

# Fundamentals of R

Data Boot Camp  
Lesson 15.2



# The Big Picture



# This Week: Fundamentals of R

---

By the end of this week, you'll know how to:



Load, clean up, and reshape datasets using tidyverse in R



Visualize datasets with basic plots such as line, bar, and scatter plots using ggplot2



Generate and interpret more complex plots such as boxplots and heatmaps using ggplot2



Plot and identify distribution characteristics of a given dataset



Formulate null and alternative hypothesis tests for a given data problem



Implement and evaluate simple linear regression and multiple linear regression models, as well as a chi-squared test for a given dataset



Implement and evaluate the one-sample t-tests, two-sample t-tests, and analysis of variance (ANOVA) models for a given dataset



## **This Week's Challenge**

Using the skills learned throughout the week, students will use linear regression to predict vehicle MPG and create summary statistics to compare vehicle manufacturing lots.

Module 15

# Today's Agenda

# Today's Agenda

---

By completing today's activities, you'll learn the following skills:

01

Using one-sample and two-sample t-tests

02

Using ANOVA

03

Performing linear and multiple linear regression

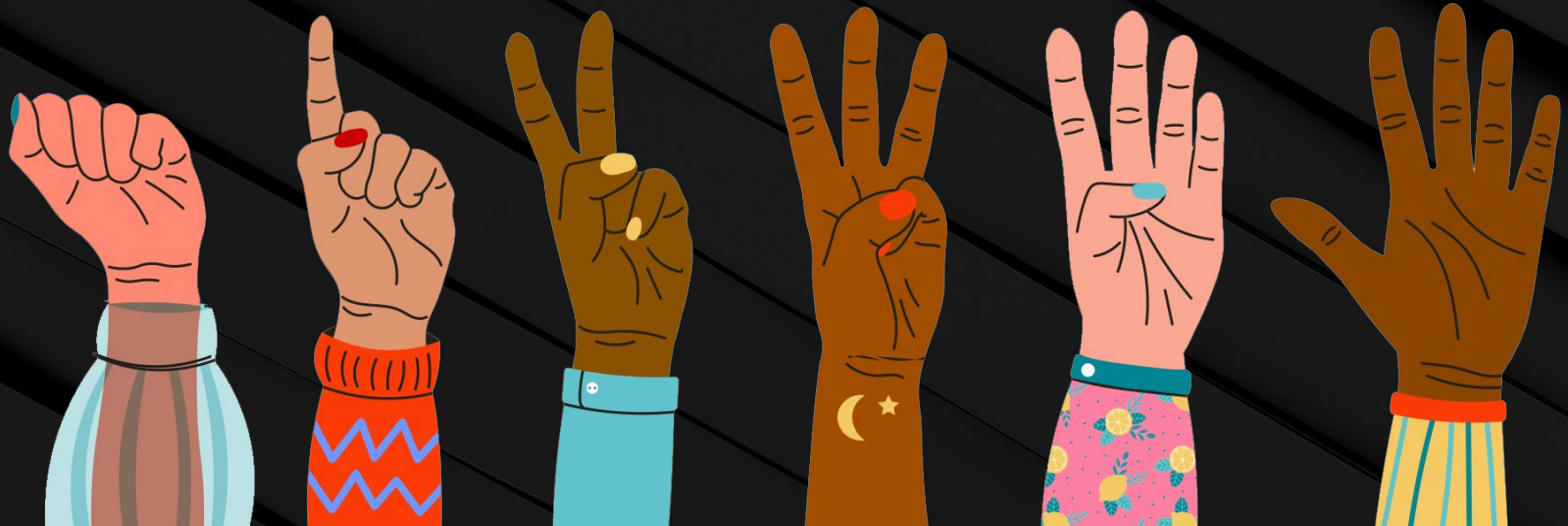


Make sure you've downloaded  
any relevant class files!

## FIST TO FIVE:

---

How comfortable do you feel with this topic?

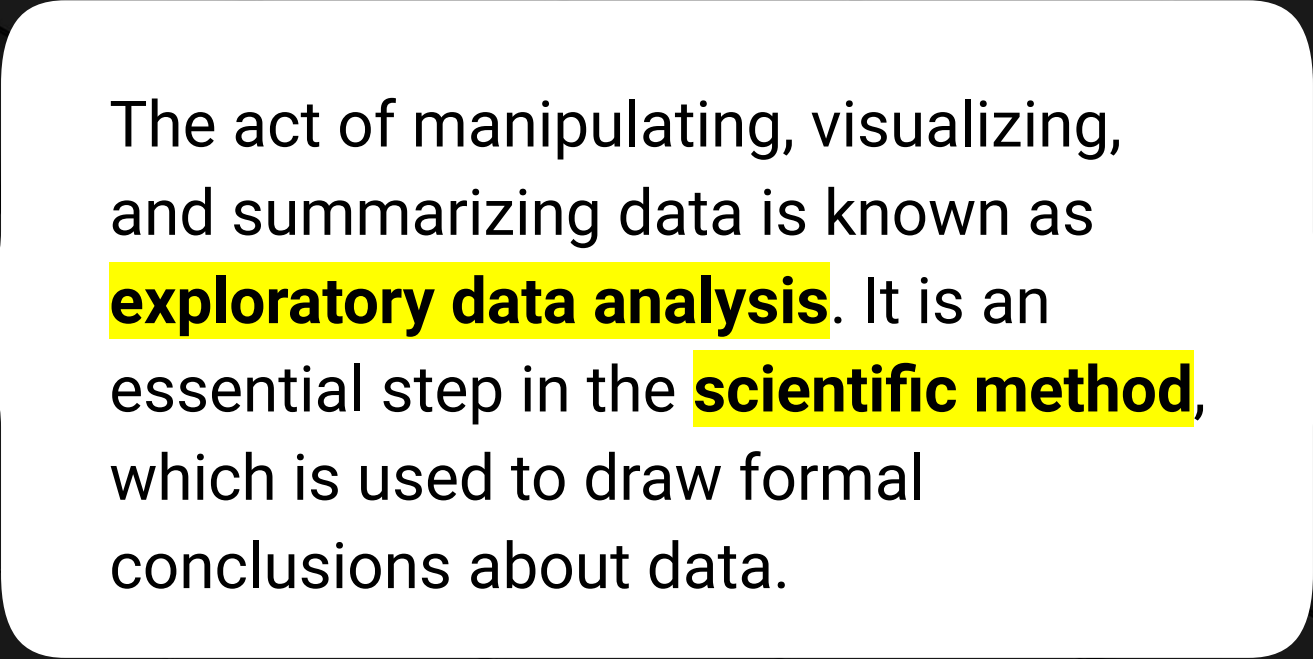


# Hypothesis Testing





**In data science, we are constantly manipulating, visualizing, and summarizing data, regardless of whether our dataset is newly collected or curated from years of data collection.**

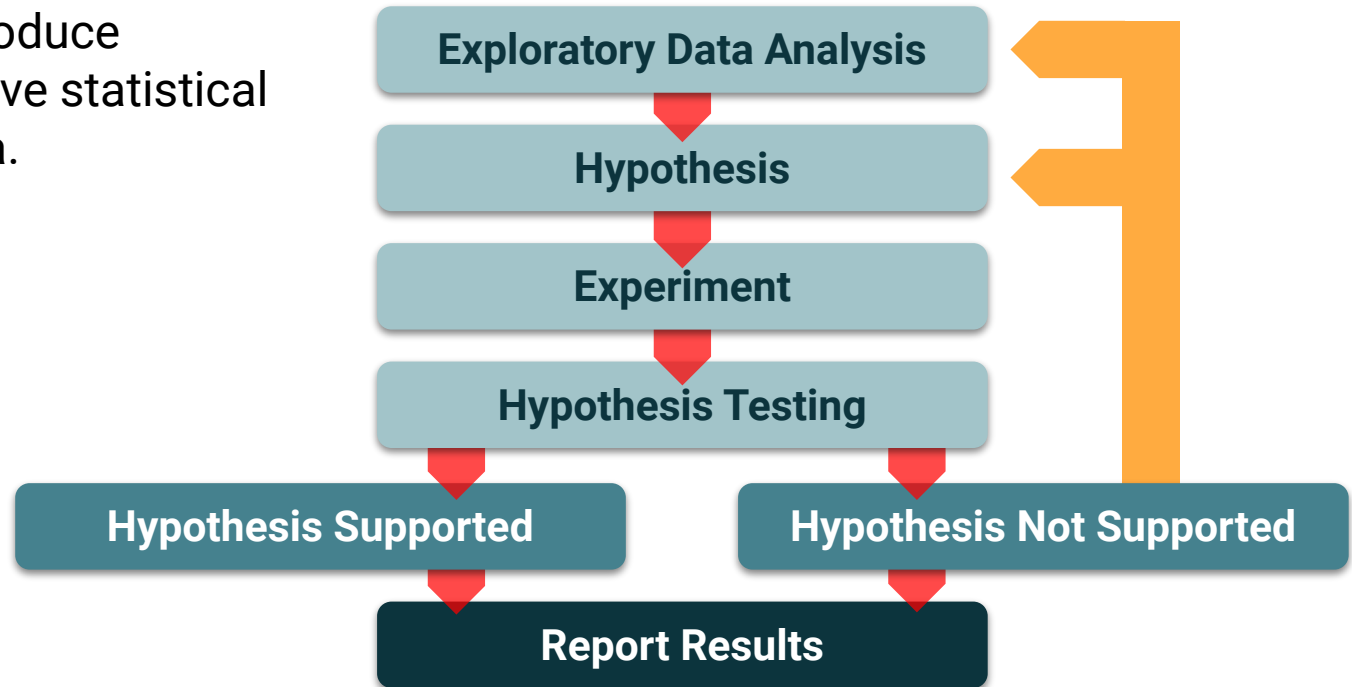


The act of manipulating, visualizing, and summarizing data is known as **exploratory data analysis**. It is an essential step in the **scientific method**, which is used to draw formal conclusions about data.

