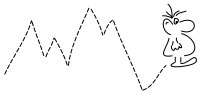
The Science and Technology of Decision−Making

David Bernstein
Princeton University

Chapter 6

Getting Promoted, the Stock Market, and the Drunken Stupor



ONE DAY in 1988 or 1989, while I was working on my Ph.D. Dissertation, I got rather frustrated. So frustrated, in fact, that I decided that there was no way that I was ever going to finish. When I got back to my apartment, there in the mailbox was the Publisher's Clearinghouse Sweepstakes™ envelope. I took it, went upstairs, got down on the floor (where I do most of my best work), and started filling it out. There I was – carefully answering all of the questions, finding all of the stickers, and putting them in all of the right places – when my wife came in. She asked what I was doing and I calmly told her "I've decided to change my career. I think that I can earn more per hour and be happier filling out sweepstakes entries full-time". My wife knew better than to question my sanity.

You've thought about it too, haven't you? It seems like you get about a million of those things. If you were to fill them all out, you'd almost have to win, wouldn't you?[1] Well, that's exactly the kind of question we're going to consider in this chapter. We're going to find out whether it was worth it to fill out that sweepstakes form, whether it makes sense to play the lottery, and a whole bunch of other things.

[1] By the way, I've had other good ideas like this. For example, have you ever seen those advertisements "This week only! Save $1000 on SuperDeluxe Refrigerators." I've always thought that I should go to the store and buy 1000 refrigerators. In total, I'd save $1,000,000 and I'd be rich. My High School girlfriend, Karen, broke-up with me when I told her about this idea. When I mentioned it to my wife, shortly **after** we got married, she took away my credit card.

6.1 You're Probably Thinking I'm Crazy

In order to do so, we need to talk about *probability*. "What is probability?", you ask. "Good question! What do you think probability is?", I respond.[2]

It turns out, that it's relatively difficult to explain what we mean when we say that something happens with a particular probability. For example, suppose I ask you what the probability is that a fair coin comes up 'heads' when you toss it. You will, undoubtedly, say $1/2$.[3] What do you mean when you say that? You don't know exactly, do you? Our conversation about it might go something like this:

Me: What do you mean when you say that the probability of a head is $1/2$?

You: If we were to toss a coin a whole bunch of times, it will come up heads half the time.

Me: Suppose we tossed it 100 times and it came up tails each time. Would that prove that you were wrong?

You: I'm not wrong. That would just be a fluke.

Me: What do you mean by "a fluke"?

You: It doesn't matter, 100 tosses isn't enough anyway.

Me: How many would be enough?

You: Well, if you tossed it an infinite number of times, it would come up heads half the time.

Me: Really? So, if we tossed it an infinite number of times it would come up heads an infinite number of times?

You: What do you mean?

Me: Well, half of infinity is infinity, and you said it would come up heads half the time.

You: Now you're just being stupid. YOU KNOW WHAT I MEAN!!!

Actually, I do know what you mean. What's confusing is that we can't say what the probability of a head "really" is. When we say that the probability of a head is $1/2$ we are stating an *axiom*. That is, we are assuming (or stipulating) it is true but we aren't proving it.

[2] Whenever I don't know the answer to a question that a student asks I always say "Good question. Who knows the answer?". If nobody knows the answer I then say "This is exactly the kind of question that might show up on the final exam. I suggest you all find the answer before then." It works quite well.

[3] Let's all agree right now that a coin can never land on its edge. Sure, I've seen it happen too, but let's just ignore it.

6.2 Be Discrete

Whenever you work with probabilities you actually conduct a thought experiment in which you know all of the possible outcomes. You'll never admit it for fear of looking like a nerd, but we both know it's true. In fact, not only do you conduct a thought experiment, you use all kinds of cool words to describe it, don't you? I can practically hear you thinking now – "I'll call each possible outcome a indexsample points *sample point* and I'll call the set of all possible sample points the *sample space*." Well, if you're not thinking that, you should be, because those are the terms I'm going to use from here on in.

It turns out that many of the sample spaces that we encounter in every day life have only a finite number of members. That is, there are only a finite number of possible outcomes to the thought experiment. For example, when we think about tossing a coin the sample space consists of the two sample points: heads (H) and tails (T). As another example, when we think about rolling a single die the sample space consists of six sample points: ⚀, ⚁, ⚂, ⚃, ⚄, and ⚅. These kinds of sample spaces are called *discrete sample spaces* and are, by far the easiest to understand and work with.

So where does the probability come in? Well, once we know all of the sample points we can **assign** each one a probability. That's right – we simply assert the probability of each sample point. There are only two rules that we have to follow. First, each probability must be greater than or equal to zero. Second, all of the probabilities have to add up to one.

If we want to look smarter, we can write these rules down mathematically. Suppose our sample space, $\mathcal{S}$, contains n sample points denoted by $O_1, O_2, O_3, \ldots O_n$, and that the probability of a particular sample point, O_i, is denoted by $P\{O_i\}$. Then, the two rules are:

$$P\{O_1\} \geq 0, P\{O_2\} \geq 0, \ldots, P\{O_n\} \geq 0 \tag{6.1}$$

and

$$P\{O_1\} + P\{O_2\} + \ldots + P\{O_n\} = 1. \tag{6.2}$$

Back to our coin tossing example, we traditionally assert that the probability of 'heads' and the probability of 'tails' is equal. That is, $P\{H\} = P\{T\}$. Combining this with (6.1) and (6.2) it follows that $P\{H\} = 1/2$ and $P\{T\} = 1/2$.

6.3 Special Events

Once we have defined the sample space and assigned probabilities to all of the outcomes we can begin to ask and answer some useful questions. To do that, we first have to define the notion of a *random variable*. A random variable is a rule that assigns a number to every element in the sample space.[4] For example, suppose you went to Atlantic City to gamble at the coin tossing table. The croupier[5] tosses a coin. If it comes up heads then you get $10.00. If it comes up tails then you get $5.00. We now have a rule that assigns numbers (i.e., $10.00 and $5.00) to every element in the sample space (i.e., H and T) so we now have a random variable.[6]

We often write random variables as capital letters (e.g., X) and the possible values that they can take as lower case letters (e.g., x_1, x_2). Then, we can talk about the probability of different *events*, which is just the probability that a random variable takes on certain values. In particular, the probability of a particular event is just the sum of the probabilities of the sample points that are in that event.

For example, let's consider rolling a die. Suppose we have a random variable, X, that assigns the value 1 to the outcome ⚀, the value 2 to the outcome ⚁, and so on. Then, one event might be $X = 1$ and another event might be $X \leq 3$. The first corresponds to the die coming up ⚀ and the second corresponds to the die coming up ⚀, ⚁, or ⚂. The probability of $X = 1$ is just the probability that the die comes up ⚀ while the probability of $X \leq 3$ is just the probability that the die comes up ⚀ plus the probability that the die comes up ⚁ plus the probability that the die comes up ⚂. So, assuming that the probability of each outcome is $1/6$, the probability of $X = 1$ is just $1/6$ and the probability of $X \leq 3$ is $1/6 + 1/6 + 1/6 = 3/6 = 1/2$.

We sometimes use a shorthand notation to refer to two special kinds of events. First, the probability that a random variable X takes on the value x is often written as $f(x)$ and is referred to as the *probability density function*. Second, the probability the a random variable X takes on some value less than or equal to x is often written as $F(x)$ and is referred to as the *cumulative distribution function*. These special events come up in a wide variety of circumstances, as you are about to see.

[4] You're thinking that you've seen a definition like this before, aren't you? Well, you're right – it's very similar to how we defined a function. In fact, if it were up to me I wouldn't use the term random variable I would use the term random function. Unfortunately, I'm not a famous probabilist so I don't get to make those kinds of decisions. Maybe we should organize a letter-writing campaign. If you're interested, give me a call.

[5] Did you ever wonder where the word 'croupier' comes from? Best I can tell it means a person that rides on the rump of a horse. I have no idea why the call the person that runs a gambling table a croupier.

[6] You're wondering where you can play this game, aren't you? It sounds like you always win, doesn't it? Don't get too excited – I didn't tell you what it costs to play!

6.4 That's Not What I Expected

Bored yet? Don't give up! Here's where I explain how you can win a million dollars at your local casino!

First, let's go back to that coin tossing game where you get $10.00 if the coin comes up heads and $5.00 if it comes up tails. If the coin is fair (i.e., if the probability of H is 0.5 and the probability of T is 0.5) then the probability density function is $f(10) = 0.5$ and $f(5) = 0.5$. That is, the probability that you get $10 is 0.5 and the probability that you get $5 is 0.5. So, how much would you be willing to pay to play this game?

Well, suppose I asked you for $20 to play, would you agree? Of course not – at best you'd lose $10 every game. Suppose I asked you for $10 to play? It's still unlikely that you'd be willing to play because the best you could do would be to break even, and that wouldn't always happen. So, I ask again, how much would you be willing to pay to play this game?

The right way to think about this is to ask about the *expected value* of the game. The expected value of a random variable, X, that can take the values $x_1, x_2, \ldots, x_n$ is usually written as $E(X)$ and is defined as follows:

$$E(X) = f(x_1) \cdot x_1 + f(x_2) \cdot x_2 + \cdots + f(x_n) \cdot x_n. \tag{6.3}$$

That is, the expected value of a random variable is the sum of the values that the random variable can take on (i.e., all of the values in the range of the random variable) weighted by (i.e., multiplied by) their respective probabilities.

