Estimation, Part 1

Sampling Distributions

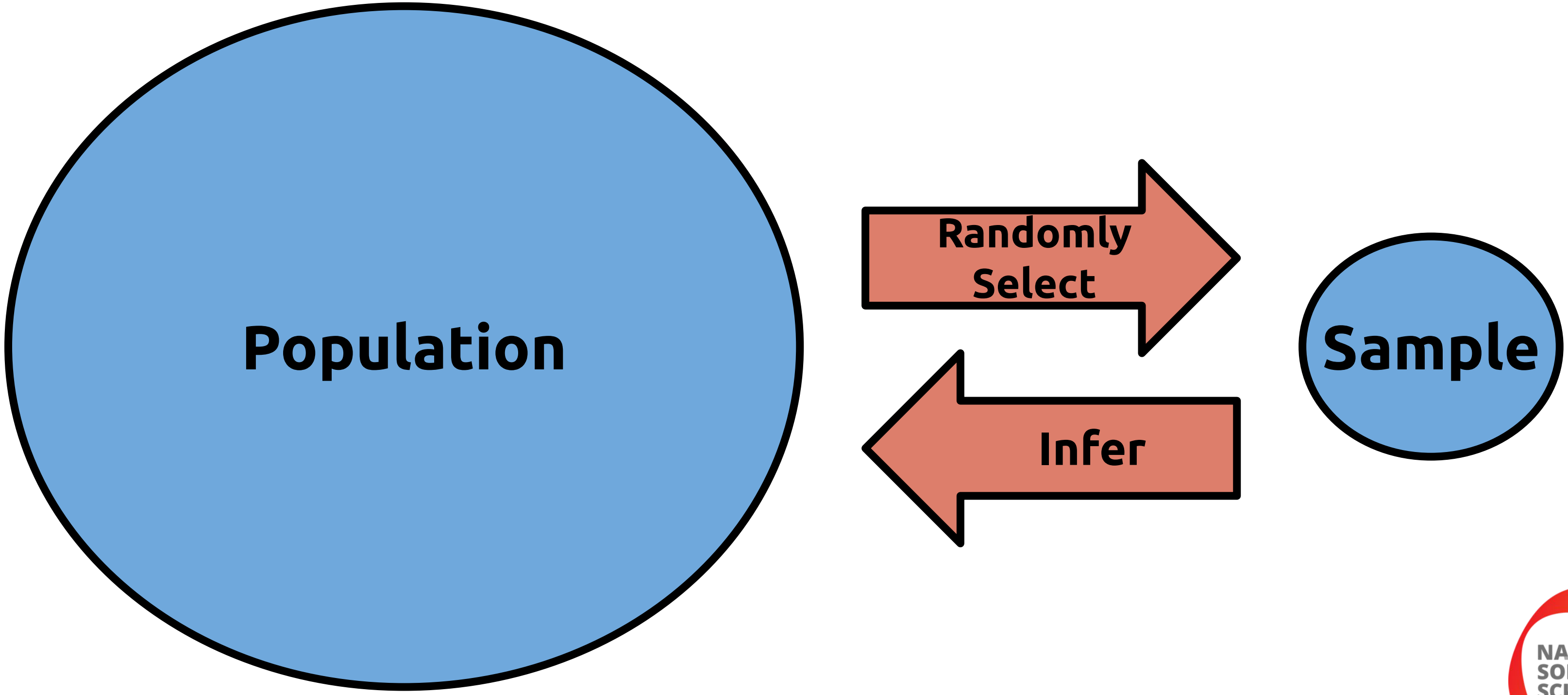
Estimation

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

Examples:

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

Estimation



Estimation

Can report a single number - a **point estimate**, as the estimate for a population parameter.

Eg. Estimating that the average household income in Putnam County is $43,000 based on a survey of 100 households.

Estimation

But, even with a large sample size, a point estimate is very unlikely to be correct.

Usually better to give some wiggle room, or a **margin or error**.

Eg. Reporting that we are highly confident that our estimate of $43,000 is off by no more than $1,500.

Estimation

Can write our estimate as

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

Or

$$\$41{,}500 < \mu < \$44{,}500$$

μ represents the true mean household income in Putnam County

Or

$$(\$41{,}500, \$44{,}500)$$

Estimation

We have created a **confidence interval**.

Confidence intervals have **confidence levels** which indicate the proportion of the time that the *procedure used to construct them* would contain the true population mean.

Eg. We could say that we are 95% confident that our interval contains the true population mean.

NASHVILLE SOFTWARE SCHOOL

Estimation

The general recipe for a confidence interval is

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

The margin of error depends on the confidence level:

- Higher confidence = wider margin of error
- Lower confidence = smaller margin of error

