Lectures: Intermediate Applied Statistics
Fall 2019

Copyright 2019 by James Bernhard
1 DESCRIPTIVE STATISTICS

Tuesday, September 3

Wednesday, September 4

Friday, September 6

Monday, September 9

Tuesday, September 10

Wednesday, September 11

Friday, September 13

2 INFERENTIAL STATISTICS
Monday, September 16

Tuesday, September 17

Wednesday, September 18

Friday, September 20
Statistics is the study of how to collect, present, and interpret data.

All three are intertwined, but in this class we’ll focus more on the last two.
Two broad areas within statistics:

1. descriptive statistics, which looks at how to describe the data that you have.
2. inferential statistics, which looks at what can be inferred from the data that you have.

We’ll start the semester by looking at descriptive statistics.
The main computational tool that we’ll use is R.

- R is an implementation of the S language, which was developed by John Chambers in 1976.
- R was initially released by Robert Gentleman and Ross Ihaka in 1995, and the first stable beta version became available in 2000.
- R is currently maintained and developed by the R Core Development team (which includes John Chambers).
- R is free and open source.

R is available at r-project.org.
The R interface is not very user-friendly, which is what led to the development of RStudio by RStudio, Inc.

- RStudio is an IDE (*integrated development environment*) for R that offers a much more user-friendly interface.
- RStudio’s beta release was in 2011; Version 1.0 was released in 2016.
- RStudio is available in many forms, some free and some commercial. We will use a free version in this class.

RStudio is available at [rstudio.com](http://rstudio.com).
R and RStudio are available for all major operating systems.

To install R on your computer for this class:

1. Install R by following a download link at r-project.org.
2. Install RStudio by following a download link at rstudio.com.

That’s it!
To use R:

1. Start RStudio on your computer.
2. Type commands at the command line (which is a > followed by a blinking cursor) in the Console tab, pressing Enter to execute the command.
R can be used for basic arithmetic: +, -, * (multiply), / (divide), ^ (caret, or shift-6: raise to a power), and \texttt{sqrt}() (square root).

\begin{verbatim}
> 2+3
[1] 5
> (1+3)^2
[1] 16
> \texttt{sqrt}(9)
[1] 3
\end{verbatim}

The > indicates that what follows is typed into R.

The [1] indicates that the output is a vector, and this is the first (and here, the only) entry of that vector.
In R, data is often stored as vectors. To create a vector, use `c()` (for *combine* or *concatenate*).

```r
> c(1, 4, 5, 2)
[1] 1 4 5 2
> c("fish", "dog", "squirrel")
[1] "fish"   "dog"    "squirrel"
```

You need quotation marks in the second example because the entries are not numbers.
Many functions in R are *vectorized*, or automatically applied to each individual entry of a vector, as in:

```r
> c(1,2,3) + c(4,5,6)
[1] 5 7 9
> c(1,4,5,2)^2
[1] 1 16 25 4
> c(1,2,3,4) * c(10, 100, 1000, 10000)
[1] 10 200 3000 40000
```
The `min()` and `max()` functions compute the minimum and maximum values in a numerical vector:

```r
> min(c(4,2,6,8))
[1] 2
> max(c(4,2,6,9,2,1))
[1] 9
```
If you type a line that is incomplete, or if code on these lecture slides takes up more than one line, you’ll see a + instead of a > at the start of each line after the first:

```r
> max(c(1, 3, 5, 7, 9, 11, + 13, 15, 17, 19, 21))
[1] 21
```

If this happens unintentionally when you’re entering code at the command prompt, you can press the `Esc` key to get out of it.

Often the cause of such a problem is mismatched parentheses.
If you’re going to use a vector (or any other form of data in R) more than once, you should assign the vector to a variable.

For this, use `<-` (*less than* followed by *hyphen*), as in:

```r
> frogHeights <- c(1,4,5,2)
```

To see that it is indeed stored under that name:

```r
> frogHeights
[1] 1 4 5 2
```

To compute its maximum value:

```r
> max(frogHeights)
[1] 5
```
Variable names:

- Can contain letters, numbers, underscore characters, and periods.
- Cannot contain spaces or hyphens.
- Cannot begin with numbers.
- Are case sensitive, so \texttt{fish} is not the same as \texttt{Fish} or \texttt{fiSH}.
Guidelines for variable names:

- Choose variable names that are *meaningful* (such as `frogHeights`).
- Avoid variable names that are generic (such as `data`).
- For variables that will be used a lot, avoid variable names that are extremely short (such as `x`), as these are not meaningful.
- Avoid having variables that differ only in capitalization (such as `Fish` and `fish`).
- Be consistent. I’ll use *camel casing* throughout: each new word after the first begins with a capital (as in `frogHeights`).
Readability of R code is very important in a statistical analysis.

If you can’t read your code in six months, you can’t modify or revise your code in six months, and you might not even be able to use it.

In order to maintain, modify, and update your statistical analyses, it is important to write a clearly commented R script.

Just entering commands at the command prompt is not enough.
To make a new R script in RStudio, select the menu item **File/New File/R Script**. This will make a new blank R script in the upper left corner of RStudio.

You can continue to enter commands in the console window (if you click on it to make it active), and when you are satisfied that you have them, then you can write them in the R script window (after clicking on it to make it active).

You can execute commands in the R script window by clicking on **Run** (for the current line or selection) or **Source** (for the whole script) in the upper right corner of the R script window.
To save an R script, select the menu item **File/Save** (or **File/Save As**).

As with most writing on the computer, you should save your R script often, so that you don’t lose your recent work if the battery runs out, etc.

RStudio will probably tack on the .R file extension for you automatically, but if it doesn’t, then you should enter that into the file name yourself.
To help your R scripts be readable by human beings, comments are crucial. The comment character in R is `#`.

The rest of the line following a `#` is ignored by R.

You can use `#` to help with visual formatting too.

Also, since most whitespace is ignored by R, use that to help with visual formatting as well.
For example, in this class, every assignment’s R script should begin with a comment header to indicate the assignment name, whose assignment it is, and the date when the R script was written:

```
############################
# Homework 0
#
#
# by James Bernhard
# 8/26/2019
############################
```
After a couple of blank lines to help readability, give an indication of which problem is next:

# Problem R.1
Then before each line of code, put a comment explaining why you are doing this. (You don’t have to explain what it does, since that is evident to anyone familiar with R.)

```
# Problem R.1
# make a vector to store frog heights (in inches)
frogHeights <- c(4, 5, 3, 3)

# find the tallest frog, to analyze later
largestHeight <- max(frogHeights)
```
Data usually comes in the form of a spreadsheet, such as from Excel or Google Sheets, etc.

