Chapter 5

Probability

5.1 Introduction to Probability

Why do we study probability?

You have likely studied hashing as a way to store data (or keys to find data) in a way that makes it possible to access that data quickly. Recall that we have a table in which we want to store keys, and we compute a function h of our key to tell us which location (also known as a "slot" or a "bucket") in the table to use for the key. Such a function is chosen with the hope that it will tell us to put different keys in different places, but with the realization that it might not. If the function tells us to put two keys in the same place, we might put them into a linked list that starts at the appropriate place in the table, or we might have some strategy for putting them into some other place in the table itself. If we have a table with a hundred places and fifty keys to put in those places, there is no reason in advance why all fifty of those keys couldn't be assigned (hashed) to the same place in the table. However someone who is experienced with using hash functions and looking at the results will tell you you'd never see this in a million years. On the other hand that same person would also tell you that you'd never see all the keys hash into different locations in a million years either. In fact, it is far less likely that all fifty keys would hash into one place than that all fifty keys would hash into different places, but both events are quite unlikely. Being able to understand just how likely or unlikely such events are is our reason for taking up the study of probability.

In order to assign probabilities to events, we need to have a clear picture of what these events are. Thus we present a model of the kinds of situations in which it is reasonable to assign probabilities, and then recast our questions about probabilities into questions about this model. We use the phrase *sample space* to refer to the set of possible outcomes of a process. For now, we will deal with processes that have finite sample spaces. The process might be a game of cards, a sequence of hashes into a hash table, a sequence of tests on a number to see if it fails to be a prime, a roll of a die, a series of coin flips, a laboratory experiment, a survey, or any of many other possibilities. A set of elements in a sample space is called an *event*. For example, if a professor starts each class with a 3 question true-false quiz the sample space of all possible patterns of correct answers is

$$\{TTT, TTF, TFT, FTT, TFF, FTF, FFT, FFF\}.$$

The event of the first two answers being true is $\{TTT, TTF\}$. In order to compute probabilities we assign a *probability weight* $p(x)$ to each element of the sample space so that the weight represents what we believe to be the relative likelihood of that outcome. There are two rules we must follow in assigning weights. First the weights must be nonnegative numbers, and second the sum of the weights of all the elements in a sample space must be one. We define the *probability* $P(E)$ of the event E to be the sum of the weights of the elements of E. Algebraically we can write

$$P(E) = \sum_{x:x\in E} p(x). \tag{5.1}$$

We read this as $p(E)$ equals the sum, over all x such that x is in E, of $p(x)$.

Notice that a probability function P on a sample space S satisfies the rules[1]

1. $P(A) \geq 0$ for any $A \subseteq S$.
2. $P(S) = 1$.
3. $P(A \cup B) = P(A) + P(B)$ for any two disjoint events A and B.

The first two rules reflect our rules for assigning weights above. We say that two events are disjoint if $A \cap B = \emptyset$. The third rule follows directly from the definition of disjoint and our definition of the probability of an event. A function P satisfying these rules is called a *probability distribution* or a *probability measure*.

In the case of the professor's three question quiz, it is natural to expect each sequence of trues and falses to be equally likely. (A professor who showed any pattern of preferences would end up rewarding a student who observed this pattern and used it in educated guessing.) Thus it is natural to assign equal weight 1/8 to each of the eight elements of our quiz sample space. Then the *probability* of an event E, which we denote by $P(E)$, is the sum of the weights of its elements. Thus the probability of the event "the first answer is T" is $\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{1}{2}$. The event "There is at exactly one True" is $\{TFF, FTF, FFT\}$, so P(there is exactly one True) is 3/8.

Some examples of probability computations

Exercise 5.1-1 Try flipping a coin five times. Did you get at least one head? Repeat five coin flips a few more times! What is the probability of getting at least one head in five flips of a coin? What is the probability of no heads?

Exercise 5.1-2 Find a good sample space for rolling two dice. What weights are appropriate for the members of your sample space? What is the probability of getting a 6 or 7 total on the two dice? Assume the dice are of different colors. What is the probability of getting less than 3 on the red one and more than 3 on the green one?

Exercise 5.1-3 Suppose you hash a list of n keys into a hash table with 20 locations. What is an appropriate sample space, and what is an appropriate weight function? (Assume the keys and the hash function are not in any special relationship to the

[1] These rules are often called "the axioms of probability." For a finite sample space, we could show that if we started with these axioms, our definition of probability in terms of the weights of individual elements of S is the only definition possible. That is, for any other definition, the probabilities we would compute would still be the same if we take $w(x) = P(\{x\})$.

number 20.) If n is three, what is the probability that all three keys hash to different locations? If you hash ten keys into the table, what is the probability that at least two keys have hashed to the same location? We say two keys *collide* if they hash to the same location. How big does n have to be to insure that the probability is at least one half that has been at least one collision?

In Exercise 5.1-1 a good sample space is the set of all 5-tuples of Hs and Ts. There are 32 elements in the sample space, and no element has any reason to be more likely than any other, so a natural weight to use is $\frac{1}{32}$ for each element of the sample space. Then the event of at least one head is the set of all elements but $TTTTT$. Since there are 31 elements in this set, its probability is $\frac{31}{32}$. This suggests that you should have observed at least one head pretty often!

Complementary probabilities

The probability of no heads is the weight of the set $\{TTTTT\}$, which is $\frac{1}{32}$. Notice that the probabilities of the event of "no heads" and the opposite event of "at least one head" add to one. This observation suggests a theorem. The *complement* of an event E in a sample space S, denoted by $S-E$, is the set of all outcomes in S but not E. The theorem tells us how to compute the probability of the complement of an event from the probability of the event.

Theorem 5.1 *If two events E and F are complementary, that is they have nothing in common ($E \cap F = \emptyset$) and their union is the whole sample space ($E \cup F = S$), then*

$$P(E) = 1 - P(F).$$

Proof: The sum of all the probabilities of all the elements of the sample space is one, and since we can break this sum into the sum of the probabilities of the elements of E plus the sum of the probabilities of the elements of F, we have

$$P(E) + P(F) = 1,$$

which gives us $P(E) = 1 - P(F)$.■

For Exercise 5.1-2 a good sample space would be pairs of numbers (a, b) where $(1 \leq a, b \leq 6)$. By the product principle[2], the size of this sample space is $6 \cdot 6 = 36$. Thus a natural weight for each ordered pair is $\frac{1}{36}$. How do we compute the probability of getting a sum of six or seven? There are 5 ways to roll a six and 6 ways to roll a seven, so our event has eleven elements each of weight 1/36. Thus the probability of our event is is 11/36. For the question about the red and green dice, there are two ways for the red one to turn up less than 3, and three ways for the green one to turn up more than 3. Thus, the event of getting less than 3 on the red one and greater than 3 on the green one is a set of size $2 \cdot 3 = 6$ by the product principle. Since each element of the event has weight 1/36, the event has probability 6/36 or 1/6.

[2]From Section 1.1.

Probability and hashing

In Exercise 5.1-3 an appropriate sample space is the set of n-tuples of numbers between 1 and 20. The first entry in an n-tuple is the position our first key hashes to, the second entry is the position our second key hashes to, and so on. Thus each n tuple represents a possible hash function, and each hash function, applied to our keys, would give us one n-tuple. The size of the sample space is 20^n (why?), so an appropriate weight for an n-tuple is $1/20^n$. To compute the probability of a collision, we will first compute the probability that all keys hash to different locations and then apply Theorem 5.1 which tells us to subtract this probability from 1 to get the probability of collision.

To compute the probability that all keys hash to different locations we consider the event that all keys hash to different locations. This is the set of n tuples in which all the entries are different. (In the terminology of functions, these n-tuples correspond to one-to-one hash functions). There are 20 choices for the first entry of an n-tuple in our event. Since the second entry has to be different, there are 19 choices for the second entry of this n-tuple. Similarly there are 18 choices for the third entry (it has to be different from the first two), 17 for the fourth, and in general $20 - i + 1$ possibilities for the ith entry of the n-tuple. Thus we have

$$20 \cdot 19 \cdot 18 \cdot \cdots \cdot (20 - n + 1) = 20^{\underline{n}}$$

elements of our event.[3] Since each element of this event has weight $1/20^n$, the probability that all the keys hash to different locations is

$$\frac{20 \cdot 19 \cdot 18 \cdot \cdots \cdot (20 - n + 1)}{20^n} = \frac{20^{\underline{n}}}{20^n.}$$

In particular if n is 3 the probability is $(20 \cdot 19 \cdot 18)/20^3 = .855$.

We show the values of this function for n between 0 and 20 in Table 5.1. Note how quickly the probability of getting a collision grows. As you can see with $n = 10$, the probability that there have been no collisions is about .065, so the probability of at least one collision is .935.

If $n = 5$ this number is about .58, and if $n = 6$ this number is about .43. By Theorem 5.1 the probability of a collision is one minus the probability that all the keys hash to different locations. Thus if we hash six items into our table, the probability of a collision is more than 1/2. Our first intuition might well have been that we would need to hash ten items into our table to have probability 1/2 of a collision. This example shows the importance of supplementing intuition with careful computation!

The technique of computing the probability of an event of interest by first computing the probability of its complementary event and then subtracting from 1 is very useful. You will see many opportunities to use it, perhaps because about half the time it is easier to compute directly the probability that an event doesn't occur than the probability that it does. We stated Theorem 5.1 as a theorem to emphasize the importance of this technique.

The Uniform Probability Distribution

In all three of our exercises it was appropriate to assign the same weight to all members of our sample space. We say P is the *uniform probability measure* or *uniform probability distribution*

[3] Using the notation for falling factorial powers that we introduced in Section 1.2.

n	Prob of empty slot	Prob of no collisions
1	1	1
2	0.95	0.95
3	0.9	0.855
4	0.85	0.72675
5	0.8	0.5814
6	0.75	0.43605
7	0.7	0.305235
8	0.65	0.19840275
9	0.6	0.11904165
10	0.55	0.065472908
11	0.5	0.032736454
12	0.45	0.014731404
13	0.4	0.005892562
14	0.35	0.002062397
15	0.3	0.000618719
16	0.25	0.00015468
17	0.2	3.09359E-05
18	0.15	4.64039E-06
19	0.1	4.64039E-07
20	0.05	2.3202E-08

Table 5.1: The probabilities that all elements of a set hash to different entries of a hash table of size 20.

when we assign the same probability to all members of our sample space. The computations in the exercises suggest another useful theorem.

Theorem 5.2 *Suppose P is the uniform probability measure defined on a sample space S. Then for any event E,*

$$P(E) = |E|/|S|,$$

the size of E divided by the size of S.

Proof: Let $S = \{x_1, x_2, \ldots, x_{|S|}\}$. Since P is the uniform probability measure, there must be some value p such that for each $x_i \in S, P(x_i) = p$. Combining this fact with the second and third probability rules, we obtain

$$\begin{aligned} 1 &= P(S) \\ &= P(x_1 \cup x_2 \cup \cdots \cup x_{|S|}) \\ &= P(x_1) + P(x_2) + \ldots + P(x_{|S|}) \\ &= p|S| \ . \end{aligned}$$

Equivalently

$$p = \frac{1}{|S|} \ . \tag{5.2}$$

E is a subset of S with $|E|$ elements and therefore

$$P(E) = \sum_{x_i \in E} p(x_i) = |E|p \ . \tag{5.3}$$

Combining equations 5.2 and 5.3 gives that $P(E) = |E|p = |E|(1/|S|) = |E|/|S|$. ■

Exercise 5.1-4 What is the probability of an odd number of heads in three tosses of a coin? Use Theorem 5.2.

Using a sample space similar to that of first example (with "T" and "F" replaced by "H" and "T"), we see there are three sequences with one H and there is one sequence with three H's. Thus we have four sequences in the event of "an odd number of heads come up." There are eight sequences in the sample space, so the probability is $\frac{4}{8} = \frac{1}{2}$.

It is comforting that we got one half because of a symmetry inherent in this problem. In flipping coins, heads and tails are equally likely. Further if we are flipping 3 coins, an odd number of heads implies an even number of tails. Therefore, the probability of an odd number of heads, even number of heads, odd number of tails and even number of tails must all be the same. Applying Theorem 5.1 we see that the probability must be 1/2.

A word of caution is appropriate here. Theorem 5.2 applies only to probabilities that come from the equiprobable weighting function. The next example shows that it does not apply in general.

Exercise 5.1-5 A sample space consists of the numbers 0, 1, 2 and 3. We assign weight $\frac{1}{8}$ to 0, $\frac{3}{8}$ to 1, $\frac{3}{8}$ to 2, and $\frac{1}{8}$ to 3. What is the probability that an element of the sample space is positive? Show that this is not the result we would obtain by using the formula of Theorem 5.2.

The event "x is positive" is the set $E = \{1, 2, 3\}$. The probability of E is

$$P(E) = P(1) + P(2) + P(3) = \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = \frac{7}{8} \ .$$

However, $\frac{|E|}{|S|} = \frac{3}{4}$.

The previous exercise may seem to be "cooked up" in an unusual way just to prove a point. In fact that sample space and that probability measure could easily arise in studying something as simple as coin flipping.

Exercise 5.1-6 Use the set $\{0, 1, 2, 3\}$ as a sample space for the process of flipping a coin three times and counting the number of heads. Determine the appropriate probability weights $P(0)$, $P(1)$, $P(2)$, and $P(3)$.

There is one way to get the outcome 0, namely tails on each flip. There are, however, three ways to get 1 head and three ways to get two heads. Thus $P(1)$ and $P(2)$ should each be three times $P(0)$. There is one way to get the outcome 3—heads on each flip. Thus $P(3)$ should equal $P(0)$. In equations this gives $P(1) = 3P(0)$, $P(2) = 3P(0)$, and $P(3) = P(0)$. We also have the equation saying all the weights add to one, $P(0) + P(1) + P(2) + P(3) = 1$. There is one and only one solution to these equations, namely $P(0) = \frac{1}{8}$, $P(1) = \frac{3}{8}$, $P(2) = \frac{3}{8}$, and $P(3) = \frac{1}{8}$. Do you notice a relationship between $P(x)$ and the binomial coefficient $\binom{3}{x}$ here? Can you predict the probabilities of 0, 1, 2, 3, and 4 heads in four flips of a coin?

Together, the last two exercises demonstrate that we must be careful not to apply Theorem 5.2 unless we are using the uniform probability measure.

Important Concepts, Formulas, and Theorems

1. *Sample Space.* We use the phrase *sample space* to refer to the set of possible outcomes of a process.

2. *Event.* A set of elements in a sample space is called an *event*.

3. *Probability.* In order to compute probabilities we assign a *weight* to each element of the sample space so that the weight represents what we believe to be the relative likelihood of that outcome. There are two rules we must follow in assigning weights. First the weights must be nonnegative numbers, and second the sum of the weights of all the elements in a sample space must be one. We define the *probability* $P(E)$ of the event E to be the sum of the weights of the elements of E.

4. *The axioms of Probability.* Three rules that a probability measure on a finite sample space must satisfy could actually be used to define what we mean by probability.

   (a) $P(A) \geq 0$ for any $A \subseteq S$.

   (b) $P(S) = 1$.

   (c) $P(A \cup B) = P(A) + P(B)$ for any two disjoint events A and B.

5. *Probability Distribution.* A function which assigns a probability to each member of a sample space is called a (discrete) *probability distribution*.

6. *Complement.* The *complement* of an event E in a sample space S, denoted by $S - E$, is the set of all outcomes in S but not E.

7. The Probabilities of Complementary Events. If two events E and F are complementary, that is they have nothing in common ($E \cap F = \emptyset$), and their union is the whole sample space ($E \cup F = S$), then
$$P(E) = 1 - P(F).$$

8. *Collision, Collide (in Hashing).* We say two keys *collide* if they hash to the same location.

9. *Uniform Probability Distribution.* We say P is the *uniform probability measure* or *uniform probability distribution* when we assign the same probability to all members of our sample space.

10. *Computing Probabilities with the Uniform Distribution.* Suppose P is the uniform probability measure defined on a sample space S. Then for any event E,
$$P(E) = |E|/|S|,$$
the size of E divided by the size of S. This *does not* apply to general probability distributions.

Problems

1. What is the probability of exactly three heads when you flip a coin five times? What is the probability of three or more heads when you flip a coin five times?

2. When we roll two dice, what is the probability of getting a sum of 4 or less on the tops?

3. If we hash 3 keys into a hash table with ten slots, what is the probability that all three keys hash to different slots? How big does n have to be so that if we hash n keys to a hash table with 10 slots, the probability is at least a half that some slot has at least two keys hash to it? How many keys do we need to have probability at least two thirds that some slot has at least two keys hash to it?

4. What is the probability of an odd sum when we roll three dice?

5. Suppose we use the numbers 2 through 12 as our sample space for rolling two dice and adding the numbers on top. What would we get for the probability of a sum of 2, 3, or 4, if we used the equiprobable measure on this sample space. Would this make sense?

6. Two pennies, a nickel and a dime are placed in a cup and a first coin and a second coin are drawn.

   (a) Assuming we are sampling without replacement (that is, we don't replace the first coin before taking the second) write down the sample space of all ordered pairs of letters P, N, and D that represent the outcomes. What would you say are the appropriate weights for the elements of the sample space?

   (b) What is the probability of getting eleven cents?

7. Why is the probability of five heads in ten flips of a coin equal to $\frac{63}{256}$?

8. Using 5-element sets as a sample space, determine the probability that a "hand" of 5 cards chosen from an ordinary deck of 52 cards will consist of cards of the same suit.

9. Using 5 element permutations as a sample space, determine the probability that a "hand" of 5 cards chosen from an ordinary deck of 52 cards will have all the cards from the same suit

10. How many five-card hands chosen from a standard deck of playing cards consist of five cards in a row (such as the nine of diamonds, the ten of clubs, jack of clubs, queen of hearts, and king of spades)? Such a hand is called a straight. What is the probability that a five-card hand is a straight? Explore whether you get the same answer by using five element sets as your model of hands or five element permutations as your model of hands.

11. A student taking a ten-question, true-false diagnostic test knows none of the answers and must guess at each one. Compute the probability that the student gets a score of 80 or higher. What is the probability that the grade is 70 or lower?

12. A die is made of a cube with a square painted on one side, a circle on two sides, and a triangle on three sides. If the die is rolled twice, what is the probability that the two shapes we see on top are the same?

13. Are the following two events equally likely? Event 1 consists of drawing an ace and a king when you draw two cards from among the thirteen spades in a deck of cards and event 2 consists of drawing an ace and a king when you draw two cards from the whole deck.

14. There is a retired professor who used to love to go into a probability class of thirty or more students and announce "I will give even money odds that there are two people in this classroom with the same birthday." With thirty students in the room, what is the probability that all have different birthdays? What is the minimum number of students that must be in the room so that the professor has at least probability one half of winning the bet? What is the probability that he wins his bet if there are 50 students in the room. Does this probability make sense to you? (There is no wrong answer to that question!) Explain why or why not.

5.2 Unions and Intersections

The probability of a union of events

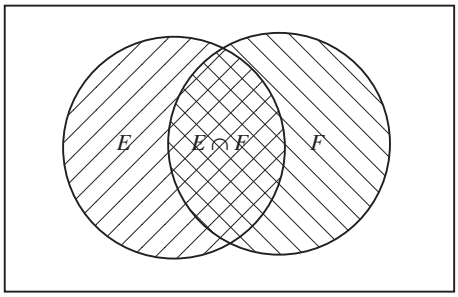
Exercise 5.2-1 If you roll two dice, what is the probability of an even sum or a sum of 8 or more?

Exercise 5.2-2 In Exercise 5.2-1, let E be the event "even sum" and let F be the event "8 or more." We found the probability of the union of the events E and F. Why isn't it the case that $P(E \cup F) = P(E) + P(F)$? What weights appear twice in the sum $P(E) + P(F)$? Find a formula for $P(E \cup F)$ in terms of the probabilities of E, F, and $E \cap F$. Apply this formula to Exercise 5.2-1. What is the value of expressing one probability in terms of three?

Exercise 5.2-3 What is $P(E \cup F \cup G)$ in terms of probabilities of the events E, F, and G and their intersections?

In the sum $P(E) + P(F)$ the weights of elements of $E \cap F$ each appear twice, while the weights of all other elements of $E \cup F$ each appear once. We can see this by looking at a diagram called a Venn Diagram, as in Figure 5.1. In a *Venn diagram*, the rectangle represents the sample space, and the circles represent the events. If we were to shade both E and F, we would wind

Figure 5.1: A Venn diagram for two events.



up shading the region $E \cap F$ twice. In Figure 5.2, we represent that by putting numbers in the regions, representing how many times they are shaded. This illustrates why the sum $P(E)+P(F)$ includes the probability weight of each element of $E \cap F$ twice. Thus to get a sum that includes the probability weight of each element of $E \cup F$ exactly once, we have to subtract the weight of $E \cap F$ from the sum $P(E) + P(F)$. This is why

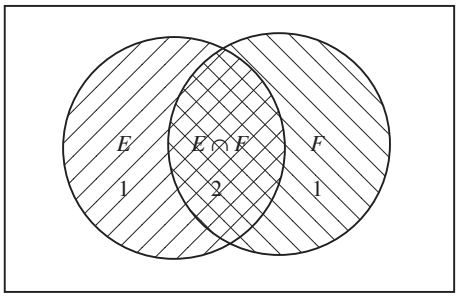
$$P(E \cup F) = P(E) + P(F) - P(E \cap F) \tag{5.4}$$

.

We can now apply this to Exercise 5.2-1 by noting that the probability of an even sum is 1/2, while the probability of a sum of 8 or more is

$$\frac{1}{36} + \frac{2}{36} + \frac{3}{36} + \frac{4}{36} + \frac{5}{36} = \frac{15}{36}.$$

Figure 5.2: If we shade each of E and F once, then we shade $E \cap F$ twice



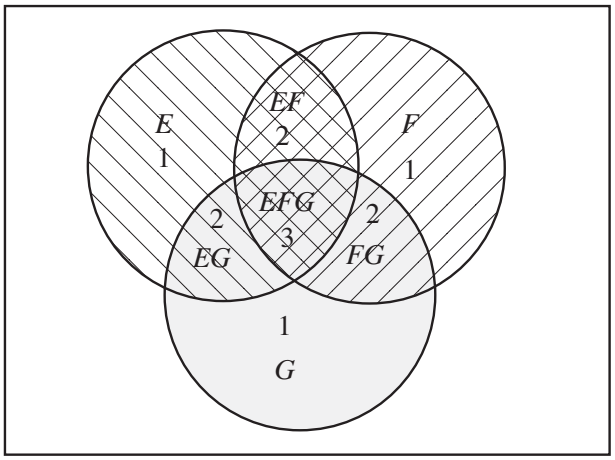
From a similar sum, the probability of an even sum of 8 or more is 9/36, so the probability of a sum that is even or is 8 or more is

$$\frac{1}{2} + \frac{15}{36} - \frac{9}{36} = \frac{2}{3}.$$

(In this case our computation merely illustrates the formula; with less work one could add the probability of an even sum to the probability of a sum of 9 or 11.) In many cases, however, probabilities of individual events and their intersections are more straightforward to compute than probabilities of unions (we will see such examples later in this section), and in such cases our formula is quite useful.

