Modeling with Probability

P. Howard

Fall 2009

Contents

> "We see that the theory of probability is at bottom only common sense reduced to calculation; it makes us appreciate with exactitude what reasonable minds feel by a sort of instinct, often without being able to account for it....It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge....The most important questions of life are, for the most part, really only problems of probability."
>
> Pierre Simon, Marquis de Laplace, *Théorie Analytique des Probabilités*, 1812

> "The gambling passion lurks at the bottom of every heart."
>
> Honoré de Balzac

1 Introduction

Though games of chance have been around in one form or another for thousands of years, the first person to attempt the development of a systematic theory for such games seems to have been the Italian mathematician, physician, astrologer and—yes—gambler Gerolamo Cardano (1501–1506). Cardano is perhaps best known for his study of cubic and quartic algebraic equations, which he solved in his 1545 text *Ars Magna*—solutions which required his keeping track of $\sqrt{-1}$. He did not develop a theory of complex numbers, but is largely regarded as the first person to recognize the possibility of using what has now become the theory of complex numbers. He is also remembered as an astrologer who made many bold predictions, including a horoscope of Jesus Christ (1554). He is also known for having predicted the precise day on which he would die and then (or at least as the story goes) committing suicide on that day.[1]

In 1654, Antoine Gombaud Chevalier de Mere, a French nobleman and professional gambler called Blaise Pascal's (1623–1662) attention to a curious game of chance: was it worthwhile betting even money (original bet is either doubled or lost) that double sixes would turn up at least once in 24 throws of a pair of fair dice. This led to a long correspondence between Pascal and Pierre de Fermat (1601–1665, of Fermat's Last Theorem fame) in which the fundamental principles of probability theory were formulated for the first time. In these notes we will review a handful of their key observations.

2 Definitions and Axioms

Suppose we flip a fair coin twice and record each time whether it turns up heads (H) or tails (T). The list of all possible outcomes for this experiment is

$$S = \{(HH), (HT), (TH), (TT)\},$$

which constitutes a set that we refer to as the *sample space* for this experiment. Each member of S is referred to as an *outcome*. In this case, finding the probability of any particular outcome is straightforward. For example, $\Pr\{(HH)\} = 1/4$. Any subset, E, of the sample space is an *event*. In the example above, $E = \{(HH), (HT)\}$ is the event that heads appears on the first flip. Suppose $F = \{(TH), (TT)\}$; that is, the event that tails appears on the first flip. We have:

Definition 2.1. For any two sets (events) A and B, we define the following:

1. (Intersection) $A \cap B$ = all outcomes in both A and B (in our example, $E \cap F = \emptyset$, the empty set).

2. (Union) $A \cup B$ = all outcomes in either A or B (in our example, $E \cup F = S$).

3. (Complement) A^c = all outcomes in S but not in A (in our example, $E^c = F$).

One of the first men to systematically develop the theory of probability was Pierre Simon Laplace (1749–1827), who famously said, "At the bottom, the theory of probability is only common sense expressed in numbers." This is at least true in the sense that we develop our theory under the assumption of a set of axioms that cannot be proven from earlier principles, but which we regard as somehow self-evident.[2]

Axioms of Probability. For any sample space S, we have

[1] Now that's dedication to your profession.

[2] Another apropos quote here is from the (Welsh-born) English philosopher and mathematician Bertrand Russell (1872–1970), who wrote, "The axiomatic method has many advantages over honest work."

Axiom 1. $0 \leq Pr\{E\} \leq 1$, for all events E in S.
Axiom 2. $Pr\{S\} = 1$.
Axiom 3. If $E_1, E_2, ...$ are *mutually exclusive events* in S (that is, $E_k \cap E_j = \emptyset$ for $k \neq j$) then

$$\Pr\{E_1 \cup E_2 \cup ...\} = \Pr\{E_1\} + \Pr\{E_2\} + ...$$

We observe that in our simple calculation $\Pr\{(HH)\} = 1/4$, we have used Axioms 2 and 3. Without stating this explicitly, we used Axiom 3 to obtain the relation,

$$\Pr\{(HH) \cup (HT) \cup (TH) \cup (TT)\} = \Pr\{(HH)\} + \Pr\{(HT)\} + \Pr\{(TH)\} + \Pr\{(TT)\}.$$

According, then, to Axiom 2, we have

$$\Pr\{(HH)\} + \Pr\{(HT)\} + \Pr\{(TH)\} + \Pr\{(TT)\} = 1,$$

and we finally conclude our calculation by further assuming that each outcome is equally likely.

3 Counting Arguments

In our example in which a fair coin is flipped twice, we can compute probabilities simply by counting. To determine, for instance, the probability of getting both a head and a tail, we count the number of ways in which a head and a tail can both occur (2), and divide by the total number of possible outcomes (4). The probability is 1/2.

In these notes, we consider *permutations* and *combinations*, both of which can be understood in terms of the following simple rule for counting.

Basic Principle of Counting. If N experiments are to be carried out, and there are n_1 possible outcomes for the first experiment, n_2 possible outcomes for the second experiment and so on up to n_N possible outcomes for the final experiment, then altogether there are

$$n_1 \cdot n_2 \cdots n_N$$

possible outcomes for the set of N experiments.

3.1 Permutations

Consider the following question: how many different numbers can be made by rearranging the digits 1, 2, and 3. We have 123, 132, 213, 231, 312, and 321, six in all. We refer to each of these arrangements as a permutation. As a general rule, we have the following:

Rule of Permutations. For n distinct objects, there will be $n!$ (read: n *factorial*) permutations, where

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1.$$

(Of course the 1 can be omitted, but it provides symmetry and closure, so I've included it.) The rule of permutations follows immediately from the basic principle of counting, through the observation that there are n ways to choose the first item (in our example, 1, 2 or 3), $n-1$ ways to choose the second item (once the first item has been eliminated) and so on.

Example 3.1. How many four-letter permutations can be made from the letters in *okra*?

We have four different letters, so we can have $4! = 24$ permutations. $\triangle$

In the event that an object is repeated, we can similarly establish a similar rule. Consider, for example, the following question: how many different four-digit numbers can be made by rearranging the digits 1, 2, 2 and 3. We have twelve total: 2213, 2231, 2123, 2321, 2132, 2312, 1223, 3221, 1322, 3122, 1232, 3212. In order to develop a formula in this case, it's useful to begin by noting that if we had four distinct objects (say $1, A, B, 3$) the number of permuations would be 4!. However, permuations such as 1AB3 would actually be the same as 1BA3. In particular, *we overcount by a factor equal to the number of ways we can permute A and B.* In this way the number permuations is

$$\frac{4!}{2!} = 12.$$

The general rule is as follows.

Permutations with repeated objects. For n objects for which n_1 are identical, n_2 are identical, etc., with n_N also identical, the number of permutations is given by

$$\frac{n!}{n_1! n_2! \cdots n_N!}.$$

Example 3.2. How many nine-letter permutations can be made from the letters in *evergreen.*

We have nine letters with one repeated twice (r) and one repeated four times (e). This gives

$$\frac{9!}{2!4!} = 7560$$

permutations. $\triangle$

3.2 Combinations

We are often interested in counting the number of subsets we can create from some set of objects. For example, we might ask how many groups of three letters could be selected from the five letters A, B, C, D, and E. If order matters, we argue that there are five ways to select the first letter (we have five possible options), four ways to select to the second (once the first letter has been chosen, we only have four remaining to choose from), and three ways to select the third. That is, the number of possible selections can be computed as

$$5 \cdot 4 \cdot 3 = 60.$$

We observe, however, that this assumes the order of selection matters; that is, that the combination ABC is different from the combination BCA. When we talk about combinations, we will assume that order of selection *does not* matter, so the calculation above overcounts. In order to count the number of un-ordered combinations, we determine the number of ways in which the calculation above overcounts. For example, how many combinations have we counted that contain the letters A, B, and C. This is a permutation problem—how many ways can we permute A, B, and C—and the answer is $3! = 6$. Of course, we are overcounting every other combination of three letters by the same amount, so the total number of combinations is really

$$\frac{5 \cdot 4 \cdot 3}{3 \cdot 2 \cdot 1} = \frac{60}{6} = 10.$$

Rule of Combinations. In general, if we have n distinct objects and choose r, we have the number of combinations

$$\frac{n(n-1)(n-2)\cdots(n-(r-1))}{r!} = \frac{n!}{r!(n-r)!}.$$

We make the following definition, typically read n *choose* r,

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

Example 3.3. The Texas Megamillions lottery works as follows: Six numbers are chosen, each between 1 and 52. Five of the numbers must be different from one another, while one can be anything. How many possible combinations are there?

First, we determine the number of combinations for five different numbers, selected from 52 possible. This is a standard combination problem, and we have,

$$\binom{52}{5} = \frac{52!}{5!(52-5)!} = 2598960.$$

There remain now 52 ways we can combine this numbers with our final number, and so the total number of possible number is

$$52 \cdot 2598960 = 135,145,920,$$

which is to say that a player's chances of choosing the correct one is

$$\frac{1}{135,145,920}.$$

△

While we're talking about the lottery, consider the following oddity. On February 20, 2004 the jackpot for the Texas Megamillions lottery was \$230, 000, 000. Tickets for this lottery are \$1.00, so in theory a player could play all 135, 145, 920 numbers and be assured of winning. In fact, of course, this is a dubious plan, since if someone else happens to pick the number as well, the player will have to share his winnings. Not to mention the logistics of buying this many tickets.

Example 3.4. Determine the probability of obtaining each of the following poker hands if five cards are dealt from a standard deck of 52 cards: (1) straight flush, (2) four of a kind (quad), (3) full house[3], (4) flush, (5) straight, (6) three of a kind (trip), (7) two pair[4], (8) one pair, (9) high-card hand (i.e., none of the above).

First, since there are 52 distinct cards in a standard deck the total number of possible five card hands is

$$\binom{52}{5} = 2,598,960.$$

1. Straight flush. Straights can be categorized by highest card: there are precisely ten rank arrangements possible, with high cards 5 through A. For each of these ten rank arrangements there are four suit arrangements that will make the straight a straight flush. This means:

$$\text{Number possible straight flushes} = 10 \cdot 4 = 40.$$

Consequently the probability of getting a straight flush is

$$\Pr\{\text{straight flush}\} = \frac{40}{2598960} = .000015.$$

We observe that there are precisely 4 ways to get a royal flush.

2. Four of a kind. There are 13 ways to rank the quad (i.e., 13 possible things to have four of) and 48 ways to choose the fifth card, so we have:

$$\text{Number possible quads} = 13 \cdot 48 = 624.$$

Consequently the probability of getting four of a kind is

$$\Pr\{\text{Four of a kind}\} = \frac{624}{2598960} = .000240.$$

3. Full house. For a full house there are 13 ways to rank the trip and $\binom{4}{3} = 4$ ways to arrange it, then 12 ways to rank the pair and $\binom{4}{2} = 6$ ways to arrange it. We have:

$$\text{Number possible full houses} = (13 \cdot 4) \cdot (12 \cdot 6) = 3744.$$

Consequently the probability of getting a full house is

$$\Pr\{\text{Full house}\} = \frac{3744}{2598960} = .001441.$$

4. Flush (and not a straight flush). We have four ways to choose the suit for a flush and $\binom{13}{5} = 1287$ ways to arrange the ranks. Finally, we will subtract off the 40 straight flushes we've already considered. We have

$$\text{Number possible flushes (not straight flushes)} = 4 \cdot 1287 - 40 = 5,108.$$

[3]A full house is sometimes referred to as a *boat*, but while the phrases four-of-a-kind and three-of-a-kind seem clunky enough to warrant an abbreviation, full house does not.

[4]The word pair can be pluralized as either pairs or pair, with (I would suggest) pairs preferred in most circumstances. In the case of two pair poker hands it's traditional to pluralize with pair.

Consequently the probability of getting a flush (and not a straight flush) is

$$\Pr\{\text{Flush (and not a straight flush)}\} = \frac{5108}{2598960} = .001965.$$

5. Straight (and not a straight flush). As in our discussion of straight flushes, we first note that there are ten possible rank arrangements for a straight, corresponding with top cards 5 through A. For each such rank arrangement there are four ways to suit each of the five cards. Finally, we subtract off the straight flushes. This gives:

$$\text{Number of straights (not straight flushes)} = 10 \cdot 4^5 - 40 = 10,200.$$

Consequently the probability of getting a straight (and not a straight flush) is

$$\Pr\{\text{Straight (and not a straight flush)}\} = \frac{10200}{2598960} = .003925.$$

6. Three of a kind. We have 13 ways to choose the rank of the trip, $\binom{4}{3} = 4$ ways to arrange the trip, $\binom{12}{2} = 66$ ways to choose two new ranks (which must be different to avoid a full house), and 4 ways to suit each of these ranks. (Notice that we have no problem overcounting flushes or straights since no hand with a trip can be a flush or a straight.) We have, then

$$\text{Number possible trips}{=}13 \cdot 4 \cdot 66 \cdot 4^2 = 54,912.$$

Consequently the probability of getting three of a kind is

$$\Pr\{\text{Three of a kind}\} = \frac{54912}{2598960} = .021128.$$

7. Two pair. We have $\binom{13}{2} = 78$ ways to choose the two different ranks to be paired, $\binom{4}{2} = 6$ ways to arrange each pair, and 44 ways to pick the remaining card. In total,

$$\text{Number possible two pair hand} = 78 \cdot 6^2 \cdot 44 = 123,552.$$

Consequently the probability of getting two pair is

$$\Pr\{\text{two pair hand}\} = \frac{123552}{2598960} = .047539.$$

8. One pair. We have 13 ways to rank the pair and $\binom{4}{2} = 6$ ways to arrange it. There are $\binom{12}{3} = 220$ ways to rank the remaining three cards so that no two of the ranks are the same, and four ways to suit each. We have:

$$\text{Number possible one pair hands} = 13 \cdot 6 \cdot 220 \cdot 4^3 = 1,098,240.$$

Consequently the probability of getting one pair is

$$\Pr\{\text{One pair}\} = \frac{1098240}{2598960} = .422569.$$

9. High-card hand. Summing all hands considered in (1) through (8) we have 1,296,420 hands. This leaves 1,302,540 hands that are ranked only by the ranks of the individual cards. We refer to these as high card hands. We have:

$$\Pr\{\text{High card hand}\} = \frac{1302540}{2598960} = .501177.$$

Alternatively, we can check our calculations by computing the number of high-card hands directly. There are $\binom{13}{5} = 1287$ ways to choose five distinct ranks and four ways to suit each of the ranks. We must subtract flushes, straights, and straight flushes, all of which are created with five distinct ranks. That is:

$$\text{Number high card hands} = 1287 \cdot 4^5 - 40 - 5108 - 10200 = 1302540,$$

as above. △

4 Discrete Random Variables

Suppose that in our experiment of flipping a coin twice, we assigned a numerical value to each outcome, referred to as X: $X(HH) = 1$, $X(HT) = 2$, $X(TH) = 3$, and $X(TT) = 4$. For instance, we might be considering a game in which X represents the payoff for each possible outcome. (Below, we will refer to this game as the "two-flip game.") We refer to functions defined on sample spaces as *random variables*. We refer to the values random variables can take as *realizations*. Random variables will be our main probabilistic interest in these notes. They represent such processes as:

- The value of a stock at a given time
- The number that wins in a game of roulette
- The time it takes to check out at a supermarket

Random variables that can take only a countable number of values are called *discrete*. (Recall that a set is said to be *countable* if its elements can be enumerated 1, 2, 3, ... The set of all rational numbers (integer fractions) is countable; the set of all real numbers is not.)

We will define events with respect to random variables in the forms $\{X = 1\}$, $\{X \leq 3\}$, $\{X \geq 2\}$ etc., by which we mean the event in our sample space for which X satisfies the condition in brackets. For example, $\{X = 1\} = \{(HH)\}$ and $\{X \leq 2\} = \{(HH), (HT)\}$.

4.1 Cumulative Distribution Functions

The *cumulative distribution function*, $F(x)$, for a random variable X is defined for all real $-\infty < x < +\infty$ as

$$F(x) = \Pr\{X \leq x\}.$$

For X as in the two-flip game above, we have

$$F(x) = \begin{cases} 0, & -\infty < x < 1 \\ 1/4, & 1 \le x < 2 \\ 1/2, & 2 \le x < 3 \\ 3/4, & 3 \le x < 4 \\ 1, & 4 \le x < \infty, \end{cases}$$

depicted graphically in Figure 4.1.

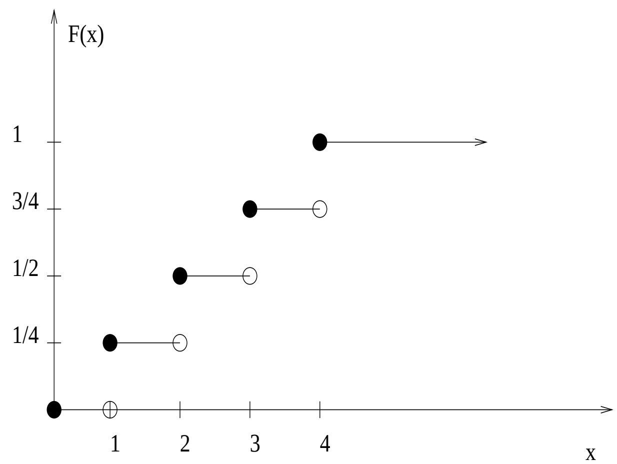


Figure 4.1: Cumulative distribution function for the two-flip game.

Below, we list for easy reference four critical properties of cumulative distribution functions, $F(x)$:

1. $F(x)$ is a non-decreasing function.
2. $\lim_{x\to+\infty} F(x) = 1$.
3. $\lim_{x\to-\infty} F(x) = 0$.
4. $F(x)$ is "right-continuous": $\lim_{y\to x^+} F(y) = F(x)$.

4.2 Probability Mass Function

The *probability mass function,* $p(x)$, for a discrete random variable X is defined by the relation

$$p(x) = \Pr\{X = x\}.$$

For example, in the two-flip game, $p(1) = p(2) = p(3) = p(4) = 1/4$. Below, we list for easy reference three critical properties of probability mass functions, $p(x)$.

1. $p(x)$ is 0 except at realizations of the random variable X.
2. $\sum_{\text{All possible } x} p(x) = 1$.

3. $F(y) = \sum_{x \le y} p(x)$.

Important examples of probability mass functions include the *Poisson*, the *Bernoulli* the *binomial*, and the *geometric.*

1. Poisson probability mass function. A random variable N that takes values $0, 1, 2, 3, ...$ is said to be a *Poisson* random variable with parameter a if for some $a > 0$ its probability mass function is given by

$$p(k) = \Pr\{N = k\} = e^{-a}\frac{a^k}{k!}.$$

2. Bernoulli probability mass function. A random variable N that takes values 0 and 1 is said to be a *Bernoulli* random variable with probability p if its probability mass function is given by

$$p(k) = \Pr\{N = k\} = \begin{cases} 1-p, & k = 0 \\ p, & k = 1. \end{cases}$$

A single flip of a coin is a Bernoulli process with $p = \frac{1}{2}$.

3. Binomial Probability mass function. A random variable X that takes values $0, 1, 2, ..., n$ is said to be a *binomial* random variable with sample size n and probability p if its probability mass function is given by

$$p(k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{n-k},$$

where $\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$ and is read as n *choose* k. The binomial random variable counts the number of probability p events in n trials of a Bernoulli process with probability p. For example, suppose we would like to determine the probability that 3 ones turn up in 5 rolls of a fair die. In this case, $p = \frac{1}{6}$ (the probability of a one) and $n = 5$, the number of rolls. We have

$$p(3) = \left(\begin{array}{c} 5 \\ 3 \end{array} \right) (\frac{1}{6})^3(\frac{5}{6})^2 = \frac{5!}{2!3!}(\frac{1}{6})^3(\frac{5}{6})^2 = .032.$$

In order to see that the binomial probability mass function satisfies condition (2) above, we recall the binomial expansion for any integers a, b, and n,

$$(a+b)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) a^k b^{n-k}.$$

We have, then, for the binomial probability mass function,

$$\sum_{k=0}^{n} p(k) = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{n-k} = (p + (1-p))^n = 1^n = 1. \tag{4.1}$$

4. Geometric probability mass function. A random variable X that takes values $1, 2, 3, ...$ is said to be a *geometric* random variable with probability p if its probability mass function is given by

$$p(k) = (1-p)^{k-1}p.$$

A geometric random variable counts the number of trials required of a Bernoulli process with probability p to get a probability p event. In this case, $E[X] = \frac{1}{p}$ and $\mathrm{Var}[X] = \frac{1-p}{p^2}$.

4.3 Conditional Probability

Often, we would like to compute the probability that some event occurs, given a certain amount of information. In our two-flip game above, suppose that we are given the information that the first flip is heads (H). The probability that $X = 2$ given that the first flip is heads is

$$\Pr\{X = 2|\text{First flip heads}\} = \frac{1}{2}.$$

Definition 4.1. (Conditional probability) Suppose A and B are events on a sample space S and $\Pr\{B\} \neq 0$. Then we define the conditional probability of event A given that event B has occurred as

$$\Pr\{A|B\} = \frac{\Pr\{A \cap B\}}{\Pr\{B\}}.$$

For the example above, $A = \{HT\}$ and $B = \{(HT), (HH)\}$ so that $A \cap B = \{HT\}$, $\Pr\{A \cap B\} = \frac{1}{4}$, and $\Pr\{B\} = \frac{1}{2}$.

Justification for the definition. The main observation here is that new information reduces the size of our sample space. In the example above, the information that the first flip is heads reduces the sample space from $\{(HH), (HT), (TH), (TT)\}$ to $\{(HH), (HT)\}$. Prior to reduction, the probability that $X = 2$ is $\frac{1}{4}$—one chance in the four possibilities. After reduction, the probability is increased to $\frac{1}{2}$—one chance in two possibilities.

In general, the sample space is reduced by discarding each event not in B. Since B certainly occurs, A will only occur in the event that $A \cap B$ does. Therefore, $\Pr\{A|B\}$ is the probability of $A \cap B$ relative to this reduced sample space. For a sample space consisting of equally likely outcomes, we have,

$$\Pr\{B\} = \frac{\text{\# of outcomes in B}}{\text{\# of outcomes in S}}; \quad \Pr\{A \cap B\} = \frac{\text{\# of outcomes in both A and B}}{\text{\# of outcomes in S}},$$

so that

$$\begin{aligned}\Pr\{A|B\} &= \frac{\text{\# of outcomes in both A and B}}{\text{\# of outcomes in B}} \\ &= \frac{\text{\# of outcomes in both A and B}}{\text{\# of outcomes in S}} \cdot \frac{\text{\# of outcomes in S}}{\text{\# of outcomes in B}}.\end{aligned}$$

Keep in mind here that an event is said to occur if any outcome in that event occurs.

Example 4.1. Suppose two fair dice are rolled. What is the probability that at least one lands on six, given that the dice land on different numbers?

