HSCI 50 LAB 2: Basics of Probability ANSWER KEY

INSTRUCTIONS

The data set that we will be using for the entire course is resampled data from the MIT COVID-19 Beliefs, Behaviors & Norms Survey (https://covidsurvey.mit.edu/api.html). This is a multi-country, online survey that examined different COVID-19 perceptions across time, from July 6, 2020 to March 28, 2021. We will be using data from the Philippines aged 20 to 60.

In this lab, you will practice calculating probabilities given aggregate statistics from this data set.

You have many options to submit this worksheet. Either you work on this by hand and scan/take a clear photo of your submission and save as PDF, or type your responses in a Word processor or PowerPoint presentation. You do not have to copy the questions again, but please number them accordingly.

Part A. The following table summarizes the responses of vaccine acceptance by age group in the Philippines during the survey's final wave, conducted last March 14 to 28, 2021. Answer the questions that follow. Express your answers as fractions. You may use a calculator.

Table 1: COVID-19 vaccine acceptance across age groups in the Philippines, March 14-18, 2021

Accept COVID-19 vaccine	Aged 20-30	Aged 31-40	Aged 41-50	Aged 51-60
Yes	179	166	123	110
No	83	63	39	26
Don't Know	157	89	44	32
Vaccinated	18	5	10	1

1. The total sample size for this survey is 1,145 respondents. If a person is chosen at random from this pool of respondents, what is the probability that:

a. This person has already been vaccinated
b. This person has not been vaccinated
c. This person is not 51-60 years old
d. This person does not accept the COVID-19 vaccine **or** is 20-30 years old
e. This person accepts the COVID-19 vaccine **or** has already been vaccinated
f. This person accepts the COVID-19 vaccine **and** has already been vaccinated
g. This person accepts the COVID-19 vaccine **and** is 31-40 years old
h. This person accepts the COVID-19 vaccine **given** they are 31-40 years old
i. This person does not accept the COVID-19 vaccine or is unsure about it **given** they are 41-50 years old
j. This person accepts the COVID-19 vaccine or has already been vaccinated **given** they are 40 years old and below.

The answers are as follows:

a. This person has already been vaccinated

$$\frac{\text{Vaccinated}}{\text{Total}} = \frac{18+5+10+1}{1145} = \frac{34}{1145}$$

b. This person has not been vaccinated

This is simply the complement of (a), or you can add the other three categories (more work, though)

$$\frac{\text{Not been vaccinated}}{\text{Total}} = 1 - \frac{18+5+10+1}{1145} = 1 - \frac{34}{1145} = \frac{1111}{1145}$$

c. This person is not 51-60 years old

This is the complement of the probability those who are not 51-60 years old

$$\frac{\text{Not } 51-60 \text{ years old}}{\text{Total}} = 1 - \frac{110+26+32+1}{1145} = 1 - \frac{169}{1145} = \frac{976}{1145}$$

d. This person does not accept the COVID-19 vaccine **or** is 20-30 years old

Remember the rejoinder OR means that we count both categories and subtract it with whatever overlaps between them (what is indicated by the rejoinder AND) so that we do not count the overlaps twice. In this case, we add all those who do not accept the COVID-19 vaccine, then add those 20-30 years old, and make sure that we count those who do not accept the COVID-19 vaccine AND 20-30 years old only once. $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

$$\frac{\text{Do not accept vaccine OR } 20-30 \text{ years old}}{\text{Total}} = \frac{(83+63+39+26)+(179+83+157+18)-83}{1145} = \frac{565}{1145}$$

e. This person accepts the COVID-19 vaccine **or** has already been vaccinated

In this case, the rejoinder OR is counting two disjoint (mutually exclusive) events, so there are no overlaps, i.e. $P(A \text{ and } B) = 0$

$$\frac{\text{Accepts vaccine OR Vaccinated}}{\text{Total}} = \frac{(179+166+123+110)+(18+5+10+1)}{1145} = \frac{612}{1145}$$

f. This person accepts the COVID-19 vaccine **and** has already been vaccinated

In this case, the rejoinder AND is counting two disjoint (mutually exclusive) events, as we have established in (e). Therefore, the answer is 0.

