R Notebook for Heights of Men and Women in the US

Contents

In this notebook we will learn aboput 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

```
rm(list=ls(all=TRUE))

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

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

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

Import the data

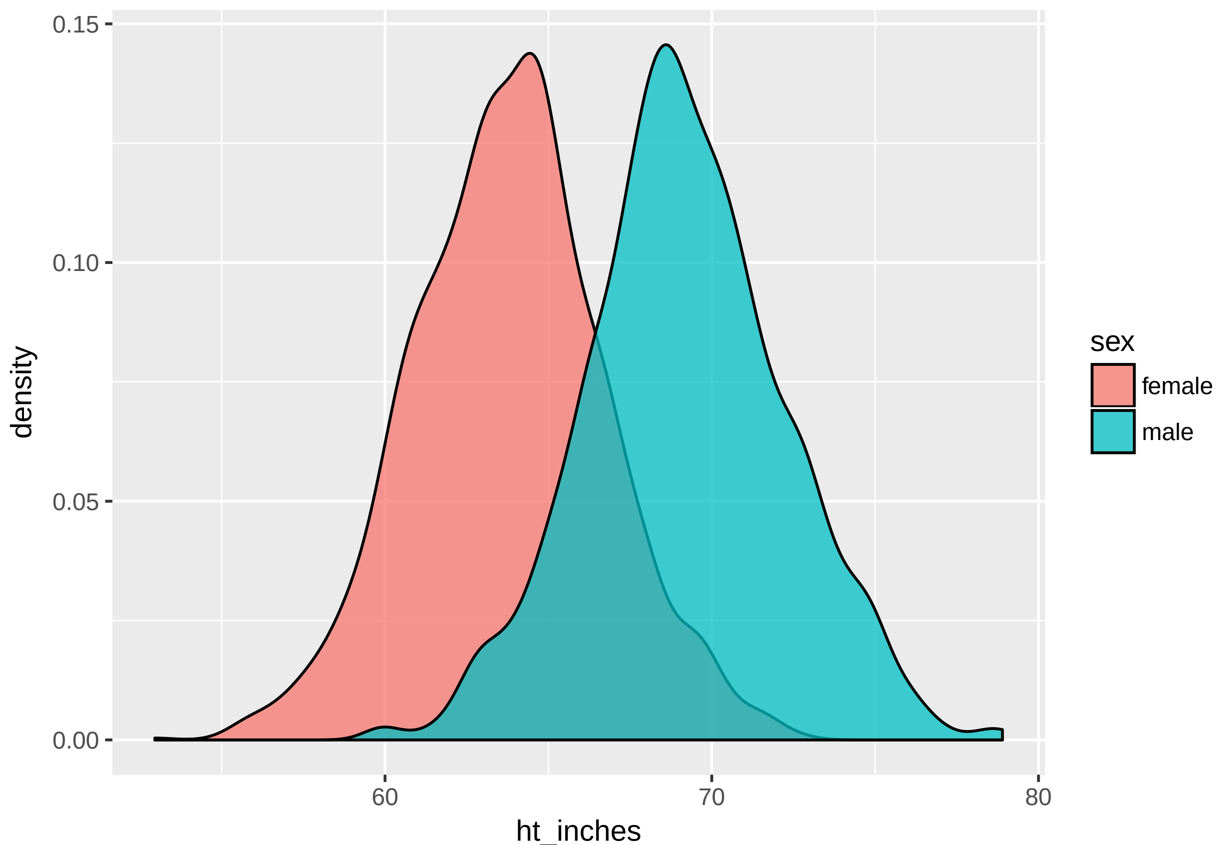
```
nhanes <- read_csv("nhanes.csv")
```

```r
height <- nhanes %>%
  filter(Age >= 20) %>%
  mutate(ht_inches = Height/2.54,
         sex = factor(Gender)) %>% #making sex a factor variable and making it a categorical factor var
  select(ht_inches, sex) %>%
  na.omit()

#na.omit = removes all rows with missing data
```

Make a desnity plot of height, group by sex

