Part 1: Hypothesis Testing Recap

Include the answers in `individual_answers.md`

1. State which test should be used for the following scenarios to calculate p-values. Explain your choice.
    - You randomly select 50 dogs and 80 cats from a large animal shelter, and want to know if dogs and cats have the same weight.
    - A random sample of San Franciscans and Oaklanders were surveyed about their favorite baseball team, and you want to determine if the same proportion of people like the SF Giants.
2. A study attempted to measure the influence of patients' astrological signs on their risk for heart failure. 12 groups of patients (1 group for each astrological sign) were reviewed and the incidence of heart failure in each group was recorded.

    For each of the 12 groups, the researchers performed a z-test comparing the incidence of heart failure in one group to the incidence among the patients of all the other groups (i.e. 12 tests). The group with the highest rate of heart failure was Pisces, which had a p-value of .026 when assessing the null hypothesis that it had the same heart failure rate as the group with the lowest heart failure rate, Leo. What is the the problem with concluding from this p-value that Pisces have a higher rate of heart failure than Leos at a significance level of 0.05? How might you adjust your interpretation of this p-value?

Part 2: Analyzing Click Through Rate

Please submit your final code in `individual.py`

Download the data here.

We will use hypothesis testing to analyze Click Through Rate (CTR) on the New York Times website. CTR is defined as the number of clicks a

user makes per impression that is made upon the user. We are going to determine if there is statistically significant difference between the mean CTR for the following groups:

```
1. Signed in users v.s. Not signed in users
2. Male v.s. Female
3. Each of 7 age groups against each other (7 choose 2 = 21 tests)
```

1. Calculate the adjustment needed to account for multiple testing at the 0.05 significance level.
2. Load `data/nyt1.csv` in a pandas dataframe.

   Use `data.info()` to make sure the data types are valid and there are no null values. This data has been cleaned for you, but generally it is good practice to check for those.
3. Make a new column `CTR` using the `Impressions` and the `Clicks` columns. Remember to remove the rows with `0` impressions.
4. Plot the distribution of each column in the dataframe. Do that using `data.hist()`. Check out the arguments you can use with the function here . Set the `figsize=(12,5)` to make sure the graph is readable.
5. Make 2 dataframes - one a dataframe of 'users who are signed in' and a second of 'users who are not signed in'. Plot the distributions of the columns in each of the dataframes. By visually inspecting the two sets of distributions, describe the differences between users who are signed in and not signed in?
6. Use a Welch's t-test to determine if the mean CTR between the signed-in users and the non-signed-in users is statistically different. Explain how you arrive at your conclusion.

   The Welch's t-test assumes the two populations in which the samples are drawn from have different variances.

   ```
   scipy.stats.ttest_ind(a, b, equal_var=False)
   ```
7. Determine if the mean CTR between male users and female users is statistically different. Use only the rows where the users are

signed in. Is the difference in mean CTR between signed-in users and non-signed-in users more worthy of further investigation than that between male and female? Explain your answer. `Male: 1, Female: 0`

8. Calculate a new column called AgeGroup, which bins Age into the following buckets `'(18, 24]', '(24, 34]', '(34, 44]', '(44, 54]', '(54, 64]', '(64, 1000]', '(7, 18]'`

   Use only the rows where the users are signed in. The non-signed in users all have age 0, which indicates the data is not available.

   Use pandas' cut function. `pandas.cut(signin_data['Age'], [7, 18, 24, 34, 44, 54, 64, 1000])`

9. Determine the pairs of age groups where the difference in mean CTR is statistically significant. Collect the p-values and the difference of the means for each pair. Store these results in a `DataFrame`.

   Rank (in descending order) the difference in mean CTR for the pairs that are statistically significant. Comment on the trend you observe for groups `(64, 1000]`, `(54, 64]` and `(7, 18]`. Feel free to include additional trends you observe.

   Rank (in ascending order) the difference in mean CTR for the pairs that are *statistically insignificant*. State the 3 groups that are the least different in mean CTR and provide an explanation for why you think this is true.