Hypothesis Testing and Statistical Tests

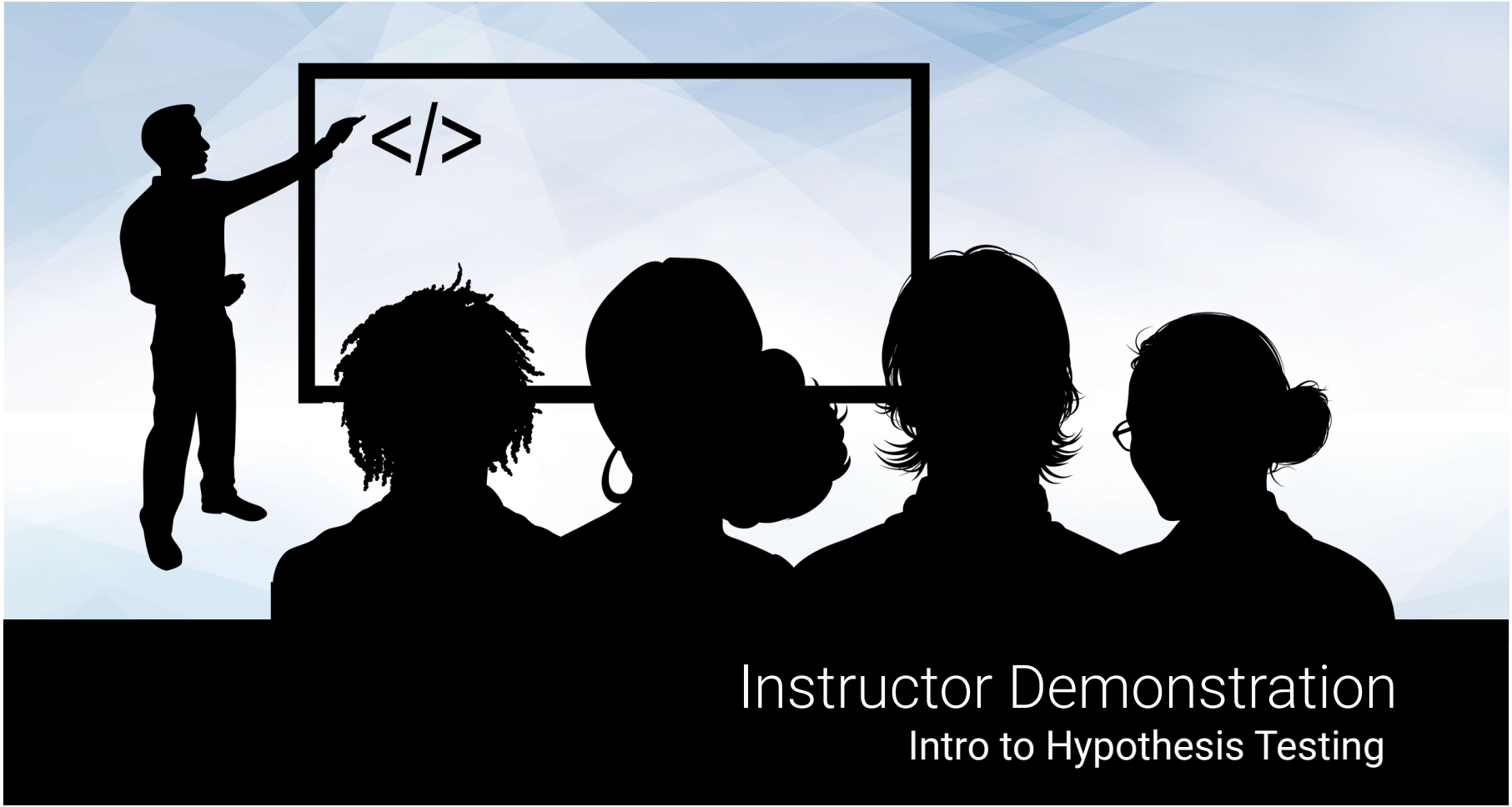
Data Boot Camp

Lesson 7.3

Class Objectives

By the end of today's class you will be able to:

- Understand the difference between null hypothesis and alternate hypothesis.
- Apply one-sample t-test to identify significant difference between sample and population data.
- Apply two sample t-test to identify significant differences between two groups.
- Apply ANOVA to compare the means of three or more groups.
- Perform Chi Square Test to compare distribution of categorical data.



