Homework # 3

Elliott Pryor

9/11/2020

Problem 1

Statement

Abott Architect SARS-CoV-2 IgG is a serology test for the antibodies or SARS-CoV2 (the virus that causes COVID-19) that is manufactured by Abbott pharmaceuticals. If a person has antibodies for the virus, it can indicate that they will not fall ill from the same virus again (although the research related to immunity after exposure to COVID-19 is still developing). Abbott claims that their test does at least this well with respect detecting antibodies when they are there: 95% of the time the test will detect antibodies when they are there (i.e., the true-positive rate or sensitivity of the test is at the very least 0.95), and detects antibodies in a person that does not have them 1% of the time (i.e., the false-positive rate of the test is at most 0.01). See the FDA fact sheet here (click this text it's a hyperlink). It is impossible to know at this time the rate of prevalence of COVID-19 in any given community (the data to estimate this just don't exist yet), so we'll consider a few different scenarios. Let's focus on the MSU community, and let's assume that within the MSU community 1 in 100 people have been exposed to COVID-19 (i.e., they have recovered from the illness)

1. If a randomly selected MSU community member tests positive for the antibodies (using Abbott's test) what is the probability that the person had actually suffered from COVID-19 and recovered (i.e., truly has the antibodies to SARS-Cov-2)?
2. If the rate of exposure within the MSU community was instead 10%, what is the probability that a randomly selected person had actually suffered from COVID-19 and recovered?
3. Briefly discuss how the results in parts (a) and (b). Use your results to discuss why it is imperative (from a public health standpoint) that we have a good estimate of the true rate of exposure (i.e., those who have been ill and then recovered) within the MSU community. One idea would be to discuss this with respect to the "trustworthiness" of the test, but that is not the only way to address this question!

Solution

1. We know that the person tested positive for the antibodies. So we are looking for P(covid | antibodies). Let C be the event that someone has covid, and A be the event that they test positive for antibodies. Clearly $C, \bar{C}$ form a partition of the sample space. So using Bayes rule:

$$P(C|A) = \frac{P(A|C)P(C)}{P(A|C)P(C) + P(A|\bar{C})P(\bar{C})} = \frac{0.95 * 0.01}{0.95 * 0.01 + 0.01 * 0.99} = 0.4897 = 48.97\%$$

This means that with a positive test result, there is only a 48.97% chance that they actually have had covid.

2. We use the same equation as above, but with a higher rate of exposure.

$$P(C|A) = \frac{P(A|C)P(C)}{P(A|C)P(C) + P(A|\bar{C})P(\bar{C})} = \frac{0.95 * 0.10}{0.95 * 0.10 + 0.01 * 0.90} = 0.9135 = 91.35\%$$

The higher exposure rate significantly changed the probability that they actually had covid.

3. Clearly the probability is vastly different depending on the exposure rate in 1. and 2. So estimating the true exposure rate is vital for determining the reliability of the test. At low exposure rates, the relatively high false positive rate and the very high number of healthy individuals drives the equation. The $P(A|\bar{C})P(\bar{C})$ term is much larger than the $P(A|C)P(C)$ term. However as the exposure rate increases, this term decreases rapidly, and the expression relies on the high true positive rate as the driving factor. Figure 1 shows that the probability that an individual has had COVID (reliability of the test) varies sharply with the exposure rate. So a small change in exposure rate can drastically change the probability that a positive individual has had COVID.

```
curve((0.95 * x / (0.95 * x + 0.01 * (1-x))), 0, 1, ylab = "Probability individual has had COVID",
      xlab = "Exposure rate")
```

