# 2014 American Community Survey Exercise 3.2 Week 3

Maxim Bilenkin

2024-12-09

“‘markdown

```
options(repos = c(CRAN = "https://cloud.r-project.org"))

# Loading the file
loadedfile <- read.csv("C:/Users/maxim/Downloads/acs-14-1yr-s0201.csv")
```

```
# Displaying the data structure
invisible(capture.output(str(loadedfile)))
```

```
library(knitr)
library(kableExtra)

# Extracting variables, data type from loaded file and manually writing intent.
fields_info <- data.frame(
  Field = names(loadedfile),
  Data_Type = sapply(loadedfile, class),
  Intent = c(
    "Unique identifier for each row",
    "Simplified version of Id column that only list last four digits.",
    "Identifies geographic location",
    "Identifies specific population of the study",
    "Label of the studied population group",
    "Number of persons reported their race",
    "Percent holding high school diploma",
    "Percent holding Bachelors degree"
  )
)

# 1) List the name of each field and what you believe the data type and intent
#    is of the data included in each field.
# Print data in friendly clean format
kable(fields_info, align='l', row.names = FALSE,
      caption = "Field/Data Type/Intent") %>%
  kable_styling(position = "left", full_width = FALSE,
                bootstrap_options = c("striped", "hover", "condensed"))%>%
    column_spec(1:3, bold = FALSE, extra_css = "text-align: left;")
```

Table 1: Field/Data Type/Intent

| Field | Data_Type | Intent |
|---|---|---|
| Id | character | Unique identifier for each row |
| Id2 | integer | Simplified version of Id column that only list last four digits. |

| | | |
|---|---|---|
| Geography | character | Identifies geographic location |
| PopGroupID | integer | Identifies specific population of the study |
| POPGROUP.display.label | character | Label of the studied population group |
| RacesReported | integer | Number of persons reported their race |
| HSDegree | numeric | Percent holding high school diploma |
| BachDegree | numeric | Percent holding Bachelors degree |

```
# 2) Run the following functions and provide the results: str(); nrow(); ncol()

# Displaying data structure.
str(loadedfile, width=80, strict.width="cut")
```

```
## 'data.frame':    136 obs. of  8 variables:
##  $ Id                     : chr  "0500000US01073" "0500000US04013" "0500000US0"..
##  $ Id2                    : int  1073 4013 4019 6001 6013 6019 6029 6037 6059 6..
##  $ Geography              : chr  "Jefferson County, Alabama" "Maricopa County,"..
##  $ PopGroupID             : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ POPGROUP.display.label: chr  "Total population" "Total population" "Total "..
##  $ RacesReported          : int  660793 4087191 1004516 1610921 1111339 965974 ..
##  $ HSDegree               : num  89.1 86.8 88 86.9 88.8 73.6 74.5 77.5 84.6 80...
##  $ BachDegree             : num  30.5 30.2 30.8 42.8 39.7 19.7 15.4 30.3 38 20...
```

