Lists, Decisions and Graphs

With an Introduction to Probability

Unit Fn: Functions

Edward A. Bender
S. Gill Williamson

Preface

The material in this unit of study was, over several years, presented by the authors to lower division undergraduates in the Department of Mathematics and the Department of Computer Science and Engineering at the University of California, San Diego (UCSD). All material has been classroom tested by the authors and other faculty members at UCSD.

The first course of a two quarter sequence was chosen from six units of study: Boolean Functions (Unit BF), Logic (Unit Lo), Number Theory and Cryptography (Unit NT), Sets and Functions (Unit SF), and Equivalence and Order (Unit EO)

The second course of the sequence was chosen from four units of study: Counting and Listing (Unit CL), Functions (Unit Fn), Decision Trees and Recursion (Unit DT), and Basic Concepts in Graph Theory (Unit GT).

The order of presentation of units within the first six, as well as those within the second four, can be varied for students with a good high school background in mathematics.

Discrete mathematics has become an essential tool in computer science, economics, biology, mathematics, chemistry, and engineering. Each area introduces its own special terms for shared concepts in discrete mathematics. The only way to keep from reinventing the wheel from area to area is to know the precise mathematical ideas behind the concepts being applied by these various fields. Our course material is dedicated to this task.

At the end of each unit is a section of multiple choice questions: **Multiple Choice Questions for Review**. These questions should be read before reading the corresponding unit, and they should be referred to frequently as the units are read. We encouraged our students to be able to work these multiple choice questions and variations on them with ease and understanding. At the end of each section of the units are exercises that are suitable for written homework, exams, or class discussion.

Table of Contents

Starred sections (*) indicate more difficult and/or specialized material.

Unit Fn

Functions

Section 1: Some Basic Terminology

Functions play a fundamental role in nearly all of mathematics. Combinatorics is no exception. In this section we review the basic terminology and notation for functions. Permutations are special functions that arise in a variety of ways in combinatorics. Besides studying them for their own interest, we'll see them as a central tool in other topic areas.

Except for the real numbers $\mathbb{R}$, rational numbers $\mathbb{Q}$ and integers $\mathbb{Z}$, our sets are normally finite. The set of the first n positive integers, $\{1, 2, \ldots, n\}$ will be denoted by $\underline{n}$.

Recall that $|A|$ is the number of elements in the set A. When it is convenient to do so, we linearly order the elements of a set A. In that case we denote the ordering by $a_1, a_2, \ldots, a_{|A|}$ or by $(a_1, a_2, \ldots, a_{|A|})$. Unless clearly stated otherwise, the ordering on a set of numbers is the numerical ordering. For example, the ordering on $\underline{n}$ is $1, 2, 3, \ldots, n$.

A review of the terminology concerning sets will be helpful. When we speak about sets, we usually have a "universal set" U in mind, to which the various sets of our discourse belong. Let U be a set and let A and B be subsets of U.

- The sets $A \cap B$ and $A \cup B$ are the *intersection* and *union* of A and B.
- The set $A \setminus B$ or $A - B$ is the *set difference* of A and B; that is, the set $\{x : x \in A, x \notin B\}$.
- The set $U \setminus A$ or A^c is the *complement* of A (relative to U). The complement of A is also written A' and $\sim A$.
- The set $A \oplus B = (A \setminus B) \cup (B \setminus A)$ is *symmetric difference* of A and B; that is, those x that are in *exactly one* of A and B. We have $A \oplus B = (A \cup B) \setminus (A \cap B)$.
- $\mathcal{P}(A)$ is the set of all subsets of A. (The notation for $\mathcal{P}(A)$ varies from author to author.)
- $\mathcal{P}_k(A)$ the set of all subsets of A of size (or cardinality) k. (The notation for $\mathcal{P}_k(A)$ varies from author to author.)
- The Cartesian product $A \times B$ is the set of all ordered pairs built from A and B:

$$A \times B = \{\, (a, b) \mid a \in A \text{ and } \ b \in B \,\}.$$

  We also call $A \times B$ the *direct product* of A and B.

If $A = B = \mathbb{R}$, the real numbers, then $\mathbb{R} \times \mathbb{R}$, written $\mathbb{R}^2$, is frequently interpreted as coordinates of points in the plane. Two points are the same if and only if they have the same coordinates, which says the same thing as our definition, $(a, b) = (a', b')$ if $a = a'$ and $b = b'$. Recall that the direct product can be extended to any number of sets. How can $\mathbb{R} \times \mathbb{R} \times \mathbb{R} = \mathbb{R}^3$ be interpreted?

Definition 1 (Function) *If A and B are sets, a function from A to B is a rule that tells us how to find a* **unique** *$b \in B$ for each $a \in A$. We write $f : A \to B$ to indicate that f is a function from A to B.*

Functions

We call the set A the domain of f and the set B the range[1] *or, equivalently, codomain of f. To specify a function completely you must give its domain, range and rule.*

The set of all functions from A to B is written B^A, for a reason we will soon explain. Thus $f\colon A \to B$ and $f \in B^A$ say the same thing.

In calculus you dealt with functions whose ranges were $\mathbb{R}$ and whose domains were contained in $\mathbb{R}$; for example, $f(x) = 1/(x^2 - 1)$ is a function from $\mathbb{R} - \{-1, 1\}$ to $\mathbb{R}$. You also studied functions of functions! The derivative is a function whose domain is all differentiable functions and whose range is all functions. If we wanted to use functional notation we could write $D(f)$ to indicate the function that the derivative associates with f.

Definition 2 (One-line notation) *When A is ordered, a function can be written in one-line notation as $(f(a_1), f(a_2), \ldots, f(a_{|A|}))$. Thus we can think of a function as an element of $B \times B \times \ldots \times B$, where there are $|A|$ copies of B. Instead of writing $B^{|A|}$ to indicate the set of all functions, we write B^A. Writing $B^{|A|}$ is incomplete because the domain A is not specified. Instead, only its size $|A|$ is given.*

Example 1 (Using the notation) To get a feeling for the notation used to specify a function, it may be helpful to imagine that you have an envelope or box that contains a function. In other words, this envelope contains all the information needed to completely describe the function. Think about what you're going to see when you open the envelope.

You might see

$$P = \{a, b, c\}, \qquad g\colon P \to \underline{4}, \qquad g(a) = 3, \quad g(b) = 1 \quad \text{and} \quad g(c) = 4.$$

This tells you that name of the function is g, the domain of g is P, which is $\{a, b, c\}$, and the range of g is $\underline{4} = \{1, 2, 3, 4\}$. It also tells you the values in $\underline{4}$ that g assigns to each of the values in its domain. Someone else may have put

$$g\colon \underline{4}^{\{a,b,c\}}, \qquad \text{ordering: } a, b, c, \qquad g = (3, 1, 4).$$

in the envelope instead. This describes the same function. It doesn't give a name for the domain, but we don't need a name like P for the set $\{a, b, c\}$ — we only need to know what is in the set. On the other hand, it gives an order on the domain so that the function can be given in one-line form. Can you describe other possible envelopes for the same function?

What if the envelope contained only $g = (3, 1, 4)$? You've been cheated! You *must* know the domain of g in order to known what g is. What if the envelope contained

$$\text{the domain of } g \text{ is } \{a, b, c\}, \qquad \text{ordering: } a, b, c, \qquad g = (3, 1, 4)?$$

We haven't specified the range of g, but is it necessary since we know the values of the function? Our definition included the requirement that the range be specified, so this is not a complete definition. On the other hand, in some discussions the range may not be important; for example, if $g = (3, 1, 4)$ all that may matter is that the range is large enough to contain 1, 3 and 4. In such cases, we'll be sloppy and accept this as if it were a complete specification. □

[1] *Some people define "range" to be the values that the function actually takes on. Most people call that the image, a concept we will discuss a bit later.*

Example 2 (Counting functions) By the Rule of Product, $\left|B^A\right| = |B|^{|A|}$. We can represent a subset S of A by a unique function $f: A \to \{0,1\}$ where $f(x) = 0$ if $x \notin S$ and $f(x) = 1$ if $x \in S$. This proves that there are $2^{|A|}$ such subsets. For example, if $A = \{a, b, d\}$, then the number of subsets of A is $2^{|\{a,b,d\}|} = 2^3 = 8$.

We can represent a multiset S formed from A by a unique function $f: A \to \mathbb{N} = \{0, 1, 2, \ldots\}$ where $f(x)$ is the number of times x appears in S. If no element is allowed to appear more than k times, then we can restrict the codomain of f to be $\{0, 1, \ldots, k\}$ and so there are $(k+1)^{|A|}$ such multisets. For example, the number of multisets of $A = \{a, b, d\}$ where each element can appear at most 4 times is $(4+1)^{|A|} = 5^3 = 125$. The particular multiset $\{a, a, a, d, d\}$ is represented by the function $f(a) = 3$, $f(b) = 0$ and $f(d) = 2$.

We can represent a k-list of elements drawn from a set B, with repetition allowed, by a unique function $f: \underline{k} \to B$. In this representation, the list corresponds to the function written in one-line notation. (Recall that the ordering on $\underline{k}$ is the numerical ordering.) This proves that there are exactly $|B|^k$ such lists. For example, the number of 4-lists that can be formed from $B = \{a, b, d\}$ is $|B|^4 = 3^4 = 81$. The 4-list (b, d, d, a) corresponds to the function $f = (b, d, d, a)$ in 1-line notation, where the domain is $\underline{4}$. □

Definition 3 (Types of functions) *Let $f: A \to B$ be a function. (Specific examples of these concepts are given after the definition.)*

- *If for every $b \in B$ there is an $a \in A$ such that $f(a) = b$, then f is called a surjection (or an onto function). Another way to describe a surjection is to say that it takes on each value in its range at least once.*
- *If $f(x) = f(y)$ implies $x = y$, then f is called an injection or a one-to-one function. Another way to describe an injection is to say that it takes on each value in its range at most once. The injections in $S^{\underline{k}}$ correspond to k-lists without repetitions.*
- *If f is both an injection and a surjection, it is a called a bijection.*
- *The bijections of A^A are called the permutations of A.*
- *If $f: A \to B$ is a bijection, we may talk about the inverse of f, written f^{-1}, which reverses what f does. Thus $f^{-1}: B \to A$ and $f^{-1}(b)$ is that unique $a \in A$ such that $f(a) = b$. Note that $f(f^{-1}(b)) = b$ and $f^{-1}(f(a)) = a$.*[2]

Example 3 (Types of functions) Let $A = \{1, 2, 3\}$ and $B = \{a, b\}$ be the domain and range of the function $f = (a, b, a)$. The function is a surjection because every element of the range is "hit" by the function. It is not an injection because a is hit twice.

Now consider the function g with domain B and range A given by $g(a) = 3$ and $g(b) = 1$. It is not a surjection because it misses 2; however, it is an injection because each element of A is hit at most once.

Neither f nor g is a bijection because some element of the range is either hit more than once or is missed. The function h with domain B and range $C = \{1, 3\}$ given by $h(a) = 3$ and $h(b) = 1$ is a bijection. At first, it may look like g and h are the same function. They

[2] *Do not confuse f^{-1} with $1/f$. For example, if $f: \mathbb{R} \to \mathbb{R}$ is given by $f(x) = x^3 + 1$, then $1/f(x) = 1/(x^3+1)$ and $f^{-1}(y) = (y-1)^{1/3}$.*

are not because they have different ranges. You can tell if a function is an injection without knowing its range, but you *must* know its range to decide if it is a surjection.

The inverse of the bijection h has domain C and range B it is given by $h^{-1}(1) = b$ and $h^{-1}(3) = a$.

The function f with domain and range $\{a, b, c, d\}$ given in 2-line form by

$$f = \begin{pmatrix} a & b & c & d \\ b & c & a & d \end{pmatrix}$$

is a permutation. You can see this immediately because the domain equals the range and the bottom line of the 2-line form is a rearrangement of the top line. The 2-line form is convenient for writing the inverse—just switch the top and bottom lines. In this example, $f^{-1} = \begin{pmatrix} b & c & a & d \\ a & b & c & d \end{pmatrix}$. □

Example 4 (Functions as relations) There is another important set-theoretic way of defining functions. Let A and B be sets. A *relation from A to B* is a subset of $A \times B$. For example:

If $A = \underline{3}$ and $B = \underline{4}$, then $R = \{(1,4), (1,2), (3,3), (2,3)\}$ is a relation from A to B.

If the relation R satisfies the condition that, for all $x \in A$ there is a *unique* $y \in B$ such that $(x, y) \in R$, then the relation R is called a *functional relation*. In the notation from logic, this can be written

$$\forall\, x \in A\ \exists !\, y \in B\ \ni\ (x, y) \in R.$$

This mathematical shorthand is well worth knowing:

- "$\forall$" means "for all",
- "$\exists$" means "there exists",
- "$\exists\,!$" means "there exists a unique", and
- "$\ni$" means "such that."

In algebra or calculus, when you draw a graph of a real-valued function $f : \mathbb{R} \to \mathbb{R}$ (such as $f(x) = x^3$), you are attempting a pictorial representation of the set $\{(x, f(x)) : x \in \mathbb{R}\}$, which is the subset of $\mathbb{R} \times \mathbb{R}$ that is the "functional relation from $\mathbb{R}$ to $\mathbb{R}$." In general, if $R \subset A \times B$ is a functional relation, then the function f corresponding to R has domain A and codomain B and is given by the ordered pairs $\{(x, f(x)) \mid x \in A\} = R$.

If you think of the "envelope game," Example 1, you will realize that a functional relation is yet another thing you might find in the envelope that describes a function. When a subset is defined it is formally required in mathematics that the "universal set" from which it has been extracted to form a subset also be described. Thus, in the envelope, in addition to R, you must also find enough information to describe completely $A \times B$. As you can see, a function can be described by a variety of different "data structures."

Given any relation $R \subseteq A \times B$, the inverse relation R^{-1} from B to A is defined to be $\{(y, x) : (x, y) \in R\}$. Recall the example in the previous paragraph where $A = \underline{3}$, $B = \underline{4}$, and

$R = \{(1,4),(1,2),(3,3),(2,3)\}$, The inverse relation is $R^{-1} = \{(4,1),(2,1),(3,3),(3,2)\}$. Notice that all we've had to do is reverse the order of the elements in the ordered pairs $(1,4),\ldots,(2,3)$ of R to obtain the ordered pairs $(4,1),\ldots,(3,2)$ of R^{-1}.

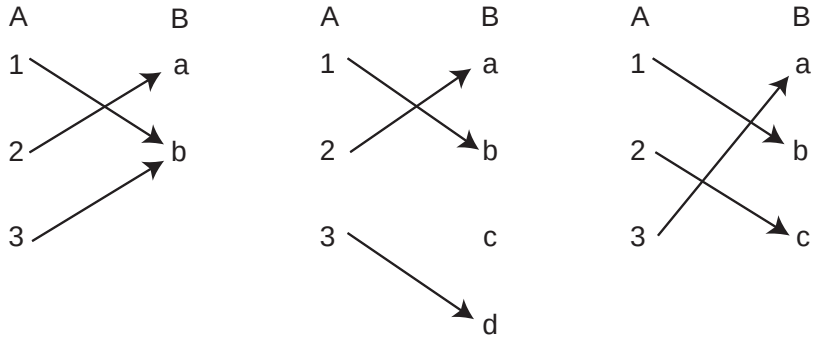
Note that neither R nor R^{-1} is a functional relation in the example in the previous paragraph. You should make sure you understand why this statement is true (Hint: R fails the "$\exists!$" test and R^{-1} fails the "$\forall$" part of the definition of a functional relation). Note also that if both R and R^{-1} are functional relations then $|A| = |B|$. In this case, R (and R^{-1}) are bijections in the sense of Definition 3. □

Example 5 (Two-line notation) Since one-line notation is a simple, brief way to specify functions, we'll use it frequently. If the domain is not a set of numbers, the notation is poor because we must first pause and order the domain. There are other ways to write functions which overcome this problem. For example, we could write $f(a) = 4$, $f(b) = 3$, $f(c) = 4$ and $f(d) = 1$. This could be shortened up somewhat to

$$a \to 4,\ b \to 3,\ c \to 4 \text{ and } d \to 1.$$

By turning each of these sideways, we can shorten it even more: $\begin{pmatrix} a & b & c & d \\ 4 & 3 & 4 & 1 \end{pmatrix}$. For obvious reasons, this is called *two-line notation*. Since x always appears directly over $f(x)$, there is no need to order the domain; in fact, we need not even specify the domain separately since it is given by the top line. If the function is a bijection, its inverse function is obtained by interchanging the top and bottom lines.

