Worksheet 08 (Solutions)

1. (Statistical Inference) In pairs (or triples), select six marbles of one color (we will call this color C) and four marbles of another color. In one of the bags (we will call this bag F, for the first bag), place 4 C and 1 non-C marbles. In the other bag, place 2 C and 3 non-C marbles. We are going to consider an experiment where you select one bag at random and then randomly select a marble from the choosen bag. The idea is that we want to see how well you can estimate which bag the marble came from based on the color of the marble. (a) Compute the two-by-two table of probabilities where C is the event of selecting a marble of color C and F is the event of selecting the first bag. (b) What are the probabilities $\mathbb{P}(F|C)$ and $\mathbb{P}(F^c|C)$? (c) What are the probabilities $\mathbb{P}(F|C^c)$ and $\mathbb{P}(F^c|C^c)$? (d) What is the best guess for a bag if you have a C marble and what is the best guess if you have a non-C marble? (e) Let R be the event that your guess of the correct bag is right. What is $\mathbb{P}(R)$?

(f) Now, we are going to simulate the game 12 times.[1] To do this, have one person close their eyes. The other person rolls the die. If odd, they select the first bag and hand it to the other person, who then selects a marble. If even, the pick the second bag and give that instead. Keep track of the number of correct guesses. Switch roles mid-way through the simulation. We will aggregate across the class and see if we can get close to the analytical answer.

Solution: (a) The table will just be the number of marbles in each intersection, divided by the total number of marbles (10). So:

	C	C^c	Total
F	0.4	0.1	0.5
F^c	0.2	0.3	0.5
Total	0.6	0.4	1.00

Parts (b) and (c) come right off of the table. You can either calculate

[1] It may seem silly to go through this exercise, but I find it really helpful to have the perspective of the guesser in which you are updating your understanding of the bag probabilties with the data.

each pair of probabilities, or realize that $\mathbb{P}(F^c|C) = 1 - \mathbb{P}(F|C)$.

$$\begin{aligned}
\mathbb{P}(F|C) &= \frac{\mathbb{P}(F \cap C)}{\mathbb{P}(C)} = \frac{.4}{.6} = \frac{2}{3} \\
\mathbb{P}(F^c|C) &= \frac{\mathbb{P}(F^c \cap C)}{\mathbb{P}(C)} = \frac{.2}{.6} = \frac{1}{3} \\
\mathbb{P}(F|C^c) &= \frac{\mathbb{P}(F \cap C^c)}{\mathbb{P}(C^c)} = \frac{.1}{.4} = \frac{1}{4} \\
\mathbb{P}(F^c|C^c) &= \frac{\mathbb{P}(F^c \cap C^c)}{\mathbb{P}(C^c)} = \frac{.3}{.4} = \frac{3}{4}
\end{aligned}$$

From this, and just some basic intuition, we see that (d) we would guess the first bag if we have the color C and the second bag if we have the color not-C.

Part (e) is a bit trickier. We need to split the probability up into the two different bags, keeping in mind that the probability of picking either bag is 0.5:

$$\begin{aligned}
\mathbb{P}(R) &= \mathbb{P}(R \cap F) + \mathbb{P}(R \cap F^c) \\
&= \mathbb{P}(R|F) \cdot \mathbb{P}(F) + \mathbb{P}(R|F^c) \cdot \mathbb{P}(F^c) \\
&= \frac{1}{2} \cdot [\mathbb{P}(R|F) + \mathbb{P}(R|F^c)]
\end{aligned}$$

We will be right when we sampled from the first bag if we selected a C marble (probability of 0.8) and right when sampling from the second bag if we selected a non-C marble (probability 0.6). So:

$$\begin{aligned}
\mathbb{P}(R) &= \frac{1}{2} \cdot [\mathbb{P}(R|F) + \mathbb{P}(R|F^c)] \\
&= \frac{1}{2} \cdot [0.8 + 0.6] \\
&= 0.7
\end{aligned}$$

So, you should be correct 70% of the time with your guess. Much better than guessing by chance.

(f) There will be some variation with your specific 12 trials. Aggregating across the class should yield something close to the correct probability.