Now let's consider the case for three events and draw a Venn diagram and fill in the numbers for shading all E, F, and G. So as not to crowd the figure we use EF to label the region corresponding to $E \cap F$, and similarly label other regions. Doing so we get Figure 5.3. Thus we

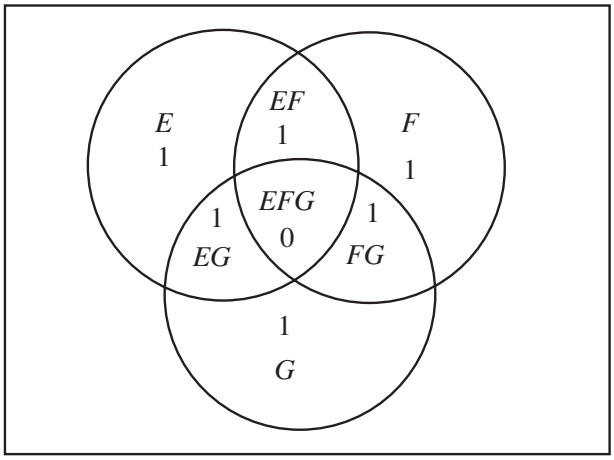
Figure 5.3: The number of ways the intersections are shaded when we shade E, F, and G.



have to figure out a way to subtract from $P(E) + P(F) + P(G)$ the weights of elements in the regions labeled EF, FG and EG once, and the the weight of elements in the region labeled EFG twice. If we subtract out the weights of elements of each of $E \cap F$, $F \cap G$, and $E \cap G$, this does more than we wanted to do, as we subtract the weights of elements in EF, FG and EG once

but the weights of elements in of EFG three times, leaving us with Figure 5.4. We then see that

Figure 5.4: The result of removing the weights of each intersection of two sets.



all that is left to do is to add weights of elements in the $E \cap F \cap G$ back into our sum. Thus we have that

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

Principle of inclusion and exclusion for probability

From the last two exercises, it is natural to guess the formula

$$P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i) - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} P(E_i \cap E_j) + \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} P(E_i \cap E_j \cap E_k) - \dots . \quad (5.5)$$

All the sum signs in this notation suggest that we need some new notation to describe sums. We are now going to make a (hopefully small) leap of abstraction in our notation and introduce notation capable of compactly describing the sum described in the previous paragraph. This notation is an extension of the one we introduced in Equation 5.1. We use

$$\sum_{\substack{i_1, i_2, \dots, i_k: \\ 1 \le i_1 < i_2 < \dots < i_k \le n}} P(E_{i_1} \cap E_{i_2} \cap \dots E_{i_k}) \quad (5.6)$$

to stand for the sum, over all sequences $i_1, i_2, \dots i_k$ of integers between 1 and n of the probabilities of the sets $E_{i_1} \cap E_{i_2} \dots \cap E_{i_k}$. More generally, $\sum_{\substack{i_1, i_2, \dots, i_k: \\ 1 \le i_1 < i_2 < \dots < i_k \le n}} f(i_1, i_2, \dots, i_k)$ is the sum of $f(i_1, i_2, \dots, i_k)$ over all increasing sequences of k numbers between 1 and n.

Exercise 5.2-4 To practice with notation, what is $\sum_{\substack{i_1, i_2, i_3: \\ 1 \le i_1 < i_2 < i_3 \le 4}} i_1 + i_2 + i_3$?

The sum in Exercise 5.2-4 is $1+2+3+1+2+4+1+3+4+2+3+4 = 3(1+2+3+4) = 30$.

With this understanding of the notation in hand, we can now write down a formula that captures the idea in Equation 5.5 more concisely. Notice that in Equation 5.5 we include probabilities of single sets with a plus sign, probabilities of intersections of two sets with a minus sign, and in general, probabilities of intersections of any even number of sets with a minus sign and probabilities of intersections of any odd number of sets (including the odd number one) with a plus sign. Thus if we are intersecting k sets, the proper coefficient for the probability of the intersection of these sets is $(-1)^{k+1}$ (it would be equally good to use $(-1)^{k-1}$, and correct but silly to use $(-1)^{k+3}$). This lets us translate the formula of Equation 5.5 to Equation 5.7 in the theorem, called the *Principle of Inclusion and Exclusion for Probability*, that follows. We will give two completely different proofs of the theorem, one of which is a nice counting argument but is a bit on the abstract side, and one of which is straightforward induction, but is complicated by the fact that it takes a lot of notation to say what is going on.

Theorem 5.3 (Principle of Inclusion and Exclusion for Probability) *The probability of the union $E_1 \cup E_2 \cup \cdots \cup E_n$ of events in a sample space S is given by*

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{\substack{i_1, i_2, \ldots, i_k: \\ 1 \le i_1 < i_2 < \cdots < i_k \le n}} P\left(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k}\right). \tag{5.7}$$

First Proof: Consider an element x of $\bigcup_{i=1}^{n} E_i$. Let $E_{i_1}, E_{i_2}, \ldots E_{i_k}$ be the set of all events E_i of which x is a member. Let $K = \{i_1, i_2, \ldots, i_k\}$. Then x is in the event $E_{j_1} \cap E_{j_2} \cap \cdots \cap E_{j_m}$ if and only if $\{j_1, j_2 \ldots j_m\} \subseteq K$. Why is this? If there is a j_r that is not in K, then $x \notin E_j$ and thus $x \notin E_{j_1} \cap E_{j_2} \cap \cdots \cap E_{j_m}$. Notice that every x in $\bigcup_{i=1}^{n} E_i$ is in at least one E_i, so it is in at least one of the sets $E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k}$.

Recall that we define $P\left(E_{j_1} \cap E_{j_2} \cap \cdots \cap E_{j_m}\right)$ to be the sum of the probability weights $p(x)$ for $x \in E_{j_1} \cap E_{j_2} \cap \cdots \cap E_{j_m}$. Suppose we substitute this sum of probability weights for $P\left(E_{j_1} \cap E_{j_2} \cap \cdots \cap E_{j_m}\right)$ on the right hand side of Equation 5.7. Then the right hand side becomes a sum of terms, each of which is plus or minus a probability weight. The sum of all the terms involving $p(x)$ on the right hand side of Equation 5.7 includes a term involving $p(x)$ for each nonempty subset $\{j_1, j_2, \ldots, j_m\}$ of K, and no other terms involving $p(x)$. The coefficient of the probability weight $p(x)$ in the term for the subset $\{j_1, j_2, \ldots, j_m\}$ is $(-1)^{m+1}$. Since there are $\binom{k}{m}$ subsets of K of size m, the sum of the terms involving $p(x)$ will therefore be

$$\sum_{m=1}^{k} (-1)^{m+1} \binom{k}{m} p(x) = \left(-\sum_{m=0}^{k} (-1)^m \binom{k}{m} p(x)\right) + p(x) = 0 \cdot p(x) + p(x) = p(x),$$

because $k \ge 1$ and thus by the binomial theorem, $\sum_{j=0}^{k} \binom{k}{j} (-1)^j = (1-1)^k = 0$. This proves that for each x, the sum of all the terms involving $p(x)$ after we substitute the sum of probability weights into Equation 5.7 is exactly $p(x)$. We noted above that for every x in $\cup_{i=1}^{n} E_i$ appears in at least one of the sets $E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k}$. Thus the right hand side of Equation 5.7 is the sum of every $p(x)$ such that x is in $\cup_{i=1}^{n} E_i$. By definition, this is the left-hand side of Equation 5.7. ■

Second Proof: The proof is simply an application of mathematical induction using Equation 5.4. When $n = 1$ the formula is true because it says $P(E_1) = P(E_1)$. Now suppose inductively

that for any family of $n-1$ sets $F_1, F_2, \ldots, F_{n-1}$

$$P\left(\bigcup_{i=1}^{n-1} F_i\right) = \sum_{k=1}^{n-1}(-1)^{k+1} \sum_{\substack{i_1,i_2,i\ldots,i_k: \\ 1\le i_1<i_2<\cdots<i_k\le n-1}} P(F_{i_1} \cap F_{i_2} \cap \ldots \cap F_{i_k}) \tag{5.8}$$

If in Equation 5.4 we let $E = E_1 \cup \ldots \cup E_{n-1}$ and $F = E_m$, we may apply Equation 5.4 to to compute $P\left(\cup_{i=1}^{n} E_i\right)$ as follows:

$$P\left(\bigcup_{i=1}^{n} E_i\right) = P\left(\bigcup_{i=1}^{n-1} E_i\right) + P(E_n) - P\left(\left(\bigcup_{i=1}^{n-1} E_i\right) \cap E_n\right). \tag{5.9}$$

By the distributive law,

$$\left(\bigcup_{i=1}^{n-1} E_i\right) \cap E_n = \bigcup_{i=1}^{n-1} \left(E_i \cap E_n\right),$$

and substituting this into Equation 5.9 gives

$$P\left(\bigcup_{i=1}^{n} E_i\right) = P\left(\bigcup_{i=1}^{n-1} E_i\right) + P\left(E_n\right) - P\left(\bigcup_{i=1}^{n-1} \left(E_i \cap E_n\right)\right).$$

Now we use the inductive hypothesis (Equation 5.8) in two places to get

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} E_i\right) &= \sum_{k=1}^{n-1}(-1)^{k+1} \sum_{\substack{i_1,i_2,i\ldots,i_k: \\ 1\le i_1<i_2<\cdots<i_k\le n-1}} P(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k}) \\ &+ P(E_n) \\ &- \sum_{k=1}^{n-1}(-1)^{k+1} \sum_{\substack{i_1,i_2,i\ldots,i_k: \\ 1\le i_1<i_2<\cdots<i_k\le n-1}} P(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k} \cap E_n). \end{aligned}$$

The first summation on the right hand side sums $(-1)^{k+1} P\left(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k}\right)$ over all lists $i_1, i_2, \ldots, i_k$ that *do not* contain n, while the $P(E_n)$ and the second summation work together to sum $(-1)^{k+1} P\left(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k}\right)$ over all lists $i_1, i_2, \ldots, i_k$ that *do* contain n. Therefore,

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{k=1}^{n}(-1)^{k+1} \sum_{\substack{i_1,i_2,i\ldots,i_k: \\ 1\le i_1<i_2<\cdots<i_k\le n}} P(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k}).$$

Thus by the principle of mathematical induction, this formula holds for all integers $n > 0$. ■

Exercise 5.2-5 At a fancy restaurant n students check their backpacks. They are the only ones to check backpacks. A child visits the checkroom and plays with the check tickets for the backpacks so they are all mixed up. If there are 5 students named Judy, Sam, Pat, Jill, and Jo, in how many ways may the backpacks be returned so that Judy gets her own backpack (and maybe some other students do, too)? What is the probability that this happens? What is the probability that Sam gets his backpack (and maybe some other students do, too)? What is the probability that Judy and Sam both get their own backpacks (and maybe some other students do, too)? For any particular

two element set of students, what is the probability that these two students get their own backpacks (and maybe some other students do, too)? What is the probability that at least one student gets his or her own backpack? What is the probability that no students get their own backpacks? What do you expect the answer will be for the last two questions for n students? This classic problem is often stated using hats rather than backpacks (quaint, isn't it?), so it is called the *hatcheck problem*. It is also known as the *derangement problem*; a *derangement* of a set being a one-to-one function from a set onto itself (i.e., a bijection) that sends each element to something not equal to it.

For Exercise 5.2-5, let E_i be the event that person i on our list gets the right backpack. Thus E_1 is the event that Judy gets the correct backpack and E_2 is the event that Sam gets the correct backpack. The event $E_1 \cap E_2$ is the event that Judy *and* Sam get the correct backpacks (and maybe some other people do). In Exercise 5.2-5, there are 4! ways to pass back the backpacks so that Judy gets her own, as there are for Sam or any other single student. Thus $P(E_1) = P(E_i) = \frac{4!}{5!}$. For any particular two element subset, such as Judy and Sam, there are 3! ways that these two people may get their backpacks back. Thus, for each i and j, $P(E_i \cap E_j) = \frac{3!}{5!}$. For a particular group of k students the probability that each one of these k students gets his or her own backpack back is $(5-k)!/5!$. If E_i is the event that student i gets his or her own backpack back, then the probability of an intersection of k of these events is $(5-k)!/5!$. The probability that at least one person gets his or her own backpack back is the probability of $E_1 \cup E_2 \cup E_3 \cup E_4 \cup E_5$. Then by the principle of inclusion and exclusion, the probability that at least one person gets his or her own backpack back is

$$P(E_1 \cup E_2 \cup E_3 \cup E_4 \cup E_5) = \sum_{k=1}^{5} (-1)^{k+1} \sum_{\substack{i_1, i_2, \ldots, i_k: \\ 1 \le i_1 < i_2 < \cdots < i_k \le 5}} P(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k}). \tag{5.10}$$

As we argued above, for a set of k people, the probability that all k people get their backpacks back is $(5-k)!/5!$. In symbols, $P(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k}) = \frac{(5-k)!}{5!}$. Recall that there are $\binom{5}{k}$ sets of k people chosen from our five students. That is, there are $\binom{5}{k}$ lists $i_1, i_2, \ldots i_k$ with $1 < i_1 < i_2 < \cdots < i_k \le 5$. Thus, we can rewrite the right hand side of the Equation 5.10 as

$$\sum_{k=1}^{5} (-1)^{k+1} \binom{5}{k} \frac{(5-k)!}{5!}.$$

This gives us

$$\begin{aligned} P(E_1 \cup E_2 \cup E_3 \cup E_4 \cup E_5) &= \sum_{k=1}^{5} (-1)^{k-1} \binom{5}{k} \frac{(5-k)!}{5!} \\ &= \sum_{k=1}^{5} (-1)^{k-1} \frac{5!}{k!(5-k)!} \frac{(5-k)!}{5!} \\ &= \sum_{k=1}^{5} (-1)^{k-1} \frac{1}{k!} \\ &= 1 - \frac{1}{2} + \frac{1}{3!} - \frac{1}{4!} + \frac{1}{5!}. \end{aligned}$$

The probability that nobody gets his or her own backpack is 1 minus the probability that someone does, or

$$\frac{1}{2} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!}.$$

To do the general case of n students, we simply substitute n for 5 and get that the probability that at least one person gets his or her own backpack is

$$\sum_{i=1}^{n}(-1)^{i-1}\frac{1}{i!} = 1 - \frac{1}{2} + \frac{1}{3!} - \cdots + \frac{(-1)^{n-1}}{n!}$$

and the probability that nobody gets his or her own backpack is 1 minus the probability above, or

$$\sum_{i=2}^{n}(-1)^{i}\frac{1}{i!} = \frac{1}{2} - \frac{1}{3!} + \cdots + \frac{(-1)^{n}}{n!}. \tag{5.11}$$

Those who have had power series in calculus may recall the power series representation of e^x, namely

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \sum_{i=0}^{\infty}\frac{x^i}{i!}.$$

Thus the expression in Equation 5.11 is the approximation to e^{-1} we get by substituting -1 for x in the power series and stopping the series at $i = n$. Note that the result depends very "lightly" on n; so long as we have at least four or five people, then, no matter how many people we have, the probability that no one gets their hat back remains at roughly e^{-1}. Our intuition might have suggested that as the number of students increases, the probability that *someone* gets his or her own backpack back approaches 1 rather than $1 - e^{-1}$. Here is another example of why it is important to use computations with the rules of probability instead of intuition!

The principle of inclusion and exclusion for counting

Exercise 5.2-6 How many functions are there from an n-element set N to a k-element set $K = \{y_1, y_2, \ldots y_k\}$ that map nothing to y_1? Another way to say this is if I have n distinct candy bars and k children Sam, Mary, Pat, etc., in how ways may I pass out the candy bars so that Sam doesn't get any candy (and maybe some other children don't either)?

Exercise 5.2-7 How many functions map nothing to a j-element subset J of K? Another way to say this is if I have n distinct candy bars and k children Sam, Mary, Pat, etc., in how ways may I pass out the candy bars so that some particular j-element subset of the children don't get any (and maybe some other children don't either)?

Exercise 5.2-8 What is the number of functions from an n-element set N to a k element set K that map nothing to at least one element of K? Another way to say this is if I have n distinct candy bars and k children Sam, Mary, Pat, etc., in how ways may I pass out the candy bars so that some child doesn't get any (and maybe some other children don't either)?

Exercise 5.2-9 On the basis of the previous exercises, how many functions are there from an n-element set onto a k element set?

The number of functions from an n-element set to a k-element set $K = \{y_1, y_2, \ldots y_k\}$ that map nothing to y_1 is simply $(k-1)^n$ because we have $k-1$ choices of where to map each of our n elements. Similarly the number of functions that map nothing to a particular set J of j elements will be $(k-j)^n$. This warms us up for Exercise 5.2-8.

In Exercise 5.2-8 we need an analog of the principle of inclusion and exclusion for the size of a union of k sets (set i being the set of functions that map nothing to element i of the set K). Because we can make the same argument about the size of the union of two or three sets that we made about probabilities of unions of two or three sets, we have a very natural analog. That analog is the *Principle of Inclusion and Exclusion for Counting*

$$\left|\bigcup_{i=1}^{n} E_i\right| = \sum_{k=1}^{n}(-1)^{k+1} \sum_{\substack{i_1,i_2,\ldots,i_k:\\ 1\le i_1<i_2<\cdots<i_k\le n}} |E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k}|\,. \tag{5.12}$$

In fact, this formula is proved by induction or a counting argument in virtually the same way. Applying this formula to the number of functions from N that map nothing to at least one element of K gives us

$$|\bigcup_{i=1}^{k} E_i| = \sum_{k=1}^{n}(-1)^{k+1} \sum_{\substack{i_1,i_2,\ldots,i_k:\\ 1\le i_1<i_2<\cdots<i_k\le n}} |E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k}| = \sum_{j=1}^{k}(-1)^{j-1}\binom{k}{j}(k-j)^n.$$

This is the number of functions from N that map nothing to at least one element of K. The total number of functions from N to K is k^n. Thus the number of onto functions is

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

where the second equality results because $\binom{k}{0}$ is 1 and $(k-0)^n$ is k^n.

Important Concepts, Formulas, and Theorems

1. *Venn Diagram.* To draw a *Venn diagram*, for two or three sets, we draw a rectangle that represents the sample space, and two or three mutually overlapping circles to represent the events.

2. *Probability of a Union of Two Events.* $P(E \cup F) = P(E) + P(F) - P(E \cap F)$

3. *Probability of a Union of Three Events.* $P(E \cup F \cup G) = P(E) + P(F) + P(G) - P(E \cap F) - P(E \cap G) - P(F \cap G) + P(E \cap F \cap G)$.

4. *A Summation Notation.* $\displaystyle\sum_{\substack{i_1,i_2,\ldots,i_k:\\ 1\le i_1<i_2<\cdots<i_k\le n}} f(i_1, i_2, \ldots, i_k)$ is the sum of $f(i_1, i_2, \ldots, i_k)$ over all increasing sequences of k numbers between 1 and n.

5. *Principle of Inclusion and Exclusion for Probability.* The probability of the union $E_1 \cup E_2 \cup \cdots \cup E_n$ of events in a sample space S is given by

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{k=1}^{n}(-1)^{k+1} \sum_{\substack{i_1,i_2,\ldots,i_k:\\ 1\le i_1<i_2<\cdots<i_k\le n}} P\left(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k}\right).$$

6. *Hatcheck Problem.* The *hatcheck problem* or *derangement problem* asks for the probability that a bijection of an n element set maps no element to itself. The answer is

$$\sum_{i=2}^{n}(-1)^i\frac{1}{i!} = \frac{1}{2} - \frac{1}{3!} + \cdots + \frac{(-1)^n}{n!},$$

the result of truncating the power series expansion of e^{-1} at the $\frac{(-1)^n}{n!}$. Thus the result is very close to $\frac{1}{e}$, even for relatively small values of n.

7. *Principle of Inclusion and Exclusion for Counting.* The *Principle of inclusion and exclusion for counting* says that

$$\left|\bigcup_{i=1}^{n} E_i\right| = \sum_{k=1}^{n}(-1)^{k+1}\sum_{\substack{i_1,i_2,\ldots,i_k:\\ 1\le i_1<i_2<\cdots<i_k\le n}} \left|E_{i_1}\cap E_{i_2}\cap\cdots\cap E_{i_k}\right|.$$

Problems

1. Compute the probability that in three flips of a coin the coin comes heads on the first flip or on the last flip.

2. The eight kings and queens are removed from a deck of cards and then two of these cards are selected. What is the probability that the king or queen of spades is among the cards selected?

3. Two dice are rolled. What is the probability that we see a die with six dots on top?

4. A bowl contains two red, two white and two blue balls. We remove two balls. What is the probability that at least one is red or white? Compute the probability that at least one is red.

5. From an ordinary deck of cards, we remove one card. What is the probability that it is an Ace, is a diamond, or is black?

6. Give a formula for the probability of $P(E\cup F\cup G\cup H)$ in terms of the probabilities of E,F, G, and H, and their intersections.

7. What is

$$\sum_{\substack{i_1,i_2,i_3:\\ 1\le i_1<i_2<i_3\le 4}} i_1 i_2 i_3 \ ?$$

8. What is

$$\sum_{\substack{i_1,i_2,i_3:\\ 1\le i_1<i_2<i_3\le 5}} i_1 + i_2 + i_3 \ ?$$

9. The boss asks the secretary to stuff n letters into envelopes forgetting to mention that he has been adding notes to the letters and in the process has rearranged the letters but not the envelopes. In how many ways can the letters be stuffed into the envelopes so that nobody gets the letter intended for him or her? What is the probability that nobody gets the letter intended for him or her?

10. If we are hashing n keys into a hash table with k locations, what is the probability that every location gets at least one key?

11. From the formula for the number of onto functions, find a formula for $S(n,k)$ which is defined in Problem 12 of Section 1.4. These numbers are called *Stirling numbers (of the second kind).*

12. If we roll 8 dice, what is the probability that each of the numbers 1 through 6 appear on top at least once? What about with 9 dice?

13. Explain why the number of ways of distributing k identical apples to n children is $\binom{n+k-1}{k}$. In how many ways may you distribute the apples to the children so that Sam gets more than m? In how many ways may you distribute the apples to the children so that no child gets more than m?

14. A group of n married couples sits a round a circular table for a group discussion of marital problems. The counselor assigns each person to a seat at random. What is the probability that no husband and wife are side by side?

15. Suppose we have a collection of m objects and a set P of p "properties," an undefined term, that the objects may or may not have. For each subset S of the set P of all properties, define $N_a(S)$ (a is for "at least") to be the number of objects in the collection that have at least the properties in S. Thus, for example, $N_a(\emptyset) = m$. In a typical application, formulas for $N_a(S)$ for other sets $S \subseteq P$ are not difficult to figure out. Define $N_e(S)$ to be the number of objects in our collection that have exactly the properties in S. Show that

$$N_e(\emptyset) = \sum_{K:K\subseteq P} (-1)^{|K|} N_a(K).$$

    Explain how this formula could be used for computing the number of onto functions in a more direct way than we did it using unions of sets. How would this formula apply to Problem 9 in this section?

16. In Problem 14 of this section we allow two people of the same sex to sit side by side. If we require in addition to the condition that no husband and wife are side by side the condition that no two people of the same sex are side by side, we obtain a famous problem known as the *mènage* problem. Solve this problem.

17. In how many ways may we place n distinct books on j shelves so that shelf one gets at least m books? (See Problem 7 in Section 1.4.) In how many ways may we place n distinct books on j shelves so that no shelf gets more than m books?

18. In Problem 15 in this section, what is the probability that an object has none of the properties, assuming all objects to be equally likely? How would this apply Problem 10 in this section?

5.3 Conditional Probability and Independence

Conditional Probability

Two cubical dice each have a triangle painted on one side, a circle painted on two sides and a square painted on three sides. Applying the principal of inclusion and exclusion, we can compute that the probability that we see a circle on at least one top when we roll them is $1/3 + 1/3 - 1/9 = 5/9$. We are experimenting to see if reality agrees with our computation. We throw the dice onto the floor and they bounce a few times before landing in the next room.

Exercise 5.3-1 Our friend in the next room tells us both top sides are the same. Now what is the probability that our friend sees a circle on at least one top?

Intuitively, it may seem as if the chance of getting circles ought to be four times the chance of getting triangles, and the chance of getting squares ought to be nine times as much as the chance of getting triangles. We could turn this into the algebraic statements that $P(\text{circles}) = 4P(\text{triangles})$ and $P(\text{squares}) = 9P(\text{triangles})$. These two equations and the fact that the probabilities sum to 1 would give us enough equations to conclude that the probability that our friend saw two circles is now 2/7. But does this analysis make sense? To convince ourselves, let us start with a sample space for the original experiment and see what natural assumptions about probability we can make to determine the new probabilities. In the process, we will be able to replace intuitive calculations with a formula we can use in similar situations. This is a good thing, because we have already seen situations where our intuitive idea of probability might not always agree with what the rules of probability give us.

