Exercise 3.2: American Community Survey Exercise

Sreenivasulu Somu

Department of Data Science, Bellevue University

DSC520-T301 Statistics for Data Science (2247-1)

Chase Denton

June 23, 2024

Exercise 3.2: American Community Survey Exercise

I. List the name of each field and what you believe the data type and intent is of the data included in each field (Example: Id - Data Type: varchar (contains text and numbers) Intent: unique identifier for each row)

Field Name	Data Type	Intent
Id	varchar	Unique identifier for each row, representing a specific geographic area.
Id2	number	Secondary identifier, likely a number representing the geographic area
Geography	varchar	Name of a county
PopGroupID	number	Code identifying the population group
POPGROUP.display-label	varchar	Human-readable label for the population group
RacesReported	number	Number of races reported in the data
HSDegree	number	Percentage of population with a high school degree
BachDegree	number	Percentage of population with a bachelor's degree

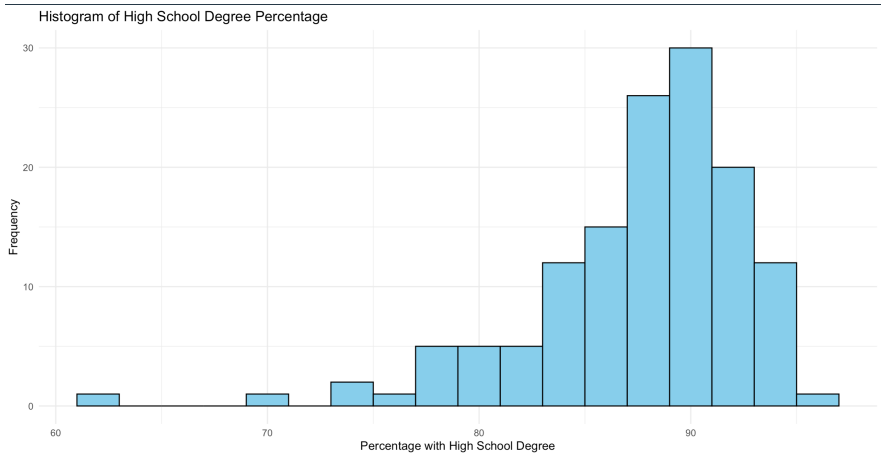
II. Run the following functions and provide the results: str(); nrow(); ncol()

```
> # Read the CSV file
> csv_data <- read.csv("acs-14-1yr-s0201.csv")
> # Use str() to print the structure of the csv data
> cat("Structure of the data frame:",str(csv_data))
'data.frame':	136 obs. of  8 variables:
 $ Id                : chr  "0500000US01073" "0500000US04013" "0500000US04019"
"0500000US06001" ...
 $ Id2               : int  1073 4013 4019 6001 6013 6019 6029 6037 6059 6065 ...
 $ Geography         : chr  "Jefferson County, Alabama" "Maricopa County, Arizona" "Pima County,
Arizona" "Alameda County, California" ...
 $ PopGroupID        : int  1 1 1 1 1 1 1 1 1 1 ...
 $ POPGROUP.display.label: chr  "Total population" "Total population" "Total population" "Total
population" ...
 $ RacesReported     : int  660793 4087191 1004516 1610921 1111339 965974 874589 10116705
3145515 2329271 ...
 $ HSDegree          : num  89.1 86.8 88 86.9 88.8 73.6 74.5 77.5 84.6 80.6 ...
 $ BachDegree        : num  30.5 30.2 30.8 42.8 39.7 19.7 15.4 30.3 38 20.7 ...
Structure of the data frame:
> # Use nrow() to get the number of rows
> cat("Number of rows:", nrow(csv_data))
Number of rows: 136
> # Use ncol() to get the number of columns
> cat("Number of columns:", ncol(csv_data))
Number of columns: 8
>
```

III. Create a Histogram of the HSDegree variable using the ggplot2 package.

a. Set a bin size for the Histogram that you think best visuals the data (the bin size will determine how many bars display and how wide they are)

b. Include a Title and appropriate X/Y axis labels on your Histogram Plot.



