# Intervals

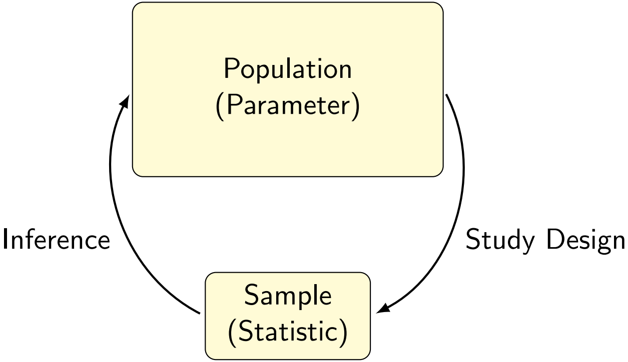
Grinnell College

February 28, 2024

# Review

sampling bias inference variability normal distribution

# The Statistical Framework



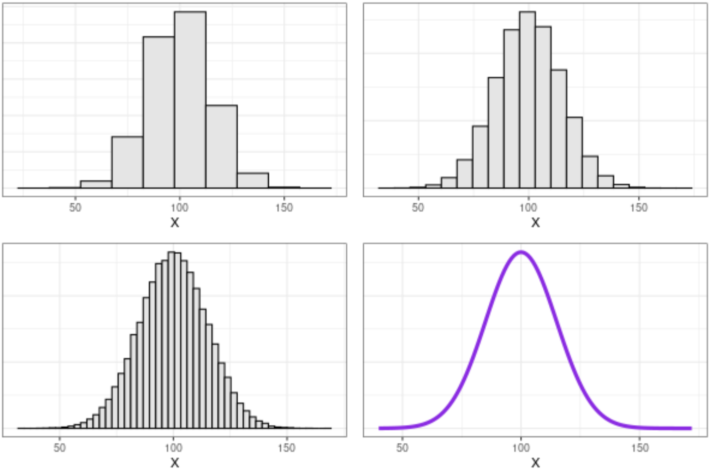
# Some Terms to Know

**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 *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

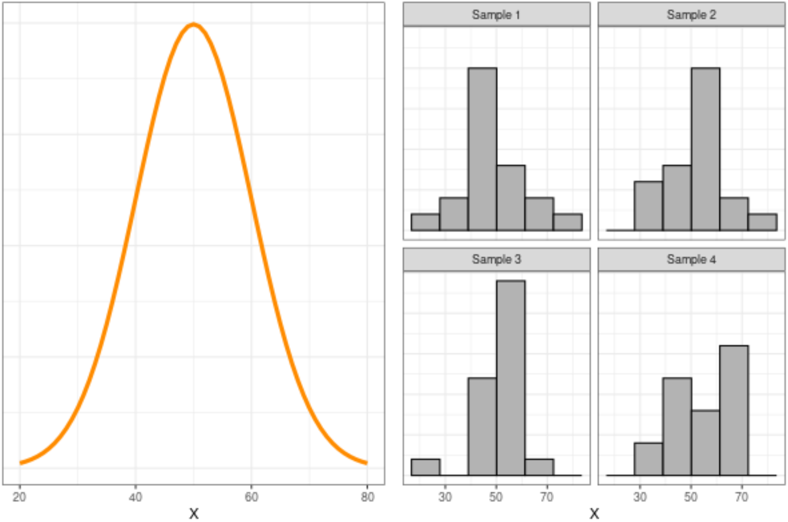
# Normal Distribution

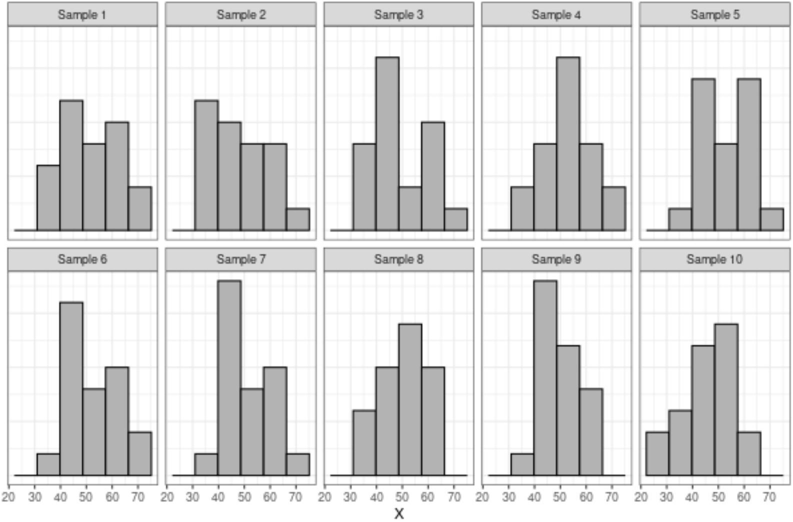
$$X \sim N(\mu, \sigma^2)$$



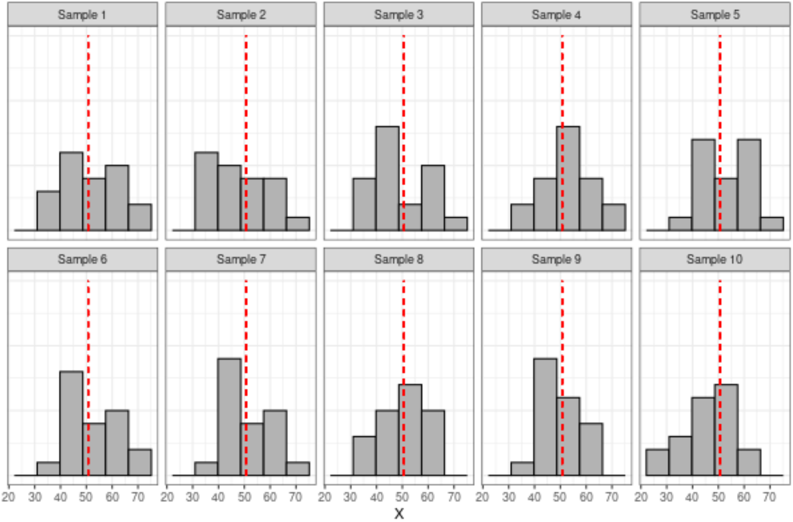
