520-Wk3-AmericanCommunitySurvey

Brandon Sams

12/15/2019

Assignment Overview

This is your second exercise with real data. This time, instead of a bank of test scores, we will use the 2014 American Community Survey. These data are maintained by the US Census Bureau and are designed to show how communities are changing.

Through asking questions of a sample of the population, it produces national data on more than 35 categories of information, such as education, income, housing, and employment.

First, lets make a histogram!

Question 1

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

```r
library(readr)
acs2014 <- read_csv("http://content.bellevue.edu/cst/dsc/520/id/resources/acs-14-1yr-s0201.csv")
```

```
## `curl` package not installed, falling back to using `url()`
```

```
## Parsed with column specification:
## cols(
##   Id = col_character(),
##   Id2 = col_double(),
##   Geography = col_character(),
##   PopGroupID = col_double(),
##   `POPGROUP.display-label` = col_character(),
##   RacesReported = col_double(),
##   HSDegree = col_double(),
##   BachDegree = col_double()
## )
```

Question 2

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

```r
str(acs2014)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 136 obs. of  8 variables:
##  $ Id                    : chr  "0500000US01073" "0500000US04013" "0500000US04019" "0500000US06001"
##  $ Id2                   : num  1073 4013 4019 6001 6013 ...
##  $ Geography             : chr  "Jefferson County, Alabama" "Maricopa County, Arizona" "Pima County,
##  $ PopGroupID            : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ POPGROUP.display-label: chr  "Total population" "Total population" "Total population" "Total popul
##  $ RacesReported         : num  660793 4087191 1004516 1610921 1111339 ...
##  $ 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 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_character(),
##   ..   Id2 = col_double(),
##   ..   Geography = col_character(),
##   ..   PopGroupID = col_double(),
##   ..   `POPGROUP.display-label` = col_character(),
##   ..   RacesReported = col_double(),
##   ..   HSDegree = col_double(),
##   ..   BachDegree = col_double()
##   .. )
```

```r
nrow(acs2014)
```

```
## [1] 136
```

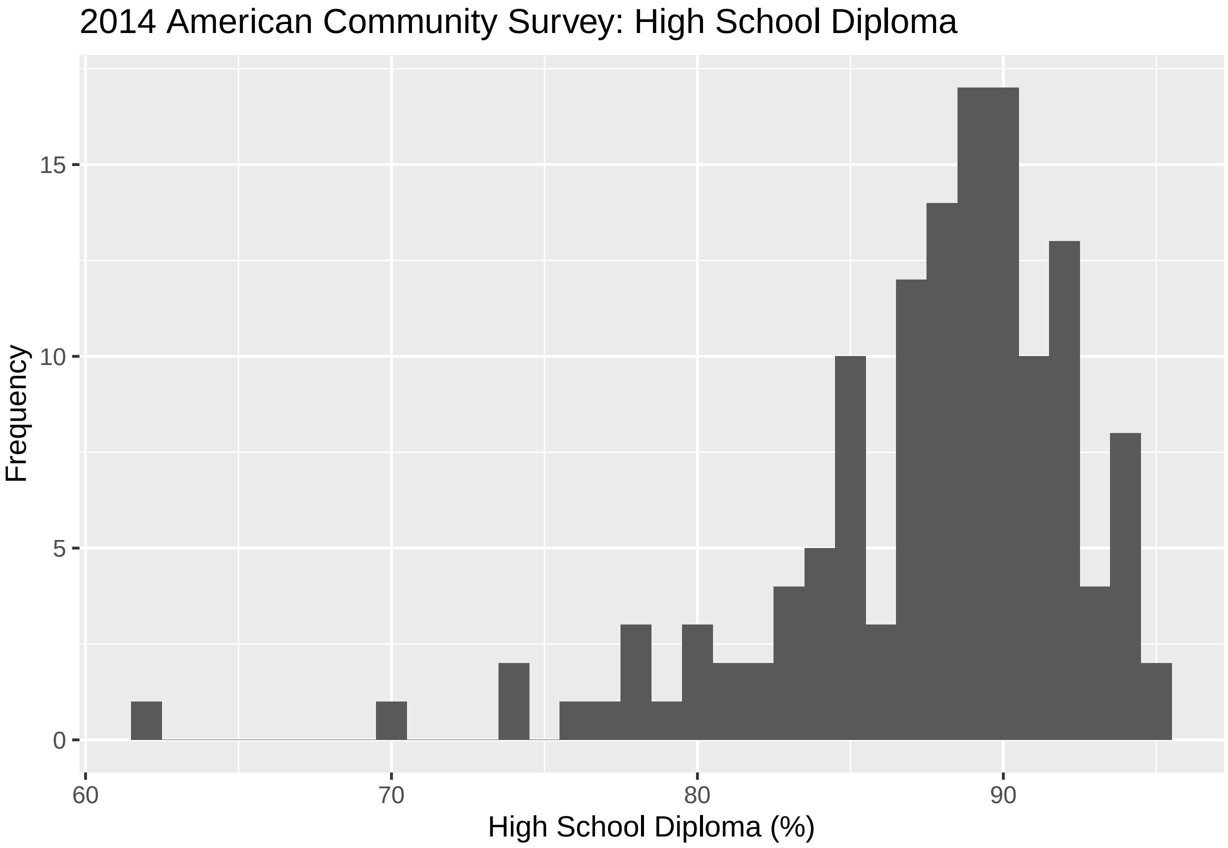
```r
ncol(acs2014)
```

```
## [1] 8
```

