> # Assignment: ASSIGNMENT 2.1 Test Scores

> # Name: Blasdell, Blaine

> # Date: 2020-06-14

>

> ## Load the `readxl` library

> library(readxl)

> library(ggplot2)

> library(pastecs)

> library(e1071)

>

> ## Read the file `data/acs-14-1yr-s0201.csv` to `census\_df`

> ## Q1. What are the elements in your data (including the categories and data types)?

> **## A1. Census Date {ID, ID2, Geographic location, Population Group ID, Population Group Label, Races Reported, High School Degree(%), BS Degree (%)}**

> census\_df <- read.csv("data/acs-14-1yr-s0201.csv")

> head(census\_df, 10)

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

7 0500000US06029 6029 Kern County, California 1 Total population 874589 74.5 15.4

8 0500000US06037 6037 Los Angeles County, California 1 Total population 10116705 77.5 30.3

9 0500000US06059 6059 Orange County, California 1 Total population 3145515 84.6 38.0

10 0500000US06065 6065 Riverside County, California 1 Total population 2329271 80.6 20.7

>

>

> ## Q2. 2. Please provide the output from the following functions: str(); nrow(); ncol()

> ## A2. See Below

**> str(census\_df)**

**'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 ...**

**> cat("\nNumber of Rows:", nrow(census\_df),"\n")**

**Number of Rows: 136**

> cat('\n')

> cat("\nNumber of columns:",ncol(census\_df))

**Number of columns: 8>**

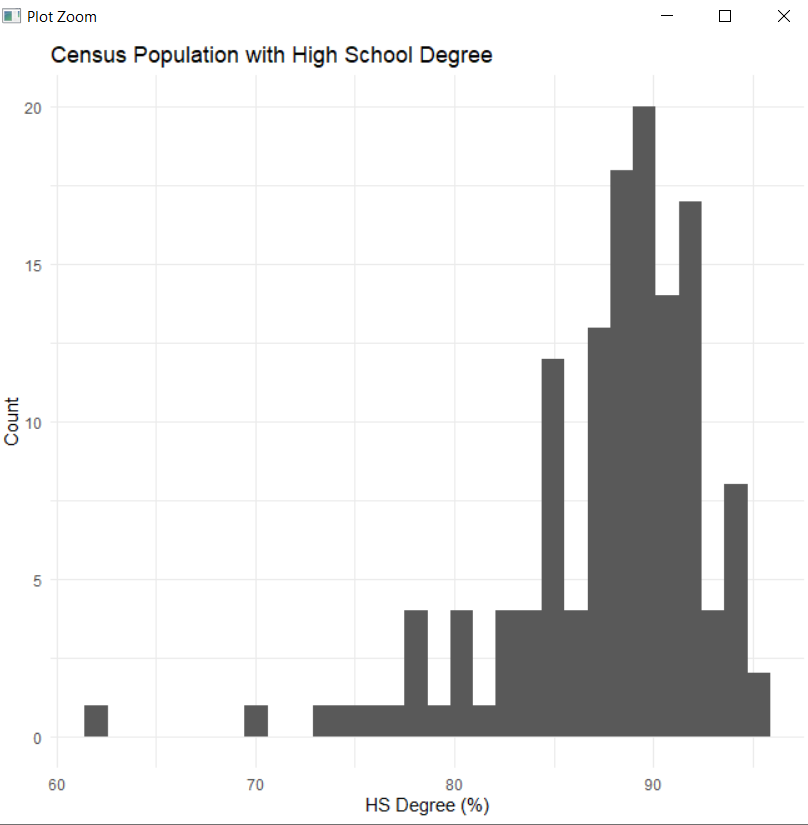
> ## Q3. Create a Histogram of the HSDegree variable using the ggplot2 package.

> ## Q3a. Set a bin size for the Histogram.

> ## Q3b. Include a Title and appropriate X/Y axis labels on your Histogram Plot.

>

> ggplot(census\_df, aes(HSDegree)) + geom\_histogram(bins=30) + ggtitle('Census Population with High School Degree') + xlab('HS Degree (%)') + ylab('Count')

> 

>

> ## Q4. Answer the following questions based on the Histogram produced:

> ## Q4a. Based on what you see in this histogram, is the data distribution unimodal?

**> ## A4a. Yes. it is unimodal.**

>

> ## Q4b. Is it approximately symmetrical?

> **## A4b. No. It is not symmetrical as it is a Left Skewed Distribution. Additionally Outliers in the 60s seem to prevent this.**

>

> ## Q4c. Is it approximately bell-shaped?

> **## A4c. No. It is unimodal which is what a bell curve is (although it is not uniform) and it is skewed.**

>

> ## Q4d. Is it approximately normal?

