A STATISTICAL CORRELATION ANALYSIS BETWEEN PERCENTAGE OF PEOPLE (AGED 18 AND UNDER) WITHOUT HEALTH INSURANCE PER COUNTY AND POPULATION PER COUNTY IN THE UNITED STATES

Team Research and Development Project

A\_group-118

|  |  |
| --- | --- |
| Name | Student ID |
| MALLESH DHARMARAJ | 20070455 |

Contents

[Introduction 3](#_Toc93095650)

[Description of the data set 3](#_Toc93095651)

[Research Question: 3](#_Toc93095652)

[Null Hypothesis: 4](#_Toc93095653)

[Alternative Hypothesis: 4](#_Toc93095654)

[Visualization 4](#_Toc93095655)

[RStudio script for scatterplot: 4](#_Toc93095656)

[RStudio script for histogram: 5](#_Toc93095657)

[Analysis 6](#_Toc93095658)

[RStudio script for analysis 6](#_Toc93095659)

[Conclusion 7](#_Toc93095660)

## Introduction

The number of people without health insurance is growing high in some countries. So far in US, there are more than 29.6 million health uninsured as per the Census Bureau reporting. It’s a huge issue within the US. There can be some relationship between this number and why it’s increasing this way. So, our topic is to find the count of people (age 18 and under) without health insurance is dependent on the population (age 18 and under).

The research is based on the question of “Is there a correlation between the percentage of people (age under 18) without health insurance per county and the population (age under 18) per county in the United States?”. As per the given introduction, the target is to find any relation to the population. The null hypothesis is to set no correlation between the number of people (age under 18) health uninsured and the population (age under 18) per county in the United States. And then the alternative hypothesis is set to be there is a correlation between the number of people (age under 18) health uninsured and the population (age under 18) per county in the United States.

### Description of the data set

The data set was gone through from the “data.world”. "data.world" is home to the world’s largest collaborative data community, which is free and open to the public. It’s where people discover data, share analysis and team up on everything from social bot detection to award-winning data journalism. The data was inherited from US citizens about the health insurance and the age with their counties. And the source of the collection was from *NCHS Urban-Rural Classification Scheme for Counties (*www.cdc.gov. (2019)). There are the following 10 attributes in the data set.

The attributes in the dataset are described as follows.

|  |
| --- |
| Year : Data collected of year  Fips : General geographical identification code in USA  state\_fips : State identification of FIPS (Same as post code)  county\_fips : County identification of FIPS (Same as post code)  county\_name : County name  state\_name : State name  state\_abbr : State short name  metro\_nonmetro : Whether the county is metropolitan or non-metropolitan  uninsured : Number of uninsured people |

population\_18under : Age under 18 population

From the above attributes fips, state\_fips, county\_fips, uninsured, population\_18under columns contain integer values. All others contain text and year data columns. The dependent variable of the dataset is the percentage of people without health insurance in the county. The number of uninsured people column has a mean value of 1,874.317 and the population has a mean value of 24,375.125.

### Research Question:

Is there a correlation between the percentage of people (age 18 and under) without health insurance per county and the population (age 18 and under) per county in the United States?

### Null Hypothesis:

There is no correlation between the percentage of people (age 18 and under) without health insurance per county and the population (age 18 and under) per county in the United States.

### Alternative Hypothesis:

There is a correlation between the percentage of people (age 18 and under) without health insurance per county and the population (age 18 and under) per county in the United States.

## Visualization

The correlation is meant to be the connection between two variables in a dataset on the view on statistical. The scatter plot is the graph that has all XY pairs of the dataset mocked in one place. This looks somewhat like a hundred dots and the design is dependent on the dataset distribution as most counties have lesser than 20000 population.

The figure 01 we have plotted the Population (Age 18 and under) per county against the percentage of people (Age 18 and under) without health insurance per county. When we analyze Figure1, it shows when the number of populations is increasing, the percentage of health uninsured people was gradually decreasing.

