208269: Statistics for Computer Science
Problem Set #1

July 10, 2023

- Submit on Mango by 23:59 on Monday, July 24th.
- **Collaboration policy:** You are encouraged to discuss problem-solving strategies with each other as well as the course staff, but you must write up your own solutions and submit individual work.
- **For each problem, briefly explain/justify how you obtained your answer.** Brief explanations of your answer are necessary to get full credit for a problem even if you have the correct numerical answer. The explanations help us determine your understanding of the problem whether or not you got the correct answer. Moreover, in the event of an incorrect answer, we can still try to give you partial credit based on the explanation you provide. It is fine for your answers to include summations, products, factorials, exponentials, or combinations; you don't need to calculate those all out to get a single numeric answer.
- If you handwrite your solutions, you are responsible for making sure that you can produce **clearly legible** scans of them for submission. You may also use any word processing software you like for writing up your solutions.

1. Say in Silicon Valley, 35% of engineers program in Java and 28% of the engineers who program in Java also program in C++. Furthermore, 40% of engineers program in C++.

    a. What is the probability that a randomly selected engineer programs in Java and C++?

    b. What is the conditional probability that a randomly selected engineer programs in Java given that they program in C++?

Answer.

2. A website wants to detect if a visitor is a robot or a human. They give the visitor five CAPTCHA tests that are hard for robots but easy for humans. If the visitor fails one of the tests, they are flagged as a robot. The probability that a human succeeds at a single test is 0.95, while a robot only succeeds with probability 0.3. Assume all tests are independent. The percentage of visitors on this website that are robots is 5%; all other visitors are human.

   a. If a visitor is actually a robot, what is the probability they get flagged (the probability they fail at least one test)?

   b. If a visitor is human, what is the probability they get flagged?

   c. Suppose a visitor gets flagged. Using your answers from part (a) and (b), what is the probability that the visitor is a robot?

Answer.

3. Two cards are randomly chosen without replacement from an ordinary deck of 52 cards. Let E be the event that both cards are Aces. Let F be the event that the Ace of Spades is one of the chosen cards, and let G be the event that at least one Ace is chosen.

   a. Compute $P(E \mid F)$.

   b. Are E and F independent? Justify your answer using your response to part (a).

   c. Compute $P(E \mid G)$.

Answer.

4. The color of a person's eyes is determined by a pair of eye-color genes, as follows:

   - if both of the eye-color genes are blue-eyed genes, then the person will have blue eyes
   - if one or more of the genes is a brown-eyed gene, then the person will have brown eyes

   A newborn child independently receives one eye-color gene from each of its parents, and the gene it receives from a parent is equally likely to be either of the two eye-color genes of that parent. Suppose William and both of his parents have brown eyes, but William's sister (Claire) has blue eyes. (We assume that blue and brown are the only eye-color genes.)

   a. What is the probability that William possesses a blue-eyed gene?
   b. Suppose that William's wife has blue eyes. What is the probability that their first child will have blue eyes?

Answer.

5. Consider the following two investment options and their gain/loss forecasts from various experts:

   Stock

   - Complete loss: 40% chance
   - No gain or loss: 15% chance
   - 100% gain: 15% chance
   - 400% gain: 15% chance
   - 900% gain: 15% chance

   Exchange-traded fund (ETF)

   - Complete loss: 1% chance
   - No gain or loss: 35% chance
   - 10% gain: 59% chance
   - 20% gain: 5% chance

   (a) What is the expectation of investing ฿1,000 in the stock?

   (b) What is the expectation of investing ฿1,000 in the exchange-traded fund?

   (c) Would you rather invest in the stock or the exchange-traded fund? Explain your answer.

Answer.

6. **[Coding]** After the Ebola outbreak of 2015, there was an urgent need to learn more about the virus. You have been asked to uncover how a particular group of bat genes impact an important trait: whether the bat can carry Ebola. Nobody knows the underlying mechanism; it is up to you to hypothesize what is going on. For 100,000 independently sampled bats you have collected data of whether or not five genes are expressed, and whether or not the bat can carry Ebola. You can find the data in a file called `bats.csv`. A value of 1 denotes True, whereas a value of 0 denotes False. Each row in the file corresponds to **one bat** and has 6 columns representing Boolean values:

   - Boolean 0: Whether the 1st gene is expressed in the bat (G_0)
   - Boolean 1: Whether the 2nd gene is expressed in the bat (G_1)
   - Boolean 2: Whether the 3rd gene is expressed in the bat (G_2)
   - Boolean 3: Whether the 4th gene is expressed in the bat (G_3)
   - Boolean 4: Whether the 5th gene is expressed in the bat (G_4)
   - Boolean 5: Whether the trait is expressed in the bat; i.e., the bat can carry Ebola (T)

   Follow the instructions in each subpart of this question to either write code or submit written answers. You'll write code in the file `stat269_pset2.py`. Submit only that file to Mango, and do not modify the name of that file

   You will submit parts (a) and (b) as code, and write up answer to part (c) in this document.

   (a) **(Coding)** Calculate $P(T = 1)$ along with $P(G_i = 1)$ for each gene i. In this part, you will implement the function `calculate_probs`. Your function should return a numpy array with shape `(6,)`. The elements at indices 0 through 4 will be $P(G_0 = 1), P(G_1 = 1), \ldots, P(G_4 = 1)$, respectively, and the element at index 5 will be $P(T = 1)$.

   Here are some Python tips that you might find useful:

   - We strongly recommend using numpy in this (and subsequent) questions. You can load a csv file into a numpy array named data by using the function `np.genfromtext`:

     ```
     data = np.genfromtxt(filename, delimiter=',')
     ```

   - You can get the `i`-th column from data using slicing by writing `data[:,i]`. That returns an array of shape `(n, )`, where `n` is the number of rows in the csv file.
   - You can take the mean of a numpy array using the `np.mean` method. Example:

     ```
     arr = np.array([1, 2, 3])
     print(arr.shape)    # Output: (3, )
     print(np.mean(arr)) # Output: 2.0
     ```

   - **If you leverage the `axis` parameter in `np.mean`**, your function will be just a few lines long.

   (b) **(Coding)** For each gene i, calculate $P(T = 1|G_i = 1)$. In this part, you'll implement the function `calculate_cond_probs`. Your function will return a numpy array with shape `(5, )`, where the element at index i is $P(T = 1|G_i = 1)$. Some more Python tips to complement the ones listed above:

- Check out the `np.where` method. Specifically, if your csv is stored in a numpy array named data, you can store a subset of the rows where the `i`-th column is:

```
subset = data[np.where(data[:, i] == 1)]
```

- Check out the `np.logical_and` method.

(c) **(Writing)** For each gene i, decide whether or not you think that is would be reasonable to assume that G_i is independent of T. Support your argument with numbers. Remember that our probabilities are based on 100,000 bats, not infinite bats, and are therefore only estimates of the true probabilities.

Answer (only write your answer to part (c))**.**