Section 1.0 Basic Probability and Randomness

In order to fully understand probabilistic modeling, you will need a good sense of the nature of ***random*** behavior. All of us have basic intuition about probability and randomness. This intuition develops over time from our experiences. These include what we watch on television, observe at sporting events, and read on the Internet. The problem is that research has shown that our intuition about randomness and probability is often flawed. For example, a manager being told that each and every one of his 10 suppliers is 95% reliable often believes the system he has in place is in good shape. In this chapter, you will learn why the manager needs to be concerned.

Because of the many misconceptions, the development of probability analysis skills requires demonstrating the flaws in our intuition. Only after this is accomplished, can you move on to learn formal approaches to probability. The first example we explore involves understanding what is often our preconceived notion of a random pattern. Specifically, we will ask you to imagine and write down a typical sequence of heads and tails when flipping a coin. You and your classmates will then be asked to actually flip a coin. By comparing the lists, we hope to dislodge a misconception you may have about random events.

One of the primary goals of this section and all future sections is to begin understanding how much **variability** there may be in a simple random process. This will be accomplished by comparing and contrasting the results of your coin flipping experiment with those of your neighbors and those of the class as a whole. For example, we will explore how much difference there is in the percentage of heads and tails in your list and those of each and every one of your classmates. We will then look at the average for the entire class.

One challenge managers and we personally face in a variety of situations involves understanding and interpreting fluctuations around a long-term average. Is a rise in absenteeism or lateness simply a reflection of a random fluctuation or an indication of a developing problem that needs to be addressed? Biostatisticians regularly are on the lookout for occurrences of disease that are out of the “normal” range. A report comes in that a small city has twice the national average rate of a specific type of cancer. There are tens of thousands of communities. Is it likely or unlikely purely by random fluctuation to see one or more cities with a rate twice the national average? If it is likely, no action should be taken, no detailed study initiated. If however, probability theory suggests this high rate for a small city is extremely rare, public health officials may commission an expensive study to gather more data to try to identify the contributing factors.

You will explore the pattern of which conference won the Super Bowl in an attempt to draw a conclusion about whether the data suggest that one conference is superior to the other. You will also look at data for a customer service telephone line to determine whether there is sufficient evidence that the service is not meeting the timeliness standard established by management. You will assist the faculty advisor to the school newspaper to better understand why they are having problems publishing the paper on time. Finally, you will assist a plant manager to decide how many spare workers to have on duty. These are needed to maintain productivity while coping with random fluctuations in the number of workers who are absent from work each day.

A key question you will always be addressing is, "What is the source of variability?" The fluctuation around the mean may be merely random variation or it could be the result of some external influence. Discerning whether the pattern of variation is random or due to some factor we might be able to control is a critical component of probabilistic and statistical analysis. We need these tools because, in trying to make this distinction, our own intuition often leads us to unfounded conclusions.

Section 1.1 Super Bowl – Conference Dominance

Every year since 1967, the American Football Conference (AFC) and National Football Conference (NFC) champions of the National Football League compete in a championship game known as the Super Bowl. Table 1.1.1 identifies by conference the winner of each of the 48 Super Bowls from I through XLVIII. The question is:

Super Bowl					
Year	**Winning Conference**	**Year**	**Winning Conference**	**Year**	**Winning Conference**
1967	NFC	1983	NFC	1999	AFC
1968	NFC	1984	AFC	2000	NFC
1969	AFC	1985	NFC	2001	AFC
1970	AFC	1986	NFC	2002	AFC
1971	AFC	1987	NFC	2003	NFC
1972	NFC	1988	NFC	2004	AFC
1973	AFC	1989	NFC	2005	AFC
1974	AFC	1990	NFC	2006	AFC
1975	AFC	1991	NFC	2007	AFC
1976	AFC	1992	NFC	2008	NFC
1977	AFC	1993	NFC	2009	AFC
1978	NFC	1994	NFC	2010	NFC
1979	AFC	1995	NFC	2011	NFC
1980	AFC	1996	NFC	2012	NFC
1981	AFC	1997	NFC	2013	AFC
1982	NFC	1998	AFC	2014	NFC

Table 1.1.1: Super Bowl winners by conference

What is it about the sequence of conference wins that has led most sports writers to believe that one conference or the other was superior at different extended periods in time?

