Real World Mini-Project 1: Bayes' Rule

The tools of this class are useful to computer scientists, but many of them are useful beyond just "classic" computer science. In this assignment you will consider an application of Bayes' Rule in the real-world and then try to model a new scenario (different from the one we provide) with the Bayes' Rule.

This assignment is a mix of technical tasks (appropriately applying theorems) and non-technical ones (considering trade-offs between various real-world effects and groups). The technical aspects can be "right" or "wrong", but the non-technical aspects are unlikely to be simply "right" or "wrong" — we won't have to **agree** with the non-technical aspects of your analysis to consider them a good analysis. Our evaluation will be based on how well they connect to the technical aspects, as well as the depth of reasoning demonstrated.

Submission: You must upload a **pdf** of your written solutions to Gradescope under "Real World Mini-Project 1". The use of LaTeX is *highly recommended*. (Note that if you want to hand-write your solutions, you'll need to scan them. We will take off points for hand-written solutions that are difficult to read due to poor handwriting and neatness.)

Please put each numbered problem on its own page of the pdf (this will make selecting pages easier when you submit), and ensure that your pdfs are oriented correctly (e.g. not upside-down or sideways).

Remember that you must tag your written problems on Gradescope.

Due Date: This assignment is due on Wednesday, July 21 at 11:59 PM (Seattle time, i.e. GMT-7).

Collaboration Policy

You are to conduct your own search and analysis for this assignment. While you may get feedback from other students on your writing, you cannot just use the results of another student's search.

1. Understanding Vaccination

Given the numerous discussions around vaccines and clinical trials that we have had this year, we would like you to get a deeper understanding of what some of these terms mean. We will look at a simplified version of vaccine trial information and try to understand what vaccine efficacy really means. There is much more that can be extrapolated from vaccine trial data, but we will be limiting ourselves to a small subset.

Table 1 is a simplified version of Table 3 from the paper that discusses the efficacy of BNT162b2, the Pfizer-made vaccine for COVID-19:

Efficacy Endpoint Subgroup	Number of COVID Cases in Vaccine Group (N=17,397)	Number of COVID Cases in Placebo Group (N=17,498)	Vaccine Efficacy, %
Overall	8	162	95.0
Age Group			
16-55 years	5 (N=9897)	114 (N=9955)	95.6
>55 years	3 (N=7500)	48 (N=7543)	93.7
>65 years	1 (N=3848)	19 (N=3880)	94.7
>75 years	0 (N=774)	5 (N=785)	100.0

Table 1: Trial data for the Pfizer made vaccine for COVID-19. Source: Link

1.1. Starting with the basics

Answer the questions using table 1. Let the event V denote that a random participant was in the vaccine group, and the event S denote that the participant got sick.

(a) What's the probability that a random participant gets sick, regardless of what group they were in? What about for each participant group (vaccine and placebo)? Clearly state the symbols you use for each event and the conditional and unconditional probability.

(b) Given that a participant got sick, what is the probability that they were in the vaccine group? This should be a direct Bayes' Rule calculation.

(c) Is this enough data to convince you to get the vaccine? Is there anything important that this table is missing?

1.2. Thinking about age

(a) If you don't read the table carefully, you might come to a strange (and incorrect) conclusion: you are more likely to get sick if you are younger. From the table, the probability of getting sick if you were between 16 and 55 is $\frac{(5+114)}{(9,897+9,955)} \approx 0.006$, while the probability of getting sick if you in the other age groups is $\frac{(3+1+0+48+19+15)}{(7,500+3,848+774+7,543+3,880+785)} \approx 0.003$. What is the issue with this calculation?

(b) If you pick three distinct participants, what is the probability that they all fall in the same age range? Here we define distinct age ranges to be — 16-55, 55-65, 65-75, and >75.

(c) What is the probability that at least two of three randomly chosen participants got the vaccine?

(d) What is the probability that at least two of three randomly chosen participants got the vaccine AND that they all fell into the same age group?

1.3. So what is "Efficacy" anyway?

(a) "Vaccine Efficacy %" is defined as $100 * \left(1 - \frac{\mathbb{P}(S \mid V)}{\mathbb{P}(S \mid V^C)}\right)$, where S is the event that a random patient gets sick, and V is the event that a random patient is in the vaccine group. Unfortunately, most people confuse efficacy this with $1 - \mathbb{P}(S \mid V)$. What is the difference between these two equations?

(b) Why do you think we use vaccine efficacy % rather than $1 - \mathbb{P}(S \mid V)$ to measure how well a vaccine is protecting against a disease? Can you come up with a scenario that illustrates why the former metric is more useful than the latter? If not, in your own words explain why efficacy is calculated as $100 * \left(1 - \frac{\mathbb{P}(S \mid V)}{\mathbb{P}(S \mid V^C)}\right)$ and not $1 - \mathbb{P}(S \mid V)$.

