POLS 205
Political Science as a Social Science

Making Inferences from Samples

Christopher Adolph

University of Washington, Seattle

May 10, 2010

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 true mean, $\bar{x}^{\text{true}}$

$$\mathrm{E}\left(\bar{x} - \bar{x}^{\text{true}}\right) = ?$$

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?

To answer this, we'll need to build up several tools...

Outline

Constructing a Sample

Probability Distributions

Inference about the Population Mean

Inferences about Differences in the Mean

Populations & Samples

We will consider groups of observations at three distinct levels:

Superpopulation All the cases in the world we think our theory applies to
A population of populations
Example: Average support $\breve{\alpha}$ of all Americans over time and space for the income tax

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 α of all Washingtonians in April 2010 for the income tax

Sample The units of analysis actually collected for our study
Usually a subset of the population
Example: Average support $\hat{\alpha}$ of 500 randomly selected Washingtonians in April 2010 for the income tax

Sampling Frames

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

Usually, we must instead make inferences about the population (and superpopulation) 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

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

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

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

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

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

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 next 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 more probability theory

Statistical Independence

We say that two events are independent if the occurence of one doesn't affect the probability that the other occurs

In math, independence implies the conditional probability of an event equals the marginal probability:

$$\Pr(a|b) = \Pr(a)$$

Another way to think of independence is that knowing how the first event turns out doesn't help us predict the second

Statistical Independence

For example, suppose we flip a coin twice.

The second flip doesn't depend on the first:

$$\Pr(\text{Second coin is heads}|\text{First coin is heads}) = \Pr(\text{Second coin is heads})$$

Gambler's Fallacy: "If a coin flip comes out heads many times in a row, the next flip is *more* likely to be heads because 'it's due' to be heads"

In fact, after a dozen straight heads, the probability flip thirteen will be heads is still 1/2.

Probability Distributions

We say a variable is random when there is some probability that it takes on any of the possible values

The mathematical function which relates those probabilities to each value is the probability distribuiton function (pdf)

We can construct many different kinds of pdfs, but it helps to start small

A probability distribution for binary variables

Consider a single flip of a coin. The sample space is $\Omega_{\text{coin flip}} = \{H, T\}$

That is, there is some probability $\Pr(H)$ that we see a head when we flip, and some probability $\Pr(T)$ that we see a tail

Based on probability assumption 1, we know that:

$$0 \leq \Pr(H) \leq 1$$

$$0 \leq \Pr(T) \leq 1$$

A probability distribution for binary variables

If H and T are the only possible outcomes, we know from assumption 2 that:

$$\Pr(H) + \Pr(T) = 1$$

so

$$\Pr(T) = 1 - \Pr(H)$$

That is, if we know $\Pr(H)$, we know everything there is to know about the probability distribution of our coin flip

A probability distribution for binary variables

Let's call the probability of a head $\Pr(H) = \pi$, following the statistics convention the we write all unknown parameters as Greek letters

And let's call our random variable (whether the flip comes out heads or tails) x, following the statistics convention that known data variables are written as Roman letters

We can summarize the probability distribution for a flip of a coin, *or any other binary variable*, in a single equation:

$$f_{\text{Bernoulli}}(x|\pi) = \begin{cases} 1 - \pi & \text{if } x = 0 \\ \pi & \text{if } x = 1 \end{cases}$$

We can summarize the probability distribution for a flip of a coin, *or any other binary variable*, in a single equation:

$$f_{\text{Bernoulli}}(x|\pi) = \begin{cases} 1-\pi & \text{if } x = 0 \\ \pi & \text{if } x = 1 \end{cases}$$

This equation is clear, but unwieldy. Using exponents, we can reduce it to a single line:

$$f_{\text{Bernoulli}}(x|\pi) = \pi^x (1-\pi)^{1-x}$$

This is the pdf of the Bernoulli distribution, which applies to *all* binary variables

The first two moments of this distribution are:

$\mathrm{E}(x) = \pi$ $\qquad$ $\mathrm{var}(x) = \pi(1-\pi)$

What about continuous data?

The Bernoulli distribution is helpful if we are talking about binary data.
In fact, it's the only choice available!

But what about a continuous random variable?
It takes on far more than two possible values

Unfortunately, there are many possible distributions for continuous variables, and choosing one is much more controversial

We will discuss three different choices:
the Normal distribution, the χ^2 distribution, and the t distribution

The Normal Distribution

Suppose we have a large number of additive or ratio level variables with unknown (ie, arbitrary) distributions

These variables need not be related to one another

Indeed, they should be *independent*; ie, uncorrelated with each other

Let us call these variables $x_{1i}, x_{2i}, x_{3i}, \ldots, x_{ki}$

They might be how much each American i spends on each product & service k for sale in the economy

Now suppose we add together the spending of each American to create X_i

The Normal Distribution

According to the Central Limit Theorem, as $k \to \infty$,
X_i will follow the so-called Normal distribution:

$$f_{\text{Normal}}(X|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left[\frac{-(X_i - \mu)^2}{2\sigma^2}\right]$$

The Normal Distribution

According to the Central Limit Theorem, as $k \to \infty$,
X_i will follow the so-called Normal distribution:

$$f_{\text{Normal}}(X|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left[\frac{-(X_i - \mu)^2}{2\sigma^2}\right]$$

Moments: $\mathrm{E}(X) = \mu \quad \mathrm{Var}(X) = \sigma^2$

The Normal Distribution

According to the Central Limit Theorem, as $k \to \infty$,
X_i will follow the so-called Normal distribution:

$$f_{\text{Normal}}(X|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left[\frac{-(X_i - \mu)^2}{2\sigma^2}\right]$$

Moments: $\mathrm{E}(X) = \mu \quad \mathrm{Var}(X) = \sigma^2$

The Normal distribution is continuous and symmetric,
with positive probability everywhere from $-\infty$ to ∞

The Normal Distribution

According to the Central Limit Theorem, as $k \to \infty$,
X_i will follow the so-called Normal distribution:

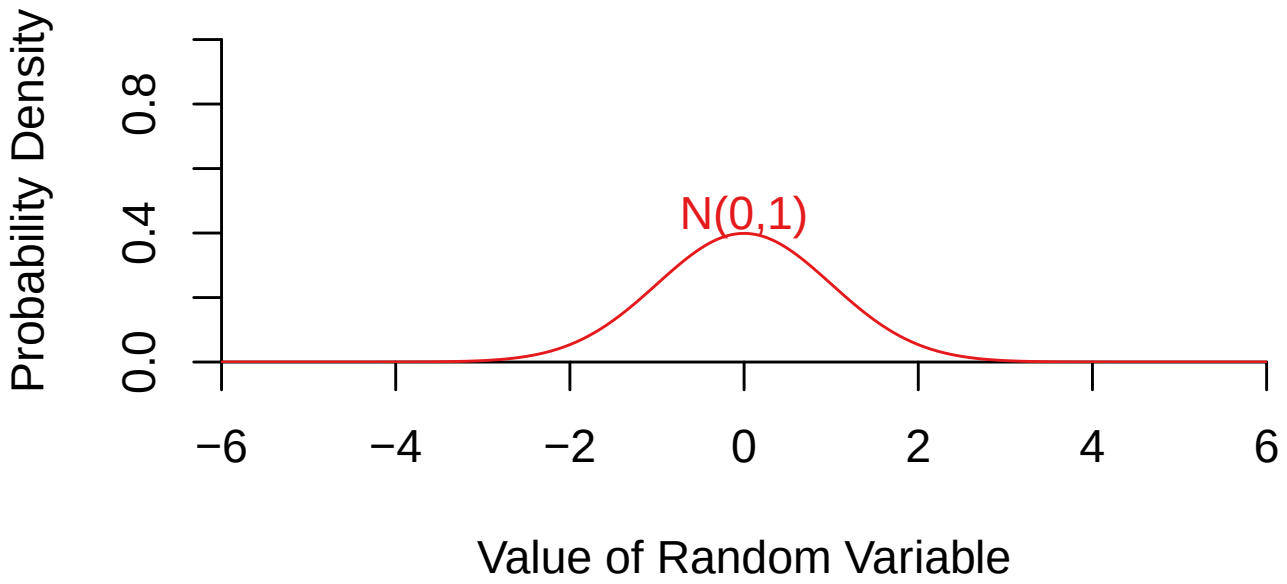
$$f_{\text{Normal}}(X|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left[\frac{-(X_i - \mu)^2}{2\sigma^2}\right]$$

Moments: $\mathrm{E}(X) = \mu \quad \mathrm{Var}(X) = \sigma^2$

The Normal distribution is continuous and symmetric,
with positive probability everywhere from $-\infty$ to ∞

Also called the Gaussian distribution. (A better name, since it avoids the implication that it is "Normal" for a variable to follow this distribution.)

