CSSS/SOC/STAT 321
Case-Based Social Statistics I

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 on 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?

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

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.

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 correclty, 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 correclty, 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 correclty, 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 correclty, 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

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

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

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

Comnvenience 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 sample help make it so?

When sampling goes wrong

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

When sampling goes wrong

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

If a stratified sample has the wrong weights, will adding more samples make it representative?

When sampling goes wrong

If a random sample is non-representative, will adding more random sample 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 sample 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 sample 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 scientificially 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 need two new concepts:

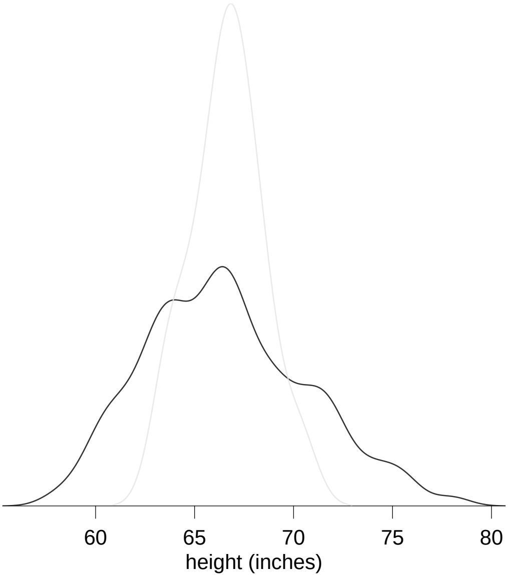
the standard error and t-statistic,

and two new continuous probability distributions:

the χ^2 distribution and t distribution

Today, we focus on the *standard error*

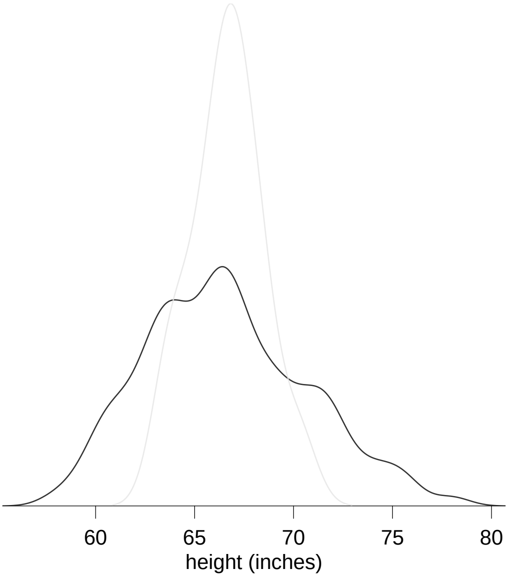
Student height example



In a class of 179 students, 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

Student height example

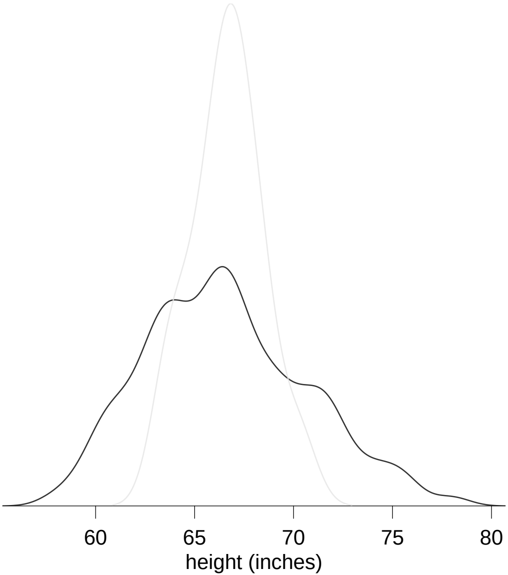


About 2/3s of the class submitted heights.

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?

Student height example



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

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)

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*

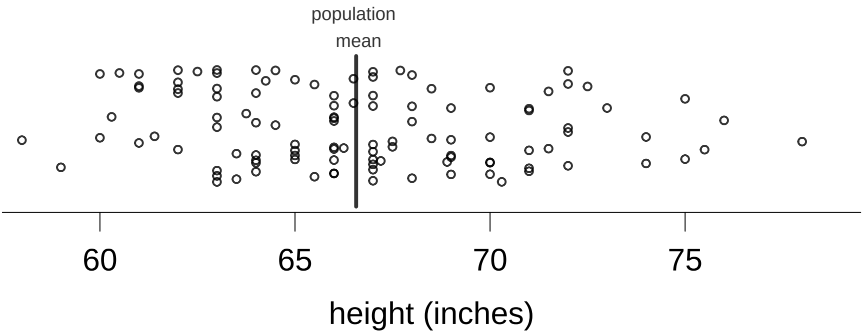
Error The difference between the sample mean (estimate) and population mean (“truth”)

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.

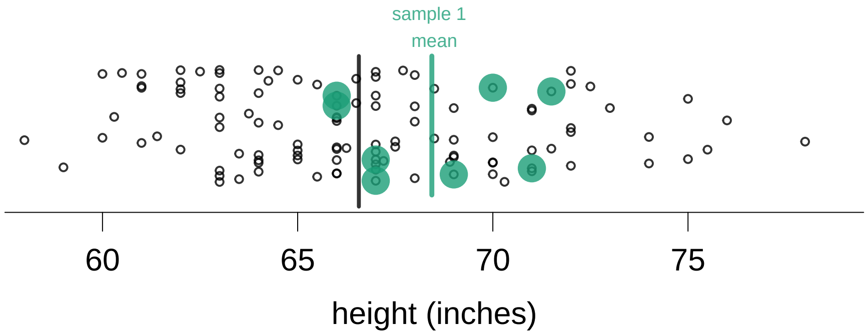
Your book also calls this the *standard deviation of the sampling mean* (Chapter 7)



Above are the submitted height data and their mean

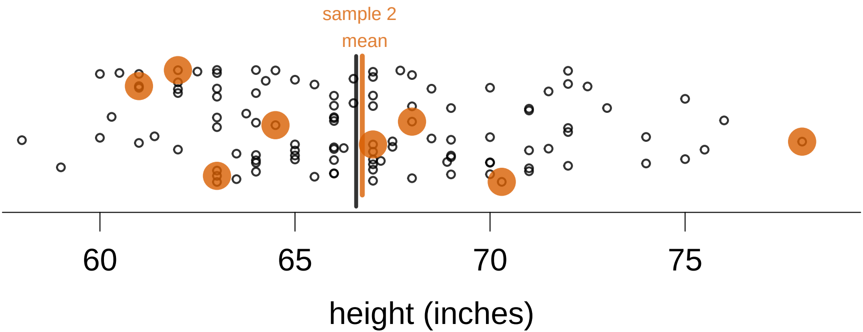
We will replicate this exercise in our (smaller) class

While we record the heights of several samples from our class, . . .