So, what is the expected value of the coin tossing game? Easy, it's just $0.5 \cdot 10 + 0.5 \cdot 5 = 7.5$. Loosely speaking, this means that you can "expect" to get $7.5 dollars if you play. Will you definitely get $7.5? No! In fact, if you only play once it is impossible for you to get $7.5, since the random variable only takes on the values 10 and 5. However, in a probabilistic sense, $7.5 is what you can "expect" to get. So, would you be willing to pay $6 to play this game? I would since the expected value is more than 6.[7] Am I guaranteed to win? No!

Now let's consider a more complicated game – roulette. There are 38 slots on a roulette wheel numbered 1-36, 0, and 00. 0 and 00 are colored green, have of the remaining numbers are colored black, and the other half of the remaining numbers red. The croupier spins the wheel and then places a small ball on it. The ball bounces around for some time and finally drops into a slot. We'll assume that the probability of each individual outcome is $1/38$ (i.e., that the wheel is fair) and that it costs $1 to play.

[7] For many people, a bird in the hand is worth two in the bush. That is, they would rather have $6 for sure than a 50-50 chance of either having $5 or $10. In fact, I think that I'm one of those people. It's still important to know what the expected value is, however.

In one variant of roulette you bet on a single number. If that number comes up you get your dollar back plus \$35 more. Suppose you bet on the number 13. Then, the random variable associated with this game is:

$$X = \begin{cases} 0 & \text{if the outcome is 00} \\ 0 & \text{if the outcome is 0} \\ 0 & \text{if the outcome is 1} \\ \vdots & \\ 0 & \text{if the outcome is 12} \\ 36 & \text{if the outcome is 13} \\ 0 & \text{if the outcome is 14} \\ \vdots & \\ 0 & \text{if the outcome is 36} \end{cases} . \tag{6.4}$$

In this game, there are 37 different outcomes that correspond to the event $X = 0$ and there is 1 outcome that corresponds to the event $X = 36$. Since the probability of each event is $1/38$ this means that the expected value of X is $37/38 \cdot 0 + 1/38 \cdot 36 = 36/38 = 0.94736842$. Obviously, this means that the expected value of the game is less than the cost of playing (i.e., \$1).

In another variant of roulette you bet on either even or odd (where 0 and 00 are condidered neither even nor odd). If you bet on even and the ball lands in an even-numbered slot you get your dollar plus \$1 more. Similarly, if you bet on odd and the ball lands in an odd-numbered slot you get your dollar plus \$1 more. Suppose you bet on odd. Then, the random variable associated with this game is:

$$Y = \begin{cases} 0 & \text{if the outcome is 00} \\ 0 & \text{if the outcome is 0} \\ 0 & \text{if the outcome is 2} \\ 0 & \text{if the outcome is 4} \\ \vdots & \\ 0 & \text{if the outcome is 36} \\ 2 & \text{if the outcome is 1} \\ 2 & \text{if the outcome is 3} \\ \vdots & \\ 2 & \text{if the outcome is 35} \end{cases} . \tag{6.5}$$

In this game, there are 20 different outcomes that correspond to the event $X = 0$ and there are 18 outcome that corresponds to the event $X = 1$. Since the probability of each event is $1/38$ this means that the expected value of X is $20/38 \cdot 0 + 18/38 \cdot 2 = 36/38 = 0.94736842$. Again, the expected value of the game is less than the cost of playing (i.e., \$1).

6.5 There Ought to be a Law

As you can see, the expectation of a random variable can be a pretty useful number to know. But something is still troubling you, isn't it? You're bothered by this notion that probabilities are assumed/assigned and not observed. If that's the case, why can't somebody say that the probability of a head is $1/5$? They can (as long as they are also willing to say that the probability of a tail is $4/5$). Fortunately, there are ways that we can evaluate the reasonableness of these kinds of claims. One method is to use the strong*law of large numbers*.

Suppose we flip a coin a whole bunch of times and we let the random variable X_i equal 1 if the ith flip is a head and 0 if it is a tail. If we flip the coin n times we will then have a sequence of random variables, $X_1, X_2, \ldots, X_n$. Now, if the probability of a head is defined to be a (which means that the probability of a tail is $1 - a$ since the two must sum to 1) then the expected value of each of these random variables is just $a \cdot 1 + (1 - a) \cdot 0 = a$.

Now, let's create a new random variable, Z_n, which is just the sum of the Xs. That is, $Z_n = X_1 + X_2 + \cdots + X_n$ which simply says that Z_n is the number of heads obtained in the n tosses. What do you think the expected value of Z_n should be? Well, if the expected value of each toss is a then the expected value of n tosses should be $n \cdot a$, and it is.

Okay so far? Good! Now what we want to do is work backwards. Suppose, somebody gives you a coin and says that the probability of a head is a. You toss the coin n times and it turns out that the number of heads is Z_n. What can you conclude about a? Well, you'd like to say that $a = Z_n/n$. So, for example, if you tossed the coin 1000 times (i.e., $n = 1000$) and you got 450 heads (i.e., $Z_n = 450$) you'd like to say that the probability of a head is $450/1000 = 0.45$.

Unfortunately, you can't conclude that $a = Z_n/n$. In fact, the coin tossing example makes it clear that you can't because it's not unreasonable to get 450 heads out of 1000 tosses of a fair coin, but you "know" that the probability is 0.5. Intuitively, what you're really saying is that the probability of a head must be approximately 0.45. Unfortunately, the word 'approximately' is not very precise. Fortunately, the strong law of large numbers let's us say something precise.

In particular, the strong law of large numbers says that:

$$\lim_{n \to \infty} \frac{Z_n}{n} = a. \tag{6.6}$$

That is, as n gets large the fraction Z_n/n approaches the true probability.

6.6 Where Do We Go From Here

So, now you know two things. You know how to gamble and you know how to determine whether the probability that a person assigns to an outcome is reasonable. By now you're probably wondering what this has to do with the stock market and getting promoted. To answer that question, we need to introduce one more cool word.

A *stochastic process* is a collection of random variables with the same sample space. The collection of random variables is often written as $X_1, X_2, \ldots, X_T$ and we often refer to a generic element in this collection as X_t. We use t because a stochastic process is often something that evolves over time.

For example, suppose we watch T people at a checkout line to see whether they pay with cash or not.[8] For each person, t, we let $X_t = 1$ if the person pays with cash and 0 if he or she doesn't. From our perspective, each X_t is a random variable and the collection, $X_1, X_2, \ldots, X_T$ is a stochastic process.

There are many different stochastic processes, but among the most important are *Markov chains*. A Markov chain is a sequence of random variables that have the property that the next random variable (i.e., X_{t+1}) is independent (more on this later) of the past random variables (i.e., $X_1, X_2, \ldots, X_{t-1}$) provided that we know the present random variable (i.e., X_t).

Though they sound a little complicated, Markov chains are actually very easy to understand and use because they can be represented in tabular form. For example, let's consider what happens when somebody that has had too much to drink tries to walk down a long corridor as shown in Figure 6.1.

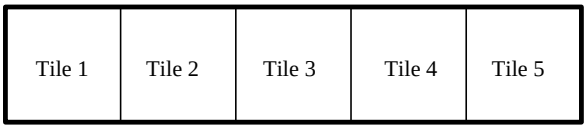


Figure 6.1: A Long Corridor

Let's assume that, because of her/his inebriated condition, the person will walk to the right one tile with probability 0.75 and will walk to the left one tile with probability 0.25. The only exception is when the person bumps into the wall at either end. Then he or she will walk in the opposite direction for sure.

The table, or *transition matrix*, that describes this Markov chain is as follows:

[8] Take my advice, if you're going to do this yourself you should check with the manager first and explain what you're doing to the person working the cash register. The only time I've done this the checkout person thought I was evaluating her performance. After about 45 minutes she started yelling at me that I should "take this job and shove it".

	Tile 1	Tile 2	Tile 3	Tile 4	Tile 5
Tile 1	0	1	0	0	0
Tile 2	0.25	0	0.75	0	0
Tile 3	0	0.25	0	0.75	0
Tile 4	0	0	0.25	0	0.75
Tile 5	0	0	0	1	0

The rows of this table represent the tiles that the drunk can be standing on at time t and the columns represent the tiles that the drunk can be standing on at time $t + 1$. Each entry in the table is the probability that the drunk moves from one tile to another. So, the first row says that if the drunk is on tile 1 at time t then he or she will, with probability 1, move to tile 2 at time $t + 1$. The second row says that if the drunk is on tile 2 at time t then he or she will move to tile 1 with probability 0.25 and to tile 3 with probability 0.75. The other rows can be interpreted in similar ways.

This particular Markov chain is called a *random walk*, and it has been applied in a variety of different ways (e.g., to the stock market). However, it is not necessary that Markov chains have this structure. All that is necessary is that all of the entries in the table are nonnegative and that the entries in each row sum to 1.

