Chapter 1

INTRODUCTION AND REVIEW OF PROBABILITY

1.1 Probability models

Probability theory is a central field of mathematics, widely applicable to scientific, technological, and human situations involving uncertainty. The most obvious applications are to situations, such as games of chance, in which repeated trials of essentially the same procedure lead to differing outcomes. For example, when we flip a coin, roll a die, pick a card from a shuffled deck, or spin a ball onto a roulette wheel, the procedure is the same from one trial to the next, but the outcome (heads (H) or tails (T) in the case of a coin, one to six in the case of a die, etc.) varies from one trial to another in a seemingly random fashion.

For the case of flipping a coin, the outcome of the flip could be predicted from the initial position, velocity, and angular momentum of the coin and from the nature of the surface on which it lands. Thus, in one sense, a coin flip is deterministic rather than random and the same can be said for the other examples above. When these initial conditions are unspecified, however, as when playing these games, the outcome can again be viewed as random in some intuitive sense.

Many scientific experiments are similar to games of chance in the sense that multiple trials of apparently the *same* procedure lead to results that *vary* from one trial to another. In some cases, this variation is due to slight variations in the experimental procedure, in some it is due to noise, and in some, such as in quantum mechanics, the randomness is generally believed to be fundamental. Similar situations occur in many types of systems, especially those in which noise and random delays are important. Some of these systems, rather than being repetitions of a common basic procedure, are systems that evolve over time while still containing a sequence of underlying similar random occurrences.

This intuitive notion of randomness, as described above, is a very special kind of uncertainty. Rather than involving a lack of understanding, it involves a type of uncertainty that can lead to probabilistic models with precise results. As in any scientific field, the models might or might not correspond to reality very well, but when they do correspond to reality, there

is the sense that the situation is completely understood, while still being random.

For example, we all feel that we understand flipping a coin or rolling a die, but still accept randomness in each outcome. The theory of probability was developed particularly to give precise and quantitative understanding to these types of situations. The remainder of this section introduces this relationship between the precise view of probability theory and the intuitive view as used in applications and everyday language.

After this introduction, the following sections review probability theory as a mathematical discipline, with a special emphasis on the laws of large numbers. In the final section of this chapter, we use the theory and the laws of large numbers to obtain a fuller understanding of the relationship between theory and the real world.[1]

Probability theory, as a mathematical discipline, started to evolve in the 17th century and was initially focused on games of chance. The importance of the theory grew rapidly, particularly in the 20th century, and it now plays a central role in risk assessment, statistics, data networks, operations research, information theory, control theory, theoretical computer science, quantum theory, game theory, neurophysiology, and many other fields.

The core concept in probability theory is that of a *probability model*. Given the extent of the theory, both in mathematics and in applications, the simplicity of probability models is surprising. The first component of a probability model is a *sample space*, which is a *set* whose elements are called *outcomes* or *sample points*. Probability models are particularly simple in the special case where the sample space is finite,[2] and we consider only this case in the remainder of this section. The second component of a probability model is a class of *events*, which can be considered for now simply as the class of all subsets of the sample space. The third component is a *probability measure*, which can be regarded for now as the assignment of a nonnegative number to each outcome, with the restriction that these numbers must sum to one over the sample space. The probability of an event is the sum of the probabilities of the outcomes comprising that event.

These probability models play a dual role. In the first, the many known results about various classes of models, and the many known relationships between models, constitute the essence of probability theory. Thus one often studies a model not because of any relationship to the real world, but simply because the model provides a building block or example useful for the theory and thus ultimately useful for other models. In the other role, when probability theory is applied to some game, experiment, or some other situation involving randomness, a probability model is used to represent the experiment (in what follows, we refer to all of these random situations as experiments).

For example, the standard probability model for rolling a die uses $\{1, 2, 3, 4, 5, 6\}$ as the sample space, with each possible outcome having probability $1/6$. An *odd* result, *i.e.*, the subset $\{1, 3, 5\}$, is an example of an event in this sample space, and this event has probability

[1] It would be appealing to show how probability theory evolved from real-world random situations, but probability theory, like most mathematical theories, has evolved from complex interactions between theoretical developments and initially over-simplified models of real situations. The successes and flaws of such models lead to refinements of the models and the theory, which in turn suggest applications to totally different fields.

[2] A number of mathematical issues arise with infinite sample spaces, as discussed in the following section.

1/2. The correspondence between model and actual experiment seems straightforward here. Both have the same set of outcomes and, given the symmetry between faces of the die, the choice of equal probabilities seems natural. On closer inspection, there is the following important difference between the model and the actual rolling of a die.

The above model corresponds to a single roll of a die, with a probability defined for each possible outcome. In a real-world experiment where a single die is rolled, an outcome k from 1 to 6 occurs, but there is no *observable* probability for k.

Our intuitive notion of rolling dice, however, involves an experiment with repeated rolls of a die or rolls of different dice. With n rolls altogether, there are are 6^n possible outcomes, one for each possible n-tuple of individual die outcomes. The standard probability model for this repeated-roll experiment is to assign probability 6^{-n} to each possible n-tuple. In this n-repetition experiment, the real-world relative frequency of k, *i.e.*, the fraction of rolls for which the result is k, can be compared with the sample value of the relative frequency of k in the model for repeated rolls. The sample value of the relative frequency of k in this n-repetition model resembles the probability of k in the single-roll experiment in a way to be explained later. This relationship through relative frequencies in a repeated experiment helps overcome the non-observable nature of probabilities in the real world.

1.1.1 The sample space of a probability model

An *outcome* or *sample point* in a probability model corresponds to a complete result (with all detail specified) of the experiment being modeled. For example, a game of cards is often appropriately modeled by the arrangement of cards within a shuffled 52 card deck, thus giving rise to a set of 52! outcomes (incredibly detailed, but trivially simple in structure), even though the entire deck might not be played in one trial of the game. A poker hand with 4 aces is an *event* rather than an *outcome* in this model, since many arrangements of the cards can give rise to 4 aces in a given hand. The possible outcomes in a probability model (and in the experiment being modeled) are mutually exclusive and collectively constitute the entire sample space (space of possible results). An outcome is often called a *finest grain* result of the model in the sense that an outcome ω contains no subsets other than the empty set ϕ and the singleton subset $\{\omega\}$. Thus events typically give only partial information about the result of the experiment, whereas an outcome fully specifies the result.

In choosing the sample space for a probability model of an experiment, we often omit details that appear irrelevant for the purpose at hand. Thus in modeling the set of outcomes for a coin toss as $\{H, T\}$, we ignore the type of coin, the initial velocity and angular momentum of the toss, etc. We also omit the rare possibility that the coin comes to rest on its edge. Sometimes, conversely, the sample space is enlarged beyond what is relevant in the interest of structural simplicity. An example is the above use of a shuffled deck of 52 cards.

The choice of the sample space in a probability model is similar to the choice of a mathematical model in any branch of science. That is, one simplifies the physical situation by eliminating detail of little apparent relevance. One often does this in an iterative way, using a very simple model to acquire initial understanding, and then successively choosing more detailed models based on the understanding from earlier models.

The mathematical theory of probability views the sample space simply as an abstract set of elements, and from a strictly mathematical point of view, the idea of doing an experiment and getting an outcome is a distraction. For visualizing the correspondence between the theory and applications, however, it is better to view the abstract set of elements as the set of possible outcomes of an idealized experiment in which, when the idealized experiment is performed, one and only one of those outcomes occurs. The two views are mathematically identical, but it will be helpful to refer to the first view as a probability model and the second as an idealized experiment. In applied probability texts and technical articles, these idealized experiments, rather than real-world situations are often the primary topic of discussion.[3]

1.1.2 Assigning probabilities for finite sample spaces

The word *probability* is widely used in everyday language, and most of us attach various intuitive meanings[4] to the word. For example, everyone would agree that something virtually impossible should be assigned a probability close to 0 and something virtually certain should be assigned a probability close to 1. For these special cases, this provides a good rationale for choosing probabilities. The relationship between *virtually* and *close to* are unclear at the moment, but if there is some implied limiting process, we would all agree that, in the limit, certainty and impossibility correspond to probabilities 1 and 0 respectively.

Between virtual impossibility and certainty, if one outcome appears to be closer to certainty than another, its probability should be correspondingly greater. This intuitive notion is imprecise and highly subjective; it provides little rationale for choosing numerical probabilities for different outcomes, and, even worse, little rationale justifying that probability models bear any precise relation to real-world situations.

Symmetry can often provide a better rationale for choosing probabilities. For example, the symmetry between H and T for a coin, or the symmetry between the the six faces of a die, motivates assigning equal probabilities, $1/2$ each for H and T and $1/6$ each for the six faces of a die. This is reasonable and extremely useful, but there is no completely convincing reason for choosing probabilities based on symmetry.

Another approach is to perform the experiment many times and choose the probability of each outcome as the relative frequency of that outcome (*i.e.*, the number of occurrences of that outcome divided by the total number of trials). Experience shows that the relative frequency of an outcome often approaches a limiting value with an increasing number of trials. Associating the probability of an outcome with that limiting relative frequency is certainly close to our intuition and also appears to provide a testable criterion between model and real world. This criterion is discussed in Sections 1.6.1 and 1.6.2 and provides a very concrete way to use probabilities, since it suggests that the randomness in a single

[3] This is not intended as criticism, since we will see that there are good reasons to concentrate initially on such idealized experiments. However, readers should always be aware that modeling errors are the major cause of misleading results in applications of probability, and thus modeling must be seriously considered before using the results.

[4] It is popular to try to define probability by likelihood, but this is unhelpful since the words are essentially synonyms.

trial tends to disappear in the aggregate of many trials. Other approaches to choosing probability models will be discussed later.

1.2 The axioms of probability theory

As the applications of probability theory became increasingly varied and complex during the 20th century, the need arose to put the theory on a firm mathematical footing. This was accomplished by an axiomatization of the theory, successfully carried out by the great Russian mathematician A. N. Kolmogorov [14] in 1932. Before stating and explaining these axioms of probability theory, the following two examples explain why the simple approach of the last section, assigning a probability to each sample point, often fails with infinite sample spaces.

Example 1.2.1. Suppose we want to model the phase of a sine wave, where the phase is viewed as being "uniformly distributed" between 0 and 2π. If this phase is the only quantity of interest, it is reasonable to choose a sample space consisting of the set of real numbers between 0 and 2π. There are uncountably[5] many possible phases between 0 and 2π, and with any reasonable interpretation of uniform distribution, one must conclude that each sample point has probability zero. Thus, the simple approach of the last section leads us to conclude that any event in this space with a finite or countably infinite set of sample points should have probability zero. That simple approach does not help in finding the probability, say, of the interval $(0, \pi)$.

For this example, the appropriate view is that taken in all elementary probability texts, namely to assign a probability density $\frac{1}{2\pi}$ to the phase. The probability of an event can then usually be found by integrating the density over that event. Useful as densities are, however, they do not lead to a general approach over arbitrary sample spaces.[6]

Example 1.2.2. Consider an infinite sequence of coin tosses. The usual probability model is to assign probability 2^{-n} to each possible initial n-tuple of individual outcomes. Then in the limit $n \to \infty$, the probability of any given sequence is 0. Again, expressing the probability of an event involving infinitely many tosses as a sum of individual sample-point probabilities does not work. The obvious approach (which we often adopt for this and similar situations) is to evaluate the probability of any given event as an appropriate limit, as $n \to \infty$, of the outcome from the first n tosses.

We will later find a number of situations, even for this almost trivial example, where working with a finite number of elementary experiments and then going to the limit is very awkward. One example, to be discussed in detail later, is the strong law of large numbers (SLLN). This

[5] A set is uncountably infinite if it is infinite and its members cannot be put into one-to-one correspondence with the positive integers. For example the set of real numbers over some interval such as $(0, 2\pi)$ is uncountably infinite. The Wikipedia article on countable sets provides a friendly introduction to the concepts of countability and uncountability.

[6] It is possible to avoid the consideration of infinite sample spaces here by quantizing the possible phases. This is analogous to avoiding calculus by working only with discrete functions. Both usually result in both artificiality and added complexity.

law looks directly at events consisting of infinite length sequences and is best considered in the context of the axioms to follow.

Although appropriate probability models can be generated for simple examples such as those above, there is a need for a consistent and general approach. In such an approach, rather than assigning probabilities to sample points, which are then used to assign probabilities to events, *probabilities must be associated directly with events.* The axioms to follow establish consistency requirements between the probabilities of different events. The axioms, and the corollaries derived from them, are consistent with one's intuition, and, for finite sample spaces, are consistent with our earlier approach. Dealing with the countable unions of events in the axioms will be unfamiliar to some students, but will soon become both familiar and consistent with intuition.

The strange part of the axioms comes from the fact that defining the class of events as the set of *all* subsets of the sample space is usually inappropriate when the sample space is uncountably infinite. What is needed is a class of events that is large enough that we can almost forget that some very strange subsets are excluded. This is accomplished by having two simple sets of axioms, one defining the class of events,[7] and the other defining the relations between the probabilities assigned to these events. In this theory, all events have probabilities, but those truly weird subsets that are not events do not have probabilities. This will be discussed more after giving the axioms for events.

The axioms for events use the standard notation of set theory. Let Ω be the set of all sample points for a given experiment. The events are subsets of the sample space. The union of n subsets (events) $A_1, A_2, \cdots, A_n$ is denoted by either $\bigcup_{i=1}^{n} A_i$ or $A_1 \bigcup \cdots \bigcup A_n$, and consists of all points in at least one of $A_1, \ldots, A_n$. Similarly, the intersection of these subsets is denoted by either $\bigcap_{i=1}^{n} A_i$ or[8] $A_1 A_2 \cdots A_n$ and consists of all points in all of $A_1, \ldots, A_n$.

A *sequence* of events is a collection of events in one-to-one correspondence with the positive integers, *i.e.*, $A_1, A_2, \ldots$, ad infinitum. A countable union, $\bigcup_{i=1}^{\infty} A_i$ is the set of points in one or more of $A_1, A_2, \ldots$. Similarly, a countable intersection $\bigcap_{i=1}^{\infty} A_i$ is the set of points in all of $A_1, A_2, \ldots$. Finally, the complement A^c of a subset (event) A is the set of points in Ω but not A.

1.2.1 Axioms for events

Given a sample space Ω, the class of subsets of Ω that constitute the set of events satisfies the following axioms:

1. Ω is an event.
2. For every sequence of events $A_1, A_2, \ldots$,, the union $\bigcup_{n=1}^{\infty} A_n$ is an event.
3. For every event A, the complement A^c is an event.

There are a number of important corollaries of these axioms. First, the empty set ϕ is an event. This follows from Axioms 1 and 3, since $\phi = \Omega^c$. The empty set does not correspond

[7] A class of elements satisfying these axioms is called a σ-algebra or, less commonly, a σ-field.

[8] Intersection is also sometimes denoted as $A_1 \bigcap \cdots \bigcap A_n$, but is usually abbreviated as $A_1 A_2 \cdots A_n$.

to our intuition about events, but the theory would be extremely awkward if it were omitted. Second, every finite union of events is an event. This follows by expressing $A_1 \bigcup \cdots \bigcup A_n$ as $\bigcup_{i=1}^{\infty} A_i$ where $A_i = \phi$ for all $i > n$. Third, every finite or countable intersection of events is an event. This follows from deMorgan's law,

$$\left[\bigcup_n A_n\right]^c = \bigcap_n A_n^c.$$

Although we will not make a big fuss about these axioms in the rest of the text, we will be careful to use only complements and countable unions and intersections in our analysis. Thus subsets that are not events will not arise.

Note that the axioms do not say that all subsets of Ω are events. In fact, there are many rather silly ways to define classes of events that obey the axioms. For example, the axioms are satisfied by choosing only the universal set Ω and the empty set ϕ to be events. We shall avoid such trivialities by assuming that for each sample point ω, the singleton subset $\{\omega\}$ is an event. For finite sample spaces, this assumption, plus the axioms above, imply that all subsets are events.

For uncountably infinite sample spaces, such as the sinusoidal phase above, this assumption, plus the axioms above, still leaves considerable freedom in choosing a class of events. As an example, the class of all subsets of Ω satisfies the axioms but surprisingly does not allow the probability axioms to be satisfied in any sensible way. How to choose an appropriate class of events requires an understanding of measure theory which would take us too far afield for our purposes. Thus we neither assume nor develop measure theory here.[9]

From a pragmatic standpoint, we start with the class of events of interest, such as those required to define the random variables needed in the problem. That class is then extended so as to be closed under complementation and countable unions. Measure theory shows that this extension is always possible, and we simply accept that as a known result.

1.2.2 Axioms of probability

Given any sample space Ω and any class of events $\mathcal{E}$ satisfying the axioms of events, a probability rule is a function $\Pr\{\}$ mapping each $A \in \mathcal{E}$ to a (finite[10]) real number in such a way that the following three probability axioms[11] hold:

1. $\Pr\{\Omega\} = 1$.

2. For every event A, $\Pr\{A\} \geq 0$.

[9] There is no doubt that measure theory is useful in probability theory, and serious students of probability should certainly learn measure theory at some point. For application-oriented people, however, it seems advisable to acquire more insight and understanding of probability, at a graduate level, before concentrating on the abstractions and subtleties of measure theory.

[10] The word *finite* is redundant here, since the set of real numbers, by definition, does not include $\pm\infty$. The set of real numbers with $\pm\infty$ appended, is called the set of *extended* real numbers

[11] Sometimes finite additivity, (1.3), is added as an additional axiom. This addition is quite intuitive and avoids the technical and somewhat peculiar proofs given for (1.2) and (1.3).

3. The probability of the union of any sequence $A_1, A_2, \ldots$ of disjoint events is given by

$$\Pr\Big\{\bigcup_{n=1}^{\infty} A_n\Big\} = \sum_{n=1}^{\infty} \Pr\{A_n\}, \tag{1.1}$$

where $\sum_{n=1}^{\infty} \Pr\{A_n\}$ is shorthand for $\lim_{m\to\infty} \sum_{n=1}^{m} \Pr\{A_n\}$.

The axioms imply the following useful corollaries:

$$\Pr\{\phi\} = 0 \tag{1.2}$$

$$\Pr\Big\{\bigcup_{n=1}^{m} A_n\Big\} = \sum_{n=1}^{m} \Pr\{A_n\} \qquad \text{for } A_1, \ldots, A_m \text{ disjoint} \tag{1.3}$$

$$\Pr\{A^c\} = 1 - \Pr\{A\} \qquad \text{for all } A \tag{1.4}$$

$$\Pr\{A\} \le \Pr\{B\} \qquad \text{for all } A \subseteq B \tag{1.5}$$

$$\Pr\{A\} \le 1 \qquad \text{for all } A \tag{1.6}$$

$$\sum_n \Pr\{A_n\} \le 1 \qquad \text{for } A_1, \ldots, \text{ disjoint} \tag{1.7}$$

$$\Pr\Big\{\bigcup_{n=1}^{\infty} A_n\Big\} = \lim_{m\to\infty} \Pr\Big\{\bigcup_{n=1}^{m} A_n\Big\} \tag{1.8}$$

$$\Pr\Big\{\bigcup_{n=1}^{\infty} A_n\Big\} = \lim_{n\to\infty} \Pr\{A_n\} \qquad \text{for } A_1 \subseteq A_2 \subseteq \cdots. \tag{1.9}$$

To verify (1.2), consider a sequence of events, $A_1, A_2, \ldots$, for which $A_n = \phi$ for each n. These events are disjoint since ϕ contains no outcomes, and thus has no outcomes in common with itself or any other event. Also, $\bigcup_n A_n = \phi$ since this union contains no outcomes. Axiom 3 then says that

$$\Pr\{\phi\} = \lim_{m\to\infty} \sum_{n=1}^{m} \Pr\{A_n\} = \lim_{m\to\infty} m\Pr\{\phi\}.$$

Since $\Pr\{\phi\}$ is a real number, this implies that $\Pr\{\phi\} = 0$.

To verify (1.3), apply Axiom 3 to the disjoint sequence $A_1, \ldots, A_m, \phi, \phi, \ldots$.

One might reasonably guess that (1.3), along with Axioms 1 and 2 implies Axiom 3. Exercise 1.3 shows why this guess is incorrect.

To verify (1.4), note that $\Omega = A \bigcup A^c$. Then apply (1.3) to the disjoint sets A and A^c.

To verify (1.5), note that if $A \subseteq B$, then $B = A \bigcup (B - A)$ where $B - A$ is an alternate way to write $B \bigcap A^c$. We see then that A and $B - A$ are disjoint, so from (1.3),

$$\Pr\{B\} = \Pr\Big\{A \bigcup (B - A)\Big\} = \Pr\{A\} + \Pr\{B - A\} \ge \Pr\{A\},$$

where we have used Axiom 2 in the last step.

To verify (1.6) and (1.7), first substitute Ω for B in (1.5) and then substitute $\bigcup_n A_n$ for A.

Finally, (1.8) is established in Exercise 1.4, part (e), and (1.9) is a simple consequence of (1.8).

The axioms specify the probability of any *disjoint* union of events in terms of the individual event probabilities, but what about a finite or countable union of arbitrary events? Exercise 1.4 (b) shows that in this case, (1.3) can be generalized to

$$\Pr\Big\{\bigcup_{n=1}^{m} A_n\Big\} = \sum_{n=1}^{m} \Pr\{B_n\}\,, \tag{1.10}$$

where $B_1 = A_1$ and for each $n > 1$, $B_n = A_n - \bigcup_{j=1}^{n-1} A_j$ is the set of points in A_n but not in any of the sets $A_1, \ldots, A_{n-1}$. The probability of a countable union is then given by (1.8). In order to use this, one must know not only the event probabilities for $A_1, A_2 \ldots$, but also the probabilities of their intersections. The union bound, which is derived in Exercise 1.4 (c), depends only on the individual event probabilities, and gives the following frequently useful upper bound on the union probability.

$$\Pr\Big\{\bigcup_{n} A_n\Big\} \leq \sum_{n} \Pr\{A_n\} \qquad \text{(Union bound)}. \tag{1.11}$$

1.3 Probability review

1.3.1 Conditional probabilities and statistical independence

Definition 1.3.1. *For any two events A and B (with $\Pr\{B\} > 0$), the conditional probability of A, conditional on B, is defined by*

$$\Pr\{A|B\} = \Pr\{AB\}\,/\Pr\{B\}\,. \tag{1.12}$$

One visualizes an experiment that has been partly carried out with B as the result. Then $\Pr\{A|B\}$ can be viewed as the probability of A normalized to a sample space restricted to event B. Within this restricted sample space, we can view B as the sample space (*i.e.*, as the set of outcomes that remain possible upon the occurrence of B) and AB as an event within this sample space. For a fixed event B, we can visualize mapping each event A in the original space to event AB in the restricted space. It is easy to see that the event axioms are still satisfied in this restricted space. Assigning probability $\Pr\{A|B\}$ to each event AB in the restricted space, it is easy to see that the axioms of probability are satisfied when B is regarded as the entire sample space. In other words, everything we know about probability can also be applied to such a restricted probability space.

Definition 1.3.2. *Two events, A and B, are statistically independent (or, more briefly,* independent*) if*

$$\Pr\{AB\} = \Pr\{A\}\Pr\{B\}\,.$$

For $\Pr\{B\} > 0$, this is equivalent to $\Pr\{A|B\} = \Pr\{A\}$. This latter form corresponds to our intuitive view of independence, since it says that the observation of B does not change the probability of A. Such intuitive statements about "observation" and "occurrence" are helpful in reasoning probabilistically, but sometimes cause confusion. For example, Bayes law, in the form $\Pr\{A|B\}\Pr\{B\} = \Pr\{B|A\}\Pr\{A\}$, is an immediate consequence of the

definition of conditional probability in (1.12). However, if we can only interpret $\Pr\{A|B\}$ when B is 'observed' or occurs 'before' A, then we cannot interpret $\Pr\{B|A\}$ and $\Pr\{A|B\}$ together. This caused immense confusion in probabilistic arguments before the axiomatic theory was developed.

The notion of independence is of vital importance in defining, and reasoning about, probability models. We will see many examples where very complex systems become very simple, both in terms of intuition and analysis, when appropriate quantities are modeled as statistically independent. An example will be given in the next subsection where repeated independent experiments are used to understand arguments about relative frequencies.

Often, when the assumption of independence turns out to be oversimplified, it is reasonable to assume conditional independence, where A and B are said to be *conditionally independent* given C if $\Pr\{AB|C\} = \Pr\{A|C\}\Pr\{B|C\}$. Most of the stochastic processes to be studied here are characterized by particular forms of independence or conditional independence.

For more than two events, the definition of statistical independence is a little more complicated.

Definition 1.3.3. *The events $A_1, \ldots, A_n$, $n > 2$ are statistically independent if for each collection S of two or more of the integers 1 to n.*

$$\Pr\Big\{\bigcap\nolimits_{i\in S} A_i\Big\} = \prod\nolimits_{i\in S} \Pr\{A_i\}\,. \tag{1.13}$$

This includes the entire collection $\{1, \ldots, n\}$, so one necessary condition for independence is that

$$\Pr\Big\{\bigcap\nolimits_{i=1}^{n} A_i\Big\} = \prod\nolimits_{i=1}^{n} \Pr\{A_i\}\,. \tag{1.14}$$

It might be surprising that (1.14) does not imply (1.13), but the example in Exercise 1.5 will help clarify this. This definition will become clearer (and simpler) when we see how to view independence of events as a special case of independence of random variables.

1.3.2 Repeated idealized experiments

Much of our intuitive understanding of probability comes from the notion of repeating the same idealized experiment many times (*i.e.*, performing multiple trials of the same experiment). However, the axioms of probability contain no explicit recognition of such repetitions. The appropriate way to handle n repetitions of an idealized experiment is through an extended experiment whose sample points are n-tuples of sample points from the original experiment. Such an extended experiment is viewed as n *trials* of the original experiment. The notion of multiple trials of a given experiment is so common that one sometimes fails to distinguish between the original experiment and an extended experiment with multiple trials of the original experiment.

