# Assignment: ASSIGNMENT 3

# Name: MacDonald, Mariana

# Date: 2022-04-03

I - What are the elements in your data (including the categories and data types)?

**Answer:**

**Nominal: Id: chr, Geography: chr, POPGROUP.display.label: chr**

**Ordinal: Id2:int , PopGroupID : int , RacesReported: int , HSDegree: num, BachDegree : num**

## Load the ggplot2 package

library(ggplot2)

theme\_set(theme\_minimal())

## Set the working directory to the root of your DSC 520 directory

setwd("/Users/marianamacdonald/Documents/DATA SCIENCE/DSC 520/Statistics R/Week 2/dsc520")

acs <- read.csv("http://content.bellevue.edu/cst/dsc/520/id/resources/acs-14-1yr-s0201.csv")

## II - Please provide the output from the following functions: str(); nrow(); ncol()

str(acs)

nrow(acs)

ncol(acs)

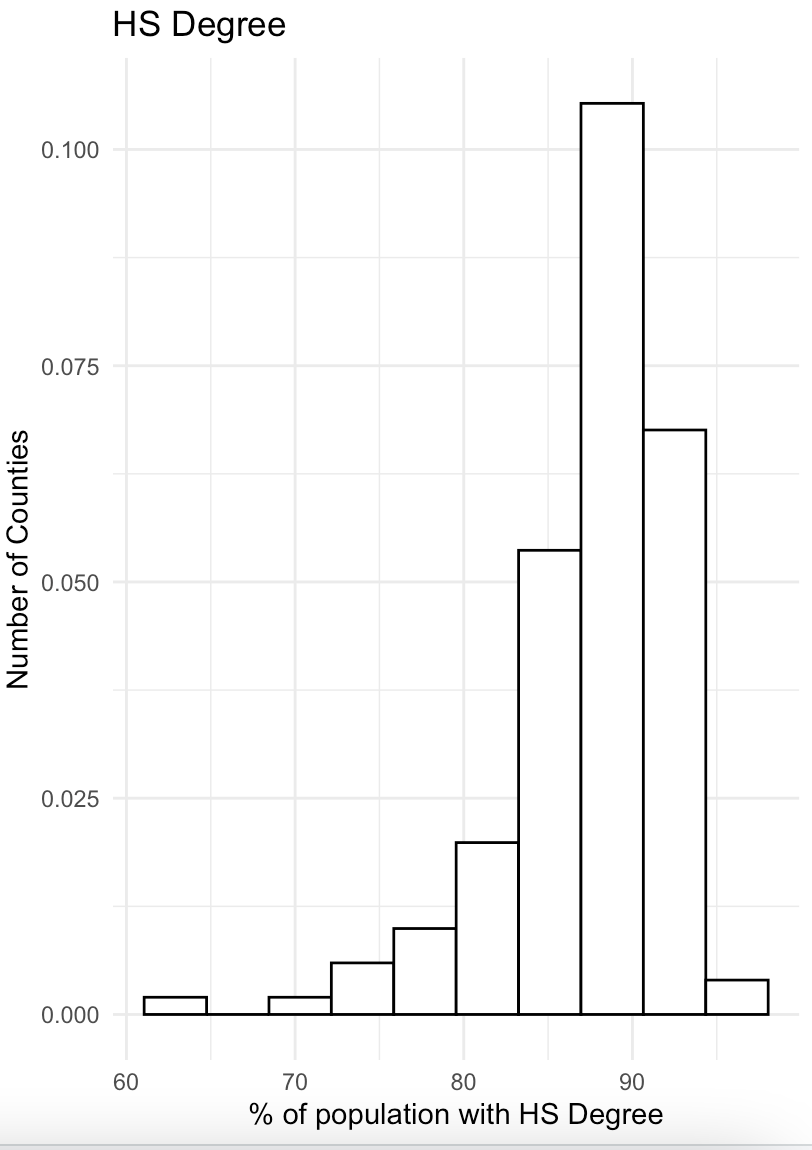
## III - Create a Histogram of the HSDegree variable using the ggplot2 package.

##1. Set a bin size for the Histogram.