Examples of the Normal Distribution

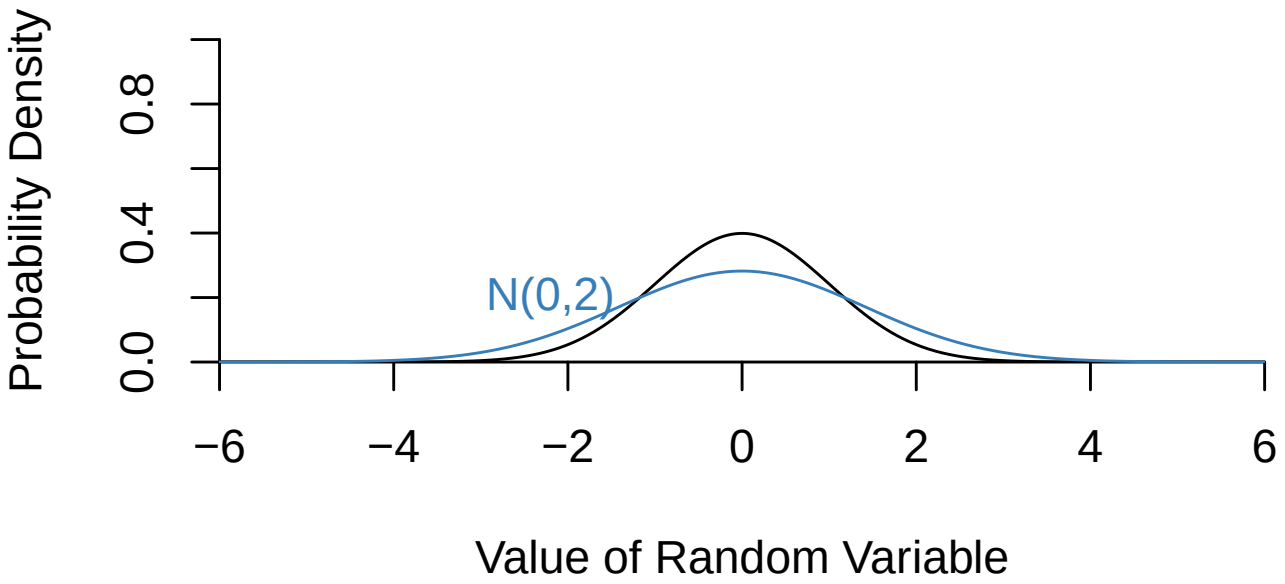


This is the Normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$.

Known as the Standard Normal. Also the Bell Curve.

67% of the density is within ± 1 sd's of the mean;
95% in ± 2 sd's; and 99% in ± 3 sd's.

Examples of the Normal Distribution

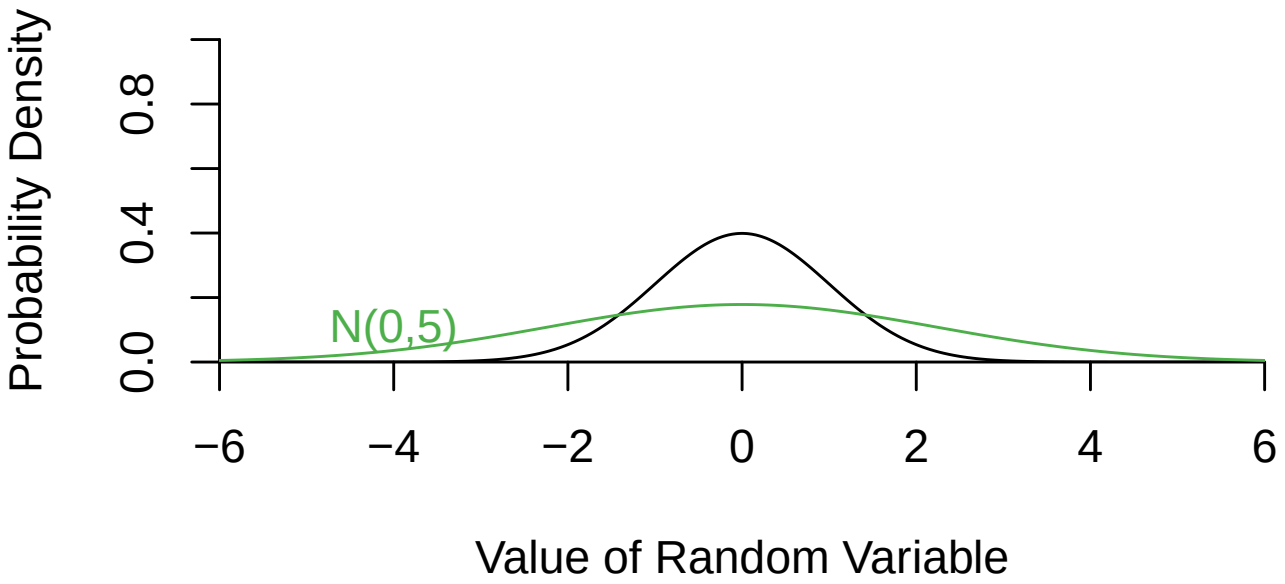


This is the Normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 2$.

The larger variance has spread out the distribution.

Still the case that: 67% of the density is within ± 1 sd's of the mean; 95% in ± 2 sd's; and 99% in ± 3 sd's.

Examples of the Normal Distribution

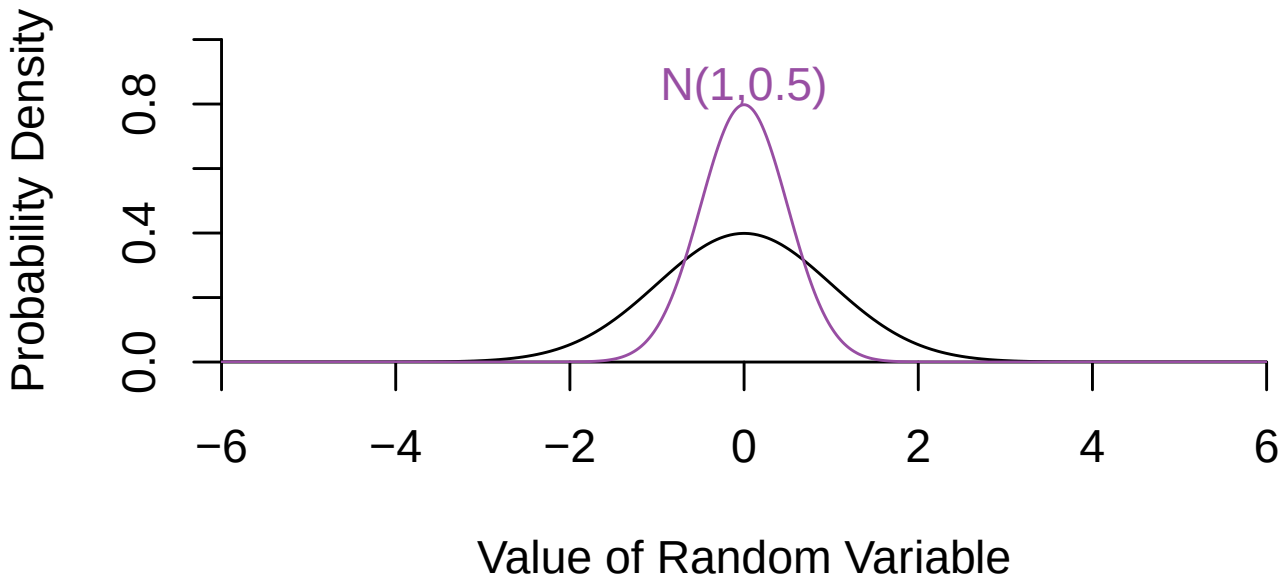


This is the Normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 5$.