To be more specific, given an original sample space Ω, the sample space of an n-repetition model is the Cartesian product

$$\Omega^{\times n} = \{(\omega_1, \omega_2, \ldots, \omega_n) \,:\, \omega_i \in \Omega \text{ for each } i, 1 \le i \le n\}, \tag{1.15}$$

i.e., the set of all n-tuples for which each of the n components of the n-tuple is an element of the original sample space Ω. Since each sample point in the n-repetition model is an n-tuple of points from the original Ω, it follows that an event in the n-repetition model is a subset of $\Omega^{\times n}$, *i.e.*, a collection of n-tuples $(\omega_1, \ldots, \omega_n)$, where each ω_i is a sample point from Ω. This class of events in $\Omega^{\times n}$ should include each event of the form $\{(A_1 A_2 \cdots A_n)\}$, where $\{(A_1 A_2 \cdots A_n)\}$ denotes the collection of n-tuples $(\omega_1, \ldots, \omega_n)$ where $\omega_i \in A_i$ for $1 \leq i \leq n$. The set of events (for n-repetitions) mult also be extended to be closed under complementation and countable unions and intersections.

The simplest and most natural way of creating a probability model for this extended sample space and class of events is through the assumption that the n-trials are statistically independent. More precisely, we assume that for each extended event $\{(A_1 A_2 \cdots A_n)\}$ contained in $\Omega^{\times n}$, we have

$$\Pr\{(A_1 A_2 \cdots A_n)\} = \prod_{i=1}^{n} \Pr\{A_i\}, \tag{1.16}$$

where $\Pr\{A_i\}$ is the probability of event A_i in the original model. Note that since Ω can be substituted for any A_i in this formula, the subset condition of (1.13) is automatically satisfied. In other words, for any probability model, there is an extended independent n-repetition model for which the events in each trial are independent of those in the other trials. In what follows, we refer to this as the probability model for n independent identically distributed (IID) trials of a given experiment.

The niceties of how to create this model for n IID arbitrary experiments depend on measure theory, but we simply rely on the existence of such a model and the independence of events in different repetitions. What we have done here is very important conceptually. A probability model for an experiment does not say anything directly about repeated experiments. However, questions about independent repeated experiments can be handled directly within this extended model of n IID repetitions. This can also be extended to a countable number of IID trials.

1.3.3 Random variables

The outcome of a probabilistic experiment often specifies a collection of numerical values such as temperatures, voltages, numbers of arrivals or departures in various time intervals, etc. Each such numerical value varies, depending on the particular outcome of the experiment, and thus can be viewed as a mapping from the set Ω of sample points to the set $\mathbb{R}$ of real numbers (note that $\mathbb{R}$ does not include $\pm\infty$). These mappings from sample points to real numbers are called random variables.

Definition 1.3.4. *A random variable (rv) is essentially a function X from the sample space Ω of a probability model to the set of real numbers $\mathbb{R}$. Three modifications are needed to make this precise. First, X might be undefined or infinite for a subset of Ω that has 0 probability.*[12] *Second, the mapping $X(\omega)$ must have the property that $\{\omega \in \Omega : X(\omega) \leq x\}$*

[12] For example, suppose Ω is the closed interval $[0, 1]$ of real numbers with a uniform probability distribution over $[0, 1]$. If $X(\omega) = 1/\omega$, then the sample point 0 maps to ∞ but X is still regarded as a rv. These subsets of 0 probability are usually ignored, both by engineers and mathematicians. Thus, for example, the set $\{\omega \in \Omega : X(\omega) \leq x\}$ means the set for which $X(\omega)$ is both defined and satisfies $X(\omega) \leq x$.

is an event[13] *for each* $x \in \mathbb{R}$. *Third, every finite set of rv's* $X_1, \ldots, X_n$ *has the property that* $\{\omega : X_1(\omega) \leq x_1, \ldots, X_n(\omega) \leq x_n\}$ *is an event for each* $x_1 \in \mathbb{R}, \ldots, x_n \in \mathbb{R}$.

As with any function, there is often confusion between the function itself, which is called X in the definition above, and the value $X(\omega)$ the function takes on for a sample point ω. This is particularly prevalent with random variables (rv's) since we intuitively associate a rv with its sample value when an experiment is performed. We try to control that confusion here by using X, $X(\omega)$, and x, respectively, to refer to the rv, the sample value taken for a given sample point ω, and a generic sample value.

Definition 1.3.5. *The distribution function*[14] $\mathsf{F}_X(x)$ *of a random variable (rv)* X *is a function,* $\mathbb{R} \to \mathbb{R}$, *defined by* $\mathsf{F}_X(x) = \Pr\{\omega \in \Omega : X(\omega) \leq x\}$. *The argument* ω *is usually omitted for brevity, so* $\mathsf{F}_X(x) = \Pr\{X \leq x\}$.

Note that x is the argument of $\mathsf{F}_X(x)$ and the subscript X denotes the particular rv under consideration. As illustrated in Figure 1.1, the distribution function $\mathsf{F}_X(x)$ is nondecreasing with x and must satisfy the limits $\lim_{x\to-\infty} \mathsf{F}_X(x) = 0$ and $\lim_{x\to\infty} \mathsf{F}_X(x) = 1$. It is not hard to show, from the axioms, that $\mathsf{F}_X(x)$ is continuous from the right (*i.e.*, that for every $x \in \mathbb{R}$, $\lim_{k\to\infty} \mathsf{F}_X(x + 1/k) = \mathsf{F}_X(x)$.

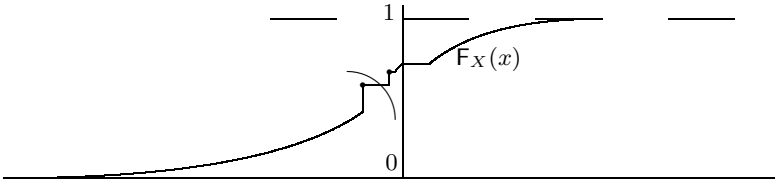

Figure 1.1: Example of a distribution function for a rv that is neither continuous nor discrete. If $\mathsf{F}_X(x)$ has a discontinuity at some x_o, it means that there is a discrete probability at x_o equal to the magnitude of the discontinuity. In this case $\mathsf{F}_X(x_o)$ is given by the height of the upper point at the discontinuity.

Because of the definition of a rv, the set $\{X \leq x\}$ for any rv X and any real number x must be an event, and thus $\Pr\{X \leq x\}$ must be defined for all real x.

The concept of a rv is often extended to complex random variables (rv's) and vector rv's. A *complex random variable* is a mapping from the sample space to the set of finite complex numbers, and a *vector random variable (rv)* is a mapping from the sample space to the finite vectors in some finite dimensional vector space. Another extension is that of defective rvs. X is *defective* if there is an event 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 Ω to $\mathbb{R}$.

[13] These last two modifications are technical limitations connected with measure theory. They can usually be ignored, since they are satisfied in all but the most bizarre conditions. However, just as it is important to know that not all subsets in a probability space are events, one should know that not all functions from Ω to $\mathbb{R}$ are rv's.

[14] The distribution function is sometimes referred to as the cumulative distribution function.

If X has only a finite or countable number of possible sample values, say $x_1, x_2, \ldots$, the probability $\Pr\{X = x_i\}$ of each sample value x_i is called the probability mass function (PMF) at x_i and denoted by $\mathsf{p}_X(x_i)$; such a random variable is called *discrete*. The distribution function of a discrete rv is a 'staircase function,' staying constant between the possible sample values and having a jump of magnitude $\mathsf{p}_X(x_i)$ at each sample value x_i. Thus the PMF and the distribution function each specify the other for discrete rv's.

If the distribution function $\mathsf{F}_X(x)$ of a rv X has a (finite) derivative at x, the derivative is called the *probability density* (or the density) of X at x and denoted by $\mathsf{f}_X(x)$; for sufficiently small δ, $\delta\mathsf{f}_X(x)$ then approximates the probability that X is mapped to a value between x and $x + \delta$. If the density exists for all x, the rv is said to be *continuous*. More generally, if there is a function $\mathsf{f}_X(x)$ such that, for each $x \in \mathbb{R}$, the distribution function satisfies $\int_{-\infty}^{x} \mathsf{f}_X(y)\, dy$, then the rv is said to be continuous and f_X is the probability density. This generalization allows the density to be discontinuous. In other words, a continuous rv requires a little more than a continuous distribution function and a little less than a continuous density.

Elementary probability courses work primarily with the PMF and the density, since they are convenient for computational exercises. We will often work with the distribution function here. This is partly because it is always defined, partly to avoid saying everything thrice, for discrete, continuous, and other rv's, and partly because the distribution function is often most important in limiting arguments such as steady-state time-average arguments. For distribution functions, density functions, and PMF's, the subscript denoting the rv is often omitted if the rv is clear from the context. The same convention is used for complex rv's and vector rv's.

Appendix A lists the PMF's of a number of widely used discrete rv's and the densities of some equally popular continuous rv's. The mean, variance and moment generating functions of these variables are also listed for ready reference.

1.3.4 Multiple random variables and conditional probabilities

Often we must deal with multiple random variables (rv's) in a single probability experiment. If $X_1, X_2, \ldots, X_n$ are rv's or the components of a vector rv, their joint distribution function is defined by

$$\mathsf{F}_{X_1 \cdots X_n}(x_1, x_2, \ldots, x_n) = \Pr\{\omega \in \Omega : X_1(\omega) \leq x_1, X_2(\omega) \leq x_2, \ldots, X_n(\omega) \leq x_n\}. \quad (1.17)$$

This definition goes a long way toward explaining why we need the notion of a sample space Ω when all we want to talk about is a set of rv's. The distribution function of a rv fully describes the individual behavior of that rv, but Ω and the above mappings are needed to describe how the rv's interact.

For a vector rv $\boldsymbol{X}$ with components $X_1, \ldots, X_n$, or a complex rv X with real and imaginary parts X_1, X_2, the distribution function is also defined by (1.17). Note that $\{X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n\}$ 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 rv (called a marginal distribution) is given by

$$\mathsf{F}_{X_i}(x_i) = \mathsf{F}_{X_1 \cdots X_{i-1} X_i X_{i+1} \cdots X_n}(\infty, \ldots, \infty, x_i, \infty, \ldots, \infty).$$

If the rv's are all discrete, there is a joint PMF which specifies and is specified by the joint distribution function. It is given by

$$\mathsf{p}_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \Pr\{X_1 = x_1, \ldots, X_n = x_n\}.$$

Similarly, if the joint distribution function is differentiable everywhere, it specifies and is specified by the joint probability density,

$$\mathsf{f}_{X_1 \ldots X_n}(x_1, \ldots, x_n) = \frac{\partial^n \mathsf{F}(x_1, \ldots, x_n)}{\partial x_1 \partial x_2 \cdots \partial x_n}.$$

.

Two rv's, say X and Y, are *statistically independent* (or, more briefly, *independent*) if

$$\mathsf{F}_{XY}(x, y) = \mathsf{F}_X(x)\mathsf{F}_Y(y) \qquad \text{for each } x \in \mathbb{R}, y \in \mathbb{R}. \tag{1.18}$$

If X and Y are discrete rv's then the definition of independence in (1.18) is equivalent to the corresponding statement for PMF's,

$$\mathsf{p}_{XY}(x_i y_j) = \mathsf{p}_X(x_i)\mathsf{p}_Y(y_j) \qquad \text{for each value } x_i \text{ of } X \text{ and } y_j \text{ of } Y.$$

Since $\{X = x_i\}$ and $\{Y = y_j\}$ are events, the conditional probability of $\{X = x_i\}$ conditional on $\{Y = y_j\}$ (assuming $\mathsf{p}_Y(y_j) > 0$) is given by (1.12) to be

$$\mathsf{p}_{X|Y}(x_i \mid y_j) = \frac{\mathsf{p}_{XY}(x_i, y_j)}{\mathsf{p}_Y(y_j)}.$$

If $\mathsf{p}_{X|Y}(x_i \mid y_j) = \mathsf{p}_X(x_i)$ for all i, j, then it is seen that X and Y are independent. This captures the intuitive notion of independence better than (1.18) for discrete rv's , since it can be viewed as saying that the PMF of X is not affected by the sample value of Y.

If X and Y have a joint density, then (1.18) is equivalent to

$$\mathsf{f}_{XY}(x, y) = \mathsf{f}_X(x)\mathsf{f}_Y(y) \qquad \text{for each } x \in \mathbb{R}, y \in \mathbb{R}.$$

If $\mathsf{f}_Y(y) > 0$, the conditional density can be defined as $\mathsf{f}_{X|Y}(x \mid y) = \frac{\mathsf{f}_{XY}(x,y)}{\mathsf{f}_Y(y)}$. Then statistical independence can be expressed as

$$\mathsf{f}_{X|Y}(x|y) = \mathsf{f}_X(x) \qquad \text{where } \mathsf{f}_Y(y) > 0. \tag{1.19}$$

This captures the intuitive notion of statistical independence for continuous rv's better than (1.18), but it does not quite say that the density of X, conditional on $Y = y$ is the same as the marginal density of X. The event $\{Y = y\}$ has zero probability for a continuous rv, and we cannot condition on events of zero probability. If we look at the derivatives defining these densities, the conditional density looks at the probability that $\{x \leq X \leq x + \delta\}$ given

that $\{y \leq Y \leq y+\epsilon\}$ in the limit $\delta, \epsilon \to 0$. At some level, this is a very technical point and the intuition of conditioning on $\{Y{=}y\}$ works very well. Furthermore, problems are often directly modeled in terms of conditional probability densities, so that viewing a conditional density as a limit is less relevant.

More generally the probability of an arbitrary event A conditional on a given value of a continuous rv Y is given by

$$\Pr\{A \mid Y = y\} = \lim_{\delta \to 0} \frac{\Pr\{A, Y \in [y, y+\delta]\}}{\Pr\{Y \in [y, y+\delta]\}}.$$

We next generalize the above results about two rv's to the case of n rv's $\boldsymbol{X} = X_1, \ldots, X_n$. Statistical independence is then defined by the equation

$$\mathsf{F}_{\boldsymbol{X}}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \Pr\{X_i \leq x_i\} = \prod_{i=1}^{n} \mathsf{F}_{X_i}(x_i) \qquad \text{for all values of } x_1, \ldots, x_n. \tag{1.20}$$

In other words, $X_1, \ldots, X_n$ are independent if the events $X_i \leq x_i$ for $1 \leq i \leq n$ are independent for all choices of $x_1, \ldots, x_n$. If the density or PMF exists, (1.20) is equivalent to a product form for the density or mass function. A set of rv's is said to be pairwise independent if each pair of rv's in the set is independent. As shown in Exercise 1.20, pairwise independence does not imply that the entire set is independent.

Independent rv's are very often also identically distributed, *i.e.*, they all have the same distribution function. These cases arise so often that we abbreviate independent identically distributed by IID. For the IID case (1.20) becomes

$$\mathsf{F}_{\boldsymbol{X}}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \mathsf{F}_X(x_i). \tag{1.21}$$

1.3.5 Stochastic processes and the Bernoulli process

A stochastic process (or random process[15]) is an infinite collection of rv's, usually indexed by an integer or a real number often interpreted as time.[16] Thus each sample point of the probability model maps to an infinite collection of sample values of rv's. If the index is regarded as time, then each sample point maps to a function of time called a sample path or sample function. These sample paths might vary continuously with time or might vary only at discrete times, and if they vary at discrete times, those times might be deterministic or random.

In many cases, this collection of rv's comprising the stochastic process is the only thing of interest. In this case the sample points of the probability model can be taken to be the

[15]Stochastic and random are synonyms, but *random* has become more popular for rv's and *stochastic* for stochastic processes. The reason for the author's choice is that the common-sense intuition associated with randomness appears more important than mathematical precision in reasoning about rv's, whereas for stochastic processes, common-sense intuition causes confusion much more frequently than with rv's. The less familiar word *stochastic* warns the reader to be more careful.

[16]This definition is deliberately vague, and the choice of whether to call a sequence of rv's a process or a sequence is a matter of custom and choice.

sample paths of the process. Conceptually, then, each event is a collection of sample paths. Often these events are defined in terms of a finite set of rv's.

As an example of sample paths that vary at only discrete times, we might be concerned with the times at which customers arrive at some facility. These 'customers' might be customers entering a store, incoming jobs for a computer system, arriving packets to a communication system, or orders for a merchandising warehouse.

The Bernoulli process is an example of how such customers could be modeled and is perhaps the simplest non-trivial stochastic process. We define this process here and develop a few of its many properties. We will frequently return to it, both to use it as an example and to develop additional properties.

Example 1.3.1. A *Bernoulli process* is a sequence, $Y_1, Y_2, \ldots$, of IID binary random variables.[17] Let $p = \Pr\{Y_i = 1\}$ and $1 - p = \Pr\{Y_i = 0\}$. We usually visualize a Bernoulli process as evolving in discrete time with the event $\{Y_i = 1\}$ representing an arriving customer at time i and $\{Y_i = 0\}$ representing no arrival. Thus at most one arrival occurs at each integer time. We visualize the process as starting at time 0, with the first opportunity for an arrival at time 1.

When viewed as arrivals in time, it is interesting to understand something about the intervals between successive arrivals, and about the aggregate number of arrivals up to any given time (see Figure 1.2). These interarrival times and aggregate numbers of arrivals are rv's that are functions of the underlying sequence $Y_1, Y_2, \ldots$,. The topic of rv's that are defined as functions of other rv's (*i.e.*, whose sample values are functions of the sample values of the other rv's) is taken up in more generality in Section 1.3.7, but the interarrival times and aggregate arrivals for Bernoulli processes are so specialized and simple that it is better to treat them from first principles.

First, consider the first interarrival time, X_1, which is defined as the time of the first arrival. If $Y_1 = 1$, then (and only then) $X_1 = 1$. Thus $\mathsf{p}_{X_1}(1) = p$. Next, $X_1 = 2$ if and only $Y_1 = 0$ and $Y_2 = 1$, so $\mathsf{p}_{X_1}(2) = pq$. Continuing, we see that X_1 has the *geometric* PMF,

$$\mathsf{p}_{X_1}(j) = p(1-p)^{j-1}.$$

Each subsequent interarrival time X_k can be found in this same way.[18] It has the same geometric PMF and is statistically independent of $X_1, \ldots, X_{k-1}$. Thus the sequence of interarrival times is an IID sequence of geometric rv's.

It can be seen from Figure 1.2 that a sample path of interarrival times also determines a sample path of the binary arrival rv's, $\{Y_i;\ i \geq 1\}$. Thus the Bernoulli process can also be characterized as a sequence of IID geometric rv's.

For our present purposes, the most important rv's in a Bernoulli process are the partial sums $S_n = \sum_{i=1}^{n} Y_i$. Each rv S_n is the number of arrivals up to and includng time n, *i.e.*,

[17] We say that a sequence $Y_1, Y_2, \ldots$, of rv's are IID if for each integer n, the rv's $Y_1, \ldots, Y_n$ are IID. There are some subtleties in going to the limit $n \to \infty$, but we can avoid most such subtleties by working with finite n-tuples and going to the limit at the end.

[18] This is one of those maddening arguments that, while intuitively obvious, requires some careful reasoning to be completely convincing. We go through several similar arguments with great care in Chapter 2, and suggest that skeptical readers wait until then to prove this rigorously.

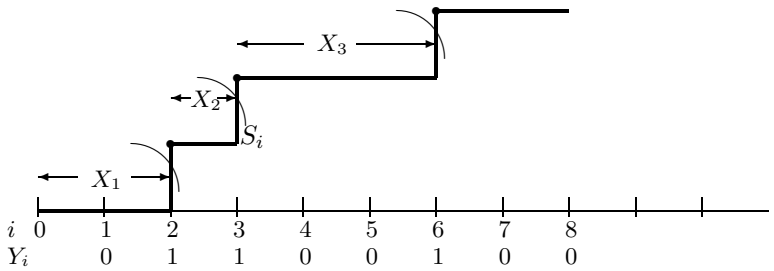

Figure 1.2: Illustration of a sample path for a Bernoulli process: The sample values of the binary rv's Y_i are shown below the time instants. The sample value of the aggregate number of arrivals, $S_n = \sum_{i=1}^{n} Y_i$, is the illustrated step function, and the interarrival intervals are the intervals between steps.

S_n is simply the sum of n binary rv's and thus has the binomial distribution. The PMF $\mathsf{p}_{S_n}(k)$ is the probability that k out of n of the Y_i's have the value 1. There are $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ arrangements of n binary numbers with k 1's, and each has probability $p^k q^{n-k}$. Thus

$$\mathsf{p}_{S_n}(k) = \binom{n}{k} p^k q^{n-k}. \tag{1.22}$$

We will use the binomial PMF extensively as an example in explaining the laws of large numbers later in this chapter, and will often use it in later chapters as an example of a sum of IID rv's. For these examples, we need to know how $\mathsf{p}_{S_n}(k)$ behaves asymptotically as $n \to \infty$ and $k \to \infty$ with k/n essentially constant. The relative frequency k/n will be denoted as $\tilde{p}$. We make a short digression here to state and develop an approximation to the binomial PMF that makes this asymptotic behavior clear.

Lemma 1.3.1. *Let $\mathsf{p}_{S_n}(\tilde{p}n)$ be the PMF of the binomial distribution for an underlying binary PMF $\mathsf{p}_Y(1) = p > 0$, $\mathsf{p}_Y(0) = q > 0$. Then for each integer $\tilde{p}n$, $1 \le \tilde{p}n \le n-1$,*

$$\mathsf{p}_{S_n}(\tilde{p}n)) < \sqrt{\frac{1}{2\pi n\tilde{p}(1-\tilde{p})}}\, \exp\big[n\phi(p,\tilde{p})\big] \qquad \textit{where} \tag{1.23}$$

$$\phi(p,\tilde{p}) = \tilde{p}\ln(\frac{p}{\tilde{p}}) + (1-\tilde{p})\ln(\frac{1-p}{1-\tilde{p}}) \;\le\; 0. \tag{1.24}$$

Also, $\phi(p,\tilde{p}) < 0$ for all $\tilde{p} \neq p$. Finally, for any $\epsilon > 0$, there is an $n(\epsilon)$ such that for $n > n(\epsilon)$,

$$\mathsf{p}_{S_n}(\tilde{p}n) > \left(1 - \frac{1}{\sqrt{n}}\right) \sqrt{\frac{1}{2\pi n\tilde{p}(1-\tilde{p})}}\, \exp\big[n\phi(p,\tilde{p})\big] \qquad \text{for } \epsilon \le \tilde{p} \le 1-\epsilon \tag{1.25}$$

Discussion: The parameter $\tilde{p} = k/n$ is the relative frequency of 1's in the n-tuple $Y_1, \ldots, Y_n$. For each n, $\tilde{p}$ on the left of (1.23) is restricted so that $\tilde{p}n$ is an integer. The lemma then says that $p_{S_n}(\tilde{p}n)$ is upper bounded by an exponentially decreasing function of n for each $\tilde{p} \neq p$.

If $\tilde{p}$ is bounded away from 0 and 1, the ratio of the upper and lower bounds on $p_{S_n}(\tilde{p}n)$ approaches 1 as $n \to \infty$. A bound that is asymptotically tight in this way is denoted as

$$\mathsf{p}_{S_n}(\tilde{p}n)) \sim \sqrt{\frac{1}{2\pi n\tilde{p}(1-\tilde{p})}}\, \exp\left[n\phi(p,\tilde{p})\right] \qquad \text{for } \epsilon < \tilde{p} < 1-\epsilon \tag{1.26}$$

where the symbol $\sim$ means that the ratio of the left and right side approach 1 as $n \to \infty$

Proof*:[19] The factorial of any positive integer n is bounded by the *Stirling* bounds,[20]

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n < n! < \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{1/12n}. \tag{1.27}$$

The ratio $\sqrt{2\pi n}(n/e)^n/n!$ is monotonically increasing with n toward the limit 1, and the ratio $\sqrt{2\pi n}(n/e)^n \exp(1/12n)/n!$ is monotonically decreasing toward 1. The upper bound is more accurate, but the lower bound is simpler and known as the Stirling approximation.

Since $\sqrt{2\pi n}(n/e)^n/n!$ is increasing in n, we see that $n!/k! < \sqrt{n/k}\, n^n k^{-k} e^{-n+k}$ for $k < n$. Combining this with (1.27) applied to $n-k$,

$$\binom{n}{k} < \sqrt{\frac{n}{2\pi k(n-k)}}\, \frac{n^n}{k^k (n-k)^{n-k}}. \tag{1.28}$$

Using (1.28) in (1.22) to upper bound $\mathsf{p}_{S_n}(k)$,

$$\mathsf{p}_{S_n}(k) < \sqrt{\frac{n}{2\pi k(n-k)}}\, \frac{p^k q^{n-k} n^n}{k^k (n-k)^{n-k}}.$$

Replacing k by $\tilde{p}n$, we get (1.23) where $\phi(p,\tilde{p})$ is given by (1.24). Applying the same argument to the right hand inequality in (1.27),

$$\begin{aligned}\binom{n}{k} &> \sqrt{\frac{n}{2\pi k(n-k)}}\, \frac{n^n}{k^k (n-k)^{n-k}} \exp\left(-\frac{1}{12k} - \frac{1}{12(n-k)}\right) \\ &> \sqrt{\frac{n}{2\pi k(n-k)}}\, \frac{n^n}{k^k (n-k)^{n-k}} \left[1 - \frac{1}{12n\tilde{p}(1-\tilde{p})}\right].\end{aligned} \tag{1.29}$$

For $\epsilon < \tilde{p} < 1-\epsilon$, the term in brackets in (1.29) is lower bounded by $1 - 1/(12n\epsilon(1-\epsilon)$, which is further lower bounded by $1 - 1/\sqrt{n}$ for all sufficiently large n, establishing (1.25).