R can’t read those spreadsheets directly, so we save those files as **csv** (comma separated value) files by using *Save as* in a spreadsheet program.

To read the csv file into R, you need to know where R is looking for it.

The name of the folder where R is currently looking for files is the **working directory**.
If you open an R script file by double-clicking on it, the working directory will be the folder that the file was in.

Or, to set the working directory yourself in RStudio:

1. Click on the **File** tab in the lower right window.
2. Click on the little ... near the upper right corner of the lower right window.
3. Navigate to the folder that you want to be the working directory.
4. Click on the small downward-pointing triangle to the right of **More** near the top of the lower right window.
5. Select **Set as working directory**.

Your working directory will last until you change it or close RStudio.
To set the default working directory in RStudio:

1. Select **Tools/Global options** from the menu at the top.
2. Under **General**, click on the **Browse** button just to the right of the **Default working directory (when not in a project)** option.
3. Navigate to and select the folder you want.
4. Click **OK**.

The folder you selected will be the working directory whenever you open RStudio.

But if you open RStudio by double-clicking on an R script file, the working directory is still automatically set to that file’s folder.
Once your working directory is set as the folder containing your csv data file, you can read it into R with:

```
> myDataFrame <- read.csv("myFile.csv")
```

Here `myDataFrame` is the name that you want the data file to have in R, and `myFile.csv` is the name of the csv data file you’re reading in.

The file name needs to have quotation marks around it so that R doesn’t think that it’s a variable name.

The file name needs to have the `.csv` extension even if you don’t see that in the file name on your computer.
In order for `read.csv("myFile.csv")` to work properly, your csv file must:

- Have the column names in the first row.
- Have one observation per row after the first row.
- Be named `myFile.csv` (although you may not see the `.csv` extension in the file name).
- Be in the folder that is current working directory.

Because of the second item, it is important that all the columns have the same number of entries.

The `read.csv("myFile.csv")` function has lots of options to read different types of files. We’ll discuss some of those options later.
To check that your data file read into R correctly, you can view it with

```r
> View(myDataFrame)
```

This will open a tab displaying `myDataFrame`.

Hovering on the column name will display additional information about the column, which we will explore later.

Your data is not editable, although the view can be sorted.

At the bottom of the window, you can see the number of rows (entries) and columns in your data frame.

You can get back to the **Console** tab by clicking on it.
For example, to view the built-in `sleep` data frame:

```
> View(sleep)
```
It’s sometimes useful to be able to compute the number of rows and columns in a data frame. For this, use `nrow(myDataFrame)` and `ncol(myDataFrame)`.

For example, to find the number of rows and columns in the built-in `sleep` data frame:

```r
> nrow(sleep)
[1] 20
> ncol(sleep)
[1] 3
```

This should agree with what you saw with `View(sleep)`.
Sometimes we need to work directly with a column of a data frame.

The usual way to access a column `myColumn` in a data frame `myDataFrame` is:

```r
> myDataFrame$myColumn
```

For example, here is the built-in `anscombe` data frame:
> anscombe

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><code>x1</code></td>
<td><code>x2</code></td>
<td><code>x3</code></td>
<td><code>x4</code></td>
<td><code>y1</code></td>
<td><code>y2</code></td>
<td><code>y3</code></td>
<td><code>y4</code></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>8.04</td>
<td>9.14</td>
<td>7.46</td>
<td>6.58</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>6.95</td>
<td>8.14</td>
<td>6.77</td>
<td>5.76</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>8</td>
<td>7.58</td>
<td>8.74</td>
<td>12.74</td>
<td>7.71</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>8.81</td>
<td>8.77</td>
<td>7.11</td>
<td>8.84</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>8</td>
<td>8.33</td>
<td>9.26</td>
<td>7.81</td>
<td>8.47</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>8</td>
<td>9.96</td>
<td>8.10</td>
<td>8.84</td>
<td>7.04</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>7.24</td>
<td>6.13</td>
<td>6.08</td>
<td>5.25</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>19</td>
<td>4.26</td>
<td>3.10</td>
<td>5.39</td>
<td>12.50</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>8</td>
<td>10.84</td>
<td>9.13</td>
<td>8.15</td>
<td>5.56</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>4.82</td>
<td>7.26</td>
<td>6.42</td>
<td>7.91</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>8</td>
<td>5.68</td>
<td>4.74</td>
<td>5.73</td>
<td>6.89</td>
</tr>
</tbody>
</table>

To access the `y1` column, use:
(The vector has all 11 entries, although they are off the slide here.)
To access a specific set of columns, use
subset(myDataFrame, select=c("myColumn1", "myColumn2", etc.)).

For example, to access the x1, y1, x2, y2 columns of anscombe:
```r
> subset(anscombe,
       select=c("x1", "y1", "x2", "y2"))
```

<table>
<thead>
<tr>
<th></th>
<th>x1</th>
<th>y1</th>
<th>x2</th>
<th>y2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>8.04</td>
<td>10</td>
<td>9.14</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>6.95</td>
<td>8</td>
<td>8.14</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>7.58</td>
<td>13</td>
<td>8.74</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>8.81</td>
<td>9</td>
<td>8.77</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>8.33</td>
<td>11</td>
<td>9.26</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
<td>9.96</td>
<td>14</td>
<td>8.10</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>7.24</td>
<td>6</td>
<td>6.13</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>4.26</td>
<td>4</td>
<td>3.10</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>10.84</td>
<td>12</td>
<td>9.13</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>4.82</td>
<td>7</td>
<td>7.26</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>5.68</td>
<td>5</td>
<td>4.74</td>
</tr>
</tbody>
</table>
To access a set of rows satisfying a condition, use `subset(myDataFrame, myCondition)`.

For example, to access only those rows of the built-in `sleep` data frame for which `extra` is less than 0:

```r
> subset(sleep, extra < 0)
   extra group ID
 2  -1.6  1  2
 3  -0.2  1  3
 4  -1.2  1  4
 5  -0.1  1  5
15 -0.1  2  5
```
To specify conditions, you can use:

- `>` (or `>=`) greater than (or less than or equal to)
- `<` (or `<=`) less than (or less than or equal to)
- `==` equals (note the double equals sign!)
- `!=` does not equal
- `&` and
- `|` or
For example:

```r
> subset(sleep, extra >= 0 & extra <= 1)
```

will include all rows of `sleep` for which `extra` is between 0 and 1, including those endpoints.