Question 3

Create a Histogram of the HSDegree variable using the ggplot2 package. Set a bin size for the Histogram. Include a Title and appropriate X/Y axis labels on your Histogram Plot.

```r
library(ggplot2)
ggplot(acs2014,aes(x = HSDegree)) + geom_histogram(binwidth = 1) + xlab("High School Diploma (%)") + yla
```



Now that we have a Histogram

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

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

Based on the histogram, I would say that the data distribution is, in fact, unimodal. There is one peak at 90%.

b. Is it approximately symmetrical?

Based on the histogram, I would say that the data distribution is not symmetrical. The left side of the histogram has a far longer tail than the right side.

c. Is it approximately bell-shaped?

Based on the histrogram, I would say that the data distribution is bell shaped.

d. Is it approximately normal?

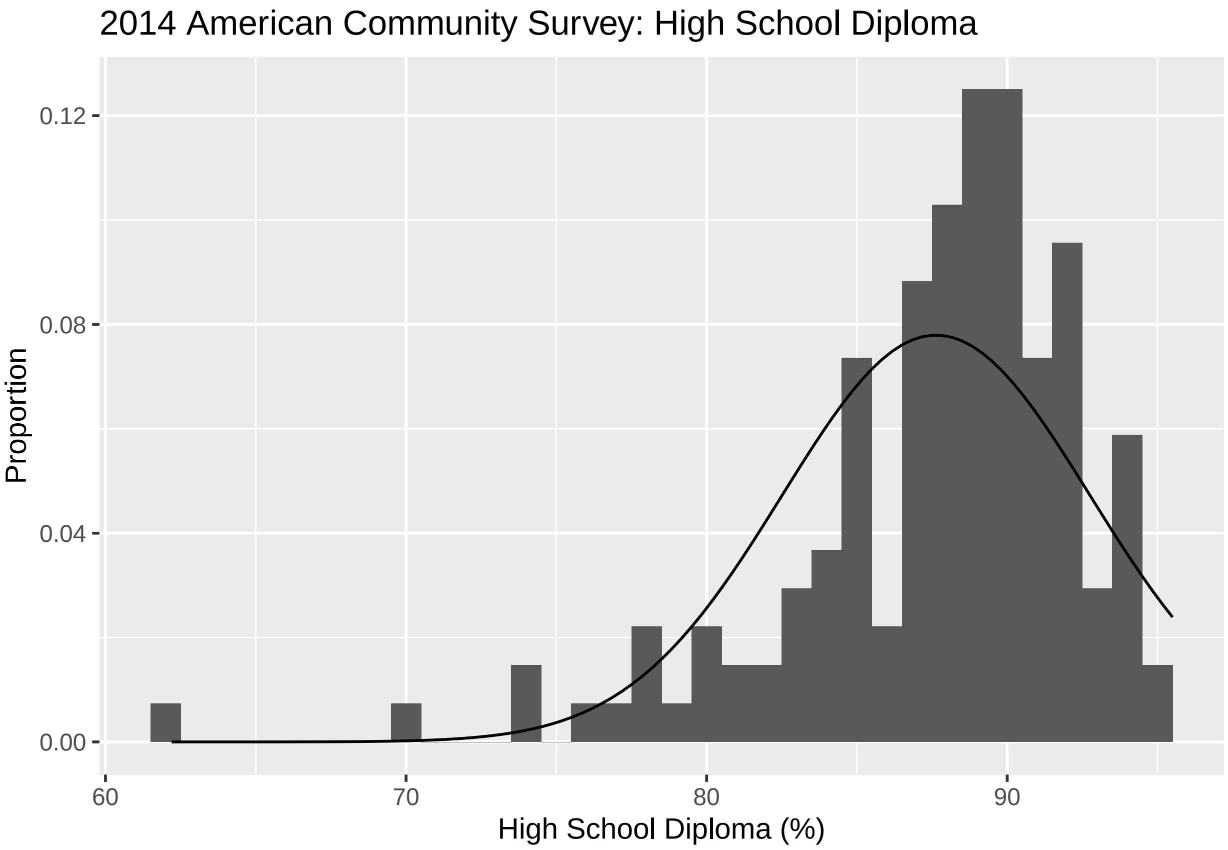
The distribution could be approximated as a normal distribution, but would likely be better approximated as a different type of distribution, such as a gamma distribution.

