Essential Probability[1]

Michael Woodroofe

The University of Michigan

Contents

Chapter 1

Combinatorial Probability

1.1 The Classical Model

In its simplest form, the theory of probability may be motivated by games of chance such a card games and dice games. The resulting models are both simple and interesting. They provide a good introduction to the subject and have wider applicability. To understand the nature of the problems, consider the game of roulette.

Example 1.1 In this game a ball must fall into one of 38 boxes labelled 00, 0, 1,2,$\cdots$,36 of which 00 and 0 are green, 1,3,$\cdots$,35 are red, and 2,4,$\cdots$,36 are black. A player may bet even money on either a red or black outcome against the house. If the player bets on red, then there are thirty-eight possible outcomes, 00, 0,1,2,$\cdots$,36 of which eighteen, 1,3,$\cdots$,35, are favorable (red), and it seems natural to call the ratio $18/38 = 9/19$ the probability of winning. What does this number mean? To answer this question, it is necessary understand the assumptions that have been made. It was tacitly assumed that the thirty-eight outcomes are *equally-likely*; that is, that in many repetitions of the game, the various outcomes will occur with approximately the same frequency. The assumption seems reasonable here, if the wheel is balanced and the boxes of equal size. It then follows that if the player plays a large number of games, say N, and if W denotes the number of games that he/she wins, then the ratio W/N will approximate the probability of winning; that is $W/N \approx 9/19$. The

qualifiecation "approximately" is important here, since one should not expect exact equality. The accuracy of this approximation is assessed in Example ?.?. ♢

A Mathematical Model. With the example in mind, consider a game which must result in one of a finite number of possible outcomes which are equally likely. Denote the set of possible outcomes by Ω, so that $\Omega = \{00, 0, 1, 2, \cdots, 36\}$ in the example. Subsets $A \subseteq \Omega$, are called *events*, and an event is said to *occur* if the outcome is an element of it. For example, $A = \{1, 3, \cdots, 35\}$ is the event of a red outcome in the example. If A is any event, denote the number of outcomes in A by $\#A$. Then the probability of an event A is defined to be

$$P(A) = \frac{\#A}{\#\Omega} \tag{1.1}$$

for all events $A \subseteq \Omega$; that is, the probability of an event is the ratio of the number of favorable outcomes to the total number of outcomes. The model defined by (1) is called the *Classical Model*.

Example 1.2 If two balanced dice are tossed, what is the probability that the sum of spots is equal to seven? In this case an outcome may be represented by an ordered pair $\omega = (i, j)$, where i and j are integers between one and six representing the numbers of spots that appear on the two dices. Thus

$$\begin{aligned}\Omega = \{&(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\\ &(2,1),(2,2),(2,3),(2,4),(2,5),(2,6)\\ &(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)\\ &(4,1),(4,2),(4,3),(4,4),(4,5),(4,6)\\ &(5,1),(5,2),(5,3),(5,4),(5,5),(5,6)\\ &(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}\end{aligned},$$

and $\#\Omega = 36$, by inspection. The event that the sum of spots is seven is the off diagonal, $A = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$, and $\#A = 6$. So, $P(A) = 6/36 = 1/6$. *diamonduit*

As defined by (1.1), probability is a function defined on the *power set* of Ω, the set of all subsets of Ω. Of course, several operations may be performed on sets, like unions, intersection, and complementation. That is $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$, $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$, and $A' = \{\omega \in \Omega : \omega \notin A\}$ for events $A, B \subseteq \Omega$. Probability interacts with these operations and has the following obvious properties:

$$0 \leq P(A) \leq 1, \text{ for all } A \subseteq \Omega, \tag{1.2}$$

$$P(\Omega) = 1 \tag{1.3}$$

and

$$P(A \cup B) = P(A) + P(B), \text{ if } A \cap B = \emptyset, \tag{1.4}$$

where $\emptyset$ denotes the emptyset.

Some simple consequences of (1.2), (1.3), and (1.4) are needed below. First, if A' denotes the complement of an event A. Then $A \cup A' = \Omega$ and $A \cap A' = \emptyset$. So, $P(A) + P(A') = P(\Omega) = 1$, by (3) and (4), and

$$P(A') = 1 - P(A). \tag{1.5}$$

Events A and B are said to be *mutually exclusive* if $A \cap B = \emptyset$. Using (4) and mathematical induction, it is easy to show that if $A_1, \cdots, A_m$ are (pairwise) mutually exclusive, then

$$P(A_1 \cup \cdots \cup A_m) = P(A_1) + \cdots + P(A_m). \tag{1.6}$$

1.2 Combinatorics

Calculations within the classical model are conceptually straightforward: To compute the probability of an event A using Equation (1.1), one has only to count the number outcomes in A, count the number of outcomes in Ω, and divde, as in Examples 4.5 and 1.2. The actual counting may be complicated, however, especially if Ω is large. See Example 2 below. In order to use Equation (1.1) in complex situations, an effecient method of counting is needed. The efficient method of counting is called *Combinatorics* and is the subject of this section.

Lists and Permutations. Some things are easy to count. If Z is a non-empty set, then one may form an ordered pairs (x, y), where $x, y \in Z$. More generally, if n is a positive integer, then a *list of n-elements of Z* is an array $(z_1, \cdots, z_n)$ with $z_i \in Z$ for all $i = 1, \cdots, n$. The order in which elements are listed in important here, so that two lists, $(z_1, \cdots, z_n)$ and $(w_1, \cdots, w_m)$ say, are equal iff $m = n$ and $w_i = z_i$ for all $i = 1, \cdots, n$. A *permutation of n-elements of Z* is a special type of list in which no element appears more than once; that is, $z_i \neq z_j$ whenever $i \neq j$. For example, if $Z = \{1, 2, 3\}$, then there are nine ordered pairs (1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), (3,3) of which six are permuations, (1,2), (1,3), (2,1), (2,3), (3,1), (3,2).

Lists and permutations are easy to count, and many interesting calculations may be done using the following simple rule.

The Basic Principal of Combinatorics. *Suppose that objects, x and y say, are to be chosen from sets X and Y. If there are M choices for x and then N choices for y, then there are $M \times N$ choices for the ordered pair (x, y).*

More generally, suppose that objects z_i are to be chosen from sets Z_i, $i = 1, \cdots, n$. If there are N_i choices for z_i for $i = 1, \cdots, n$, then there are $N_1 \times N_2 \times \cdots \times N_n$ choices for the list $(z_1, \cdots, z_n)$.

In the Basic Principal, the set Y from which y is chosen may depend on x; only the number of choices N must be fixed in advance. For a small example, nine ordered pairs (x, y) may be chosen from $\{1, 2, 3\}$, since there are three choices for each of x and y; and there are six permutations (x, y) with $x \neq y$, since then there are three choices for x, but only two for y, which must be different from x Here is a larger example.

Example 1.3 In a certain state, automobile license plates consist of a list of two letters $(a, \cdots, z)$ followed by four digits $(0. \cdots, 9)$. Thus license plates may be identified with lists $(z_1, \cdots, z_6)$, where z_1 and z_2 are letters and z_3, z_4, z_5, and z_6 are digits.

a) There are $26^2 = 676$ ways to choose the two letters, since there are $N_i = 26$ choices for z_1 and z_2. Similarly, there are $10^4 = 10,000$ ways to choose the four digits, since there are 10 choices for each of $z_3, \cdots, z_6$. So, there are $26^2 \times 10^4 = 6,760,000$ distinct license plates.

b) There are $26 \times 25 \times 10 \times 9 \times 8 \times 7 = \cdots$ license plates with distinct letters and distinct digits. For there are 26 choices for z_1, but then only 25 for z_2, which must be different from z_1. Similarly, there are 10 choices for z_3, but then only 9 for z_4, 8 for z_5, and 7 for z_6. ♢

The derivation used in the example generalizes easily. *Let n and N are positive integers and let Z be a set with N elements. Then there are $N^n = n \times \cdots \times n$ lists $(z_1, \cdots, z_n)$ of n-elements of Z*, since there are N choices for z_i for all $i = 1, \cdots, n$. *If $n \leq N$, then there are $N \times (N-1) \times \cdots \times (N-n+1)$ permutations*, since there are N choices for z_1, $N-1$ for z_2, which must be different from z_1, etc.$\cdots$. It is convenient to have a notation for the latter product. Let

$$(N)_n := N \times (N-1) \times \cdots \times (N-n+1) \tag{1.7}$$

for positive integers N and n, and let $(N)_0 = 1$. The symbol $(N)_n$ defined in (**??**) is called a it descending product. Observe that $(N)_n = 0$ if $n > N$, since then one of the factors is zero. Equation (**??**) is called the *Permutations Formula*, because there are $(N)_n$ permutations of length n from a set of N elements. In the special case $N = n$, the product $(n)_n$ is denoted by $n!$ (read "n-factorial"); thus, $0! = 1$,

$$n! = n \times (n-1) \times \cdots \times 2 \times 1 \tag{1.8}$$

for positive integers $n = 1, 2, \cdots$, and there are $n!$ permutations of a set of n elements. The numbers $n!$ grow very rapidly. For example, $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$, and $10! = 3,628,800$. *Stirling's Formula* asserts that

$$n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}, \tag{1.9}$$

as $n \to \infty$, where $\sim$ means that the ratio of the two sides approaches one. It is accurate to within 1% for $n \geq 10$. A derivation is later.

Combinations Permutations may be contrasted with *combinations*. If Z is a non-empty set, then a subset $\{z_1, \cdots, z_n\}$ of n distinct elements of Z is called a *combination of n-elements of Z*. Like permutations, combinations must have distinct elements, but for combinations the order in which these elements are listed is not

important. That is, two combinations, $\{w_1, \cdots, w_m\}$ and $\{z_1, \cdots, z_n\}$ are equal iff $w_i \in \{z_1, \cdots, z_n\}$ for all $i = 1 \cdots, m$ and $z_i \in \{w_1, \cdots, w_m\}$ for all $i = 1, \cdots, n$. This implies $m = n$. For example, the two combinations $\{1, 2, 3\}$ and $\{3, 2, 1\}$ are the same, but the permutations $(1, 2, 3)$ and $(3, 2, 1)$ are different. In fact, six differenct permutations may be obtained from the combination $\{1, 2, 3, \}$.

If n and N is positive integers for which $n \leq N$ and if Z is a set with N elements, then there are $(N)_n/n!$ combinations of n-elements from Z. To see this observe that a permutation of n-elements of Z may be chosen in two step: first select a combination of n-elements; then arrange them in a definite order. Let C denote the number of combinations. Then, since there are $(N)_n$ permutations and $n!$ ways to arrange it in a definite order, $(N)_n = C \times n!$, by the Basic Principal. So, $C = (N)_n/n!$.

Example 1.4 a) If a committee of three is to be chosen from a group of nine people, then thre are $(9)_3/3! = 9 \times 8 \times 7/6 = 84$ possible choices, if that all members of the committee have equal status.

b) A menu at a resturant lists 30 entrees. If a large group decides to order 10 different ones, then there are $(30)_{10}/10! = 30,045,015$ possible choices.

c) A bridge hand consists of a combination of thirteen card from a standard deck. There are $(52)_{13}/13! = 635,013,559,600$ bridge hands. ♢

Binomial Coefficients. The numbers on the right side of are called *binomial coefficients* and denoted by

$$\binom{N}{n} = \frac{(N)_n}{n!} \tag{1.10}$$

for $n = 1, \cdots, N$. When $n = 0$, the right side of (5) is one, since $0! = 1$, and if $n > N$, then it is zero. Further, it is convenient to let $\binom{N}{n} = 0$ if $n < 0$. Then $\binom{N}{n}$ combinations of n elements may be formed from a set of N elements for all integer n and N. Equation(**??**) called the *Combinations Formula* below. Using the relation $N! = (N)_n \times (N - n)!$, the number of combinations may be written

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} \tag{1.11}$$

for $n = 0, \cdots, N$. The reader has undoubtedly encountered the binomial coefficients in the context of the *Binomial Theorem* which states that

$$(a+b)^N = \sum_{n=0}^{N} \binom{N}{n} a^n b^{n-n} \tag{1.12}$$

for real numbers a and b and non-negative integers N. In fact, the Binomial Theorem follows easily from the combinations formula. For if the product $(a+b)^N = (a+b) \times \cdots \times (a+b)$ is expanded, then $a^n b^{N-n}$ will appear as often as a may be chosen from n of the factors and b from the remaining $N-n$, and there $\binom{N}{n}$ such terms by the Combinations Formula.

Partitions. If Z is a set and r is a positive integer, then *a partition of Z into r subsets* is a list $(Z_1, \cdots, Z_r)$, where $Z_i,\ i = 1, \cdots, r$ are mutually exclusive sets whose union is Z; that is, $Z_i \cap Z_j = \emptyset$ whenever $i \neq j$ and $\cup_{i=1}^{r} Z_i = Z$. The numbers of elements $N_i = \#Z_i,\ i = 1, \cdots, r$ in the sets are called the *partition numbers*. Observe that the order in which the sets $Z_1, \cdots, Z_r$ are written is important, but the order in which elements are written within each Z_i is not.

For an example, suppose that six toys are to be divided among three children in such a manner that the youngest gets 3, the middle child gets 2, and the oldest gets 1. Then there are $\binom{6}{3} = 20$ ways to select three toys for the youngest. Then there are $\binom{3}{2} = 3$ ways to select two toys for the middle child from the remaining three toys, and then there is only one toy left for the oldest child. So, the number of possible partitions is $20 \times 3 \times 1 = 60$.

It is possible to count the number of partitions with given partition numbers quite generally. *If Z is a set with N elements, and if $N_1, \cdots, N_r$ are non-negative integers for which $N_1 + \cdots + N_r = N$, then there are*

$$\binom{N}{N_1, \cdots, N_r} := \frac{N!}{N_1! \times \cdots \times N_r!} \tag{1.13}$$

partitions $(Z_1, \cdots, Z_r)$ of Z for which $\#Z_i = N_i$ for all $i = 1, \cdots, r$. To see this observe that there are $\binom{N}{N_1}$ possible choices for Z_1, which may be any combination of size N_1 from Z. Then there are $\binom{N-N_1}{N_2}$ possible choices for Z_2, which may be any combination of size N_2 from $Z - N_1$, and so on. By the Basic Principal, the number

of partitions is the product

$$\begin{aligned}
&\binom{N}{N_1} \times \binom{N-N_1}{N_2} \times \cdots \times \binom{N-N_1-\cdots-N_{r-1}}{N_r} \\
&\quad = \frac{N!}{N_1!(N-N_1)!} \times \frac{(N-N_1)!}{N_2!(N-N_1-N_2)!} \times \cdots \times \frac{(N-N_1-\cdots-N_{r-1})!}{N_r!(N-N_1-\cdots-N_r)!} \\
&\quad = \frac{N!}{N_1! \times \cdots \times N_r!}.
\end{aligned}$$

Example 1.5 How many distinguishable configurations can be made from the letters in the word MISSISSIPPI? Label the positions of the eleven letter by $1, 2, \cdots, 11$. Then a distinguishable configuration consists of a partition of the eleven places into four subsets of sizes 4 for the I, 1 for the M, 2 for the P, and 4 for the S. Using (1.13), the answer is

$$\frac{11!}{4! \times 1! \times 2! \times 4!} = \cdots .$$

The symbol defined in (**??**) is called a *multinomial coefficient*, and the *Multinomial Theorem* asserts that

$$(a_1 + \cdots + a_r)^N = \sum \binom{N}{N_1, \cdots, N_r} a_1^{N_1} \times \cdots \times a_r^{N_r}$$

for real $a_1, \cdots, a_r$, where the summation extends over all non-negative integers $N_1, \cdots, N_r$ for which $N_1 + \cdots + N_r = N$.

1.3 Sampling Experiments

The combinatorial formulas from the previous section may be combined with Equation (1.1) to compute many interesting probabilities. In this section, they are used to study sampling experiments. These are experiments in which a smaller group of objects, called the sample, is selected from a larger group, called the population, and examined in some way. The object of the exercise may be to learn about the population, or simply to select a sample fairly. Here are two examples to illustrate the nature of the calculations.

Example 1.6 *a).* If a committee of size five is selected at random from a group of nine Democrats and six Republicans, what is the probability that the Committee

consists of three Deomcrats and two Republicans? In this case, and outcome to the experiment is a combination of five of the nine fifteen. So, there are $\#\Omega = \binom{15}{5} = 3003$ possible outcomes. Let A the the event that there are three Democrats and two Republicans on the committee. Then $\#A = \binom{9}{3} \times \binom{6}{2} = 84 \times 15 = 1260$, since there are $\binom{9}{3}$ ways to select two of the six Democrats, and $\binom{6}{2}$ ways to select two of the six Republicans. So, the desired probability is

$$P(A) = \frac{\binom{9}{3}\binom{6}{2}}{\binom{15}{5}} = \frac{1260}{3003} = .4196,$$

by (**??**).

b). If the group consists of six Democrats, three Independents, and six Republicans, what is the probability that the committee consists of two Democrats, one Independent, and two Republicans. Let B be the latter event. Then $\#B = \binom{6}{2} \times \binom{3}{1} \times \binom{6}{2} = 675$, since there are $\binom{6}{2}$ ways to select two Democrats, $\binom{3}{1}$ ways to select an Independent, and $\binom{6}{2}$ ways to select two Republicans. So,

$$P(B) = \frac{\binom{6}{2}^2\binom{3}{1}}{\binom{15}{5}} = .2248.$$

Box-Ticket Models for Sampling. In the last two examples, a smaller group is selected from a larger group. There are many examples of this nature, and it is useful to have some uniform terminology to discuss them. The larger group may be pictured as a set of tickets in a box–for example, movie tickets in a shoe box. The box is given a vigorous shake and some of the tickets are removed. The set of tickets in the box is called the *population*, and those selected the *sample*. Several types of samples may be identified. Let N denote the population size, the number of tickets in the box, and let n denote the sample size.

Unordered Samples. If the tickets are drawn all at once, or if the order in which they are drawn is unimportant, then an outcome may be described by a combination of n of the N tickets, and the sample space Ω consists of all such combinations. Thus there are $\#\Omega = \binom{N}{n}$ possible outcomes. Example 1 is of this nature, if the fifteen people are identified with tickets in a box.

Ordered Samples. Alternatively, the tickets may be drawn one at a time, and the order in which they are drawn recorded. In this case, an outcome consists of a list of

the n tickets drawn. For ordered samples, there is a further distinction to be drawn. The tickets may be replaced after each drawing, or not.

Ordered Sampling With Replacement. If the tickets are replaced after each drawing, then all lists of n tickets are possible outcomes. In this case, the sample space Ω consists of all lists of n of the N tickets, and $\#\Omega = N^n$.

Ordered Sampling Without Replacement. If $n \leq N$ and the tickets are not replaced after each drawing, then only lists of distinct tickets are possible outcomes. So, the sample space Ω consists of all permutations of n of the N tickets in this case, and $\#\Omega = (N)_n$.

Interesting mathematical models arise when all samples of a given size and type are equally likely to be drawn. Then the sample is said to be a *simple random sample*, and the sampling is said to be *at random*. To justify this assumption, there should be an explicit act of randomization, like shaking the box, and the box should be shaken after each drawn in the case of sampling with replacement. Further examples conclude this section. The first is a famous one with a surprizing answer.

Example 1.7 *The Birthday Problem*

If twenty-five people gather at a party, what is the probability that at least two of them have the same birthday. The answer is $.56\cdots$, not $25/365$. To see why suppose that n people gather at the party. Order them by order of arrival, say, and regard the birthdays of the n people as a simple random sample with replacement from the 365 day of the year, ignoring leap years and effects like snowstorms. Then the sample space Ω consists of all lists of length n from the days of the year, and $\#\Omega = 365^n$. Let A be the event that at least two people have the same birthday. In this example, it is easier to compute the probability of the complement A', and the probability of A may be recovered from (1.5). Here A' is the event that at no people have the same birthday and, therefore consists of all permutations of length n from the 365 days of the year. So, $\#A' = (365)_n$, by the Permutations Formula, $P(A') = (365)_n/365^n$, and

$$P(A) = 1 - P(A') = 1 - \frac{(365)_n}{365^n},$$

by (1.5). For $n = 25,\ P(A) = .5687$. $\diamondsuit$

Binomial and Hypergeometric Probabilities. If a simple random sample of size n is drawn from a box that contains R red tickets and $N-R$ white tickets, what is the probability that the sample contains exactly r red tickets? The answer depends on the type of sample drawn (unsurprisingly). For unordered sampling, the derivation is similar to that of Example 1.6(a): The sample space Ω consists of all combinations of n of the N tickets in the box, and $\#\Omega = \binom{N}{n}$. Let A_r be the event that the sample contains exactly r red tickets. Then elements of A_r consist of combination of r red tickets joined with $n-r$ white tickets. The r red tickets may be chosen in $\binom{R}{r}$ ways and the $n-r$ white tickets in $\binom{N-R}{n-r}$ ways. So, $\#e_r = \binom{R}{r}\binom{N-R}{n-r}$, using the basic principle , and

$$P(e_r) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}} \tag{1.14}$$

The terms on the right side of (1.14) are called the *hypergeometric probabilities*. Example 1.6(a) provides a numerical illustration with $N = 15$, $R = 9$, and $n = 5$.