IV. Answer the following questions based on the Histogram produced:

a. Based on what you see in this histogram, is the data distribution unimodal?

The data distribution appears to be unimodal, there is a single peak in the histogram.

b. Is it approximately symmetrical?

The distribution is approximately symmetrical, not perfect. There is a slight lean towards the left side.

c. Is it approximately bell-shaped?

Yes, the distribution is approximately bell-shaped and has a central peak with tapering tails on both sides.

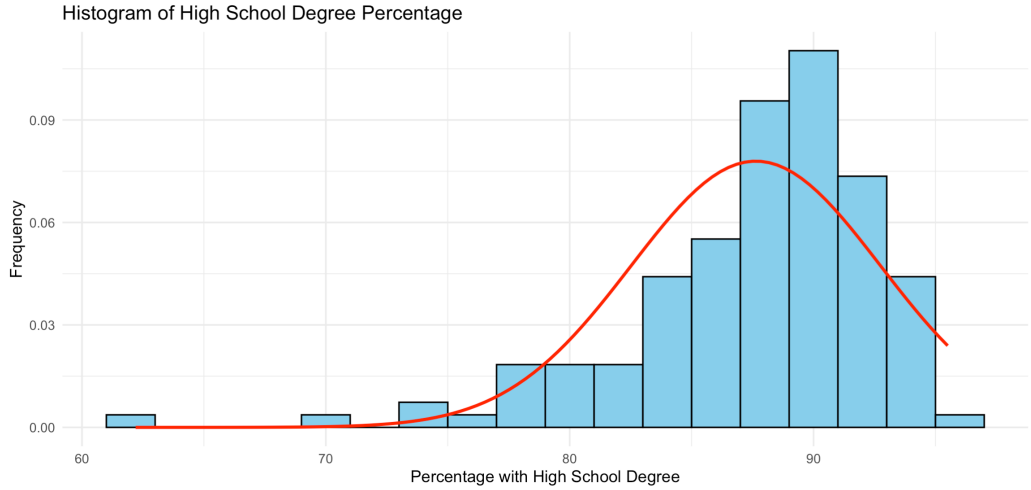
d. Is it approximately normal?

The distribution is close to normal, not exactly normal.

e. If not normal, is the distribution skewed? If so, in which direction?

The distribution shows a slight negative skew to the left side. It's clear from the longer tail on the left side of the histogram and the peak is slightly to the right of center.

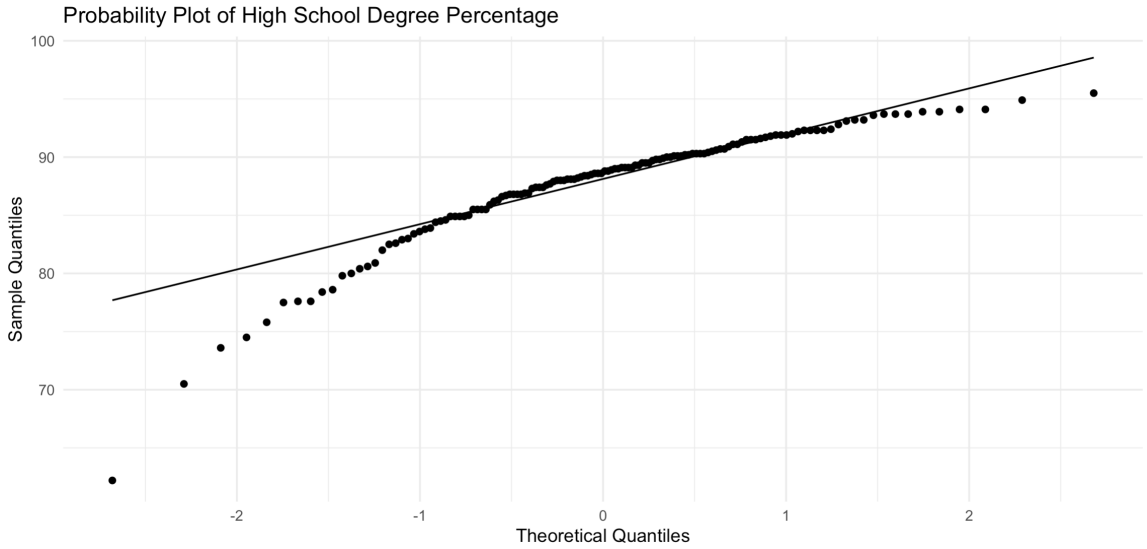
f. Include a normal curve to the Histogram that you plotted.



