Statistics Introduction

Instructor

Lester Leung

CFI Instructor

Data Science Professional

- Leading a startup fintech data science team
- Holding VP roles in quantitative analysis and strategy at Wells Fargo
- Various data consulting roles in other Fortune 500 companies

What is Statistics?

Statistics are applied all around us:

- Search Results
- Inventory
- Weather Forecasting
- Election Forecasting
- Customer Research
- Market Analysis
- Budget Planning
- Scenario Exploration
- Risk Management
- Machine Learning

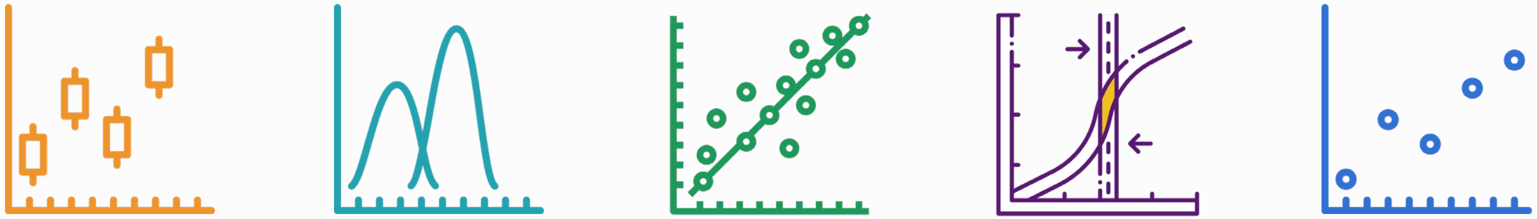
What is Statistics?

Statistics is a branch of mathematics that deals with the collection, analysis and interpretation, and presentation of data.

Collection → **Analysis and Interpretation** → **Presentation**

What is Statistics

The Statistical Toolbox



Statistical tools can be applied to:

- Make predictions
- Measure uncertainty
- Evaluate claims
- Make improvements to systems
- Inform or educate us about a population

Course Objectives

Descriptive Statistics

Key statistics used for describing data sets and popular visualizations

Central Limit Theorem and Law of Large Numbers

An overview the central limit theorem and law of large number—two critical statistical principles

Constructing a Sample

Populations, samples, and strategies for preventing bias

Hypothesis Testing

Performing a two-tail and one-tail hypothesis test

Error Estimations

Estimating the possibilities of false positives or false negatives

Effect Size & Power Analysis

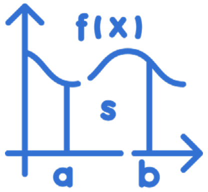
Estimating effect sizes and conducting a power analysis to determine a better sample size

Why Use Statistics?

Challenges with Statistics

Statistics requires collecting unbiased samples ,applying mathematical principles to analyze data, and rarely provides perfectly precise answers—often it leaves uncertainty which can be expressed as a margin of error or confidence level.

Finding Good Samples

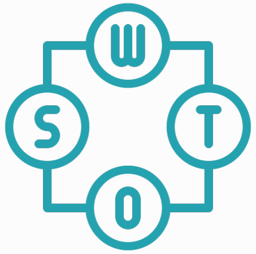


Applying Complex Principles

Still Being Unsure

Other Approaches to Decision Making

Statistics is a tool, but its one tool of many. Business leaders might at times prefer rely on a conceptual framework like SWOT analysis, or their gut instinct, or experience. These can sometimes reach decisions faster and with less resources spent.



Strengths of Using a Statistical Approach

Statistics are applied math. As math, they can easily be replicated, shared, and critiqued by others.

When used properly, statistics can greatly reduce the influence of bias.

Statistical models are easy to tweak or fine-tune over time as new data becomes available.

Statistic models can be generated and updated using automation for increased speed and scale.

The Branches of Statistics

Statistics

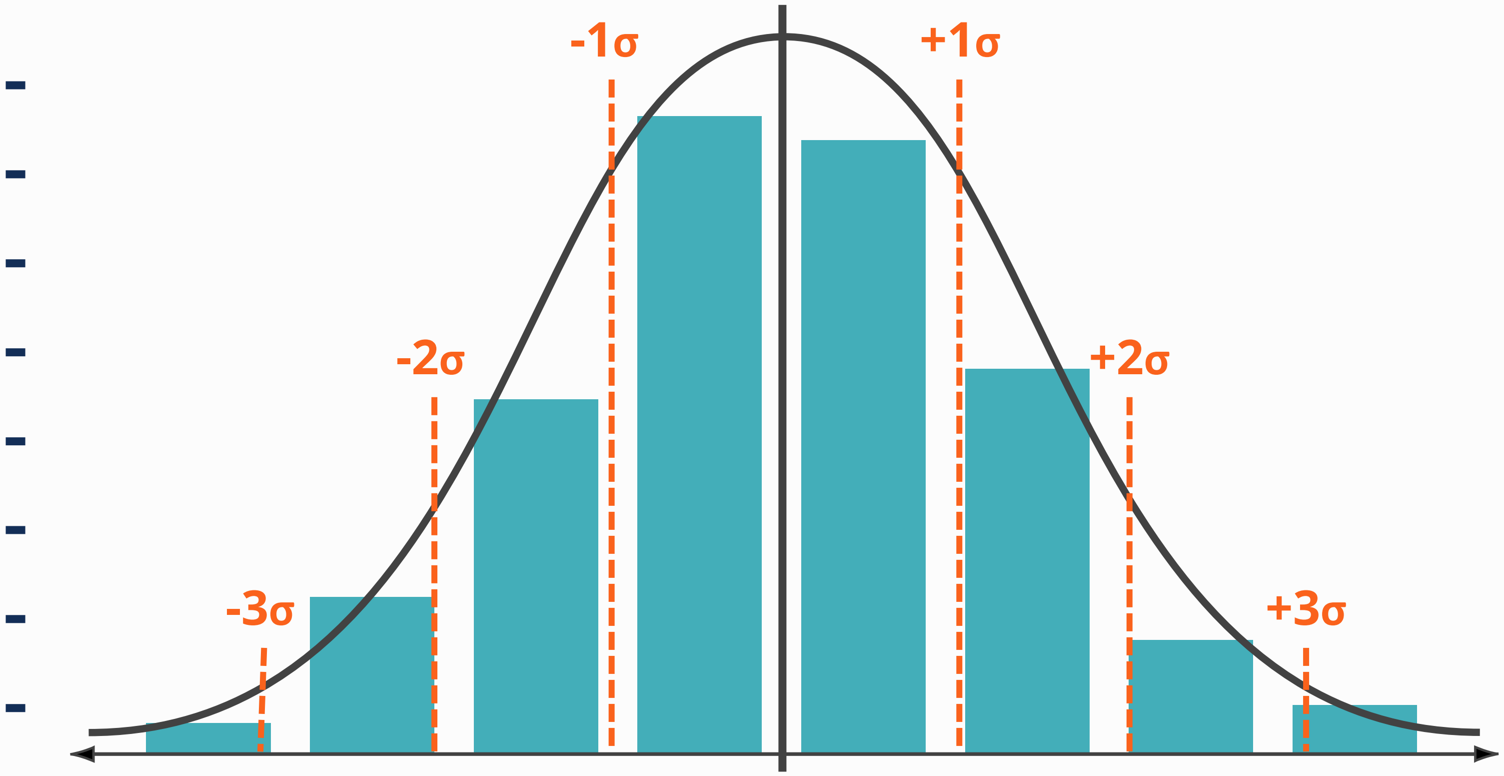
Descriptive Statistics

- Describes or summarizes data
- **Examples:** mean or variance

Inferential Statistics

- Infers a characteristic of a population
- **Examples:** probability and confidence measures

Applying Inferential Statistics



Chapter 2: Descriptive Statistics

Scenario

Scenario

Scenario

The CEO wants a 10-minute summary of the last 3 years of an acquired company's sales data.

Goals

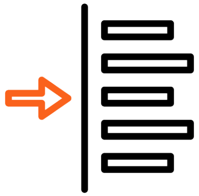
Keep it quick.

Make it visual.

Select the appropriate summaries.
Don't misrepresent the data.

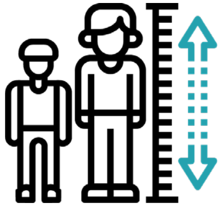
Descriptive Statistics

Descriptive statistics broadly describe data through single values.



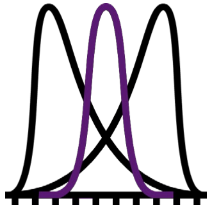
Measures of Central Tendency

- Mean
- Median
- Mode



Measures of Dispersion

- Range
- Variance
- Standard deviation



Shapes of Distribution

- Skewness
- Kurtosis

Course Objectives

Measures of Central Tendency

Explore measures that summarize the central point of our data

Measures of Dispersion

Explore measures that summarize the dispersion of our data.

Shapes of Distribution

Explore measures that summarize the shape of the distribution of our data.

Excel Data Analysis

Find measures through Excel's Data Analysis ToolPak

Visualizations

Learn basic visualizations used commonly in statistics.

Measures of Central Tendency

Measures of Central Tendency

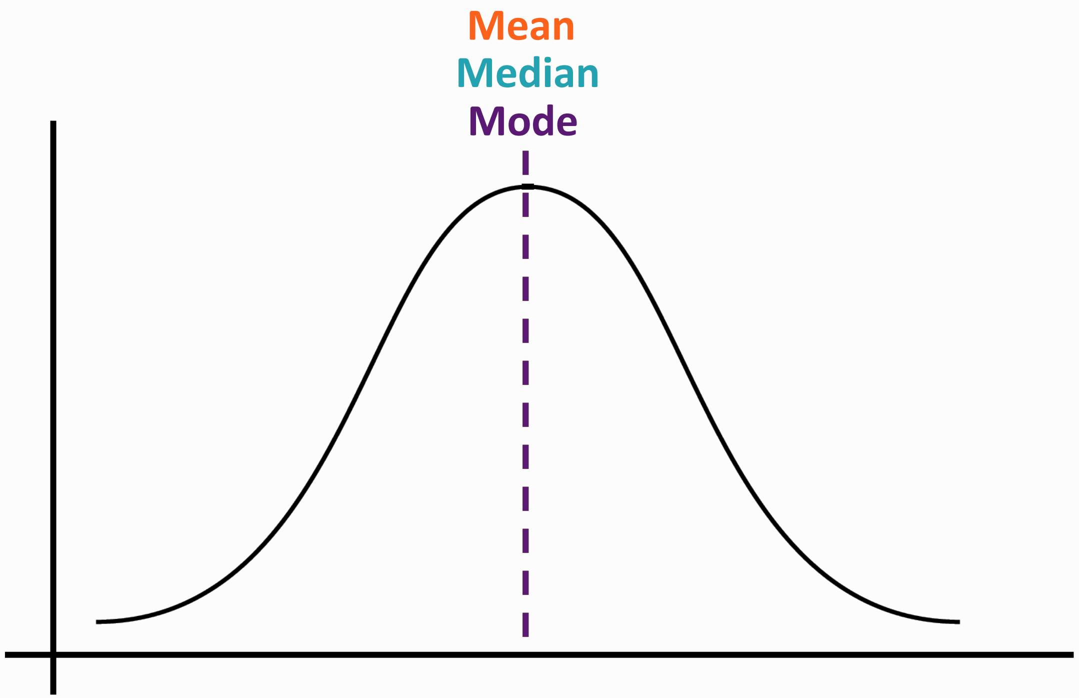
Measures of central tendency are single values that attempt to describe the central position of a set of data.

1 6 13.57

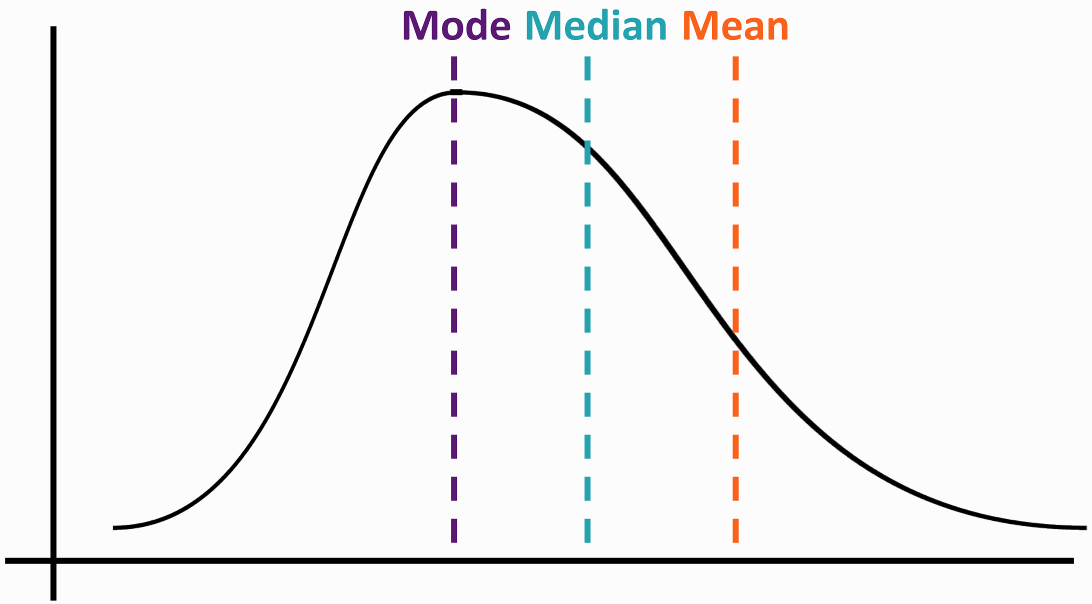
0, 1, 1, 1, 4, 5, 6, 6, 7, 10, 10, 27, 42, 70

Mode Median Mean

Measures of Central Tendency



Mean, Median, Mode under normal distribution



Mean, Median, Mode under skewed distribution

Measures of Central Tendency

Mean (Average):

- Used a lot (for example, in our t-tests in Chapter 4)
- Is most meaningful when we have data that is relatively normally distributed

Use cases:

- Average performance over three years is a popular measure of company performance
- Expected default rates for loans

Measures of Central Tendency

Median:

- Diminishes the effect of outliers

Use cases:

- Employee salaries
- S&P Daily Returns

Measures of Central Tendency

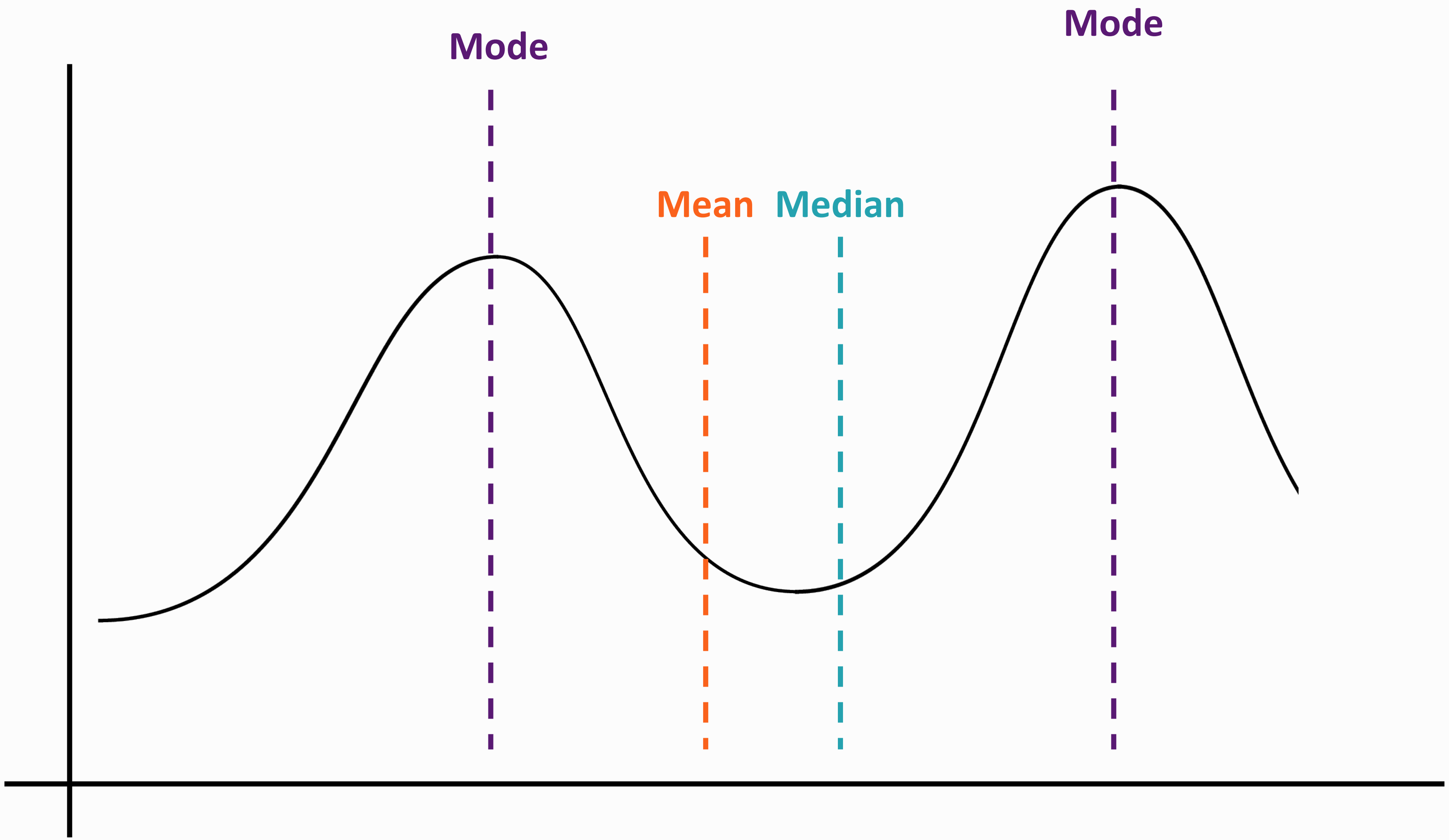
Mode:

- Best used when a curve is apparent
- Shouldn't be applied to small datasets or flat distributions
- With a good dataset, it will describe the most likely outcome

Use cases:

- When dealing with bi-modal distributions
  - Call-center busy hours
  - Restaurant busy hours
  - Marathon organizers
  - Businesses dealing with seasonality
- When dealing with categorical data (e.g., measuring favorite colors)

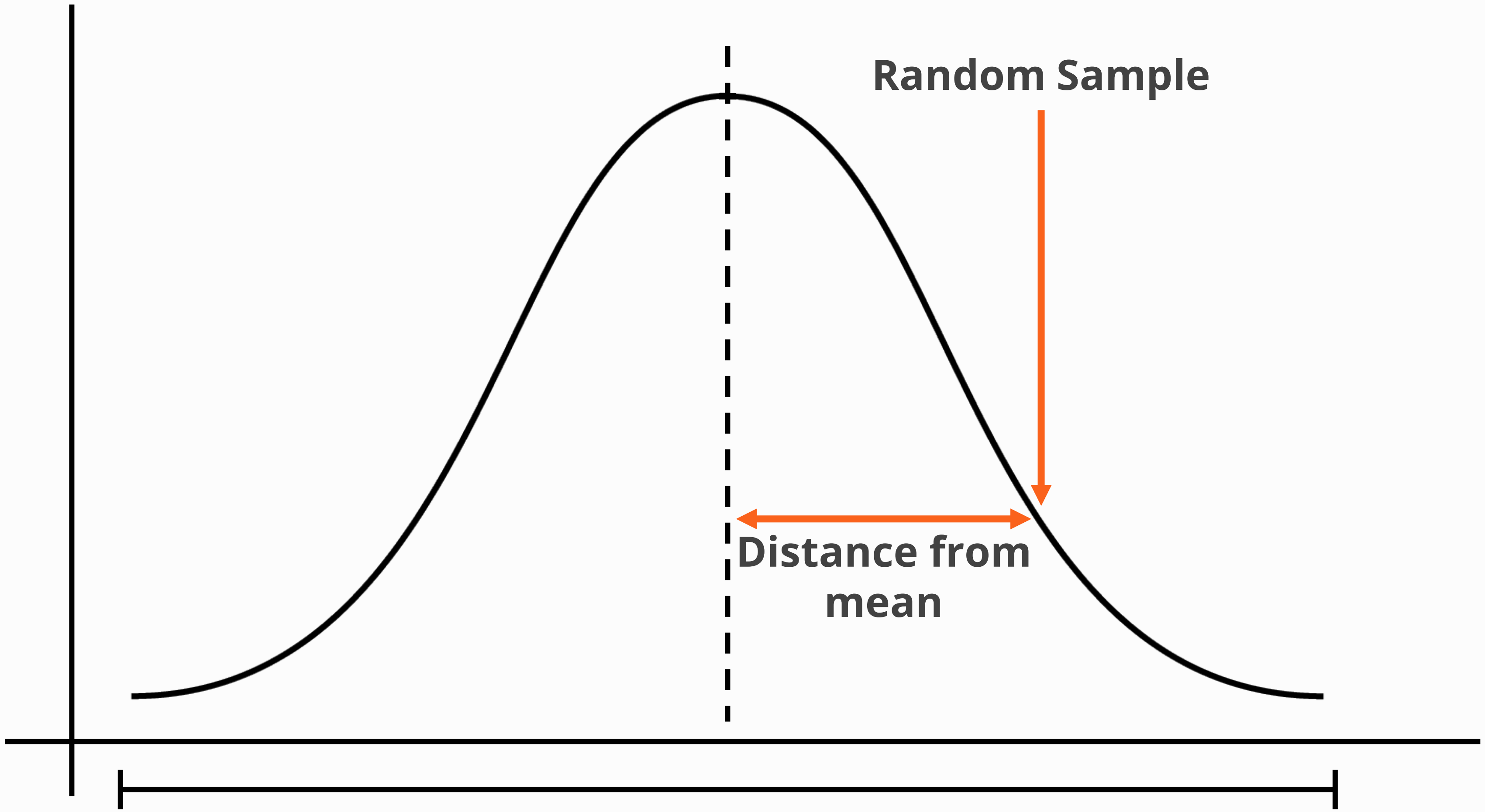
Measures of Central Tendency



Measures of Dispersion

Measure of Dispersions

Measures of dispersion describe how much our data is either spread out or squeezed. The two most important dispersions we will cover are variance and standard deviation.



Minimum, Maximum, and Range

The simplest measures of dispersion are the **minimum, maximum,** and **range**.

Minimum

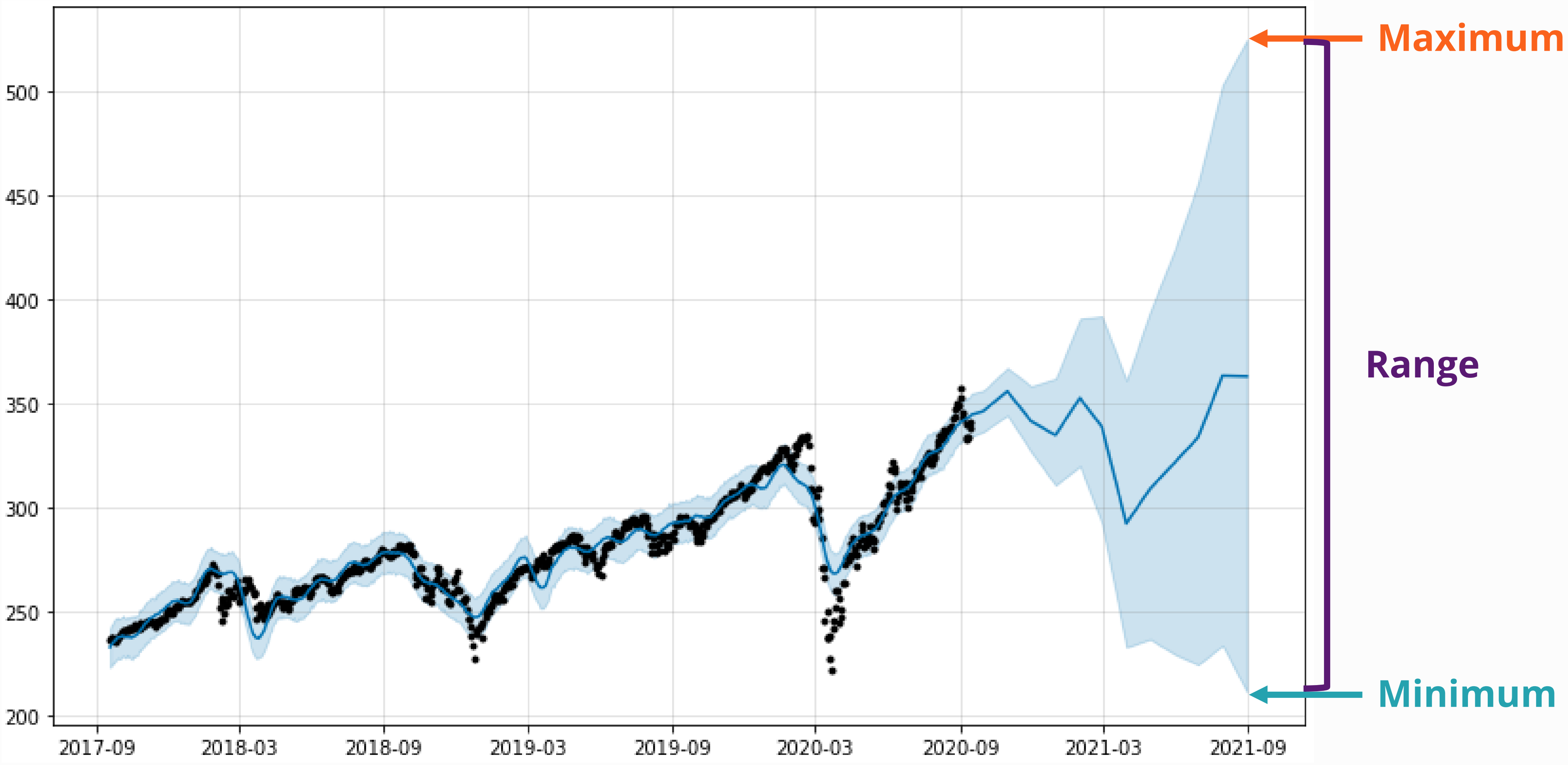
Maximum

0, 1, 1, 1, 4, 5, 6, 6, 7, 10, 10, 27, 42, 70

70

Range

Range (Monte Carlo)



Quartiles

Quartiles are another way of dividing distributions. Whereas median is just the middle of our data, quartiles take the middle of that middle—dividing the data into four quartiles.

0	1	6	10	70
Min	Q1	Median / Q2	Q3	Max

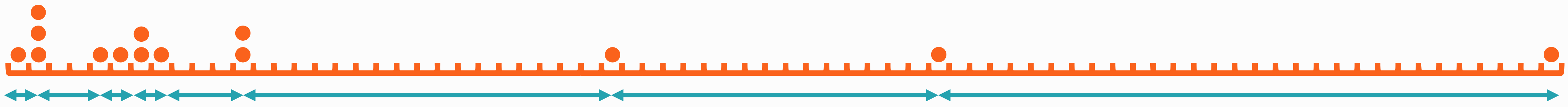
0, 1, 1, 1, 4, 5, 6, 6, 7, 10, 10, 27, 42, 70

1st Quartile | 2nd Quartile | 3rd Quartile | 4th Quartile

Variance

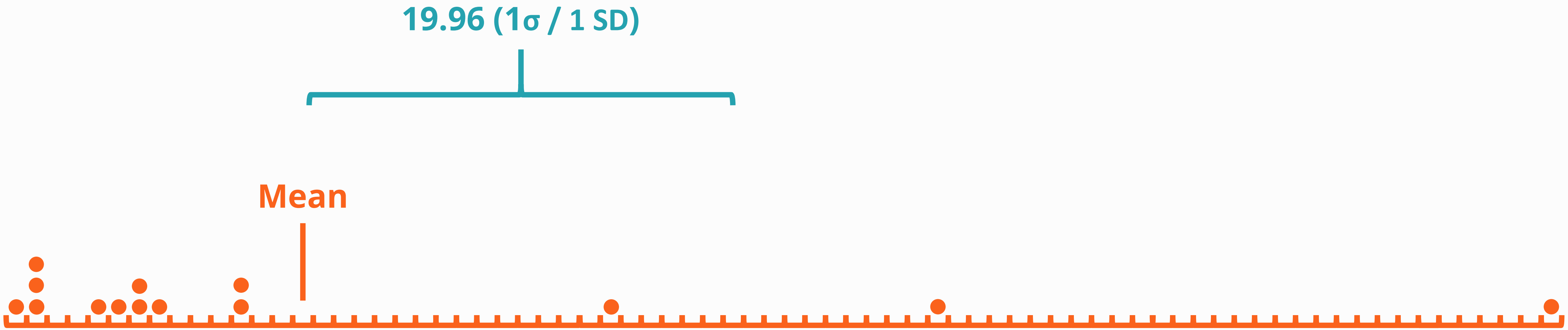
Variance is a measure of how spread-out all our data points are. A high variance suggests a high volatility, while a low variance suggests a low volatility.

