Making Inferences from Samples

Christopher Adolph

Department of Political Science

and

Center for Statistics and the Social Sciences

University of Washington, Seattle
Motivation

How do we know what the average American thinks about an issue?

Usual approach: conduct an opinion poll, randomly sample 1000 or so people, and present the average of their opinions.

But how do we know this matches the average opinion of all Americans?
Motivation

In particular, how do we know how far the sample mean, $\bar{x}$, is from the population mean, $\bar{x}_{\text{population}}$?
Motivation

In particular, how do we know how far the sample mean, \( \bar{x} \), is from the population mean, \( \bar{x}_{\text{population}} \)?

\[
\bar{x} - \bar{x}_{\text{population}} = ?
\]

If our sample isn’t very representative of the population, these might be far apart.
Motivation

In particular, how do we know how far the sample mean, $\bar{x}$, is from the population mean, $\bar{x}_{\text{population}}$?

$$\bar{x} - \bar{x}_{\text{population}} = ?$$

If our sample isn’t very representative of the population, these might be far apart.

Without knowing anything but the sample, can we estimate the deviation between the sample mean and the population mean?
We will consider groups of observations at two distinct levels:

**Population**  All the potential units of analysis in our chosen research design

Ideally we’d like to analyze a census, or complete set, of these observations

Example: Average support $\overline{x}^{\text{population}}$ of all Washingtonians for same-sex marriage
Populations & Samples

We will consider groups of observations at two distinct levels:

**Population**  All the potential units of analysis in our chosen research design

Ideally we’d like to analyze a census, or complete set, of these observations

Example: Average support $\bar{x}^{\text{population}}$ of all Washingtonians for same-sex marriage

**Sample**  The units of analysis actually collected for our study

Usually a subset of the population

Example: Average support $\bar{x}$ of 500 randomly selected Washingtonians for same-sex marriage
Sampling Frames

In an ideal situation, our sample and population will contain the same cases (a census)

Usually, we must instead make inferences about the population using a subset, or sample, of cases

Can select this sample in different ways
Sampling Frames

**Random sample**  Make a list of the full population and randomly select by identification number.

E.g., Random Digit Dialling of phone numbers.

If done correctly, makes inference "easy"

Stratified sample  If we can't randomly sample properly, but have detailed information on the population, we could re-weight our flawed random sample based on identifiable strata

E.g., If a phone survey fails to reach enough people who work at night, we could give the few we reach extra weight based on their known population frequency

If done correctly, produces something close to a random sample
Sampling Frames

**Random sample** Make a list of the full population and randomly select by identification number.

E.g., Random Digit Dialling of phone numbers.
Sampling Frames

**Random sample** Make a list of the full population and randomly select by identification number.

E.g., Random Digit Dialling of phone numbers.

If done correctly, makes inference “easy”
Sampling Frames

**Random sample**  Make a list of the full population and randomly select by identification number.

E.g., Random Digit Dialling of phone numbers.

If done correctly, makes inference “easy”

**Stratified sample**  If we can’t randomly sample properly, but have detailed information on the population, we could re-weight our flawed random sample based on identifiable strata
Sampling Frames

**Random sample**  Make a list of the full population and randomly select by identification number.

E.g., Random Digit Dialling of phone numbers.

If done correctly, makes inference “easy”

**Stratified sample**  If we can’t randomly sample properly, but have detailed information on the population, we could re-weight our flawed random sample based on identifiable strata

E.g., If a phone survey fails to reach enough people who work at night, we could give the few we reach extra weight based on their known population frequency
Sampling Frames

**Random sample** Make a list of the full population and randomly select by identification number.

E.g., Random Digit Dialling of phone numbers.

If done correctly, makes inference “easy”

**Stratified sample** If we can’t randomly sample properly, but have detailed information on the population, we could re-weight our flawed random sample based on identifiable strata

E.g., If a phone survey fails to reach enough people who work at night, we could give the few we reach extra weight based on their known population frequency

If done correctly, produces something close to a random sample
**Convenience sample** If we can’t form any sort of random sample, we might take people non-randomly who are close at hand.
Convenience sample If we can’t form any sort of random sample, we might take people non-randomly who are close at hand

E.g., When studying a hard to reach population, we might ask each member we find to nominate other members, forming a snowball sample
Sampling Frames

**Convenience sample**  If we can’t form any sort of random sample, we might take people non-randomly who are close at hand

E.g., When studying a hard to reach population, we might ask each member we find to nominate other members, forming a snowball sample

Convenience samples do *not* allow scientific inference to the population parameters
When sampling goes wrong

If a random sample is non-representative, will adding more random samples help make it so?
When sampling goes wrong

If a random sample is non-representative, will adding more random samples help make it so? Yes
When sampling goes wrong

If a random sample is non-representative, will adding more random samples help make it so? Yes

If a stratified sample has the wrong weights, will adding more samples make it representative? No
When sampling goes wrong

If a random sample is non-representative, will adding more random samples help make it so? Yes

If a stratified sample has the wrong weights, will adding more samples make it representative? No
When sampling goes wrong

If a random sample is non-representative, will adding more random samples help make it so? Yes

If a stratified sample has the wrong weights, will adding more samples make it representative? No

Are convenience samples more likely to be representative as they get larger?
When sampling goes wrong

If a random sample is non-representative, will adding more random samples help make it so? Yes

If a stratified sample has the wrong weights, will adding more samples make it representative? No

Are convenience samples more likely to be representative as they get larger? NO! No matter how large a convenience sample, they are likely to be sampled with huge and unknown selection bias
Sampling Inference

Our goal is to make scientifically valid inferences from the random or representative sample we’ve collected.

Standard scientific practice requires that we quantify the uncertainty introduced by sampling.

To learn how to do this, we will eventually learn new concepts:

the standard error and $t$-statistic,

and new continuous probability distributions:

the $\chi^2$ distribution and $t$ distribution

Today, we focus on the standard error.
119 students submitted their heights in inches, with a mean of 66.6 inches, and a standard deviation of 4.1 inches.

The class distribution of heights is shown
About 2/3s of the class submitted heights.

Getting that kind of response from a population is expensive and impractical in most cases.

Suppose we just wanted to know the average height in the class?
How well could we estimate the class mean from a sample of a specific size?

And how can we tell if our sample estimate of the mean is reliable?

We need a way to predict how much our sample mean will diverge from the population mean.
**Error** The difference between an estimate (based on a sample) and the true value of a quantity (in the population)
Useful concepts for estimating uncertainty (statistical inference)

**Error**  The difference between an estimate (based on a sample) and the true value of a quantity (in the population)

**Root mean squared error (RMSE)**  The average amount of error observed across repeated samples from the population.

(To avoid cancellation of equal and opposite errors, we “average” error by squaring first, then taking the mean, then the square root)
Useful concepts for estimating uncertainty (statistical inference)

**Error**  The difference between an estimate (based on a sample) and the true value of a quantity (in the population)

**Root mean squared error (RMSE)**  The average amount of error observed across repeated samples from the population.

(To avoid cancellation of equal and opposite errors, we “average” error by squaring first, then taking the mean, then the square root)

**Standard error**  An *estimate* of the error in our sample’s estimate of the population quantity.

Thus the standard error is the best guess from a single sample of what the RMSE would turn out to be if we could afford to take many samples
Concepts for statistical inference applied to the *sample mean*

**Error** The difference between the sample mean (estimate) and population mean ("truth")
Concepts for statistical inference applied to the *sample mean*

**Error** The difference between the sample mean (estimate) and population mean ("truth")

**Root mean squared error (RMSE)** The average amount of error *observed* between the sample means and the population mean.

(To avoid cancellation of equal and opposite errors, we “average” error by squaring first, then taking the mean, then the square root)
Concepts for statistical inference applied to the *sample mean*

**Error**  The difference between the sample mean (estimate) and population mean (“truth”)

**Root mean squared error (RMSE)**  The average amount of error *observed* between the sample means and the population mean.

(To avoid cancellation of equal and opposite errors, we “average” error by squaring first, then taking the mean, then the square root)

**Standard error**  An *estimate* of the error in our sample’s estimate of the population mean.

Your book also calls this the *standard deviation of the sampling mean* (Chapter 11)
Above are the submitted height data and their mean.

We will treat the rows of this classroom as samples.

While you collect your data on your row...
I will explore a set of 14 pre-selected samples of 8 heights each

Note that the means and variances of the samples can differ from the full population...
I will explore a set of 14 pre-selected samples of 8 heights each.

Note that the means and variances of the samples can differ from the full population... 

...but can also resemble it fairly closely.
Sometimes an individual sample can be so far off that it would mislead us considerably.

This is more likely the smaller the sample.
Sometimes an individual sample can be so far off that it would mislead us considerably.

This is more likely the smaller the sample.

But even with samples as tiny as 8 students, most of the time the sample mean is fairly close to the population mean.
Moreover, the sample mean seems just as likely to be below the population mean . . .
Moreover, the sample mean seems just as likely to be below the population mean . . .