I will explore a set of 14 pre-selected samples of 8 heights each, using the old 179 class data

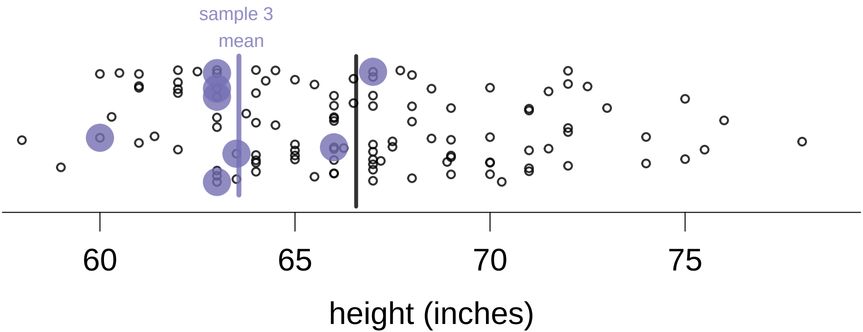
Note that the means and variances of the samples can differ from the full population. . .



I will explore a set of 14 pre-selected samples of 8 heights each

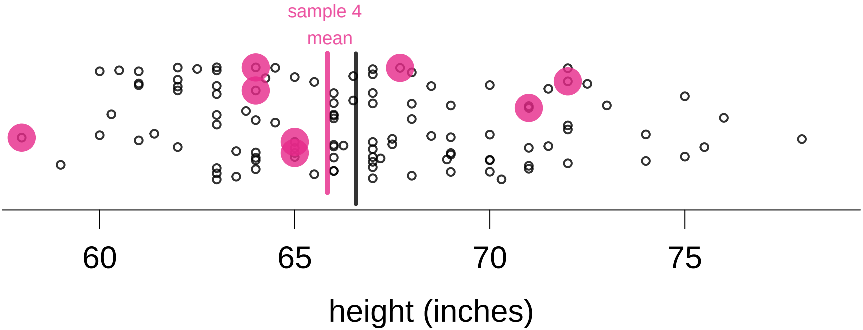
Note that the means and variances of the samples can differ from the full population…

…but can also resemble it fairly closely



Sometimes an individual sample can be so far off that it would mislead us considerably

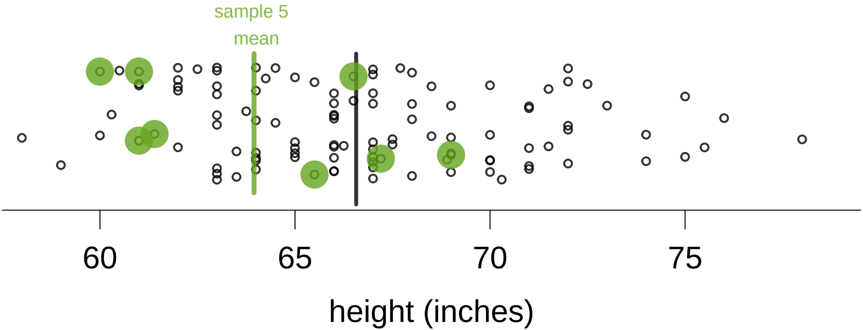
This is more likely the smaller the sample



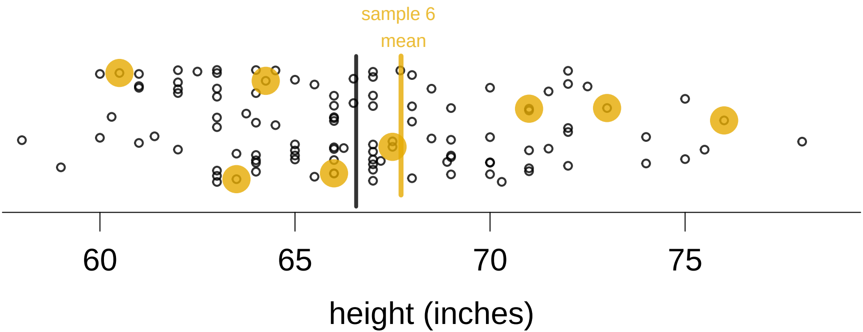
Sometimes an individual sample can be so far off that it would mislead us considerably

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

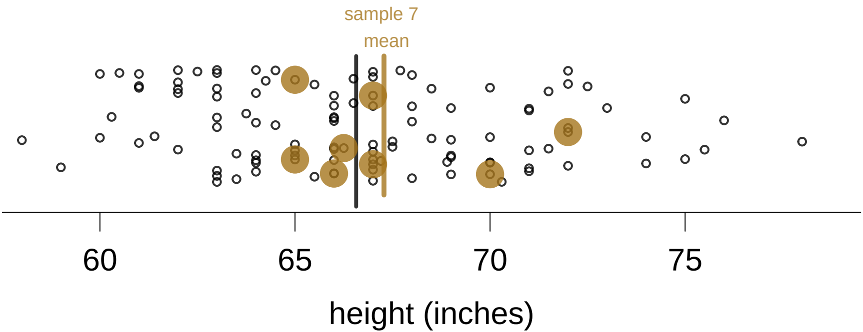


Moreover, the sample mean seems just as likely to be below the population mean . . .



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

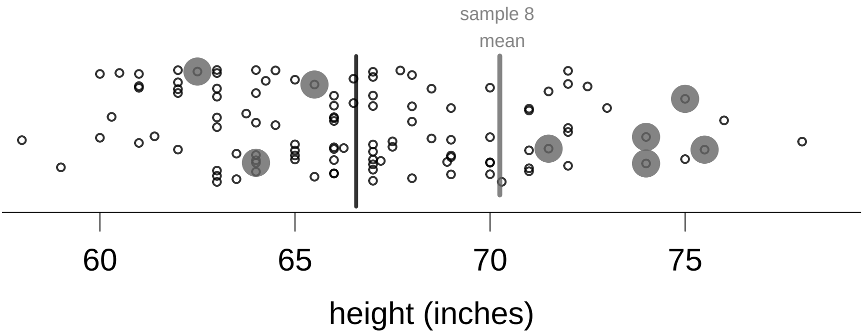
. . . as above it.



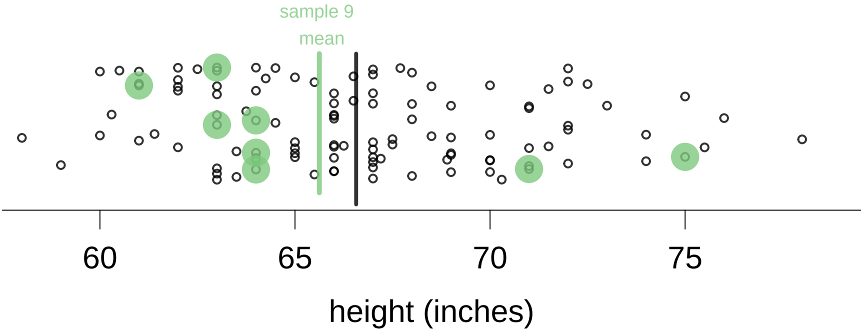
Moreover, the sample mean seems just as likely to be below the population mean . . .

. . . as above it.