# Null and Alternative Hypotheses

---

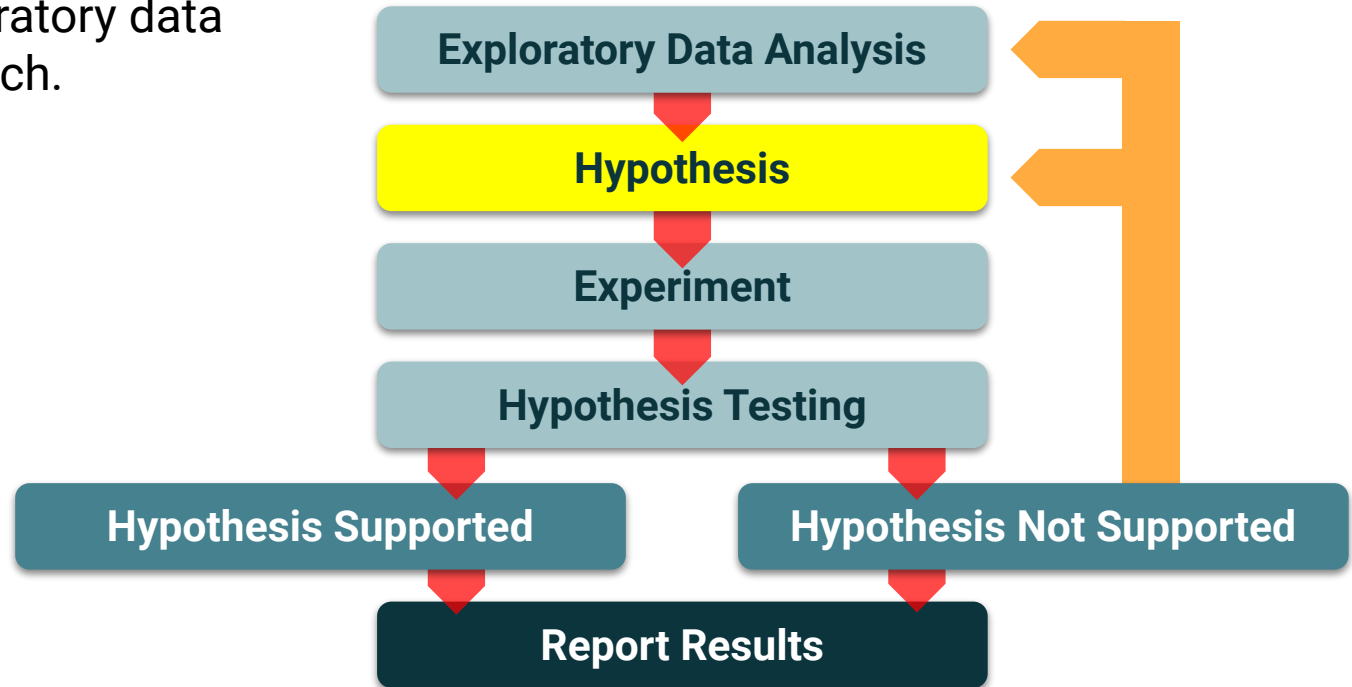
We will learn how to perform the remaining scientific method steps in order to produce powerful, quantitative statistical analysis of our data.

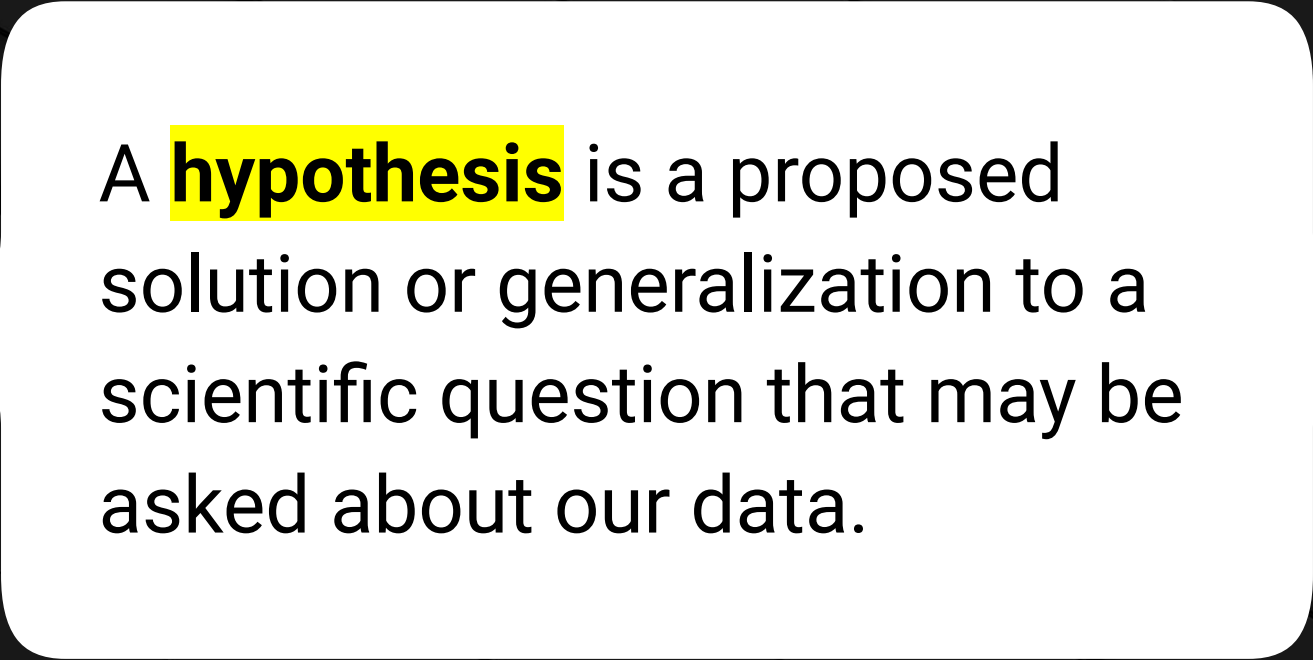


# Null and Alternative Hypotheses

---

The next step in the scientific method is building a hypothesis based on our exploratory data analysis and research.





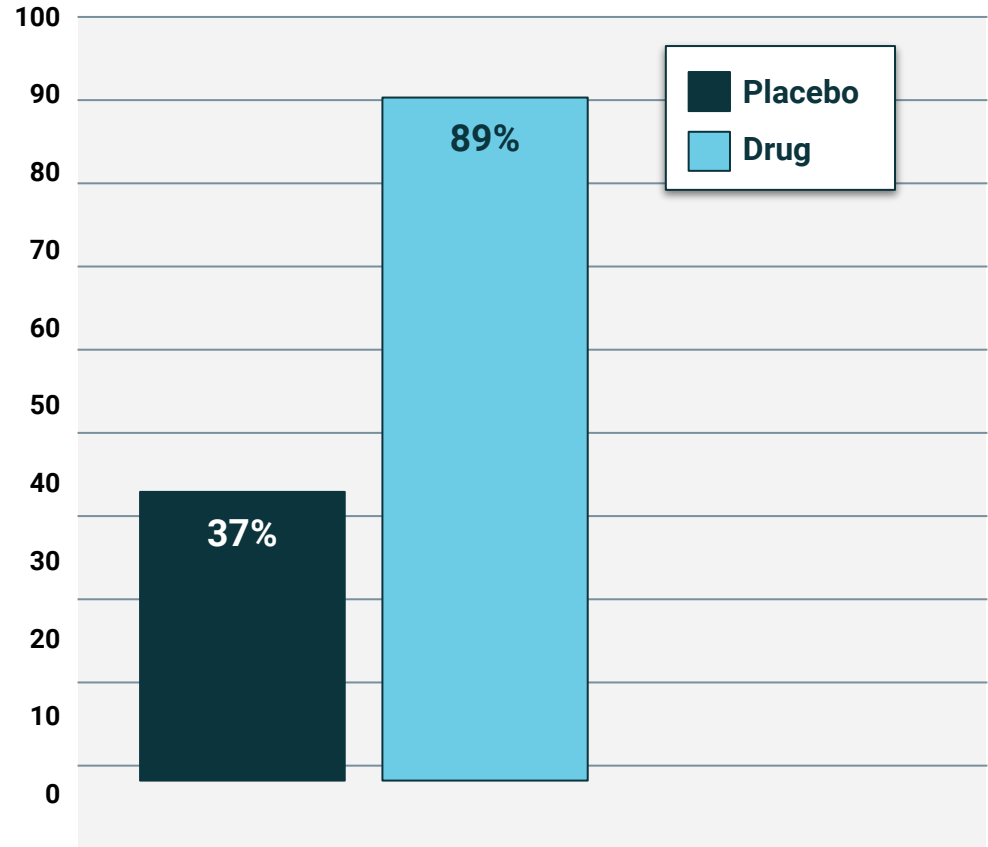
A **hypothesis** is a proposed solution or generalization to a scientific question that may be asked about our data.