. . . as above it.
Moreover, the sample mean seems just as likely to be below the population mean . . .

. . . as above it.

When an estimate is neither systematically too high or too low, it is unbiased.
Unbiased estimators can still sometimes be far from the true value
Unbiased estimators can still sometimes be far from the true value

But on average, over many samples, unbiased estimators will equal the truth, rather than show systematic *bias* up or down
We also want our estimates to be *close* to the truth most of the time, a separate issue from bias.
We also want our estimates to be *close* to the truth most of the time, a separate issue from bias.

Sample estimates which are usually close to the population value are said to have low *error* or to be *efficient*. 

**sample 11 mean**
Notice that although the sample mean bounces back and forth,
Notice that although the sample mean bounces back and forth, it tends to stay close to the population mean.
Notice that although the sample mean bounces back and forth, it tends to stay close to the population mean and doesn’t range as far as the data itself.
That is, the standard deviation of the sample means is smaller than the standard deviation of the data itself:

$$sd(\bar{x}) < sd(x)$$
We now overlay the distribution of sampling means in red.

Note that the distribution of sampling means looks quite Normal, even though the distribution of heights is only approximately Normal.
This is the Central Limit Theorem again:

As we take more and more samples, the distribution of the sampling means of *any* random variable approaches the Normal (with few exceptions)
To estimate the mean of population well, we want to make the distribution of sampling means (in red) as narrow as possible: ideally, a spike right over the true population mean.

How can we do this?
The Law of Large Numbers

*When sampling from a population, our estimates of features of that population get better the more data we sample*

What do we mean by better estimates?
The Law of Large Numbers

*When sampling from a population, our estimates of features of that population get better the more data we sample*

What do we mean by better estimates?

An estimate with smaller *standard error* (expected deviation from the truth)
The Law of Large Numbers

When sampling from a population, our estimates of features of that population get better the more data we sample

What do we mean by better estimates?

An estimate with smaller *standard error* (expected deviation from the truth)

Formula for the standard error of the mean:

$$se(\bar{x}) = \frac{sd(x)}{\sqrt{n}}$$

The Law of Large Numbers applies to estimating the mean of a population:

Our estimate of the mean, $\bar{x}$ gets closer to the truth, and its standard error, $se(\bar{x})$ gets smaller as the sampled $n$ increases
The Square Root Law

Formula for the standard error of the mean:

\[ se(\bar{x}) = \frac{sd(x)}{\sqrt{n}} \]

Remember that the smaller \( se(\bar{x}) \) is, the better our estimate

Making \( n \) bigger—adding more observations—will indeed shrink \( se(\bar{x}) \), but there are diminishing returns

Because \( se(\bar{x}) \) depends on \( \sqrt{n} \),
to halve the amount of error we must quadruple the amount of data

If our se is 1 inch of height with 100 observations,
to reduce our expected error to 0.5 inches, we need 400 total observations

Note a surprise: the size of the population *does not appear* in this formula, and does not affect the precision of our estimates!
The population mean, 66.6 inches, is the average height in the class. Below it are the means of 14 samples of 8 students I drew to simulate “rows” of the classroom.
<table>
<thead>
<tr>
<th></th>
<th>Mean (inches)</th>
<th>Sampling Error</th>
<th>Std Dev</th>
<th>Std Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>66.6</td>
<td>—</td>
<td>4.1</td>
<td>—</td>
</tr>
<tr>
<td>Sample 1</td>
<td>68.4</td>
<td>1.9</td>
<td>2.2</td>
<td>0.8</td>
</tr>
<tr>
<td>Sample 2</td>
<td>66.7</td>
<td>0.2</td>
<td>5.5</td>
<td>2.0</td>
</tr>
<tr>
<td>Sample 3</td>
<td>63.6</td>
<td>−3.0</td>
<td>2.1</td>
<td>0.8</td>
</tr>
<tr>
<td>Sample 4</td>
<td>65.8</td>
<td>−0.7</td>
<td>4.4</td>
<td>1.6</td>
</tr>
<tr>
<td>Sample 5</td>
<td>64.0</td>
<td>−2.6</td>
<td>3.5</td>
<td>1.2</td>
</tr>
<tr>
<td>Sample 6</td>
<td>67.7</td>
<td>1.2</td>
<td>5.2</td>
<td>1.9</td>
</tr>
<tr>
<td>Sample 7</td>
<td>67.3</td>
<td>0.7</td>
<td>2.5</td>
<td>0.9</td>
</tr>
<tr>
<td>Sample 8</td>
<td>70.3</td>
<td>3.7</td>
<td>5.4</td>
<td>1.9</td>
</tr>
<tr>
<td>Sample 9</td>
<td>65.6</td>
<td>−0.9</td>
<td>4.8</td>
<td>1.7</td>
</tr>
<tr>
<td>Sample 10</td>
<td>66.4</td>
<td>−0.2</td>
<td>3.9</td>
<td>1.4</td>
</tr>
<tr>
<td>Sample 11</td>
<td>67.1</td>
<td>0.5</td>
<td>3.9</td>
<td>1.4</td>
</tr>
<tr>
<td>Sample 12</td>
<td>64.9</td>
<td>−1.7</td>
<td>3.8</td>
<td>1.3</td>
</tr>
<tr>
<td>Sample 13</td>
<td>66.8</td>
<td>0.2</td>
<td>2.9</td>
<td>1.0</td>
</tr>
<tr>
<td>Sample 14</td>
<td>68.5</td>
<td>1.9</td>
<td>1.7</td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Sample Mean</th>
<th>Pop Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>66.6</td>
<td>66.6</td>
</tr>
<tr>
<td></td>
<td>Avg SE</td>
<td>RMSE</td>
</tr>
<tr>
<td></td>
<td>1.4</td>
<td>1.8</td>
</tr>
</tbody>
</table>

The population mean, 66.6 inches, is the average height in the class. The average sample mean matches the population mean almost exactly.
<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std Dev</th>
<th>Std Error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population</strong></td>
<td>66.6</td>
<td>4.1</td>
<td>—</td>
</tr>
<tr>
<td><strong>Sample 1</strong></td>
<td>68.4</td>
<td>2.2</td>
<td>0.8</td>
</tr>
<tr>
<td><strong>Sample 2</strong></td>
<td>66.7</td>
<td>5.5</td>
<td>2.0</td>
</tr>
<tr>
<td><strong>Sample 3</strong></td>
<td>63.6</td>
<td>2.1</td>
<td>0.8</td>
</tr>
<tr>
<td><strong>Sample 4</strong></td>
<td>65.8</td>
<td>4.4</td>
<td>1.6</td>
</tr>
<tr>
<td><strong>Sample 5</strong></td>
<td>64.0</td>
<td>3.5</td>
<td>1.2</td>
</tr>
<tr>
<td><strong>Sample 6</strong></td>
<td>67.7</td>
<td>5.2</td>
<td>1.9</td>
</tr>
<tr>
<td><strong>Sample 7</strong></td>
<td>67.3</td>
<td>2.5</td>
<td>0.9</td>
</tr>
<tr>
<td><strong>Sample 8</strong></td>
<td>70.3</td>
<td>5.4</td>
<td>1.9</td>
</tr>
<tr>
<td><strong>Sample 9</strong></td>
<td>65.6</td>
<td>4.8</td>
<td>1.7</td>
</tr>
<tr>
<td><strong>Sample 10</strong></td>
<td>66.4</td>
<td>3.9</td>
<td>1.4</td>
</tr>
<tr>
<td><strong>Sample 11</strong></td>
<td>67.1</td>
<td>3.9</td>
<td>1.4</td>
</tr>
<tr>
<td><strong>Sample 12</strong></td>
<td>64.9</td>
<td>3.8</td>
<td>1.3</td>
</tr>
<tr>
<td><strong>Sample 13</strong></td>
<td>66.8</td>
<td>2.9</td>
<td>1.0</td>
</tr>
<tr>
<td><strong>Sample 14</strong></td>
<td>68.5</td>
<td>1.7</td>
<td>0.6</td>
</tr>
</tbody>
</table>

| Sample Mean    | 66.6 | Avg SE | 1.4     |
| Pop Mean       | 66.6 | RMSE   | 1.8     |

Sampling Error is how much this row of students differs from the class mean:

How wrong (&in what direction) you’d be if you used your row to estimate the class average height.
<table>
<thead>
<tr>
<th>Population</th>
<th>Mean</th>
<th>Sampling Error</th>
<th>Std Dev</th>
<th>Std Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>68.4</td>
<td>1.9</td>
<td>2.2</td>
<td>0.8</td>
</tr>
<tr>
<td>Sample 2</td>
<td>66.7</td>
<td>0.2</td>
<td>5.5</td>
<td>2.0</td>
</tr>
<tr>
<td>Sample 3</td>
<td>63.6</td>
<td>−3.0</td>
<td>2.1</td>
<td>0.8</td>
</tr>
<tr>
<td>Sample 4</td>
<td>65.8</td>
<td>−0.7</td>
<td>4.4</td>
<td>1.6</td>
</tr>
<tr>
<td>Sample 5</td>
<td>64.0</td>
<td>−2.6</td>
<td>3.5</td>
<td>1.2</td>
</tr>
<tr>
<td>Sample 6</td>
<td>67.7</td>
<td>1.2</td>
<td>5.2</td>
<td>1.9</td>
</tr>
<tr>
<td>Sample 7</td>
<td>67.3</td>
<td>0.7</td>
<td>2.5</td>
<td>0.9</td>
</tr>
<tr>
<td>Sample 8</td>
<td>70.3</td>
<td>3.7</td>
<td>5.4</td>
<td>1.9</td>
</tr>
<tr>
<td>Sample 9</td>
<td>65.6</td>
<td>−0.9</td>
<td>4.8</td>
<td>1.7</td>
</tr>
<tr>
<td>Sample 10</td>
<td>66.4</td>
<td>−0.2</td>
<td>3.9</td>
<td>1.4</td>
</tr>
<tr>
<td>Sample 11</td>
<td>67.1</td>
<td>0.5</td>
<td>3.9</td>
<td>1.4</td>
</tr>
<tr>
<td>Sample 12</td>
<td>64.9</td>
<td>−1.7</td>
<td>3.8</td>
<td>1.3</td>
</tr>
<tr>
<td>Sample 13</td>
<td>66.8</td>
<td>0.2</td>
<td>2.9</td>
<td>1.0</td>
</tr>
<tr>
<td>Sample 14</td>
<td>68.5</td>
<td>1.9</td>
<td>1.7</td>
<td>0.6</td>
</tr>
<tr>
<td>Sample Mean</td>
<td>66.6</td>
<td></td>
<td>Avg SE</td>
<td>1.4</td>
</tr>
<tr>
<td>Pop Mean</td>
<td>66.6</td>
<td></td>
<td>RMSE</td>
<td>1.8</td>
</tr>
</tbody>
</table>

RMSE is “root mean squared error”

RMSE is the average error we would make if we predicted the class from each row sample in turn
<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Sampling Error</th>
<th>Std Dev</th>
<th>Std Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>66.6</td>
<td>—</td>
<td>4.1</td>
<td>—</td>
</tr>
<tr>
<td>Sample 1</td>
<td>68.4</td>
<td>1.9</td>
<td>2.2</td>
<td>0.8</td>
</tr>
<tr>
<td>Sample 2</td>
<td>66.7</td>
<td>0.2</td>
<td>5.5</td>
<td>2.0</td>
</tr>
<tr>
<td>Sample 3</td>
<td>63.6</td>
<td>−3.0</td>
<td>2.1</td>
<td>0.8</td>
</tr>
<tr>
<td>Sample 4</td>
<td>65.8</td>
<td>−0.7</td>
<td>4.4</td>
<td>1.6</td>
</tr>
<tr>
<td>Sample 5</td>
<td>64.0</td>
<td>−2.6</td>
<td>3.5</td>
<td>1.2</td>
</tr>
<tr>
<td>Sample 6</td>
<td>67.7</td>
<td>1.2</td>
<td>5.2</td>
<td>1.9</td>
</tr>
<tr>
<td>Sample 7</td>
<td>67.3</td>
<td>0.7</td>
<td>2.5</td>
<td>0.9</td>
</tr>
<tr>
<td>Sample 8</td>
<td>70.3</td>
<td>3.7</td>
<td>5.4</td>
<td>1.9</td>
</tr>
<tr>
<td>Sample 9</td>
<td>65.6</td>
<td>−0.9</td>
<td>4.8</td>
<td>1.7</td>
</tr>
<tr>
<td>Sample 10</td>
<td>66.4</td>
<td>−0.2</td>
<td>3.9</td>
<td>1.4</td>
</tr>
<tr>
<td>Sample 11</td>
<td>67.1</td>
<td>0.5</td>
<td>3.9</td>
<td>1.4</td>
</tr>
<tr>
<td>Sample 12</td>
<td>64.9</td>
<td>−1.7</td>
<td>3.8</td>
<td>1.3</td>
</tr>
<tr>
<td>Sample 13</td>
<td>66.8</td>
<td>0.2</td>
<td>2.9</td>
<td>1.0</td>
</tr>
<tr>
<td>Sample 14</td>
<td>68.5</td>
<td>1.9</td>
<td>1.7</td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Sample Mean</th>
<th>Pop Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>66.6</td>
<td>66.6</td>
</tr>
</tbody>
</table>

Because errors can be + or −, we square error before averaging, then take the square root of the sum of squared errors to get RMSE. We need lots of samples to calculate RMSE.
<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Sampling Error</th>
<th>Std Dev</th>
<th>Std Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>66.6</td>
<td>—</td>
<td>4.1</td>
<td>—</td>
</tr>
<tr>
<td>Sample 1</td>
<td>68.4</td>
<td>1.9</td>
<td>2.2</td>
<td>0.8</td>
</tr>
<tr>
<td>Sample 2</td>
<td>66.7</td>
<td>0.2</td>
<td>5.5</td>
<td>2.0</td>
</tr>
<tr>
<td>Sample 3</td>
<td>63.6</td>
<td>-3.0</td>
<td>2.1</td>
<td>0.8</td>
</tr>
<tr>
<td>Sample 4</td>
<td>65.8</td>
<td>-0.7</td>
<td>4.4</td>
<td>1.6</td>
</tr>
<tr>
<td>Sample 5</td>
<td>64.0</td>
<td>-2.6</td>
<td>3.5</td>
<td>1.2</td>
</tr>
<tr>
<td>Sample 6</td>
<td>67.7</td>
<td>1.2</td>
<td>5.2</td>
<td>1.9</td>
</tr>
<tr>
<td>Sample 7</td>
<td>67.3</td>
<td>0.7</td>
<td>2.5</td>
<td>0.9</td>
</tr>
<tr>
<td>Sample 8</td>
<td>70.3</td>
<td>3.7</td>
<td>5.4</td>
<td>1.9</td>
</tr>
<tr>
<td>Sample 9</td>
<td>65.6</td>
<td>-0.9</td>
<td>4.8</td>
<td>1.7</td>
</tr>
<tr>
<td>Sample 10</td>
<td>66.4</td>
<td>-0.2</td>
<td>3.9</td>
<td>1.4</td>
</tr>
<tr>
<td>Sample 11</td>
<td>67.1</td>
<td>0.5</td>
<td>3.9</td>
<td>1.4</td>
</tr>
<tr>
<td>Sample 12</td>
<td>64.9</td>
<td>-1.7</td>
<td>3.8</td>
<td>1.3</td>
</tr>
<tr>
<td>Sample 13</td>
<td>66.8</td>
<td>0.2</td>
<td>2.9</td>
<td>1.0</td>
</tr>
<tr>
<td>Sample 14</td>
<td>68.5</td>
<td>1.9</td>
<td>1.7</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Sample Mean   66.6  Avg SE   1.4
Pop Mean      66.6  RMSE    1.8

