CHOICE, CHANCE, AND INFERENCE

An Introduction to Combinatorics, Probability and Statistics

Carl Wagner
Department of Mathematics
The University of Tennessee
Knoxville, TN 37996-1300

Contents

FOREWORD

In 1961 the public television program Continental Classroom featured a course on probability and statistics taught by the eminent Harvard statistician Frederick Mosteller. Then a senior in high school, I listened to Mosteller's lectures each weekday morning before heading off to school, an experience that began my lifelong fascination with combinatorics, probability, and statistics.

Those familiar with the book, *Probability with Statistical Applications*, written by Mosteller, Rourke, and Thomas to accompany the Continental Classroom course, will immediately note the strong influence of that wonderful old text, now out of print, on the following notes. Like the book, those notes contain a much more extensive treatment of combinatorics and probability than typical introductory statistics texts. Both contain a careful discussion of Chebyshev's inequality and use this inequality to construct crude solutions to problems later solved by using normal approximations to the relevant probabilities.

The statistical applications presented in these notes are limited to some simple hypothesis tests, carried out rather informally, and to constructing confidence intervals for a population percentage. Those planning to employ statistics in their research, or to teach statistics in high school will thus need an additional course purely devoted to that subject. My hope is that the thorough treatment of probability in these notes will prepare students to understand the logic underlying statistical methods and dispel the notion that uncertain inference simply amounts to applying one of the many varieties of statistical software.

To conclude, let me mention a few notable features of the treatment of probability here. I motivate the probability axioms by noting that empirical probabilities (observed relative frequencies) clearly satisfy those axioms, and hence probability models (predicted relative frequencies) ought to satisfy the axioms as well. In my experience, students have a more visceral feel for positive and negative relevance than they do for independence, and so I introduce the former ideas (and their alter egos, overrepresentation and underrepresentation) before the latter. I also attempt to dispel the notion that positive relevance, though in some sense an attenuated implication relation, shares all the properties of implication, citing both Simpson's paradox and the symmetry of positive relevance to make the point. Finally, in addition to presenting Bayes' rule in the usual form, I give a particularly salient form of the rule, "Bayes' rule in odds form," which features the likelihood ratio $P(E|H)/P(E|\bar{H})$, arguably the best single numerical summary of the impact of evidence E on hypothesis H.

Preparation of these notes was supported in part by the Appalachian Math Science Partnership (AMSP). The notes remain a work in progress, and I would be grateful for comments from readers.

CHAPTER 1

BASIC ENUMERATIVE COMBINATORICS

1.1 Multisets and Sets

A *multiset* is a collection of objects, taken without regard to order, and with repetition of the same object allowed. For example, $M = \{1, 1, 1, 2, 2, 2, 2, 3\}$ is a multiset. Since the order in which objects in a multiset are listed is immaterial, we could also write, among many other possibilities, $M = \{1, 3, 2, 2, 1, 2, 2, 1\}$. The list of objects constituting a multiset, called the *elements* of the multiset, is always enclosed by a pair of curly brackets.

A *set* is just a special kind of multiset, namely, a multiset in which there are no repetitions of the same object. For example, $S = \{1, 2, 3\} = \{1, 3, 2\} = \{2, 1, 3\} = \{2, 3, 1\} = \{3, 1, 2\} = \{3, 2, 1\}$ is a set. As an alternative to listing the elements of a set, one can specify them by a defining property. So, for example, one could write $S = \{x : x^3 - 6x^2 + 11x - 6 = 0\}$.

The following sets occur with such frequency in mathematics as to warrant special symbols:

(a) $\mathbb{P} = \{1, 2, 3, \ldots\}$, the set of *positive integers*

(b) $\mathbb{N} = \{0, 1, 2, \ldots\}$, the set of *nonnegative integers*

(c) $\mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$, the set of *integers*

(d) $\mathbb{Q} = \{m/n : m, n \in \mathbb{Z} \text{ and } n \neq 0\}$, the set of *rational numbers*

(e) $\mathbb{R}$, the set of *real numbers*

(f) $\mathbb{C}$, the set of *complex numbers*

(Remark: It is advisable to avoid using the term "natural numbers," since this term is used by some to denote positive integers and by others to denote nonnegative integers.)

As usual, the empty set is denoted by $\emptyset$. Following notation introduced by Richard Stanley, we shall, for each $n \in \mathbb{P}$, denote the set $\{1, 2, \ldots, n\}$ by the simple notation $[n]$.

A set A is *finite* if $A = \emptyset$ or if there exists some $n \in \mathbb{P}$ such that there is a one-to-one correspondence between A and $[n]$. In the latter case, we write $|A| = n$ and say that A has *cardinality* n. The empty set has cardinality 0, i.e., $|\emptyset| = 0$. A set is *infinite* if it is not finite.

Let us recall some basic operations and relations on sets. Given sets A and S, we say that A is a *subset of* S, denoted $A \subseteq S$, if every element of A is also an element of S. A is a *proper subset of* S, denoted $A \subset S$, if $A \subseteq S$, but $A \neq S$, i.e., if every element of A is also an element of S, but there is at least one element of S that is not an element of A. If

$A \subseteq S$, where S contains all elements of interest in the problem at hand (i.e., where S is the "universal set"), then the *complement of* A, denoted $\bar{A}$, is the set of all elements of S that do not belong to A, i.e.,

$$\bar{A} = \{x \in S : x \notin A\}.$$

Alternative notations for $\bar{A}$ are A^c and A'.

Given any sets A and B, their *union*, $A \cup B$, and their *intersection*, $A \cap B$ are defined by

$$A \cup B = \{x : x \in A \text{ and/or } x \in B\}$$

and

$$A \cap B = \{x : x \in A \text{ and } x \in B\}.$$

If $A \cap B = \emptyset$, we say that A and B are *disjoint* or *mutually exclusive.*

1.2 The Addition Rule

The following is perhaps the most basic principle of enumerative combinatorics:

Addition Rule. If A and B are finite, disjoint sets, then $A \cup B$ is finite and $|A \cup B| = |A| + |B|$.

In order to establish this rule, it is necessary to pursue a careful axiomatic analysis of the nonnegative integers. Rather than spending time on this somewhat tedious task, we shall simply take the Addition Rule as an axiom. In any case, it is perfectly obvious intuitively that this rule is correct.

Here is a trivial application of the addition rule. Suppose that there are two sections of a probability course. To determine the total number of students enrolled in the course, we need only find out from the instructors of the two sections how many students are enrolled in their sections and take the sum of these two numbers. We do *not* need to ask that all students enrolled in the course show up *en masse* to be counted.

Here is a less trivial application. Let n be a positive integer. A *composition of* n is a way of writing n as an ordered sum of one or more positive integers (called *parts*). For example, the compositions of 3 are: $1+1+1$, $1+2$, $2+1$, and 3. Let $f(n) =$ the number of compositions of n in which all parts belong to the set $\{1, 2\}$. Let us determine $f(n)$. Listing and counting the acceptable compositions of n for some small values of n yields:

$f(1) = 1$, the acceptable compositions of 1 being: 1

$f(2) = 2$, the acceptable compositions of 2 being: $1 + 1$; 2

$f(3) = 3$, the acceptable compositions of 3 being: $1 + 1 + 1$, $1 + 2$; $2 + 1$

$f(4) = 5$, the acceptable compositions of 4 being: $1 + 1 + 1 + 1$, $1 + 1 + 2$, $1 + 2 + 1$; $2 + 1 + 1$, $2 + 2$.

Further listing and counting reveals that $f(5) = 8$ and $f(6) = 13$. So it looks like maybe $f(n) = F_n$, the n^{th} Fibonacci number. We can prove this if we can show that, for all $n \geq 3$,

$$f(n) = f(n-1) + f(n-2). \tag{1.1}$$

Let C be the set of all compositions of n in which all parts belong to the set $\{1, 2\}$. Let A be the subset of C consisting of those compositions with initial part equal to 1, and let B be the subset of C consisting of those compositions with initial part equal to 2. Obviously, $C = A \cup B$, and so $|C| = |A \cup B|$. But since $A \cap B = \emptyset$, the Addition Rule tells us that $|A \cup B| = |A| + |B|$. So

$$|C| = |A| + |B|. \tag{1.2}$$

By definition of $f(n)$, $|C| = f(n)$. Moreover, $|A| = f(n-1)$, for the acceptable compositions of n with initial part equal to 1 consist of a 1, followed by any acceptable composition of $n - 1$. Similarly, $|B| = f(n-2)$, for the acceptable compositions of n with initial part equal to 2 consist of a 2, followed by any acceptable composition of $n - 2$. Combining these observations with (1.2) yields (1.1).

1.3 Generalizations of the Addition Rule

The sets $A_1, A_2, \ldots, A_n$ are said to be *pairwise disjoint* if $A_i \cap A_j = \emptyset$ whenever $i \neq j$. The following theorem generalizes the Addition Rule.

Theorem 1.1. (Extended Addition Rule) If the sets $A_1, A_2, \ldots, A_n$ are finite and pairwise disjoint, then the set $A_1 \cup A_2 \cup \cdots \cup A_n$ is finite and

$$|A_1 \cup A_2 \cup \cdots \cup A_n| = |A_1| + |A_2| + \cdots + |A_n|. \tag{1.3}$$

Proof. This theorem may be easily proved by induction on n using the Addition Rule. □

Here is an application of Theorem 1.1. Recall that a composition of the positive integer n is a way of writing n as an ordered sum of one or more positive integers. Let $c(n)$ denote the number of compositions of n. By listing and counting, we can easily discover the values $c(1) = 1$, $c(2) = 2$, $c(3) = 4$, $c(4) = 8$, and $c(5) = 16$, which suggests the following result.

Theorem 1.2. For every $n \in \mathbb{P}$, $c(n) = 2^{n-1}$.

Proof. We first show that for all $n \geq 1$

$$c(n+1) = c(n) + c(n-1) + \cdots + c(1) + 1. \tag{1.4}$$

To show (1.4), let A be the set of all compositions of $n + 1$, and let A_i be the set of all compositions in A with initial part equal to i, for $i = 1, \ldots, n+1$. Clearly $A = A_1 \cup \cdots \cup A_n \cup A_{n+1}$ and the sets $A_1, \ldots, A_{n+1}$ are pairwise disjoint. So $|A| = |A_1| + \cdots + |A_{n+1}|$. But

$|A| = c(n+1)$, $|A_i| = c(n+1-i)$ for $i = 1, \ldots, n$ [why?], and $|A_{n+1}| = 1$ [why?]. This establishes (1.4).

Using (1.4), we prove that $c(n) = 2^{n-1}$ by induction on n. We have already seen that $c(1) = 1 = 2^0$. Suppose that $c(k) = 2^{k-1}$ for $k = 1, \ldots, n$. From (1.4) we then get

$$c(n+1) = 2^{n-1} + 2^{n-2} + \cdots + 1 + 1 = 2^n,$$

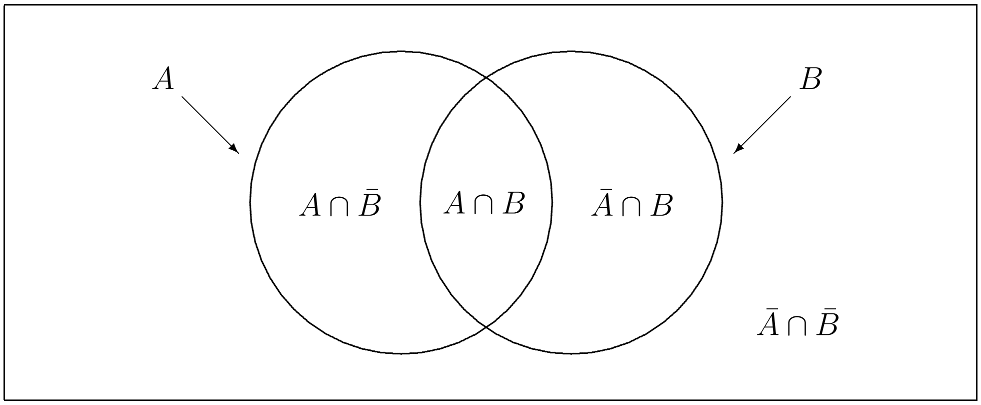
by the well known formula for the sum of a geometric progression. □

Both the Addition Rule and the Extended Addition Rule specify the cardinality of a union of pairwise disjoint sets. When sets are not necessarily pairwise disjoint the cardinality of their union is determined by the *Principle of Inclusion and Exclusion*, the simplest case of which is stated in the next theorem.

Theorem 1.3. If A and B are finite sets, then $A \cup B$ is finite, and

$$|A \cup B| = |A| + |B| - |A \cap B|. \tag{1.5}$$

Proof. From the Venn diagram below,



and the addition rule, we have

$$\begin{aligned} |A \cup B| &= |A| + |\bar{A} \cap B| \\ &= |A| + |\bar{A} \cap B| + |A \cap B| - |A \cap B| \\ &= |A| + |B| - |A \cap B|. \qquad \square \end{aligned}$$

Note that the Addition Rule states the special case of Theorem 1.3 in which $A \cap B = \emptyset$. Using Theorem 1.3, one can prove that for any finite sets A, B, C,

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|, \tag{1.6}$$

and, in turn, that for any finite sets A, B, C, D,

$$|A \cup B \cup C \cup D| = |A| + |B| + |C| + |D| - |A \cap B| - |A \cap C| - |A \cap D|$$

$$\begin{aligned}&-|B\cap C|-|B\cap D|-|C\cap D|+|A\cap B\cap C|\\ &+|A\cap B\cap D|+|A\cap C\cap D|+|B\cap C\cap D|-|A\cap B\cap C\cap D|.\end{aligned}\tag{1.7}$$

More generally, one may prove by induction on n that if $A_1, \ldots, A_n$ are finite sets, then $A_1 \cup \cdots \cup A_n$ is finite, and

$$\begin{aligned}|A_1\cup\cdots\cup A_n| = \sum_{1\leq i\leq n}|A_i| - \sum_{1\leq i<j<n}|A_i\cap A_j| + \sum_{1\leq i<j<k\leq n}|A_i\cap A_j\cap A_k|\\ -\cdots+(-1)^{n-1}|A_1\cap\cdots\cap A_n|.\end{aligned}\tag{1.8}$$

1.4 Sequences and Words

A *sequence* is an ordered list of objects, with repetitions of the same object allowed. The objects constituting a sequence, called the *terms* of the sequence, are separated by commas, enclosed in parentheses. A sequence may be finite, e.g., $(1, 2, 3, 2)$, or infinite, e.g.,$(1, 3, 5, 7, \ldots)$. Sequences consisting of the same terms, but listed in a different order, are considered to be different sequences. So while $\{1, 2, 3\} = \{1, 3, 2\}$, for example, $(1, 2, 3) \neq (1, 3, 2)$. A sequence of the form $(t_1, t_2, \ldots, t_k)$ is called a sequence of length k and a sequence of the form $(t_1, t_2, \ldots)$ is called a *sequence of infinite length* or, more simply, an *infinite sequence.* Note that the length of a sequence equals the number of terms in the sequence, *not* the number of distinct terms in the sequence. So, for example, $(1, 2, 3, 2)$ is a sequence of length 4 and $(1, 1, 1, \ldots)$ is an infinite sequence.

If all of the terms of a sequence are elements of a set A, the sequence is said to be a *sequence in A* or an *A-sequence.* So, for example, $(1, 2, 3, 2)$ is a sequence in $\mathbb{P}$ (Of course, it is also a sequence in $\mathbb{N}$, in $\mathbb{Z}$, in $\mathbb{Q}$, in $\mathbb{R}$, and in $\mathbb{C}$!).

A sequence in the set A is also called a *word* in the "alphabet" A. When a sequence is construed as a word, it is typically written without enclosing parentheses, and with no commas. Thus the sequence $(t_1, t_2, \ldots, t_k)$ is written $t_1t_2\cdots t_k$ when construed as a word. Also, whereas one refers to t_i as the i^{th} *term* of the sequence $(t_1, \ldots, t_k)$, t_i is called the i^{th} *letter* of the word $t_1t_2\cdots t_k$.

1.5 The Multiplication Rule

The following rule for counting sequences is one of the two pillars (the other being the Addition Rule) of enumerative combinatorics. While it is possible to give a rigorous proof of this rule, we shall simply take it as an axiom.

The Multiplication Rule. Let s be the number of sequences $(t_1, t_2, \ldots, t_k)$ that can be constructed under the following restrictions:

1° there are n_1 possible values of t_1,

2° whatever the value chosen for t_1, there are n_2 possible values of t_2

3° whatever the values chosen for t_1 and t_2, there are n_3 possible values of t_3,

$\vdots$

k° whatever the values chosen for $t_1, t_2, \ldots, t_{k-1}$, there are n_k possible values of t_k.

Then $s = n_1 n_2 \cdots n_k$.

The following theorems are obvious consequences of the multiplication rule.

Theorem 1.4. Let n and k be any positive integers. If $|A| = n$, there are n^k sequences of length k in A (equivalently, n^k words of length k in the "alphabet" A).

Proof. Straightforward. □

Theorem 1.5. Let n and k be any positive integers. There are n^k ways to distribute k distinct balls among n distinct urns.

Proof. Suppose that the balls are labeled $1, 2, \ldots, k$ and the urns are labeled $1, 2, \ldots, n$. Each distribution can be represented as a sequence $(t_1, t_2, \ldots, t_k)$, where t_1 is the number of the urn in which ball 1 is placed, t_2 is the number of the urn in which ball 2 is placed, etc. Since each t_i can take any of the n values $1, 2, \ldots, n$, the result follows from the Multiplication Rule. □

A sequence in which all of the terms are distinct (or a word in which all of the letters are distinct) is called a *permutation.* In particular, if $|A| = n$, a sequence of length k in A in which all of the terms are distinct (or word of length k in the alphabet A in which all of the letters are distinct) is called a *permutation of n things* (namely, the n things in A) *taken k at a time.* A permutation of n things taken n at a time is simply called a *permutation of n things.*

Before stating the next theorem, we need to introduce some notation. If n and k are nonnegative integers, the notation $n^{\underline{k}}$ is read "n falling factorial k." It is defined as follows:

(i) $n^{\underline{0}} = 1$ for all $n \in \mathbb{N}$, and

(ii) $n^{\underline{k}} = n(n-1)\cdots(n-k+1)$ for all $n \in \mathbb{N}$ and all $k \in \mathbb{P}$.

In particular, $0^{\underline{0}} = 1^{\underline{0}} = 2^{\underline{0}} = \cdots = 1$, $n^{\underline{1}} = n$ for all $n \in \mathbb{N}$, $n^{\underline{2}} = n(n-1)$ for all $n \in \mathbb{N}$, etc. Recalling that $0! = 1$ and $n! = n(n-1)\cdots(1)$ for all $n \in \mathbb{P}$, we see that $n^{\underline{n}} = n!$ for all $n \in \mathbb{N}$. On the other hand, if $0 \leq n < k$, then $n^{\underline{k}} = 0$. For example, $3^{\underline{5}} = (3)(2)(1)(0)(-1) = 0$.

Theorem 1.6. For all positive integers n and k, there are $n^{\underline{k}}$ permutations of n things taken k at a time.

Proof. If $k > n$, there are no permutations of n things taken k at a time [why?]. As observed above, $n^{\underline{k}} = 0$ if $k > n$, so the formula $n^{\underline{k}}$ is correct in this case.

If $1 \leq k \leq n$, we must construct a sequence $(t_1, \ldots, t_k)$ of distinct terms from the elements of a set A, where $|A| = n$. Clearly, we have n choices for t_1, then $(n-1)$ choices for t_2, then

$(n-2)$ choices for $t_3, \ldots$, and finally $n-k+1$ choices for t_k. The result now follows from the Multiplication Rule. □

Corollary 1.6.1. For all positive integers n, there are $n!$ permutations of n things.

Proof. By Theorem 1.6 there are $n^{\underline{n}}$ permutations of n things. But, as observed above, $n^{\underline{n}} = n!$. □

Theorem 1.7. Let n and k be any positive integers. There are $n^{\underline{k}}$ ways to distribute k distinct balls among n distinct urns with at most one ball per urn.

Proof. Represent distributions as in the proof of Theorem 1.5 and observe that the acceptable distributions correspond to permutations of $1, \ldots, n$, taken k at a time. Then invoke Theorem 1.6. □

1.6 Useful Counting Strategies

In counting a family of sequences (or words) of length k, it is useful to begin with k slots, (optionally) labeled $1, 2, \ldots, k$:

$$\underset{(1)}{\underline{\qquad}}\ \ \underset{(2)}{\underline{\qquad}}\ \ \underset{(3)}{\underline{\qquad}}\ \ \cdots\ \ \underset{(k)}{\underline{\qquad}}$$

One then fills in the slot labeled (i) with the number n_i of possible values of the i^{th} term of the sequence (or i^{th} letter of the word), given the restrictions of the problem, and multiplies:

$$\underset{(1)}{\underline{n_1}} \times \underset{(2)}{\underline{n_2}} \times \underset{(3)}{\underline{n_3}} \times \cdots \times \underset{(k)}{\underline{n_k}}$$

Example 1. There are 3 highways from Knoxville to Nashville, and 4 from Nashville to Memphis. How many roundtrip itineraries are there from Knoxville to Memphis via Nashville? How many itineraries are there if one never travels the same highway?

Solution. A round trip itinerary is a sequence of length 4, the i^{th} term of which designates the highway taken on the i^{th} leg of the trip. The solution to the first problem is thus $\underline{3} \times \underline{4} \times \underline{4} \times \underline{3} = 144$, and the solution to the second problem is $\underline{3} \times \underline{4} \times \underline{3} \times \underline{2} = 72$.

Remark. If there are 2 highways from Knoxville to Asheville and 5 from Asheville to Durham, the number of itineraries one could follow in making a round trip from Knoxville to Memphis via Nashville *or* from Knoxville to Durham via Asheville would be calculated, using both the addition and multiplication rules as

$$\underline{3} \times \underline{4} \times \underline{4} \times \underline{3} \quad + \quad \underline{2} \times \underline{5} \times \underline{5} \times \underline{2} \quad = \quad 244$$

In employing the multiplication rule, one need not fill in the slots from left to right. In fact *one should always fill in the number of alternatives so that the slot subject to the most restrictions is filled in first, the slot subject to the next most restrictions is filled in next*, etc.

Example 2. How many odd, 4-digit numbers are there having no repeated digits?

Solution. Fill in the slots below in the order indicated and multiply:

$$\underset{(2)}{\underline{\;8\;}} \quad\times\quad \underset{(3)}{\underline{\;8\;}} \quad\times\quad \underset{(4)}{\underline{\;7\;}} \quad\times\quad \underset{(1)}{\underline{\;5\;}}$$

(The position marked (1) can be occupied by any of the 5 odd digits; the position marked (2) can be occupied by any of the 8 digits that are different from 0 and from the digit in position (1); etc.; etc.).

Note what happens if we try to work the above problem left to right.

$$\underset{(1)}{\underline{\;9\;}} \quad\times\quad \underset{(2)}{\underline{\;9\;}} \quad\times\quad \underset{(3)}{\underline{\;8\;}} \quad\times\quad \underset{(4)}{\underline{\;?\;}}$$

We get off to a flying start, but cannot fill in the last blank, since the number of choices here depends on how many odd digits have been chosen to occupy slots (1), (2), and (3).

Sometimes, no matter how cleverly we choose the order in which slots are filled in, a counting problem simply must be decomposed into two or more subproblems.

Example 3. How many even, 4 digit numbers are there having no repeated digits?

Solution. Following the strategy of the above example, we can easily fill in slot (1) below, but the entry in slot (2) depends on whether the digit 0 or one of the digits 2, 4, 6, 8 is chosen as the last digit:

$$\underset{(2)}{\underline{\;?\;}} \quad \underset{(3)}{\underline{\;\;\;}} \quad \underset{(4)}{\underline{\;\;\;}} \quad \underset{(1)}{\underline{\;5\;}}.$$

So we decompose the problem into two subproblems. First we count the 4 digit numbers in question with the last digit equal to zero:

$$\underset{(2)}{\underline{\;9\;}} \quad\times\quad \underset{(3)}{\underline{\;8\;}} \quad\times\quad \underset{(4)}{\underline{\;7\;}} \quad\times\quad \underset{(1)}{\underline{\;1\;}} \quad=\quad 504.$$

Then we count those with last digit equal to 2, 4, 6, or 8:

$$\underset{(2)}{\underline{\;8\;}} \quad\times\quad \underset{(3)}{\underline{\;8\;}} \quad\times\quad \underset{(4)}{\underline{\;7\;}} \quad\times\quad \underset{(1)}{\underline{\;4\;}} \quad=\quad 1792.$$

The total number of even, 4-digit numbers with no repeated digits is thus 504 + 1792 = 2296.

Problems involving the enumeration of sequences with prescribed or forbidden adjacencies may be solved by “pasting” together objects that must be adjacent, so that they form a single object.

Example 4. In how many ways may Graham, Knuth, Stanley, and Wilf line up for a photograph if Graham wishes to stand next to Knuth?

Solution: There are 3! permutations of $\underset{\smile}{\mathrm{GK}}$, S, and W and 3! permutations of $\underset{\smile}{\mathrm{KG}}$, S, and W. So there are $3! + 3! = 12$ ways to arrange these individuals (all of whom are famous combinatorists) under the given restriction.

Remark: To count arrangements with forbidden adjacencies, simply count the arrangements in which those adjacencies occur, and subtract from the total number of arrangements. Thus, the number of ways in which the 4 combinatorists can line up for a photograph with Graham and Knuth *not* adjacent is $4! - 12 = 12$.

1.7 Problems

1. In how many ways can a family of 6 line up for a photograph a) with no restrictions, and b) if the parents stand next to each other in the middle?

2. How many 5-digit numbers can be formed from the integers $1, 2, 4, 6, 7, 8$, if no integer can be used more than once? How many of these numbers will be even? How many odd?

3. A "binary word" is a word constructed from the "alphabet" $\{0, 1\}$. How many binary words of length 5 are there? How many of these words contain at least one 0 and at least one 1?

4. How many 3 digit numbers are there? How many of these are even and have no repeated digits?

5. How many permutations of the numbers $1, 2, \ldots, 8$ are there in which the numbers 1 and 2 are not adjacent?

6. How many ways are there to distribute balls labeled $1, 2, \ldots, 10$ among urns labeled $1, 2, 3$?

7. A teacher wishes to assign each of 6 students a different book to read. If 12 different books are available, in how many ways may this be done?

8. Twelve boys try out for the basketball team. Two can play only at center, four only as right or left guard, and the rest can play only as right or left forward. In how many ways could the coach assign a team?

9. How many numbers, each with at least 3 digits, can be formed from the 5 digits $1, 2, 3, 4, 5$, if, in each number, no digit may be used more than once?

10. In how many ways can 5 boys and 5 girls be seated alternately in a row of 10 chairs, numbered from 1 to 10, if a boy always occupies chair number one?

11. A *function* from a set A to a set B is a rule which assigns to each $a \in A$ an element, denoted $f(a)$, belonging to B. How many different functions from $\{a, b, c, d\}$ to $\{1, 2, 3\}$ are there?

12. An encyclopedia consists of 12 volumes numbered 1 to 12. In how many ways can the 12 volumes be lined up on a shelf so that some or all of the volumes are out of order?

13. How many 5-digit numbers can be formed? How many of these begin with 2 and end with 4? How many do not contain the digit 5? How many are divisible by 5?

14. From a committee of 10 people, it is necessary to choose a chair, vice-chair, and secretary. In how many ways may this be done?

15. A farm is divided into 49 plots of land. An agricultural experimenter wishes to compare, on this farm, the yield of 5 varieties of corn using 3 kinds of insecticide and 4 different fertilizers. Will he have enough plots to compare all possible combinations of corn, insecticide, and fertilizer?

16. How many ways are there to distribute 10 distinct balls among 5 distinct urns? How many ways if we can put at most one ball per urn?

17. How many ways are there to distribute 8 different gifts among 9 children, with no restrictions on the number of gifts given to any child? How many ways if no child gets more than one gift?

18. How many permutations of 3 things, taken 5 at a time, are there? How many permutations of 5 things taken 3 at a time?

19. How many ways are there to distribute 3 distinct balls among 2 distinct urns if at least one ball must be placed in each urn?

1.8 Binomial Coefficients

The *power set* of a set A, denoted 2^A (we'll see why shortly) is the set of all subsets of A, i.e.,

$$2^A = \{B : \ B \subseteq A\}.$$

For example, $2^{\{a,b,c\}} = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a,b\}, \{a,c\}, \{b,c\}, \{a,b,c\}\}$.

Theorem 1.8. If $|A| = n$, where $n \in \mathbb{N}$, then A has 2^n subsets, i.e., $|2^A| = 2^{|A|}$.

Proof. The assertion is true for $n = 0$, since $2^{\emptyset} = \{\emptyset\}$ and $|\{\emptyset\}| = 1$. Suppose $n \in \mathbb{P}$. Let $A = \{a_1, a_2, \ldots, a_n\}$. Each subset B of A may be represented uniquely by a sequence $(t_1, \ldots, t_n)$ in $\{0, 1\}$, the so-called *bit string representation* of B, where $t_i = 0$ if $a_i \notin B$ and $t_i = 1$ if $a_i \in B$. (For example, the bit string representations of $\emptyset$, $\{x_1, x_3\}$, and A are, respectively, $(0, 0, \ldots, 0)$, $(1, 0, 1, 0, \ldots, 0)$, and $(1, 1, \ldots, 1)$.) By the Multiplication Rule, there are obviously 2^n bit strings of length n, and hence 2^n subsets of an n-element set. □

Remark. In light of Theorem 1.8, we see that the notation 2^A for the power set of A is a mnemonic for the result $|2^A| = 2^{|A|}$.

Let $|A| = n$, where $n \in \mathbb{N}$. For every $k \in \mathbb{N}$, let $\binom{n}{k}$ denote the number of subsets of A that have k elements, i.e.,

$$\binom{n}{k} = |\{B : B \subseteq A \text{ and } |B| = k\}|.$$

The symbol $\binom{n}{k}$ is read " n choose k," or "the k^{th} binomial coefficient of order n." Certain values of $\binom{n}{k}$ are immediately obvious. For example, $\binom{n}{0} = 1$, since the only subset of an n-set having cardinality 0 is the empty set. Also, $\binom{n}{n} = 1$, since the only subset of an n-set A having cardinality n is A itself. Moreover, it is clear that $\binom{n}{k} = 0$ if $k > n$, for the subsets of an n-set must have cardinality less than or equal to n.

Theorem 1.9. For all $n \in \mathbb{N}$ and all $k \in \mathbb{N}$,

$$\binom{n}{k} = \frac{n^{\underline{k}}}{k!},$$

where $n^{\underline{0}} = 1$ and $0! = 1$.

Proof. Let $|A| = n$. If $0 \leq n < k$, or $n = k = 0$, this result follows from earlier remarks. If $1 \leq k \leq n$, all of the $n^{\underline{k}}$ permutations of the n things in A taken k at a time arise from (i°) choosing a subset $B \subseteq A$ with $|B| = k$ and (ii°) permuting the elements of B in one of the $k!$ possible ways. Hence

$$n^{\underline{k}} = \binom{n}{k} k!.$$

Solving for $\binom{n}{k}$ yields $\binom{n}{k} = \frac{n^{\underline{k}}}{k!}$. □

Remark. A subset of cardinality k of the n-element set A is sometimes called a *combination of the n things in A taken k at a time*. The essence of the above proof is that every combination of n things taken k at a time gives rise to $k!$ permutations of n things taken k at a time. Hence there are $k!$ times as many permutations of n things taken k at a time as there are combinations of n things taken k at a time.

Since $n^{\underline{k}} = 0$ if $k > n$, it follows that $\binom{n}{k} = \frac{n^{\underline{k}}}{k!} = 0$ if $k > n$, as we previously observed. If $0 \leq k \leq n$, there is an alternative formula for $\binom{n}{k}$. We have in this case

$$\binom{n}{k} = \frac{n^{\underline{k}}}{k!} = \frac{n(n-1)\cdots(n-k+1)}{k!} \, \frac{(n-k)!}{(n-k)!} = \frac{n!}{k!(n-k)!}.$$

Let us make a partial table of the binomial coefficients, the so-called *Pascal's Triangle* (Blaise Pascal, 1623-1662).

k n	0	1	2	3	4	5	6
0	1	0	0	0	0	0	0
1	1	1	0	0	0	0	0
2	1	2	1	0	0	0	0
3	1	3	3	1	0	0	0
4	1	4	6	4	1	0	0
5	1	5	10	10	5	1	0
6	1	6	15	20	15	6	1

Theorem 1.10. For all $n, k \in \mathbb{N}$ such that $0 \leq k \leq n$, $\binom{n}{k} = \binom{n}{n-k}$.

Proof. There is an easy algebraic proof of this using the formula $\binom{n}{k} = n!/k!(n-k)!$. □

Pascal's triangle is generated by a simple recurrence relation.

Theorem 1.11. For all $n \in \mathbb{N}$, $\binom{n}{0} = 1$ and for all $k \in \mathbb{P}$, $\binom{0}{k} = 0$. For all $n, k \in \mathbb{P}$,

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

Proof. Let $A = \{a_1, \ldots, a_n\}$. The k-element subsets of A belong to one of two disjoint, exhaustive classes (i) the class of k-element subsets which contain a_1 as a member, and (ii) the class of k-element subsets, which do not contain a_1 as a member. Then are clearly $\binom{n-1}{k-1}$ subsets in the first class (choose $k-1$ additional elements from $\{a_2, \ldots, a_n\}$ to go along with a_1) and $\binom{n-1}{k}$ subsets in the second class (since a_1 is excluded as a member, choose all k elements of the subset from $\{a_2, \ldots, a_n\}$). By the addition rule it follows that $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$. □

Theorem 1.12. For all $n \in \mathbb{N}$,

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n.$$

Proof. You are probably familiar with the proof of Theorem 1.12 based on the binomial theorem, which we shall prove later. Actually, there is a much simpler combinatorial proof. By Theorem 1.8, 2^n counts the total number of subsets of an n-element set. But so does $\sum_{k=0}^{n} \binom{n}{k}$, counting these subsets in $n+1$ disjoint exhaustive classes, the class of subsets of cardinality $k = 0, k = 1, \ldots, k = n$. By the extended addition rule we must therefore have $\sum_{k=0}^{n} \binom{n}{k} =$ the total number of subsets of an n-set $= 2^n$. □

The following is an extremely useful identity rarely mentioned in combinatorics texts.

Theorem 1.13. For all $n, k \in \mathbb{P}$,

$$\binom{n}{k} = \frac{n}{k}\binom{n-1}{k-1}.$$

Proof. By Theorem 1.9,

$$\begin{aligned}\binom{n}{k} &= \frac{n^{\underline{k}}}{k!} = \frac{n}{k}\frac{(n-1)(n-2)\cdots(n-k+1)}{(k-1)!}\\ &= \frac{n}{k}\frac{(n-1)((n-1)-1)\cdots((n-1)-(k-1)+1)}{(k-1)!}\\ &= \frac{n}{k}\frac{(n-1)^{\underline{k-1}}}{(k-1)!} = \frac{n}{k}\binom{n-1}{k-1}.\end{aligned}$$

Here is a combinatorial proof. (A combinatorial proof that $a = b$ is based on the observation that a and b both count the same set of objects, and, hence, must be equal.) It is equivalent to prove that for all $n, k \in \mathbb{P}$

$$\binom{n}{k}k = n\binom{n-1}{k-1}.$$

But each side of the above counts the number of ways to choose from n people a committee of k people, with one committee member designated as chair. The LHS counts such committees by first choosing the k members ($\binom{n}{k}$ ways) and then one of these k as chair (k ways). The RHS counts them by first choosing the chair (n ways), then $k-1$ additional members from the $n-1$ remaining people ($\binom{n-1}{k-1}$ ways). □

Here is an application of the above theorem.

Corollary 1.13.1. For all $n \in \mathbb{N}$,

$$\sum_{k=0}^{n} k\binom{n}{k} = n \cdot 2^{n-1}.$$

Proof. One can prove this identity by induction on n, but that is not very enlightening. We give a proof that discovers the result as well as proving it. First note that the identity holds for $n = 0$. So assume $n \geq 1$. Then

$$\begin{aligned}\sum_{k=0}^{n} k\binom{n}{k} &= \sum_{k=1}^{n} k\binom{n}{k} = \sum_{k=1}^{n} k \cdot \frac{n}{k}\binom{n-1}{k-1}\\ &= n\sum_{k=1}^{n}\binom{n-1}{k-1} = n\sum_{j=0}^{n-1}\binom{n-1}{j} = n \cdot 2^{n-1}. \qquad \square\\ &\qquad\qquad (\text{let } j = k-1)\end{aligned}$$

We conclude this section with a famous binomial coefficient identity known as Vandermonde's identity (Abnit-Theophile Vandermonde, 1735-1796).