Estimation

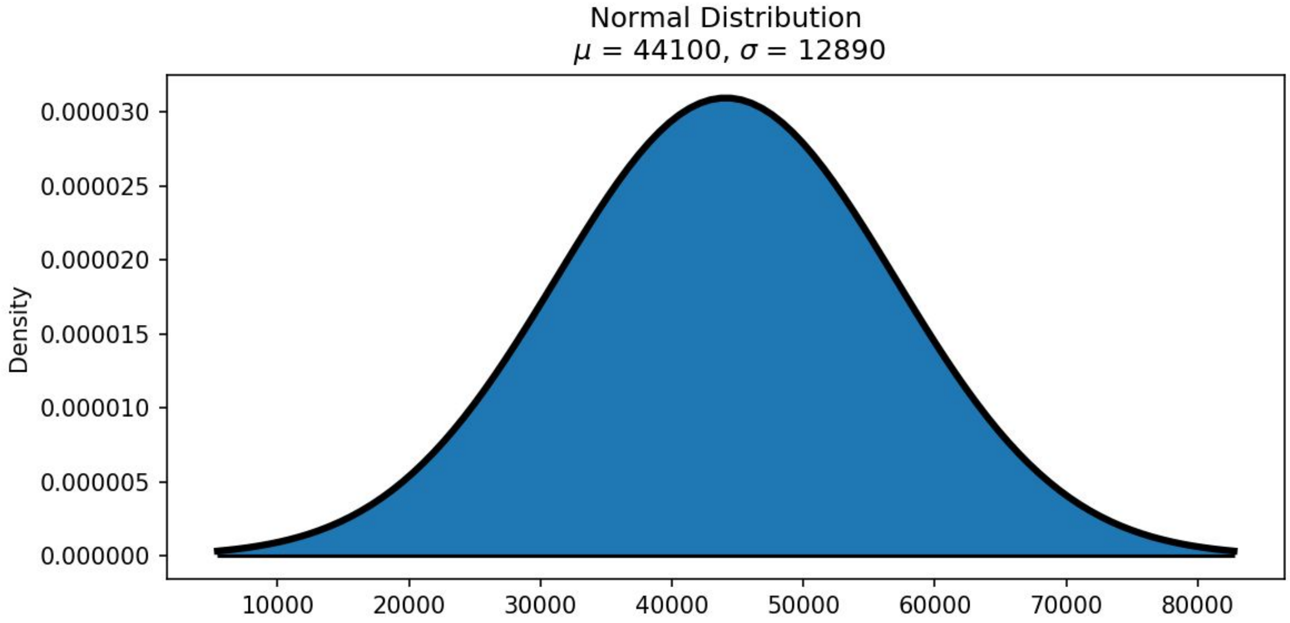
How do we find the margin of error?

We need to temporarily pretend that we know the real population distribution.

Let's say that for Putnam County, household incomes are distributed normally with a mean of \$44,100 and a standard deviation of \$12,890.

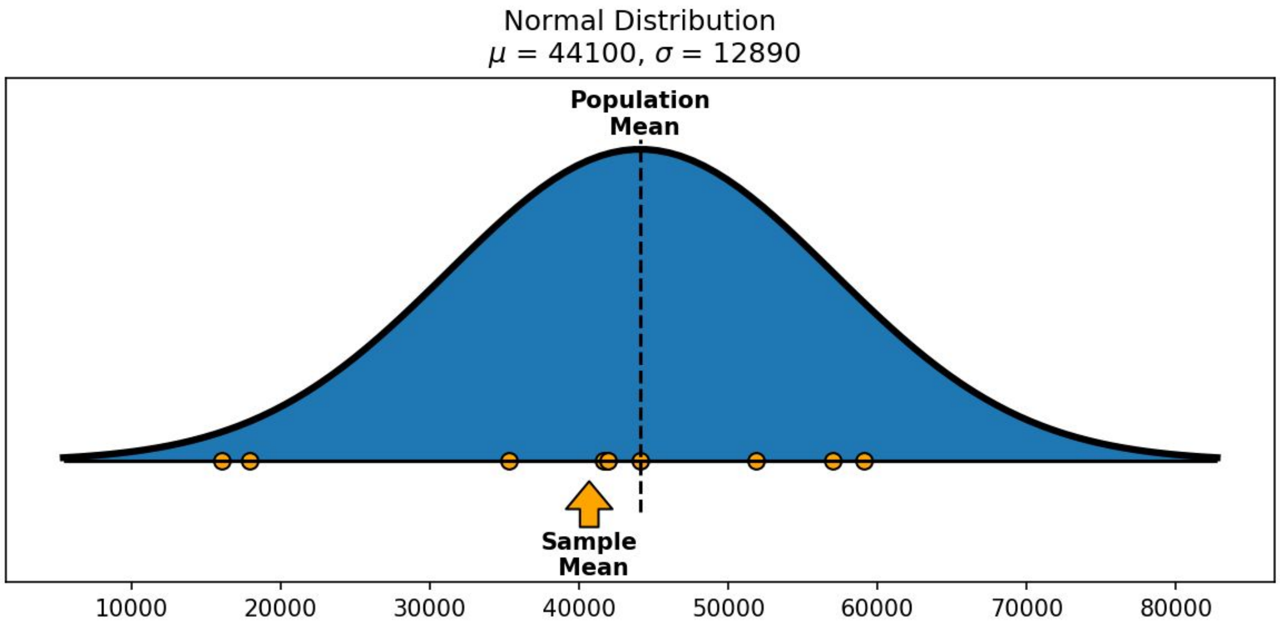
(Note: this is not a good approximation of income distribution, but this is only a thought experiment.)

