3.2 Exercise

Felipe Rodriguez

2022-12-18

Data Type Dictionary

Id - Data Type: character
Definition: Unique identifier for the record

Id2 - Data Type: integer
Definition: Last 5 digits of Id

Geography - Data Type: character
Definition: County and State for the record

POPGROUPID - Data Type: integer
Definition: Unique Identifier for population group

POPGROUP.display.label - Data Type: character
Definition: Population group label

HSDegree - Data Type: numeric
Definition: Percentage amount of population with High School Degree

BachDegree - Data Type: numeric
Definition: Percentage amount of population with Bachelors Degree

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

```
library(ggplot2)
library(qqplotr)
```

```
##
## Attaching package: 'qqplotr'

## The following objects are masked from 'package:ggplot2':
##
##     stat_qq_line, StatQqLine
```

```
library(pastecs)
theme_set(theme_minimal())

setwd("/Users/feliperodriguez/OneDrive - Bellevue University/Github/dsc520/")

acs_data <- read.csv("data/acs-14-1yr-s0201.csv")

colnames(acs_data)
```

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

```
lapply(acs_data, class)
```

```
## $Id
## [1] "character"
##
## $Id2
## [1] "integer"
##
## $Geography
## [1] "character"
##
## $PopGroupID
## [1] "integer"
##
## $POPGROUP.display.label
## [1] "character"
##
## $RacesReported
## [1] "integer"
##
## $HSDegree
## [1] "numeric"
##
## $BachDegree
## [1] "numeric"
```

```
str(acs_data)
```

```
## '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(acs_data)
```

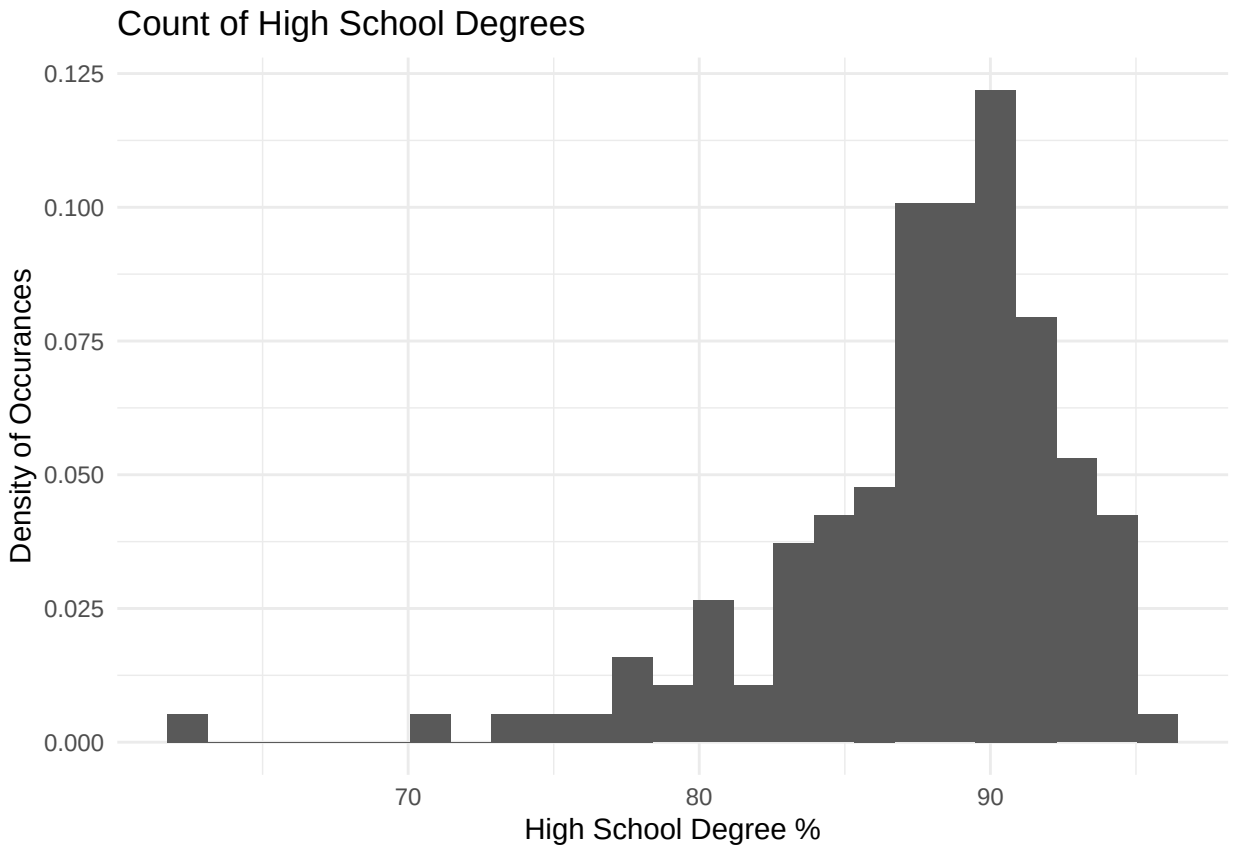
```
## [1] 136
```

```
ncol(acs_data)
```

