Cordero_week3.2

Joaquin Cordero

2024-06-24

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stats)
library(pastecs)
```

```
##
## Attaching package: 'pastecs'

## The following objects are masked from 'package:dplyr':
##
##     first, last
```

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

#Id - Data Type: integer , Intent: unique identifier for each row

#Id2 - Data Type: integer ,
#Intent: last 4-5 digits of 'Id' and unique identifier for each row
```

```r
#Geography - Data Type: text ,
#Intent: unique identifier for each location

#PopGroupID - Data Type: integer ,
#Intent: unique identifier for 'Total population'

#POPGROUP.display.label - Data Type: text ,
#Intent: unique identifier for 'PopGroupID'

#RacesReported - Data Type: integer ,
#Intent: represents the population in each geography

#HSDegree - Data Type: integer ,
#Intent: percent of the population that has HS degree

#BachDegree - Data Type: integer ,
#Intent: percent of the population that has bachelor degree

str(acs_dataset)
```

```
## '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 ...
```

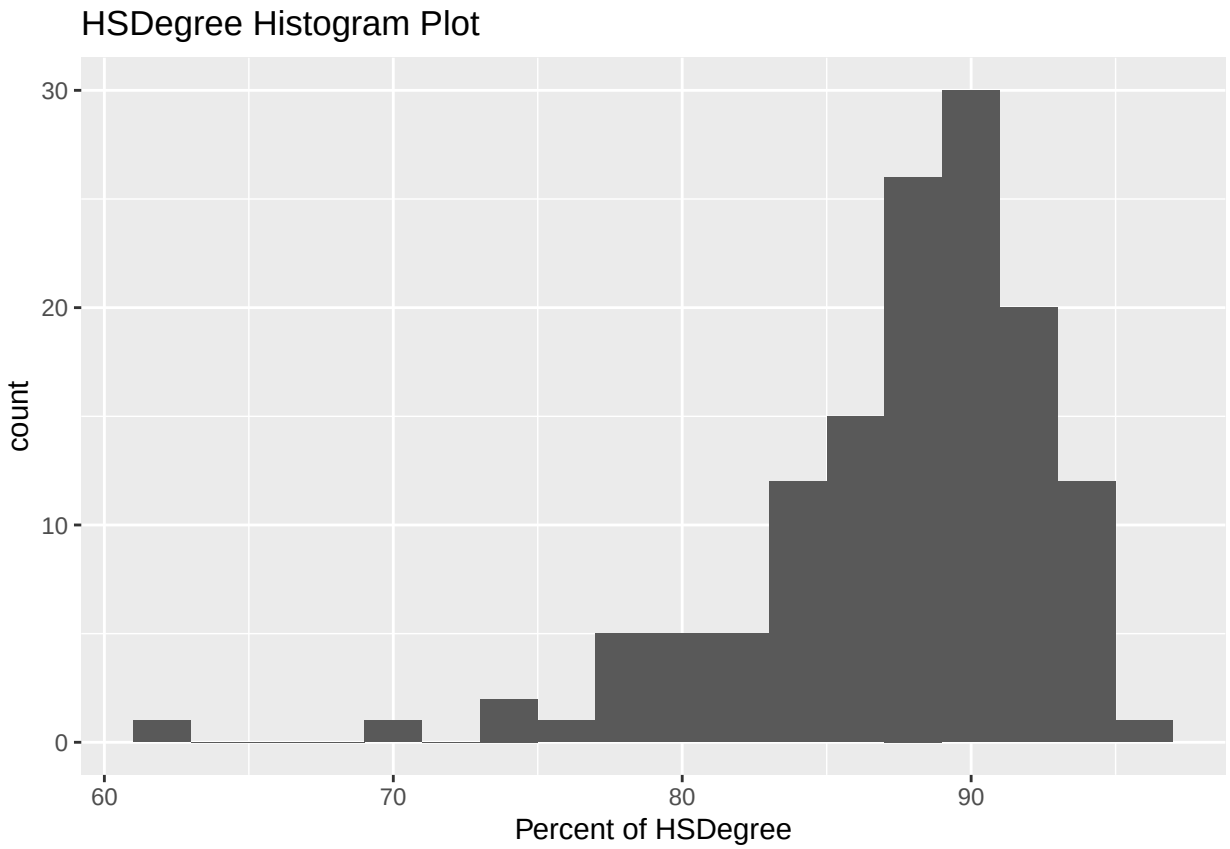
```r
nrow(acs_dataset)
```

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

```r
ncol(acs_dataset)
```