Estimation



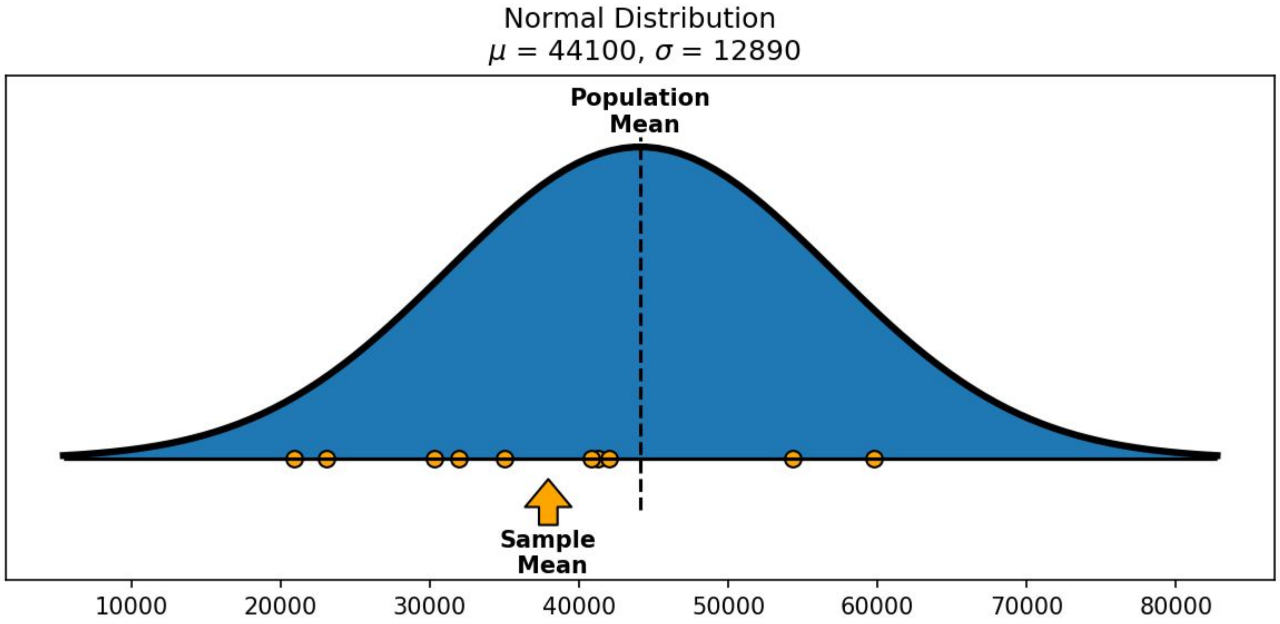
What might a sample of size 10 from this population look like?

