STOCHASTIC PROCESSES
CONCEPTS FOR APPLICATIONS

R. G. Gallager

July 20, 2004

Chapter 1

INTRODUCTION AND PROBABILITY REVIEW

1.1 Introduction

A stochastic process (or random process) is a probabilistic experiment of model that evolves in time. That is, each sample point (i.e., possible outcome) of the experiment is a function of time that is called a sample function. The sample space is the set of possible sample functions, and events are subsets of sample functions. Finally, there is a rule for determining the probabilities of various events. As an example, we might be concerned with arrivals to some system. The arrivals might be incoming jobs for a computer system, arriving packets to a communication system, patients in a health care system, or orders for some merchandising warehouse.

Example 1.1 Suppose, to look at a particularly simple case, that arrivals occur only at integer instants of time, at most one arrival occurs at any given integer time, and the time between arrivals is one with probability 1/2 and two with probability 1/2; the time between arrival n and arrival $n+1$ is independent of all previous inter-arrival intervals. Finally, the first arrival occurs at time 1 or 2, with equal probability.

A sample point for such a process could be regarded as the set of instants at which arrivals occur. Alternatively, we could view a sample function as $\{n(t); t \geq 0\}$, where $n(t)$ is the number of arrivals up to and including time t (see Figure 1.1. One can then ask questions such as: for any given t, what is the probability distribution of the number of arrivals from 0 to t? what is the probability of an arrival at t?, is it meaningful to talk about the average arrival rate, $n(t)/t$, in the limit $t \to \infty$?

A more general form of the above simple process is to allow the inter-arrival intervals (i.e., the times between successive arrivals) to be independent, identically distributed random variables with an arbitrary discrete or continuous distribution function. Such stochastic

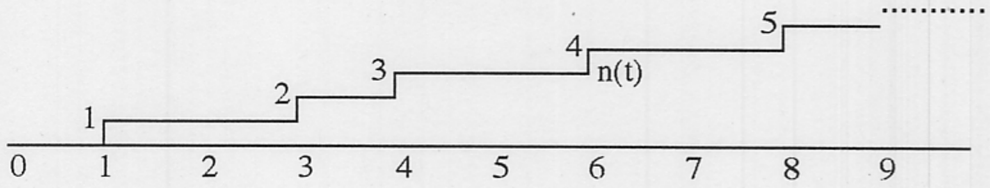

Figure 1.1: A sample function of the arrival process in example 1.1

processes are called renewal processes and play a central role in the theory of stochastic processes.

Renewal processes are examples of *discrete stochastic processes.* The distinguishing characteristic of such processes is that interesting things (arrivals, departures, changes of state) occur at discrete instants of time separated by deterministic or random intervals. Discrete stochastic processes are to be distinguished from noise-like stochastic processes in which changes are continuously occurring and the sample functions are continuously varying functions of time. The description of discrete stochastic processes above is not intended to be precise. The various types of stochastic processes developed in subsequent chapters are all discrete in the above sense, however, and we refer to these processes, somewhat loosely, as discrete stochastic processes in what follows.

Discrete stochastic processes find wide and diverse applications in operations research, communication, control, computer systems, management science, etc. Paradoxically, we shall spend relatively little of our time discussing these particular applications, and rather spend our time developing results and insights about these processes in general. We shall discuss a number of examples drawn from the above fields, but the examples will be "toy" examples, designed to enhance understanding of the material rather than to really understand the application areas.

One possible approach to applying probabilistic models to an application area is to first gather statistics about the area, then construct the most accurate model possible from those statistics, and then analyze the model. Unfortunately, this approach almost invariably fails, the reason being that real problems are too complicated to model completely. A better approach is to choose a simple and well understood model that is intuitively related to the application area. Analysis and understanding of the model couples with knowledge and some simple statistics about the application area, thensuggest refinements of the model. Both the original choice of model and successive refinements require thorough understanding and insight about a broad class of potential models. This means that the key to successful modeling is the acquisition of insight and deep understanding about a broad class of models. It is for this reason that we stress models here rather than application areas.

The above discussion of modeling is applicable to any class of problems amenable to mathematical analysis. It is particularly applicable to stochastic processes for a number of reasons. First, the range of applications is diverse. Second, many of the most important practical results depend on various limiting operations; understanding these results require time and patience, more than is required in many engineering subjects where many applications lie closer to the surface. Third, there are many paradoxical results that are counter-intuitive until they are thoroughly understood.

1.2 Probability Review

A *probability model* (or probability experiment) has three ingredients—a sample space, a set of events, and a rule for assigning probabilities to events. The sample space is the set of possible outcomes (sample points) of the experiment. When the experiment is performed, one (and only one) of these outcomes occurs. An event is a subset of the sample space and the probability of that event is a real number between 0 and 1. Note that the technical meanings of outcome and event are not quite the same as their ordinary usage in English. An outcome completely determines the result of the experiment, specifying a single sample point, whereas an event partially determines the result, specifying only that the outcome lies in the subset of the sample space corresponding to the event.

For the simplest type of experiment, the number of sample points is finite or countably infinite (i.e., the sample points can be listed $(\omega_1, \omega_2, \ldots)$). Each sample point, or, more precisely, each event consisting of a single sample point, has some probability associated with it. For an arbitrary event, the probability of that event is the sum of the probabilities of the sample points making up the event. All the probabilities are non-negative and the sum of the probabilities of all sample points is 1.

Example 1.2 Consider an experiment of two flips of a coin. The sample space consists of four sample points $\{(HH), (HT), (TH), (TT)\}$ and the probability of each sample point is taken to be 1/4. Each subset of the sample space corresponds to an event. For example the event that the first coin flip is heads is the subset $\{(HH), (HT)\}$ and has probability 1/2. The event that at least one flip results in heads is the subset $\{(HH), (HT), (TH)\}$ and has probability 3/4.

Note that by choosing the sample space as above, we have ruled out the possibility of the coin getting lost before the second toss or of coming to rest on its edge. By choosing the probabilities as above, we have ruled out the possibility of a bias toward heads or tails. That is, once a probability model is selected, all questions about the physical situation have been resolved (at least for as long as one continues to use that model).

It would seem that the class of experiments involving only a finite number of sample points is trivial, involving nothing more than defining the probability of each sample point and then adding these probabilities appropriately. What makes some of these problems non-trivial is the very large number of sample points and their combinatorial aspects.

Example 1.3 Suppose we modify example 1.1 to consider only the first 1000 arrivals for the process. Then a sample point becomes a sequence of 1000 arrival instants and there are 2^{1000} sample points. Finding the probability of an arrival at time 1000 by adding up the probabilities of individual sample points is highly impractical.

For some sample spaces, it is not only impractical but also meaningless to determine the probability of events from the probabilities of individual sample points. This is illustrated in the next example, and is one reason why we focus on the probabilities of events rather than sample points.

Example 1.4 Consider a sample space consisting of the set of real numbers between 0 and 1, and assume that the outcome is uniformly distributed over this interval. Each sample point then has zero probability, and there is no way to add up these zero probability single point events in a meaningful way. It is reasonable in this case to take the probability of any interval as the size of that interval, and to take the probability of the union of disjoint intervals as the sum of the sizes of the intervals.

For examples such as example 1.4, one must assign probabilities to events directly. Along with the restriction that the probability of each event must lie between 0 and 1 and the probability of the sample space itself is 1, there is the final restriction that for any sequence of disjoint events $E_1, E_2, \ldots$, the probability of the union of these events is given by

$$\Pr(\cup_i E_i) = \sum_i \Pr(E_i). \tag{1.1}$$

These three restrictions can be regarded as the axioms of probability[1]. For a countable[2] set of outcomes, (1.1) specifies the probability of an event as the sum of the included sample point probabilities. For more general cases, as in Example 1.4, one must find a way to assign probabilities to enough events that (1.1) suffices to determine the probabilities of all events.

We are not interested here in showing how all of the rules for manipulating probabilities follow from these axioms, but we give three examples to indicate why (1.1) says more than is immediately apparent. First, for any event A, let A^c be the complementary event (i.e., the set of all sample points not in A). Then A and A^c are disjoint and their union is the entire space, which has probability 1. Thus $\Pr(A^c) = 1 - \Pr(A)$. Second if A and B are non-disjoint (i.e., contain sample points in common), then $A \cup B = AB^c \cup AB \cup A^cB$ (where, for example AB^c denotes the intersection of subsets A and B^c) and these events are all disjoint. Thus $\Pr(A \cup B) = \Pr(AB^c) + \Pr(AB) + \Pr(A^cB)$. Third, since $B = AB \cup A^cB$, we can rewrite $\Pr(A \cup B)$ as $\Pr(A) + \Pr(B) - \Pr(AB)$, which gives a convenient relation between the probabilities of unions an intersections.

A sample point of the stochastic process described in Example 1.1 is a time function $n(t)$ $(0 \le t < \infty)$ which is constant except for unit increases at integer times separated by one or two. Each sample point corresponds to an infinite sequence of binary choices between inter-arrivals of 1 or 2, and thus each sample point has probability 0. The independent identically distributed (IID) inter-arrival intervals, which are 1 or 2 with equal probability, allow one to calculate the probability of any event of interest. For example, the probability of an arrival at time 2 is 3/4 (with probability 1/2, the first arrival occurs then, and with probability 1/4, the second arrival occurs then).

Note that many questions concerning the process of Example 1.1 can be answered by considering the process of Example 1.3, thus avoiding the necessity of dealing with sample

[1] One must add the axioms of set theory to this and specify the class of events.

[2] A set is countable if it has a finite number of elements or if its elements can be put in one to one correspondence with the positive integers. See any elementary text on set theory.

points having zero probability. We shall see, as we proceed, that using such artifices is unnatural and counterproductive—it is like avoiding calculus by approximating all functions as discrete functions.

A subtle question arises both in Example 1.1 and 4; namely can we define probabilities for all subsets of the sample space, consistent with the axioms above? The answer, unfortunately, is no, and understanding this requires mathematics beyond the scope of this text. Fortunately, the subsets that cause problems are so bizarre that they do not arise in any of the practical or conceptual problems of interest here. The set of events is now defined as those subsets for which probabilities exist. These subsets are always constructed in such a way that countable unions and intersections of events are also events. Thus one cannot create a subset of undefined probability in the course of ordinary analysis.

There is another simpler type of subtlety that arises in examples 1.1 and 1.4; the probability of each sample point is 0, and thus, from (1.1), any event containing only a finite or countably infinite set of sample points has probability zero. Typically there are also many other kinds of events that have zero probability. We shall find that many of the most important results in stochastic processes are true for all sample points except some such set of probability zero. An event including all sample points except some set of probability zero is referred to as an *event of probability 1*. Fortunately, we can work with these special events of zero probability or probability 1 without a great deal of mathematical sophistication.