Also,

```r
> subset(sleep, extra < 0 | ID==2)
```

will include all rows of `sleep` for which `extra` is less than 0, as well as those for which `ID` equals 2.
You can combine accessing subsets of rows and columns as well:

```r
> subset(sleep, extra < 0 | ID==2, 
       select=c("ID", "extra"))

   ID extra
2  2   -1.6
3  3   -0.2
4  4   -1.2
5  5   -0.1
12 2   0.8
15 5   -0.1
```
In order to do statistics, we need some concepts and terminology.

**Definitions**

A *random procedure* is a repeatable procedure that has a specific set of possible results. Conducting a random procedure once is called a **trial** of that random procedure. The result of a random procedure is called an **outcome**. The set of all possible outcomes is called the **sample space** of the random procedure.
**Example**

We usually model flipping a coin as a random procedure, with outcomes $H$ (in which the coin lands *heads*) and $T$ (in which the coin lands *tails*). The sample space is \{H, T\}. 
EXAMPLE

Drawing at random a student out of all students enrolled at the University of Puget Sound can be viewed as a random procedure. What is an outcome, and what is the sample space?
An outcome is a specific student, and the sample space is the set of all the students enrolled at the University of Puget Sound.
**Example**

Drawing at random 10 students out of all students enrolled at the University of Puget Sound can be viewed as a random procedure. What is an outcome, and what is the sample space?
An outcome is a set of 10 students enrolled at the University of Puget Sound. The sample space is the set of all possible sets of 10 students who are enrolled at the University of Puget Sound.
**Definition**

A variable is a characteristic that can be measured.
EXAMPLE

The height of a squirrel is a variable. The height of a student selected at random from all currently enrolled students at the University of Puget Sound is a variable. The average height of 10 students selected at random from all currently enrolled students at the University of Puget Sound is a variable.
DEFINITIONS

A variable is called

- **numerical** if it would make sense to use it in the context of arithmetic operations.

- **categorical** if it is not numerical and if it has a finite set of possible values that it can take on. A value that a categorical variable can take on is called a *level*.

- **Boolean** or **logical** if it can take on only two values: **TRUE** and **FALSE**.

In R, categorical variables are called **factors**, a term is also used in other ways in statistics.
If you use `View(myDataFrame)` to view a data frame in R, hovering over the triangles next to the column headers will tell you the type of variable that the column is.

For example, try this with the built-in `sleep` data frame:

```r
> View(sleep)
```

Hovering shows that the `extra` column is numerical, while the `group` and `ID` columns are categorical (factors), even though they both have numbers in them.
**Definitions**

To do whatever is necessary to produce a value of a variable is to **observe** the variable. The variable’s value resulting from observing the variable is called an **observation** of the variable.

**Example**

Let $X$ be the variable whose value is the average height of 10 students selected at random from all currently enrolled students at the University of Puget Sound. To observe $X$, we select 10 students at random from all currently enrolled students at the University of Puget Sound. The average height of those students is the observation of $X$ that results.
DEFINITION
A set of observations of a variable is called independent if knowing the value of one observation gives no additional information (even probabilistic information) about any of the other observations in the set.

EXAMPLE
Let $X$ be the variable whose value is 0 if a coin flip lands tails and 1 if heads. If we observe $X$ by flipping the coin twice, we would ordinarily model the resulting observations as independent, since neither gives any information about the other.
**Example**

Let $X$ be the variable whose value is the number of a card that we pull from a standard deck (aces being 1 and face cards being 11 through 13). If we observe $X$ by choosing two cards from the top of the deck, the resulting observations are *not* independent. For example, the second observation can’t be the same as the first. Knowing the first card gives some information about what the second might be.

If we have enough cards in the deck, the information gleaned from knowing the first card is usually deemed to be negligible, so we model the observations as being (approximately) independent.

As a rough rule of thumb, we often view this as “close enough” to independent when we observe less than 10% of the set that we are drawing from.
DEFINITIONS

To **sample** a variable is to generate a collection of independent observations of that variable. Similarly, a **sample** of a variable is a collection of independent observations of that variable. The number of observations in a sample is called the **size** of the sample, and is frequently denoted generically by $n$.

Note that the term **sample** can also be used to refer to the **collection of individuals** themselves that have been selected in order to observe a particular variable.
**Definitions**

A **random variable** is a numerical variable whose value is determined by the outcome of a random process. A **discrete** random variable has only a finite (or at most countable) number of possible values. A **continuous** random variable has values that include an interval in the real number line.

Any random variable can be considered simply as a variable, simply by ignoring the fact that it comes from a random process.

To specify a random variable, you should specify both its random process and how its value is determined by this process.
**Example**

The height of a squirrel is not a random variable, since it doesn’t involve a random process. The height of a student selected at random from all currently enrolled students at the University of Puget Sound is a random variable. The average height of 10 students selected at random from all currently enrolled students at the University of Puget Sound is a random variable.
**DEFINITION**

The **distribution** of a random variable is the collection of all information about the probability that the random variable will take on values (both individual values and ranges of values).

For a discrete random variable, we give the distribution by listing the variable’s possible values and the probabilities of those values.

In this class, we’ll mostly use continuous random variables though.

Their distributions are usually described by a type of function that we now define...
**Definition**

The **probability density function** of a continuous random variable $X$ is a function $f$ with that property that for any two real numbers $a$ and $b$, the area under the graph of $f$ between $a$ and $b$ equals the probability that $a < X \leq b$.

To specify the distribution of a continuous random variable, we ordinarily give its probability density function.

By the way that the probability density function is defined, the value of $X$ will tend to be in the regions where the probability density function is large.
By its definition, a probability density function:

- Is never negative.
- Must have a total area underneath it and above the \( x \) axis equal to 1.
- Can be arbitrarily large, but only for short stretches (because of the first two properties).
- Must tend to 0 as \( x \to -\infty \) and as \( x \to \infty \).
**Definition**

A **statistic** is a random variable whose random process is that of sampling some other random variable (often with a specific sample size).
EXAMPLE

The average height of 10 students selected at random from all currently enrolled students at the University of Puget Sound is a statistic: its value depends on sampling the height of a randomly selected student (with sample size 10).
**Definition**

The **expected value** (or *random variable mean*) of a random variable $X$ is denoted by $E(X)$. It captures the idea of the *center* or *average* of the distribution of $X$.

