Progress Report: Week 1

Reading:

This week, I read Chapter 1 of *Think Bayes*. At the end of this chapter, I was pretty sure that I understood how to solve a problem using Bayes' theorem, but I was less clear on the big-picture idea of Bayesian statistics, or why someone would want to use Bayesian reasoning, or what the benefits of Bayesian reasoning were.

As a result, I also read the optional reading: Yudkowsky's "An Intuitive Explanation of Bayes' Theorem". Like the title suggests, the explanation of Bayes' theorem was intuitive, and I thought that stating the same problem multiple different ways, as he did with the problem about women having breast cancer was an effective way to not only explain what the correct answer to the problem, but also why this particular problem lends itself well to a solution using Bayes' theorem.

In addition, I also read Chapter 2 of *Think Bayes*. This chapter walked through the most of the same problems as in chapter 1, and some new ones, but solved them used python to do the calculations, instead of making tables, like in chapter 1. This chapter was more confusing to me than the first chapter, and I think that has to do with the fact that I'm not entirely confident with solving problems in this particular way (with python as opposed to on paper).

I read Downey's blog post "Frank is a scoundrel, probably" which contains an example problem of someone flipping a coin to determine whether or not the coin is fair. This article explained that the Frank, because of the test he used to test for whether a coin was fair, was probably trying to mislead Betsy by saying that the coin was not fair. I thought that the post highlighted some reasons to be skeptical of repeated tests, but that I didn't actually understand the math done in the blog post until I expressed the Bayesian update in the form of a table.

I also read Chapter 3 of *Think Bayes*. This chapter included examples of problems where there were multiple fairly likely answers to the problem and what to do if this was the case. In addition, there was also information about how to choose the priors for a problem based on background knowledge and to get more data to refine the posteriors. I thought that this chapter of the book as fairly easy to understand, and I liked that this chapter included a real-world problem, the German tank problem, as an example.

Exercises:

This week, I tried the sample problems:

Question: Read about Hume's article *Of Miracles*. Now suppose that a good friend of yours, who is generally sober, reliable, and responsible, reports to you that she is quite certain that she saw Elvis Presley, alive and 79 years old, at the Piggly Wiggly. What effect does this report have on your subjective degree of belief that Elvis is alive? Express your Bayesian update in the form of a table.

My Hypotheses are:

A: Elvis is alive

B:Elvis is not alive

My Evidence:

D: My friend thinks that Elvis was at the Piggly Wiggly

	Prior P(H)[1]	Likelihood P(D|H)[2]	P(H)*P(D|H)	Posterior P(H|D)
A	1/100	1/1000	1	1/991
B	99/100	10/1000	990	990/991

1: I am fairly certain that Elvis is not alive, so the likelihood of my hypotheses that Elvis is alive is very small, but not zero

2: If my friend were 100% reliable (which people aren't according to Hume's article Of Miracles), then the likelihood that that Elvis was at the Piggly Wiggly given that Elvis is dead would be 0. Given that people aren't 100% reliable, though, the chance that my friend thinks that Elvis was at the Piggly Wiggly given that he is not alive is not 0, but is rather small compared to the likelihood that Elvis was at the Piggly Wiggly given that he is alive (which is also rather small).

Question: Elvis Presley had a twin brother who died at birth. What is the probability that Elvis was an identical twin?

My Hypotheses are:

A: Elvis is an identical twin

B:Elvis is a fraternal twin

My Evidence:

D: Elvis had a twin was male

	Prior P(H)[1]	Likelihood P(D|H)[2]	P(H)*P(D|H)	Posterior P(H|D)
A	8	1	8	4/27
B	92	1/2	46	23/27

1: Identical twins make up about 8% of all twins, according to data from the sample problem

2: If Elvis were an identical twin, the likelihood of his twin being male would be 100%. If he were a fraternal twin, then his twin could be either male or female. Assuming that two male fraternal twins are no less common than a male and female twin, the likelihood that his twin is male given that Elvis is a fraternal twin is 50%.

