Statistics Terminology

Like any field of inquiry,
statistics assigns very specific meaning to some everyday words.

- sample (data), statistic
- population, parameter
- dataset: case, label, variable, value
- variable: quantitative, categorical
- distribution: variance, skew

Example: American College Football Fumbles I

```
glimpse(fumbles)
```

```
Observations: 120
Variables: 7
$ team  (fctr) Air Force, Akron, Alabama, Arizona, Ariz...
$ rank  (int) 53, 19, 68, 31, 94, 46, 60, 94, 18, 94, 8...
$ W     (int) 8, 1, 9, 7, 5, 9, 4, 6, 12, 4, 7, 10, 6, ...
$ L     (int) 4, 11, 3, 4, 6, 2, 7, 5, 0, 8, 5, 1, 5, 1...
$ week1 (int) 4, 2, 0, 1, 2, 0, 0, 3, 1, 2, 5, 3, 0, 1,...
$ week2 (int) 2, 3, 3, 0, 1, 1, 0, 2, 1, 2, 2, 2, 2, 1,...
$ week3 (int) 2, 2, 2, 2, 3, 0, 4, 0, 0, 2, 1, 2, 4, 2,...
```

Terms: popn vs. sample, cases vs. labels, variables vs. values

Variables: quantitative vs. categorical

```
help(fumbles)
```

Example: American College Football Fumbles II

Description

This data frame gives the number of fumbles by each NCAA FBS team for the first three weeks in November, 2010.

Format

A data frame with 120 observations on the following 7 variables.

- team NCAA football team
- rank rank based on fumbles per game through games on November 26, 2010
- W number of wins through games on November 26, 2010
- L number of losses through games on November 26, 2010
- week1 number of fumbles on November 6, 2010
- week2 number of fumbles on November 13, 2010
- week3 number of fumbles on November 20, 2010

Details

The fumble counts listed here are total fumbles, not fumbles lost. Some of these fumbles were recovered by the team that fumbled.

Source

http://www.teamrankings.com/college-football/stat/fumbles-per-game

Example: American College Football Fumbles III

- ✓ **Cases** are the objects described by a set of data. Cases may be customers, companies, experimental subjects, or other objects.
- ✓ A **variable** is a special characteristic of a case.
- ✓ A **label** is a special variable used in some data sets to distinguish between cases.
- ✓ Different cases can have different **values** of a variable.

Example: American College Football Fumbles IV

dis·tri·bu·tion

/ˌdistrəˈbyo͞oSH(ə)n/

noun

the action of sharing something out among a number of recipients.
"she had it printed for distribution among her friends"
synonyms: giving out, dealing out, doling out, handing out/around, issue, issuing, dispensation; More

- the way in which something is shared out among a group or spread over an area.
  "changes undergone by the area have affected the distribution of its wildlife"
  synonyms: dispersal, dissemination, spread; More
- the action or process of supplying goods to stores and other businesses that sell to consumers.
  "a manager has the choice of four types of distribution"
  synonyms: supply, supplying, delivery, transport, transportation
  "centers of food distribution"

Example: American College Football Fumbles V

What is the distribution of number of team fumbles in week #1?

The sample (or population) **distribution**
of a variable has two components:

1. the set values observed (or possible to observe)
2. the relative frequency of occurrence for those values

- A **categorical** variable places each case into one of several groups, or categories.

- A **quantitative** variable takes numerical values for which arithmetic operations such as adding and averaging make sense.

- The **distribution** of a variable tells us the values that a variable takes and how often it takes each value.

Example: American College Football Fumbles VI

```
head(fumbles)
```

```
         team rank W  L week1 week2 week3
1   Air Force   53 8  4     4     2     2
2       Akron   19 1 11     2     3     2
3     Alabama   68 9  3     0     3     2
4     Arizona   31 7  4     1     0     2
5  Arizona St   94 5  6     2     1     3
```

```
tail(fumbles)
```

```
           team rank  W L week1 week2 week3
116 Wake Forest   23  2 9     1     1     1
117  Wash State   53  2 9     3     1     4
118  Washington   41  4 6     1     0     0
119   Wisconsin    4 10 1     1     1     0
120     Wyoming   28  3 9     0     3     1
```

...and how was the `rank` variable determined?

It just looks wrong.

```
help(fumbles)
```

Example: American College Football Fumbles VII

Description

This data frame gives the number of fumbles by each NCAA FBS team for the first three weeks in November, 2010.

Format

A data frame with 120 observations on the following 7 variables.

- team NCAA football team
- rank rank based on fumbles per game through games on November 26, 2010
- W number of wins through games on November 26, 2010
- L number of losses through games on November 26, 2010
- week1 number of fumbles on November 6, 2010
- week2 number of fumbles on November 13, 2010
- week3 number of fumbles on November 20, 2010

Details

The fumble counts listed here are total fumbles, not fumbles lost. Some of these fumbles were recovered by the team that fumbled.

Source

http://www.teamrankings.com/college-football/stat/fumbles-per-game

The rank was based on fumbles per game over the whole season, not on just the first 3 weeks (not on winning percentage either).

Example: American College Football Fumbles VIII

What is the distribution of number of team fumbles in week #1?

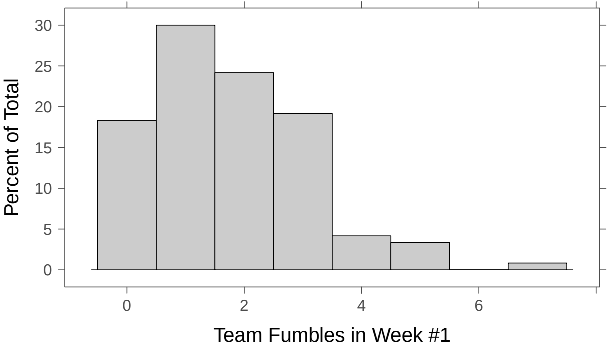
```
tally(  week1, data=fumbles)
```

```
 0  1  2  3  4  5  7
22 36 29 23  5  4  1
```

Example: American College Football Fumbles IX

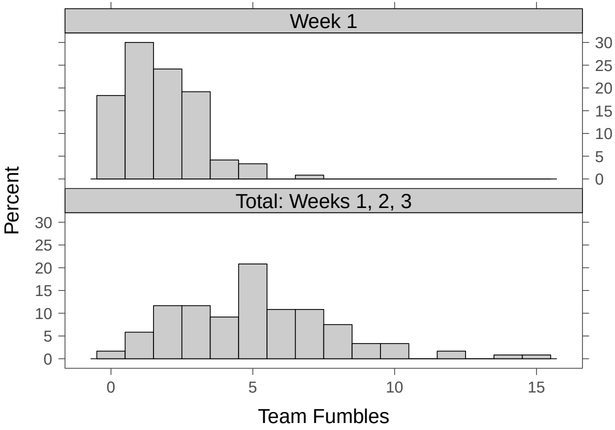
Qualitatively describing the distribution of a quantitative variable: center, spread, and shape

```
histogram(~ week1, data=fumbles, type="percent",
          xlab="Team Fumbles in Week #1")
```



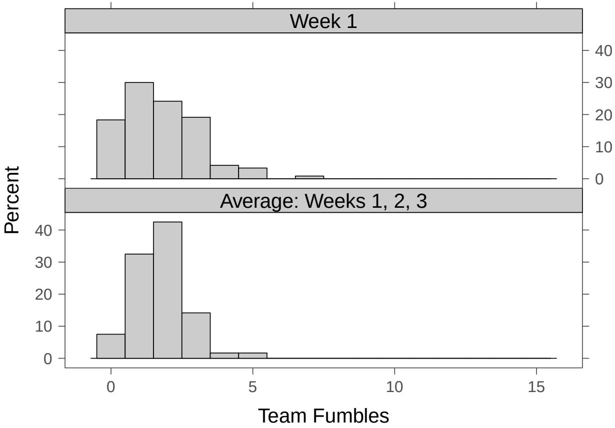
Example: American College Football Fumbles X

Distribution of the **total** team fumbles over 3 games:



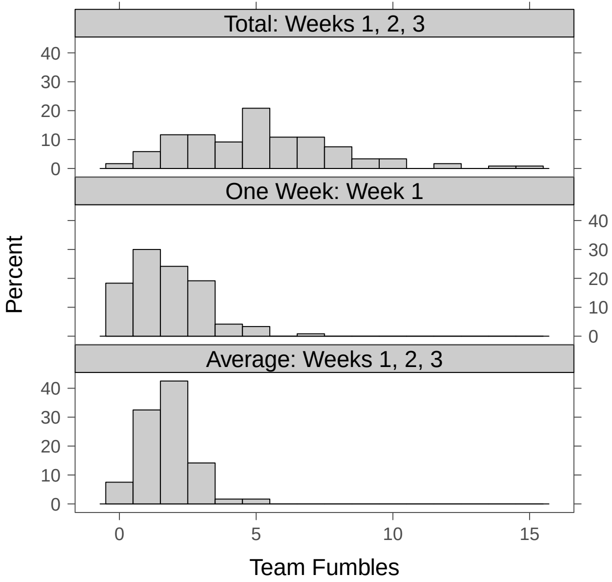
Example: American College Football Fumbles XI

Distribution of the **average** team fumbles over 3 games:



Center of averages similar to individuals, but less spread and skew.

