Assignment #6: Expected value, variance, Bayesian Inference Name: _____

1. Faked numbers in tax returns, payment records, invoices, expense account claims, and many other settings often display patterns that aren't present in legitimate records. Some patterns, like too many round numbers, are obvious and easily avoided by a clever crook. Other patterns are more subtle. It is a striking fact that the first digits of numbers in legitimate records often follow a distribution known as Benford's Law. Here is that distribution:

First Digit	1	2	3	4	5	6	7	8	9
Probability:	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Calculate the <u>expected value</u> for the first digit (according to Benford's Law).

2. Calculate the <u>variance</u> of the first digit under Benford's Law. If you don't want to calculate the entire thing, just write out the beginning and end of the formula.

3. As we calculated in class, the expected value of rolling a single die is 3.5 (and the variance is 2.9167). Suppose I want to replace all the pips (dots) from the die with the numbers 4, 7, 10, 13, 16, and 19. In other words, I <u>tripled</u> the number of pips on each side and <u>added one</u>. Calculate the expected value and variance of this transformed die.

E(x) = 3.5 Var(x) = 2.9167

E(3x + 1) = _____ Var(3x + 1) = _____

Bayesian Inference

In assignment #4, you were introduced to Bayes' Theorem: $P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}$

and used the Law of Total Probability to calculate the denominator: $P(A) = P(A|B)P(B) + P(A|B')P(B')$

We did this using tables. In this assignment, you'll learn how to employ some basic Bayesian methods for inference.

Scenario: Let's play a somewhat convoluted game.

In my hands, I have 2 dice: one 6-sided; the other 12-sided. I will keep one die in each hand.

You'll pick a hand (left or right) and I'll roll the die that's in that hand.

If the die shows a number greater than or equal to 4, you win $1. If the die shows a 1, 2, or 3, you lose $1.

I will never change which die is in which hand, but we'll keep playing this game until you believe you have enough information to tell me which hand is holding the 12-sided die. If you guess correctly, you'll win $100. If you guess incorrectly, you'll have to return all the money you've made in the game. Notice that since you <u>can</u> lose money each round, you'll want to guess in as few rounds as possible.

4. Let's find the probability that you win $1 each round. For both the 6-sided and 12-sided dice, calculate the probability of rolling a number greater than or equal to 4.

$P(\text{roll} \geq 4 \mid \text{6-sided die}) =$ _____ $P(\text{roll} \geq 4 \mid \text{12-sided die}) =$ _____

5. Before we play the first round, you don't know which hand holds the 12-sided die. Fill-in the following:

H_1: The <u>left</u> hand holds the 12-sided die. $P(H_1) =$ _____

H_2: The <u>right</u> hand holds the 12-sided die $P(H_2) =$ _____

We'll call these your **prior probabilities** for the two competing hypotheses. These **priors** represent your beliefs before seeing any data (playing any rounds of the game).

6. Suppose we play 12 rounds of this game and get the results listed in the table below. Based on those results, which hand would you guess holds the 12-sided die? Briefly explain your decision.

Round	You choose… (Left or Right)	Result (Win or Loss)
1	L	Loss
2	L	Win
3	L	Win
4	R	Loss
5	R	Loss
6	R	Win
7	R	Loss
8	L	Win
9	L	Loss
10	L	Win
11	R	Loss
12	R	Win

I believe the 12-sided die is in the _____ hand.

Brief explanation:

7. Remember, before we begin playing, we believe the following:

   - P(chosen hand holds the 12-sided die) = 0.50 - - - - - - - - - - - - - - - - - - (based on your answer to question #5)
   - P(we lose | chosen hand holds the 12-sided die) = 0.25 - - - - - - - - - - - (based on your answer to question #4)
   - P(we lose | chosen hand holds the 6-sided die) = 0.50 - - - - - - - - - - - - (based on your answer to question #4)

   The table on the previous page shows we chose the left hand and lost. Let's see how that info changes our beliefs.

   Bayes' theorem calculates this change in our beliefs: $P(B|A) = \dfrac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B')P(B')}$

   Using notation from this specific example, Bayes' theorem would be:

$$P(\text{12-sided die in chosen hand}|\text{lose}) = \frac{P(\text{lose}|\text{12-sided in chosen})P(\text{12-sided in chosen})}{P(\text{lose}|\text{12-sided in chosen})P(\text{12-sided in chosen}) + P(\text{lose}|\text{6-sided in chosen})P(\text{6-sided in chosen})}$$

   Go ahead and calculate this probability:

   P(12-sided die in chosen hand | lose round) = _____

8. Like we've done in class with many of our discrete probability problems, we can simplify things a bit via a table. I've attempted to explain the values in the tables through steps A, B, and C.

