Assignment_2.1_BreitbachScott

Scott Breitbach

6/11/2020

2.1 Assignment: 2014 American Community Survey

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

```
class(acs2014$Id)
```

```
## [1] "character"
```

```
class(acs2014$Id2)
```

```
## [1] "integer"
```

```
class(acs2014$Geography)
```

```
## [1] "character"
```

```
class(acs2014$PopGroupID)
```

```
## [1] "integer"
```

```
class(acs2014$POPGROUP.display.label)
```

```
## [1] "character"
```

```
class(acs2014$RacesReported)
```

```
## [1] "integer"
```

```
class(acs2014$HSDegree)
```

```
## [1] "numeric"
```

```
typeof(acs2014$HSDegree)
```

```
## [1] "double"
```

```r
class(acs2014$BachDegree)
```

```
## [1] "numeric"
```

```r
typeof(acs2014$BachDegree)
```

```
## [1] "double"
```

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

```r
str(acs2014)
```

```
## '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,
##  $ PopGroupID            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ POPGROUP.display.label: chr  "Total population" "Total population" "Total population" "Total popul
##  $ RacesReported         : int  660793 4087191 1004516 1610921 1111339 965974 874589 10116705 314551
##  $ 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 ...
```

```r
nrow(acs2014)
```

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

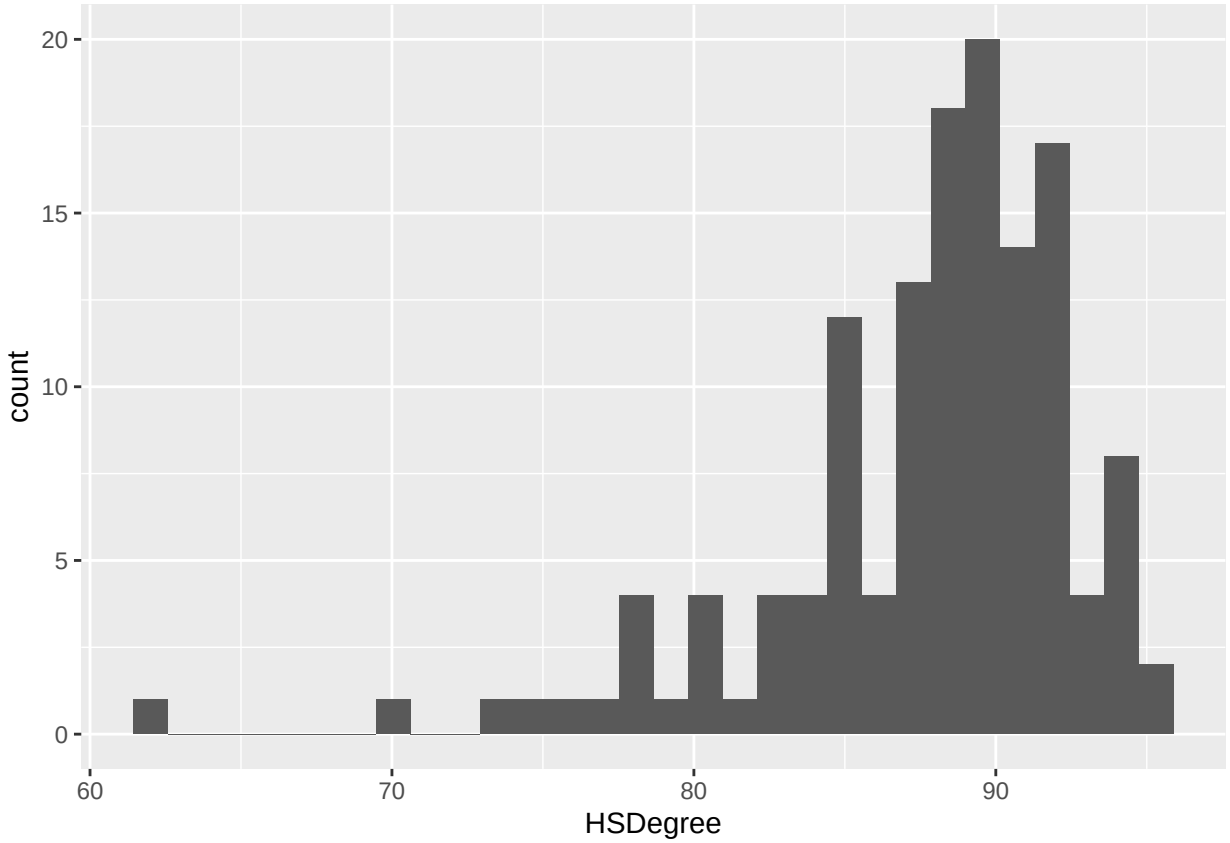
```r
ncol(acs2014)
```

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

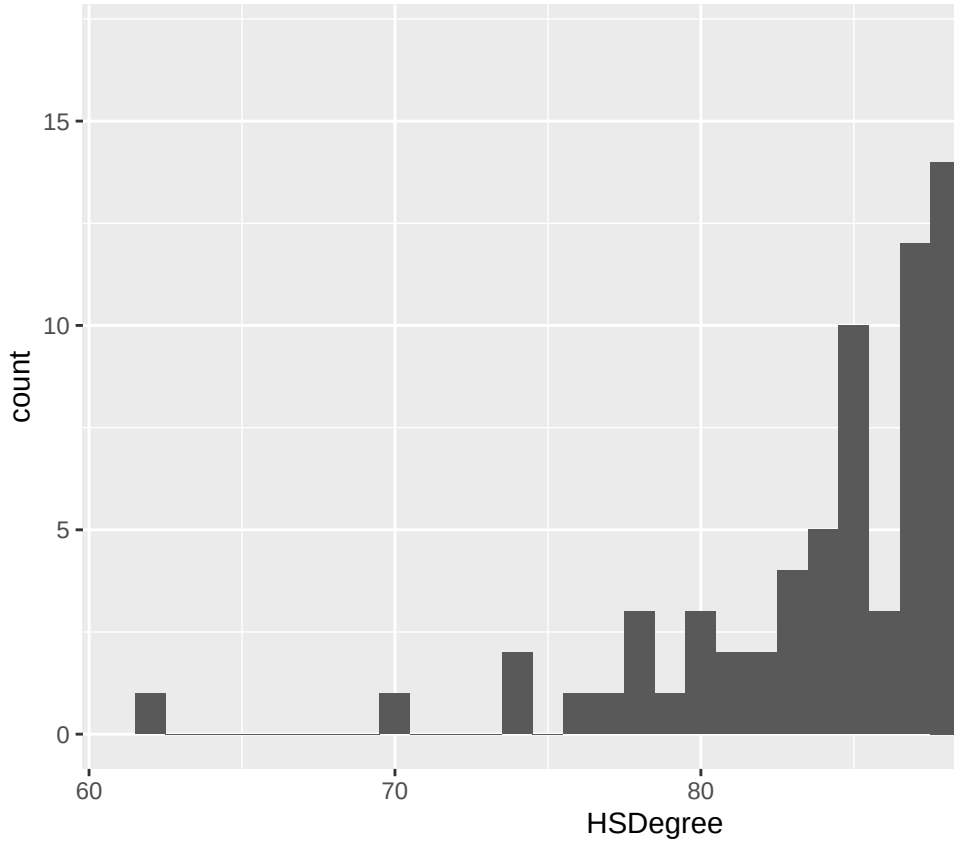
3. Create a Histogram of the HSDegree variable using the ggplot2 package.

```r
acs14Histogram <- ggplot(acs2014, aes(HSDegree))
acs14Histogram + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



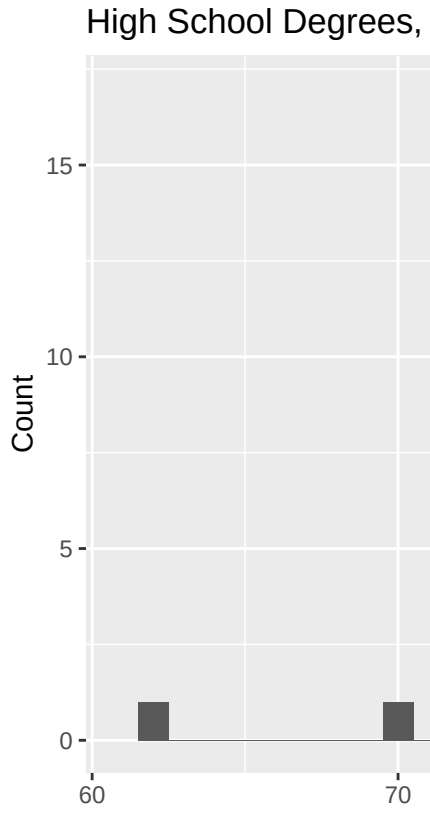

```
acs14Histogram + geom_histogram(binwidth = 1)
```



a. Set a bin size for the Histogram.

