1. Probabilities are sensitive to the form of the question that was used to generate the answer.

   My neighbor has two children. Assuming that the gender of a child is like a coin flip, it is most likely, a priori, that my neighbor has one boy and one girl, with probability 1/2. The other possibilities—two boys or two girls—have probabilities 1/4 and 1/4.

   (a) Suppose I ask him whether he has any boys, and he says yes. What is the probability that one child is a girl?

   There are four possibilities $\{(B,B),(B,G),(G,B),(G,G)\}$. The (G,G) possibility is eliminated based on his response. There are two remaining outcomes that include at least one girl, so the probability is $\frac{2}{3}$.

   (b) Suppose instead that I happen to see one of his children run by, and it is a boy. What is the probability that the other child is a girl?

   The gender of the other child is independent from from the fact that one child is a boy. So here the possibilities are $\{B,G\}$, and the probability is simply $\frac{1}{2}$.

2. Legal Reasoning. Suppose a crime has been committed. Blood is found at the scene for which there is no innocent explanation. It is of a type which is present in 1% of the population.

   (a) The prosecutor claims: "There is a 1% chance that the defendant would have the crime blood type if he were innocent. Thus there is a 99% chance that he guilty". This is known as the **prosecutor's fallacy**. What is wrong with this argument?

   Let G indicate the outcome where the defendant is guilty, I the outcome where the defendant is innocent, and M the outcome where the defendant's blood type matches blood type at the scene. Clearly $P(G)+P(I)=1$ and $P(M)=0.01$. Also, it seems reasonable to assume that $P(M \mid G)=1$.

   The prosecutor's fallacy here is to confuse $P(I|M)$ with $P(M)$. $P(I|M)$ is shown below by Bayes Law.

   $$P(I \mid M) = \frac{P(M \mid I) \times P(I)}{P(M)}$$

   So to find $P(I|M)$, one also needs $P(M \mid I)$ and $P(I)$. And $P(G|M)$ would be 1 minus this value, not $1-P(M)$.

   (b) The defender claims: "The crime occurred in a city of 800,000 people. The blood type would be found in approximately 8000 people. The evidence has provided a probability of just 1 in 8000 that the defendant is guilty, and thus has no relevance." This is known as the **defender's fallacy**. What is wrong with this argument?

   The defender seems to be arguing that the $P(G)=1/800,000$. Then we could calculate as follows:

   $$P(G \mid M) = \frac{P(M \mid G) \times P(G)}{P(M)} = \frac{1 \times 1/800000}{1/100} = 1/8000$$

However, the point is that $P(G \mid M)$ is 100fold higher than $P(G)$. Is it relevant that the defendant is now 100 times more likely to have commited the crime than before the blood match was announced? I would think so.

3. Bayes rule for medical diagnosis. After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e., the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you dont have the disease). The good news is that this is a rare disease, striking only one in 10,000 people. What are the chances that you actually have the disease? (Show your calculations as well as giving the final result.)

   Let D indicate having the disease and T testing positive. We have:

$$P(D) = 1/10000 \qquad P(\overline{D}) = 9999/10000$$
$$P(T \mid D) = 99/100 \qquad P(T \mid \overline{D}) = 1/100$$

   Bayes law:

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)}$$

   The numerator is clearly $99/100 \times 1/10000$. What is the denominator?

$$P(T) = P(T|D)P(D) + P(T|\overline{D})P(\overline{D}) = (.99)(.0001) + (.01)(.9999) = 0.010098$$

   Putting all this together, we have

$$P(D|T) = \frac{(.99)(.0001)}{0.010098} = 0.009803922$$

   So the chances of actually having the disease is about 0.98%.

4. The Monty Hall problem. On a game show, a contestant is told the rules as follows:

   There are three doors, labelled 1, 2, 3. A single prize has been hidden behind one of them. You get to select one door. Initially your chosen door will not be opened. Instead, the gameshow host will open one of the other two doors, and he will do so in such a way as not to reveal the prize. For example, if you first choose door 1, he will then open one of doors 2 and 3, and it is guaranteed that he will choose which one to open so that the prize will not be revealed. At this point, you will be given a fresh choice of door: you can either stick with your first choice, or you can switch to the other closed door. All the doors will then be opened and you will receive whatever is behind your final choice of door.

