2

Fundamentals of Probability and Information Theory

The study of modern digital transmission practice is in many ways a study of the theory of probability and statistical inference. Messages to be sent through a system, as well as the actions of the channel upon these messages, are regarded as outcomes from some grand underlying experiment with a certain probability structure. Indeed the concept of information to a communication theorist is fundamentally linked to probability—loosely speaking, we regard information transfer as having occurred when our prior uncertainty associated with selection of a message is reduced.

We begin then by reviewing the basic concepts of probability and information theory. Our coverage is intended as a survey of those probability and information-theoretic ideas essential to the rest of the book and is by no means comprehensive. Readers wishing to delve further into the theory of probability will find solid treatments with an engineering orientation in the texts by Papoulis [1], Larson and Shubert [2], Gray and Davisson [3], and Leon-Garcia [4]. Texts that are more expansive on information theory are those of Gallager, McEliece, Blahut, Viterbi and Omura, and Cover and Thomas cited at the end of Chapter 1. Sections 2.1, 2.2, and 2.5 summarize standard material in probability and random processes and can be skipped for those with prior background in this material.

2.1 PROBABILITY

Probability theory is a body of mathematics, derived from an axiomatic basis, that seeks to describe apparently random, or uncertain, phenomena. We should realize that probability descriptions are models of the environment we seek to describe; these models are derived from prior experience or careful reasoning about the problem. Whether these models provide a sufficient description of real behavior is something that always should be questioned.

The elementary concepts of probability may be conveyed through die-rolling or balls-in-urns problems, but we shall develop a more formal understanding that suffices to handle all problems of typical engineering interest.

We first speak of an ***experiment*** that has results that seem, at least at our level of understanding or observation, to be random. Every experiment has a ***sample space***, Ω, which is the set of all possible outcomes from an experiment.[1] The sample space may be a finite set, countably infinite (i.e., one-to-one indexable by the positive integers), or noncountably infinite. Possible sample spaces could be the set {head, tail}; the set of real numbers; the set of all finite-energy waveforms $x(t)$, $-\infty < t < \infty$; or the set of binary sequences of length 23. Elementary outcomes or members of the sample space are denoted by ω, and we indicate this by $\omega \in \Omega$. ***Events***, A_i, are formed as subsets of Ω, which we denote $A_i \subseteq \Omega$. For example, the singleton set {head} and the set of rational numbers are subsets of the first two of the previous sample spaces. We note that the sample space is a subset of itself, an event called the ***sure event***, and the empty set, or ***null set***, $\emptyset$, is another subset of Ω, sometimes called the impossible event.

The ***union*** of events A and B, written $A \cup B$, is the set of outcomes ω such that $\omega \in A$ or $\omega \in B$, or both. Likewise, the ***intersection*** of events, written $A \cap B$, is the set of outcomes $\{\omega : \omega \in A \text{ and } \omega \in B\}$. We say A and B are ***disjoint*** if they have no points in common or, equivalently if $A \cap B = \emptyset$. Finally, we denote the ***complement*** of an event A, that is, the set of points in Ω, but not in A, by A^c.

For mathematical consistency, probability statements are made only about certain events that together form a field. A ***field*** or ***algebra*** F of events is a collection of events $A_1, A_2, \ldots, A_n$ such that, for every event A_i and A_j contained in the field F,

$$A_i \cup A_j \in F \tag{2.1.1}$$

and

$$A_i^c \in F. \tag{2.1.2}$$

Thus, a field is a collection of sets closed under the set operations of union and complementation. Every field F will therefore include, since $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$, the sure event Ω and the null set $\emptyset$.

In many cases of importance, we need to deal with infinite collections of events, for which we require that (2.1.1) extend to countable unions; that is, for a countable set

[1] Sets will be denoted either by enumeration, for example, {red, blue, yellow, green}, formally by listing the set property, for example, {x : x a positive integer}, or in longhand form, for example, the set of rational numbers.

of events $A_1, A_2, \ldots$ contained in F

$$\bigcup_{i=1}^{\infty} A_i \in F. \tag{2.1.1a}$$

If the family of events F satisfies this stronger requirement, in addition to (2.1.2), it is commonly called a ***sigma field*** or ***sigma algebra***.

A primary example of a sigma field is the situation where the sample space Ω is the real-line R^1, with the field F formed by the collection of open intervals (a, b), their complements, and countable unions of these. This field includes every event of possible interest, including sets with single elements, the set of rational numbers, and so on. (This field with an infinite number of constituent events is called a *Borel field.*)

We should note that the construction of the field F is partly determined by the ultimate interests of the probability analyst. For example, if the only question of interest in a real-line experiment ($\Omega = R^1$) is the probability of the event $A = \{\omega : \omega \geq 0\}$, then we could make this event be in our field F, along with the event consisting of the negative numbers, the entire real-line event Ω, and the empty set, $\emptyset$. This collection of four events is clearly a valid field and is much smaller than the Borel field (which also includes the event in question).

Having described the sample space and a collection of subsets defined on the sample space, we assign to every $A_i \in F$ a ***probability***, or set measure, $P(A_i)$, such that three basic axioms are satisfied:

$$P(A_i) \geq 0, \tag{2.1.3}$$

$$P(\Omega) = 1, \tag{2.1.4}$$

$$P(A_i \cup A_j) = P(A_i) + P(A_j) \quad \text{when } A_i \cap A_j = \emptyset. \tag{2.1.5}$$

Furthermore, in the countably infinite case, we require that

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i), \tag{2.1.5a}$$

provided the events are pairwise disjoint. These axioms simply require that probability measures be positive, normalized, additive (or countably additive) set measures.

Any probability measure satisfying these axioms is mathematically valid, and its choice then completes the ***probability system*** (Ω, F, P) that describes the experiment. Figure 2.1.1 summarizes the elements of such a system. Of course, for probability theory to be a useful and relevant theory, probability assignments must reflect physical laws, previously observed behavior, or good judgment about an experiment. There is also usually some flexibility in specifying a probability system for a given experiment. We naturally should seek the most tractable description.

Two examples will serve to illustrate probability systems. The first has a finite sample space and the second possesses an infinite sample space.

Example 2.1 Binary Sequences

Let Ω be the set of binary 4-tuples, that is, $\Omega = \{0000, 0001, \ldots, 1111\}$. The sample space contains 16 elementary events. The set of all subsets (there are $2^{16} = 65{,}536$ of them, including the null set $\emptyset$) is a field F closed under union and complementation, and the

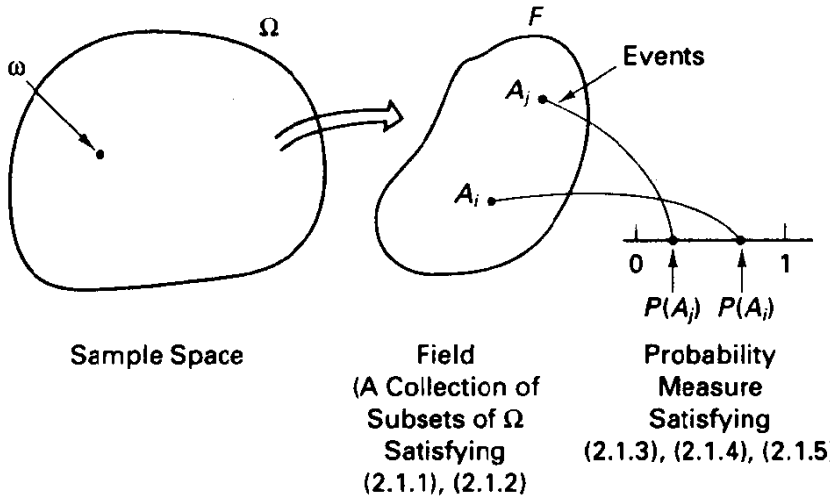


Figure 2.1.1 Components of a probability system (Ω, F, P).

probability assignment to events in F *could* be

$$P(A_i) = \frac{1}{16}(\text{number of elementary outcomes in } A_i). \tag{2.1.6}$$

Thus if A_i is the event described by "first three places are 0," or $\{0000, 0001\}$, $P(A_i) = \frac{1}{8}$. We don't need to make this probability assignment to have (2.1.3), (2.1.4), and (2.1.5) hold, but the assignment, given that all elementary outcomes are equally likely, represents a typical model.

Example 2.2 Receiver Noise

Let the experiment involve the specification of a noise voltage at one instant in a communication receiver. We take the sample space Ω to be the interval $[-5, 5]$. As the field of events, we might adopt the Borel field F consisting of all open intervals of Ω, their complements, and finite or countable unions of these. We might *choose* to assign probability to intervals to be proportional to the interval length. This constitutes a uniform probability assignment on the interval $[-5, 5]$, again not a necessary assumption. The constant of proportionality must ensure that $P(\Omega) = 1$. Therefore, to an interval (a, b) we assign probability $(b - a)/10$. Note in particular that the probability of the event $\{\omega = 0\}$, or any other singleton point, is zero in this system. This begins to illustrate the need for care in dealing with experiments with an infinity of outcomes. In fact, there are infinite sets in F, for example, the set of rational numbers in Ω, that have zero probability.

We might ask whether the "family of all subsets of Ω" is a useful field. This class of events meets the requirements of a field, by definition, and its adoption thereby would avoid the need for care in specifying the field F. However, in cases where the sample space is infinite, in particular the real-line situation, there are some events in this superfield that are unable to be "measured" or assigned probability consistent with the axioms (2.1.3) through (2.1.5a). This is a topic for a course in mathematical analysis.

Now we consider the union of events A and B, not necessarily disjoint events. Note that, by (2.1.1), $A \cup B \in F$; hence it is proper to assign to this event a probability,

which is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B), \tag{2.1.7}$$

which may be easily shown by application of elementary set operations and the axioms of probability (see Exercise 2.1.1).

Since probability is a nonnegative measure, we then have from (2.1.7) that

$$P(A \cup B) \leq P(A) + P(B), \tag{2.1.8}$$

with equality if A and B are disjoint or, more generally, if $A \cap B$ is an event of zero probability. This is a special case of a more general result, which will be used repeatedly in this book, the ***union bound:***

$$P\left(\bigcup_i A_i\right) \leq \sum_i P(A_i), \tag{2.1.9}$$

with equality if the events are disjoint. This bound follows by mathematical induction on (2.1.8).

2.1.1 Conditional Probability

The ***conditional probability*** of an event A, given the occurrence of the event B, is defined as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{2.1.10}$$

provided the denominator is nonzero. [If it is zero, the conditional probability in question is formally undefined, but it is perhaps better to say that $P(A \mid B)$ can be arbitrarily chosen as any number in $[0, 1]$. In this way an equivalent form of (2.1.10)

$$P(A \cap B) = P(A \mid B)P(B) \tag{2.1.11}$$

can be thought of as always holding, even when $P(B) = 0$.]

We should emphasize that conditional probabilities are measures on (conditioned) events and as such must satisfy the axioms stated earlier. For example, if conditioned on the event C, the events A and B are disjoint; then we insist that $P(A \cup B \mid C) = P(A \mid C) + P(B \mid C)$.

Since $P(A \cap B) = P(B \cap A) = P(B \mid A)P(A)$, we have the rule, sometimes known as ***Bayes's rule:***

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \tag{2.1.12}$$

This proves to be a useful computational tool in probability calculations.

From repeated application of (2.1.11) we can derive the ***chain rule for probabilities:***

$$\begin{aligned} P(A_1 \cap A_2 \ldots \cap A_n) = {} & P(A_1 \mid A_2 \ldots A_n)P(A_2 \mid A_3 \ldots A_n) \\ & \ldots P(A_{n-1} \mid A_n)P(A_n). \end{aligned} \tag{2.1.13}$$

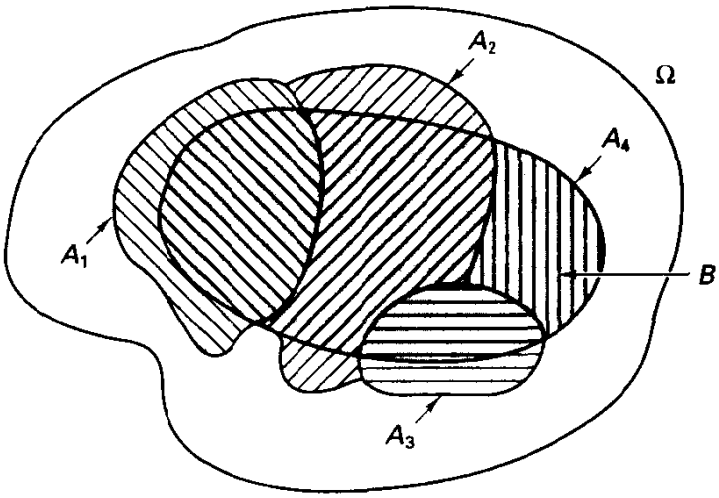


Figure 2.1.2 Venn diagram illustration of law of total probability. $P(B) = P(A_1, B) + P(A_2, B) + B(A_3, B) + P(A_4, B)$.

Also, if events $A_1, A_2, \ldots, A_m$ form a partition of B, that is, $B = \cup_{j=1}^{m} A_j$, but $A_i \cap A_j = \emptyset$ for all $i \neq j$ (see Figure 2.1.2), then the ***law of total probability*** expresses $P(B)$ as

$$P(B) = \sum_{j=1}^{m} P(B \cap A_j) = \sum_{j=1}^{m} P(B \mid A_j)P(A_j). \tag{2.1.14}$$

This rule allows us to find the probability of an event by analyzing disjoint constituent events and also allows Bayes's rule to be rewritten as

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{m} P(B \mid A_j)P(A_j)}. \tag{2.1.15}$$

Example 2.3 Noisy Channel with Binary Input, Binary Output

A simple communication situation will illustrate these relations. Let a binary channel have a single input and single output, both in the set $\{0, 1\}$. We proceed to define the sample space as the set of input/output pairs, $\Omega = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. We designate the events (subsets) corresponding to, respectively, sending 0 and 1 as

$$A_1 = \{(0, 0), (0, 1)\} \quad \text{and} \quad A_2 = \{(1, 0), (1, 1)\}.$$

Similarly, we designate the events of receiving 0 and 1 as B_1 and B_2. Suppose the conditional probabilities $P(B_1 \mid A_1)$ and $P(B_2 \mid A_2)$ are both defined to be 0.9, while the "error" probabilities are $P(B_2 \mid A_1) = P(B_1 \mid A_2) = 0.1$. Also, let us assume that the input probabilities are $P(A_1) = 0.4$ and $P(A_2) = 0.6$. This model is summarized in Figure 2.1.3.

By the law of total probability, the probability that the output is 0 is

$$P(B_1) = P(A_1)P(B_1 \mid A_1) + P(A_2)P(B_1 \mid A_2) = 0.42,$$

and application of (2.1.12) gives that the a posteriori[2] probability

$$P(\text{input} = 0 \mid \text{output} = 0) = P(A_1 \mid B_1) = \frac{(0.9)(0.4)}{0.42} = 0.857.$$

[2]"A posteriori" and "a priori" are Latin for "after the fact," and "before the fact," respectively.