Theorem 1.14. For all m, n, and $r \in \mathbb{N}$,

$$\binom{m+n}{r} = \sum_{j=0}^{r} \binom{m}{j}\binom{n}{r-j}.$$

Proof. Given a set of m men and n women, there are $\binom{m+n}{r}$ ways to select a committee with r members from this set. The RHS of the above identity counts these committees in subclasses corresponding to $j = 0, \ldots, r$, when j denotes the number of men on a committee. □

Remark. A noncombinatorial proof of Theorem 1.14 based on the binomial theorem, expands $(1+x)^m$, $(1+x)^n$, and multiplies, comparing the coefficient of x^r in the product with the coefficient of x^r in the expansion of $(1+x)^{m+n}$. This proof is considerably more tedious than the combinatorial proof.

Corollary 1.14.1. For all $n \in \mathbb{N}$

$$\sum_{j=0}^{n} \binom{n}{j}^2 = \binom{2n}{n}.$$

Proof. Let $m = r = n$ in Theorem 1.14 and use Theorem 1.10:

$$\sum_{j=0}^{n} \binom{n}{j}^2 = \sum_{j=0}^{n} \binom{n}{j}\binom{n}{n-j} = \binom{2n}{n}. \qquad \square$$

1.9 The Binomial Theorem

In how many ways may 3 (indistinguishable) x's and 4 (indistinguishable) y's be arranged in a row? Equivalently, how many words of length 7 in the alphabet $\{x, y\}$ are there in which x appears 3 times and y appears 4 times? The multiplication rule is not helpful here. One can fill in the number of possible choices in the first 3 slots of the word

$$\underset{(1)}{\underline{\;2\;}} \times \underset{(2)}{\underline{\;2\;}} \times \underset{(3)}{\underline{\;2\;}} \quad \underset{(4)}{\underline{\;?\;}} \quad \underset{(5)}{\underline{\quad}} \quad \underset{(6)}{\underline{\quad}} \quad \underset{(7)}{\underline{\quad}},$$

but the number of choices for slot (4) depends on how many x's were chosen to occupy the first 3 slots. Going back and decomposing the problem into subcases results in virtually listing all the possible arrangements, something we want to avoid.

The solution to this problem requires a different approach. Consider the slots

$$\underset{(1)}{\underline{\quad}} \quad \underset{(2)}{\underline{\quad}} \quad \underset{(3)}{\underline{\quad}} \quad \underset{(4)}{\underline{\quad}} \quad \underset{(5)}{\underline{\quad}} \quad \underset{(6)}{\underline{\quad}} \quad \underset{(7)}{\underline{\quad}}.$$

Once we have chosen the three slots to be occupied by x's, we have completely specified a word of the required type, for we must put the y's in the remaining slots. And how many ways are there to choose 3 of the 7 slots in which to place x's? The question answers itself, "7 choose 3", i.e., $\binom{7}{3}$ ways. So there are $\binom{7}{3}$ linear arrangements of 3 x's and 4 y's. Of course, we could also write the answer as $\binom{7}{4}$ or as $7!/3!4!$. The general rule is given by the following theorem.

Theorem 1.15. The number of words of length n, consisting of n_1 letters of one sort and n_2 letters of another sort, where $n = n_1 + n_2$, is $\binom{n}{n_1} = \binom{n}{n_2} = \frac{(n_1+n_2)!}{n_1!n_2!}$.

Proof. Obvious. □

Remark 1. A word or sequence consisting of n_1 letters of the sort and n_2 letters of another sort is an example of words or sequences of several sorts of things *with prescribed frequencies* (the frequencies here being, of course, n_1 and n_2).

Remark 2. Theorem 1.15 may be formulated as a result about distributions, as follows: The number of ways to distribute n labeled balls among 2 urns, labeled u_1 and u_2, so that n_i balls are placed in u_i, $i = 1, 2$ $(n_1 + n_2 = n)$ is $\binom{n}{n_1} = \binom{n}{n_2} = \frac{(n_1+n_2)!}{n_1!n_2!}$. This sort of distribution is called a distribution *with prescribed occupancy numbers* (the occupancy numbers here being, of course, n_1 and n_2).

If we expand $(x+y)^2 = (x+y)(x+y)$, we get, before simplification,

$$(x+y)^2 = xx + xy + yx + yy,$$

the sum of all 4 words of length 2 in the alphabet $\{x, y\}$. Similarly

$$(x+y)^3 = xxx + \underline{xxy} + \underline{xyx} + xyy + \underline{yxx} + yxy + yyx + yyy,$$

the sum of all 8 words of length 3 in the alphabet $\{x, y\}$. After simplification, we get the familiar formulas

$$\begin{aligned} (x+y)^2 &= x^2 + 2xy + y^2, \qquad\qquad \text{and} \\ (x+y)^3 &= x^3 + \underline{3}x^2y + 3xy^2 + y^3. \end{aligned}$$

Where did the coefficient 3 of x^2y come from? It came from the 3 underlined words in the presimplified expansion of $(x+y)^3$. And how could we have predicted that this coefficient would be 3, without writing out the presimplified expansion? Well, the number of words of length 3 consisting of 2 x's and 1 y is $\binom{3}{1} = \binom{3}{2} = 3$, that's how. This leads to a proof of the binomial theorem:

Theorem 1.16. For all $n \in \mathbb{N}$,

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

Proof. If $n = 0$, the result is obvious, so suppose that $n \in \mathbb{P}$. Before simplification the expansion of $(x+y)^n$ consists of the sum of all 2^n words of length n in the alphabet $\{x, y\}$. The number of such words consisting of k x's and $n - k$ y's is $\binom{n}{k}$ by Theorem 1.15. □

The formula

$$(x+n)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} = y^n + \binom{n}{1} x y^{n-1} + \binom{n}{2} x^2 y^{n-2} + \cdots + x^n$$

is an expansion in *ascending powers of* x. We can of course also express the binomial theorem as

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k = x^n + \binom{n}{1} x^{n-1} y + \binom{n}{2} x^{n-2} y^2 + \cdots + y^n,$$

an expansion in *descending powers of* x.

By substituting various numerical values for x and y in the binomial theorem, one can generate various binomial coefficient identities:

(i) Setting $x = y = 1$ in Theorem 1.16 yields

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n,$$

which yields an alternative proof of Theorem 1.12.

(ii) Setting $x = -1$ and $y = 1$ yields, for $n \in \mathbb{P}$,

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} = \binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \cdots + (-1)^n \binom{n}{n} = 0,$$

i.e.,

$$\binom{n}{0} + \binom{n}{2} + \cdots = \binom{n}{1} + \binom{n}{3} + \cdots ,$$

i.e.,

$$\sum_{\substack{0 \le k \le n \\ k \text{ even}}} \binom{n}{k} = \sum_{\substack{0 \le k \le n \\ k \text{ odd}}} \binom{n}{k}.$$

It is important to learn to recognize when a sum is in fact a binomial expansion, or nearly so. The following illustrate a few tricks of the trade:

(iii) Simplify $\sum_{k=0}^{n} \binom{n}{k} a^k$. Solution:

$$\sum_{k=0}^{n} \binom{n}{k} a^k = \sum_{k=0}^{n} \binom{n}{k} a^k 1^{n-k} = (a+1)^n.$$

(iv) Simplify $\sum_{k=1}^{17}(-1)^k\binom{17}{k}13^k$. Solution:

$$\begin{aligned}\sum_{k=1}^{17}(-1)^k\binom{17}{k}13^k &= \sum_{k=1}^{17}\binom{17}{k}(-13)^k(1)^{17-k}\\ &= \sum_{k=0}^{17}\binom{17}{k}(-13)^k(1)^{17-k} - \binom{17}{0}(-13)^0(1)^{17}\\ &= (-13+1)^{17} - 1 = (-12)^{17} - 1.\end{aligned}$$

(v) Simplify $\sum_{k=1}^{17}(-1)^k\binom{17}{k}13^{17-k}$. Solution:

$$\begin{aligned}\sum_{k=1}^{17}(-1)^k\binom{17}{k}13^{17-k} &= \sum_{k=0}^{17}\binom{17}{k}(-1)^k(13)^{17-k} - \binom{17}{0}(-1)^0(13)^{17}\\ &= (-1+13)^{17} - 13^{17} = 12^{17} - 13^{17}.\end{aligned}$$

1.10 Trinomial Coefficients

Recall from Theorem 1.15 ff. that if $n_1 + n_2 = n$, then $\binom{n}{n_1} = \binom{n}{n_2} = \frac{n!}{n_1!n_2!}$ counts

(i) the number of sequences of n_1 x's and n_2 y's,

(ii) the number of distributions of n labeled balls among 2 urns, u_1 and u_2, such that n_i balls are placed in u_i, $i = 1, 2$,

We wish to extend these results in this section. This extension will preserve the parallels between sequences with prescribed frequency of types of terms and distributions with prescribed occupancy numbers.

We start with a concrete example: In how many ways may 2 x's, 3 y's, and 5 z's be arranged in a sequence? Using the strategy of §1.9, consider the problem of filling the 10 slots below with appropriate letters.

$$\overline{(1)}\ \overline{(2)}\ \overline{(3)}\ \overline{(4)}\ \overline{(5)}\ \overline{(6)}\ \overline{(7)}\ \overline{(8)}\ \overline{(9)}\ \overline{(10)}$$

First we choose 2 of the slots (in any of the $\binom{10}{2}$ possible ways) in which to place the 2 x's. Then, from the remaining 8 slots we choose 3 slots (in any of the $\binom{8}{3}$ possible ways) in which to place the 3 y's. The 5 z's go in the remaining 5 slots. The solution is therefore

$$\binom{10}{2}\binom{8}{3} = \frac{10!}{2!8!}\cdot\frac{8!}{3!5!} = \frac{10!}{2!3!5!}.$$

Clearly, this is also the number of ways to distribute 10 labeled balls among 3 urns, u_1, u_2 and u_3, such that 2 balls are placed in u_1, 3 balls in u_2, and 5 balls in u_3.

The number $\frac{10!}{2!3!5!}$ is a trinomial coefficient of order 10, often denoted $\binom{10}{2,3,5}$. In general, if $n, n_1, n_2, n_3 \in \mathbb{N}$ and $n_1 + n_2 + n_3 = n$,

$$\binom{n}{n_1, n_2, n_3} := \binom{n}{n_1}\binom{n-n_1}{n_2} = \frac{n!}{n_1!n_2!n_3!}$$

is termed a *trinomial coefficient of order n.*

Theorem 1.17. If $n, n_1, n_2, n_3 \in \mathbb{N}$ and $n_1 + n_2 + n_3 = n$, then $\binom{n}{n_1,n_2,n_3}$ counts

(i) the number of sequences of n_1 x's, n_2 y's, and n_3 z's, and

(ii) the number of distributions of n labeled balls among 3 urns, labeled u_1, u_2 and u_3, such that n_i balls are placed in u_i, $i = 1, 2, 3$.

Proof. Choose n_1 of the n slots (balls) in which to place the n_1 x's (which will be placed in u_1) in any of the $\binom{n}{n_1}$ possible ways. Choose n_2 of the remaining $n - n_1$ slots (balls) in which to place the n_2 y's (which will be placed in u_2) in any of the $\binom{n-n_1}{n_2}$ possible ways. Place the n_3 z's in the remaining $n - n_1 - n_2 = n_3$ slots (place the remaining $n - n_1 - n_2 = n_3$ balls in u_3.) Multiply

$$\begin{aligned}\binom{n}{n_1}\binom{n-n_1}{n_2} &= \frac{n!}{n_1!(n-n_1)!}\frac{(n-n_1)!}{n_2!(n-n_1-n_2)!} \\ &= \frac{n!}{n_1!n_2!n_3!} = \binom{n}{n_1, n_2, n_3}.\end{aligned}$$

□

Remark. Statisticians often use the abbreviated notation $\binom{n}{n_1,n_2}$ for the above trinomial coefficient, in analogy with the binomial coefficient notation $\binom{n}{n_1}$. We shall *not* follow this practice. If we write $\binom{n}{n_1,n_2}$, it will be the case that $n_1 + n_2 = n$, with $\binom{n}{n_1,n_2}$ simply being a more elaborate notation for $\binom{n}{n_1} = \binom{n}{n_2} = \frac{n!}{n_1!n_2!}$.

Corresponding to the symmetry property $\binom{n}{k} = \binom{n}{n-k}$ of binomial coefficients, we have

$$\binom{n}{n_1, n_2, n_3} = \binom{n}{m_1, m_2, m_3}$$

whenever (m_1, m_2, m_3) is a rearrangement of (n_1, n_2, n_3).

The sum of all binomial coefficients of order n is 2^n. The sum of all trinomial coefficients of order n is given by the following theorem.

Theorem 1.18. For all $n \in \mathbb{N}$,

$$\sum_{\substack{n_1+n_2+n_3=n \\ n_i \in \mathbb{N}}} \binom{n}{n_1, n_2, n_3} = 3^n.$$

Proof. Each side of the above equation counts the class of distributions of n labeled balls among 3 urns u_1, u_2 and u_3. The RHS counts this family of distributions by Theorem 1.5. The LHS counts this family in pairwise disjoint subclasses, one for each sequence (n_1, n_2, n_3) of occupancy numbers for u_1, u_2, and u_3. For by Theorem 1.17, there are $\binom{n}{n_1,n_2,n_3}$ distributions with occupancy numbers (n_1, n_2, n_3). □

Remark. One can also give a sequence counting argument for the above identity.

1.11 The Trinomial Theorem

If we expand $(x + y + z)^2 = (x + y + z)(x + y + z)$, we get, before simplification,

$$(x + y + z)^2 = xx + xy + xz + yx + yy + yz + zx + zy + zz,$$

the sum of all $3^2 = 9$ words of length 2 in the alphabet $\{x, y, z\}$. Similarly,

$$\begin{aligned}(x + y + z)^3 = {} & xxx + xxy + xxz + xyx + xyy + xyz + xzx + xzy + xzz + yxx \\ & + yxy + yxz + yyx + yyy + yyz + yzx + yzy + yzz + zxx + zxy \\ & + zxz + zyx + zyy + zyz + zzx + zzy + zzz,\end{aligned}$$

the sum of all $3^3 = 27$ words of length 3 in the alphabet $\{x, y, z\}$. After simplification, we get

$$(x + y + z)^2 = x^2 + 2xy + 2xz + y^2 + 2yz + z^2$$

and

$$(x + y + z)^3 = x^3 + 3x^2y + 3x^2z + 3xy^2 + 6xyz + 3xz^2 + y^3 + 3y^2z + 3yz^2 + z^3.$$

More generally, before simplification, the expansion of $(x+y+z)^n$ involves the sum of all 3^n words of length n in the alphabet $\{x, y, z\}$. The following "trinomial theorem" describes the situation after simplification.

Theorem 1.19. For all $n \in \mathbb{N}$,

$$(x + y + z)^n = \sum_{\substack{n_1+n_2+n_3=n \\ n_i \in \mathbb{N}}} \binom{n}{n_1, n_2, n_3} x^{n_1} y^{n_2} z^{n_3}.$$

Proof. If $n = 0$, the result is obvious, so suppose that $n \in \mathbb{P}$. By Theorem 1.17, there are $\binom{n}{n_1,n_2,n_3}$ words of length n in the alphabet $\{x, y, z\}$ in which x appears n_1 times, y appears n_2 times, and z appears n_3 times. □

In the foregoing, we have written the expansions of $(x + y + z)^2$ and $(x + y + z)^3$ in a particular order. But that order is not dictated by the formula

$$(X + y + z)^n = \sum_{\substack{n_1+n_2+n_3=n \\ n_i \in \mathbb{N}}} \binom{n}{n_1, n_2, n_3} x^{n_1} y^{n_2} z^{n_3},$$

which simply tells us to find all sequences (n_1, n_2, n_3) in $\mathbb{N}$ summing to n, form the associated terms $\binom{n}{n_1,n_2,n_3}x^{n_1}y^{n_2}z^{n_3}$, and sum these terms in whatever order we like.

In contrast, writing the binomial theorem in the form

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}$$

dictates that the expansion be written in the order $y^n + \binom{n}{1}xy^{n-1} + \cdots + \binom{n}{n-1}x^{n-1}y + x^n$. We could write the binomial theorem in such a way that the ordering of terms is not dictated, as

$$(x+y)^n = \sum_{\substack{n_1+n_2=n \\ n_i \in \mathbb{N}}} \binom{n}{n_1, n_2} x^{n_1} y^{n_2},$$

although this is rarely done.

Can we write the trinomial theorem in such a way that the ordering of the terms is dictated? The answer is affirmative, but we require double summation, as follows:

$$\begin{aligned}
(x+y+z)^n &= \sum_{i=0}^{n}\sum_{j=0}^{n-i} \binom{n}{i, j, n-i-j} x^i y^j z^{n-i-j} \\
&= \left[\binom{n}{0,0,n}x^0y^0z^n + \binom{n}{0,1,n-1}x^0y^1z^{n-1} + \cdots + \binom{n}{0,n,0}x^0y^nz^0\right] \\
&\quad + \left[\binom{n}{1,0,n-1}x^1y^0z^{n-1} + \binom{n}{1,1,n-2}x^1y^1z^{n-2} + \cdots + \binom{n}{1,n-1,0}x^1y^{n-1}z^0\right] \\
&\quad + \left[\binom{n}{2,0,n-2}x^2y^0z^{n-2} + \cdots\right] + \cdots + \left[\binom{n}{n,0,0}x^ny^0z^0\right]
\end{aligned}$$

In this expansion we first set $i = 0$ and run j from 0 to n, then set $i = 1$ and run j from 0 to $n-1$, then set $i = 2$ and run j from 0 to $n-2$, etc.

Just as we can generate various binomial coefficient identities from the binomial theorem by substituting particular values for x and y, we can generate various identities involving trinomial coefficients from the trinomial theorem. For example, setting $x = y = z = 1$ in Theorem 1.19 yields

$$\sum_{\substack{n_1+n_2+n_3=n \\ n_i \in \mathbb{N}}} \binom{n}{n_1, n_2, n_3} = 3^n,$$

which provides an alternative to the combinatorial proof of this identity offered in Theorem 1.18.

1.12 Multinomial Coefficients and the Multinomial Theorem

For all $n \in \mathbb{N}$, all $k \in \mathbb{P}$, and for every sequence $(n_1, n_2, \ldots, n_k)$ in $\mathbb{N}$ summing to n, the k-nomial coefficient of order n, $\binom{n}{n_1,n_2,\ldots,n_k}$ is defined by

$$\binom{n}{n_1, n_2, \ldots, n_k} := \binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n-n_1-n_2-\cdots-n_{k-2}}{n_{k-1}}$$

$$= \frac{n!}{n_1!n_2!\cdots n_k!}.$$

The following three theorems generalize Theorems 1.17, 1.18, and 1.19.

Theorem 1.20. For all $n \in \mathbb{N}$, all $k \in \mathbb{P}$, and for every sequence $(n_1, n_2, \ldots, n_k)$ in $\mathbb{N}$ summing to n, the k-nomial coefficient $\binom{n}{n_1,n_2,\ldots,n_k}$ counts

(i) the number of sequences of n_1 x_1's, n_2 x_2's, …, and n_k x_k's, and

(ii) the number of distributions of n labeled balls among k urns, labeled $u_1, u_2, \ldots, u_k$, such that n_i balls are placed in u_i, $i = 1, \ldots, k$.

Proof. Straightforward generalization of the proof of Theorem 1.17. □

Theorem 1.21. For all $n \in \mathbb{N}$ and all $k \in \mathbb{P}$,

$$\sum_{\substack{n_1+n_2+\cdots+n_k=n \\ n_i\in\mathbb{N}}} \binom{n}{n_1, n_2, \ldots, n_k} = k^n.$$

That is, the sum of all k-nomial coefficients of order n is k^n.

Proof. Straightforward generalization of the proof of Theorem 1.18. □

Theorem 1.22. (Multinomial theorem). For all $n \in \mathbb{N}$ and all $k \in \mathbb{P}$,

$$(x_1 + x_2 + \cdots + x_k)^n = \sum_{\substack{n_1+n_2+\cdots+n_k=n \\ n_i\in\mathbb{N}}} \binom{n}{n_1, n_2, \ldots, n_k} x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k}.$$

Proof. The proof is straightforward generalization of the proof of Theorem 1.19. □

Remark. Theorem 1.22 does not dictate the order in which the terms of the simplified expansion of $(x_1+x_2+\cdots+x_k)^n$ appear. One way to dictate that order is to use $(k-1)$-fold summation, e.g.,

$$(x_1 + x_2 + \cdots + x_k)^n =$$
$$\sum_{n_1=0}^{n}\sum_{n_2=0}^{n-n_1}\sum_{n_3=0}^{n-n_1-n_2}\cdots\sum_{n_{k-1}=0}^{n-n_1-\cdots-n_{k-2}} \binom{n}{n_1, n_2, \ldots, n_{k-1}, n-n_1-\cdots-n_{k-1}} \times$$
$$\times\, x_1^{n_1} x_2^{n_2} \cdots x_{k-1}^{n_{k-1}} x_k^{n-n_1-\cdots-n_{k-1}}.$$

1.13 Problems

1. In how many ways can a committee of 5 be chosen from 8 people?

2. In how many ways can a selection of fruit be made from 7 plums, 4 lemons, and 9 oranges? (Assume that the 7 plums are indistinguishable; likewise for the lemons and for the oranges.)

3. Answer question 2 if at least 2 plums, at least one lemon, and at least 3 oranges must be chosen.

4. Ten points are marked on the circumference of a circle. How many chords can be drawn by joining them in all possible ways? With these 10 points as vertices, how many triangles can be drawn? How many hexagons?

5. In how many ways can a selection of 4 CD's be made from 9? If a certain CD must be chosen, in how many ways can the selection be made? In how many ways can it be made if a certain CD is excluded?

6. From 20 sailors, 3 must be assigned to navigation, 5 to communications, and 12 to maintainance. In how many ways can such an assignment be made?

7. From 8 men and 10 women, in how many ways may a committee consisting of 3 men and 3 women be chosen?

8. A publisher has 90 copies of one book and 60 of another. In how many ways can these books be divided among 2 booksellers if no bookseller gets more than two-thirds of the copies of either book?

9. How many 5-letter words, each consisting of 3 consonants and 2 vowels, can be formed from the letters of the word *equations*?

10. Find the number of arrangements of the letters of the word *committee*, using all the letters in each arrangement.

11. How many different numbers can be obtained by arranging the digits 2233344455, all together, in all possible ways?

12. How many sequential arrangements of the letters of the word *institution* are there? How many of these begin with t and end with s?

13. Determine the size of the smallest set having at least 100 proper, nonempty subsets.

14. How many subsets of [1000] have odd cardinality?

15. In how many ways can a selection of at least one book be made from 8 different books?

16. In how many ways may a search party of 3 or more people be chosen from 10 individuals?

17. Wendy's advertises "We fix 'em 1024 ways!" How many hamburger toppings (lettuce, tomato, etc.) must they stock to make this claim?

18. How many distributions of balls $1, 2, \ldots, 25$ among urns 1, 2, 3, and 4 are there such that urn 1 gets 2 balls, urn 2 gets 5 balls, urn 3 gets 7 balls, and urn 4 gets 11 balls?

19. What is the cardinality of the smallest set having at least 1000 subsets of even cardinality?

20. What is the sum of all 8-nomial coefficients of order 3?

21. How many 5-element subsets of $[10]$ contain at least one of the members of $[3]$? How many 5-element subsets of $[10]$ contain 2 odd and 3 even numbers?

22. Simplify the following sums

$$\text{(a)} \quad \sum_{j=1}^{513} (-1)^j \binom{513}{j} a^j \qquad\qquad \text{(b)} \quad \sum_{r=0}^{13} (-1)^r \binom{13}{r} 5^{13-r}$$

$$\text{(c)} \quad \sum_{k=0}^{18} \binom{18}{k} e^{-k} \qquad\qquad \text{(d)} \quad \sum_{k=0}^{25} \binom{37}{k} \binom{63}{25-k}.$$

23. Simplify the sum

$$\sum_{i=0}^{10} \sum_{j=0}^{10-i} (-1)^i \frac{10!}{i!j!(10-i-j)!} 2^j.$$

24. How many ways are there to distribute n balls, labeled $1, \ldots, n$, among 2 urns, labeled u_1 and u_2, if at least one ball must be placed in each urn?

25. How many ways are there to distribute 5 balls, labeled $1, \ldots, 5$, among 3 urns, labeled u_1, u_2, and u_3, if at least one ball must be placed in each urn?

CHAPTER 2

BASIC PROBABILITY THEORY

2.1 Observed Relative Frequencies

The notion of probability is closely connected with the notion of relative frequency. Consider, for example, the *experiment* of tossing a six-sided die with sides numbered 1 through 6 and recording the number showing on the top side. The *set of possible outcomes* of this experiment is $S = \{1, 2, 3, 4, 5, 6\}$. One also calls S the *sample space* of the experiment or the *frame of discernment* of the experiment. There is some latitude in choosing S, depending on how specifically we choose to describe outcomes. If we only cared about the parity of the outcome we might choose to use the sample space $S' = \{\text{odd, even}\}$ instead. So, strictly speaking, we should refer to *a* set of possible outcomes (sample space, frame of discernment) of an experiment, rather than to *the* set of possible outcomes.

Suppose that we repeat the above experiment 25 times, and record the outcomes, getting, say, the *sequence of observed outcomes*

$$(4, 2, 2, 6, 5, 1, 6, 6, 4, 4, 4, 5, 4, 5, 4, 3, 1, 5, 5, 1, 4, 2, 2, 4, 2). \tag{2.1}$$

For each $E \subseteq S$, define the *observed relative frequency of E in the sequence* (2.1), denoted ORF (E), by

$$\text{ORF } (E) = \frac{1}{25}(\#\text{ of times an outcome of the sequence (2.1) is an element of } E) \tag{2.2}$$

Using (2.2), we may calculate ORF $(\{4\}) = 8/25$, ORF $(\{1, 2\}) = 8/25$, ORF $(\{1, 3, 5\}) = 9/25$, etc. Let us note some basic properties of the function ORF. We have

$$0 \le \text{ORF } (E) \le 1 \qquad \text{for all } E \subseteq S, \tag{2.3}$$

$$\text{ORF } (\emptyset) = 0 \quad \text{and} \quad \text{ORF } (S) = 1, \text{and} \tag{2.4}$$

$$\text{if } E_1, E_2 \subseteq S \text{ and } E_1 \cap E_2 = \emptyset, \text{ then} \tag{2.5}$$

$$\text{ORF } (E_1 \cup E_2) = \text{ORF } (E_1) + \text{ORF } (E_2).$$

2.2 Predicted Relative Frequencies

Suppose now that, instead of having actually tossed the die in question some number of times, we were about to toss it some number of times, and wanted to *predict* the relative frequencies that will be observed when the die is actually tossed. Much of this course will be devoted to strategies that can be used to make such predictions. Supppose that a person employs one of these strategies, thereby assigning each $E \subseteq S$ a number PRF$\,(E)$, the *predicted relative frequency of* E in the sequence of tosses about to be performed. Since PRF$\,(E)$ is that person's best guess as to what the observed relative frequency of E will turn out to be once the tosses are actually performed, it is clear that the function PRF should have the same basic properties as ORF, as listed in (2.3)-(2.5) above. That is, any function worthy of the name PRF ought to satisfy the following properties:

(2.6) $$0 \leq \text{PRF}\,(E) \leq 1 \quad \text{for all } E \subseteq S,$$

(2.7) $$\text{PRF}\,(\emptyset) = 0 \quad \text{and} \quad \text{PRF}\,(S) = 1, \text{ and}$$

(2.8) $$\text{if } E_1, E_2 \subseteq S \text{ and } E_1 \cap E_2 = \emptyset,$$
$$\text{then PRF}\,(E_1 \cup E_2) = \text{PRF}\,(E_1) + \text{PRF}\,(E_2).$$

2.3 Probability Measures

What we've said above applies of course to many situations in which some "experiment" has been or will be performed some number of times. The experiment might involve games of chance (tossing a die or pair of dice, choosing one or more cards at random from a deck, spinning a roulette wheel) or more serious things like making a scientific measurement, recording the outcome of a random selection from some population of things or individuals, observing the lifespan of an individual or piece of equipment, etc., etc.

In all such cases the conceptual apparatus is the same. We choose a set S of possible outcomes of a single instance of the experiment. Any set $E \subseteq S$ is called an *event*. The notions of observed relative frequency of E in an actual sequence of experiments and of predicted relative frequency of E in an about-to-be performed sequence of experiments are defined just as they are in the die tossing example above. In particular, these functions will satisfy, respectively, (2.3)-(2.5) and (2.6)-(2.8).

It gets kind of tedious saying "observed relative frequency" and "predicted relative frequency." There is, however, a simple term that covers both sorts of functions. That term is *probability measure*. Here is a simultaneous treatment of observed and predicted relative frequencies in "probabilistic" terms: Let S be a sample space (set of possible outcomes, frame of discernment) of an experiment. A *probability measure* is a function P that assigns to each event $E \subseteq S$ a number $P(E)$, called *the probability of* E, satisfying the following axioms:

(2.9) $$0 \leq P(E) \leq 1 \text{ for all events } E \subseteq S,$$

(2.10) $$P(\emptyset) = 0 \text{ and } P(S) = 1, \text{ and}$$

(2.11) $$\text{for all events } E, F \subseteq S, \text{ if } E \cap F = \emptyset,$$
$$\text{then } P(E \cup F) = P(E) + P(F).$$

In addition, we require for certain technical reasons that will not play a large role in this course, that

(2.12) $$\text{if there is an infinite sequence of events } E_1, E_2, \ldots \text{ with } E_i \cap E_j = \emptyset$$

$$\text{whenever } i \neq j, \text{ then } P(E_1 \cup E_2 \cup \cdots) = P(E_1) + P(E_2) + \cdots$$
$$[\text{convergent infinite series}].$$

We call (2.11) the *additivity axiom* and (2.12) the *countable additivity axiom.* This axiomatic approach to probability was developed by the Russian mathematician Kolmogorov in the 1930's. For much of the development of probability theory, we can treat P, as one says, "abstractly." That is, we need not be concerned whether the function P in question represents observed or predicted relative frequencies. But if we do want to make this distinction, we will do it using simpler language. If P represents observed relative frequencies, we'll call P an *empirical probability.* If P represents predicted relative frequencies, we'll call P a *theoretical probability* or *probability model.*

2.4 Basic Theorems of Probability Theory

Using Axioms (2.9)-(2.12) we can prove as theorems certain other basic properties that hold for every probability measure P.

Theorem 2.1. For all events E,

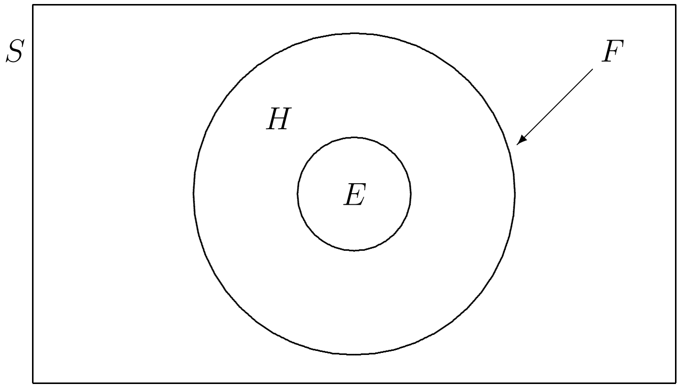
$$P(E) + P(\bar{E}) = 1. \tag{2.13}$$

Proof. We have $E \cap \bar{E} = \emptyset$, so $P(E \cup \bar{E}) = P(E) + P(\bar{E})$ by (2.11). But $E \cup \bar{E} = S$, and $P(S) = 1$ by (2.10), and so $P(E) + P(\bar{E}) = 1$. □

Remark. This theorem is usually used in the form $P(\bar{E}) = 1 - P(E)$.

Theorem 2.2. If E and F are events and $E \subseteq F$, then $P(E) \leq P(F)$.

Proof. Consider the following Venn Diagram:



By (2.11), $P(F) = P(E) + P(H)$, where $H = F \cap \bar{E}$. So $P(F) - P(E) = P(H)$. Since $P(H) \geq 0$ by (2.9), we have $P(F) - P(E) \geq 0$, i.e., $P(F) \geq P(E)$. □

Theorem 2.3. If $E_1, E_2, \ldots, E_n$ are events with $E_i \cap E_j = \emptyset$ whenever $i \neq j$, then

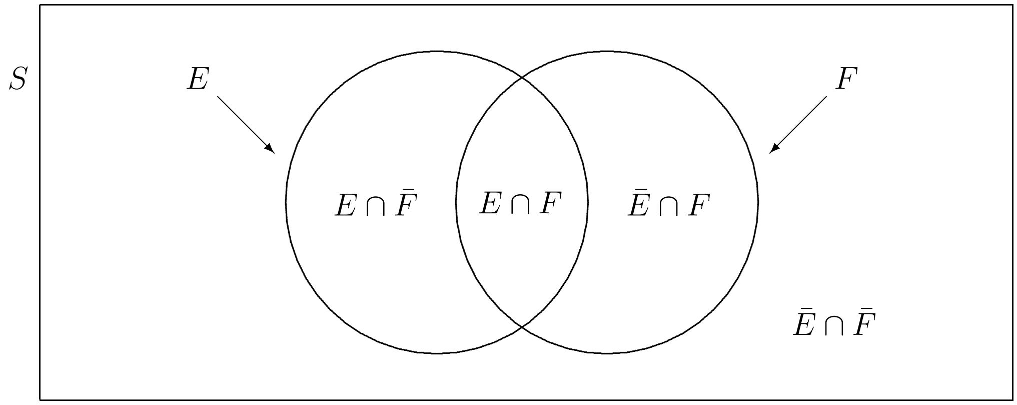
$$P(E_1 \cup E_2 \cup \cdots \cup E_n) = P(E_1) + P(E_2) + \cdots + P(E_n). \tag{2.14}$$

Proof. By induction on n, using (2.11). □

Theorem 2.4. For all events E and F

$$P(E \cup F) = P(E) + P(F) - P(E \cap F). \tag{2.15}$$

Proof. Consider the following Venn diagram:



By (2.11),

$$\begin{aligned} P(E \cup F) &= P(E) + P(\bar{E} \cap F) \\ &= P(E) + \big(P(\bar{E} \cap F) + P(E \cap F)\big) - P(E \cap F) \\ &= P(E) + P(F) - P(E \cap F). \qquad \square \end{aligned} \tag{2.16}$$

Remark. Using Theorem 2.4, one can prove that

$$P(E \cup F \cup G) = P(E) + P(F) + P(G) - P(E \cap F) - P(E \cap G) - P(F \cap G) + P(E \cap F \cap G), \tag{2.17}$$

and, more generally, that

$$\begin{aligned} P(E_1 \cup E_2 \cup \cdots \cup E_n) &= \sum_{1 \le i \le n} P(E_i) - \sum_{1 \le i < j \le n} P(E_i \cap E_j) \\ &\quad + \sum_{1 \le i < j < k \le n} P(E_i \cap E_j \cap E_k) - \cdots + (-1)^{n-1} P(E_1 \cap E_2 \cap \cdots \cap E_n). \end{aligned} \tag{2.18}$$

Formula (2.18) is called the *principle of inclusion and exclusion for probability measures.*

2.5 The Uniform Probability Model for a Finite Sample Space

We will study many different probability models in this course and try to develop intuitions for when a given model is likely to be appropriate. We'll start with the simplest of these models, the *uniform probability model for a finite sample space.*

Let S be a finite sample space. The *uniform probability model* P for S is defined by

$$P(E) = \frac{|E|}{|S|}. \tag{2.19}$$

It is easy to check that P, as defined by (2.19) satisfies the axioms for a probability measure. If we adopt this model, we are predicting that in a sequence of experiments with sample space S the observed relative frequency of E *in the sequence of experiments* will be equal to the relative frequency, $|E|/|S|$, of E *in the set* S. In particular, for each outcome $s \in S$, we have $P(\{s\}) = \frac{1}{|S|}$, or as one says, "All outcomes in S are equally likely (i.e., equally probable)." Incidentally, for events $E = \{s\}$ consisting of a single outcome s, we will be careless in our notation and write $P(s)$ instead of the correct, but cumbersome, $P(\{s\})$. Events consisting of a single outcome are sometimes called *elementary events.*

When is it appropriate to use the uniform probability model? Here are two classes of cases:

1° Let S be any finite set. The experiment consists of choosing an element of S "at random." Obviously, the set of possible outcomes is S itself. The phrase "at random" tells us to use the uniform model. Incidentally, events E will often be described in words, rather than by a list of their elements.