**Does this new medication  
help patients recover faster  
compared to a placebo group?**

# Hypothesis

**Hypothesis:** If we give patients a dose of our new drug, they should recover at least 50% faster than those patients who do not receive the drug.





**Is the cost of living higher in  
my city compared to another?**



# Hypothesis

**Hypothesis:** If we collect different metrics around cost of living and total them together, on average the total cost of living in my city should be significantly higher than in our neighboring city.

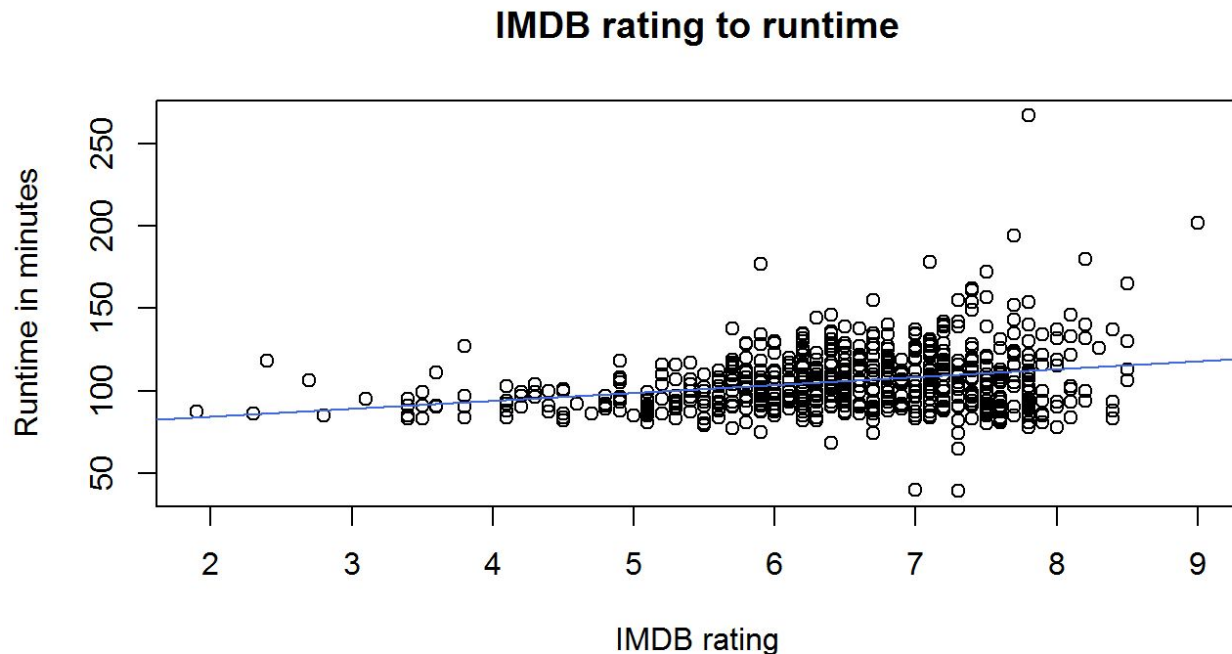
Cost of Living Indexes	New York, NY	Philadelphia, PA	Difference
Overall Index: Homeowner, No Child Care, Taxes Not Considered	187.2	101.2	45.9% less
Food & Groceries	116.6	102.5	12.1% less
Housing (Homeowner)	294.3	66.3	77.5% less
Median Home Cost	\$680,500	\$153,400	77.5% less



**Can we predict an IMDB score  
based on the length of the movie?**

# Hypothesis

**Hypothesis:** There is a linear relationship between the length of the movie and the IMDB score; therefore, the slope of the line should be nonzero.



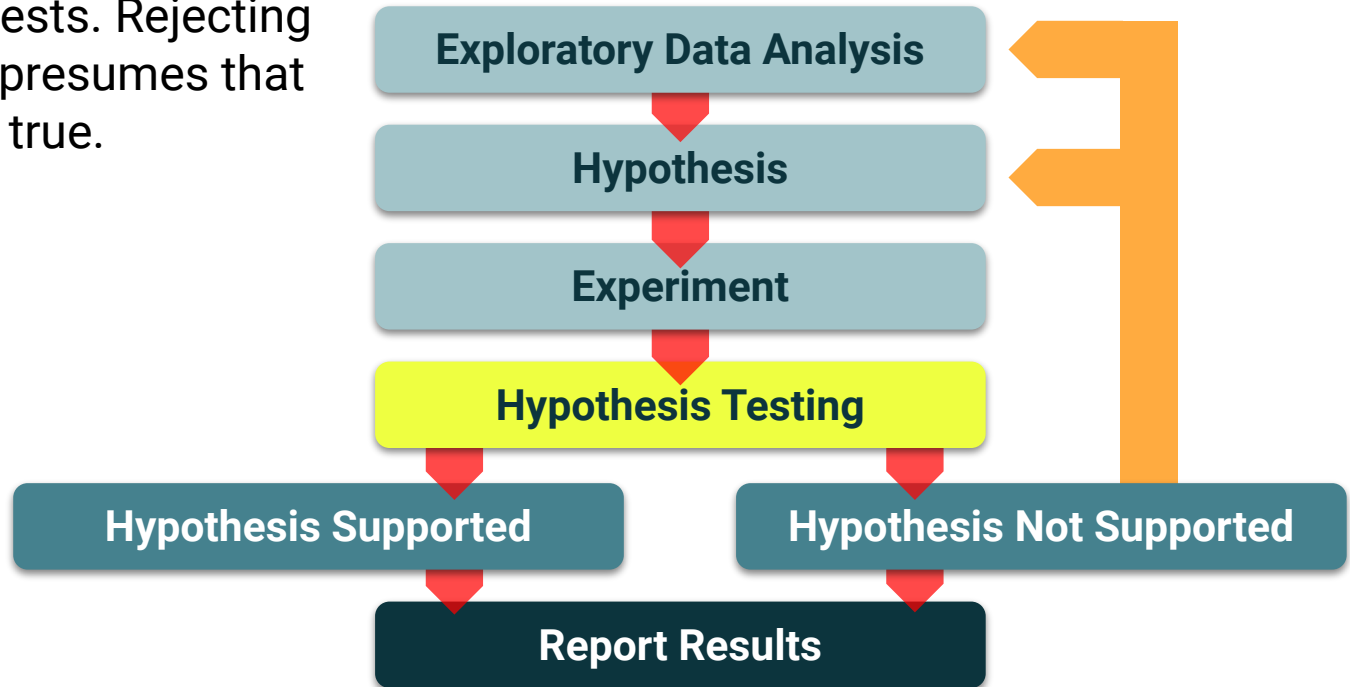


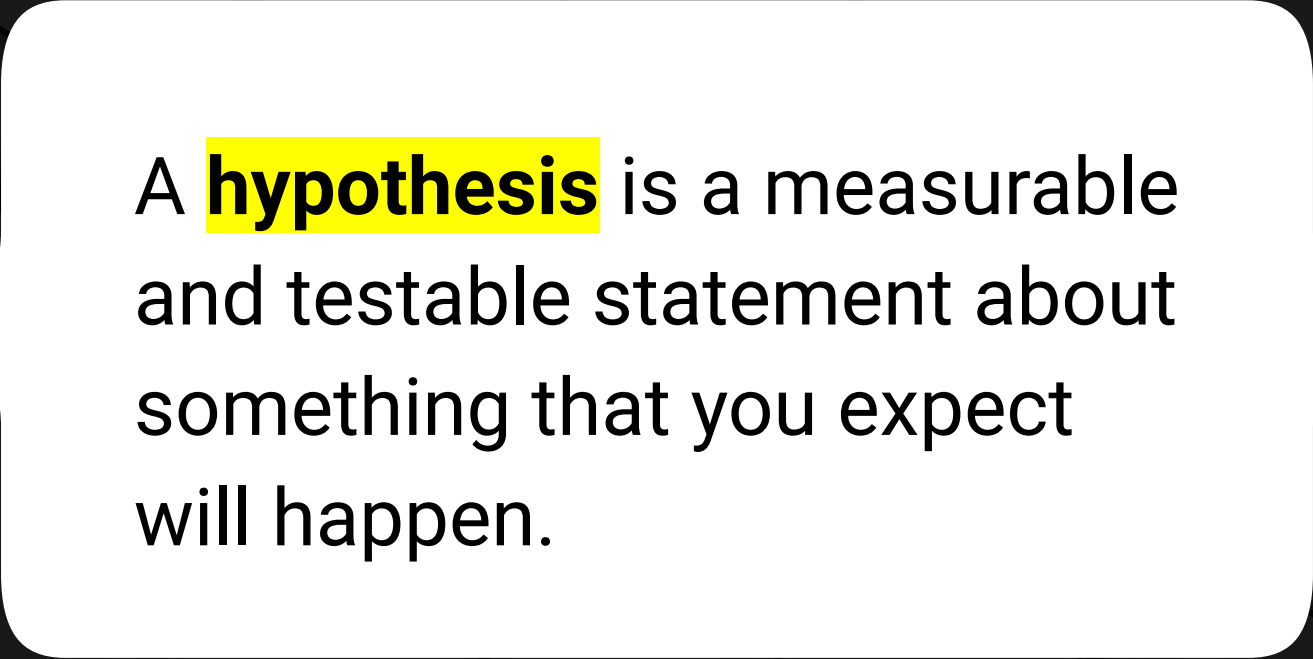
**These hypotheses should be as specific as possible to address a proposed question—the more specific the hypothesis, the easier it is to analyze.**

