Key Ideas

- For a while, instead of talking about statistics, we'll be talking about probability. The two subjects are very closely related; there is little you can do in one without the other. So, we're going to learn some basic probability and then later turn to statistics to apply what we've learned.

- Abstraction: Unfortunately, the definitions in probability theory can seem a bit abstract. However, you ought to understand that this is not done simply to be annoying. The ability to manipulate abstract mathematical concepts is an important part of mathematical reasoning. I could argue (but won’t waste your time in class) that without such abstraction we would not have cell phones or computers or airplanes or satellites or...well, you get the picture.

- Trials and Outcomes: In basic probability we consider a trial to be a specific description of something that is going to happen in which it is not certain what will happen. Flipping a coin or giving a patient with cancer a risky treatment are two examples. The different things that can happen are called outcomes and we generally label them with capital letters. Note that if we talk about repeating a trial, we will assume that the likelihood of any given outcome remains the same from one trial to another. For example, the likelihood that a flipped coin lands on heads is not affected by whether it landed on heads in previous flips. This is known as independence and it is contrary to many people’s intuitions, so please keep it in mind.

- Sample Space: The set of all possible outcomes for a trial is called its sample space. That is, if there are three possible outcomes for a trial – A, B and C – then the sample space is the set \( \{A, B, C\} \). It is important to note that when the trial is done, exactly one of those outcomes (no more no less) will happen. If that is not the case, then you have not really specified the sample space properly.

**Question 1:** What are the sample spaces for these trials, and how large are they?

a) A patient in a coma is given a drug that sometimes wakes them.
b) A card is selected from a standard deck.
c) A student takes a quiz with three questions on it. The student’s answer is either correct or incorrect. We can then use a string like “CCI” to indicate that the student got the first two right but not the third one. What is the sample space and how large is it?

- For (c) above, it may be useful to draw a tree diagram such as the one on page 222 of your textbook.

(What if the trial is “A bottle is spun and we record which direction it is pointing”? That is beyond the scope of this chapter of the book, where we consider only finite sample spaces. We will see that probability works just fine for infinite sample spaces in Chapter Six.)

- Events: Perhaps the most abstract thing we’ll be considering is the notion of an event. According to its definition, an event is just a subset of the sample space. If the event is a single element of the sample space then we say it is a simple event. An event, like an outcome, is often labeled by a capital letter, but we usually describe an event in words.
For example: If the sample space for selecting a playing card from a regular deck is \( \{A\heartsuit, 2\heartsuit, 3\heartsuit, \ldots, Q\clubsuit, K\clubsuit\} \) then a simple event \( E \) would be something like \( E = \{2\spadesuit\} \), which is “the card selected is the two of spades”. But a non-simple event would be \( S \), “the card selected is a spade”. As a subset it would look like \( S = \{A\spadesuit, 2\spadesuit, 3\spadesuit, \ldots, Q\spadesuit, K\spadesuit\} \). (Note that this is a subset of the sample space.)

- **The Complement of an Event:** Given any event, \( E \), we can talk about its complement which we denote by \( E^c \). This is, by definition, just the set of all outcomes that are not in \( E \). (So, continuing the example above, \( E^c \) would be the set of all of the outcomes except for \( 2\spadesuit \) and \( S^c \) would be the set of all outcomes for which the suit is a club, heart or diamond.)

- **Probability:** We can assign a number between 0 and 1 to an event. This number reflects the likelihood of that event occurring with 0 reflecting that the event is certain not to happen and 1 being the case in which it is guaranteed to occur. According to the definition on page 218:

  The probability of a particular outcome is the proportion of times that the outcome would occur in a long run of observations.

  (This is mathematically justified by Jacob Bernoulli’s 1689 rigorous proof of “the law of large numbers”.)

- **Finding Probability of an Event if Outcomes are Equally likely:** If each outcome in the sample space is equally likely to occur, then the probability of event \( A \) is:

\[
P(A) = \frac{\text{Number of outcomes in } A}{\text{Number of outcomes in sample space}}
\]

**Question 2:** What is the probability of selecting an ace from a regular deck of cards?

- **Empirically Determined Probabilities:** Because of the definition above, if the frequency of outcomes in the sample space is known for a sufficiently large series of prior trials, then the probability of the event \( A \) can be estimated as

\[
P(A) = \frac{\text{Frequency of event } A}{\text{Number of trials conducted}}
\]

**Question 3:** A study followed the health of 896,577 randomly selected Americans through their entire lives. It found that 206,244 of them died from some form of cancer. Based on this, estimate the probability that an American will die from cancer.
The book also mentions “subjective probability”, but I warn that this is often “wrong probability”. A famous example is the “Monty Hall Problem” (see http://kasmana.people.cofc.edu/MATHFICT/montyhall.html) where intuition disagrees with both theory and practice. A more personal example resulted in a paper that I wrote with my wife (J. Virology 76 11 (2002) 557-5564) in which we showed that biologists intuition about how likely it is for a virus to infect bacteria was way off base. (The good news is that using our formula it is possible for doctors to use viruses to kill bacteria infecting a patient!)

- **Probabilities Sum to One:** The sum of the probabilities of all of the outcomes in the sample space must be equal to one. (This is equivalent to the previously stated fact that we know that exactly one of the outcomes will actually be realized.) A closely related fact is that the sum of the probabilities of \( E \) and \( E^c \) (its complement) are always equal to one, which can be written as: \( P(E) + P(E^c) = 1 \) (or \( P(E^c) = 1 - P(E) \)).