For ordered samples with replacement, the sample space Ω consists of all lists $\omega = (\omega_1, \cdots, \omega_n)$, and $\#\Omega = N^n$. again, let E_r denote the event that exactly r red tickets are drawn. Then $P(E_r) = \#E_r/\#\Omega$, and $\#E_r$ is required. An element of $E_r$maybe chosen in steps. Fir select a combination $C = \{I_1, \cdots, I_r\}$ of r of the indices from $\{1, \cdots, n\}$. Then draw red tickets on draws labelled $i \in C$ and white on the other draws. The combination may be chosen in $\binom{n}{r}$ ways, by the combinations formula. Then the tickets can be drawn in $R^r \times (N-R)^{n-r}$ ways, since there are r choices for ω_i when $i \in C$ and $N = R$ when $i \notin C$. So,

$$P(E_r) = \binom{n}{r}\frac{R^r \times (N-R)^{n-r}}{N^n} \tag{1.15}$$

Let $p = R/N$. Then (**??** may be writted as

$$P(E_r) = \binom{n}{r}p^r(1-p)^{n=}, \tag{1.16}$$

The terms on the right side of *1.16) are called *the Binomial probabilities* with parameters n and p.

Forordered samples without replacement a similar argument shows that

$$P(E_r) = \binom{n}{r}\frac{R_r N - R)_{n-r}}{(N)_n}, \tag{1.17}$$

and simple algebra shows that

$$\binom{n}{r}\frac{R_r N-R)_{n-r}}{(N)_n}=\frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}}, \tag{1.18}$$

the same answer obtained for unordered samples.

Example 1.8 *Sampling Inspection*

Suppose that a manufacturer markets its product in lots of size N, that each lot may contain and unknown number, R say, of defective items, and that a lot is regarded as acceptable if it does not contain too may defectives, say $R \leq R_0$. Suppose further that testing is expensive. Then a potential customer might wish to test only some of the items in a lot before deciding whether to buy it. Suppose that the customer decides to test a n of the items, selected at random, and to accept the lot iff the number of defectives in the sample is at most r_0. Then the probability that a lot is accepted is

$$a=\sum_{k=0}^{r_0}\binom{n}{k}\frac{(R)_k(N-R)_{n-k}}{(N)_n}.$$

Here R is unknown. If $R > R_0$, then a is called the *consumer's risk*, since it is the probability of accepting a bad lot. Similarly, if $R \leq R_0$, then $1-a$ is the probability of rejecting a good lot and is called the *producer's risk*. The values of n and r may be chosen to control these risks. Figure 1 shows a graph of a for the case in which $N = 100,\ R = 10$, $n = 25$, and r.

Example 1.9 *Capture-Recapture Estimation*

Consider the problem of estimating the size of an animal population–for example, the number of fish in a lake. One popular method makes essential use of the hypergeometric probabilities. Suppose that there are an unknown number N of fish in the lake. Suppose also that R of these are caught, tagged, and returned to the lake. Later a second batch of n fish are caught. If the second batch is regarded as a random sample, then the probability that it includes exactly r of the tagged fish is given by (2), since the tagged fish may be regarded as red tickets. This is denoted by $q_r(N)$, since the dependence on N is important. Thus,

$$q_r(N)=\binom{n}{r}\frac{(R)_r(N-r)_{n-r}}{(N)_n}. \tag{4}$$

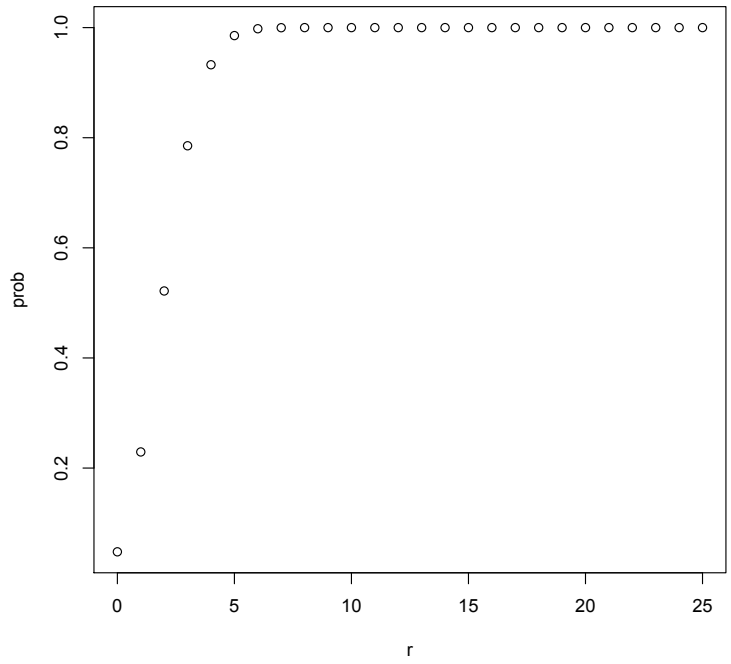

Figure 1.1: The probability of obtaining r or fewer defective in a sample of size 25 from a lot containing 100 items of which 10 are defective.

In (4), $n,\ r,$, and R are known, but N is not. To estimate N, it seems reasonable to take the value which maximizes $q_r(N)$. The maximizing value may be shown to be an integer adjacent to

$$\hat{N} = \frac{nR}{r},$$

which serves as the estimator for N. For example, if $R = 100$ fish are tagged, $n = 100$ are caught in the second batch, and $r = 25$ of the fish in the second batch had been tagged, then the estimated number of fish in the lake is $\hat{N} = 400$.

To see that $q_r(N)$ is maximized at $\hat{N}$, suppose that $r \geq 1$ and consider the ratios

$$\frac{q_r(N)}{q_r(N-1)} = \frac{(N-R)_{n-r}(N-1)_n}{(N-R-1)_{n-r}(N)_n} = \frac{(N-R)(N-n)}{[N-R-(n-r)]N}.$$

The ratio is bigger than one iff $(N-R)(N-n) > [N-R-(n-r)]N$ or, equivalently, $(n-r)N > (N-R)n$. This in turn is equivalent to $rN < Rn$ or $N < \hat{N}$. Thus,

$q_r(N) >$ or $\leq q_r(N-1)$ accordingly as $N <$ or $\geq \hat{N}$ and, therefore, $q_r(N)$ is maximized when N is an integer adjacent to $\hat{N}$. $\diamondsuit$

1.4 Problems and Complements

Problems

1. How many four letter words can be formed if every list of four letters is a word; if each word must contain at least one vowel (a,e,i,o, or u)?

Ans: $456,976$ and $262,495$.

2. How many four letter words may be formed with distinct letters; distinct letters and at least one vowel?

3. How many committees of size five can be formed from 15 people if all members of the committee have the same status; if there is a chair and four others of equal status; if there is a chair, a secretary, and three others of equal status?

Ans: 3003; $15,015$; and $60,060$

4. How many ways can 10 students be divided into two teams of size 5 each if each team has a captain and four others of equal status and the teams are identified only by the names of their captains?

5. A child has two indistinguishable red blocks, three indistinguishable white blocks, and four indistinguishable blue blocks. If he/she arranges them in a row, how many distinguishable configurations can be made?

Ans: 1260

6. In how many ways can 10 one dollar coins be divided among five children if each child is to get at least one dollar?

7. If two balanced dice are rolled, what is the probability that the sum of spots is equal to five? Determine the sample space; identify the event in question as a subset of the sample space; and compute its probablity.

Ans: 1/9

8. In the previous problem, what is the probability that the absolute difference between the numbers of spots on the two dice (larger less smaller) is equal to 1?

9. What is the probability that a poker hand contains a pair (least two cards of the same denomination–i.e. at least two aces, or two two's, or $\cdots$)?

Ans: .4929

10. A drawer contains four pairs of socks–for example, a red pair, a blue pair, a black pair, and white pair? If four socks are selected at random, what is the probability that the four socks include at least on pair?

Complements

1. Supply the simple algebra for (1.18).

2. Supply the similar argument leading to (1.17).

3. Show that $\sum_{k=1}^{n} \binom{n}{k} (=1)^{k-1} = 1$ for any $n \geq 1$

Chapter 2

Axiomatic Probability

2.1 Probabily Models

With its insistence on a finite number of equally likely outcomes, the classical model presented in the previous chapter is not sufficiently flexible accomodate many interesting applications. For example, in many sampling experiments, members of the population are of different sizes, and it is natural and/or desirable to select larger items with higher probability than small ones. In other examples, there are infinitely many possible outcomes. A more general and flexible model is presented in this section. The more general model starts with axioms and allows multiple interpretations of results derived from it.

The term *experiment* is used to describe an activity or phenomena for which the outcome is unpredictable, but the set of possible outcomes can be specified. Examples include observing the number of traffic accidents at a given point during a given time interval, observing stock prices over a given time period, and observing responses to an experimental medical treatment, in addition to the games and sampling experiments of the previous chapter. As in Chapter 1, the set of all possible outcomes is denoted by Ω and called the *sample space*, and *events* are subsets of Ω.

A novel feature of the general model is that not all subsets of Ω are required to be events. Rather there is a distinguished class $\mathcal{A}$ of subsets of Ω, and only those subsets $A \subseteq \Omega$ which are members of $\mathcal{A}$ are called events. As in Chapter 1, set

theoretic operations like union, intersection, and complementation may be performed on events, and the class of event will be required to closed under these operations. Initially, it is required that $\Omega \in \mathcal{A}$, that $A' \in \mathcal{A}$ whenever $A \in \mathcal{A}$, and that $A \cup B \in \mathcal{A}$ whenever $A, B \in \mathcal{A}$. Such a class is called a *(Boolean) algebra* of subsets of Ω. Clearly, the class of all subsets of Ω satisfies the conditions and is, therefore, and Boolean algebra. A less transparent examples is given below.

The third element of the model is a *probability function*. Probability is a function P from the algebra $\mathcal{A}$ of events into the real numbers $\mathcal{R}$, so that $\mathcal{A}$ is the domain of P, and P is required to satisfy the following conditions:

$$P(\Omega) = 1, \tag{2.1}$$

$$0 \leq P(A) \leq 1, \tag{2.2}$$

for all $A \in \mathcal{A}$, and

$$P(A \cup B) = P(A) + P(B), \tag{2.3}$$

whenever A and B are mutually exclusive events ($A \cap B = \emptyset$, where $\emptyset$ denotes the emptyset). Here (2.1), (2.2), and (2.3) are not a definition of probability but only a limitation on possible definitions. A whole class of example which satisfy (2.1), (2.2), and (2.3)) is described next.

Discrete Spaces. Let Ω denote a finite or countably infinite set[1] say $\Omega = \{\omega_1, \omega_2 \cdots\}$; let $\mathcal{A}$ denote the class of all subsets of Ω; and let $p_1, p_2, \cdots$ denote non-negative real numbers for which $p_1 + p_2 + \cdots = 1$. Then a function P may bedefined by

$$P(A) = \sum_{i:\omega_i \in A} p_i, \tag{2.4}$$

for $A \subseteq \Omega$, where the summation extends over those i for which $\omega_i \in A$; and P satisfies (2.1), (2.2), and (2.3 For example, to establish (2.1) and (2.2) observe that $P(\Omega) = p_1 + p_2 + \cdots = 1$, by assumption, and $0 \leq P(A) \leq P(\Omega)$, since $p_1, p_2, \cdots$ are non-negative. Condition (3) is equally easy. The p_i may be recovered from P by

[1] A set is said to be countably infinite it there is a one-to-one correspondence between the set and the positive integers $\{1, 2, 3, \cdots\}$.

$p_j = P(\{\omega_j\})$ for $j = 1, 2, \cdots$, where $\{x\}$ denotes the set whose only element is x. Thus, if P is deined by (4), then p_j is the probability that the outcome will be ω_j.

Example 2.1 *:Some Discrete Spaces. a): Classical Models.* If Ω is a finite sets, say $\Omega = \{\omega_1, \cdots, \omega_M\}$ and $p_i = 1/M$ for all $i = 1, \cdots, M$, then (4) becomes

$$P(A) = \frac{1}{M} \times \#A = \frac{\#A}{\#\Omega}$$

for all $A \subseteq \Omega$, and the Classical Model of Chapter 1 is recovered.

b): Sampling Proportional to Size. Suppose that $\Omega = \{\omega_1, \cdots, \omega_M\}$, where $\omega_i > 0$ for $i = 1, \cdots, M$, and let $p_i = \omega_i/s$ for $i = 1, \cdots, M$, where $s = \omega_1 + \cdots + \omega_M$. Then $p_1 + \cdots + p_M = 1$, and a probabilty function may be defined by (4). This model arises in accounting, for example. Faced with a several invoices, an accountant may wish to examine only one; he/she may also want to select larger invoices with higher probability. For a numerical example, suppose that $M = 10$ and $\omega_i = i$ for $i = 1, \cdots, 10$, in which case $s = 55$. To illustrate the use of (4), the the probability of selecting one of the three largest invoices is almost $1/2$, since

$$P(\{8, 9, 10\}) = \frac{8}{55} + \frac{9}{55} + \frac{10}{55} = \frac{27}{55}.$$

Of course, the accountant may wish to examine more than one invoice, if M is large. See Problem ?.?.

In the next example, use is made of the geometric series,

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}, \ -1 < x < 1. \tag{2.5}$$

This is a special case of the Generalized Binomial Theorem with $\alpha = 1$, though simpler derivations are possible. See Problem ?.?.

Example 2.2 *:An Infinite Sample Space.* Suppose that a balanced coin is tossed until a head appears, and the number of required tosses is recorded. Then $\Omega = \{1, 2, \cdots\}$, an infinite set. Now the first head appears on the n^{th} toss iff there are $n-1$ tails followed by a head. Intuitively, the probability of this is $(1/2)^n = (1/2) \times \cdots \times (1/2)$. This suggests letting $p_k = (1/2)^k$ for $k = 1, 2, \cdots$. It is easily seen from

(2.5) that $p_1 + p_2 + \cdots = 1$. So, a probability function may be defined by (eq:dscrt). To illustrate the use of (eq:dscrt), consider the event $\{1, 3, \cdots\}$ that an odd number of tosses is required. By (4),

$$P(\{1,3,\cdots\}) = \sum_{k=0}^{\infty} (\frac{1}{2})^{2k+1} = \frac{1}{2} \sum_{k=0}^{\infty} (\frac{1}{2})^{2k} = \frac{1}{2} \times \frac{1}{1-\frac{1}{4}} = \frac{2}{3}.$$

A function P which satisfies (1), (2), and (3) is called a *probability content*. These axioms work splendidly if Ω is a finite set but do not lead to a sufficiently rich mathematical theory if Ω is an infinite set, and a stronger version of (2.3) is needed. The stronger version concerns infinite unions of the form $\cup_{k=1}^{\infty} A_k = \{\omega \in \Omega : \omega \in A_k \text{ for some } k = 1, 2, \cdots\}$. A function P defined on a class of $\mathcal{A}$ of events is called a *probability measure* if it satisfies (2.1), (2.2) and

$$P(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} P(A_k), \tag{2.6}$$

whenever $A_1, A_2, \cdots$ are (pairwise) mutually exclusive events whose union is also an event. It may be shown (2.6) implies (2.3) and that any P of the form (2.4) satisfies (2.6). See $\cdots$. The algebra $\mathcal{A}$ of events is called a *sigma-algebra* if it is closed under the formation of such unions; that is, $\cup_{k=1}^{\infty} A_k \in \mathcal{A}$ whenever $A_1, A_2, \cdots \in \mathcal{A}$.

A *probability space* consists of a triple $(\Omega, \mathcal{A}, P)$, where Ω is a non-empty set, called the *sample space*, $\mathcal{A}$ is sigma-algebra, called the class of *events*, and P is a probability measure, called the *probability function*. This is the general model for probability.

The next example illustrates the need to restrict the class of events. The sample space is an interval, and the following notation for intervals is employed: if $-\infty \leq a \leq b \leq \infty$, then

$$(a,b) = \{x \in \mathbb{R} : a < x < b\}$$
$$(a,b] = \{x \in \mathbb{R} : a < x \leq b\}$$
$$[a,b) = \{x \in \mathbb{R} : a \leq x < b\}$$
$$[a,b] = \{x \in \mathbb{R} : a \leq x \leq b\},$$

where $\mathbb{R}$ denotes the set of real numbers.

Example 2.3 . An arrow is free to rotate about an axle, as in Figure 1. The arrow is spun and comes to rest, and the angle (in radians) which it makes with a fixed

direction, say ω, is recorded. Then the set of possible outcomes is $\Omega = (-\pi, \pi] = \{\omega : -\pi < \omega \leq \pi\}$. Intuitively, the probability that the outcome falls in a subset of Ω should be proportional to its length. This suggests letting

$$P((a, b]) = \frac{b - a}{2\pi} \tag{2.7}$$

for all $-\pi < a \leq b \leq \pi$. The problem here is the class of events, since the class of intervals is not closed under the formation of unions–for example, $(-1, 0] \cup (1, 2]$ is not an interval. The resolution of this difficulty is described next.

Absolutely Continuous Spaces. Let Ω be an interval and let f be a Riemann integrable function for which $f(\omega) \geq 0$ for all $\omega \in \Omega$ and

$$\int_{\Omega} f(\omega) d\omega = 1. \tag{2.8}$$

For example, $\Omega = (-\pi, \pi]$ and $f(\omega) = 1/2\pi$ for $\omega \in \Omega$. *Then there is a sigma-algebra $\mathcal{A}$, which contains all subintervals of Ω, and a probability measure P defined on $\mathcal{A}$ for which*

$$P(I) = \int_{I} f(\omega) d\omega \tag{2.9}$$

for all subintervals $I \subseteq \Omega$. This is the basic result for specifying probability models in which the sample space is a continuum, like and interval. For example, it shows the existence of a probability measure P for which (6) holds in Example 2.3

Here is an amusing consequence of (8). If $\omega \in \Omega$, then $\{\omega\} = [\omega, \omega]$ is an interval, and $P(\{\omega\}) = \int_{\omega}^{\omega} f(\theta) d\theta = 0$. That is, every $\omega \in \Omega$ has zero probability. Next, if Ω_0 is any countably infinite set, so that $\Omega_0 = \{\omega_1, \omega_2, \cdots\}$, then $\Omega_0 = \cup_{k=1}^{\infty} \{\omega_k\}$ and, therefore,

$$P(\Omega_0) = \sum_{k=1}^{\infty} P(\{\omega_k\}) = 0. \tag{2.10}$$

In particular, the probability of a rational outcome is zero, since the set of rational numbers is countably infinite.

The existence of P is in (8) called the *Extension Theorem*. The idea behind it proof is quite simple. First P is defined on a class of subintervals by (8). Then it is extended to a larger class using the axioms of probability (1),(2),(3), and (3').

The details of the extension process are complicated, however. They are discussed in Section 2.6. Unfortunately, there is no simple description for a typical event in the Extension Theorem. One knows only that all intervals are events and that the class of events $\mathcal{A}$ is closed under complementation and the formation of (countable) unions, so that any set that can be constructed from intervals by the operations is an event. Equation (9) provieds an example.

2.2 Consequences of the Axioms

Some simple consequences of the axioms (1.1),(1.2), and (1.3') are detailed in this section. Throughout $(\Omega, \mathcal{A}, P)$ denote a probability space.

Complements. The *difference* between two events, A and B say, is defined by $B - A = A' \cap B$, the event that B occurs but A does not. If one event is a subset of another, say $A \subseteq B$, then A is said to imply B, since the outcome must be in B whenever it is in A. In this case, B may be written in the form $B = A \cup (B - A)$, and $A \cap (B - A) \subseteq A \cap A' = \emptyset$. So, $P(B) = P(A) + P(B - A)$, by (1.3), and, therefore,

$$P(B - A) = P(B) - P(A), \tag{2.11}$$

whenever $A \subseteq B$. There are several corollaries to this simple relation. First,

$$P(A) \leq P(B), \tag{2.12}$$

whenever $A \subseteq B$, since $P(B - A) \geq 0$, by (2.1). Next, letting $B = \Omega$, leads to $B - A = \Omega - A = A'$ and $P(\Omega) = 1$; that is,

$$P(A') = 1 - P(A) \tag{2.13}$$

for all event A. Finally, letting $A = \Omega$ and observing that $\Omega' = \emptyset$,

$$P(\emptyset) = 0. \tag{2.14}$$

The use of (3) was illustrated in Example 1.3.1, the Birthday Problem, and Example 1.4.2, Sampling Inspection.

Example 1. $\cdots$

Unions. It is possible to use (1) to refine (1.3) by showing

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{2.15}$$

for any two events A and B. To see this first observe that $A \cup B = A \cup (B - A \cap B)$. Here A and $B - A \cap B$ are mutually exclusive, since $A \cap (B - A \cap B] \subseteq A \cap A' = \emptyset$. So, $P(A \cup B) = P(A) + P(B - A \cap B)$, by (2.3), and $P(B - A \cap B) = P(B) - P(A \cap B)$, by (**??**eq:cnsq1). Equation **??**) follows. The *symmetric difference* between two events, A and B, is defined to be $A \Delta B = A' \cap B \cup A \cap B'$, the event that one of A or B occurs, but the other doesn't. Equivalently, $A \Delta B = A \cup B - A \cap B$. So, $P(A \Delta B) = P(A \cup B) - P(A \cap B)$, by 2.11), since $A \cap B \subset A \cup B$ and, therefore,

$$P(A \Delta B) = P(A) + P(B) - 2P(A \cap B). \tag{2.16}$$

Equations (**??**) and (2.16) are the starting point for the Inclusion-Exclusion Formulas, discussed in Section 2.4.