The arrows we introduced in the last paragraph can be used to help visualize different properties of functions. Imagine that you've listed the elements of the domain A in one column and the elements of the range B in another column to the right of the domain. Draw an arrow from a to b if $f(a) = b$. Thus the heads of arrows are labeled with elements of B and the tails with elements of A. Here are some arrow diagrams.



In all three functions, the domain $A = \{1,2,3\}$; however, the range B is different for each function. Since each diagram represents a function f, no two arrows have the same tail. If f is an injection, no two arrows have the same head. Thus the second and third diagrams are injections, but the first is not. If f is a surjection, every element of B is on the head of some arrow. Thus the first and third diagrams are surjections, but the second is not. Since the third diagram is both an injection and a surjection, it is a bijection. You should be able to describe the situation with the arrowheads when f is a bijection. How can you tell if a diagram represents a permutation? □

Exercises for Section 1

1.1. This exercise lets you check your understanding of the definitions. In each case below, some information about a function is given to you. Answer the following questions and give reasons for your answers:

- Have you been given enough information to specify the function?
- Can you tell whether or not the function is an injection? a surjection? a bijection?
- If possible, give the function in two-line form.

(a) $f \in \underline{3}^{\{>,<,+,?\}}, \quad f = (3, 1, 2, 3)$.

(b) $f \in \{>, <, +, ?\}^{\underline{3}}, \quad f = (?, <, +)$.

(c) $f \in \underline{4}^{\underline{3}}, \quad 2 \to 3, \quad 1 \to 4, \quad 3 \to 2$.

1.2. Let A and B be finite sets and $f\colon A \to B$. Prove the following claims. Some are practically restatements of the definitions, some require a few steps.

(a) If f is an injection, then $|A| \leq |B|$.

(b) If f is a surjection, then $|A| \geq |B|$.

(c) If f is a bijection, then $|A| = |B|$.

(d) If $|A| = |B|$, then f is an injection if and only if it is a surjection.

(e) If $|A| = |B|$, then f is a bijection if and only if it is an injection or it is a surjection.

1.3. Let S be the set of students attending a large university, let I be the set of student ID numbers for those students, let D be the set of dates for the past 100 years (month/day/year), let G be the set of 16 possible grade point averages between 2.0 and 3.5, rounded to the nearest tenth. For each of the following, decide whether or not it is a function. If it is, decide whether it is an injection, bijection or surjection. Give reasons for your answers.

(a) The domain is S, the codomain is I and the function maps each student to his or her ID number.

(b) The domain is S, the codomain is D and the function maps each student to his or her birthday.

(c) The domain is D, the codomain is I and the function maps each date to the ID number of a student born on that date. If there is more than one such student, the lexicographically least ID number is chosen.

(d) The domain is S, the codomain is G and the function maps each student to his or her GPA rounded to the nearest tenth.

(e) The domain is G, the codomains is I and the function maps each GPA to the ID number of a student with that GPA. If there is more than one such student, the lexicographically least ID number is chosen.

1.4. Let $A = \{1, 2, 3\}$ and $B = \{a, b, d\}$. Consider the following subsets of sets.

$$\{(3, a),\ (2, b),\ (1, a)\}, \qquad \{(1, a),\ (2, b),\ (3, c)\},$$
$$\{(1, a),\ (2, b),\ (1, d)\}, \qquad \{(1, a),\ (2, b)\ (3, d),\ (1, b)\}.$$

Which of them are relations on $A \times B$? Which of the are functional relations? Which of their inverses are functional relations?

Section 2: Permutations

Before beginning our discussion, we need the notion of composition of functions. Suppose that f and g are two functions such that the values f takes on are contained in the domain of g. We can write this as $f\colon A \to B$ and $g\colon C \to D$ where $f(a) \in C$ for all $a \in A$. We define the *composition* of g and f, written $gf\colon A \to D$ by $(gf)(x) = g(f(x))$ for all $x \in A$. The notation $g \circ f$ is also used to denote composition. Suppose that f and g are given in two-line notation by

$$f = \begin{pmatrix} p & q & r & s \\ P & R & T & U \end{pmatrix} \qquad g = \begin{pmatrix} P & Q & R & S & T & U & V \\ 1 & 3 & 5 & 2 & 4 & 6 & 7 \end{pmatrix}.$$

Then $gf = \begin{pmatrix} p & q & r & s \\ 1 & 5 & 4 & 6 \end{pmatrix}$. To derive $(gf)(p)$, we noted that $f(p) = P$ and $g(P) = 1$. The other values of gf were derived similarly.

The set of permutations on a set A is denoted in various ways in the literature. Two notations are $\mathrm{PER}(A)$ and $\mathcal{S}(A)$. Suppose that f and g are permutations of a set A. Recall that a permutation is a bijection from a set to itself and so it makes sense to talk about f^{-1} and fg. We claim that fg and f^{-1} are also permutations of A. This is easy to see if you write the permutations in two-line form and note that the second line is a rearrangement of the first if and only if the function is a permutation. You may want to look ahead at the next example which illustrates these ideas.

The permutation f given by $f(a) = a$ for all $a \in A$ is called the *identity permutation*. Notice that $f \circ f^{-1}$ and $f^{-1} \circ f$ both equal the identity permutation. You should be able to show that, if f is any permutation of A and e is the identity permutation of A, then $f \circ e = e \circ f = f$.

Again suppose that f is a permutation. Instead of $f \circ f$ or ff we write f^2. Note that $f^2(x)$ is not $(f(x))^2$. (In fact, if multiplication is not defined in A, $(f(x))^2$ has no meaning.) We could compose three copies of f. The result is written f^3. In general, we can compose k copies of f to obtain f^k. A cautious reader may be concerned that $f \circ (f \circ f)$ may not be the same as $(f \circ f) \circ f$. They are equal. In fact, $f^{k+m} = f^k \circ f^m$ for all nonnegative integers k and m, where f^0 is defined by $f^0(x) = x$ for all x in the domain. This is based

on the "associative law" which states that $f \circ (g \circ h) = (f \circ g) \circ h$ whenever the compositions make sense. We'll prove these results.

To prove that the two functions are equal, it suffices to prove that they take on the same values for all x in the domain. Let's use this idea for $f \circ (g \circ h)$ and $(f \circ g) \circ h$. We have

$$\begin{aligned}(f \circ (g \circ h))(x) &= f((g \circ h)(x)) && \text{by the definition of } \circ,\\ &= f(g(h(x))) && \text{by the definition of } \circ.\end{aligned}$$

Similarly

$$\begin{aligned}((f \circ g) \circ h)(x) &= (f \circ g)(h(x)) && \text{by the definition of } \circ,\\ &= f(g(h(x))) && \text{by the definition of } \circ.\end{aligned}$$

More generally, one can use this approach to prove by induction that $f_1 \circ f_2 \circ \cdots \circ f_n$ is well defined. This result then implies that $f^{k+m} = f^k \circ f^m$. Note that we have proved that the associative law for any three functions f, g and h for which the domain of f contains the values taken on by g and the domain of g contains the values taken on by h.

Example 6 (Composing permutations) We'll use the notation. Let f and g be the permutations

$$f = \begin{pmatrix}1&2&3&4&5\\2&1&4&5&3\end{pmatrix} \qquad g = \begin{pmatrix}1&2&3&4&5\\2&3&4&5&1\end{pmatrix}.$$

We can compute fg by calculating all the values. This can be done fairly easily from the two-line form: For example, $(fg)(1)$ can be found by noting that the image of 1 under g is 2 and the image of 2 under f is 1. Thus $(fg)(1) = 1$. You should be able to verify that

$$fg = \begin{pmatrix}1&2&3&4&5\\1&4&5&3&2\end{pmatrix} \qquad gf = \begin{pmatrix}1&2&3&4&5\\3&2&5&1&4\end{pmatrix} \neq fg$$

and that

$$f^2 = \begin{pmatrix}1&2&3&4&5\\1&2&5&3&4\end{pmatrix} \quad f^3 = \begin{pmatrix}1&2&3&4&5\\2&1&3&4&5\end{pmatrix} \qquad g^5 = f^6 = \begin{pmatrix}1&2&3&4&5\\1&2&3&4&5\end{pmatrix}.$$

Note that it is easy to get the inverse, simply interchange the two lines. Thus

$$f^{-1} = \begin{pmatrix}2&1&4&5&3\\1&2&3&4&5\end{pmatrix} \quad \text{which is the same as } f^{-1} = \begin{pmatrix}1&2&3&4&5\\2&1&5&3&4\end{pmatrix},$$

since the order of the columns in two-line form does not matter. □

Let f be a permutation of the set A and let $n = |A|$. If $x \in A$, we can look at the sequence

$$x, f(x), f(f(x)), \ldots, f^k(x), \ldots,$$

D which is often written as

$$x \to f(x) \to f(f(x)) \to \ldots \to f^k(x) \to \cdots.$$

Since the range of f has n elements, this sequence will contain a repeated element in the first $n+1$ entries. Suppose that $f^s(x)$ is the first sequence entry that is ever repeated and that $f^{s+p}(x)$ is the first time that it is repeated. Thus $f^s(x) = f^{s+p}(x)$. Apply $(f^{-1})^s$ to both sides of this equality to obtain $x = f^p(x)$ and so, in fact, $s = 0$. It follows that the sequence cycles through a pattern of length p forever since

$$f^{p+1}(x) = f(f^p(x)) = f(x), \quad f^{p+2}(x) = f^2(f^p(x)) = f^2(x), \quad \text{and so on.}$$

We call $(x, f(x), \ldots, f^{p-1}(x))$ the *cycle* containing x and call p the *length of the cycle*. If a cycle has length p, we call it a p-cycle.[3] Cyclic shifts of a cycle are considered the same; for example, if (1,2,6,3) is the cycle containing 1 (as well as 2, 3 and 6), then (2,6,3,1), (6,3,1,2) and (3,1,2,6) are other ways of writing the cycle (1,2,6,3). A cycle looks like a function in one-line notation. How can we tell them apart? Either we will be told or it will be clear from the context.

Example 7 (Using cycle notation) Consider the permutation

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 2 & 4 & 8 & 1 & 5 & 9 & 3 & 7 & 6 \end{pmatrix}.$$

Since $1 \to 2 \to 4 \to 1$, the cycle containing 1 is (1,2,4). We could equally well write it (2,4,1) or (4,1,2); however, (1,4,2) is a different cycle since it corresponds to $1 \to 4 \to 2 \to 1$ The usual convention is to list the cycle starting with its smallest element. The cycles of f are (1,2,4), (3,8,7), (5) and (6,9). We write f in *cycle form* as

$$f = (1,2,4)\ (3,8,7)\ (5)\ (6,9).$$

It is common practice to omit the cycles of length one and write $f = (1,2,4)(3,8,7)(6,9)$. The inverse of f is obtained by reading the cycles backwards because $f^{-1}(x)$ is the lefthand neighbor of x in a cycle. Thus

$$f^{-1} = (4,2,1)(7,8,3)(9,6) = (1,4,2)(3,7,8)(6,9). \qquad \square$$

Cycle form is useful in certain aspects of the branch of mathematics called "finite group theory." Here's an application.

[3] If $(a_1, a_2, \ldots, a_p)$ is a cycle of f, then

$$f(a_1) = a_2, \quad f(a_2) = a_3, \quad \ldots, \quad f(a_{p-1}) = a_p, \quad f(a_p) = a_1.$$

Example 8 (Powers of permutations) With a permutation in cycle form, it's very easy to calculate a power of the permutation. For example, suppose we want the tenth power of the permutation whose cycle form (including cycles of length 1) is $(1,5,3)(7)(2,6)$. To find the image of 1, we take ten steps: $1 \to 5 \to 3 \to 1 \cdots$. Where does it stop after ten steps? Since three steps bring us back to where we started (because 1 is in a cycle of length three), nine steps take us around the cycle three times and the tenth takes us to 5. Thus $1 \to 5$ in the tenth power. Similarly, $5 \to 3$ and $3 \to 1$. Clearly $7 \to 7$ regardless of the power. Ten steps take us around the cycle $(2,6)$ exactly five times, so $2 \to 2$ and $6 \to 6$. Thus the tenth power is $(1,5,3)(7)(2)(6)$. □

Suppose we have a permutation in cycle form whose cycle lengths all divide k. The reasoning in the previous example shows that the kth power of that permutation will be the identity permutation; that is, all the cycles will be 1-long and so every element is mapped to itself (i.e., $f(x) = x$ for all x). In particular, if we are considering permutations of an n-set, every cycle has length at most n and so we can take $k = n!$, regardless of the permutation. We have shown

Theorem 1 (A fixed power of n-permutations is the identity) *Given a set S, there are $k > 0$ depending on $|S|$ such that f^k is the identity permutation for every permutation f of S. Furthermore, $k = |S|!$ is one such k.*

***Example 9 (Involutions)** An *involution* is a permutation which is equal to its inverse. Since $f(x) = f^{-1}(x)$, we have $f^2(x) = f(f^{-1}(x)) = x$. Thus involutions are those permutations which have all their cycles of lengths one and two. How many involutions are there on $\underline{n}$? Let's count the involutions with exactly k 2-cycles and use the Rule of Sum to add up the results. We can build such an involution as follows:

- Select $2k$ elements for the 2-cycles AND
- partition these $2k$ elements into k blocks that are all of size 2 AND
- put the remaining $n - 2k$ elements into 1-cycles.

Since there is just one 2-cycle on two given elements, we can interpret each block as 2-cycle. This specifies f. The number of ways to carry out the first step is $\binom{n}{2k}$. For the second step, we might try the multinomial coefficient $\binom{2k}{2,\ldots,2} = (2k)!/2^k$. This is almost right! In using the multinomial coefficient, we're assuming an ordering on the pairs even though they don't have one. For example, with $k = 3$ and the set $\underline{6}$, there are just 15 possible partitions as follows.

$$\begin{array}{lll}
\{\{1,2\},\{3,4\},\{5,6\}\} & \{\{1,2\},\{3,5\},\{4,6\}\} & \{\{1,2\},\{3,6\},\{4,5\}\} \\
\{\{1,3\},\{2,4\},\{5,6\}\} & \{\{1,3\},\{2,5\},\{4,6\}\} & \{\{1,3\},\{2,6\},\{4,5\}\} \\
\{\{1,4\},\{2,3\},\{5,6\}\} & \{\{1,4\},\{2,5\},\{3,6\}\} & \{\{1,4\},\{2,6\},\{3,5\}\} \\
\{\{1,5\},\{2,3\},\{4,6\}\} & \{\{1,5\},\{2,4\},\{3,6\}\} & \{\{1,5\},\{2,6\},\{3,4\}\} \\
\{\{1,6\},\{2,3\},\{4,5\}\} & \{\{1,6\},\{2,4\},\{3,5\}\} & \{\{1,6\},\{2,5\},\{3,4\}\}
\end{array}$$

This is smaller than $\binom{6}{2,2,2} = 6!/2!2!2! = 90$ because all 3! ways to order the three blocks in each partition are counted differently to obtain the number 90. This is because we've

chosen a first, second and third block instead of simply dividing $\underline{6}$ into three blocks of size two.

