Project: Basic Probability
(AKA, single probabilities)

At this point in the term in class, we’re talking about statistics. Statistics, as you’ve learned, is the science of analyzing and presenting data. It’s a tremendously important branch of mathematics, one that you experience every day, and one that influences the majority of your decisions, whether you realize it or not.

The first two projects in MTH 234 deal with statistics’ fraternal twin, probability. In reality, much of what folks lump under the umbrella of “statistics” is actually part statistics, and part probability. Where statistics is taking data and analyzing, probability is measuring how likely something is to happen. I call the two branches “fraternal twins” because they often get used in tandem when analyzing results and making predictions. More on that later.

You might also ask yourself, “Why have this as a project (or two)? I notice that the book has a whole chapter on probability!” Good call! Yes, it does (and most stat textbooks do, as well). In fact, there are multiple chapters on probability. However, the introduction to probability on which most textbooks focus is too cumbersome, and it often focuses unbearably on small details that are better served with things called distributions, which we will do, in class, after the first two projects.

Part 1. The nuts and bolts.

Let’s begin with an interesting dataset: According to an article in Chance magazine (Vol. 14, no. 2), tenants at a large apartment complex filed a lawsuit alleging racial steering. Racial steering occurs when real estate agents show prospective buyers homes only in neighborhoods already dominated by that family’s race (a practice that violates the Fair Housing Act of 1968). The complex of interest was divided into two parts…call them Section A and Section B. The plaintiffs claimed that Caucasian potential renters were steered to Section A, while African American potential renters were steered to Section B (there were no other races represented in this apartment complex). The table below shows the data presented in court to show the locations of recently rented apartments.

<table>
<thead>
<tr>
<th>New Renters</th>
<th>Caucasian</th>
<th>African American</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section A</td>
<td>87</td>
<td>8</td>
</tr>
<tr>
<td>Section B</td>
<td>83</td>
<td>34</td>
</tr>
</tbody>
</table>

This table is called a contingency table (or CT, for short). A contingency table is a way of clearly delineating the contents of a data set into disjoint, or mutually exclusive, categories (each data point placed into a “disjoint” categorical CT falls into one, and only one, of the categories). For example, in the CT above, there are 4 disjoint categories: a) “Caucasian & Section A”, b) “African American & Section A”, c) “Caucasian & Section B”, and d) “African American & Section B”. Any person included in the dataset will fall into exactly one of these categories.

Let’s ask a few questions about this data before continuing:

Example 1. How many new renters were included in this study?
Answer 1. Adding up all of the disjoint categories, we find that there were 212 new renters in this dataset (87 + 83 + 8 + 34).

Example 2. How many new renters moved into section A?
Answer 2. 95 (87 Caucasians and 8 African Americans)

Example 3. How many new renters moved into section B?
Answer 3. 117 (please note that the sum of Answer 2 and Answer 3 is 212. That should seem reasonable).
Example 4. How many new renters were Caucasian?
Answer 4. 170 (87 in Section A and 83 in Section B)

Example 5. How many new renters were African American?
Answer 5. 42 (again, check to see that the sum of Answers 4 and 5 equal Answer 1)

Example 6. What percentage of new renters were Caucasian?
Answer 6. Well, percentages you can handle, right? It’s just a matter of “part of whole”. In this data, there are 212 new renters. Of those, 170 were Caucasian...so the percentage of new renters who were Caucasian was \( \frac{170}{212} \), or about 80%.

Now...our first probability question:

Example 7. Suppose we conduct an experiment where we randomly select one new renter from the 212. What is the probability (or chance) that this renter is Caucasian?

Answer 7. Ah – HA! The plot thickens. Actually, it’s the same plot, as the answer is still \( \frac{170}{212} \), or about 80%. But why?

When we made the assertion “randomly select one new renter from the 212”, we put into place the denominator of the probability fraction (the 212, in this case). This denominator’s numerical value is the size of the sample space, and it represents the number of all possible outcomes. Since there are 212 new renters, and we’re picking one at random, there are 212 possible ways this experiment can proceed.

Now, the numerator of any probability fraction is called the event space (well, that’s what I call it, anyway). Its value represents the number of ways that the event of interest can occur. In this case, the event of interest is “getting a Caucasian”. Since there are 170 Caucasians among the new renters, the event space is 170. So we have the probability of getting a Caucasian among a new renter as \( \frac{170}{212} \), or 80%.