   B) From question #4, we know P(win | 12-sided) = 0.75. We can find: P(win AND 12-sided) = P(win | 12-sided)P(12-sided)
   We can do a similar calculation to find P(win AND 6-sided)

	Win round	Lose round	Total
12-sided in chosen hand	.75 x .50 = .375	.50 – .375 = .125	0.50
6-sided in chosen hand	.50 x .50 = .250	.50 – .25 = .250	0.50
Total	.375 + .25 = .625	.125 + .25 = .375	1.00

   A) Our prior beliefs were that the 12-sided die had a 50/50 chance of being in either hand.

   C) Now we simply add/subtract to find all the other cell values.

   From the probabilities in this table, verify your answer to question #7:

   P(12-sided die in chosen hand | lose round) = _____

9. Your answer to the previous two questions is called a **posterior probability**. It represents a belief that has been updated based on results from an experiment.

    - Prior probability = P(chosen hand holds the 12-sided die) = 0.50 - - - - based on your answer to question #5)
    - Posterior probability = P(chosen hand holds the 12-sided die) = your answer to questions 7-8

    We can think of a posterior probability as: P(hypothesis | data). We now believe, based on the results of the first round of this game, that there is a 1/3 chance the left hand holds the 12-sided die. To summarize our beliefs:

    - P(left hand holds 12-sided die) = 1/3
    - P(win | chosen hand holds the 12-sided die) = 0.750
    - P(win | chosen hand holds the 6-sided die) = 0.50.

    Let's see how the results from round #2 updates our belief. In round 2, you chose the left hand and won. We could, once again, use the formula for Bayes' theorem, but let's try the table method. Notice that the far right column has been filled-in with our updated beliefs about the probability of the 12-sided die being in the left hand.

    Complete this table:

	Win round	Lose round	Total
12-sided in chosen hand			1/3
6-sided in chosen hand			2/3
Total			1.000

    From this table, update our belief. Remember, we won this round.

    P(12-sided die in chosen hand | won round) = _____

10. Let's move on to round #3, where you chose the left hand and won the round. Complete the table:

	Win round	Lose round	Total
12-sided in chosen hand			
6-sided in chosen hand			
Total			1.000

→ Remember, your answer to the previous question represents P(12-sided die is in chosen hand).

    Calculate P(12-sided die in chosen hand | won round) = _____

11. Let's do this one last time. In round #4, you chose the right hand and lost the round. Notice this is the first time your chosen hand is the right hand. Try to complete the table and calculate the probability:

	Win round	Lose round	Total
12-sided in chosen hand			
6-sided in chosen hand			
Total			1.000

→ Your answer to the previous question is the probability the 6-sided die is in the chosen hand

Let's rephrase the probability to continue focusing on the left hand.

P(12-sided die is in left hand | lost round) = P(6-sided die is in chosen hand | lost round) = _____.

12. This is becoming tedious, but I hope you see how we're simply updating our beliefs when we get new results from each round of the game. Using a computer, I went ahead and updated our beliefs after each of the 12 rounds in the game. You can check your previous answers with the results from this table.

Round	Choice (Left or Right)	Result (Win or Loss)	P(left hand holds 12-sided die)
(prior)	---	---	0.500
1	L	Loss	0.333
2	L	Win	0.428
3	L	Win	0.529
4	R	Loss	0.692
5	R	Loss	0.818
6	R	Win	0.750
7	R	Loss	0.857
8	L	Win	0.900
9	L	Loss	0.818
10	L	Win	0.871
11	R	Loss	0.931
12	R	Win	0.900

Based on the results from all 12 rounds, I believe the 12-sided die is in the _____ hand.

Note: I had a computer simulate this game and it turned out the 12-sided die was in the left hand.

Recap: Bayesian information takes advantage of prior information. No matter what we're investigating, we could take advantage of information from prior studies or physical models.

This process naturally integrates data as we collect it and updates our prior beliefs.

As we'll learn, Bayesian methods avoid counterintuitive notions like p-values (the probability of observing something as or more extreme than what we observed | the null hypothesis is true). Instead, Bayesian methods base decisions on posterior probabilities (the probability that the null hypothesis is true | our observed data). This is something we'll explore throughout the semester.

Bayesian methods depend on good prior probabilities, which can be difficult to obtain.

