Chapter 1

Probability

1.1 Getting Started

I begin with one of my favorite quotes from one my favorite sources.

> Predictions are tough, especially about the future.—Yogi Berra.

Probability theory is used by mathematicians, scientists and statisticians to quantify uncertainty about the future.

We begin with the notion of a **chance mechanism**. This is a two-word technical expression. It is very important that we use technical expressions exactly as they are defined. In every day life you may have several meanings for some of your favorite words, for example phat, but in this class technical expressions mean what they mean. Uniquely. In these notes the first occurrence of a technical expression/term will be in bold-faced type.

Both words in 'chance mechanism' (CM) are meaningful. The second word reminds us that the CM, when **operated**, produces an **outcome**. The first word reminds us that the outcome cannot be predicted *with certainty*.

Several examples will help.

1. CM: A coin is tossed. Outcome: The face that lands up, either heads or tails.
2. CM: A (six-sided) die is cast. Outcome: The face that lands up, either 1, 2, 3, 4, 5 or 6.
3. CM: A man with AB blood and a woman with AB blood have a child. Outcome: The blood type of the child, either A, B or AB.
4. CM: The next NFL season's Super Bowl game. Outcome: The winner of the game, which could be any one of the 32 NFL teams (well, perhaps not my childhood favorite, the Detroit Lions).

The next idea is the **sample space**, usually denoted by $\mathcal{S}$. The sample space is the collection of all possible outcomes of the CM. Below are the sample spaces for the CM's listed above.

1. CM: Coin. $\mathcal{S} = \{H, T\}$.

2. CM: Die. $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$.

3. CM: Blood. $\mathcal{S} = \{A, B, AB\}$.

4. CM: Super Bowl. $\mathcal{S} =$ A list of the 32 NFL teams.

An **event** is a collection of outcomes; that is, it is a subset of the sample space. Events are typically denoted by upper case letters, usually from the beginning of the alphabet. Below are some events for the CM's listed above.

1. CM: Coin. $A = \{H\}$, $B = \{T\}$.

2. CM: Die. $A = \{5, 6\}$, $B = \{1, 3, 5\}$.

3. CM: Blood. $C = \{A, B\}$.

4. CM: Super Bowl. $A = \{$Vikings, Packers, Bears, Lions$\}$.

Sometimes it is convenient to describe an event with words. As examples of this: For the die, event A can described as 'the outcome is larger than 4,' and event B can be described as 'the outcome is an odd integer.' For the Super Bowl, event A can described as 'the winner is from the NFC North Division.'

Here is where I am going with this: *Before* a CM is operated, nobody knows what the outcome will be. In particular, for any event A that is not the entire sample space, we don't know whether the outcome will be a member of A. *After* the CM is operated we can determine/see whether the actual outcome is a member of an event A; if it is, we say that the event A has **occurred**; if not, we say that the event A has **not occurred**. Below are some examples for our CM's above.

1. CM: Coin. If the coin lands heads, then event A has occurred and event B has not occurred.

2. CM: Die. If the die lands 5, both A and B have occurred. If the die lands 1 or 3, B has occurred, but A has not. If the die lands 6, A has occurred, but B has not. Finally, if the die lands 2 or 4, both A and B have not occurred.

3. CM: Blood. If the child has AB blood, then the even C has not occurred.

4. CM: Super Bowl. If the Packers win the Super Bowl, then the event A has occurred.

Before the CM is operated, the **probability** of the event A, denoted by $P(A)$, is a number that measures the *likelihood* that A will occur. This incredibly vague statement raises three questions that we will answer.

1. *How* are probabilities assigned to events?

2. What are the *rules* that these assignments obey?

3. If I say, for example, that $P(A) = 0.25$, what does this *mean*?

First, the assignment of probabilities to events *always* is based on *assumptions* about the operation of the world. As such, it is a *scientific*, not a *mathematical* exercise. There are always assumptions, whether they are expressed or tacit; implicit or explicit. My advice is to always do your best to be aware of any assumptions you make. (This is, I believe, good advice for outside the classroom too.)

The most popular assumption for a CM is the assumption of the **equally likely case (ELC)**. As the name suggests, in the ELC we assume that each possible outcome is equally likely to occur. Another way to say this is that it is impossible to find two outcomes such that one outcome is more likely to occur than the other. I will discuss the ELC for the four CM's we have been considering in this section.