Estimation



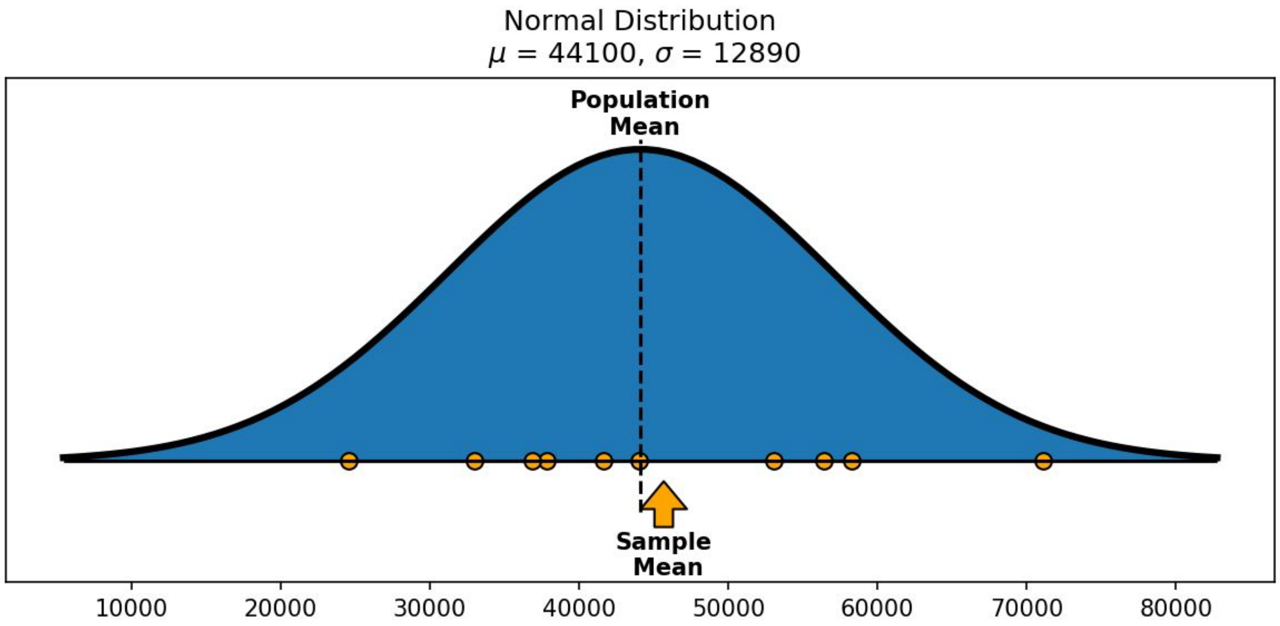
One Possible Sample

Estimation



Another Possible Sample

Estimation



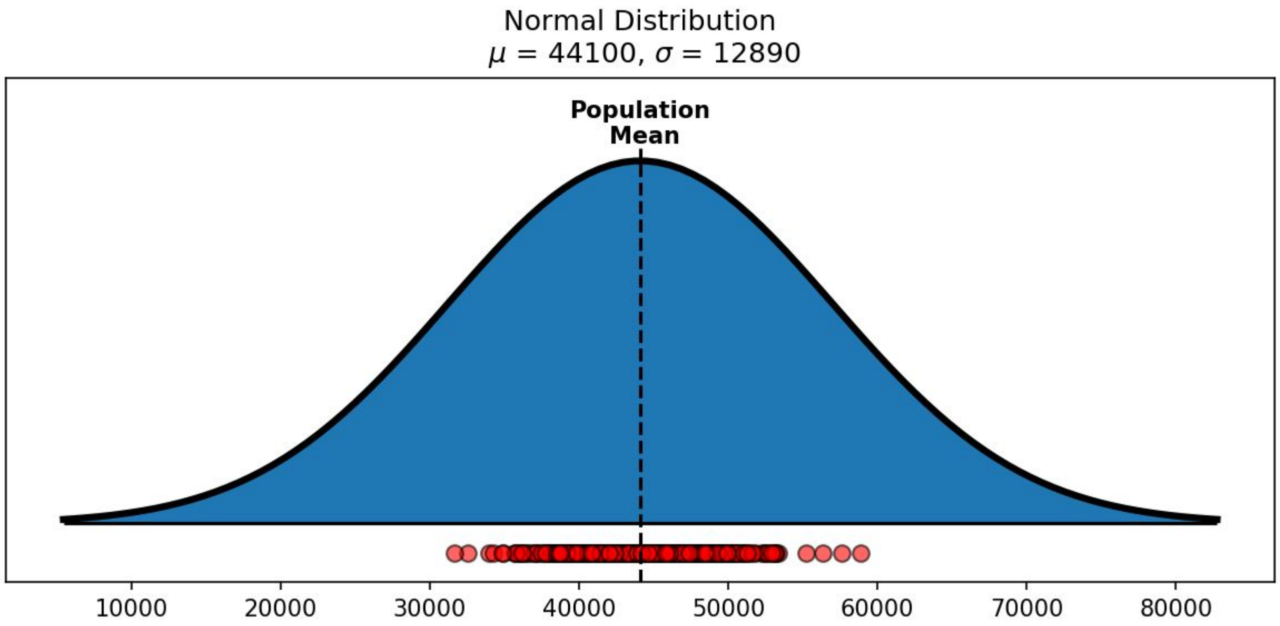
Yet Another Possible Sample

Estimation

In reality, we only get our one sample, from which we can calculate a sample mean $\bar{x}$

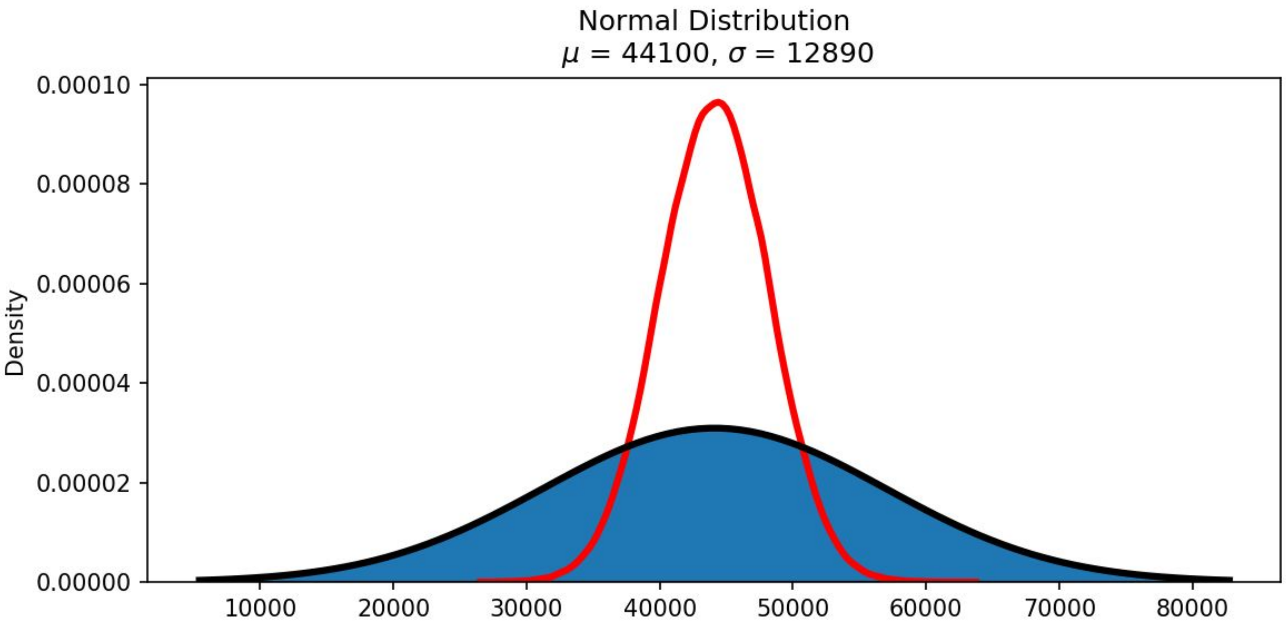
If we treat $\bar{x}$ as a random variable, we can try and understand its distribution.