Example 2.4 All students at Classical University must take either Greek or Latin. If 75% take Greek and 55% take Latin, how many take both. How many take only one? Let A be the event that a randomly selected student takes Greek, and let B be the event that he/she takes Latin. Then $P(A) = .75,\ P(B) = .55$, and $P(A \cup B) = 1$. So, $P(A \cap B) = P(A) + P(B) - 1 = .30$. Thus, 30% of the students take both, and 70% take just one. $\diamondsuit$

If $A_1, \cdots, A_m$ are mutually exclusive events, then

$$P(A_1 \cup \cdots \cup A_m) = P(A_1) + \cdots + P(A_m), \tag{2.17}$$

either by (2.3) and mathematical induction or by setting $A_i = \emptyset$ for all $i > m$ in (2.6). If $A_1, \cdots, A_m$ are any m events, not necessarily mutually exclusive,

$$P(A_1 \cup \cdots \cup A_m) \leq P(A_1) + \cdots + P(A_m), \tag{2.18}$$

To see this, let $B_1 = A_1$ and $B_k = A_k - (A_1 \cup \cdots \cup A_{k-1})$ for $k = 2, \cdots, m$. Then $B_1, \cdots, B_m$ are mutually exclusive. For if $j < k$, then $B_k \subseteq B_j'$ and, therefore,

$B_j \cap B_k = 0$. Moreover, $B_1 \cup \cdots \cup B_m = A_1 \cup \cdots \cup A_m$. For if $\omega \in A_1 \cup \cdots \cup A_m$, then there is a smallest k for which $\omega \in A_k$, and then $\omega \in B_k$. It follows that $P(A_1 \cup \cdots \cup A_m) = P(B_1 \cup \cdots \cup B_m) = P(B_1) + \cdots + P(B_m) \leq P(A_1) + \cdots + P(A_m)$, where the last two steps follows from (**??**) and (2.12). Relation (2.18) follows. Relation (2.18) is known as *Boole's Inequality*. The infinite version of Boole's Inequality, in which $A_1, \cdots, A_m$ is replaced by an infinite sequence is also true; that is,

$$P(\cup_{k=1}^{\infty} A_k) \leq \sum_{k=1}^{\infty} P(A_k) \tag{2.19}$$

for every infinite sequence $A_1, A_2, \cdots$. The proof of (2.19) is virtually identical to that of (**??**), but uses (2.6) in place of (2.17).

Example 2.5 In bridge, a void is the absence of one of the four suits. What is the probability of a void. Let A (respectively, B, C, C) be the event that the hand contains no spades (respectively, hearts, diamonds, clubs), so that $A \cup B \cup C \cup D$ is the event that the hand contains a void. Then $P(A) = \binom{39}{13} / \binom{52}{13}$, since a hand with no spades must be chosen from 39 non-spades. Similarly, $P(B) = P(C) = P(D) = \binom{39}{13} / \binom{52}{13}$. So,

$$\begin{aligned} P(A \cup B \cup C \cup D) &\leq P(A) + P(B) + P(C) + P(D) \\ &= 4\binom{39}{13} / \binom{52}{13} = \cdots . \end{aligned}$$

In this case the upper bound is quite close the exact answer, $\cdots$, which is derived in Problem 2.?.

Monotone Sequences. So far, only (2.19) used the stronger axiom (2.6) in an essential way. The next result makes additional use of (2.6). The result is known as he *Monotone Sqequnces Theorem* A sequence $B_1, B_2, \cdots$ is said to be *increasing* if $B_1 \subseteq B_2 \subseteq B_3 \subseteq \cdots$, and *decreasing* if $B_1 \supseteq B_2 \supseteq B_3 \supseteq \cdots$. Examples of such sequences are easy to find. If $A_1, A_2, \cdots$, is any sequence of events, then $B_n = A_1 \cup \cdots \cup A_n$, $n = 1, 2, \cdots$ defines an increasing sequence, and $C_n = A_1 \cap \cdots \cap A_n$, $n = 1, 2, \cdots$ defines a decreasing sequence. A sequence is said to be *monotone* if it is either increasing

or decreasing. The following result is known as he *Monotone Sqequnces Theorem*: *If* $B_1, B_2, \cdots$ *is increasing, then*

$$P(\bigcup_{k=1}^{\infty} B_k) = \lim_{n\to\infty} P(B_n); \tag{2.20}$$

and if $C_1, C_2, \cdots$ *is decreasing, then*

$$P(\bigcap_{k=1}^{\infty} C_k) = \lim_{n\to\infty} P(C_n) \tag{2.21}$$

To establish (2.20), let $A_1 = B_1$ and let $A_k = B_k - B_{k-1}$ for $k = 2, 3, \cdots$. Then $A_1, A_2, \cdots$ are mutually exclusive; for if $j < k$, then $A_j \subseteq B_j$ and $B_k \subseteq B_j^c$. Moreover, $\cup_{k=1}^{\infty} A_k = \cup_{k=1}^{\infty} B_k$; for if $\omega \in \cup_{k=1}^{\infty} B_k$, then there is a smallest k for which $\omega \in B_k$ in which case $\omega \in A_k$. It follows that

$$P(\cup_{k=1}^{\infty} B_k) = \cup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} P(A_k).$$

But

$$\sum_{k=1}^{\infty} P(A_k) = \lim_{n\to\infty} \sum_{k=1}^{n} P(A_k) = \lim_{n\to\infty} P(B_n).$$

This establishes (2.20), and (2.21) then follows by taking complements. For if $C_1, C_2, \cdots$ is decreasing, then $B_n := C_n^c,\ n \geq 1$ is increasing and $(\cap_{k=1}^{\infty} C_k)' = \cup_{k=1}^{\infty} B_k$, so that

$$\begin{aligned} P(\bigcap_{k=1}^{\infty} C_k) &= 1 - P(\bigcup_{k=1}^{\infty} B_k) \\ &= 1 - \lim_{n\to\infty} P(B_n) = \lim_{n\to\infty} P(C_n). \end{aligned}$$

The following simple corollary t (2.20), and (2.21) is useful: if $A_1, A_2, \cdots$ is any sequence of events, then

$$P(\bigcup_{k=1}^{\infty} A_k) = \lim_{n\to\infty} P(\bigcup_{k=1}^{n} A_k) \tag{2.22}$$

and

$$P(\bigcap_{k=1}^{\infty} A_k) = \lim_{n\to\infty} P(\bigcap_{k=1}^{n} A_k). \tag{2.23}$$

For example, (2.22) follows directly by applying (2.20) to the sequence $B_n = \cup_{k=1}^{n} A_k,\ n \geq 1$.

2.3 Significance Tests

It is best to begin with an example.

Example 2.6 *: Proving Unfairness.*There are twenty-five registered voters in Small County USA of whom thirteen are Hatfields and twelve are Mc Coys. A jury was recently selected to adjudicate a dispute between the two families. It consisted of ten Hatfields and ony two McCoys. By law, juries are supposed to be selected at random from the list of registered voters. The county clerk, a Hatfield, says that she of course followed the law. "Nonsene," reply the Mccoys, "That's incredible:"A random selection would not produce such a skewed jury"." Who's right? Is the county clerks' assertion credible? The short answer is "No." To see why, let P_0 denote the probability function under which all $\binom{25}{12}$ juries of size twelve are equally likely, and let T denote the number of Hatfields on a jury. Then

$$P_0[T = t] = \frac{\binom{13}{t}\binom{12}{12-t}}{\binom{25}{12}}$$

for $t = 0, 1, 2, \cdots 12.$, and

$$P_0[T \geq 10] = P_0[T = 10] + P_0[T = 11] + P_0[T = 12] = 000381 \cdots \tag{2.24}$$

That is less than four in ten-thousand random juries would have ten or more Hatfields. So, the clerk's claim of random selection requires believing that an even to probabiliy less than .0004 has occurred. The latter is so small that the possibility can safely be neglected. The observed value T=10 is not consistent with random selection. $\diamondsuit$

To formalize the calculation in the example suppose consider a probability model $(\Omega, \mathcal{A}, P_0)$ and a function $T : \Omega \to \mathbb{R}$ of which large values are inconsistent with P_0. Let H_0 denote the *hypothesis* that the data were drawn from P_0; that is, that $(\Omega, \mathcal{A}, P_0)$ is the appropriate model. Suppose next that the result of an experiment ω_{obs} has been observed and let $t_{\text{obs}} = T(\omega_{\text{obs}})$ denote the observed value of T. Further, let

$$p^* = P_0[T \geq t_{\text{obs}}] \tag{2.25}$$

Then p^* is called the *p value*, and small values of p^* are regarded as evidence against H_0. How small does p^* have to be before the outcome is declared to be inconsistent with H_0? In the medical and social sciences, $p^* < .05$ is generally regarded as inconistent with H_0. In Physics $p^* < .01$ is often demanded to claim inconsistency.

Example 2.7 *: A Clinical Trial.* As background the FDA (theFood and Drug Administration) requires that a treatment be shown to be safe and effective before it can be markreted. In a study of an experimental treatment for pain relief, twent-five patients were asked to score their level of pain both before and after receiving the treatment. For a given a patient, success is defined as scoring less pain after the treatment than before. Le p denote the probabiltiy of success for a patient (assumed to be the same for all twenty-five patients), and let T denote the number of successes in the trial. Then, as explained in $\cdots$,

$$P[T = t[= \binom{(25}{t} p^t (1-p)^{10-t}$$

for $t = 0, \cdots, 10$, and

$$P[T \geq t] = \sum_{k=t}^{25} \binom{25}{k} p^k (1-p)^{135-k}$$

The assertion that the treatment is ineffective can be stated $H_0 : p = \frac{1}{2}$. Suppose now that eighteen of the twenty-five patients report success. Then the p-value is

$$p^* = P_0[T \geq 18] = \sum_{k=18}^{25} \binom{25}{t} (\frac{1}{2})^k (\frac{1}{w})^{25-k} = \sum_{k=18}^{25} \binom{25}{k} (\frac{1}{2})^{25} = .0216 < .05.$$

So, the data are inconsistent with H_0 and the treatment's effectiveness has been established.

The reader should note the logic used in the last example in which effectiveness is shown, by showing that the treatment is not ineffective. Logic of this nature is common. The logic of significance tests may be compared to that of *proof by contraction* where one may disprove an assertion by assuming it and then deriving something that is obviously false. With significance tests one assumes an hypothesis and then shows that something that is very unlikely has occurred. The sterngth of the conclusion then depends on how unlikely, and this is measured by the p value.

2.4 Combinations of Events

If $A_1, \cdots, A_n$ are events and $J \subseteq \{1, \cdots, n\}$, then

$$B_j = \bigcap_{i \in J} A_i \tag{2.26}$$

is the event that A_i occurs for all $i \in J$, Let

$$s_j = \sum_{\#J=j} P(B_J), \tag{2.27}$$

where the summation extends over all $\binom{n}{j}$ subsets of size j. Thus,

$$\begin{aligned}
s_1 &= \sum_{I=1}^{n} P(A_i), \\
s_2 &= \sum \sum_{i<j} P(A_I \cap A_j), \\
&\cdots, \\
s_k &= \sum \cdots \sum_{i_1<i_2\cdots<i_k} P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}), \\
&\cdots \\
s_n &= P(A_1 \cap \cdots \cap A_n).
\end{aligned}$$

Then the probability of a union may be written in terms of $s_1, \cdots, s_n$: *If $A_1, \cdots, A_n$ are any n events, then*

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{k=1}^{n} (-1)^{k-1} s_k. \tag{2.28}$$

The proof of (3.15) will be described in the discrete case only. Thus suppose that Ω is a finite set and that $P(B) = \sum_{\omega \in B} p(\omega)$ for all subsets $B \subseteq \Omega$; equvalently $P(B) = \sum_{\omega \in \Omega} \mathbf{1}_B(\omega) p(\omega)$, where $\mathbf{1}_B$ denotes the indicator of $B, \mathbf{1}_B(\omega) = 1$ if $\omega \in B$ and $\mathbf{1}_B(\omega) = 0$ otherwise Let $A = A_1 \cup \cdots \cup A_n$ denote the union on the left side fo (**??**. If $\omega \notin A$, then clearly $\mathbf{1}_{B_J}(\omega) = 0$ for all non-empy $J \subseteq \{1, \cdots, n\}$, and the right side of (**??**) may be written

$$\sum_{k=1}^{n} (-1)^{k-1} \sum_{\#J=k} \sum_{\omega \in A} \mathbf{1}_{B_J}(\omega) = \sum_{\omega \in A} \left[\sum_{k=1}^{n} (-1)^{k-1} \sum_{\#J=k} \mathbf{1}_{B_J}(\omega) \right] p(\omega)$$

by reversing the order of summation.So, it suffices to show that the term in brackets (the coefficient of $p(\omega)$) is one. The inner sum $\sum_{\#J=k} \mathbf{1}_{B_J}(\omega).$ is just the number of subsets of size k that can be drawn from $\{1, \cdots, n\}$, so that

$$\sum_{\#J=k} \mathbf{1}_{B_J}(\omega). = \binom{n}{k},$$

So,

$$\sum_{k=1}^{n}(-1)^{k-1} \sum_{\#J=k} \mathbf{1}_{B_J}(\omega) = \sum_{k=1}^{n}(-1)^{k-1}\binom{n}{k}\binom{n}{1} - \binom{n}{2} + \cdots \pm \binom{n}{n}$$

which is one by Complement 1.3. ♢

The formula (3.15) is especially useful in problems that exhibit some symmety in the form

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{\cap} \cdots \cap A_k) \tag{2.29}$$

for all choices of $1 \leq i_1 < \cdots < i_k \leq n$. Then the s_k symplify to

$$s_k = \binom{n}{k} P(A_{\cap} \cdots \cap A_k).$$

Example 2.8 *: A Gift Exchange.* Friends agree to exchange gifts for a holiday. Each person writes his/her name on a slip of paper and places the slip in a box. The after a vigorous shake each person draws a slip from the box, and buys a present for the person identified on the slip. Of course the exercise will fail if anyone draws his/her own name. What is the probabiliy of this? Let n denote the number of people. Surpisingly, the answer is not highly sensitive to n. It is about .63 provided that $n \geq 6$. To see why label the people $1 \cdots, n$ by the order in which they draw; let Ω denote the set of all permutation $\omega = (i_1, \cdots, i_n)$ of$\{1, \cdots, n\}$; regard the list of people drawn from the box as a random permutation;and suppose that all $n!$ permutations are equally likely. Then

$$A_j = \{\omega : i_j = j\}$$

is the event that the j^{th} person draws his/her own name, and $A := \cup_{I=1}^{n} A_i$ is the event that someone draws his/her own name. It is not dificult to see that the symmetry condition (**??**) is satified in this example. Moreover,

$$P(A_1 \cap \cdots \cap A_k) = \frac{(n-k)!}{n!}$$

since A_k specifies that $i_j = j$ for $j = 1, \cdots, k$ and allows $i_{k+1}, \cdots, i_n$ to be permuted arbitrarily. So,

$$s_k = \binom{n}{k} P(A_1 \cap \cdots \cap A_k) = \binom{n}{k} \times \frac{(n-k)!}{n!} = \frac{1}{k!},$$

and

$$P(A) = \sum_{k=1}^{n} (-1)^{k-1}/k!.$$

Next,set $x = 1$ in the Taylor series expansion

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \tag{2.30}$$

to obtain $e^{-1} = \sum_{K=0}^{\infty} (-1)^k/k!$. Then rewrite the expression for $P(A)$ as

$$P(A) = 1 - \sum_{k=0}^{n} \frac{(-1)^k}{k!} \approx 1 - \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} = 1 - \frac{1}{e} = .6321 \cdots .$$

The approximation is excellent if $n \geq 6$

Example 2.9 *: The Coupon Collector's Problem.* A manufacter gives away coupons of various types with its product. If a consumer collects n coupons. what is the probability that he/she collects at least one of all the diferent types? Regard the n coupons collected as a sample of size with replacement from the population of all coupons. (If there are a large number of coupons, then there is not much difference between sampling with and without replacement. About the population, suppose that there are t different types of coupon and and equal numbers of all types, so that the probability that a single coupon is of type i sis $1/t$ for all $i = 1, \cdots, t$. Let A_i be the event that the n coupons collected do not include any of type i. Then $A = A_1 \cup \cdots \cup A_t$ is the even that at least one type is misssing from the sample, and A^c is the event that all types are represented in the sample. The probability of$A$ can b ecomputed from (**??**), and the probabilty that all types are collected is then $1 - P(A)$, *In this case* $P(A_i) = (1 - 1/t)^n$, $P(A_i \cap A_j) = (1 - 2/t)^n$ *for* $i \neq j$, *and*

$$P(B_J) = \left(1 - \frac{k}{t}\right)^n$$

for subsets $J \subseteq \{1, \cdots, t\}$,*of size* $\#J = k$. *So,*

$$s_k = \binom{t}{k}\left(1 - \frac{k}{t}\right)^n$$

and

$$P(A) = \sum_{k=1}^{t}(-1)^{k-1}\binom{t}{k}\left(1 - \frac{k}{t}\right)^n.$$

For example, if $n = t = 10$, *so that the sample size is equal to number of types, then the probability that all types are included in the sample is less than* .0001

2.5 Problems and Complements

Problems

1 Define an appropriate sample space for each of the following experiments:

(a) A (six sided) die is rolled, and the number of spots that appear is recorded.

(b) A die is rolled until an ace appears, and the number of rolls is recored.

(c) The number of traffic accidents in a given city on a given day is recorded.

(d) You look up the year of Newton's birth.

2 Define an appropriate sample space for each of the following experiments:

(a) A ra radio-active substance is observed and the number of emissions (clicks on a geiger-counter) during a given time interval is recorded.

(b) The time required for a radio-active substance to emit a particle is recorded.

(c) The annual precipitation in Seattle is recorded.

(d) The closing value of Apple stock is recorded each day for a week. **3** A person is selected from the population of a given city. Let A be the event that the person is female, B be the event that the person is under 30, and C be the event that he/she speaks a foreign language. Describe in symbols: (a) a male who speaks a foreign language; (b) a female who is under 30 and speaks a foreign language; a person who is male or under 30 but not both.

4 In the previous problem describe the following event in word: $A \cap C$; $A \cup (B \cap C)$; $A \cap B^c \cup B \cap A^c$; $A \cup B - C$.

5 A die is so loaded that the probability that k spots appear when it is rolled is proportional to k for $k = 1, \cdots, 6$. What is the probability that an odds number of spots appear?

6 The probability that a typist commits exactly k errors on a given page is proportional to $1/k!$ for $k = 0, 1, 2, \cdots$. What is the probability that he/she commits no errors?

7. The symmetric difference between two events, A and B say, is defined to be $A\Delta B = A^c \cap B \cup A \cap B^c$. Show that $P(A\Delta B) = P(A) + P(B) - 2P(A \cap B)$

8. Using only (2.16), show that $P(A \cup B \cup C) = P(A) + P(B) + (C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$.

Chapter 3

Conditional Probability and Independence

3.1 Conditional Probability

In some problems, partial information about the outcome may become available, and probability must be modified to take account of the additional information. Suppose, for example, that a family is known to have two children and that the four possible sex distributions, bb, bg, gb, gg are regarded as equally likely. Suppose also that the parents are seen shopping for girls' clothes (so that one of the children must be a girl). What is the probability that the other child is a boy? The answer is 2/3, not 1/2 for reasons that will be explained below.

Consider a probability model with sample space Ω and probability function P, say, and let A and B be events for which $P(A) > 0$. Then *the conditional probability of B given A* is defined by

$$P(B|A) = \frac{P(A \cap B)}{P(A)}. \tag{3.1}$$

It may be regarded as an appropriate modification of probability, if it is learned (only) that A has occurred. In the example, the event of interest is $B = \{(b, b), (b, g), (g, b)\}$ that the family has at least one boy, and the information is that there is at least one girl, so that $A = \{(b, g), (g, b), (g, g)\}$. Thus, $P(A) = 3/4$, $A \cap B = \{(b, g), (g, b)\}$,

$P(A \cap B) = 1/2$, and $P(B|A) = 2/3$.

Further examples, may help the reader develop intuition for conditional probabilities.

Example 3.1 An automobile insurance company classifies it policy holders as *experienced* or *inexperience*. The following table shows the number of accidents during the past year for each classification.

	Accident	No Accident	Total
Experienced	10%	70%	80%
Inexperienced	5%	15%	20%
Total	15%	85%	

For example, 10% of the policy holders were experienced and had had and accident during the past year. According to the table, experienced drivers had more accidents during the past year than inexperienced ones. An experienced driver is not more likely to have an accident, however. To see why, suppose that a policy holder is chosen at random from the group; let E be the event that the chosen person is experienced; and let A be the event that he/she had an accident. Then $P(E) = .80$, $P(E \cap A) = .10$, and $P(A|E) = .10/.80 = .125$, while $P(E^c) = .20$, $P(E^c \cap A) = .05$, and $P(A|E^c) = .05/.20 = .25$. *That is,* 25% *of inexperience drivers had an accident, compared to* 12.5% *of experienced ones.* ♢

If Ω is a finite set and $P(C) = \#C/\#\Omega$ for all $C \subseteq \Omega$, then

$$P(B|A) = \frac{\#(A \cap B)/\#\Omega}{\#A/\#\Omega} = \frac{\#(A \cap B)}{\#A} \tag{3.2}$$

for all B and non-empty A. In effect, the sample space has been reduced to A and the remaining outcomes are still equally likely. Example 3.1 above illustrates the process. In some cases, it is possible to calculate conditional probabilities directly, using (3.2).