Finally, to show that $\phi(p,\tilde{p}) \leq 0$, with strict inequality for $\tilde{p} \neq p$, we take the first two derivatives of $\phi(p,\tilde{p})$ with respect to $\tilde{p}$.

$$\frac{\partial \phi(p,\tilde{p})}{\partial \tilde{p}} = \ln\left(\frac{p(1-\tilde{p})}{\tilde{p}(1-p)}\right) \qquad \frac{\partial f^2(p,\tilde{p})}{\partial \tilde{p}^2} = \frac{-1}{\tilde{p}(1-\tilde{p})}.$$

Since the second derivative is negative for $0 < \tilde{p} < 1$, the maximum of $\phi(p,\tilde{p})$ with respect to $\tilde{p}$ is 0, achieved at $\tilde{p} = p$. Thus $\phi(p,\tilde{p}) < 0$ for $\tilde{p} \neq p$. Furthermore, $\phi(p,\tilde{p})$ decreases as $\tilde{p}$ moves in either direction away from p. □

Various aspects of this lemma will be discussed later with respect to each of the laws of large numbers.

[19] Proofs with an asterisk can be omitted without an essential loss of continuity

[20] See Feller [7] for a derivation of these results about the Stirling bounds. Feller also shows that that an improved lower bound to $n!$ is given by $\sqrt{2\pi n}(n/e)^n \exp[\frac{1}{12n} - \frac{1}{360n^3}]$.

We have seen that the Bernoulli process can also be characterized as a sequence of IID geometric interarrival intervals. An interesting generalization of this arises by allowing the interarrival intervals to be arbitrary discrete or continuous nonnegative IID rv's rather than geometric rv's. These processes are known as *renewal processes* and are the topic of Chapter 3. Poisson processes are special cases of renewal processes in which the interarrival intervals have an exponential PDF. These are treated in Chapter 2 and have many connections to Bernoulli 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 paths 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.

Discrete stochastic processes find wide and diverse applications in operations research, communication, control, computer systems, management science, finance, etc. Paradoxically, we shall spend relatively little of our time discussing these particular applications, and rather develop results and insights about these processes in general. Many examples drawn from the above fields will be discussed, but the examples will be simple, avoiding many of the complications that require a comprehensive understanding of the application area itself.

1.3.6 Expectation

The *expected value* $\mathsf{E}[X]$ of a random variable X is also called the *expectation* or the *mean* and is frequently denoted as $\overline{X}$. Before giving a general definition, we discuss several special cases. First consider nonnegative discrete rv's. The expected value $\mathsf{E}[X]$ is then given by

$$\mathsf{E}[X] = \sum_x x\, \mathsf{p}_X(x). \tag{1.30}$$

If X has a finite number of possible sample values, the above sum must be finite since each sample value must be finite. On the other hand, if X has a countable number of nonnegative sample values, the sum in (1.30) might be either finite or infinite. Example 1.3.2 illustrates a case in which the sum is infinite. The expectation is said to *exist* only if the sum is finite (*i.e.*, if the sum converges to a real number), and in this case $\mathsf{E}[X]$ is given by (1.30). If the sum is infinite, we say that $\mathsf{E}[X]$ does not exist, but also say[21] that $\mathsf{E}[X] = \infty$. In other words, (1.30) can be used in both cases, but $\mathsf{E}[X]$ is said to *exist* only if the sum is finite.

Example 1.3.2. This example will be useful frequently in illustrating rv's that have an infinite expectation. Let N be a positive integer-valued rv with the distribution function

[21] It almost seems metaphysical to say that something has the value infinity when it doesn't exist. However, the word 'exist' here is shorthand for 'exist as a real number,' which makes it quite reasonable to also consider the value in the extended real number system, which includes $\pm\infty$.

$\mathsf{F}_N(n) = n/(n+1)$ for each integer $n \geq 1$. Then N is clearly a positive rv since $\mathsf{F}_N(0) = 0$ and $\lim_{N\to\infty} \mathsf{F}_N(n) = 1$. For each $n \geq 1$, the PMF is given by

$$\mathsf{p}_N(n) = \mathsf{F}_N(n) - \mathsf{F}_N(n-1) = \frac{n}{n+1} - \frac{n-1}{n} = \frac{1}{n(n+1)}. \tag{1.31}$$

Since $\mathsf{p}_N(n)$ is a PMF, we see that $\sum_{n=1}^{\infty} 1/[n(n+1)] = 1$, which is a frequently useful sum. The following equation, however, shows that $\mathsf{E}[N]$ does not exist and has infinite value.

$$\mathsf{E}[N] = \sum_{n=1}^{\infty} n\,\mathsf{p}_N(n) = \sum_{n=1}^{\infty} \frac{n}{n(n+1)} = \sum_{n=1}^{\infty} \frac{1}{n+1} = \infty,$$

where we have used the fact that the harmonic series diverges.

We next derive an alternative expression for the expected value of a nonnegative discrete rv. This new expression is given directly in terms of the distribution function. We then use this new expression as a general definition of expectation which applies to all nonnegative rv's, whether discrete, continuous, or arbitrary. It contains none of the convergence questions that could cause confusion for arbitrary rv's or for continuous rv's with very wild densities.

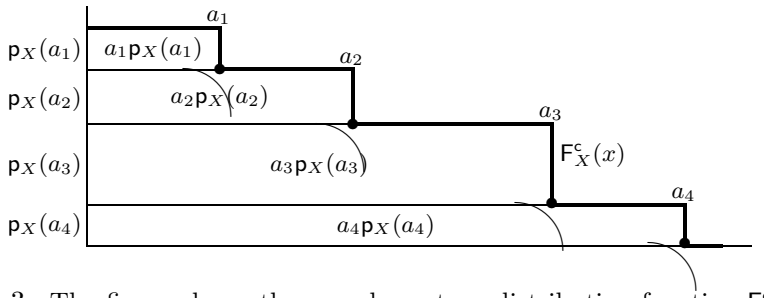

Figure 1.3: The figure shows the complementary distribution function $\mathsf{F}^{\mathsf{c}}_X$ of a nonnegative discrete rv X. For this example, X takes on four possible values, $0 < a_1 < a_2 < a_3 < a_4$. Thus $\mathsf{F}^{\mathsf{c}}_X(x) = \Pr\{X > x\} = 1$ for $x < a_1$. For $a_1 \leq x < a_2$, $\Pr\{X > x\} = 1 - \mathsf{p}_X(a_1)$, and $\Pr\{X > x\}$ has similar drops as x reaches a_2, a_3, and a_4. $\mathsf{E}[X]$, from (1.30), is $\sum_i a_i \mathsf{p}_X(a_i)$, which is the sum of the rectangles in the figure. This is also the area under the curve $\mathsf{F}^{\mathsf{c}}_X(x)$, *i.e.*, $\int_0^\infty \mathsf{F}^{\mathsf{c}}_X(x)\,dx$. It can be seen that this argument applies to any nonnegative rv, thus verifying (1.32).

For a nonnegative discrete rv X, Figure 1.3 illustrates that (1.30) is simply the integral of the complementary distribution function, where the *complementary distribution function* F^{c} of a rv is defined as $\mathsf{F}^{\mathsf{c}}_X(x) = \Pr\{X > x\} = 1 - \mathsf{F}_X(x)$.

$$\mathsf{E}[X] = \int_0^\infty \mathsf{F}^{\mathsf{c}}_X\,dx = \int_0^\infty \Pr\{X > x\}\,dx. \tag{1.32}$$

Although Figure 1.3 only illustrates the equality of (1.30) and (1.32) for one special case, one easily sees that the argument applies to any nonnegative discrete rv, including those with countably many values, by equating the sum of the indicated rectangles with the integral.

For a continuous nonnegative rv X, the conventional definition of expectation is given by

$$\mathsf{E}[X] = \lim_{b\to\infty} \int_0^b x \mathsf{f}_X(x)\, dx. \tag{1.33}$$

Suppose the integral is viewed as a limit of Riemann sums. Each Riemann sum can be viewed as the expectation of a discrete approximation to the continuous rv. The corresponding expectation of the approximation is given by (1.32) using the approximate F_X. Thus (1.32), using the true F_X, yields the expected value of X. This can also be seen using integration by parts. There are no mathematical subtleties in integrating an arbitrary nonnegative nonincreasing function, and this integral must have either a finite or infinite limit. This leads us to the following fundamental definition of expectation for nonnegative rv's:

Definition 1.3.6. *The expectation $\mathsf{E}[X]$ of a nonnegative rv X is defined by (1.32). The expectation is said to exist if and only if the integral is finite. Otherwise the expectation is said to not exist and is also said to be infinite.*

Next consider rv's with both positive and negative sample values. If X has a finite number of positive and negative sample values, say $a_1, a_2, \ldots, a_n$ the expectation $\mathsf{E}[X]$ is given by

$$\begin{aligned}\mathsf{E}[X] &= \sum_i a_i \mathsf{p}_X(a_i) \\ &= \sum_{a_i\le 0} a_i\, \mathsf{p}_X(a_i) + \sum_{a_i>0} a_i\, \mathsf{p}_X(a_i).\end{aligned} \tag{1.34}$$

If X has a countably infinite set of sample values, then (1.34) can still be used if each of the sums in (1.34) converges to a finite value, and otherwise the expectation does not exist (as a real number). It can be seen that each sum in (1.34) converges to a finite value if and only if $\mathsf{E}[|X|]$ exists (*i.e.*, converges to a finite value) for the nonnegative rv $|X|$.

If $\mathsf{E}[X]$ does not exist (as a real number), it still might have the value ∞ if the first sum converges and the second does not, or the value $-\infty$ if the second sum converges and the first does not. If both sums diverge, then $\mathsf{E}[X]$ is undefined, even as an extended real number. In this latter case, the partial sums can be arbitrarily small or large depending on the order in which the terms of (1.34) are summed (see Exercise 1.7).

As illustrated for a finite number of sample values in Figure 1.4, the expression in (1.34) can also be expressed directly in terms of the distribution function and complementary distribution function as

$$\mathsf{E}[X] = -\int_{-\infty}^0 \mathsf{F}_X(x)\, dx + \int_0^\infty \mathsf{F}^{\mathsf{c}}_X(x)\, dx. \tag{1.35}$$

Since $\mathsf{F}^{\mathsf{c}}_X(x) = 1 - \mathsf{F}_X(x)$, this can also be expressed as

$$\mathsf{E}[X] = \int_{-\infty}^\infty \big[u(x) - \mathsf{F}_X(x)\big]\, dx$$

where $u(x)$ is the unit step, $u(x) = 1$ for $x \ge 0$ and $u(x) = 0$ otherwise.

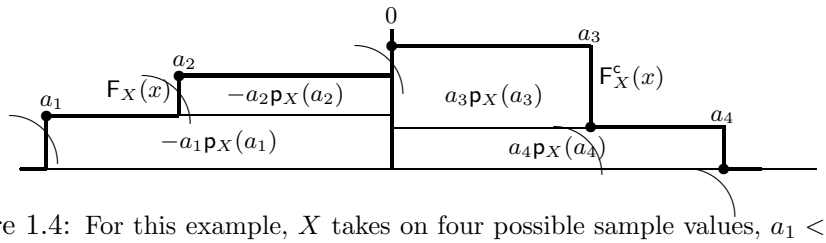

Figure 1.4: For this example, X takes on four possible sample values, $a_1 < a_2 < 0 < a_3 < a_4$. The figure plots $\mathsf{F}_X(x)$ for $x \leq 0$ and $\mathsf{F}^{\mathsf{c}}_X(x)$ for $x > 0$. As in Figure 1.3, $\int_{x\geq 0} \mathsf{F}^{\mathsf{c}}_X(x)\,dx = a_3\mathsf{f}_X(a_3)+a_4\mathsf{f}_X(a_4)$. Similarly, $\int_{x<0} \mathsf{F}_X(x)\,dx = -a_1\mathsf{f}_X(a_1)-a_2\mathsf{f}_X(a_2)$.

The first integral in (1.35) corresponds to the negative sample values and the second to the positive sample values, and $\mathsf{E}[X]$ exists if and only if both integrals are finite (*i.e.*, if $\mathsf{E}[|X|]$ is finite.

For continuous valued rv's with positive and negative sample values, the conventional definition of expectation (assuming that $\mathsf{E}[|X|]$ exists) is given by

$$\mathsf{E}[X] = \int_{-\infty}^{\infty} x\,\mathsf{f}_X(x)\,dx. \tag{1.36}$$

This is equal to (1.35) by the same argument as with nonnegative rv's. Also, as with nonnegative rv's, (1.35) also applies to arbitrary rv's. We thus have the following fundamental definition of expectation:

Definition 1.3.7. *The expectation $\mathsf{E}[X]$ of a rv X exists, with the value given in (1.35), if each of the two terms in (1.35) is finite. The expectation does not exist, but has value ∞ ($-\infty$), if the first term is finite (infinite) and the second infinite (finite). The expectation does not exist and is undefined if both terms are infinite.*

We should not view the general expression in (1.35) for expectation as replacing the need for the conventional expressions in (1.36) and (1.34) for continuous and discrete rv's respectively. We will use all of these expressions frequently, using whichever is most convenient. The main advantages of (1.35) are that it applies equally to all rv's and that it poses no questions about convergence.

Example 1.3.3. The *Cauchy* rv X is the classic example of a rv whose expectation does not exist and is undefined. The probability density is $\mathsf{f}_X(x) = \frac{1}{\pi(1+x^2)}$. Thus $x\mathsf{f}_X(x)$ is proportional to $1/x$ both as $x \to \infty$ and as $x \to -\infty$. It follows that $\int_0^\infty x\mathsf{f}_X(x)\,dx$ and $\int_{-\infty}^0 -x\mathsf{f}_X(x)\,dx$ are both infinite. On the other hand, we see from symmetry that the Cauchy principal value of the integral in (1.36) is given by

$$\lim_{A\to\infty} \int_{-A}^{A} \frac{x}{\pi(1+x)^2}\,dx = 0.$$

There is usually little motivation for considering the upper and lower limits of the integration to have the same magnitude, and the Cauchy principal value usually has little significance for expectations.

1.3.7 Random variables as functions of other random variables

Random variables (rv's) are often defined in terms of each other. For example, if h is a function from $\mathbb{R}$ to $\mathbb{R}$ and X is a rv, then $Y = h(X)$ is the random variable that maps each sample point ω to the composite function $h(X(\omega))$. The distribution function of Y can be found from this, and the expected value of Y can then be evaluated by (1.35).

It is often more convenient to find $\mathsf{E}[Y]$ directly using the distribution function of X. Exercise 1.16 indicates that $\mathsf{E}[Y]$ is given by $\int h(x)\mathsf{f}_X(x)\,dx$ for continuous rv's and by $\sum_x h(x)\mathsf{p}_X(x)$ for discrete rv's. In order to avoid continuing to use separate expressions for continuous and discrete rv's, we express both of these relations by

$$\mathsf{E}[Y] = \int_{-\infty}^{\infty} h(x)\,d\mathsf{F}_X(x), \tag{1.37}$$

This is known as a Stieltjes integral, which can be used as a generalization of both the continuous and discrete cases. For most purposes, we use Stieltjes integrals[22] as a notational shorthand for either $\int h(x)\mathsf{f}_X(x)\,dx$ or $\sum_x h(x)\mathsf{p}_X(x)$.

Knowing that $\mathsf{E}[X]$ exists does not guarantee that $\mathsf{E}[Y]$ exists, but we will treat the question of existence as it arises rather than attempting to establish any general rules.

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

$$\sigma_X^2 = \mathsf{E}\left[(X-\overline{X})^2\right] = \mathsf{E}\left[X^2\right] - \overline{X}^2. \tag{1.38}$$

The *standard deviation* σ_X of X is the square root of the variance and provides a measure of dispersion of the rv around the mean. Thus the mean is a rough measure of typical values for the outcome of the rv, and σ_X is a measure of the typical difference between X and $\overline{X}$. There are other measures of typical value (such as the 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 1.21) is that $\mathsf{E}\left[(X-a)^2\right]$ is minimized over α when $\alpha = \mathsf{E}[X]$.

Next suppose X and Y are rv's and consider the rv[23] $Z = X + Y$. If we assume that X

[22] More specifically, the Riemann-Stieltjes integral, abbreviated here as the Stieltjes integral, is denoted as $\int_a^b h(x)dF_X(x)$. This integral is defined as the limit of a generalized Riemann sum, $\lim_{\delta\to 0}\sum_n h(x_n)[\mathsf{F}(y_n) - \mathsf{F}(y_{n-1})]$ where $\{y_n; n \geq 1\}$ is a sequence of increasing numbers from a to b satisfying $y_n - y_{n-1} \leq \delta$ and $y_{n-1} < x_n \leq y_n$ for all n. The Stieltjes integral exists over finite limits if the limit exists and is independent of the choices of $\{y_n\}$ and $\{x_n\}$ as $\delta \to 0$. It exists over infinite limits if it exists over finite lengths and a limit over the integration limits can be taken. See Rudin [18] for an excellent elementary treatment of Stieltjes integration, and see Exercise 1.12 for some examples.

[23] The question whether a real-valued function of rv's is itself a rv is usually addressed by the use of measure theory, and since we neither use nor develop measure theary in this text, we usually simply assume (within the limits of common sense) that any such function is itself a rv. However, the sum $X + Y$ of rv's is so important throughout this subject that Exercise 1.10 provides a guided derivation of this result for $X + Y$. In the same way, the sum $S_n = X_1 + \cdots + X_n$ of any finite collection of rv's is also a rv.

and Y are independent, then the distribution function of $Z = X + Y$ is given by[24]

$$\mathsf{F}_Z(z) = \int_{-\infty}^{\infty} \mathsf{F}_X(z-y)\, d\mathsf{F}_Y(y) = \int_{-\infty}^{\infty} \mathsf{F}_Y(z-x)\, d\mathsf{F}_X(x). \tag{1.39}$$

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

$$\mathsf{f}_Z(z) = \int_{-\infty}^{\infty} \mathsf{f}_X(z-y)\mathsf{f}_Y(y)\, dy = \int_{-\infty}^{\infty} \mathsf{f}_Y(z-x)\mathsf{f}_X(x)\, dx. \tag{1.40}$$

Eq. (1.40) is the familiar convolution equation from linear systems, and we similarly refer to (1.39) as the convolution of distribution functions (although it has a different functional form from (1.40)). If X and Y are nonnegative random variables, then the integrands in (1.39) and (1.40) are non-zero only between 0 and z, so we often use 0 and z as the limits in (1.39) and (1.40).

If $X_1, X_2, \ldots, X_n$ are independent rv's, then the distribution of the rv $S_n = X_1 + X_2 + \cdots + X_n$ 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 $S_n = X_1 + X_2 + \cdots + X_n$ satisfies

$$\mathsf{E}[S_n] = \mathsf{E}[X_1 + X_2 + \cdots + X_n] = \mathsf{E}[X_1] + \mathsf{E}[X_2] + \cdots + \mathsf{E}[X_n]. \tag{1.41}$$

This says that the expected value of a sum is equal to the sum of the expected values, whether or not the rv's are independent (see exercise 1.11). The following example shows how this can be a valuable problem solving aid with an appropriate choice of rv's.

Example 1.3.4. Consider a switch with n input nodes and n output nodes. Suppose each input is randomly connected to a single output in such a way that each output is also connected to a single input. That is, each output is connected to input 1 with probability $1/n$. Given this connection, each of the remaining outputs are connected to input 2 with probability $1/(n-1)$, and so forth.

An input node is said to be *matched* if it is connected to the output of the same number. We want to show that the expected number of matches (for any given n) is 1. Note that the first node is matched with probability $1/n$, and therefore the expectation of a match for node 1 is $1/n$. Whether or not the second input node is matched depends on the choice of output for the first input node, but it can be seen from symmetry that the *marginal distribution* for the output node connected to input 2 is $1/n$ for each output. Thus the expectation of a match for node 2 is also $1/n$. In the same way, the expectation of a match for each input node is $1/n$. From (1.41), the expected total number of matchs is the sum over the expected number for each input, and is thus equal to 1. This exercise would be quite difficult without the use of (1.41).

[24] See Exercise 1.12 for some peculiarties about this definition.

If the rv's $X_1, \ldots, X_n$ are independent, then, as shown in exercises 1.11 and 1.18, the variance of $S_n = X_1 + \cdots + X_n$ is given by

$$\sigma^2_{S_n} = \sum_{i=1}^{n} \sigma^2_{X_i}. \tag{1.42}$$

If $X_1, \ldots, X_n$ are also identically distributed (*i.e.*, $X_1, \ldots, X_n$ are IID) with variance σ^2_X, then $\sigma^2_{S_n} = n\sigma^2_X$. Thus the standard deviation of S_n is $\sigma_{S_n} = \sqrt{n}\sigma_X$. Sums of IID rv's appear everywhere in probability theory and play an especially central role in the laws of large numbers. It is important to remember that the mean of S_n is linear in n but the standard deviation increases only with the square root of n. Figure 1.5 illustrates this behavior.

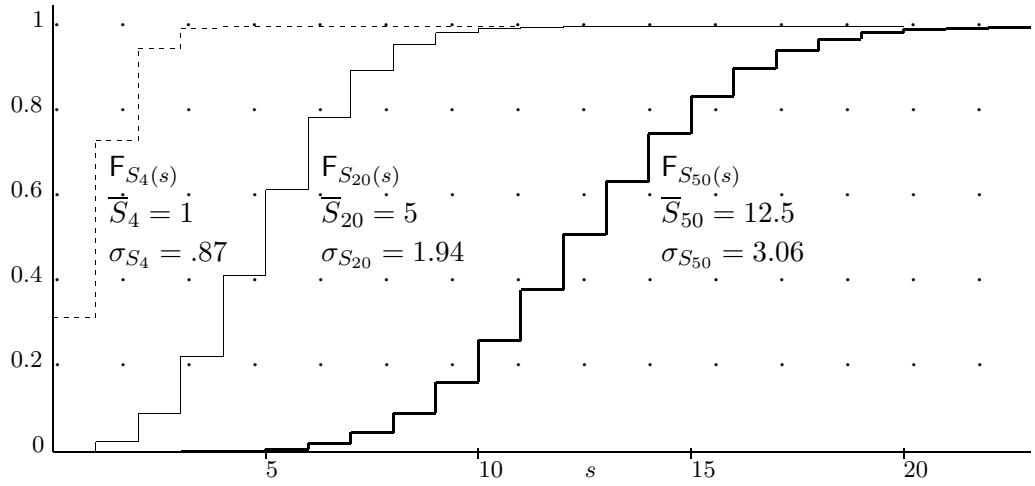

Figure 1.5: The distribution function $\mathsf{F}_{S_n}(s)$ of $S_n = X_1 + \cdots + X_n$ where $X_1, \ldots, X_n$ are typical IID rv's and n takes the values 4, 20, and 50. The particular rv in the figure is binary with $\mathsf{p}_X(1) = 1/4, \mathsf{p}_X(0) = 3/4$. Note that the mean of S_n is proportional to n and the standard deviation to $\sqrt{n}$.

1.3.8 Conditional expectations

Just as the *conditional distribution* of one rv conditioned on a sample value of another rv is important, the *conditional expectation* of one rv based on the sample value of another is equally important. Initially let X be a positive discrete rv and let y be a sample value of another discrete rv Y such that $\mathsf{p}_Y(y) > 0$. Then the conditional expectation of X given $Y = y$ is defined to be

$$\mathsf{E}\,[X \mid Y{=}y] = \sum_x x\, \mathsf{p}_{X|Y}(x \mid y). \tag{1.43}$$

This is simply the ordinary expected value of X using the conditional probabilities in the reduced sample space corresponding to $Y = y$. This value can be finite or infinite as before. More generally, if X can take on positive or negative values, then there is the possibility that

the conditional expectation is undefined. In other words, for discrete rv's, the conditional expectation is exactly the same as the ordinary expectation, except that it is taken using conditional probabilities over the reduced sample space.

More generally yet, let X be an arbitrary rv and let y be a sample value of a discrete rv Y with $\mathsf{p}_Y(y) > 0$. The conditional distribution function of X conditional on $Y = y$ is defined as

$$\mathsf{F}_{X|Y}(x \mid y) = \frac{\Pr\{X \le x, Y = y\}}{\Pr\{Y = y\}}.$$

Since this is an ordinary distribution function in the reduced sample space where $Y = y$, (1.35) expresses the expectation of X conditional on $Y = y$ as

$$\mathsf{E}[X \mid Y = y] = -\int_{-\infty}^{0} \mathsf{F}_{X|Y}(x \mid y)\, dx + \int_{0}^{\infty} \mathsf{F}_{X|Y}(x \mid y)\, dx. \tag{1.44}$$

The forms of conditional expectation in (1.43) and (1.44) are given for individual sample values of Y for which $\mathsf{p}_Y(y) > 0$.

We next show that the conditional expectation of X conditional on a discrete rv Y can also be viewed as a rv. With the possible exception of a set of zero probability, each $\omega \in \Omega$ maps to $\{Y = y\}$ for some y with $\mathsf{p}_Y(y) > 0$ and $\mathsf{E}[X \mid Y = y]$ is defined for that y. Thus we can define $\mathsf{E}[X \mid Y]$ as[25] a rv that is a function of Y, mapping ω to a sample value, say y of Y, and mapping that y to $\mathsf{E}[X \mid Y = y]$. Regarding a conditional expectation as a rv that is a function of the conditioning rv is a powerful tool both in problem solving and in advanced work. For now, we use this to express the unconditional mean of X as

$$\mathsf{E}[X] = \mathsf{E}\big[\mathsf{E}[X \mid Y]\big], \tag{1.45}$$

where the inner expectation is over X for each value of Y and the outer expectation is over the rv $\mathsf{E}[X \mid Y]$, which is a function of Y.

Example 1.3.5. Consider rolling two dice, say a red die and a black die. Let X_1 be the number on the top face of the red die, and X_2 that for the black die. Let $S = X_1 + X_2$. Thus X_1 and X_2 are IID integer rv's, each uniformly distributed from 1 to 6. Conditional on $S = j$, X_1 is uniformly distributed between 1 and $j - 1$ for $j \le 7$ and between $j - 6$ and 6 for $j \ge 7$. For each $j \le 7$, it follows that $\mathsf{E}[X_1 \mid S = j] = j/2$. Similarly, for $j \ge 7$, $\mathsf{E}[X_1 \mid S = j] = j/2$. This can also be seen by the symmetry between X_1 and X_2.