Scenario: The American Cancer Society estimates that about 1.7% of women have breast cancer.
(Source: http://www.cancer.org/cancer/cancerbasics/cancer-prevalence)

The Susan G. Komen For the Cure Foundation reports that mammography correctly identifies about 78% of women who truly have breast cancer.
(Source: http://ww5.komen.org/BreastCancer/AccuracyofMammograms.html)

A 2003 article suggests that up to 10% of all mammograms are false positives.
(Source: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1360940)

13. A woman undergoes a routine mammogram and finds a positive result (indicating she has breast cancer). Given this positive result, what's the probability this woman has breast cancer? You may want to use a table.

Prior probability that the woman has breast cancer = _____

Updated probability = P(breast cancer | positive result) = _____

	+ Result	– Result	Total
cancer			
No			
Total			

14. Using this updated probability of breast cancer, calculate the probability that she has breast cancer given a second test yields a positive result? Once again, you may want to use a table like the one shown below.

Updated probability = P(breast cancer | a second positive result) = _____

	+ Result	– Result	Total
cancer			
No			
Total			

Scenario: Suppose a student in this class send me an email containing the word "radom."

Having never seen the word "radom" before, I guess the student could have either:

- intended to write the word "radom"
- intended to write the word "random" but, due to a typo, forgot the "n"
- intended to write the word "radon" but, due to a typo, changed the "n" to an "m"

We're going to attempt to find the likelihood of the student doing each of those three things.

Source: Gelman, A., et. al (2013). Bayesian Data Analysis, 3rd edition. ISBN: 978-1439840955

15. Let's first take a look at the probability we will compute. Using Bayes' Theorem, we have:

$$P(\text{intended word} \mid \text{"radom"}) = \frac{P(\text{"radom"} \mid \text{intended word})P(\text{intended word})}{P(\text{"radom"})}$$

Take a look at the denominator. How can we calculate the probability that the student typed "radom?" In this scenario, there are 3 different ways the student could have written that word. He could have written it (1) on purpose, (2) mistakenly instead of "random," or (3) mistakenly instead of "radon." Using the Law of Total Probability, we can expand the denominator of this formula and rewrite Bayes' formula:

$$\frac{P(\text{"radom"} \mid \text{intended word})P(\text{intended word})}{P(\text{"radom"} \mid \text{intended radom})P(\text{intended radom}) + P(\text{"radom"} \mid \text{intended random})P(\text{intended random}) + P(\text{"radom"} \mid \text{intended radon})P(\text{intended radon})}$$

We can then use this formula 3 times to estimate the likelihood of each of our 3 scenarios.

But how can we calculate the numerator of this formula? How could we possibly know the probability of a student typing radom, random, or radon in an email? Researchers at Google found the relative frequencies of these words in a large database of emails. They found the following probabilities:

P(radom) = 0.000000312
P(random) = 0.0000760
P(radon) = 0.00000605

Wait a second! How could the probability of "radom" be so close to the probabilities of the other two words? To figure this out, a quick look at Wikipedia showed me that "radom" is a medium-sized city in Poland (home to the largest air show in Poland) and the word is also an unofficial name for a semiautomatic 9mm Para pistol of Polish design. It looks like some people really do write "radom" intentionally in emails.

To use Bayes' formula, we also need to know P("radom" | intended word) for each of our 3 intended words. Thankfully, Google comes to our rescue once again. Researchers at Google provide the following conditional probabilities (based on their spelling and typing errors model):

P("radom" | radom) = 0.975
P("radom" | random) = 0.00193
P("radom" | radon) = 0.000143

To me, these values seem reasonable. There's a 97.5% that the word was typed correctly, a 0.2% chance the word "random" was misspelled, and a 0.01% chance the word "radon" was misspelled. I'll trust that these are good estimates of these conditional probabilities.

Using the given information and Bayes' formula, calculate each of the following:

$P(\text{intended "radom" | "radom"}) =$ _____

$P(\text{intended "random" | "radom"}) =$ _____

$P(\text{intended "radon" | "radom"}) =$ _____

Instead of using the formula for Bayes' theorem, you can try the calculations via a table:

	Intended radom	Intended random	Intended radon
Use *radom* in email			
Use *random* in email			
Use radon in email			

Based on your calculations, which word did the student most likely intend to type in the email?

The student most likely intended to type the word: _____

16. If you did the calculations correctly, you should have found the likelihood of the student intending to write "radom" was more than twice as much as the likelihood that the student intended to write either "random" or "radon." Do you really believe the most likely scenario is one in which the student (a student from this class) intended to write "radom"? If not, explain what went wrong. Do you have reason to not believe in some of the given information supplied in this scenario?