Example 3.2 If South has 6 spades in a bridge game, what is the probability that North (South's partner) has at least two. Given South's hand, North's hand may

be regarded as a sample of size 13 from a deck containing 7 spades and 32 non-spades. The conditional probability that North has no spades is $\binom{32}{13}/\binom{39}{13} = .0428$, since there are $\binom{39}{13}$ possible hands for North of which $\binom{32}{13}$ have no aces. Similarly, the probability that North has one spade is $\binom{7}{1}\binom{32}{12}/\binom{39}{13} = .1946$. So, the desired conditional probability is

$$1 - \frac{\binom{7}{1}\binom{32}{12}}{\binom{39}{13}} - \frac{\binom{32}{13}}{\binom{39}{13}} = .2374.$$

$P(B|A)$ is a Probability Function in B; that is, $0 \le P(B|A) \le 1 = P(\Omega|A)$, and $P(B \cup C|A) = P(B|A) + P(C|A)$, if $B \cap C = \emptyset$. More generally, if $B_1, B_2, \cdots$ are (pairwise) mutually exclusive, then

$$P(\bigcup_{k=1}^{\infty} B_k|A) = \sum_{k=1}^{\infty} P(B_k|A).$$

The first two assertions are clear. To verify the finite additivity, observe that if $B \cap C = \emptyset$, then $(B \cap A) \cap (C \cap A) = \emptyset$ and, therefore,

$$\begin{aligned} P(B \cup C|A) &= \frac{P[(B \cup C) \cap A]}{P(A)} \\ &= \frac{P[(B \cap A) \cup (C \cap A)]}{P(A)} \\ &= \frac{P(B \cap A) + P(C \cap A)}{P(A)} = P(B|A) + P(C|A). \end{aligned}$$

The fourth condition may be verified similarly. As a corollary,

$$P(B^c|A) = 1 - P(B|A)$$

for all events B, since this is true of every probability function. This relation was implicitly used in Example 3.2.

3.2 Three Formulas

There are three simple formulas which relate conditional probabilities to unconditional ones. If A and B are events for which $0 < P(A) < 1$, then

$$P(A \cap B) = P(B|A)P(A), \tag{3.3}$$

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c), \tag{3.4}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}. \tag{3.5}$$

The first of these is clear from the definition (3.1). For the second write $B = (A \cap B) \cup (A^c \cap B)$ and

$$P(B) = P(A \cap B) + P(A^c \cap B) = P(B|A)P(A) + P(B|A^c)P(A^c),$$

by (3.3) applied to both A and A^c. The third relation then follows by writing $P(A|B) = P(A \cap B)/P(B)$ and substituting for $P(A \cap B)$ and $P(B)$. Equation (3.3) is called the *Product Rule*. Equations (3.4) and (3.5) are called the *Law of Total Probability* and *Bayes Formula*.

Example 3.3 On a True-False Examination, a student knows the answer with probability .6 and guesses otherwise. What is the probability that he/she answers a given question correctly? Given a correct answer, what is the conditional probability that the student knew the answer? Let A be the event that the student knows the answer, and let C be the event that he/she answers the questions correctly. Then $P(A) = .60$, $P(C|A) = 1$, and $P(C|A^c) = .5$ from the information above. So, the desired probabilities are

$$P(C) = P(C|A)P(A) + P(C|A^c)P(A^c) = 1 \times .60 + .50 \times .40 = .80,$$

$$P(A|C) = \frac{1 \times .60}{.80} = .75.$$

Example 3.4 [1] Consider a routine diagnostic test for a rare disease–for example, X-Rays and Lung Cancer. For a given patient, let D be the event that the disease is present and E be the event that the test indicates the disease to be present. Suppose, that the test is good but not perfect–for example, that $P(D) = .001, P(E|D) = .99$, and $P(E|D^c) = .025$. Then

$$P(E) = (.99)(.001) + (.025)(.999) = .00099 + .09990 = .02597,$$

[1] Books have been written about this example. A good one is *Matters of Life and Death*, by $\cdots$, Stanford University Press.

$$P(D|E) = \frac{.00099}{.02597} = .03813 < .04.$$

That is, less than four percent of people for whom the test indicates the disease to be present actually have the disease.

Here $P(E|D^c)$ is called the *false positive rate*, and $P(E^c|D)$ is called the *false negative rate*. Even though these error rates are low, in the example, the test is unreliable in that most of the people who test positive are not sick. To understand this apparent paradox, recall that the disease is very rare. For a person who tested positive, the probability that the disease is present has increased from .1% to almost 4%, but this probability is still low in absolute terms. ♢

Example 3.5 Terrorists hold five hostages and have agreed to exchange two for food. The two to be released are to be chosen by drawing lots. Is there an advantage to drawing first? More generally, suppose that two cards are drawn in order without replacement from a deck R red and $N-R$ white cards–for example, $R = 2$ and $N = 5$. Let A be the event that the first card is red; and let B be the event that the second card is red. Then $P(A) = R/N$, $P(A^c) = (N-R)/N$, $P(B|A) = (R-1)/(N-1)$, and $P(B|A^c) = R/(N-1)$. So,

$$P(B) = \frac{(R-1)}{(N-1)} \times \frac{R}{N} + \frac{R}{(N-1)} \times \frac{(N-R)}{N} = \frac{R}{N}.$$

$$P(A|B) = \frac{(R-1)}{(N-1)} \times \frac{R}{N} / \frac{R}{N} = \frac{(R-1)}{(N-1)}.$$

In the example, there is no advantage, or disadvantage to drawing first.

Similar, results may be obtained for sampling with replacement. Then $P(B) = R/N$ and $P(A|B) = R/N$. ♢

Several Events. There are some simple extensions of Equations (3.3), (3.4), and (3.5). First, if $A_1, \cdots, A_n$ are any n events, then

$$P(\cap_{i=1}^m A_i) = P(A_1) \prod_{j=2}^{m} P(A_m | \cap_{i=1}^{j-1} A_i). \tag{3.6}$$

Moreover, $A_1, \cdots, A_m$ are mutually exclusive events for which $P(A_i) > 0$, $i = 1, \cdots, m$ and $\cup_{i=1}^m A_i = \Omega$, then

$$P(B) = \sum_{i=1}^{m} P(B|A_i)P(A_i) \tag{3.7}$$

and

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^m P(B|A_i)P(A_i)} \tag{3.8}$$

for every other event B and all $j = 1, \cdots, m$. As above, (3.6) is called the *Product Rule*, and (3.7) and (3.8) are called the *Law of Total Probability* and *Bayes Rule*. The proof of (3.6) is left as an exercise. For (3.7), let B be any event. Then $B = B \cap \Omega = B \cap (\cup_{i=1}^m A_i) = \cup_{i=1}^m (A_i \cap B)$. So,

$$P(B) = \sum_{i=1}^{m} P(A_i \cap B) = \sum_{i=1}^{m} P(B|A_i)P(A_i),$$

by the additivity of probability and (3.3), applied to each A_i. Equation (3.8) then follows by writing $P(A_j|B) = P(A_j \cap B)/P(B)$ and computing $P(A_j \cap B)$ by (3.3) and $P(B)$ by (3.7).

Example 3.6 Box 1 contains two gold coins, Box 2 contains one gold and one silver coin, and Box 3 contains two silver coins. One box is selected at random, and then the two coins are drawn out sequentially with all choices being equally likely at each stage. Let B be the event that the first coin drawn is gold, and C the event that the second coin drawn in gold. What is the conditional probability of C given B? By definition, it is $P(BC)/P(B)$. The numerator and denominator may be computed using (3.7):

$$\begin{aligned} P(B) &= P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_2)P(B|A_2) \\ &= \frac{1}{3} \times 1 + \frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times 0 = \frac{1}{2} \end{aligned}$$

and

$$\begin{aligned} P(BC) &= P(A_1)P(BC|A_1) + P(A_2)P(BC|A_2) + P(A_2)P(BC|A_2) \\ &= \frac{1}{3} \times 1 + \frac{1}{3} \times 0 + \frac{1}{3} \times 0 = \frac{1}{3}. \end{aligned}$$

So,

$$P(C|B) = \frac{P(BC)}{P(B)} = \frac{2}{3}.$$

The reader may find this mildly surprising: If a gold coin is selected at the first stage, then the coin has to have been drawn from Box 1 or 2, suggesting that the conditional probability of drawing second gold coin is 1/2. The explanation is that if a gold coin is selected on the first draw, it is more likely to have been drawn from Box 1. For

$$P(A_1|B) = \frac{PA_1)P(B|A_1)}{P(B)} = \frac{1/3}{1/2} = \frac{2}{3}$$

by (3.8) ♢

Simplson's Paradox. It is possible to have Events $A_1, \cdots, A_m$ and B and two different probability functions P_1 and P_2 for which

$$P_1(B|A_k) < P_2|B|A_k) \tag{3.9}$$

for all $k = 1, \cdots, m$ and

$$P_2(B) < P_1(B). \tag{3.10}$$

Here is an example.

Example 3.7 The following table gives the death rates in 1930 per 100,000 for several ages ranges in each of two states along with the percentage of people in each range.

	State 1		State 2	
Age	Percent	Deaths	Percent	Deaths
0-4	9.4	2056	11.4	2392
5-14	19.3		26.7	
15-24	16.2		21.8	
25-34	13.3	391	12.6	871
35-44	12.7	545	11.0	1242
45-54	11.3	1085	8.3	1994
55-64	9.1	2036	4.6	3313
65-74	5.8	5219	2.3	6147
75-	2.8 7	13,645	1.0	14,136

Suppose that a person is selected a random from each of the two states and let P_1 and P_2 be the probability function for States 1 and 2. Further, let A_i be the event that he/she is in the i^{th} age category, and let B be the event that he/she died during the year. Then, for example, $P_1(A_1) = .094$, $P_1(B|A_1) = .02056$, $P_2(A_1) = .114$, and $P_2(B|A_1) = .02392$. It is clear from inspection of the table that (3.9) holds. The unconditional probabilities $P_1(B)$ and $P_2(B)$ may be computed from (3.7), and $P_2(B) = \cdots < \cdots = P_1(B)$. In words, the age adjusted death rates in State 1 were lower in all age categories, but the overall death rate in State 2 was lower. $\diamondsuit$

3.3 Independence

For a given probability model, with sampe space Ω and probability function P, events A and B are said to be *independent* iff

$$P(A \cap B) = P(A)P(B). \tag{1}$$

Thus, if $P(A) > 0$, then A and B are independent iff $P(B|A) = P(B)$, since $P(A \cap B) = P(B|A)P(A)$. Intuitively, independence means that the occurrence or non-occurrence of one event does not affect the probability that the other occurs.

Example 3.8 If a single card is drawn from a standard deck, then the events $A = \{$an ace$\}$ and $B = \{$a spade$\}$ are independent, since $P(A) = 4/52 = 1/13, P(B) = 13/52 = 1/4$, and $P(A \cap B) = 1/52 = (1/13)(1/4) = P(A)P(B)$. $\diamondsuit$

Example 3.9 If two tickets are drawn with replacement from a box containing R red and $N-R$ white tickets, then the events $A = \{$red on the first draw$\}$ and $B = \{$red on the second draw$\}$ are independent, since $P(A \cap B) = (R \times R)/(N \times N) = P(A)P(B)$.

If the sampling were without replacement, then A and B are not independent, since $P(B|A) = (R-1)/(N-1) < R/N = P(B)$. $\diamondsuit$

Example 3.10 Let $\Omega = (0, 1] = \{\omega : 0 < \omega \leq 1\}$ and suppose that $P((a, b]) = b - a$ for all subintervals $(a, b] \subseteq \Omega$, as in Example 2.?. Then $A = (0, \frac{1}{2}] = \{\omega : 0 < \omega \leq \frac{1}{2}\}$

and $B = (0, \frac{1}{4}] \cup (\frac{1}{2}, \frac{3}{4}]$ are independent. To see this observe that $P(A) = \frac{1}{2},\ P(B) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$, and $P(A \cap B) = P((0, \frac{1}{4}]) = \frac{1}{4} = P(A)P(B)$. In this example, A is the event that the first binary digit of ω is zero, and B is the event that the second binary digit is zero. $\diamondsuit$

As Example 2 illustrates, independence of events depends on P as well as the events.

Several Events. Events $A_1, ..., A_n$ are said to be *(mutually) independent* iff

$$P(A_i \cap A_j) = P(A_i)P(A_j),\ \forall\ i < j, \tag{3.11}$$

$$P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k),\ \forall\ i < j < k,$$

$$P(A_{i_1} \cap ... \cap A_{i_k}) = P(A_{i_1}) \times ... \times P(A_{i_k}),\ \forall 1_1 < ... < i_k, \tag{3.12}$$

for all $k = 2, \cdots, n$. In words, $A_1, \cdots, A_n$ are mutually independent if the probability of any subcollection of $A_1, \cdots, A_n$ is the product of their probabilities. Mathematically, (mutual) independence is a stringent condition. Equation (2) imposes $\binom{n}{k}$ constraints on the probabilities of interesections of $A_1, \cdots, A_n$, and it must hold for all $k = 1, \cdots, n$. Events $A_1, ..., A_n$ are said to be *pairwise independent* if A_i and A_j are independent for all $i \neq j$. This is equivalent to (3.11). It is clear that (mutual) independence implies pairwise independence. That pairwise independence does does not imply mutual independence is shown in Example (**??**) below. In the sequel, the unqualified term "independent" means mutually independent.

Example 3.11 If n tickets are drawn with replacement from a box containing R red and $N - R$ white tickets, then the events $A_i = \{\text{red on the } i^{th} \text{ draw}\}, i = 1, ..., n$, are (mutually) independent. In fact, if $2 \leq k \leq n$, then $P(A_{i_1} \cap ... \cap A_{i_k}) = R^k/N^k = P(A_{i_1}) \times ... \times P(A_{i_k})$ for all $i_1 < ... < i_k$. $\diamondsuit$

Example 3.12 Suppose that a single card is drawn at random from a box contains four cards labelled $1, 2, 3, 4$, and let $A_i = \{i, 4\}$, the event that either card i or card 4 is drawn for $i = 1, 2, 3$. Then clearly, $P(A_1) = P(A_2) = P(A_3) = 1/2$. If $i \neq j$, then $A_i \cap A_j = \{4\}$ and, therefore, $P(A_i \cap A_j) = 1/4 = P(A_i)P(A_j)$. So, A_1, A_2, A_3

are pairwise independent. However, $P(A_1 \cap A_2 \cap A_3) = P(\{4\}) = 1/4 \neq 1/8 = P(A_1)P(A_2)P(A_3)$, so that A_1, A_2, A_3 are not mutually independent. $\diamondsuit$

It is clear that independence is preserved by relabeling of $A_1, \cdots, A_n$ and that any subcollection of independent events is again independent: *If $A_1, ..., A_n$ are independent and if $B_i = A_i$ or A_i^c for all $i = 1, ..., n$, then $B_1, ..., B_n$ are independent.* The proof of this assertion is supplied in the next section.

Series and Parallel Connections If $A_1, \cdots, A_n$ are events, then $\cup_{i=1}^n A_i = A_1 \cup ... \cup A_n$ and $\cap_{i=1}^n A_i = A_1 \cap ... \cap A_n$ denote their union and intersection. Then *De Morgan's Laws* assert

$$(\cup_{i=1}^n A_i)^c = \cap_{i=1}^n A_i^c \tag{3.13}$$

and

$$(\cap_{i=1}^n A_i)^c = \cup_{i=1}^n A_i^c. \tag{3.14}$$

If $A_1, ..., A_n$ are independent and if $p_i = P(A_i), i = 1, ..., n$, then

$$P(A_1 \cap ... \cap A_n) = p_1 \times ... \times p_n \tag{3.15}$$

$$P(A_1 \cup ... \cup A_n) = 1 - (1 - p_1) \times ... \times (1 - p_n) \tag{3.16}$$

The first assertion is clear. For the second, write $P(\cup_{i=1}^n A_i) = 1 - P[(\cup_{i=1}^n A_i)^c] = 1 - P(\cap_{i=1}^n A_i^c)$ and $P(\cap_{i=1}^n A_i^c) = P(A_1^c) \times ... \times P(A_n^c) = (1 - p_1) \times ... \times (1 - p_n)$. For example, if $p_i = p \ \forall \ i = 1, ..., n$, then the two probabilities are p^n and $1 - (1 - p)^n$. These may be quite different. For $p = .99$ and $n = 100$, they are .3679... and $1 - 10^{-200}$.

Consider a group of electrical devices, for example, a string of lights, each of which may fail with a small probability. If the devices are connected in series, then current will flow (from the source to the ground) iff every device operates properly. So, if A_i denotes the event that the i^{th} device operates properly, $i = 1, \cdots, n$, then the event that current follows is $\cap_{i=1}^n A_i$, and the probability of this event is given by (3.15). If the devices are connected in parallel, then current flows iff at least one of the devices operates properly. In this case, the event that current flows is $\cup_{i=1}^n A_i$, and its probability is given (3.16).

It is possible to construct more complicated connections; for example, devices with subsystems may be connected in parallel, but subsystems may be connected in series. Here is an example, phrased in a different language.

Example 3.13 Small University has three students, four professors, and five deans. On any given day, the students show up for class with probability .75 each, the professors come to work with probability .5 each, and the deans appear for work with probability .25 each. Class are held iff at least one member of each group comes to class or work. Assuming independence, what is the probability that classes are held. The probability that at least one student comes to class is $1-(.25)^3 = .9849$; the probability that at least one professor comes to work is $1-(.05)^4 = .9375$; and the probability that at least one dean comes to work is $1-(.75)^5 = .7627$. So, the probability that classes are held is .7039. $\diamondsuit$

Product Spaces and Repeated Trials. Let $(\Omega_i, \mathcal{A}_i, P_i), i = 1, \cdots, n$ $(n \geq 2)$ denote probability space, regarded as models for experiments $\mathcal{E}_1, \cdots, \mathcal{E}_n$. Then it is possible to construct a probability space $(\Omega, \mathcal{A}, P)$ that provides a description for doing all n experiments in such a manner that the outcome of one does not affect those of the others. In this construction Ω denotes the Cartesian procuct. $\Omega = \Omega_1 \times \cdots, \times\Omega_n$. Thus, elements of Ω are lists $\omega = (\omega_1, \cdots, \omega_n)$ with $\omega_i \in \Omega_i,\ i = 1, \cdots, n$. If $A_i \in \mathcal{A}_i$, so that $A_i \subseteq \Omega_i,\ i = 1, \cdots, n$, then their Cartesian product $A_1 \times \cdots \times A_n = \{\omega \in \Omega : \omega_i \in A_i, \text{for all } i = 1, \cdots, n\}$ is called a *measurable rectangle. There are a sigma=algrebra$\mathcal{A}$ of subsets of* Ω *and a probability measure* P *defined on* $\mathcal{A}$ *or which*

$$P(A_1 \times \cdots \times A_n) = P_1(A_1) \times \cdots \times P_n(A_n) \tag{3.17}$$

for all measurable rectangles. .

$$\tilde{A} = \{\omega \in \Omega : \omega_i \in A\},$$

so that $A \subseteq \Omega$. This operation injects A into the larger space Ω

3.4 Mendel's Laws

inheritable characteristics of plants and animals are carried by *genes*, which occur in pairs within an organism . Pairs of genes are located on long strands, called *chromosomes*, and a pair of genes may be identified with its location on the chromosome. In sexual reproduction, each parent contributes one of its two genes to the offspring. The genes have different form called *alleles*. For example, garden peas may produce either round or wrinkled seeds, depending on the forms taken by the genes. In the simplest cases, there are only two alleles for each gene. Call them A and a. Since genes occur in pairs, there are three possibilities, AA, Aa, and aa, there being no distinction between Aa and aA. There is an important distinction between the *genotype*, the genetic composition of an organism, and its *phenotype*, its observable characteristics. In some cases, AA, Aa, and aa may be expressed by distinct phenotypes. For example, AA, Aa, and aa might be expressed by red pink and white flowers. In other case Aa and AA may have the same phenotype. In such cases, A is said to be *dominant over* a, and a is said to be *recessive*. For example, round is dominant over wrinkled.

The science of genetics began with the findings of Gregor Mendel, an Austrian monk who studied garden peas., a sexually reproducing plant. Mendel's Laws may be stated:

• *The Law of Independent Assortment*: The genes contributed by each parent are selected independently with all choices equally likely.

• *The Law of Independent Aggregation*: The contributions for different genes are independent.

Something important should be noted here: The laws of science (genetics in this case) are expressed in probabilistic terms.

What does the The Law of Independent Assortment predict? If there are two alleges and two hybrids (Aa) are crossed, then each parent independently contributes either an A or an a with probability 1/2 each. So AA, Aa aA and aa each have probability $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$.

If also A is dominate over a, so that AA and Aa are both expressed by the domi-

nant phenotype, then the probability of the dominant phenotype is $1/4 + 1/2 = 3/4$. So, in a large number of crosses about 3/4 of the offspring should show the dominant phenotype. The predictions agree with experimental data[2]. In 7324 (74.7%)[3] crosses of garden peas,, 5474 (74.7%) of the offspring produced round seeds, and 1850 wrinkled ones.