Question: Two people have left traces of their own blood at the scene of a crime. A suspect, Oliver, is tested and found to have type O blood. The blood groups of the two traces are found to be of type O (a common type in the local population, having frequency 60%) and of type AB (a rare type, with frequency 1%). Do these data (the blood types found at the scene) give evidence in favour [sic] of the proposition that Oliver was one of the two people whose blood was found at the scene?

My Hypotheses are:

A: Oliver was at the crime scene

B: Oliver was not at the crime scene

My Evidence:

D: The blood types at the crime scene were O and AB

We're trying to find p(D|H)

p(D|A): If Oliver was at the crime scene, then all we need to account for is the AB blood. The chance of there being AB blood at the crime scene is 1%.

p(D|B): If Oliver wasn't at the crime scene, then we just need to find the probability that an O blood and AB blood was at the crime scene: .6*0.01* 2=1.2% (multiply by 2 to account for the fact that it could be AB and O blood or O and AB blood.)

This suggests that Oliver might not be guilty.

Question: According to the CDC, "Compared to nonsmokers, men who smoke are about 23 times more likely to develop lung cancer and women who smoke are about 13 times more likely." If you learn that a woman has been diagnosed with lung cancer, and you know nothing else about her, what is the probability that she is a smoker?

My Hypotheses are:

A: The woman is a smoker

B: The woman is not a smoker

My Evidence:

D: The woman has lung cancer

	Prior P(H)[1]	Likelihood P(D|H)[2]	P(H)*P(D|H)	Posterior P(H|D)
A	16.4	20.3	3329.2	71.9%
B	83.4	1.56	1301.04	28.1%

1: These are the percentages of women who do and don't smoke according to the CDC

2: To calculate the chance that a woman who does or doesn't smoke will get lung cancer, we get the total probability that anyone will get lung cancer. And then using the fact that we know smokers are 13 times more likely than non-smokers to get lung cancer. Doing a little algebra, we can determine that there is 1.56% chance that a non-smoker will get lung cancer and a 20.3% chance that a smoker will get lung cancer.

Question: Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say Door A [but the door is not opened], and the host, who knows what's behind the doors, opens Door B, which has a goat. He then says to you, "Do you want to pick Door C?" Is it to your advantage to switch your choice?

My Hypotheses are:

A: The car is behind door A

B: The car is behind door B

C: The car is behind door C

My Evidence:

D: Door B is opened

	Prior P(H)[1]	Likelihood P(D|H)[2]	P(H)*P(D|H)	Posterior P(H|D)
A	1/3	1/2	1/6	1/3
B	1/3	0	0	0
C	1/3	1	1/3	2/3

1: Before, there's an equal chance that the car is behind any door

2: If the car is behind door A, then it doesn't matter which door is opened, so there's a 50% chance that door B is opened. If the car is behind door B, then the host won't open door B. If the car is behind door C, then the host will open door B.

Question: If you meet a man from Scotland with (naturally) red hair, what is the probability that neither of his parents has red hair?

My Hypotheses are:

A: The parents (at least 1) have red hair

B: Neither parent has red hair

My Evidence:

D: The son has red hair

	Prior P(H)[1]	Likelihood P(D|H)[2]	P(H)*P(D|H)	Posterior P(H|D)
A	0.2431	0.0525	397	68%
B	0.7569	0.0768	187	32%

1: If 13% of people have red hair, then the probability of neither parent having red hair is $(1-.13)^2$. The probability of at least 1 parent having red hair is $1 – (1-0.13)^2$.