# Representative Samples

Our data $X$ for today is normally distributed with a mean value of 50 and a standard deviation of 15





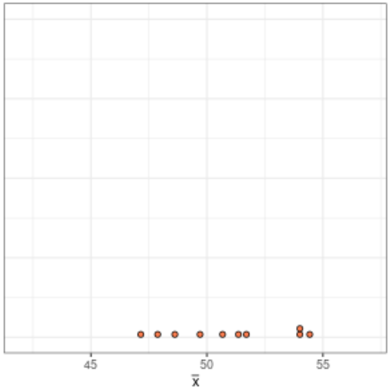
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
Sample 9
Sample 10
20 30 40 50 60 70
X


Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
Sample 9
Sample 10
20 30 40 50 60 70
X

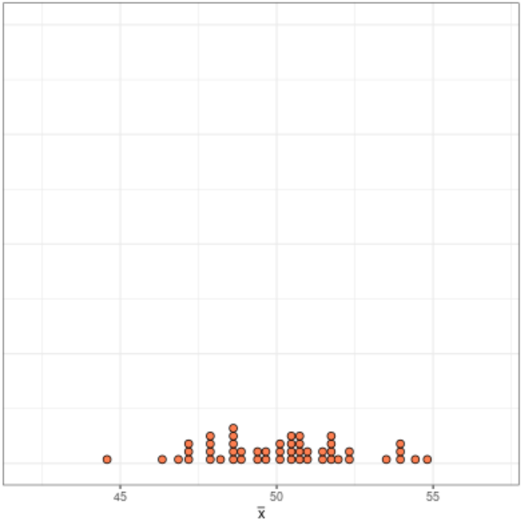
| Sample | Mean |
|---|---|
| Sample 1 | 51.13 |
| Sample 2 | 48.38 |
| Sample 3 | 49.11 |
| Sample 4 | 53.37 |
| Sample 5 | 53.25 |
| Sample 6 | 52.64 |
| Sample 7 | 51.41 |
| Sample 8 | 50.16 |
| Sample 9 | 50.47 |
| Sample 10 | 46.43 |