g. This person accepts the COVID-19 vaccine **and** is 31-40 years old

$$\frac{\text{Accepts vaccine AND } 31-40 \text{ years old}}{\text{Total}} = \frac{166}{1145}$$

h. This person accepts the COVID-19 vaccine **given** they are 31-40 years old

Now we are measuring conditional probabilities (as indicated by the rejoinder GIVEN). Our denominator is not the total sample anymore, but whatever is the conditional statement.

$$\frac{\text{Accepts vaccine among } 31-40 \text{ years old}}{31-40 \text{ years old}} = \frac{166}{166+63+89+5} = \frac{166}{323}$$

i. This person does not accept the COVID-19 vaccine or is unsure about it **given** they are 41-50 years old

Take note that the numerator has the rejoinder OR, but these are disjoint events so there are no overlaps.

$$\frac{\text{Does not vaccine OR Unsure among } 41-50 \text{ years old}}{41-50 \text{ years old}} = \frac{39+44}{123+39+44+10} = \frac{83}{216}$$

j. This person accepts the COVID-19 vaccine or has already been vaccinated **given** they are 40 years old and below

Take note that the denominator covers those who are below 40 years old. In other words, 20-30 years old OR 31-40 years old

$$\frac{\text{Accepts vaccine OR Vaccinated among 40 years old and below}}{\text{40 years old and below}} = \frac{(179+18)+(166+5)}{(179+83+157+18)+(166+63+89+5)} = \frac{368}{760}$$

2. Let's define vaccine hesitancy as not accepting or being unsure about the COVID-19 vaccine. Based on this definition, does our data show that vaccine hesitancy and age are independent? Why or why not? If it is not independent, what direction is the relationship trending towards? Hint: calculate conditional probabilities of vaccine hesitancy per age group and express your answers as percentages.

First, we check the conditional probabilities of vaccine hesitancy across age groups. The basic formula is

$$\frac{\text{Vaccine hesitancy among age group}}{\text{Total age group}}$$

The conditional probabilities of each age group are as follows:

- 20-30 years old: $\frac{83+157}{179+83+157+18} = 0.55$
- 31-40 years old: $\frac{63+89}{166+63+89+5} = 0.47$
- 41-50 years old: $\frac{39+44}{123+39+44+10} = 0.38$
- 51-60 years old: $\frac{26+32}{110+26+32+1} = 0.34$

Statistical independence is established when the occurence of one event does not affect the occurence of another. In our example, we would see statistical independence if the conditional probability of vaccine hesitancy for each age group were not different from one another. We can clearly see that the probabilities vary widely across age groups. Therefore, the relationship is not independent. The trend is that younger age groups are more vaccine hesitant than older age groups. From a public health perspective, we should be concerned about vaccine hesitancy across all age groups, as around 2 in 5 are vaccine hesitant, but more so among younger people.

PART B. Read the following problem below and construct a probability tree, a 2x2 table, or use Bayes' Theorem to answer the question.

Now consider the Philippine government's testing policy for COVID-19. The government said that in the context of a widespread outbreak and the declaration of a strict lockdown, a rapid antigen test validated against World Health Organization (WHO) standards may be used as a confirmatory COVID-19 test instead of the gold standard reverse transcriptase - polymerase chain reaction (RT-PCR) test.

At the minimum, a validated rapid antigen test has 80% sensitivity, or the probability of receiving a positive test result given having the disease, and 97% specificity, or the probability of receiving a negative test result given not having the disease. Assume that the true prevalence of COVID-19 in a widespread outbreak is 30%. What is the positive predictive value, or the probability of having the disease given a positive test result? Round off your answer to the nearest percent.

The question has three givens:

$$Pr(\text{Antigen } + \mid \text{COVID}) = 0.80$$
$$Pr(\text{Antigen } - \mid \text{NoCOVID}) = 0.97$$
$$Pr(\text{COVID}) = 0.30$$

Method 1: 2x2 tables

Assume 1000 people in the entire sample. With a prevalence of 30%, 300 people have COVID, 700 people don't.