**However, if a hypothesis is too specific, it may not yield meaningful results or generalize well.**

# Hypothesis

The goal of hypothesis testing is to reject the null hypothesis through statistical tests. Rejecting the null hypothesis presumes that the hypothesis was true.





A **hypothesis** is a measurable and testable statement about something that you expect will happen.

# Hypothesis

---

## Null Hypothesis

The null hypothesis typically states that **NO** differences exist between the variables or groups of interest.

*"If San Diego, CA, is not warmer than Austin, TX, in July, then there will be no difference in the average daily temperatures."*

## Alternative Hypothesis

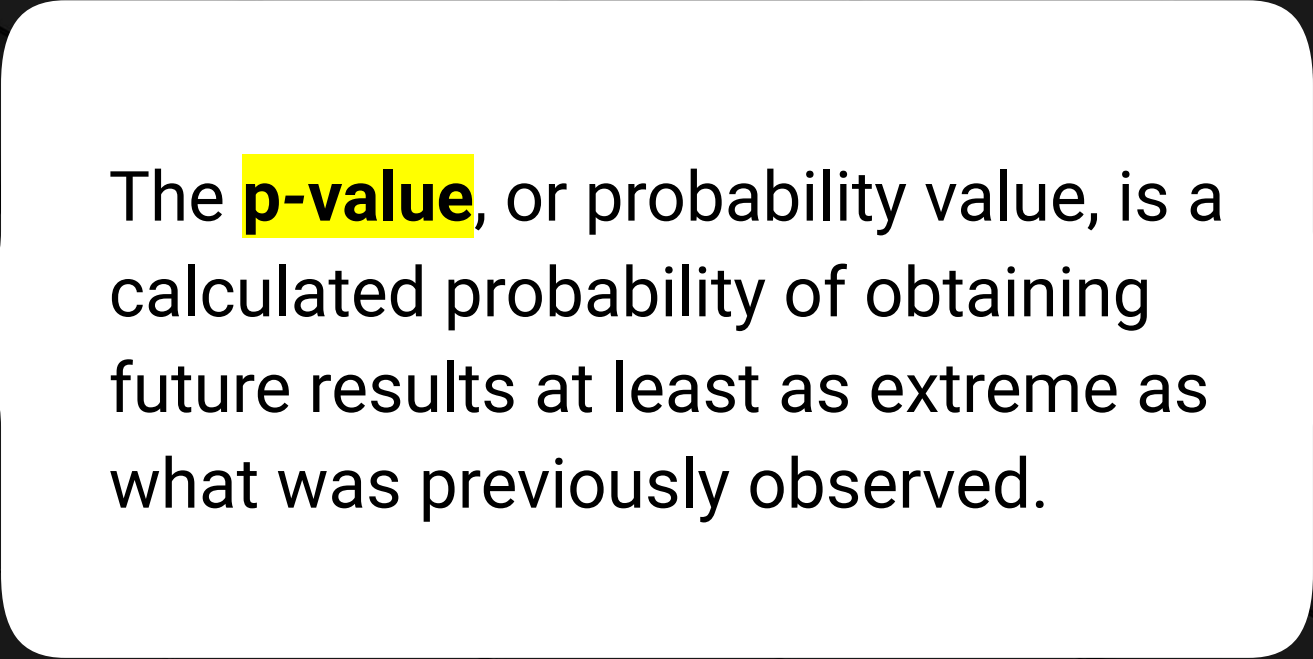
The hypothesis is often expressed as an If/Then statement.

*"If San Diego, CA, is warmer than Austin, TX, in July, then the average daily temperature will be higher."*



**Rejecting the null hypothesis is  
never absolute**

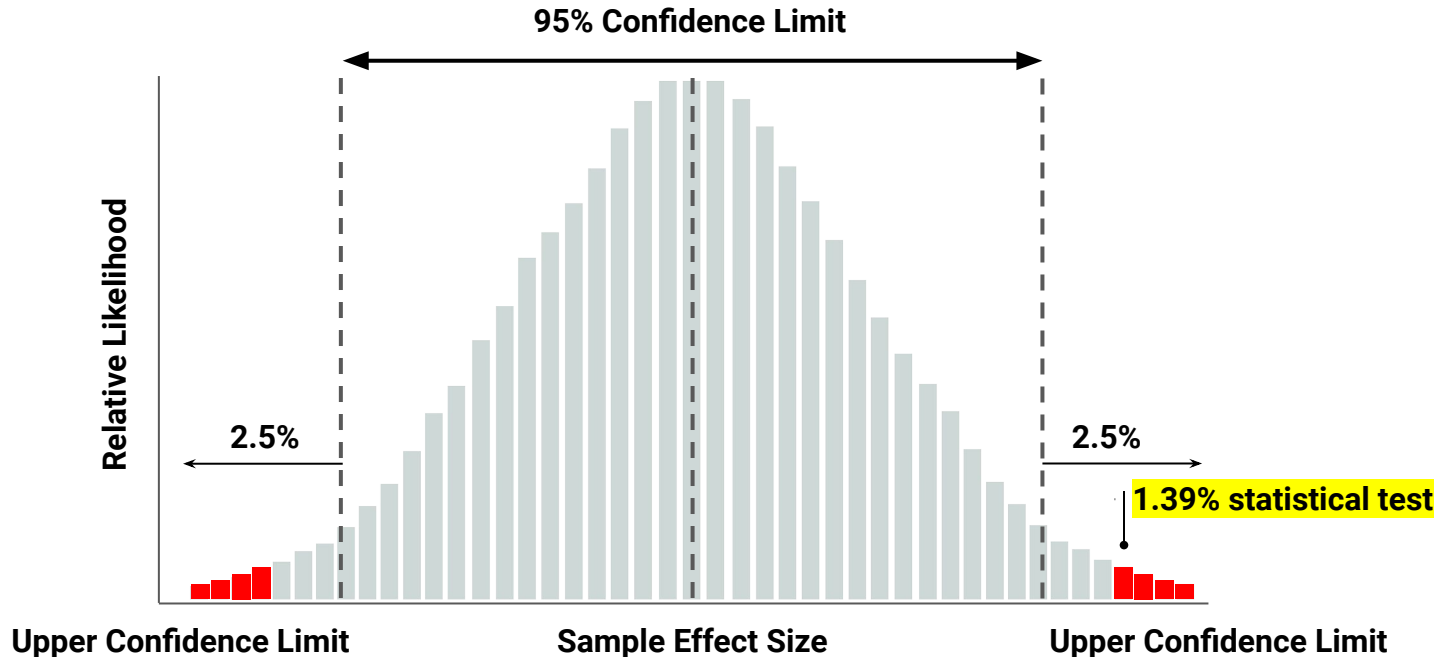




The **p-value**, or probability value, is a calculated probability of obtaining future results at least as extreme as what was previously observed.

# p-value

The p-value is compared to a fixed significance level to determine if the null hypothesis can be rejected. A smaller p-value indicates stronger evidence against the null hypothesis.



# Hypothesis Testing

---

Steps for hypothesis testing:

01

Determine the hypothesis and null hypothesis.

02

Identify the appropriate statistical test.

03

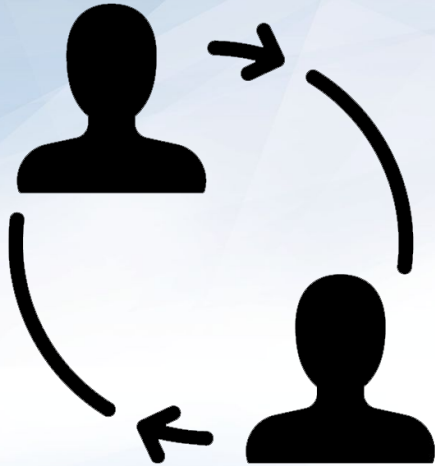
Determine the acceptable significance value.

04

Compute the p-value.

05

Determine if the p-value rejects the null hypothesis by comparing it to the significance value, which is typically  $< 0.05$ .



