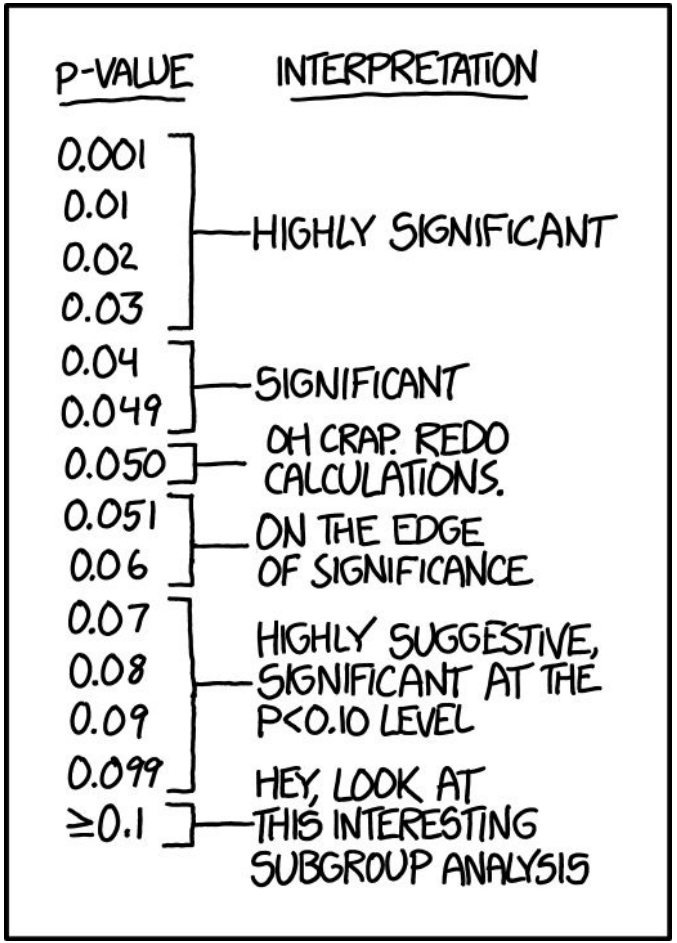


Image Source

Welcome to Week 15, Lecture 02!

Hypothesis Testing with Python

06/08/22

Agenda

- Assignments & Announcements
- Belt Exam (Eligibility & Material Covered)
- Hypothesis Testing with Python:
  - Quick Overview
  - Apply Hands-On with CodeAlong

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!**

📣 Announcements

- **Reminder:**
  - 🥋 **Optional Class - Reviewing Mock Belt Exam Solution**
    - Tomorrow, 06/09/22 @ 5 pm PST.
    - It is completely optional & will be recorded.

- 🤩 **Bonus Lecture Next Week:**
  - It is completely optional & will be recorded.
  - Possible Topics:
    - Deep Dive into Object-Oriented Programming
    - Creating Your Own Python Package/Module
    - Creating Dashboards with Streamlit
  - **Please answer the following poll to decide the topic and time slot:** https://forms.gle/bUxX4rznSEkpgaPR9
    - I will check Friday morning to determine the winner.

Belt Exam

Belt Exam Eligibility

- **Final deadline for Belt Exam eligibility is Friday 06/10 at 9 AM PST.**
- Requirements:
  - All week 13 and week 14 assignments **submitted**.
  - All week 13 **resubmits corrected.**
    - It is YOUR responsibility to check your feedback doc for resubmission requests.
  - No more than 1 missed lecture.

- **If you have a Friday one on one:**
  - DO NOT UNLOCK YOUR EXAM BEFORE WE MEET!!!

What is on the belt exam?

- **You will NOT have to make API calls** on the exam
  - Instead, 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 downloaded from github.com
    - The .ipynb notebook
    - The .sql file

- **You WILL NOT need to:**
  - normalize the tables.
  - create an ERD.
  - Do any cleaning or analysis beyond what is stated/needed to perform the hypothesis test

