Intermediate Statistics Supplementary Notes

Spring 2015

David Lyons
Mathematical Sciences
Lebanon Valley College

Intermediate Statistics
Supplementary Notes

Spring 2015

David Lyons
Mathematical Sciences
Lebanon Valley College

Contents

The vocabulary of sets and functions is fundamental to all of mathematics, theoretical and applied. We present basic terminology in the first section of these notes.

1 Sets and Functions

1.1 Sets

A ***set*** is a collection of objects. The objects belonging to a set are called its ***elements*** or ***members***. To indicate that an object x is an element of the set A, we write $x \in A$, and pronounce those symbols "x is an element (or member) of A" or "x belongs to A". We write $x \notin A$ to denote that the object x is not an element of the set A.

Sets are specified by listing or describing the elements inside curly braces. For example, we write $A = \{x, y, z\}$ to specify that the set A consists of elements x, y, and z. We write {even whole numbers} or $\{\ldots, -4, -2, 0, 2, 4, \ldots\}$ to specify the set of even whole numbers. The ellipsis symbol "$\ldots$" indicates that the reader should infer an obvious pattern. Order and redundancy of the list inside curly braces are irrelevant. For example, for the set $A = \{x, y, z\}$, we have

$$A = \{x, y, z\} = \{y, z, x\} = \{x, x, y, z\}.$$

The colon symbol or vertical bar inside curly braces denotes the phrase "such that". For example, we may write $\{n^2 : n = 1, 2, 3\}$ or $\{n^2 \mid n = 1, 2, 3\}$ to describe the set $\{1, 4, 9\}$.

It is convenient to have a special set, called the ***empty set***, that contains no members. It plays a role among sets analogous to the role of zero among numbers. The empty set is denoted by an empty pair of curly braces { } or by the symbol $\emptyset$.

We write $A \subseteq B$ or $A \subset B$ to indicate that all the members of set A are also members of set B, and we express this by saying "A is a ***subset*** of B", "A is ***contained in*** B" or "B ***contains*** A". According to this definition, every set is a subset of itself. Less intuitive, but also a consequence of the defintion, is that the empty set is a subset of any other set. Some examples of subsets: the sets $\emptyset$, $\{y\}$, $\{z, x\}$, and $\{x, y, z\}$ are subsets of $\{x, y, z\}$. To indicate that A is a subset of B but not equal to B, we write $A \subsetneq B$ and say A is a ***proper subset*** of B, or A is ***properly contained*** in B.

We write $A \cap B$, pronounced "the ***intersection*** of A with B" or "A intersect B", to denote the set

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

of all objects that are members both of set A and also of set B. We write $A \cup B$, pronounced "the ***union*** of A with B" or "A union B", to denote the set

$$A \cup B = \{x \colon x \in A \text{ or } x \in B\}$$

of all objects that are members of either set A or of set B or both. An important feature of this definition is that in mathematics, the word "or" is *always* used in the inclusive sense. That is, "or" means "one or the other or both". We

write $A \setminus B$ or $A - B$, pronounced "the ***complement*** of B ***relative to*** A" or "A minus B", to denote the set

$$A \setminus B = \{x\colon x \in A \text{ and } x \notin B\}$$

of all objects which are members of the set A *and are not* members of the set B. Sets A and B are ***disjoint*** or ***mutually exclusive*** if their intersection is the empty set. An example: let $A = \{x, y, z\}$ and let $B = \{a, b, y, z\}$. Then we have $A \cap B = \{y, z\}$, $A \cup B = \{a, b, x, y, z\}$, $A \setminus B = \{x\}$, and $B \setminus A = \{a, b\}$.

An ***ordered pair of elements from the set*** A is an ordered list (x, y) of two elements from A, where we allow the possibility that x equals y. It is important to not confuse ordered pairs with sets containing two elements. For example, in the set $A = \{x, y, z\}$, the symbols (y, x) denote the ordered list with y first and x second, which is different from the ordered pair (x, y). Both of these are different from the two-element set $\{x, y\}$. The set $A \times B$, called the ***(Cartesian) product*** of A and B, is the set of all ordered pairs (a, b) of elements from $A \cup B$ such that $a \in A$ and $b \in B$. Here is an example.

$$\begin{aligned}\{x, y, z\} \times \{a, b, y, z\} \;=\; &\{(x, a), (x, b), (x, y), (x, z),\\ &(y, a), (y, b), (y, y), (y, z),\\ &(z, a), (z, b), (z, y), (z, z)\}\end{aligned}$$

A technical consequence of this definition is that for any set A, we have $A \times \emptyset = \emptyset$ because there are *no* ordered pairs (a, b) with $a \in A$ and $b \in \emptyset$. We use the notation A^2 (pronounced "A squared") to denote the product $A \times A$ of a set A with itself. Given a finite collection of sets $A_1, A_2, \ldots, A_n$, the ***n-fold (Cartesian) product*** $A_1 \times A_2 \times \cdots \times A_n$ is the set of all ordered lists, also called ***n-tuples***, of the form $(a_1, a_2, \ldots, a_n)$, where $a_k \in A_k$ for every k in the range $1 \le k \le n$. We write A^n to denote the n-fold product of a set A with itself. For example,

$$\begin{aligned}\{a, b\}^3 \;=\; &\{(a, a, a), (a, a, b), (a, b, a), (a, b, b),\\ &(b, a, a), (b, a, b), (b, b, a), (b, b, b)\}.\end{aligned}$$

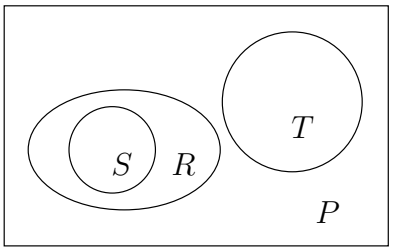


Figure 1
Euler diagram example

It is often helpful to use pictures to visualize the relationships between sets. ***Euler diagrams*** depict sets as 2-dimensional regions in the plane. Figure 1 shows an Euler diagram illustrating the relationships between the set R of all rectangles, the set S of all squares, the set T of all triangles and the set P of all polygons. A special type of Euler diagram called a ***Venn diagram*** is used to visualize unions, intersections, and complements of sets. Figure 2 shows an example.

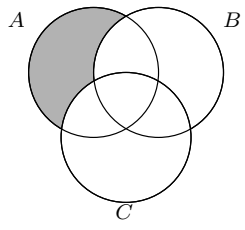


Figure 2
Venn diagram showing
$A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C)$

Some important sets

Certain sets are so widely used in mathematics that they have standard names and symbols. One of the most important of these is the set $\mathbf{R}$ of real numbers, which is the set of points on a line. The name "real" indicates the notion that $\mathbf{R}$ is an appropriate set to represent quantities which can be measured in the "real" physical world, such as time, distance, temperature, etc. The set $\mathbf{R}^2 = \mathbf{R} \times \mathbf{R} = \{(x, y) : x, y \in \mathbf{R}\}$ is called the ***x, y-coordinate plane*** or the ***Euclidean plane***, named after Euclid (ca. 300 BC) because it is the

setting for classical plane geometry. The set of ***natural numbers*** is the set $\mathbf{N} = \{1, 2, 3, \ldots\}$ of counting numbers. The ***integers*** or ***whole numbers***, is the set $\mathbf{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$. The ***rational numbers*** or ***fractions***, denoted $\mathbf{Q}$, is the subset of all real numbers which can be written in the form m/n, where m and n are integers and $n \neq 0$.

Notes on terminology

The symbol $\mathbf{Z}$ for the integers comes from the German "Zahlennummern," which means "counting numbers." The symbol $\mathbf{Q}$ for the rationals comes from the word "quotient." The word "rational" comes from the root for *ratio*, meaning proportion. Beware that the notation for an open interval $(a, b) = \{x : a < x < b\}$ of the real line is identical to the notation for the point (a, b) in the x, y-plane; if context does not make clear which is meant, then some additional comment is appropriate on the part of the user.

1.2 Functions

A function is a mathematical model for a process or machine that takes "inputs," does something to them, then produces "outputs." While the idea is not complicated, it is difficult to give a precise definition in everyday language, so some formality is required. The collections of inputs and outputs are modeled by sets. The function itself is modeled by pairs of the form (input value, output value). The machine is not allowed to be ambiguous; for an input value a, there must be exactly one output value b. Here is the formal definition, using the language of sets, that captures this idea.

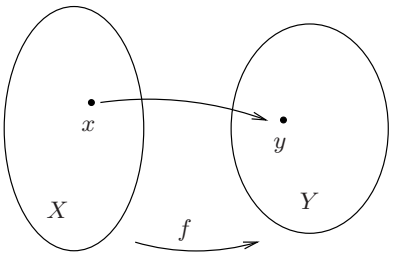


Figure 3
Schematic diagram of a function

A ***function*** f ***from a set*** X ***to a set*** Y, denoted $f\colon X \to Y$, is a subset of $X \times Y$ in which each element $x \in X$ appears in exactly one ordered pair. That is, if $(x, y) \in f$ and $(x, y') \in f$, then it must be that $y = y'$. We write $f(x) = y$ or $x \mapsto y$ to mean $(x, y) \in f$. The set X is called the ***domain*** of f, and the set Y is called the ***codomain***. The arrows in the symbols $f\colon X \to Y$ and $x \mapsto y$ remind us that the machine takes input $x \in X$ and produces output $y = f(x) \in Y$.

Figure 3 shows a schematic representation of a function $f\colon X \to Y$. We use Euler diagrams for the domain and codomain sets with an arrow labeled f to indicate direction. An arrow from a point x in X to a point y in Y indicates that $f(x)$ is y. Figure 4 shows a version of a commonly used diagram that illustrates the conceptualization of a function as a machine.

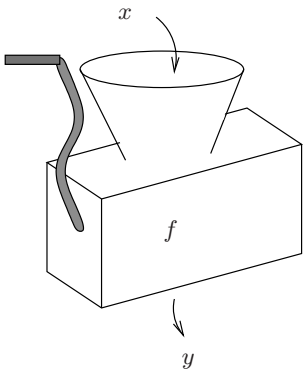


Figure 4
A function "machine"

Functions are often specified by equations. For example, we write $f(x) = x^2$ or $g(x) = 2x + 3$ to define functions f and g whose domains are sets of real numbers. We often refer to "the function $f(x) = x^2$," or simply, "the function x^2," to mean the function f defined by the equation $f(x) = x^2$. This language carries the potential to confuse the function f with its value $f(x)$; care must be taken when the distinction makes a difference.

Usually, when a function is specified by an equation, the domain is not explicitly given. For example, one might say "the function $h(x) = \sqrt{x - 2}$". The convention in such cases is that the domain is all real x for which the equation specifying the function yields a meaningful real number. In this example, the domain for h is $\{x : 2 \leq x\}$.

The terms ***map*** and ***mapping*** are synonyms for the term *function*. We may also write $X \xrightarrow{f} Y$ to denote the function $f\colon X \to Y$. We write $x \mapsto y$ (pronounced "x goes to y" or "x maps to y") to denote that $f(x) = y$. We refer to $y = f(x)$ as the ***value of*** f ***at*** x or the ***image of*** x ***under*** f. The ***image of a function*** $f\colon X \to Y$ is the subset $\mathrm{Im}(f)$ of the codomain Y given by

$$\mathrm{Im}(f) = \{f(x) : x \in X\}.$$

Note: the term *range of a function* may refer to either (1) the image of the function or (2) the codomain; since range as two meanings, care must be taken to avoid ambiguity.

Given a subset A of the domain X of the function $f\colon X \to Y$, the ***image of*** A is defined to be the set

$$f(A) = \{f(x) : x \in A\}.$$

It is worth noting that the image of f is the same thing as $f(X)$. Given a subset B of the codomain Y, the ***preimage of*** B or the ***inverse image*** of B, is the set

$$f^{-1}(B) = \{x \in X : f(x) \in B\}.$$

For a point $y \in Y$, we write $f^{-1}(y)$ to denote the preimage $f^{-1}(\{y\})$. A function $f\colon X \to Y$ is called ***one-to-one*** or ***injective*** if $f^{-1}(y)$ has no more than 1 element for every $y \in Y$. The function f is called ***onto*** or ***surjective*** if $f^{-1}(y)$ has at least 1 element for every $y \in Y$. The function f is called a ***one-to-one correspondence*** or ***bijective*** if $f^{-1}(y)$ has exactly 1 element for every $y \in Y$, that is, if f is both one-to-one and onto. Examples: consider $f\colon \mathbf{Z} \to \mathbf{Z}$, $g\colon \mathbf{N} \to \mathbf{Z}$, and $h\colon \mathbf{Z} \to \mathbf{Z}$ given by $f(n) = g(n) = n^2$ and $h(n) = -n$. The function f is not one-to-one because the set $f^{-1}(4) = \{-2, 2\}$ has more than one element. The function g is one-to-one because the set $g^{-1}(n)$ is either empty (if n is not a perfect square) or is the 1-element set $\{\sqrt{n}\}$ (if n is a perfect square). Neither f nor g is onto because $f^{-1}(3) = g^{-1}(3) = \emptyset$. The function h is both one-to-one and onto because $h^{-1}(n) = \{-n\}$ for every n in $\mathbf{Z}$.