The larger variance has spread out the distribution even more

Still the case that: 67% of the density is within ± 1 sd's of the mean; 95% in ± 2 sd's; and 99% in ± 3 sd's.

Examples of the Normal Distribution

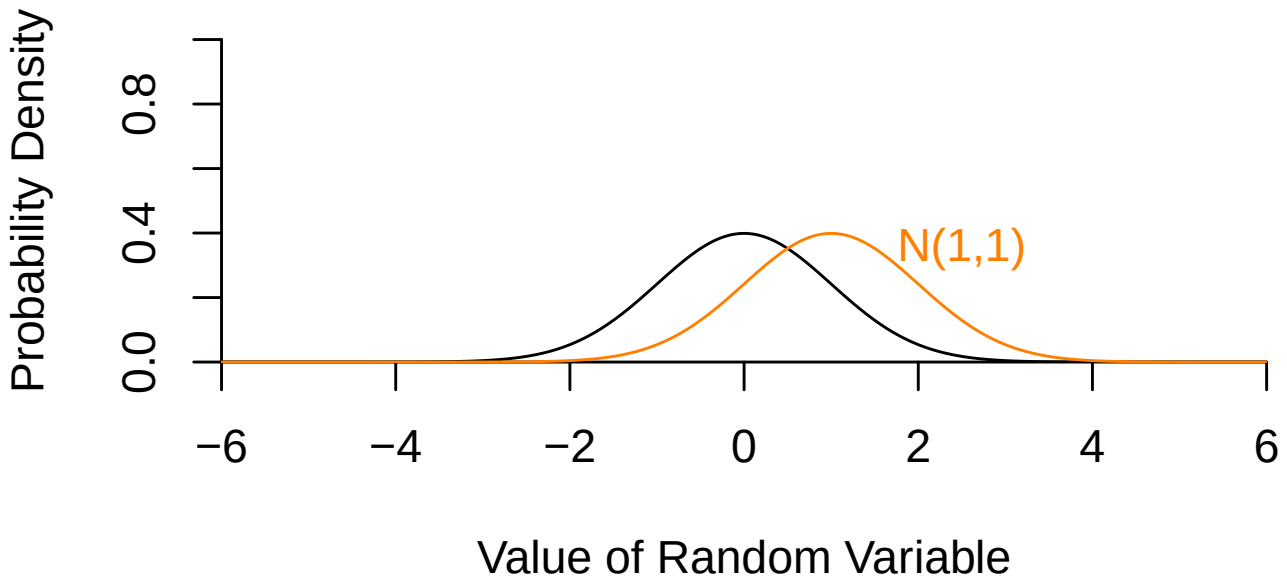


This is the Normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 0.25$.

Smaller variance tightens distribution to a spike over the mean

Still the case that: 67% of the density is within ± 1 sd's of the mean; 95% in ± 2 sd's; and 99% in ± 3 sd's.

Examples of the Normal Distribution



This is the Normal distribution with mean $\mu = 1$ and variance $\sigma^2 = 1$.

Increasing the mean just shifts the distribution rightward

Still the case that: 67% of the density is within ± 1 sd's of the mean; 95% in ± 2 sd's; and 99% in ± 3 sd's.

The χ^2 distribution

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

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

Sum of a finite set of Normal random variables, so the Normal only applies approximately

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,

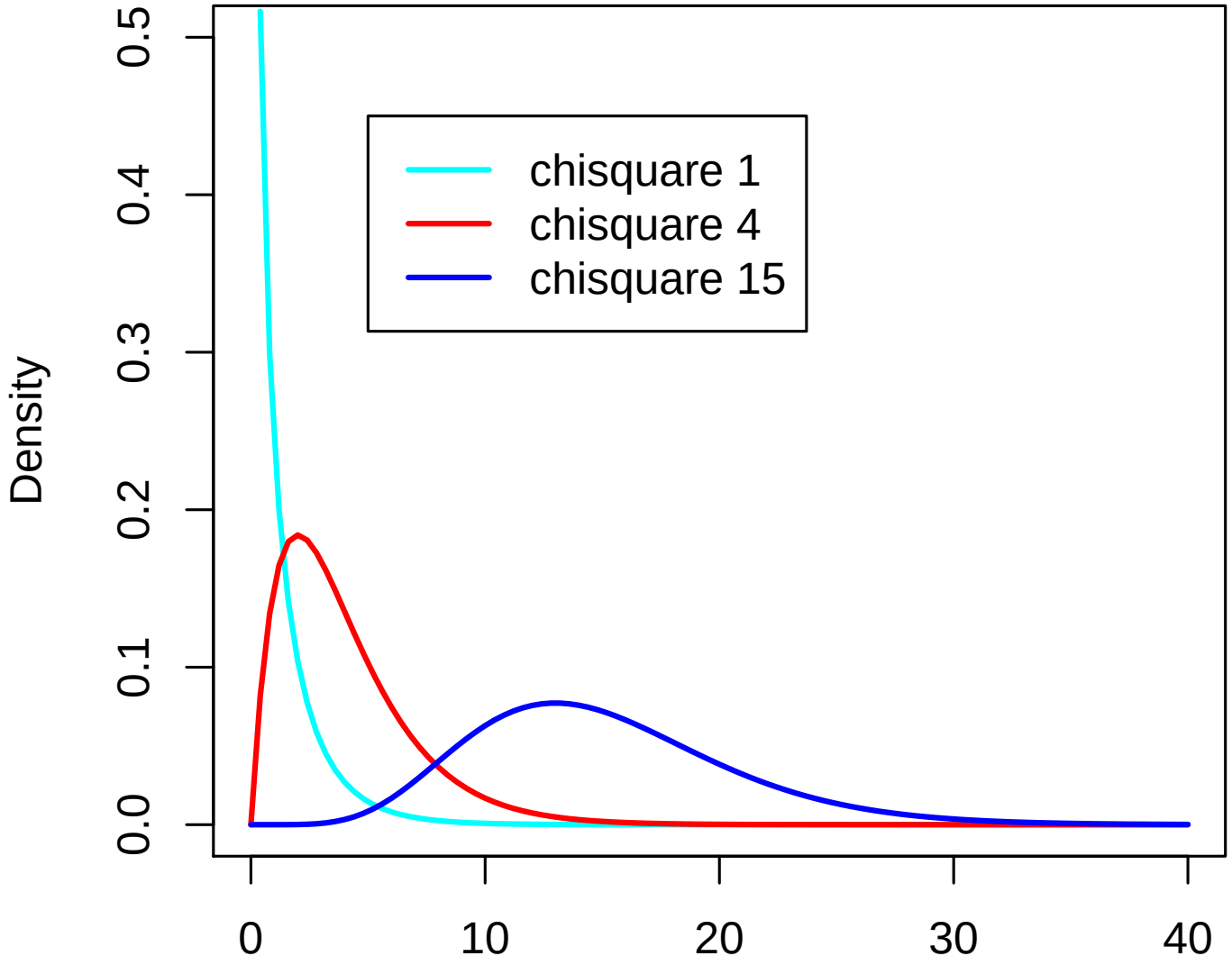
$$\chi^2(X_n^2) = \frac{1}{2^{n/2}\Gamma(n/2)}(X^2)^{(n-2)/2}\exp(-X/2)$$