```r
ggplot(height, aes(x=ht_inches, group=sex, fill=sex)) +
  geom_density(alpha=0.75) #"alpha=" changes the transparency
```



Full numeric summary of ht_inches (descriptr package)

```r
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
```

Description of the numeric summaries

Central Tendancy

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.

Figure 1:

Dispersion

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.

Figure 2:

Normal distribution explanation

Emperical Rule

Full numeric summary of ht_inches by sex

```r
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|
```

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.

Figure 3:

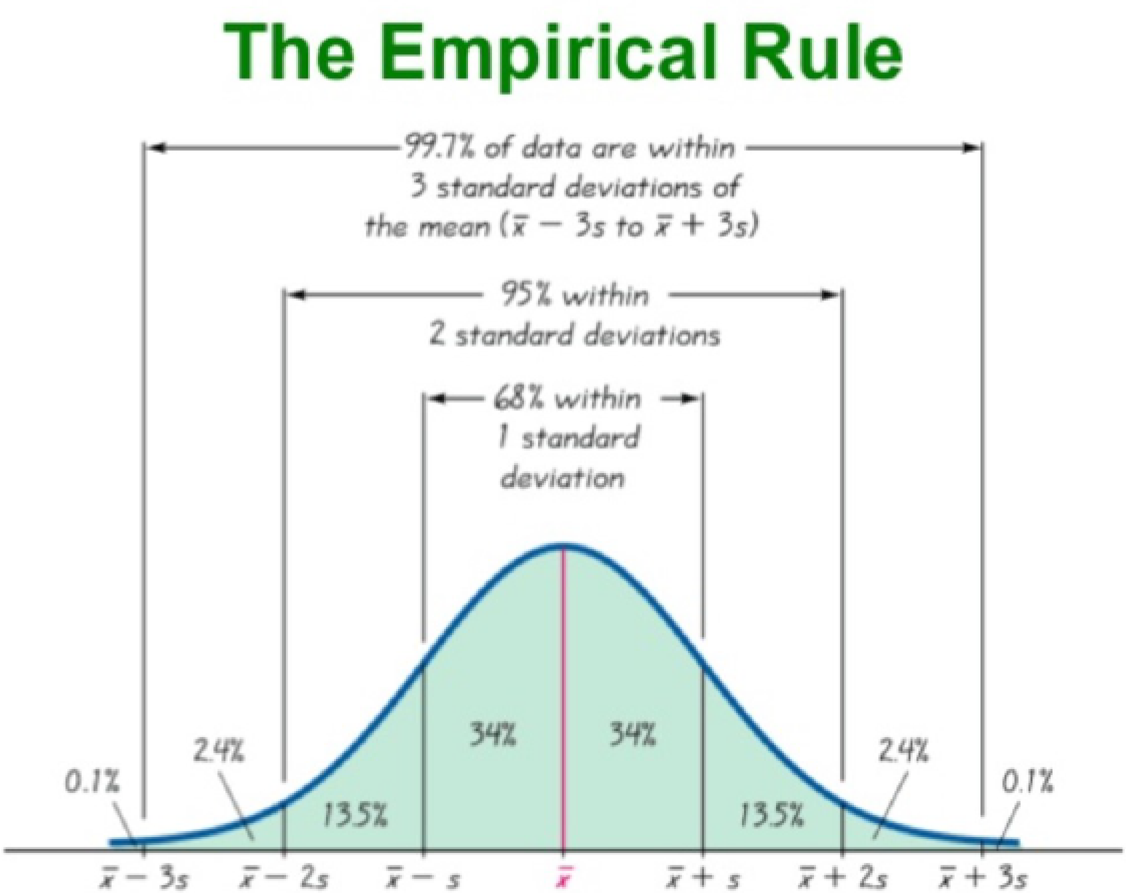


Figure 4:

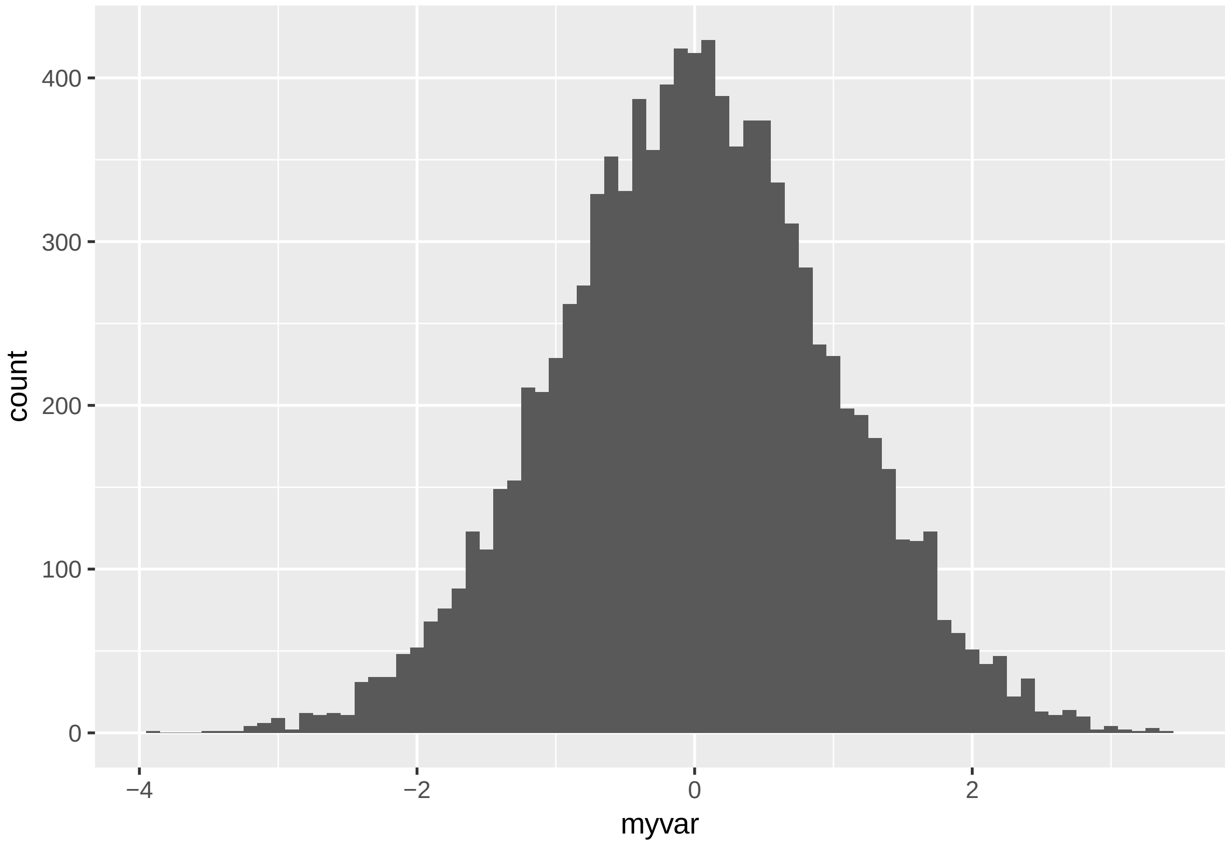
```
## |                  Mean|                63.76|                69.19|
## |                Median|                63.82|                69.02|
## |                  Mode|                63.27|                68.19|
## |        Std. Deviation|                 2.91|                 3.01|
## |              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|
## -------------------------------------------------------------------
```

Generate and explore a normal distribution