```
acs14Histogram <- acs14Histogram + geom_histogram(binwidth = 1)
```

```
acs14Histogram + labs(x = "% with High School Degree", y = "Count",
                      title = "High School Degrees, 2014")
```



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

```
acs14Histogram <- acs14Histogram +
  labs(x = "% with High School Degree", y = "Count",
       title = "High School Degrees, 2014")
```

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

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

The data distribution appears to be unimodal, somewhere around 89-90%.

b. Is it approximately symmetrical?

The data appears to be negatively skewed, not symmetrical.

c. Is it approximately bell-shaped?

Though skewed, the data does appear approximately bell-shaped.

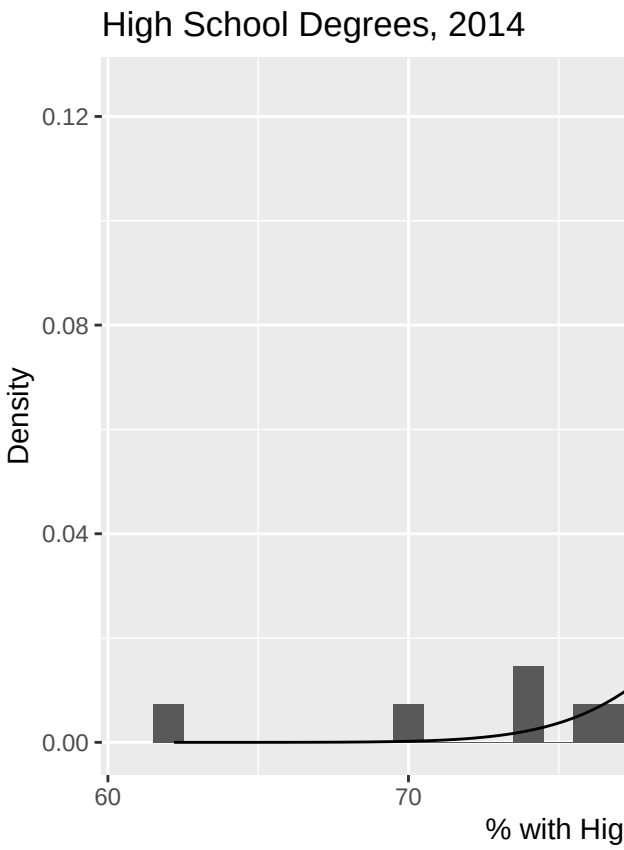
d. Is it approximately normal?

I would say the skew deviates the curve from a normal distribution, but the text says that 'the sampling distribution will tend to be normal regardless of the population distribution in samples of 30 or more. Based on this criteria, I would say it is a normal distribution because we have 137 samples.

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

The data is clustered toward the higher end and so is negatively skewed.