2. (Simpson's Paradox) There are two physicians named Dr. A and Dr. Z. Each of them performs two types of procedures: band-aid removal (B) and heart surgery (H). Their recent performance is given by the following tables:

	Heart	Band-Aid
Success	70	10
Failure	20	0

Dr. A

	Heart	Band-Aid
Success	2	81
Failure	8	9

Dr. Z

For this question, we will use the *empirical probability* of each event. That is, the value of every probability $\mathbb{P}E$ is given by the proportion of procedures from the data for which E occurs. (a) Compute the probability that a procedure done by Dr. A is successful and the probability that a procedure done by Dr. Z is successful. Who seems to be the better physician? (b) Compute the probabilities that each procedure is successful, conditioned on the doctor doing the procedure and which procedure is being done. Who seems to be the better physician now? (c) What paradox seems to exist? Can you explain why this happens?

Solution: The first two parts are just counting. (a) We have:

$$\mathbb{P}[S|A] = \frac{80}{100} = 0.80$$
$$\mathbb{P}[S|Z] = \frac{83}{100} = 0.83$$

So it seems that Dr. Z has a slightly higher percentage of their procedures being successful. For (b), we have:

$$\mathbb{P}[S|A \cap B] = \frac{10}{10} = 1.00$$
$$\mathbb{P}[S|Z \cap B] = \frac{81}{90} = 0.90$$
$$\mathbb{P}[S|A \cap H] = \frac{70}{90} = 0.78$$
$$\mathbb{P}[S|Z \cap H] = \frac{2}{10} = 0.20$$

So Dr. A seems to be better at both band-aid removal and heart surgery.

(c) The paradox, if you want to call it that, is that Dr. A seems to be better at both procedures but worse when we combine the data. A simple way to explain this is that Dr. A does a higher proportion of the heart surgery procedures than Dr. Z. Since heart surgery is slightly more difficult than band-aid removal, even though Dr. A is better at both, the mixture of their procedures causes Dr. Z's overall rate to be higher. This is a notorously common phenomenon in data science and is extremely difficult to identify. The general trouble is that we often do not know or have easy access to the confounding variable—here, the procedure type—and often need to have a deep understanding of the problem domain to understand what confounding variables may be missing from an analysis.[2]

3. (Monty Hall) This is a probability question so well-known my guess is that most of you have already heard of it. But let's see how we can answer it in a formal, systematic way. There are three doors, one randomly has a car behind it and the other two have goats. A contestant is playing a game in which they want to win the car. In the first round,

[2] The extreme values here and silly example of equating band-aid removal with heart surgury makes the result seem more obvious, which is the intent. These things actually happen in real-life medicine, where the best physicians get the most challenging cases and often have the worst outcome metrics. Similarly, experimental procedures may be done only on patients with the worst prognosis, making them look much worse even if they are actually better.

they stand in front of a door that the are thinking of picking. The host of the game, Monty Hall, selects one of the other doors that he knows has a goat behind it and opens it for everyone to see. The contestant now has to pick the door they actually want to open and then open it. What is the probability that they will win if they switch their choice from the first selection? Hint: We can assume that the contestant selects door 1.[3] Let C_1, C_2, and C_3 be the events that the car is behind door 1, 2, and 3, respectively and W be the event that the contestant wins if they switch their choice.

Solution: We want to get the probability $\mathbb{P}(W)$. We can split this like we did on our tables, but this time between three mutually exclusive events:

$$\begin{aligned}
\mathbb{P}(W) &= \mathbb{P}(W \cap C_1) + \mathbb{P}(W \cap C_2) + \mathbb{P}(W \cap C_3) \\
&= \mathbb{P}(C_1) \cdot \mathbb{P}(W|C_1) + \mathbb{P}(C_2) \cdot \mathbb{P}(W|C_2) + \mathbb{P}(C_3) \cdot \mathbb{P}(W|C_3) \\
&= \frac{1}{3} \times 0 + \frac{1}{3} \times 1 + \frac{1}{3} \times 1 \qquad\qquad\qquad\qquad\qquad\qquad = 0 + \frac{1}{3} + \frac{1}{3} \\
&= 2/3
\end{aligned}$$

So in general there is a 2/3 chance that switching doors will result in a win. The reason that $\mathbb{P}(W|C_1)$ is zero is because, if the car is behind door number 1, then switching will always lose. However, if the car is behind either of the other doors, switching will always win, so we get $\mathbb{P}(W|C_2) = \mathbb{P}(W|C_3) = 1$.

4. (Monty Hall, revisited) Consider a variation of the previous problem where there are seven doors, all equally likely to have the prize. In the second round, Monty Hall randomly selects three goat doors that are you not in front of to open. There is now a closed door you are in front of and three other remaining doors. What is the probability that you will win if you switch to one of the other three doors?