Instructor Demonstration

Intro to Hypothesis Testing

Hypothesis testing can be confusing at times, mostly because you **must create** your **null** and **alternative hypothesis** before performing any analysis.

Intro to Hypothesis Testing

- Hypothesis is an **educated guess** about something.
- **Hypothesis Statement:**
  - The Hypothesis is often expressed as an **If/Then** statement.
- **Hypothesis Testing:**
  - Hypothesis testing in statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results.
  - We test against for two mutually exclusive outcomes - **null** and **alternative** hypothesis

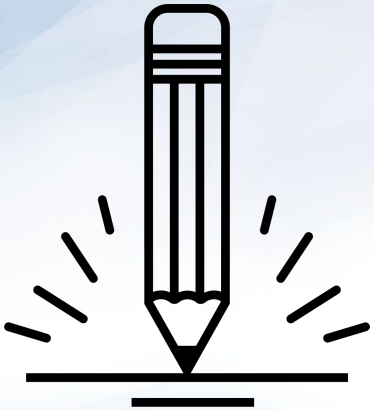
Intro to Hypothesis Testing

- **Null Hypothesis (H_0):**
  - Null hypothesis is the hypothesis that we are trying to disprove due to no statistical significance between the two variables.
  - In short, you null hypothesis assumes that your results **happened by chance**.
- **Alternative Hypothesis (H_a):**
  - Alternative hypothesis is the opposite of the null hypothesis, it assumes there is some factor influencing the results. The hypothesis assumes your results **did not happen by chance.**

Steps for Hypothesis Testing:

Intro to Hypothesis Testing

1. Determine the Hypothesis and Null Hypothesis.
2. Identify the appropriate statistical test.
3. Determine the acceptable significance value.
4. Compute the P-value.
5. Determine if the P-value rejects the Null Hypothesis by comparing it to the significance value (Typically < 0.05)



Activity: Forming a Null Hypothesis

In this activity, you and your partner will take two given questions into an Hypothesis and Null Hypothesis.

Suggested Time:
10 Minutes

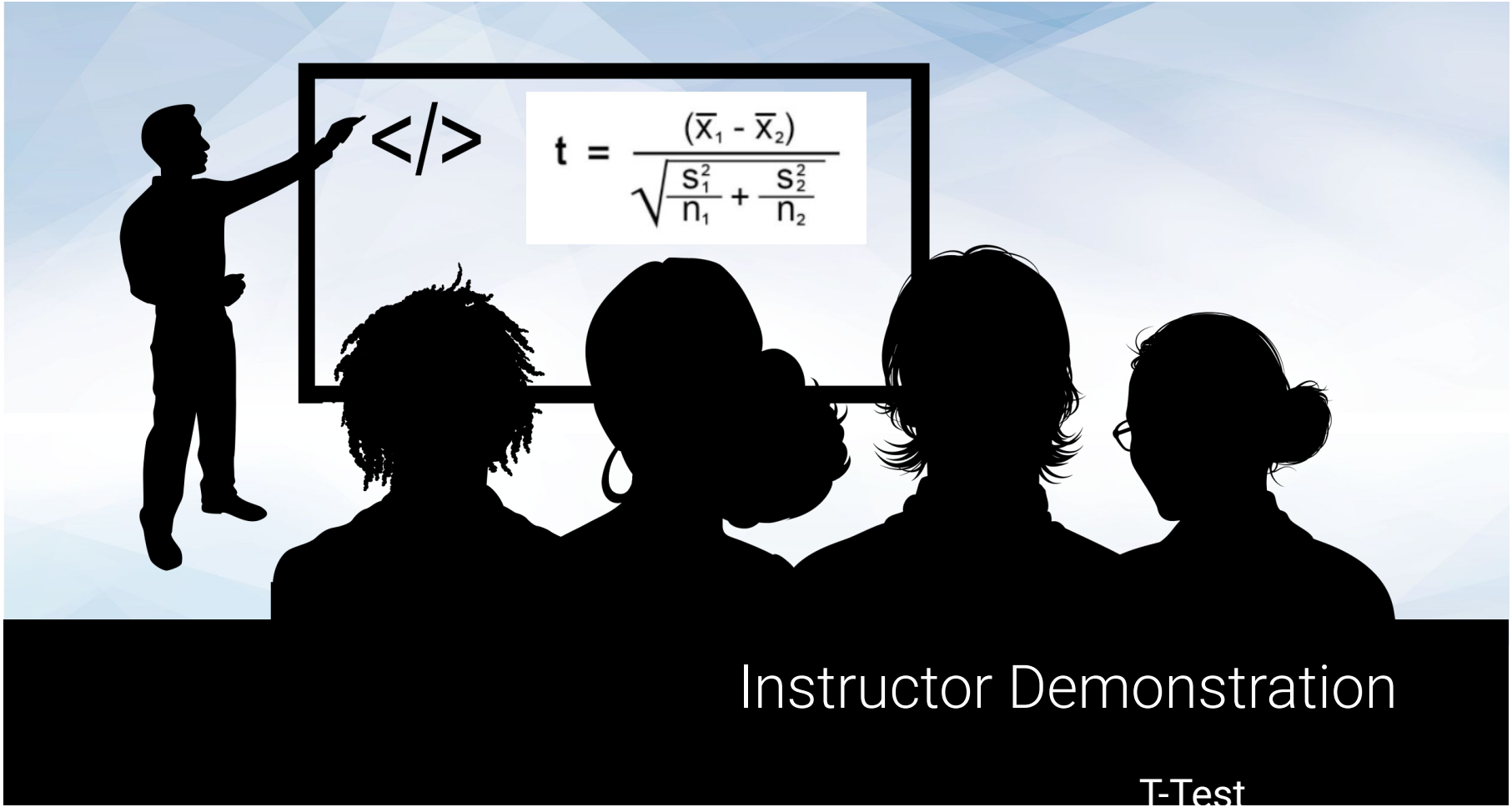
Instructions:

Activity: Forming a Null Hypothesis

- Convert the following Questions into an Hypothesis and Null Hypothesis.

    - Does Dark Chocolate affect arterial function in healthy individuals?
    - Does Coffee have anti-aging properties?

Time’s Up! Let’s Review.



Instructor Demonstration

T-Test

t-test

Calculating t-test

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- $\bar{X}_1$ = mean of the first data set.
- $\bar{X}_2$ = mean of the second data set.
- $S1^2$ = standard deviation of the first data set.
- $S2^2$ = standard deviation of the second data set.
- N_1 = number of elements in the first data set.
- N_2 = number of elements in the second data set.

t-test

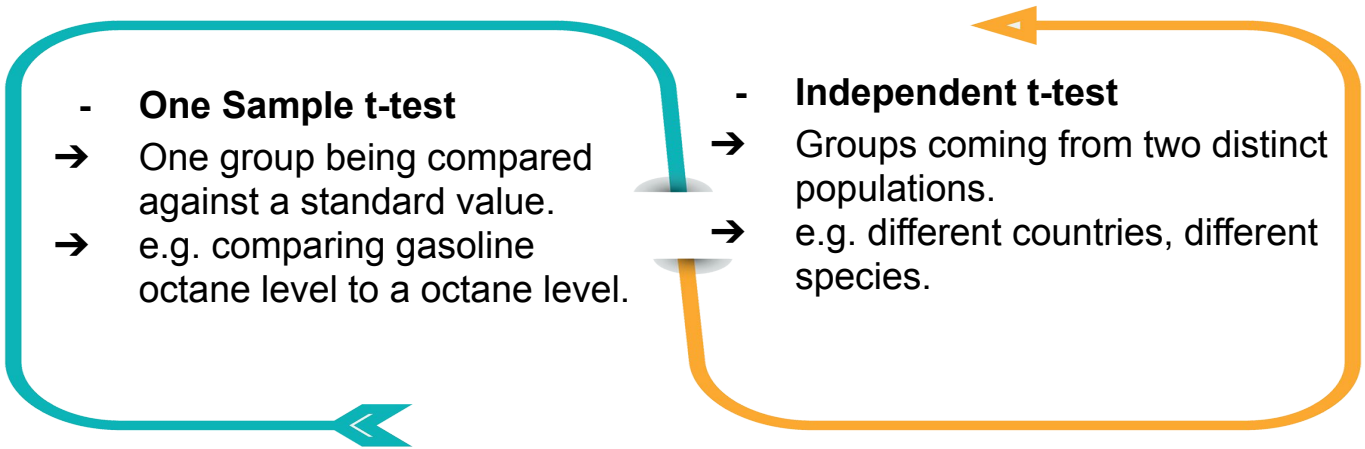
- A t-test tells you how significant the differences between groups are.
- It lets you know if the differences, measured in means/average, could have happened by chance.

