DSC520 Week 02 Assignment

Kurt Stoneburner

6/11/2020

1. There are 8 elements in the dataset

- Id : Categorical, used as an identifier
- Id2 : Categorical, used as an identifier
- Geography : Categorical
- PopGroupID : Categorical
- RacesReported : Quantitative
- HSDegree : Quantitative
- BachDegree : Quantitative

2.

str(survey_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 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 ...
```

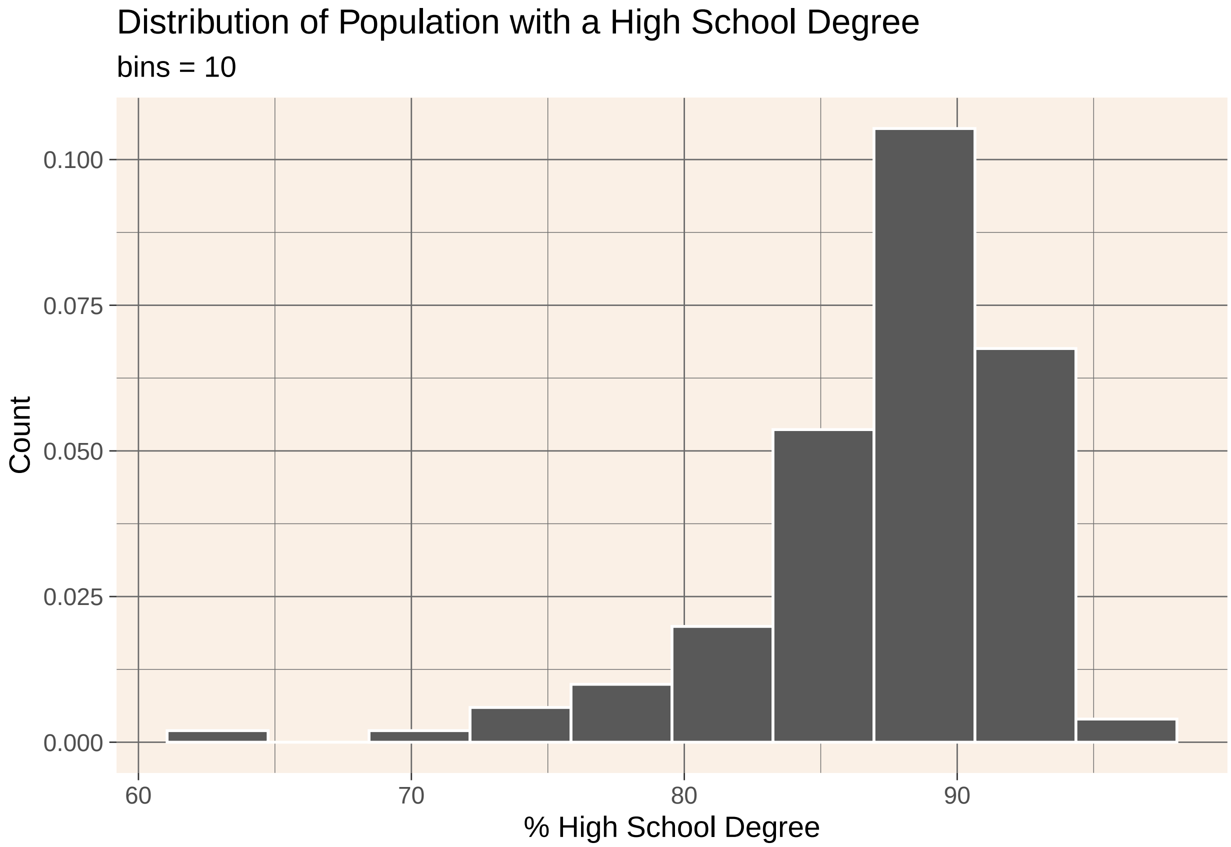
nrow(survey_df)

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

ncol(survey_df)

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

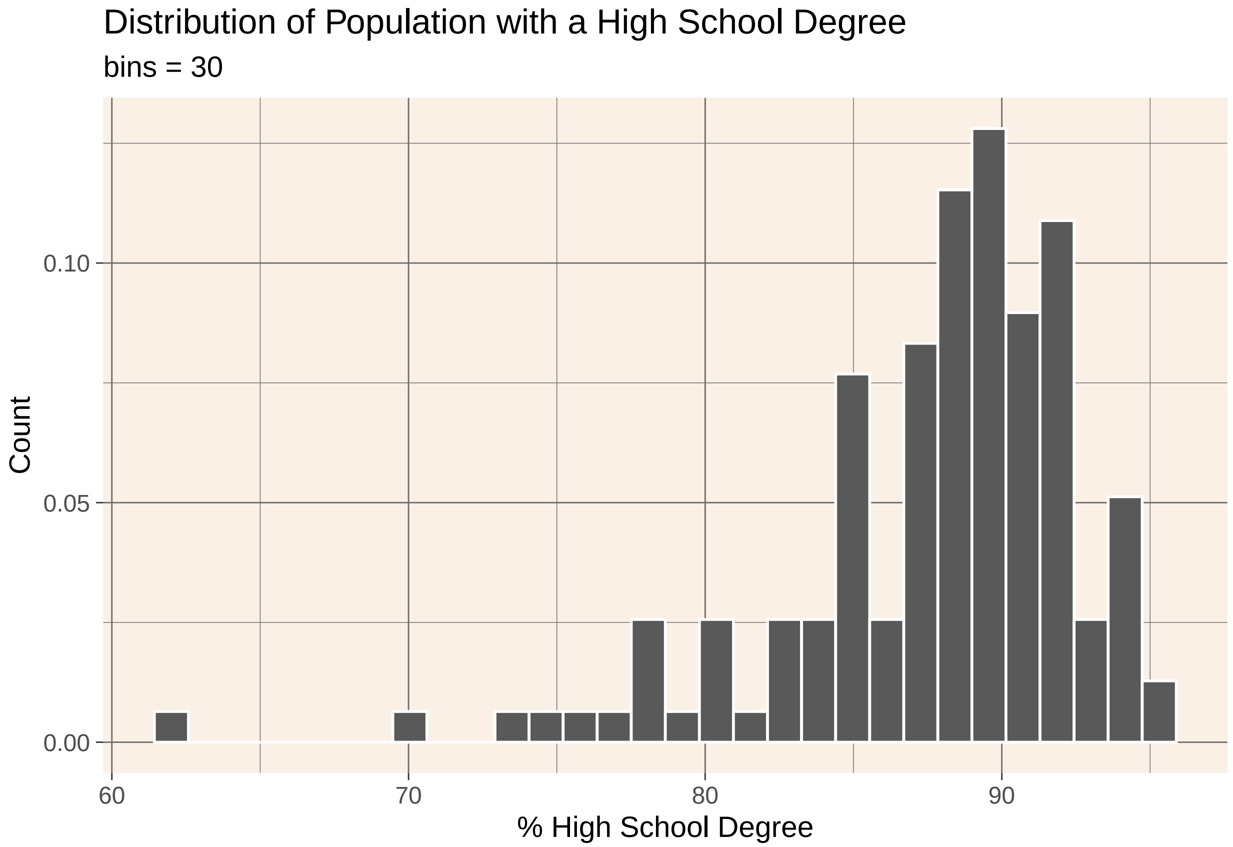
3. bins = 10. I used Rice's rule (cube root of the observations * 2) .