The rv $\mathsf{E}[X_1 \mid S]$ is thus a discrete rv taking on values from 1 to 6 in steps of 1/2 as the sample value of S goes from 2 to 12. The PMF of $\mathsf{E}[X_1 \mid S]$ is given by $\mathsf{p}_{\mathsf{E}[X_1|S]}(j/2) = \mathsf{p}_S(j)$. Using (1.45), we can then calculate $\mathsf{E}[X_1]$ as

$$\mathsf{E}[X_1] = \mathsf{E}\big[\mathsf{E}[X_1 \mid S]\big] = \sum_{j=2}^{12} \frac{j}{2}\, \mathsf{p}_S(j) = \frac{\mathsf{E}[S]}{2} = \frac{7}{2}.$$

This example is not intended to show the value of (1.45) in calculating expectation, since $\mathsf{E}[X_1] = 7/2$ is initially obvious from the uniform integer distribution of X_1. The purpose is simply to illustrate what the rv $\mathsf{E}[X_1 \mid S]$ means.

[25] This assumes that $\mathsf{E}[X \mid Y = y]$ is finite for each y, which is one of the reasons that expectations are said to exist only if they are finite.

To illustrate (1.45) in a more general way, while still assuming X to be discrete, we can write out this expectation by using (1.43) for $\mathsf{E}[X \mid Y=y]$.

$$\begin{aligned}\mathsf{E}[X] &= \mathsf{E}\big[\mathsf{E}[X \mid Y]\big] = \sum_y \mathsf{p}_Y(y)\mathsf{E}[X \mid Y=y] \\ &= \sum_y \mathsf{p}_Y(y) \sum_x x\, \mathsf{p}_{X|Y}(x|y). \end{aligned} \tag{1.46}$$

Operationally, there is nothing very fancy in the example or in (1.45). Combining the sums, (1.46) simply says that $\mathsf{E}[X] = \sum_{y,x} x\, \mathsf{p}_{YX}(y,x)$. As a concept, however, viewing the conditional expectation $\mathsf{E}[X \mid Y]$ as a rv based on the conditioning rv Y is often a useful theoretical tool. This approach is equally useful as a tool in problem solving, since there are many problems where it is easy to find conditional expectations, and then to find the total expectation by averaging over the conditioning variable. For this reason, this result is sometimes called either the total expectation theorem or the iterated expectation theorem. Exercise 1.17 illustrates the advantages of this approach, particularly where it is initially unclear whether or not the expectation is finite. The following cautionary example, however, shows that this approach can sometimes hide convergence questions and give the wrong answer.

Example 1.3.6. Let Y be a geometric rv with the PMF $\mathsf{p}_Y(y) = 2^{-y}$ for integer $y \geq 1$. Let X be an integer rv that, conditional on Y, is binary with equiprobable values $\pm 2^y$ given $Y = y$. We then see that $\mathsf{E}[X \mid Y=y] = 0$ for all y, and thus, (1.46) indicates that $\mathsf{E}[X] = 0$. On the other hand, it is easy to see that $\mathsf{p}_X(2^k) = \mathsf{p}_X(-2^k) = 2^{-k-1}$ for each integer $k \geq 1$. Thus the expectation over positive values of X is ∞ and that over negative values is $-\infty$. In other words, the expected value of X is undefined and (1.46) is incorrect.

The difficulty in the above example cannot occur if X is a nonnegative rv. Then (1.46) is simply a sum of a countable number of nonnegative terms, and thus it either converges to a finite sum independent of the order of summation, or it diverges to ∞, again independent of the order of summation.

If X has both positive and negative components, we can separate it into $X = X^+ + X^-$ where $X^+ = \max(0, X)$ and $X^- = \min(X, 0)$. Then (1.46) applies to X^+ and $-X^-$ separately. If at most one is infinite, then (1.46) applies to X, and otherwise X is undefined. This is summarized in the following theorem:

Theorem 1.3.1 (Total expectation). *Let X and Y be discrete rv's. If X is nonnegative, then $\mathsf{E}[X] = \mathsf{E}\big[\mathsf{E}[X \mid Y]\big] = \sum_y \mathsf{p}_Y(y)\mathsf{E}[X \mid Y=y]$. If X has both positive and negative values, and if at most one of $\mathsf{E}[X^+]$ and $\mathsf{E}[-X^-]$ is infinite, then $\mathsf{E}[X] = \mathsf{E}\big[\mathsf{E}[X \mid Y]\big] = \sum_y \mathsf{p}_Y(y)\mathsf{E}[X \mid Y=y]$.*

We have seen above that if Y is a discrete rv, then the conditional expectation $\mathsf{E}[X|Y=y]$ is little more complicated than the unconditional expectation, and this is true whether X is discrete, continuous, or arbitrary. If X and Y are continuous, we can essentially extend these results to probability densities. In particular, defining $\mathsf{E}[X \mid Y=y]$ as

$$\mathsf{E}[X \mid Y=y] = \int_{-\infty}^{\infty} x\, \mathsf{f}_{X|Y}(x \mid y), \tag{1.47}$$

we have

$$\mathsf{E}[X] = \int_{-\infty}^{\infty} \mathsf{f}_Y(y)\mathsf{E}[X \mid Y{=}y]\, dy = \int_{-\infty}^{\infty} \mathsf{f}_Y(y) \int_{-\infty}^{\infty} x\, \mathsf{f}_{X|Y}(x \mid y)\, dx\, dy. \tag{1.48}$$

We do not state this as a theorem because the details about the integration do not seem necessary for the places where it is useful.

1.3.9 Indicator random variables

For any event A, the *indicator random variable* of A, denoted $\mathbb{I}_A$, is a binary rv that has the value 1 for all $\omega \in A$ and the value 0 otherwise. It then has the PMF $\mathsf{p}_{\mathbb{I}_A}(1) = \Pr\{A\}$ and $\mathsf{p}_{\mathbb{I}_A}(0) = 1 - \Pr\{A\}$. The corresponding distribution function $\mathsf{F}_{\mathbb{I}_A}$ is then illustrated in Figure 1.6. It is easily seen that $\mathsf{E}[\mathbb{I}_A] = \Pr\{A\}$.

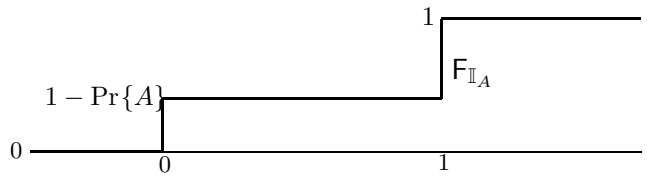

Figure 1.6: The distribution function $\mathsf{F}_{\mathbb{I}_A}$ of an indicator random variable $\mathbb{I}_A$.

Indicator rv's are useful because they allow us to apply the many known results about rv's and particularly binary rv's to events. For example, the laws of large numbers are expressed in terms of sums of rv's, and those results all translate into results about relative frequencies through the use of indicator functions.

1.3.10 Moment generating functions and other transforms

The *moment generating function* (MGF) for a rv X is given by

$$\mathsf{g}_X(r) = \mathsf{E}\left[e^{rX}\right] = \int_{-\infty}^{\infty} e^{rx}\, d\mathsf{F}_X(x). \tag{1.49}$$

where r is a real variable. The integrand is nonnegative, and we can study where the integral exists (*i.e.*, where it is finite) by separating it as follows:

$$\mathsf{g}_X(r) = \int_{0}^{\infty} e^{rx}\, d\mathsf{F}_X(x) + \int_{-\infty}^{0} e^{rx}\, d\mathsf{F}_X(x). \tag{1.50}$$

Both integrals exist for $r = 0$, since the first is $\Pr\{X > 0\}$ and the second is $\Pr\{X \leq 0\}$. The first integral is increasing in r, and thus if it exists for one value of r, it also exists for all smaller values. For example, if X is a nonnegative exponential rv with the density $\mathsf{f}_X(x) = e^{-x}$, then the first integral exists if and only $r < 1$, where it has the value $\frac{1}{1-r}$. As another example, if X satisfies $\Pr\{X > A\} = 0$ for some finite A, then the first integral is at most e^{rA}, which is finite for all real r.

Let $r_+(X)$ be the supremum of values of r for which the first integral exists. The first integral exists for all $r < r_+(X)$, and $0 \leq r_+(X) \leq \infty$. In the same way, let $r_-(X)$ be the infimum of values of r for which the the second integral exists. The second integral exists for all $r > r_-(X)$ and $r_-(X)$ satisfies $0 \geq r_-(X) \geq -\infty$.

Combining the two integrals, the region over which the MGF of an arbitrary rv exists is an interval $I(X)$ from $r_-(X) \leq 0$ to $r_+(X) \geq 0$. Either or both of the end points, $r_-(X)$ and $r_+(X)$, might be included in $I(X)$, and either or both might be either 0 or infinite. We denote these quantities as I, r_-, and r_+ when the rv X is clear from the context. Appendix A gives the interval I for for a number of standard rv's and Exercise 1.22 illustrates $I(X)$ further.

If $\mathsf{g}_X(r)$ exists in an open region of r around 0 (*i.e.*, if $r_- < 0 < r_+$), then derivatives[26] of all orders exist in that region. They are given by

$$\frac{\partial^k \mathsf{g}_X(r)}{\partial r^k} = \int_{-\infty}^{\infty} x^k e^{rx} d\mathsf{F}_X(x) \quad ; \qquad \left.\frac{\partial^k \mathsf{g}_X(r)}{\partial r^k}\right|_{r=0} = \mathsf{E}\left[X^k\right]. \tag{1.51}$$

This shows that finding the moment generating function often provides a convenient way to calculate the moments of a random variable. If any moment fails to exist, however, then the MGF must also fail to exist over each open interval containing 0 (see Exercise 1.31).

Another convenient feature of moment generating functions is their use in treating sums of independent rv's. For example, let $S_n = X_1 + X_2 + \cdots + X_n$. Then

$$\mathsf{g}_{S_n}(r) = \mathsf{E}\left[e^{rS_n}\right] = \mathsf{E}\left[\exp\left(\sum_{i=1}^n rX_i\right)\right] = \mathsf{E}\left[\prod_{i=1}^n \exp(rX_i)\right] = \prod_{i=1}^n \mathsf{g}_{X_i}(r). \tag{1.52}$$

In the last step here, we have used a result of Exercise 1.11, which shows that for independent rv's, the mean of the product is equal to the product of the means. If $X_1, \ldots, X_n$ are also IID, then

$$\mathsf{g}_{S_n}(r) = [\mathsf{g}_X(r)]^n. \tag{1.53}$$

We will use this property frequently in treating sums of IID rv's. Note that this also implies that the region over which the MGF's of S_n and X exist are the same, *i.e.*, $I(S_n) = I(X)$.

The real variable r in the MGF can also be viewed as a complex variable, giving rise to a number of other transforms. A particularly important case is to view r as a pure imaginary variable, say $i\omega$ where $i = \sqrt{-1}$ and ω is real. The MGF is then called the *characteristic function*. Since $|e^{i\omega x}|$ is 1 for all x, $\mathsf{g}_X(i\omega)$ exists for all rv's X and all real ω, and its magnitude is at most one. Note that $\mathsf{g}_X(-i\omega)$ is the Fourier transform of the density of X, so the Fourier transform and characteristic function are the same except for this small notational difference.

The Z-transform is the result of replacing e^r with z in $\mathsf{g}_X(r)$. This is useful primarily for integer valued rv's, but if one transform can be evaluated, the other can be found

[26] This result depends on interchanging the order of differentiation (with respect to r) and integration (with respect to x). This can be shown to be permissible because $\mathsf{g}_X(r)$ exists for r both greater and smaller than 0, which in turn implies, first, that $1 - \mathsf{F}_X(x)$ must approach 0 exponentially as $x \to \infty$ and, second, that $\mathsf{F}_X(x)$ must approach 0 exponentially as $x \to -\infty$.

immediately. 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 rv's. The simplicity of taking products of transforms is a major reason that transforms are so useful in probability theory.

1.4 Basic inequalities

Inequalities play a particularly fundamental role in probability, partly because many of the models we study are too complex to find exact answers, and partly because many of the most useful theorems establish limiting rather than exact results. In this section, we study three related inequalities, the Markov, Chebyshev, and Chernoff bounds. These are used repeatedly both in the next section and in the remainder of the text.

1.4.1 The Markov inequality

This is the simplest and most basic of these inequalities. It states that if a nonnegative random variable Y has a mean $\mathsf{E}[Y]$, then, for every $y > 0$, $\Pr\{Y \geq y\}$ satisfies[27]

$$\Pr\{Y \geq y\} \leq \frac{\mathsf{E}[Y]}{y} \quad \text{Markov Inequality for nonnegative } Y. \tag{1.54}$$

Figure 1.7 derives this result using the fact (see Figure 1.3) that the mean of a nonnegative rv is the integral of its complementary distribution function, i.e., of the area under the curve $\Pr\{Y > z\}$. Exercise 1.28 gives another simple proof using an indicator random variable.

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 half of the population have a height exceeding 3.2 meters. We see from this example that the Markov inequality is often very weak. However, for any $y > 0$, we can consider a rv that takes on the value y with probability ϵ and the value 0 with probability $1 - \epsilon$; this rv satisfies the Markov inequality at the point y with equality. Figure 1.7 (as elaborated in Exercise 1.40) also shows that, for any nonnegative rv Y with a finite mean,

$$\lim_{y \to \infty} y \Pr\{Y \geq y\} = 0. \tag{1.55}$$

This will be useful shortly in the proof of Theorem 1.5.3.

[27] The distribution function of any given rv Y is known (at least in principle), and thus one might question why an upper bound is ever preferable to the exact value. One answer is that Y might be given as a function of many other rv's and that the parameters such as the mean in a bound are much easier to find than the distribution function. Another answer is that such inequalities are often used in theorems which state results in terms of simple statistics such as the mean rather than the entire distribution function. This will be evident as we use these bounds.

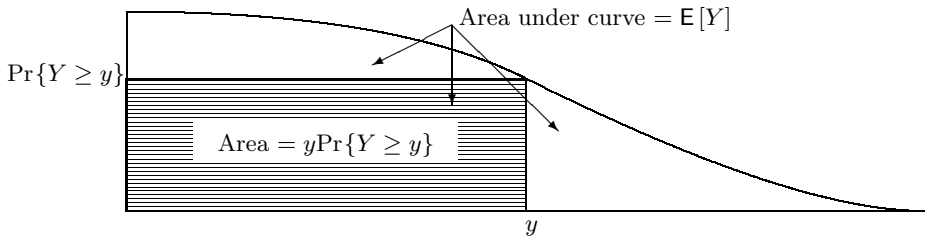

Figure 1.7: Demonstration that $y\Pr\{Y \geq y\} \leq \mathsf{E}[Y]$. By letting $y \to \infty$, it can also be seen that the shaded area becomes a negligible portion of the area $\mathsf{E}[Y]$, so that $\lim_{y\to\infty} y\Pr\{Y > y\} = 0$ if $\mathsf{E}[Y] < \infty$

1.4.2 The Chebyshev inequality

We now use the Markov inequality to establish the well-known Chebyshev inequality. Let Z be an arbitrary rv with finite mean $\mathsf{E}[Z]$ and finite variance σ_Z^2, and define Y as the nonnegative rv $Y = (Z - \mathsf{E}[Z])^2$. Thus $\mathsf{E}[Y] = \sigma_Z^2$. Applying (1.54),

$$\Pr\big\{(Z - \mathsf{E}[Z])^2 \geq y \quad \leq \frac{\sigma_Z^2}{y} \qquad \text{for any } y > 0.$$

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

$$\Pr\{|Z - \mathsf{E}[Z]| \geq \epsilon\} \leq \frac{\sigma_Z^2}{\epsilon^2} \qquad \text{(Chebyshev inequality).} \tag{1.56}$$

Note that the Markov inequality bounds just the upper tail of the distribution function and applies only to nonnegative rv's, 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).

The Chebyshev inequality is particularly useful when Z is the sample average, $(X_1 + X_2 + \cdots + X_n)/n$, of a set of IID rv's. This will be used shortly in proving the weak law of large numbers.

1.4.3 Chernoff bounds

Chernoff (or exponential) bounds are another variation of the Markov inequality in which the bound on each tail of the distribution function goes to 0 exponentially with distance from the mean. For any given rv Z, let $I(Z)$ be the interval over which the MGF $\mathsf{g}_Z(r) = \mathsf{E}\left[e^{Zr}\right]\}$ exists. Letting $Y = e^{Zr}$ for any $r \in I(Z)$, the Markov inequality (1.54) applied to Y is

$$\Pr\{\exp(rZ) \geq y\} \leq \frac{\mathsf{g}_Z(r)}{y} \qquad \text{for any } y > 0.$$

This takes on a more meaningful form if y is replaced by e^{rb}. Note that $\exp(rZ) \geq \exp(rb)$ is equivalent to $Z \geq b$ for $r > 0$ and to $Z < b$ for $r < 0$. Thus, for any real b, we get the following two bounds, one for $r > 0$ and the other for $r < 0$:

$$\begin{aligned} \Pr\{Z \geq b\} &\leq \mathsf{g}_Z(r)\exp(-rb) \quad ; \qquad \text{(Chernoff bound for } 0 < r \in I(Z) \qquad (1.57) \\ \Pr\{Z \leq b\} &\leq \mathsf{g}_Z(r)\exp(-rb) \quad ; \qquad \text{(Chernoff bound for } 0 > r \in I(Z)). \qquad (1.58) \end{aligned}$$

This provides us with a family of upper bounds on the tails of the distribution function, using values of $r > 0$ for the upper tail and $r < 0$ for the lower tail. For fixed $0 < r \in I(Z)$, this bound on $\Pr\{Z \geq b\}$ decreases exponentially[28] in b at rate r. Similarly, for each $0 > r \in I(Z)$, the bound on $\Pr\{Z \leq b\}$ decreases exponentially at rate r as $b \to -\infty$. We will see shortly that (1.57) is useful only when $b > \mathsf{E}[X]$ and (1.58) is useful only when $b < \mathsf{E}[X]$).

The most important application of these Chernoff bounds is to sums of IID rv's. Let $S_n = X_1 + \cdots + X_n$ where $X_1, \ldots, X_n$ are IID with the MGF $\mathsf{g}_X(r)$. Then $\mathsf{g}_{S_n}(r) = [\mathsf{g}_X(r)]^n$, so (1.57) and (1.58) (with b replaced by na) become

$$\begin{aligned} \Pr\{S_n \geq na\} &\leq [\mathsf{g}_X(r)]^n \exp(-rna) \quad ; \qquad (\text{ for } 0 < r \in I(Z)) \qquad (1.59) \\ \Pr\{S_n \leq na\} &\leq [\mathsf{g}_X(r)]^n \exp(-rna) \quad ; \qquad (\text{ for } 0 > r \in I(Z)). \qquad (1.60) \end{aligned}$$

These equations are easier to understand if we define the *semi-invariant MGF*, $\gamma_X(r)$, as

$$\gamma_X(r) = \ln g_X(r). \tag{1.61}$$

The semi-invariant MGF for a typical rv X is sketched in Figure 1.8. The major features to observe are, first, that $\gamma'_X(0) = \mathsf{E}[X]$ and, second, that $\gamma''_X(r) \geq 0$ for r in the interior of $I(X)$..

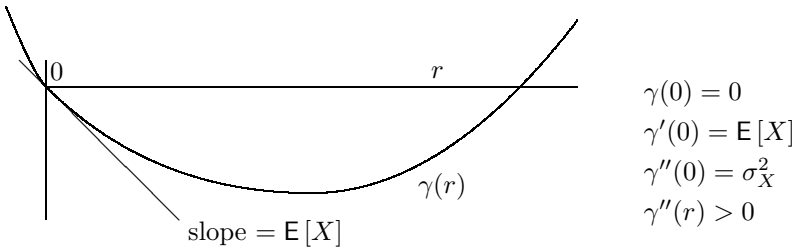

Figure 1.8: Semi-invariant moment-generating function $\gamma(r)$ for a typical rv X assuming $r_- < 0 < r_+$. Since $\gamma(r) = \ln g(r)$, we see that that $\frac{\partial}{\partial r}\gamma(r) = \frac{1}{g(r)}\frac{\partial}{\partial r}g(r)$. Thus $\gamma'(0) = \mathsf{E}[X]$. Also, for r in the interior of $I(X)$, Exercise 1.24 shows that $\gamma''(r) \geq 0$, and in fact, $\gamma''(r)$ is strictly positive except in the uninteresting case where X is deterministic (takes on a single value with probability 1). As indicated in the figure, the straight line of slope $\mathsf{E}[X]$ through the origin is tangent to $\gamma(r)$.

[28] This seems paradoxical, since Z seems to be almost arbitrary. However, since $r \in I(Z)$, we have $\int e^{rb} d\mathsf{F}_Z(b) < \infty$.

In terms of $\gamma_X(r)$, (1.59) and (1.60) become

$$\Pr\{S_n \geq na\} \leq \exp(n[\gamma_X(r) - ra]) \quad ; \quad (\text{ for } 0 < r \in I(X) \tag{1.62}$$

$$\Pr\{S_n \leq na\} \leq \exp(n[\gamma_X(r) - ra]) \quad ; \quad (\text{ for } 0 > r \in I(X). \tag{1.63}$$

These bounds are geometric in n for fixed a and r, so we should ask what value of r provides the tightest bound for any given a. Since $\gamma''_X(r) > 0$, the tightest bound arises either at that r for which $\gamma'(r) = a$ or at one of the end points, r_- or r_+, of $I(X)$. This minimum value is denoted by

$$\mu_X(a) = \inf_r [\gamma_X(r) - ra].$$

Note that $(\gamma_X(r) - ra)|_{r=0} = 0$ and $\frac{\partial}{\partial r}(\gamma_X(r) - ra)|_{r=0} = \mathsf{E}[X] - a$. Thus if $a > \mathsf{E}[X]$, then $\gamma_X(r) - ra$ must be negative for sufficiently small positive r. Similarly, if $a < \mathsf{E}[X]$, then $\gamma_X(r) - ra$ is negative for negative r sufficiently close[29] to 0. In other words,

$$\Pr\{S_n \geq na\} \leq \exp(n\mu_X(a)) \; ; \qquad \text{where } \mu_X(a) < 0 \text{ for } a > \mathsf{E}[X] \tag{1.64}$$

$$\Pr\{S_n \leq na\} \leq \exp(n\mu_X(a)) \; ; \qquad \text{where } \mu_X(a) < 0 \text{ for } a < \mathsf{E}[X]\,. \tag{1.65}$$

This is summarized in the following lemma:

Lemma 1.4.1. *Assume that 0 is in the interior of $I(X)$ and let S_n be the sum of n IID rv's each with the distribution of X. Then $\mu_X(a) = \inf_r[\gamma_X(r) - ra] < 0$ for all $a \neq \mathsf{E}[X]$. Also, $\Pr\{S_n \geq na\} \leq e^{n\mu_X(a)}$ for $a > \mathsf{E}[X]$ and $\Pr\{S_n \leq na\} \leq e^{n\mu_X(a)}$ for $a < \mathsf{E}[X]$.*

Figure 1.9 illustrates the lemma and gives a graphical construction to find[30] $\mu_X(a) = \inf_r[\gamma_X(r) - ra]$.

These Chernoff bounds will be used in the next section to help understand several laws of large numbers. They will also be used extensively in Chapter 7 and are useful for detection, random walks, and information theory.

The following example evaluates these bounds for the case where the IID rv's are binary. We will see that in this case the bounds are exponentially tight in a sense to be described.

Example 1.4.1. Let X be binary with $\mathsf{p}_X(1) = p$ and $\mathsf{p}_X(0) = q = 1 - p$. Then $\mathsf{g}_X(r) = q + pe^r$ for $-\infty < r < \infty$. Also, $\gamma_X(r) = \ln(q + pe^r)$. To be consistent with the expression for the binomial PMF in (1.23), we will find bounds to $\Pr\{S_n \geq \tilde{p}n\}$ and $\Pr\{S_n \leq \tilde{p}n\}$ for $\tilde{p} > p$ and $\tilde{p} < p$ respectively. Thus, according to Lemma 1.4.1, we first evaluate

$$\mu_X(\tilde{p}) = \inf_r [\gamma_X(r) - \tilde{p}r].$$

The minimum occurs at that r for which $\gamma'_X(r) = \tilde{p}$, *i.e.*, at

$$\frac{pe^r}{q + pe^r} = \tilde{p}.$$

[29] In fact, for r sufficiently small, $\gamma(r)$ can be approximated by a second order power series, $\gamma(r) \approx \gamma(0) + r\gamma'(0) + (r^2/2)\gamma''(0) = r\overline{X} + (r^2/2)\sigma_X^2$. It follows that $\mu_X(a) \approx -(a - \overline{X})^2/2\sigma_X^2$ for very small r.

[30] As a special case, the infimum might occur at the edge of the interval of convergence, *i.e.*, at r_- or r_+. As shown in Exercise 1.23, the infimum can be at r_+ (r_-) only if $g_X(r_+)$ ($g_X(r_-)$ exists, and in this case, the graphical techinique in Figure 1.9 still works.