Given a set X, the ***identity function*** on X is the function $\mathrm{id}\colon X \to X$ given by $x \mapsto x$ for all x in X. A ***constant function*** is a function $f\colon X \to Y$ for which there is an element y_0 in Y such that $f(x) = y_0$ for all x in X.

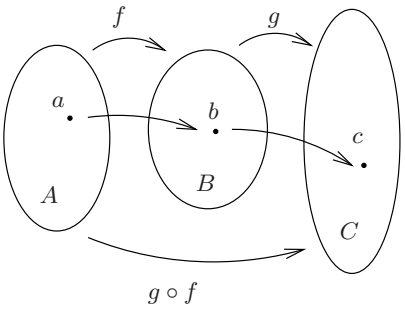


Figure 5
Composition of functions

Composition

Given two functions $f\colon A \to B$ and $g\colon B \to C$, the ***composition*** $g \circ f$ is the function $g \circ f\colon A \to C$ defined by $(g \circ f)(a) = g(f(a))$. Note that the order matters; $g \circ f$ is not the same as $f \circ g$. Figure 5 shows a schematic diagram.

Inverse functions

Functions operate on input values to produce output values. It is often worthwhile to reverse this process. For example, suppose we have a function f which tells us the dollar value $A = f(t)$ of an investment at time t. A practical problem would be to determine the time value needed to realize a given investment value. In other words, we are seeking a "reversing function," say g, that operates on dollar values and produces the corresponding time values. To say that

g reverses the procedure f is to say $g(A) = t$ whenever $f(t) = A$. For any time value t or dollar value A, we would have the following.

$$g(f(t)) = t \qquad\qquad f(g(A)) = A$$

The above equations say that the composition $g \circ f$ is the identity function on the domain of time values of the function f, and $f \circ g$ is the identity function on the domain of dollar values of the function g. This example motivates the official definition of inverse functions.

(1.2.1) **Definition of Inverse Function.** Suppose $f\colon X \to Y$ and $g\colon Y \to X$ satisfy the equations

$$g \circ f = \mathrm{id}_X \qquad\qquad f \circ g = \mathrm{id}_Y$$

where id_X and id_Y denote the identity functions on X and Y, respectively. Then we say f and g are ***inverses*** of one another, and we write $g = f^{-1}$ and $f = g^{-1}$. The functions f and g are also called ***invertible***.

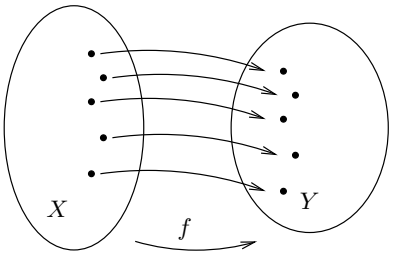


Figure 6
An invertible function

(1.2.2) **Comment on notation clash, and an important fact about invertible functions.** The alert reader will have noticed that we have now given two different usages of the symbol f^{-1}. We write $f^{-1}(y)$ to denote the *preimage* set of an element y in the codomain, which is defined for *any* function f, and we write $f^{-1}(y)$ to denote the *image* of the element $y \in Y$ under the inverse function for f, which is defined only if f is invertible. Happily, the two meanings have a harmonious resolution when the latter is defined. If the preimage $f^{-1}(\{y\})$ is the 1-element set $\{x\}$, then the image $f^{-1}(y)$ of y under the inverse function f^{-1} is the element x, and vice-versa. It is an important fact that f is invertible if and only if f is one-to-one and onto.

A visual representation of an invertible function $f\colon X \to Y$ (see Figure 6) shows the assignments made by f as arrows matching the elements of X and Y in a one-to-one manner. The picture of the inverse function f^{-1} is obtained by simply reversing the direction of all the arrows.

(1.2.3) **Examples of inverse functions.** Suppose X is a finite set, and $f\colon X \to Y$ is an invertible function. Since f matches the elements of X with the elements of Y in a one-to-one manner, Y must also be a finite set with the same number of elements as X.
Let $s\colon [0,\infty) \to [0,\infty)$ be the squaring function given by $x \mapsto x^2$. The inverse of s is the square root function $r\colon [0,\infty) \to [0,\infty)$ given by $x \mapsto \sqrt{x}$. Note that the domains are important here. Let $f\colon \mathbf{R} \to \mathbf{R}$ also be the squaring function $x \mapsto x^2$, but on the domain of all reals. The square root function is *not* an inverse for f because $r(f(-2)) = \sqrt{(-2)^2} = 2 \neq -2$. A lesson here is that a function given by an equation may be invertible with one domain, but not invertible with another domain.

Operations on real-valued functions

A ***real-valued function*** is a function whose codomain is a subset of the real numbers. Given two functions $f\colon A \to \mathbf{R}$, $g\colon A \to \mathbf{R}$ and a constant real number k, we define the functions kf, $f+g$, $f-g$, $f \cdot g$ and f/g by the following formulas, for all a in A (note that the last equation is defined only when $g(a) \neq 0$, so the

function f/g is defined to have domain $\{a \in A : g(a) \neq 0\}$).

$$\begin{aligned}(kf)(a) &= kf(a)\\ (f+g)(a) &= f(a)+g(a)\\ (f-g)(a) &= f(a)-g(a)\\ (f\cdot g)(a) &= f(a)g(a)\\ (f/g)(a) &= f(a)/g(a)\end{aligned}$$

Note: The notation fg is sometimes used to mean the product function $f\cdot g$, and sometimes to mean the composition $f \circ g$. Care should be taken when context does not make clear which is meant.

Summation, Intersection, and Union Notation

Let m, n be nonnegative whole numbers with $n \geq m$, and let $f\colon \{m, m+1, \ldots, n\} \to R$ be a function. We write $\sum_{i=m}^{n} f(i)$ to denote the ***sum***

$$\sum_{i=m}^{n} f(i) = f(m) + f(m+1) + \cdots + f(n).$$

The symbol $\sum$ is the capital Greek letter sigma, and denotes a sum. The variable i is called the ***index*** of the sum. More generally, we write $\sum_{i=m}^{n} x_i$ to denote the sum

$$x_m + x_{m+1} + x_{m+2} + \cdots + x_n$$

where m, n are integers with $m \leq n$.

Given a collection $A_1, A_2, \ldots, A_n$ of sets, we write $\bigcap_{i=1}^{n} A_i$, $\bigcup_{i=1}^{n} A_i$, to denote the ***intersection*** and ***union***, respectively, of the sets, defined as follows.

$$\bigcap_{i=1}^{n} A_i = \{x\colon x \in A_i \text{ for all } i, 1 \leq i \leq n\}$$

$$\bigcup_{i=1}^{n} A_i = \{x\colon x \in A_i \text{ for some } i, 1 \leq i \leq n\}$$

1.3 Exercises

1. List all the subsets of the set $X = \{a, b, c, d\}$.
2. Let $A = \{1, 3, 5, 7, 9\}$, let $B = \{2, 3, 4, 5, 6\}$ and let $D = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ be the set of all ten digits. Find $A \cup B$, $A \cap B$, $A \setminus B$, and the $D \setminus A$. Draw a single Euler diagram showing the relationships between A, B, D, and the set $C = \{0, 2, 6\}$.
3. Sketch a Venn diagram for the *symmetric difference* $(A \setminus B) \cup (B \setminus A)$ of two sets A and B for the sets A and B in the previous problem.

4. Use a Venn diagram to illustrate the following property of set operations (one of *DeMorgan's Laws*).

$$A \setminus (B \cap C) = (A \setminus B) \cup (A \setminus C)$$

5. Let $A = \{a, b\}$ and $B = \{x, y, z\}$.
   (a) List all members of the set $A \times B$.
   (b) List all members of the set B^2.

6. Draw a single Euler diagram illustrating the relationships between the following sets: $\mathbf{R}$, $\mathbf{Q}$, $\mathbf{Z}$, and $[0, 1)$.

7. Let A and B be subsets of the real line $\mathbf{R}$ given by $A = \{x : -2 \leq x < 3\}$ and $B = \{x : 1 < x \leq 5\}$. Write in interval notation, set notation and sketch a picture of A, B, $A \cup B$, $A \cap B$, $A \setminus B$ and $\mathbf{R} \setminus A$. Example: in interval notation, set A is written $A = [-2, 3)$, in set notation $A = \{x : -2 \leq x < 3\}$, and the sketch of A is the shaded region of the real line marked with endpoint brackets at -2 on the left and 3 on the right.

8. Let $f(x) = x^2$ and $g(x) = x + 2$ define functions f and g from the reals to the reals.
   (a) Find $(f \circ g)(3)$
   (b) Find $(g \circ f)(3)$
   (c) Find $(g \cdot f)(3)$
   (d) Find $(f/g)(3)$
   (e) Find $(3f + g)(3)$
   (f) Write an equation for $(f \circ g)(x)$
   (g) Write an equation for $(g \circ f)(x)$

9. Let $X = \{a, b, c\}$ and $Y = \{1, 2, 3\}$. Describe all possible one-to-one correspondences between X and Y.

10. (a) Does there exist a function $f\colon \emptyset \to X$, where X is nonempty? If so, give an example. If not, explain.
    (b) Does there exist a function $f\colon X \to \emptyset$, where X is nonempty? If so, give an example. If not, explain.

11. (a) Evaluate $\displaystyle\sum_{i=1}^{10}(2i + 1)$.
    (b) Write the sum

$$(1^2 + 2 \cdot 1 + 3) + (2^2 + 2 \cdot 2 + 3) + (3^2 + 2 \cdot 3 + 3) + \cdots + (15^2 + 2 \cdot 15 + 3)$$

    using summation notation.

12. Let $A = \{x, y, z\}$ and let $f\colon A \to \mathbf{R}$ and $g\colon A \to \mathbf{R}$ be given by the following table of values.

element of A	value of f	value of g
x	2	5
y	0	1
z	-1	2

Let $s_1 = x, s_2 = y, s_3 = z$ be an ordered list of the elements in A. Find the following.

(a) $\sum_{i=1}^{3} f(s_i)$

(b) $\sum_{i=1}^{2} (f+g)(s_i)$

(c) $\sum_{i=2}^{3} (f \cdot g)(s_i)$

2 Counting

In this section we present some basic principles of counting.

2.1 Basic Principles

Given a set X, we say that X is ***finite*** if there exists a natural number n such that there exists a bijection between X and the set $\{1, 2, \ldots, n\}$. In this case we write $|X| = n$ and refer to n as the ***number of elements*** in X or the ***size*** of X. Here are several intuitive facts.

(2.1.1) **Bijection Principle.** *If $f\colon X \to Y$ is a bijection between finite sets, then $|X| = |Y|$.*

(2.1.2) **Addition Principle.** *Given finite sets A, B, we have*

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

(2.1.3) **Multiplication Principle.** *Given finite sets A, B, we have*

$$|A \times B| = |A||B|.$$

A consequence of the Addition Principle is that if A and B are disjoint finite sets, then the size of their union is the sum of their sizes. Applying this to a collection of disjoint sets leads to the following.

(2.1.4) **Multiplication Principle 2.** *If a finite set X is the disjoint union of r subsets of size s, then $|X| = rs$.*

Multiplication Principle 2 can also be derived from the ordinary Multiplication Principle, as follows. Given a finite set X, let $A_1, A_2, \ldots, A_r$ be r disjoint subsets of X whose union is X. In each set A_i, label the s distinct elements $a_{i1}, a_{i2}, \ldots, a_{is}$, for $1 \leq i \leq r$. Then we have a bijection $X \to \{1, 2, \ldots, r\} \times \{1, 2, \ldots, s\}$ given by $x \mapsto (i, j)$, where $x = a_{ij}$, that is, x is the jth element of the ith subset. Then by the Bijection Principle and the Multiplication Principle, we have $|X| = rs$.

It is also easy to derive the ordinary Multiplication Principle from Multiplication Principle 2. Given finite sets A, B, let $S_a = \{(a, y) : y \in B\}$ for each a in A. Clearly, the sets S_a are disjoint, their union is $A \times B$, the number of S_as is $|A|$, and each S_a has $|B|$ elements, so by Multiplication Principle 2, we have $|A \times B| = |A||B|$.

Multiplication Principle 2 can be rephrased as follows. Choosing an element of the set X that is the disjoint union of r sets, each of size s, can be thought of as a two-step decision process: first choose one of the r subsets, then choose one of the s elements in that set. Here is a version of this "decision sequence" version of the multiplication principle.

(2.1.5) **Multiplication Principle 3.** *Suppose that a sequence of two decisions must be made, where the first decision chooses one from among r choices, and having made the first decision, the second decision chooses one from among s choices (although the s choices need not be the same for each of the r initial choices). Then the total number of ways to make the pair of decisions is rs.*

It is useful to have "multi-step" versions of the multiplication principle. Here are two of them.