Estimation



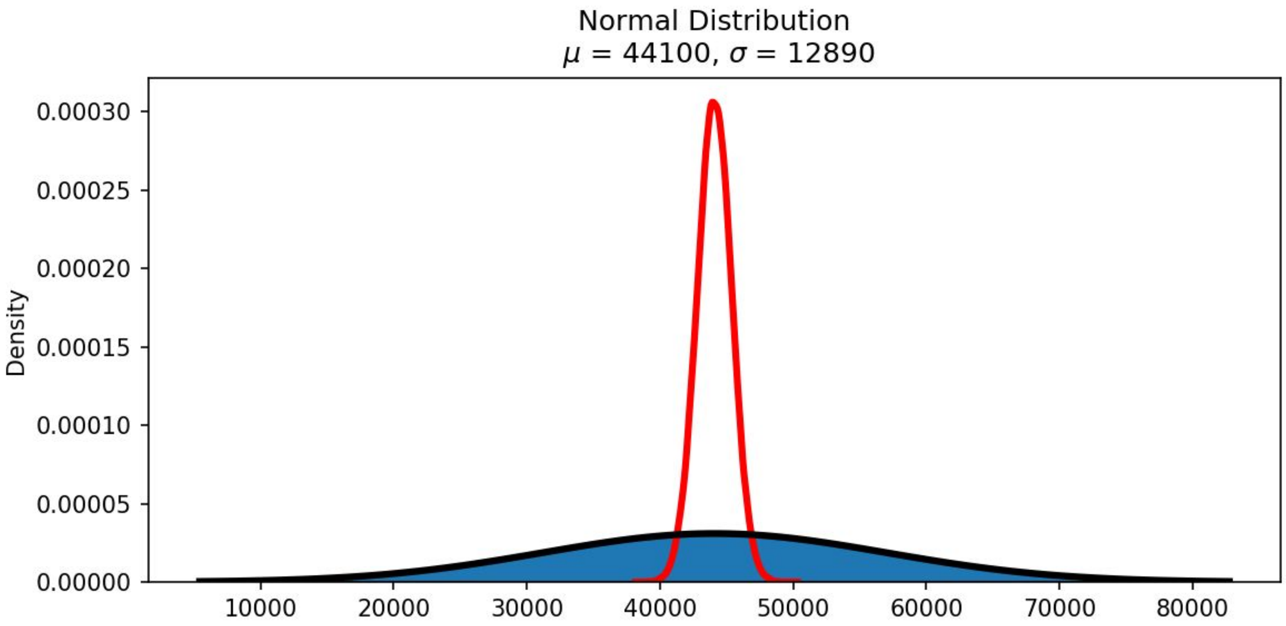
Each red dot corresponds to one of 500 different sample means $\bar{x}$

Estimation



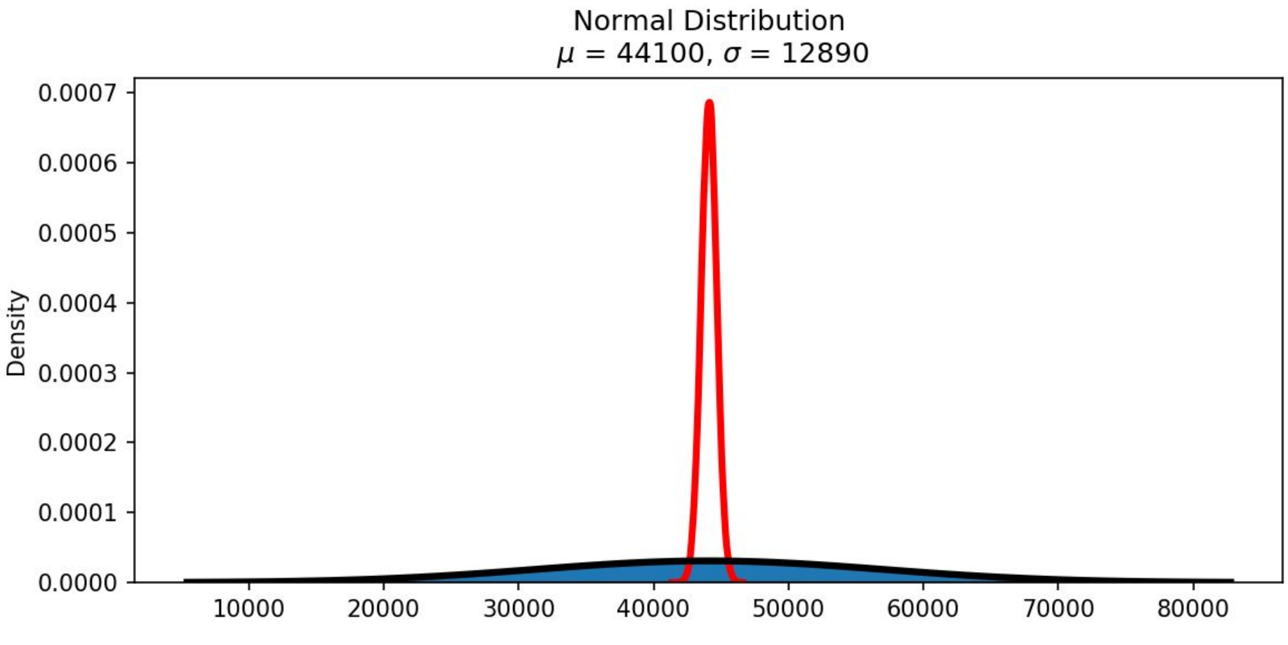
KDE for the distribution of sample means

Estimation



If we increase the sample size to 100, we get a tighter distribution.

Estimation



Sample size of 1000 gives an even narrower distribution of sample means.

Estimation

Let's look at confidence intervals widget to see how knowing about the distribution of sample means can help us construct a confidence interval.

Estimation

Central Limit Theorem: For the random variable x with mean μ and standard deviation σ, the distribution of sample means of size n is approximately normal mean mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$

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

The quantity $\frac{\sigma}{\sqrt{n}}$ is called the **standard error of the mean.**

Estimation

Problem: To use the Central Limit Theorem, we need to know the standard deviation, $\boldsymbol{\sigma}$, of the *population*.

At best, we usually only know 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.