The standard error is how much we expect this sample to miss by

$$sd(\text{height}_{\text{sample}}) \sqrt{\frac{1}{8}}$$

An estimate of RMSE we construct just from one sample
<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Sampling Error</th>
<th>Std Dev</th>
<th>Std Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>66.6</td>
<td>—</td>
<td>4.1</td>
<td>—</td>
</tr>
<tr>
<td>Sample 1</td>
<td>68.4</td>
<td>1.9</td>
<td>2.2</td>
<td>0.8</td>
</tr>
<tr>
<td>Sample 2</td>
<td>66.7</td>
<td>0.2</td>
<td>5.5</td>
<td>2.0</td>
</tr>
<tr>
<td>Sample 3</td>
<td>63.6</td>
<td>—3.0</td>
<td>2.1</td>
<td>0.8</td>
</tr>
<tr>
<td>Sample 4</td>
<td>65.8</td>
<td>—0.7</td>
<td>4.4</td>
<td>1.6</td>
</tr>
<tr>
<td>Sample 5</td>
<td>64.0</td>
<td>—2.6</td>
<td>3.5</td>
<td>1.2</td>
</tr>
<tr>
<td>Sample 6</td>
<td>67.7</td>
<td>1.2</td>
<td>5.2</td>
<td>1.9</td>
</tr>
<tr>
<td>Sample 7</td>
<td>67.3</td>
<td>0.7</td>
<td>2.5</td>
<td>0.9</td>
</tr>
<tr>
<td>Sample 8</td>
<td>70.3</td>
<td>3.7</td>
<td>5.4</td>
<td>1.9</td>
</tr>
<tr>
<td>Sample 9</td>
<td>65.6</td>
<td>—0.9</td>
<td>4.8</td>
<td>1.7</td>
</tr>
<tr>
<td>Sample 10</td>
<td>66.4</td>
<td>—0.2</td>
<td>3.9</td>
<td>1.4</td>
</tr>
<tr>
<td>Sample 11</td>
<td>67.1</td>
<td>0.5</td>
<td>3.9</td>
<td>1.4</td>
</tr>
<tr>
<td>Sample 12</td>
<td>64.9</td>
<td>—1.7</td>
<td>3.8</td>
<td>1.3</td>
</tr>
<tr>
<td>Sample 13</td>
<td>66.8</td>
<td>0.2</td>
<td>2.9</td>
<td>1.0</td>
</tr>
<tr>
<td>Sample 14</td>
<td>68.5</td>
<td>1.9</td>
<td>1.7</td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Sampling Error</th>
<th>Std Dev</th>
<th>Std Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Mean</td>
<td>66.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pop Mean</td>
<td>66.6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The average standard error is how much we expect to miss the population mean in repeated samples. Avg SE should match RMSE pretty closely.
<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Sampling Error</th>
<th>Std Dev</th>
<th>Std Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>66.6</td>
<td>—</td>
<td>4.1</td>
<td>—</td>
</tr>
<tr>
<td>Sample 1</td>
<td>68.4</td>
<td>1.9</td>
<td>2.2</td>
<td>0.8</td>
</tr>
<tr>
<td>Sample 2</td>
<td>66.7</td>
<td>0.2</td>
<td>5.5</td>
<td>2.0</td>
</tr>
<tr>
<td>Sample 3</td>
<td>63.6</td>
<td>−3.0</td>
<td>2.1</td>
<td>0.8</td>
</tr>
<tr>
<td>Sample 4</td>
<td>65.8</td>
<td>−0.7</td>
<td>4.4</td>
<td>1.6</td>
</tr>
<tr>
<td>Sample 5</td>
<td>64.0</td>
<td>−2.6</td>
<td>3.5</td>
<td>1.2</td>
</tr>
<tr>
<td>Sample 6</td>
<td>67.7</td>
<td>1.2</td>
<td>5.2</td>
<td>1.9</td>
</tr>
<tr>
<td>Sample 7</td>
<td>67.3</td>
<td>0.7</td>
<td>2.5</td>
<td>0.9</td>
</tr>
<tr>
<td>Sample 8</td>
<td>70.3</td>
<td>3.7</td>
<td>5.4</td>
<td>1.9</td>
</tr>
<tr>
<td>Sample 9</td>
<td>65.6</td>
<td>−0.9</td>
<td>4.8</td>
<td>1.7</td>
</tr>
<tr>
<td>Sample 10</td>
<td>66.4</td>
<td>−0.2</td>
<td>3.9</td>
<td>1.4</td>
</tr>
<tr>
<td>Sample 11</td>
<td>67.1</td>
<td>0.5</td>
<td>3.9</td>
<td>1.4</td>
</tr>
<tr>
<td>Sample 12</td>
<td>64.9</td>
<td>−1.7</td>
<td>3.8</td>
<td>1.3</td>
</tr>
<tr>
<td>Sample 13</td>
<td>66.8</td>
<td>0.2</td>
<td>2.9</td>
<td>1.0</td>
</tr>
<tr>
<td>Sample 14</td>
<td>68.5</td>
<td>1.9</td>
<td>1.7</td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Avg SE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Mean</td>
<td>66.6</td>
<td>1.4</td>
<td>1.8</td>
</tr>
<tr>
<td>Pop Mean</td>
<td>66.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hypothesis testing

A framework for using a \textit{sample} to test whether the mean of a population is on one side of a \textit{reference point}

Invented by Jerzy Neyman & Egon Pearson, using concepts by R A Fisher
Hypothesis testing

A framework for using a \textit{sample} to test whether the mean of a population is on one side of a \textit{reference point}

Invented by Jerzy Neyman & Egon Pearson, using concepts by R A Fisher

This framework tends to mislead if not precisely understood, but is still widely used in teaching statistics and in older publications. Be warned!
Hypothesis testing

A framework for using a sample to test whether the mean of a population is on one side of a reference point

Invented by Jerzy Neyman & Egon Pearson, using concepts by R A Fisher

This framework tends to mislead if not precisely understood, but is still widely used in teaching statistics and in older publications. Be warned!

Hypothesis testing terms:

**Null hypothesis** The reference point we *arbitrarily* chose. Denote as $\mu_0$. 
Hypothesis testing

A framework for using a *sample* to test whether the mean of a population is on one side of a *reference point*

Invented by Jerzy Neyman & Egon Pearson, using concepts by R A Fisher

This framework tends to mislead if not precisely understood, but is still widely used in teaching statistics and in older publications. Be warned!

Hypothesis testing terms:

**Null hypothesis**  The reference point *we arbitrarily chose*. Denote as $\mu_0$.

**Alternative hypothesis**  The side of the reference point which *we think* contains the population mean
Hypothesis testing

Null hypothesis  The reference point *we arbitrarily chose*. Denote as $\mu_0$.

Alternative hypothesis  The side of the reference point which *we think* contains the population mean

Hypothesis testing attempts to reject the Null hypothesis in favor of the alternative hypothesis

Hypothesis testing does *not* directly test the alternative hypothesis, but attempts to *reject* a reference point far away from it
Hypothesis testing

**Null hypothesis**  The reference point *we arbitrarily chose*. Denote as $\mu_0$.

**Alternative hypothesis**  The side of the reference point which we *think* contains the population mean

For example, we might try to reject $\overline{x}_{\text{population}} = \mu_0$ in favor of $\overline{x}_{\text{population}} > \mu_0$

Hypothesis testing does *not* directly test the alternative hypothesis, but attempts to *reject* a reference point far away from it
Hypothesis testing

**Null hypothesis**  The reference point *we arbitrarily chose*. Denote as $\mu_0$.

**Alternative hypothesis**  The side of the reference point which we *think* contains the population mean.

Notice the italicized words. Our own theory is (somewhat) summarized by the alternative hypothesis,

But everything will hinge on an arbitrary reference point.
Hypothesis testing

Null hypothesis The reference point *we arbitrarily chose*. Denote as $\mu_0$.

Alternative hypothesis The side of the reference point which *we think* contains the population mean

Suppose we wanted to know if the average UW student works more than 10 hours a week.

We could randomly sample 1000 students;
Hypothesis testing

**Null hypothesis**  The reference point *we arbitrarily chose*. Denote as $\mu_0$.

** Alternative hypothesis**  The side of the reference point which we *think* contains the population mean

Suppose we wanted to know if the average UW student works more than 10 hours a week.

We could randomly sample 1000 students; ask how many hours they work per week, $H_i$;
Hypothesis testing

**Null hypothesis**  The reference point *we arbitrarily chose*. Denote as $\mu_0$.

**Alternative hypothesis**  The side of the reference point which we *think* contains the population mean

Suppose we wanted to know if the average UW student works more than 10 hours a week.

We could randomly sample 1000 students; ask how many hours they work per week, $H_i$; calculate $H$ for the sample;
Hypothesis testing

Null hypothesis The reference point we arbitrarily chose. Denote as $\mu_0$.

Alternative hypothesis The side of the reference point which we think contains the population mean.

Suppose we wanted to know if the average UW student works more than 10 hours a week.

We could randomly sample 1000 students; ask how many hours they work per week, $H_i$; calculate $\bar{H}$ for the sample; in order to learn about the average population spending $\bar{H}_{\text{population}}$. 
Hypothesis testing

**Null hypothesis**  The reference point *we arbitrarily chose*. Denote as $\mu_0$.

**Alternative hypothesis**  The side of the reference point which *we think* contains the population mean.

To see if $\bar{H}^{\text{population}} > 10$,
we could test against the null hypothesis that $\bar{H}^{\text{population}} = 10$ exactly.
Hypothesis testing

**Null hypothesis**  The reference point *we arbitrarily chose*. Denote as $\mu_0$.

**Alternative hypothesis**  The side of the reference point which *we think* contains the population mean.

To see if $H_{\text{population}} > 10$, we could test against the null hypothesis that $H_{\text{population}} = 10$ exactly.

Suppose we reject the null hypothesis of 10 hours or less. We can say that we have rejected that possibility, or that $H_{\text{population}}$ is *statistically significantly* greater than 10 hours.
Hypothesis testing

**Null hypothesis**  The reference point *we arbitrarily chose*. Denote as $\mu_0$.

**Alternative hypothesis**  The side of the reference point which *we think* contains the population mean

Could the true population mean hours still be 10.01?
Hypothesis testing

Null hypothesis  The reference point *we arbitrarily chose*. Denote as $\mu_0$.

Alternative hypothesis  The side of the reference point which *we think* contains the population mean

Could the true population mean hours still be 10.01?
Yes. Hypothesis tests are sharp and arbitrary.

The truth could be *any* value on this side of the null hypothesis.
Take care in selecting the null and interpreting the meaning of the test.
From $z$-scores to $t$-statistics

How do we perform a hypothesis test?

We need some way to standardize the distance between our sample mean $\bar{x}$ and the null hypothesis $\mu_0$
From $z$-scores to $t$-statistics

How do we perform a hypothesis test?

We need some way to standardize the distance between our sample mean $\bar{x}$ and the null hypothesis $\mu_0$.