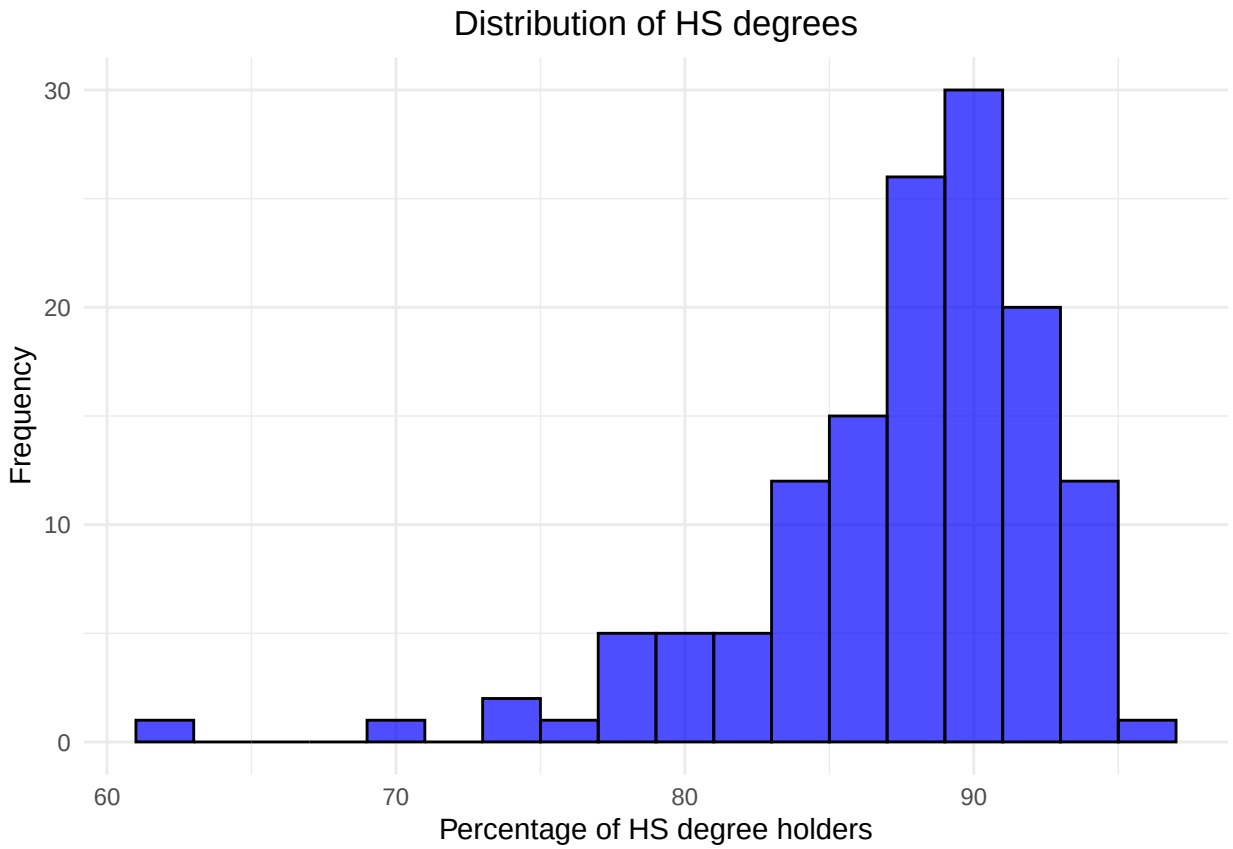
```
# Displaying number of rows.
nrow(loadedfile)
```

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

```
# Displaying number of columns.
cat(ncol(loadedfile))
```