```
set.seed(12345) #you NEED this in order for us to create the same result everytime

myvar <- rnorm(n=10000, m=0, sd=1) #rnorm = function will generate data under normal distribution n= ge
example <- data.frame(myvar) #turning the

#Plot the distribution of the example dataframe we created
ggplot(example, aes(x = myvar)) +
  geom_histogram(binwidth = .1)
```




```
example <- example %>%
mutate(within1 = ifelse(myvar <= 1 & myvar >= -1, 1, 0), #we're creating a new variable that is testing
  within2 = ifelse(myvar <= 2 & myvar >= -2, 1, 0), #we're creating a new variable that is testing if i
  within3 = ifelse(myvar <= 3 & myvar >= -3, 1, 0)) #we're creating a new variable that is testing if i

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

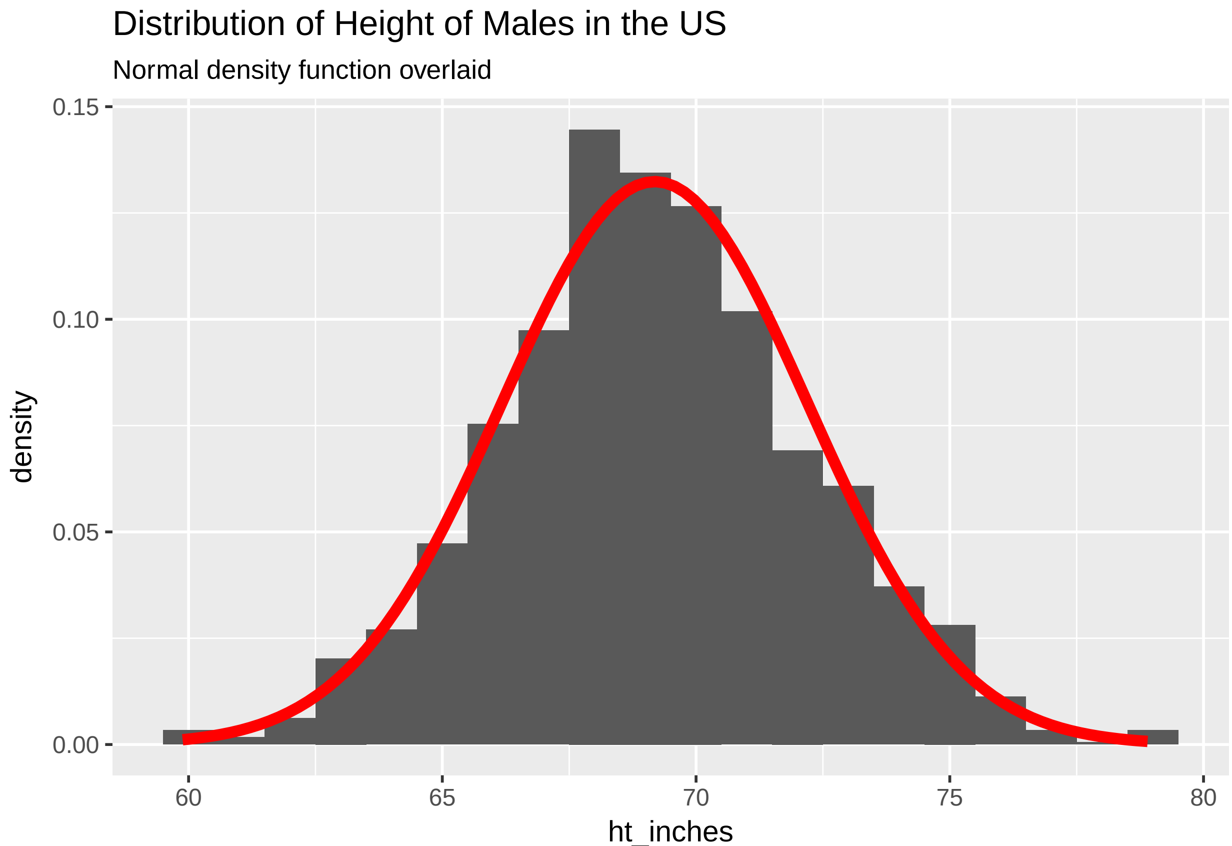
```
##   prop_within1 prop_within2 prop_within3
## 1        0.684       0.9528       0.9975
```

Compare the height distributions to a perfect normal distribution

This is a good tool to see if your data is normally distributed or not.