For a discrete random variable $X$ with possible values $x_1, \ldots, x_k$ having probabilities $p_1, \ldots, p_k$, the random variable mean is

$$E(X) = p_1 x_1 + p_2 x_2 + \cdots + p_k x_k.$$  

For a continuous random variable with probability density function $f$, the random variable mean is

$$E(X) = \int_{-\infty}^{\infty} x f(x) \, dx.$$
**DEFINITION**

The **random variable standard deviation** of a random variable $X$ is denoted by $\sigma_X$. It captures the idea of *variability* or *volatility* of $X$.

The standard deviation of $X$ is defined as:

$$\sigma_X = \sqrt{E((X - E(X))^2)}.$$
The *random variable variance* of a random variable $X$ is $(\sigma_X)^2$.

It is mathematically more convenient than the random variable standard deviation, but because of its units it is harder to interpret.

We will not use it much in this class.
**Definition**

For any $p$ with $0 < p < 1$, the **random variable $p$ quantile** of a random variable $X$ can be thought of as the real number $q$ for which $P(X \leq q) = p$.

In actuality, there may be many real numbers with this property, so a more rigorous definition is needed.

However, for our purposes in this class, this conceptual definition will suffice.

A (random variable) **quantile** is simply a (random variable) **percentile** divided by 100.
The random variable 0.5 quantile is called the **random variable median**. The random variable 0.25 quantile is called the **random variable first quartile**. The random variable 0.75 quantile is called the **random variable third quartile**.
A random variable $X$ is said to have a **normal distribution** with mean $\mu$ and standard deviation $\sigma$ if its probability density function is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}.$$  

In this case, we write $X \sim N[\mu, \sigma]$. The normal distribution with mean 0 and standard deviation 1 is called the **standard normal distribution**.
Graphs of normal distribution functions are *bell-shaped*:
Facts about normal distributions:

- The mean and median of a normal distribution are both at the point directly beneath the top of the bell.
- Their probability density functions change concavity at $\mu - \sigma$ and $\mu + \sigma$. 

\[ \mu - \sigma \text{ and } \mu + \sigma \]
**PROPOSITION (68-95-99.7 rule)**

The probability that an observed value of a normally distributed random variable will lie within 2 standard deviations of the variable's mean is approximately 0.95.

This means that on average for a normally distributed random variable:

- About 1 in 20 observed values of $X$ will lie further than 2 standard deviations from the mean.
- About 3 in 1000 observed values of $X$ will lie further than 3 standard deviations from the mean.
Computer lab.
**Definition**

Let \( x_1, x_2, \ldots, x_n \) be a sample of a random variable \( X \). Then the sample mean of this sample is derived from the values in the sample, and it gives information about where the center of the sample is. It is defined to be

\[
\frac{x_1 + x_2 + \cdots + x_n}{n}.
\]

The sample mean is usually what most people think of as the “average” of the values in the sample (but average is a somewhat vague and nontechnical term).

If a penny were placed on a horizontal real number line at the values in the sample, the point at which the line would balance is the sample mean.
In R, the sample mean of a vector is computed with the `mean(myVector)` function. For example:

```r
> mean(sleep$extra)
[1] 1.54
> frogHeight <- c(1,4,5,2)
> mean(frogHeight)
[1] 3
```

When you interpret this in a write-up, make sure that you have the correct units. Without the units, the result is meaningless.

The units of the sample mean are the same as the units of the observations in the sample.
R does not keep store or keep track of units in any of its computations.

So it is up to you to keep track of the units in the computations that you do!

Be sure to include units in your write-ups whenever you are doing statistics. Your results are meaningless without them.
**Definitions**

Sample standard deviation and sample variance are statistics that give information about the variation of the values in a sample about its sample mean.

The sample standard deviation of a sample $x_1, x_2, \ldots, x_n$ is defined to be

$$\sqrt{\frac{(x_1 - m)^2 + (x_2 - m)^2 + \cdots + (x_n - m)^2}{n - 1}}.$$  

The sample variance is the square of the sample standard deviation.
Sample standard deviation and sample variance give exactly the same information.

Sample variance is usually more convenient mathematically, but the sample standard deviation has units that are easier to interpret.

We will usually use the sample standard deviation rather than the sample variance in this course.
The sample standard deviation is computed with the `sd(myVector)` function. For example:

```r
> sd(sleep$extra)
[1] 2.01792
> frogHeight <- c(1,4,5,2)
> sd(frogHeight)
[1] 1.825742
```

As usual, it is up to you to supply the units when you interpret the R output.

The units of the sample standard deviation are the same as the units of the observations in the sample.
**Definition**

For any $p$ with $0 < p < 1$, the **sample $p$ quantile** can be thought of as the real number $q$ for which the proportion of the sample that is less than or equal to $q$ equals $p$.

In actuality, there may be many real numbers with the property given here, so a more rigorous definition is needed.

However, for our purposes in this class, this conceptual definition will suffice.

A (sample) **quantile** is simply a (sample) **percentile** divided by 100.
**Definitions**

Several quantiles have their own names:

- The sample 0.25 quantile is called the **sample first quartile**.
- The sample 0.5 quantile is called the **sample median**.
- The sample 0.75 quantile is called the **sample third quartile**.

Also, the **five-number summary** of a sample consists of:

1. the minimum value in the sample,
2. the sample first quartile,
3. the sample median,
4. the sample third quartile, and
5. the maximum value in the sample.
To compute the sample median, use the `median(myVector)` function:

```r
> median(sleep$extra)
[1] 0.95
> frogHeight <- c(1,4,5,2)
> median(frogHeight)
[1] 3
```

Notice how R interpolates: since the sample has an even number of observations, there is no “middle value”, so the median is halfway between the two middle values.
To compute the sample $myProb$ quantile, use the `quantile(myVector, prob=myProb)` function.

For example, to compute the sample 0.3 quantile of the `extra` column of the built-in `sleep` data frame, use:

```r
> quantile(sleep$extra, prob=0.3)
  30%
0.07
```

The output of this function is labeled as a percentile rather than a quantile. To figure out which quantile it is, divide the percent value by 100.
If you don’t enter the `prob=myProb` argument, you will obtain the *five-number summary*:

```r
> quantile(sleep$extra)
       0%      25%      50%      75%     100%
-1.600   -0.025   0.950    3.400    5.500
```

The minimum and maximum values in the vector are labeled as *0%* and *100%* here.
**Definition**

A **density plot** of a vector is a graph whose height indicates how many observed values are near each value below the graph on the $x$ axis. It can be thought of intuitively as a smoothed out histogram. The $y$ values on this graph are always nonnegative.

