DESCRIBING DATA

Research Methods in Psychology I & II ▪ Department of Psychology ▪ Colorado State University

BY THE END OF THIS SECTION YOU WILL HAVE A REFRESHED MEMORY ON:

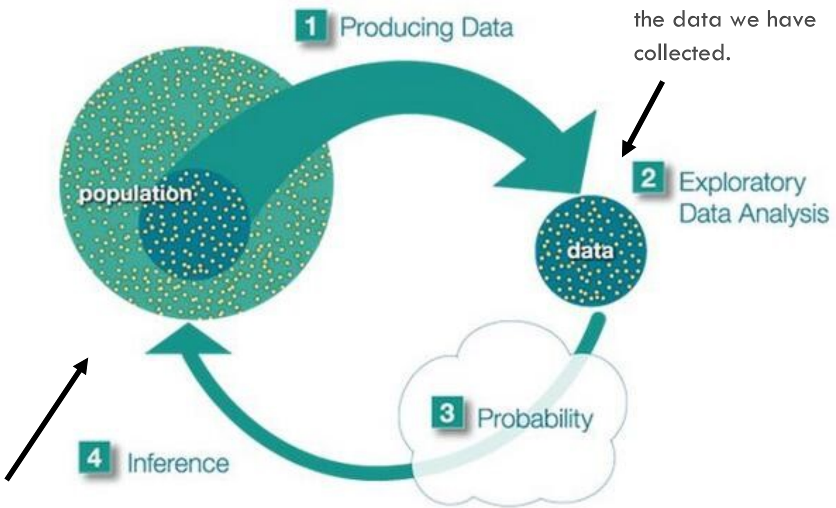
1. The difference between descriptive and inferential statistics.
2. Methods for describing central tendency.
3. Methods for describing variability.
4. The normal distribution, z-scores, and the empirical rule.

Descriptive & Inferential Statistics

Descriptive statistics are used to organize, summarize, simplify, and present data — usually about a sample that we have collected from a population. Inferential statistics are used to generalize from our sample to the larger population, to test hypotheses, and to make predictions (and understand the accuracy of those predictions).

In this unit we will focus on descriptive statistics, in Unit 3, we will focus on inferential statistics.

Descriptive statistics are used to describe the data we have collected.



Inferential statistics are used to make inferences or generalizations about the population based on our sample.

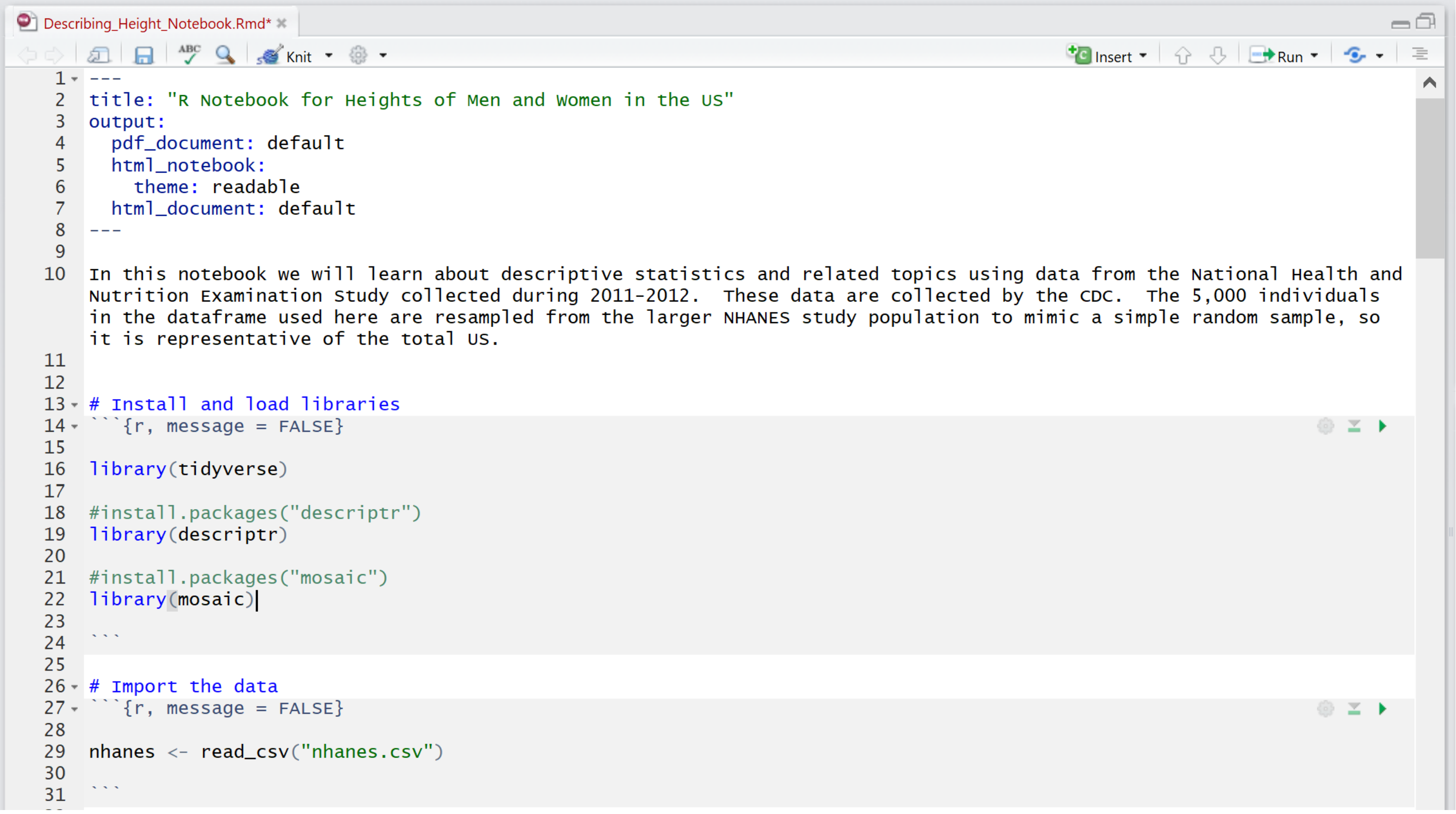
Image produced by Carnegie Melon University, Open Learning Institute.