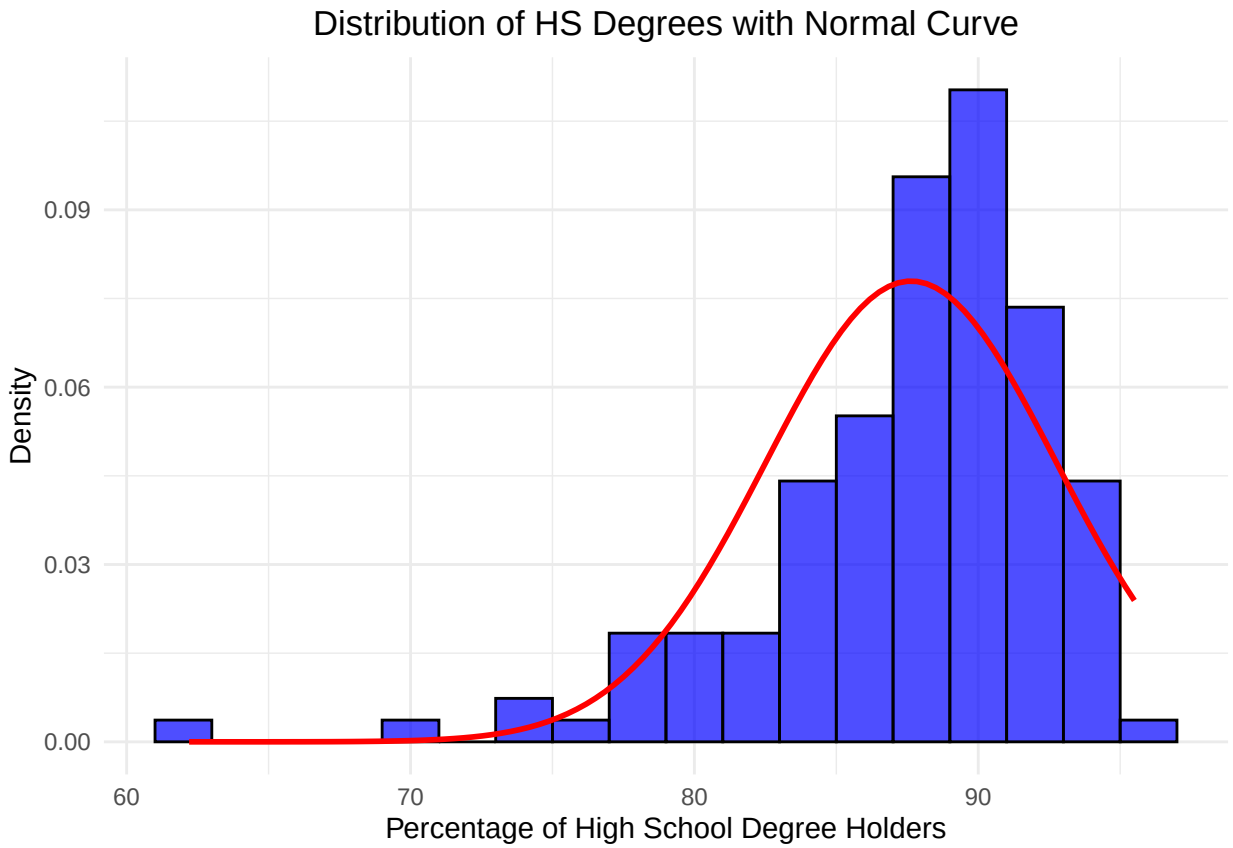
```
## 8
```

```
# 3) Create a Histogram of the HSDegree variable using the ggplot2 package.
# installing ggplot2 package.
library(ggplot2)

# Setting a bin size for the Histogram.
bin_size <- 2

# Creating a histogram.
histogram_plot <- ggplot(loadedfile, aes(x = HSDegree)) +
    geom_histogram(binwidth = bin_size, fill = "blue", color = "black",
                   alpha = 0.7) +
    labs(title = "Distribution of HS degrees",
         x = "Percentage of HS degree holders",
         y = "Frequency") +
    theme_minimal() + theme(plot.title = element_text(hjust = 0.5))

print(histogram_plot)
```



