Problem Set 1

Daniel Shapiro

Question 1 Background:

Our colleague Marc Meredith published a paper in 2020 estimating the prevalence of double voting in U.S. presidential elections. The core insight is that two John Smiths with the same birthday could very well be two different voters, because John Smith is a common name. Two people named Exa Dark Siderael (Elon Musk's daughter's name) with the same birthday are probably the same person voting twice.

Let's say we are given access to the voter files of American men. American men in this universe are all named Hugo or Gus (short for Augustus). We cannot see their birth year, but we know the date of their birth (e.g. July 23). We know the following about the distribution of the names Hugo and Gus:

a. All months are equally likely (eg, men are equally likely to be born in February as December).

b. Within any given month, every day is equally likely (eg, men are equally likely to be born on March 9 as March 26).

c. Hugo and Gus are equally popular names over the course of a given year (50% of men are Hugos, and 50% are Guses).

d. However, parents are more likely to name their child Augustus in August. 15% of Guses are born in August.

So, defining what we already know:

- P(Man born in August|Gus) = P(A|G) = $\frac{3}{20}$,
- P(G) = $\frac{1}{2}$,
- P(A) = $\frac{1}{12}$

1a: What is the overall probability of being born in August?

$\frac{1}{12}$: We already know that all months are equally likely.

1b: What is P(Name=Gus|Man born in August)? That is, conditional on being born in August, what is the probability that a man is named Gus?

By Bayes' Rule, we know that P(A|B) = $\frac{P(B|A)*P(A)}{P(B)}$. So here, we can do P(G|A) = $\frac{P(A|G)*P(G)}{P(A)} = \frac{\frac{3}{20}*\frac{1}{2}}{\frac{1}{12}} = \frac{\frac{3}{40}}{\frac{1}{12}} = \frac{36}{40} = \frac{9}{10} = 90\%$.

1c: What is P(Name=Gus|Man born in July)?

If we assume that in each month other than August, "Gus"es are spread out evenly, then we can say the following: if the likelihood of a "Gus" being born in August is 15% ($\frac{3}{20}$), the likelihood for the rest of the year is ($\frac{17}{20}$). Then, divided by the remaining 11 months, we get ($\frac{17}{220}$). We can write this as P(J|G), or the probability that a man was born in July, given that his name is Gus. This gives us enough information to use Bayes' Rule again:

$P(G|J) = \frac{P(J|G)*P(G)}{P(J)} = \frac{\frac{17}{220}*\frac{1}{2}}{\frac{1}{12}} = \frac{\frac{17}{440}}{\frac{1}{12}} = \frac{204}{440} = 46.36\%$.

1d: Imagine that there are 600 men born in 1982. Given the conditions above, what is the probability that at least two baby *Guses* are born on August 1, 1982?

1e: What is the probability that at least two baby *Hugos* are born on August 1, 1982?

1f: Given your answers to (d) and (e), what would you say to a pundit who pointed to two *Guses* with the same August 1, 1982 birthday as evidence of voter fraud in the form of double voting? What would you say to a pundit who pointed to two *Hugos* with the same August 1, 1982 birthday as evidence of voter fraud in the form of double voting?

I think that one cannot "conclude" anything either way, but there's a much higher chance that there are two Guses who were born on that day than that there are two Hugos.

Question 2 Background:

Does social environment affect the political development of preadults? There has been some work [1] *that shows that socializing agents do transmit certain political attitudes and values. Specifically, the authors argue that there exists a transmission of political values from parents to children, as manifested in their views during late adolescence. We will use this idea to solve the following questions.*

Take that a parent can share (or not) a certain political attitude – here, we'll say xenophobia. If the parent is xenophobic, the children will share this same view with probability $3/5$*. On the other hand, if the parent does not have this political view (is not xenophobic), the children will not have it either. Take a parent, who has probability* $1/3$ *of being xenophobic, and has* 2 *children.*

So, defining what we already know:

- P(Parent is Xenophobic) = P(PX) = $\frac{1}{3}$
- P(Parent is Not Xenophobic) = P(PNX) = $\frac{2}{3}$
- P(Children are Xenophobic | PX) = P(CX|PX) = $\frac{3}{5}$
- P(Children are Not Xenophobic | PNX) = P(CNX|PNX) = 1
- P(CNX|PX) = $\frac{2}{5}$ (just the inverse of P(CX|PX))
- P(CX|PNX) = 0 (just the inverse of P(CNX|PNX))

2a: What is the sample space for this experiment?

The sample space is: {(PX, C_1X, C_2X), (PX, C_1X, C_2NX), (PX, C_1NX, C_2X), (PX, C_1NX, C_2NX), (PNX, C_1X, C_2X), (PNX, C_1X, C_2NX), (PNX, C_1NX, C_2X), (PNX, C_1NX, C_2NX)}

2b: What is the probability that the younger child is xenophobic if their parent is xenophobic?

$\frac{3}{5}$ (as shown above)

[1] Jennings, M. K., and Niemi, R. G. (1968). "The transmission of political values from parent to child",* American Political Science Review *62(1), 169-184.

2c: What is the probability that neither child are xenophobic?

There are two options here. Either the parent is xenophobic and neither child is xenophobic, or the parent is not xenophobic and neither child is xenophobic. Either way, we have to multiply and add some probabilities together. First, if the parent is xenophobic, we know that there is a 40% chance that the child will not be xenophobic. So, using the information we know already (as illustrated above), we can multiply $\frac{1}{3} * \frac{2}{5} * \frac{2}{5}$, which equals $\frac{4}{75}$. Then, if the parent is not xenophobic, we know that there is a 100% chance that the child will also not be xenophobic, so we can say that $\frac{2}{3} * 1 * 1 = \frac{2}{3}$. $\frac{2}{3} = \frac{50}{75}$; $\frac{50}{75} + \frac{4}{75} = \frac{54}{75}$, or 72%.