0, 1, 1, 1, 4, 5, 6, 6, 7, 10, 10, 27, 42, 70

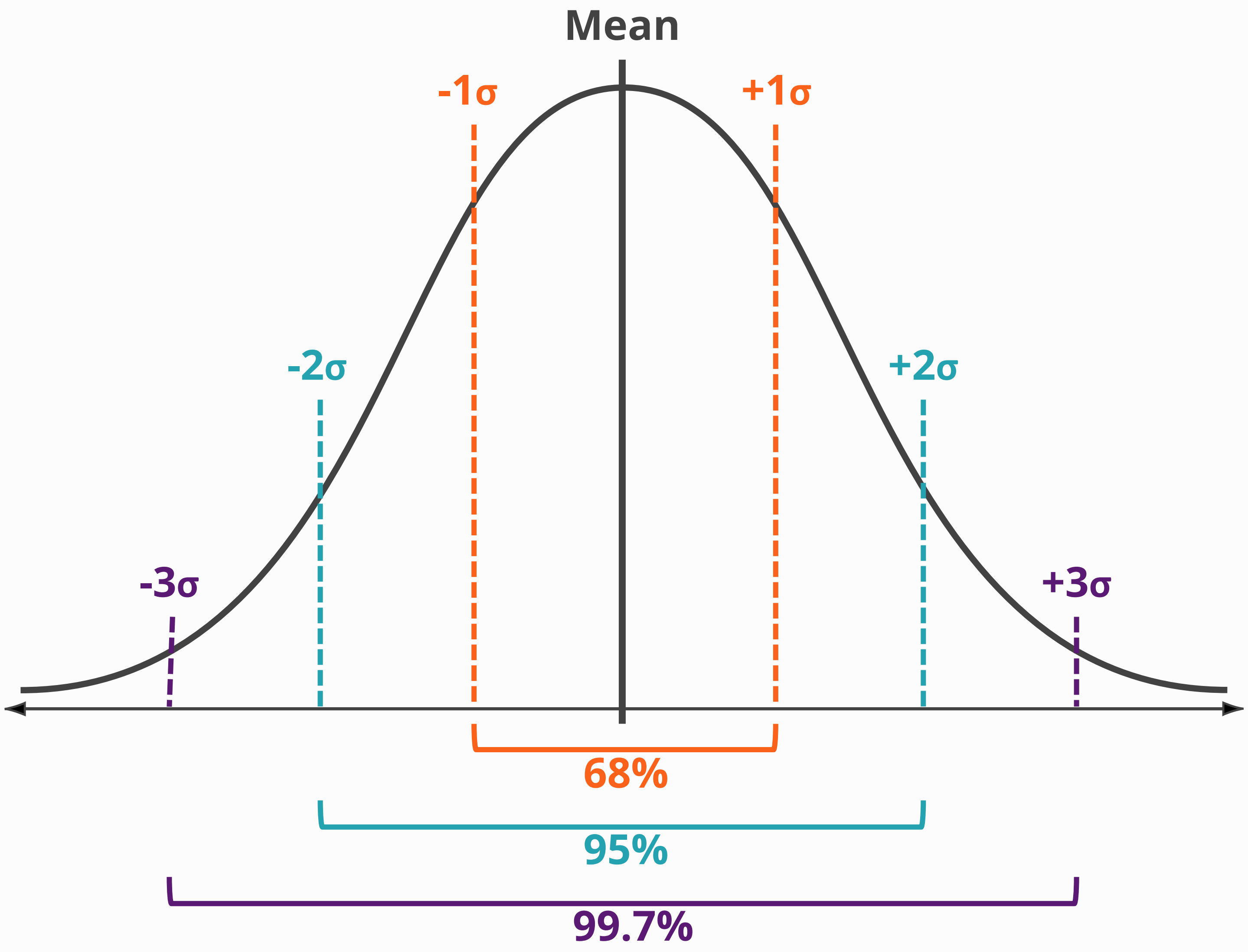


Standard Deviation

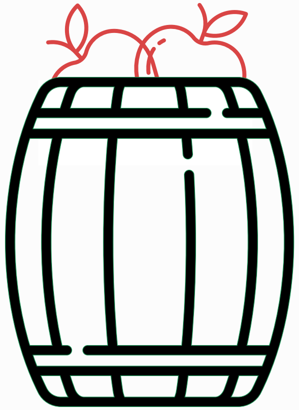
Standard deviation (σ) is our last measure of spread. It's very commonly used in statistics. The standard deviation describes the average distance we can expect an individual data point—or sample—to fall from the mean.



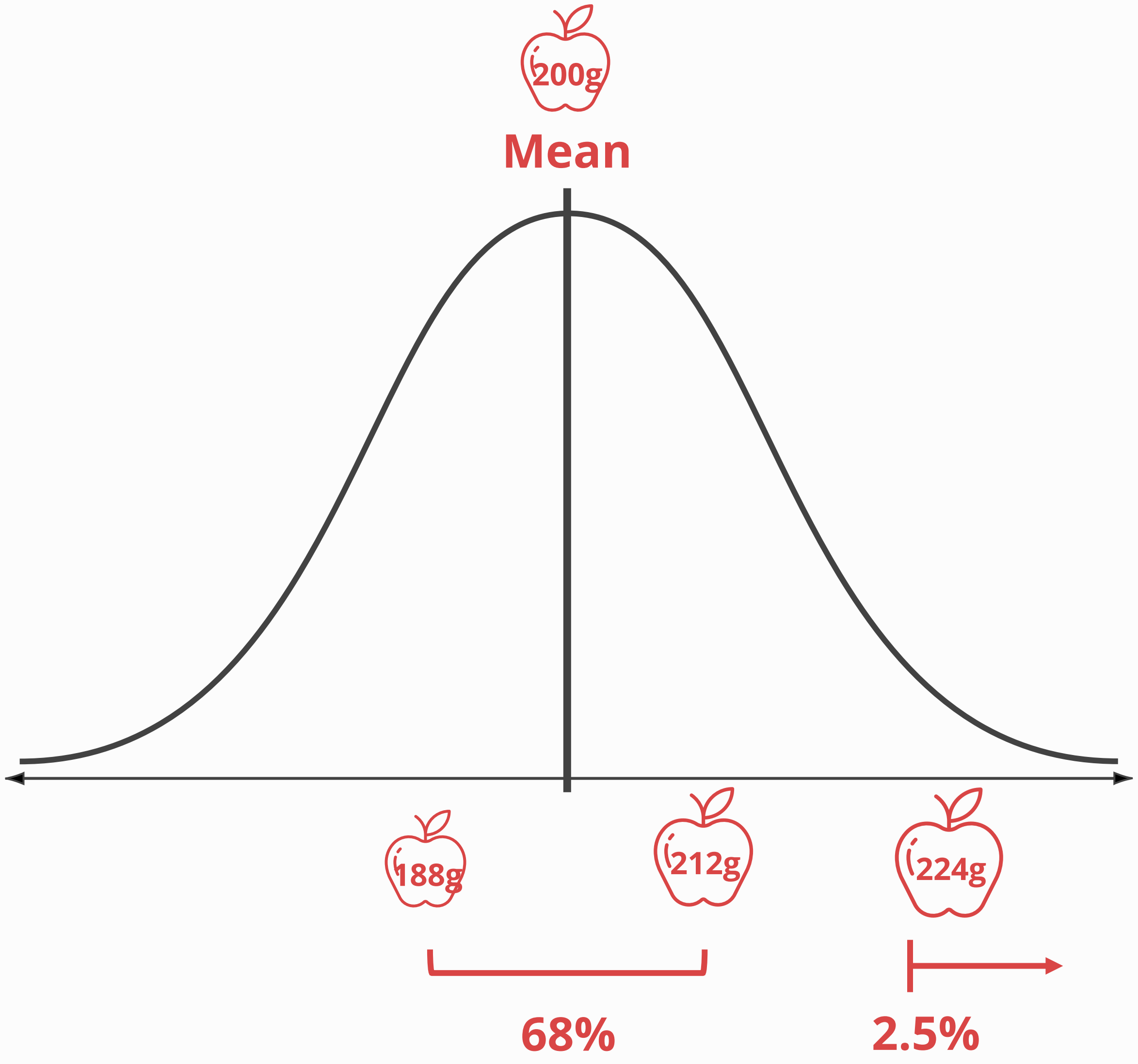
Standard Deviation and the Empirical Rule



Standard Deviation Applied to a Bucket of Apples



Standard Deviation = 12



Calculating Variance and Standard Deviation

Calculating Variance | Step 1

1 **Find the deviation from the mean [Sample – Mean] for each data point**

Samples	Deviation	Deviation^2
0	0 – 13.57 =	
1	1 – 13.57 =	
1	1 – 13.57 =	
1	1 – 13.57 =	
4	4 – 13.57 =	
5	5 – 13.57 =	
6	6 – 13.57 =	
6	6 – 13.57 =	
7	7 – 13.57 =	
10	10 – 13.57 =	
10	10 – 13.57 =	
27	27 – 13.57 =	
42	42 – 13.57 =	
70	70 – 13.57 =	

Calculating Variance | Step 1

1 Find the deviation from the mean [Sample – Mean] for each data point

Samples	Deviation	Deviation^2
0	-13.57	
1	-12.57	
1	-12.57	
1	-12.57	
4	-9.57	
5	-8.57	
6	-7.57	
6	-7.57	
7	-6.57	
10	-3.57	
10	-3.57	
27	13.43	
42	28.43	
70	56.43	
SUM	0	

Calculating Variance | Step 2

1. Find the deviation from the mean [Sample – Mean] for each data point
2. **Square the deviations**

Samples	Deviation	$Deviation^2$
0	-13.57	**184.17**
1	-12.57	**158**
1	-12.57	**158**
1	-12.57	**158**
4	-9.57	**91.58**
5	-8.57	**73.44**
6	-7.57	**57.3**
6	-7.57	**57.3**
7	-6.57	**43.16**
10	-3.57	**12.74**
10	-3.57	**12.74**
27	13.43	**180.36**
42	28.43	**808.26**
70	56.43	**3,184.34**

Calculating Variance | Step 3

1. Find the deviation from the mean [Sample – Mean] for each data point
2. Square the deviations
3. **Sum the squared deviations**

Samples	Deviation	Deviation^2
0	-13.57	184.17
1	-12.57	158
1	-12.57	158
1	-12.57	158
4	-9.57	91.58
5	-8.57	73.44
6	-7.57	57.3
6	-7.57	57.3
7	-6.57	43.16
10	-3.57	12.74
10	-3.57	12.74
27	13.43	180.36
42	28.43	808.26
70	56.43	3,184.34
	SUM	**5,179.43**

Calculating Variance | Step 4

1. Find the deviation from the mean [Sample – Mean] for each data point
2. Square the deviations
3. Sum the squared deviations
4. **Divide sum by the number of data points (n) – 1**

$$\frac{5{,}179.43}{14-1} = Variance(\sigma^2) = \mathbf{398.42}$$

Calculating Standard Deviation

The standard deviation equals the square root of our variance and is represented by a lowercase sigma. Sometimes it is even referred as the ‘sigma’.

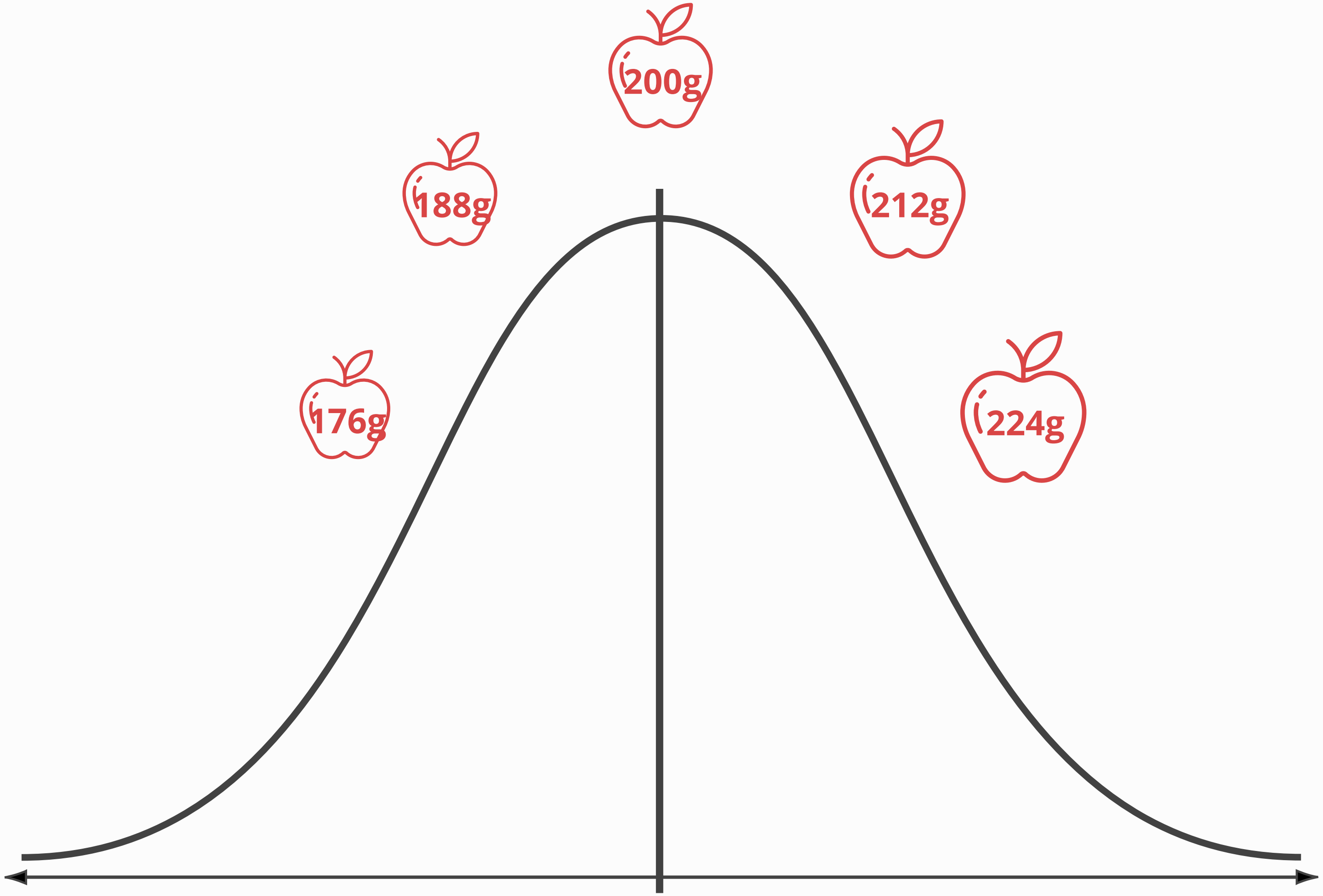
Often you’ll see the variance written simply as the standard deviation squared. If we square our standard deviation, we get right back to our variance.

$$\textit{Standard Deviation}\ (\sigma) = \sqrt{\sigma^2} = \sqrt{398.42} = 19.96$$

$$\textit{Variance}\ (\sigma^2) = 19.96^2$$

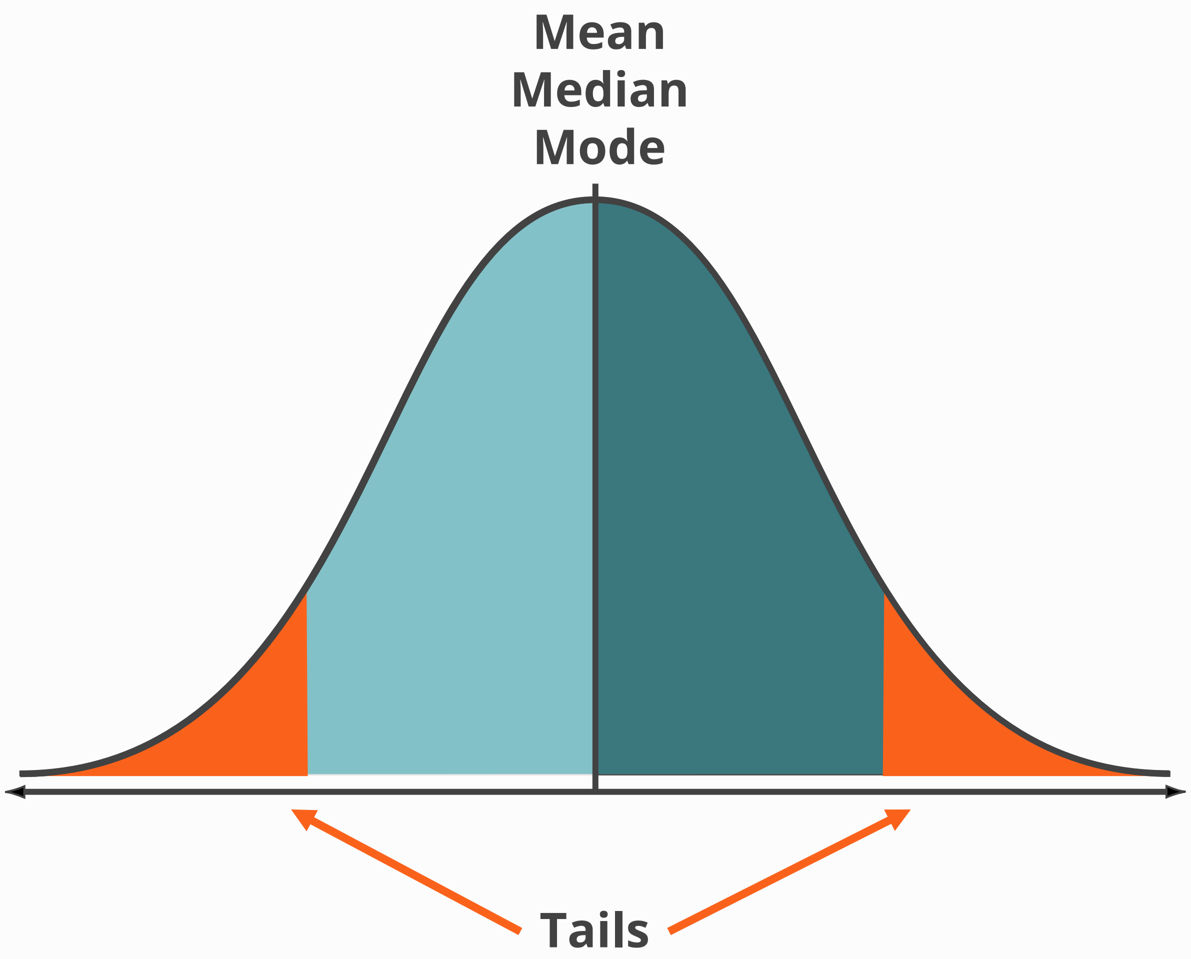
Shapes of Distribution

Normal (Gaussian) Distribution



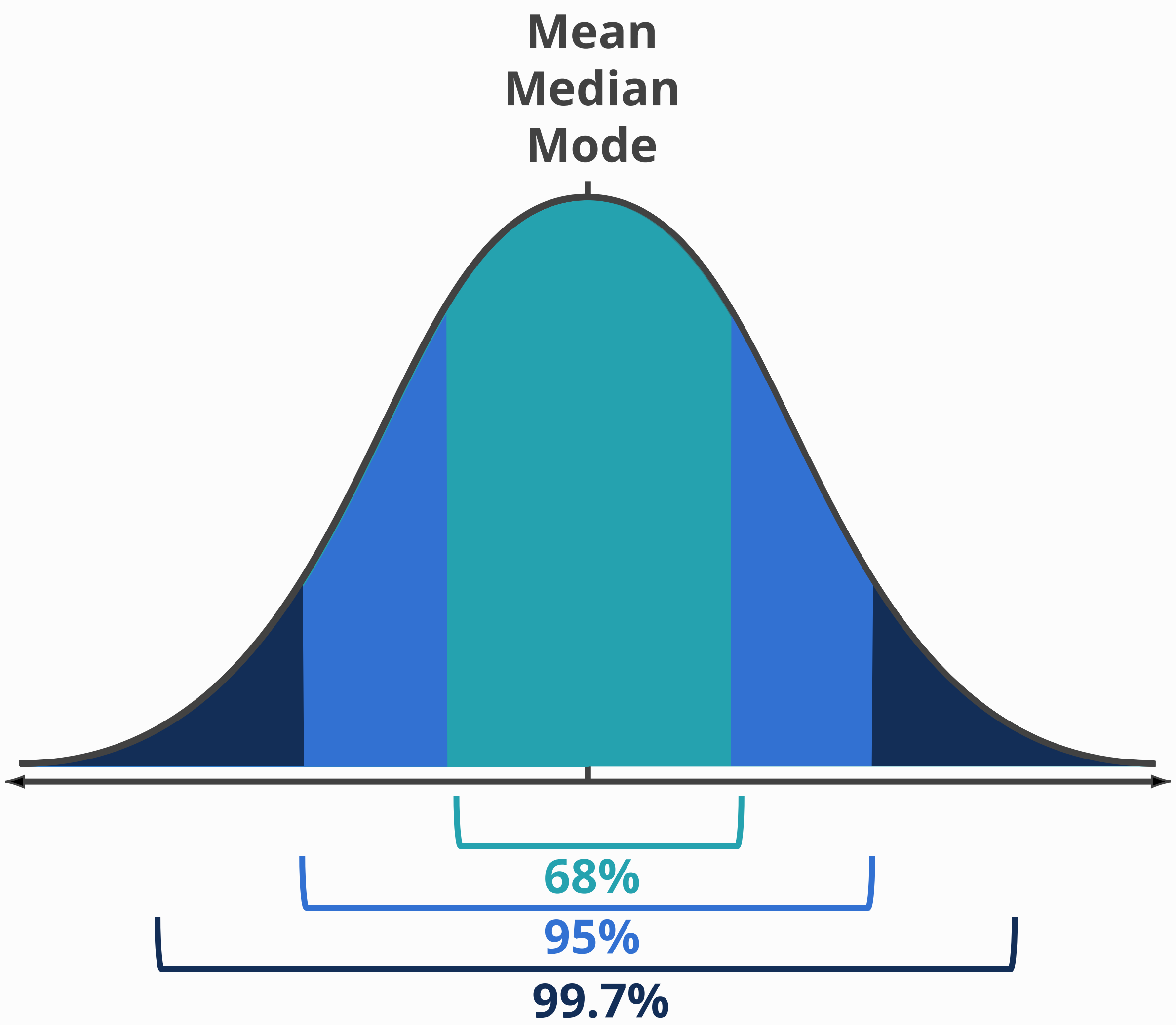
Normal (Gaussian) Distribution

1. **Symmetrical**
2. **The mean, median, mode are the same.**
3. **Has tails that approach the x-axis.**
4. Distribution follows the Empirical Rule.

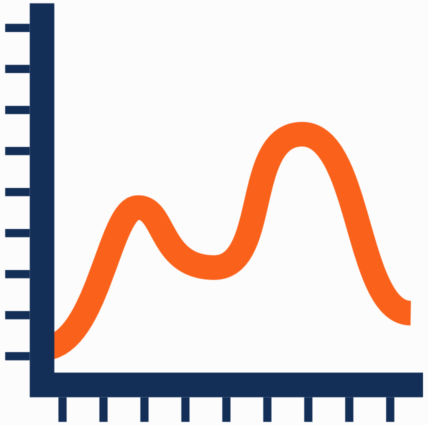


Normal (Gaussian) Distribution

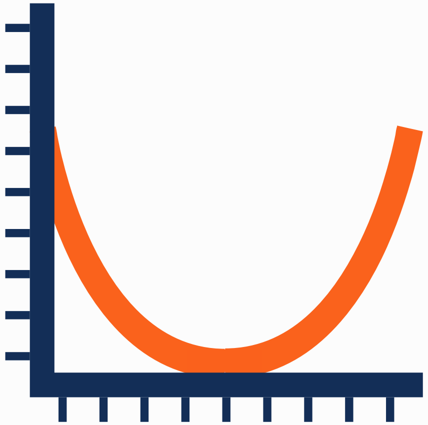
1. Symmetrical
2. The mean, median, mode are the same.
3. Has tails that approach the x-axis.
4. **Distribution follows the Empirical Rule.**