### RStudio script for scatterplot:

*library(tidyverse)*

*my\_data <- read.csv("Health insurance (18 and under).csv")*

*#Normalizing data*

*#selecting only the data of year 2015*

*HtInsur <- my\_data %>% filter(year==2015)*

*#creating a column of percentage of uninsured people*

*HtInsur$uninsured1=(HtInsur$uninsured/HtInsur$population\_18under)\*100*

*#assigning data*

*population\_under18 <- HtInsur$population\_18under*

*percent\_Uninsured\_people\_under18 <- HtInsur$uninsured1*

*#finding the correlation*

*cor(population\_under18,percent\_Uninsured\_people\_under18,use="pairwise.complete.obs")*

*#creating a scatter plot*

*pdf("Scatterplot.pdf")*

*plot(population\_under18,*

*percent\_Uninsured\_people\_under18,*

*main = "% of people (Age 18 and under) without health insurance per \n county Vs Population (Age 18 and under) per county in the US",*

*xlab = "Population (Age 18 and under) per county",*

*ylab = "% of people (Age 18 and under) without health insurance per county",*

*pch = 19)*

*legend(x="topright",*

*legend =c("correlation"),*

*fill=c("red"))*

*#creating a correlation line*

*model <- lm(percent\_Uninsured\_people\_under18 ~ population\_under18, data = HtInsur)*