# 10 Samples

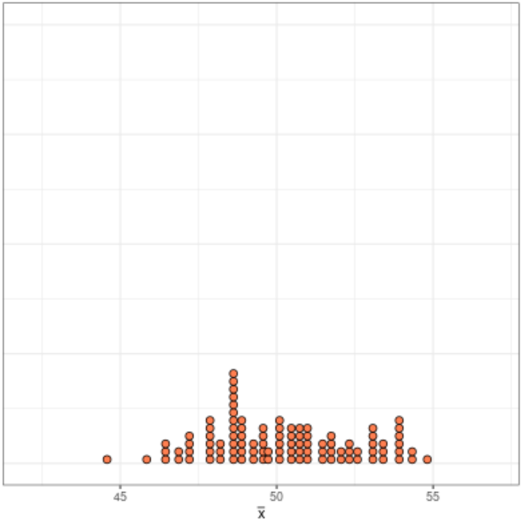
| Sample | Mean |
|---|---|
| Sample 1 | 51.13 |
| Sample 2 | 48.38 |
| Sample 3 | 49.11 |
| Sample 4 | 53.37 |
| Sample 5 | 53.25 |
| Sample 6 | 52.64 |
| Sample 7 | 51.41 |
| Sample 8 | 50.16 |
| Sample 9 | 50.47 |
| Sample 10 | 46.43 |



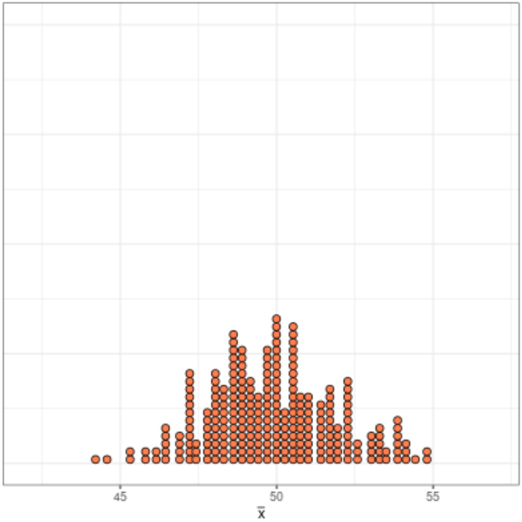
# 50 samples



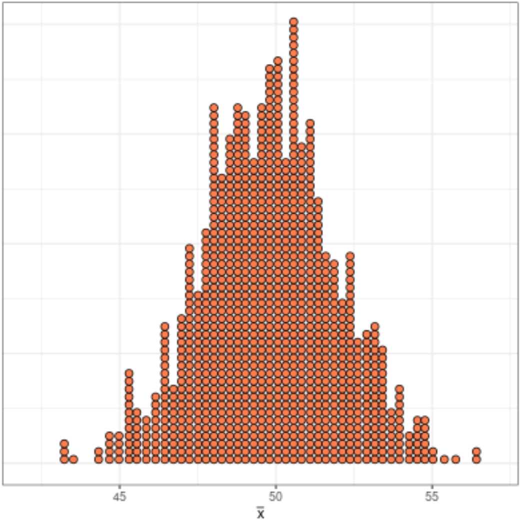
# 100 samples



# 250 samples

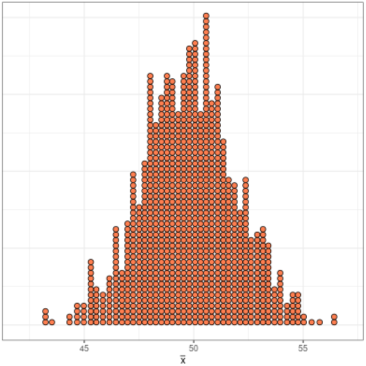


# 1000 samples



# Sampling Distribution

- This is an example of a **sampling distribution**
  - What values?
  - How frequently?
- What can we do with this?



