Confidence Intervals

Grinnell College

March 6, 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 $\overline{X}$?

Keep in Mind

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:

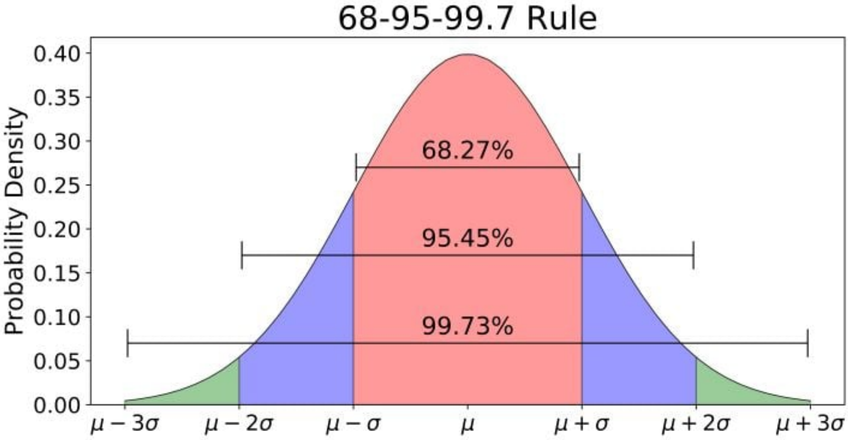
Suppose I have gone out and collected a sample:

- 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 μ and standard deviation σ?

Empirical Rule



Determining Range

If we can determine that the probability of our standardized values being between -1 and 1 to be

$$P(-1 < Z < 1) = 0.68$$

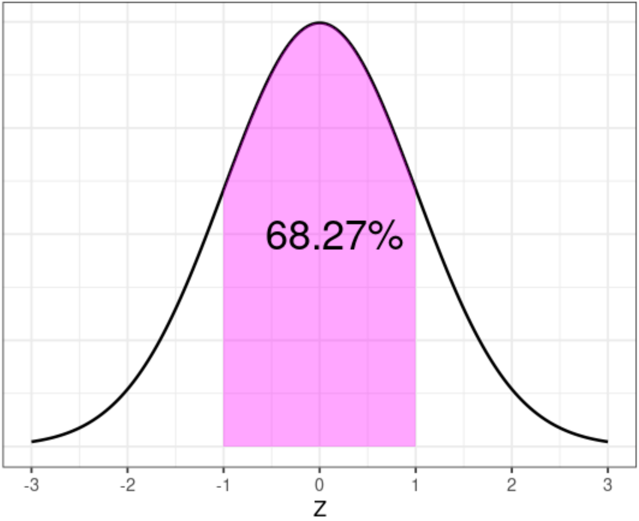
and

$$P(-2 < Z < 2) = 0.9545$$

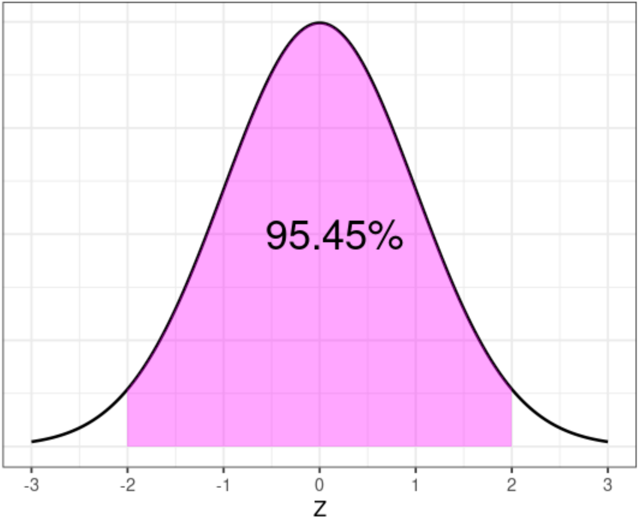
Then theoretically to get the middle $M\%$ of values we should be able to find values C such that