Let us take as our sample space for this experiment the ordered pairs shown in Table 5.2 along with their probabilities.

Table 5.2: Rolling two unusual dice

TT	TC	TS	CT	CC	CS	ST	SC	SS
$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{9}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{4}$

We know that the event {TT, CC, SS} happened. Thus we would say while it used to have probability

$$\frac{1}{36} + \frac{1}{9} + \frac{1}{4} = \frac{14}{36} = \frac{7}{18} \tag{5.13}$$

this event now has probability 1. Given that, what probability would we now assign to the event of seeing a circle? Notice that the event of seeing a circle now has become the event CC. Should we expect CC to become more or less likely in comparison than TT or SS just because we know now that one of these three outcomes has occurred? Nothing has happened to make us expect that, so whatever new probabilities we assign to these two events, they should have the same ratios as the old probabilities.

Multiplying all three old probabilities by $\frac{18}{7}$ to get our new probabilities will preserve the ratios and make the three new probabilities add to 1. (Is there any other way to get the three new probabilities to add to one and make the new ratios the same as the old ones?) This gives

us that the probability of two circles is $\frac{1}{9} \cdot \frac{18}{7} = \frac{2}{7}$. Notice that nothing we have learned about probability so far told us what to do; we just made a decision based on common sense. When faced with similar situations in the future, it would make sense to use our common sense in the same way. However, do we really need to go through the process of constructing a new sample space and reasoning about its probabilities again? Fortunately, our entire reasoning process can be captured in a formula. We wanted the probability of an event E given that the event F happened. We figured out what the event $E \cap F$ was, and then multiplied its probability by $1/P(F)$. We summarize this process in a definition.

We define the *conditional probability* of E given F, denoted by $P(E|F)$ and read as "the probability of E given F" by

$$P(E|F) = \frac{P(E \cap F)}{P(F)}. \tag{5.14}$$

Then whenever we want the probability of E knowing that F has happened, we compute $P(E|F)$. (If $P(F) = 0$, then we cannot divide by $P(F)$, but F gives us no new information about our situation. For example if the student in the next room says "A pentagon is on top," we have no information except that the student isn't looking at the dice we rolled! Thus we have no reason to change our sample space or the probability weights of its elements, so we define $P(E|F) = P(E)$ when $P(F) = 0$.)

Notice that we did not prove that the probability of E given F is what we said it is; we simply defined it in this way. That is because in the process of making the derivation we made an additional assumption that the relative probabilities of the outcomes in the event F don't change when F happens. This assumption led us to Equation 5.14. Then we chose that equation as our definition of the new concept of the conditional probability of E given F.[4]

In the example above, we can let E be the event that there is more than one circle and F be the event that both dice are the same. Then $E \cap F$ is the event that both dice are circles, and $P(E \cap F)$ is, from the table above, $\frac{1}{9}$. $P(F)$ is, from Equation 5.13, $\frac{7}{18}$. Dividing, we get the probability of $P(E|F)$, which is $\frac{1}{9}/\frac{7}{18} = \frac{2}{7}$.

Exercise 5.3-2 When we roll two ordinary dice, what is the probability that the sum of the tops comes out even, given that the sum is greater than or equal to 10? Use the definition of conditional probability in solving the problem.

Exercise 5.3-3 We say E is *independent* of F if $P(E|F) = P(E)$. Show that when we roll two dice, one red and one green, the event "The total number of dots on top is odd" is independent of the event "The red die has an odd number of dots on top."

Exercise 5.3-4 Sometimes information about conditional probabilities is given to us indirectly in the statement of a problem, and we have to derive information about other probabilities or conditional probabilities. Here is such an example. If a student knows 80% of the material in a course, what do you expect her grade to be on a (well-balanced) 100 question short-answer test about the course? What is the probability that she answers a question correctly on a 100 question true-false test if she guesses at each question she does not know the answer to? (We assume that she knows what

[4]For those who like to think in terms of axioms of probability, we could give an axiomatic definition of conditional probability, and one of our axioms might be that for events E_1 and E_2 that are subsets of F, the ratio of the conditional probabilities $P(E_1|F)$ and $P(E_2|F)$ is the same as the ratio of $P(E_1)$ and $P(E_2)$.

she knows, that is, if she thinks that she knows the answer, then she really does.) What do you expect her grade to be on a 100 question True-False test to be?

For Exercise 5.3-2 let's let E be the event that the sum is even and F be the event that the sum is greater than or equal to 10. Thus referring to our sample space in Exercise 5.3-2, $P(F) = 1/6$ and $P(E \cap F) = 1/9$, since it is the probability that the roll is either 10 or 12. Dividing these two we get 2/3.

In Exercise 5.3-3, the event that the total number of dots is odd has probability 1/2. Similarly, given that the red die has an odd number of dots, the probability of an odd sum is 1/2 since this event corresponds exactly to getting an even roll on the green die. Thus, by the definition of independence, the event of an odd number of dots on the red die and the event that the total number of dots is odd are independent.

In Exercise 5.3-4, if a student knows 80% of the material in a course, we would hope that her grade on a well-designed test of the course would be around 80%. But what if the test is a True-False test? Let R be the event that she gets the right answer, K be the event that she knows that right answer and $\overline{K}$ be the event that she guesses. Then $R = P(R \cap K) + P(R \cap \overline{K})$. Since R is a union of two disjoint events, its probability would be the sum of the probabilities of these two events. How do we get the probabilities of these two events? The statement of the problem gives us implicitly the conditional probability that she gets the right answer given that she knows the answer, namely one, and the probability that she gets the right answer if she doesn't know the answer, namely 1/2. Using Equation 5.14, we see that we use the equation

$$P(E \cap F) = P(E|F)P(F) \tag{5.15}$$

to compute $P(R \cap K)$ and $P(R \cap \overline{K})$, since the problem tells us directly that $P(K) = .8$ and $P(\overline{K}) = .2$. In symbols,

$$\begin{aligned} P(R) &= P(R \cap K) + P(R \cap \overline{K}) \\ &= P(R|K)P(K) + P(R|\overline{K})P(\overline{K}) \\ &= 1 \cdot .8 + .5 \cdot .2 = .9. \end{aligned}$$

We have shown that the probability that she gets the right answer is .9. Thus we would expect her to get a grade of 90%.

Independence

We said in Exercise 5.3-3 that E is independent of F if $P(E|F) = P(E)$. The *product principle for independent probabilities* (Theorem 5.4) gives another test for independence.

Theorem 5.4 *Suppose E and F are events in a sample space. Then E is independent of F if and only if $P(E \cap F) = P(E)P(F)$.*

Proof: First consider the case when F is non-empty. Then, from our definition in Exercise 5.3-3

$$E \text{ is independent of } F \quad \Leftrightarrow \quad P(E|F) = P(E).$$

(Even though the definition only has an "if", recall the convention of using "if" in definitions, even though "if and only if" is meant.) Using the definition of $P(E|F)$ in Equation 5.14, in the right side of the above equation we get

$$\begin{aligned} & P(E|F) = P(E) \\ \Leftrightarrow \quad & \frac{P(E \cap F)}{P(F)} = P(E) \\ \Leftrightarrow \quad & P(E \cap F) = P(E)P(F). \end{aligned}$$

Since every step in this proof was an if and only if statement we have completed the proof for the case when F is non-empty.

If F is empty, then E is independent of F and both $P(E)P(F)$ and $P(E \cap F)$ are zero. Thus in this case as well, E is independent of F if and only if $P(E \cap F) = P(E)P(F)$. ■

Corollary 5.5 *E is independent of F if and only if F is independent of E.*

When we flip a coin twice, we think of the second outcome as being independent of the first. It would be a sorry state of affairs if our definition of independence did not capture this intuitive idea! Let's compute the relevant probabilities to see if it does. For flipping a coin twice our sample space is $\{HH, HT, TH, TT\}$ and we weight each of these outcomes 1/4. To say the second outcome is independent of the first, we must mean that getting an H second is independent of whether we get an H or a T first, and same for getting a T second. This gives us that $P(H \text{ first}) = 1/4+1/4 = 1/2$ and $P(H \text{ second}) = 1/2$, while $P(\text{H first and H second}) = 1/4$. Note that

$$P(H \text{ first})P(H \text{ second}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = P(\text{H first and H second}).$$

By Theorem 5.4, this means that the event "H second" is independent of the event "H first." We can make a similar computation for each possible combination of outcomes for the first and second flip, and so we see that our definition of independence captures our intuitive idea of independence in this case. Clearly the same sort of computation applies to rolling dice as well.

Exercise 5.3-5 What sample space and probabilities have we been using when discussing hashing? Using these, show that the event "key i hashes to position p" and the event "key j hashes to position q" are independent when $i \neq j$. Are they independent if $i = j$?

In Exercise 5.3-5 if we have a list of n keys to hash into a table of size k, our sample space consists of all n-tuples of numbers between 1 and k. The event that key i hashes to some number p consists of all n-tuples with p in the ith position, so its probability is $\left(\frac{1}{k}\right)^{n-1} / \left(\frac{1}{k}\right)^{n} = \frac{1}{k}$. The probability that key j hashes to some number q is also $\frac{1}{k}$. If $i \neq j$, then the event that key i hashes to p and key j hashes to q has probability $\left(\frac{1}{k}\right)^{n-2} / \left(\frac{1}{k}\right)^{n} = \left(\frac{1}{k}\right)^{2}$, which is the product of the probabilities that key i hashes to p and key j hashes to q, so these two events are independent. However if $i = j$ the probability of key i hashing to p and key j hashing to q is zero unless $p = q$, in which case it is 1. Thus if $i = j$, these events are not independent.

Independent Trials Processes

Coin flipping and hashing are examples of what are called "independent trials processes." Suppose we have a process that occurs in stages. (For example, we might flip a coin n times.) Let us use x_i to denote the outcome at stage i. (For flipping a coin n times, $x_i = H$ means that the outcome of the ith flip is a head.) We let S_i stand for the set of possible outcomes of stage i. (Thus if we flip a coin n times, $S_i = \{H, T\}$.) A process that occurs in stages is called an **independent trials process** if for each sequence $a_1, a_2, \ldots, a_n$ with $a_i \in S_i$,

$$P(x_i = a_i | x_1 = a_1, \ldots, x_{i-1} = a_{i-1}) = P(x_i = a_i).$$

In other words, if we let E_i be the event that $x_i = a_i$, then

$$P(E_i | E_1 \cap E_2 \cap \cdots \cap E_{i-1}) = P(E_i).$$

By our product principle for independent probabilities, this implies that

$$P(E_1 \cap E_2 \cap \cdots E_{i-1} \cap E_i) = P(E_1 \cap E_2 \cap \cdots E_{i-1}) P(E_i). \tag{5.16}$$

Theorem 5.6 *In an independent trials process the probability of a sequence $a_1, a_2, \ldots, a_n$ of outcomes is $P(\{a_1\})P(\{a_2\}) \cdots P(\{a_n\})$.*

Proof: We apply mathematical induction and Equation 5.16. ■

How do independent trials relate to coin flipping? Here our sample space consists of sequences of n Hs and Ts, and the event that we have an H (or T) on the ith flip is independent of the event that we have an H (or T) on each of the first $i-1$ flips. In particular, the probability of an H on the ith flip is $2^{n-1}/2^n = .5$, and the probability of an H on the ith flip, given a particular sequence on the first $i-1$ flips is $2^{n-i-1}/2^{n-i} = .5$.

How do independent trials relate to hashing a list of keys? As in Exercise 5.3-5 if we have a list of n keys to hash into a table of size k, our sample space consists of all n-tuples of numbers between 1 and k. The probability that key i hashes to p and keys 1 through $i-1$ hash to q_1, q_2,... q_{i-1} is $\left(\frac{1}{k}\right)^{n-i} / \left(\frac{1}{k}\right)^n$ and the probability that keys 1 through $i-1$ hash to q_1, q_2,... q_{i-1} is $\left(\frac{1}{k}\right)^{n-i+1} / \left(\frac{1}{k}\right)^n$. Therefore

$$P(\text{key } i \text{ hashes to } p | \text{keys 1 through } i-1 \text{ hash to } q_1, q_2, \ldots q_{i-1}) = \frac{\left(\frac{1}{k}\right)^{n-i} / \left(\frac{1}{k}\right)^n}{\left(\frac{1}{k}\right)^{n-i+1} / \left(\frac{1}{k}\right)^n} = \frac{1}{k}.$$

Therefore, the event that key i hashes to some number p is independent of the event that the first $i-1$ keys hash to some numbers q_1, q_2,... q_{i-1}. Thus our model of hashing is an independent trials process.

Exercise 5.3-6 Suppose we draw a card from a standard deck of 52 cards, replace it, draw another card, and continue for a total of ten draws. Is this an independent trials process?

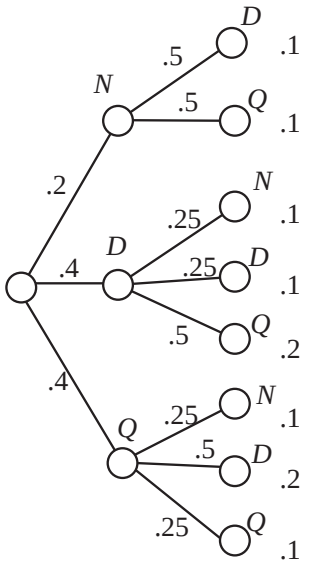
Exercise 5.3-7 Suppose we draw a card from a standard deck of 52 cards, discard it (i.e. we do not replace it), draw another card and continue for a total of ten draws. Is this an independent trials process?

In Exercise 5.3-6 we have an independent trials process, because the probability that we draw a given card at one stage does not depend on what cards we have drawn in earlier stages. However, in Exercise 5.3-7, we don't have an independent trials process. In the first draw, we have 52 cards to draw from, while in the second draw we have 51. In particular, we do not have the same cards to draw from on the second draw as the first, so the probabilities for each possible outcome on the second draw depend on whether that outcome was the result of the first draw.

Tree diagrams

When we have a sample space that consists of sequences of outcomes, it is often helpful to visualize the outcomes by a tree diagram. We will explain what we mean by giving a tree diagram of the following experiment. We have one nickel, two dimes, and two quarters in a cup. We draw a first and second coin. In Figure 5.3 you see our diagram for this process. Notice that in probability theory it is standard to have trees open to the right, rather than opening up or down.

Figure 5.5: A tree diagram illustrating a two-stage process.



Each level of the tree corresponds to one stage of the process of generating a sequence in our sample space. Each vertex is labeled by one of the possible outcomes at the stage it represents. Each edge is labeled with a conditional probability, the probability of getting the outcome at its right end given the sequence of outcomes that have occurred so far. Since no outcomes have occurred at stage 0, we label the edges from the root to the first stage vertices with the probabilities of the outcomes at the first stage. Each path from the root to the far right of the tree represents a possible sequence of outcomes of our process. We label each leaf node with the probability of the sequence that corresponds to the path from the root to that node. By the definition of conditional probabilities, the probability of a path is the product of the probabilities along its edges. We draw a probability tree for any (finite) sequence of successive trials in this way.

Sometimes a probability tree provides a very effective way of answering questions about a

process. For example, what is the probability of having a nickel in our coin experiment? We see there are four paths containing an N, and the sum of their weights is .4, so the probability that one of our two coins is a nickel is .4.

Exercise 5.3-8 How can we recognize from a probability tree whether it is the probability tree of an independent trials process?

Exercise 5.3-9 In Exercise 5.3-4 we asked (among other things), if a student knows 80% of the material in a course, what is the probability that she answers a question correctly on a 100 question True-False test (assuming that she guesses on any question she does not know the answer to)? (We assume that she knows what she knows, that is, if she thinks that she knows the answer, then she really does.) Show how we can use a probability tree to answer this question.

Exercise 5.3-10 A test for a disease that affects 0.1% of the population is 99% effective on people with the disease (that is, it says they have it with probability 0.99). The test gives a false reading (saying that a person who does not have the disease is affected with it) for 2% of the population without the disease. We can think of choosing someone and testing them for the disease as a two stage process. In stage 1, we either choose someone with the disease or we don't. In stage two, the test is either positive or it isn't. Give a probability tree for this process. What is the probability that someone selected at random and given a test for the disease will have a positive test? What is the probability that someone who has positive test results in fact has the disease?

A tree for an independent trials process has the property that at each level, for each node at that level, the (labeled) tree consisting of that node and all its children is identical to each labeled tree consisting of another node at that level and all its children. If we have such a tree, then it automatically satisfies the definition of an independent trials process.

In Exercise 5.3-9, if a student knows 80% of the material in a course, we expect that she has probability .8 of knowing the answer to any given question of a well-designed true-false test. We regard her work on a question as a two stage process; in stage 1 she determines whether she knows the answer, and in stage 2, she either answers correctly with probability 1, or she guesses, in which case she answers correctly with probability 1/2 and incorrectly with probability 1/2. Then as we see in Figure 5.3 there are two root-leaf paths corresponding to her getting a correct answer. One of these paths has probability .8 and the other has probability .1. Thus she actually has probability .9 of getting a right answer if she guesses at each question she does not know the answer to.

In Figure 5.3 we show the tree diagram for thinking of Exercise 5.3-10 as a two stage process. In the first stage, a person either has or doesn't have the disease. In the second stage we administer the test, and its result is either positive or not. We use D to stand for having the disease and ND to stand for not having the disease. We use "pos" to stand for a positive test and "neg" to stand for a negative test, and assume a test is either positive or negative. The question asks us for the conditional probability that someone has the disease, given that they test positive. This is

$$P(D|\text{pos}) = \frac{P(D \cap \text{pos})}{P(\text{pos})}.$$

Figure 5.6: The probability of getting a right answer is .9.

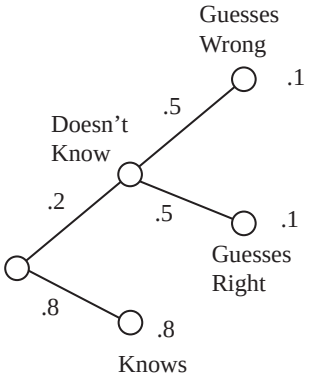
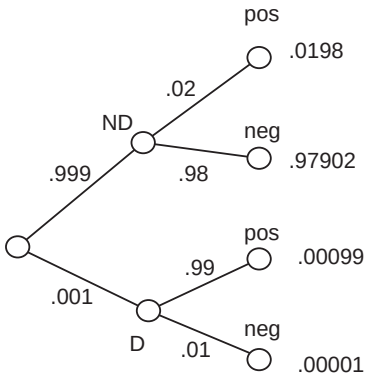


Figure 5.7: A tree diagram illustrating Exercise 5.3-10.



From the tree, we read that $P(D \cap \text{pos}) = .00099$ because this event consists of just one root-leaf path. The event "pos" consists of two root-leaf paths whose probabilities total $.0198 + .00099 = .02097$. Thus $P(D|\text{pos}) = P(D \cap \text{pos})/P(\text{pos}) = .00099/.02097 = .0472$. Thus, given a disease this rare and a test with this error rate, a positive result only gives you roughly a 5% chance of having the disease! Here is another instance where a probability analysis shows something we might not have expected initially. This explains why doctors often don't want to administer a test to someone unless that person is already showing some symptoms of the disease being tested for.

We can also do Exercise 5.3-10 purely algebraically. We are given that

$$P(\text{disease}) = .001, \tag{5.17}$$

$$P(\text{positive test result}|\text{disease}) = .99, \tag{5.18}$$

$$P(\text{positive test result}|\text{no disease}) = .02. \tag{5.19}$$

We wish to compute

$$P(\text{disease}|\text{positive test result}).$$

We use Equation 5.14 to write that

$$P(\text{disease}|\text{positive test result}) = \frac{P(\text{disease} \cap \text{positive test result})}{P(\text{positive test result})}. \tag{5.20}$$

How do we compute the numerator? Using the fact that $P(\text{disease} \cap \text{positive test result}) = P(\text{positive test result} \cap \text{disease})$ and Equation 5.14 again, we can write

$$P(\text{positive test result}|\text{disease}) = \frac{P(\text{positive test result} \cap \text{disease})}{P(\text{disease})}.$$

Plugging Equations 5.18 and 5.17 into this equation, we get

$$.99 = \frac{P(\text{positive test result} \cap \text{disease})}{.001}$$

or $P(\text{positive test result} \cap \text{disease}) = (.001)(.99) = .00099$.

To compute the denominator of Equation 5.20, we observe that since each person either has the disease or doesn't, we can write

$$P(\text{positive test}) = P(\text{positive test} \cap \text{disease}) + P(\text{positive test} \cap \text{no disease}). \tag{5.21}$$

We have already computed $P(\text{positive test result} \cap \text{disease})$, and we can compute the probability $P(\text{positive test result} \cap \text{no disease})$ in a similar manner. Writing

$$P(\text{positive test result}|\text{no disease}) = \frac{P(\text{positive test result} \cap \text{no disease})}{P(\text{no disease})},$$

observing that $P(\text{no disease}) = 1 - P(\text{disease})$ and plugging in the values from Equations 5.17 and 5.19, we get that $P(\text{positive test result} \cap \text{no disease}) = (.02)(1 - .001) = .01998$ We now have the two components of the right hand side of Equation 5.21 and thus $P(\text{positive test result}) = .00099 + .01998 = .02097$. Finally, we have all the pieces in Equation 5.20, and conclude that

$$P(\text{disease}|\text{positive test result}) = \frac{P(\text{disease} \cap \text{positive test result})}{P(\text{positive test result})} = \frac{.00099}{.02097} = .0472.$$

Clearly, using the tree diagram mirrors these computations, but it both simplifies the thought process and reduces the amount we have to write.

Important Concepts, Formulas, and Theorems

1. *Conditional Probability.* We define the *conditional probability* of E given F, denoted by $P(E|F)$ and read as "the probability of E given F" to be

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

2. *Independent.* We say E is *independent* of F if $P(E|F) = P(E)$.

3. *Product Principle for Independent Probabilities.* The *product principle for independent probabilities* (Theorem 5.4) gives another test for independence. Suppose E and F are events in a sample space. Then E is independent of F if and only if $P(E \cap F) = P(E)P(F)$.

4. *Symmetry of Independence.* The event E is independent of the event F if and only if F is independent of E.

5. *Independent Trials Process.* A process that occurs in stages is called an *independent trials process* if for each sequence $a_1, a_2, \ldots, a_n$ with $a_i \in S_i$,

$$P(x_i = a_i | x_1 = a_1, \ldots, x_{i-1} = a_{i-1}) = P(x_i = a_i).$$

6. *Probabilities of Outcomes in Independent Trials.* In an independent trials process the probability of a sequence $a_1, a_2, \ldots, a_n$ of outcomes is $P(\{a_1\})P(\{a_2\}) \cdots P(\{a_n\})$.

7. *Coin Flipping.* Repeatedly flipping a coin is an independent trials process.

8. *Hashing.* Hashing a list of n keys into k slots is an independent trials process with n stages.

9. *Probability Tree.* In a probability tree for a multistage process, each level of the tree corresponds to one stage of the process. Each vertex is labeled by one of the possible outcomes at the stage it represents. Each edge is labeled with a conditional probability, the probability of getting the outcome at its right end given the sequence of outcomes that have occurred so far. Each path from the root to a leaf represents a sequence of outcomes and is labelled with the product of the probabilities along that path. This is the probability of that sequence of outcomes.

Problems

1. In three flips of a coin, what is the probability that two flips in a row are heads, given that there is an even number of heads?

2. In three flips of a coin, is the event that two flips in a row are heads independent of the event that there is an even number of heads?

3. In three flips of a coin, is the event that we have at most one tail independent of the event that not all flips are identical?

4. What is the sample space that we use for rolling two dice, a first one and then a second one? Using this sample space, explain why it is that if we roll two dice, the event "i dots are on top of the first die" and the event "j dots are on top of the second die" are independent.

5. If we flip a coin twice, is the event of having an odd number of heads independent of the event that the first flip comes up heads? Is it independent of the event that the second flip comes up heads? Would you say that the three events are mutually independent? (This hasn't been defined, so the question is one of opinion. However you should back up your opinion with a reason that makes sense!)