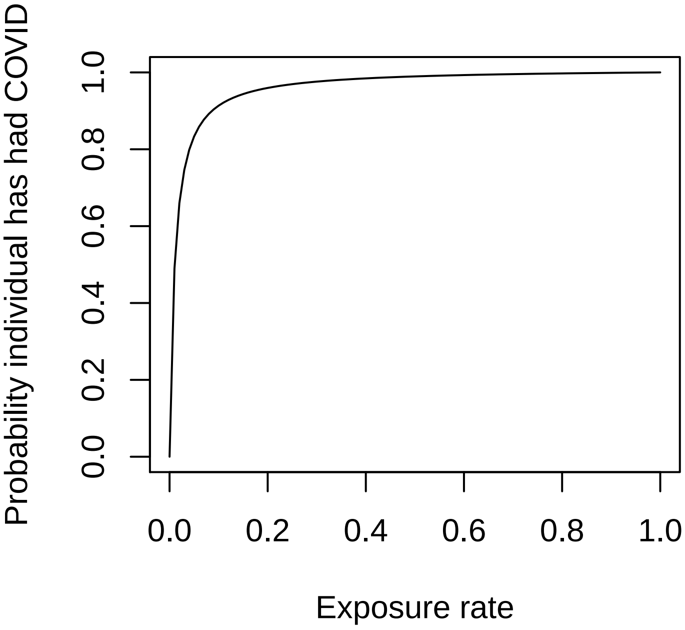

Figure 1: Probability that a person is positive for COVID-19 with varying exposure rates

Problem 2

Statement

An appliance store receives a shipment of 30 microwave ovens, 5 of which are (unknown to the manager) defective. The store manager selects 4 ovens at random, without replacement, and tests to see if they are defective. Let $X=$ number of defectives found

1. Find the probability distribution of X and plot it
2. Find the cumulative distribution of X and plot it
3. Find $E(X)$ and $Var(X)$

Solution

1. We first list the support of X: $S_X = \{0, 1, 2, 3, 4\}$. We first compute the probabilities:

$P(X = 0) = \frac{\binom{25}{4}}{\binom{30}{4}} = 0.4616$ - The probability that no microwaves are defective.

$P(X = 1) = \frac{\binom{5}{1}\binom{26}{3}}{\binom{30}{4}} = 0.4196$ - The probability that exactly one microwave is defective.

$P(X = 2) = \frac{\binom{5}{2}\binom{26}{2}}{\binom{30}{4}} = 0.1095$ - The probability that exactly two microwaves are defective.

$P(X = 3) = \frac{\binom{5}{3}\binom{26}{1}}{\binom{30}{4}} = 0.0091$ - The probability that exactly three microwaves are defective.

$P(X = 4) = \frac{\binom{5}{4}}{\binom{30}{4}} = 0.0002$ - The probability that all four microwaves are defective.

then $P(X = \text{other}) = 0$

So

$$f_X = \begin{cases} 0.4616 & x = 0 \\ 0.4196 & x = 1 \\ 0.1095 & x = 2 \\ 0.0091 & x = 3 \\ 0.0002 & x = 4 \\ 0 & \text{otherwise} \end{cases}$$

Figure 2 shows a plot of the probability distribution of X. The probability spikes are at single point values. This makes physical sense as it isn't possible to have part of a microwave broken, so all the probabilities are discrete and occur at integer values.

```r
P <- function(x){
  if(x == 0){return(0.4616)}
  if(x == 1){return(0.4196)}
  if(x == 2){return(0.1095)}
  if(x == 3){return(0.0091)}
  if(x == 4){return(0.002)}

  return(0)
}
vP <- Vectorize(P)
curve(vP, from = -1, to = 5, xlab = "Number of defective microwaves", ylab = "Probability", n= 6001)
```

2.

$$F_X = \begin{cases} 0 & x < 0 \\ 0.4616 & 0 \leq x < 1 \\ 0.8812 & 1 \leq x < 2 \\ 0.9907 & 2 \leq z < 3 \\ 0.9998 & 3 \leq x < 4 \\ 1 & 4 \leq x \end{cases}$$

Figure 3 shows a plot of the cumulative distribution of X. This is the probability of $P(X <= x)$, or the probability that fewer than x microwaves are broken. As this goes to infinity this probability is 1. This is good, otherwise there would be some probability that infinite microwaves are broken and that is not physically possible as we must have 0, 1, 2, 3, or 4 defective microwaves in the sample.

