sigma(x_2)\sigma(x_p) & \cdots \quad \cdots \quad \sigma^2(x_p) \end{bmatrix}.$$

Equation (A.19.1) is called the multivariate normal density function of a vector $\mathbf{X}$ with mean $\mathbf{\Lambda}$ and variance-covariance V.

The following theorems illustrate useful properties of multivariate normal density functions.

> Theorem 1: If the random variables $x_1, x_2 \ldots x_p$ are jointly normal, the joint marginal distribution of any subset of $s < p$ of the random variables is the s-variate normal. For a proof, see Chapter 3 of Graybill (1961).
>
> Theorem 2: If the $p \times 1$ vector $\mathbf{X}$ has the multivariate normal distribution with mean $\mathbf{\Lambda}$ and covariance V, then the components, x_i, are jointly independent if and only if the covariance of x_i and x_j for all $i \neq j$ is zero, that is, if and only if the covariance matrix V is diagonal.

To prove this, we must show that the joint density function factors into the product of the marginal density functions, which by Theorem 1 are also normal. Hence, we must show that

$$f(x_1, x_2 \ldots x_p) = f(x_1) f(x_2) \ldots f(x_p) \tag{A.19.4}$$

where

$$f(x_1, x_2 \ldots x_p) = \kappa\, e^{-1/2\,(\mathbf{X}-\mathbf{\Lambda})^T V^{-1}\,(\mathbf{X}-\mathbf{\Lambda})}$$

and

$$f(x_i) = \frac{1}{\sqrt{2\pi\, V_{ii}}} e^{-1/2\,\frac{(x_i-\lambda_i)^2}{V_{ii}}}$$

$$\kappa = \frac{1}{(2\pi)^{p/2}|V|^{1/2}}$$

where V_{ii} indicates the diagonal elements of V.

If $V_{ij} = 0$ for $i \neq j$, we have

$$(\mathbf{X} - \mathbf{\Lambda})^T V^{-1} (\mathbf{X} - \mathbf{\Lambda}) = \sum_{i=1}^{p} (x_i - \lambda_i)^2 V_{ii}^{-1}.$$

Also,

$$|V|^{1/2} = \prod_{i=1}^{p} (V_{ii})^{1/2}$$

since V is a diagonal matrix. Consequently, the joint density function can be factored as indicated by Eq. (A.19.4) and the elements of X are independent.

A.19.1 The Conditional Distribution for Multivariate Normal Variables

Theorem 3: If the $p \times 1$ vector $\mathbf{X}$ is normally distributed with mean $\mathbf{\Lambda}$ and covariance V and if the vector $\mathbf{X}$ is partitioned into two subvectors such that

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}, \mathbf{\Lambda} = \begin{bmatrix} \mathbf{\Lambda}_1 \\ \mathbf{\Lambda}_2 \end{bmatrix}, \text{and} \quad V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$$

then the conditional distribution of the $q \times 1$ vector $\mathbf{X}_1$, given that the vector $\mathbf{X}_2 = \mathbf{x}_2$, is the multivariate normal distribution with mean $\mathbf{\Lambda}_1 + V_{12} V_{22}^{-1} (\mathbf{x}_2 - \mathbf{\Lambda}_2)$ and the covariance matrix $(V_{11} - V_{12} V_{22}^{-1} V_{21})$ (Graybill, 1961); that is,

$$g(\mathbf{X}_1/\mathbf{X}_2 = \mathbf{x}_2) = \frac{1}{\kappa^*} e^{-1/2[\mathbf{X}_1 - \mathbf{\Lambda}_1 - V_{12} V_{22}^{-1}(\mathbf{x}_2 - \mathbf{\Lambda}_2)]^T [V_{11} - V_{12} V_{22}^{-1} V_{21}]^{-1} [\mathbf{X}_1 - \mathbf{\Lambda}_1 - V_{12} V_{22}^{-1} (\mathbf{x}_2 - \mathbf{\Lambda}_2)]} \tag{A.19.5}$$

where

$$\kappa^* = (2\pi)^{q/2} | V_{11} - V_{12} V_{22}^{-1} V_{21} |^{1/2}.$$

Theorem 4: The covariance matrix of the conditional distribution of $\mathbf{X}_1$, given $\mathbf{X}_2 = \mathbf{x}_2$ does not depend on $\mathbf{x}_2$.