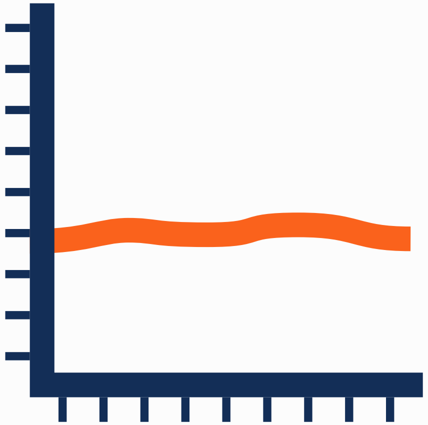
Non-normal Distributions



Bi-modal

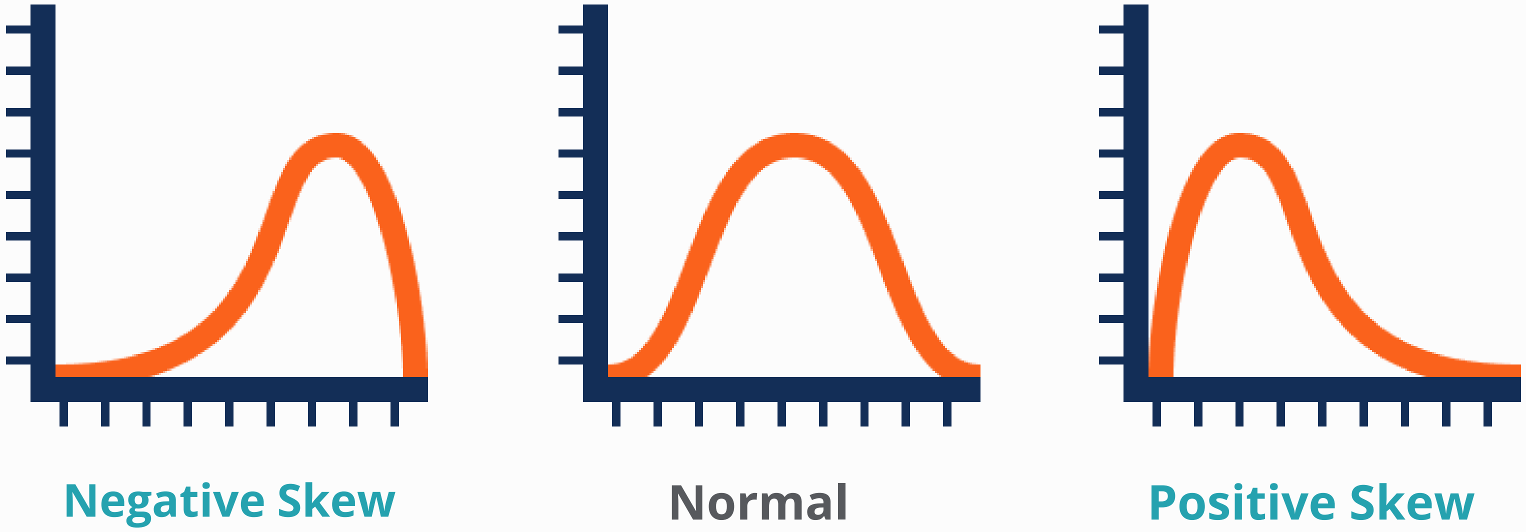


Parabolic

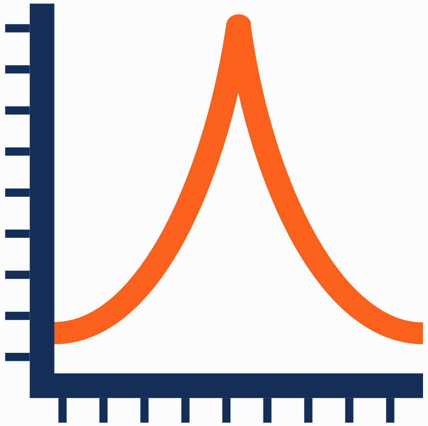


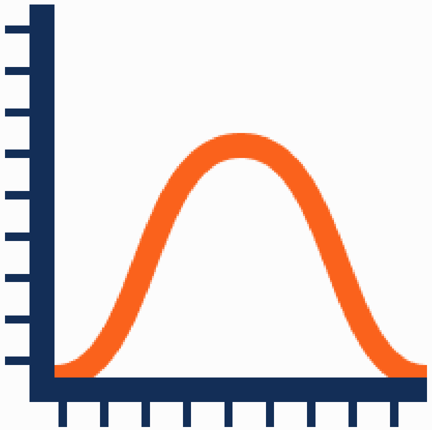
Flat

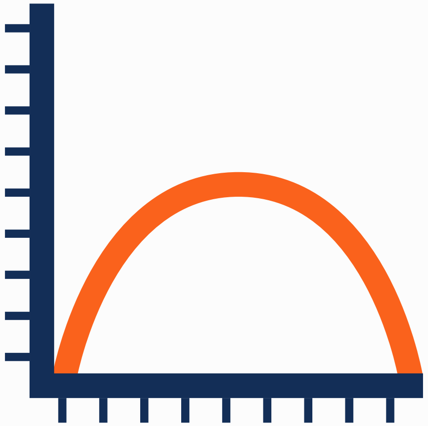
Skewness



Kurtosis







Leptokurtic

Mesokurtic

Platykurtic

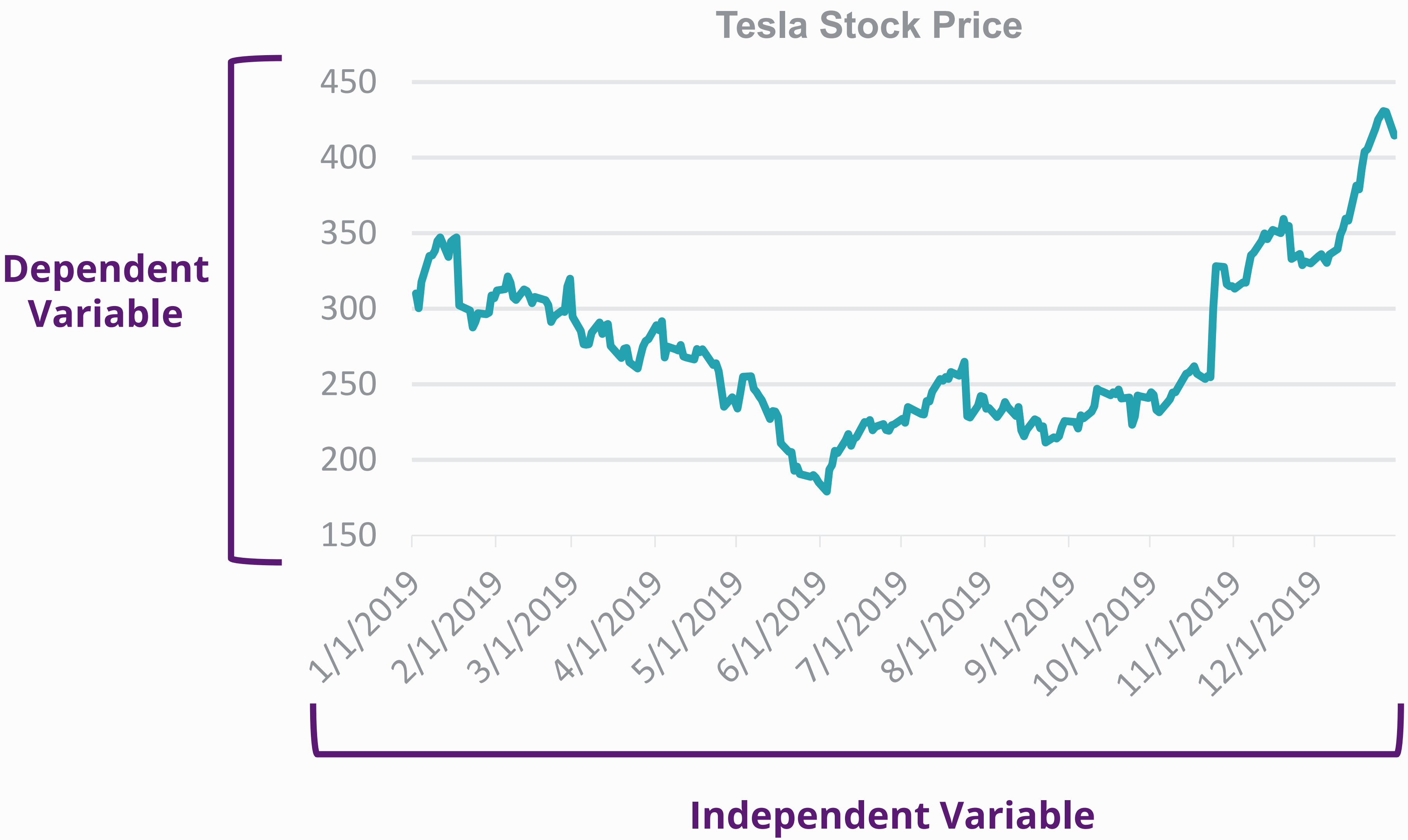
Example: S&P 500
(riskier; big swings)

Example: Bonds
(stable; few swings)

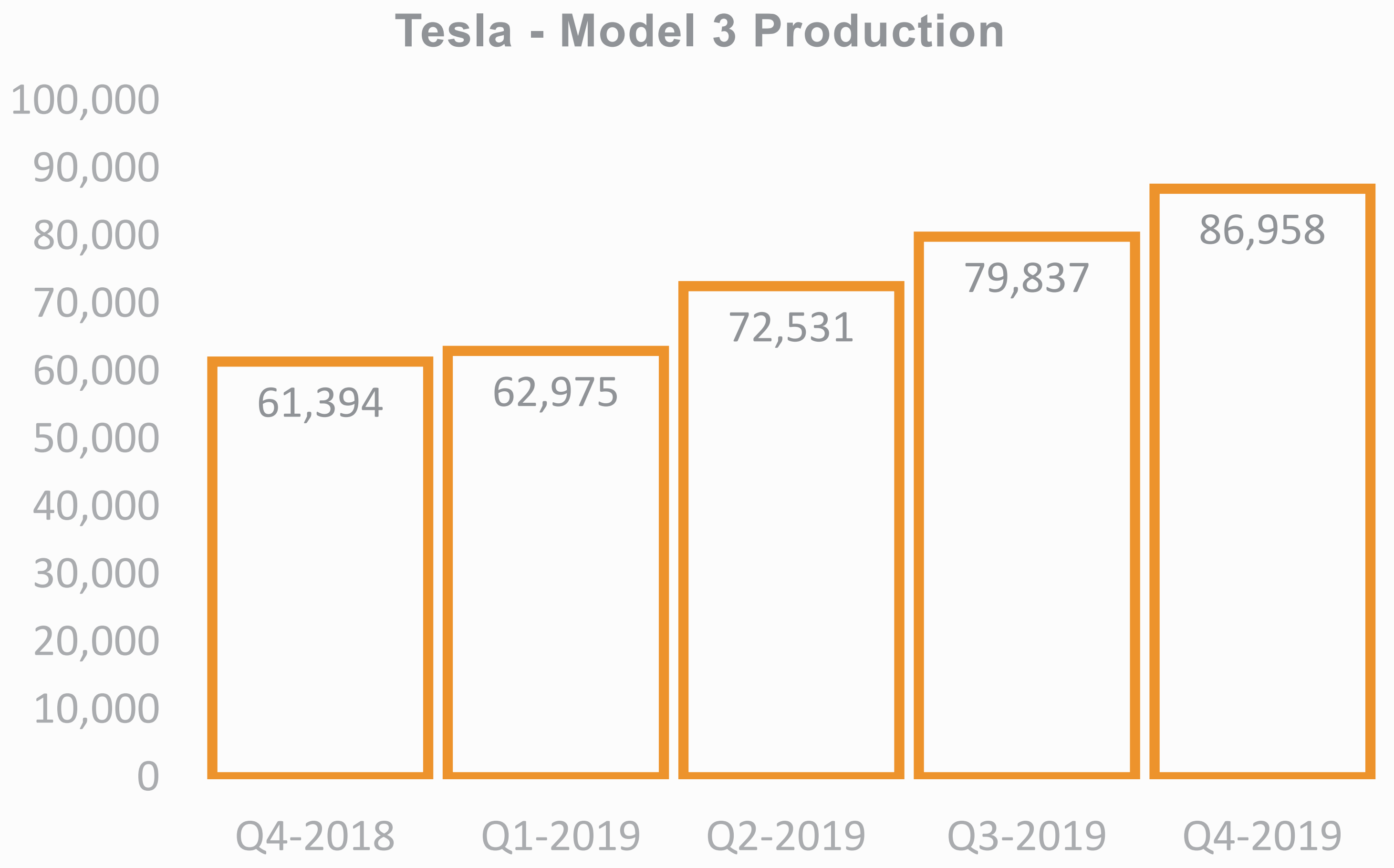
Basic Data Visualizations

Common Visualizations

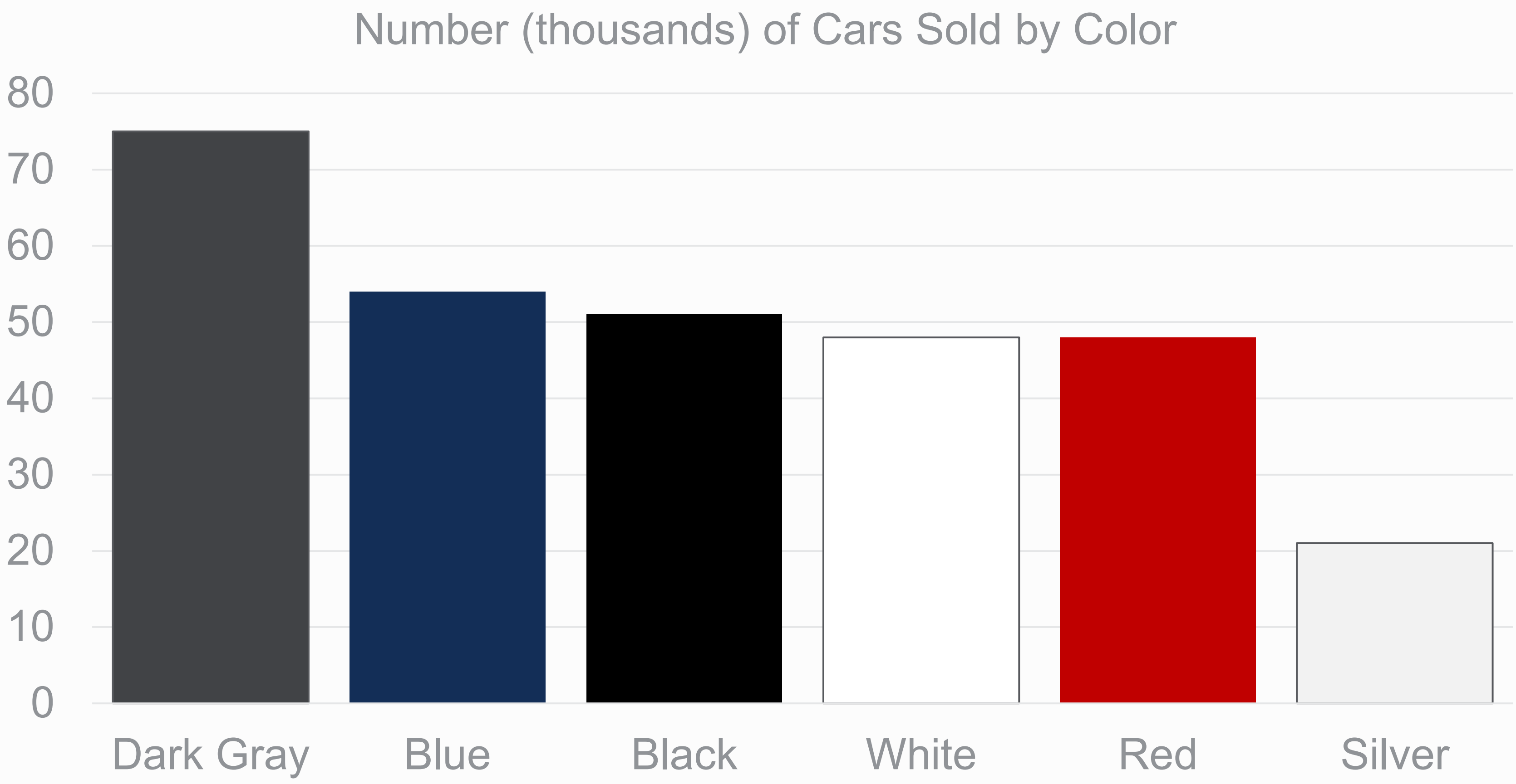
Line Plot Example



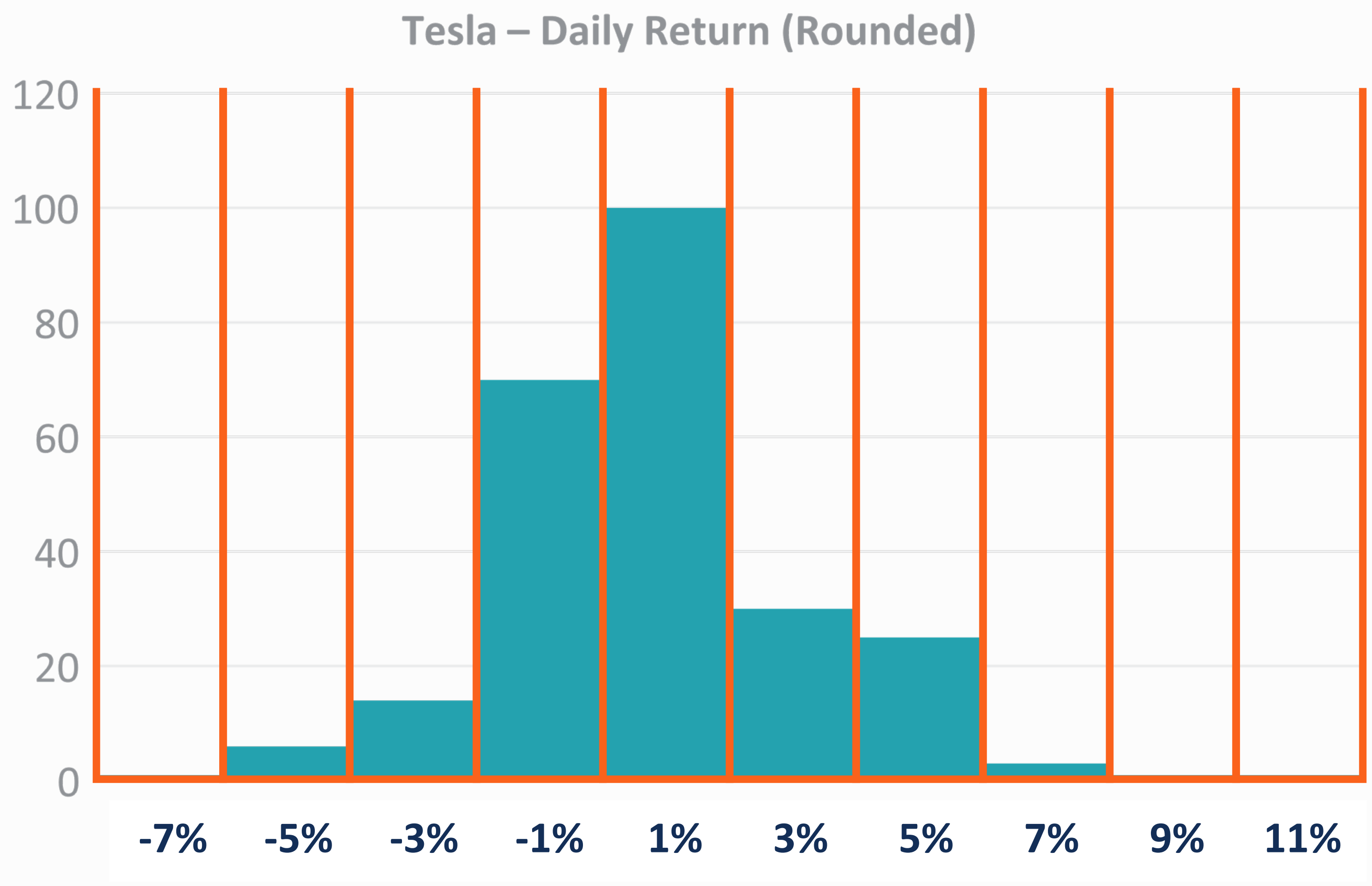
Bar Chart Example



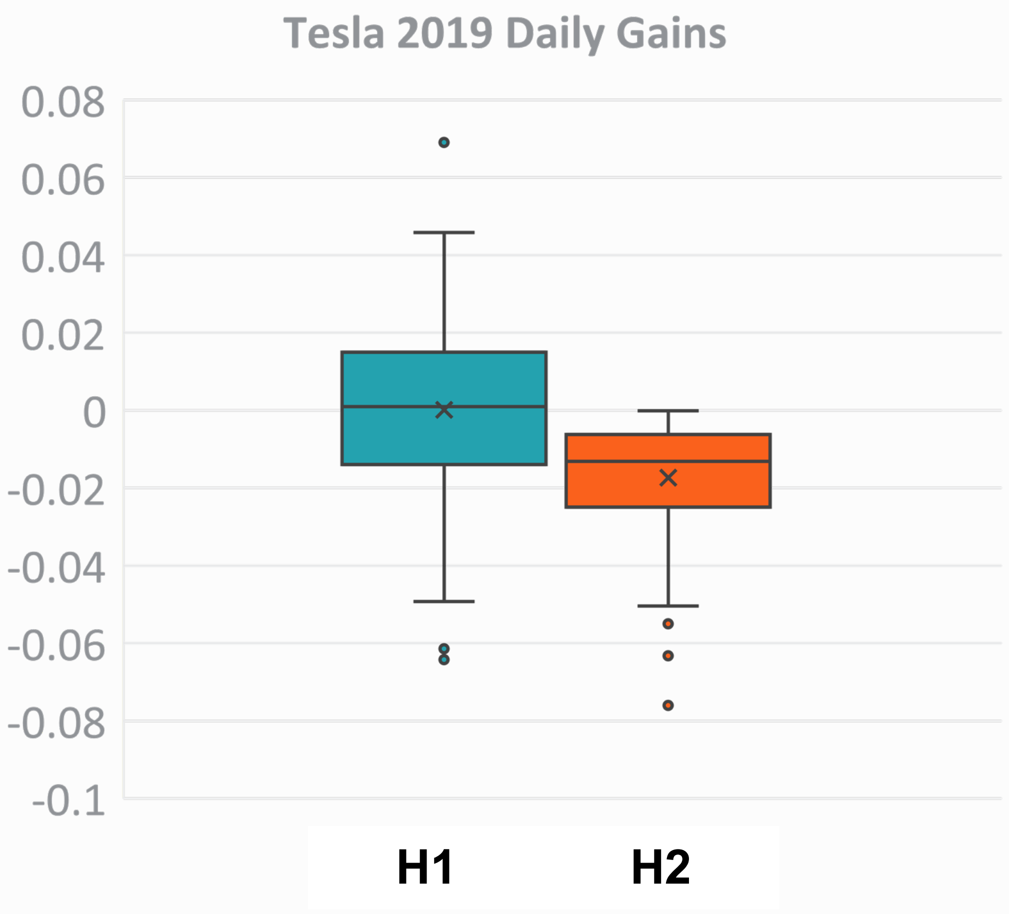
Bar Chart Example



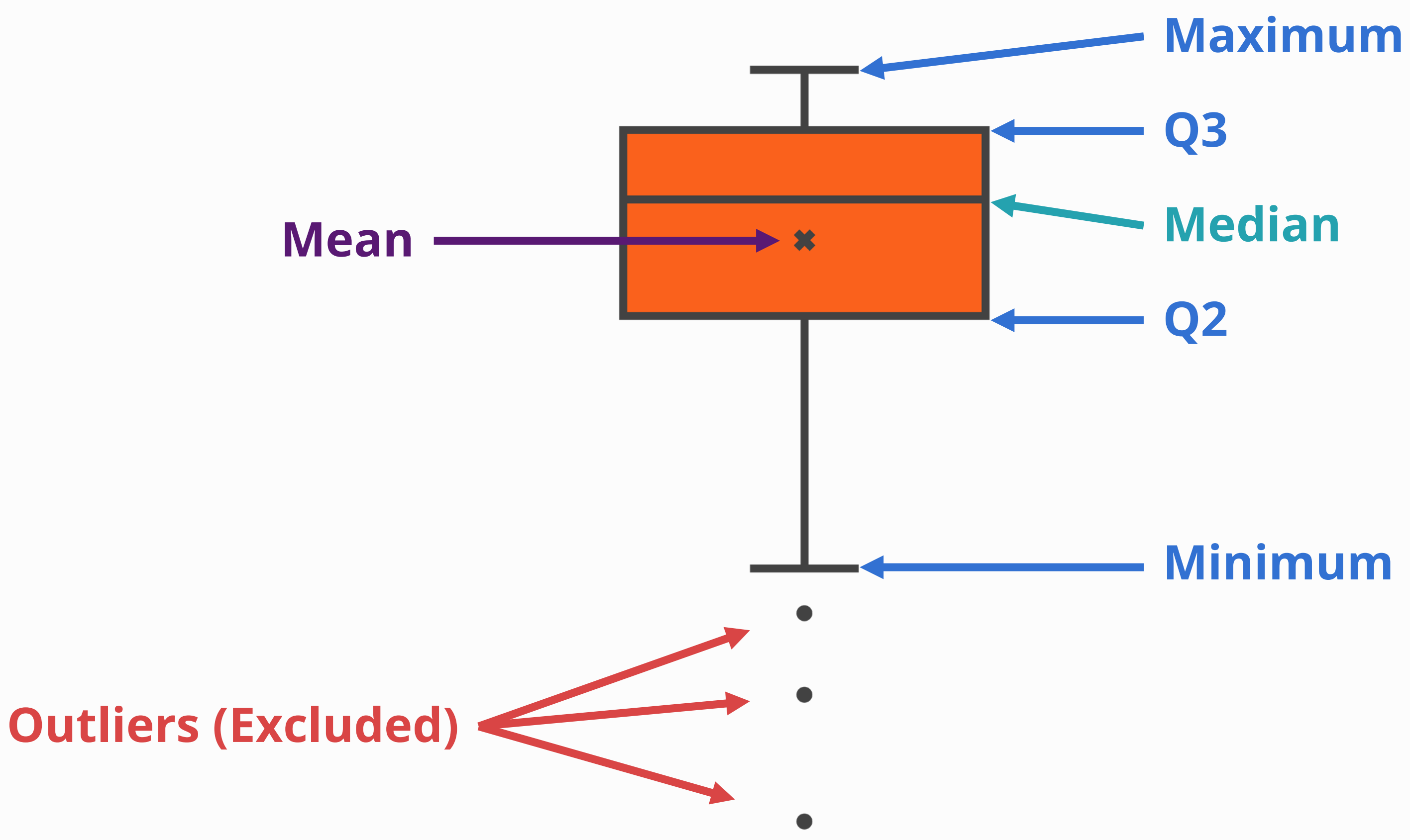
Histogram Example



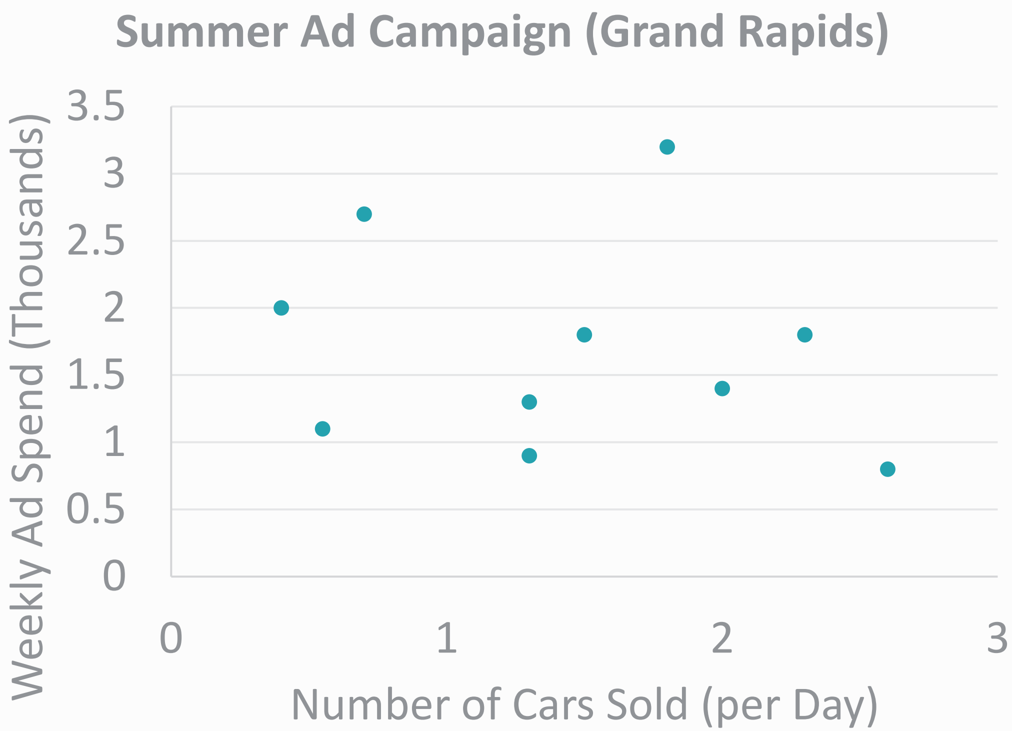
Box and Whisker Plot Example



Box and Whisker Anatomy



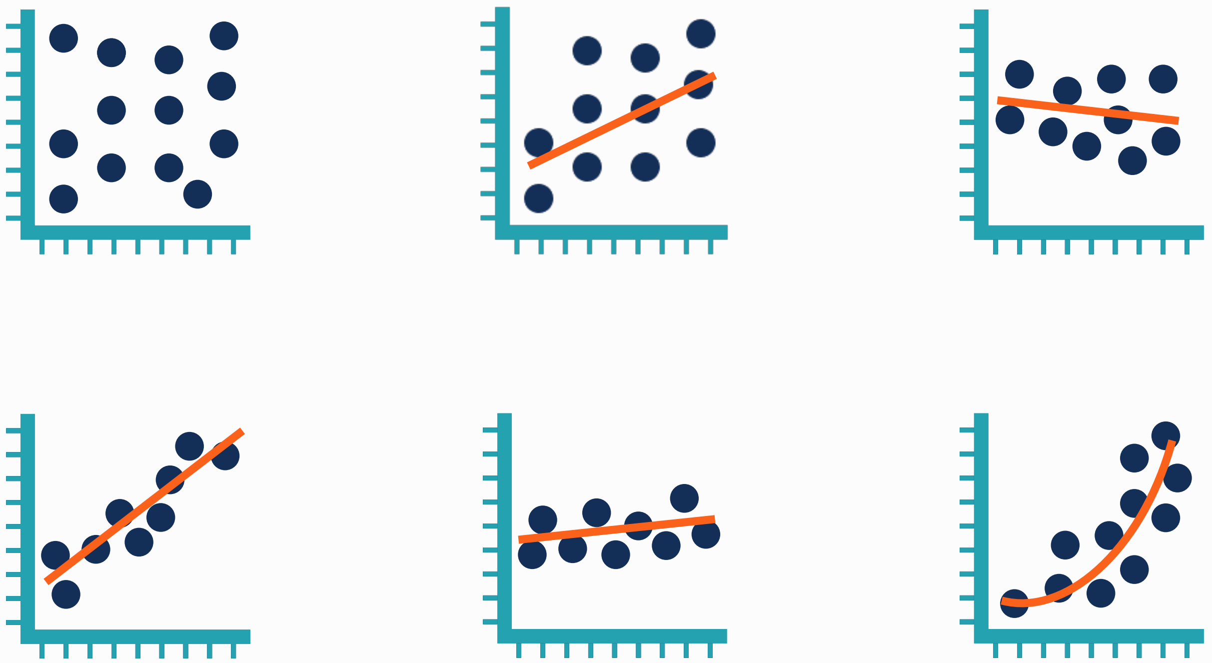
Scatter Plot



Quantitative Variable #1

Quantitative Variable #2

Scatter Plot Example



Conclusion

Chapter 3: Sampling

Course Objectives

Samples & Populations

Samples and populations and learn their importance for inferential statistics

Sample Sizes (and the Law of Large Numbers)

