2014 American Community Survey_YoungMelissa.Rmd

Melissa Young

June 24, 2022

Load the packages

library(ggplot2) library(pastecs) library(psych)

```
library(ggplot2)
library(pastecs)
library(psych)
```

```
##
## Attaching package: 'psych'

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

```
library(formatR)
```

Load the csv fileto

df <- read.csv("acs-14-1yr-s0201.csv")

```
df <- read.csv("acs-14-1yr-s0201.csv")
```

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

```
summary(df)
```

```
##       Id                 Id2          Geography           PopGroupID
##  Length:136         Min.   : 1073   Length:136         Min.   :1
##  Class :character   1st Qu.:12082   Class :character   1st Qu.:1
##  Mode  :character   Median :26112   Mode  :character   Median :1
##                     Mean   :26833                      Mean   :1
##                     3rd Qu.:39123                      3rd Qu.:1
##                     Max.   :55079                      Max.   :1
##  POPGROUP.display.label RacesReported         HSDegree       BachDegree
##  Length:136             Min.   :  500292   Min.   :62.20   Min.   :15.40
##  Class :character       1st Qu.:  631380   1st Qu.:85.50   1st Qu.:29.65
##  Mode  :character       Median :  832708   Median :88.70   Median :34.10
##                         Mean   : 1144401   Mean   :87.63   Mean   :35.46
##                         3rd Qu.: 1216862   3rd Qu.:90.75   3rd Qu.:42.08
##                         Max.   :10116705   Max.   :95.50   Max.   :60.30
```

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

```
str(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,
##  $ 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 ...
```