*abline(model, col="red")*

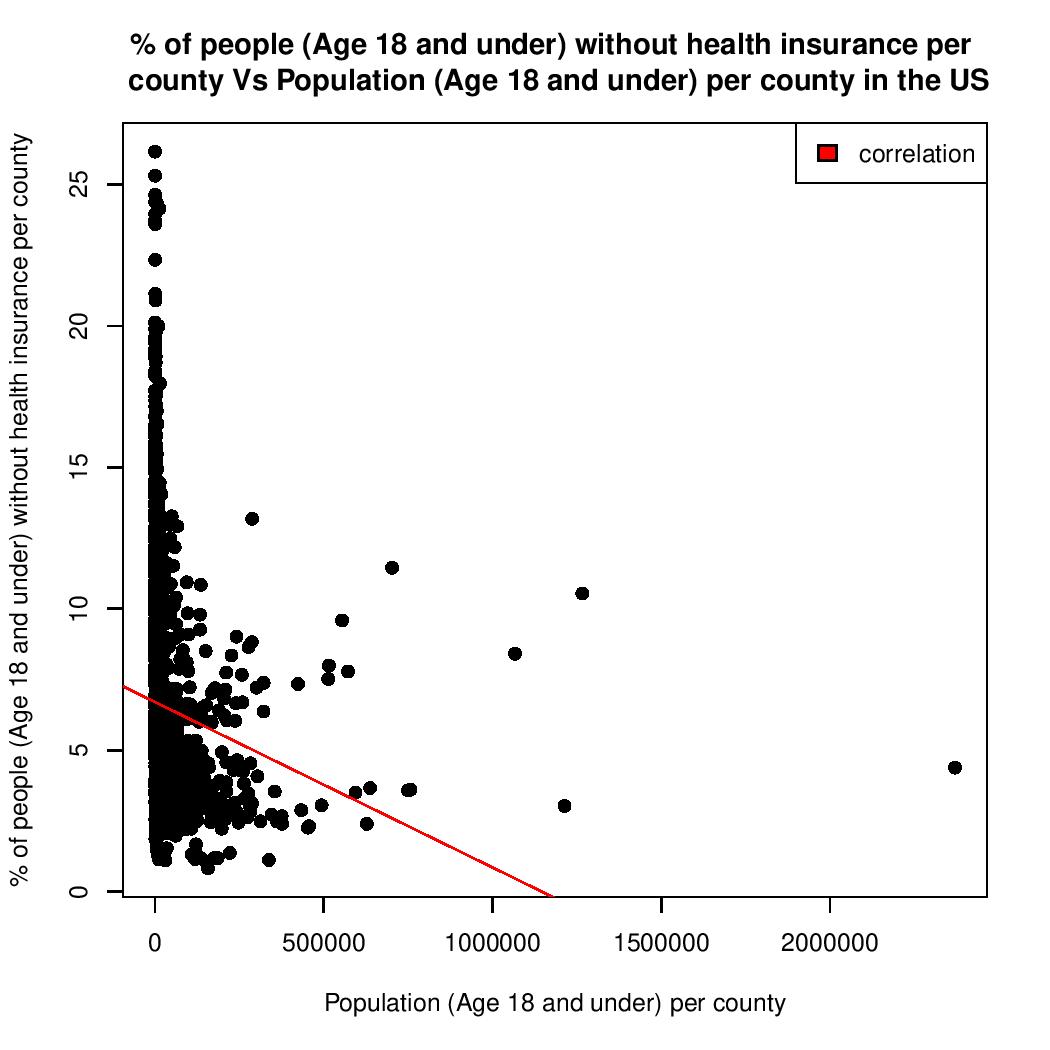
*dev.off()*

Figure 1 Percentage of people without health insurance per county vs population

The histogram shows the distribution of the dataset referencing column with using series of bars. Each bar covers the whole set of data with numerical values. This is somewhat like the bar charts. Figure 02 shows a histogram that shows the distribution of frequency distribution for percentage of people (Age 18 and under) without health insurance per county in the US

### RStudio script for histogram:

*dt<- HtInsur$uninsured1*

*dtMin=min(dt,na.rm=TRUE)*

*dtMax=max(dt,na.rm=TRUE)*

*dtMean=mean(dt,na.rm=TRUE)*

*dtSd=sd(dt,na.rm=TRUE)*

*#creating a histogram*

*pdf("Histogram.pdf")*

*h<-hist(dt, breaks=20, density=50,*

*main = "Frequency distribution for % of people (Age 18 and under) without \n health insurance per county in the US",*