Discuss the theory and what it tells us about our sample size

Central Limit Theorem

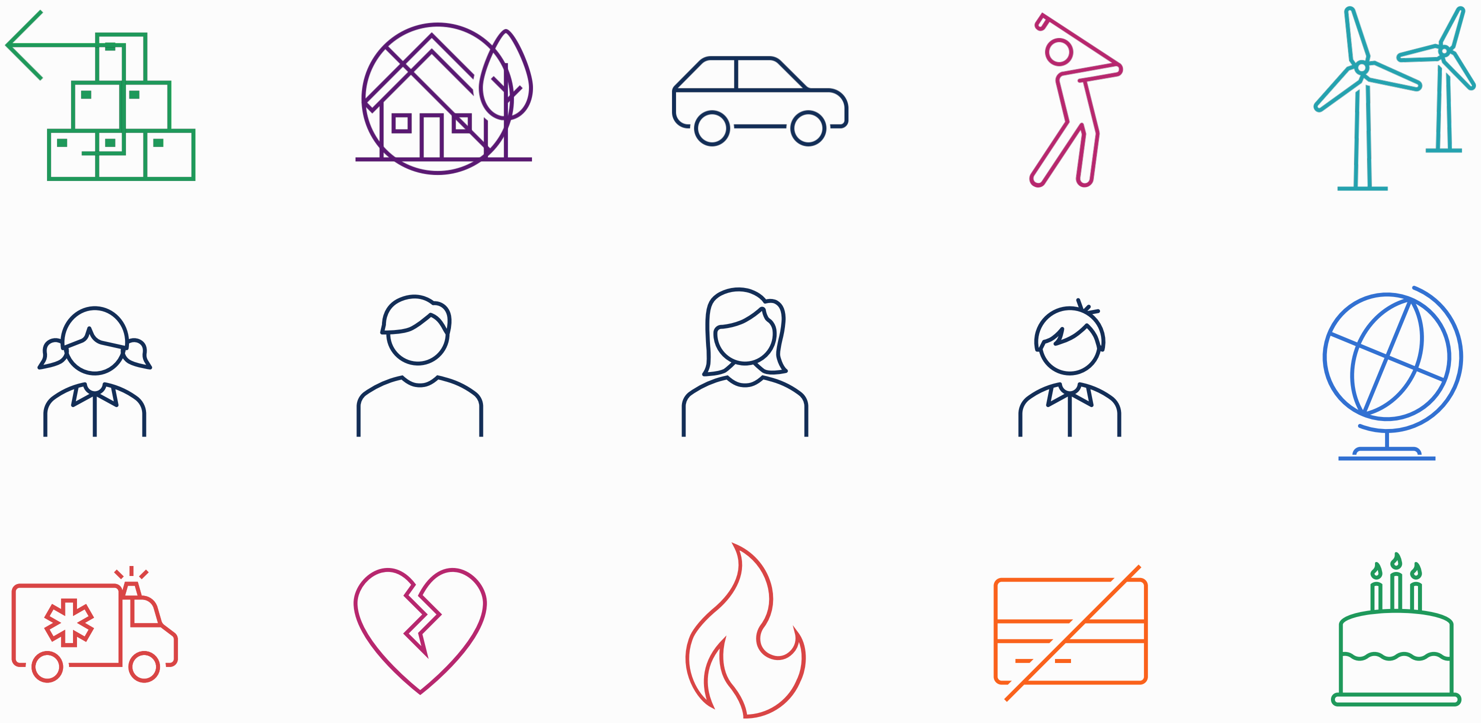
Explore how samples from non-normally distributed data become normally distributed as our sample size increases

Controlling for Bias

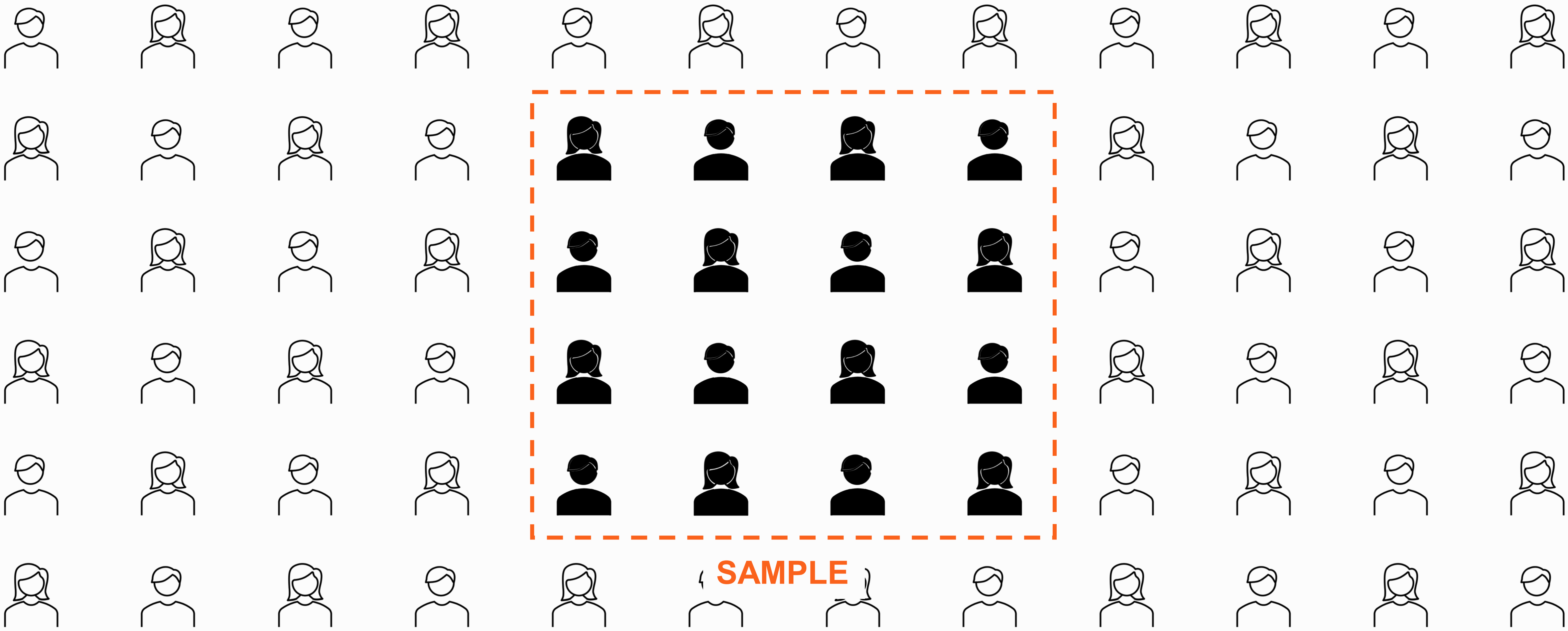
Discover the two basic methods for controlling for bias

Samples & Populations

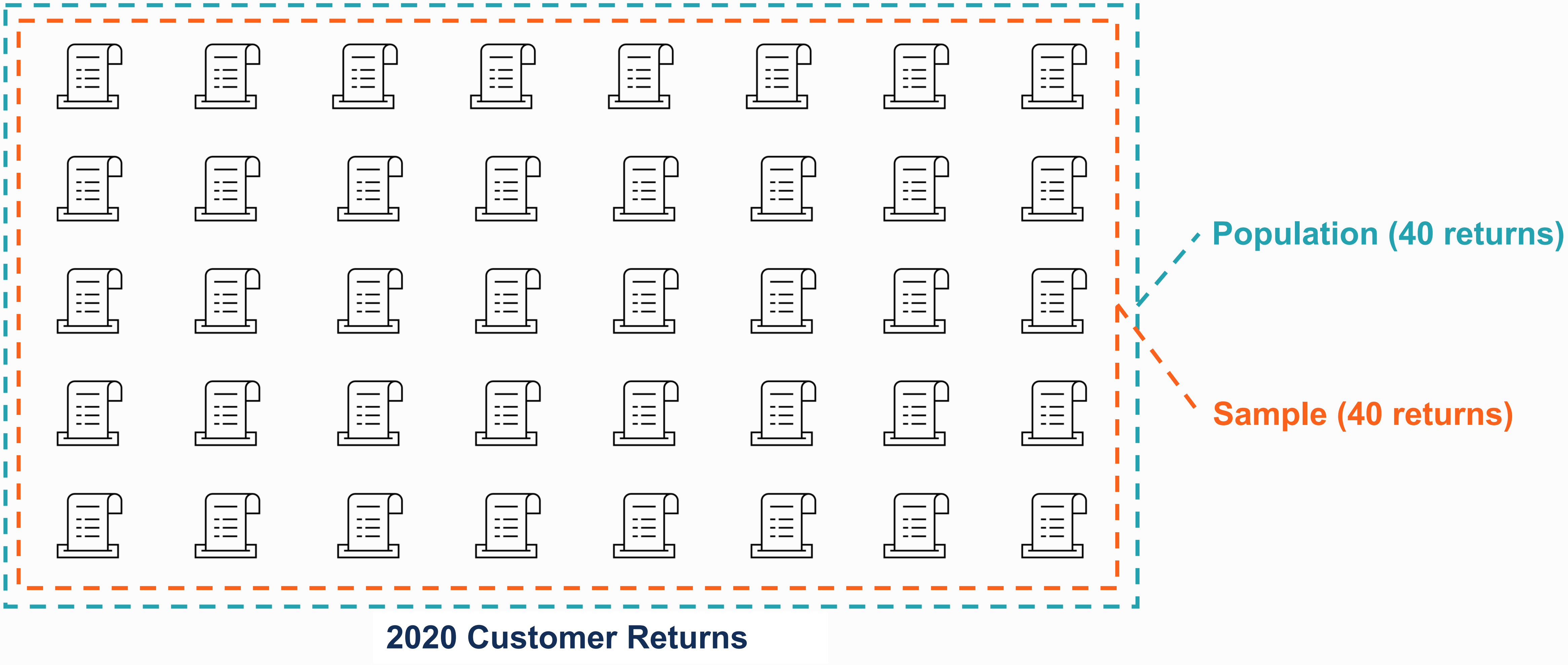
In statistics, our population is what we want to analyze and learn about.



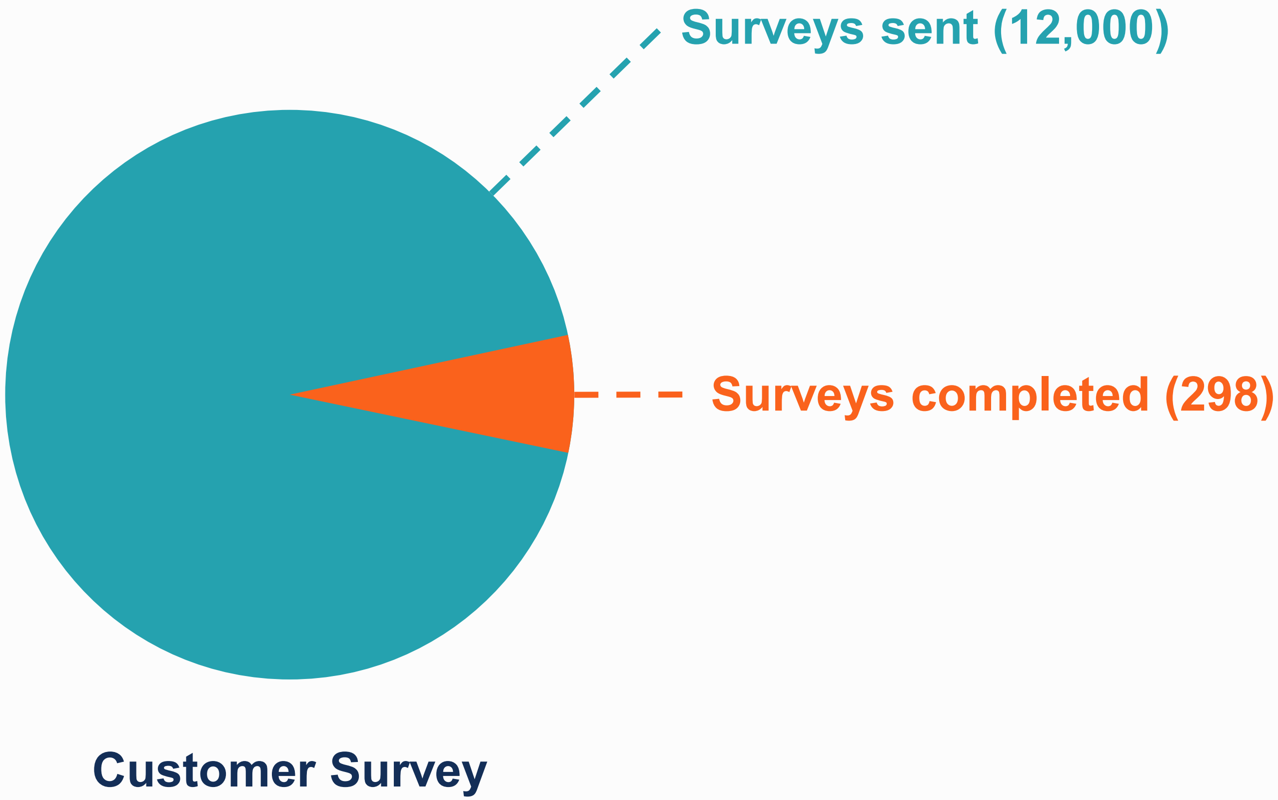
Samples & Populations



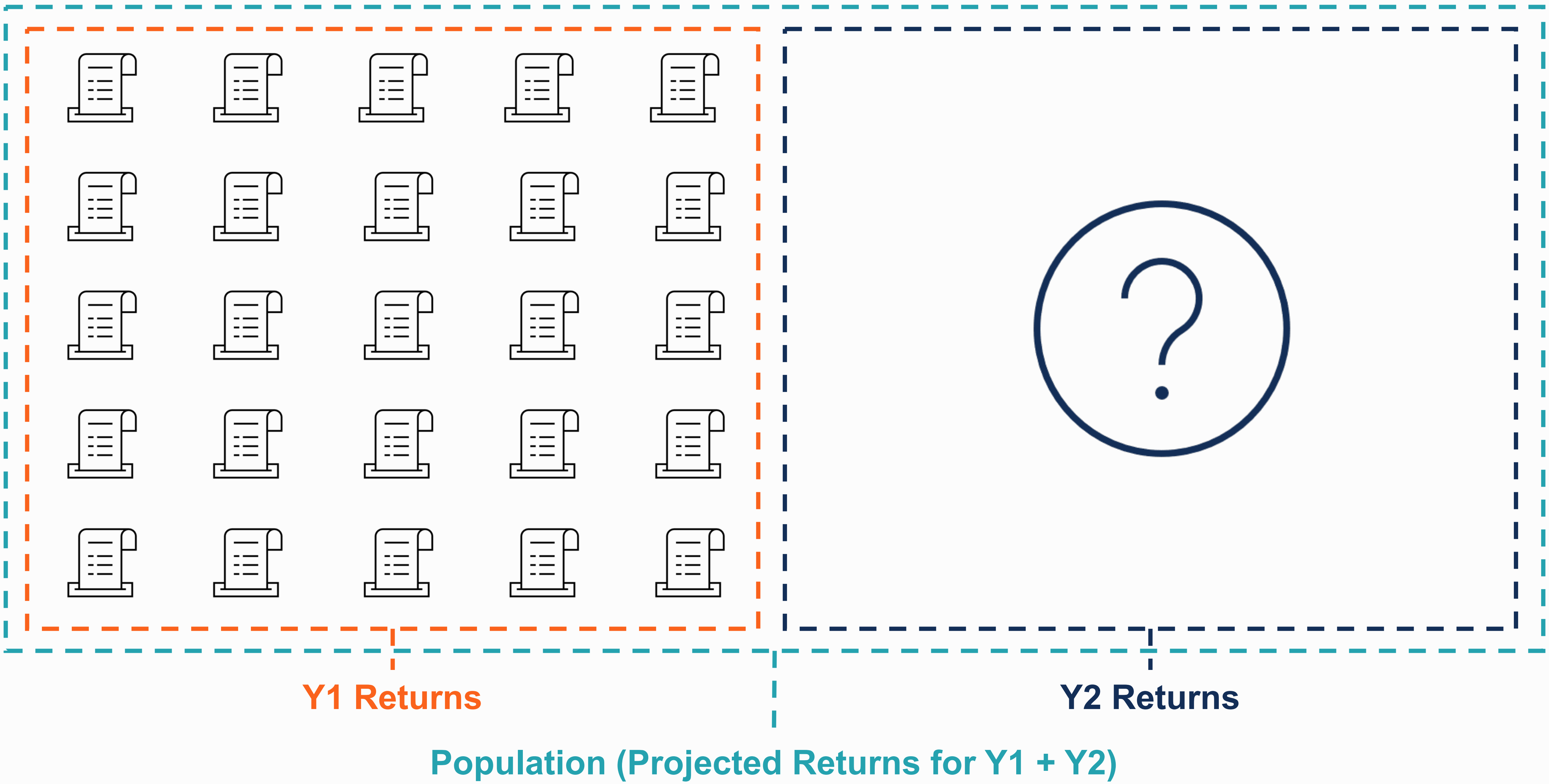
Complete Sample



Partial Sample



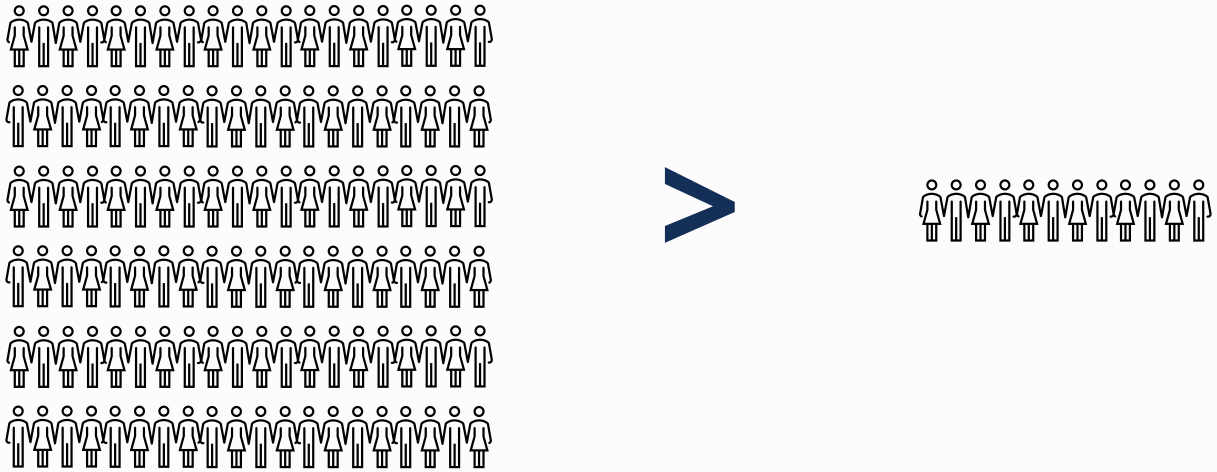
Prior Data



Y1 Returns
Y2 Returns
Population (Projected Returns for Y1 + Y2)

Sample Size & Law of Large Numbers

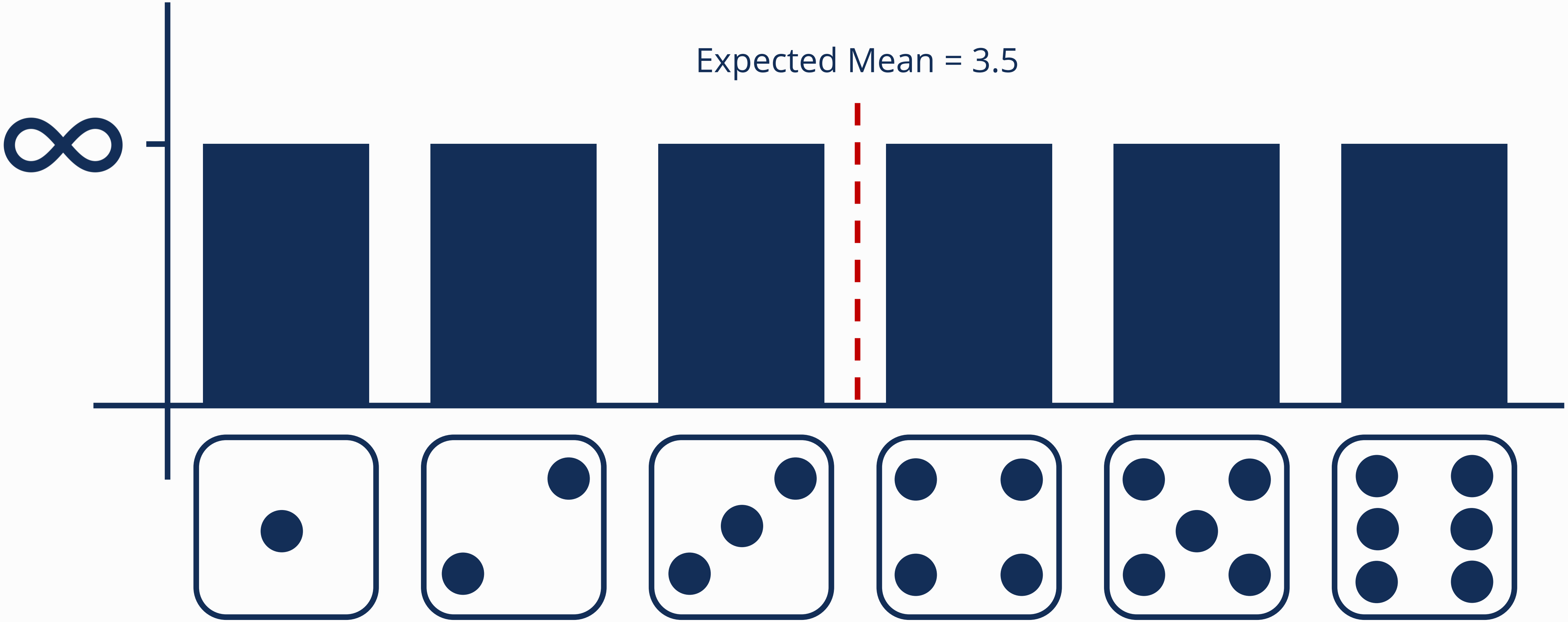
Law of Large Numbers



Law of Large Numbers:

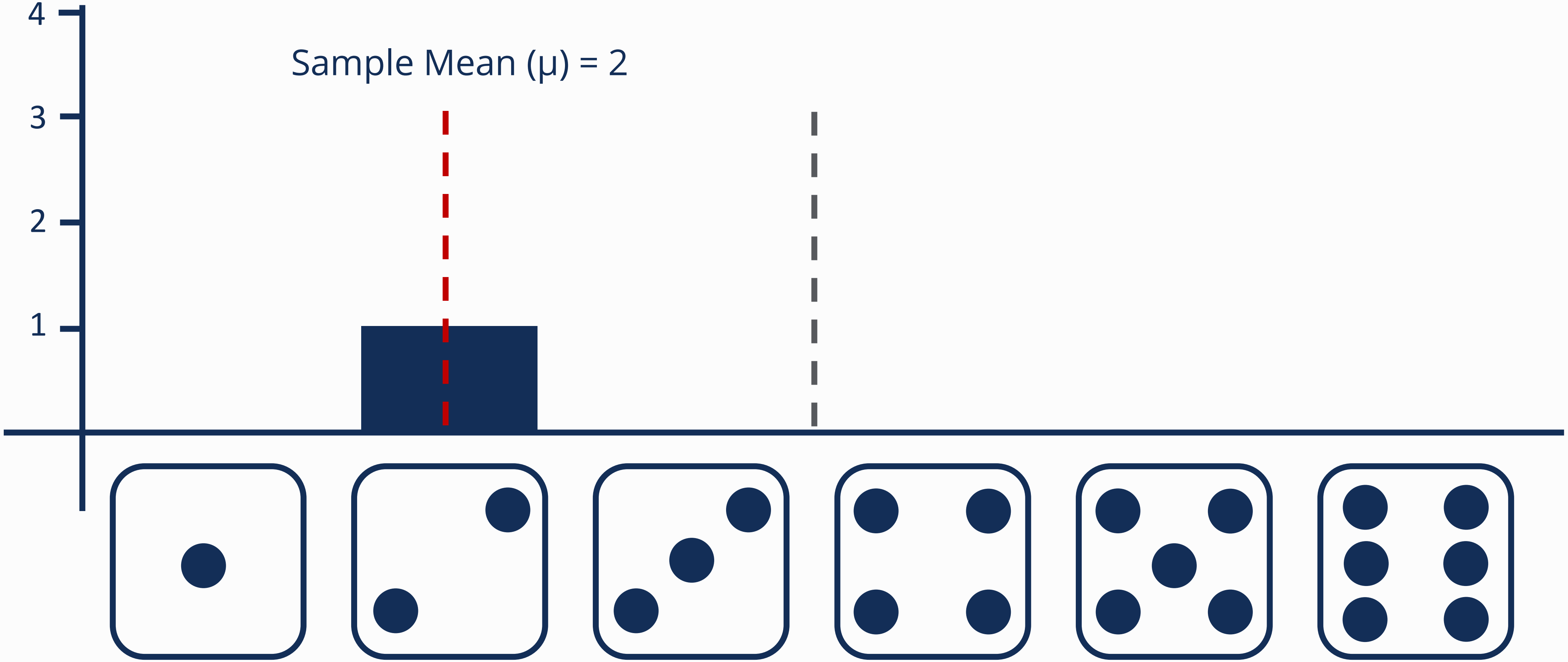
As our sample size grows, its mean tends towards the average of the whole population.