```
acs14Density <- ggplot(acs2014, aes(HSDegree)) +
  geom_histogram(aes(y = ..density..), binwidth = 1) +
  labs(x = "% with High School Degree", y = "Density",
       title = "High School Degrees, 2014") +
  stat_function(fun = dnorm, args = list(mean = mean(acs2014$HSDegree),
                                         sd(acs2014$HSDegree)))
acs14Density
```



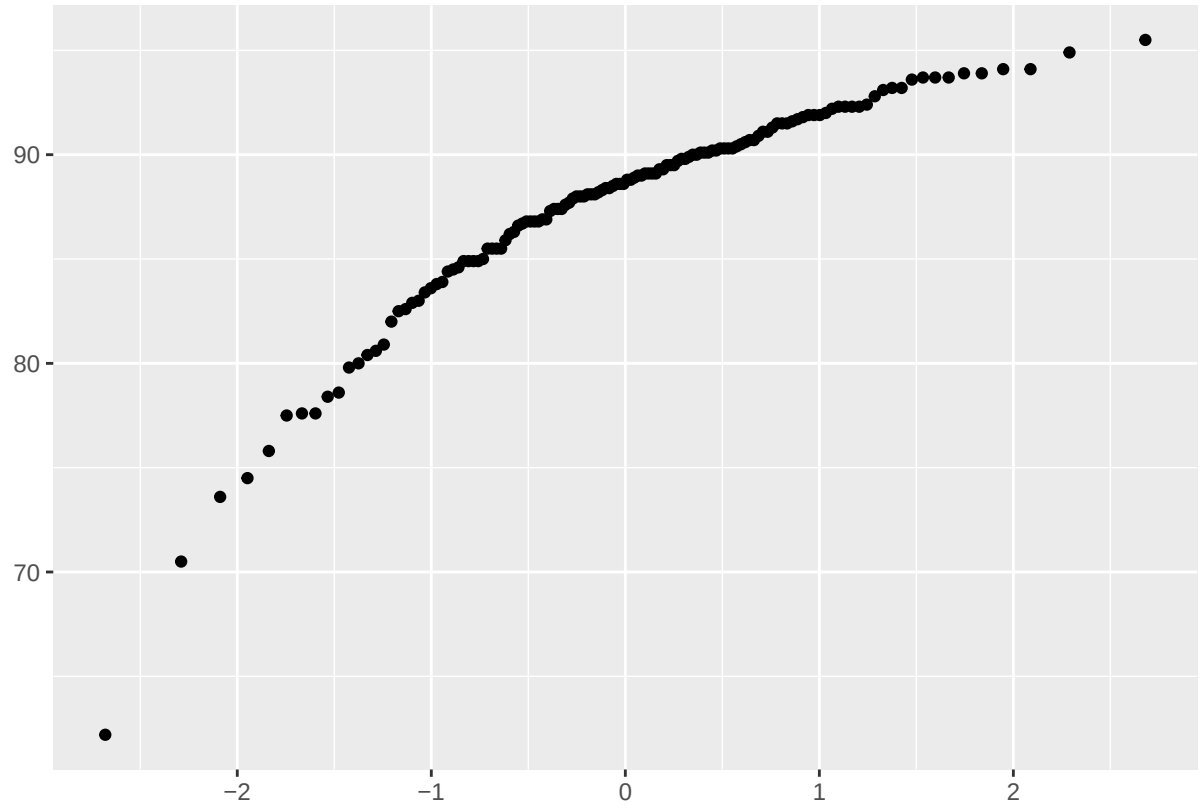
f. Include a normal curve to the Histogram that you plotted.

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

The normal curve does not appear to fit the data very well; it appears non-symmetrical and to be negatively skewed. Additionally, it appears not to fit the normal curve very well as the data is too pointy (positive kurtosis).

5. Create a Probability Plot of the HSDegree variable.

```
qq.acs14.HSDegree <- qplot(sample = acs2014$HSDegree)
qq.acs14.HSDegree
```



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.

This distribution does not appear normal. If the data matched a normal distribution, we would expect the points to form a straght line. Devations from this line indicate deviations from normality.

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

The data deviates from normal creating an upward curve, which indicates the data is clustered toward the high end of the scale, or negatively skewed.

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.

```
round(stat.desc(acs2014$HSDegree, basic = FALSE, norm = TRUE), digits = 3)
```

```
##       median         mean      SE.mean CI.mean.0.95          var      std.dev
##       88.700       87.632        0.439        0.868       26.193        5.118
##     coef.var     skewness     skew.2SE     kurtosis     kurt.2SE   normtest.W
##        0.058       -1.675       -4.030        4.353        5.274        0.877
##   normtest.p
##        0.000
```

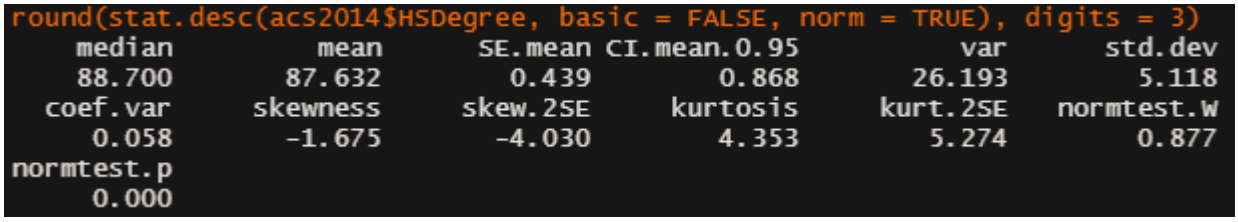


Figure 1: screen capture

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?

In a normal distribution, we would expect the values for skew and kurtosis to be at or very near zero. In these results, there is a negative skewness (-1.675) indicating the overall scores tend toward the higher end. Kurtosis is positive (4.353) indicating a pointy and heavily-tailed distribution. The z-scores for each are related to the skew.2SE and kurt.2SE values. For smaller sample sizes, one might look at these and say that they are significant if they are >2, however we have well over 200 samples in our data set, which will make these appear more significant than they are. Because of this, we should not use z-scores as a metric and instead rely more on the skew and kurtosis values as well as visually inspecting the histogram.