Example: American College Football Fumbles XII



Software Installation: RStudio (and R) I

RStudio = the work environment
R = the engine (a statistical programming language)

To use the R code suggested for homework, you must install the `mosaic` package **once** at the start of the quarter.

```
install.packages("mosaic", ...and other packages)
```

Then, **every** time you start RStudio, type

```
require(mosaic)
```

Software installation instructions:
`http://statistics.uchicago.edu/~collins/Rinstall`

Example: Bicycle weight and commuting time I

```
glimpse(myBikeCommute)
```

```
Observations: 56
Variables: 7
$ Bike     (fctr) Steel, Carbon, Steel, Carbon, Carbon,...
$ Date     (fctr) 20/01/10, 21/01/10, 25/01/10, 26/01/1...
$ Distance (dbl) 27.20, 27.46, 27.20, 27.52, 27.51, 27....
$ Minutes  (dbl) 115.1, 115.6, 115.8, 113.9, 119.2, 108...
$ AvgSpeed (dbl) 14.10, 14.25, 14.10, 14.49, 13.84, 14....
$ TopSpeed (dbl) 31.50, 30.64, 30.92, 33.02, 30.92, 32....
$ Month    (fctr) 1Jan, 1Jan, 1Jan, 1Jan, 2Feb, 2Feb, 2...
```

Thanks to Dr. Jeremy Groves for providing his personal data.

`http://www.bmj.com/content/341/bmj.c6801` Groves, J. Bicycle weight and commuting time: randomised trial, *British Medical Journal*, BMJ 2010;341:c6801.

Example: Bicycle weight and commuting time II

```
head(myBikeCommute)
```

```
     Bike     Date Distance Minutes AvgSpeed TopSpeed Month
1   Steel 20/01/10    27.20   115.1    14.10    31.50  1Jan
2  Carbon 21/01/10    27.46   115.6    14.25    30.64  1Jan
3   Steel 25/01/10    27.20   115.8    14.10    30.92  1Jan
4  Carbon 26/01/10    27.52   113.9    14.49    33.02  1Jan
5  Carbon 27/01/10    27.51   119.2    13.84    30.92  2Feb
6   Steel 01/02/10    27.17   108.7    14.99    32.09  2Feb
7   Steel 03/02/10    27.16   117.7    13.84    32.09  2Feb
8  Carbon 03/02/10    27.49   123.3    13.37    29.58  2Feb
9  Carbon 08/02/10    27.48   112.5    14.65    34.02  2Feb
10  Steel 09/02/10    27.09   112.6    14.43    32.71  2Feb
11 Carbon 11/02/10    27.44   117.7    13.99    32.00  3Mar
12 Carbon 01/03/10    27.49   108.6    15.18    32.71  3Mar
13 Carbon 03/03/10    27.49   110.9    14.82    34.71  3Mar
```

Why not alternating Steel, Carbon, Steel, Carbon, Steel, etc.?

Terms: popn vs. sample, cases vs. labels, variables vs. values

Variables: quantitative vs. categorical

```
help(BikeCommute)
```

Example: Bicycle weight and commuting time III

Description

Commute times for two kinds of bicycle

Format

A dataset with 56 observations on the following 9 variables.

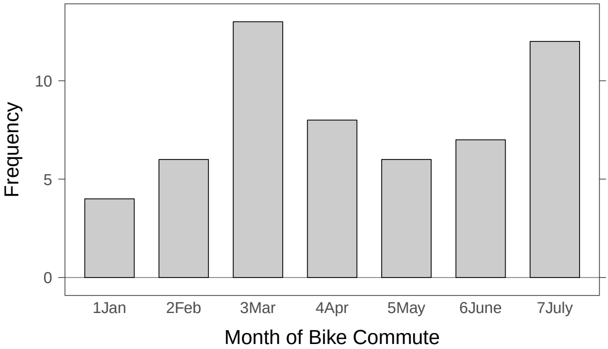
`Bike`	Type of material `Carbon` or `Steel`
`Date`	Date of the bike commute
`Distance`	Length of commute (in miles)
`Time`	Total commute time (hours:minutes:seconds)
`Minutes`	Time converted to minutes
`AvgSpeed`	Average speed during the ride (miles per hour)
`TopSpeed`	Maximum speed (miles per hour)
`Seconds`	Time converted to seconds
`Month`	Categories: `1Jan` `2Feb` `3Mar` `4Apr` `5May` `6June` `7July`

Details

Data from a personal experiment to compare commuting time based on a randomized selection between two bicycles made of different materials.

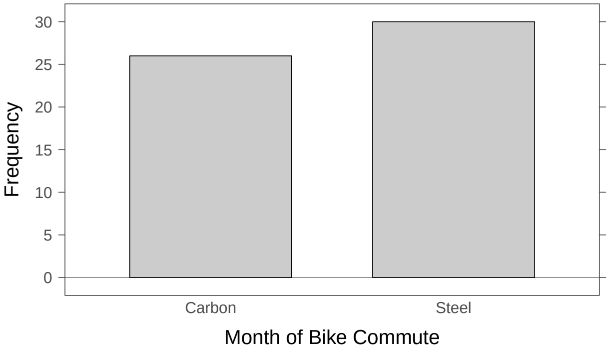
Example: Bicycle weight and commuting time IV

```
bargraph(~ Month, data=myBikeCommute,
         xlab="Month of Bike Commute")
```



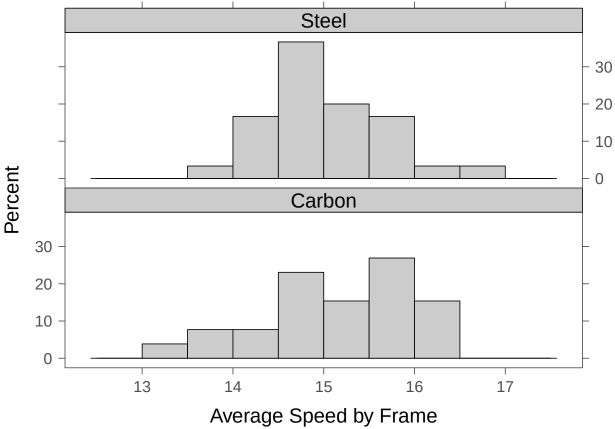
Example: Bicycle weight and commuting time V

```
bargraph(~ Bike, data=myBikeCommute,
         xlab="Month of Bike Commute")
```

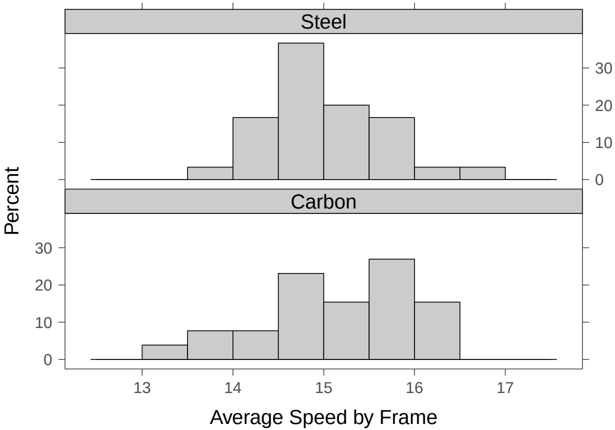


Example: Bicycle weight and commuting time VI

```
histogram(~ AvgSpeed | Bike, data=myBikeCommute, type="percent",
  xlab="Average Speed by Frame", ylab="Percent", layout=c(1,2))
```



Example: Bicycle weight and commuting time VII



Compare speed distributions: center, spread, shape

Steel: same average?, less spread, right skewed

Carbon: same average?, more spread, left skewed

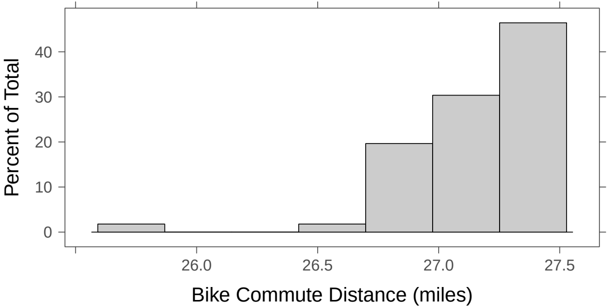
Example: Bicycle weight and commuting time VIII

Compare centers and spreads of speed distributions

```
mean(~ AvgSpeed | Bike, data=myBikeCommute)

Carbon  Steel
 15.19  15.04

favstats(~ AvgSpeed | Bike, data=myBikeCommute)

    Bike   min    Q1 median    Q3   max  mean     sd  n
1 Carbon 13.37 14.60  15.22 15.91 16.28 15.19 0.8102 26
2  Steel 13.84 14.57  14.96 15.45 16.55 15.04 0.6457 30
  missing
1       0
2       0

IQR(~ AvgSpeed | Bike, data=myBikeCommute)

Carbon  Steel
1.3100 0.8725
```

Example: Bicycle weight and commuting time IX

```
histogram(~ Distance, data=myBikeCommute, type="percent",
          xlab="Bike Commute Distance (miles)")
```



Why does the commute distance vary from ride to ride?

Isn't it the same route to work every day?

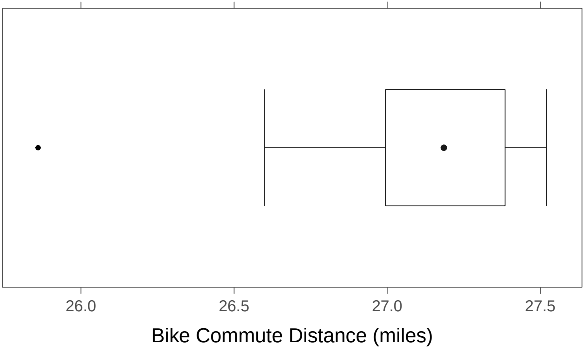
Why is one commute so much shorter than the others?

