Estimation, Part 1: Sampling Distributions

Estimation

When doing statistics, the goal is usually to *infer* something about a population *parameter* using only a sample from that population.

Estimation

When doing statistics, the goal is usually to *infer* something about a population *parameter* using only a sample from that population.

Examples:

- Estimating the average household income in Putnam County, based on a sample.

Estimation

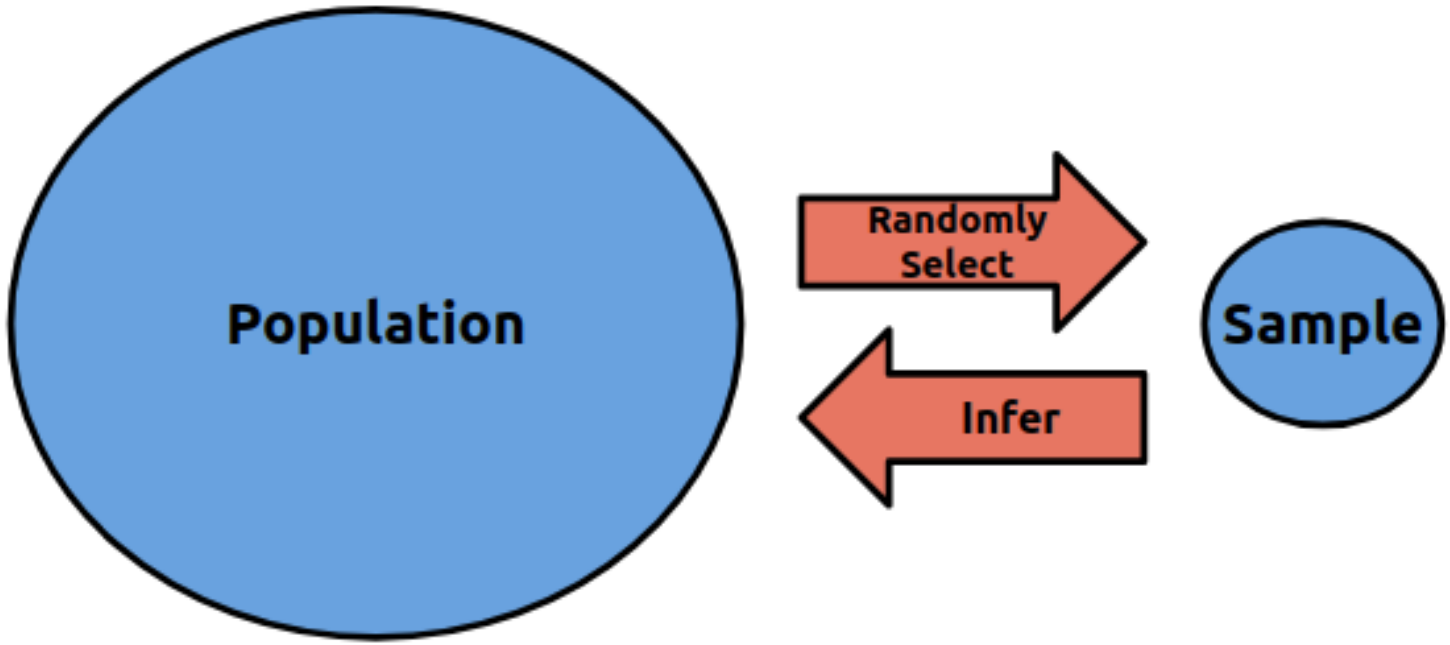
When doing statistics, the goal is usually to *infer* something about a population *parameter* using only a sample from that population.

Examples:

- Estimating the average household income in Putnam County, based on a sample.
- Predicting the percentage of votes a particular candidate will receive in an upcoming election, based on a poll.

Estimation

We are trying to say something about the entire population just by examining a subset.



Estimation

One option that we have is to report a **point estimate**. That is, we can report just a single number as our estimate for the population parameter.

For example, we might survey 100 households in Putnam County and estimate, based on our sample, that the average household income in Putnam County is $43,000.

Estimation

However, if we are only reporting a single number, we are very unlikely to be correct. We would usually be better off reporting our estimate, along with how close it is likely to be to the correct number.

Estimation

However, if we are only reporting a single number, we are very unlikely to be correct. We would usually be better off reporting our estimate, along with how close it is likely to be to the correct number.

To improve our income estimate above, we might say that we are fairly confident that our original estimate of household income being $43,000 is off by no more that $1,500. If we do this, we are indicating that our estimate has **margin of error** $1,500.

Estimation

However, if we are only reporting a single number, we are very unlikely to be correct. We would usually be better off reporting our estimate, along with how close it is likely to be to the correct number.

To improve our income estimate above, we might say that we are fairly confident that our original estimate of household income being $43,000 is off by no more that $1,500. If we do this, we are indicating that our estimate has **margin of error** $1,500.

We can frame our estimate and margin of error in another way, by saying that we are confident that the true mean household income is in the interval

$$\$43,000 \pm \$1,500$$

Estimation

However, if we are only reporting a single number, we are very unlikely to be correct. We would usually be better off reporting our estimate, along with how close it is likely to be to the correct number.

To improve our income estimate above, we might say that we are fairly confident that our original estimate of household income being $43,000 is off by no more that $1,500. If we do this, we are indicating that our estimate has **margin of error** $1,500.

We can frame our estimate and margin of error in another way, by saying that we are confident that the true mean household income is in the interval

$$\$43,000 \pm \$1,500$$

That is, we are highly confident that the true population mean is between $41,500 and $44,500.

Estimation

Another way to write this is that a confidence interval for μ is given by

$$41500 < \mu < 44500$$

Here, μ represents the true average household income of Putnam County.

Confidence Intervals

A **confidence interval** is an interval that is used to estimate the value of a parameter.

Confidence Intervals

A **confidence interval** is an interval that is used to estimate the value of a parameter.