1. CM: Coin. If I select an ordinary coin from my pocket and plan to toss it, I would assume that the two outcomes, heads and tails, are equally likely to occur. This seems to be a popular assumption in our culture b/c 'tossing a coin' is often used as a way to decide which of two persons/teams is allowed to make a choice. For example, football games typically begin with a coin toss and the winner gets to make a choice involving direction of attack or initial possession of the ball. Note, however, that I would *not* make this assumption w/o thinking about it. In particular, the path of a coin is governed by the laws of physics and presumably if I could always apply *exactly* the same forces to the coin it would always land the same way. I am an extremely minor friend of a famous person named Persi Diaconis. Persi has been a tenured professor at Stanford, Harvard, Cornell and Stanford again, and he was a recipient of a MacArthur Foundation 'no strings attached genius' fellowship a number of years ago. More relevant for this discussion is that while a teenager, Persi worked as a small acts magician. Thus, it is no surprise to learn that Persi has unusually good control of his hands and reportedly can make heads much more likely than tails when *he* tosses a coin. My willingness to assume that heads and tails are equally likely when *I* toss a coin reflects my belief about how coins are balanced and my limited ability to control my hands.

2. CM: Die. Again, if I take an ordinary die from a board game I am willing to assume that the six sides are equally likely to land facing up when I cast the die. Certainly, the casinos of Las Vegas believe that the ELC is reasonable for their dice b/c their payoffs in the game of *craps* could result in their losing large sums of money if the ELC does not apply. I own, however, two round cornered dice (ordinary dice have squared corners) which I will tell you about later in the notes. In particular, based on data I collected, we will conclude that the ELC is not reasonable for either of my round cornered dice.

3. CM: Blood. The three possible blood types for the child are *not* equally likely. There is a version of the ELC lurking in this problem and we will analyze it later.

4. CM: Super Bowl. I would certainly **never** assume the ELC for 'who wins the Super Bowl.' I would like to see my childhood favorite team win, but I believe they are much less likely to win than say . . . , well, just about any other team.

Please draw the following lessons from the above discussions. We should carefully consider how (we believe) the world operates and decide whether the ELC seems reasonable. These con-

siderations are a matter of science, not mathematics.

Let us suppose that we are willing to assume the ELC for a CM. What happens next?

If we assume the ELC, then we assign probabilities to events as follows. For any event A,

$$P(A) = \frac{\text{The number of outcomes in } A}{\text{The number of outcomes in } \mathcal{S}}.$$

Let's see how this formula works.

1. CM: Coin. The probability of obtaining heads is $1/2 = 0.50$. The probability of obtaining tails also is $1/2 = 0.50$.

2. CM: Die. The probability of obtaining the '1' is $1/6$. The probability of obtaining the '2' is $1/6$. In fact, the probability of obtaining any particular integer from 1, 2, ... 6, is $1/6$. They are equally likely and have the same probability of occurring.

   When considering this CM earlier, I defined the events $A = \{5, 6\}$, and $B = \{1, 3, 5\}$. We can now see that $P(A) = 2/6$ and $P(B) = 3/6$.

The obvious question is: If I am not willing to assume the ELC, how do I assign probabilities to events?

First, we need to discuss the nature of the sample space. For our few examples to date the sample space has consisted of a finite number of elements. There are actually three possibilities of interest to us for the nature of the sample space.

- The sample space can consist of a finite number of elements.
- The sample space can consist of a sequence of elements.
- The sample space can consist of an interval of numbers.

Let me give examples of the latter two of these possibilities.

1. CM: I cast a die until I obtain a 6. Outcome: The number of casts I perform.
   $\mathcal{S} = \{1, 2, 3, \ldots\}$.

2. CM: I hit a golf ball off a tee with my 9-iron. Outcome: The distance the ball travels before coming to rest, measured in yards. The sample space can be taken as the interval of numbers $(0, 300)$.

Let me add a few comments about this last CM. The key feature is that the outcome is a measurement. Measurements occur often in science, for example distance, weight (or mass), time, area, and volume are examples of measurements. It is true that we could treat measurements as counts simply by rounding off the measurement. For example, a person's weight is a measurement and it could be (and usually is in our culture) rounded to the nearest pound. It turns out, however, that the mathematics are actually easier to study for a measurement than for a count. Thus, instead of approximating a measurement as a count, the temptation is to approximate a count as a measurement. This issue will be revisited later in this course.

If you are a golfer or know much about golf, you will realize that nobody hits a golf ball 300 yards with a 9-iron. (Well, unless one is hitting down a very steep hill.) Thus, you might wonder why I put 300 yards as the upper limit in my sample space. As long as I put a number large enough to include all possibilities, it does not matter how large I make the upper bound. In fact, curious as it might seem, usually we don't worry too much about the upper or, indeed, lower bound. For example, if I select an adult male at random and measure his height, I would typically take the sample space to be all numbers larger than 0!

In terms of assigning probabilities to events, measurements require a very different method of study and they will be discussed later in this course.

Finite and countably infinite—the technical term used by mathematicians for a sample space that is a sequence—sample spaces are handled the same way. I will describe it now.