With $z$-scores, we standardized using $z = (x - \mu) / \sigma$.
From $z$-scores to $t$-statistics

How do we perform a hypothesis test?

We need some way to standardize the distance between our sample mean $\bar{x}$ and the null hypothesis $\mu_0$.

With $z$-scores, we standardized using $z = (x - \mu) / \sigma$.

Here, we do something similar, using the standard error, to standardize the gap:

$$t = \frac{\text{sample statistic of interest} - \mu_0}{\text{se(sample statistic of interest)}}$$

The $t$ statistic of an estimate is:
the estimate itself, minus a hypothetical level, divided by the standard error of the estimate.
The $t$-statistic

In the case of the sample mean, the $t$-statistic is given by:

$$t = \frac{\bar{x} - \mu_0}{se(\bar{x})}$$
The *t*-statistic

In the case of the sample mean,

\[
t = \frac{\bar{x} - \mu_0}{\text{se}(\bar{x})} = \frac{\bar{x} - \mu_0}{\text{sd}(x)/\sqrt{n}}
\]

We are converting the gap between the data and the null into standard error units.
The $t$-statistic

In the case of the sample mean,

\[ t = \frac{\bar{x} - \mu_0}{se(\bar{x})} = \frac{\bar{x} - \mu_0}{sd(x)/\sqrt{n}} \]

We are converting the gap between the data and the null into standard error units.

We will often set our hypothetical comparison level $\mu_0 = 0$, so this frequently reduces to:

\[ t = \frac{\bar{x}}{sd(x)/\sqrt{n}} \]
The \( t \)-statistic

\[
t = \frac{\bar{x} - \mu_0}{\text{se}(\bar{x})}
\]

As with \( z \)-scores, our goal is to say how extreme or “unusual” the observed \( t \) is with reference to the distribution of \( t \).

\( t \) is a random variable with mean 0; the further away \( t \) is from its mean, the lower the chance of seeing a sample like our if the null hypothesis is true.

But \( t \) isn’t Normally distributed, so can’t use the method we used for \( z \)-scores (looking up the quantiles of the Normal distribution).

Instead, we need to look up quantiles of the \( t \)-distribution (Table C in Moore 6th ed.)
The $t$ distribution

Originally discovered by William Gosset, a statistician working at Guinness Brewery in the 1908 on the problem of measuring the quality of beer

If you sample random bottles of beer, how many bad samples do you need before you decide the production line is faulty?

Guinness was a pioneer of early statistical quality control, but forbade its statisticians from publishing (trade secrets!)

Gosset published his discovery under the pseudonym “Student”. Hence this is Student’s $t$-test
The $t$ distribution

$t$ distribution is a continuous distribution, with density from $-\infty$ to $+\infty$

Usually assume mean is 0

$t$ distribution has a “degrees of freedom” parameter

As the degrees of freedom does up, the $t$ distribution looks like the Normal
Example $t$ distributions

![Graph showing three $t$ distributions: $t_1$, $t_4$, and $t_{15}$.](image)

Figure 4: $t$ probability density functions.
The *t* distribution

Suppose we have a variable $t$ that is $t$-distributed with mean 0 and 5 degrees of freedom.

That is, $P(t) = t(5)$
The $t$ distribution

Suppose we have a variable $t$ that is $t$-distributed with mean 0 and 5 degrees of freedom

That is, $P(t) = t(5)$

What are the “critical” values of $t$ we would see just
The \( t \) distribution

Suppose we have a variable \( t \) that is \( t \)-distributed with mean 0 and 5 degrees of freedom.

That is, \( P(t) = t(5) \)

What are the “critical” values of \( t \) we would see just

- once in 10 draws?
- once in 20 draws?
- once in 100 draws?

Put still another way, which critical values will bound the 90% (or 95%, or 99%) most ordinary \( t \) draws?
The $t$ distribution

Suppose we have a variable $t$ that is $t$-distributed with mean 0 and 5 degrees of freedom.

That is, $P(t) = t(5)$

What are the “critical” values of $t$ we would see just

- once in 10 draws?
- once in 20 draws?
The \( t \) distribution

Suppose we have a variable \( t \) that is \( t \)-distributed with mean 0 and 5 degrees of freedom.

That is, \( P(t) = t(5) \)

What are the “critical” values of \( t \) we would see just

- once in 10 draws?
- once in 20 draws?
- once in 100 draws?

Put still another way, which critical values will bound the 90% (or 95%, or 99%) most ordinary \( t \) draws?
Areas under the $t$ distribution with 5 degrees of freedom

90% of the mass of $t(5)$ is between -2.015 and 2.015

Another 5 percent of the mass, for a total of 10 percent

Values outside the critical values are “unusual”.

We expect to see these values rarely, and may even suspect we have the wrong distribution if we see them often.
Areas under the $t$

Note that the above distribution is a $t$ with 5 degrees of freedom

Degrees of freedom roughly here reflect how many independent pieces of information helped create the $t$-ratio
Areas under the $t$

More information makes $t$ “better behaved”, so that extreme values occur less often,

More dfs thus make the tails thinner, and make critical values smaller
Areas under the $t$ distribution with 5 degrees of freedom

2.5 percent of the mass

Another 2.5 percent of the mass, for a total of 5 percent

95% of the mass of $t(5)$ is between -2.571 and 2.571

Going back to the $df = 5$ case, notice we can choose what constitutes unusual

Here, we’ve raise the bar: only the 5% most extreme values are unusual, so the critical values increase
Areas under the $t$ distribution with $\infty$ degrees of freedom

95% of the mass of $t(\infty)$ is between -1.96 and 1.96

Another 2.5 percent of the mass, for a total of 5 percent

The infinite degrees of freedom critical values for the 95% case

This is the most widely used standard for whether a $t$-ratio is unusual
Areas under the $t$ distribution with 5 degrees of freedom

99% of the mass of $t(5)$ is between -4.032 and 4.032

Another 0.5 percent of the mass, for a total of 1 percent

The most stringent commonly used standard is 99%

In this case, a draw from the $t$ must be in the 1% most extreme region to be considered unusual
**Areas under the $t$**

The infinite degrees of freedom case for 99%

Quick check: what do the critical values here mean?
Critical values of the $t$ distribution

We can state how unusual a $t$-ratio is under the assumption that it is distributed $t(n)$

<table>
<thead>
<tr>
<th>Test level</th>
<th>Interval</th>
<th>df = 5</th>
<th>df = $\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1 level</td>
<td>90%</td>
<td>2.015</td>
<td>1.645</td>
</tr>
<tr>
<td>0.05 level</td>
<td>95%</td>
<td>2.571</td>
<td>1.960</td>
</tr>
<tr>
<td>0.01 level</td>
<td>99%</td>
<td>4.032</td>
<td>2.576</td>
</tr>
</tbody>
</table>

These will be very useful for quantifying the uncertainty of estimates
The $t$-statistic

We can use the $t$-test to assess how likely it is that the truth deviates from a hypothetical value, given the sample estimate and standard error.
The \( t \)-statistic

We can use the t-test to assess how likely it is that the truth deviates from a hypothetical value, given the sample estimate and standard error.

That is, given \( \bar{x} - \mu_0 \) as large as the one we saw,
The $t$-statistic

We can use the $t$-test to assess how likely it is that the truth deviates from a hypothetical value, given the sample estimate and standard error.

That is, given $\bar{x} - \mu_0$ as large as the one we saw, and uncertainty of that estimate $sd(x)/\sqrt{n}$,
The $t$-statistic

We can use the t-test to assess how likely it is that the truth deviates from a hypothetical value, given the sample estimate and standard error

That is, given $\bar{x} - \mu_0$ as large as the one we saw, and uncertainty of that estimate $sd(x)/\sqrt{n}$, how likely is it that the population mean of $x$ is actual $\mu_0$ or smaller?
The $t$-statistic

We can use the t-test to assess how likely it is that the truth deviates from a hypothetical value, given the sample estimate and standard error.

That is, given $\bar{x} - \mu_0$ as large as the one we saw, and uncertainty of that estimate $sd(x)/\sqrt{n}$, how likely is it that the population mean of $x$ is actual $\mu_0$ or smaller?

Large $t$ could occur in one of two way:

- A unusual random sample far from the true population mean, which happens to be close to $\mu_0$
The \( t \)-statistic

We can use the t-test to assess how likely it is that the truth deviates from a hypothetical value, given the sample estimate and standard error.

That is, given \( \bar{x} - \mu_0 \) as large as the one we saw, and uncertainty of that estimate \( \text{sd}(x)/\sqrt{n} \), how likely is it that the population mean of \( x \) is actual \( \mu_0 \) or smaller?

Large \( t \) could occur in one of two way:

1. A unusual random sample far from the true population mean, which happens to be close to \( \mu_0 \)

2. A typical sample from a population mean that is larger than \( \mu_0 \)
We will never know which situation we are in, and we don’t know the probability of the latter case at all.
The \( t \)-statistic

We will never know which situation we are in, and we don’t know the probability of the latter case at all

But we can calculate the probability we would see a \( t \) as large as the one we saw by chance.
The \( t \)-statistic

We will never know which situation we are in, and we don’t know the probability of the latter case at all.