Associated with a confidence interval is a **confidence level**. A confidence level is a percentage between 0% and 100% that measures the success rate of the method used to construct the confidence interval. If we were to draw many samples and use each one to construct a confidence interval, then in the long run, the percentage of confidence intervals that cover the true value would be equal to the confidence level.

Confidence Intervals

A **confidence interval** is an interval that is used to estimate the value of a parameter.

Associated with a confidence interval is a **confidence level**. A confidence level is a percentage between 0% and 100% that measures the success rate of the method used to construct the confidence interval. If we were to draw many samples and use each one to construct a confidence interval, then in the long run, the percentage of confidence intervals that cover the true value would be equal to the confidence level.

To make our household income estimate into a true confidence interval, we need to attach a confidence level. For example, I could say that I am 95% confident that the true mean household income is in the interval \$43,000 $\pm$ \$1,500.

Confidence Intervals

The general formula for a confidence interval is

$$\text{point estimate} \pm \text{margin of error}$$

Confidence Intervals

The general formula for a confidence interval is

$$\text{point estimate} \pm \text{margin of error}$$

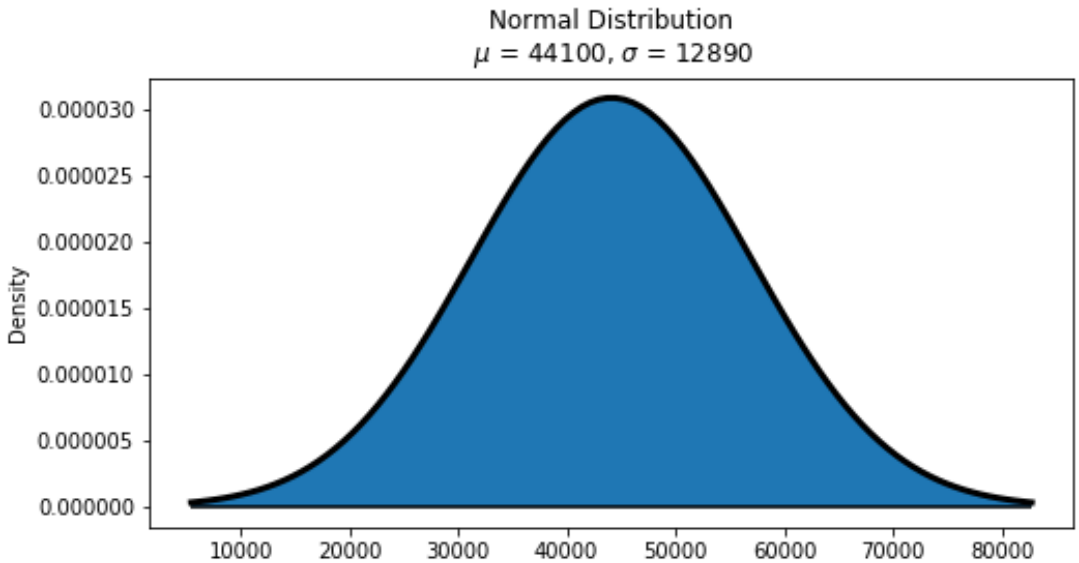
For example, for the Putnam County mean household income, our point estimate was \$43,000 and our margin of error was \$1,500.

Confidence Intervals

How do we find the margin of error?

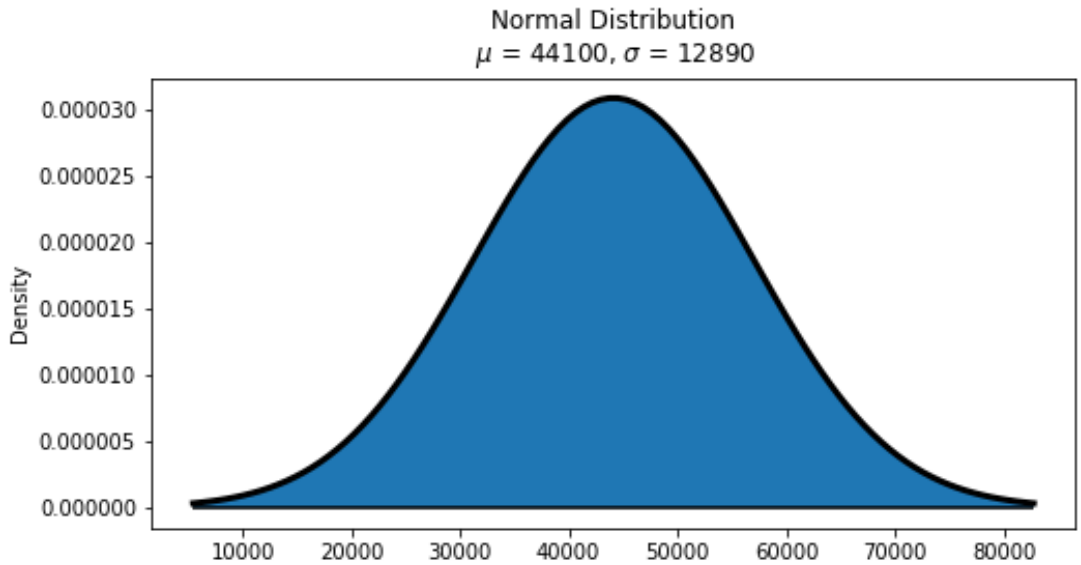
To come up with a general recipe, we have to temporarily pretend like we know something about the entire population of interest.

Let's pretend that in Putnam County, household incomes are distributed normally with a mean of $44,100 and a standard deviation of $12,890. (Note: this is not a good approximation of income distribution, but this is only a thought experiment.)



