**Name: Wiltse, Dan**

**Assignment: Week 3 Assignment**

**Date: 12-9-2019**

1. **What are the elements in your data (including the categories and data types)**

|  |  |  |
| --- | --- | --- |
| **Element** | **Category** | **Data Type** |
| **ID** | **Nominal** | **str** |
| **ID2** | **Ordinal** | **int** |
| **Geography** | **Nominal** | **str** |
| **PopGroupID** | **Ordinal** | **int** |
| **PopGroup.display-label** | **Nominal** | **str** |
| **RacesReported** | **Ratio** | **int** |
| **HSDegree** | **Ratio** | **decimal/int** |
| **BachDegree** | **Ratio** | **decimal/int** |

1. **2. Please provide the output from the following functions: str(); nrow(); ncol()**

**> str(acs)**

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

$ Id : Factor w/ 136 levels "0500000US01073",..: 1 2 3 4 5 6 7 8 9 10 ...

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

$ Geography : Factor w/ 136 levels "Alameda County, California",..: 56 70 98 1 20 43 62 68 92 106 ...

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

$ POPGROUP.display.label: Factor w/ 1 level "Total population": 1 1 1 1 1 1 1 1 1 1 ...

$ 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 ...

**> nrow(acs)**

[1] 136

**> ncol(acs)**

[1] 8

1. **Create a Histogram of the HSDegree variable using the ggplot2 package.**

**a. Set a bin size for the Histogram.**

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

See R code and plots below

**4. Answer the following questions based on the Histogram produced:**

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

Yes, unimodal as there aren’t multiple humps in the curve

**b. Is it approximately symmetrical?**

While the main part of the curve appears to be approximately symmetrical, as it appears it would look the same if you split it down the middle, there is a much longer tail on the left side of the chart than the right, so the overall curve appears to be skewed in one direction over the other, which is explained in more detail below.

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

It is approximately shaped like a bell in that it peaks in the middle and lower probability in each of the tails.

**d. Is it approximately normal?**

No

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

The distribution is left skewed, as it has negative skewness, and the mean is to the left of the median.

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

See R code and plots below

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

I would say normal distribution can’t be used as a model for this data because the data isn’t evenly distributed across the histogram, the majority of the data in the right side of of chart (with negative skewness), so its not normally distributed across the entire population.

Also, when running the shapiro test, which is based on the corresponding normal scores and the data, I get a p value of -0.000000003194, which means that distribution of the data is significantly different from normal distribution.

Shapiro-Wilk normality test

data: acs$HSDegree

W = 0.87736, p-value = 3.194e-09

**R code and Output for Question #4:**

#load library for ggplot

library(ggplot2)

#ggplot with binwidth determined

ggplot(acs, aes(x=HSDegree)) +

geom\_histogram(binwidth=0.6)

#histogram using hist function

hist(acs$HSDegree, freq=TRUE, col="gray", xlab="Pct w/ HS Degree", main="Distribution of HS Degrees")

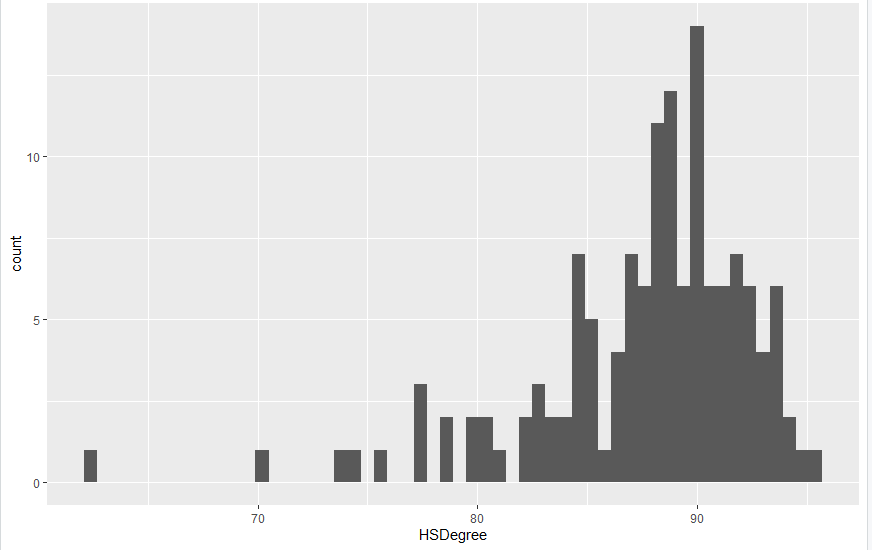
# adding a normal distribution line in histogram