Height of Adult Males and Females in the US

In this section we will learn about descriptive statistics and related topics using data from the National Health and Nutrition Examination Study that were collected during 2011-2012 by the CDC. We will consider the height of the participants in the study. The 5,000 individuals in the dataframe are resampled from the larger NHANES study population to mimic a simple random sample, so the data are representative of the total US population.

The data are in the Unit 2 folder in dropbox, the file is called nhanes.csv. There is also a file called nhanes data dictionary which summarizes these data. Please copy both of these into your MyClassActivities folder. Next, open up your MyClassActivities project in RStudio, and then create a new R Notebook. Call this Notebook: Describing_Height_Notebook.

To begin our notebook, add the following two code chunks. Notice that you need to install two new packages — descriptr and mosaic. Once you execute this code chunk, you can put a hashtag in front of install.packages so R doesn't reinstall each time you execute this chunk.



~~~
---
title: "R Notebook for Heights of Men and Women in the US"
output:
  pdf_document: default
  html_notebook:
    theme: readable
  html_document: default
---

In this notebook we will learn about descriptive statistics and related topics using data from the National Health and
Nutrition Examination Study collected during 2011-2012.  These data are collected by the CDC.  The 5,000 individuals
in the dataframe used here are resampled from the larger NHANES study population to mimic a simple random sample, so
it is representative of the total US.


# Install and load libraries
```{r, message = FALSE}

library(tidyverse)

#install.packages("descriptr")
library(descriptr)

#install.packages("mosaic")
library(mosaic)

~~~

Import the data
```{r, message = FALSE}

nhanes <- read_csv("nhanes.csv")

```
```
```

Prepare the Dataframe for Exploration

Let's begin by transforming Height, which is expressed in cm, to a new variable called ht_inches that is expressed in inches. We will also create a new variable called sex, that is a copy of gender, and we will specify it to be a factor. We are interested here in only adults (i.e., people who are likely done growing), so we will use filter to choose only people 20 years or older. Last, we will subset to keep just the variables we will use today. Let's do all of this with a pipe.

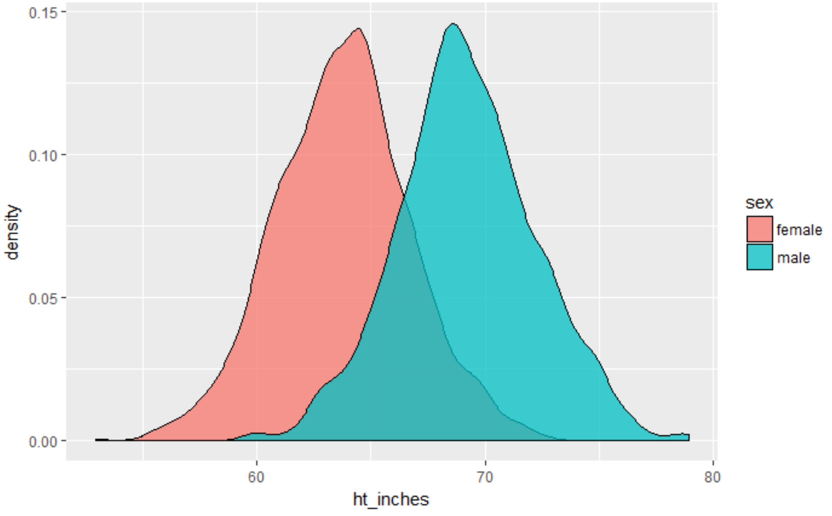
```
height <- nhanes %>%
  filter(Age >= 20) %>%
  mutate(ht_inches = Height/2.54, sex = factor(Gender)) %>%
  select(ht_inches, sex) %>%
  na.omit()
```

na.omit() takes the height dataframe and removes all rows with missing data on any of the variales — i.e., ht_inches and sex, since we subsetted to include just these variables in the prior step.

Create a Density Plot of Height by Sex

As a first step in describing height, we will create a density plot of height, and group our plot by sex. This will give us an idea of the distribution of height among adult males and females.