1. What do you notice about these results?

2. Does one conference appear stronger than the other? Why or why not?

3. If neither conference is superior to the other, what number would it make sense to use as the probability that the winner will come from a given conference?

4. What was the overall percentage of wins by each conference? Each column represents a 16-year period. What was the winning percentage in each 16-year period?

In order to explore your conception of randomness, we will model the Super Bowl results since 1967. Assume that the two conferences are equally strong. If so, each conference winner has the same 50% chance of winning the Super Bowl in any year. We will use a coin flip to represent the 50% chance. First you will imagine and write down a typical sequence of heads and tails. Only afterwards, will you actually flip a coin.

5. On a sheet of paper, list the numbers from 1 to 48 to represent each of the Super Bowls from 1967 to 2014. Table 1.1.2 shows how to set up the list before entering the Hs and Ts. You will use an H to represent an AFC win and a T for an NFC win. Now imagine a random sequence of wins assuming the two conferences are equally strong. Record next to each number, an H or a T.

1		17		33	
2		18		34	
3		19		35	
4		20		36	
5		21		37	
6		22		38	
7		23		39	
8		24		40	
9		25		41	
10		26		42	
11		27		43	
12		28		44	
13		29		45	
14		30		46	
15		31		47	
16		32		48	
Number of Heads		Number of Heads		Number of Heads	
Percent of Heads		Percent of Heads		Percent of Heads	

Table 1.1.2: A format for recording simulated coin flips

After recording all 48 Hs and/or Ts, fill in the bottom rows of the table: the number of heads and percent of heads in each column. Also calculate the percent of Hs for all 48 imagined events.

6. Compare the percent of heads in your three columns and the overall percent. How much difference is there among the column percentages?

7. Compare the percent of heads in your list with those of two or three other students sitting close by. Are your results similar, or is there a lot of difference among them?

Next, make another list just like the one in Table 1.1.2 to record the results of your actual coin flips. After recording all 48 Hs and/or Ts, fill in the bottom rows of the table: the number of heads and percent of heads in each column. Also calculate the percent of Hs for all 48 flips.

8. Compare the percent of heads in your three columns and the overall percent. How much difference is there among the column percentages?

9. Compare the percent of heads in your list with those of two or three other students sitting close by. Are your results similar, or is there a lot of difference among them?

In the following questions, you will compare your imagined sequence with the sequence of the actual coin flips. Be sure to label your lists, so that you'll know which is which.

10. Was there more variability in the column percentages among the imagined lists or the coin flip lists?

11. What were the maximum and minimum percentages for you and your neighbors for each column for the imagined lists and coin flip lists? Are you surprised at how far these numbers are from 50% in the actual coin flip data?

12. What were the maximum and minimum percent of Hs and Ts for you and your neighbors for all 48 imagined events and coin flips? What differences do you notice about the ranges?

13. For the entire class, what was the minimum percent of Hs observed when flipping the coins 48 times? What was the maximum? Are you surprised at how far these numbers are from 50%?

14. Did anyone in the class have a percentage that low or that high in the imagined list?

15. For the entire class, what was the overall percent of Hs and Ts for the coin flip random experiment?

We will now return to the actual data for the Super Bowl in Table 1.2.1. Find the total the number of times in each of the three columns that the NFC won and calculate the percent.

16. Look at the percentage of NFC wins in the first column of the actual Super Bowl data. In your coin flip table, was there any percentage of Hs or Ts that low? Did anyone in the entire class have a percentage of Hs or Ts that low in a column of coin flips?

17. Look at the percentage of NFC wins in the second column of the actual Super Bowl data. In your coin flip table, was there any percentage of Hs or Ts that high? Did anyone in the entire class have a percentage of Hs or Ts that high in a column of coin flips?