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

```
acs_hist <- ggplot(acs_data, aes(HSDegree)) +
  ggtitle('Count of High School Degrees') +
  xlab('High School Degree %') +
  ylab('Density of Occurances') +
  geom_histogram(aes(y=..density..), bins = 25)
acs_hist
```

```
## Warning: The dot-dot notation (‘..density..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(density)‘ instead.
```

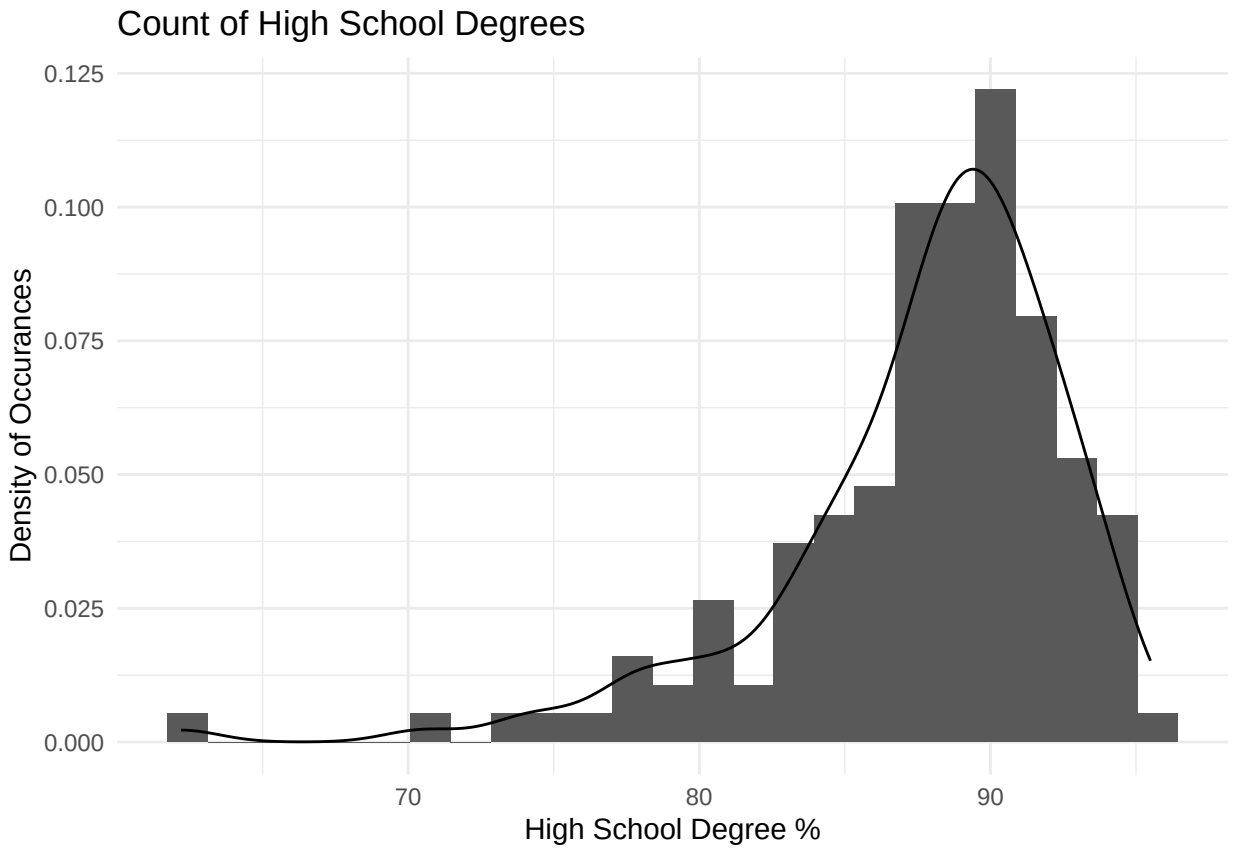


Based on what you see in this histogram, is the data distribution unimodal?
-Yes, the data distribution is unimodal.
Is it approximately symmetrical?
-The shape is not symmetrical.
Is it approximately bell-shaped?
-The shape appears to be bell-shaped but is not equal on either side.
Is it approximately normal?