How else might Markov chains be applied. Well suppose you were responsible for doing the manpower planning[9] for your firm. You could use a Markov chain to describe promotion probabilities as follows:

	VP	AVP	Proj. Ldr.	Sen. An.	Analyst	Quit
VP	0.83	0	0	0	0	0.17
AVP	0.23	0.65	0	0	0	0.12
Proj. Ldr.	0	0.04	0.86	0	0	0.10
Sen. An.	0	0	0.08	0.89	0	0.03
Analyst	0	0	0	0.03	0.93	0.04

You could then make predictions about how top heavy or bottom heavy the firm will become.

For example, suppose in year 1, the company has 100 analysts, 30 senior analysts, 10 project leaders, 5 assistant vice presidents, and 2 vice presidents. The expected number of VPs in year 2 is the expected number of VPs that stay VPs plus the number of AVPs that get promoted to VP. That is, the expected number of VPs in year 2 is $0.83 \cdot 2 + 0.23 \cdot 5 = 2.81$. Similarly, the expected number of AVPs in year 2 is $0.65 \cdot 5 + 0.04 \cdot 10 = 3.65$. Performing these same kinds of calculations for the other

[9]Now, before we go any further (or should that be farther), let me apologize for using the term "manpower". I know it's sexist, I really do. I even know, and freely admit, that "men are scum". (I have Carmen, Denise, and Alice to thank for this, women I went to graduate school with. They devoted the better part of two years to making sure that, if nothing else, I learned that. I am deeply indebted to them.) It's just too hard to keep saying 'man/woman-power' and personpower just sounds like some motivational thing.

categories results in 11 project leaders in year 2, 29.7 senior analysts in year 2, and 93 analysts in year 3.

In year 3, the expected number of VPs is $0.83 \cdot 2.81 + 0.23 \cdot 3.65 = 3.17$. For the other categories, the expected values in year 3 are 2.81, 11.83, 29.22 and 86.49. So, by year 3 the expected number of total employees is 133.53 (down from 147) and the expected number of VPs is 3.17 (up from 2).

6.7 It Gets Immeasurably Harder

Not too hard, is it? All these years you've been impressed by probability theorists for nothing, haven't you? Well, let's see how good you really are.

First an easy one. I just picked an integer that is greater than or equal to 1 and less than or equal to 10. How many sample points are there? Ten. What is the probability of each sample point? We have to assign the probabilities. What probabilities are the most "natural"? I think, and you probably do too, that each sample point is equally likely and, since the probabilities must sum to one, this means that the probability associated with each sample point should be $1/10$.

Now, a harder one. I just picked a real number that is greater than or equal to 0 and less than or equal to 1. How many sample points are there? There are an uncountable infinity. What is the probability of each sample point? We have to assign the probabilities. What probabilities are the most "natural"? Well? Do you have an answer? If I'm not mistaken, you're now waging a battle with yourself that is going something like this:

You: In the first example I divided 1 into the number of sample points to get $1/10$. So, I'll do the same thing here and get $1/\infty$ which means the probability of each sample point should be 0.

You: Do I really think the probability that he picked any number should be 0?

You: Sure! There are an infinite number of possible outcomes. The probability of any one should be zero.

You: But then the probabilities don't sum to one!

You: What am I talking about?

You: If I add up zero an infinite number of times I get zero.

You: So, it can't be zero, it must be a very small number. That's it! The probability must be some small number ϵ.

You: That can't be it. If I add up an infinite number of ϵs I get infinity. The sum is supposed to be 1.

You: Why am I reading this book anyway?

It turns out that you're right. And so are you! No, maybe you're wrong. It certainly is a conundrum, isn't it? You think that the probabilities should be the same for all of the sample points since they're all equally likely. However, if you make the probability of each event positive then the probabilities sum to infinity (which isn't one) and if you make them all zero then they sum to zero (which also isn't one). And, your only choices are zero or some positive number. What should you do?

It turns out that you need to assign the probabilities differently. One way to proceed is to assign probabilities to intervals, rather than individual real numbers. For example, what is the "natural" probability that the number I chose was between 0 and $1/2$? That's easy, $1/2$. What about the the "natural" probability that the number I chose was between $1/2$ and 1? Again, it should be $1/2$. Is this internally consistent? Yes! These two events are *mutually exclusive* (i.e., they don't have any sample points in common) and they encompass all possible outcomes. So the sum of their probabilities should be 1 and it is. This is illustrated in Figure 6.2.

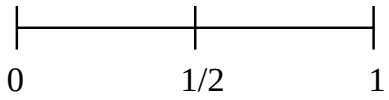


Figure 6.2: Two Intervals

In fact, what we really need to do is assign a probability to each and every interval. How many intervals are there? There are an uncountable infinity. Fortunately, we can still write down a rule for assigning probabilities to intervals.

Recall that when we had a discrete sample space we needed to ensure that the probability of each sample point was greater than or equal to zero and that all of the probabilities added up to 1. In a *continuous sample space* (i.e., when the number of sample points is uncountably infinite) we also need obey two rules. Specifically, we need to define f in such a way that if we were to draw a curve representing f it would always be greater than or equal to zero and the area underneath the curve would equal 1.

In this example, we could define f as follows:

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if } x > 1 \end{cases} . \tag{6.7}$$

This says that $f(x)$ is 0 if x is less than 0 or greater than 1, and that $f(x)$ is 1 if x is greater than or equal to 0 and less than or equal to 1. This is called the *uniform* density function on the interval $[0, 1]$. A plot of f is shown in Figure 6.3. Clearly f is

always greater than 0. What about the area under f? Since the curve representing f is a rectangle it's very easy to calculate this area. Do you remember your geometry? The area of a rectangle is just the length times, l, the height, h. In this case, the $l = 1$ and $h = 1$ so the area is just $l \cdot h - 1 \cdot 1 = 1$.

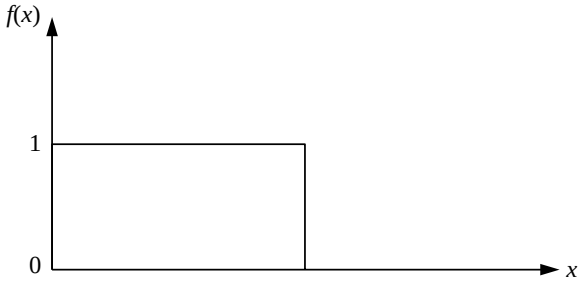


Figure 6.3: The Uniform Density Function on [0,1]

How do we use this to calculate probabilities? It's easy – the probability of being between two numbers x_L and x_R is just the area under f between x_L and x_R. For example, suppose $x_L = 0.125$ and $x_R = 0.750$ as shown in Figure 6.4. Then, the height of the rectangle is given by $h = 1$ and the length of the rectangle is given by $l = (x_R - x_L) = 0.750 - 0.125 = 0.625$. So the area, and hence the probability, is $l \cdot h = 0.625 \cdot 1 = 0.625$.

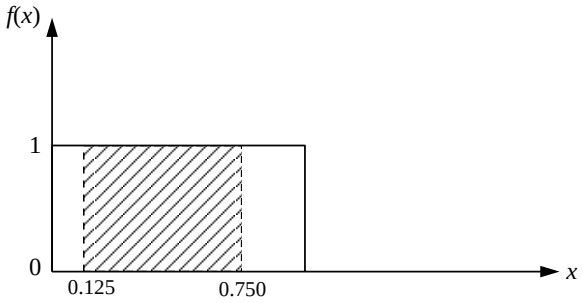


Figure 6.4: The Probability of $0.125 \leq X \leq 0.750$

Easy, right? Okay, then you're ready for a quiz. Given this density function, f, what is the cumulative distribution function, F? I'll give you a hint. Recall that $F(x)$ is the probability that the random variable takes on some value less than or equal to x. So, for x between 0 and 1 $F(x)$ is just the area under f between 0 and x. Since this rectangle has a length of $l = x - 0$ and a height of $h = 1$, it follows that its area is just $l \cdot h = (x - 0) \cdot 1 = x \cdot 1 = x$. Furthermore, for x less than 0 the area under f is 0 and for x greater than 1 the area under f is the same as the area under f at 1. So, it follows that the cumulative uniform distribution on $[0, 1]$ is given by:

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if } x > 1 \end{cases} . \tag{6.8}$$

What about the probability of a particular number? This is essentially the same as asking about the area of a rectangle that has a length of 0. More formally, we say that a point has *zero measure*. Hence, it has a probability of zero.

6.8 Green Events and Ham

About now I can hear you cry, "You're wasting our time with nonsense! Why?". I promise you, I would not lie, measure theory will be useful, by-the-by. We'll get there soon, don't you sigh, but now I have other eggs to fry.[10]

What we have to worry about now are relationships between events. Which events go together and which do not, and how we calculate the probabilities of multiple events.

Let's assume that we have two events, A and B. We often want to know things like "the probability that A or B (or both) occurs" and "the probability that both A and B occur". It turns out that this is easier than you might think, as long as you are careful. The key is to remember that events are sets of sample points.

There are two cases that we need to consider, and these are illustrated in the *Venn diagrams* in Figure 6.5. On the left, the two events, A and B, are *mutually exclusive*. That is, they have no sample points in common. On the right, the two events *intersect*. That is, they have sample points in common. Simple right?