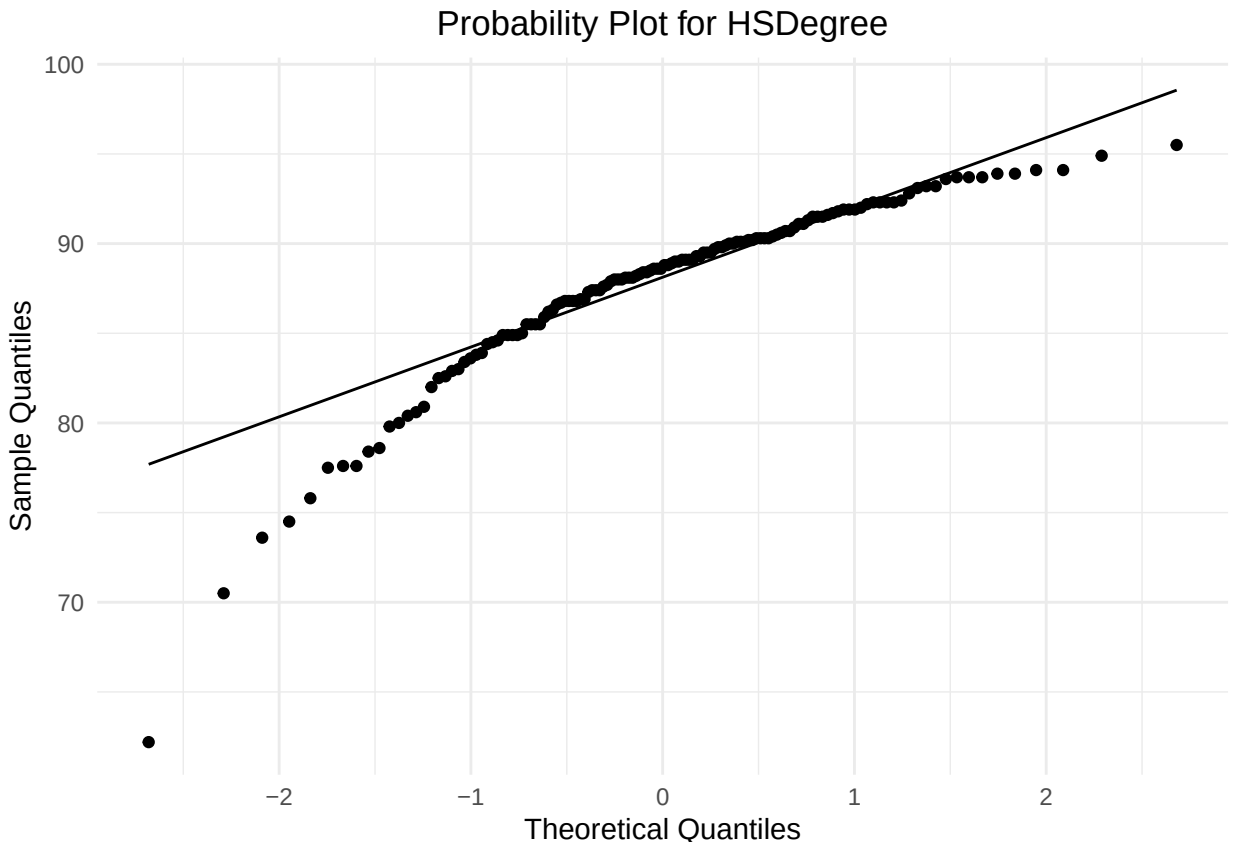

```
# 4) Answer the following questions based on the Histogram produced:
# Based on what you see in this histogram, is the data distribution unimodal?

# Answer: This distribution seems unimodal because there is only one highest
#         pick at 90% on the x-axis and 30 frequency on the y-axis.

# Is it approximately symmetrical?

# Answer: This histogram not approximately symmetrical. More distribution on
#         the left side further away from the mean.

# Is it approximately bell-shaped?

# Answer: This distribution not approximately bell-shaped.

# Is it approximately normal?

# Answer: This distribution not approximately normal.

# If not normal, is the distribution skewed? If so, in which direction?

# Answer: This is a negatively skewed distribution. The left tail is longer
#         than the right tail.

# Include a normal curve to the Histogram that you plotted.
mean_hsdegree <- mean(loadedfile$HSDegree, na.rm = TRUE)
sd_hsdegree <- sd(loadedfile$HSDegree, na.rm = TRUE)
```

```r
ggplot(loadedfile, aes(x = HSDegree)) +
    geom_histogram(aes(y = after_stat(density)), binwidth = bin_size,
    fill = "blue", color = "black", alpha = 0.7) +
    stat_function(fun = dnorm,
    args = list(mean = mean_hsdegree, sd = sd_hsdegree), color = "red",
    linewidth = 1) + labs(title ="Distribution of HS Degrees with Normal Curve",
     x = "Percentage of High School Degree Holders",
     y = "Density") +
    theme_minimal() + theme(plot.title = element_text(hjust = 0.5))
```




```r
# Explain whether a normal distribution can accurately be used as a model
# for this data.

# Answer: The normal distribution model cannot be accurately used for this data
#         because it is negatively skewed to the left.


# 5) Create a Probability Plot of the HSDegree variable.
probability_plot <- ggplot(loadedfile, aes(sample = HSDegree)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Probability Plot for HSDegree",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles") +
  theme_minimal() +
```

```
  theme(plot.title = element_text(hjust = 0.5))

# Print the Q-Q plot
print(probability_plot)
```




```
# 6) Answer the following questions based on the Probability Plot:
#    Based on what you see in this probability plot, is the distribution
#    approximately normal? Explain how you know.

# Answer: The probability plot doesn't look approximately normal because
#         the end tails deviate substantially from the 45 degree straight line
#         that follows hypothetical normal distribution.


#    If not normal, is the distribution skewed? If so, in which direction?
#    Explain how you know.

# Answer: The distribution is negatively skewed. This plot only confirms that
#          this is a left-sided distribution because the tails deviate toward
#          down direction on both side from the straight line. If the data from
#          American survey followed a normal distribution then all or most
#          points would be plotted on the line or very close to the line. We can
#          see from the image the area between -1 and 1.5 on the x-axis follows
#          approximately normal distribution since all the points plot on or
#          very close to the line. But outside of the area the data points plot
#          away down from the line and the distribution exhibits negative
#          skewness.
```

```r
# 7) 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.

# Install and load the pastecs package
if (!require(pastecs)) {
  install.packages("pastecs")
}
library(pastecs)

# Calculating descriptive statistics for HSDegree
stats <- stat.desc(loadedfile$HSDegree, basic = TRUE, desc = TRUE, norm = TRUE)

# Converting results to a data frame and storing to the variable.
stats_df <- as.data.frame(stats)

# Creating and printing data in a readable formatted table vertically.
kable(stats_df, caption = "Descriptive Statistics for HSDegree", digits = 2) %>%
    kable_styling(full_width = FALSE, position = "left",
                  bootstrap_options = c("striped", "hover", "condensed")) %>%
    column_spec(1, width = "5cm")
```

