

# Confidence Intervals

Grinnell College

October 8, 2025

# Warmup

- ▶ If I have 1,000 observations, how many of them will fall between the 10th and 90th percentile?
- ▶ How does sampling distribution differ from distribution of a sample?
- ▶ What are distributional parameters of normal distribution?
- ▶ If I have two samples with:
  - ▶ Sample 1:  $n_1 = 25$  and  $\sigma_1 = 10$
  - ▶ Sample 2:  $n_2 = 50$  and  $\sigma_2 = 15$

which sample will have the least variability in its estimate of  $\bar{X}$ ?

# Review

The **Law of Large Numbers** guarantees that, as the number of observations  $n$  in my sample increases, my estimate of the parameter will converge to the true value

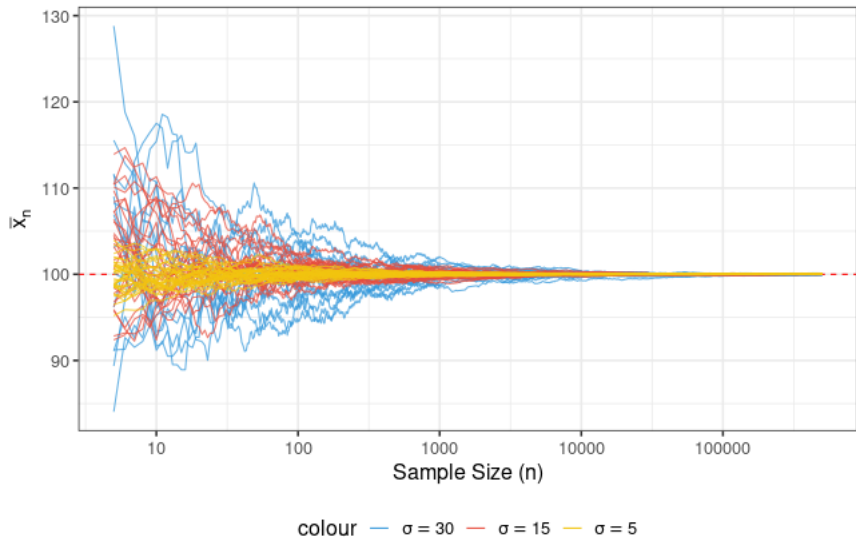
A **sampling distribution** refers to the distribution of a sample statistic (i.e.,  $\bar{X}$ ) if we were to repeatedly sample from a population and recompute the statistic

- ▶ What values would they take?
- ▶ How frequently would they appear?

The **Central Limit Theorem** states that if my statistic is an average or a proportion, then the sampling distribution of my statistic will be approximately normal, with

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

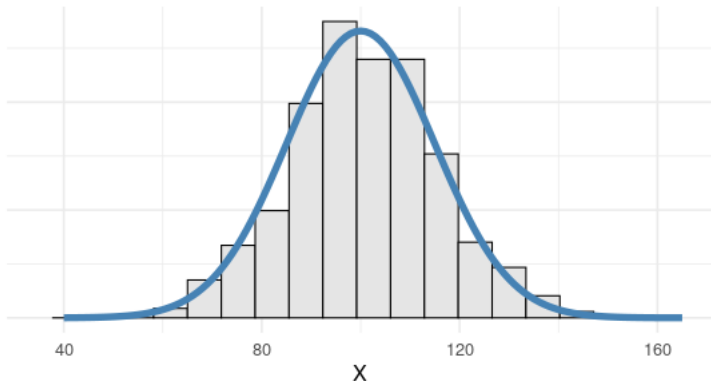
## Different Sample SD



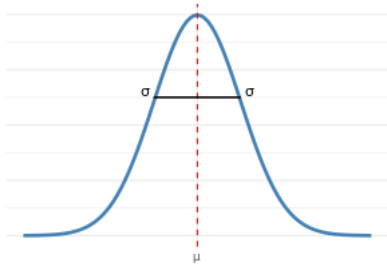
# Notes on Normal

The **normal distribution** describes a distribution that is

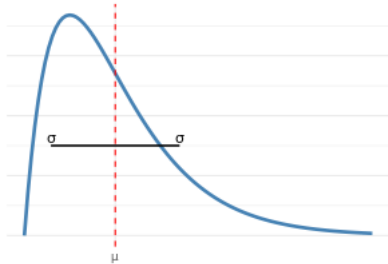
- ▶ Bell-shaped
- ▶ Symmetric about the mean
- ▶ Has two distributional parameters, the mean  $\mu$  and standard deviation  $\sigma$