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 a more useful distribution

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

Define

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

which is distributed t with n degrees of freedom:

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

The t distribution

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

Moments:

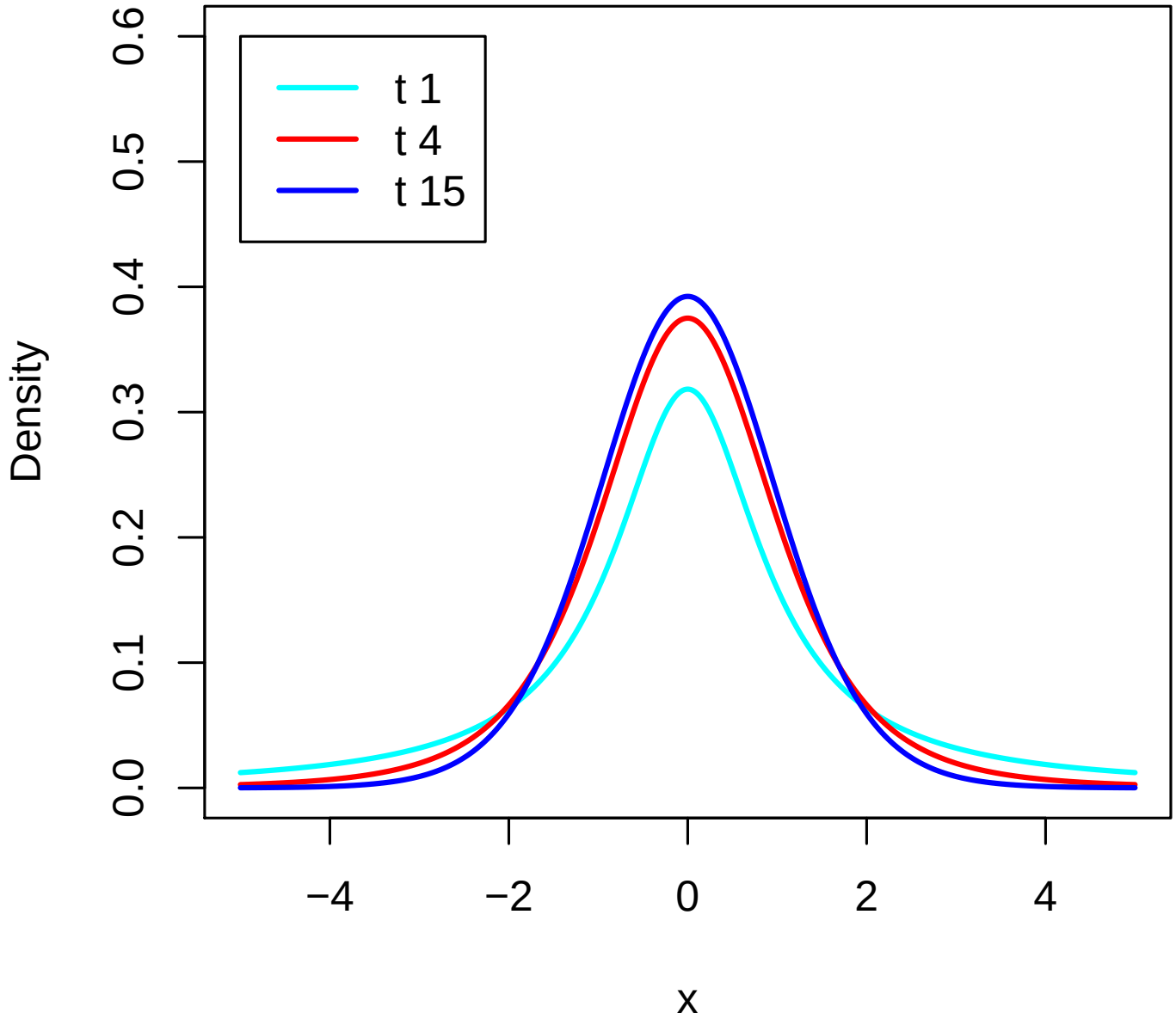
$\mathrm{E}(t) = 0$ (we could change this)

$\mathrm{Var}(t) = n/(n-2)$ 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, $\mathrm{P}(t) = f_t(5)$

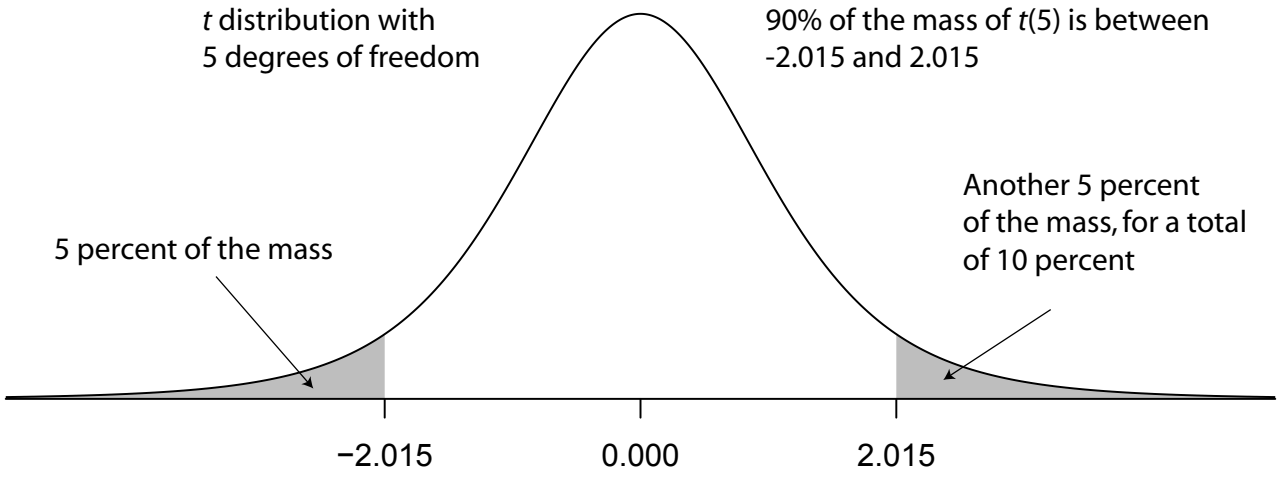
How large would t need to be for us to doubt it came from this distribution?