How can we solve the dilemma? Actually, the discussion of what went wrong contains the key to the solution: The multinomial coefficient counts ordered collections of k blocks and we want unordered collections. Since the blocks in a partition are all distinct, there are $k!$ ways to order the blocks and so the multinomial coefficient counts each unordered collection $k!$ times. Thus we must simply divide the multinomial coefficient by $k!$. If this dividing by $k!$ bothers you, try looking at it this way. Let $f(k)$ be the number of ways to carry out the second step, partition the $2k$ elements into k blocks that are all of size 2. Since the k blocks can be permuted in $k!$ ways, the Rule of Product tells us that there are $f(k)\,k!$ ways to select k ordered blocks of 2 elements each. Thus $f(k)\,k! = \binom{2k}{2,\ldots,2}$.

Since there is just one way to carry out Step 3, the Rule of Product tells us that the number of involutions with exactly k 2-cycles is

$$\binom{n}{2k}\,\frac{1}{k!}\,\binom{2k}{2,\ldots,2} = \frac{n!}{(2k)!(n-2k)!}\,\frac{1}{k!}\,\frac{(2k)!}{(2!)^k}.$$

Simplifying and using the Rule of Sum to combine the various possible values of k, we obtain a formula for involutions.

We have just proved: The number of involutions of $\underline{n}$ is

$$\sum_{k=0}^{\lfloor n/2\rfloor} \frac{n!}{(n-2k)!2^k k!}$$

where $\lfloor n/2\rfloor$ denotes the largest integer less than or equal to $n/2$. Let's use this to compute the number of involutions when $n = 6$. Since $\lfloor 6/2\rfloor = 3$, the sum has four terms:

$$\begin{aligned}
&\frac{6!}{(6-0)!\,2^0 0!} + \frac{6!}{(6-2)!\,2^1 1!} + \frac{6!}{(6-4)!\,2^2 2!} + \frac{6!}{(6-6)!\,2^3 3!} \\
&\quad = \; 1 + \frac{6\times 5}{2} + \frac{6\times 5\times 4\times 3}{4\times 2!} + \frac{6\times 5\times 4}{8} \\
&\quad = \; 1 + 15 + 45 + 15 \;=\; 76.
\end{aligned}$$

The last term in the sum, namely $k = 3$ corresponds to those involutions with three 2-cycles (and hence no 1-cycles). Thus it counts the 15 partitions listed earlier in this example. □

If you're familiar with the basic operations associated with matrices, the following example gives a correspondence between matrix multiplication and composition of permutations.

***Example 10 (Permutation matrices)** Suppose f and g are permutations of $\underline{n}$. We can define an $n \times n$ matrix F to consist of zeroes except that the $(i,j)^{\text{th}}$ entry, $F_{i,j}$, equals one whenever $f(j) = i$. Define G similarly. Then

$$(FG)_{i,j} = \sum_{k=1}^{n} F_{i,k}G_{k,j} = F_{i,g(j)},$$

since $G_{k,j} = 0$ except when $g(j) = k$. By the definition of F, this entry of F is zero unless $f(g(j)) = i$. Thus $(FG)_{i,j}$ is zero unless $(fg)(j) = i$, in which case it is one. We've proven that FG corresponds to fg. In other words:

Composition of permutations corresponds to multiplication of matrices.

It is also easy to prove that f^{-1} corresponds to F^{-1}. Using this correspondence, we can prove things such as $(fg)^{-1} = g^{-1}f^{-1}$ and $(f^k)^{-1} = (f^{-1})^k$ by noting that they are true for matrices F and G.

As an example, let f and g be the permutations

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 4 & 5 & 3 \end{pmatrix} \qquad g = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 4 & 5 & 1 \end{pmatrix}.$$

We computed fg in Example 6. We obtained

$$fg = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 4 & 5 & 3 & 2 \end{pmatrix}.$$

Using our correspondence, we obtain

$$F = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \qquad G = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

You should multiply these two matrices together and verify that you get the matrix FG corresponding to

$$fg = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 4 & 5 & 3 & 2 \end{pmatrix}.$$ □

***Example 11 (Derangements)** A *derangement* is a permutation f with no fixed points; i.e., $f(x) \neq x$ for all x. We first show that the probability is $1/n$ that a permutation f, selected uniformly at random from all permutations of $\underline{n}$, has $f(k) = k$. If $f(k) = k$, then the elements of $\underline{n} - \{k\}$ can be permuted in any fashion. This can be done in $(n-1)!$ ways. Thus, $(n-1)!$ is the cardinality of the set of all permutations of $\underline{n}$ that satisfy $f(k) = k$. Since there are $n!$ permutations, the probability that $f(k) = k$ is $(n-1)!/n! = 1/n$. Hence the probability that $f(k) \neq k$ is $1 - 1/n$. If we toss a coin with probability p of heads for n tosses, the probability that no heads occurs in n tosses is $(1-p)^n$. This is because each toss is "independent" of the prior tosses. If we, incorrectly, treat the n events $f(1) \neq 1, \ldots, f(n) \neq n$ as independent in this same sense, the probability that $f(k) \neq k$ for $k = 1, \ldots, n$, would be $(1-1/n)^n$. One of the standard results in calculus is that $(1-1/n)^n$ approaches $1/e$ as $n \to \infty$. (You can prove it by writing $(1-1/n)^n = \exp(\ln(1-1/n)/(1/n))$, setting $1/n = x$ and using l'Hôpital's Rule.) Thus, we might expect approximately $n!/e$ derangements of $\underline{n}$ for large n. Although our argument is wrong, the result is right! We get partial credit for this example. □

Exercises for Section 2

2.1. This exercise lets you check your understanding of cycle form. A permutation is given in one-line, two-line or cycle form. Convert it to the other two forms. Give its inverse in all three forms.

(a) (1,5,7,8) (2,3) (4) (6).

(b) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 8 & 3 & 7 & 2 & 6 & 4 & 5 & 1 \end{pmatrix}$.

(c) (5,4,3,2,1), which is in one-line form.

(d) (5,4,3,2,1), which is in cycle form.

2.2. A carnival barker has four cups upside down in a row in front of him. He places a pea under the cup in the first position. He quickly interchanges the cups in the first and third positions, then the cups in the first and fourth positions and then the cups in the second and third positions. This entire set of interchanges is done a total of five times. Where is the pea?
Hint: Write one entire set of interchanges as a permutation in cycle form.

2.3. Let f be a permutation of $\underline{n}$. The cycle of f that contains 1 is called the *cycle generated by* 1.

(a) Prove that the number of permutations in which the cycle generated by 1 has length n is $(n-1)!$.

(b) For how many permutations does the cycle generated by 1 have length k? (Remember that a permutation must be defined on *all* elements of its domain $\underline{n}$.)

(c) If your answer to (b) is correct, when it is summed over $1 \leq k \leq n$ it should equal $n!$, the total number of permutations of $\underline{n}$. Why? Show that your answer to (b) has the correct sum.

2.4. This exercise deals with powers of permutations. All our permutations will be written in cycle form.

(a) Compute $(1,2,3)^{300}$.

(b) Compute $\big((1,3)(2,5,4)\big)^{300}$.

(c) Show that for every permutation f of $\underline{5}$, we have f^{60} is the identity permutation. What is f^{61}?

Section 3: Other Combinatorial Aspects of Functions

This section contains two independent parts. The first deals with the concept of the inverse of a general function. The second deals with functions related to computer storage of unordered lists.

The Inverse of an Arbitrary Function

Again, let $f: A \to B$ be a function. The *image* of f is the set of values f actually takes on: $\text{Image}(f) = \{\, f(a) \mid a \in A \,\}$. The definition of a surjection can be rewritten $\text{Image}(f) = B$ because a surjection was defined to be a function $f : A \to B$ such that, for every $b \in B$ there is an $a \in A$ with $f(a) = b$.

For each $b \in B$, the *inverse image* of b, written $f^{-1}(b)$ is the set of those elements in A whose image is b; i.e.,

$$f^{-1}(b) = \{\, a \mid a \in A \text{ and } f(a) = b \,\}.$$

This extends our earlier definition of f^{-1} from bijections to all functions; however, such an f^{-1} can't be thought of as a function from B to A unless f is a bijection because it will not give a unique $a \in A$ for each $b \in B$.[4]

Suppose f is given by the functional relation $R \subset A \times B$. Then $f^{-1}(b)$ is all those a such that $(a, b) \in R$. Equivalently, $f^{-1}(b)$ is all those a such that $(b, a) \in R^{-1}$.

Definition 4 (Coimage) *Let $f: A \to B$ be a function. The collection of nonempty inverse images of elements of B is called the coimage of f. In mathematical notation*

$$\text{Coimage(f)} = \{f^{-1}(b) \mid b \in B, f^{-1}(b) \neq \emptyset\} = \{f^{-1}(b) \mid b \in \text{Image}(f)\}.$$

We claim that the coimage of f is the partition of A whose blocks[5] are the maximal subsets of A on which f is constant. For example, if $f \in \{a, b, c\}^{\underline{5}}$ is given in one line form as (a, c, a, a, c), then

$$\text{Coimage}(f) = \{f^{-1}(a), f^{-1}(c)\} = \big\{\{1, 3, 4\}, \{2, 5\}\big\},$$

f is a on $\{1, 3, 4\}$ and is c on $\{2, 5\}$.

We now prove the claim. If $x \in A$, let $y = f(x)$. Then $x \in f^{-1}(y)$ and the set $f^{-1}(y)$ is an element of $\text{Coimage}(f)$. Hence the union of the nonempty inverse images contains A. Clearly it does not contain anything which is not in A. If $y_1 \neq y_2$, then we cannot have $x \in f^{-1}(y_1)$ and $x \in f^{-1}(y_2)$ because this would imply $f(x) = y_1$ and $f(x) = y_2$, a contradiction of the definition of a function. Thus $\text{Coimage}(f)$ is a partition of A. Since the value of $f(x)$ determines the block to which x belongs, x_1 and x_2 belong to the same block if and only if $f(x_1) = f(x_2)$. Hence a block is a maximal set on which f is constant.

---

[4] There is a slight abuse of notation here: If $f: A \to B$ is a bijection, our new notation is $f^{-1}(b) = \{a\}$ and our old notation is $f^{-1}(b) = a$.

[5] Recall that a partition of a set S is an unordered collection of disjoint nonempty subsets of S whose union is S. These subsets are called the blocks of the partition.

Example 12 (f^{-1} as a function) Let $f: A \to B$ be a function. For each $b \in B$,

$$f^{-1}(b) = \{\, a \mid a \in A \text{ and } f(a) = b \,\}.$$

Thus, for each $b \in B$, $f^{-1}(b) \in \mathcal{P}(A)$. Hence f^{-1} is a function with domain B and range (codomain) $\mathcal{P}(A)$, the set of all subsets of A. This is true for any function f and does not require f to be bijection. For example, if $f \in \{a, b, c\}^{\underline{5}}$ is given in one-line form as (a, c, a, a, c), then, f^{-1}, in two-line notation is

$$\begin{pmatrix} a & b & c \\ \{1,3,4\} & \emptyset & \{2,5\} \end{pmatrix}$$

If, we take the domain of f^{-1} to be Image(f), instead of all of B, then f^{-1} is a bijection from Image(f) to Coimage(f). In the case of our example (a, c, a, a, c), we get, in two-line notation

$$\begin{pmatrix} a & c \\ \{1,3,4\} & \{2,5\} \end{pmatrix}$$

for the image–coimage bijection associated with f^{-1}. If we are only given the coimage of a function then we don't have enough information to specify the function. For example, suppose we are given only that $\{\{1,3,4\},\{2,5\}\}$ is the coimage of some function g with codomain $\{a, b, c\}$. We can see immediately that the domain of g is $\underline{5}$. But what is g? To specify g we need to know the elements x and y in $\{a, b, c\}$ that make

$$\begin{pmatrix} x & y \\ \{1,3,4\} & \{2,5\} \end{pmatrix}$$

the correct two-line description of g^{-1} (restricted to its image). There are $(3)_2 = 6$ choices[6] for xy, namely, ab, ac, bc, ba, ca, and cb. In general, suppose $f: A \to B$ and we are given that a particular partition of A with k blocks is the coimage of f. Then, by comparison with our example ($A = \underline{5}$, $B = \{a, b, c\}$), it is easy to see that there are exactly $(|B|)_k$ choices for the function f. □

We can describe the image and coimage of a function by the arrow pictures introduced in Example 5. Image(f) is the set of those $b \in B$ which appear as labels of arrowheads. A block in Coimage(f) is the set of labels on the tails of those arrows that all have their heads pointing to the same value; for example, the block of Coimage(f) arising from $b \in$ Image(f) is the set of labels on the tails of those arrows pointing to b.

Example 13 (Counting functions with specified image size) How many functions in B^A have an image with exactly k elements? You will need to recall that the symbol $S(n, k)$, stands for the number of partitions of set of size n into k blocks. (The $S(n, k)$ are called the Stirling numbers of the second kind and are discussed in Unit CL. See the index for page numbers.) If $f \in B^A$ has k elements in its image, then this means that the coimage of f is a partition of A having exactly k blocks. Suppose that $|A| = a$ and $|B| = b$. There

[6] Recall that $(n)_k = n(n-1)\cdots(n-k+1) = n!/(n-k)!$ is the number of k-lists without repeats that can be made from an n-set.

are $S(a,k)$ ways to choose the blocks of the coimage. The partition of A does not fully specify a function $f \in B^A$. To complete the specification, we must specify the image of the elements in each block (Example 12). In other words, an injection from the set of k blocks to B must be specified. This is an ordered selection of size k without replacement from B. There are $(b)_k = b!/(b-k)!$ such injections, independent of which k block partition of A we are considering. By the Rule of Product, there are $S(a,k)(b)_k$ functions $f \in B^A$ with $|\text{Image(f)}| = k$. For example, when the domain is $\underline{5}$ and the range is $\{a,b,c\}$, the number of functions with $|\text{Image(f)}| = 2$ is $S(5,2)(3)_2 = 15 \times 6 = 90$, where the value of $S(5,2)$ was obtained from the table in the discussion of Stirling numbers in Unit CL. Example 12 gave one of these 90 possibilities.

We consider some special cases.

- Suppose $k = a$.
  - If $b < a$, there are no functions f with $|\text{Image(f)}| = a$ because the size a of the image is at most the size b of the codomain.
  - If $b \geq a$ there are $(b)_a$ functions with $|\text{Image(f)}| = a$.
  - If $b = a$, the previous formula, $(b)_a$, reduces to $a!$ and the functions are injections from A to B.
- Suppose $k = b$.
  - If $b > a$ there are no functions f with $|\text{Image(f)}| = b$ because the size of the image is at most the size of the domain.
  - If $b \leq a$ then there are $S(a,b)(b)_b = S(a,b)\, b!$ functions $f \in B^A$ with $|\text{Image(f)}| = b$. These functions are exactly the surjections. □

Monotonic Lists and Unordered Lists

In computers, all work with data structures requires that the parts of the data structure be ordered. The most common orders are arrays and linked lists.

Sometimes the order relates directly to an order associated with the corresponding mathematical objects. For example, the one-line notation for a function is simply an ordered list, which is an array. Thus there is a simple correspondence (i.e., bijection) between lists and functions: A k-list from S is a function $f\colon \underline{k} \to S$. Thus functions (mathematical objects) are easily stored as ordered lists (computer objects).

Sometimes the order is just an artifact of the algorithm using the structures. In other words, the order is imposed by the designer of the algorithm. Finding such a "canonical" ordering[7] is essential if one wants to work with unordered objects efficiently in a computer.

[7] In mathematics, people refer to a unique thing (or process or whatever) that has been selected as *canonical*.

Since sets and multisets[8] are basic unordered mathematical objects, it is important to have ways of representing them in a computer. We'll discuss a canonical ordering for k-sets and k-multisets whose elements lie in an n-set.