```
nrow(df)
```

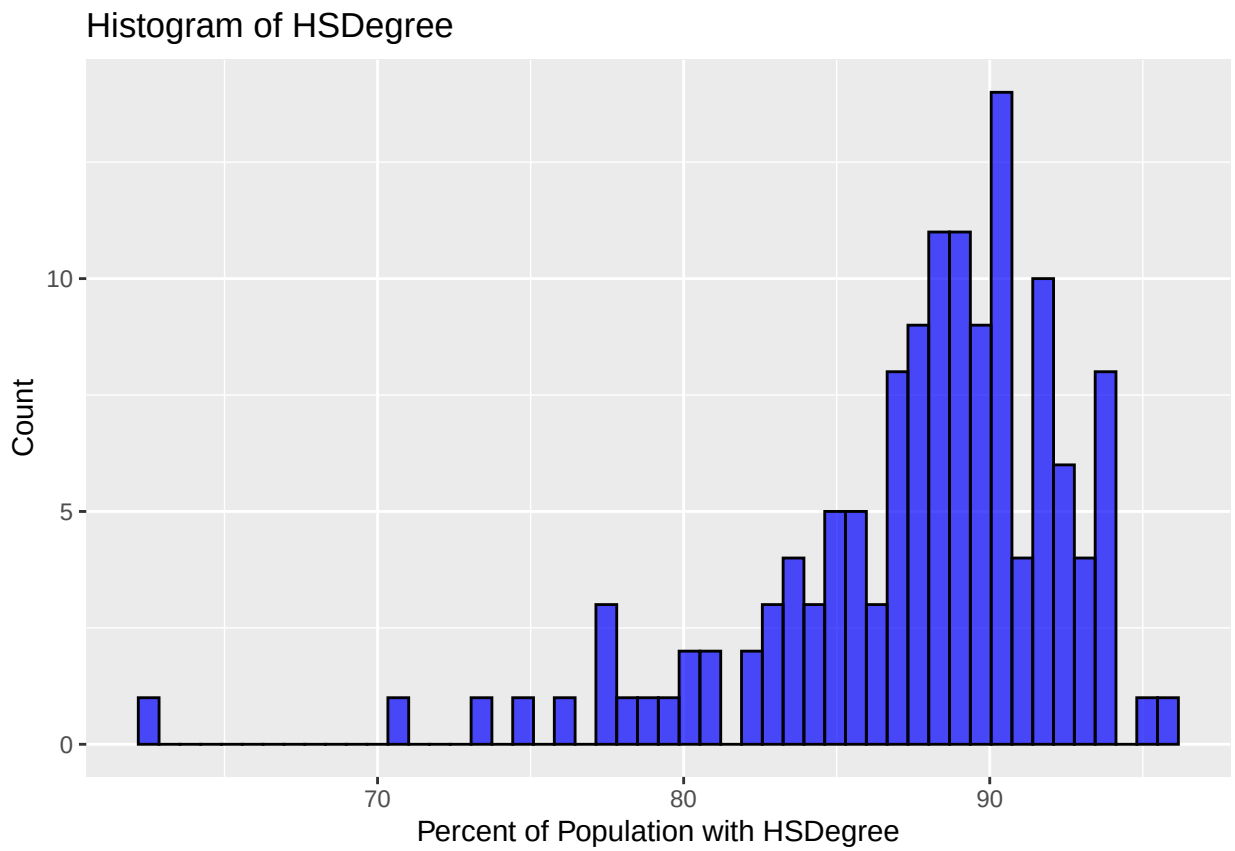
```
## [1] 136
```

```
ncol(df)
```

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

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

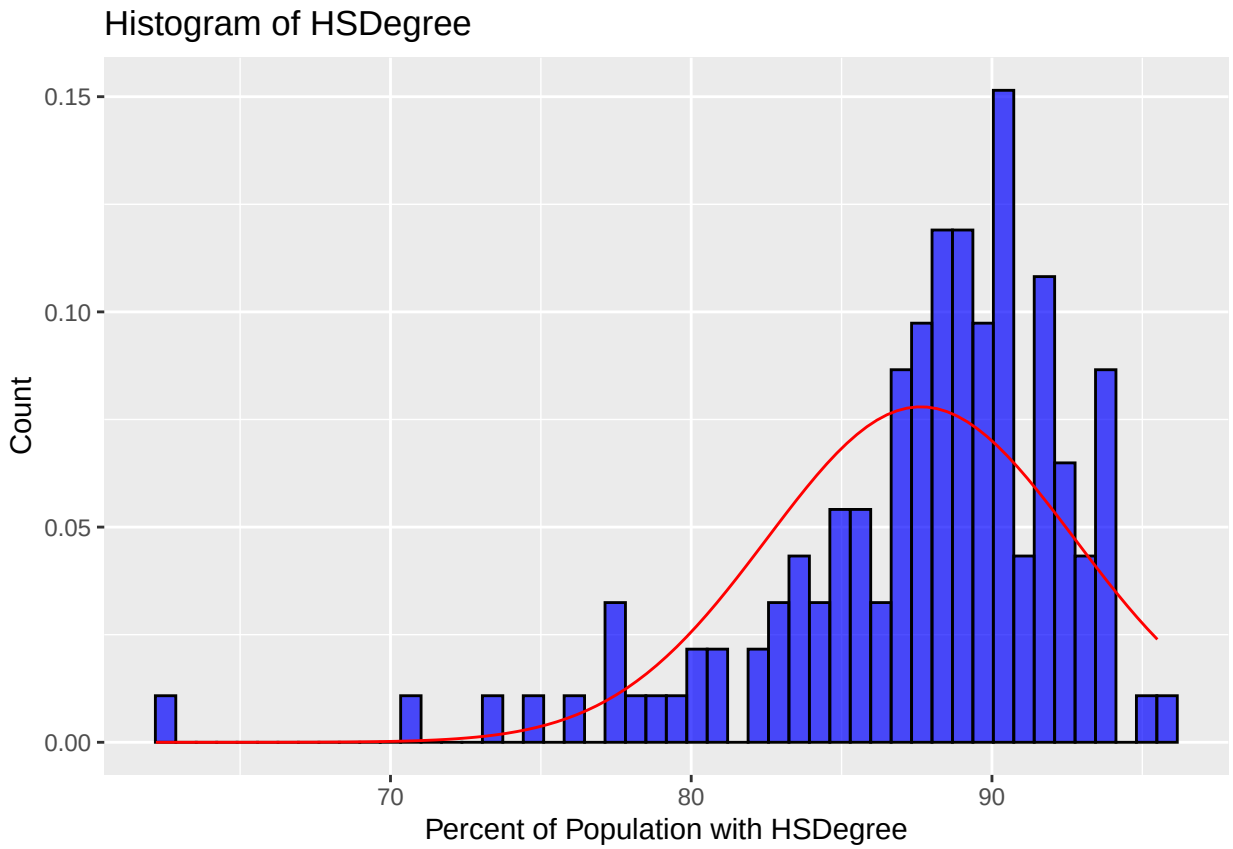
```
ggplot(data = df, aes(x = HSDegree)) + geom_histogram(bins = 50,
    color = "black", fill = "blue", alpha = 0.7) + ggtitle("Histogram of HSDegree") +
    xlab("Percent of Population with HSDegree") + ylab("Count")
```



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

##4.1 Based on what you see in this histogram, is the data distribution unimodal?
Answer - Yes, there is only one hump.
##4.2 Is it approximately symmetrical?
Answer - No, it seems to be negatively skewed to the left.
##4.3 Is it approximately bell-shaped?
Answer - No
##4.4 Is it approximately normal?