Put another way, what are the "critical" values of t we would see just

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

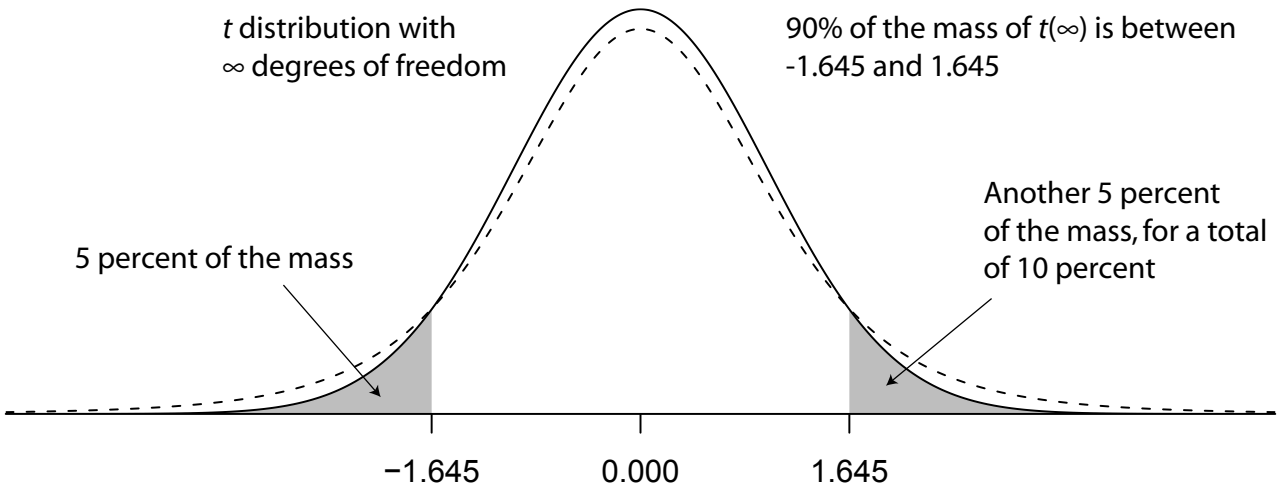
Put still another way,
which critical values will bound the 90% (or 95%, or 99%) most ordinary t draws?

Areas under the t



A unusual value is one in the tails. Critical values = cutoff for “unusualness”