Example: Bicycle weight and commuting time X

```
quantile(~ Distance, data=myBikeCommute)

   0%   25%   50%   75%  100%
25.86 27.00 27.19 27.38 27.52


c(Q1, Q3, IQR, 1.5*IQR, Q1 - 1.5*IQR, Q3 + 1.5*IQR)

[1] 27.00 27.38  0.39  0.58 26.42 27.96


sort(myBikeCommute$Distance)

 [1] 25.86 26.60 26.74 26.88 26.90 26.91 26.91 26.91 26.92
[10] 26.94 26.94 26.94 26.95 26.99 27.00 27.00 27.01 27.02
[19] 27.02 27.03 27.03 27.05 27.06 27.09 27.10 27.16 27.16
[28] 27.17 27.20 27.20 27.27 27.29 27.31 27.31 27.32 27.32
[37] 27.33 27.34 27.34 27.36 27.36 27.38 27.39 27.40 27.40
[46] 27.43 27.44 27.45 27.46 27.48 27.49 27.49 27.49 27.51
[55] 27.52 27.52
```

Example: Bicycle weight and commuting time XI

```
bwplot(~ Distance, data=myBikeCommute,
      xlab="Bike Commute Distance (miles)")
```



Measuring Center of Data Distribution: Average I

```
mean(~AvgSpeed | Bike, data = myBikeCommute)
```

```
Carbon  Steel
 15.19  15.04
```

Average of average speed is about the same for both frame types.

```
mean(~Distance, data = myBikeCommute)
```

```
[1] 27.16
```

The average distance is close to claimed distance: 27 miles

Measuring Center of Data Distribution: Average II

Definition:

$$\text{sample average} \;=\; \overline{x} = \text{ ``x-bar'' } = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$n =$ sample size

Measuring Center of Data Distribution: Median

```
median(~AvgSpeed | Bike, data = myBikeCommute)
```

```
Carbon  Steel
 15.22  14.96
```

```
median(~Distance, data = myBikeCommute)
```

```
[1] 27.19
```

The median distance is close to claimed distance: 27 miles

```
sort(myBikeCommute$Distance)
```

```
 [1] 25.86 26.60 26.74 26.88 26.90 26.91 26.91 26.91 26.92
[10] 26.94 26.94 26.94 26.95 26.99 27.00 27.00 27.01 27.02
[19] 27.02 27.03 27.03 27.05 27.06 27.09 27.10 27.16 27.16
[28] 27.17 27.20 27.20 27.27 27.29 27.31 27.31 27.32 27.32
[37] 27.33 27.34 27.34 27.36 27.36 27.38 27.39 27.40 27.40
[46] 27.43 27.44 27.45 27.46 27.48 27.49 27.49 27.49 27.51
[55] 27.52 27.52
```

The Average is the Balancing Point

Consider the data $x_1 = 9, x_2 = 3, x_3 = 15, x_4 = 1$

- What is the average of these values?

The Average is the Balancing Point

Consider the data $x_1 = 9, x_2 = 3, x_3 = 15, x_4 = 1$

- What is the average of these values?
- What are the deviations of the data from the average?

The Average is the Balancing Point

Consider the data $x_1 = 9, x_2 = 3, x_3 = 15, x_4 = 1$

- What is the average of these values?
- What are the deviations of the data from the average?
- What is the sum of the deviations from the average?

The Average is the Balancing Point

Consider the data $x_1 = 9, x_2 = 3, x_3 = 15, x_4 = 1$

- What is the average of these values?
- What are the deviations of the data from the average?
- What is the sum of the deviations from the average?
- The average is the "balancing point" of the data, the "center of mass" (assigning each data value the same mass $= 1/4$)

The Average is the Balancing Point

Consider the data $x_1 = 9, x_2 = 3, x_3 = 15, x_4 = 1$

- What is the average of these values?
- What are the deviations of the data from the average?
- What is the sum of the deviations from the average?
- The average is the "balancing point" of the data, the "center of mass" (assigning each data value the same mass $= 1/4$)

Talk a moment with your neighbor. See if you can come up an equation to express this "balancing point" property of the average.

The Average is the Balancing Point

Consider the data $x_1 = 9, x_2 = 3, x_3 = 15, x_4 = 1$

- What is the average of these values?
- What are the deviations of the data from the average?
- What is the sum of the deviations from the average?
- The average is the "balancing point" of the data, the "center of mass" (assigning each data value the same mass $= 1/4$)

Talk a moment with your neighbor. See if you can come up an equation to express this "balancing point" property of the average.

Proof: Show that for **any** sample of size n, $\sum_{i=1}^{n} (x_i - \overline{x}) = 0$

How to Prove the Math Stuff I

- A proof is a “paragraph” of mathematical “sentences”,
- written in order to make logical sense to the reader.
  ...just like you do in the Core all the time!
- It’s your personal argument as to why a claim must be true.
- Justify each step ("sentence") using statistics
  (and using results already proven in the course).

How to Prove the Math Stuff II

OK. Our first proof is to confirm an equation.

Proof: Show that for **any** sample of size n, $\displaystyle\sum_{i=1}^{n}(x_i - \overline{x}) = 0$

Start on the left side: $\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})$

$=$ rewrite

$=$ and rewrite

$=$ and rewrite again

$=$ until arriving at the right side $= 0$

In groups: Write down a first step.

Our First Proof!

Four common starting points.

Three are great, but one is incorrect. Which one? Why?

1. $\displaystyle\sum_{i=1}^{n}(x_i - \overline{x}) = (x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x})$
2. $\displaystyle\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n}\left[x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right]$
3. $\displaystyle\sum_{i=1}^{n}(x_i - \overline{x}) = 0$
4. $\displaystyle\sum_{i=1}^{n}(x_i - \overline{x}) = \left[\sum_{i=1}^{n} x_i\right] - \left[\sum_{i=1}^{n} \overline{x}\right]$

Is "Σ" confusing you? Read Chapter 0 (Math Supplement).

Our First Proof!

Starting with the first option:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \overline{x}) &= (x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x}) \\
&= \\
&= \\
&= \\
&=
\end{aligned}$$

Our First Proof!

Starting with the first option:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = (x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x})$$

$$= (x_1 + x_2 + \cdots + x_n) - \underbrace{(\overline{x} + \overline{x} + \cdots + \overline{x})}_{n \text{ times}}$$

$$=$$

$$=$$

$$=$$

Our First Proof!

Starting with the first option:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \overline{x}) &= (x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x}) \\
&= (x_1 + x_2 + \cdots + x_n) - \underbrace{(\overline{x} + \overline{x} + \cdots + \overline{x})}_{n \text{ times}} \\
&= \left[\sum_{i=1}^{n} x_i\right] - n\overline{x} \\
&= \\
&=
\end{aligned}$$

Our First Proof!

Starting with the first option:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = (x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x})$$

$$= (x_1 + x_2 + \cdots + x_n) - \underbrace{(\overline{x} + \overline{x} + \cdots + \overline{x})}_{n \text{ times}}$$

$$= \left[\sum_{i=1}^{n} x_i\right] - n\overline{x} \quad = \quad \left[\frac{n}{n}\sum_{i=1}^{n} x_i\right] - n\overline{x}$$

$$=$$

$$=$$

Our First Proof!

Starting with the first option:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = (x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x})$$

$$= (x_1 + x_2 + \cdots + x_n) - \underbrace{(\overline{x} + \overline{x} + \cdots + \overline{x})}_{n \text{ times}}$$

$$= \left[\sum_{i=1}^{n} x_i\right] - n\overline{x} \quad = \quad \left[\frac{n}{n}\sum_{i=1}^{n} x_i\right] - n\overline{x}$$

$$= n\overline{x} - n\overline{x}$$

$$=$$

Our First Proof!

Starting with the first option:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = (x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x})$$

$$= (x_1 + x_2 + \cdots + x_n) - \underbrace{(\overline{x} + \overline{x} + \cdots + \overline{x})}_{n \text{ times}}$$

$$= \left[\sum_{i=1}^{n} x_i\right] - n\overline{x} \quad = \quad \left[\frac{n}{n}\sum_{i=1}^{n} x_i\right] - n\overline{x}$$

$$= n\overline{x} - n\overline{x} \quad \text{since } \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{(Justification required!)}$$

$$=$$

Our First Proof!

Starting with the first option:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \overline{x}) &= (x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x}) \\
&= (x_1 + x_2 + \cdots + x_n) - \underbrace{(\overline{x} + \overline{x} + \cdots + \overline{x})}_{n \text{ times}} \\
&= \left[\sum_{i=1}^{n} x_i\right] - n\overline{x} \quad = \quad \left[\frac{n}{n}\sum_{i=1}^{n} x_i\right] - n\overline{x} \\
&= n\overline{x} - n\overline{x} \qquad \text{since } \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{(Justification required!)} \\
&= 0
\end{aligned}$$

Let's agree that $b - b = 0$ for any real number b. :)

Measuring Spread of Data Distribution I

The average devation $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})$ **always** $= 0$!

Need a different measure for "typical size of deviations" (spread)

There are many measures of spread:

- mean squared deviation (MSD or "variance"),
- mean absolute deviation (MAD),
- standard deviation (SD) = root $MSD = RMSD = \sqrt{MSD}$,
- interquartile range (IQR= range of middle 50% of data)
- range,
- ...and more (not covered in this course).

Measuring Spread of Data Distribution II

- No matter what number we might choose to measure center,
- we are summarizing an entire distribution with one number.
- There is a cost.
- We lose information.
- We should measure that loss and be aware of its magnitude.
- The mean and the median minimize the loss of information in some sense.
- Statisticians measure loss numerically with a "loss function"
- A loss function measures the distance of the data from the one-number summary (the "center").

A loss functions can be thought of as a measure of spread.