g. Explain whether a normal distribution can accurately be used as a model for this data.

Based on the histogram with the overlay normal curve, we can observe that a normal distribution would provide a rough approximation of the data and not perfect fit.

V. Create a Probability Plot of the HSDegree variable.



VI. Answer the following questions based on the Probability Plot:

a. Based on what you see in this probability plot, is the distribution approximately normal? Explain how you know.

The distribution of the HSDegree variable is approximately normal as most of the data points fall close to the reference line, which is represented by the theoretical quantiles of a normal distribution. The data points are aligned closely with the line indicating that the sample quantiles are similar to the theoretical quantiles, suggesting that the data follows in a normal distribution. Minor deviations are common at the tails and would not have significant deviation from the overall distribution.

b. If not normal, is the distribution skewed? If so, in which direction? Explain how you know.

Although the distribution is approximately normal, there is a slight indication of left skewness (negative skew) and is observed in the probability plot where the data points deviate slightly below the reference line at the lower end and above the line at the upper end. These patterns indicate that the lower quantiles are less than expected under normality, and the upper quantiles are greater, indicating a longer left tail.

VII. Now that you have looked at this data visually for normality, you will now quantify normality with numbers using the stat.desc() function. Include a screen capture of the results produced.

136 0 0 62.2 95.5 33.3 11918 88.7 87.63235 0.4388598 0.8679296 26.19332 5.117941 0.05840241

VIII. In several sentences provide an explanation of the result produced for skew, kurtosis, and z-scores. In addition, explain how a change in the sample size may change your explanation?

Skew: Negative skew indicates that the distribution has a longer tail on the left side, with more extreme low values than expected in a normal distribution.

Kurtosis: Positive kurtosis suggests that the distribution has heavier tails and a higher, sharper peak compared to a normal distribution.

Z-scores: Helps to quantify how many standard deviations an observation is from the mean. Extreme z-scores for skew and kurtosis (> 1.96) suggests a significant deviation from the normality.

Sample size: As sample size increases, the small deviations from normality could become statistically significant. In large samples, the z-scores for skew and kurtosis are likely to exceed the critical values, potentially leading to the rejection of normality assumptions.

Output of R program

```
> # Load required libraries
> library(ggplot2)
> # Load required libraries
> library(ggplot2)
> library(pastecs)
>
> # Read the CSV file
> csv_data <- read.csv("acs-14-1yr-s0201.csv")
> head(csv_data)
          Id  Id2                  Geography PopGroupID POPGROUP.display.label RacesReported
HSDegree BachDegree
1 0500000US01073 1073      Jefferson County, Alabama         1      Total population       660793
89.1       30.5
2 0500000US04013 4013       Maricopa County, Arizona         1      Total population      4087191
86.8       30.2
3 0500000US04019 4019          Pima County, Arizona         1      Total population      1004516
88.0       30.8
4 0500000US06001 6001      Alameda County, California         1      Total population      1610921
86.9       42.8
5 0500000US06013 6013 Contra Costa County, California         1      Total population
1111339     88.8       39.7
6 0500000US06019 6019       Fresno County, California         1      Total population       965974
73.6       19.7
>
> # II.    Run the following functions and provide the results: str(); nrow(); ncol()
>
> # Use str() to print the structure of the csv data
> cat("Structure of the data frame:",str(csv_data))
```

'data.frame': 136 obs. of 8 variables:

$ Id : chr "0500000US01073" "0500000US04013" "0500000US04019" "0500000US06001" ...