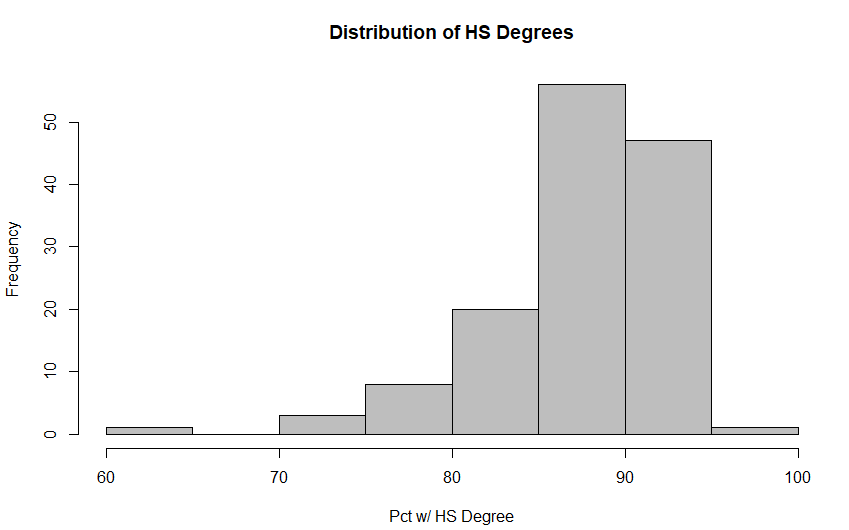
hist(acs$HSDegree, freq=FALSE, col="gray", xlab="Pct w/ HS Degree", main=" Distribution of HS Degrees")

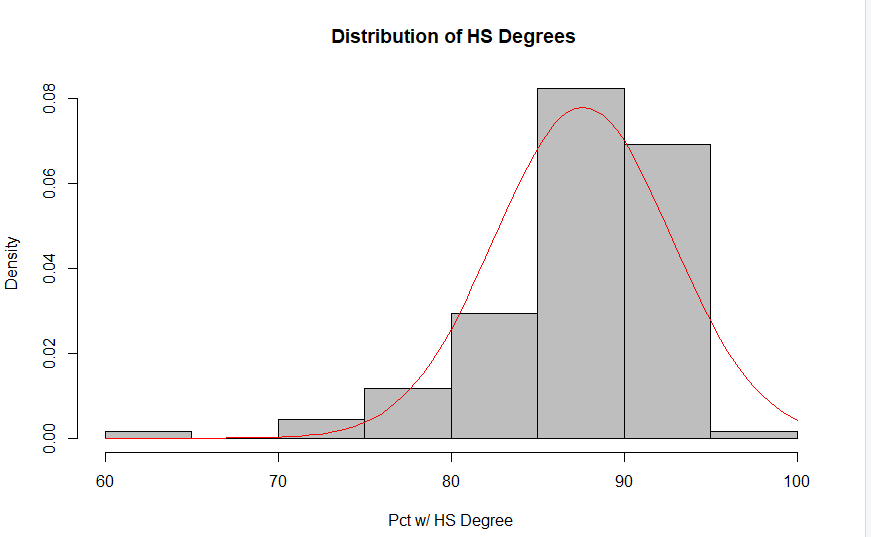
curve(dnorm(x, mean=mean(acs$HSDegree), sd=sd(acs$HSDegree)), add=TRUE, col="red") #line

shapiro.test(acs$HSDegree)

Plots:







**5. Create a Probability Plot of the HSDegree variable.**

See R Code and Output Below

**6. 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.**

It is not normally distributed, as observed values don’t fall on a straight line, since in a normal distribution probability plot, the observed values in the data set would match exactly with what you would expect in a normally distributed data set. This was summarized on page 184 of the textbook by Field (2012).