Example 1. A permutation of the numbers $1, 2, 3$ is selected at random. What is the probability that 1 and 2 are adjacent? Here $S = \{123, 132, 213, 231, 312, 321\}$. The event $E = \{123, 213, 312, 321\}$. So $P(1 \text{ and } 2 \text{ adjacent}) = P(E) = |E|/|S| = 4/6 = 2/3$.

Remark. We could have determined $|S|$ and $|E|$ using combinatorics instead of brute force listing, and you should do this whenever you can.

2° In games of chance, whenever we are told that the gaming devices (dice, coins, roulette wheel, etc.) are *fair*, we use the uniform model.

Example 2. A fair die is tossed. What is the probability that an odd number shows?

Here $S = \{1, 2, 3, 4, 5, 6\}$ and the event E in question is $E = \{1, 3, 5\}$. So

$$P(\text{odd outcome}) = P(\{1, 3, 5\}) = 3/6 = 1/2.$$

Of course, our predicted relative frequency may differ from the observed relative frequency once the experiment is carried out a number of times. In the very first example in these notes, I generated the sequence in question by simulating a fair die, but got for that sequence $\mathrm{ORF}(\{1,3,5\}) = 9/25$, considerably less than the $1/2$ predicted by the uniform model. For a larger number of repetitions of the experiment, we may expect prediction to more closely approximate observation. This is the essence of the famous *law of large numbers* of probability theory, a result that we will take up later.

CAUTION. Formula (2.19) is sometimes presented in high school of *the* definition of probability. In fact, it defines just one of many possible probability models. *In a large number of interesting experiments, the outcomes in S are not equally likely and* (2.19) *is inapplicable.* Chapters 3 and 4 of these notes introduce a variety of non-uniform probability models.

We conclude this section with a famous problem.

Example 3 (the birthday problem). There are n people in a room. What is the probability that at least two of these people have the same birthday (i.e., have birthdays on the same day and month of the year)? What is the smallest value of n such that the probability is $1/2$ or better that two or more people have the same birthday?

Solution. We neglect February 29 and assume a year of 365 days. The sample space consists of 365^n n-tuples $(b_1, b_2, \ldots, b_n)$ where b_i is the birthday of the i^{th} person in the room. Let p_n be the probability in question. Clearly,

$$\begin{aligned} p_n &= 1 - P(\text{all } n \text{ individuals have different birthdays}) \\ &= 1 - \frac{365^{\underline{n}}}{365^n}, \end{aligned}$$

assuming that all n-tuples $(b_1, \ldots, b_n)$ are equally likely, which seems reasonable. Here is a table of some values of p_n:

$\underline{n}$	$\underline{p_n}$
5	.027
10	.117
20	.411
23	.507
30	.706
40	.891
60	.994

Remarkably, in a room with just 23 people, the probability is greater than $1/2$ that two or more people have the same birthday.

2.6 Problems

1. If we choose at random a permutation of (all of) the letters a, b, c, and d, what is the probability that the letters a and b will *not* be adjacent?

2. From a group of 8 men and 5 women a random selection of 4 individuals is made.

   a. What is the probability that 2 men and 2 women are chosen?

   b. What is the probability that all of the individuals chosen are women?

   c. What is the probability that at least one of the individuals chosen is a man?

3. a. Suppose that 30% of the legislature voted for term limits and 50% voted to cut taxes. If 40% voted against both these measures, what percentage voted for both measures?

   b. If 10% of the objects in a bin are blue spheres and 50% of the objects are blue and/or spherical, what percentage of the objects are

   (i) neither blue nor spherical?

   (ii) either blue or spherical, but not both?

4. If 5 individuals, a, b, c, d, and e line up at random at a ticket counter what is the probability that

   a. individual a is first in line? 3^{rd} in line?

   b. neither individual a nor individual b is 2^{nd} in line?

5. A red die and a clear die are tossed. Write out the sample space S consisting of all possible outcomes (r, c) where $1 \leq r \leq 6$ and $1 \leq c \leq 6$. Assuming that each of these outcomes is equally likely (i.e., assuming the uniform probability model), determine

   a. the probability that $r > c$.

   b. the probability that $r + c = 10$.

   c. the probability that $r = c$.

   d. the probability that $r \geq 4$ and $c \geq 3$.

   e. the probability that $r \geq 4$ or $c \geq 3$ (note: in mathematics, "or" *always* means "and/or").

   f. the probability that r is even.

6. If 30% of the people staying at a conference hotel are females, 50% are mathematicians, and 70% are females or mathematicians, what percentage are female mathematicians?

7. If I select a permutation of $\{1, 2, 3, 4\}$ at random, what is the probability that no number is in its "natural" position?

8. In a room of n people, what is the probability q_n that at least two have the same birth month? Find the smallest n such that $q_n \geq 1/2$ and the smallest n such that $q_n = 1$.

9. Let S be a finite set. For all $E \subseteq S$, define $P(E) = |E|/|S|$. Prove that P is a probability measure.

10. Prove that if P is a probability measure and E_1 and E_2 are any events, then

    a. $P(E_1 \cup E_2) \leq P(E_1) + P(E_2)$

    b. $P(E_1 \cap E_2) \geq P(E_1) + P(E_2) - 1$

2.7 Conditional Relative Frequencies

Suppose that, among a set S of 200 voters, 60 are female Republicans, 50 are female Democrats, 60 are male Republicans, and 30 are male Democrats. These data are entered in the table below, with row and column sums entered, respectively, in the right hand and bottom *margins* of the table:

(2.20)

	R	D	
F	60	50	110
M	60	30	90
	120	80	200

The above table is an example of a *two-way frequency table*, specifically, a 2×2 (*not* 3×3) *frequency table.* If we divide each entry in the above table by 200, we get the associated 2×2 *relative frequency table* (also known as a 2×2 *contingency table*) below:

(2.21)

	R	D	
F	.30	.25	.55
M	.30	.15	.45
	.60	.40	1.00

The entries in the above table are, of course, probabilities (specifically, empirical probabilities). For example, $P(F) = .55$ is the fraction (expressed as a decimal) of females in S, $P(R) = .60$ the fraction of Republicans in S, and $P(F \cap D) = .25$ the fraction of female Democrats in S.

$P(F \cap D)$, the fraction of female Democrats in S, should not be confused with the fraction of females *among* Democrats in S, which is denoted by $P(F|D)$. From table (2.20) it is clear that $P(F|D) = |F \cap D|/|D| = 50/80 \approx .63$. On the other hand, $P(F|R)$, the fraction of females among Republicans in S, is given by $P(F|R) = |F \cap R|/|R| = 60/120 = .50$. So $P(F|R) < P(F) < P(F|D)$. This result is described in words by saying that females are *underrepresented* among Republicans (in the sense that females constitute a smaller fraction of Republicans than they do of the population S) and *overrepresented* among Democrats (in the sense that females constitute a larger fraction of Democrats than they do of the population S).

$P(F|D)$, the fraction of females among Democrats in S, should not be confused with $P(D|F)$, the fraction of Democrats among females. We have $P(D|F) = |D \cap F|/|F| = 50/110 \approx .45$. Since $P(D) = .40$, we see that $P(D|F) > P(D)$, i.e., that Democrats are overrepresented among females. This is no surprise, for as we shall prove later, a category A is overrepresented in a category B if and only if B is overrepresented in A. Similarly, A is underrepresented in B if and only if B is underrepresented in A.

Consider the following variant of table (2.20):

(2.22)

	R	D	
F	72	48	120
M	48	32	80
	120	80	200

The relative frequency table associated with (2.3) is

$$\begin{array}{cccc} & R & D & \\ F & .36 & .24 & .60 \\ M & .24 & .16 & .40 \\ & .60 & .40 & 1.00 \end{array} . \tag{2.23}$$

Here $P(F|R) = P(F) = P(F|D)$, i.e., females are *proportionally represented* among Republicans, and also among Democrats. Readers may wish to check that, conversely, Republicans are proportionally represented among females, and that Democrats are also proportionally represented among females.

Remark. In the foregoing, we have read an expression of the form $P(A|B)$ as *the fraction of* A*'s among* B*'s*, since this locution is simple and clear. Alternatively, $P(A|B)$ is sometimes called the *conditional relative frequency of* A *in* B, or the *conditional empirical probability of* A*, given* B.

The general 2×2 frequency table takes the form

$$\begin{array}{cccc} & B & \bar{B} & \\ A & |A \cap B| & |A \cap \bar{B}| & |A| \\ \bar{A} & |\bar{A} \cap B| & |\bar{A} \cap \bar{B}| & |\bar{A}| \\ & |B| & |\bar{B}| & |S| \end{array} \tag{2.24}$$

where A and B are subsets of the finite set S. The 2×2 relative frequency table associated to (2.24) is

$$\begin{array}{cccc} & B & \bar{B} & \\ A & P(A \cap B) & P(A \cap \bar{B}) & P(A) \\ \bar{A} & P(\bar{A} \cap B) & P(\bar{A} \cap \bar{B}) & P(\bar{A}) \\ & P(B) & P(\bar{B}) & 1.00 \end{array} , \tag{2.25}$$

where $P(A \cap B) = |A \cap B|/|S|$, $P(A \cap \bar{B}) = |A \cap \bar{B}|/|S|$, etc.

Up to this point, we have computed a conditional relative frequency, such as $P(A|B)$ by the formula

$$P(A|B) = \frac{|A \cap B|}{|B|}. \tag{2.26}$$

But if we divide the numerator and denominator of the right hand side of (2.26) by $|S|$, we see that $P(A|B)$ can also be computed by the formula

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{2.27}$$

This is an important observation. In addition to establishing that conditional relative frequencies can be computed directly from a relative frequency table, it motivates the general definition of conditional probability that we shall introduce in the next section.

In conclusion, we mention that there are frequency and relative freqency tables of arbitrary finite dimension. Given a finite set S, a sequence $(A_1, \ldots, A_m)$ of subsets of S is called an *ordered partition of* S if the A_i's are nonempty and pairwise disjoint, and $A_1 \cup \cdots \cup A_m = S$. Given ordered partitions $(A_1, \ldots, A_m)$ and $(B_1, \ldots, B_n)$ of S, the associated $m \times n$ frequency table records the number $|A_i \cap B_j|$ in the cell in the i^{th} row and j^{th} column of the table, with row and column sums recorded, respectively, in the right-hand and bottom margins of the table. In the associated $m \times n$ relative frequency table, the number $P(A_i \cap B_j) = |A_i \cap B_j|/|S|$ is recorded in the cell in the i^{th} row and j^{th} column.

2.8 Conditional Probability

A conditional relative frequency $P(A|B)$ may be thought of as a "revision" of the relative frequency $P(A)$ that is appropriate given a narrowing of our focus from the population S to the sub-population B. A similar revision is called for in the case of probability models, when we discover partial information about the outcome of an experiment, namely that the outcome is an element of the event B. Assuming that $P(B) > 0$, we would revise the *prior probability* $P(A)$ to the *conditional probability of* A*, given* B, again denoted by $P(A|B)$, and defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{2.28}$$

The justification for using this formula when P is a probability model (i.e., when P records predicted relative frequencies) is, once again, that predicted relative frequencies ought to have the same properties as observed relative frequencies.

Here is an application of (2.28). We have constructed a probability model for the outcome of one toss of a fair die. Our sample space is $S = \{1, 2, 3, 4, 5, 6\}$ and it is endowed with the uniform probability measure P. Let A = "outcome is prime" = $\{2, 3, 5\}$. Then $P(A) = 1/2$. Suppose that we are told that the outcome of the toss in question was a number ≥ 4. Let $B = \{4, 5, 6\}$. Then

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{5\})}{P(\{4, 5, 6\})} = \frac{1/6}{3/6} = \frac{1}{3}.$$

Knowing that the event B has occurred has prompted us to revise the probability of a prime outcome downward from 1/2 to 1/3.

Just as the basic theory of probability can be treated abstractly, without the necessity of differentiating empirical probabilities from probability models, the same is true of the basic theory of conditional probability. We use formula (2.28) regardless of whether P is an observed or predicted relative frequency.

The first thing to observe is that conditional probabilities satisfy the same basic properties as ordinary probabilities:

Theorem 2.5. Let P be a probability measure on the sample space S, and suppose that $P(B) > 0$. Define $P(A|B) = P(A \cap B)/P(B)$ for all events $A \subseteq S$. Then

(2.29) $$0 \leq P(A|B) \leq 1 \text{ for all events } A \subseteq S,$$

(2.30) $$P(\emptyset|B) = 0 \text{ and } P(S|B) = 1, \text{ and}$$

(2.31) $$P(E \cup F|B) = P(E|B) + P(F|B) \text{ for all events } E, F \subseteq S \text{ with } E \cap F = \emptyset.$$

Proof. By (2.9), $P(A \cap B) \geq 0$. Since $P(B) > 0$, $P(A|B) = P(A \cap B)/P(B) \geq 0$. Since $A \cap B \subset B$, it follows from Theorem 2.2 that $P(A \cap B) \leq P(B)$. Hence $P(A|B) \leq 1$. This proves (2.29).

We next prove (2.30). By (2.10), $P(\emptyset|B) = P(\emptyset \cap B)/P(B) = P(\emptyset)/P(B) = 0/P(B) = 0$. In fact, $P(E|B) = 0$ whenever $E \cap B = \emptyset$, which makes sense intuitively [why?]. In particular, $P(\bar{B}|B) = 0$. Also, $P(S|B) = P(S \cap B)/P(B) = P(B)/P(B) = 1$. In fact, $P(E|B) = 1$ whenever $B \subseteq E$, which also makes sense intuitively [why?].

To prove (2.31), we note that $(E \cup F) \cap B = (E \cap B) \cup (F \cap B)$, and that $E \cap B$ and $F \cap B$ are disjoint since E and F are disjoint. So

$$\begin{aligned} P(E \cup F|B) &= \frac{P((E \cup F) \cap B)}{P(B)} = \frac{P((E \cap B) \cup (F \cap B))}{P(B)} \\ &= \frac{P(E \cap B) + P(F \cap B)}{P(B)} = \frac{P(E \cap B)}{P(B)} + \frac{P(F \cap B)}{P(B)} \\ &= P(E|B) + P(F|B). \qquad \square \end{aligned}$$

Remark. Using a proof like that of (2.31) one can show that if $(E_1, E_2, \ldots)$ is any infinite sequence of pairwise disjoint events, then

(2.32) $$P(E_1 \cup E_2 \cup \cdots |B) = P(E_1|B) + P(E_2|B) + \cdots .$$

Another way to put (2.29)-(2.32), is that if P is a probability measure on S with $P(B) > 0$ and we define Q for all events $A \subseteq S$ by

$$Q(A) = P(A|B),$$

then Q is a probability measure on S. This enables us to deduce immediately the following conditional analogues of Theorems 2.1, 2.2, 2.3, and (2.4):

(2.33) $$P(A|B) + P(\bar{A}|B) = 1, \quad \text{for all events } A \subseteq S,$$

(2.34) $$\text{If } E \subseteq F, \text{ then } P(E|B) \leq P(F|B), \text{ for all events } E, F \subseteq S.$$

(2.35) If $E_1, \ldots, E_n$ are pairwise disjoint events, then
$$P(E_1 \cup \cdots \cup E_n|B) = P(E_1|B) + \cdots + P(E_n|B).$$

(2.36) For all events $E, F \subseteq S$,
$$P(E \cup F|B) = P(E|B) + P(F|B) - P(E \cap F|B).$$

2.9 Representation and Relevance

The following theorem states an important set of equivalences:

Theorem 2.6. If P is a probability measure on S and $A, B \subseteq S$, the following inequalities are equivalent:

$$P(A|B) > P(A) \tag{2.37}$$

$$P(A \cap B) > P(A)P(B) \tag{2.38}$$

$$P(A|B) > P(A|\bar{B}) \tag{2.39}$$

$$P(A \cap B)P(\bar{A} \cap \bar{B}) > P(A \cap \bar{B})P(\bar{A} \cap B). \tag{2.40}$$

Proof. Up to this point we have only used two-way tables to represent empirical probabilities. But they are useful for representing arbitrary probabilities. Suppose here that $P(A \cap B) = a$, $P(A \cap \bar{B}) = b$, $P(\bar{A} \cap B) = c$, and $P(\bar{A} \cap \bar{B}) = d$. In tabular form we have

$$P: \quad \begin{array}{cccc} & B & \bar{B} & \\ A & a & b & a+b \\ \bar{A} & c & d & c+d \\ & a+c & b+d & 1.00 \end{array} . \tag{2.41}$$

To show the equivalence of (2.37) and (2.38), we note that

$$\begin{aligned} P(A|B) > P(A) &\Leftrightarrow \frac{P(A \cap B)}{P(B)} > P(A) \\ &\Leftrightarrow P(A \cap B) > P(A)P(B). \end{aligned}$$

To show the equivalence of (2.39) and (2.40), we note that

$$\begin{aligned} P(A|B) > P(A|\bar{B}) &\Leftrightarrow \frac{P(A \cap B)}{P(B)} > \frac{P(A \cap \bar{B})}{P(\bar{B})} \\ &\Leftrightarrow \frac{a}{a+c} > \frac{b}{b+d} \\ &\Leftrightarrow ab + ad > ab + bc \\ &\Leftrightarrow ad > bc \\ &\Leftrightarrow P(A \cap B)P(\bar{A} \cap \bar{B}) > P(A \cap \bar{B})P(\bar{A} \cap B). \end{aligned}$$

We complete the proof by showing the equivalence of (2.38) and (2.40), as follows:

$$\begin{aligned} P(A \cap B) > P(A)P(B) &\Leftrightarrow a > (a+b)(a+c) \\ &\Leftrightarrow a > a^2 + ac + ab + bc \\ &\Leftrightarrow a > a(a+b+c) + bc \\ &\Leftrightarrow a > a(1-d) + bc \end{aligned}$$

$$\begin{aligned}&\Leftrightarrow ad > bc\\ &\Leftrightarrow P(A\cap B)P(\bar{A}\cap\bar{B}) > P(A\cap\bar{B})P(\bar{A}\cap B). \qquad \square\end{aligned}$$

If P is any probability measure for which any (hence, all) of the inequalities (2.37)-(2.40) hold, we say that B is *positively relevant to* A. We can use this terminology regardless of whether P is empirical or a probability model. As noted in §2.7, however, when P is empirical, we also say in case (2.37)-(2.40) hold that A is *overrepresented in* B.

From (2.38) it is pretty clear that positive relevance is a symmetric relation on events. This is established in the following theorem.

Theorem 2.7. If B is positively relevant to A with respect to the probability measure P, then A is positively relevant to B with respect to P.

Proof. The slick proof of this simply uses the "symmetry" of (2.38) in A and B, i.e., the fact that (2.38) remains true when A and B are interchanged.

A lengthier, but perhaps more salient, proof goes as follows:

B is positively relevant to $A \Leftrightarrow P(A|B) > P(A) \Leftrightarrow \frac{P(A\cap B)}{P(B)} > P(A) \Leftrightarrow$ $P(A\cap B) > P(A)P(B) \Leftrightarrow P(B\cap A) > P(A)P(B) \Leftrightarrow \frac{P(B\cap A)}{P(A)} > P(B) \Leftrightarrow P(B|A) > P(B)$ $\Leftrightarrow A$ is positively relevant to B. □

Remark 1. By (2.40), (2.41), and Theorem 2.7, A and B are positively relevant to each other if and only if $ad > bc$, i.e., if and only if $ad - bc > 0$. In other words, A and B are positively relevant to each other if and only if the determinant of the matrix

$$\begin{bmatrix} a & b\\ c & d\end{bmatrix}$$

is positive.

Remark 2. The event A is said to *imply* the event B if $A \subseteq B$. In other words, A implies B if the occurrence of A guarantees the occurrence of B. Of course, implication is not a symmetric relation (e.g., being a resident of Tennessee implies that one is a resident of the USA, but not conversely.) Positive relevance is an attenuated sort of implication, since if A is positively relevant to B, the occurrence of A raises the probability of the occurrence of B. Unlike implication, however, positive relevance *is* a symmetric relation.

The following theorem states another important set of equivalences:

Theorem 2.8. If P is a probability measure on S and $A, B \subseteq S$, the following inequalities are equivalent:

(2.42) $$P(A|B) < P(A)$$

(2.43) $$P(A\cap B) < P(A)P(B)$$

(2.44) $$P(A|B) < P(A|\bar{B})$$

(2.45) $$P(A\cap B)P(\bar{A}\cap\bar{B}) < P(A\cap\bar{B})P(\bar{A}\cap B).$$

Proof. The proof of this theorem may be gotten from the proof of Theorem 2.6 simply by replacing each occurrence of the symbol $>$ with the symbol $<$. □

If P is any probability measure for which any (hence, all) of the inequalities (2.42)-(2.45) hold, we say that B is *negatively relevant to* A. We can use this terminology regardless of whether P is empirical or a probability model. As noted in §2.7, however, when P is empirical, we also say in case (2.42)-(2.45) hold that A is *underrepresented in* B.

Of course, negative relevance is also a symmetric relation on events.

Theorem 2.9. If B is negatively relevant to A with respect to the probability measure P, then A is negatively relevant to B with respect to P.

Proof. Replace each occurrence of the symbol $>$ in the proof of Theorem 2.7 with the symbol $<$. □

Remark. By (2.40), (2.45), and Theorem 2.9, A and B are negatively relevant to each other if and only if $\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} < 0$.

The following theorem states yet another important set of equivalences:

Theorem 2.10. If P is a probability measure on S, and $A, B \subseteq S$, the following identities are equivalent:

$$P(A|B) = P(A) \tag{2.46}$$

$$P(A \cap B) = P(A)P(B) \tag{2.47}$$

$$P(A|B) = P(A|\bar{B}) \tag{2.48}$$

$$P(A \cap B)P(\bar{A} \cap \bar{B}) = P(A \cap \bar{B})P(\bar{A} \cap B). \tag{2.49}$$

Proof. Replace each occurrence of the symbol $>$ in the proof of Theorem 2.6 with the symbol $=$. □

If P is any probability measure for which any (hence, all) of the identities (2.46)-(2.49) hold, one would expect, based on our earlier choice of terminology, to say that B is *irrelevant to* A. While some people use this terminology, it is much more common to say in this case that A is *independent of* B. We can use this terminology regardless of whether P is empirical or a probability model. As noted in §2.7, however, when P is empirical, we also say in case (2.46)-(2.49) hold that A is *proportionally represented in* B.

Of course, independence is also a symmetric relation on events.

Theorem 2.11. If A is independent of B with respect to the probability measure P, then B is independent of A with respect to P.

Proof. Replace each occurrence of the symbol $>$ in the proof in Theorem 2.7 with the symbol $=$. □

Remark 1. By (2.40), (2.49), and Theorem 2.11, A and B are independent of each other if and only if $\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = 0$.

Remark 2. It is customary simply to say that A and B are independent, omitting the phrase "of each other."

Remark 3. Most probability texts simply *define* A and B to be independent with respect to P if $P(A \cap B) = P(A)P(B)$. This is in fact the formulation of independence that we shall most often use in the remainder of the course. This definition has the virtue of applying to empty as well as nonempty events. In particular, as you are asked to prove in the problems, it is the case that $\emptyset$ and A are independent for every event A.

Theorems 2.7, 2.9, and 2.11, enable us to deduce certain relevance relations from other relevance relations. The following theorems demonstrate that once the relevance relation between A and B is known, the relations between A and $\bar{B}$, between $\bar{A}$ and B, and between $\bar{A}$ and $\bar{B}$ follow automatically.

Theorem 2.12. If A and B are positively relevant to each other with respect to the probability measure P, then

1° A and $\bar{B}$ are negatively relevant to each other,

2° $\bar{A}$ and B are negatively relevant to each other, and

3° $\bar{A}$ and $\bar{B}$ are positively relevant to each other.

Proof. To prove 1°, it suffices (by the symmetry of the negative relevance relation) to prove that $P(\bar{B}|A) < P(\bar{B})$. By hypothesis, $P(B|A) > P(B)$. Hence $1 - P(B|A) < 1 - P(B)$, i.e., $P(\bar{B}|A) < P(\bar{B})$.

To prove 2°, we show by a similar argument that $P(\bar{A}|B) < P(\bar{A})$.

By symmetry of negative relevance, it follows from the preceding inequality that $P(B|\bar{A}) < P(B)$. Hence, $1 - P(B|\bar{A}) > 1 - P(B)$, i.e., $P(\bar{B}|\bar{A}) > P(\bar{B})$. By symmetry of positive relevance, this establishes 3°. □

Theorem 2.13. If A and B are negatively relevant to each other with respect to the probability measure P, then

1° A and $\bar{B}$ are positively relevant to each other,

2° $\bar{A}$ and B are positively relevant to each other, and

3° $\bar{A}$ and $\bar{B}$ are negatively relevant to each other.

Proof. The proof is similar to that of Theorem 2.12. □

Theorem 2.14. If A and B are independent with respect to the probability measure P, then

1° A and $\bar{B}$ are independent,

2° $\bar{A}$ and B are independent,

3° $\bar{A}$ and $\bar{B}$ are independent.

Proof. The proof is similar to that of Theorem 2.12. □

2.10 Simpson's Paradox

Consider the following admissions data for 1200 applicants to the graduate school of a certain university (F = females, M = males, A = acceptees, R = rejectees):

(2.50)

	A	R	
F	210	390	600
M	275	325	600
	485	715	1200

(university-wide data)

It appears from this data that females do not fare as well as males in gaining admission to this graduate school. For example, the overall acceptance rate $P(A) = 485/1200 \approx .40$, whereas the acceptance rate for females $P(A|F) = 210/600 = .35$, and the acceptance rate for males $P(A|M) = 275/600 \approx .46$. That is, the set of acceptees is underrepresented in the class of female applicants and overrepresented in the class of male applicants. By symmetry, it follows of course that $P(F|A) < P(F)$ and $P(M|A) > P(M)$, i.e., that females are underrepresented among acceptees and males are overrepresented among acceptees.

The administration, wishing to investigate this state of affairs, asks the admission office to furnish separate admissions data for the humanities and science divisions of the graduate school. Here is what that office reported for the humanities:

(2.51)

	A	R	
F	150	350	500
M	25	75	100
	175	425	600

(humanities data)

In the humanities, the overall acceptance rate $P_h(A) = 175/600 \approx .27$, the acceptance rate for females $P_h(A|F) = 150/500 = .30$ and the acceptance rate for males $P_h(A|M) = 25/100 = .25$. So females fare better than males with respect to admission to graduate programs in the humanities. So it must be, must it not, that females are faring worse than males in admissions to the science programs?

But wait. Here is what the office reported for the sciences:

(2.52)

	A	R	
F	60	40	100
M	250	250	500
	310	290	600

(sciences data)

In the sciences, the overall acceptance rate $P_s(A) = 310/600 \approx .52$, the acceptance rate for females $P_s(A|F) = 60/100 = .60$, and the acceptance rate for males $P_s(A|M) = 250/500 = .50$. So females fare better than males with respect to admission to graduate programs in the sciences as well!

The above is a striking example of a phenomenon known as *Simpson's Paradox*. (Incidentally, a paradox is a *seeming* contradiction, not an *actual* contradiction. Faced with a paradox, one's intellectual responsibility is to dispel the air of contradiction by giving a clear account of the state of affairs in question.) In general, we are said to have a case of Simpson's paradox whenever the merger of two or more 2×2 tables results in the reversal of the relevance relations obtaining in those tables.

In the admissions case, we have the positive relevance of F to A in tables (2.51) and (2.52) reversed to a case of negative relevance when these tables are merged into table (2.50). How can we account for this situation? The explanation is not hard to see. It is much tougher at this university to be admitted to a graduate program in the humanities division (acceptance rate = .27) than in the science division (acceptance rate = .52). And 5 out of every 6 females are applying to this division, generating a large number of female rejectees, so many, that for the merged data, females are underrepresented among acceptees.

By the way, a case similar to this actually occurred at the University of California at Berkeley [see Bickel, Hammel, and O'Connell, "Sex Bias in Graduate Admissions: Data from Berkeley," *Science* 187 (Feb. 1975), 398-404]. Based on the merged data, certain individuals accused the university of discrimination against women. But a departmental breakdown of the data showed that women were in every case proportionally represented or overrepresented among acceptees. They were just applying in greater numbers to departments with high rejection rates.

It is, incidentally, possible to have more than one reversal of relevance occur. In testing drug A against drug B as a cure for a certain disease, it is possible that 1° the cure rate of A is higher than that of B at every hospital in a given state, 2° when the data for individual hospitals are merged by county, the cure rate for B is higher than that of A in each county and 3° when the data are merged for the entire state, the cure rate for A is again higher than that of B. In fact, one can cook up data for which there are four, five, or, indeed, any finite number of relevance reversals.

2.11 Problems

1. A set S of 3000 professionally employed individuals is categorized by sex (F = the set of females in S; M = the set of males) and occupation type (T = the set of people with technical positions; $\bar{T}$ = the set of people with nontechnical positions). The results are shown in the following two-way frequency table.

	T	$\bar{T}$	
F	150	850	1000
M	600	1400	2000
	750	2250	3000

a. Find the associated two-way *relative* frequency table.

b. Give a clear verbal description of the quantities $P(T \cap F)$, $P(T|F)$, and $P(F|T)$ and calculate these quantities.

c. Compare $P(F|T)$ with $P(F)$ and $P(T|F)$ with $P(T)$ and give a clear verbalization of the results.

d. Somebody says "Males are more likely to go into technical professions than females in this sample, because there are 600 male technical professionals and only 150 female technical professionals." Explain why, although the assertion is correct, the reasoning is faulty.

2. Choose at random a sequential arrangement of three 4's and two 3's.

a. What is the probability that the 3's are adjacent?

b. What is the (conditional) probability that the 3's are adjacent, given that the number chosen is even?

c. Are the events "3's adjacent" and "even number chosen" independent?

3. Let S = set of all males ≥ 19 years old in the USA. Let L be the subset of those less than 6 feet tall and let B = the subset of those who are professional basketball players. Let P be the measure of relative frequency on S. Which of the following is true, and why.

1° $P(L|B) < P(L)$

2° $P(L|B) > P(L)$

3° $P(L|B) = P(L)$

Same question for

1_a° $P(B|L) < P(B)$

2_a° $P(B|L) > P(B)$

3_a° $P(B|L) = P(B)$

Don't say "No data have been given on the basis of which I can answer this." Think about what the world is like.

4. Given any probability measure P on S and any event $A \subseteq S$, prove that A and $\emptyset$ are independent and A and S are independent.

5. Generalize the above results by showing that for any probability measure P on S and any event $A \subseteq S$,

a. if $P(B) = 0$, then A and B are independent

b. if $P(B) = 1$, then A and B are independent. (Hint: Prove that $P(A \cap \bar{B}) = 0$.)

6. Construct data for five baseball seasons in which player X has a higher batting average than player Y in each of those five seasons, and yet player Y has the higher batting average when the data for the five seasons are merged. Can this be done under the constraint that X is at bat the same number of times as Y each season?

7. a. Suppose that events $A, B \subseteq S$ are mutually exclusive, i.e., disjoint. Prove that if $P(A) > 0$ and $P(B) > 0$, then A and B are negatively relevant to each other (hence, *not* independent).

   b. Prove that if A and B are mutually exclusive, and also independent, then $P(A) = 0$ or $P(B) = 0$.

8. A pair of fair dice, one red and one blue, are tossed. Let A = "red die comes up odd" and let B = "blue die comes up even." Use common sense to determine, without calculation, if A and B are independent. Then verify your answer with a calculation.

9. In a certain school, 10% of the students failed mathematics, 12% failed English, and 2% failed both mathematics and English. Are the students who failed mathematics a) overrepresented b) underrepresented or c) proportionally represented among students who failed English? Are the students who failed English a) overrepresented b) underrepresented or c) proportionally represented among students who failed mathematics?

10. Under the conditions of problem 8 above, let A = "red die comes up 4" and B = "sum of numbers showing on the two dice is 6." Decide the nature of the relevance relation between A and B.

11. If $P(E) = P(F) = .40$ and $P(\bar{E} \cap \bar{F}) = .24$, are E and F independent?

2.12 The Product Rule for Probabilities

Give that $P(A) > 0$, we have defined $P(B|A)$, the conditional probability of B, given A, by

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)}. \tag{2.53}$$

It follows from (2.53) that

$$P(A \cap B) = P(A)P(B|A). \tag{2.54}$$

Formula (2.54) holds for all events A and B and all probability measures P such that $P(A) > 0$. It is the simplest case of the *product rule for probabilities*. When A and B are independent, (2.54) reduces to the special case $P(A \cap B) = P(A)P(B)$ since in that case $P(B|A) = P(A)$.

One uses (2.54) to evaluate $P(A \cap B)$ when it is clear what $P(A)$ is and where one can assess $P(B|A)$ directly (i.e., without breaking it down as the ratio $P(B \cap A)/P(A)$, which would involve one in a vicious circle.) Here are some examples:

Example 1. An urn contains 3 red and 5 blue balls. A ball is chosen at random, without replacement, and another ball is chosen at random from the remaining balls. Find the probability that the first ball chosen is red, and the second is blue.

Solution: $P(\text{1st red} \cap \text{2nd blue}) = P(\text{1st red})\ P(\text{2nd blue}|\ \text{1st red}) = \left(\frac{3}{8}\right)\left(\frac{5}{7}\right) = \frac{15}{56}$.

Example 2. A fair die is tossed. If it comes up 1 or 6, a ball is drawn at random from urn I; otherwise a ball is drawn at random from urn II. Urn I contains 2 red balls, 1 white ball, and 2 blue balls. Urn II contains 3 white balls, 1 blue ball, and no red balls. What is the probability that a red ball is drawn? a white ball? a blue ball?

Solution. It is useful to fill out a 2×3 contingency table for this problem:

	R	W	B	
urn I	2/15	1/15	2/15	1/3
urn II	0	2/4	2/12	2/3
	2/15	17/30	18/60	1

One calculates $P(I \cap R)$ by the formula $P(I \cap R) = P(I)P(R|I) = (1/3)(2/5) = 2/15$. $P(E \cap W)$, $P(I \cap B)$, $P(II \cap R)$, $P(II \cap W)$, and $P(II \cap B)$ are computed similarly. Summing the columns yields $P(R)$, $P(W)$, and $P(B)$, as desired.

The following is a generalization of (2.54).

Theorem 2.15. Let $A_1, A_2, \ldots, A_n$ be events for which $P(A_1)$, $P(A_1 \cap A_2)$, …, and $P(A_1 \cap A_2 \cap \cdots \cap A_{n-1})$ are all positive. Then

$$(2.55)\quad P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

Proof. We prove (2.55) for all $n \geq 2$ by induction on n. When $n = 2$, (2.55) is just (2.54), which we have already established. Suppose (2.55) is true for $n = 2, \ldots, k-1$. Then

$$\begin{aligned} P(A_1 \cap \cdots \cap A_n) &= P((A_1 \cap \cdots \cap A_{k-1}) \cap A_k) \\ &= P(A_1 \cap \cdots \cap A_{k-1})P(A_k|A_1 \cap \cdots \cap A_{k-1}) \\ &= P(A_1)P(A_2|A_1) \cdots P(A_{k-1}|A_1 \cap \cdots \cap A_{k-2})P(A_k|A_1 \cap \cdots \cap A_{k-1}), \end{aligned}$$

which completes the induction. □

Example 3. From an urn containing 3 red and 5 blue balls, 3 balls are chosen at random, in succession and without replacement. What is the probability that all three are red?

Solution. $P(\text{1st red} \cap \text{2nd red} \cap \text{3rd red}) = P(\text{1st red})\ P(\text{2nd red}|\ \text{1st red})\ P(\text{3rd red}|\ \text{1st red} \cap \text{2nd red}) = \frac{3}{8} \cdot \frac{2}{7} \cdot \frac{1}{6} = \frac{1}{56}$.

2.13 Bayes' Rule

Let S be a sample space equipped with the probability measure P. If E and H are events (i.e., subsets) of S, the conditional probabilities $P(H|E)$ and $P(E|H)$ are, in general, unequal. There is, however, a simple formula connecting $P(H|E)$ and $P(E|H)$.

