Project Part 5

Nhi Nguyen

12/3/2021

```
WhiteL = data.frame(data$UnemploymentRate[data$WhitePopulation<=65])
WhiteH = data.frame(data$UnemploymentRate[data$WhitePopulation>=65])
colnames(WhiteL) = c("LowWhitePop")
colnames(WhiteH) = c("HighWhitePop")
```

Research Question

Are the unemployment rates of each county in California influenced by high or low White population?

Data Description, Sources, and Appropriateness

The data set contains information like the 2020 unemployment rate ("Unemployment Rate and Labor Force Data for California Areas Detailed"), education attainment rates for White and Asians ("Educational Attainment"), and the population of White and Asians ("California Remained Most Populous State But Growth Slowed Last Decade"). This information is available for each county in California and all the values are listed as percentages. This data is appropriate because it contains White population rates which will then be broken into two different samples of "High Population" and "Low Population" using the threshold of 65 (=< 65 "low" and >= 65 "high"). These two samples and their means will be compared to see if the unemployment rate will be influenced by high or low White population.

Test Identification

Two-sample randomization test

Test Appropriateness

After creating Q-Q plots for all the variables involved in the research, the plots show signs of abnormality which indicates that it won't be possible to conduct a test that doesn't require Bootstrapping. Since the choices have been narrowed down to only randomization tests, the final decision depends on the details of the research question. Since the question is testing if unemployment will be influenced by "High White Population" or "Low White Population," the test that makes the most sense is the two-sample randomization test. Two samples can also be tested with a paired randomization test, but the samples here are independent which means that the assumptions for a paired test are not met.

Test Characteristics and Explanations

Characteristics: Quantitative, Two-tailed test, and Testing the means The data set is quantitative which means that it requires a test that deals with quantitative data. Since my research question tests if the response variable would be different if it received another "treatment," a two-tailed test makes the most sense. The statistics getting tested are the means because the data doesn't have any concerning outliers so there is no need to use the median.