```
ggplot(height, aes(x = ht_inches, group = sex, fill = sex)) +
    geom_density(alpha = .75)
```



As we would expect, males are, on average, taller than females.

Descriptive Statistics: Central Tendency & Dispersion

When we approach a new analysis, we typically begin by examining the distribution of our variables. For a continuous variable (like the height of people), assessment of central tendency and dispersion is a good place to start.

A measure of central tendency captures a central or typical value for the distribution of the variable. Common measures of central tendency include the arithmetic mean, the median, and the mode.

Dispersion describes the extent to which a distribution is stretched or squeezed. Common measures of dispersion are the variance, the standard deviation, the range, and interquartile range.

For a thorough display of measures of central tendency and dispersion, I like the descriptr package.

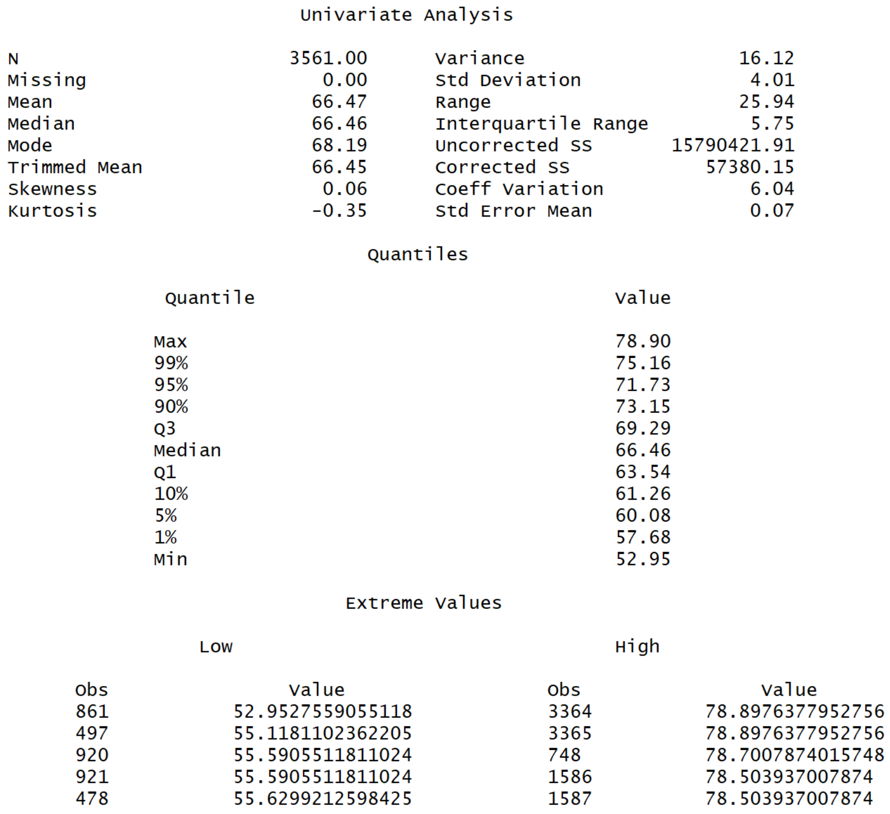
Let's use the summary_stats function to obtain descriptive statistics for ht_inches.

In many R functions, a "data =" argument is not available, and in these cases, we need to specify the name of the dataframe and the variable of interest inside the dataframe using the following convention:

data_frame_name$variable_name

Where data_frame_name is the name of your dataframe, and variable_name is the name of your variable.

```
summary_stats(height$ht_inches)
```

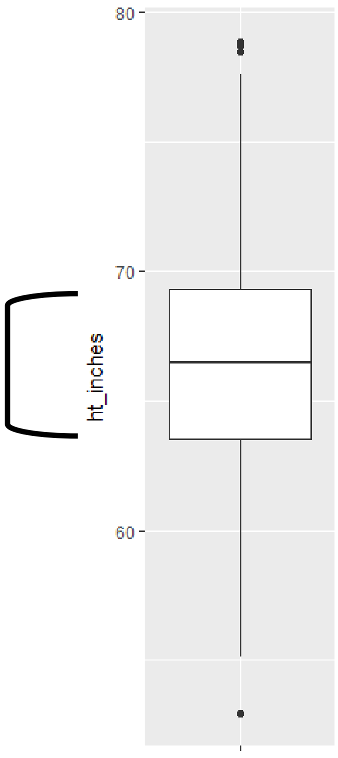


```
                         Univariate Analysis

N                       3561.00      Variance                 16.12
Missing                    0.00      Std Deviation             4.01
Mean                      66.47      Range                    25.94
Median                    66.46      Interquartile Range       5.75
Mode                      68.19      Uncorrected SS     15790421.91
Trimmed Mean              66.45      Corrected SS          57380.15
Skewness                   0.06      Coeff Variation           6.04
Kurtosis                  -0.35      Std Error Mean            0.07

                              Quantiles

             Quantile                              Value

            Max                                    78.90
            99%                                    75.16
            95%                                    71.73
            90%                                    73.15
            Q3                                     69.29
            Median                                 66.46
            Q1                                     63.54
            10%                                    61.26
            5%                                     60.08
            1%                                     57.68
            Min                                    52.95