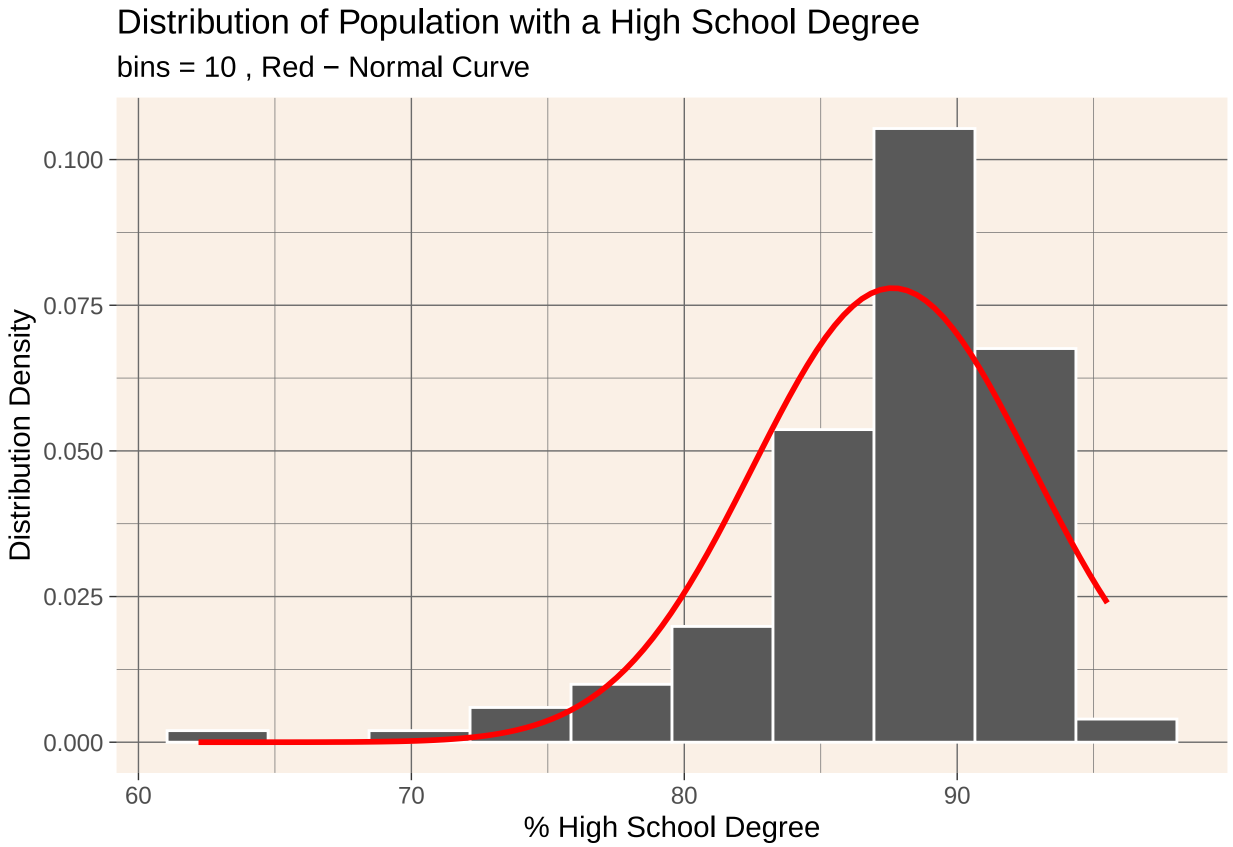
4.

- a. The Data appears to be modal. The measured values are continuous with a specificity that is generally greater than what is needed for categorical evaluation (There isn't a difference between 98% and 98.3%). The degree of unimodality is directly related to the bins size. Once the bins setting is greater than 30, the data no longer appears to be unimodal. Sampling with a higher bins value doesn't seem to have significance since it distorts the histogram by highlighting the variance (and making the graph noisy).