1.3 Conditional Probabilities

For any two events A and B (with $\Pr(B) > 0$), the *conditional probability* of A, conditional on B, is defined by $\Pr(A|B) = \Pr(AB)/\Pr(B)$. One visualizes an experiment that has been partly carried out with B as the result. $\Pr(A|B)$ is then the probability of A within the sample space restricted to the event B. It is important to recognize that anything we know about probability can also be applied to such a restricted probability space.

Two events, A and B, are said to be *independent* if $\Pr(AB) = \Pr(A)\Pr(B)$. For $\Pr(B) > 0$, this is equivalent to $\Pr(A|B) = \Pr(A)$; i.e., A and B are independent if the observation of B does not change the probability of A. Such intuitive statements about "observation" and "occurrence" are helpful in reasoning probabilistically, but also can lead to great confusion. For $\Pr(B) > 0$, $\Pr(A|B)$ is defined without any notion of B being observed "before" A. For example $\Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A)$ is simply a consequence of the definition of conditional probability and has nothing to do with causality or observations. This issue caused immense confusion in probabilistic arguments before the theory was placed on a firm mathematical foundation.

The notion of independence is of vital importance in defining, and reasoning about, probability models. One needs a way to assign probability to all events, and this is often done by assuming independence between events. In Example 1.1, we assumed such independence without even defining it. Often, when events are not independent, they are conditionally independent, where A and B are said to be *conditionally independent* given

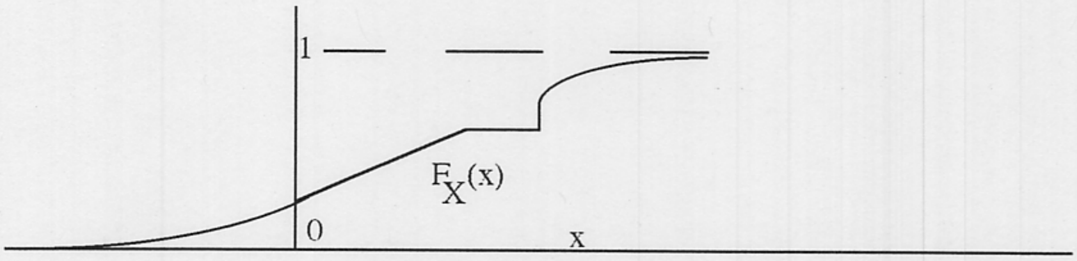

Figure 1.2: Example of a distribution function for a random variable that is neither continuous nor discrete.

C if $\Pr(AB|C) = \Pr(A|C)\Pr(B|C)$. Most of the stochastic process to be studies here are characterized by particular forms of independence or conditional independence.

1.4 Random Variables

The outcome of a probabilistic experiment often contains a collection of numerical values such as temperatures, voltages, numbers of arrivals or departures in various intervals, etc. Each such numerical value varies depending on the particular outcome of the experiment, and thus is a mapping from the set of sample points to the set of numerical values. These variable numerical outcomes, as mapped from the sample points, are called random variables. More precisely, a *random variable* X is defined as a function that maps each point ω in the sample space S (or generally, each point except perhaps those in some subset of probability 0) into the set of finite real numbers. Thus, for any random variable X and any real number x, there is an event $X \le x$; this event is the subset of S whose elements are mapped by X into real numbers less than or equal to x, i.e.,

$$\Pr(X \le x) = \Pr(\{\omega \in S : X(\omega) \le x\}). \tag{1.2}$$

Note that $\Pr(X \le x)$ is a function of the real variable x. It is monotonic nondecreasing, from 0 to 1, as x goes from $-\infty$ to $+\infty$. It is called the *distribution function* of the random variable X and will usually be denoted by $F_X(x)$ (see Figure 1.2).

Note that x is the argument of the function, whereas the subscript X denotes the particular random variable under consideration; if the random variable is clear from the context, we will omit the subscript. We shall always denote random variables by capital letters; this convention is almost universally observed in the field of probability.

As mentioned above, it is permissible for a random variable X to be undefined over some set of sample points comprising an event of zero probability. We shall see many examples of this later, and the student is best advised to postpone concern about it until faced with such examples. These events of zero probability have no effect on the distribution function of X, and, as will be more clear later, no effect on anything we do with random variables.

The concept of a random variable must sometimes be extended to complex random variables and vector random variables. A *complex random variable* is a mapping from the sample space to the set of finite complex numbers, and a *vector random variable* is a mapping from the sample space to the finite vectors in some finite dimensional vector space. Another extension is that of defective random variables. X is *defective* if there is a set of sample points of *positive* probability for which the mapping is either undefined or defined to be either $+\infty$ or $-\infty$. When we refer to random variables in this text (without any modifier such as complex, vector, or defective), we explicitly restrict attention to the original definition, i.e., a function from the sample space (except a subset of zero probability) to the finite real numbers.

If the distribution function $F_X(x)$ of a random variable X has a derivative it is called the *probability density* (or just density) of X and denoted by $f_X(x)$; for sufficiently small δ, $\delta f_X(x)$ approximates the probability that X is mapped into a value between x and $x + \delta$. If the density exists and is finite everywhere, we say that the random variable is continuous. Similarly, if X has a finite or countable number of possible outcomes $x_1, x_2, \dots$, the probability of outcome i is denoted by $\{P_X(x_i)$ and the set of these probabilities, $\{P_X(x_i); i \geq 1$ is called the probability mass function (PMF) of X; such a random variable is called *discrete.* Finally, the *distribution* of a random variable is any rule from which the distribution function can be determined; thus the distribution of X is specified by the density or the PMF or the distribution function.

Elementary probability courses work primarily with density and the PMF, since they are convenient for computational exercises. We will mostly work with the distribution function here. This is partly to avoid saying everything thrice, once for discrete random variables, once for variables with a density, and once for other variables, and partly because it is the distribution function that is most important in limiting arguments such as going to steady state and dealing with time averages.

Often we must deal with multiple random variables in a single probability experiment. For Example 1.1 above, each inter-arrival interval can be regarded as a random variable, taking the value 1 or 2 with equal probability. If $X_1, X_2, \dots, X_n$ are n random variables, their joint distribution function is defined by

$$F_{X_1,\dots,X_n}(x_1, x_2, \dots, x_n) = \Pr(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n). \tag{1.3}$$

For a vector random variable $\vec{X}$ with components $X_1, \dots, X_n$, or a complex random variable X with real and imaginary parts X_1, X_2, the distribution function is also defined by (1.3). We will often leave out the subscript $X_1, \dots, X_n$ in such expressions if the meaning is clear from the context. Note that the expression $(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n)$ is an event and the corresponding probability is nondecreasing in each argument x_i. Also the distribution function of any subset of random variables is obtained by setting the other arguments to $+\infty$. For example, the distribution of a single variable (often called a marginal distribution) is given by

$$F_{X_i}(x_i) = F_{X_1,\dots,X_{i-1},X_i,X_{i+1},\dots X_n}(\infty, \dots, \infty, x_i, \infty, \dots, \infty).$$

The joint probability density $f(x_1, \dots, x_n)$, if it exists, is given by the derivative $\frac{\partial^n F(x_1,\dots,x_n)}{\partial x_1 \partial x_2 \cdots \partial x_n}$. Similarly, the joint PMF is given by $\Pr(X_1 = x_1, \dots, X_n = x_n)$.

The random variables $X_1, \ldots, X_n$ are said to be *independent* if, for all $x_1, \ldots, x_n$,

$$F(x_1, \ldots, x_n) = \prod_{i=1}^{n} \Pr(X_i \leq x_i). \tag{1.4}$$

Another way to state this is that $X_1, \ldots, X_n$ are independent if the events $X_i \leq x_i$ for $1 \leq i \leq n$ are independent for all choices of $x_1, \ldots, x_n$. Similarly, events $A_1, \ldots, A_k$ and random variables $X_1, \ldots, X_n$ are independent if the events $A_1, \ldots, A_k, X_1 \leq x_1, \ldots, X_n \leq x_n$ are independent for all choices of $x_1, \ldots, x_n$. If the density or mass function exists, (1.4) is equivalent to a product form for the density or mass function. A set of random variables is said to be pairwise independent if each pair of random variables in the set is independent. As shown in Exercise 1.9, the pairwise independence does not imply that the entire set is independent.

1.5 Expectations

The *expected value* (or the mean) of a continuous or discrete random variable respectively is defined to be

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx; \quad E[X] = \sum_x x P_X(x). \tag{1.5}$$

We shall denote the expected value of a random variable both by $E[X]$ and by $\overline{X}$. In order to avoid rewriting things for discrete, continuous, and mixed cases, we shall generally express the expected valued as a Stieltjes integral,

$$E[X] = \overline{X} = \int_{-\infty}^{\infty} x\, dF_X(x). \tag{1.6}$$

The Stieltjes integral can be defined by

$$\int_{-\infty}^{\infty} g(x)\, dF(x) = \lim_{\delta \to 0} \sum_{n=-\infty}^{\infty} \sup_{n\delta < x \leq n\delta + \delta} g(x)[F(n\delta + \delta) - F(n\delta)] \tag{1.7}$$

The integral is said to exist only if the limit exists and if the limit remains the same if sup[3] above is replaced by inf. It can be seen that if a density exists, $F(n\delta + \delta) - F(n\delta) \approx \delta f(n\delta)$, so that, aside from limiting questions, (1.7) is the same as $\int g(x) f(x)\, dx$. For our purposes, we shall not often be concerned with peculiar functions for which these limit questions are serious. Thus, we will regard (1.6) mostly as shorthand for the expressions in (1.5). One can also view $dF_X(x)$ as standing for $(dF_x(x)/dx)\, dx$ or $f_X(x)\, dx$. In this view, $f_X(x)$ should be extended to include impulse functions to deal with the discrete case. The integral in (1.7) (and similarly the integral and sum in (1.5)) are said to exist only if both the integral from

[3] The *sup*, or *supremum*, of a set of numbers is the smallest number greater than or equal to all members of the set. It is essentially the maximum of the set, but takes care of situations where the max doesn't exist. For example the sup of real numbers x satisfying $x < 2$ is 2, whereas there is no maximum x less than 2. The *inf*, or *infimum*, is defined similarly as the largest number less than or equal to all members of the set. It is essentially the minimum of the set.