We need to think of a unique way to order the set or multiset, say $s_1, s_2, \ldots, s_k$ so that we have an ordered list. (A mathematician would probably speak of a canonical ordering of the multiset rather than a unique ordering; however, both terms are correct.)

Let's look at a small example, the 3-element multisets whose elements are chosen from $\underline{5}$. Here are the $\binom{5+3-1}{3} = 35$ such multisets.[9] An entry like 2,5,5 stands for the multiset containing one 2 and two 5's.

1,1,1	1,1,2	1,1,3	1,1,4	1,1,5	1,2,2	1,2,3	1,2,4	1,2,5	1,3,3
1,3,4	1,3,5	1,4,4	1,4,5	1,5,5	2,2,2	2,2,3	2,2,4	2,2,5	2,3,3
2,3,4	2,3,5	2,4,4	2,4,5	2,5,5	3,3,3	3,3,4	3,3,5	3,4,4	3,4,5
3,5,5	4,4,4	4,4,5	4,5,5	5,5,5					

We've simply arranged the elements in each 3-multiset to be in "weakly increasing order." Let $(b_1, b_2, \ldots, b_k)$ be an ordered list. We say the list is in *weakly increasing order* if the values are not decreasing as we move from one element to the next; that is, if $b_1 \le b_2 \le \cdots \le b_k$. The list of lists we've created can be thought of as a bijection from

(i) the 3-multisets whose elements lie in $\underline{5}$ to

(ii) the weakly increasing functions in $\underline{5}^{\underline{3}}$ written in one-line notation.

Thus, 3-multisets with elements in $\underline{5}$ correspond to weakly increasing functions in $\underline{5}^{\underline{3}}$. For example the multiset $\{2, 5, 5\}$ corresponds to the weakly increasing function $f = (2, 5, 5)$ in 1-line form.

Since we have seen that functions with domain $\underline{k}$ can be viewed as k-lists, we say that $f \in \underline{n}^{\underline{k}}$ is a *weakly increasing function* if its one-line form is weakly increasing; that is, $f(1) \le f(2) \le \cdots \le f(k)$. In a similar fashion we say that the list $b_1, b_2, \ldots, b_k$ is in

$$\left.\begin{array}{c}\textit{weakly decreasing}\\ \textit{strictly decreasing}\\ \textit{strictly increasing}\end{array}\right\}\ \textit{order if}\ \begin{cases} b_1 \ge b_2 \ge \cdots \ge b_k; \\ b_1 > b_2 > \cdots > b_k; \\ b_1 < b_2 < \cdots < b_k. \end{cases}$$

Again, this leads to similar terminology for functions. All such functions are also called *monotone functions*.

In the bijection we gave for our 35 lists, the lists without repetition correspond to the strictly increasing functions. Thus 3-subsets of $\underline{5}$ correspond to strictly increasing functions. From our previous list of multisets, we can read off these functions in 1-line form:

$$\begin{array}{ccccc}(1,2,3) & (1,2,4) & (1,2,5) & (1,3,4) & (1,3,5)\\ (1,4,5) & (2,3,4) & (2,3,5) & (2,4,5) & (3,4,5)\end{array}$$

We can interchange the strings "decreas" and "increas" in the previous paragraphs and read the functions in the list backwards. For example, the bijection between 3-subsets of $\underline{5}$ and strictly decreasing functions is given by

$$3,2,1 \quad 4,2,1 \quad 4,3,1 \quad 4,3,2 \quad 5,2,1 \quad 5,3,1 \quad 5,3,2 \quad 5,4,1 \quad 5,4,2 \quad 5,4,3.$$

[8] Recall that a multiset is like a set except that repeated elements are allowed.

[9] A later example explains how we got this number.

The function $(4,3,1)$ corresponds to the 3-subset $\{4,3,1\} = \{1,3,4\}$ of $\underline{5}$.

All these things are special cases of the following 2-in-1 theorem.

Theorem 2 (Sets, unordered lists and monotone functions) *Read either the top lines in all the braces or the bottom lines in all the braces. There are bijections between each of the following:*

- $\left\{\begin{matrix} k\text{-multisets} \\ k\text{-sets} \end{matrix}\right\}$ *whose elements lie in* $\underline{n}$,
- *the* $\left\{\begin{matrix} \text{weakly} \\ \text{strictly} \end{matrix}\right\}$ *increasing ordered* k*-lists made from* $\underline{n}$,
- *the* $\left\{\begin{matrix} \text{weakly} \\ \text{strictly} \end{matrix}\right\}$ *increasing functions in* $\underline{n}^{\underline{k}}$.

In these correspondences, the items in the list are the elements of the (multi)-set and are the values of the function.
In the correspondences, "increasing" can be replaced by "decreasing."

For example, reading the top lines in the braces with $k = 3$ and $n = 5$ and, in the last one, replacing "increasing" with "decreasing," we have: There are bijections between

(a) 3-multisets whose elements lie in $\underline{5}$,

(b) the weakly increasing ordered 3-lists in $\underline{5}$ and

(c) the weakly decreasing functions in $\underline{5}^{\underline{3}}$.

In these bijections, $\{2,5,5\}$ corresponds to list $(2,5,5)$ and the function $f = (5,5,2)$.

Example 14 (Counting multisets) Earlier we said there were $\binom{5+3-1}{3} = 35$ different 3-element multisets whose elements come from $\underline{5}$ and gave the list

1,1,1	1,1,2	1,1,3	1,1,4	1,1,5	1,2,2	1,2,3	1,2,4	1,2,5	1,3,3
1,3,4	1,3,5	1,4,4	1,4,5	1,5,5	2,2,2	2,2,3	2,2,4	2,2,5	2,3,3
2,3,4	2,3,5	2,4,4	2,4,5	2,5,5	3,3,3	3,3,4	3,3,5	3,4,4	3,4,5
3,5,5	4,4,4	4,4,5	4,5,5	5,5,5					

How did we know this? To see the trick, do the following to each 3-list:

- add 0 to the first item,
- add 1 to the second item, and
- add 2 to the third item.

Thus the first ten become

1,2,3	1,2,4	1,2,5	1,2,6	1,2,7	1,3,4	1,3,5	1,3,6	1,3,7	1,4,5

You should be able to see that you've created strictly increasing 3-lists from $\underline{7}$. In other words, you have listed all subsets of $\underline{7}$ of size 3. We know there are $\binom{7}{3} = 35$ such subsets and hence there were 35 multisets in the original list. In general, suppose we have listed

all weakly increasing k-lists from $\underline{n}$. Suppose each such k-list is in weakly increasing order. If, as in the above example, we add the list $0, 1, \ldots, k-1$ to each such k-list element by element, we get the strictly increasing k-lists from $\underline{n+k-1}$. By Theorem 2, this is a list of all k-subsets of $\underline{n+k-1}$. Thus, the number of weakly increasing $k-$ lists from $\underline{n}$ is $\binom{n+k-1}{k}$. By Theorem 2, this is also the number of k-multisets from $\underline{n}$. □

Exercises for Section 3

3.1. This exercise lets you check your understanding of the definitions. In each case below, some information about a function is given to you. Answer the following questions and give reasons for your answers:

- Have you been given enough information to specify the function; i.e., would this be enough data for a function envelope?
- Can you tell whether or not the function is an injection? a surjection? a bijection? If so, what is it?

(a) $f \in \underline{4}^{\underline{5}}$, $\quad \text{Coimage}(f) = \{\{1,3,5\},\{2,4\}\}$.

(b) $f \in \underline{5}^{\underline{5}}$, $\quad \text{Coimage}(f) = \{\{1\},\{2\},\{3\},\{4\},\{5\}\}$.

(c) $f \in \underline{4}^{\underline{5}}$, $\quad f^{-1}(2) = \{1,3,5\}, \quad f^{-1}(4) = \{2,4\}\}$.

(d) $f \in \underline{4}^{\underline{5}}$, $\quad \big|\text{Image}(f)\big| = 4$.

(e) $f \in \underline{4}^{\underline{5}}$, $\quad \big|\text{Image}(f)\big| = 5$.

(f) $f \in \underline{4}^{\underline{5}}$, $\quad \big|\text{Coimage}(f)\big| = 5$.

3.2. Let A and B be finite sets and let $f\colon A \to B$ be a function. Prove the following claims.

(a) $|\text{Image}| = |\text{Coimage}(f)|$.

(b) f is an injection if and only if $|\text{Image}| = |A|$.

(c) f is a surjection if and only if $|\text{Coimage}(f)| = |B|$.

3.3. In each case we regard the specified functions in one-line form. (As strings of integers, such functions are ordered lexicographically.)

(a) List all strictly decreasing functions in $\underline{5}^{\underline{3}}$ in lexicographic order. Note that this lists all subsets of size 3 from $\underline{5}$ in lexicographic order.

(b) In the list for part (a), for each string $x_1\,x_2\,x_3$ compute $\binom{x_1-1}{3} + \binom{x_2-1}{2} + \binom{x_3-1}{1} + 1$. What do these integers represent in terms of the list of part (a)?

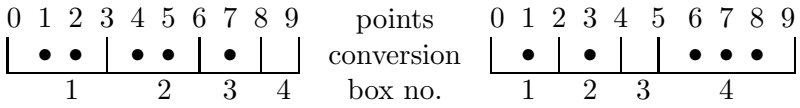
(c) List all strictly increasing functions in $\underline{5}^{\underline{3}}$ in lexicographic order. Note that this also lists all subsets of size 3 from $\underline{5}$ in lexicographic order.

(d) What is the analog, for part (c) of the formula of part (b)?
Hint: For each $x_1\,x_2\,x_3$ in the list of part (c), form the list $(6-x_1)(6-x_2)(6-x_3)$.

(e) In the lexicographic list of all strictly decreasing functions in $\underline{9}^{\underline{5}}$, find the successor and predecessor of 98321.

(f) How many elements are there before 98321 in this list?

3.4. We study the problem of listing all ways to put 5 balls (unlabeled) into 4 boxes (labeled 1 to 4). Consider ten consecutive points, labeled $0\cdots 9$. Some points are to be converted to box boundaries, some to balls. Points 0 and 9 are always box boundaries (call these *exterior box boundaries*). From the remaining points labeled $1\cdots 8$, we can arbitrarily pick three points to convert to *interior* box boundaries, five points to convert to balls. Here are two examples:



In this way, the placements of 5 balls into 4 boxes are made to correspond to subsets of size 3 (the box boundaries we can select) from $\underline{8}$. Lexicographic order on subsets of size 3 from $\underline{8}$, where the subsets are listed as strictly decreasing strings of length 3 from $\underline{8}$, correspondingly lex orders all placements of 5 balls into three boxes.

(a) Find the successor and predecessor of each of the above placements of balls into boxes.

(b) In the lex order of 5 balls into 4 boxes, which placement of balls into boxes is the last one in the first half of the list? The first one in the second half of the list?
Hint: The formula $p(x_1,x_2,x_3)=\binom{x_1-1}{3}+\binom{x_2-1}{2}+\binom{x_3-1}{1}+1$ gives the position of the string $x_1x_2x_3$ in the list of decreasing strings of length three from $\underline{8}$. Try to solve the equation $p(x_1,x_2,x_3)=\binom{8}{3}/2=28$ for the variables x_1,x_2,x_3.

3.5. Listing all of the partitions of an n set can be tricky and, of course, time consuming as there are lots of them. This exercise shows you a useful trick for small n. We define a class of functions called *restricted growth* functions that have the property that their collection of coimages is exactly the set of all partitions of $\underline{n}$.

(a) Call a function $f\in\underline{n}^{\underline{n}}$ a *restricted growth function* if $f(1)=1$ and $f(i)-1$ is at most the maximum of $f(k)$ over all $k<i$. Which of the following functions in one-line form are restricted growth functions? Give reasons for your answers.

$$2,2,3,3\quad 1,2,3,3,2,1\quad 1,1,1,3,3\quad 1,2,3,1.$$

(b) List, in lexicographic order, all restricted growth functions for $n=4$. Use one-line form and, for each one, list its coimage partition.

(c) For $n = 5$, list in lexicographic order the first fifteen restricted growth functions. Use one-line form. For the functions in positions 5, 10, and 15, list their coimage partitions.

3.6. How many functions f are there from $\underline{6}$ to $\underline{5}$ with $|\text{Image}(f)| = 3$?

3.7. How many ways can 6 different balls be placed into 5 labeled cartons in such a way that exactly 2 of the cartons contain no balls?

3.8. Count each of the following

(a) the number of multisets of size 6 whose elements lie in $\{a, b, c, d\}$,

(b) the number of weakly increasing functions from $\underline{6}$ to $\underline{4}$,

(c) the number of weakly decreasing ordered 6-lists made from $\underline{4}$,

(d) the number of strictly increasing functions from $\underline{6}$ to $\underline{9}$.

Section 4: Functions and Probability

In this section we look at various types of functions that occur in elementary probability theory. For the most part, we deal with finite sets. The functions we shall define are not difficult to understand, but they do have special names and terminology common to the subject of probability theory. We describe these various functions with a series of examples.

Probability functions have already been encountered in our studies. To review this idea, let U be a finite sample space (that is, a finite set) and let P be a function from U to $\mathbb{R}$ such that $P(t) \geq 0$ for all $t \in U$ and $\sum_{t \in U} P(t) = 1$. Then P is called a *probability function* on U. For any event $E \subseteq U$, define $P(E) = \sum_{t \in E} P(t)$. $P(E)$ is called the *probability of the event* E. The pair (U, P) is called a *probability space*.

Random Variables

Consider tossing a coin. The result is either heads or tails, which we denote by H and T. Thus the sample space is $\{H, T\}$. Sometimes we want to associate numerical values with elements of the sample space. Such functions are called "random variables." The function $X : \{H, T\} \to \mathbb{R}$, defined by $X(H) = 1$, $X(T) = 0$, is a random variable. Likewise the function $Y(H) = 1$, $Y(T) = -1$, same domain and range, is a random variable.

As another example, consider the sample space $U = \times^4\{H, T\}$, which contains the possible results of four coin tosses. The random variable $X(t_1, t_2, t_3, t_4) = |\{i \mid t_i = H\}|$ counts the number of times H appears in the sequence t_1, t_2, t_3, t_4. The function X has $\text{Image}(X) = \{0, 1, 2, 3, 4\}$, which is a subset of the set of real numbers.

Definition 5 (Random variable) *Let (U, P) be a probability space, and let $g : U \to \mathbb{R}$ be a function with domain U and range (codomain) $\mathbb{R}$, the real numbers. Such a function g is called a random variable on (U, P). The term "random variable" informs us that the range is the set of real numbers and that, in addition to the domain U, we also have a probability function P.*

Random variables are usually denoted by capital letters near the end of the alphabet. Thus, instead of g in the definition of random variable, most texts would use X.

By combining the two concepts, random variable and probability function, we obtain one of the most important definitions in elementary probability theory, that of a distribution function.

Definition 6 (Distribution function of a random variable) *Let $X : U \to \mathbb{R}$ be a random variable on a sample space U with probability function P. For each real number $t \in \text{Image}(X)$, let $X^{-1}(t)$ be the inverse image of t. Define a function $f_X : \text{Image}(X) \to \mathbb{R}$ by $f_X(t) = P(X^{-1}(t))$. The function f_X is called the probability distribution function of the random variable X. The distribution function is also called the density function.*

Since $P(X^{-1}(t))$ is the probability of the set of events $E = \{e \mid X(e) = t\}$, one often writes $P(X = t)$ instead of $P(X^{-1}(t))$.

Example 15 (Some distribution functions) Suppose we roll a fair die. Then $U = \{1, 2, 3, 4, 5, 6\}$ and $P(i) = 1/6$ for all i.

