1. Calculate the probability of flipping a balanced coin four times and getting each pattern: HTTH, HHHH and TTHH.

Since flipping the coin is independent event each time. The probability of getting head or tail each time is same, which is 0.5. Therefore, see below

HTTH = 0.5 x 0.5 x 0.5 x 0.5 = 0.0625

HHH = 0.5 x 0.5 x 0.5 x 0.5 = 0.0625

TTHH = 0.5 x 0.5 x 0.5 x 0.5 = 0.0625

1. If a list of people has 24 women and 21 men, then the probability of choosing a man from the list is 21/45. What is the probability of not choosing a man?

The probability of not choosing a man is 24/45

1. The probability that Bernice will travel by plane sometime in the next year is 10%. The probability of a plane crash at any time is .005%. What is the probability that Bernice will be in a plane crash sometime in the next year?

The probability is 10% x 0.005% = 0.00005%

1. A data scientist wants to study the behavior of users on the company website. Each time a user clicks on a link on the website, there is a 5% chance that the user will be asked to complete a short survey about their behavior on the website. The data scientist uses the survey data to conclude that, on average, users spend 15 minutes surfing the company website before moving on to other things. What is wrong with this conclusion?

Surfing the company website may be not an independent event since there are 5% of user will take time to finish the short survey which the time is part of the 15 minutes that they use to surf the website.

Now it's time to use Bayes' rule to compute some conditional probabilities. First look over the numbers and estimate each of the four probabilities, using your intuition. Then, calculate the probabilities using Bayes' rule. Keep track of your work in a Google document or markdown file that you can share with your mentor.

A diagnostic test has a 98% probability of giving a positive result when applied to a person suffering from Thripshaw's Disease, and 10% probability of giving a (false) positive when applied to a non-sufferer. It is estimated that 0.5 % of the population are sufferers. Suppose that the test is now administered to a person whose disease status is unknown. Calculate the probability that the test will:

1. Be positive
2. Correctly diagnose a sufferer of Thripshaw's
3. Correctly identify a non-sufferer of Thripshaw's
4. Misclassify the person
5. Positive = P(Positive/Sufferer) \* P(Sufferer) + P(Positive/non-sufferer) \* P(Non-sufferer)

P = 0.98 x 0.005 + 0.1 x 0.995 = 0.1044

1. Already given in the question, which is 0.98
2. Already given in the question, which is 0.9
3. Correctly classify = P(positive/sufferer) \* P(sufferer) + P(correct/non-sufferer) \* P(non-sufferer)

Correctly classify = 0.98 x 0.005 + 0.9 x 0.995 = 0.9004

Misclassify = 1 – 0.9004 = 0.0996

In each of the scenarios, find possible shortcomings of the theoretical or actual data sources to answer the given question. What could be done to either adjust the analysis or reframe the question so that you can answer it accurately?

Data Source: Amsterdam availability data scraped from AirBnB on December 24th. Question: What are the popular neighborhoods in Amsterdam?

Probably there is no availability for this data since this is the date before Christmas and Airbnb is fully booked. So it should scraped data from Airbnb on a random date, or multiple random dates.

Data Source: Mental health services use on September 12, 2001 in San Francisco, CA and New York City, NY. Question: How do patterns of mental health service use vary between cities?

First this data source is a bit outdated so it’s not suitable for recent research. Second this data source only includes San Francisco and NYC, which is not enough to represent all other cities in the U.S. Third, September 12, 2001 is the date after 9/11, which will change the data of NYC dramatically. The data source should be pulled before 9/11.

Data Source: Armenian Pub Survey. Question: What are the most common reasons Armenians visit local pubs?

Hangout with friends. However, most of the samples in this data source are students. Therefore, should change to What are the most common reasons American students visit local pubs.

|  |  |  |
| --- | --- | --- |
| Door 1 | Door 2 | Door 3 |
| Goat | Goat | Car |
| Goat | Car | Goat |
| Car | Goat | Goat |

Three doors, you get above combination. For cobo#1, let’s assume we pick door 1, and the host knows the car is at door#3, so he can’t open door#3. The host opens door#2, if we stay, we get a goat, **if we switch, we get a car.**

Cobo#2, let’s pick door#1 again. Host cannot open door#2 because it’s a car, so he opens door#3, it’s a goat. If we stay, we get a goat, **if we switch, we get a car.**

Combo#3, let’s also pick door#1. Host will open either door #2 or door#3, both are goat. If we stay, we get a car. **If we switch, we get a goat.**

If we look at the example above, 2 out of 3 switches, we get a car, and only 1 switch gets a goat. The probability of switching and getting a car is 2/3, and staying and get a car is 1/3. So mathematically, switching have the higher probability to get a car. This is very similar to penalty kick in soccer. If the goalkeeper stays in the middle, and not going to left or right, the goalkeeper has higher probability to block this penalty kick. But in real life, goalkeepers normally don’t stay in the middle.