It will help if I begin with a new example of a finite sample space with a small number of elements. Let us once again consider the next NFL season, but now the outcome will be the team that wins the NFC North Division title. Note that there will be exactly one team that wins this title. If a single team has the best record, it is the winner. If two, three or all four teams tie for the best record, the NFL has a *tie-breaking* procedure that will determine a unique winner of the NFC North Division. With this understanding, the sample space consists of four elements: Bears, Lions, Packers and Vikings, denoted CB, DL, GBP and MV, respectively.

Now it is very important to remember that there is a distinction between mathematics and science. What I am about to show you is the *mathematically valid* way to assign probabilities to the events. The scientific validity will be discussed later.

First, let me state the obvious, namely that I am *not* willing to assume the ELC for this CM. For a finite sample space, if I am not willing to assume the ELC, I refer to the situation as the **general case**. Here is what I must do in the general case.

1. To every outcome in the sample space, assign a nonnegative number, called its probability. When summed over all outcomes in the sample space, these probabilities yield 1.

2. The probability of any event is equal to the sum of the probabilities of its outcomes.

This is actually quite simple, as I will now demonstrate.

1. I assign the following probabilities to the elements of the sample space: P(CB)=0.10, P(DL)=0.03, P(GBP)=0.40 and P(MV)=0.47.

2. There are sixteen possible events (including the sample space and the empty event) so I won't list all of them, but I will give you a few examples.

$$P(\text{GBP or MV}) = 0.40 + 0.47 = 0.87.$$

$$P(\text{CB or GBP or MV}) = 0.10 + 0.40 + 0.47 = 0.97.$$

$$P(\text{DL or CB}) = 0.03 + 0.10 = 0.13.$$

The assignments I have made, whether you agree with them or not, are mathematically valid. So is the following assignment, which a football fan would likely find absurd scientifically:

$$P(\text{CB}) = 0.01, P(\text{DL}) = 0.97, P(\text{GBP}) = 0.01 \text{ and } P(\text{MV}) = 0.01.$$

As you can no doubt imagine, if two persons have different assignments of probabilities to outcomes and if they like to gamble, they can agree on a mutually satisfactory wager.

We now turn to the second of our three questions about probability: What rules do they obey? B/c these rules are so important, we will number them. By the way, I will prove that Rules 1–3 are true for the ELC and the general case. They also can be proven to be true for our method of handling measurement outcomes which, as mentioned previously, will be covered later in this course.

By contrast, Rules 4–6 are logical consequences of Rules 1–3, which means that, for their proofs, we won't need to keep referring to how probabilities were initially assigned to events.

Rule 1. Called the rule of total probability. The probability of the sample space equals 1. I will prove this rule for both of our methods of assigning probabilities.

For the ELC, the probability of any event is the number of outcomes in the event divided by the number of outcomes in the sample space. Apply this definition with 'event' replaced by 'sample space' and the result is that the numerator and denominator coincide, making the quotient equal to 1.

For the general case, the probability of the sample space is the sum of the probabilities of all of its outcomes. By definition, these probabilities sum to 1.

The sample space is often called the **certain event** b/c the outcome of the operation of the CM *must* be a member of the sample space. (By definition, the sample space contains all possible outcomes.) Rule 1 states that certainty corresponds to a probability of one.

Rule 2. For any event A, $0 \leq P(A) \leq 1$.

Proof: For the equally likely case,

$$P(A) = \frac{\text{The number of outcomes in } A}{\text{The number of outcomes in the sample space}}.$$

The numerator is nonnegative and the denominator is positive; thus, the ratio cannot be negative. The numerator cannot exceed the denominator, so the ratio is at most one.

For the general case, $P(A)$ equals the sum of the probabilities of its outcomes. By definition, this sum cannot be negative and it cannot exceed one.

The consequences of Rule 2 are: Probabilities cannot be negative and no event can have more probability than the certain event.

To summarize, probability is a measure with extremes 0 and 1, where 0 corresponds to an impossible event and 1 to the certain event.

Before I can state and prove Rule 3, I need to remind you of some definitions for set operations.

If A and B are events, then (A or B) is the event that contains all elements that are in A and/or B; (A and B) is the event that contains all elements that are in both A and B. In the old days, what we call (A or B) was called the union of A and B and what we call (A and B) was called the intersection of A and B. Also, b/c many of us in the math sciences are lazy/frugal, (A and B) is typically written as AB.

Two events, A and B, are called disjoint or mutually exclusive if they have no elements in common; in other words, if AB is the empty set.

Rule 3. Called the addition rule for disjoint events. If A and B are disjoint events, then

$$P(A \textbf{ or } B) = P(A) + P(B).$$

Proof: For the ELC, the number of elements in (A or B) equals the number in A added to the number in B and the result follows. For the general case, adding the probabilities of the outcomes in A to the probabilities of the outcomes in B give us the total of the probabilities of the outcomes in (A or B).