6. Assume that on a true-false test, students will answer correctly any question on a subject they know. Assume students guess at answers they do not know. For students who know 60% of the material in a course, what is the probability that they will answer a question correctly? What is the probability that they will know the answer to a question they answer correctly?

7. A nickel, two dimes, and two quarters are in a cup. We draw three coins, one at a time, without replacement. Draw the probability tree which represents the process. Use the tree to determine the probability of getting a nickel on the last draw. Use the tree to determine the probability that the first coin is a quarter, given that the last coin is a quarter.

8. Write down a formula for the probability that a bridge hand (which is 13 cards, chosen from an ordinary deck) has four aces, given that it has one ace. Write down a formula for the probability that a bridge hand has four aces, given that it has the ace of spades. Which of these probabilities is larger?

9. A nickel, two dimes, and three quarters are in a cup. We draw three coins, one at a time without replacement. What is the probability that the first coin is a nickel? What is the probability that the second coin is a nickel? What is the probability that the third coin is a nickel?

10. If a student knows 75% of the material in a course, and a 100 question multiple choice test with five choices per question covers the material in a balanced way, what is the student's probability of getting a right answer to a given question, given that the student guesses at the answer to each question whose answer he or she does not know?

11. Suppose E and F are events with $E \cap F = \emptyset$. Describe when E and F are independent and explain why.

12. What is the probability that in a family consisting of a mother, father and two children of different ages, that the family has two girls, given that one of the children is a girl? What is the probability that the children are both boys, given that the older child is a boy?

5.4 Random Variables

What are Random Variables?

A **random variable** for an experiment with a sample space S is a function that assigns a number to each element of S. Typically instead of using f to stand for such a function we use X (at first, a random variable was conceived of as a variable related to an experiment, explaining the use of X, but it is very helpful in understanding the mathematics to realize it actually is a function on the sample space).

For example, if we consider the process of flipping a coin n times, we have the set of all sequences of n Hs and Ts as our sample space. The "number of heads" random variable takes a sequence and tells us how many heads are in that sequence. Somebody might say "Let X be the number of heads in 5 flips of a coin." In that case $X(HTHHT) = 3$ while $X(THTHT) = 2$. It may be rather jarring to see X used to stand for a function, but it is the notation most people use.

For a sequence of hashes of n keys into a table with k locations, we might have a random variable X_i which is the number of keys that are hashed to location i of the table, or a random variable X that counts the number of collisions (hashes to a location that already has at least one key). For an n question test on which each answer is either right or wrong (a short answer, True-False or multiple choice test for example) we could have a random variable that gives the number of right answers in a particular sequence of answers to the test. For a meal at a restaurant we might have a random variable that gives the price of any particular sequence of choices of menu items.

Exercise 5.4-1 Give several random variables that might be of interest to a doctor whose sample space is her patients.

Exercise 5.4-2 If you flip a coin six times, how many heads do you expect?

A doctor might be interested in patients' ages, weights, temperatures, blood pressures, cholesterol levels, etc.

For Exercise 5.4-2, in six flips of a coin, it is natural to expect three heads. We might argue that if we average the number of heads over all possible outcomes, the average should be half the number of flips. Since the probability of any given sequence equals that of any other, it is reasonable to say that this average is what we expect. Thus we would say we expect the number of heads to be half the number of flips. We will explore this more formally later.

Binomial Probabilities

When we study an independent trials process with two outcomes at each stage, it is traditional to refer to those outcomes as successes and failures. When we are flipping a coin, we are often interested in the number of heads. When we are analyzing student performance on a test, we are interested in the number of correct answers. When we are analyzing the outcomes in drug trials, we are interested in the number of trials where the drug was successful in treating the disease. This suggests a natural random variable associated with an independent trials process with two outcomes at each stage, namely the number of successes in n trials. We will analyze in general

the probability of exactly k successes in n independent trials with probability p of success (and thus probability $1-p$ of failure) on each trial. It is standard to call such an independent trials process a *Bernoulli trials process.*

Exercise 5.4-3 Suppose we have 5 Bernoulli trials with probability p of success on each trial. What is the probability of success on the first three trials and failure on the last two? Failure on the first two trials and success on the last three? Success on trials 1, 3, and 5, and failure on the other two? Success on any particular three trials, and failure on the other two?

Since the probability of a sequence of outcomes is the product of the probabilities of the individual outcomes, the probability of any sequence of 3 successes and 2 failures is $p^3(1-p)^2$. More generally, in n Bernoulli trials, the probability of a given sequence of k successes and $n-k$ failures is $p^k(1-p)^{n-k}$. However this is not the probability of having k successes, because many different sequences could have k successes.

How many sequences of n successes and failures have exactly k successes? The number of ways to choose the k places out of n where the successes occur is $\binom{n}{k}$, so the number of sequences with k successes is $\binom{n}{k}$. This paragraph and the last together give us Theorem 5.7.

Theorem 5.7 *The probability of having exactly k successes in a sequence of n independent trials with two outcomes and probability p of success on each trial is*

$$P(\textit{exactly } k \textit{ successes}) = \binom{n}{k} p^k (1-p)^{n-k}$$

.

Proof: The proof follows from the two paragraphs preceding the theorem.■

Because of the connection between these probabilities and the binomial coefficients, the probabilities of Theorem 5.7 are called **binomial probabilities**, or the **binomial probability distribution**.

Exercise 5.4-4 A student takes a ten question objective test. Suppose that a student who knows 80% of the course material has probability .8 of success an any question, independently of how the student did on any other problem. What is the probability that this student earns a grade of 80 or better (out of 100)? What grade would you expect the student to get?

Exercise 5.4-5 Recall the primality testing algorithm from Section 2.4. Here we said that we could, by choosing a random number less than or equal to n, perform a test on n that, if n was not prime, would certify this fact with probability 1/2. Suppose we perform 20 of these tests. It is reasonable to assume that each of these tests is independent of the rest of them. What is the probability that a non-prime number is certified to be non-prime?

Since a grade of 80 or better on a ten question test corresponds to 8, 9, or 10 successes in ten trials, in Exercise 5.4-4 we have

$$P(\text{80 or better}) = \binom{10}{8}(.8)^8(.2)^2 + \binom{10}{9}(.8)^9(.2)^1 + (.8)^{10}.$$

Some work with a calculator gives us that this sum is approximately .678. The grade we would expect the student to get is 80.

In Exercise 5.4-5, we will first compute the probability that a non-prime number is not certified to be non-prime. If we think of success as when the number is certified non-prime and failure when it isn't, then we see that the only way to fail to certify a number is to have 20 failures. Using our formula we see that the probability that a non-prime number is not certified non-prime is just $\binom{20}{20}(.5)^{20} = 1/1048576$. Thus the chance of this happening is less than one in a million, and the chance of certifying the non-prime as non-prime is 1 minus this. Therefore the probability that a non-prime number will be certified non-prime is 1048575/1048576, which is more than .999999, so a non-prime number is almost sure to be certified non-prime.

A Taste of Generating Functions We note a nice connection between the probability of having exactly k successes and the binomial theorem. Consider, as an example, the polynomial $(H+T)^3$. Using the binomial theorem, we get that this is

$$(H+T)^3 = \binom{3}{0}H^3 + \binom{3}{1}H^2T + \binom{3}{2}HT^2 + \binom{3}{3}T^3.$$

We can interpret this as telling us that if we flip a coin three times, with outcomes heads or tails each time, then there are $\binom{3}{0} = 1$ way of getting 3 heads, $\binom{3}{2} = 3$ ways of getting two heads and one tail, $\binom{3}{1} = 3$ ways of getting one head and two tails and $\binom{3}{3} = 1$ way of getting 3 tails.

Similarly, if we replace H and T by px and $(1-p)y$ we would get the following:

$$(px+(1-p)y)^3 = \binom{3}{0}p^3x^3 + \binom{3}{1}p^2(1-p)x^2y + \binom{3}{2}p(1-p)^2xy^2 + \binom{3}{3}(1-p)^3y^3.$$

Generalizing this to n repeated trials where in each trial the probability of success is p, we see that by taking $(px+(1-p)y)^n$ we get

$$(px+(1-p)y)^n = \sum_{k=0}^{k}\binom{n}{k}p^k(1-p)^{n-k}x^ky^{n-k}.$$

Taking the coefficient of x^ky^{n-k} from this sum, we get exactly the result of Theorem 5.7. This connection is a simple case of a very powerful tool known as *generating functions*. We say that the polynomial $(px+(1-p)y)^n$ *generates* the binomial probabilities. In fact, we don't even need the y, because

$$(px+1-p)^n = \sum_{i=0}^{n}\binom{n}{i}p^i(1-p)^{n-i}x^i.$$

In general, the *generating function* for the sequence $a_0, a_1, a_2, \ldots, a_n$ is $\sum_{i=1}^{n} a_ix^i$, and the *generating function* for an infinite sequence $a_0, a_1, a_2, \ldots, a_n, \ldots$ is the infinite series $\sum_{i=1}^{\infty} a_ix^i$.

Expected Value

In Exercise 5.4-4 and Exercise 5.4-2 we asked about the value you would expect a random variable(in these cases, a test score and the number of heads in six flips of a coin) to have. We haven't yet defined what we mean by the value we expect, and yet it seems to make sense in the places we asked about it. If we say we expect 1 head if we flip a coin twice, we can explain our reasoning by taking an average. There are four outcomes, one with no heads, two with one head, and one with two heads, giving us an average of

$$\frac{0+1+1+2}{4} = 1.$$

Notice that using averages compels us to have some expected values that are impossible to achieve. For example in three flips of a coin the eight possibilities for the number of heads are 0, 1, 1, 1, 2, 2, 2, 3, giving us for our average

$$\frac{0+1+1+1+2+2+2+3}{8} = 1.5.$$

Exercise 5.4-6 An interpretation in games and gambling makes it clear that it makes sense to expect a random variable to have a value that is not one of the possible outcomes. Suppose that I proposed the following game. You pay me some money, and then you flip three coins. I will pay you one dollar for every head that comes up. Would you play this game if you had to pay me \$2.00? How about if you had to pay me \$1? How much do you think it should cost, in order for this game to be fair?

Since you expect to get 1.5 heads, you expect to make \$1.50. Therefore, it is reasonable to play this game as long as the cost is at most \$1.50.

Certainly averaging our variable over all elements of our sample space by adding up one result for each element of the sample space as we have done above is impractical even when we are talking about something as simple as ten flips of a coin. However we can ask how many times each possible number of heads arises, and then multiply the number of heads by the number of times it arises to get an average number of heads of

$$\frac{0\binom{10}{0} + 1\binom{10}{1} + 2\binom{10}{2} + \cdots + 9\binom{10}{9} + 10\binom{10}{10}}{1024} = \frac{\sum_{i=0}^{10} i\binom{10}{i}}{1024}. \tag{5.22}$$

Thus we wonder whether we have seen a formula for $\sum_{i=0}^{n} i\binom{n}{i}$. Perhaps we have, but in any case the binomial theorem and a bit of calculus or a proof by induction show that

$$\sum_{i=0}^{n} i\binom{n}{i} = 2^{n-1}n,$$

giving us $512 \cdot 10/1024 = 5$ for the fraction in Equation 5.22. If you are asking "Does it have to be that hard?" then good for you. Once we know a bit about the theory of expected values of random variables, computations like this will be replaced by far simpler ones.

Besides the nasty computations that a simple question lead us to, the average value of a random variable on a sample space need not have anything to do with the result we expect. For instance if we replace heads and tails with right and wrong, we get the sample space of possible

results that a student will get when taking a ten question test with probability .9 of getting the right answer on any one question. Thus if we compute the average number of right answers in all the possible patterns of test results we get an average of 5 right answers. This is not the number of right answers we expect because averaging has nothing to do with the underlying process that gave us our probability! If we analyze the ten coin flips a bit more carefully, we can resolve this disconnection. We can rewrite Equation 5.22 as

$$0\frac{\binom{10}{0}}{1024} + 1\frac{\binom{10}{1}}{1024} + 2\frac{\binom{10}{2}}{1024} + \cdots + 9\frac{\binom{10}{9}}{1024} + 10\frac{\binom{10}{10}}{1024} = \sum_{i=0}^{10} i\frac{\binom{10}{i}}{1024}. \tag{5.23}$$

In Equation 5.23 we see we can compute the average number of heads by multiplying each value of our "number of heads" random variable by the probability that we have that value for our random variable, and then adding the results. This gives us a "weighted average" of the values of our random variable, each value weighted by its probability. Because the idea of weighting a random variable by its probability comes up so much in Probability Theory, there is a special notation that has developed to use this weight in equations. We use $P(X = x_i)$ to stand for the probability that the random variable X equals the value x_i. We call the function that assigns $P(x_i)$ to the event $P(X = x_i)$ the *distribution function* of the random variable X. Thus, for example, the binomial probability distribution is the distribution function for the "number of successes" random variable in Bernoulli trials.

We define the **expected value** or **expectation** of a random variable X whose values are the set $\{x_1, x_2, \ldots x_k\}$ to be

$$E(X) = \sum_{i=1}^{k} x_i P(X = x_i).$$

Then for someone taking a ten-question test with probability .9 of getting the correct answer on each question, the expected number of right answers is

$$\sum_{i=0}^{10} i\binom{10}{i}(.9)^i(.1)^{10-i}.$$

In the end of section exercises we will show a technique (that could be considered an application of generating functions) that allows us to compute this sum directly by using the binomial theorem and calculus. We now proceed to develop a less direct but easier way to compute this and many other expected values.

Exercise 5.4-7 Show that if a random variable X is defined on a sample space S (you may assume X has values x_1, x_2, $\ldots x_k$ as above) then the expected value of X is given by

$$E(X) = \sum_{s:s\in S} X(s)P(s).$$

(In words, we take each member of the sample space, compute its probability, multiply the probability by the value of the random variable and add the results.)

In Exercise 5.4-7 we asked for a proof of a fundamental lemma

Lemma 5.8 *If a random variable X is defined on a (finite) sample space S, then its expected value is given by*

$$E(X) = \sum_{s:s\in S} X(s)P(s).$$

Proof: Assume that the values of the random variable are x_1, x_2, $\ldots x_k$. Let F_i stand for the event that the value of X is x_i, so that $P(F_i) = P(X = x_i)$. Then, in the sum on the right-hand side of the equation in the statement of the lemma, we can take the items in the sample space, group them together into the events F_i and and rework the sum into the definition of expectation, as follows:

$$\begin{aligned}
\sum_{s:s\in S} X(s)P(s) &= \sum_{i=1}^{k} \sum_{s:s\in F_i} X(s)P(s) \\
&= \sum_{i=1}^{k} \sum_{s:s\in F_i} x_i P(s) \\
&= \sum_{i=1}^{k} x_i \sum_{s:s\in F_i} P(s) \\
&= \sum_{i=1}^{k} x_i P(F_i) \\
&= \sum_{i=1}^{k} x_i P(X = x_i) = E(X).
\end{aligned}$$

■

The proof of the lemma need not be so formal and symbolic as what we wrote; in English, it simply says that when we compute the sum in the Lemma, we can group together all elements of the sample space that have X-value x_i and add up their probabilities; this gives us $x_iP(x_i)$, which leads us to the definition of the expected value of X.

Expected Values of Sums and Numerical Multiples

Another important point about expected value follows naturally from what we think about when we use the word "expect" in English. If a paper grader expects to earn ten dollars grading papers today and expects to earn twenty dollars grading papers tomorrow, then she expects to earn thirty dollars grading papers in these two days. We could use X_1 to stand for the amount of money she makes grading papers today and X_2 to stand for the amount of money she makes grading papers tomorrow, so we are saying

$$E(X_1 + X_2) = E(X_1) + E(X_2).$$

This formula holds for any sum of a pair of random variables, and more generally for any sum of random variables on the same sample space.

Theorem 5.9 *Suppose X and Y are random variables on the (finite) sample space S. Then*

$$E(X + Y) = E(X) + E(Y).$$

Proof: From Lemma 5.8 we may write

$$E(X+Y) = \sum_{s:s\in S} (X(s)+Y(s))P(s) = \sum_{s:s\in S} X(s)P(s) + \sum_{s:s\in S} Y(s)P(s) = E(X)+E(Y).$$

■

If we double the credit we give for each question on a test, we would expect students' scores to double. Thus our next theorem should be no surprise. In it we use the notation cX for the random variable we get from X by multiplying all its values by the number c.

Theorem 5.10 *Suppose X is a random variable on a sample space S. Then for any number c, $E(cX) = cE(X)$.*

Proof: Left as a problem.■

Theorems 5.9 and 5.10 are very useful in proving facts about random variables. Taken together, they are typically called *linearity of expectation*. (The idea that the expectation of a sum is the same as the sum of expectations is called the *additivity of expectation*.) The idea of linearity will often allow us to work with expectations much more easily than if we had to work with the underlying probabilities.

For example, on one flip of a coin, our expected number of heads is .5. Suppose we flip a coin n times and let X_i be the number of heads we see on flip i, so that X_i is either 0 or 1. (For example in five flips of a coin, $X_2(HTHHT) = 0$ while $X_3(HTHHT) = 1$.) Then X, the total number of heads in n flips is given by

$$X = X_1 + X_2 + \cdots X_n, \tag{5.24}$$

the sum of the number of heads on the first flip, the number on the second, and so on through the number of heads on the last flip. But the expected value of each X_i is .5. We can take the expectation of both sides of Equation 5.24 and apply Lemma 5.9 repeatedly (or use induction) to get that

$$\begin{aligned} E(X) &= E(X_1 + X_2 + \cdots + X_n) \\ &= E(X_1) + E(X_2) + \cdots + E(X_n) \\ &= .5 + .5 + \cdots + .5 \\ &= .5n \end{aligned}$$

Thus in n flips of a coin, the expected number of heads is $.5n$. Compare the ease of this method with the effort needed earlier to deal with the expected number of heads in ten flips! Dealing with probability .9 or, in general with probability p poses no problem.

Exercise 5.4-8 Use the additivity of expectation to determine the expected number of correct answers a student will get on an n question "fill in the blanks" test if he or she knows 90% of the material in the course and the questions on the test are an accurate and uniform sampling of the material in the course.

In Exercise 5.4-8, since the questions sample the material in the course accurately, the most natural probability for us to assign to the event that the student gets a correct answer on a given

question is .9. We can let X_i be the number of correct answers on question i (that is, either 1 or 0 depending on whether or not the student gets the correct answer). Then the expected number of right answers is the expected value of the sum of the variables X_i. From Theorem 5.9 see that in n trials with probability .9 of success, we expect to have $.9n$ successes. This gives that the expected number of right answers on a ten question test with probability .9 of getting each question right is 9, as we expected. This is a special case of our next theorem, which is proved by the same kind of computation.

Theorem 5.11 *In a Bernoulli trials process, in which each experiment has two outcomes and probability p of success, the expected number of successes is np.*

Proof: Let X_i be the number of successes in the ith of n independent trials. The expected number of successes on the ith trial (i.e. the expected value of X_i) is, by definition,

$$p \cdot 1 + (1 - p) \cdot 0 = p.$$

The number of successes X in all n trials is the sum of the random variables X_i. Then by Theorem 5.9 the expected number of successes in n independent trials is the sum of the expected values of the n random variables X_i and this sum is np.■

The Number of Trials until the First Success

Exercise 5.4-9 How many times do you expect to have to flip a coin until you first see a head? Why? How many times to you expect to have to roll two dice until you see a sum of seven? Why?

Our intuition suggests that we should have to flip a coin twice to see a head. However we could conceivably flip a coin forever without seeing a head, so should we really expect to see a head in two flips? The probability of getting a seven on two dice is 1/6. Does that mean we should expect to have to roll the dice six times before we see a seven?

In order to analyze this kind of question we have to realize that we are stepping out of the realm of independent trials processes on finite sample spaces. We will consider the process of repeating independent trials with probability p of success until we have a success and then stopping. Now the possible outcomes of our multistage process are the infinite set

$$\{S, FS, FFS, \ldots, F^iS, \ldots\},$$

in which we have used the notation F^iS to stand for the sequence of i failures followed by a success. Since we have an infinite sequence of outcomes, it makes sense to think about whether we can assign an infinite sequence of probability weights to its members so that the resulting sequence of probabilities adds to one. If so, then all our definitions make sense, and in fact the proofs of all our theorems remain valid.[5] There is only one way to assign weights that is consistent with our knowledge of (finite) independent trials processes, namely

$$P(S) = p, \quad P(FS) = (1 - p)p, \quad \ldots, \quad P(F^iS) = (1 - p)^ip, \quad \ldots.$$

[5]for those who are familiar with the concept of convergence for infinite sums (i.e. infinite series), it is worth noting that it is the fact that probability weights cannot be negative and must add to one that makes all the sums we need to deal with for all the theorems we have proved so far converge. That doesn't mean all sums we might want to deal with will converge; some random variables defined on the sample space we have described will have infinite expected value. However those we need to deal with for the expected number of trials until success do converge.

Thus we have to hope these weights add to one; in fact their sum is

$$\sum_{i=0}^{\infty}(1-p)^i p = p\sum_{i=0}^{\infty}(1-p)^i = p\frac{1}{1-(1-p)} = \frac{p}{p} = 1.$$

Therefore we have a legitimate assignment of probabilities and the set of sequences

$$\{F, FS, FFS, FFFS, \ldots, F^iS, \ldots\}$$

is a sample space with these probability weights. This probability distribution, $P(F^iS) = (1-p)^i p$, is called a *geometric* distribution because of the geometric series we used in proving the probabilities sum to 1.

Theorem 5.12 *Suppose we have a sequence of trials in which each trial has two outcomes, success and failure, and where at each step the probability of success is p. Then the expected number of trials until the first success is* $1/p$.

Proof:

We consider the random variable X which is i if the first success is on trial i. (In other words, $X(F^{i-1}S)$ is i.) The probability that the first success is on trial i is $(1-p)^{i-1}p$, since in order for this to happen there must be $i-1$ failures followed by 1 success. The expected number of trials is the expected value of X, which is, by the definition of expected value and the previous two sentences,

$$\begin{aligned} E[\text{number of trials}] &= \sum_{i=0}^{\infty} p(1-p)^{i-1} i \\ &= p\sum_{i=0}^{\infty}(1-p)^{i-1} i \\ &= \frac{p}{1-p}\sum_{i=0}^{\infty}(1-p)^i i \\ &= \frac{p}{1-p}\frac{1-p}{p^2} \\ &= \frac{1}{p} \end{aligned}$$

To go from the third to the fourth line we used the fact that

$$\sum_{j=0}^{\infty} j x^j = \frac{x}{(1-x)^2}, \tag{5.25}$$

true for x with absolute value less than one. We proved a finite version of this equation as Theorem 4.6; the infinite version is even easier to prove. ■

Applying this theorem, we see that the expected number of times you need to flip a coin until you get heads is 2, and the expected number of times you need to roll two dice until you get a seven is 6.

Important Concepts, Formulas, and Theorems

1. *Random Variable.* A *random variable* for an experiment with a sample space S is a function that assigns a number to each element of S.

2. Bernoulli Trials Process. An independent trials process with two outcomes, success and failure, at each stage and probability p of success and $1-p$ of failure at each stage is called a *Bernoulli trials process.*

3. Probability of a Sequence of Bernoulli Trials. In n Bernoulli trials with probability p of success, the probability of a given sequence of k successes and $n-k$ failures is $p^k(1-p)^{n-k}$.

4. *The Probability of k Successes in n Bernoulli Trials* The probability of having exactly k successes in a sequence of n independent trials with two outcomes and probability p of success on each trial is

$$P(\text{exactly } k \text{ successes}) = \binom{n}{k} p^k (1-p)^{n-k}$$

.