Given, however, that the output is 0, the a posteriori probability of a 1 input is

$$P(A_2 \mid B_1) = \frac{(0.1)(0.6)}{0.42} = 0.143.$$

This serves to illustrate that conditioning can either raise or lower probabilities relative to the unconditioned, or a priori values. This example also introduces the ubiquitous ***binary symmetric channel***, which we shall frequently encounter throughout the book.

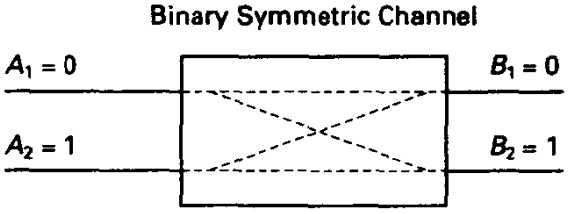


Figure 2.1.3 Probability model for Example 2.3.

2.1.2 Independence

Independence is one of the most fundamental concepts in all probability. Two events A and B are ***independent*** if and only if

$$P(A \cap B) = P(A)P(B), \tag{2.1.16}$$

which in conditional probability terms, by (2.1.10), means that

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = P(A). \tag{2.1.17}$$

In words, the conditioning upon B does nothing to alter the probability of the occurrence of A, if A and B are independent, which is certainly a reasonable interpretation of the independence of two events. Notice that the event A_1 ("input is 0") is not independent of the event B_1 ("output is 0") in Example 2.3; only for useless channels will it be the case that all input and output events are independent.

For multiple events $A_1, A_2, \ldots, A_n$, we say the events are ***jointly-independent*** if for *every* choice of subsets of these events, say $A_i, A_j, \ldots, A_m$

$$P(A_i \cap A_j \ldots \cap A_m) = P(A_i)P(A_j) \ldots P(A_m). \tag{2.1.18}$$

Joint independence naturally implies pairwise independence of the considered events, but the converse is not true in general. (Exercises 2.1.2 and 2.1.3 treat simple counterexamples.)

2.2 RANDOM VARIABLES: DISCRETE AND CONTINUOUS

In communication theory, we are normally interested in probabilistic experiments for which the outcomes are numerical, for example, the index number of a message or the voltage associated with a noise signal. Additionally, we will encounter the transformation

of random numerical data in the processing of signals. We now proceed to build on the previous theory to develop the calculus for handling such cases.

A scalar ***random variable*** X is formally defined as a mapping from a sample space Ω to the real line; that is, to every ω in Ω, we associate an image $X(\omega)$ on the real line. Thus, Ω is the domain of the function X, and the real line, or a subset of the real line, is the range of the function.[3] Figure 2.2.1 illustrates this abstraction. Provided this mapping is well behaved (mathematicians would say measurable), events in the sample space Ω define events on the real line, where the natural field of events is the Borel field cited earlier. Such an event is that consisting of all $\omega \in \Omega$ such that $X(\omega) \leq \frac{1}{2}$. In other words, a random variable defined on a probability space defines a new probability system. Usually, in applications we suppress the explicit functional relation and work directly with probability descriptions on the real line. Regarding notation, uppercase letters, for example, X, will denote random variables, while lowercase characters, for example, x, will indicate specific values of the random variable. (Incidentally, do not be misled by the term random variable; the functional relation is not random, but only the value of the function is.)

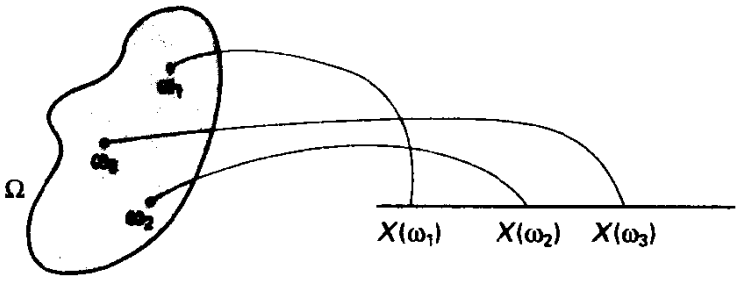


Figure 2.2.1 Abstract notion of a scalar random variable: a mapping from Ω to R.

If the random variable, which we shall often abbreviate as r.v., takes on a finite, or perhaps a countably infinite, number of values, x_i, we say X is a ***discrete*** r.v. Examples are the binary random variable taking on values 0 or 1 and the random variable corresponding to the number of transmissions (1, 2, ...) before the next error. On the other hand, when the values of X form an interval, or several disjoint intervals, we call X a ***continuous*** r.v. An example is the random phase angle Θ attached to a transmitted sinusoidal signal; normally, we would say this angle is a real number $0 \leq \Theta < 2\pi$. Occasionally, we encounter ***mixed*** random variables, where X exhibits attributes of both discrete and continuous r.v.'s.

2.2.1 Discrete Random Variables

A scalar discrete r.v. X is specified by its possible values $x_i, i = 1, 2, \ldots$, and by its ***probability mass function:***

$$P_X(x_i) = P(X = x_i), i = 1, 2, \ldots \tag{2.2.1a}$$

[3]More general range spaces are possible, but will not be employed here.

where by $P(X = x_i)$ we formally understand the probability of the event in the field F whose included outcomes ω satisfy $X(\omega) = x_i$, that is,

$$P_X(x_i) = P(\omega : X(\omega) = x_i). \tag{2.2.1b}$$

The axioms of probability imply that $P_X(x_i) \geq 0$ and $\sum_i P_X(x_i) = 1$.

Equivalently, we can describe a discrete r.v. by its ***cumulative distribution function (c.d.f.):***

$$F_X(x) = P_X(X \leq x) = \sum_{i:x_i \leq x} P_X(x_i) = \sum_i P_X(x_i)u(x - x_i), \tag{2.2.2}$$

where $u(x)$ is the unit step function. From the definition (2.2.2), we see that $F_X(x)$ is a nondecreasing function of x, with jump discontinuities at those x_i having nonzero probability. Also, from the definition, $F_X(-\infty) = 0$ and $F_X(\infty) = 1$.

We have earlier indicated we may just skip the functional understanding and work with distribution functions on the real line. This is all mathematically acceptable because we can show that, given any nonnegative, real-valued function $g(x)$ defined on the real numbers $x_1, x_2, \ldots$, which sums to 1, we can always set up a probability system (Ω, F, P) on which a random variable can be defined such that $P_X(x_i) = g(x_i)$.

Example 2.4 Binary R.V.'s and Related Distributions

Certainly, one of the most relevant examples for digital communications is the binary r.v. X taking on values $x_1 = 0$ and $x_2 = 1$ with probability p and $1 - p$, respectively, where $0 \leq p \leq 1$. [In the formalism of random variables, for a coin-flipping experiment with Ω = {head, tail} we might assign X(head) = 1 and X(tail) = 0 with P(head) = $1 - p$.] The probability mass function and distribution function for the r.v. X are shown in Figure 2.2.2a.

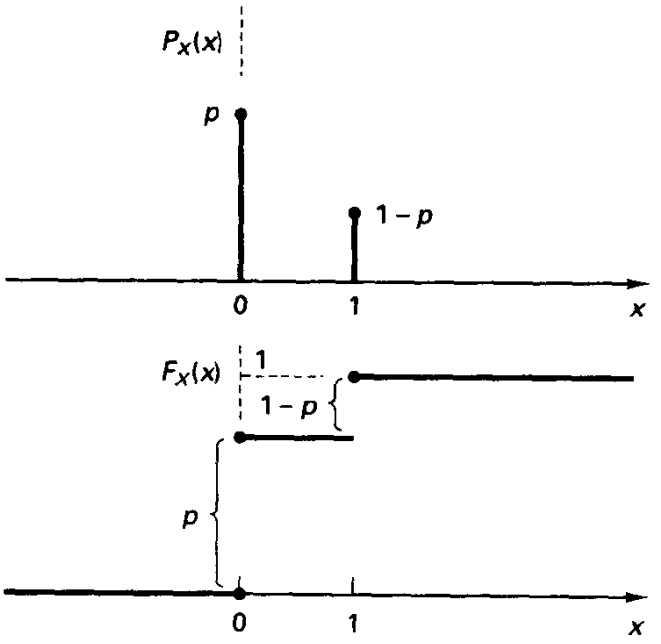


Figure 2.2.2a Probability mass function and distribution function for a binary random variable.

A related distribution is the ***binomial distribution***, giving the probability for the number of 0's (or tails in coin tossing) appearing in n independent trials[4] of a binary experiment. On the binary symmetric channel, the binomial distribution foretells the probability of having k errors in n uses of the channel if p is the error probability. Letting X be the random variable representing the number of 0's, we have

$$P_X(k) = C_k^n p^k (1-p)^{n-k}, \qquad k = 0, 1, 2, \ldots, n, \tag{2.2.3}$$

where $C_k^n = n!/k!(n-k)!$ is the binomial coefficient. Figure 2.2.2b illustrates the probability mass function for the binomial r.v. with $n = 10$ and $p = 0.25$.

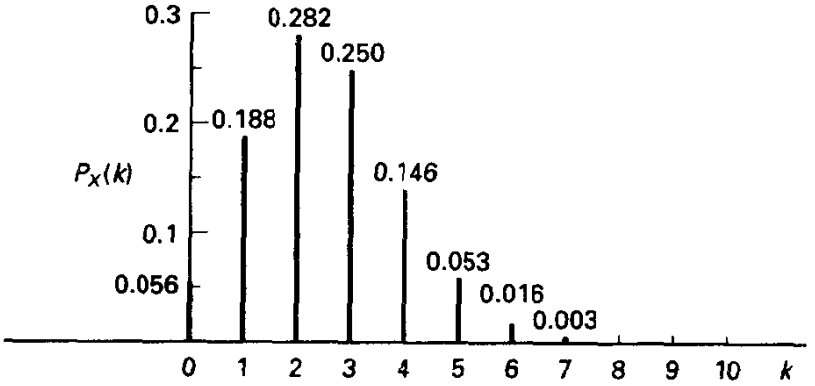


Figure 2.2.2b Binomial probability mass function when $n = 10$, $p = 0.25$.

As a check on the validity of the probability mass function (2.2.3), we note that summability to 1 follows from the binomial formula:

$$\sum_{j=0}^{n} C_k^n p^j (1-p)^{n-j} = [p + (1-p)]^n = 1. \tag{2.2.4}$$

Still another related random variable has the ***geometric distribution***, relating waiting-time probabilities in independent binary trials. In coin tossing, this relates the number of tosses k until the next head appears, while on the BSC, we have the distribution for the random time until the next error. For the next error to occur exactly k trials later, we must have $k-1$ correct transmissions followed by an error outcome. These events are independent, so we find the probability that the waiting time is k units is

$$P_X(k) = p(1-p)^{k-1}, \qquad k = 1, 2, \ldots, \tag{2.2.5}$$

where again $0 \le p \le 1$. Here summability to 1 follows from the expression for the sum of a geometric progression.

2.2.2 Continuous Random Variables

For the continuous random variable case the distribution function $F_X(x)$ defined as in (2.2.2) is, by definition, continuous and nondecreasing in x. If $F_X(x)$ is differentiable at all x, $f_X(x) = dF_X(x)/dx$ is the ***probability density function***, or p.d.f., for

[4]Often called Bernoulli trials.

the random variable X. We shall take derivatives in the right-hand-limit sense when it matters to define the value of $f_X(x_0)$, as in Example 2.5.

From the nondecreasing and end-point properties of distribution functions, it follows that a probability density function satisfies

$$f_X(x) \geq 0 \tag{2.2.6a}$$

and

$$\int_{-\infty}^{\infty} f_X(x)\, dx = F_X(\infty) - F_X(-\infty) = 1. \tag{2.2.6b}$$

Any function $f_X(x)$ defined on the real number line satisfying (2.2.6) constitutes the p.d.f. for a valid random variable.

It should be emphasized that probability density functions yield probability quantities only when integrated over intervals; that is,

$$P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)\, dx. \tag{2.2.7}$$

A differential calculus interpretation is that $f_X(x)\, dx$ is the probability that the random variable X lies in the interval $(x, x + dx)$.

Example 2.5 Uniform Random Variable

If X is equally likely to lie anywhere in the interval $[a, b]$, we say X is uniformly distributed on $[a, b]$, writing for shorthand $X \sim U[a, b]$. The cumulative distribution and probability density functions are, respectively,

$$F_X(x) = \begin{cases} 0, & x \leq b, \\ \dfrac{x-a}{b}, & a < x \leq b, \\ 1, & x > b, \end{cases} \tag{2.2.8}$$

and

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x < b \\ 0, & \text{otherwise.} \end{cases} \tag{2.2.9}$$

Both functions are shown in Figure 2.2.3.

Example 2.6 Gaussian Random Variable

The preeminent continuous random variable in probability and statistics literature, as well as in communication theory, is the ***Gaussian random variable***, named after K. F. Gauss, who called its distribution "normal."[5] Its central importance to probability theory is rooted in limit theorems, which, under mild restrictions, hold that sums of random variables converge in distribution to the Gaussian form as the number of summed variables becomes large. (This will be studied in Section 2.4.) This behavior makes the Gaussian distribution plausible as a model for electronic noise in communication systems, since most electrical noise processes are due to the aggregate effect of huge numbers of charge carriers undergoing

[5]Gauss argued that this distribution was the typical one emerging in measurements involving many error sources.

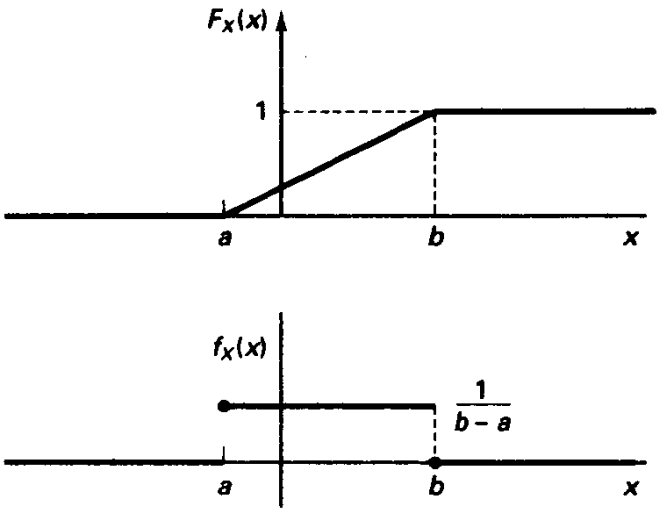


Figure 2.2.3 Cumulative distribution function and probability density function for a uniform random variable.

random motion. It is also true that the Gaussian assumption, when invoked in systems analysis, allows much easier mathematics, and this promotes a bit of overuse of the Gaussian model!