Proof: The proof of this is obvious from an examination of Eq. (A.19.5); that is, $V_{\mathbf{X}_1/\mathbf{X}_2 = \mathbf{x}_2} = V_{11} - V_{12} V_{22}^{-1} V_{21}$. From Theorem 3, we also have

$$E(\mathbf{X}_1/\mathbf{X}_2 = \mathbf{x}_2) = \Lambda_1 + V_{12} V_{22}^{-1} (\mathbf{x}_2 - \mathbf{\Lambda}_2). \tag{A.19.6}$$

If we were attempting to estimate $\mathbf{X}_1$, its mean value given by Eq. (A.19.6) would be a likely value to choose. Also, because the covariance of the conditional density function is independent of $\mathbf{x}_2$, we could generate the covariance without actually knowing the values of $\mathbf{X}_2$. This would allow us to perform an accuracy assessment of $\mathbf{X}_1$ without knowing the values of $\mathbf{X}_2$.

A.20 THE CENTRAL LIMIT THEOREM

If we have n independent random variables x_i, $i = 1 \ldots n$ that are identically distributed with common means $E[x_i] = \lambda$ and (finite) variance $\sigma^2(x_i) = \sigma^2$, and we form the sum

$$W = x_1 + x_2 + \ldots x_n$$

whose mean and variance are given by

$$E[W] = n\lambda$$
$$E(W - E(W))^2 = \sigma^2(W) = n\sigma^2 .$$

The central limit theorem states that as $n \to \infty$ the standardized random variable of the sum

$$Z = \frac{W - E(W)}{\sigma(W)}$$

is normally distributed with mean 0 and variance 1 (Freeman, 1963). The important point is that W also is distributed normally. Hence, any random variable made up of the sum of enough independent random components from the same distribution will be distributed normally. Furthermore, if $n > 30$, Z (and W) will be distributed normally no matter what the shape of the distribution (Walpole *et al.*, 2002).

Another way of stating the theorem is that if sets of random samples are taken from any population, the means of these samples will tend to be distributed normally as the size of the samples becomes large (Davis, 1986).

The utility of the central limit theorem for orbit determination is that it gives some assurance that observation errors in tracking systems will tend to be distributed normally. This is because tracking system errors are usually the sum of a number of small random errors from a number of sources, including the hardware, the electronics, the mountings, and so on. This is a fundamental assumption in our development of statistical estimation algorithms.

A.21 BAYES THEOREM

One form of Bayes theorem is simply a statement of the conditional density functions given by Eq. (A.12.2),

$$g(x/y) = \frac{f(x, y)}{h(y)}$$

and

$$h(y/x) = \frac{f(x, y)}{g(x)} . \qquad \text{(A.21.1)}$$

Hence,

$$g(x/y) = \frac{h(y/x)g(x)}{h(y)}.$$

The last equations for $g(x/y)$ is the most elementary form of Bayes theorem. It is a useful starting point in the development of statistically based estimation criteria.

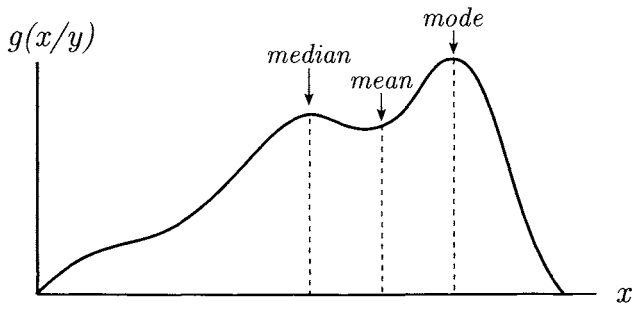
If we define X to be the state vector and Y to be the observation vector, then

$g(x/y) \equiv$ *a posteriori* density function

$h(y/x) \equiv$ *a priori* density function.

From a Bayesian viewpoint, we wish to develop a filter to propagate as a function of time the probability density function of the desired quantities conditioned on knowledge of the actual data coming from the measurement devices. Once such a conditional density function is propagated, the optimal estimate can be defined. Possible choices for the optimal estimate include:

1. The *mean* – the "center of the probability mass" distribution.
2. The *mode* – the value of x that has the highest probability, locating the peak of the density function.
3. The *median* – the value of x such that half the probability weight lies to the left and half to the right.



By generating the density function, some judgment can be made as to which criterion defines the most reasonable estimate for a given purpose.

