Assignment3.2_BrownLincoln.R

x

2021-06-27

```
# Assignment3.2
# Lincoln Brown
# DSC520-T301
# Dr. Bushart

# Load the data
acs <- read.csv("/media/x/disk/School/DSC520/Wk3/acs-14-1yr-s0201.csv")

# Import ggplot2
library(ggplot2)
library(pastecs)
library(moments)
# Set theme
theme_set(theme_dark())

# 1. What are the elements in your data (including the categories and data types)?
# The columns: Id, Id2, Geography, and POPGROUP.display.label are categorical.
# The column PopGroupID is either binary or categorical.
# The columns RacesReported, HSDegree, and BachDegree are numeric.

# Column Name | Data Type
#_________________________
# Id               chr
# Id2              int
# Geography        chr
# PopGroupID       int
# POPGROUP         chr
# RacesReported    int
# HSDegree         num
# BachDegree       num

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

```
## '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 3145515
##  $ 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 ...
```

```
colnames(acs)
```

```
## [1] "Id"                     "Id2"                    "Geography"
## [4] "PopGroupID"             "POPGROUP.display.label" "RacesReported"
## [7] "HSDegree"               "BachDegree"
```

```
rownames(acs)
```

```
##   [1] "1"   "2"   "3"   "4"   "5"   "6"   "7"   "8"   "9"   "10"  "11"  "12"
##  [13] "13"  "14"  "15"  "16"  "17"  "18"  "19"  "20"  "21"  "22"  "23"  "24"
##  [25] "25"  "26"  "27"  "28"  "29"  "30"  "31"  "32"  "33"  "34"  "35"  "36"
##  [37] "37"  "38"  "39"  "40"  "41"  "42"  "43"  "44"  "45"  "46"  "47"  "48"
##  [49] "49"  "50"  "51"  "52"  "53"  "54"  "55"  "56"  "57"  "58"  "59"  "60"
##  [61] "61"  "62"  "63"  "64"  "65"  "66"  "67"  "68"  "69"  "70"  "71"  "72"
##  [73] "73"  "74"  "75"  "76"  "77"  "78"  "79"  "80"  "81"  "82"  "83"  "84"
##  [85] "85"  "86"  "87"  "88"  "89"  "90"  "91"  "92"  "93"  "94"  "95"  "96"
##  [97] "97"  "98"  "99"  "100" "101" "102" "103" "104" "105" "106" "107" "108"
## [109] "109" "110" "111" "112" "113" "114" "115" "116" "117" "118" "119" "120"
## [121] "121" "122" "123" "124" "125" "126" "127" "128" "129" "130" "131" "132"
## [133] "133" "134" "135" "136"
```