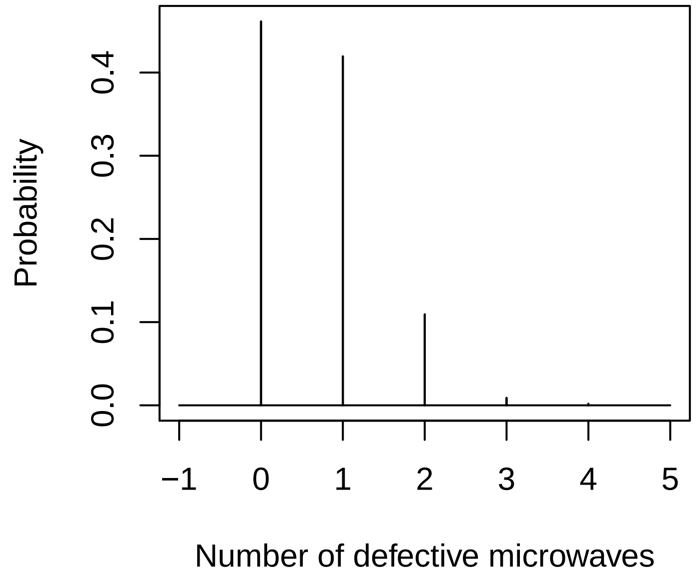

Figure 2: Probability Distribution of X

```r
P <- function(x){
  if(x < 0){return(0)}
  if(0 <= x && x < 1){return(0.4616)}
  if(1 <= x && x < 2){return(0.8812)}
  if(2 <= x && x < 3){return(0.9907)}
  if(3 <= x && x < 4){return(0.9998)}
  if(x >= 4){return(1)}

  return(0)
}
vP <- Vectorize(P)
curve(vP, from = -1, to = 5, xlab = "Number of defective microwaves",
      ylab = "Cumulative Probability", n= 6001)
```

3. By definition, $E(X) = \sum_{x\in S_X} x * P(X = x) = \sum_{x\in S_X} x * f_X(x)$. So we have $E(X) = 0 * 0.4616 + 1 * 0.4196 + 2 * 0.1095 + 3 * 0.0091 + 4 * 0.0002 = 0.6667$. So we expect 0.6667 microwaves to be broken, or in other words, the average number of broken microwaves is 0.6667

We can then compute the variance of X. $V(x) = E(X^2) - \bar{x}^2$ where $\bar{x}$ is the mean. $\bar{x} = E(x)$ so $V(x) = E(X^2) - (E(X))^2$. We can compute $E(X^2) = \sum_{x\in S_X} y^2 f_Y(y) = 0 * 0.4616 + 1 * 0.4196 + 4 * 0.1095 + 9 * 0.0091 + 16 * 0.0002 = 0.9427$. So the variance is $V(X) = 0.9427 - 0.6667 = 0.276$. This value is fairly small, so we do not expect much deviation from the mean. Ie. most of the samples of microwaves will be close to the mean value.

This means that most of the time only 0 or 1 microwave will be defective. Which is probably not good for the store as there is a high likeli-hood that defective microwaves will go untested.