Law of Large Numbers



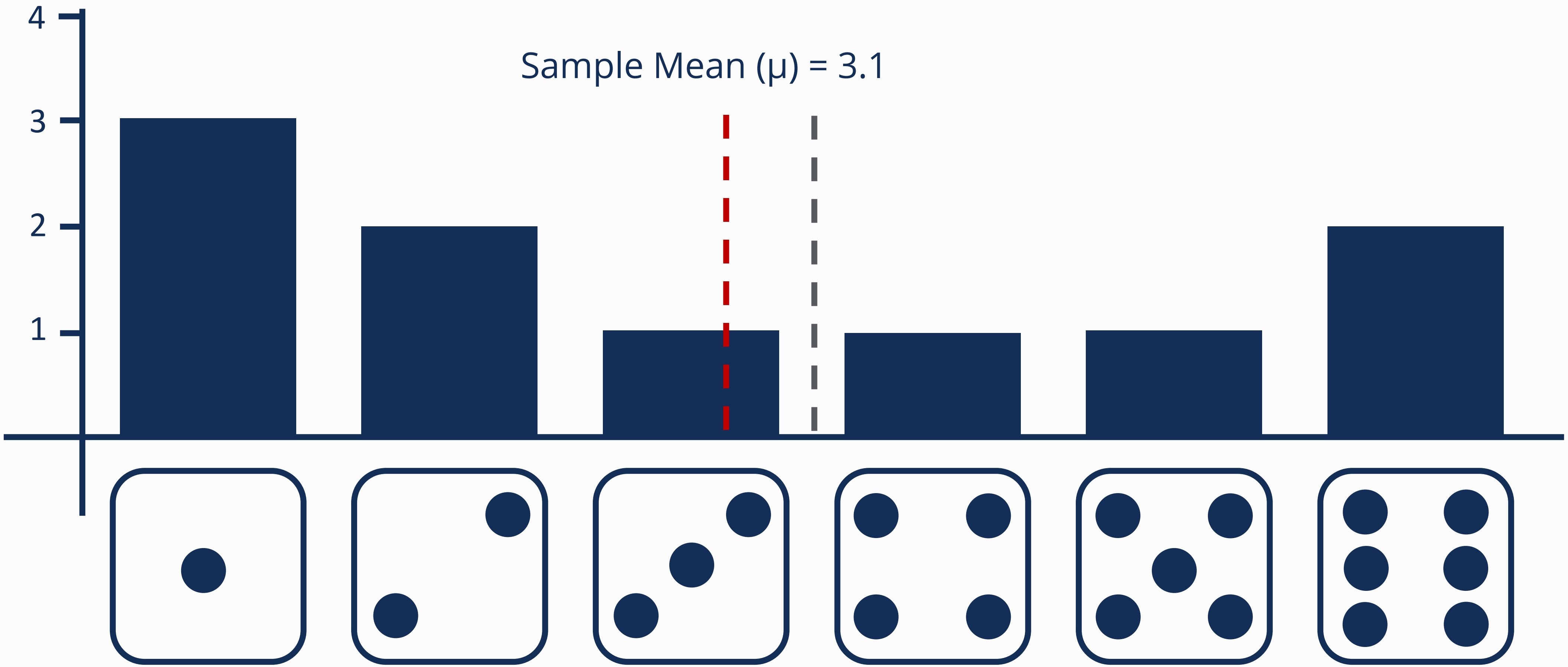
Probability: 16.67% 16.67% 16.67% 16.67% 16.67% 16.67%

Law of Large Numbers (Average after 1 roll)



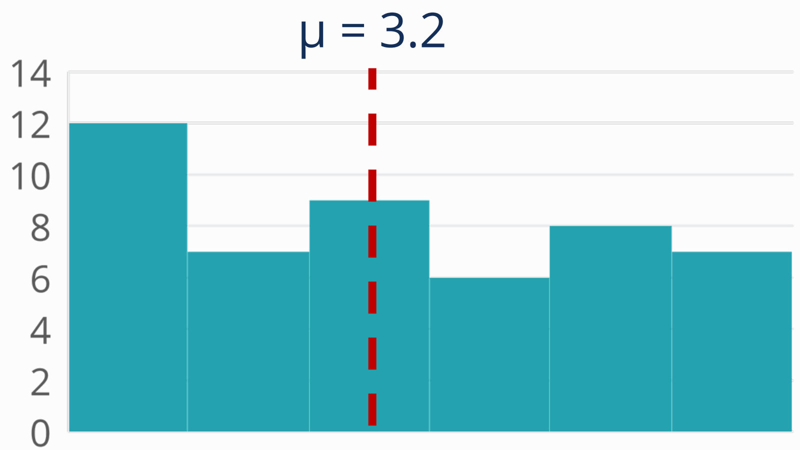
Number of samples (n) = **1**

Law of Large Numbers (Average after 10 rolls)

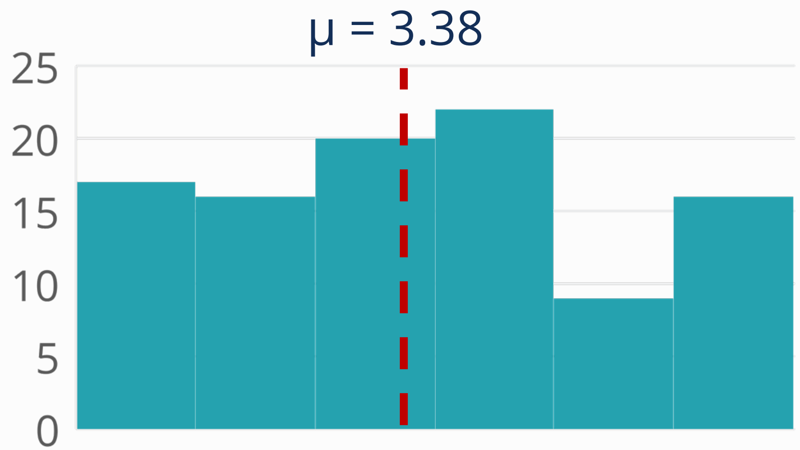


Number of samples (n) = **10**

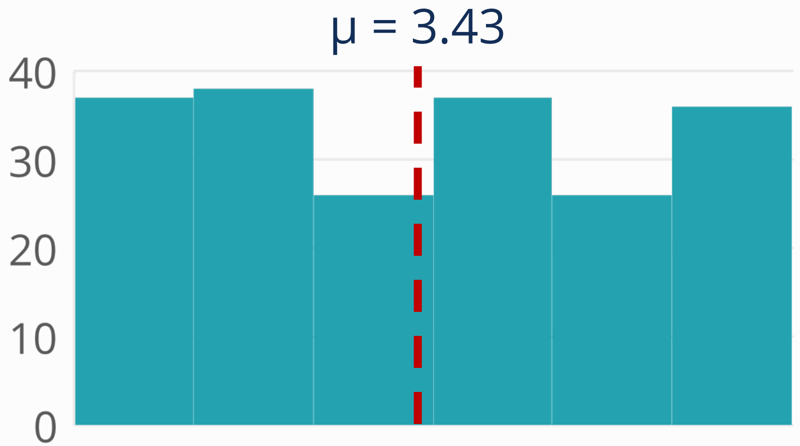
Law of Large Numbers (Simulated with Dice Rolls)



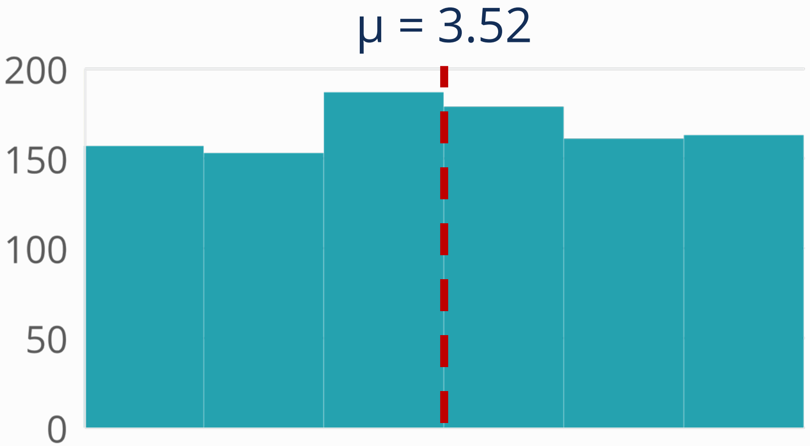
Simulation #1 (n = 50)



Simulation #2 (n = 100)



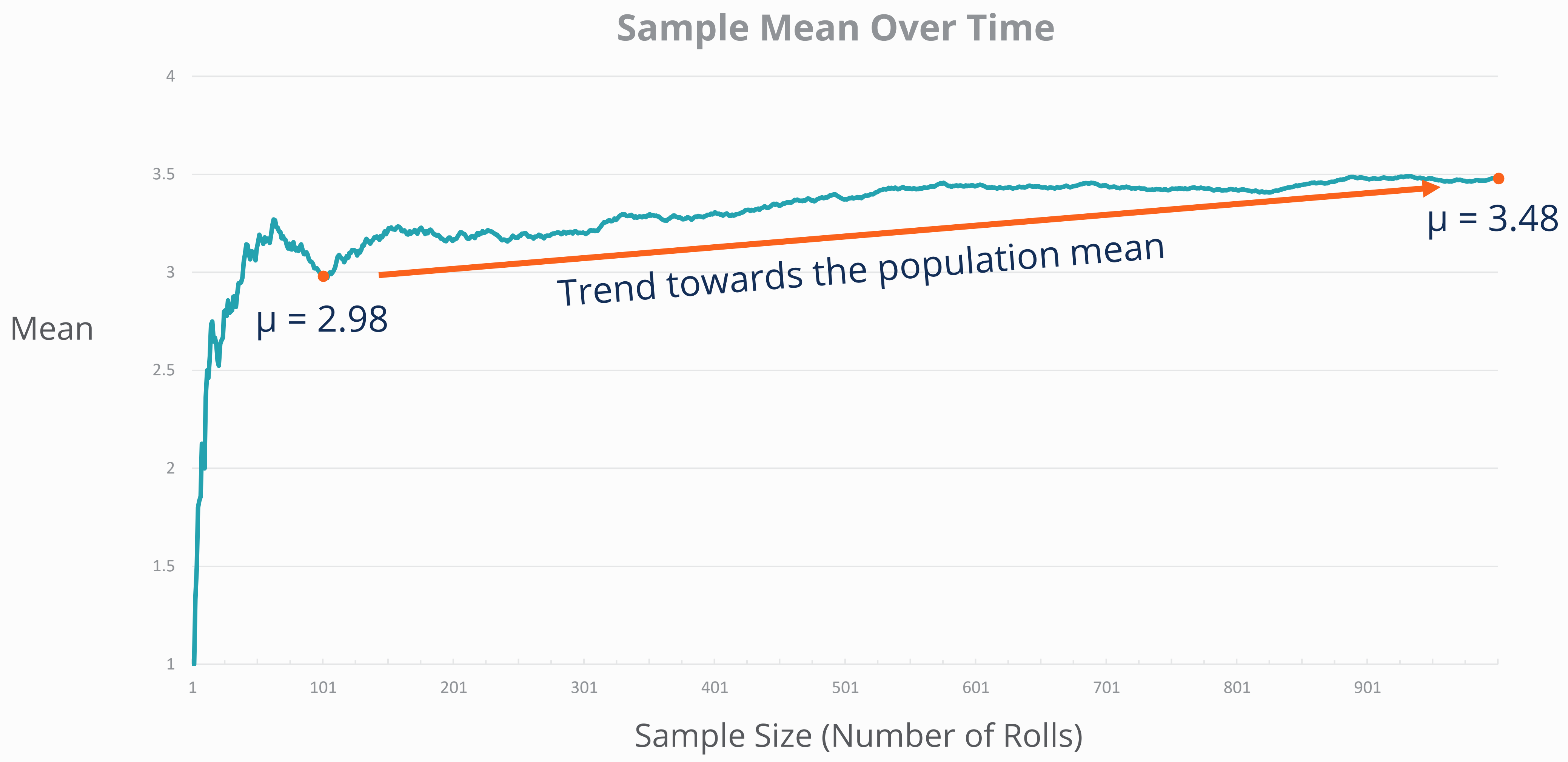
Simulation #3 (n = 200)



Simulation #4 (n = 1000)

Note: 'μ' stands for 'sample mean'

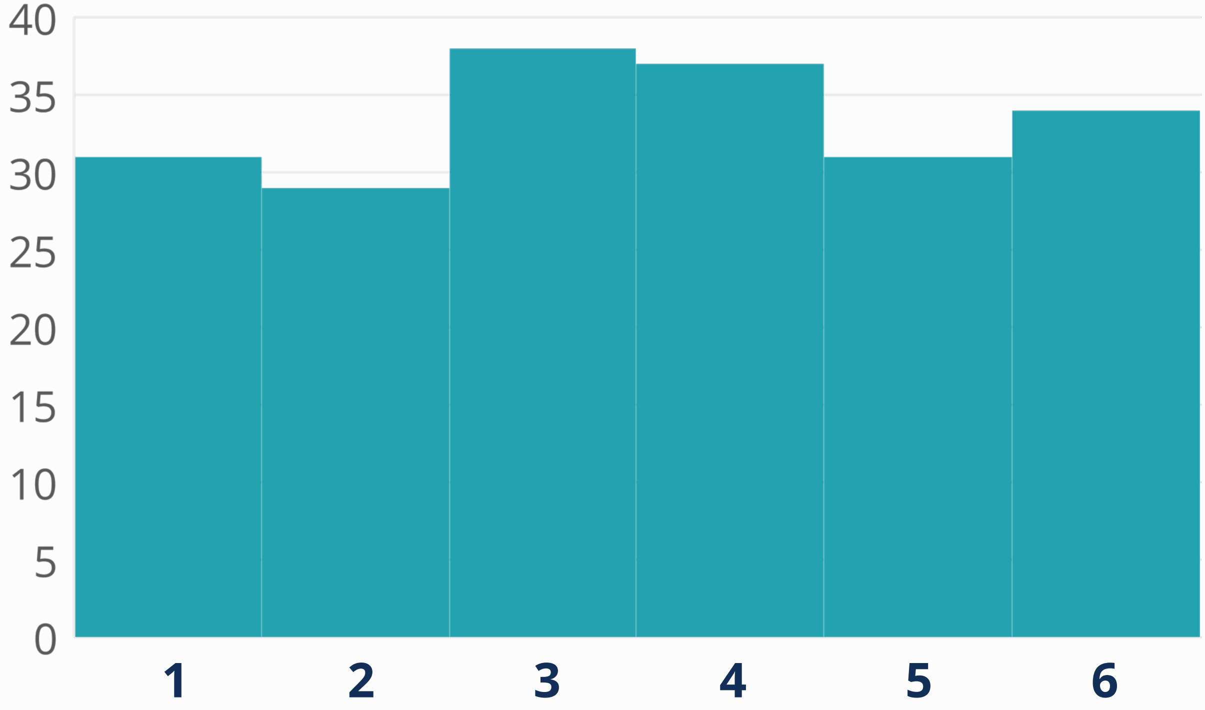
Law of Large Number (1-1000 dice rolls; mean over time)



Central Limit Theorem

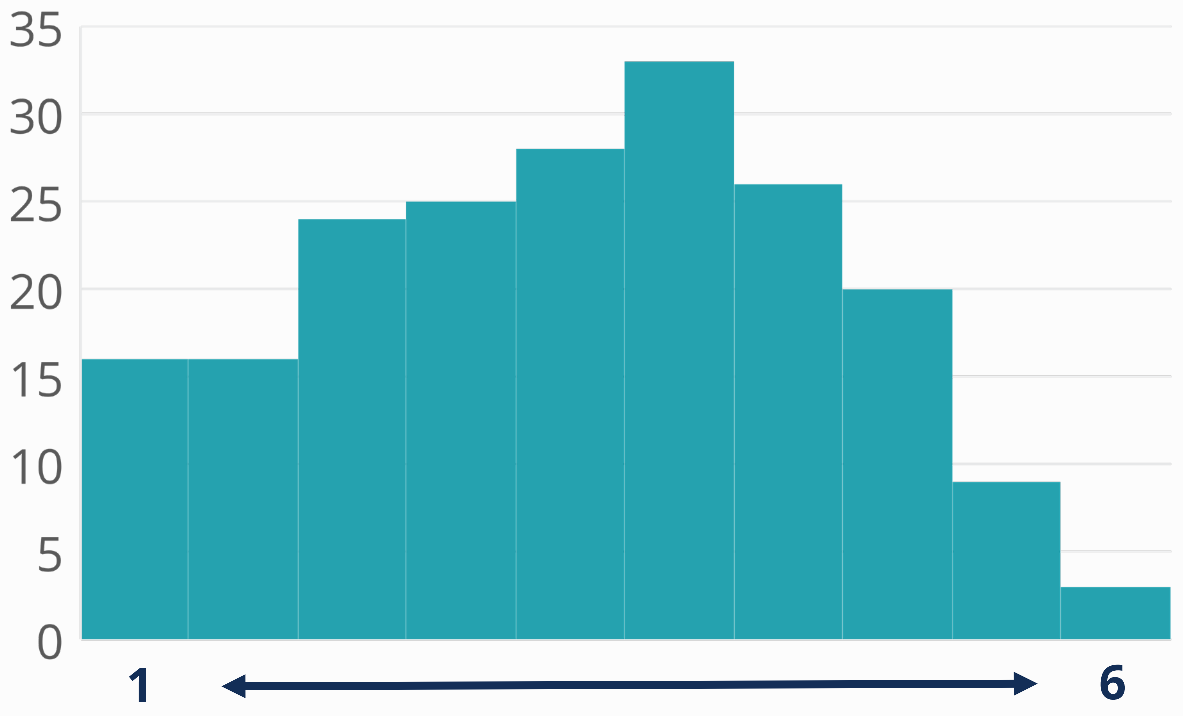
200 samples: Sample size (n) = 1

As we increase our sample size, the distribution of the **mean of our samples** changes.



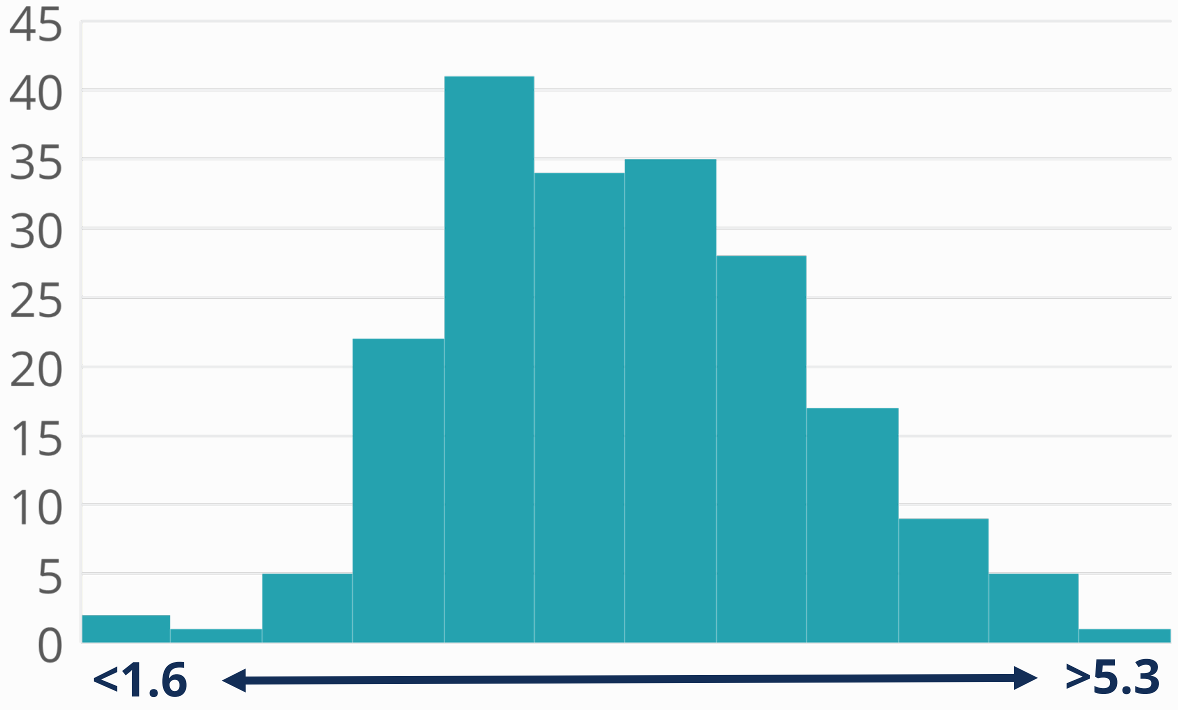
Distribution of sample means (n = 1) after 200 rolls

200 samples: Sample size (n) = 2



Distribution of sample means (n = 2)
after 200 rolls

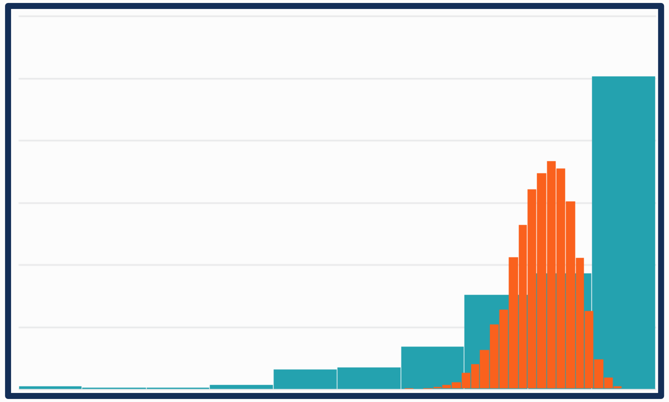
200 samples: Sample size (n) = 6



Distribution of sample means (n = 6)
after 200 rolls

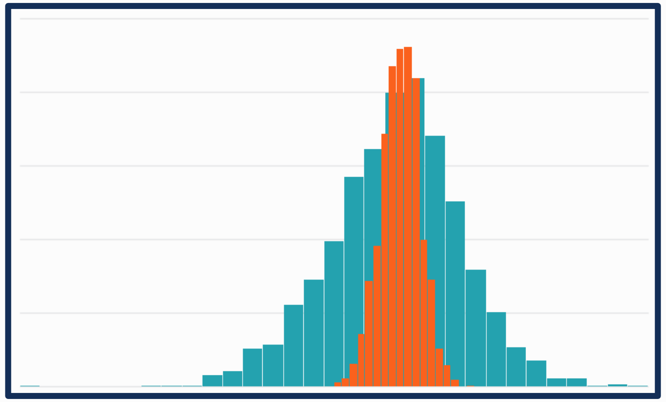
Sampled Means of Other Distributions

Services Company Customer NPS Responses



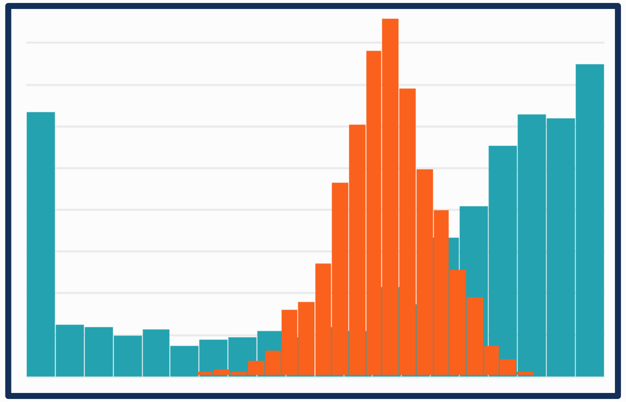
(10,000 samples; n = 16)

US Treasury Bond (2014-2019) Daily Price Change %



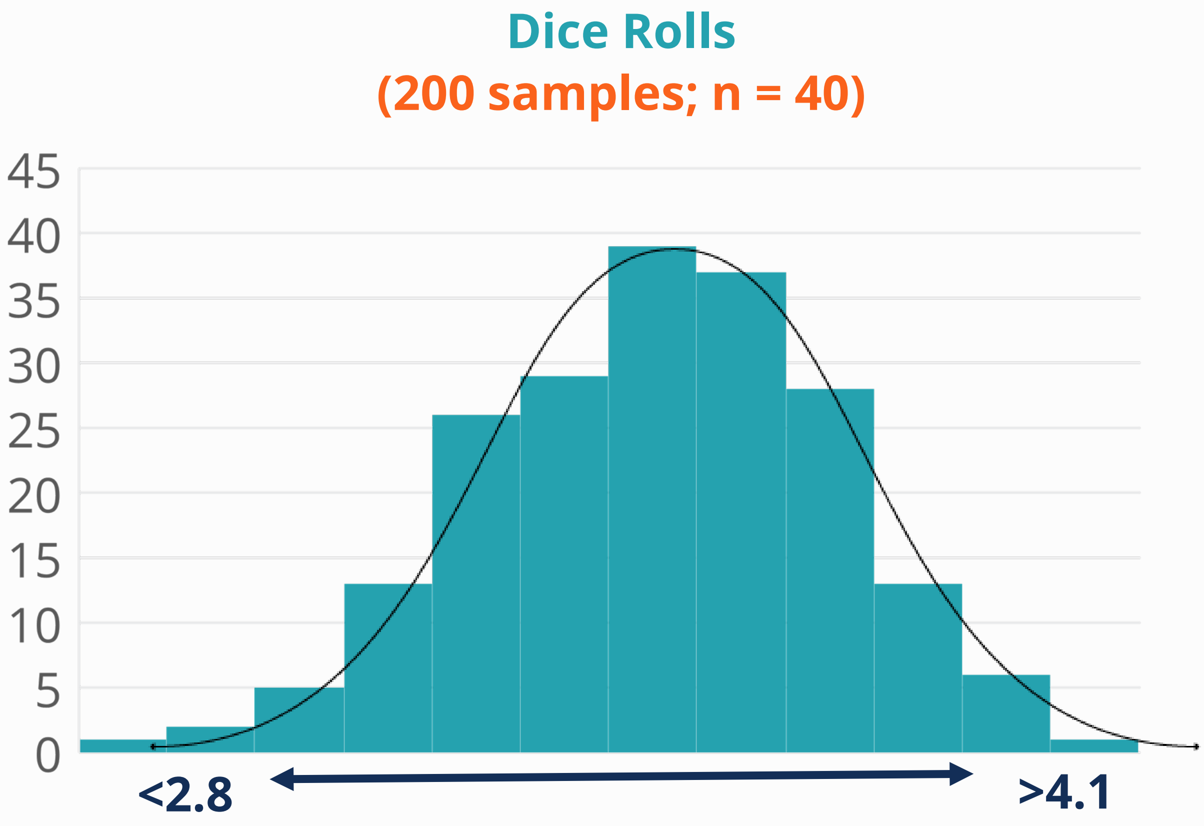
(1,510 samples; n = 12)

University Mid-term Exam Scores



(1,090 samples; n = 25)

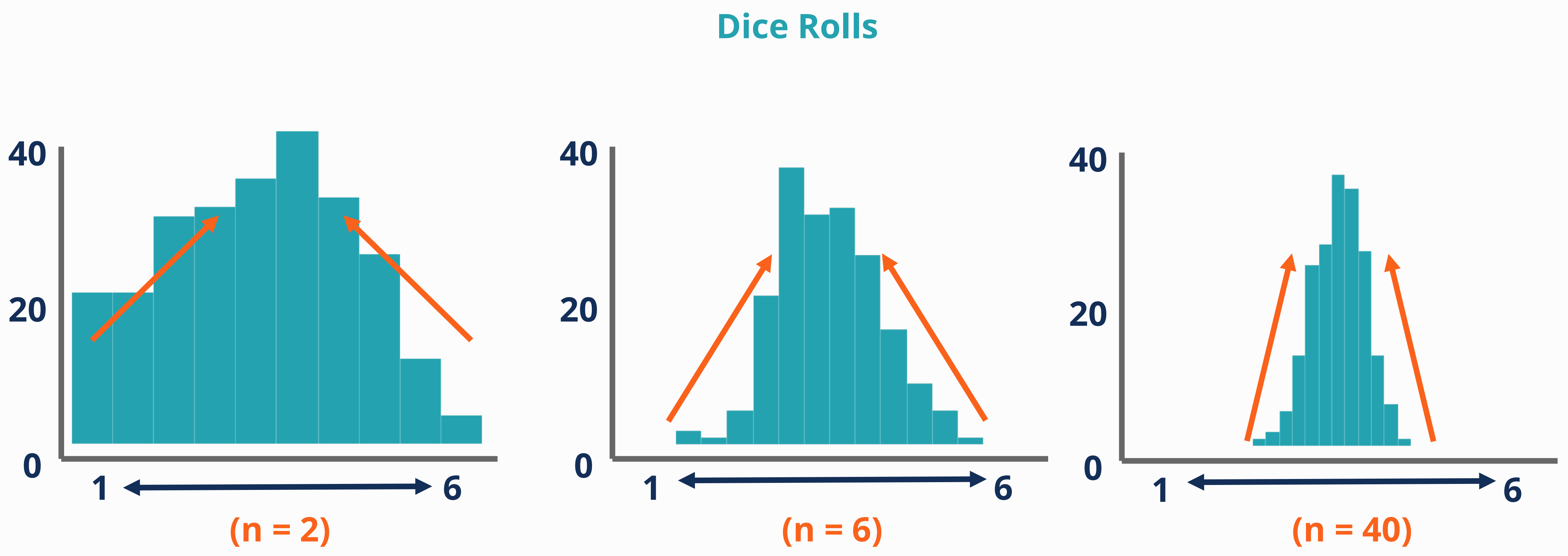
Central Limit Theorem



Central Limit Theorem:

*The distribution of **sample means** approaches the normal distribution as the sample size gets larger—regardless of the shape of the population distribution.*

Central Limit Theorem



Determining a Sample Size

Large Samples vs Resource Limitations

$$\frac{\textbf{Budget}}{\textbf{Cost per Sample}} = \text{Number of samples (you can afford)}$$

Rules-of-Thumb and Tables

Rule of Thumb Recommendations:

A general rule for determining sample sizes is that, under most conditions, a sample of **30+** is sufficient. If you can push it, **100+** is even better.

Sample Size Recommendation Tables:

Acceptable Margin of Error	Size of Population			
	Large	5000	1000	200
±20%	24	24	23	22
±10%	96	94	88	65
±5%	384	357	278	132
±3%	1067	880	516	169

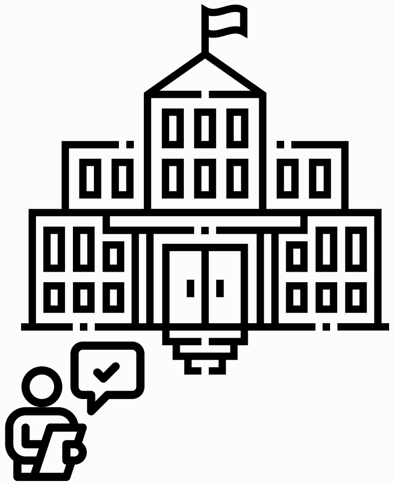
Beyond Rules of Thumb

There are a few factors that may require you to use larger samples than is typically recommended:

1. Wanting a high degree of confidence.
2. When measuring variables with a high variance.
3. When measuring for variables that occur very infrequently in the population

Controlling for Bias

Scenario



1. We're only capturing students from one area.
2. We're not capturing private or home school students.
3. We may not capture shy or busy students who ignore us.

Sampling bias occurs when our data is collected in a way that certain members of our population are either overrepresented or underrepresented.

Bias samples won't be representative of our population.

Representative vs. Random Sample

Economists and researchers seek to reduce sampling bias when employing statistical analysis.