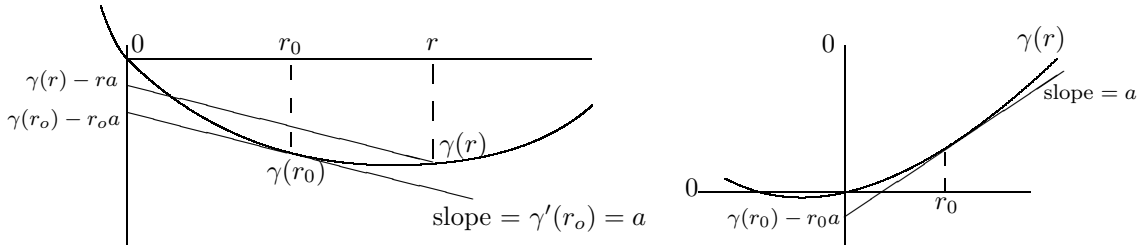

Figure 1.9: Graphical minimization of $\gamma(r) - ar$: For any $r \in I(X)$, $\gamma(r) - ar$ is the vertical axis intercept of a line of slope a through the point $(r, \gamma(r))$. The minimum occurs when the line of slope a is tangent to the curve. The two examples show one case where $\mathsf{E}[X] < 0$ and another where $\mathsf{E}[X] > 0$.

Rearranging terms,

$$e^r = \frac{\tilde{p}q}{p\tilde{q}} \qquad \text{where } \tilde{q} = 1 - \tilde{p}. \tag{1.66}$$

Substituting this minimizing value of r into $\ln(q + pe^r) - r\tilde{p}$ and rearranging terms,

$$\mu_X(\tilde{p}) = \tilde{p} \ln \frac{p}{\tilde{p}} + \tilde{q} \ln \frac{\tilde{q}}{q}. \tag{1.67}$$

Substituting this into (1.64), and (1.65), we get the following Chernoff bounds for binary IID rv's. As shown above, they are exponentially decreasing in n.

$$\Pr\{S_n \geq n\tilde{p}\} \leq \exp\left\{n\left[\tilde{p}\ln\frac{p}{\tilde{p}} + \tilde{q}\ln\frac{q}{\tilde{q}}\right]\right\}; \qquad \text{for } \tilde{p} > p \tag{1.68}$$

$$\Pr\{S_n \leq n\tilde{p}\} \leq \exp\left\{n\left[\tilde{p}\ln\frac{p}{\tilde{p}} + \tilde{q}\ln\frac{q}{\tilde{q}}\right]\right\}; \qquad \text{for } \tilde{p} < p. \tag{1.69}$$

So far, it seems that we have simply developed another upper bound on the tails of the distribution function for the binomial. It will then perhaps be surprising to compare this bound with the asymptotically correct value (repeated below) for the binomial PMF in (1.26).

$$\mathsf{p}_{S_n}(k) \sim \sqrt{\frac{1}{2\pi n\tilde{p}\tilde{q}}}\, \exp n\left[\tilde{p}\ln(p/\tilde{p}) + \tilde{q}\ln(q/\tilde{q})\right] \qquad \text{for } \tilde{p} = \frac{k}{n}. \tag{1.70}$$

For any integer value of $n\tilde{p}$ with $\tilde{p} > p$, we can lower bound $\Pr\{S_n \geq n\tilde{p}\}$ by the single term $\mathsf{p}_{S_n}(n\tilde{p})$. Thus $\Pr\{S_n \geq n\tilde{p}\}$ is both upper and lower bounded by quantities that decrease exponentially with n at the same rate. The lower bound is asymptotic in n and has the coefficient $1/\sqrt{2\pi n\tilde{p}\tilde{q}}$. These differences are essentially negligible for large n compared to the exponential term. We can express this analytically by considering the log of the upper bound in (1.68) and the lower bound in (1.70).

$$\lim_{n\to\infty} \frac{\ln \Pr\{S_n \geq n\tilde{p}\}}{n} = \left[\tilde{p}\ln\frac{p}{\tilde{p}} + \tilde{q}\ln\frac{q}{\tilde{q}}\right] \qquad \text{where } \tilde{p} > p. \tag{1.71}$$

In the same way, for $\tilde{p} < p$,

$$\lim_{n\to\infty} \frac{\ln \Pr\{S_n \le n\tilde{p}\}}{n} = \left[\tilde{p} \ln \frac{p}{\tilde{p}} + \tilde{q} \ln \frac{q}{\tilde{q}}\right] \} \quad \text{where } \tilde{p} < p. \tag{1.72}$$

In other words, these Chernoff bounds are not only upper bounds, but are also exponentially correct in the sense of (1.71) and (1.72). In Chapter 7 we will show that this property is typical for sums of IID rv's . Thus we see that the Chernoff bounds are not 'just bounds,' but rather are bounds that when optimized provide the correct asymptotic exponent for the tails of the distribution of sums of IID rv's. In this sense these bounds are very different from the Markov and Chebyshev bounds.

1.5 The laws of large numbers

The laws of large numbers are a collection of results in probability theory that describe the behavior of the arithmetic average of n rv's for large n. For any n rv's, $X_1, \ldots, X_n$, the *arithmetic average* is the rv $(1/n)\sum_{i=1}^{n} X_i$. Since in any outcome of the experiment, the sample value of this rv is the arithmetic average of the sample values of $X_1, \ldots, X_n$, this random variable is usually called the *sample average*. If $X_1, \ldots, X_n$ are viewed as successive variables in time, this sample average is called the time-average. Under fairly general assumptions, the standard deviation of the sample 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 section, we develop and discuss two of these results, the weak and the strong law of large numbers for independent identically distributed rv's. The strong law requires considerable patience to understand, but it is a basic and essential result in understanding stochastic processes. We also discuss the central limit theorem, partly because it enhances our understanding of the weak law, and partly because of its importance in its own right.

1.5.1 Weak law of large numbers with a finite variance

Let $X_1, X_2, \ldots, X_n$ be IID rv's with a finite mean $\overline{X}$ and finite variance σ_X^2. Let $S_n = X_1 + \cdots + X_n$, and consider the sample average S_n/n. We saw in (1.42) that $\sigma_{S_n}^2 = n\sigma_X^2$. Thus the variance of S_n/n is

$$\mathsf{VAR}\left[\frac{S_n}{n}\right] = \mathsf{E}\left[\left(\frac{S_n - n\overline{X}}{n}\right)^2\right] = \frac{1}{n^2}\,\mathsf{E}\left[\left(S_n - n\overline{X}\right)^2\right] = \frac{\sigma_X^2}{n}. \tag{1.73}$$

This says that the standard deviation of the sample average S_n/n is $\sigma/\sqrt{n}$, which approaches 0 as n increases. Figure 1.10 illustrates this decrease in the standard deviation of S_n/n with

increasing n. In contrast, recall that Figure 1.5 illustrated how the standard deviation of S_n increases with n. From (1.73), we see that

$$\lim_{n\to\infty} \mathsf{E}\left[\left(\frac{S_n}{n} - \overline{X}\right)^2\right] = 0. \tag{1.74}$$

As a result, we say that S_n/n *converges in mean square* to $\overline{X}$.

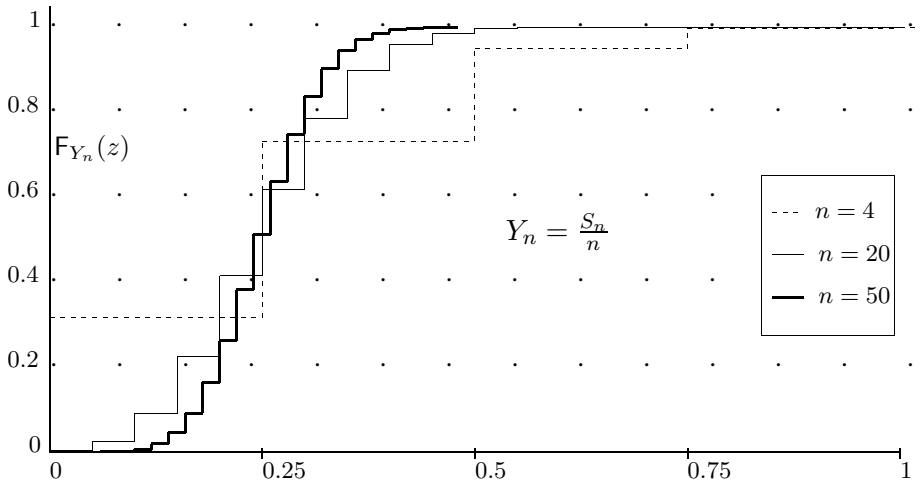

Figure 1.10: The same distribution as Figure 1.5, scaled differently to give the distribution function of the sample average Y_n. It can be visualized that as n increases, the distribution function of Y_n becomes increasingly close to a unit step at the mean, 0.25, of the variables X being summed.

This convergence in mean square says that the sample average, S_n/n, differs from the mean, $\overline{X}$, by a random variable whose standard deviation approaches 0 with increasing n. This convergence in mean square is one sense in which S_n/n approaches $\overline{X}$, but the idea of a sequence of rv's (*i.e.*, a sequence of functions) approaching a constant is clearly much more involved than a sequence of numbers approaching a constant. The laws of large numbers bring out this central idea in a more fundamental, and usually more useful, way. We start the development by applying the Chebyshev inequality (1.56) to the sample average,

$$\Pr\left\{\left|\frac{S_n}{n} - \overline{X}\right| > \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}. \tag{1.75}$$

This is an upper bound on the probability that S_n/n differs by more than ϵ from its mean, $\overline{X}$. This is illustrated in Figure 1.10 which shows the distribution function of S_n/n for various n. The figure suggests that $\lim_{n\to\infty} \mathsf{F}_{S_n/n}(y) = 0$ for all $y < \overline{X}$ and $\lim_{n\to\infty} \mathsf{F}_{S_n/n}(y) = 1$ for all $y > \overline{X}$. This is stated more cleanly in the following weak law of large numbers, abbreviated WLLN

Theorem 1.5.1 (WLLN with finite variance). *For each integer $n \geq 1$, let $S_n = X_1 + \cdots + X_n$ be the sum of n IID rv's with a finite variance. Then the following holds:*

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{S_n}{n} - \overline{X}\right| > \epsilon\right\} = 0 \qquad \textit{for every } \epsilon > 0. \tag{1.76}$$

Proof: For every $\epsilon > 0$, $\Pr\{|S_n/n - \overline{X}| > \epsilon$ is bounded between 0 and $\sigma^2/n\epsilon^2$. Since the upper bound goes to 0 with increasing n, the theorem is proved. □

Discussion: The algebraic proof above is both simple and rigorous. However, the geometric argument in Figure 1.11 probably provides more intuition about how the limit takes place. It is important to understand both.

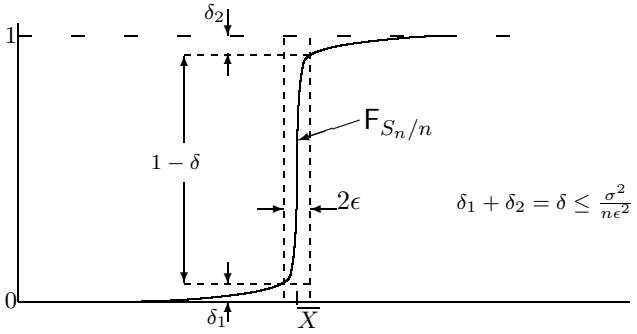

Figure 1.11: Approximation of the distribution function $\mathsf{F}_{S_n/n}$ of a sample average by a step function at the mean: From (1.75), the probability δ that S_n/n differs from $\overline{X}$ by more than ϵ (*i.e.*, $\Pr\{|S_n/n - \overline{X}| \geq \epsilon$) is at most $\sigma^2/n\epsilon^2$. The complementary event, where $|S_n/n - \overline{X}| < \epsilon$, has probability $1 - \delta \geq 1 - \sigma^2/n\epsilon^2$. This means that we can construct a rectangle of width 2ϵ centered on $\overline{X}$ and of height $1 - \delta$ such that $\mathsf{F}_{S_n/n}$ enters the rectangle at the lower left (say at $(\overline{X} - \epsilon, \delta_1)$) and exits at the upper right, say at $(\overline{X} + \epsilon, 1 - \delta_2)$). Now visualize increasing n while holding ϵ fixed. In the limit, $1 - \delta \to 1$ so $\Pr\{|S_n/n - \overline{X}| \geq \epsilon \to 0$. Since this is true for every $\epsilon > 0$ (usually with slower convergence as ϵ gets smaller), $\mathsf{F}_{S_n/n}(y)$ approaches 0 for every $y < \overline{X}$ and approaches 1 for every $y > \overline{X}$, *i.e.*, $\mathsf{F}_{S_n/n}$ approaches a unit step at $\overline{X}$. Note that there are two 'fudge factors' here, ϵ and δ and, since we are approximating an entire distribution function, neither can be omitted, except by directly going to a limit as $n \to \infty$.

We refer to (1.76) as saying that S_n/n converges to $\overline{X}$ in probability. To make sense out of this, we should view $\overline{X}$ as a deterministic random variable, *i.e.*, a rv that takes the value $\overline{X}$ for each sample point of the space. Then (1.76) says that the probability that the absolute difference, $|S_n/n - \overline{X}|$, exceeds any given $\epsilon > 0$ goes to 0 as $n \to \infty$.[31]

One should ask at this point what (1.76) adds to the more specific bound in (1.75). In particular (1.75) provides an upper bound on the rate of convergence for the limit in (1.76). The answer is that (1.76) remains valid when the theorem is generalized. For variables that are not IID or have an infinite variance, (1.75) is no longer necessarily valid. In some situations, as we see later, it is valuable to know that (1.76) holds, even if the rate of convergence is extremely slow or unknown.

One difficulty with the bound in (1.75) is that it is extremely loose in most cases. If S_n/n actually approached $\overline{X}$ this slowly, the weak law of large numbers would often be more a mathematical curiosity than a highly useful result. If we assume that the MGF of X exists

[31] Saying this in words gives one added respect for mathematical notation, and perhaps in this case, it is preferable to simply understand the mathematical statement (1.76).

in an open interval around 0, then (1.75) can be strengthened considerably. Recall from (1.64) and (1.65) that for any $\epsilon > 0$,

$$\Pr\{S_n/n - \overline{X} \geq \epsilon \quad \leq \quad \exp(n\mu_X(\overline{X} + \epsilon)) \tag{1.77}$$

$$\Pr\{S_n/n - \overline{X} \leq -\epsilon \quad \leq \quad \exp(n\mu_X(\overline{X} - \epsilon)), \tag{1.78}$$

where from Lemma 1.4.1, $\mu_X(a) = \inf_r\{\gamma_X(r) - ra\} < 0$ for $a \neq \overline{X}$. Thus, for any $\epsilon > 0$,

$$\Pr\{|S_n/n - \overline{X}| \geq \epsilon \quad \leq \exp[n\mu_X(\overline{X} + \epsilon)] + \exp[n\mu_X(\overline{X} - \epsilon)]. \tag{1.79}$$

The bound here, for any given $\epsilon > 0$, decreases geometrically in n rather than harmonically. In terms of Figure 1.11, the height of the rectangle must approach 1 at least geometrically in n.

1.5.2 Relative frequency

We next show that (1.76) can be applied to the relative frequency of an event as well as to the sample average of a random variable. Suppose that A is some event in a single experiment, and that the experiment is independently repeated n times. Then, in the probability model for the n repetitions, let A_i be the event that A occurs at the ith trial, $1 \leq i \leq n$. The events $A_1, A_2, \ldots, A_n$ are then IID.

If we let $\mathbb{I}_{A_i}$ be the indicator rv for A on the ith trial, then the rv $S_n = \mathbb{I}_{A_1} + \mathbb{I}_{A_2} + \cdots + \mathbb{I}_{A_n}$ is the number of occurrences of A over the n trials. It follows that

$$\text{relative frequency of } A = \frac{S_n}{n} = \frac{\sum_{i=1}^{n} \mathbb{I}_{A_i}}{n}. \tag{1.80}$$

Thus the relative frequency of A is the sample average of the binary rv's $\mathbb{I}_{A_i}$, and everything we know about the sum of IID rv's applies equally to the relative frequency of an event. In fact, everything we know about the sums of IID *binary* rv's applies to relative frequency.

1.5.3 The central limit theorem

The weak law of large numbers says that with high probability, S_n/n is close to $\overline{X}$ for large n, but it establishes this via an upper bound on the tail probabilities rather than an estimate of what $\mathsf{F}_{S_n/n}$ looks like. If we look at the shape of $\mathsf{F}_{S_n/n}$ for various values of n in the example of Figure 1.10, we see that the function $\mathsf{F}_{S_n/n}$ becomes increasingly compressed around $\overline{X}$ as n increases (in fact, this is the essence of what the weak law is saying). If we normalize the random variable S_n/n to 0 mean and unit variance, we get a normalized rv, $Z_n = (S_n/n - \overline{X})\sqrt{n}\,/\sigma$. The distribution function of Z_n is illustrated in Figure 1.12 for the same underlying X as used for S_n/n in Figure 1.10. The curves in the two figures are the same except that each curve has been horizontally scaled by $\sqrt{n}$ in Figure 1.12.

Inspection of Figure 1.12 shows that the normalized distribution functions there seem to be approaching a limiting distribution. The critically important *central limit theorem* states that there is indeed such a limit, and it is the normalized Gaussian distribution function.

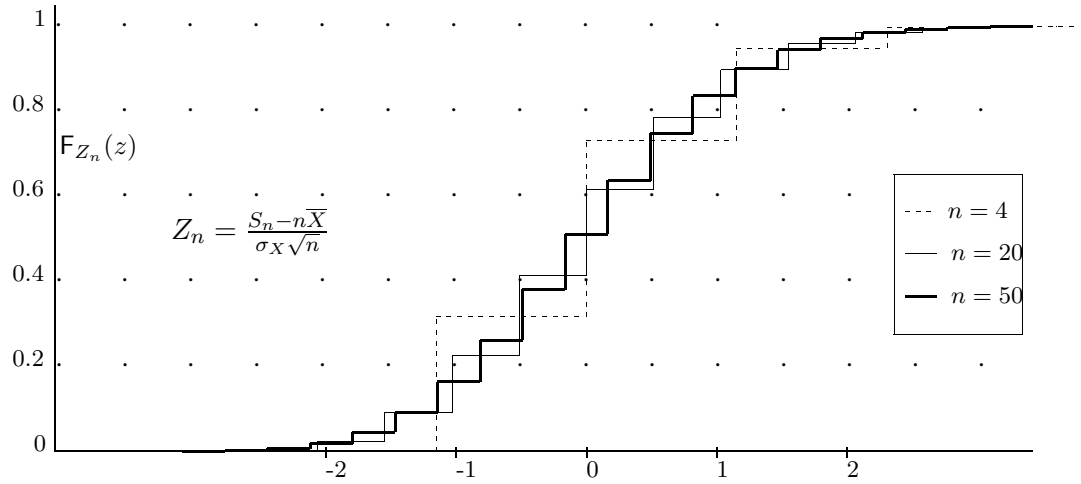

Figure 1.12: The same distribution functions as Figure 1.5 normalized to 0 mean and unit standard deviation, *i.e.*, the distribution functions of $Z_n = (S_n/n - \overline{X})\frac{\sqrt{n}}{\sigma_X}$ for $n = 4$, 20, 50. Note that as n increases, the distribution function of Z_n slowly starts to resemble the normal distribution function.

Theorem 1.5.2 (Central limit theorem (CLT)). *Let $X_1, X_2, \ldots$ be IID rv's with finite mean $\overline{X}$ and finite variance σ^2. Then for every real number z,*

$$\lim_{n\to\infty} \Pr\left\{ \frac{S_n - n\overline{X}}{\sigma\sqrt{n}} \leq z \right\} = \Phi(z), \tag{1.81}$$

where $\Phi(z)$ is the normal distribution function, i.e.*, the Gaussian distribution with mean 0 and variance 1,*

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\frac{y^2}{2})\, dy.$$

Discussion: The rv's $Z_n = (S_n - n\overline{X})/(\sigma\sqrt{n})$ for each $n \geq 1$ on the left side of (1.81) each have mean 0 and variance 1. The central limit theorem (CLT), as expressed in (1.81), says that the sequence of distribution functions, $\mathsf{F}_{Z_1}(z), \mathsf{F}_{Z_2}(z), \ldots$ converges at each value of z to $\Phi(z)$ as $n \to \infty$. In other words, $\lim_{n\to\infty} \mathsf{F}_{Z_n}(z) = \Phi(z)$ for each $z \in \mathbb{R}$. This is called convergence in distribution, since it is the sequence of distribution functions, rather than the sequence of rv's that is converging. The theorem is illustrated by Figure 1.12.

The reason why the word *central* appears in the CLT can be seen by rewriting (1.81) as

$$\lim_{n\to\infty} \Pr\left\{ \frac{S_n}{n} - \overline{X} \leq \frac{\sigma z}{\sqrt{n}} \right\} = \Phi(z), \tag{1.82}$$

Asymptotically, then, we are looking at a sequence of sample averages that differ from the mean by a quantity going to 0 with n as $1/\sqrt{n}$, *i.e.*, by sample averages very close to the mean for large n. This should be contrasted with the optimized Chernoff bound in (1.64)

and (1.65) which look at a sequence of sample averages that differ from the mean by a constant amount for large n. These latter results are exponentially decreasing in n and are known as large deviation results.

The CLT says nothing about speed of convergence to the normal distribution. The Berry-Esseen theorem (see, for example, Feller, [8]) provides some guidance about this for cases in which the third central moment $\mathsf{E}\left[|X-\overline{X}|^3\right]$ exists. This theorem states that

$$\left|\Pr\left\{\frac{S_n - n\overline{X})}{\sigma\sqrt{n}} \le z\right\} - \Phi(z)\right| \;\le\; \frac{C\mathsf{E}\left[|X-\overline{X}|^3\right]}{\sigma^3\sqrt{n}}. \tag{1.83}$$

where C can be upper bounded by 0.766. We will come back shortly to discuss convergence in greater detail.

The CLT helps explain why Gaussian rv's 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, $(S_n - n\overline{X})/\sigma\sqrt{n}$ need not have a density that is approximately Gaussian. In fact, if the underlying variables are discrete, the normalized sum is discrete and has no density. The PMF of the normalized sum can have very detailed fine structure; this does not disappear as n increases, but becomes "integrated out" in the distribution function.

The CLT tell us quite a bit about how $\mathsf{F}_{S_n/n}$ converges to a step function at $\overline{X}$. To see this, rewrite (1.81) in the form

$$\lim_{n\to\infty} \Pr\left\{\frac{S_n}{n} - \overline{X} \le \frac{\sigma z}{\sqrt{n}}\right\} = \Phi(z). \tag{1.84}$$

This is illustrated in Figure 1.13 where we have used $\Phi(z)$ as an approximation for the probability on the left.

A proof of the CLT requires mathematical tools that will not be needed subsequently. Thus we give a proof only for the binomial case. Feller ([7] and [8]) gives a thorough and careful exposition and proof of several versions of the CLT including those here.[32]

Proof*[33] of Theorem 1.5.2 (binomial case): We first establish a somewhat simpler result which uses finite limits on both sides of $(S_n - n\overline{X})/\sigma\sqrt{n}$, *i.e.*, we show that for any finite $y < z$,

$$\lim_{n\to\infty} \Pr\left\{y < \frac{S_n - n\overline{X}}{\sigma\sqrt{n}} \le z\right\} = \int_y^z \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-u^2}{2}\right) du. \tag{1.85}$$

[32] Many elementary texts provide 'simple proofs,' using transform techniques, but, among other problems, they usually indicate that the normalized sum has a density that approaches the Gaussian density, which is incorrect for all discrete rv's.

[33] This proof can be omitted without loss of continuity. However, it is an important part of achieving a thorough understanding of the CLT.

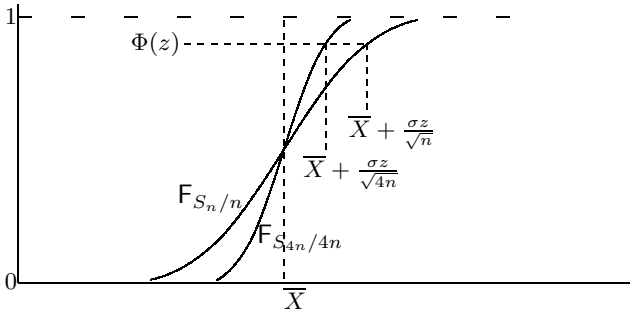

Figure 1.13: Approximation of the distribution function $\mathsf{F}_{S_n/n}$ of a sample average by a Gaussian distribution of the same mean and variance. Whenever n is increased by a factor of 4, the curve is horizontally scaled inward toward $\overline{X}$ by a factor of 2. The CLT says both that these curves are scaled horizontally as $1/\sqrt{n}$ and also that they are better approximated by the Gaussian of the given mean and variance as n increases.