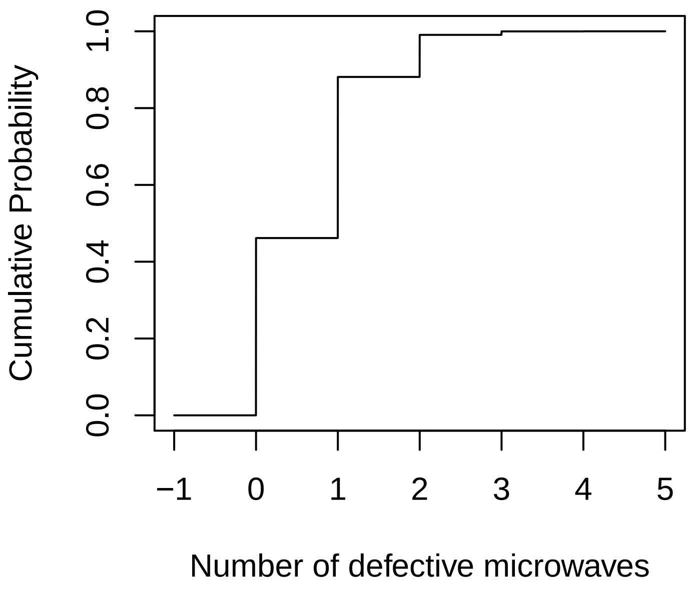

Figure 3: Cumulative Distribution of X

Problem 3

Statement

Next week (week of 9/13/20) we will be discussing commonly used discrete distributions and their applications. One such distribution that we have already used is the Bernoulli(p) distribution. In fact, we could have used the Bernoulli distribution to quickly simulate trials of 100 coin flips instead of using the sample function. We can also use random draws from the Bernoulli distribution to generate realizations from the Binomial(n,p), Geometric(p), Negative Binomial(r,p), and even the Poisson(λ) distributions. Let X be a Bernoulli(p) random variable with probability of success p, and W be a RV representing the waiting time until the first head occurs (i.e., the number of tosses until the first head occurs.) Work through the example code BEFORE attempting the following.

1. Write a function that generates 1 realization of W; assume p= 0.1. This will require some modifications to rg_fun but the function will be very similar! Then, use the replicate function to generate 1000 realizations of W
2. Display your results from part (a) in a table and a plot. Comment on the“reasonableness” of your results. Did anything surprise you?

Solution

1. Below is the modified rg_fun. We only modified the probability in the binomial function, as well as the index of the which() function to select the first occurence.

```
rg_fun <-function() {# generate a sequence of 100 independent Bernoulli trials
  samp <-rbinom(n = 100, size = 1, prob = 0.5)# find the location of the 2nd 1 - corresponds to the num
  # the second success
  g <-which(samp==1)[1]
  return(g)}

g_vec <-replicate(n = 1000, expr =rg_fun())
```

2. Now we show the results from (1.). It was strange seeing how much smaller the maximum number of observed runs there were when compared to the example code where probability of heads was 0.2. This difference makes sense. The results are fairly expected. For the first flip we have almost 1/2 the flips being heads, then 1/4, then 1/8, etc. This is roughly what we would expect of a fair coin.

```
table(g_vec)
```

```
## g_vec
##   1   2   3   4   5   6   7   8   9  10
## 509 261 119  49  31  15   7   3   2   4
```

```
tab_g  <-tabulate(g_vec,max(g_vec))

plot(tab_g, type = "h",xlab = "w",ylab = "Frequency (W = w)")# add points at the observed frequencies
points(1:max(g_vec), tab_g, pch = 20)
```

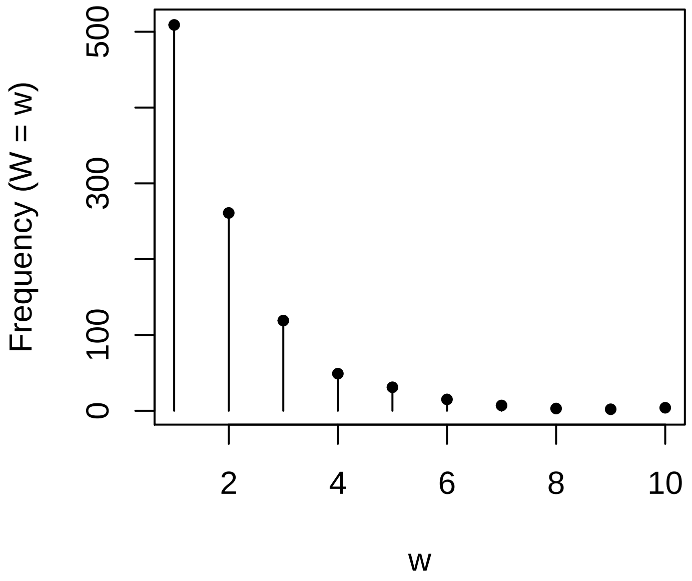

Figure 4: Plot of Results