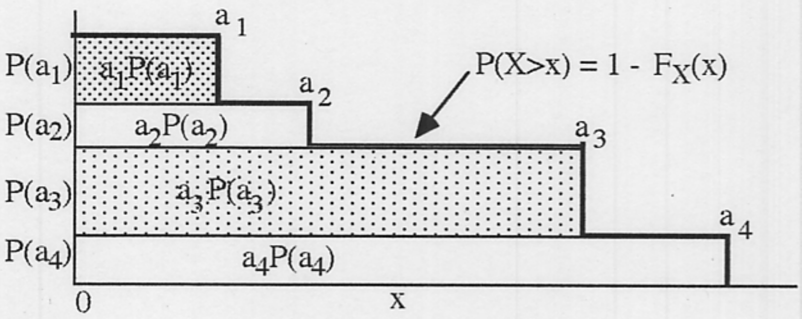

Figure 1.3: For this example, X takes on four possible values, $a_1 < a_2 < a_3 < a_4$, with $a_1 \geq 0$. Thus $\Pr(X > x) = 1$ for all $x \leq a_1$. For $a_1 < x \leq a_2$, $\Pr(X > x) = 1 - P_X(a_1)$, and $\Pr(X > x)$ has similar drops as x reaches a_2, a_3, and a_4. $E[X]$, from (1.2), is $\sum_i a_i P_X(a_i)$, which is the sum of the rectangles in the figure. This is also the area under the curve $1 - F_X(x)$, verifying (1.8).

0 to ∞ and also the integral from $-\infty$ to 0 exist. Thus $E[X]$ exists if and only if $E[|X|]$ exists. This is illustrated later in Example 1.6.

For a non-negative random variable (i.e., a random variable for which $F_X(x) = 0$ for $x < 0$), the expectation can be found in a particularly simple way, namely as the integral of the complementary distribution function, where the *complementary distribution function* of a random variable is defined as $\Pr(X > x) = 1 - F_X(x)$.

$$E[X] = \int_0^\infty [1 - F_X(x)]\, dx. \tag{1.8}$$

This relationship is even more important for conceptual purposes than for computational purposes. To derive it for a discrete random variable, consider sketching the complementary distribution function, $\Pr(X > x) = 1 - F_X(x)$ as in Figure 1.3. The figure shows that $\sum a_i P_X(a_i)$ is equal to the area under the curve $1 - F_X(x)$ from $x = 0$ to ∞.

For an arbitrary distribution, we can visualize quantizing the distribution function, using the above argument, and then passing to the limit of arbitrarily fine quantizing. Since there are no mathematical subtleties in integrating a non-negative decreasing function, many people prefer defining expectation in this way (see Exercise 1.1 for the generalization to arbitrary rather than non-negative random variables). We shall use (1.8) frequently throughout the text.

Random variables are often defined in terms of each other. For example if g is a function from real numbers to real numbers and X is a random variable, then $Y = g(X)$ is the random variable that maps each sample point ω into $g(X(\omega))$. As indicated in Exercise 1.6, one can find the expected value of Y (if it exists) in either of the following ways:

$$E[Y] = \int_{-\infty}^{\infty} y\, dF_Y(y) = \int_{-\infty}^{\infty} g(x)\, dF_X(x). \tag{1.9}$$

Particularly important examples of such expected values are the moments $E[X^n]$ of a random variable X (which is simply the expected value of the random variable X^n) and central moments $E[(X - \overline{X})^n]$ of X where $\overline{X}$ is the mean $E[X]$. The second central moment is called the *variance*, denoted by VAR(X) of σ_X^2. It is given by

$$\mathrm{VAR}(X) = E[(X - \overline{X})^2] = E[X^2] - \overline{X}^2. \tag{1.10}$$

The *standard deviation* of X, σ_X, is the square root of the variance and provides a measure of dispersion of the random variable around the mean. Thus the mean is a rough measure of the typical values for the outcome of the random variable, and σ_X is a measure of how close one expects to come to that typical value. There are other measure of typical value (such as median and the mode) and other measures of dispersion, but mean and standard deviation have a number of special properties that make them important. One of these (see Exercise 1.10 is that $E[X]$ is the value of z that minimizes $E[(X - z)^2]$.

If X and Y are random variables, then the sum $Z = X + Y$ is also a random variable. If X and Y are independent, then the distribution function of Z is given by

$$F_Z(z) = \int_{-\infty}^{\infty} F_X(z - y)\, dF_Y(y) = \int_{-\infty}^{\infty} F_Y(z - x)\, dF_X(x). \tag{1.11}$$

If X and Y both have densities, this can be rewritten as

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y)\, dy = \int_{-\infty}^{\infty} f_Y(z - x) f_X(x)\, dx. \tag{1.12}$$

Eq. (1.12) is the familiar convolution equation from linear systems, and we similarly refer to (1.11) as the convolution of distribution functions (although it has a different functional form from (1.12)). If X and Y are non-negative random variables, then the integrands in (1.11) and (1.12) are non-zero only between 0 and z, so we often use 0 and z as the limits in (1.11) and (1.12). Note, however, that if $\Pr(Y = 0) \neq 0$, then the lower limit must be 0^-, i.e., in terms of (1.11), it must include the jump in $F_Y(y)$ at $y = 0$, and in terms of (1.12), it must include the impulse at $y = 0$. One can avoid confusion of this type by always keeping infinite limits until actually calculating something.

If $X_1, X_2, \ldots, X_n$ are independent random variables, then the distribution of the random variable $S_n = X_1 + X_2 + \ldots + X_n$ can be found by first convolving the distributions of X_1 and X_2 to get the distribution of S_2 and then, for each $i \geq 2$, convolving the distribution of S_i and X_{i+1} to get the distribution of S_{i+1}. The distributions can be convolved in any order to get the same resulting distribution.

Whether or not $X_1, X_2, \ldots, X_n$ are independent, the expected value of $S_n = X_1 + X_2 + \ldots + X_n$ satisfies

$$E[S_n] = E[X_1 + X_2 + \ldots + X_n] = E[X_1] + E[X_2] + \ldots + E[X_n]. \tag{1.13}$$

This says that the expected value of a sum is equal to the sums of the expected values whether or not the random variables are independent (see exercise 1.3). The following example shows how this can be a valuable problem solving aid with appropriate choice of random variables.

Example 1.5 In packet networks, a packet can be crudely modeled as a string of IID binary digits with $P(0) = P(1) = 1/2$. Packets are usually separated from each other by a special bit pattern, 01111110, called a flag. If this special pattern appears within a packet, it could be interpreted as a flag indicating the end of the packet. To prevent this problem, an extra binary digit of value 0 is inserted after each appearance of 011111 in the original string (this can be deleted after reception). Suppose we want to find the expected number of inserted bits in a string of length n. For each position $i \geq 6$ in the original string, define X_i as a random variable whose value is 1 if an insertion occurs after the i^{th} data bit. The total number of insertions is then just the sum of X_i from $i = 6$ to n inclusive. Since $E[X_i] = 2^{-6}$, the expected number of insertions is $(n-5)2^{-6}$. Note that the positions in which the insertions occur are highly dependent, and the problem would be quite difficult if one didn't use (1.13) to avoid worrying about the dependence. If the random variables $X_1, \ldots, X_n$ are independent, then, as shown in exercises 1.3 and 1.7, the variance of $S_n = X_1 + \cdots + X_n$ is given by

$$\sigma^2_{S_n} = \sum_{i=1}^{n} \sigma^2_{X_i} \tag{1.14}$$

If $X_1, \ldots, X_n$ are also identically distributed, then $\sigma^2_{S_n} = n\sigma^2_{X_i}$ and the standard deviation of S_n is $\sigma_{S_n} = \sqrt{n}\sigma_{X_i}$. It is important to remember that the standard deviation of S_n increases with n, although only with the square root of n. Similarly, if $X_1, \ldots, X_n$ are independent, exercise 1.3 shows that

$$E[\prod_{i=1}^{n} X_i] = \prod_{i=1}^{n} E[X_i] \tag{1.15}$$

1.6 Transforms

The *moment generating function* for a random variable X is given by

$$g_X(r) = E[e^{rX}] = \int_{-\infty}^{\infty} e^{rx}\, dF_X(x) \tag{1.16}$$

In (1.16), we can view $g_X(r)$ as a function of a complex variable r. If r is pure imaginary, then the magnitude of e^{rx} is 1 for all x, and the magnitude of $g_X(r)$ is at most one. For values of r with a positive real part, $g_X(r)$ only exists if $1 - F_X(x)$ approaches 0 at least exponentially as $x \to \infty$. Similarly, for values of r with a negative real part, gX(r) exists only if FX(x) approaches 0 at least exponentially as $x \to -\infty$. If $g_X(r)$ exists in a region of real r around 0, then derivatives of all orders also exist, given by

$$\frac{\partial^n g_X(r)}{\partial r^n} = \int_{-\infty}^{\infty} x^n e^{rx} dF_X(x) \quad ; \qquad \left.\frac{\partial^n g_X(r)}{\partial r^n}\right|_{r=0} = e[X^n] \tag{1.17}$$

This shows that finding the moment generating function often provides a convenient way to calculate the moments of a random variable. Another convenient feature of moment generating functions is their use in dealing with sums of independent random variables. For example, suppose $S = X_1 + X_2 + \ldots + X_n$. Then

$$g_S(r) = E\left[e^{rS}\right] = E\left[\exp\left(\sum_{i=1}^{n} rX_i\right)\right] = E\left[\prod_{i=1}^{n} \exp(rX_i)\right] = \prod_{i=1}^{n} g_{X_i}(r) \qquad (1.18)$$

In the last step here, we have used (1.15). There are many other similar types of transforms. For example, if we replace e^r with z, we get the z transform; this is mainly useful for integer valued random variables, but if one transform can be evaluated, the other can be found immediately. If we use $j\omega$ in place of r, where $j = \sqrt{-1}$ and ω is real, we get the characteristic function; it is the inverse Fourier transform of the density function of X and (as pointed out before), it always exists. Finally, if we use $-s$, viewed as a complex variable, in place of r, we get the two sided Laplace transform of the density of the random variable. Note that for all of these transforms, multiplication in the transform domain corresponds to convolution of the distribution functions or densities, and summation of independent random variables. It is the simplicity of taking products of transforms that make transforms so useful in probability theory. We will use transforms sparingly here, since, along with simplifying computational work, they frequently obscure underlying probabilistic insights.

1.7 Weak Law of Large Numbers

The laws of large numbers are a collection of results in probability that describe the behavior of the arithmetic average of n random variables, for n large. Under fairly general assumptions, the standard deviation of the average goes to 0 with increasing n, and, in various ways, depending on the assumptions, the sample average approaches the mean. These results are central to the study of stochastic processes because they allow us to relate time averages (i.e., the average over time of individual sample paths) to ensemble averages (i.e., the mean of the value of the process at a given time). In this and the next section, we discuss two of these results, the weak and the strong law of large numbers for independent identically distributed random variables. The strong law requires considerable patience to understand, but it will be used frequently throughout the text. We start with some basic inequalities which are useful in their own right and which lead us directly to the weak law of large numbers.

