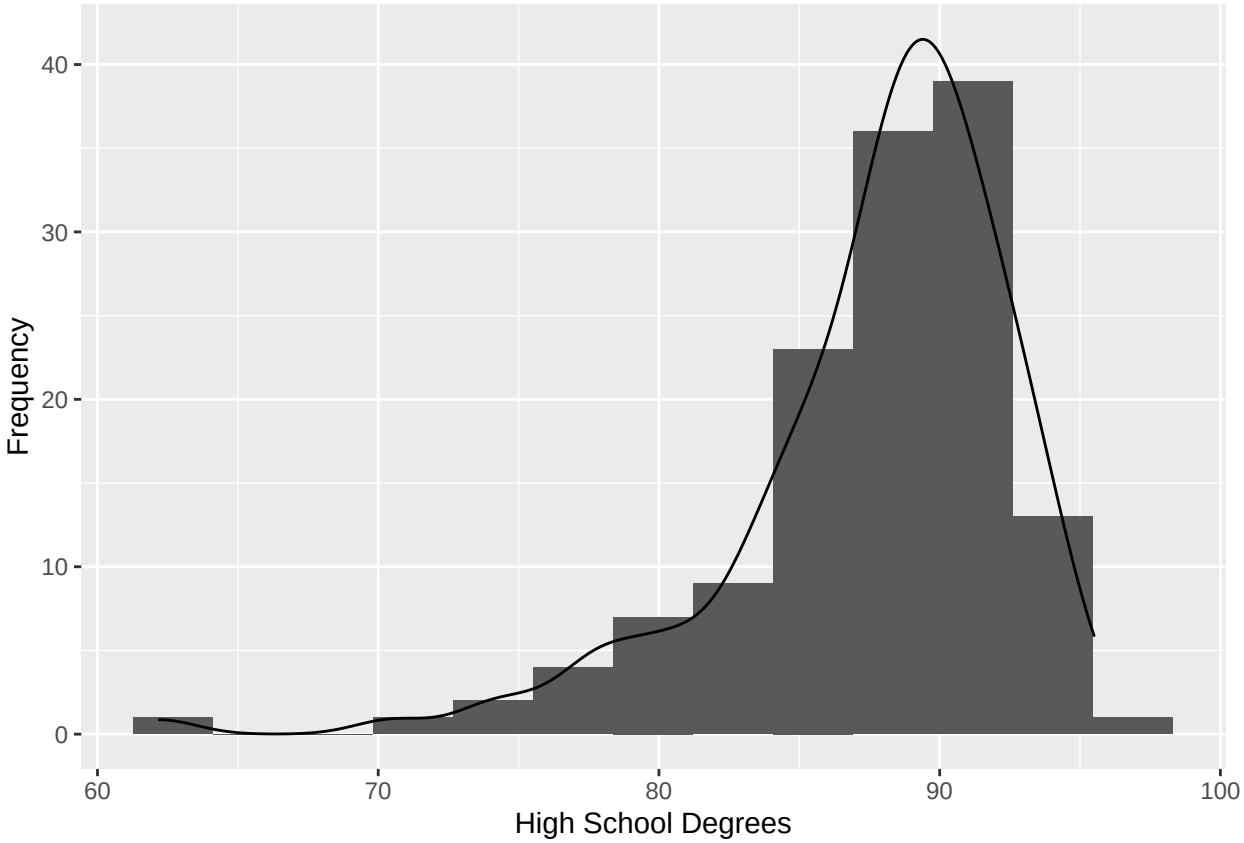
ACS_Dinkins_Darius

2022-06-23

The American Community Survey in 2014 contains 7 columns of data: Id: Ordinal Id2: Ordinal PopGroup: Ordinal Geography: Categorical POPGROUP.display-label: Categorical RacesReported:interval variable HSDegree: interval variable BachDegree: interval variable

3. Create a Histogram of the HSDegree variable using the ggplot2 package
4. Set a bin size for the Histogram Bins determined by sqrt of data points. Bin Width determined by (max - min) / bins
5. Include a Title and appropriate X/Y axis labels on your Histogram Plot.
6. Answer the following questions based on the Histogram produced:
7. Based on what you see in this histogram, is the data distribution unimodal?
   Yes, X appears the most in the data set
8. Is it approximately symmetrical? No. a vertical cannot be drawn in such a way that the shape to the left or right are similar.
9. Is it approximately bell-shaped? No, it skews toward a negative distribution
10. Is it approximately normal? No, it skews toward a negative distribution
11. If not normal, is the distribution skewed? If so, in which direction? The data skews toward a negative distribution
12. Include a normal curve to the Histogram that you plotted.