## **Partner Activity:** Forming Null and Alternative Hypotheses

In this activity, you will work with a partner to create hypothesis statements to some example questions we could ask in a scientific study or in an analysis industry job.

**Suggested Time:**  
15 minutes





**Let's Review**

# Does Dark Chocolate Affect Arterial Function in Healthy Individuals?

---

**Hypothesis:** If dark chocolate is related to arterial function in healthy individuals, then consuming 30g of dark chocolate daily for 1 year will result in improved arterial function.

## Null Hypothesis

If dark chocolate is not related to arterial function in healthy individuals, then consuming 30 g of dark chocolate daily for 1 year will result in no improvement in arterial function.

## Alternative Hypothesis

If dark chocolate is related to arterial function in healthy individuals, then consuming 30 g of dark chocolate daily for 1 year will result in improvement in arterial function.

# Does Coffee Have Anti-aging Properties?

---

**Hypothesis:** If coffee consumption is related to anti-aging properties, then consuming 400 mg of coffee daily will reduce mortality from age-related disease such as heart disease.

## Null Hypothesis

If coffee consumption is not related to anti-aging properties, then consuming 400 mg of coffee daily will not result in a reduction in age-related disease such as heart disease.

## Alternative Hypothesis

If coffee consumption is related to anti-aging properties, then consuming 400 mg of coffee daily will result in a reduction in age-related disease such as heart disease.

# Is Biodiesel Better for the Environment Than Fossil Fuel?

---

**Hypothesis:** If we assume that carbon emission is the largest contributing factor to environmental damage from burning fuel, biodiesel should produce significantly less carbon than fossil fuel.

## Null Hypothesis

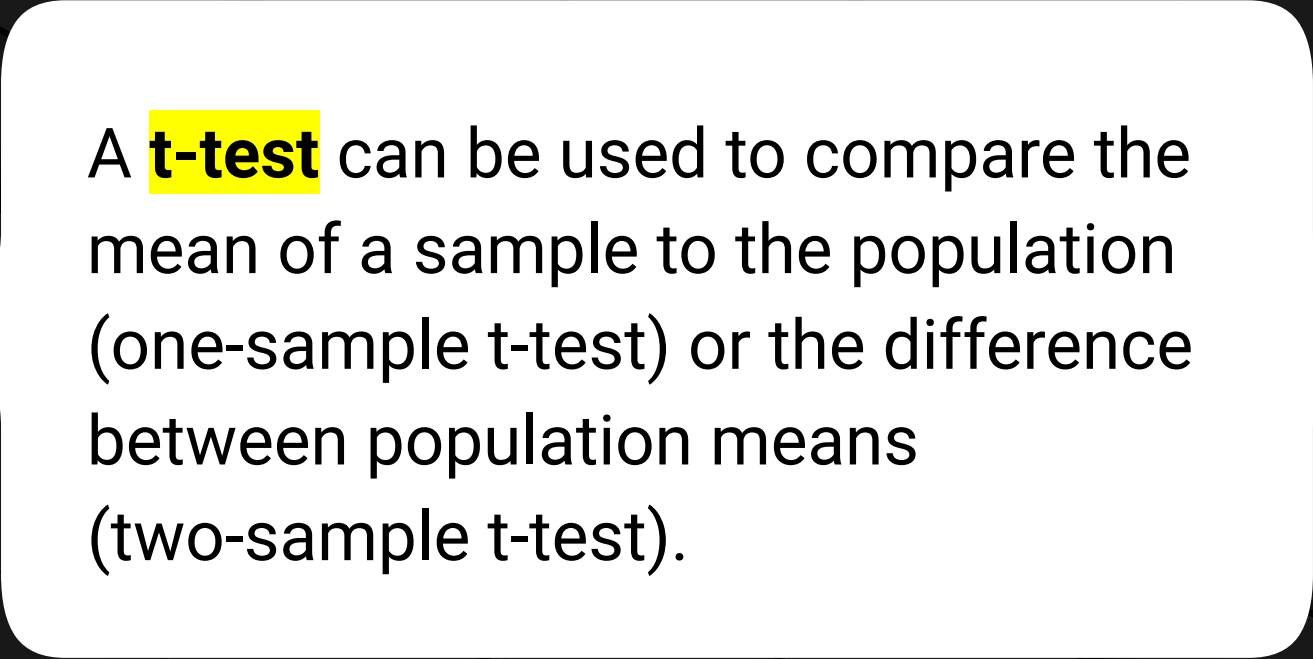
If we burn one gallon of both fuels, fossil fuel will produce the same or less carbon than biodiesel.

## Alternative Hypothesis

If we burn one gallon of both fuels, fossil fuel will produce more carbon than biodiesel.



# T-Tests



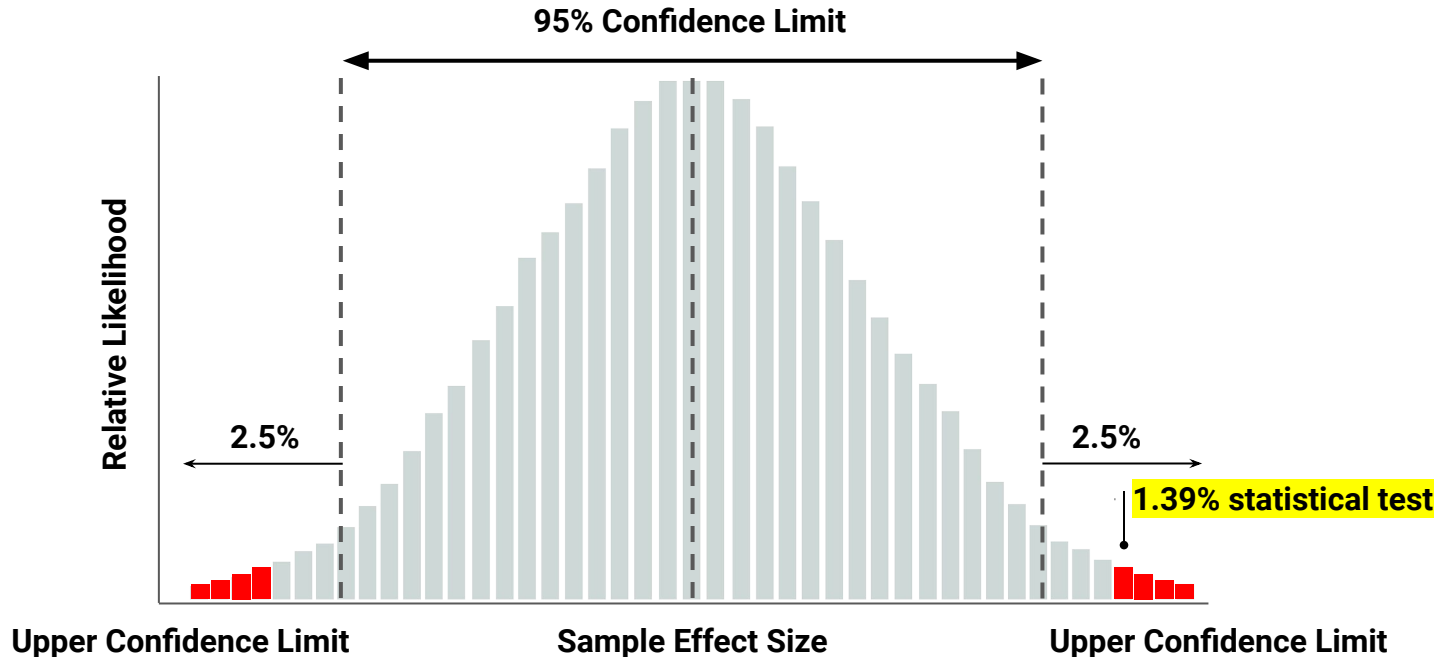
A **t-test** can be used to compare the mean of a sample to the population (one-sample t-test) or the difference between population means (two-sample t-test).



The null hypothesis assumes that there is no meaningful difference between the two means. Therefore, the goal of the t-test is to reject the null hypothesis.

# p-value

The p-value is the probability of seeing a meaningful difference between the two means by random chance, assuming that the null hypothesis cannot be rejected.