The Gaussian p.d.f. is given by

$$f_X(x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{(2\pi\sigma^2)^{1/2}}, \qquad -\infty < x < \infty \tag{2.2.10}$$

where the two parameters μ and σ specify the random variable. This Gaussian p.d.f. is depicted in Figure 2.2.4a. Notice that μ determines the location of the p.d.f., while σ controls the width. We use the notation $X \sim N(\mu, \sigma^2)$ to connote that X has p.d.f. given by (2.2.10).

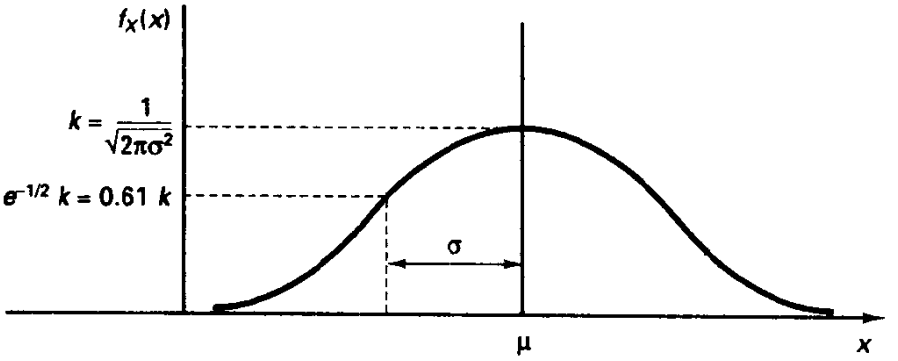


Figure 2.2.4a Gaussian probability density function.

If $f_X(x)$ is a Gaussian p.d.f. as in (2.2.10), with $\mu = 0$ and $\sigma^2 = 1$, then

$$F_X(x) = \int_{-\infty}^{x} f_X(z)\, dz = 1 - \int_{x}^{\infty} \frac{1}{(2\pi)^{1/2}} e^{-z^2/2}\, dz. \tag{2.2.11}$$

The latter integral is not expressible in terms of elementary functions, but is widely tabulated and available in many computer subroutine libraries. In this text, we adopt the

so-called Q-function notation to refer to such integrals and define

$$Q(x) = \int_x^\infty \frac{1}{(2\pi)^{1/2}} e^{-z^2/2}\, dz. \tag{2.2.12}$$

Thus, $Q(x)$ is simply the tail integral of the standard ($\mu = 0, \sigma = 1$) Gaussian density function (see Figure 2.2.4b). Note also that $Q(-x) = 1 - Q(x)$, so we only need to evaluate $Q(x)$ for $x \geq 0$. Also, it is obvious from the area interpretation of the integral that $Q(0) = \frac{1}{2}$ and $Q(\infty) = 0$.

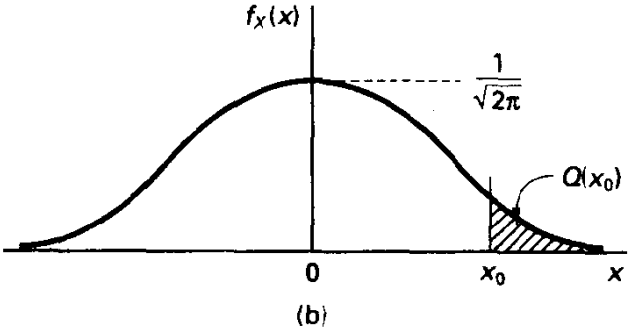


Figure 2.2.4b Standard Gaussian p.d.f. and definition of $Q(x)$.

When we are interested in probability quantities for a general Gaussian random variable, as in (2.2.10), a simple change of variables shows that

$$P(X \leq x) = F_X(x) = 1 - Q\left(\frac{x - \mu}{\sigma}\right). \tag{2.2.13}$$

In Table 2.1, a short table of the function $Q(x)$ is provided; more exhaustive tabulations are found in any standard mathematical handbook.

TABLE 2.1 GAUSSIAN TAIL INTEGRAL, $Q(x)$, AND UPPER BOUNDS

x	$Q(x)$	$\frac{e^{-x^2/2}}{(2\pi)_x^{1/2}}$	$\frac{e^{-x^2/2}}{2}$
0.0	0.5	—	5.00×10^{-1}
0.5	3.08×10^{-1}	6.21×10^{-1}	3.89×10^{-1}
1.0	1.59×10^{-1}	2.42×10^{-1}	3.03×10^{-1}
1.5	6.68×10^{-2}	8.63×10^{-2}	1.62×10^{-1}
2.0	2.28×10^{-2}	2.70×10^{-2}	6.77×10^{-2}
2.5	6.21×10^{-3}	7.01×10^{-3}	2.20×10^{-2}
3.0	1.35×10^{-3}	1.48×10^{-3}	5.55×10^{-3}
3.5	2.30×10^{-4}	2.50×10^{-4}	1.09×10^{-3}
4.0	3.17×10^{-5}	3.35×10^{-5}	1.70×10^{-4}
1.23	1×10^{-1}		
2.34	1×10^{-2}		
3.10	1×10^{-3}		
3.72	1×10^{-4}		
4.27	1×10^{-5}		
4.77	1×10^{-6}		

CAUTION

Other forms similar to $Q(x)$ abound in the literature, notably the complementary error function,

$$\operatorname{erfc}(x) = \frac{2}{\pi^{1/2}} \int_x^\infty e^{-z^2}\, dz. \tag{2.2.14a}$$

By a simple change of variables we have that

$$Q(x) = \frac{1}{2}\operatorname{erfc}\frac{x}{\sqrt{2}}. \tag{2.2.14b}$$

Special care should be used in reference to tables of the Gaussian integral, as well as in comparing expressions found here, such as probability of error, with other texts. Both definitions are widely used.

Because of the frequent need to evaluate the $Q(x)$ function or to manipulate $Q(x)$ to obtain analytical bounds on a problem, several approximations to $Q(x)$ are now presented, all of which become tight as the argument x increases.

We first apply integration by parts to the definition of $Q(x)$:

$$\begin{aligned} Q(x) &= \int_x^\infty \frac{1}{(2\pi)^{1/2}} e^{-z^2/2}\, dz = \int_x^\infty \frac{z e^{-z^2/2}}{z(2\pi)^{1/2}}\, dz \\ &= \frac{e^{-x^2/2}}{(2\pi)^{1/2} x} - \int_x^\infty \frac{e^{-z^2/2}}{(2\pi)^{1/2} z^2}\, dz, \qquad x > 0. \end{aligned} \tag{2.2.15a}$$

Since the second integral is positive, we obtain

$$Q(x) < \frac{e^{-x^2/2}}{(2\pi)^{1/2} x}, \qquad x > 0. \tag{2.2.15b}$$

The upper bound becomes tighter as the argument x increases.

Next, to establish a lower bound, we note that the last integral in (2.2.15a) we just observed to be positive is upper bounded by

$$\int_x^\infty \frac{e^{-z^2/2}}{(2\pi)^{1/2} z^2}\, dz < \int_x^\infty \frac{z e^{-z^2/2}}{z^3 (2\pi)^{1/2}}\, dz < \frac{1}{x^3} \int_x^\infty \frac{z e^{-z^2/2}}{(2\pi)^{1/2}}\, dz = \frac{e^{-x^2/2}}{x^3 (2\pi)^{1/2}}. \tag{2.2.16}$$

Using (2.2.16) in (2.2.15a) gives the lower bound:

$$Q(x) > \left(1 - \frac{1}{x^2}\right) \frac{e^{-x^2/2}}{(2\pi)^{1/2} x}, \qquad x > 0. \tag{2.2.17}$$

The upper and lower bounds of (2.2.15b) and (2.2.17) are in agreement to within 10% for argument x greater than 3, but neither are close approximations to $Q(x)$ for small x.

Sometimes more analytically convenient, and in fact a better upper bound than (2.2.15b) for small x, is [5, p. 123]

$$Q(x) \le \frac{1}{2} e^{-x^2/2}, \qquad x \ge 0. \tag{2.2.18}$$

The upper bounds, (2.2.15b) and (2.2.18), are also tabulated in Table 2.1, giving an indication of their accuracy. Another very accurate approximation amenable to machine evaluation is presented in Exercise 2.2.3.

Example 2.7 Rician–Rayleigh Random Variables

Another example of special importance in communication engineering is the random variable that arises in the envelope, or noncoherent, detection of noisy signals and as a model for certain fading processes. The ***Rayleigh***[6] random variable has a single parameter p.d.f. given by

$$f_X(x) = \begin{cases} \dfrac{x}{\sigma^2} e^{-x^2/2\sigma^2}, & x \geq 0, \\ 0, & x < 0, \end{cases} \tag{2.2.19a}$$

and the c.d.f. is obtained by integration to be

$$F_X(x) = \begin{cases} 1 - e^{-x^2/2\sigma^2}, & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{2.2.19b}$$

As we shall find shortly, the Rayleigh r.v. arises as the root-sum square of independent Gaussian variables; that is, $X = \left(X_1^2 + X_2^2\right)^{1/2}$, where X_1 and X_2 are independent Gaussian random variables having common parameters $\mu = 0$ and σ^2.

If we change the formulation slightly to allow one of these variables, say X_1, to have $\mu \neq 0$, but define the problem otherwise as before, then X is ***Rician distributed***,[7] with p.d.f. given by

$$f_X(x) = \frac{x}{\sigma^2} I_0\left(\frac{\mu x}{\sigma^2}\right) e^{-(x^2+\mu^2)/2\sigma^2}, \qquad x \geq 0, \tag{2.2.20}$$

where $I_0(z)$ is the modified Bessel function of zero order. Since $I_0(z) = 1$ at $z = 0$, it is obvious that (2.2.20) reduces to (2.2.19a) when $\mu = 0$, as it should from the definition of the Rician variable. Thus, the Rayleigh random variable is a special case ($\mu = 0$) of the Rician random variable. Figure 2.2.5a shows the Rayleigh p.d.f. and the Rician p.d.f. for a case where $\mu/\sigma = 3$. When μ/σ becomes large, it can be appreciated that the Rician p.d.f. approaches the Gaussian form, with parameters μ and σ. Figure 2.2.5b depicts the joint p.d.f. for independent Rayleigh and Rician variates.

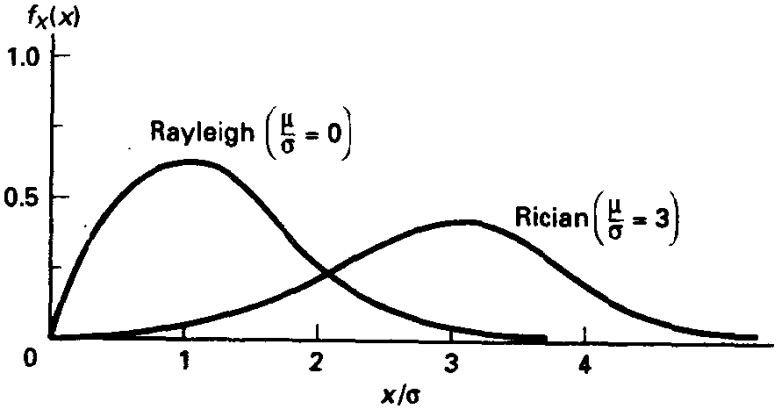


Figure 2.2.5a Rayleigh and Rician probability density functions.

[6]After Lord Rayleigh, who worked on electromagnetic scattering problems, among many others.

[7]Named for S. O. Rice, one of the foremost contributors to the mathematical description of noise processes.

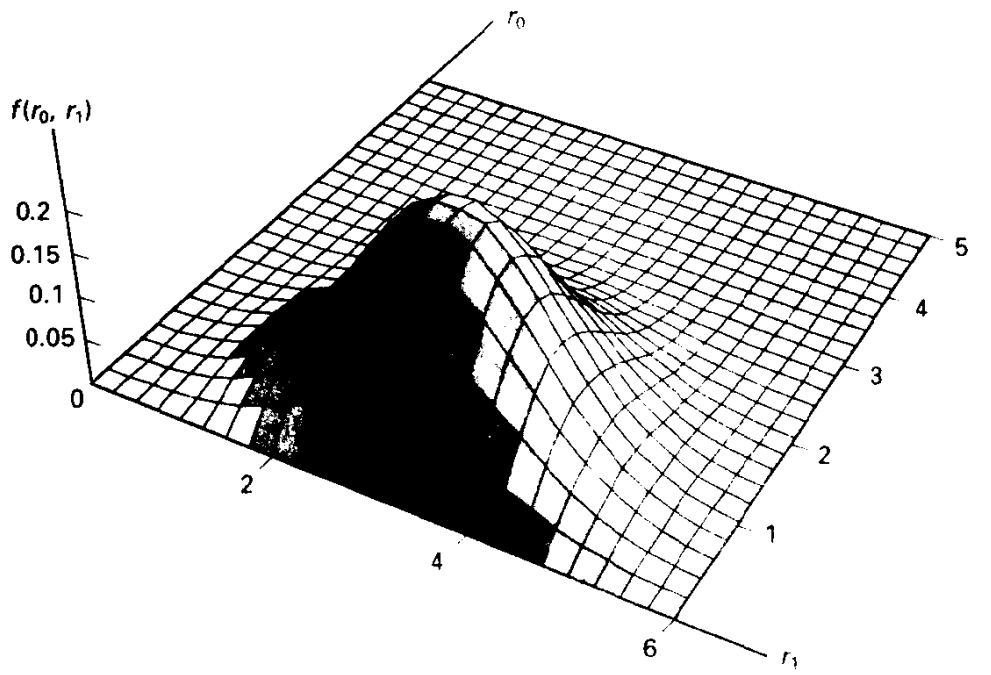


Figure 2.2.5b Joint p.d.f. for independent Rician and Rayleigh r.v.'s.

Discrete and continuous random variables have thus far been specified using separate notation. We can unify the description of discrete and continuous r.v.'s by invoking the Dirac impulse, or delta, function, $\delta(x)$, to handle derivatives of discontinuous distribution functions. Thereby, we can define a p.d.f. for discrete random variables totally comprised of impulse components.

The Dirac impulse is best defined through its *sifting integral* property:

$$\int_{-\infty}^{x} f(z)\delta(z - z_0)\,dz = \begin{cases} 0, & x < z_0, \\ f(z_0), & x \geq z_0. \end{cases} \tag{2.2.21}$$

implying that the function $\delta(x)$ has zero width and unit area. To apply this notation, consider a mixed continuous/discrete r.v. having the distribution function shown in Figure 2.2.6. Note there is nonzero probability that the variable X takes on values -1 and 1. The p.d.f. for this r.v. is also indicated, with impulse contributions having strength (or area) equaling the size of the discontinuity in the distribution function. With this extension, we may describe virtually any random variable by its probability density function, realizing it may contain impulse terms.

