# Unit 1 Homework

w203: Statistics for Data Science

## 0. Homework infrastructure

We encourage you to try writing your solutions using Latex. [This cheatsheet] has most of the symbols you'll need. Latex is a useful language for writing equations, and we may use it to jot down ideas in live session.

It does take time to get used to LaTeX, so you may also handwrite your solutions.

Please use a separate page or a separate file for each question. You will submit four different files on Gradescope.

## 1. (9 points total) Gas Station Analytics

At a certain gas station, 40% of customers use regular gas (event R), 35% use mid-grade (event M), and 25% use premium (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?
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 Toy Bin

Suppose that there is a collection of toys that have the following characteristics:

- 2/5 are red, 3/5 are waterproof; 1/2 are cool.
- 1/5 are red and waterproof; 1/5 are red and cool; 3/10 are waterproof and cool.
- 1/10 are neither red, waterproof, nor cool.
- Each toy has an equal chance of being selected.

With this as the setup, answer the following questions:

1. (3 points) Draw an area diagram to represent these events.
2. (3 points) What is the probability of getting a red, waterproof, cool toy?
3. (3 points) 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 red or waterproof, what is the probability that it is cool?

## 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 $P(A|B)$?

## 4. (6 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 (sometimes called sensitivity). If a person does not have coronavirus, the test gives the correct response with probability 0.96 (sometimes called specificity).

1. (3 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 that takes population prevalence as an argument and returns the false discovery rate. Check that your function works by comparing to the table below, drawn from this article [link here] 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% |

```r
false_discovery_rate <- function(population_prevalence) {

}
```

3. (3 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

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

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