Representative Sample

- A group or set chosen from a larger population
- Should adequately replicate the chosen characteristic or quality of the larger group

Random Sample

- A group or set chosen from a larger population randomly
- Should be an unbiased representation of the larger population

Representative vs. Random Sample Example

Population		Representative Sample		Random Sample
All the daily stock prices of AAPL since IPO	→	A period which well represents all phases of AAPL stock—with phases of upward, downward, and stagnant trends.	→	Stock prices from 100 random days over the period since IPO
All the customers living in the state of California	→	100 selected from different parts of California. The number selected from each area is proportionate to the number from our population who live in those areas.	→	100 randomly selected customers who live in California

Chapter 4: Hypothesis Testing

Session Outline

Hypothesis

Determine the null and alternative hypothesis for a test

Confidence Levels

Select an appropriate confidence level

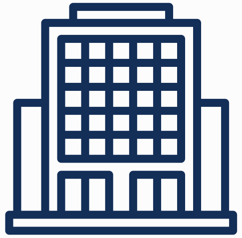
Test statistics

Select an appropriate test to apply

Applying the t-test and getting your P-value

Determine the p-value for your test and accept or reject the null hypothesis based on that data

Scenario



Vacuums that last a lifetime!

Scenario

A

B

Hypothesis Testing

Scenario

Which warranty sells our vacuums better?

Test Results

	5-Year (Full)	7-Year (Limited)
Sample Size	750	750
Total Revenue	$42,222	$47,631
Mean Per-visitor	$56.30	$63.51

Hypothesis Test

Hypothesis tests attempt to provide an answer to questions such as "How likely is an observation just random chance?"

Hypothesis Test

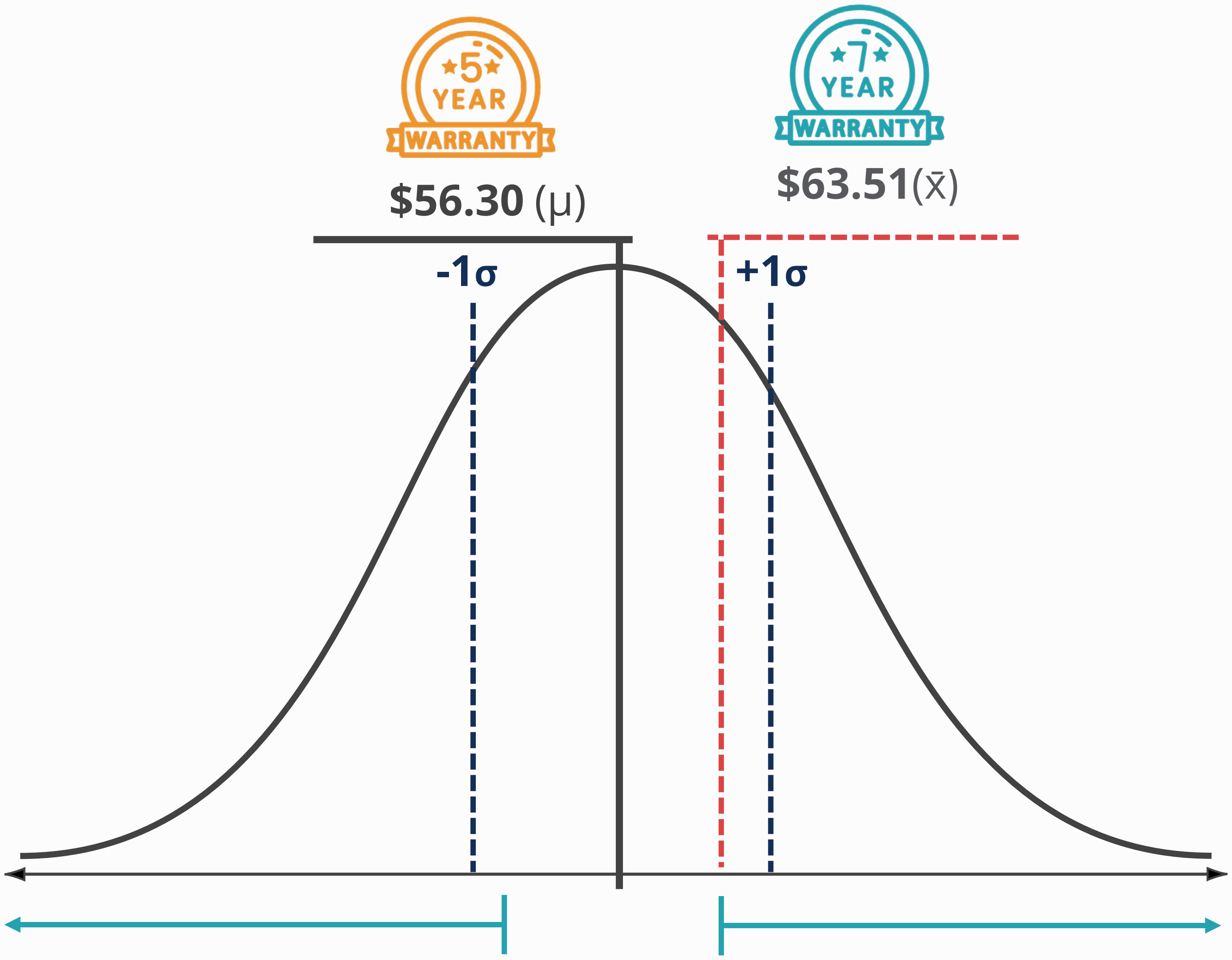
Assumption 1: Our population mean and variance is the same as sample 1 (the 5-year sample).

Assumption 2: The difference observed in our second sample is the result of random chance. There is no real difference between sample 1 and sample 2.

What the test tells us:
Our test will tell us how improbable it is to see a difference this extreme, if our assumptions are true.

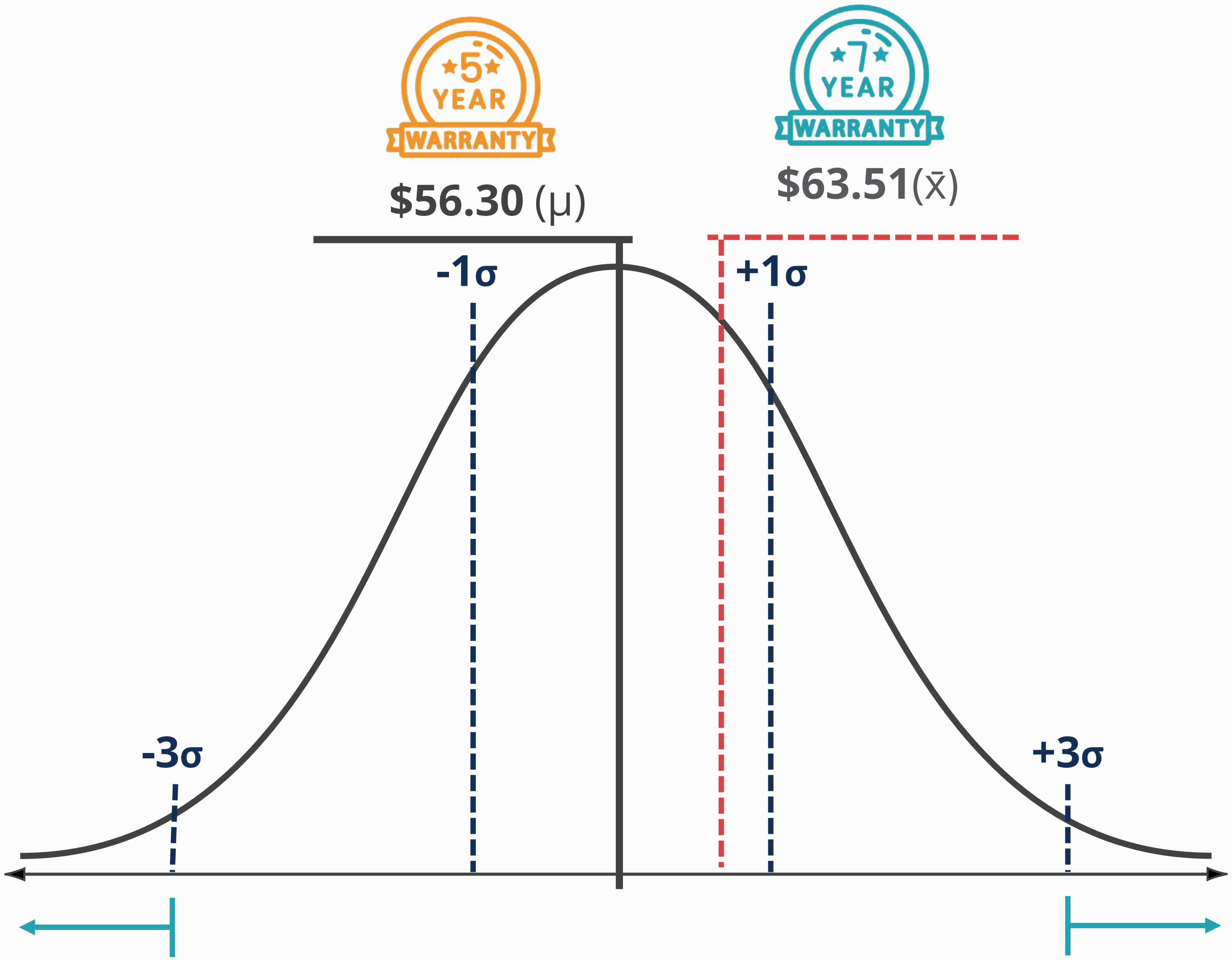
If it is highly improbable, it is likely our assumptions aren't true. There is a high probability it is more than random chance.

Hypothesis Test Explained



Likelihood of a result this extreme, given our assumptions ≈ **0.62**

Hypothesis Test Explained

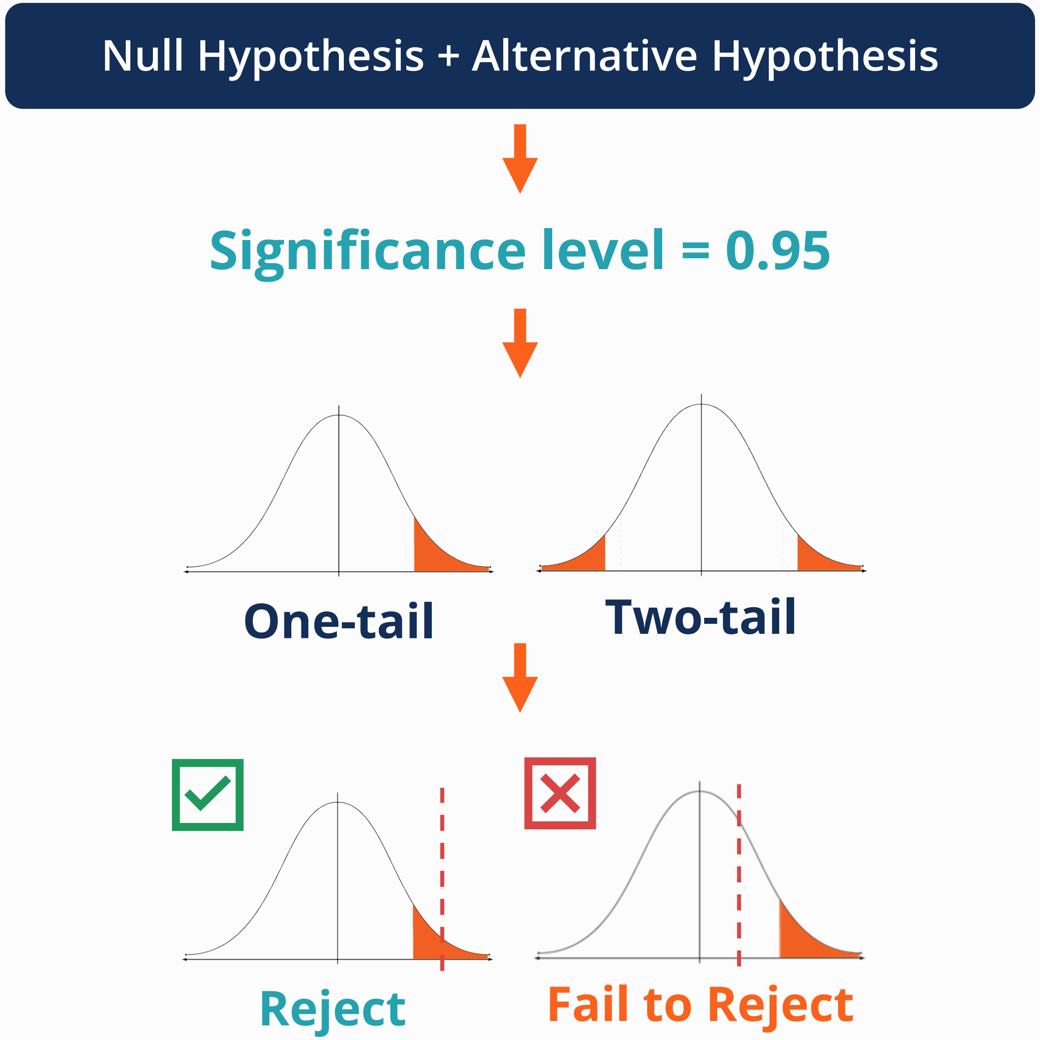


Likelihood of a result this extreme, given our assumptions ≈ **0.003**

Steps for Conducting a Hypothesis Test

The 5 Steps for Conducting a Hypothesis Test

1. State the hypotheses.
2. State the significance level.
3. Select a statistical test.
4. Apply the test and calculate the p-value.
5. Draw a conclusion about the null hypothesis.



Stating the Hypotheses

Null Hypothesis (H_0) vs Alternative Hypothesis (H_a)

Null Hypothesis (H_0)

- States that there is no difference between two parameters

Alternative Hypothesis (H_a)

- There is a difference between the two parameters
- **Sometimes, we may declare which direction that difference leans**

$\neq$

Setting Your Confidence Level and Alpha

Confidence Levels

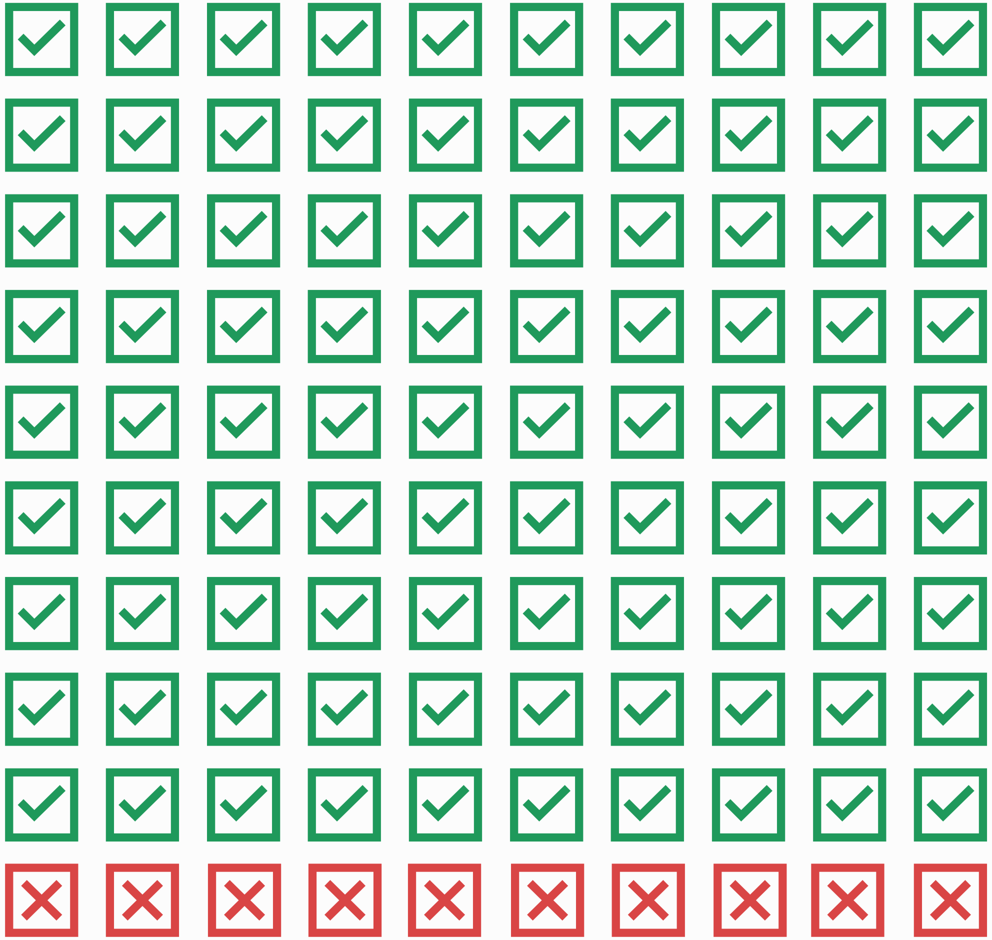
Our **confidence level** sets the probability of our test correctly failing to reject the null hypothesis—assuming the null hypothesis is true.

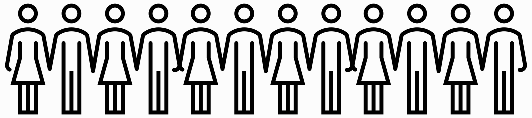
Confidence level = 95%

If there is no difference between the populations, our test will correctly conclude this 19 out of 20 times.