Let A be the event of at least one die landing on six, and let B be the event that the dice land on different numbers. We immediately see by counting outcomes that $\Pr\{B\} = \frac{30}{36} = \frac{5}{6}$. On the other hand, the probability of 1 six and 1 non-six is the number of possible combinations with exactly one six (10) divided by the total number of possible combinations (36):

$$\Pr\{A \cap B\} = \Pr\{\text{1 six, 1 not-six}\} = \frac{\text{\# combinations with 1 six, 1 not-six}}{\text{36 total combination possible}} = \frac{10}{36} = \frac{5}{18}.$$

Conseqently,

$$\Pr\{A|B\} = \frac{\Pr\{A \cap B\}}{\Pr\{B\}} = \frac{\frac{5}{18}}{\frac{5}{6}} = \frac{1}{3}.$$

△

Lemma 4.2. (Bayes' Lemma) Suppose the events $A_1, A_2, ..., A_n$ form a partition of a sample space S. (That is, the events are mutually exclusive (see Axiom 3), and $\bigcup_{k=1}^n A_k = A_1 \cup A_2 \cup ... \cup A_n = S$.) Then if B is an event in S and $\Pr\{B\} \neq 0$, we have,

$$\Pr\{A_k|B\} = \frac{\Pr\{B|A_k\}\Pr\{A_k\}}{\sum_{j=1}^n \Pr\{B|A_j\}\Pr\{A_j\}}.$$

Proof. We first observe that since the A_k form a partition of S, we can write,

$$B = B \cap S = B \cap (\bigcup_{k=1}^n A_k) = \bigcup_{k=1}^n (B \cap A_k).$$

According to Axiom 3, then,

$$\Pr\{B\} = \Pr\{\bigcup_{j=1}^n (B \cap A_j)\} = \sum_{j=1}^n \Pr\{B \cap A_j\} = \sum_{j=1}^n \Pr\{B|A_j\}\Pr\{A_j\}, \tag{4.2}$$

where the final equality follows from the definition of conditional probability. We have, then,

$$\Pr\{A_k|B\} = \frac{\Pr\{A_k \cap B\}}{\Pr\{B\}} = \frac{\Pr\{B|A_k\}\Pr\{A_k\}}{\sum_{j=1}^n \Pr\{B|A_j\}\Pr\{A_j\}},$$

where the numerator is due again to the definition of conditional probability and the denominator is precisely (4.2). □

Example 4.2. (The infamous Monty Hall problem.[5]) Consider a game show in which a prize is hidden behind one of three doors. The contestant chooses a door, and then the host opens one of the two unchosen doors, showing that the prize is not behind it. (He never opens the door that the prize is behind.) The contestant then gets the option to switch doors. Given this scenario, should a contestant hoping to optimize his winnings, (1) always switch doors, (2) never switch doors, or (3) doesn't matter?

Though we can argue a solution to this problem on intuitive grounds (be warned: the argument might not be the first one you think of), we will work through the details as an application of Bayes' lemma. We will determine the probability that the prize is behind the first door the contestant selects, given that the host opens one of the other doors. We begin by defining a set of appropriate events, in which the doors the constestant does not open are generically labeled *alternative door number 2* and *alternative door number 3*:

$$\begin{aligned} A_1 &= \text{event that prize is behind first door selected} \\ A_2 &= \text{event that prize is behind alternative door 1} \\ A_3 &= \text{event that prize is behind alternative door 2} \\ B &= \text{event that host opens alternative door 1} \end{aligned}$$

[5] See http://www.shodor.org/interactivate/activities/monty3/

According to our Bayes' Lemma, we have

$$\begin{aligned}\Pr\{A_1|B\} =&\frac{\Pr\{B|A_1\}\Pr\{A_1\}}{\Pr\{B|A_1\}\Pr\{A_1\}+\Pr\{B|A_2\}\Pr\{A_2\}+\Pr\{B|A_3\}\Pr\{A_3\}}\\ =&\frac{\frac{1}{2}\cdot\frac{1}{3}}{\frac{1}{2}\cdot\frac{1}{3}+0+1\cdot\frac{1}{3}}=\frac{\frac{1}{6}}{\frac{1}{2}}=\frac{1}{3}.\end{aligned}$$

Observe that since the host's opening alternative door 1 is entirely arbitrary, we can make the same calculation given that he opens alternative door 2. Therefore, whichever door he opens, the contestant who sticks with his initial choice only has a 1/3 chance of being right. Which, of course, means that the contestant should always switch doors, giving him a 2/3 chance of winning.

Intuitively, we can regard this game as follows: If the contestant chooses not to switch doors, his only chance of winning is if his original choice was correct, and his odds for that are clearly 1/3. On the other hand, since the host removes one of the two doors not selected, if the contestant switches doors he wins so long as his original choice was incorrect. △

Remark. In many applications of Bayes' Lemma, it is not immediately clear that the conditioning set B is in the sample space S. For example, in the Monty Hall problem, the events that partition S correspond only with possible locations of the prize and seem to have nothing to do with which door the host opens. In such cases, it is important to keep in mind that while a set of outcomes entirely describes S, a set of events can hide individual outcomes. In order to see this more clearly, notice that

$$B^c = \text{event the host opens alternative 2}.$$

Then we can write the sample space for the Monty Hall problem as follows:

$$S = \{(A_1\cap B), (A_1\cap B^c), (A_2\cap B^c), (A_3\cap B)\}.$$

Here $A_1 = \{(A_1\cap B), (A_1\cap B^c)\}$, $A_2 = \{(A_2\cap B^c)\}$, $A_3 = \{(A_3\cap B)\}$, and so the A_k clearly partition S. Moreover, $B = \{(A_1\cap B), (A_3\cap B)\}$, which is clearly a subset of S.

4.4 Independent Events and Random Variables

An important concept in the study of probability theory is that of *independence*. Loosely speaking, we say that two random variables are independent if the outcome of one is in no way correlated with the outcome of the other. For example, if we flip a fair coin twice, the result of the second flip is independent of the result of the first flip. On the other hand, the random variable X in our two-flip game is certainly not independent of the outcome of the first flip. More precisely, we have the following definition.

Definition 4.3. (Independence) Two discrete random variables X and Y are said to be independent if

$$\Pr\{X=x|Y=y\} = \Pr\{X=x\}.$$

Likewise, two events A and B are said to be independent if

$$\Pr\{A|B\} = \Pr\{A\}.$$

From the definition of conditional probability, we can derive the critical relation

$$\Pr\{\{X = x\} \cap \{Y = y\}\} = \Pr\{X = x|Y = y\}\Pr\{Y = y\} = \Pr\{X = x\}\Pr\{Y = y\}.$$

Example 4.3. Compute the probability that double sixes will turn up at least once in 24 throws of a pair of fair dice.

We begin this calculation by computing the probability that double sixes does not occur even once in the 24 throws. On each trial the probability that double sixes will not occur is $\frac{35}{36}$, and so by independence the probability of the intersection of 24 of these events in a row is $(\frac{35}{36})^{24}$. We conclude that the probability that double sixes *do* turn up once in 24 throws is (to four decimal places of accuracy)

$$1 - (\frac{35}{36})^{24} = .4914.$$

$\triangle$

4.5 Expected Value

Often, we would like to summarize information about a particular random variable. For example, we might ask how much we could expect to make playing the two-flip game. Put another way, we might ask, how much would we make on average if we played this game repeatedly a sufficient number of times. In order to compute this *expected value*, we multiply the amount we win from each outcome with its probability and sum. In the case of the two-flip game, we have

$$\text{Expectation} = \$1.00 \times \frac{1}{4} + \$2.00 \times \frac{1}{4} + \$3.00 \times \frac{1}{4} + \$4.00 \times \frac{1}{4} = \$2.50.$$

It's important to notice that we will never actually make \$2.50 in any single play of the game. But if we play it repeatedly for a sufficient length of time, our average winnings will be \$2.50. Denoting expectation by E, we summarize this critical expression as

$$E[X] = \sum_{x \text{ Possible}} x\Pr\{X = x\}.$$

Example 4.4. Suppose a man counting cards at the blackjack table knows the only cards not yet dealt are a pair of fours, three nines, a ten, two Jacks, and a King. What is the expected value of his next card?

Keeping in mind that tens, Jacks and Kings are all worth ten points, while fours and nines are worth face value, we compute

$$E[\text{Next card drawn}] = 4 \cdot \frac{2}{9} + 9 \cdot \frac{3}{9} + 10\frac{4}{9} = \frac{25}{3}.$$

$\triangle$

Example 4.5. The game of (American) roulette employs a large wheel with 38 slots, thirty-six labeled nonsequentially by the numbers 1–36, and two house slots, labeled 0 and 00.[6] Exactly half of the thirty-six numbered slots are red, while the other half are black. (The two house slots are typically shaded green.) In each play of the game, players make bets on certain numbers or sequences of numbers until the *croupier*[7] calls out "No more bets," usually while the wheel is already in motion, but before it would be possible for anyone to guess where the ball he's spun into it might land.

The simplest bet in roulette is on red or black, which returns even money (the player betting one dollar either loses his dollar or wins one dollar). The expected value of a one dollar bet on black can easily be computed as,

$$E[\text{One dollar on black}] = 1 \cdot \frac{18}{38} - 1 \cdot \frac{20}{38} = -.0526 \text{ cents},$$

with variance

$$Var[\text{One dollar bet on black}] = 1.0526^2 \cdot \frac{18}{38} + .9474^2 \cdot \frac{20}{38} = .9972.$$

On the other extreme, we could bet on a single number, which pays 35:1 ($35 dollar win for $1 bet) giving an expected value

$$E[\text{Single Number}] = 35 \cdot \frac{1}{38} - 1 \cdot \frac{37}{38} = -.0526.$$

One famous gambling stragedy that is often used in roulette, though is applicable to just about any game, is known as the "double-up", or *Martingale* strategy. The idea behind the Martingale strategy is extremely simple: At every play of the game, if the player wins, he collects his money and repeats his initial bet, starting over; if he loses, he doubles his bet *on the same thing*. For example, a player might use this strategy betting on black in roulette. He begins by betting, say, one dollar on black. If he wins, he collects the dollar he wins and starts over, betting a single dollar on black again. If he loses, he bets two dollars on black. Now, if he wins, he collects two dollars from the croupier, which even counting the bill he's already lost leaves him ahead by $1. If he loses, he doubles his bet again to four dollars. Assuming black eventually turns up, then no matter how long it takes for him to get there, the player will eventually make $1. He can't lose, right? Well...unfortunately, there's some fine print. True, the mathematical expectation of this strategy is $1—you really do, in theory, make a buck every time you use it. The problem is that casinos have betting limits (and gamblers have capital limits), so eventually the doubling has to stop. In order to see what effect betting limits have on the expectation, let's suppose there's an extremely small cap of $4. If we let X be the player's final payout from one series with the strategy, we have

$$E[X] = 1 \cdot \Big(\frac{18}{38} + \frac{20}{38} \cdot \frac{18}{38} + (\frac{20}{38})^2 \frac{18}{38}\Big) - 7(\frac{20}{38})^3 = -.1664.$$

(Notice that the probability in parentheses is the probability that the spin lands black exactly once in three consecutive plays.)

[6]In European roulette, there is only one house slot, 0.

[7]French for "dealer."

At first glance, it might appear that the player's expected value is getting worse, but the situation isn't quite that bad. Notice that the player's bet is no longer a single dollar, but rather varies depending upon the turn of the wheel. His expected bet under the assumption of a \$4 cap is given by,

$$E[B] = 1 \cdot \frac{18}{38} + 3 \cdot \frac{20}{38} \cdot \frac{18}{38} + 7 \cdot (1 - \frac{18}{38} - \frac{20}{38} \cdot \frac{18}{38}) = 3.1607.$$

(Notice that in this calculation the first addened on the right-hand side represents the player's bet in the event that he bets exactly one dollar, which only occurs in the event that he wins on the first roll, and the others are similar.) The player's expectation *per dollar bet* becomes

$$\text{Expectation per dollar bet} = \frac{E[X]}{E[B]} = \frac{-.1664}{3.1607} = -.0526.$$

So, in fact, the player hasn't helped (or hurt) himself at all. △

Example 4.6. (St. Petersburg Paradox, suggested by Daniel and Nicolaus Bernoulli around 1725.) Suppose a dealer says that he will flip a fair coin until it turns up heads and will pay you 2^n dollars, where n is the number of flips it takes for the coin to land heads. How much would you be willing to pay in order to play this game?

I ask this question each semester, and so far no one has offered to pay more than five dollars. Most students won't go much above two. In order to determine the expected value of this game, let the random variable X represent the payoff. If the coin lands heads on the first flip the payoff is \$2, with probability $\frac{1}{2}$. If the coin does not land heads until the second flip, the payoff is $2^2 = 4$, with probability $\frac{1}{4}$—the probability of a tail followed by a head. Proceeding similarly, we see that

$$E[X] = 2 \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} + 8 \cdot \frac{1}{8} + ... = 1 + 1 + 1 + ... = \infty.$$

The expected value of this game is infinite! Which means that *ideally* we should be willing to pay any amount of money to play it. But almost no one is willing to pay more than about five bucks. This is what the brothers Bernoulli considered a paradox.

In order to resolve this, we need to keep in mind that the expected value of a game reflects our average winnings if we continued to play a game for a sufficient length of time. Suppose we pay five dollars per game. Half the time we will lose three dollars ($2^1 - 5 = -3$), while another quarter of the time we will lose one dollar ($2^2 - 5 = -1$). On the other hand, roughly one out of every sixty-four times (6 flips) we will make $2^6 - 5 = 59$. The point is that though we lose more often than we win, we have the chance to win big. Practically speaking, this means that two things come into play when thinking about the expected value of a game: the expected value itself and the number of times you will get a chance to play it. Yet one more way to view this is as follows. The fact that this game has infinite expectation means that no matter how much the dealer charges us to play—\$5.00, \$5 million, etc.—the game is worth playing (i.e., we will eventually come out ahead) *so long as we can be sure that we will be able to play it enough times.* △

Example 4.7. Compute the expected value for the binomial distribution.

Recall from Section 4.2 that a random variable X that takes values $0, 1, 2, ..., n$ is said to be a *binomial* random variable with sample size n and probability p if its probability mass function is given by

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

We compute

$$\begin{aligned} E[X] &= \sum_{k=0}^{n} kp(k) = \sum_{k=1}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} k \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k}. \end{aligned}$$

Letting $l = k - 1$, we have

$$E[X] = \sum_{l=0}^{n-1} n \frac{(n-1)!}{l!((n-1)-l)!} p^l p (1-p)^{(n-1)-l} = np,$$

where in the last step we have used (4.1). $\triangle$

4.6 Properties of Expected Value

In what follows, we require the following preliminary observation.

Claim. For any two discrete random variables X and Y, we have

$$\sum_x \Pr\{\{X = x\} \cap \{Y = y\}\} = \Pr\{Y = y\}.$$

Proof of claim. We first observe that

$$\bigcup_x \{X = x\} = S,$$

from which Axiom 2 provides,

$$\begin{aligned} \sum_x \Pr\{\{X = x\} \cap \{Y = y\}\} &= \Pr\{\bigcup_x (\{X = x\} \cap \{Y = y\})\} \\ &= \Pr\{\{Y = y\} \cap (\bigcup_x \{X = x\})\} \\ &= \Pr\{\{Y = y\} \cap S\} = \Pr\{Y = y\}. \end{aligned}$$

$\square$

Lemma 4.4. For any constant c and random variable X,

$$E[cX] = cE[X].$$

Proof. Define $Y = cX$. According to the definition of expectation in the discrete case,

$$E[Y] = \sum_y y\Pr\{Y = y\} = \sum_x cx\Pr\{cX = cx\}$$
$$= \sum_x cx\Pr\{X = x\} = c\sum_x x\Pr\{X = x\} = cE[X].$$

□

Lemma 4.5. For any two random variables X and Y,

$$E[X + Y] = E[X] + E[Y].$$

Proof. While the lemma is true for both discrete and continuous random variables, we carry out the proof only in the discrete case. Computing directly from our definition of expected value, and using our claim from the beginning of this section, we have

$$\begin{aligned} E[X + Y] &= \sum_{x,y}(x + y)\Pr(\{X = x\} \cap \{Y = y\}) \\ &= \sum_{x,y} x\Pr(\{X = x\} \cap \{Y = y\}) + \sum_{x,y} y\Pr(\{X = x\} \cap \{Y = y\}) \\ &= \sum_x x \sum_y \Pr(\{X = x\} \cap \{Y = y\}) + \sum_y y \sum_x \Pr(\{X = x\} \cap \{Y = y\}) \\ &= \sum_x x\Pr\{X = x\} + \sum_y y\Pr\{Y = y\} = E[X] + E[Y]. \end{aligned}$$

□

Lemma 4.6. For any discrete random variable X and continuous function $g(x)$,

$$E[g(X)] = \sum_x g(x)\Pr\{X = x\}.$$

Proof. Observing that $Y = g(X)$ is a new random variable, we compute directly from our definition of expected value,

$$E[g(X)] = \sum_{g(x)} g(x)\Pr\{g(X) = g(x)\}.$$

We notice that by the continuity of $g(x)$, $\Pr\{g(X) = g(x)\} \geq \Pr\{X = x\}$. That is, if $X = x$, then $g(X) = x$, but there may be several values of x that give the same value of $g(x)$ (e.g., $+1$ and -1 for $g(x) = x^2$). We observe, however, that

$$g(x)\Pr\{g(X) = g(x)\} = \sum_{\{y:g(x)=g(y)\}} g(y)\Pr\{X = y\},$$

which establishes the lemma. □

Lemma 4.7. For any two independent random variables X and Y,

$$E[XY] = E[X]E[Y].$$

Proof. See homework for a proof in the discrete case. □

4.7 Conditional Expected Value

As with probability, we often would like to compute the expected value of a random variable, given some information. For example, suppose that in the two-flip game we know that the first flip lands heads. The expected value of X given this information is computed as,

$$E[X|\text{First flip heads}] = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2} = 1.5.$$

More generally, we have that for any event A

$$E[X|A] = \sum_x x\Pr\{X = x|A\}.$$

If we define our event in terms of the value of a random variable Y, we can write this as

$$E[X|Y = y] = \sum_x x\Pr\{X = x|Y = y\}.$$

In order to understand this, we simply observe that $Z = [X|Y = y]$ is a new random variable whose expectation can be computed directly from the definition as

$$E[X|Y = y] = \sum_x x\Pr\{(X|Y = y) = x\} = \sum_x x\Pr\{X = x|Y = y\}.$$

Conditional expectation can also be a critical tool in the calculation of ordinary expectation. In this regard, we have the following lemma.

Lemma 4.8. Suppose the events $A_1, A_2, ...$ form a partition of some sample space. Then for any random variable X

$$E[X] = \sum_k E[X|A_k]\Pr\{A_k\}.$$

Equivalently, for any discrete random variables X and Y,

$$E[X] = \sum_y E[X|Y = y]\Pr\{Y = y\}.$$

Proof. See homework. □

Example 4.7 (The frustrated mouse). A certain mouse is placed in the center of a maze, surrounded by three paths that open with varying widths. The first path returns him to the center after two minutes; the second path returns him to the center after four minutes; and the third path leads him out of the maze after one minute. Due to the differing widths, the mouse chooses the first path 50% of the time, the second path 30% of the time, and the third path 20% of the time. Determine the expected number of minutes it will take for the mouse to escape.

Let M be a random variable representing the number of minutes until the mouse escapes, and let D represent which door the mouse chooses. Expected values conditioned on D are easy to calculate. For Door 3, $E[M|D = 3] = 1$. That is, if the mouse chooses door 3, we know he will escape in 1 minute. On the other hand, for doors 1 and 2 the mouse will wander through the maze and then find himself back where he started. We represent this situation by writing, for example, $E[M|D=1] = 2 + E[M]$. The expected number of minutes it takes for the mouse to escape the maze is the two minutes he spends getting back to his starting point plus his expected value of starting over. (We assume the mouse doesn't learn anything from taking the wrong doors.) By conditioning on D, we find,

$$\begin{aligned}
E[M] =& \sum_{d=1}^{3} E[M|D=d]\Pr\{D=d\} \\
=& E[M|D=1]\cdot .5 + E[M|D=2]\cdot .3 + E[M|D=3]\cdot .2 \\
=& (2+E[M])\cdot .5 + (4+E[M])\cdot .3 + 1\cdot .2,
\end{aligned}$$

which is an algebraic equation that can be solved for $E[M] = 12$. △

Example 4.8. Compute the expected number of roles of a pair of fair dice until a pair of sixes appears.

Problems like this are so easy to solve by conditioning, it almost seems like we're cheating. Let N be the number of roles required to get a pair of sixes, and let E be the event that two sixes appear on the first roll. (The complement of E, denoted E^c represents the event that at least one of the dice is not a six on the first roll.) We compute

$$\begin{aligned}
E[N] =& E[N|E]\Pr\{E\} + E[N|E^c]\Pr\{E^c\} \\
=& 1\cdot\frac{1}{36} + (1+E[N])\cdot\frac{35}{36},
\end{aligned}$$

which is an algebraic equation that can be solved for $E[N] = 36$. △

4.8 Variance and Covariance

Consider the following three random variables,

$$W = 0,\ \text{prob } 1; \qquad Y = \begin{cases} -1, & \text{prob } \frac{1}{2} \\ +1, & \text{prob } \frac{1}{2} \end{cases}; \qquad Z = \begin{cases} -100, & \text{prob } \frac{1}{2} \\ +100, & \text{prob } \frac{1}{2} \end{cases}.$$

We see immediately that though these three random variables are very different, the expected value of each is the same, $E[W] = E[Y] = E[Z] = 0$. The problem is that the expected value of a random variable does not provide any information about how far the values the random variable takes on can deviate from one another. We measure this with *variance*, defined as

$$\mathrm{Var}[X] = E[(X - E[X])^2].$$

That is, we study the squared difference between realizations of the random variable and the mean of the random variable. Computing directly from this definition, we find the variance of each random variable above,

$$\begin{aligned}
\mathrm{Var}[W] =& (0-0)^2\cdot 1 = 0, \\
\mathrm{Var}[Y] =& (-1-0)^2\cdot\frac{1}{2} + (1-0)^2\cdot\frac{1}{2} = 1, \\
\mathrm{Var}[Z] =& (-100-0)^2\cdot\frac{1}{2} + (100-0)^2\cdot\frac{1}{2} = 100^2.
\end{aligned}$$

Typically, a more intuitive measure of such deviation is the square root of variance, which we refer to as the *standard deviation,* and typically denote σ. We notice that by expanding the square in our definition of var

Example 4.9. (Computing variance by conditioning) In this example, we employ a conditioning argument to determine the variance on the number of roles required in Example 3.6 for a pair of sixes to emerge.