Answer - No, would be more bell shaped and symmetrical.
##4.5 If not normal, is the distribution skewed? If so, in which direction?
Answer - It's negatively skewed to the left. ##4.6 Include a normal curve to the Histogram that you plotted.

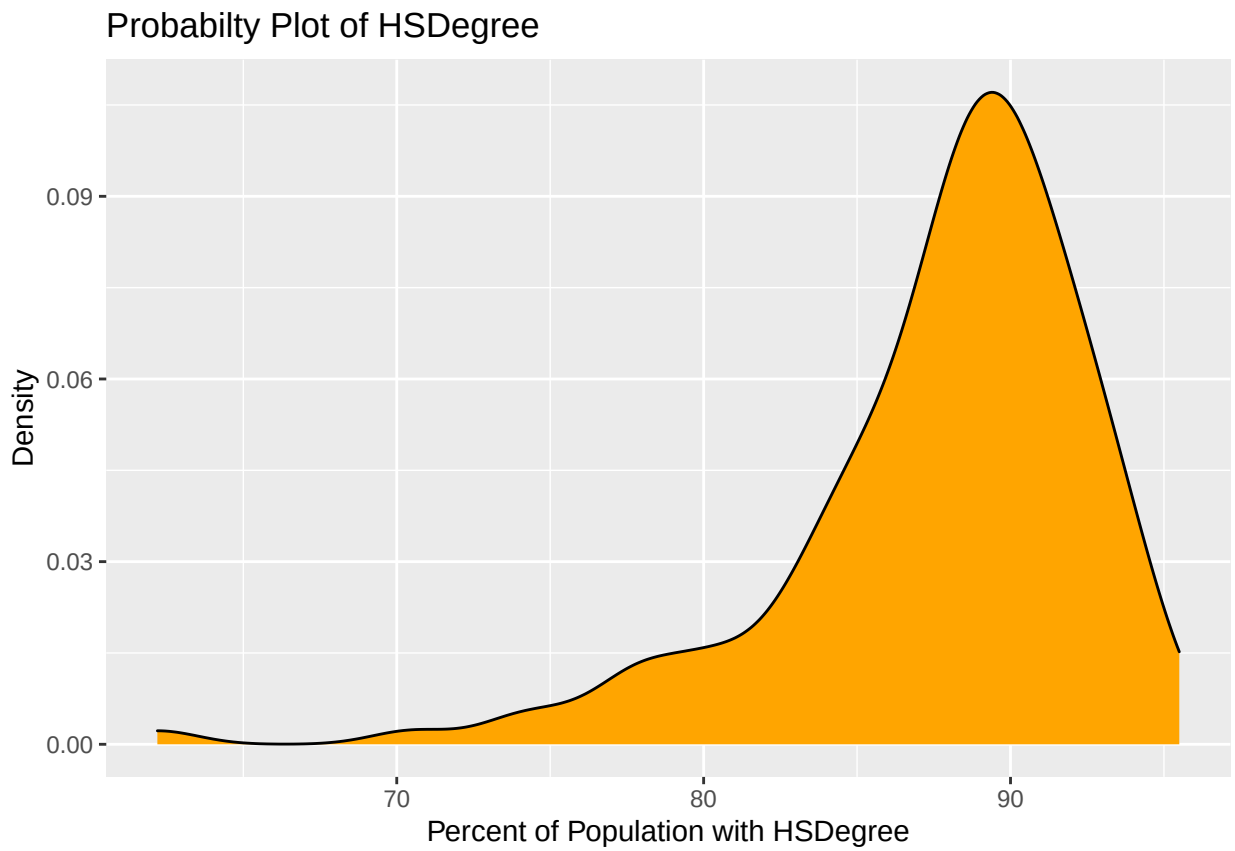
```
ggplot(data = df, aes(x = HSDegree)) + geom_histogram(aes(y = ..density..),
    bins = 50, colour = "black", fill = "blue", alpha = 0.7) +
    stat_function(fun = dnorm, args = list(mean = mean(df$HSDegree),
        sd = sd(df$HSDegree)), color = "red") + ggtitle("Histogram of HSDegree") +
    xlab("Percent of Population with HSDegree") + ylab("Count")
```



##4.7 Explain whether a normal distribution can accurately be used as a model for this data.
Answer - A normal distribution would not work with this dataset, as it is skewed.

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