When an estimate is neither systematically too high or too low, it is unbiased

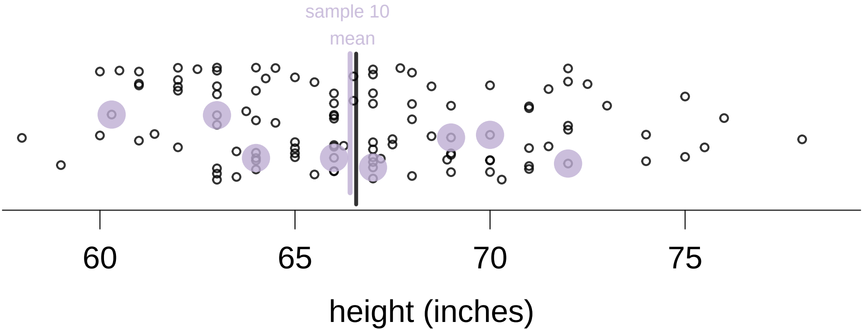


Unbiased estimators can still sometimes be far from the true value

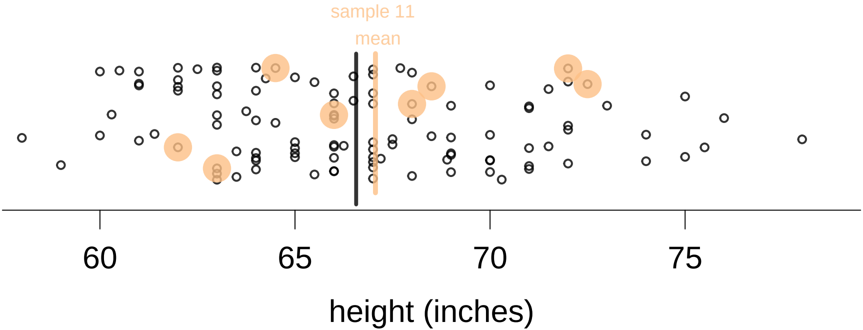


Unbiased estimators can still sometimes be far from the true value

But on average, over many samples, unbiased estimators will equal the truth, rather than show systematic *bias* up or down

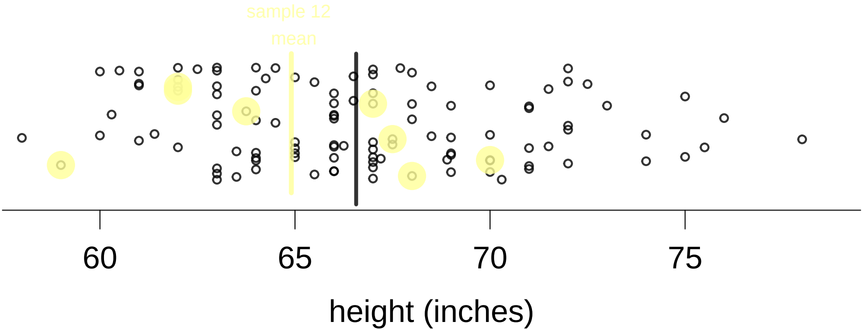


We also want out estimates to be *close* to the truth most of the time, a separate issue from bias

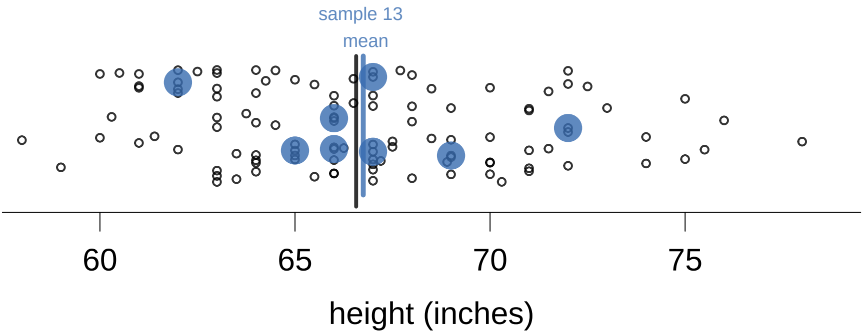


We also want out estimates to be *close* to the truth most of the time, a separate issue from bias

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

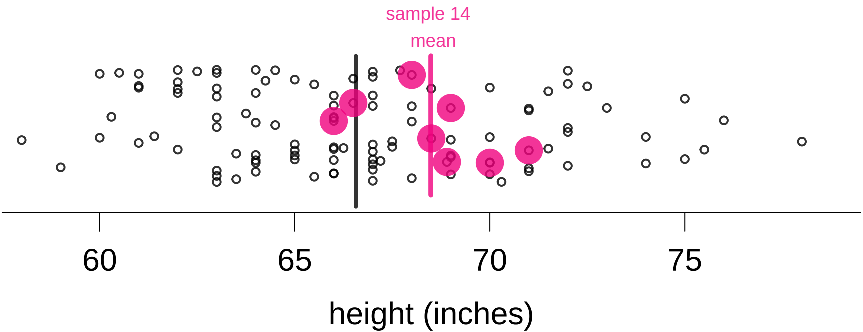


Notice that although the sample mean bounces back and forth,



Notice that although the sample mean bounces back and forth,

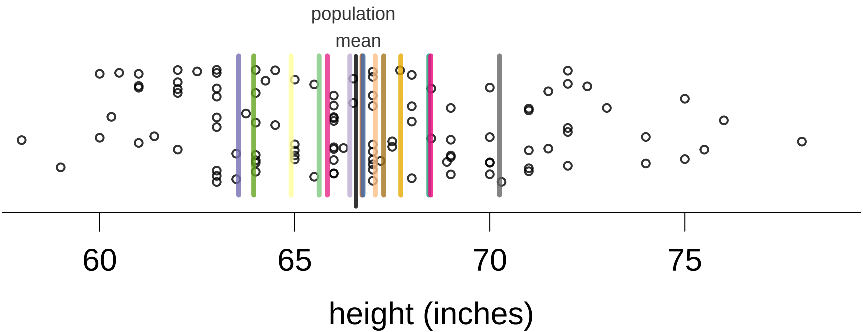
it tends to stay close to the population mean



Notice that although the sample mean bounces back and forth,

it tends to stay close to the population mean

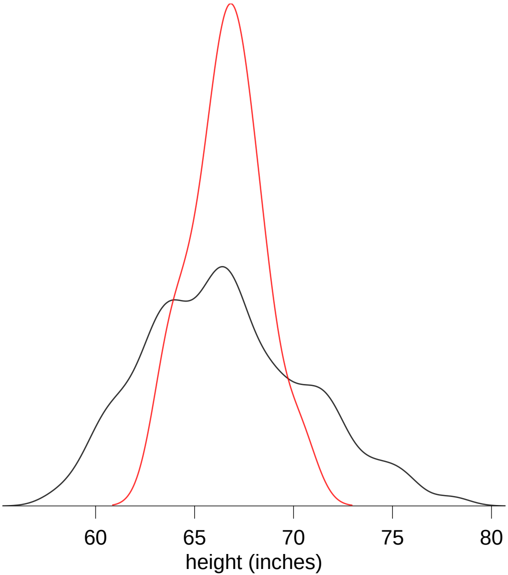
and doesn't range as far as the data itself