Important Belt Exam Reminders

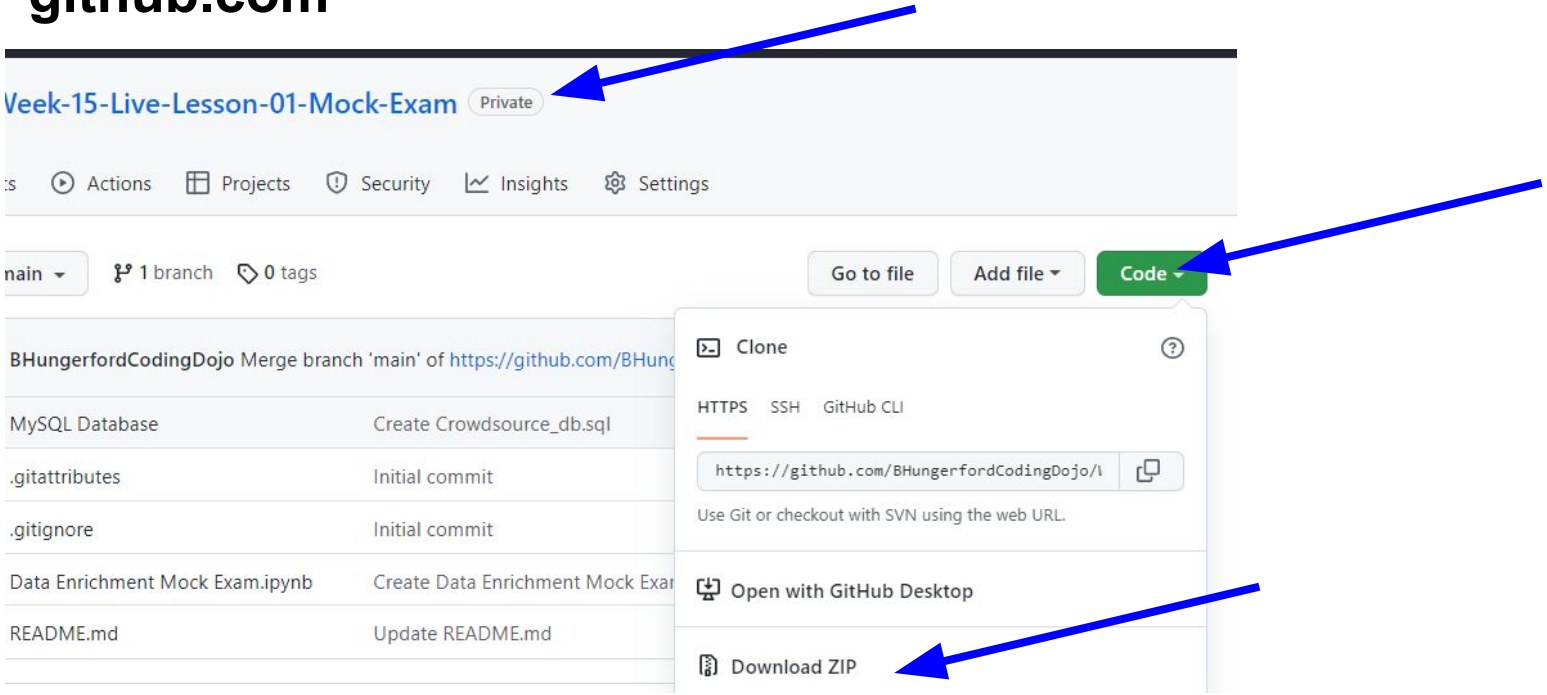
- **Do NOT use a Public repository for your belt exam!!!**
  - When you publish your repo, **keep the "Keep Code Private" box checked!**
  - **If you accidentally make it Public, change it to Private IMMEDIATELY!**
    - See this help article for how to change the visibility after publishing

- **Use a private repository, commit and push your final work to GitHub, and then follow the instructions on the next slide to download a zip file of your private repo** to turn in as your final submission.

- **Your repo/zip file should contain:**
  - a jupyter notebook (.ipynb file) with your work
  - a .sql file of your exported database. (Instructions for exporting database from mysql workbench: https://login.codingdojo.com/m/376/12528/88194)

- **If anything goes wrong when submitting your belt exam,** IMMEDIATELY email me your zip file as a backup. jirving@codingdojo.com
  - Email submissions sent more than 24 hours after you unlock your exam will not be graded.

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



Hypothesis Testing with Python

Hypothesis Testing: Overview

- Before we can start performing a hypothesis test:
  - We need to have a well-defined hypothesis and null hypothesis.
  - We need to determine WHICH test would be appropriate for our question

- Once we know which test is appropriate, we will know what assumptions we need to check.

- After we've checked the assumptions, we conclude if we can run the test that we planned.
  - If not, there is an alternative test that answers the same question, but uses different calculation to get the p-value.

Selecting the Right Test

Everything covered in the next few slides can be found in the:
"Guide: Choosing the Right Hypothesis Test" lesson.

Test Selection Overview

- **STEP 1: Stating our Hypothesis**

- **STEP 2: Determine the category/type of test based on your data.**
  - Answer 2 questions about what you are comparing to determine which test.
  - Once you know which test, check the assumptions of the test.