	COVID	No COVID	Total
Antigen +			
Antigen -			
Total	300	700	1000

Given that the sensitivity is 80%, the number of true positives is 0.80 * 300 = 240 and false negatives is 0.20 * 300 = 60

	COVID	No COVID	Total
Antigen +	240		
Antigen -	60		
Total	300	700	1000

Given that the specificity is 97%, the number of true negatives is 0.97 * 700 = 679 and false positives is 0.03 * 700 = 21. We also fill in the rest of the totals

	COVID	No COVID	Total
Antigen +	240	21	251
Antigen -	60	679	739
Total	300	700	1000

From the table, we can calculate the positive predictive value as $\frac{240}{251} = 92\%$

Method 2: Probability Tree

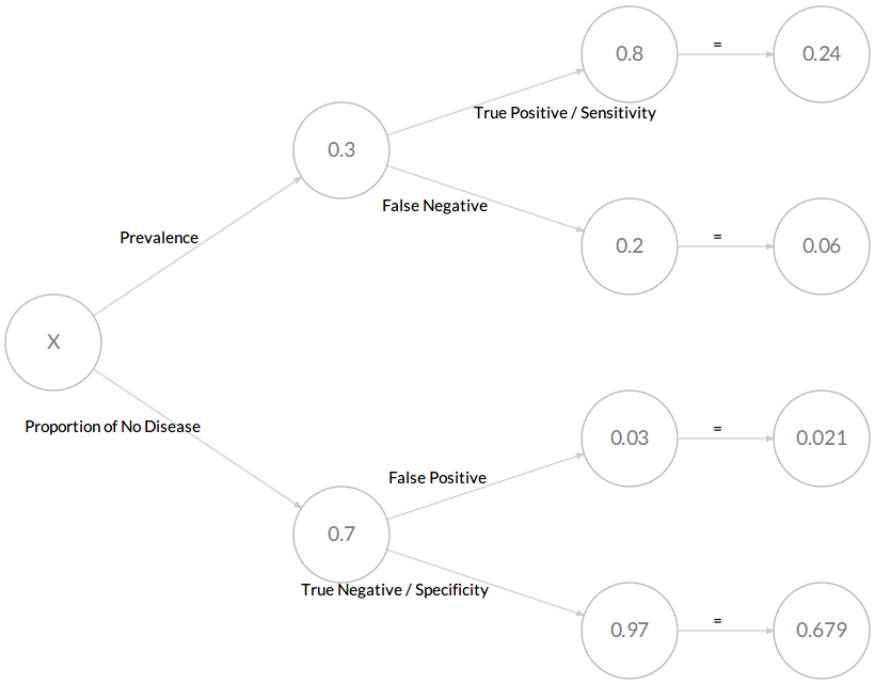

Figure 1: Probability Tree for 30% COVID-19 Prevalence.

Based on the figure above, the positive predictive value is $\frac{0.24}{0.24+0.021} = 92\%$

Method 3: Bayes' Theorem

$$
\begin{aligned}
Pr(\text{COVID} \mid \text{Antigen} +) &= \frac{Pr(\text{Antigen} + \mid \text{COVID}) \times Pr(\text{COVID})}{Pr(\text{Antigen} + \mid \text{COVID}) \times Pr(\text{COVID}) + Pr(\text{Antigen} + \mid \text{No COVID}) \times Pr(\text{No COVID})} \\
&= \frac{0.8 \times 0.3}{(0.8 \times 0.3) + (0.03 \times 0.7)} \\
&= 92\%
\end{aligned}
$$

Public health implication

The government seemed to have made a fairly reasonable policy decision in allowing the use of rapid antigen testing kits in areas with widespread community infection. A positive predictive value of 92% seems acceptable given that rapid antigen test results can give results as quickly as 15 minutes as compared to the gold standard RT-PCR test which may take hours to days. The government may not admit that 30% of the population at any given time (that's almost 1 out of 3!) has COVID-19, that is a reasonable assumption to

make given data from other countries through studies that we call seroprevalence studies. Seroprevalence studies measure antibody levels in a random sample of a population, and many countries like India and Spain do these seroprevalence studies regularly for the entire population to get a sense of how prevalent COVID-19 was around two to three months prior to the conduct of the seroprevalence study. There were estimates in India last year that seroprevalence reached as high as 70%, which made the government there think that the country had reached natural herd immunity. We now know of course that the new variants of concern, such as the Delta variant, has the capacity to reinfect previously infected persons.

END OF LAB