2.2.3 Multidimensional Random Variables or Random Vectors

Frequently, we need to describe situations where the outcomes are collections of real variables, for example, X_1, X_2, X_3. This may arise in either of two cases: (1) single performance of an experiment that produces multiple outputs or (2) repeated trials of

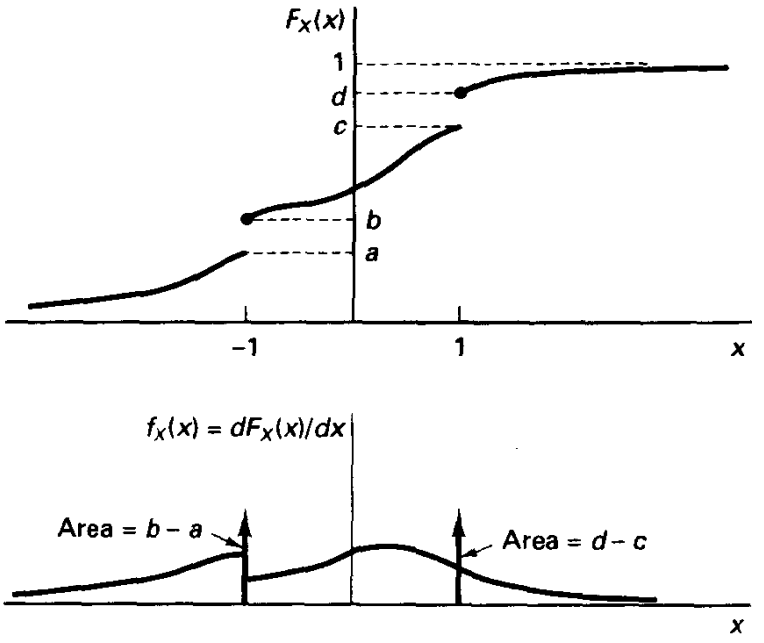


Figure 2.2.6 Distribution and density function for random variables of mixed type.

an experiment that produces a single output. We denote a real n-dimensional random vector by $\mathbf{X} = (X_1, X_2, \ldots, X_n)$[8] and by $F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$ its multidimensional ***joint distribution function.*** The value of $F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$ is the probability of the joint event "$X_1 \leq x_1$ and $X_2 \leq x_2 \ldots$ and $X_n \leq x_n$." This distribution function is everywhere nonnegative and is nondecreasing in each of its n arguments. These interpretations and properties are direct generalizations of the scalar r.v. case.

The n-dimensional ***joint probability density function*** is written $f_{X_1,X_2,\ldots,X_n}(x_1, \ldots, x_n)$, and is a nonnegative function of n dimensions that has n-dimensional volume of 1. The n-dimensional p.d.f. is related to the distribution function through partial derivatives:

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \frac{\partial^n [F_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n)]}{\partial x_1 \ldots \partial x_n}, \tag{2.2.22}$$

provided these partial derivatives exist. Figure 2.2.7a depicts a possible two-dimensional, or bivariate, p.d.f.

Especially in the multi-dimensional case, it is obvious that the notation becomes cumbersome, and we shall often omit subscripts denoting the random variable's name in density functions and distribution functions when the argument can be used to unambiguously convey the variable name. For example, we will frequently write $f(x)$ to mean $f_X(x)$ or $f(x, y)$ to mean $f_{XY}(x, y)$, but we cannot write $f(x - y)$ since its meaning is ambiguous: for what random variable is $x - y$ an outcome?

[8] Vectors will be denoted by boldface type to distinguish these from scalars.

Example 2.8 Three-dimensional Density Function

We consider the case $n = 3$ and define the random variables to have joint density function

$$f(x_1, x_2, x_3) = \frac{1}{(2\pi)^{3/2}} e^{-(x_1^2+x_2^2+x_3^2)/2}. \tag{2.2.23}$$

This p.d.f. has surfaces of constant probability density that are spheres in three-dimensional space, and the function can be visualized as a cloud having spherical symmetry about the origin, with the cloud density decreasing exponentially with the square of the distance from the origin. This density could be integrated over various regions of three-space to obtain probabilities that the three-dimensional random vector would lie in a certain volume. In particular, the integral over all space is 1 (as it must be). Other calculations are treated in Exercise 2.2.4.

If n-dimensional probability descriptions are given either in the form of distribution or density functions, we can obtain reduced or ***marginal*** descriptions by eliminating unneeded variables. For example, the distribution function for X_1 can be obtained from a bivariate distribution function for X_1, X_2 by

$$\begin{aligned} F_{X_1}(x_1) &= P(X_1 \le x_1) = P(X_1 \le x_1, X_2 < \infty) \\ &= F_{X_1, X_2}(x_1, \infty) \end{aligned} \tag{2.2.24}$$

The marginal p.d.f. for X_1 is obtained by

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1, X_2}(x_1, x_2)\, dx_2. \tag{2.2.25}$$

The term marginal is properly suggestive; we are either evaluating c.d.f.'s on the "margin" of the plot (2.2.24) or are projecting p.d.f's onto the margin by a process of integration (2.2.25). It may be simply verified by integration that, in Example 2.8, all first-order densities are of the Gaussian form (2.2.10) with $\mu = 0$ and $\sigma = 1$.

2.2.4 Conditional Distributions and Densities

Conditional distributions and densities are defined in keeping with the definition for conditional probability of events, with appropriate definitions of events in terms of random variables. Let $X \le x$ be an event and $Y \le y$ be another. Then the ***conditional distribution function*** for X, given $Y \le y$, is

$$\begin{aligned} F_{X|Y \le y}(x) &= \frac{P(X \le x, Y \le y)}{P(Y \le y)} = \frac{F_{XY}(x, y)}{F_Y(y)} \\ &= \frac{F_{XY}(x, y)}{F_{XY}(\infty, y)}. \end{aligned} \tag{2.2.26}$$

This should be interpreted as a function of x, for any fixed y. As for any cumulative distribution function, $F_{X|Y \le y}(x)$ must be nondecreasing in x, for any y, and must have limiting values of 0 and 1 as $x \to -\infty$ and $x \to \infty$, respectively.

For any fixed y, we could differentiate with respect to x, obtaining a certain form of conditional density function:

$$f_{X|Y \le y}(x) = \frac{\partial F_{X|Y \le y}(x)}{\partial x}. \tag{2.2.27}$$

On the other hand, suppose the conditioning on Y is that $y < Y \leq y + \delta y$. The distribution function for X, given that Y lies in this interval, is, from the preceding arguments

$$\begin{aligned} F_{X|y<Y\leq y+\delta y}(x) &= \frac{P(X \leq x, y < Y \leq y + \delta y)}{P(y < Y \leq y + \delta y)} \\ &= \frac{F_{XY}(x, y+\delta y) - F_{XY}(x, y)}{F_Y(y+\delta y) - F_Y(y)}. \end{aligned} \tag{2.2.28}$$

In terms of density functions, (2.2.28) could be expressed as

$$F_{X|y<Y\leq y+\delta y}(x) = \frac{\int_{-\infty}^{x}\int_{y}^{y+\delta y} f_{XY}(u, v)\, du\, dv}{\int_{y}^{y+\delta y} f_Y(v)\, dv}. \tag{2.2.29}$$

Differentiating (2.2.29) with respect to x using Leibniz's rule, we obtain the conditional density function for X, given that Y lies in a small interval $(y, y + \delta y)$:

$$f_{X|y<Y\leq y+\delta y}(x) = \frac{\int_{y}^{y+\delta y} f_{XY}(x, v)dv}{\int_{y}^{y+\delta y} f_Y(v)dv} \approx \frac{f_{XY}(x, y)\delta y}{f_Y(y)\delta y}, \tag{2.2.30}$$

assuming the density functions involved are approximately constant over the interval. In the limit as $\delta y \to 0$, we then have that the ***conditional density function*** for the random variable X, given $Y = y$, is

$$f_{X|Y=y}(x) = \frac{f_{XY}(x, y)}{f_Y(y)}, \tag{2.2.31}$$

provided that $f_Y(y) \neq 0$. Rather than write the cumbersome expression $f_{X|Y=y}(x)$, it is conventional to write instead

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x, y)}{f_Y(y)}, \tag{2.2.32}$$

but we should remember that this is a density function only for a fixed y. As y is varied, we obtain a family of conditional p.d.f.'s for the random variable X, and for any fixed y, $f_{X|Y}(x \mid y)$ must be nonnegative and integrate to 1. As earlier remarked, we shall sometimes omit the explicit naming of the random variables so that (2.2.32) could be denoted simply as $f(x \mid y)$.

A graphical interpretation of a conditional density $f_{X|Y}(x \mid y)$ is depicted in Figure 2.2.7a. Consider the curve formed by the intersection of the joint density surface and the plane $y = y_0$. This function of x is clearly nonnegative and gives the correct shape as a function of x for the conditional density $f_{X|Y}(x \mid y_0)$, but must be scaled by the marginal density at y_0, as shown in (2,2.32), to provide the integral constraint on the conditional density function. Figure 2.2.7b shows this conditional p.d.f., as well as another conditioned by $y = y_1$.

Manipulation of conditional density functions is often aided by using the density function version of (2.1.12), that is, if $f_Y(y) \neq 0$,

$$f_{X|Y}(x \mid y) = \frac{f_{Y|X}(y \mid x)f_X(x)}{f_Y(y)}, \tag{2.2.33}$$

although we should be careful to interpret this not as a ratio of probabilities, but of probability densities.

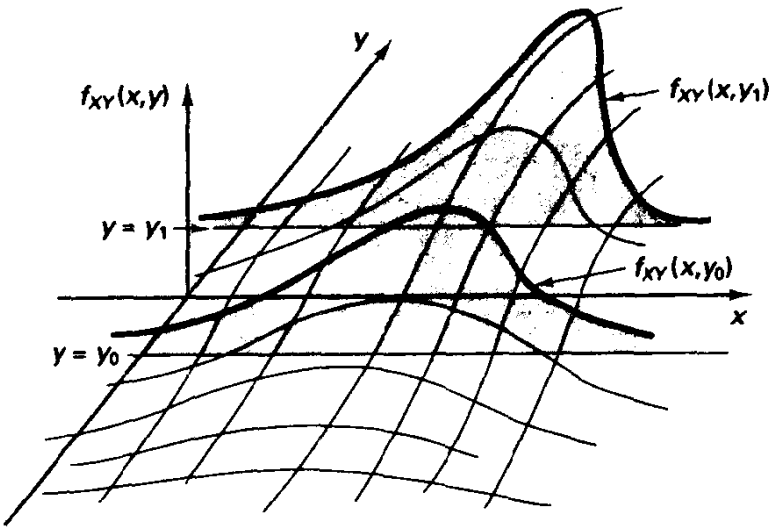


Figure 2.2.7a Bivariate density function.

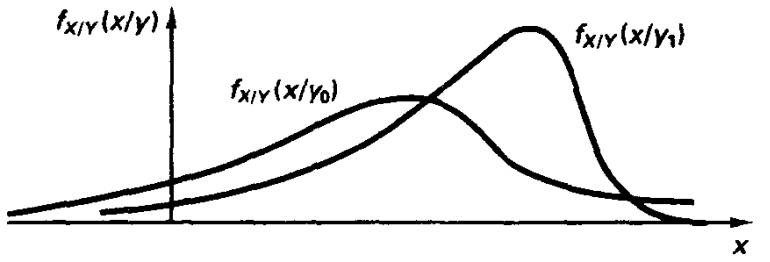


Figure 2.2.7b Conditional probability density functions for joint p.d.f. of Figure 2.2.7a.

2.2.5 Independence of Random Variables

We have already encountered the concept of independence of events. Realizing that the distribution function for n random variables is a probability of a joint event $X_1 \leq x_1, \ldots, X_n \leq x_n$, we can immediately define independence of n random variables as follows: $X_1, X_2, \ldots, X_n$ are ***independent*** if and only if

$$F(x_1, x_2, \ldots, x_n) = F(x_1)F(x_2)\cdots F(x_n) \tag{2.2.34}$$

for all choices of arguments $x_1, x_2, \ldots, x_n$. That is, the joint distribution function must factor into product form. Note that (2.2.34) implies that

$$\begin{aligned} F(x_1, x_2, \ldots, x_{n-1}) &= F(x_1, x_2, \ldots, x_{n-1}, x_n = \infty) \\ &= F(x_1)F(x_2)\cdots F(x_{n-1})F(x_n = \infty) \\ &= F(x_1)F(x_2)\cdots F(x_{n-1}), \end{aligned} \tag{2.2.35}$$

so, unlike the case in (2.1.17), we do not need to specify further that (2.2.34) holds for all subsets of arguments. Thus, independence of n random variables implies any pair (or other subset) of these are independent; the converse statement is, however, not true (see Exercise 2.2.3).

The corresponding statement for the joint density function of n independent variables is that $X_1, \ldots, X_n$ are independent if and only if

$$f(x_1, x_2, \ldots, x_n) = f(x_1) f(x_2) \cdots f(x_n) \tag{2.2.36}$$

for all choices of $x_1, x_2, \ldots, x_n$. Clearly, the three-dimensional density function in Example 2.8 has this property, so in that case X_1, X_2, and X_3 are independent.

In the case that two random variables X_1 and X_2 are independent, by (2.2.32) and (2.2.36) we have that

$$f_{X_1|X_2}(x_1 \mid x_2) = f_{X_1}(x_1), \tag{2.2.37}$$

again meaning that conditioning upon $X_2 = x_2$ does not influence the density function for X_1, and vice versa.

Example 2.9 Computation with Rayleigh and Rician Random Variables

To provide a result useful later in our study, as well as to provide facility with computation involving probability density functions, we compute the probability that, in Example 2.7, the value of the Rayleigh random variable exceeds the value of the Rician random variable, assuming these are *independent*. We let R_0 denote the Rayleigh variate with p.d.f. given in (2.2.19) and R_1 denote the Rician variate with p.d.f. given by (2.2.21). To find the desired probability, we must integrate the joint p.d.f. over the region defined by the event, that is, $r_0 \geq r_1$. Thus, using independence, the two p.d.f.'s, and integration, we obtain

$$\begin{aligned} P(R_0 > R_1) &= \int_0^\infty \int_{r_1}^\infty f(r_0, r_1)\, dr_0\, dr_1 \\ &= \int_0^\infty f(r_1) \left[\int_{r_1}^\infty f(r_0)\, dr_0 \right] dr_1 \\ &= \int_0^\infty f(r_1) \left[\int_{r_1}^\infty \frac{r_0}{\sigma^2} e^{-r_0^2/2\sigma^2}\, dr_0 \right] dr_1 \\ &= \int_0^\infty \frac{r_1}{\sigma^2} I_0\left(\frac{\mu r_1}{\sigma^2}\right) e^{-(r_1^2+\mu^2)/2\sigma^2} \left[e^{-r_1^2/2\sigma^2} \right] dr_1. \end{aligned} \tag{2.2.38}$$

By combining exponents, removing $e^{-\mu^2/4\sigma^2}$ from the integrand, and changing variables, we are left with

$$\begin{aligned} P(R_0 > R_1) &= \frac{1}{2} e^{-\mu^2/4\sigma^2} \int_0^\infty \frac{y}{\sigma^2} I_0\left(\frac{\mu y}{2^{1/2}\sigma^2}\right) e^{-\left(y^2+\frac{\mu^2}{2}\right)/2\sigma^2} dy \\ &= \frac{1}{2} e^{-\mu^2/4\sigma^2}, \end{aligned} \tag{2.2.39}$$

since the integrand is the Rician density function (2.2.20), albeit with a parameter $\mu' = \mu/\sqrt{2}$. It is remarkable that the rather formidable looking integral in (2.2.38) reduces to such a compact result.