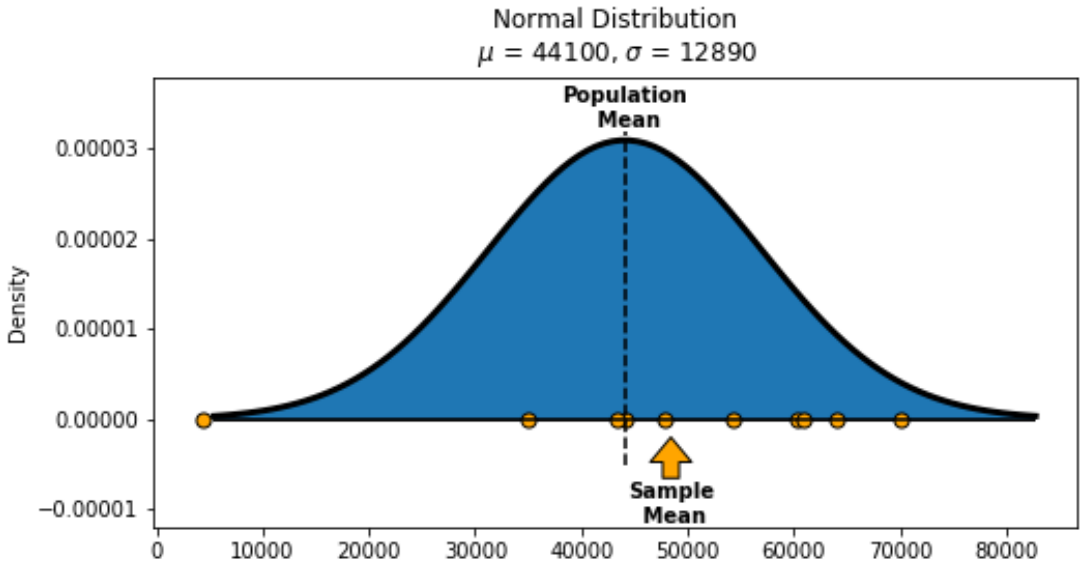
Confidence Intervals

Note that the pdf describes the probability of a *single (random) observation* landing in a particular range.
Let's take a random sample of size 10 and see where it lands on this distribution.



Confidence Intervals

Let's take a random sample of size 10 and see where it lands on this distribution.

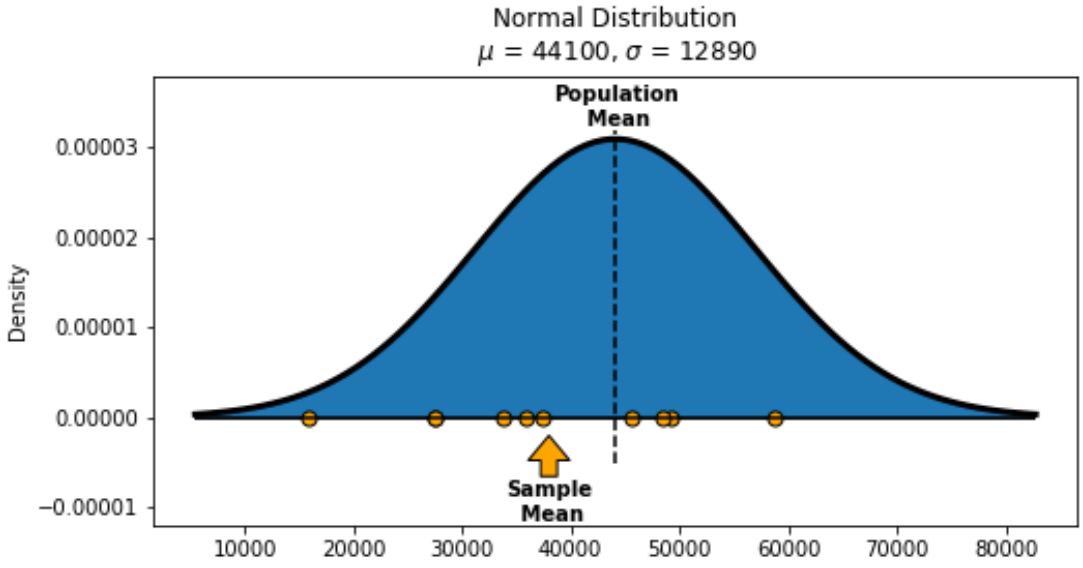


Even though individual elements of the sample can be quite far from the population mean, the sample mean ends up close to (but not exactly equal to) the population mean.
Let's see what happens if we take a different sample of size 10.

NASHVILLE SOFTWARE SCHOOL

Confidence Intervals

Let's take a random sample of size 10 and see where it lands on this distribution.

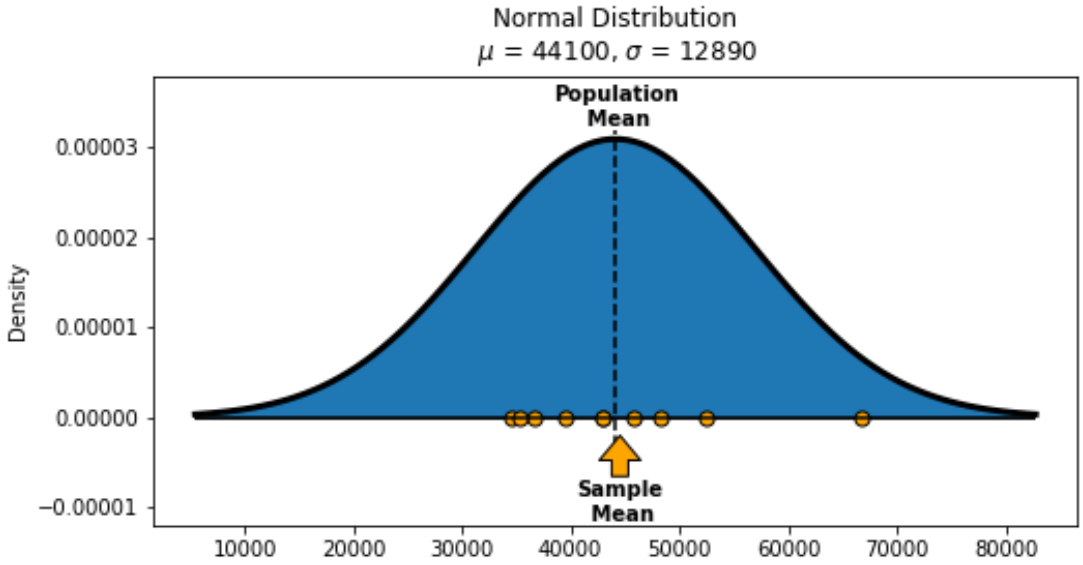


This time the sample mean ends up further away from the population mean.
Let's see what happens if we take a different sample of size 10.

