Confidence Intervals

Grinnell College

October 16, 2024

Review

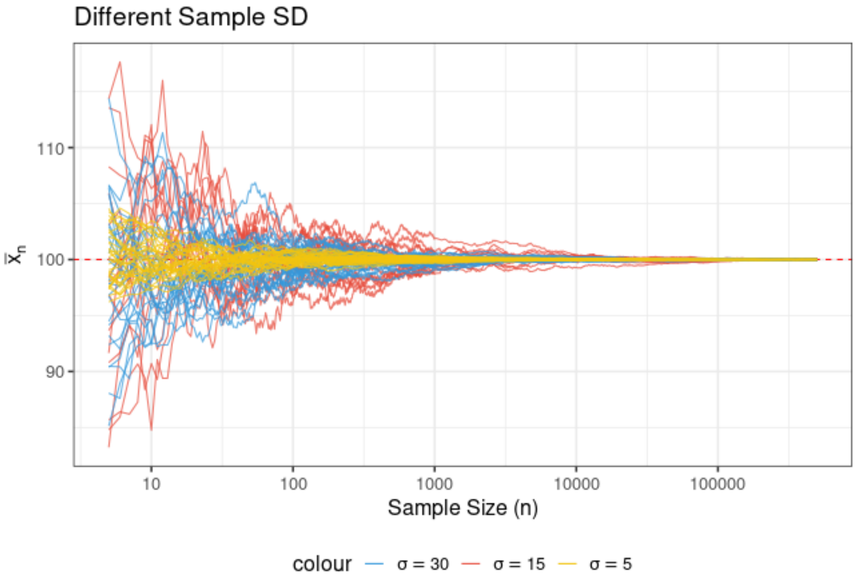
A **sampling distribution** refers to the distribution of a sample statistic (i.e., $\overline{x}$) if we were to repeatedly sample from a population and recompute the statistic

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

The **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

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

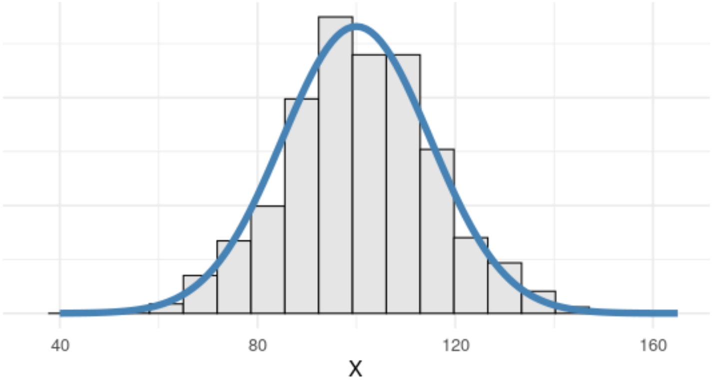
$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

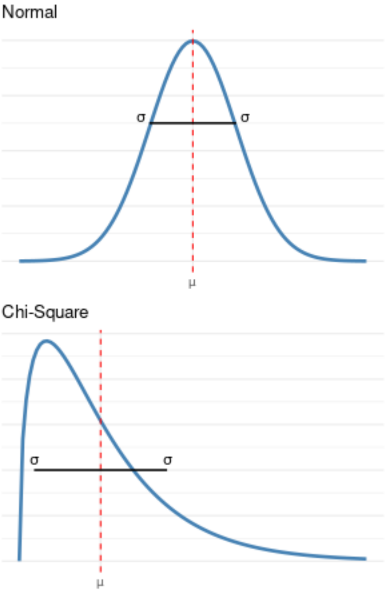

Different Sample SD
$\bar{x}_n$
110
100
90
10
100
1000
10000
100000
Sample Size (n)
colour
σ = 30
σ = 15
σ = 5

Notes on Normal

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

- Bell-shaped
- Symmetric about the mean
- Has two distributional parameters, the mean μ and standard deviation σ