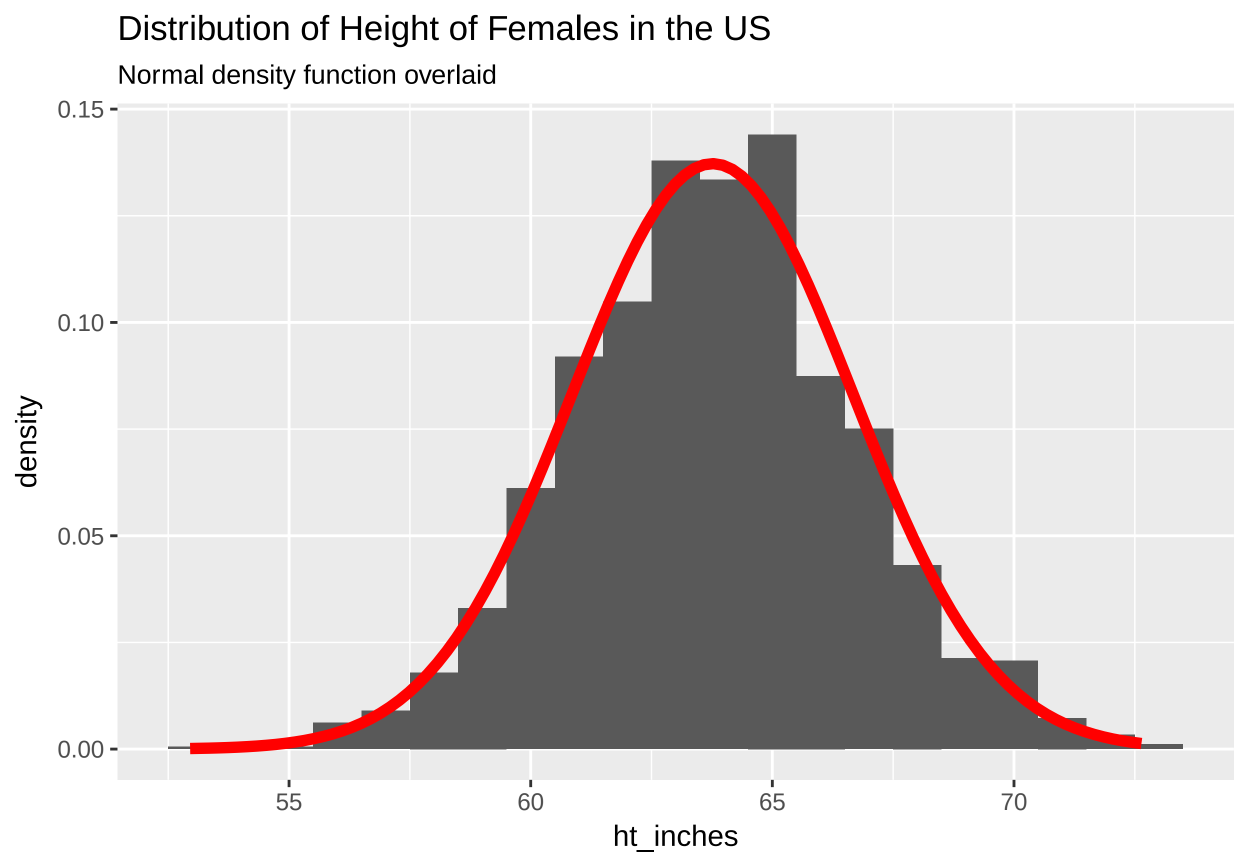
```
# for males
males <- filter(height, sex == "male")

ggplot(males, aes(x = ht_inches)) +
  geom_histogram(aes(y = ..density..), binwidth = 1) + #here with the "y=..density.." we are indicating
  stat_function(fun = dnorm,  #stat_function() indicates that we want to add a stat function.... "dnorm
  args = list(mean = mean(males$ht_inches), sd = sd(males$ht_inches)), #this is indicatign what we want
  lwd = 2, #linewidth
  col = 'red') + #color
  labs(title = "Distribution of Height of Males in the US", subtitle = "Normal density function overlai
```




```
# 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 overla
```