   Imagine that the contestant chooses door 1 first; then the gameshow host opens door 3, revealing nothing behind the door, as promised. Should the contestant (a) stick with door 1, or (b) switch to door 2, or (c) does it make no difference? You may assume that initially, the prize is equally likely to be behind any of the 3 doors. Hint: use Bayes rule.

Let $D1$ indicate the prize is behind door 1, $D2$ indicate the prize is behind door 2, and $D3$ indicate the prize is behind door 3. Let $O3$ be the host opening door 3. The contestant has chosen door 1.

We know that$P(D1) = P(D2) = P(D3) = 1/3$**. We also know that** $P(O3|D1) = 1/2$**,** $P(O3|D2) = 1$**, and** $P(O3|D3) = 0$**. We again apply Bayes law and solve.**

$$P(D2|O3) = \frac{P(O3|D2)P(D2)}{P(O3)}$$

This time $P(O3)$ **expands as follows**

$$P(O3) = P(O3|D1)P(D1) + P(O3|D2)P(D2) + P(O3|C)P(O3) = 1/2 \times 1/3 + 1/3 + 0 = 1/2$$

Thus we have

$$P(D2|O3) = \frac{1/3}{1/2} = 2/3$$

So there is now a $2/3$ **chance that the prize is behind door 2, so the contestant should switch.**

5. Conditional Independence

   (a) Let $H \in \{1, \ldots, K\}$ be a discrete random variable, and let e_1 and e_2 be the observed values of two other random variables E_1 and E_2. Suppose we wish to calculate the vector

   $$\vec{P}(H \mid e_1, e_2) = (P(H = 1 \mid e_1, e_2), \ldots, P(H = K \mid e_1, e_2))$$

   Which of the following sets of numbers are sufficient for the calculation?

   i. $P(e_1, e_2), P(H), P(e_1 \mid H), P(e_2 \mid H)$

   ii. $P(e_1, e_2), P(H), P(e_1, e_2 \mid H)$

   iii. $P(e_1 \mid H), P(e_2 \mid H), P(H)$

   Bayes Law strikes again.

   $$\vec{P}(H \mid e_1, e_2) = \frac{\vec{P}(H \mid e_1, e_2) \times \vec{P}(H)}{P(e_1, e_2)}$$

   The terms in (ii) are sufficient then to solve for $\vec{P}(H \mid e_1, e_2)$**. The terms in (i) are insufficient. In particular we cannot obtain from** $P(e_1, e_2 \mid H)$ **from** $P(e_1 \mid H)$ **and** $P(e_2 \mid H)$ **unless we know they are independent, which we do not. And the terms in (iii) are strictly less than the terms in (i).**

(b) Now suppose we now assume $e_1 \perp e_2 \mid H$ (i.e., e_1 and e_2 are conditionally independent given H). Which of the above 3 sets are sufficent now?

Since $e_1 \perp e_2 \mid H$, we know that $P(e_1, e_2 \mid H) = P(e_1 \mid H) \times P(e_2 \mid H)$. Thus, the terms in (i) are now sufficient. Whether the terms in (iii) are sufficient comes down to whether we can determine $P(e_1, e_2)$ from the terms in (iii).

For each $h \in H$, we have

$$P(e_1, e_2) = \sum_h P(e_1, e_2, h) = \sum_h P(h) P(e_1, e_2 \mid h)$$

Again we can use the conditional independence assumption to show that

$$P(e_1, e_2) \sum_h P(h) P(e_1 \mid h) P(e_2 \mid h)$$

Thus even (iii) now has enough information to solve for $\vec{P}(H \mid e_1, e_2)$.

Show your calculations as well as giving the final result. Hint: use Bayes rule.

6. Do exercise 1.17.2 “Swirl” on page 24 of Winter 2020.

7. Do exercise 1.17.3 “Spot-the-Error #1” on page 24 of Winter 2020.

8. Do exercise 1.17.4 “Spot-the-Error #2” on page 25 of Winter 2020.

9. Do exercise 1.17.5 “Spot-the-Error #3” on page 25 of Winter 2020.

10. Do exercise 2.11.3. “Subsetting Data Frames with Tidyverse Function” on page 51 of Winter 2020.

11. Do exercise 2.11.4. “Creating a Pipeline” on page 51 of Winter 2020.

12. Do exercise 3.10.3. “Using the 68%-95% to Interpret Research Papers” on page 68 of Winter 2020.