                            Extreme Values

                Low                                High

      Obs              Value                 Obs            Value
      861        52.9527559055118            3364     78.8976377952756
      497        55.1181102362205            3365     78.8976377952756
      920        55.5905511811024            748      78.7007874015748
      921        55.5905511811024            1586     78.503937007874
      478        55.6299212598425            1587     78.503937007874
```

Measures of Central Tendency

Mean	The average value. The mean can be highly affected by outliers.
Median	The central value of an ordered distribution.
Mode	The value that occurs most often.
Trimmed Mean	Extreme cases are discarded, and the average is computed on the remainder. The descriptr package trims the lowest 5% of cases and the highest 5% of cases.

A boxplot displays many of the key descriptive statistics.



Interquartile Range = Q3 - Q1 = 5.75

Q3 = 69.29

Q2, Median = 66.46

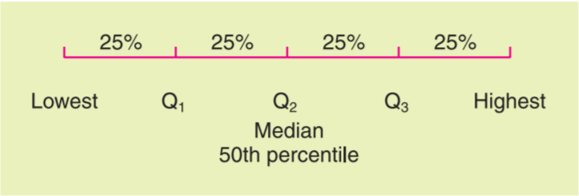
Q1 = 63.54

In R, the default is for the whiskers to extend to the extremes (min and max), but no further than 1.5 times the IQR (i.e., 1.5 ▪ IQR = 1.5 ▪ 5.75 = 8.63) above Q3 or 1.5 times the IQR below Q1. In other words, the upper whisker is located at the **smaller** of the maximum y value and Q3 + 1.5 IQR, and the lower whisker is located at the **larger** of the smallest y value and Q1 – 1.5 IQR. You can change this behavior with the range argument in geom_boxplot. An outlier (the individual points that you see) is a value that is outside of the defined whiskers. Therefore the boxplot also displays that minimum and maximum values.

Measures of Dispersion

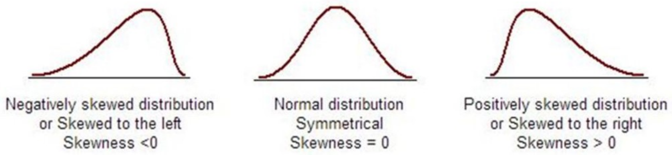
Range	The difference between the largest and smallest value (max - min = range)
Quantile Scores	Quantiles are the values of a variable that divide a distribution into equal parts. Quartiles are commonly used. Quartiles divide the distribution into 4 equals parts. The first quartile Q1 is the 25th percentile, the second quartile Q2 is the median, and the third quartile Q3 is the 75th percentile. See the Quartiles figure below.
Variance	The average of the squared differences between each value and the mean. It captures how far a set of numbers are spread out from the mean.
Standard Deviation (SD)	The square root of the variance.
Uncorrected SS (Sum of Squares)	Sum of the squared values.
Corrected SS	Sum of the squared differences between each value and the mean.
Coefficient of Variation	The ratio of the standard deviation to the mean, expressed as a percentage, so (SD/Mean) ▪ 100. It captures the extent of variability of the variable in relation to the mean.
Skewness	Measures the degree and direction of asymmetry in the distribution of the variable. A symmetric distribution has a skewness of 0. A distribution that is skewed to the left (i.e., the mean is less than the median) has a negative skewness, while a distribution that is skewed to the right has a positive skewness. See skewness figure below.
Kurtosis	Measures the heaviness of the tails of a distribution. Given the way kurtosis is scaled here (type 1), a normal distribution has kurtosis 0. Kurtosis is positive if the tails are heavier than for a normal distribution (leptokurtic) and negative if the tails are lighter than for a normal distribution (platykurtic). See Kurtosis figure below.
Standard Error of the Mean	The estimated standard deviation of the sampling distribution. This isn't a descriptive statistic, but rather an inferential statistic. We'll cover this in the next unit.

Quartiles



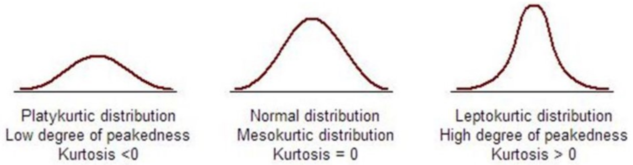
Skewness

The coefficient of Skewness is a measure for the degree of symmetry in the variable distribution.



Kurtosis

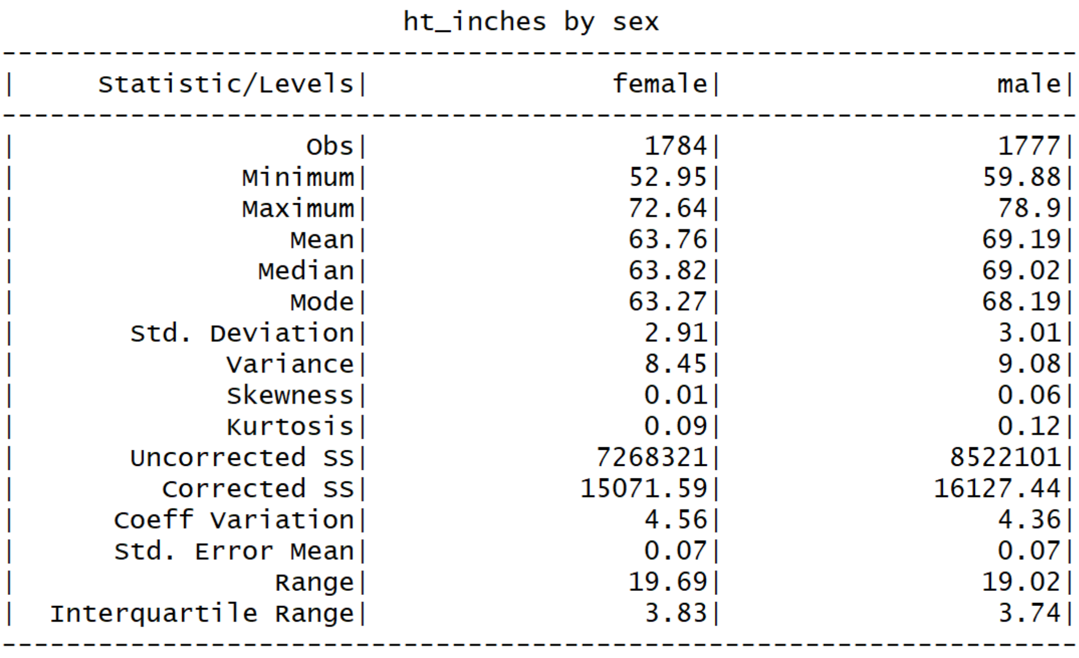
The coefficient of Kurtosis is a measure for the degree of peakedness/flatness in the variable distribution.



Descriptive Statistics for Height by Sex

Let’s request descriptive statistics for males and females separately.