Areas under the t

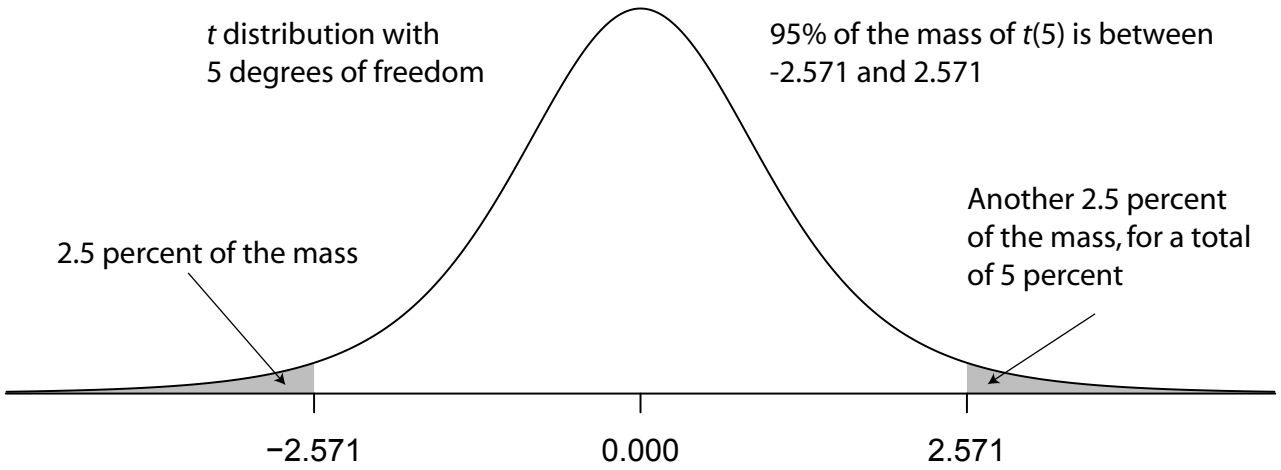


The degrees of freedom reflect how much information we have

More information makes the tails thinner

Critical values shrink; estimates get more certain

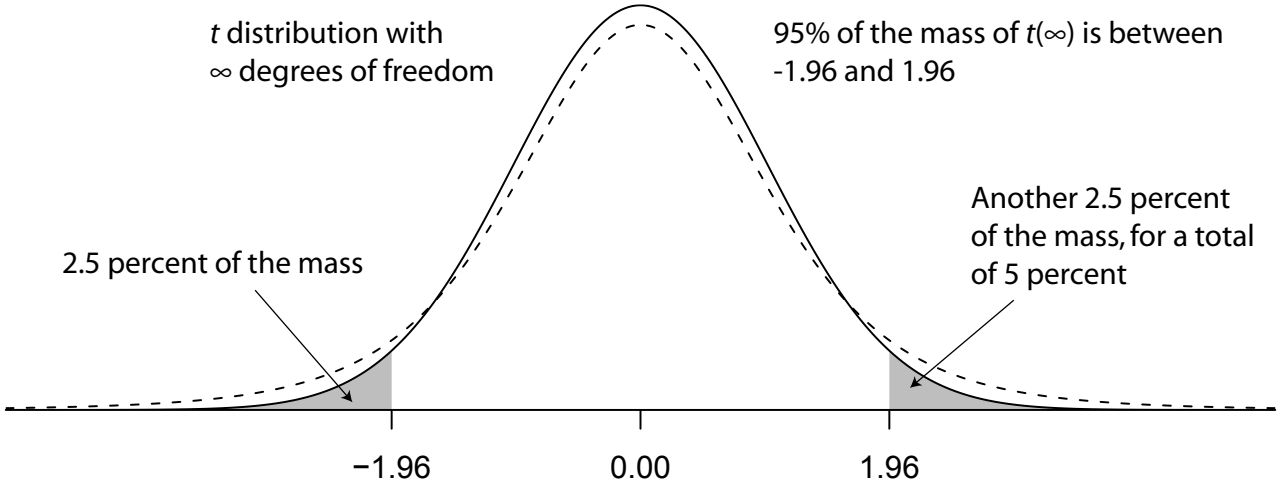
Areas under the t



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

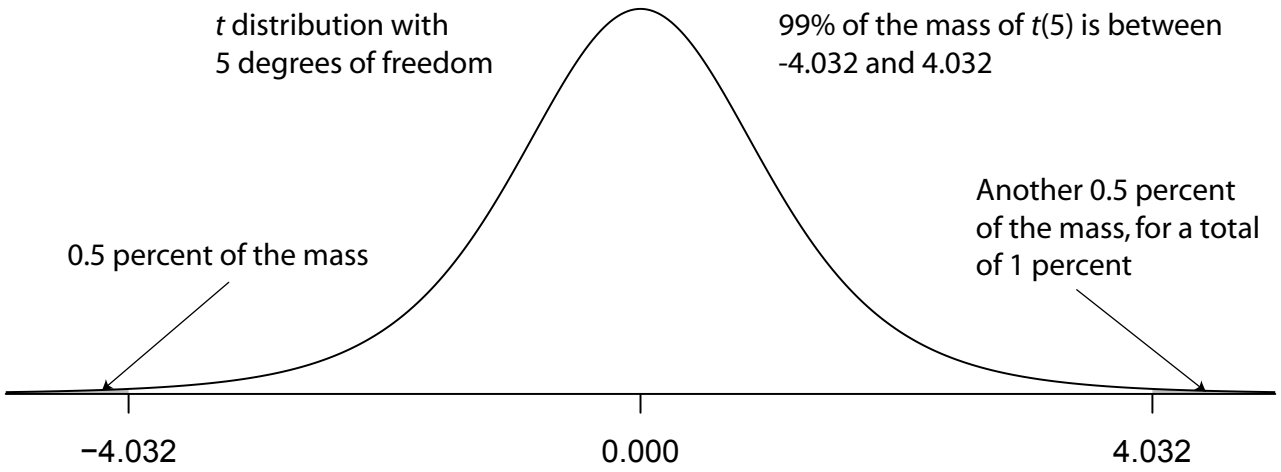
Areas under the t



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

This is the most widely used standard for whether a result is unusual

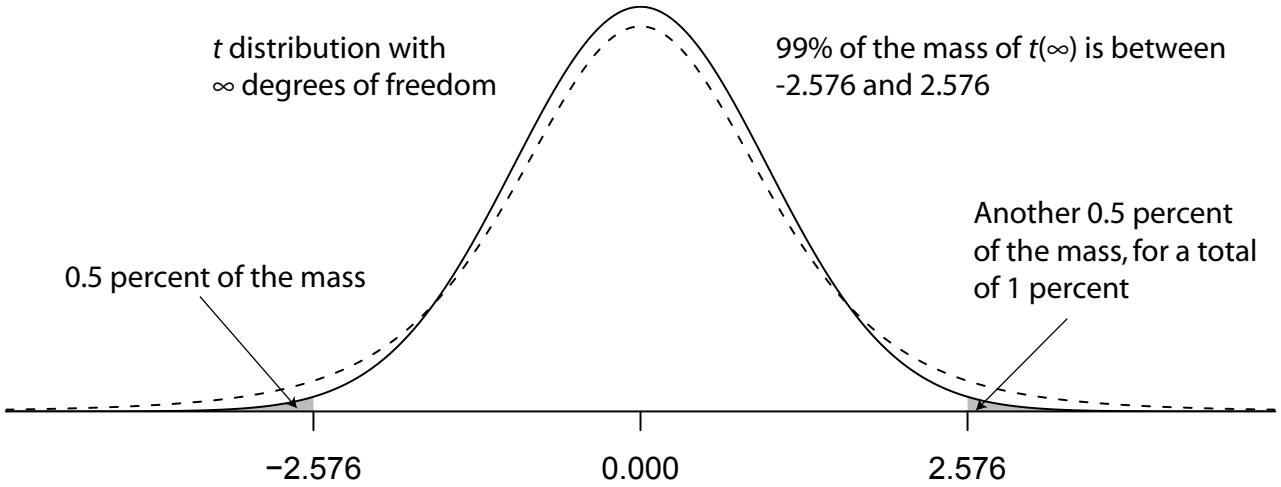
Areas under the t



The most stringent 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%

Critical values of the t distribution

We can state how unusual an observation is under the assumption that it is distributed $t(n)$

Test level	$\mathrm{df} = 5$	$\mathrm{df} = \infty$
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 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 *error* (expected deviation from the truth):

$$\mathrm{E}\left((\mathrm{Estimate} - \mathrm{Truth})^2\right)$$

$$\mathrm{E}\left((\mathrm{Estimate} - \mathrm{E}(\mathrm{Estimate}))^2\right)$$

$$\mathrm{var}(\mathrm{Estimate})$$

We have a special name for the square root of the variance of an error

We call it this special standard deviation the standard error of the estimate, or $\mathrm{se}(\mathrm{Estimate})$

The Law of Large Numbers

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

When our estimate of the mean, $\bar{x}$ gets closer to the truth, its standard error, $\mathrm{se}(\bar{x})$ gets smaller

To see this, we need to derive $\mathrm{se}(\bar{x})$, which means we need to first derive $\mathrm{var}(\bar{x})$

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

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

Derivation of the standard error of the mean

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

Derivation of the standard error of the mean

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

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

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:

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

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:

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

The Square Root Law

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

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 500 dollars of wealth with 100 observations,
to reduce our expected error to 250 dollars, we need 400 total observations

The t-statistic

The t statistic of an estimate is the estimate, minus a hypothetical level, divided by the standard error of the estimate

For the mean, $\bar{x}$, this is

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

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

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

The t-statistic

Note that the t-statistic should be t distributed!

1. $\bar{x}$: The mean of x_i is the sum of a large number of independent variables, and thus will tend to be Normally distributed, by the Central Limit Theorem
2. σ^2: The variance of x_i 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

That is, given $\bar{x} - \mu_0$ as large as the one we saw, uncertainty of that estimate $\sigma/\sqrt{n}$, how likely is it that the population mean of x is actual μ_0 or smaller?

Large t could occur for one of two reasons:

1. A unusual random sample far from the true population mean (which is 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

But we can calculate how often 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, how much information we have, $n - 1$)

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

A statistical significance test checks whether the p-value associated with a t-test is below this level, usually 0.05

Significance tests are tests against a specific null hypothesis, and a "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 II error at the expense of allowing for more Type I 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 II error at the expense of allowing for more Type I 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

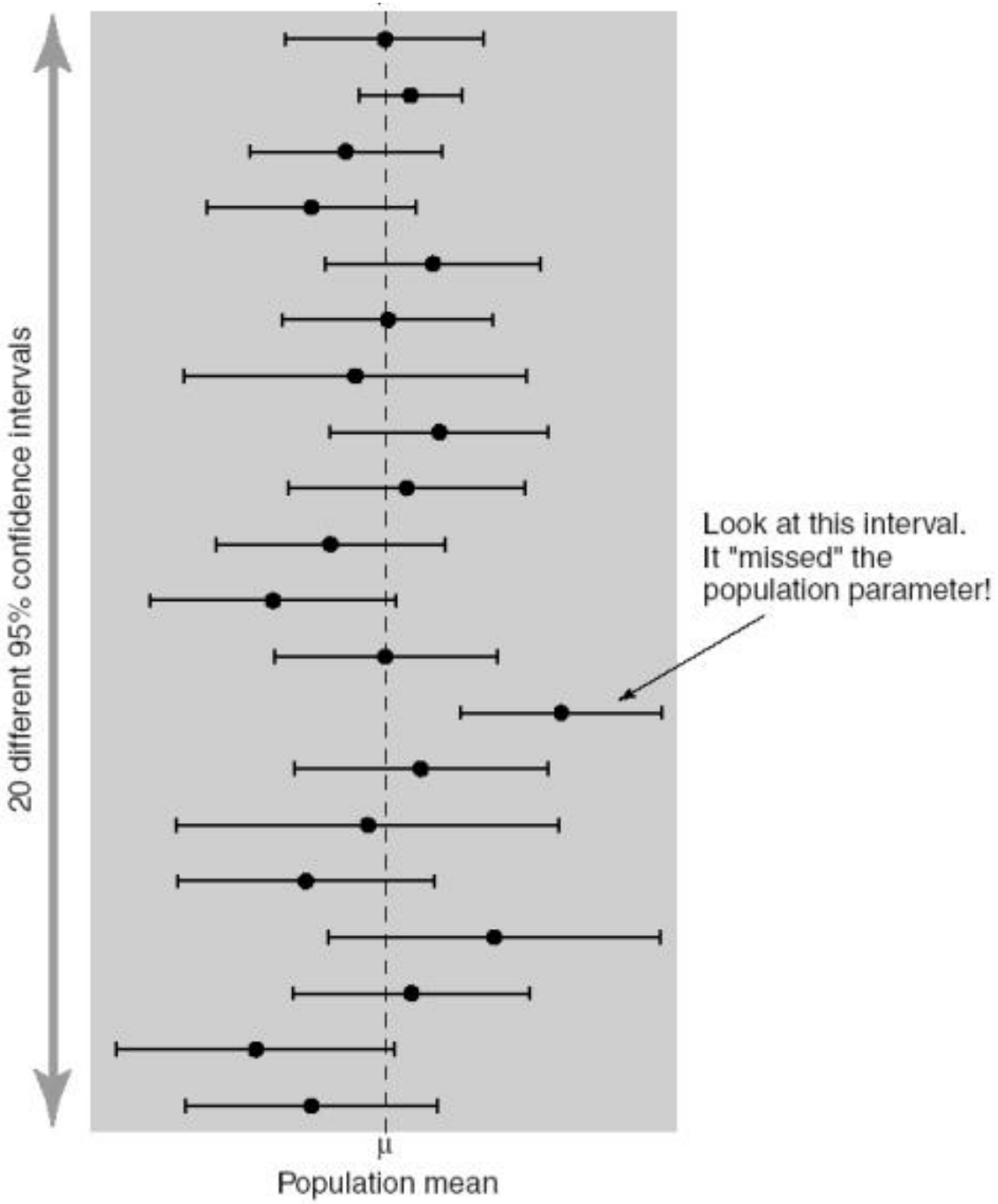
Confidence intervals

An 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:

$$\begin{aligned} \bar{x}^{\text{lower}} &= \bar{x} - t_{n-1}\hat{\sigma}_x \\ \bar{x}^{\text{upper}} &= \bar{x} + t_{n-1}\hat{\sigma}_x \end{aligned}$$

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

Example: Washington State Income Tax

Bill Gates Sr. has proposed a state income tax for the November ballot

On April 21, 2010, SurveyUSA sampled 500 Washington adults in order to estimate the statewide support, asking the following:

"A proposed initiative would create an income tax in Washington state on people making $200,000 per year and on couples making twice that. It would also cut the state's portion of the property tax by 20%, and end the business and occupation tax for small businesses. Do you support or do you oppose this proposed initiative?"