1.7.1 Basic Inequalities

The Markov inequality states that if a non-negative random variable Y has a mean $E[Y]$, then the probability that the outcome exceeds any given number y satisfies

$$P(Y \geq y) \leq \frac{E[Y]}{y} \quad \text{Markov Inequality} \qquad (1.19)$$

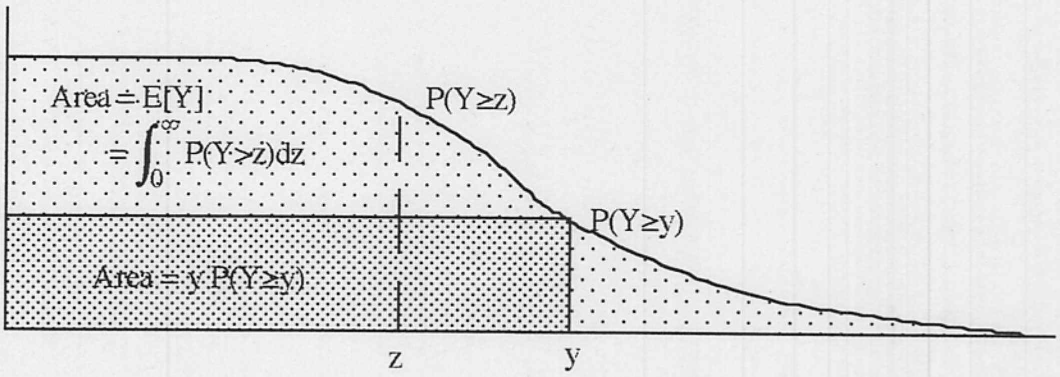

Figure 1.4: Demonstration that $yP(Y \geq y) \leq E[Y]$.

Figure 1.4 derives this result using the fact (see figure 1.3) that the mean of a non-negative random variable is the integral of its complementary distribution function, i.e., of the area under the curve $P(Y > z)$ or equivalently under the curve $P(Y \geq z)$.

As an example of this inequality, assume that the average height of a population of people is 1.6 meters. Then the Markov inequality states that at most one half of the population have a height exceeding 3.2 meters. We see from the example that this inequality is typically very weak. However, for any $y > 0$, we can consider a random variable that takes on the value y with probability ϵ and the value 0 with probability $1-\epsilon$; this random variable satisfies the Markov inequality at the point y with equality. Another application of figure 1.4 is the observation that, for any given non-negative random variable with finite mean, (i.e., with finite area under the curve $P(Y \geq y)$),

$$\lim_{y\to\infty} y\,P(Y \geq y) = 0 \tag{1.20}$$

This is shown in exercise 1.18 and will be useful shortly in the proof of theorem 1. Next, let Z be an arbitrary random variable with finite mean $E[Z]$ and finite variance σ_Z^2, and define Y as the non-negative random variable $Y = (Z - E[Z])^2$. Thus $E[Y] = \sigma_Z^2$. Applying (1.19),

$$P\left((Z - E[Z])^2 \geq y\right) \leq \frac{\sigma_Z^2}{y}$$

Replacing y with ϵ^2 (for any $\epsilon > 0$) and noting that the event $(Z - E[Z])^2 \geq \epsilon^2$ is the same as $|Z - E[Z]| \geq \epsilon$, this becomes the well known Chebyshev inequality,

$$P(|Z - E[Z]| \geq \epsilon) \leq \frac{\sigma_Z^2}{\epsilon^2} \quad \text{Chebyshev inequality} \tag{1.21}$$

Note that the Markov inequality bounds just the upper tail of the distribution function and applies only to non-negative random variables, whereas the Chebyshev inequality bounds both tails of the distribution function. The more important difference, however, is that the Chebyshev bound goes to zero inversely with the square of the distance from the mean, whereas the Markov bound goes to zero inversely with the distance from 0 (and thus asymptotically with distance from the mean).

There is another variant of the Markov inequality, known as an exponential bound or a Chernoff bound, in which the bound goes to 0 exponentially with distance from the mean.

Let $Y = \exp(rZ)$ for some arbitrary random variable Z that has a moment generating function, $g_Z(r) = E\exp(rZ)]$ over some open interval of real values of r including $r = 0$. Then, for r in that interval, (1.19) becomes

$$P(\exp(rZ \geq y) \leq \frac{g_Z(r)}{y}$$

Letting $y = \exp(ra)$ for some constant a, we have the two inequalities,

$$P(Z \geq a) \leq g_Z(r)\exp(-ra); \text{ for } r \geq 0. \quad \text{(Exponential bound)} \tag{1.22}$$

$$P(Z \geq a) \leq g_Z(r)\exp(-ra); \quad \text{for } r \leq 0. \quad \text{(Exponential bound)} \tag{1.23}$$

These bounds can be optimized over r to get the strongest bound; the bound in (1.22), however, is exponentially decreasing in a for fixed $r > 0$, and the bound in (1.23) is exponentially decreasing in $-a$ for fixed $r < 0$. These bounds will be used extensively in Chapter ?? and are useful in information theory, detection theory, and random walk theory.

1.7.2 Weak Law assuming a Finite Variance

Let $X_1, X_2, \ldots, X_n$ be IID random variables with a finite mean $\overline{X}$ and finite variance σ_X^2, let $S_n = X_1 + \ldots + X_n$, and consider the sample average S_n/n. We saw in (1.14) that $\sigma_{S_n}^2 = n\sigma_X^2$. Thus the variance of S_n/n is

$$\text{VAR}\,(S_n/s) = E\left[\left(\frac{S_n - n\overline{X}}{n}\right)^2\right] = \frac{1}{n^2}E\left[\left(S_n - n\overline{X}\right)^2\right] = \frac{\sigma^2}{n} \tag{1.24}$$

This says that the standard deviation of the sample average S_n/n is $\sigma/\sqrt{n}$. Thus, the standard deviation of S_n approaches ∞ as $n \to \infty$, whereas the standard deviation of S_n/n approaches 0 as $n \to \infty$. Since $\lim_{n\to\infty} E[(S_n/n - \overline{X})^2] = 0$, S_n/n is said to *converge to $E[X]$ in the mean square sense.*

Applying the Chebyshev inequality in (1.21) to the sample average S_n/n, we have

$$P\Big(\Big|\frac{S_n}{n} - \overline{X}\Big| \geq \epsilon\Big) \leq \frac{\sigma^2}{n\epsilon^2} \tag{1.25}$$

For any $\epsilon > 0$, we can pass to the limit $n \to \infty$, getting

$$\lim_{n\to\infty} P\left(\Big|\frac{S_n}{n} - \overline{X}\Big| \geq \epsilon\right) = 0 \tag{1.26}$$

Eqs. (1.25) and (1.26) are alternate forms of the weak law of large numbers for IID random variables $\{X_i\}$ that have a mean and variance. For any $\epsilon > 0$, (1.26) says that $\lim_{n\to\infty} P(A_n) = 0$ where A_n is the event that the sample average S_n/n differs from the true mean by more than ϵ. This means (see Figure 1.5) that the distribution function of S_n/n approaches a unit step function with the step at $\overline{X}$ as $n \to \infty$. Because of (1.26), S_n/n is said to *converge to $E[X]$ in probability.*

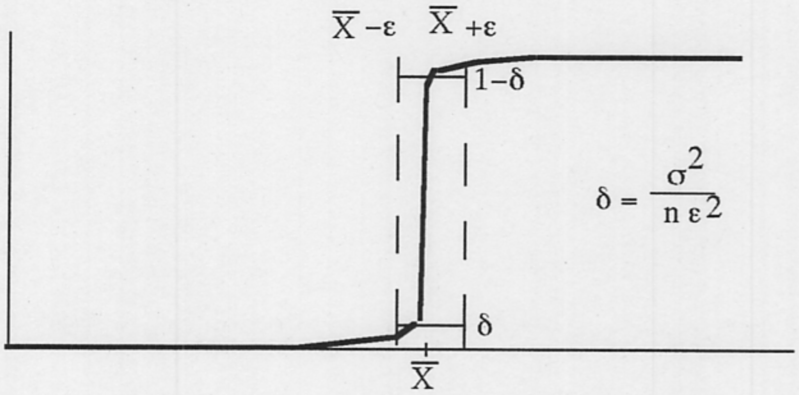

Figure 1.5: Approximation of a distribution function by a step function. Note the rectangle of width 2ϵ and height $1 - 2\delta$ in the figure. Eq. (1.25) asserts that the distribution function of S_n/n enters this rectangle from below and exits from above, thus approximating a unit step.

To get a better appreciation of (1.25) and (1.26), consider figures 1.6, 1.7, and 1.8. These figures show the distribution functions of S_n, $(S_n - n\overline{X})/n$, and $(S_n - n\overline{X})/\sqrt{n}$ as a function of n for a typical random variable X that is binary with $P(X = 0) = 3/4$, $P(X = 1) = 1/4$. There are two important effects to be observed in figure 1.6. First, $E[S_n]$ equals $n\overline{X}$, which is linear in n. Second, the standard deviation of S_n is $\sqrt{n}\sigma$, which gives rise to a spreading of the distribution with $\sqrt{n}$. In figure 1.7, note that the standard deviation of S_n/n *decreases* as $1/\sqrt{n}$, and note the corresponding compression of the distribution function with increasing n. Note finally that with the normalization of figure 1.8, the distribution function neither compresses or expands, but simply becomes smoother with increasing n.

1.7.3 The Central Limit Theorem

The law of large numbers does *not* say that S_n is close to n with high probability as $n \to \infty$. In fact, the standard deviation of S_n is $\sigma\sqrt{n}$, which increases with n. It can be seen that the standard deviation of $S_n/\sqrt{n}$ is σ and does not vary with n. In fact, the celebrated central limit theorem states that if $\sigma^2 < \infty$, then

$$\lim_{n\to\infty} \left[P\left(\frac{S_n - n\overline{X}}{\sqrt{n}\,\sigma} \le y \right) \right] = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2} \right) dx \qquad (1.27)$$

Note that the random variable $(S_n - n\overline{X})/(\sqrt{n}\,\sigma)$ in (1.27) has mean 0 and variance 1 for all n. The central limit theorem says that this random variable tends to the distribution on the right hand side of (1.27) as $n \to \infty$. This is the normal distribution, or normalized Gaussian distribution. The theorem is illustrated by figure 1.8. The difference between the right side of (1.27) and the term in brackets on the left goes to 0 as $1/\sqrt{n}$ if X has a third moment and it goes to zero, but perhaps slowly, if X does not have a third moment. For large negative y, the right side of (1.27) is very close to 0, so, for moderate values of n, the ratio of the left to right side of (1.27) might be (and typically is) far from 1, even though

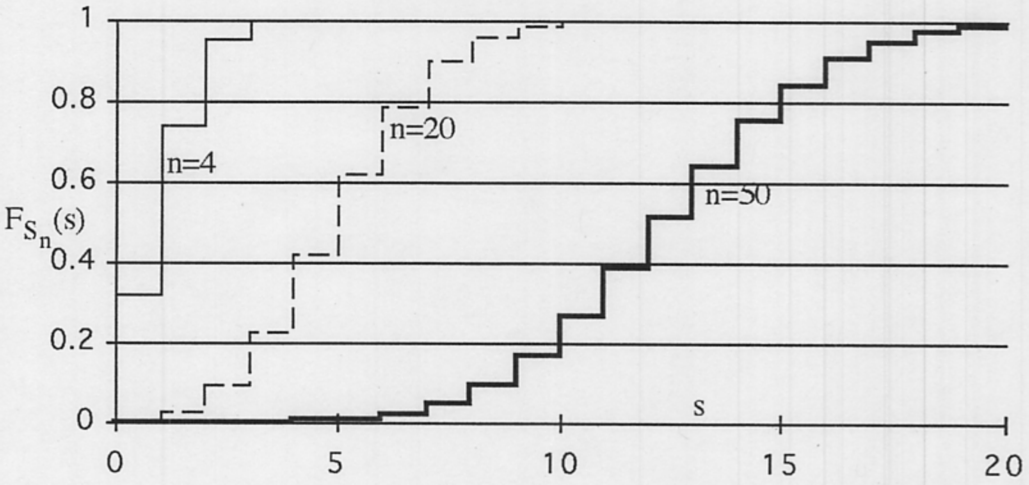

Figure 1.6: The distribution function of S_n as a function of n. In the figure, S_n is the sum of n binary random variables with $P(1) = 1/4$, $P(0) = 3/4$.

the difference between left and right is very small. For large values of y, both sides of (1.27) are close to 1, so that again (1.27) does not yield a very good approximation for reasonable values of n. It is for this reason that the word central appears in the name "central limit theorem."