```
group_summary(height$ht_inches, fvar = height$sex)
```



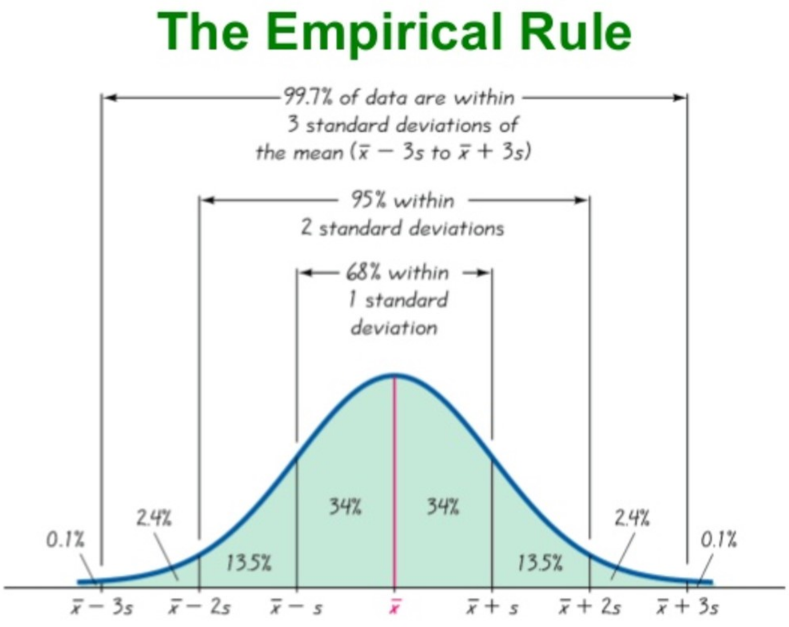
```
                         ht_inches by sex
-------------------------------------------------------------------
|     Statistic/Levels|               female|                 male|
-------------------------------------------------------------------
|                  Obs|                 1784|                 1777|
|              Minimum|                52.95|                59.88|
|              Maximum|                72.64|                 78.9|
|                 Mean|                63.76|                69.19|
|               Median|                63.82|                69.02|
|                 Mode|                63.27|                68.19|
|       Std. Deviation|                 2.91|                 3.01|
|             Variance|                 8.45|                 9.08|
|             Skewness|                 0.01|                 0.06|
|             Kurtosis|                 0.09|                 0.12|
|       Uncorrected SS|              7268321|              8522101|
|         Corrected SS|             15071.59|             16127.44|
|      Coeff Variation|                 4.56|                 4.36|
|      Std. Error Mean|                 0.07|                 0.07|
|                Range|                19.69|                19.02|
|  Interquartile Range|                 3.83|                 3.74|
-------------------------------------------------------------------
```

What is a Normal Distribution and Why is it Important?

A random variable with a Gaussian (e.g., bell-shaped) distribution is said to be normally distributed. A normal distribution is a symmetrical distribution. The mean, median and mode are in the same location and at the center of the distribution. The empirical rule provides a quick estimate of the spread of data in a normal distribution given the mean and standard deviation. Specifically, the empirical rule states that for a normal distribution:

- 68% of the data will fall within about one standard deviation of the mean.
- 95% of the data will fall within about two standard deviations of the mean.
- Almost all (99.7%) of the data will fall within about three standard deviations of the mean.

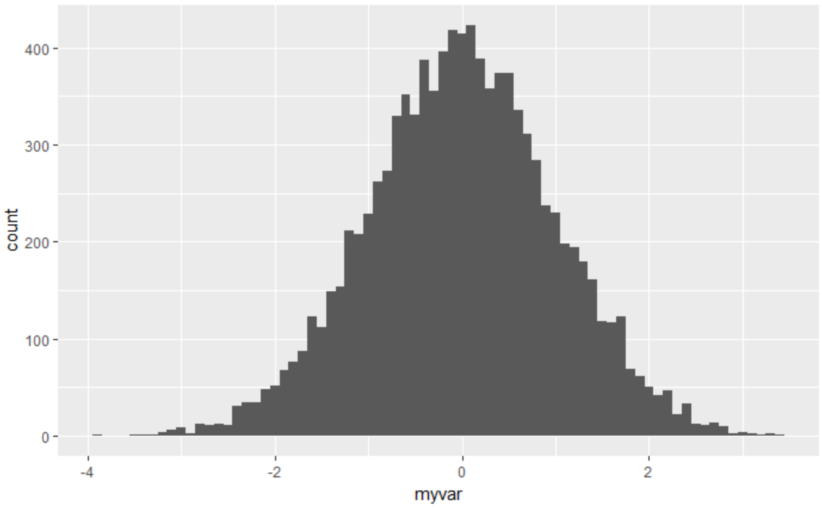
The empirical rule helps us to gain a sense of the distribution of scores in our dataframe. For example, if all we knew was that the average height for a female is 63.76 inches, with a standard deviation of 2.91, we would know that about 95% of all females are between 57.95 inches and 69.58 inches (that is, 63.76 ± 2 ▪ 2.91). This premise will serve as the basis for the inferential statistics that we will cover this semester, so it is important to understand.

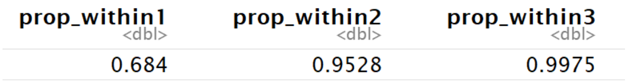

The Empirical Rule
99.7% of data are within 3 standard deviations of the mean ($\bar{x} - 3s$ to $\bar{x} + 3s$)
95% within 2 standard deviations
68% within 1 standard deviation
0.1%
2.4%
13.5%
34%
34%
13.5%
2.4%
0.1%
$\bar{x} - 3s$
$\bar{x} - 2s$
$\bar{x} - s$
$\bar{x}$
$\bar{x} + s$
$\bar{x} + 2s$
$\bar{x} + 3s$

Simulate and Explore a Normal Distribution

We can use the rnorm function in R to simulate data according to a normal distribution. Here, we create a sample of size 10,000, with a mean of 0 and a standard deviation of 1. We can plot the data to confirm that the distribution is normal. Finally, we can compute the number of cases within 1 SD, 2 SDs, and 3SDs from the mean, and then summarize these new variables to see what proportion of cases fall into each range.