(2.1.6) **Multiplication Principle 4.** *Given finite sets $A_1, A_2, \ldots, A_m$, we have*

$$|A_1 \times A_2 \times \cdots \times A_m| = |A_1||A_2| \cdots |A_m|.$$

(2.1.7) **Multiplication Principle 5.** *Suppose that a sequence of m decisions must be made, where the kth decision chooses one from among r_k choices for $1 \leq k \leq m$. Then the total number of ways to make the m decisions is $r_1 r_2 \cdots r_m$.*

2.2 Permutations

A ***permutation of a set*** X is a bijection from X to itself. The set of ***permutations of*** n ***items*** is the set of permutations of the set $\{1, 2, \ldots, n\}$. A permutation $f\colon \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ determines an ordered n-tuple $(i_1, i_2, \ldots, i_n)$ of distinct elements of $\{1, 2, \ldots, n\}$ by setting $i_j = f(j)$ for $1 \leq j \leq n$. Conversely, an ordered n-tuple $(i_1, i_2, \ldots, i_n)$ of distinct elements of $\{1, 2, \ldots, n\}$ determines a bijection $f\colon \{1, 2, \ldots, n\} \to \{1, 2 \ldots, n\}$ by assigning $f(j) = i_j$ for $1 \leq j \leq n$. We can use Multiplication Principle 5 to count the number of n-tuples of distinct elements, and hence the number of permutations, of $\{1, 2, \ldots, n\}$, as follows, by counting sequences of n decisions: first choose i_1, then choose i_2, and so on, until the final choice of i_n. There are n choices for i_1, then $n-1$ remaining choices for i_2, and so on, and finally there is 1 choice for i_n. Thus we have $n(n-1)(n-2)\cdots(3)(2)(1)$ permutations of $\{1, 2, \ldots, n\}$. This number is so widely used that it has special notation and a name. We write $n!$, pronounced "n ***factorial***", for the quantity $n(n-1)(n-1)\cdots(3)(2)(1)$, for all natural numbers n. Because there is 1 bijection from the empty set to itself (think about that!), we define 0! to be 1.

We can modify the above counting procedure to obtain the number $P(n, k)$ of k-tuples of distinct elements of $\{1, 2, \ldots, n\}$.

$$P(n,k) = \underbrace{n(n-1)(n-2)\cdots(n-k+1)}_{k \text{ factors}} = \frac{n!}{(n-k)!} \tag{2.2.1}$$

We call $P(n, k)$ the ***number of permutations of*** n ***items chosen*** k ***at a time***.

2.3 Combinations

We write $C(n, k)$ or $\binom{n}{k}$, pronounced "n choose k", to denote the number of subsets of size k in a set of size n. For example, the subsets of size 2 of the set $\{a, b, c, d\}$ are $\{a, b\}, \{a, c\}, \{a, d\}, \{b, c\}, \{b, d\}, \{c, d\}$, so $\binom{4}{2} = 6$. The number $\binom{n}{k}$ is also called the ***number of combinations of*** n ***items chosen*** k ***at a time***, and is also called a ***binomial coefficient***. The latter name comes from the formula

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k} \tag{2.3.2}$$

for the nth power of the binomial expression $a + b$. Combinations are related to permutations as follows. To specify an ordered k-tuple in an n-element set X, we perform the following two-step decision process. First, choose a subset of X of size k. Second, choose an ordering of that subset. The first decision has $C(n, k)$ choices, and the second decision has $k!$ choices, so by Multiplication Principle 3 we have $P(n, k) = C(n, k) \cdot k!$, and hence, solving for $C(n, k)$

$$C(n, k) = \frac{P(n, k)}{k!} = \frac{n!}{k!(n - k)!}. \tag{2.3.3}$$

2.4 Exercises

1. How many integers are in the open interval of the real line $\{x\colon 53 < x < 186\}$?

   Hint: consider the function $f\colon \mathbf{Z} \to \mathbf{Z}$ given by $f(n) = n - 53$. Observe that f is a bijection and count the elements in $f(I)$, where I is the given interval of integers.

2. How many integers are in the closed interval of the real line $\{x\colon 53 \leq x \leq 186\}$?

3. How many distinct telephone numbers are there under the following rules: a telephone number has 10 digits (3 digit area code plus 7 digit local number), the first digit (first digit of the area code) must not be a zero or a one, and the fourth digit (first digit of the local number) must not be a zero or a one?

4. How many different ways are there to make an ordered list of 20 people?

5. A DJ has recordings of 50 songs. How many different playlists (where order of the songs is part of the playlist) of 20 songs can she possibly make, with no songs repeated?

6. In the problem above, how many playlists are there if the DJ can repeat any number of songs any number of times?

7. Twenty players sit on a bench. How many different teams of 11 players are there?

3 Probability

The theory of probability was developed to analyze processes involving chance. In this section we give the basic vocabulary, properties, and examples of the simplest type of the probability model, called a *finite discrete probability space*. The general case of infinite probability spaces is much more technical. However, it is fair to say that many aspects of the general case are approximated by large finite probability spaces.

Notation preliminaries

In the definitions below, we use the Greek letter capital omega Ω as the name of a finite set. We use lower case omega ω to denote an element in Ω. Given a finite set S consisting of distinct elements $s_1, s_2, \ldots, s_N$, and a function $f\colon S \to \mathbf{R}$, we write $\sum_{s\in S} f(s)$ to denote the sum $f(s_1) + f(s_2) + \ldots + f(s_N)$.

3.1 Probability function and probability space

A function $P\colon \Omega \to \mathbf{R}$ on a finite set Ω is called a ***probability function*** if it satisfies the following properties.

(i) $P(\omega) \geq 0$ for all ω in Ω

(ii) $\sum_{\omega\in\Omega} P(\omega) = 1$

The set Ω is called the ***probability space*** or ***sample space*** of the probability function P. Elements of Ω are called ***outcomes*** and subsets of Ω are called ***events***. Given an event E, we define the ***probability of*** E, denoted $P(E)$, to be the sum

$$P(E) = \sum_{\omega\in E} P(\omega).$$

We define the probability of the empty set to be zero. Given a single outcome ω, we call $P(\omega) = P(\{\omega\})$ the ***probability of*** ω. If P is a constant function given by $P(\omega) = 1/N$ for all ω, where N is the number of elements in Ω, we say P is a ***uniform*** probability function.

(3.1.1) **Examples.**

1. Tossing a fair coin is modeled by the probability space $\Omega = \{H, T\}$ with probability function p given by $P(H) = P(T) = 1/2$.
2. A single roll of a fair die is modeled by the probability space $\Omega = \{1, 2, 3, 4, 5, 6\}$ with the uniform probability distribution $p(\omega) = 1/6$ for all six outcomes.
3. A gambling game in which your chance of winning is 25% is modeled by the probability space $\Omega = \{W, L\}$ with $P(W) = .25$ and $P(L) = .75$.

Given events E and F in a probability space Ω, we refer to the intersection $E \cap F$ as the event "E and F," and refer to the union $E \cup F$ as the event "E or F," where "or" has the meaning "one or the other or both." The event

$\Omega \setminus E = \{\omega \in \Omega \colon \omega \notin E\}$, also denoted E^c, is called the ***complement*** or ***opposite*** of E. Two events that have no outcomes in common (that is, subsets of the sample space whose intersection is the empty set) are called ***disjoint*** or ***mutually exclusive***.

(3.1.2) **Example.** In example 2 above that models a single roll of a fair die. Let $E = \{2, 4, 6\}$ be the event "roll an even number" and $F = \{4, 5, 6\}$ be the event "roll a number greater than 3." Then the event "E and F" is the set $E \cap F = \{4, 6\}$. In words, this is the set of all rolls that fit the description "roll an even number that is also greater than 3." The event "E or F" is the set $E \cup F = \{2, 4, 5, 6\}$. The is the set of all rolls that fit the description "roll an even *or* roll a number greater than 3." The opposite of E is the event $\{1, 3, 5\}$ described by the phrase "roll an odd number."

3.2 The Addition Rule

Given events E and F in a sample space Ω, regrouping the terms in the sum

$$P(E) + P(F) = \sum_{\omega \in E} P(\omega) + \sum_{\omega \in F} P(\omega)$$

yields the sum

$$\sum_{\omega \in E \cup F} + \sum_{\omega \in E \cap F} = P(E \cup F) + P(E \cap F).$$

Solving for $P(E \cup F)$ gives the ***union rule***, also called the ***sum rule*** or ***addition rule***.

(3.2.1) $$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

When E and F are mutually exclusive, that is, when $E \cap F$ is empty, the union rule takes on a simpler form, sometimes called the *sum rule* or *addition rule* for mutually exclusive events.

(3.2.2) $$P(E \cup F) = P(E) + P(F) \quad (E, F \text{ mutually exclusive})$$

When $F = E^c$, we have $E \cup F = \Omega$ and the sum rule yields

(3.2.3) $$1 = P(S) = P(E \cup E^c) = P(E) + P(E^c).$$

Rearranging by solving for $P(E)$, we get a formula called the ***complement rule*** or ***opposite rule***.

(3.2.4) $$P(E) = 1 - P(E^c)$$

In words, the opposite rule says that the probability of an event equals one hundred percent minus the probability of the opposite event. This rule is useful because sometimes $P(E^c)$ is easier to determine than $P(E)$. See example 2 in 3.5 below.

3.3 Conditional Probability and Multiplication Rule

Let E and F be two events in a sample space Ω, with $P(F) \neq 0$. The ***conditional probability of*** E ***given*** F, denoted $P(E|F)$, is defined to be

(3.3.1) $$P(E|F) = \frac{P(E \cap F)}{P(F)}.$$

Conditional probability has the following meaning. Suppose it is known that event F has occurred in the game of chance modeled by Ω, and we wish to assess the probability that the selected outcome belongs to the set E. In effect, the knowledge that the given outcome lies in F allows us to reduce the sample space from Ω to F. We replace $P\colon \Omega \to \mathbf{R}$ by $P'\colon F \to \mathbf{R}$ given by $P'(\omega) = P(\omega)/P(F)$. We ignore all the outcomes in E that don't belong to F, which means that we replace E by $E \cap F$ and calculate the "new" probability of E as $\frac{P(E\cap F)}{P(F)}$. In words, $P(E|F)$ is the probability that the event E occurs ***given that*** the event F has occurred.

A rearrangement of (3.3.1) is called the ***multiplication rule***. Given events E, F with $P(F) \neq 0$, we have

$$P(E \cap F) = P(E|F)P(F). \tag{3.3.2}$$

In words, the multiplication rule says that the probability that events E and F both happen is equal to the probability that E happens given that F has happened, times the probability that F happens. This rule is useful because it gives us a way to calculate $P(E \cap F)$, which is often complicated, in terms of $P(E|F)$ and $P(F)$, which are often simpler.

(3.3.3) **Example.** Problem: Two cards are dealt from a well-shuffled standard 52 card deck. What is the probability that both cards are red?

Comment: When we use the multiplication rule (3.3.2) we think of F as "happening first" and work out the right hand side from right to left. Here is the solution for this example.

Solution: The probability we wish to find is the probability that the first card is red and the second card is red. Working on the right hand side of (3.3.2) from right to left, let F be the event "first card is red" and E be the event "second card is red." It is easy to see that $P(F) = 1/2$ (there are 26 red cards in the 52 card deck; each card has the same chance of being drawn first) and that $P(E|F) = 25/51$ (once F has happened, there are 51 cards left, 25 of which are red; each of the 51 cards has the same chance of being drawn next). Therefore, the multiplication rule gives us

$$P(E \cap F) = 1/2 \cdot 25/51 = 25/102.$$

3.4 Independence and Dependence

Two events E, F are called ***independent*** if

$$P(E \cap F) = P(E)P(F) \tag{3.4.1}$$

and are called ***dependent*** otherwise. If $P(F) \neq 0$, dividing both sides of (3.4.1) by $P(F)$ yields $P(E|F) = P(E)$. The intuitive practical interpretation of the equation $P(E|F) = P(E)$ is that the likelihood of E occurring does not depend on knowledge that F has occurred. Conversely, if the likelihood of E occurring *does* depend on knowledge that F has occurred, that is, if $P(E|F) \neq P(E)$, then E, F are dependent.

We use equation (3.4.1) in two distinct ways: as the *definition* of independence for two events; and as a *formula* where we use the right hand side to find the unknown probability on the left hand side when E, F are known to be

independent. In the latter case, we think of (3.4.1) as a special case of the multiplication rule (3.3.2).