# p-value

---

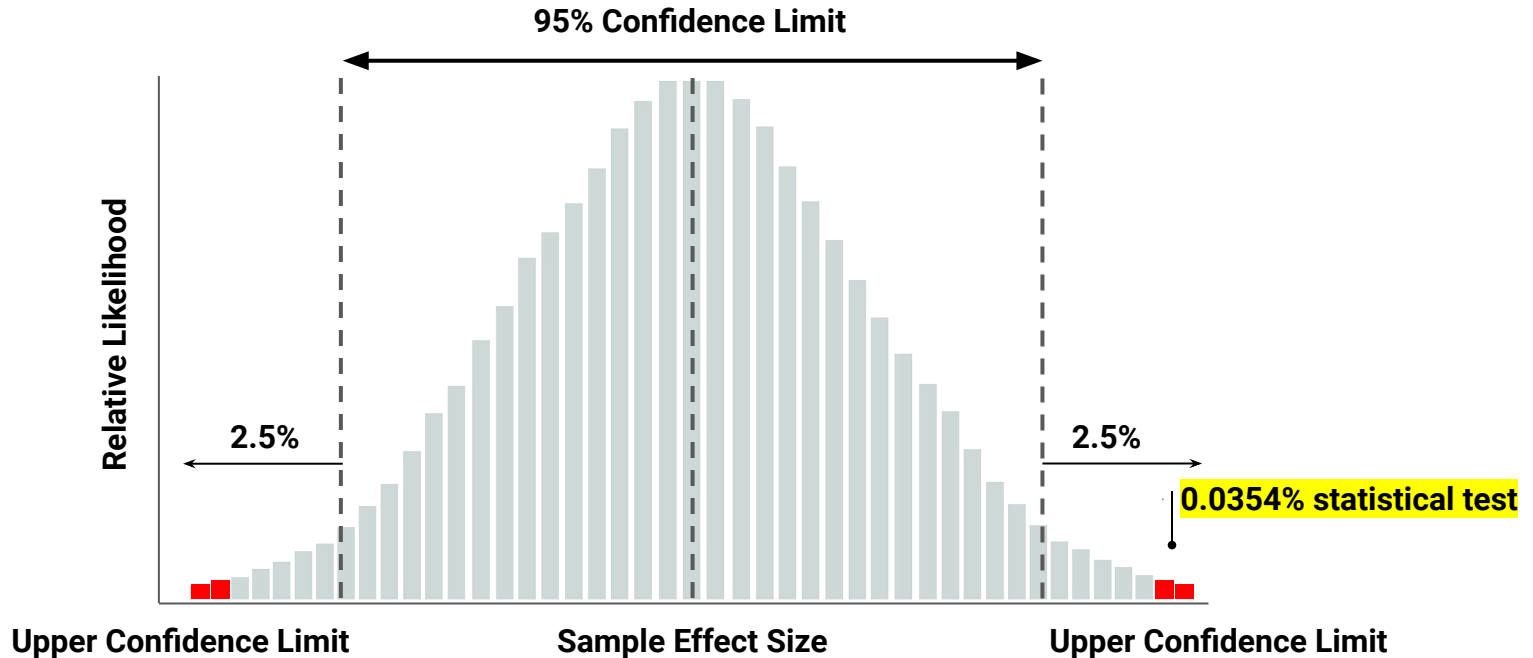
## For example:

Say that an acne medication, ActinoX, is undergoing a trial to see whether it eliminates acne more quickly than an older medication, Renovagene.



# p-value

The trial data indicates that ActinoX is effective, with a p-value of 0.0354.



# p-value

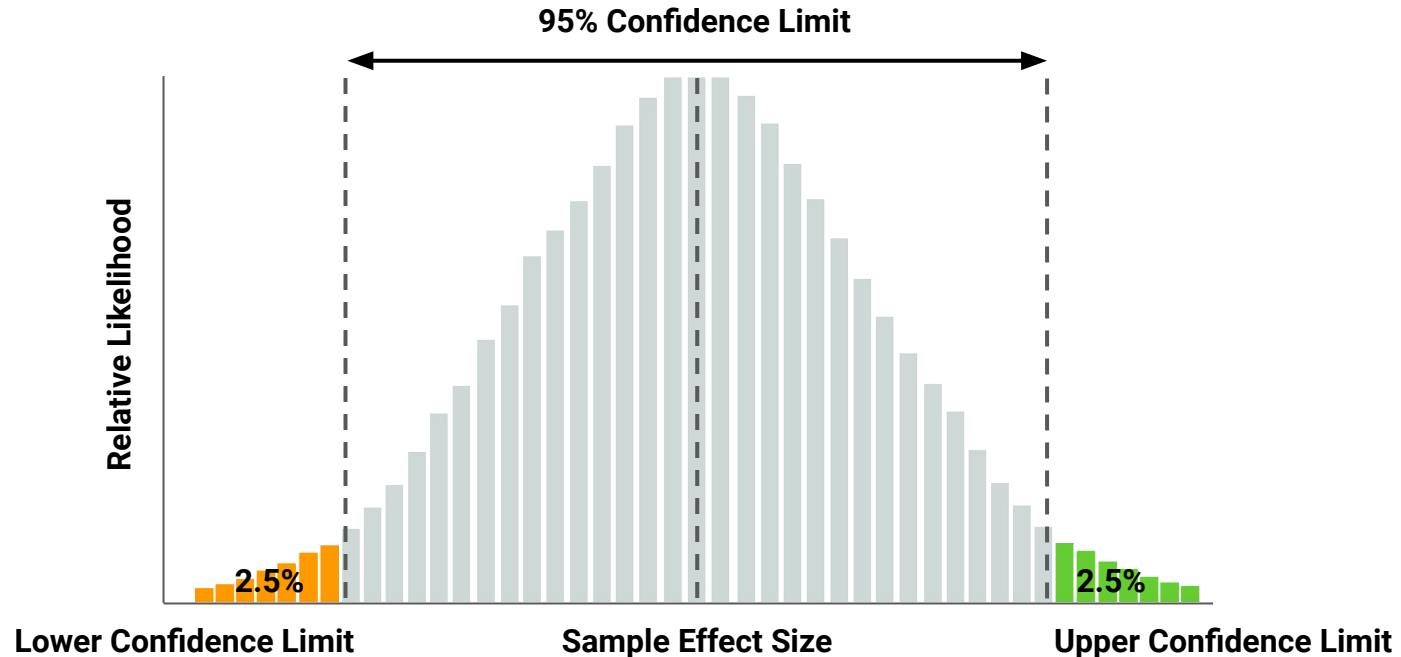
---

This means that there is a 96.46% chance that using ActinoX leads to a meaningful difference in the outcome compared to Renovagene; it also means there is a 3.54% chance that the improvements observed in the trial can be attributed to random chance.



# p-value

If the level of significance was set at 0.05, the p-value of 0.0354 is low enough to reject the null hypothesis.







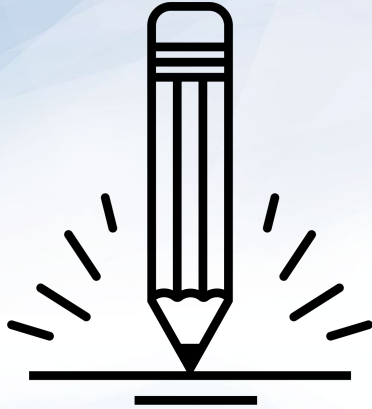
The 0.05 cutoff for the significance isn't written in stone; it is arbitrarily chosen. Another value could have been chosen beforehand to determine the level of significance required to reject the null hypothesis.



# Instructor Demonstration

---

## T-Tests



## **Activity: T-Test**

In this activity, you will determine if there is a statistically significant difference in the number of vertebrae of adult sardines in Alaska vs. San Diego.

**Suggested Time:**  
15 minutes





**Let's Review**

# ANOVA

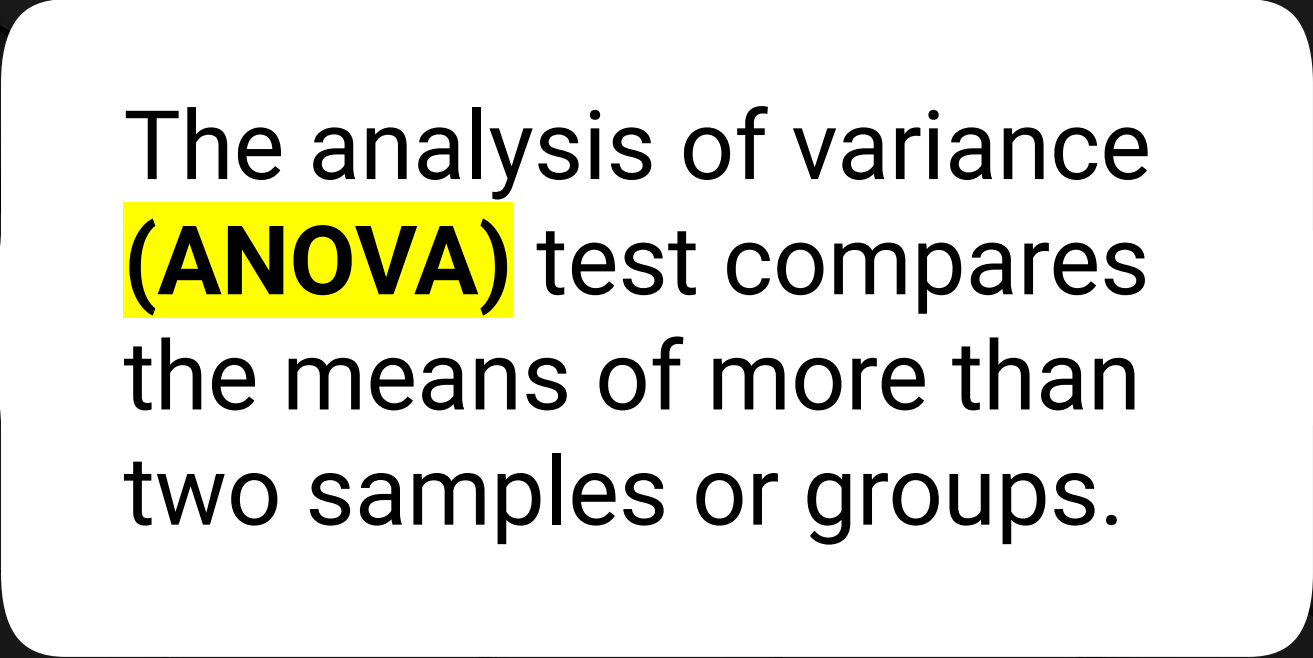


Now that we have learned how to compare the means of one or two normal samples, it's time to start thinking bigger!