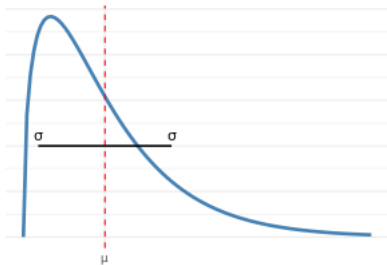
Normal



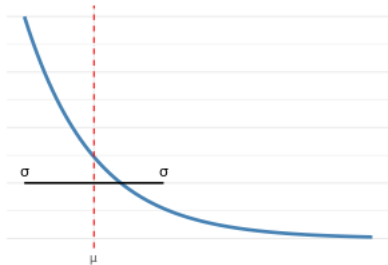
Gamma



Chi-Square



Exponential



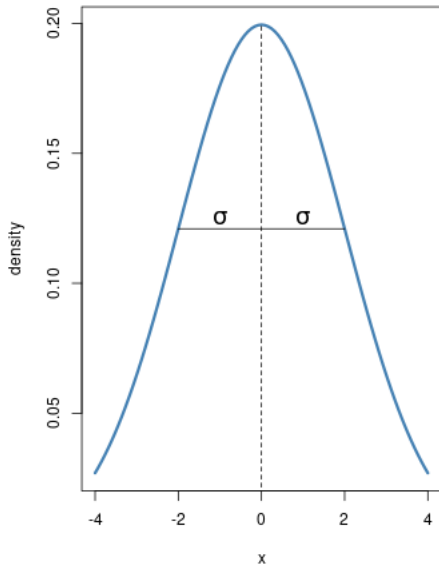
# Some Terms to Know

**Standard Deviation:** A description of the variability in our *observations* describing average distances from the average or mean. It is often denoted  $\sigma$

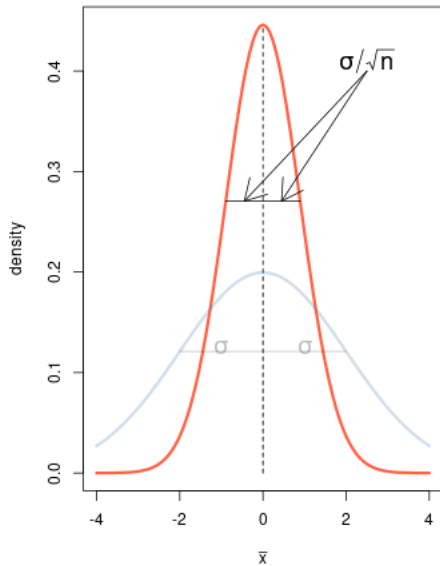
**Standard Error:** A description of variability in our *sampling distribution*. We will denote standard error as  $SE$ , with  $SE = \sigma/\sqrt{n}$ , where  $n$  is the number of observations in our sample.

Note that the standard error *is* the standard deviation of the sampling distribution

### Standard Deviation



### Standard Error





One issue we have with the CLT is that it utilizes two parameters that we generally do not know:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

For now, we will rely on two simple facts:

1. The LLN suggests that our best guess for  $\mu$ , given our data, is  $\bar{X}$
2. Our best guess for  $\sigma$  will be  $\hat{\sigma}$ , which I can estimate from my sample
3. My estimate of the standard error, then, will be  $\hat{\sigma}/\sqrt{n}$ , where  $n$  is the size of my sample

Our basic goal is this: *can I use information from my sample to provide a range of reasonable estimates for my population parameter?*

In actuality, this represents two competing goals:

1. I want an interval that is narrow enough to provide a useful estimate of my parameter
2. I want an interval that is wide enough that it is likely to contain the true value