That is, the standard deviation of the sample means is smaller than the standard deviation of the data itself:

$$\text{sd}(\bar{x}) < \text{sd}(x)$$

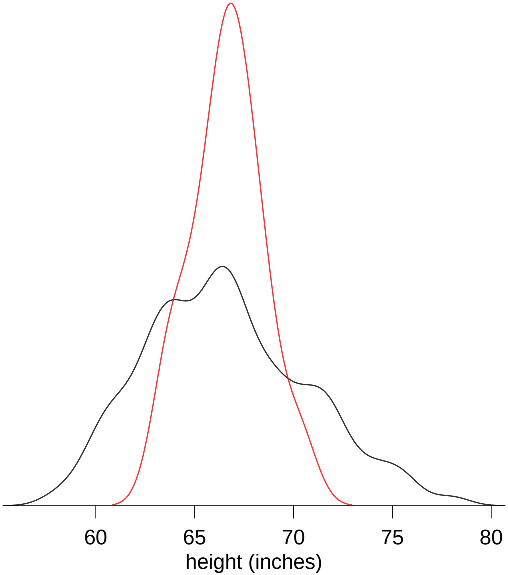
Student height example



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.

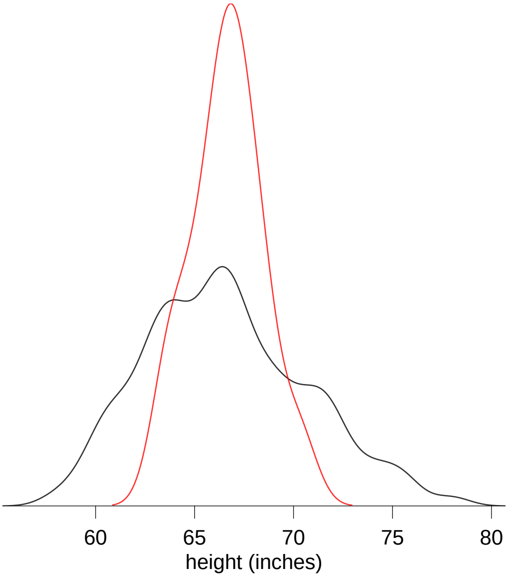
Student height example



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)

Student height example



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:

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

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, $\mathrm{se}(\bar{x})$ gets smaller as the sampled n increases

The Square Root Law

Formula for the standard error of the mean:

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

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

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

Because $\mathrm{se}(\bar{x})$ 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!

	Mean	Sampling Error	Std Dev	Std Error
Population	66.6	—	4.1	—
Sample 1	68.4	1.9	2.2	0.8
Sample 2	66.7	0.2	5.5	2.0
Sample 3	63.6	−3.0	2.1	0.8
Sample 4	65.8	−0.7	4.4	1.6
Sample 5	64.0	−2.6	3.5	1.2
Sample 6	67.7	1.2	5.2	1.9
Sample 7	67.3	0.7	2.5	0.9
Sample 8	70.3	3.7	5.4	1.9
Sample 9	65.6	−0.9	4.8	1.7
Sample 10	66.4	−0.2	3.9	1.4
Sample 11	67.1	0.5	3.9	1.4
Sample 12	64.9	−1.7	3.8	1.3
Sample 13	66.8	0.2	2.9	1.0
Sample 14	68.5	1.9	1.7	0.6
Sample Mean	66.6		Avg SE	1.4
Pop Mean	66.6		RMSE	1.8

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

	Mean	Sampling Error	Std Dev	Std Error
Population	66.6	—	4.1	—
Sample 1	68.4	1.9	2.2	0.8
Sample 2	66.7	0.2	5.5	2.0
Sample 3	63.6	−3.0	2.1	0.8
Sample 4	65.8	−0.7	4.4	1.6
Sample 5	64.0	−2.6	3.5	1.2
Sample 6	67.7	1.2	5.2	1.9
Sample 7	67.3	0.7	2.5	0.9
Sample 8	70.3	3.7	5.4	1.9
Sample 9	65.6	−0.9	4.8	1.7
Sample 10	66.4	−0.2	3.9	1.4
Sample 11	67.1	0.5	3.9	1.4
Sample 12	64.9	−1.7	3.8	1.3
Sample 13	66.8	0.2	2.9	1.0
Sample 14	68.5	1.9	1.7	0.6
Sample Mean	66.6		Avg SE	1.4
Pop Mean	66.6		RMSE	1.8

The population mean, 66.6 inches, is the average height in the class

The average sample mean matches the population mean almost exactly

	Mean	Sampling Error	Std Dev	Std Error
Population	66.6	—	4.1	—
Sample 1	68.4	1.9	2.2	0.8
Sample 2	66.7	0.2	5.5	2.0
Sample 3	63.6	−3.0	2.1	0.8
Sample 4	65.8	−0.7	4.4	1.6
Sample 5	64.0	−2.6	3.5	1.2
Sample 6	67.7	1.2	5.2	1.9
Sample 7	67.3	0.7	2.5	0.9
Sample 8	70.3	3.7	5.4	1.9
Sample 9	65.6	−0.9	4.8	1.7
Sample 10	66.4	−0.2	3.9	1.4
Sample 11	67.1	0.5	3.9	1.4
Sample 12	64.9	−1.7	3.8	1.3
Sample 13	66.8	0.2	2.9	1.0
Sample 14	68.5	1.9	1.7	0.6
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

	Mean	Sampling Error	Std Dev	Std Error
Population	66.6	—	4.1	—
Sample 1	68.4	1.9	2.2	0.8
Sample 2	66.7	0.2	5.5	2.0
Sample 3	63.6	−3.0	2.1	0.8
Sample 4	65.8	−0.7	4.4	1.6
Sample 5	64.0	−2.6	3.5	1.2
Sample 6	67.7	1.2	5.2	1.9
Sample 7	67.3	0.7	2.5	0.9
Sample 8	70.3	3.7	5.4	1.9
Sample 9	65.6	−0.9	4.8	1.7
Sample 10	66.4	−0.2	3.9	1.4
Sample 11	67.1	0.5	3.9	1.4
Sample 12	64.9	−1.7	3.8	1.3
Sample 13	66.8	0.2	2.9	1.0
Sample 14	68.5	1.9	1.7	0.6
Sample Mean	66.6		Avg SE	1.4
Pop Mean	66.6		RMSE	1.8