- **One Sample t-test:**
  - Determines whether the sample mean statistically differ from a known or hypothesized population mean.

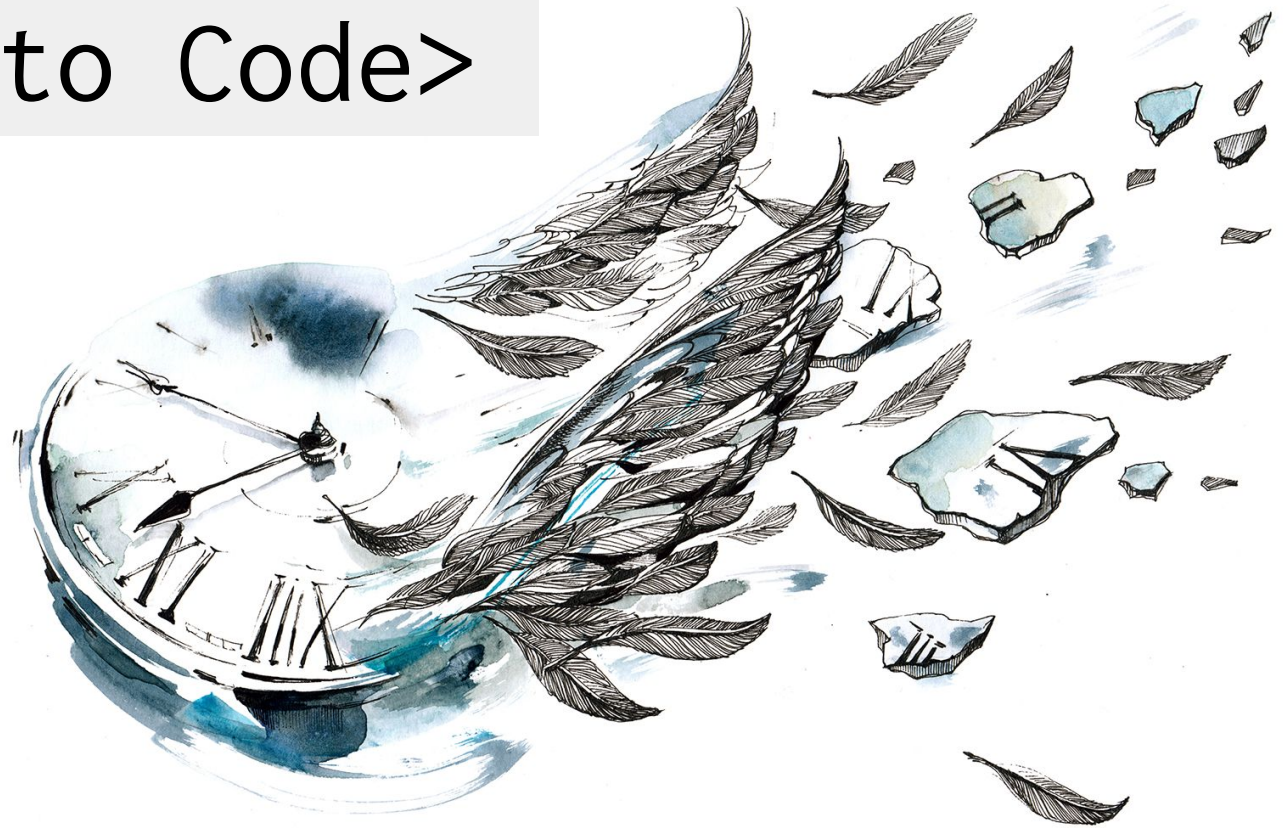
- **Independent t-test:**
  - Also known as **two sample t-test** , determines whether there is a statistically significant difference between the means in **two unrelated** groups.

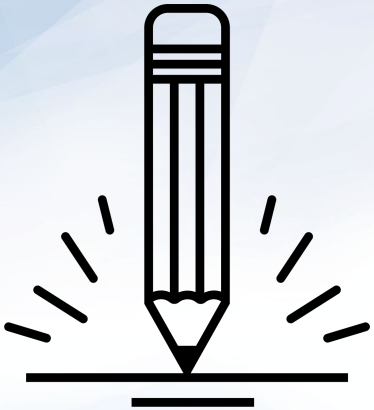
What type of t-test Should I use?