Theorem 2.16. (Bayes' rule - simple form)

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)}. \tag{2.56}$$

Proof. By definition, $P(H|E) = P(H \cap E)/P(E)$. By (2.54), $P(H \cap E) = P(H)P(E|H)$. Combining these results yields (2.56). □

The following extended form of (2.56) is often useful:

Theorem 2.17. (Bayes' rule - extended form)

$$P(H|E) = \frac{P(H)P(E|H)}{P(H)P(E|H) + P(\bar{H})P(E|\bar{H})}. \tag{2.57}$$

Proof. Since $H \cap E$ and $\bar{H} \cap E$ are disjoint and $(H \cap E) \cup (\bar{H} \cap E) = E$, we have $P(E) = P(H \cap E) + P(\bar{H} \cap E)$. Expanding $P(H \cap E)$ and $P(\bar{H} \cap E)$ by (2.54) yields

$$P(E) = P(H)P(E|H) + P(\bar{H})P(E|\bar{H}). \tag{2.58}$$

Substituting (2.58) in (2.56) for $P(E)$ yields (2.57). □

Example. (Screening test for a disease). As you know, a "positive" result on such a test is evidence (though usually far from decisive evidence, taken all by itself) that a person has the disease and a "negative" result is evidence (again, not decisive, taken all by itself) that the person does not have the disease. Suppose that the characteristics of a given test, established by medical research, are

$$P(\text{positive}|\text{healthy}) = \frac{1}{100}, \quad \text{and} \tag{2.59}$$

$$P(\text{negative}|\text{diseased}) = \frac{2}{100}. \tag{2.60}$$

This looks like (and is) a pretty good test. Rarely (1 out of 100 times) does it label a healthy person as diseased, and rarely (2 out of 100 times) does it give a diseased person a clean bill of health.

Someone is given this screening test and it comes out positive. What is the probability that the person in fact has the disease, i.e., what is $P(\text{diseased}|\text{positive})$? By (2.57) with $H = \text{diseased}$ and $E = \text{positive}$, we have

$$P(\text{diseased}|\text{positive}) \tag{2.61}$$

$$
\begin{aligned}
&= \frac{P(\text{diseased})P(\text{positive}|\text{diseased})}{P(\text{diseased})P(\text{positive}|\text{diseased}) + P(\text{healthy})P(\text{positive}|\text{healthy})} \\
&= \frac{P(\text{diseased})(.98)}{P(\text{diseased})(.98) + (1 - P(\text{diseased}))(.01)}.
\end{aligned}
$$

We see that to calculate $P(\text{diseased}|\text{positive})$ we need one more datum, namely $P(\text{diseased})$, the relative frequency (or "prevalence") of this disease in the population to which the individual belongs. This might differ from country to country or from subculture to subculture (it is a subtle issue to determine the appropriate "reference population" for an individual). The drug company that developed the test can furnish you with the numbers $P(\text{positive}|\text{healthy})$ and $P(\text{negative}|\text{diseased})$ based on broad experimentation. But the calculation of $P(\text{diseased}|\text{positive})$ is a "local matter" and requires *local* epidemiological data. As an example, suppose

1° $P(\text{diseased}) = \frac{1}{1000}$. Plugging in (2.61) we get $P(\text{diseased}|\text{positive}) = \frac{98}{1097} \approx .09$.

2° $P(\text{diseased}) = \frac{2}{1000}$. Then $P(\text{diseased}|\text{positive}) = \frac{196}{1194} \approx .16$.

3° $P(\text{diseased}) = \frac{10}{1000} = \frac{1}{100}$. $P(\text{diseased}|\text{positive}) = \frac{98}{197} \approx .50$.

Suppose that a person's test comes out negative. What is the probability $P(\text{healthy}|\text{negative})$ that the person is in fact free of the disease? Again by (2.57), with H = healthy and E = negative, we have

$$
\begin{aligned}
(2.62) \qquad & P(\text{healthy}|\text{negative}) \\
&= \frac{P(\text{healthy})P(\text{negative}|\text{healthy})}{P(\text{healthy})P(\text{negative}|\text{healthy}) + P(\text{diseased})P(\text{negative}|\text{diseased})} \\
&= \frac{(1 - P(\text{diseased})(.99)}{(1 - P(\text{diseased}))(.99) + P(\text{diseased})(.02)}.
\end{aligned}
$$

Again, we need to know $P(\text{diseased})$ to complete the computation. If, for example, $P(\text{diseased}) = \frac{1}{1000}$, then $P(\text{healthy}|\text{negative}) = \frac{98901}{98903} \approx .99998$.

Remark 1. In the medical profession, the following terminology is used

1° sensitivity of the test = $P(\text{positive}|\text{diseased})$.

2° specificity of the test = $P(\text{negative}|\text{healthy})$.

3° predictive value positive of the test (PV^+) = $P(\text{diseased}|\text{positive})$.

4° predictive value negative of the test (PV^-) = $P(\text{healthy}|\text{negative})$.

5° prevalence of disease = $P(\text{diseased})$ (note: while "prevalence" and "incidence" are synonyms in ordinary English, these terms have different meanings in medicine. The latter term is the probability of developing a *new* case of the disease over some period of time for an individual who does not have the disease at the beginning of that time interval).

Remark 2. Not making a distinction between the sensitivity of a test and its predictive value positive can have tragic consequences. There is at least one case on record, for example, of an individual who tested positive for HIV and committed suicide based on the high sensitivity (0.999) of the test, not realizing that the relevant conditional probability, PV^+, could be considerably smaller.

Remark 3. In the formulation of Bayes' rule, H is often termed the *hypothesis*, E the *evidence*, $P(H)$ the *prior probability* of H, and $P(H|E)$, the *posterior probability of H, given that E has been observed.*

Remark 4. If $P(H) = a$, the *odds in favor of H* are a to $1 - a$ (or ma to $m(1 - a)$ for any number m). For example, if $P(H) = 3/5$, then the odds in favor of H are 3/5 to 2/5 (or 3 to 2, or 6 to 4, or 6000 to 4000, etc., etc.). (If P is uniform, the odds in favor of H can be thought of as: # favorable cases to # unfavorable cases.). Going from odds to probability, if the odds in favor of H are r to s, the $P(H) = r/(r + s)$. The following is a formulation of Bayes' Rule in odds form:

Theorem 2.18. (Bayes' Rule - odds form)

$$\begin{array}{ccccc} \dfrac{P(H|E)}{P(\bar{H}|E)} & = & \dfrac{P(H)}{P(\bar{H})} & \times & \dfrac{P(E|H)}{P(E|\bar{H})} \\ \downarrow & & \downarrow & & \downarrow \\ \text{posterior} & & \text{prior} & & \text{likelihood ratio, or} \\ \text{odds} & & \text{odds} & & \text{Bayes factor.} \end{array} \tag{2.63}$$

Proof. Using the definition of conditional probability, show that each side of (2.63) is equal to $P(H \cap E)/P(\bar{H} \cap E)$. □

In the preceding discussion, hypothesis H is competing with the alternative $\bar{H}$. Given evidence E, we compute $P(H|E)$ by Bayes' rule. We could also compute $P(\bar{H}|E)$ by Bayes' rule, but it is simpler to use the relation $P(\bar{H}|E) = 1 - P(H|E)$, as established in (2.33). In many cases, however, there are more than two competing hypotheses. Suppose that $H_1, \ldots, H_n$ are competing hypotheses, where $H_i \cap H_j = \emptyset$ if $i \neq j$ and $H_1 \cup \cdots \cup H_n = S$, i.e., where the H_i's are pairwise disjoint and exhaustive. Given evidence E, we need to "update" each of the prior probabilities $P(H_i)$ to the posterior probability $P(H_i|E)$. The following theorem shows how to do this.

Theorem 2.19. (Bayes' rule - multiple hypotheses) For each $i = 1, 2, \ldots, n$,

$$P(H_i|E) = \frac{P(H_i)P(E|H_i)}{P(H_1)P(E|H_1) + \cdots + P(H_n)P(E|H_n)}. \tag{2.64}$$

Proof. By (2.56),

$$P(H_i|E) = P(H_i)P(E|H_i)/P(E). \tag{2.65}$$

But

$$E = (H_1 \cap E) \cup (H_2 \cap E) \cup \cdots \cup (H_n \cap E), \tag{2.66}$$

with $(H_i \cap E) \cap (H_j \cap E) = \emptyset$ if $i \neq j$. Hence

$$\begin{aligned} P(E) &= P(H_1 \cap E) + P(H_2 \cap E) + \cdots + P(H_n \cap E) \\ &= P(H_1)P(E|H_1) + P(H_2)P(E|H_2) + \cdots + P(H_n)P(E|H_n), \end{aligned} \tag{2.67}$$

by Theorem 2.3 and (2.54). Substituting this expression for $P(E)$ in (2.65) yields (2.64). □

Remark 1. Since $P(H_1|E) + \cdots + P(H_n|E) = 1$ [why?], we can save a little work by using (2.64) to calculate $P(H_1|E), \ldots, P(H_{n-1}|E)$ and then computing $P(H_n|E)$ by the relation $P(H_n|E) = 1 - (P(H_1|E) + \cdots + P(H_n|E))$. On the other hand, doing this deprives us of the possibility of using the relation $P(H_1|E) + \cdots + P(H_n|E) = 1$ as a partial check on our computations.

Remark 2. When all of the hypotheses are equally likely prior to the discovery of E (i.e., when $P(H_1) = \cdots = P(H_n) = \frac{1}{n}$), formula (2.64) reduces to the strikingly simple form

$$P(H_i|E) = \frac{P(E|H_i)}{P(E|H_1) + \cdots + P(E|H_n)}. \tag{2.68}$$

Example. In 60-year-old never smoking males the probability of normal lungs (H_1) is .99, the probability of lung cancer (H_2) is .001, and the probability of sarcoidosis (H_3), a fairly common nonfatal lung disease, is .009. Such a male, complaining of a chronic cough, is biopsied. Let $E = \{$chronic cough, results of the biopsy$\}$ and suppose that $P(E|H_1) = .001$, $P(E|H_2) = .9$ and $P(E|H_3) = .9$. Find $P(H_1|E)$, $P(H_2|E)$, and $P(H_3|E)$.

Solution.

$$\begin{aligned} P(H_1|E) &= \frac{P(H_1)P(E|H_1)}{P(H_1)P(E|H_1) + P(H_2)P(E|H_2) + P(H_3)P(E|H_3)} \\ &= \frac{(.99)(.001)}{(.99)(.001) + (.001)(.9) + (.009)(.9)} \\ &= .099. \end{aligned}$$

Similarly, $P(H_2|E) = .090$ and $P(H_3|E) = .811$. So sarcoidosis is the most probable hypothesis.

2.14 Independence for Three or More Events

Let S be a sample space equipped with a probability measure P. Given three events A, B, and C in S, how should one define the *independence of* A, B, *and* C? Perhaps surprisingly, we require, in addition to the condition

$$P(A \cap B \cap C) = P(A)P(B)P(C), \tag{2.69}$$

the conditions

$$P(A \cap B) = P(A)P(B), \quad P(A \cap C) = P(A)P(C), \text{ and} \\ P(B \cap C) = P(B)P(C) \tag{2.70}$$

as well. The reason for this is as follows. Recall (Theorem 2.14) that the independence of A and B with respect to P (i.e., $P(A \cap B) = P(A)P(B)$) implies $P(A \cap \bar{B}) = P(A)P(\bar{B})$, $P(\bar{A} \cap B) = P(\bar{A})P(B)$ and $P(\bar{A} \cap B) = P(\bar{A})P(\bar{B})$. We want to define the independence of A, B, and C in such a way that, in particular, such independence implies $P(E_1 \cap E_2 \cap E_3) = P(E_1)P(E_2)P(E_3)$ where $E_1 = A$ or $\bar{A}$, $E_2 = B$ or $\bar{B}$ and $E_3 = C$ or $\bar{C}$. It turns out that (2.69) and (2.70) are necessary and sufficient to guarantee this result. We shall not prove this result in full, since the proof is rather tedious. But here is a proof that $P(A \cap B \cap \bar{C}) = P(A)P(B)P(\bar{C})$ based on (2.69) and (2.70): Clearly,

$$P(A \cap B) = P(A \cap B \cap C) + P(A \cap B \cap \bar{C}), \tag{2.71}$$

and so by (2.69) and (2.70),

$$P(A)P(B) = P(A)P(B)P(C) + P(A \cap B \cap \bar{C}), \tag{2.72}$$

and so

$$\begin{aligned} P(A \cap B \cap \bar{C}) &= P(A)P(B)[1 - P(C)] \\ &= P(A)P(B)P(\bar{C}). \end{aligned} \tag{2.73}$$

In general, we say that events $A_1, A_2, \ldots, A_n$ are independent if, for each $k = 2, \ldots, n$ and every sequence $1 \le i_1 < i_2 < \cdots < i_k \le n$,

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}). \tag{2.74}$$

Note that checking independence of $A_1, \ldots, A_n$ involves verifying $2^n - n - 1$ conditions [why?].

One can prove that if $A_1, \ldots, A_n$ are independent, then

$$P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_1)P(E_2) \cdots P(E_n), \tag{2.75}$$

where, for each $i = 1, \ldots, n$, $E_i = A_i$ or $\bar{A}_i$, and, more generally, that for any sequence $1 \le i_1 < i_2 < \cdots < i_k \le n$, $P(E_{i_1} \cap \cdots \cap E_{i_k}) = P(E_{i_1}) \cdots P(E_{i_k})$, where, for $j = 1, \ldots, k$, $E_{i_j} = A_{i_j}$ or $\bar{A}_{i_j}$.

2.15 Problems

1. From twelve tickets numbered 1 through 12, two tickets are drawn at random, one after the other, without replacement. What is the probability that one of the numbers drawn is even and the other is odd?

2. In a certain factory, machine A produces 30% of the output, machine B 25%, and machine C the rest. One percent of the output of machine A is defective, as is 1.2% of B's output, and 2% of C's. What percent of the factory's total daily output is defective?

3. A fair die is tossed. If it comes up "1", we pick at random a ball from Urn I, which contains 4 white balls and 1 blue ball. If it comes up "2", "3", "4", "5", or "6", we pick a ball at random from Urn II, which contains 5 white balls and 5 blue balls. If the result of carrying out this experiment is that a white ball is chosen, what is the probability that it came from Urn I?

4. A fair coin, a two headed coin, and a coin weighted so that heads comes up with probability $\frac{1}{4}$ are in a box. We select a coin from this box at random, flip it, and it comes up heads. What is the probability that a) it is the fair coin b) it is 2-headed c) it is weighted?

5. A fair coin is tossed twice. Let E_1 = heads on 1st toss, E_2 = heads on 2nd toss, and E_3 = results of both tosses the same. Show that, while E_1, E_2, and E_3 are pairwise independent, they are not independent.

6. The sensitivity and specificity of the ELISA test for HIV are both equal to 0.999. If 1 out of every 1000 individuals is infected with HIV, determine PV^+ and PV^-.

7. In the July 11, 1996 issue of the *New York Review of Books*, in an article entitled "Goodbye to Affirmative Action?", Andrew Hacker reports that, among freshman entering UCLA in 1994, Asians constituted 42.2% of the class, Whites 30.7%, Hispanics 20.0%, and Blacks 7.1%. On the other hand, Asians constituted 51.1% of those admitted solely on academic criteria, while Whites constituted 42.7% of that group, Hispanics 5% and Blacks 1.2%. Of the entire entering class, 67% of the students were admitted solely on academic criteria. Hacker claims that Asian admissions are based more on academic merit than White admissions. Evaluate Hacker's claim, including as part of your analysis a discussion of which conditional probabilities are relevant to deciding the extent to which admissions of particular groups are based solely on academic merit.

CHAPTER 3

DISCRETE RANDOM VARIABLES

3.1 Random Variables

Let S be a sample space equipped with some probability measure. A *random variable on* S is a function $X : S \to \mathbb{R}$. The random variable X may be thought of as assigning to each outcome $s \in S$ the "numerical label" $X(s)$. Typically, this label records some interesting numerical property of the outcome s.

Example 1. $S = \{(r, b) : r, b \in [6]\}$ consists of the possible outcomes of tossing a red and a blue die. There are many different random variables that can be defined on S, e.g., $X(r, b) = r + b$, $X(r, b) = \max\{r, b\}$, $X(r, b) = \min\{r, b\}$, or even $X(r, b) = \sqrt{r^2 + b^2}$.

Example 2. $S = \{hh, ht, th, tt\}$ consists of the possible outcomes of tossing a coin twice. Define $X(hh) = 2$, $X(ht) = X(th) = 1$, and $X(tt) = 0$. Here X records the number of heads appearing in a given outcome. One could, alternatively, record the number of tails, or the number of heads minus the number of tails, etc., etc.

Example 3. $S = \{1, 2, 3, 4, 5, 6\}$, the set of possible outcomes of tossing a die one time, with $X(s) = s$ for all $s \in S$. When the possible outcomes of an experiment are individual real numbers, as in this case, one often simply takes for X the identity function on S, as we did here. Of course one need not do this. Perhaps after tossing the die in question we plan to construct a cube of side s. The random variable $X(s) = s^3$, recording the volume of such a randomly generated cube, might very well be of interest in such a case.

A random variable is *discrete* if its range is finite or countably infinite, i.e., if it takes on finitely many possible values $x_1, \ldots, x_n$ for some $n \in \mathbb{P}$, or an infinite sequence of possible values $x_1, x_2, \ldots$. Suppose that $X : S \to \mathbb{R}$ is a discrete random variable on S, where S is equipped with the probability measure P. For each possible value x_i of X, define

$$f_X(x_i) = P(X = x_i) = P(\{s \in S : X(s) = x_i\}). \tag{3.1}$$

The function f_X, denoted simply by f is no confusion arises thereby, is called the *probability density function* (abbreviated pdf) *of* X. As with probabilities, $f_X(x_i)$ may record either the observed or the predicted relative frequency with which X takes the value x_i. The pdf f_X of a discrete random variable X taking on the finitely many possible values $x_1, \ldots, x_n$ has the property that $0 \leq f_X(x_i) \leq 1$ for all i, and

$$\sum_{i=1}^{n} f_X(x_i) = 1, \tag{3.2}$$

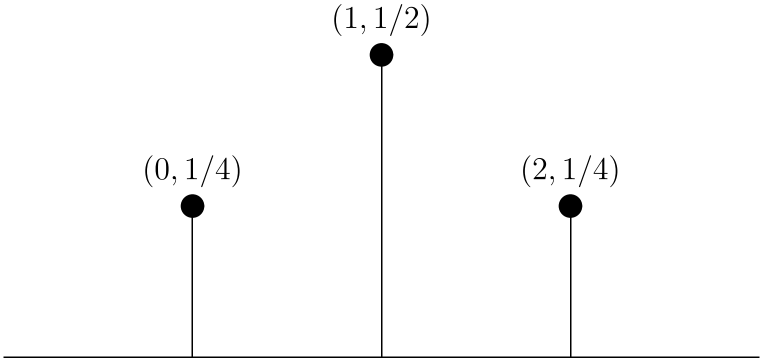
a result which follows from the additivity of P. If X takes on the infinite sequence of possible values $x_1, x_2, \ldots$, then, again, $0 \leq f_X(x_i) \leq 1$ for all i, and

$$\sum_{i=1}^{\infty} f_X(x_i) = 1, \tag{3.3}$$

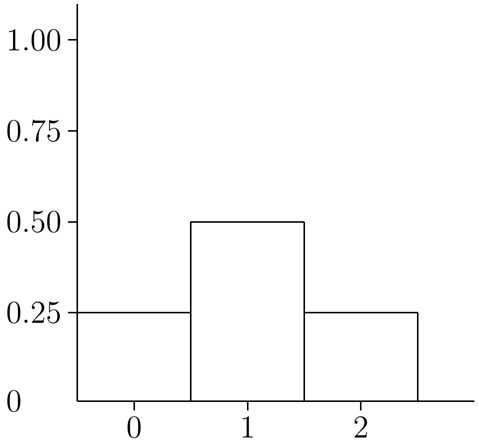
which follows from the countable additivity of P.

Example 4. Suppose that in Example 2 above each outcome has probability $\frac{1}{4}$. Then the possible values of X are $x_1 = 0$, $x_2 = 1$, and $x_3 = 2$ and $f_X(0) = P(\{tt\}) = 1/4$, $f_X(1) = P(\{ht, th\}) = 1/2$, and $f_X(2) = P(\{hh\}) = 1/4$.

The pdf of a discrete random variable (sometimes called a "discrete density function" for short) can be represented by a *discrete graph* or by a *histogram.* The discrete graph of the pdf of Example 4 is



The histogram of this pdf is

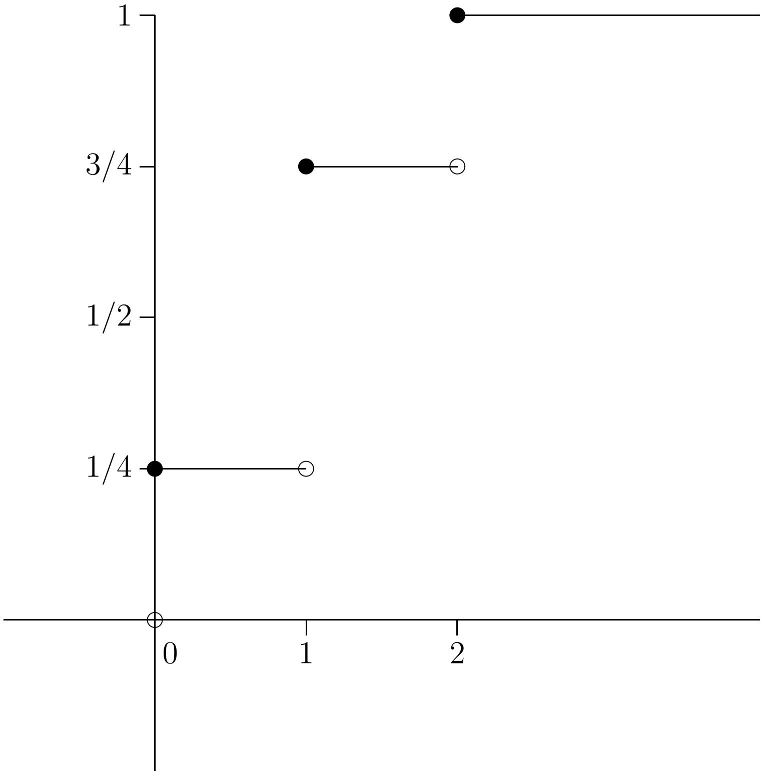


where, for $k = 0, 1, 2$, $f_X(k)$ is represented by the area of a block of base 1, centered at k, and height $f_X(k)$. (If X takes on non-integral values, constructing a histogram for its pdf is more complicated. The only discrete random variables that we will consider take on exclusively integral values.)

Associated with any random variable X (discrete or not) is a function F_X, called the *cumulative distribution function* (abbreviated cdf) *of* X, and defined for all $x \in \mathbb{R}$ by

$$F_X(x) = P(X \le x) = P(\{s \in S : X(s) \le x\}). \tag{3.4}$$

If no confusion arises thereby, the function F_X is simple written as F. The cdf of a discrete random variable is always a step function. In the case of Example 4 above, the graph of F_X is given below:



We have $F_X(x) = 0$ if $x < 0$, $F_X(x) = 1/4$ if $0 \le x < 1$, $F_X(x) = 3/4$ if $1 \le x < 2$, and $F_X(x) = 1$ if $x \ge 2$.

3.2 Binomial and Geometric Random Variables

Suppose that we have a coin weighted so that it comes up heads with probability p and (hence) tails with probability $1 - p$. The sample space S of possible outcomes of n tosses of this coin consists of the set of 2^n words in the alphabet $\{h, t\}$. For example, if $n = 2$, $S = \{hh, ht, th, tt\}$.

Define the random variable X on S by X (each word) = the number of h's in that word. In other words, X labels each outcome in S with the number of heads that occurred as part of that outcome. The possible values that X can take on are clearly $0, 1, \ldots, n$, so X is a discrete random variable. Let us determine its pdf. Since the results of different tosses of the coin have no causal relation to each other (or, as one also says, these results are "physically independent"), it is reasonable in constructing a probability measure on S to assume that the events "heads on toss 1," "heads on toss 2," ..., and "heads on toss n" are independent in the sense described in §2.14. In particular, this means that to get the probability of any particular sequence of results, one simply multiplies the probabilities of those results. So, for example, P(all heads) $= p^n$, P(heads on tosses 1 and 2 and tails on other tosses) $= p^2(1-p)^{n-2}$, but also P(heads on toss 1 and tails on toss 3) $= p(1-p)$ etc., etc.

Now that we have equipped S with a probability measure P, we can determine the pdf of X. For $k = 0, 1, \ldots, n$

$$f_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{3.5}$$

The reason for (3.5) is clear. By our independence assumption, every outcome consisting of k heads and (hence) $n-k$ tails has probability $p^k(1-p)^{n-k}$. But there are $\binom{n}{k}$ such outcomes since by Theorem 1.15, there are $\binom{n}{k}$ words of length n comprised of k h's and $n-k$ t's. By the binomial theorem

$$\sum_{k=0}^{n} f_X(k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1, \tag{3.6}$$

so f_X, as defined by (3.5), is indeed a discrete density function.

The above example represents a prototypical case of a *binomial experiment.* In such an experiment there are n *independent trials*, with the results of each trial labeled "success" or "failure." On each trial the probability of success is p and (hence) the probability of failure is $1-p$. The sample space S consists of the set of 2^n words in the alphabet $\{s, f\}$ and the random variable X on S is defined by X (each word) = the number of s's in that word. The pdf f_X is again given by (3.5). We say that such a random variable X is a *binomial random variable with parameters n and p*, abbreviating this with the notation $X \sim$ binomial (n, p). We also say that X has a *binomial distribution with parameters n and p*, and that f_X is a *binomial density function with parameters n and p.*

Example 1. A pair of fair dice are tossed 10 times. What is the probability of getting "snake eyes" (i.e., two 1's) exactly 3 times? at least once?

Solution. Let X record the number of times (out of 10) that snake eyes (success) occurs. Clearly, $X \sim$ binomial $(10, \frac{1}{36})$. So for $k = 0, \ldots, 10$, $P(X = k) = \binom{10}{k} \left(\frac{1}{36}\right)^k \left(\frac{35}{36}\right)^{10-k}$. Setting $k = 3$ answers the first question above. The answer to the second question is $P(X \geq 1) = \sum_{k=1}^{10} P(X = k)$, with $P(X = k)$ given as above, but this answer is excessively complicated. More simply, $P(X \geq 1) = 1 - P(X = 0) = 1 - \left(\frac{35}{36}\right)^{10}$. In general, if $X \sim$ binomial (n, p), then $P(X \geq 1) = 1 - (1-p)^n$.

The binomial distribution arises whenever we have a finite number of physically independent trials with the result of each trial being one of two types, and the probability of a result of each type remaining constant over all the trials. A common example of this is the case of *random sampling with replacement from a finite population of two types of things.* Suppose the population has N members, A of type 1 and $N - A$ of type 2, and we select at random, and with replacement after each selection, n things from the population. Let X record the number of times (out of n) that we select a thing of type 1. Then clearly, $X \sim$ binomial $(n, A/N)$.

Example 2. An urn contains 10 red and 90 white balls. How many times must one select a ball at random, with replacement, from this urn in order to guarantee that the odds in favor of selecting at least one red ball are 100 to 1 or better?

Solution. With $X \sim$ binomial $(n, 1/10)$, $P(X \geq 1) = 1 - (.9)^n \geq \frac{100}{101}$ whenever $n \geq 45$. (You can solve this inequality using logarithms, or just experiment with your calculator until you find the smallest n that works.)

Example 3. Suppose that, in flight, airplane engines fail with probability q, independently from engine to engine, and that a plane lands safely if at least half of its engines run. For what values of q is a two-engine plane preferable to a four-engine plane?

Solution. Let X be the number of engines that do not fail. For the two-engine plane, $X \sim$ binomial $(2, 1-q)$ and $P(\text{lands safely}) =$
$P(X \geq 1) = 1 - P(X = 0) = 1 - q^2$. For the four-engine plane, $X \sim$
binomial $(4, 1-q)$ and $P(\text{lands safely}) = P(X \geq 2) = 1 - P(X = 0) - P(X = 1) =$
$1 - q^4 - 4(1-q)q^3 = 1 - 4q^3 + 3q^4$. The two-engine plane is preferable to the four-engine plane if and only if

$$\begin{aligned} 1 - q^2 > 1 - 4q^3 + 3q^4 &\Leftrightarrow q^2(1 - 4q + 3q^2) < 0 \\ &\Leftrightarrow q^2(1-q)(1-3q) < 0 \\ &\Leftrightarrow 1 - 3q < 0 \\ &\Leftrightarrow q > \frac{1}{3}, \end{aligned}$$

since $0 < q < 1$.

The Geometric Distribution. Suppose that we perform a sequence of independent trials, with the probability of success on each trial being p. Unlike the case of a binomial experiment in which the number of trials is fixed, in this case we shall continue to perform the trials until the first success occurs. So the sample space $S = \{s, fs, ffs, fffs, \ldots\}$. Let X record the number of the trial on which the first success occurs. The possible values of X are $1, 2, 3, \ldots$.

Let us determine the pdf f_X of X. Clearly, $f_X(1) = P(X = 1) = P(\text{success on trial 1}) = p$, $f_X(2) = P(X = 2) = P(\text{failure on trial 1 and success on trial 2}) = (1 - p)p$, and in general, for all $k \in \mathbb{P}$,

$$\begin{aligned} f_X(k) &= P(X = k) \\ &= P(\text{failure on 1st } k-1 \text{ trials and success on trial } k) \\ &= (1-p)^{k-1}p. \end{aligned} \tag{3.7}$$

Note that

$$\begin{aligned} \sum_{k=1}^{\infty} f_X(k) &= \sum_{k=1}^{\infty}(1-p)^{k-1}p = p\sum_{j=0}^{\infty}(1-p)^j \\ &= p \times \frac{1}{1-(1-p)} = 1, \end{aligned} \tag{3.8}$$

so f_X, as defined by (3.7), is indeed a discrete density function. This is our first example of a discrete random variable that takes on a countably infinite number of possible values. A random variable X with pdf given by (3.7) is called a *geometric random variable with parameter* p, abbreviated $X \sim$ geometric (p). We also say that X has a *geometric distribution with parameter* p.

Example 4. If we toss a pair of dice, what is the probability that the first "7" appears on the 5th toss? before the 5th toss? after the 10th toss?

Solution. The probability p of getting a "7" on any toss is $\frac{6}{36} = \frac{1}{6}$. So we are asking, with $X \sim$ geometric $(1/6)$, for $P(X = 5)$, $P(X \leq 4)$, and $P(X \geq 11)$. For each $k \in \mathbb{P}$, $P(X = k) = (5/6)^{k-1}(1/6)$. Setting $k = 5$ answers the first question. Clearly, $P(X \leq 4) = \frac{1}{6}\sum_{k=1}^{4}\left(\frac{5}{6}\right)^{k-1} = 1 - \left(\frac{5}{6}\right)^4$, by the formula for the sum of a finite geometric series. Finally,

$$\begin{aligned} P(X \geq 11) &= \frac{1}{6}\sum_{k=11}^{\infty}\left(\frac{5}{6}\right)^{k-1} \\ &= \left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^{10}\sum_{j=0}^{\infty}\left(\frac{5}{6}\right)^j \\ &= \left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^{10}\frac{1}{1-\frac{5}{6}} \\ &= \left(\frac{5}{6}\right)^{10}. \end{aligned}$$

A simpler derivation of this result follows from the fact that $P(X \geq 11) = P(\text{non-7 on trials } 1, \ldots, 10) = \left(\frac{5}{6}\right)^{10}$. Similarly, above, $P(X \leq 4) = 1 - P(X > 4) = 1 - \left(\frac{5}{6}\right)^4$.

3.3 Hypergeometric Random Variables

Just as binomial random variables arise in connection with random sampling with replacement from a finite population of two types of things, *hypergeometric random variables* arise in connection with random sampling *without replacement* from a finite population of two types of things. Suppose the population has N members, A of type 1 and $N-A$ of type 2, and we select at random, and without replacement, n things from that population. Let X record the number of times (out of n) that we select a thing of type 1. We say in such a case that X is *hypergeometric with parameters* n, A, *and* N, abbreviated $X \sim$ hypergeometric (n, A, N). Let us determine the pdf of X. We first derive this result by selecting the n objects one at a time.

For $k = 0, 1, \ldots, n$,

$$
\begin{aligned}
(3.9) \qquad f_X(k) &= P(X = k) \\
&= \frac{\#\text{ of permutations of } N \text{ things taken } n \text{ at a time, } k \text{ of type 1}}{\#\text{ of permutations of } N \text{ things taken } n \text{ at a time}} \\
&= \underset{(1)}{\binom{A}{k}} \underset{(2)}{\binom{N-A}{n-k}} \underset{(3)}{n!} / N^{\underline{n}} = \binom{A}{k}\binom{N-A}{n-k} \Big/ \binom{N}{n}
\end{aligned}
$$

(1) choose k of the A things of type 1
(2) choose $n-k$ of the $N-A$ things of type 2
(3) arrange the n chosen things in a permutation

The final form of this pdf can be derived by selecting the sample of n objects simultaneously (i.e., by viewing the sample as a set rather than as a sequence). There are $\binom{N}{n}$ ways to select n of N things and $\binom{A}{k}\binom{N-A}{n-k}$ of these involve choosing k things of type 1.

Note that

$$
\begin{aligned}
\sum_{k=0}^{n} f_X(k) &= \sum_{k=0}^{n} \binom{A}{k}\binom{N-A}{n-k} \Big/ \binom{N}{n} \\
(3.10) \qquad &= \binom{N}{n} \Big/ \binom{N}{n} = 1,
\end{aligned}
$$

by Vandermonde's Theorem (Theorem 1.14), so f_X, as given by (3.9), is indeed a discrete density function.

Remark. Typically the sample size n is much smaller than A and $N-A$, so that $f_X(k) > 0$ for every $k = 0, \ldots, n$. But our formula works for every $k = 0, \ldots, n$ no matter how large

n is (of course, we must have $n \le N$). For example, suppose $n = N$, i.e., suppose we take as our sample the entire population of N objects. Then formula (3.9) yields $P(X = k) = 0$ for $k \ne A$ [why?] and $P(X = A) = 1$, which makes sense, since we are certain to get all A objects of type 1.

Example. There are 30 pennies and 70 nickels in a bowl. Suppose that 15 of these coins are selected at random, without replacement. What is the probability that exactly 10 pennies are selected? What is the probability that at least one penny is chosen?

Solution. If X records the number of pennies selected, then
$X \sim$ hypergeometric (15, 30, 100). So for $k = 0, \ldots, 15$

$$P(X = k) = \binom{30}{k}\binom{70}{15-k}\Big/\binom{100}{15}.$$

The answer to the first question is gotten by setting $k = 10$ in this formula. The answer to the second question is, of course

$$P(X \ge 1) = \sum_{k=1}^{15} \binom{30}{k}\binom{70}{15-k}\Big/\binom{100}{15},$$

but this is needlessly complicated. More simply,

$$P(X \ge 1) = 1 - P(X = 0) = 1 - \binom{70}{15}\Big/\binom{100}{15}.$$

Remark. When n is small relative to A and $N - A$, which is typically the case, for example, in opinion polling, the pdfs of the hypergeometric (n, A, N) random variable and the binomial $\left(n, \frac{A}{N}\right)$ random variable are very close, i.e., for $k = 0, \ldots, n$

$$\frac{\binom{A}{k}\binom{N-A}{n-k}}{\binom{N}{n}} \approx \binom{n}{k}\left(\frac{A}{N}\right)^k \left(\frac{N-A}{N}\right)^{n-k}.$$

Since the binomial distribution is easier to work with, it is often used as an approximation to the hypergeometric distribution in such cases. As we shall see, we will make extensive use of this approximation when we study statistical methods of estimation based on random sampling, and also when we study hypothesis testing.

3.4 Elementary Hypothesis Testing

Example 1. A psychic claims the ability to predict the results of a sequence of coin flips carried out in another location. In a test of this ability, he correctly predicts the results of 7 out of 10 flips. To evaluate this claim, we consider the so-called “null hypothesis” that the individual is just guessing and calculate the probability that a person would get 7 or more correct predictions under that hypothesis. The smaller that probability is, the less plausible the null hypothesis is, i.e., the more likely it is that he is not just guessing (that doesn't mean that he necessarily has psychic powers - he may be cheating).