# Benefits of a distribution

## Empirical Distribution:

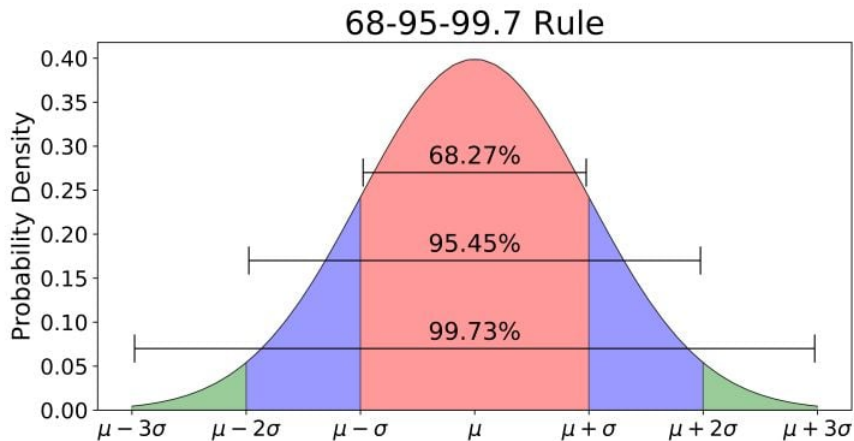
Suppose I have gone out and collected a sample:

- ▶ If I wanted to find the median of this dataset, what would I do?
- ▶ What if I wanted to find Q1 and Q3 of this dataset?

## Named Distribution:

What if instead I wanted to find the median and Q1 and Q3 of a normal distribution with mean value  $\mu$  and standard deviation  $\sigma$ ?

# Empirical Rule



# Confidence

We may ask ourselves: if I were to sample from this population over and over, where would the middle  $p\%$  of my estimates fall?

If we were to create our interval around the middle 99.999%, we would nearly certainly have an interval that contains the true value of the mean, but our interval of possible values would be very large

Alternatively, if we elected to construct an interval containing only the middle 50% of our sampling distribution, we would have a very small interval but may very well construct an interval that does not contain the true parameter

We call the amount of certainty based on percentiles our **confidence**

# Intervals

The range of values we might consider realistic in estimating our population parameter will be based on two things:

- ▶ The amount of uncertainty in our data (standard error)
- ▶ How conservative we wish for our estimate to be

These two ideas will come together to give us an estimate that looks like:

$$\bar{X} \pm \text{Margin of Error}$$

# Using the normal distribution

Because the CLT grants that our statistic is normally distributed, we can express our *confidence* in terms of multiples standard errors

For example, an interval based on

$$\bar{x} \pm 2 \times SE$$

will contain about 95% of our sampled statistics

# Confidence Intervals

Suppose the mean of my population is  $\mu = 50$ , with  $\sigma = 15$  and  $n = 30$ . The statistics from a collected sample are

$$\bar{x} = 46.35, \quad \hat{\sigma} = 15.281$$

From here, we can construct a **95% confidence interval** of:

$$\begin{aligned} 95\%CI &= \text{Point estimate} \pm \text{Margin of Error} \\ &= \bar{x} \pm 2 \times \hat{\sigma} / \sqrt{n} \\ &= 46.35 \pm 2 \times 2.79 \\ &= (40.75, 51.93) \end{aligned}$$

What does this even mean?

- ▶ 95% what?
- ▶ We are 95% sure it contains the mean?
- ▶ The probability of the mean being there is 95%?
- ▶ Or something else?



## Repeated Samples

Suppose that I had repeated this process of collecting samples of size  $n = 30$  and constructing intervals many times

Sample	$\bar{x}$	$\hat{\sigma}$	95% CI	Contains $\mu$
1	46.49	2.93	(40.64, 52.34)	TRUE
2	50.06	2.69	(44.68, 55.43)	TRUE
3	48.74	2.68	(43.38, 54.11)	TRUE
4	48.26	2.34	(43.59, 52.93)	TRUE
5	52.48	2.84	(46.79, 58.17)	TRUE
6	49.75	3.13	(43.48, 56.01)	TRUE
7	50.37	2.85	(44.66, 56.07)	TRUE
8	47.20	3.07	(41.06, 53.35)	TRUE
9	55.06	2.20	(50.65, 59.47)	FALSE
10	46.47	3.39	(39.68, 53.25)	TRUE