SurveyUSA found 66 percent supported the measure.

How certain are we that the referendum would pass if it were held today?

Example: Washington State Income Tax

How likely is it that a survey of 500 random individuals from a population would find 66% 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 State Income Tax

How likely is it that a survey of 500 random individuals from a population would find 66% support for a measure when really only 50% or less support the measure

Let's use a *t*-test:

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

Example: Washington State Income Tax

How likely is it that a survey of 500 random individuals from a population would find 66% support for a measure when really only 50% or less support the measure

Let's use a *t*-test:

$$
\begin{aligned}
t &= \frac{\bar{x} - \mu_0}{\mathrm{se}(\bar{x})} \\
&= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \\
&= \frac{0.66 - 0.5}{0.474/\sqrt{500}}
\end{aligned}
$$

Example: Washington State Income Tax

How likely is it that a survey of 500 random individuals from a population would find 66% support for a measure when really only 50% or less support the measure

Let's use a *t*-test:

$$
\begin{aligned}
t &= \frac{\bar{x} - \mu_0}{\mathrm{se}(\bar{x})} \\
&= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \\
&= \frac{0.66 - 0.5}{0.474/\sqrt{500}} \\
&= 7.545
\end{aligned}
$$

Example: Washington State Income Tax

How likely is it that a survey of 500 random individuals from a population would find 66% support for a measure when really only 50% or less support the measure

Let's use a t-test:

$$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{\mathrm{se}(\bar{x})} \\ &= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \\ &= \frac{0.66 - 0.5}{0.474/\sqrt{500}} \\ &= 7.545 \end{aligned}$$

A t this big would appear by chance only 1 in 4,620,000,000,000 random samples, (1 in 4.6 trillion), for a $p = 0.000000000000216$

Example: Washington State Income Tax

A t this big would appear by chance only 1 in 4,620,000,000,000 random samples, (1 in 4.6 trillion), for a $p = 0.000000000000216$

Example: Washington State Income Tax

A t this big would appear by chance only 1 in 4,620,000,000,000 random samples, (1 in 4.6 trillion), for a $p = 0.000000000000216$

Why is this so unlikely? Suppose that on April 21, a bare majority of Washington adults really did oppose the income tax.

Example: Washington State Income Tax

A t this big would appear by chance only 1 in 4,620,000,000,000 random samples, (1 in 4.6 trillion), for a $p = 0.000000000000216$

Why is this so unlikely? Suppose that on April 21, a bare majority of Washington adults really did oppose the income tax.

Then to get 66% approval, instead of the correct 50% approval, SurveyUSA would have need to sample $500 \times (0.66 - 0.50) = 80$ more supporters than we would expect on average in 500 random draws.

Example: Washington State Income Tax

A t this big would appear by chance only 1 in 4,620,000,000,000 random samples, (1 in 4.6 trillion), for a $p = 0.000000000000216$

Why is this so unlikely? Suppose that on April 21, a bare majority of Washington adults really did oppose the income tax.

Then to get 66% approval, instead of the correct 50% approval, SurveyUSA would have need to sample $500 \times (0.66 - 0.50) = 80$ more supporters than we would expect on average in 500 random draws.

That's as unlikely as flipping a coin 500 times and getting 330 head and 170 tails.

Example: Washington State Income Tax

Another way to summarize the uncertainty in our polling results is to calculate the confidence interval

We can also state with 95% confidence that the actual level of support for the income tax among all Washington adults is between 61.8% and 70.2%

Notice these numbers are 66 ± 4.2, which also happens to be the reported "margin of error" for the poll (what journalists call a confidence interval).

"Margin of error" is misnamed:
errors can be bigger than this, & are guaranteed to be 5% of the time!

Example: Washington State Income Tax

SurveyUSA's sample of Washington voters includes 120 Republicans, 5 7% percent of whom supported the income tax (!)

Is this result certain?

Judging by the published "margin of error", we might think so:
$57\% - 4.2\% = 52.8\%$, still a majority of Republicans.

Example: Washington State Income Tax

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

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

Example: Washington State Income Tax

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

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

Example: Washington State Income Tax

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

$$
\begin{aligned}
t &= \frac{\bar{x} - \mu_0}{\mathrm{se}(\bar{x})} \\
&= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \\
&= \frac{0.57 - 0.5}{0.497/\sqrt{120}}
\end{aligned}
$$

Example: Washington State Income Tax

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

$$
\begin{aligned}
t &= \frac{\bar{x} - \mu_0}{\mathrm{se}(\bar{x})} \\
&= \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \\
&= \frac{0.57 - 0.5}{0.497/\sqrt{120}} \\
&= 1.468
\end{aligned}
$$

This is a pretty small t-statistic, one we would see by chance in 1 out of 7 random samples. The p-value is 0.150.

We find that the 95% confidence interval ranges from 48% to 66%, which is equal to our estimate of 57% by $\pm$9%.

We are not at all certain that Washington Republicans support the income tax.

Example: Washington State Income Tax

Why is this different from our last example?

Example: Washington State Income Tax

Why is this different from our last example?

Two reasons:

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 a bare majority of Washington Republicans actually oppose the income tax.

Change in Size of Sample

Suppose a bare majority of Washington Republicans actually oppose the income tax.

Then, to get 57% of Republicans in favor in a sample of 120, SurveyUSA would need to have randomly sampled $120 \times (0.57 - 0.50) = 8$ more Republicans in favor than they would expect to on average

Change in Size of Sample

Suppose a bare majority of Washington Republicans actually oppose the income tax.

Then, to get 57% of Republicans in favor in a sample of 120, SurveyUSA would need to have randomly sampled $120 \times (0.57 - 0.50) = 8$ more Republicans in favor than they would expect to on average

This is exactly the same as flipping a coin 120 times and getting 68 heads and 52 tails. Unlikely, but not that unlikely.

Change in Size of Sample

Suppose a bare majority of Washington Republicans actually oppose the income tax.

Then, to get 57% of Republicans in favor in a sample of 120, SurveyUSA would need to have randomly sampled $120 \times (0.57 - 0.50) = 8$ more Republicans in favor than they would expect to on average

This is exactly the same as flipping a coin 120 times and getting 68 heads and 52 tails. Unlikely, but not that unlikely.

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

Change in Size of Sample

Suppose a bare majority of Washington Republicans actually oppose the income tax.

Then, to get 57% of Republicans in favor in a sample of 120, SurveyUSA would need to have randomly sampled $120 \times (0.57 - 0.50) = 8$ more Republicans in favor than they would expect to on average

This is exactly the same as flipping a coin 120 times and getting 68 heads and 52 tails. Unlikely, but not that 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 thw whole sample

Change in Size of Sample

Suppose a bare majority of Washington Republicans actually oppose the income tax.

Then, to get 57% of Republicans in favor in a sample of 120, SurveyUSA would need to have randomly sampled $120 \times (0.57 - 0.50) = 8$ more Republicans in favor than they would expect to on average

This is exactly the same as flipping a coin 120 times and getting 68 heads and 52 tails. Unlikely, but not that 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 thw whole sample

The smaller the *n*, the bigger the confidence interval, the less certain the finding

Change in Variance of the Sample

The t-statistic gets bigger the smaller the variance

Change in Variance of the Sample

The t-statistic gets bigger the smaller the variance

Is the variance for our Republican sample smaller or larger than the whole sample variance?