How do we calculate the above probability? Our null hypothesis says that the number X of correct predictions is such that $X \sim$ binomial $(10, 1/2)$. We just need to calculate $P(X \geq 7)$ for such a random variable. We can do this by using the pdf of X or by consulting a table of binomial probabilities. In any case, we get $P(X \geq 7) = .172$. What this tells us is that, just by guessing, a person would make 7 or more correct predictions out of 10 more than 17% of the time. So this performance would surely not prompt us to reject the hypothesis that the individual is just guessing.

Suppose, on the other hand, that he correctly predicts the results of 14 out of 20 flips. You might be inclined to think that this performance is no more impressive than getting 7 out of 10 correct (his "batting average" in each case is .700, after all). But let's not prejudge the situation. Let's go ahead and determine $P(X \geq 14)$ under the hypothesis that $X \sim$ binomial $(20, \frac{1}{2})$. It turns out in this case that $P(X \geq 14) = .058$, i.e., one would make 14 or more correct predictions out of 20 less than 6% of the time just by guessing. So this performance casts considerably more doubt on the hypothesis that he is just guessing.

Deciding how small the probability of the observed performance must be in order to "reject the null hypothesis" is partly a philosophical matter. Many people reject the null hypothesis when the probability of the observed outcome (or an outcome more extreme) under that hypothesis is less than 5%. This is called "testing the hypothesis at the 5% significance level." Others are more conservative and use a significance level of 1%.

Example 2. From a group of 10 men and 10 women, 4 individuals are selected, supposedly at random, to receive an all-expense paid trip to a conference. The men suspect that in fact the selection process is biased in favor of women. When the selection is made, all four individuals chosen are women. The men complain that this result casts doubt on the claim (hypothesis) of random selection. Do they have a point? Under the hypothesis, the number X of men selected satisfies $X \sim$ hypergeometric $(4, 10, 20)$. For such a random variable X,

$$P(X = 0) = \frac{\binom{10}{0}\binom{10}{4}}{\binom{20}{4}} = .043$$

We would reject the claim of random selection at the 5% significance level, though not at the 1% significance level.

Example 3. In Lake Wobegone, 30% of the residents are Protestant and 70% are Roman Catholic. A supposedly random selection of 24 individuals for jury duty is carried out, but the Protestants suspect that the selection process is in fact biased in favor of Catholics. When the selection is carried out, just 4 Protestants are chosen. Do the Protestants have a basis for doubting the randomness of the selection?

Here, strictly speaking, a hypergeometric model applies, since they are obviously choosing without replacement. But we don't know the precise *number* of Protestants and Catholics, just their proportions. In any case, since the sample size is small relative to the number of Protestants and to the number of Catholics, we'll use the binomial approximation to the hypergeometric here.

Letting X denote the number of Protestants chosen, the hypothesis of random selection specifies that $X \sim \text{binomial}(24, .3)$. Given such an X, $P(X \leq 4) = .112$. While this result is somewhat improbable under the hypothesis of random selection, it would not be sufficient to reject the hypothesis at the 5% significance level. On the other hand, the selection of just 3 Protestants would lead to the rejection of the hypothesis of random selection at the 5% level. For $P(X \leq 3) = .043$. These probabilities were calculated using the table of binomial probabilities following Remark 2 below. You should use this table in working the relevant problems in §3.6.

Remark 1. In the examples above, the parameter n (sample size) was kept relatively modest in size so that we could calculate the relevant probabilities quickly by hand, or look them up in the table of binomial probabilities. For larger n, one could program a sophisticated calculator or computer to calculate the relevant probabilities. Alternatively, one can use what is called the "normal approximation to the binomial distribution" to calculate these probabilities. We shall discuss, and make extensive use of, that approximation in Chapter 4. Still another possibility, when p is "small" and n is "large," is to use what is called the "Poisson approximation to the binomial distribution." We shall discuss the latter approximation in §3.7.

Remark 2. In the above examples, the hypotheses in question were all tested against "one-sided" alternatives. In example 1, the hypothesis that the psychic is just guessing (i.e. that he has probability $p = 1/2$ of correctly identifying the result of any particular flip) is clearly being tested against the alternative $p > 1/2$. Sufficiently large values of X, the number of correctly identified flips, tend to cast doubt on the hypothesis and support the alternative. We decide whether the observed value of X is sufficiently large by calculating, under the hypothesis $p = 1/2$, the probability that X takes a value as large or larger than the value observed in the test. The smaller that probability is, the more doubt is cast on the hypothesis [why?]. In example 2 (respectively 3) the alternative is that the selection process is biased toward women(respectively, Catholics). In each case, small values of X tend to support those alternatives.

In some cases, the alternative to a given hypothesis is "two-sided," in the sense that both large and small values of the relevant random variable tend to support the alternative. Suppose, for example, that we are testing the hypothesis that a coin is fair ($P(\text{heads}) = p = 1/2$) against the two-sided alternative that it is biased in an unspecified direction ($p \neq 1/2$. The coin is flipped 24 times and comes up heads 7 times. In this case, we conclude for $X \sim \text{binomial } (24, 1/2)$ the probability that X takes a value "at least as far away" from its expected value 12 (see §3.5 below) as 7 is, i.e. we calculate $P(|X - 12| \geq 5) = P(X \leq 7) + P(X \geq 17) = .033 + .033 = .066$. Here we would not reject the hypothesis $p = 1/2$ at the 5% significance level. If, on the other hand, we were testing $p = 1/2$ against the one-sided alternative $p < 1/2$, we would reject the hypothesis at this level. This illustrates the general phenomenon that one is less likely to reject a hypothesis when it is tested against a two-sided alternative than when it is tested against a one-sided alternative.

Note that one formulates the alternative hypothesis *prior* to observing the value of the test random variable X. This standard statistical practice ensures that one does not cook up *ad hoc* alternatives after seeing the data.

Table of Density Functions for $X \sim \text{binomial}(24, p)$. The entry in row k and column p is $P(X = k)$.

p k	.01	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95	.99
0	.786	.292	.080	.005	0+	0+	0+	0+	0+	0+	0+	0+	0+
1	.190	.369	.213	.028	.002	0+	0+	0+	0+	0+	0+	0+	0+
2	.022	.223	.272	.081	.010	.001	0+	0+	0+	0+	0+	0+	0+
3	.002	.086	.221	.149	.031	.003	0+	0+	0+	0+	0+	0+	0+
4	0+	.024	.129	.196	.069	.010	.001	0+	0+	0+	0+	0+	0+
5	0+	.005	.057	.196	.118	.027	.003	0+	0+	0+	0+	0+	0+
6	0+	.001	.020	.155	.160	.056	.008	0+	0+	0+	0+	0+	0+
7	0+	0+	.006	.100	.176	.096	.021	.002	0+	0+	0+	0+	0+
8	0+	0+	.001	.053	.160	.136	.044	.005	0+	0+	0+	0+	0+
9	0+	0+	0+	.024	.122	.161	.078	.014	.001	0+	0+	0+	0+
10	0+	0+	0+	.009	.079	.161	.117	.032	.003	0+	0+	0+	0+
11	0+	0+	0+	.003	.043	.137	.149	.061	.008	0+	0+	0+	0+
12	0+	0+	0+	.001	.020	.099	.161	.099	.020	.001	0+	0+	0+
13	0+	0+	0+	0+	.008	.061	.149	.137	.043	.003	0+	0+	0+
14	0+	0+	0+	0+	.003	.032	.117	.161	.079	.009	0+	0+	0+
15	0+	0+	0+	0+	.001	.014	.078	.161	.122	.024	0+	0+	0+
16	0+	0+	0+	0+	0+	.005	.044	.136	.160	.053	.001	0+	0+
17	0+	0+	0+	0+	0+	.002	.021	.096	.176	.100	.006	0+	0+
18	0+	0+	0+	0+	0+	0+	.008	.056	.160	.155	.020	.001	0+
19	0+	0+	0+	0+	0+	0+	.003	.027	.118	.196	.057	.005	0+
20	0+	0+	0+	0+	0+	0+	.001	.010	.069	.196	.129	.024	0+
21	0+	0+	0+	0+	0+	0+	0+	.003	.031	.149	.221	.086	.002
22	0+	0+	0+	0+	0+	0+	0+	.001	.010	.081	.272	.223	.022
23	0+	0+	0+	0+	0+	0+	0+	0+	.002	.028	.213	.369	.190
24	0+	0+	0+	0+	0+	0+	0+	0+	0+	.005	.080	.292	.786

3.5 The Expected Value of a Discrete Random Variable

Suppose that X takes on the possible values $x_1, \ldots, x_n$ and that f_X is the pdf of X. The *expected value of* X, denoted $E(X)$, is defined by

$$E(X) = \sum_{i=1}^{n} x_i f_X(x_i) = \sum_{i=1}^{n} x_i P(X = x_i). \tag{3.11}$$

If X takes on countably infinitely many possible values $x_1, x_2, \ldots$, then $E(X)$ is defined as the sum of the infinite series (if it converges absolutely — it may not)

$$E(X) = \sum_{i=1}^{\infty} x_i f_X(x_i). \tag{3.12}$$

The expected value of X is also called the *mean of* X, and it is also denoted by μ_X (or just by μ if no confusion arises thereby). From now on, we shall treat the finite and countably infinite cases together, writing in place of (3.11) and (3.12)

$$E(X) = \sum_{i} x_i f_X(x_i). \tag{3.13}$$

The expected value of X is a key *parameter* associated with this random variable, being a weighted average of the possible values of X. Indeed, if X takes on the possible values $x_1, \ldots, x_n$, each with probability $1/n$, then $E(X)$ is simply the arithmetic mean $(x_1 + \cdots + x_n)/n$. One should not be misled by the terminology "expected" value into thinking that $E(X)$ is the most probable value of X. Indeed, $E(X)$ may very well be a value that X *never* takes on. For example, if $P(X = 0) = P(X = 1) = 1/2$, then $E(X) = 1/2$.

What, then, is the correct way of interpreting $E(X)$? As with probabilities, which can be empirical (observed relative frequencies) or theoretical (predicted relative frequencies), there are two cases to consider. If $f_X(x_i)$ records the *actual* relative frequency with which each of the values x_i of X has occurred, then $E(X)$ is simply the average of all of the observed values of X (see section 3.12 below for further details). On the other hand, if $f_X(x_i)$ is our *prediction* of the relative frequency with which the value x_i of X will occur in a yet-to-be-made sequence of observations, then $E(X)$ is our prediction of the average of all those yet-to-be-observed values.

The following theorems specify the expected values of binomial, geometric, and hypergeometric random variables.

Theorem 3.1. If $X \sim \text{binomial}(n, p)$ then $E(X) = np$.

Proof. . We have, using (3.5), Theorem 1.13, and the binomial theorem

$$E(X) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$$

$$\begin{aligned}
&= \sum_{k=1}^{n} k\binom{n}{k} p^k(1-p)^{n-k} \\
&= \sum_{k=1}^{n} k\frac{n}{k}\binom{n-1}{k-1} p^k(1-p)^{n-k} \\
&= np\sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1}(1-p)^{n-k} \\
(j = k-1) &= np\sum_{j=0}^{n-1} \binom{n-1}{j} p^{j}(1-p)^{n-1-j} \\
&= np(p+(1-p))^{n-1} \\
&= np.
\end{aligned}$$

□

Remark 1. We say that x is a *modal value*, or *mode*, of the discrete random variable X if $P(X = x) \geq P(X = y)$ for all y. Every discrete random variable has at least one modal value [why?]. We noted above that μ_X need not be a modal value of X, or even a value taken on by X at all. When $X \sim$ binomial(n, p), however, there is always a modal value of X *near* $\mu_X = np$. Specifically, one can prove that for $0 < p < 1$, there is a modal value of X at $m = \lfloor np + p \rfloor$, the greatest integer less than or equal to $np + p$. If $np + p$ is itself an integer (e.g., when n is odd and $p = 1/2$), then both $m - 1$ and m are modal values of X. Otherwise, m is the only modal value of X. In particular, if $\mu_X = np$ is itself an integer, then (since in that case $m = \lfloor np + p \rfloor = np$), $\mu_X = np$ is the only modal value of X.

Theorem 3.2. If $X \sim$ geometric(p), then $E(X) = 1/p$.

Proof. By (3.7) we have

$$E(X) = \sum_{k=1}^{\infty} k(1-p)^{k-1}p = p\sum_{k=1}^{\infty} k(1-p)^{k-1} \tag{3.14}$$

Recall that for $|x| < 1$,

$$\sum_{k=0}^{\infty} x^k = (1-x)^{-1} \tag{3.15}$$

Differentiating each side of (3.15) yields

$$\sum_{k=0}^{\infty} kx^{k-1} = \sum_{k=1}^{\infty} kx^{k-1} = (1-x)^{-2}. \tag{3.16}$$

Substituting $1-p$ for x in (3.16) enables us to simplify (3.14) to

$$E(X) = p(1-(1-p))^{-2} = 1/p. \qquad \square \tag{3.17}$$

Theorem 3.3. If $X \sim$ hypergeometric(n, A, N) then $E(X) = n\left(\frac{A}{N}\right)$.

Proof. We have, using (3.9), Theorem 1.13, and Vandermonde's Theorem (Theorem 1.14),

$$\begin{aligned}
E(X) &= \sum_{k=0}^{n} k \binom{A}{k}\binom{N-A}{n-k}\Big/\binom{N}{n} \\
&= \frac{1}{\binom{N}{n}} \sum_{k=1}^{n} k\, \frac{A}{k}\binom{A-1}{k-1}\binom{N-A}{n-k} \\
&= \frac{A}{\binom{N}{n}} \sum_{k=1}^{n} \binom{A-1}{k-1}\binom{N-A}{n-k} \\
&= \frac{A}{\binom{N}{n}} \sum_{j=0}^{n-1} \binom{A-1}{j}\binom{N-A}{n-1-j} \\
&(j = k-1) \\
&= \frac{A}{\binom{N}{n}} \binom{N-1}{n-1} \\
&= n\left(\frac{A}{N}\right). \qquad \square
\end{aligned}$$

Remark 2. If we select at random, with replacement, n things from a population of N things, A of type 1 and $N-A$ of type 2, and let X record the number of times a thing of type 1 is selected, then, as already mentioned, $X \sim$ binomial$(n, A/N)$. So $E(X) = n\left(\frac{A}{N}\right)$. So the expected number of times a thing of type 1 is selected is the same, regardless of whether we select with or without replacement.

Remark 3. Recall that, just as a probability model predicts relative frequencies of various events, the pdf of a random variable defined on a sample space equipped with a probability model predicts relative frequencies of the possible values of that random variable in a sequence of yet-to-be-performed experiments. Furthermore, the expected value of such a random variable predicts the average value of the numerical characteristic recorded by that random variable in that sequence of experiments. So Theorem 3.1 may be interpreted as predicting that, in many repetitions of a binomial experiment with parameters n and p, the average number of successes should be around np. Theorem 3.2 says that in many repetitions of an experiment where we record the number of the trial on which the first success occurs (with the probability of success on any trial equal to p), the average number of the trial on which the first success occurs should be around $1/p$. Theorem 3.3 says that in many samples of size n from a population of size N, A of which are of type 1, the average number of things of type 1 selected should be around $n\left(\frac{A}{N}\right)$. All of these results are intuitively reasonable.

3.6 Problems

1. From an urn containing 3 red, 4 white, and 5 blue balls, 12 balls are selected at random with replacement.

   a. What is the probability that at least one blue ball is selected?

   b. What is the answer to part a if the balls are selected without replacement?

2. A number is selected at random from the integers 1 through 20. Let X record the number of divisors of the number selected. Determine the pdf of X and evaluate $P(X \geq 4)$. Find $E(X)$.

3. Under the assumptions of Example 3 in §3.2, determine the values of q for which a one-engine plane is preferable to a two-engine plane.

4. Under the assumptions of Example 2 in §3.2, how many times must one select a ball at random, with replacement, from this urn in order to guarantee that the odds in favor of selecting at least *two* red balls are 100 to 1 or better?

5. If three dice are tossed repeatedly, what is the probability that a sum of 18 occurs for the first time after the 20th toss? on or before the 50th toss? What is the expected number of the toss on which one attains a sum of 18 for the first time?

6. A supposedly fair coin is tossed 24 times and it comes up heads 20 times. Test the hypothesis that the coin is fair at the 5% significance level, against a) the two-sided alternative that the coin is not fair, b) the one-sided alternative that the coin is biased in favor of heads, and c) the one-sided alternative that the coin is biased against heads.

7. A political candidate claims that at least 60% of the voters in his district prefer him to his opponent. In a random sample of 24 voters, just 10 prefer this candidate. Test the 60% hypothesis at the 5% significance level.

8. From a group of 10 men and 15 women, 5 individuals are selected, allegedly at random, to receive a special benefit. The women suspect that the selection process is biased in favor of men. It turns out that only one woman is selected. Test the hypothesis of random selection at 5% level.

9. From an urn containing 600 white and 400 red balls, 50 balls are selected at random. What is the expected number of white balls drawn if

   a. the selection is made with replacement?

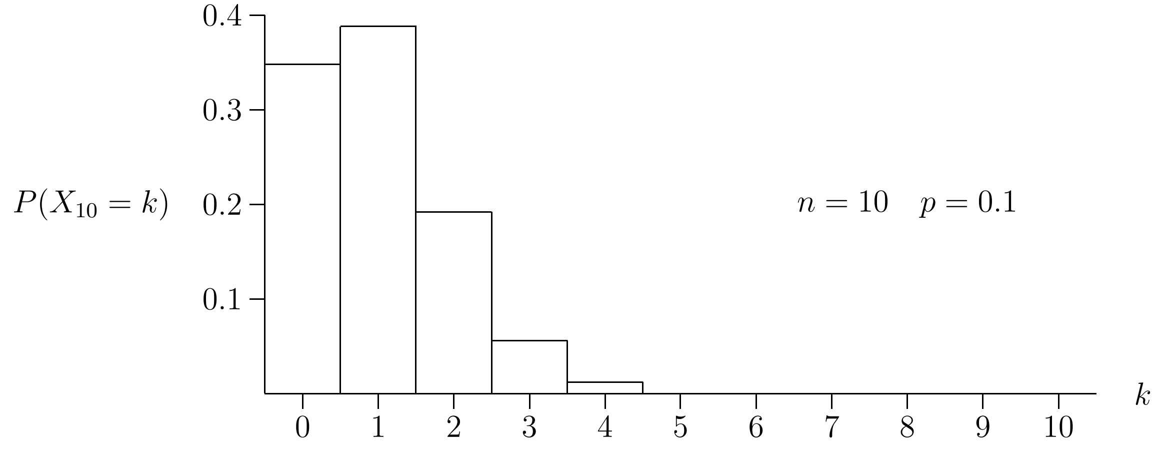
   b. the selection is made without replacement?

10. Prove that if $X \sim$ geometric (p) and m and $n \in \mathbb{P}$, with $m \leq n$, then $P(m \leq X \leq n) = (1-p)^{m-1}[1-(1-p)^{n-m+1}]$.

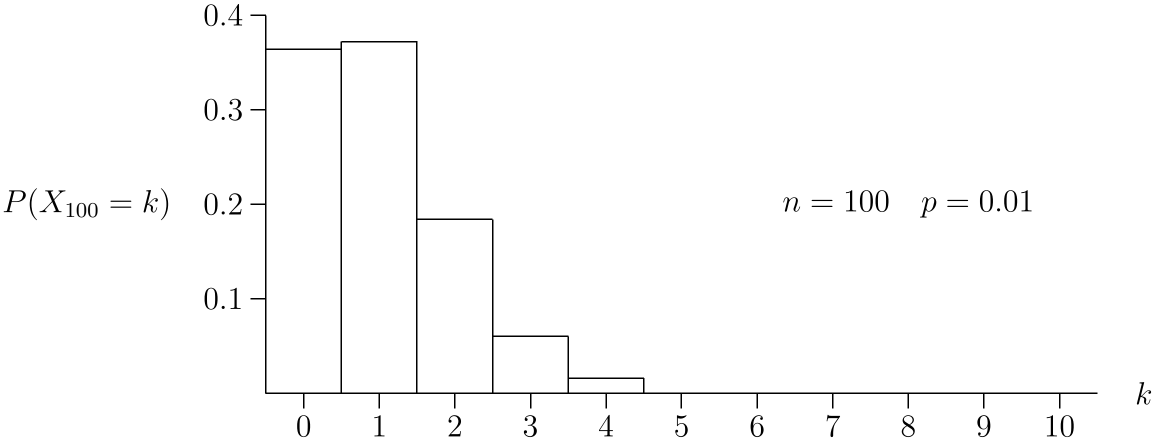
3.7 Poisson Random Variables

Consider a sequence $X_1, X_2, \ldots$ of binomial random variables, where $X_1 \sim \text{binomial}(1,1)$, $X_2 \sim \text{binomial}(2, \frac{1}{2})$, and, in general, $X_n \sim \text{binomial}(n, \frac{1}{n})$. Note that the expected value of each of these random variables is equal to 1. Partial histograms of the pdf's of X_{10}, X_{100}, and X_{1000} are shown below. They are remarkably similar, with most of the probability concentrated on values near the mean value $\mu = 1$.

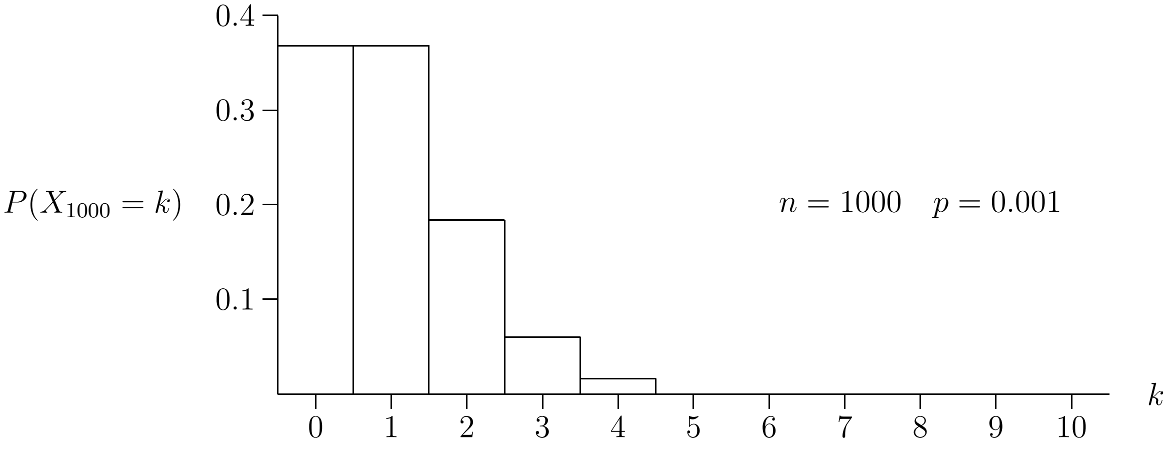
Example 1. The binomial (10, 1/10) distribution.



Example 2. The binomial (100, 1/100) distribution



Example 3. The binomial (1000, 1/1000) distribution



This suggests that these histograms approach some limiting shape as $n \to \infty$. In fact, we can prove that, for all $k \in \mathbb{N}$,

$$\lim_{n\to\infty} f_{X_n}(k) = \lim_{n\to\infty} P(X_n = k) = \frac{e^{-1}}{k!}. \tag{3.18}$$

More generally, suppose that $X_1, X_2, \ldots$ is a sequence of binomial random variables with $X_n \sim \text{binomial}(n, \lambda/n)$ for some $\lambda > 0$. Note that $E(X_n) = \lambda$ for all n. The histograms of the pdf's of these random variables also have a limiting shape.

Theorem 3.4. If $X_n \sim \text{binomial}(n, \lambda/n)$ for all $n \in \mathbb{P}$, where $\lambda > 0$, then for all $k \in \mathbb{N}$

$$\lim_{n\to\infty} f_{X_n}(k) = \lim_{n\to\infty} P(X_n = k) = \frac{\lambda^k}{k!}e^{-\lambda}. \tag{3.19}$$

Proof. We have

$$\begin{aligned}
\lim_{n\to\infty} f_{X_n}(k) &= \lim_{n\to\infty} \binom{n}{k}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{k!(n-k)!}\lambda^k n^{-k}\left(1 - \frac{\lambda}{n}\right)^{-k}\left(1 - \frac{\lambda}{n}\right)^{n} \\
&= \frac{\lambda^k}{k!} \lim_{n\to\infty} \frac{n!}{(n-k)!}(n-\lambda)^{-k}\left(1 - \frac{\lambda}{n}\right)^{n} \\
&= \frac{\lambda^k}{k!}e^{-\lambda},
\end{aligned}$$

since, for all $x \in \mathbb{R}$,

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

and

$$\lim_{n\to\infty}\frac{n!}{(n-k)!}(n-\lambda)^{-k} = \lim_{n\to\infty}\frac{n(n-1)\cdots(n-k+1)}{(n-\lambda)^k} = 1. \qquad \Box$$

The practical application of the above theorem is as follows: If $X \sim \text{binomial}(n, p)$, where n is "large" and p is "small," then for $k = 0, \dots, n$

$$P(X = k) \approx \frac{\lambda^k}{k!}e^{-\lambda}, \tag{3.20}$$

where $\lambda = np$. This is called the *Poisson approximation* (Denis Poisson, 1781-1840) *to the binomial distribution.*

Remark. It can in fact be shown that the accuracy of this approximation depends largely on the value of p, and hardly at all on the value of n. The errors in using this approximation are of the same order of magnitude as p, roughly speaking.

Example 1. Given 400 people, estimate the probability that 3 or more will have a birthday on July 4.

Solution. Assuming a year of 365 days, each equally likely to be the birthday of a randomly chosen individual, if X denotes the number of people with birthday on July 4 among 400 randomly chosen individuals, then $X \sim \text{binomial}(400, 1/365)$. The exact answer to this question is

$$P(X \geq 3) = 1 - P(X \leq 2) = 1 - \sum_{k=0}^{2}\binom{400}{k}\left(\frac{1}{365}\right)^k\left(\frac{364}{365}\right)^{400-k}.$$

The Poisson approximation of this quantity is

$$\begin{aligned}1 - \sum_{k=0}^{2}\frac{(400/365)^k}{k!}e^{-400/365} &= 1 - e^{-1.096}(1 + 1.096 + (1.096)^2/2)\\ &\approx .099.\end{aligned}$$

Example 2. Suppose that 1% of the balls in an urn are white. If we choose 10 balls at random, with replacement, what is the approximate probability that one or more will be white?

Solution. If X denotes the number of white balls chosen, then $X \sim \text{binomial}(10, .01)$. So the exact answer is

$$P(X \geq 1) = 1 - P(X = 0) = 1 - (.99)^{10} = .09561.$$

The Poisson approximation of $1 - P(X = 0)$ is

$$1 - e^{-0.1} = .09516$$

which is very close, despite the fact that n is only equal to 10.

Poisson Random Variables. Up to this point, for a fixed $\lambda > 0$, the quantities $\frac{\lambda^k}{k!}e^{-\lambda}$ for $k = 0, 1, \ldots$ have served simply as approximations of certain binomial probabilities. But note that

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}e^{-\lambda} = e^{-\lambda}\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1. \tag{3.21}$$

Thus if we define $f(k) = \frac{\lambda^k}{k!}e^{-\lambda}$ for $k = 0, 1, 2, \ldots$, f has the properties of a discrete density function for a random variable X taking as possible values $k = 0, 1, 2, \ldots$. A random variable X with this pdf is called a *Poisson random variable with parameter* λ, abbreviated $X \sim \text{Poisson}(\lambda)$. In particular, if $X \sim \text{Poisson}(\lambda)$, then $P(X = 0) = e^{-\lambda}$ and $P(X \geq 1) = 1 - e^{-\lambda}$. As you might expect, the expected value of such a random variable is λ.

Theorem 3.5. If $X \sim \text{Poisson}(\lambda)$, then $E(X) = \lambda$.

Proof. We have

$$\begin{aligned}
E(X) &= \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} \\
&= \sum_{k=1}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} \\
&= \lambda e^{-\lambda}\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\
&= \lambda e^{-\lambda}\sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \\
&(j = k-1) \\
&= \lambda e^{-\lambda}e^{\lambda} \\
&= \lambda. \qquad \square
\end{aligned}$$

While the Poisson distribution arose simply as an approximation to binomial distributions for small p, it has turned out to be extremely useful in modeling the occurrences of radioactive emissions, outbreaks of wars, accidents, etc. within some fixed period of time, and also the occurrence of stars within a fixed volume of space, misprints per page, etc. As in the case of the binomial, we are interested, in such problems, in the number of "successes" of some type, but with the difference being that the number of trials is some unspecified large number (indeed, usefully regarded as being infinite) and, instead of knowing the probability of success on a given trial, we know the average number of successes within the unit of time or space that concerns us.

Example 3. A radioactive source is monitored for 2 hours, during which time 482 alpha particles are emitted. What is the probability that exactly three particles will be emitted within a given minute? Three or fewer?

Solution. We would expect on the average 482/120 $\approx$ 4.02 emissions per minute. If X denotes the number of emissions per minute, it is reasonable to take $X \sim$ Poisson(4.02) whence

$$P(X = 3) = \frac{(4.02)^3}{3!} e^{-4.02} = 0.195$$

and

$$P(X \leq 3) = e^{-4.02} \sum_{k=0}^{3} \frac{(4.02)^k}{k!} = 0.430.$$

Example 4. In a certain book with 390 pages, 520 typographical errors occur. What is the probability that a randomly selected page is error-free?

Solution. The average number of errors per page is 520/390 = 1.33. It is reasonable to assume that the number, X, of errors on a single page satisfies $X \sim$ Poisson(1.33). Hence, $P(X = 0) = e^{-1.33} = 0.264$.

Readers are referred to §4.2 of *An Introduction to Mathematical Statistics and its Applications* by Larsen and Marx (Prentice-Hall) for further discussion of the Poisson distribution.

3.8 The Variance of a Discrete Random Variable

Recall that if X is a random variable taking the possible values $x_1, x_2, \ldots$, then

$$E(X) = \sum_i x_i f_X(x_i),$$

where $f_X(x_i) = P(X = x_i)$. Consider now a new random variable $Y = h(X)$. Suppose we wish to find $E(Y)$. Based on what we know so far, it would appear that to find $E(Y)$, we need to identify the possible values $y_1, y_2, \ldots$ of Y, find the pdf f_Y of Y and apply the formula

$$E(Y) = \sum_j y_j f_Y(y_j).$$

For example, suppose $P(X = -2) = P(X = -1) = P(X = 0) = P(X = 1) = P(X = 2) = P(X = 3) = 1/6$. Let $Y = X^2$. The possible values of Y are $0, 1, 4, 9$, with $P(Y = 0) = P(X = 0) = 1/6$, $P(Y = 1) = P(X = -1) + P(X = 1) = 1/3$, $P(Y = 4) = P(X = -2) + P(X = 2) = 1/3$, and $P(y = 9) = P(X = 3) = 1/6$. So

$$E(Y) = 0 \cdot \frac{1}{6} + 1 \cdot \frac{1}{3} + 4 \cdot \frac{1}{3} + 9 \cdot \frac{1}{6} = 19/6.$$

But notice that we could get this same result, without finding $f_Y(y)$, by evaluating the sum

$$(-2)^2 \frac{1}{6} + (-1)^2 \frac{1}{6} + 0^2 \frac{1}{6} + 1^2 \frac{1}{6} + 2^2 \frac{1}{6} + 3^2 \frac{1}{6}.$$

The following theorem tells us that we can always avoid finding f_Y.

Theorem 3.6. If X is a discrete random variable taking the possible values $x_1, x_2, \ldots$ and h is any real valued function of a real variable then

$$E(h(X)) = \sum_i h(x_i) f_X(x_i). \tag{3.22}$$

Proof. Let $Y = h(X)$ and suppose Y takes the possible values $y_1, y_2, \ldots$. Then

$$\begin{aligned} \sum_i h(x_i) f_X(x_i) &= \sum_j y_j \sum_{\substack{i: \\ h(x_i)=y_j}} P(X = x_i) \\ &= \sum_j y_j P(Y = y_j) = E(Y) = E(h(X)). \qquad \square \end{aligned}$$

Remark. This theorem is called (by mathematicians) the *law of the unconscious statistician* (LOTUS) because statisticians often mistakenly take (3.22) as a *definition* of $E(h(X))$ instead of realizing that it requires proof, even though it is intuitively reasonable.

Theorem 3.6 enables us to prove some basic properties of expected value.

Theorem 3.7. If X is a discrete random variable and $a, b \in \mathbb{R}$, then

$$E(aX + b) = aE(X) + b. \tag{3.23}$$

Proof. If X takes the possible values $x_1, x_2, \ldots$, then by LOTUS,

$$\begin{aligned} E(aX + b) &= \sum_i (ax_i + b) f_X(x_i) \\ &= a \sum_i x_i f_X(x_i) + b \sum_i f_X(x_i) \\ &= aE(X) + b. \qquad \square \end{aligned}$$

Theorem 3.8. If X is a discrete random variable and $h_1, h_2, \ldots, h_r$ are real valued functions of a real variable then

$$E(h_1(X) + h_2(X) + \cdots + h_r(X)) = E(h_1(X)) + E(h_2(X)) + \cdots + E(h_r(X)). \tag{3.24}$$

Proof. Exercise. (hint: Prove it for $r = 2$ and then use induction on r.) $\square$

The Variance of a Discrete Random Variable. Consider the random variables X and Y, where $P(X = 4) = P(X = 6) = 1/2$ and $P(Y = 2) = P(Y = 3) = P(Y = 4) =$

$P(Y=5) = p(Y=6) = P(Y=7) = P(Y=8) = 1/7$. Although $E(X) = 5 = E(Y)$ the distributions (i.e., the pdf's) of X and Y are rather different. The distribution of X is more "concentrated" around its mean value, whereas the distribution of Y is more "spread out." We would like to devise a numerical measure of how spread out the distribution of a random variable is around its mean value. Perhaps the first thing that comes to mind would be to calculate the *mean* (i.e., expected) *deviation of X from μ_X*, i.e., $E(X - \mu_X)$. But by Theorem 3.7,

$$E(X - \mu_X) = E(X) - \mu_X = \mu_X - \mu_X = 0$$

for every discrete random variable X. So $E(X - \mu_X)$ is not a good measure of spread. The problem is that values of X less than μ_X result in negative values of $X - \mu_X$ and values of X greater than μ_X result in positive values of $X - \mu_X$ and these positive and negative values, weighted by the appropriate probabilities, cancel each other out.

A natural response to this would be to switch to the *mean absolute deviation of X from μ_X*, i.e., $E(|X - \mu_X|)$ to avoid this cancellation phenomenon. While this measure of spread has some virtues, it turns out to be difficult to work with theoretically. Instead, mathematicians and statisticians almost universally use as a measure of spread the *mean squared deviation of X from μ_X*, $E((X - \mu_X)^2)$. This quantity is usually simply called the *variance of X*, denoted $\mathrm{Var}(X)$, so

$$\mathrm{Var}(X) := E((X - \mu_X)^2). \tag{3.25}$$

An alternative notation for $\mathrm{Var}(X)$ is σ_X^2, written simply as σ^2 if no confusion arises thereby.

The nonnegative square root $\sqrt{\mathrm{Var}(X)}$ of the variance of X is called the *standard deviation of X*, and denoted by $\mathrm{SD}(X)$ or by σ_X (or just by σ, if no confusion arises thereby).

Let us develop some basic properties of $\mathrm{Var}(X)$ and $\mathrm{SD}(X)$. The first result is a simpler formula for $\mathrm{Var}(X)$.

Theorem 3.9. For every discrete random variable X,

$$\mathrm{Var}(X) = E(X^2) - \mu^2. \tag{3.26}$$

Proof. By (3.25) and Theorems 3.8 and 3.7,

$$\begin{aligned}\mathrm{Var}(X) &= E((X-\mu)^2) = E(X^2 - 2\mu X + \mu^2)\\ &= E(X^2) - 2\mu E(X) + \mu^2\\ &= E(X^2) - 2\mu^2 + \mu^2\\ &= E(X^2) - \mu^2. \qquad \square\end{aligned}$$

The counterpart of Theorem 3.7 is given below.