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

	Mean	Sampling Error	Std Dev	Std Error
Population	66.6	—	4.1	—
Sample 1	68.4	1.9	2.2	0.8
Sample 2	66.7	0.2	5.5	2.0
Sample 3	63.6	−3.0	2.1	0.8
Sample 4	65.8	−0.7	4.4	1.6
Sample 5	64.0	−2.6	3.5	1.2
Sample 6	67.7	1.2	5.2	1.9
Sample 7	67.3	0.7	2.5	0.9
Sample 8	70.3	3.7	5.4	1.9
Sample 9	65.6	−0.9	4.8	1.7
Sample 10	66.4	−0.2	3.9	1.4
Sample 11	67.1	0.5	3.9	1.4
Sample 12	64.9	−1.7	3.8	1.3
Sample 13	66.8	0.2	2.9	1.0
Sample 14	68.5	1.9	1.7	0.6
Sample Mean	66.6		Avg SE	1.4
Pop Mean	66.6		RMSE	1.8

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

	Mean	Sampling Error	Std Dev	Std Error
Population	66.6	—	4.1	—
Sample 1	68.4	1.9	2.2	0.8
Sample 2	66.7	0.2	5.5	2.0
Sample 3	63.6	−3.0	2.1	0.8
Sample 4	65.8	−0.7	4.4	1.6
Sample 5	64.0	−2.6	3.5	1.2
Sample 6	67.7	1.2	5.2	1.9
Sample 7	67.3	0.7	2.5	0.9
Sample 8	70.3	3.7	5.4	1.9
Sample 9	65.6	−0.9	4.8	1.7
Sample 10	66.4	−0.2	3.9	1.4
Sample 11	67.1	0.5	3.9	1.4
Sample 12	64.9	−1.7	3.8	1.3
Sample 13	66.8	0.2	2.9	1.0
Sample 14	68.5	1.9	1.7	0.6
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

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

An estimate of RMSE we construct just from one sample

	Mean	Sampling Error	Std Dev	Std Error
Population	66.6	—	4.1	—
Sample 1	68.4	1.9	2.2	0.8
Sample 2	66.7	0.2	5.5	2.0
Sample 3	63.6	−3.0	2.1	0.8
Sample 4	65.8	−0.7	4.4	1.6
Sample 5	64.0	−2.6	3.5	1.2
Sample 6	67.7	1.2	5.2	1.9
Sample 7	67.3	0.7	2.5	0.9
Sample 8	70.3	3.7	5.4	1.9
Sample 9	65.6	−0.9	4.8	1.7
Sample 10	66.4	−0.2	3.9	1.4
Sample 11	67.1	0.5	3.9	1.4
Sample 12	64.9	−1.7	3.8	1.3
Sample 13	66.8	0.2	2.9	1.0
Sample 14	68.5	1.9	1.7	0.6
Sample Mean	66.6		Avg SE	1.4
Pop Mean	66.6		RMSE	1.8

The average standard error is how much we expect to miss the population mean in repeated samples

Avg SE should match RMSE pretty closely

	Mean	Sampling Error	Std Dev	Std Error
Population	66.6	—	4.1	—
Sample 1	68.4	1.9	2.2	0.8
Sample 2	66.7	0.2	5.5	2.0
Sample 3	63.6	−3.0	2.1	0.8
Sample 4	65.8	−0.7	4.4	1.6
Sample 5	64.0	−2.6	3.5	1.2
Sample 6	67.7	1.2	5.2	1.9
Sample 7	67.3	0.7	2.5	0.9
Sample 8	70.3	3.7	5.4	1.9
Sample 9	65.6	−0.9	4.8	1.7
Sample 10	66.4	−0.2	3.9	1.4
Sample 11	67.1	0.5	3.9	1.4
Sample 12	64.9	−1.7	3.8	1.3
Sample 13	66.8	0.2	2.9	1.0
Sample 14	68.5	1.9	1.7	0.6
Sample Mean	66.6		Avg SE	1.4
Pop Mean	66.6		RMSE	1.8

Standard errors tell us how much we can trust our sample estimates

If standard errors are close to RMSE, then they are close to the true error in the estimate

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

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

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 μ_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 μ_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 μ_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 μ_0.

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

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

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 μ_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 μ_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 μ_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 μ_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;

Hypothesis testing

Null hypothesis The reference point *we arbitrarily chose*. Denote as μ_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 μ_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 μ_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

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 μ_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 μ_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 μ_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 μ_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 μ_0

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

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

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

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,

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

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,

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

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:

$$t = \frac{\bar{x}}{\mathrm{sd}(x)/\sqrt{n}}$$

The t-statistic

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

As with z-scores, our goal is to say how "unusual" the observed t is with reference to the distribution of t

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)

We need to take a detour back to probability theory to figure out its distribution

The χ^2 distribution

Recall the Normal distribution describes the behavior of the sum of an infinite number of independent random variables

The χ^2 distribution

Recall the Normal distribution describes the behavior of the sum of an infinite number of independent random variables

What if we have a variable X^2 that is the sum of $n < \infty$ *squared* independent standard Normal random variables?

$$X^2 = x_1^2 + x_2^2 + \dots x_n^2$$

This is the sum of a finite set of Normal random variables, so the Normal doesn't quite apply

What distribution does this sum really follow?

The χ^2 distribution

$$X^2 = x_1^2 + x_2^2 + \ldots x_k^2, \qquad n < \infty$$

follows a χ^2 (chi-squared) distribution, with this pdf:

The χ^2 distribution

$$X^2 = x_1^2 + x_2^2 + \ldots x_k^2, \qquad n < \infty$$

follows a χ^2 (chi-squared) distribution, with this pdf:

$$\chi^2(X_n^2) = \frac{1}{2^{n/2}\Gamma(n/2)}(X^2)^{(n-2)/2}\exp(-X/2)$$

The χ^2 distribution

$$X^2 = x_1^2 + x_2^2 + \ldots x_k^2, \qquad n < \infty$$

follows a χ^2 (chi-squared) distribution, with this pdf:

$$\chi^2(X_n^2) = \frac{1}{2^{n/2}\Gamma(n/2)}(X^2)^{(n-2)/2}\exp(-X/2)$$

which has "degrees of freedom" n

The χ^2 distribution

$$X^2 = x_1^2 + x_2^2 + \ldots x_k^2, \qquad n < \infty$$

follows a χ^2 (chi-squared) distribution, with this pdf:

$$\chi^2(X_n^2) = \frac{1}{2^{n/2}\Gamma(n/2)}(X^2)^{(n-2)/2}\exp(-X/2)$$

which has “degrees of freedom” n

$\Gamma(\cdot)$ is the Gamma function, an interpolated factorial

The χ^2 distribution

$$X^2 = x_1^2 + x_2^2 + \dots x_k^2, \qquad n < \infty$$

follows a χ^2 (chi-squared) distribution, with this pdf:

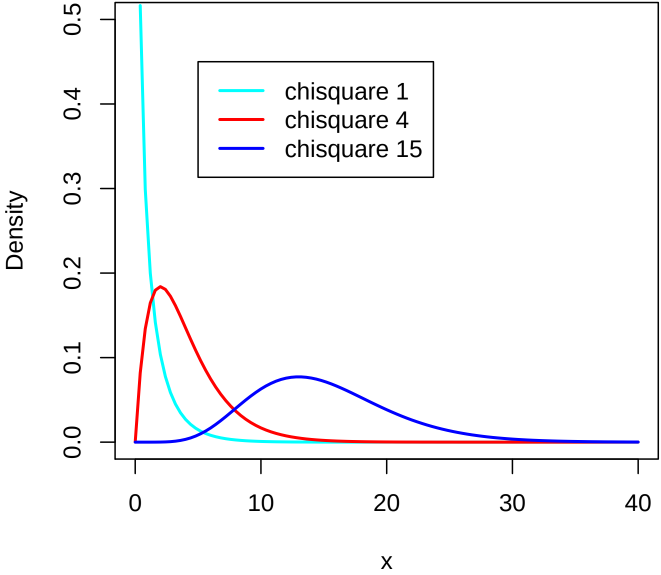
$$\chi^2(X_n^2) = \frac{1}{2^{n/2}\Gamma(n/2)}(X^2)^{(n-2)/2}\exp(-X/2)$$

which has "degrees of freedom" n

$\Gamma(\cdot)$ is the Gamma function, an interpolated factorial

Moments: $\mathrm{E}(\chi^2) = n$ and $\mathrm{Var}(\chi^2) = 2n$

χ^2 approaches the Normal as k increases



The t distribution

The χ_2 is a key building block for another useful distribution, which describes the behavior of a very specific ratio:

$$\frac{Z}{\sqrt{X^2/n}}$$

where Z is Normally distributed and
X^2 is distributed χ^2 with n degrees of freedom.

The t distribution

The χ_2 is a key building block for another useful distribution, which describes the behavior of a very specific ratio:

$$\frac{Z}{\sqrt{X^2/n}}$$

where Z is Normally distributed and
X^2 is distributed χ^2 with n degrees of freedom.

The ratio above follows the Student's t distribution with n degrees of freedom:

$$t(n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\Gamma(n/2)}} \times \frac{1}{(1 + t^2/n)^{(n+1)/2}}$$

The t distribution

$$t(n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\Gamma(n/2)}} \times \frac{1}{(1+t^2/n)^{(n+1)/2}}$$

Moments:

$\mathrm{E}(t) = 0$ (we could change this)

The t distribution

$$t(n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\Gamma(n/2)}} \times \frac{1}{(1 + t^2/n)^{(n+1)/2}}$$

Moments:

$\mathrm{E}(t) = 0$ (we could change this)

$\mathrm{Var}(t) = n/(n-2)$ for $n > 2$. Not defined for $n = 1$.

The t distribution

$$t(n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\Gamma(n/2)}} \times \frac{1}{(1+t^2/n)^{(n+1)/2}}$$

Moments:

$\mathrm{E}(t) = 0$ (we could change this)

$\mathrm{Var}(t) = n/(n-2)$ for $n > 2$. Not defined for $n = 1$.

As the degrees of freedom grow, the t distribution approximates the Normal

The t distribution

$$t(n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\Gamma(n/2)}} \times \frac{1}{(1 + t^2/n)^{(n+1)/2}}$$

Moments:

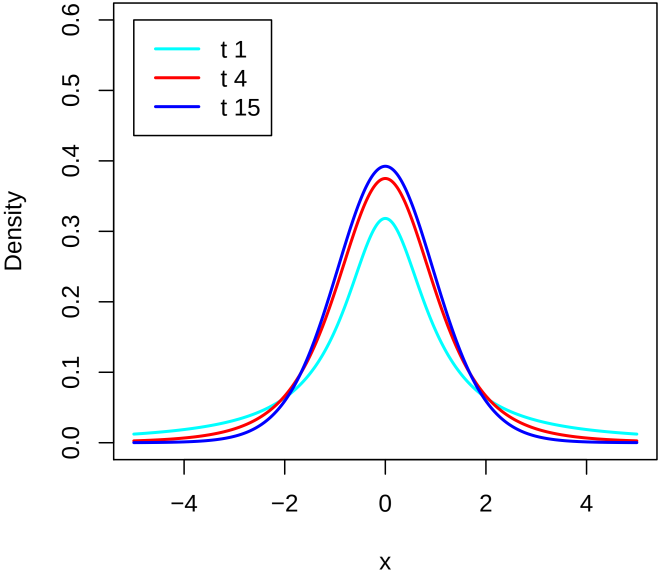
$\mathrm{E}(t) = 0$ (we could change this)

$\mathrm{Var}(t) = n/(n-2)$ for $n > 2$. Not defined for $n = 1$.

As the degrees of freedom grow, the t distribution approximates the Normal

For low degrees of freedom, the t has fatter tails

Example t distributions



The t distribution

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

That is, $\mathrm{P}(t) = t(5)$

The t distribution

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

That is, $\mathrm{P}(t) = t(5)$

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, $\text{P}(t) = t(5)$

What are the "critical" values of t we would see just

- once in 10 draws?

The t distribution

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

That is, $\mathrm{P}(t) = t(5)$

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

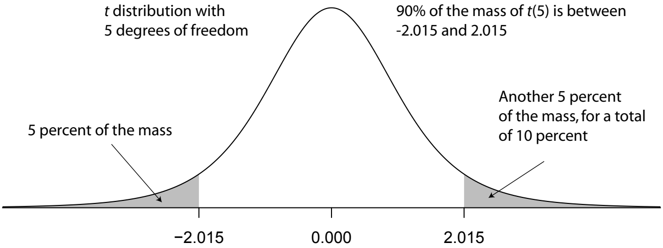
That is, $\mathrm{P}(t) = t(5)$

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?