Above points on the $x$ axis where there are lots of observed values nearby, the density plot will be tall.

Above points on the $x$ axis where there are not many observed values nearby, the density plot will be short.

If the vector is a sample of a random variable, then a density plot can be thought of as an estimate of the underlying random variable’s probability density function.
For example, here is a density plot of the `Height` column of the built-in `trees` data set:
What to look for in a density plot:

1. The **shape** of the plot.
2. The **center** of the plot.
3. The **spread** of the plot.
4. Any **outliers**.
Regarding the shape:

• How many modes does it have? (A mode is the $x$ coordinate of the top of a big hill. We often speak of unimodal and bimodal graphs, which have one or two big hills respectively.)

• Is it symmetric, skewed left, skewed right, or none of these? Symmetric means that each side (left and right) is approximately a mirror image of the other. Skewed left (right) means that the graph swings out to the left (right), with a big hill on the other side.
What can you say about this density plot’s shape?
It is unimodal, and it is skewed right.

At first glance, you might think it is bimodal, but the smaller hill on the right involves only 4 or 5 points, so it doesn’t count as a “big hill”.

Wednesday, September 11
What can you say about this density plot’s shape?
It is unimodal, and it is symmetric.

At first glance, you might think it is skewed left, but although the right side descends more smoothly, the two sides are similar enough that most statisticians would consider this to be symmetric.
For the center of a density plot, you might try estimating the mean or median.

This is not easy, but one can often come up with a rough estimate at least.

What can you say about the center of the previous two density plots?
In the first plot, we might estimate that the median is around 30 sq mi. (A direct computation reveals that it is actually at 29.0 sq mi.)

In the first plot, we might estimate that the mean is around 40 sq mi. (A direct computation reveals that it is actually at 34.8 sq mi.)

The median is easier to estimate: estimate how many points there are, and then go half that number of points up.

In any case, we know that the mean is to the right of the median due to the right skew: points far out to the right “pull” the mean much more than the median.
In the second plot, we might estimate that the median is around 70 ft to 80 ft. (A direct computation reveals that it is actually at 76.0 ft.)

In the second plot, we might estimate that the mean is also around 70 ft to 80 ft. (A direct computation reveals that it is actually at 76.0 ft.)

In any case, we know that the mean and median are pretty near to each other because this plot is symmetric.
For the spread, one could use the difference between the minimum and maximum as a measure, or one course use the difference between the first and third quartiles (which is more common).

One could also use the standard deviation (which is more informative in some cases than others).

In general though, just be aware of how much variability there is.
**DEFINITION**

An **outlier** is an observation that doesn’t fit the pattern of the rest of the observations.

Determining outliers is rather subjective.

Be clear about what pattern you’re referring to when you speak of an observation as an outlier.

Even when two people agree on the pattern, they may not agree that an observation is an outlier.

An observation may be an outlier relative to one pattern, but not relative to another.

You may have heard of some rules of thumb for detecting outliers, but there aren’t any that are very reliable, so we won’t use them here.
We will discuss outliers more later, but for now, please be aware that:

Outliers are *not* bad!

Sometimes outliers are the most important observations in your entire data set.
To make density plots in R, we use a function from the \textit{lattice} package.

Packages are additional modules for R. There are \textit{tons} of them, so you wouldn’t want to try to put all of them on your computer.

When you need one, install it on your computer. After that, it can be used on your computer without installing it again.
To install a package in RStudio:

1. Click on the **Packages** tab at the top of the window in the lower right corner of RStudio.
2. Click on the **Install** option in the upper left corner of the window in the lower right corner of RStudio.
3. In the dialog box that opens, select a repository if one isn’t already selected. Any one will do; nearer repositories will be faster.
4. Enter the name of the package in the **Packages** area of the dialog box.
5. Make sure that the **Install dependencies** box is checked.
6. Click **Install**.
To use functions from a specific package, type the name of the package, followed by two colons (that is, ::), and then the name of the function.

There are other ways to use functions from packages (such as with `library()`), but this is the method I would like you to use in this class.
This brings us back to making density plots in R.

To use the `densityplot( myColumnName, data=myDataFrame, xlab="My label")` function from the `lattice` package:

```
> lattice::densityplot(~myColumnName, data=myDataFrame, xlab="My label")
```

In this:

- **myDataFrame** is the name of the data frame containing the column to be plotted.
- **myColumnName** is the name of the column in that data frame that is to be plotted.
- **My label** is the x-axis label that you want.
The $x$ axis label should give the non-R name of the variable (with only the first letter of the first word capitalized) and the units, ordinarily but not necessarily abbreviated, in parentheses.

For example, to make a density plot of the Height column from the built-in trees data frame:
> lattice::densityplot(~Height, data=trees, 
  xlab="Height (ft)")
Density plots some big advantages over histograms:

- The individual data points are shown in the “rug” along the $x$ axis.
- Multiple histograms can plotted and viewed simultaneously. For this, include two additional arguments: `group` and `auto.key`. 
To view more than one density plot simultaneously:

```r
> lattice::densityplot(~myColumnName,
                       group=groupingColumnName,
                       data=myDataFrame, xlab="My label",
                       auto.key=TRUE)
```

Here `groupingColumnName` is the name of the column that tells which group each observation belongs to.

This column must be a `factor` in R.

For example:
> lattice::densityplot(~len, group=supp, 
                   data=ToothGrowth, xlab="Length (mm)",
                   auto.key=TRUE)
Density plots work well for looking at a single numerical variable at a time.

To look at two numerical variables together, we often use the following:

**Definition**

Let $X, Y$ be variables, and let $(x_1, y_1), \ldots, (x_n, y_n)$ be a set of $n$ simultaneous observations of these two variables. A **scatterplot** of $Y$ versus $X$ for these observations is a graphical display in which the points $(x_1, y_1), \ldots, (x_n, y_n)$ are plotted in a coordinate plane.

This allows us to explore *associations* between the two variables.