```
dim(acs)
```

```
## [1] 136   8
```

```
acs[c(2,7)]
```

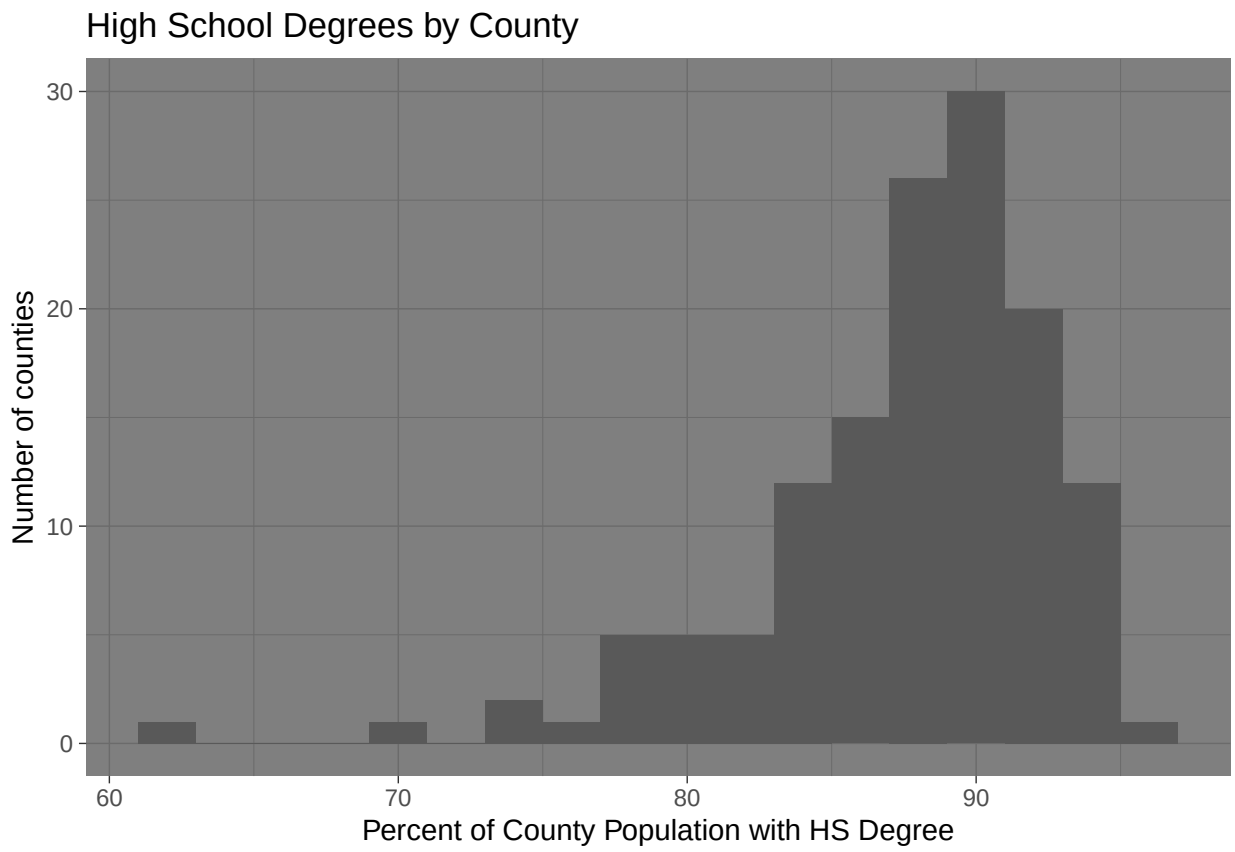
```
##       Id2 HSDegree
## 1    1073     89.1
## 2    4013     86.8
## 3    4019     88.0
## 4    6001     86.9
## 5    6013     88.8
## 6    6019     73.6
## 7    6029     74.5
## 8    6037     77.5
## 9    6059     84.6
## 10   6065     80.6
## 11   6067     86.8
## 12   6071     78.6
## 13   6073     86.6
## 14   6075     88.1
## 15   6077     77.6
## 16   6081     88.1
## 17   6085     87.4
## 18   6097     87.6
## 19   6099     78.4
## 20   6111     83.6
## 21   8005     91.9
## 22   8031     85.5
## 23   8041     92.8
## 24   8059     94.1
## 25   9001     89.8
## 26   9003     89.3
## 27   9009     89.5
## 28  10003     90.1
```

```
## 29  11001     90.2
## 30  12009     91.6
## 31  12011     88.4
## 32  12031     89.0
## 33  12057     87.3
## 34  12071     86.3
## 35  12086     80.9
## 36  12095     87.9
## 37  12099     87.7
## 38  12103     90.1
## 39  12105     84.9
## 40  12127     88.9
## 41  13067     90.3
## 42  13089     88.4
## 43  13121     91.3
## 44  13135     88.0
## 45  15003     91.8
## 46  17031     85.5
## 47  17043     92.3
## 48  17089     82.9
## 49  17097     90.3
## 50  17197     90.7
## 51  18097     85.0
## 52  20091     95.5
## 53  20173     88.8
## 54  21111     88.5
## 55  24003     91.9
## 56  24005     90.4
## 57  24031     90.9
## 58  24033     85.5
## 59  24510     84.4
## 60  25005     82.5
## 61  25009     89.1
## 62  25017     92.3
## 63  25021     94.1
## 64  25023     92.2
## 65  25025     83.9
## 66  25027     90.1
## 67  26081     89.1
## 68  26099     89.3
## 69  26125     93.6
## 70  26163     84.9
## 71  27053     93.2
## 72  27123     89.9
## 73  29095     90.0
## 74  29189     93.2
## 75  31055     88.2
## 76  32003     84.5
## 77  34003     91.5
## 78  34007     88.3
## 79  34013     85.5
## 80  34017     83.4
## 81  34023     89.1
## 82  34025     93.1
```