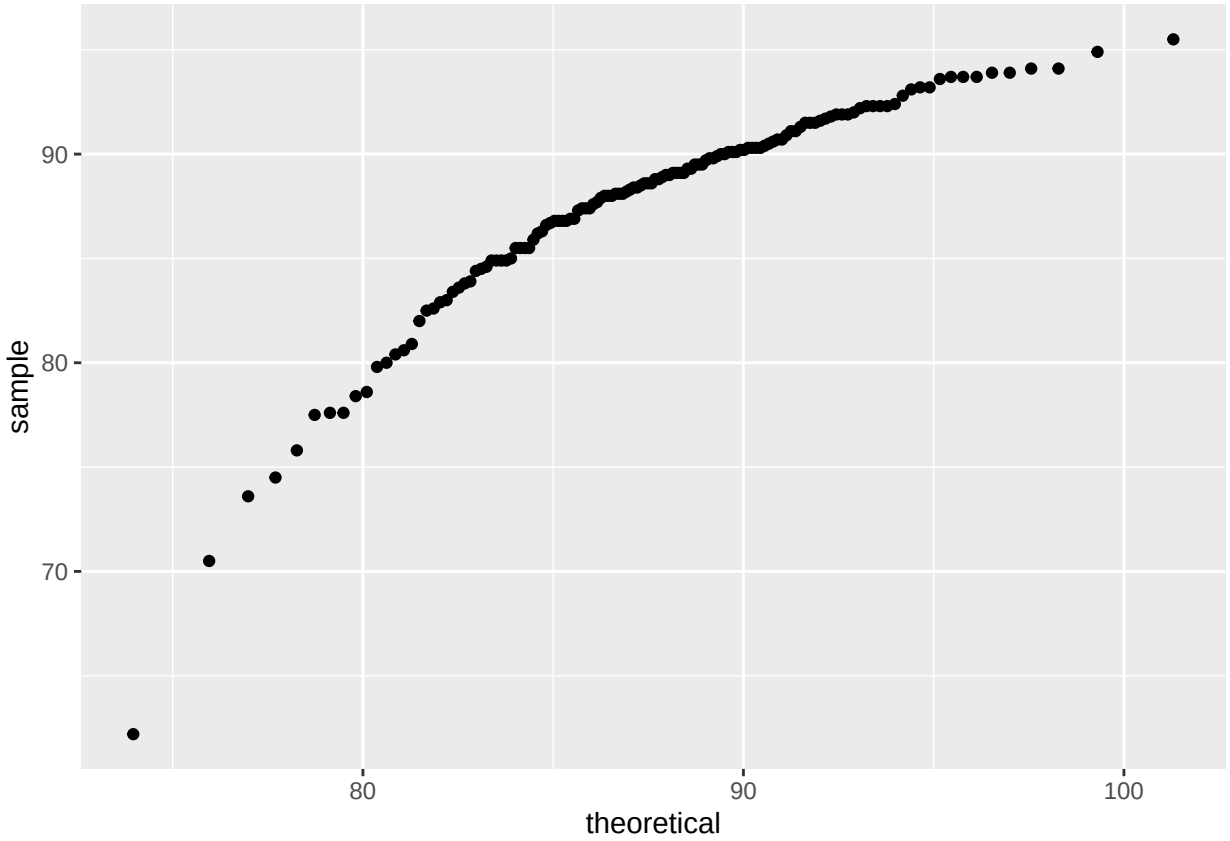
```
ggplot(survey_2014, aes(x=HSDegree)) + geom_histogram(bins = 12, binwidth = 2.85) +
  xlab('High School Degrees')  + ylab('Frequency') + geom_histogram(bins = 12, binwidth = 2.85)  +
  geom_density(aes(y = 2.85*..count..))
```



7. Explain whether a normal distribution can accurately be used as a model for this data. Normal distribution can not be accurately used as a model for this data. The data skews in negative manner. The majority of the data points exist toward the max.

8. Create a Probability Plot of the HSDegree variable. "'{r, include = false} ggplot(survey_2014)

```
ggplot(survey_2014, aes(sample =HSDegree)) + stat_qq_point()
```



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

7. Based on what you see in this probability plot, is the distribution approximately normal? Explain how you know. The distribution is not normal. It does not form a straight line. 2.If not normal, is the distribution skewed? If so, in which direction? Explain how you know. The distribution is skewed to the left. If you drew a straight line down the middle, the data points will skew towards the left.

8. 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.

```
library(pastecs)
```

```
## Warning: package 'pastecs' was built under R version 4.2.1
```

```
round(stat.desc(survey_2014[,"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
```

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?

Skew: Negative skew value indicates a majority of the data points on the right. This indicates that a large portion of the population has HS degrees

Kurtosis: A positive kurtosis indicates a pointy and heavy-tailed distribution. This indicates that a large portion of the population has degrees.

Z-Scores: Skew = (-1.675 - 87.632) = -89.307 / 5.118 = -17.44

Z-Scores: kurtosis = (4.353 - 87.632) = -83.279 / 5.118 = -16.272

The z scores are less than 1 by a large margin. This makes since as a large size of the population will produce large statisitical difference within the data.