- If $X(t) = t$, then $f_X(t) = P(X^{-1}(t)) = P(t) = 1/6$ for $t = 1, 2, 3, 4, 5, 6$.
- If $X(t) = 0$ when t is even and $X(t) = 1$ when t is odd, then $f_X(0) = P(X^{-1}(0)) = P(\{2, 4, 6\}) = 1/2$ and $f_X(1) = 1/2$.
- If $X(t) = -1$ when $t \leq 2$ and $X(t) = 1$ when $t > 2$, then $f_X(-1) = P(\{1, 2\}) = 1/3$ and $f_X(1) = 2/3$. □

The function f_X, in spite of its fancy sounding name, is nothing but a probability function on the set $\text{Image}(X)$. Why is that?

- Since P is a nonnegative function with range $\mathbb{R}$, so is f_X.
- Since $\text{Coimage}(X) = \{X^{-1}(t) \mid t \in \text{Image}(X)\}$ is a partition of the set U,

$$1 = P(U) = \sum_{t \in \text{Image}(X)} P(X^{-1}(t)) = \sum_{t \in \text{Image}(X)} f_X(t)\,.$$

Thus, f_X is a nonnegative function from $\text{Image}(X)$ to $\mathbb{R}$ such that

$$\sum_{t \in \text{Image}(X)} f_X(t) = 1\,.$$

This is exactly the definition of a probability function on the set $\text{Image}(X)$.

Example 16 (Distribution of six cards with replacement) First one card and then a second card are selected at random, with replacement, from 6 cards numbered 1 to 6. The basic sample space is $S \times S = \{(i,j) : 1 \le i \le 6, 1 \le j \le 6\}$. Every point (i,j) in this sample space is viewed as equally likely: $P(i,j) = 1/36$. Define a random variable X on $S \times S$ by $X(i,j) = i + j$. $S \times S$ can be visualized as a 6×6 rectangular array and X can be represented by inserting $X(i,j)$ in the position represented by row i and column j. This can be done as follows:

	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

It is evident that $\text{Image}(X) = \{2,3,4,5,6,7,8,9,10,11,12\}$. The blocks of the $\text{Coimage}(X)$ are the sets of pairs (i,j) for which $i + j$ is constant. Thus,

$$X^{-1}(5) = \{(1,4),(2,3),(3,2),(4,1)\}$$

and

$$X^{-1}(8) = \{(2,6),(3,5),(4,4),(5,3),(6,2)\}\,.$$

Since every point in $S \times S$ has probability $1/36$, $P(X^{-1}(i)) = |X^{-1}(i)|/36$. Thus, with a little counting in the previous figure, the distribution function of X, $f_X = |X^{-1}(i)|/36$, in two-line form is

$$f_X = \begin{pmatrix} 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 1/36 & 2/36 & 3/36 & 4/36 & 5/36 & 6/36 & 5/36 & 4/36 & 3/36 & 2/36 & 1/36 \end{pmatrix}. \quad \square$$

Suppose we have two random variables X and Y defined on the same sample space U. Since the range of both X and Y is the real numbers, it makes sense to add the two random variables to get a new random variable: $Z = X + Y$; that is, $Z(t) = X(t) + Y(t)$ for all

$t \in U$. Likewise, if r is a real number, then $W = rX$ is a random variable, $W(t) = rX(t)$ for all $t \in U$. Thus we can do basic arithmetic with random variables.

For any random variable, X on U, we have the following important definition:

Definition 7 (Expectation of a random variable) *Let X be a random variable on a sample space U with probability function P. The expectation $E(X)$, or expected value of X, is defined by*

$$E(X) = \sum_{t \in U} X(t)P(t)\,.$$

$E(X)$ is often denoted by μ_X and referred to as the mean of X.

If we collect terms in preceding sum according to the value of $X(t)$, we obtain another formula for the expectation:

$$E(X) = \sum_{t \in U} X(t)P(t) = \sum_{r} r\Bigg(\sum_{\substack{t \in U \\ X(t)=r}} P(t) \Bigg) = \sum_{r} r f_X(r)\,.$$

The expectation E is a function whose arguments are functions. Another example of such a function is differentiation in calculus. Such functions, those whose arguments are themselves functions, are sometimes called "operators." Sometimes you will see a statement such as $E(2) = 2$. Since the arguments of the expectation operator E are functions, the "2" inside the parentheses is interpreted as the constant function whose value is 2 on all of U. The second 2 in $E(2) = 2$ is the number 2.

If X and Y are two random variables on a sample space U with probability function P, and $Z = X + Y$, then

$$\begin{aligned} E(Z) &= \sum_{t \in U} Z(t)P(t) \\ &= \sum_{t \in U} (X(t) + Y(t))P(t) \\ &= \sum_{t \in U} X(t)P(t) + \sum_{t \in U} Y(t)P(t) = E(X) + E(Y)\,. \end{aligned}$$

Similarly, for any real number r, $E(rX) = rE(X)$. Putting these observations together, we have proved the following theorem:

Theorem 3 (Linearity of expectation) *If X and Y are two random variables on a sample space U with probability function P and if a and b are real numbers, then $E(aX + bY) = aE(X) + bE(Y)$.*

We now introduce some additional functions defined on random variables, the covariance, variance, standard deviation, and correlation.

Definition 8 (Covariance, variance, standard deviation, correlation) *Let U be a sample space with probability function P, and let X and Y be random variables on U.*

- *Then the covariance of X and Y, $\mathrm{Cov}(X,Y)$, is defined by* $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$.
- *The variance of X, $\mathrm{Var}(X)$ (also denoted by σ_X^2), is defined by* $\mathrm{Var}(X) = \mathrm{Cov}(X,X) = E(X^2) - (E(X))^2$.
- *The standard deviation of X is* $\sigma_X = (\mathrm{Var}(X))^{1/2}$.
- *Finally, the correlation, $\rho(X,Y)$ of X and Y is* $\rho(X,Y) = \mathrm{Cov}(X,Y)/\sigma_X\sigma_Y$, *provided $\sigma_X \neq 0$ and $\sigma_Y \neq 0$.*

Example 17 (Sampling from a probability space) Consider the probability space (U,P) consisting of all possible outcomes from tossing a fair coin three times. Let the random variable X be the number of heads. Suppose we now actually toss a fair coin three times and record the result. This is called "sampling from the distribution." Given our sample $e \in U$, we can compute $X(e)$. Suppose we sample many times and average the values of $X(e)$ that we compute. As we increase the number of samples, our average moves around; e.g., if our values of $X(e)$ are $1,2,0,1,2,3,\ldots$ our averages of the first one, first two, first three, etc., are $1, 3/2, 3/3, 4/4, 6/5, 9/6, \ldots$, which reduce to $1, 1.5, 1, 1, 1.2, 1.5, \ldots$. What can we say about the average? When we take a large number of samples the average will tend to be close to μ_X, which is 1.5 in this case. Thus without knowing P we can estimate μ_x by sampling many times computing X and averaging the results.

The same idea applies to other functions of random variables: Sample many times, compute the function of the random variable(s) for each sample and average the results. In this way, we can estimate $E(X^2)$. In the previous paragraph, we estimated $E(X)$. Combining these estimates, we obtain an estimate for $\mathrm{Var}(X) = E(X^2) - (E(X))^2$. Refinements of this procedure are discussed in statistics classes. □

Theorem 4 (General properties of covariance and variance) *Let X, Y and Z be random variables on a probability space (U,P) and let a, b and c be real numbers. The covariance and variance satisfy the following properties:*

(1) (symmetry) $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$

(2) (bilinearity) $\mathrm{Cov}(aX+bY,Z) = a\mathrm{Cov}(X,Z) + b\mathrm{Cov}(Y,Z)$ *and*
$\mathrm{Cov}(X,aY+bZ) = a\mathrm{Cov}(X,Y) + b\mathrm{Cov}(X,Z)$.

Thinking of a as the constant function equal to a and likewise for b, we have

(3) $\mathrm{Cov}(a,X) = 0$ *and* $\mathrm{Cov}(X+a,Y+b) = \mathrm{Cov}(X,Y)$.

In particular, $\mathrm{Cov}(X,Y) = E((X-\mu_X)(Y-\mu_Y))$ *and* $\mathrm{Var}(X) = E((X-\mu_X)^2)$.

(4) $\begin{aligned}\mathrm{Var}(aX+bY+c) &= \mathrm{Var}(aX+bY)\\ &= a^2\mathrm{Var}(X) + 2ab\,\mathrm{Cov}(X,Y) + b^2\mathrm{Var}(Y).\end{aligned}$

The last two formulas in (3) are sometimes taken as the definition of the covariance and variance. In that case, the formulas for them in Definition 8 would be proved as a theorem. Note that the formula $\mathrm{Var}(X) = E((X-\mu_X)^2)$ says that $\mathrm{Var}(X)$ tells us something about how far X is likely to be from its mean.

Proof: Property (1) follows immediately from the fact that $XY = YX$ and $E(X)E(Y) = E(Y)E(X)$. The following calculations prove (2).

$$\begin{aligned}
\mathrm{Cov}(aX+bY,Z) &= E((aX+bY)Z) - E(aX+bY)E(Z) && \text{by definition}\\
&= aE(XZ) + bE(YZ) - \big(aE(X)E(Z) + bE(Y)E(Z)\big) && \text{by Theorem 3}\\
&= a\big(E(XZ) - E(X)E(Z)\big) + b\big(E(YZ) - E(Y)E(Z)\big)\\
&= a\mathrm{Cov}(X,Z) + b\mathrm{Cov}(Y,Z) && \text{by definition.}
\end{aligned}$$

We now turn to (3). By definition, $\mathrm{Cov}(a,X) = E(aX) - E(a)E(X)$. Since $E(aX) = aE(X)$ and $E(a) = a$, $\mathrm{Cov}(a,X) = 0$. Using the various parts of the theorem,

$$\begin{aligned}
\mathrm{Cov}(X+a,Y+b) &= \mathrm{Cov}(X,Y+b) + \mathrm{Cov}(a,Y+b) = \mathrm{Cov}(X,Y+b)\\
&= \mathrm{Cov}(X,Y) + \mathrm{Cov}(X,b) = \mathrm{Cov}(X,Y).
\end{aligned}$$

The particular results follow when we set $a = -\mu_X$ and $b = -\mu_Y$.

You should be able to prove (4) by using (1), (2), (3) and the definition $\mathrm{Var}(Z) = \mathrm{Cov}(Z,Z)$. □

Example 18 (General properties of correlation) Recall that the correlation of X and Y is $\rho(X,Y) = \mathrm{Cov}(X,Y)/\sigma_X\sigma_Y$, provided $\sigma_x \neq 0$ and $\sigma_Y \neq 0$. Since

$$\mathrm{Cov}(X+c,Y+d) = \mathrm{Cov}(X,Y) \quad \text{and} \quad \mathrm{Var}(X+c) = \mathrm{Var}(X)$$

for any constant functions c and d, we have $\rho(X+c,Y+d) = \rho(X,Y)$ for any constant functions c and d. Suppose that X and Y both have mean zero. Note that

$$0 \leq E((X+tY)^2) = E(X^2) + 2E(XY)t + E(Y^2)t^2 = f(t)$$

defines a nonnegative polynomial of degree 2 in t. From high school math or from calculus, the minimum value of a polynomial of the form $A + 2Bt + Ct^2$ $(C > 0)$ is

$$A - B^2/C = E(X^2) - (E(XY))^2/E(Y^2).$$

Since for all t, $f(t) \geq 0$, we have $E(X^2) - (E(XY))^2/E(Y^2) \geq 0$. Since X and Y have mean zero,

$$\mathrm{Cov}(X,Y) = E(XY), \quad \mathrm{Var}(X) = E(X^2), \quad \text{and} \quad \mathrm{Var}(Y) = E(Y^2).$$

Thus,

$$(E(XY))^2/E(X^2)E(Y^2) = (\rho(X,Y))^2 \leq 1$$

or, equivalently, $-1 \leq \rho(X,Y) \leq +1$. If the means of X and Y are not zero then replace X and Y by $X - \mu_X$ and $Y - \mu_Y$ respectively, to obtain $-1 \leq \rho(X,Y) \leq +1$. □

We have just proved the following theorem about the correlation of random variables.

Theorem 5 (Bounds on the correlation of two random variables) *Let X and Y be random variables on a sample space U and let $\rho(X,Y)$ be their correlation. Then $-1 \leq \rho(X,Y) \leq +1$.*

The intuitive interpretation of the correlation $\rho(X,Y)$ is that values close to 1 mean points in U where X is large also tend to have Y large. Values close to -1 mean the opposite. As extreme examples, take Y=X. Then $\rho(X,Y) = 1$. If we take $Y = -X$ then $\rho(X,Y) = -1$.

Tchebycheff's inequality[10] is another easily proved inequality for random variables. It relates the tendency of a random variable to be far from its mean to the size of its variance. Such results are said to be "measures of central tendency."

Theorem 6 (Tchebycheff's inequality) *Let X be a random variable on a probability space (U,P). Suppose that $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Let $\epsilon > 0$ be a real number. Then*

$$P\Big(\{u \mid \ |X(u) - \mu| \geq \epsilon\}\Big) \ \leq \ \frac{\sigma^2}{\epsilon^2}\,.$$

The left side of the inequality contains the set of all u for which $|X(u) - \mu| \geq \epsilon$. Thus it can be thought of as the probability that the random variable X satisfies $|X - \mu| \geq \epsilon$.

The most important aspect of Tchebycheff's inequality is the universality of its applicability: the random variable X is arbitrary.

Proof: Let's look carefully at the computation of the variance:

$$\mathrm{Var}(X) \ = \ E[(X-\mu)^2] \ = \sum_{\{u \mid \ |X-\mu| \geq \epsilon\}} (X(u)-\mu)^2 P(u) \ + \sum_{\{u \mid \ |X-\mu| < \epsilon\}} (X(u)-\mu)^2 P(u)\,.$$

In breaking down the variance into these two sums, we have partitioned U into two disjoint sets $\{u \mid \ |X-\mu| \geq \epsilon\}$ and $\{u \mid \ |X-\mu| < \epsilon\}$. Since all terms are positive, $\mathrm{Var}(X)$ is greater than or equal to either one of the above sums. In particular,

$$\mathrm{Var}(X) = E[(X-\mu)^2] \ \geq \sum_{\{u \mid \ |X-\mu| \geq \epsilon\}} (X(u)-\mu)^2 P(u)\,.$$

Note that

$$\sum_{\{u| \ |X-\mu| \geq \epsilon\}} (X(u)-\mu)^2 P(u) \ \geq \ \epsilon^2 \Bigg(\sum_{\{u \mid \ |X-\mu| \geq \epsilon\}} P(u) \Bigg) \ = \ \epsilon^2 P\Big(\{u \ \mid \ |X-\mu| \geq \epsilon\}\Big).$$

Putting all of this together proves the theorem. □

[10] Tchebycheff is also spelled Chebyshev, depending on the system used for transliterating Russian.