(3.4.2) **Examples.**

1. (Using (3.4.1) as a formula for an unknown probability) What is the chance that a fair coin comes up heads in two consecutive tosses? It is intuitively clear that the probability that the second toss comes up heads is $1/2$ whether or not we know that the first toss comes up heads. In other words, the events $E =$ "first toss is heads" and $F =$ "second toss is heads" are independent. We use (3.4.1) to get

$$P(E \cap F) = P(E)P(F) = 1/2 \cdot 1/2 = 1/4.$$

2. (Using (3.4.1) as a definition to check independence) Problem: Let S be a sample space consisting of twelve marbles labeled by the letters A through L. Let U be the event consisting of the letters A through F, let V be the event consisting of the letters D through I, and let W be the letters H, I, J. For each pair U, V, U, W and V, W, decide if the pair is independent or dependent.

   Solution: We have $U \cap V = \{\text{D,E,F}\}$, $U \cap W = \emptyset$ and $V \cap W = \{\text{H,I}\}$. We see that $P(U \cap V) = 3/12 = 1/4$ and $P(U)P(V) = 1/2 \cdot 1/2 = 1/4$, so U and V are independent of one another. We have $P(U \cap W) = 0$, whereas $P(U)P(W)$ is certainly not zero, so U and W are dependent events. Finally, we have $P(V \cap W) = 2/12 = 1/6$, but $P(V)P(W) = 1/2 \cdot 3/12 = 1/8$, so V and W are dependent.

Many independent events

To say that events in a finite collection $\mathcal{E}$ are ***independent*** means that the multiplication rule

$$P(E_1 \cap E_2 \cap \cdots \cap E_k) = P(E_1)P(E_2) \cdots P(E_k)$$

holds for for every subset $\{E_1, E_2, \ldots, E_k\}$ of $\mathcal{E}$.

The classic example is in a game of many tosses, say n, of a fair coin. Let E_i be the event "get a head on toss i". The events $E_1, \ldots, E_n$ are independent.

3.5 Example problems

1. Three cards are dealt from a well-shuffled, standard 52 card deck. Find the probability that all three are diamonds.

   Solution: Working from right to left, let G be the event "first card is a diamond," let F be "second card is a diamond," and let E be "third card is a diamond." We wish to find $P(E \cap F \cap G)$. Using the multiplication rule, we have

$$P(E \cap F \cap G) = P(E|(F \cap G)) \cdot P(F \cap G).$$

   Using the multiplication rule again, we have

$$P(F \cap G) = P(F|G) \cdot P(G).$$

Putting this together, we get

$$P(E \cap F \cap G) = P(E|(F \cap G)) \cdot P(F|G) \cdot P(G).$$

The probabilities on the right are easy. Since there are 13 diamonds in the deck of 52 cards, we have $P(G) = 13/52 = 1/4$. If we know the first card is a diamond, there are 12 diamonds left among 51 cards, so $P(F|G) = 12/51$. If we know the first and second cards are diamonds, there are 11 diamonds left among 50 cards, so $P(E|(F \cap G)) = 11/50$. Putting these together (in reverse order), we get

$$P(E \cap F \cap G) = 13/52 \cdot 12/51 \cdot 11/50.$$

2. A fair die is rolled three times. Find the probability that one or more of the rolls are sixes.

   Solution: Let E be the event "one or more of the rolls are sixes." Notice that the opposite event E^c is "none of the rolls are sixes," which is the same thing as "all of the rolls are in the range one through five." Let A be the event "get a non-six on the first roll," let B be "get a non-six on the second roll," and let C be "get a non-six on the third roll," so that $E^c = A \cap B \cap C$. Using the opposite rule, we have

   $$P(E) = 1 - P(E^c) = 1 - P(A \cap B \cap C).$$

   Since A, B and C are independent (the rolls of the die do not know about or affect one another) we have,

   $$P(A \cap B \cap C) = P(A)P(B)P(C) = 5/6 \cdot 5/6 \cdot 5/6 = (5/6)^3.$$

   So we have our solution

   $$P(E) = 1 - (5/6)^3.$$

3. In five card poker, a *full house* is a collection of five cards of which three have the same face value and the remaining two have the same face value, but all five cards do not share the same face value. What is the probability that five cards dealt from a well-shuffled, standard 52 card deck is a full house?

   Solution (using counting techniques from §2): Let S be the set of all five card hands, and let E be the subset of S which is the set of full houses, so $P(E) = |E|/|S|$. The size of S is $|S| = \binom{52}{5}$. To count the size of E, we see that we can make a full house by the following sequence of choices: first, choose a face value for three cards; second, choose three suits for the three cards; third, choose a face value for the two remaining cards; last, choose two suits for the two cards. Clearly, there are 13 ways to make the first choice, $\binom{4}{3} = 4$ ways to make the second choice, 12 ways to make the third choice, and $\binom{4}{2} = 6$ ways to make the final choice. Thus we have a total of $13 \cdot 12 \cdot 4 \cdot 6 = 3744$ full houses, and so $P(E) = 3744/\binom{52}{5} \approx 0.144\%$, or about 1 chance in 700.

4. A sample space consists of 4 outcomes called a, b, c, d. Outcome a is twice as likely as outcome b, and outcomes b, c, d are equally likely. What is the probability of outcome a?

Solution: We know $P(a) + P(b) + P(c) + P(d) = 1$, that $P(a) = 2P(b)$, and that $P(b) = P(c) = P(d)$. Substituting, we get

$$P(a) + \frac{1}{2}P(a) + \frac{1}{2}P(a) + \frac{1}{2}P(a) = 1.$$

Solving for $P(a)$, we get $P(a) = 2/5$.

3.6 Exercises

1. In a family of four children, what is more likely: two boys and two girls, or three of one gender and one of the other?

2. A box contains five letters A, B, C, D, and E, all equally likely to be drawn at random. Two draws are made at random with replacement. (Drawing "with replacement" means that after the first draw, the item drawn in replaced and the box is reshuffled, so that all the letters have the same chance of being drawn on the second draw as they did on the first.)

   (a) Find the probability that the two letters drawn are different.

   (b) Find the probability that at least one of the letters drawn is a vowel.

   (c) Work the same probability problems as in parts (a) and (b) where the two draws are taken *without* replacement. ("Without replacement" means that the first item is *not* placed back in the box after the first draw.)

3. Four cards are dealt from a well-shuffled, standard 52 card deck.

   (a) Find the probability that none of the cards are diamonds.

   (b) Find the probability that all four cards are diamonds.

   (c) Find the probability that all four cards are the same suit.

4. A box contains 1 red and 5 green marbles; each marble is equally likely to be selected on a random draw from the box. You draw four marbles from the box at random.

   (a) Find the probability that a red appears exactly three times if the draws are made with replacement.

   (b) Find the probability that a red appears exactly three times if the draws are made without replacement.

5. Let S be the sample space of all possible outcomes for the experiment which is making 5 random draws, with replacement, from a jar containing 2 red marbles, 3 green marbles and 4 blue marbles. For each individual draw, all marbles have the same chance of being selected. Outcomes in S are strings of 5 letters using the letters R, G and B, where the order counts. For example, the outcome RGGBR indicates draws of red, green, green, blue, and red marbles, in that order.

   (a) How many outcomes are in S?

   (b) Let E be the event "draw exactly 2 reds in 5 draws." How many outcomes are in E?

(c) Is $P(E)$ equal to the number you found in (b) divided by the number you found in (a)? Why or why not? And if not, find $P(E)$. Give $P(E)$ as a decimal approximation correct to 3 significant digits.

(d) Let F be the event "draw exactly 2 blues in 5 draws." Are E and F independent? Show your calculations to support your answer. Give $P(F)$ and $P(E \cap F)$ as decimal approximations correct to 3 significant digits.

6. A fair coin is flipped 100 times.

   (a) Suppose you know that at least 99 of the flips are tails. What are the chances that all 100 are tails?

   (b) Suppose you know that the first 99 tosses are tails. What are the chances that all 100 are tails?

7. You play a lottery where you win if you correctly predict the six numbers that will be randomly chosen (without replacement) from the numbers 1–36. What are your chances of winning?

8. Four cards are dealt face down from a well-shuffled, standard 52 card deck.

   (a) Find the probability that the first two cards are red and the second two cards black.

   (b) Find the probability that exactly two of the cards are red.

9. 100 marbles are in a box; 54 are red, 46 are blue. Each marble has the same chance of being selected on a random draw from the box. Let E be the event "get exactly 3 reds in 5 random draws with replacement," and let F be the event "get a red on the first and last of 5 random draws with replacement."

   (a) Find $P(E)$.

   (b) Find $P(F)$.

   (c) Are E and F dependent or independent?

10. A fair die is rolled 5 times. Give the probabilities for each of the following events as a decimal approximation correct to 3 significant digits.

    (a) All rolls have an even number of spots.

    (b) It is not the case that all rolls have an even number of spots.

    (c) All rolls have an odd number of spots.

    (d) Exactly 3 of the rolls show an even number of spots.

11. Given $\Omega = \{a, b, c, d\}$, with $p(a) = 1/2$ and $p(b) = 1/4$. Find $p(c)$ and $p(d)$ if events $E = \{a, c\}$ and $F = \{c, d\}$ are independent.

12. Two cards are dealt from a well-shuffled standard 52-card deck.

    (a) What is the probability that either both cards are black or both cards are hearts?

    (b) What is the probability that either both cards are black or both cards are aces?

13. A card is drawn at random from a well-shuffled standard 52 card deck, replaced, then the deck is reshuffled. Imagine this procedure is repeated N times. What is the smallest value of N for which there is a 70% or better chance that a king was drawn at least once?

14. A sample space S consists of 3 outcomes a, b, and c. Events U and V in S are given by $U = \{a, b\}$ and $V = \{b, c\}$.

    (a) Suppose the 3 outcomes are equally likely. Are U and V dependent or independent? Explain.

    (b) If $p(a) = 1/2$, find values for $p(b)$ and $p(c)$ so that U and V are independent, or explain why this cannot be done.

    (c) If possible, give a complete list of events in S.

15. Independent events E and F in a sample space Ω have probabilities $P(E) = 2/3$ and $P(F) = 1/3$. Find the probability of the event $E \cup F$, or say why there is not enough information.

16. A fair coin is flipped ten times.

    (a) Which is more likely: (i) getting at least one head in the first five flips or (ii) getting at least two heads in all ten flips? Show your work.

    (b) Find the probability of getting exactly two heads on the first five flips and exactly four heads on the last five flips. Give a decimal approximation correct to three significant digits.

17. Five cards are dealt face down from a well shuffled standard 52 card deck.

    (a) Find the probability that all the cards are red.

    (b) Find the probability that the cards are, in order, the ace, king, queen, jack and ten of diamonds.

    (c) Find the probability that the five cards are the ace, king, queen, jack and ten of diamonds, in any order.

18. A fair coin is tossed 10 times. Find the probability that exactly 3 of the first five tosses are heads, while at least 1 of the last five tosses is a head. Calculate and give the probability as a decimal approximation correct to 2 decimal places.

4 Descriptive Statistics

In this section we present the basic vocabulary and properties of random variables on finite discrete probability spaces presented in the previous section.

4.1 Random Variables

A ***random variable*** is a function whose domain is a probability space. When the codomain is a set of numbers, a random variable is called ***quantitative***. When the codomain is something other than a set of numbers the random variable is called ***qualitative***. There are many familiar examples for which the probability space is a set of people. In this setting, a random variable is a characteristic of or measurement made on those people, such as: eye color, marital status, income, height, weight, and IQ score. The first two in the preceding list are qualitative, and the last four are quantitative. A more abstract example, but nonetheless useful, is a model for winning and losing money in roulette. To model one play of the game "bet \$1 on red", we use the sample space $\Omega = \{R, B, G\}$ with probabilities $P(R) = P(B) = 18/38$, and $P(G) = 2/38$. The random variable X for the net gain from the game is given by$X(R) = +\$1$ and $X(B) = -\$1 = X(G)$.

4.2 Any Data List is a Random Variable

The basic object of descriptive statistics is the ***data list***, which is simply an ordered list of numbers $X = (x_1, x_2, \ldots, x_N)$. In practice, data lists do not arise in a vacuum; there is a context that connects the data list to a set of concrete objects. For example, a list of heights, incomes, or test scores comes from a set of people. Thus height, income, and test score are functions whose domains are the set of people, and whose values are the entries of data lists. For every data list $X = (x_1, \ldots, x_N)$, there is a set $\Omega = \{\omega_1, \ldots, \omega_N\}$ such that $x_i = X(\omega_i)$ for $1 \leq i \leq N$. If Ω is not given explicitly, we can always choose $\Omega = \{1, 2, \ldots, N\}$. If there is no explicit probability function $P\colon \Omega \to \mathbf{R}$ on Ω, we assume the default uniform function given by $P(\omega) = 1/N$ for all $\omega \in \Omega$. In this way, we will always consider a data list to be a random variable.