The central limit theorem (CLT) helps explain why Gaussian random variables play such a central role in probability theory. In fact, many of the cookbook formulas of elementary statistics are based on the tacit assumption that the underlying variables are Gaussian, and the CLT helps explain why these formulas often give reasonable results.

One should be careful to avoid reading more into the CLT than it says. For example, the normalized sum, $[S_n - n\overline{X}]/(\sqrt{n}\sigma)$ need not have a density that is approximately Gaussian (in fact, if the underlying variables are discrete, the normalized sum is also, and does not have a density at all). What is happening is that the normalized sum can have very detailed fine structure; this does not disappear as n increases, but becomes "integrated out" in the distribution function. We will not use the CLT extensively here, and will not prove it (See Feller[4] for a thorough and careful exposition on various forms of the central limit theorem). Giving a proof from first principles is quite tricky; many elementary texts on probability give "heuristic" proofs indicating that the normalized sum has a density that tends to Gaussian (thus indicating that both the heuristics and the manipulations are wrong).

Since the central limit theorem gives such explicit information on the behavior of S_n as $n \to \infty$, one wonders why the law of large numbers should be studied at all. There are three answers—the first is that sometimes one doesn't care about the added information in (1.27), and that the added information obscures some issues of interest. The next is that (1.27) is only meaningful if the variance σ^2 of X is finite, whereas, as we soon show, (1.26) holds whether or not the variance is finite. One might think that variables with infinite variance are of purely academic interest. Unfortunately, it is precisely in renewal theory

[4] Feller, *An Introduction to Probability Theory and its Application*, Vol. I and II, Wiley, 1968 and 1966.

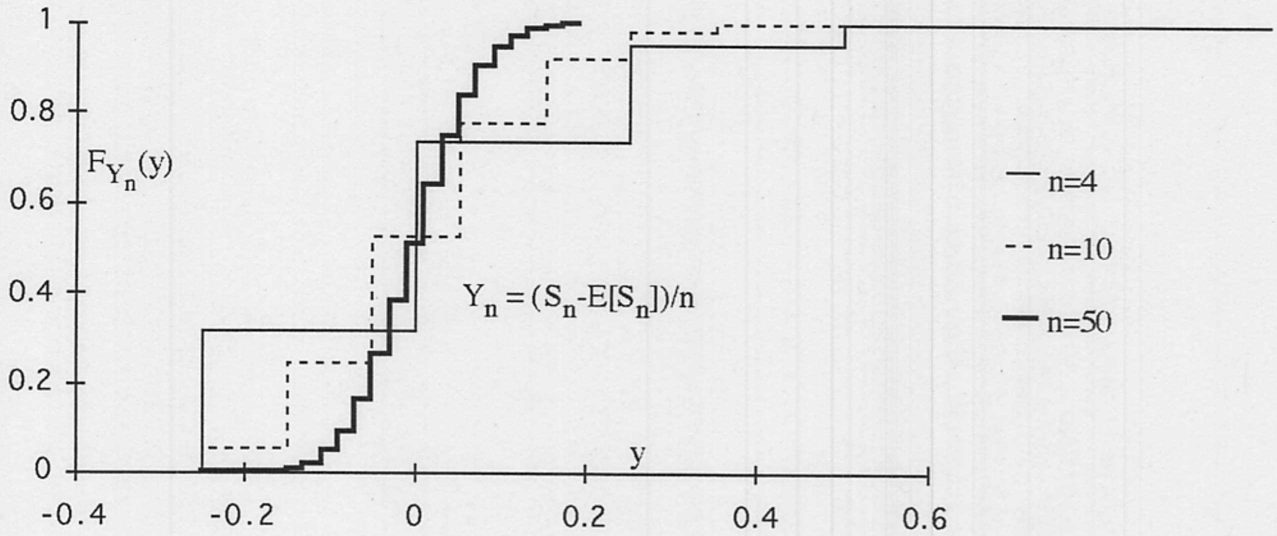

Figure 1.7: The same distribution as figure 1.6, scaled differently to focus on sample average.

that random variables without variances are sometimes important.

The third reason for interest in laws of large numbers is that they hold in many situations more general than that of sums of IID random variables[5]. These more general laws of large numbers can usually be interpreted as saying that some time average (i.e., an average over time of a sample function of a stochastic process) approaches the expected value of the process at a given time. Since expected values only exist within a probability model, but time averages can be evaluated from a sample function of the actual process being modeled, *the relationship between time averages and expected values is often the link between a model and the reality being modeled.*

1.7.4 Relative Frequency and Indicator Functions

We next show that (1.25) and (1.26) can be applied to the relative frequency of an event as well as to the sample average. For any event A, we define I_A, the *indicator function* of A, to be a random variable that has the value 1 for all sample points in A and has the value 0 otherwise. Thus $P(A) = E[I_A]$. Indicator functions are useful because they allow us to apply many of the results we know about random variables to events. For the situation here, we view an experiment containing an event A as being independently repeated n times (i.e., all events in each repetition are independent of events in each other repetition). Let A_i be the event A in the i^{th} repetition. The relative frequency of A over the n experiments is then the number of times that A occurs divided by n. This is best expressed as

$$\text{relative frequency of } A = \frac{\sum_{i=1}^{n} I_{A_i}}{n} \tag{1.28}$$

[5] Central limit theorems also hold in many of these more general situations, but they usually do not have quite the generality of the laws of large numbers

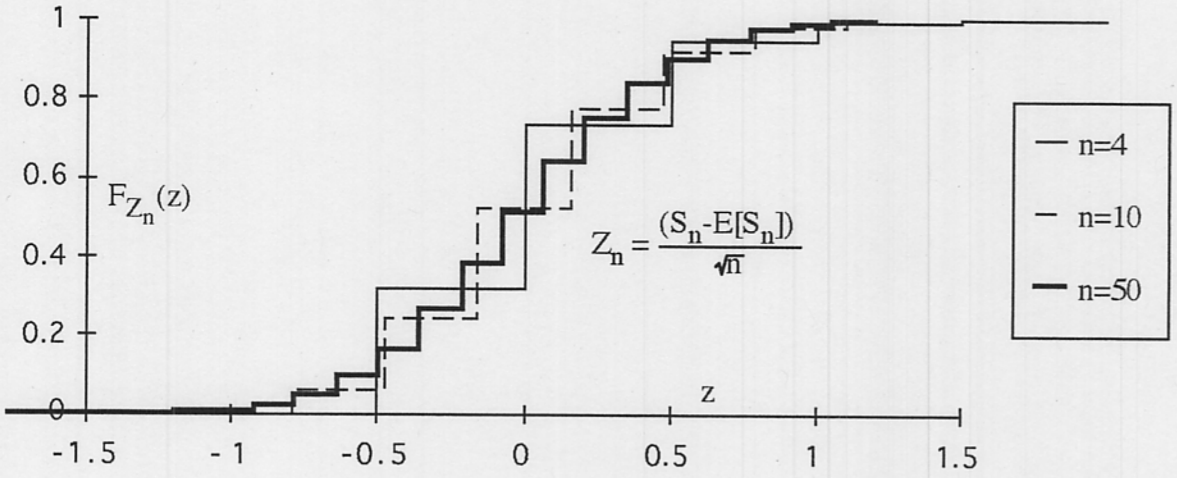

Figure 1.8: The same distribution as figure 1.6 scaled for constant standard deviation.

Thus the relative frequency is the sample average of the indicator functions, and, from (1.25)

$$P(|\text{ relative frequency of } A - P[A] \mid \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2} \tag{1.29}$$

where σ^2 is the variance of I_A, which is $P(A)[1 - P(A)]$.

1.7.5 Digression

One usually motivates the use of probability theory for real world applications by first making an analogy between the probability of an event A (theory) and relative frequency (in the real world). One also makes an analogy between independence (theory) and physical independence of experiments (real world). It then appears that (1.29) justifies the connection between probability and relative frequency.

A sceptic, listening to this argument, will rightfully call "FOUL," because of the somewhat circular nature of the argument. That is, the real world must "act like" the probability model in performing the n independent experiments, and we cannot precisely define what that means for the real world.

For example, we could determine, for a sample of integrated circuits, the relative frequency of those that work correctly. However, different integrated circuits in the sample might be manufactured on different machines, at different times of the day, with different operators, etc. Asserting that these are independent and identically distributed depends on understanding both the physical processes and the theory.

What the law of large numbers does is to relate relative frequency within the theory to the probability of an event *within the theory.* It provides a type of consistency, within the theory, between relative frequency and probability. This is all that can be expected. Theorems establish rigorous results within a model of reality, but cannot prove things about the real world itself. These theorems, however, provide us with the framework to understand and interpret the behavior of the real world. They both allow us to improve our models, and to predict future behavior to the extent that the models reflect reality.

1.7.6 Weak Law with Infinite Variance

We now establish the law of large numbers without assuming a finite variance.