Joint Distributions

A useful concept for working with pairs of random variables is the *joint distribution function*:

Definition 9 (Joint distribution function) *Let X and Y be random variables on a sample space U with probability function P. For each $(i,j) \in \text{Image}(X) \times \text{Image}(Y)$, define $h_{X,Y}(i,j) = P(X^{-1}(i) \cap Y^{-1}(j))$. The function $h_{X,Y}$ is called the joint distribution function of X and Y.*

Recalling the meaning of the distribution functions f_X and f_Y (Definition 6) you should be able to see that

$$f_X(i) = \sum_{j \in \text{Image}(Y)} h_{X,Y}(i,j) \quad \text{and} \quad f_Y(j) = \sum_{i \in \text{Image}(X)} h_{X,Y}(i,j).$$

Example 19 (A joint distribution for coin tosses) A fair coin is tossed three times, recording H if heads, T if tails. The sample space is

$$U = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

Let X be the random variable defined by $X(t_1t_2t_3) = 0$ if $t_1 = T$, $X(t_1t_2t_3) = 1$ if $t_1 = H$. Let Y be the random variable that counts the number of times T occurs in the three tosses. $\text{Image}(X) \times \text{Image}(Y) = \{0,1\} \times \{0,1,2,3\}$. We compute $h_{X,Y}(0,2)$. $X^{-1}(0) = \{THH, THT, TTH, TTT\}$. $Y^{-1}(2) = \{HTT, THT, TTH\}$. $X^{-1}(0) \cap Y^{-1}(2) = \{THT, TTH\}$. Thus, $h_{X,Y}(0,2) = 2/8$. Computing the rest of the values of $h_{X,Y}$, we can represent the results in the following table:

$h_{X,Y}$	Y=0	Y=1	Y=2	Y=3	f_X
X=0	0	1/8	2/8	1/8	1/2
X=1	1/8	2/8	1/8	0	1/2
f_Y	1/8	3/8	3/8	1/8	

In this table, the values of $h_{X,Y}(i,j)$ are contained in the submatrix

$$\begin{matrix} 0 & 1/8 & 2/8 & 1/8 \\ 1/8 & 2/8 & 1/8 & 0 \end{matrix}$$

The last column gives the distribution function f_X and the last row gives the distribution function f_Y. These distributions are called the *marginal distributions* of the joint distribution $h_{X,Y}$. You should check that $E(X) = 1/2$, $E(Y) = 3/2$, $\text{Var}(X) = 1/4$, $\text{Var}(Y) = 3/4$. Compute also that $E(XY) = \sum ijh_{X,Y}(i,j) = 1/2$, where the sum is over $(i,j) \in \{0,1\} \times \{0,1,2,3\}$. Thus, $\text{Cov}(X,Y) = -1/4$. Putting this all together gives $\rho(X,Y) = -3^{-1/2}$. It should be no surprise that the correlation is negative. If X is "large" (i.e., 1) then Y should be "small," since the first toss came up H, making the total number of T's at most 2. □

Example 20 (Another joint distribution for coin tosses) As before, a fair coin is tossed three times, recording H if heads, T if tails. The sample space is

$$U = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

Again, let X be the random variable defined by $X(t_1t_2t_3) = 0$ if $t_1 = T$, $X(t_1t_2t_3) = 1$ if $t_1 = H$. However, now let Y be the random variable that counts the number of times T occurs in the last two tosses. Image$(X) \times$ Image$(Y) = \{0,1\} \times \{0,1,2\}$. We compute $h_{X,Y}(0,2)$. $X^{-1}(0) = \{THH, THT, TTH, TTT\}$. $Y^{-1}(2) = \{HTT, TTT\}$. $X^{-1}(0) \cap Y^{-1}(2) = \{TTT\}$. Thus, $h_{X,Y}(0,2) = 1/8$. Computing the rest of the values of $h_{X,Y}$, we can represent the results in the following table:

$h_{X,Y}$	Y=0	Y=1	Y=2	f_X
X=0	1/8	2/8	1/8	1/2
X=1	1/8	2/8	1/8	1/2
f_Y	1/4	1/2	1/4	

You should compute that $E(X) = 1/2$, $E(Y) = 1$, $\mathrm{Var}(X) = 1/4$, $\mathrm{Var}(Y) = 1/2$ and $\mathrm{Cov}(X,Y) = 0$. Since all the previous numbers were nonzero, it is rather surprising that the covariance is zero. This is a consequence of "independence," which we will study next. □

Independence

If the expectation of the sum of random variables is the sum of the expectations, then maybe the expectation of the product is the product of the expectations? Not so. We'll look a simple example with $Y(T) = X(T)$. Let $U = \{H, T\}$ and $P(H) = P(T) = 1/2$ and let $X(T) = 0$ and $X(H) = 1$ be a random variable on (U, P). Then $E(X) = X(T)(1/2) + X(H)(1/2) = 1/2$. Since $X^2 = X$, we have $E(XX) = E(X^2) = E(X) = 1/2$. This does not equal $E(X)^2$. In order to give some general sufficient conditions for when $E(XY) = E(X)E(Y)$ we need the following definition.

Definition 10 (Independence) *Let (U, P) be a probability space.*

- *If $A \subseteq U$ and $B \subseteq U$ are events (subsets) of U such that $P(A \cap B) = P(A)P(B)$. Then A and B are said to be a pair of independent events.*
- *If X and Y are random variables on U such that*

  $$\text{for } \underline{\text{all}}\ s \in \text{Image}(X) \text{ and } \underline{\text{all}}\ t \in \text{Image}(Y),$$
  $$X^{-1}(s) \text{ and } Y^{-1}(t) \text{ are a pair of independent events,}$$

  then X and Y are said to be a pair of independent random variables.

The definition of independence for events and random variables sounds a bit technical. In practice we will use independence in a very intuitive way.

Recall the definitions of the marginal and joint distributions f_X, f_Y and $h_{X,Y}$ in Definition 9. Using that notation, the definition of independence can be rephrased as

$$\begin{gathered} X \text{ and } Y \text{ are independent random variables} \\ \text{if and only if} \\ h_{X,Y}(i,j) = f_X(i) f_Y(j) \text{ for all } (i,j) \in \text{Image}(X) \times \text{Image}(Y). \end{gathered}$$

You should verify that X and Y are independent in Example 20. Intuitively, knowing what happens on the first toss gives us no information about the second and third tosses. We explore this a bit in the next example.

Example 21 (Independent coin tosses) Suppose a fair coin is tossed twice, one after the other. In everyday language, we think of the tosses as being independent. We'll see that this agrees with our mathematical definition of independence.

The sample space is $U = \{(H,H),(H,T),(T,H),(T,T)\}$. Note that $U = \{H,T\} \times \{H,T\}$ and p is the uniform probability function on U; i.e., $P(e) = 1/4$ for each $e \in U$.

Let A be the event that the first toss is H and let B be the event that the second is H. Thus $A = \{HH, HT\}$ and $B = \{HH, TH\}$. You should be able to see that $P(A) = 1/2$, $P(B) = 1/2$ and

$$P(A \cap B) = P(HH) = 1/4 = (1/2)(1/2)$$

and so A and B are independent.

What about independent random variables? Let

$$X(t_1,t_2) = \begin{cases} 0, & \text{if } t_1 = T, \\ 1, & \text{if } t_1 = H. \end{cases}$$

and

$$Y(t_1,t_2) = \begin{cases} 0, & \text{if } t_2 = T, \\ 1, & \text{if } t_2 = H. \end{cases}$$

Thus X "looks at" just the first toss and Y "looks at" just the second. You should be able to verify that $X^{-1}(1) = A$, $X^{-1}(0) = A^c$, $Y^{-1}(1) = B$ and $Y^{-1}(0) = B^c$. To see that X and Y are independent, we must verify that each of the following 4 pairs of events is independent

$$A \text{ and } B \qquad A \text{ and } B^c \qquad A^c \text{ and } B \qquad A^c \text{ and } B^c.$$

In the previous paragraph, we saw that A and B are independent. You should be able to do the other 3.

This seems like a lot of work to verify the mathematical notion of independence, compared with the obvious intuitive notion. Why bother? There are two reasons. First, we want to see that the two notions of independence are the same. Second, we can't do any calculations with an intuitive notion, but the mathematical definition will allow us to obtain useful results. □

The preceding example can be generalized considerably. The result is an important method for building up new probability spaces from old ones.

***Example 22 (Product spaces and independence)** Let (U_1, P_1) and (U_2, P_2) be probability spaces. Define the *product space* (U, P) by

$$\begin{aligned} U &= U_1 \times U_2 && \text{(Cartesian product)} \\ P(e_1, e_2) &= P_1(e_1) \times P_2(e_2) && \text{(multiplication of numbers)} \end{aligned}$$

Suppose $A = A_1 \times U_2$ and $B = U_1 \times B_2$. We claim that A and B are independent events. Before proving this, let's see how it relates to the previous example.

Suppose $U_1 = U_2 = \{H, T\}$ and that P_1 and P_2 are the uniform probability functions. Then (U_1, P_1) describes the first toss and (U_2, P_2) describes the second. Also, you should check that (U, P) is the same probability space as in the previous example. Check that, if $A_1 = \{H\}$ and $B_2 = \{H\}$, then A and B are the same as in the previous example.

We now prove that A and B are independent in our general setting. We have

$$\begin{aligned} P(A) &= \sum_{e \in A} P(e) && \text{definition of } P(A) \\ &= \sum_{e \in A_1 \times U_2} P(e) && \text{definition of } A \\ &= \sum_{\substack{e_1 \in A_1 \\ e_2 \in U_2}} P(e_1, e_2) && \text{definition of Cartesian product} \\ &= \sum_{\substack{e_1 \in A_1 \\ e_2 \in U_2}} P_1(e_1) P_2(e_2) && \text{definition of } P \\ &= \left(\sum_{e_1 \in A_1} P_1(e_1) \right) \times \left(\sum_{e_2 \in U_2} P_2(e_2) \right) && \text{algebra} \\ &= \sum_{e_1 \in A_1} P_1(e_1) \times 1 && \text{definition of probability} \\ &= P_1(A_1) && \text{definition of } P_1(A_1). \end{aligned}$$

Similarly, $P(B) = P_2(B_2)$. You should verify that $A \cap B = A_1 \times B_2$. By doing calculations similar to what we did for $P(A)$, you should show that $P(A_1 \times B_2) = P_1(A_1)P_2(B_2)$. This proves independence.

Suppose X is a random variable such that $X(e_1, e_2)$ depends only on e_1. What does $X^{-1}(r)$ look like? Suppose $X(e_1, e_2) = r$ so that $(e_1, e_2) \in X^{-1}(r)$. Since X does not depend on e_2, $X(e_1, u_2) = r$ for all $u_2 \in U_2$. Thus $\{e_1\} \times U_2 \subseteq X^{-1}(r)$. Proceeding in this way, one can show that $X^{-1}(r) = A_1 \times U_2$ for some $A_1 \subseteq U_1$.

Suppose Y is a random variable such that $Y(e_1, e_2)$ depends only on e_2. We then have $Y^{-1} = U_1 \times B_2$ for some $B_2 \subseteq U_2$. By our earlier work in this example, it follows that $X^{-1}(r)$ and $Y^{-1}(s)$ are independent events and so X and Y are independent.

What made this work? X "looked at" just the first component and Y "looked at" just the second.

This can be generalized to the product of any number of probability spaces. Random variables X and Y will be independent if the components that X "looks at" are disjoint

from the set of components that Y "looks at." For example, suppose a coin is tossed "independently" 20 times. Let X count the number of heads in the first 10 tosses and let Y count the number of tails in the last 5 tosses. □

We now return to $E(XY)$, with the assumption that X and Y are independent random variables on a sample space U with probability function P. We also look at variance, covariance, and correlation.

Theorem 7 (Properties of independent random variables) *Suppose that U is a sample space with probability function P and that X and Y are independent random variables on U. Then the following are true*

- $E(XY) = E(X)E(Y)$,
- $\mathrm{Cov}(X,Y) = 0$ *and* $\rho(X,Y) = 0$,
- $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$,
- *if $f, g : \mathbb{R} \to \mathbb{R}$, then $f(X)$ and $g(Y)$ are independent random variables.*

Proof: First note that there are two ways to compute $E(Z)$ for a random variable Z:

$$E(Z) = \sum_{u\in U} Z(u)P(u) \qquad \text{(the first way)}$$

$$= \sum_{k\in \mathrm{Image}(Z)} kP(Z^{-1}(k)) = \sum_{k\in \mathrm{Image}(Z)} kf_Z(k) \quad \text{(the second way).}$$

For $Z = XY$, we use the second way. If $Z = k$, then $X = i$ and $Y = j$ for some i and j with $ij = k$. Thus

$$E(Z) = \sum_{k\in \mathrm{Image}(Z)} kP(Z^{-1}(k))$$

$$= \sum_{\substack{i\in \mathrm{Image}(X)\\ j\in \mathrm{Image}(Y)}} ijP(X^{-1}(i)\cap Y^{-1}(j)).$$

From the definition of independence, $P(X^{-1}(i)\cap Y^{-1}(j)) = P(X^{-1}(i))P(Y^{-1}(j))$, and hence

$$\sum_{\substack{i\in \mathrm{Image}(X)\\ j\in \mathrm{Image}(Y)}} ijP(X^{-1}(i)\cap Y^{-1}(j)) = \sum_{\substack{i\in \mathrm{Image}(X)\\ j\in \mathrm{Image}(Y)}} ijP(X^{-1}(i))\ P(Y^{-1}(j))$$

$$= \left(\sum_{i\in \mathrm{Image}(X)} iP(X^{-1}(i))\right) \times \left(\sum_{j\in \mathrm{Image}(Y)} jP(Y^{-1}(j))\right).$$

The right hand side of the above equation is just $E(X)E(Y)$. This proves the first part of the theorem.

By Definition 8 and the fact that we have just proved $E(XY) = E(X)E(Y)$, it follows that $\mathrm{Cov}(X,Y) = 0$. It then follows from Definition 8 that $\rho(X,Y) = 0$. You should use some of the results in Theorem 4 to show that $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

We omit the proof of the last part of the theorem; however, you should note that $f(X)$ is a *function* on U because, for $u \in U$, $f(X)(u) = f(X(u))$. □

***Example 23 (Generating random permutations)** Suppose we want to generate permutations of $\underline{n}$ randomly so that each permutation is equally likely to occur. How can we do so? We now look at a simple, efficient method for doing this. The procedure is based on a bijection between two sets:

- Seq_n, the set of all sequences $a_2, a_3, \ldots, a_n$ of integers with $1 \le a_k \le k$ for all k and
- Perm_n, the set of all permutations of $\underline{n} = \{1, 2, \ldots, n\}$.

We can tell there must be such a bijection even without giving one! Why is this? Since there are k choices for a_k, the number of sequences $a_2, a_3, \ldots, a_n$ is $2 \times 3 \times \cdots \times n = n!$. We know that there are $n!$ permutations. Thus $|\text{Seq}_n| = |\text{Perm}_n|$. Because Seq_n and Perm_n have the same size, there must be a bijection between them. Since there's a bijection, what's the problem? The problem is to find one that is easy to use.

Before providing the bijection, let's look at how to use it to generate permutations uniformly at random. It is easy to generate the sequences uniformly at random: Choose a_k uniformly at random from $\underline{k}$ and choose each of the a_k independently. This makes Seq_n into a probability space with the uniform probability distribution. Once we have a sequence, we use the bijection to construct the permutation that corresponds to the sequence. This makes Perm_n into a probability space with the uniform probability distribution.

Now we want to specify a bijection. There are lots of choices. (You should be able to show that there are $(n!)!$ bijections.) Here's one that is easy to use:

Step 1. Write out the sequence $1, 2, \ldots, n$ and set $k = 2$

Step 2. If $a_k \neq k$, swap the elements in positions k and a_k in the sequence. If $a_k = k$, do nothing.

Step 3. If $k < n$, increase k by 1 and go to Step 2. If $k = n$, stop.

The result is a permutation in one-line form.

For example, suppose $n = 5$ and the sequence of a_k's is 2,1,3,3. Here's what happens, where "next step" tells which step to use to produce the permutation on the next line:

action	permutation	information	next step
the start (Step 1)	1,2,3,4,5	$k = 2$, $a_k = 2$	Step 3
do nothing	1,2,3,4,5	$k = 3$, $a_k = 1$	Step 2
swap at 3 and 1	3,2,1,4,5	$k = 4$, $a_k = 3$	Step 2
swap at 4 and 3	3,2,4,1,5	$k = 5$, $a_k = 3$	Step 2
swap at 5 and 3	3,2,5,1,4		all done

Thus, the sequence $2, 1, 3, 3$ corresponds to the permutation $3, 2, 5, 1, 4$.

How can we prove that this is a bijection? We've described a function F from Seq_n to Perm_n. Since $|\text{Seq}_n| = |\text{Perm}_n|$, we can prove that F is a bijection if we can show that it is a surjection. (This is just Exercise 1.2(c).) In other words, we want to show that, for every permutation p in Perm_n, there is a sequence $\mathbf{a}$ in Seq_n such that $F(\mathbf{a}) = p$.

