DSC520: Week 3, Part 2

Kristie Kooken

2022-06-28

```
library(ggplot2)
theme_set(theme_minimal())
setwd("C:/Users/kkooken/Documents/EDU/520/R/dsc520-1")
## Load the 2014 American Community Survey data
## `data/acs-14-1yr-s0201.csv` to
acs_df <- read.csv("data/acs-14-1yr-s0201.csv")
```

Question 1: What are the elements in your data (including the categories and data types)?

```
sapply(acs_df, class)
```

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

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

```
str(acs_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, Arizona" "Alameda County, Californ
##  $ PopGroupID            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ POPGROUP.display.label: chr  "Total population" "Total population" "Total population" "Total population" ...
##  $ RacesReported         : int  660793 4087191 1004516 1610921 1111339 965974 874589 10116705 3145515 2329271 ...
##  $ 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_df)
```

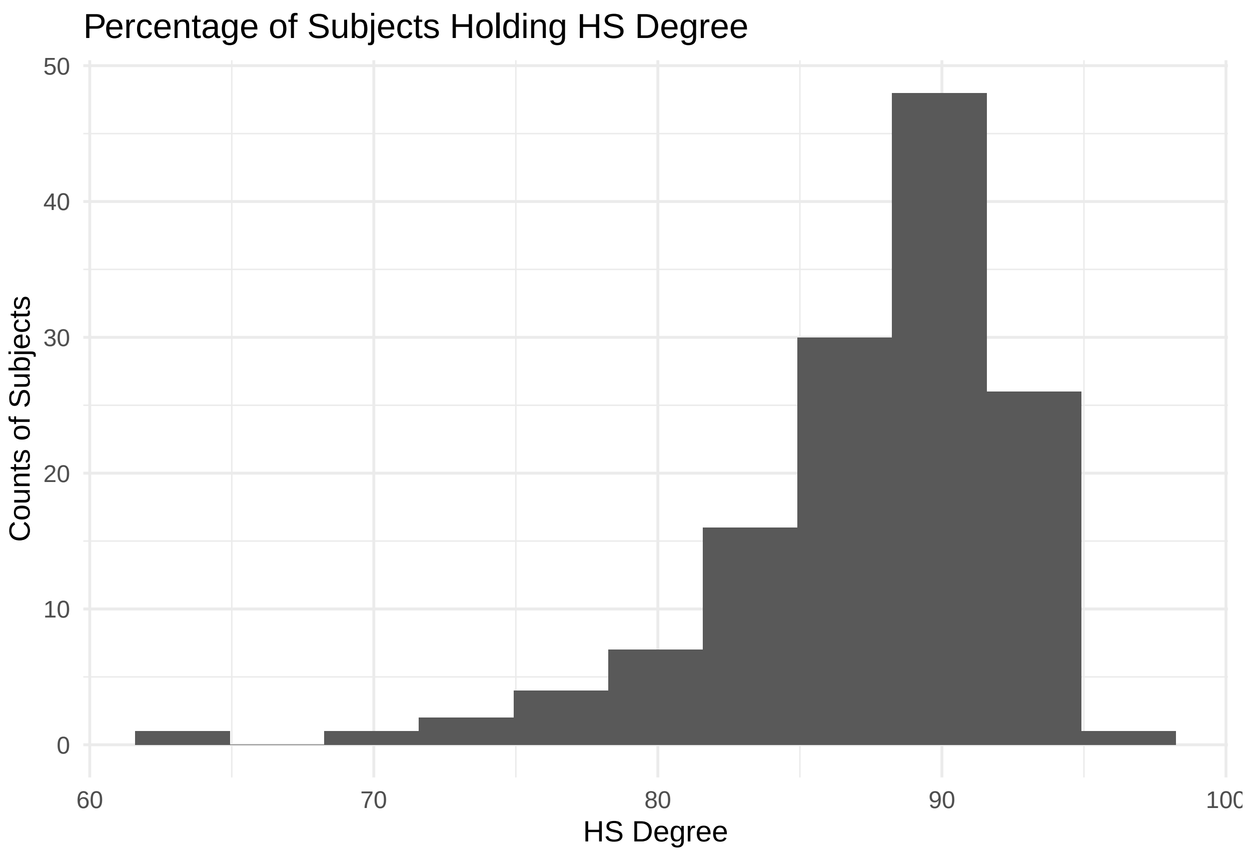
```
## [1] 136
```

```
ncol(acs_df)
```

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

Question 3: Create a Histogram of the HS Degree variable using the ggplot2 package including bin, title and axis labels

```
ggplot(acs_df, aes(HSDegree)) + geom_histogram(bins = 11) + ggtitle("Percentage of Subjects Holding HS Degree") +
    xlab("HS Degree") + ylab("Counts of Subjects")
```