$$P(-C < Z < C) = M\%$$

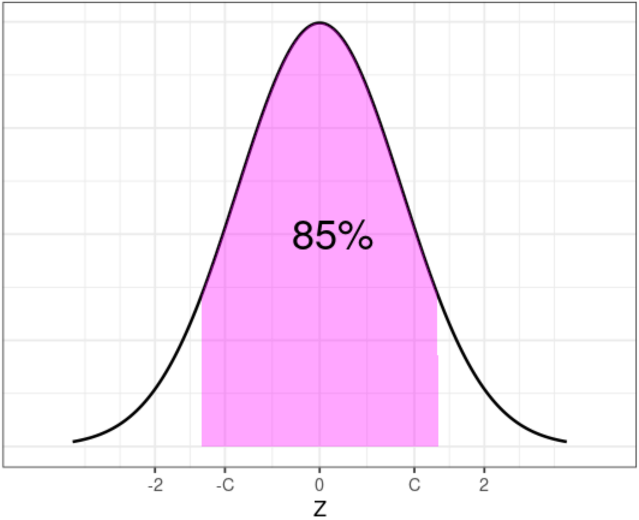
$$P(-1 < Z < 1) = 0.68$$



$$P(-2 < Z < 2) = 0.95$$



$$P(-C < Z < C) = 0.85$$



Critical Values

In general, we can create a symmetric interval around 0 (why?) such that a randomly collected standardized sample mean Z will fall within this interval with some probability

$$P(-C < Z < C) = M$$

The selected probability, M, is defined as our **confidence**; this is the frequency with which our statistic will fall within this interval

The bounds of this interval, C, is called our **critical value**. *The choice of C is the only thing we choose that determines our level of confidence*

Determining Confidence

Again: *The choice of C is the only thing we choose that determines our level of confidence*

Why doesn't the sample size n or the standard error, $\sigma/\sqrt{n}$ impact this width? *Because Z has already been standardized by these values*

Also again: the point is not to find an interval that might contain our sample mean; the point is to use information to construct a plausible interval for the true value of our parameter μ

But we don't know μ

Suppose we wanted to construct an interval that 68% of the time would contain the true parameter μ:

$$\begin{aligned}
68\% &= P(-1 < Z < 1) \\
&= P\left(-1 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1\right) \\
&= P\left(-\frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < \frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(-\overline{X} - \frac{\sigma}{\sqrt{n}} < -\mu < -\overline{X} + \frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(\overline{X} - \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + \frac{\sigma}{\sqrt{n}}\right)
\end{aligned}$$

In general,

$$M\% = P(-C < Z < C)$$

$$= P\left(-C < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < C\right)$$

$$= P\left(-C < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < C\right)$$

$$= P\left(\overline{X} - C\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + C\frac{\sigma}{\sqrt{n}}\right)$$

This gives us a general method for constructing an $M\%$ confidence around μ of the form

$$\overline{X} \pm C \times \frac{\sigma}{\sqrt{n}}$$

A few notes

Given statistics derived from our sample, specifically $\overline{X}$, n, and σ, we are able to construct a range of values for μ of the form

$$\overline{X} \pm \text{Margin of error}$$

where this margin of error is determined by:

1. The sample size
2. The population standard deviation
3. The amount of confidence we wish to have in our interval

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**

Confidence Intervals

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

$$\overline{x} = 46.35, \qquad \hat{\sigma} = 15.281$$

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

$$\begin{aligned} 95\% CI &= \text{Point estimate} \pm \text{Margin of Error} \\ &= \overline{x} \pm 1.96 \times \hat{\sigma}/\sqrt{n} \\ &= 46.35 \pm 1.96 \times 2.79 \\ &= (40.75, 51.93) \end{aligned}$$

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	$\overline{x}$	$\hat{\sigma}$	95% CI	Contains μ
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