- There are couple of things to consider before performing t-test:
  - Whether the compared groups comes from a single population or two distinct population.
  - Whether you want to est the difference in a specific direction.



<Time to Code>





Activity: t-test

In this activity, you will use a t-test to compare the difference in Adult Sardine Vertebrae counts from two different locations.

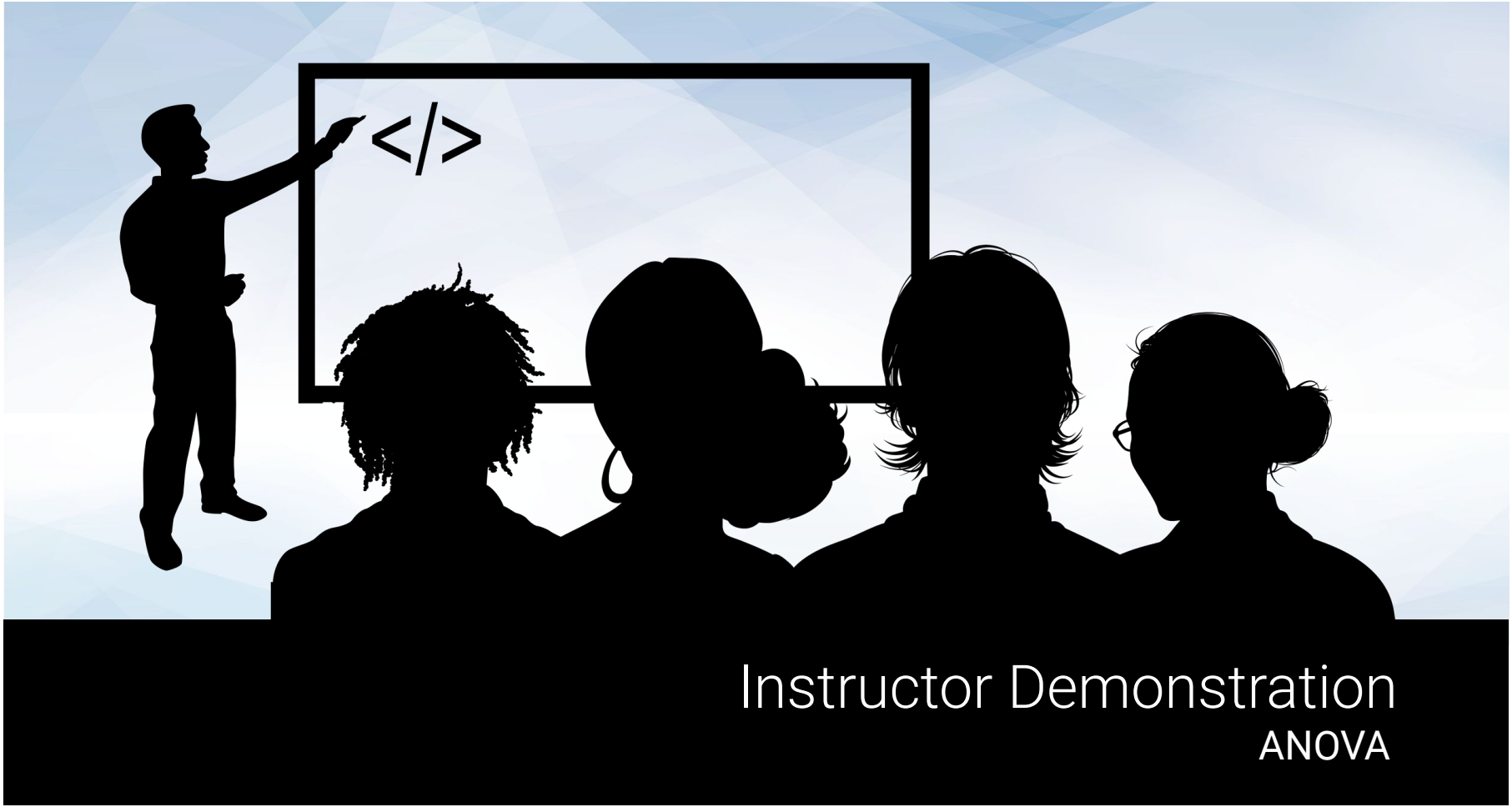
Suggested Time:
15 Minutes

Instructions:

Activity: t-test

- Calculate the mean for each population.
- Use a t-Test to determine if there is a statistically significant difference in the number of vertebrae of Adult Sardines in Alaska vs. San Diego.
- It is up to you to determine if you should use a One Sample independent T-Test.