The probability on the left of (1.85) can be rewritten as

$$\begin{aligned}\Pr\left\{y < \frac{S_n - n\overline{X}}{\sigma\sqrt{n}} \le z\right\}\} &= \Pr\left\{\frac{\sigma\, y}{\sqrt{n}} < \frac{S_n}{n} - \overline{X} \le \frac{\sigma\, z}{\sqrt{n}}\right\}\} \\ &= \sum_k \mathsf{p}_{S_n}(k) \quad \text{for} \quad \frac{\sigma\, y}{\sqrt{n}} < \frac{k}{n} - p \le \frac{\sigma\, z}{\sqrt{n}}. \end{aligned} \tag{1.86}$$

Let $\tilde{p} = k/n$, $\tilde{q} = 1 - \tilde{p}$, and $\epsilon(k,n) = \tilde{p} - p$. We abbreviate $\epsilon(k,n)$ as ϵ where k and n are clear from the context. From (1.23), we can express $\mathsf{p}_{S_n}(k)$ as

$$\begin{aligned}\mathsf{p}_{S_n}(k) &\sim \frac{1}{\sqrt{2\pi n\tilde{p}\tilde{q}}} \exp n\left[\tilde{p}\ln(p/\tilde{p}) + \tilde{q}\ln(q/\tilde{q})\right] \\ &= \frac{1}{\sqrt{2\pi n(p+\epsilon)(q-\epsilon)}} \exp\left[-n(p+\epsilon)\ln\left(1+\frac{\epsilon}{p}\right) - n(q-\epsilon)\ln\left(1-\frac{\epsilon}{q}\right)\right]\} \\ &= \frac{1}{\sqrt{2\pi n(p+\epsilon)(q-\epsilon)}} \exp\left[-n\left(\frac{\epsilon^2}{2p} - \frac{\epsilon^3}{6p^2} + \cdots + \frac{\epsilon^2}{2q} + \frac{\epsilon^3}{6q^2}\cdots\right)\right]. \end{aligned} \tag{1.87}$$

where we have used the power series expansion, $\ln(1+u) = u - u^2/2 + u^3/3\cdots$. From (1.86), $\sigma y/\sqrt{n} < p + \epsilon(k,n) \le \sigma z/\sqrt{n}$, so the omitted terms in (1.87) go to 0 uniformly over the range of k in (1.86). The term $(p+\epsilon)(q-\epsilon)$ also converges (uniformly in k) to pq as $n \to \infty$. Thus

$$\begin{aligned}\mathsf{p}_{S_n}(k) &\sim \frac{1}{\sqrt{2\pi npq}} \exp\left[-n\left(\frac{\epsilon^2}{2p} + \frac{\epsilon^2}{2q}\right)\right]\} \\ &= \frac{1}{\sqrt{2\pi npq}} \exp\left[\frac{-n\epsilon^2}{2pq}\right]\} \\ &= \frac{1}{\sqrt{2\pi npq}} \exp\left[\frac{-u^2(k,n)}{2}\right]\} \qquad \text{where } u(k,n) = \frac{k-np}{\sqrt{pqn}}. \end{aligned} \tag{1.88}$$

Since the ratio of the right to left side in (1.88) approaches 1 uniformly in k over the given range of k,

$$\sum_k \mathsf{p}_{S_n}(k) \sim \sum_k \frac{1}{\sqrt{2\pi npq}} \exp\left[\frac{-u^2(k,n)}{2}\right]. \tag{1.89}$$

Since $u(k,n)$ increases in increments of $1/\sqrt{pqn}$, we can view the sum on the right above as a Riemann sum approximating (1.85). Since the terms of the sum get closer as $n \to \infty$, and since the Riemann integral exists, (1.85) must be satisfied in the limit.

To complete the proof, note from the Chebyshev inequality that

$$\Pr\left\{|S_n - \overline{X}| \geq \frac{\sigma|y|}{\sqrt{n}}\right\} \leq \frac{1}{y^2},$$

and thus

$$\Pr\left\{y < \frac{S_n - n\overline{X}}{\sigma\sqrt{n}} \leq z\right\} \leq \Pr\left\{\frac{S_n - n\overline{X}}{\sigma\sqrt{n}} \leq z\right\} \leq \frac{1}{y^2} + \Pr\left\{y < \frac{S_n - n\overline{X}}{\sigma\sqrt{n}} \leq z\right\}$$

Choosing $y = -n^{1/4}$, we see that $1/y^2 \to 0$ as $n \to \infty$. Also $n\epsilon^3(k,n)$ in (1.87) goes to 0 as $n \to \infty$ for all (k,n) satisfying (1.86) with $y = -n^{1/4}$. Taking the limit as $n \to \infty$ then proves the theorem. □

If we trace through the various approximations in the above proof, we see that the error in $\mathsf{p}_{S_n}(k)$ goes to 0 as $1/n$. This is faster than the $1/\sqrt{n}$ bound in the Berry-Esseen theorem. If we look at Figure 1.3, however, we see that the distribution function of S_n/n contains steps of order $1/\sqrt{n}$. These vertical steps cause the binomial result here to have the same slow $1/\sqrt{n}$ convergence as the general Berry-Esseen bound. It turns out that if we evaluate the distribution function only at the midpoints between these steps, *i.e.*, at $z = (k + 1/2 - np)/\sigma\sqrt{n}$, then the convergence in the distribution function is of order $1/n$.

Since the CLT provides such explicit information about the convergence of S_n/n to $\overline{X}$, it is reasonable to ask why the weak law of large numbers (WLLN) is so important. The first reason is that the WLLN is so simple that it can be used to give clear insights to situations where the CLT could confuse the issue. A second reason is that the CLT requires a variance, where as we see next, the WLLN does not. A third reason is that the WLLN can be extended to many situations in which the variables are not independent and/or not identically distributed.[34] A final reason is that the WLLN provides an upper bound on the tails of $\mathsf{F}_{S_n/n}$, whereas the CLT provides only an approximation.

1.5.4 Weak law with an infinite variance

We now establish the WLLN without assuming a finite variance.

[34]Central limit theorems also hold in many of these more general situations, but they do not hold as widely as the WLLN.

Theorem 1.5.3 (WLLN). *For each integer $n \geq 1$, let $S_n = X_1 + \cdots + X_n$ where $X_1, X_2, \ldots$ are IID rv's satisfying $\mathsf{E}[|X|] < \infty$. Then for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{S_n}{n} - \mathsf{E}[X]\right| > \epsilon\right\} = 0. \tag{1.90}$$

Proof:[35] We use a truncation argument; such arguments are used frequently in dealing with rv's that have infinite variance. The underlying idea in these arguments is important, but some less important details are treated in Exercise 1.35. Let b be a positive number (which we later take to be increasing with n), and for each variable X_i, define a new rv $\breve{X}_i$ (see Figure 1.14) by

$$\breve{X}_i = \begin{cases} X_i & \text{for } \mathsf{E}[X] - b \leq X_i \leq \mathsf{E}[X] + b \\ \mathsf{E}[X] + b & \text{for } X_i > \mathsf{E}[X] + b \\ \mathsf{E}[X] - b & \text{for } X_i < \mathsf{E}[X] - b. \end{cases} \tag{1.91}$$

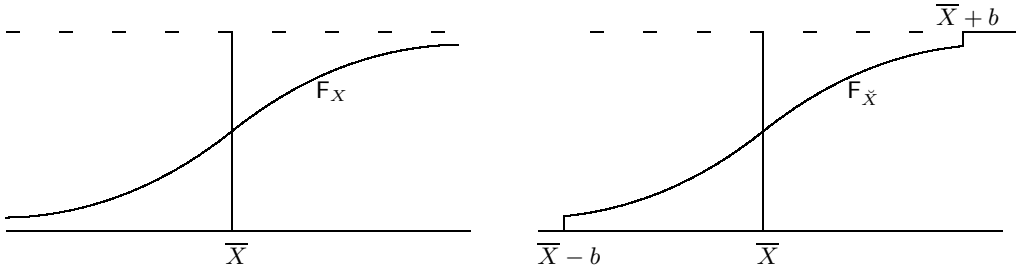

Figure 1.14: The truncated rv $\breve{X}$ for a given rv X has a distribution function which is truncated at $\overline{X} \pm b$.

The truncated variables $\breve{X}_i$ are IID and, because of the truncation, must have a finite second moment. Thus the WLLN applies to the sample average $\breve{S}_n = \breve{X}_1 + \cdots \breve{X}_n$. More particularly, using the Chebshev inequality in the form of (1.75) on $\breve{S}_n/n$, we get

$$\Pr\left\{\left|\frac{\breve{S}_n}{n} - \mathsf{E}\left[\breve{X}\right]\right| > \frac{\epsilon}{2}\right\} \leq \frac{4\sigma^2_{\breve{X}}}{n\epsilon^2} \leq \frac{8b\mathsf{E}[|X|]}{n\epsilon^2},$$

where Exercise 1.35 demonstrates the final inequality. Exercise 1.35 also shows that $\mathsf{E}\left[\breve{X}\right]$ approaches $\mathsf{E}[X]$ as $b \to \infty$ and thus that

$$\Pr\left\{\left|\frac{\breve{S}_n}{n} - \mathsf{E}[X]\right| > \epsilon\right\} \leq \frac{8b\mathsf{E}[|X|]}{n\epsilon^2}, \tag{1.92}$$

for all sufficiently large b. This bound also applies to S_n/n in the case where $S_n = \breve{S}_n$, so we have the following bound (see Exercise 1.35 for further details):

$$\Pr\left\{\left|\frac{S_n}{n} - \mathsf{E}[X]\right| > \epsilon\right\} \leq \Pr\left\{\left|\frac{\breve{S}_n}{n} - \mathsf{E}[X]\right| > \epsilon\right\} + \Pr\left\{S_n \neq \breve{S}_n\right\}. \tag{1.93}$$

[35] The details of this proof can be omitted without loss of continuity. However, the general structure of the truncation argument should be understood.

The original sum S_n is the same as $\breve{S}_n$ unless one of the X_i has an outage, *i.e.*, $|X_i - \overline{X}| > b$. Thus, using the union bound, $\Pr\left\{S_n \neq \breve{S}_n\right\} \leq n\Pr\{|X_i - \overline{X}| > b$. Substituting this and (1.92) into (1.93),

$$\Pr\left\{\left|\frac{S_n}{n} - \mathsf{E}[X]\right| > \epsilon\right\} \leq \frac{8b\mathsf{E}[|X|]}{n\epsilon^2} + \frac{n}{b}\left[b\Pr\{|X - \mathsf{E}[X]| > b\}\right]. \tag{1.94}$$

We now show that for any $\epsilon > 0$ and $\delta > 0$, $\Pr\{|S_n/n - \overline{X}| \geq \epsilon \leq \delta$ for all sufficiently large n. We do this, for given ϵ, δ, by choosing $b(n)$ for each n so that the first term in (1.94) is equal to $\delta/2$. Thus $b(n) = n\delta\epsilon^2/16\mathsf{E}[|X|]$. This means that $n/b(n)$ in the second term is independent of n. Now from (1.55), $\lim_{b\to\infty} b\Pr\{|X - \overline{X}| > b = 0$, so by choosing $b(n)$ sufficiently large (and thus n sufficiently large), the second term in (1.94) is also at most $\delta/2$. □

1.5.5 Convergence of random variables

This section has now developed a number of results about how the sequence of sample averages, $\{S_n/n;\ n \geq 1\}$ for a sequence of IID rv's $\{X_i;\ i \geq 1\}$ approach the mean $\overline{X}$. In the case of the CLT, the limiting distribution around the mean is also specified to be Gaussian. At the outermost intuitive level, *i.e.*, at the level most useful when first looking at some very complicated set of issues, viewing the limit of the sample averages as being essentially equal to the mean is highly appropriate.

At the next intuitive level down, the meaning of the word *essentially* becomes important and thus involves the details of the above laws. All of the results involve how the rv's S_n/n change with n and become better and better approximated by $\overline{X}$. When we talk about a sequence of rv's (namely a sequence of functions on the sample space) being approximated by a rv or numerical constant, we are talking about some kind of *convergence*, but it clearly is not as simple as a sequence of real numbers (such as $1/n$ for example) converging to some given number (0 for example).

The purpose of this section, is to give names and definitions to these various forms of convergence. This will give us increased understanding of the laws of large numbers already developed, but, equally important, It will allow us to develop another law of large numbers called the *strong law of large numbers* (SLLN). Finally, it will put us in a position to use these convergence results later for sequences of rv's other than the sample averages of IID rv's.

We discuss four types of convergence in what follows, convergence in distribution, in probability, in mean square, and with probability 1. For the first three, we first recall the type of large-number result with that type of convergence and then give the general definition.

For convergence with probability 1 (WP1), we first define this type of convergence and then provide some understanding of what it means. This will then be used in Chapter 4 to state and prove the SLLN.

We start with the central limit theorem, which, from (1.81) says

$$\lim_{n\to\infty} \Pr\left\{ \frac{S_n - n\overline{X}}{\sqrt{n}\,\sigma} \le z \right\} = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) dx \qquad \text{for every } z \in \mathbb{R}.$$

This is illustrated in Figure 1.12 and says that the sequence (in n) of distribution functions $\Pr\left\{ \frac{S_n - n\overline{X}}{\sqrt{n}\sigma} \le z \right\}$ converges at every z to the normal distribution function at z. This is an example of *convergence in distribution*.

Definition 1.5.1. *A sequence of random variables, $Z_1, Z_2, \ldots$, converges in distribution to a random variable Z if $\lim_{n\to\infty} \mathsf{F}_{Z_n}(z) = \mathsf{F}_Z(z)$ at each z for which $\mathsf{F}_Z(z)$ is continuous.*

For the CLT example, the rv's that converge in distribution are $\{\frac{S_n - n\overline{X}}{\sqrt{n}\sigma};\ n \ge 1\}$, and they converge in distribution to the normal Gaussian rv.

Convergence in distribution does not say that the rv's themselves converge in any reasonable sense, but only that their distribution functions converge. For example, let $Y_1, Y_2, \ldots$, be IID rv's, each with the distribution function F_Y. For each $n \ge 1$, if we let let $Z_n = Y_n + 1/n$, then it is easy to see that $\{Z_n;\ n \ge 1\}$ converges in distribution to Y. However (assuming Y has variance σ_Y^2 and is independent of each Z_n), we see that $Z_n - Y$ has variance $2\sigma_Y^2$. Thus Z_n does not get close to Y as $n \to \infty$ in any reasonable sense, and $Z_n - Z_m$ does not get small as n and m both get large.[36] As an even more trivial example, the sequence $\{Y_n\ n \ge 1\}$ converges in distribution to Y.

For the CLT, it is the rv's $\frac{S_n - n\overline{X}}{\sqrt{n}\sigma}$ that converge in distribution to the normal. As shown in Exercise 1.38, however, the rv $\frac{S_n - n\overline{X}}{\sqrt{n}\sigma} - \frac{S_{2n} - 2n\overline{X}}{\sqrt{2n}\sigma}$ is not close to 0 in any reasonable sense, even though the two terms have distribution functions that are very close for large n.

For the next type of convergence of rv's, the WLLN, in the form of (1.90), says that

$$\lim_{n\to\infty} \Pr\left\{ \left| \frac{S_n}{n} - \overline{X} \right| > \epsilon \right\} = 0 \qquad \text{for every } \epsilon > 0.$$

This is an example of *convergence in probability,* as defined below:

Definition 1.5.2. *A sequence of random variables $Z_1, Z_2, \ldots$, converges in probability to a rv Z if $\lim_{n\to\infty} \Pr\{|Z_n - Z| > \epsilon\} = 0$ for every $\epsilon > 0$.*

For the WLLN example, Z_n in the definition is the relative frequency S_n/n and Z is the constant rv $\overline{X}$. It is probably simpler and more intuitive in thinking about convergence of rv's to think of the sequence of rv's $\{Y_n = Z_n - Z;\ n \ge 1\}$ as converging to 0 in some sense.[37] As illustrated in Figure 1.10, convergence in probability means that $\{Y_n;\ n \ge 1\}$ converges in distribution to a unit step function at 0.

[36] In fact, saying that a sequence of rv's converges in distribution is unfortunate but standard terminology. It would be just as concise, and far less confusing, to say that a sequence of distribution functions converge rather than saying that a sequence of rv's converge in distribution.

[37] Definition 1.5.2 gives the impression that convergence to a rv Z is more general than convergence to a constant or convergence to 0, but converting the rv's to $Y_n = Z_n - Z$ makes it clear that this added generality is quite superficial.

An equivalent statement, as illustrated in Figure 1.11, is that $\{Y_n;\, n \geq 1\}$ converges in probability to 0 if $\lim_{n\to\infty} \mathsf{F}_{Y_n}(y) = 0$ for all $y < 0$ and $\lim_{n\to\infty} \mathsf{F}_{Y_n}(y) = 1$ for all $y > 0$. This shows that convergence in probability is a special case of convergence in distribution, since with convergence in probability, the sequence F_{Y_n} of distribution functions converges to a unit step at 0. Note that $\lim_{n\to\infty} \mathsf{F}_{Y_n}(y)$ is not specified at $y = 0$. However, the step function is not continuous at 0, so the limit there need not be specified for convergence in distribution.

Convergence in probability says quite a bit more than convergence in distribution. As an important example of this, consider the difference $Y_n - Y_m$ for n and m both large. If $\{Y_n;\, n \geq 1\}$ converges in probability to 0, then Y_n and Y_m are both close to 0 with high probability for large n and m, and thus close to each other. More precisely, $\lim_{m\to\infty,n\to\infty} \Pr\{|Y_n - Y_m| > \epsilon\} = 0$ for every $\epsilon > 0$. If the sequence $\{Y_n;\, n \geq 1\}$ merely converges in distribution to some arbitrary distribution, then, as we saw, $Z_n - Z_m$ can be large with high probability, even when n and m are large. Another example of this is given in Exercise 1.38.

It appears paradoxical that the CLT is more explicit about the convergence of S_n/n to $\overline{X}$ than the weak law, but it corresponds to a weaker type of convergence. The resolution of this paradox is that the sequence of rv's in the CLT is $\{\frac{S_n - n\overline{X}}{\sqrt{n}\sigma};\, n \geq 1\}$. The presence of $\sqrt{n}$ in the denominator of this sequence provides much more detailed information about how S_n/n approaches $\overline{X}$ with increasing n than the limiting unit step of $\mathsf{F}_{S_n/n}$ itself. For example, it is easy to see from the CLT that $\lim_{n\to\infty} \mathsf{F}_{S_n/n}(\overline{X}) = 1/2$, which can not be derived directly from the weak law.

Yet another kind of convergence is *convergence in mean square* (MS). An example of this, for the sample average S_n/n of IID rv's with a variance, is given in (1.74), repeated below:

$$\lim_{n\to\infty} \mathsf{E}\left[\left(\frac{S_n}{n} - \overline{X}\right)^2\right] = 0.$$

The general definition is as follows:

Definition 1.5.3. *A sequence of rv's $Z_1, Z_2, \ldots$, converges in mean square (MS) to a rv Z if $\lim_{n\to\infty} \mathsf{E}\left[(Z_n - Z)^2\right] = 0$.*

Our derivation of the weak law of large numbers (Theorem 1.5.1) was essentially based on the MS convergence of (1.74). Using the same approach, Exercise 1.37 shows in general that convergence in MS implies convergence in probability. Convergence in probability does not imply MS convergence, since as shown in Theorem 1.5.3, the weak law of large numbers holds without the need for a variance.

Figure 1.15 illustrates the relationship between these forms of convergence, *i.e.*, mean square convergence implies convergence in probability, which in turn implies convergence in distribution. The figure also shows convergence with probability 1 (WP1), which is the next form of convergence to be discussed.

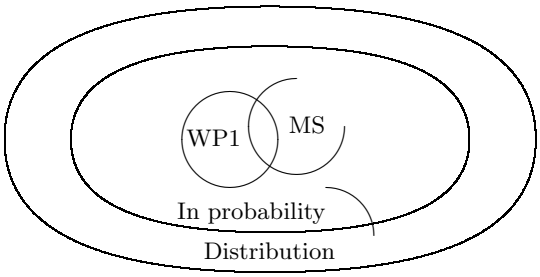

Figure 1.15: Relationship between different kinds of convergence: Convergence in distribution is the most general and is implied by all the others. Convergence in probability is the next most general and is implied by convergence with probability 1 (WP1) and by mean square (MS) convergence, neither of which imply the other.

1.5.6 Convergence with probability 1

We next define and discuss *convergence with probability 1*, abbreviated as convergence WP1. Convergence WP1 is often referred to as convergence a.s. (almost surely) and convergence a.e. (almost everywhere). The strong law of large numbers, which is discussed briefly in this section and further discussed and proven in various forms in Chapters 4 and 7, provides an extremely important example of convergence WP1. The general definition is as follows:

Definition 1.5.4. *Let $Z_1, Z_2, \ldots$, be a sequence of rv's in a sample space Ω and let Z be another rv in Ω. Then $\{Z_n; n \geq 1\}$ is defined to converge to Z with probability 1 (WP1) if*

$$\Pr\Big\{\omega \in \Omega \;:\; \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\Big\} = 1. \tag{1.95}$$

The condition $\Pr\{\omega \in \Omega : \lim_{n\to\infty} Z_n(\omega) = Z(\omega)\} = 1$ is often stated more compactly as $\Pr\{\lim_n Z_n = Z\} = 1$, and even more compactly as $\lim_n Z_n = Z$ WP1, but the form here is the simplest for initial understanding. As discussed in Chapter 4. the SLLN says that if $\{X_i; i \geq 1\}$ are IID with $\mathsf{E}[|X|] < \infty$, then the sequence of sample averages, $\{S_n/n; n \geq 1\}$ converges WP1 to $\overline{X}$.

In trying to understand (1.95), note that each sample point ω of the underlying sample space Ω maps to a sample value $Z_n(\omega)$ of each rv Z_n, and thus maps to a sample path $\{Z_n(\omega);\ n \geq 1\}$. For any given ω, such a sample path is simply a sequence of real numbers. That sequence of real numbers might converge to $Z(\omega)$ (which is a real number for the given ω), it might converge to something else, or it might not converge at all. Thus there is a set of ω for which the corresponding sample path $\{Z_n(\omega); n \geq 1\}$ converges to $Z(\omega)$, and a second set for which the sample path converges to something else or does not converge at all. Convergence WP1 of the sequence of rv's is thus defined to occur when the first set of sample paths above has probability 1.

For each ω, the sequence $\{Z_n(\omega);\ n \geq 1\}$ is simply a sequence of real numbers, so we briefly review what the limit of such a sequence is. A sequence of real numbers $b_1, b_2, \ldots$ is said to

have a limit b if, for every $\epsilon > 0$, there is an integer m_ϵ such that $|b_n - b| \leq \epsilon$ for all $n \geq m_\epsilon$. An equivalent statement is that $b_1, b_2, \ldots$, has a limit b if, for every integer $k \geq 1$, there is an integer $m(k)$ such that $|b_n - b| \leq 1/k$ for all $n \geq m(k)$.

Figure 1.16 illustrates this definition for those, like the author, whose eyes blur on the second or third 'there exists', 'such that', etc. in a statement. As illustrated, an important aspect of convergence of a sequence $\{b_n;\, n \geq 1\}$ of real numbers is that b_n becomes close to b for large n and stays close for all sufficiently large values of n.

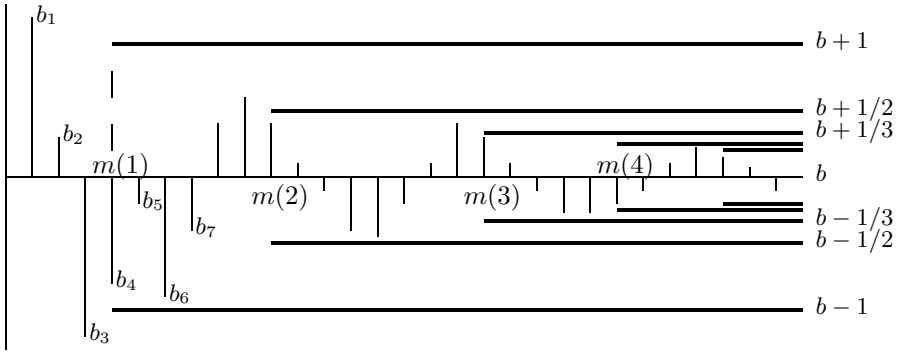

Figure 1.16: Illustration of a sequence of real numbers $b_1, b_2, \ldots$ that converge to a number b. The figure illustrates an integer $m(1)$ such that for all $n \geq m(1)$, b_n lies in the interval $b \pm 1$. Similarly, for each $k \geq 1$, there is an integer $m(k)$ such that b_n lies in $b \pm 1/k$ for all $n \geq m(k)$. Thus $\lim_{n\to\infty} b_n = b$ means that for a sequence of ever tighter constraints, the kth constraint can be met for all sufficiently large n, (*i.e.*, all $n \geq m(k)$). Intuitively, convergence means that the elements $b_1, b_2, \ldots$ get close to b *and stay close*. The sequence of positive integers $m(1), m(2), \ldots$ is nondecreasing, but otherwise arbitrary, depending only on the sequence $\{b_n;\, n \geq 1\}$. For sequences that converge very slowly, the integers $m(1), m(2), \ldots$ are simply correspondingly larger.

Figure 1.17 gives an example of a sequence of real numbers that does not converge. Intuitively, this sequence is close to 0 (and in fact identically equal to 0) for most large n, but it doesn't stay close.

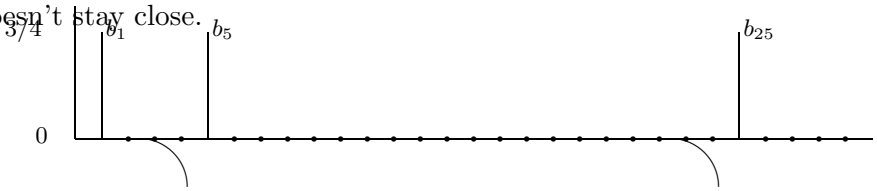

Figure 1.17: Illustration of a non-convergent sequence of real numbers $b_1, b_2, \ldots$. The sequence is defined by $b_n = 3/4$ for $n = 1,\ 5,\ 25, \ldots, 5^j, \ldots$ for all integer $j \geq 0$. For all other n, $b_n = 0$. The terms for which $b_n \neq 0$ become increasingly rare as $n \to \infty$. Note that $b_n \in [1, 1]$ for all n, but there is no $m(2)$ such that $b_n \in [-\frac{1}{2}, \frac{1}{2}]$ for all $n \geq m(2)$. Thus the sequence does not converge.

The following example illustrates how a sequence of rv's can converge in probability but not converge WP1. The example also provides some clues as to why convergence WP1 is important.

Example 1.5.1. Consider a sequence $\{Z_n; n \geq 1\}$ of rv's for which the sample paths constitute the following slight variation of the sequence of real numbers in Figure 1.17. In particular, as illustrated in Figure 1.18, the nonzero term at $n = 5^j$ in Figure 1.17 is replaced by a nonzero term at a randomly chosen n in the interval $[5^j, 5^{j+1})$.

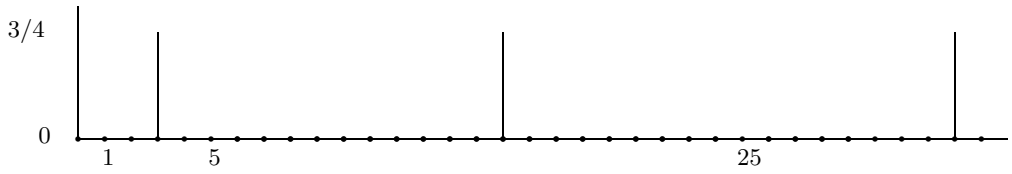

Figure 1.18: Illustration of a sample path of a sequence of rv's $\{Y_n; n \geq 0\}$ where, for each $j \geq 0$, $Y_n = 1$ for an equiprobable choice of $n \in [5^j, 5^{j+1})$ and $Y_n = 0$ otherwise.