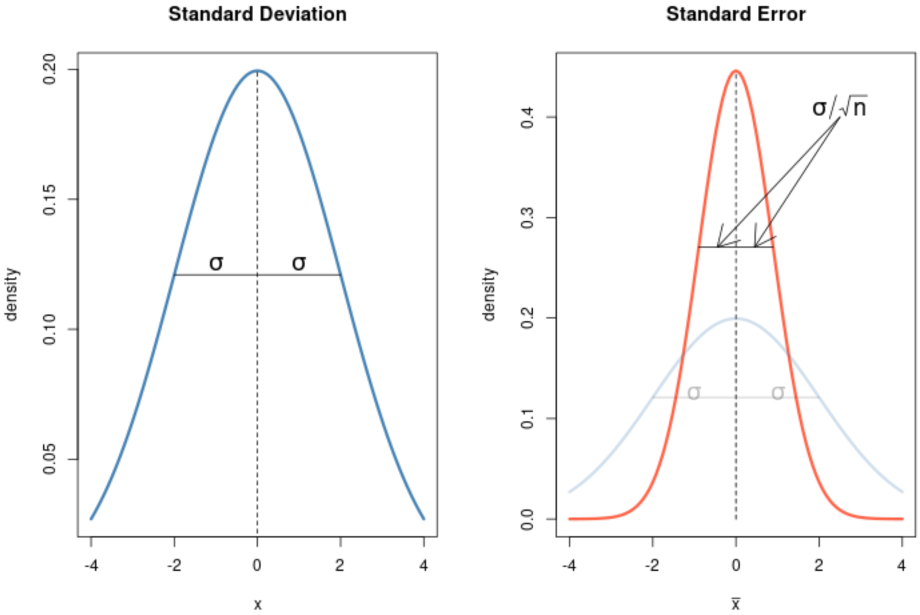



Normal
σ
σ
μ
Gamma
σ
σ
μ
Chi-Square
σ
σ
μ
Exponential
σ
σ
μ

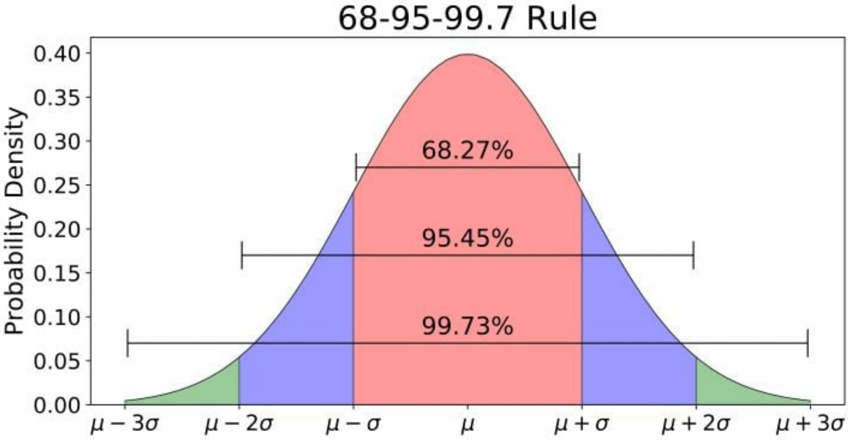
Some Terms to Know

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

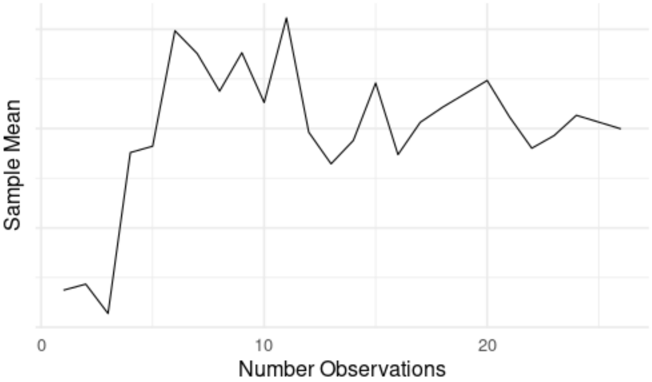
Standard Error: A description of variability in our *estimates* of a parameter (such as the mean). We will denote standard error as SE, with $SE = \sigma/\sqrt{n}$, where n is the number of observations in our sample


Standard Deviation
Standard Error
density
0.05
0.10
0.15
0.20
σ
σ
-4
-2
0
2
4
x
0.0
0.1
0.2
0.3
0.4
σ/√n
$\bar{x}$

Empirical Rule



Suppose we conduct a study to estimate a population mean, and we collect sample of size $n = 30$. How could we use this to estimate the mean?