Time’s Up! Let’s Review.

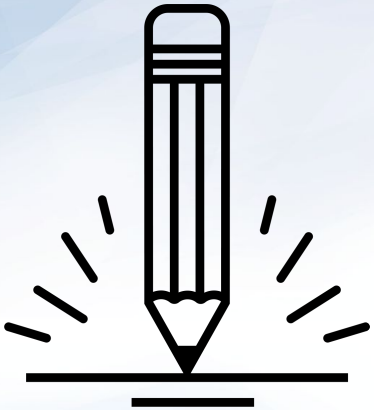


Instructor Demonstration

ANOVA

ANOVA

- In short, **ANOVA** is an extension of a t-Test. As we previously covered, with t-Test you can test two **groups** to see if it is difference in means.
- An **ANOVA** test is a way to find out if survey or experiment results are significant.
- They help you to figure out whether to **reject** the **null hypothesis** or **to fail to reject it.**



Activity: ANOVA

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

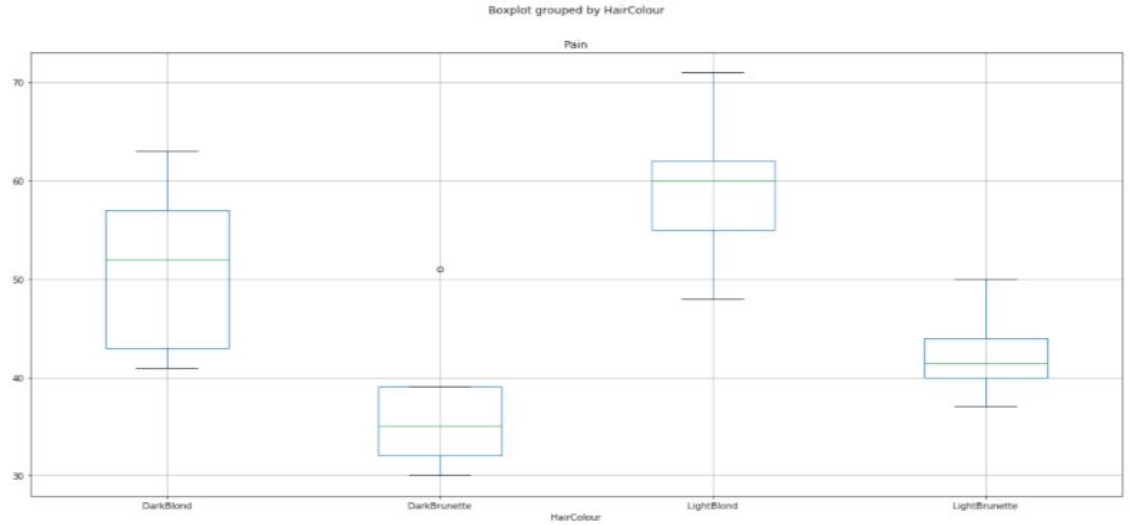
Suggested Time:
10 Minutes

Instructions:

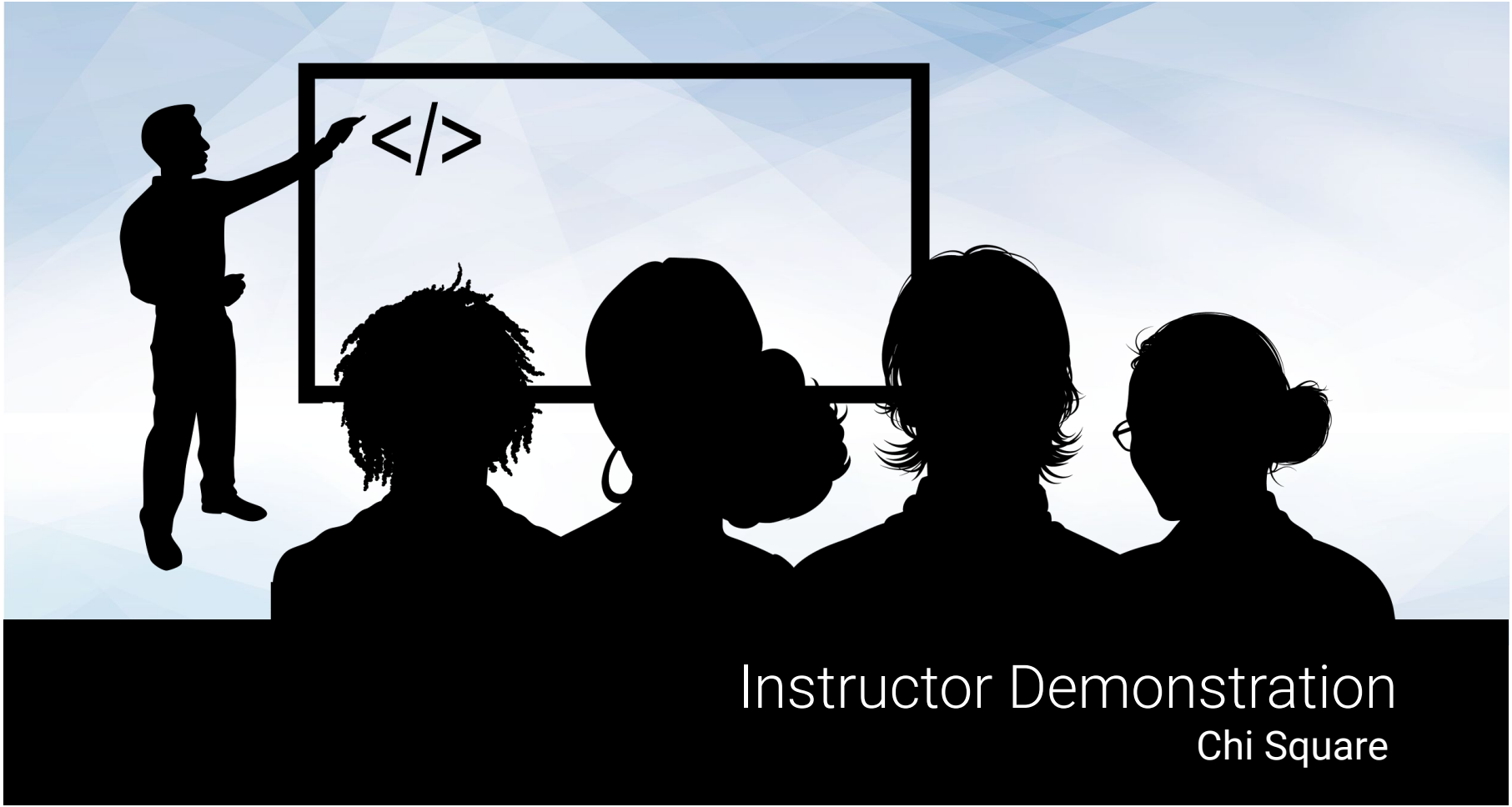
Activity: ANOVA

- Perform a one-way ANOVA test to determine if there are any significant differences in Hair Color vs. Pain Threshold.
- Create a Boxplot to show the distribution of pain tolerances for each hair color.

```
Out[4]: <matplotlib.axes._subplots.AxesSubplot at 0x112628978>
```



Time’s Up! Let’s Review.



Instructor Demonstration

Chi Square

The Chi-Square Test

What's used for?

To answer the question: Is the distribution of frequencies in the dataset meaningful?

In other words, does the data match our expectations?

In still other words, do we reject the null hypothesis or fail to reject it?

The Chi-Square Test

Example: Out of 300 dinosaurs,

220 eat everything

55 eat only meat

25 eat only plants

Null hypothesis:
No statistical significance exists in the distribution of omnivores, carnivores, and herbivores. That is, this data can be explained by random distribution.

The chi-square test can help us reject or fail to reject the null hypothesis.

The Chi-Square Python Function

How is the chi-square test used:

01 Determine chi-square value.

02 Determine degree of freedom.

03 Choose a p-value.

04 Determine critical value.

05 Make a decision.

Degree of Freedom

To determine the degree of freedom (df), take the number of rows and subtract 1:

Omnivores: 220
Carnivores: 55
Herbivores: 25

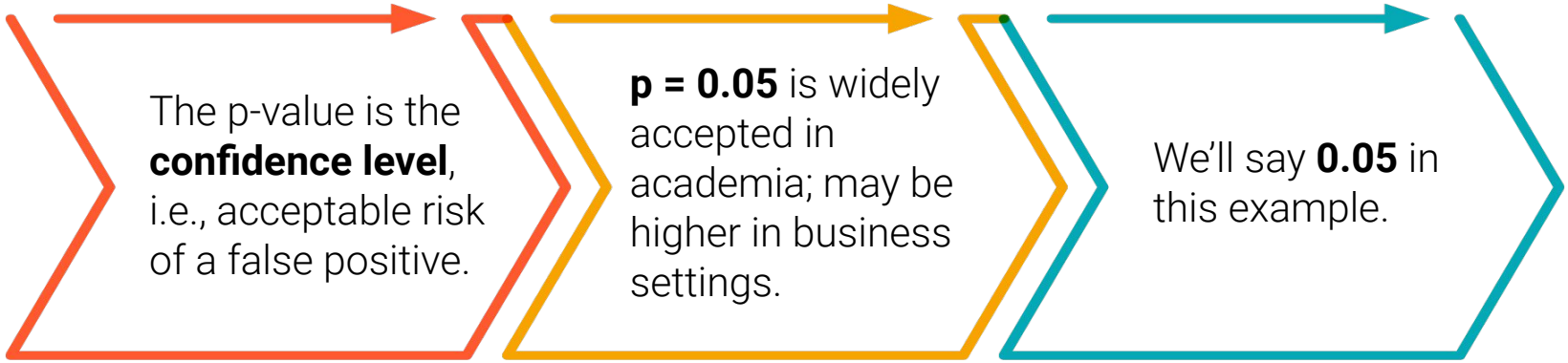
There are three rows, so the degree of freedom is

3 – 1 = 2

The degree of freedom is the number of figures required to fill out the table (like Sudoku).

If we have two of the numbers, we can figure out the value of the third.

P-value



Importance of Findings	Significance Level	Probability of Being Wrong
Low	0.1	1 in 10
Normal	**0.05**	**5 in 100**
High	0.01	1 in 100
Very high	0.001	1 in 1,000
Extreme	0.0001	1 in 10,000

The Chi-Square Test Formula

A few more considerations:

- The chi-squared test is used to test categorical variables; it can't be used on continuous data.
- The categories must be mutually exclusive.
- We have covered using the chi-square test formula to test goodness of fit.
- It can also be used to test independence. (Feel free to explore this on your own.)

Using the Chi-Squared Test In Python

01 Import the **scipy.stats** module.

```
# The statistical module used to run chi square test
import scipy.stats as stats
```

02 Determine the **critical value.**

```
# The degree of freedom is 3-1 = 2
# With a p-value of 0.05, the confidence level is 1.00-0.05 = 0.95.
critical_value = stats.chi2.ppf(q = 0.95, df = 2)
```

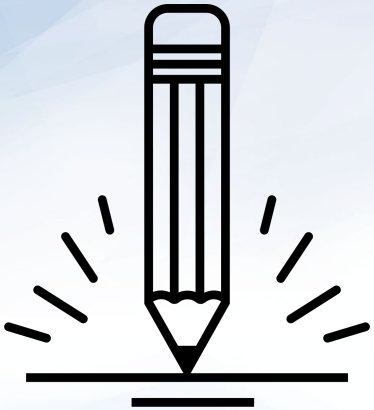
```
# The critical value
critical_value
```

```
5.99146454710798
```

03 Run the **chi-squared test.**

```
# Run the chi square test with stats.chisquare()
stats.chisquare(df['observed'], df['expected'])
```

```
Power_divergenceResult(statistic=220.5, pvalue=1.3153258948574585e-48)
```



Activity: Chi-Square

In this activity, you will perform the Chi-Square test: First in Python, then by hand.

Suggested Time:
10 Minutes

Instructions:

Activity: Chi-Square

- You are the owner of four cafés in a town of avid coffee drinkers.
- Using a Chi-square goodness-of-fit test, determine whether the results suggest that customers are more likely to frequent one cafe over another.
- Perform the necessary calculations by using Python.
- Then perform the calculations by hand to verify your findings.
- Use a p-value of 0.05.
- Consult a Chi-square table to find your critical value: https://www.medcalc.org/manual/chi-square-table.php
- On your student repository, open: Stu-Cafes.ipynb

Time’s Up! Let’s Review.

Countdown timer

40:00

(with alarm)


TRADING PLATFORM
ETH USD LTC
190.34 +3.44 +1.88%
199.31 +0.31 +0.05% 19:55 Jun 01.06
month
SELL
BUY
Sales
UP 23%
DOWN 14%
Month
190.34
BTC USD


The End