2.2.6 Transformations of Random Variables

Frequently, in communication systems analysis we encounter the transformation of random variable(s) $X_1, \ldots, X_n$, producing new random variables $Y_1, \ldots, Y_m$. Dependent on the nature of the functional transformation and the nature of the input random variables, the new random variables may be discrete, continuous, or mixed random variables.

The most general case is provided by a vector-valued function of a vector

$$\mathbf{Y} = \mathbf{g}(\mathbf{X}), \tag{2.2.40}$$

which represents a mapping from R^n to R^m. We shall concern ourselves, however, with the case of a scalar output variable $Y = g(\mathbf{X})$. Some transformations of this form are

$$Y = \text{Qnt}(X), \tag{2.2.41a}$$

where $\text{Qnt}(X)$ designates a quantizing, or discretizing, operation on the input random variable. Quantizing is further described shortly. Other routine transformations are

$$Y = |X| \tag{2.2.41b}$$

and

$$Y = \sum_{j=1}^{n} X_j^2. \tag{2.2.41c}$$

A general procedure[9] for describing Y is to find its distribution function $F_Y(y)$ or density function $f_Y(y)$ by direct appeal to the definition of these functions. That is, we determine

$$F_Y(y) = P(Y \leq y) = P(\mathbf{X} : g(\mathbf{x}) \leq y). \tag{2.2.42}$$

The latter can be computed as a multidimensional integral involving the n-dimensional joint p.d.f. of the random vector $\mathbf{X}$. We shall illustrate such analyses with two examples. The first, a quantizer, produces a discrete output random variable, while the second example yields a continuous random variable from a continuous function defined on a continuous random variable.

Example 2.10 Uniform Scalar Quantizing

Quantizers appear commonly in digital communication systems as the interface between continuous, or analog, signals (observations) and digital processors. As we discussed in Chapter 1, quantizers are also a frequent first step in the digital transmission pathway, forming a discrete information source from an analog source.

A typical uniform scalar quantizer operates as follows. $N = 2^b$ output levels (voltages) are assigned as possible output approximations of the input signal X. These can then be uniquely identified by a b-bit message. The output levels are designated $y_i = -A + (i + \frac{1}{2})\Delta, i = 0, 1, \ldots, N-1$, where A is a constant and Δ is the step size of the quantizer. Thresholds, η_i, are placed on the input interval at evenly spaced values: $\eta_i = -A + i\Delta, i = 1, 2, \ldots, N-2$. The outer thresholds η_0 and η_N are normally taken to be $\pm\infty$, allowing the quantizer to in principle accept arbitrarily large signal magnitudes. The mapping $\text{Qnt}(X)$ is such that output y_i is produced if $\eta_i \leq x < \eta_{i+1}$. The input/output characteristic is shown in Figure 2.2.8a.

Consider the problem of quantizing a Gaussian random variable with $\mu = 0$ and $\sigma = 1$ and optimizing the step size of the quantizer for a given number of bits, b, to achieve a best approximation. The result depends on the criterion of optimality, but a common choice, and one that is readily solved, is to use the minimum-mean-square-error criterion. That is, we choose Δ so that the probability weighted squared error,

$$\overline{e^2} = \int_{-\infty}^{\infty} f_X(x)\left[(x - \text{Qnt}(x))^2\right] dx \tag{2.2.43}$$

[9] See [1], Chapter 6, for other methods of handling transformations.

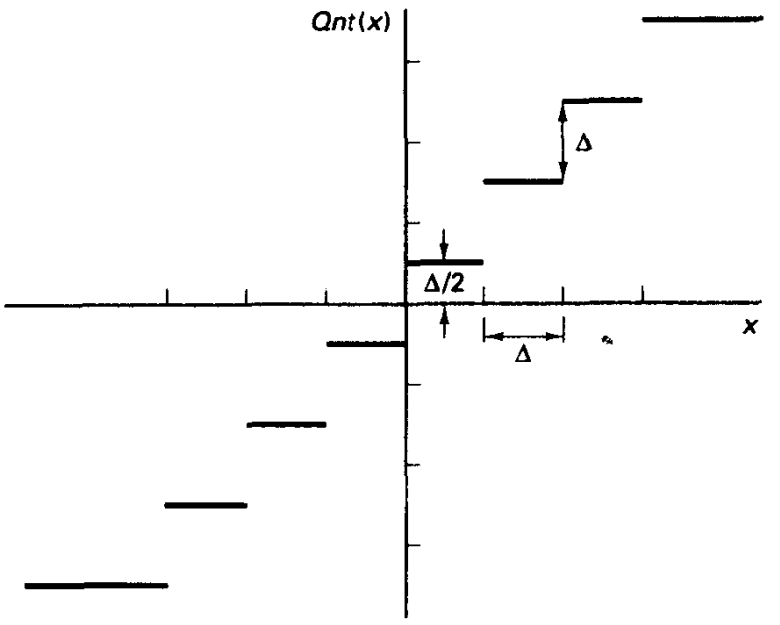


Figure 2.2.8a Uniform quantizer input/output characteristic.

is minimized. (This is the second moment of the quantizing error, as discussed in Section 2.3.) Max [6] performed this calculation for varying numbers of levels N (he also performed the optimization when the uniform interval requirement is removed).[10] For $N = 8$ (or $b = 3$), the optimal step size is $\Delta = 0.586$. This gives the thresholds shown in Figure 2.2.8b along the input axis. The probabilities of the eight output values y_i are given by integrals of the Gaussian p.d.f. over the appropriate interval:

$$P(Y = y_i) = Q(\eta_{i+1}) - Q(\eta_i), \qquad i = 0, 1, \ldots, 7, \tag{2.2.44}$$

where $Q(x)$ is again the Gaussian tail integral defined in (2.2.12).

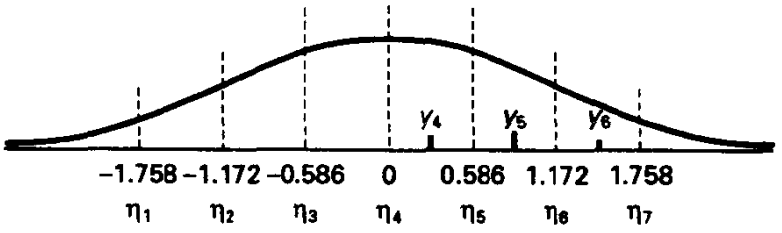


Figure 2.2.8b Quantizer spacings for $N = 8$ levels, Gaussian source, mean-square-error criterion.

Example 2.11 Development of Rayleigh Random Variable from Transformation

Let X_1 and X_2 be independent, Gaussian r.v.'s having $\mu = 0$ and identical parameters σ^2. Define

$$Y = \left(X_1^2 + X_2^2\right)^{1/2}, \qquad Y \geq 0, \tag{2.2.45}$$

[10] Uniform quantizers are easier to implement than nonuniform quantizers. The latter effect can be obtained with nonlinear mappings ahead of the quantizer and after the inverse quantizer.

to be a continuous random variable formed by these two variables. To find the distribution function for Y, we have

$$F_Y(y) = P(Y \le y) = P(\mathbf{x} : (x_1^2 + x_2^2)^{1/2} \le y). \tag{2.2.46}$$

The latter probability is obtained by integrating the joint probability density function for X_1 and X_2 over a disc centered at the origin, with radius y. Thus,

$$F_Y(y) = \iint_{(x_1+x_2)^{1/2} \le y} \frac{1}{2\pi\sigma^2} e^{-(x_1^2+x_2^2)/2\sigma^2}\, dx_1\, dx_2. \tag{2.2.47}$$

(The joint density function in the integrand is the product of marginal densities here.) By a rectangular-to-polar coordinate change, this becomes

$$\begin{aligned} F_Y(y) &= \int_0^{2\pi} \int_0^{y} \frac{\rho e^{-\rho^2/2\sigma^2}}{2\pi\sigma^2}\, d\rho\, d\theta \\ &= \int_0^{y} \frac{\rho e^{-\rho^2/2\sigma^2}}{\sigma^2}\, d\rho = 1 - e^{-y^2/2\sigma^2}, \qquad y \ge 0. \end{aligned} \tag{2.2.48}$$

The density function then follows from differentiation:

$$f_Y(y) = \frac{y}{\sigma^2} e^{-y^2/2\sigma^2}, \qquad y \ge 0, \tag{2.2.49}$$

which is the earlier defined p.d.f. for the Rayleigh r.v. (2.2.19a). Thus, the root-sum-square value of two independent Gaussian variables with $\mu = 0$ and equal σ is Rayleigh distributed. Equation (2.2.45) provides a recipe then for generating Rayleigh random variables for simulation purposes, given the ability to construct Gaussian random variables. Alternatively, this development leads to the Box–Muller method for producing two Gaussian r.v.'s from two uniform r.v.'s, as described in Exercise 2.2.7.

Proceeding further, we could ask for the distribution of $Z = Y^2 = X_1^2 + X_2^2$:

$$\begin{aligned} F_Z(z) &= P(Z \le z) = P(Y^2 \le z) = P(Y \le z^{1/2}) \\ &= \int_0^{z^{1/2}} \frac{y}{\sigma} e^{-y^2/2\sigma^2}\, dy = 1 - e^{-z/2\sigma^2}, \qquad z \ge 0. \end{aligned} \tag{2.2.50}$$

It follows from differentiation that

$$f_Z(z) = \frac{1}{2\sigma^2} e^{-z/2\sigma^2}, \qquad z \ge 0, \tag{2.2.51}$$

which is the p.d.f. for the ***one-sided exponential random variable.***

This result is a special case of a more general result, which we shall merely state (see Papoulis [1]). Let

$$Z = X_1^2 + X_2^2 + \cdots + X_n^2 \tag{2.2.52}$$

where X_i are independent, $\mu = 0$, Gaussian random variables. Then Z is said to be a ***chi-squared random variable with n degrees of freedom.*** Its p.d.f. is

$$f_Z(z) = \frac{1}{2^{n/2}\sigma^n \Gamma(n/2)} z^{(n-2)/2} e^{-z/2\sigma^2}, \qquad z \ge 0, \tag{2.2.53}$$

where $\Gamma(x)$ is the gamma function, evaluated by

$$\Gamma(x) = (x-1)!, \quad x \text{ an integer} \tag{2.2.54a}$$

or

$$\Gamma\left(x+\frac{1}{2}\right)=\frac{1\cdot 3\cdot 5\cdots(2x-1)}{2^n}\sqrt{\pi}, \quad x \text{ an integer.} \tag{2.2.54b}$$

As we have already stated, the Rician r.v. is obtained by

$$Y=\left[(\alpha+X_1)^2+X_2^2\right]^{1/2}, \tag{2.2.55}$$

where α is a constant and X_1, X_2 are Gaussian independent variables with $\mu = 0$ and common parameter σ. Equivalently, we could simply define one of the variables to have centering parameter α. The p.d.f. was expressed in (2.2.20) and can be obtained using similar methods, but eventual appeal to special functions, that is, $I_0(x)$, is necessary.

The sum of squares of n independent, identically distributed Rician r.v.'s, $Z = Y_1^2 + \cdots + Y_n^2$, has a ***noncentral chi-squared distribution with 2n degrees of freedom***, whose p.d.f. is given by [1]

$$f_Z(z)=\frac{1}{2\sigma^2}\left(\frac{z}{s^2}\right)^{(n-1)/2} e^{-(z+s^2)/2\sigma^2} I_{n-1}\left(\frac{z^{1/2}s}{\sigma^2}\right), \qquad z \geq 0, \tag{2.2.56}$$

where $s^2 = n\alpha^2/2$ is the *noncentrality parameter* of the distribution, and $I_n(x)$ is the modified Bessel function of order n. In particular, the p.d.f. for the square of a single Rician variable is noncentral chi-squared with two degrees of freedom:

$$f_Y(y)=\frac{1}{2\sigma^2}e^{-(y+s^2)/2\sigma^2} I_0\left(\frac{sy^{1/2}}{\sigma^2}\right) \tag{2.2.57}$$

2.3 EXPECTATIONS AND MOMENTS

Expectations, or expected values, are simply probabilistic averages of random variables (or functions of variables) in an experiment. We begin with the case of a single random variable, after which generalization to the multivariate case is simple.

Let $g(X)$ be a function of a random variable X with a specified probability density function or probability mass function. Thus, $Y = g(X)$ is another r.v., which may be either discrete or continuous depending on the nature of the random variable X and the function $g(X)$. The ***expected value*** of Y, written $E[Y]$, is defined as

$$E[Y]=E[g(X)]=\int_{-\infty}^{\infty} g(x) f_X(x)\,dx. \tag{2.3.1a}$$

In the case of discrete r.v.'s, we replace the integral with a summation:

$$E[Y]=\sum_i g(x_i) P_X(x_i). \tag{2.3.1b}$$

The relation in (2.3.1) is sometimes called the *fundamental theorem of expectation*, but we take it as a definition. In words, the expected value is simply the probability-weighted average of values of $g(X)$, and the notation $E[\ ]$ is merely shorthand for an integral operator.

2.3.1 First and Second Moments

Important special cases of this general definition are obtained when $g(X) = X$, the identity function, and when $g(X) = X^2$. In the former case

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

or

$$E[X] = \sum_i x_i P_X(x_i) \tag{2.3.2}$$

is called the ***first moment*** of X or, more commonly, the ***mean*** of X or the ***expected value*** of X. For shorthand, the expected value of X will be represented by $\overline{X}$ or occasionally by m when the random variable name is clear.

To recall an analogy in mechanics, if we let $f_X(x)$ be the mass per unit length of a thin rod, then $\overline{X}$ is the center of mass. Clearly, if the density function, or probability mass function, is symmetric about some value x_0, then $\overline{X} = x_0$, provided the integral or sum in (2.3.2) exists. This allows us to say by inspection that $(a + b)/2$ is the mean of the uniform r.v. in Example 2.5, while μ is the mean of the Gaussian random variable defined in Example 2.6.

When $g(X) = X^2$, we have

$$E[X^2] = \int x^2 f_X(x)\, dx, \tag{2.3.3}$$

which is called the ***second moment***, or ***mean-square value*** of X.[11] Often this is denoted by $\overline{X^2}$. (To pursue the mechanical analogy, $\overline{X^2}$ is the moment of inertia of the rod about the point $x = 0$.)

A related moment is the ***variance***, or second central moment:

$$\text{Var}[X] = E[(X - \overline{X})^2] = \int_{-\infty}^{\infty} (x - m)^2 f_X(x)\, dx. \tag{2.3.4}$$

(In the physical analogy, $\text{Var}[X]$ is just the moment of inertia about the center of mass). Direct expansion and integration in (2.3.4) gives

$$\begin{aligned} \text{Var}[X] &= \overline{X^2} - m^2 - m^2 + m^2 \\ &= \overline{X^2} - m^2. \end{aligned} \tag{2.3.5}$$

Thus, the variance of a random variable is equivalent to its second moment minus the square of the first moment.