5. Binomial Probability Distribution. The probabilities of of k successes in n Bernoulli trials, $\binom{n}{k}p^k(1-p)^{n-k}$, are called *binomial probabilities*, or the *binomial probability distribution.*

6. *Generating Function.* The *generating function* for the sequence $a_0, a_1, a_2, \ldots, a_n$ is $\sum_{i=1}^{n} a_i x^i$, and the *generating function* for an infinite sequence $a_0, a_1, a_2, \ldots, a_n, \ldots$ is the infinite series $\sum_{i=1}^{\infty} a_i x^i$. The polynomial $(px + 1 - p)^n$ is the generating function for the binomial probabilities for n Bernoulli trials with probability p of success.

7. *Distribution Function.* We call the function that assigns $P(x_i)$ to the event $P(X = x_i)$ the *distribution function* of the random variable X.

8. *Expected Value.* We define the *expected value* or *expectation* of a random variable X whose values are the set $\{x_1, x_2, \ldots x_k\}$ to be

$$E(X) = \sum_{i=1}^{k} x_i P(X = x_i).$$

9. *Another Formula for Expected Values.* If a random variable X is defined on a (finite) sample space S, then its expected value is given by

$$E(X) = \sum_{s:s\in S} X(s)P(s).$$

10. *Expected Value of a Sum.* Suppose X and Y are random variables on the (finite) sample space S. Then

$$E(X + Y) = E(X) + E(Y).$$

This is called the *additivity of expectation.*

11. *Expected Value of a Numerical Multiple.* Suppose X is a random variable on a sample space S. Then for any number c, $E(cX) = cE(X)$. This result and the additivity of expectation together are called the *linearity of expectation.*

12. *Expected Number of Successes in Bernoulli Trials.* In a Bernoulli trials process, in which each experiment has two outcomes and probability p of success, the expected number of successes is np.

13. *Expected Number of Trials Until Success.* Suppose we have a sequence of trials in which each trial has two outcomes, success and failure, and where at each step the probability of success is p. Then the expected number of trials until the first success is $1/p$.

Problems

1. Give several random variables that might be of interest to someone rolling five dice (as one does, for example, in the game Yatzee).

2. Suppose I offer to play the following game with you if you will pay me some money. You roll a die, and I give you a dollar for each dot that is on top. What is the maximum amount of money a rational person might be willing to pay me in order to play this game?

3. How many sixes do we expect to see on top if we roll 24 dice?

4. What is the expected sum of the tops of n dice when we roll them?

5. In an independent trials process consisting of six trials with probability p of success, what is the probability that the first three trials are successes and the last three are failures? The probability that the last three trials are successes and the first three are failures? The probability that trials 1, 3, and 5 are successes and trials 2, 4, and 6 are failures? What is the probability of three successes and three failures?

6. What is the probability of exactly eight heads in ten flips of a coin? Of eight or more heads?

7. How many times do you expect to have to role a die until you see a six on the top face?

8. Assuming that the process of answering the questions on a five-question quiz is an independent trials process and that a student has a probability of .8 of answering any given question correctly, what is the probability of a sequence of four correct answers and one incorrect answer? What is the probability that a student answers exactly four questions correctly?

9. What is the expected value of the constant random variable X that has $X(s) = c$ for every member s of the sample space? We frequently just use c to stand for this random variable, and thus this question is asking for $E(c)$.

10. Someone is taking a true-false test and guessing when they don't know the answer. We are going to compute a score by subtracting a percentage of the number of incorrect answers from the number of correct answers. When we convert this "corrected score" to a percentage score we want its expected value to be the percentage of the material being tested that the test-taker knows. How can we do this?

11. Do Problem 10 of this section for the case that someone is taking a multiple choice test with five choices for each answer and guesses randomly when they don't know the answer.

12. Suppose we have ten independent trials with three outcomes called good, bad, and indifferent, with probabilities p, q, and r, respectively. What is the probability of three goods, two bads, and five indifferents? In n independent trials with three outcomes A, B, and C, with probabilities p, q, and r, what is the probability of i As, j Bs, and k Cs? (In this problem we assume $p + q + r = 1$ and $i + j + k = n$.)

13. In as many ways as you can, prove that

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

14. Prove Theorem 5.10.

15. Two nickels, two dimes, and two quarters are in a cup. We draw three coins, one after the other, without replacement. What is the expected amount of money we draw on the first draw? On the second draw? What is the expected value of the total amount of money we draw? Does this expected value change if we draw the three coins all together?

16. In this exercise we will evaluate the sum

$$\sum_{i=0}^{10} i\binom{10}{i}(.9)^i(.1)^{10-i}$$

that arose in computing the expected number of right answers a person would have on a ten question test with probability .9 of answering each question correctly. First, use the binomial theorem and calculus to show that

$$10(.1 + x)^9 = \sum_{i=0}^{10} i\binom{10}{i}(.1)^{10-i}x^{i-1}$$

Substituting in $x = .9$ gives us almost the sum we want on the right hand side of the equation, except that in every term of the sum the power on .9 is one too small. Use some simple algebra to fix this and then explain why the expected number of right answers is 9.

17. Give an example of two random variables X and Y such that $E(XY) \neq E(X)E(Y)$. Here XY is the random variable with $(XY)(s) = X(s)Y(s)$.

18. Prove that if X and Y are independent in the sense that the event that $X = x$ and the event that $Y = y$ are independent for each pair of values x of X and y of Y, then $E(XY) = E(X)E(Y)$. See Exercise 5-17 for a definition of XY.

19. Use calculus and the sum of a geometric series to show that

$$\sum_{j=0}^{\infty} jx^j = \frac{x}{(1-x)^2}$$

as in Equation 5.25.

20. Give an example of a random variable on the sample space $\{S, FS, FFS, \ldots, F^iS, \ldots\}$ with an infinite expected value.

5.5 Probability Calculations in Hashing

We can use our knowledge of probability and expected values to analyze a number of interesting aspects of hashing including:

1. expected number of items per location,
2. expected time for a search,
3. expected number of collisions,
4. expected number of empty locations,
5. expected time until all locations have at least one item,
6. expected maximum number of items per location.

Expected Number of Items per Location

Exercise 5.5-1 We are going to compute the expected number of items that hash to any particular location in a hash table. Our model of hashing n items into a table of size k allows us to think of the process as n independent trials, each with k possible outcomes (the k locations in the table). On each trial we hash another key into the table. If we hash n items into a table with k locations, what is the probability that any one item hashes into location 1? Let X_i be the random variable that counts the number of items that hash to location 1 in trial i (so that X_i is either 0 or 1). What is the expected value of X_i? Let X be the random variable $X_1 + X_2 + \cdots + X_n$. What is the expected value of X? What is the expected number of items that hash to location 1? Was the fact that we were talking about location 1 special in any way? That is, does the same expected value apply to every location?

Exercise 5.5-2 Again we are hashing n items into k locations. Our model of hashing is that of Exercise 5.5-1. What is the probability that a location is empty? What is the expected number of empty locations? Suppose we now hash n items into the same number n of locations. What limit does the expected fraction of empty places approach as n gets large?

In Exercise 5.5-1, the probability that any one item hashes into location 1 is $1/k$, because all k locations are equally likely. The expected value of X_i is then $1/k$. The expected value of X is then n/k, the sum of n terms each equal to $1/k$. Of course the same expected value applies to any location. Thus we have proved the following theorem.

Theorem 5.13 *In hashing n items into a hash table of size k, the expected number of items that hash to any one location is n/k.*

Expected Number of Empty Locations

In Exercise 5.5-2 the probability that position i will be empty after we hash 1 item into the table will be $1 - \frac{1}{k}$. (Why?) In fact, we can think of our process as an independent trials process with two outcomes: the key hashes to slot i or it doesn't. From this point of view, it is clear that the probability of nothing hashing to slot i in n trials is $(1 - \frac{1}{k})^n$. Now consider the original sample space again and let X_i be 1 if slot i is empty for a given sequence of hashes or 0 if it is not. Then the number of empty slots for a given sequence of hashes is $X_1 + X_2 + \cdots + X_k$ evaluated at that sequence. Therefore, the expected number of empty slots is, by Theorem 5.9, $k(1 - \frac{1}{k})^n$. Thus we have proved another nice theorem about hashing.

Theorem 5.14 *In hashing n items into a hash table with k locations, the expected number of empty locations is $k(1 - \frac{1}{k})^n$.*

Proof: Given above.■

If we have the same number of slots as places, the expected number of empty slots is $n(1-\frac{1}{n})^n$, so the expected fraction of empty slots is $(1-\frac{1}{n})^n$. What does this fraction approach as n grows? You may recall that $\lim_{n\to\infty}(1+\frac{1}{n})^n$ is e, the base for the natural logarithm. In the problems at the end of the section, we show you how to derive from this that $\lim_{n\to\infty}(1-\frac{1}{n})^n$ is e^{-1}. Thus for a reasonably large hash table, if we hash in as many items as we have slots, we expect a fraction $1/e$ of those slots to remain empty. In other words, we expect n/e empty slots. On the other hand, we expect $\frac{n}{n}$ items per location, which suggests that we should expect each slot to have an item and therefore expect to have no empty locations. Is something wrong? No, but we simply have to accept that our expectations about expectation just don't always hold true. What went wrong in that apparent contradiction is that our definition of expected value doesn't imply that if we have an expectation of one key per location then every location must have a key, but only that empty locations have to be balanced out by locations with more than one key. When we want to make a statement about expected values, we must use either our definitions or theorems to back it up. This is another example of why we have to back up intuition about probability with careful analysis.

Expected Number of Collisions

We say that we have a *collision* when we hash an item to a location that already contains an item. How can we compute the expected number of collisions? The number of collisions will be the number n of keys hashed minus the number of occupied locations because each occupied location will contain one key that will not have collided in the process of being hashed. Thus, by Theorems 5.9 and 5.10,

$$E(\text{collisions}) = n - E(\text{occupied locations}) = n - k + E(\text{empty locations}) \tag{5.26}$$

where the last equality follows because the expected number of occupied locations is k minus the expected number of unoccupied locations. This gives us yet another theorem.

Theorem 5.15 *In hashing n items into a hash table with k locations, the expected number of collisions is $n - k + k(1 - \frac{1}{k})^n$.*

Proof: We have already shown in Theorem 5.14 that the expected number of empty locations is $k(1-\frac{1}{k})^n$. Substituting this into Equation 5.26 gives our formula. ■

Exercise 5.5-3 In real applications, it is often the case that the hash table size is not fixed in advance, since you don't know, in advance, how many items you will insert. The most common heuristic for dealing with this is to start k, the hash table size, at some reasonably small value, and then when n, the number of items, gets to be greater than $2k$, you double the hash table size. In this exercise we propose a different idea. Suppose you waited until every single slot in the hash table had at least one item in it, and then you increased the table size. What is the expected number of items that will be in the table when you increase the size? In other words, how many items do you expect to insert into a hash table in order to ensure that every slot has at least one item? (Hint: Let X_i be the number of items added between the time that there are $i-1$ occupied slots for the first time and the first time that there are i occupied slots.)

For Exercise 5.5-3, the key is to let X_i be the number of items added between the time that there are $i-1$ full slots for the first time and i full slots for the first time. Let's think about this random variable. $E(X_1) = 1$, since after one insertion there is one full slot. In fact X_1 itself is equal to 1.

To compute the expected value of X_2, we note that X_2 can take on any value greater than 1. In fact if we think about it, what we have here (until we actually hash an item to a new slot) is an independent trials process with two outcomes, with success meaning our item hashes to an unused slot. X_2 counts the number of trials until the first success. The probability of success is $(k-1)/k$. In asking for the expected value of X_2, we are asking for expected number of steps until the first success. Thus we can apply Lemma 5.12 to get that it is $k/(k-1)$.

Continuing, X_3 similarly counts the number of steps in an independent trials process (with two outcomes) that stops at the first success and has probability of success $(k-2)/k$. Thus the expected number of steps until the first success is $k/(k-2)$.

In general, we have that X_i counts the number of trials until success in an independent trials process with probability of success $(k-i+1)/k$ and thus the expected number of steps until the first success is $k/(k-i+1)$, which is the expected value of X_i.

The total time until all slots are full is just $X = X_1 + \cdots + X_k$. Taking expectations and using Lemma 5.12 we get

$$\begin{aligned} E(X) &= \sum_{j=1}^{k} E(X_j) \\ &= \sum_{j=1}^{k} \frac{k}{k-j+1} \\ &= k\sum_{j=1}^{k} \frac{1}{k-j+1} \\ &= k\sum_{k-j+1=1}^{k} \frac{1}{k-j+1} \end{aligned}$$

$$= \quad k\sum_{i=1}^{k}\frac{1}{i},$$

where the last line follows just by switching the variable of the summation, that is, letting $k-j+1=i$ and summing over i.[6] Now the quantity $\sum_{i=1}^{k}\frac{1}{i}$ is known as a *harmonic number*, and is sometimes denoted by H_k. It is well known (and you can see why in the problems at the end of the section) that $\sum_{i=1}^{k}\frac{1}{i}=\Theta(\log k)$, and more precisely

$$\frac{1}{4}+\ln k \le H_k \le 1+\ln k, \tag{5.27}$$

and in fact,

$$\frac{1}{2}+\ln k \le H_k \le 1+\ln k, \tag{5.28}$$

when k is large enough. As n gets large, $H_n-\ln n$ approaches a limit called *Euler's constant*; Euler's constant is about .58. Equation 5.27 gives us that $E(X)=O(k\log k)$.

Theorem 5.16 *The expected number of items needed to fill all slots of a hash table of size k is between $k\ln k+\frac{1}{2}k$ and $k\ln k+k$.*

Proof: Given above. ■

So in order to fill every slot in a hash table of size k, we need to hash roughly $k\ln k$ items. This problem is sometimes called the *coupon collectors problem.*

The remainder of this section is devoted to proving that if we hash n items into a hash table with n slots, then the expected number of items per slot is $O(\log n/\log\log n)$. It should be no surpise that a result of this form requires a somewhat complex proof. The remainder of this section can be skipped without losss of continuity.

Expected maximum number of elements in a slot of a hash table (Optional)

In a hash table, the time to find an item is related to the number of items in the slot where you are looking. Thus an interesting quantity is the expected maximum length of the list of items in a slot in a hash table. This quantity is more complicated than many of the others we have been computing, and hence we will only try to upper bound it, rather than compute it exactly. In doing so, we will introduce a few upper bounds and techniques that appear frequently and are useful in many areas of mathematics and computer science. We will be able to prove that if we hash n items into a hash table of size n, the expected length of the longest list is $O(\log n/\log\log n)$. One can also prove, although we won't do it here, that with high probability, there will be some list with $\Omega(\log n/\log\log n)$ items in it, so our bound is, up to constant factors, the best possible.

Before we start, we give some useful upper bounds. The first allows us to bound terms that look like $(1+\frac{1}{x})^x$, for any positive x, by e.

Lemma 5.17 *For all $x>0$, $(1+\frac{1}{x})^x \le e$.*

[6]Note that $k-j+1$ runs from k to 1 as j runs from 1 to k, so we are describing exactly the same sum.

Proof: $\lim_{x\to\infty}(1+\frac{1}{x})^x = e$, and $(1+\frac{1}{x})^x$ has positive first derivative. ■

Second, we will use an approximation called Stirling's formula,

$$x! = \left(\frac{x}{e}\right)^x \sqrt{2\pi x}(1+\Theta(1/n)),$$

which tells us, roughly, that $(x/e)^x$ is a good approximation for $x!$. Moreover the constant in the $\Theta(1/n)$ term is extremely small, so for our purposes we will just say that

$$x! = \left(\frac{x}{e}\right)^x \sqrt{2\pi x}.$$

(We use this equality only in our proof of Lemma 5.18. You will see in that Lemma that we make the statement that $\sqrt{2\pi} > 1$. In fact, $\sqrt{2\pi} > 2$, and this is more than enough to make up for any lack of accuracy in our approximation.) Using Stirling's formula, we can get a bound on $\binom{n}{t}$,

Lemma 5.18 *For $n > t > 0$,*

$$\binom{n}{t} \leq \frac{n^n}{t^t(n-t)^{n-t}}.$$

Proof:

$$\binom{n}{t} = \frac{n!}{t!(n-t)!} \tag{5.29}$$

$$= \frac{(n/e)^n\sqrt{2\pi n}}{(t/e)^t\sqrt{2\pi t}((n-t)/e)^{n-t}\sqrt{2\pi(n-t)}} \tag{5.30}$$

$$= \frac{n^n\sqrt{n}}{t^t(n-t)^{n-t}\sqrt{2\pi}\sqrt{t(n-t)}} \tag{5.31}$$

Now if $1 < t < n-1$, we have $t(n-t) \geq n$, so that $\sqrt{t(n-t)} \geq \sqrt{n}$. Further $\sqrt{2\pi} > 1$. We can use these facts to upper bound the quantity marked 5.31 by

$$\frac{n^n}{t^t(n-t)^{n-t}}.$$

When $t = 1$ or $t = n-1$, the inequality in the statement of the lemma is $n \leq n^n/(n-1)^{n-1}$ which is true since $n-1 < n$. ■

We are now ready to attack the problem at hand, the expected value of the maximum list size. Let's start with a related quantity that we already know how to compute. Let H_{it} be the event that t keys hash to slot i. $P(H_{it})$ is just the probability of t successes in an independent trials process with success probability $1/n$, so

$$P(H_{it}) = \binom{n}{t}\left(\frac{1}{n}\right)^t\left(1-\frac{1}{n}\right)^{n-t}. \tag{5.32}$$

Now we relate this known quantity to the probability of the event M_t that the maximum list size is t.

Lemma 5.19 *Let M_t be the event that t is the maximum list size in hashing n items into a hash table of size n. Let H_{1t} be the event that t keys hash to position 1. Then*

$$P(M_t) \leq nP(H_{1t})$$

Proof: We begin by letting M_{it} be the event that the maximum list size is t and this list appears in slot i. Observe that that since M_{it} is a subset of H_{it},

$$P(M_{it}) \leq P(H_{it}). \tag{5.33}$$

We know that, by definition,

$$M_t = M_{1t} \cup \cdots \cup M_{nt},$$

and so

$$P(M_t) = P(M_{1t} \cup \cdots \cup M_{nt}).$$

Therefore, since the sum of the probabilities of the individual events must be at least as large as the probability of the union,

$$P(M_t) \leq P(M_{1t}) + P(M_{2t}) + \cdots + P(M_{nt}). \tag{5.34}$$

(Recall that we introduced the Principle of Inclusion and Exclusion because the right hand side overestimated the probability of the union. Note that the inequality in Equation 5.34 holds for any union, not just this one: it is sometimes called *Boole's inequality.*)

In this case, for any i and j, $P(M_{it}) = P(M_{jt})$, since there is no reason for slot i to be more likely than slot j to be the maximum. We can therefore write that

$$P(M_t) = nP(M_{1t}) \leq nP(H_{1t}).$$

■

Now we can use Equation 5.32 for $P(H_{1t})$ and then apply Lemma 5.18 to get that

$$\begin{aligned} P(H_{1t}) &= \binom{n}{t}\left(\frac{1}{n}\right)^t\left(1-\frac{1}{n}\right)^{n-t} \\ &\leq \frac{n^n}{t^t(n-t)^{n-t}}\left(\frac{1}{n}\right)^t\left(1-\frac{1}{n}\right)^{n-t}. \end{aligned}$$

We continue, using algebra, the fact that $(1-\frac{1}{n})^{n-t} \leq 1$ and Lemma 5.17 to get

$$\begin{aligned} &\leq \frac{n^n}{t^t(n-t)^{n-t}n^t} \\ &= \frac{n^{n-t}}{t^t(n-t)^{n-t}} \\ &= \left(\frac{n}{n-t}\right)^{n-t}\frac{1}{t^t} \\ &= \left(1+\frac{t}{n-t}\right)^{n-t}\frac{1}{t^t} \\ &= \left(\left(1+\frac{t}{n-t}\right)^{\frac{n-t}{t}}\right)^t\frac{1}{t^t} \\ &\leq \frac{e^t}{t^t}. \end{aligned}$$

We have shown the following:

Lemma 5.20 *$P(M_t)$, the probability that the maximum list length is t, is at most ne^t/t^t.*

Proof: Our sequence of equations and inequalities above showed that $P(H_{1t}) \leq \frac{e^t}{t^t}$. Multiplying by n and applying Lemma 5.19 gives our result.■

Now that we have a bound on $P(M_t)$, we can compute a bound on the expected length of the longest list, namely

$$\sum_{t=0}^{n} P(M_t)t.$$

However, if we think carefully about the bound in Lemma 5.20, we see that we have a problem. For example when $t = 1$, the lemma tells us that $P(M_1) \leq ne$. This is vacuous, as we know that any probability is at most 1. We could make a stronger statement that $P(M_t) \leq \max\{ne^t/t^t, 1\}$, but even this wouldn't be sufficient, since it would tell us things like $P(M_1) + P(M_2) \leq 2$, which is also vacuous. All is not lost however. Our lemma causes this problem only when t is small. We will split the sum defining the expected value into two parts and bound the expectation for each part separately. The intuition is that when we restrict t to be small, then $\sum P(M_t)t$ is small because t is small (and over all t, $\sum P(M_t) \leq 1$). When t gets larger, Lemma 5.20 tells us that $P(M_t)$ is very small and so the sum doesn't get big in that case either. We will choose a way to split the sum so that this second part of the sum is bounded by a constant. In particular we split the sum up by

$$\sum_{t=0}^{n} P(M_t)t \leq \sum_{t=0}^{\lfloor 5\log n/\log\log n\rfloor} P(M_t)t + \sum_{t=\lceil 5\log n/\log\log n\rceil}^{n} P(M_t)t \tag{5.35}$$

For the sum over the smaller values of t, we just observe that in each term $t \leq 5\log n/\log\log n$ so that

$$\begin{aligned}
\sum_{t=0}^{5\log n/\log\log n} P(M_t)t &\leq \sum_{t=0}^{5\log n/\log\log n} P(M_t)5\log n/\log\log n && (5.36)\\
&= 5\log n/\log\log n \sum_{t=0}^{5\log n/\log\log n} P(M_t) && (5.37)\\
&\leq 5\log n/\log\log n && (5.38)
\end{aligned}$$

(Note that we are not using Lemma 5.20 here; only the fact that the probabilities of disjoint events cannot add to more than 1.) For the rightmost sum in Equation 5.35, we want to first compute an upper bound on $P(M_t)$ for $t = (5\log n/\log\log n)$. Using Lemma 5.20, and doing a bit of calculation we get that in this case $P(M_t) \leq 1/n^2$. Since the bound on $P(M_t)$ from Lemma 5.20 decreases as t grows, and $t \leq n$, we can bound the right sum by

$$\sum_{t=5\log n/\log\log n}^{n} P(M_t)t \leq \sum_{t=5\log n/\log\log n}^{n} \frac{1}{n^2}n \leq \sum_{t=5\log n/\log\log n}^{n} \frac{1}{n} \leq 1. \tag{5.39}$$

Combining Equations 5.38 and 5.39 with 5.35 we get the desired result.

Theorem 5.21 *If we hash n items into a hash table of size n, the expected maximum list length is $O(\log n/\log\log n)$.*

The choice to break the sum into two pieces here—and especially the breakpoint we chose—may have seemed like magic. What is so special about $\log n / \log\log n$? Consider the bound on $P(M_t)$. If you asked what is the value of t for which the bound equals a certain value, say $1/n^2$, you get the equation $ne^t/t^t = n^{-2}$. If we try to solve the equation $ne^t/t^t = n^{-2}$ for t, we quickly see that we get a form that we do not know how to solve. (Try typing this into Mathematica or Maple, to see that it can't solve this equation either.) The equation we need to solve is somewhat similar to the simpler equation $t^t = n$. While this equation does not have a closed form solution, one can show that the t that satisfies this equation is roughly $c \log n / \log\log n$, for some constant c. This is why some multiple of $\log n / \log\log n$ made sense to try as the the magic value. For values much less than $\log n / \log\log n$ the bound provided on $P(M_t)$ is fairly large. Once we get past $\log n / \log\log n$, however, the bound on $P(M_t)$ starts to get significantly smaller. The factor of 5 was chosen by experimentation to make the second sum come out to be less than 1. We could have chosen any number between 4 and 5 to get the same result; or we could have chosen 4 and the second sum would have grown no faster than the first.