NASHVILLE SOFTWARE SCHOOL

Confidence Intervals

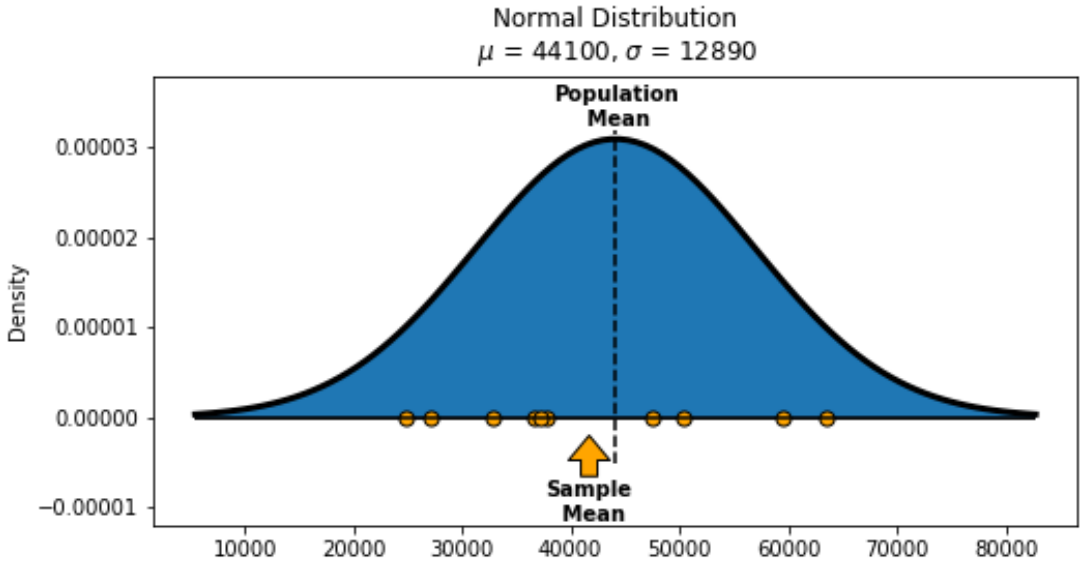
Let's take a random sample of size 10 and see where it lands on this distribution.



This time we end up quite close (but not exactly equal to) the population mean.
Let's see what happens if we take a different sample of size 10.

Confidence Intervals

Let's take a random sample of size 10 and see where it lands on this distribution.



Again, we end up close to the population mean.

Confidence Intervals

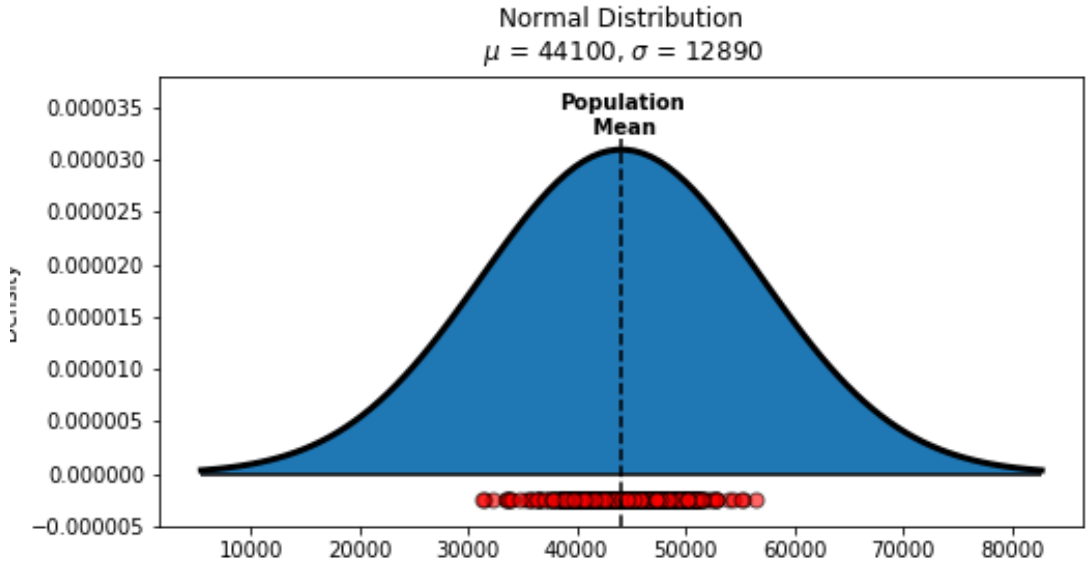
In practice, the difficult thing is that we don't get to know the true population distribution, so we can't see how close our sample is. We really only get a single shot and have to try and determine how close we are likely to be.

From what we've seen, the sample means have been pretty good estimates of the population mean, but there is some variance.

We can treat the sample mean $\bar{x}$ as a random variable and try to understand the distribution of $\bar{x}$.

Confidence Intervals

Let's try and see what this distribution looks like. In the following picture, each red dot corresponds to the sample mean from a sample of size 10. There were 500 different samples picked.