Let's try an example. Suppose we have the permutation $4, 1, 3, 2, 6, 5$. What sequence does it come from? This is a permutation of $\underline{6}$. The only way 6 could move from the last place is because $a_6 \neq 6$. In fact, since 6 is in the fifth place, we must have had $a_6 = 5$.

Which caused us to swap the fifth and sixth positions. So, just before we used $a_6 = 5$, we had the permutation $4, 1, 3, 2, 5, 6$. None of $a_2, \ldots, a_5$ can move what's in position 6 since they are all less than 6, so $a_2, \ldots, a_5$ must have rearranged $1, 2, 3, 4, 5, (6)$ to give $4, 1, 3, 2, 5, (6)$. (We've put 6 in parentheses to remember that it's there and that none of $a_2, \ldots, a_5$ can move it.) How did this happen? Well, only a_5 could have affected position 5. Since 5 is there, it didn't move and so $a_5 = 5$. Now we're back to $4, 1, 3, 2, (5, 6)$ and trying to find a_4. Since 4 is in the first position, $a_4 = 1$. So, just before using, a_4 we had $2, 1, 3, (4, 5, 6)$. Thus $a_3 = 3$ and we're back to $2, 1, (3, 4, 5, 6)$. Finally $a_2 = 1$. We've found that the sequence $1, 3, 1, 5, 5$ gives the permutation $4, 1, 3, 2, 6, 5$. You should apply the algorithm described earlier in Steps 1, 2, and 3 and so see for yourself that the sequence gives the permutation.

The idea in the previous paragraph can be used to give a proof by induction on n. For those of you who would like to see it, here's the proof. The fact that F is a surjection is easily checked for $n = 2$: There are two sequences, namely 1 and 2. These correspond to the two permutations $2, 1$ and $1, 2$, respectively. Suppose $n > 2$ and $p_1, \ldots, p_n$ is a permutation of $\underline{n}$. We need to find the position of n in the permutation. The position is that k for which $p_k = n$. So we set $a_n = k$ and define a new permutation $q_1, \ldots, q_{n-1}$ of $\{1, 2, \ldots, n-1\}$ to correspond to the situation just before using $a_n = k$:

- If $k = n$, then $q_i = p_i$ for $1 \leq i \leq n - 1$.
- If $k \neq n$, the $q_k = p_n$ and $q_i = p_i$ for $1 \leq i < k$ and for $k < i \leq n - 1$.

You should be able to see that $q_1, \ldots, q_{n-1}$ is a permutation of $\underline{n-1}$. By induction, there is a sequence $a_2, \ldots, a_{n-1}$ that gives $q_1, \ldots, q_{n-1}$ when we apply our 3-step procedure to $1, 2, 3, \ldots, (n-1)$. After that, we must apply $a_n = k$ to $q_1, \ldots, q_{n-1}, n$. What happens? You should be able to see that it gives us $p_1, \ldots, p_n$. This completes the proof. □

Some Standard Distributions

We now take a look at some examples of random variables and their distributions that occur often in applications. The first such distribution is the *binomial distribution*.

Example 24 (Binomial distribution) Suppose we toss a coin, sequentially and independently, n times, recording H for heads and T for tails. Suppose the probability of H in a single toss of the coin is p. Define

$$P^*(t) = \begin{cases} p, & \text{if } t = H, \\ q = 1 - p, & \text{if } t = T. \end{cases}$$

Our sample space is $U = \times^n \{H, T\}$ and the probability function P is given by $P(t_1, \ldots, t_n) = P^*(t_1) \cdots P^*(t_n)$ because of independence. This is an example of a product space. We discussed product spaces in Example 22.

Define the random variable $X(t_1, \ldots, t_n)$ to be the number of H's in the sequence $(t_1, \ldots, t_n)$. This is a standard example of a *binomial random variable*.

We want to compute $P(X = k)$ for $k \in \mathbb{R}$. Note that Image$(X) = \{0, \ldots, n\}$. Hence $P(x = k) = 0$ if k is not in $\{0, \ldots, n\}$. Note that $(t_1, \ldots, t_n) \in X^{-1}(k)$ if and only if $(t_1, \ldots, t_n)$ contains exactly k heads (H's). In this case, $P(t_1, \ldots, t_n) = p^k q^{n-k}$. Since all elements of $X^{-1}(k)$ have the same probability $p^k q^{n-k}$, it follows that $f_X(k) = |X^{-1}(k)|\, p^k q^{n-k}$. What is the value of $|X^{-1}(k)|$. It is the number of sequences with exactly k heads. Since the positions for k heads must be chosen from among the n tosses, $|X^{-1}(k)| = \binom{n}{k}$. Thus $f_X(k) = \binom{n}{k} p^k q^{n-k}$. This is the *binomial distribution* function. A common alternative notation for this distribution function is $b(k; n, p)$. This notation has the advantage of explicitly referencing the parameters, n and p.

An alternative way of thinking about the random variable X is to write it as a sum, $X = X_1 + \cdots + X_n$, of n independent random variables. The random variable X_i is defined on the sample space $U = \times^n \{H, T\}$ by the rule

$$X_i(t_1, \ldots, t_n) = \begin{cases} 1, & \text{if } t_i = H, \\ 0, & \text{if } t_i = T. \end{cases}$$

Using this representation of X, we can compute $E(X) = E(X_1) + \cdots + E(X_n)$, and $\mathrm{Var}(X) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)$. Computation gives

$$E(X_i) = 1 \times P(X_i = 1) + 0 \times P(X_i = 0) = p$$

and

$$\mathrm{Var}(X_i) = E(X_i^2) - E(X_i)^2 = p - p^2 = p(1 - p),$$

where we have used $X_i^2 = X_i$ because X_i must be 0 or 1. Thus, we obtain $E(X) = np$ and $\mathrm{Var}(X) = np(1 - p) = npq$. □

Of course, the binomial distribution is not restricted to coin tosses, but is defined for any series of outcomes that

- are restricted to two possibilities,
- are independent, and
- have a fixed probability p of one outcome, $1 - p$ of the other outcome.

Our next example is a random variable X that is defined on a countably infinite sample space U. This distribution, the Poisson, is associated with random distributions of objects.

Example 25 (Poisson distribution and its properties) Suppose a 500 page book has 2,000 misprints. If the misprints are distributed randomly, what is the probability of exactly k misprints appearing on page 95? (We want the answers for $k = 0, 1, 2, \ldots$.)

Imagine that the misprints are all in a bag. When we take out a misprint, it appears on page 95 with probability 1/500. Call the case in which a misprint appears on page 95 a "success" and the case when it does not a "failure." We have just seen that, for a randomly selected misprint, the probability of success is $p = 1/500$. Since we have assumed the misprints are independent, we can use the binomial distribution. Our answer is therefore that the probability of exactly k misprints on page 95 is $b(k; 2000, 1/500)$.

Thus we have our answer: $b(k; 2000, 1/500) = \binom{2000}{k}(1/500)^k(1-1/500)^{2000-k}$. Unfortunately, its hard to use: for large numbers the binomial distribution is awkward to work with because there is a lot of calculation involved and numbers can be very large or very small. Can we get a more convenient answer? Yes. There is a nice approximation which we will now discuss.

The function $f_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$ is also denoted by $p(k;\lambda)$ and is called the *Poisson distribution*. Clearly the $p(k;\lambda)$ are positive. Also, they sum to one:

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

We have used the Taylor Series expansion, obtained in calculus courses, $\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{\lambda}$. In a similar manner, it can be shown that

$$E(X) = \lambda \quad \text{and} \quad \mathrm{Var}(X) = \lambda.$$

Thus, a Poisson distributed random variable X has the remarkable property that $E(X) = \lambda$ and $\mathrm{Var}(X) = \lambda$ where $\lambda > 0$ is the parameter in the distribution function $P(X = k) = p(k;\lambda) = e^{-\lambda}\lambda^k/k!$.

We now return to our binomial distribution $b(k; 2000, 1/500)$. The Poisson is a good approximation to $b(k; n, p)$ when n is large and np is not large. In this case, take $\lambda = np$, the mean of the binomial distribution. For our problem, $\lambda = 2000(1/500) = 4$, which is not large when compared to the other numbers in the problem, namely 2,000 and 500. Let's compute some estimates for P_k, the probability of exactly k errors on page 95.

$$P_0 = e^{-4} = 0.0183, \quad P_1 = 4e^{-4} = 0.0733, \quad P_3 = 4^3e^{-4}/3! = 0.1954,$$

and so on. □

Our final example of a random variable X has its underlying sample space $U = \mathbb{R}$, the real numbers. Rather than starting with a description of X itself, we start with the distribution function $f_X(x) = \phi_{\mu,\sigma}(x)$, called the *normal distribution function with mean μ and standard deviation σ*.

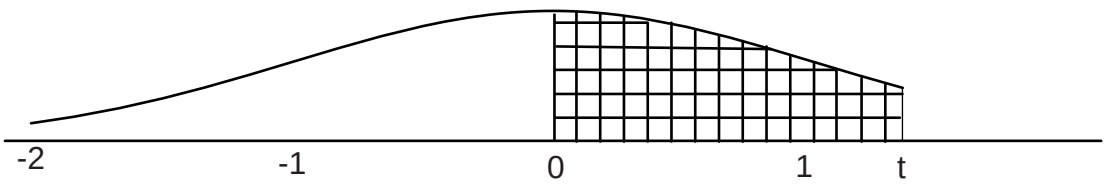
$$\phi_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}\,.$$

For computations concerning the normal distribution, it suffices in most problems, to work with the special case when $\mu = 0$ and $\sigma = 1$. In this case, we use the notation

$$\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$$

where $\phi(x) = \phi_{0,1}(x)$ is called the *standard normal distribution*.

The function $\phi(x)$ is defined for $-\infty < x < \infty$ and is symmetric about $x = 0$. The maximum of $\phi(x)$ occurs at $x = 0$ and is about 0.4. Here is a graph of $\phi(x)$ for $-2 \le x \le t$:



In this graph of $\phi(x)$ shown above, the area between the curve and the interval from 0 to t on the x-axis is shaded. This area, as we shall discuss below, represents the probability that a random variable with distribution function ϕ lies between 0 and t. For $t = 1$ the probability is about 0.34, for $t = 1.5$ the probability is about 0.43, and for $t = 2$, the probability is about 0.48. Since this is a probability distribution, the area under the whole curve is 1. Also, since the curve is symmetric, the area for $x < 0$ is 1/2. We'll use these values in the examples and problems, so you will want to refer back to them.

Example 26 (The normal distribution and probabilities) The way the normal curve relates to probability is more subtle than in the finite or discrete case. If a random variable X has $\phi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ as its distribution function then we compute the probability of any event of the form $[a, b] = \{x \mid a \leq x \leq b\}$ by computing the area under the curve $\phi_{\mu,\sigma}(x)$ and above the interval $[a, b]$.

How can we compute this area? Tables and computer programs for areas below $y = \phi(x)$ are available. Unfortunately $\phi_{\mu,\sigma}$ and ϕ are different functions unless $\mu = 0$ and $\sigma = 1$. Fortunately, there is a simple recipe for converting one to the other. Let $h(t) = (t - \mu)/\sigma$. The area below $\phi_{\mu,\sigma}(x)$ above the interval $[a, b]$ equals the area below ϕ above the interval $[h(a), h(b)]$.

A farmer weighs some oranges from his crop and comes to you for help. From his data you notice that the mean weight is 8 ounces and the standard deviation is 0.67 ounces. You've read somewhere (Was it here?) that for such things a normal distribution is a good approximation to the weight. The farmer can sell oranges that weigh at least 9 ounces at a higher price per ounce, so he wants to estimate what fraction of his crop weighs at least 9 ounces. Using our recipe, $h(9) = (9 - 8)/0.67 = 1.5$. We know that the area under $\phi(x)$ for the interval $[0, 1.5]$ is 0.43. Since the area under $\phi(x)$ for $x < 0$ is 1/2, the area for $x \leq 1.5$ is $0.43 + 0.5 = 0.93$. Since these are the "underweight" oranges, the farmer can expect about 7% of his crop to be at least 9 ounces. □

Example 27 (Approximating the binomial distribution) Recall the binomial distribution from Example 24: $b(k; n, p)$ is the probability of exactly k heads in n tosses and p is the probability of a head on one toss. We derived the formula $b(k; n, p) = \binom{n}{k} p^k q^{n-k}$, where $q = 1 - p$. We also found, that for a binomial random variable X, $E(X) = np$ and $\mathrm{Var}(X) = npq$. How does the random variable behave when n is large? We already saw in Example 25 how to use the Poisson approximation when $E(X)$ is not large. When $E(X)$ and $\mathrm{Var}(X)$ are large, a better approximation is given by the normal distribution $\phi_{\mu,\sigma}$ with $\mu = np$ and $\sigma = \sqrt{npq}$.

Suppose that our book in Example 25 is a lot worse: About one word in ten is wrong. How can we estimate the probability of at most 30 errors on page 95? If the errors are independent, the distribution is a binomial with $p = 0.1$ and n equal to the number of words on page 95. We estimate that n is about 400. Thus we are dealing with $b(k; 400, 0.1)$. We have

$$\mu = 400 \times 0.1 = 40 \quad \text{and} \quad \sigma = \sqrt{400 \times 0.1 \times 0.9} = \sqrt{36} = 6.$$

Thus we want the area under $\phi(x)$ for $x < h(30) = (30 - 40)/6 \approx -1.5$. By the symmetry of ϕ, this is the area under $\phi(x)$ for $x > 1.5$, which is $0.5 - 0.43 = 7\%$.

We've done some rounding off here, which is okay since our estimates are rather crude. There are ways to improve the estimates, but we will not discuss them. □

Approximations like those in the preceding example are referred to as "limit theorems" in probability theory. The next example discusses the use of an important limit theorem, the Central Limit Theorem, for estimating how close an average of measurements is to the true value of a number. This is often used in experimental science when estimating a physical constant.

***Example 28 (The Central Limit Theorem and the normal distribution)** Suppose a student must estimate a quantity, say the distance between two buildings on campus. The student makes a number n of measurements. Each measurement can be thought of as a sample of a random variable. Call the random variable for measurement i X_i. If the student is not influenced by the previous measurements, we can think of the random variables as being independent and identically distributed. The obvious thing to do is average these measurements. How accurate is the result?

Let's phrase this in probabilistic terms. We have a new random variable given by $X = (X_1 + \cdots + X_n)/n$ and our average is a sample of the value of the random variable X. What can we say about X?

We can approximate X with a normal distribution. This approximation is a consequence of the *Central Limit Theorem*. Let A_1 be the average of the n measurements and let A_2 the average of the squares of the n measurements. Then we estimate μ and σ by A_1 and $\sqrt{(A_2 - (A_1)^2)/(n-1)}$, respectively.[11] We could now use $\phi_{\mu,\sigma}$ to estimate the distribution of the random variable X.

This can be turned around, $\phi_{\mu,\sigma}$ can also be used to estimate the true mean of the random variable X. You might have thought that A_1 was the mean. No. It is just the average of some observed values. Thus, the probability that the mean of X lies in $[\mu - \sigma, \mu + \sigma]$ equals $0.34 + 0.34 = 0.68$. □

We've looked at several different distributions: binomial, normal, Poisson and marginal. What do we use when? How are they related?

The binomial distribution occurs when you have a sequence of repeated independent events and want to know how many times a certain event occurred. For example, the probability of k heads in n tosses of a coin. The coin tosses are the repeated independent events and the heads are the events we are interested in.

The normal distribution is usually an approximation for estimating a number whose value is the sum of a lot of (nearly) independent random variables. For example, let X_i be 1 or 0 according as the i-th coin toss is a head or tail. We want to know the probability that $X_1 + X_2 + ... + X_n$ equals k. The exact answer is the binomial distribution. The normal distribution gives an approximation.