1.4. Debunking Myths

Table 2 is from the same paper about the Pfizer vaccine. It concerns the occurrence of "adverse events" — e.g., negative reactions after a vaccine. Note that the number of people in the trial here is higher because some people in this table dropped out mid-way through the trials.

Event Type	Vaccine Group (N=21,621)	Placebo (N=21,631)
Any (including mild reactions, e.g. fevers or muscle soreness)	5770	2638
Severe	240	139
Life-threatening	21	24

Table 2: Vaccine side effects data from clinical trials. Note: The data is cumulative — "Life-threatening" is included in the "Severe" and "Any" categories and "Severe" is included in the "Any" category.

(a) Some people have argued that taking a vaccine is not worth it because the chance of developing a severe adverse event after a vaccine $\frac{240}{21,621}$ is larger than the chance of actually getting COVID-19 without a vaccine $\frac{162}{17,498}$. Using the symbols, V and S above, as well as A_s to denote the event of a participant getting a severe adverse event, what two conditional probabilities are we comparing here? State both the conditional probabilities as $\mathbb{P}(A \mid B)$ clearing stating the events A and B. Hint: you can use events you have already defined in the previous parts.

(b) What is a logical flaw in the argument above? How would you argue that getting a vaccine is still worth it? Note: there are several answers to this question! Be creative!

2. Make Another Argument

In this portion of the assignment, we want you to do some data gathering on your own in order to further see how Bayes' Rule and the probability techniques you have learned can be applied to real world scenarios.

For example, you could look at sports statistics, analyze political forecasts, or even consider what the average American thinks about "Star Wars". A great place for data (and has data about all the above topics) is https://data.fivethirtyeight.com/, although this is by no means the only source of data you may consider. [1]

If you are stuck, you could consider keeping with the theme of the assignment and doing an analysis on COVID-19 statistics: you could answer questions about mask or lockdown efficacy, analyze case rates across countries, compare the data above with the data from a different vaccine, or look at the accuracy rate of different COVID tests.

When you are looking at a scenario, please bear in mind that you will need to use Bayes' Rule for this part of the assignment.

You are allowed (and encouraged) to do your own research toward this question.

(a) Write up the data you will be using in a table, and make sure to cite your sources.

(b) Define events A and B on which you'll apply Bayes' Rule (along with any other events you need).

(c) State probabilities (or probability estimates) for three of the four quantities you need to use Bayes rule. For those estimates, either cite a source for the numbers that you think is reliable or give a justification for your estimate. Then using Bayes' Rule find the missing probability.

(d) What is your takeaway from this calculation? How did you increase your understanding of the problem?

(e) Discuss at least one way you would like to extend your analysis, but cannot with the limited data, and/or just the help of discrete probability, counting and independence. You may (but are not required to) have a look at the course schedule to think of topics that may help you in further analysis.

3. Sample Solution

Since we haven't asked you to do tasks exactly like these before, here is a sample of the kind of answers we'd be expecting for an application. When doing research, scientists often use statistical significance testing. In that framework one writes a hypothesis ("smoking causes cancer"), and then asks for the p-value: the probability of seeing the data in the study, if the hypothesis is false. $p = 0.05$ (or less) is usually taken as statistically significant.

[1] We will be quite lenient about what counts as real world — the hope is that you will pick something you care about. If it's just the probability that the second and third card of a deck of cards is the same value, it's probably not "real-world." But if you're an avid poker player, and you want to use Bayes' Rule to analyze a particular game scenario, that would definitely count.

(a) A data table.

(b) Let H be the event "the hypothesis is true" and D be the event "we saw data like this in an experiment"

(c) We'll analyze a statistically significant experiment, so $\mathbb{P}\left(D \mid \overline{H}\right) = 0.05$. We'll consider an experiment where the result would be surprising — one where before running the experiment, $\mathbb{P}(H) = 0.2$. Furthermore, we'll suppose the data show only a weak effect, so $\mathbb{P}(D \mid H) = 0.5$.

Applying Bayes Rule:

$$\mathbb{P}(H \mid D) = \frac{\mathbb{P}(D \mid H) \cdot \mathbb{P}(H)}{\mathbb{P}(D)} = \frac{0.5 \cdot 0.2}{0.5 \cdot 0.2 + 0.05 \cdot 0.8} \approx 0.714$$

(d) The chances of the hypothesis in the paper being true are pretty good — but not nearly the 95% one might imagine if you misinterpret the meaning of the p-value.

(e) Estimating the probability of a hypothesis being true without experimenting would be quite difficult in the real-world. With a lower starting value, the probability of accuracy would drop; one should perhaps be careful when reading papers (even ones with statistically significant results). Particularly when they have a surprising claim.