Letting N and E be as in example 3.6, we have

$$\mathrm{Var}[N] = E[N^2] - E[N]^2,$$

for which we compute

$$\begin{aligned} E[N^2] &= E[N^2|E]\Pr\{E\} + E[N^2|E^c]\Pr\{E^c\} \\ &= 1 \cdot \frac{1}{36} + E[(1+N)^2] \cdot \frac{35}{36} = \frac{1}{36} + E[1 + 2N + N^2]\frac{35}{36} \\ &= 1 + 2\frac{35}{36}E[N] + \frac{35}{36}E[N^2]. \end{aligned}$$

We already know $E[N]$ from Example 3.7, so we can now solve for

$$E[N^2] = 71 \cdot 36.$$

We have, finally,

$$\mathrm{Var}[N] = 36 \cdot 35.$$

△

Example 4.10. Show that the variance of a binomial random variable is $\mathrm{Var}[X] = np(1-p)$.

Recall from Section 4.2 that a random variable X that takes values $0, 1, 2, ..., n$ is said to be a *binomial* random variable with sample size n and probability p if its probability mass function is given by

$$p(k) = \binom{n}{k} p^k(1-p)^{n-k}.$$

We showed in Example 4.2 that $E[X] = np$, so we need only compute $E[X^2]$. We have

$$\begin{aligned} E[X^2] &= \sum_{k=0}^{n} k^2 \binom{n}{k} p^k(1-p)^{n-k} = \sum_{k=1}^{n} k^2 \frac{n!}{(n-k)!k!} p^k(1-p)^{n-k} \\ &= \sum_{k=1}^{\infty} k \frac{n!}{(n-k)!(k-1)!} p^k(1-p)^{n-k}, \end{aligned}$$

where we have observed that the $k = 0$ term gives no contribution. Setting now $l = k - 1$ we find

$$E[X^2] = \sum_{l=0}^{n-1} (l+1)np\binom{n-1}{l} p^l(1-p)^{(n-1)-l} = npE[Y+1],$$

where Y denotes a binomial random variable that takes values 0, 1, 2, ..., $n-1$. I.e., $E[Y+1] = E[Y] + E[1] = (n-1)p + 1$. We have, then,

$$E[X^2] = np(np - p + 1),$$

so that

$$\text{Var}[X] = E[X^2] - E[X]^2 = (np)^2 + np(1-p) - (np)^2 = np(1-p),$$

which is the claim. △

We can generalize the idea of variance to two random variables X and Y. We define the *covariance* of X and Y as

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E(Y))] = E[XY] - E[X]E[Y].$$

5 Game Theory

Despite its misleading connotations the expression game theory has become standard terminology for the systematic study of situations of conflict and cooperation. We regard a game as a situation with the following four properties[8]:

1. There are at least two players. A player may be an individual, but it may also be a more general entity such as a company, a nation, or a biological species.
2. Each player has a number of possible strategies: courses of actions that he or she may choose to follow.
3. The strategies chosen by each player determine the outcome of the game.
4. Associated to each possible outcome of the game is a collection of numerical payoffs, one to each player. These payoffs represent the value of the outcome to different players.

5.1 Zero-Sum Games

Suppose two players, Rose and Colin, play a game in which Rose has three possible strategies and Colin has two possible strategies. We often depict such a game in a *payoff table*, where the values in parentheses can be read as

(payoff to Rose, payoff to Colin).

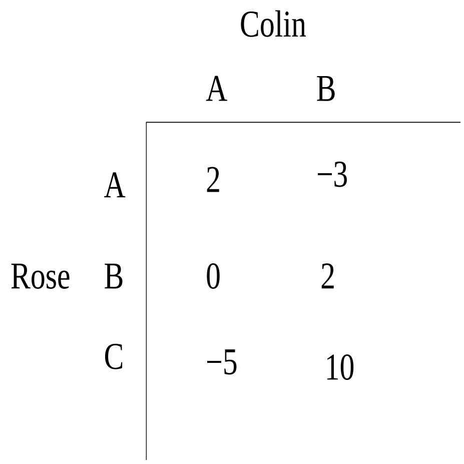
Example 5.1. In the game depicted in Table 5.1 if Rose plays strategy C and Colin plays strategy A then the payoff is (-5,5) so Rose loses 5 and Colin gains 5. We refer to a game such as this in which the values of each outcome sum to zero as a zero-sum game.

Since Colin's payoff is entirely determined by Rose's in such games we often record only Rose's payoffs. The payoff table Table 5.1 often appears as follows.

[8] Taken from [S]

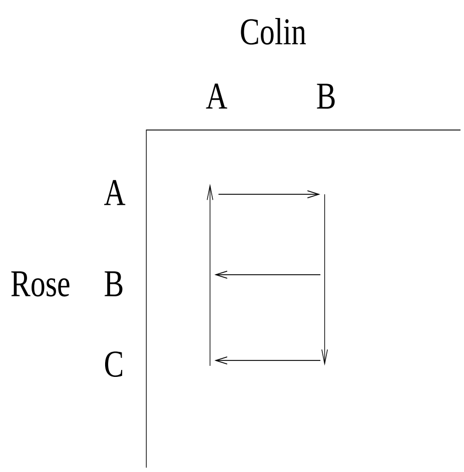
Rose\Colin	A	B
A	(2,-2)	(-3,3)
B	(0,0)	(2,-2)
C	(-5,5)	(10,-10)

Table 5.1: Payoff table



Definition 5.1. *We refer to a two-person zero-sum game as a* matrix game.

We will assume Rose and Colin are both *rational* players in the sense that they make reasonable decisions. For example, Rose would like to win 10, so she might start by playing strategy C, but as soon as Colin realizes that Rose is consistently playing strategy C he will begin playing strategy A. Of course, when Rose realizes that Colin is consistently playing strategy A she will begin playing strategy A. We can graphically depict the constantly changing choice of strategies with an arrow diagram, drawn as follows: in each column we draw arrows from each entry to the largest entry in the column (this indicates what Rose's inclination will be), and in each row we we draw an arrow from each entry to the smallest entry in the row (indicating what Colin's inclination will be). For example, the arrow diagram corresponding with Example 5.1 would be as follows.



For example, the arrow down the second column indicated Rose's inclination for Strategy C if Colin is playing strategy B, while the arrow across the third row indicates Colin's inclination for strategy A if Rose is playing strategy C. We notice in particular that the path has no stopping point, which indicates that there is no single strategy pair that can be viewed as optimal for both Colin and Rose. We will soon return to this example, but first let's consider a case in which a stopping point exists. We refer to such a point as a saddle point.

5.1.1 Dominate Strategies

The following discussion will be clarified if we have an example to refer to, so we give a convenient game in Example 5.2.

Example 5.2. Consider the following zero-sum game.

		Colin			
		A	B	C	D
Rose	A	12	−1	1	0
	B	5	1	7	−20
	C	3	2	4	3
	D	−16	0	0	16

Definition 5.2. *A strategy S is said to dominate a strategy T if every outcome in S is at least as good as the corresponding outcome in T, and at least one outcome in S is strictly better than the corresponding outcome in T.*

In Example 5.2. Colin strategy B clearly dominates Colin strategy C in each component. That is, -1 is better for Colin than 1, 1 is better than 7, 2 is better than 4, and 0 is no worse than 0.

Dominance Principle. *A rational player, in contest with a rational player, should never play a dominated strategy.*

While this principle is straightforward, it's worth investigating just a little further. Cleverly, Colin might consider playing strategy C in the hope of inducing Rose to maximize her profits with strategy B. Then Colin will switch to strategy D and make a killing. This is where we assume Rose is also a rational player, and as such she recognizes her strategy against Colin C should be Rose C, which over time should force Colin to Colin B. According to the Dominance Principle we are justified in eliminating Colin strategy C from the table (it will never be used), and we obtain the reduced problem given below.

		Colin		
		A	B	D
	A	12	−1	0
Rose	B	5	1	−20
	C	3	2	3
	D	−16	0	16

We will return to this example in the next subsection, but first we consider a game that can be solved entirely by dominance.

Example 5.3. Rose and Colin each put a \$1.00 ante in a pot, and then each is dealt a single card from a limitless deck consisting only of kings and aces. After looking at his card Colin must decide either to bet \$2.00 or to fold. If Colin bets then Rose must decide whether to call or fold. (We don't allow a raise in this game.) If Rose calls there is a showdown of cards in which ace beats king, and in the event of a tie no money exchanges hands. Determine optimal strategies for both Rose and Colin.

First, we want to cast this as a matrix problem. Collin's possible strategies are

Colin A: Bet only with the ace
Colin B: Always bet

Notice that Colin's strategy B allows for the possibility of bluffing in the sense that a king can't possibly be the best hand. Likewise, Rose's possible strategies are

Rose A: Call only with an ace
Rose B: Always call

We now construct a matrix game by considering Rose's expected payoff for each of the four strategy combinations.[9] For this game there are four possible equally likely events from the deal,

$$\begin{aligned}
A_1 &= \text{event Rose gets an ace and Colin gets an ace}\\
A_2 &= \text{event Rose gets an ace and Colin gets a king}\\
A_3 &= \text{event Rose gets a king and Colin gets an ace}\\
A_4 &= \text{event Rose gets a king and Colin gets a king}
\end{aligned}$$

Using the fact that $\Pr\{A_j\} = \frac{1}{4}$ for all $j = 1, 2, 3, 4$, we can compute Rose's expected values as follows[10]:

[9]Later, we will work with the expected value associated with each of Rose's strategies, and it's worth stressing here that that will be a fundamentally different thing.

[10]In this expressions we continue to write out $\Pr\{A_j\}$ to clarify the case.

Rose A, Colin A.

$$E[R] = 0 \cdot \Pr\{A_1\} + 1 \cdot \Pr\{A_2\} - 1 \cdot \Pr\{A_3\} + 1 \cdot \Pr\{A_4\} = \frac{1}{4}.$$

Rose A, Colin B.

$$E[R] = 0 \cdot \Pr\{A_1\} + 3 \cdot \Pr\{A_2\} - 1 \cdot \Pr\{A_3\} - 1 \cdot \Pr\{A_4\} = \frac{1}{4}.$$

Rose B, Colin A.

$$E[R] = 0 \cdot \Pr\{A_1\} + 1 \cdot \Pr\{A_2\} - 3 \cdot \Pr\{A_3\} + 1 \cdot \Pr\{A_4\} = -\frac{1}{4}.$$

Rose B, Colin B.

$$E[R] = 0 \cdot \Pr\{A_1\} + 3 \cdot \Pr\{A_2\} - 3 \cdot \Pr\{A_3\} + 0 \cdot \Pr\{A_4\} = 0.$$

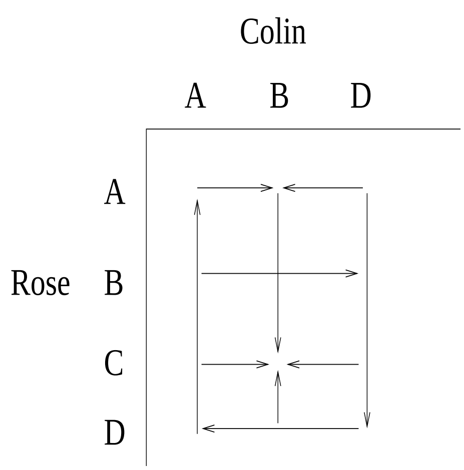
The matrix game associated with these values is given below.

		Colin	
		A	B
Rose	A	1/4	1/4
	B	−1/4	0

Rose's strategy A clearly dominates her strategy B, so she should play only Strategy A, and this will ensure that she gets $\frac{1}{4}$ per game. △

5.1.2 Saddle Points

Consider now the arrow diagram for the reduced game from Example 5.2:



Clearly, this diagram places emphasis on the joint strategy (C,B). (In joint strategies, the row strategy will be given first, the column strategy second.) If Rose knows Colin will play B then she will certainly play C, and likewise if Colin knows Rose will play C then he will certainly play B. In this way (C,B) is regarded as an equilibrium strategy, referred to as a *saddle point* in game theory.

Definition 5.3. *An outcome in a matrix game (with payoffs recorded to the row player) is called a saddle point if the entry at the outcome is both its row minimum and its column maximum.*

Of course a matrix game can have multiple saddle points.

Example 5.4. Consider the following matrix game.

Colin

		A	B	C	D
	A	1	5	1	2
Rose	B	0	−2	−5	10
	C	−5	16	0	−10
	D	1	13	1	2

First, we see that our algorithm for drawing an arrow diagram gets thrown off because two rows have a repeated minimum and two columns have a repeated maximum.[11] On the other hand, it's clear that all four ones (for joint strategies (A,A), (A,C), (D,A), and (D,C)) are saddle points.

Definition 5.4. *In a matrix game if there is a value v so that Rose has a strategy that guarantees that she will win at least v, and Colin has a strategy that guarantees Rose will win no more than v then v is called the value of the game.*

Theorem 5.5. *We have the following:*
(i) Any two saddle points in a matrix game have the same value.
(ii) If a matrix game has at least one saddle point its value is the value of the game.

Proof. For (i) we first observe that if two saddle points are in the same row or the same column they must have identical values, because two different values can neither both be

[11]Of course we could still draw an arrow diagram consisting of arrows from all row entries that are not minimal to all that are and from all column entries that are not maximal to all that are, but we won't pursue this.

a row minimizer or a column maximizer. Suppose, then, that a and d are any two saddle values that are not in either the same row or the same column:

$$\begin{matrix} a & \cdots & b \\ \vdots & \vdots & \vdots \\ c & \cdots & d \end{matrix}$$

We know by definition of a saddle point that we have

$$\begin{aligned} c \le a \le & b \\ b \le d \le & c, \end{aligned}$$

which gives

$$c \le a \le b \le d \le c.$$

This clearly asserts $a = b = c = d$.

Finally, we observe that (ii) is simply a restatement of the definition of a saddle point. Since it is the minimum value in a row Rose is guaranteed to achieve it by playing that row, and since it is the maximum value in its column Colin is guaranteed that Rose cannot do better by switching strategies. □

Saddle Point Principle. *If a matrix game between two rational players has at least one saddle point both players should play a strategy that contains a saddle point.*

The reasoning is clear: for example, if Rose plays anything other than a saddle, Colin can ensure that she will not do as well as she would if she played a saddle.

5.1.3 Minimax Method

While the graphical (arrow diagram) method for locating saddle points can work in principle on any matrix game with saddle points it's convenient to have an analytic approach that, for example, could easily be implemented numerically for large systems. For this, we simply think through what the arrow diagram does. First, the row arrows identify the minimum value(s) across each row. Next, the column arrows identify the maximum values across each column. If an entry is *both* its row minimizer and its column maximizer it is a saddle point. This means that in locating saddle points, we can proceed as follows: (1) Find the minimum across each row and compute the maximin of these minimizers (i.e., the maximin), and (2) Find the maximum down each column and compute the minimum of these maximizers (i.e., the minimax). A value is a saddle point if and only if the maximin and the minimax are the same value. That is, if v is a maximin then it is certainly a row minimizer, and if it is a minimax it is certainly a column maximizer. If it is both a maximin and a minimax then it is both a row minimizer and a column maximizer, and that, be definition, is a saddle point. (In fact, what this says is that any value that is in both the list of row minimizers and the list of column maximizers is a saddle, but since the smallest of the maximizers will necessarily be greater than or equal to the largest of the minimizers, the two sets of numbers can only agree if the maximin is the minimax.)

As an example, let's carry out the minimax method for the reduced version of the game in Example 5.2.

Colin

		A	B	D	Row mins	
	A	12	−1	0	−1	
Rose	B	5	1	−20	−20	
	C	3	2	3	2	maximin
	D	−16	0	16	−16	
Column maxes		12	2	16		
			minimax			

The saddle point value is clearly 2, with corresponding joint strategy (C,B).

5.1.4 Mixed Stategies

We have seen that in games in which there is at least one saddle point each player should play a strategy that contains a saddle point, and this process will select a particular joint strategy (or perhaps a collection of joint strategies with equivalent values for both players). In other words, we now completely understand two-person zero-sum games in which there is a saddle point. Next, we turn to the case of games in which there is no saddle point.

Example 5.5. Consider the matrix game given below.

Colin

		A	B
Rose	A	2	−3
	B	0	3

In this case it's easy to see from an arrow diagram that there is no saddle point, but to clarify what happens with the minimax method in the absence of a saddle point, we illustrate the method below.

		Colin A	Colin B	Row mins	
Rose	A	2	−3	−3	
	B	0	3	0	← maximin
Column maxes		2	3		
		minimax			

Since the maximin does not agree with the minimax there cannot be a saddle point.

The absence of a saddle point means there is no single strategy that is optimal for both Rose and Colin, and so we must begin to think about how Rose and Colin should rationally proceed in such a situation. We begin by observing that if Colin consistently plays strategy A then Rose can profitably play strategy A, while if Colin consistently plays strategy B then Rose can profitably play strategy B. Even if Colin switches from one strategy to the other, Rose can play profitably so long as she can intelligently guess which strategy Colin will play next. In particular, notice that Rose does not have to guess correctly every time: she can play profitably so long as she guesses Colin correctly more than half the time. This suggests two things: (1) Colin should play a mixed strategy that includes both A and B, and (2) While Colin might play strategy A a different percentage of the time than he plays strategy B, he should make the choice probabilistically so that it's impossible (for all practical purposes) for Rose to guess anything about what he will play next.

In order to see how this works, let's suppose Colin plays strategy A with probability x and strategy B with probability $1-x$ (i.e., he always plays one or the other). In this case Rose's outcome from each of her strategies becomes a random variable, depending on what Colin plays. Let R_A denote Rose's outcome while playing strategy A and R_B denote Rose's outcome while playing strategy B. Then we have

$$\begin{aligned} E[R_A] &= 2x - 3(1-x) = 5x - 3 \\ E[R_B] &= 3(1-x) = -3x + 3. \end{aligned}$$

If one of these is larger than the other then Rose can profitably proceed by choosing the strategy with the larger expected outcome, so Colin should proceed by choosing x in such a way that these two expected values give the same outcome. To visualize this, we refer to the MATLAB plot in Figure 5.1 of the lines $y = 5x - 3$ and $y = -3x + 3$.

In principle, Rose can always choose the strategy that gives her the expected value associated with the upper line. For example, if Colin chooses $x = .4$, the dashed line ($y = -3x + 3$) is above the solid line, so Rose would choose strategy B. In light of this, we see that Colin's optimal choice will be the value of x where the two lines intersect. That is, if he chooses any other value for x then Rose can certainly play a strategy that will give her a higher outcome than the intersection value. We find x by solving

$$5x - 3 = -3x + 3 \Rightarrow x = \frac{3}{4}.$$

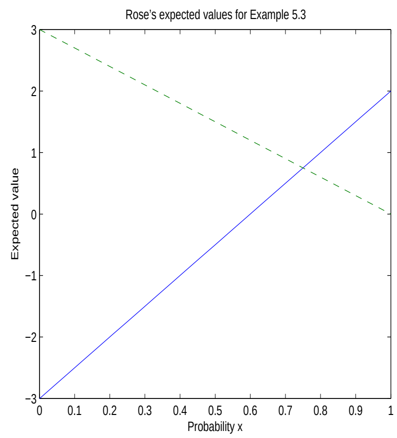


Figure 5.1: Plot of Rose's expected values for Example 5.3.

We conclude that Colin should play strategy A $\frac{3}{4}$ of the time and strategy B $\frac{1}{4}$ of the time. For example, whenever it is Colin's time to make a decision he might glance at his watch. If the seconds hand is in the first quarter of the face he can play strategy B, while if it is in the latter three quarters of the face he can play strategy A. As long as Rose can't see his watch it doesn't even matter if she knows what he's doing, or even if she knows what his value of x is. If Colin plays this strategy then Rose's expectation will be the same for either of her strategies,

$$5(\frac{3}{4}) - 3 = \frac{3}{4},$$

where we emphasize that the equivalence here of Colin's probability x and Rose's expected value is coincidental.

Now of course Rose wants to think about the game in exactly the same way, so we let y denote the probability she plays strategy A and $1 - y$ the probability she plays strategy B. In this case Colin's expectations when playing strategies A and B are

$$\begin{aligned} E[C_A] &= -2y \\ E[C_B] &= 3y - 3(1-y) = 6y - 3. \end{aligned}$$

(Many authors change the sign here (i.e., use the same signs as in the matrix), so that these expectations actually correspond with Rose's payoffs.) Arguing precisely as we did for Colin, we reason that Rose should equate these to expressions,

$$-2y = 6y - 3 \Rightarrow y = \frac{3}{8}.$$

We conclude that Rose should play strategy A $\frac{3}{8}$ of the time and strategy B $\frac{5}{8}$ of the time. Colin's outcome will be

$$-2(\frac{3}{8}) = -\frac{3}{4},$$

as we knew it should be for a zero-sum game. Notice that in our terminology $\frac{3}{4}$ is the value of this game. $\triangle$

5.1.5 The Method of Oddments

It's possible to analyze 2×2 matrix games in general. Consider the general case

		Colin	
		A	B
Rose	A	a	b
	B	c	d

For this calculation we will assume that the game *does not* have a saddle, and we stress that this needs to be checked before the method is applied. First, suppose the two largest entries are those in the first row, a and b. Then clearly the column arrows in an arrow diagram will both point upward and the minimum of a and b will be a saddle. Likewise, if the two largest values occur in any row or column there will be a saddle. Suppose, then, that the largest values occur on diagonals, either a and d or b and c. If we let x denote the probability Colin plays strategy A so that $1-x$ denotes the probability he plays strategy B, Rose's general expected values become

$$
\begin{aligned}
E[R_A] &= ax + b(1-x) = (a-b)x + b \\
E[R_B] &= cx + d(1-x) = (c-d)x + d.
\end{aligned}
$$

Upon setting $E[R_A] = E[R_B]$ we find

$$
x = \frac{(d-b)}{(a-b)-(c-d)} = \frac{(d-b)}{(a-c)+(d-b)}.
$$

Here, since we assume the largest values are diagonal from one another, we know that $a-c$ and $d-b$ have the same sign, and so each term in parentheses on the far right of our expression for x is positive. I.e.,

$$
x = \frac{|b-d|}{|a-c|+|b-d|}.
$$

This says we can proceed as follows to compute Colin's probabilities: we compute the absolute values of the differences in the columns, $|a-c|$ and $|b-d|$, exchange them (after the exchange these values are referred to as Colin's *oddments*), and write

$$
\begin{aligned}
\Pr\{\text{Colin plays A}\} &= \frac{|b-d|}{|a-c|+|b-d|} \\
\Pr\{\text{Colin plays B}\} &= \frac{|a-c|}{|a-c|+|b-d|}.
\end{aligned}
$$

Proceeding similarly for Rose, we obtain a general method, described schematically below for Example 5.3.