Confidence Levels

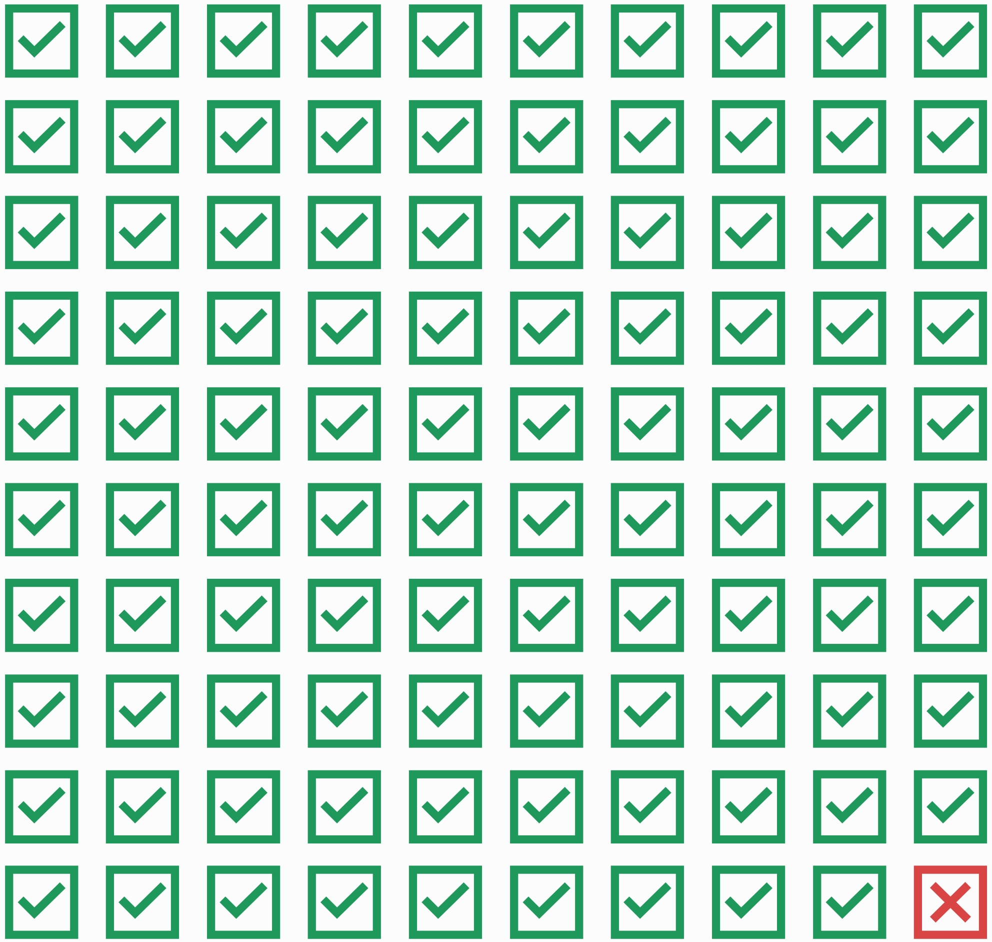
Confidence level = 90%

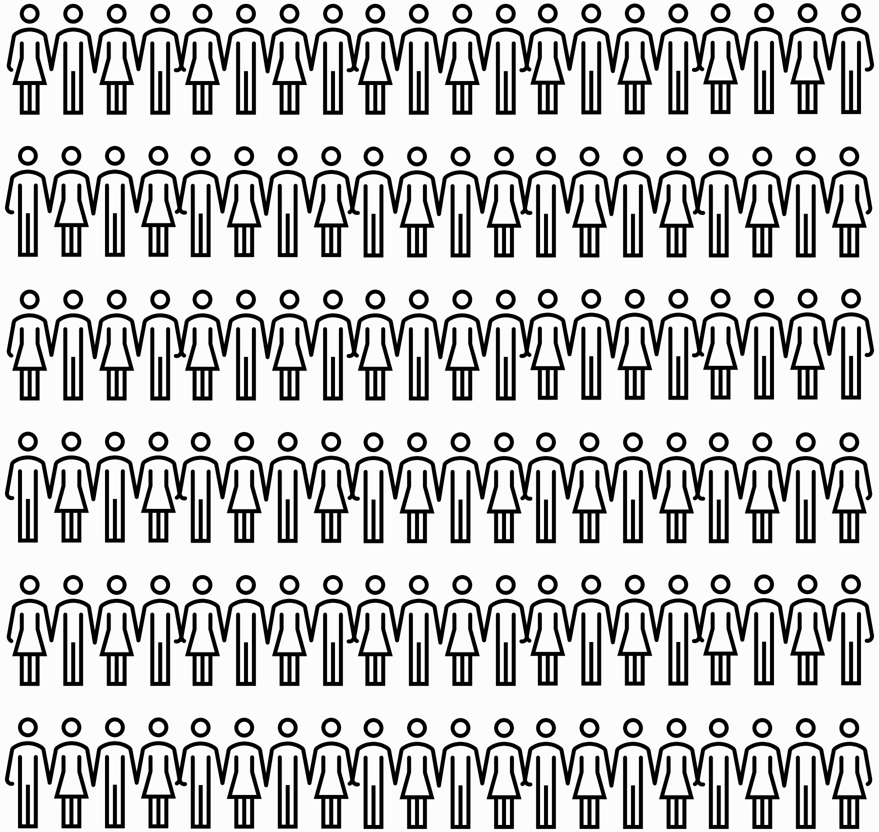




Confidence Levels

Confidence level = 99%





Common Confidence Levels

90%

α = 0.1

- Cheap and quick
- 1 in 10 results are expected to be Type I Errors

95%

α = 0.05

- Commonly used
- 1 in 20 results are expected to be Type I Errors

99%

α = 0.01

- Provides higher confidence
- 1 in 100 results are expected to be Type I Errors

Alpha Value

1 – confidence level = **alpha value**

↓

1 – 0.95 = **0.05**

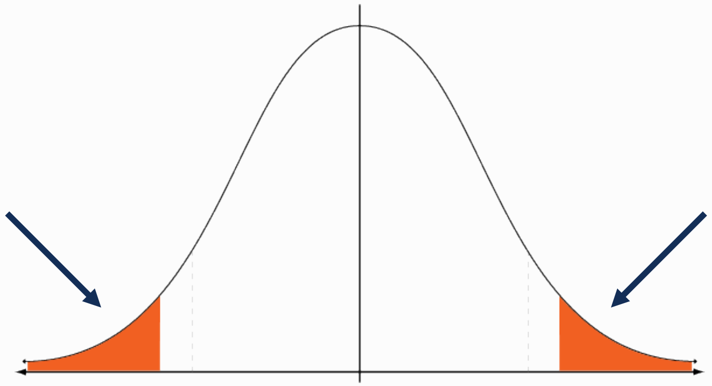
Two-Tail Test

Selecting Your T-test

H_a = mean of 7-year warranty ≠ mean of 5-year warranty

Confidence level = 95%

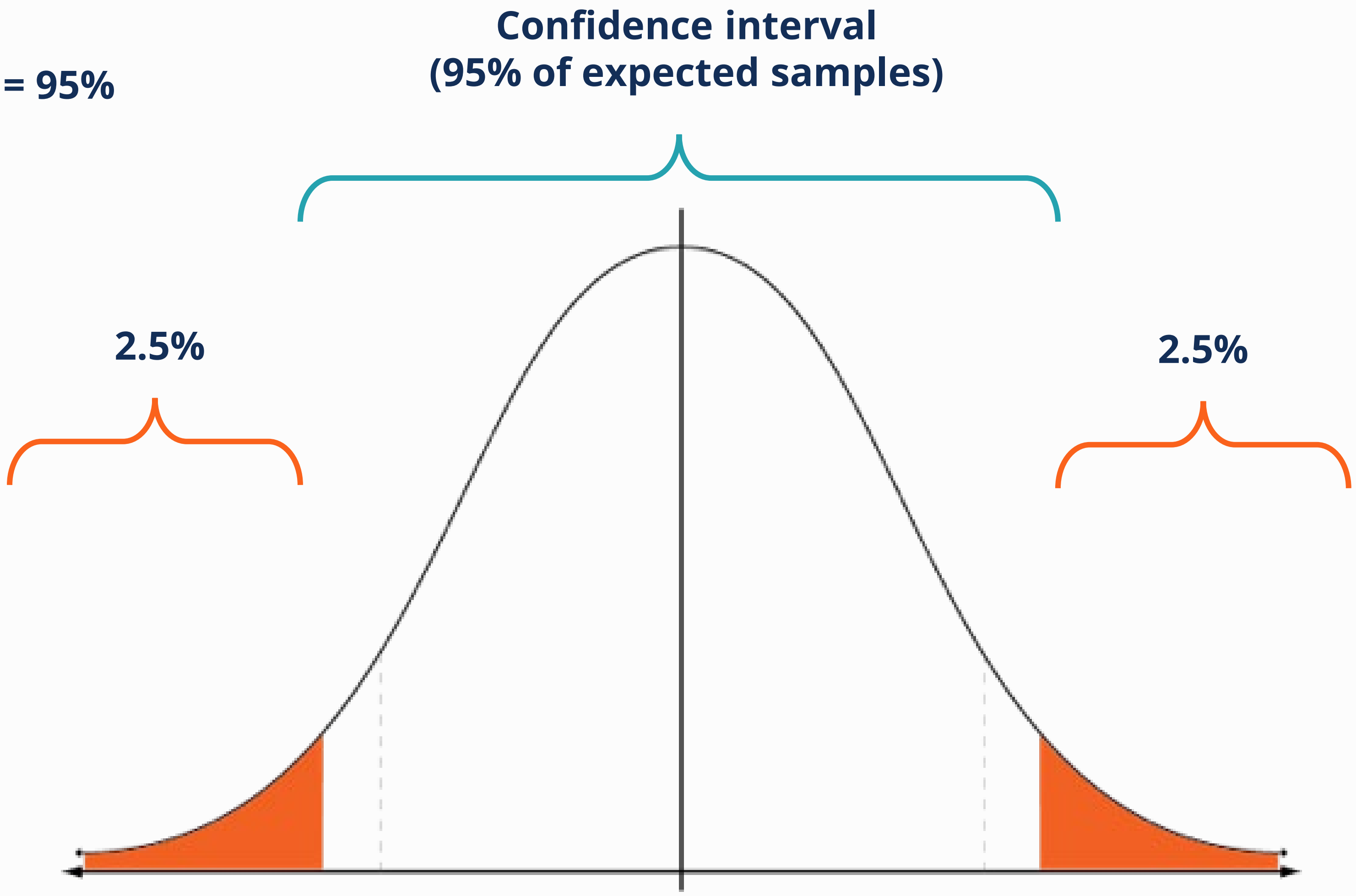
Alpha value = 5%



Two-Tail

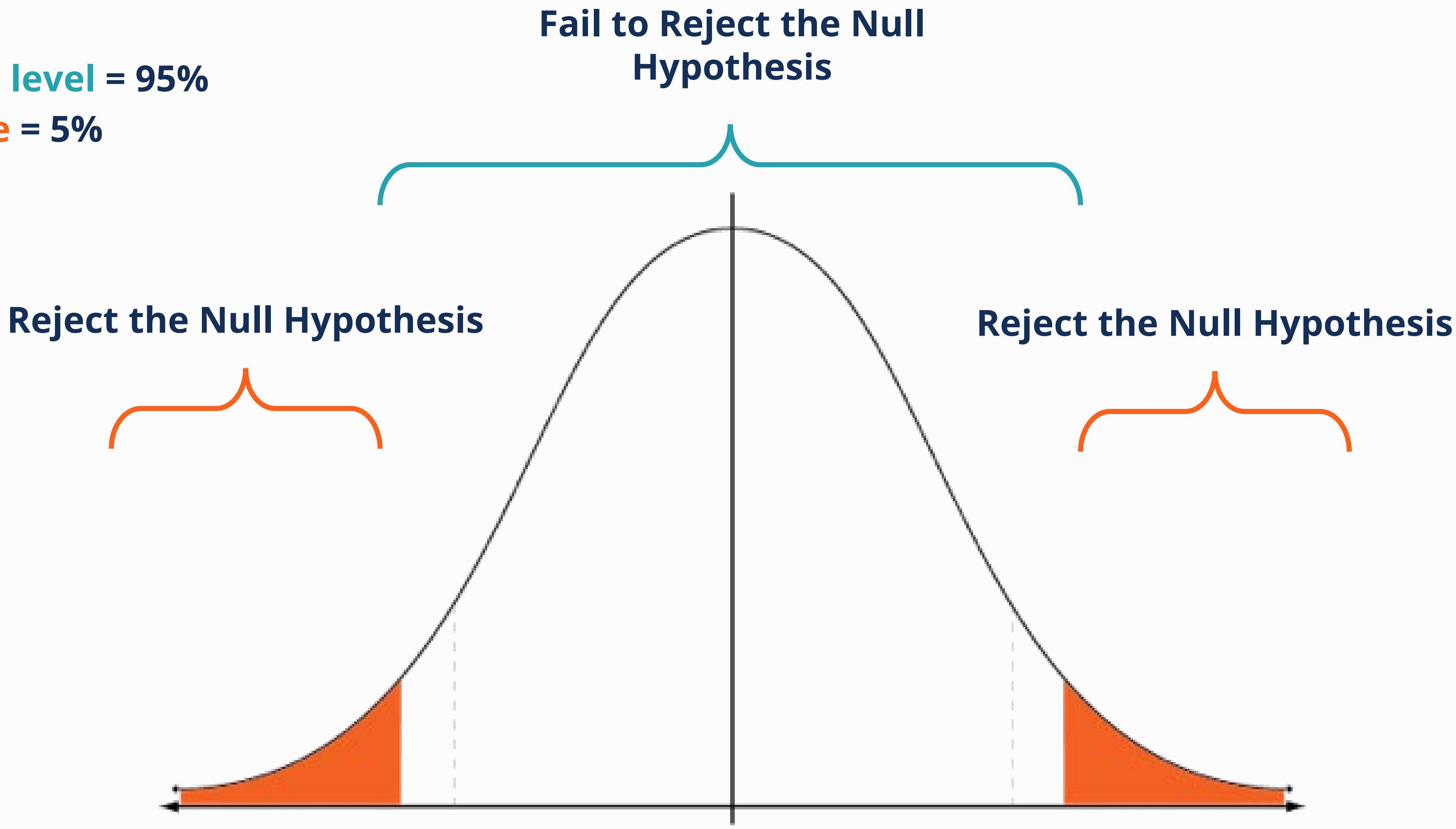
Rejecting and Failing to Reject H_0

Confidence level = 95%
Alpha value = 5%



Rejecting and Failing to Reject H_0

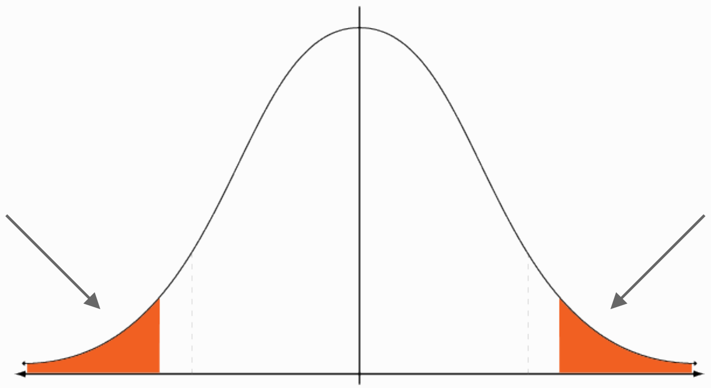
Confidence level = 95%
Alpha value = 5%



Selecting Your T-test

Two-Tail

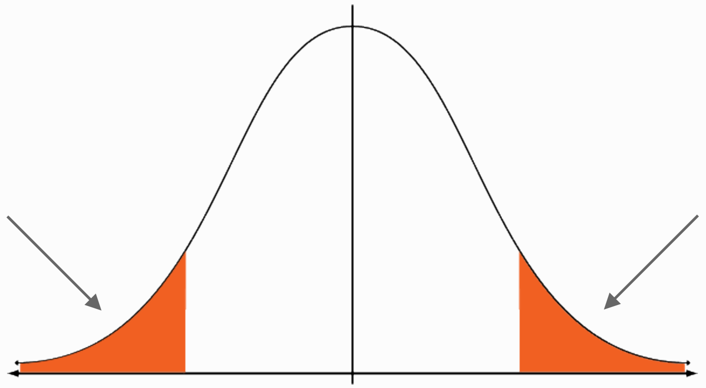
- H_0 is 5-year full = 7-year limited
- H_a is 5-year full ≠ 7-year full



- The null hypothesis is rejected if the sample falls outside a range greater or less than a certain value.

One-Tail

- H_0 is no warranty $\lesseqgtr$ 5-year full
- H_a is no warranty $\lessgtr$ 5-year full



- The null hypothesis is rejected if the result is greater than or less than a certain value, but not both.

Calculating Our P-Value

Calculating Our P-Value

+

0.05

+

Test Data | **Alpha (α)** | **Statistical Test**

Calculating Our P-Value

Option 1

P-value Table

	Alpha		
N	0.05	0.025	0.01
10	2.036	2.345	2.863
20	2.004	2.189	2.54
30	2.075	2.521	2.548
45	2.146	2.349	2.733
60	2.155	2.352	2.66
90	2.216	2.455	2.719
120	2.152	2.427	2.727

Scientific Calculator

Option 2

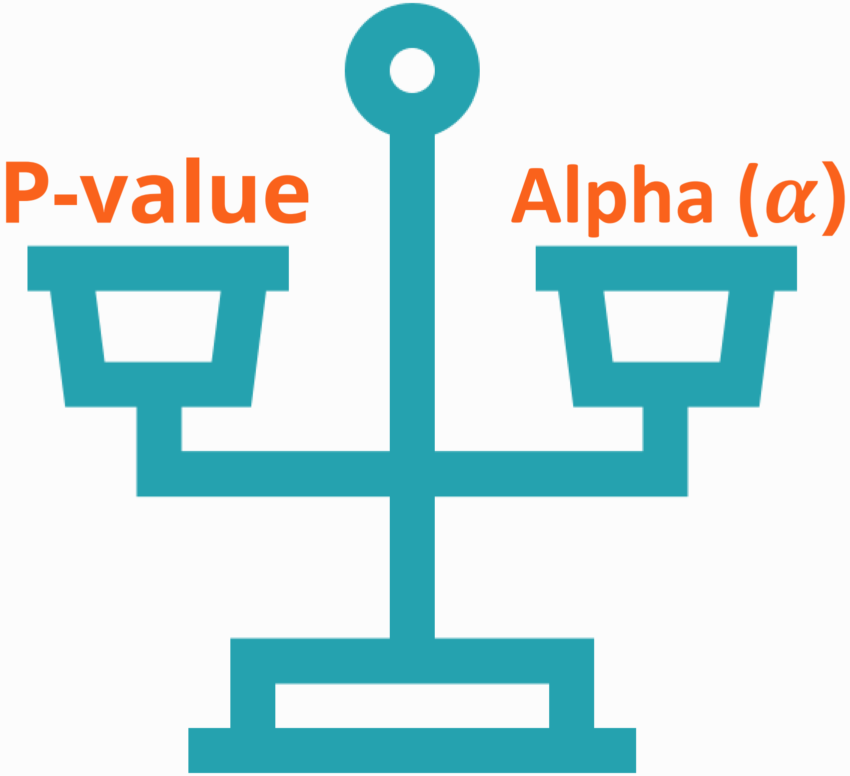
Python

R

Excel

Drawing Our Conclusion

Calculating Our P-Value



Rejecting or Failing to Reject H_0

If P-value $\leq$ Alpha (α), we **reject** the null hypothesis.

If P-value $>$ Alpha (α), we **fail to reject** the null hypothesis.

Reject the Null Hypothesis

The null hypothesis is unlikely. There is sufficient evidence that there is a meaningful difference between the samples.

Fail to Reject the Null Hypothesis

The sample did not provide sufficient evidence to confidently conclude the alternative hypothesis.

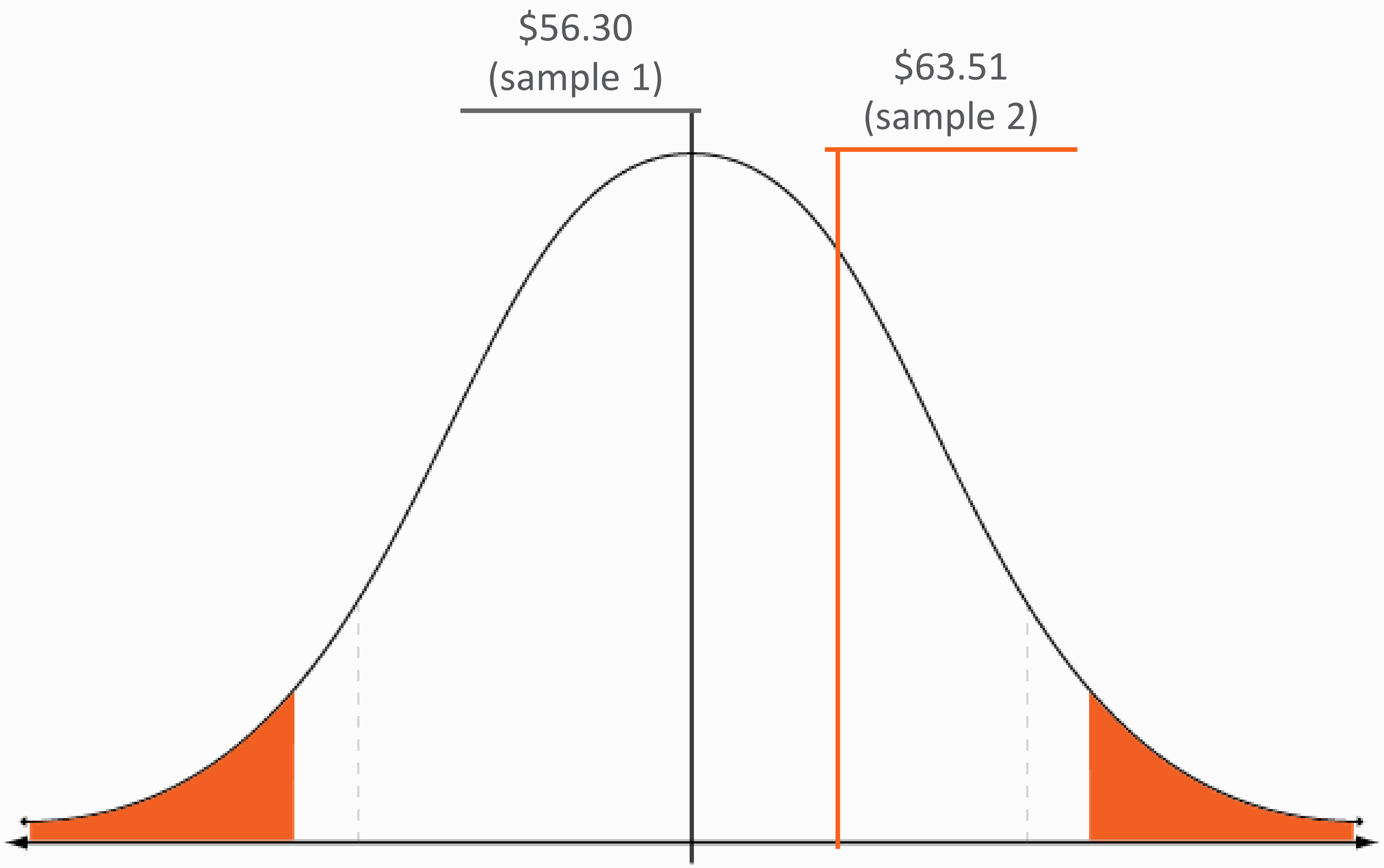
Rejecting or Failing to Reject H_0

$$0.423 > 0.05$$

Fail to Reject the Null Hypothesis

The sample did not provide sufficient evidence to confidently conclude the alternative hypothesis.

Results



Rejecting or Failing to Reject H_0

Reject the Null Hypothesis

Fail to Reject the Null Hypothesis

~~**Accept** the Null Hypothesis~~

Rejecting or Failing to Reject H_0

All flamingos are pink.

Proving this statement requires observing ALL the flamingos in the world.

Rejecting it requires observing one flamingo that isn't pink.

Reporting Our Results

Our Results

Failed to reject the null hypothesis.

- Despite a $7 difference between means, our test wasn't sensitive.
- We aren't confident this difference isn't simply random chance.

Reporting Our Results

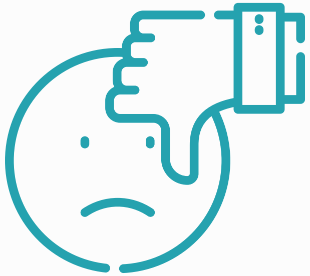
Normally a sample sizes over 100 would be quite good, but:

- **Populations with a low rate of positive cases can require a larger sample size.** Our population has a low rate of positive cases (~10%).
- **Populations with a higher variance also generally require higher sample sizes.** Our population has a large variance. The average variation between orders is over $174.
- **Together, a high variance and low rate of positive cases will really hurt the sensitivity of the test.**

Avoid Re-dos

As a best practice always **record and report** findings. **DO NOT** do any of the following when you are unhappy with the results:

- Pull a new sample of the same size to see if our answer changes.
- Go back and remove outliers to see if it becomes significant.
- Change up the variables, like looking at only results from the US.
- Add a few more samples in to see if it becomes significant.



These are examples of p-hacking. Each of these actions dramatically increase the chance of a false positive.

One-tail T-test Example

Do Warranties Sell Vacuums Better?

Test #2

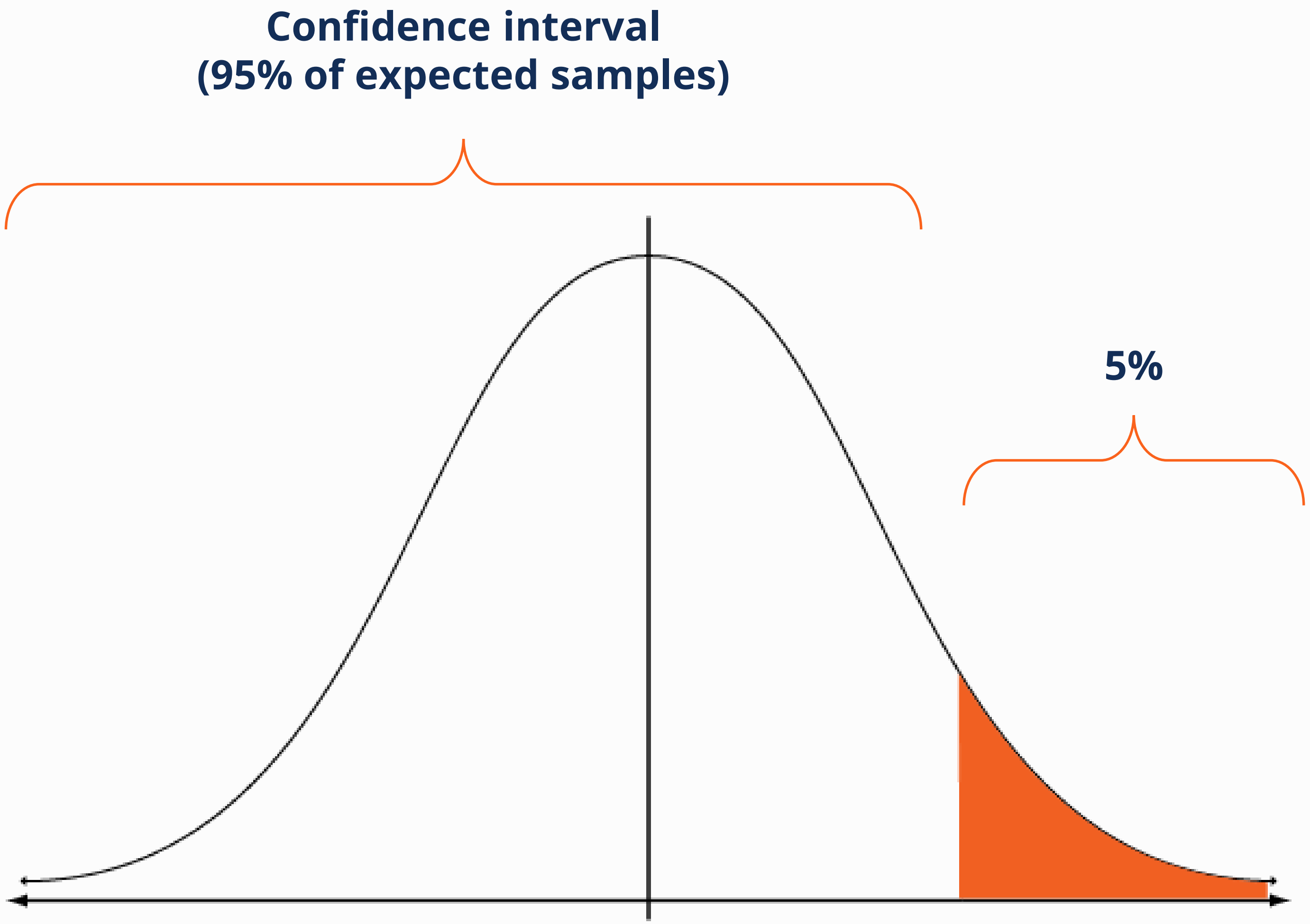
Sample size = 750 visitors

H_0 = No Difference

H_a = Warranty > No warranty

	A	B	C
1	Subject ID	Sample Group	Order Value
2	1	5-year full	0
3	2	none	0
4	3	7-year limited	0
5	4	5-year full	0
6	5	none	593
7	6	none	0
8	7	7-year limited	0

One-tail T-test



Chapter 5: Errors and Estimations

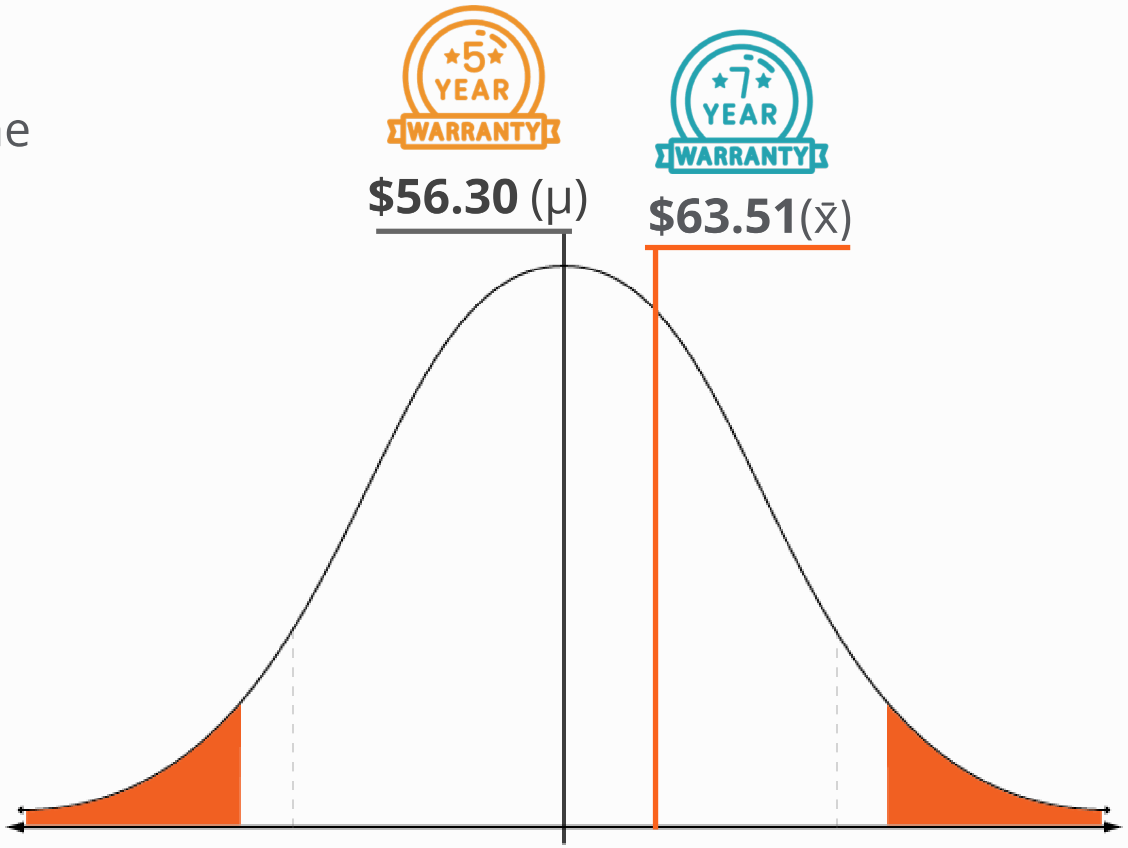
Two-tail T-test Results

Test 1 | 5 Year Full-warranty vs 7-year Limited-warranty

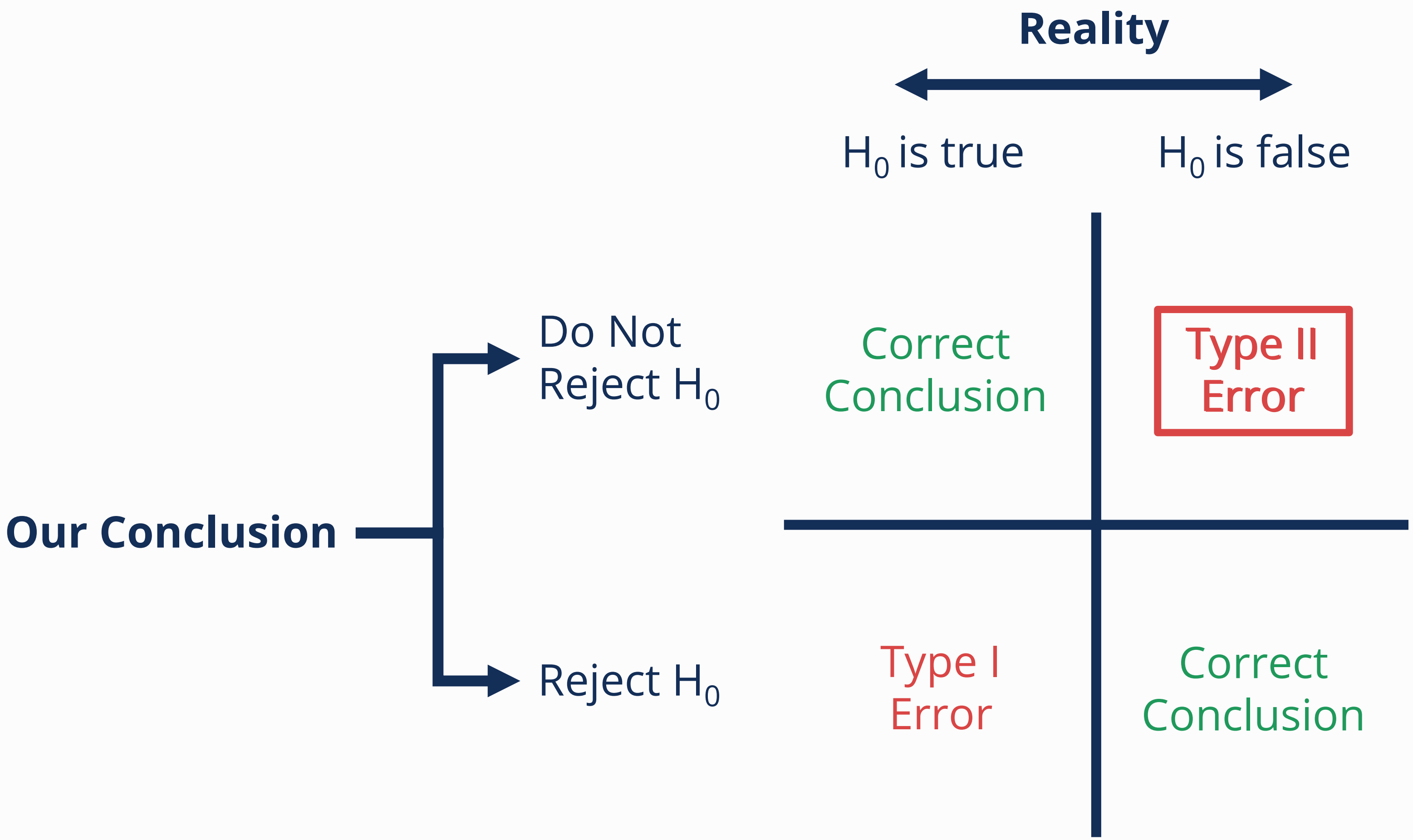
Failed to reject the null hypothesis.

Even though we failed to reject the null, the null hypothesis is **NOT** proven.

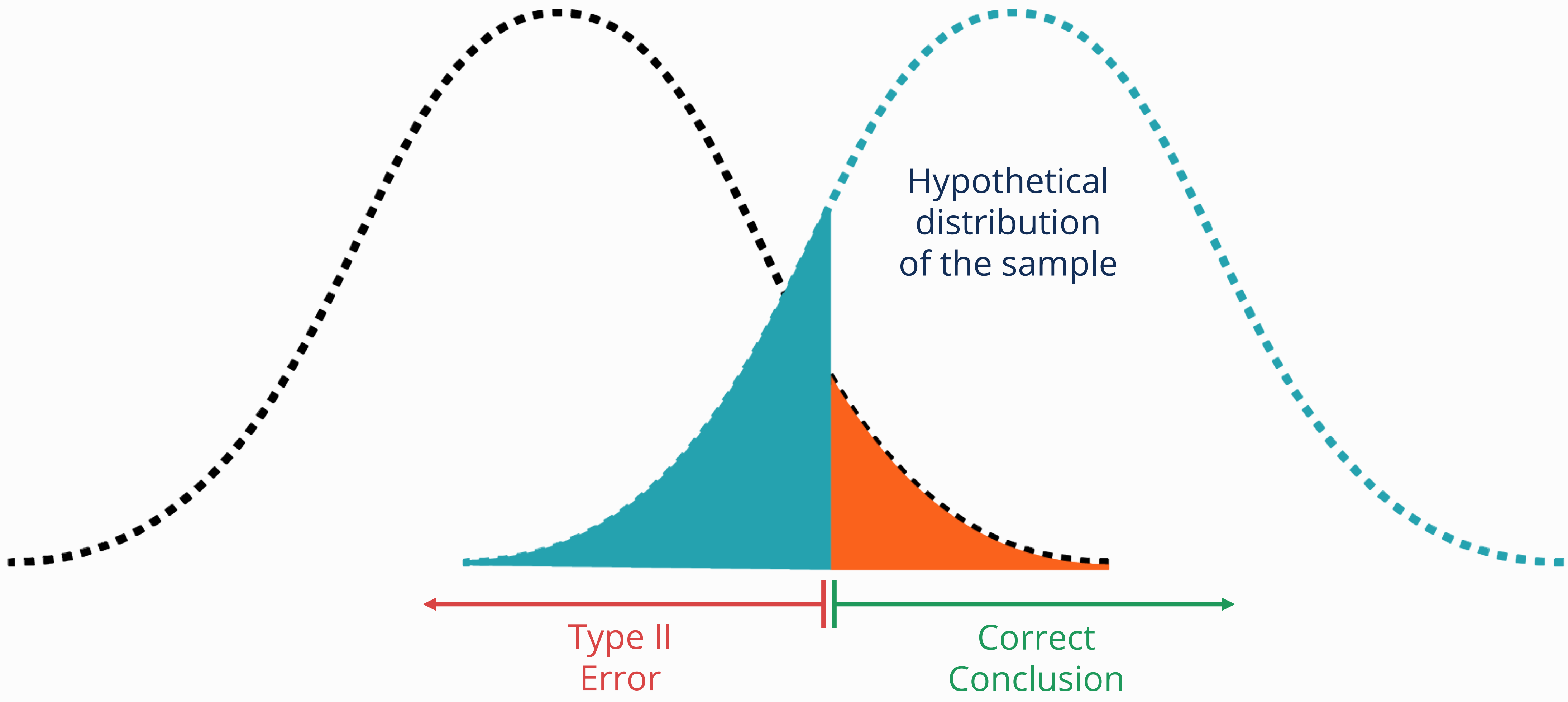
A more sensitive test might still lead us to conclude the **alternative hypothesis**.