4.3 Distribution Function, Percentile, and Histogram

Given a random variable $X\colon \Omega \to \mathbf{R}$ and a set A of real numbers, we use the shorthand "$X \in A$" to describe the event $X^{-1}(A) = \{\omega \in \Omega\colon X(\omega) \in A\}$. For example, if Ω is a population of people, and $X(\omega)$ is the height in centimeters of person ω, the event $X \leq 150$ is the set of all people whose height is 150 centimeters or less. For a single real number a, the event $X = a$ is the set $X^{-1}(\{a\})$, which is the set of all elements of the probability space whose X measurement is a. If X is a data list arising from measurements on a uniform probability space, then $P(X \in A)$ is the percent of the data list in the range A.

The ***distribution function of a random variable*** X, denoted $F_X : \mathbf{R} \to \mathbf{R}$, is defined by

$$F_X(a) = P(X \leq a)$$

for any number a. In words, $F_X(a)$ is the probability that the measurement made by X on a randomly selected object in Ω is less than or equal to a. The ***percentile rank*** of a number a is defined to be $100F_X(a)$. If p is the percentile rank of a, the number a is said to be p***th percentile***. For example, if we say that 1200 is the 80th percentile SAT score for a certain population, that means 80% of the population scored at or below 1200.

Given an interval $(u, v]$ of the real line, the probability $P(X \in (u, v])$ that X lies in the interval $(u, v]$ can be computed from the distribution function F_X as follows.

$$P(X \in (u, v]) = F_X(v) - F_X(u)$$

The quantity $\frac{P(X \in (u,v])}{v-u} = \frac{F_X(v)-F_X(u)}{v-u}$ is called the ***average probability density*** of X on $(u, v]$. Given a set of contiguous, nonoverlapping intervals

$$(x_0, x_1], (x_1, x_2], (x_2, x_3], \ldots, (x_{r-1}, x_r]$$

whose union contains the image $X(\Omega)$ of X, we define a piecewise constant ***average probability density function*** f by setting $f(x)$ equal to the average probability density of X on the interval in which x lies. The graph of f is called a ***histogram*** of X, and the chosen intervals are called the ***class intervals*** of the histogram. Some histograms use a choice of class intervals that reverses the left-open-right-closed convention. An example is academic grades, for which the "standard 10-point grade scale" class intervals

$$[0, 60), [60, 70), [70, 80), [80, 90), [90, 100]$$

are widely used.

Interpreting a histogram relies on the following key observation. Suppose that a, b are real numbers with $a < b$, that a is the left endpoint of a class interval, and b is the right endpoint of a class interval (not necessarily the same as for a). Then we have the following.

> The area under the histogram f over the interval $(a, b]$ is the probability that the random variable X is in that interval. In symbols,
>
> $$\int_a^b f(x)dx = P(a < X \leq b) = F_X(b) - F_X(a).$$

This formula suggests that f, F_X are a derivative-antiderivative pair of functions. This is in fact the case, but requires a technical extension of the introductory calculus theory of derivatives and integrals beyond the scope of these notes.

4.4 Expected Value, Variance, Standard Deviation

The ***expected value*** of a random variable $X : \Omega \to \mathbf{R}$, denoted $E(X)$, is defined to be

$$E(X) = \sum_{\omega \in \Omega} X(\omega)P(\omega) = \sum_{a \in X(\Omega)} aP(X = a) \tag{4.4.1}$$

This is also called the ***average*** or ***mean*** of X, and is also denoted $\mathrm{AVE}(X)$ and μ_X. The physical interpretation is this: at each value a in the image $X(\Omega)$ of

X, place a point-sized object with mass $P(X = a)$ on the (massless) real line. The expected value $E(X)$ is the balance point or center of mass of this system of masses.

Expected value has the following basic properties. Let $K\colon \Omega \to \mathbf{R}$ be a constant random variable given by $K(\omega) = k$, for some constant k, and let X, Y be two random variables on Ω. The following hold.

(4.4.2) **Properties of expected value.**

1. $E(K) = k$
2. $E(kX) = kE(X)$
3. $E(X + Y) = E(X) + E(Y)$

In words, property 1 says that if a random variable has only one value, then the average is that value. Property 2 says that if you rescale a random variable by multiplying by a constant, the average of the rescaled variable is the constant times the average of the original variable. Property 3 says that if you add two random variables, the average for the sum is the sum of the two original averages. One might be tempted to guess that $E(XY) = E(X)E(Y)$, however, this is not true in general (find an example!).

The mean measures the center of mass of the values of a random variable; the ***variance*** measures the extent to which values are spread out about the center of mass. The variance of a random variable X, denoted $\mathrm{var}(X)$, is defined to be

$$\mathrm{var}(X) = E((X - \mu_X)^2). \tag{4.4.3}$$

Since $X - \mu_X$ measures the distance from the mean of a value of X, a random variable whose values are clustered tightly about its mean has a smaller variance than does a random variable whose values are more spread out about its mean.

The ***standard deviation*** of X, denoted σ_X or $\mathrm{SD}(X)$, is the positive square root of $\mathrm{var}(X)$.

$$\sigma_X = \sqrt{\mathrm{var}(X)} \tag{4.4.4}$$

Here is an intuitive way to think of standard deviation. Suppose we play a game in which we randomly draw an item from a probability space S, then measure the item with the random variable X, and suppose we are asked to predict the distance between the value of X on a random pick and the mean $E(X)$. The (absolute size of the) the distance between $X(s)$ and $E(X)$ for a randomly chosen element s is in S is $|X(s) - E(X)| = |X(s) - \mu_X| = \sqrt{(X(s) - \mu_X)^2}$. Therefore the average distance of values of X from $E(X)$ is $E(\sqrt{(X - \mu_X)^2})$. If we could switch the order of the average with the square root in the last expression, we would have $\sqrt{E((X - \mu_X)^2)} = \mathrm{SD}(X)$. It turns out that the square root of the average is not exactly the same thing as the average of the square root, but they are close. This justifies describing the SD as a guess for the size of a "typical deviation from the mean" for a randomly chosen value of X.

As an exercise, the reader should verify the following formula for standard deviation and variance.

$$\sigma_X^2 = \mathrm{var}(X) = E(X^2) - E(X)^2 \tag{4.4.5}$$

It will be useful to have formulas for variance analogous to those above for expected value. The proofs are left as an exercise.

(4.4.6) **Properties of Variance.** Let X, Y be random variables on a sample space S, let k be a constant, and let K denote the random variable on S that has the same value k for every outcome.

1. $\text{var}(K) = 0$
2. $\text{var}(kX) = k^2\text{var}(X)$
3. $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2(E(XY) - E(X)E(Y))$
4. (special case of property 3) $\text{var}(X + K) = \text{var}(X)$

Property 1 says that a random variable whose values have no deviation at all from the mean has zero variance. Property 4 says that translating all the values of a random variable by the same constant value does not affect the amount of spread about the mean. Property 3 shows how variance is more complicated than expected value for the sum of two random variables. The quantity $E(XY) - E(X)E(Y)$ turns out to be a useful measure relating the two variables X and Y. We call this quantity the ***covariance*** of X and Y, and denote it by $\text{cov}(X, Y)$. It is commonly defined in two ways.

(4.4.7) $$\text{cov}(X, Y) = E(XY) - E(X)E(Y) \qquad \text{or}$$

(4.4.8) $$\text{cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y))$$

As an exercise, the reader should verify that the two definitions are equivalent.

4.5 Standard Units

Suppose we have a random variable X with average μ and standard deviation σ. For any number a, the number $x = a\sigma + \mu$ is said to measure *a standard units*. For example, if X is a random variable with $\mu = 10$ and $\sigma = 2$, then $a = 3$ standard units corresponds to the number $16 = 3{\cdot}2{+}10$. The value of a measures how far $x = a\sigma + \mu$ lies from the mean, relative to the standard deviation. To convert a number x to standard units, solve the equation $x = a\sigma + \mu$ for a. When $\sigma \neq 0$, we have $a = \dfrac{x - \mu}{\sigma}$. For the example above, to convert the number 7 to standard units, compute $a = \frac{7-10}{2} = -1.5$.

Converting all the values of a random variable (with $\sigma \neq 0$) to standard units makes the new random variable $X_{\text{std}} = \dfrac{X - \mu}{\sigma}$, called the *standardized* version of X.

Here are some convenient properties of standardized variables. Let S be a standardized variable (i.e., $S = X_{\text{std}}$ for some X). Then we have

(4.5.1) $$\mu_S = 0 \qquad \text{and} \qquad \sigma_S = 1.$$

The reader should prove this as an exercise.

4.6 Bernoulli Variables

In practical problems we often use a random variable which takes only the values 0 and 1. Such a random variable is called a ***Bernoulli variable*** or a *Bernoulli trial*.

Given a Bernoulli variable X with $p = P(X = 1)$ and $1 - p = P(X - 0$, the expected value of X is $E(X) = 0 \cdot (1 - p) + 1 \cdot p = p$. The variance of X is $E((X - \mu_X)^2) = E(X^2 - 2\mu_X X + \mu_X^2)$. Since $X = X^2$, and $\mu_X = p$, we have $\text{var}(X) = p - 2p + p^2 = p(1 - p)$. Thus the standard deviation of X is $\text{SD}(X) = \sqrt{p(1-p)} = \sqrt{P(X = 1)P(X = 0)}$.

4.7 Correlation and Regression

Given two random variables X and Y on the same domain sample space S, the ***scatter diagram for*** X ***and*** Y is the set of points

$$\{(X(s), Y(s)) \colon s \in S\}$$

in the x, y plane, The ***regression line*** for the scatter diagram is the line which minimizes the average of squares of vertical distances of all points in the scatter diagram to the line.

Here is how to find the regression line for two *standard* variables X and Y (recall that X and Y are standard means that $E(X) = E(Y) = 0$ and $\text{SD}(X) = \text{SD}(Y) = 1$). Let us denote by rms^2 the quantity to be minimized, and let $y = mx + b$ be a candidate for the regression line. We have

$$\begin{aligned} \text{rms}^2 &= E((Y - (mX + b))^2) \\ &= E(Y^2 - 2Y(mX + b) + (mX + b)^2) \\ &= E(Y^2 - 2YmX - 2Yb + m^2X^2 + 2mXb + b^2) \\ &= E(Y^2) - 2mE(XY) - 2bE(Y) + m^2E(X^2) + 2mbE(X) + E(b^2) \end{aligned}$$

These steps are all accomplished using the basic properties (4.4.2) of expected value. Now use the fact that X and Y are standard to simplify much further. Let us denote by r the quantity $E(X_{\text{std}}Y_{\text{std}})$. The reader should check that we get

$$\text{rms}^2 = m^2 - 2rm + b^2 + 1. \tag{4.7.1}$$

This quadratic in m has a minimum value when $m = r$ and $b = 0$. Thus the regression line for standard variables has the equation

$$y = rx \qquad \text{(for standard variables)} \tag{4.7.2}$$

and the rms error is given by

$$\text{rms} = \sqrt{1 - r^2} \qquad \text{(for standard variables).} \tag{4.7.3}$$

When X and Y are not necessarily already standard, we get the following equations for the regression line and rms.

$$y = r\left(\frac{x - \mu_X}{\sigma_X}\right)\sigma_Y + \mu_Y \tag{4.7.4}$$

(4.7.5) $$\text{rms} = \sqrt{1 - r^2}\, \sigma_Y$$

The number r is called the *correlation coefficient for X and Y* and is given by

(4.7.6) $$r = E(X_{\text{std}} Y_{\text{std}}) = E\left(\left(\frac{X - \mu_X}{\sigma_X}\right)\left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right).$$

The correlation coefficient can be expressed in terms of covariance.

(4.7.7) $$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\text{cov}(X, Y)}{\sqrt{\text{cov}(X, X)\text{cov}(Y, Y)}}$$

From this we see that the correlation r between two random variables X and Y is zero precisely when $\text{cov}(X, Y)$ is zero.

4.8 Independence of Random Variables

Two random variables X and Y are ***independent*** if

$$P(X = a \text{ and } Y = b) = P(X = a)P(Y = b)$$

for all pairs a, b or real numbers. Similar to independence for events, the practical interpretation is of independence for random variables is that knowing the value of one of them does not help you to predict the value of the other. Conversely, if X and Y are dependent, knowing the value of one of the variables should help you predict the value of the other.

Dependence between X and Y manifests itself in the scatter diagram in the form of recognizable pattern or trend. If the correlation r between X and Y is positive, then the scatter diagram must have an upward sloping trend from left to right. If r is negative, the scatter diagram has a downward sloping trend. In other words, if the correlation r is not zero, X and Y are dependent. It follows that if X and Y are independent, then their correlation, and hence their covariance, must be zero.

(4.8.1) **Correlation and covariance of independent random variables.** If X and Y are independent random variables, their correlation r and their covariance $\text{cov}(X, Y)$ are both zero.

We conclude with an important consequence of (4.8.1) which tells us how to combine individual variances (or standard deviations) for the sum of two independent random variables.

(4.8.2) **Variance and SD for the Sum of Independent Variables.**
If X and Y are independent random variables, then

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y).$$

Equivalently,

$$\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}.$$

To see why the first equation is true, combine (4.4.6) property (3) with (4.8.1). For the second equation, take square roots.