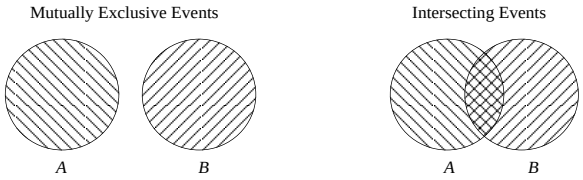


Figure 6.5: Mutually Exclusive and Intersecting Events

Okay, let's talk about a specific application. Suppose we roll a die, and event A is defined by $\{x \leq 2\}$ (i.e., the event in which the outcome is either a 1 or a two) and event B is defined by $\{x \leq 4\}$ (i.e., the event in which the outcome is 4 or less). Are these events mutually exclusive or overlapping? Obviously, they are overlapping.

[10] Watch out Dr. Seuss. Dr. Bernstein is on the loose!

If you understand this, you're ready. First, let's consider the probability that A or B or both occur. Well, if $P(A)$ denotes the probability of the event A and $P(B)$ denotes the probability of the event B, then it must be the case that:

$$P(A \text{ or } B \text{ or both}) = P(A) + P(B) - P(A \text{ and } B). \tag{6.9}$$

Why? Just think about what's going on. The events A and B are just sets of sample points. So, if we want to know the probability of A or B or both what we want to know is the probability of the *union* of the two sets A and B. You remember how the union of two sets is defined, don't you. If you have two sets, S_1 and S_2, then the union of S_1 and S_2 is denoted by $S_1 \cup S_2$ and is defined as the set containing the members of both S_1 and S_2. So, since A and B are just sets of sample points, $P(A \text{ or } B \text{ or both}) = P(A \cup B)$. Finally, since sets don't contain duplicates, $P(A \cup B) = P(A) + P(B) - P(A \text{ and } B)$.

Now you see why I started out by talking about mututally exclusive and intersecting sets. What happens when A and B are mutually exclusive? Then $P(A \text{ and } B)$ is 0! Why? Because if they have no sample points in common then the probability of both events must be zero. For example, let's think about rolling a die again and suppose that our two events are $\{x \leq 2\}$ and $\{x \geq 5\}$. What's the probability of one roll of the die yielding both a number that is less than or equal to 2 and greater than or equal to 5? Zero!

So, what can we conclude? When two events are mutually exclusive, $P(A \text{ or } B \text{ or both}) = P(A) + P(B) - P(A \text{ and } B) = P(A) + P(B) - 0 = P(A) + P(B)$.

What about when the two events aren't mutually exclusive? For that, we need to think more about $P(A \text{ and } B)$. But that's okay, we wanted to do that anyway!

To get started, we can consider the easy case. In fact, you already know this case. Suppose we toss two coins, and the probabilty of a head is $1/2$ for each. What's the probability that they both come up heads? I'm sure you've seen this before, probably when you studied genetics. The answer is $1/2 \cdot 1/2 = 1/4$. But why?

Well, it turns out that you've made an important, but reasonable, assumption. You've assumed that the two random variables are *independent*. In fact, this is the definition of independence. Two random variables, X_1 and X_2, are said to be independent if and only if $P(X_1 = a_1 \text{ and } X_2 = a_2) = P(X_1 = a_1) \cdot P(X_2 = a_2)$. That is, they are independent if the probability that the first random variable takes on the value a_1 and the second random variable takes on the value a_2 is exactly equal to the product of the two probabilities. So, if X_1 and X_2 are our two coin tosses, a_1 represents a head on coin 1, and a_2 represents a head on coin 2, then the probability that both coins come up head is $1/2 \cdot 1/2 = 1/4$.

What if the two random variables are not independent? That's the more complicated case. To talk about this case we need to define the notion of a *conditional probability*. A conditional probability is just the probability that one event occurs given that another

occurs. Unfortunately, we can't really define a conditional probability directly. We can only do it indirectly.

So let's try. First, let's let $P(A|B)$ denote the conditional probability (of event A given event B). Then, $P(A|B)$ is a number between 0 and 1 (inclusive) that satisfies the following condition:

$$P(A \text{ and } B) = P(A|B) \cdot P(B). \tag{6.10}$$

Now, you're probably wondering why I didn't just define $P(A|B)$ as $P(A \text{ and } B)/P(B)$. The reason is technical but important. If $P(B) > 0$ then $P(A|B)$ does equal $P(A \text{ and } B)/P(B)$. But what about when $P(B) = 0$? Then $P(A|B)$ is not uniquely defined, it is any number between 0 and 1 (inclusive).

In fact, if you think about it, it even makes some intuitive sense. What's the probability of A given that B occurs when $P(B) = 0$? It's the probability of A given that B occurs when B doesn't occur. We can't say it's ∞ because probabilities must be between 0 and 1. So, we define it to be any number between 0 and 1.

Now, let's try and combine independence and conditional probability. We know that when are two random variables, X_1 and X_2 are independent $P(X_1 = a_1 \text{ and } X_2 = a_2) = P(X_1 = a_1) \cdot P(X_2 = a_2)$. We also know that, in general, $P(X_1 = a_1 \text{ and } X_2 = a_2) = P(X_1 = a_1|X_2 = a_2) \cdot P(X_2 = a_2)$. So, when X_1 and X_2 are independent it must be the case that $P(X_1 = a_1|X_2 = a_2) = P(X_1 = a_1)$. In fact, that makes perfect sense. The probability that $X_1 = a_1$ given that $X_2 = a_2$ should simply be the probability that $X_1 = a_1$ when X_1 and X_2 are independent. After all, X_1 isn't influenced by X_2.

What about when the two random variables are not independent? Then, all we know is that $P(X_1 = a_1 \text{ and } X_2 = a_2) = P(X_1 = a_1|X_2 = a_2) \cdot P(X_2 = a_2)$. There's just not much else that we can say.

So, when the events are mutually exclusive we can answer both of the questions we started with. Also, when the random variables are independent we can answer both of these questions. In other cases, things get pretty darn difficult.

6.9 Slow and Steady gets the A

When using probability you have to be very careful both about the mathematics and the language used. Suppose, for example, that you are work at Shenandoah Valley Regional Airport (near Harrisonburg, VA) and know that there are two flights to Washington's Dulles International Airport each day – flight 0 and 00. For each flight there are two outcomes, a flight can be "on time" or "late". The probability that a flight is

"on time" is 0.8. Somebody now asks you the question "What is the probability that either of the flights is on time?".

To answer this question you first have to be careful about the language used. That is, you have to deal with the ambiguities of English. In this case, after a little back and forth, you determine that the question can be re-phrased as "What is the probability that one or the other or both of the flights is on time?".

Next, you have to be careful about the mathematics. In this case, what you want to find is $P(0 \text{ is on time or } 00 \text{ is on time})$. Here's where you have to be careful. You might be inclined to say that the answer is the sum of the probabilities of the individual events. If you did, you'd conclude that the answer was $P(0 \text{ is on time}\} + P\{00 \text{ is on time}) = 0.8 + 0.8 = 1.6$.

Hopefully you're a little air-sick right now. You know that all probabilities are in the interval $[0, 1]$ and 1.6 is clearly greater than one. Hence, this can't be the right answer. Unfortunately, you can't count on finding this kind of mistake this way. For example, suppose the probability that a flight is "on time" were 0.3. Then, you would have gotten an answer of 0.6 which would have seemd fine.

Instead, you have to avoid making these kinds of mistakes. That is, you have to be slow and steady (i.e., careful and methodical). What mistake led to your queasiness? You (that is, I) didn't apply the "or rule" carefully. That is, you had to remember that:

$$\begin{aligned} P(0 \text{ is on time or } 00 \text{ is on time}) &= P(0 \text{ is on time}) + P(00 \text{ is on time}) \\ &\quad - P(0 \text{ is on time and } 00 \text{ is on time}) \\ &= 0.8 + 0.8 - (0.8 \cdot 0.8) = 1.60 - 0.64 = 0.96 \end{aligned}$$

6.10 Flip it Good

One way to make sure that you have the correct answer is to flip a problem around. Continuing with this same example, we could have worked with the outcomes "late" rather than the outcomes "on time". Since we know that the probabilities of all outcomes in the sample space must sum to 1, we know that:

$$P(0 \text{ is late}) + P(0 \text{ is on time}) = 1 \tag{6.11}$$

Hence, we know that:

$$P(0 \text{ is late}) = 1 - P(0 \text{ is on time}) = 1 - 0.8 = 0.2 \tag{6.12}$$

Similarly, we know that $P(0 \text{ is late}) = 0.2$.

Now, assuming that the two flights are indepenent:

$$\begin{aligned} P(0 \text{ is late and } + P(00 \text{ is late}) &= P(0 \text{ is late}) \cdot P(00 \text{ is late}) && (6.13) \\ &= 0.2 \cdot 0.2 = 0.04. && (6.14) \end{aligned}$$

So, it follows that the probability that neither is late (i.e., that one or the other or both is on time) is $1 - 0.04 = 0.96$, which is precisely the answer we got before.

6.11 It Will Make Sense Event-ually

At this point you might be wondering why I numbered the flights 0 and 00.[11] It's because I wanted to have an opportunity to discuss the difference between the probability of an individual event and the probability of multiple events.