*xlab = "% of people (Age 18 and under) without health insurance per county",*

*ylab = "Frequency",*

*xlim=c(0,dtMax),*

*ylim=c(0,1000),*

*col ="gray")*

*legend(x="topright",*

*legend =c("normal curve"),*

*fill=c("blue"))*

*#creating a normal curve*

*x<-seq(dtMin,dtMax,1)*

*y1<-dnorm(x, mean=dtMean, sd=dtSd)*

*y1<-y1\*diff(h$mids[1:2])\*length(dt)*

*lines(x,y1,col="blue")*

*dev.off()*

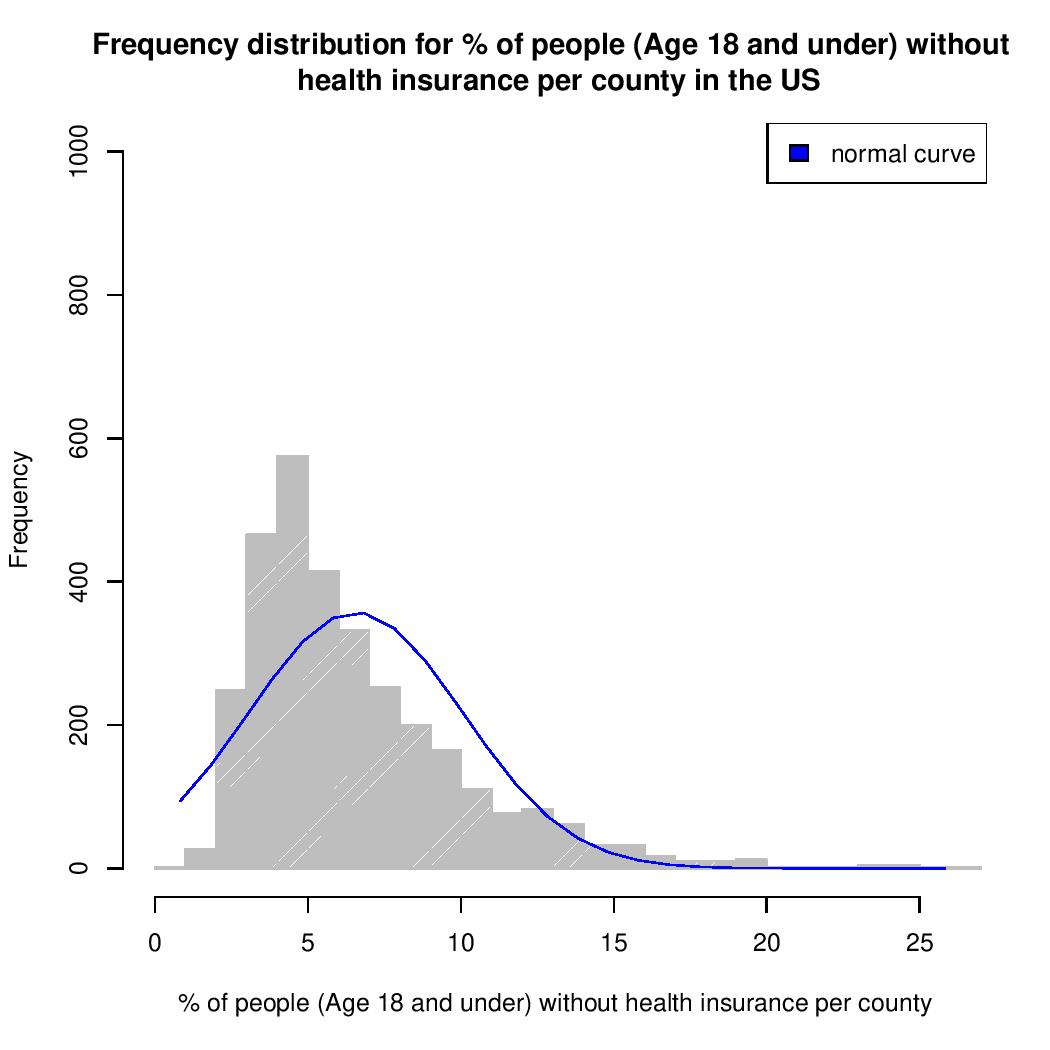
**

Figure 2 Frequency distribution for percentage of people (Age 18 and under) without health insurance per county in the US

## Analysis

Spearman’s Rank Correlation Coefficient (ρ): a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function. (Wikipedia Contributors (2019))

### RStudio script for analysis

*library(tidyverse)*

*my\_data <- read.csv("Health insurance (18 and under).csv")*

*HtInsur <- my\_data %>% filter(year==2015)*

*HtInsur$uninsured1=(HtInsur$uninsured/HtInsur$population\_18under)\*100*

*x <- as.numeric(HtInsur$population\_18under)*

*y <- as.numeric(HtInsur$uninsured1)*

*cor(y,x,use="pairwise.complete.obs",method="spearman")*

*result <- cor.test(y, x, method="spearman", exact = FALSE)*

*print(result)*

The Output of the Spearman’s Rank Correlation Coefficient Test Statistic performed in RStudio is given as follows:

***Spearman's rank correlation rho***

***data: y and x***

***S = 7528680135, p-value < 2.2e-16***

***alternative hypothesis: true rho is not equal to 0***

***sample estimates:***

***rho***

***-0.4576939***

In this Analysis we used Spearsman’s rank correlation as the data seems to be not normally distributed. The correlation is to be found between the percentage of people (age 18 and under) without health insurance per county and the population (age 18 and under) per county in the United States as the p-value < 2.2e-16 and the value of spearman’s coefficient Rho = 0.4576939, which is a significant Negative correlation.

It shows that the direction of the association between X (the independent variable) and Y (the dependent variable). If Y tends to increase when X increases, the Spearman correlation coefficient is Negative. The two variables here are the percentage of people (age 18 and under) without health insurance per county and the population (age 18 and under). This method is giving an output of a p-value less than 0.05 meaning that it rejects the null hypothesis, to better understand that this method is giving us good results one can check using different correlation methodologies that can confirm which assumption is true in the hypothesis.

## Conclusion

There are several measures that one needs to follow to find the correlation between two variables, one is independent and the other dependent variable. Firstly, it is important to collect the data and understand how the raw data can be pre-processed and then move on to the stage where further computation is considered for the columns that need to be tested for association. In the case of this research question, the columns that need to be assessed are the percentage of people (age under 18) without health insurance per county and the population (age under 18) per county which are plotted in a diagram to understand the distribution, from the analysis we have statistical evidence to reject the null hypothesis (H0) as P < 0.05.

## References

data.world. (n.d.). data.world. [online] Available at: https://data.world/dc-data-journalism/urban-rural-health-and-demographic-data/workspace/file?filename=Health+insurance+%2818+and+under%29.csv [Accessed 3 Jan. 2022].

www.cdc.gov. (2019). Data Access - Urban Rural Classification Scheme for Counties. [online] Available at: https://www.cdc.gov/nchs/data\_access/urban\_rural.htm. [Accessed 3 Jan. 2022].

Wikipedia Contributors (2019). Spearman’s rank correlation coefficient. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Spearman%27s\_rank\_correlation\_coefficient. [Accessed 3 Jan. 2022].

data.world. (n.d.). About us. [online] Available at: https://data.world/company/about-us [Accessed 14 Jan. 2022].