But we can calculate the probability we would see a \( t \) as large as the one we saw by chance.

This probability is known as the \( p \)-value.

To look it up in a table or stat package, we need to know the degrees of freedom.
The \( t \)-statistic

We will never know which situation we are in, and we don’t know the probability of the latter case at all

But we can calculate the probability we would see a \( t \) as large as the one we saw by chance.

This probability is known as the \( p \)-value

To look it up in a table or stat package, we need to know the degrees of freedom

Roughly, dfs are how much information we have, in this case, \( n - 1 \), since calculating \( \bar{x} \) uses up a degree of freedom
We call an estimate **statistically significant** when we would only expect to see such a large $t$ by chance less often than a prespecified significance level $\alpha$. 
Significance tests

We call an estimate **statistically significant** when we would only expect to see such a large $t$ by chance less often than a prespecified significance level $\alpha$.

A statistical significance test checks whether the $p$-value associated with a $t$-test is below $\alpha$, which is most often set to 0.05.
Significance tests

We call an estimate statistically significant when we would only expect to see such a large $t$ by chance less often than a prespecified significance level $\alpha$.

A statistical significance test checks whether the $p$-value associated with a $t$-test is below $\alpha$, which is most often set to 0.05.

Significance tests are tests against a specific null hypothesis, and are “conservative” in the sense of being likely to favor the null over our own hypothesis.
Are significance tests “really” conservative?

**Type I error**  Probability of falsely rejecting the null

**Type II error**  Probability of falsely accepting the null

Significance tests minimize the chance of Type I error at the expense of allowing for more Type II error
Are significance tests “really” conservative?

**Type I error**  Probability of falsely rejecting the null

**Type II error**  Probability of falsely accepting the null

Significance tests minimize the chance of Type I error at the expense of allowing for more Type II error

Is this a good idea?
Are significance tests “really” conservative?

**Type I error**  Probability of falsely rejecting the null

**Type II error**  Probability of falsely accepting the null

Significance tests minimize the chance of Type I error at the expense of allowing for more Type II error

Is this a good idea?

The null hypothesis is usually arbitrary, and our prior belief is usually that it is unlikely.

Significance tests may lead to excessive contrarianism, which is not “conservative” at all
Confidence intervals

An better alternative to $p$-values which conveys the same information is the confidence interval
Confidence intervals

An better alternative to $p$-values which conveys the same information is the confidence interval

In repeated samples from the same population, the 95% confidence interval contains the true population mean 95% of the time
Confidence intervals

An better alternative to \( p \)-values which conveys the same information is the confidence interval

In repeated samples from the same population, the 95% confidence interval contains the true population mean 95% of the time

Warning! We cannot say the truth lies in the confidence interval we calculate with 95% probability—we don’t know in this specific case
An better alternative to $p$-values which conveys the same information is the confidence interval.

In repeated samples from the same population, the 95% confidence interval contains the true population mean 95% of the time.

Warning! We cannot say the truth lies in the confidence interval we calculate with 95% probability—we don’t know in this specific case.

But if we conduct 20 studies, and in each report a 95% confidence interval, we will expect to be “wrong” in only one study (1 in 20).
Look at this interval. It "missed" the population parameter!
Calculating the confidence interval

We pick a confidence level, such as 95%

Then, we look up the critical value of $t$ containing that 95% of the $t$ distribution, and calculate:

$$\bar{x}_{\text{lower}} = \bar{x} - t_{n-1}^* \times \text{se}(\bar{x})$$
$$\bar{x}_{\text{upper}} = \bar{x} + t_{n-1}^* \times \text{se}(\bar{x})$$
Calculating the confidence interval

We pick a confidence level, such as 95%

Then, we look up the critical value of $t$ containing that 95% of the $t$ distribution, and calculate:

\[
\bar{x}_{\text{lower}} = \bar{x} - t^*_{n-1} \times \text{se}(\bar{x}) \\
\bar{x}_{\text{upper}} = \bar{x} + t^*_{n-1} \times \text{se}(\bar{x})
\]

Note that for the 95% CI, the critical value with infinite degrees of freedom is $\pm 1.96$, so 95% CIs are roughly $\pm 2$ standard errors from the estimate.
Example: Washington Same-Sex Marriage Referendum

This week, Governor Gregoire signed legislation recognizing same-sex marriage in Washington State.

Opponents promised to petition to put the law on the November ballot.

Will same-sex marriage stand, or be repealed?
The Washington Poll, October 2011, asked a prescient question of 983 Washington registered voters

Next year the legislature could pass a law allowing gay and lesbian couples to get married. If that happens, there could be a referendum in which voters would be asked to approve or reject the law.

If such a referendum were held today: Would you vote YES – that is, to keep a law in place allowing gay and lesbian couples to marry OR would you vote NO, against the law – to make it so that gay and lesbian couples could not marry?

The Washington Poll found that 55 percent of registered voters would keep same-sex marriage, and 38 percent would not.

(The Washington Poll is conducted by my colleagues, Matt Barreto and Christopher Parker, and Betsy Cooper, of UW Political Science. See www.washingtonpoll.org.)
Example: Washington Same-Sex Marriage Referendum

The Washington Poll found that 55 percent of registered voters would keep same-sex marriage, and 38 percent would not.

How certain is this result?

We will use the raw data from this survey to investigate

Some caveats:

1. Original survey was stratified, and weighted some groups more heavily; we will ignore weights

2. To simplify, we will ignore non-response and “I don’t knows”.

Because of the above (especially 2) our proportions in this lecture differ from the official results of the poll.
Example: Washington Same-Sex Marriage Referendum

Our initial $N$ of people responding YES or NO on the same-sex referendum is 979.

Of these, 61.6% say YES, they would vote to keep SSM.

Assuming the caveats above pose no problems, how certain are we the referendum will pass based on this sample?
Example: Washington Same-Sex Marriage Referendum

How likely is it that a survey of 979 random individuals from a population would find 61.6% support for a measure when really only 50% or less support the measure?

Let’s use a $t$-test:

$$ t = \frac{\bar{x} - \mu_0}{se(\bar{x})} $$

$A$ this big would appear by chance only 1 in 504,000,000,000 random samples of 979 people, (1 in 504 billion), for a $p = 0.0000000000001985$. 

Chris Adolph (UW) Statistical Inference
Example: Washington Same-Sex Marriage Referendum

How likely is it that a survey of 979 random individuals from a population would find 61.6% support for a measure when really only 50% or less support the measure?

Let’s use a $t$-test:

$$t = \frac{\bar{x} - \mu_0}{se(\bar{x})}$$

$$= \frac{\bar{x} - \mu_0}{sd(x)/\sqrt{n}}$$
Example: Washington Same-Sex Marriage Referendum

How likely is it that a survey of 979 random individuals from a population would find 61.6% support for a measure when really only 50% or less support the measure?

Let’s use a $t$-test:

$$t = \frac{\bar{x} - \mu_0}{se(\bar{x})} = \frac{\bar{x} - \mu_0}{sd(x)/\sqrt{n}} = \frac{0.616 - 0.5}{0.487/\sqrt{979}}$$
How likely is it that a survey of 979 random individuals from a population would find 61.6% support for a measure when really only 50% or less support the measure?

Let’s use a $t$-test:

$$
t = \frac{\bar{x} - \mu_0}{se(\bar{x})} = \frac{\bar{x} - \mu_0}{sd(x)/\sqrt{n}}
$$

$$
= \frac{0.616 - 0.5}{0.487/\sqrt{979}} = 7.454
$$
Example: Washington Same-Sex Marriage Referendum

How likely is it that a survey of 979 random individuals from a population would find 61.6% support for a measure when really only 50% or less support the measure?

Let’s use a $t$-test:

\[
    t = \frac{\bar{x} - \mu_0}{se(\bar{x})} = \frac{\bar{x} - \mu_0}{sd(x)/\sqrt{n}} = \frac{0.616 - 0.5}{0.487/\sqrt{979}} = 7.454
\]

A $t$ this big would appear by chance only 1 in 504,000,000,000 random samples of 979 people, (1 in 504 billion), for a $p = 0.0000000000001985$
Example: Washington Same-Sex Marriage Referendum

A $t$ this big would appear by chance only 1 in $504,000,000,000$ random samples of 979 people, (1 in 504 billion), for a $p = 0.00000000000001985$
Example: Washington Same-Sex Marriage Referendum

A $t$ this big would appear by chance only 1 in 504,000,000,000 random samples of 979 people, (1 in 504 billion), for a $p = 0.0000000000001985$

Why is this so unlikely? Suppose that in October, a bare majority of Washington registered voters really did oppose same-sex marriage.
Example: Washington Same-Sex Marriage Referendum

A $t$ this big would appear by chance only 1 in 504,000,000,000 random samples of 979 people, (1 in 504 billion), for a $p = 0.0000000000001985$

Why is this so unlikely? Suppose that in October, a bare majority of Washington registered voters really did oppose same-sex marriage.