- **STEP 3: Does the data meet the assumptions of the selected test?**
  - If you meet the assumptions, run the desired test.
  - If you don't meet the assumptions, run the non-parametric equivalent test.

- **STEP 4: Perform Test & Interpret Result**
  - Run the selected statistical test to get your p-value and interpret it.
  - Do you reject your null hypothesis or did you fail to reject it?

- **STEP 5: Post-hoc multiple comparison tests (if needed)**

STEP 1: Stating our Hypothesis

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

STEP 2: Determine the category/type of test based on your data.

Test Selection Question 1

- **Question 1: Is my data Categorical or Numerical?**
  - What type of data is the target of your question?
  - Are you asking about differences in a measured value or differences in which category/group something belongs to?

- **Examples of Numeric Data:**
  - A company wants to compare the time spent on different versions of their homepage. Does one version make users stay on the page significantly longer?

  - An instructor expects an exam average to be roughly 85%, and wants to know if the actual scores line up with this expectation. Was the test actually too easy or too hard?

- **Examples of Categorical Data:**
  - A pollster wants to know if men and women have significantly different flavor preferences for ice cream. Does a result where more men more often answer "chocolate" as their favorite reflect a significant difference in the population?
  - Are men and women equally likely to be smokers?

Test Selection Question 2

- **Question 2: How many samples/groups am I comparing?**
    - 1 Sample (i.e., comparing to an ideal target)
        - i.e., comparing an actual result against a desired target or Key Performance Indicator (KPI)
    - 2 Samples
        - i.e., comparing a control and treatment group or an A/B test
    - More than 2 Samples
        - i.e., comparing three different variants of a landing page

Select the Right Test Using your Answers

What type of comparison?	Numeric Data	Categorical Data
Sample vs Known Quantity/Target	1 Sample T-Test	Binomial Test
2 Samples	2 Sample T-Test	Chi-Square
More than 2	ANOVA and/or Tukey	Chi-Square

STEP 3: Does the data meet the assumptions of the selected test?

Test Assumptions

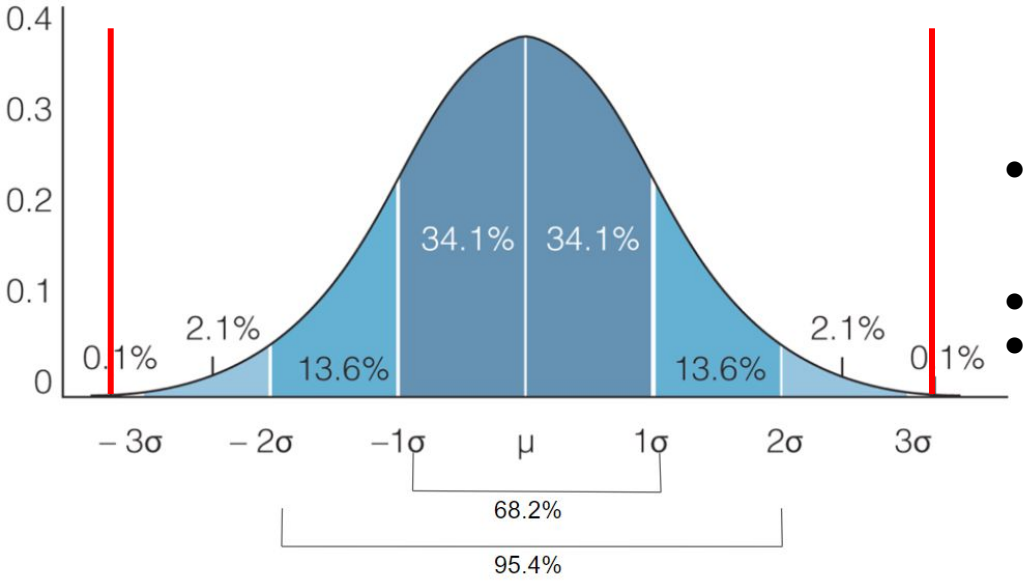
- Each statistical test was built with **specific assumptions about the data.**
- **If we do not meet** a test's assumptions, **we cannot trust the result.**

- **For numeric data, the assumptions are usually:**
  - No significant outliers
  - Normally distributed GROUPS
  - Equal Variance between groups.

- **For categorical data, the assumptions are usually:**
  - The outcome is binary (e.g.: two groups, success vs failure)
  - The trials/observations are independent.

- **For a list of the assumptions for each test:**
  - See the Choosing the Right Hypothesis Test lesson.