# Sampling Distribution

Specifically, this is a *distribution of sample means*

The same variability that we expect to see in our sampling process is reflected in the distribution of sample means

We will discuss in detail soon:

- How is sampling distribution of $\overline{X}$ related to population distribution of $X$?
- How does sample size impact variability?
- Normality assumptions
- Standard error vs standard deviations

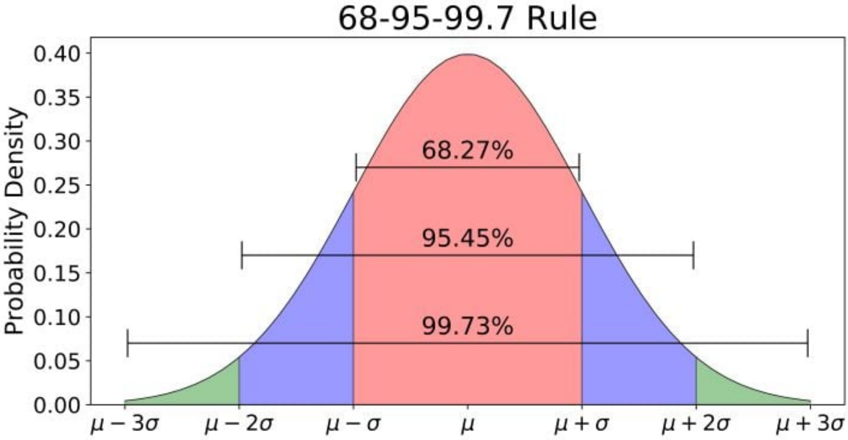
# Intervals

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

If we cannot rely on only our point estimate, $\overline{X}$, perhaps we can find a range of reasonable values:

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

# Good place to start



# 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

Consider Sample 4 from our experiment $X_4$ where we find:

$$\overline{X}_4 = 46.35, \qquad \text{SE}_4 = 2.79$$

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

$$\begin{aligned} 95\% CI &= \text{Point estimate} \pm \text{Margin of Error} \\ &= \overline{X}_4 \pm 2 \times \text{SE}_4 \\ &= 46.35 \pm 2 \times 2.79 \\ &= (40.75, 51.93) \end{aligned}$$

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 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

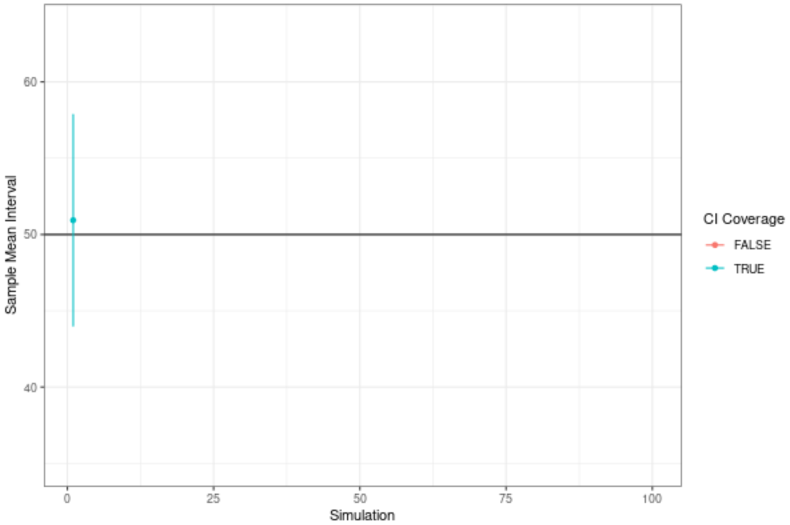
Consider the confidence interval that we constructed on a previous slide from Sample 4:

- It was constructed according to the procedure
  Point estimate $\pm$ Margin of Error
- It was made to present a reasonable range of values for the *parameter* $\mu$ 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

