ASSIGNMENT 5

Simek, Kathryn

2020-09-20

### Question 1: What are the elements in your data (including the categories and data types)?

#### Answer:

* Id - quantitative - character
* Id2 - quantitative - integer
* Geography - categorical - character
* PopGroupId - quantitative - integer
* POPGROUP.display-label - categorical - character
* RacesReported - quantitative - integer
* HSDegree - quantitative - numeric
* BachDegree - quantitative - numeric

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

## Set the working directory to the root of your DSC 520 directory  
setwd("C:/Users/katie/OneDrive/Documents/GitHub/dsc520")  
  
## Load the ggplot2 package  
library(ggplot2)  
theme\_set(theme\_minimal())  
  
acs\_df <- read.csv("data/acs-14-1yr-s0201.csv")

### str():

str(acs\_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 ...

### nrow():

nrow(acs\_df)

## [1] 136

### ncol():

ncol(acs\_df)

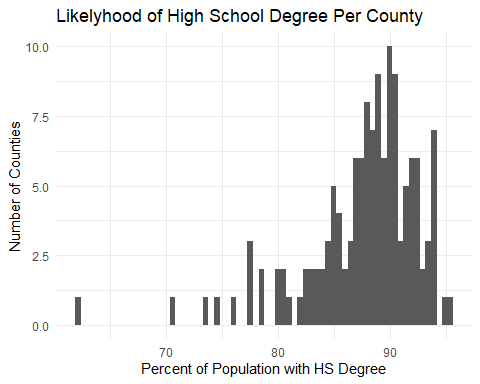
## [1] 8

### Question 3. Create a Histogram of the HSDegree variable.

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

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

ggplot(acs\_df, aes(HSDegree))+geom\_histogram(bins = 68) +  
 ggtitle("Likelyhood of High School Degree Per County") +  
 xlab("Percent of Population with HS Degree") +  
 ylab("Number of Counties")



### Question 4. Answer the following questions based on the Histogram produced:

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

#### Answer: Yes, the data distribution is unimodal.

### b. Is it approximately symmetrical?

#### Answer: No, the data distribution is not symmetrical.

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

#### Answer: Yes, the data is approximately bell-shaped.

### d. Is it approximately normal?

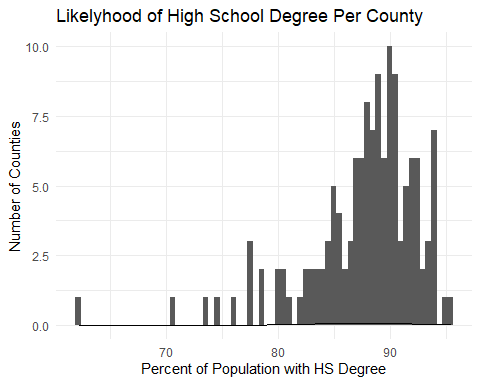
#### Answer: No, it is not approximately normal.

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

#### Answer: The distribution is skewed negatively.

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

ggplot(acs\_df, aes(HSDegree)) + geom\_histogram(bins = 68) +  
 ggtitle("Likelyhood of High School Degree Per County") +  
 xlab("Percent of Population with HS Degree") +  
 ylab("Number of Counties") +   
 stat\_function(fun = dnorm, args = list(mean = mean(acs\_df$HSDegree,   
 na.rm = TRUE),   
 sd = sd(acs\_df$HSDegree,   
 na.rm = TRUE)))

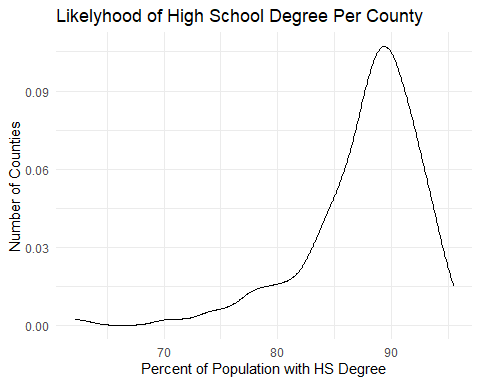


### g. Explain whether a normal distribution can accurately be used as a model for this data.

#### Answer: No, the normal curve is significantly different from the probability plot. The plot is skewed quite a bit.

### Question 5. Create a Probability Plot of the HSDegree variable.

ggplot(acs\_df, aes(HSDegree))+geom\_density() +  
 ggtitle("Likelyhood of High School Degree Per County") +  
 xlab("Percent of Population with HS Degree") +  
 ylab("Number of Counties")



### Question 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.

#### Answer: The plot does not have a symetrical shape, so the distribution is not normal.

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

#### Answer: The distribution is skewed negatively, the tail on the left side of the peak is longer and thinner than the right side of the curve.