Measuring Spread of Data Distribution III

Let's consider two common loss functions (measures of spread)

- The mean of absolute deviations:

$$MAD(w) = \frac{1}{n}\sum_{i=1}^{n} |x_i - w|$$

- The mean of squared deviations:

$$MSD(w) = \frac{1}{n}\sum_{i=1}^{n} (x_i - w)^2$$

What value of w should we choose using MAD? Using MSD?

It seems reasonable that w should be in the "center" of the data for each measure. But which value in the middle would be best?

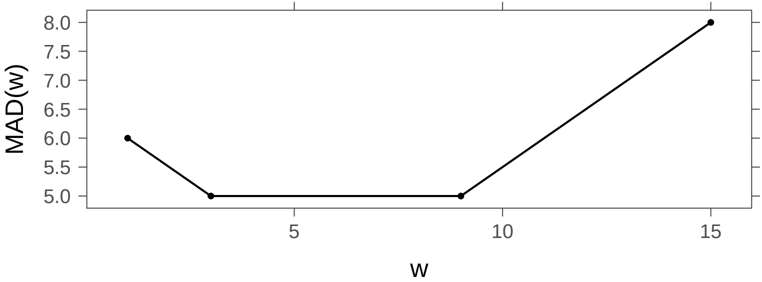
One optimality criteria: Choose w that minimizes MAD or MSD.

What is so special about the median? I

Consider again the data $x_1 = 9, x_2 = 3, x_3 = 15, x_4 = 1$

What does the $MAD(w)$ function look like for these data?

```
x <- c(9,3,15,1)
MAD <- function(w) { mean( abs(x-w) ) }
```



Where is the function $MAD(w)$ smallest (minimized)?

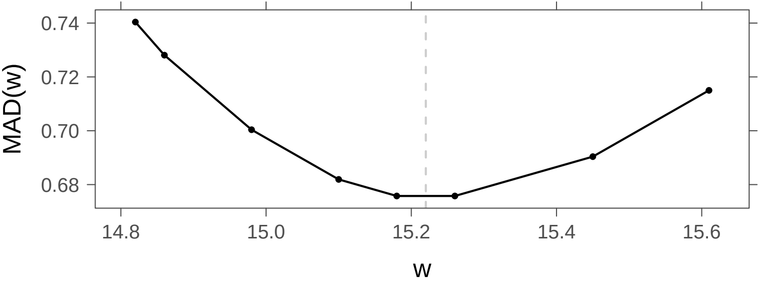
What is so special about the median? II

Consider again the bike commute data: Carbon frame `AvgSpeed`

What does the $MAD(w)$ function look like for these data?

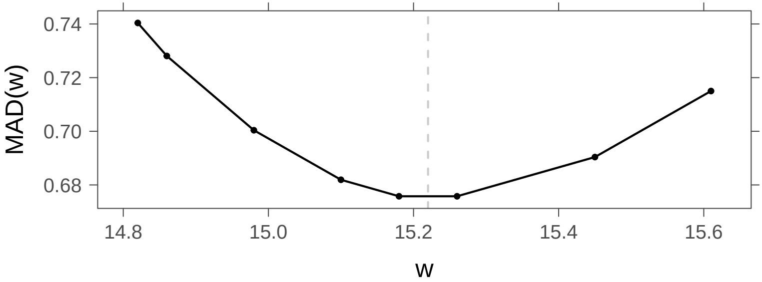
```
x <- with(myBikeCommute, AvgSpeed[Bike=="Carbon"])
favstats(x)
```

```
   min    Q1 median    Q3   max  mean     sd  n missing
 13.37 14.6  15.22 15.91 16.28 15.19 0.8102 26       0
```



Where is the function $MAD(w)$ smallest (minimized)?

What is so special about the median? III



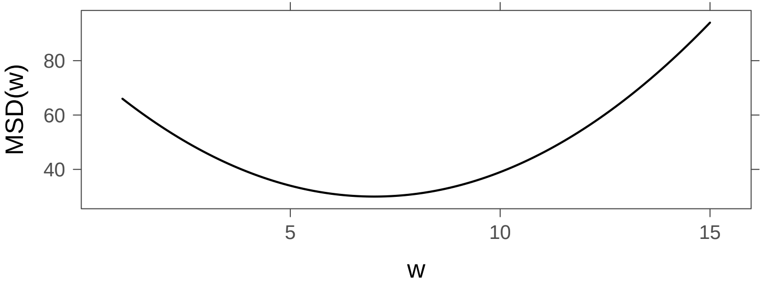

```
sort(x)

 [1] 13.37 13.84 13.99 14.25 14.49 14.54 14.58 14.65 14.82
[10] 14.86 14.98 15.10 15.18 15.26 15.45 15.61 15.64 15.75
[19] 15.78 15.95 15.96 15.99 16.15 16.15 16.25 16.28

median(x)

[1] 15.22
```

What is so special about the average? I

Consider again the data $x_1 = 9, x_2 = 3, x_3 = 15, x_4 = 1$

What does the $MSD(w)$ function look like for these data?

```
x <- c(9,3,15,1)
MSD <- function(w) { mean( (x-w)^2 ) }
```



Note: In this case, $MSD(W) = w^2 - 14w + 79$

Where is the function $MSD(w)$ smallest (minimized)?

What is so special about the average? II

What value w minimizes $MSD(w)$ for **any** sample: $x_1, x_2, \ldots, x_n$?

We want to minimize the following function with respect to w:

$$f(w) = MSD(w) = \frac{1}{n}\sum_{i=1}^{n}(x_i - w)^2$$

On your own: Show that $w = \overline{x}$ (average) minimizes $MSD(w)$.

Check that the average is the *unique* minimum (not just one of several values that attain the minimum, as for the median).

Solution: See Section 1.3 (Math Supplement)

We say that $\overline{x}$ is a **"least squares"** statistic.

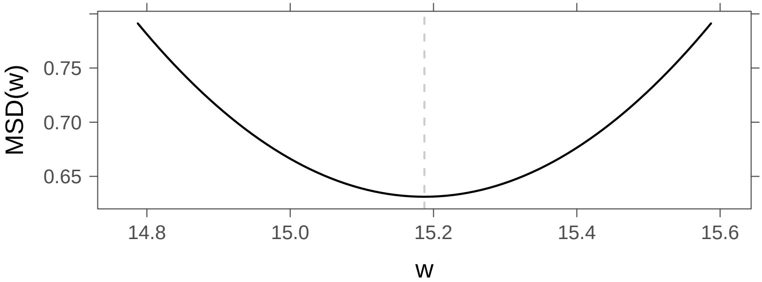
What is so special about the average? III

Consider again the bike commute data: Carbon frame `AvgSpeed`

What does the $MAD(w)$ function look like for these data?

```
x <- with(myBikeCommute, AvgSpeed[Bike=="Carbon"])
xbar <- mean(x)
xbar
```

```
[1] 15.19
```



Where is the function $MAD(w)$ smallest (minimized)?

Formulas for Sample Average, Variance, SD

$$\text{sample average } = \overline{x} = \text{ “x-bar” } = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\text{sample variance } = s^2 = \text{ “s-squared” } = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$\text{sample standard deviation } = s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$= \text{ “typical” distance from the average}$$

Why divide by $(n-1)$ instead of n for sample variance and SD?

Why divide by $(n-1)$ for sample variance and SD? I

Variance has a particular meaning in statistics:
mean squared distance from the average

$$MSD_n(\,\overline{x}\,) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Why collect data (**statistics**)?
To learn about the population (**parameters**).

$$\text{population mean} = \mu = \text{“myoo”} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\text{popn variance} = \sigma^2 = \text{“sigma squared”} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Why divide by $(n-1)$ for sample variance and SD? II

truth = popn variance $= \sigma^2 = MSD_N(\,\mu\,) = \dfrac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$

If we know the true popn mean (μ) and had a sample of n, use

$$\text{estimate} = \widehat{\sigma}^2_{\mu} = MSD_n(\,\mu\,) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 \qquad (1)$$

But, we almost never know μ! That's why we sample!

$$\text{realistic estimate} = \widehat{\sigma}^2_{\overline{x}} = MSD_n(\,\overline{x}\,) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 \qquad (2)$$

The problem: (2) $\leq$ (1). Why? ...and why is this a problem?
How does dividing by $(n-1)$ for (2) help? solve the problem?

Why divide by $(n-1)$ for sample variance and SD? III

OK. So, we should divide by a number smaller than n.
But, why $(n-1)$ in particular?

Claim: Just $(n-1)$ observations and $\overline{x}$ are sufficient to determine the one remaining observation.

Proof: We know $n\overline{x} = x_1 + x_2 + \cdots + x_n$, since $\overline{x} = \frac{1}{n}\sum x_i$
So, $x_n = n\overline{x} - (x_1 + x_2 + \cdots + x_{n-1})$.