Theorem 3.10. For every discrete random variable X, if $a, b \in \mathbb{R}$, then

$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X). \tag{3.27}$$

Proof. One can prove this result using either (3.25) or (3.26). We'll use (3.26), along with Theorem 3.8 and 3.7. Substituting $aX + b$ for X in (3.26) yields

$$\begin{aligned}\operatorname{Var}(aX+b) &= E((aX+b)^2) - (E(aX+b))^2 \\ &= E(a^2X^2 + 2abX + b^2) - (aE(X)+b)^2 \\ &= a^2E(X^2) + 2abE(X) + b^2 - a^2(E(X))^2 - 2abE(X) - b^2 \\ &= a^2(E(X^2) - \mu^2) \\ &= a^2\operatorname{Var}(X). \qquad \square\end{aligned}$$

Corollary 3.10.1. For every discrete random variable X, if $a, b \in \mathbb{R}$, then

$$\operatorname{SD}(aX+b) = |a|\operatorname{SD}(X). \tag{3.28}$$

Proof. By the definition of SD and Theorem 3.10,

$$\begin{aligned}\operatorname{SD}(aX+b) &= \sqrt{\operatorname{Var}(aX+b)} = \sqrt{a^2\operatorname{Var}(X)} \\ &= \sqrt{a^2}\,\sqrt{\operatorname{Var}(X)} = |a|\operatorname{SD}(X). \qquad \square\end{aligned}$$

Remark 1. Theorem 3.10 and Corollary 3.10.1 are intuitively reasonable. Shifting the possible values of a random variable by the distance b results in no change in the "spread" of the distribution. Multiplying these values by a, on the other hand compresses the distribution if $|a| < 1$ and expands it if $|a| > 1$.

Remark 2. Recall that in the case of expected value of a discrete random variable taking an infinite number of possible values, the infinite series representing that expected value may fail to converge. Similarly, the variance of such a random variable may be infinite.

In the following theorems, we determine the variances of the discrete random variables that we have studied so far.

Theorem 3.11. If $X \sim \text{binomial}(n, p)$, then

$$\operatorname{Var}(X) = np(1-p). \tag{3.29}$$

Proof. By (3.26) and Theorem 3.1,

$$\operatorname{Var}(X) = E(X^2) - (E(X))^2 = E(X^2) - (np)^2. \tag{3.30}$$

To find $E(X^2)$, we use the trick of writing $X^2 = X(X-1) + X$, so that

$$E(X^2) = E(X(X-1)) + E(X) = E(X(X-1)) + np. \tag{3.31}$$

By LOTUS, (3.5), and Theorem 1.13 (twice!),

$$\begin{aligned}
E(X(X-1)) &= \sum_{k=0}^{n} k(k-1)\binom{n}{k} p^k(1-p)^{n-k} \\
&= \sum_{k=2}^{n} \frac{k(k-1)(n)(n-1)}{k(k-1)}\binom{n-2}{k-2} p^k(1-p)^{n-k} \\
&= n(n-1)p^2 \sum_{k=2}^{n} \binom{n-2}{k-2} p^{k-2}(1-p)^{n-k} \\
&= n(n-1)p^2 \sum_{j=0}^{n-2} \binom{n-2}{j} p^{j}(1-p)^{n-2-j} \\
&(j = k-2) \\
&= n(n-1)p^2(p+(1-p))^{n-2} \\
&= n(n-1)p^2.
\end{aligned}$$

Substituting this result in (3.31) yields

$$E(X^2) = n(n-1)p^2 + np, \tag{3.33}$$

and substituting this result in (3.30) yields

$$\begin{aligned}
\mathrm{Var}(X) &= n(n-1)p^2 + np - (np)^2 \\
&= np((n-1)p + 1 - np) \\
&= np(1-p). \qquad \square
\end{aligned} \tag{3.34}$$

Theorem 3.12. If $X \sim \mathrm{geometric}(p)$, then

$$\mathrm{Var}(X) = \frac{1-p}{p^2}. \tag{3.35}$$

Proof. We use the same approach as in the preceding theorem, along with Theorem 3.2. This yields

$$\begin{aligned}
\mathrm{Var}(X) &= E(X^2) - (E(X))^2 \\
&= E(X(X-1) + X) - (E(X))^2 \\
&= E(X(X-1)) + E(X) - (E(X))^2 \\
&= E(X(X-1)) + \frac{1}{p} - \frac{1}{p^2}.
\end{aligned} \tag{3.36}$$

Now by LOTUS and (3.7)

$$E(X(X-1)) = \sum_{k=1}^{\infty} k(k-1)(1-p)^{k-1}p \tag{3.37}$$

$$= (1-p)p\sum_{k=2}^{\infty} k(k-1)(1-p)^{k-2}.$$

Recall that for $|x| < 1$

$$\sum_{k=0}^{\infty} x^k = (1-x)^{-1}. \tag{3.38}$$

Differentiating each side of (3.38) with respect to x twice yields

$$\sum_{k=0}^{\infty} k(k-1)x^{k-2} = \sum_{k=2}^{\infty} k(k-1)x^{k-2} = 2(1-x)^{-3}. \tag{3.39}$$

Substituting $1-p$ for x in (3.39) yields

$$\sum_{k=2}^{\infty} k(k-1)(1-p)^{k-2} = 2p^{-3}, \tag{3.40}$$

and substituting this result in (3.37) yields

$$E(X(X-1)) = 2(1-p)p^{-2}. \tag{3.41}$$

Finally, substituting this result in (3.36) yields

$$\mathrm{Var}(X) = \frac{2(1-p)}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}. \qquad \square \tag{3.42}$$

Theorem 3.13. If $X \sim$ hypergeometric(n, A, N), then

$$\mathrm{Var}(X) = n\left(\frac{A}{N}\right)\left(1-\frac{A}{N}\right)\left(\frac{N-n}{N-1}\right). \tag{3.43}$$

Proof. We leave the proof as an (honors) exercise. One uses the same approach as in the preceding two theorems, along with Theorem 1.14. □

Remark. Given a population of N things, A of type 1 and $N-A$ of type 2, let a random sample of size n be chosen from this population. Let X record the number of things of type 1 that are chosen. If we choose *with replacement*, then $X \sim$ binomial$(n, \frac{A}{N})$ so by Theorems 3.1 and 3.11,

$$E(X) = n\left(\frac{A}{N}\right) \text{ and } \mathrm{Var}(X) = n\left(\frac{A}{N}\right)\left(1-\frac{A}{N}\right). \tag{3.44}$$

If we choose *without replacement*, then $X \sim \text{hypergeometric}(n, A, N)$ and so by Theorems 3.3 and 3.13,

$$E(X) = n\left(\frac{A}{N}\right) \text{ and } \mathrm{Var}(X) = n\left(\frac{A}{N}\right)\left(1 - \frac{A}{N}\right)\left(\frac{N-n}{N-1}\right). \tag{3.45}$$

So, while the expected number of things of type 1 chosen is the same, regardless of whether we sample with or without replacement, the variance is smaller (if $n \geq 2$) when we sample without replacement. Indeed, if $n = N$ and we sample without replacement, the variance of the number of things of type 1 chosen is equal to zero. You should think about why these results are intuitively reasonable.

Theorem 3.14. If $X \sim \text{Poisson}(\lambda)$, then

$$\mathrm{Var}(X) = \lambda. \tag{3.46}$$

Proof. Using the above approach, along with Theorem 3.5, we have

$$\begin{aligned} \mathrm{Var}(X) &= E(X(X-1)) + E(X) - (E(X))^2 \\ &= E(X(X-1)) + \lambda - \lambda^2. \end{aligned} \tag{3.47}$$

By LOTUS and the fact that $P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$ for all $k \in \mathbb{N}$, we have

$$\begin{aligned} E(X(X-1)) &= \sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k}{k!}e^{-\lambda} \\ &= \sum_{k=2}^{\infty} k(k-1)\frac{\lambda^k}{k!}e^{-\lambda} \\ &= \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} \\ (j = k-2) &= \lambda^2 e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \\ &= \lambda^2 e^{-\lambda} e^{\lambda} = \lambda^2. \end{aligned} \tag{3.48}$$

Substituting this result in (3.47) yields

$$\mathrm{Var}(X) = \lambda^2 + \lambda - \lambda^2 = \lambda. \qquad \square \tag{3.49}$$

3.9 Problems

1. Suppose $X \sim \text{Poisson}(\lambda)$. Find $P(X \geq 1)$, $P(X \geq 2)$, $P(X \geq 1 | X \geq 2)$, and $P(X \geq 2 | X \geq 1)$. Are the events "$X \geq 1$" and "$X \geq 2$" independent?

2. If 1% of the population has a certain rare blood type, find the approximate probability that, in a random sample of 100 individuals, more than 3 will have this blood type. Compare your answer to the probability of this event based on the binomial distribution.

3. Suppose 3 fair dice are tossed 648 times. Find the approximate probability that a sum greater than 3 comes up (exactly) 646 times. Compare your answer to the exact probability of this event.

4. If the random variable X has a Poisson distribution and $P(X = 1) = P(X = 2)$, find $P(X = 4)$.

5. Prove that if $X \sim \text{Poisson}(\lambda)$, then $P(X \text{ is even}) = \frac{1}{2}(1 + e^{-2\lambda})$.

6. Flaws in a particular kind of metal sheeting occur at an average rate of one per 10 ft^2. What is the probability of 2 or more flaws in a 5 × 8 ft. sheet?

7. Granny claims that the average number of raisins in one of her muffins is 8. You suspect that the average number is smaller than 8. You get a muffin with 3 raisins. Evaluate Granny's claim at the 5% significance level, using an appropriate Poisson distribution.

8. From a population of 100 men and 150 women a random sample of size 40 is drawn. Find the mean and standard deviation of the number of women selected if a) we sample with replacement and b) we sample without replacement.

9. Suppose X is a discrete random variable and $\text{Var}(X) = 0$. Describe the pdf of X.

3.10 Chebyshev's Theorem

As we shall see, the standard deviation σ of a random variable X is the perfect unit of measurement for expressing the distance between possible values of X and its expected value μ. The following theorem of Chebyshev (Pavnuti L. Chebyshev, 1829-1894) supports this assertion.

Theorem 3.15. Let X be a discrete random variable with mean μ and standard deviation $\sigma > 0$. Then for all $h > 0$,

$$P(|X - \mu| \geq h\sigma) \leq \frac{1}{h^2}, \quad \text{and so} \tag{3.50}$$

$$P(|X - \mu| \leq h\sigma) \geq 1 - \frac{1}{h^2}. \tag{3.51}$$

Proof. By (3.25) and LOTUS,

$$\sigma^2 = \sum_i (x_i - \mu)^2 f(x_i), \tag{3.52}$$

where f is the pdf of X. Now the sum in (3.52) can be broken into two parts. In one part we group together all the terms for which $|x_i - \mu| < h\sigma$, and in the other we group together all the terms for which $|x_i - \mu| \geq h\sigma$. This yields

$$\sigma^2 = \sum_{\substack{i: \\ |x_i - \mu| < h\sigma}} (x_i - \mu)^2 f(x_i) + \sum_{\substack{i: \\ |x_i - \mu| \geq h\sigma}} (x_i - \mu)^2 f(x_i). \tag{3.53}$$

Each of the sums on the RHS of (3.53) is nonnegative and so, clearly,

$$\sum_{\substack{i: \\ |x_i - \mu| \geq h\sigma}} (x_i - \mu)^2 f(x_i) \leq \sigma^2. \tag{3.54}$$

But if $|x_i - \mu| \geq h\sigma$, then $(x_i - \mu)^2 \geq h^2\sigma^2$, so

$$\sum_{\substack{i: \\ |x_i - \mu| \geq h\sigma}} (x_i - \mu)^2 f(x_i) \geq \sum_{\substack{i: \\ |x_i - \mu| \geq h\sigma}} h^2\sigma^2 f(x_i) = \tag{3.55}$$

$$= h^2\sigma^2 \sum_{\substack{i: \\ |x_i - \mu| \geq h\sigma}} f(x_i) = h^2\sigma^2 P(|X - \mu| \geq h\sigma)$$

Combining (3.55) and (3.54) yields

$$h^2\sigma^2 P(|X - \mu| \geq h\sigma) \leq \sigma^2, \tag{3.56}$$

and dividing each side of (3.56) by the positive quantity $h^2\sigma^2$ yields

$$P(|X - \mu| \geq h\sigma) \leq \frac{1}{h^2}. \tag{3.57}$$

Finally, since $|X - \mu| < h\sigma$ implies $|X - \mu| \leq h\sigma$, and is the complement of the event $|X - \mu| \geq h\sigma$,

$$P(|X - \mu| \leq h\sigma) \geq P(|X - \mu| < h\sigma) = 1 - P(|X - \mu| \geq h\sigma) \geq 1 - \frac{1}{h^2}, \tag{3.58}$$

by (3.57). □

Remark. When $h \leq 1$, Theorem 3.15 tells us nothing of interest [Why?]. But when $h > 1$, (3.50) gives a useful upper bound on the probability that X takes on a value at least h standard deviations away from its mean μ, and (3.51) gives a useful lower bound on the probability that X takes on a value within h standard deviations of μ. Setting $h = 2$ in (3.50), for example, yields the result that, for any discrete random variable X, the probability that X takes on a value at least 2 standard deviations away from its mean is always less than or equal to $\frac{1}{4}$. Similarly, (3.51) tells us that the probability that X takes on a value within 2 standard deviations of its mean (i.e., in the interval $[\mu - 2\sigma, \mu + 2\sigma]$) is always greater than or equal to $\frac{3}{4}$.

Example 1. If a fair coin is flipped 100 times, find a lower bound on the probability that the number of heads is between 34 and 66 inclusive.

Solution. If X denotes the number of heads, then $X \sim \text{binomial}(100, \frac{1}{2})$ and so $\mu = 50$ and $\sigma = 5$ [why?]. So

$$\begin{aligned} P(34 \le X \le 66) &= P(-16 \le X - 50 \le 16) = P(|X - 50| \le 16) \\ &= P(X - 50| \le \frac{16}{5} \cdot 5) = P(|X - \mu| \le \frac{16}{5}\sigma) \\ &\ge 1 - \frac{25}{256} = \frac{231}{256} \approx 0.90. \end{aligned}$$

As stated, (3.51) furnishes a lower bound on the probability that X takes on a value inside a closed interval symmetrical about μ. The next example treats the case of an asymmetrical interval.

Example 2. For the problem above, find a lower bound on the probability that the number of heads is between 30 and 66 inclusive.

Solution. Shrink the interval to the largest symmetrical interval around μ and proceed as in Example 1:

$$P(30 \le X \le 66) \ge P(34 \le X \le 66) \ge \frac{231}{256}.$$

The event $34 \le X \le 66$ implies (i.e., is a subset of) the event $30 \le X \le 66$, so the latter has probability greater than or equal to the former.

The bounds given by Chebyshev's Theorem can be fairly crude. The following examples illustrate this.

Example 3. If $X \sim \text{Poisson}(5)$, find a lower bound on $P(1 \le X \le 9)$.

Solution. Since $\mu = 5$ and $\sigma = \sqrt{5}$ [why?],

$$\begin{aligned} P(1 \le X \le 9) &= P(-4 \le X - 5 \le 4) = P(|X - 5| \le 4) \\ &= P((X - 5) \le \frac{4}{\sqrt{5}}\sqrt{5}) = P(|X - \mu| \le \frac{4}{\sqrt{5}}\sigma) \\ &\ge 1 - \frac{5}{16} \approx .69. \end{aligned}$$

This should be compared to the exact value, $P(1 \le X \le 9) = \sum_{k=1}^{9} e^{-5}\frac{k^5}{5!} \approx .96$.

Example 4. A fair coin is tossed 100 times. Using Chebyshev's Theorem, find the shortest closed interval I that is symmetric around 50 and has the property that the probability that the number of heads lies in I is greater than or equal to .95.

Solution. If X records the number of heads, then $X \sim \text{binomial}(100, \frac{1}{2})$, so $E(X) = 50$ and $SD(X) = 5$. For each $h > 0$

$$(*) \qquad P(|X - 50| \leq h \cdot 5) \geq 1 - \frac{1}{h^2}.$$

To make the LHS of $(*) \geq .95$, it suffices to make $1 - \frac{1}{h^2} \geq .95$, i.e., to make $h \geq 2\sqrt{5}$. We want I as short as possible so we take $h = 2\sqrt{5}$ yielding $I = [50 - 10\sqrt{5}, 50 + 10\sqrt{5}] = [27.64, 72.36]$. Since X takes only integral values, we can shorten this to $[28, 72]$. Using the normal approximation to the binomial, we'll shorten this later to $[40, 60]$.

3.11 Bernoulli's Law of Large Numbers

Let $X \sim \text{binomial}(n, p)$. The random variable X records the *number of successes* in n independent trials, on each of which the probability of success is p. Given X, let us define a new random variable $\bar{X}$ (sometimes denoted by $\hat{p}$ in statistics texts) by the formula

$$(3.59) \qquad \bar{X} = \frac{X}{n}.$$

Clearly, $\bar{X}$ records the *fraction of successes* in the aforementioned n trials. Combining Theorems 3.1 and 3.7, it follows that

$$(3.60) \qquad E(\bar{X}) = \frac{1}{n}E(X) = \frac{1}{n}(np) = p$$

and combining Theorems 3.11 and Corollary 3.10.1, it follows that

$$(3.61) \qquad \begin{aligned} SD(\bar{X}) &= \frac{1}{n}SD(X) = \frac{1}{n}\sqrt{np(1-p)} \\ &= \sqrt{\frac{p(1-p)}{n}}. \end{aligned}$$

The number $SD(\bar{X})$ is often called the *standard error of the mean* in statistics texts.

An important consequence of Chebyshev's Theorem is the following rough specification of the behavior of $\bar{X}$.

Theorem 3.16. Let $X \sim \text{binomial}(n, p)$ and let $\bar{X} := X/n$, the fraction of successes in n trials. Then, for each fixed $d > 0$,

$$(3.62) \qquad P(|\bar{X} - p| \geq d) \leq \frac{p(1-p)}{nd^2},$$

and so,

$$(3.63) \qquad P(|\bar{X} - p| \leq d) \geq 1 - \frac{p(1-p)}{nd^2}.$$

Proof. By (3.50) with $\bar{X}$ substituted for X and $d/\sqrt{\frac{p(1-p)}{n}}$ substituted for h,

$$P(|\bar{X}-p| \geq d) = P\left(|\bar{X}-p| \geq \frac{d}{\sqrt{\frac{p(1-p)}{n}}}\sqrt{\frac{p(1-p)}{n}}\right)$$

$$= P\left(|\bar{X}-\mu_{\bar{X}}| \geq \frac{d}{\sqrt{\frac{p(1-p)}{n}}}\sigma_{\bar{X}}\right) \leq \frac{p(1-p)}{nd^2}.$$

The same substitution in (3.51) yields

$$P(|\bar{X}-p| \leq d) \leq 1 - \frac{p(1-p)}{nd^2}. \qquad \square$$

Example 1. How many times should we toss a fair die in order to ensure with probability at least .95 that the fraction of aces (1's) observed is within .01 of the true probability, $\frac{1}{6}$, of an ace?

Solution. If we toss the die n times and let X record the number of aces, then $X \sim$ binomial$(n, \frac{1}{6})$. With $\bar{X} := X/n$, and $d = .01$, we have by (3.63) above that

$$P\left(\left|\bar{X}-\frac{1}{6}\right| \leq .01\right) \geq 1 - \frac{\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)}{n(.01)^2}. \tag{3.64}$$

To make the LHS of (3.64) greater than or equal to .95 it therefore suffices to make the RHS of (3.64) greater than or equal to .95. Solving the inequality

$$1 - \frac{\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)}{n(.01)^2} \geq .95$$

for n yields $n \geq \left(\frac{1}{6}\right)\left(\frac{5}{6}\right)/(.01)^2(.05) \approx 27,778$.

Remark. While taking $n \geq 27,778$ is sufficient to yield the desired result, it is in fact not necessary to take n this large. This is the best we can do using Chebyshev's Theorem. Later, when we study the normal approximation to the binomial, we shall see that $n \geq 5336$ suffices to yield the desired result.

In Example 1 above, we took a fairly narrow symmetric interval around $\frac{1}{6}$ and showed that by making n sufficiently large we could make the probability that $\bar{X}$ takes a value in that interval greater than or equal to .95. It is pretty clear no matter how narrow the interval (as long as it has some positive radius d) and no matter how large the probability (as long as it's less than 1), problems of the type posed in this example can be solved. The general statement covering this phenomenon is called *Bernoulli's Law of Large Numbers* (Jacob Bernoulli, 1654-1705), which is itself a special case of a more general result called the *weak law of large numbers.* Bernoulli's Law of Large Numbers specifies the behavior of the fraction of successes in a binomial experiment with n trials as $n \to \infty$. It is stated below.

Theorem 3.17. Let $X \sim \text{binomial}(n, p)$ and let $\bar{X} := X/n$, the fraction of successes in n trials. Then, for each fixed $d > 0$,

$$\lim_{n\to\infty} P(|\bar{X} - p| \le d) = 1. \tag{3.66}$$

Proof. By (3.63) and the fact that probabilities are always less than or equal to 1, we have

$$1 - \frac{p(1-p)}{nd^2} \le P(|\bar{X} - p| \le d) \le 1. \tag{3.67}$$

Since $\lim_{n\to\infty} p(1-p)/nd^2 = 0$, we can make $1 - p(1-p)/nd^2$ and hence $P(|\bar{X} - p| \le d)$ as close as we like to 1 (though never equal to 1) by taking n sufficiently large. This proves (3.66). □

Remark. Our proof of Theorem 3.17 is a straightforward application of (3.63), which is, in essence, Chebyshev's Theorem particularized to the random variable $\bar{X}$. When Bernoulli proved this result, Chebyshev had not yet been born. Bernoulli had to construct a very subtle and clever argument to prove Theorem 3.17.

Theorem 3.17 and various generalizations thereof are often referred to as the *law of averages.* It is important not to read more into such statements than is warranted. If you flip a fair coin 10 times and it comes up heads 8 of those times, this does *not* raise the probability of "tails" on the next toss (the coin has no memory!) As William Feller, an eminent probabilist, once said, "The law of averages works by swamping, not by compensation." We do not have to make up (i.e. compensate for) the "deficit" of 3 tails after 10 flips. It is washed out (swamped) when we divide by an increasingly large n. Also, we do not necessarily get closer to p with each trial. See the next page, which illustrates how much fluctuation can occur.

100 Tosses of a Fair Coin
(1=heads, 0=tails, CRF=cumulative relative frequency of heads)

Toss	Result	CRF	Toss	Result	CRF	Toss	Result	CRF
1	0	0	35	0	.457	69	0	.478
2	0	0	36	0	.444	70	1	.486
3	1	.333	37	0	.432	71	1	.493
4	1	.500	38	1	.447	72	0	.486
5	0	.400	39	1	.462	73	1	.493
6	0	.333	40	1	.475	74	1	.500
7	0	.286	41	1	.488	75	1	.507
8	0	.250	42	1	.500	76	1	.513
9	1	.333	43	0	.488	77	0	.506
10	1	.400	44	1	.500	78	1	.513
11	1	.455	45	1	.511	79	0	.506
12	0	.417	46	1	.522	80	0	.500
13	0	.385	47	0	.511	81	0	.494
14	0	.357	48	0	.500	82	0	.488
15	0	.333	49	0	.490	83	1	.494
16	1	.375	50	1	.500	84	0	.488
17	0	.353	51	1	.510	85	1	.494
18	0	.333	52	0	.500	86	0	.488
19	1	.368	53	0	.491	87	0	.483
20	0	.350	54	0	.481	88	0	.477
21	1	.381	55	0	.473	89	1	.483
22	0	.364	56	1	.482	90	1	.489
23	0	.348	57	1	.491	91	0	.484
24	1	.375	58	0	.483	92	1	.489
25	1	.400	59	1	.492	93	1	.495
26	0	.385	60	1	.500	94	1	.500
27	1	.407	61	0	.492	95	0	.495
28	1	.429	62	0	.484	96	0	.490
29	1	.448	63	1	.492	97	1	.495
30	0	.433	64	0	.484	98	0	.490
31	0	.419	65	1	.492	99	0	.485
32	1	.438	66	1	.500	100	0	.480
33	1	.455	67	0	.493			
34	1	.471	68	0	.485			

The average number of successes $\bar{X}$ in a binomial experiment with n trials is, by Theorem 3.17, as close to the true probability p of success as we wish, with probability as close to 1 as we wish, for sufficiently large n. Thus we may use the observed value of $\bar{X}$ as an *estimate* of p when p is unknown. How many observations n should we make in order to yield a given level of accuracy with a given probability? Formula (3.63), which we repeat below as (3.68), is relevant to this problem:

$$P(|\bar{X}-p| \leq d) \geq 1 - \frac{p(p-1)}{nd^2}. \tag{3.68}$$

You may think that this can't be useful since it contains the unknown parameter p. But wait. Using a little calculus, it is easy to see that on the interval $[0,1]$ the function $p(p-1)$ attains its maximum value of $\frac{1}{4}$ when $p = \frac{1}{2}$, i.e., that $p(p-1) \leq \frac{1}{4}$ for all $p \in [0,1]$. But this means that $p(p-1)/nd^2 \leq 1/4nd^2$ for all $p \in [0,1]$ and so

$$1 - \frac{p(p-1)}{nd^2} \geq 1 - \frac{1}{4nd^2} \text{ for all } p \in [0,1]. \tag{3.69}$$

Combined with (3.68), this yields

$$P(|\bar{X}-p| \leq d) \geq 1 - \frac{1}{4nd^2}, \tag{3.70}$$

and (3.70) is very useful indeed, as the following example illustrates.

Example 5. It is desired to use the observed value of $\bar{X}$ as an estimate for the unknown binomial parameter p, with the probability being at least .97 that the error is no more than .05. How large should n be?

Solution. By (3.70), to make $P(|\bar{X}-p| \leq .05) \geq .97$, it suffices to make

$$1 - \frac{1}{4n(.05)^2} \geq .97. \tag{3.71}$$

Solving (3.71) for n yields $n \geq 1/(.03)(4)(.05)^2 \approx 3333$.

Remark 1. In fact, the value $n = 3333$ is much larger than necessary. This is the best we can do using Chebyshev's Theorem. Using the normal approximation to the binomial, we shall see later that $n \geq 471$ suffices to yield the desired accuracy with the desired probability. We shall develop this example further when we study confidence intervals in Chapter 4.

Remark 2. It is important not to read into Example 5 more than is warranted. The fact that $n \geq 3333$ implies that $P(|\bar{X}-p| \leq .05) \geq .97$ does *not* mean that if we carry out *one* set of 3333 trials and record the fraction q of successes on these trials, then $P(|q-p| \leq .05) \geq .97$. Indeed, the statement $|q-p| \leq .05$ cannot meaningfully be assigned a probability — it is either true or false. What the analysis of Example 5 tells us is that if we carry out *many* (say, N, for some large N) sets of 3333 trials and record the fraction of successes, say, $q_1, q_2, \ldots, q_N$

for each of these sets of trials, then $|q_i - p| \le .05$ for around 97% of the q_i's. The probability .97 is thus a *performance characteristic* associated with the *general method of estimating p by carrying out 3333 trials*, not a "score" assigned to an estimate materializing from a single application of that method.

3.12 Data

Suppose that a die is tossed 25 times and that the sequence of outcomes observed is

$$(4, 2, 2, 6, 5, 1, 6, 6, 4, 4, 4, 5, 4, 5, 4, 3, 1, 5, 5, 1, 4, 2, 2, 4, 2) \tag{3.72}$$

This is an example of *ungrouped data*, the general case of which is a sequence

$$x = (x_1, x_2, \dots, x_n) \tag{3.73}$$

of observed numerical outcomes of n repetitions of some experiment. Given (3.73) define the *mean* $\bar{x}$ of this data sequence by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{3.74}$$

the *variance* s_x^2 (denoted simply by s^2 if no confusion arises thereby) by

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2. \tag{3.75}$$

and the *standard deviation* s_x (denoted simply by s if no confusion arises thereby) by

$$s_x = \sqrt{s_x^2}. \tag{3.76}$$

In the case of (3.72), $n = 25$ and, as you can check, $\bar{x} = 3.64$, $s^2 = 2.47$, and $s = 1.57$.

An alternative way to record the outcomes of the 25 tosses of the die is to display these outcomes as *grouped data*, as in the following table

(3.77)

outcome	frequency
1	3
2	5
3	1
4	8
5	5
6	3

In the general case of grouped data, $x_1, \dots, x_t$ represent *distinct* outcomes, with x_i occurring n_i times in the data:

(3.78)

outcome	frequency
x_1	n_1
x_2	n_2
$\vdots$	$\vdots$
x_t	n_t,

where $n = n_1 + \cdots + n_t$. When data are grouped, formula (3.74) reduces to

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{t} x_i n_i, \tag{3.79}$$

and formula (3.75) to

$$s_x^2 = \frac{1}{n} \sum_{i=1}^{t} (x_i - \bar{x})^2 n_i = \frac{1}{n} \sum_{i=1}^{t} x_i^2 n_i - (\bar{x})^2. \tag{3.80}$$

The quantities $\bar{x}$, s^2, and s are sometimes called the *sample mean*, the *sample variance*, and the *sample standard deviation* (even though we may not, literally, be generating the numerical observations in question by sampling from some population; in the case of the die we are in some sense "sampling the behavior" of the die). We now make an important observation. Given the grouped data in (3.78), let us define a random variable X taking the possible values $x_1, \ldots, x_t$ with

$$f_X(x_i) = P(X = x_i) = \frac{n_i}{n}. \tag{3.81}$$

Clearly, (3.81) defines a pdf since $n_1 + \cdots + n_t = n$. It is easy to check that

$$E(X) = \bar{x}, \quad \operatorname{Var}(X) = s_x^2 \quad \text{and} \quad SD(X) = s_x, \tag{3.82}$$

i.e., the sample mean, sample variance, and sample standard deviation are simply the mean, variance, and standard deviation of the "empirical random variable" X naturally associated with the sample data.

This means, of course, that all of our theorems about means, variances, and standard deviations of random variables yield as special cases theorems about sample means, sample variances, and sample standard deviations.

Example. Given a linear transformation $y = ax + b$ of the "data variable" x (i.e., a transformation of data value x_i to $y_i = ax_i + b$ for each i), it follows from Theorem 3.7 that

$$\bar{y} = a\bar{x} + b \tag{3.83}$$

and from Theorem 3.10 and Corollary 3.10.1 that

$$s_y^2 = a^2 s_x^2, \quad \text{and} \tag{3.84}$$

$$s_y = |a| s_x. \tag{3.85}$$

Example. The computationally simplest formulas for s_x^2, namely,

$$s_x^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - (\bar{x})^2 \tag{3.86}$$

for ungrouped data and

$$(3.87) \qquad s_x^2 = \frac{1}{n}\sum_{i=1}^{t} x_i^2 n_i - (\bar{x})^2$$

for grouped data are simple consequences of Theorem 3.9.

Remark. In most statistics texts, one finds the quantity

$$(3.88) \qquad s_x^2{+} := \frac{1}{n-1}\sum_{i=1}^{r}(\bar{x}-x_i)^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{n}{n-1}(\bar{x})^2$$

being called the sample variance (and being denoted s_x^2, not s_x^2+). As an estimate of the variance of a population based on a sample, s_x^2+ is slightly preferable to s_x^2 (note that s_x^2+ is slightly larger than s_x^2), but the difference is negligible for large samples.

As a special case of Theorem 3.15, we get

Chebyshev's Theorem for Data: In any sequence of numerical observations, the fraction of observations that lie at least h standard deviations from the sample mean is less than or equal to $1/h^2$; the fraction of observations that lie at or within h standard deviations from the sample mean is greater than or equal to $1 - 1/h^2$.

Example. For the data in (3.72), or, equivalently, (3.77), the interval of values at or within 1.5 standard deviations from the sample mean 3.64 is $[1.285, 5.995]$. There are 19 observations in this interval (all the 2's, 3's, 4's, and 5's) and, as asserted, $19/25 \geq 1 - 1/\left(\frac{3}{2}\right)^2 = 5/9$.

3.13 Problems

1. From a set of 30 red and 70 white balls, 50 balls are selected at random. Find a lower bound on the probability that the number of red balls selected lies in the interval $[10, 20]$ if a. sampling is done with replacement and b. sampling is done without replacement.

2. Let X be a discrete random variable taking only nonnegative values, with $E(X) = \mu$. Prove, by an argument similar to the proof of Chebyshev's theorem, that if $a > 0$, $P(X \geq a) \leq \mu/a$. The result is known as *Markov's inequality.*

3. The class average on a certain test was 60 and your score was 80. Your percentile rank (the percentage of students with scores strictly lower than yours) is at least ______. Hint: use Markov's inequality.

4. Prove that if X is a discrete random variable with $E(X) = \mu$ and $SD(X) = \sigma > 0$, then for each fixed $d > 0$,

$$P(|X-\mu| \geq d) \leq \frac{\sigma^2}{d^2}.$$

5. Referring to the results of problem 10 (§3.6), explain why one does not need to call upon the probability estimates furnished by Chebyshev's Theorem in connection with problems about geometric random variables.

6. A fair coin is flipped 64 times. Find a lower bound on the probability that the number of heads observed is between 25 and 39 inclusive.

7. A fair coin is flipped 64 times. Find the shortest closed interval I that is symmetric around 32 and has the property that the probability that the number of heads lies in I is at least .80.

8. How many times should we toss a fair coin so that the probability is at least .60 that the fraction of heads observed is within .02 of $\frac{1}{2}$?

9. We want to estimate the fraction of voters favoring a ban on assault weapons. How large a sample should we take so that the probability is at least .95 that the error is no more than .10? (We will use the fraction of the sample favoring the ban as our estimate of the fraction of all voters favoring the ban).

10. Suppose noontime temperatures are recorded each day for a year in Knoxville and that the average temperature is 60° F, with a standard deviation of 10° F. If all the daily temperatures observed were converted to Centigrade, what would the average and standard deviations of the transformed temperatures be?

CHAPTER 4

CONTINUOUS RANDOM VARIABLES

4.1 Introduction

In the previous chapter we considered random variables whose set of possible values was finite or countably infinite. In this chapter we consider random variables whose set of possible values is uncountably infinite. Random variables that record "continuous" quantities such as time (e.g., arrival times or lifetimes), mass, and length furnish examples of the latter type of random variable. We say that X is a *continuous random variable* if there is a function $f : \mathbb{R} \to [0, \infty)$ such that

$$\int_{-\infty}^{\infty} f(x)dx = 1, \tag{4.1}$$

and, more generally, for every set $B \subseteq \mathbb{R}$,

$$P(X \in B) = \int_B f(x)dx. \tag{4.2}$$

The function f (sometimes denoted f_X for clarity) is called the probability density function (abbreviated pdf) of X. Any function $f : \mathbb{R} \to [0, \infty)$ satisfying (4.1) is called a *continuous probability density function.*

An important special case of (4.2) is the formula

$$P(a \le X \le b) = \int_a^b f(x)dx. \tag{4.3}$$

Letting $a = b$ in (4.3), we obtain

$$P(X = a) = \int_a^a f(x)dx = 0, \tag{4.4}$$

i.e., *the probability that X takes on a given value a is zero, for every $a \in \mathbb{R}$!* This points up a crucial difference between discrete and continuous pdf's. If f is the pdf of a *discrete* random variable X taking possible values $x_1, x_2, \ldots$, then

$$P(X = x_i) = f(x_i), \quad i = 1, 2, \ldots\ . \tag{4.5}$$

If, however, f is the pdf of a continuous random variable, then, for each $x \in \mathbb{R}$

$$P(X = x) = 0, \tag{4.6}$$

which is generally not equal to $f(x)$. Indeed, it is perfectly possible for a continuous pdf to take values greater than 1!

Example. Let $f(x) = \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$ Then f is a continuous[1] pdf, for $f(x) \geq 0$ for all $x \in \mathbb{R}$ and

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)dx &= \int_{-\infty}^{0} 0dx + \int_{0}^{1} 2xdx + \int_{1}^{\infty} 0dx \\ &= 0 + 1 + 0 = 1.\end{aligned}$$

Notice that, for example, $f(\frac{3}{4}) = 1.5$. So it is clear that $P(X = \frac{3}{4}) \neq f(\frac{3}{4})$. Indeed, as we know,

$$P(X = \frac{3}{4}) = \int_{\frac{3}{4}}^{\frac{3}{4}} 2xdx = 0.$$

Example. Suppose that X is a continuous random variable whose pdf is given by

$$f(x) = \begin{cases} c(4x - 2x^2) & 0 \leq x \leq 2 \\ 0 & \text{elsewhere.} \end{cases}$$

Determine c and find $P(X \geq 1)$.