In roulette, the event "Green" consists of the two sample points, 0 and 00. That is:

$$\text{Green} = \{0, 00\} \tag{6.15}$$

Hence, because the probability of an event is just the sum of the probabilities of the sample points in that event:

$$P(\text{Green}) = P\{0\} + P\{00\} \tag{6.16}$$

On the other hand, in the case of flights 0 and 00, the probability that either flight is on time involves two events – 0 being on time and 00 being on time. Hence, we have to use the "or rule":

$$\begin{aligned} P(0 \text{ is on time or } 00 \text{ is on time}) &= P(0 \text{ is on time}) + P(00 \text{ is on time}) \\ &\quad - P(0 \text{ is on time and } 00 \text{ is on time}) \end{aligned}$$

So, some final advice – don't take any shortcuts. Always start by identifying the outcomes of the thought experiement, then assign probabilities to those ouctomes, then

[11] Shame on those of you who were sure that it was a statement about my personality and/or the number of dates I had in college!

construct the events, then create the random variables, then use the “and rule”, “or rule” or the definition of the expected value to perform some calculations. After that, make sure your answers are reasonable and, if possible, flip the problem around and make sure you get the same answer. Though this may not be the shortest path to the answer, you’ll get there eventually and you’ll be sure you’re getting to the right answer.

Chapter 8

Why you wait in line!

THERE ARE few things in life that annoy me more than waiting in line. In fact, I hate it so much that I do everything that I can to avoid it. I don't travel to work during the rush hour, I do all of my banking electronically or at off-peak hours, I get my hair cut (what's left of it) during the middle of the day, I only go to movies after they've been out for a few weeks, I shop for groceries at ridiculous hours, etc. . .. Why do I hate waiting in line so much? I think James, one of my college roommates, summed it up best with his immortal "Got a lot to do, don't wanna rush". At the time I never knew what he was talking about, but it makes perfect sense to me now.

Oddly enough, while I hate waiting in line[1], I like talking to people about waiting in line. When I have conversations with people about this topic[2] they often go something like this:

Me: Can I ask you a question?

Him: Well...

Me: Why do you wait in line?

Him: Because I don't have any choice!

Me: No, I know why you wait in line when there is a line, but why do you think that there are lines?

[1] I can hear some of you now. "People don't wait in line, they wait on line". I'm not sure what's 'proper' English and I don't care. I've said 'in line' my whole life and I'm not changing now.

[2] I suspect that you are now starting to worry that you and I might get invited to the same cocktail party at some point. Don't worry – I don't get invited to many anymore. Surprised? Read on.

Him: Because if there weren't lines we'd constantly be fighting about whose turn it was.

Me: No, I know why there are lines for us to wait in, but why do you think you have to wait?

Him: Because there is somebody in front of me!

(At this point, I often go running to my wife to ask her if there is something wrong with me. She generally says "Yes, but why do you ask?". I explain and she takes out a pad of paper and starts making notes. You see, my wife is a theoretical linguist and I am the source of a wealth of data.[3] This particular exchange is either an example of a structural ambiguity or a lexical ambiguity. I can never remember which.)

Me: But why is there always somebody in front of you?

Him: Why didn't you ask that to begin with? Because there are too few checkout people/toll booths/tellers/....

Everybody I talk to ultimately has the same opinion – that there are too few checkout people/toll booths/tellers/.... Is that really the problem? That's what we're going to talk about in this chapter.

8.1 Check it Out

I stopped at the grocery store on my way home from work a few weeks ago because I wanted to pick up something for dinner.[4] It was about 5:30PM and there was an enormous line at every cashier. In fact, I was just about the last person in line.

What caused this line (or *queue*)? I think the explanation is simple, and is shown graphically in Figure 8.1, where t represents the time of day, and $Q(t)$ represents the number of people in the queueing system[5]) at time t. At about 5:00 everybody left work and about 200 people headed to the grocery store. They each grabbed the one or two items that they needed and raced to the cashiers. There were 5 cashiers working, each of whom could check-out 50 people per hour. That meant that by 5:30 (when I arrived) they had checked-out 125 people (5 cashiers times 50 people per hour times 1/2 hour) and the remaining 75 were still waiting in line (about 15 in each line).

[3] So, you see, it's not just my fault that we don't get invited to many cocktail parties. Between my constant "Did you ever wonder why..." and her constant scribbling, we're about as popular as the plague.

[4] Two things. First, I told you earlier that I take the bus to work, and I usually do. However, I have been known to miss it (it only comes once an hour), in which case I have to beg my wife to use this car. This was one of those days. Second, we don't cook any more. We have two grocery stores – Davidson's and McCaffery's – that do all of our cooking for us. It works out quite well.

[5] Pretty clever use of the letter 'Q' isn't it?

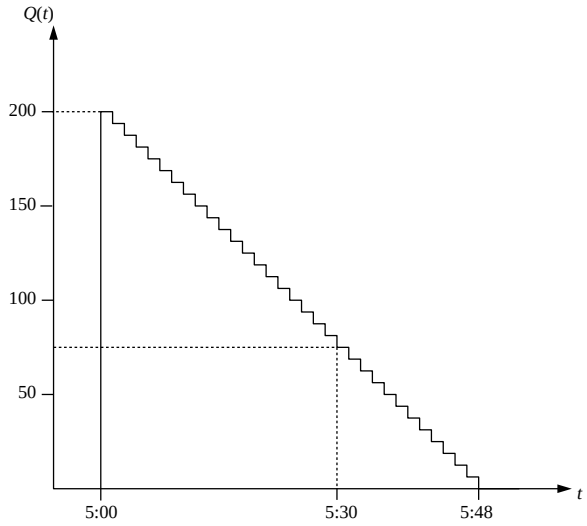


Figure 8.1: My Experience at the Grocery Store

So, why was there a line when I arrived? Because at 5:00 there were not nearly enough cashiers to handle the customers. Hence, a line formed at 5:00 and it didn't dissipate before I arrived. In essence, the line was created by a dramatic increase in the number of customers.

What's interesting about this kind of line is that it almost can't be avoided – a line almost has to form in these kinds of situations. Why? Because the grocery store can't possibly have enough cashiers. In order to have prevented a line from forming the grocery store would have had to have about as many cashiers as customers. This is not to say that the line had to last as long as it did. For example, if they'd had 10 cashiers working the line would have dissipated by 5:24, and if they'd had 20 cashiers working it would have dissipated by 5:12. But, in some sense, a line almost had to form.

8.2 It's a Small World

Is this the only reason for lines? Not really. To understand why, consider the following example.

I was at a conference[6] in Atlanta recently. One night I went out to dinner with Kelley, Thanasis, Srinivas, Haris, and Elise. Since none of us had ever been there, we decided

[6]The conference was organized by the Institute for Operations Research and the Management Sciences (INFORMS). An entire hotel filled with people that study and apply the kinds of methods discussed in the book. Makes you sorry you missed it, doesn't it?

to go to Planet Hollywood^{TM} . We had to wait for a few minutes to get a table.

The restaurant seemed to have about 60 tables and the average group seemed to take about an hour to eat. So, a table was coming empty about every minute. Groups seemed to be arriving about every minute and a half. In other words, unlike the earlier example, there wasn't a dramatic increase in the number of customers. So, why the line?

I think that the key to thinking about this example is my use of the word 'about'. Tables came empty "about" every minute. Groups arrived "about" every minute and a half. Was it exactly a minute and a minute and a half? Of course not – there was some randomness.

A hypothetical version of this is illustrated in Figure 8.2. Each '+' indicates a group arriving at the restaurant and each '-' represents a group leaving. As you can see, the arrival rate and the departure rate are the same for the period shown (there are 12 arrivals and 12 departures). Nevertheless, because the arrivals and departures are random, a line forms (and grows and shrinks).

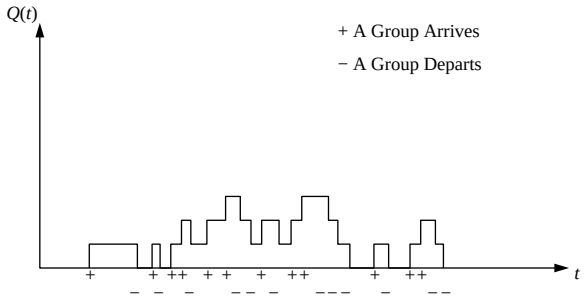


Figure 8.2: Arrivals and Departures at a Restaurant

In other words, we had to wait in line at Planet Hollywood^{TM} simply because of randomness. It's not that they didn't have enough staff, or that the staff worked too slow.

8.3 Queues 'R Us

How do we model these kinds of queueing systems? We usually start with the basic framework illustrated in Figure 8.3. Staying within the context of the two examples above, we have people arriving (e.g., at a grocery store or restaurant) at some rate. If there is a queue, they join it. Each person is then served somehow (e.g., they pay for their groceries or finish eating their meal), at which point they depart.

When the *arrival rate* and *service rate* are deterministic it is very easy to figure out how long the queue will be at any point in time, and hence, how long each person has

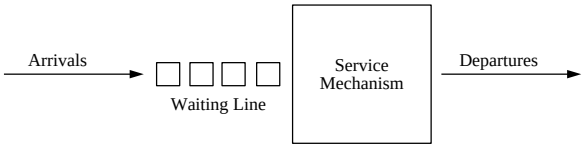


Figure 8.3: Waiting Lines

to wait in line. For example, if people arrive at a rate of 60 per hour and are served at a rate of 30 per hour then the queue will build at a rate of 30 people per hour. Hence, after 10 minutes there will be 5 people in the queue, after 30 minutes there will be 15 people in the queue, etc.... Obviously, as long as the arrival rate is larger than the service rate the queue will grow. Once the arrival rate is less than the service rate the queue can begin to shrink.