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

```r
mean_HSDegree <- mean(acs_dataset$HSDegree)
sd_HSDegree <- sd(acs_dataset$HSDegree)

hsd_plot <- ggplot(acs_dataset, aes(x = HSDegree,)) +
  geom_histogram(binwidth = 2)
hsd_plot <- hsd_plot + labs(x = "Percent of HSDegree", y = "count",
                            title = "HSDegree Histogram Plot" )
print(hsd_plot)
```




```
'1.Based on what you see in this histogram, is the data distribution unimodal?
-Yes, based on the histogram the data distribution is unimodal.'
```

```
## [1] "1.Based on what you see in this histogram, is the data distribution unimodal?\n-Yes, based on t
```

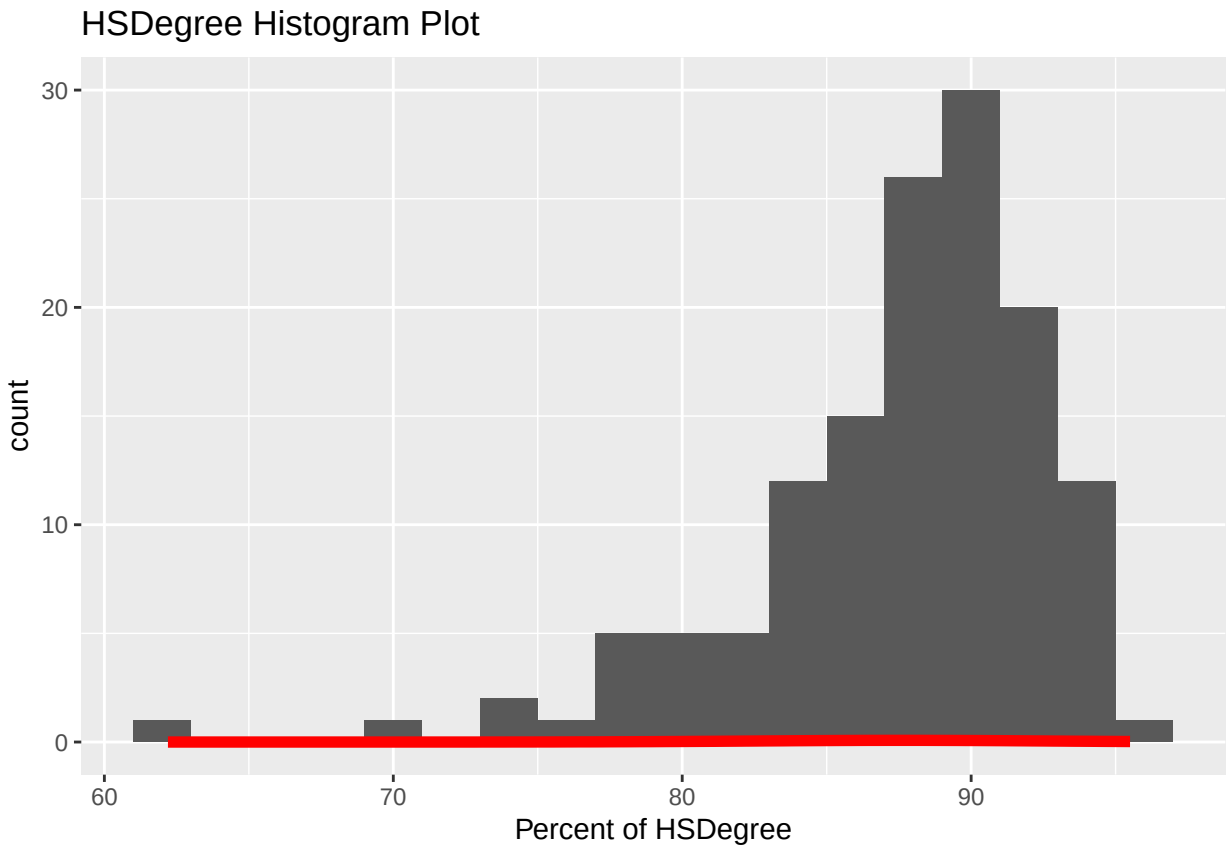
```
'2.Is it approximately symmetrical?
-No, it is not symmetrical?'
```

```
## [1] "2.Is it approximately symmetrical?\n-No, it is not symmetrical?"
```