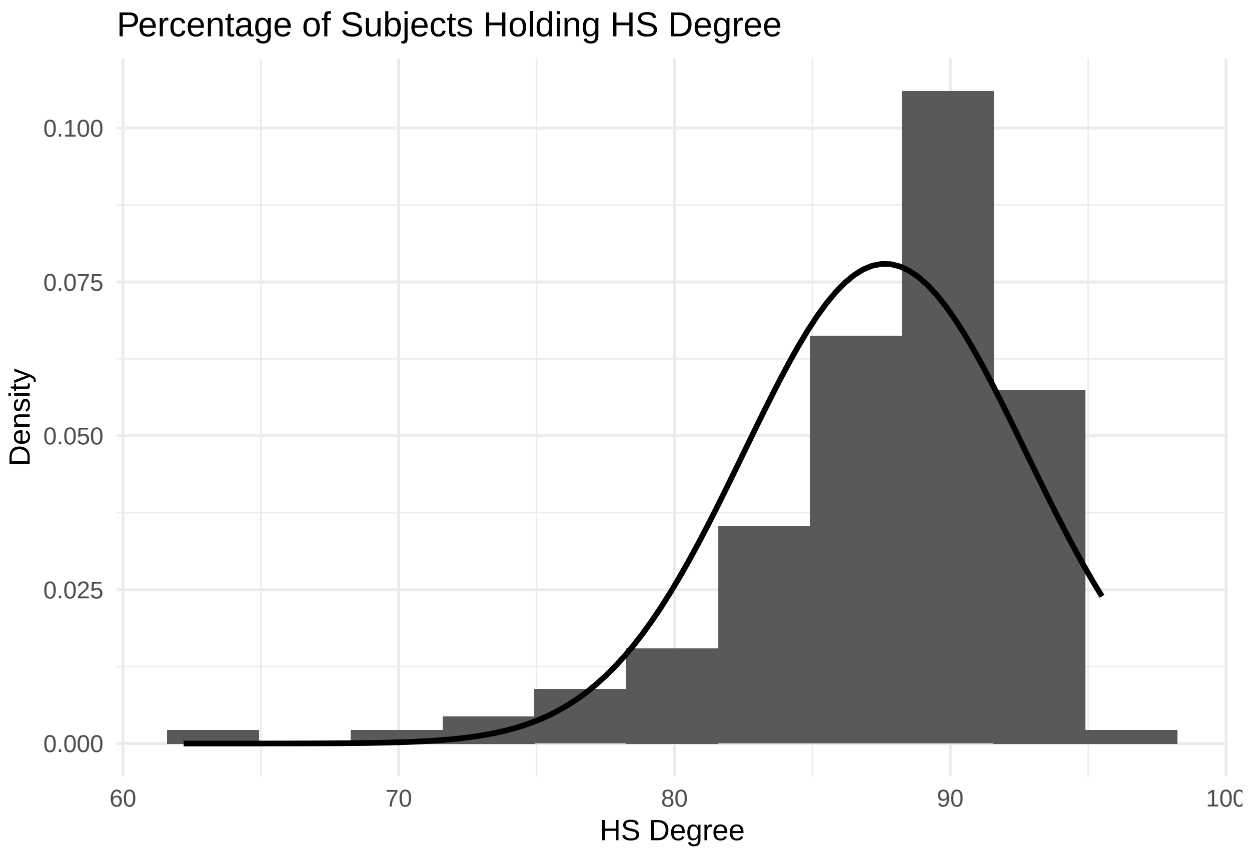
Question 4.1-4.5: Answers to questions about the produced histogram

This histogram of the high school data is unimodel, not symmetrical and is not bell shaped. The plot is negatively skewed to the left.

Question 4.6 Below is the same plot with a normal curve added

```
gplot1 <- ggplot(acs_df, aes(HSDegree)) + geom_histogram(aes(y = ..density..),
    bins = 11) + ggtitle("Percentage of Subjects Holding HS Degree") +
    xlab("HS Degree") + ylab("Density")

gplot1 + stat_function(fun = dnorm, args = list(mean = mean(acs_df$HSDegree,
    na.rm = TRUE), sd = sd(acs_df$HSDegree, na.rm = TRUE)), color = "black",
    size = 1)
```