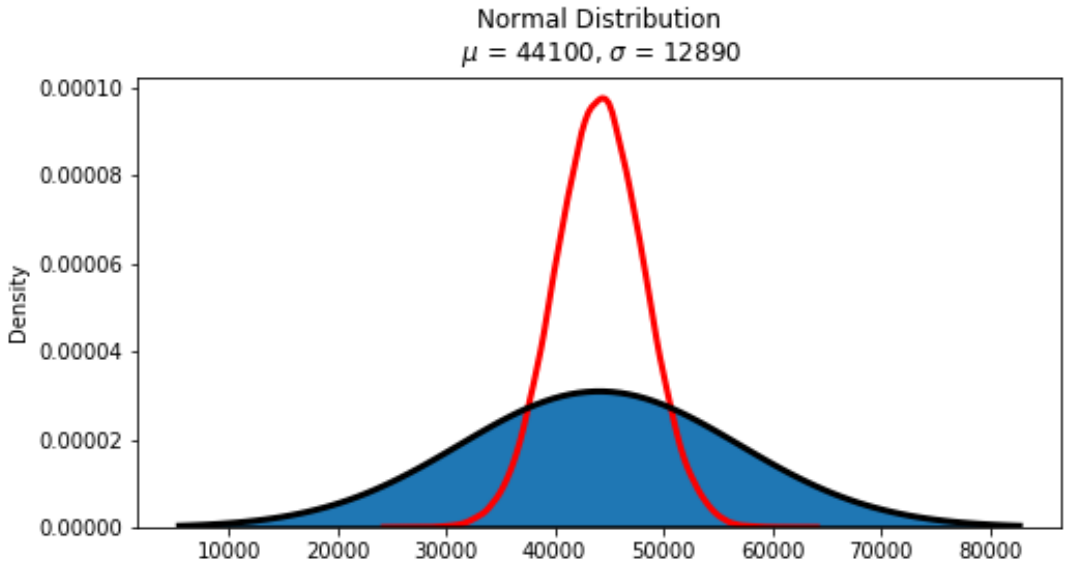


Notice that not a single time out of 500 samples was the sample mean less than 30000 or above 60000.

NASHVILLE SOFTWARE SCHOOL

Confidence Intervals

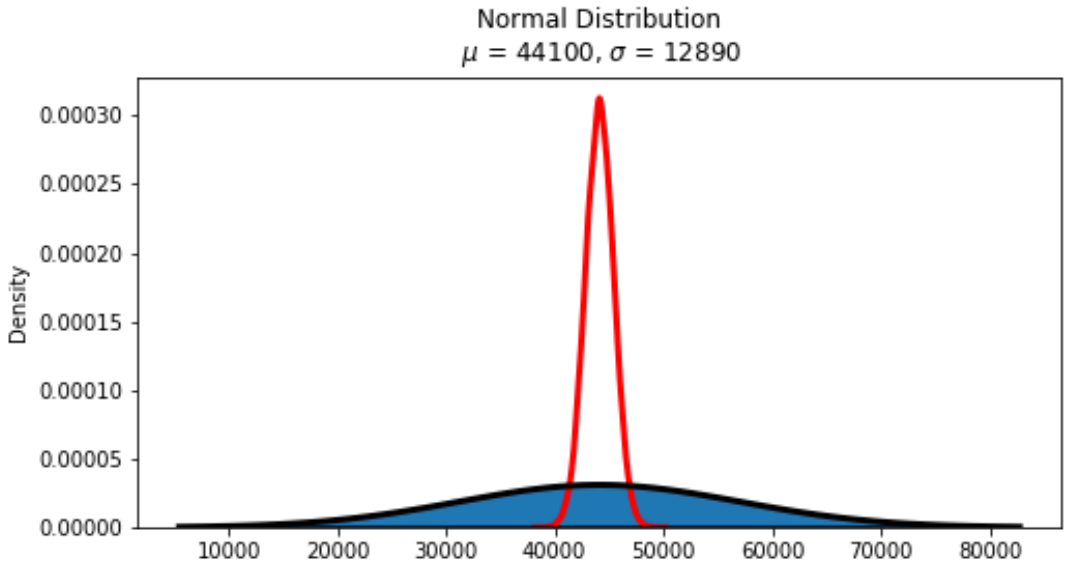
Let's look at a KDE of the distribution of sample means.



What we can see is that the distribution of sample means is centered around the population mean but is much narrower than the population distribution.

Confidence Intervals

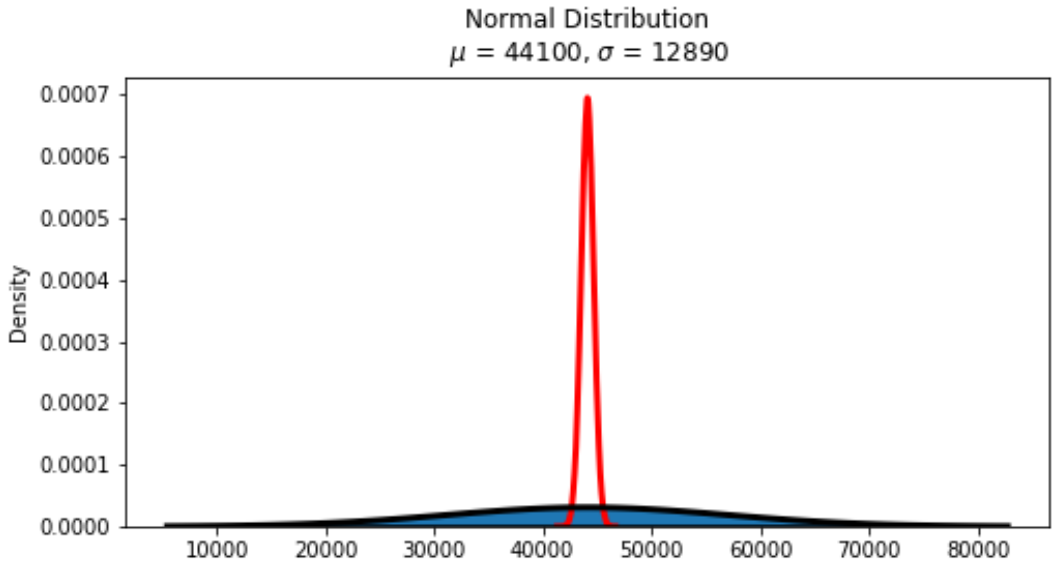
Remember that the distribution we were looking at was for samples of size 10. What if we increase the sample size to 100?