Change in Variance of the Sample

The t-statistic gets bigger the smaller the variance

Is the variance for our Republican sample smaller or larger than the whole sample variance?

Note that our outcome is a binary variable

Change in Variance of the Sample

The t-statistic gets bigger the smaller the variance

Is the variance for our Republican sample smaller or larger than the whole sample variance?

Note that our outcome is a binary variable

Recall the variance of a binary variable is always

$$\mathrm{var}(x) = \pi(1 - \pi) = \pi - \pi^2$$

Change in Variance of the Sample

The t-statistic gets bigger the smaller the variance

Is the variance for our Republican sample smaller or larger than the whole sample variance?

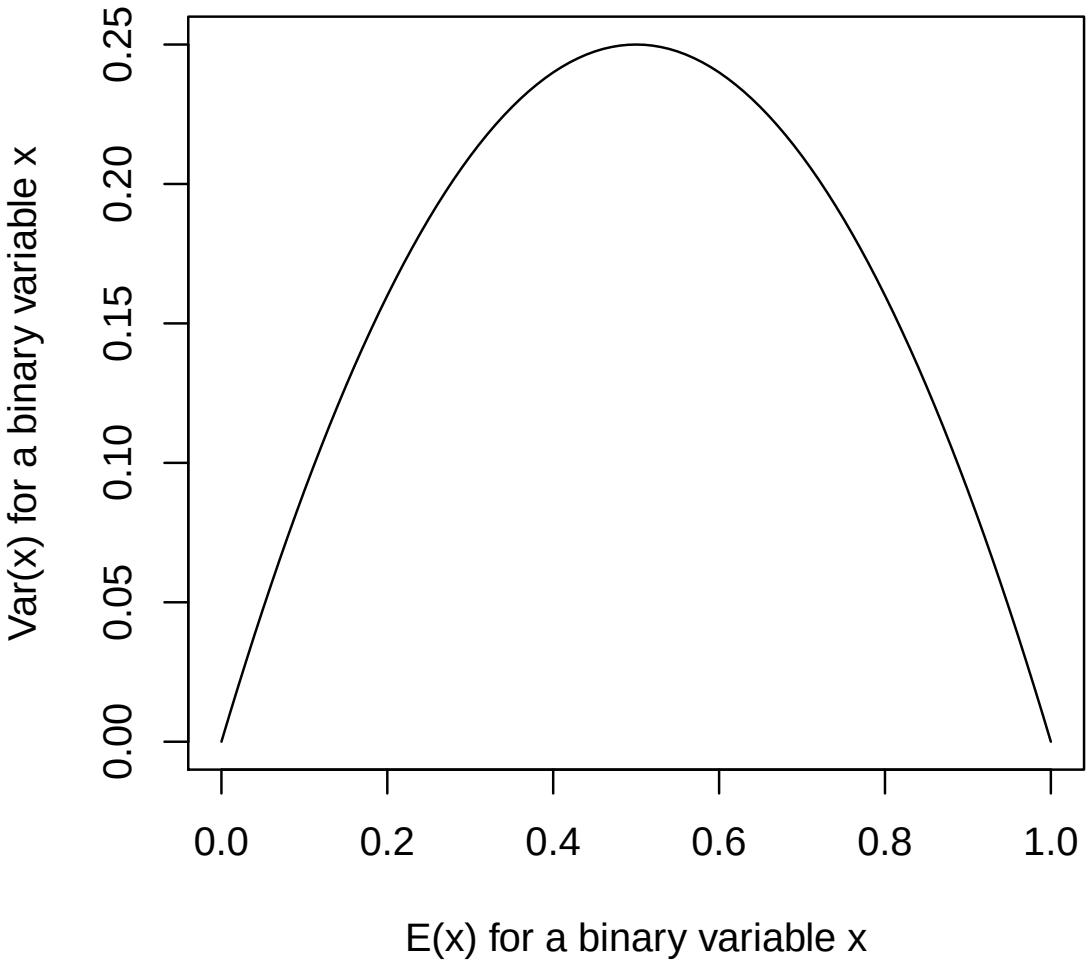
Note that our outcome is a binary variable

Recall the variance of a binary variable is always

$$\mathrm{var}(x) = \pi(1 - \pi) = \pi - \pi^2$$

This is a parabola maximized at 0.5

Change in Variance of the Sample



Change in Variance of the Sample

Thus because the estimated probability a Republican supports the income tax is closer to 0.5 than the probability for all surveyed adults,

The uncertainty of the proportion of Republicans supporting is also greater

Bad news: Not only do “margins of error” reported in the press only apply to the full sample, they also only apply to one specific question!

Good news: With minimal calculation, you can find the right “margin of error” on you own

Worse news: If error is maximized for probabilities ≈ 0.5, what does that mean for predicting election outcomes?

On confidence versus significance

There are two ways we could report our finding on Republicans' support for the income tax:

Significance test Based on a survey of Washington adults, we estimate 57% of Republicans support the income tax. However, this estimate is not statistically significantly different from 50% at the 0.05 level.

Confidence interval Based on a survey of Washington adults, we estimate 57% of Republicans support the income tax. The 95% confidence interval for this estimate ranges from 48% to 66%, suggesting anywhere from a slight majority against the tax 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 Republicans support the tax,

it is very likely that half or more do,

and very unlikely that a large percentage of Republicans are opposed

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

$$t = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

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 an income tax differ from 0.5?

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

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

e.g., how much does the average support for an income tax among women differ from support among men?

t-test for comparison of means

As with a single mean, we will calculate a t-statistic:

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

then check if the t-statistic exceeds the chosen critical value or simply calculate the probability of seeing so large a t

The form of the standard error here is a bit messy:

$$\mathrm{se}(\bar{x} - \bar{y}) = \sqrt{\left(\frac{(n_x - 1)\hat{\sigma}_x^2 + (n_y - 1)\hat{\sigma}_y^2}{n_x + n_y - 2}\right) \times \left(\frac{1}{n_x} + \frac{1}{n_y}\right)}$$

t-test for comparison of means

Unfortunately, the number of degrees of freedom, ν, is now ambiguous, since the samples could be different sizes

An estimate of the dfs for the comparison of means of different-sized samples is:

$$\hat{\nu} = \frac{\left(\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}\right)^2}{\frac{\hat{\sigma}_x^4}{n_x^2(n_x-1)} + \frac{\hat{\sigma}_y^4}{n_y^2(n_y-1)}}$$

(Don't worry, you'll never need to do this by hand)

Example: Washington State Income Tax

The SurveyUSA sample of Washington adults includes 245 men, 62% of whom support the income tax proposal, and 255 women, 70% of whom support

How certain are we that on average, Washington women are actually more likely to support the income tax than men are?

That is, how certain are we that $\mathrm{E}(\mathrm{support}|\mathrm{female}) - \mathrm{E}(\mathrm{support}|\mathrm{male}) > 0$?

We can do a comparison of means *t*-test. (The math for this is too hairy to show here; just let STATA do it)

We find $t = 1.8323$, which implies a *p*-value of 0.068.

That's not significant at the 0.05 level, but only just.

The 95% confidence interval is ranges from -0.5% to 16%. (What does this mean?)

Example: Washington State Income Tax

What if we considered a different level of confidence?

	t-stat	lower	upper
90% level	1.83	0.7%	14.7%
95% level	1.83	-0.6%	16.1%
99% level	1.83	-3.2%	18.7%

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. Probability distributions
2. Standard errors of estimates
3. t-tests and confidence intervals for a sample mean
4. 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 spurrious.