### Question 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.

library(psych)

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

describe(acs\_df)

## vars n mean sd median trimmed  
## Id\* 1 136 68.50 39.40 68.5 68.50  
## Id2 2 136 26833.13 15429.11 26112.0 26542.96  
## Geography\* 3 136 68.50 39.40 68.5 68.50  
## PopGroupID 4 136 1.00 0.00 1.0 1.00  
## POPGROUP.display.label\* 5 136 1.00 0.00 1.0 1.00  
## RacesReported 6 136 1144400.99 1090507.89 832707.5 927231.74  
## HSDegree 7 136 87.63 5.12 88.7 88.28  
## BachDegree 8 136 35.46 9.51 34.1 35.23  
## mad min max range skew kurtosis  
## Id\* 50.41 1.0 136.0 135.0 0.00 -1.23  
## Id2 20778.64 1073.0 55079.0 54006.0 0.05 -1.34  
## Geography\* 50.41 1.0 136.0 135.0 0.00 -1.23  
## PopGroupID 0.00 1.0 1.0 0.0 NaN NaN  
## POPGROUP.display.label\* 0.00 1.0 1.0 0.0 NaN NaN  
## RacesReported 314163.68 500292.0 10116705.0 9616413.0 4.98 33.50  
## HSDegree 3.78 62.2 95.5 33.3 -1.67 4.35  
## BachDegree 8.23 15.4 60.3 44.9 0.33 -0.28  
## se  
## Id\* 3.38  
## Id2 1323.04  
## Geography\* 3.38  
## PopGroupID 0.00  
## POPGROUP.display.label\* 0.00  
## RacesReported 93510.28  
## HSDegree 0.44  
## BachDegree 0.82

library(moments)  
skewness(acs\_df$HSDegree)

## [1] -1.69341

library(moments)  
kurtosis(acs\_df$HSDegree)

## [1] 7.462191

hsdegree\_z <-(acs\_df$HSDegree - mean(acs\_df$HSDegree)) / sd(acs\_df$HSDegree)  
hsdegree\_z

## [1] 0.286765161 -0.162634350 0.071834960 -0.143095241 0.228147834  
## [6] -2.741796762 -2.565944779 -1.979771504 -0.592494752 -1.374059119  
## [11] -0.162634350 -1.764841303 -0.201712568 0.091374069 -1.960232394  
## [16] 0.091374069 -0.045399695 -0.006321476 -1.803919521 -0.787885844  
## [21] 0.833860218 -0.416642769 1.009712201 1.263720620 0.423538925  
## [26] 0.325843380 0.364921598 0.482156253 0.501695362 0.775242891  
## [31] 0.149991397 0.267226052 -0.064938804 -0.260329896 -1.315441791  
## [36] 0.052295851 0.013217633 0.482156253 -0.533877424 0.247686943  
## [41] 0.521234471 0.149991397 0.716625563 0.071834960 0.814321109  
## [46] -0.416642769 0.912016655 -0.924659608 0.521234471 0.599390908  
## [51] -0.514338315 1.537268149 0.228147834 0.169530506 0.833860218  
## [56] 0.540773581 0.638469126 -0.416642769 -0.631572970 -1.002816045  
## [61] 0.286765161 0.912016655 1.263720620 0.892477546 -0.729268516  
## [66] 0.482156253 0.286765161 0.325843380 1.166025074 -0.533877424  
## [71] 1.087868638 0.443078035 0.462617144 1.087868638 0.110913179  
## [76] -0.612033861 0.755703781 0.130452288 -0.416642769 -0.826964062  
## [81] 0.286765161 1.068329528 0.794782000 -0.748807625 -0.279869005  
## [86] 0.071834960 -3.347509146 0.579851799 -1.491293774 0.521234471  
## [91] 0.599390908 -0.162634350 -1.413137337 0.423538925 -0.045399695  
## [96] 0.267226052 0.364921598 0.931555764 0.091374069 0.462617144  
## [101] 0.560312690 0.403999816 0.677547345 -0.162634350 0.189069615  
## [106] 0.677547345 0.501695362 1.224642402 1.224642402 0.912016655  
## [111] 0.755703781 -0.533877424 1.185564183 -0.983276935 -1.100511591  
## [116] -0.182173459 -0.045399695 -0.905120499 1.185564183 -1.960232394  
## [121] 0.833860218 -2.311936360 0.189069615 -1.530371992 -4.969255208  
## [126] -0.338486333 -0.533877424 0.189069615 0.364921598 1.185564183  
## [131] 0.755703781 0.912016655 0.521234471 0.853399327 1.420033494  
## [136] -0.143095241

### Question 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?

#### Answer: The skew of -1.69 is greater than a normal skew range of -1 to 1 so the skew is not normal, it is significantly negatively skewed. The kurtosis of 7.46 is greater than the -3 to 3 range so the kurtosis is not normal, it is significantly more positive, or leptokurtic. With the z-scores, we can see that most values are pretty significant, with a large range in values, which describes why we have heavy tails and many outliers in our distribution.