Now consider two genes with two alleles each, say A or a and B or b.. If two hybrids ($AaBb$ are crossed, then the offspring have each of the 16 possible genotypes with probability 1/16, by the Law of Independent Aggregation. Similarly, if A and B are dominant, then the offspring with have the two dominant phenotypes with probability $\frac{3}{4} \times \frac{3}{4} = \frac{9}{16}(56.15.25\%)$. They will have one dominant and one recessive trait with probability 3/16(18.75%) and both recessive traits with probability 1/16(6.25%).

3.5 Problems and Complements

1. A committee of size three is to be selected from a group of ten men and ten women. What is the conditional probability that both sexes are represented, given that there is at least one man on the committee? *Ans: .8824*

2. In Problem 1, suppose that the committee size is four. What is the conditional probability that the committee contains two men and two women, given that it contains at least one person of each sex?

3. In a game of bridge, North and South have nine spades in their combined hands. What is the conditional probability that the remaining four spades are evenly divided between East and West? *Ans:* $\binom{4}{2}\binom{22}{11}/\binom{26}{13} = .407$

4. If South has no aces, what is the conditional probability that North has at least two aces?

5. In a certain court, cases are decided by a single judge. Suppose that the judge finds an innocent person guilty with probability .2 and finds a guilty person guilty with probability .9. Suppose also that 60% of defendants are guilty. What proportion

[2]In fact, it may agree a bit too well. See Fisher (1936), *Annals of Science*. **1**, 115-137

[3]Source: *General Genectics*, by Srb, Owen, and Edgar

of defendants are convicted (found guilty)? What proportion of people convicted are actually guilty? *Ans: .62 and .871*

6. A test for a rare disease has a false positive rate of 2.5% and a false negative rate of 5%. Suppose that .5% of the population have the disease. If a person takes the test as part of a routine physical examination (no symptoms), what is the probability that the test will indicate the disease to be present? Given that the test is indicates the disease, what is conditional probability that the person has it?

7. One card is drawn from a standard deck. Let A be the event that the card is either a spade or a club; let B be the event that the card is either a heart or a club; and let C be the event that the card is either a diamond or a club. Show that A and B are independent but that A, B, and C are not mutually independent.

8. Two cards are drawn from a standard deck without replacement. Which pairs of the following events are independent? A: the first card is an ace; B: the first card is a spade; C: the second card is a heart. Why?

Chapter 4

Discrete Random Variables

4.1 Probability Mass Functions

Suppose, for example, that there is interest in the sum of spots on two dice, as in Example 1.2. Then an outcome is a pair $\omega = (i, j)$, where $1 \leq i, j \leq 6$, and the the sum of spots, $X(i, j) = i + j$, defines a function on the sample space. The event that the sum of spots has a given values, for example seven, is then a subset of the sample space, $(\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\})$. This is called the event that $X = 7$ and denoted $\{X = 7\}$, and its probability is

$$P[X = 7] = P(\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}) = \frac{1}{6}.$$

since there are 36 equally likely outcomes.

The example may be generalized. Consider a probability model with sample space Ω and probability function P. A *random variable* is a function

$$X : \Omega \to \mathbb{R}.$$

Less formally, a random variable is a rule that associates a real number $X(\omega)$ with each outocme ω. A random variable is said to be *discrete* is its range $\mathcal{X} = \{X(\omega) : \omega \in \Omega\}$ is a finite or countably infinite set; that is, $\mathcal{X} = \{x_1.x_2, \cdots\}$ (finite or infinte). The sum of spots example just given is of this nature with $\mathcal{X} = \{2, 3, \cdots, 11, 12\}$. If X is a random variable and $B \subseteq \mathbb{R}$, then it is convenient to write $\{X \in B\}$ for

$\{\omega : X(\omega) \in B\}$ and call $\{X \in B\}$ the *event that X is in B*. The probability of this event is denoted by

$$P[X \in B] = P(\{\omega \in \Omega : X(\omega) \in B\}), \tag{4.1}$$

with natural simplifications, like writing $P[X = c]$ for the probability that $X \in \{c\}$. The sum of spots example illustrates the use of this notations.

If X is a discrete random variable, then the *probability mass function of X* is defined by

$$f_{\mathbf{x}}(x) = P[X = x] \tag{4.2}$$

for $x \in \mathbb{R}$. In words, $f(x)$ is the probability of the set of outcomes for which $X(\omega) = x$. In the dice example $f(7) = 1/6$. Similar calculation show that $f(6) = f(8) = 5/36$, $f(5) = f(9) = 4/36$, and

$$f(x) = \frac{6 - |x - 6|}{36}$$

for $x = 2, \cdots, 12$ and $f(x) = 0$ for other values of x. The probability mass function of a random variable X may also be denoted by f_X if there is danger of confusion.

Example 4.1 Recall that a bridge hand is a combination of 13 cards from a standard deck. If a bridge hand is chosen at random, then the sample space consists of all bridge hands, and $\#\Omega = \binom{52}{13}$. The number of aces in a hand is a well defined random variable X. There is no simple mathematical expression for this random variable, but it is a well defined rule that associates numbers to outcomes. Clearly $\mathcal{X} = \{0, 1, 2, 3, , 4\}$. For $x \in \mathcal{X}$ the number of hands containing exactly x aces is $\binom{4}{x}\binom{48}{13-x}$, since there are $\binom{4}{x}$ to choose x aces and $\binom{48}{13-x}$ to choose $13 - x$ non-aces. So,

$$f_{\mathbf{x}}(x) = P[X = x] = \frac{\binom{4}{x}\binom{48}{13-x}}{\binom{52}{13}}$$

for $x = 0, 1, 2, 3, 4$ and $f(x) = 0$ otherwise. Numerically $f(0) = .3038, \cdots$. ♢

Probability mass functions have certain characteristic properties. *If $f = f_{\mathbf{x}}$ is the probability mass function of a random variable X, with rangle $\mathcal{X}$, then*

$$f(x) \geq 0 \text{ for all } x \in \mathbb{R}, \tag{4.3}$$

$$f(x) = 0 \text{ unless } x \in \mathcal{X}, \tag{4.4}$$

$$\sum_{x \in \mathcal{X}} f(x) = 1, \tag{4.5}$$

and

$$P[X \in B] = \sum_{x \in B} f(x) \tag{4.6}$$

for all subsets $B \subseteq \mathbb{R}$. Conversely, if f is any function that satisfies (4.3),(4.4), and (4.5), then there is a random variable X with probability mass function f. To see why, suppose first that f is the probability mass function of a random variable X. Then $f(x)$ is a probability and, therefore, non-negative. If $x \notin \mathcal{X}$, then there are no ω for which $X(\omega) = x$, so that $f(x)$ is the probability of the emptyset and, therefore, $f(x) = 0$. For (4.6), let $B \subseteq \mathbb{R}$. Then $B \cap \mathcal{X} = \{x'_1, x'_2, \cdots\}$, where x'_k are the x_k that are in B. So, $B = \cup_k \{X = x'_k\}$, and

$$P[X \in B] = P[X \in \bigcup_k \{x'_k\}] = \sum_k P[X = x'_k] = \sum_{x \in B} f(x).$$

This establishes (4.6), and (4.5) then follows by letting $B = \mathcal{X}$ and noting that $P[X \in \mathcal{X}] = 1$. For the converse,

...

4.2 The Mean and Variance

If f is a probability mass function then the mean of f is defined by

$$\mu = \sum_{x \in \mathcal{X}} x f(x), \tag{1}$$

where $\mathcal{X} = \{x \in \mathbb{R} : f(x) > 0\} = \{x_1, x_2, \cdots\}$, provided that the sum converges absolutely[1] (if $\mathcal{X}$ is an infinite set). If X is a random variable with probability mass function f, then the mean of f is also called the mean of X and may be denoted by $\mu_{\mathbf{x}}$. Thus the mean is a weighted average of its possible values of the possible values of X with weights $f(x)$.

[1]That is, provided that $\sum_{x \in \mathcal{X}} |x| f(x) < \infty$

Example 4.2 If A is an event, then its indicator $\mathbf{1}_A$ is defined by

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & if\ \omega \in A \\ 0 & if\ \omega \notin A \end{cases}.$$

Thus, $\mathbf{1}_A$ is a random variable for which $f(1) = P\{\mathbf{1}_A = 1\} = P(A)$, $f(0) = P(A^c)$, and $f(x) = 0$ for other values of x. So, $\mu = 1P(A) + 0P(A^c) = P(A)$. In words, the mean value of an indicator variable is the probability of the event. The example shows that the language of random variables contains the language of events as a special case: Any question that could be ask or answered in terms of events could also be ask or answered in terms of random variables. ♢

If f is a probability mass function with mean μ, then *the variance of* f is defined by

$$\sigma^2 = \sum_{x \in \mathcal{X}} (x - \mu)^2 p(x). \tag{2}$$

In this case, the variance is to be interpreted as ∞ if the sum does not converge. If X has probability mass function p, then σ^2 is called the variance of X and may be denoted by σ_X^2. The mean may be regarded as the center of the distribution, and the variance and standard deviation measure the tendency of X to deviate from μ. The square root $\sigma = \sqrt{\sigma^2}$ of the variance is called the *standard deviation of* f *or of* X.

Example3.2:*Continued* The probability mass function of the number of aces in a bridge hand is given in the table below, along with the calculation of the mean and variance:

...

There is an alternative expression for the variance. The *moments* of a probability mass function f are

$$\mu_k = \sum_{x \in \mathcal{X}} x^k f(x)$$

for $k = 0, 1, 2, \cdots$, provided that the sum converges. Thus, $\mu_0 = 1$, by (4.5), and $\mu_1 = \mu$, the mean. The alternative expression of σ^2 is then

$$\sigma^2 = \mu_2 - \mu^2$$

For, writing $(x-\mu)^2 = x^2 - 2\mu x + \mu^2$,

$$\begin{aligned}\sigma^2 &= \sum_{x\in\mathcal{X}}(x^2 - 2\mu x + \mu^2)p(x)\\ &= \sum_{x\in\mathcal{X}} x^2 f(x) - 2\mu\sum_{x\in\mathcal{X}} x f(x) + \mu^2\sum_{x\in\mathcal{X}} f(x)\\ \sigma^2 = \qquad &= \sum_{x\in\mathcal{X}} x^2 f(x) - 2\mu^2 + \mu^2\\ &= \sum_{x\in\mathcal{X}} x^2 f(x) - \mu^2 = \mu_2 - \mu^2,,\end{aligned}$$

where the next to last equality uses the definition of μ and Equation (4.5).

Example 4.2: *Continued.* For an indicator variable, $\mu_2 = 1^2 f(1) + 0^2 f(0) = f(1) = P(A)$ and, therefore, $\sigma^2 = \mu_2 - \mu^2 = P(A) - P(A)^2 = P(A)[1 - P(A)]$. ♢

Example 4.3 If an n-sided balanced die is rolled once, then the number of spots X is random variable with probability mass function

$$f(x) = \begin{cases} 1/n & if\ x = 1, 2, \cdots, n \\ 0 & if\ otherwise \end{cases} \tag{4.7}$$

Thus,

$$\mu = \sum_{x=1}^{n} \frac{x}{n} = \frac{n+1}{2},$$

since $1 + 2 + \cdots + n = n(n+1)/2$. Similarly

$$\sum_{k=1}^{n} \frac{k^2}{n} = \frac{n(n+1)(2n+1)}{6n} = \frac{(n+1)(2n+1)}{6},$$

since $1 + 4 + 9 + \cdots + n^2 = n(n+1)(2n+1)/6$, and

$$\sigma^2 = \frac{(n+1)(2n+1)}{6} - (\frac{n+1}{2})^2 = \frac{n^2-1}{12},$$

after some algebra. The probability mass function (4.7) is called the *discrete uniform with parameter n.* ♢

Moment Generating Functions. If f, then the *moment generating function of* f is defined by

$$M(t) = \sum_{x\in\mathcal{X}} e^{tx} f(x) \tag{4.8}$$

for those $t \in \mathbb{R}$ for which the sum converges. Observe that if $t = 0$, them $M(0) = 1$ for any f, since $e^0 = 1$. The name derives from the following simple formula: *$M(t) < \infty$ for all t in some neighborhood of 0, then the moments of p are*

$$\mu_k = M^{(k)}(0) := \frac{d^k}{dt^k} M(t)\big|_{t=0}, \tag{4.9}$$

the k^{th} derivative of M at $t = 0$, for all $k = 1, 2, \cdots$. This is easily seen if $\mathcal{X}$ is a finite set, for then

$$\frac{d^k}{dt^k} M(t) = \frac{d^k}{dt^k} \sum_{x \in \mathcal{X}} e^{tx} f(x) = \sum_{x \in \mathcal{X}} \frac{d^k}{dt^k} e^{tx} f(x) = \sum_{x \in \mathcal{X}} x^k e^{tx} f(x)$$

for all $t \in \mathbb{R}$, so that

$$M^{(k)}(0) = \sum_{x \in \mathcal{X}} x^k e^{tx} p(x) = \mu_k$$

for all $k = 1, 2, \cdots$. These relations are still valid when $\mathcal{X}$ is an infinite set, but they require some justification. This is provided in $\cdots$.

Thus the mean and variance may be expressed in terms of the moment generating function as

$$\mu = M'(0),$$
$$\sigma^2 = M''(0) - M'(0)^2.$$

The logarithm of M i s called *cumulant function of f* and denote by κ. Thus, $\kappa(t) = \log[M(t)]$, $kappa'(t) = M'(t)/M(t)$, and $\kappa''(t) = [M''(t0 - M'(t)^2]/M(t)^2$, so that

$$\mu = \kappa'(0) \qquad \text{and} \qquad \kappa''(0) = \sigma^2. \tag{4.10}$$

Examples are provided in the next section.

4.3 Special Discrete Distributions

Observe that (4.7) defines a entire family of probability mass functions, one for each integer n. Four more important families are introduced in this section.

Hypergeometric Distributions. Example 3.2 generalizes easily. If a sample of size n is drawn without replacement from a box containing R red tickets and $N - R$

white tickets, then the number of red tickets in the sample is a random variable X with probability mass function

$$f(x) = \frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{n}} \tag{4.11}$$

for $x = 0, \cdots, n$ and $f(x) = 0$ otherwise. The derivation of this formula simply replaces the number 4, 48, and 13 with the symbols R, $N-R$, and n, and is left a an exercise. The probability mass function (4.11) is called the *Hypergeomeric probability mass function with parameters N, R, and n.*

To compute the mean of f, observe first that

$$\binom{N}{n} = \frac{N}{n}\binom{N-1}{n-1} \qquad \text{and} \qquad k\binom{R}{k} = R\binom{R-1}{k-1}$$

$1 \le k \le n$. So,

$$\mu = \sum_{k=0}^{n} kf(k) = \sum_{k=1}^{n} \frac{Rn}{N} \frac{\binom{R-1}{k-1}\binom{N-R}{n-k}}{\binom{N-1}{n-1}} = \frac{nR}{N},$$

since the last sum is just the sum of the Hypergeometric probability mass function with parameters $N-1$, $R-1$, and $n-1$. The variance can be computed similarly, and

$$\sigma^2 = \left(\frac{N-n}{N-1}\right)\frac{RWn}{N^2}.$$

The details are omitted here, since the result will be derived by quite different methods in Chapter **??**.

Binomial Distributions. Suppose now that there are independent events $A_1, \cdots, A_n$ with the same probabilty, $P(A_i) = p$, $i = 1, \cdots, n$ and interest in the number of occurrences X of $A_1, \cdots, A_n$. If the occurrence of an A_i is regarded as a success, then X is simply the number of successes. Examples include independent plays of a game, like roulette, in which A_i is the event that the i^{th} game is won. Here is a simple example.

Example 4.4 Suppose that a gambler plays four games of roulette, always betting on red. Let W_i be the even that he/she wins the i^{th} game and $L_i = W_i^c$ the event that he/she loses. Thus, $P(W_i) = 9/19$ for $i = 1, \cdots, 4$. Then the event that $X = 2$

is the union

$$\{X = 2\} = W_1W_2L_3L_4 \cup W_1L_2W_3L_4 \cup W_1L_2L_3W_4 \\ \cup L_1W_2W_3L_4 \cup L_1W_2L_3W_4 \cup L_1L_2W_3W_4.$$

since there are 6 ways to win two games and lose two, and

$$\begin{aligned} P[X = 2] &= P(W_1W_2L_3L_4) + \cdots + P(L_1L_2W_1W_2) \\ &= P(W_1)P(W_2)P(L_3)P(L_4) + \cdots + P(L_1)P(L_2)P(W_1)P(W_2) \\ &= (\frac{9}{19})^2(\frac{10}{19})^2 + \cdots + (\frac{9}{19})^2(\frac{10}{19})^2 \\ &= 6(\frac{9}{19})^2(\frac{10}{19})^2 \end{aligned}$$

since the 6 are mutually exclusive and the events W_1, W_2, W_3, W_4 are independent with the same probability ♢

This simple calculation generalizes easily. Let $A_1, \cdots, A_n$ denote independent events with the same probabilty, $P(A_i) = p,\ i = 1, \cdots, n$, and let X denote the number of occurrences. Then

$$X = \mathbf{1}_{A_1} + \cdots + \mathbf{1}_{A_n},$$

since the sum of indicator functions just counts occurrence. The probability mass function of X is then

$$f(x) = \binom{n}{x} p^x(1-p)^{n-x} \tag{4.12}$$

for $x = 0, \cdots, n$ and $f(x) = 0$ otherwise. To see this simply observe that the event $\{X = x\}$ is simply the event that for some combination C of the indices $\{1, \cdots, n\}$ A_i occurs when $i \in C$ and A_i^c occurs when $i \notin C$. For any given combination, this has probability $p^x(1-p)^{n-x}$, and (4.12) then follows since there are $\binom{n}{x}$ combinations of size x.

Example 4.5 *: Continued.* If the gambler plays n games, what is the probability that he/she is a net winner. Letting m the least integer that exceeds $n/2$, the probability that $X > m$ is required. The probability of winning x games is given by (4.12)

with $p = 9/19$. So,

$$P[X > m] = \sum_{x=m}^{n} \binom{n}{x} (\frac{9}{19})^x (\frac{10}{19})^{n-x}$$

The following table gives some numerical values. From this table, the probability that the gambler is a net winner (wins six or more times) is .1310.

Table 4.1: default

m	prob
1	.9837
2	.9242
3	.7816
4	.5568
5	.3112
6	.1310
7	.0385
8	.0069
9	.0006
10	.0000

Example 4.6 *:Bridge.* In an evening of bridge, South receives two or more aces on five of nine hands. Do his/her opponents have reason to complain that he/she was just lucky? The question may seem complicated, but is quite simple if approached properly. First the probability that South receives two or more aces on any given hand is

$$p = \frac{\binom{4}{2}\binom{48}{11} + \binom{4}{3}\binom{48}{10} + \binom{4}{4}\binom{48}{9}}{\binom{52}{13}} = \cdots .$$

So, the probability that he/she receives two or more aces on at least five of nine independent hands is

$$\sum_{x=5}^{9} \binom{9}{x} p^x (1-p)^{n-x} = \cdots ,$$

and the answer is, "No, South has not been especially lucky." ♢

The probability mass function (4.12) is called *the binomial probability mass function with parameters n and p*, and denoted by $b_{n,p}(x)$. A graph is included below.