$ Id2 : int 1073 4013 4019 6001 6013 6019 6029 6037 6059 6065 ...

$ Geography : chr "Jefferson County, Alabama" "Maricopa County, Arizona" "Pima County, Arizona" "Alameda County, California" ...

$ PopGroupID : int 1 1 1 1 1 1 1 1 1 1 ...

$ POPGROUP.display.label: chr "Total population" "Total population" "Total population" "Total population" ...

$ RacesReported : int 660793 4087191 1004516 1610921 1111339 965974 874589 10116705 3145515 2329271 ...

$ HSDegree : num 89.1 86.8 88 86.9 88.8 73.6 74.5 77.5 84.6 80.6 ...

$ BachDegree : num 30.5 30.2 30.8 42.8 39.7 19.7 15.4 30.3 38 20.7 ...

Structure of the data frame:>

> # Use nrow() to get the number of rows

> cat("Number of rows:", nrow(csv_data))

Number of rows: 136>

> # Use ncol() to get the number of columns

> cat("Number of columns:", ncol(csv_data))

Number of columns: 8>

> # III. Create a Histogram of the HSDegree variable using the ggplot2 package.

> # a. Set a bin size for the Histogram that you think best visuals the data (the bin size will determine how many bars display and how wide they are)

> # b. Include a Title and appropriate X/Y axis labels on your Histogram Plot.

```
> ggplot(csv_data, aes(x = HSDegree)) +
+   geom_histogram(binwidth = 2, fill = "skyblue", color = "black") +
+   labs(title = "Histogram of High School Degree Percentage",
+        x = "Percentage with High School Degree",
+        y = "Frequency") +
+   theme_minimal()
```

> # IV. Answer the following questions based on the Histogram produced:

> # Create histogram with normal curve

```
> #   a.   Based on what you see in this histogram, is the data distribution unimodal?
> #   b.   Is it approximately symmetrical?
> #   c.   Is it approximately bell-shaped?
> #   d.   Is it approximately normal?
> #   e.   If not normal, is the distribution skewed? If so, in which direction?
> #   f.   Include a normal curve to the Histogram that you plotted.
> ggplot(csv_data, aes(x = HSDegree)) +
+   geom_histogram(aes(y = ..density..), binwidth = 2, fill = "skyblue", color = "black") +
+   stat_function(fun = dnorm, args = list(mean = mean(csv_data$HSDegree),
+                                          sd = sd(csv_data$HSDegree)),
+                 color = "red", size = 1) +
+   labs(title = "Histogram of High School Degree Percentage",
+        x = "Percentage with High School Degree",
+        y = "Frequency") +
+   theme_minimal()
> # g.    Explain whether a normal distribution can accurately be used as a model for this data.
> # V.    Create a Probability Plot of the HSDegree variable.
>
> # Probability plot
> ggplot(csv_data, aes(sample = HSDegree)) +
+   stat_qq() +
+   stat_qq_line() +
+   labs(title = "Probability Plot of High School Degree Percentage",
+        x = "Theoretical Quantiles",
+        y = "Sample Quantiles") +
+   theme_minimal()
> # VI.   Answer the following questions based on the Probability Plot:
> #   a.   Based on what you see in this probability plot, is the distribution approximately normal?
Explain how you know.
> # b.    If not normal, is the distribution skewed? If so, in which direction? Explain how you
know.
```

> # VII. Now that you have looked at this data visually for normality, you will now quantify normality with numbers using the stat.desc() function. Include a screen capture of the results produced.

> # VIII. In several sentences provide an explanation of the result produced for skew, kurtosis, and z-scores. In addition, explain how a change in the sample size may change your explanation?

>

> # Calculate descriptive statistics

> desc_stats <- stat.desc(csv_data$HSDegree)

> cat(desc_stats)

136 0 0 62.2 95.5 33.3 11918 88.7 87.63235 0.4388598 0.8679296 26.19332 5.117941 0.05840241

>