```
set.seed(12345)
myvar <- rnorm(n=10000, m=0, sd=1)
example <- data.frame(myvar)

ggplot(example, aes(x = myvar)) +
  geom_histogram(binwidth = .1)

example <- example %>%
  mutate(within1 = ifelse(myvar <= 1 & myvar >= -1, 1, 0),
         within2 = ifelse(myvar <= 2 & myvar >= -2, 1, 0),
         within3 = ifelse(myvar <= 3 & myvar >= -3, 1, 0))

summarize(example, prop_within1 = mean(within1), prop_within2 = mean(within2), prop_within3 = mean(within3))
```



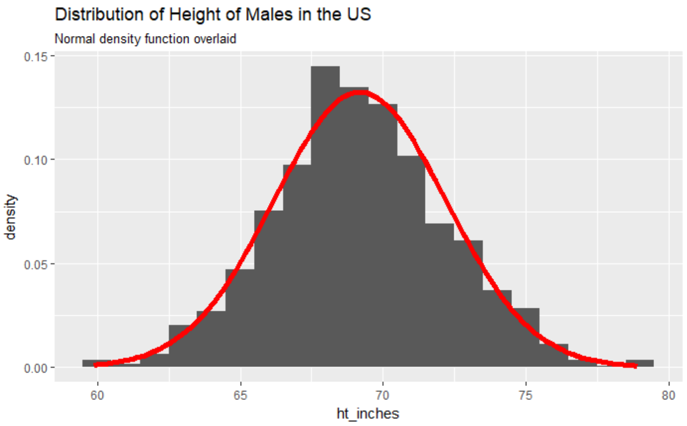


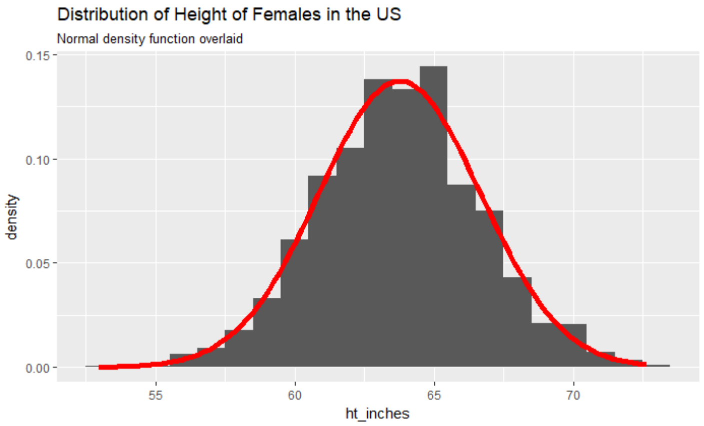
```
prop_within1  prop_within2  prop_within3
       <dbl>         <dbl>         <dbl>
       0.684        0.9528        0.9975
```

Do these match what we would expect based on the empirical rule?

Compare our Distributions to a Normal Distribution Function