The ***standard deviation*** of a random variable X is defined as the positive square root of the variance and is a common measure of scatter or dispersion. The standard deviation of the random variable in Example 2.5 is $(b-a)/12^{1/2}$, while for the Gaussian random variable of Example 2.6, the standard deviation is σ. These can be verified by carrying out the integration in (2.3.4).

---

[11] We will formulate expressions for the continuous r.v. case; the discrete case has an obvious analogous expression.

To extend the definition of expectation to the multidimensional case, we let $Y = g(\mathbf{X})$ be a scalar-valued function of an n-dimensional random vector $\mathbf{X}$. We define the expected value of Y to be

$$E[Y] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x}, \tag{2.3.6}$$

where we interpret the integral to require an n-dimensional integration.

We now develop one important operational rule for the expectation operator. Consider the sum of n random variables, $Y = X_1 + X_2 + \cdots + X_n$. The expected value of Y is

$$\begin{aligned} E[Y] = E[X_1 + \cdots + X_n] &= \int (x_1 + x_2 + \cdots + x_n) f(\mathbf{x})\, d\mathbf{x} \\ &= \int x_1 f(\mathbf{x})\, d\mathbf{x} + \cdots + \int x_n f(\mathbf{x})\, d\mathbf{x} \\ &= E[X_1] + E[X_2] + \cdots + E[X_n]. \end{aligned} \tag{2.3.7}$$

The last step follows from the definition of expectation (2.3.6), or we may first find the appropriate marginal densities for each variable in turn and then use (2.3.2). Therefore, no matter what the nature of the joint density of the n r.v.'s, we have that *the expectation of the sum is the sum of the expected values.*

2.3.2 Correlation and Covariance

An expectation of great importance in the study of random processes and certainly in communications and signal processing is the ***correlation*** between two random variables X_1 and X_2, defined as the expectation of their product:

$$\text{Corr}(X_1, X_2) = E[X_1 X_2] = \int\int x_1 x_2 f(x_1, x_2)\, dx_1\, dx_2. \tag{2.3.8}$$

Even more useful is the ***covariance*** between X_1 and X_2:

$$\begin{aligned} \text{Cov}(X_1, X_2) &= E[(X_1 - m_1)(X_2 - m_2)] \\ &= E[X_1 X_2] - m_1 m_2 - m_1 m_2 + m_1 m_2 \\ &= \text{Corr}(X_1, X_2) - m_1 m_2, \end{aligned} \tag{2.3.9}$$

where we have defined m_i as the mean of random variable X_i. Note that $\text{Cov}(X, X) = \text{Var}(X)$ and that $\text{Cov}(X_1, X_2) = \text{Cov}(X_2, X_1)$.

It follows from (2.3.8) that if X_1, X_2 are independent, then their correlation is the product of their means, $m_1 m_2$, giving a covariance of zero in (2.3.9). When the covariance is zero, the variables are said to be ***uncorrelated.*** (A more apt and less confusing term would be *uncovarianced.*)

Now let's return to the sum of n independent variables, $Y = \sum_{i=1}^{n} X_i$ and consider the variance of Y. For simplicity, let us assume $E[X_i] = 0$ for all i, and thus $E[Y] = 0$.

The variance of Y is then

$$\begin{aligned}\text{Var}[Y] = E[Y^2] &= E\left[\sum_{i=1}^{n} X_i \sum_{j=1}^{n} X_j\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} E[X_i X_j] = \sum_{i=1}^{n} E\left[X_i^2\right] = \sum_{i=1}^{n} \text{Var}[X_i].\end{aligned} \tag{2.3.10}$$

We have invoked the result just developed that $E[X_iX_j] = E[X_i]E[X_j], i \neq j$, for independent variables, as well as the zero-mean assumption. Therefore, the variance of the sum of *independent* random variables is obtained by summing the variances of each variable in the sum. This result holds even when the summed variables have nonzero mean.

Independent random variables are uncorrelated, but the converse is not in general true (see Exercise 2.3.3 for a counterexample). An important case where the converse does hold is the case of jointly Gaussian r.v.'s, as we now discuss.

Example 2.12 Jointly Gaussian Random Variables

Random variables $X_1, X_2, \ldots, X_n$ are ***jointly Gaussian*** (or jointly normal) if the n-dimensional density function is of the form

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi\,|\,\mathbf{K}\,|)^{n/2}} \exp\left[\frac{-(\mathbf{x}-\mathbf{m})\mathbf{K}^{-1}(\mathbf{x}-\mathbf{m})^T}{2}\right], \tag{2.3.11}$$

where $\mathbf{x}$ is the row vector representation of the n-tuple, $\mathbf{m} = (m_1, m_2, \ldots, m_n)$ is the vector of means, $\mathbf{x}^T$ denotes the vector transpose, and $\mathbf{K}$ is the $n \times n$ covariance matrix, defined as

$$\mathbf{K} = [K_{ij}] = [E[(X_i - m_i)(X_j - m_j)]]. \tag{2.3.12}$$

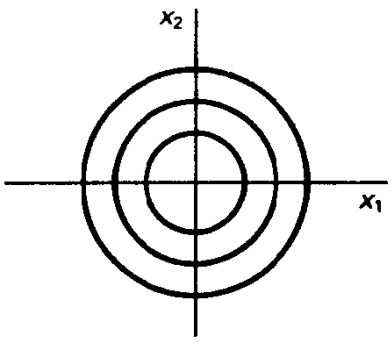
The exponent in (2.3.11) is a quadratic form, which implies that surfaces of constant probability density are n-dimensional ellipsoids. In the particular case of two variables, the joint p.d.f. can be expressed in terms of five parameters: two means, two variances, and a ***correlation coefficient***

$$\begin{aligned}\rho &= \frac{E[(X_1 - m_1)(X_2 - m_2)]}{\sigma_1\sigma_2} \\ &= \frac{\text{Cov}(X_1, X_2)}{\sigma_1\sigma_2}.\end{aligned} \tag{2.3.13}$$

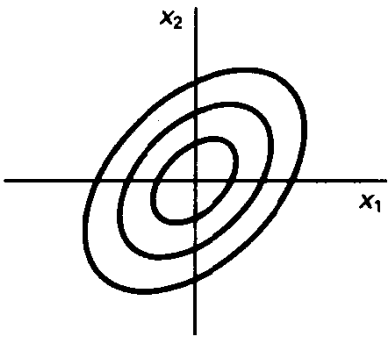
The correlation coefficient is defined in like manner for non-Gaussian variables and may be shown to lie in the interval $[-1, 1]$. Figure 2.3.1 depicts level contours, or contours of constant p.d.f. in the bivariate Gaussian case for several values of ρ.

Returning to the n-dimensional case, we can see that if the random variables are pair-wise uncorrelated, then $\mathbf{K}$ is a diagonal matrix, with entries $\sigma_1^2, \ldots, \sigma_n^2$, and the density function becomes

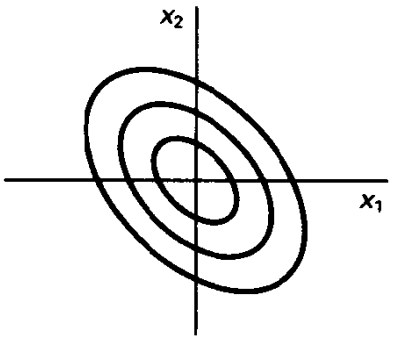
$$\begin{aligned}f(x_1, \ldots, x_n) &= \prod_{i=1}^{n} \frac{1}{\left(2\pi\sigma_i^2\right)^{1/2}} e^{(-x_i - m_i)^2/2\sigma_i^2} \\ &= f(x_1)f(x_2)\cdots f(x_n),\end{aligned} \tag{2.3.14}$$



(a) $\rho = 0$



(b) $\rho = 0.7$



(c) $\rho = -0.7$

Figure 2.3.1 Level contours of bivariate Gaussian p.d.f.'s. All have $m_1 = m_2 = 0$, $\sigma_1 = \sigma_2$.

proving independence. To emphasize the point, uncorrelated Gaussian random variables are independent.

Additional analytical convenience pertains to Gaussian r.v.'s. First, marginal densities of Gaussian random variables are also of Gaussian form, and conditional p.d.f.'s for one of the Gaussian variables, conditioned upon knowledge of another, are also of Gaussian form. Furthermore, if we let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ be obtained by any linear transformation of $\mathbf{X}$, that is, $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$, where $\mathbf{A}$ is a real invertible $n \times n$ matrix and $\mathbf{b}$ is a real $1 \times n$ vector, then $\mathbf{Y}$ is still jointly Gaussian, albeit with a new mean vector and new covariance matrix. This is easily demonstrated by solving for $\mathbf{x}$ in terms of $\mathbf{y}$ and then substituting into (2.3.12) and observing the quadratic form exponent. (See also Exercise 2.3.4.) Thus, analysis of Gaussian variables undergoing

linear transformation requires only consideration of the mean vector and covariance matrix.

As an extension of the correlation of two random variables, consider the random variable Z defined by the product of n variables, $Z = X_1 X_2 \dots X_n$. Its expectation is

$$E[Z] = E[X_1 X_2 \cdots X_n] = \int \cdots \int x_1 x_2 \cdots x_n f(\mathbf{x})\, d\mathbf{x}, \tag{2.3.15}$$

which cannot in general be further simplified. However, if the variables are independent, then we may factor the n-dimensional p.d.f. and obtain

$$\begin{aligned} E[X_1 \cdots X_n] &= \int x_1 f(x_1)\, dx_1 \cdots \int x_n f(x_n)\, dx_n \\ &= E[X_1] \cdots E[X_n]. \end{aligned} \tag{2.3.16}$$

Thus, *provided independence holds*, the expected value of a product of random variables is the product of their expected values.

2.3.3 Characteristic Functions

Another expectation, whose importance will shortly be apparent, is that of the random variable defined by the transformation $g(X) = e^{j\omega X}$, where $j = -1^{1/2}$ [12] This gives

$$E\left[e^{j\omega X}\right] = \Phi_X(\omega) = \int_{-\infty}^{\infty} f_X(x) e^{j\omega x}\, dx, \tag{2.3.17}$$

which is a complex function in ω and is known as the ***characteristic function*** of X. Its usefulness will be seen subsequently in handling of sums of r.v.'s, especially independent ones, and in finding moments of r.v.'s. In regard to moments, note that

$$\frac{d\Phi_X(\omega)}{d\omega} = \int_{-\infty}^{\infty} jx f_X(x) e^{j\omega x}\, dx \tag{2.3.18}$$

and thus

$$\overline{X} = -j \left. \frac{d\Phi_X(\omega)}{d\omega} \right|_{\omega = 0} \tag{2.3.19}$$

Extension of this argument relates higher-order moments to the characteristic function:

$$\overline{X^n} = (-j)^n \left. \frac{d^n \Phi_X(\omega)}{d\omega^n} \right|_{\omega = 0} \tag{2.3.20}$$

Those familiar with the Fourier transform of linear system theory will recognize the characteristic function $\Phi_X(\omega)$ as (within a sign in the exponent) the Fourier transform of the probability density function $f_X(x)$. As expected, the inverse transform gives the

[12] The variable ω here should not be confused with the earlier usage for outcomes in a sample space.

p.d.f in terms of the characteristic function:

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi_X(\omega)e^{-j\omega x}\,d\omega. \tag{2.3.21}$$

Example 2.13 Characteristic Function of Gaussian R.V.

By substituting the Gaussian p.d.f. (2.2.10) into (2.3.17), we obtain

$$\Phi_X(\omega) = \int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-(x-\mu)^2/2\sigma^2} e^{j\omega x}\,dx. \tag{2.3.22}$$

After completing the square in the exponent and then recognizing that the integral of a Gaussian p.d.f. is 1, we obtain

$$\Phi_X(\omega) = e^{j\mu\omega - \sigma^2\omega^2/2}. \tag{2.3.23}$$

As a check, the first two moments by (2.3.19) and (2.3.20) are $\overline{X} = \mu$ and $\overline{X^2} = \mu^2 + \sigma^2$, as we earlier determined without the use of characteristic functions.

Characteristic functions provide an easy route to another important probability density law: the p.d.f of the random variable $Y = \sum_{i=1}^{n} X_i$, where X_i are independent, is obtained by *convolving* the density functions for the variables X_i. That is,

$$f_Y(u) = f_{X_1}(u) * f_{X_2}(u) * \cdots * f_{X_n}(u). \tag{2.3.24}$$

where * denotes the ***convolution*** of two functions:

$$g(u) * h(u) = \int_{-\infty}^{\infty} g(z)h(u - z)\,dz. \tag{2.3.25}$$

To demonstrate this result, we use

$$\begin{aligned} \Phi_Y(\omega) &= E\left[e^{j\omega\sum X_i}\right] = E\left[\prod_{i=1}^{n} e^{j\omega X_i}\right] \\ &= \prod_{i=1}^{n} E\left[e^{j\omega X_i}\right] = \prod_{i=1}^{n} \Phi_{X_i}(\omega). \end{aligned} \tag{2.3.26}$$

Equation (2.3.16) was used in the third step of (2.3.26).

Now recall that the characteristic function and the probability density function are a Fourier transform pair. The convolution theorem of Fourier transform calculus would hold that

$$\begin{aligned} f_Y(u) &= F^{-1}(\Phi_Y(\omega)) \\ &= F^{-1}\left(\prod_{i=1}^{n} \Phi_{X_i}(\omega)\right) = f_{X_1}(u) * \cdots * f_{X_n}(u). \end{aligned} \tag{2.3.27}$$

In Figure 2.3.2, we show the p.d.f. for the sum of two and three independent uniform random variables obtained by convolution. The p.d.f. for the sum of three such variables is already suggestive of the Gaussian density; this is no special occurrence, as we will demonstrate in the following section.

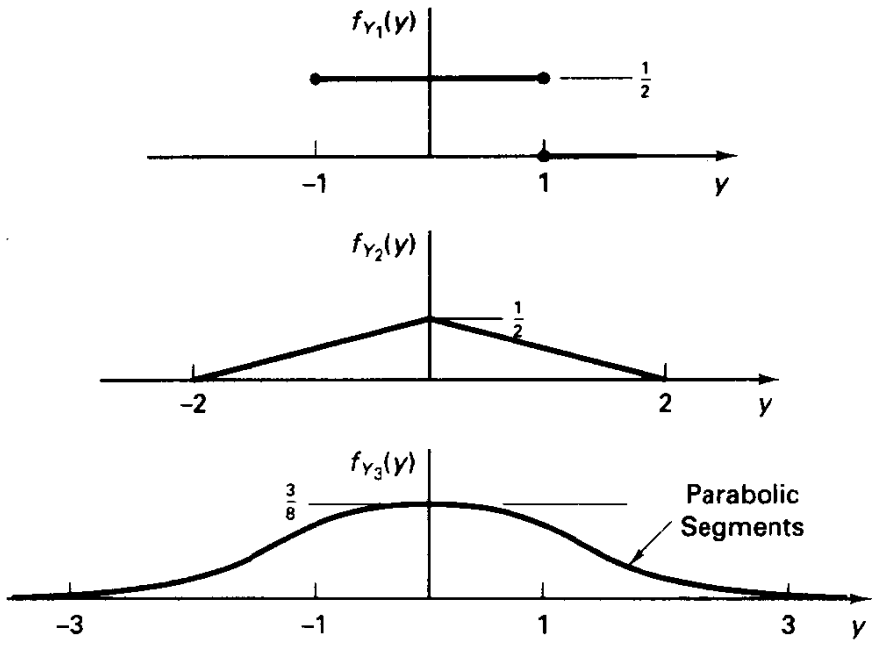


Figure 2.3.2 Probability density functions for sum of one, two, and three independent uniform r.v.'s.