# Confidence Intervals

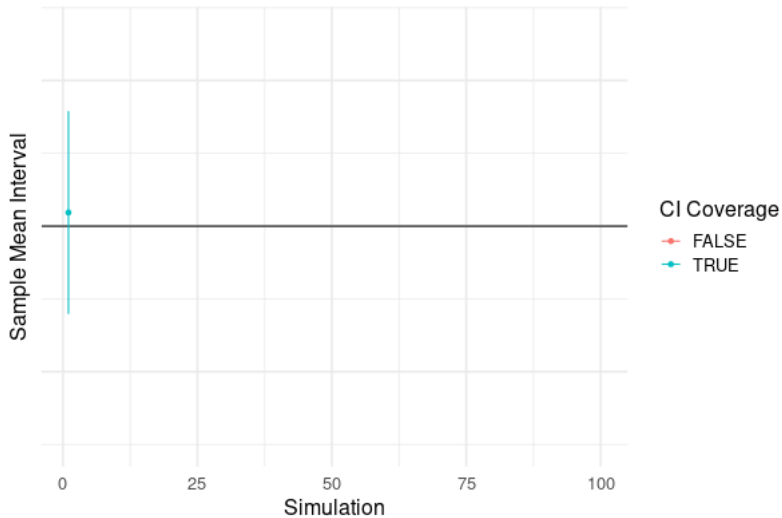
A **confidence interval** is an interval that has the following properties:

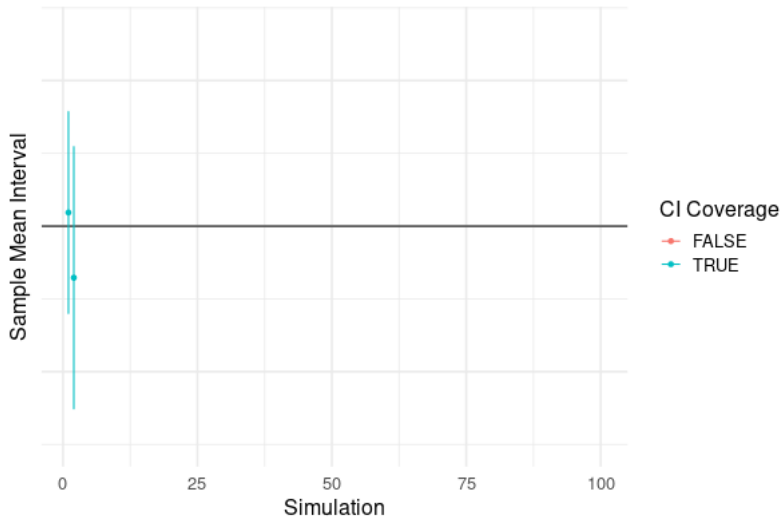
- ▶ It is the result of a *random process*
- ▶ It is constructed according to a procedure or set of rules
- ▶ It is made with the intention of giving a plausible range of values for a *parameter* based on a *statistic*
- ▶ There is no probability associated with a confidence interval; *it is either correct or it is incorrect*