Theorem 1.1 (Weak Law of Large Numbers) Let $S_n = X_1 + \ldots + X_n$ where $X_1, X_2, \ldots$ are IID random variables with a finite mean $E[X]$. Then for any $e > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - E[X]\right| \geq \epsilon\right) = 0 \tag{1.30}$$

Proof:*[6] We use a truncation argument; such arguments are used frequently in dealing with random variables that have infinite variance. Let b be a real number (which we later take to be increasing with n), and for each variable X_i, define a new random variable i (see Figure 1.9) by

$$\begin{aligned} \widetilde{X}_i &= X_i && \text{if } X_i - E[X] \geq b \\ \widetilde{X}_i &= E[X] + b && \text{if } Xi - E[X] \geq b \\ \widetilde{X}_i &= E[X] - b && \text{if } X_i - E[X] \leq -b \end{aligned} \tag{1.31}$$

The variables $\widetilde{X}_i$ are IID and we let $E[\widetilde{X}]$ be the mean of $\widetilde{X}_i$. As shown in exercise 1.10, the variance of $\widetilde{X}$ can be upper bounded by the second moment around any other value, so $\sigma^2_{\widetilde{X}} \leq E[(\widetilde{X} - E[X])^2]$. This can be further upper bounded by

$$\sigma^2_{\widetilde{X}} \leq E\left[\left(\widetilde{X} - E[X]\right)^2\right] = \int_{-\infty}^{\infty} (x - E[X]^2 dF_{\widetilde{X}}(x) \leq b \int_{-\infty}^{\infty} |x - E[X]| dF_{\widetilde{X}}(x)$$

The last inequality follows since $|x - E[X]| \leq b$ over the range of $\widetilde{X}$. We next use the fact that $F_{\widetilde{X}}(x) = F_X(x)$ for $E[X] - b < x < E[X] + b$ to upper bound the final integral.

$$\sigma^2_{\widetilde{X}} \leq b \int_{-\infty}^{\infty} |x - E[X]| dF_X(x) = b\alpha \text{ where } \alpha = \int_{-\infty}^{\infty} |x - E[X]| dF_X(x) \tag{1.32}$$

The quantity α in (1.32) is the mean of $|X - E[X]|$ and must exist since we assume that X has a mean (see example 1.6). Now, letting $\widetilde{S}_n = \widetilde{X}_1 + \ldots + \widetilde{X}_n$, (and using $\epsilon/2$ in place of ϵ), ((1.25) becomes

$$P\left(\left|\frac{\widetilde{S}_n}{n} - E[\widetilde{X}]\right| \geq \frac{\epsilon}{2}\right) \leq \frac{4\sigma^2_{\widetilde{X}}}{n\epsilon^2} \leq \frac{4b\alpha}{n\epsilon^2}$$

[6] Proofs and sections marked with an asterisk, while instructive, can be omitted without loss of continuity.

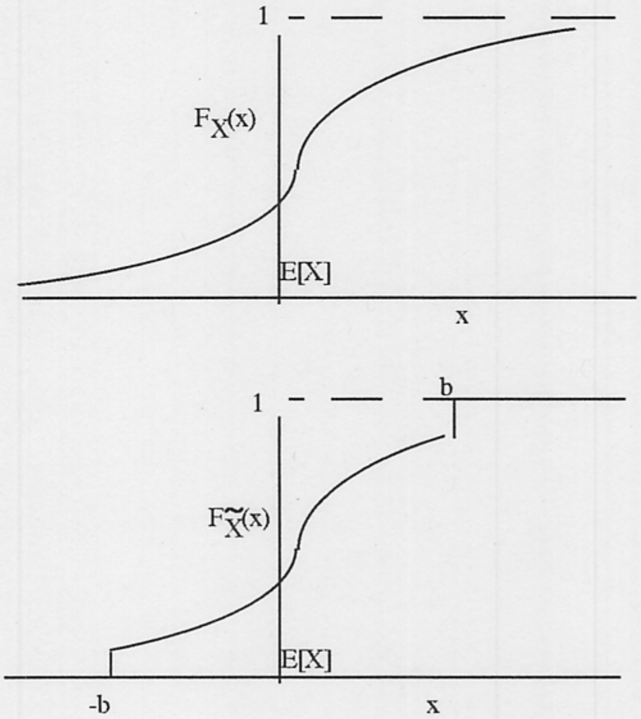

Figure 1.9: Truncated variable $\widetilde{X}$.

As b increases, $E[\widetilde{X}]$ approaches $E[X]$. Thus for sufficiently large b, $|E[\widetilde{X}] - E[X]| < \epsilon/2$ and

$$P\left(\left|\frac{\widetilde{S}_n}{n} - E[X]\right| \geq \epsilon\right) \leq \frac{4b\alpha}{n\epsilon^2} \tag{1.33}$$

Now $\widetilde{S}_n$ and S_n have the same value for sample points where $|X_i - E[X]| \leq b$ for all i, $1 \leq i \geq n$. Thus, using the union bound (which says that the probability of a union of events is less than or equal to the sum of the probabilities of the individual events),

$$P(\widetilde{S}_n \neq S_n) \leq n\,P(|X - E[X]| > b) \tag{1.34}$$

The event $\{|(S_n/n) - E[X]| \geq \epsilon\}$ can only occur if either $|(\widetilde{S}_n/n) - E[X]| \geq \epsilon$ or if $\widetilde{S}_n \neq S_n$. Thus, combining (1.33) and (1.34), and letting $\delta = b/n$, we have

$$P\left(\left|\frac{S_n}{n} - E[X]\right| \geq \epsilon\right) \leq \frac{4\delta\alpha}{\epsilon^2} + \frac{1}{\delta}\left[\delta n\,P\left(|X - E[X]| > \delta n\right)\right] \tag{1.35}$$

Since (1.33) and (1.34) are valid for arbitrary n and sufficiently large $b > 0$, (1.35) is valid for arbitrary $d > 0$ and sufficiently large n. For any given $\epsilon > 0$, we now choose δ to make the first term on the right of (1.35) as small as desired. From (1.20), the final term in brackets in (1.35) can be made arbitrarily small by choosing n large enough, and thus the right hand side of (1.35) can be made arbitrarily small by choosing n large enough, thus completing the proof.

Example 1.6 The Cauchy random variable Z has the probability density $f_Z(z) = 1/[\pi(1+z^2)]$. The mean of Z does not exist, and Z has the very peculiar property that $[Z_1 + Z_2 + \ldots + Z_n]/n$ has the same density as Z for all n. Thus the law of large numbers does not hold for the Cauchy distribution, which is not surprising since the mean doesn't exist. Recall that the mean of a random variable exists only if

$$\begin{aligned} \int_{-\infty}^{0} x dF_X(x) \quad &> \quad -\infty \quad \text{and} \quad \int_{0}^{\infty} x dF_X(x) < \infty, \quad \text{or equivalenty,} \\ \int_{-infty}^{\infty} |x| dF_X(x) \quad &< \quad \infty \end{aligned} \tag{1.36}$$

From symmetry, we note that, for the Cauchy distribution, the integral $\int_{z=-b}^{b} dF_Z(z)$ is zero for all b, and thus the integral exists in the Cauchy principal value sense, but not in the ordinary sense of (1.36). In this text, the existence of integrals always refers to existence in the ordinary rather than Cauchy principal value sense.

1.8 Strong Law of Large Numbers

We next discuss the strong law of large numbers. We will not prove this result here, but will prove a slightly weaker form of it after discussing martingales in chapter ??.

Theorem 1.2 (Strong Law of Large Numbers (Version 1)) Let $S_n = X_1 + \ldots + X_n$ where $X_1, X_2, \ldots$ are IID random variables with a finite mean $\overline{X}$. Then for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\sup_{m\geq n} \left|\frac{S_m}{m} - \overline{X}\right| > \epsilon\right) = 0 \tag{1.37}$$

The notation *sup* above stands for supremum (see footnote 2). The supremum of a set $\{Y_i; i \geq 1\}$ of random variables is a random variable. For each sample point ω in the sample space, $Y_i(\omega)$ is the sample value of Y_i for that ω, and $\sup_i Y_i(\omega)$ is the supremum of those sample values. Thus, $\sup_i Y_i$ is the random variable that maps each sample point ω into $\sup_i Y_i(\omega)$.

We can then interpret the event, $\{\sup_{m\geq n} |S_m/m - \overline{X}| > \epsilon\}$ as the event that $|S_m/m - \overline{X}|$ exceeds ϵ for at least one value of $m \geq n$. The theorem states that the probability of this event approaches 0 with increasing n. Thus, the theorem states that, for large n, it is unlikely that the sequence $|S_m/m - \overline{X}|$ for $m \geq n$ ever exceeds ϵ. An alternative way to express this limit is the statement that for each $\epsilon > 0$ and each $\delta > 0$, there is an integer $n(\epsilon, \delta)$ such that

$$P\left(\sum_{m\geq n(\epsilon,\delta} \left|\frac{S_m}{m} - \overline{X}\right| > \epsilon\right) \leq \delta \tag{1.38}$$

The weak law states that it is increasingly unlikely (with increasing n) for S_n/n to deviate from the mean by any given $\epsilon > 0$. The strong law states that the *sequence* $S_1, S_2/2, S_3/3, \ldots$

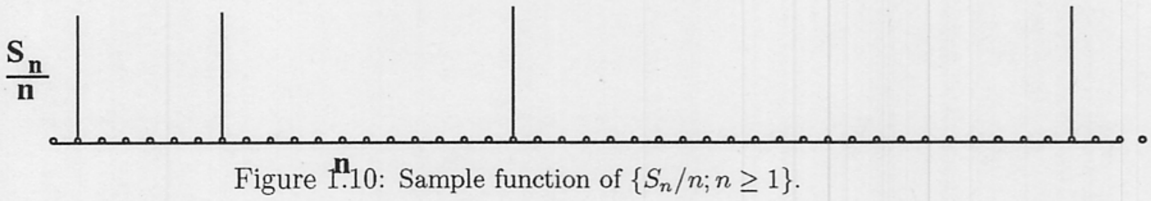

Figure 1.10: Sample function of $\{S_n/n; n \geq 1\}$.

has the property that it is increasingly unlikely for *any* term in the sequence beyond n to deviate by more than ϵ from the mean. The following example illustrates the difference between the strong and weak law. Note that in this example, the random variables $X_i; i \geq 1$ are neither independent nor equally distributed. Since both the strong and weak law hold for IID variables, we cannot illustrate the difference for the IID case.

Example 1.7 Let $\{S_n; n \geq 1\}$ be a sequence of independent random variables for which S_n has value n with probability $1/n$ and value 0 otherwise (see Figure 1.10). Then $E[S_n] = 1$ and $E[S_n/n] = 1/n$. If one wishes, S_n can be interpreted as a sum of dependent random variables, $X_1 + \ldots + X_n$ where $X_1 = S_1$ and $X_i = S_i - S_{i-1}$, but the argument really just concerns the sequence $\{S_n; n \geq 1\}$. For any ϵ, $0 < \epsilon < 1$, the Markov inequality shows that $P(S_n/n > \epsilon) \geq (n\epsilon)^{-1}$, so $\lim_{n\to\infty} P(S_n/n > \epsilon) = 0$. Next, we see that $\sup_{m\geq n} S_m/m$ takes on only the values 0 and 1, and takes on the value 0 only if $S_m/m = 0$ for all $m \geq n$. This is an event of probability $\Pi_{m\geq n}(1 - 1/m)$, which (see exercise 1.19) can be seen to be zero for all n. Thus

$$\lim_{n\to\infty} P\left(\frac{S_n}{n} > \epsilon\right) = 0 \text{ but } \lim_{n\to\infty} P\left(\sup_{m\geq n} \frac{S_m}{m} > \epsilon\right) = 1$$

This says that S_n/n is likely to be small for large n (and in fact likely to be 0), but that if one waits long enough beyond n, a sample sequence has a value exceeding ϵ (and in fact equal to 1) with probability one.