Here is why Rule 3 is important. Unlike the first two rules, Rule 3 allows us to determine new probabilities from given probabilities without going back to first principles.

There are three more rules that we will need. I will use Rules 1–3 to prove these rules, so I won't need to keep referring to the ELC or the general case.

First, I need to remind you of another definition from sets. If A is any event, then its **complement**, denoted A^c, is the event which consists of all elements that are not in A.

Rule 4. The rule of complements.

$$P(A^c) = 1 - P(A).$$

Proof: A and A^c are disjoint events whose union is the sample space. Thus, by Rule 1, $1 = P(\mathcal{S}) = P(A \text{ or } A^c) = P(A) + P(A^c)$, by Rule 3, and the result follows.

Like Rule 3, Rule 4 allows us to calculate new probabilities from ones we already know.

If we have two events, A and B, we say that A is a **subset** of B if and only if every every element of A is in B. (B might have additional elements that are not in A.)

Rule 5. The subset rule. If A is a subset of B, then

$$P(A) \leq P(B).$$

Proof: B is equal to the union of A and A^cB, two disjoint sets. (It might help if you draw a picture.) Thus,

$$P(B) = P(A) + P(A^cB) \geq P(A),$$

b/c by Rule 2 all probabilities are nonnegative.

Rule 5 is important b/c it shows that more likely means larger probability. Event B is clearly more likely to occur than A b/c A occurring implies that B must occur.

Rule 6, our last rule, is a generalization of Rule 3 to events that are not disjoint.

Rule 6. The general addition rule for probabilities. For any events A and B,

$$P(A \textbf{ or } B) = P(A) + P(B) - P(AB).$$

Proof: It will definitely help if you draw a picture. The event (A or B) is the union of the following three disjoint events: AB^c, AB, and A^cB. Thus, by Rule 3,

$$P(A \text{ or } B) = P(AB^c) + P(AB) + P(A^cB) =$$

$$P(AB^c) + P(AB) + P(A^cB) + P(AB) - P(AB).$$

Now, referring to your picture,

$$P(A) = P(AB^c) + P(AB), \text{ and } P(B) = P(A^cB) + P(AB).$$

The result follows.

1.2 Independent, Identically Distributed Trials

Above we considered the operation of a CM. Many, but not all, CMs can be operated more than once. For example, a coin can be tossed or a die cast many times. By contrast, the next NFL season will operate only once.

In this section we consider repeated operations of a CM.

Let us return to the 'Blood type' CM of Section 1. Previously, I described the situation as follows: A man with AB blood and a woman with AB blood will have a child. The outcome is the child's blood type. The sample space consists of A, B and AB. I stated that these three outcomes are not equally likely, but that the ELC is lurking in this problem. We get the ELC by viewing the problem somewhat differently, namely as two operations of a CM.

The first operation is the selection of the allele that Dad gives to the child. The second operation is the selection of the allele that Mom gives to the child. For each operation, the possible outcomes are A and B and it seems reasonable to assume that these are equally likely. Consider the following display of the possibilities for the child's blood type.

	Allele from Mom	
Allele from Dad	A	B
A	A	AB
B	AB	B

I am willing to make the following assumption.

- The allele contributed by Dad (Mom) has no influence on the allele contributed by Mom (Dad).

Based on this assumption, and the earlier assumption of the ELC for each operation of the CM, I conclude that the four entries in the cells of the table above are equally likely. As a result, we have the following probabilities for the blood type of the child: $P(A) = P(B) = 0.25$ and $P(AB) = 0.50$.

Here is another example. I cast a die twice and I am willing to make the following assumptions.

Table 1.1: All Possible Outcomes For Casting a Pair of Dice.

Number from first cast	Number from second cast					
	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Table 1.2: The Probability Distribution of the Outcome Obtained When Casting a Balanced Die.

Value	Probability
1	$1/6$
2	$1/6$
3	$1/6$
4	$1/6$
5	$1/6$
6	$1/6$
Total	1

- The number obtained on the first cast is equally likely to be 1, 2, 3, 4, 5 or 6.
- The number obtained on the second cast is equally likely to be 1, 2, 3, 4, 5 or 6.
- The number obtained on the first (second) cast has no influence on the number obtained on the second (first) cast. We summarize this by saying that the outcomes on the two casts are **(statistically) independent**.

The 36 possible ordered results of the two casts are displayed in Table 1.1, where, for example, $(5, 3)$ means that the first die landed 5 and the second die landed 3. This is different from $(3, 5)$. Just like in the blood type example, b/c of my assumptions, I conclude that these 36 possibilities are equally likely. We will do a number of calculations now.

For ease of presentation, define X_1 to be the number obtained on the first cast of the die and let X_2 denote the number obtained on the second cast of the die.