		Colin A	Colin B	differences	oddments	Rose probabilities
Rose	A	2	−3	2−(−3)=5	3	$\frac{3}{8}$
	B	0	3	0−3=−3	5	$\frac{5}{8}$
differences		2−0=2	−3−3=−6			
oddments		6	2			
Colin probabilities		$\frac{6}{8}$	$\frac{2}{8}$			

5.1.6 The Minimax Theorem

Before proceeding with a few more examples, we give one of the first important results developed in the study of game theory, established by John von Neumann in 1928.

Minimax Theorem. *Every $m \times n$ matrix game has a solution in the following sense: there is a unique game value v and optimal strategies (pure or mixed) for Rose and Colin (i.e., for the two players in the game) so that the following hold.*

(i) If Rose plays her optimal strategy her expected payoff will be greater than or equal to v no matter what Colin does.

(ii) If Colin plays his optimal strategy Rose's expected payoff will be less than or equal to v no matter what Rose does.

(iii) The solution can always be found as the solution to some square (i.e. $k \times k$ for some positive integer $k \leq \max(m, n)$) subgame of the original game.

(iv) The roles of Rose and Colin can be reversed, and the same statements obtained.

Example 5.6. Consider again the game described in Example 5.1,

		Colin	
		A	B
	A	2	−3
Rose	B	0	2
	C	−5	10

The Minimax Theorem asserts that there exists a solution (game value plus optimal strategies) for this game, and that it is the solution for either a 1×1 or 2×2 submatrix. We have already seen that there are no saddle points for this example, and so the solution cannot be that of a 1×1 matrix (which clearly must correspond with a saddle point). We see, then, from the Minimax Theorem that the solution will be a mixed strategy that solves one of the three possible 2×2 matrix subgames. In order to understand which subgame we should solve, we let x denote the probability Colin plays strategy A so that $1 - x$ denotes the strategy he plays strategy B, and we compute Rose's expected values

$$\begin{aligned}
E[R_A] &= 2x - 3(1-x) = 5x - 3\\
E[R_B] &= 2(1-x) = -2x + 2\\
E[R_C] &= -5x + 10(1-x) = -15x + 10.
\end{aligned}$$

Next, we plot these three lines together in Figure 5.2.

For each value x that Colin chooses, a vertical line through x will cross all three of Rose's expected value lines. Rose can obtain the expected value associated with the largest y-value at which these lines intersect. So Colin must choose the value x at the lowest intersection so that no lines are above the intersection. We refer to the curve created by taking the top line at each point us the upper *envelope* for these lines. In this terminology, Colin should choose x to correspond with the minimum value on the upper envelope. This point is indicated on the graph for Example 5.4 with a dark dot. We find that the opimal strategy will be the mixed solution for Rose's strategies A and B. Graphically, the value of x has already been located, but to be precise we will solve the associated matrix game

		Colin	
		A	B
Rose	A	2	−3
	B	0	2

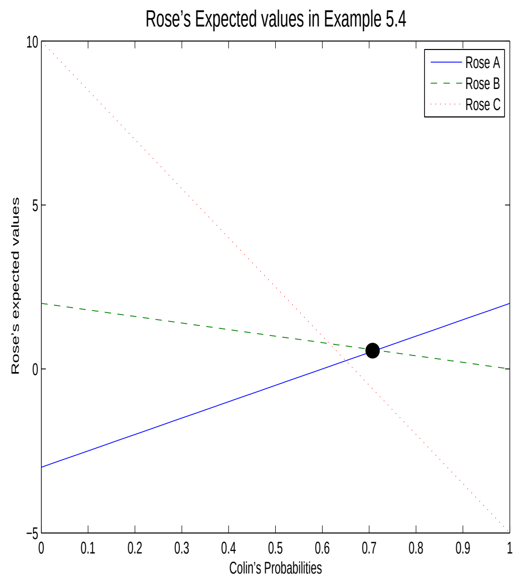


Figure 5.2: Plot for Example 5.4.

This game is close, but not identical to the game considered in Example 5.4. We solve it by the method of oddments. Rose's oddments are respectively 2 and 5, so

$$\Pr\{\text{Rose plays A}\} = \frac{2}{7}$$
$$\Pr\{\text{Rose plays B}\} = \frac{5}{7},$$

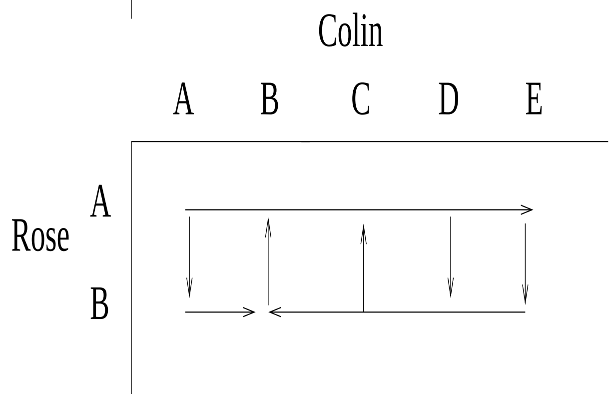
and Colin's oddments are respectively 5 and 2 so

$$\Pr\{\text{Colin plays A}\} = \frac{5}{7}$$
$$\Pr\{\text{Colin plays B}\} = \frac{2}{7}.$$

$\triangle$

Example 5.7. Consider the following game, for which we see from an arrow diagram that there are no saddle points.

		Colin				
		A	B	C	D	E
Rose	A	−2	5	1	0	−4
	B	3	−3	−1	3	8



If we let y denote the probability that Rose plays strategy A so that $1-y$ denotes the probability Rose plays strategy B, then Colin's expected values are

$$\begin{aligned}
E[C_A] &= 2y - 3(1-y) = 5y - 3 \\
E[C_B] &= -5y + 3(1-y) = -8y + 3 \\
E[C_C] &= -y + (1-y) = -2y + 1 \\
E[C_D] &= -3(1-y) = 3y - 3 \\
E[C_E] &= 4y - 8(1-y) = 12y - 8.
\end{aligned}$$

We depict this graphically in Figure 5.3.

In this case Rose's choice of y will correspond with a horizontal line through that value, and Colin will be able to bank an expected value corresponding with the largest value (measured along the horizontal axis) at which one of his expected value lines intersects Rose's horizontal line. In this way, Rose should choose y as the value that minimizes the right most envelope of Colin's expected value lines. This point is indicated with the comically small dot in Figure 5.3. We see that the optimal strategy is obtained as a solution to the 2×2 subgame corresponding the Colin's strategies A and C. Accordingly, we must solve

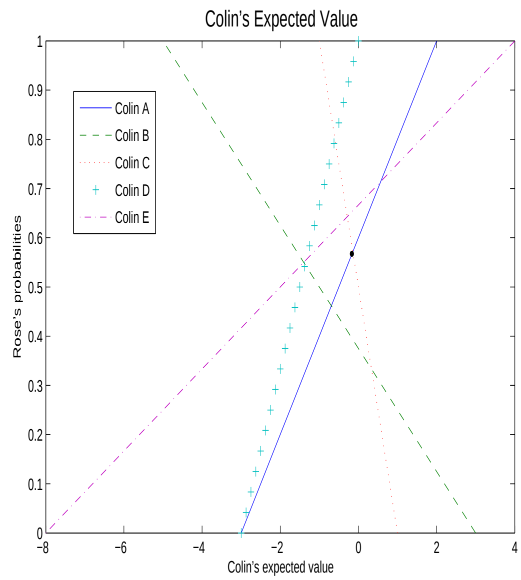


Figure 5.3: Figure for Example 5.5.

		Colin A	Colin C
Rose	A	−2	1
	B	3	−1

Using the method of oddments, we immediately find

$$\Pr\{\text{Rose plays A}\} = \frac{4}{7}$$
$$\Pr\{\text{Rose plays B}\} = \frac{3}{7},$$

and

$$\Pr\{\text{Colin plays A}\} = \frac{2}{7}$$
$$\Pr\{\text{Colin plays C}\} = \frac{5}{7}.$$

$\triangle$

6 Continuous Random Variables

For some random variables, the collection of possible realizations can be regarded as continuous. For example, the time between scores in a soccer match or the price of a stock at a given time are such random variables.

6.1 Cumulative Distribution Functions

As with discrete random variables, the cumulative distribution function, $F(x)$, for a continuous random variable X is defined by $F(x) = \Pr\{X \leq x\}$.

Example 4.1. Suppose U is a random variable that takes real values between 0 and 1, and that we have no reason to believe that the probability of U taking any one value is different from the probability that it will take any other. Write down an expression for the cumulative distribution function of U.

Since U is equally likely to take on any value in the interval $[0,1]$, we observe that it has the same likelihood of being above 1/2 as being below. That is,

$$F(\frac{1}{2}) = \Pr\{U \leq 1/2\} = 1/2.$$

Similarly,

$$F(\frac{1}{3}) = \Pr\{U \leq 1/3\} = 1/3$$

and so on,

$$F(\frac{1}{n}) = \Pr\{U \leq 1/n\} = 1/n. \tag{6.1}$$

In general, we have the relationship that if F is the CDF associated with U, then

$$F(x) = \begin{cases} 0, & x \leq 0 \\ x, & 0 \leq x \leq 1 \\ 1, & x \geq 1. \end{cases}$$

Finally, observe that if we take the limit in (6.1) as $n \to \infty$, we get

$$\Pr\{U = 0\} = 0.$$

Arguing similarly for any point $u \in [0,1]$, we have

$$\Pr\{U = u\} = 0.$$

This odd relationship arises because each outcome of U is equally likely, but U has an infinite number of possible outcomes. If each of these is anything besides 0, the probability of the entire sample space would be infinite. On the other hand, the probability that U lies on any interval $[a,b] \subset [0,1]$ is clearly $b-a$. In this way, we will be interested in computing the probability that a uniform random variable lies on a certain interval rather than the probability that it takes on a certain value. One reasonable way to think about this is to consider a soccer match assumed to last for exactly one hour, and to let U be the time of the first score in hours (so U is between 0 and 1, and roughly uniform). Suppose now that we are asked to predict a value for U, and that in order for our value to be considered correct, it must be accurate to every decimal place. That is, if we choose the value $u = .5$, we really mean

$$u = .50000000000\ldots.$$

In this way, if the score occurs at the time

$$u = .5000000000000000000001,$$

we are still incorrect. Assuming time can be measured as accurately as we like, our odds of choosing the single correct time are 0. △

6.2 Probability Density Functions

A useful tool in the study of continuous random variables is the *probability density function*, defined in terms of the CDF as

$$f(x) = F'(x),$$

wherever $F(x)$ is differentiable.

Example 4.1 cont. For the random variable U from Example 4.1, we can compute $f(x)$ at all points except $x = 0$ and $x = 1$ at which $F(x)$ is not differentiable. We find,

$$f(x) = \begin{cases} 0, & x < 0 \\ 1, & 0 < x < 1 \\ 0, & x > 1, \end{cases}$$

where we observe that since $f(x)$ is discontinuous at the points $x = 0$ and $x = 1$, we do not define it there. △

According to the Fundamental Theorem of Calculus, we also have the integral relationship,

$$F(x) = \int_{-\infty}^{x} f(y)dy.$$

6.3 Properties of Probability density functions

Let $f(x)$ be the probability density function associated with random variable X. Then the following hold:

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

2. $\Pr\{a \leq X \leq b\} = \int_a^b f(x)dx$.

3. (Generalization of (2)) For any set of real number I, $\Pr\{X \in I\} = \int_I f(x)dx$.

4. $E[X] = \int_{-\infty}^{+\infty} xf(x)dx$.

5. (Generalization of (4)) For any continuous function $g(x)$, $E[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)dx$.

6.4 Identifying Probability Density Functions

A critical issue in the study of probability and statistics regards determining the probability density function for a given random variable. We typically begin this determination through consideration of a histogram, simply a bar graph indicating the number of realizations that fall into each of a predetermined set of intervals.

Example 4.2. Suppose we are hired by Lights, Inc. to study the lifetime of lightbulbs (a continuous random variable). We watch 100 bulbs and record times to failure, organizing them into the following convenient ranges (in hours):

Time range	0–400	400–500	500–600	600-700	700–800	800–900
# Failed	0	2	3	5	10	10

900–1000	1000–1100	1100–1200	1200–1300	1300–1400	1400–1500	1500–1600
20	20	10	10	5	3	2

Table 6.1: Data for Lights, Inc. example.

This data is recorded in the MATLAB M-file *lights.m.*

```
%LIGHTS: Script file that defines times T at which
%lightbulbs failed.
T=[401 402 501 502 503 601 602 603 604 605 701 702 703 704 705 706 707
708 709 710 ...
801 802 803 804 805 806 807 808 809 810 901 902 903 904 905 906 907 908
909 910 ...
911 912 913 914 915 916 917 918 919 920 1001 1002 1003 1004 1005 1006 1007
1008 1009 1010 ...
1011 1012 1013 1014 1015 1016 1017 1018 1019 1020 1101 1102 1103 1104
1105 ...
1106 1107 1108 1109 1110 1201 1202 1203 1204 1205 1206 1207 1208 1209
1210...
1301 1302 1303 1304 1305 1401 1402 1403 1501 1502];
```

Of course, we could analyze the times more carefully in intervals of 50 hours or 10 hours etc., but for the purposes of this example, intervals of 100 hours will suffice. Define now the function

$$f(x) = \begin{cases} 0, & 0 \le x \le 400 \\ .0002, & 400 \le x \le 500 \\ .0003, & 500 \le x \le 600 \\ .0005, & 600 \le x \le 700 \\ \vdots & \vdots \\ .0002, & 1500 \le x \le 1600 \end{cases},$$

Let T represent the time to failure of the lightbulbs. Then we can compute the probability that a lightbulb will fail in some interval $[a, b]$ by integrating $f(x)$ over that interval. For example,

$$\Pr\{400 \le T \le 500\} = \int_{400}^{500} f(x)dx = \int_{400}^{500} .0002dx = .02.$$
$$\Pr\{600 \le T \le 800\} = \int_{600}^{800} f(x)dx = \int_{600}^{700} .0005dx + \int_{700}^{800} .001dx = .15.$$

Recalling our properties of the probability density function, we see that $f(x)$ is an approximation to the PDF for the random variable T. (It's not precise, because it is only precisely accurate on these particular intervals, and a PDF should be accurate on all intervals.)

The function $f(x)$ is a histogram for our data, scaled by a value of 10,000 to turn numerical counts into probabilities (more on this scale in a minute). If the vector T contains all the failure times for these lights, then the MATLAB command *hist(T,12)* creates a histogram with twelve *bins* (bars) (see Figure 6.1). Notice that this histogram is precisely a scaled version of

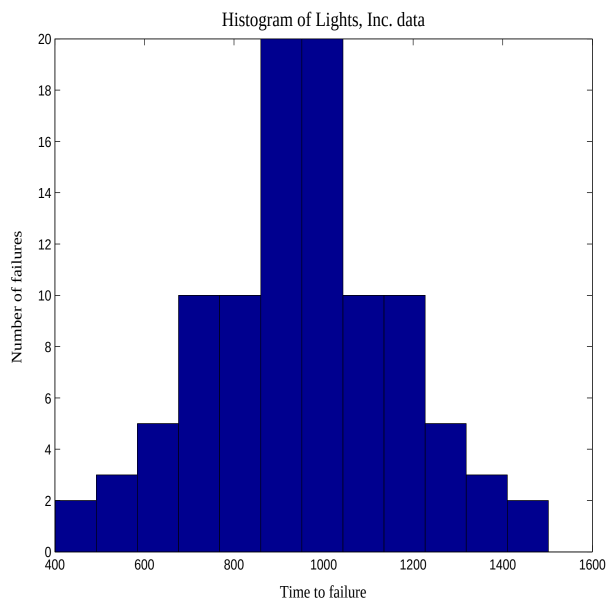


Figure 6.1: Histogram of data from Example 4.2.

$f(x)$. As for the scaling, we choose it so that $\int_{-\infty}^{+\infty} f(x)dx = 1$, which can be accomplished by dividing the height of each bar in the histogram by binwidth $\times$ Total number of data points. Here, we divide by $100 \times 100 = 10000$, giving $f(x)$. In general, it can be difficult to determine the exact binwidths MATLAB chooses. $\triangle$

6.5 Useful Probability Density Functions

We typically proceed by looking at a histogram of our data, and trying to match it to the form of a smooth probability density function, and preferably one that is easy to work with. (Though as computing power becomes better and better, researchers are becoming less concerned with ease of computation.) In these notes, we will focus on the following distributions.

1. Gaussian
2. Uniform
3. Exponential
4. Weibull
5. Beta
6. Gamma
7. Mixture

1. Gaussian distribution. One of the most common probability density functions is the *Gaussian* distribution—also known as the normal distribution, or somewhat infamously as the bell curve. The Gaussian probability density function for a random variable X with mean μ ($E[X] = \mu$) and standard deviation σ ($\sigma^2 = \mathrm{Var}[X]$), takes the form

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Example 4.2 cont. For the case of our Lights, Inc. data we compute $f(x)$ with the following MATLAB script.

```
>>mu=mean(T)
mu =
956.8800
>>sd=std(T)
sd =
234.6864
>>x=linspace(400,1600,100);
>>f=1/(sqrt(2*pi)*sd)*exp(-(x-mu).^2/(2*sd^2));
>>plot(x,f,'–')
```

In order to compare our fit with our data, we will have to scale the Gaussian distribution so that it is roughly the same size as the histogram. More precisely, we know that the Gaussian distribution must integrate to 1 while an integral over our histogram is given by a sum of the areas of its rectangles. In the event that these rectangles all have the same width, which is the generic case for histograms, this area is

$$\text{Total area} = [\text{width or a single bar}] \times [\text{total points}].$$

In MATLAB, if the histogram command *[n,c]=hist(T,12)* will return a vector n containing the number of data points in each bin, and another vector c containing the center point of each bin. The binwidth can be computed from c as, for example, $c(2) - c(1)$. The scaling can be computed, then, as

$$\text{Scale} = (c(2) - c(1)) * \text{sum}(n).$$

Assuming x, *mu* and *sd* are defined as above, we use the following MATLAB script.

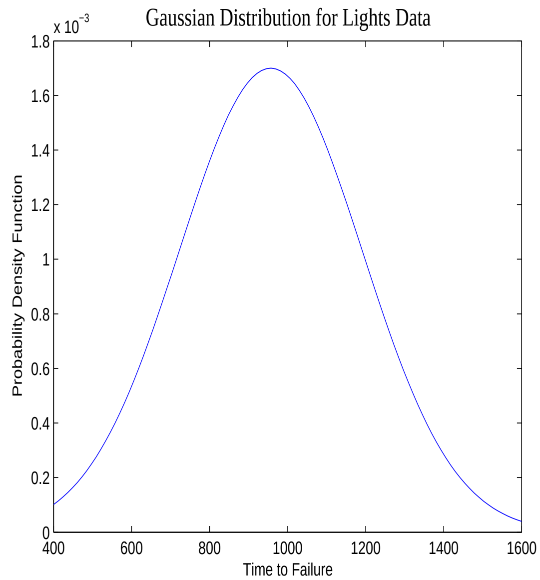


Figure 6.2: Gaussian distribution for Lights, Inc. data.

```
>>[n,c]=hist(T);
>>hist(T)
>>f=(c(2)-c(1))*sum(n)/(sqrt(2*pi)*sd)*exp(-(x-mu).^2/(2*sd^2));
>>hold on
>>plot(x,f,'r')
```

(see Figure 6.3).

The Gaussian distribution is typically useful when the values a random variable takes are clustered near its mean, with the probability that the value falls below the mean equivalent to the probability that it falls above the mean. Typical examples include the height of a randomly selected man or woman, the grade of a randomly selected student, and the velocity of a molecule of gas.

2. Uniform Distribution. The *uniform* probability density function has the form

$$f(x) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{otherwise} \end{cases},$$

and is applicable to situations for which all outcomes on some interval $[a, b]$ are equally likely. The mean of a uniformly distributed random variable is $\frac{a+b}{2}$, while the variance is $\frac{(b-a)^2}{12}$.

Example 4.3. Consider the game of American roulette, in which a large wheel with 38 slots is spun in one direction and a small white ball is spun in a groove at the top of the wheel in the opposite direction. Though Newtonian mechanics could ideally describe the outcome of roulette exactly, the final groove in which the ball lands is for all practical purposes random. Of course, roulette is a discrete process, but its probability density function can be well approximated by the uniform distribution. First, we create a vector R that contains the outcomes of 5000 spins:

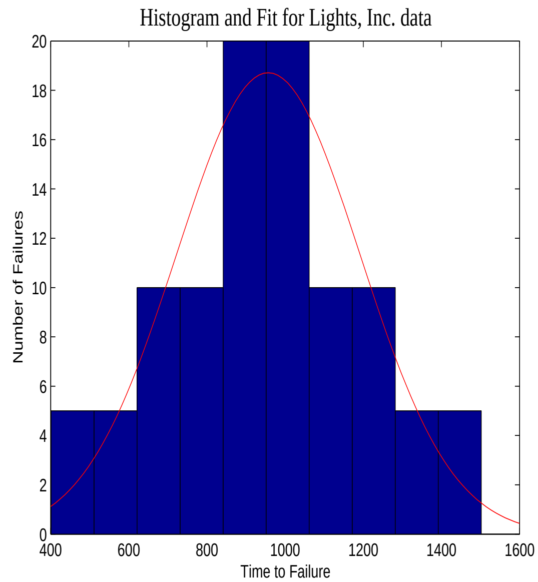


Figure 6.3: Lights, Inc. data with Gaussian distribution.

```
>>R=ceil(rand([5000,1])*38);
```

The following MATLAB code compares a histogram of this data with its probability density function (see Figure 6.4).

```
>>[n,c]=hist(R,38)
>>hist(R,38)
>>x=linspace(1,39,25);
>>f=(c(2)-c(1))*sum(n)*sign(x);
>>hold on
>>plot(x,f,'–')
```

3. Exponential distribution. The *exponential* probability density function is given by

$$f(x) = \begin{cases} ae^{-ax}, & x > 0 \\ 0, & x < 0 \end{cases},$$

where $a > 0$. This distribution is often employed as a model for the random variable time in situations in which the time remaining until the next event is independent of the time since the previous event. Examples include the time between goals in a soccer match and the time between arrivals in a waiting line. The mean of an exponentially distributed random variable is $1/a$, while the variance is $1/a^2$.

Example 4.4. Consider the number of rolls between sixes on a fair die, where two sixes in a row correspond with zero roles. The M-file *roles1.m* creates a vector R containing 10,000 realizations of this random variable.