2.4 PROBABILITY BOUNDS AND LIMIT THEOREMS

We will usually be unable, at least without Herculean effort, to evaluate exactly the probability of an event of interest, for example, the probability of a transmission error in sending a message through a noisy channel. In such cases, upper and/or lower bounds on the probability are often acceptable, especially if the underlying mathematics or our experience shows the bounds are reasonably tight. Beyond being a computational tool, probability bounds are useful as well in proving many of the major results of statistical theory and information theory.

Later in the section, we focus on results of paramount importance to communications and information theory: laws of large numbers related to sums of random variables and the distribution of such sums as the number of variables increases, the central limit theorems.

2.4.1 Bounds Based on First and Second Moments

Simple, but generally not very tight, bounds can be stated in terms of only first and second moments of the random variable under study. Let Y be a *nonnegative* random variable with p.d.f. $f(y)$. Then, for $\alpha > 0$, we have

$$\begin{aligned} P(Y \geq \alpha) &= \int_{y \geq \alpha} f(y)\,dy \\ &= \int_{y \geq 0} u(y-\alpha) f(y)\,dy = E[u(Y-\alpha)], \end{aligned} \tag{2.4.1}$$

where $u(x)$ is the unit-step function:

$$u(x) = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{2.4.2}$$

Since $f(y)$ is nonnegative, and $u(y - \alpha) \leq y/\alpha$ for $y \geq 0$, we obtain

$$P(Y \geq \alpha) \leq \int_{y \geq 0} \frac{y}{\alpha} f(y)\, dy = \frac{E[Y]}{\alpha}. \tag{2.4.3}$$

This result is known as the ***Markov inequality.*** A similar bound may be derived for the case of discrete random variables, with sums replacing integrals and probabilities replacing p.d.f.'s.

The fact that the bound depends only on the mean of Y is both good and bad news—it's easy to compute, but because it requires only knowledge of the mean of the random variable, the bound is unlikely to be very tight. This inequality says, for example, that if we select a person at random from an adult population whose mean height is 1.6 meters the probability that person has height 3.2 meters or greater is less than $\frac{1}{2}$! On the other hand, it is possible to construct random variables for which the bound gives exact results; a random variable taking on values 0 or a positive constant, α, will produce equality in (2.4.3).

A variation on (2.4.3), which is perhaps more familiar, is obtained by letting Y be the squared deviation of a r.v. X from its mean; that is, $Y = (X - \overline{X})^2$. Y is obviously a nonnegative r.v., and (2.4.3) gives

$$\begin{aligned} P((X - \overline{X})^2 \geq \alpha) &\leq \frac{E\left[(X - \overline{X})^2\right]}{\alpha} \\ &= \frac{\mathrm{Var}[X]}{\alpha}. \end{aligned} \tag{2.4.4}$$

We could just as well take $\beta = \alpha^{1/2}$ and obtain

$$P(|X - \overline{X}| \geq \beta) \leq \frac{\mathrm{Var}[X]}{\beta^2}, \tag{2.4.5}$$

which is known as the ***Chebyshev inequality***. Consider again the heights of randomly selected persons. If, in addition to specifying the mean height of 1.6 meters, we state that the standard deviation of height is 0.2 meter, then the probability that a person is shorter than 1 meter or taller than 2.2 meters (i.e., more than three standard deviations in either direction from the mean height) is less than $(0.2)^2/[3(0.2)]^2 = 0.111$. This is a slightly more realistic prediction than made earlier, at least given our prior knowledge about the distribution of heights in typical populations. Notice we have employed first- and second-moment information, but nothing else.

Despite the apparent weakness of these bounds, they are strong enough to assert certain laws of large numbers, which in turn are at the very heart of information theory. We will develop this further later in this chapter.

2.4.2 Chernoff Bounds

Whereas the upper bounds just presented are often loose, the Chernoff bounding technique [7] often yields much tighter numerical estimates, especially in estimating probabilities

having to do with the tail of a distribution, and has become a fundamental tool in communication systems analysis.

We formulate the Chernoff bound by reinspecting the development of the previous bounds. In the first case, the desired probability was written as the expectation of a function $u(Y - \alpha)$ of the random variable Y, and we proceeded to overbound this function by a linear function $g(Y) = Y/\alpha$. [The second form bound used a quadratic upper bound $g(Y) = (Y/\alpha)^2$.] We are at liberty to choose any function $g(Y)$ overbounding the step function $u(Y - \alpha)$: a computationally simple choice is to let

$$g(Y) = e^{s(Y-\alpha)}, \qquad s > 0, \tag{2.4.6}$$

as shown in Figure 2.4.1, where s is a free (nonnegative) parameter. Furthermore, we no longer require that Y be a nonnegative random variable or that $\alpha > 0$.

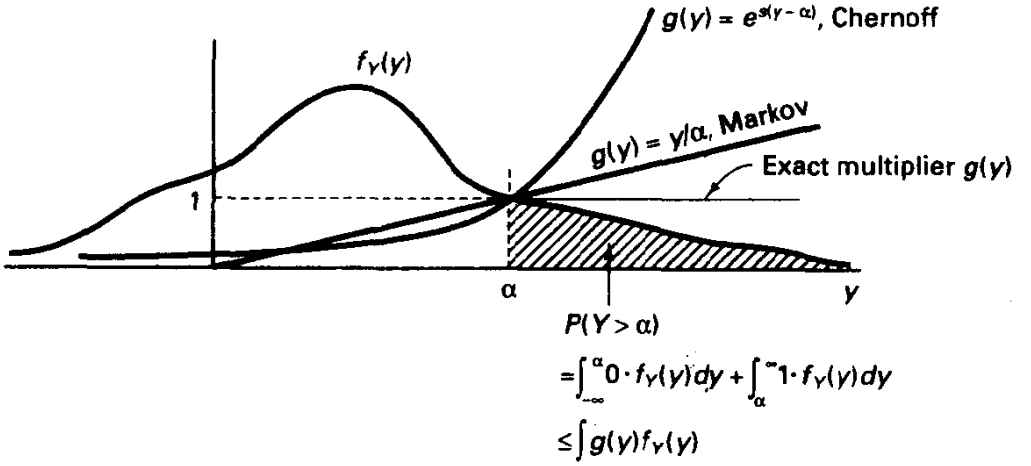


Figure 2.4.1 Bounding of tail probabilities.

Using this choice for $g(Y)$, we have

$$P(Y \geq \alpha) \leq E\left[e^{s(Y-\alpha)}\right] = e^{-s\alpha}E[e^{sY}], \qquad s > 0, \tag{2.4.7}$$

which, after the expectation is taken, is only a function of s for a given α. To obtain the tightest upper bound, we can minimize the right-hand side of (2.4.7) with respect to s, yielding the general ***Chernoff bound***

$$P(Y \geq \alpha) \leq \min_{s>0} e^{-s\alpha}E[e^{sY}]. \tag{2.4.8}$$

Example 2.14 Bounds on Binomial Probabilities

Consider the independent transmission of 100 binary symbols through the error-prone channel of Example 2.3, with the channel error probability now changed to 0.01. We inquire about the probability that more than four errors will occur in the transmission, for we may have a code embedded in the transmission that is capable of correcting four or fewer errors in the transmission of 100 symbols.

Let Y be the random number of errors. Y is the sum of 100 zero/one random variables, X_i, with $X_i = 1$ denoting the error event. Since each r.v. X_i has an expected value of 0.01, we observe by (2.3.7) that $E[Y] = 100(0.01) = 1$. In other words, the expected number of errors in 100 transmissions is 1.

Before developing bounds, we observe that it is not difficult to evaluate the *exact probability in this case. It is*

$$\begin{aligned} P(Y > 4) &= 1 - P(Y \leq 4) \\ &= 1 - \sum_{i=0}^{4} C_i^{100}(0.01)^i(0.99)^{100-i} = 0.0034. \end{aligned} \tag{2.4.9}$$

The Markov inequality would say that, because Y is nonnegative and $\overline{Y} = 1$,

$$P(Y > 4) = P(Y \geq 5) = P(Y \geq 5\overline{Y}) \leq \frac{1}{5}, \tag{2.4.10}$$

which is pessimistic by nearly two orders of magnitude. (The reader is invited to compute the Chebyshev bound for this same question in Exercise 2.4.3; the result is 0.0618.)

To determine the Chernoff bound, we first compute

$$E\left[e^{sY}\right] = E\left[e^{s\sum X_i}\right] = E\left[\prod_i e^{sX_i}\right] = \prod_{i=1}^{100} E\left[e^{sX_i}\right]. \tag{2.4.11}$$

The last step follows from the independence of the 100 variables $X_1, \ldots, X_{100}$. Each term in the final product is the same and evaluates to $0.99+0.01e^s$ since X_i is 0 or 1. Substituting this into (2.4.8), we have

$$P(Y \geq 5) \leq \min_{s>0} e^{-5s}(0.99 + 0.01e^s)^{100} \tag{2.4.12}$$

Minimizing (2.4.12) with respect to s yields a best value of $s = 1.651$, giving $P(Y \geq 5) = 0.016$, a much more reasonable approximation to the true probability 0.0034.

Example 2.15 Chernoff Bound on Gaussian Tail Integral

Suppose X is a Gaussian r.v. with zero mean and unit variance. Earlier, the Gaussian tail integral was defined in (2.2.12) as

$$P(X \geq x) = Q(x) = \int_x^\infty \frac{e^{-z^2/2}}{(2\pi)^{1/2}}\, dz.$$

The Chernoff bound on this probability is

$$\begin{aligned} P(X \geq x) &\leq \min_{s>0} e^{-sx} E\left[e^{sX}\right], \qquad s > 0 \\ &= \min_{s>0} e^{-sx} \int_{-\infty}^{\infty} e^{sz} \frac{e^{-z^2/2}}{(2\pi)^{1/2}}\, dz. \end{aligned} \tag{2.4.13}$$

By completing the square on the integrand exponent, we have

$$\begin{aligned} P(X \geq x) &\leq \min_{s>0} e^{-sx} \int_{-\infty}^{\infty} \frac{e^{-(z-s)^2/2}}{(2\pi)^{1/2}} e^{s^2/2}\, dz \\ &= \min_{s>0} e^{-sx} e^{s^2/2}. \end{aligned} \tag{2.4.14}$$

This expression is minimized when $s = x$, giving the resultant bound

$$P(X \geq x) \leq e^{-x^2/2}. \tag{2.4.15}$$

We have earlier seen in Section 2.2 that $e^{-x^2/2}/2$ is an upper bound to $Q(x)$, and in fact a tighter upper bound for large x was $e^{-x^2/2}/(2\pi)^{1/2}x$. Thus, the Chernoff bound does not strengthen these results for large x (which were after all obtained with some care

involving the actual p.d.f.), but it should be observed that the three expressions all have the *same exponential dependence* on the parameter x. This will always be the case with the Chernoff bounding procedure: it gives the correct and strongest exponential dependence on the parameters of the problem, if an exponential dependence exists.

2.4.3 Sequences, Sums, and Laws of Large Numbers

We now turn to sequences of random variables, their partial sums, and the asymptotic behavior of these for large numbers of variables. Our aims are to formulate ***laws of large numbers*** and the ***central limit theorem***, the tendency for normalized sums of r.v.'s to approach, in cumulative distribution, the Gaussian random variable.

Let $X_1, X_2, \ldots, X_i, \ldots$ be a semiinfinite sequence of random variables, which for simplicity we model as independent, having common first-order density $f_X(x)$, mean m, and variance σ^2. For example, these variables might be {0, 1} binary variates associated with an information source, or they might have a uniform density on (a, b). We let

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad n = 1, 2, \ldots, \tag{2.4.16}$$

be the *running averages* associated with the X_i sequence. Note that S_n is another sequence of random variables induced by the sequence X_i. However, despite independence of successive X_i's, the running average sequence $S_1, S_2, \ldots, S_n, \ldots$ is strongly correlated. Sample functions of the random sequence S_n are illustrated in Figure 2.4.2.

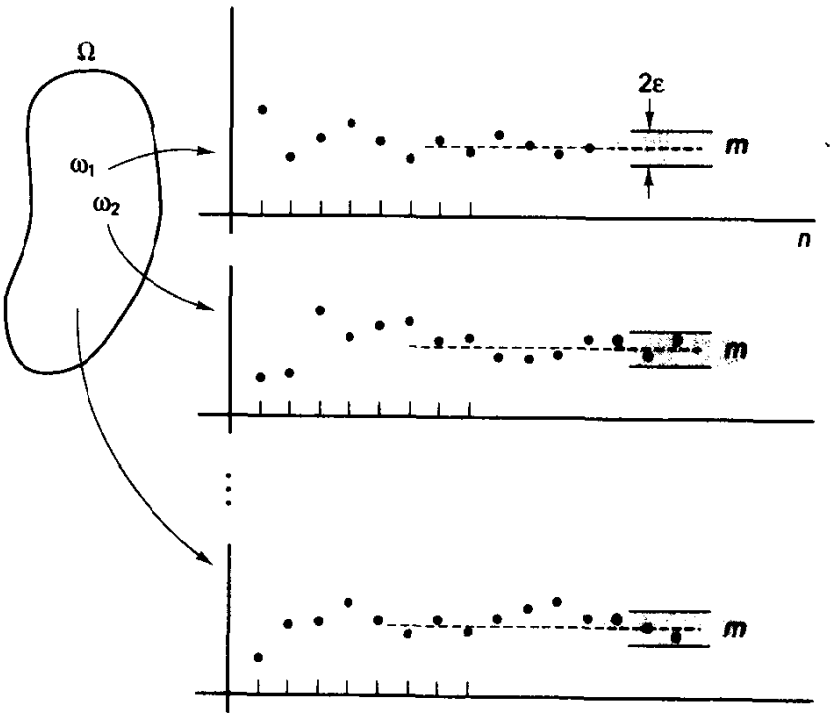


Figure 2.4.2 Convergence of arithmetic averages of r.v.'s, $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Even a casual understanding of probability theory prompts us to speculate that, as n becomes large, the r.v. S_n should be near the expected value of X. For example, in sums of the output of an equiprobable 0/1 binary source, we should see a tendency for S_n to eventually fluctuate near $\frac{1}{2}$. (We should be wary though of the oft-heard misinterpretation of the following kind: after observing nine consecutive tosses of a fair coin to be heads, the "law of averages must catch up with us," implying that a tail is sure, or at least more probable than 1/2, on the next toss. If we believe in independence of r.v.'s, this is certainly fallacious reasoning.)