We get an even narrower distribution around the mean.

Confidence Intervals

What if we increase the sample size to 500?



We can see that at this sample size, there is not a lot of variance at all in the sampling distribution.

This can give us a pretty good idea of how far off our estimate is likely to be if we use the sample mean. We can see how we can use this information to build a considence interval using the confidence interval widget.

Confidence Intervals

So the widget shows us that if we know what the sampling distribution of the mean (or whatever parameter we are interested in) looks like, we can use this to construct our confidence interval.

Confidence Intervals

So the widget shows us that if we know what the sampling distribution of the mean (or whatever parameter we are interested in) looks like, we can use this to construct our confidence interval.

Central Limit Theorem: For the random variable $\bar{x}$, the sample mean of a random sample of size n from a given population having mean μ and standard deviation σ, if n is sufficiently large, then the distribution of $\bar{x}$ is approximately normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Thus, $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ approximately follows a standard normal distribution.

Confidence Intervals

There is a problem with using the Central Limit Theorem. In order to use it to construct a confidence interval, we need to know $\frac{\sigma}{\sqrt{n}}$. But if we don't know the population mean μ, why would we know the population standard deviation σ. The best we could do would be to approximate it with the sample standard deviation, s.

Fact: If either the population is normally distributed or we have a large enough sample (usually 30 will do), then

$$\frac{\bar{x} - \mu}{s/\sqrt{n}}$$

follows a Student's t-distribution with $n - 1$ degrees of freedom.

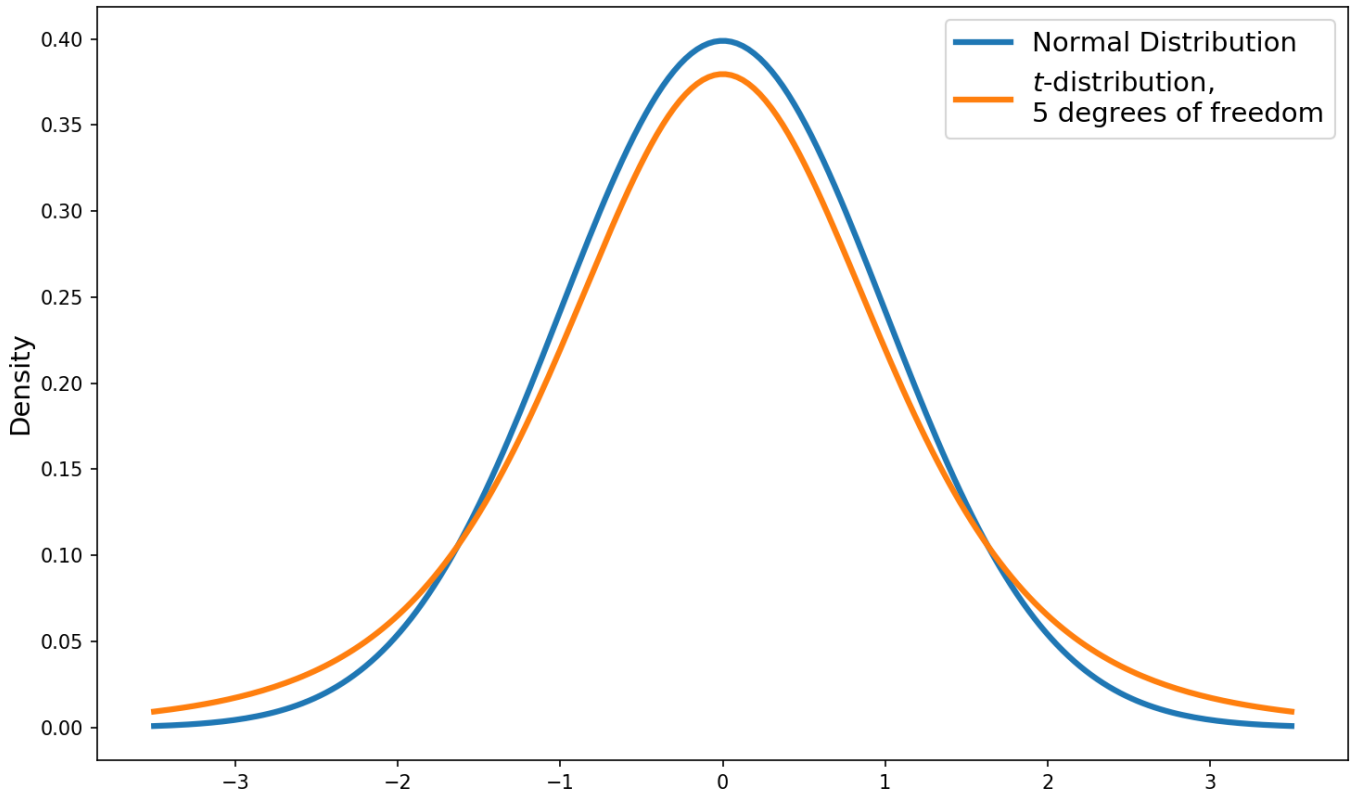
Confidence Intervals

The family of Student's t-distribution is named after statistician William Sealy Gosset, who published his research under the pseudonym "Student".

Gosset worked for the Guinness Brewery where he worked on determining the quality of raw materials. Gosset was interested in the problem of small samples, as he would sometimes have to draw inferences from samples with as few as just 3 observations.