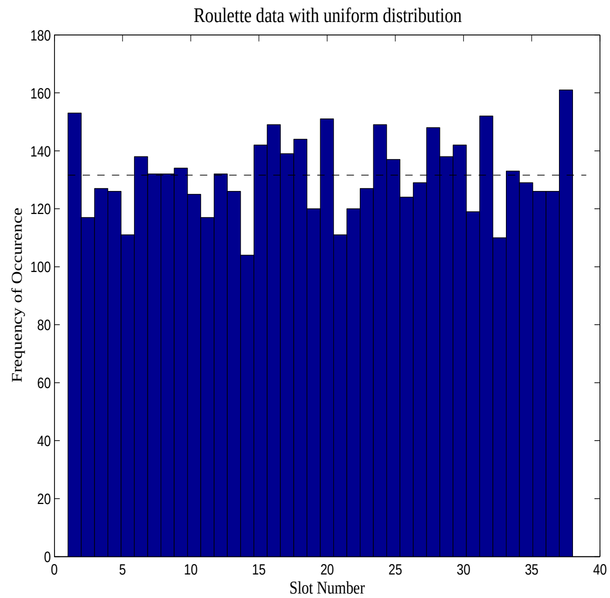


Figure 6.4: Uniform distribution with roulette data.

```
%ROLES1: Creates a list R of number of roles of
%a six-sided die between occurrences of a 6.
N=10000; %Number of sixes
clear R;
for k=1:N
m = 0; %Number of roles since last 6 (0 for 2 sixes in a row)
num = rand*6; %Random number between 1 and 6.
while num <= 5
m = m + 1;
num = rand*6; %Next role
end
R(k) = m;
end
```

The following MATLAB code produces Figure 6.5.

```
>>[n,c]=hist(R,max(R)+1)
>>hist(R,max(R)+1)
>>mu=mean(R)
mu =
4.9493
>>x=linspace(0,max(R),max(R));
>>f=(c(2)-c(1))*sum(n)*(1/mu)*exp(-x/mu);
>>hold on
>>plot(x,f,'–')
```

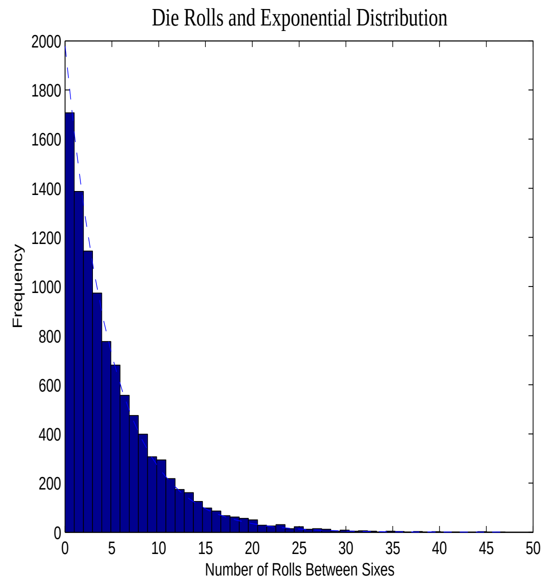


Figure 6.5: Histogram and exponential pdf for Example 5.8.

4. Weibull Distribution. The probability density function for the *Weibull* distribution is given by

$$f(x) = \begin{cases} \lambda^\beta \beta x^{\beta-1} e^{-(\lambda x)^\beta}, & x > 0, \\ 0, & x < 0 \end{cases},$$

where $\lambda > 0$ and $\beta > 0$, with mean and variance

$$E[X] = \frac{1}{\lambda}\Gamma(1+\frac{1}{\beta}), \quad \text{and} \quad \mathrm{Var}[X] = \frac{1}{\lambda^2}\Big(\Gamma(1+\frac{2}{\beta}) - \Gamma(1+\frac{1}{\beta})^2\Big),$$

where $\Gamma(\cdot)$ is the *gamma* function,

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt.$$

(MATLAB has a built-in gamma function, *gamma()*.) Observe that in the case $\beta = 1$ the Weibull distribution reduces to the exponential distribution. Named for the Swedish mechanical engineer Walloddi Weibull (1887–1979) who first suggested it, the Weibull distribution is widely used as a model for times to failure; for example, in the case of automotive parts. The Weibull probability density function for $\beta = 2$, $\lambda = 1$ and for $\beta = \frac{1}{2}$, $\lambda = 1$ is given in Figure 6.6.

5. Beta Distribution. The probability density function for the *beta* distribution is given by

$$f(x) = \begin{cases} \frac{1}{\beta(a,b)} x^{a-1}(1-x)^{b-1}, & 0 < x < 1, \\ 0, & \text{otherwise} \end{cases},$$

where $a > 0$ and $b > 0$, and where the *beta* function is defined as

$$\beta(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx.$$

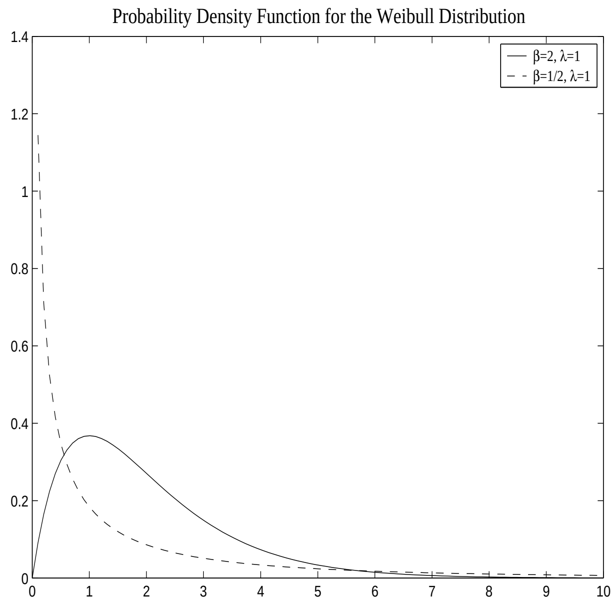


Figure 6.6: Probability density function for Weibull distribution.

(MATLAB has a built-in beta function, *beta()*.) The expected value and variance for beta random variables, X, are

$$E[X] = \frac{a}{a+b}; \quad \mathrm{Var}[X] = \frac{ab}{(a+b)^2(a+b+1)}.$$

The beta random variable is useful in the event of slow tails; that is, when the probability density function decays at algrebraic rate rather than exponential. The beta distribution for values $a = 2$, $b = 4$ and for $a = \frac{1}{2}$, $b = 2$ are given in Figure 6.7.

6. Gamma Distribution. The probability density function for the gamma distribution is given by

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{n-1}}{\Gamma(n)}, & x \geq 0 \\ 0, & x < 0, \end{cases}$$

for some $\lambda > 0$ and $n > 0$, where $\Gamma(\cdot)$ is the gamma function as defined in the discussion of the Weibull distribution above. The mean and variance of the gamma distribution are

$$E[X] = \frac{n}{\lambda}; \quad \mathrm{Var}[X] = \frac{n}{\lambda^2}.$$

When n is an integer, the gamma distribution is the distribution of the sum of n independent exponential random variables with parameter λ. The case $\lambda = 1, n = 2$ is depicted in Figure 6.8.

7. Mixture Distributions. Often, a random phenomenon will be divided into two or more characteristic behaviors. For example, if the random variable T represents service time in a certain coffee house, the time for specialty drinks may satisfy an entirely different distribution than the time for simple coffee.

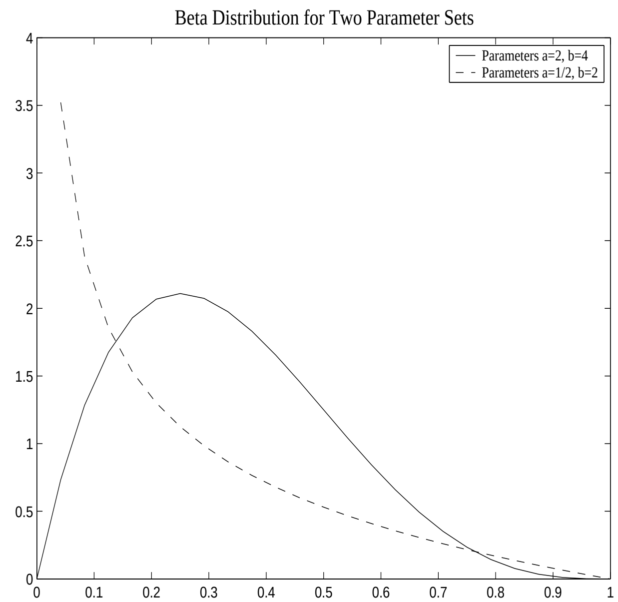


Figure 6.7: Probability Density Functions for Beta distribution.

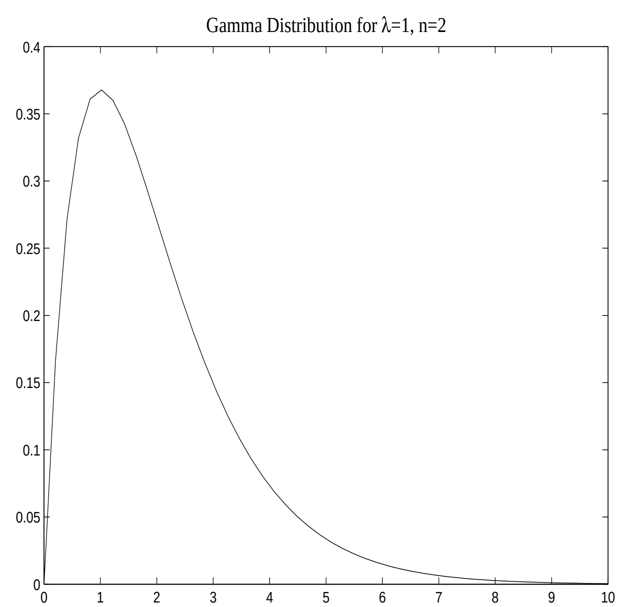


Figure 6.8: Probability density function for the gamma distribution.

Example 4.5. Consider a game that begins with two flips of a fair coin. If the two flips are both heads, then a coin is flipped 50 times and \$1.00 is paid for each tail, \$2.00 for each head. On the other hand, if the initial two flips are not both heads, then a fair six-sided die is rolled fifty times and \$1.00 is paid for each 1, \$2.00 is paid for each 2, \$3.00 for each 3 etc.

If we let R represent a vector containing 1,000 plays of the game. We can compute such a vector with the MATLAB M-file *mixture.m*. (For a discussion of simulating random variables, see Section 4.9.)

```
%MIXTURE: Script file to run example of a mixture
%random variable.
%Experiment: Flip a coin twice. If HH, flip the coin
%fifty more times and make $1.00 for each head, and
%$2.00 for each tail. If not HH, role a fair
%six-sided die 50 times and make $1.00 for
%each 1, $2.00 for each 2, etc.
global R; %Need R and N for mix1.m
global N;
N = 1000; %Number of times to run experiment.
for j=1:N
m=0;
if rand <= .25 %Flip two heads
for k=1:50
m = m + round(rand*2+.5);
end
else
for k=1:50
m = m + round(rand*6+.5);
end
end
R(j) = m;
end
```

The MATLAB command *hist(R,max(R))* creates the histogram in Figure 6.9.

As expected, the payoff from the fifty coin flips satisfies a distribution entirely different from that of the payoff from the fifty die rolls. In order to analyze this data, we first need to split the data into two parts, one associated with the coin flips and the other associated with the die rolls. Generally, the two (or more) distributions will run into one another, so this step can be quite difficult, but here it is clear that we should take one set of data for payoffs below, say, 110, and the other set of data above 110. We will refer to the former as S for *small* and the latter as B for *big*. Letting μ_s and σ_s represent the mean and standard deviation for S, and letting μ_b and σ_b represent the mean and standard deviation for B, we will fit each clump of data separately to a Gaussian distribution. We have,

$$f_s(x) = \frac{1}{\sqrt{2\pi}\sigma_s} e^{-\frac{(x-\mu_s)^2}{2\sigma_s^2}}; \quad f_b(x) = \frac{1}{\sqrt{2\pi}\sigma_b} e^{-\frac{(x-\mu_b)^2}{2\sigma_b^2}}.$$

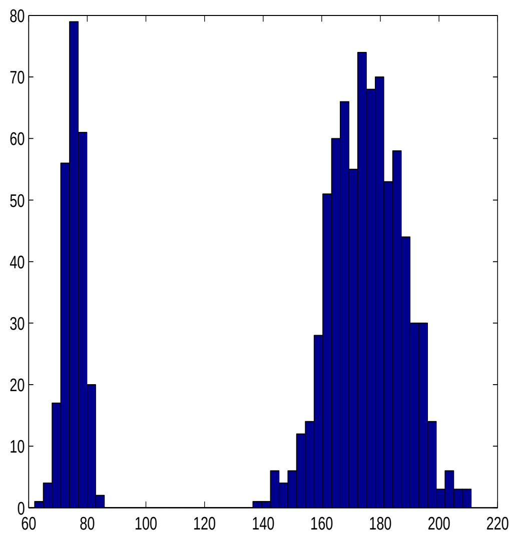


Figure 6.9: Histogram with data from Example 4.4.

In combining these into a single mixture distribution, we must be sure the integral of the final distribution over $\mathbb{R}$ is 1. (An integral over $f_s + f_b$ is clearly 2.) Letting p represent the probability that a play of our game falls into S, we define our mixture distribution by

$$f(x) = pf_s(x) + (1-p)f_b(x),$$

for which

$$\begin{aligned}\int_{-\infty}^{+\infty} f(x)dx &= \int_{-\infty}^{+\infty} \Big(pf_s(x) + (1-p)f_b(x)\Big)dx \\ &= p\int_{-\infty}^{+\infty} f_s(x)dx + (1-p)\int_{-\infty}^{+\infty} f_b(x) = p + (1-p) = 1.\end{aligned}$$

We first parse our data into two sets, S for the smaller numbers and B for the bigger numbers.

```
%MIX1: Companion file for mixture.m, cleans up the data.
global N;
global R;
i = 0;
l = 0;
for k=1:N
if R(k) <= 110
i = i + 1;
S(i) = R(k);
else
l = l + 1;
B(l) = R(k);
```

```
end
end
```

The following M-file now creates Figure 6.10.

```
%MIX1PLOT: MATLAB script M-file for comparing mixture
%distribution with histogram for data created in mixture.m
hist(R,50);
mus = mean(S); sds = std(S); mub = mean(B); sdb = std(B);
p = length(S)/(length(S)+length(B));
x = linspace(0, max(B), max(B));
fs = 1/(sqrt(2*pi)*sds)*exp(-(x-mus).^2/(2*sds^2));
fb = 1/(sqrt(2*pi)*sdb)*exp(-(x-mub).^2/(2*sdb^2));
[n,c]=hist(R, 50);
f = sum(n)*(c(2)-c(1))*(p*fs+(1-p)*fb);
hold on
plot(x,f,'r')
```

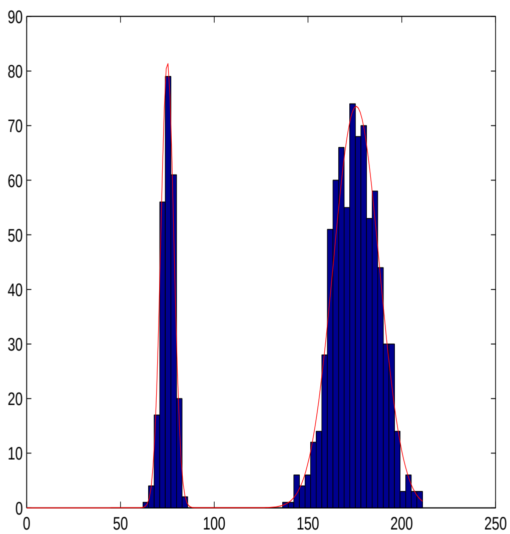


Figure 6.10: Mixture distribution with data from Example 4.5.

6.6 More Probability Density Functions

In this section, we list for convenient reference ten additional PDF, though detailed discussions are omitted. We continue our numbering scheme from Section 4.5.

8. Cauchy Distribution. The PDF for the Cauchy distribution is

$$f(x) = \frac{1}{\pi\beta[1+(\frac{x-\alpha}{\beta})^2]},$$

where $-\infty < \alpha < \infty$, and $\beta > 0$. The expected value and variance for the Cauchy distribution are infinite.

9. Lognormal Distribution. The PDF for the lognormal distribution is

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, \quad x > 0,$$

where $-\infty < \mu < \infty$ and $\sigma > 0$. The lognormal distribution arises through exponentiation of the Gaussian distribution. That is, if G is a Gaussian random variable with mean μ and variance σ^2, then the random variable $X = e^G$ has the PDF given above. In this case, we have

$$\begin{aligned} E[X] &= e^{\mu + \frac{1}{2}\sigma^2} \\ \mathrm{Var}[X] &= e^{2\mu + 2\sigma^2} - e^{2\mu + \sigma^2}. \end{aligned}$$

This random variable plays a fundamental role in the modeling of stock prices.

10. Double Exponential (or Laplace) Distribution. The PDF for the double exponential distribution is

$$f(x) = \frac{1}{2\beta} e^{-\frac{|x-\alpha|}{\beta}},$$

where $-\infty < \alpha < \infty$ and $\beta > 0$. If X has a double exponential distribution, then

$$\begin{aligned} E[X] &= \alpha \\ \mathrm{Var}[X] &= 2\beta^2. \end{aligned}$$

11. Logistic Distribution. The PDF for the logistic distribution is

$$F(x) = \frac{1}{\beta} \frac{e^{-\frac{x-\alpha}{\beta}}}{(1 + e^{-\frac{x-\alpha}{\beta}})^2},$$

where $-\infty < \alpha < \infty$ and $\beta > 0$. If X has a logistic distribution, then

$$\begin{aligned} E[X] &= \alpha \\ \mathrm{Var}[X] &= \frac{\beta^2 \pi^2}{3}. \end{aligned}$$

12. Rayleigh Distribution. The PDF for the Rayleigh distribution is

$$f(x) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}, \quad x > 0$$

where $\sigma > 0$. If X has a Rayleigh distribution, then

$$\begin{aligned} E[X] &= \sigma \sqrt{\frac{\pi}{2}} \\ \mathrm{Var}[X] &= 2\sigma^2 (1 - \frac{\pi}{4}). \end{aligned}$$

The Rayleigh distribution is used in the modeling of communications systems and in reliability theory.

13. Pareto Distribution. The PDF for the Pareto distribution is

$$f(x) = \frac{\beta\alpha^\beta}{x^{\beta+1}}, \quad x > \alpha,$$

where $\alpha > 0$ and $\beta > 0$. If X has a Pareto distribution, then

$$E[X] = \frac{\beta\alpha}{\beta - 1}, \quad \beta > 1$$
$$\mathrm{Var}[X] = \frac{\beta\alpha^2}{(\beta-1)^2(\beta-2)}, \quad \beta > 2,$$

where for $0 < \beta \leq 1$ the expected value is infinite and for $0 < \beta \leq 2$ the variance is infinite.

14. Extreme value (or Gumbel) Distribution. The PDF for the extreme value distribution is

$$f(x) = e^{-e^{-\frac{x-\alpha}{\beta}}} \frac{1}{\beta} e^{-\frac{x-\alpha}{\beta}},$$

where $-\infty < \alpha < \infty$ and $\beta > 0$. If X has an extreme value distribution, then

$$E[X] = \alpha + \beta\gamma$$
$$\mathrm{Var}[X] = \frac{\pi^2\beta^2}{6}.$$

Here $\gamma \cong .577216$ is Euler's constant.

15. Chi-square Distribution. The PDF for the chi-square distribution is

$$f(x) = \frac{1}{\Gamma(k/2)}(\frac{1}{2})^{k/2} x^{k/2-1} e^{-\frac{1}{2}x}, \quad x > 0,$$

where $k = 1, 2, \ldots$ is called the *number of degerees of freedom*. If X has a chi-square distribution, then

$$E[X] = k$$
$$\mathrm{Var}[X] = 2k.$$

The chi-square distribution is often the distribution satisfied by a test statistic in hypothesis testing. It is precisely the distribution of the sum of the squares of k independent standard normal random variables,

$$X = N_1^2 + N_2^2 + \cdots + N_k^2.$$

16. t distribution. The PDF for the t distribution is

$$f(x) = \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} \frac{1}{\sqrt{k\pi}} \frac{1}{(1 + x^2/k)^{(k+1)/2}},$$

where $k > 0$. If X has a t distribution, then

$$E[X] = 0, \quad k > 1$$
$$\mathrm{Var}[X] = \frac{k}{k-2}, \quad k > 2.$$

The t distribution is the distribution for

$$X = \frac{N}{\sqrt{\chi/k}},$$

where N is a standard normal random variable, and χ is a chi-square random variable with k degrees of freedom.

17. F distribution. The PDF for the F distribution is

$$f(x) = \frac{\Gamma(\frac{r_1+r_2}{2})}{\Gamma(\frac{r_1}{2})\Gamma(\frac{r_2}{2})} r_1^{r_1/2} r_2^{r_2/2} \frac{x^{r_1/r_2-1}}{(r_2+r_1x)^{(r_1+r_2)/2}}, \quad x > 0,$$

where $m = 1, 2, \ldots$ and $n = 1, 2, \ldots$. If X has an F distribution, then

$$E[X] = \frac{r_2}{r_2-2}, \quad r_2 > 2$$
$$\mathrm{Var}[X] = \frac{2r_2^2(r_1+r_2-2)}{r_1(r_2-2)^2(r_2-4)}, \quad r_2 > 4.$$

The F distribution is the distribution for

$$X = \frac{\chi_1}{r_1}\frac{\chi_2}{r_2},$$

where χ_1 and χ_2 are independent chi-square random variables with r_1 and r_2 degrees of freedom respectively.

6.7 Joint Probability Density Functions

Definition. Given two random variables X and Y, we define the joint cumulative probability distribution function as

$$F_{X,Y}(x, y) = \Pr\{X \leq x, Y \leq y\}.$$

Definition. We say that X and Y are *jointly continuous* if there exists a function $f(x, y)$ defined for all real x and y so that for every set C of pairs of real numbers,

$$\Pr\{(X, Y) \in C\} = \int\int_{(x,y)\in C} f(x, y)dxdy.$$

The function $f(x, y)$ is called the *joint probability density function* of X and Y, and whenever F is twice differentiable, we have the relationship

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

6.8 Maximum Likelihood Estimators

Once we decide on a probability density function that we expect will reasonably fit our data, we must use our data to determine values for the parameters of the distribution. One method for finding such parameter values is that of maximum likelihood estimation.

6.8.1 Maximum Likelihood Estimation for Discrete Random Variables

Example 4.6. Consider a coin, known to be unfair, with probability of heads either $p = .6$ or $p = .4$, and suppose we would like to use experimental data to determine which is correct. That is, we are trying to estimate the value of the parameter p.

First, consider the case in which we are only given one flip on which to base our decision. Let X be a random variable representing the number of heads that turn up in a given flip, and let $f(x; p)$ be the probability density function associated with X. (Since X is a discrete random variable, f would often be referred to here as a probability mass function: $f(x; p) = \Pr\{X = x|p\}$.) We have two possibilites for f,

$$f(x; .6) = \begin{cases} 1, & \text{prob } .6 \\ 0, & \text{prob } .4 \end{cases}, \quad f(x; .4) = \begin{cases} 1, & \text{prob } .4 \\ 0, & \text{prob } .6. \end{cases}$$

It will be useful to think in terms of a table of possible outcomes Table 6.2.