Example:

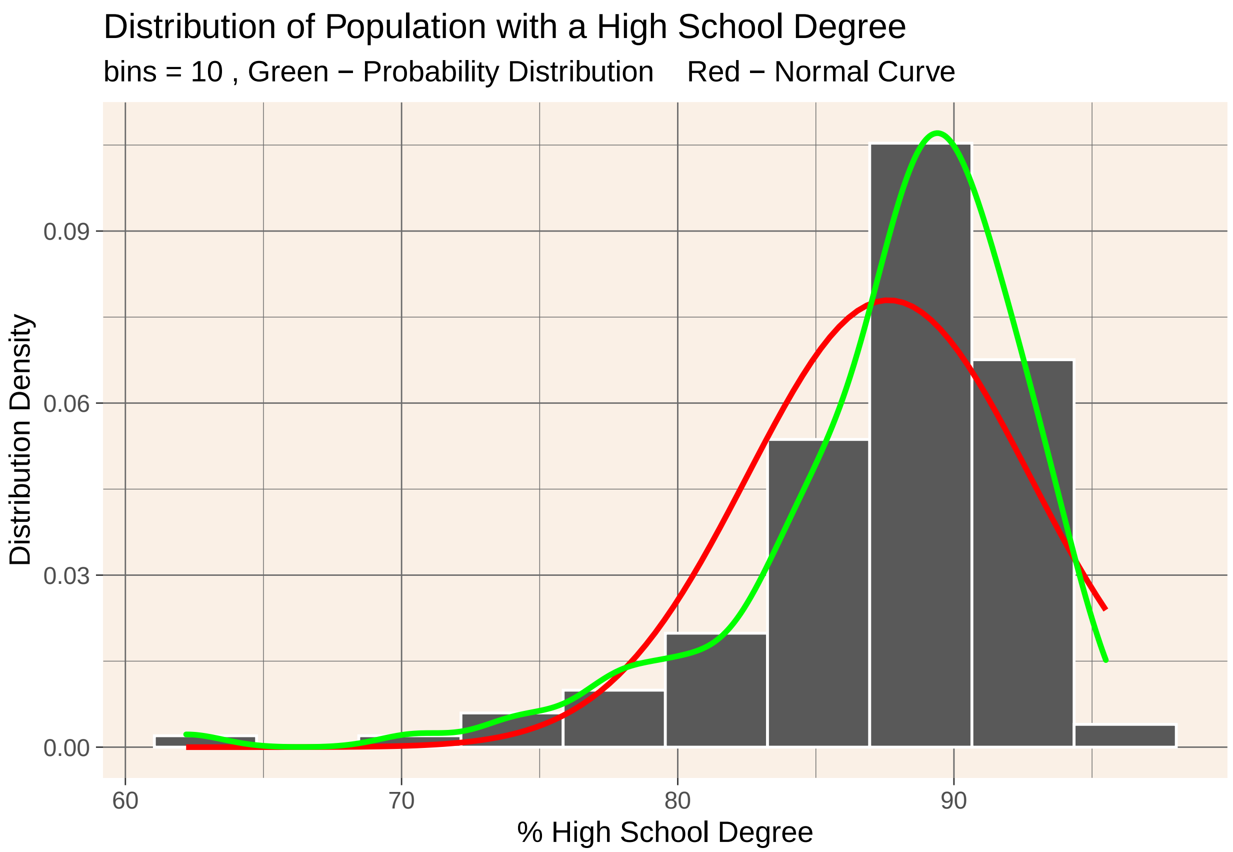


- b. The distribution does not appear to be symmetrical
- c. The distribution does not appear to be bell shaped.
- d. The distribution does not appear to be normal
- e. The distribution appears to be negatively skewed.
- f.



- g. The data does not appear to be normally distributed, therefore the normal distribution model cannot be used for this data.

5.



6.
   - a. The probability plot is not approximately normal. The shape of the probability plot has a sharper, more peaked curve compared to the normal curve. I plotted both and the differences are readily apparent.
   - b. The probability plot is negatively skewed because it trails off to the left. There isn't enough data to finish the plot to the right of the peak which adds to the appearance of negative skewness.

7.