When arrivals and departures are random variables, on the other hand, things get a little bit more complicated. One way to think about this situation is illustrated in Figure 8.4. Here each of the circles represents a state of the system, which is to say the number of people in the system. The lines with arrows represent state transitions, and λ and μ are the arrival and service rates, respectively. In this figure we are assuming that only one person (or group) can arrive or depart at a time. Hence, we are implicitly looking at the state transitions during a relatively small time period.

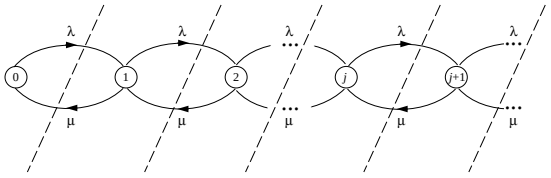


Figure 8.4: State Transitions in Queues

At arbitrary points in time it is relatively difficult to say anything about the number of people in line. However, we can say something about the steady states. In particular, let's suppose that the system is "stable" in the sense that:

$$\lambda \cdot P\{Q(t) = j\} = \mu \cdot P\{Q(t) = j + 1\} \text{ for all } j. \tag{8.1}$$

Intuitively, this says that the rate of flow across the dotted lines in Figure 8.4 is zero. In such steady states we can sometimes say something about sizes and times.

First, let's make some assumptions about the arrival and service rates. In particular, let's assume that the service times, which are random variables, are independent and have identical exponential distributions. Similarly, let's assume that the interarrival

times (i.e., the amount of time between arrivals), are independent and have identical exponential distributions. This means that all of the distributions are memoryless (loosely, that the past has no impact on the future).

Now, let's simplify the notation somewhat. Since we are in a steady state we needn't worry about t. So, let's just use P_j to denote the steady-state probability that the system is in state j. That is, P_j is the probability that there are j people in the system when the system has settled down. Then, we know from (8.1) that:

$$P_{j+1} = \frac{\lambda}{\mu} P_j \text{ for all } j. \tag{8.2}$$

In particular, this means that:

$$\begin{array}{rcl} P_1 & = & \frac{\lambda}{\mu} P_0 \\ P_2 & = & \frac{\lambda}{\mu}\left(\frac{\lambda}{\mu} P_0\right) = \left(\frac{\lambda}{\mu}\right)^2 P_0 \\ & \vdots & \\ P_n & = & \left(\frac{\lambda}{\mu}\right)^n P_0 \end{array} . \tag{8.3}$$

Or, letting $\rho = \lambda/\mu$ we have:

$$P_n = \rho^n \cdot P_0. \tag{8.4}$$

At this point we have everything we need – almost. We're still missing P_0, and all of our calculations depend on P_0. How do we figure out what it is? Well, we know that the sum of the probabilities has to be 1. That is:

$$P_0 + P_1 + P_2 + P_3 + \cdots = 1. \tag{8.5}$$

So, substituting in we get:

$$P_0 + \rho P_0 + \rho^2 P_0 + \rho^3 P_0 + \cdots = 1 \tag{8.6}$$

or:

$$P_0 \cdot \left(1 + \rho + \rho^2 + \rho^3 + \cdots\right) = 1. \tag{8.7}$$

Finally, this means that:

$$P_0 = \frac{1}{(1 + \rho + \rho^2 + \rho^3 + \cdots)}. \tag{8.8}$$

In and of itself this doesn't really do us much good since we have an infinite number of terms in the denominator of this fraction. However, it turns out that we know something about the infinite *geometric series* $1 + \rho + \rho^2 + \rho^3 + \cdots$. In particular, it turns out that:

$$1 + \rho + \rho^2 + \rho^3 + \cdots = \frac{1}{1 - \rho} \text{ if } \rho < 1. \tag{8.9}$$

Of course, now we have to ask whether ρ is, in fact, less than 1. The answer is that, while it needn't be, the problem isn't interesting if ρ is not less than 1. Why? Well, $\rho > 1$ when $\lambda > \mu$ (i.e., when the arrival rate is greater than the service rate). But, if this is the case then we know what happens – the queue continues to grow and there is no steady state. Thus, the only interesting cases are when $\lambda < \mu$.

Restricting ourselves to such cases, we see that $P_0 = (1 - \rho)$, and hence that:

$$P_n = (1 - \rho) \cdot \rho^n. \tag{8.10}$$

What about the expected number of people in the system in the steady state? That's easy, it's just $0 \cdot P_0 + 1 \cdot P_1 + 2 \cdot P_2 + \cdots$. With a little bit of algebra, you can show that this is simply $\rho/(1 - \rho)$ which is equal to $\lambda/(\mu - \lambda)$. What about the expected time spend in the system in the steady state? The expected number of people should equal the arrival rate times the expected time. Hence, the expected time in the system is just the expected number of people in the system divided by the arrival rate or $1/(\mu - \lambda)$.

What does all of this mean? Let's consider a simple example. Suppose that people arrive at a rate of 30 per hour and are served at a rate of 40 per hour. Then, the expected number of people in the system is just 30 / (40-30) = 30/10 = 3 people and the expected time in the system is 1 / (40 - 30) = 1/10 hours (i.e., 6 minutes).

Amazing, isn't it? The service rate is significantly larger than the arrival rate and yet the expected number of people in the system (in the steady-state) is 3 people and the expected time spent in the system is 6 minutes. All because of randomness!

8.4 Quotidian Queues

So far, we've only talked about two specific waiting lines. But, queues are quite common. In fact, we experience queues every day.[7] For example, we bump into traditional

[7] Hence the title of this section. You mean, you don;t know what "quotidian" means? I'm shocked, it's such an ordinary, everyday kind of word. In fact, if I remember correctly, my six year old niece Joy used it

waiting lines at:

- The bank;
- The barbershop;[8]
- The cafeteria;
- Toll booths;

  ⋮

In addition, there are many other real-world situations where queueing arises. For example:

The Judicial System. The court calendar is essentially a big queue. Cases arrive, enter a queue, and wait to be heard or settled.[9]

Secretarial Typing Pools. Jobs arrive at the typing pool, enter a queue, and wait to be completed.

Computer Operating Systems. Many computer operating systems can handle multiple users or multiple tasks "at the same time". In many cases, the computer can really only do one thing at a time, however it appears to be doing many things at the same time because it cycles through the different users/tasks. In essence, tasks are submitted to the central processing unit (CPU) of the computer, they sit in a queue, get partially processed, and then get returned to the queue.

Production Processes. Many production processes involve multiple machines and multiple operations. Work items arrive, enter queues, get completed and then move on to the next stage of the process.

A great many people spend a great deal of time trying to reduce the queueing delays inherent in these various situations. They have several things they have to decide, some of which can change over time.

First, they have to make decisions about the service mechanism. Specifically, they have to determine the number of servers, and the number of waiting lines, the service rate

---

the last time we played hang man.

[8]For those of you with hair, I'm referring to old-fashioned kinds of hair stylists. Those of us that are losing our hair don't need to have any styling done. While we are on this rather sensitive subject, I should tell you that I had shoulder length hair when I was a kid and my father was about 80% bald. I was unmerciful. Now I'm losing my hair and he just can't be happier. Is my family well-adjusted, or what?

[9]There's no reason to make snide comments about the judicial system. Both my brother Glen and my sister-in-law Anne are lawyers and they're nice people. Really! They are! I'm not just saying this for fear of being sued.

(if it is under their control). Obviously, if you increase the number of servers and/or the service rate the expected queue length and expected waiting time both go done. However, neither of these values can, in general, be increased costlessly.

What about the number of waiting lines? In essence, there are two possible approaches. You can either have as many waiting lines as you have servers or you can have one waiting line. These two approaches are illustrated in Figure 8.5 using two hypothetical toll plazas. In the top plaza there are three servers (i.e., toll booths) and three waiting lines (i.e., approach lanes). Vehicles enter a particular a lane and cannot change.[10] In the bottom plaza there are three toll booths but there is only one approach lane. In this case, vehicles wait in a single lane until a toll plaza becomes available and then proceed (in an orderly fashion).

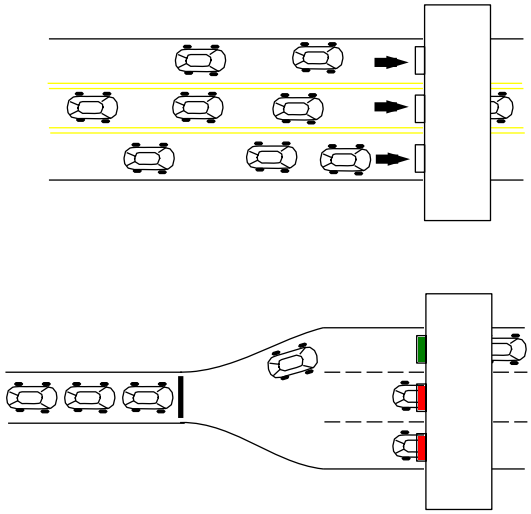


Figure 8.5: Changing the Number of Waiting Lines

Systems with a single waiting have an obvious advantage over those with multiple waiting lines – the servers are never idle when there is somebody in line. When there are multiple waiting lines a server might be idle while people are still in line because they're in another line.[11]

So why doesn't everybody use a single waiting line? Best I can tell it's a space issue. Figure 8.6 shows how much space is required in the waiting area when there are 9 people in 3 waiting lines (the top illustration) and 1 waiting line (the bottom illustration).