Since each sample path contains a single one in each segment $[5^j, 5^{j+1})$, and contains zero's elsewhere, none of the sample paths converge. In other words, $\Pr\{\omega : \lim Z_n(\omega) = 0\} = 0$ rather than 1. On the other hand $\Pr\{Z_n = 0\} = 1 - 5^{-j}$ for $5^j \leq n < 5^{j+1}$, so $\lim_{n\to\infty} \Pr\{Z_n = 0\} = 1$.

Thus this sequence of rv's converges to 0 in probability, but does not converge to 0 WP1. This sequence also converges in mean square and (since it converges in probability) in distribution. Thus we have shown (by example) that convergence WP1 is not implied by any of the other types of convergence we have discussed. We will show in Section 4.2 that convergence WP1 does imply convergence in probability and in distribution but not in mean square (as illustrated in Figure 1.15).

The interesting point in this example is that this sequence of rv's is not bizarre (although it is somewhat specialized to make the analysis simple). Another important point is that this definition of convergence has a long history of being accepted as the 'useful,' 'natural,' and 'correct' way to define convergence for a sequence of real numbers. Thus it is not surprising that convergence WP1 will turn out to be similarly useful for sequences of rv's.

There is a price to be paid in using the concept of convergence WP1. We must then look at the entire sequence of rv's and can no longer analyze finite n-tuples and then go to the limit as $n \to \infty$. This requires a significant additional layer of abstraction, which involves additional mathematical precision and initial loss of intuition. For this reason we put off further discussion of convergence WP1 and the SLLN until Chapter 4 where it is needed.

1.6 Relation of probability models to the real world

Whenever experienced and competent engineers or scientists construct a probability model to represent aspects of some system that either exists or is being designed for some application, they must acquire a deep knowledge of the system and its surrounding circumstances, and concurrently consider various types of probability models used in probabilistic analysis of the same or similar systems. Usually very simple probability models help in understanding

the real-world system, and knowledge about the real-world system helps in understanding what aspects of the system are well-modeled by a given probability model. For a text such as this, there is insufficient space to understand the real-world aspects of each system that might be of interest. We must use the language of various canonical real-world systems for motivation and insight when studying probability models for various classes of systems, but such models must necessarily be chosen more for their tutorial than practical value.

There is a danger, then, that readers will come away with the impression that analysis is more challenging and important than modeling. To the contrary, for work on real-world systems, modeling is almost always more difficult, more challenging, and more important than analysis. The objective here is to provide the necessary knowledge and insight about probabilistic models so that the reader can later combine this with a deep understanding of particular real application areas. This will result in a useful interactive use of models, analysis, and experimentation.

In this section, our purpose is not to learn how to model real-world problems, since, as said above, this requires deep and specialized knowledge of whatever application area is of interest. Rather it is to understand the following conceptual problem that was posed in Section 1.1. Suppose we have a probability model of some real-world experiment involving randomness in the sense expressed there. When the real-world experiment being modeled is performed, there is an outcome, which presumably is one of the outcomes of the probability model, but there is no observable probability.

It appears to be intuitively natural, for experiments that can be carried out repeatedly under essentially the same conditions, to associate the probability of a given event with the relative frequency of that event over many repetitions. We now have the background to understand this approach. We first look at relative frequencies within the probability model, and then within the real world.

1.6.1 Relative frequencies in a probability model

We have seen that for any probability model, an extended probability model exists for n IID idealized experiments of the original model. For any event A in the original model, the indicator function $\mathbb{I}_A$ is a random variable, and the relative frequency of A over n IID experiments is the sample average of n IID rv's each with the distribution of $\mathbb{I}_A$. From the weak law of large numbers, this relative frequency converges in probability to $\mathsf{E}[\mathbb{I}_A] = \Pr\{A\}$. By taking the limit $n \to \infty$, the strong law of large numbers says that the relative frequency of A converges with probability 1 to $\Pr\{A\}$.

In plain English, this says that for large n, the relative frequency of an event (in the n-repetition IID model) is essentially the same as the probability of that event. The word *essentially* is carrying a great deal of hidden baggage. For the weak law, for any $\epsilon, \delta > 0$, the relative frequency is within some ϵ of $\Pr\{A\}$ with a confidence level $1 - \delta$ whenever n is sufficiently large. For the strong law, the ϵ and δ are avoided, but only by looking directly at the limit $n \to \infty$. Despite the hidden baggage, though, relative frequency and probability are related as indicated.

1.6.2 Relative frequencies in the real world

In trying to sort out if and when the laws of large numbers have much to do with real-world experiments, we should ignore the mathematical details for the moment and agree that for large n, the relative frequency of an event A over n IID trials of an idealized experiment is essentially $\Pr\{A\}$. We can certainly visualize a real-world experiment that has the same set of possible outcomes as the idealized experiment and we can visualize evaluating the relative frequency of A over n repetitions with large n. If that real-world relative frequency is essentially equal to $\Pr\{A\}$, and this is true for the various events A of greatest interest, then it is reasonable to hypothesize that the idealized experiment is a reasonable model for the real-world experiment, at least so far as those given events of interest are concerned.

One problem with this comparison of relative frequencies is that we have carefully specified a model for n IID repetitions of the idealized experiment, but have said nothing about how the real-world experiments are repeated. The IID idealized experiments specify that the conditional probability of A at one trial is the same no matter what the results of the other trials are. Intuitively, we would then try to isolate the n real-world trials so they don't affect each other, but this is a little vague. The following examples help explain this problem and several others in comparing idealized and real-world relative frequenices.

Example 1.6.1. Coin tossing: Tossing coins is widely used as a way to choose the first player in various games, and is also sometimes used as a primitive form of gambling. Its importance, however, and the reason for its frequent use, is its simplicity. When tossing a coin, we would argue from the symmetry between the two sides of the coin that each should be equally probable (since any procedure for evaluating the probability of one side should apply equally to the other). Thus since H and T are the only outcomes (the remote possibility of the coin balancing on its edge is omitted from the model), the reasonable and universally accepted model for coin tossing is that H and T each have probability $1/2$.

On the other hand, the two sides of a coin are embossed in different ways, so that the mass is not uniformly distributed. Also the two sides do not behave in quite the same way when bouncing off a surface. Each denomination of each currency behaves slightly differently in this respect. Thus, not only do coins violate symmetry in small ways, but different coins violate it in different ways.

How do we test whether this effect is significant? If we assume for the moment that successive tosses of the coin are well-modeled by the idealized experiment of n IID trials, we can essentially find the probability of H for a particular coin as the relative frequency of H in a sufficiently large number of independent tosses of that coin. This gives us slightly different relative frequencies for different coins, and thus slightly different probability models for different coins.

The assumption of independent tosses is also questionable. Consider building a carefully engineered machine for tossing coins and using it in a vibration-free environment. A standard coin is inserted into the machine in the same way for each toss and we count the number of heads and tails. Since the machine has essentially eliminated the randomness, we would expect all the coins, or almost all the coins, to come up the same way — the more precise the machine, the less independent the results. By inserting the original coin in a random

way, a single trial might have equiprobable results, but successive tosses are certainly not independent. The successive trials would be closer to independent if the tosses were done by a slightly inebriated individual who tossed the coins high in the air.

The point of this example is that there are many different coins and many ways of tossing them, and the idea that one model fits all is reasonable under some conditions and not under others. Rather than retreating into the comfortable world of theory, however, note that we can now find the relative frequency of heads for any given coin and essentially for any given way of tossing that coin.[38]

Example 1.6.2. Binary data: Consider the binary data transmitted over a communication link or stored in a data facility. The data is often a mixture of encoded voice, video, graphics, text, etc., with relatively long runs of each, interspersed with various protocols for retrieving the original non-binary data.

The simplest (and most common) model for this is to assume that each binary digit is 0 or 1 with equal probability and that successive digits are statistically independent. This is the same as the model for coin tossing after the trivial modification of converting $\{H, T\}$ into $\{0, 1\}$. This is also a rather appropriate model for designing a communication or storage facility, since all n-tuples are then equiprobable (in the model) for each n, and thus the facilities need not rely on any special characteristics of the data. On the other hand, if one wants to compress the data, reducing the required number of transmitted or stored bits per incoming bit, then a more elaborate model is needed.

Developing such an improved model would require finding out more about where the data is coming from — a naive application of calculating relative frequencies of n-tuples would probably not be the best choice. On the other hand, there are well-known data compression schemes that in essence track dependencies in the data and use them for compression in a coordinated way. These schemes are called *universal data-compression* schemes since they don't rely on a probability model. At the same time, they are best analyzed by looking at how they perform for various idealized probability models.

The point of this example is that choosing probability models often depends heavily on how the model is to be used. Models more complex than IID binary digits are usually based on what is known about the input processes. Measuring relative frequencies and Associating them with probabilities is the basic underlying conceptual connection between real-world and models, but in practice this is essentially the relationship of last resort. For most of the applications we will study, there is a long history of modeling to build on, with experiments as needed.

Example 1.6.3. Fable: In the year 2008, the financial structure of the USA failed and the world economy was brought to its knees. Much has been written about the role of greed on Wall Street and incompetence in Washington. Another aspect of the collapse, however, was a widespread faith in stochastic models for limiting risk. These models encouraged

[38] We are not suggesting that distinguishing different coins for the sake of coin tossing is an important problem. Rather, we are illustrating that even in such a simple situation, the assumption of identically prepared experiments is questionable and the assumption of independent experiments is questionable. The extension to n repetitions of IID experiments is not necessarily a good model for coin tossing. In other words, one has to question both the original model and the n-repetition model.

people to engage in investments that turned out to be far riskier than the models predicted. These models were created by some of the brightest PhD's from the best universities, but they failed miserably because they modeled everyday events very well, but modeled the rare events and the interconnection of events poorly. They failed badly by not understanding their application, and in particular, by trying to extrapolate typical behavior when their primary goal was to protect against highly atypical situations. The moral of the fable is that brilliant analysis is not helpful when the modeling is poor; as computer engineers say, "garbage in, garbage out."

The examples above show that the problems of modeling a real-world experiment are often connected with the question of creating a model for a set of experiments that are not exactly the same and do not necessarily correspond to the notion of independent repetitions within the model. In other words, the question is not only whether the probability model is reasonable for a single experiment, but also whether the IID repetition model is appropriate for multiple copies of the real-world experiment.

At least we have seen, however, that if a real-world experiment can be performed many times with a physical isolation between performances that is well modeled by the IID repetition model, then the relative frequencies of events in the real-world experiment correspond to relative frequencies in the idealized IID repetition model, which correspond to probabilities in the original model. In other words, under appropriate circumstances, the probabilities in a model become essentially observable over many repetitions.

We will see later that our emphasis on IID repetitions was done for simplicity. There are other models for repetitions of a basic model, such as Markov models, that we study later. These will also lead to relative frequencies approaching probabilities within the repetition model. Thus, for repeated real-world experiments that are well modeled by these repetition models, the real world relative frequencies approximate the probabilities in the model.

1.6.3 Statistical independence of real-world experiments

We have been discussing the use of relative frequencies of an event A in a repeated real-world experiment to test $\Pr\{A\}$ in a probability model of that experiment. This can be done essentially successfully if the repeated trials correpond to IID trials in the idealized experiment. However, the statement about IID trials in the idealized experiment is a statement about probabilities in the extended n-trial model. Thus, just as we tested $\Pr\{A\}$ by repeated real-world trials of a single experiment, we should be able to test $\Pr\{A_1, \ldots, A_n\}$ in the n-repetition model by a much larger number of real-world repetitions of n-tuples rather than single trials.

To be more specific, choose two large integers, m and n, and perform the underlying real-world experiment mn times. Partition the mn trials into m runs of n trials each. For any given n-tuple $A_1, \ldots, A_n$ of successive events, find the relative frequency (over m trials of n tuples) of the n-tuple event $A_1, \ldots, A_n$. This can then be used essentially to test the probability $\Pr\{A_1, \ldots, A_n\}$ in the model for n IID trials. The individual event probabilities can also be tested, so the condition for independence can be tested.

The observant reader will note that there is a tacit assumption above that successive n tuples can be modeled as independent, so it seems that we are simply replacing a big problem with a bigger problem. This is not quite true, since if the trials are dependent with some given probability model for dependent trials, then this test for independence will essentially reject the independence hypothesis for large enough n. In other words, we can not completely verify the correctness of an independence hypothesis for the n-trial model, although in principle we could eventually falsify it if it is false.

Choosing models for real-world experiments is primarily a subject for statistics, and we will not pursue it further except for brief discussions when treating particular application areas. The purpose here has been to treat a fundamental issue in probability theory. As stated before, probabilities are non-observables — they exist in the theory but are not directly measurable in real-world experiments. We have shown that probabilities essentially become observable in the real-world via relative frequencies over repeated trials.

1.6.4 Limitations of relative frequencies

Most real-world applications that are modeled by probability models have such a large sample space that it is impractical to conduct enough trials to choose probabilities from relative frequencies. Even a shuffled deck of 52 cards would require many more than $52! \approx 8 \times 10^{67}$ trials for most of the outcomes to appear even once. Thus relative frequencies can be used to test the probability of given individual events of importance, but are usually impractical for choosing the entire model and even more impractical for choosing a model for repeated trials.

Since relative frequencies give us a concrete interpretation of what probability means, however, we can now rely on other approaches, such as symmetry, for modeling. From symmetry, for example, it is clear that all 52! possible arrangements of a card deck should be equiprobable after shuffling. This leads, for example, to the ability to calculate probabilities of different poker hands, etc., which are such popular exercises in elementary probability classes.

Another valuable modeling procedure is that of constructing a probability model where the possible outcomes are independently chosen n-tuples of outcomes in a simpler model. More generally, most of the random processes to be studied in this text are defined as various ways of combining simpler idealized experiments.

What is really happening as we look at modeling increasingly sophisticated systems and studying increasingly sophisticated models is that we are developing mathematical results for simple idealized models and relating those results to real-world results (such as relating idealized statistically independent trials to real-world independent trials). The association of relative frequencies to probabilities forms the basis for this, but is usually exercised only in the simplest cases.

The way one selects probability models of real-world experiments in practice is to use scientific knowledge and experience, plus simple experiments, to choose a reasonable model. The results from the model (such as the law of large numbers) are then used both to hypothesize results about the real-world experiment and to provisionally reject the model

when further experiments show it to be highly questionable. Although the results about the model are mathematically precise, the corresponding results about the real-world are at best insightful hypotheses whose most important aspects must be validated in practice.

1.6.5 Subjective probability

There are many useful applications of probability theory to situations other than repeated trials of a given experiment. When designing a new system in which randomness (of the type used in probability models) is hypothesized, one would like to analyze the system before actually building it. In such cases, the real-world system does not exist, so indirect means must be used to construct a probability model. Often some sources of randomness, such as noise, can be modeled in the absence of the system. Often similar systems or simulation can be used to help understand the system and help in formulating appropriate probability models. However, the choice of probabilities is to a certain extent subjective.

Another type of situation, of which a canonic example is risk analysis for nuclear reactors, deals with a large number of very unlikely outcomes, each catastrophic in nature. Experimentation clearly cannot be used to establish probabilities, and it is not clear that probabilities have any real meaning here. It can be helpful, however, to choose a probability model on the basis of subjective beliefs which can be used as a basis for reasoning about the problem. When handled well, this can at least make the subjective biases clear, leading to a more rational approach to the problem. When handled poorly, it can hide the arbitrary nature of possibly poor decisions.

We will not discuss the various, often ingenious methods to choose subjective probabilities. The reason is that subjective beliefs should be based on intensive and long term exposure to the particular problem involved; discussing these problems in abstract probability terms weakens this link. We will focus instead on the analysis of idealized models. These can be used to provide insights for subjective models, and more refined and precise results for objective models.

1.7 Summary

This chapter started with an introduction into the correspondence between probability theory and real-world experiments involving randomness. While almost all work in probability theory works with established probability models, it is important to think through what these probabilities mean in the real world, and elementary subjects rarely address these questions seriously.

The next section discussed the axioms of probability theory, along with some insights about why these particular axioms were chosen. This was followed by a review of conditional probabilities, statistical independence, random variables, stochastic processes, and expectations. The emphasis was on understanding the underlying structure of the field rather than reviewing details and problem solving techniques.

This was followed by a fairly extensive treatment of the laws of large numbers. This involved

a fair amount of abstraction, combined with mathematical analysis. The central idea is that the sample average of n IID rv's approaches the mean with increasing n. As a special case, the relative frequency of an event A approaches $\Pr\{A\}$. What the word *approaches* means here is both tricky and vital in understanding probability theory. The strong law of large numbers and convergence WP1 requires mathematical maturity, and is postponed to Chapter 4 where it is first used.

The final section came back to the fundamental problem of understanding the relation between probability theory and randomness in the real-world. It was shown, via the laws of large numbers, that probabilities become essentially observable via relative frequencies calculated over repeated experiments.

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 in this chapter. We recommend Bertsekas and Tsitsiklis [2], both for a careful statement of the fundamentals and for a wealth of well-chosen and carefully explained examples.

Texts that cover similar material to that here are [16] and [11]. Kolmogorov [14] is readable for the mathematically mature and is also of historical interest as the translation of the 1933 book that first put probability on a firm mathematical basis. Feller [7] is the classic extended and elegant treatment of elementary material from a mature point of view. Rudin [17] is an excellent text on measure theory for those with advanced mathematical preparation.

1.8 Exercises

Exercise 1.1. Consider a sequence $A_1, A_2, \ldots$ of events each of which have probability zero.

a) Find $\Pr\{\sum_{n=1}^{m} A_n\}$ and find $\lim_{m\to\infty} \Pr\{\sum_{n=1}^{m} A_n\}$. What you have done is to show that the sum of a countably infinite set of numbers each equal to 0 is perfectly well defined as 0.

b) For a sequence of possible phases, $a_1, a_2, \ldots$ between 0 and 2π, and a sequence of singleton events, $A_n = \{a_n\}$, find $\Pr\{\bigcup_n A_n\}$ assuming that the phase is uniformly distributed.

c) Now let A_n be the empty event ϕ for all n. Use (1.1) to show that $\Pr\{\phi\} = 0$.

Exercise 1.2. Let A_1 and A_2 be arbitrary events and show that $\Pr\{A_1 \bigcup A_2\} + \Pr\{A_1 A_2\} = \Pr\{A_1\} + \Pr\{A_2\}$. Explain which parts of the sample space are being double counted on both sides of this equation and which parts are being counted once.

Exercise 1.3. Let $A_1, A_2, \ldots$, be a sequence of disjoint events and assume that $\Pr\{A_n\} = 2^{-n-1}$ for each $n \geq 1$. Assume also that $\Omega = \bigcup_{n=1}^{\infty} A_n$.

a) Show that these assumptions violate the axioms of probability.

b) Show that if (1.3) is substituted for the third of those axioms, then the above assumptions satisfy the axioms.

This shows that the countable additivity of Axiom 3 says something more than the finite additivity of (1.3).

Exercise 1.4. This exercise derives the probability of an arbitrary (non-disjoint) union of events, derives the union bound, and derives some useful limit expressions.

a) For 2 arbitrary events A_1 and A_2, show that

$$A_1 \bigcup A_2 = A_1 \bigcup (A_2 - A_1) \qquad \text{where } A_2 - A_1 = A_2 A_1^c.$$

Show that A_1 and $A_2 - A_1$ are disjoint Hint: This is what Venn diagrams were invented for.

b) For an arbitrary sequence of events, $\{A_n;\, n \geq 1\}$, let $B_1 = A_1$ and for each $n \geq 2$ define $B_n = A_n - \bigcup_{j=1}^{n-1} A_j$. Show that $B_1, B_2, \ldots$, are disjoint events and show that for each $m \geq 2$, $\bigcup_{n=1}^{m} A_n = \bigcup_{n=1}^{n} B_n$. Hint: Use induction.

c) Show that

$$\Pr\Big\{\bigcup_{n=1}^{\infty} A_n\Big\} = \Pr\Big\{\bigcup_{n=1}^{\infty} B_n\Big\} = \sum_{n=1}^{\infty} \Pr\{B_n\}\,.$$

Hint: Use the axioms of probability for the second equality.

d) Show that for each n, $\Pr\{B_n\} \leq \Pr\{A_n\}$. Use this to show that

$$\Pr\Big\{\bigcup_{n=1}^{\infty} A_n\Big\} \leq \sum_{n=1}^{\infty} \Pr\{A_n\}\,.$$

e) Show that $\Pr\{\bigcup_{n=1}^{\infty} A_n\} = \lim_{m\to\infty} \Pr\{\bigcup_{n=1}^{m} A_n\}$. Hint: Combine parts c) and b).. Note that this says that the probability of a limit of unions is equal to the limit of the probabilities. This might well appear to be obvious without a proof, but you will see situations later where similar appearing interchanges cannot be made.

f) Show that $\Pr\{\bigcap_{n=1}^{\infty} A_n\} = \lim_{n\to\infty} \Pr\{\bigcap_{i=1}^{n} A_i\}$. Hint: Remember deMorgan's equalities.

Exercise 1.5. Consider a sample space of 8 equiprobable sample points and let A_1, A_2, A_3 be three events each of probability 1/2 such that $\Pr\{A_1A_2A_3\} = \Pr\{A_1\}\Pr\{A_2\}\Pr\{A_3\}$.

a) Create an example where $\Pr\{A_1A_2\} = \Pr\{A_1A_3\} = \frac{1}{4}$ but $\Pr\{A_2A_3\} = \frac{1}{8}$. Hint: Make a table with a row for each sample point and a column for each event and try different ways of assigning sample points to events (the answer is not unique).

b) Show that, for your example, A_2 and A_3 are not independent. Note that the definition of statistical independence would be very strange if it allowed A_1, A_2, A_3 to be independent while A_2 and A_3 are dependent. This illustrates why the definition of independence requires (1.13) rather than just (1.14).

Exercise 1.6. Suppose X and Y are discrete rv's with the PMF $\mathsf{p}_{XY}(x_i, y_j)$. Show (a picture will help) that this is related to the joint distribution function by

$$\mathsf{p}_{XY}(x_i, y_j) = \lim_{\delta>0,\delta\to 0} [\mathsf{F}(x_i, y_j) - \mathsf{F}(x_i - \delta, y_j) - \mathsf{F}(x_i, y_j - \delta) + \mathsf{F}(x_i - \delta, y_j - \delta)]\,.$$

Exercise 1.7. A variation of Example 1.3.2 is to let M be a random variable that takes on both positive and negative values with the PMF

$$\mathsf{p}_M(m) = \frac{1}{2|m|\,(|m|+1)}.$$

In other words, M is symmetric around 0 and $|M|$ has the same PMF as the nonnegative rv N of Example 1.3.2.

a) Show that $\sum_{m\geq 0} m\mathsf{p}_M(m) = \infty$ and $\sum_{m<0} m\mathsf{p}_M(m) = -\infty$. (Thus show that the expectation of M not only does not exist but is undefined even in the extended real number system.)

b) Suppose that the terms in $\sum_{m=-\infty}^{\infty} m\mathsf{p}_M(m)$ are summed in the order of 2 positive terms for each negative term (*i.e.*, in the order $1, 2, -1, 3, 4, -2, 5, \cdots$). Find the limiting value of the partial sums in this series. Hint: You may find it helpful to know that

$$\lim_{n\to\infty} \left[\sum_{i=1}^{n} \frac{1}{i} - \int_1^n \frac{1}{x}\,dx\right] = \gamma,$$

where γ is the Euler-Mascheroni constant, $\gamma = 0.57721\cdots$.

c) Repeat part b) where, for any given integer $k > 0$, the order of summation is k positive terms for each negative term.

Exercise 1.8. a) For any given rv Y, express $\mathsf{E}[|Y|]$ in terms of $\int_{y<0} \mathsf{F}_Y(y)\,dy$ and $\int_{y\geq 0} \mathsf{F}^{\mathsf{c}}_Y(y)\,dy$.

b) Show that $\mathsf{E}[|Y-\alpha|]$ is minimized by setting α equal to the median of Y (*i.e.*, the value of Y for which $\mathsf{F}_Y(y) = 1/2$). Hint: Use a graphical argument.

Exercise 1.9. Let X be a rv with distribution function $\mathsf{F}_X(x)$. Find the distribution function of the following rv's.

a) The maximum of n IID rv's, each with distribution function $\mathsf{F}_X(x)$.

b) The minimum of n IID rv's, each with distribution $\mathsf{F}_X(x)$.

c) The difference of the rv's defined in a) and b); assume X has a density $\mathsf{f}_X(x)$.

Exercise 1.10. Let X and Y be rv's in some sample space Ω and let $Z = X + Y$, *i.e.*, for each $\omega \in \Omega$, $Z(\omega) = X(\omega) + Y(\omega)$.

a) Show that the set of ω for which $Z(\omega) = \pm\infty$ has probability 0.

b) To show that $Z = X+Y$ is a rv, we must show that for each real number α, the set $\{\omega \in \Omega : X(\omega)+Y(\omega) \leq \alpha\}$ is an event. We proceed indirectly. For an arbitrary positive integer n and an arbitrary integer k, let $B(n,k) = \{\omega : X(\omega) \leq k\alpha/n\} \bigcap \{Y(\omega) \leq (n+1-k)\alpha/n\}$. Let $D(n) = \bigcup_k B(n,k)$ and show that $D(n)$ is an event.

c) On a 2 dimensional sketch for a given α, show the values of $X(\omega)$ and $Y(\omega)$ for which $\omega \in D(n)$. Hint: This set of values should be bounded by a staircase function.

d) Show that

$$\{\omega : X(\omega) + Y(\omega) \leq \alpha\} \;=\; \bigcap\nolimits_n B(n)$$

Explain why this shows that $Z = X + Y$ is a rv.

