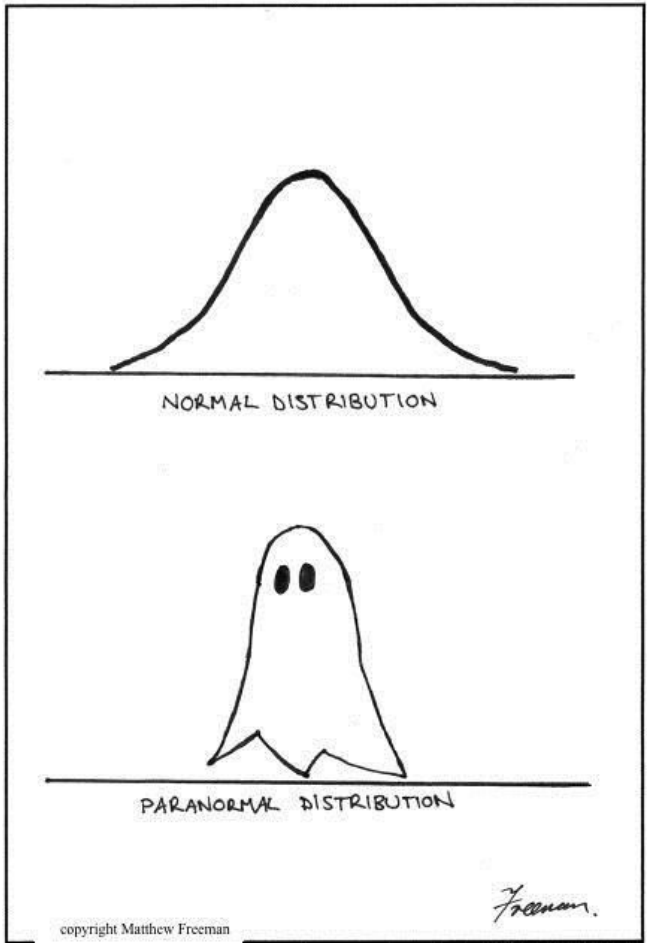
Welcome to Week 15, Lecture 01!

Intro to Hypothesis Testing & Mock Belt Exam

Agenda

- Intro to Hypothesis Testing
- Mock Exam



Assignments

This week's assignments:

- Project 3 Part 3(Core)
- Describing Distributions (Core)
- Hypothesis Testing with Insurance Data(Core)

Remember that Belt Exam eligibility is this Friday at 9AM

- **Make sure you have all of week 1 and 2 assignments submitted and any resubmits from week 1.**
- **Note that content from this week (week 3 of the stack) is assessed on the exam!**

Hypothesis Testing


Reject,
or fail to reject.
There is no accept.

Hypothesis Testing: Getting Started

- Is there actually a *significant* difference between two groups or are the differences just due to randomness?

- Example:
  - A survey goes out and asks students to rate Jupyter notebooks and Google Colab each on a scale of 1-5.

  - On average, Jupyter scores higher, but can we conclude that students really do prefer Jupyter notebooks or is it just due to randomness?

State the null and alternate hypothesis

Null Hypothesis:

- The null hypothesis is the one that seems like it doesn't need to be stated! It is that there is nothing special going on. In our case:

- **Null Hypothesis (H_0):** There is NO difference between ratings for Jupyter notebooks and Google Colab

Alternate Hypothesis:

- The alternate hypothesis is usually what you would just think of as "the" hypothesis! It states that something significant is going on.

- **Alternate Hypothesis: (H_a)** There is a significant difference between ratings for Jupyter notebooks and Google Colab

Establish the significance Level (alpha)

- The most common significance value is alpha = 0.05.
- This means that if the likelihood of the results due to random chance (p) is less than 5%, we consider the results significant (and not just random).
- If we use a significance value of 0.01, it means than the likelihood of the results due to randomness (p) must be less than 1% in order for us to claim it is significant.
- Setting an alpha value establishes your willingness to accept Type 1 or Type 2 errors, and really it depends on your data and situation. When in doubt, consult a SME!

Examples of Type 1 and 2 Errors with Hypothesis Testing

Perform the test to get a p value

There are many different statistical tests, but the interpretation is generally the same!

- Whichever statistical test you use, you will be given a p-value in your results.
- The p-value is the probability of your situation (like the difference in ratings for Jupyter and Colab) occurring at random.
- If your p is < alpha, you have a significant result and reject the null hypothesis.
  - You conclude there is a significant difference in the ratings.
- If you p is > alpha, you do not have a significant results and “fail to reject” the null hypothesis.
  - You conclude there is NOT a significant difference in the ratings.