4.9 Exercises

1. (a) Verify equation (4.4.5).
   (b) Verify the equivalence of definitions (4.4.7) and (4.4.8) for covariance.
   (c) Verify (4.5.1).

2. Sketch a histogram for the data given in the table below. Use a density scale for the vertical axis.

Range	Count
0–10	5
10–50	20
50–200	16
200–500	4

3. Let $S = \{a, b, c, d, e\}$ be a sample space for a weighted box model and let X be a random variable on S, with values given in the table below.

element s of S	$p(s)$	$X(s)$
a	1/4	4
b	1/4	2
c	1/4	3
d	1/8	4
e	1/8	8

   (a) Find $E(X)$.
   (b) Find $\mathrm{SD}(X)$.
   (c) Sketch the distribution function F_X for X.
   (d) Sketch a histogram for X using class intervals 1 unit wide, centered on whole number values.

4. 100 marbles are in a box; 54 are red, 46 are blue. Each marble has the same chance of being selected on a random draw from the box. Let $\Omega = \{R, B\}$ be the sample space of the two possible outcomes for the game "one random draw." Let X be the random variable on Ω given by $X(R) = 2$, $X(B) = -1$.
   (a) Calculate $E(X)$.
   (b) Calculate $\mathrm{SD}(X)$.
   (c) Suppose the probabilities for R and B are altered by removing some of the red marbles. What is the least number of reds you would have to remove from the box in order to make $E(X)$ be negative?

5. Consider the list of numbers: 1, 10.
   (a) Assign weights to each number in the list so that the weighted average is 9, or explain why this cannot be done.
   (b) Assign weights to each number in the list so that the weighted average is 12, or explain why this cannot be done.

6. A company with 500 employees has recorded the number of years each worker has been with the company. The data is summarized in the table

below.

Years Employed	Number of Workers
0–1	50
1–2	75
2–5	125
5–10	150
10–20	100

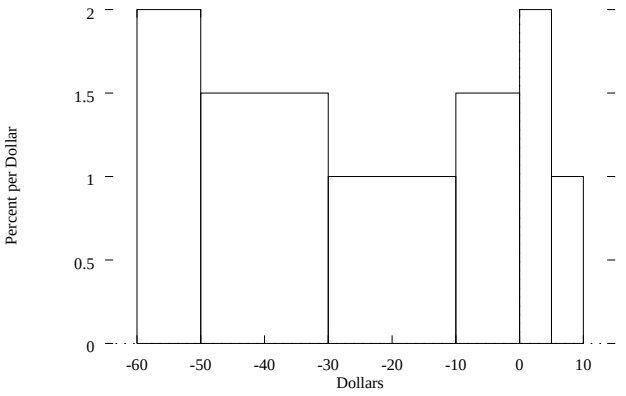
(a) Sketch a histogram for the data using a density scale for the vertical axis.

(b) Estimate the median number of years employment.

(c) Estimate the 25th percentile number of years employment.

(d) How does the average number of years employment compare with the median? Is there enough information to decide?

7. Below is a histogram showing the distribution of winnings (or loss) for 5000 bingo players at a recent bingo convention.



(a) Estimate the median amount of winnings.

(b) Estimate the 70th percentile amount of winnings.

(c) About how many bingo players won amounts in the range -30 to $+5$ dollars?

8. Sketch a histogram for the data in the table below. Use a density scale for the vertical axis.

Class Interval	**Frequency**
5–10	8
10–30	22
30–40	10

9. A slightly more general Bernoulli variable is a random variable that takes two values, say A, B, with $B > A$. Let X be such a variable, and let $p = P(X = B)$ and $q = 1 - p = P(X = A)$. Derive $E(X)$ and $\mathrm{SD}(X)$. Hint: note that $Y = (X - A)/(B - A)$ is a standard Bernoulli variable that takes values $0, 1$ with $p = P(Y = 1)$ and $q = P(Y = 0)$.

5 Sampling

5.1 Model for a Random Sample

A ***random sample of size*** n from a sample space Ω is an n-tuple $(\omega_1, \omega_2, \ldots, \omega_n)$ of elements from Ω. We think of ω_i as the result of the ith random draw in a sequence of n random draws from a box containing the elements of Ω. There are two basic ways to make such a sequence of draws. Sampling ***with replacement*** refers to a situation where each random draw starts with a "fresh copy" of Ω. Sampling ***without replacement*** implies that if ω is the item selected on the first draw, then ω is no longer eligible for subsequent draws; in other words, if the first k draws in a random sample without replacement are $\omega_1, \ldots, \omega_k$, then the sample space for the $(k+1)$st draw is $\Omega \setminus \{\omega_1, \ldots, \omega_k\}$. Formally, given a sample space Ω, a random sample with replacement of size n is an element of the set Ω^n. A random sample of without replacement of size n is an element of the set $\mathrm{Perm}(\Omega, n)$, which is the subset of elements $(\omega_1, \ldots, \omega_n)$ of Ω^n with no elements repeated, that is, we have $\omega_i \neq \omega_j$ for $i \neq j$.

Given the probability function $P\colon \Omega \to \mathbf{R}$ on Ω, the probability functions for sampling with and without replacement are, respectively, $P_{\mathrm{indep}}\colon \Omega^n \to \mathbf{R}$ and $P_{\mathrm{dep}}\colon \mathrm{Perm}(\Omega, n) \to \mathbf{R}$, given by

$$P_{\mathrm{indep}}((\omega_1, \omega_2 \ldots, \omega_n)) = P(\omega_1)P(\omega_2)\cdots P(\omega_n) \tag{5.1.1}$$

$$\begin{aligned} P_{\mathrm{dep}}((\omega_1, \omega_2 \ldots, \omega_n)) &= P(\omega_1)P(\omega_2|\{\omega_1\}^c)P(\omega_3|\{\omega_1, \omega_2\}^c) \\ &\quad \cdots P(\omega_n|\{\omega_1, \omega_2, \ldots, \omega_{n-1}\}^c). \end{aligned} \tag{5.1.2}$$

A ***random sample of size*** n of a random variable $X\colon \Omega \to \mathbf{R}$ is a collection of n random variables $X_i\colon \Omega^n \to \mathbf{R}$, $1 \leq i \leq n$, given by

$$X_i(\omega_1, \omega_2, \ldots, \omega_n) = X(\omega_i)$$

for $1 \leq i \leq n$. That is, variable X_i performs the measurement given by our random variable X on the outcome obtained on the ith random draw. We say that the random sample of X is ***with replacement*** or ***without replacement*** to indicate which of $P_{\mathrm{indep}}, P_{\mathrm{dep}}$ is used as the probability measure on Ω^n (respectively, $\mathrm{Perm}(\Omega, n)$). A random sample taken without replacement is also called a ***simple random sample***.

5.2 Sample Data, Statistics and Parameters

The term ***sample data*** or ***observed data*** refers to the values of the sample variables $X_1, X_2, \ldots, X_n$. Numbers computed from those values are called sample ***statistics***. Here are the most important sample statistics. The ***sample average***, denoted $\overline{X}$, is the random variable

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \cdots + X_n}{n}. \tag{5.2.1}$$

The ***sample standard deviation***, denoted s, is the random variable

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}. \tag{5.2.2}$$

As you might expect, the variable s^2 is called the ***sample variance***, and the variable $n\overline{X}$ is called the ***sample sum***.

Properties of the random variable X, such as $E(X)$ and $\mathrm{SD}(X)$, are called ***parameters***. The general goal of sampling theory is to use knowledge of parameters to make predictions about sample measurements, and conversely, to use sample statistics to make predictions about unknown parameters. In the remainder of this subsection, we present the fundamental formulas that relate statistics and parameters. We begin with a key observation about the variables X_i.

(5.2.3) **Sample variables are copies of X.** For each of the sample variables X_i, $1 \le i \le n$, the distribution of X_i is the same as the distribution of X. In particular, we have

$$\begin{aligned} E(X_i) &= E(X), \text{ and} \\ \mathrm{SD}(X_i) &= \mathrm{SD}(X). \end{aligned}$$

Furthermore, when the sample is taken with replacement, the variables X_i are independent.

The intuitive idea for (5.2.3) is simple. The variable X_i performs the measurement X on the ith draw in the sample $(\omega_1, \omega_2, \ldots, \omega_n)$. Since X_i "sees" only the ith draw, we can pretend that X_i is simply a random variable on the sample space Ω and not on all of Ω^n. Thus X_i looks like a "copy" of X. Since the draws are taken with replacement, the variables X_i are independent of one another.

Here are the essential formulas that relate sample statistics to parameters. For sample variables $X_1, X_2, \ldots, X_n$ (taken with replacement) of the random variable X, with sample average $\overline{X} = (1/n)\sum X_i$ and sample variance $s^2 = (1/n)\sum(X_i - \overline{X})$, we have the following.

$$E(\overline{X}) = E(X) \tag{5.2.4}$$

$$E\left(\sum_{i=1}^{n} X_i\right) = nE(X) \tag{5.2.5}$$

$$\mathrm{SD}(\overline{X}) = \frac{\mathrm{SD}(X)}{\sqrt{n}} \tag{5.2.6}$$

$$\mathrm{SD}\left(\sum_{i=1}^{n} X_i\right) = \sqrt{n}\,\mathrm{SD}(X) \tag{5.2.7}$$

$$E(s^2) = \left(\frac{n-1}{n}\right)\mathrm{var}(X) \tag{5.2.8}$$

Here is the derivation for (5.2.4).

$$\begin{aligned} E(\overline{X}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \qquad \text{(by (4.4.2))} \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X) \qquad \text{(by (5.2.3))} \\ &= E(X) \end{aligned}$$

The fact that $E(\overline{X}) = E(X)$ justifies our practice of using a value of the sample average to estimate the average value of the random variable.

Next we derive (5.2.6). To do this we first calculate $\text{var}(\overline{X})$, then take the square root. We use the fact (4.8.2) that the variance of a sum of independent variables is the sum of their variances.

$$\begin{aligned}
\text{var}(\overline{X}) &= \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\text{var}\left(\sum_{i=1}^{n} X_i\right) \qquad \text{(by (4.4.6) part (2))} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(X_i) \qquad \text{(by (4.8.2))} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(X) \qquad \text{(by (5.2.3))} \\
&= \frac{\text{var}(X)}{n}
\end{aligned}$$

Thus we have

$$\text{var}(\overline{X}) = \frac{\text{var}(X)}{n}.$$

Taking square roots yields (5.2.6).

Equation (5.2.6) says that the variance of the sample average decreases as the sample size n increases. This confirms our intuition that "the larger the sample, the higher the accuracy" of the sample average for estimating the population average $E(X)$.

Equations (5.2.5) and (5.2.7) follow immediately from (5.2.4) and (5.2.6) by multiplying both sides of the latter two equations by n. Equation (5.2.7) is called the "square root law" in the text *Statistics* by Friedman et al. Friedman refers to $\text{SD}(\overline{X})$ as the "SE for the average of the draws" and refers to $\text{SD}\left(\sum_{i=1}^{n} X_i\right)$ as the "SE for the sum of the draws" from a box model.

We conclude with a discussion of (5.2.8), the proof of which is outlined in exercise 1 below. Multiplying both sides by $n/(n-1)$ and bringing that constant inside the expected value on the left hand side yields

$$E(\frac{n}{n-1}s^2) = \text{var}(X).$$

This explains why we use $ns^2/(n-1)$ to estimate $\text{var}(X)$, or taking square roots, we use the quantity

$$s\sqrt{\frac{n}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}} \tag{5.2.9}$$

called the SD^+ of the sample[1] to estimate σ_X.

[1] In some statistics texts, the quantity (5.2.9) is called the *sample standard deviation*, and its square is called the *sample variance*. Regrettably, this requires anyone who uses the term "sample SD" to say which of (5.2.2) or (5.2.9) she means.

The discussion in this paragraph explains why we use a value of $\frac{n\,s^2}{n-1}$ to estimate $\mathrm{var}(X)$, or taking square roots, SD^+ to estimate the "population SD" σ_X.

5.3 Exercises

1. Complete the following proof of equation (5.2.8) by putting the lines into proper form using complete sentences, with justifications for each step.

   (i)

   $$\begin{aligned} s^2 &= \frac{1}{n}\sum_i (X_i - \overline{X})^2 \\ &= \frac{1}{n}\sum_i X_i^2 - \overline{X}^2 \end{aligned}$$

   (ii)

   $$E(s^2) = \frac{1}{n}\sum_i E(X_i^2) - E(\overline{X}^2)$$

   (iii)

   $$\begin{aligned} E(X_i^2) &= \mathrm{var}(X_i) + E(X_i)^2 = \mathrm{var}X + \mu^2 \\ E(\overline{X}^2) &= \mathrm{var}(\overline{X}) + E(\overline{X})^2 = \frac{\mathrm{var}X}{n} + \mu^2 \end{aligned}$$

   (iv)

   $$\begin{aligned} E(s^2) &= = \frac{1}{n} n(\mathrm{var}X + \mu^2) - \left(\frac{\mathrm{var}X}{n} + \mu^2\right) \\ &= \mathrm{var}X\left(1 - \frac{1}{n}\right) \\ &= \left(\frac{n-1}{n}\right)\mathrm{var}X \end{aligned}$$

6 Central Limit Theorem

This section gives an outline of a proof of a version of the Central Limit Theorem for finite sample spaces.