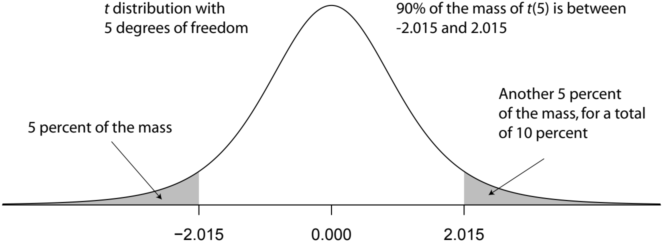
Areas under the t



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

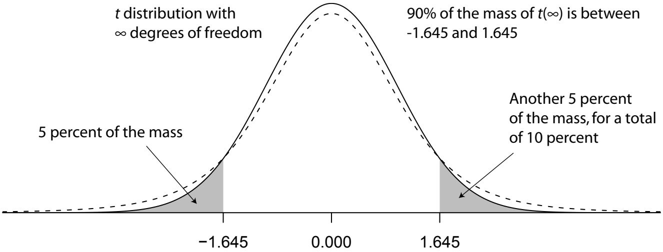
Areas under the t



Note that the above distribution is a t with 4 degrees of freedom

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

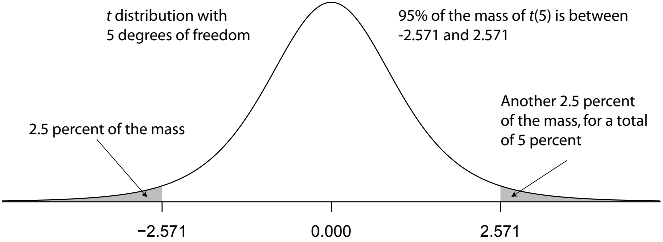
Areas under the t



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

More dfs thus make the tails thinner, and make critical values smaller

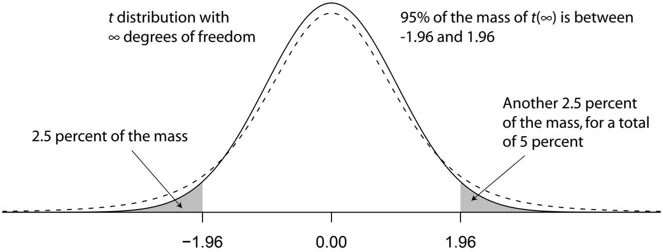
Areas under the t



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

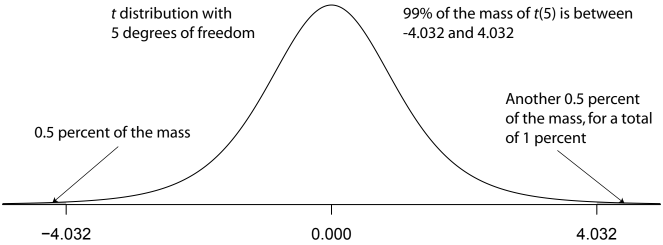
Areas under the t



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

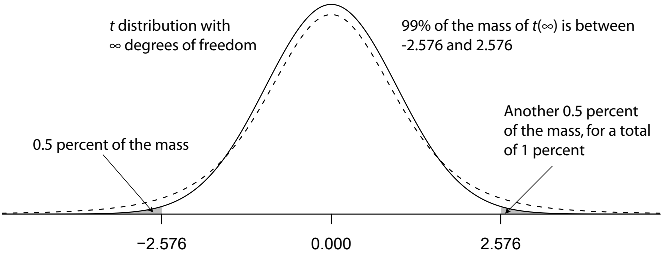
Areas under the t



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 $t(n)$

Test level	Interval	$\text{df} = 5$	$\text{df} = \infty$
0.1 level	90%	2.015	1.645
0.05 level	95%	2.571	1.960
0.01 level	99%	4.032	2.576

These will be very useful for quantifying the uncertainty of estimates

The t-statistic

Note that the t-statistic should be t distributed!

1. $\bar{x}$: The mean of x is the sum of a large number of independent variables, and thus will tend to be Normally distributed, by the Central Limit Theorem
2. $\mathrm{var}(x)$: The variance of x is the sum of n squared variables, and is thus χ^2 distributed
3. The ratio of a Normal variable and the square root of a χ^2 variable is t-distributed

The t-statistic

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

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-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, uncertainty of that estimate $\mathrm{sd}(x)/\sqrt{n}$, how likely is it that the population mean of x is actual μ_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, uncertainty of that estimate $\mathrm{sd}(x)/\sqrt{n}$, how likely is it that the population mean of x is actual μ_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 μ_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, uncertainty of that estimate $\mathrm{sd}(x)/\sqrt{n}$, how likely is it that the population mean of x is actual μ_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 μ_0
2. A typical sample from a population mean that is larger than μ_0

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

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

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 siginificance level α

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 siginificance level α

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 siginificance level α

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

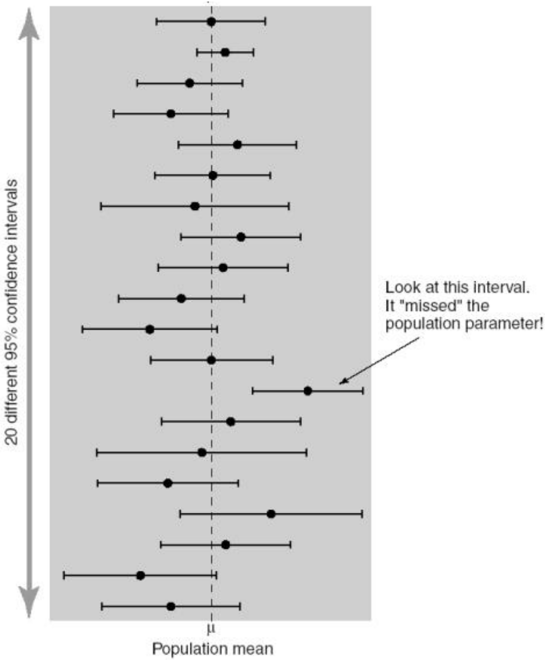
Confidence intervals

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

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)


20 different 95% confidence intervals
Look at this interval.
It "missed" the
population parameter!
μ
Population mean

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:

$$\begin{aligned} \bar{x}^{\text{lower}} &= \bar{x} - t^*_{n-1} \times \text{se}(\bar{x}) \\ \bar{x}^{\text{upper}} &= \bar{x} + t^*_{n-1} \times \text{se}(\bar{x}) \end{aligned}$$

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:

$$\begin{aligned} \bar{x}^{\text{lower}} &= \bar{x} - t^*_{n-1} \times \text{se}(\bar{x}) \\ \bar{x}^{\text{upper}} &= \bar{x} + t^*_{n-1} \times \text{se}(\bar{x}) \end{aligned}$$

Note that for the 95% CI, the critical value with infinite degrees of freedom is ± 1.96, so 95% CIs are roughly ± 2 standard errors from the estimate

Example: Washington Same-Sex Marriage Referendum

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

Opponents successfully petitioned to put the law on the November ballot.

By the time you hear this lecture, this issue will be decided...
Did the poll one year ago get it right?

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:

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

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:

$$
\begin{aligned}
t &= \frac{\bar{x} - \mu_0}{\mathrm{se}(\bar{x})} \\
&= \frac{\bar{x} - \mu_0}{\mathrm{sd}(x)/\sqrt{n}}
\end{aligned}
$$

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:

$$
\begin{aligned}
t &= \frac{\bar{x} - \mu_0}{\mathrm{se}(\bar{x})} \\
&= \frac{\bar{x} - \mu_0}{\mathrm{sd}(x)/\sqrt{n}} \\
&= \frac{0.616 - 0.5}{0.487/\sqrt{979}}
\end{aligned}
$$

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:

$$
\begin{aligned}
t &= \frac{\bar{x} - \mu_0}{\text{se}(\bar{x})} \\
&= \frac{\bar{x} - \mu_0}{\text{sd}(x)/\sqrt{n}} \\
&= \frac{0.616 - 0.5}{0.487/\sqrt{979}} \\
&= 7.454
\end{aligned}
$$

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:

$$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{\mathrm{se}(\bar{x})} \\ &= \frac{\bar{x} - \mu_0}{\mathrm{sd}(x)/\sqrt{n}} \\ &= \frac{0.616 - 0.5}{0.487/\sqrt{979}} \\ &= 7.454 \end{aligned}$$

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 $p = 0.000000000001985$

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 $p = 0.000000000001985$

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 $p = 0.000000000001985$

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 $p = 0.000000000001985$

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 $p = 0.000000000001985$

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.