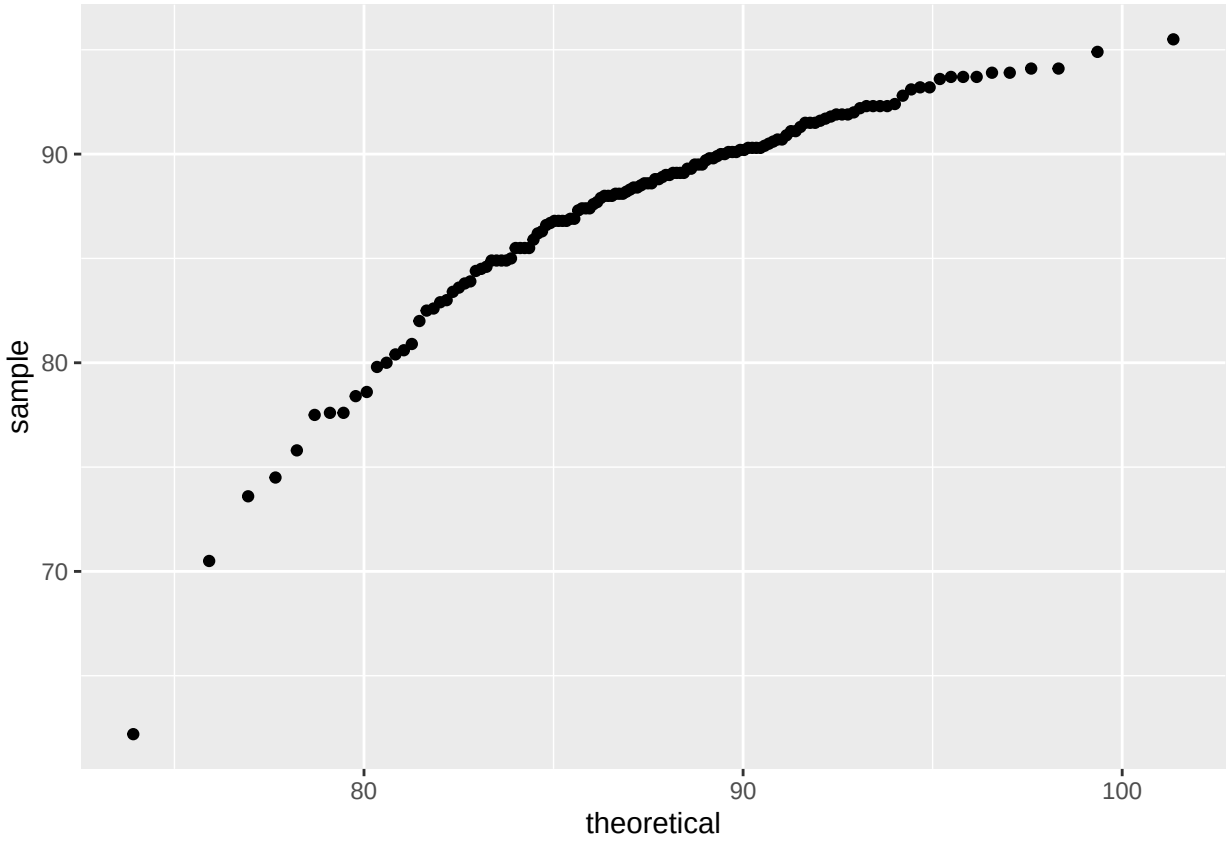
```
#3.Is it approximately bell-shaped?
#-Yes, it is approximately bell-shaped

#4.Is it approximately normal?
#-No, it is not normal shaped

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

#6.Include a normal curve to the Histogram that you plotted.
hsd_plot <- hsd_plot +
  stat_function(fun = dnorm,
                args = list(mean = mean_HSDegree, sd = sd_HSDegree),
                color = "red", size = 2)
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```r
print(hsd_plot)
```




```r
#7.Explain whether a normal distribution can accurately be used as a
#model for this data.
#-No, I do not believe a normal distribution can accurately
#be used as a model for this data'

hsd_prob_plot <- ggplot(acs_dataset, aes(sample = HSDegree)) +
  stat_qq(distribution = qnorm,
          dparams = list(mean = mean(acs_dataset$HSDegree),
                         sd = sd(acs_dataset$HSDegree)))
print(hsd_prob_plot)
```




```
#1.Based on what you see in this probability plot,
#is the distribution approximately normal? Explain how you know.
#-No, because the plotted points bends down

#2.If not normal, is the distribution skewed? If so, in which direction?
#Explain how you know.
#-The distribution is skewed left becuase the plotted points curve down

stat.desc(acs_dataset$HSDegree)
```

```
##      nbr.val     nbr.null       nbr.na          min          max        range
## 1.360000e+02 0.000000e+00 0.000000e+00 6.220000e+01 9.550000e+01 3.330000e+01
##          sum       median         mean      SE.mean CI.mean.0.95          var
## 1.191800e+04 8.870000e+01 8.763235e+01 4.388598e-01 8.679296e-01 2.619332e+01
##      std.dev     coef.var
## 5.117941e+00 5.840241e-02
```

```
#Skew measures the dataset's asymmetry and the probability distribution.
#Kurtosis measures the outliers in the dataset.
#Z-score can measure how many standard deviation away from the dataset's
#mean a data point is. A change in sample size can change how accurate these
#are because smaller sample sizes will tend to show less of the overall picture.
```