**What happens if our data is separated into more than two samples or groups?**

**What if there were 3, 5, or even 10 samples to compare?**



The analysis of variance  
**(ANOVA)** test compares  
the means of more than  
two samples or groups.



# Applying ANOVA Models

We wouldn't want to make pairwise comparisons for each set of samples just to find out that all samples are statistically similar.

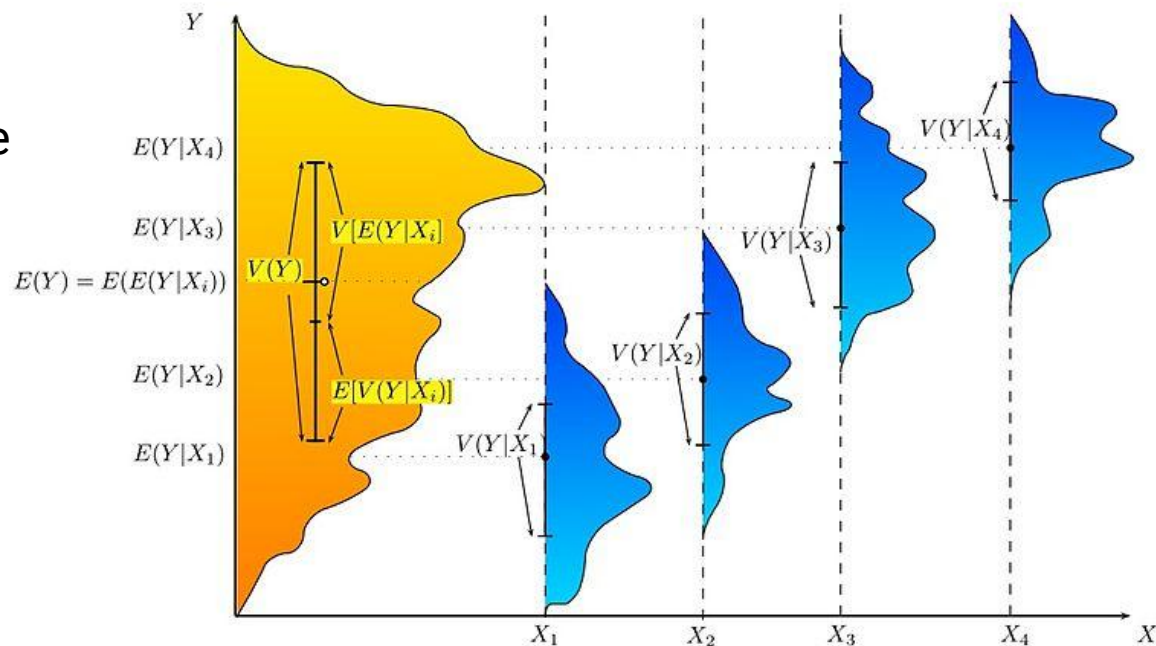


Figure 1: ANOVA : Fair fit

# Applying ANOVA Models

---

If any of these assumptions are violated, we need to use a nonparametric test:



ANOVA compares the means across all samples and determines whether there is a significant difference in at least one sample.



An ANOVA test is based on a null hypothesis that there are no differences between any groups.



If the p-value is smaller than the significance level, then at least one sample mean is statistically different from the other means.



ANOVA tests assume that the samples/groups are independent, that the samples all come from normally distributed population data, and that the standard deviation of the population data is equal for all groups.



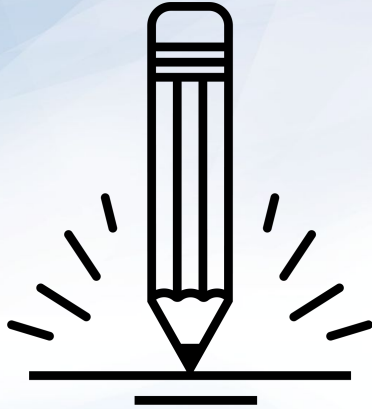
# Instructor Demonstration

---

## ANOVA

# Questions?





## Activity: ANOVA

In this activity, you will use ANOVA to compare the differences in pain threshold for people with different hair colors.

**Suggested Time:**  
15 minutes





**Let's Review**

# Fits and Regressions

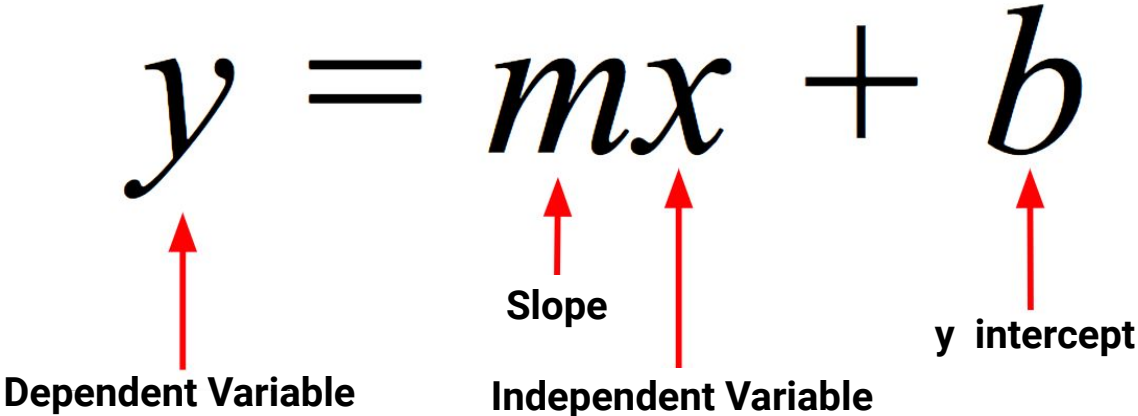
# Equation of a Line

---

The equation of a line defines the relationship between x values and y values.

**Equation of a Line**

$$y = mx + b$$

  
The diagram shows the equation  $y = mx + b$  with three red arrows pointing upwards to its components. The arrow under  $y$  points to the label "Dependent Variable". The arrow under  $m$  points to the label "Slope". The arrow under  $x$  points to the label "Independent Variable". The arrow under  $b$  points to the label "y intercept".

Dependent Variable

Slope

Independent Variable

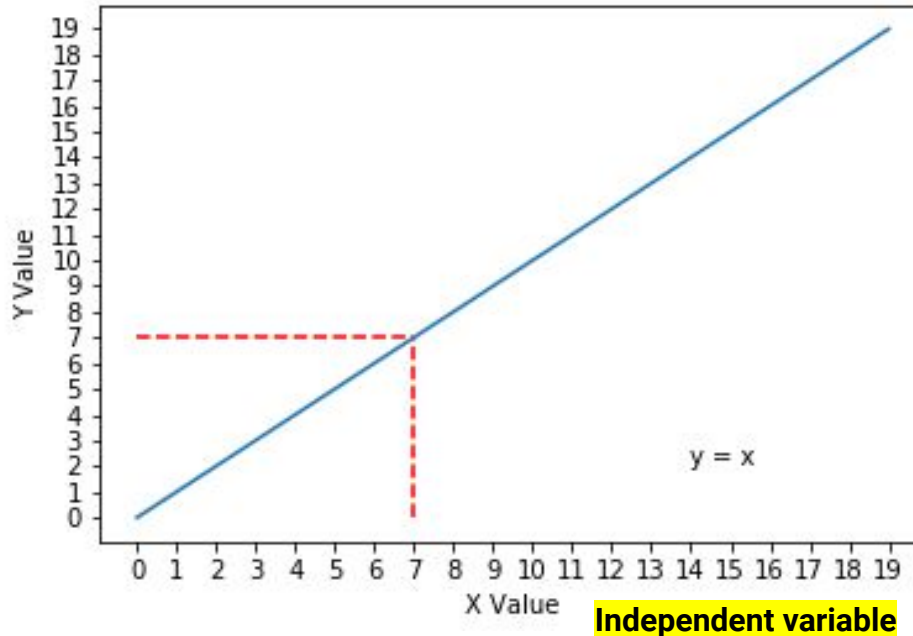
y intercept



# Equation of a Line

When it comes to variables in the equation, we refer to the  $x$  in the equation as the **independent variable** and the  $y$  as the **dependent variable**.