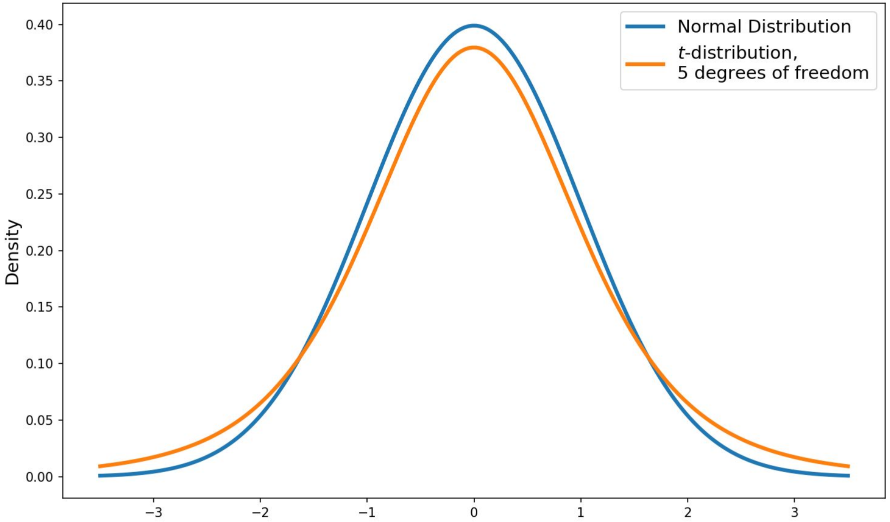
Estimation

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 3 observations.

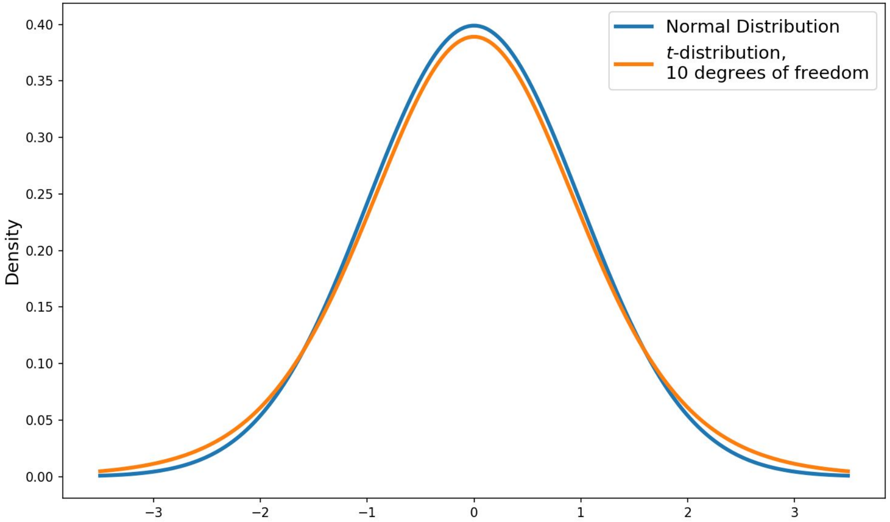
Estimation

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.



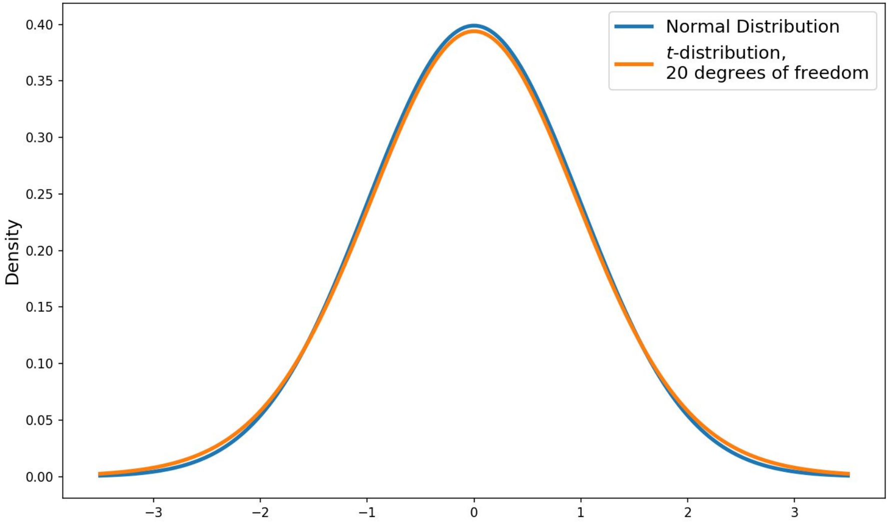
Estimation

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.



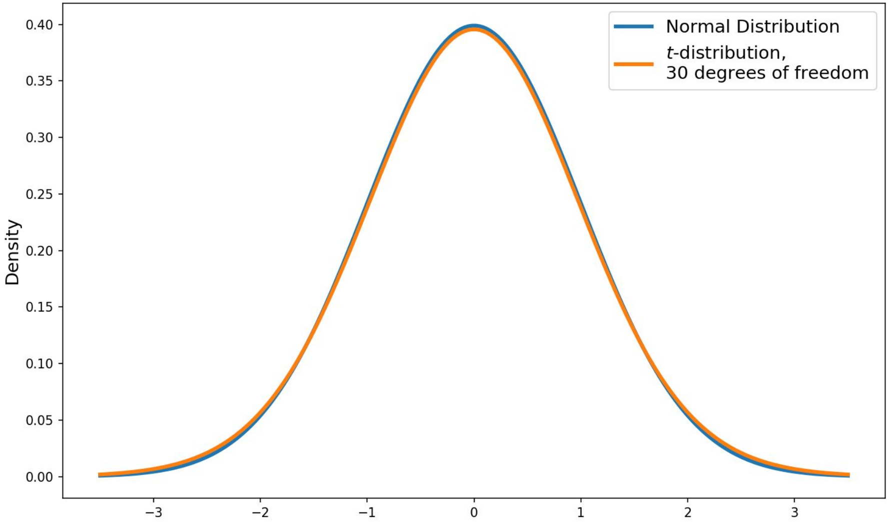
Estimation

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.