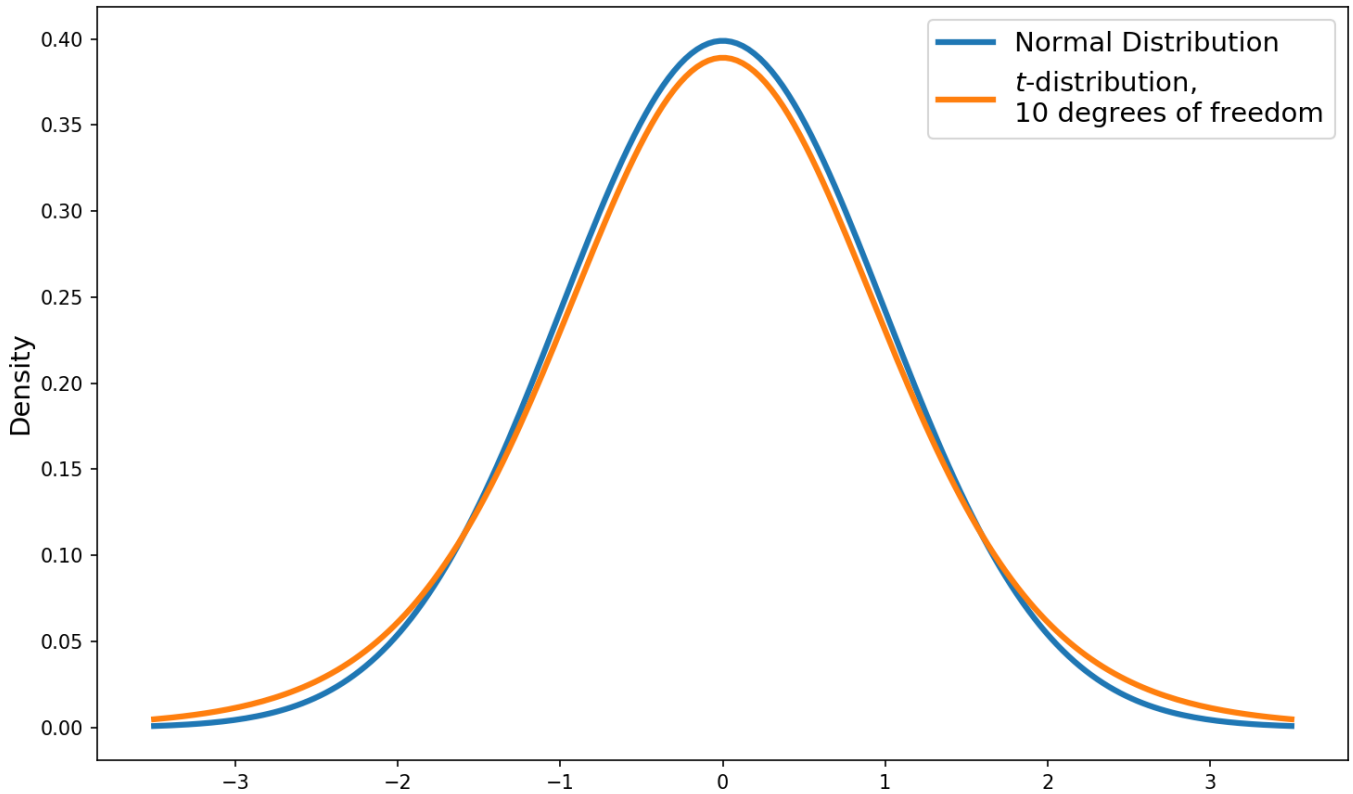
Confidence Intervals

The family of Student's t-distributions is a family of symmetric, bell-shaped distributions, which are similar to normal distributions, but have wider tails. That is, more extreme observations are more common. This family is parametrized by the number of degrees of freedom.