We call X_1 and X_2 **random variables**, which means that to each possible outcome of the CM they assign a number. Every random variable has a **probability distribution** which is simply a listing of its possible values along with the probability of each value. Note that X_1 and X_2 have the same probability distribution; a fact we describe by saying that they are **identically distributed**, Table 1.2 presents the common probability distribution for X_1 and X_2. As we have seen, X_1 and

X_2 each has its own probability distribution (which happens to be the same). It is also useful to talk about their **joint probability distribution** which is concerned with how they interact. I will illustrate this idea with a number of computations.

$$P(X_1 = 3 \text{ and } X_2 = 4) = 1/36,$$

b/c, by inspection exactly one of the 36 ordered pairs in the earlier table have a 3 in the first position and a 4 in the second position.

Before we proceed, I want to invoke my laziness again. It is too much bother to type, say,

$$P(X_1 = 3 \text{ and } X_2 = 4).$$

It is much easier to type,

$$P(X_1 = 3, X_2 = 4).$$

In fact, provided it is not confusing, it is easier still to type simply $P(3, 4)$. To summarize: a comma within a probability statement represents the word 'and.'

For future reference, note that

$$P(X_1 = 3, X_2 = 4) = 1/36, \text{ as does } P(X_1 = 3)P(X_2 = 4).$$

In words, the word 'and' within a probability statement tells us to multiply.

Here is another example.

$$P(X_1 \leq 4, X_2 \geq 4) = 12/36,$$

b/c, as you can see from the table below, exactly 12 of the 36 pairs have the required property.

	X_2					
X_1	1	2	3	4	5	6
1				X	X	X
2				X	X	X
3				X	X	X
4				X	X	X
5						
6						

Note again that

$P(X_1 \leq 4, X_2 \geq 4) = 12/36$ gives the same answer as $P(X_1 \leq 4)P(X_2 \geq 4) = (4/6)(3/6) = 12/36$.

This last property is called **the multiplication rule for independent random variables**. It is a very important result. It says that if we have two random variables that are independent, then we can compute joint probabilities by using individual probabilities. In simpler words, if we want to know the probability of X_1 doing something **and** X_2 doing something, then we can calculate two individual probabilities, one for X_1 and one for X_2 and then take the **product** of these two individual probabilities. As we shall see repeatedly in this class, the multiplication rule for independent random variables is a great labor saving device.

The above ideas for two casts of a die can be extended to any number of casts of a die. In particular, define

- X_1 to be the number obtained on the first cast of the die;
- X_2 to be the number obtained on the second cast of the die;
- X_3 to be the number obtained on the third cast of the die;
- and so on, in general, X_k is the number obtained on the 'kth' cast of the die.

If we assume that the casts are independent–that is, that no outcome has any influence on another outcome–then we can use the multiplication rule to calculate probabilities. Some examples are below.

You might be familiar with the popular dice game Yahtzee. In this game, a player casts five dice. If all dice show the same number, then the player has achieved a Yahtzee. One of the first things you learn upon playing the game Yahtzee is that the event Yahtzee occurs only rarely. We will calculate its probability.

Let's find the probability of throwing five 1's when casting five dice. We write this as:

$$P(1, 1, 1, 1, 1) = (1/6)^5.$$

Now, the probability of a Yahtzee is:

$$P(Y_1 \text{ or } Y_2 \text{ or } Y_3 \text{ or } Y_4 \text{ or } Y_5 \text{ or } Y_6),$$

where 'Y_k' means all five dice land with the side 'k' facing up; in words, 'Y_k' means a Yahtzee on the number 'k.' Clearly all of the 'Y_k' have the same probability. Thus, by Rule 3, the probability of a Yahtzee is:

$$6(1/6)^5 = 1/1296 = 0.000772.$$

There is a slicker way to calculate the probability of a Yahtzee. Imagine that you cast the dice one-at-a-time. (Don't play Yahtzee this way; it will *really* annoy your friends.) No matter how the first die lands, you can still get a Yahtzee. (An ESPN anchor might announce, 'Ralph is on pace for a Yahtzee!') To obtain a Yahtzee, your last four dice must match your first one. The probability of this event is $(1/6)^4 = 0.000772$, as above.

Here is our general definition of **independent and identically distributed trials**, abbreviated **i.i.d.**:

> **Random variables $X_1, X_2, X_3, \ldots$ all have the same probability distribution *and* these random variables are independent.**

But how do we know we have independent random variables? This question is a bit tricky. Often times, we simply assume it to be true. This is indeed what I did above when I assumed that the outcome of the first cast of the die has no influence on the outcome of the second cast. In other problems, however, we will need to work from first principles to determine whether or not two random variables are independent. Also, when we try to apply these ideas to scientific problems that are more complex than tossing coins or casting dice, we will need to give careful consideration to whether or not independence makes sense scientifically. Finally, we will learn how to use data to investigate whether the assumption of independence is reasonable. (See Chapter 6.)