6.1 The normal distribution

Careful plots of the probability histograms for binomial distributions (say, for the number of heads in n tosses of a fair coin) suggest that there is some limit histogram as $n \to \infty$. Indeed this is exactly what de Moivre noticed in the mid-1700s. In the late 1700s and early 1800s, Laplace and Gauss derived the formula for this limit histogram curve, called the ***normal distribution***, the ***bell curve***, the the ***Laplacian*** distribution, or the ***Gaussian*** distribution. Here is a way to see how to find the curve that roughly parallels Gauss' published account.

Consider the probability histogram for the number of heads in n tosses of a fair coin. For a given number of heads m, the probability histogram has a rectangle R with a base of width 1, centered at m, and height $\binom{n}{m}\frac{1}{2^n}$, which is the probability of getting m heads in n tosses. Now let $x = x(m)$ denote the standardized version of m, that is,

$$x = \frac{m - E(m)}{\mathrm{SD}(m)} = \frac{m - n/2}{\sqrt{n}/2} = \frac{2m - n}{\sqrt{n}}.$$

In the histogram with standard units, the rectangle R transforms into R' with base of width $2/\sqrt{n}$ and height $\binom{n}{m}\frac{1}{2^n}\frac{\sqrt{n}}{2}$ (the total area remains the same).

Let $y = y(x)$ denote the unknown continuous limit curve of the histogram for coin tosses in standard units, that is,

$$y(x) = \lim_{n\to\infty} \binom{n}{m}\frac{1}{2^n}\frac{\sqrt{n}}{2}. \tag{6.1.1}$$

Notice that $m = m(x) = \frac{x\sqrt{n}+n}{2}$ is not constant, and worse, this value is not likely to be an integer. We could hope that

$$y(x) = \lim_{n\to\infty} \binom{n}{\lfloor \frac{x\sqrt{n}+n}{2} \rfloor}\frac{1}{2^n}\frac{\sqrt{n}}{2}.$$

might work, but even so, the above limit turns out to be resistant to direct attack. So we need an indirect method. Success comes, as it so often does, by considering rates of change. We will continue to work with the approximation $w_n(x)$ for $y(x)$ given by

$$w_n(x) = \binom{n}{\lfloor \frac{x\sqrt{n}+n}{2} \rfloor}\frac{1}{2^n}\frac{\sqrt{n}}{2}.$$

It is natural in science to examine *relative change* of a function, which is the quantity $\frac{\Delta y}{y} \approx \frac{y'\,dx}{y}$. This in turn motivates examining the function y'/y. The idea is that if you can discover a law of the form

$$\frac{y'}{y} = f(x)$$

then you can integrate to get a solution.

$$\ln y = \int f(x)\, dx + C$$
$$y = Ae^{\int f(x)\, dx}$$

We will show that the limit function y in (6.1.1) satisfies $y'/y = -x$ by showing that an approximation for the left side approaches x as $n \to \infty$, that is,

$$\lim_{n\to\infty} \frac{w_n'}{w_n} = -x. \tag{6.1.2}$$

First we need to approximate w_n' from it's discrete values. Using x values $x(u)$ and $x(u+1)$ (for some integer $u = \lfloor m(x) \rfloor$, so u depends on x and n) we have

$$w_n' \approx \frac{\Delta w_n}{\Delta x} = \frac{w_n(x(u+1)) - w_n(x(u))}{1/(\sqrt{n}/2)} \approx \frac{\left[\binom{n}{u+1} - \binom{n}{u}\right]}{2^n} \frac{n}{4}. \tag{6.1.3}$$

Then we have

$$\frac{w_n'}{w_n} \approx \frac{\binom{n}{u+1} - \binom{n}{u}}{\binom{n}{u}} \frac{\sqrt{n}}{2} = \frac{n-2u-1}{u+1} \frac{\sqrt{n}}{2}.$$

Now we substitute $u = \frac{x\sqrt{n}+n}{2}$ and we have (after simplifying)

$$\frac{w_n'}{w_n} \approx -\frac{xn+1}{n+x\sqrt{n}+2}$$

so we see that $w_n'/w_n \to -x$ as $n \to \infty$. Thus it is reasonable to suppose that if the limit curve y exists, then it must satisfy

$$\frac{y'}{y} = -x \tag{6.1.4}$$

from which we readily find the solution

$$y = Ae^{-x^2/2}. \tag{6.1.5}$$

A standard exercise in multivariable calculus shows that

$$\int_{-\infty}^{\infty} e^{-x^2} 2\, dx = \sqrt{2\pi}$$

so we have $A = 1\sqrt{2\pi}$. This is the normal curve.

6.2 Central Limit Theorem: statement and proof strategy

Informally speaking, the Central Limit Theorem says that sample sums and averages of any random variable have an approximately normal distribution when the sample size is large. More precise versions of the Central Limit Theorem say that the limit, as sample size increases, of the histogram for a sample sum or average (in standard units), approaches the histogram of the normal curve. Still more precise versions of the Central Limit Theorem give estimates for error in using the normal curve to estimate probabilities for sample sums and averages.

The preceding section develops an outline for the proof of the Central limit theorem for the random variable $X = 0, 1$ with uniform probability distribution. In the subsections that follow, we develop the following argument for the general case.

(i) To every random variable X, we associate a function $\psi_X\colon \mathbf{R} \to \mathbf{R}$, with the property that if $\psi_X \approx \psi_Y$, then the histograms for X and Y are close. It turns out that for this normal distribution this function is $e^{t^2/2}$.

(ii) Given a random variable X, let Y_n denote the average, in standard units, of a random sample of size n of X. That is, $Y_n = \frac{(X_1+\cdots+X_n)-\mu}{\sigma/\sqrt{n}}$, where $E(X) = \mu$ and $\mathrm{SD}(X) = \sigma$. We show that *no matter what X is*,

$$\lim_{n\to\infty} \psi_{Y_n}(t) = e^{t^2/2}.$$

(iii) We "conclude" that the distribution of any standardized random sample average for any random variable X tends to the normal distribution as the sample size grows.

Again we admit that this reasoning constitutes intuitive plausibility, but not a proof.

6.3 Moment generating functions

Given a random variable $X\colon \Omega \to \mathbf{R}$ and a function $f\colon \mathbf{R} \to \mathbf{R}$, the expression $f(X)$ denotes the random variable $f \circ X\colon \Omega \to \mathbf{R}$. An important and useful example is e^X. This building block is used to define the ***moment generating function*** $\psi_X\colon \mathbf{R} \to \mathbf{R}$ of a random variable X, given by

$$\psi_X(t) = E(e^{tX}). \tag{6.3.6}$$

The way to interpret the symbols on the right hand side is this. Given a real number t_0, the function t_0X is the random variable X rescaled by a factor of t_0. The expression e^{t_0X} is the random variable that sends $\omega \in \Omega$ to $e^{t_0X(\omega)}$, and finally, $\psi_X(t_0) = E(e^{t_0X})$ is the expected value of that latter random variable.

Given a random variable X, the expected value $E(X^k)$ of the kth power of X is called the ***kth moment of*** X. The following proposition explains the name of the moment generating function.

(6.3.7) **Proposition.** Let $X\colon \Omega \to \mathbf{R}$ be a random variable. Then we have

$$\psi_X^{(k)}(0) = E(X^k)$$

where $\psi_X^{(k)}$ denotes the kth derivative of ψ_X.

To see why the collection of moments $\{E(X^k)\}_{k\in\mathbf{N}}$ determines X, start with the case of a 2-element probability space $\Omega = \{a, b\}$, with probabilities $P(a) = p, P(b) = q$. If we know all the moments of X, we have equations

$$\begin{aligned} E(X) &= pX(a) + qX(b) \\ E(X^2) &= pX(a)^2 + qX(b)^2 \\ E(X^3) &= pX(a)^3 + qX(b)^3 \end{aligned}$$

and so on. Let $x = X(a), y = X(b)$. The first equation says (x, y) lies on a certain line. The second equation say (x, y) lies on a certain ellipse. The third constrains (x, y) to a cubic curve, and so on. It is intuitively clear that these

equations will uniquely determine the point (x, y), and therefore X. A similar arguments works for any finite Ω.

The following propositions lay the groundwork for using moment generating functions to prove the Central Limit Theorem.

(6.3.8) **Proposition.** If random variables $X, Y : \Omega \to \mathbf{R}$ are independent, then $e^{t_0 X}, e^{t_0 Y}$ are independent. It follows that

$$\psi_{X+Y}(t) = \psi_X(t)\psi_Y(t).$$

(6.3.9) **Proposition.** Let $X : \Omega \to \mathbf{R}$ be a random variable and let

$$X_1, \ldots, X_n : \Omega^n \to \mathbf{R}$$

be sample variables for a sample of size n of X. Then $\psi_{X_i}(t) = \psi_X$ for $1 \leq i \leq n$.

(6.3.10) **Corollary.** Let $X : \Omega \to \mathbf{R}$ be a random variable with $E(X) = \mu$ and $\mathrm{SD}(X) = \sigma$, let

$$X_1, \ldots, X_n : \Omega^n \to \mathbf{R}$$

be sample variables for a sample of size n of X, and let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample average. Let $Y_n = \sqrt{n}(\overline{X}_n - \mu)/\sigma$ be the standardized version of $\overline{X_n}$ (that is, Y_n has $E(Y_n) = 0$ and $\mathrm{SD}(Y_n) = 1$) and let $W = (X - \mu)/(\sigma\sqrt{n})$. We have the following.

$$\psi_{X_1+\cdots+X_n}(t) = (\psi_X(t))^n$$
$$\psi_{Y_n}(t) = (\psi_W(t))^n$$

(6.3.11) **Proposition.** Let $X, X_1, \ldots, X_n, Y_n, W$ be as above in the Corollary. We have

$$\lim_{n\to\infty} (\psi_W(t))^n = e^{t^2/2}.$$

Outline for (6.3.11). Using the facts that $E(W) = 0$, $E(W^2) = 1/n$, and $E(W^k) = E(X_{\mathrm{std}}^k)/n^{k/2}$ (where X_{std} denotes $(X-\mu)/\sigma$, which is X in standard units) we have

$$\begin{aligned}(\psi_W(t))^n &= (E(e^{tW}))^n \\ &= (E(1 + tW + \frac{t^2W^2}{2} + (\text{higher order terms})))^n \\ &= \left(1 + \frac{t^2}{2n} + (\text{higher order terms } \tfrac{t^k X_{\mathrm{std}}^k}{k! n^{k/2}})\right)^n.\end{aligned}$$

Now use the fact that

$$\lim_{n\to\infty} \left(1 + \frac{a}{n} + \mathcal{O}\left(\frac{1}{n^{3/2}}\right)\right)^n = e^a.$$

Solutions to Exercises

Note: Most of the "solutions" posted here are not solutions at all, but are merely final answer keys, although some are complete. These are provided so that you can check your work; reading the answer keys is not a substitute for working the problems yourself.

1.3 Sets and Functions Solutions

1.

$$\emptyset, \{a\}, \{b\}, \{c\}, \{d\}, \{a,b\}, \{a,c\}, \{a,d\}, \{b,c\}, \{b,d\}, \{c,d\},$$
$$\{a,b,c\}, \{a,b,d\}, \{a,c,d\}, \{b,c,d\}, \{a,b,c,d\}$$

2.

$$\begin{aligned} A \cup B &= \{1,2,3,4,5,6,7,9\} \\ A \cap B &= \{3,5\} \\ A \setminus B &= \{1,7,9\} \\ D \setminus A &= \{0,2,4,6,8\} \end{aligned}$$

(also show an Euler diagram)

3. (Venn diagram)

4. (Venn diagram)

5. (a) $A \times B = \{(a,x),(a,y),(a,z),(b,x),(b,y),(b,z)\}$
   (b) $B^2 = \{(x,x),(x,y),(x,z),(y,x),(y,y),(y,z),(z,x),(z,y),(z,z)\}$

6. (Euler diagram)

7.

$$\begin{aligned} A &= \{x\colon -2 \le x < 3\} = [-2,3) \\ B &= \{x\colon 1 < x \le 5\} = (1,5] \\ A \cup B &= \{x\colon -2 \le x \le 5\} = [-2,5] \\ A \cap B &= \{x\colon 1 < x < 3\} = (1,3) \\ A \setminus B &= \{x\colon -2 \le x \le 1\} = (-2,1) \\ R \setminus A &= \{x\colon x < -2 \text{ or } 3 \le x\} = (-\infty,-2) \cup [3,\infty) \end{aligned}$$

(also sketch intervals)

8. Let $f(x) = x^2$ and $g(x) = x+2$ define functions f and g from the reals to the reals.
   (a) $f(g(3)) = 25$
   (b) $g(f(3)) = 11$
   (c) $(g \cdot f)(3) = 45$
   (d) $(f/g)(3) = 9/5$

(e) $(3f + g)(3) = 32$

(f) $f(g(x)) = (x + 2)^2$

(g) $g(f(x)) = x^2 + 2$

9. There are six 1-1 correspondences, indicated below.

abc	abc	abc	abc	abc	abc
123	132	213	231	312	321

10. If either A or B is the empty set, then $A \times B$ is empty because there are no ordered pairs of the form (a, b) with a in A and b in B. Thus any subset of $A \times B$ is empty when one or both of A or B are empty. We check to see if the empty subset of $A \times B$ satisfies the definition of a function in two cases.