```
# for males
males <- filter(height, sex == "male")

ggplot(males, aes(x = ht_inches)) +
  geom_histogram(aes(y = ..density..), binwidth = 1) +
  stat_function(fun = dnorm,
               args = list(mean = mean(males$ht_inches), sd = sd(males$ht_inches)),
               lwd = 2,
               col = 'red') +
   labs(title = "Distribution of Height of Males in the US", subtitle = "Normal density function overlaid")


# for females
females <- filter(height, sex == "female")

ggplot(females, aes(x = ht_inches)) +
  geom_histogram(aes(y = ..density..), binwidth = 1) +
  stat_function(fun = dnorm,
               args = list(mean = mean(females$ht_inches), sd = sd(females$ht_inches)),
               lwd = 2,
               col = 'red') +
   labs(title = "Distribution of Height of Females in the US", subtitle = "Normal density function overlaid")
```





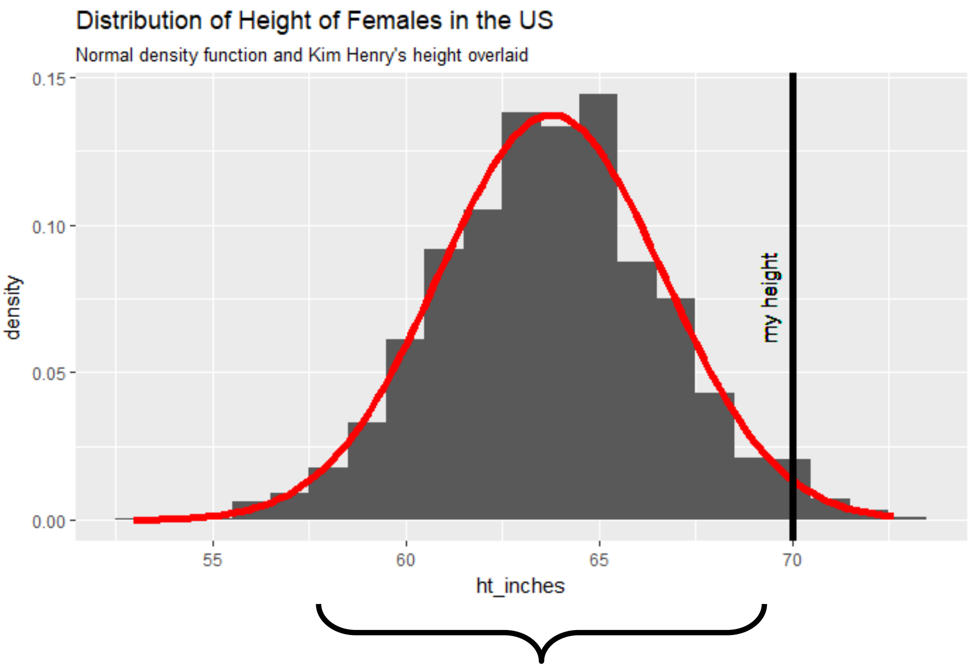
Both of the height distributions match a normal distribution very well. When this is the case, then we can use the principles of the empirical rule to consider the expected number of cases in various parts of the distribution. For example:

What percentage of males do we expect to be shorter than 65 inches?/What's the probability that we would randomly select a male from the population who is shorter than 65 inches?

What percentage of females do we expect to be between 60 and 70 inches tall?/What's the probability that we would randomly select a female from the population who is between 60 and 70 inches tall?

Where are You on the Distribution of Height?

```
# for females
females <- filter(height, sex == "female")
ggplot(females, aes(x = ht_inches)) +
  geom_histogram(aes(y = ..density..), binwidth = 1) +
  stat_function(fun = dnorm,
                args = list(mean = mean(females$ht_inches), sd = sd(females$ht_inches)),
                lwd = 2,
                col = 'red') +
  geom_vline(xintercept = 70, colour = "black", lwd = 2) +
  geom_text(aes(x=70, label="my height", y=.075), colour="black", angle=90, vjust = -1, text=element_text(size=11)) +
  labs(title = "Distribution of Height of Females in the US", subtitle = "Normal density function and Kim Henry's height overlaid")
```



The empirical rule tells us that about 95% of females will be within 2 standard deviations of the mean, so 57.95 inches to 69.58 inches. Only about 2.5% of the population is likely to be shorter than 57.95, and only about 2.5% of the population is likely to be taller than 69.58 inches.

Mean ± 2 ▪ SD = 63.76 ± 2 ▪ 2.91 = 57.95, 69.58

Raw Scores and Z-Scores

A Z-Score is a standardized score based on some raw variable (e.g., height in inches, ht_inches) that has been transformed by first subtracting the mean, and then dividing by the standard deviation. Let's consider Kim Henry's height (70 inches):

Z-Score = (70 - 63.76) / 2.91 = 2.15

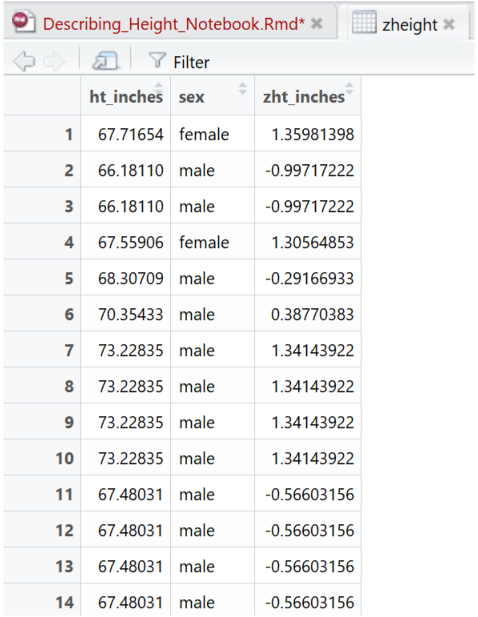
So, Kim Henry is 2.15 standard deviations above the mean height for females. What is your Z-Score for height?