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

hist <- ggplot(acs, aes(HSDegree)) + geom\_histogram (bins = 10, aes(y = ..density..), colour = "black", fill = "white") + ggtitle("HS Degree") + labs(x = "% of population with HS Degree", y = "Number of Counties")

hist



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

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

**Yes, there is a peak at around 90%**

**2. Is it approximately symmetrical?**

**No. the mean is not approximately equal to the median.**

**3. Is it approximately bell-shaped?**

**No.**

**4. Is it approximately normal?**

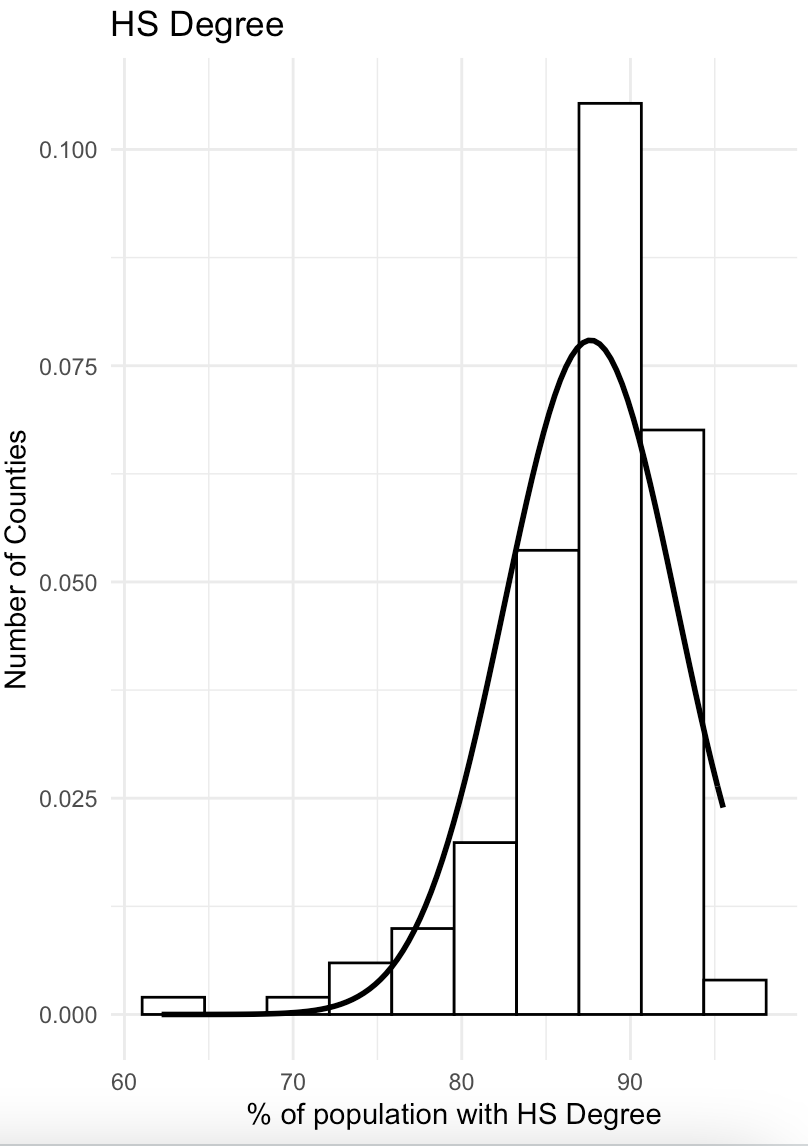
**No.**

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

**Negative, left skewed**

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

hist + stat\_function(fun = dnorm, args = list (mean = mean(acs$HSDegree, na.rm = TRUE), sd = sd(acs$HSDegree, na.rm = TRUE)), colour = "black", size = 1)