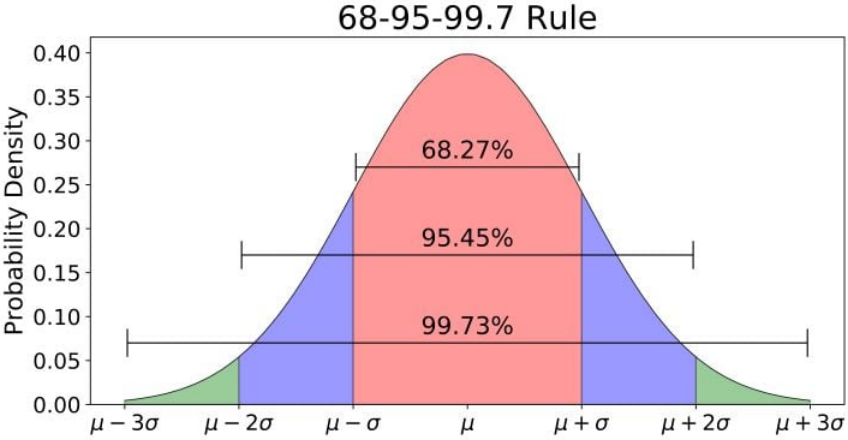
Intervals

For now, recall that our goal here is to determine the mean of our *population*.

If we can make an estiamte of the uncertainty associated with our statistic, $\overline{X}$, perhaps we can find a range of reasonable values:

$$\text{Point Estimate} \pm \text{Margin of Error}$$

Empirical Rule



Intervals

95% seems a reasonable thing to do, which is $\mu \pm 2\sigma$ from the previous slide

If our *point estimate* is $\overline{x}$ and our margin of error for measuring parameters is the standard error, then perhaps

$$\overline{x} \pm 2 \times \text{SE}$$

would create for us a suitable interval of plausible values

Confidence Intervals

The mean of my population from the previous slide is $\mu = 50$, with $\sigma = 15$. The statistics from my sample were

$$\overline{x} = 46.35, \qquad s = 2.79$$

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

$$\begin{aligned} 95\% CI &= \text{Point estimate} \pm \text{Margin of Error} \\ &= \overline{x} \pm 2 \times s \\ &= 46.35 \pm 2 \times 2.79 \\ &= (40.75, 51.93) \end{aligned}$$

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?

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*

Confidence Intervals

Consider the confidence interval that we constructed on a previous slide from Sample 4:

- It was constructed according to the procedure

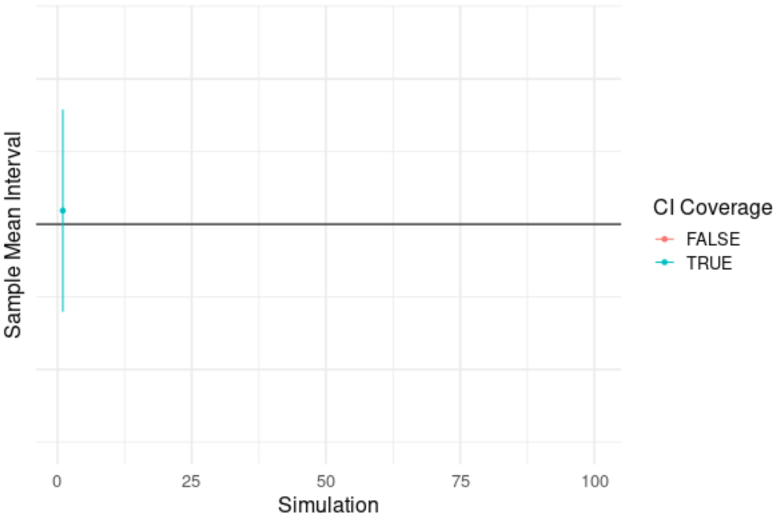
$$\text{Point estimate} \pm \text{Margin of Error}$$

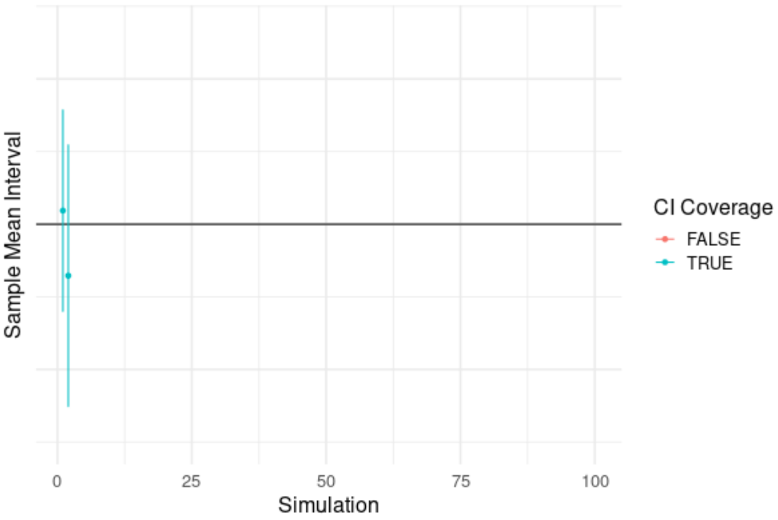
- It was made to present a reasonable range of values for the *parameter* μ as estimated by the *statistic* $\overline{X}$
- The interval was $(40.75, 51.93)$. As our true mean is $\mu = 50$, this interval *is* correct and it *does* contain our true parameter