Table 1.3: 10,000 (100,000) Simulated Casts of a Fair Die.

	10,000 Casts				100,000 Casts		
Value	Freq.	Rel. Freq.	Prob.	Absolute Difference	Rel. Freq.	Prob.	Absolute Difference
1	1,681	0.1681	0.1667	0.0014	0.1655	0.1667	0.0012
2	1,675	0.1675	0.1667	0.0008	0.1682	0.1667	0.0015
3	1,676	0.1676	0.1667	0.0009	0.1644	0.1667	0.0023
4	1,693	0.1693	0.1667	0.0026	0.1669	0.1667	0.0002
5	1,674	0.1674	0.1667	0.0007	0.1676	0.1667	0.0009
6	1,601	0.1601	0.1667	0.0066	0.1674	0.1667	0.0007
Total	10,000	1.0000	1.0002		1.0000	1.0002	

I am ready to answer the third of our three questions about probability, posed long ago on page 2. Namely, if I determine/state $P(A) = 0.25$, what does this mean?

There is a famous result in probability theory that answers this question. Sort of. In a limited situation. It is called the **Law of Large Numbers (LLN)**. I will try to explain it.

Let X_1 be a random variable and let A be some event whose occurrence is determined by the value of X_1. Let $X_1, X_2, X_3, \ldots X_n$ be independent and identically distributed trials. Clearly, for each of $X_2, X_3, \ldots X_n$ we can determine whether or not the event A occurs. Then:

1. Count the number of times that A occurs in the first n trials.

2. Divide the frequency you obtained in step 1 by n, to obtain the relative frequency of occurrence of event A in n trials.

The LLN states that in the limit, as n becomes larger without bound, the relative frequency of occurrence of event A converges to $P(A)$.

Here is an example. I programmed my computer to **simulate** 10,000 independent trials for casting a balanced die (ELC). Then I had my computer repeat the process, but with 100,000 independent trials. The results are summarized in Table 1.3. Look at the results for 10,000 casts first; these are in the five columns to the left of the vertical line segment in the table. We see that the simulated frequencies ranged from a low of 1,601 for the outcome '6' to a high of 1,693 for the outcome '4.' Thus, obviously, the relative frequencies range from 0.1601 to 0.1693. These relative frequencies are all 'close' to the probability of each outcome: 0.1667. This last statement is supported by the values in the column 'Absolute Difference' which lists the absolute values of relative frequency minus probability. The largest discrepancy (absolute difference) is 0.0066 for the outcome '6;' three of the discrepancies are smaller than 0.0010. Thus, the LLN seems to be 'working;' for a large value of n, in this case 10,000, the relative frequencies are close to the probabilities. Well, if you agree with my notion of close.

With 100,000 casts (to the right of the vertical line segment in the table) the largest discrepancy is 0.0023 and, again, there are three that are smaller than 0.0010. Generally speaking, this table

shows that the LLN is better 'overall' for $n = 100{,}000$ than it is is for $n = 10{,}000$, but even this statement is open to debate. For example, in our table the relative frequencies of 2, 3 and 5 are all closer to the probability for $n = 10{,}000$ than they are for $n = 100{,}000$. Learning about the usefulness and oddities of approximations are an important part of this course.

For any event larger than a single outcome, the relative frequency of the event will be the sum of the appropriate relative frequencies, which, from the table above, will be close to its probability. For example, consider the event $A = \{1, 2, 3\}$. Given the ELC, $P(A) = 3/6 = 0.5000$. From the table above, the relative frequency of A for $n = 10{,}000$ is: $0.1681 + 0.1675 + 0.1676 = 0.5032$.

To summarize, when we have independent and identically distributed trials, then the probability of an event is approximately equal to its long-run-relative frequency. This result is used in 'both directions.' If we know the probability, we can predict the long-run-relative frequency of occurrence. If, however, we do not know the probability, we can approximate it by performing a large computer simulation and calculating the relative frequency of occurrence. This latter use is much more important to us and we will use it many times in this course.

One of the main users of the 'first application' of the LLN are gambling casinos. I will give a brief example of this.

An American roulette wheel has 38 slots, each slot with a number and a color. For this example, I will focus on the color. Two slots are colored green, 18 are red and 18 are black. Red is a popular bet and the casino pays 'even money' to a winner.

If we assume that the ELC is appropriate for the roulette wheel, then the probability that a red bet wins is $18/38 = 0.4737$. But a gambler is primarily concerned with his/her relative frequency of winning. Suppose that one gambler places a very large number, n, of one dollar bets on red. By the LLN, the relative frequency of winning bets will be very close to 0.4737 and the relative frequency of losing bets will be very close to $1 - 0.4737 = 0.5263$. In simpler terms, in the long run, for every \$100 bet on red, the casino pays out $2(47.37) = 94.74$ dollars, for a net profit of \$5.26 for every \$100 bet.