The Poisson distribution is associated with rare events. For example, if light bulbs fail at random (we're not being precise here) and have an average lifetime L, then the number

[11] The estimate for σ is a result from statistics. We cannot derive it here.

of failures in a time interval T is roughly Poisson if $\lambda = T/L$ is not too big or too small. Another example is errors in a text, which are rare and have a distribution associated with them that is like the binomial.

Unlike the previous three distributions, which exist by themselves, a marginal distribution is always derived from some given distribution. In our coin toss experiment, let X be the number of heads and let Y be the number of times two or more tails occur together. We could ask for the distribution given by $P(X = k \text{ and } Y = j)$. This is called a "joint distribution" for the random variables X and Y. *Given* the joint distribution, we could ask for the distribution of just one of the random variables. These are "marginal distributions" associated with the joint distribution. In this example, $P(X = k)$ and $P(Y = j)$ are marginal distributions. The first one (the probability of k heads) is the sum of $P(X = k \text{ and } Y = j)$ over all j and the second (the probability of two or more tails together happening j times) is the sum of $P(X = k \text{ and } Y = j)$ over all k.

Exercises for Section 4

4.1. A fair coin is tossed four times, recording H if heads, T if tails. Let X be the random variable defined by $X(t_1t_2t_3t_4) = |\{i \mid t_i = H\}|$. Let Y be the random variable defined by

$$Y(t_1t_2t_3t_4) = \begin{cases} 0, & \text{if } t_i = T \text{ for all } i = 1,\ 2,\ 3,\ 4; \\ \max\{k \mid H = t_i = t_{i+1} = \cdots = t_{i+k-1}, i = 1, 2, 3, 4\}, & \text{otherwise.} \end{cases}$$

The random variable X equals the number of H's. The random variable Y equals the length of the longest consecutive string of H's. Compute

(a) the joint distribution function $h_{X,Y}$,

(b) the marginal distributions f_X and f_Y,

(c) the covariance $\mathrm{Cov}(X, Y)$, and

(d) the correlation $\rho(X, Y)$.

Give an intuitive explanation of the value of $\rho(X, Y)$.

4.2. Let X and Y be random variables on a sample space U and let a and b be real numbers.

(a) Show that $\mathrm{Cov}(aX + bY, aX - bY)$ is $a^2\mathrm{Var}(X) - b^2\mathrm{Var}(Y)$.

(b) What is $\mathrm{Var}((aX - bY)(aX + bY))$?

4.3. Let X be random variable on a sample space U and let a and b be real numbers. What is $E((aX + b)^2)$ if

(a) X has the binomial distribution $b(k; n, p)$?

(b) X has the Poisson distribution $e^{-\lambda}\lambda^k/k!$?

4.4. A 100 page book has 200 misprints. If the misprints are distributed uniformly throughout the book, show how to use the Poisson approximation to the binomial distribution to calculate the probability of there being less than 4 misprints on page 8.

4.5. Let X and Y be independent random variables and let a and b be real numbers. Let $Z = aX + bY$. Then, for all $\epsilon > 0$, Tchebycheff's inequality gives an upper bound for $P(|Z - E(Z)| \geq \epsilon)$. Give this upper bound for the cases where

(a) X and Y have Poisson distribution $p(k;\gamma)$ and $p(k;\delta)$ respectively.

(b) X and Y have binomial distribution $p(k;n,r)$ and $p(k;n,s)$ respectively.

4.6. Each time a customer checks out at Super Save Groceries, a wheel with nine white and one black dot, symmetrically placed around the wheel, is spun. If the black dot is uppermost, the customer gets the least expensive item in their grocery cart for free. Assuming the probability of any dot being uppermost is 1/10, what is the probability that out of the first 1000 customers, between 85 and 115 customers get a free item? Write the formula for the exact solution and show how the normal distribution can be used to approximate this solution. You need not compute the values of the normal distribution.

4.7. Let $X_1, \ldots, X_n$ be independent random variables each having mean μ and variance σ^2. (These could arise by having one person repeat n times an experiment that produces an estimate of a number whose value is μ. See Example 28.) Let $X = (X_1 + \cdots + X_n)/n$.

(a) Compute the mean and variance of X.

(b) Explain why an observed value of X could be used as an estimate of μ.

(c) It turns out that the error we can expect in approximating μ with X is proportional to the value of σ_X. Suppose we want to reduce this expected error by a factor of 10. How much would we have to increase n. (In other words, how many more measurements would be needed.)

Multiple Choice Questions for Review

1. In each case some information is given about a function. In which case is the information *not* sufficient to define a function?

(a) $f \in \underline{4}^{\underline{3}}, \quad 2 \to 3, \quad 1 \to 4, \quad 3 \to 2.$

(b) $f \in \{>, <, +, ?\}^{\underline{3}}, \quad f = (?, <, +).$

(c) $f \in \underline{3}^{\{>,<,+,?\}}, \quad f = (3, 1, 2, 3).$

(d) $f \in \underline{3}^{\{>,<,+,?\}}, \quad f = (3, 1, 2, 3)$. Domain ordered as follows: $>$, $<$, $+$, ?.

(e) $f \in \{>, <, +, ?\}^{\underline{3}}, \quad f = (?, <, +)$. Domain ordered as follows: $3, 2, 1$.

2. The following function is in two line form: $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 9 & 3 & 7 & 2 & 6 & 4 & 5 & 1 & 8 \end{pmatrix}$. Which of the following is a correct cycle form for this function?

(a) $(1, 8, 9)(2, 3, 7, 5, 6, 4)$

(b) $(1, 9, 8)(2, 3, 5, 7, 6, 4)$

(c) $(1, 9, 8)(2, 3, 7, 5, 4, 6)$

(d) $(1, 9, 8)(2, 3, 7, 5, 6, 4)$.

(e) $(1, 9, 8)(3, 2, 7, 5, 6, 4)$

3. In each case some information about a function is given to you. Based on this information, which function is an injection?

(a) $f \in \underline{6}^{\underline{5}}, \quad \text{Coimage}(f) = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}\}$

(b) $f \in \underline{6}^{\underline{6}}, \quad \text{Coimage}(f) = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5, 6\}\}$

(c) $f \in \underline{5}^{\underline{5}}, \quad f^{-1}(2) = \{1, 3, 5\}, \quad f^{-1}(4) = \{2, 4\}\}$

(d) $f \in \underline{4}^{\underline{5}}, \quad \big|\text{Image}(f)\big| = 4$

(e) $f \in \underline{5}^{\underline{5}}, \quad \text{Coimage}(f) = \{\{1, 3, 5\}, \{2, 4\}\}$

4. The following function is in two line form: $f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 8 & 5 & 9 & 2 & 4 & 1 & 3 & 6 & 7 \end{pmatrix}$. Which of the following is a correct cycle form for $h = f^3 \circ f^{-1}$?

(a) $(1, 6, 8)(2, 3, 7)(5, 6, 4)$

(b) $(1, 6, 8)(2, 4, 5)(3, 7, 9)$

(c) $(1, 8, 6)(2, 3, 7)(5, 9, 4)$

(d) $(1, 9, 8)(2, 3, 5)(7, 6, 4)$

(e) $(8, 5, 9, 2, 4, 1, 3, 6, 7)$

5. The following permutation is in two line form: $f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 8 & 6 & 4 & 7 & 2 & 9 & 1 & 3 & 5 \end{pmatrix}$. The permutation $g = (1, 2, 3)$ is in cycle form. Let $h = f \circ g$ be the composition of f and g. Which of the following is a correct cycle form for h?

(a) $(1,6,9,5,2,4,7)(3,8)$

(b) $(1,8,3,4,7,2,6)(5,9)$

(c) $(1,8,3,7,4,2,6)(9,5)$

(d) $(1,8,4,3,7,2,6)(9,5)$

(e) $(8,6,4,7,9,1,2)(3,5)$

6. We want to find the smallest integer $n > 0$ such that, for every permutation f on $\underline{4}$, the function f^n is the identity function on $\underline{4}$. What is the value of n?

(a) 4 (b) 6 (c) 12 (d) 24 (e) It is impossible.

7. In the lexicographic list of all strictly decreasing functions in $\underline{9}^{\underline{5}}$, find the successor of 98432.

(a) 98431 (b) 98435 (c) 98521 (d) 98532 (e) 98543

8. The 16 consecutive points $0, 1, \ldots, 14, 15$ have 0 and 15 converted to exterior box boundaries. The interior box boundaries correspond to points $1, 5, 7, 9$. This configuration corresponds to

(a) 9 balls into 5 boxes

(b) 9 balls into 6 boxes

(c) 10 balls into 5 boxes

(d) 10 balls into 6 boxes

(e) 11 balls into 4 boxes

9. The 16 consecutive points $0, 1, \ldots, 14, 15$ have 0 and 15 converted to exterior box boundaries. The interior box boundaries correspond to the strictly increasing functions $1 \le x_1 < x_2 < x_3 < x_4 \le 14$ in lex order. How many configurations of balls into boxes come before the configuration $\bullet|\,|\,|\,|\,\bullet\bullet\bullet\bullet\bullet\bullet\bullet\bullet\bullet$? (Exterior box boundaries are not shown.)

(a) $\binom{13}{3}$ (b) $\binom{13}{4}$ (c) $\binom{14}{3}$ (d) $\binom{14}{4}$ (e) $\binom{15}{3}$

10. Suppose $f \in \underline{7}^{\underline{6}}$. How many such functions have $|\text{Image}(f)| = 4$?

(a) $S(7,4)$ (b) $S(7,4)(6)_4$ (c) $S(6,4)(7)_4$ (d) $S(4,7)(6)_4$ (e) $S(7,4)\,6!$

11. Let X be a random variable with distribution $b(k; n, p)$, $q = 1 - p$. Let $Y = (X+1)^2$. Then $E(Y) =$?

(a) $npq + (np+1)^2$

(b) $2npq + (np+1)^2$

(c) $npq + 2(np+1)^2$

(d) $(npq)^2 + (np+1)^2$

(e) $2npq(np+1)^2$

12. Let X and Y be independent random variables with distribution $b(k; n, a)$ and $b(k; n, b)$ respectively. Let $Z = X + 2Y$. Then, for all $\epsilon > 0$, Tchebycheff's inequality guarantees that $P(|Z - na - 2nb| \ge \epsilon)$ is always less than or equal to what?

(a) $(na(1-a)+nb(1-b))/\epsilon^2$

(b) $(na(1-a)+2nb(1-b))/\epsilon^2$

(c) $(na(1-a)+4nb(1-b))/\epsilon^2$

(d) $(na(1-a)+2nb(1-b))/\epsilon^3$

(e) $(na(1-a)+4nb(1-b))/\epsilon^3$

13. An 800 page book has 400 misprints. If the misprints are distributed uniformly throughout the book, and the Poisson approximation to the binomial distribution is used to calculate the probability of exactly 2 misprints on page 16, which of the following represents the correct use of the Poisson approximation?

(a) $e^{0.5}/8$ (b) $e^{-0.5}/8$ (c) $e^{0.5}/16$ (d) $e^{-0.5}/16$ (e) $e^{-0.5}/32$

14. For 40 weeks, once per hour during the 40 hour work week, an employee of Best Cars draws a ball from an urn that contains 1 black and 9 white balls. If black is drawn, a \$10 bill is tacked to a bulletin board. At the end of the 40 weeks, the money is given to charity. What is the expected amount of money given?

(a) 1000 (b) 1200 (c) 1400 (d) 1600 (e) 1800

15. For 40 weeks, once per hour during the 40 hour work week, an employee of Best Cars draws a ball from an urn that contains 1 black and 9 white balls. If black is drawn, \$10 is tacked to a bulletin board. At the end of the 40 weeks, the money is given to charity. Using the normal approximation, what interval under the standard normal curve should be used to get the area which equals the probability that \$1800 or more is given?

(a) from 1.67 to ∞

(b) from 0 to 1.67

(c) from 0.6 to ∞

(d) from 0 to 0.6

(e) from 0.6 to 1.67

16. A fair coin is tossed three times. Let X be the random variable which is one if the first throw is T (for tails) and the third throw is H (for heads), zero otherwise. Let Y denote the random variable that is one if the second and third throws are both H, zero otherwise. The covariance, $\mathrm{Cov}(X,Y)$ is

(a) 1/8 (b) $-1/8$ (c) 1/16 (d) $-1/16$ (e) 1/32

17. A fair coin is tossed three times. Let X be the random variable which is one if the first throw is T (for tails) and the third throw is H (for heads), zero otherwise. Let Y denote the random variable that is one if the second and third throws are both H, zero otherwise. The correlation, $\rho(X,Y)$ is

(a) 0 (b) 1/3 (c) $-1/3$ (d) 1/8 (e) $-1/8$

18. A fair coin is tossed three times and a T (for tails) or H (for heads) is recorded, giving us a 3-long list. Let X be the random variable which is zero if no T has another T adjacent to it, and is one otherwise. Let Y denote the random variable that counts

the number of T's in the three tosses. Let $h_{X,Y}$ denote the joint distribution of X and Y. $h_{X,Y}(1,2)$ equals

(a) 5/8 (b) 4/8 (c) 3/8 (d) 2/8 (e) 1/8

19. Which of the following is equal to $\text{Cov}(X+Y, X-Y)$, where X and Y are random variables on a sample space S?

(a) $\text{Var}(X) - \text{Var}(Y)$

(b) $\text{Var}(X^2) - \text{Var}(Y^2)$

(c) $\text{Var}(X^2) + 2\text{Cov}(X,Y) + \text{Var}(Y^2)$

(d) $\text{Var}(X^2) - 2\text{Cov}(X,Y) + \text{Var}(Y^2)$

(e) $(\text{Var}(X))^2 - (\text{Var}(Y))^2$

20. Which of the following is equal to $\text{Var}(2X-3Y)$, where X and Y are random variables on S?

(a) $4\text{Var}(X) + 12\text{Cov}(X,Y) + 9\text{Var}(Y)$

(b) $2\text{Var}(X) - 3\text{Var}(Y)$

(c) $2\text{Var}(X) + 6\text{Cov}(X,Y) + 3\text{Var}(Y)$

(d) $4\text{Var}(X) - 12\text{Cov}(X,Y) + 9\text{Var}(Y)$

(e) $2\text{Var}(X) - 6\text{Cov}(X,Y) + 3\text{Var}(Y)$

21. The strictly decreasing functions in $\underline{100}^{\underline{3}}$ are listed in lex order. How many are there before the function (9,5,4)?

(a) 18 (b) 23 (c) 65 (d) 98 (e) 180

22. All but one of the following have the same answer. Which one is different?

(a) The number of multisets of size 20 whose elements lie in $\underline{5}$.

(b) The number of strictly increasing functions from $\underline{20}$ to $\underline{24}$.

(c) The number of subsets of size 20 whose elements lie in $\underline{24}$.

(d) The number of weakly decreasing 4-lists made from $\underline{21}$.

(e) The number of strictly decreasing functions from $\underline{5}$ to $\underline{24}$.

23. Let X be a random variable with Poisson distribution $p(k;\lambda)$ Let $Y = (X+2)(X+1)$. What is the value of $E(Y)$?

(a) $\lambda^2 + 3\lambda + 1$

(b) $\lambda^2 + 3\lambda + 2$

(c) $\lambda^2 + 4\lambda + 2$

(d) $3\lambda^2 + 3\lambda + 2$

(e) $4\lambda^2 + 4\lambda + 2$

Answers: **1** (c), **2** (d), **3** (a), **4** (b), **5** (a), **6** (c), **7** (c), **8** (c), **9** (a), **10** (c), **11** (a), **12** (c), **13** (b), **14** (d), **15** (a), **16** (c), **17** (b), **18** (d), **19** (a), **20** (d), **21** (c), **22** (e), **23** (c).

Notation Index

Subject Index