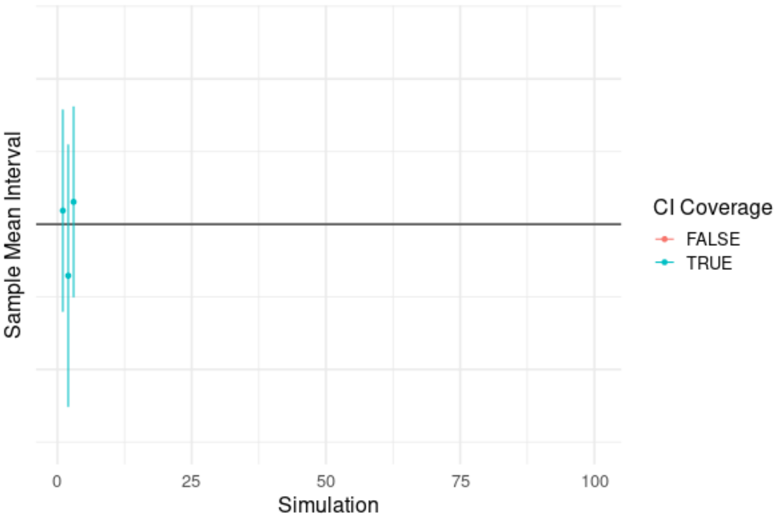
95% Confidence Interval

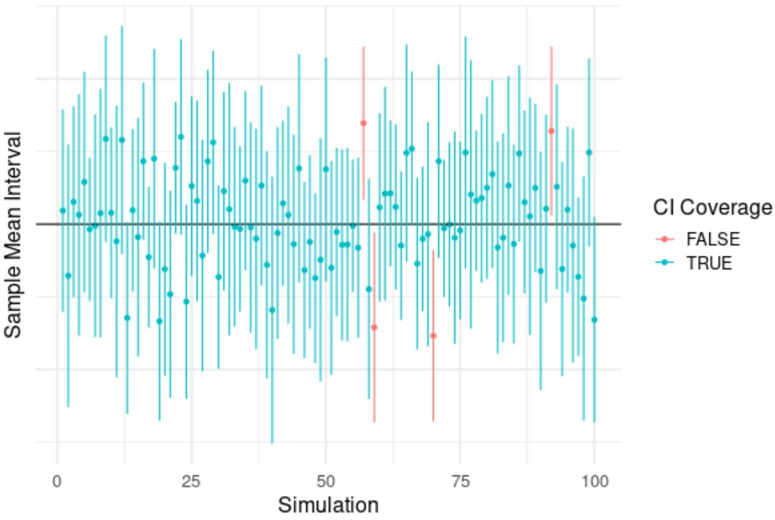
When we say something has a 95% confidence interval, what we mean is:

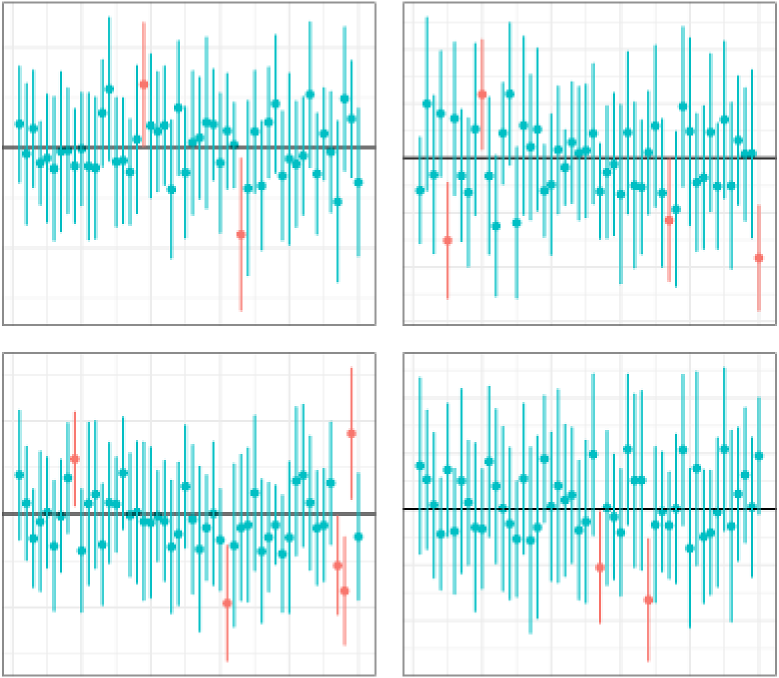
The process that constructed this interval has the property that, on average, it contains the true value of the parameter 95 times out of 100