[10] Pretty realistic, huh? Just humor me!

[11] The other advantage is that when there is only one line you can't pick the "wrong line". Boy, that aggravates me! I get in the line with two people in it and it takes longer than the one with ten people in it. But, you keep thinking, "it has to get better, it has to get better, . . .".

Obviously, when there is only one waiting line, it is three times as long.

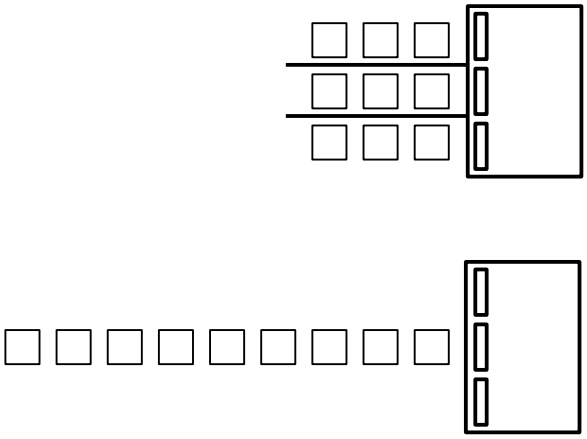


Figure 8.6: The Size of the Waiting Area

The only way to resolve this is to use a winding waiting line, as shown in Figure 8.7. This kind of line seems to trouble many of the store-owners that I've spoken to about it. They seem to think that winding lines are difficult to administer and look more unruly.

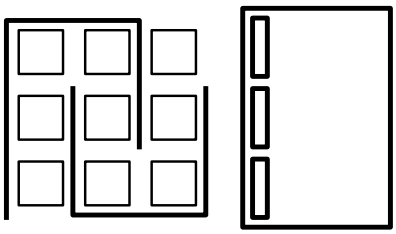


Figure 8.7: A Winding Waiting Line

In this regard, a decision also needs to be made about the maximum length of the line. To some extent, this may not be under the designers control, since people may "balk" when the line gets too long and leave the system. However, in a purely physical sense, the designer of a queueing system can force people to balk by not allowing a line to exceed a certain length. Examples include: the size of the waiting area, the number of people that can be placed on hold in a telephone-based customer service center, and the number of users allowed to login to a computer. Obviously, the trade-off here is the cost of allowing for a large queue versus the cost of turning away customers.

The final big decision that needs to be made when designing a queueing system is the *queueing discipline* for each waiting line. There are several obvious options, and some not so obvious options. One alternative, perhaps the one we see most often, is

the *first-come-first-served* discipline. Another is the *last-come-first-served* discipline. This is the discipline that gets used in after-hours video returns. You drop a video in the slot, someone else drops on one top of yours, etc. . .. The last one that gets dropped in the slot is the first one to be processed. This is also the way milk is purchased in grocery stores. People always seem to buy the carton with the latest possible expiration date, which is, of course, the carton that was delivered last. A third approach is *priority and/or preemptive service*. This is quite common in hospital emergency rooms. Cases are prioritized and handled accordingly. Life-and-death cases preempt all others. Another alternative is to use the *shortest remaining processing time* approach. I often fall back on this approach to process my list of daily "things to do". That is, I do the easy ones first. As you can imagine, each of these approaches has advantages and disadvantages in different situations.

Appendix A

My dog's smarter than your dog!

I REMEMBER it as if it happened only yesterday. I was seven years old and we had just moved into our new house. The kids across the street, Paul and Peter, came over to challenge me to a karate chopping contest. They humiliated me! In fact, to this day my hand hurts just thinking about it. In order to save my self-esteem, I took a quarter out of my pocket and said "Oh yeah? Well, I have more money than you do". Paul said he had 10 dollars in the bank. I then said that I had 20 dollars in the bank. He said 100. I said 1000. Paul, thinking he had me beaten, said he had infinity. Knowing I was beaten but with as sincere an expression as I could muster, I said I had two infinities. He said there's no such thing. I said there is. He said there isn't. I said there is. He said, he had three infinities. Well, you get the idea. It ended with my going into the house and wishing that we had never left Staten Island.

So, who was right?[1] How many numbers are there? How big is infinity? Well, it turns out, as silly as it sounds, that these questions come up several times throughout this book. So, just in case you've never thought about, or in case you have thought about it but forgotten what you concluded, I thought I'd spend a little time talking about numbers.

[1] Go ahead and admit it. You've had a similar conversation at some point in your life, haven't you? I still have them. In fact, I had one with my niece Ashley just the other day. So, sue us! We were at a wedding, we were bored. . .. It kept us amused for almost an hour!

A.1 How Big is Big?

If I were to ask you to count as high as you could, starting at zero, you'd probably think that I had lost my mind. Then, just to humor me (I could be dangerous after all), you'd say: 0, 1, 2, 3, 4, ... At some point (though it might take awhile)[2], I'd get bored and stop you. Now, why did you choose those numbers? I mean, why did you leave out 0.1, 0.2, 0.3, 0.4, . . .? For that matter, why did you leave out 0.01, 0.02, 0.03, 0.04, . . .? And how about numbers like 1/3, 1/9, 1/11? And why would you leave out your old friend π? (Doesn't look familiar? You probably remember it as pi. π is just the symbol for the Greek letter pi. Why use Greek letters? It makes you look smarter and, as I mentioned in the Preface, that's one of the most important things you'll get out of this book.)

Well, in the words of my old teacher Ms. Gunther, you decided to use the "counting numbers" (sometimes also called the "whole numbers"). You left out all of the other numbers – the numbers that Ms. Gunther called the "decimal numbers" and the numbers that she called the "fractions".

Now, before we go on, it's again time to make you feel and sound smarter. Instead of the term "counting numbers" we're going to use the word *integers*, instead of the word "fractions" we're going to use the word *rationals*, and instead of the term "decimal numbers" we're going to use the word *reals*. Why? Because these are the words that mathematicians use, and everybody knows that mathematicians are smart!

	1	**2**	**3**	**4**	**5**	**. . .**
1	1/1	1/2	1/ 3	1/ 4	1/5	
2	2/1	2/2	2/ 3	2/ 4	2/5	
3	3/1	3/2	3/ 3	3/ 4	3/5	
4	4/1	4/2	4/ 3	4/ 4	4/5	
5	5/1	5/2	5/ 3	5/ 4	5/5	
.						
.						
.						

Figure A.1: The Rational Numbers

[2] Just ask my wife, we've played these kinds of games. You know, it starts harmlessly enough. "I don't know why, but I'm tired", I'll say. "That's funny, I'm tired too" she'll say. "Oh really, well then I'm tired three", I'll say. Thinking I'll stop, she'll say "That may be, but I'm tired four". But, I won't stop. I'll go on for ever. And it's not just stubbornness, either. I actually enjoy it.

So, back to the questions at hand. What's the biggest number? How many numbers are there? Are there the same number of integers, rationals, and reals?

First, let's start with the integers. What's the biggest integer? There isn't one! You can always take any integer and add 1 to it. What about infinity? Well, infinity isn't an integer (or a real or a rational for that matter). We just use the term 'infinity' to indicate that there is no largest integer (i.e., we say that there are an infinite number of integers).

Now, how about the reals? Since every integer is also real, it follows that there are at least as many reals as there are integers. But, somehow, it feels like there are a lot more reals than integers, even though there are an infinite number of integers. I mean, just think about how many reals there are between every two integers (say between 0 and 1). There are an infinite number! That means that if I could somehow start "counting" the reals at zero and worked my way up I wouldn't ever get to 1. Even worse, it's hard to imagine how I would even count the reals! Starting at 0, what's the next number you would "count"? If you said 0.1, I'd ask why you left out 0.01. If you said 0.01, I'd ask why you left out 0.001. In fact, no matter how hard you try, you can't come up with the "next" real after 0. Why not, because there isn't one! (I'll talk more about this shortly.)

For this reason, we like to distinguish between the number of reals and the number of integers. Specifically, we say that there are a *countable infinity* of integers and an *uncountable infinity* of reals. In other words, though you can't actually count all of the integers, you can at least describe a process for doing so. On the other hand, there are so many more real numbers than integers, that it's not even possible to describe a process for counting them all.

	1	**2**	**3**	**4**	**5**	**. . .**
1	1/1 →	1/2 →	1/ 3 →	1/ 4 →	1/5 →	
2	2/1	2/2	2/ 3	2/ 4	2/5	
3	3/1	3/2	3/ 3	3/ 4	3/5	
4	4/1	4/2	4/ 3	4/ 4	4/5	
5	5/1	5/2	5/ 3	5/ 4	5/5	
.						
.						
.						

Figure A.2: A Bad Way to Try and Count the Rational Numbers

What's the practical significance of all of this? Perhaps most importantly, if you ever bump into my old friend Paul and he says he has an infinite number of dollars in the bank, you can say "Oh yeah, well I have an uncountably infinite number of dollars in

the bank" and you'll win. In addition, it turns out that because of the different properties of the reals and integers, it is sometimes "easier" to make decisions involving integers and other times it is "easier" to make decisions involving reals. Don't believe me? Read the rest of the book!