-The shape is not normal since it is not centered and it is tilted to the left.

If not normal, is the distribution skewed? If so, in which direction?
-The distribution is skewed to the left.

Include a normal curve to the Histogram that you plotted.

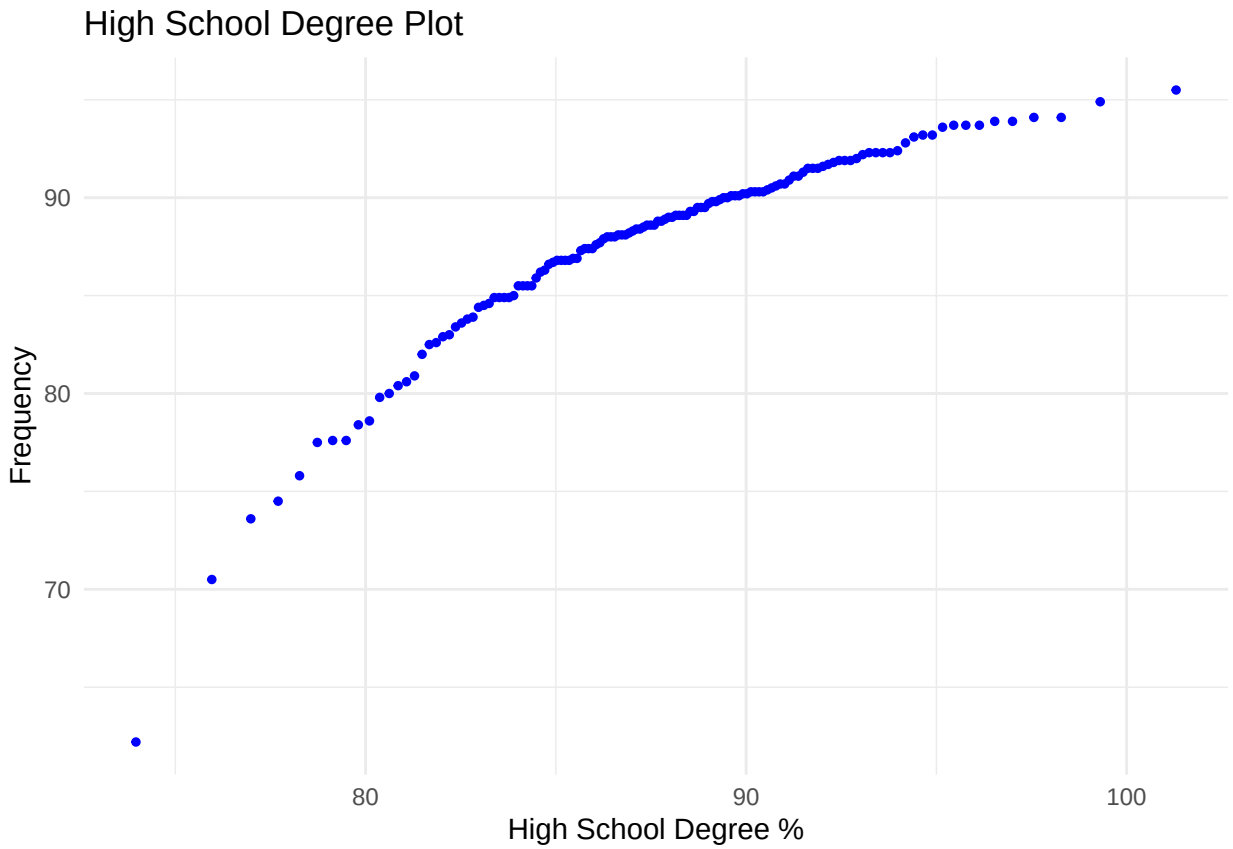
```
acs_hist + geom_density()
```



Explain whether a normal distribution can accurately be used as a model for this data. -This distribution can be used as a model because it is contains the majority of the data in its distribution. The density of values below 85 are a lot lower and not within the peak. The values above 95 are are also not within the peak so the middle values provide a good representation of distribution and can be used as a model.