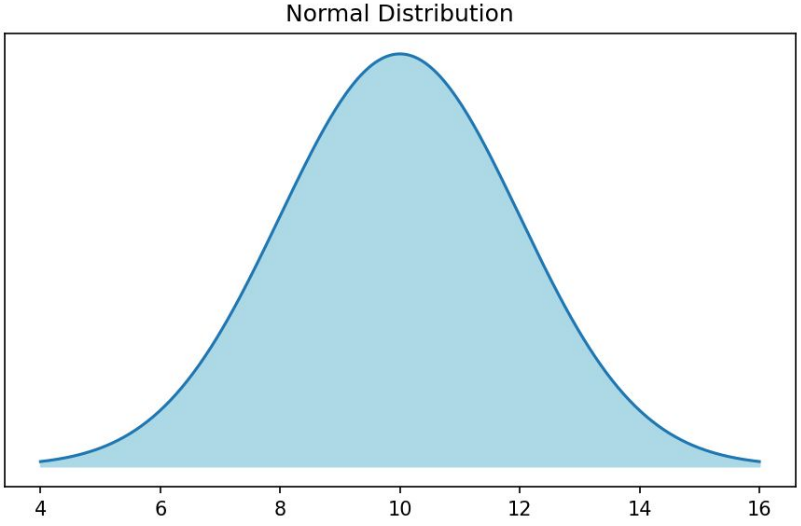


Estimation

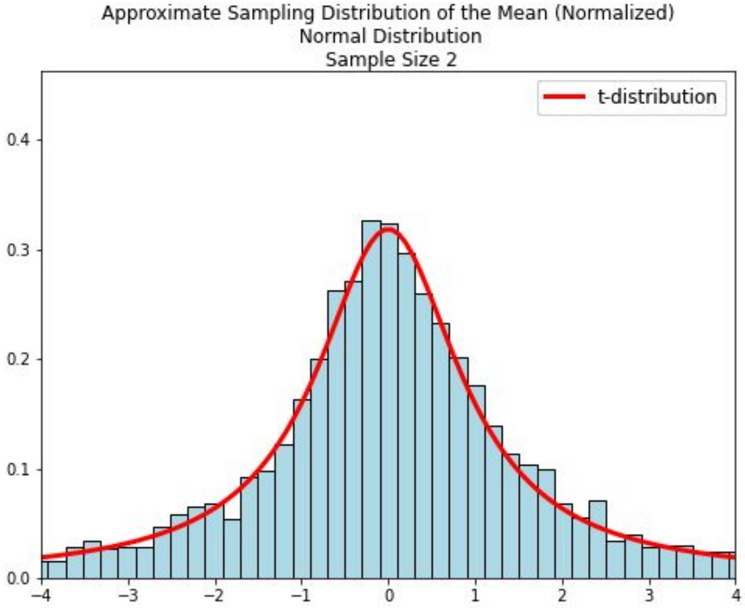
The family of Student's *t*-distributions is a family of symmetric, bell-shaped distributions, which are similar to normal distributions, but have wider tails. That is, more extreme observations are more common. This family is parametrized by the number of degrees of freedom.