In a sense, $\sum(x_i - \overline{x})^2$ adds up $(n-1)$ "independent" values.
We say that the sum $\sum(x_i - \overline{x})^2$ has $(n-1)$ **degrees of freedom** .
So, the sample average squared deviation (variance) is defined as

$$s^2 = \text{"s-squared"} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Linear Transformation of Data I

Sometimes we want to analyze data in different units

- Temperature: $\text{Celsius} = \frac{5}{9}(\text{Fahrenheit} - 32)$
- Curve: $\text{exam} = \text{score} + (0.25)(100 - \text{score})$

  This curve adds back 25% of exam points missed.
- Standardized Score: $z_i = \dfrac{x_i - \overline{x}}{s}$

Claim: All 3 are examples of linear transformations: $y = a + bx$

- Temperature: $\text{Celsius} = -\left(\dfrac{160}{9}\right) + \left(\dfrac{5}{9}\right)\text{Fahrenheit}$
- Curve: $\text{exam} = 25 + (0.75)\,\text{score}$
- Standardized Score: $z_i = -\left(\dfrac{\overline{x}}{s}\right) + \left(\dfrac{1}{s}\right)x_i$

Linear Transformation of Data II

High temperature in Chicago last 5 days of December

```
Fahrenheit

[1] 39 39 29 28 31

mean(Fahrenheit)

[1] 33.2

Celsius = -(160/9) + (5/9)*Fahrenheit

rbind(Fahrenheit, Celsius)

            [,1]  [,2]  [,3]  [,4]   [,5]
Fahrenheit 39.00 39.00 29.00 28.00 31.000
Celsius     3.89  3.89 -1.67 -2.22 -0.556

mean(Celsius)

[1] 0.667
```

Linear Transformation of Data III

```
mean(Celsius)
```

```
[1] 0.667
```

```
-(160/9) + (5/9) * mean(Fahrenheit)
```

```
[1] 0.667
```

Claim: If data $x_1, x_2, \ldots, x_n$
are linearly transformed to $y_i = a + bx_i$

Then, $\overline{y} = a + b\overline{x}$.

Proof: In class, if time.
A proof appears in Section 1.4 (Math Supplement).

Linear Transformation of Data IV

```
sd(Celsius)
[1] 3
(5/9) * sd(Fahrenheit)
[1] 3
```

Claim: If data $x_1, x_2, \ldots, x_n$
are linearly transformed to $y_i = a + bx_i$

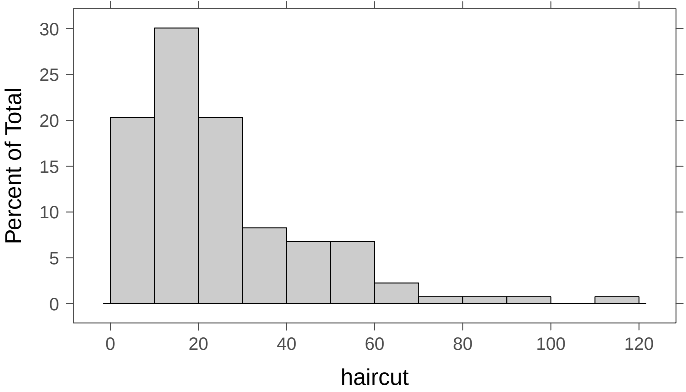
Then, $SD(y) = s_y = |b|s_x = |b|SD(x)$.

Proof: On your own for HW #2.

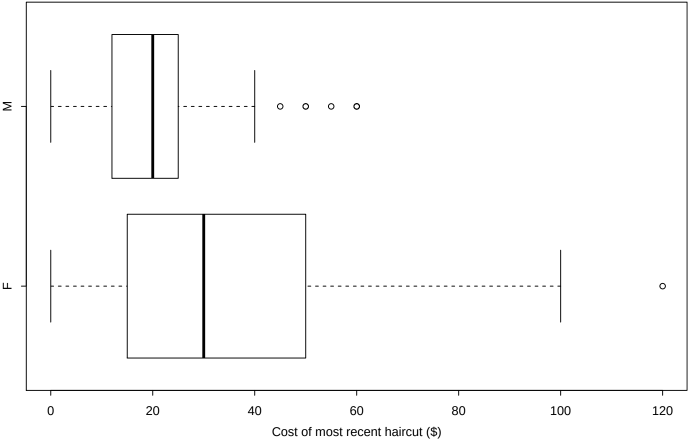
Class Survey Data

```
glimpse(surveyData)

Observations: 133
Variables: 13
$ student    (int) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12...
$ ageguess   (int) 62, NA, NA, 60, 50, 48, 45, 45, NA, 5...
$ section    (int) 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, NA, ...
$ priorStat (fctr) NotMuch, NotMuch, NotMuch, NotMuch, ...
$ math      (fctr) M196, M204, M133-153-163, M201, M196...
$ division  (fctr) SOC, PSD, BSD, NA, SOC, PSD, SOC, HU...
$ econ       (int) 1, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0...
$ height     (dbl) 63, 67, 64, 67, 68, 73, 63, 73, 70, 6...
$ momht      (dbl) 62, NA, NA, 60, 64, 66, NA, 59, NA, 6...
$ dadht      (dbl) 71, 66, NA, 69, 68, 76, NA, 67, NA, 7...
$ gender    (fctr) F, M, M, M, M, M, F, M, M, F, F, M, ...
$ haircut    (dbl) 12.0, 0.0, 8.0, 13.0, 12.0, 0.0, 21.0...
$ sibs       (int) 1, 2, 2, 1, 2, 6, 0, 1, 1, 0, 0, 1, 4...
```

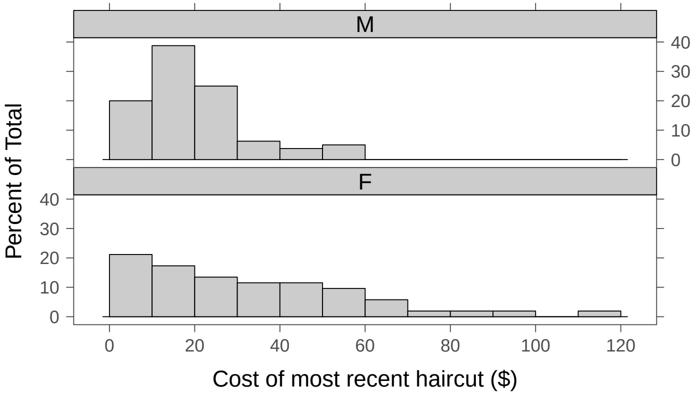
Is the cost of a haircut related to gender? I



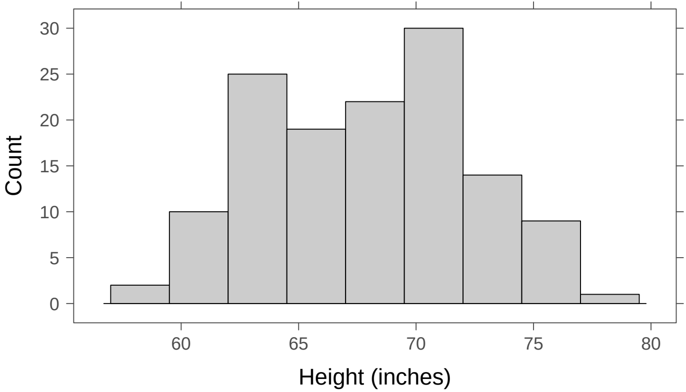
Is the cost of a haircut related to gender? II



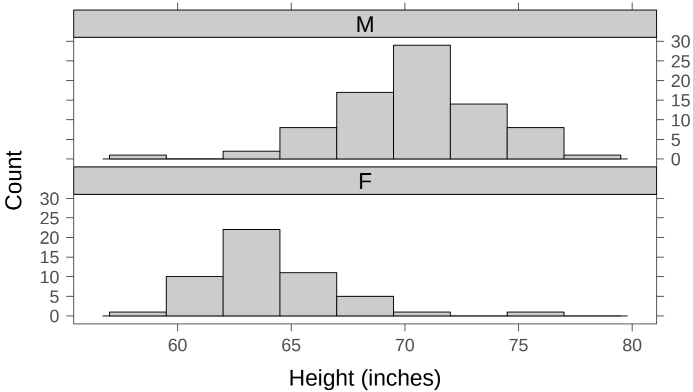
Is the cost of a haircut related to gender? III



The distribution of heights I



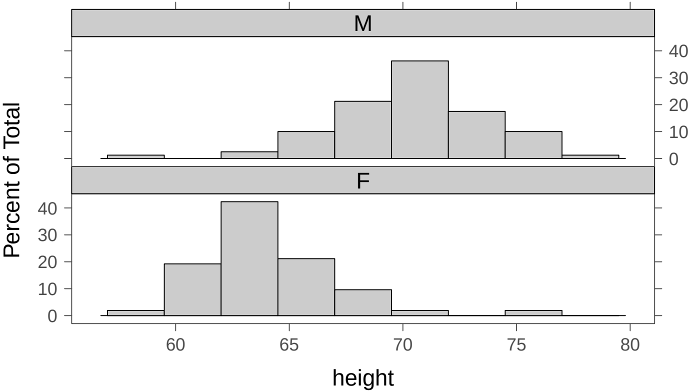
The distribution of heights II




```
gender
 F  M
52 80
```

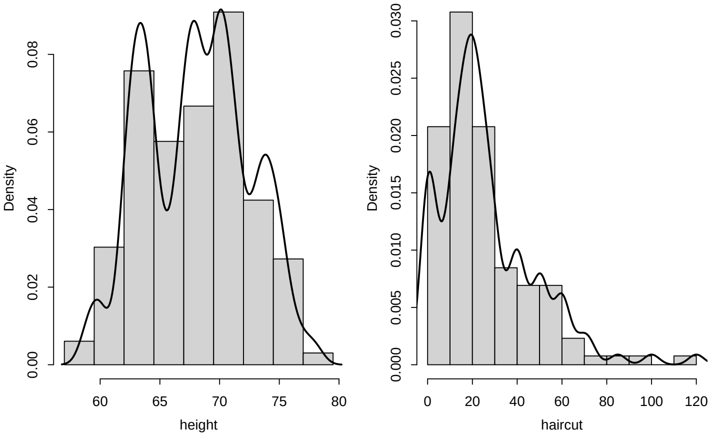
The distribution of heights III

Let's make the comparison based on percentages, not counts



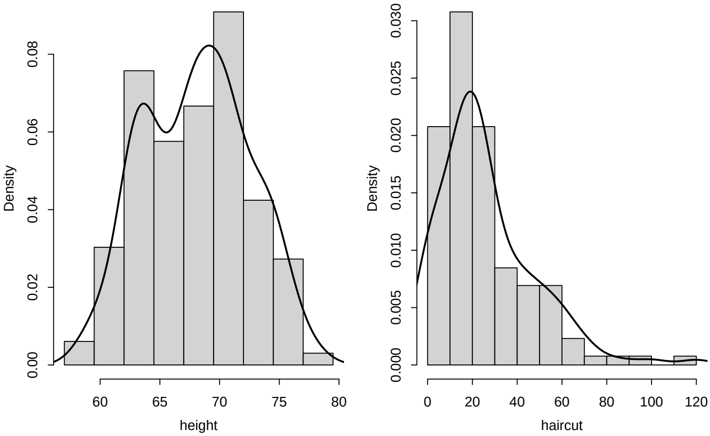
Getting a feel for the shape of a distribution I

Too much "detail"? More than is really available in the data?