Test 1 | 5 Year Full-warranty vs 7-year Limited-warranty

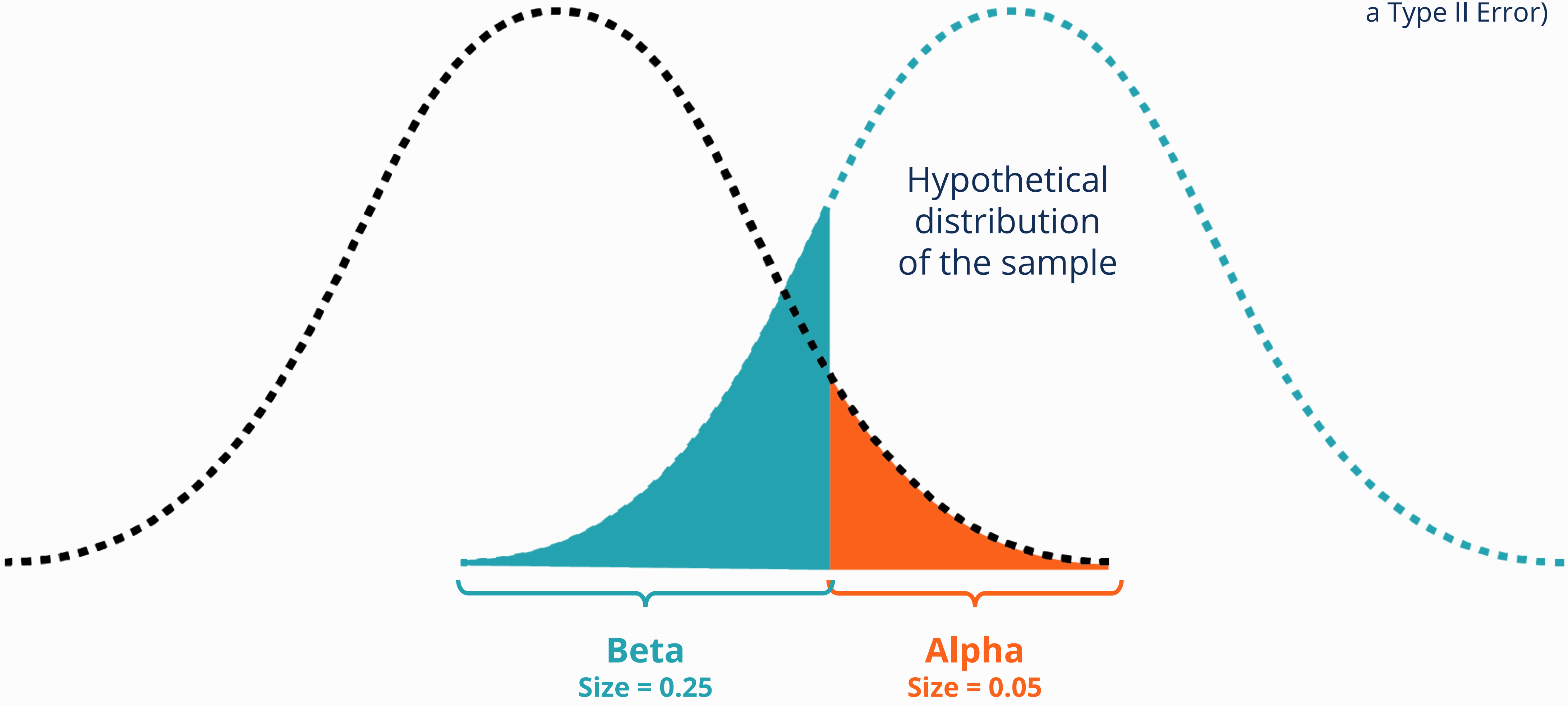


Test 1 | 5 Year Full-warranty vs 7-year Limited-warranty



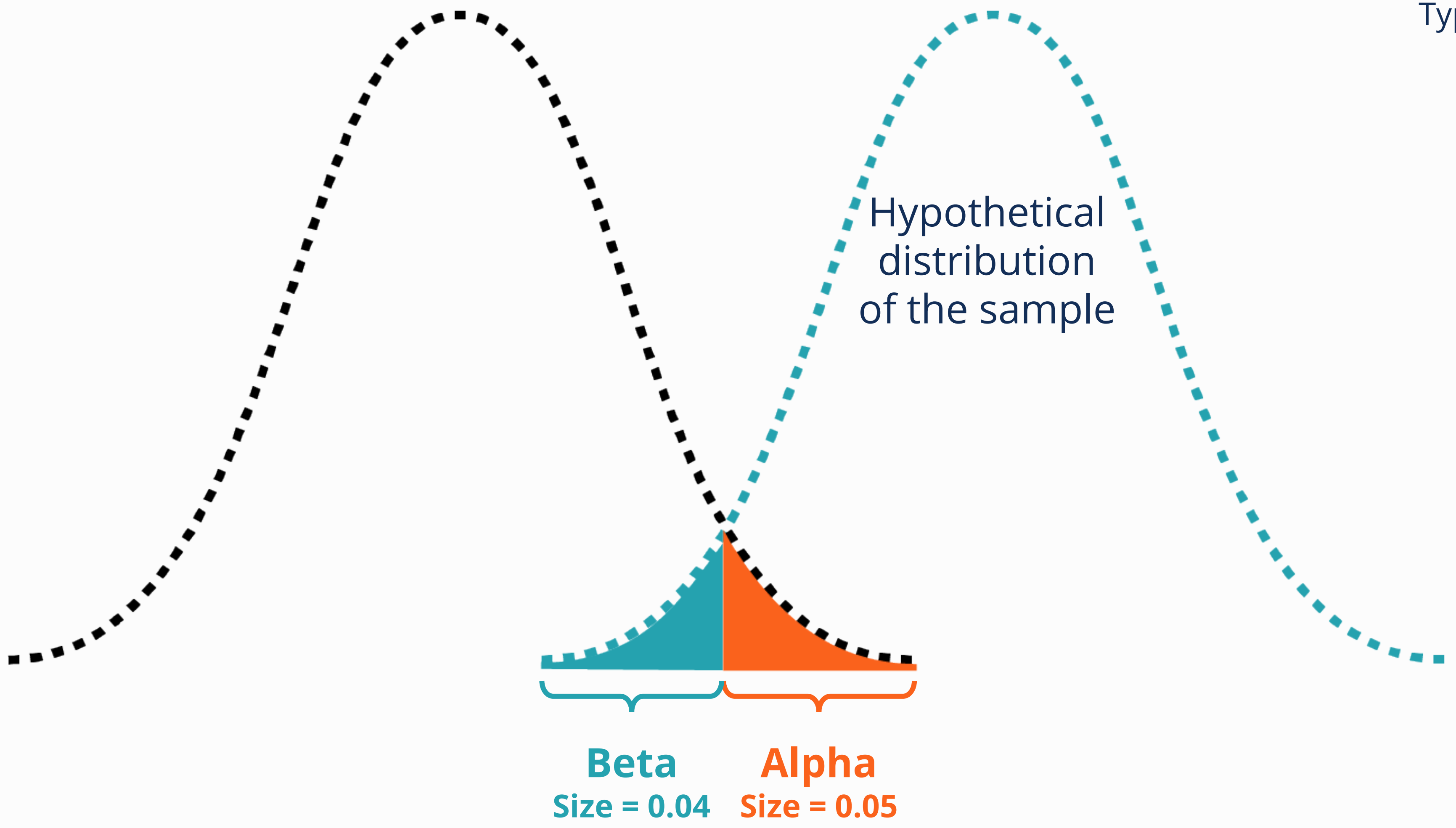
Test 1 | 5 Year Full-warranty vs 7-year Limited-warranty

1 - beta = **0.75**
(75% chance of avoiding a Type II Error)



Test 1 | 5 Year Full-warranty vs 7-year Limited-warranty

1 - **beta** = **0.96**
(96% chance of avoiding Type II Error)

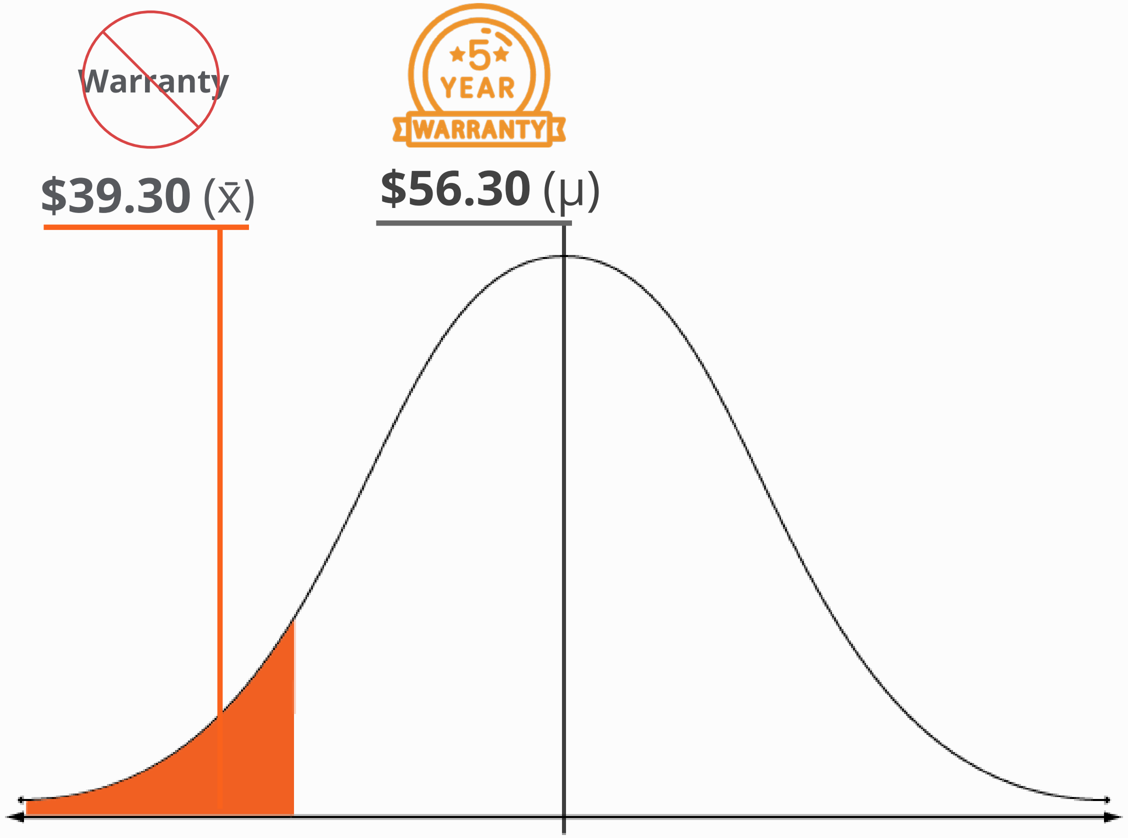


One-tail T-test Results

Test 2 | 5 Year Full-warranty vs No Warranty

Rejected the null hypothesis.

It is unlikely the measured difference between these samples was not random.

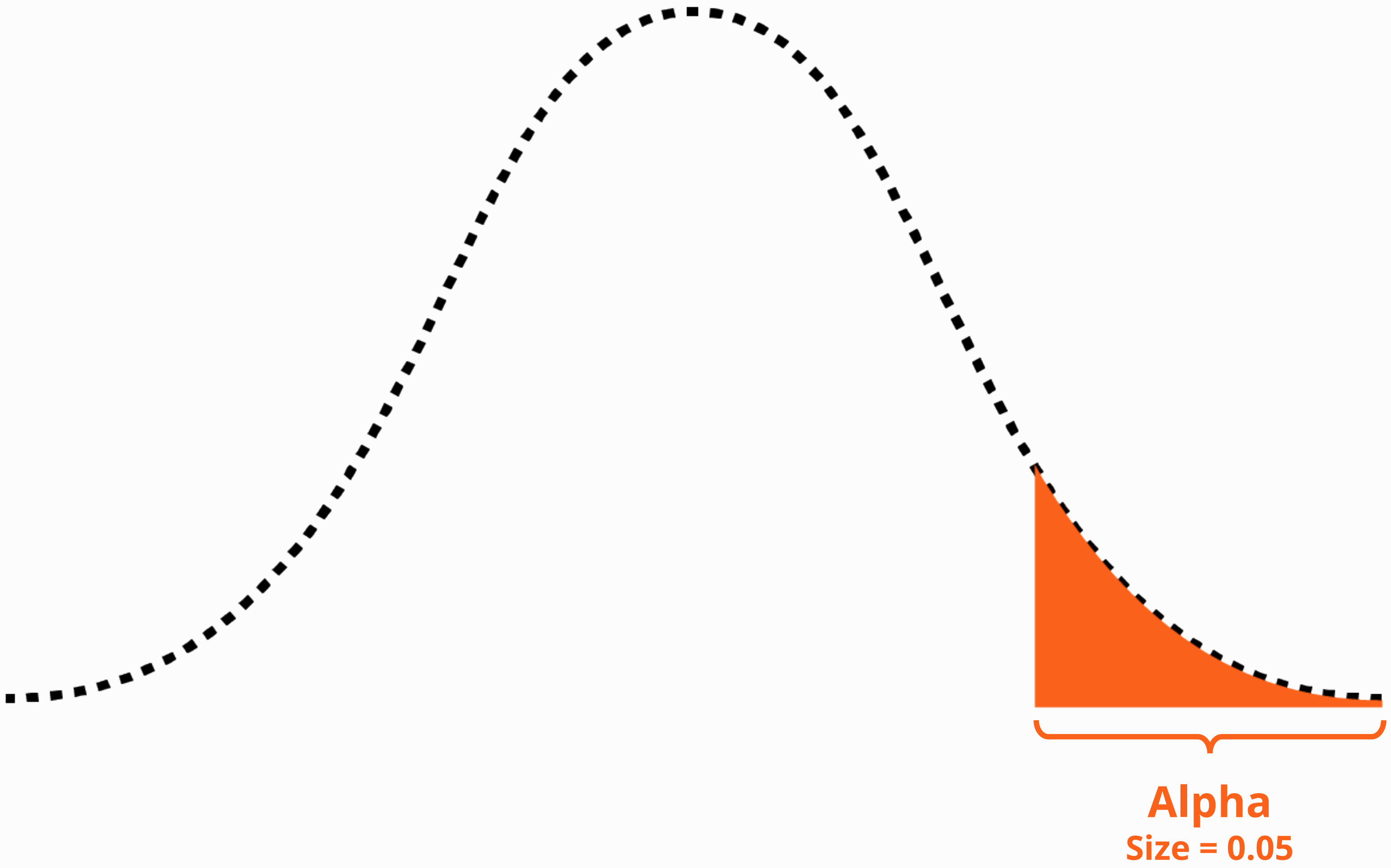


Test 2 | 5 Year Full-warranty vs No Warranty

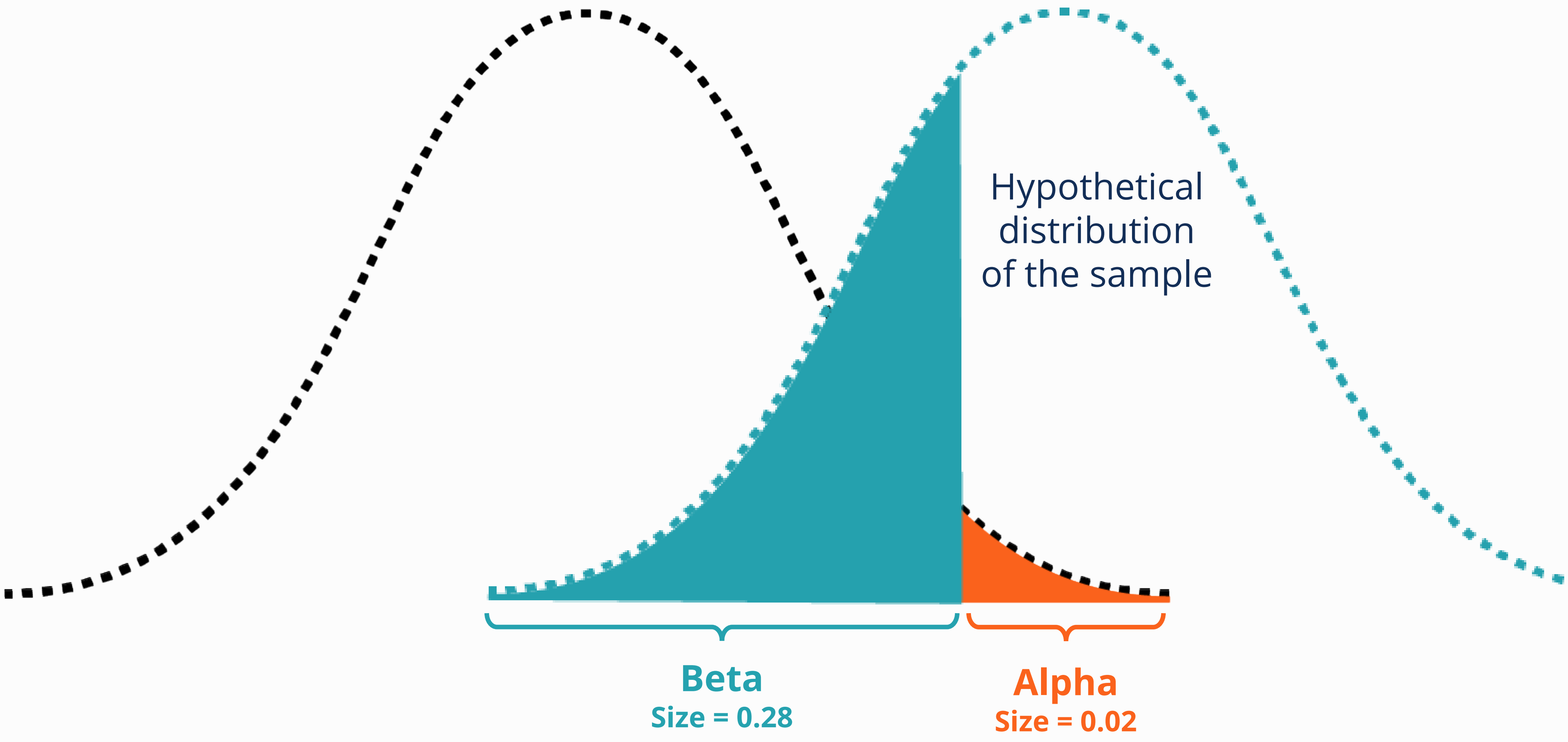
Reality

Our Conclusion	H_0 is true	H_0 is false
Do Not Reject H_0	Correct Conclusion	Type II Error
Reject H_0	Type I Error	Correct Conclusion

Test 2 | 5 Year Full-warranty vs No Warranty



Test 2 | 5 Year Full-warranty vs No Warranty



Type I Error vs Type II Error

Scenario: Type I Error vs Type II Error

Imagine you're in a courtroom and you've been asked to run a hypothesis test. The results will be used to decide if an alleged bank robber is guilty.

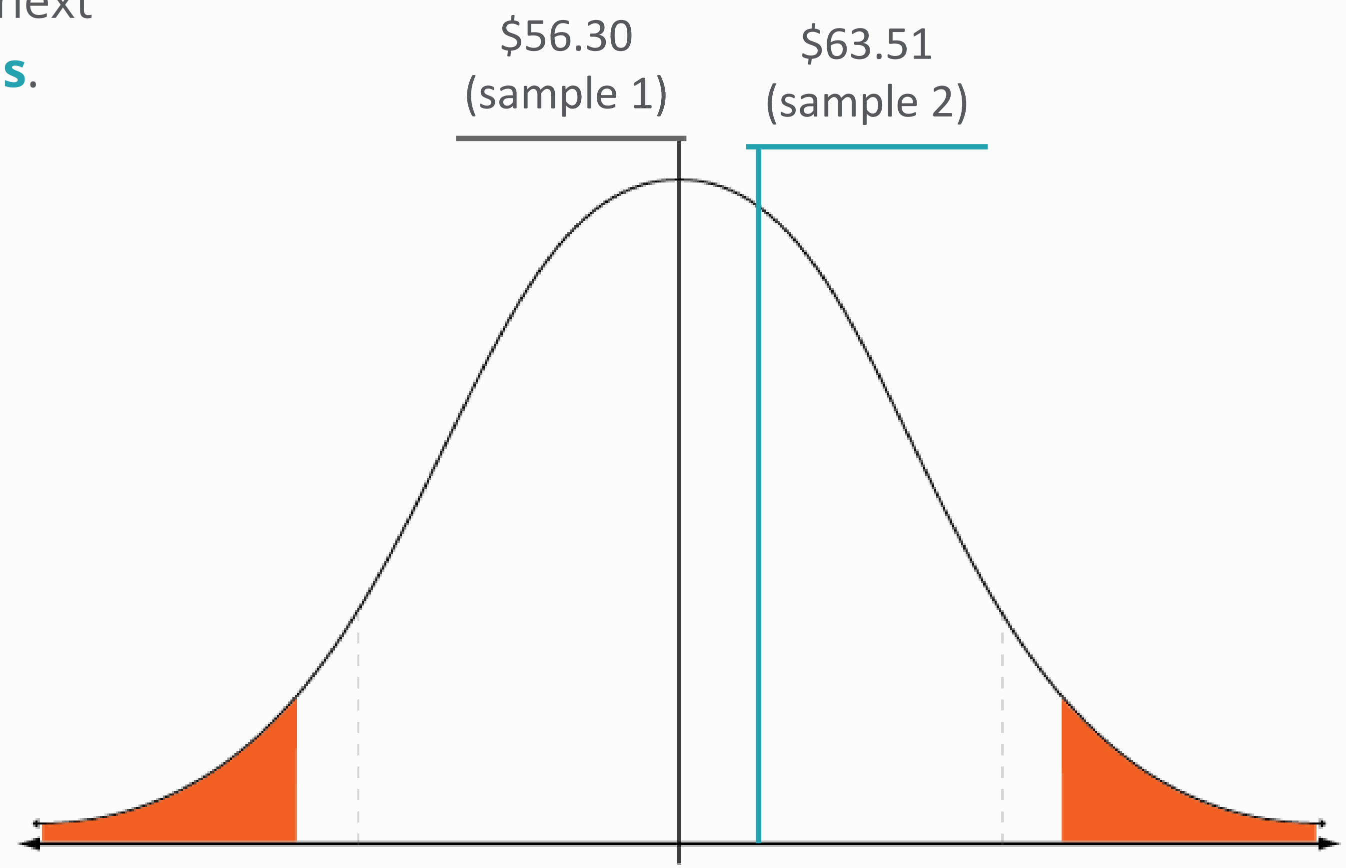
Type II Error = a guilty person goes free.

Type I Error = an innocent person is punished.

Power Analysis: Effect Size

Running a Power Analysis

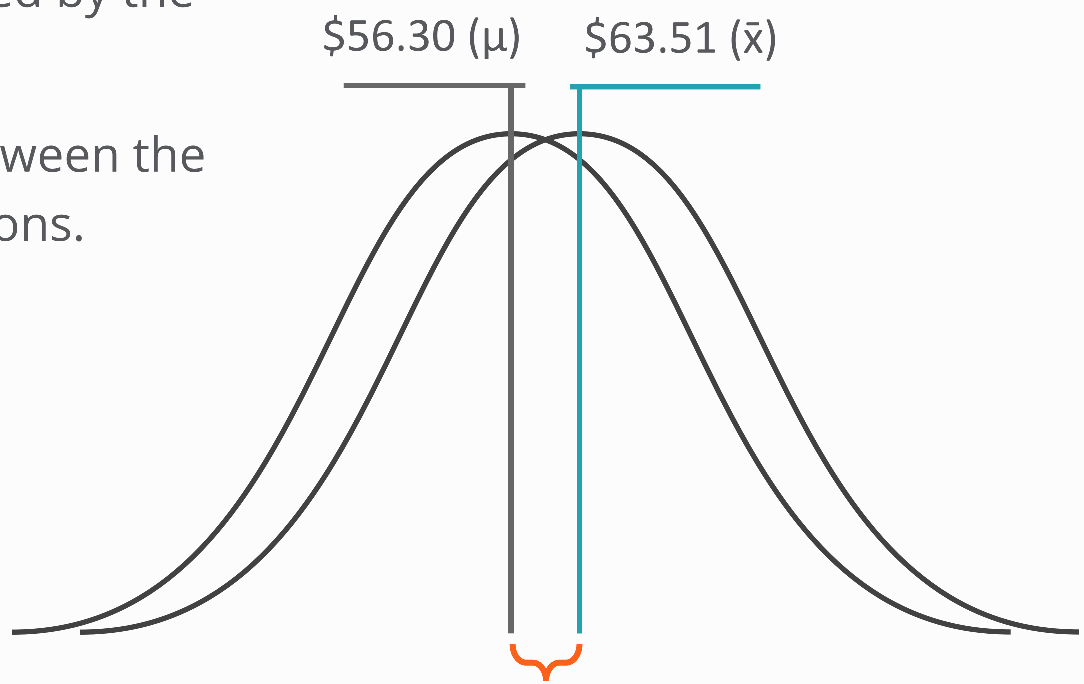
To determine the sample size for our next experiment we'll use a **Power Analysis**.



Determining Effect Size

To run a **Power Analysis**, we need to know the distance between our means and the overlap of our samples distributions. This is expressed by the **Effect Size**.

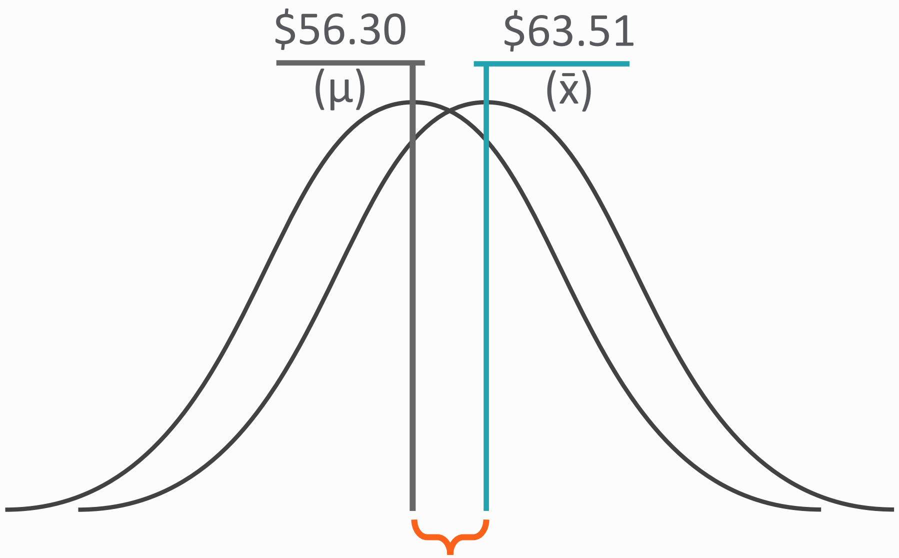
The **Effect size** describes the distance between the means of our samples in standard deviations.



Effect size = distance between means in standard deviations

Cohen's D

$$\textbf{Effect size (cohen's d)} = \frac{Mean\ 2 - Mean\ 1}{Standard\ deviation} = \frac{63.51 - 56.3}{174.3} = \mathbf{0.041}$$



Effect size = 0.041

Cohen's D

Relative size	Effect size	% of sample 1 below the mean of sample 2
	0.0	50%
Small	0.2	58%
Medium	0.5	69%
Large	0.8	79%
Huge	1.4	92%

Power Analysis: Statistical Power

Statistical Power

Statistical power, or just **power**, is the chance that our statistical test correctly rejects the null hypothesis if the difference is real.

Power = 1 – **beta (β)**

0.8

β = 0.2

- Commonly used
- 1 in 5 results are expected to be Type II Errors

0.9

β = 0.1

- Gives test a higher chance of finding a difference if it exists
- 1 in 10 results are expected to be Type II Errors

Elements for Our Power Analysis

Mean 1 (μ) = 56.296

Mean 2 ($\bar{x}$) = 63.508

Standard Deviation (σ)* = 174.3

Effect size (Cohen's d) = $\frac{Mean\ 2 - Mean\ 1}{Standard\ deviation}$ = 0.042

Power ($1-\beta$) = 0.8

Alpha (α) = 0.05

*Our standard deviation is based on the pooled variance (the average between both samples)

Power Analysis: Solution