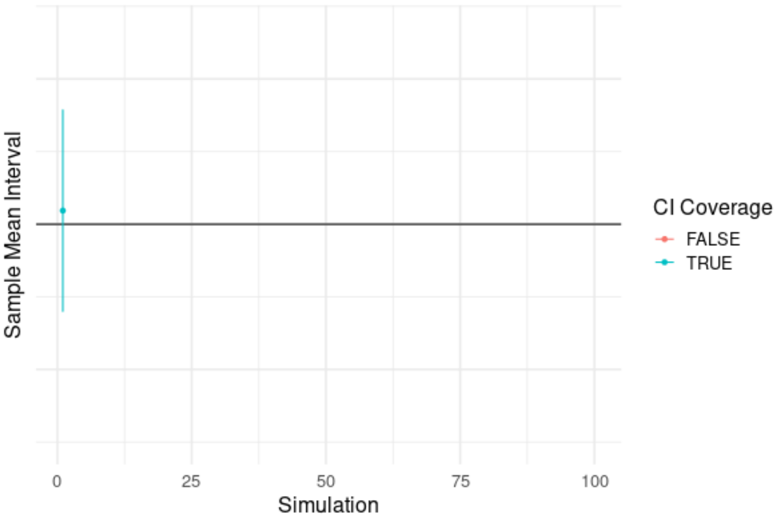
A **confidence interval** is an interval that has the following properties:

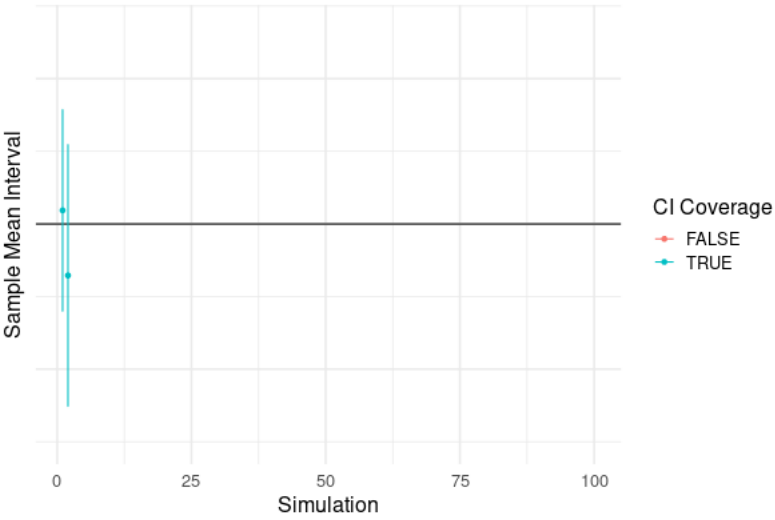
- It is the result of a *random process*
- It is constructed according to a procedure or set of rules
- It is made with the intention of giving a plausible range of values for a *parameter* based on a *statistic*
- There is no probability associated with a confidence interval; *it is either correct or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