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

The distribution is skewed to the right

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

```
ggplot(acs2014,aes(x = HSDegree)) + geom_histogram(binwidth = 1,aes(y=..density..))  + xlab("High Schoo
```



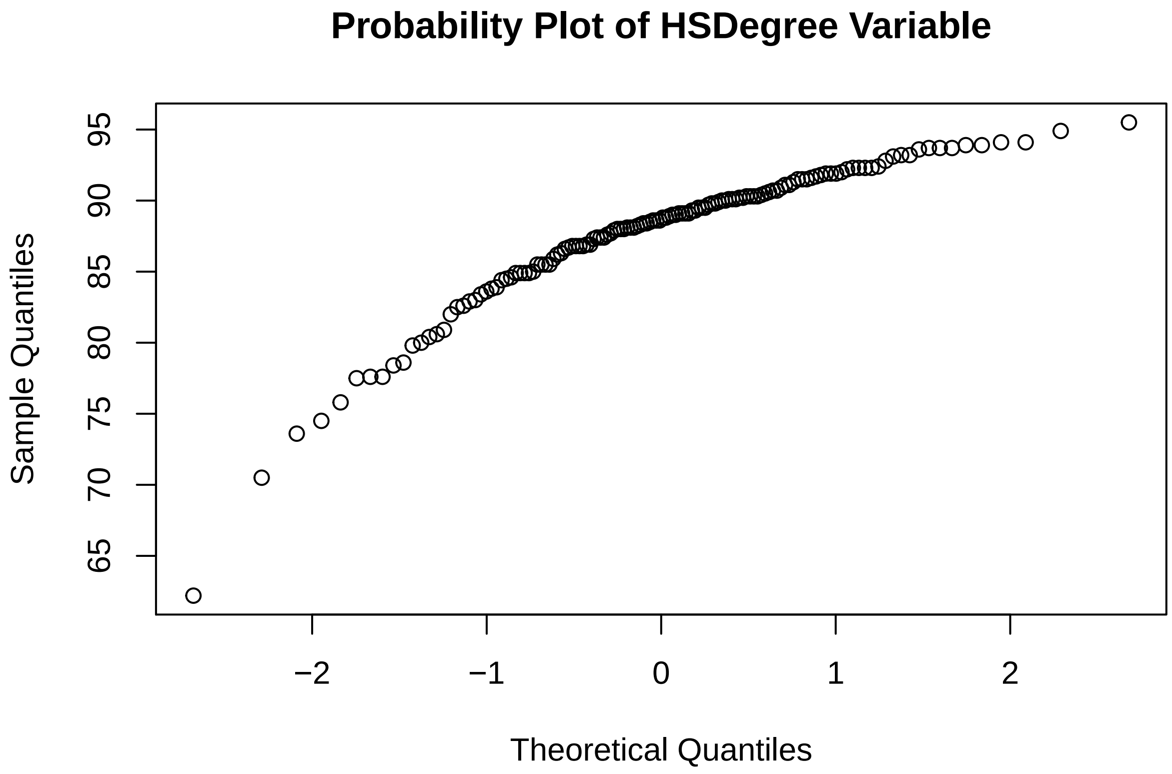
g. Explain whether a normal distribution can accurately be used as a model for this data.

It does not appear to be an accurage model for this data. The height of the peak is far too low on the normal curve, and the mean is too far off from what was observed.

Probabiliy Plot

1. Create a Probability Plot of the HSDegree variable.

```
qqnorm(acs2014$HSDegree,main = "Probability Plot of HSDegree Variable")
```



2. 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.

I do not think this distribution is approximately normal. All the probability plots I found online showed the diagonal line formed by the points in the probability plot to be very straight.

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

The distribution is skewed to the right, as that is where the diagonal line starts to level off. More values are present above the mean than expected.

1. 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
library(pastecs)
stat.desc(acs2014$HSDegree)
```

```
##      nbr.val     nbr.null       nbr.na          min          max        range 
## 1.360000e+02 0.000000e+00 0.000000e+00 6.220000e+01 9.550000e+01 3.330000e+01 
##          sum       median         mean      SE.mean CI.mean.0.95          var 
## 1.191800e+04 8.870000e+01 8.763235e+01 4.388598e-01 8.679296e-01 2.619332e+01 
##      std.dev     coef.var 
## 5.117941e+00 5.840241e-02
```

2. 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 result produced skewed to the high end of the range. With a mean centered at about 87% of people having a High School Diploma, this left very little room for a long tail to exist above the mean. This is different from the values below 87%, of which there is plenty of room for a long tail. This would be unchanged if additional data was collected.

As for the Kurtosis, the peak itslef was not very sharp. The standard deviation for this distribution was more than 5 percent, which appears to be a large value. If more data was collected, I do not expect the kurtosis value to change.

The z-score is a measure of how far, in standard deviations, is a point from the mean. Due to the nature of the z-score being a measure for a single value, it is difficult to say why the z-score is the way it is. I would wager a guess that it likely had lot to do with the distribution being skewed, and the case where the average z score on the left being different than an average z-score for a value to the right of the mean. If more data was collected, I do not expect the shape or location to change. Therefore, the z scores would likely be largely unchanged as well.