Important Concepts, Formulas, and Theorems

1. *Expected Number of Keys per Slot in Hash Table.* In hashing n items into a hash table of size k, the expected number of items that hash to any one location is n/k.

2. *Expected Number of Empty Slots in Hash Table.* In hashing n items into a hash table with k locations, the expected number of empty locations is $k(1 - \frac{1}{k})^n$.

3. *Collision in Hashing.* We say that we have a *collision* when we hash an item to a location that already contains an item.

4. *The Expected Number of Collisions in Hashing.* In hashing n items into a hash table with k locations, the expected number of collisions is $n - k + k(1 - \frac{1}{k})^n$.

5. *Harmonic Number.* The quantity $\sum_{i=1}^{k} \frac{1}{i}$ is known as a *harmonic number*, and is sometimes denoted by H_k. It is a fact that that $\sum_{i=1}^{k} \frac{1}{i} = \Theta(\log k)$, and more precisely

$$\frac{1}{2} + \ln k \leq H_k \leq 1 + \ln k.$$

6. *Euler's Constant.* As n gets large, $H_n - \ln n$ approaches a limit called *Euler's constant*; Euler's constant is about .58.

7. *Expected Number of Hashes until all Slots of a Hash Table are Occupied.* The expected number of items needed to fill all slots of a hash table of size k is between $k \ln k + \frac{1}{2}k$ and $k \ln k + k$.

8. *Expected Maximum Number of Keys per Slot.* If we hash n items into a hash table of size n, the expected maximum list length is $O(\log n / \log\log n)$.

9. *Stirling's Formula for n!. (Optional.)* $n!$ is approximately $\left(\frac{n}{e}\right)^n \sqrt{2\pi n}$.

Problems

1. A candy machine in a school has d different kinds of candy. Assume (for simplicity) that all these kinds of candy are equally popular and there is a large supply of each. Suppose that c children come to the machine and each purchases one package of candy. One of the kinds of candy is a Snackers bar. What is the probability that any given child purchases a Snackers bar? Let Y_i be the number of Snackers bars that child i purchases, so that Y_i is either 0 or 1. What is the expected value of Y_i? Let Y be the random variable $Y_1 + Y_2 + \cdots + Y_c$. What is the expected value of Y? What is the expected number of Snackers bars that is purchased? Does the same result apply to any of the varieties of candy?

2. Again as in the previous exercise, we have c children choosing from among ample supplies of d different kinds of candy, one package for each child, and all choices equally likely. What is the probability that a given variety of candy is chosen by no child? What is the expected number of kinds of candy chosen by no child? Suppose now that $c = d$. What happens to the expected number of kinds of candy chosen by no child?

3. How many children do we expect to have to observe buying candy until someone has bought a Snackers bar?

4. How many children to we expect to have to observe buying candy until each type of candy has been selected at least once?

5. If we have 20 kinds of candy, how many children have to buy candy in order for the probability to be at least one half that (at least) two children buy the same kind of candy?

6. What is the expected number of duplications among all the candy the children have selected?

7. Compute the values on the left-hand and right-hand side of the inequality in Lemma 5.18 for $n = 2$, $t = 0, 1, 2$ and for $n = 3$, $t = 0, 1, 2, 3$.

8. When we hash n items into k locations, what is the probability that all n items hash to different locations? What is the probability that the ith item is the first collision? What is the expected number of items we must hash until the first collision? Use a computer program or spreadsheet to compute the expected number of items hashed into a hash table until the first collision with $k = 20$ and with $k = 100$.

9. We have seen a number of occasions when our intuition about expected values or probability in general fails us. When we wrote down Equation 5.26 we said that the expected number of occupied locations is k minus the expected number of unoccupied locations. While this seems obvious, there is a short proof. Give the proof.

10. Write a computer program that prints out a table of values of the expected number of collisions with n keys hashed into a table with k locations for interesting values of n and k. Does this value vary much as n and k change?

11. Suppose you hash n items into a hash table of size k. It is natural to ask about the time it takes to find an item in the hash table. We can divide this into two cases, one when the item is not in the hash table (an unsuccessful search), and one when the item is in the hash table (a successful search). Consider first the unsuccessful search. Assume the keys hashing

to the same location are stored in a list with the most recent arrival at the beginning of the list. Use our expected list length to bound the expected time for an unsuccessful search. Next consider the successful search. Recall that when we insert items into a hash table, we typically insert them at the beginning of a list, and so the time for a successful search for item i should depend on how many entries were inserted after item i. Carefully compute the expected running time for a successful search. Assume that the item you are searching for is randomly chosen from among the items already in the table. (Hint: The unsuccessful search should take roughly twice as long as the successful one. Be sure to explain why this is the case.)

12. Suppose I hash $n \log n$ items into n buckets. What is the expected maximum number of items in a bucket?

13. The fact that $\lim_{n\to\infty}(1+\frac{1}{n})^n = e$ (where n varies over integers) is a consequence of the fact that $\lim_{h\to 0}(1+h)^{\frac{1}{h}} = e$ (where h varies over real numbers). Thus if h varies over negative real numbers, but approaches 0, the limit still exists and equals e. What does this tell you about $\lim_{n\to-\infty}(1+\frac{1}{n})^n$? Using this and rewriting $(1-\frac{1}{n})^n$ as $(1+\frac{1}{-n})^n$ show that

14. What is the expected number of empty slots when we hash $2k$ items into a hash table with k slots? What is the expected fraction of empty slots close to when k is reasonably large?

15. Using whatever methods you like (hand calculations or computer), give upper and/or lower bounds on the value of the x satisfying $x^x = n$.

16. Professor Max Weinberger decides that the method proposed for computing the maximum list size is much too complicated. He proposes the following solution. Let X_i be the size of list i. Then what we want to compute is $E(\max_i(X_i))$. Well

$$E(\max_i(X_i)) = \max_i(E(X_i)) = \max_i(1) = 1.$$

What is the flaw in his solution?

17. Prove as tight upper and lower bounds as you can on $\sum_{i=1}^{k}\frac{1}{i}$. For this purpose it is useful to remember the definition of the natural logarithm as an integral involving $1/x$ and to draw rectangles and other geometric figures above and below the curve.

18. Notice that $\ln n! = \sum_{i=1}^{n}\ln i$. Sketch a careful graph of $y = \ln x$, and by drawing in geometric figures above and below the graph, show that

$$\sum_{i=1}^{n}\ln i - \frac{1}{2}\ln n \le \int_1^n \ln x\,dx \le \sum_{i=1}^{n}\ln i.$$

Based on your drawing, which inequality do you think is tighter? Use integration by parts to evaluate the integral. What bounds on $n!$ can you get from these inequalities? Which one do you think is tighter? How does it compare to Stirling's approximation? What big Oh bound can you get on $n!$?

5.6 Conditional Expectations, Recurrences and Algorithms

Probability is a very important tool in algorithm design. We have already seen two important examples in which it is used—primality testing and hashing. In this section we will study several more examples of probabilistic analysis in algorithms. We will focus on computing the running time of various algorithms. When the running time of an algorithm is different for different inputs of the same size, we can think of the running time of the algorithm as a random variable on the sample space of inputs and analyze the expected running time of the algorithm. This gives us a different understanding from studying just the worst case running time for an input of a given size. We will then consider *randomized algorithms*, algorithms that depend on choosing something randomly, and see how we can use recurrences to give bounds on their expected running times as well.

For randomized algorithms, it will be useful to have access to a function which generates random numbers. We will assume that we have a function `randint(i,j)`, which generates a random integer uniformly between i and j (inclusive) [this means it is equally likely to be any number between i and j] and `rand01()`, which generates a random real number, uniformly between 0 and 1 [this means that given any two pairs of real numbers (r_1, r_2) and (s_1, s_2) with $r_2 - r_1 = s_2 - s_1$ and r_1, r_2, s_1 and s_2 all between 0 and 1, our random number is just as likely to be between r_1 and r_2 as it is to be between s_1 and s_2]. Functions such as `randint` and `rand01` are called *random number generators*. A great deal of number theory goes into the construction of good random number generators.

When Running Times Depend on more than Size of Inputs

Exercise 5.6-1 Let A be an array of length $n-1$ (whose elements are chosen from some ordered set), sorted into increasing order. Let b be another element of that ordered set that we want to insert into A to get a sorted array of length n. Assuming that the elements of A and b are chosen randomly, what is the expected number of elements of A that have to be shifted one place to the right to let us insert b?

Exercise 5.6-2 Let $A(1:n)$ denote the elements in positions 1 to n of the array A. A recursive description of insertion sort is that to sort $A(1:n)$, first we sort $A(1:n-1)$, and then we insert $A(n)$, by shifting the elements greater than $A(n)$ each one place to the right and then inserting the original value of $A(n)$ into the place we have opened up. If $n = 1$ we do nothing. Let $S_j(A(1:j))$ be the time needed to sort the portion of A from place 1 to place j, and let $I_j(A(1:j), b)$ be the time needed to insert the element b into a sorted list originally in the first j positions of A to give a sorted list in the first $j+1$ positions of A. Note that S_j and I_j depend on the actual array A, and not just on the value of j. Use S_j and I_j to describe the time needed to use insertion sort to sort $A(1:n)$ in terms of the time needed to sort $A(1:n-1)$. Don't forget that it is necessary to copy the element in position i of A into a variable b before moving elements of $A(1:i-1)$ to the right to make a place for it, because this moving process will write over $A(i)$. Let $T(n)$ be the expected value of S_n; that is, the expected running time of insertion sort on a list of n items. Write a recurrence for $T(n)$ in terms of $T(n-1)$ by taking expected values in the equation that corresponds to your previous description of the time needed to use insertion sort on a particular array. Solve your recurrence relation in big-Θ terms.

If X is the random variable with $X(A, b)$ equal to the number of items we need to move one place to the right in order to insert b into the resulting empty slot in A, then X takes on the values $0, 1, \ldots, n-1$ with equal probability $1/n$. Thus we have

$$E(x) = \sum_{i=0}^{n-1} i\frac{1}{n} = \frac{1}{n}\sum_{i=0}^{n-1} i = \frac{1}{n}\frac{(n-1)n}{2} = \frac{n-1}{2}.$$

Using $S_j(A(1:j))$ to stand for the time to sort the portion of the array A from places 1 to j by insertion sort, and $I_j(A(1:j), b)$ to stand for the time needed to insert b into a sorted list in the first j positions of the array A, moving all items larger than j to the right one place and putting b into the empty slot, we can write that for insertion sort

$$S_n(A(1:n)) = S_{n-1}(A(1:n-1)) + I_{n-1}(A(1:n-1), A(n)) + c_1.$$

We have included the constant term c_1 for the time it takes to copy the value of $A(n)$ into some variable b, because we will overwrite $A(n)$ in the process of moving items one place to the right. Using the additivity of expected values, we get

$$E(S_n) = E(S_{n-1}) + E(I_{n-1}) + E(c_1).$$

Using $T(n)$ for the expected time to sort $A(1:n)$ by insertion sort, and the result of the previous exercise, we get

$$T(n) = T(n-1) + c_2\frac{n-1}{2} + c_1.$$

where we include the constant c_2 because the time needed to do the insertion is going to be proportional to the number of items we have to move plus the time needed to copy the value of $A(n)$ into the appropriate slot (which we will assume we have included in c_1). We can say that $T(1) = 1$ (or some third constant) because with a list of size 1 we have to realize it has size 1, and then do nothing. It might be more realistic to write

$$T(n) \leq T(n-1) + cn$$

and

$$T(n) \geq T(n-1) + c'n,$$

because the time needed to do the insertion may not be exactly proportional to the number of items we need to move, but might depend on implementation details. By iterating the recurrence or drawing a recursion tree, we see that $T(n) = \Theta(n^2)$. (We could also give an inductive proof.) Since the best-case time of insertion sort is $\Theta(n)$ and the worst case time is $\Theta(n^2)$, it is interesting to know that the expected case is much closer to the worst-case than the best case.

Conditional Expected Values

Our next example is cooked up to introduce an idea that we often use in analyzing the expected running times of algorithms, especially randomized algorithms.

Exercise 5.6-3 I have two nickels and two quarters in my left pocket and 4 dimes in my right pocket. Suppose I flip a penny and take two coins from my left pocket if it is heads, and two coins from my right pocket if it is tails. Assuming I am equally likely to choose any coin in my pocket at any time, what is the expected amount of money that I draw from my pocket?

You could do this problem by drawing a tree diagram or by observing that the outcomes can be modeled by three tuples in which the first entry is heads or tails, and the second and third entries are coins. Thus our sample space is HNQ, HQN, HQQ, HNN, TDD The probabilities of these outcomes are $\frac{1}{6}$, $\frac{1}{6}$, $\frac{1}{12}$, $\frac{1}{12}$, and $\frac{1}{2}$ respectively. Thus our expected value is

$$30\frac{1}{6} + 30\frac{1}{6} + 50\frac{1}{12} + 10\frac{1}{12} + 20\frac{1}{2} = 25.$$

Here is a method that seems even simpler. If the coin comes up heads, I have an expected value of 15 cents on each draw, so with probability 1/2, our expected value is 30 cents. If the coin comes up tails, I have an expected value of ten cents on each draw, so with probability 1/2 our expected value is 20 cents. Thus it is natural to expect that our expected value is $\frac{1}{2}30 + \frac{1}{2}20 = 25$ cents. In fact, if we group together the 4 outcomes with an H first, we see that their contribution to the expected value is 15 cents, which is 1/2 times 30, and if we look at the single element which has a T first, then its contribution to the sum is 10 cents, which is half of 20 cents.

In this second view of the problem, we took the probability of heads times the expected value of our draws, given that the penny came up heads, plus the probability of tails times the expected value of our draws, given that the penny came came up tails. In particular, we were using a new (and as yet undefined) idea of *conditional expected value*. To get the conditional expected value if our penny comes up heads, we could create a new sample space with four outcomes, NQ, QN, NN, QQ, with probabilities $\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{6}$, and $\frac{1}{6}$. In this sample space the expected amount of money we draw in two draws is 30 cents (15 cents for the first draw plus 15 cents for the second), so we would say the conditional expected value of our draws, given that the penny came up heads, was 30 cents. With a one-element sample space $\{DD\}$, we see that we would say that the conditional expected value of our draws, given that the penny came up tails, is 20 cents.

How do we define conditional expected value? Rather than create a new sample space as we did above, we use the idea of a new sample space (as we did in discovering a good definition for conditional probability) to lead us to a good definition for conditional expected value. In particular, to get the conditional expected value of X given that an event F has happened we use our conditional probability weights for the elements of F, namely $P(s)/P(F)$ is the weight for the element s of F, and pretend F is our sample space. Thus we define the **conditional expected value** of X given F by

$$E(X|F) = \sum_{x:s\in F} X(x)\frac{P(x)}{P(F)}. \tag{5.40}$$

Remember that we defined the expected value of a random variable X with values $x_1, x_2, \ldots x_k$ by

$$E(X) = \sum_{i=1}^{k} x_i P(X = x_i),$$

where $X = x_i$ stands for the event that X has the value x_i. Using our standard notation for conditional probabilities, $P(X = x_i|F)$ stands for the conditional probability of the event $X = x_i$ given the event F. This lets us rewrite Equation 5.40 as

$$E(X|F) = \sum_{i=1}^{k} x_i P(X = x_i|F).$$

Theorem 5.22 *Let X be a random variable defined on a sample space S and let F_1, F_2, ... F_n be disjoint events whose union is S (i.e. a partition of S). Then*

$$E(X) = \sum_{i=1}^{n} E(X|F_i)P(F_i).$$

Proof: The proof is simply an exercise in applying definitions. ■

Randomized algorithms

Exercise 5.6-4 Consider an algorithm that, given a list of n numbers, prints them all out. Then it picks a random integer between 1 and 3. If the number is 1 or 2, it stops. If the number is 3 it starts again from the beginning. What is the expected running time of this algorithm?

Exercise 5.6-5 Consider the following variant on the previous algorithm:

```
funnyprint(n)
if (n == 1)
     return
for i = 1 to n
     print i
x = randint(1,n)
if (x > n/2)
     funnyprint(n/2)
else
     return
```

What is the expected running time of this algorithm?

For Exercise 5.6-4, with probability 2/3 we will print out the numbers and quit, and with probability 1/3 we will run the algorithm again. Using Theorem 5.22, we see that if $T(n)$ is the expected running time on a list of length n, then there is a constant c such that

$$T(n) = \frac{2}{3}cn + \frac{1}{3}(cn + T(n)),$$

which gives us $\frac{2}{3}T(n) = cn$. This simplifies to $T(n) = \frac{3}{2}cn$.

Another view is that we have an independent trials process, with success probability 2/3 where we stop at the first success, and for each round of the independent trials process we spend $\Theta(n)$ time. Letting T be the running time (note that T is a random variable on the sample space $1, 2, 3$ with probabilities $\frac{1}{3}$ for each member) and R be the number of rounds, we have that

$$T = R \cdot \Theta(n)$$

and so

$$E(T) = E(R)\Theta(n).$$

Note that we are applying Theorem 5.10 since in this context $\Theta(n)$ behaves as if it were a constant[7], since n does not depend on R. By Lemma 5.12, we have that $E(R) = 3/2$ and so $E(T) = \Theta(n)$.

In Exercise 5.6-5, we have a recursive algorithm, and so it is appropriate to write down a recurrence. We can let $T(n)$ stand for the *expected* running time of the algorithm on an input of size n. Notice how we are changing back and forth between letting T stand for the running time of an algorithm and the expected running time of an algorithm. Usually we use T to stand for the quantity of most interest to us, either running time if that makes sense, or expected running time (or maybe worst-case running time) if the actual running time might vary over different inputs of size n. The nice thing will be that once we write down a recurrence for the expected running time of an algorithm, the methods for solving it will be those for we have already learned for solving recurrences. For the problem at hand, we immediately get that with probability 1/2 we will be spending n units of time (we should really say $\Theta(n)$ time), and then terminating, and with probability 1/2 we will spend n units of time and then recurse on a problem of size $n/2$. Thus using Theorem 5.22, we get that

$$T(n) = n + \frac{1}{2}T(n/2)$$

Including a base case of $T(1) = 1$, we get that

$$T(n) = \begin{cases} \frac{1}{2}T(n/2) + n & \text{if } n > 1 \\ 1 & \text{if } n = 1 \end{cases}.$$

A simple proof by induction shows that $T(n) = \Theta(n)$. Note that the Master Theorem (as we originally stated it) doesn't apply here, since $a < 1$. However, one could also observe that the solution to this recurrence is no more than the solution to the recurrence $T(n) = T(n/2) + n$, and then apply the Master Theorem.

Selection revisited

We now return to the selection algorithm from Section 4.6. The purpose of the algorithm is to select the ith smallest element in a set with some underlying order. Recall that in this algorithm, we first picked an an element p in the middle half of the set, that is, one whose value was simultaneously larger than at least 1/4 of the items and smaller than at least 1/4 of the items. We used p to partition the items into two sets and then recursed on one of the two sets. If you recall, we worked very hard to find an item in the middle half, so that our partitioning would

[7]What we mean here is that $T \geq Rc_1 n$ for some constant c_1 and $T \leq Rc_2 n$ for some other constant c_2. Then we apply Theorem 5.10 to both these inequalities, using the fact that if $X > Y$, then $E(X) > E(Y)$ as well.

work well. It is natural to try instead to just pick a partition element at random, because, with probability 1/2, this element will be in the middle half. We can extend this idea to the following algorithm:

```
RandomSelect(A,i,n)
(selects the ith smallest element in set A, where n = |A| )
if (n = 1)
     return the one item in A
else
     p = randomElement(A)
     Let H be the set of elements greater than p
     Let L be the set of elements less than or equal to p
     If (H is empty)
          put p in H
     if (i ≤ |L|)
          Return RandomSelect(L, i, |L|)
     else
          Return RandomSelect(H, i − |L|, |H|).
```

Here randomElement(A) returns one element from A uniformly at random. We use this element as our partition element; that is, we use it to divide A into sets L and H with every element less than the partition element in L and every element greater than it in H. We add the special case when H is empty, to ensure that both recursive problems have size strictly less than n. This simplifies a detailed analysis, but is not strictly necessary. At the end of this section we will show how to get a recurrence that describes fairly precisely the time needed to carry out this algorithm. However, by being a bit less precise, we can still get the same big-O upper bound with less work.

When we choose our partition element, half the time it will be between $\frac{1}{4}n$ and $\frac{3}{4}n$. Then when we partition our set into H and L, each of these sets will have no more than $\frac{3}{4}n$ elements. The other half of the time each of H and L will have no more than n elements. In any case, the time to partition our set into H and L is $O(n)$. Thus we may write

$$T(n) \leq \begin{cases} \frac{1}{2}T(\frac{3}{4}n) + \frac{1}{2}T(n) + bn & \text{if } n > 1 \\ d & \text{if } n = 1. \end{cases}$$

We may rewrite the recursive part of the recurrence as

$$\frac{1}{2}T(n) \leq \frac{1}{2}T\left(\frac{3}{4}n\right) + bn,$$

or

$$T(n) \leq T\left(\frac{3}{4}n\right) + 2bn = T\left(\frac{3}{4}n\right) + b'n.$$

Notice that it is possible (but unlikely) that each time our algorithm chooses a pivot element, it chooses the worst one possible, in which case the selection process could take n rounds, and thus take time $\Theta(n^2)$. Why, then, is it of interest? It involves far less computation than finding the median of medians, and its expected running time is still $\Theta(n)$. Thus it is reasonable to suspect that on the average, it would be significantly faster than the deterministic process. In fact, with good implementations of both algorithms, this will be the case.

Exercise 5.6-6 Why does every solution to the recurrence

$$T(n) \leq T\left(\frac{3}{4}n\right) + b'n$$

have $T(n) = O(n)$?

By the master theorem we know that any solution to this recurrence is $O(n)$, giving a proof of our next Theorem.

Theorem 5.23 *Algorithm RandomSelect has expected running time $O(n)$.*

Quicksort

There are many algorithms that will efficiently sort a list of n numbers. The two most common sorting algorithms that are guaranteed to run in $O(n \log n)$ time are MergeSort and HeapSort. However, there is another algorithm, Quicksort, which, while having a worst-case running time of $O(n^2)$, has an expected running time of $O(n \log n)$. Moreover, when implemented well, it tends to have a faster running time than MergeSort or HeapSort. Since many computer operating systems and programs come with Quicksort built in, it has become the sorting algorithm of choice in many applications. In this section, we will see why it has expected running time $O(n \log n)$. We will not concern ourselves with the low-level implementation issues that make this algorithm the fastest one, but just with a high-level description.

Quicksort actually works similarly to the RecursiveSelect algorithm of the previous subsection. We pick a random element, and then use it to partition the set of items into two sets L and H. In this case, we don't recurse on one or the other, but recurse on both, sorting each one. After both L and H have been sorted, we just concatenate them to get a sorted list. (In fact, Quicksort is usually done "in place" by pointer manipulation and so the concatenation just happens.) Here is a pseudocode description of Quicksort:

```
Quicksort(A,n)
if (n = 1)
     return the one item in A
else
     p = randomElement(A)
     Let H be the set of elements greater than p; Let h = |H|
     Let L be the set of elements less than or equal to p; Let ℓ = |L|
     If (H is empty)
          put p in H
     A_1 = QuickSort(H,h)
     A_2 = QuickSort(L,ℓ)
     return the concatenation of A_1 and A_2.
```

There is an analysis of Quicksort similar to the detailed analysis of RecursiveSelect at the end of the section, and this analysis is a problem at the end of the section. Instead, based on the preceding analysis of RandomSelect we will think about modifying the algorithm a bit in order

to make the analysis easier. First, consider what would happen if the random element was the median each time. Then we would be solving two subproblems of size $n/2$, and would have the recurrence

$$T(n) = \begin{cases} 2T(n/2) + O(n) & \text{if } n > 1 \\ O(1) & \text{if } n = 1 \end{cases}$$

and we know by the master theorem that all solutions to this recurrence have $T(n) = O(n \log n)$. In fact, we don't need such an even division to guarantee such performance.