As a side note, when a person goes to a casino, he/she can see that every table game has a range of allowable bets. For example, there might be a roulette wheel that states that the minimum bet allowed is \$5 and the maximum is \$500. Well, a regular person likely pays no attention to the maximum, but it is very important to the casino. As a silly and extreme example, suppose Bill Gates walks into a casino and wants to place a \$50 billion bet on red. No casino could/would accept the bet. Why? And, of course, I have seen no evidence that Mr. Gates would want to place such a bet either.

1.3 Sums of i.i.d. Random Variables

This section provides you with practice on computing probabilities for i.i.d. random variables and will illustrate why approximations are so important.

Refer to Table 1.1. As earlier, define X_1 (X_2) to be the number that will be obtained on the first (second) cast. Define $X = X_1 + X_2$; in words, X is the sum of the numbers on the two dice. We will now learn how to obtain the probability distribution of X; by the way, the probability distribution of X is usually called its **sampling distribution**. The obvious question is: Why do

statisticians adopt two expressions for the same thing? Answer: The term sampling distribution is reserved for random variables, like X, that are functions of two or more (in the present case two) other random variables, while probability distribution implies that we have a random variable that is not such a function. Thus, when a statistician hears/reads probability/sampling distribution he/she knows a bit more about what is going on. We will see other examples of this idea later.

Anyways, a sampling distribution consists of two sets of numbers: a listing of all possible values of the random variable and a listing of the probabilities of the possible values. Typically, the first of these lists is far easier to determine.

In the present case, clearly the possible value of X are: 2, 3, 4, ... 12. Finding the probabilities is not difficult, but it is time consuming and tedious. The ingredients we need are: determination, Table 1.1 and the addition and multiplication rules. I will determine a few of the probabilities and then, when my determination flags, I will tell you the rest of them.

I begin with $P(X = 2)$. In lecture, I have told you the story of the mathematician in the kitchen and it applies now. We write the event of interest, $(X = 2)$, as $(1, 1)$ which means, recall, that the first and second dice both landed '1.' Now, we can use the multiplication rule:

$$P(1, 1) = P(X_1 = 1)P(X_2 = 1) = (1/6)(1/6) = 1/36.$$

We could easily change $1/36$ to a decimal, $1/36 = 0.0278$, but b/c this example is primarily for illustration, we won't bother. Next, we note that the event $(X = 3)$ is the same as the event $[(1, 2)$ or $(2, 1)]$. Thus, $P(X = 3)$ equals

$$P((1, 2) \text{ or } (2, 1)) = P(1, 2) + P(2, 1) = 1/36 + 1/36 = 2/36.$$

Again, we could simplify $2/36$ to $1/18$ or write it as a decimal, 0.0556, but we won't bother.

Next, the event $(X = 4)$ is the same as the event $[(1, 3)$ or $(2, 2)$ or $(3, 1)]$. Thus, $P(X = 4)$ equals

$$P((1, 3) \text{ or } (2, 2) \text{ or } (3, 1)) = P(1, 3) + P(2, 2) + P(3, 1) = 3/36.$$

Continuing in this way (my determination has flagged), we get the entire sampling distribution for X, given below.

x :	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$:	$1/36$	$2/36$	$3/36$	$4/36$	$5/36$	$6/36$	$5/36$	$4/36$	$3/36$	$2/36$	$1/36$

In principle, the above method can be extended from two casts of the die to any number, n, of casts. But the method is extremely tedious and time-consuming, even with the aid of a computer. Here is a quick (?) example. Suppose we want to cast the die $n = 5$ times and we want to determine $P(X = 9)$, where X is the sum of the five numbers obtained.

Well, first, I won't try to draw a picture like the one we have in Table 1.1 b/c it would be too tedious and difficult. But here is the important point: with five casts of the die, each of the $6^5 = 7776$ possible five-tuples are equally likely to occur. (Five-tuple is just a generalization of the words pair, e.g. (1,3) and triple, e.g. (1,3,2); in general, statisticians talk about n-tuples for an ordered list of n numbers.) Thus, we simply need to count how many of these five-tuples yield a sum of 9. To do this, we list possibilities. This listing goes much better if we are clever.

Table 1.4: Results from a Computer Simulation with 10,000 Runs for the Value of X, the Sum of the Numbers on Five Casts of a Fair Die.