Mock Belt Exam

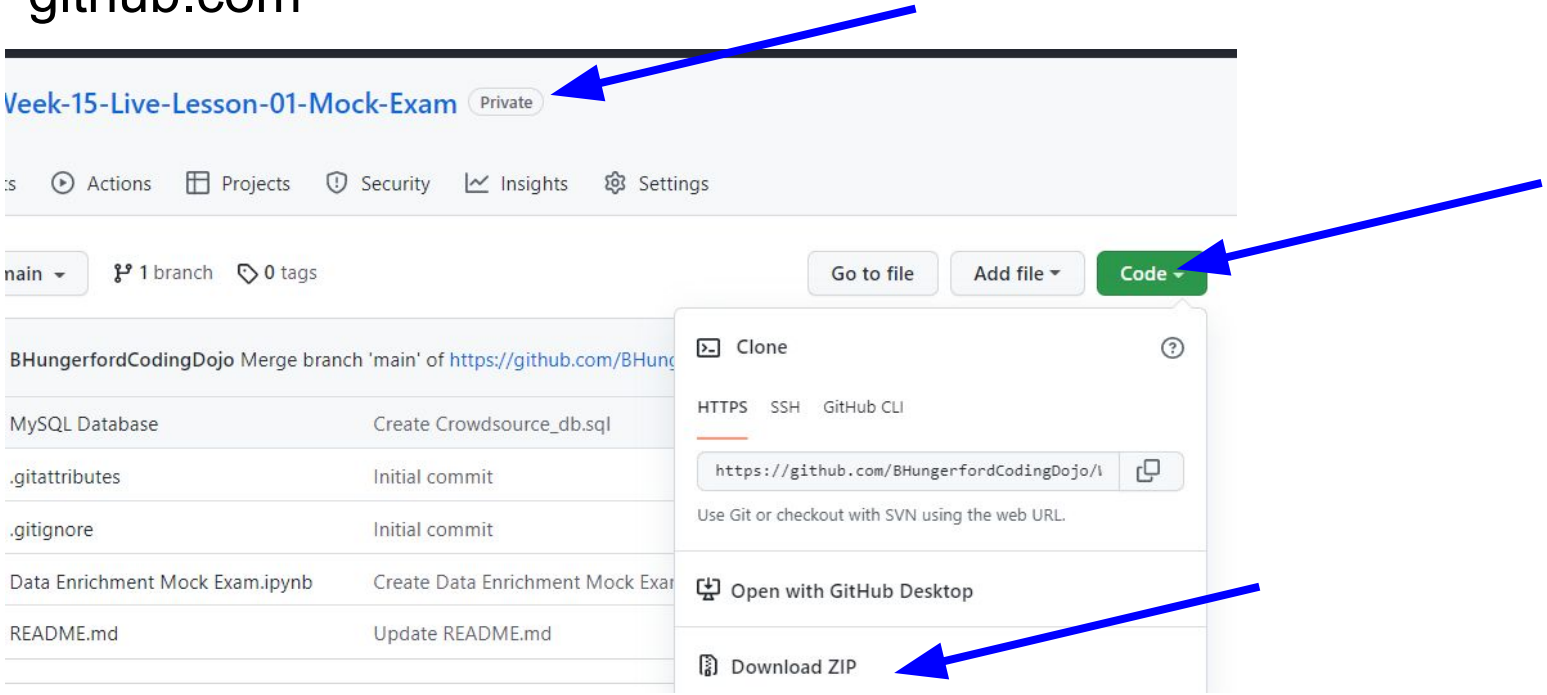
What is on the belt exam?

- You will NOT have to make API calls on the exam
- You WILL be given a .json file that simulates the results of making API calls
- You will be asked to:
  - navigate the .json file to find information (EXTRACT)
  - convert the json records into pandas dataframes
  - make some stated transformations to the features (TRANSFORM)
  - create a database using SQLAlchemy and add the pandas dataframes as tables in the database using Python
  - Open the database in MySQL workbench and export it as .sql file (LOAD)
  - Perform a hypothesis test on the data
  - Submit your final repository as a zipped file on github.com
    - The .ipynb notebook
    - The .sql file
- You WILL NOT need to normalize the tables
- You WILL NOT need to create an ERD
- You WILL NOT need to do cleaning or analysis beyond what is stated or what is needed to effectively perform the hypothesis test

The API results for your mock exam can be found here.

1. Open the .json file and explore it
2. Find 4 records and convert them into dataframes
3. Make the following transformation:
    a. In the financials record, remove '$' from funded_amount and change to numeric
4. Create a database with SQAlchemy and add the tables to the database
5. Open MySQL workbench and export the .sql file
6. Perform a hypothesis test to determine if there is a significant difference between the funded amount when it is all males and when there is at least one female in the group
7. Push your .ipynb and .sql files to your github repository and download as a zip file (see next slide)

How to Download your final repository as a zip file on github.com



MOCK BELT EXAM SOLUTION

- Solution Repository:
  - Note: the solution is **A** solution and is not the only way of accomplishing the mock belt exam's task. There are several ways of doing some of the steps and I tried to demonstrate some of them as different "approaches" - indicated in the headers.
  - Solution Repository: https://github.com/coding-dojo-data-science/data-enrichment-mock-belt-exam