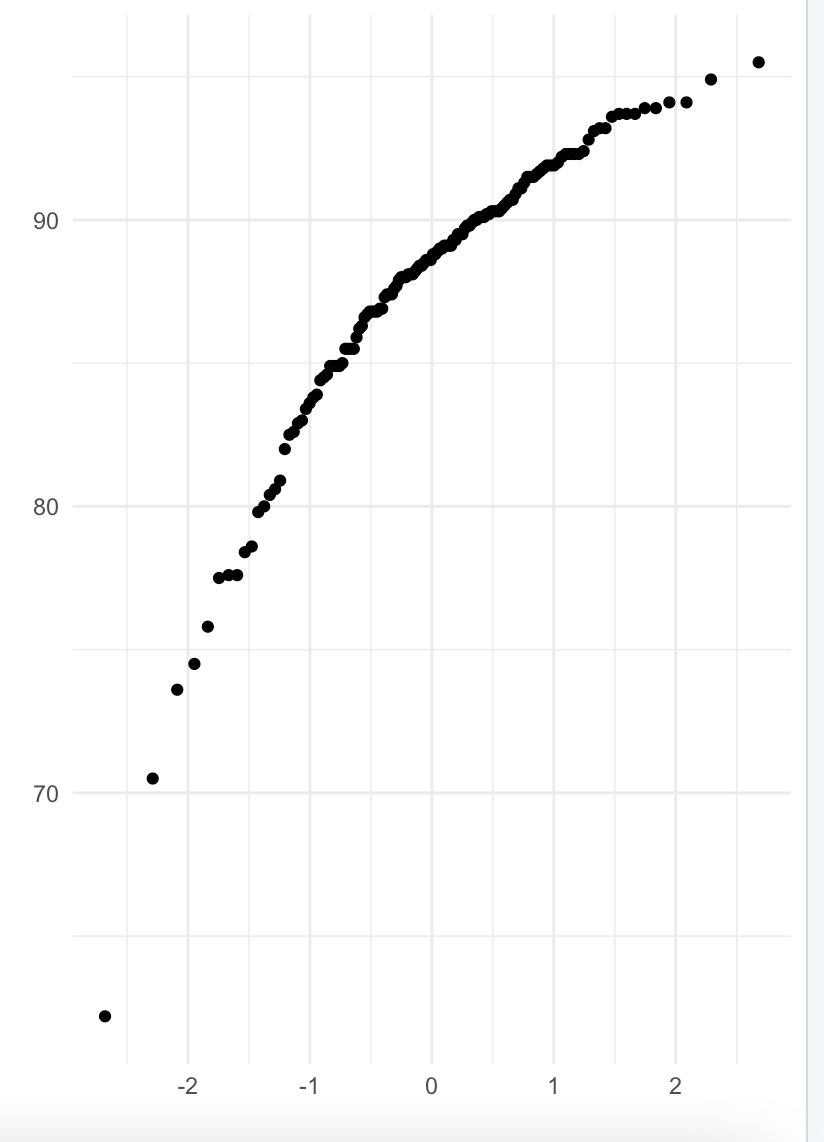
7. Explain whether a normal distribution can accurately be used as a model for this data.

No, because it's not normal, it's skewed.

## V - Create a Probability Plot of the HSDegree variable.

qqplot.HSDegree <- qplot(sample = acs$HSDegree, stat = "qq")

qqplot.HSDegree



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

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

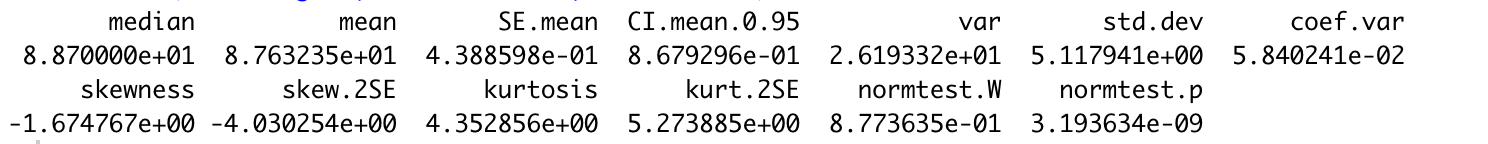
2. If not normal, is the distribution skewed? If so, in which direction? Explain how you know.

No, the data appears curved, which means it's a left skewed. If it was normal, it would be a straight line and it if was positive or right skewed, it would be curved but the opposite way.

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

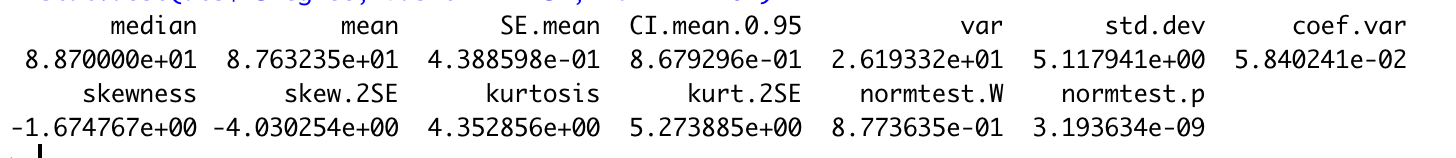
library(pastecs)

stat.desc(acs$HSDegree, basic = FALSE, norm = TRUE)



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

The



Mean: 87.63

Median: 88.70

Std. Dev: 5.11

Skewness -1.67

Kurtosis 4.35

The mean is less than the median, indicting negatively skewed data.

High Kurtosis (the tails extend farther than the three standard deviations of the normal bell-curved distribution.

The values of skew and kurtosis should be zero in a normal distribution.

Negative values indicate a pile-up on the right.

Negative values indicate a flat and light-tailed distribution. The further the value is from zero, the more likely it is that the data are not normally distributed.