Sample Mean Interval
0
25
50
75
100
Simulation
CI Coverage
FALSE
TRUE


Sample Mean Interval
0
25
50
75
100
Simulation
CI Coverage
FALSE
TRUE


Sample Mean Interval
0
25
50
75
100
Simulation
CI Coverage
FALSE
TRUE


Sample Mean Interval
0
25
50
75
100
Simulation
CI Coverage
FALSE
TRUE



Confidence Intervals

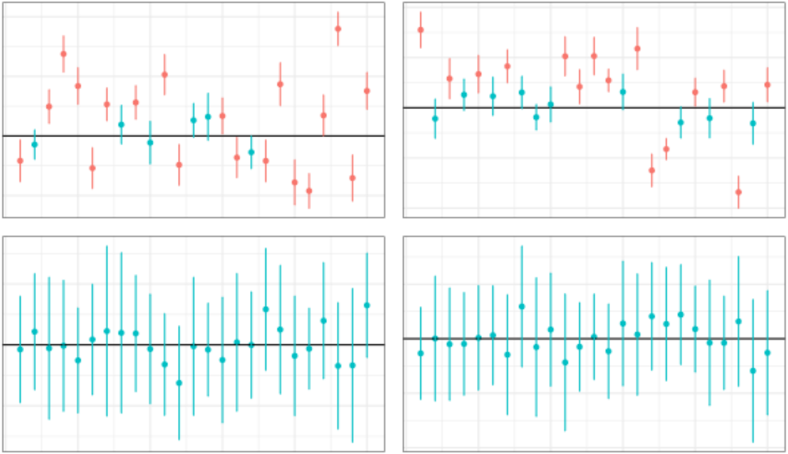
To be absolutely clear: we will *never* know if the confidence interval we construct contains the true value of the parameter

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

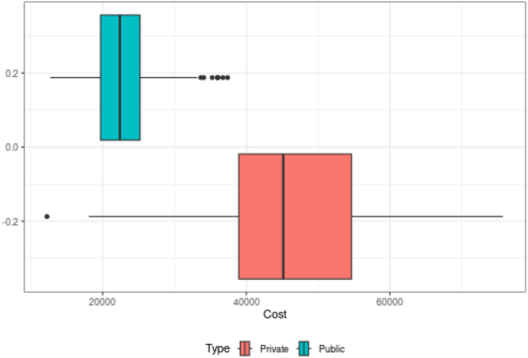
Confidence Intervals

It is also worth observing that we can *alter* our process to acheive different results. There is a tradeoff 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	47073
Public	22766

Example

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

Type	$\overline{X}$	Std. Error
Sample Private (N = 50)	44947	1467
Sample Public (N = 50)	22833	684

$$\begin{aligned}
\text{95\% CI for Private} &= \text{Point estimate} \pm \text{Margin of Error} \\
&= \overline{X} \pm 2 \times \text{SE} \\
&= 44947 \pm 2 \times 1467 \\
&= (42013,\ 47882)
\end{aligned}$$

Example

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

Type	$\overline{X}$	Std. Error
Sample Private (N = 50)	44947	1467
Sample Public (N = 50)	22833	684

$$\begin{aligned}
\text{95\% CI for Public} &= \text{Point estimate} \pm \text{Margin of Error} \\
&= \overline{X} \pm 2 \times \text{SE} \\
&= 22833 \pm 2 \times 684 \\
&= (21464,\ 24201)
\end{aligned}$$

Example

Type	95% Conf Int.	True Mean
Private	(42013, 47882)	47,073
Public	(21464, 24201)	22,766

Review

- **Standard deviation** (σ) 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