Now consider the implications of the Chebyshev inequality. From linearity of the expectation operator, expressed in (2.3.7), we find that at time n

$$E[S_n] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = m \tag{2.4.17a}$$

and, by virtue of independence of the summed variables,

$$\mathrm{Var}[S_n] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[X_i] = \frac{\sigma^2}{n}. \tag{2.4.17b}$$

Then, from (2.4.5), we have that

$$P(|S_n - m| \geq t) \leq \frac{\sigma^2}{nt^2}, \tag{2.4.18}$$

which holds that for any interval $(m - t, m + t)$; as n increases, the normalized partial sum random variable S_n is increasingly probable to lie in the interval. In probability theory, we say that the sequence of random variables S_n ***converges in probability*** to m, since the probability of S_n being within t units of the mean goes to 1 as $n \to \infty$, for any $t > 0$. This also provides a statement of a ***weak law of large numbers***, so called because the convergence in probability is a weak form of convergence of random sequences, saying only that the probability of the sum being "typical" at any time n is high. The weak law claims nothing about the convergence behavior of individual sample paths in the ensemble of partial sum sequences. (Figure 2.4.3 illustrates a contrived example wherein some of the partial sum sample paths do not converge to m in the classical sense of convergence, yet the probability of selecting a sample path that is within the tolerance band increases with time.)

Actually, it is possible to say stronger things about the convergence of the sequence S_n under the assumptions above. Because the variance of S_n decreases monotonically toward zero by (2.4.17b), we say S_n ***converges in mean square*** to m. (This form of convergence implies convergence of probability, by (2.4.19).) Still stronger laws of large numbers are unnecessary for our purposes, but they would reveal that S_n ***converges with probability one*** (also called almost everywhere convergence), meaning that virtually all experiment outcomes would have sample sums which converge to m in the usual sense of convergence for deterministic sequences.

Although we fashioned the problem in a rather restricted vein, requiring independence and a common distribution, it may be shown that considerably weaker conditions suffice for the kinds of convergence we have seen here. The reader is referred to the ex-

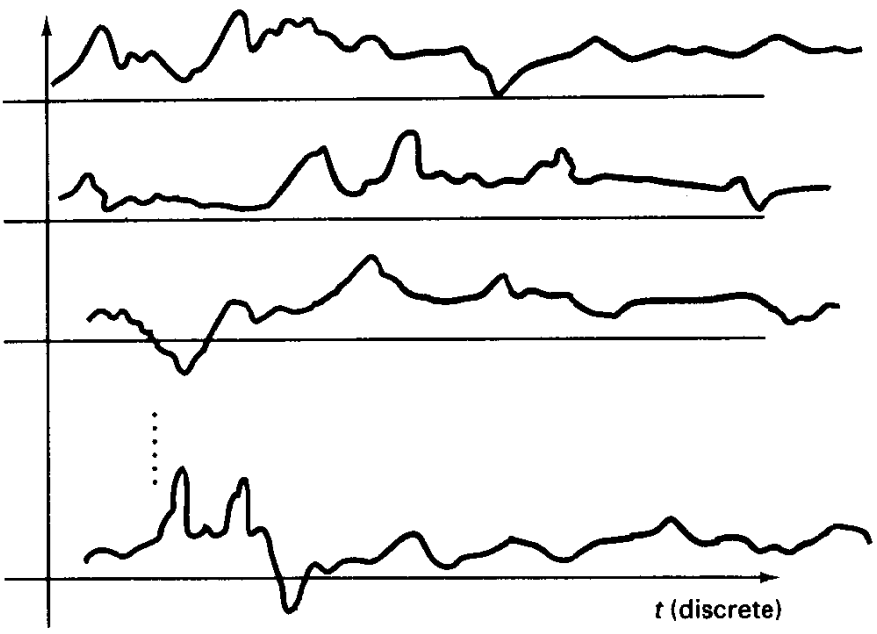


Figure 2.4.3 Process that converges in probability, $P[X(t)$ in shaded box$] \to 1$ as $t \to \infty$, but that does not converge with probability 1; that is, each sample function may wander outside shaded band occasionally forever.

cellent treatments of stochastic convergence found in Papoulis [1] and Gray and Davisson [3] for further information on this topic.

2.4.4 Central Limit Theorem

This classical theorem, with its several variations, imbues the Gaussian distribution with special significance in probability theory. Loosely stated, it holds that an appropriately normalized sum of independent random variables has a distribution tending to the Gaussian distribution as the number of summed variables becomes large. We shall demonstrate the result for a special case—where the random variables are independent and identically distributed (i.i.d.).

Theorem. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with finite mean m and finite variance σ^2 and with higher-order moments all finite. Let $Y_i = (X_i - m)/\sigma$ be a normalized random variable. Then the random variable

$$Z_n = n^{-1/2} \sum_{i=1}^{n} Y_i = n^{-1/2} \sum_{i=1}^{n} \frac{X_i - m}{\sigma} \tag{2.4.19}$$

converges in distribution to the standard Gaussian random variable having zero mean and unit variance; that is, $F_{Z_n}(z) \to 1 - Q(z)$ as $n \to \infty$, where $Q(x)$ is as defined in (2.2.10).

Before providing the proof, we noted in the previous section that, given the theorem conditions, the arithmetic average S_n (that is with normalization by n) converges in probability to the mean (the weak law of large numbers). Here we normalize differently

so that the limiting random variable does not collapse to a degenerate one (a constant). Also, we observe that the translation and scaling make each of the variables Y_j zero mean with unit variance. This in turn implies that Z_n will be zero mean and have unity variance for all n as well, the latter following from independence of the Y_i.

Proof of Theorem:

Our proof relies on characteristic functions. The characteristic function of Z_n is

$$\begin{aligned}\Phi_{Z_n}(\omega) &= E\left[e^{j\omega Z_n}\right] = E\left[e^{j\omega n^{-1/2}\sum Y_i}\right] = E\left[\prod_{i=1}^{n} e^{j\omega n^{-1/2}Y_i}\right] \\ &= \prod_{i=1}^{n} E\left[e^{j\omega n^{-1/2}Y_i}\right] = \left[\Phi_Y(\omega n^{-1/2})\right]^n\end{aligned} \tag{2.4.20}$$

since the Y_i variables are independent and identically distributed.

The characteristic function in brackets may be expanded as a power series in $\omega n^{-1/2}$:

$$\Phi_Y(\omega n^{-1/2}) = 1 + \frac{\omega}{n^{1/2}}\Phi_Y'(0) + \left(\frac{\omega}{n^{1/2}}\right)^2 \frac{\Phi_Y''(0)}{2!} + \cdots, \tag{2.4.21}$$

where the superscript primes denote differentiation with respect to ω.

Next, recalling how moments were linked to derivatives of the characteristic function in Section 2.3, we have that

$$\Phi_Y(\omega n^{-1/2}) = 1 + j\frac{\omega}{n^{1/2}}m_Y - \frac{\omega^2}{2n}\sigma_Y^2 + \frac{j\omega^3}{6n^{3/2}}E\left[Y^3\right] + \cdots. \tag{2.4.22}$$

But since Y has zero mean and unit variance, (2.4.22) becomes

$$\Phi_Y(\omega n^{-1/2}) = 1 - \frac{\omega^2}{2n} + \frac{1}{n^{3/2}}r(n), \tag{2.4.23}$$

where $r(n)$ accounts for the remaining terms involving third- and higher-order moments of Y_n. By the theorem conditions, $r(n)$ is finite and will not increase with n.

Now, substituting (2.4.23) into (2.4.20) and taking logarithms of both sides of the equation gives

$$\log_e \Phi_{Z_n}(\omega) = n\log_e\left[1 - \frac{\omega^2}{2n} + \frac{r(n)}{n^{3/2}}\right]. \tag{2.4.24}$$

The expansion $\log_e(1+x) = x - (x^2/2!) + \cdots$ applied to the right-hand side yields

$$\log_e \Phi_{Z_n}(\omega) = n\left[-\frac{\omega^2}{2n} + \frac{r(n)}{n^{3/2}} - \frac{1}{2}\frac{\omega^4}{4n^2} + \cdots\right]. \tag{2.4.25}$$

and as $n \to \infty$ we have

$$\lim_{n\to\infty} \log_e \Phi_{Z_n}(\omega) = -\frac{\omega^2}{2}. \tag{2.4.26}$$

This implies that the limiting characteristic function for Z_n is

$$\lim_{n\to\infty} \Phi_{Z_n}(\omega) = e^{-\omega^2/2} \tag{2.4.27}$$

which can be recognized from (2.3.23) as the characteristic function of a standard $\mu = 0, \sigma^2 = 1$ Gaussian random variable, proving that Z_n converges in distribution to the standard Gaussian random variable.

The theorem conditions may be considerably relaxed without changing the fundamental conclusion: Most importantly, the independence assumption may be relaxed in favor of a *mixing* property. Essentially, what is required is that the variables involved in the sum are asymptotically uncorrelated for large positional separation. While the limiting distribution is still Gaussian, the rate of convergence in n is somewhat slower when the sequence of random variables is not independent.

It is easy to see by example how the Gaussian distribution emerges from sums of familiar random variables. If Y_i are uniform on $[-3^{1/2}, 3^{1/2}]$, thereby having zero mean and unit variance, the p.d.f. for the sum of two and three of these independent random variables was shown in Figure 2.3.2. These are obtained by repeated convolutions of the underlying density as described in (2.3.27). The piecewise-quadratic density for $n = 3$ already looks remarkably similar to the Gaussian density; the departure of the density from Gaussian in the tails is significant, however, and often crucial in analysis.

Visitors to science museums often encounter a demonstration of the convergence of sums of *binary* random variables to the Gaussian distribution. A large number of balls is allowed to ripple downward under gravitational action through a layered network of pegs into a number of bins, the ball taking either a left or right direction at each successive peg independently of the previous trajectory. The final bin placement relative to the initial horizontal position may then be regarded as the sum of several binary random variables. After a large number of balls have traversed the maze of pegs, the histogram formed by the heights of the piles in each compartment is seen to mimic the Gaussian density function, albeit a discrete approximation to it. This points out that the limiting distribution revealed by the central limit theorem is Gaussian in integral form. Sums of discrete random variables will always have a probability density function comprised of impulse functions and thus cannot converge in the usual sense to a continuous density function, but the cumulative distribution function will converge to that of the Gaussian random variable. The same holds for the distribution of the number of errors occurring in n uses of a BSC, as discussed in Example 2.14, provided n is large and $np \gg 1$.

2.5 STOCHASTIC PROCESSES

In the analysis of digital communications systems, we need to describe signals that evolve randomly over time, such as channel noise impairments. This evolution may be either a continuous-time or discrete-time process. The theory of stochastic, or random, processes provides us the analysis methods.

Stochastic processes should be understood as a natural extension of the concept of a random variable or random vector. Recall the view of a random variable—the assignment of a real number $x(\omega)$ (or perhaps vector of numbers) to an outcome ω in the sample space Ω. These could be called *realizations* of the random variable. In a stochastic process we associate a function of time, $x(\omega, t) - \infty < t \infty$ with each point $\omega \in \Omega$ and call each a ***sample function*** or realization of the random process. The entire

collection of these sample functions is called the ***ensemble*** or simply the ***process***. When we speak of a random process, we figuratively have in mind the entire ensemble, not a single sample function, although in practice we generally deal with one sample function. We write $x(\omega, t)$ when we wish to explicitly denote the sample function assigned to ω. This raises an important practical issue, which we shall address shortly.

More formally, a stochastic process $X(t)$ is merely an infinite collection of random variables, indexed by time in some index set I; that is, $X(t) = \{X(t_i), t_i \in I\}$. If the index set is discrete, we refer to the process as a discrete-time process or a random sequence, while if the index set is a continuous variable, we say $X(t)$ is a continuous-time process.

Figure 2.5.1a illustrates this ensemble viewpoint for a continuous-time process and also the concept that if we freeze time, say at $t = t_0$, then the collection of sample values $x(\omega_i, t_0)$ is a random variable, just as described previously. This random variable has all the attributes we have discussed—a distribution function, density function, moments, and so on. One distinction is important, however: the exact nature of these quantities may depend on the choice of t_0. For example, the first-order distribution function should be written as $F_{X_0}(x_0; t_0)$ to indicate explicitly this time dependence. Likewise, probability densities and quantities derived from them should carry an explicit time tagging.

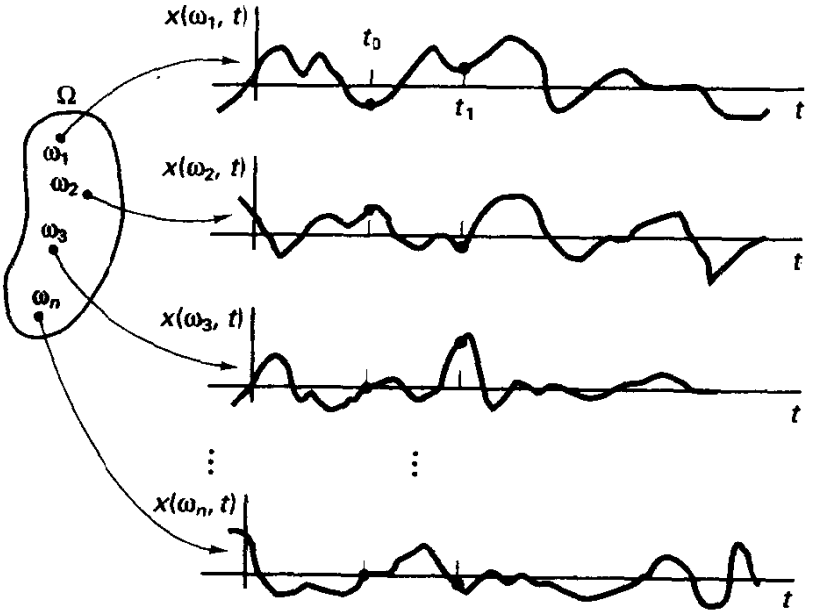


Figure 2.5.1a Ensemble of sample functions.

To generalize this, consider two instants t_0 and t_1. The process values at these two times have bivariate cumulative distribution and probability density functions that depend in general on the exact values of these times. For example, the joint distribution function would be written as $F_{X_0,X_1}(x_0, x_1; t_0, t_1)$, where X_i denotes the random variable defined by $X(\omega, t_i)$. The function specifies the probability that at times t_0 and t_1 the associated random variables $X(\omega, t_0)$ and $X(\omega, t_1)$ are, respectively, less than x_0 and x_1. This lends a *ceiling function* interpretation, as indicated in Figure 2.5.1b.