Value:	5	6	7	8	9	10	11	12	13	14	15	16	17
Freq:	1	2	15	38	78	175	250	384	507	699	804	945	1004
Value:	18	19	20	21	22	23	24	25	26	27	28	29	30
Freq:	1012	978	854	720	524	407	286	169	89	28	19	10	2

First, I note that 9 is a pretty small total to obtain when one performs five casts. So, we begin by considering lots of 1's in the five-tuples. All 1's will give us a total of 5, which is no good. Four 1's will work if they are matched with a 5, such as (1,1,1,1,5). There are **5** such five-tuples, one for each choice of the position of the 5. The next possibility is to have three 1's. Three 1's can lead to a total of 9 if they are matched with: a 2 and a 4; or two 3's. There are **20** five-tuples that arrange 1,1,1,2,4 and **10** five-tuples that arrange 1,1,1,3,3. Two 1's can lead to a total of 9 if they are matched with 2,2,3. There are **30** five-tuples that arrange 1,1,2,2,3. Finally, one 1 can lead to a total of 9 if it is matched with four 2's; there are **5** such five-tuples.

If we sum the counts in bold-faced type in the previous paragraph, we find that there are **70** five-tuples that will yield a total of 9 on the five casts. Thus,

$$P(X = 9) = 70/7776 = 0.0090.$$

I now will show you a way to approximate this $P(X = 9)$ and similar probabilities.

Consider the CM: Perform $n = 5$ i.i.d. trials of casting a balanced die; compute the sum of the five numbers obtained, X. I programmed my computer to operate this CM 10,000 times. I learned a great deal from this **computer simulation**; much more than just an approximation to $P(X = 9)$. Thus, I will present the entire results of it in Table 1.4.

First, note that the total 9 occurred on 78 runs; thus, the computer simulation approximation of $P(X = 9)$ is 0.0078, the relative frequency of occurrence of 9. Recall, that we determined the actual probability to be 0.0090. In my opinion (feel free to disagree) 0.0078 is a good approximation of 0.0090.

Usually in practice, we would *not* know the true probability of $P(X = 9)$; we would simply have its approximation, in this case 0.0078. A natural question is: How close is the approximation to the truth? Well, obviously, we can give a definitive answer to this question *only if* we know the truth; knowing the truth I can state, "The approximation is too small by 0.0012." Not knowing the truth, below is the best we can do.

Denote our approximation, which is a relative frequency, by $\hat{r}$. Denote the truth, the actual probability, by r. Let m denote the number of runs in our computer simulation; in the present example, m equals 10,000. (Remember that if we don't like the answer we get below we can improve it by increasing m.) Calculate the interval:

$$\hat{r} \pm 3\sqrt{\hat{r}(1-\hat{r})/m}.$$

We can be *pretty certain* that the true probability, r, is in this interval. In Chapter 2 the notion of *pretty certain* will be made more precise. Let's see how this interval works.

In our example, $\hat{r} = 0.0078$ and the interval is

$$0.0078 \pm 3\sqrt{0.0078(0.9922)/10000} = 0.0078 \pm 0.0026 = [0.0052, 0.0104].$$

In this example, b/c we know that $r = 0.0090$ we know that the interval is correct; i.e. it contains r.

As a further example, I supplemented my original computer simulation of 10,000 runs with an additional 30,000 runs, bringing my total to m equals 40,000 runs. The total 9 occurred 342 times in these 40,000 runs, giving a relative frequency of $\hat{r} = 342/40000 = 0.00855$ and an interval of

$$0.00855 \pm 3\sqrt{0.00855(0.99145)/40000} = 0.00855 \pm 0.00138 = [0.00717, 0.01093].$$

This interval is correct b/c it contains $r = 0.0090$. The result of a four-fold increase in m is to make the interval, roughly, one-half as wide.

The lesson to be learned here: We can use a computer simulation to approximate r. We can always get a more precise approximation by increasing the number of runs in the computer simulation.

I will now give you another example of computing probabilities for a sum. Assume we have i.i.d. trials with the following probability distribution: the possible values are 0, 1 and 2, with probabilities 0.5, 0.3 and 0.2, respectively. This is similar to our die example, but easier b/c there are fewer possible values (3 versus 6), but more difficult b/c we no longer have the ELC.

Let X be the sum of $n = 3$ trials from this probability distribution. We will find the sampling distribution of X.

The possible values of X are 0, 1, 2, 3, 4, 5 and 6. I will calculate one of the probabilities for you. To find $P(X = 4)$ we note that a total of 4 can occur by: 0,2,2 or 1,1,2.

$$P(0, 2, 2) = 0.5(0.2)(0.2) = 0.020.$$

Similarly, $P(2, 0, 2) = P(2, 2, 0) = 0.020$. Next,

$$P(1, 1, 2) = 0.3(0.3)(0.2) = 0.018.$$

Similarly, $P(1, 2, 1) = P(2, 1, 1) = 0.018$. Adding all of these guys, we find that $P(X = 4) = 0.114$. The entire sampling distribution of X is given in the following table.

x :	0	1	2	3	4	5	6
$P(X = x)$:	0.125	0.225	0.285	0.207	0.114	0.036	0.008