Notes thus far:

a) All probabilities are of the same form: \( \frac{\text{size of event space}}{\text{size of sample space}} \).

b) Statisticians get tired of writing the phrase “What’s the probability of such-and-such an event of interest happening?” They use (and you will, too) a shorthand that looks like this:

\[
p(\text{Caucasian}) = \frac{170}{212} \approx 80%.
\]

Let’s deconstruct this a bit:

i) The “p” out front just stands for “probability”; I know it looks like function notation from MTH 095 and MTH 111, but it’s not.

ii) The words in parentheses represent the event of interest. Unless otherwise noted (which you’ll see in a little bit), the sample space is understood to be all available data.

iii) There’s no one correct way to write these events in the parentheses; you can get more or less wordy, so long as you and your intended audience understands them. For example, all of these are clear to me:

\[
p(\text{Caucasian}) \quad p(\text{Caucasian in new renters}) \quad p(\text{Caucasian randomly selected from new renters})
\]
c) Note that I used some squiggly equals signs up there; that’s because $\frac{170}{212}$ isn’t exactly equal to 80%. It’s approximately equal to 80%.

d) You can choose to use fractions, percentages or decimals when you answer probability questions.

e) The smallest value the event space can be is 0; therefore, the smallest probability you can ever have is 0. For example:

**Example 8.** Find $p(\text{Mexican American selected from new renters}).$

**Answer 8.** The sample space is still 212, since we’re still selecting from the same group of folks. However, the event space is 0, since there are no Mexican Americans represented in this dataset (remember, they were either Caucasian or African American). Therefore, $p(\text{Mexican American selected from new renters}) = \frac{0}{212} = 0.$

f) The largest possible probability is 1, or 100%. The event space can never be larger than your sample space; however, it can be equal to it. For example:

**Example 9.** Find $p(\text{new renter lives in either Section A or Section B}).$

**Answer 9.** All of the renters included in the dataset live in either Section A or Section B...so no matter who you pick, they will live in one the two places of interest. So, $p(\text{new renter lives in either Section A or Section B}) = \frac{212}{212} = 100\% = 1.$

g) The word “or” when dealing with disjoint events means to add.

**Example 9 (revisited).** Find $p(\text{new renter lives in either Section A or Section B}).$

**Answer 9 (alternate).** 95 renters moved into Section A while 117 moved into section B. “or” means to add events, so we can answer the question as such:

$$p(\text{new renter lives in either Section A or Section B}) = \frac{95 + 117}{212} = \frac{212}{212} = 100\%.$$

But be careful when adding disjoint probabilities, as the following “or” problem shows:

**Example 10.** Find $p(\text{new renter lives in either Section A or is Caucasian}).$

**Answer 10.** At first glance, you might say, “Well, let’s just add up all of those in section A (95), and then add the Caucasians (170), and call it good.” So you proceed:

$$p(\text{new renter lives in either Section A or is Caucasian}) = \frac{95 + 170}{212} = \frac{265}{212} = \ldots \text{wait a minute...how can your event space be larger than your sample space? That goes in direct violation of note 5) above. So what gives?}$$

Look at your CT again:

<table>
<thead>
<tr>
<th>New Renters</th>
<th>Caucasian</th>
<th>African American</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Section A</strong></td>
<td>87</td>
<td>8</td>
</tr>
<tr>
<td><strong>Section B</strong></td>
<td>83</td>
<td>34</td>
</tr>
</tbody>
</table>
Your sample space is still 212. Your event space needs to be all of those who live in Section A, or who are Caucasian. Looking at the CT, that means our event space consists of those who are **bolded and in red** (as they are the ones who satisfy either being Caucasian or living in Section A):

<table>
<thead>
<tr>
<th>New Renters</th>
<th>Caucasian</th>
<th>African American</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section A</td>
<td>87</td>
<td>8</td>
</tr>
<tr>
<td>Section B</td>
<td>83</td>
<td>34</td>
</tr>
</tbody>
</table>

**Answer 10 (the correct way).** So, 

\[
p(\text{new renter lives in either Section A or is Caucasian}) = \frac{87 + 83 + 8}{212} = \frac{178}{212} \approx 84%.
\]

So what went wrong with our previous attempt? We broke rule g) above:

**g) The word “or”, when dealing with disjoint events, means to add.**

The events “Section A” and “Caucasian” are not disjoint, since a data point can be both Caucasian AND live in Section A (in fact, 87 of them do). “Double Counting”, as I call this, is an easy – to – make mistake, but fortunately, CT’s make it easy to avoid: simple add up the disjoint cells. Voila! Impossible to double count.