Python Simulations of Sampling Distributions

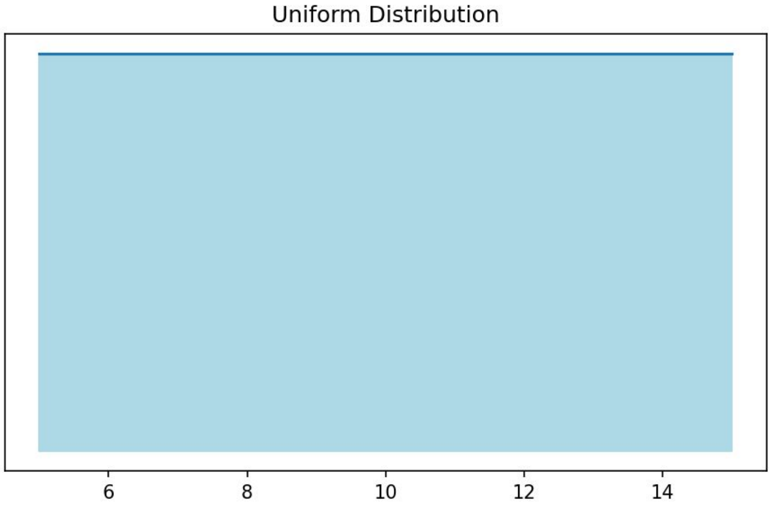


Population Distribution

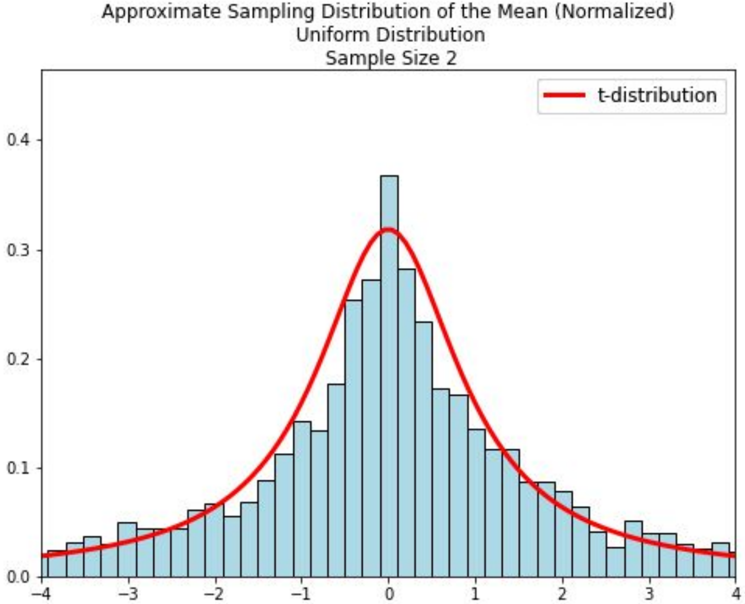


Distribution of $\dfrac{\bar{x} - \mu}{s/\sqrt{n}}$

Python Simulations of Sampling Distributions

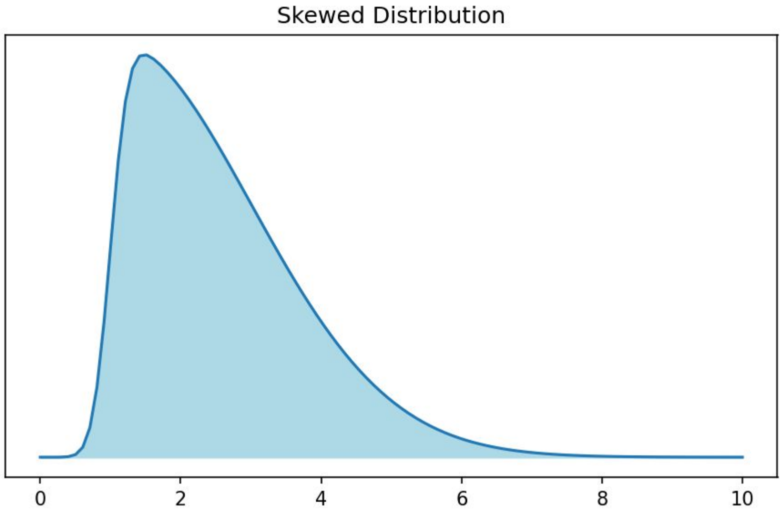


Population Distribution

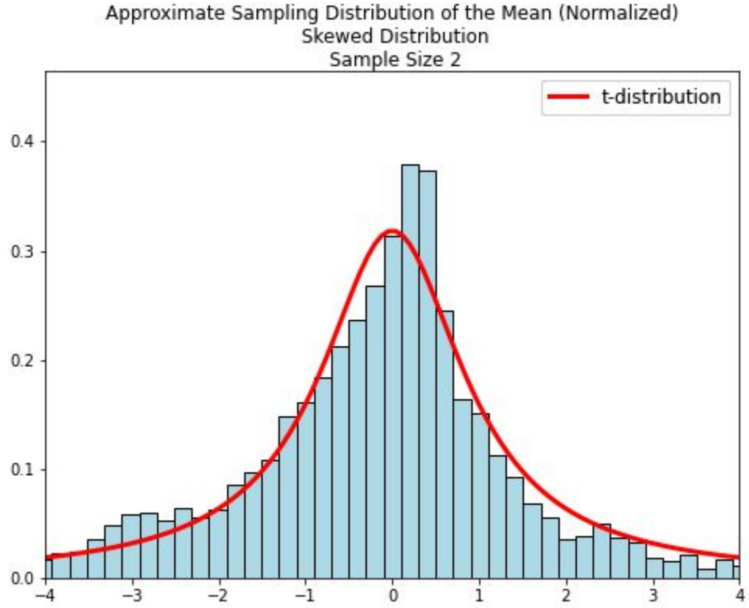


Distribution of $\dfrac{\bar{x} - \mu}{s/\sqrt{n}}$

Python Simulations of Sampling Distributions



Population Distribution



Distribution of $\dfrac{\bar{x} - \mu}{s/\sqrt{n}}$

Estimation

Big Idea: If we want a 95% confidence interval for the mean, 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).

Then, multiply by the standard deviation of the sampling distribution, $\frac{s}{\sqrt{n}}$

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

Point Estimate

Margin of Error

Estimation

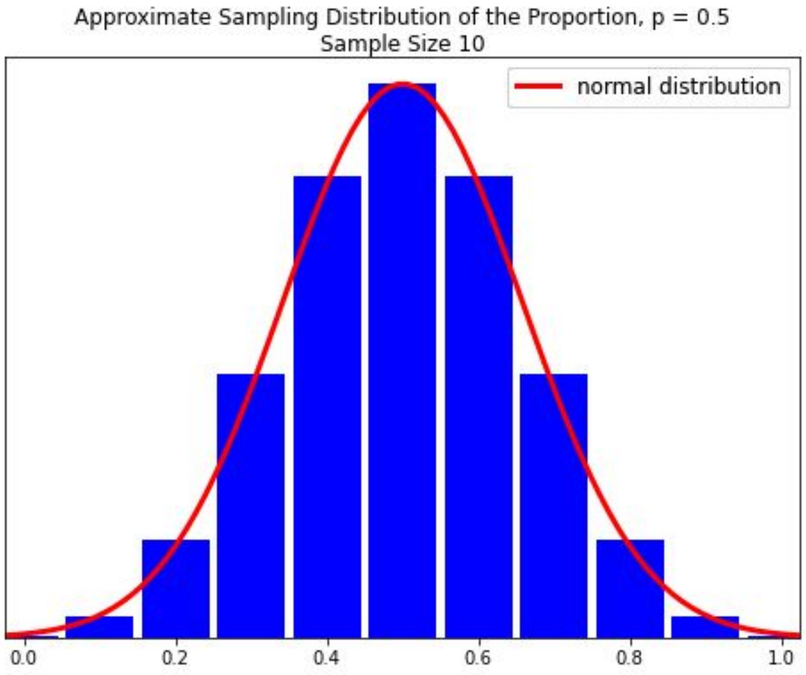
For proportions, we can use sample proportion $\hat{p}$ to estimate the population proportion *p*

Fact: For a sample of size *n*, the quantity

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p})/n}}$$

approximately follows a standard normal distribution.

Python Simulations of Sampling Distributions



Python Simulations of Sampling Distributions