```
## 83  34029     91.7
## 84  34031     83.8
## 85  34039     86.2
## 86  35001     88.0
## 87  36005     70.5
## 88  36029     90.6
## 89  36047     80.0
## 90  36055     90.3
## 91  36059     90.7
## 92  36061     86.8
## 93  36081     80.4
## 94  36103     89.8
## 95  36119     87.4
## 96  37081     89.0
## 97  37119     89.5
## 98  37183     92.4
## 99  39035     88.1
## 100 39049     90.0
## 101 39061     90.5
## 102 39113     89.7
## 103 39153     91.1
## 104 40109     86.8
## 105 40143     88.6
## 106 41051     91.1
## 107 41067     90.2
## 108 42003     93.9
## 109 42017     93.9
## 110 42029     92.3
## 111 42045     91.5
## 112 42071     84.9
## 113 42091     93.7
## 114 42101     82.6
## 115 44007     82.0
## 116 47037     86.7
## 117 47157     87.4
## 118 48029     83.0
## 119 48085     93.7
## 120 48113     77.6
## 121 48121     91.9
## 122 48141     75.8
## 123 48157     88.6
## 124 48201     79.8
## 125 48215     62.2
## 126 48339     85.9
## 127 48439     84.9
## 128 48453     88.6
## 129 49035     89.5
## 130 49049     93.7
## 131 51059     91.5
## 132 53033     92.3
## 133 53053     90.3
## 134 53061     92.0
## 135 55025     94.9
## 136 55079     86.9
```

```r
#3. Create a Histogram of the HSDegree variable using ggplot2
# Set a bin size for the histogram
# Include a Title and appropriate X/Y axis labels on the plot.

g <- ggplot(data = acs, aes(HSDegree))
g + geom_histogram(binwidth=2) + ggtitle("High School Degrees by County") +
                                          xlab("Percent of County Population with HS Degree") +
                                          ylab("Number of counties")
```




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

# 4a. Based on what you see in this histogram, is the data distribution unimodal?
(q4a <- c("Yes this distribution is unimodal"))
```

```
## [1] "Yes this distribution is unimodal"
```

```r
# 4b.  Is it approximately symmetrical?
(q4b <- c("No, this distribution is negatively skewed."))
```

```
## [1] "No, this distribution is negatively skewed."
```

```r
# 4c.  Is it approximately bell-shaped?
(q4c <- c("Yes it is approximately bell-shaped."))
```

```
## [1] "Yes it is approximately bell-shaped."
```

```r
# 4d. Is it approximately normal?
(q4d <- c("No, it is not normal."))
```