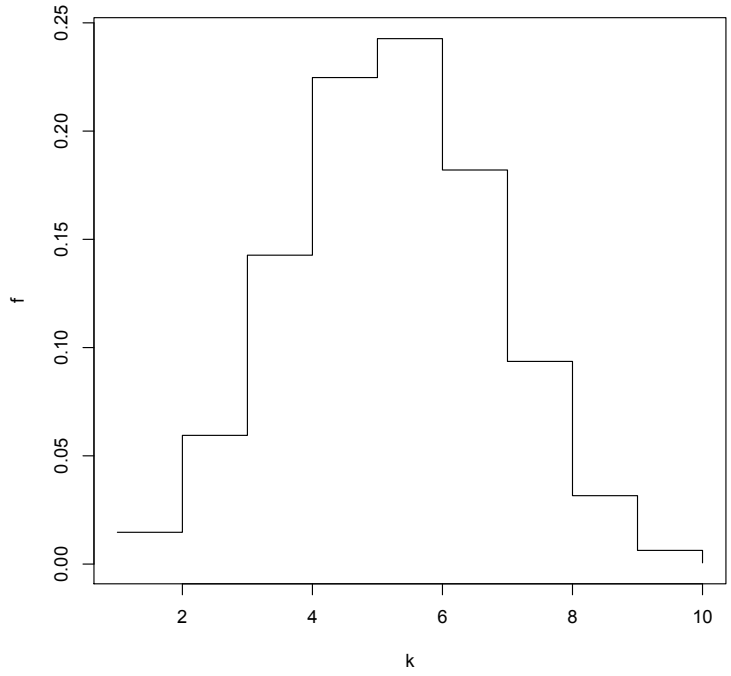

Figure 4.1: The binomial probability mass function with $n = 10$ and p = 9/19

The moment generating function of the binomial is

$$M(t) = \left[1 + p(e^t - 1)\right]^n . \tag{4.13}$$

For

$$\sum_{x=0}^{n} e^{tx} b_{n,p}(x) = \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} = \left[pe^t + 1 - p\right]^n ,$$

which is the right side of (**??**). Differentiation then yields $M'(t) = npe^t[1+p(e^t-1)]^{n-1}$ and $M''(t) = npe^t[1 + p(e^t - 1)]^{n-1} + n(n-1)p^2e^{2t}[1 + p(e^t - 1)]^{n-1}$; and then setting $t = 0$ yields

$$\mu = M'(0) = np$$

and

$$\sigma^2 = M''(0) - M'(0)^2 = [np + n(n-1)p^2] - (np)^2 = np(1-p),$$

Poisson Distirbutions.In the first instance, Poisson distributions arise as limits of binomial distributions

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

when n is large and p is small. The derivation depends on the following two representations of the exponential function: for all $x \in \Re$,

$$e^x = \lim_{n\to\infty} (1 + \frac{x}{n})^n \tag{4.14}$$

and

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}. \tag{4.15}$$

Here is the main result: *If $n \to \infty$ and $p = p_n \to 0$ in such a manner that $\lambda = np$ remains constant, then*

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda} \tag{4.16}$$

for $k = 0, 1, 2, \cdots$. To see why (4.16) holds, write $p = \lambda/n$ and

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{(n)_k!}{k!} (\frac{\lambda}{n})^k (1 - \frac{\lambda}{n})^{n-k} = \frac{1}{k!} \frac{(n)_k}{n^k} \lambda^k (1 - \frac{\lambda}{n})^{n-k}.$$

As $n \to \infty$, the terms on the right converge to $1/k!$, 1, λ^k, and $e^{-\lambda}$, using (**??**). The result (4.16) follows. Observe that

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{\lambda} e^{-\lambda} = 1.$$

So, the function f defined by

$$f(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \tag{4.17}$$

for $k = 1, 2, \cdots$ is a probability mass function, called the *Poisson probability mass function with parameter* λ.

Less formally Equation (4.16) asserts that if n is large, p is small, and $\lambda = np$ is moderate, then the binomial probabilities may be approximated by the right side of (3). Thus, the Poisson distribution provides a good model for the number of occurrences of a large number (n large) of improbable (p small) events and is useful in problems involving accidents, coincidences, mistakes, etc. $\cdots$.

Example 4.7 . Suppose that the probability of winning \$5000 in a lottery is $p = .0001$. If 30000 people play on a given day, what is the probability that more than 5 win. The number of winners X has a Poisson distribution with parameter $\lambda = 30000 \times .0001 = 3$. Thus, the probability that exactly two people win is $P\{X = 2\} = (3^2/2)e^{-3} = .224\cdots$, and the probability that more than five do is

$$P\{X > 5\} = 1 - P\{X \leq 5\} = 1 - \sum_{k=0}^{5} \frac{3^k}{k!} e^{-3} = .086\cdots .$$

Example 4.8 .A book has an average of 1.5 typographical error per page. What is the probability that there are more than two errors on a given page. Let X be the number of errors on the given page. Then it is reasonable to suppose that X has the Poisson distribution with $\lambda = 1.5$, since there are many words on each page and each has a small probability of being in error. Then

$$P\{X \leq 2\} = e^{-1.5} + (1.5)e^{-1.5} + \frac{(1.5)^2}{2} e^{-1.5} = .251\cdots$$

$$P\{X > 2\} = 1 - P\{X \leq 2\} = .191\cdots .$$

Moments. Since Poisson distributions are limits of binomial distributions, and since the mean of a binomial distribution is np, it seems clear that the mean of a Poisson distribution is

$$\mu = \lambda. \tag{4}$$

This may be seen analytically. The moment generating function of a Poisson distribution is easily derived; for

$$M(t) = \sum_{k=0}^{\infty} e^{tk} \frac{1}{k!} \lambda^k e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{1}{k!} (\lambda e^t)^k = e^{-\lambda} e^{-\lambda e^t}.$$

for all $t \in \Re$. That is,

$$M(t) = e^{\lambda(e^t - 1)}, \ \forall \ t \in \Re.$$

In this case, the cumulant function is $\kappa(t) = \log[M(t)] = \lambda(e^t - 1)$, and the derivatives of κ are $\kappa'(t) = \lambda e^t$, $\kappa''(t) = \lambda e^t$, etc.$\cdots$. Thus, the mean is $\mu = \kappa'(0) = \lambda$, as asserted in (4), and the variance is $\sigma^2 = \kappa''(0) = \lambda$.

Radioactive Decay. Suppose that a radio-active substance, such as Carbon-14, is observed with a Geiger counter for a given time period, say t time units. Let X_t denote the number of disintegrations. Then each unstable atom has a small probability, say p_t, of decaying, and there are many atoms, say N. In this case, $p_t = ct$, where c depends on the substance. So, $Np_t = Nct = \lambda t$, say, where $\lambda = Nc$, and it is reasonable to suppose that X has aPoisson distribution with parameter λt–that is,

$$P\{X = k\} = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \ \forall \ k = 0, 1, \cdots .$$

Derivation. First observe that for fixed k,

$$\frac{(n)_k}{n^k} = \frac{n \times (n-1) \times \cdots (n-k+1)}{n \times n \times \cdots \times n} \to 1,$$

as $n \to \infty$. Next, recall that $\lambda = np$ is fixed and write $p = \lambda/n$. So,

$$\begin{aligned} \binom{n}{k} p^k (1-p)^{n-k} &= \frac{1}{k!} \times \frac{(n)_k}{n^k} (np)^k (1-p)^{n-k} \\ &= \frac{1}{k!} \times \frac{(n)_k}{n^k} \lambda^k (1 - \frac{\lambda}{n})^n \times (1 - \frac{\lambda}{n})^{-k}. \end{aligned}$$

As $n \to \infty$ and $p \to 0$ with $\lambda = np$ fixed, the terms in the last line converge to

$$\frac{1}{k!} \times 1 \times \lambda^k \times e^{-\lambda} \times 1,$$

as asserted.

Negative Binomial Distributions. Suppose that a gambler plays a series of games winning each with probability p and losing with probabilility $q = 1 - p$ Let A_k be the event that the gambler wins the k^{th} game. Then$A_1, A_2 \cdots$, areindependent and $P(A_k) = p$ for all k. How many games will the gambler have to play before he/she wins? Let Y denote the number of games required. Thus, Y is the smallest n for which A_n occurs. The event that $Y = 1$ is simply A_1. Similarly, the event that $Y = 2$ is $A_1^c \cap A_2$, and the event that $Y = k$ is $\{Y = n\} = A_1^c \cap \cdots \cap A_{n-1}^c \cap A_n$. So, the probability mass function of Y is

$$f_{\mathbf{y}}(n) = P[Y = n] = P(A_1^c \cap \cdots \cap A_{n-1}^c \cap A_n.) = PA_1^c) \times \cdots \times PA_{n-1}^c) P(\cap A_n) = q^{n-1} \times p.$$

For example, if $p = 9/19$, then $P[Y = 3] = p \times q^2 = \cdots$. Define g by

$$g(n(= p \times q^{-1}$$

for $n = 1, 32 \cdots$ and $g(x) = 0$ for other values of x. Then g is a probability mass function, since $g(x) \geq 0$ for all x, and

$$\sum_{n=1}^{\infty} g(n) = p \sum_{n=1}^{\infty} q^{n-} = p \times \frac{1}{1-q} = 1;$$

and g is called the *geometric probability mass function with parameter p*

This simple calculation can be generalized. Let $A_1, A_2, \cdots$ be independent events with probbiliity $P(A_k) = p$ for all k; regard the occucence of A_k a success and the index k at time; and let

$$X_n = \sum_{j=1}^{n} \mathbf{1}_{A_j},$$

the number of successes by time n. For a given $r \geq 1$, let Y_r be the smallest n for which $X_n = r$, the number of trials required to obtain r successes. The probability mass function of Y_r is derived below, but first an example.

Example 4.9 . The ABD Corporation needs to hire three new engineers. From past experience, it knows that each interview leads to a success hire with probability $p = .3$ What is the probability that exactly nine interviews are required to hire three engineers? The event in question requires success on the ninth interview and exactly two successes on the first eight. So,

$$P[Y_3 = 9] = P[X_8 = 2\text{and } A_9] = P[X_8 = 2]P(A_9) = \binom{8}{2} p^2 q^6 \times p = \binom{8}{2} p^3 q^6$$

For a general $r \geq 1$, the event that $Y_r = n$ requires success on the n^{th} trial and exactly $r - 1$successes on the first $n - 1$ trials. So,

$$P[Y_r = n] = P(\{X_{n-1} = r-1\} \cap A_n) = P[X_{n-1} = r-1] \times PA_n) = \binom{n-1}{r-1} p^r qn - r.$$

The last example provides a special case with $r = 3$ and $n = 9$. Define g by

$$g(n) = \binom{n-1}{r-1} p^r q^{n-r}, \tag{4.18}$$

for $n = r, r+1, \cdots$. and $g(x) = 0$ for other values of x. Then $g(y) = P[Y = y]$ for all y, but it is not immediately clear that g is a mass function, because there is the apparent possility that $X_n < r$ for all n, in which case Y is undefined. It is clear, however, that $\{Y \leq n\} \cup \{X_n < r\} = \Omega$ for all n. So,

$$\sum_{k=r}^{n} g(k) + \sum_{j=0}^{r-1} \binom{n}{j} p^j q^{n-j} = P[Y \leq n] + P[X_n < r] = 1 \tag{4.19}$$

for every $n > r$. Moreover, $\lim_{n\to\infty} P[X_n < r] = 0$, becuse $\binom{n}{j} p^j q^{n-j} \leq n^j q^n \times (p/q)^j \to 0$ as $n \to \infty$ for all $j \geq 0$. Letting $n \to \infty$ in (4.19) then shows tha

$$\sum_{k=r}^{\infty} g(k) = 1 \tag{4.20}$$

so that g is a valid probability mass function. It is called the *negative binomial with parameters p and r*. In the special case that $r = 1$, the negative binomial reduces to the geometric. For the moment generating function, consider a t for which $pe^t <$ ‘1 and let $q_t = qe^t$and $p_t = 1 - q'$. Then

$$M(t) = \sum_{n=r+1}^{\infty} e^{nt} g(n) = e^{rt} \sum_{n=r}^{\infty} \binom{n-1}{r-1} p^r (qe^t)^{n-r} = (\frac{pe^t}{p_t})^r \sum_{n=r}^{\infty} \binom{n-1}{r-1} p_t^r (q_t)^{n-r}$$

The last sum here onwe ibecsause it is the sum of the negative binomial probability mass function with parameter p_t and r. So,

$$M(t) = (\frac{pe^t}{p_t})^r = (\frac{pe^t}{1 - qe^t})^r \tag{4.21}$$

The mean and variance of the a negative binomial distribution

$$\mu = \frac{r}{p} \qquad \text{and} \qquad \sigma^2 = \frac{rq}{p^2}. \tag{4.22}$$

can then be obtained by differentiation.

4.4 Problems and Complements

Problems

1. Let X be the absolute difference between the number of spots that appear when two balanced (six-sided) dice are tossed (larger less smaller). Represent X as a function on an appropriate sample space and find the probability mass function of X.

2. Suppose that two tickets are drawn without replacement from a box containing ten tickets labeled "1,2,$\cdots$,10." Let X be the absolute difference between the numbers on the two tickets. Represent X as a function on an appropriate sample space and find the probability mass function of X.

3. If X has the geometric distribution with parameter $0 < p < 1$, find $P\{X\text{is even}\}$ in terms of p.

4 For what value of c does $p(k) = c/k^2,\ k = 1, 2, \cdots$ and $p(a) = 0$ for other values of a define a probability mass function?

5. Let $p(k) = 1/k(k+1)$ for $k = 1, 2, \cdots$ and $p(a) = 0$ for other values of a. Show that p is a probability mass function.

6. Find the mean and variance of X in Problem 1. *Ans: 1.833$\cdots$ and 1.091$\cdots$.*

7. Find the mean and variance of X in Problem 2.

8. In an evening of bridge South plays eight hands. What is the probability that he/she receives no aces on exactly four of those hands; on at least four hands?

9. There were 1095 marriages in a certain town last year. Let X denote the number of couples whose birthdays fall on the same day of the year. Find $P\{X = 3\}$ and $P\{X \geq 3\}$. What assumptions are you making? *Ans: .224 and .577.*

10. The probability of winning a prize in a lottery is 1/100. If a person plays every day for a year, what is the probability that he/she wins exactly three prizes; at least three?

11. Show that if X has the geometric distribution with parameter $0 < p < 1$, then X has distribution function $F(x) = 1 - q^{\lfloor x \rfloor}$, where $q = 1 - p$ and $\lfloor x \rfloor$ denotes the smallest integer that is less than or equal to X.

12. Tickets are drawn without replacement from a box containing R red tickets and $N - R$ white tickets. Let X denote the number of the draw on which the first

red ticket appears. Find the distribution function of X and use (1.?) to compute its probability mass function.

Chapter 5

Distribution Functions and Densities

5.1 Distribution Functions

If X is any random variable, then the *distribution function of* X is defined by

$$F_{\mathbf{x}}(x) = P[X \leq x], \ \forall \ a \in \mathbb{R}. \tag{5.1}$$

fpr $x \in \mathbb{R}$. The notation her is (very) case sensitive: X is the random variable whose distribution function is defined by (5.1). x is a symbol used in the definition. It could be changed without affecting the meaning of (5.1). For example, the equation ,$F_{\mathbf{x}}(t) = P[X \leq t]$ for $t \in \mathbb{R}$, has the same meaning as (5.1). Unsupisingly, the concept is simplest in the discrete case. If X is discrete with possible values $\mathcal{X} = \{x_1, x_2, \cdots\}$, then

$$F(a) = \sum_{i:x_i \leq a} f_{\mathbf{x}}(x_i), \ \forall \ a \in \mathbb{R} \tag{5.2}$$

where $f_{\mathbf{x}}$ denotes the probability mass functions of X (that is, $f_{\mathbf{x}}(x_i) = P[X = x_i]$ for $i = 1, 2, \cdots$), by (4.1.?).

Example 5.1 Suppose that X has the discrete uniform distribution with parameter n, so that $p(a) = 1/n$ for $a = 1, \cdots, n$ and $p(a) = 0$ for other values of a. If $1 \leq a \leq n$, then $F_{\mathbf{x}}(a) = P\{X \leq a\} = \sum_{i \leq a}(1/n) = \lfloor a \rfloor / n$, where $\lfloor a \rfloor$ denotes the

greatest integer which is less than or equal to a. Also, $F(a) = 0$ for $a < 1$, since the event $\{X < 1\}$ is impossible, and similarly, $F(a) = 1$ for $a > n$. $\diamondsuit$

Here are two examples of a different nature. Recall the notation for intervals: $(a, b) = \{x \in \mathbb{R} : a < x < b\}$, $(a, b] = \{x \in \mathbb{R} : a < x \leq b\}$, $[a, b) = \{x \in \mathbb{R} : a \leq x < b\}$, and $[a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$ for $-\infty \leq a \leq b \leq \infty$.

Example 5.2 . Let $\Omega = (-\pi, \pi]$ and $P((a, b]) = (b - a)/2\pi$ for $-\pi < a < b \leq \pi$, as in Example 2.3?.

a). If $X(\omega) = \omega$ for $\omega \in \Omega$, then the possible values of X are $\mathcal{X} = (-\pi, \pi]$, and $P\{a < X \leq b\} = P((a, b]) = (b - a)/2\pi$ for all $-\pi < a \leq b \leq \pi$. In particular, $P\{X = a\} = (a - a)/2\pi = 0$ for all $-\pi < a \leq \pi$, so that the notion of a probability mass functions is not useful concept. The distribution function may be computed, however: if $-\pi < a \leq \pi$, then

$$F_{\mathbf{x}}(a) = P\{X \leq a\} = P((-\pi, a]) = \frac{a + \pi}{2\pi}.$$

Also, $F(a) = 0$ for $a \leq -\pi$, since the event $\{X \leq -\pi\}$ is impossible, and $F(a) = 1$ for all $a \geq \pi$.

b). Let $Y(\omega) = \tan(\omega)$, if ω is not a multiple of $\pi/2$, and let $Y(\omega) = 0$ otherwise. Then Y has distribution function

$$F_{\mathbf{y}}(y) = \frac{1}{2} + \frac{1}{\pi} \arctan(y), \ -\infty < y < \infty, \tag{5.3}$$

This distribution function is called the *standard Cauchy.*

To see (5.3) it is convenient to consider the cases $y > 0$ and $y < 0$ separately. If $y > 0$, then the event that $\{Y \leq y\}$ is $\{\omega : Y(\omega) \leq y\} = (-\pi, -\pi + a] \cup -\frac{\pi}{2}, a] \cup [\frac{\pi}{2}, \pi]$ where $a = \arctan(y)$, the solution to $\tan(\omega) = y$ for which

$$\begin{aligned} P\{Y \leq y\} &= P((-\pi, -\pi + a]) + P([-\frac{\pi}{2}, a]) + P([\frac{\pi}{2}, \pi]) \\ &= \frac{a}{2\pi} + \frac{a + \pi/2}{2\pi} + \frac{\pi/2}{2\pi} \\ &= \frac{1}{2} + \frac{1}{\pi} \arctan(y). \end{aligned}$$

The same final answer is obtained for negative y, though the picture is a little different. Another derivation is included in Problem ?.?. ♢

Computing Probabilities from Distribution Functions. If X has distribution function F, then probability that X falls in an interval, $(a, b]$ say, can be computed directly from the distribution function as

$$P\{a < X \leq b\} = F(b) - F(a), \ \forall \ b \in \mathbb{R}. \tag{3}$$

To see this, simply observe that $F(b) = P\{X \leq b\} = P\{X \leq a\} + P\{a < X \leq b\} = F(a) + P\{a < X \leq b\}$ for all $a < b$. For an example, suppose that Y has the standard Cauchy distribution G of (2), then $P\{0 < Y \leq 1\} = G(1) - G(0) = \cdots$.

An arbitrary choice was made in (1) by defining $F(a)$ to be $P\{X \leq a\}$, instead of $P\{X < a\}$. With the definition (1), it may be shown that

$$P\{X < b\} = F(b-) = \lim_{a \to b, a < b} F(a) \tag{5.4}$$

for all $b \in \mathbb{R}$. See Section 5.? for the proof. The probability that X falls in any subinterval may then be computed by retracing the derivation of (3). For example, $P\{a \leq X \leq b\} = P\{X \leq b\} - P\{X < a\} = F(b) - F(a-)$ for all $a \leq b$. An interesting consequence of this formula is that

$$P\{X = a\} = F(a) - F(a-), \ \forall \ a \in \mathbb{R}.$$

So, F is continuous at a given a iff $P\{X = a\} = 0$. Discrete uniform distribution functions illstrate this point. They are step functions with discontinuities at integer values. In particular, if F is continuous at a and b, then $P\{a \leq X \leq b\} = P\{a < X \leq b\} = P\{a \leq X < b\} = P\{a < X < b\}$. Finally, if X has distribution function F, then $P\{X > b\} = 1 - P\{X \leq b\} = 1 - F(b)$ and $P\{X \geq b\} = 1 - F(b-)$ for all $b \in \mathbb{R}$.

Characteristic Properties. As detailed in section 5.4, distribution functions have certain characteristic properties. If F is the distribution function of random variable, X say, then: F is non-decreasing; that is, $F(a) \leq F(b)$ when $a \leq b$; F is continuous from the right; that is, $F(a) = \lim_{b \to a, b > a}, \ \forall \ a \in \mathbb{R}$; also $\lim_{x \to -\infty} F(x) = 0$ and

$\lim_{x\to\infty} F(x) = 1$. Conversely, any such function is the distribution function of some random variable. The proofs of thes assertions are deferred to Sections 5.4 and 5.5. Any function F for which *a), b)*, and *c)* hold is called a *distribution function*, since it is then the distribution function of some random variable. Here is an example to illustrate the use of the conditions.

Example 5.3 *: Exponential Distributions.* If $0 < \lambda < \infty$, then the function F defined by

$$F(x) = \begin{cases} 0 & if\ x \leq 0 \\ 1 - e^{-\lambda x} & if\ x > 0 \end{cases}$$

is a distribution function. For F satisfies *a), b)*, and *c)* above, as is easily checked by drawing the its graph, Figure 5.3. ♢

5.2 Densities

If a distribution function F, say, is differentiable, then its derivative

$$f(x) = F'(x) = \frac{d}{dx}F(x) \tag{5.5}$$

is of interest. For example, if X is a random variable with distribution F, then

$$P\{a < X \leq b\} = F(b) - F(a) = \int_a^b f(x)dx \tag{5.6}$$

for all $-\infty < a < b < \infty$, by the Fundamental Theorem of Calculus. This formula may be regarded as a continuous analogue of Equation (4.6). Letting $a \to -\infty$ in (5.6) and using c) of Section 1,

$$F(b) = \int_{-\infty}^b f(x)dx,\ \forall\ b \in \mathbb{R}, \tag{5.7}$$

so that F may be recovered from f.