For example...
Since these two variables are always greater than 0, we often start the axes at 0, as in...
Friday, September 13

Descriptive statistics

Inferential statistics
Things to look for in scatterplots:

1. The **form** of the association between the two variables depicted.
2. The **direction** of the association between the two variables depicted.
3. The **strength** of the association between the two variables depicted.
4. Which points, if any, are **outliers**.
The *form* of an association refers to the general underlying relationship between the variables, such as:

- linear
- quadratic
- exponential
- sinusoidal
- etc.

Due other factors contributing fluctuations, a scatterplot will not take the *exact* shape of the underlying relationship, but will only approximate it.
What is the form of this association?
Most statisticians would classify this association as linear.

Many relationships are linear, at least to a first approximation: if you view a quadratic, exponential, or sinusoidal relationship up close near a point, it looks approximately linear.

In thinking about the form of an association, knowledge of the subject matter that the variables relate to often also helps.
The *direction* of an association applies only when the form has a direction (as do linear and exponential forms, but not quadratic, sinusoidal, etc.).

When the form has a direction, the direction is either:

- *positive*, meaning that when $X$ is larger, $Y$ tends to be larger,
- *negative*, meaning that when $X$ is larger, $Y$ tends to be smaller, or
- *zero*, meaning that when $X$ is larger, $Y$ exhibits no particular tendency.

When the form is linear, the direction corresponds to the slope being positive, negative, or zero.

What is the direction in the previous scatterplot?
The direction of the association shown in the previous scatterplot is positive.

This makes sense, since when the girth of a tree is larger, its height tends to be larger.
The *strength* of an association refers to how closely the points follow the underlying form of the association.

Looking at a graph, it may be difficult to assess the strength of the association in absolute terms, but often one can compare the strength of association in two graphs.

Even that can be tricky though, especially if the units on the graphs aren’t the same...
With a different scale, the strength looks different.
**Definition**

Let $X, Y$ be variables, and let $(x_1, y_1), \ldots, (x_n, y_n)$ be a set of $n$ simultaneous observations of these two variables. The **sample correlation** $r$ between these samples of $X$ and $Y$ is a real number between $-1$ and $1$ measuring of the strength of linear association exhibited in the samples. It has the properties that:

- The larger the $|r|$, the larger the linear association exhibited in the samples, the extremes being $r = 0$ (exhibiting no linear association) and $r = \pm 1$ (all points lie perfectly in a line).

- If the association is linear, then the sign of $r$ gives the direction of association.

- $r$ is unitless and can be used to compare associations between variables that don’t have the same units.

- Order doesn’t matter: correlation between samples of $X$ and $Y$ is the same as between samples of $Y$ and $X$. 
The sample correlation of the samples shown in the scatterplot of height versus girth is 0.519.

The sample correlation of the samples shown in the scatterplot of Anscombe \( y \) versus Anscombe \( x \) is 0.816.

Since the absolute value of their correlation is larger, the \( x_1 \) and \( y_1 \) samples in the \texttt{anscombe} data set exhibit a stronger linear association than the \texttt{Girth} and \texttt{Height} samples in the \texttt{trees} data set.
Caution: sample correlation measures only *linear* association, not any other type of association.
An even bigger caution:
Association does not imply causation.
For example: a tree’s girth being large does not cause the tree to be taller. Nor does a tree’s being tall cause the tree to have a wider girth.

The list of such examples goes on and on. Look for it throughout your life, and you’ll find lots of them.
To make a scatterplot of \texttt{myYColumn} versus \texttt{myXColumn} in R, use the \texttt{lattice::xyplot(myYColumn \sim myXColumn, data=myData)} function.

In this, \texttt{myXColumn} and \texttt{myYColumn} are the columns containing the simultaneous samples to be displayed, and \texttt{myData} is the data frame containing those columns.

This function also has \texttt{xlab} and \texttt{ylab} arguments to specify the labels on the $x$ and $y$ axes.

For example...
> lattice::xyplot(Height~Girth, data=trees, 
xlab="Girth (in)", ylab="Height (ft)")

Descriptive
statistics

Inferential
statistics
Before we get to the next topic, a little wrap-up from last week...
To compute a sample correlation in R, use `cor(myXVector, myYVector)`. For example, to compute the sample correlations for the two examples we have seen so far:

```r
> cor(trees$Girth, trees$Height)
[1] 0.5192801
> cor(anscombe$x1, anscombe$y1)
[1] 0.8164205
```
To view a matrix of scatterplots of all pairs of columns in a data frame, make a scatterplot matrix.

To do this in R, use the `lattice::splom(subset(myData, select=c("myColumn1", "myColumn2", etc.) ) )` function.

Here `myData` is the data frame with the columns to plot, and `myColumn1`, `myColumn2`, etc. are the columns to plot in the scatterplot matrix.

Since scatterplots make sense only for numerical variables, you should select only numerical columns for a scatterplot matrix.

For example, to plot the `mpg`, `hp`, and `wt` columns of the built-in `mpg` data set in a scatterplot matrix...
> lattice::splom(subset(mtcars, select=c("mpg", "hp", "wt")))
Inferential statistics

There are (at least) two ways to view inferential statistics:

1. *Population* point of view: inferential statistics is the branch of statistics in which we seek to learn about a population by observing a small number of individuals selected from the population.

2. *Random variable* point of view: inferential statistics is the branch of statistics in which we seek to learn about the distribution of a random variable by observing the random variable a number of times.

It's good to be familiar with both of these.
In inferential statistics, we use the term *sample* in (at least) two ways:

1. A sample *from a population* is the set of individuals that we actually observe and measure.
2. A sample *of a random variable* is the collection of observed values of a random variable.

If the random process behind a random variable is choosing a set of individuals to observe, then these two are related but still not the same.

The first refers to the individuals themselves; the second refers to the value that the random variable takes on more those individuals.
In this class, we’ll often look at a specific type of inferential statistics: *estimation*.

Estimation refers to one of (at least) two things:

1. *Population* point of view: the process of estimating the value of number(s) whose true value is derived from an entire population of individuals, by observing a sample from that population.

2. *Random variable* point of view: the process of estimating the value of number(s) whose true value is derived from the distribution of a random variable, by observing that random variable a number of times.

In either case, we refer to the number that we’re trying to estimate as a *parameter*. 
For the moment, we’ll focus on the random variable point of view.

In an upcoming lecture, we’ll talk about the population point of view as well.
In inferential statistics, we compute a single-number estimate (or **point estimate**) of a parameter first.