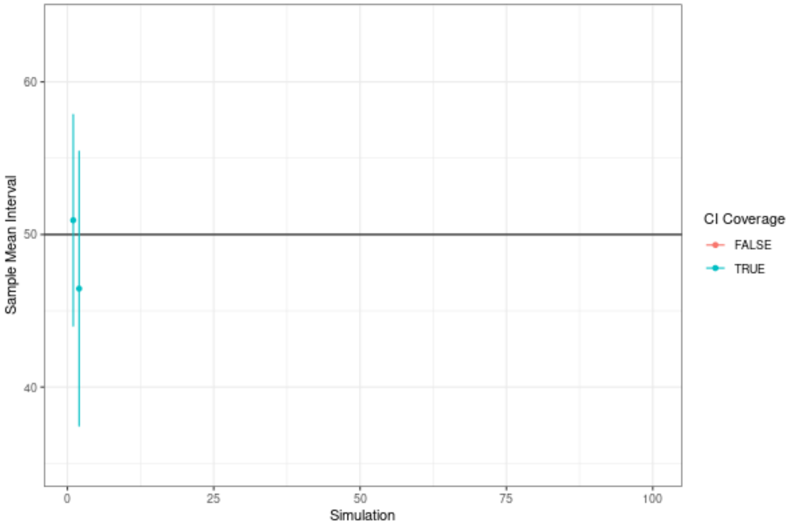
# 95% Confidence Interval

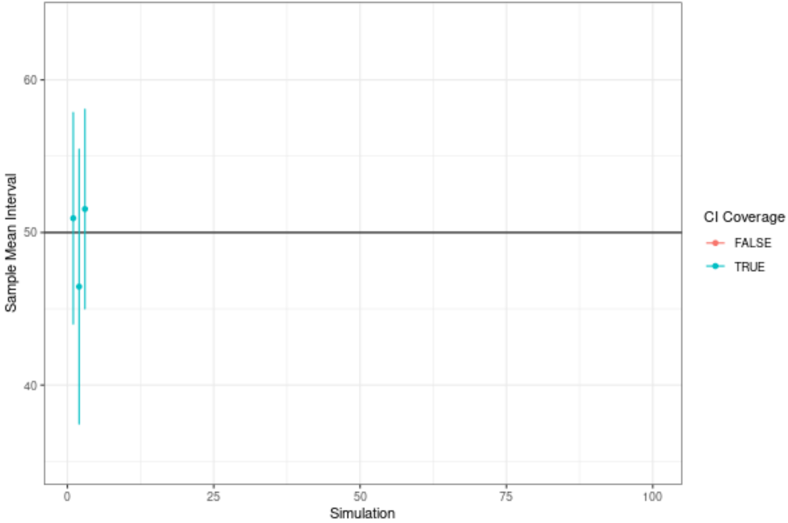
When we say something has a 95% confidence interval, what we mean is:

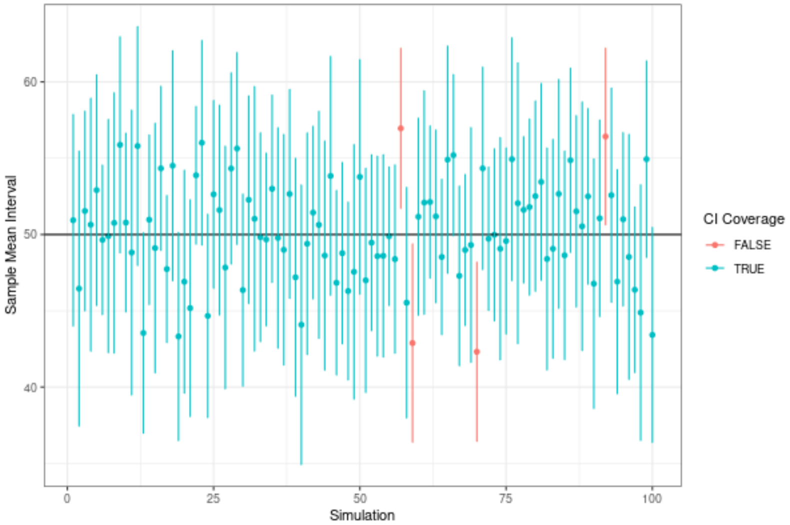
*The process that constructed this interval has the property that, on average, it contains the true value of the parameter 95 times out of 100*



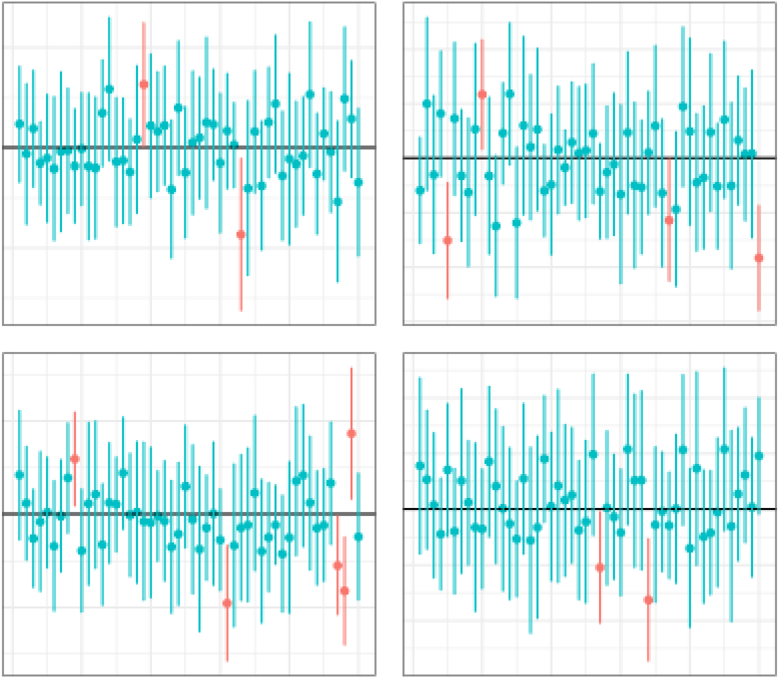
60
50
40
Sample Mean Interval
0
25
50
75
100
Simulation
CI Coverage
FALSE
TRUE