Letting $b \to \infty$ in (5.7) and using charactersitic Property *c)* of the previous section, $\lim_{b\to\infty} F(b) = 1$,

$$\int_{-\infty}^{\infty} f(x)dx = 1, \tag{5.8}$$

and any non-negative function f for which (5.8) holds is called a *density*. Thus the derivative of any differentiable distribution function is a density. If X is a random variable with a distribution function F of the form (3) then X and F are said to be *absolutely continuous with density* f. There is a converse: if f is a density, then the function defined by (5.7) is a distribution function–that is satisfies a), b), and c) of Section 1. For any such F is continuous and non-decreasing; and c) follows easily from (4). So, *if f is any density, then there is a random variable X with density f.* This allows modeling directly in terms of random variables, distribution functions, and densities, avoiding sample spaces. Examples 1.2, 1.3, and 1.4 were of this nature.

It is convenient to write $X \sim F$ (read "X is distributed as F"), when X is a random variable with distribution function F–that is, when $F = F_{\mathbf{x}}$..

Example 5.4 *: Uniform Densities.* If $-\infty < \alpha < \beta < \infty$, then the function

$$f(x) = \begin{cases} 1/(\beta - \alpha) & if\ \alpha < x \leq\ \beta \\ 0 & else \end{cases}$$

is a density. For the graph of f is a rectangle, and the area under the graph is one. The distribution function corresponding to f though (3) is

$$F(x) = \begin{cases} 0 & if\ x \leq \alpha \\ (x - \alpha)/(\beta - \alpha) & if\ \alpha < x \leq \beta \\ 1 & if\ x > \beta \end{cases}$$

by simple integration. This distribution function and density are called *uniform with parameters α and β*; and when $\alpha = 0$ and $\beta = 1$, it is called the *Standard Uniform.* Example **??**. a special case with $\alpha = -\pi$ and $\beta = \pi$. Then $f(x) = 1/2\pi$ for $-\pi < x \leq \pi$. The standard uniform density is displayed in Figure 5.4 $\diamondsuit$

Example 5.5 *: Exponential Densities* If F is the exponential distribution function with failure rate $\lambda > 0$, so that $F(x) = 0$ for $x \leq 0$ and $F(x) = 1 - e^{-\lambda x}$ for $x \geq 0$, then

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & if 0 \leq x < \infty, \\ 0, & otherwise, \end{cases}$$

again using (1). ♢

Example 5.6 *: The Cauchy Density.* The standard Cauchy distribution function is $G(y) = 1/2 + (1/\pi)\arctan(y),\ \forall\ y \in \mathbb{R}$; and G has derivative

$$g(y) = \frac{1}{\pi(1+y^2)},\ \forall\ y \in \mathbb{R}.$$

Example 5.7 *: Bilateral Exponential Densities.* For any $\lambda > 0$, the function f defined by

$$f(x) = \frac{1}{2}\lambda e^{-\lambda|x|} \tag{5.9}$$

for $x \in \mathbb{R}$ is a density, because $f(x) \geq 0$ for all $x \in \mathbb{R}$, and

$$\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{\infty} \lambda e^{-\lambda x}dx = -e^{-\lambda x}|_0^{\infty} = 1.$$

At a technical level, the term "differentiable" in (1) means piecewise continuously differentiable; that is, the function F must be continuous and continuosly differentiable, except possibly at a finite number of points, where the derivative may fail to exist. The function f may be defined arbitrarily at this finite number of points, subject only to the condition $f(x) \geq 0$. Examples 1 and 2 are of this nature. In Example, the derivative fails to exist at $x = \alpha$ and $x = \beta$ and it Example 2, it fails to exist at $x = 0$.

Densities and Mass Functions. There are strong analogies between densities (of absolutely continuous distributions) and probability mass functions (of discrete ones). For example, (1) is a continuous analogue of (1.2). There is also an important difference: the values of probability mass functions are probabilities; those of densities are derivatives of probabilities. To illustrate the latter point, let X be absolutely continuous with distribution function F and density f. Then $P\{X = a\} = 0$ for all $a \in \mathbb{R}$, since F must be a continuous function. The probability that X is close to a is related to $f(a)$, however. To see this let $h > 0$ be small. Then $P\{a < X \leq a + h\} = F(a + h) - F(a)$; and $f(a) = F'(a)$, then $F(a + h) - F(a) \approx f(a)h$, the tangent approximation to the graph of F. So,

$$P\{a < X \leq a + h\} \approx f(a) \times h \tag{5.10}$$

for small h.

Failure Rates. Now let X denote a positive random variable and regard X as the time until failure of a mechanical device, or the lifetime of a biological organism. If $t, h > 0$ and $P\{X > t\} > 0$, then the conditional probability that the device fails during the time interval $[t, t+h]$ given that it is still operating at time t is $P\{t < X \leq t+h | X > t\} = P\{t < X \leq t+h\}/P\{|X > t\}$. Letting F denote the distribution function of X, so that $P\{t < X \leq t+h\} = F(t+h) - F(t)$ and $P\{X > t\} = 1 - F(t)$, the conditional probability may be written

$$P\{t < X \leq t+h | X > t\} = \frac{F(t+h) - F(t)}{1 - F(t)}.$$

If h is small, then (6) may be used to approximate the numerator, and

$$P\{t < X \leq t+h | X > t\} \approx \lambda(t)h,$$

where

$$\lambda(t) = \frac{f(t)}{1 - F(t)} \tag{5.11}$$

That is, the probability of failure in a short interval $[t, t+h]$ is approximately $\lambda(t)h$. Here $\lambda(t)$ is called the *failure rate at* t, and the function defined by (7) is called the *failure rate*.

A distribution function is uniquely determined by its failure rate. To see this first observe that

$$\lambda(t) = \frac{d}{dt} \log \left[\frac{1}{1 - F(t)}\right],$$

by the chain rule. Since $F(0) = P\{X \leq 0\} = 0$, it then follows that $\log[1 - F(t)] = -\int_0^t \lambda(s)ds$ and, therefore, that

$$F(t) = 1 - \exp[-\int_0^t \lambda(s)ds] \tag{5.12}$$

for all $t > 0$ for which $F(t) < 1$.

Example 5.8 *: Exponential Densities* If F is the exponential distribution, $F(t) = 1 - e^{-\lambda t}$ for $t \geq 0$, then $f(t) = \lambda e^{-\lambda t}$ and

$$\lambda(t) = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$$

for all $0 < t < \infty$. That is, exponential distributions have constant failure rates. Conversely, any distribution function F with a constant failure rate is an exponential distribution. For, if $\lambda(t) = \lambda > 0$ for all $0 < t < \infty$, then $\int_0^t \lambda(s)ds = \lambda t$ and, therefore, $F(t) = 1 - e^{-\lambda t}$ for $t \geq 0$, by (5.12).

Example 5.9 *: The Rayleigh Distribution.* Suppose that the lifetime of a given device in years has failure rate $\lambda(t) = t$. What is the probability that the device last more than one year before failing. Let X denote the lifetime and let F denote its distribution function. Then $\int_0^t \lambda(s)ds = \int_0^t sds = t^2/2$ and, therefore,

$$F(t) = 1 - e^{-\frac{1}{2}t^2}, \ \forall \ t > 0, \tag{5.13}$$

by (8). So, $P\{X > 1\} = 1 - F(1) = e^{-\frac{1}{2}} = .607$. The distribution function in (9) is called the *standard Rayleigh distribution function* and is a special case of the *Weibull distributions*, described in Problems $\cdots$. $\diamondsuit$

—

Means f is a probability density function and X is a random variable with density f, then the *mean* of X is defined by

$$\mu = \int_{-\infty}^{\infty} xf(x)dx, \tag{5.14}$$

provided that the integral converges absolutely. Thus the mean of a random variable depends only of its density; and the expectation of X may also be called the *mean of* f. The mean provides one notion of the center of a distribution.

Example 5.10 *Examples 1. a): Uniform.* If f is uniform on an interval $(\alpha, \beta]$, then

$$\mu = \int_{\alpha}^{\beta} \frac{x}{\beta - \alpha}dx = \frac{\beta^2 - \alpha^2}{2(\beta - \alpha)} = \frac{\alpha + \beta}{2},$$

the midpoint of the interval.

b): Exponential. If f is exponential with failure rate $\lambda > 0$, then $f(x) = \lambda e^{-\lambda x}$ for $0 \leq x < \infty$ and, therefore,

$$\mu = \int_0^{\infty} x\lambda e^{-\lambda x}dx = -xe^{-\lambda x}\big|_{x=0}^{\infty} + \int_0^{\infty} e^{-\lambda x}dx = -\frac{1}{\lambda}e^{-\lambda x}\big|_{x=0}^{\infty} = \frac{1}{\lambda}.$$

In words, the mean of an exponential distribution is the reciprocal of the failure rate.
$\diamondsuit$ *endex*

Example 5.11 *: Symmetric Distributions.* If f is a bilateral exponential density, say $f(x) = \frac{1}{2}\lambda e^{-\lambda|x|},\ x \in \mathbb{R}$, where $\lambda > 0$, then f is symmertic; that is, $f(-x) = f(x)$ for all $x \in \mathbb{R}$. It then follows that $\mu = 0$ in (1). The mean of any symmetric density is zero, provided that the integral in (1) converges. The standard Cauchy density $f(x) = 1/\pi(1 + x^2),\ x \in \mathbb{R}$, is also symmetric, but in this case the mean is not defined. See Problem ?.?.

Variances. *If f is a density with mean μ,say, then the variance of f is defined by*

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx. \tag{5.15}$$

In this case, the variance is to be interpreted as ∞ if the integral does not converge. If X has density f, then σ^2 is called the variance of X and may be denoted by $Var(X)$ or σ_X^2. The square root $\sigma = \sqrt{\sigma^2}$ of the variance is called the standard deviation of f or of X. The mean may be regarded as the center of the distribution, and the variance and standard deviation measure the tendency of X to deviate from μ. For a physical analogy, imagine a long thin wire with phsyical mass density f. Then μ is the center of gravity and σ^2 is the moment of inertia. It is covenient to develop Equation (3) below, before considering examples. **Moments**. *If X has density f, then the moments of X, or of f are defined by*

$$\mu_k = \int_{-\infty}^{\infty} x^k f(x) dx, \tag{5.16}$$

provided that the integral converges absolutely. Thus, $\mu_1 = \mu$ is the mean. As in the discrete case, the variance may be recovered from the mean μ and the second moment μ_2 by the simple formula

$$\sigma^2 = \mu_2 - \mu^2, \tag{5.17}$$

provided that μ_2 is finite. The proof is similar to the proof of (4.2.?) and is left as an exercise.

Example 5.12 *:a): Exponential.* If f is the exponential density with failure rate λ, then

$$\mu_2 = \int_0^\infty x^2 \lambda e^{-\lambda x} dx = x^2 e^{-\lambda x}\Big|_{x=0}^\infty + \int_0^\infty 2xe^{-\lambda x} dx = \frac{2}{\lambda^2},$$

integrating by parts as in Example 1-b). So, $\sigma^2 = 2\lambda^{-2} - \lambda^{-2} = \lambda^{-2}$.

b): Uniform. If f is uniform on $(\alpha, \beta]$, then

$$\mu_2 = \int_\alpha^\beta \frac{x^2 dx}{\beta - \alpha} = \frac{\beta^3 - \alpha^3}{3(\beta - \alpha)} = \frac{\alpha^2 + \alpha\beta + \beta^2}{3}$$

and

$$\sigma^2 = \frac{\alpha^2 + \alpha\beta + \beta^2}{3} - \frac{\alpha^2 + 2\alpha\beta + \beta^2}{4} = \frac{(\beta - \alpha)^2}{12},$$

after some simple algebra.

Moment Generating Functions. *As in the discrete case, the moment generating function of a density f is defined by*

$$M(t) = \int_{-\infty}^\infty e^{tx} f(x) dx$$

for those t for which the integral is finite; and if X is a random variable with density f, then M may be called the moment generating function of X too. As in the discrete case, the moments of f may be computed from M by the formula

$$\mu_k = M^{(k)}(0) := \frac{d^k}{dt^k} M(t)\Big|_{t=0}$$

for all $k = 1, 2, \cdots$, provided that $M(t)$ is finite for all t is some interval containing 0. Examples 4. If f is the exponential density with failure rate λ, then

$$\begin{aligned} M(t) &= \int_0^\infty e^{tx} \lambda e^{-\lambda x} dx \\ &= \lambda \int_0^\infty e^{-(\lambda - t)x} dx \\ &= -\frac{\lambda}{\lambda - t} e^{-(\lambda - t)x}\Big|_{x=0}^\infty = \frac{\lambda}{\lambda - t} \end{aligned}$$

for all $t < \lambda$. The mean and variance may be recovered by differentiation. $\diamondsuit$

5.3 Induced Distributions

Let X denote a random variable $\mathcal{X} \subseteq \mathbb{R}$ a set for which $P[X \in \mathcal{X}] = 1$, and $w : \mathcal{X} \to \mathbb{R}$ a function defined on $\mathcal{X}$. If Y is a random variable fof the form $Y = w(X)$, then the distribution functions $F_{\mathbf{x}}$ and $F_{\mathbf{y}}$ of X and Y are related. The relationship is describe in this section. Let $\mathcal{Y} = \{w(x) : x \in \mathcal{X}\}$, the range of w. Then clearly $P[Y \in \mathcal{Y}] = P[X \in \mathcal{X}] = 1$. Moe generally, let

$$w^{-1}(B) = \{x \in \mathcal{X} : w(x) \in B\} \tag{5.18}$$

for subsets $B \subseteq \mathcal{Y}$. Thus, w^{-1} maps subsets of the range $\mathcal{Y}$ into subsets of the domain $\mathcal{X}$. Then $Y = w(X) \in B$ if and only if $X \in w^{-1}(B)$ and, therefore,

$$P[Y \in B] = P[X \in w^{-1}(B)], \tag{5.19}$$

for subsets $B \subseteq \mathcal{Y}$ for which the right side is defined. Unsurprizingly, this relation is simplest in the discrete case, when $\mathcal{X}$ is a finite set. Then the probabilry mass functions $p_{\mathbf{x}}$ and $p_{\mathbf{y}}$ are related by

$$p_{\mathbf{y}}(y) = P[Y = y] = P[X \in w^{-1}(\{y\})] = \sum_{x \in w^{-1}(\{y\})} p_{\mathbf{x}}(x). \tag{5.20}$$

That is ,$p_{\mathbf{y}}$ is the sum of $p_{\mathbf{x}}(x)$ over all solutions to the equation $w(x) = y$.

Example 5.13 . Recall that if r is an integer, then any inetger, n say, may be written $n = k \times r + j$, where $0 \leq j \leq r - 1$. The relation between n and $j4$ is written $j = n \pmod r$. If X is an integer valued random variable, so that $\mathcal{X} = \{0, 1, 2, \cdots\}$ and $w(x) = x \pmod r$, then, the the solutions to the equation $w(x) = y$ are $x = nr+y$, where n is a positive integer. So,

$$p_{\mathbf{y}}(y) = \sum_{n=0}^{\infty} p_{\mathbf{x}}(nr + y) \tag{5.21}$$

for $y = 0, \cdots r - 1$. If X has a geometric distribution, so that $p_{\mathbf{x}]}x) = pq^{x-1}$for $x = 0, 1, 2, \cdots$, then the later sum may be computed in closed form as

$$p_{\mathbf{y}}(y) = \sum_{n=0}^{\infty} pq^{nr-1} = (\frac{p}{q})(\frac{1}{1` - q^r}).$$

Next consider the distribution functions.

Next consider the distribution functions. Since the event $\{Y \leq y\}$ *may be written* $\{Y \in (-\infty, y]\}$,

$$F_{\mathbf{y}}(y) = P[Y \leq y] = P[Y \in (-\infty, y]] = P[X \in w^{-1}((-\infty, y])] \tag{5.22}$$

Direct use of (5.22) can be cumbersome, because it requires one to solve the inequality $w(x) \leq y$, *but it can be use uuseful in special cases.*

Two Special Cases. Suppose first that Y *is a linear function of* X, *say* $Y = aX + b$, *where* $a > 0$. *If* $y \in \mathbb{R}$, *then, clearly,* $Y \leq y$ *iff* $X \leq (y - b)/a$. *So, the distribution functions of* X *and* Y *are related by*

$$F_{\mathbf{y}}(y) = P\{Y \leq y\} = P\{X \leq \frac{y-b}{a}\} = F_{\mathbf{x}}(\frac{y-b}{a}) \tag{5.23}$$

for all $y \in \mathbb{R}$. *If* X *has a density* f, *then*

$$f_{\mathbf{y}}(y) = F'_{\mathbf{y}}(y) = \frac{1}{a}F'_{\mathbf{x}}(\frac{y-b}{a}) = \frac{1}{a}f_{\mathbf{x}}(\frac{y-b}{a}), \tag{5.24}$$

at least at continuity points of g. *Observe that the factor* $1/a$ *arises from the differentiation.*

Ffor a second example, suppose that $Y = X^2$. *If* $y > 0$, *then* $Y \leq y$ *iff* $-\sqrt{y} \leq X \leq \sqrt{y}$. *So,*

$$F_{\mathbf{y}}(y) = P\{Y \leq y\} = P\{-\sqrt{y} \leq X \leq \sqrt{y}\} = F_{\mathbf{x}}(\sqrt{y}) - F_{\mathbf{x}}(-\sqrt{y}-), \tag{5.25}$$

where $F_{\mathbf{x}}(x-) = \lim_{z \to x, z<x} F_{\mathbf{x}}(z)$. *If* X *has a density* f, *then* $F_{\mathbf{x}}$ *is continuous, so that* $F_{\mathbf{x}}(-\sqrt{y}-) = F_{\mathbf{x}}(-\sqrt{y})$, *and*

$$f_{\mathbf{y}}(y) = F_{\mathbf{y}}(y) = \frac{d}{dy}[F_{\mathbf{x}}(\sqrt{y}) - F_{\mathbf{x}}(-\sqrt{y})] = \frac{f_{\mathbf{x}}(\sqrt{y}) + f_{\mathbf{x}}(-\sqrt{y})}{2\sqrt{y}}.$$

Observe that if f *is any density, then Equations (5.24)) and generate an entire family of distributions, one distribution for each choice of* a *and* b. *In such cases* a *and* b *may be called scale and location parameters.*

Example 5.14 Suppose that X has the standard Cauchy distribution, with density $f_{\mathbf{x}}(x) = 1/\pi(1 + x^2),\ x \in \mathbb{R}$.

a) If $Y = aX + b$, where $a > 0$, then Y has density

$$f_{\mathbf{y}}(y) = \frac{a}{\pi[a^2 + (y-b)^2]},\ y \in \mathbb{R}.$$

This density is called *Cauchy with location and scale parameters.*

b). If $Z = X^2$, then Z has density

$$f_{\mathbf{y}}(y) = \frac{1}{\pi\sqrt{z}(1+|z|)},\ 0 < z < \infty.$$

Monotone Functions. Here is a generalization of (5.23) and (5.24). Let $\mathcal{X}$ denote an interval for which $P\{X \in \mathcal{X}\} = 1$ and suppose that w is continuous and strictly increasing on $\mathcal{X}$; that is, $w(x_1) < w(x_2)$ whenever $x_1, x_2 \in \mathcal{X}$ and $x_1 < x_2$. Let $\mathcal{Y} = w(I)$ denote the range of w. Then $\mathcal{Y}$ is an interval and w has a well defined inverse function v, defined on $\mathcal{Y}$; that is, $v(y)$ is the unique solution to the equation $w(x) = y$ for each $y \in J$. See Figure 1. For example, if $w(x) = ax + b$, where $a > 0$, then $v(y) = (y-b)/a$. From Figure 1, it is clear that if $y \in J$, $Y = w(X) \leq y$ iff $X \leq v(y)$. So,

$$F_{\mathbf{y}}(y) = P[Y \leq y] = P\{X \leq v(y)\} = F_{\mathbf{x}}[v(y)] \tag{5.26}$$

for all $y \in J$. A similar result holds if w is continuous and decreasing (that is, $w(x_1) > 2(x_2)$ whenever $x_1 < x_2$). Then w has a well defined inverse again, but in this case $Y \leq y$ iff $X \geq v(y)$, so that

$$F_{\mathbf{y}}(y) = P\{Y \leq y\} = P\{X \geq v(y)\} = 1 - F_{\mathbf{x}}[v(y)-][v(y)-] \tag{5.27}$$

for $y \in J$. If X has a piecewise continuous density f and if v is differentiable, the density of Y may be obtained by differentiation. For example, if w is increasing, then Y has density $g(y) = G'(y) = F'[v(y)]v'(y) = f[v(y)]v'(y)$; and if w is decreasing then $g(y) = -f[v(y)]v'(y)$. Since $v' \geq 0$ for increasing w, and $v' \leq 0$ for decreasing w, the two cases can be combined in the simple formula,

$$g(y) = f[v(y)]|v'(y)| \tag{5.28}$$

for $y \in J$ and $g(y) = 0$ for $y \in J^c$

Example 5.15 . If X has the uniform distribution on $(0,1]$, then $f(x) = 1$ for $0 < x \leq 1$ and $f(x) = 0$ otherwise. If $Y = -\log(X)$, then $I = (0,1]$, $w(x) = -\log(x), J = [0,\infty)$, $v(y) = e^{-y}$, and $v'(y) = -e^{-y}$ in (6). So, Y has density

$$f_{\mathbf{y}}(y) = f_{\mathbf{x}}(e^{-y})|-e^{-y}| = e^{-y} \tag{5.29}$$

for $0 < y < \infty$. That is, Y has the standard exponential distribution.

t

Example 5.16 If X has the standard exponential distribution, with density $f_{\mathbf{x}}(x) = e^{-x}$ for $0 < x < \infty$, and if $Y = \sqrt{2X}$, then $I = J = (0,\infty)$, $w(x) = \sqrt{2x}$, $v(y) = y^2/2$, $v'(y) = y$, and

$$f_{\mathbf{y}}(y) = f_{\mathbf{x}}(\frac{y^2}{2})2y = ye^{-\frac{1}{2}y^2}$$

for $0 < y < \infty$. That is, Y has the standard Rayleigh distribution. ♢

Example 5.17 *The Probability Integral Transformation.* If the $F_{\mathbf{x}}$ s continuous and strictly increasing on an interval $\mathcal{X}$ for which $P\{X \in \mathcal{X}\} = 1$, then $F_{\mathbf{x}}(X)$ has the standard uniform distribution. For if $0 < u < 1$, then

$$P[F_{\mathbf{x}}(X) \leq u] = P[X \leq F_{\mathbf{x}}^{-1}(u)] = F_{\mathbf{x}}[F_{\mathbf{x}}^{-1}(u)] = u, \tag{5.30}$$

where F^{-1} denotes the inverse function. In fact, only the continuity is essential, though the proof is harder if F is not strictly increasing. See $\cdots$.