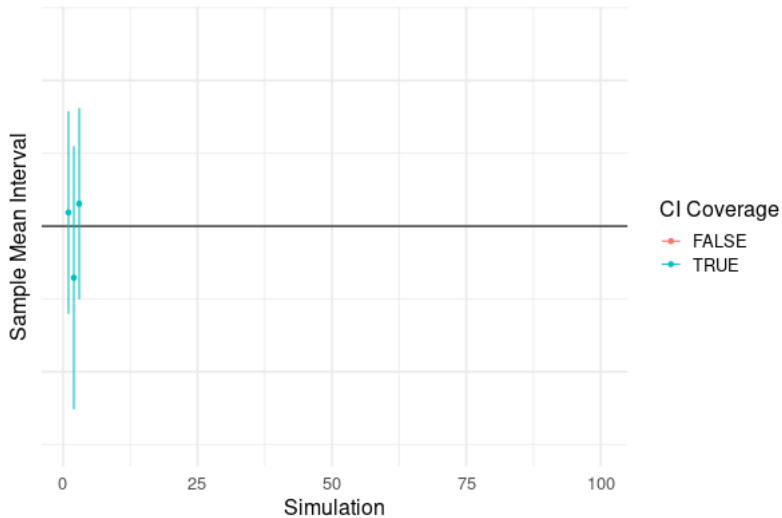
# 95% Confidence Interval

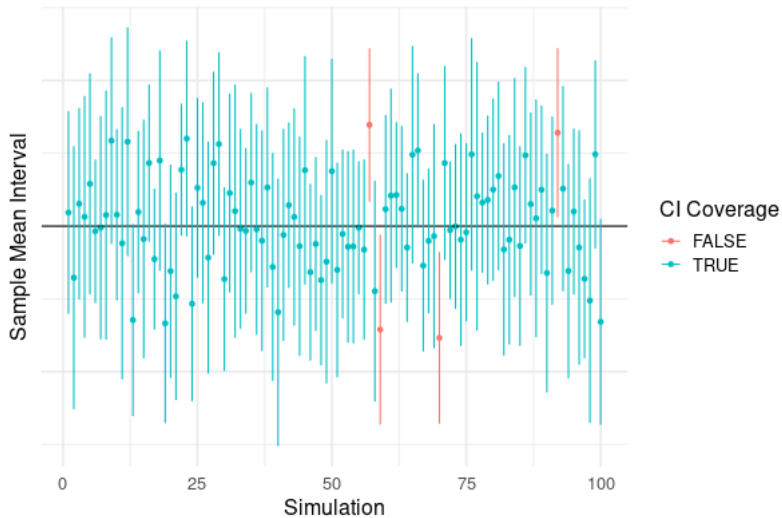
When we say something has a 95% confidence interval, what we mean is:

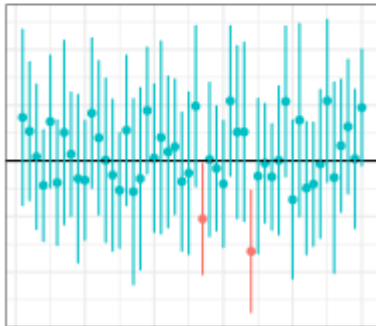
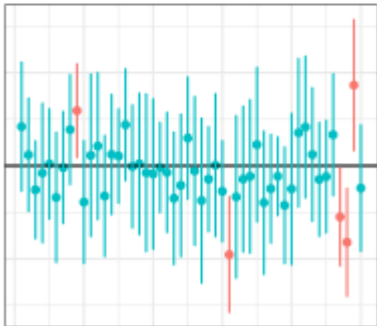
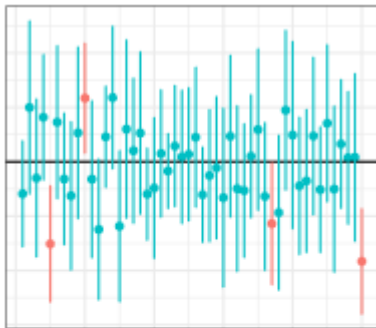
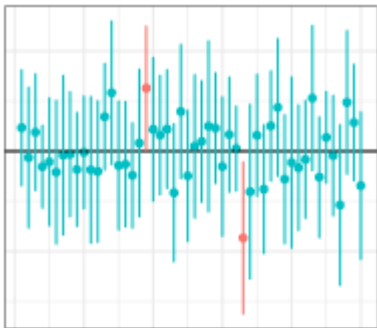
*The process that constructed this interval has the property that, on average, it contains the true value of the parameter 95 times out of 100*