18. Complete the Table 1.1.3 for the class as a whole. In this table we want the frequency of either a large number of Hs or Ts. In the summary of the 16 year periods, we would be surprised if either of the divisions won an unusually large proportion of times. Thus, in assessing the likelihood of one conference winning, for example, 11 or more times in 16 years, we need to consider both Hs and Ts.

In order to complete the table we will need the total number of times a set of 16 coin flips were recorded by the students in the class. It should be three times the number of students. The actual frequency of each outcome listed in Table 1.1.3 is then divided by this total to determine thhe relative frequency.

> The ***relative frequency*** of an event is equal to the ratio of the number of times the event occurred to the total number of observations.

	Record the class' total number of sets of 16 coin flips =					
	Large number of Hs		Large number of Ts		Hs and Ts Combined	
Number of Hs or Ts	Frequency	Relative Frequency	Frequency	Relative Frequency	Frequency	Relative Frequency
11						
12						
13						
14						
15						
16						

Table 1.1.3: Class frequency of high percentages

19. Based on Table 1.1.3, what is the likelihood that one conference would win 11 or more Super Bowls in a 16 year period? 12 or more? 13 or more? 14 or more?

20. Based on your comparisons between your actual coin flips and the Super Bowl wins, do the percentages strongly support the news writers' claims of dominance in any of the 16 year periods?

Up to this point we have explored different aspects of the percent of Hs and Ts. We have compared the imagined sequence with that generated by actual coin flips. We have also compared minimum and maximum percentages for columns and for the total. By including your classmates, you were able to consider larger and larger sets of data. For the class as a whole, you should have seen a wider range of percentages for the column percentages than you found in your data and your neighbor's. In addition, there should have been less variability in the total percent of conference wins over 48 years. The total percent for the entire class would likely be close to 50%.

Now we are going to look at the sequences of consecutive Hs or Ts. These sequences are called strings. There have been numerous class room experiments comparing made up lists and randomly generated lists. In general, the lists of imagined sequences of Hs and Ts tend to have shorter strings than randomly generated sequences. In addition, imagined lists rarely have any long strings of four more. We are going to see if your experience and those of your classmates replicate the research.

Record the lengths of every string of consecutive heads or tails in your imagined list. Then do the same for your coin flip list. For example, in Table 1.1.4, the first column records a sequence of Hs and Ts. The second column records the length of each string. In Table 1.1.4 the longest string is 3.

T	2
T	
H	1
T	1
H	2
H	
T	1
H	3
H	
H	
T	3
T	
T	
H	2
H	
T	1

Table 1.1.4: Length of strings for imagined list

21. Compare the lengths of the strings in your two lists. Which list has the longest string? How long is that string? Compare your answers with those of two or three neighbors. In your group, what is the length of the longest string? Which list did it come from?

22. Compare your answers with the entire class. What is the length of the longest string? Which list did it come from?

23. In the Super Bowl results, what was the longest string of wins for one conference? Did any of the coin flip lists have a string this long?

24. Based on your comparisons do the data support the sports writers' claims of dominance?

25. Have you discovered any misconceptions in your own thinking about randomness?

1.1.1 Percentage, Long Strings, and Probabilistic Thinking

Every alternative sequence of 48 heads and tails has exactly the same probability of occurrence, one in 281.5 trillion. Thus the likelihood of seeing any particular sequence of alternating heads and tails is the same as the probability of seeing a sequence of all heads or all tails. However, the likelihood of producing exactly 24 heads out of the total of 48 is orders of magnitude higher than the chance of 48 heads in a row. That is due to the fact that there are many sequences of heads and tails that result in exactly 24 heads, but there is only one way to obtain 48 heads in a row. Thus, 48 heads in a row might lead you to believe that the coin is two-headed.

This same principle points to an interesting phenomenon. When random flips of a coin are carried out, the sequence often includes multiple examples of long runs of four or more heads or four or more tails. However, individuals recording their own made-up sequence have a tendency to believe that randomness means frequently alternating back and forth between heads and tails. Consequently, they tend to create lists that have few long strings of heads or tails. As a result, an experienced probabilistic thinker can often identify which of the two lists was generated by flipping a coin and which was generated by a human mind. The human mind tends to look at the list as it develops and work to balance the distribution of heads and tails.