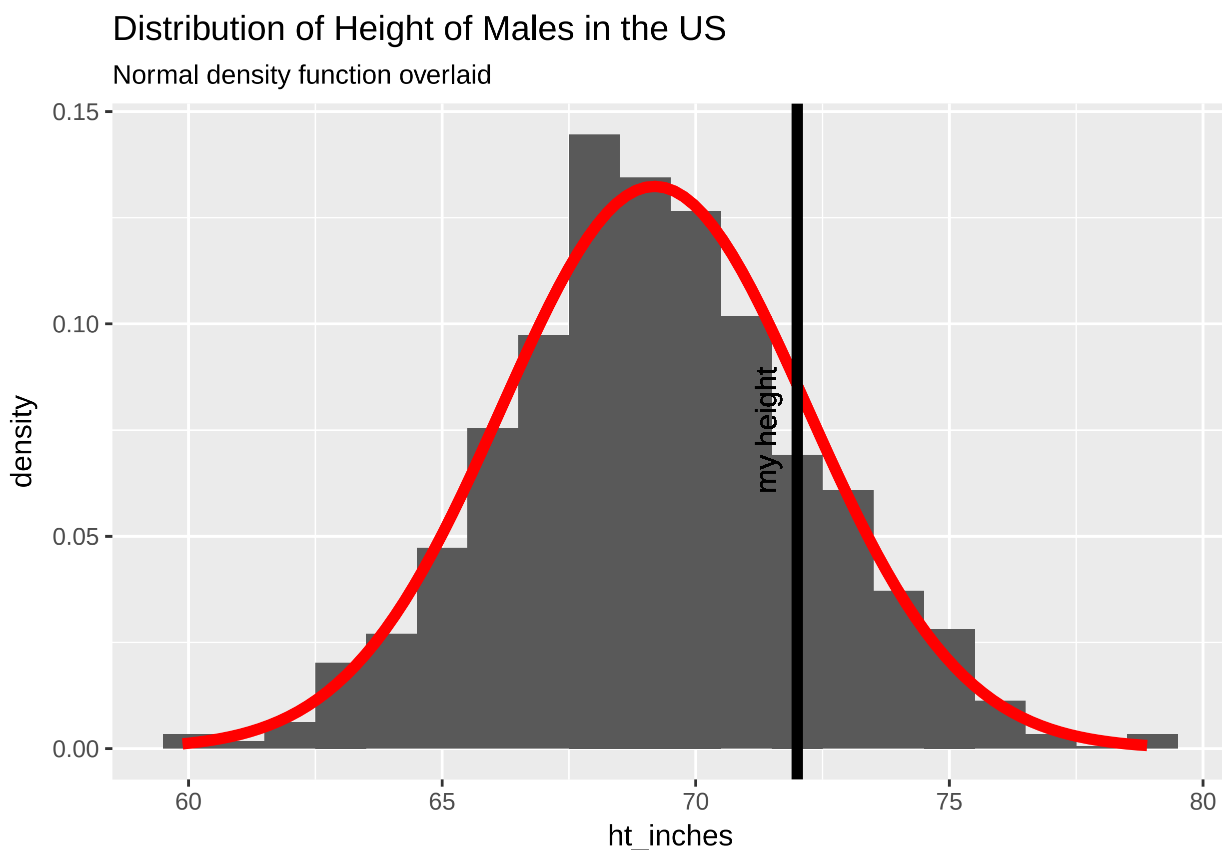


Add your own height to the graph geom_vline()

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

ggplot(males, aes(x = ht_inches)) +
  geom_histogram(aes(y = ..density..), binwidth = 1) + #here with the "y=..density.." we are indicating
  stat_function(fun = dnorm,  #stat_function() indicates that we want to add a stat function.... "dnorm
  args = list(mean = mean(males$ht_inches), sd = sd(males$ht_inches)), #this is indicatign what we want
  lwd = 2, #linewidth
  col = 'red') + #color
  geom_vline(xintercept=72, colour="black", lwd=2) +
  geom_text(aes(x=72, label="my height", y=.075), colour="black", angle=90, vjust = -1, text=element_te
  labs(title = "Distribution of Height of Males in the US", subtitle = "Normal density function overlai
```

```
## Warning: Ignoring unknown parameters: text
```



Store mean and standard deviation of height for males and females

```
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 <- (72 - mean_m)/sd_m
```

create z-scores of height variables

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

What is the probability that a randomly selected male is less than 65 inches?

We can determine this by using the pnorm function!

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

```
## [1] 0.0823984
```