Testing for/Removing Outliers



- **<u>Z-Score Rule:</u>**
  - Any value that has a z-score more than more than 3 standard deviations away from the mean is considered an outlier.
- Outliers must be removed from each group SEPARATELY!
- Use **scipy.stats.zscore** to calculate z-scores.
- Any z-score that has an absolute value >3 is an outlier.

Testing for Normally Distributed Groups - 1

- Use a statistical test to check for if groups normally distributed:
  - D'Agostino-Pearson's normality test: Scipy.stats.normaltest

- To interpret the p-value from the normal test you must check the null hypothesis of the normaltest.

scipy.stats.normaltest

`scipy.stats.normaltest(a, axis=0, nan_policy='propagate')` [source]

Test whether a sample differs from a normal distribution.

This function tests the null hypothesis that a sample comes from a normal distribution. It is based on D'Agostino and Pearson's [1], [2] test that combines skew and kurtosis to produce an omnibus test of normality.

Testing for Normally Distributed Groups - 2

- However, if your group n's are large enough, its ok if you do not pass a normaltest.

Parametric Test	Sample size guidelines for non-normal data
1-sample t-test	Greater than 20
2-sample t-test	Each group should be greater than 15
One-Way ANOVA	If have 2-9 groups, each group n >= 15. If have 10-12 groups, each group n>20.

Testing for Equal Variance

- Use a statistical test to check for if groups have equal variance:
  - Levene's Test: Scipy.stats.levene

- To interpret the p-value from Levene's test you must check the null hypothesis of the test:

scipy.stats.levene

`scipy.stats.levene(*args, center='median', proportiontocut=0.05)` ¶ [source]

Perform Levene test for equal variances.

The Levene test tests the null hypothesis that all input samples are from populations with equal variances.

Final Test Selection

- Then we select the non-parametric equivalent of our original parametric test.
- All functions below are located in the stats module of scipy.
    - See this cheat sheet for more info on each of them cheatsheet: Hypothesis Testing with Scipy

```
from scipy import stats
```

Parametric tests (means)	Function	Nonparametric tests (medians)	Function
1-sample t test	`stats.ttest_1samp()`	**1-sample Wilcoxon**	`scipy.stats.wilcoxon`
2-sample t test	`stats.ttest_ind()`	**Mann-Whitney U test**	`scipy.stats.mannwhitneyu()`
One-Way ANOVA	`stats.f_oneway()`	**Kruskal-Wallis**	`stats.kruskal`
Binomial test	`stats.binom_test()`	N/A	N/A
Chi-Square test	`stats.chi2_contingency()`	N/A	N/A

- **Perform the test to get your test-statistic and the associated p-value.**

STEP 4: Perform Test & Interpret Result


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

Perform Final Test & Interpret

- **Perform the final test selection after checking assumptions:**
  - See this cheat sheet for more info on using the scipy functions. cheatsheet: Hypothesis Testing with Scipy

- Whichever statistical test you use, you will be given a p-value in your results.

  - The p-value is the probability of your situation occurring due to random chance.

- **Interpret Your P-Value:**
  - If the p-value is > alpha:
    - We fail to reject the null hypothesis. There is no significant difference between groups.

  - If the p-value is < alpha:
    - Reject the null hypothesis. There is a significant difference between groups. We have supported the alternative hypothesis.

    - **If you have multiple groups (i.e. ANOVA, Kruskal-Wallis),** see Step 4: Post-Hoc Tests in order to determine which groups were different.

STEP 5: Post-hoc multiple comparison tests (if needed)

Post-Hoc Tests

- When our hypothesis includes more than 2 groups, our p-value indicates there IS a significant difference between groups, but we not WHICH groups.
  - We must run a pairwise Tukey's test to know which groups were significantly different.
- Tukey pairwise comparison test
    - Statsmodels.stats.multicomp.pairwise_tukeyhsd
  - Tukey's test will run separate tests on pair of groups to get a separate p-value for each. But it does it in a smart way that prevents false positives.

Activity

Hypothesis Testing CodeAlong/Activity

- We will be answering questions about crowdfunded loans from the website Kiva.
- The data and questions to answer are already in the activity repo:
  - https://github.com/coding-dojo-data-science/data-enrichment-wk15-lect02-activity
- You are welcome to Code Along with me, but the second hypothesis will be more time-intensive so don't feel bad if you can't keep up with Hypothesis #2!

- Branch from class:
  - "06-08-22-class"
  - Notebook: https://github.com/coding-dojo-data-science/data-enrichment-wk15-lect02-activity/blob/06-08-22-class/Class-Wk15-L02-Hypothesis-Testing.ipynb