Problem Set 1

Daniel Shapiro

Question 1 Background:[1]

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) = $\frac{P(A|G)*P(G)}{P(A)} = \frac{\frac{3}{20}*\frac{1}{2}}{\frac{1}{12}} = \frac{\frac{3}{40}}{\frac{1}{12}} = \frac{36}{40} = \frac{9}{10} = 90\%$.

[1]For code for this PDF, see my Github repository at https://github.com/dfshapir/ps1.

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?

First, 600/12 is 50. So we can assume that 50 men were likely born in August. Of those, we know that there is a 90% chance that the baby will be named Gus. So we can assume that we have 45 Guses. We also know that there is a 1/31 chance of being born on any given day (days are spread evenly).

We can get the probability by doing $1 - P(Aug1)^c$, where $P(Aug1)$ is defined as the probability of two or more Guses being born on the same day. The inverse of this $(P(Aug1)^c$ can be defined as the event in which there are either one or zero Guses born on that day. Thus, we get the following equation:

$\frac{30}{31}^{45} + 45(\frac{1}{31}(\frac{30}{31})^{44})$. This equals .571. 1-.571 = .428, or about a 43% chance.

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

We take a similar process here. Out of the 50 men that were born in August, we know that there is a 10% chance that the baby will be named Hugo. So we can assume that we have 5 Guses. Thus, here we can do the same process as 1d), but with 5 instead of 45:

$\frac{30}{31}^{5} + 5(\frac{1}{31}(\frac{30}{31})^{4})$. This equals about .990. 1-.990 = .001, or essentially a 1% chance.

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. Thus, if there are two Hugos there is a much higher chance that there is voter fraud.

Question 2 Background:

Does social environment affect the political development of preadults? There has been some work [2] *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.*

[2] 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.

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_1NX, C_2NX)}

Also, I know these are not possible, but there are also technically three more combinations with a probability of 0: {(PNX, C_1X, C_2X), (PNX, C_1X, C_2NX), (PNX, C_1NX, C_2X)}. I included them because it felt strange not to, and I include them in my answers for the remaining questions when applicable as well just for procedural purposes.

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

$\frac{3}{5}$. The given information indicates that if the parent is xenophobic, there is a 3/5 chance that the child is xenophobic.

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 about 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:

```
set.seed(6800)
parent <- sample(0:1, size = 1000, replace = TRUE, prob = c((2/3), (1/3)))

# I have already set my seed to 6800 in my initial chunk, but I did it again just to be on the
#safe side. I also turned off the warning because sometimes I'd get some very lengthy ones
#that would just slow things down and not change anything. I sample between 0 and 1 and put
#replace to TRUE; otherwise I'd just have two values. Size is 1000 (size of the vector) and
#the probability of being non-xenophobic (labeled as 0) is 2/3 for the parent. The
#probability of being xenophobic (labeled as 1) is 1/3.

child1 <- c(rep(1, 1000))
child2 <- c(rep(1, 1000))

# Created two empty vectors with all values of one, then bound them together
#into a data frame.

families <- data.frame(parent, child1, child2)

# Below are the conditions that the problem required. Equations for both kids are the
#same, although the results may not be.

for(i in 1:1000) {
  if(families$parent[i] == 0){
    families$child1[i] <- sample(c(0, 1), replace = TRUE, prob = c(1, 0))}
  else{families$child1[i] <- sample(c(0, 1), replace = TRUE, prob = c(.4, .6))}
  }

for(i in 1:1000) {
    if(families$parent[i] == 0){
    families$child2[i] <- sample(c(0, 1), replace = TRUE, prob = c(1, 0))}
  else{families$child2[i] <- sample(c(0, 1), replace = TRUE, prob = c(.4, .6))}
}
```

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

I read this question as asking "no child holds the political attitude *of the parent*." So to figure this one out, first, I filtered out all of the values in which families$parent == 0, because we know that if the parent is non-xenophobic, the kid will not be xenophobic either. Then, I just kept the instances in which both child1 and child2 equaled 0, which left me with 63 families. The proportion, thus, is **63/1000, or 0.063, or 6.3%.**

```
threea <- families %>%
  filter(parent != 0) %>%
  filter(child1 == 0) %>%
  filter(child2 == 0)

print(nrow(threea)/nrow(families))
```

```
## [1] 0.063
```

3b: Among families with an elder child who is not xenophobic, what proportion of parents are xenophobic?