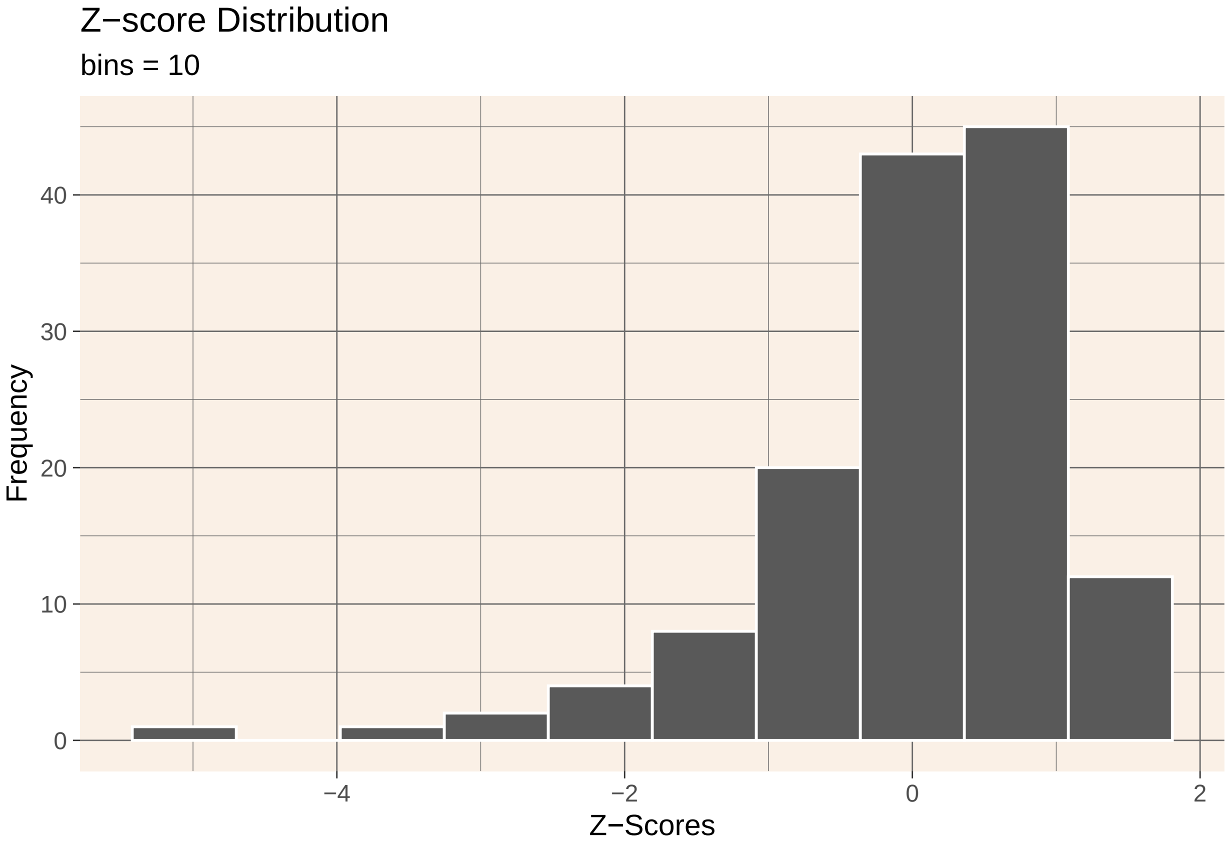
```
##        median          mean       SE.mean  CI.mean.0.95           var
##  8.870000e+01  8.763235e+01  4.388598e-01  8.679296e-01  2.619332e+01
##       std.dev      coef.var      skewness      skew.2SE      kurtosis
##  5.117941e+00  5.840241e-02 -1.674767e+00 -4.030254e+00  4.352856e+00
##      kurt.2SE    normtest.W    normtest.p
##  5.273885e+00  8.773635e-01  3.193634e-09
```

8. In a normal distribution the values for skew and kurtosis are 0. The further the values are from 0, the more likely the data is not normally distributed. The data has a kurtosis 4.352 and a skewness of -1.674. These values indicate the data is not normally distributed.

The data may be skewed because the sample size is not sufficiently large to represent a normal distribution according to the Central Limit Theorem. Increasing the sample size to 200 may result in a more normal distribution.

Non-statistically speaking, the data is measuring the frequency of a high school education across subsection of the US population. the quality of education varies greatly by geography in the United States and is directly linked to individual prosperity or the lack thereof measured by inequality. Given socio-economic issues endemic to the United States I hypothesize that basic education levels are not normally distributed (in the statistical sense) as a whole.

I am unsure of how to interpret the z-scores in this context. I understand an individual z-scores can be used to indicate the probability of an individual score. But I'm unclear on how to use and interpret the z-scores as a whole. I plotted the z-scores below, but that didn't seem significant.


Z−score Distribution
bins = 10
Frequency
Z−Scores