Project 1

Aidan Dunleavy

10/16/2020

The goal of this project is to determine if the total COVID-19 cases in a state is proportional to the total population of that state. The reason behind why this is being done is due to the recent COVID-19 pandemic, bringing more interest toward the topic.

Data title: United States COVID-19 Cases and Deaths by State

Date generated: Fri Oct 16 2020 15:23:08 GMT-0500 (Central Daylight Time)

Data source link: <https://covid.cdc.gov/covid-data-tracker/#cases_casesper100klast7days>

covid = read.csv("https://raw.githubusercontent.com/AidanDunleavy/Fall-2020-STAT-40001/master/united\_states\_covid19\_cases\_and\_deaths\_by\_state.csv", header = TRUE)

Link to population data: <https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html>

Data title: Population, Population Change, and Estimated Components of Population Change: April 1, 2010 to July 1, 2019 (NST-EST2019-alldata)

Data information: <https://www2.census.gov/programs-surveys/popest/technical-documentation/file-layouts/2010-2019/nst-est2019-alldata.pdf>

populations = read.csv("https://raw.githubusercontent.com/AidanDunleavy/Fall-2020-STAT-40001/master/nst-est2019-alldata.csv", header = TRUE)  
  
# Note that all data had to be edited for merging purposes only.  
# No data values were changed, rather they were either removed or reordered.  
# The data is restricted to the classic 50 states, Puerto Rico, and the  
# District of Columbia.  
# there is a total of 14 variables and 52 observations in the  
# merged data set. However, only two variables will be used:  
# total cases (Oct 16 2020) and population estimates (2019).  
  
covid2 = cbind(covid, populations$POPESTIMATE2019)  
head(covid2)

## State.Territory Total.Cases Confirmed.Cases Probable.Cases  
## 1 Alaska 10323 null null  
## 2 Alabama 169162 148659 20503  
## 3 Arkansas 96524 null null  
## 4 Arizona 228748 223692 5056  
## 5 California 858401 null null  
## 6 Colorado 81918 76377 5541  
## Cases.in.Last.7.Days Case.Rate.per.100000 Total.Deaths Confirmed.Deaths  
## 1 1318 1400 65 null  
## 2 7187 3461 2756 2590  
## 3 6379 3203 1645 null  
## 4 5347 3190 5789 5502  
## 5 23601 2170 16757 null  
## 6 6133 1438 2162 1783  
## Probable.Deaths Deaths.in.Last.7.Days Death.Rate.per.100000  
## 1 null 5 8  
## 2 166 119 56  
## 3 null 138 54  
## 4 287 46 80  
## 5 null 396 42  
## 6 379 67 37  
## Case.Rate.per.100000.in.Last.7.Days Death.Rate.per.100K.in.Last.7.Days  
## 1 25.5 0.1  
## 2 21.0 0.4  
## 3 30.2 0.7  
## 4 10.7 0.1  
## 5 8.5 0.1  
## 6 15.4 0.2  
## populations$POPESTIMATE2019  
## 1 731545  
## 2 4903185  
## 3 3017804  
## 4 7278717  
## 5 39512223  
## 6 5758736

# to statistically determine the proportionality we shall use a   
# Chi-squared GOF test.

The parameter is the true distribution of COVID-19 cases in a state

null hypothesis: The number of COVID-19 cases in each state is proportional to the population of that state.

alternative hypothesis: The number of COVID-19 cases in each state is NOT proportional to the population of that state.

We shall conduct a X^2 GOF test at alpha = 0.05.

chitest = chisq.test(covid2$Total.Cases, p = covid2$`populations$POPESTIMATE2019`/sum(covid2$`populations$POPESTIMATE2019`))  
cntrb = (chitest$observed - chitest$expected)^2/chitest$expected  
chitest

##   
## Chi-squared test for given probabilities  
##   
## data: covid2$Total.Cases  
## X-squared = 814287, df = 51, p-value < 2.2e-16

Our p-value is low (lower than alpha = 0.05), so we reject the null hypothesis in favor of the alternative hypothesis. Thus, we conclude with 95% confidence that the number of COVID-19 cases in each state is NOT proportional to the population of that state.

CONCLUSIONs:

After performing the test, it is quite apparent that there is extremely little chance that the number of COVID-19 cases in each state is proportional to the population of that state. Some things that could be improved is the accuracy of the population data when the 2020 census is complete. As currently the population data used in the test is an estimation for 2019 based on a census from 2010.