**Example 11.** Find \( p(\text{African American}) \).

**Answer 11 (one way).** We can use the idea of probability (event space over sample space):

\[
p(\text{African American}) = \frac{42}{212} \approx 20%.
\]

**Answer 11 (a slicker way).** From Question 7, we knew that \( p(\text{Caucasian}) = 80% \). Since there are only two races represented in this data set, it stands to reason that

\[
p(\text{Caucasian}) + p(\text{African American}) = 100%
\]

In this dataset, “Caucasian” and “African American” are called **complementary events**. Complementary events are always mutually exclusive, and, if one of them doesn’t happen, the other one does.

\[a\] Using that little above equation, 

\[
p(\text{Caucasian}) + p(\text{African American}) = 100%
\]

\[
80% + p(\text{African American}) = 100%
\]

\[
p(\text{African American}) = 20%
\]

Remember, however...this only works with complementary events!

Let’s look back at that last probability one more time:

\[a\] Your text, I’ve noticed, uses a fairly confusing notation for complementary events. They would write, “\( p(\overline{\text{Caucasian}}) \),” with the overscore implying the opposite of Caucasian, which in this dataset, is African American. This notation too closely resembles the over score in \( \overline{X} \), in my opinion, so I won’t use it here. Other texts use different notations...a pseudo - popular one is “\( p(\text{Caucasian}') \),” with the “C” meaning complementary... but, honestly, I think it’s easier to just write out the events.
What I want to know is this: what does that *mean*? For example, when someone tells you that there’s a 20% chance of, say, rain today, what does that *mean*? Or, if you have a 5-question multiple choice test question, and you guess at the answer, you can be sure that the chance you guess the correct one is 20%...but, again...what does that *mean*?

Well, here’s one more way that that probability and statistics are fraternal twins...a 20% chance of some event happening simply means that, on average, if you conduct an experiment, 20% of the time you’ll get the event in which you’re interested. So:

- If \( p(\text{rain}) = 20\% \), then, 20% of the time, when the atmospheric conditions are the way they are right now, it will rain.
- If \( p(\text{you guess the answer correctly}) = 20\% \), then, if you randomly guess, over and over, at many, many 5-option questions, you’ll get the right answer about 20% of the time.
- If you randomly select one person from the 212 represented in the apartment study, time and time again, about 20% of the time, you’ll select someone of African American descent.

So, here’s a synopsis:

<table>
<thead>
<tr>
<th><strong>Definition of Probability</strong></th>
<th>The probability of some event happening is found by calculating the ratio of the event space to the sample space.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Meaning of Probability</strong></td>
<td>The value of the probability found in the previous definition can be interpreted as the percentage of time you will successfully observe the event in question over repeated trials...the more trials, the closer you get to the percentage.</td>
</tr>
</tbody>
</table>

#### Part 2. One more nut and bolt, and some tightening.

OK, here’s our original data set one more time:

<table>
<thead>
<tr>
<th>New Renters</th>
<th>Caucasian</th>
<th>African American</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section A</td>
<td>87</td>
<td>8</td>
</tr>
<tr>
<td>Section B</td>
<td>83</td>
<td>34</td>
</tr>
</tbody>
</table>

The last bit in this project will introduce us to a very important probabilities idea. Let’s look back at the prior point of this data set, though, first...remember what it was?

> “[T]enants at a large apartment complex filed a lawsuit alleging racial steering. Racial steering occurs when real estate agents show prospective buyers homes only in neighborhoods already dominated by that family’s race (a practice which violates the Fair Housing Act of 1968).”

Now, we’ve calculated a bunch of probabilities so far, but have any of them come close to answering the question, “Does it appear that the apartment complex was involved in racial steering?" Let’s see:

\[\text{\textsuperscript{b} This is the kind of question we answer daily (heck, 10 – minutely), in MTH 244, a course you all should take, for sure.}\]
- \( P(\text{Caucasian}) \)... nah... that misses the association with where they live.
- \( P(\text{Section A or Section B}) \)... again, there’s an association missing (their race).
- \( p(\text{new renter lives in either Section A or is Caucasian}) \).... feels closer, as we’ve got both race and where they live in there, but something's still not right; it’s not specific enough.