2: If we apply the Hardy Weinberg principle, and we know that $q^2 = 0.13$, then we do algebra to find that q = 0.36 and p = 0.64. Then, to find the probability that the son has red hair if neither parent does, is the probability that both parents without red hair are pq and then that they have a son with red hair. This is 2*p*q*2*p*q = 5.25%

Conversely, the probability that a son has red hair if one or both parents does is the probability that both have red hair $0.13^2 + 2*0.13*0.46/2$. The 0.13^2 term means the likelihood that both parents have red hair. If this happens, then they'll definitely have a child with red hair. The second term is the likelihood that one person has red hair multiplied by the likelihood that a parent doesn't have red hair, but is has one recessive gene, multiplied by 2 because either the mother or the father could have red hair divided by 2 because these parents have a 50% chance of having a child with red hair. This is 7.68%

From the Yudkowsky reading, I did the following problem:

Question: Suppose that a barrel contains many small plastic eggs. Some eggs are painted red and some are painted blue. 40% of the eggs in the bin contain pearls, and 60% contain nothing. 30% of eggs containing pearls are painted blue, and 10% of eggs containing nothing are painted blue. What is the probability that a blue egg contains a pearl? For this example the arithmetic is simple enough that you may be able to do it in your head, and I would suggest trying to do so.

My Hypotheses are:

A: The Egg contains a pearl

B: The Egg does not contain a pearl

My Evidence:

D: The egg was blue

	Prior P(H)[1]	Likelihood P(D|H)[2]	P(H)*P(D|H)	Posterior P(H|D)
A	40	30	1200	2/3
B	60	10	600	1/3

After reading chapter 2 of *Think Bayes*, to make sure I understood how the code described in the chapter worked, I implemented the above problem in python. The complete code can be found in my github repository at: https://github.com/phiaseitz/ThinkBayes2/blob/master/code/eggs.py

To do this, I followed the format of the cookie example in *Think Bayes* with a few minor changes. Most of these changes, other than having appropriate variable names, were in the init method.

```
 '''Instantiates the egg class'''
def __init__(self, hypos, priors):
    Pmf.__init__(self)
    '''Set the priors (for each H, the corrisponding p(H)'''
    for i in range(len(hypos)):
        self.Set(hypos[i],priors[i])
    self.Normalize()
```

The reason for these changes is that, unlike the cookie example, there is not an equal likelihood that the egg contains a pearl or does not contain a pearl. As a result, I passed in both the hypotheses and the prior probabilities into the init method. Then for each hypotheses, I set the corresponding probability.

The update method remained exactly the same.

I also changed the mixes variable to reflect my hypotheses and the distributions of egg colors for eggs with and without pearls

```
  mixes = {
        'Pearl':dict(red=0.7, blue=0.3),
        'No Pearl':dict(red=0.9, blue=0.1),
        }
```

Then, I wrote the code to instantiate the egg class and run a Bayesian update.

```
def main():
    '''Instantiate the hypotheses and the priors'''
    hypos = ['Pearl', 'No Pearl']
    priors = [0.4, 0.6]

    '''Create the pmf object'''
    pmf = Egg(hypos,priors)

    '''Assuming that we got a blue egg'''
    pmf.Update('blue')
```

Here, I also needed to set the priors because the number of eggs with pearls was not the same as the eggs without pearls. Then assuming that we got a blue egg, we calculate the probability of having a pearl and not having a pearl.

The output, after printing each hypothesis and probability was:

```
Pearl 0.666666666667
No Pearl 0.333333333333
```

This is what I got when I did the Bayesian update by hand. Yay!

Case Study:

I haven't started any work on the case study yet.

Reflection:

In the sample problems (solutions above) the part that confused me the most at first was not necessarily finding the actual probabilities of p(D|H) or p(H) figuring out what D was. When I looked through the solutions, that was what I most often got wrong, or that was often where I got stuck. After doing the all the sample problems and some of the problems in Yudkowsky's *An Intuitive Explanation of Bayes' Theorem*, I think I've gotten more of a handle on that. The thing that I am most struggling with is how to go from the table to using python, but doing the egg problem over again in python really helped with that.