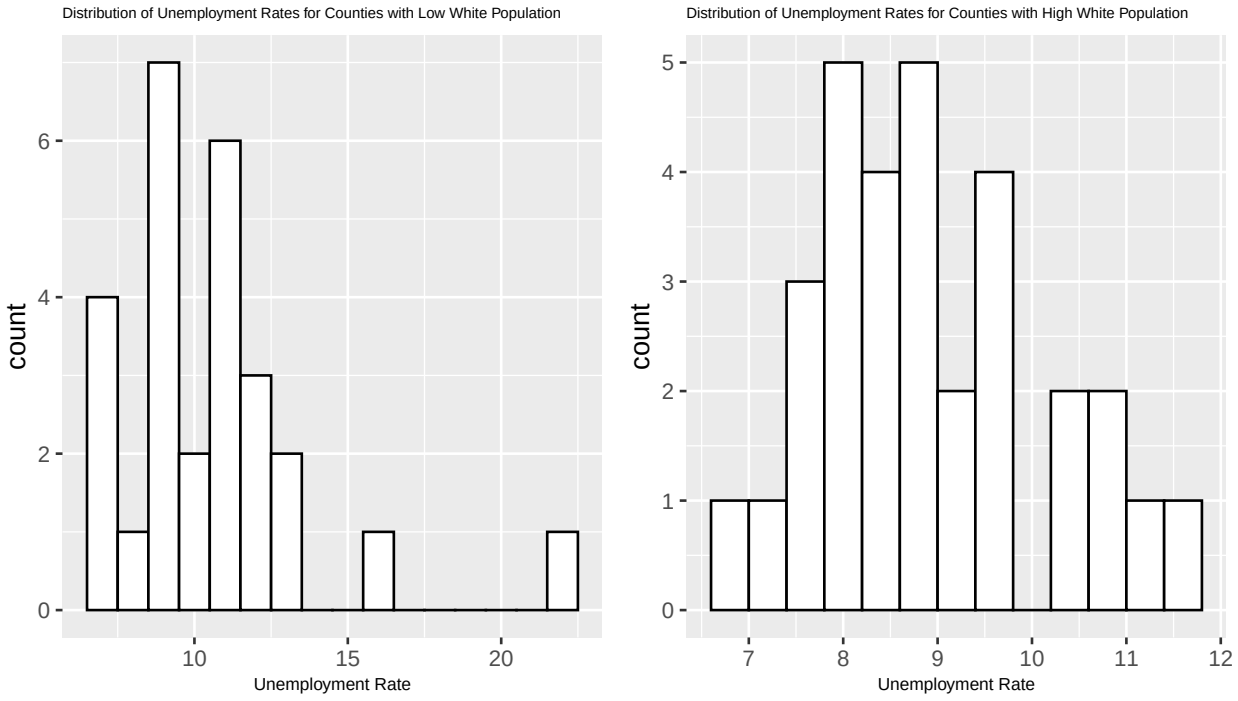
Test Assumptions and Validities

1. The sampled data represent an independent, representative sample from the population.

Each of the observations from the data are different counties in California which means one county's unemployment and White population rate would have no effect on the unemployment and White population rate of another. The collectors of the data (Employment Development Department) claimed that this data is representative of the population because they collected unemployment data from every county in California.

2. Bootstrapping will not work as expected if the underlying population has very heavy tails.

```
plot1 = ggplot(WhiteL, aes(x=LowWhitePop))+geom_histogram(binwidth=1, fill="white", color="black")+
  labs(title="Distribution of Unemployment Rates for Counties with Low White Population",
                    x="Unemployment Rate")+
  theme(plot.title=element_text(size=6),axis.title.x=element_text(size=7))
plot2 = ggplot(WhiteH, aes(x=HighWhitePop))+geom_histogram(binwidth=0.4, fill="white", color="black")+
  labs(title="Distribution of Unemployment Rates for Counties with High White Population",
                    x="Unemployment Rate")+
  theme(plot.title=element_text(size=6),axis.title.x=element_text(size=7))
grid.arrange(plot1, plot2, ncol=2)
```



To test if this assumption was met, two histograms, one for "High White Population" and one for "Low White Population," were created. Neither of the histograms took on the shape of a uniform distribution (straight across) so there was no indication of the populations being heavy-tailed.

Test Hypotheses

Null Hypothesis

H0: muWhiteH = muWhiteL
Null hypothesis: The mean for high White population is equal to mean for low White population.

Alternative Hypothesis

HA: muWhiteH != muWhiteL
Alternative hypothesis: The mean for high White population is not equal to mean for low White population.

Two-sample Randomization Test

```
set.seed(11302001)
B = 10000
meanWhiteL = mean(WhiteL$LowWhitePop)
meanWhiteH = mean(WhiteH$HighWhitePop)
samp_diff = meanWhiteH - meanWhiteL
rand.test <- function(x){
  rand_comb <- sample(c(WhiteL$LowWhitePop, WhiteH$HighWhitePop))
  bmean1 <- mean(rand_comb[1:x])
  bmean2 <- mean(rand_comb[(x+1):(length(WhiteL$LowWhitePop)+length(WhiteH$HighWhitePop))])
  bmean1 - bmean2
}
boot_diffs_null <- replicate(B, rand.test(length(WhiteL$LowWhitePop)))
sum(boot_diffs_null <= samp_diff | boot_diffs_null >= 2*mean(boot_diffs_null) - samp_diff)/B
```

```
## [1] 0.0042
```

Test Results, Conclusions, and Generalization

The test results in a p-value of 0.0042 which means the null hypothesis is rejected as there is evidence that the alternative hypothesis is true. This means that the means for low and High White population are not equal. In the context of the research question, this means that the amount of White people living in a county (high or low) does impact the unemployment rate of that county.

My data and research question specifically focused on California because the state has one of the most diverse populations in the United States. Since White people hold the population majority in almost all of U.S. states, this conclusion will probably hold true if the data was extended to include counties from the other fifty states.

References

"Unemployment Rate and Labor Force Data for California Areas Detailed":
https://www.labormarketinfo.edd.ca.gov/data/unemployment-and-labor-force.html
https://www.labormarketinfo.edd.ca.gov/file/lfhist/20aacou.pdf

"California Remained Most Populous State But Growth Slowed Last Decade":
https://www.census.gov/library/stories/state-by-state/california-population-change-between-census-decade.html

"Educational Attainment":
https://data.census.gov/cedsci/table?q=S1501&g=0400000US06%240500000&tid=ACSST5Y2019.S1501&hidePreview=true&tp=true