Then to get 61.6% approval, instead of the correct 50% approval, the Washington Poll would have needed to sample $979 \times (0.616 - 0.500) = 114$ more supporters than we would expect on average in 979 random draws.
Example: Washington Same-Sex Marriage Referendum

A $t$ this big would appear by chance only 1 in $504,000,000,000$ random samples of 979 people, (1 in 504 billion), for a $p = 0.0000000000001985$

Why is this so unlikely? Suppose that in October, a bare majority of Washington registered voters really did oppose same-sex marriage.

Then to get $61.6\%$ approval, instead of the correct $50\%$ approval, the Washington Poll would have needed to sample $979 \times (0.616 - 0.500) = 114$ more supporters than we would expect on average in 979 random draws.

That’s as unlikely as flipping a coin 979 times and getting 603 heads and 376 tails.
Another way to summarize the uncertainty in our polling results is to calculate a confidence interval.

We can state with 95% confidence that the actual level of support for same-sex marriage among all Washington RVs in April was between 58.5% and 64.6%.

Notice these numbers are $61.6\% \pm 3.1\%$, which also happens to be the “margin of error” for the poll (journalists’ name for a confidence interval).

Unfortunately, “margin of error” is a misleading name: errors can be bigger than this margin, & are guaranteed to be 5% of the time!
The Washington Poll’s sample of Washington voters includes 317 voters over the age of 65, 53.9% percent of whom said they would support SSM.

Do a majority of older Washingtonians actually support SSM? Or is this a sampling error?

If we made the mistake of judging by the “margin of error” for the whole survey, we might think a majority of older voters did support SSM:

$$53.9\% - 3.1\% = 50.8\%$$
Example: Washington Same-Sex Marriage Referendum

Let’s do our own $t$-test to be sure:

\[ t = \frac{\bar{x} - \mu_0}{\text{se}(\bar{x})} \]

This is a pretty small $t$-statistic, one we would see by chance in 1 out of 6 random samples. The $p$-value is 0.161.

We find that the 95% confidence interval ranges from 48.4% to 59.5%, which is equal to our estimate of 53.9% by $\pm 5.5\%$.

We are not certain that Washington 65+'s supported SSM in October.
Example: Washington Same-Sex Marriage Referendum

Let’s do our own $t$-test to be sure:

$$
t = \frac{\bar{x} - \mu_0}{se(\bar{x})}$$

$$= \frac{\bar{x} - \mu_0}{sd(x) / \sqrt{n}}$$

This is a pretty small $t$-statistic, one we would see by chance in 1 out of 6 random samples. The $p$-value is 0.161.

We find that the 95% confidence interval ranges from 48.4% to 59.5%, which is equal to our estimate of 53.9% by $\pm 5.5$%.

We are not certain that Washington 65+’s supported SSM in October.
Example: Washington Same-Sex Marriage Referendum

Let’s do our own $t$-test to be sure:

$$t = \frac{\bar{x} - \mu_0}{\text{se}(\bar{x})}$$

$$= \frac{\bar{x} - \mu_0}{\text{sd}(x)/\sqrt{n}}$$

$$= \frac{0.539 - 0.500}{0.499/\sqrt{317}}$$

This is a pretty small $t$-statistic, one we would see by chance in 1 out of 6 random samples. The $p$-value is 0.161. We find that the 95% confidence interval ranges from 48.4% to 59.5%, which is equal to our estimate of 53.9% by $\pm 5.5\%$. We are not certain that Washington 65+'s supported SSM in October.
Example: Washington Same-Sex Marriage Referendum

Let’s do our own $t$-test to be sure:

\[
\begin{align*}
t &= \frac{\bar{x} - \mu_0}{\text{se}(\bar{x})} \\
&= \frac{\bar{x} - \mu_0}{\text{sd}(x)/\sqrt{n}} \\
&= \frac{0.539 - 0.500}{0.499/\sqrt{317}} \\
&= 1.406
\end{align*}
\]

This is a pretty small $t$-statistic, one we would see by chance in 1 out of 6 random samples. The $p$-value is 0.161.

We find that the 95% confidence interval ranges from 48.4% to 59.5%, which is equal to our estimate of 53.9% by ±5.5%.

We are not certain that Washington 65+’s supported SSM in October.
Example: Washington Same-Sex Marriage Referendum

1. Uncertainty depends on the size of the sample (which has changed)

2. Uncertainty depends on the variance of the sample (which has changed)
Change in Size of Sample

Suppose Washington elderly were evenly split on the same-sex marriage in October.

Then, to get 53.9% of 65+'s in favor in a sample of 317, The Washington Poll would need to have randomly sampled 

\[ 317 \times \left( 0.539 - 0.500 \right) = 12 \]

more people in favor than they would expect to on average. This is exactly the same as flipping a coin 317 times and getting 170 heads and 147 tails. A little unlikely, but not very unlikely.

The margin of error reported with a survey applies only to the full population. Any average we calculate for a subgroup (the young, women, Republicans, Hispanics, etc.) will have a unique confidence interval, always bigger than that for the whole sample.

The smaller the \( n \), the bigger the confidence interval, the less certain the finding.
Change in Size of Sample

Suppose Washington elderly were evenly split on the same-sex marriage in October.

Then, to get 53.9% of 65+’s in favor in a sample of 317, The Washington Poll would need to have randomly sampled \( 317 \times (0.539 - 0.500) = 12 \) more people in favor than they would expect to on average.
Change in Size of Sample

Suppose Washington elderly were evenly split on the same-sex marriage in October.

Then, to get 53.9% of 65+’s in favor in a sample of 317, The Washington Poll would need to have randomly sampled $317 \times (0.539 - 0.500) = 12$ more people in favor than they would expect to on average.

This is exactly the same as flipping a coin 317 times and getting 170 heads and 147 tails. A little unlikely, but not very unlikely.
Change in Size of Sample

Suppose Washington elderly were evenly split on the same-sex marriage in October.

Then, to get 53.9% of 65+’s in favor in a sample of 317, The Washington Poll would need to have randomly sampled $317 \times (0.539 - 0.500) = 12$ more people in favor than they would expect to on average.

This is exactly the same as flipping a coin 317 times and getting 170 heads and 147 tails. A little unlikely, but not very unlikely.

*The margin of error reported with a survey applies only to the full population.*
Change in Size of Sample

Suppose Washington elderly were evenly split on the same-sex marriage in October.

Then, to get 53.9% of 65+’s in favor in a sample of 317, The Washington Poll would need to have randomly sampled $317 \times (0.539 - 0.500) = 12$ more people in favor than they would expect to on average.

This is exactly the same as flipping a coin 317 times and getting 170 heads and 147 tails. A little unlikely, but not very unlikely.

*The margin of error reported with a survey applies only to the full population*

Any average we calculate for a subgroup (the young, women, Republicans, Hispanics, etc.) will have a unique confidence interval, always bigger than that for the whole sample.
Change in Size of Sample

Suppose Washington elderly were evenly split on the same-sex marriage in October.

Then, to get 53.9% of 65+’s in favor in a sample of 317, The Washington Poll would need to have randomly sampled $317 \times (0.539 - 0.500) = 12$ more people in favor than they would expect to on average.

This is exactly the same as flipping a coin 317 times and getting 170 heads and 147 tails. A little unlikely, but not very unlikely.

The margin of error reported with a survey applies only to the full population.

Any average we calculate for a subgroup (the young, women, Republicans, Hispanics, etc.) will have a unique confidence interval, always bigger than that for the whole sample.

The smaller the $n$, the bigger the confidence interval, the less certain the finding.
On confidence versus significance

There are two ways we could report our finding on older voters support for same-sex marriage:

**Significance test**  Based on a survey of Washington registered voters, we estimate 54% of voters over 65 years supported same-sex marriage in October. However, this estimate is not statistically significantly different from 50% at the 0.05 level.

**Confidence interval**  Based on a survey of Washington registered voters, we estimate 53.9% of voters over 65 years supported same-sex marriage in October. The 95% confidence interval for this estimate ranges from 48.4% to 59.5%, suggesting anywhere from a slight majority against same-sex marriage to a large majority in favor.
On confidence versus significance

There are two ways we could report our finding on older voters support for same-sex marriage:

**Significance test**  Based on a survey of Washington registered voters, we estimate 54% of voters over 65 years supported same-sex marriage in October. However, this estimate is not statistically significantly different from 50% at the 0.05 level.

**Confidence interval**  Based on a survey of Washington registered voters, we estimate 53.9% of voters over 65 years supported same-sex marriage in October. The 95% confidence interval for this estimate ranges from 48.4% to 59.5%, suggesting anywhere from a slight majority against same-sex marriage to a large majority in favor.
On confidence versus significance

These write-ups present the same results. They rely on the same math and the same statistical theory.

The significance test presentation obscures the substantive impact of the result in jargon, and makes it appear ignorable.