    (a) Suppose $A = \emptyset$ and $B = X \neq \emptyset$. We see that every element of A is the first coordinate of exactly one ordered pair of the empty set, so the empty set is indeed a function from A to B.

    (b) Suppose $A = X \neq \emptyset$ and $B = \emptyset$. Let a be an element of A. Since a is not the first coordinate of any ordered pair in the empty set, we conclude that the empty set is *not* a function from A to B, and hence that there are *no* functions from A to B.

11. (a) 120

    (b) $\sum_{i=1}^{15}(i^2 + 2i + 3)$

12. (a) 1

    (b) 8

    (c) -2

2.4 Counting Solutions

1. 132
2. 134
3. $8^2 \times 10^8 = 6.4$ billion
4. $20! \approx 2.4 \times 10^{18}$
5. $P(50, 20) \approx 1.15 \times 10^{32}$
6. $50^{20} \approx 9.54 \times 10^{33}$
7. $\binom{20}{11} = 167,960$

3.6 Probability Solutions

1. $P(\text{even gender split}) = \binom{4}{2}/16 = 6/16 = 3/8$
   $P(\text{3-1 gender split}) = 2\binom{4}{3}/16 = 8/16 = 1/2$, the greater of the two

2. (a)

$$\begin{aligned} & P(\text{2 draws different}) \\ = \; & P(\text{1st draw is a letter AND 2nd draw is something else}) \\ = \; & P(\text{1st draw is a letter})P(\text{2nd is something else}|\text{1st is something}) \\ = \; & 1 \cdot 4/5 = 4/5 \end{aligned}$$

   (b)

$$\begin{aligned} & P(\text{at least one vowel}) \\ = \; & 1 - P(\text{no vowels}) \\ = \; & 1 - P(\text{2 consonants}) \\ = \; & 1 - P(\text{1st draw consonant AND 2nd draw consonant}) \\ = \; & 1 - 3/5 \cdot 3/5 = 16/25 \end{aligned}$$

   (c) $P(\text{2 draws different}) = 1$,

$$\begin{aligned} & P(\text{at least one vowel}) \\ = \; & 1 - P(\text{no vowels}) \\ = \; & 1 - P(\text{2 consonants}) \\ = \; & 1 - P(\text{1st draw consonant AND 2nd draw consonant}) \\ = \; & 1 - P(\text{1st draw consonant})P(\text{2nd draw consonant}|\text{1st draw consonant}) \\ = \; & 1 - 3/5 \cdot 2/4 = 7/10 \end{aligned}$$

3. (a) $P(\text{no diamonds}) = 39/52 \cdot 38/51 \cdot 37/50 \cdot 36/49$
   (b) $P(\text{all diamonds}) = 13/52 \cdot 12/51 \cdot 11/50 \cdot 10/49$
   (c) $P(\text{all same suit}) = 4 \cdot (\text{above})$

4. (a) (with replacement) $P(\text{exactly 3 R in 4 draws}) = \binom{4}{3} \cdot (1/6)^3 \cdot 5/6$
   (b) (without replacement) $P(\text{exactly 3 R in 4 draws}) = 0$

5. (a) $3^5 = 243$
   (b) $\binom{5}{2} \cdot 2^3 = 80$
   (c) $P(E) = \binom{5}{2} \cdot (2/9)^2 \cdot (7/9)^3 \approx .232$, this does not equal $80/243 \approx .329$
   (d) $P(F) = \binom{5}{2} \cdot (4/9)^2 \cdot (5/9)^3 \approx .339$,
   $|E \cap F| = \binom{5}{2} \cdot \binom{3}{2} = 30$,
   $P(E \cap F) = 30 \cdot (2/9)^2 \cdot (4/9)^2 \cdot 3/9 \approx .098$,
   $P(E \cap F)$ does not equal $P(E)P(F)$, so E, F are dependent

6. (a) $P(\text{100 T}|\text{at least 99 T}) = 1/101$
   (b) $P(\text{100 T}|\text{first 99 T}) = 1/2$

7. $P(\text{win lottery}) = 1/\binom{36}{6}$

8. (a) $P(\text{RRBB}) = 26/52 \cdot 25/51 \cdot 26/50 \cdot 25/49$
   (b) $P(\text{exactly 2R}) = \binom{4}{2} \cdot (\text{above})$

9. (a) $P(E) = \binom{5}{3}(.54)^3(.46)^2$
   (b) $P(F) = (.54)^2$
   (c) Event $E \cap F$ consists of the 3 outcomes RRBBR, RBRBR, and RBBRR. Each of these outcomes has probability $(.54)^3(.46)^2$, so $P(E \cap F) = 3(.54)^3(.46)^2$, which does not equal $P(E)P(F)$, so we conclude that E and F are dependent.

10. (a) $P(\text{all rolls even}) = (1/2)^5 = 1/32 \approx .0313$
    (b) $P(\text{not all rolls even}) = 1 - 1/32 \approx .969$
    (c) $P(\text{all rolls odd}) = 1/32 \approx .0313$
    (d) $P(\text{exactly 3 even}) = \binom{5}{3}(1/2)^3(1/2)^2 \approx .313$

11. $P(c) = 1/6$, $P(d) = 1/12$

12. (a)

$$\begin{aligned} & P(\text{both black OR both hearts}) \\ = \; & P(\text{both black}) + P(\text{both hearts}) \\ = \; & 26/52 \cdot 25/51 + 13/52 \cdot 12/51 \end{aligned}$$

    (b)

$$\begin{aligned} & P(\text{both black OR both aces}) \\ = \; & P(\text{both black}) + P(\text{both aces}) - P(\text{both black aces}) \\ = \; & 26/52 \cdot 25/51 + 4/52 \cdot 3/51 - 2/52 \cdot 1/51 \end{aligned}$$

13.

$$\begin{aligned} P(\text{draw at least one K in N draws}) &= 1 - P(\text{draw no K in N draws}) \\ &= 1 - (12/13)^N \end{aligned}$$

So we want to solve $1 - (12/13)^N \geq .7$, or equivalently $.3 \leq (12/13)^N$. You can find N by trial and error, or solve as follows.

$$\begin{aligned} \log(.3) &\leq N \log(12/13) \quad (\text{take log both sides}) \\ \log(.3)/\log(12/13) &\geq N \quad (\text{divide both sides by } \log(12/13)) \end{aligned}$$

Get N slightly less than 16, so the smallest whole number solution is $N = 16$.

14. (a) Dependent. $P(U \cap V) = P(b) = 1/3$. This does not equal $P(U)P(V) = 2/3 \cdot 2/3 = 4/9$.
    (b) $P(b) = 1/2$ and $P(c) = 0$ makes U and V independent.
    (c) $\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}$

15.

$$\begin{aligned} P(E \cup F) &= P(E) + P(F) - P(E \cap F) \\ &= 2/3 + 1/3 - 2/3 \cdot 1/3 = 7/9 \end{aligned}$$

16. (a) $P(i) = 1 - P(\text{no heads}) = 1 - (1/2)^5 = 31/32$
$P(ii) = 1 - P(\text{zero H or one H}) = 1 - (1/2)^1 0 - \binom{10}{1}(1/2)^1 0 = 1013/1024$
$P(ii)$ is greater

    (b) $P(\text{exactly 2 H in first 5 flips AND exactly 4 H in last 5 flips}) = \binom{5}{2}(1/2)^5\binom{5}{4}(1/2)^5 = 50/1024 \approx .0488$

17. (a) $P(\text{all red}) = 26/52 \cdot 25/51 \cdot 24/50 \cdot 23/49 \cdot 22/48$

    (b) $P(\text{AKQJten diamonds in order}) = 1/(52 \cdot 51 \cdot 50 \cdot 49 \cdot 48)$

    (c) $P(\text{AKQJten diamonds in any order}) = 5! \cdot (\text{above})$

18. $\binom{5}{3}(1/2)^5(1 - (1/2)^5) = 310/(32^2) \approx .303$

4.9 Random Variables Solutions

1. (a)

$$\begin{aligned}
\sigma_X^2 &= \text{var}(X) = E((X - \mu_X)^2) \quad \text{(by definition)} \\
&= E(X^2 - 2X\mu_X + \mu_X^2) \quad \text{(squaring)} \\
&= E(X^2) - E(2X\mu_X) + E(\mu_X^2) \quad \text{(property 3 of (4.4.2))} \\
&= E(X^2) - 2\mu_X E(X) + \mu_X^2 \quad \text{(properties 1,2 of (4.4.2))} \\
&= E(X^2) - E(X)^2 \quad \text{(simplifying)}
\end{aligned}$$

    (b) (similar to above, expand and simplify using properties of expected value)

    (c) Let $S = (X - \mu_X)/\sigma_X$.

$$\begin{aligned}
\mu_S &= E((X - \mu_X)/\sigma_X) \\
&= (1/\sigma_X)E(X - \mu_X) \\
&= (1/\sigma_X)(E(X) - \mu_X) = 0 \\
\sigma_S &= \sqrt{\text{var}((X - \mu_X)/\sigma_X)} \\
&= \sqrt{(1/\sigma_X^2)\text{var}(X - \mu_X)} \\
&= \sqrt{(1/\sigma_X^2)\text{var}(X)} \\
&= \sqrt{\sigma_X^2/\sigma_X^2} = 1
\end{aligned}$$

2. The heights of the four blocks on a density scale are the following.

class interval	height of rectangle
0--10	(5/45)/(10 - 0) = approx .011 or 1.1%/unit
10--50	(20/45)/(50 - 10) = approx .011 or 1.1%/unit
50--200	(16/45)/(200 - 50) = approx .0024 or .24%/unit
200--500	(4/45)/(500 - 200) = approx .0003 or .03%/unit

3. (a) $E(X) = 2(1/4) + 3(1/4) + 4(3/8) + 8(1/8) = 3.75$

(b) $\mathrm{SD}(X) = \sqrt{1.75^2(1/4) + .75^2(1/4) + .25^2(3/8) + 4.25^2(1/8)} \approx 1.78$

(c) F_X is a step function with heights of steps (horizontal line segments) as follows.

interval	height of step
less than 2	0
[2,3)	1/4
[3,4)	1/2
[4,8)	7/8
8 and higher	1

(d) The histogram consists of four rectangles with bases and heights as follows.

base interval	height of rectangle
[1.5,2.5]	1/4
[2.5,3.5]	1/4
[3.5,4.5]	3/8
[7.5,8.5]	1/8

4. (a) $E(X) = (2)(.54) + (-1)(.46) = .62$

(b) $\mathrm{SD}(X) = \sqrt{E(X^2) - E(X)^2} \approx 1.4952$

(c) After removing n red marbles, we have

$$E(X) = 2\left(\frac{54-n}{100-n}\right) + (-1)\left(\frac{46}{100-n}\right) = \frac{62-n}{100-n}$$

so $E(X) < 0$ when $62 < 2n$. The smallest possible value of n to make this happen is $n = 32$.

5. (a) Call the weights x and y. Solve equations $x + 10y = 9$ and $x + y = 1$ to get $x = 1/9$, $y = 8/9$.

(b) This cannot be done becauase the weighted average of a list cannot exceed the maximum value in the list.

6. (a)

class interval	height of rectangle
0--1	10
1--2	15
2--5	8 1/3
5--10	6
10--20	2

(b) 5 yrs

(c) 2 yrs

(d) average is greater than the median since distribution has long right tail

7. (a), (b) and (c)
Reading the histogram from right to left, the six blocks contain 20, 30, 20, 15, 10, and 5 percent of the data. Thus the median is at −\$30, the 70th percentile at −\$10, and the range from −\$30 to +\$5 accounts for 45% or 2250 bingo players.

8. The block over 5–10 is 4 PPU (percent per unit) tall, the block over 10–30 is 2.75 PPU tall, and the block over 30–40 is 2.5 PPU tall.

9. You should get $E(X) = pB + qA = A + p(B - A)$, and $\mathrm{SD}(X) = (B - A)\sqrt{pq}$.