Exercise 5.6-7 Suppose you had a recurrence of the form

$$T(n) \leq \begin{cases} T(a_n n) + T((1 - a_n)n) + cn & \text{if } n > 1 \\ d & \text{if } n = 1, \end{cases}$$

where a_n is between 1/4 and 3/4. Show that all solutions of a recurrence of this form have $T(n) = O(n \log n)$. What do we really need to assume about a_n in order to prove this upper bound?

We can prove that $T(n) = O(n \log n)$ by induction, or via a recursion tree, noting that there are $O(\log n)$ levels, and each level has at most $O(n)$ work. (The details of the recursion tree are complicated somewhat by the fact that a_n varies with n, while the details of an inductive proof simply use the fact that a_n and $1 - a_n$ are both no more than 3/4.) So long as we know there is some positive number $a < 1$ such that $a_n < a$ for every n, then we know we have at most $\log_{(1/a)} n$ levels in a recursion tree, with at most cn units of work per level for some constant c, and thus we have the same upper bound in big-O terms.

What does this tell us? As long as our problem splits into two pieces, each having size at least 1/4 of the items, Quicksort will run in $O(n \log n)$ time. Given this, we will modify our algorithm to enforce this condition. That is, if we choose a pivot element p that is not in the middle half, we will just pick another one. This leads to the following algorithm:

```
Slower Quicksort(A,n)
if (n = 1)
     return the one item in A
else
     Repeat
          p = randomElement(A)
          Let H be the set of elements greater than p; Let h = |H|
          Let L be the set of elements less than or equal to p; Let ℓ = |L|
     Until (|H| ≥ n/4) and (|L| ≥ n/4)
     A_1 = QuickSort(H,h)
     A_2 = QuickSort(L,ℓ)
     return the concatenation of A_1 and A_2
```

Now let's analyze this algorithm. Let r be the number of times we execute the loop to pick p, and let $a_n n$ be the position of the pivot element. Then if $T(n)$ is the expected running time for a list of length n, then for some constant b

$$T(n) \leq E(r)bn + T(a_n n) + T((1 - a_n)n),$$

since each iteration of the loop takes $O(n)$ time. Note that we take the expectation of r, because $T(n)$ stands for the expected running time on a problem of size n. Fortunately, $E(r)$ is simple to compute, it is just the expected time until the first success in an independent trials process with success probability 1/2. This is 2. So we get that the running time of Slower Quicksort satisfies the recurrence

$$T(n) \leq \begin{cases} T(a_n n) + T((1-a_n))n + b'n & \text{if } n > 1 \\ d & \text{if } n = 1 \end{cases},$$

where a_n is between 1/4 and 3/4. Thus by Exercise 5.6-7 the running time of this algorithm is $O(n \log n)$.

As another variant on the same theme, observe that looping until we have($|H| \geq n/4$ and $|L| \geq n/4$, is effectively the same as choosing p, finding H and L and then calling Slower Quicksort(A,n) once again if either H or L has size less than $n/4$. Then since with probability 1/2 the element p is between $n/4$ and $3n/4$, we can write

$$T(n) \leq \frac{1}{2}T(n) + \frac{1}{2}(T(a_n n) + T((1-a_n)n) + bn),$$

which simplifies to

$$T(n) \leq T(a_n n) + T((1-a_n)n) + 2bn,$$

or

$$T(n) \leq T(a_n n) + T((1-a_n)n) + b'n.$$

Again by Exercise 5.6-7 the running time of this algorithm is $O(n \log n)$.

Further, it is straightforward to see that the expected running time of Slower Quicksort is no less than half that of Quicksort (and, incidentally, no more than twice that of Quicksort) and so we have shown:

Theorem 5.24 *Quicksort has expected running time $O(n \log n)$.*

A more exact analysis of RandomSelect

Recall that our analysis of the RandomSelect was based on using $T(n)$ as an upper bound for $T(|H|)$ or $T(|L|)$ if either the set H or the set L had more than $3n/4$ elements. Here we show how one can avoid this assumption. The kinds of computations we do here are the kind we would need to do if we wanted to try to actually get bounds on the constants implicit in our big-O bounds.

Exercise 5.6-8 Explain why, if we pick the kth element as the random element in RandomSelect ($k \neq n$), our recursive problem is of size no more than $\max\{k, n-k\}$.

If we pick the kth element, then we recurse either on the set L, which has size k, or on the set H, which has size $n-k$. Both of these sizes are at most $\max\{k, n-k\}$. (If we pick the nth element, then $k = n$ and thus L actually has size $k-1$ and H has size $n-k+1$.)

Now let X be the random variable equal to the rank of the chosen random element (e.g. if the random element is the third smallest, $X = 3$.) Using Theorem 5.22 and the solution to Exercise 5.6-8, we can write that

$$T(n) \leq \begin{cases} \sum_{k=1}^{n-1} P(X=k)(T(\max\{k, n-k\}) + bn) + P(X=n)(T(\max\{1, n-1\} + bn) & \text{if } n > 1 \\ d & \text{if } n = 1. \end{cases}$$

Since X is chosen uniformly between 1 and n, $P(X = k) = 1/n$ for all k. Ignoring the base case for a minute, we get that

$$\begin{aligned} T(n) &\leq \sum_{k=1}^{n-1} \frac{1}{n}(T(\max\{k, n-k\}) + bn) + \frac{1}{n}(T(n-1) + bn) \\ &= \frac{1}{n}\left(\sum_{k=1}^{n-1} T(\max\{k, n-k\})\right) + bn + \frac{1}{n}(T(n-1) + bn). \end{aligned}$$

Now if n is odd and we write out $\sum_{k=1}^{n-1} T(\max\{k, n-k\})$, we get

$$T(n-1) + T(n-2) + \cdots + T(\lceil n/2 \rceil) + T(\lceil n/2 \rceil) + \cdots + T(n-2) + T(n-1),$$

which is just $2\sum_{k=\lceil n/2 \rceil}^{n-1} T(k)$. If n is even and we write out $\sum_{k=1}^{n-1} T(\max\{k, n-k\})$, we get

$$T(n-1) + T(n-2) + \cdots + T(n/2) + T(1 + n/2) + \cdots + T(n-2) + T(n-1),$$

which is less than $2\sum_{k=n/2}^{n-1} T(k)$. Thus we can replace our recurrence by

$$T(n) \leq \begin{cases} \frac{2}{n}\left(\sum_{k=n/2}^{n-1} T(k)\right) + \frac{1}{n}T(n-1) + bn & \text{if } n > 1 \\ d & \text{if } n = 1. \end{cases} \tag{5.41}$$

If n is odd, the lower limit of the sum is a half-integer, so the possible integer values of the dummy variable k run from $\lceil n/2 \rceil$ to $n-1$. Since this is the natural way to interpret a fractional lower limit, and since it corresponds to what we wrote in both the n even and n odd case above, we adopt this convention.

Exercise 5.6-9 Show that every solution to the recurrence in Equation 5.41 has $T(n) = O(n)$.

We can prove this by induction. We try to prove that $T(n) \leq cn$ for some constant c. By the natural inductive hypothesis, we get that

$$\begin{aligned} T(n) &\leq \frac{2}{n}\left(\sum_{k=n/2}^{n-1} ck\right) + \frac{1}{n}c(n-1) + bn \\ &= \frac{2}{n}\left(\sum_{k=1}^{n-1} ck - \sum_{k=1}^{n/2-1} ck\right) + \frac{1}{n}c(n-1) + bn \\ &\leq \frac{2c}{n}\left(\frac{(n-1)n}{2} - \frac{(\frac{n}{2}-1)\frac{n}{2}}{2}\right) + c + bn \\ &= \frac{2c}{n}\frac{\frac{3n^2}{4} - \frac{n}{2}}{2} + c + bn \\ &= \frac{3}{4}cn + \frac{c}{2} + bn \\ &= cn - (\frac{1}{4}cn - bn - \frac{c}{2}) \end{aligned}$$

Notice that so far, we have only assumed that there is some constant c such that $T(k) < ck$ for $k < n$. We can choose a larger c than the one given to us by this assumption without changing

the inequality $T(k) < ck$. By choosing c so that $\frac{1}{4}cn - bn - \frac{c}{2}$ is nonnegative (for example $c \geq 8b$ makes this term at least $bn - 2b$ which is nonnegative for $n \geq 2$), we conclude the proof, and have another proof of Theorem 5.23.

This kind of careful analysis arises when we are trying to get an estimate of the constant in a big-O bound (which we decided not to do in this case).

Important Concepts, Formulas, and Theorems

1. *Expected Running Time.* When the running time of an algorithm is different for different inputs of the same size, we can think of the running time of the algorithm as a random variable on the sample space of inputs and analyze the expected running time of the algorithm. This gives us a different understanding from studying just the worst case running time.

2. *Randomized Algorithm.* A*randomized algorithm* is an algorithm that depends on choosing something randomly.

3. *Random Number Generator.* A random number generator is a procedure that generates a number that appears to be chosen at random. Usually the designer of a random number generator tries to generate numbers that appear to be uniformly distributed.

4. *Insertion Sort.* A recursive description of insertion sort is that to sort $A(1 : n)$, first we sort $A(1 : n-1)$, and then we insert $A(n)$, by shifting the elements greater than $A(n)$ each one place to the right and then inserting the original value of $A(n)$ into the place we have opened up. If $n = 1$ we do nothing.

5. *Expected Running Time of Insertion Sort.* If $T(n)$ is the expected time to use insertion sort on a list of length n, then there are constants c and c' such that $T(n) \leq T(n-1) + cn$ and $T(n) \geq T(n-1) + c'n$. This means that $T(n) = \Theta(n^2)$. However the best case running time of insertion sort is $\Theta(n)$.

6. *Conditional Expected Value.* We define the *conditional expected value* of X given F by $E(X|F) = \sum_{x:x\in F} X(x)\frac{P(x)}{P(F)}$. This is equivalent to $E(X|F) = \sum_{i=1}^{k} x_i P(X = x_i|F)$.

7. *Randomized Selection Algorithm.* In the randomized selection algorithm to select the ith smallest element of a set A, we randomly choose a pivot element p in A, divide the rest of A into those elements that come before p (in the underlying order of A) and those that come after, put the pivot into the smaller set, and then recursively apply the randomized selection algorithm to find the appropriate element of the appropriate set.

8. *Running Time of Randomized Select.* Algorithm RandomSelect has expected running time $O(n)$. Because it does less computation than the deterministic selection algorithm, on the average a good implementation will run faster than a good implementation of the deterministic algorithm, but the worst case behavior is $\Theta(n^2)$.

9. *Quicksort.* Quicksort is a sorting algorithm in which we randomly choose a pivot element p in A, divide the rest of A into those elements that come before p (in the underlying order of A) and those that come after, put the pivot into the smaller set, and then recursively apply the Quicksort algorithm to sort each of the smaller sets, and concatenate the two sorted lists. We do nothing if a set has size one.

10. *Running Time of Quicksort.* Quicksort has expected running time $O(n \log n)$. It has worst case running time $\Theta(n^2)$. Good implementations of Quicksort have proved to be faster on the average than good implementations of other sorting algorithms.

Problems

1. Given an array A of length n (chosen from some set that has an underlying ordering), we can select the largest element of the array by starting out setting $L = A(1)$, and then comparing L to the remaining elements of the array one at a time, replacing L by $A(i)$ if $A(i)$ is larger than L. Assume that the elements of A are randomly chosen. For $i > 1$, let X_i be 1 if element i of A is larger than any element of $A(1 : i-1)$. Let $X_1 = 1$. Then what does $X_1 + X_2 + \cdots + X_n$ have to do with the number of times we assign a value to L? What is the expected number of times we assign a value to L?

2. Let $A(i : j)$ denote the array of items in positions i through j of the Array A. In selection sort, we use the method of Exercise 5.6-1 to find the largest element of the array A and its position k in the array, then we exchange the elements in position k and n of Array A, and we apply the same procedure recursively to the array $A(1 : n-1)$. (Actually we do this if $n > 1$; if $n = 1$ we do nothing.) What is the expected total number of times we assign a value to L in the algorithm selection sort?

3. Show that if H_n stands for the nth harmonic number, then

$$H_n + H_{n-1} + \cdots + H_2 = \Theta(n \log n).$$

4. In a card game, we remove the Jacks, Queens, Kings, and Aces from a deck of ordinary cards and shuffle them. You draw a card. If it is an Ace, you are paid a dollar and the game is repeated. If it is a Jack, you are paid two dollars and the game ends; if it is a Queen, you are paid three dollars and the game ends; and if it is a King, you are paid four dollars and the game ends. What is the maximum amount of money a rational person would pay to play this game?

5. Why does every solution to $T(n) \leq T(\frac{2}{3}n) + bn$ have $T(n) = O(n)$?

6. Show that if in Algorithm Random Select we remove the instruction

```
If H is empty
     put p in H,
```

then if $T(n)$ is the expected running time of the algorithm, there is a constant b such that $T(n)$ satisfies the recurrence

$$T(n) \leq \frac{2}{n-1} \sum_{k=n/2}^{n-1} T(k) + bn.$$

Show that if $T(n)$ satisfies this recurrence, then $T(n) = O(n)$.

7. Suppose you have a recurrence of the form

$$T(n) \leq T(a_n n) + T((1 - a_n)n) + bn \text{ if } n > 1,$$

where a_n is between $\frac{1}{5}$ and $\frac{4}{5}$. Show that all solutions to this recurrence are of the form $T(n) = O(n \log n)$.

8. Prove Theorem 5.22.

9. A tighter (up to constant factors) analysis of Quicksort is possible by using ideas very similar to those that we used for the randomized selection algorithm. More precisely, we use Theorem 5.6.1, similarly to the way we used it for select. Write down the recurrence you get when you do this. Show that this recurrence has solution $O(n \log n)$. In order to do this, you will probably want to prove that $T(n) \leq c_1 n \log n - c_2 n$ for some constants c_1 and c_2.

10. It is also possible to write a version of the randomized Selection algorithm analogous to Slower Quicksort. That is, when we pick out the random pivot element, we check if it is in the middle half and discard it if it is not. Write this modified selection algorithm, give a recurrence for its running time, and show that this recurrence has solution $O(n)$.

11. One idea that is often used in selection is that instead of choosing a random pivot element, we choose three random pivot elements and then use the median of these three as our pivot. What is the probability that a randomly chosen pivot element is in the middle half? What is the probability that the median of three randomly chosen pivot elements is in the middle half? Does this justify the choice of using the median of three as pivot?

12. Is the expected running time of Quicksort $\Omega(n \log n)$?

13. A random binary search tree on n keys is formed by first randomly ordering the keys, and then inserting them in that order. Explain why in at least half the random binary search trees, both subtrees of the root have between $\frac{1}{4}n$ and $\frac{3}{4}n$ keys. If $T(n)$ is the expected height of a random binary search tree on n keys, explain why $T(n) \leq \frac{1}{2}T(n) + \frac{1}{2}T(\frac{3}{4}n) + 1$. (Think about the **definition** of a binary tree. It has a root, and the root has two subtrees! What did we say about the possible sizes of those subtrees?) What is the expected height of a one node binary search tree? Show that the expected height of a random binary search tree is $O(\log n)$.

14. The expected time for an unsuccessful search in a random binary search tree on n keys (see Problem 13 for a definition) is the expected depth of a leaf node. Arguing as in Problem 13 and the second proof of Theorem 5.6.2, find a recurrence that gives an upper bound on the expected depth of a leaf node in a binary search tree and use it to find a big Oh upper bound on the expected depth of a leaf node.

15. The expected time for a successful search in a random binary search tree on n nodes (see problem 13 for a definition) is the expected depth of a node of the tree. With probability $\frac{1}{n}$ the node is the root, which has depth 0; otherwise the expected depth is one plus the expected depth of one of its subtrees. Argue as in Problem 13 and the first proof of Theorem 5.23 to show that if $T(n)$ is the expected depth of a node in a binary search tree, then $T(n) \leq \frac{n-1}{n}(\frac{1}{2}T(n) + \frac{1}{2}T(\frac{3}{4}n)) + 1$. What big Oh upper bound does this give you on the expected depth of a node in a random binary search tree on n nodes?

16. Consider the following code for searching an array A for the maximum item:

```
max = -∞
for i = 1 to n
    if (A[i] > max)
        max = A[i]
```

    If A initially consists of n nodes in a random order, what is the expected number of times that the line $max = A[i]$ is executed? (Hint: Let X_i be the number of times that $max = A[i]$ is executed in the ith iteration of the loop.)

17. You are a contestant in the game show "Let's make a Deal." In this game show, there are three curtains. Behind one of the curtains is a new car, and behind the other two are cans of spam. You get to pick one of the curtains. After you pick that curtain, the emcee, Monte Hall, who we assume knows where the car is, reveals what is behind one of the curtains that you did not pick, showing you some cans of spam. He then asks you if you would like to switch your choice of curtain. Should you switch? Why or why not? Please answer this question carefully. You have all the tools needed to answer it, but several math Ph.D.'s are on record (in Parade Magazine) giving the wrong answer.

5.7 Probability Distributions and Variance

Distributions of random variables

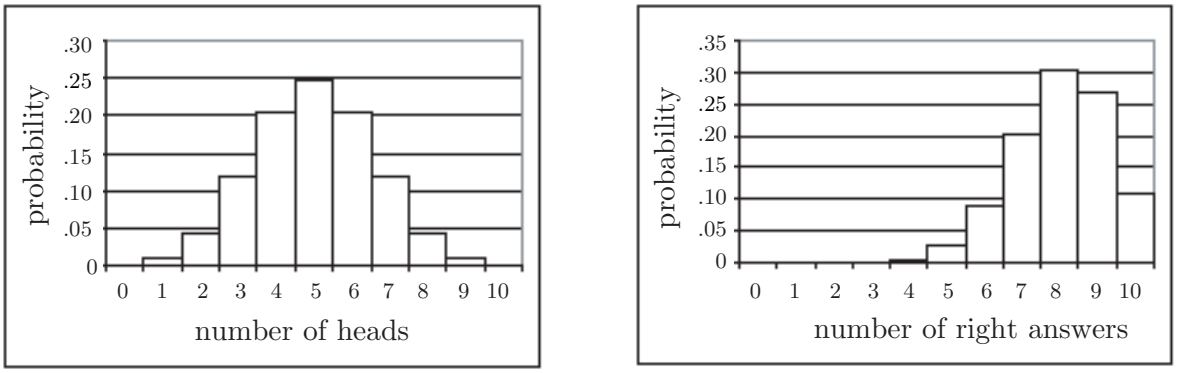
We have given meaning to the phrase expected value. For example, if we flip a coin 100 times, the expected number of heads is 50. But to what extent do we expect to see 50 heads. Would it be surprising to see 55, 60 or 65 heads instead? To answer this kind of question, we have to analyze how much we expect a random variable to deviate from its expected value. We will first see how to analyze graphically how the values of a random variable are distributed around its expected value. The *distribution function* D of a random variable X is the function on the values of X defined by

$$D(x) = P(X = x).$$

You probably recognize the distribution function from the role it played in the definition of expected value. The distribution function of the random variable X assigns to each value x of X the probability that X achieves that value. (Thus D is a function whose domain is the set off values of X.) When the values of X are integers, it is convenient to visualize the distribution function with a diagram called a *histogram*. In Figure 5.8 we show histograms for the distribution of the "number of heads" random variable for ten flips of a coin and the "number of right answers" random variable for someone taking a ten question test with probability .8 of getting a correct answer. What is a histogram? Those we have drawn are graphs which show for for each integer value x of X a rectangle of width 1, centered at x, whose height (and thus area) is proportional to the probability $P(X = x)$. Histograms can be drawn with non-unit width rectangles. When people draw a rectangle with a base ranging from $x = a$ to $x = b$, the area of the rectangle is the probability that X is between a and b.

The function $D(a, b) = P(a \leq X \leq b)$ is often called a *cumulative distribution function*. When sample spaces can be infinite, it doesn't always make sense to assign probability weights to individual members of our sample space, but cumulative distribution functions still make sense. Thus for infinite sample spaces, the treatment of probability is often based on random variables and their cumulative distribution functions. Histograms are a natural way to display information about the cumulative distribution function.

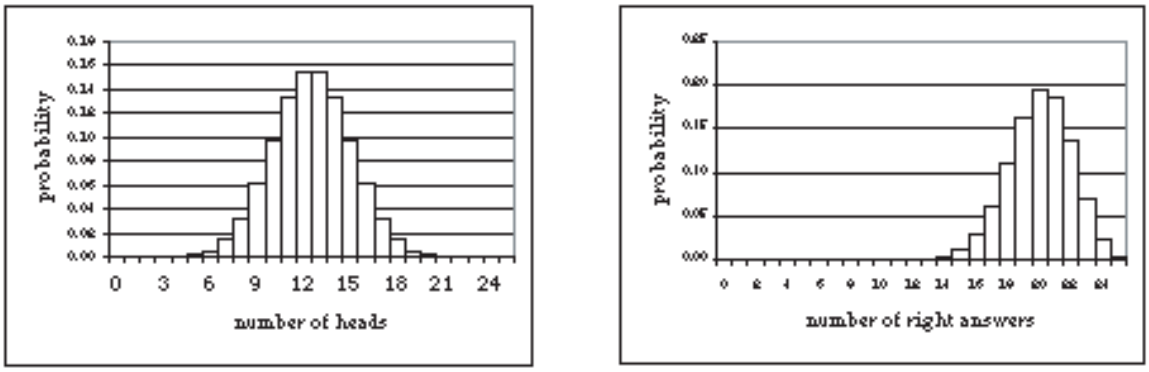
Figure 5.8: Two histograms.



From the histograms in Figure 5.8 you can see the difference between the two distributions.

You can also see that we can expect the number of heads to be somewhat near the expected number, though as few heads as 2 or as many as 8 are not out of the question. We see that the number of right answers tends to be clustered between 6 and ten, so in this case we can expect to be reasonably close to the expected value. With more coin flips or more questions, however, will the results spread out? Relatively speaking, should we expect to be closer to or farther from the expected value? In Figure 5.9 we show the results of 25 coin flips or 25 questions. The expected number of heads is 12.5. The histogram makes it clear that we can expect the vast majority of our results to have between 9 and 16 heads. Essentially all the results lie between 5 and 20 Thus the results are not spread as broadly (relatively speaking) as they were with just ten flips. Once again the test score histogram seems even more tightly packed around its expected value. Essentially all the scores lie between 14 and 25. While we can still tell the difference between the shapes of the histograms, they have become somewhat similar in appearance.

Figure 5.9: Histograms of 25 trials



In Figure 5.10 we have shown the thirty most relevant values for 100 flips of a coin and a 100 question test. Now the two histograms have almost the same shape, though the test histogram is still more tightly packed around its expected value. The number of heads has virtually no chance of deviating by more than 15 from its expected value, and the test score has almost no chance of deviating by more than 11 from the expected value. Thus the spread has only doubled, even though the number of trials has quadrupled. In both cases the curve formed by the tops of the rectangles seems quite similar to the bell shaped curve called the normal curve that arises in so many areas of science. In the test-taking curve, though, you can see a bit of difference between the lower left-hand side and the lower right hand side.