Exercise 1.11. a) Let $X_1, X_2, \ldots, X_n$ be rv's with expected values $\overline{X}_1, \ldots, \overline{X}_n$. Show that $\mathsf{E}[X_1 + \cdots + X_n] = \overline{X}_1 + \cdots + \overline{X}_n$. You may assume that the rv's have a joint density function, but do not assume that the rv's 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 1.12. (Stieltjes integration) **a)** Let $h(x) = u(x)$ and $\mathsf{F}_X(x) = u(x)$ where $u(x)$ is the unit step, *i.e.*, $u(x) = 0$ for $-\infty < x < 0$ and $u(x) = 1$ for $x \geq 0$. Using the definition of the Stieltjes integral in Footnote 22, show that $\int_{-1}^{1} h(x) d\mathsf{F}_X(x)$ does not exist. Hint: Look at the term in the Riemann sum including $x = 0$ and look at the range of choices for $h(x)$ in that interval. Intuitively, it might help initially to view $d\mathsf{F}_X(x)$ as a unit impulse at $x = 0$.

b) Let $h(x) = u(x - a)$ and $\mathsf{F}_X(x) = u(x - b)$ where a and b are in $(-1, +1)$. Show that $\int_{-1}^{1} h(x) d\mathsf{F}_X(x)$ exists if and only if $a \neq b$. Show that the integral has the value 1 for $a < b$ and the value 0 for $a > b$. Argue that this result is still valid in the limit of integration over $(-\infty, \infty)$.

c) Let X and Y be independent discrete rv's, each with a finite set of possible values. Show that $\int_{-\infty}^{\infty} \mathsf{F}_X(z - y) d\mathsf{F}_Y(y)$, defined as a Stieltjes integral, is equal to the distribution of $Z = X + Y$ at each z other than the possible sample values of Z, and is undefined at each sample value of Z. Hint: Express F_X and F_Y as sums of unit steps. Note: This failure of Stieltjes integration is not a serious problem; $\mathsf{F}_Z(z)$ is a step function, and the integral is undefined at its points of discontinuity. We automatically define $\mathsf{F}_Z(z)$ at those step values so that F_Z is a distribution function (*i.e.*, is continuous from the right). This problem does not arise if either X or Y is continuous.

Exercise 1.13. Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of IID continuous rv's with the common probability density function $\mathsf{f}_X(x)$; note that $\Pr\{X{=}\alpha\} = 0$ for all α and that $\Pr\{X_i{=}X_j\} = 0$ for all $i \neq j$. For $n \geq 2$, define X_n as a *record-to-date* of the sequence if $X_n > X_i$ for all $i < n$.

a) Find the probability that X_2 is a record-to-date. Use symmetry to obtain a numerical answer without computation. A one or two line explanation should be adequate.

b) Find the probability that X_n is a record-to-date, as a function of $n \geq 1$. Again use symmetry.

c) Find a simple expression for the expected number of records-to-date that occur over the first m trials for any given integer m. Hint: Use indicator functions. Show that this expected number is infinite in the limit $m \to \infty$.

Exercise 1.14. (Continuation of Exercise 1.13)

a) Let N_1 be the index of the *first* record-to-date in the sequence. Find $\Pr\{N_1 > n\}$ for each $n \geq 2$. Hint: There is a far simpler way to do this than working from part b) in Exercise 1.13.

b) Show that N_1 is a rv.

c) Show that $\mathsf{E}[N_1] = \infty$.

d) Let N_2 be the index of the *second* record-to-date in the sequence. Show that N_2 is a rv. Hint: You need not find the distribution function of N_2 here.

e) Contrast your result in part c) to the result from part c) of Exercise 1.13 saying that the expected number of records-to-date is infinite over an an infinite number of trials. Note:

this should be a shock to your intuition — there is an infinite expected wait for the first of an infinite sequence of occurrences, each of which must eventually occur.

Exercise 1.15. (Another direction from Exercise 1.13) **a)** For any given $n \geq 2$, find the probability that X_n and X_{n+1} are both records-to-date. Hint: The idea in part b) of 1.13 is helpful here, but the result is not.

b) Is the event that X_n is a record-to-date statistically independent of the event that X_{n+1} is a record-to-date?

c) Find the expected number of adjacent pairs of records-to-date over the sequence $X_1, X_2, \ldots$. Hint: A helpful fact here is that $\frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1}$.

Exercise 1.16. a) Assume that X is a nonnegative discrete rv taking on values $a_1, a_2, \ldots$, and let $Y = h(X)$ for some nonnegative function h. Let $b_i = h(a_i)$, $i \geq 1$ be the i^{th} value taken on by Y. Show that $\mathsf{E}[Y] = \sum_i b_i \mathsf{p}_Y(b_i) = \sum_i h(a_i)\mathsf{p}_X(a_i)$. Find an example where $\mathsf{E}[X]$ exists but $\mathsf{E}[Y] = \infty$.

b) Let X be a nonnegative continuous rv with density $\mathsf{f}_X(x)$ and let $h(x)$ be differentiable, nonnegative, and strictly increasing in x. Let $A(\delta) = \sum_n h(n\delta)[\mathsf{F}(n\delta) - \mathsf{F}(n\delta - \delta)]$, *i.e.*, $A(\delta)$ is the δth order approximation to the Stieltjes integral $\int h(x) d\mathsf{F}(x)$. Show that if $A(1) < \infty$, then $A(2^{-k}) \leq A(2^{k-1}) < \infty$. Show from this that $\int h(x)\, d\mathsf{F}(x)$ converges to a finite value. Note: this is a very special case, but it can be extended to many cases of interest. It seems better to consider these convergence questions as required rather than consider them in general.

Exercise 1.17. a) Consider a positive, integer-valued rv whose distribution function is given at integer values by

$$\mathsf{F}_Y(y) = 1 - \frac{2}{(y+1)(y+2)} \qquad \text{for integer } \ y > 0.$$

Use (1.35) to show that $\mathsf{E}[Y] = 2$. Hint: Note the PMF given in (1.31).

b) Find the PMF of Y and use it to check the value of $\mathsf{E}[Y]$.

c) Let X be another positive, integer-valued rv. Assume its conditional PMF is given by

$$\mathsf{p}_{X|Y}(x|y) = \frac{1}{y} \qquad \text{for } 1 \leq x \leq y.$$

Find $\mathsf{E}[X \mid Y = y]$ and show that $\mathsf{E}[X] = 3/2$. Explore finding $\mathsf{p}_X(x)$ until you are convinced that using the conditional expectation to calculate $\mathsf{E}[X]$ is considerably easier than using $\mathsf{p}_X(x)$.

d) Let Z be another integer-valued rv with the conditional PMF

$$\mathsf{p}_{Z|Y}(z|y) = \frac{1}{y^2} \qquad \text{for } 1 \leq z \leq y^2.$$

Find $\mathsf{E}[Z \mid Y = y]$ for each integer $y \geq 1$ and find $\mathsf{E}[Z]$.

Exercise 1.18. a) Show that, for uncorrelated rv's, the expected value of the product is equal to the product of the expected values (by definition, X and Y are uncorrelated if $\mathsf{E}\left[(X-\overline{X})(Y-\overline{Y})\right] = 0$).

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 rv's are uncorrelated.

e) Let X, Y be identically distributed ternary valued random variables with the PMF $\mathsf{p}_X(-1) = \mathsf{p}_X(1) = 1/4$; $\mathsf{p}_X(0) = 1/2$. 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 rv's 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 1.19. Suppose X has the Poisson PMF, $\mathsf{p}_X(n) = \lambda^n \exp(-\lambda)/n!$ for $n \geq 0$ and Y has the Poisson PMF, $\mathsf{p}_Y(m) = \mu^n \exp(-\mu)/n!$ for $n \geq 0$. Assume that X and Y are independent. Find the distribution of $Z = X + Y$ and find the conditional distribution of Y conditional on $Z = n$.

Exercise 1.20. a) Suppose X, Y and Z are binary rv's, 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 $\mathsf{p}_{XYZ}(x,y,z)$ for your example. Hint: In the simplest examle, only 4 of the joint values for x, y, z have positive probabilities.

b) Is pairwise statistical independence enough to ensure that

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

for a set of rv's $X_1, \ldots, X_n$?

Exercise 1.21. Show that $\mathsf{E}[X]$ is the value of α that minimizes $\mathsf{E}\left[(X-\alpha)^2\right]$.

Exercise 1.22. For each of the following random variables, find the interval (r_-, r_+) over which the moment generating function $g(r)$ exists. Determine in each case whether $g_X(r)$ exists at the end points r_- and r_+. For parts a) and b) you should also find and sketch $g(r)$. For part c), $g(r)$ has no closed form.

a) Let λ, θ, be positive numbers and let X have the density

$$\mathsf{f}_X(x) = \frac{1}{2}\lambda \exp(-\lambda x);\ x \geq 0; \quad \mathsf{f}_X(x) = \frac{1}{2}\theta \exp(\theta x);\ x < 0.$$

b) Let Y be a Gaussian random variable with mean m and variance σ^2.

c) Let Z be a nonnegative random variable with density

$$\mathsf{f}_Z(z) = k(1+z)^{-2}\ \exp(-\lambda z); \quad z \geq 0.$$

where $\lambda > 0$ and $k = [\int_{z\geq 0}(1+z)^2\ \exp(-az)dz]^{-1}$. Hint: Do not try to evaluate $g_Z(r)$. Instead, investigate values of r for which the integral is finite and infinite.

Exercise 1.23. Recall that the MGF of the nonnegative exponential rv with density e^{-x} is $(1-r)^{-1}$ for $r < r_+ = 1$. In other words, $g(r_+)$ does not exist and $\lim_{r\to r_+} g(r) = \infty$, where the limit is over $r < r_+$. In this exercise, you are to assume that X is an arbitrary rv for which $g(r_+)$ does not exist and show that $\lim_{r\to r_+} g(r) = \infty$ where the limit is over $r < r_+$.

a) Explain why

$$\lim_{A\to\infty} \int_0^A e^{xr_+}\, d\mathsf{F}(x) = \infty.$$

b) Show that for any $\epsilon > 0$ and any $A > 0$,

$$g(r_+ - \epsilon) \geq e^{-\epsilon A} \int_0^A e^{xr_+}\, d\mathsf{F}(x).$$

c) Choose $A = 1/\epsilon$ and show that

$$\lim_{\epsilon\to 0} g(r_+ - \epsilon) = \infty.$$

Exercise 1.24. a) Assume that the MGF of the random variable X exists (*i.e.*, is finite) in the interval (r_-, r_+), $r_- < 0 < r_+$, and assume $r_- < r < r_+$ throughout. For any finite constant c, express the moment generating function of $X - c$, *i.e.*, $\mathsf{g}_{(X-c)}(r)$, in terms of $\mathsf{g}_X(r)$ and show that $\mathsf{g}_{(X-c)}(r)$ exists for all r in (r_-, r_+). Explain why $\mathsf{g}''_{(X-c)}(r) \geq 0$.

b) Show that $\mathsf{g}''_{(X-c)}(r) = [\mathsf{g}''_X(r) - 2c\mathsf{g}'_X(r) + c^2\mathsf{g}_X(r)]e^{-rc}$.

c) Use **a)** and **b)** to show that $\mathsf{g}''_X(r)\mathsf{g}_X(r) - [\mathsf{g}'_X(r)]^2 \geq 0$, Let $\gamma_X(r) = \ln \mathsf{g}_X(r)$ and show that $\gamma''_X(r) \geq 0$. Hint: Choose $c = \mathsf{g}'_X(r)/\mathsf{g}_X(r)$.

d) Assume that X is non-deterministic, *i.e.*, that there is no value of α such that $\Pr\{X = \alpha\} = 1$. Show that the inequality sign "$\geq$" may be replaced by "$>$" everywhere in a), b) and c).

Exercise 1.25. 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 rv with the value 1 if user i can use the system and 0 otherwise.

Exercise 1.26. Suppose the rv X is continuous and has the distribution function $\mathsf{F}_X(x)$. Consider another rv $Y = \mathsf{F}_X(X)$. That is, for each sample point ω such that $X(\omega) = x$, we have $Y(\omega) = \mathsf{F}_X(x)$. Show that Y is uniformly distributed in the interval 0 to 1.

Exercise 1.27. Let Z be an integer valued rv with the PMF $\mathsf{p}_Z(n) = 1/k$ for $0 \leq n \leq k-1$. Find the mean, variance, and moment generating function of Z. Hint: An elegant way to do this is to let U be a uniformly distributed continuous rv 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 1.28. a) Let Y be a nonnegative rv and $y > 0$ be some fixed number. Let A be the event that $Y \geq y$. Show that $y\,\mathbb{I}_A \leq Y$ (*i.e.*, that this inequality is satisfied for every $\omega \in \Omega$).

b) Use your result in part a) to prove the Markov inequality.

Exercise 1.29. a) Show that for any $0 < k < n$

$$\binom{n}{k+1} \leq \binom{n}{k} \frac{n-k}{k}.$$

b) Extend part a) to show that, for all $\ell \leq n - k$,

$$\binom{n}{k+\ell} \leq \binom{n}{k} \left[\frac{n-k}{k}\right]^{\ell}.$$

c) Let $\tilde{p} = k/n$ and $\tilde{q} = 1 - \tilde{p}$. Let S_n be the sum of n binary IID rv's with $\mathsf{p}_X(0) = q$ and $\mathsf{p}_X(1) = p$. Show that for all $\ell \leq n - k$,

$$\mathsf{p}_{S_n}(k+\ell) \leq \mathsf{p}_{S_n}(k) \left(\frac{\tilde{q}p}{\tilde{p}q}\right)^{\ell}.$$

d) For $k/n > p$, show that $\Pr\{S_n \geq kn\} \leq \frac{\tilde{p}q}{\tilde{p}-p}\,\mathsf{p}_{S_n}(k)$.

e) Now let ℓ be fixed and $k = \lceil n\tilde{p} \rceil$ for fixed $\tilde{p}$ such that $1 > \tilde{p} > p$. Argue that as $n \to \infty$,

$$\mathsf{p}_{S_n}(k+\ell) \sim \mathsf{p}_{S_n}(k) \left(\frac{\tilde{q}p}{\tilde{p}q}\right)^{\ell} \qquad \text{and} \quad \Pr\{S_n \geq kn\} \sim \frac{\tilde{p}q}{\tilde{p}-p}\,\mathsf{p}_{S_n}(k).$$

Exercise 1.30. A sequence $\{a_n;\, n \geq 1\}$ of real numbers has the limit 0 if for all $\epsilon > 0$, there is an $m(\epsilon)$ such that $|a_n| \leq \epsilon$ for all $n \geq m(\epsilon)$. Show that the sequences in parts a) and b) below satisfy $\lim_{n\to\infty} a_n = 0$ but the sequence in part c) does not have a limit.

a) $a_n = \frac{1}{\ln(\ln(n+1))}$

b) $a_n = n^{10} \exp(-n)$

c) $a_n = 1$ for $n = 10^{\ell}$ for each positive integer ℓ and $a_n = 0$ otherwise.

d) Show that the definition can be changed (with no change in meaning) by replacing ϵ with either $1/k$ or 2^{-k} for every positive integer k.

Exercise 1.31. Consider the moment generating function of a rv X as consisting of the following two integrals:

$$\mathsf{g}_X(r) = \int_{-\infty}^{0} e^{rx} d\mathsf{F}(x) + \int_{0}^{\infty} e^{rx} d\mathsf{F}(x).$$

In each of the following parts, you are welcome to restrict X to be either discrete or continuous.

a) Show that the first integral always exists (*i.e.*, is finite) for $r \geq 0$ and that the second integral always exists for $r \leq 0$.

b) Show that if the second integral exists for a given $r_1 > 0$, then it also exists for all r in the range $0 \leq r \leq r_1$.

c) Show that if the first integral exists for a given $r_2 < 0$, then it also exists for all r in the range $r_2 \leq r \leq 0$.

d) Show that the range of r over which $\mathsf{g}_X(r)$ exists is an interval from some $r_2 \leq 0$ to some $r_1 \geq 0$ (the interval might or might not include each endpoint, and either or both end point might be 0 or ∞).

e) Find an example where $r_1 = 1$ and the MGF does not exist for $r = 1$. Find another example where $r_1 = 1$ and the MGF does exist for $r = 1$. Hint: Consider $\mathsf{f}_X(x) = e^{-x}$ for $x \geq 0$ and figure out how to modify it to $\mathsf{f}_Y(y)$ so that $\int_0^\infty e^y \mathsf{f}_Y(y)\, dy < \infty$ but $\int_0^\infty e^{y+\epsilon y} \mathsf{f}_Y(y) = \infty$ for all $\epsilon > 0$.

Exercise 1.32. Let $\{X_n;\, n \geq 1\}$ be a sequence of independent but not identically distributed rv's. We say that the weak law of large numbers (WLLN) holds for this sequence if for all $\epsilon > 0$

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{S_n}{n} - \frac{\mathsf{E}\,[S_n]}{n}\right| \geq \epsilon\right\} = 0 \quad \text{where } S_n = X_1 + X_2 + \cdots + X_n. \qquad \text{(WL)}.$$

a) Show that the WLLN holds if there is some constant A such that $\sigma^2_{X_n} \leq A$ for all n.

b) Suppose that $\sigma^2_{X_n} \leq A\, n^{1-\alpha}$ for some $\alpha < 1$ and for all n. Show that the WLLN holds in this case.

Exercise 1.33. Let $\{X_i;\ i \geq 1\}$ be IID binary rv's. Let $\Pr\{X_i = 1\} = \delta$, $\Pr\{X_i = 0\} = 1 - \delta$. Let $S_n = X_1 + \cdots + X_n$. Let m be an arbitrary but fixed positive integer. Think! then evaluate the following and explain your answers:

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

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

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

Exercise 1.34. Use the Berry-Esseen result, (1.83), to prove the WLLN under the restriction that $\mathsf{E}\left[|X|^3\right]$ exists. Note: This is not intended as a reasonable way to prove the WLLN. Rather, it is to better understand what the convergence result of (1.83) implies. It appears that the CLT, without something extra about convergence, does not establish the WLLN.

Exercise 1.35. (Details in the proof of Theorem 1.5.3)

a) Show that if $X_1, X_2, \ldots$, are IID, then the truncated versions $\breve{X}_1, \breve{X}_2, \ldots$, are also IID.

b) Show that each $\breve{X}_i$ has a finite mean $\mathsf{E}\left[\breve{X}\right]$ and finite variance $\sigma^2_{\breve{X}}$. Show that the variance is upper bounded by the second moment around the original mean $\overline{X}$, *i.e.*, show that $\sigma^2_{\breve{X}} \leq \mathsf{E}\left[\,|\,\breve{X} - \mathsf{E}\left[X\right]\,|^2\right]$.

c) Assume that $\breve{X}_i$ is X_i truncated to $\overline{X} \pm b$. Show that $|\breve{X} - \overline{X}| \leq b$ and that $|\breve{X} - \overline{X}| \leq |X - \overline{X}|$. Use this to show that $\sigma^2_{\breve{X}} \leq b\mathsf{E}\left[|\breve{X} - \overline{X}|\right] \leq 2b\mathsf{E}\left[|X|\right]$.

d) Let $\breve{S}_n = \breve{X}_1 + \cdots + \breve{X}_n$ and show that for any $\epsilon > 0$,

$$\Pr\left\{\left|\frac{\breve{S}_n}{n} - \mathsf{E}\left[\breve{X}\right]\right| \geq \frac{\epsilon}{2}\right\} \leq \frac{8b\mathsf{E}\left[|X|\right]}{n\epsilon^2}.$$

e) Sketch the form of $\mathsf{F}_{\breve{X}-\overline{X}}(x)$ and use this, along with (1.35), to show that for all sufficiently large b, $\left|\,\mathsf{E}\left[\breve{X} - \overline{X}\right]\,\right| \leq \epsilon/2$. Use this to show that

$$\Pr\left\{\left|\frac{\breve{S}_n}{n} - \mathsf{E}\left[X\right]\right| \geq \epsilon\right\} \leq \frac{8b\mathsf{E}\left[|X|\right]}{n\epsilon^2} \qquad \text{for all large enough } b.$$

f) Use the following equation to justify (1.94).

$$\begin{aligned}\Pr\left\{\left|\frac{S_n}{n} - \mathsf{E}\left[X\right]\right| > \epsilon\right\} \;=\; & \Pr\left\{\left|\frac{S_n}{n} - \mathsf{E}\left[X\right]\right| > \epsilon \bigcap S_n = \breve{S}_n\right\} \\ & + \Pr\left\{\left|\frac{S_n}{n} - \mathsf{E}\left[X\right]\right| > \epsilon \bigcap S_n \neq \breve{S}_n\right\}.\end{aligned}$$

Exercise 1.36. Let $\{X_i;\ i \geq 1\}$ be IID rv's with mean 0 and infinite variance. Assume that $\mathsf{E}\left[|X_i|^{1+h}\right] \leq \beta$ for some given h, $0 < h < 1$ and some finite β. Let $S_n = X_1 + \cdots + X_n$.

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

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

Show that $\mathsf{E}\left[\breve{X}^2\right] \leq \frac{2\beta b^{1-h}}{1-h}$ Hint: For a nonnegative rv Z, $\mathsf{E}\left[X^2\right] \leq \int_0^\infty 2\,z\Pr\{Z \geq z\}\,dz$ (you can establish this, if you wish, by integration by parts).

c) Let $\breve{S}_n = \breve{X}_1 + \ldots + \breve{X}_n$. Show that $\Pr\left\{S_n \neq \breve{S}_n\right\} \leq n\beta b^{-1-h}$

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

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

Exercise 1.37. (**MS convergence → convergence in probability**) Assume that $\{Z_n;\, n \geq 1\}$ is a sequence of rv's and α is a number with the property that $\lim_{n\to\infty} \mathsf{E}\left[(Z_n - \alpha)^2\right] = 0$.

a) Let $\epsilon > 0$ be arbitrary and show that for each $n \geq 0$,

$$\Pr\{|Z_n - \alpha| \geq \epsilon\} \leq \frac{\mathsf{E}\left[(Z_n - \alpha)^2\right]}{\epsilon^2}.$$

b) For the ϵ above, let $\delta > 0$ be arbitrary. Show that there is an integer m such that $\mathsf{E}\left[(Z_n - \alpha)^2\right] \leq \epsilon^2\delta$ for all $n \geq m$.

c) Show that this implies convergence in probability.

Exercise 1.38. Let $X_1, X_2 \ldots$, be a sequence of IID rv's each with mean 0 and variance σ^2. Let $S_n = X_1 + \cdots + X_n$ for all n and consider the random variable $S_n/\sigma\sqrt{n} - S_{2n}/\sigma\sqrt{2n}$. Find the limiting distribution function for this sequence of $rv's$ as $n \to \infty$. The point of this exercise is to see clearly that the distribution function of $S_n/\sigma\sqrt{n}$ is converging but that the sequence of rv's is not converging.

Exercise 1.39. A town starts a mosquito control program and the rv Z_n is the number of mosquitos at the end of the nth 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 rv's with the PMF $\Pr\{X{=}2\} = 1/2$; $\Pr\{X{=}1/2\} = 1/4$; $\Pr\{X{=}1/4\} = 1/4$. 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 $\mathsf{E}[Z_n]$ as a function of n and find $\lim_{n\to\infty} \mathsf{E}[Z_n]$.

b) Let $W_n = \log_2 X_n$. Find $\mathsf{E}[W_n]$ and $\mathsf{E}[\log_2(Z_n/Z_0)]$ 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 rv's might be more significant that the expected value of the product itself.

Exercise 1.40. Use Figure 1.7 to verify (1.55). Hint: Show that $y\Pr\{Y{\geq}y\} \leq \int_{z\geq y} z d\mathsf{F}_Y(z)$ and show that $\lim_{y\to\infty} \int_{z\geq y} z d\mathsf{F}_Y(z) = 0$ if $\mathsf{E}[Y]$ is finite.

Exercise 1.41. Show that $\prod_{m\geq n}(1-1/m)=0$. Hint: Show that

$$\left(1-\frac{1}{m}\right)\Big\} = \exp\left(\ln\left(1-\frac{1}{m}\right)\right)\Big\} \leq \exp\left(-\frac{1}{m}\right).$$

Exercise 1.42. Consider a discrete rv X with the PMF

$$\begin{aligned} \mathsf{p}_X(-1) &= (1-10^{-10})/2, \\ \mathsf{p}_X(1) &= (1-10^{-10})/2, \\ \mathsf{p}_X(10^{12}) &= 10^{-10}. \end{aligned}$$

a) Find the mean and variance of X. Assuming that $\{X_m;\, m\geq 1\}$ is an IID sequence with the distribution of X and that $S_n = X_1+\cdots+X_n$ for each n, find the mean and variance of S_n. (no explanations needed.)

b) Let $n=10^6$ and describe the event $\{S_n \leq 10^6\}$ in words. Find an exact expression for $\Pr\{S_n \leq 10^6 \quad = \mathsf{F}_{S_n}(10^6)$.

c) Find a way to use the union bound to get a simple upper bound and approximation of $1-\mathsf{F}_{S_n}(10^6)$.

d) Sketch the distribution function of S_n for $n=10^6$. You can choose the horizontal axis for your sketch to go from -1 to $+1$ or from -3×10^3 to 3×10^3 or from -10^6 to 10^6 or from 0 to 10^{12}, whichever you think will best describe this distribution function.

e) Now let $n=10^{10}$. Give an exact expression for $\Pr\{S_n \leq 10^{10}$ and show that this can be approximated by e^{-1}. Sketch the distribution function of S_n for $n=10^{10}$, using a horizontal axis going from slightly below 0 to slightly more than 2×10^{12}. Hint: First view S_n as conditioned on an appropriate rv.

d) Can you make a qualitative statement about how the distribution function of a rv X affects the required size of n before the WLLN and the CLT provide much of an indication about S_n.

6.262 Discrete Stochastic Processes
Spring 2011