Again, this is straightforward to do with filter(). First, I filtered out the instances where the elder child (child1) was xenophobic. I saved that as a separate dataframe because I knew I would have to use it later. Then I only kept instances where the parent was xenophobic and saved that. The final proportion came out to **145/790, or .184, or about 18.4%.**

```
threeb <- families %>%
  filter(child1 == 0)

threebctd <- threeb %>%
  filter(parent == 1)

print(nrow(threebctd)/nrow(threeb))
```

```
## [1] 0.1835443
```

3c: Among families where no child is xenophobic, what is the proportion of families where the parent is also not xenophobic?

The setup for 3c is very similar to that of 3b, as you can see below. The final proportion came out to **645/708, or .911, or about 91.1%.**

```
threec <- families %>%
  filter(child1 == 0) %>%
  filter(child2 == 0)

threecctd <- threec %>%
  filter(parent == 0)

print(nrow(threecctd)/nrow(threec))
```

```
## [1] 0.9110169
```

3d: Is whether the elder child xenophobic independent of whether the younger child is xenophobic?

Here, we have to do the same things as 2d. We do: Does P($A \bigcap B$) = P(A) * P(B)? But we use the data we got from our sampling endeavor, instead of the original stuff. 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 can use filter() to determine the number of instances of each in our data and divide them by 1000 to determine what R gives us.

Then we take P(B). 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)}. The last two are impossible given our parameters, so we can filter for only the first two potential outcomes.

As we can see below, P(A)*P(B) = .045.

```
# First, P(A).

# First on P(A): (PX, $C_1$X, $C_2$X):

threeda <- families %>%
  filter(parent == 1) %>%
  filter(child1 == 1) %>%
  filter(child1 == 1) %>%
  nrow()

# Second, (PX, $C_1$X, $C_2$NX)

threedb <- families %>%
  filter(parent == 1) %>%
  filter(child1 == 1) %>%
  filter(child1 == 0) %>%
  nrow()

pa <- ((threedb + threeda)/1000)

# This gives us 21%. The other two options have a probability of 0.

# Next, P(B).

# First on P(B): (PX, $C_1$X, $C_2$X)

threedc <- families %>%
  filter(parent == 1) %>%
  filter(child1 == 1) %>%
  filter(child2 == 1) %>%
  nrow()

# Next: (PX, $C_1$NX, $C_2$X)

threedd <- families %>%
  filter(parent == 1) %>%
  filter(child1 == 0) %>%
  filter(child2 == 1) %>%
  nrow()

pb <- ((threedc + threedd)/1000)

pa*pb
```

```
## [1] 0.04494
```

Now, we need to see if this is equal to P($A \bigcap B$). ($A \bigcap B$) = {(PX, C_1X, C_2X), (PNX, C_1X, C_2X)}. The second outcome is impossible, so P($A \bigcap B$), according to our simulation, equals .132:

```
threedint <- families %>%
  filter(parent == 1) %>%
  filter(child1 == 1) %>%
  filter(child2 == 1) %>%
```

```
  nrow()

threedint/1000
```

```
## [1] 0.132
```

.132 and .045 are not equal to one another, so **NO**, they are not independent.

3e: Conditional on the parent's attitude, is the elder child independent of the younger child?

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)}

In R:

```
threee <- families %>%
  filter(parent == 1)

nrow(threee)
```

```
## [1] 355
```

Note that now we are dividing by 355 instead of 1000.

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:

```
threeea <- threee %>%
  filter(child1 == 1) %>%
  filter(child2 == 1) %>%
  nrow()

threeeb <- threee %>%
  filter(child1 == 1) %>%
  filter(child2 == 0) %>%
  nrow()

pa <- ((threeea + threeeb)/355)
```

Next, 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:

```
threeec <- threee %>%
  filter(child1 == 1) %>%
  filter(child2 == 1) %>%
  nrow()

threeed <- threee %>%
```

```
  filter(child1 == 0) %>%
  filter(child2 == 1) %>%
  nrow()

pb <- ((threeec + threeed)/355)

pa*pb
```

```
## [1] 0.3565959
```

$(A \bigcap B)$ for this smaller sample space, meanwhile, is {(PX, C_1X, C_2X)}. P$(A \bigcap B)$ thus equals:

```
threeefinal <- threee %>%
  filter(child1 == 1) %>%
  filter(child2 == 1) %>%
  nrow()

threeefinal/355
```

```
## [1] 0.371831
```

This is difficult to tell. .3566 and .3718 are APPROXIMATELY equal, and considering this is from sampled data, and we know that arithmetically the two numbers should be the same (.36 and .36, as shown in 2E), we can say that it is *probably* (not a technical term!) equal. By this logic and this set of assumptions, we can say that **YES, they are probably independent** given these conditions.

How well do these simulation results match with the analytic results in Problem 2 above?

For the most part, they are very similar – not exactly the same, but very similar. It makes sense that they are not exactly the same; after all, this is meant to be a random sample.