**> ## A4d. No it is not normal as it is not symmetrical.**

>

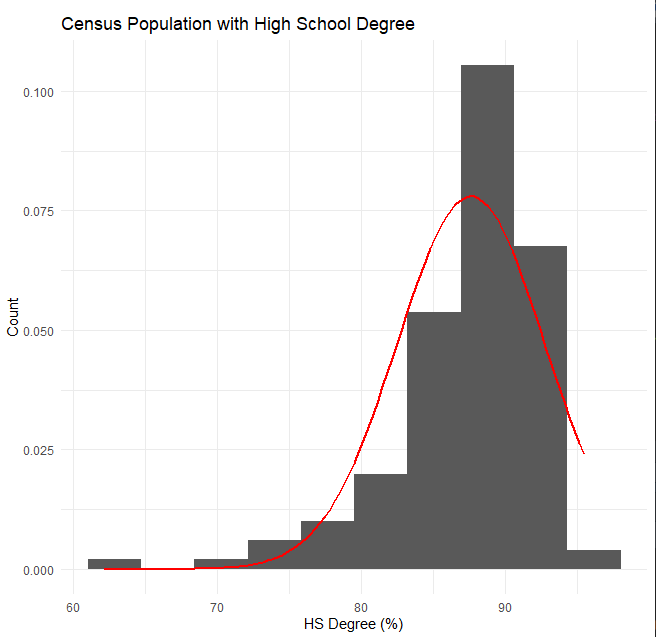
> ## Q4e. If not normal, is the distribution skewed? If so, in which direction?

**> ## A4e. Yes it is Left Skewed.**

>

> ## Q4f. Include a normal curve to the Histogram that you plotted.

> ggplot(census\_df, aes(HSDegree)) + geom\_histogram(aes(y=..density..), bins = 10) + ggtitle('Census Population with High School Degree') + xlab('HS Degree (%)') + ylab('Count') + stat\_function(fun = dnorm, args = list(mean = mean(census\_df$HSDegree), sd = sd(census\_df$HSDegree)),color="red", size=1, )

> 

> ## Q4g. Explain whether a normal distribution can accurately be used as a model for this data.

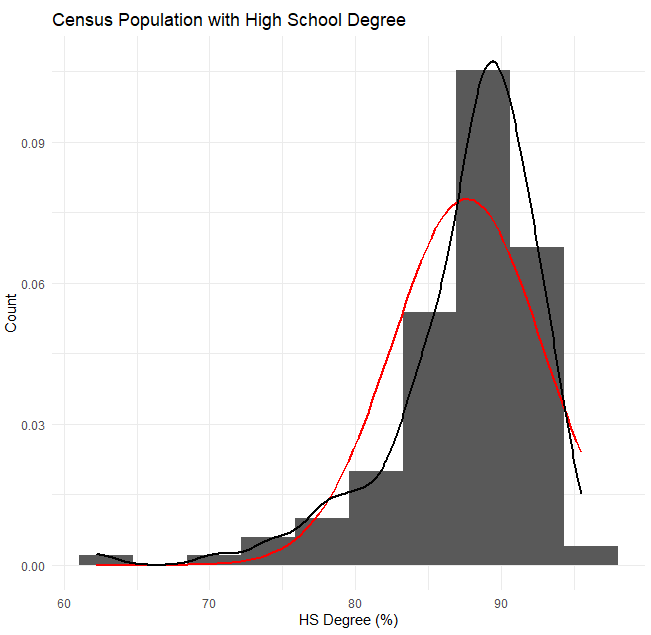
**> ## A4g. As this is skewed and not symmetrical a normal distribution would not work as a model.**

>

> ## Q5. Create a Probability Plot of the HSDegree variable.

> plt <- ggplot(census\_df, aes(HSDegree)) + geom\_histogram(aes(y=..density..), bins = 10) + ggtitle('Census Population with High School Degree') + xlab('HS Degree (%)') + ylab('Count') + stat\_function(fun = dnorm, args = list(mean = mean(census\_df$HSDegree), sd = sd(census\_df$HSDegree)),color="red", size=1, ) + geom\_density(color="black", size=1)

> plt

> 

>

> ## Q6. Answer the following questions based on the Probability Plot:

> ## Q6a. Based on what you see in this probability plot, is the distribution approximately normal? Explain how you know.

> ## Q6b. If not normal, is the distribution skewed? If so, in which direction? Explain how you know.

**> ## A6. The probability plot (Black Line) still appears to be skewed left. The mean is 87.6 and the median is 88.7.**

**> ## A6. The graph is neither symmetrical or normal at those points although the probability plot it is approaching normal.**

>

> ## Q7. Now that you have looked at this data visually for normality, you will now quantify normality with numbers using the stat.desc() function.

> ## Q7. Include a screen capture of the results produced.

**> stat.desc(census\_df$HSDegree, basic = TRUE, norm = FALSE)**

**nbr.val nbr.null nbr.na min max range sum median mean SE.mean**

**1.360000e+02 0.000000e+00 0.000000e+00 6.220000e+01 9.550000e+01 3.330000e+01 1.191800e+04 8.870000e+01 8.763235e+01 4.388598e-01**

**CI.mean.0.95 var std.dev coef.var**

**8.679296e-01 2.619332e+01 5.117941e+00 5.840241e-02**

>

> ##Q8. 8. 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?

**> ## Z-score for 70: 74.5 - 87.6/5.12 = - 2.56 (Used Excel to calculate based on book formula)**

**> ## Z-scoree for 90: 90.1-87.6/5.12 = .488**

**> ##A8. As the points are clustered on the right side of the distribution, it is negatively skewed. The outliers in the 60s and 70s tend to contribute to the outcome. With a Kurtosis of a positive value of 4.35, it is considered to be a leptokurtic or a fat-tailed distribution. The z-score helps us predict the probability that a person in this census would or would not have a High School Diploma. This later will also help us determine the confidence of the data at different levels. This is a relatively small sample size. As the size of the sample were to increase the closer we should approach a normal distribution.**

**>**

**kurtosis(census\_df$HSDegree)**

**[1] 4.352856**