One of the primary goals of probabilistic analysis is to attempt to assess whether a pattern is simply random or if some factor is contributing to the pattern. It is never possible to resolve this issue with certainty, but an understanding of probability theory can help decide which driving force is more likely.

Section 1.4 Getting *The Lancer* to Press

Each month Al Mitchell, the faculty advisor for a school newspaper, oversees the production of the newspaper, *The Lancer*. He has 10 student writers, five of whom are editors and five are staff writers. In order to get to press on time, it is necessary that Al's students finish their articles by the required deadline. That is, he needs each and every student to meet his or her deadline.

1.4.1 The Problem

Over the years, Al has found that that student writers meet their deadlines only 95% of the time. On the surface, this does not seem bad, but *The Lancer* frequently goes to press late. He wants to change that. Al recalls something from his high school probability class called the ***multiplication principle***: The probability that a set of ***independent*** events will occur can be found by multiplying the probabilities of each event occurring. Formally stated, the multiplication principle says that:

If X and Y are *independent* events, then the probability of both X and Y occurring is equal to the product of the probabilities of X occurring and Y occurring.

$$P(X \cap Y) = P(X) \cdot P(Y)$$

Al wonders what independent events would mean in this context. One particular writer meeting a deadline does not seem to affect the likelihood of any other writer meeting the deadline. Thus, the events that writers X and Y meet the deadline are *independent*. Therefore, the multiplication principle can be used. If each of his 10 writers meets the deadline 95% of the time, the probability that all of them meet the deadline is given below.

$$\begin{aligned} P(\text{going to press on time}) &= (0.95)(0.95)(0.95)(0.95)(0.95)(0.95)(0.95)(0.95)(0.95)(0.95) \\ &= 0.95^{10} \\ &\approx 0.599 \end{aligned}$$

Now, for the paper to go to press on time, every one of the 10 writers must meet the deadline. If even one of the ten does not meet the deadline, the printing of the paper is delayed. This calculation shows that *The Lancer* would go to press on time approximately 60% of the time! No wonder it has been going to press late so often. Al wants to change this situation to ensure the paper goes to press on time 90% of the time.

Each writer meeting deadlines 95% of the time seems to be a reasonable expectation. However, having each of the10 students meet the deadlines 95% of the time results in a big problem for *The Lancer*. To improve this situation, Al believes he can get his editors to meet the deadlines 99% of the time. They have the most experience and are likely to work better under pressure compared to the staff writers who have just joined *The Lancer*'s staff.

Al's idea is to have his five editors meet deadlines 99% of the time, and the less experienced staff writers can continue to meet deadlines 95% of the time. The multiplication principle allows Al to calculate the probability of going to press on time.

$$\begin{aligned} P(\text{going to press on time}) &= (0.99)(0.99)(0.99)(0.99)(0.99)(0.95)(0.95)(0.95)(0.95)(0.95) \\ &= (0.99)^5 \cdot (0.95)^5 \\ &\approx 0.736 \end{aligned}$$

Going to press on time about 74% of the time is far better than the 60% it was previously. Nevertheless, he would still not achieve his 90% goal.

Al goes back to the drawing board and believes that he can push his editors to be on time 100% of the time. He feels the editors can handle that level of responsibility. If the five editors meet the deadlines 100% of the time and the staff writers meet the deadlines 95% of the time, the probability of going to press on time is shown below.

$$\begin{aligned} P(\text{going to press on time}) &= (1)(1)(1)(1)(1)(0.95)(0.95)(0.95)(0.95)(0.95) \\ &= (1)^5 \cdot (0.95)^5 \\ &\approx 0.775 \end{aligned}$$

Getting the editors to meet their deadlines 100% of the time barely made a difference!

Al sees that the only way to get *The Lancer* to press on time is to push also the staff writers to meet their deadlines at a higher rate. If the editors meet the deadlines 100% of the time, what must the probability be for the staff writers in order for *The Lancer* to not be delayed 90% of the time?