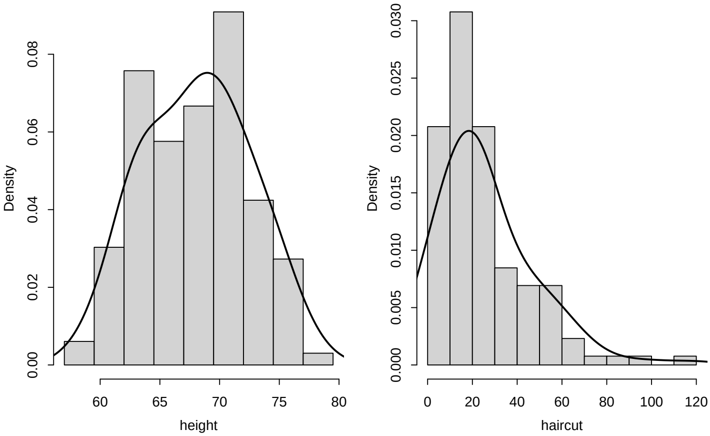
Getting a feel for the shape of a distribution II

The smoothing R does as default



Getting a feel for the shape of a distribution III

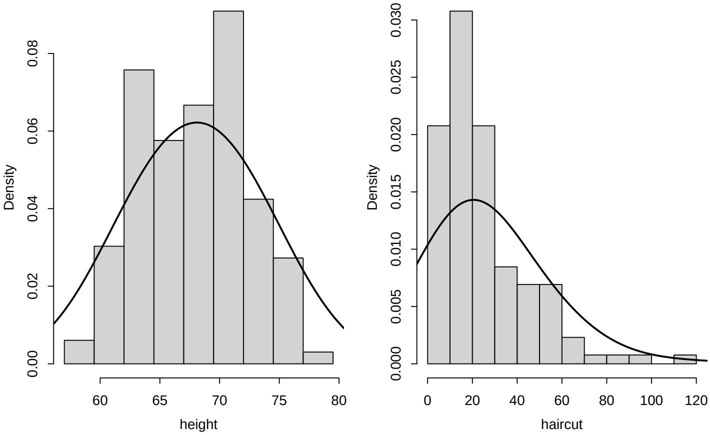
You could make smooth things out more to get a feel for shape



Getting a feel for the shape of a distribution IV

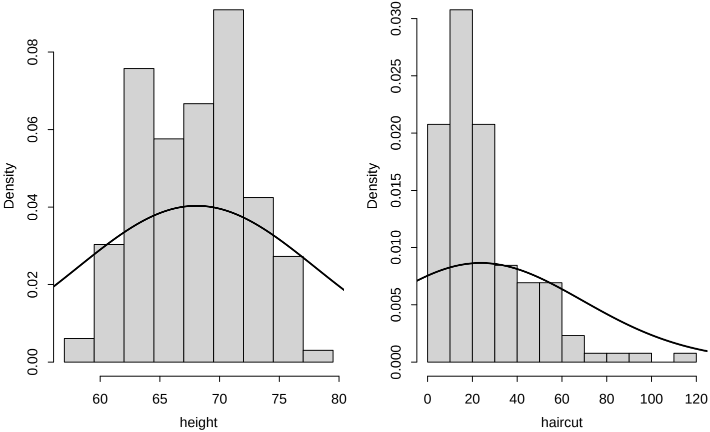
...and smooth some more. Too much?

The smooth curve no longer represents the shape?



Getting a feel for the shape of a distribution V

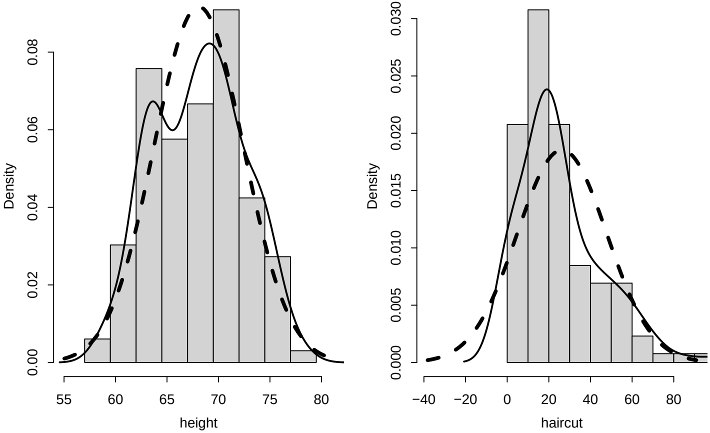
Way too much smoothing!



The "normal density" model I

The 68-95-99.7 rule for the normal distribution.

The normal density a good "fit" for these data distributions?



The "normal density" model II

I would have thought the distribution of height would be more symmetric and mound-shaped.

It seems to have two humps (bimodal).

We'll deal with that later...

The "normal density" model (haircut cost) I

The "normal density" model.

A good fit for these data distributions?

A numerical look

What percent of area under standard normal density is above/below 1?

The "normal density" model (haircut cost) II

```
pnorm(-1, m=0, s=1)

[1] 0.1587

pnorm(-1)  # the default is mean=0, sd=1 ("standard" normal)

[1] 0.1587

pnorm(1)

[1] 0.8413

1 - pnorm(1)

[1] 0.1587
```

The "normal density" model (haircut cost) III

What percent of area under *any* normal density is above/below 1 sd from mean?

```
mhair+shair
```

```
[1] 47.81
```

```
1 - pnorm(mhair + shair, m=mhair, s=shair)
```

```
[1] 0.1587
```

```
pnorm(mhair - shair, m=mhair, s=shair)
```

```
[1] 0.1587
```

The "normal density" model (haircut cost) IV

What percent of the observed data are right/left of 1 sd from mean?

```
n = length(na.omit(haircut));    n

[1] 130

sum(haircut >= mhair + shair, na.rm=TRUE) / n

[1] 0.1692

sum(haircut <= mhair - shair, na.rm=TRUE) / n

[1] 0.1462
```

The "normal density" model (haircut cost) V

What percent of the model/data are 2 sd to the right mean?

```
1 - pnorm(2)
```

```
[1] 0.02275
```

```
sum(haircut >= mhair + 2*shair, na.rm=TRUE) / n
```

```
[1] 0.04615
```

What percent of the model/data are 2 sd to the left of mean?

```
pnorm(-2)
```

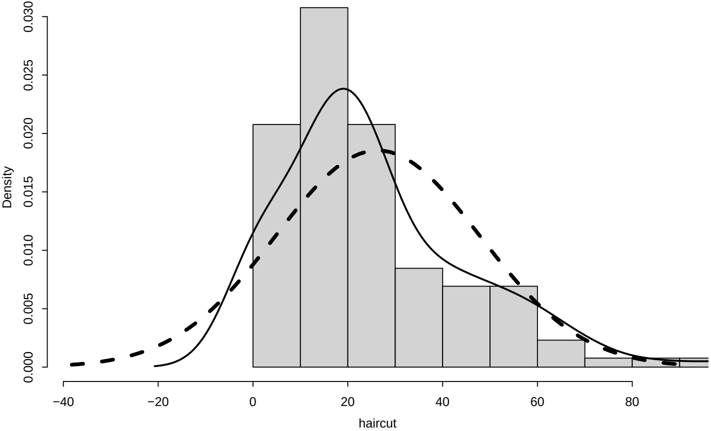
```
[1] 0.02275
```

```
sum(haircut <= mhair - 2*shair, na.rm=TRUE) / n
```

```
[1] 0
```

The "normal density" model (haircut cost) VI

Does this difference between data and model make sense?



