Computational Statistics
Assignment 1

by James D. Wilson (University of San Francisco)

Directions: For all questions in this assignment, fully answer any question that is asked. Late assignments will automatically have 10 points deducted for each day that they are late.

Reading Questions

Read Chapter 1 of Doing Data Science by Cathy O'Neil and Rachel Schutt entitled "What is Data Science?" here: https://www.safaribooksonline.com/library/view/doing-data-science/9781449363871/ch01.html.

1. From reading the article, what in your opinion is the biggest issue facing the field of "data science," and what do you think can be done to help fix this issue?

2. What, in your opinion, is the biggest difference between traditional statistics and data science? How does technology play a role in this difference?

3. Who is responsible for coining the term "data scientist?"

4. What, in your opinion, is the biggest difference between data scientists in academia and industry?

5. Based on this article, a major obstacle in understanding data science in the first place is understanding what data science actually *is*. From what you've read, and what you understand so far, how would you define "data science"?

Quantitative Questions

1. Let E, F, and G be three events. Find expressions for the events so that, of E, F, and G,
   (a) both E and G, but not F, occur
   (b) at least two of the events occur
   (c) at most one of the events occurs

2. Find the simplest expression for the following events:
$$(E \cup F) \cap (E^c \cup F) \cap (E \cup F^c)$$

3. Let $(S, \mathbb{P})$ be a valid probability model. Using the axioms of probability, prove that $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$ for any event $E \subset S$. Then use this to show that $P(\emptyset) = 0$.

4. Consider the experiment of flipping a coin. The sample space of this experiment is $S = \{H, T\}$ and possible events include $\{H\}$, $\{T\}$, S, $\emptyset$. Fix $p \in (0, 1)$ and assign $\mathbb{P}(\{H\}) = p$, $\mathbb{P}(\{T\}) = 1 - p$, $\mathbb{P}(S) = 1$, $\mathbb{P}(\emptyset) = 0$.
   (a) Verify that the three axioms of probability hold for $\mathbb{P}(\cdot)$.

(b) If $\mathbb{P}(\{T\}) = 1 - \frac{p}{2}$, which of the axioms are not satisfied?

5. A recent college graduate is planning to take the first 3 actuarial exams starting with the first one in June. If she passes this she will take the 2nd one in July and if she passes the 2nd exams she will take the third in September. If she fails an exam then she is not allowed to take the next exam(s). Let F (S, T) be event she passes the first (respectively second and third exam). Then,

$$P(F) = .9, \qquad P(S|F) = .8, \qquad P(T|F \cap S) = .7$$

   (a) Compute the probability she passes all three exams.

   (b) What is the conditional probability she failed the 2nd exam, if we know that she does not pass all three exams?

6. Suppose that we toss 2 fair dice (with 6 sides each).

   (a) Let E_1 denote the event that the sum of the dice is 6 and F denote the event that the first die equals 4. Are E_1 and F independent?

   (b) Now, suppose that we let E_2 be the event that the sum of the dice equals 7. Is E_2 independent of F?

7. Go back to the *Conditional probability and medicine* example on slide 38 of Lecture 2. Show that the probability that the patient has breast cancer is actually 0.09.

1 Computational Questions

1. **The Monty Hall Problem**: Suppose you're given the choice of three doors: behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to switch to door No. 2?" The question is whether or not it to your advantage to switch your choice.

   (a) Simulate an experiment that tests this question and calculate your "best guess" of what your probability of winning is if you switch doors. Note that whether or not you win is a random variable, so histograms are helpful here. Is this surprising?

   (b) Analytically calculate the probability of winning if you switch doors.

2. **The Birthday Problem**: Our class has around 30 students. Assume that all birthdays are equally likely and you may assume no one in the class was born on February 29th. You are interested in calculating the probability that there are at least two people with the same birthday.

   (a) Simulate an experiment that enables you can estimate the probability that there are at least two people with the same birthday in a class of 30. What probability do you obtain?

   (b) Generalize the above simulation for a class of size n, for any $n > 1$. By estimating the probabilities for a range of n, what is the minimum size of class required for there to be a probability of 0.50 or higher of having two students with the same birthday?