Since we needed about 30 values to see the most relevant probabilities for these curves, while we needed 15 values to see most of the relevant probabilities for independent trials with 25 items, we might predict that we would need only about 60 values to see essentially all the results in four hundred trials. As Figure 5.11 shows, this is indeed the case. The test taking distribution is still more tightly packed than the coin flipping distribution, but we have to examine it closely to find any asymmetry. These experiments are convincing, and they suggest that the spread of a distribution (for independent trials) grows as the square root of the number of trials, because each time we quadruple the number of elements, we double the spread. They also suggest there is some common kind of bell-shaped limiting distribution function for at least the distribution of successes in independent trials with two outcomes. However without a theoretical foundation we

Figure 5.10: One hundred independent trials

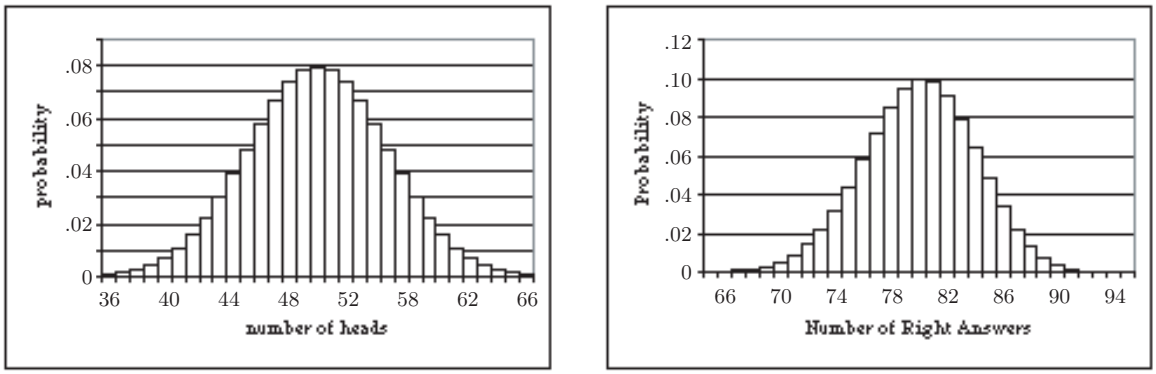
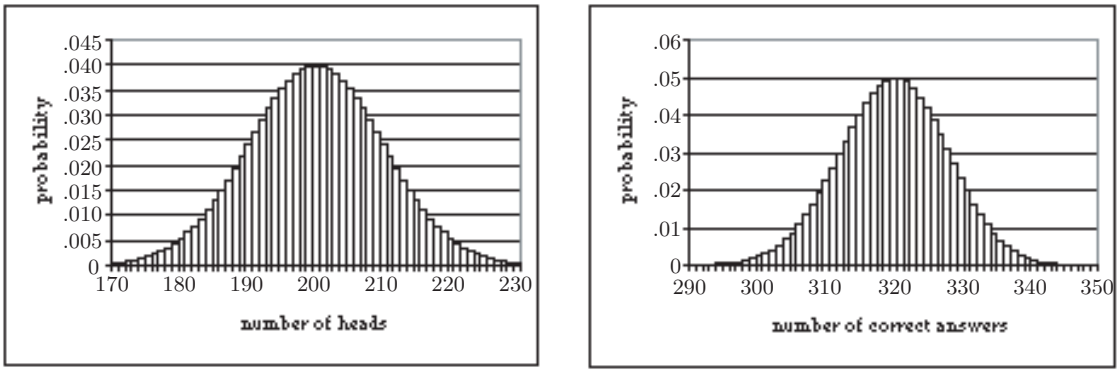


Figure 5.11: Four hundred independent trials



don't know how far the truth of our observations extends. Thus we seek an algebraic expression of our observations. This algebraic measure should somehow measure the difference between a random variable and its expected value.

Variance

Exercise 5.7-1 Suppose the X is the number of heads in four flips of a coin. Let Y be the random variable $X - 2$, the difference between X and its expected value. Compute $E(Y)$. Does it effectively measure how much we expect to see X deviate from its expected value? Compute $E(Y^2)$. Try repeating the process with X being the number of heads in ten flips of a coin and Y being $X - 5$.

Before answering these questions, we state a trivial, but useful, lemma (which appeared as Problem 9 in Section 5.4 of this chapter) and corollary showing that the expected value of an expectation is that expectation.

Lemma 5.25 *If X is a random variable that always takes on the value c, then $E(X) = c$.*

Proof: $E(X) = P(X = c) \cdot c = 1 \cdot c = c$. ■

We can think of a constant c as a random variable that always takes on the value c. When we do, we will just write $E(c)$ for the expected value of this random variable, in which case our lemma says that $E(c) = c$. This lemma has an important corollary.

Corollary 5.26 $E(E(X)) = E(X)$.

Proof: When we think of $E(X)$ as a random variable, it has a constant value, μ. By Lemma 5.25 $E(E(x)) = E(\mu) = \mu = E(x)$. ■

Returning to Exercise 5.7-1, we can use linearity of expectation and Corollary 5.26 to show that

$$E(X - E(X)) = E(X) - E(E(X)) = E(X) - E(X) = 0. \tag{5.42}$$

Thus this is not a particularly useful measure of how close a random variable is to its expectation. If a random variable is sometimes above its expectation and sometimes below, you would like these two differences to somehow add together, rather than cancel each other out. This suggests we try to convert the values of $X - E(X)$ to positive numbers in some way and then take the expectation of these positive numbers as our measure of spread. There are two natural ways to make numbers positive, taking their absolute value and squaring them. It turns our that to prove things about the spread of expected values, squaring is more useful. Could we have guessed that? Perhaps, since we see that the spread seems to grow with the square root, and the square root isn't related to the absolute value in the way it is related to the squaring function. On the other hand, as you saw in the example, computing expected values of these squares from what we know now is time consuming. A bit of theory will make it easier.

We define the **variance** $V(X)$ of a random variable X as the expected value of $E\left((X - E(X))^2\right)$. We can also express this as a sum over the individual elements of the sample space S and get that

$$V(X) = E(X - E(X))^2 = \sum_{s:s\in S} P(s)(X(s) - E(X))^2. \tag{5.43}$$

Now let's apply this definition and compute the variance in the number X of heads in four flips of a coin. We have

$$V(X) = (0-2)^2 \cdot \frac{1}{16} + (1-2)^2 \cdot \frac{1}{4} + (2-2)^2 \cdot \frac{3}{8} + (3-2)^2 \cdot \frac{1}{4} + (4-2)^2 \cdot \frac{1}{16} = 1.$$

Computing the variance for ten flips of a coin involves some very inconvenient arithmetic. It would be nice to have a computational technique that would save us from having to figure out large sums if we want to compute the variance for ten or even 100 or 400 flips of a coin to check our intuition about how the spread of a distribution grows. We saw before that the expected value of a sum of random variables is the sum of the expected values of the random variables. This was very useful in making computations.

Exercise 5.7-2 What is the variance for the number of heads in one flip of a coin? What is the sum of the variances for four independent trials of one flip of a coin?

Exercise 5.7-3 We have a nickel and quarter in a cup. We withdraw one coin. What is the expected amount of money we withdraw? What is the variance? We withdraw two coins, one after the other without replacement. What is the expected amount of money we withdraw? What is the variance? What is the expected amount of money and variance for the first draw? For the second draw?

Exercise 5.7-4 Compute the variance for the number of right answers when we answer one question with probability .8 of getting the right answer (note that the number of right answers is either 0 or 1, but the expected value need not be). Compute the variance for the number of right answers when we answer 5 questions with probability .8 of getting the right answer. Do you see a relationship?

In Exercise 5.7-2 we can compute the variance

$$V(X) = (0 - \frac{1}{2})^2 \cdot \frac{1}{2} + (1 - \frac{1}{2})^2 \cdot \frac{1}{2} = \frac{1}{4}.$$

Thus we see that the variance for one flip is 1/4 and sum of the variances for four flips is 1. In Exercise 5.7-4 we see that for one question the variance is

$$V(X) = .2(0 - .8)^2 + .8(1 - .8)^2 = .16$$

For five questions the variance is

$$4^2 \cdot (.2)^5 + 3^2 \cdot 5 \cdot (.2)^4 \cdot (.8) + 2^2 \cdot 10 \cdot (.2)^3 \cdot (.8)^2 + 1^2 \cdot 10 \cdot (.2)^2 \cdot (.8)^3 + 0^2 \cdot 5 \cdot (.2)^1 \cdot (.8)^4 + 1^2 \cdot (.8)^5 = .8.$$

The result is five times the variance for one question.

For Exercise 5.7-3 the expected amount of money for one draw is \$.15. The variance is

$$(.05 - .15)^2 \cdot .5 + (.25 - .15)^2 \cdot .5 = .01.$$

For removing both coins, one after the other, the expected amount of money is \$.30 and the variance is 0. Finally the expected value and variance on the first draw are \$.15 and .01 and the expected value and variance on the second draw are \$.15 and .01.

It would be nice if we had a simple method for computing variance by using a rule like "the expected value of a sum is the sum of the expected values." However Exercise 5.7-3 shows that the variance of a sum is not always the sum of the variances. On the other hand, Exercise 5.7-2 and Exercise 5.7-4 suggest such a result might be true for a sum of variances in independent trials processes. In fact slightly more is true. We say random variables X and Y are *independent* when the event that X has value x is independent of the event that Y has value y, regardless of the choice of x and y. For example, in n flips of a coin, the number of heads on flip i (which is 0 or 1) is independent of the number of heads on flip j. To show that the variance of a sum of independent random variables is the sum of their variances, we first need to show that the expected value of the product of two *independent* random variables is the product of their expected values.

Lemma 5.27 *If X and Y are independent random variables on a sample space S with values $x_1, x_2, \ldots, x_k$ and $y_1, y_2, \ldots, y_m$ respectively, then*

$$E(XY) = E(X)E(Y).$$

Proof: We prove the lemma by the following series of equalities. In going from (5.44) to (5.45), we use the fact that X and Y are independent; the rest of the equations follow from definitions and algebra.

$$\begin{aligned}
E(X)E(Y) &= \sum_{i=1}^{k} x_i P(X = x_i) \sum_{j=1}^{m} y_j P(Y = y_j) \\
&= \sum_{i=1}^{k} \sum_{j=1}^{m} x_i y_j P(X = x_i) P(y = y_j) \\
&= \sum_{z:\, z \text{is a value of } XY} z \sum_{(i,j):x_i y_j = z} P(X = x_i) P(Y = y_j) \qquad (5.44) \\
&= \sum_{z:\, z \text{is a value of } XY} z \sum_{(i,j):x_i y_j = z} P((X = x_i) \wedge (Y = y_j)) \qquad (5.45) \\
&= \sum_{z:\, z \text{is a value of } XY} z P(XY = z) \\
&= E(XY).
\end{aligned}$$

■

Theorem 5.28 *If X and Y are independent random variables then*

$$V(X + Y) = V(X) + V(Y).$$

Proof: Using the definitions, algebra and linearity of expectation we have

$$\begin{aligned}
V(X + Y) &= E\left((X + Y - E(X + Y))^2\right) \\
&= E\left((X - E(X) + Y - E(Y))^2\right) \\
&= E\left((X - E(X))^2 + 2(X - E(X))(Y - E(Y)) + (Y - E(Y))^2)\right) \\
&= E\left((X - E(X))^2\right) + 2E((X - E(X))(Y - E(Y))) + E\left((Y - E(Y))^2\right).
\end{aligned}$$

Now the first and last terms are just the definitions of $V(X)$ and $V(Y)$ respectively. Note also that if X and Y are independent and b and c are constants, then $X - b$ and $Y - c$ are independent (See Problem 8 at the end of this section.) Thus we can apply Lemma 5.27 to the middle term to obtain

$$= V(X) + 2E(X - E(X))E(Y - E(Y)) + V(Y).$$

Now we apply Equation 5.42 to the middle term to show that it is 0. This proves the theorem. ■

With this theorem, computing the variance for ten flips of a coin is easy; as usual we have the random variable X_i that is 1 or 0 depending on whether or not the coin comes up heads. We saw that the variance of X_i is $1/4$, so the variance for $X_1 + X_2 + \cdots + X_{10}$ is $10/4 = 2.5$.

Exercise 5.7-5 Find the variance for 100 flips of a coin and 400 flips of a coin.

Exercise 5.7-6 The variance in the previous problem grew by a factor of four when the number of trials grew by a factor of 4, while the spread we observed in our histograms grew by a factor of 2. Can you suggest a natural measure of spread that fixes this problem?

For Exercise 5.7-5 recall that the variance for one flip was 1/4. Therefore the variance for 100 flips is 25 and the variance for 400 flips is 100. Since this measure grows linearly with the size, we can take its square root to give a measure of spread that grows with the square root of the quiz size, as our observed "spread" did in the histograms. Taking the square root actually makes intuitive sense, because it "corrects" for the fact that we were measuring expected squared spread rather than expected spread.

The square root of the variance of a random variable is called the *standard deviation* of the random variable and is denoted by σ, or $\sigma(X)$ when there is a chance for confusion as to what random variable we are discussing. Thus the standard deviation for 100 flips is 5 and for 400 flips is 10. Notice that in both the 100 flip case and the 400 flip case, the "spread" we observed in the histogram was ± 3 standard deviations from the expected value. What about for 25 flips? For 25 flips the standard deviation will be 5/2, so ± 3 standard deviations from the expected value is a range of 15 points, again what we observed. For the test scores the variance is .16 for one question, so the standard deviation for 25 questions will be 2, giving us a range of 12 points. For 100 questions the standard deviation will be 4, and for 400 questions the standard deviation will be 8. Notice again how three standard deviations relate to the spread we see in the histograms.

Our observed relationship between the spread and the standard deviation is no accident. A consequence of a theorem of probability known as the central limit theorem is that the percentage of results within one standard deviation of the mean in a relatively large number of independent trials with two outcomes is about 68%; the percentage within two standard deviations of the mean is about 95.5%, and the percentage within three standard deviations of the mean is about 99.7%.

The central limit theorem tells us about the probability that the sum of independent random variables with the same distribution function is between two numbers. When the number of random variables we are adding is sufficiently large, the theorem tells us the approximate probability that the sum is between a and b standard deviations from its expected value. (For example if $a = -1.5$ and $b = 2$, the theorem tells us an approximate probability that the sum is between 1.5 standard deviations less than its expected value and 2 standard deviations more than its expected value.) This approximate value is $\frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx$.[8] The distribution given by that multiple of the integral is called the *normal distribution*. Since many of the things we observe in nature can be thought of as the outcome of multistage processes, and the quantities we measure are often the result of adding some quantity at each stage, the central limit theorem "explains" why we should expect to see normal distributions for so many of the things we do measure. While weights can be thought of as the sum of the weight change due to eating and exercise each week, say, this is not a natural interpretation for blood pressures. Thus while we shouldn't be particularly surprised that weights are normally distributed, we don't have the same basis for predicting that blood pressures would be normally distributed[9], even though they are!

Exercise 5.7-7 If we want to be 95% sure that the number of heads in n flips of a coin is within $\pm 1\%$ of the expected value, how big does n have to be?

Exercise 5.7-8 What is the variance and standard deviation for the number of right answers for someone taking a 100 question short answer test where each answer is graded

[8] Still more precisely, if we let μ be the expected value of the random variable X_i and σ be its standard deviation (all X_i have the same expected value and standard distribution since they have the same distribution) and scale the sum of our random variables by $Z = \frac{X_1 + X_2 + \ldots . X_n - n\mu}{\sigma\sqrt{n}}$, then the probability that $a \leq Z \leq b$ is $\frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx$.

[9] Actually, this is a matter of opinion. One might argue that blood pressures respond to a lot of little additive factors.

either correct or incorrect if the person knows 80% of the subject material for the test the test and answers correctly each question she knows? Should we be surprised if such a student scores 90 or above on the test?

Recall that for one flip of a coin the variance is $1/4$, so that for n flips it is $n/4$. Thus for n flips the standard deviation is $\sqrt{n}/2$. We expect that 95% of our outcomes will be within 2 standard deviations of the mean (people always round 95.5 to 95) so we are asking when two standard deviations are 1% of $n/2$. Thus we want an n such that $2\sqrt{n}/2 = .01(.5n)$, or such that $\sqrt{n} = 5 \cdot 10^{-3}n$, or $n = 25 \cdot 10^{-6}n^2$. This gives us $n = 10^6/25 = 40,000$.

For Exercise 5.7-8, the expected number of correct answers on any given question is .8. The variance for each answer is $.8(1-.8)^2 + .2(0-.8)^2 = .8 \cdot .04 + .2 \cdot .64 = .032 + .128 = .16$. Notice this is $.8 \cdot (1-.8)$. The total score is the sum of the random variables giving the number of points on each question, and assuming the questions are independent of each other, the variance of their sum is the sum of their variances, or 16. Thus the standard deviation is 4. Since 90% is 2.5 standard deviations above the expected value, the probability of getting that a score that far from the expected value is somewhere between .05 and .003 by the Central Limit Theorem. (In fact it is just a bit more than .01). Assuming that someone is just as likely to be 2.5 standard deviations below the expected score as above, which is not exactly right but close, we see that it is quite unlikely that someone who knows 80% of the material would score 90% or above on the test. Thus we should be surprised by such a score, and take the score as evidence that the student likely knows more than 80% of the material.

Coin flipping and test taking are two special cases of Bernoulli trials. With the same kind of computations we used for the test score random variable, we can prove the following.

Theorem 5.29 *In Bernoulli trials with probability p of success, the variance for one trial is $p(1-p)$ and for n trials is $np(1-p)$, so the standard deviation for n trials is $\sqrt{np(1-p)}$.*

Proof: You are asked to give the proof in Problem 7.■

Important Concepts, Formulas, and Theorems

1. *Histogram. Histograms* are graphs which show for for each integer value x of a random variable X a rectangle of width 1, centered at x, whose height (and thus area) is proportional to the probability $P(X = x)$. Histograms can be drawn with non-unit width rectangles. When people draw a rectangle with a base ranging from $x = a$ to $x = b$, the area of the rectangle is the probability that X is between a and b.

2. *Expected Value of a Constant.* If X is a random variable that always takes on the value c, then $E(X) = c$. In particular, $E(E(X)) = E(X)$.

3. *Variance.* We define the *variance* $V(X)$ of a random variable X as the expected value of $(X - E(X))^2$. We can also express this as a sum over the individual elements of the sample space S and get that $V(X) = E\left((X - E(X))^2\right) = \sum_{s:s\in S} P(s)(X(s) - E(X))^2$.

4. *Independent Random Variables.* We say random variables X and Y are *independent* when the event that X has value x is independent of the event that Y has value y, regardless of the choice of x and y.

5. *Expected Product of Independent Random Variables.* If X and Y are independent random variables on a sample space S, then $E(XY) = E(X)E(Y)$.

6. *Variance of Sum of Independent Random Variables.* If X and Y are independent random variables then $V(X + Y) = V(X) + V(Y)$.

7. *Standard deviation.* The square root of the variance of a random variable is called the *standard deviation* of the random variable and is denoted by σ, or $\sigma(X)$ when there is a chance for confusion as to what random variable we are discussing.

8. *Variance and Standard Deviation for Bernoulli Trials.* In Bernoulli trials with probability p of success, the variance for one trial is $p(1-p)$ and for n trials is $np(1-p)$, so the standard deviation for n trials is $\sqrt{np(1-p)}$.

9. *Central Limit Theorem.* The central limit theorem says that the sum of independent random variables with the same distribution function is approximated well as follows. The probability that the random variable is between a and b is an appropriately chosen multiple of $\int_a^b e^{-cx^2}\,dx$, for some constant c, when the number of random variables we are adding is sufficiently large. This implies that the probability that a sum of independent random variables is within one, two, or three standard deviations of its expected value is approximately .68, .955, and .997.

Problems

1. Suppose someone who knows 60% of the material covered in a chapter of a textbook is taking a five question objective (each answer is either right or wrong, not multiple choice or true-false) quiz. Let X be the random variable that for each possible quiz, gives the number of questions the student answers correctly. What is the expected value of the random variable $X - 3$? What is the expected value of $(X - 3)^2$? What is the variance of X?

2. In Problem 1 let X_i be the number of correct answers the student gets on question i, so that X_i is either zero or one. What is the expected value of X_i? What is the variance of X_i? How does the sum of the variances of X_1 through X_5 relate to the variance of X for Problem 1?

3. We have a dime and a fifty cent piece in a cup. We withdraw one coin. What is the expected amount of money we withdraw? What is the variance? Now we draw a second coin, without replacing the first. What is the expected amount of money we withdraw? What is the variance? Suppose instead we consider withdrawing two coins from the cup together. What is the expected amount of money we withdraw, and what is the variance? What does this example show about whether the variance of a sum of random variables is the sum of their variances.

4. If the quiz in Problem 1 has 100 questions, what is the expected number of right answers, the variance of the expected number of right answers, and the standard deviation of the number of right answers?

5. Estimate the probability that a person who knows 60% of the material gets a grade strictly between 50 and 70 in the test of Exercise 5.7-4

6. What is the variance in the number of right answers for someone who knows 80% of the material on which a 25 question quiz is based? What if the quiz has 100 questions? 400 questions? How can we "correct" these variances for the fact that the "spread" in the histogram for the number of right answers random variable only doubled when we multiplied the number of questions in a test by 4?

7. Prove Theorem 5.29.

8. Show that if X and Y are independent and b and c are constant, then $X - b$ and $Y - c$ are independent.

9. We have a nickel, dime and quarter in a cup. We withdraw two coins, first one and then the second, without replacement. What is the expected amount of money and variance for the first draw? For the second draw? For the sum of both draws?

10. Show that the variance for n independent trials with two outcomes and probability p of success is given by $np(1-p)$. What is the standard deviation? What are the corresponding values for the number of failures random variable?

11. What are the variance and standard deviation for the sum of the tops of n dice that we roll?

12. How many questions need to be on a short answer test for us to be 95% sure that someone who knows 80% of the course material gets a grade between 75% and 85%?

13. Is a score of 70% on a 100 question true-false test consistent with the hypothesis that the test taker was just guessing? What about a 10 question true-false test? (This is not a plug and chug problem; you have to come up with your own definition of "consistent with.")

14. Given a random variable X, how does the variance of cX relate to that of X?

15. Draw a graph of the equation $y = x(1 - x)$ for x between 0 and 1. What is the maximum value of y? Why does this show that the variance (see Problem 10 in this section) of the "number of successes" random variable for n independent trials is less than or equal to $n/4$?

16. This problem develops an important law of probability known as *Chebyshev's law*. Suppose we are given a real number $r > 0$ and we want to estimate the probability that the difference $|X(x) - E(X)|$ of a random variable from its expected value is more than r.

    (a) Let $S = \{x_1, x_2, \ldots, x_n\}$ be the sample space, and let $E = \{x_1, x_2, \ldots, x_k\}$ be the set of all x such that $|X(x) - E(X)| > r$. By using the formula that defines $V(X)$, show that
    $$V(X) > \sum_{i=1}^{k} P(x_i)r^2 = P(E)r^2$$

    (b) Show that the probability that $|X(x) - E(X)| \geq r$ is no more than $V(X)/r^2$. This is called *Chebyshev's law*.

17. Use Problem 15 of this section to show that in n independent trials with probability p of success,
    $$P\left(\left|\frac{\#\ \text{ of successes} - np}{n}\right| \geq r\right) \leq \frac{1}{4nr^2}$$

18. This problem derives an intuitive law of probability known as the *law of large numbers* from Chebyshev's law. Informally, the law of large numbers says if you repeat an experiment many times, the fraction of the time that an event occurs is very likely to be close to the probability of the event. In particular, we shall prove that for any positive number s, no matter how small, by making the number n independent trials in a sequence of independent trials large enough, we can make the probability that the number X of successes is between $np - ns$ and $np + ns$ as close to 1 as we choose. For example, we can make the probability that the number of successes is within 1% (or 0.1 per cent) of the expected number as close to 1 as we wish.

    (a) Show that the probability that $|X(x) - np| \geq sn$ is no more than $p(1-p)/s^2n$.

    (b) Explain why this means that we can make the probability that $X(x)$ is between $np - sn$ and $np + sn$ as close to 1 as we want by making n large.

19. On a true-false test, the score is often computed by subtracting the number of wrong answers from the number of right ones and converting that number to a percentage of the number of questions. What is the expected score on a true-false test graded this way of someone who knows 80% of the material in a course? How does this scheme change the standard deviation in comparison with an objective test? What must you do to the number of questions to be able to be a certain percent sure that someone who knows 80% gets a grade within 5 points of the expected percentage score?

20. Another way to bound the deviance from the expectation is known as Markov's inequality. This inequality says that if X is a random variable taking only non-negative values, then, for any $k \geq 1$,

$$P(X > kE(X)) \leq \frac{1}{k}.$$

    Prove this inequality.