Normal quantile plot I

A special plot can help us to compare all quantiles/percentiles of the data and the normal model (from 1% to 100%)

... the "normal probability plot" or "normal quantile plot"

Let's 1st draw this plot on our own and then let R draw the fancy plot

Normal quantile plot II

standard normal density quantiles (percentiles)

```
p = c(0.01, 0.025, 0.16, 0.25, 0.50, 0.75, 0.84, 0.975, 0.99)
modelQuantile = qnorm(p)
modelQuantile
```

```
[1] -2.3263 -1.9600 -0.9945 -0.6745  0.0000  0.6745  0.9945
[8]  1.9600  2.3263
```

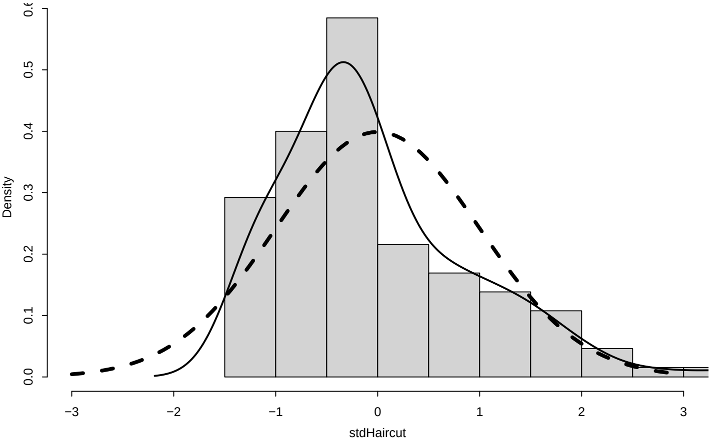
Strong suggestion (always do this)

Standardize data first to make comparison to "standard normal" model easier

z = (x - xbar) / s

Normal quantile plot III

`stdHaircut = (haircut - mhair) / shair`



Normal quantile plot IV

Data quantiles (percentiles)

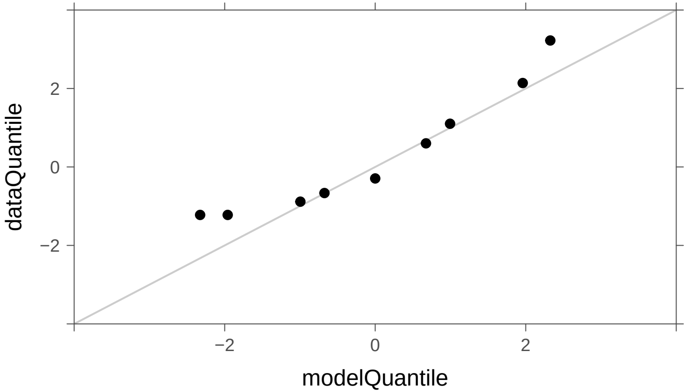
```
dataQuantile = quantile(stdHaircut, p, na.rm=TRUE)
dataQuantile
```

```
     1%    2.5%     16%     25%     50%     75%     84%
-1.2228 -1.2228 -0.8843 -0.6649 -0.2930  0.6019  1.1016
  97.5%     99%
 2.1395  3.2238
```

```
rbind(dataQuantile, modelQuantile)
```

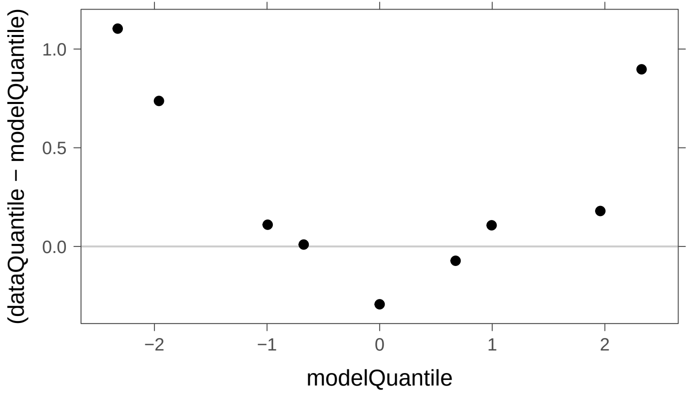
```
                 1%   2.5%     16%     25%    50%    75%
dataQuantile  -1.223 -1.223 -0.8843 -0.6649 -0.293 0.6019
modelQuantile -2.326 -1.960 -0.9945 -0.6745  0.000 0.6745
                 84% 97.5%   99%
dataQuantile  1.1016  2.14 3.224
modelQuantile 0.9945  1.96 2.326
```

Normal quantile plot V



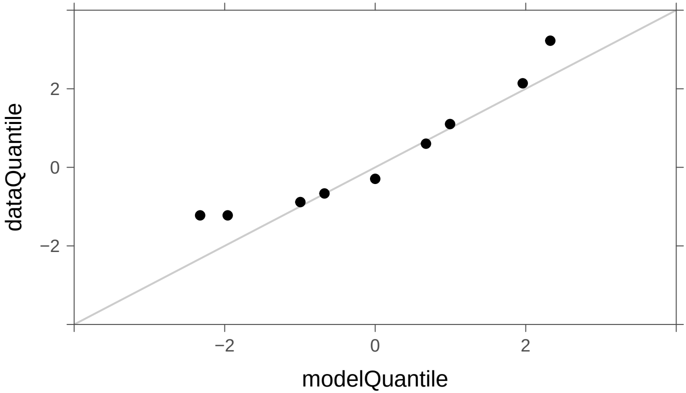
Normal quantile plot VI

I wish the "normal probability plot" was actually plotted like this (much easier to read)



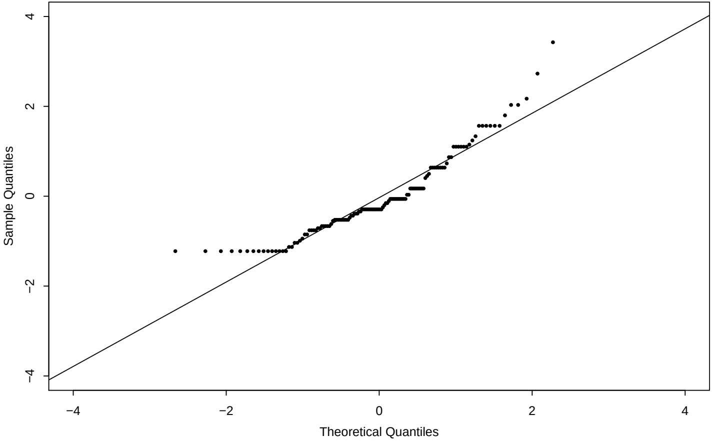
Normal quantile plot VII

But here is the style of plot traditionally called the "normal probability plot" or "normal quantile plot"



Normal quantile plot VIII

OK. Let R calculate the quantile (percentile) for ALL data points and compare to the quantiles (percentiles) of the normal distribution



interpreting a normal quantile plot I

How could we decide when a normal density might be a reasonable model (or not) for the population from which the data came?
Do the data fall "too far" from the line for it to make sense that the data came from a normal model?
What would n data points look like if they ACTUALLY came from a normal density model?

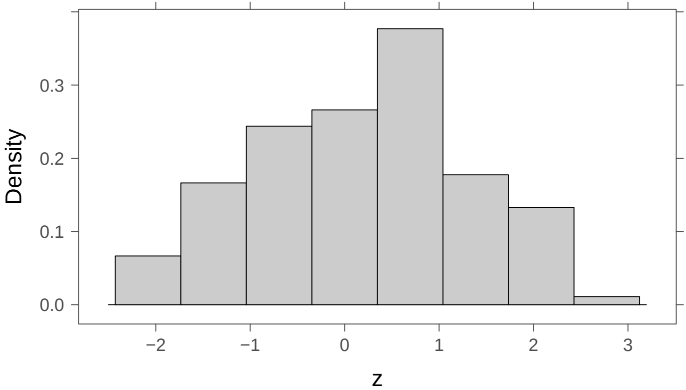
```
n
```

```
[1] 130
```

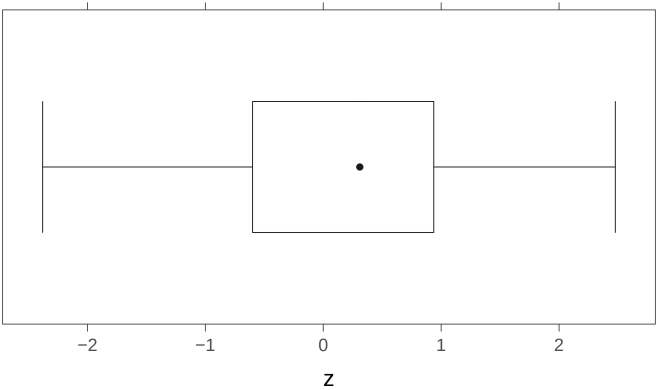
```
z <- rnorm(n)
```

```
                   1%   2.5%     16%     25%    50%    75%
               -2.282 -1.842 -1.0054 -0.5966 0.3099 0.9249
modelQuantile  -2.326 -1.960 -0.9945 -0.6745 0.0000 0.6745
                  84%  97.5%     99%
               1.4218  2.175   2.292
modelQuantile  0.9945  1.960   2.326
```