2d: Is whether the elder child has the political attitude independent of whether the younger child shares it?

The question here comes down to: Does P($A \bigcap B$) = P(A) * P(B)?

We will define A as C_1X, which is comprised of {(PX, C_1X, C_2X), (PX, C_1X, C_2NX),(PNX, C_1X, C_2X), (PNX, C_1X, C_2NX)}. Then, we use the same operations as in 2c. So P(A) will equal $(\frac{1}{3} * \frac{3}{5} * \frac{3}{5}) + (\frac{1}{3} * \frac{3}{5} * \frac{2}{5}) + (\frac{2}{3} * 0 * 0) + (\frac{2}{3} * 0 * 100) = \frac{9}{75} + \frac{6}{75} = \frac{1}{5}$.

We will define B as C_2X, which is comprised of {(PX, C_1X, C_2X), (PX, C_1NX, C_2X), (PNX, C_1X, C_2X), (PNX, C_1NX, C_2X)}. Again, we use the same operations as in 2c, so P(B) will equal $(\frac{1}{3} * \frac{3}{5} * \frac{3}{5}) + (\frac{1}{3} * \frac{2}{5} * \frac{3}{5}) + (\frac{2}{3} * 0 * 0) + (\frac{2}{3} * 100 * 0) = \frac{9}{75} + \frac{6}{75} = \frac{1}{5}$.

P(A) * P(B) = $\frac{1}{5} * \frac{1}{5} = \frac{1}{25}$.

Meanwhile, $(A \bigcap B)$ = {(PX, C_1X, C_2X), (PNX, C_1X, C_2X)}. Thus, P($A \bigcap B$) = $\frac{9}{75} + 0$, which equals $\frac{9}{75}$ or $\frac{3}{25}$. $\frac{3}{25} \neq \frac{1}{25}$. So **NO**, they are not independent – at least mathematically.

2e: Conditional on the parent holding the attitude, is the elder child's viewpoint independent of the younger child's?

For this question, I just removed the sets from the sample space in which the parent was NOT xenophobic. Thus, I ended up with a sample space of:

{(PX, C_1X, C_2X), (PX, C_1X, C_2NX), (PX, C_1NX, C_2X), (PX, C_1NX, C_2NX)}

Now, same formula: Does Does P($A \bigcap B$) = P(A) * P(B)?

A is defined as C_1X, which is comprised of {(PX, C_1X, C_2X), (PX, C_1X, C_2NX)}. Since PX is now 100%, we can remove that from the probability, and only focus on working with the other probabilities. Here, our expression will look like P(C_1X) = $(\frac{3}{5} * \frac{3}{5}) + (\frac{3}{5} * \frac{2}{5}) = \frac{9}{25} + \frac{6}{25} = \frac{15}{25} = \frac{3}{5}$.

B is defined as C_2X, which is comprised of {(PX, C_1X, C_2X), (PX, C_1NX, C_2X)}. Our expression will look like: P(C_2X) = $(\frac{3}{5} * \frac{3}{5}) + (\frac{2}{5} * \frac{3}{5}) = \frac{9}{25} + \frac{6}{25} = \frac{15}{25} = \frac{3}{5}$.

P(C_1X) * P(C_2X) = $\frac{9}{25}$.

$(A \bigcap B)$ for this smaller sample space, meanwhile, is {(PX, C_1X, C_2X)}. P($A \bigcap B$) thus equals $\frac{9}{25}$ as well, meaning that conditional on the parent holding the attitude, the elder child's viewpoint **IS** independent of the younger child's.

2f: We find an elder child who is not xenophobic. How does this change your beliefs about whether the parent is xenophobic? Show numerically.

Here, we can use Bayes' Rule. We are looking for P(PX|CNX). We already know from 2d that P(CX)=$\frac{1}{5}$, so the inverse, P(CNX)=$\frac{4}{5}$. Thus, we can make the following equation:

$P(PX|CNX) = \frac{\frac{2}{5} * \frac{1}{3}}{\frac{4}{5}} = \frac{10}{60} = \frac{1}{6}$. Given that the original probability of the parent being xenophobic was $\frac{1}{3}$, we see here that if we can observe that one child is not xenophobic, the probability that the parent is xenophobic splits in half.

2g: We also find out that that the younger child does not share the value. How does this change your evaluation of whether the parent is xenophobic? Show numerically.

Here, we will define P(2CNX) as the probability that both children are not xenophobic. From 2c, we know that P(2CNX) is $\frac{54}{75}$. I want to know P(PNX|2CNX) – the probability that the parent is not xenophobic if both children are also not xenophobic. We also know that P(PNX) = $\frac{2}{3}$, and that if the parent is not xenophobic, the child will not be xenophobic either, meaning that P(2CNX|PNX) = 1.

Thus, we can use Bayes' Rule. $P(PNX|2CNX) = \frac{1 * \frac{2}{3}}{\frac{54}{75}} = \frac{150}{162}$ = about 92.6%. So there is a very high percent chance that if both children are not xenophobic, the parent is not xenophobic either.

Question 3 Background:

Now let's verify whether our results hold doing simulations. In R, create three empty vectors for parent, kid1, and kid2. Simulate 1,000 "families" where the probabilities of each individual holding the policy preference correspond to the probabilities given in Problem 2 above. After generating these 3 vectors of length 1,000, bind them together and assemble a "families" dataset using {data.frame(parent = vector1, child1 = vector2, child2 = vector3)}.

Based on my simulation:

```
parent <- c(1, 2, 3)
```

3a: What is the proportion of families in which no child holds the political attitude?

My answer will eventually be here.