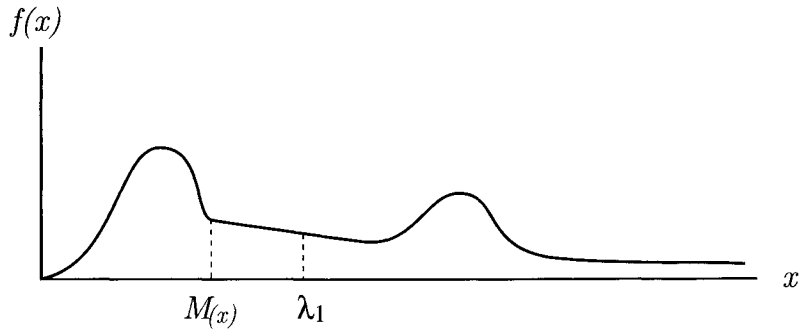
It is useful to examine the difference between the mean and median of a density function. Recall that the mean is defined as

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

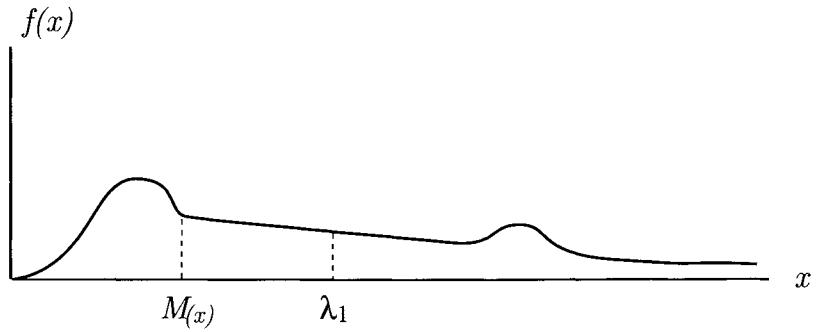
and the median, M, is defined as

$$M = \int_{-\infty}^{M} f(x)dx = 1/2\,. \tag{A.21.2}$$

As an example, consider a bimodal distribution:



If we move the second mode further out without changing the density function between the origin and the median, the median remains fixed but the mean moves to the right.



A.22 STOCHASTIC PROCESSES

Previously we have considered random variables that were not functions of time. Assume that we are given an experiment β specified by its outcomes η forming the space S and by the probabilities of these events. To every outcome η we now assign a time function $X(t, \eta)$. We have thus created a family of functions, one for each η. This family is called a *stochastic process*. A stochastic process is a function of two variables, t and η. The domain of η is S and the domain of t is a set of real numbers the time axis (Papoulis, 1991).

For a specific outcome η_i, the expression $X(t, \eta_i)$ signifies a single time function. For a specific time t_i, $X(t_i, \eta)$ is a quantity depending on η (i.e., a random variable). Finally, $X(t_i, \eta_i)$ is a mere number. In the future, $X(t, \eta)$ will be written $X(t)$.

As an example, consider a coin tossing experiment and define $X(t)$ so that

$$X(t) = \sin t\,, \text{ if } \eta = \text{heads}; X(t) = \cos t\,, \text{ if } \eta = \text{tails}.$$

Thus $X(t)$ consists of two regular curves; it is nevertheless a stochastic process.

From the preceding, we see that $X(t, \eta)$ represents four different things:

1. A family of time functions (t and η variable).
2. A single time function (t variable, η fixed).
3. A random variable (t fixed, η variable).
4. A single number (t fixed, η fixed).

We shall assume that $X(t)$ is a real process. For a specific t, $X(t)$ is a random variable. As in the case of random variables, we have a distribution function given by

$$F(x, t) = p(X(t) \leq x). \tag{A.22.1}$$

This is interpreted as: Given two real numbers x and t, the function $F(x, t)$ equals the probability of the event $\{X(t) \leq x\}$ consisting of all outcomes η such that at the specified time t, the functions $X(t)$ of our process do not exceed the given number x.

Associated with the distribution function is the corresponding density function

$$f(x, t) = \frac{\partial F(x, t)}{\partial x}. \tag{A.22.2}$$

These results hold at a given time and are known as the first order distribution and density function, respectively.

Now given two time instances t_1 and t_2, consider the random variable $X(t_1)$ and $X(t_2)$. Their joint distribution depends, in general, on t_1 and t_2 and will be denoted by $F(x_1, x_2, t_1, t_2)$

$$F(x_1, x_2, t_1, t_2) = p\{X(t_1) \leq x_1,\ X(t_2) \leq x_2\} \tag{A.22.3}$$

and will be called the second-order distribution function of the process $X(t)$. The corresponding density function is

$$f(x_1, x_2, t_1, t_2) = \frac{\partial^2 F}{\partial x_1\, \partial x_2}(x_1, x_2, t_1, t_2). \tag{A.22.4}$$