- **Reading Probabilities of Two Events from a Table:** Let us return to the study on which we based our cancer probability above. It considered other features also, such as whether the person smoked, as shown in this table:

<table>
<thead>
<tr>
<th></th>
<th>Smoked ((S))</th>
<th>Didn't Smoke ((S^c))</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Died of Cancer ((C))</td>
<td>107,228</td>
<td>99,016</td>
<td>206,244</td>
</tr>
<tr>
<td>Didn’t Die of Cancer ((C^c))</td>
<td>125,891</td>
<td>564,442</td>
<td>690,333</td>
</tr>
<tr>
<td>Totals</td>
<td>233,119</td>
<td>663,467</td>
<td>896,577</td>
</tr>
</tbody>
</table>

**Question 4:** What is the probability that an American smokes? (Approximate a value empirically from the data.)

In fact, we can turn this entire table into probabilities by dividing each entry by the 896,577 (the total number of people):

<table>
<thead>
<tr>
<th></th>
<th>Smoked ((S))</th>
<th>Didn’t Smoke ((S^c))</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Died of Cancer ((C))</td>
<td>.120</td>
<td>.110</td>
<td>.230</td>
</tr>
<tr>
<td>Didn’t Die of Cancer ((C^c))</td>
<td>.140</td>
<td>.630</td>
<td>.770</td>
</tr>
<tr>
<td>Totals</td>
<td>.260</td>
<td>.740</td>
<td>1</td>
</tr>
</tbody>
</table>

Note that the two in the last column are \( P(C) \) and \( P(C^c) \) and the two in the last row are \( P(S) \) and \( P(S^c) \). What are the others?

- **A and B / The Intersection of Two Events:** If \( A \) and \( B \) are two subsets of a sample space, then the intersection of \( A \) and \( B \) is the subset of outcomes that are in both \( A \) and \( B \) (see Figure 5.6a on page 228). This is also an event, which we call “\( A \) and \( B \)”.

**Question 5:** Which of the entries in the table above is \( P(S \text{ and } C) \)? Which is \( P(S \text{ and } C^c) \)?

- **Disjoint Events:** If \( P(A \text{ and } B) = 0 \) then we say that the events are disjoint. In the language of sets, this just means that there are none of the outcomes in \( A \) are in \( B \) also. Or, another way of thinking about it is that \( A \) and \( B \) cannot both occur. For instance, \( A \) and \( A^c \) are always disjoint.

**Question 6:** Suppose the trial is that we are going to randomly select a person from the Charleston phone book. Let \( F \) be the event “the person is female” and \( C \) be the event “the person is a Catholic priest”. What is \( P(F \text{ and } C) \) and why?
• **A or B / The Union of Two Events:** If $A$ and $B$ are two subsets of a sample space, then the union of $A$ and $B$ is the subset of outcomes that are in either $A$ or $B$ – including those that may be in both (see Figure 5.6a on page 228). This is also an event, which we call “$A$ or $B$”.

**Question 7:** Suppose we randomly select a marble from a bag containing 5 small green marbles, 10 small white marbles, 10 large white marbles and 10 large blue marbles. If $S$ is the event “the marble is small” and $G$ is the event “the marble is white”, what is the event “$S$ or $G$”? What is the $P(S \text{ or } G)$?

**Question 8:** Go back to the first table and try to figure out what $P(S \text{ or } C)$ would be. (Be careful not to double-count the smokers who died from cancer.)

• **A Formula for $P(A \text{ or } B)$:** Amazingly, the following simple formula is always true for any two events:

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

(Figure 5.7 on page 229 shows why this works. Basically, if you add $P(A)$ to $P(B)$ you are counting the intersection $A$ and $B$ twice. Subtracting it off once gives just the right value!)

Note that if the events $A$ and $B$ are disjoint, then this formula gets even simpler since the term being subtracted off is zero. For instance if $P(F) = .502$ and $P(C) = .0001$ then when $F$ and $C$ are disjoint $P(F \text{ or } C) = .502 + .0001 = .5021$. But, you must be careful to only apply this simple “addition formula” when the events really are disjoint.

• **Is there a formula for $P(A \text{ and } B)$?** Yes, there is a formula like the one above that works for any intersection. However, we will not see it until next time (Section 5.3). Instead, we have a formula like the simpler one above that only works in special cases.

• **Independence of Events:** We say events $A$ and $B$ are independent

$$P(A \text{ and } B) = P(A) \times P(B).$$

(We will learn next time what it means that the events are independent. For now, just accept this formula as the definition.)

**Question 9:** Using the table, find $P(S \text{ and } C)$, then find $P(S) \times P(C)$. Are they equal? What do you think this means? (Answer: They are not equal, and this means that the probability of dying from cancer is not the same for smokers and non-smokers!)

---

**Homework**

**Read:** Read Sections 5.1 and 5.2 in the book.

**Register:** Go to http://www.coursecompass.com and click on “register” to sign up for the online homework software. Use “kasman69631” as the Course ID.

**Do:** Answer the 10 questions in the online homework assignment “Orientation / Sections 5.1-5.2”. It is due on Thursday at 5PM. So, try to do it right away and ask me about it in class next time if there are any problems.