PDF/Number of Heads that turn up in experiment	0	1
$f(x; .6)$	.4	.6
$f(x; .4)$	.6	.4

Table 6.2: Analysis of an unfair coin with a single flip.

Clearly, the only conclusion we can make from a single flip is that if $X = 1$ (the coin turns up heads), we take $\hat{p} = .6$, while if $X = 0$, we take $\hat{p} = .4$, where in either case $\hat{p}$ represents our estimator of p. Looking down each column, we observe that we are choosing $\hat{p}$ so that

$$f(x; \hat{p}) \geq f(x; p).$$

That is, if the experiment turns up $X = 0$, we have two possible values, .4 and .6, and we choose $\hat{p}$ so that we get .6, the larger of the two. Looking at the third row of our table, we see that this corresponds with the choice $\hat{p} = .4$. We proceed similarly for the case $X = 1$.

In the case that we have two flips to base our estimate on, the probability density function becomes $f(x_1; p)f(x_2; p)$, where f is the PDF for a single flip. Proceeding as above, we have Table 6.3.

PDF/Number of heads	$x_1 = 0$, $x_2 = 0$	$x_1 = 0$, $x_2 = 1$	$x_1 = 1$, $x_2 = 0$	$x_1 = 1$, $x_2 = 1$
$f(x_1; .6)f(x_2; .6)$	$.4^2$	$.4 \cdot .6$	$.6 \cdot .4$	$.6^2$
$f(x_1; .4)f(x_2; .4)$	$.6^2$	$.6 \cdot .4$	$.4 \cdot .6$	$.4^2$

Table 6.3: Analysis of an unfair coin two flips.

Proceeding exactly as in the case of a single flip, we determine that in the case $x_1 = 0$, $x_2 = 0$ the MLE is .4, while in the case $x_1 = 1$, $x_2 = 1$, the MLE is .6. In the remaining two cases, the experiment does not favor one value over another. $\triangle$

Remark. In the expression

$$f(x; \hat{p}) \geq f(x; p), \tag{6.2}$$

the variable x denotes the actual outcome. Since x is what happened, we assume it is the most likely thing to have happened. By choosing $\hat{p}$ so that (6.2) holds, we are choosing $\hat{p}$ so that x is the most likely thing to have happened.

6.8.2 Maximum Likelihood Estimation for Continuous Random Variables

The main observation we take from our discussion of MLE in the case of discrete random variables is that we choose $\hat{p}$ so that

$$f(x; \hat{p}) \geq f(x; p).$$

In this way, we observe that finding a maximum likelihood estimator is a maximization problem, and in the continuous case, we will be able to use methods from calculus.

Example 4.7. Suppose an experimental set of measurements x_1, x_2, ..., x_n appears to have arisen from an exponential distribution. Determine an MLE for the parameter a.

As in Example 4.6, we first consider the case of a single measurement, x_1. The PDF for the exponential distribution is

$$f(x; a) = ae^{-ax}, \quad x > 0,$$

and so we search for the value of a that maximizes

$$L(a) = f(x_1; a),$$

which is called the *likelihood function*. Keep in mind that our rationale here is precisely as it was in the discrete case: since x_1 is the observation we get, we want to choose a so as to make this as likely as possible—the assumption being that in experiments we most often see the most likely events. Here, we have

$$L'(a) = e^{-ax_1} - ax_1 e^{-ax_1} = 0 \Rightarrow e^{-ax_1}(1 - ax_1) = 0 \Rightarrow a = \frac{1}{x_1}.$$

In the case of n measurements x_1, x_2, ..., x_n, the likelihood function becomes the joint pdf

$$\begin{aligned} L(a) &= f(x_1; a) f(x_2; a) \cdots f(x_n; a) \\ &= \prod_{k=1}^{n} f(x_k; a) \\ &= \prod_{k=1}^{n} ae^{-ax_k} \\ &= a^n e^{-a \sum_{k=1}^{n} x_k}. \end{aligned}$$

In this case,

$$\frac{\partial L}{\partial a} = na^{n-1}e^{-a\sum_{k=1}^n x_k} - a^n(\sum_{k=1}^n x_k)e^{-a\sum_{k=1}^n x_k},$$

so that we have,

$$a = \frac{n}{\sum_{k=1}^n x_k}.$$

(Notice that since we are maximizing $L(a)$ on the domain $0 \le a < \infty$, we need only check that $L(0) = 0$ and $\lim_{a\to\infty} L(a) = 0$ to see that this is indeed a maximum.) In order to simplify calculations of this last type, we often define the *log-likelihood* function,

$$L^*(a) = \ln(L(a)),$$

simply the natural logarithm of the likelihood function. The advantage in this is that since natural logarithm is monotonic, L and L^* are maximized by the same value of a, and due to the rule of logarithms, if $L(a)$ is a product, $L^*(a)$ will be a sum. Also, in the case of a large number of data point, L can become quite large. In the current example,

$$L^*(a) = \ln(a^n e^{-a\sum_{k=1}^n x_k}) = \ln(a^n) + \ln(e^{-a\sum_{k=1}^n x_k}) = n\ln a - a\sum_{k=1}^n x_k,$$

so that

$$\frac{\partial L^*}{\partial a} = \frac{n}{a} - \sum_{k=1}^n x_k = 0,$$

which gives the same result as obtained above. △

More generally, we compute MLE with the aid of MATLAB.

Example 4.8. Given a set of 100 data points that appear to arise from a process that follows a Weibull distribution, determine maximum likelihood estimators for λ and β.

The data for this example can be computed with the M-file *weibsim.m:* (For a discussion of simulating random variables, see Section 4.9.)

```
%WEIBSIM: MATLAB script M-file that simulates 100
%Weibull random variables
lam = .5; bet = 2;
for k=1:100
X(k) = (1/lam)*(-log(1-rand))^(1/bet);
end
```

We first observe that the PDF for the Weibull distribution is

$$f(x) = \begin{cases} \lambda^\beta \beta x^{\beta-1} e^{-(\lambda x)^\beta}, & x > 0, \\ 0, & x < 0 \end{cases},$$

and so the likelihood function is

$$L(\lambda, \beta) = \prod_{k=1}^n \lambda^\beta \beta x_k^{\beta-1} e^{-(\lambda x_k)^\beta}.$$

In theory, of course, we can find values of λ and β by solving the system of equations

$$\frac{\partial L}{\partial \lambda} = 0$$
$$\frac{\partial L}{\partial \beta} = 0.$$

In practice, however, we employ MATLAB. First, we record the likelihood function in a function M-file, which takes values λ, β, and $\{x_k\}_{k=1}^n$, and returns values of the log-likelihood function $\ln(L)$ (see *weiblike.m*). (In this case, we use the log-likelihood function because of how large this product of 100 points becomes.)

```
function value = weiblike(p, X);
%WEIBLIKE: MATLAB function M-file that compute the likelihood function
%for a Weibull distrubution with data vector D
%Note: D is passed as a parameter
%p(1) = lambda, p(2) = beta
value = 0;
for k=1:length(X)
value=value+log(p(1)^(p(2))*p(2)*X(k)^(p(2)-1)*exp(-(p(1)*X(k))^(p(2))));
end
value = -value;
```

Observe that since MATLAB's optimization routines are for minimization, we multiply the function by a negative sign for maximization. In order to find the optimal values and plot our fit along with a histogram, we use *weibfit.m*.

```
function weibfit(X)
%WEIBFIT: MATLAB function M-file that fits data to a Weibull
%distribution, and checks fit by plotting a scaled PDF along
%with a histogram.
hold off;
guess = [1 1];
options=optimset('MaxFunEvals',10000);
[p, LL]=fminsearch(@weiblike,guess,options,X)
%Plotting
hist(X,12)
[n,c]=hist(X,12);
hold on;
x = linspace(min(X),max(X),50);
fweib = sum(n)*(c(2)-c(1))*p(1)^(p(2))*p(2)*x.^(p(2)-1).*exp(-(p(1)*x).^(p(2)));
plot(x,fweib,'r')
```

We obtain the fit given in Figure 6.11.

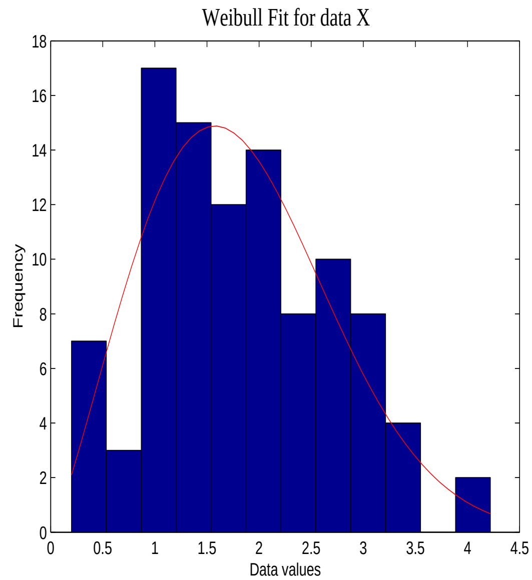


Figure 6.11: MLE PDF fit for Weibull distribution.

6.9 Simulating a Random Process

Sometimes the best way to determine how a certain phenomenon will play out is to simulate it several times and simply watch what happens. Generally, this is referred to as the *Monte Carlo* method, after a famous casino by that name in Monaco (a French province on the Mediterranean).[12]

Example 4.9. What is the expected number of flips of a fair coin until it turns up heads?

At first glance, this might look like a difficult problem to study analytically. The problem is that if we begin computing the expected value from the definition, we get an infinite series,

$$E[N] = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + ...$$

That is, the probability that it takes one flip is $\frac{1}{2}$; the probability that it takes two flips is $\frac{1}{4}$ etc. In fact, we have already developed a method for analyzing problems like this in Section 3, but for now let's suppose we want to take an alternative approach. One method is to simply pull out a coin and begin flipping it, counting how many flips it takes for it to land heads. This should lead to some sort of list, say 2, 4, 1, 3, 2 for five trials. The average of these should give us an approximation for the expected value:

$$E[N] \cong \frac{2+4+1+3+2}{5} = \frac{12}{5}.$$

[12] Apparently, this name was given to the method by Nicholas Metropolis, a researcher at Los Alamos National Laboratory (located in New Mexico) in the 1940s, and was used to describe simulations they were running of nuclear explosions. The first appearance of the method appears to be in the 1949 paper, "The Monte Carlo Method," published by Metropolis and Stanislow Ulam in the Journal of the American Statistical Association.

The more trials we run, the better our approximation should be.

In general, we would like to carry out such simulations on the computer. We accomplish this through *pseudo random numbers,* which behave randomly so long as we don't watch them too carefully.[13] Our fundamental random variable generator from MATLAB will be the built-in function *rand*, which creates a real number, fifteen digits long, uniformly distributed on the interval [0,1]. (The fact that this number has a finite length means that it is not really a continuous random variable, but with fifteen digits, errors will only crop up in our calculations if we run 10^{15} simulations, which we won't.) In the following MATLAB code, we take *rand<=.5* to correspond with the coin landing on heads and *rand>.5* to correspond with it landing tails.

```
function ev = flips(n)
%FLIPS: MATLAB function M-file that simulates
%flipping a coin until it turns up heads. The
%input, n, is number of trials, and the
%output, ev, is expected value.
for k=1:n
m=1; %m counts the number of flips
while rand > .5 %while tails
m=m+1;
end
R(k)=m;
end
ev=mean(R);
```

We now compute as follows in the MATLAB Command Window.

```
>>flips(10)
ans =
1.6000
>>flips(100)
ans =
1.8700
>>flips(1000)
ans =
2.0430
>>flips(10000)
ans =
1.9958
```

Observe that as we take more trials, the mean seems to be converging to 2. △

[13] There is an enormous amount of literature regarding pseudo random variables, which we will ignore entirely. For our purposes, the random variables MATLAB creates will be sufficient. If, however, you find yourself doing serious simulation (i.e., getting paid for it) you should *at least* understand the generator you are using.

6.10 Simulating Uniform Random Variables

As mentioned above, the MATLAB built-in function *rand* creates pseudo random numbers uniformly distributed on the interval $[0, 1]$. In order to develop a new random variable, U, uniformly distributed on the interval $[a, b]$, we need only use *U=a+(b-a)*rand*.

6.11 Simulating Discrete Random Variables

We can typically build discrete random variables out of uniform random variables by either conditioning (through *if* or *while* statements) or rounding. Suppose we want to simulate a random variable that is 0 with probability 1/2 and 1 with probability 1/2. We can either condition,

```
if rand < .5
X = 0;
else
X = 1;
end
```

or round

```
X=round(rand)
```

Notice that in either case we ignore the subtle problem that the probability that *rand<.5* is slightly smaller than the complementary probability that *rand>=.5*, which includes the possibility of equality. Keep in mind here that MATLAB computes rand to fifteen decimal places of accuracy, and so the probability that *rand* is precisely .5 is roughly 10^{-14}.

As another example, suppose we want to simulate the role of a fair die. In this case, our *if* statement would grow to some length, but we can equivalently use the single line,

```
R=round(6*rand+.5)
```

Notice that the addition of .5 simply insures that we never get a roll of 0.

Another option, similar to using *round*, is the use of MATLAB's function *ceil* (think *ceiling*), which rounds numbers to the nearest larger integer. See also *floor*.

6.12 Simulating Gaussian Random Variables

MATLAB also has a built-in Gaussian random number generator, *randn*, which creates pseudo random numbers from a Gaussian distribution with mean 0 and variance 1 (such a distribution is also referred to as the *standard normal* distribution). In order to see how we can generate more general Gaussian random numbers, we let N represent a standard normal random variable; that is, $E[N] = 0$ and $\text{Var}[N] = E[(N - E[N])^2] = E[N^2] = 1$. Introducing the new random variable $X = \mu + \sigma N$, we have

$$E[X] = E[\mu + \sigma N] = E[\mu] + \sigma E[N] = \mu,$$

and

$$\begin{aligned}
\mathrm{Var}[X] =& E[(X - E[X])^2] = E[X^2 - 2\mu X + \mu] = E[(\mu + \sigma N)^2 - 2\mu(\mu + \sigma N) + \mu^2] \\
=& E[\mu^2 + 2\sigma\mu N + \sigma^2 N^2 - 2\mu^2 - 2\mu\sigma N + \mu^2] = \sigma^2 E[N^2] = \sigma^2,
\end{aligned}$$

from which we *suspect* that X is a Gaussian random variable with mean μ and standard deviation σ. In order to create pseudo random numbers from such a distribution in MATLAB, with mean *mu* and standard deviation *sigma*, we simply use *X=mu+sigma*randn*.

In the previous discussion, we have not actually proven that X is a Gaussian random variable, only that it has the correct expected value and variance. In order to prove that it is Gaussian distributed, we must show that it has the Gaussian PDF. In order to do this, we will compute its CDF and take a derivative. We compute

$$\begin{aligned}
F(x) =& \Pr\{X \leq x\} = \Pr\{\mu + \sigma N \leq x\} = \Pr\{N \leq \frac{x - \mu}{\sigma}\} \\
=& \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{y^2}{2}} dy.
\end{aligned}$$

We have, then, according to the fundamental theorem of calculus

$$f(x) = F'(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

which is indeed the PDF for a Gaussian distribution with mean μ and variance σ^2.

6.13 Simulating More General Random Variables

In order to simulate general random variables, we will require two theorems.

Theorem 4.1. Suppose the random variable X has a cumulative distribution function $F(x)$, where $F(x)$ is continuous and strictly increasing whenever it is not 0 or 1. Then $X \stackrel{d}{=} F^{-1}(Y)$, where Y is uniformly distributed on $[0, 1]$ and by $\stackrel{d}{=}$ we mean *equal in distribution*: that each random variable has the same distribution.

Proof. First, observe that for $y \in [0, 1]$ and Y uniformly distributed on $[0, 1]$, we have $\Pr\{Y \leq y\} = y$. Next, note that our assumptions of continuity and monotonicity on $F(x)$ require it to behave somewhat like the example cumulative distribution function sketched in Figure 6.12.

We have, then, that the cumulative distribution function for $X = F^{-1}(Y)$ is given by

$$F_{F^{-1}(Y)}(x) = \Pr\{F^{-1}(Y) \leq x\} = \Pr\{Y \leq F(x)\} = F(x),$$

where the first equality follows from the definition of cumulative distribution function, the second follows from continuity and monotonicity, and the third follows from our first observation of the proof. $\square$

Example 4.10. Assuming Y is a uniformly distributed random variable on $[0, 1]$, develop a random variable X in terms of Y that satisfies the exponential distribution.

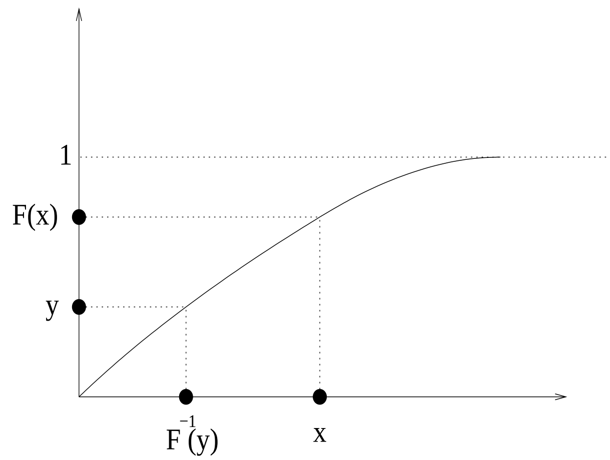


Figure 6.12: $F(x)$ continuous and strictly increasing.

First, we compute $F(x)$ for the exponential distribution by integrating over the probability density function,

$$F(x) = \int_0^x ae^{-ay}dy = -e^{-ay}|_0^x = 1 - e^{-ax}, \quad x \geq 0.$$

Clearly, $F(x)$ satisfies the conditions of Theorem 4.1, so all that remains is to find F^{-1}. We write $y = 1 - e^{-ax}$ and solve for x to find $x = -\frac{1}{a}\log(1-y)$, or in terms of X and Y, $X = -\frac{1}{a}\log(1-Y)$. △

The Rejection Method. A still more general method for simulating random variables is the *rejection* method. Suppose we can simulate random variables associated with some probability density function $g(x)$, and would like to simulate random variables from a second probability density function $f(x)$. The rejection method follows the steps outlined below.

1. Let c be a constant so that
$$\frac{f(x)}{g(x)} \leq c \quad \text{for all } x \in \mathbb{R},$$
where for efficiency c is to be chosen as small as possible.

2. Simulate both a random variable Y with density $g(y)$ and a random variable U uniformly distributed on $[0, 1]$.

3. If $U \leq \frac{f(Y)}{cg(Y)}$, set $X = Y$. Otherwise, repeat Step 2.

Theorem 4.2. The random variable X created by the rejection method has probability density function $f(x)$.

Proof. Let X be the random variable created by the rejection method, and compute its associated cumulative distribution function,

$$F_X(x) = \Pr\{X \leq x\} = \Pr\{Y \leq x | U \leq \frac{f(Y)}{cg(Y)}\} = \frac{\Pr\{\{Y \leq x\} \cap \{U \leq \frac{f(Y)}{cg(Y)}\}\}}{\Pr\{U \leq \frac{f(Y)}{cg(Y)}\}}.$$

Since Y and U are independent random variables, the joint probability density function of Y and U is

$$p(y,u) = g(y) f_U(u),$$

so that

$$\Pr\{\{Y \le x\} \cap \{U \le \frac{f(Y)}{cg(Y)}\}\} = \int_{-\infty}^{x} \int_{0}^{\frac{f(y)}{cg(y)}} g(y) dy du = \int_{-\infty}^{x} \frac{f(y)}{c} dy.$$

We have, then

$$\Pr\{X \le x\} = \frac{1}{c\Pr\{U \le \frac{f(Y)}{cg(Y)}\}} \int_{-\infty}^{x} f(y) dy.$$

Taking the limit as $x \to \infty$, we see that $c\Pr\{U \le \frac{f(Y)}{cg(Y)}\} = 1$, and consequently $F_X(x) = \int_{-\infty}^{x} f(y)dy$. □

Example 4.11. Develop a MATLAB program that simulates a beta random variable (in the case that the beta PDF is bounded).

For simplicity, we take our known probability density function $g(y)$ to be uniformly distributed on the interval $[0,1]$; i.e.,

$$g(y) = \begin{cases} 1, & 0 \le y \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

For c depending on the values of a and b, we simulate both Y and U as (independent) uniformly distributed random variables on $[0,1]$ and check of $U \le \frac{f(Y)}{c}$, where $f(Y)$ represents the beta probability density function evaluated at Y. If $U \le \frac{f(Y)}{c}$, we set $X = Y$, otherwise we repeat Step 2.

```
function b = ranbeta(a,b,c);
%RANBETA: function file for simulating a random variable
%with beta distribution. The value of c must be greater than
%the maximum of the beta distribution you are simulating,
%though not by much, or you will go through too many
%iterations. Employs the rejection method with comparison
%PDF g uniform on [0,1]; i.e., identically 1.
m = 0;
while m<1
var = rand; %Simulates Y
f = (1/beta(a,b))*var.^(a - 1).*(1 - var).^(b - 1);
if rand <= f/c %rand is simulating U
b = var;
m = 1;
end
end
```

As an example implementation, we will take $a = 2$ and $b = 4$. In this case,

$$f(x) = 20x(1-x)^3.$$

In order to select an appropriate value for c, we can find the maximum value of f. Setting $f'(x) = 0$, we find that the maximum occurs at $x = 1/4$, which gives $|f(x)| \leq 2.1094$. We can choose $c = 2.2$. We will simulate data and plot a histogram of the data along with a scaled pdf with the M-file *betafit.m*.

```
%BETAFIT: MATLAB script M-file that uses betasim.m to simulate
%a beta random variable and tests the result by plotting
%a scaled PDE along with a histogram
a = 2; b = 4; c = 2.2;
for k=1:1000
B(k) = ranbeta(a,b,c);
end
hist(B,12)
[n,c]=hist(B,12);
hold on;
x = linspace(0,1,100);
fbeta = sum(n)*(c(2)-c(1))/(beta(a,b))*x.^(a-1).*(1-x).^(b-1);
plot(x,fbeta,'r')
```

The resulting figure is Figure 6.13.