**Dependent variable**

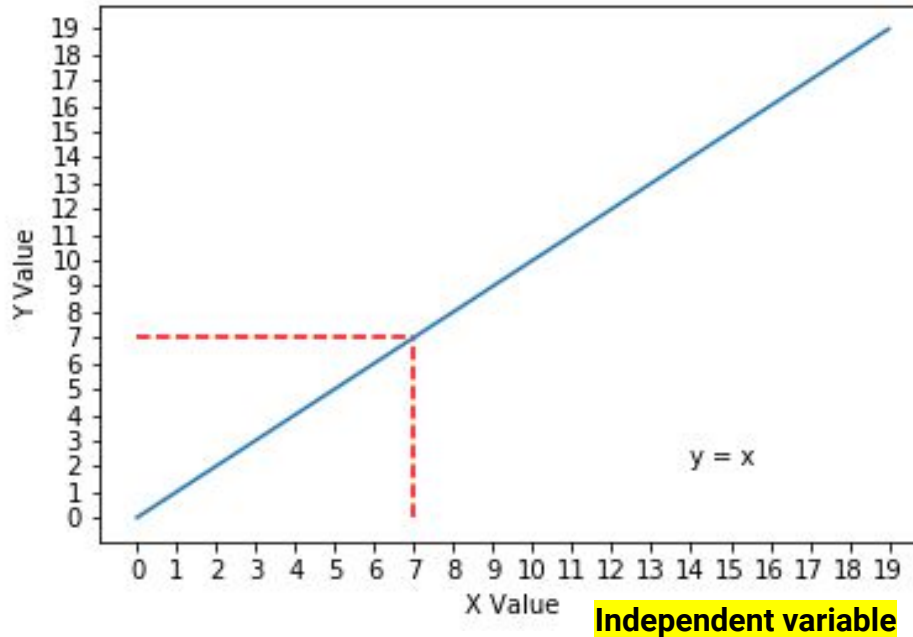


**Independent variable**

# Equation of a Line

The first plot is considered the ideal linear relationship of  $y$  and  $x$ , where the  $x$  and  $y$  values are the same value.

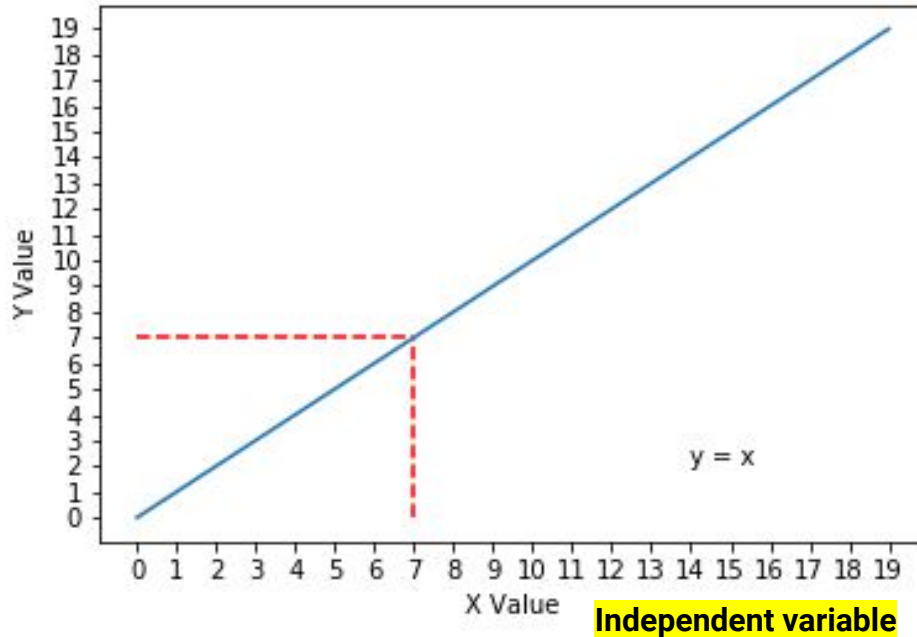
Dependent variable



# Equation of a Line

In this plot, the equation for line is  $y = x$  because the slope is equal to 1, and the  $y$  intercept is equal to 0.

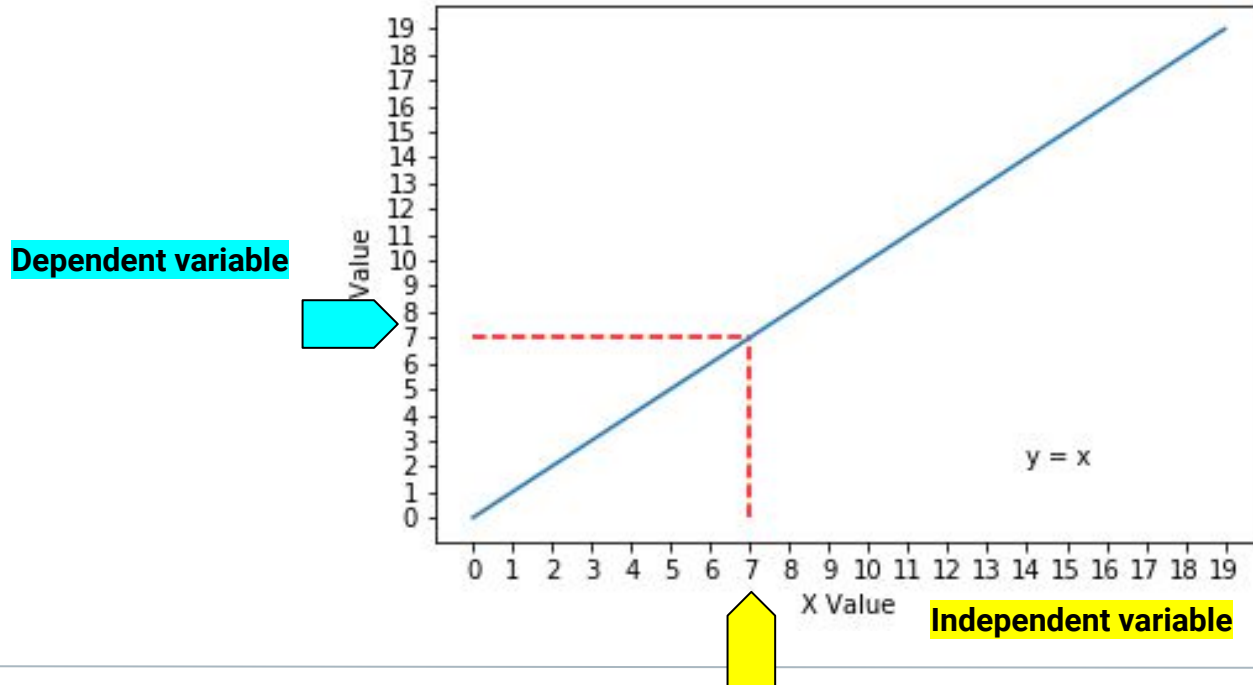
**Dependent variable**



**Independent variable**

# Equation of a Line

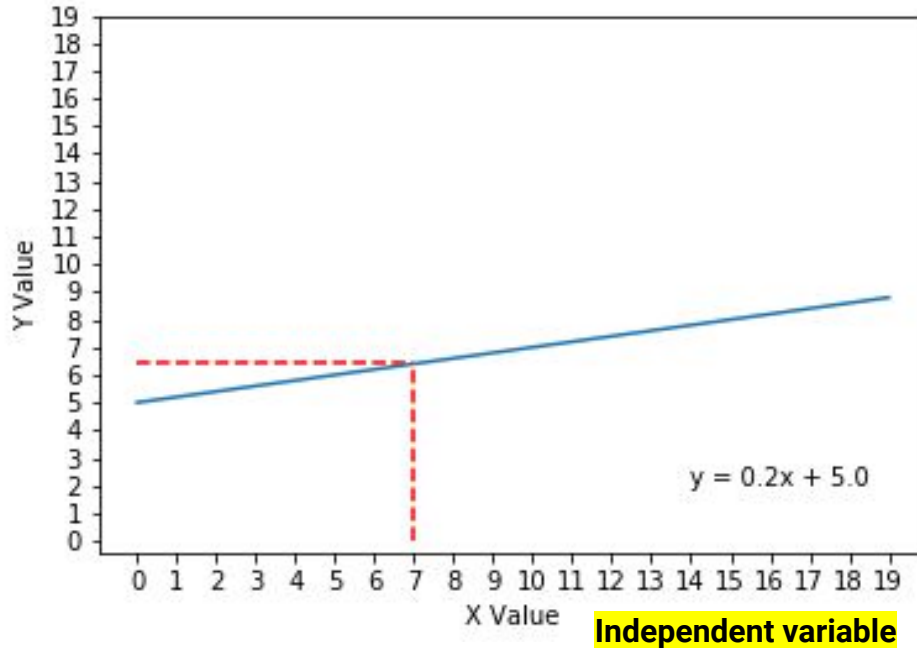
If we look at the  $x$  value of 7 (denoted by the vertical dashed line), the corresponding  $y$  value is also 7 (denoted by the horizontal dashed line).



# Equation of a Line

In this linear relationship between  $x$  and  $y$ , the slope is much smaller, but the  $y$  intercept is much larger.

**Dependent variable**



**Independent variable**



# Time to Code



## Fits and Regression

Suggested Time:

15 minutes

# Questions?