But wait!!! You're thinking that I've forgotten about the rationals. Actually, I haven't, they just don't play a very big role in this book. So, I was going to ignore them. But, since you've brought it up...

How many rational numbers are there? Well, let's see. Every integer is a rational number (since $1 = 1/1$, $2 = 2/1$, $3 = 3/1$, etc...), so there are at least as many rationals as there are integers. In addition, there are a whole bunch of rationals that are not integers (e.g., $1/2$, $1/3$, $1/4$, etc...). So, there must be more rationals than integers. Right? Wrong!

Remember, the real question we need to ask is whether we can describe a process for counting the rationals. If so, there are a countable infinity of them, if not there are an uncountable infinity of them. So, can we describe a process for counting the rationals? At first glance, it seems like we can't.

First, let's put all of the rational numbers in a table. In particular, label the columns with all of the integers, label the rows with all of the integers, and let the cells be the row label divided by the column label. This is illustrated in Figure A.1. Though there will be some duplicates, this will clearly give us all of the rationals.

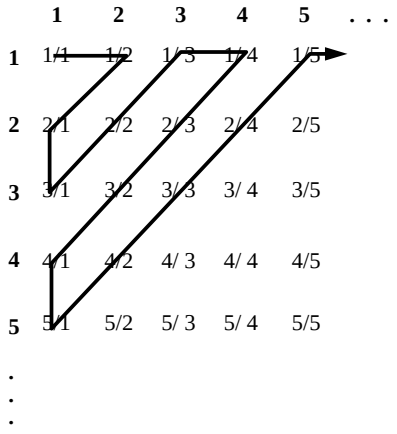


Figure A.3: A Good Way to Count the Rational Numbers

Now, about counting them. The most obvious way to count would be to start at row 1, column 1 (R1C1) and start counting across. This is illustrated in Figure A.2. The problem with this approach is that you'll never finish row 1. Hence, you can't possibly count all of the rationals this way.

But, it turns out that there is a way to count them. Again, start at row 1, column 1. Then, go to R1C2. After that, go to: R2C1, R3C1, R2C2, R1C3, R1C4, etc. . .. This is illustrated in Figure A.3. At least in principle, it is possible to count the rationals this way. Thus, there are the same number of rationals as there are integers, a countable infinity.

A.2 How Small is Small?

Not long after the karate chopping incident, on a lazy summer afternoon, my new-found friend Paul and I had the following conversation:

Paul: What do you want to do?

Me: I don't know, what do you want to do?

Paul: **I** don't know, what do **you** want to do?

Me: I don't **know**, what do you want to **do**?

Mom: **Knock it off and find something to do!!!!**

Me: OK, mom!

Me: So, what do you want to do?

Anyway, I'm sure this went on for quite some time until I finally suggested that we pitch pennies. It turned out that Paul didn't know what I was talking about. So, being a street-wise city kid that had just moved out to the suburbs, I pounced on my opportunity and explained the rules. We get a bunch of kids together on some pavement and toss pennies at a wall.[3] After everybody tosses, the person whose penny is closest to the wall gets to keep all of the pennies.

Thinking this was a good idea, we gathered up Pete, Matt, Tommy and some other kids and started playing. Things were going fine until, on one round, Paul and I both tossed pennies really close to the wall. So, as any two kids would do, we decided to measure. Paul said his penny was only 1/10 of an inch from the wall. I said my penny was only 1/100 of an inch from the wall. He said his was only 1/1000 of an inch away. I said 1/10000. He said 1/100000. Anyway, you get the idea.[4]

[3] Actually, I said that you get a bunch of kids together on the sidewalk and toss pennies at the stoop. Paul then asked what a 'stoop' was. I explained that stoops are the stairs leading into your building. He then asked where we could find a sidewalk. I then realized that I had moved from Staten Island to some foreign country.

[4] Needless to say, throughout this whole discussion, we only had a ruler that went down to eighths of an inch!

Now, for some reason, neither one of us ever said 'one infinity-eth' of an inch.[5] Instead, it ended with us fighting over whether 'one b-zillionth' was a number or not and, if it was, whether it was bigger or smaller than 'one quadrillionth'.

Of course, that fight was never resolved, but it did start me thinking about finding the closest number to zero. It never even occurred to me that there might not be a closest number to zero. However, it is now easy to see that there isn't one. No matter what (positive) number you say, no matter how close to zero it is, I can always divide it by 2 (or 3, or 10, or . . .) and find a smaller number.

A.3 Numbers Unlimited

If there isn't a biggest number, and there isn't a smallest number, and there isn't a number closest to zero (or closest to any number, for that matter), how should you talk about really, really big numbers or really, really small numbers? One way to do it is to use *limits*.[6] When using limits you ask yourself what happens as some number gets larger or what happens as some number gets smaller. In other words, you consider a *sequence* of events.

If you don't understand, an example should clear things right up. What happens to the fraction $1/n$ as n gets larger and larger? Well, starting at $n = 1$ (and assuming that n is an integer), as n gets larger and larger we get a sequence of numbers $1/1$, $1/2$, $1/3$, $1/4$, ... Thus, what happens as n gets larger and larger is that $1/n$ gets smaller and smaller. In fact, $1/n$ gets closer and closer to zero.

Congratulations! You've just taken your first limit. How would you write this down? There are two ways. One way is to write: if $n \to \infty$ then $1/n \to 0$ (i.e., if n goes to infinity then $1/n$ goes to zero). Another way is to write:

$$\lim_{n\to\infty} \frac{1}{n} = 0 \tag{A.1}$$

which says that the limit of $1/n$ as n goes to infinity is zero.

So, how do you talk about situations where some number is getting really large or really small? You talk about what happens "in the limit". So, where would we have gotten if Paul and I had continued our fight about who's penny was closest? In the limit, we would have gotten to zero! That is, we would have kept saying numbers that got closer and closer to zero, even though we never would have said zero (since our pennies were clearly not touching the wall).

[5] I like to think it's because we had matured in the four days since the karate chopping incident.

[6] Take my word for it, using limits really makes you look smart. Even the notation looks cool!

Prefix	Power of 10	Description	Number
kilo	3	thousands	1,000
mega	6	millions	1,000,000
giga	9	billions	1,000,000,000
tera	12	trillions	1,000,000,000,000
peta	15	quadrillions	1,000,000,000,000,000
exa	18	quintillions	1,000,000,000,000,000,000
zetta	21	sextillions	1,000,000,000,000,000,000,000
yotta	24	septillions	1,000,000,000,000,000,000,000,000

Table A.1: Big Numbers

Prefix	Power of 10	Description	Number
milli	-3	thousandths	0.001
micro	-6	millionths	0.000001
nano	-9	billionths	0.000000001
pico	-12	trillionths	0.000000000001
femto	-15	quadrillionths	0.000000000000001
atto	-18	quintillionths	0.000000000000000001
zepto	-21	sextillionths	0.000000000000000000001
yocto	-24	septillionths	0.000000000000000000000001

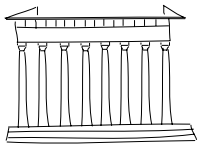
Table A.2: Small Numbers

A.4 That does not Compute

One final thing about numbers and computer manufacturers. Computer manufacturers just love to use coolwords to describe really big numbers and really small numbers. There are kilobytes and megabytes, milliseconds and microseconds. What does all of this mean? These numbers are explained in Tables A.4 amd A.4.

Appendix B

It's All Greek To Me



IT'S USUALLY a good idea to use single letters like a or b (set in a special math-italic typeface) to represent variables in equations. Since we distinguish between lowercase and uppercase letters (i.e., A is different from a) this means that you can distinguish between 52 different variables using the English alphabet.

In general, that's more than enough. In fact, if you use more than 52 different variables in any one paper you're probably doing something wrong. Who could possibly keep track of all of those variables?

However, even though the English alphabet (really the Roman alphabet) provides more than enough variable names it is very common to use Greek letters as well (and even the occasional Hebrew letter). Why? Most people working in mathematical fields will tell you that they use the different alphabets to distinguish between different types or categories of variables.

I'm not convinced!! I'm pretty sure that people use Greek letters because it makes them look smarter. I know that's why I use them.

With that in mind, Table B.1 provides you with a list of Greek letters and how they are spelled in English.

"How are they pronounced?", you ask. That's a good question. It turns out that we use the Roman pronunciation of Greek letters, not the Greek pronunciations. So, for example, we pronounce β as 'bay-tuh' whereas the Greeks pronounce it as 'bee-tuh'. Pretty weird, huh?

Lowercase Letter	Uppercase Letter	English Spelling
α	A	alpha
β	B	beta
γ	Γ	gamma
δ	Δ	delta
ϵ or ε	E	epsilon
ζ	Z	zeta
η	H	eta
θ or ϑ	Θ	theta
ι	I	iota
κ	K	kappa
λ	Λ	lambda
μ	M	mu
ν	N	nu
ξ	Ξ	xi
o	O	omicron
π or ϖ	Π	pi
ρ or ϱ	R	rho
σ or ς	Σ	sigma
τ	T	tau
υ	Υ	upsilon
ϕ	Φ	phi
χ	X	chi
ψ	Ψ	psi
ω	Ω	omega

Table B.1: Greek Letters