Solution: We can define C_j in the same way as before. Let's define W to be the event of winning if we switch. Then:

$$\begin{aligned}
\mathbb{P}(W) &= \mathbb{P}(W \cap C_1) + \mathbb{P}(W \cap C_2) + \cdots + \mathbb{P}(S \cap C_7) \\
&= \mathbb{P}(C_1) \cdot \mathbb{P}(W|C_1) + \sum_{i=2}^{7} \mathbb{P}(C_i) \cdot \mathbb{P}(W|C_i) \\
&= \frac{1}{7} \times \mathbb{P}(W|C_1) + \frac{1}{7} \times \sum_{i=2}^{7} \mathbb{P}(W|C_i)
\end{aligned}$$

As before, $\mathbb{P}(W|C_1)$ is zero, since not switching will always lose. How about the other doors? If we are conditioning on the event C_i, where $i \neq 1$, we know that switching will result in a win 1/3 of the time. Why?

[3] If not, relabel the doors so their choice is called door 1.

We are guessing between the three remaining open doors; there are three of them, and we are equally likely to get the answer correct. So:

$$\begin{aligned}\mathbb{P}(W) &= \frac{1}{7} \times \sum_{i=2}^{7} \mathbb{P}(W|C_i) \\ &= \frac{1}{7} \times \sum_{i=2}^{7} \frac{1}{3} \\ &= \frac{1}{7} \times \frac{6}{3} = \frac{2}{7} \approx 0.286\end{aligned}$$

As in the original problem, you double your chances by switching, but here the overall chances are lower.

5. (Hard!/Fun?) Consider an airplane with 100 seats assigned to each of 100 passengers. The first person to board has had too much to drink and selects their seat at random. Everyone else sits in their assigned seat unless it is already occupied, in which case they select a seat at random from the remaining empty seats. What is the probability that the last person to board will sit in their own seat?[4]

Solution: This is difficult question, but it is too fun not to put on a worksheet. It was asked as a question on the statistics Ph.D. qualifying exam the year before I took my qualifiers.

Trying to directly count the probabilities here is essentially impossible. The trick is to consider the following: what possible seats could the last person find themself sitting in? The only possible seats are (1) their own seat or (2) the first person's seat. Why? Passengers 2-99 always sit in their assigned seat if it's free when they board, so there is no way their seat could still be free when the last person boards. Now, notice that to everyone but the last person to board there is no difference between the first person's seat and the last person's seat. The first person just sits at random and everyone else only distinguishes their seat from everything else. Therefore, by symmetry, the probability that the last person is left with their own seat must be equal to the probability that the last person is left with the first person's seat. Since these are the only two options, the probabilities must add up to 1. Therefore, there is a probability of 0.5 that the last person sits in their own seat. The total number of seats is a red herring, of course, and the result is the same regardless of the total number of passengers.

6. (Prosecutor's Falacy) There are 10000 people living in a remote rural town. One night, a chicken is stolen from the town barn. The person who stole the chicken accidentally cut themselves on some barb wire escaping the scene, leaving just enough evidence to determine that

[4] Try to do this with a much smaller number of passengers and see if you can find a pattern. This question is better to do with some basic logic rather than formal probability manipulations. I include it here because it fits the general theme of the other questions.

the blood type of the thief is B positive. The next morning, a man is arrested for the crime who has B positive blood type. Based on knowledge that this type of blood (B positive) is only present in 8.5% of people in the US (and this town in particular), the prosectutor for the case argues that there is a 91.5% chance that arrested man commited the crime. (a) Write out the problem using some probabilistic notation. (b) Find a better measurement of the man's guilt, given that there is no other evidence against him.

Solution: (a) Let B be the event that a randomly selected person in the town is B positive and G be the event that the arrested man is guilty. We have $\mathbb{P}B = 0.085$ and $\mathbb{P}G = 0.0001$, the latter coming from the assumption that there is one person in the entire town who is guilty. We can also assume that $\mathbb{P}(B|G) = 1$, since we know the blood type of the guilty party. The prosecutor is trying to estimate the probability $\mathbb{P}(G|B)$.

(b) The correct calculation is given by Bayes's Rule:

$$\begin{aligned}\mathbb{P}(G|B) &= \mathbb{P}(B|G) \times \frac{\mathbb{P}(G)}{\mathbb{P}(B)} \\ &= 1 \times \frac{0.0001}{0.085} \approx 0.0012\end{aligned}$$

So, just a bit more than 0.1%. The prosecutor is making a *base rate falacy* in forgetting the account for the very low prior chance that any given individual is guilt, something that real-life prosecutor's seem to routinely have trouble understanding.