```
ggplot(data = df, aes(x = HSDegree)) + geom_density(fill = "Orange") +
    ggtitle("Probabilty Plot of HSDegree") + xlab("Percent of Population with HSDegree") +
    ylab("Density")
```



##6. Answer the following questions based on the Probability Plot:

##6.1 Based on what you see in this probability plot, is the distribution approximately normal? Explain how you know.
Answer - No, this plot is not normal because it is not symmetrical. There is a tail from the left.
##6.2 If not normal, is the distribution skewed? If so, in which direction? Explain how you know.
Answer - Yes, skewed to the left. This plot is negative skew, with the longer tail on the left of the distribution.

##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(stats <- stat.desc(df$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
```

z_score <- round((df$HSDegree − mean(df$HSDegree))/sd(df$HSDegree), digits = 3)

```
z_score <- round((df$HSDegree - mean(df$HSDegree))/sd(df$HSDegree),
    digits = 3)
z_score
```

```
##   [1]  0.287 -0.163  0.072 -0.143  0.228 -2.742 -2.566 -1.980 -0.592 -1.374
##  [11] -0.163 -1.765 -0.202  0.091 -1.960  0.091 -0.045 -0.006 -1.804 -0.788
##  [21]  0.834 -0.417  1.010  1.264  0.424  0.326  0.365  0.482  0.502  0.775
##  [31]  0.150  0.267 -0.065 -0.260 -1.315  0.052  0.013  0.482 -0.534  0.248
##  [41]  0.521  0.150  0.717  0.072  0.814 -0.417  0.912 -0.925  0.521  0.599
##  [51] -0.514  1.537  0.228  0.170  0.834  0.541  0.638 -0.417 -0.632 -1.003
##  [61]  0.287  0.912  1.264  0.892 -0.729  0.482  0.287  0.326  1.166 -0.534
##  [71]  1.088  0.443  0.463  1.088  0.111 -0.612  0.756  0.130 -0.417 -0.827
##  [81]  0.287  1.068  0.795 -0.749 -0.280  0.072 -3.348  0.580 -1.491  0.521
##  [91]  0.599 -0.163 -1.413  0.424 -0.045  0.267  0.365  0.932  0.091  0.463
## [101]  0.560  0.404  0.678 -0.163  0.189  0.678  0.502  1.225  1.225  0.912
## [111]  0.756 -0.534  1.186 -0.983 -1.101 -0.182 -0.045 -0.905  1.186 -1.960
## [121]  0.834 -2.312  0.189 -1.530 -4.969 -0.338 -0.534  0.189  0.365  1.186
## [131]  0.756  0.912  0.521  0.853  1.420 -0.143
```

##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 - For the Skew we see a value of -1.675 this indicates that the distribution is highly negatively skewed to the left.
The Kurtosis is 4.353, this indicates a large tail with outliers.
The z-scores show high levels of variability with the outliers.
If you change to sample size, the summary stastics can be impacted by significant outliers and the denominator changes etc.