Example: Washington Same-Sex Marriage Referendum

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!

Example: Washington Same-Sex Marriage Referendum

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:

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

Example: Washington Same-Sex Marriage Referendum

Let's do our own t-test to be sure:

$$
\begin{aligned}
t &= \frac{\bar{x} - \mu_0}{\text{se}(\bar{x})} \\
&= \frac{\bar{x} - \mu_0}{\text{sd}(x)/\sqrt{n}}
\end{aligned}
$$

Example: Washington Same-Sex Marriage Referendum

Let's do our own t-test to be sure:

$$
\begin{aligned}
t &= \frac{\bar{x} - \mu_0}{\mathrm{se}(\bar{x})} \\
&= \frac{\bar{x} - \mu_0}{\mathrm{sd}(x)/\sqrt{n}} \\
&= \frac{0.539 - 0.500}{0.499/\sqrt{317}}
\end{aligned}
$$

Example: Washington Same-Sex Marriage Referendum

Let's do our own t-test to be sure:

$$
\begin{aligned}
t &= \frac{\bar{x} - \mu_0}{\mathrm{se}(\bar{x})} \\
&= \frac{\bar{x} - \mu_0}{\mathrm{sd}(x)/\sqrt{n}} \\
&= \frac{0.539 - 0.500}{0.499/\sqrt{317}} \\
&= 1.406
\end{aligned}
$$

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

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.

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.

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 you own null

And shows how robust your findings are to slight changes in the null

The irrelevance of population size

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

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

	Have contact?		
	Yes	No	Total
Support SSM	399	204	603
Oppose SSM	183	193	376
Total	582	397	979

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

Let's convert to column percentages

A simple cross-tab

	Have contact? Yes	No	Mean
Support SSM	68.6%	51.4%	61.6%
Oppose SSM	31.4%	48.6	38.4%
Total	100.0%	100.0%	100.0%

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:

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

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:

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

t-test for comparison of means

Our difference-of-means t-statistic is:

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

Calculating this t by hand is hard, so we'll let the computer do it.

Then we'll check if this t-statistic exceeds the chosen critical value or simply calculate the probability of seeing so large a 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 Weath 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 Weath 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

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 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 $\sigma^2 = 0.5$

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)

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

Derivation of the standard error of the mean

$$\mathrm{var}(\bar{x}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

Derivation of the standard error of the mean

$$\begin{aligned}
\mathrm{var}(\bar{x}) &= \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\
&= \mathrm{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\right)^2\right)
\end{aligned}$$

Derivation of the standard error of the mean

$$
\begin{aligned}
\mathrm{var}(\bar{x}) &= \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\
&= \mathrm{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\right)^2\right) \\
&= \mathrm{E}\left(\frac{1}{n^2}\left(\sum_{i=1}^{n} x_i - \mathrm{E}\left(\sum_{i=1}^{n} x_i\right)\right)^2\right)
\end{aligned}
$$

Derivation of the standard error of the mean

$$
\begin{aligned}
\text{var}(\bar{x}) &= \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\
&= \text{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \text{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\right)^2\right) \\
&= \text{E}\left(\frac{1}{n^2}\left(\sum_{i=1}^{n} x_i - \text{E}\left(\sum_{i=1}^{n} x_i\right)\right)^2\right) \\
&= \frac{1}{n^2}\text{E}\left(\left(\sum_{i=1}^{n} x_i - \text{E}\left(\sum_{i=1}^{n} x_i\right)\right)^2\right)
\end{aligned}
$$

Derivation of the standard error of the mean

$$
\begin{aligned}
\text{var}(\bar{x}) &= \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\
&= \text{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \text{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\right)^2\right) \\
&= \text{E}\left(\frac{1}{n^2}\left(\sum_{i=1}^{n} x_i - \text{E}\left(\sum_{i=1}^{n} x_i\right)\right)^2\right) \\
&= \frac{1}{n^2}\text{E}\left(\left(\sum_{i=1}^{n} x_i - \text{E}\left(\sum_{i=1}^{n} x_i\right)\right)^2\right) \\
&= \frac{1}{n^2}\text{var}\left(\sum_{i=1}^{n} x_i\right)
\end{aligned}
$$

Derivation of the standard error of the mean

Now we make use of the fact that for uncorrelated $x_1, \ldots, x_i, \ldots x_n$,
$\text{var}\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} \text{var}\left(x_i\right)$,

Derivation of the standard error of the mean

Now we make use of the fact that for uncorrelated $x_1, \ldots, x_i, \ldots x_n$, $\mathrm{var}\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} \mathrm{var}\left(x_i\right)$, and write:

$$\mathrm{var}(\bar{x}) = \frac{1}{n^2}\mathrm{var}\left(\sum_{i=1}^{n} x_i\right)$$

Derivation of the standard error of the mean

Now we make use of the fact that for uncorrelated $x_1, \ldots, x_i, \ldots x_n$, $\mathrm{var}\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} \mathrm{var}\left(x_i\right)$, and write:

$$\begin{aligned}
\mathrm{var}(\bar{x}) &= \frac{1}{n^2}\mathrm{var}\left(\sum_{i=1}^{n} x_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}\left(x_i\right)
\end{aligned}$$

Derivation of the standard error of the mean

Now we make use of the fact that for uncorrelated $x_1, \ldots, x_i, \ldots x_n$, $\text{var}\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} \text{var}\left(x_i\right)$, and write:

$$\begin{aligned}
\text{var}(\bar{x}) &= \frac{1}{n^2}\text{var}\left(\sum_{i=1}^{n} x_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \text{var}\left(x_i\right) \\
&= \frac{1}{n^2} n \times \text{var}\left(x\right)
\end{aligned}$$

Derivation of the standard error of the mean

Now we make use of the fact that for uncorrelated $x_1, \ldots, x_i, \ldots x_n$, $\mathrm{var}\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} \mathrm{var}\left(x_i\right)$, and write:

$$
\begin{aligned}
\mathrm{var}(\bar{x}) &= \frac{1}{n^2}\mathrm{var}\left(\sum_{i=1}^{n} x_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}\left(x_i\right) \\
&= \frac{1}{n^2} n \times \mathrm{var}\left(x\right) \\
&= \frac{\mathrm{var}\left(x\right)}{n}
\end{aligned}
$$

Derivation of the standard error of the mean

Now we make use of the fact that for uncorrelated $x_1, \ldots, x_i, \ldots x_n$, $\mathrm{var}\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} \mathrm{var}\left(x_i\right)$, and write:

$$
\begin{aligned}
\mathrm{var}(\bar{x}) &= \frac{1}{n^2}\mathrm{var}\left(\sum_{i=1}^{n} x_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}\left(x_i\right) \\
&= \frac{1}{n^2} n \times \mathrm{var}\left(x\right) \\
&= \frac{\mathrm{var}\left(x\right)}{n} \\
\mathrm{se}(\bar{x}) &= \frac{\mathrm{sd}\left(x\right)}{\sqrt{n}}
\end{aligned}
$$