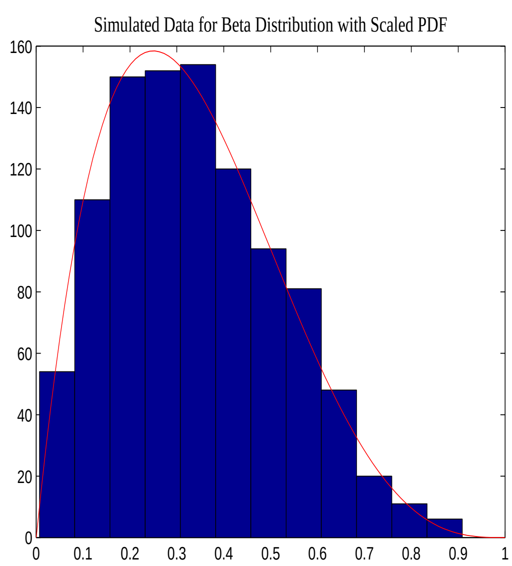


Figure 6.13: Plot of simulated beta distributed data along with scaled pdf.

6.14 Limit Theorems

When proceeding by simulation we typically make the following pair of assumptions:

1. If we take a large number of observations X_1, X_2, ..., X_n (n large) of the same process X, the average of these observations will be a good approximation for the average of the process:

$$E[X] \approx \frac{1}{n} \sum_{k=1}^{n} X_k.$$

2. If we define a random variable Y as the sum of n repeatable observations,

$$Y = \sum_{k=1}^{n} X_k,$$

then Y will be approximately Gaussion.

In this section we will investigate conditions under which these statements are justified. Results of this kind are typically referred to as Limit Theorems: those of Type 1 are "laws of large numbers, " while those of type 2 are "central limit theorems."

Lemma 4.3. (Markov's Inequality) If X is a random variable that takes only non-negative values, then for any $a > 0$

$$\Pr\{X \geq a\} \leq \frac{E[X]}{a}.$$

Proof. For $a > 0$ set

$$I_{\{x \geq a\}}(x) = \begin{cases} 1 & x \geq a \\ 0 & x < a \end{cases},$$

which is typically referred to as the *indicator function* for the set $\{x \geq a\}$. Since $X \geq a$, we clearly have

$$I_{\{x \geq a\}}(X) \leq 1 \leq \frac{X}{a},$$

and so

$$E[I_{\{x \geq a\}}(X)] \leq \frac{E[X]}{a}.$$

By the definition of expected value,

$$E[I_{\{x \geq a\}}(X)] = 0 \cdot \Pr\{X < a\} + 1 \cdot \Pr\{X \geq a\},$$

and the result is clear. □

Lemma 4.4. (Chebyshev's Inequality) If X is a random variable with finite mean μ and variance σ^2, then for any $k > 0$

$$\Pr\{|X - \mu| \geq k\} \leq \frac{\sigma^2}{k^2}.$$

Proof. We can prove this by applying Markov's inequality to the non-negative random variable $Z = (X - \mu)^2$. That is,

$$\Pr\{(X - \mu)^2 \geq k^2\} \leq \frac{E[(X - \mu)^2]}{k^2},$$

which is equivalent to the claim. $\square$

The importance of these last two inequalities is that they allow us to obtain information about certain probabilities without knowing the exact distributions of the random variables involved.

Example 4.12. Suppose that a certain student's exam average in a class is 75 out of 100, and that each exam is equally difficult. Find an upper bound on the probability that this student will make a 90 on the final.

Using Markov's inequality, we compute

$$\Pr\{X \geq 90\} \leq \frac{E[X]}{90} = \frac{75}{90} = \frac{5}{6}.$$

(Note that this calculation hinges on the fairly implausible assumption that the student's expected exam score is precisely $E[X] = 75$. On the other hand, we do not have to know anything about how X is distributed.) $\triangle$

Example 4.13. For the student discussed in Example 4.12 suppose the variance for X is 25. What is the probability the student's grade on the final will be between 60 and 90?

Using Chebyshev's inequality, we have

$$\Pr\{|X - 75| \geq 15\} \leq \frac{25}{15^2} = \frac{1}{9},$$

and so

$$\Pr\{|X - 75| < 15\} = \frac{8}{9}.$$

It should be fairly clear that both Markov's inequality and Chebyshev's inequality can give crude results.

Example 4.14. Suppose the random variable X from Examples 4.12 and 4.13 is known to be Gaussian. Compute $\Pr\{X \geq 90\}$.

In this case $\mu = 75$ and $\sigma^2 = 25$, so the Gaussian probability density function is

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

and consequently

$$\Pr\{X \geq 90\} = \int_{90}^{\infty} \frac{1}{\sqrt{50\pi}} e^{-\frac{(x-75)^2}{50}} dx = .0013,$$

or .13%. $\triangle$

Theorem 4.5. (The Weak Law of Large Numbers) Let X_1, X_2, $\ldots$ denote a sequence of independent and identically distributed random variables, each having the same finite mean

$$E[X_j] = \mu, \quad j = 1, 2, ..., n.$$

Then for any $\epsilon > 0$

$$\lim_{n\to\infty} \Pr\{|\frac{1}{n}\sum_{j=1}^{n} X_j - \mu| \geq \epsilon\} = 0.$$

Note. The weak law of large numbers was first proven by the Swiss mathematician Jacob Bernoulli (1654–1705) for the special case of Bernoulli random variables, and in the form stated here by the Russian Mathematician Aleksandr Yakovlevich Khintchine (1894–1959).

Proof. Though the theorem is true regardless of whether or not the variance associated with these random variables in finite, we prove it only for the case of finite variance.

It can be shown by direct calculation that if $E[X_j] = \mu$ and $\text{Var}[X_j] = \sigma^2$ then

$$E[\frac{1}{n}\sum_{j=1}^{n} X_j] = \mu \quad \text{and} \quad \text{Var}[\frac{1}{n}\sum_{j=1}^{n} X_j] = \frac{\sigma^2}{n}.$$

We now apply Chebyshev's Inequality to the random variable $Z = \frac{1}{n}\sum_{j=1}^{n} X_j$, giving

$$\Pr\{|Z - \mu| \geq k\} \leq \frac{\frac{\sigma^2}{n}}{k^2} = \frac{\sigma^2}{nk^2}. \tag{6.3}$$

Fixing now $\epsilon = k$, we see immediately that as $n \to \infty$ the probability goes to 0. □

In practical applications, inequality (6.3) is often more useful than the full theorem.

Example 4.15. Suppose a trick coin is to be flipped n times, and we would like to determine from this experiment the probability that it will land heads. We can proceed by defining a random variable X as follows:

$$X = \begin{cases} 1 & \text{if the coin lands heads} \\ 0 & \text{if the coin lands tails} \end{cases}.$$

Letting now X_1 denote the outcome of the first flip, X_2 the outcome of the second flip etc., we expect the probability that the coin lands heads to be approximately

$$p = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

For example, if $n = 100$ and heads turns up 59 times then

$$p = \frac{1}{100}\overbrace{(1 + 1 + \cdots + 1)}^{59 \text{ of these}} = \frac{59}{100} = .59.$$

We would now like to answer the following question regarding this approximation to the true probability of this coin's landing heads: What is the probability that the error for this approximation is larger than some threshold value, say .25? According to equation (6.3) we have

$$\Pr\{|.59 - \mu| \geq .1\} \leq \frac{\sigma^2}{100(.25)^2},$$

where since $0 \leq X \leq 1$ $\sigma^2 \leq 1$, and we have

$$\Pr\{|.59 - \mu| \geq .1\} \leq \frac{1}{6.25} = .16.$$

I.e., the probability of having an error this large is less than 16%. (Here $\mu = E[X]$ is the theoretically precise probability, which of course we don't know.)

More typically, we would like to turn this around and ask the following question: Find the minimum number of flips n required to ensure that the probability of a large error is small. (Notice that we must scrap the old value for p now, as we will have to make a new approximation with a larger value of n.) For example, let's find the number of flips required to ensure that with probability .95 the error will be less than .01. We need to find n so that

$$\Pr\{|Z - \mu| \geq .01\} = \frac{\sigma^2}{n(.01)^2} \leq \frac{1}{.0001n},$$

and we need to choose n large enough so that $\frac{1}{.0001n} \leq .05$. We choose

$$n \geq \frac{1}{.05(.0001)} = 200,000.$$

In order to verify that this is reasonable, let's simulate 200,000 flips of a fair coin and check that the error (on an expected probability of .5) is less than .01. In MATLAB,

```
>>sum(round(rand([200000 1])))/200000
ans =
0.4982
```

We see that the error is $|.5-.4982| = .0018$, which is certainly smaller than .01. (Bear in mind, however, that it is possible for the error to be larger than .01, but unlikely. Probabilistically speaking, it is possible to flip a coin 200000 times and get heads each time.)

7 Hypothesis Testing

Once we have determined which probability density function appears to best fit our data, we need a method for testing how good the fit is. In general, the analysis in which we test the validity of a certain statistic is called *hypothesis testing*. Before considering the case of testing an entire distribution, we will work through a straightforward example involving the test of a mean value.

7.1 General Hypothesis Testing

Example 5.1, Part 1. In the spring of 2003 the pharmaceutical company VaxGen published the results of a three-year study on their HIV vaccination. The study involved 5,009 volunteers from the United States, Canada, Puerto Rico, and the Netherlands. Overall, 97 out of 1679 placebo recipients became infected, while 190 out of 3330 vaccine recipients became infected. Of the 498 non-white, non-hispanic participants, 17 out of 171 placebo recipients became infected while 12 out of 327 vaccine recipients became infected. Determine whether it is reasonable for VaxGen to claim that their vaccination is successful.

Let N represent the number of participants in this study who were vaccinated ($N = 3330$), and let p_0 represent the probability of a placebo recipient becoming infected ($p_0 = \frac{97}{1679} = .058$). (Note that p_0 is the probability of infection in the absence of vaccination. We expect the the probability of a vaccine recipient becoming infected to be less than p_0.) Next, consider the possibility of repeating exactly the same study, and let the random variable X represent the number of vaccinated participants who become infected and the random variable p the probability that a vaccinated participant in the new study becomes infected ($p = \frac{X}{N}$). The goal of hypothesis testing in this case is to determine how representative p_0 is of the values that the random variable p can assume. (VaxGen would like to demonstrate that $p < p_0$, hence that the vaccine is effective.) Our benchmark hypothesis, typically referred to as the *null hypothesis* and denoted H_0, is that the vaccination is *not* effective. That is,

$$H_0: \qquad p = p_0 = .058.$$

We test our null hypothesis against our *alternative hypothesis*, typically denoted H_1, that the vaccine *is* effective. That is,

$$H_1: \qquad p < p_0.$$

In order to do this, we first observe that the random variable X is a binomial random variable with probability p and sample size N. That is, X counts the number of probability p events in N participants, where only two outcomes are possible for each participant, infection or non-infection. Here's the main idea. We are going to assume H_0, that $p = p_0 = .058$ is fixed. Our experimental observation, however, is that $p = p_1 = \frac{190}{3330} = .057$, a little better. We will determine the probability that our random sample determined $p \leq p_1$ given that the true underlying value is p_0. If this is highly unlikely, we reject the null hypothesis.[14]

In order to determine the probability that $p \leq p_1$, we need to develop a probability density function for p. Recalling that X is arises from a binomial distribution with sample size $N = 3330$ and probability $p = .058$. The MATLAB M-file *binomial.m* simulates binomial random variables with this distribution.

```
function value = binomial(p,N)
%BINOMIAL: Simulate a binomial random variable
%given its probability p and the number of
%trials N.
X = 0; %X represents the number of occurrences
for k=1:N
X=X+round(rand+(p-1/2));
end
value = X;
```

[14] Think about flipping a fair coin ten times and counting the number of times it lands heads. If the coin is fair, $p_0 = .5$, but for each ten-flip experiment $p_1 = \frac{\text{Number of heads}}{10}$ will be different. If $p_1 = .1$, we might question whether or not the coin is genuinely fair.

Typing *binomial(.058,3330)* in the MATLAB Command Window simulates the entire three-year VaxGen study once, and returns a value for the number of participants infected. In order to get a feel for the distribution of p, we will simulate 5000 such studies and look at a histogram of the results (see Figure 7.1).

```
>>for k=1:5000
X(k)=binomial(.058,3330);
end
>>hist(X,max(X))
```

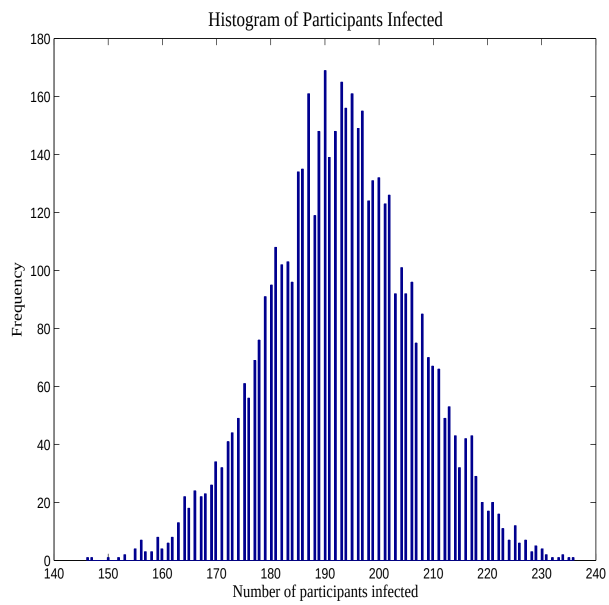


Figure 7.1: Histogram of number of participants infected.

We observe that X is well described by a Gaussian distribution. Since $p = \frac{X}{N}$, with N constant, p must also be well described by a Gaussian distribution. Though we could obtain the mean and variance of X or p directly from our simulation, the standard method is to use our observation that X is a binomial random variable to compute $E[X] = Np_0$ and $\text{Var}[X] = Np_0(1 - p_0)$ (see Section 3.2), from which $E[p] = E[\frac{X}{N}] = \frac{1}{N}E[X] = p_0$ and $\text{Var}[p] = \text{Var}[\frac{X}{N}] = \frac{1}{N^2}\text{Var}[X] = \frac{p_0(1-p_0)}{N}$. Setting, then

$$\mu = p_0 = .058, \quad \text{and} \quad \sigma = \sqrt{\frac{p_0(1-p_0)}{N}} = \sqrt{\frac{.058(1-.058)}{3330}} = .004,$$

we can compute probabilities on p from the Gaussian distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

In particular, we are interested in the probability that $p \leq p_1$. We have,

$$\Pr\{p \leq .057\} = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{.057} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = .4013,$$

where the integration has been carried out with MATLAB. We see that we have a 40% chance of getting $p \leq .057$ even if the vaccine is not effective at all. This is *not* good enough to reject H_0 and we conclude that VaxGen cannot legitimately claim that their vaccine is effective.

Example 5.1, Part 2. When this study came out, VaxGen understandably played down the main study and stressed the results in the non-white, non-hispanic subset of participants, for which $p_0 = \frac{17}{171} = .099$ and $p_1 = \frac{12}{327} = .037$. In this case,

$$\mu = p_0 = .099 \quad \text{and} \quad \sigma = \sqrt{\frac{.099(1 - .099)}{327}} = .0165,$$

from which we compute

$$\Pr\{p \leq .037\} = .00009.$$

In this case, the probability that $p \leq p_1$ is .009% and we reject the null hypothesis and claim that the vaccination is effective in this subset of participants.[15] △

A critical question becomes, how low does the probability of the event have to be before the null hypothesis is rejected. In Example 5.2, Part 1, the probability was over 40%, not that much better than a coin toss. We certainly cannot reject it based on that. In Example 5.2, Part 2, the probability was .009%: an anomaly like this might occur 9 times in 100,000 trials. In this case, we are clearly justified in rejecting the null hypothesis. What about cases in between: 10%, 5%, 1%? In the end, decisions like this are largely made on the requirement of accuracy.

7.2 Hypothesis Testing for Distributions

In the case of hypothesis testing for distributions, our idea will be to check our theoretical cumulative distribution function against the *empirical distribution function.*

7.2.1 Empirical Distribution Functions

For a random variable X and a set of n observations $\{x_1, x_2, ..., x_n\}$, we define the *empirical* distribution function $F_e(x)$ by

$$F_e(x) = \frac{\text{Number of } x_k \leq x}{n}.$$

Example 5.2. Consider again the Lights, Inc. data given in Table 6.1, and let $F_e(x)$ represent the empirical distribution function for the times to failure. Zero lights failed between 0 and 400 hours, so

$$F_e(400) = \frac{0}{n} = 0.$$

[15]Though several factors suggest that significantly more study is necessary, since the sample size is so small in this subset, and since in any set of data, subsets will exist in which results are favorable.

By 500 hours 2 lights had failed, while by 600 hours 5 lights had failed, and we have

$$F_e(500) = \frac{2}{100} = .02$$
$$F_e(600) = \frac{5}{100} = .05.$$

The MATLAB M-file *edf.m* takes a vector of observations and a point and calculates the empirical distribution of those observations at that point.

```
function f = edf(x,D);
%EDF: Function file which returns the empirical
%distribution function at a point x of a set
%of data D.
j = 1; m = 0;
l = length(D); %Compute length of D once
S = sort(D); %Puts data in ascending order
while S(j) <= x
m = m + 1;
j = j + 1;
if j == l + 1
break
end
end
f = m/l;
```

If T represents the data for Lights, Inc., the following MATLAB diary file shows the usage for *edf.m*. (Recall that the data for Light, Inc. is recorded in the MATLAB M-file *lights.m*.)

```
>>lights
>>edf(400,T)
ans =
0
>>edf(500,T)
ans =
0.0200
>>edf(600,T)
ans =
0.0500
>>edf(1500,T)
ans =
0.9800
>>edf(1600,T)
ans =
1
```

Recall that our theoretical distribution for this data was Gaussian with $\mu = 956.88$ and $\sigma = 234.69$. Our theoretical cumulative distribution function is

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy.$$

The MATLAB script M-file *edfplot.m* compares the theoretical distribution with the empirical distribution (see Figure 7.2).

```
%EDFPLOT: MATLAB script M-file for plotting
%the EDF along with the CDF for Lights, Inc. data.
x=linspace(400,1600,100);
for k=1:length(x);
Fe(k)=edf(x(k),T);
F(k)=quad('1/(sqrt(2*pi)*234.69)*exp(-(y-956.88).^2/(2*234.69^2))',0,x(k));
end
plot(x,Fe,x,F,'--')
```

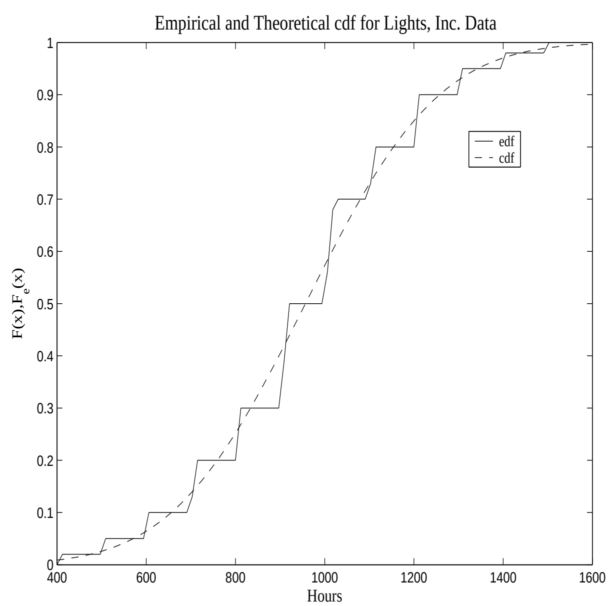


Figure 7.2: Empirical and Theoretical cumulative distribution functions for Lights, Inc. data.

Though certainly not perfect, the fit of our Gaussian distribution at least seems plausible. Determining whether or not we accept this fit is the subject of *hypothesis testing*. $\triangle$

Example 5.3. Consider again our data from Lights, Inc., summarized in Table 6.1 and denoted by the vector T. After looking at a histogram of this data, we determined that it was well described by a Gaussian distribution with $\mu = 956.88$ and $\sigma = 234.69$. Here, we consider whether or not this distribution is indeed a reasonable fit.

First, we require some quantifiable measure of how closely our distribution fits the data. (So far we've simply been glancing at the data and judging by the shape of its histogram.)

One method for doing this is to compare the proposed cumulative distribution function, $F(x)$, with the data's empirical distribution function, $F_e(x)$ (see Figure 7.2). Two standard tests are the *Kolmogorov–Smirnov* test and the *Cramer–von Mises* test.

1. Kolmogorov–Smirnov statistic. The simplest test statistic for cumulative distribution functions is the *Kolmogorov–Smirnov* statistic, defined by

$$D = \sup_{x \in \mathbb{R}} |F_e(x) - F(x)|.$$

Referring, for example, to Figure 7.2, the Kolmogorov–Smirnov statistic simply follows the entire plot of both functions and determines the greatest distance between them. In the event that F_e and F have been defined as vectors as in Section 5.1, we can compute the K–S statistic in MATLAB with the single command *D=max(abs(Fe-F))*. For our Lights, Inc. data $D = .0992$.[16]

2. Cramer–von Mises statistic. A second test statistic that measures the distance between $F(x)$ and $F_e(x)$ along the entire course of the functions is called the *Cramer-von Mises* statistic and is given by

$$W^2 = N \int_{-\infty}^{+\infty} (F_e(x) - F(x))^2 f(x) dx,$$

where N is the number of data points and $f(x) = F'(x)$ is the (proposed) probability density function. The C–vM statistic is slightly more difficult to analyze than the K–S statistic, largely because the empirical distribution function, $F_e(x)$, can be cumbersome to integrate. The primary thing to keep in mind is that integrands in MATLAB must accept vector data. We compute the integrand for the C–vM test with the MATLAB M-file *cvm.m*. Observe that the command *lights* simply defines the vector T (i.e., we could have replaced this command with *T=[415,478,...]*, the entire vector of data).

```
function value = cvm(x,T)
%CVM: Returns the Cramer-von Mises integrand
%for the given data and distribution.
for k=1:length(x)
F(k) = quad('1/(sqrt(2*pi)*234.69)*exp(-(y-956.88).^2/(2*234.69^2))',0,x(k));
Fe(k) = edf(x(k),T);
f(k) = 1/(sqrt(2*pi)*234.69)*exp(-(x(k)-956.88).^2/(2*234.69^2));
end
value = (F-Fe).^2.*f;
```

We now compute W^2 as *Wsq=length(T)*quad(@cvm,400,1600,[],[],T)*. We find $W^2 = .1347$.

Finally, we use the K–S and C–vM statistics to test the adequacy of our proposed distribution. Our null hypothesis in this case is that our proposed distribution adequately describes the data,

$$H_0 : F_d(x) = F(x),$$

[16]Observe that D can be calculated more accurately by refining x. Accomplish this by taking more points in the *linspace* command.

while our alternative hypothesis is that it does not,

$$H_1 : F_d(x) \neq F(x).$$

We test our null hypothesis by testing one of the test statistics described above. In order to accomplish this, we first need to determine the (approximate) distribution of our test statistic. (In our HIV example, our test statistic was p and its approximate distribution was Gaussian.)