# Confidence Intervals

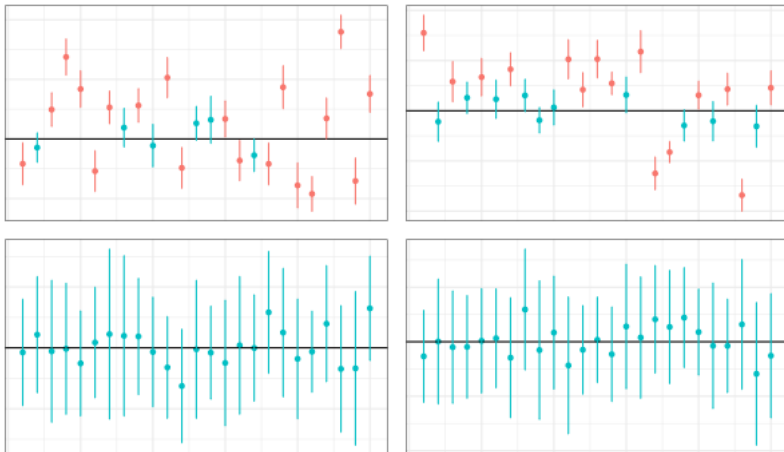
To be absolutely clear: we will *never* know if the confidence interval we construct contains the true value of the parameter

This is akin to throwing a dart but never seeing the target

This is the nature of statistical inference: we can describe properties of the *process* that created our intervals, but we can never conclusively speak about the interval itself

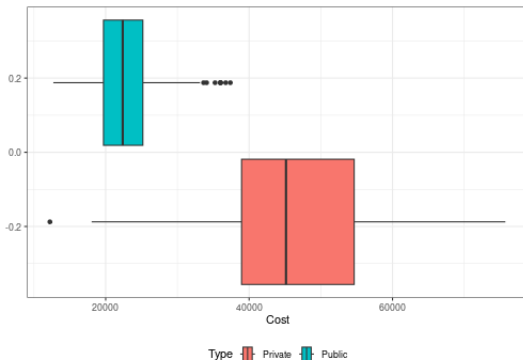
# Confidence Intervals

It is also worth observing that we can *alter* our process to achieve different results. There is a trade-off between how frequently we are correct and how much uncertainty we allow in our prediction



## Example

Our college dataset, which represents a population, contains 1,095 observations, with 647 private schools and 448 public schools. The distributions and true average cost of each group is given below:



Type	Average Cost
Private	47,073
Public	22,766

## Example

Let's randomly collect a sample of 50 schools from each group and create a confidence interval for the mean

Type	$\bar{X}$	Std. Error
Sample Private (N = 50)	44,947	1,467
Sample Public (N = 50)	22,833	684

$$\begin{aligned} 95\% \text{ CI for Private} &= \text{Point estimate} \pm \text{Margin of Error} \\ &= \bar{X} \pm 2 \times \text{SE} \\ &= 44,947 \pm 2 \times 1,467 \\ &= (42,013, 47,882) \end{aligned}$$

## Example

Let's randomly collect a sample of 50 schools from each group and create a confidence interval for the mean

Type	$\bar{X}$	Std. Error
Sample Private (N = 50)	44,947	1,467
Sample Public (N = 50)	22,833	684

$$\begin{aligned} 95\% \text{ CI for Public} &= \text{Point estimate} \pm \text{Margin of Error} \\ &= \bar{X} \pm 2 \times \text{SE} \\ &= 22,833 \pm 2 \times 684 \\ &= (21,464, 24,201) \end{aligned}$$

# Example

Type	95% Conf Int.	True Mean
Private	(42,013, 47,882)	47,073
Public	(21,464, 24,201)	22,766

# Review

- ▶ **Standard deviation** ( $\sigma$ ) is an estimate of the amount of variability in our sample, while **standard error** ( $\sigma/\sqrt{n}$ ) is an estimate of the variability in estimating a parameter
- ▶ A **sampling distribution** describes the distribution of a statistic or parameter estimate if we could repeat the sampling process as many times as we wish
- ▶ Approximations to the normal distribution generally follow the **66-95-99 rule** with 1/2/3 standard deviations of the mean
- ▶ If these properties hold, we can create a reasonable interval of possible parameter values of the form Point Estimate  $\pm$  Margin of Error
- ▶ A **confidence interval** is an interval with the properties that:
  - ▶ It is constructed according to a procedure or set of rules
  - ▶ It is intended to give plausible range of values for a *parameter* based on a *statistic*
  - ▶ It has no probability; the interval either contains the true value or it does not