One sees from this example that the strong law is saying something about a sample outcome of an infinite sequence of random variables. If one views X_i as a time sequence, then the sample output from the sequence of sample averages S_n/n can be viewed as a sequence of more and more elaborate attempts to estimate the mean. There is clearly some advantage to being able to say that this sequence of attempts not only gets close to the mean with high probability but also stays close to the mean.

Despite the above rationalization, the difference between the strong and weak law almost appears to be mathematical nit picking. On the other hand, we shall discover, as we use these results, that the strong law is often much easier to use than the weak law. The useful form of the strong law, however, is the following theorem. The statement of this theorem is deceptively simple, and it will take some care to understand what the theorem is saying.

Theorem 1.3 (Strong Law of Large Numbers (Version 2)) Let $S_n = X_1 + \ldots + X_n$ where $X_1, X_2, \ldots$ are IID random variables with a finite mean $\overline{X}$. Then with probability 1,

$$\lim_{n\to\infty} \frac{S_n}{n} = \overline{X} \tag{1.39}$$

For each sample point ω, $S_n(\omega)/n$ is a sequence of real numbers that might or might not have a limit. If this limit exists for all sample points, then $\lim_{n\to\infty} S_n/n$ is a random variable that maps each sample point ω into $\lim_{n\to\infty} S_n(\omega)/n$. Usually this limit does not exist for all sample points, but the theorem implicitly asserts that the limit does exist for all sample points except a set of probability 0. Thus $\lim_{n\to\infty} S_n/n$ is still regarded as a random variable. The theorem asserts not only that $\lim_{n\to\infty} S_n(\omega)/n$ exists for all sample points except a set of zero probability, but also asserts that the limit is equal to $\overline{X}$ for all sample points except a set of probability 0. A sequence of random variables S_n/n that converges in the sense of (1.39) is said to *converge with probability* 1.

Example 1.8 Suppose the X_i are Bernoulli with equiprobable ones and zeros. Then $\overline{X} = 1/2$. We can easily construct sequences for which the sample average is not 1/2; for example the sequence of all zeros, the sequence of all ones, sequences with 1/3 zeros and 2/3 ones, and so forth. The theorem says, however, that collectively those sequences have zero probability.

Proof of theorem 3: We assume theorem 2 (which we do not prove until chap. 7) in order to prove theorem 3. Consider the event illustrated in figure 1.11 in which tighter and tighter bounds are placed on successive elements of the sequence $\{S_n/n; n \geq 1\}$. In particular, for some increasing set of positive integers $n_1, n_2, \ldots$, we consider the bound $|S_n/n - \overline{X}| \geq 2^{-k}$ for $n_k \leq n < n_{k+1}$. For any sample point ω, if $\{S_n(\omega); n \geq 1\}$ satisfies all these constraints, then $\lim_{n\to\infty} S_n(\omega)/n = \overline{X}$. The probability of the complementary set of sample points for which one of these bounds is unsatisfied is given by

$$P\left\{\cup_{k\geq 1}\left[\cup_{n_k\leq n<n_{k+1}}\left(\left|\frac{S_n}{n} - \overline{X}\right| > 2^{-k}\right)\right]\right\} = P\left\{\cup_{k\geq 1}\left[\cup_{n_k\leq n}\left(\left|\frac{S_n}{n} - \overline{X}\right| > 2^{-k}\right)\right]\right\} \tag{1.40}$$

$$\leq \sum_{k=1}^{\infty} P\left[\cup_{n\geq n_k}\left(\left|\frac{S_n}{n} - \overline{X}\right| > 2^{-k}\right)\right] \tag{1.41}$$

$$= \sum_{k=1}^{\infty} P\left(\sup_{n\geq n_k}\left|\frac{S_n}{n} - \overline{X}\right| > 2^{-k}\right) \tag{1.42}$$

The first equality is most easily visualized in figure 1.11; if $|S_n/n - \overline{X}| > 2^{-k}$ for one value of k, then $|S_n/n - \overline{X}| > 2^{-k'}$ for all $k' \geq k$. In going from (1.40) to (1.41), we have used the union bound; this says that the probability of a union of events is less than or equal to the sum of the probabilities of the individual events. Finally (1.42) follows because the supremum of a sequence exceeds 2^{-k} if and only if one of the elements exceeds 2^{-k}.

From (1.38), for any ϵ, $d > 0$, there is an $n(\epsilon,\delta)$ such that $P(\sup_{n\geq n(\epsilon,\delta)} |S_n/n - \overline{X}| > \epsilon) \leq \delta$. For given δ_0, we then choose n_k in the bound above as $n_k = n(2^{-k}, \delta_0 2^{-k})$, i.e., so that $P(\sup_{n\geq n_k} |S_m/m - \overline{X}| > 2^{-k}) \leq \delta_0 2^{-k}$. Substituting this in (1.42), we have

$$\sum_{k=1}^{\infty} P\left(\sup_{n\geq n_k}\left|\frac{S_n}{n} - \overline{X}\right| > 2^{-k}\right) \leq \sum_{k=1}^{\infty} \delta_0 2^{-k} = \delta_0 \tag{1.43}$$

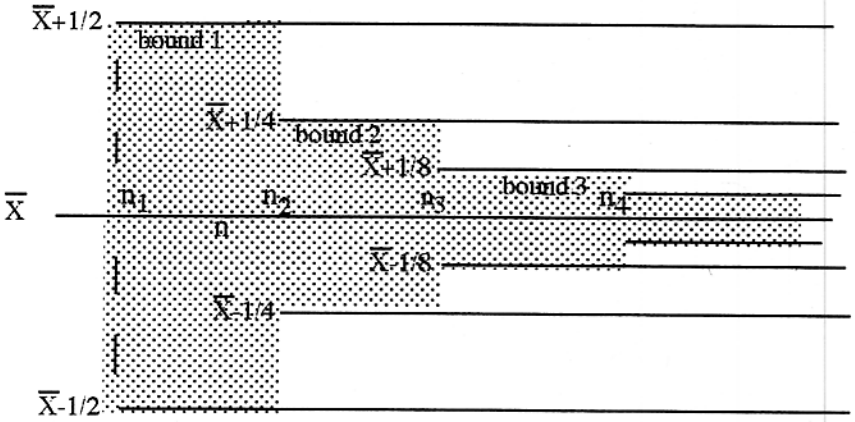

Figure 1.11: Illustration of the union of events in (1.40); the k^{th} sub-event in (1.40) is the set of sample points for which S_n/n falls outside of the k^{th} bound for some n, $n_k \leq n < n_{k+1}$; i.e., for which $|S_n/n - \overline{X}| > 2^{-k}$ for some $n_k \geq n < n_{k+1}$.

It follows that the set of sample points that do not violate these bounds has probability at least $1 - \delta_0$, and as we have seen, $\lim_{n\to\infty} S_n/n = \overline{X}$ for each of these sample points. Since this is true for any $\delta_0 > 0$, the set of sample points for which $\lim_{n\to\infty} S_n/n = \overline{X}$ must have probability 1, completing the proof.

Note that as δ_0 is decreased, the integers $n_1, n_2, \dots$ become larger, thus enlarging the set of sample points that fall within the given bounds. Thus if $\{S_n(\omega)/n; n \geq 1\}$ converges very slowly to $\overline{X}$ for a given ω, then a very small value of δ_0 is required for $\{S_n(\omega)/n; n \geq 1\}$ to stay within the bounds of figure 1.11.

1.9 Summary

This chapter has provided a brief review of elementary probability theory, starting with the basic ingredients of sample space, events, and probabilities of events, then moving to random variables, and then to laws of large numbers. The emphasis has been on understanding the underlying structure of the field rather than reviewing details and problem solving techniques. The strong law of large numbers requires mathematical maturity, and might be postponed to chapter 3 when it is first used.

There are too many texts on elementary probability to mention here, and most of them serve to give added understanding and background to the material here. [Ros94] and [Dra67] are both quite readable. [Kol50] is of historical interest (and is also readable) as the translation of the 1933 book that first put probability on a firm mathematical basis. [Fel68] is an extended and elegant treatment of elementary material from a mature point of view.

1.9.1 Table of Standard Random Variables

The following table summarizes the properties of some common random variables. If a density or PMF is specified only in a given region, it is assumed to be zero elsewhere.

Name	**Density or PMF**	**Mean**	**Variance**	**Generating functio**
(Continuous rv	$f_X(x))$			
Exponential	$\lambda \exp(-\lambda x); x \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\frac{\lambda}{\lambda - r}$
Erlang	$\frac{\lambda^n x^{n-1} \exp(-\lambda x)}{(n-1)!}; x \geq 0$	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$	$\left(\frac{\lambda}{\lambda - r}\right)^n$
Gaussian	$\frac{\exp(-x^2/2\sigma^2}{\sqrt{2\pi}\,\sigma}$	0	σ^2	$\exp(r^2 - \sigma^2/2)$
Uniform	$\frac{1}{a}; 0 \leq x \leq a$	$\frac{1}{2}$	$\frac{\alpha^2}{12}$	$\frac{\exp(ra)-1}{ra}$
Integer rv	$P_N(n))$			
Bernoulli	$P_N(0) = 1 - p; P_N(1) = p$	$p(1-p)$	$1 - p + pe^r$	
Binomial	$\frac{k!}{n!(k-n)!} p^n (1-p)^{k-n}; 0 \leq n \leq k$	kp	$kp(1-p)$	$[1 - p + pe^r]^k$
Geometric	$(1-p)p^n; n \geq 0$	$\frac{p}{1-p}$	$\frac{p}{(1-p)^2}$ $\frac{1-p}{1-pe^r}$	
Poisson	$\frac{\lambda^n \exp(-\lambda)}{n!}; n \geq 0$	λ	λ	$\exp[\lambda(e^r - 1)]$

1.10 EXERCISES

Exercise 1.1 The text shows that, for a non-negative random variable X with distribution function $F_X(x)$, $E[X] = \int_0^\infty [1 - F_X(x)]dx$.

a) Write this integral as a sum for the special case in which X is a non-negative integer random variable.

b) Generalize the above integral for the case of an arbitrary (rather than non-negative) random variable Y with distribution function $F_Y(y)$; use a graphical argument.

c) Find $E[|Y|]$ by the same type of argument.

d) For what value of α is $E[|Y - \alpha|]$ minimized? Use a graphical argument again.

