Unit 1 Homework

w203: Statistics for Data Science

This assignment is due at the time of your live session. Notice that there are two parts that are mandatory: a section called **Applied Practice**, and a section called **Proof Practice**. There is also an optional programming assignment, if you're interested.

Section 1: Applied Practice

For each of these questions, please do the work (probably with a pen and paper), and then enter your solutions into the Gradescope assignment slot that is called, "Unit 1 – Applied Practice."

1. (9 points total) Summer Roadtrip

At gas station on Route 66, 40% of customers use regular gas (event R), 35% use mid-grade gas (event M), and 25% use premium gas (event P). Of the customers that use regular gas, 30% fill their tanks (event F). Of the customers that use mid-grade gas, 60% fill their tanks, while of those that use premium, 50% fill their tanks. Assume that each customer is drawn independently from the entire pool of customers.

1. (3 points) What is the probability that the next customer will request regular gas and fill the tank?
2. (3 points) What is the probability that the next customer will fill the tank with any kind of gas?
3. (3 points) Given that the next customer fills the tank, what is the conditional probability that they use regular gas?

2. (12 points total) The Claaaaaw

Suppose that in a claw game at an arcade, there is a collection of toys that have the following characteristics:

- $2/5$ are red;
- $3/5$ are waterproof;
- $1/2$ are cool.

(When we write that $2/5$ are red, this means that $3/5$ are not red. But the facts above do not provide any information about the intersection between the different traits.)

Furthermore:

- $1/5$ are both red *and* waterproof;
- $1/5$ are both red *and* cool;
- $3/10$ are both waterproof *and* cool.

(When we write that $1/5$ are both red *and* waterproof, this contains no information about how cool they are.)

Finally:

- $1/10$ are neither red, waterproof, nor cool. (These are pretty lame toys.)

Since working those claws is so hard, suppose that any toy in the game has an equal chance of being selected.

1. (3 points) Draw an area diagram to represent these events. For as many of the events that you can, compute the probability (for example, $P(R \cap C \cap W)$)).
2. (3 points) What is the probability of drawing a toy that is red *and* waterproof *and* cool?
3. (3 points) Suppose that you pull out a toy at random, and you observe only the color, noting that it is red. Conditional on just this information, what is the probability that the toy is not cool?
4. (3 points) Given that a randomly selected toy is *either* red or waterproof, what is the probability that it is cool?

Section 2: Proof Practice

For this question please do the work (probably with pen and paper, although you should feel free to write a `.Rmd` solution if you're interested in the practice with), and then enter your solutions into the Gradescope assignment slot that is called "Unit 1 – Proof Practice."

Math proofs have a bad reputation: many of the instructors have negative memories of rote tasks about proving the Pythagorean Theorem, or some other such task. Perhaps you do too?

However, in this course and throughout many parts of data science the work that you will do will involve arguing for why the approach that you have taken is the *best* of the available options. A math proof is just one of these forms of argument.

For the question that follows, a compelling argument – a *proof* – should show the following two facts:

a. That the value you arrive at is possible (for example, for the maximum possible value, you might state that the maximum possible value is 1); but, importantly,
b. That there are no ways to arrive at a value that is higher (or lower as the case may be) that the value you arrive at.

Is part (b) that really makes this question a proof.

3. (6 points total) On the Overlap of Two Events

Suppose for events A and B, $P(A) = \frac{1}{2}$, $P(B) = \frac{3}{4}$, but we have no more information about the events.

1. (3 points) What are the maximum and minimum possible values for $P(A \cap B)$?
2. (3 points) What are the maximum and minimum possible values for $P(A|B)$?

(Optional) Programming Practice

This question is an optional question that allows you to solve a short math question and then encode your solution in a function. We are not going to grade this question, but we will provide our solution after the submission window.

4. (0 points total) Testing for Coronavirus

What we learn from a statistical test depends crucially on the population prevalence of the disease being tested for. Suppose that you are interested in a rapid assay for the coronavirus. Let **T** be the event that the **T**est comes back positives, **C** be the event that an individual has **C**oronavirus and **P** be the population prevalence of the disease.

If a person has coronavirus, the test gives the correct response with probability 0.94. If a person does not have coronavirus, the test gives the correct response with probability 0.96.

1. (0 points) You are interested in the *false discovery rate*, meaning the conditional probability that a person does not have coronavirus, given that their test is positive. Write a function **in R** that takes population prevalence as an argument and returns the false discovery rate. Check that your function works by comparing to the table below; provide the output of this function call so that your grader can verify it. The table is drawn from this article [link here] (./ebell_2020.pdf) published in the journal *American Family Physician* (Unfortunately, the paper has an error. It claims that these numbers are false positive rates; they are actually false discovery rates).

Population Prevalence	Cellex Test
1%	80.8%
5%	44.7%
10%	27.7%
20%	14.5%
30%	9.0%
50%	4.1%
70%	1.8%
90%	0.5%

```
false_discovery_rate <- function(population_prevalence) {

}
```

2. (0 points) Using the function that you have just written and the data supplied in the object `d` below, create a plot, using `ggplot` that has the following characteristics:
    - On the x-axis: The population prevalence rate
    - On the y-axis: The false discovery rate
    - Meaningful axis and plot titles

```
d <- data.frame(
  population_prevalence = seq(from = 0, to = 100, by = 0.1)
)
```

```
# d %>%             # fill this in
#   mutate() %>%    # with code that will
#   ggplot() +      # produce the desired plot
#   ...
```

- **for related information,Click here (https://en.wikipedia.org/wiki/Sensitivity_and_specificity)** and here (https://en.wikipedia.org/wiki/Base_rate_fallacy)