```
mean_m <- mean(males$ht_inches)
sd_m <- sd(males$ht_inches)

mean_f <- mean(females$ht_inches)
sd_f <- sd(females$ht_inches)

myzscore <- (70 - mean_f)/sd_f
```

```
[1] 2.145213
```

We can use the mosaic package to convert the raw scores to Z-Scores in the NHANES data. Because we use group_by first, then the sex specific mean and sd is used to form the Z-Scores for males and females respectively.

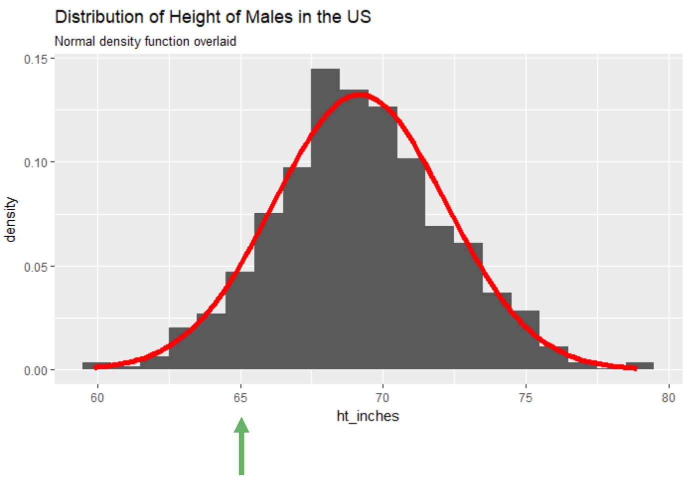
```
zheight <- height %>%
  group_by(sex) %>%
  mutate(zht_inches = zscore(ht_inches)) %>%
  ungroup()
```

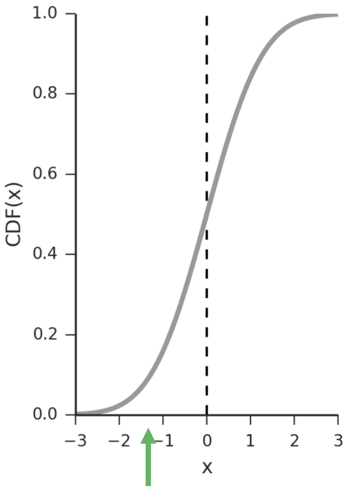


	ht_inches	sex	zht_inches
1	67.71654	female	1.35981398
2	66.18110	male	-0.99717222
3	66.18110	male	-0.99717222
4	67.55906	female	1.30564853
5	68.30709	male	-0.29166933
6	70.35433	male	0.38770383
7	73.22835	male	1.34143922
8	73.22835	male	1.34143922
9	73.22835	male	1.34143922
10	73.22835	male	1.34143922
11	67.48031	male	-0.56603156
12	67.48031	male	-0.56603156
13	67.48031	male	-0.56603156
14	67.48031	male	-0.56603156

The Cumulative Distribution Function

The cumulative distribution function (cdf) is the probability that a variable takes on a value less than or equal to x. We can use the cdf of our normal distribution to ask a question such as "What is the probability that a randomly selected male from the population will be shorter the 65 inches tall?"





```
pnorm(65, mean = mean_m, sd = sd_m, lower.tail=TRUE )
```

```
[1] 0.0823984
```

```
male65 <- (65 - mean_m)/sd_m
male65
```

```
[1] -1.389118
```

What is the probability that a randomly selected male is **taller** than 80 inches?

```
pnorm(80, mean = mean_m, sd = sd_m, lower.tail=FALSE )
```

```
[1] 0.0001662314
```

What is the probability that a randomly selected male is **between** 70 and 75 inches?

```
pnorm(75, mean = mean_m, sd = sd_m, lower.tail=TRUE) - pnorm(70, mean = mean_m, sd = sd_m, lower.tail=TRUE)
```

```
[1] 0.3666908
```

We can use qnorm to do the inverse. Here we might ask "What height does a male need to be, to be in the top 10%?"

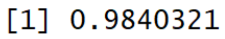
```
qnorm(.10, mean = mean_m, sd = sd_m, lower.tail=FALSE )
```

```
[1] 73.04788
```

Let’s Finish with Your Height & Z-Score

What is the probability that a randomly selected adult of your same sex will be shorter than you?

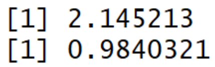
```
pnorm(70, mean = mean_f, sd = sd_f, lower.tail=TRUE )
```



```
[1] 0.9840321
```

You can obtain the same answer using Z-Scores (notice the change to mean and sd)

```
myzscore <- (70 - mean_f)/sd_f

myzscore


pnorm(myzscore, mean=0, sd=1, lower.tail = TRUE)
```



```
[1] 2.145213
[1] 0.9840321
```