No, what we need is something like this “Suppose we randomly choose someone from Section A...what’s the chance they’re Caucasian?” Let’s make that question 12, shall we?

**Example 12.** Find \( p(\text{randomly chosen Section A resident is Caucasian}) \).

**Answer 12.** OK, the first thing we're going to do is write that more concisely. Here she goes:

\[
p(\text{randomly chosen Section A resident is Caucasian}) = p(\text{Caucasian} \mid \text{Section A})
\]

The vertical line you see stands for the words “given that”... so, what you’re looking for is the “chance you get a Caucasian, given that you’ve chosen someone from Section A.” And that’s the interesting part... since you know you’ve selected someone from Section A, your sample space has changed, hasn’t it?

**New Renters**

<table>
<thead>
<tr>
<th></th>
<th>Caucasian</th>
<th>African American</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Section A</strong></td>
<td>87</td>
<td>8</td>
</tr>
<tr>
<td><strong>Section B</strong></td>
<td>83</td>
<td>34</td>
</tr>
</tbody>
</table>

Your sample space is now 95 (87 + 8), since you know you didn’t select someone from Section B. So,

\[
p(\text{Caucasian} \mid \text{Section A}) = \frac{87}{95} \approx 92\%
\]

This is called a **conditional probability** (the condition being that you selected someone from Section A). I like to think of these types as a “sample space reduction” problem, as that’s precisely what you’re doing.

**Example 13.** Find \( p(\text{African American} \mid \text{Section A}) \).

**Answer 13.** Same conditional sample space as the last one, right?

\[
p(\text{African American} \mid \text{Section A}) = \frac{8}{95} \approx 8\%
\]

**Answer 13 (more slickly done).** Do you see how the probabilities in Questions 12 and 13 are complements?

\[
p(\text{Caucasian} \mid \text{Section A}) + p(\text{African American} \mid \text{Section A}) = 100\%
\]

That seems reasonable... if you select someone from Section A, they’re either African American or Caucasian. Since you found that 92% of Section A folks are Caucasian, then the rest must be African American, so

\[
p(\text{African American} \mid \text{Section A}) = 100\% - p(\text{Caucasian} \mid \text{Section A}) = 100\% - 92\% = 8\%
\]
Example 14. Find $p(\text{Caucasian} \mid \text{Section B})$.
Answer 14. This one’s sample space is 117 (83 + 34), as shown below:

<table>
<thead>
<tr>
<th>New Renters</th>
<th>Caucasian</th>
<th>African American</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section A</td>
<td>87</td>
<td>8</td>
</tr>
<tr>
<td>Section B</td>
<td>83</td>
<td>34</td>
</tr>
</tbody>
</table>

So, $p(\text{Caucasian} \mid \text{Section B}) = \frac{83}{117} \approx 71\%$.

Example 15. Find $p(\text{African American} \mid \text{Section B})$.
Answer 15. $p(\text{African American} \mid \text{Section B}) = 100\% - p(\text{Caucasian} \mid \text{Section B}) = 100\% - 71\% = 29\%$.

Example 16. Find $p(\text{Section A} \mid \text{Caucasian})$.
Answer 16. Oooooo! That’s interesting, isn’t it? It’s still conditional, but the condition has changed….now, we know we selected a Caucasian, and we’re interested in knowing where they live. Our data table now looks like this:

<table>
<thead>
<tr>
<th>New Renters</th>
<th>Caucasian</th>
<th>African American</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section A</td>
<td>87</td>
<td>8</td>
</tr>
<tr>
<td>Section B</td>
<td>83</td>
<td>34</td>
</tr>
</tbody>
</table>

So, $p(\text{Section A} \mid \text{Caucasian}) = \frac{87}{170} \approx 51\%$.

Example 17. Find $p(\text{Section B} \mid \text{Caucasian})$.
Answer 17. $p(\text{Section B} \mid \text{Caucasian}) = 100\% - p(\text{Section A} \mid \text{Caucasian}) = 100\% - 51\% = 49\%$.

Example 18. Find $p(\text{Section A} \mid \text{African American})$.
Answer 18. A new sample space…but you figured that out, right? $p(\text{Section A} \mid \text{African American}) = \frac{8}{42} \approx 19\%$.