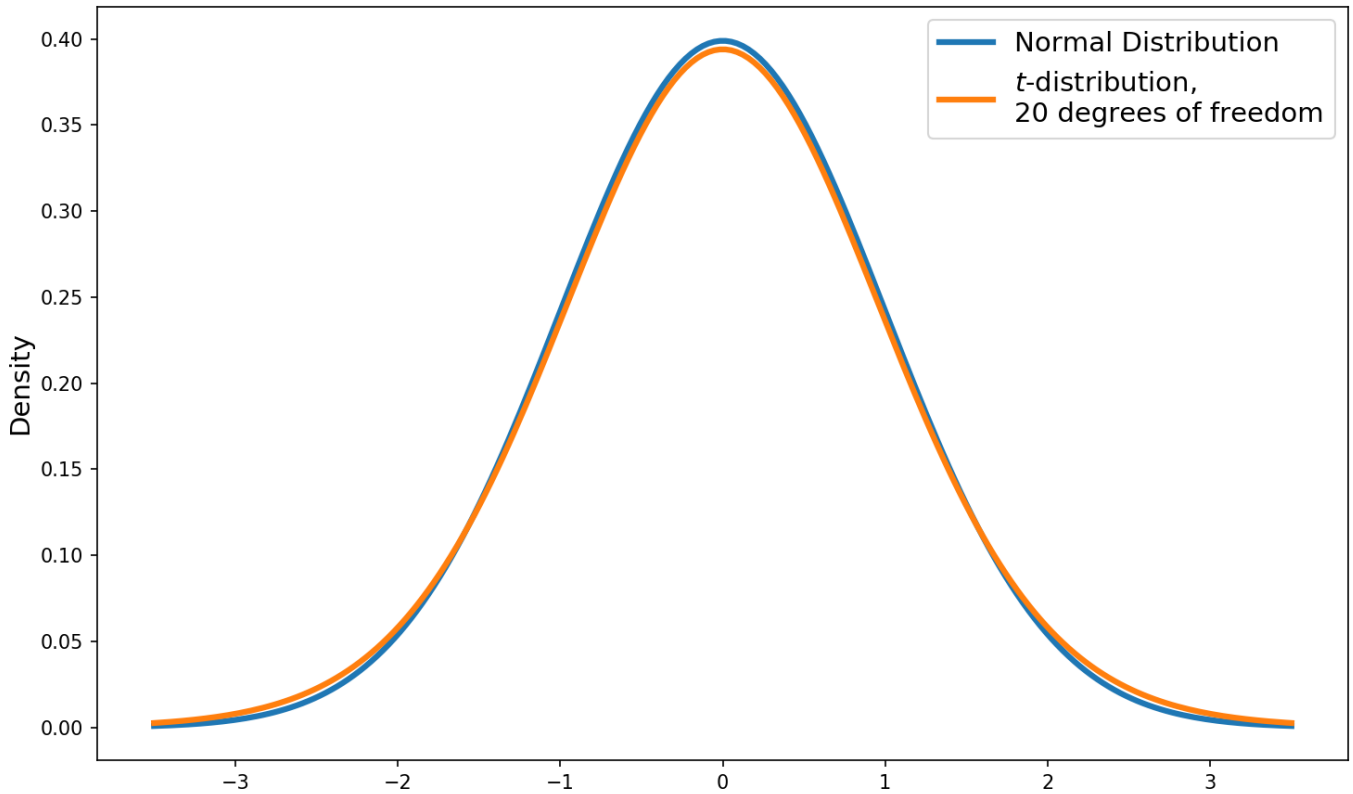
Confidence Intervals

The family of Student's t-distributions is a family of symmetric, bell-shaped distributions, which are similar to normal distributions, but have wider tails. That is, more extreme observations are more common. This family is parametrized by the number of degrees of freedom.



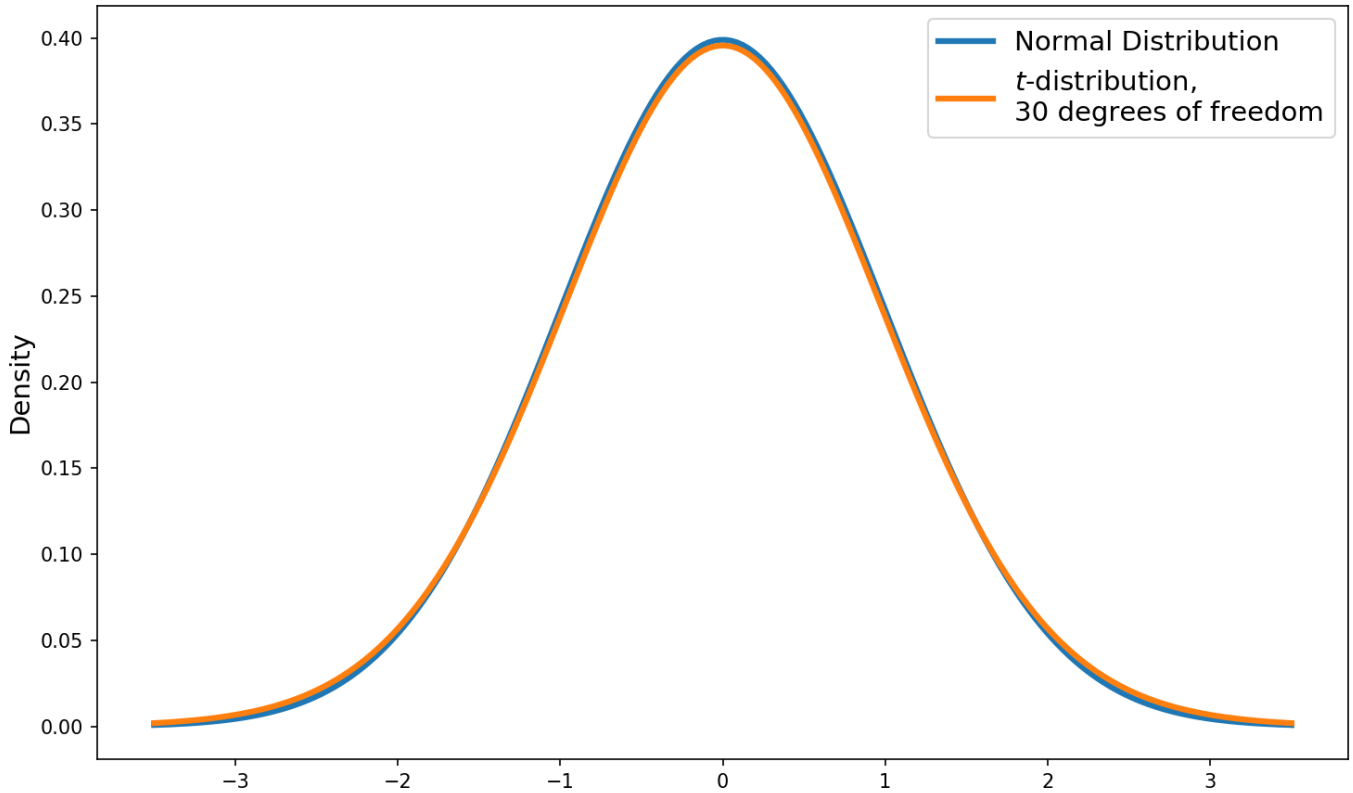
Confidence Intervals

The family of Student's t-distributions is a family of symmetric, bell-shaped distributions, which are similar to normal distributions, but have wider tails. That is, more extreme observations are more common. This family is parametrized by the number of degrees of freedom.



Confidence Intervals

The family of Student's t-distributions is a family of symmetric, bell-shaped distributions, which are similar to normal distributions, but have wider tails. That is, more extreme observations are more common. This family is parametrized by the number of degrees of freedom.



Confidence Interval

Big Idea: If we want a 95% confidence interval, we just need to find the distance $t_{0.025}$ from the center of the t distribution with $n-1$ degrees of freedom to the point where the area to the right is 0.025 (that is, $t_{0.025}$ is the 97.5th percentile).

Once we know $t_{0.025}$, we can get the margin of error by multiplying by the standard deviation of the sampling distribution, $\frac{\sigma}{\sqrt{n}}$. This means that our confidence interval is

$$\bar{x} \pm t_{0.025} \cdot \frac{\sigma}{\sqrt{n}}$$

Confidence Intervals for the Proportion

We can also construct confidence intervals for the population proportion.

Let $\hat{p}$ be the sample proportion and p be the population proportion.

Fact: The sampling distribution of the proportion is approximately normal with mean equal to the population mean and standard deviation $\sqrt{\frac{p(1-p)}{n}}$ (which we can approximate with $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$).

Thus, $\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}}$ approximately follows the standard normal distribution.