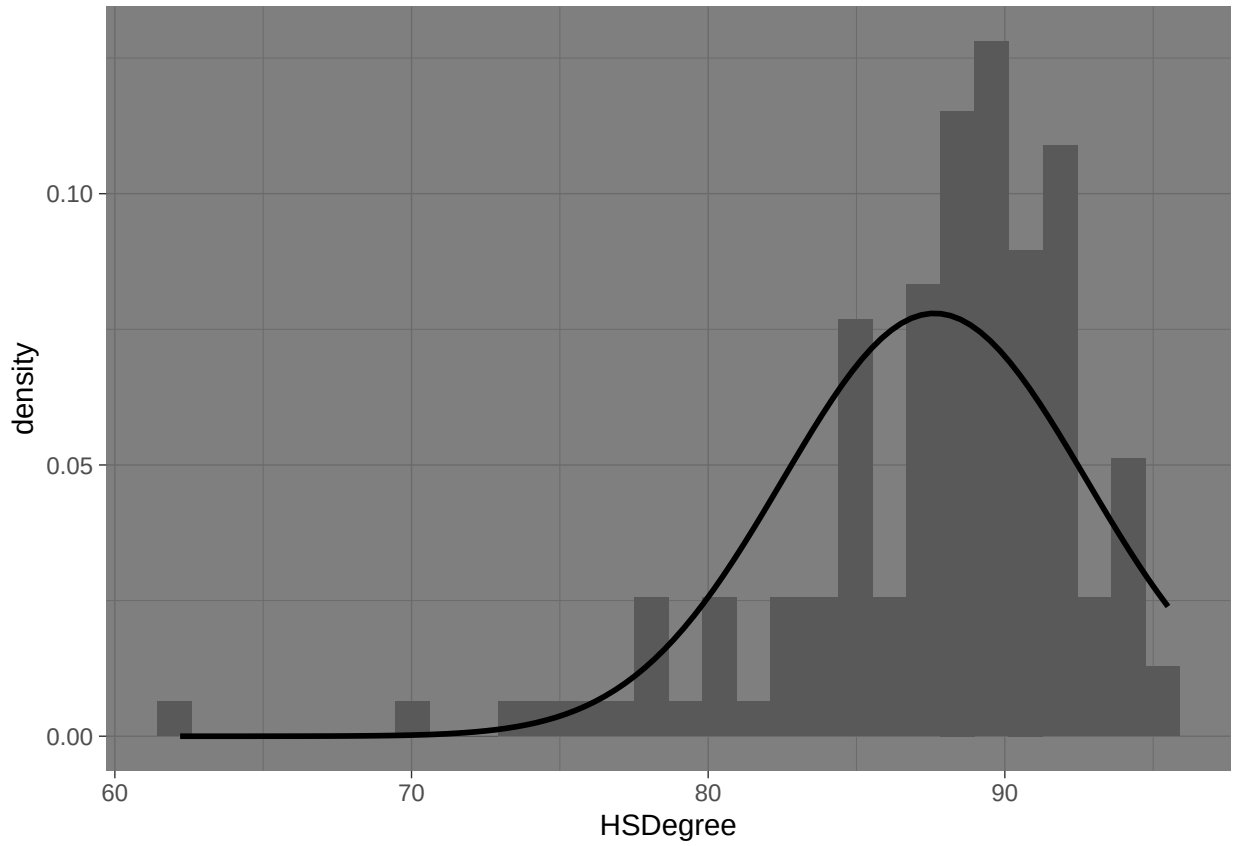
```
## [1] "No, it is not normal."
```

```
# 4e. If not normal, is the distribution skewed? If so, in which direction?
(q4e <- c("The distribution is skewed left or negatively skewed. "))
```

```
## [1] "The distribution is skewed left or negatively skewed. "
```

```
# 4f. Include a normal curve to the Histogram that you plotted.
high_school <- g + geom_histogram(aes(y = ..density..))
high_school +
  stat_function(fun = dnorm, args =
                  list(mean = mean(acs$HSDegree, na.rm = TRUE),
                       sd = sd(acs$HSDegree, na.rm = TRUE)),
                colour  = "black", size = 1)
```