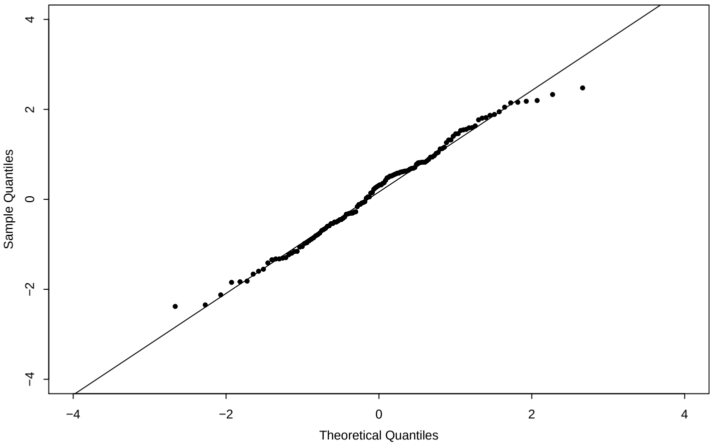
interpreting a normal quantile plot II



interpreting a normal quantile plot III

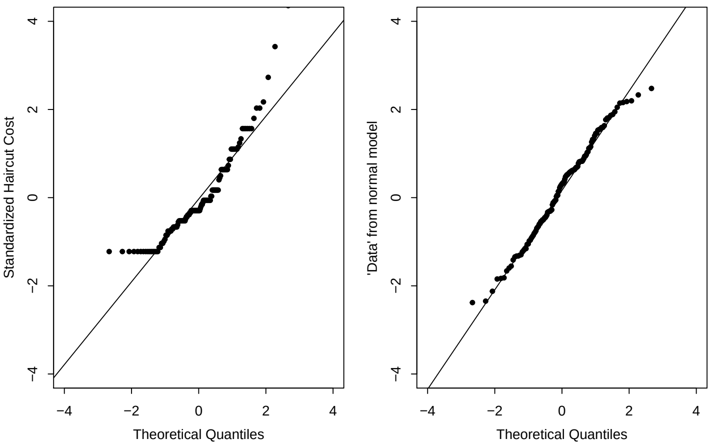


interpreting a normal quantile plot IV

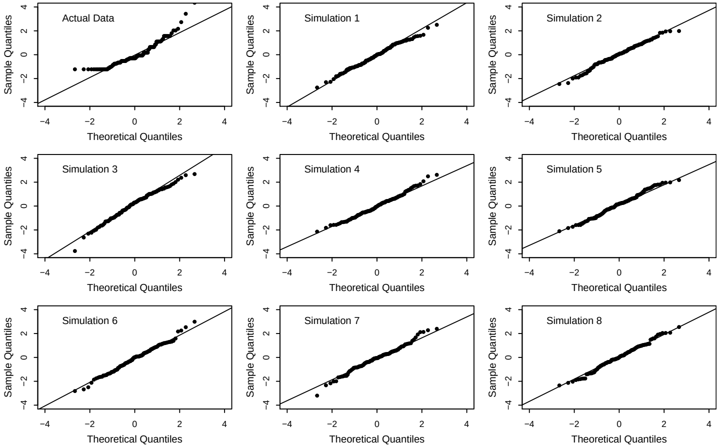


How do data actually from a normal population compare to the (standardized) haircut cost data we actually observed?

interpreting a normal quantile plot V

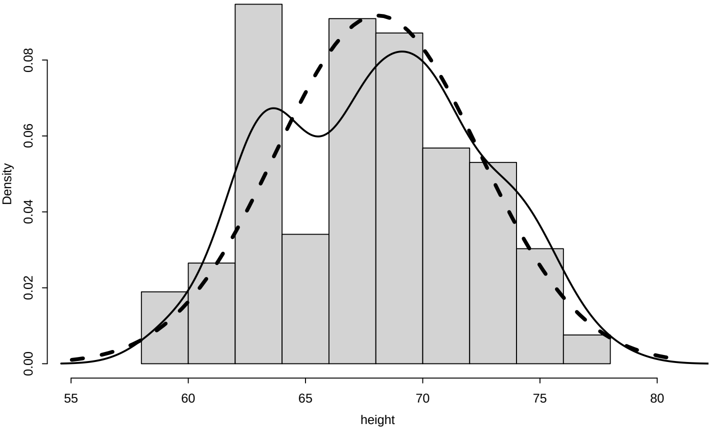


interpreting a normal quantile plot VI

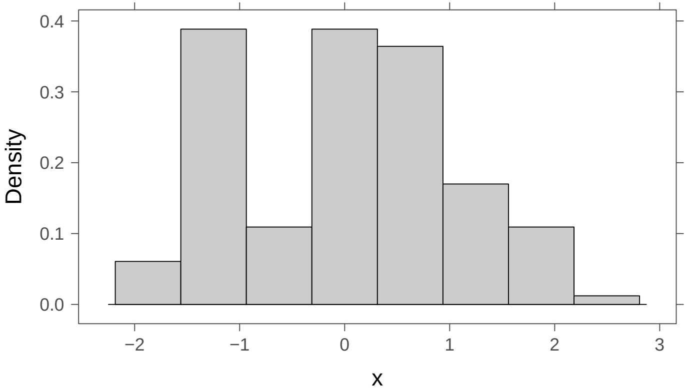


The "normal density" model (heights) I

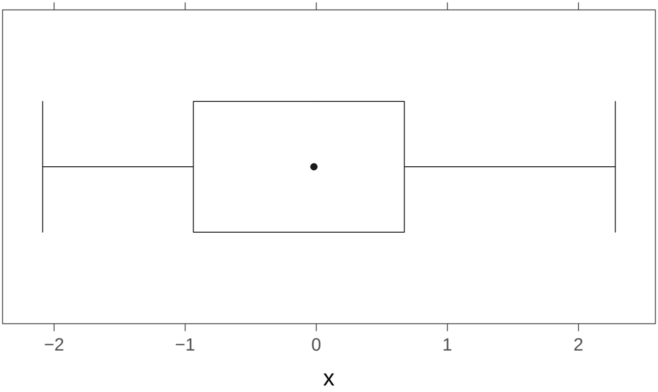
From past experience, I would expect the population distribution is approximately normal



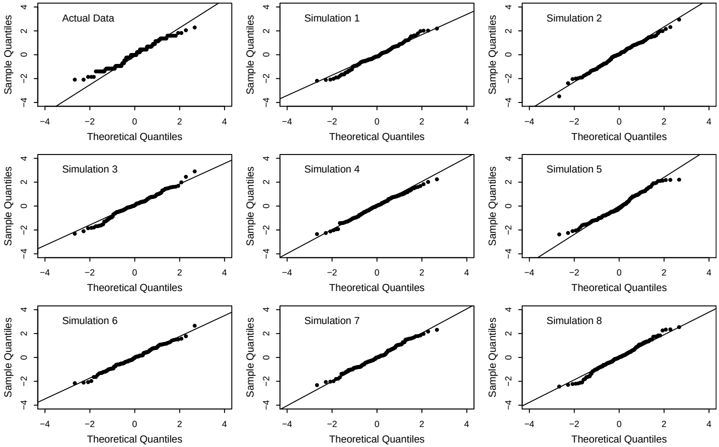
The "normal density" model (heights) II



The "normal density" model (heights) III



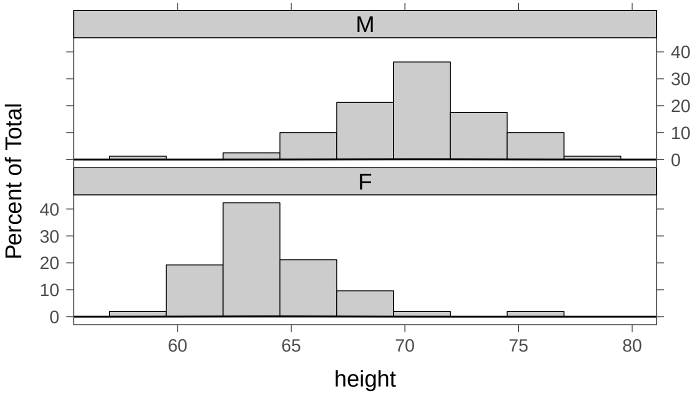
The "normal density" model (heights) IV



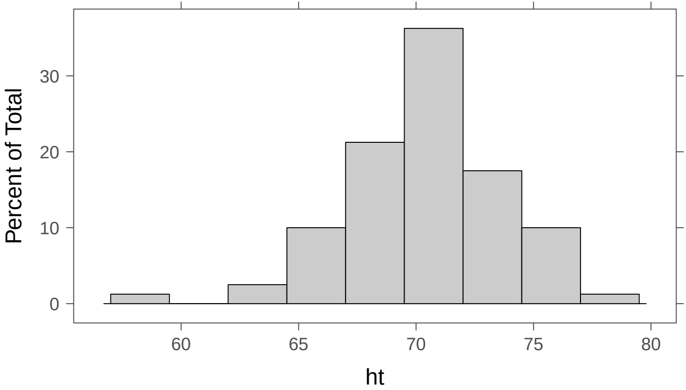
In my experience, height is pretty much symmetric and mound-shaped

What is the distribution by gender?

The "normal density" model (heights) V

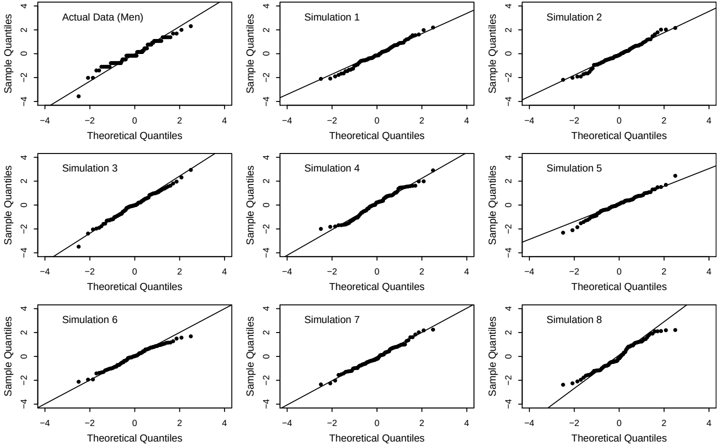


The "normal density" model (male heights) I

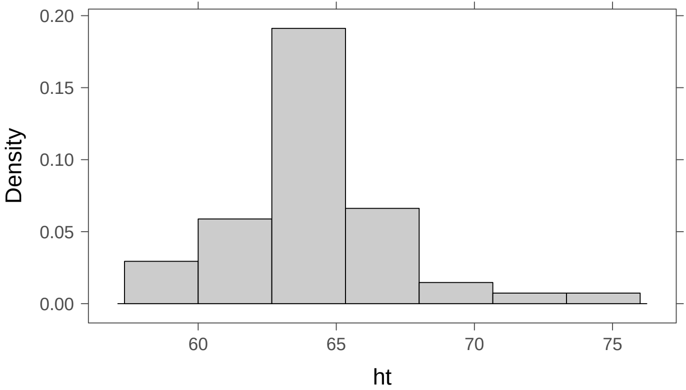



```
[1] 80
```

The "normal density" model (male heights) II



The "normal density" model (female heights) I




```
[1] 51
```

The "normal density" model (female heights) II