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

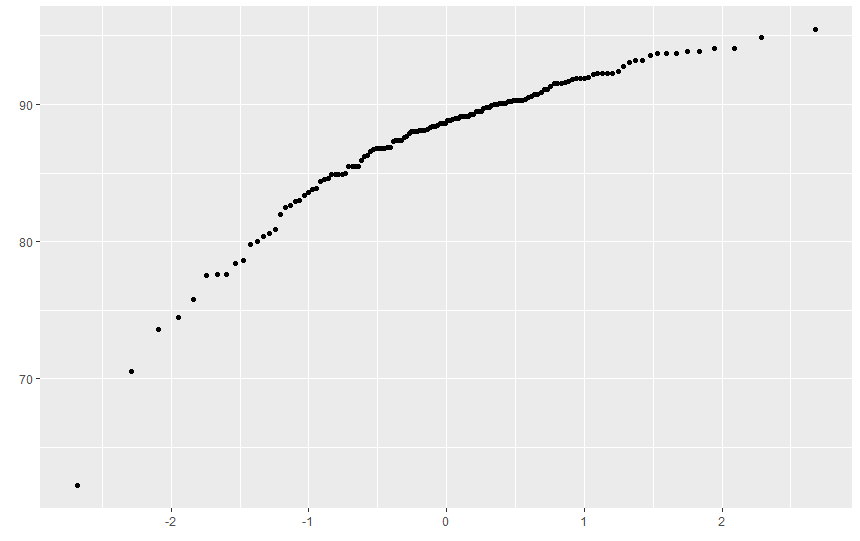
Besides the information I already knew from the histogram, the distribution of the data is left-skewed, with negative skewness, because you can see the curve of the data becomes more pronounced the further to the right of the curve.

**R Code and Output**

# Calculating Probability Curve

qplot(sample = acs$HSDegree, stat = "qq")

Probability Plot Output:



**7. 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.**

R Code:

options(scipen=100)

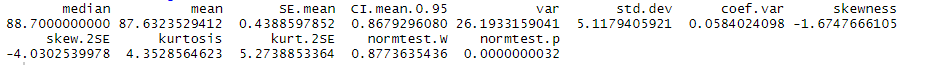
options(digits=2)

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

#zscore calc (not used but saved for later)

scale(acs$HSDegree, center = TRUE, scale = TRUE)

Output:



**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?**

Skewness is a measure of symmetry of the data, and the data has a skewness is -1.67, so it is skewed to the left or negatively skewed. The mean (87.6) is lower than the median (88.7), which is also a sign of left skewness.

Kurtosis is the sharpness of the peak of a frequency distribution curve, and with a kurtosis of 4.35 Since it has a K > 3, it indicates a leptokurtic distribution, which is one that is more peaked than a normal distribution with longer tails.

Z-scores are the number of standard deviations from the mean a data point is. So if we looked at the first record in the data set (Jefferson County, Alabama), they had a mean HS Diploma percentage of 89.1 The formula for the z score is taking the value-the mean divided by the standard deviation. So in this case the formula would be 89.1-87.63/5.118, which equals .287. So this tells us this county was .28 standard deviations above the mean.

A larger sample size would impact the explanation, as it affects the standard error of the sample of data, as mentioned in the book. Large sample sizes will lead to small standard errors, so when sample sizes are big, significant values occur from even small deviations from the norm. (Field, 2012).The larger the sample, the lower the p value you would want to use for your criteria, and he mentioned in samples of 200 or more, you should be more interested in looking at the shape of the distribution and value of skew and kurtosis like we have done in the exercises above, rather than looking at their significance values.