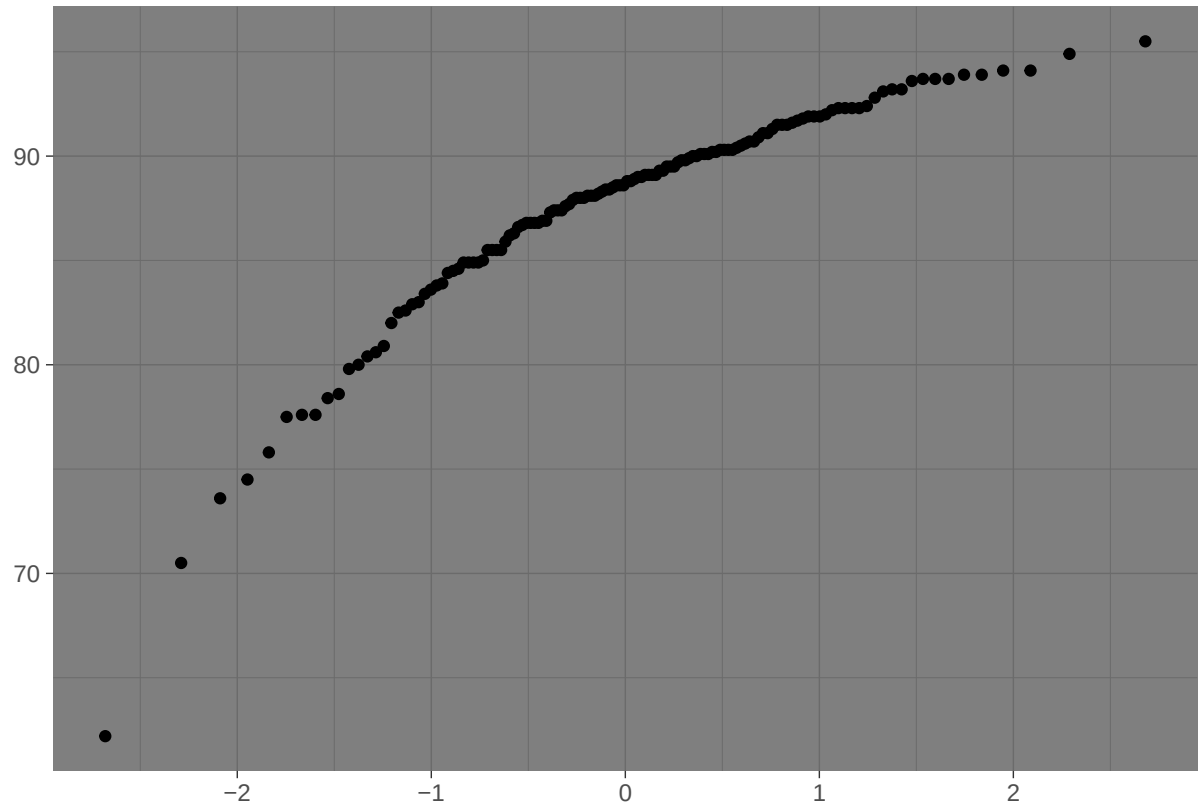
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




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

```
## Warning: `stat` is deprecated
```

```
qqplot.HSDegree
```

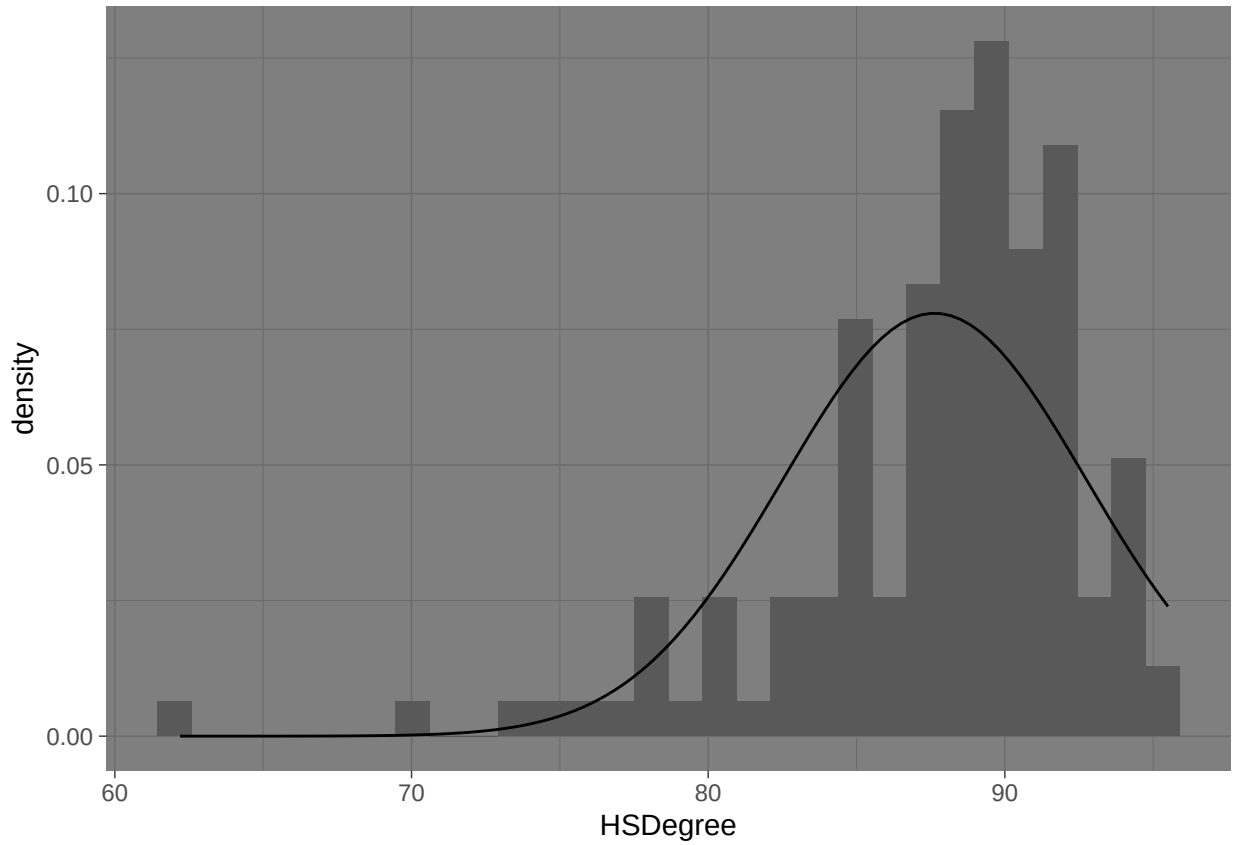



```
# 4g. Explain whether a normal distribution can accurately
#be used as a model for this data.
(q4g <- c("A normal distribution cannot be used as a model
          because it is negatively skewed."))
```

```
## [1] "A normal distribution cannot be used as a model \n          because it is negatively skewed."
```

```
# 5. Create a Probability Plot of the HSDegree variable.
g + geom_histogram(aes(y=..density..)) +
  stat_function(fun = dnorm,
                args = list(mean = mean(acs$HSDegree), sd = sd(acs$HSDegree)))
```