Let's focus first on the K–S statistic, D. Observe that D is a random variable in the following sense: each time we test 100 bulbs, we will get a different outcome and consequently a different value for D. Rather than actually testing more bulbs, we can simulate such a test, assuming H_0, i.e., that the data really is arising from a Gaussion distribution with $\mu = 956.88$ and $\sigma = 234.69$. In the MATLAB M-file *kstest.m* we simulate a set of data points and compute a new value of D. Observe that the vectors x and F have already been computed as above, and T contains the original data.

```
function value = kstest(F,x,T)
%KSTEST: Takes a vector cdf F and a vector
%x and original data T and determines
%the Kolmogorov-Smirnov
%statistic for a sample taken from a
%normal distribution.
clear G; clear Fe;
N = length(F); %Number of points to consider
%Simulate Gaussian data
mu = mean(T);
sd = std(T);
for k=1:length(T) %Always simulate the same number of data points in ex-
periment
G(k)=mu+randn*sd;
end
%Compute empirical distribution function
for k=1:N
Fe(k)=edf(x(k),G);
end
%Kolmogorov-Smirnov statistic
value = max(abs(Fe-F));
```

Working at the MATLAB Command Window prompt, we have

```
kstest(F,x,T)
ans =
0.0492
kstest(F,x,T)
ans =
0.0972
```

```
kstest(F,x,T)
ans =
0.0655
```

We next develop a vector of D values and consider a histogram (see Figure 7.3).

```
>>for j=1:1000
D(j)=kstest(F,x,T);
end
>>hist(D,50)
```

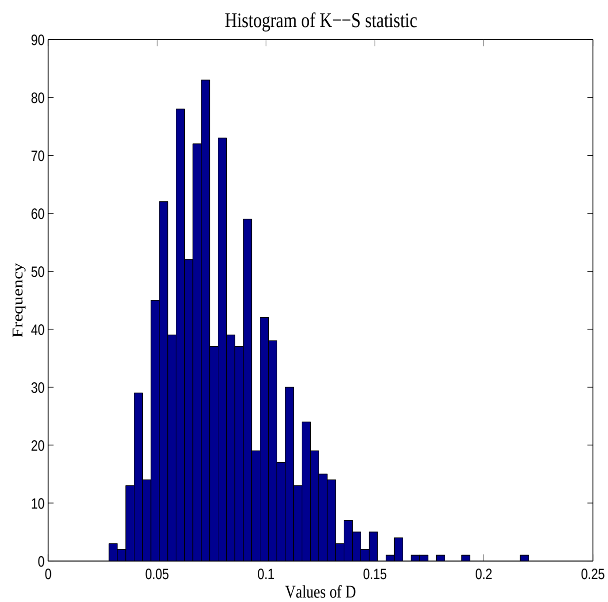


Figure 7.3: Histogram for K–S statistic, Lights, Inc. data.

Of the distributions we've considered, this data is best fit by either a beta distribution or a Weibull distribution. One method by which we can proceed is to develop an appropriate PDF for this new random variable D, and use it to compute the probabilities required for our hypothesis testing. In fact, this D is precisely the random variable D we analyzed in our section on maximum likelihood estimation. Observing, however, that we can create as many realizations of D as we like, and that as the number of observations increases, the empirical distribution function approaches the continuous distribution function, we proceed directly from the EDF. That is, we compute

$$\Pr\{D \geq .0992\} \cong 1 - F_e(.0992).$$

In MATLAB, the command *1-edf(.0992,D)* returns .23, or 23%. This signifies that if our distribution is correct, there remains a 23% chance that our statistic D will be as large as it was. Though we would certainly like better odds, this is typically considered acceptable. $\triangle$

8 Brief Compendium of Useful Statistical Functions

Mark Twain once quoted Benjamin Disraeli as having said, "There are three kinds of lies: lies, damn lies, and statistics."[17] Historically, the study of statistics arose in the fifteenth and sixteenth centuries when monarchs began to get serious about taxation and ordered census counts not only of people, but of the number of goats and sheep etc. they owned as well. At the time, it was described as "political arithmetic."

In this section, we review for ease of reference a number of basic statistical functions. To set down a definition, a *statistic* is simply an estimate or a piece of data, concerning some parameter, obtained from a sampling or experiment. Suppose, for instance, that we have some set of data T={1.7,1.75,1.77,1.79}—the height in meters of four randomly chosen men and women. In MATLAB, we could define this data as a vector and make the following computations of *mean, standard deviation, variance*, and *maximum*:

```
>>T=[1.7 1.75 1.77 1.79];
>>mean(T)
ans =
   1.7525
>>std(T)
ans =
   0.0386
>>var(T)
   0.0015
>>max(T)
ans =
   1.7900
```

For statistics such as *maximum* for which a particular element is selected from the list, it is also useful to know which index corresponds with the choice—in this case, 4. To get this information, we type

```
>>[m,k]=max(T)
m =
   1.7900
k =
    4
```

Three more routine statistics are *minimum*, *median*, and *sum*.

```
>>min(T)
ans =
   1.7000
>>median(T)
```

[17]It's cumbersome to use this "Mark Twain once quoted..." introduction, but oddly enough this quote has never been found among Disraeli's writings.

```
ans =
   1.7600
>>sum(T)
ans =
   7.0100
```

Occasionally, you will be interested in the *cumulative sum*, a vector containing the first entry of the data, the first two entries in the data summed, and the the first three entries in the data summed etc. The *product* and *cumulative product* are defined similarly.

```
>>cumsum(T)
ans =
   1.7000   3.4500   5.2200   7.0100
>>prod(T)
ans =
   9.4257
>>cumprod(T)
ans =
   1.7000   2.9750   5.2657   9.4257
```

The difference between each successive pair of data points can be computed with *diff()*.

```
>>diff(T)
ans =
   0.0500   0.0200   0.0200
```

Certainly, a tool useful in the manipulation of data is the ability to sort it. In the following example, the data stored in the variable Y is sorted into ascending order.

```
>>Y=[4 3 6 7 1];
>>sort(Y)
ans =
    1    3    4    6    7
```

9 Application to Queuing[18] Theory

In this section, we consider an application to the study of queueing theory. Suppose we want to simulate various situations in which customers randomly arrive at some service station. The following script M-file, *queue1.m*, simulates the situation in which the customer arrival times are exponentially distributed and the service times are fixed at exactly .5 minutes per customer (or $\mu = 2$ customers/minute).

[18]Sometimes spelled *Queueing* Theory, as for example in Sheldon Ross's infludential textbook *Introduction to Probability Models*, Academic Press 1989. In adopting the convenction *queuing*, I'm following Bryan A. Garner's *A Dictionary of Modern American Usage*, Oxford University Press, 1998.

```
%QUEUE1: Script file for simulating customers
%arriving at a single queue.
S=.5; %Time to service customers (1/mu)
m=.5; %Mean time between customer arrivals (1/lambda)
Q=0; %Time to service all customers remaining in queue (I.e., time until
% an arriving customer is served.)
queuewait=0; %Total time customers spent waiting in queue
systemwait=0; %Total time cusomers spent waiting in system
N0=0; %Number of times arriving customer finds no one in queue
N=1000; %Number of customers to watch
% The simulation
for k=1:N %Watch N customers
T=-m*log(1-rand); %Arrival of customer C (exponential)
Q=max(Q-T,0); %Time T has elapsed since time Q was set
if Q==0
N0=N0+1;
end
queuewait=queuewait+Q; %Total time spent waiting in queue
Q = Q + S; %Time until customer C leaves system
systemwait=systemwait+Q; %Total time spent waiting in system
end
Wq=queuewait/N
W=systemwait/N
P0=N0/N
```

You will need to go through this file line by line and make sure you understand each step. One thing you will be asked to do in the Queuing Theory project is to modify this file to accomodate a number of related situations.

10 Application to Finance

In this section, we introduce some of the basic mathematical tools involved in the study of finance.

10.1 Random Walks

Suppose two individuals flip a fair coin each day, and exchange money according to the following rule: if the coin lands with heads up, player A pays player B one dollar, while if the coin lands tails up, player B pays player A one dollar. Define the series of random variables

$$X_t = \text{Player A's earnings on day}\, t.$$

For example, we have

$$X_1 = X = \begin{cases} +1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2 \end{cases},$$

while

$$X_2 = X_1 + X = \begin{cases} +2 & \text{with probability } 1/4 \\ 0 & \text{with probability } 1/2 \\ -2 & \text{with probability } 1/4 \end{cases},$$

where X represents the random event carried out each day and the pattern continues with $X_{t+1} = X_t + X$. Defined as such, X_t is a *random walk*; more generally, any process such as X_t for which each value of t corresponds with a different random variable will be referred to as a *stochastic process*. We make the following critical observations:

1. $X_{t+1} - X_t$ is independent of $X_t - X_{t-1}$, which in turn is independent of $X_{t-1} - X_{t-2}$ etc. That is, the coin toss on any given day is entirely independent of all preceding coin tosses.
2. $E[X_t] = 0$ for all t.
3. $\text{Var}[X_t] = t$ for all t.

Observe that the recursive nature of X_t makes it particularly easy to simulate with MATLAB. For example, the following M-file serves to simulate and plot a random walk.

```
%RANWALK: MATLAB M-file written for the purpose
%of simulating and plotting a random walk.
clear X; %Initialize random variable
N = 50; %Number of steps in the random walk
X(1) = 0; %Start at 0
for k=1:N
if rand <= .5
flip = 1;
else
flip = -1;
end
X(k+1) = X(k) + flip;
end
t = 0:N;
plot(t,X)
```

The result is given in Figure 10.1.

10.2 Brownian Motion

In 1905, at the age of 26, a little known Swiss patent clerk in Berne published four landmark papers: one on the photoelectric effect (for which he would receive a Nobel prize in 1921), two on the theory of special relativity, and one on the transition density for a phenomenon that had come to be known as Brownian motion. The clerk was of course Albert Einstein, so we're in good company with our studies here.[19]

[19]Though Einstein was working in Switzerland at the time, he was born in Ulm, Germany.

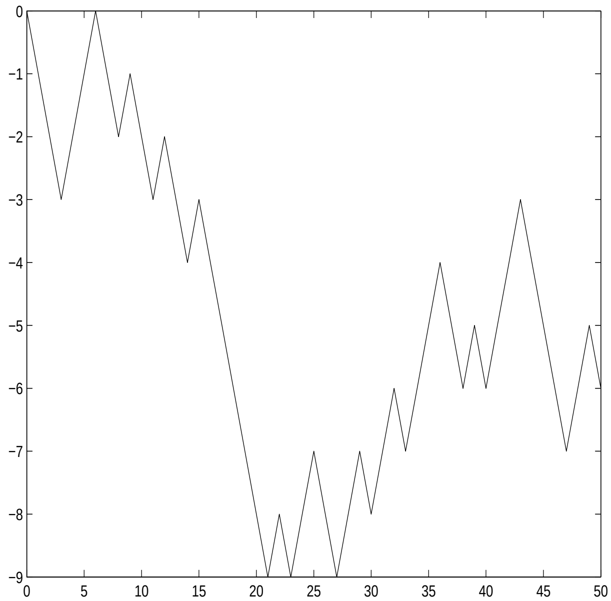


Figure 10.1: Random walk over 50 days.

Brownian motion is the name given to the irregular movement of pollen, suspended in water, observed by the Scottish botanist Robert Brown in 1828. As Einstein set forth in his paper of 1905, this movement is caused by collisions with water molecules. The dynamics are so complicated that the motion appears completely random. Our eventual goal will be to use the framework of mathematics built up around Brownian motion to model "random" behavior in the stock market.

A good intuitive way to think of Brownian motion is as a continuous time random walk. The random walk discussed in Section 8.1 is carried out at discrete times, one flip each day. Of course, we could increase the number of flips to say 2 each day, or 4 or 8. As the number of flips increases, the process looks more and more like a continuous process: at some point we find ourselves doing nothing else besides flipping this stupid coin. Rather, however, than actually flip all these coins, we'll simply build Brownian motion from a continuous distribution, the normal distribution. As it turns out, these two approaches—infinite flipping and normal distributions—give us exactly the same process (though we will make no effort to prove this).

The particular definition we'll use is due to Norbert Wiener (1894–1964), who proposed it in 1918. (Compare with Properties 1, 2, and 3 from Section 8.1.)

Definition 1. (Standard Brownian Motion)[20] A stochastic process B_t, $t \geq 0$, is said to be a standard Brownian motion if:

1. $B_0 = 0$.

[20] We omit entirely any attempt to prove that a process satisfying this definition exists. Such proofs are not for the faint of heart. Ye who yet dare are hereby referred to Chapter 2 of [I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd Ed., Springer–Verlag 1991]. Best of luck.

2. $B_{t_n} - B_{t_{n-1}}$, $B_{t_{n-1}} - B_{t_{n-2}}$, ..., $B_{t_2} - B_{t_1}$, B_{t_1} are all independent with distributions that depend (respectively) only on $t_n - t_{n-1}$, $t_{n-1} - t_{n-2}$, ..., $t_2 - t_1$, t_1; that is, only on the time interval between observations.

3. For each $t \geq 0$, B_t is normally distributed with mean 0 and variance t. ($E[B_t] = 0$, $\text{Var}[B_t] = t$.)

Critically, Property 3 is equivalent to the assertion that the probability density function for a standard Brownian motion B_t is

$$p(t, x) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}.$$

Hence, we can carry out all the usual analyses. For example, to compute the probability that B_t lies on the interval $[a, b]$, we compute

$$Pr\{a \leq B_t \leq b\} = \frac{1}{\sqrt{2\pi t}} \int_a^b e^{-\frac{x^2}{2t}} dx.$$

Just as with random walks, Brownian motion is straightforward to simulate using MATLAB. Recalling from Section 4.8 that MATLAB's command *randn* creates a normally distributed random variable with mean 0 and variance 1, we observe that $\sqrt{t}$**randn* will create a random variable with mean 0 and variance t. That is, if N represents a standard normal random variable (the random variable simulated by *randn*), we have

$$\begin{aligned} E[\sqrt{t}N] &= \sqrt{t}E[N] = 0 \\ \text{Var}[\sqrt{t}N] &= E[tN^2] - E[\sqrt{t}N]^2 = tE[N^2] = t. \end{aligned}$$

Our fundamental building block for Brownian paths will be the MATLAB function M-file *brown.m*, which will take a time t and return a random value associated with that time from the appropriate distribution.

```
function value = brown(t);
%BROWN: Returns the value of a standard Brownian motion
%at time t.
value = sqrt(t)*randn;
```

A *Brownian path* is simply a series of realizations of the Brownian motion at various times t. For example, the following M-file *snbm.m* describes a Brownian path over a period of 50 days, with a realization at each day.

```
%SNBM: Simulates a standard normal Brownian motion.
T = 50; %Number of time steps
delT = 1; %Time increment
time = 0; %Starting time
t(1) = 0; %Begin at t = 0
B(1) = 0; %Corresponds with t = 0
```

```
for k = 2:T+1;
t(k) = t(k-1) + delT;
B(k) = B(k-1) + brown(delT);
end
plot(t,B)
```

A resulting path is given in Figure 10.2.

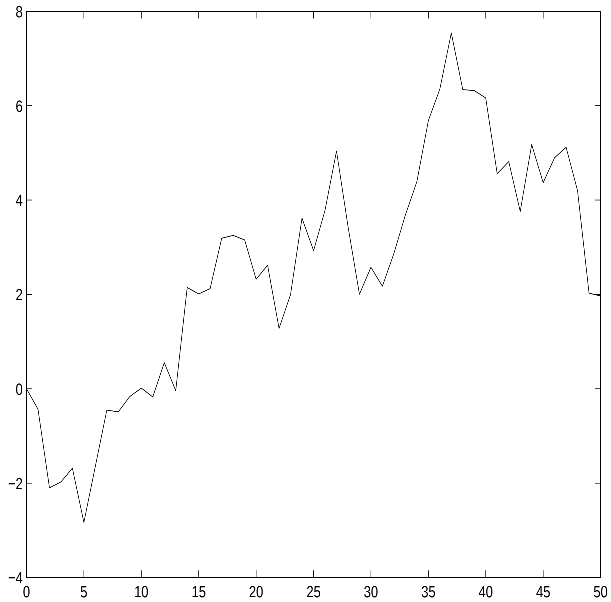


Figure 10.2: Standard Brownian motion over 50 days.

10.3 Stochastic Differential Equations

In the event that there was nothing whatsoever deterministic about the stock market, Brownian motion alone would suffice to model it. We would simply say that each day the probability of the stock's value rising a certain amount is exactly the same as the probability of its declining by the same amount. If this were the case, however, the stock market would most likely have closed a long time ago.

A popular and reasonably effective measure of overall stock market behavior is the Dow Jones industrial average, initiated by the journalist Charles Dow in 1884, along with his fellow journalist and business partner Edward Jones. In 1882, Dow and Jones formed Dow Jones & Company, which gathered and sold news important to financial institutions (this newsletter would later morph into the Wall Street Journal.) The stock market was relatively new and untested at the time, and one thing companies wanted was an idea of how it was behaving in general. Dow's idea was to follow 11 important stocks (mostly railroads back then) in hopes of finding overall market trends. On October 1, 1928, long after Dow's death, the number of stocks in his average was fixed at 30, the current number. Along with a number of companies now defunct, this list included such familiar names as Chrysler, General Electric, and Sears, Roebuck. At the end of 1928, the DJIA (Dow Jones Industrial Average) stood at roughly

300. Today (11/13/06), after a number of shaky years, it's 12,131.9. Over 78 years this corresponds with an average yearly increase of 4.9%. [21]

In order to introduce this kind of deterministic growth into our model for a stock, we might simply add a term that corresponds with this growth. Recalling that continuously compounded interest grows like Pe^{rt}, where P is the principal investment (cf. footnote) we have the stock price model

$$S_t = S_0 e^{rt} + B_t,$$

determined growth corrected by a random fluctuation.

The next step requires some motivation. Recall that in our section on mathematical biology we found it much easier to write an equation for the change in population than for the population itself. That is, if I simply asked you to write an equation of the form $p(t) = ...$, for the population of geese in Finland, you would find it quite difficult deciding how to start. However, as we've seen, the logistics model for the change in population

$$\frac{dp}{dt} = rp(1 - \frac{p}{K})$$

is straightforward to formulate. What I'm getting at here, is that we would like something akin to differential equations for these stochastic processes. Though a quick glance at the jagged turns in Figures 10.1 and 10.2 should convince you that stochastic processes don't generally have classical derivatives, we can (for the sake of all that $$ to be made gambling on the stock market) define the *differential form* (i.e., a multidimensional polynomial of differentionals)

$$dS_t = S_0 r e^{rt} dt + dB_t, \tag{10.1}$$

where we will look on dt as a small but finite increment in time. Equations of form (10.1) are called *stochastic differential equations.*

More generally, the stochastic differential equation for a reasonably well-behaved stochastic process X_t can be written in the form

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dB_t. \tag{10.2}$$

Though such equations can be analyzed exactly, we will focus on solving them numerically. Recall from your introductory course in differential equations that Euler's numerical method for solving a first order equation $x'(t) = f(t, x)$ is derived by approximating $x'(t)$ with a difference quotient:

$$x'(t) = \lim_{h \to 0} \frac{x(t+h) - x(t)}{h} \Rightarrow x'(t) \cong \frac{x(t+h) - x(t)}{h}, \quad h \,\text{small}.$$

[21] You should recall from high school algebra that a principal investment P, invested over t years at interest rate r yields a return $R_t = P(1+r)^t$. Hence, the equation I've solved here is $12131.9 = 300(1+r)^{78}$. Typically, the numbers you'll hear bandied around are more in the ballpark of 10% or 11%; for example, in his national bestseller, *A Random Walk Down Wall Street*, Burton G. Malkiel points out that between 1926 and 1994, the average growth on all common stocks has been 10.2%. The point here is simply that, historically speaking, stock behavior is not entirely random: these things are going up.

Euler's approximation becomes, then,

$$\frac{x(t+h)-x(t)}{h} = f(t,x) \Rightarrow x(t+h) = x(t) + hf(t,x),$$

an iterative equation ideal for computation. For (10.2) we note that dt plays the role of h and dB_t is the critical new issue. We have

$$X_{t+dt} = X_t + \mu(t, X_t)dt + \sigma(t, X_t)(B_{t+dt} - B_t).$$

Recalling that $B_{t+dt} - B_t \stackrel{d}{=} B_{dt}$, ($\stackrel{d}{=}$ means equal in probability, that the quantities on either side of the expression have the same probability density function), our algorithm for computation will be

$$X_{t+dt} = X_t + \mu(t, X_t)dt + \sigma(t, X_t)B_{dt},$$

where B_{dt} will be recomputed at each iteration. Assuming appropriate MATLAB function M-files *mu.m* and *sig.m* have been created (for μ and σ respectively), the following function M-file will take a time interval t and initial value z and solve (10.2) numerically.

```
function sdeplot(t,z);
%SDEPLOT: Employs Euler's method to solve and plot
%an SDE numerically. The drift is contained
%in mu.m, the diffusion in sig.m
%The variable t is time, and z is initial data
clear time;
clear X;
steps = 50; %Number of increments
dt = t/steps; %time increment
X(1) = z; %Set initial data
m = 1; %Initialization of vector entry
time(1) = 0;
for k=dt:dt:t
X(m+1) = X(m) + mu(k,X(m))*dt + sig(k,X(m))*brown(dt);
time(m+1) = time(m) + dt;
m = m + 1;
end
plot(time, X)
```

Carried out for the SDE $dX_t = .002X_t dt + .051 dB_t$, one path is shown in Figure 10.3.

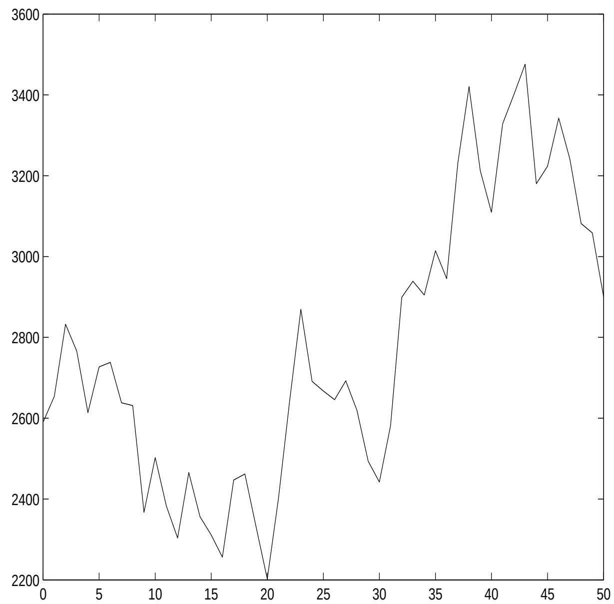


Figure 10.3: Geometic Brownian Motion

Index

References

[R] S. Ross, *A First Course in Probability*, 4th Ed. Macmillan College Publishing Co. 1994.

[S] P. Straffin, *Game Theory and Strategy*, The Mathematical Association of America 1993 (sixth printing 2006).