Sample Mean Interval
60
50
40
0
25
50
75
100
Simulation
CI Coverage
FALSE
TRUE


60
50
40
0
25
50
75
100
Sample Mean Interval
Simulation
CI Coverage
FALSE
TRUE



60
50
40
Sample Mean Interval
0
25
50
75
100
Simulation
CI Coverage
FALSE
TRUE



# Confidence Intervals

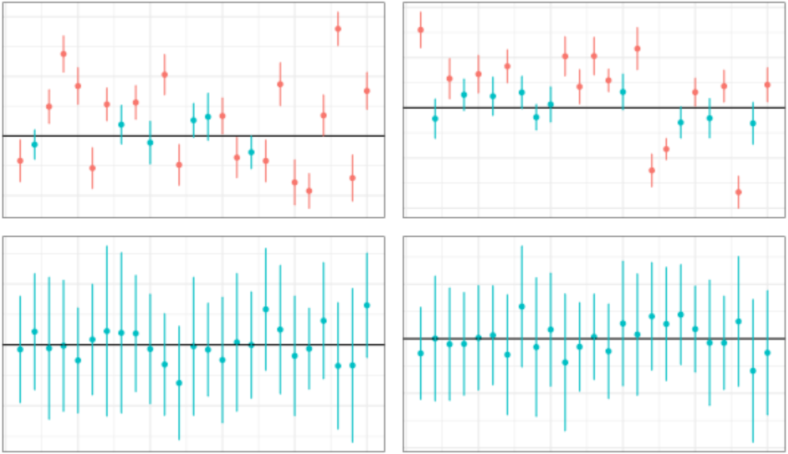
To be absolutely clear: we will *never* know if the confidence interval we construct contains the true value of the parameter

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

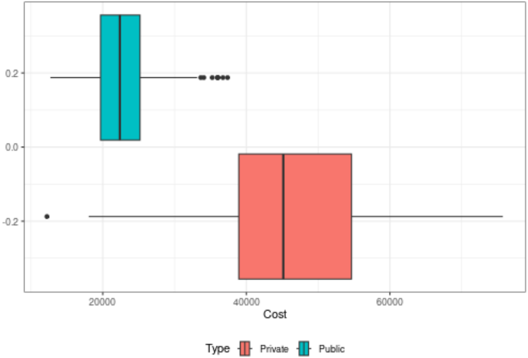
# Confidence Intervals

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** ($\sigma$) 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
- Many sampling distributions 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

Break

# Repeated Samples

The confidence intervals that we constructed of the form

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

relied on assumptions made about our *sampling process* and examined what we might expect if we could repeat it ad infinitum

We are often limited to collecting only a single sample from our population

Further, the degree to which our assumptions are met may be tenuous

# Repeated Samples

The confidence intervals we constructed of the form:

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

- Relied on assumptions (TBD) about our *sampling process*
- Examined what might happen if we could repeat sampling ad infinitum

There are, naturally, some limitations:

- We are limited to collecting a single sample
- Our assumptions may be tenuous

It would be helpful to have a more general method of constructing intervals with similar properties we had before

# Bootstrapping

Somewhat amazingly, we can get around this problem with a technique known as **bootstrapping**

Instead of drawing more samples from our *original population*, we treat our sample as an *estimate* of the population and instead draw bootstrapped samples from our original sample

# Bootstrapping

"Pick yourself up from bootstraps"