Table 2: Descriptive Statistics for HSDegree

| | stats |
|---|---|
| nbr.val | 136.00 |
| nbr.null | 0.00 |
| nbr.na | 0.00 |
| min | 62.20 |
| max | 95.50 |
| range | 33.30 |
| sum | 11918.00 |
| median | 88.70 |
| mean | 87.63 |
| SE.mean | 0.44 |
| CI.mean.0.95 | 0.87 |
| var | 26.19 |
| std.dev | 5.12 |
| coef.var | 0.06 |
| skewness | -1.67 |
| skew.2SE | -4.03 |
| kurtosis | 4.35 |
| kurt.2SE | 5.27 |
| normtest.W | 0.88 |
| normtest.p | 0.00 |

```r
# Calculating z-scores for HSDegree.
loadedfile$z_scores <- (loadedfile$HSDegree - mean_hsdegree) / sd_hsdegree

# Displaying the first 6 calculated z-scores.
head(loadedfile$z_scores, 6)
```

```
## [1]  0.28676516 -0.16263435  0.07183496 -0.14309524  0.22814783 -2.74179676
```

```
# In several sentences provide an explanation of the result produced for skew,
# kurtosis, and z-scores.

# Answer:  Based on the result the skewness is -1.67 which means this is a
#          negatively skewed distribution. It means most data points concentrate
#          on the higher end with fewer values on the lower end which makes long
#          tail on the lower end (left side). The value for kurtosis is 4.35.
#          It means it is a leptokurtic distribution because the value above 3.
#          This means the data has significant outliers.From the calculated and
#          first 6 displayed z-scores we can see that the first four z-scores
#          are near or very close to zero which mean they are not deviating much
#          from the mean. However, the last or the sixth value of -2.74179676 is
#          more than two standard derivations from below the mean which is a
#          significant deviation. This means that the data point is a outlier on
#          the left side. This type of outliers make distribution negatively
#          skewed.

# In addition, explain how a change in the sample size may change your
# explanation?

# Answer: Increasing sample size will stabilize the distribution and make it
#         more reliable. It will improve accuracy and reduce impact of outliers.
```