**Drill set 1**

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

Since the variables are independent, and have the same probability each, the probability for all of them is the same:

P(H) \* P(T) \* P(T) \* P(H) = (0.5)^4 = 0.0625 (*Or %6.25*)

P(H) \* P(H) \* P(H) \* P(H) = (0.5)^4 = 0.0625 (*Or %6.25*)

P(T) \* P(T) \* P(H) \* P(H) = (0.5)^4 = 0.0625 (*Or %6.25*)

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?

45/45 – 21/45 = 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?

P(Crash) \* P(Travel) = 0.005/100 \* .1 = 0.000005 (*Or 0.0005%*)

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?

The possibility of a user taking a survey, which is 5%, applies to clicking one link. Therefore, there is a 95% chance that each time a user clicks the link they will not take a survey.

This means that if we assume each user clicks one link, then the data scientist’s sample consists of only %5 of the population. Also, users who click more links are more likely to be surveyed, so the sample is likely to be biased towards users who actually spend more time on the website, and therefore this sample is not truly representative of the population.

**Drill set 2**

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:

P(Positive) = (P(Thripsaw)\*P(Positive | Thripsaw)) + (P(Healthy)\*P(Positive | Healthy))

P(Positive) = (0.005 \* 0.98) + (0.995 \* 0.1) = 0.1044 (*Or 10.44%*)

1. Correctly diagnose a sufferer of Thripshaw's:

P(Positive | Thripsaw) = 98%

1. Correctly identify a non-sufferer of Thripshaw's:

***## My own first try ##***

P(Healthy | Negative) = [P(Negative | Healthy) \* P(Healthy)]/ [(P(Healthy)\*P(Negative | Healthy)) + (P(Thripsaw)\*P(Negative | Thripshaw))]

P(Healthy | Negative) = [0.1\*0.995]/[(0.995\*0.1) + (0.005\*0.02) ]

P(Healthy | Negative) = [0.0995]/[0.0996] = 0.998 (Or %99.8)

***## After seeing the result is 0.9 and thinking how to get that result ##***

The probability of incorrectly identifying a healthy person is already given as 10%, this means that the only other probability is:

P(Healthy | Negative) = 1 – 0.1 = 0.9 (Or 90%)

1. Misclassify the person:

We questions 2 and 3 we have calculated probabilities of correctly classifying both Thripshaw and healthy people. However they cannot simply be added together because the probability would be more than 1, which is impossible. So I conclude that correct classification would be the probability of correctly classifying the condition WHEN that condition exists, translated as conditional probability of correctly classifying the case TIMES the probability of the case itself, like so:

P(Misclassify) = 1- P(Correctly classify)

P(Misclassify) = 1- (P(Healthy | Negative)\*P(Healthy) + P(Thripshaw | Positive)\*P(Thripshaw))

P(Misclassify) = 1 – (0.9\*0.995 + 0.98\*0.005) = 1 – 0.9004 = 0.0996 (Or 9.96%)