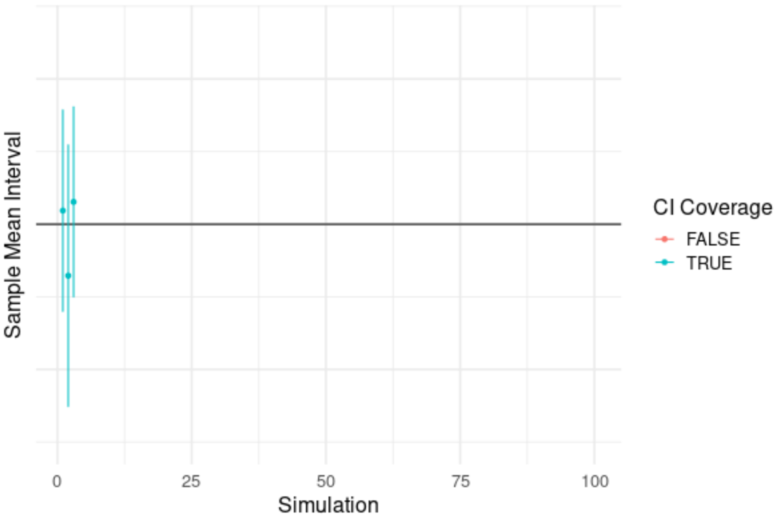
95% Confidence Interval

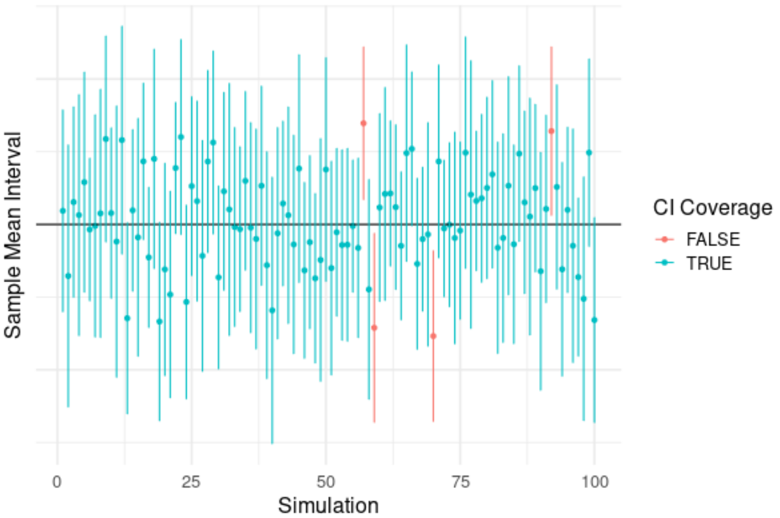
When we say something has a 95% confidence interval, what we mean is:

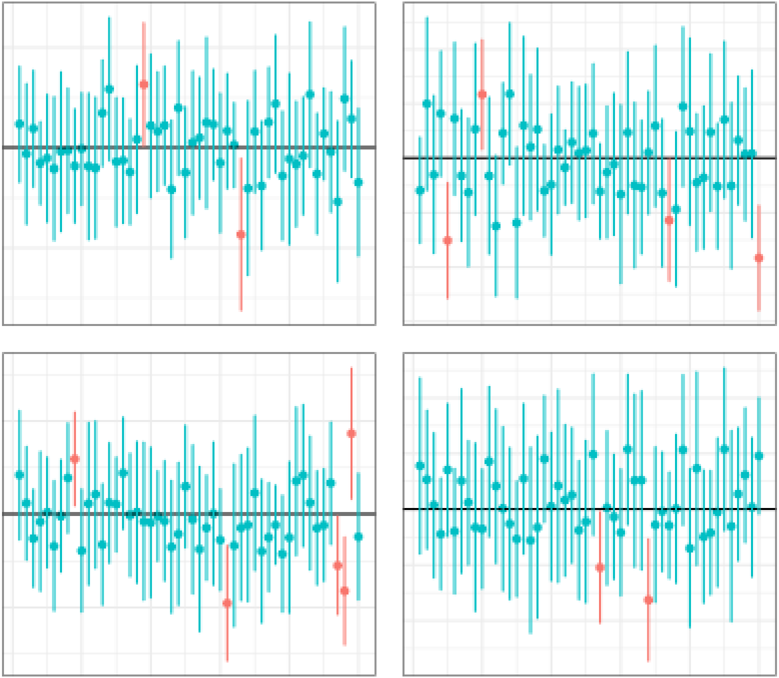
The process that constructed this interval has the property that, on average, it contains the true value of the parameter 95 times out of 100


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

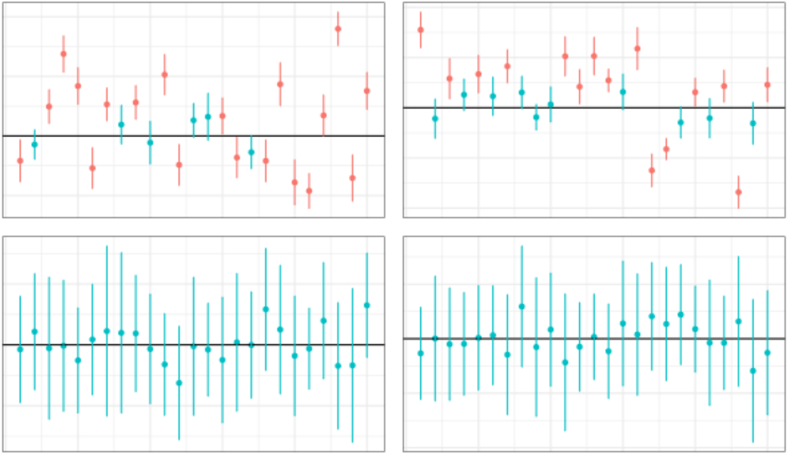
To be absolutely clear: we will *never* know if the confidence interval we construct contains the true value of the parameter

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 achieve different results. There is a trade-off between how frequently we are correct and how much uncertainty we allow in our prediction



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