Piecewise Monotone Functions. There is a far reaching generalization of (5.28) and (5.29) A function w is said to be piecewise monotone on an interval (a,b), say, if there is a partition $a = a_0 < a_1 < \cdots < a_m = b$ for which w is strictly increasing or strickly decreasing on each of the subintervals (a_{i-1}, a_i), $i = 1, \cdots, m$. A sufficient condition for this is that $|w'(x)| > 0$ for all $a_{i-1} < x < a_i$ and $i = 1, \cdots, m$. If the latter condition is satisfied, $P\{X \in (a,b)\} = 1$ and $Y = w(X)$, then Y has density

$$f_{\mathbf{y}}(y) = \sum_{x:w(x)=y} f_{\mathbf{x}}(x)\big|\frac{1}{w'(x)}\big|, \ \forall\ y \in \mathbb{R}, \tag{5.31}$$

where an empty sum is to be interpreted as zero. The square function and Example 1-b) illustrate the use of (8). Here is another example. Example 5. Suppose that X is uniformly distributed over $(-\pi, \pi]$ *and that* $Y = \sin(X)$. *Then the conditions are satisfied with* $a = a_0 = -\pi$, $a_1 = -\pi/2$, $a_2 = \pi/2$, *and* $a_3 = b = \pi$. *See Figure ?. In Equation (8),* $f(x) = 1/2\pi$ *for all* $-\pi < x \leq \pi$; *and if* $\sin(x) = y$, *then the derivative of* $\sin(x)$ *is* $\cos(x) = \pm\sqrt{1 - \sin^2(x)} = \pm\sqrt{1 - y^2}$. *So, Y has density*

$$g(y) = \sum_{x:\sin(x)=y} \frac{1}{2\pi} \frac{1}{\sqrt{1-y^2}} = \frac{1}{\pi\sqrt{1-y^2}}$$

for $-1 < y < 1$ *and* $0 < |y| <!$, *since there are two solutions two the equation* $\sin(x) = y$ *for all such y, and* $g(y) = 0$ *for* $|y| > 1$, *since there are no solutions for* $y > 1$.

5.4 Characteristic Properties of Distribution Function

Distribution functions have certain characteristic properties: If F is the distribution of a random variable X, then

a) F is nopn-decreasing;

b) F is right continuous;

c) i) $\lim_{x\to-\infty} F(x) = 0$ *and ii)* $\lim_{x\to\infty} F(x) = 1$.

Conversely, any such function is the distribution function of some random variable. The proof of this assertion is deferred to Section 5.?. Any function F for which a), b), and c) hold is called a distribution function, since it is then the distribution function of some random variable. Properties a), b), and c) are established belwow; but firsthere is an example to illustrate the use of the conditions.

For a), simply observe that if $a < b$, *then* $0 \leq P[a < X \leq b] = F(b) - F(a)$, *so that* $F(a) \leq F(b)$, *So,F is non-decreasing, and therefore, F has one sided limits* $F(x-) = \lim y \to x, y < xF(y)$ *and* $F(x+) = \lim_{y\to x, y>x} F(y)$. *Property b) can be*

restated $F(x+) = F(x)$ for all $x \in \mathbb{R}$; and this is easily seen. Let Ω denote the sample space on which X is defined. Then

$$\{\omega : X(\omega \leq x\} = \bigcap_{n=1}^{\infty} \{\omega : X(\omega) \leq x + \frac{1}{n}\},$$

*because $X(\omega) \leq x$ if and only if $X(\omega) \leq x + \frac{1}{n}$ for every $n = 1, 2, \cdots$. The events $B_n = \{\omega : X(\omega) \leq x + \frac{1}{n}\}$ are decreasing; that is $B_n \supseteq B_{n+1}$ for all n; and therefore,$P \cap_{n=1}^{\infty}$ $B_n) = \lim_{n\to\infty} P(B_n)$, by (**??**). It follows that*

$$F(sx) = P[X \leq x] = P(\bigcap_{n=1}^{\infty} B_n) = \lim_{n\to\infty} P(B_n) = \lim_{n\to\infty} F(x + \frac{1}{n}) = F(x+),$$

*eestablishing b). . The proof of (**??**) is similar. For this, let $B_n = \{\omega : X(\omega) \leq x - \frac{1}{n}\}$, the event that $X \leq x - \frac{1}{n}$. Then the B_n are increasing, $B_n \subseteq B_{n+1}$, and $\cup_{n=1}^{\infty} B_n = \{\omega : X(\omega) < x\}$, since $X(\omega) < x$ if and only if $X(\omega \leq x - \frac{1}{n}$ for some n. So,*

$$P[X < x] = P(\bigcup_{n=`1}^{|\infty} B_n) = \lim_{n\to\infty} P(B_n) = \lim_{n\to\infty}]P[X \leq x - \frac{`1}{n}] = \lim_{n\to\infty} F(x - \frac{1}{n}) = F(x-).$$

The prof of c) i also similar. For c-ii), first observe that $\lim_{n\to\infty}$ exists because F is non-decreasing. Then let $B_n = \{\omega : X(\omega) \leq n\}$, the event that $X \leq nIn$ this case the B_n are increasing, and $\cup_{n=1}^{\infty} B_n = \Omega$. since for every ω there is an n for which $X(\omega) \leq n$. So,

$$\lim_{x\to\infty} F(x) = \lim_{n\to\infty} F(n) = \lim_{n\to\infty} P(B_n) = p\left(\bigcup_{n=1}^{\infty} B_n\right) = P(\Omega) = 1,$$

as assertd. The proof of c-i) is left as an exercise.

An important consequence of a) is that A distribution function F can have a most countably many point o f discontinuity;; that is, the set

$$D_F = \{x \in \mathbb{R} : F(x-) \neq F9x0\}$$

is either finite or countably infinite. To see this first observe that $F(x-) \leq F(x)$ for all x, since $F(y) \leq F(x)$ for all $y < x$, and therefore $D_F = \{x \in \mathbb{R} : F(x) - F(x-) > 0\}$. Next, let

$$D_{F,m} = \{X \in \mathbb{R} : F(x) - F(x-) \geq \frac{1}{m}\}$$

for $m = 1, 2, \cdots$. *Then*

$$D_F = \bigcup_{M=1}^{\infty} D_{F,m}.$$

If $x_1, \cdots, x_n$ *are any* n *point in* $D_{F,m}$*, then* $x_1, \cdots, x_n$ *a can be so labelled that* $x_1 < \cdots, < x_n$ *in which case*

$$\frac{n}{m} \leq \sum_{i=2}^{n} [F(x_i) - F(x_i-)] \leq \sum_{i=2}^{n} [F(x_i) - F(x_{I-1}] = F(x_n) - F(x_1) \leq 1$$

and therefore $n \leq m + 1$*. It follows that* $\#D_{F,m} \leq m + 1$*; that is, each* $D_{F,m}$ *is a finite set. That* D_F *is countable then follows directly.*

There is an interesting Corollary in which $C_F = D_F^c$ *denotes the set fo* $x \in \mathbb{R}$ *at which* F *is continuous; that is,* $C_F = \{x \in \mathbb{R} : F(x-) = F(x)\}$*. If* F *and* G *ae two distibution funtions for which F(x) = G(x) for all* $X \in C_F \cap C_G$*, then F(x) = G(x) for all* $x \in \mathbb{R}$*. To see this first obsrve that* $(C_F \cap C_G)^c = D_F \cup D_G$*. So,* $(C_F \cap C_G)^c 4$ *is countabe, and therefore, cannot contain any non-degenerate interval, because the latter are uncountable; that is, if* $a < b$*, then* $(a, b) \cap (C_F \cap C_G) \neq \emptyset$*. So, if* $x \in \mathbb{R}$ *and* $n \geq 1$ *there in an* $x_n \in (C_F \cap C_G) \cap (x, x + 1/n)$*. Then clearly* $lim_{n\to\infty} = x$ *and* $F(x_n = G(x_n$ *for alll* n*. So, using the right continuity,*

$$F(x) = \lim_{n\to\infty} F(x_n) = \lim_{n\to\infty} G(x_n) = G(x) \tag{5.32}$$

5.5 Quantiles

If F *is a distribution function and* $0 < p < 1$*, then an* $x \in \mathbb{R}$ *is called a* p^{th} *quantile or* $100p^{\text{th}}$ *percentile of* F *if*

$$F(x-) \leq p \leq F(x), \tag{5.33}$$

where $F(x-) = \lim_{y\to x, y<x} F(y)$*.If* $X \sim F$*, then the condition may be written*

$$P[X < x] \leq p \leq P[X \leq x], \tag{5.34}$$

and x *is also called a* p^{th} *quantile of* X*. If the equation* $F(x) = p$ *has a solution, then* x *is a* p^{th} *quantile, because* F *is non-decreaing and, therefore,* $F(x-) \leq F(x)$*. But this equation need not have a solution, and a solution need not be unique, if one*

exists. When $p = 1/2$ a p^{th} quantile is called a median. Thus, a random variable falls on either side of a median with probabiliy about one-half. To solve the inequalitites (5.33), let

$$\mathcal{S}_p = \{x \in \mathbb{R} : F(x) \geq p\},$$

the set of $x \in \mathbb{R}$ for which $F(x) \geq p$. For any $0 < p < 1$, the set $\mathcal{S}_p$ is non-empty, because $\lim_{x\to\infty} F(x) = 1$, and therefore,$F(x) \geq p$ for all sufficiently large p. Similarly, there is an $a \in \mathbb{R}$ for which $F(x) < p$ for all $z \leq a$, since $\lim_{x\to-\infty} F(x) = 0$, and therfore, $(-\infty, a] \cap \mathcal{S}_P = m\emptyset$. Thus a every $x \in \mathcal{S}_p$ must be greater than or equal to (actually greater than) a, and serves as a lower bound for $\mathcal{S}_p$. So, $\mathcal{S}_p$ has a greatest lower bound $b- = glb(\mathcal{S}_p)$ for which: $x \geq b$ for all $x \in \mathcal{S}_p$, and if $x \geq b'$ for all $x \in \mathcal{S}_p$, then $b \geq b'$. Let

$$F^{\#}(p) = glb(\mathcal{S}_p),$$

Two importany properties of $F^{\#}$ are

$$F^{\#}(p) \in \mathcal{S}_p \text{ and } F^{\#}(p) \text{ is a } p^{\text{th}} \text{ quantile of } F \tag{5.35}$$

and

$$F(x) \geq p \text{ if andonly if } x \geq F^{\#}(p). \tag{5.36}$$

*For (**??**, there is a sequence $x_n \in \mathcal{S}_p$ for which $\lim_{n\to\infty} x_n = F^{\#}(p)$. Then $x_n \geq F^{\#}(p$ and $F(x_n) \geq p$, sincex$_n \in \mathcal{S}_p$, and therefore, $F[F^{\#}(p)] = \lim_{n\to\infty} F(x_n) \geq p$ So, $F^{\#}(p) \in \mathcal{S}_p$, and $x = F^{\#}(p)$ satisfies the right hand inequaltiy in (5.33). To see that it also satisfies the left hand inquality, observe that if $y < x$, then $y \notin \mathcal{S}_p$, and therefore $F(y) < p$. So, $F(x-) = \lim_{y\to x<y<x} F(y) \leq p$. The proof of (**??**) is similar. If $F(x) \geq p$, then $x \in \mathcal{S}_p$, and therefore $x \geq glb(\mathcal{S}_p) = F^{\#}(p)$. Conversely, if $x \geq F^{\#}(p)$, then $F(x) \geq F[F^{\#}(p)] \geq p$, using (5.35). Another important feature of the quantile function is; If U is uniformly distributed over $[0, 1]$, then $X := F^{\#}(U)\tilde{F}$. To see this first recall that the distribtuion function of U is $P[U \leq u] = 0, u,$ or 1, accrodlingly as $u < 0, 0 \leq u \leq 1$, or $u > 1$. Then observe that the distribution function of $X := F^{\#}(U)$. is*

$$F(X) = P[X \leq X] = P[F^{\#}(U) \leq x] - 1 - P[F^{\#}(U) > x],$$

$$P[F^{\#}(U) \geq x] = P[U \geq F(x)] = 1 - F(x),$$

*using (**??**).*

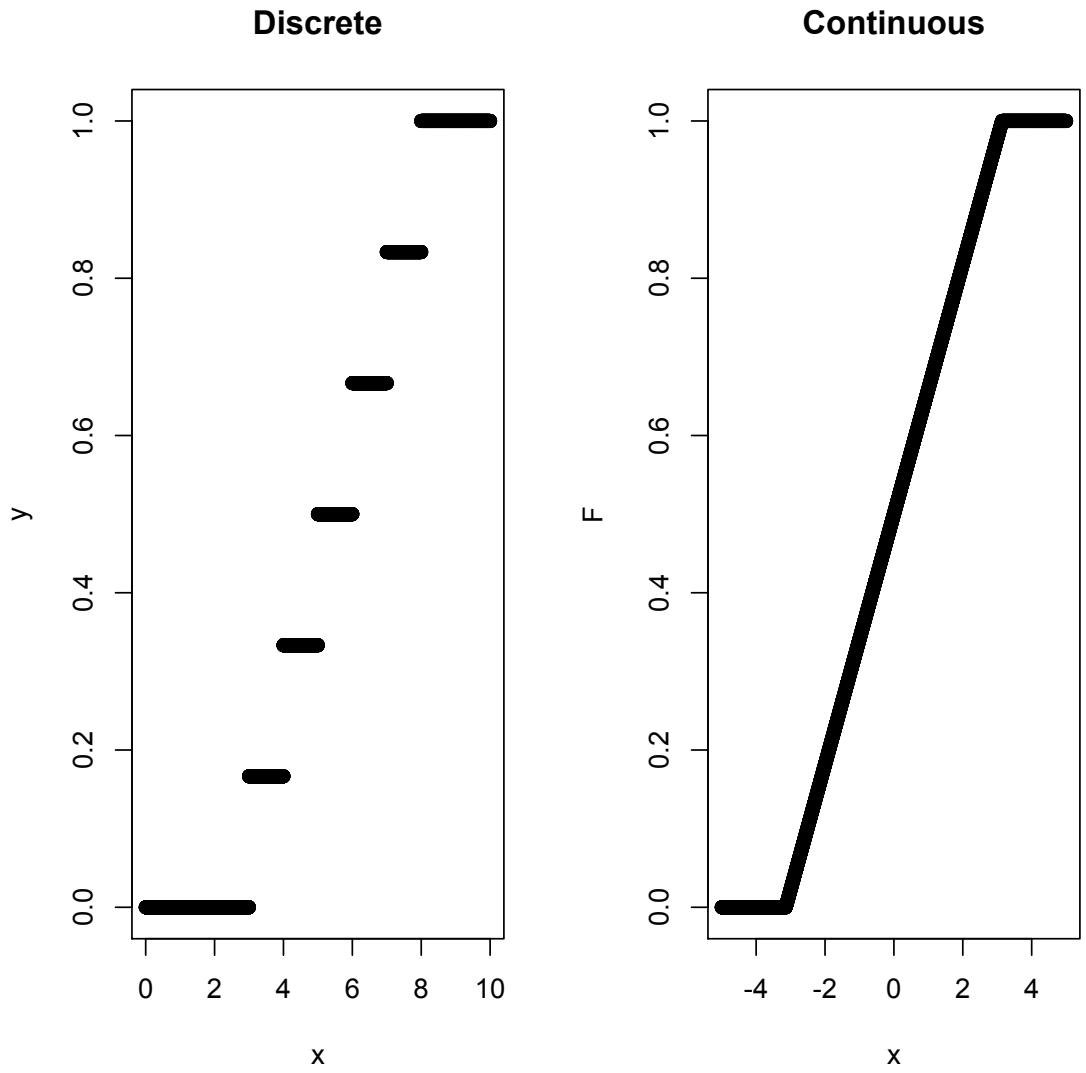

Figure 5.1: The Discrete Uniform Distibtuion Funtion with $n = 6$ and continuous uniform distribution function on $(-\pi\pi]$

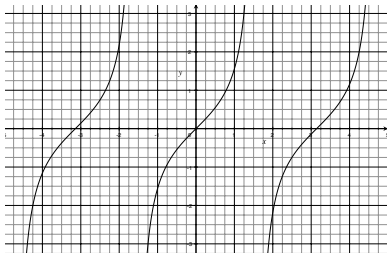

Figure 5.2: example caption

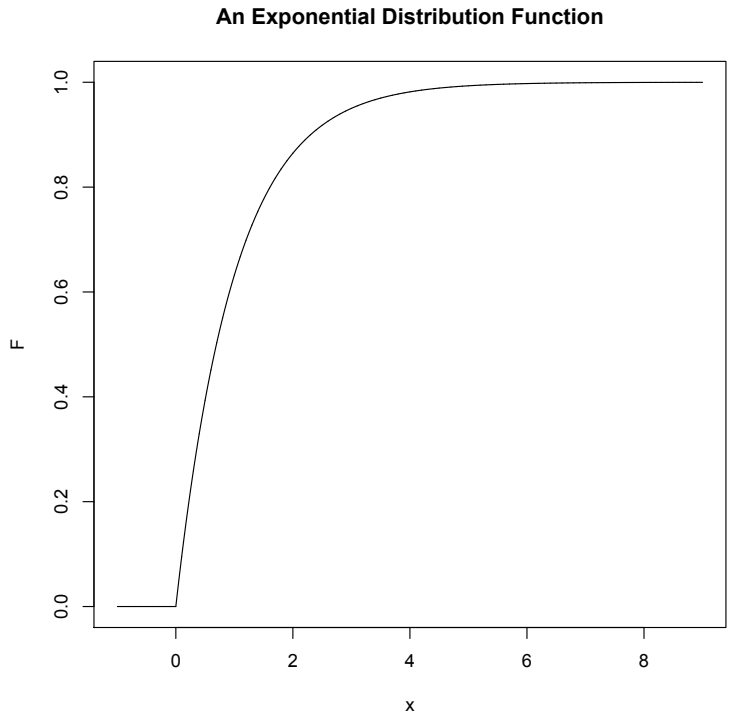

Figure 5.3: The Exponential Distribution Function with $\lambda = 1$

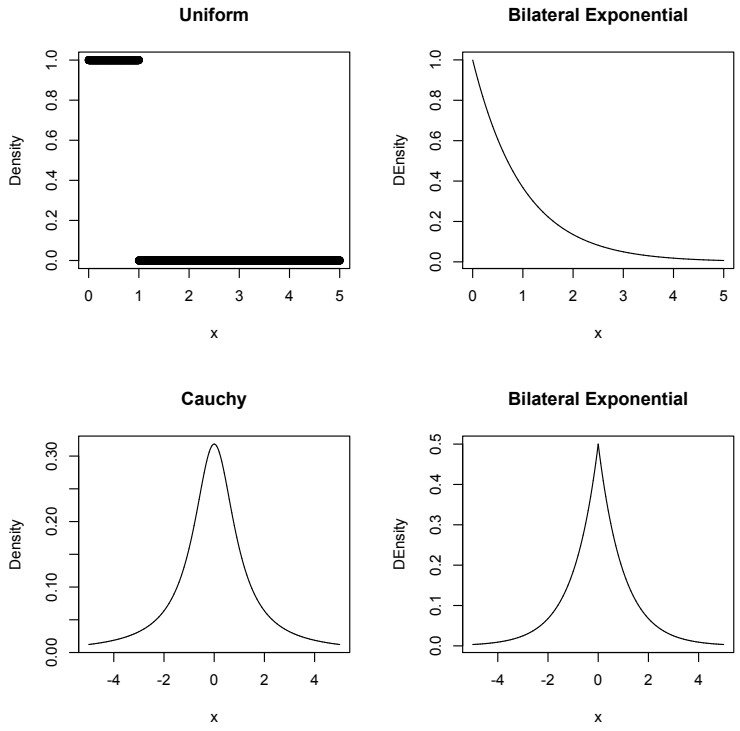
Uniform
Density
x
Bilateral Exponential
DEnsity
x
Cauchy
Density
x
Bilateral Exponential
DEnsity
x

Figure 5.4: