Week3_Assignment_AmericanCommunitySurvey

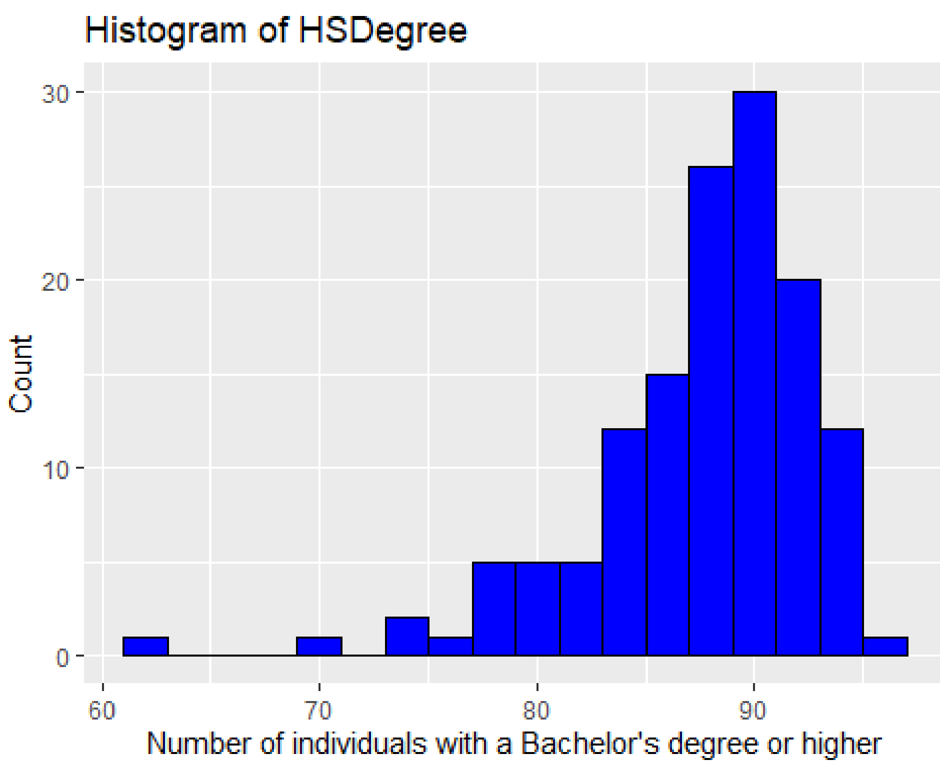
Henry Pham

2023-04-02

```
## Set the working directory to the root of your DSC 520 directory

getwd()

## [1] "C:/R/DSC520"

dir("C://R//DSC520//data")

## [1] "acs-14-1yr-s0201.csv"           "G04ResultsDetail2004-11-02.xls"
## [3] "r4ds"                           "scores.csv"
## [5] "tidynomicon"

## [1] "acs-14-1yr-s0201.csv" "G04ResultsDetail2004-11-02.xls"
## [3] "r4ds" "scores.csv"
## [5] "tidynomicon"
if (!file.exists("r4ds"))
{
 # set data working directory
setwd("C://R//DSC520//data")
}

# Load the data
acs_data <- read.csv("acs-14-1yr-s0201.csv")
head(acs_data)

##               Id  Id2                      Geography PopGroupID
## 1 0500000US01073 1073       Jefferson County, Alabama          1
## 2 0500000US04013 4013        Maricopa County, Arizona          1
## 3 0500000US04019 4019            Pima County, Arizona          1
## 4 0500000US06001 6001      Alameda County, California          1
## 5 0500000US06013 6013 Contra Costa County, California          1
## 6 0500000US06019 6019       Fresno County, California          1
##   POPGROUP.display.label RacesReported HSDegree BachDegree
## 1       Total population        660793     89.1       30.5
## 2       Total population       4087191     86.8       30.2
## 3       Total population       1004516     88.0       30.8
## 4       Total population       1610921     86.9       42.8
## 5       Total population       1111339     88.8       39.7
## 6       Total population        965974     73.6       19.7

## i.   List the name of each field and what you believe the data type and
intent is of the data included in each field (Example: Id - Data Type:
varchar (contains text and numbers) Intent: unique identifier for each row)
```

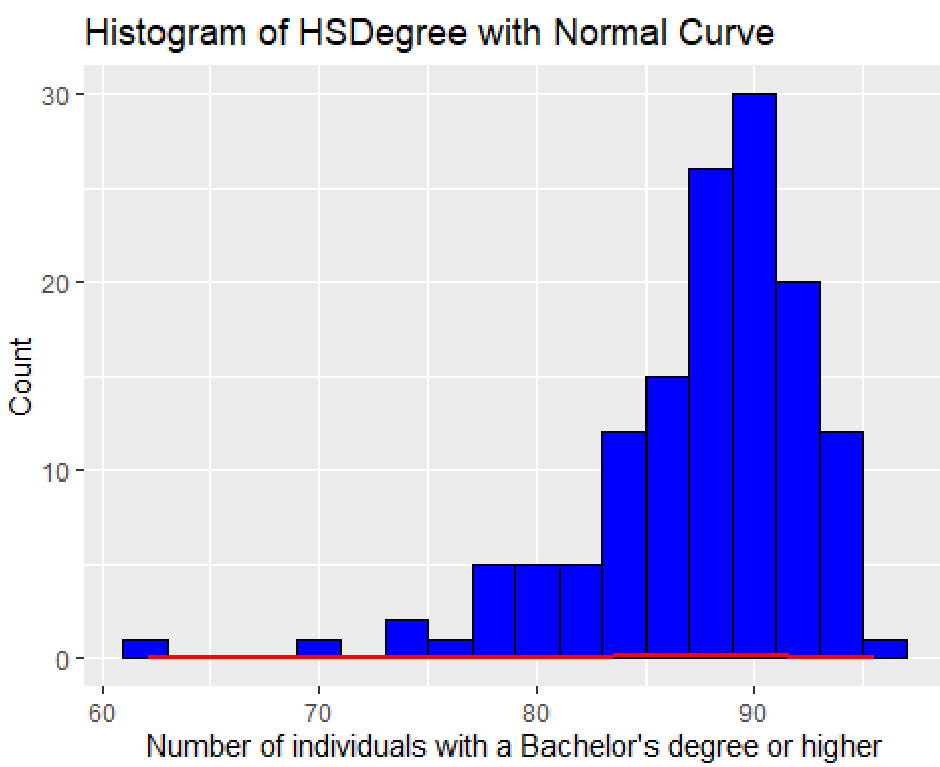
```r
# Id: unique identified number, data type: character, Intent: unique key.
# Id2: unique identified number, data type: integer, Intent: unique key.
# Geography: the geographic area (city and state), data type: character,
Intent: the geopraphic area id.
# PopGroupID: Population group id, data type = integer, Intent: classify
group id.
# POPGROUP.display-label: total population of that group id, data type:
character, Intent: total population of that group id.
# RacesReported: population of races, data type: integer, intent: population
of races.
# HSDegree: percentage of high school degree, data type: numeric, intent:
percentage of people got high school degree.
# BachDegree: percentage of bachelor degree, data type: numeric, intent:
percentage of people got bachelor degree.


# Use str() to show the structure of the data frame
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, Arizona" "Alameda County, California" ...
##  $ 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 ...

# Use nrow() to show the number of rows in the data frame
nrow(acs_data)

## [1] 136

# Use ncol() to show the number of columns in the data frame
ncol(acs_data)

## [1] 8

## Create a Histogram of the HSDegree variable using the ggplot2 package.
# Set a bin size for the Histogram that you think best visuals the data (the
bin size will determine how many bars display and how wide they are)
# Include a Title and appropriate X/Y axis labels on your Histogram Plot.
```

```r
library(ggplot2)

# Load the data
acs_data <- read.csv("acs-14-1yr-s0201.csv")

ggplot(acs_data, aes(x = HSDegree)) +
  geom_histogram(binwidth = 2, fill = "blue", color = "black") +
  ggtitle("Histogram of HSDegree") +
  xlab("Number of individuals with a Bachelor's degree or higher") +
  ylab("Count")
```



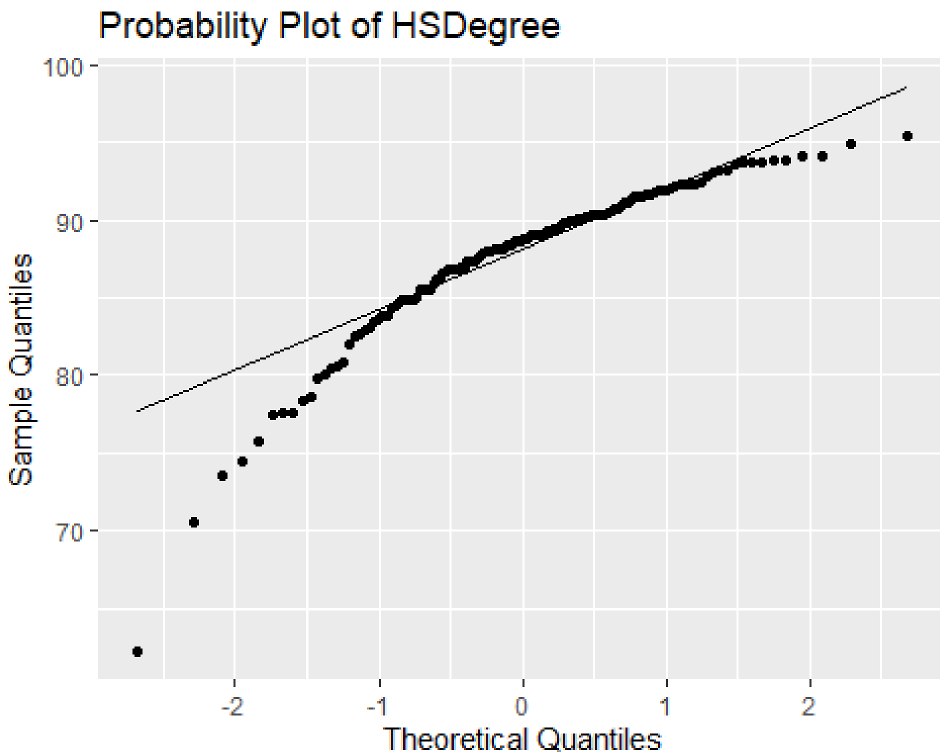

```r
# Based on the histogram, the data distribution appears to be unimodal, as
there is one clear peak in the data.

# The distribution is not perfectly symmetrical.

# The distribution is not perfectly bell-shaped.

# The distribution is not perfectly normal.

# The distribution is slightly skewed to the right.

#To add a normal curve to the histogram, we can use the stat_function()
function in ggplot2

library(ggplot2)
```

```r
# Load the data
acs_data <- read.csv("acs-14-1yr-s0201.csv")

# Create the histogram with a bin size of 2 and a normal curve
ggplot(acs_data, aes(x = HSDegree)) +
  geom_histogram(binwidth = 2, fill = "blue", color = "black") +
  stat_function(fun = dnorm, args = list(mean = mean(acs_data$HSDegree), sd =
sd(acs_data$HSDegree)),
                color = "red", linewidth = 1) +
  ggtitle("Histogram of HSDegree with Normal Curve") +
  xlab("Number of individuals with a Bachelor's degree or higher") +
  ylab("Count")
```




```r
# Based on the histogram and normal curve, it appears that the distribution
is not perfectly normal, as there is a slight skew to the right and the
distribution is not perfectly bell-shaped.

## Create a Probability Plot of the HSDegree variable.

library(ggplot2)

# Load the data
acs_data <- read.csv("acs-14-1yr-s0201.csv")

# Create the probability plot of the HSDegree variable
```

```r
ggplot(acs_data, aes(sample = HSDegree)) +
  stat_qq() +
  stat_qq_line() +
  ggtitle("Probability Plot of HSDegree") +
  xlab("Theoretical Quantiles") +
  ylab("Sample Quantiles")
```




```r
## Answer the following questions based on the Probability Plot:

# Based on the probability plot, the distribution appears to be approximately
normal, as most of the points fall close to the reference line. There are
some slight deviations from the line, particularly at the tails of the
distribution.

# There may be a slight skew to the right in the distribution, as some of the
points deviate from the reference line at the right tail of the plot.
However, the deviations are relatively small.

## 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 4.2.3
```

```r
# Load the data
acs_data <- read.csv("acs-14-1yr-s0201.csv")

# Calculate summary statistics for HSDegree
stat.desc(acs_data$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
```

```r
## 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 measures the degree of asymmetry in a distribution, with a value
of 0 indicating a perfectly symmetrical distribution. Positive and negative
values indicate a right or left skew, respectively.

# Kurtosis measures the degree of peakedness in a distribution, with a value
of 0 indicating a normal distribution. Positive values indicate a more peaked
or "heavy-tailed" distribution, while negative values indicate a more flat or
"light-tailed" distribution.

# Z-scores, also known as standard scores, indicate how many standard
deviations a data point is from the mean of a distribution. A z-score of 0
indicates a data point that is equal to the mean, while positive and negative
z-scores indicate data points that are above or below the mean, respectively.

# Changes in the sample size can affect the interpretation of these
statistics. With larger sample sizes, it is more likely that the distribution
will be closer to a normal distribution and that the skewness and kurtosis
values will be closer to 0. With smaller sample sizes, there is greater
variability in the data and the distribution may be more skewed or have a
different shape than a normal distribution. Therefore, it's important to
consider the sample size when interpreting these statistics and to use
caution when making inferences based on a small sample size.
```