1. If each staff writer meets her/his deadline 96% of the time, would The Lancer go to press on time 90% of the time? If not, what if the staff writers meet their deadlines 97% of the time?

Al begins to wonder whether it is reasonable to have a goal of 90% for *The Lancer* to go to press on time. If the editors meet their deadlines 100% of the time, Al wonders what standard should be required of each staff writer. We can take an algebraic approach to determine this standard.

Let x represent the standard for each staff writer meeting the deadline. The probability of the paper going to press on time at least 90% of the time can be represented by the following inequality. Then we need to solve the inequality for x.

$$
\begin{aligned}
P(\text{going to press on time}) &\geq 0.90 \\
(1)(1)(1)(1)(1)(x)(x)(x)(x)(x) &\geq 0.90 \\
1^5 \cdot x^5 &\geq 0.90 \\
x^5 &\geq 0.90 \\
x &\geq \sqrt[5]{0.90}
\end{aligned}
$$

Now, take the fifth-root of a number, which is equivalent to raising the number to the $\frac{1}{5}$ power. This can be done with a calculator, so

$$
\begin{aligned}
x &\geq (0.90)^{\frac{1}{5}} \\
x &\geq 0.979
\end{aligned}
$$

The solution shows us that for *The Lancer* to go to press on time 90% of the time, each of the five staff writers would have to make deadline almost 98% of the time! Al wonders if this is a realistic goal. In the next chapter, Al will evaluate a change in policy. He will consider going to press on time if one or two writers are late. He will simply leave their columns out of the paper.

1.4.2 Complementary Events

Let's consider the timely publication of *The Lancer* from a slightly different perspective. Instead of focusing on how frequently *The Lancer* goes to press on time, let's consider how often it goes to press late.

Going back to the original case, the editors and staff writers originally each met the individual deadlines 95% of the time. This led to *The Lancer* going to press on time roughly 60% of the time.

2. That being the case, how often did *The Lancer* go to press late?

If they were on time 60% of the time, it makes sense to say that they were late 40% of the time. Mathematically, this happens because going to press on time and going to press late are ***complementary events***. Complementary events are both ***mutually exclusive*** and ***collectively exhaustive***. If two events are mutually exclusive, that means they cannot occur at the same time. For example, when rolling a die, the event of getting a three and the event of getting a five are mutually exclusive. However, getting a 3 and getting an odd number are not mutually exclusive. Two events are collectively exhaustive if together they cover all possible outcomes. This implies that one of them must occur. For example, when rolling a die, the event of rolling an odd number and the event of rolling an even number are collectively exhaustive. However, rolling an odd number and rolling a 2 are not collectively exhaustive.

3. Keeping with our die-rolling context, list three pairs of complementary events.

When we consider publishing a student newspaper, one pair of complementary events is publishing on time or not publishing on time (i.e., publishing late). Because an event and its complement include all the possible outcomes of an event, the probability that one or the other of them will occur is equal to one (i.e., one of them is certain to occur). Now, because they cannot both occur (why not?), you can find the probability of an event's complement by subtracting the probability of the original event from one. This is significant because calculating the probability of the complement of an event is sometimes easier than finding the probability of the event directly. Thus, in the case of *The Lancer*,

$$P(\text{publishing on time}) + P(\text{not publishing on time}) = 1$$
$$P(\text{not publishing on time}) = 1 - P(\text{publishing on time})$$

You might think that all of the writers on time and all of the writers not on time are complementary events. All of the writers on time corresponds directly to the event that all ten writers meet their individual deadlines. The complementary event that the paper is late is not equivalent to ***all*** writers being late. It only takes ***one*** writer being late for publication of *The Lancer* to be delayed. Thus, the event the paper is late includes all of the following possibilities: 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 writers being late. Calculating directly the probability of the paper being late requires calculating the probability of each of these one through ten possibilities. The calculation of each of these is in itself complex and requires the advanced concepts discussed in the next chapter.