Example 19. Find $p(\text{Section B} \mid \text{African American})$.
Answer 19. $p(\text{Section B} \mid \text{African American}) = 100\% - p(\text{Section A} \mid \text{African American}) = 100\% - 19\% = 81\%$.

News organizations wanting to imply that the apartment complex is engaging in racial steering would be quick to point out that “an African American is four times more likely to live in Section B than in Section A, while Caucasians are just as likely to live in either Section” (these are from answers 16, 17, 18 and 19). Other new organizations, wanting to cast doubt on the racial steering allegation, would say that the racial proportion in each of the two sections is statistically the same (questions 12, 13, 14 and 15). In MTH 244, should you choose that wonderful road, you’ll see why both of these news organizations have correct points, and incorrect ones.
At this point, the most important thing to notice is the power of probability as it occurs in your everyday lives. Next project, we’ll look at joint probabilities and see how they’re all over the place, too…this stuff is awesome!

For these problems, answer as either a decimal (rounded to the nearest hundredth’s place) or a percent (rounded to the nearest whole number). In addition, show how you arrived at that decimal or percentage (w). For example, if I were answering question 18 for credit, it could look like this:

\[ p(\text{Section A} \mid \text{African American}) = \frac{8}{42} \approx 0.19 \text{ or } p(\text{Section A} \mid \text{African American}) = \frac{8}{42} \approx 19\% \]

Question 19’s answer could look like this:

Find \( p(\text{Section B} \mid \text{African American}) = 100\% - p(\text{Section A} \mid \text{African American}) = 100\% - 19\% = 81\% \).

Or

Find \( p(\text{Section B} \mid \text{African American}) = 1 - p(\text{Section A} \mid \text{African American}) = 1 - 0.19 = 0.81 \)

1. Punxatawney Phil (shown at right) is a wonderful little creature; he’s the groundhog who is constantly predicting weather on (you guessed it) Groundhog’s Day. Without going too much into the history and rigmarole of this strange event (that’s what Google is for), here’s the basic idea: Phil either predicts an early spring or he predicts a late spring. In either case, he’s either right, or he’s wrong. Here are his numbers (as many as I can find records for, anyway!):

<table>
<thead>
<tr>
<th></th>
<th>right</th>
<th>wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early spring prediction</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Late spring prediction</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

(2 points each) From these data, find the following probabilities (‘I’ll give you a few answers in red so you know you’re on the right track…remember that you still need to show me how to get those answers!’):

a. (w) \( p(\text{wrong}) \) 0.29

b. (w) \( p(\text{right}) \)

c. (w) \( p(\text{right} \mid \text{early spring prediction}) \)

d. (w) \( p(\text{right} \mid \text{late spring prediction}) \) 0.5

e. It’s often said that Punxatawney Phil is right 75% of the time. This roughly agrees with your answer to part b). However, what do your answers in parts c) and d) mean? In other words, should we feel better about Phil’s prediction being right when he predicts an early spring or a late one? Why?

f. (w) \( p(\text{right or early spring prediction}) \)
2. Start by reading this article!


This study actually covered more cities than just Boston, but the numbers crunched in Boston will do just fine for this project. Using the values found in the article, find the following probabilities:

a. (2 points) \( P(\text{Uber rider gets cancellation} \mid \text{African American sounding name}) = \)

b. (2 points) \( P(\text{Uber rider gets cancellation} \mid \text{white sounding name}) = \)

c. (4 points) Explain what the probability in part b means, without using the words “chance”, “probability” nor any other word equivalent to those. If you get stuck, look at the box where “Meaning of Probability” is explained!

d. (2 points) Complete the blank space with either a “<” or “>” sign (you’ll have to read down a bit in the article to find the context:

\[
P \left( \frac{\text{Uber rider gets cancellation}}{\text{African American sounding name}} \right) \quad \quad \quad \quad \quad P \left( \frac{\text{Lyft rider gets cancellation}}{\text{African American sounding name}} \right)
\]

Of course, the Lyft company read this analysis and same to a wonderful conclusion:

“Lyft spokesman Adrian Durbin said the company does not tolerate discrimination. Because of Lyft, people living in underserved areas — which taxis have historically neglected — are now able to access convenient, affordable rides,” he said.

e. (3 points) Read the article more carefully and give me one other reason your inequality sign in part c might point the way it points (in other words, could anything else explain the discrepancy in the cancellations between the two companies besides Mr. Durbin’s explanation?).