However, this doesn’t include information about the variability inherent in the estimation process, so we follow up our point estimate by doing one or both of the following:

1. conducting a *hypothesis test*
2. computing a *confidence interval*. 
Example

Suppose you’re wondering whether a particular way of flipping a coin is fair, meaning heads has probability 1/2. You flip the coin 10 times and find that 8 of the flips were heads. Is this evidence of unfairness?
In this case, we are testing the **null hypothesis** that *heads* has probability $1/2$, versus the **alternative hypothesis** that *heads* does not have probability $1/2$.

As a **test statistic**, we are using the count of *heads* out of 10 flips.

Since the alternative hypothesis includes both high and low probability of *heads*, then any number of heads far from 5 (what we’d expect on average for a fair coin) is evidence of unfairness.
We can compute that under the null hypothesis (meaning, if the null hypothesis is true, so the coin flipping is fair), the probability of getting a number of heads that is at least as far from 5 as 8 (our observed count) is equals 0.109.

Another way to state this is that the p-value that we obtained by conducting this test with our data is 0.109.

A small $p$-value (close to 0) indicates evidence against the null hypothesis.

A large $p$-value (close to 1) indicates a lack of evidence against the null hypothesis.
Usually evidence from a $p$-value of less than 0.05 is considered statistically significant, although the exact threshold (called the significance level) is chosen before conducting the hypothesis test.

In this case, since we found a $p$-value of 0.109 (which is greater than 0.05), we report that we did not find statistically significant evidence of unfairness.

If instead we had found a $p$-value of, say, 0.02, we would instead report that we found statistically significant evidence of unfairness.
In general, to conduct a hypothesis test, we need:

1. A **null hypothesis**, denoted by $H_0$, which is the hypothesis being tested.
2. An **alternative hypothesis**, denoted by $H_a$, which usually consists of all the cases not included in the null hypothesis. (There are exceptions to this, however.)
3. A **test statistic**, which is a statistic whose distribution under the null hypothesis is known.
4. If a “yes/no” result is desired, a **significance level**, denoted by $\alpha$ (the Greek lowercase *alpha*), which is a real number between 0 and 1.

The phrase **under the null hypothesis** means “if the null hypothesis is true”.
To conduct a hypothesis test:

1. observe the test statistic, meaning compute the value of the test statistic for the data that you have.

2. Use the alternative hypothesis to determine the meaning of the phrase “at least as extreme” (to be used in the next step). We’ll discuss this later.

3. Compute the **p-value**, which is the probability under the null hypothesis that the test statistic would be at least as extreme as the value you observed.

4. If a “yes/no” result is desired, compare the *p*-value to the significance level.

If a “yes/no” result is not desired for a hypothesis test, then the result to be reported is the *p*-value.
If a “yes/no” result is desired, the following terminology is used.

**DEFINITION**

Evidence against the null hypothesis is called **statistically significant** if the $p$-value is less than the significance level.
If a “yes/no” result is desired, you should report the $p$-value and:

- If the $p$-value is less than or equal to the significance level, report that *we did not find statistically significant evidence* against the null hypothesis.
- If the $p$-value is greater than the significance level, report that *we found statistically significant evidence* against the null hypothesis.

In reporting, should reword the non-italicized part to incorporate what the actual null hypothesis says, but do not reword the italicized part.

Later in the semester, we’ll talk about other ways to phrase this, but for now, you should leave the italicized part unaltered.
Keep in mind: the evidence that you find is always *against* the null hypothesis. It is never *in favor of* the null hypothesis.

Although evidence against the null hypothesis might be considered evidence in favor of the alternative hypothesis, you should still report it as evidence against the null hypothesis.

Again, later in the semester we’ll talk about other phrasings, but for now, the only two possibilities are that you either did or did not find statistically significant evidence against the null hypothesis.
Computer lab.

Descriptive statistics

Inferential statistics
In inferential statistics, we compute a point estimate (meaning a *single-number* estimate) of a parameter first.

Since that doesn’t include information about the variability inherent in the estimation process, we follow up our point estimate by doing one or both of the following:

1. conducting a *hypothesis test*
2. computing a *confidence interval*.

We have discussed what a hypothesis test is; now let’s look at confidence intervals.
**Definition**

Let $\beta$ (lowercase Greek *beta*) be a parameter whose true value is unknown. An interval in the real number line is called a **95% confidence interval for $\beta$** if it is produced by a process that has a 0.95 probability of producing an interval that contains the true value of $\beta$.

*Caution:* People sometimes try to simplify this by saying that a particular 95% confidence interval for $\beta$ has a 0.95 probability of containing the true value of $\beta$. This is incorrect.

*Caution:* It is incorrect to assert that something is *the* 95% confidence interval for $\beta$. For a parameter $\beta$, there are in general infinitely many 95% confidence intervals, so you shouldn’t claim that you have computed the one and only one.
**Note:** We can define an 80% (or other) confidence interval for $\beta$ similarly: it is produced by a process that has a 0.80 probability of producing an interval that contains the true value of $\beta$. The percentage chosen is called the **confidence level** of the confidence interval being computed.

**Caution:** There is no such thing as just “a confidence interval”. You need to give the other parts of the term as well: the confidence level and the parameter being estimated. The correct phrasing is “a 95% confidence interval for $\beta$”.

These cautions are not nitpicking: they are important distinctions to make in order to have a clear understanding of what a 95% confidence interval for $\beta$ is (and isn’t!).
Note that:

- Hypothesis tests require a hypothesis; confidence interval computations do not.
- Hypothesis tests don’t necessarily have to be estimations; confidence intervals necessarily are.
- The higher the confidence level, the wider the confidence interval (if all other things remain constant).
- As part of estimation, we might conduct a hypothesis test or compute a confidence interval or do both, but in any of these cases we ordinarily report a point estimate as well.
To return to the coin-flipping example from the previous lecture, suppose we obtain 7 of out 10 coin flips as *heads*.

This would ordinarily give us a point estimate of 0.7 for the probability of *heads*.
Also, we might compute (through a process we won’t discuss in this class) that a 95% confidence interval for the probability of heads is from 0.416 to 0.984.

We would report:

We estimate that the probability of heads is 0.7 (95% confidence interval from 0.416 to 0.984).

This example was unitless, but if there are units, you should always include them in your report, as in: we estimate the average gray squirrel tail length to be 208 mm (95% confidence interval from 169 mm to 257 mm).
What does the coin probability confidence interval above tell us?