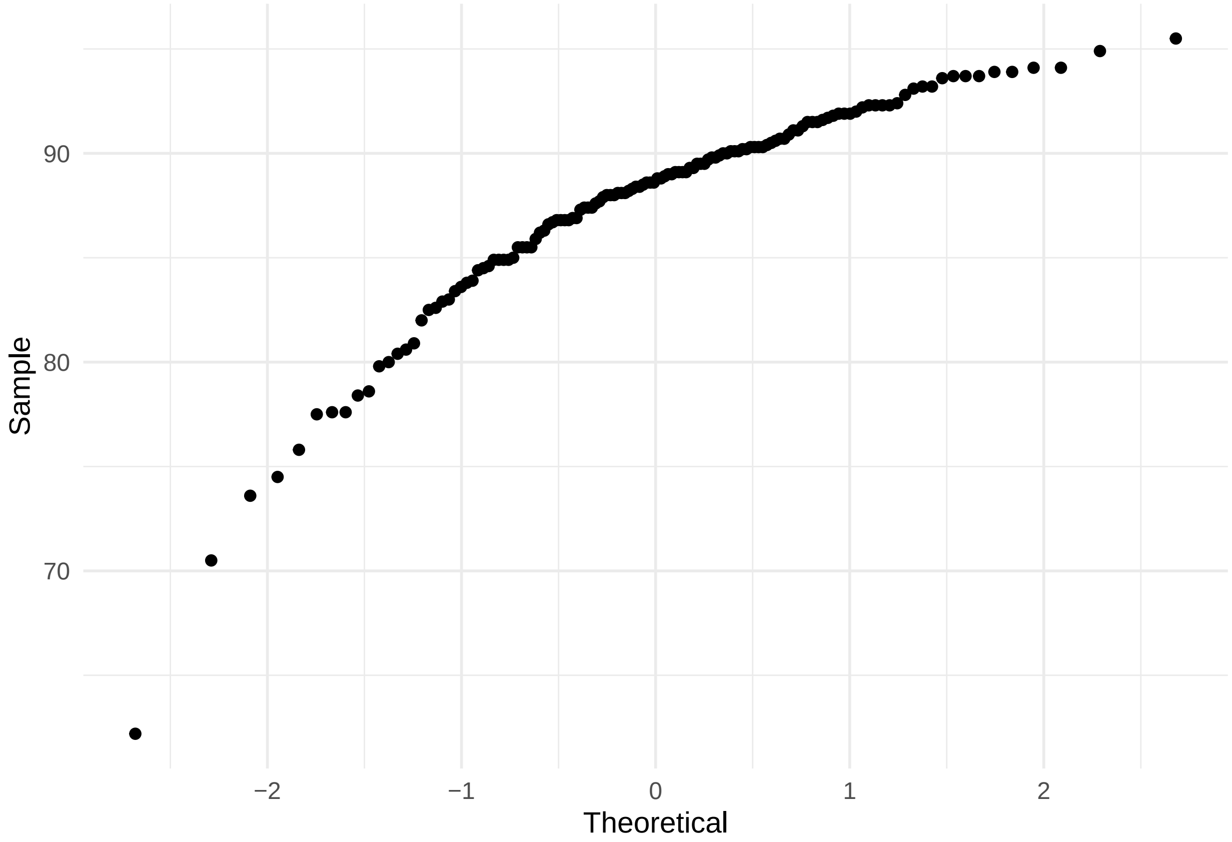
Question 4.7 Response:

Adding a normal curve to this histogram does not model this data well as the histogram is skewed to the left meaning that a higher level of occurrences happen at the higher end of the x axis.

Question 5: Create a probability plot of the high school data

```
qqplot1 <- ggplot(acs_df, aes(sample = HSDegree)) + stat_qq() +
    xlab("Theoretical") + ylab("Sample")

qqplot1
```



Question 6.1-6.2 Based on what you see in this probability plot, is the distribution approximately normal?

Based on the probability plot, this distribution appears to be not symmetrical and negatively skewed, this is because the plotted line is not symmetric and appears to curve outward like a bow or plateau at the top. A positive skew would be curved in, going upward.

Question 7 Quantify normality with numbers using the stat.desc() function

```
library(pastecs)
stat.desc(acs_df$HSDegree, basic = FALSE, norm = TRUE)
```

```
##        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
```

```
# Descriptive stats for z scores of HS Degree variable
library(psych)
describe(scale(acs_df$HSDegree))
```

```
##    vars   n mean sd median trimmed  mad   min  max range  skew kurtosis   se
## X1    1 136    0  1   0.21    0.13 0.74 -4.97 1.54  6.51 -1.67     4.35 0.09
```

Question 8: Result interpretation

For the variable HS Degree, the skew is -1.67, because this value is negative, this means that there is a build up of high scores. Likewise, the value of 4.35 for Kurtosis is a positive number meaning that the shape of plot would be pointy with heavy-tailed distribution. Likewise, since both these values are not zero, this indicates that the distribution is not normal.

For the z scores, I generated z scores on the variable HS Degree and confirmed the mean is 0 and standard deviation is 1 as expected. In order to see the z score for skew and kurtosis, the formula of

((variable-0)/SE for variable)

can be used. This equation is condensed with the outputted variables 2SE for Skew and Kurtosis. The 2SE value is either skew or kurtosis divided by 2*standard error and thus a handy way to check if the skew or kurtosis are significantly different than chance. Both of these values are greater than

1 (representing $p < 0.05$), indicating significant skew and significant kurtosis for the HS Degree variable. In fact, the z score values are significant at $p < 0.001$ (having a value greater than 1.65 which represents 1.65 times 2 - a z score of 3.29)

Our sample size is relatively small (n=136) and adding more data could make using these variables to determine the significance of skew and kurtosis challenging. This method is better for smaller sample sizes. With larger samples, histograms are better indicators of the shape of the distribution.