Solution. We must have

$$\int_{0}^{2} c(4x - 2x^2)dx = 1,$$

i.e.,

$$\begin{aligned}c\int_{0}^{2} (4x - 2x^2)dx &= c\left[2x^2 - \frac{2}{3}x^3\right]_0^2 \\ &= c \cdot \frac{8}{3} = 1.\end{aligned}$$

Hence $c = \frac{3}{8}$, and

$$P(X \geq 1) = \int_{1}^{\infty} f(x)dx = \int_{1}^{2} \frac{3}{8}(4x - 2x^2)dx = \frac{1}{2}.$$

Remark. If X is a continuous random variable, then, as we have seen, $P(X = a) = 0$ for every $a \in \mathbb{R}$. This means that *for X continuous*,

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) \tag{4.7}$$

[1] Note that f is discontinuous at $x = 1$. The terminology "continuous pdf" is shorthand for "pdf of a continuous random variable." It does not mean that f is continuous at every $x \in \mathbb{R}$.

$$= P(a < X < b)$$

(4.8) $$P(X \geq a) = P(X > a), \quad \text{and}$$

(4.9) $$P(X \leq b) = P(X < b),$$

results which of course do *not* hold in general for discrete random variables.

Example. The amount of time, in hours, that a certain machine functions before breaking down is a continuous random variable X, with pdf given by

$$f(x) = \begin{cases} \frac{1}{100}e^{-x/100} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

What is the probability that the machine functions for less than 100 hours?

Solution.

$$\begin{aligned} P(X < 100) = P(X \leq 100) &= \int_0^{100} \frac{1}{100}e^{-x/100} \\ &= -e^{-x/100}\Big|_0^{100} = 1 - e^{-1} \approx .633. \end{aligned}$$

4.2 The Cumulative Distribution Function of a Continuous Random Variable

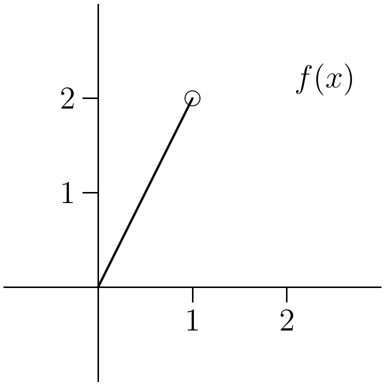
Let X be a continuous random variable with pdf f. The *cumulative distribution function* (cdf) F of X (sometimes denoted F_X for clarity) is defined for all $x \in \mathbb{R}$ by

(4.10) $$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

It follows from a basic theorem of calculus that, *at every x where f is continuous*,

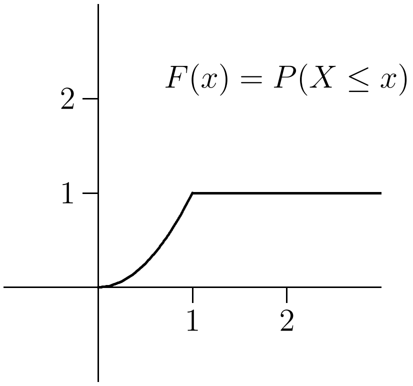
(4.11) $$f(x) = F'(x).$$

Example. Let $f(x) = \begin{cases} 2x & 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$



Then, for all $x \in \mathbb{R}$, $F(x) = \int_{-\infty}^{x} f(t)dt$. So

$$F(x) = \begin{cases} 0 & -\infty < x \leq 0 \\ x^2 & 0 < x < 1 \\ 1 & x \geq 1. \end{cases}$$



Note that f is continuous for all $x \neq 1$ and, as asserted in (4.11), $f(x) = F'(x)$ for each $x \neq 1$ (at $x = 1$, F is not differentiable since its left derivative equals 2 and its right derivative equals 0).

Remark. We recall that the cdf of a discrete random variable is always a step function. On the other hand, *the cdf of n continuous random variable is always continuous for every x.* Moreover, one can prove that

(4.12)
$$\lim_{x \to -\infty} F(x) = 0, \quad \text{and}$$
$$\lim_{x \to \infty} F(x) = 1.$$

The limits in (4.12) also hold for discrete random variables.

4.3 The Mean and Variance of a Continuous Random Variable

Recall that if X is a discrete random variable taking the possible values $x_1, x_2, \ldots$ then the expected value (or mean) $E(X)$ of X is defined by

(4.13)
$$E(X) := \sum_i x_i f(x_i),$$

where f is the pdf of X. If X is continuous, with pdf f, we define $E(X)$ by the formula

(4.14)
$$E(X) := \int_{-\infty}^{\infty} x f(x) dx.$$

As in the case of discrete random variables, $E(X)$ is also denoted by μ_X or (if no confusion arises thereby) by μ.

Our study of the mean and variance of a discrete random variable made frequent use of Theorem 3.6 ("the law of the unconscious statistician - discrete case). Similarly, our analysis of the mean and variance of a continuous random variable will make frequent use of the following theorem:

Theorem 4.1 (law of the unconscious statistician - continuous case). If X is a continuous random variable with pdf f and h is any real valued function of a real variable, then

$$E(h(X)) = \int_{-\infty}^{\infty} h(x)f(x)dx. \tag{4.15}$$

Proof. The proof is beyond the level of this course. Interested students should consult Chapter 7 of Sheldon Ross's text, *A First Course in Probability* (MacMillan Publishing Company) for the proof. □

Using the above theorem, we may prove the following continuous analogues of Theorems 3.7 and 3.8.

Theorem 4.2. If X is a continuous random variable and $a, b \in \mathbb{R}$, then

$$E(aX + b) = aE(X) + b. \tag{4.16}$$

Proof. Exercise. □

Theorem 4.3. If X is a continuous random variable and $h_1, h_2, \ldots, h_r$ are real valued functions of a real variable, then

$$E(h_1(X) + \cdots + h_r(X)) = E(h_1(X)) + \cdots + E(h_r(X)). \tag{4.17}$$

Proof. Exercise. □

The variance of a continuous random variable X is defined just as it is in the discrete case, namely,

$$\text{Var}(X) := E((X - \mu)^2), \tag{4.18}$$

where $\mu = E(X)$, as determined by (4.14). $\text{Var}(X)$ is also denoted by σ_X^2 or by σ^2. The standard deviation of X, $SD(X)$ is defined by

$$SD(X) = \sqrt{\text{Var}(X)}, \tag{4.19}$$

and is also denoted by σ_X or σ.

The continuous analogue of Theorem 3.9 is given below.

Theorem 4.4. For every continuous random variable X,

$$\text{Var}(X) = E(X^2) - \mu^2. \tag{4.20}$$

Proof. The proof is exactly like the proof of Theorem 3.9, but based on Theorems 4.3 and 4.2, along with (4.20). □

Remark. The above theorem provides a formula for Var(X) that is computationally simpler than (4.18).

The continuous analogue of Theorem 3.10 is given below.

Theorem 4.5. For every continuous random variable X, if $a, b \in \mathbb{R}$, then

$$\text{Var}(aX + b) = a^2\text{Var}(X). \tag{4.21}$$

Proof. The proof is exactly like the proof of Theorem 3.10, but based on Theorems 4.3 and 4.2. □

Corollary 4.5.1. For every continuous random variable, if $a, b \in \mathbb{R}$, then

$$SD(aX + b) = |a|SD(X). \tag{4.22}$$

Proof. The proof is exactly like the proof of Corollary 3.10.1, but based on Theorem 4.5. □

We conclude this section with a continuous analogue of Chebyshev's Theorem (Theorem 3.15).

Theorem 4.6. Let X be a continuous random variable with mean μ and standard deviation $\sigma > 0$. Then, for all $h > 0$,

$$P(|X - \mu| \geq h\sigma) \leq \frac{1}{h^2}, \tag{4.23}$$

and so

$$P(|X - \mu| \leq h\sigma) \geq 1 - \frac{1}{h^2}. \tag{4.24}$$

Proof. By (4.18) and LOTUS - continuous case,

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx, \tag{4.25}$$

where f is the pdf of X. Breaking this improper integral into 3 parts yields

$$\sigma^2 = \int_{-\infty}^{\mu - h\sigma} (x - \mu)^2 f(x)dx + \int_{\mu - h\sigma}^{\mu + h\sigma} (x - \mu)^2 f(x)dx + \int_{\mu + h\sigma}^{\infty} (x - \mu)^2 f(x)dx \tag{4.26}$$

$$\geq \int_{-\infty}^{\mu-h\sigma} (x-\mu)^2 f(x)dx + \int_{\mu+h\sigma}^{\infty} (x-\mu)^2 f(x)dx,$$

since all of the integrals in question are nonnegative. If $x \leq \mu - h\sigma$, then $(x-\mu)^2 \geq h^2\sigma^2$ and if $x \geq \mu + h\sigma$, then $(x-\mu)^2 \geq h^2\sigma^2$, so it follows from (4.26) that

$$\sigma^2 \geq h^2\sigma^2 \left[\int_{-\infty}^{\mu-h\sigma} f(x)dx + \int_{\mu+h\sigma}^{\infty} f(x)dx\right] \tag{4.27}$$

i.e.,

$$\sigma^2 \geq h^2\sigma^2 P(|X-\mu| \geq h\sigma). \tag{4.28}$$

Dividing each side of (4.28) by $h^2\sigma^2$ yields the desired result. Finally, 4.24 follows from (4.23) since $P(|X-\mu| \leq h\sigma) = P(|X-\mu| < h\sigma) = 1 - P(|X-\mu| \geq h\sigma)$. □

4.4 Problems

1. The pdf f of X is given by

$$f(x) = \begin{cases} e^{-x} & \text{if } x \geq 0 \\ 0 & \text{elsewhere.} \end{cases}$$

   Find a. $P(X \leq 2)$ b. $P(3 < X \leq 5)$ c. $E(X)$ d. $\text{Var}(X)$

2. For X as in 1. above, find $P(X \leq 3)$ and compare it to the lower bound on this probability given by Chebyshev's Theorem.

3. The pdf of X, the lifetime of an electronic device (in hours), is given by

$$f(x) = \begin{cases} \frac{10}{x^2} & \text{if } x \geq 10 \\ 0 & \text{elsewhere.} \end{cases}$$

   a. Find $P(X > 20)$.

   b. Find the cdf of X.

   c. What is the probability that, of 6 such devices, at least 3 will function for at least 15 hours?

   d. Find $P(X \geq 5)$.

4. A filling station receives gasoline once a week. If its weekly sales in thousands of gallons is a random variable with pdf

$$f(x) = \begin{cases} 5(1-x)^4 & 0 \leq x \leq 1 \\ 0 & \text{elsewhere,} \end{cases}$$

   how much gas should the manager order each week so that the probability of running out of gas in a given week is 0.01?

5. Show that

$$f(x) = \begin{cases} \frac{1}{x^2}, & x \geq 1 \\ 0 & \text{elsewhere} \end{cases}$$

is a continuous pdf. Let X be the random variable whose pdf is f. Find $E(X)$.

4.5 The Uniform Distribution on an Interval

Let a and b be real numbers, with $a < b$. We say that X is *uniformly distributed on* (or "over") *the interval* (a, b), symbolized $X \sim \text{uniform}(a, b)$ if the pdf f of X is given by

$$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{elsewhere.} \end{cases} \tag{4.29}$$

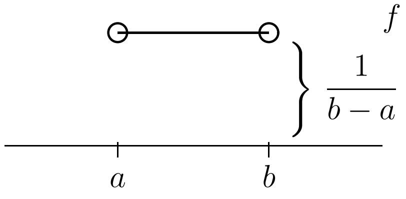
Such a random variable provides a model of the "experiment" of choosing a real number at random from the interval (a, b), with X recording the number that is selected. For any $B \subseteq \mathbb{R}$, the probability $P(X \in B)$ that the number selected belongs to the set B is given by formula

$$P(X \in B) = \int_B f(x)dx, \tag{4.30}$$

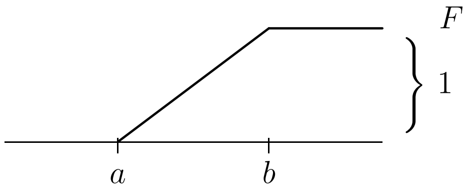
with f given by (4.29). In particular, for each $c \in \mathbb{R}$, $P(X = c) = 0$, as in the case of all continuous random variables.

Here is a graph of the pdf and cdf of X:

(4.31)



(4.32)



We have $F(x) = 0$ for $x \leq a$ and $F(x) = 1$ for $x \geq b$ and

$$F(x) = \int_a^x \frac{1}{b-a}dx = \frac{x-a}{b-a} \quad a < x < b.$$

The next result is one that could easily be guessed.

Theorem 4.7. If $X \sim \text{uniform}(a, b)$ then $E(X) = (a+b)/2$.

Proof. From (4.14) and (4.29), we have

$$\begin{aligned} E(X) &= \int_a^b x \frac{1}{b-a} dx = \frac{1}{b-a} \left[\frac{1}{2}x^2\right]_a^b \\ &= \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}. \qquad \square \end{aligned}$$

The next theorem gives a formula for the variance of a uniformly distributed random variable.

Theorem 4.8. If $X \sim \text{uniform}(a, b)$, then

$$\text{Var}(X) = \frac{(b-a)^2}{12}. \tag{4.33}$$

Proof. We use the formula $\text{Var}(X) = E(X^2) - \mu^2$, along with Theorem 4.7. By LOTUS - continuous case,

$$\begin{aligned} E(X^2) &= \int_a^b x^2 \frac{1}{b-a} dx = \frac{1}{3(b-a)} \quad x^3\Big|_a^b \\ &= \frac{b^3 - a^3}{3(b-a)}, \end{aligned}$$

and so

$$\text{Var}(X) = \frac{b^3 - a^3}{3(b-a)} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}. \qquad \square$$

Example. A policeman on patrol passes a certain intersection at random between noon and 2:30 p.m. What is the probability that he passes the corner after 1:30 p.m.? between 1:00 p.m. and 2:00 p.m.?

Solution. In such problems one needs to select a unit of measurement and decide how to label the left-hand end point of the interval in question. We'll represent noon by 0 on the x-axis and measure in terms of minutes. So X, the time at which the corner is passed, satisfies $X \sim \text{uniform}(0, 150)$. Then

$$f(x) = \begin{cases} \frac{1}{150} & 0 \le x \le 150 \\ 0 & \text{otherwise.} \end{cases}$$

So

$$P(X > 90) = \int_{90}^{150} \frac{1}{150} dx = \frac{60}{150} = 0.4$$

$$\downarrow$$

rep'n of 1:30 p.m.

and

$$P(60 < X < 120) = \int_{60}^{120} \frac{1}{150} dx = \frac{60}{150} = 0.4.$$

Notice that one doesn't really have to go through the elaborate business of setting up integrals here. If $X \sim \text{uniform}(a, b)$ and if $B \subseteq (a, b)$, then $P(X \in B)$ is just the sum of the lengths of the intervals comprising B, divided by $b - a$.

People sometimes make the mistake of thinking that if $X \sim \text{uniform}(a, b)$ and $Y = h(X)$, then Y is uniformly distributed over the interval of its possible values, but this is not generally true, as the following example shows.

Example. We will pick a real number from (0,1) at random and construct a cube with that number being the length of each side. If V denotes the volume of such a cube, find the pdf of V and $E(V)$.

Solution. Let X record the number selected from (0,1). Then $X \sim \text{uniform}(0, 1)$, so

$$f_X(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

and $V = X^3$. By LOTUS,

$$E(V) = E(X^3) = \int_0^1 x^3 \cdot 1 dx = \frac{1}{4}.$$

This shows us already that V is *not* uniformly distributed over (0,1), for if it were, the expected value would be $\frac{1}{2}$, not $\frac{1}{4}$. Let us find the pdf $f_V(v)$ of V. In all such problems, one first finds the cdf, and then differentiates. Now

$$(4.34) \qquad F_V(v) = P(V \leq v) = \begin{cases} 0 & \text{if } v \leq 0 \\ 1 & \text{if } v \geq 1 \end{cases}$$

and if $0 < v < 1$, then

$$(4.35) \qquad \begin{aligned} F_V(v) &= P(V \leq v) = P(X^3 \leq v) = P(X \leq v^{\frac{1}{3}}) \\ &= \int_0^{v^{\frac{1}{3}}} 1 dx = v^{\frac{1}{3}}. \end{aligned}$$

From (4.35) and (4.34) we get

$$(4.36) \qquad f_V(v) = \begin{cases} \frac{d}{dv} v^{\frac{1}{3}} = \frac{1}{3} v^{-\frac{2}{3}} & \text{if } 0 < v < 1 \\ 0 & \text{otherwise} \end{cases}$$

Note that, as an alternative to using LOTUS, as we did above to find $E(V)$, we can use (4.36):

$$E(V) = \int_0^1 v \cdot \frac{1}{3} v^{-\frac{2}{3}} dv = \frac{1}{4} \left[v^{\frac{4}{3}} \right]_0^1 = \frac{1}{4}.$$

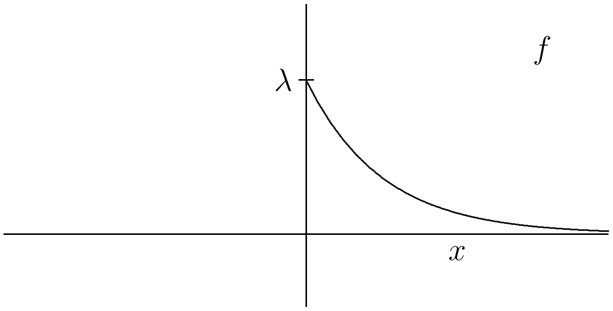
It is, of course, easier to use LOTUS, since we don't have to find the pdf of V that way.

Remark. The above example shows that one needs to be careful to avoid ambiguity in the phrasing of "random geometry" problems. Suppose someone asks you to determine the expected volume of a "randomly chosen cube" under the restriction that the length of each side be less than 1. If one gets a "randomly chosen cube" by selecting its side length at random from (0,1), then the answer, as we showed above, is $\frac{1}{4}$. But if one gets a "randomly chosen cube" by selecting its volume at random from (0,1), one would get $\frac{1}{2}$ as the expected volume. The moral of the story is that one should avoid vague phrases like "randomly chosen cube" and specify precisely what characteristic of the object in question is to be selected randomly.

4.6 Exponential Distributions

If $\lambda > 0$, the random variable X is said to be *exponentially distributed with parameter* λ (abbreviated $X \sim \text{exponential}(\lambda)$) if the pdf f of X is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.37}$$



It is easy to check that f, as given by (4.37), is a continuous density function, for

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)dx &= \int_0^{\infty} \lambda e^{-\lambda x} dx = \lim_{b \to \infty} \int_0^b \lambda e^{-\lambda x} dx \\ &= \lim_{b \to \infty} \left[-e^{-\lambda x} \right]_0^b = \lim_{b \to \infty} 1 - e^{-\lambda b} = 1. \end{aligned}$$

Theorem 4.9. If $X \sim \text{exponential}(\lambda)$, then

$$E(X) = \frac{1}{\lambda}, \quad \text{and} \tag{4.38}$$

$$\mathrm{Var}(X) = \frac{1}{\lambda^2}. \tag{4.39}$$

Proof. Using integration by parts, we get

$$\begin{aligned}\int \lambda e^{-\lambda x} x dx &= -e^{-\lambda x}x + \int e^{-\lambda x} dx \\ &= -e^{-\lambda x}x - \frac{1}{\lambda}e^{-\lambda x} = -e^{-\lambda x}\left(x + \frac{1}{\lambda}\right)\end{aligned} \tag{4.40}$$

and

$$\begin{aligned}\int \lambda e^{-\lambda x} x^2 dx &= -e^{-\lambda x}x^2 + \int 2e^{-\lambda x} x dx \\ &= -e^{-\lambda x}x^2 + \frac{2}{\lambda}\int \lambda e^{-\lambda x} x dx \\ &= -e^{-\lambda x}x^2 - \frac{2}{\lambda}e^{-\lambda x}\left(x + \frac{1}{\lambda}\right) \\ &= -e^{-\lambda x}\left(x^2 + \frac{2}{\lambda}x + \frac{2}{\lambda^2}\right).\end{aligned} \tag{4.41}$$

Hence by (4.40)

$$\begin{aligned}E(X) &= \int_0^\infty \lambda e^{-\lambda x} x dx \\ &= \lim_{b\to\infty}\left[-e^{-\lambda x}\left(x + \frac{1}{\lambda}\right)\right]_0^b \\ &= \lim_{b\to\infty}\left(\frac{1}{\lambda} - \frac{b + \frac{1}{\lambda}}{e^{\lambda b}}\right) \\ &= \frac{1}{\lambda},\end{aligned}$$

by de l'Hôpital's rule. Also, by (4.41)

$$\begin{aligned}\mathrm{Var}(X) &= E(X^2) - (E(X))^2 \\ &= \int_0^\infty \lambda^{-\lambda x} x^2 dx - \frac{1}{\lambda^2} \\ &= \lim_{b\to\infty}\left[-e^{-\lambda x}\left(x^2 + \frac{2}{\lambda}x + \frac{2}{\lambda^2}\right)\right]_0^b - \frac{1}{\lambda^2} \\ &= \lim_{b\to\infty}\left(\frac{2}{\lambda^2} - \frac{b^2 + \frac{2}{\lambda}b + \frac{2}{\lambda^2}}{e^{\lambda b}}\right) - \frac{1}{\lambda^2} \\ &= \frac{2}{\lambda^2} - \frac{1}{\lambda^2}\end{aligned}$$

$$= \frac{1}{\lambda^2},$$

again by de l'Hôpital's rule. □

Exponential distributions often arise as models of "waiting times," or "lifetimes." For example, starting from a fixed point in time, the amount of time that elapses until the next earthquake, or the next war, or the next wrong number are all random variables which tend to have exponential distributions.

Example. The length of a phone call in minutes is exponentially distributed, averaging 10 minutes. If someone arrives immediately ahead of you at a public phone, what is the probability that you must wait more than 10 minutes to use the phone?

Solution. If X records the length of this phone call (= your waiting time), then $X \sim$ exponential(1/10), since $E(X) = 10 = 1/\lambda$. So

$$P(X > 10) = \int_{10}^{\infty} \frac{1}{10} e^{-\frac{x}{10}} dx = \lim_{b\to\infty} \left[-e^{-\frac{x}{10}}\right]_{10}^{b} = e^{-1} \approx .368.$$

Exponential distributions have a peculiar property, known as the "memoryless property," described by the following theorem.

Theorem 4.10. If $X \sim$ exponential(λ), then for all $a, b \geq 0$

$$P(X > a + b \,|\, X > a) = P(X > b). \tag{4.42}$$

Proof. If $X \sim$ exponential(λ), then $F(x) = 0$ if $x < 0$, and if $x \geq 0$, then

$$\begin{aligned} F(x) = P(X \leq x) &= \int_0^x \lambda e^{-\lambda t} dt = \left[-e^{-\lambda t}\right]_0^x \qquad (4.43) \\ &= 1 - e^{-\lambda x}, \quad \text{and so} \\ P(X > x) = 1 - P(X \leq x) &= e^{-\lambda x}. \qquad (4.44) \end{aligned}$$

By (4.44),

$$\begin{aligned} P(X > a + b \,|\, X > a) &= \frac{e^{-\lambda(a+b)}}{e^{-\lambda a}} \\ &= e^{-\lambda b} = P(X > b). \qquad \square \end{aligned}$$

If X records, for example, the lifetime of some instrument, and X is exponentially distributed, then (4.42) says that the probability that the instrument survives for at least an additional b hours, given that it has already survived for at least a hours, is the same as the initial probability that it survives for at least b hours. In short, the instrument does not "remember" that it has already survived for a hours.

Example (cf. the example at the top of this page). Suppose that the length of a phone call is exponentially distributed, averaging 10 minutes. You arrive at a phone booth and find it occupied. What is the probability you must wait more than 10 minutes to use the phone?

Solution. Let X be the duration of the phone call being made by the individual occupying the booth when you arrive. Suppose he has already been talking for a minutes when you arrive. You need to calculate $P(X > a + 10|X > a)$. By Theorem 4.10, that's just the same as $P(X > 10) = e^{-1} \approx .368$.

Remark. There is an interesting connection between Poisson and exponential distributions. Suppose that the number of events of a given type occurring within a given unit interval of time has a Poisson distribution with mean λ. Let X denote the waiting time until the first such "Poisson event" occurs. Then $X \sim \text{exponential}(\lambda)$ by the following argument: Let F be the cdf of X. Clearly, $F(x) = 0$ if $x < 0$. If $x \geq 0$, then

$$F(x) = P(X \leq x) = 1 - P(X > x). \tag{4.45}$$

Now $P(X > x) =$ the probability that no "Poisson event" occurs in the interval $[0, x]$. If the number of Poisson events in a unit interval has mean λ, the number of Poisson events in $[0, x]$ has mean λx, so the probability that no Poisson event occurs in $[0, x]$ equals $e^{-\lambda x}$. By (4.45)

$$F(x) = 1 - e^{-\lambda x} \quad \text{if } x \geq 0. \tag{4.46}$$

So

$$f(x) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \geq 0 \end{cases} \tag{4.47}$$

i.e. $X \sim \text{exponential}(\lambda)$.

Remark. Exponential distributions are, in a sense, continuous analogues of geometric distributions. In the latter case, we record the number of trials elapsed until the first success is observed. In the former case, we record the time elapsed until the occurrence of the first "Poisson event."

4.7 Problems

1. Prove that if $X \sim \text{geometric}(p)$ and $a, b \in \mathbb{N}$, then

$$P(X > a + b|X > a) = P(X > b),$$

i.e., that geometric distributions are memoryless.

2. The lifetime of a light bulb is exponentially distributed with a mean of 3000 hours. What is the probability that the bulb lasts less than 300 hours?

3. If X is any random variable, a real number m is called a median of X if $P(X < m) = P(X > m)$. If X is continuous, m is a median of X if $\int_{-\infty}^{m} f(x)dx = 1/2$. A random variable may have more than one median. Show that the random variable X recording the light bulb lifetime in problem 2 above has a median of approximately 2079 hours. Why would the manufacturer of this light bulb advertise its average (i.e., mean) lifetime rather than its median lifetime?

4. If $X \sim$ exponential(λ), find $P(X > E(X))$. Show why your result implies that the mean of an exponentially distributed random variable is always greater than its median.

5. Suppose $X \sim$ uniform$(0, 120)$. Find the exact value of $P(15 \le X \le 105)$ and, using Chebyshev's Theorem, a lower bound on this probability.

6. Suppose that a piece of string is 8 inches long. A child playing with scissors cuts the string in 2 pieces, cutting at a random position on the string.

   a) What is the probability that the longer piece is at least 6 inches long?

   b) What is the probability that the shorter piece is at least 3 inches long?

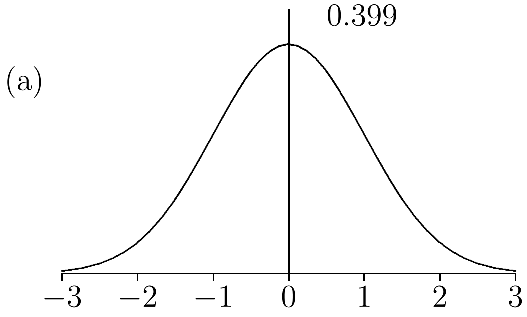
   c) What is the probability that the longer piece is at least twice as long as the shorter piece?

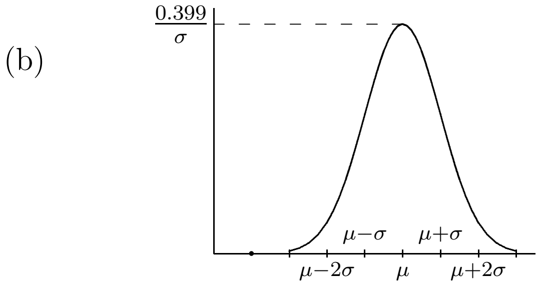
4.8 Normal Distributions

A random variable X is said to be *normally distributed with parameters μ and σ^2* (abbreviated $X \sim$ normal(μ, σ^2)) if the pdf f of X is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty. \tag{4.48}$$

Graphs of f are shown below for the case $\mu = 0$ and $\sigma = 1$, and in the general case.





The normal density function: (a) with $\mu = 0$, $\sigma = 1$; and (b) arbitrary μ, σ^2

The normal distribution was introduced in 1733 by deMoivre in order to approximate probabilities associated with binomial random variables (see §4.9 below for details), but it has turned out to have much broader applications than that, for many random phenomena may be modeled by normal distributions. The following theorem shows why we have labeled the parameters of a normal distribution as we have.

Theorem 4.11. If $X \sim \text{normal}(\mu, \sigma^2)$, then $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$.

Proof. Writing x as $(x - \mu) + \mu$ yields

$$E(X) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-\mu) e^{-(x-\mu)^2/2\sigma^2} dx + \mu \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2} dx$$

Letting $y = x - \mu$ in the first integral yields

$$E(X) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} y e^{-y^2/2\sigma^2} dy + \mu \int_{-\infty}^{\infty} f(x) dx$$

with f the pdf of X. Since the integrand of the first integral is an odd function (g is *odd* if $g(-y) = -g(y)$ for all y), the first integral equals 0, and so

$$E(X) = \mu \int_{-\infty}^{\infty} f(x) dx = \mu \cdot 1 = \mu.$$

We have assumed without proof here that $\int_{-\infty}^{\infty} f(x) dx = 1$, where f is given by (4.48).

To show that $\text{Var}(X) = \sigma^2$, we use the formula

$$\begin{aligned} \text{Var}(X) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-(x-\mu)^2/2\sigma^2} dx \\ (y = \tfrac{x-\mu}{\sigma}) &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-y^2/2} dy. \end{aligned}$$

Now integration-by-parts yields

$$\int y^2 e^{-y^2/2} dy = \int y e^{-y^2/2} y \, dy$$

$$= -e^{-y^2/2} \cdot y + \int e^{-y^2/2} dy,$$

and so

$$\operatorname{Var}(X) = -\left[\frac{\sigma^2}{\sqrt{2\pi}} y e^{-y^2/2}\right]_{-\infty}^{\infty} + \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy$$

$$= 0 + \sigma^2 \cdot 1 = \sigma^2. \qquad \square$$

If $X \sim \text{normal}(\mu, \sigma^2)$, the cdf F of X is given of course by the formula

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{(t-\mu)^2/2\sigma^2} dt. \tag{4.49}$$

Unfortunately, this antiderivative of f cannot be expressed in terms of any of the familiar functions of analysis. Values of $F(x)$ need to be calculated by numerical approximation methods. Fortunately, it is not necessary to have tables of $F(x)$ for each pair of parameters μ and σ^2. By the following theorem, a table of $F(x)$ when $X \sim \text{normal}(0, 1)$ suffices.

Theorem 4.12. If $X \sim \text{normal}(\mu, \sigma^2)$ and $Z := (X - \mu)/\sigma$, then $Z \sim \text{normal}(0, 1)$.

Proof. Let F denote the cdf of Z and f its pdf. Now for all $z \in \mathbb{R}$

$$\begin{aligned} F(z) = P(Z \le z) &= P\left(\frac{X-\mu}{\sigma} \le z\right) \\ &= P(X \le z\sigma + \mu) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{z\sigma+\mu} e^{-(x-\mu)^2/2\sigma^2} dx \\ \left(t = \tfrac{x-\mu}{\sigma}\right) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt. \end{aligned} \tag{4.50}$$

Differentiating (4.50) with respect to z yields

$$f(z) = F'(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad \forall z \in \mathbb{R}, \tag{4.51}$$

so clearly $Z \sim \text{normal}(0, 1)$. By the way, Z is referred to as *the standard normal random variable*, and its cdf is almost always denoted by Φ rather than by F_Z or by F. There are several ways to find approximate values of $\Phi(z)$.

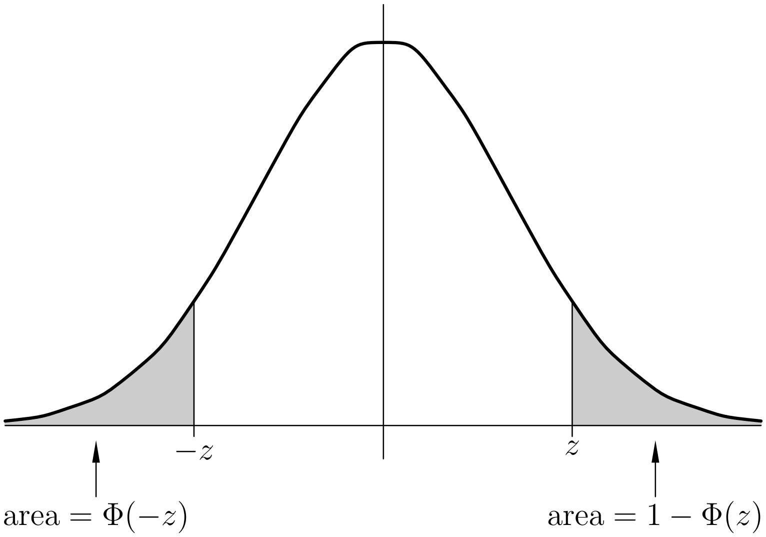
I. You may wish to program an approximate formula for $\Phi(z)$ on your calculator. The following approximation formula has, for every $z \ge 0$, an absolute error of approximation less than 2.5×10^{-4}

$$\Phi(z) \approx 1 - \frac{1}{2}\left(1 + c_1 z + c_2 z^2 + c_3 z^3 + c_4 z^4\right)^{-4} \tag{4.52}$$

for all $z \geq 0$, where $c_1 = 0.196854$, $c_2 = 0.115194$, $c_3 = 0.000344$, $c_4 = 0.019527$. To evaluate Φ for negative values, use the formula

$$\Phi(-z) = 1 - \Phi(z), \quad \forall z > 0, \tag{4.53}$$

which is intuitively clear based on the picture below:



II. Go to `http://math.uc.edu/statistics/statbook/tables.html`, select table 1 (Standard Normal Distribution) and the option “Left tail,” and then enter your z value (positive or negative) and click on the arrow pointing to the right. An approximation of $\Phi(z)$ will appear in the box labeled “probability.”

III. Use the following table of approximate values of $\Phi(z) = P(Z \leq z)$ for $z \geq 0$, supplemented by formula (4.53) above:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

If $Z \sim \text{normal}(0,1)$, then, as we have noted,

$$P(Z \le b) = \Phi(b), \qquad \text{and so} \tag{4.54}$$

$$P(Z > b) = 1 - \Phi(b), \quad \forall b \in \mathbb{R}. \tag{4.55}$$

Furthermore,

$$P(a \le Z \le b) = \Phi(b) - \Phi(a), \quad \forall a, b \in \mathbb{R} \text{ with } a \le b. \tag{4.56}$$

Suppose $X \sim \text{normal}(\mu, \sigma^2)$ and $h > 0$. By Chebyshev's Theorem, $P(|X - \mu| \leq h\sigma) \geq 1 - \frac{1}{h^2}$. The exact probability is given by

$$\begin{aligned}
P(|X-\mu| \leq h\sigma) &= P\left(\left|\frac{X-\mu}{\sigma}\right| \leq h\right) && (4.57)\\
&= P(|Z| \leq h)\\
&= P(-h \leq Z \leq h) && (4.58)\\
&= \Phi(h) - \Phi(-h), \quad \text{by Theorem 4.12.} && (4.59)
\end{aligned}$$

These values are tabulated below for $h = 1, 2, 3$, along with the Chebyshev estimates:

	normal table	Chebyshev
$P(\lvert X-\mu\rvert \leq \sigma) = P(\lvert Z\rvert \leq 1)$	0.6826("68%")	≥ 0
$P(\lvert X-\mu\rvert \leq 2\sigma) = P(\lvert Z\rvert \leq 2)$	0.9544("95%")	$\geq \frac{3}{4}$
$P(\lvert X-\mu\rvert \leq 3\sigma) = P(\lvert Z\rvert \leq 3)$	0.9974("99%")	$\geq \frac{8}{9}$

Example. The length of a human pregnancy (in days) is approximately normally distributed with $\mu = 270$ and $\sigma = 10$. Approximately what percentage of pregnancies last more than 290 days?

Solution: With $X \sim \text{normal}(270, 100)$, $P(X > 290) = P\left(\frac{X-270}{10} > \frac{290-270}{10}\right) = P(Z > 2) = 1 - \Phi(2) = 0.0228$. So 2.28% of human pregnancies last more than 290 days.

Remark. Given any continuous random variable X and any q with $0 < q < 100$, the "qth percentile" of the distribution in question is that value x_q of X satisfying $P(X < x_q) = \frac{q}{100}$. The 25th percentile is often called the "lower quartile," the 50th percentile the "median," and the 75th percentile the "upper quartile."

Example. SAT scores are normally distributed with mean 500 and standard deviation 100. a.) What percentile corresponds to a score of 650? 300? b.) What score is required to place a student at the 90th percentile?