It asserts that if we were to flip this coin 10 times over and over, and use the same method to produce a 95% confidence interval for the probability of heads, then on average about 95% of the intervals we would produce this way would contain the true probability of heads.

This does not assert that there is a 95% probability that the true probability of heads is between 0.416 and 0.984.

The difference is subtle, but you should be aware of it.

Statistical results need to be reported with precise technical wording. Otherwise they may not be correct!
Evidence of causation

We have discussed that association does not imply causation.

How *do* we establish causation?
There are many criteria for causation, one famous set (which you can look up if you’re interested) are *Hill’s criteria*.

In short, there are two things that you should know about causation:

1. A single study is never sufficient to establish causation. It takes many, many studies.
2. How you collect your data determines the extent to which your study can be used as evidence toward establishing causation.

Because of the first of these, if you report that your study has established causation, *you will be incorrect*. 

**Descriptive statistics**

**Inferential statistics**

**Wednesday, September 18**
Two common types of study are:

- **a randomized experiment**, in which the researchers deliberately apply different treatments to different individuals in the study, according to a quantifiable randomized method.

- **an observational study**, in which the researchers do not themselves apply different treatments to different individuals; rather they simply observe the individuals.

Note the term *quantifiable* in the first type: there are many ways to randomize quantifiably, but if assignment of treatments depends on anything involving human choice, quantifiability is lost (even if there is randomization).
EXAMPLE
Researchers conduct an opinion poll by selecting people at random and then asking each person the same questions. What type of study is this?
This is an observational study. Although the people were selected at random, they did not treat the interviewees differently.
**EXAMPLE**

In order to determine the efficacy of a certain medicine, researchers gather together a set of 24 volunteers. They divide the volunteers into 2 groups at random (by flipping a coin). The first group takes the medicine during the study; the second group takes a placebo (substitute for the medicine which should have no effect).

This is an example of a randomized experiment. We don’t know exactly how the volunteers were chosen, but they were subjected to two different treatments according to a quantifiably randomized method.
Not every study can be classified as one of these two types, but these two types are very common and are worth being aware of.

For us, the big difference between these two is:

- Evidence from randomized experiments carries a lot of weight as part of an argument to establish causation.
- Evidence from observational studies carries little to no weight as part of an argument to establish causation.
The reasoning behind this is that if treatments are applied randomly, then although we aren’t assured that all the other factors that might play a role average out the same in all the groups being compared, we can compute the probability that it won’t.

This is why the quantifiable part is so important.

If human choice is involved, that may or may not help to average things out, but the problem is that we can no longer quantify the probability that it won’t.
We have looked at the role of randomized *treatments*: they help determine how much causative weight the study’s results carry.

There is another place where randomization can enter a study: in the selection of the participants to be studied.
In this class, we will view the random processes of our random variables as selecting a sample (of a specified size) from a population.

The population might be real or theoretical.
A key aspect of this view is:

Statistical inference extends to the population from which a sample was drawn in a quantifiably random way, and no further.

The inferences may or may not hold for a larger population, but it is invalid to conclude from the statistics that they do.
We often draw inferences for broader populations because of our knowledge of the subject area to which the data relate (such as: biology, economics, etc.).

But it is important to recognize that those sorts of inferences are not statistical inferences.

Separate statistical and non-statistical inferences clearly in your mind, and when you report your results.
An example from *The Statistical Sleuth (3rd edition)*:

**EXAMPLE**

Researchers measured the lead content in teeth and the IQ test scores for all 3,229 children attending first and second grades in the period between 1975 and 1975 in Chelsea and Somerville, MA. The researchers found statistically significant evidence that the average IQ score among individuals with low lead concentration is higher than the average IQ score among individuals with high lead concentration. To what population does this statistical inference apply?
These students were not sampled from a larger population of students, so this statistical inference does not apply to children not in the study.

We might suspect that these children are no different from other children across the US, and so conclude that this result should apply to all children in the US, but that inference is non-statistical and is certainly open to question.
This begs the question: why are we making inferences?

Put otherwise, where is the variability? What would be different if we conducted this study again?
The variability here is in the measurement:

- If we were to measure these same students’ IQs again, they would not be exactly the same.
- If we were to measure the lead concentrations again, they would not be exactly the same.

The theoretical population from which this sample is drawn is: the set of all IQ measurements and lead concentration measurements of these exact children (at the time that this study was conducted).
It may seem strange to have such a theoretical population, but it is important to consider where the variability is.

Otherwise, you might not even know why you’re attempting to draw statistical inferences!
EXAMPLE

Suppose you wanted to estimate the average height among all students at the University of Puget Sound. You select 30 students at random (meaning truly, quantifiably randomly) from the student body, measure their heights, and use that data to compute a confidence interval for the average height. To what population does this confidence interval apply?
The confidence interval that you computed applies to the population of all currently enrolled students at the University of Puget Sound.

You might suspect that the confidence interval would apply to currently enrolled students in different years too, but that would be non-statistical and open to question.

For example, what if the university decided to recruit way more aggressively for its basketball team?

Or what if you look back 100 years? Average heights have changed since then!

Such considerations illustrate why it is important to be clear about statistical versus non-statistical inferences.
Another example from *The Statistcal Sleuth (3rd edition)*:

**EXAMPLE**

Subjects with considerable experience in creative writing were randomly assigned to one of two treatment groups: the *intrinsic* and *extrinsic*, in which they were given two different questionnaires. Subjects were then asked to write a poem about laughter. Researchers found statistically significant evidence that the *intrinsic* group scored higher on a ranking of their poem’s creativity.

1. For the population represented, does this provide weighty evidence that the *intrinsic* group treatment caused the higher creativity in the poems?
2. What population do these statistical inferences validly extend to?
Since the treatments were randomized, these statistical inferences do provide weighty evidence that the *intrinsic* group treatment caused the higher creativity in the poems.

However, since the subjects were not selected at random (if they had been, it should have been included in the description of the study), these statistical inferences extend only to the participants in the study.

One might be tempted to extend the inferences to other people on the (invalid, rather flimsy) argument that these people are somehow “similar enough to” or “a representative sample of” everyone in the world.

Doing so would be non-statistical and open to question, to say the least.
Why would it be important that the people judging the creativity not be aware of which poems came from which groups? (Another way to say this is to assert that the judges were blinded to the treatments.)
If they were aware who write which, that might affect how they judged the poem *in unquantifiable ways*.

It’s not always important to eliminate all biases, but it is important to eliminate (or at least) minimize those that are unquantifiable.