The confidence interval focuses on the substantive impact of the result, and clarifies what we can and cannot reject:

Although we aren’t sure how many older voters supported same-sex marriage in October,

it is very likely that half or more do,

and very unlikely that a large percentage of older were opposed before 2012 started.
On confidence versus significance

The significance test forces you to accept the author’s arbitrary null hypothesis.

The confidence interval allows you to choose your own null.

And shows how robust your findings are to slight changes in the null.
The irrelevance of population size

\[ t = \frac{\bar{x} - \mu_0}{sd(x)/\sqrt{n}} \]

Notice one number that doesn’t appear in this formula: the size of the population

The precision of an estimate doesn’t depend on the size of the population, only the size of the sample.

That’s why you tend to see polls using samples of 500 to 2000 respondents regardless of whether they are sampling from a small town population or the whole country.
Comparing two means

So far, we have asked how far the mean of our sample might differ from a specific value

e.g., how much does the average support for same-sex marriage differ from 0.5?

But what if we want to compare two groups in our sample?

That is, what if we want to compare two means to each other?

e.g., how much does the average support for same-sex marriage among those with a close gay friend or family member differ from support among those without (knowledge of) close contact with someone gay?
A simple cross-tab

<table>
<thead>
<tr>
<th></th>
<th>Have contact?</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td></td>
<td>Total</td>
</tr>
<tr>
<td>Support SSM</td>
<td>399</td>
<td>204</td>
<td></td>
<td>603</td>
</tr>
<tr>
<td>Oppose SSM</td>
<td>183</td>
<td>193</td>
<td></td>
<td>376</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>582</strong></td>
<td><strong>397</strong></td>
<td></td>
<td><strong>979</strong></td>
</tr>
</tbody>
</table>

Here are the two variables, support for SSM and contact, in a cross-tabulation

Let’s convert to column percentages
A simple cross-tab

<table>
<thead>
<tr>
<th></th>
<th>Have contact?</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Mean</td>
</tr>
<tr>
<td>Support SSM</td>
<td>68.6%</td>
<td>51.4%</td>
<td>61.6%</td>
</tr>
<tr>
<td>Oppose SSM</td>
<td>31.4%</td>
<td>48.6</td>
<td>38.4%</td>
</tr>
<tr>
<td>Total</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

This table shows much more support for SSM among those with contact than those without (68.6% vs. 51.4%, or 17.2% more support)

Our question:

How certain are we that this difference in mean support across groups in our sample really exists in the Washington voter population?
\( t \)-test for comparison of means

To make inferences about the *difference* in means of two samples, we can do a *difference in means* \( t \)-test

Remember the form of a \( t \)-statistic:

\[
t = \frac{\text{sample statistic of interest} - \mu_0}{\text{se(sample statistic of interest)}}
\]

Before, the sample statistic of interest was \( \bar{x} \), but now it is \( \bar{x} - \bar{y} \). We want to know if this difference is itself different from zero, so:

\[
t = \frac{(\bar{x} - \bar{y}) - 0}{\text{se(\( \bar{x} - \bar{y} \))}}
\]
\textit{t}-test for comparison of means

Our difference-of-means \textit{t}-statistic is:

\[ t = \frac{(\bar{x} - \bar{y}) - 0}{\text{se}(\bar{x} - \bar{y})} \]

Calculating this \textit{t} by hand is hard, so we’ll let the computer do it.

Then we’ll check if this \textit{t}-statistic exceeds the chosen critical value or simply calculate the probability of seeing so large a \textit{t}
Example: Washington Same-Sex Marriage Referendum

Voters in the Washington Poll sample with contact were 17.2% more likely to support same-sex marriage.

How certain are we that this difference holds in the population?

That is, how certain are we that $\Pr(\text{support}|\text{contact}) - \Pr(\text{support}|\text{no contact}) > 0$?

We can do a comparison of means $t$-test.

We find $t = 5.425$, which implies a $p$-value of $0.00000007661$.

The 95% confidence interval is ranges from $+11.0\%$ to $+23.4\%$. (What does this mean?)
Example: Household Wealth and Race

In a sample of 10,000 households, we found households headed by a self-identified white earned more, on average, than households headed by a self-identified black or Hispanic.

How certain are we that these sample results hold in the full American population?
Example: Household Wealth and Race

Let’s do a comparison-of-means $t$-test for black and white households

Average gap between black and white household wealth, in $\$k$: -496.7

t-stat: -19.8

$p$-value: 0.00000000000000022

(that’s just 1 in 4,540,000,000,000,000, or 4.5 thousand trillion)

95% CI: $-545.9$ to $-447.5$
Summing up

We’ve added several new tools to our analytic toolkit:

1. Standard errors of estimates
2. $t$-tests and confidence intervals for a sample mean
3. $t$-tests and CIs for a comparison of means
Caveats

Comparison of means tests seem especially helpful for our inference about hypotheses

We can now state whether apparent differences in conditional means are likely to be mere happenstance, or real features of the population

But are there reasons to doubt findings from a comparison of means test?
Caveats

Comparison of means tests seem especially helpful for our inference about hypotheses.

We can now state whether apparent differences in conditional means are likely to be mere happenstance, or real features of the population.

But are there reasons to doubt findings from a comparison of means test?

These tests still don’t control for confounders. So results might be spurious.
Wait! What are degrees of freedom (df)?

Degrees of freedom:
The number of separate pieces of information used to calculate a statistic

“separate” = “freely movable”

Not the same thing as the number of observations (may be the same as $N$ or less)

Relevance: how many quantities could we estimate from a set of data?

Can’t be more quantities than are left to vary!
How many things can we estimate using?

two numbers, $x_1$ and $x_2$
How many things can we estimate using?

two numbers, $x_1$ and $x_2$  

2 things
How many things can we estimate using?

- two numbers, \( x_1 \) and \( x_2 \)  
  2 things

Now I decide to add an assumption regarding the value of \( x_2 \)

- two numbers, \( x_1 \) and 3

\( \bar{x} \)
How many things can we estimate using?

two numbers, $x_1$ and $x_2$  	2 things

Now I decide to add an assumption regarding the value of $x_2$

two numbers, $x_1$ and 3  	1 thing
How many things can we estimate using?

two numbers, $x_1$ and $x_2$  
2 things

Now I decide to add an assumption regarding the value of $x_2$

two numbers, $x_1$ and 3  
1 thing

Instead, suppose I assume I know the mean?

two numbers, $x_1$ and $x_2$, and $\bar{x} = 2$
How many things can we estimate using?

two numbers, $x_1$ and $x_2$  
2 things

Now I decide to add an assumption regarding the value of $x_2$

two numbers, $x_1$ and 3  
1 thing

Instead, suppose I assume I know the mean?

two numbers, $x_1$ and $x_2$, and $\bar{x} = 2$  
1 thing
How many things can we estimate using?

two numbers, \( x_1 \) and \( x_2 \)  
2 things

Now I decide to add an assumption regarding the value of \( x_2 \)

two numbers, \( x_1 \) and 3  
1 thing

Instead, suppose I assume I know the mean?

two numbers, \( x_1 \) and \( x_2 \), and \( \bar{x} = 2 \)  
1 thing

How does this work at larger scales?

fifty numbers, \( x_1, \ldots x_{50} \), and \( \bar{x} = 2 \)
How many things can we estimate using?

- two numbers, $x_1$ and $x_2$  
  2 things

Now I decide to add an assumption regarding the value of $x_2$

- two numbers, $x_1$ and 3  
  1 thing

Instead, suppose I assume I know the mean?

- two numbers, $x_1$ and $x_2$, and $\bar{x} = 2$  
  1 thing

How does this work at larger scales?

- fifty numbers, $x_1, \ldots, x_{50}$, and $\bar{x} = 2$  
  49 things
How many things can we estimate using?

- two numbers, $x_1$ and $x_2$  

Now I decide to add an assumption regarding the value of $x_2$

- two numbers, $x_1$ and $3$  

Instead, suppose I assume I know the mean?

- two numbers, $x_1$ and $x_2$, and $\bar{x} = 2$  

How does this work at larger scales?

- fifty numbers, $x_1, \ldots, x_{50}$, and $\bar{x} = 2$  

- fifty numbers, $x_1, \ldots, x_{50}$, $\bar{x} = 2$, and $\sigma^2 = 0.5$
How many things can we estimate using?

two numbers, $x_1$ and $x_2$ 2 things

Now I decide to add an assumption regarding the value of $x_2$

two numbers, $x_1$ and 3 1 thing

Instead, suppose I assume I know the mean?

two numbers, $x_1$ and $x_2$, and $\bar{x} = 2$ 1 thing

How does this work at larger scales?

fifty numbers, $x_1, \ldots x_{50}$, and $\bar{x} = 2$ 49 things

fifty numbers, $x_1, \ldots x_{50}$, $\bar{x} = 2$, and $\sigma^2 = 0.5$ 48 things
Degrees of freedom (df): the remaining allowed ways you could move the data.

If we make as many assumptions as there are observations, nothing left to estimate.