Create a Probability Plot of the HSDegree variable.

```
acs_prob_plot <- ggplot(mapping = aes(sample = acs_data$HSDegree)) +
  stat_qq_point(size = 1,color = "blue") +
  ggtitle("High School Degree Plot") +
  xlab("High School Degree %") + ylab("Frequency")
acs_prob_plot
```



Based on what you see in this probability plot, is the distribution approximately normal? Explain how you know.
-The distribution is approximately normal. This can be identified because the majority of points are centrally found. Based on the graph, there is no indication that the points are skewed to the right or left.

Quantify normality with numbers using the stat.desc() function.

```
stats <- stat.desc(acs_data$HSDegree, norm=TRUE)
stats
```

```
##       nbr.val      nbr.null        nbr.na           min           max
##  1.360000e+02  0.000000e+00  0.000000e+00  6.220000e+01  9.550000e+01
##         range           sum        median          mean       SE.mean
##  3.330000e+01  1.191800e+04  8.870000e+01  8.763235e+01  4.388598e-01
##  CI.mean.0.95           var       std.dev      coef.var      skewness
##  8.679296e-01  2.619332e+01  5.117941e+00  5.840241e-02 -1.674767e+00
##      skew.2SE      kurtosis      kurt.2SE    normtest.W    normtest.p
## -4.030254e+00  4.352856e+00  5.273885e+00  8.773635e-01  3.193634e-09
```

```
HSDegree_mean <- mean(acs_data$HSDegree)
HSDegree_sd <- sd(acs_data$HSDegree)
HSDegree_zscore <- ((acs_data$HSDegree - HSDegree_mean)/HSDegree_sd)
matrix_zscore <- matrix(HSDegree_zscore, nrow=136)
colnames(matrix_zscore) <- c("z_score")
```

```
rownames(matrix_zscore) <- c(acs_data$Id)
head(matrix_zscore)
```

```
##                     z_score
## 0500000US01073  0.28676516
## 0500000US04013 -0.16263435
## 0500000US04019  0.07183496
## 0500000US06001 -0.14309524
## 0500000US06013  0.22814783
## 0500000US06019 -2.74179676
```

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?

Skewness is "a measure of asymmetry of the distribution of a variable" (Kim, 2013). The skewness value for HSDegree is -1.675 and a normal distribution has a value close to 0. This skewness value indicates that the distribution is shifted to the right. A change in sample size can give a wider sample set and will cause a move in skewness.

Kurtosis is a "a measure of the degree to which portfolio returns appear in the tails of our distribution" (Regenstein, 2013). The kurtosis value for HSDegree is 4.353 and a normal distribution has a kurtosis of 3. With kurtosis being above 3, it indicates that there are more tail returns than normal. A change in sample size could potentially reduce the amount of outliers that are in the dataset.

Z-score is a "measure that shows how much away (below or above) of the mean is a specific value (individual) in a given dataset" (Dhana, 2020). The z_score of HSDegree can indicate if the variable is above or below average. A change in sample size will change the mean and standard deviation. These two values will affect z_score directly.

Reference

Dhana, A. (2020, February 16). How to compute the Z-score with R: R-bloggers. R. Retrieved December 17, 2022, from https://www.r-bloggers.com/2020/02/how-to-compute-the-z-score-with-r/

Kim, H.-Y. (2013, February). Statistical notes for clinical researchers: Assessing Normal Distribution (2) using skewness and Kurtosis. Restorative dentistry & endodontics. Retrieved December 17, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3591587/

Regenstein, J. (2018, January 4). Introduction to kurtosis. · R Views. Retrieved December 17, 2022, from https://rviews.rstudio.com/2018/01/04/introduction-to-kurtosis/