The marginal distribution and density functions are given by

$$F(x_1, \infty, t_1, t_2) = F(x_1, t_1),$$
$$f(x_1, t_1) = \int_{-\infty}^{\infty} f(x_1, x_2, t_1, t_2)\, dx_2 . \tag{A.22.5}$$

The conditional density is

$$f(X(t_1)/X(t_2) = (x_2) = \frac{f(x_1, x_2, t_1, t_2)}{f(x_2, t_2)} . \tag{A.22.6}$$

A.22.1 Definitions for Stochastic Processes

Given a stochastic process $X(t)$, its mean $\eta(t)$ is given by the expected value of the random variable $X(t)$,

$$\eta(t) = E\{X(t)\} = \int_{-\infty}^{\infty} x f(x, t) dx . \tag{A.22.7}$$

It is in general a function of time.

The *autocorrelation* $R(t_1, t_2)$ of a process $X(t)$ is the joint moment of the random variables $X(t_1)$ and $X(t_2)$

$$R(t_1, t_2) = E\{X(t_1)\, X(t_2)\}$$
$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1\, x_2\, f(x_1, x_2, t_1, t_2)\, dx_1\, dx_2 \tag{A.22.8}$$

and is a function of t_1 and t_2. The autocorrelation is analogous to the moment about the origin.

The *autocovariance* of $X(t)$ is the covariance of the random variables $X(t_1)$ and $X(t_2)$,

$$C(t_1, t_2) = E\{[X(t_1) - \eta(t_1)][X(t_2) - \eta(t_2)]\} . \tag{A.22.9}$$

If we have two stochastic processes $X(t)$ and $Y(t)$, these become the cross-correlation and cross-covariance, respectively. From Eq. (A.22.9), it is seen that

$$C(t_1, t_2) = R(t_1, t_2) - \eta(t_1)\eta(t_2) . \tag{A.22.10}$$

The variance of the random variable $X(t_1)$ is given by $(t_1 = t_2)$

$$C(t_1, t_1) = R(t_1, t_1) - \eta^2(t_1) . \tag{A.22.11}$$

Two random processes $X(t)$, $Y(t)$ are called uncorrelated if for any t_1 and t_2 we have

$$R_{XY}(t_1, t_2) = E[X(t_1)\, Y(t_2)] = \eta_x(t_1)\, \eta_y(t_2); \tag{A.22.12}$$

that is,

$$C_{XY}(t_1, t_2) = 0. \tag{A.22.13}$$

They are called *orthogonal* if

$$R_{XY}(t_1, t_2) = 0. \tag{A.22.14}$$

A stochastic process $X(t)$ is *stationary* in the strict sense if its statistics are not affected by a shift in the time origin. This means that the two processes

$$X(t) \quad \text{and} \quad X(t + \epsilon)$$

have the same statistics for any ϵ.

As a result

$$f(x, t) = f(x, t + \epsilon) \tag{A.22.15}$$

and since this is true for every ϵ, we must have the first-order density

$$f(x, t) = f(x) \tag{A.22.16}$$

independent of time, and

$$E[X(t)] = \eta \ , \quad \text{a constant.}$$

The density of order two must be such that

$$f(x_1, x_2, t_1, t_2) = f(x_1, x_2, t_1 + \epsilon, t_2 + \epsilon). \tag{A.22.17}$$

Because this must be true for any ϵ, it must be a function of only $t_1 - t_2$. This can be seen by noting that $t_1 + \epsilon - (t_2 + \epsilon) = t_1 - t_2$ is not dependent on ϵ, hence

$$f(x_1, x_2, t_1, t_2) = f(x_1, x_2, \tau) \tag{A.22.18}$$

where $\tau = (t_1 - t_2)$. Thus, $f(x_1, x_2, \tau)$ is the joint density function of the random variables

$$X(t + \tau) \quad \text{and} \quad X(\tau). \tag{A.22.19}$$

A.23 REFERENCES

Davis, J. C., *Statistics and Data Analysis in Geology*, John Wiley & Sons Inc., New York, 1986.

Freeman, H., *Introduction to Statistical Inference*, Addison-Wesley, 1963.

Graybill, F. A., *An Introduction to Linear Statistical Models*, McGraw-Hill, New York, 1961.

Helstrom, C. W., *Probability and Stochastic Processes for Engineers*, MacMillan, New York, 1984.

Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York, 1991.

Walpole, R. E., R. H. Myers, S. L. Myers, and Y. Keying, *Probability and Statistics for Engineers and Scientists*, Prentice Hall, Englewood Cliffs, NJ, 2002.