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




```
# 6. Answer the following questions based on the Probability Plot:
# 6a  Based on what you see in this probability plot,
#is the distribution approximately normal? Explain how you know.
(q6a <- "The negatively skewed central tendencies of this distribution
  indicate that it is not approximately normal. ")
```

```
## [1] "The negatively skewed central tendencies of this distribution \n  indicate that it is not appro
```

```
# 6b If not normal, is the distribution skewed?
#If so, in which direction? Explain how you know.
(q6b <- "The distribution is negatively skewed, or skewed to the left.
  The tail extends further on the left side, indicating a negative skew.")
```

```
## [1] "The distribution is negatively skewed, or skewed to the left. \n  The tail extends further on t
```

```
# 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.
desc <- stat.desc(acs["HSDegree"], norm = TRUE)
kurtosis(acs["HSDegree"])
```

```
## HSDegree
## 7.462191
```

```
skewness(acs["HSDegree"])
```

```
## HSDegree
## -1.69341
```

```r
# 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?
zskew <- (desc["skewness",] - desc["SE.mean",]) / desc["std.dev",]
zkurt <- (desc["kurtosis",] - desc["SE.mean",]) / desc["std.dev",]

zskew1 <- (desc["skewness",]- 0) / desc["std.dev",]
zkurt1 <- (desc["kurtosis",]- 0) / desc["std.dev",]

# I am a little confused on which of the above formulas is
#appropriate for calculating the z scores.
# The book indicates that we should use a 0 for the mean of the distribution,
#which makes sense because we are calculating
# for a z-score which have a mean of 0 and a sd of 1.
# Therefore, I believe that zskew1 and zkurt1 are the correct formulas,
#but I would appreciate some guidance on the matter.

#The scores for kurtosis and skew in the descriptive statistics have the values:
(skew <- desc["skewness",])
```

```
## [1] -1.674767
```

```r
(kurt <- desc["kurtosis",])
```

```
## [1] 4.352856
```

```r
(excess_kurtosis <- kurt - 3)
```

```
## [1] 1.352856
```

```r
# The skew of -1.674767 indicates a significant negative skew (> 1)
# The kurtosis of 4.352856 indicates that it has excess kurtosis of 1.352856


# The skew occurs because there are more counties with a
#higher percent of citizens with HS degrees.
# I suspect that this is particularly affected by the Hidalgo, TX county where
#only 62.2% of the population has a HS diploma.
# More data points are likely to affect this by smoothing out the distribution
#and lessening the impact of the outlier.
# The kurtosis may also be affected because it is likely that more counties
#will occupy the 90% bin, resulting in a more leptokurtic distribution.
```