Lab 1 Question 3: Are survey respondents who have had someone in their home infected by COVID-19 more likely to disapprove of the way their governor is handling the pandemic?

1. Introduction

As Americans went to the polls on November 3rd 2020, almost all Americans had been affected in some form or another by the COVID-19 pandemic, but relatively few Americans had actually experienced a COVID-19 infection. In early Nov 2020, the US had recorded 10 Million confirmed COVID cases (approximately 3% of the population). Do these Americans and their loved ones feel differently about their elected officials than Americans who did not personally experience a COVID-19 infection?

In this analysis we will examine the above question, and specifically test whether survey respondents who had someone in their household infected by COVID-19 are more likely to disapprove of the way their governor is handling the pandemic when compared with respondents that did not have someone in their household infected by COVID-19. This is a critical question because respondents who had someone in their home infected by COVID-19 are, simply put, the people who were most affected by their governor's handling of the pandemic. This analysis seeks to give voice to this minority group of Americans by determining whether they responded to the survey differently.

2. Data

We begin by pulling our samples from the dataset. For our analysis we will look at responses to two questions:

1. V201624: Has anyone in your household tested positive for the coronavirus disease, COVID-19, or has no one tested positive? (1. Someone in my household tested positive; 2. No one tested positive; -5. Interview breakoff (sufficient partial IW); -9. Refused).

2. V201145: Do you approve or disapprove of the way [Governor of respondent's preloaded state] has handled the COVID-19 pandemic? (1. Approve; 2. Disapprove; -8. Don't know; -9. Refused).

For our samples we remove any respondents that failed to answer either one of the above with a 1 or 2. Per instruction from our professor, we do not attempt to apply the weights in the dataset, and thus will treat simply treat the samples as if they are independently and identically distributed random samples from the two populations in question. This is not necessarily a good assumption, and we will discuss it further in section 4 on the significance of our results.

We find that in the sample of Americans with a confirmed COVID case in their household, 44% disapproved of their governor's handling the pandemic, and in the sample of Americans with no confirmed COVID case in their household, 38% disapproved of their governor's handling the pandemic.

Table 1: Households With a Confirmed COVID Case

COVID Case in Household?	Approve of Governor?	Count	Percent
Yes	Approve	159	0.56
Yes	Disapprove	125	0.44

Table 2: Households Without a Confirmed COVID Case

COVID Case in Household?	Approve of Governor?	Count	Percent
No	Approve	4876	0.62
No	Disapprove	2978	0.38

3. Test

We want to test the difference between two population proportions. Our two populations are: 1) American's with a confirmed COVID case in their household; and 2) Americans without a confirmed COVID case in their household. We want to compare the proportion that disapproves of their Governor's handling of COVID-19 in these two populations.

To do this we can calculate a Z statistic comparing the difference in the sample proportions against the null hypothesis that there is no difference in the population proportions. The formula for running such a test is:

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{pq(\frac{1}{m} + \frac{1}{n})}}$$

Where $\hat{p}_1 = \frac{X}{m}$ with $X =$ the number of respondents that disapprove over $m=$ the number of respondents in the sample, $\hat{p}_2 = \frac{Y}{n}$ with $Y =$ the number of respondents that disapprove over $n=$ the number of respondents in the sample, $p = \frac{(X+Y)}{m+n}$, and $q = 1 - p$.

We are able to use a Z statistic and the above formula because of the central limit theorem and the fact that the two populations in question follow a know distribution (the binomial distribution), this follows ipso facto from the from the nature of the question at hand (it is a yes/no question, approve or disapprove).

To reject the null hypothesis, we will need a p-value in either tail that exceeds 0.025.

The assumptions underlying this test are:

1. The samples drawn from the two populations are random samples that are independently and identically distributed.

2. The sample size is large enough that it closely approximates a normal distribution.

3. The two populations each follow a known distribution (the binomial distribution).

```
z_statistic <- function(X,Y,m,n) {
  p1 = X/m
  p2 = Y/n
  p = (X+Y)/(m+n)
  q = 1-p
  ((p1)-(p2))/sqrt((p*q)*((1/n)+(1/m)))
}

z_statistic(X=125,Y=2978,m=284,n=7854)
```

```
## [1] 2.078242
```

```
1-pnorm(2.078242,0,1)
```

```
## [1] 0.01884354
```

We find that our p value is less than .025 and we can reject the null hypothesis.

4. Significance

Although we have rejected the null hypothesis, the significance of the result is questionable for two reasons: 1) our sampling method is highly vulnerable to bias; 2) although the 6% difference in proportion between the two samples is statistically significant, even if the two populations exactly mirrored this difference, the practical significance is relatively low. We find that in either case the majority of the two populations approve of their Governor's handling of COVID-19.

Bias occurs when the expected value of our estimator is not equal to the true value in the population $E[\hat{\theta}] \neq \theta$. This commonly occurs when one is sampling from a group that does not truly represent the population, therefore increasing the number of samples drawn from that group simply brings one closer to the true value for a different population. In our analysis there is a real danger of this. The ANES dataset takes steps to ensure its respondents represent a random sample of the population of all Americans, but we have not applied the weights and adjustments that ANES advises are necessary to achieve this. Furthermore, when we divide the ANES sample into a sample of 284 respondents that had a confirmed COVID-19 case in their household, and a sample of 7,854 respondents that did not, our samples may no longer reflect true random draws from their populations. To illustrate this problem further with a simple example: our sample of 284 respondents that had a confirmed COVID-19 case in their household could have overdrawn from 1 state with a particularly bad governor, and thus not reflect a random sample from the population of Americans overall that had a confirmed COVID-19 case in their household.