Exercise 1.2 Let X be a random variable with distribution function $F_X(x)$. Find the distribution function of the following random variables.

a) The maximum of n IID random variables with distribution function $F_X(x)$.

b) The minimum of n IID random variables with distribution $F_X(x)$.

c) The difference of the random variables defined in a) and b); assume X has a density $f_X(x)$.

Exercise 1.3 **a)** Let $X_1, X_2, \ldots, X_n$ be random variables with expected values $\overline{X}_1, \ldots, \overline{X}_n$. Prove that $E[X_1 + \ldots + X_n] = \overline{X}_1 + \ldots + \overline{X}_n$. Do not assume that the random variables are independent.

b) Now assume that $X_1, \ldots, X_n$ are statistically independent and show that the expected value of the product is equal to the product of the expected values.

c) Again assuming that $X_1, \ldots, X_n$ are statistically independent, show that the variance of the sum is equal to the sum of the variances.

Exercise 1.4 Suppose $X_1, X_2, X_3, \ldots$ is a sequence of continuous IID random variables. X_n, for a given $n > 1$, is called a local minimum of the sequence if $X_n \leq X_{n+1}$, $X_n \leq X_{n-1}$. Find the probability that X_n is a local minimum. Hint: No computation is necessary–use symmetry.

Exercise 1.5 Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of independent identically distributed (IID) continuous random variables with the common probability density function $f_X(x)$; note that $P(X{=}\alpha) = 0$ for all α and that $P(X_1{=}X_2) = 0$.

a) Find $P(X_1{\geq}X_2)$ (give a numerical answer, not an expression; no computation is required and a one or two line explanation should be adequate).

b) Find $P(X_1{\leq}X_2; X_1{\leq}X_3)$ (in other words, find the probability that X_1 is the smallest of X_1,X_2, X_3; again, think–don't compute).

c) Let the random variable N be the index of the first rv in the sequence to be less than X_1; that is, $P(N{=}n) = P(X_1{\leq}X_2;\ X_1{\leq}X_3;\ \cdots X_1{\leq}X_{n-1};\ X_1{>}X_n)$. Find $P(N > n)$ as a function of n. Hint: generalize part b.

d) Show that $E[N] = \infty$.

Exercise 1.6 **a)** Assume that X is a discrete random variable taking on values $a_1, a_2, \ldots$, and let $Y = g(X)$. Let $b_i = g(a_i), i{\geq}1$ be the i^{th} value taken on by Y. Show that $E[Y] = \sum_i b_i P_Y(b_i) = \sum_i g(a_i) P_X(a_i)$.

b) Let X be a continuous random variable with density $f_X(x)$ and let g be differentiable and monotonic increasing. Show that $E[Y] = \int y f_Y(y)dy = \int g(x) f_X(x)dx$.

Exercise 1.7 **a)** Show that, for uncorrelated random variables, the expected value of the product is equal to the product of the expected values (X and Y are uncorrelated if $E[(X - \overline{X})(Y - \overline{Y})] = 0$).

b) Show that if X and Y are uncorrelated, then the variance of $X + Y$ is equal to the variance of X plus the variance of Y.

c) Show that if $X_1, \ldots, X_n$ are uncorrelated, the the variance of the sum is equal to the sum of the variances.

d) Show that independent random variables are uncorrelated.

e) Let X, Y be identically distributed ternary valued random variables with the probability assignment $P(X{=}1) = P(X{=}-1) = 1/4; P(X{=}0) = 1/2$. Find a simple joint probability assignment such that X and Y are uncorrelated but dependent.

f) You have seen that the moment generating function of a sum of independent random variables is equal to the product of the individual moment generating functions. Give an example where this is false if the variables are uncorrelated but dependent.

Exercise 1.8 Suppose X has the Poisson PMF, $P(X{=}n) = \lambda^n \exp(-\lambda)/n!$ for $n \geq 0$ and Y has the Poisson PMF, $P(Y{=}n) = \mu^n \exp(-\mu)/n!$ for $n \geq 0$. Find the distribution of $Z = X + Y$ and find the conditional distribution of Y conditional on $Z = n$.

Exercise 1.9 **a)** Suppose X, Y and Z are binary random variables, each taking on the value 0 with probability 1/2 and the value 1 with probability 1/2. Find a simple example in which X, Y, Z are statistically dependent but are *pairwise* statistically independent (i.e., X, Y are statistically independent, X, Z are statistically independent, and Y, Z are statistically independent). Give $P_{XYZ}(x, y, z)$ for your example.

b) Is pairwise statistical independence enough to ensure that

$$E\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} E[X_i]$$

for a set of random variables $X_1, \ldots, X_n$?

Exercise 1.10 Show that $E[X]$ is the value of z that minimizes $E[(X - z)^2]$.

Exercise 1.11 A computer system has n users, each with a unique name and password. Due to a software error, the n passwords are randomly permuted internally (i.e. each of the $n!$ possible permutations are equally likely. Only those users lucky enough to have had their passwords unchanged in the permutation are able to continue using the system.

a) What is the probability that a particular user, say user 1, is able to continue using the system?

b) What is the expected number of users able to continue using the system? Hint: Let X_i be a random variable with the value 1 if user i can use the system and 0 otherwise.

Exercise 1.12 Suppose the random variable X is continuous and has the distribution function $F_X(x)$. Consider another random variable $Y = F_X(X)$. That is, for any sample point α such that $X(\alpha) = x$, we have $Y(\alpha) = F_X(x)$. Show that Y is uniformly distributed in the interval 0 to 1.

Exercise 1.13 Let Z be an integer valued random variable with the PMF $P_Z(n) = 1/k$ for $0 \leq n \leq k - 1$. Find the mean, variance, and moment generating function of Z. Hint: The elegant way to do this is to let U be a uniformly distributed continuous random variable over $(0, 1]$ that is independent of Z. Then $U + Z$ is uniform over $(0, k]$. Use the known results about U and $U + Z$ to find the mean, variance, and mgf for Z.

Exercise 1.14 Let $\{X_n; n \geq 1\}$ be a sequence of independent but not identically distributed random variables. We say that the weak law of large numbers holds for this sequence if for all $\epsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \frac{E[S_n]}{n}\right| \geq \epsilon\right) = 0 \quad \text{where } S_n = X_1 + X_2 + \ldots + X_n$$

a) Show that (a) holds if there is some constant A such that $\mathrm{VAR}(X_n) \leq A$ for all n.

b) Suppose that $\mathrm{VAR}(X_n) \leq A n^{1-\alpha}$ for some $\alpha < 1$ and for all n. Show that (a) holds in this case.

Exercise 1.15 Let $\{X_i; i \geq 1\}$ be IID Bernoulli random variables. Let $P(X_i = 1) = \delta$, $P(X_i = 0) = 1 - \delta$. Let $S_n = X_1 + \ldots + X_n$. Let m be an arbitrary but fixed positive integer. Think! then evaluate the following and explain your answers:

a) $\lim_{n\to\infty} \sum_{i:n\delta-m\leq i\leq n\delta+m} P(S_n = i)$

b) $\lim_{n\to\infty} \sum_{i:0\leq i\leq n\delta+m} P(S_n = i)$

c) $\lim_{n\to\infty} \sum_{i:n(\delta-1/m)\leq i\leq n(\delta+1/m)} P(S_n = i)$

Exercise 1.16 Let $\{X_i; i \geq 1\}$ be IID random variables with mean 0 and infinite variance. Assume that $E[|X_i|^{1+h}] = \beta$ for some given h, $0 < h < 1$ and some given β. Let $S_n = X_1 + \ldots + X_n$.

a) Show that $P(|X_i| \geq y) \leq \beta\, y^{-1-h}$

b) Let $\{\widetilde{X}_i; i \geq 1\}$ be truncated variables $\widetilde{X}_i = \begin{cases} b quad: & X_i \geq b \\ X_i quad: & -b \leq X_i \leq b \\ -b quad: & X_i \leq -b \end{cases}$

Show that $E[\widetilde{X}^2] \leq \frac{2\beta b^{1-h}}{1-h}$ Hint: For a non-negative rv Z, $E[X^2] = \int_0^\infty 2\,z\,P(Z \geq z)dz$ (you can establish this, if you wish, by integration by parts).

c) Let $\widetilde{S}_n = \widetilde{X}_1 + \ldots + \widetilde{X}_n$. Show that $P(S_n \neq \widetilde{S}_n) \leq n\beta b^{-1-h}$

d Show that $P\left(\left|\frac{S_n}{n}\right| \geq \epsilon\right) \leq \beta\left[\frac{2b^{1-h}}{(1-h)n\epsilon^2} + \frac{n}{b^{1+h}}\right]$.

e) Optimize your bound with respect to b. How fast does this optimized bound approach 0 with increasing n?

Exercise 1.17 A town starts a mosquito control program and we let the random variable Z_n be the number of mosquitos at the end of the n^{th} year ($n = 0, 1, 2, \ldots$). Let X_n be the growth rate of mosquitos in year n; i.e., $Z_n = X_n Z_{n-1}; n \geq 1$. Assume that $\{X_n; n \geq 1\}$ is a sequence of IID random variables with the PMF $P(X{=}2) = 1/2$; $P(X{=}1/2) = 1/4$; $P(X{=}1/4) = 1/4$. Suppose that Z_0, the initial number of mosquitos, is some known constant and assume for simplicity and consistency that Z_n can take on non-integer values.

a) Find $E[Z_n]$ as a function of n and find $\lim_{n\to\infty} E[Z_n]$.

b) Let $W_n = \log_2 X_n$. Find $E[W_n]$ and $E[\log_2(Z_n/Z_0)]$ as a function of n.

c) There is a constant α such that $\lim_{n\to\infty}(1/n)[\log_2(Z_n/Z_0)] = \alpha$ with probability 1. Find α and explain how this follows from the strong law of large numbers.

d) Using (c), show that $\lim_{n\to\infty} Z_n = \beta$ with probability 1 for some β and evaluate β.

e) Explain carefully how the result in (a) and the result in (d) are possible. What you should learn from this problem is that the expected value of the log of a product of IID random variables is more significant that the expected value of the product itself.

Exercise 1.18 Use figure 1.4 to verify (1.20). Hint: Show that $yP(Y{\ge}y) \ge \int_{z\ge y} z dF_Y(z)$ and show that $\lim_{y\to\infty} \int_{z\ge y} z dF_Y(z) = 0$ if $E[Y]$ is finite.

Exercise 1.19 Show that $\prod_{m\ge n}(1 - 1/m) = 0$. Hint: Show that

$$\left(1 - \frac{1}{m}\right) = \exp\left(\ln\left(1 - \frac{1}{m}\right)\right) \le \exp\left(-\frac{1}{m}\right)$$