Solution: a) With $X \sim \text{normal}(500, 100^2)$ $P(X < 650) = P\left(\frac{X-500}{100} < \frac{650-500}{100}\right) = P(Z < 1.5) = 0.9332$. So a score of 650 places a student at the 93.32th percentile. Similarly, $P(X < 300) = P\left(\frac{X-500}{100} < \frac{300-500}{100}\right) = P(Z < -2.0) = 0.0228$. So a score of 300 places a student at the 2.28th percentile.

b) Find x_{90} satisfying $P(X < x_{90}) = 0.90$. Now,

$$P(X < x_{90}) = P\left(\frac{X-500}{100} < \frac{x_{90}-500}{100}\right) = P\left(Z < \frac{x_{90}-500}{100}\right) = 0.90$$

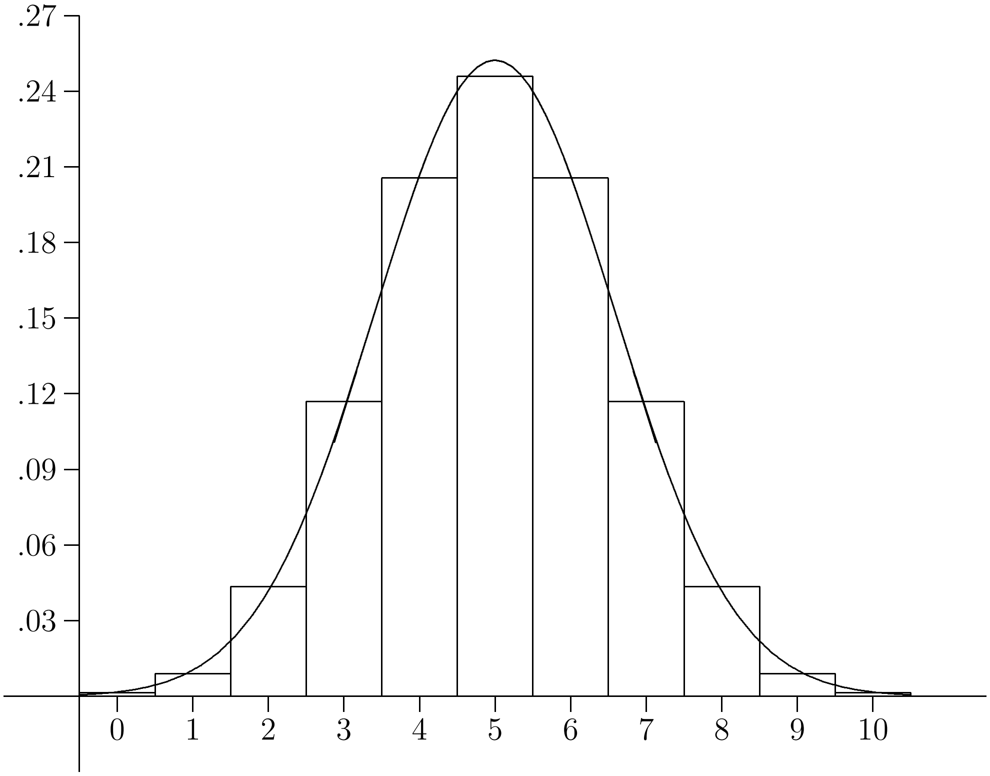
when $\frac{x_{90}-500}{100} \approx 1.28$, i.e., when $x_{90} = 500 + (1.28)(100) = 628$. Here, of course, we have read the normal table "backwards," i.e., we looked in the interior of the table for the closest probability to 0.90 and then read off the value of z corresponding to that probability. Alternatively, we could have gone to the web address listed above in II, chosen the option "Left tail," entered 0.90 in the box labeled "probability," and clicked on the arrow pointing to the left. The 90th percentile of Z (i.e., the value z for which $P(Z < z) = 0.90$ would then have appeared in the box labeled "z value."

4.9 The de Moivre-Laplace Theorem

Let $X \sim \text{binomial}(n, p)$. The main theorem of this section assures us that, for n sufficiently large, areas under the histogram of the binomial random variable X may be approximated by areas under the pdf of a normal random variable Y having the same mean and variance as X.

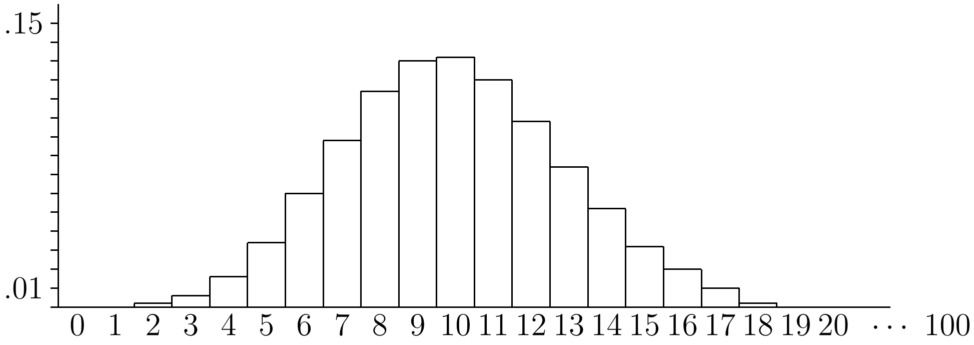
When $p = 1/2$, this claim is already quite plausible when $n = 10$. As the illustration below indicates, the histogram of the binomial random variable X with these parameters is remarkably well approximated by the pdf of the normal random variable Y with $E(Y) = E(X) = 5$ and $\text{Var}(Y) = \text{Var}(X) = 2.5$.

(4.60)



If $p \neq 1/2$ and $X \sim \text{binomial}(n, p)$, then the histogram of X is asymmetrical, notably so in certain cases where p is close to 0 or to 1, as can be seen in Examples 1, 2, and 3 in section 3.7 above. Such notably asymmetrical histograms are of course poorly approximated by normal pdfs, which are symmetrical about their means. Nevertheless, for a *fixed* p, no matter how close to 0 or 1, if n is sufficiently large and $X \sim \text{binomial}(n, p)$, the histogram if X will be only mildly asymmetrical, and well approximated by the pdf of a normal random variable Y with $E(Y) = np$ and $\text{Var}(Y) = np(1 - p)$. This is illustrated below for $X \sim \text{binomial}(100, 1/10)$.

(4.61)



In Remark 4 below, we describe the binomial distributions which are well approximated by normal distributions.

We asserted above that, *for n sufficiently large, if $X \sim$ binomial(n,p), then the distribution of X is approximately that of a normal random variable Y having the same mean and variance as X*, i.e., with $\mu_Y = \mu_X = np$ and $\sigma_Y^2 = \sigma_X^2 = np(1-p)$. Now this is the case if and only if $(X-np)/\sqrt{np(1-p)}$ has approximately the distribution of $(Y-np)/\sqrt{np(1-p)}$. But by Theorem 4.12, if $Y \sim$ normal$(np, np(1-p))$ then

$$\frac{Y-np}{\sqrt{np(1-p)}} = Z, \tag{4.62}$$

the standard normal random variable. Thus to establish the italicized assertion above, it suffices to prove the following theorem.

Theorem 4.13. Let $X \sim$ binomial(n,p). Then for all $z \in \mathbb{R}$,

$$\lim_{n\to\infty} P\left(\frac{X-np}{\sqrt{np(1-p)}} \le z\right) = \Phi(z), \tag{4.63}$$

when Φ is the cdf of the standard normal random variable Z.

Proof. We omit the proof. □

Corollary 4.13.1. If $X \sim$ binomial(n,p), then for all $\alpha, \beta \in \mathbb{R}$ such that $\alpha < \beta$,

$$P\left(\alpha \le \frac{X-np}{\sqrt{np(1-p)}} \le \beta\right) \approx \Phi(\beta) - \Phi(\alpha), \tag{4.64a}$$

for n sufficiently large.

Proof. Exercise. □

Corollary 4.13.2. Given $X \sim \text{binomial}(n,p)$ for n sufficiently large, and a and b integers with $0 \le a < b \le n$,

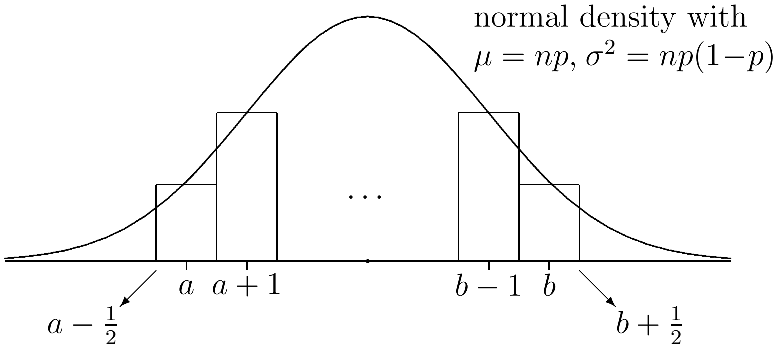
$$P(a \le X \le b) \approx \Phi\left(\frac{b-np}{\sqrt{npq}}\right) - \Phi\left(\frac{a-np}{\sqrt{npq}}\right), \text{ where } q := 1-p. \tag{4.64b}$$

Proof. A rigorous proof of this result is beyond the scope of these notes. An informal argument goes as follows: Since $a \le X \le b$ if and only if $(a-np)/\sqrt{npq} \le (X-np)/\sqrt{npq} \le (b-np)\sqrt{npq}$, and $(X-np)/\sqrt{npq}$ "behaves like" Z for large n,

$$\begin{aligned} P(a \le X \le b) &= P\left(\frac{a-np}{\sqrt{npq}} \le \frac{X-np}{\sqrt{npq}} \le \frac{b-np}{\sqrt{npq}}\right) \\ &\approx P\left(\frac{a-np}{\sqrt{npq}} \le Z \le \frac{b-np}{\sqrt{npq}}\right) = \Phi\left(\frac{b-np}{\sqrt{npq}}\right) - \Phi\left(\frac{a-np}{\sqrt{npq}}\right). \end{aligned}$$

Remark 1. Theorem 4.13 is called the *deMoivre-Laplace Theorem.* It was established for $p = 1/2$ by deMoivre and later for arbitrary p with $0 < p < 1$ by Laplace. This theorem is a special case of a much more general result known as the *Central Limit Theorem.*

Remark 2. "The continuity correction": For very large n, the approximation formula (4.64b) works fine, but for "moderate" n, we get a little better approximation using the so-called continuity correction (you might as well use this all the time). The reasoning behind this is that we are approximating a discrete distribution by a continuous one and it is the *histogram* of the pdf of the binomial distribution that is approximated by a normal pdf:



Note that $P(a \le X \le b)$ is the sum of the areas of rectangles in the histogram that run from $a - 1/2$ to $b + 1/2$, so it is appropriate to take the area under the normal curve from $a - 1/2$ to $b + 1/2$ as our approximation of $P(a \le X \le b)$. The practical approximation formulas, with continuity correction are thus as follows: If $X \sim \text{binomial}(n,p)$, with n sufficiently large, and a and b are integers with $0 \le a \le b \le n$, then

$$P(a \le X \le b) \approx \Phi\left(\frac{b+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-\frac{1}{2}-np}{\sqrt{np(1-p)}}\right), \tag{4.65}$$

which has as the special case when $a = b$,

$$P(X = a) \approx \Phi\left(\frac{a + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right). \tag{4.66}$$

Also,

$$P(X \leq b) \approx \Phi\left(\frac{b + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right), \text{ and} \tag{4.67}$$

$$P(X \geq a) = 1 - P(X \leq a - 1) \approx 1 - \Phi\left(\frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right). \tag{4.68}$$

Remark 3. When is n "sufficiently large" to use the normal approximation? It is appropriate to use the approximation formulas (4.65)-(4.68) for $X \sim \text{binomial}(n, p)$ whenever $\sigma = \sqrt{np(1-p)} \geq 3$, i.e., when $n \geq 9/p(1-p)$. Note that as p gets close to 0 or 1, n must be correspondingly larger:

(4.69)

p	minimum n
0.5	36
0.4 or 0.6	38
0.3 or 0.7	43
0.2 or 0.8	57
0.1 or 0.9	100
0.05 or 0.95	190
0.01 or 0.99	910
0.005 or 0.995	1810
0.001 or 0.999	9010

When p is close to $\frac{1}{2}$, this criterion is a little stringent. In fact, when $p = \frac{1}{2}$ using (4.65)-(4.68) for $n \geq 10$ results in an error less than 0.01, and for $n \geq 20$ in an error less than 0.005.

Of course, if p is very small or very large and n is not large enough to make $\sigma \geq 3$, we can use the Poisson approximation to the binomial.

Example. If a fair coin is tossed 25 times, what is the approximate probability of getting exactly 12 heads? 18 or more heads?

Solution. We'll use the normal approximation based on the remarks following (4.69). Let X = the number of heads in 25 tosses. Then $X \sim \text{binomial}(25, 0.5)$ and $\mu = 12.5$ and $\sigma = 2.5$. Using (4.66) and (4.68) we have

$$\begin{aligned} P(X = 12) &\approx \Phi\left(\frac{12 + \frac{1}{2} - 12.5}{2.5}\right) - \Phi\left(\frac{12 - \frac{1}{2} - 12.5}{2.5}\right) \\ &= \Phi(0) - \Phi(-0.4) = 0.5000 - 0.3446 = 0.1554. \end{aligned}$$

This may be compared to the exact result,

$$P(X = 12) = \binom{25}{12}(.5)^{25} \approx 0.1550.$$

On the other hand,

$$P(X \geq 18) \approx 1 - \Phi\left(\frac{18 - \frac{1}{2} - 12.5}{2.5}\right) = 1 - \Phi(2) = 1 - 0.9772 = 0.0228,$$

which may be compared to the exact result $P(X \geq 18) = \sum_{k=18}^{25} \binom{25}{k}(.5)^{25} \approx 0.0220$.

Example. Let $X \sim \text{binomial}(1000, 0.01)$. Estimate $P(X \leq 3)$.

Solution. Since $\sigma = \sqrt{(1000)(.01)(.99)} \approx 3.15 > 3$, we may use the normal approximation. We have

$$P(X \leq 3) \cong \Phi\left(\frac{3 + \frac{1}{2} - 10}{\sqrt{9.9}}\right) \approx \Phi(-2.07) = 0.0192.$$

As p is small, a Poisson approximation would also be appropriate, with $\lambda = 10$,

$$P(X \leq 3) \approx e^{-10}\left(\frac{10^0}{0!} + \frac{10^1}{1!} + \frac{10^2}{2!} + \frac{10^3}{3!}\right) = 0.0103$$

which is very close to the normal approximation.

Of course, one may use the normal approximation to the binomial in testing binomial hypotheses when $\sqrt{np(1-p)} \geq 3$.

Example. An allegedly fair coin is tossed 100 times and comes up heads 62 times. Test the hypothesis of fairness against the alternative that the coin is biased in favor of heads at the 1% significance level.

Solution. Let X record the number of heads in 100 tosses under the hypothesis that the coin is fair. Then $X \sim \text{binomial}(100, \frac{1}{2})$, and under this hypothesis we need to find $P(X \geq 62)$. Since $\sigma = 5 \geq 3$, we may use the normal approximation

$$P(X \geq 62) \approx 1 - \Phi\left(\frac{61.5 - 50}{5}\right) = 1 - \Phi(2.3) = 1 - 0.9893 = 0.0107 > 0.01.$$

This outcome is not sufficiently improbable to reject the hypothesis of fairness at the 1% significance level.

Example. The army claims that at most 10% of its recruits lack a high school diploma. a) Would you reject this claim at the 5% significance level if, in a random sample of 625 recruits, 75 lacked a high school diploma? b) In such a sample, what is the minimum number of non-high school graduates that would cause you to reject the army's claim at the 1% significance level?

Solution to a). Although we are sampling without replacement, we can use the binomial distribution here. If X records the number of non-high school graduates (out of 625 recruits), the army's claim implies $X \sim \text{binomial}(625, p)$ where $p \leq 0.1$. Since we must specify p exactly, we'll test the weakest version of their claim, namely, $p = 0.1$. If $X \sim \text{binomial}(625, 0.1)$, $\mu = 62.5$ and $\sigma = \sqrt{(625)(.1)(.9)} = 7.5 > 3$, so the normal approximation is appropriate. Since large values of X undermine the claim, we calculate

$$P(X \geq 75) \approx 1 - \Phi\left(\frac{74.5 - 62.5}{7.5}\right) = 1 - \Phi(1.6) = 1 - 0.9452 = 0.0548,$$

which is not sufficient to reject the army's claim at the 5% level.

Solution to b). Find the smallest integer k such that

$$P(X \geq k) \approx 1 - \Phi\left(\frac{k - \frac{1}{2} - 62.5}{7.5}\right) < 0.01,$$

i.e. such that

$$\Phi\left(\frac{k - \frac{1}{2} - 62.5}{7.5}\right) > 0.99.$$

Reading the table on p. 107 "inside-out" we see that $\Phi(2.33) = 0.9901$ so set $(k - \frac{1}{2} - 62.5)/7.5 = 2.33$, solve for k and round up to the nearest integer: $k = \lceil 80.475 \rceil = 81$.

4.10 Problems

1. If the heights of UTK students are normally distributed with mean 68.2" and standard deviation 4", what heights correspond to the lower quartile, the median, and the upper quartile?

2. Someone is indifferent between betting that a fair die tossed 72 times will come up "1" 15 or more times and betting that a fair die tossed 144 times will come up 30 or more times "because each of these events represents a 25% deviation from the expected number of "1"'s." Enlighten this individual.

3. What is the approximate probability that a fair coin tossed 400 times comes up HEADS exactly 210 times?

4. A coin tossed 10,000 times comes up HEADS 5067 times. Test the hypothesis that the coin is fair at the 5% significance level, against a) the alternative that the coin is biased in favor of heads and b) the alternative that the coin is biased in an unspecified way.

5. The glorious leader of the democratic peoples' socialist republic of Yak asserts that 90% or more of its citizens are literate. In a random sample of 900 Yaks, what is the minimum number of illiterates that would cause you to reject this claim at the 5% level?

6. a. Let X be any random variable (discrete or continuous) with $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$. Prove that $E\left(\frac{X-\mu}{\sigma}\right) = 0$ and $\text{Var}\left(\frac{X-\mu}{\sigma}\right) = 1$.

   b. In the above situation, will it be true that $P\left(\frac{X-\mu}{\sigma} \leq z\right) \approx \Phi(z)$?

4.11 Estimating a Binomial Parameter p

Let $X \sim \text{binomial}(n,p)$ and let $\bar{X} := X/n$, the fraction of successes in n independent trials, in each of which the probability of success is p. Using the fact that

$$(4.70) \qquad E(\bar{X}) = p \quad \text{and} \quad SD(\bar{X}) = \sqrt{\frac{p(1-p)}{n}},$$

we proved (Theorem 3.16) that, for each fixed $d > 0$,

$$(4.71) \qquad P(|\bar{X} - p| \leq d) \geq 1 - \frac{p(1-p)}{nd^2} \geq 1 - \frac{1}{4nd^2},$$

which we then used to prove that $\lim_{n\to\infty} P(|\bar{X} - p| \leq d) = 1$ (Bernoulli's Law of Large Numbers.)

We also used (4.71) to give (crude) solutions to three types of problems connected with an inequality of the form

$$(4.72) \qquad P(|\bar{X} - p| \leq d) \geq c,$$

namely,

1° given p, n, and d, find c (as large as possible)

2° given p, d, and c, find n (as small as possible)

3° given p, n, and c, find d (as small as possible)

so that (4.72) is satisfied. In what follows, we will see that we can get much more refined answers to 1°-3° using a certain normal approximation.

This approximation is based on the following corollary of the deMoivre-Laplace Theorem (Theorem 4.13):

Corollary 4.13.2. If $X \sim \text{binomial}(n,p)$ and $\bar{X} := X/n$, then for all $z \in \mathbb{R}$,

$$(4.73) \qquad \lim_{n\to\infty} P\left(\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z\right) = \Phi(z).$$

Proof. Use Theorem 4.13 and the fact that (dividing numerator and denominator of the LHS of (4.74) by n)

$$\frac{X - np}{\sqrt{np(1-p)}} = \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}}. \qquad \square \tag{4.74}$$

The practical effect of Corollary 4.13.2 is that the linear transformation of $\bar{X}$,

$$\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \approx Z, \tag{4.75}$$

the standard normal random variable. Here are some typical applications of (4.75). Instead of using the crude, Chebyshev-based result (4.71), we now make use of the more refined result below:

Corollary 4.13.3. With X and $\bar{X}$ as in Corollary 4.13.2 and $d > 0$,

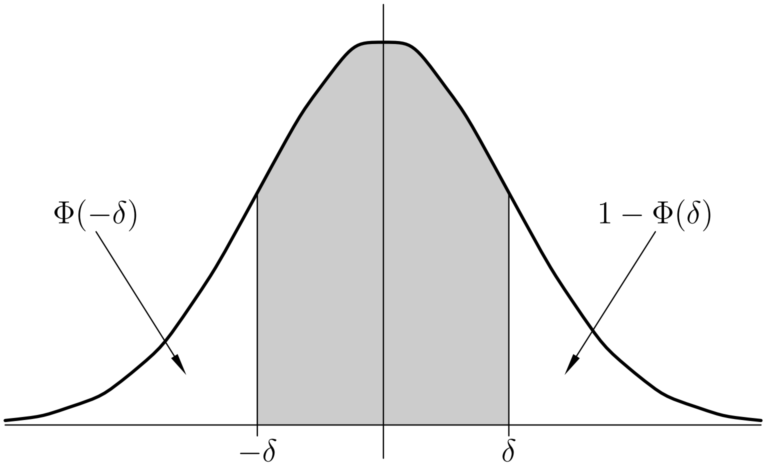
$$P(|\bar{X} - p| \leq d) \approx 2\Phi\left(\frac{d}{\sqrt{\frac{p(1-p)}{n}}}\right) - 1 \tag{4.76}$$

Proof.

$$\begin{aligned} P(|\bar{X} - p| \leq d) &= P\left(\left|\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}}\right| \leq \frac{d}{\sqrt{\frac{p(1-p)}{n}}}\right) \\ &\approx P\left(|Z| \leq \frac{d}{\sqrt{\frac{p(1-p)}{n}}}\right) \\ &= 2\Phi\left(\frac{d}{\sqrt{\frac{p(1-p)}{n}}}\right) - 1, \end{aligned}$$

since, for any $\delta > 0$,

$$\begin{aligned} P(|Z| \leq \delta) &= P(-\delta \leq Z \leq \delta) \\ &= \Phi(\delta) - \Phi(-\delta) \\ &= \Phi(\delta) - (1 - \Phi(\delta)) \\ &= 2\Phi(\delta) - 1 \end{aligned} \tag{4.77}$$



$$P(|Z| \le \delta) = 2\Phi(\delta) - 1. \qquad \square$$

Using (4.76) we can answer (with good approximations for n sufficiently large) questions related to the following equation:

$$P(|\bar{X} - p| \le d) = c. \tag{4.78}$$

1° given p, n, and d, find c

2° given p, d, and c, find n

3° given p, n, and c, find d,

so that (4.78) holds.

Example 1°. A fair coin is tossed 100 times. What is the approximate probability that the fraction of heads obtained is within 0.01 of 0.5?

Solution. With $\bar{X}$ the fraction of heads, (4.76) yields

$$\begin{aligned} P(|\bar{X} - 0.5| \le 0.01) &\approx 2\Phi(0.2) - 1 \\ &= (2)(0.5793) - 1 = 0.1586. \end{aligned}$$

Remark. Note that no continuity correction is used in connection with normal approximations of $\bar{X}$, only of the binomial random variable X.

Example 2°. How many times do we need to flip a fair coin so that the probability is 0.9 that the fraction of heads obtained is within 0.01 of 0.5?

Solution. By (4.76),

$$P(|\bar{X} - 0.5| \le 0.01) \approx 2\Phi\left(\frac{0.01}{\sqrt{\frac{1}{4n}}}\right) - 1,$$

so we solve

$$2\Phi\left(\frac{0.01}{\sqrt{\frac{1}{4n}}}\right) - 1 = 0.9, \text{ i.e.}$$

$$\Phi\left(\frac{0.01}{\sqrt{\frac{1}{4n}}}\right) = 0.95.$$

Reading the table on page 107 inside-out, we see that $\Phi(1.65) \approx 0.95$, so we solve

$$\frac{0.01}{\sqrt{\frac{1}{4n}}} = 1.65 \quad \text{for } n$$

$$\Leftrightarrow (0.02)\sqrt{n} = 1.65 \Leftrightarrow n = \left\lceil \left(\frac{1.65}{0.02}\right)^2 \right\rceil \approx 6807.$$

(Note that if we interpolate and use 1.645 instead, we get $n = \left\lceil \left(\frac{1.645}{0.02}\right)^2 \right\rceil \approx 6766.$)

Example 3°. If a fair coin is flipped 100 times, determine that distance d such that the probability is 0.9 that the fraction of heads obtained is within d of 0.5.

Solution. Obviously, d will be considerably larger than the 0.01 of example 2, which required over 6000 tosses.

Using (4.76) again, we have

$$P(|\bar{X} - 0.5| \le d) \approx 2\Phi\left(\frac{d}{\sqrt{\frac{1}{400}}}\right) - 1 = 0.9$$

$$\Leftrightarrow \Phi(20\,d) = 0.95$$

$$\Leftrightarrow 20\,d = 1.645$$

$$\Leftrightarrow d = 0.08225 \approx 0.08.$$

Random intervals

The following statements all say the same thing:

(4.79a) $$P(|\bar{X} - p| \le d) = c$$

(4.79b) $$P(-d \le \bar{X} - p \le d) = c$$

$$P(p - d \le \bar{X} \le p + d) = c \tag{4.79c}$$

$$P(|p - \bar{X}| \le d) = c \tag{4.79d}$$

$$P(-d \le p - \bar{X} \le d) = c \tag{4.79e}$$

$$P(\bar{X} - d \le p \le \bar{X} + d) = c \tag{4.79f}$$

Let us focus in particular on (4.79c) and (4.79f). In words, (4.79c) says "The probability that the random variable $\bar{X}$ takes a value in the closed interval $[p - d, p + d]$ is c." On the other hand, (4.79f) says "The probability that the *random closed interval* $[\bar{X} - d, \bar{X} + d]$ contains the parameter p is c." These are two ways of saying the same thing.

Equation (4.79f) may seem like a strange and unintuitive reformulation of (4.79c), but it turns out to be exactly what we need to give what is called a *confidence interval estimate* of an unknown binomial parameter p, based on the observed number of successes in n actual trials. Before we get to this statistical application, we need one more probabilistic result.

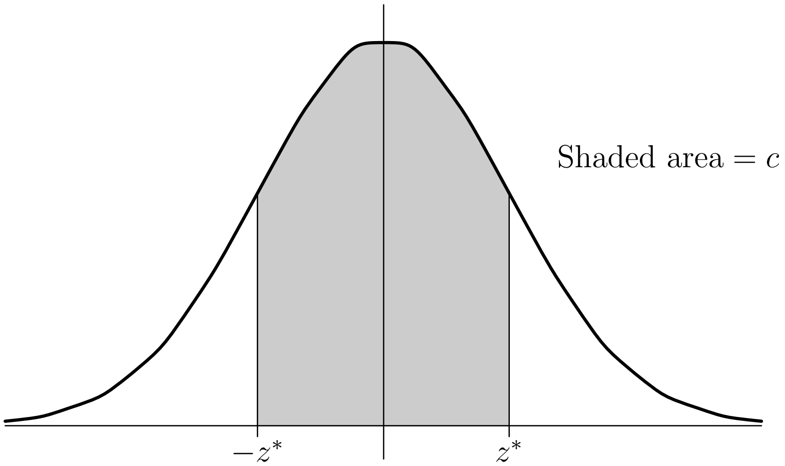
We say that the random closed interval $[\bar{X} - d, \bar{X} + d]$ is *a* 100 c% *confidence interval for* p if (4.79f) holds, i.e., if

$$P(\bar{X} - d \le p \le \bar{X} + d) = c \tag{4.80}$$

The following theorem gives a recipe for constructing an approximation of such an interval, in effect giving a general solution to problem 3° with respect to (4.78).

Theorem 4.14. Let $X \sim \text{binomial}(n, p)$ and $\bar{X} := X/n$. Given c, with $0 < c < 1$, let z^* be the unique real number having the property that

$$P(-z^* \le Z \le z^*) = 2\Phi(z^*) - 1 = c. \tag{4.81}$$



Then

$$\left[\bar{X} - z^*\sqrt{\frac{p(1-p)}{n}}, \bar{X} + z^*\sqrt{\frac{p(1-p)}{n}}\right] \tag{4.82}$$

is an approximate 100 c% confidence interval for p.

Proof.

$$\begin{aligned}
P\left(\bar{X} - z^*\sqrt{\frac{p(1-p)}{n}} \le p \le \bar{X} + z^*\sqrt{\frac{p(1-p)}{n}}\right)& \\
&= P\left(p - z^*\sqrt{\frac{p(1-p)}{n}} \le \bar{X} \le p + z^*\sqrt{\frac{p(1-p)}{n}}\right) \\
&\text{(by the equivalence of (4.79f) and (4.79c))} \\
&= P\left(-z^* \le \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \le z^*\right) \\
&\approx P(-z^* \le Z \le z^*) \\
&= c, \text{ by (4.81).} \quad \square
\end{aligned}$$

Example. Let $X \sim \text{binomial}(n, p)$. Find an approximate 90% confidence interval for p.

Solution. Find z^* such that $P(-z^* \le Z \le z^*) = 2\Phi(z^*) - 1 = 0.9$, i.e., such that $\Phi(z^*) = 0.95$. Reading the table on page 107 "inside-out" (and interpolating, since it's easy), we get $z^* = 1.645$. So

$$\left[\bar{X} - 1.645\sqrt{\frac{p(1-p)}{n}}, \bar{X} + 1.645\sqrt{\frac{p(1-p)}{n}}\right]$$

is an approximate 90% confidence interval for p. This means that if we perform n trials and calculate $\bar{X}$, the fraction of successes, and construct the above interval, and repeat this process many times (thereby constructing many such intervals), that around 90% of those intervals will contain the parameter p.

Estimating the binomial parameter p.

Everything that we have done so far in this section has been probability theory. The assumption of our theorems has always been that $X \sim \text{binomial}(n, p)$ where p is known, and the theorems have specified the behavior of $\bar{X} = X/n$. In many cases, however, we do not know (but would like to estimate) the value of p. The following method for estimating p was developed by Jerzy Neyman in the 1930's. Suppose that we perform an experiment, and that in n actual trials the fraction of successes observed is $\bar{x}$ (Note that the x here is lower case; $\bar{x}$ is a number, not the random variable $\bar{X}$. We often call $\bar{x}$ a *realization* of the random variable $\bar{X}$). If we had to estimate p on the basis of the results of these n trials, it would clearly be reasonable to use the number $\bar{x}$ as our estimate. This is called a *point estimate.*

If we want to be more cautious (which is advisable), we choose a "confidence level" c, and find z^* satisfying (4.81). In a variant of (4.82) we say that the numerical interval

$$\left[\bar{x} - z^*\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}, \bar{x} + z^*\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}\right] \tag{4.83}$$

is a *realization of an approximate* 100 c% *confidence interval for* p. Note that in addition to replacing $\bar{X}$ of (4.82) by its realization $\bar{x}$, we approximate the (here unknown) p in (4.82) by $\bar{x}$.

If we want to avoid this approximation, we replace $p(1-p)$ by $1/4$ instead of by $\bar{x}(1-\bar{x})$, since $\max\{p(1-p) : 0 \le p \le 1\} = 1/4$ by elementary calculus. This produces the possibly wider interval

$$\left[\bar{x} - z^*\sqrt{\frac{1}{4n}}, \bar{x} + z^*\sqrt{\frac{1}{4n}}\right], \tag{4.84}$$

the official name for which is a *realization of an approximate conservative* 100 c% *confidence interval for* p. That is quite a mouthful. Both in the case of (4.83) and (4.84), people generally omit the phrase "realization of an approximate," calling (4.83) simply a 100 c% confidence interval and (4.84) a conservative 100 c% confidence interval for p.

The most common example of the above, with $c = 0.95$, is constructed as follows: Find z^* satisfying $P(-z^* \le Z \le z^*) = 2\Phi(z^*) - 1 = 0.95$, i.e., $\Phi(z^*) = 0.975$. This yields $z^* = 1.96$. So if on n trials we observe a fraction of successes equal to $\bar{x}$, then

$$\left[\bar{x} - 1.96\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}, \bar{x} + 1.96\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}\right] \tag{4.85}$$

is a 95% confidence interval for p, and

$$\left[\bar{x} - 1.96\sqrt{\frac{1}{4n}}, \bar{x} + 1.96\sqrt{\frac{1}{4n}}\right] \tag{4.86}$$

is a conservative 95% confidence interval for p.

Note. Many people round up the 1.96 in (4.85) and (4.86) to 2, which produces a slightly wider interval in each case. In the case of (4.86), one gets the nice, simple interval

$$\left[\bar{x} - \frac{1}{\sqrt{n}}, \bar{x} + \frac{1}{\sqrt{n}}\right] \tag{4.87}$$

by doing this (perhaps we might call (4.87) a "very conservative 95% confidence interval for p"). This is the estimate used by most polls reported in the press, although they don't use the language of confidence intervals. They would report: "The poll indicates that 60% of voters favor stricter gun control. The possible error is 5 percentage points." In our terminology, they are saying that [0.55, 0.65] is a very conservative 95% confidence interval for p, the fraction of voters in the general population favoring stricter gun control. We can deduce how large the sample was because we know that $1/\sqrt{n} = 0.05$. Hence, $n = 400$ was the sample size.

Remark 1. "Sample size is all that matters." The accuracy of an estimate of p, as measured, say, by the length of a conservative 95% confidence interval for p depends only on n, the sample size. This is counterintuitive to many people. But it is true, for example, that a poll of 400 Tennesseans on some issue gives just as accurate a reading of the sentiment of all Tennesseans as a poll of 400 Californians does of the sentiment of all Californians. Each produces a conservative 95% confidence interval of the same length. The accuracy of a poll depends on the sample size, *not on the ratio of the sample size to the population size.* (Of course we assume here that we sample a modest number of individuals from a large population, so the binomial approximation to the hypergeometric distribution is appropriate.)

Remark 2. To say that [a,b] is a 100 c% confidence interval for p is *not* to say that the probability that p lies in [a,b] is c. Probabilities are only meaningfully ascribed to statements about the random variable $\bar{X}$. The interval [a,b] is a numerical interval, not a random interval. It either contains p or it doesn't. The most we can say is that *if* we were to construct 100 c% confidence intervals many times (which we're not doing - we only construct one) we could expect around 100 c% of those intervals to contain p. Neyman chose the name "confidence interval" rather than "probability interval" to emphasize the foregoing point.

4.12 Problems

1. Suppose that an urn contains 1 million balls, 10% of which are white and 90% blue.

   a. In a random sample of 400 balls, what is the approximate probability that the fraction of white balls in the sample is within 0.01 of 0.1?

   b How many balls must be sampled so that the fraction of white balls in the sample is within 0.02 of 0.1, with probability 0.95?

   c. For a sample of 400 balls, determine the value d for which the probability is 0.99 that the fraction of white balls in the sample is within d of 0.1.

2. In n trials we observe a fraction of successes equal to $\bar{x}$. We construct, using this data, both a 100 c$_1$% confidence interval $[a_1, b_1]$ and a 100 c$_2$% confidence interval $[a_2, b_2]$ for p. If $c_1 < c_2$, which of these intervals is contained in the other?

3. In a random sample of 1600 college students, 960 students had read *Faust.*

   a. Find a 95% confidence interval for the fraction of all college students who have read *Faust.*

   b. Find a conservative 90% confidence interval for the aforementioned fraction.

4. In n trials with unknown success probability p, we observe a fraction of successes equal to $\bar{x}$. Find a) a 99% confidence interval for p and b) a conservative 99% confidence interval for p.

5. The probability that an individual in some population has a given characteristic is an unknown p. For fixed $d > 0$ and $0 < c < 1$, how large a sample must we take so that $[\bar{x} - d, \bar{x} + d]$ is a conservative 100 c% confidence interval for p?

6. In the above case, how large a sample must be taken to establish a conservative 90% confidence interval of length 0.01 for p?

7. In Remark 1 of the preceding section I asserted that a poll of 400 Tennesseans on some issue would produce a conservative 95% confidence interval of the same length as a poll of 400 Californians on the same issue. What is that length? What is the length if n people are sampled in each state and 100 c% conservative confidence intervals are constructed, say, for the respective fractions of people in these states favoring stricter gun control.