Group Project

Yuhan Zheng

Goal

- Use statistical inference to answer **at least** one question of interest of your choice, using at least one dataset of your choice.

What’s include

- Include background on the topics, and clearly indicate the question(s) of interest.
- Describe the dataset were obtained from and describe variables in as much detail as another person would need to replicate the analysis.
- Display at least one polished, well-labeled visualization.
- Include at least one interpretation (CI or p-value).

Data

- Dataset enables you to answer a question of interest with one of the statistical inference methods we learn in this class.

Note: You may obtain a dataset from a repository such as Kaggle, but citing Kaggle as a source is not enough - you will need to find information about how the data was originally collected.

Data

https://github.com/awesomedata/awesome-public-datasets

https://www.data-is-plural.com/

https://archive.ics.uci.edu/ml/datasets.php

https://msropendata.com/categories

https://methods.sagepub.com/Datasets

https://www.pewresearch.org/download-datasets/

https://opendata.cityofnewyork.us/

https://ieee-dataport.org/datasets

https://catalog.data.gov/dataset

https://datahub.io/collections

Example of Data Description

Moon Phase Data (lunar_phases.csv)

Unlike our crime data, our moon phase data does not come from a ready-made dataset. It, instead, was built from 1096 calls to the Naval Observatory API[2]. The resulting dataset has two columns: `date` which specifies the date each row corresponds to, and `lunar_phase` which says what the moon phase was on each date at noon eastern time. The dates in the `date` column are of the same range as that in the crime dataset (i.e., January 1st, 2019 to December 31st, 2021).

Note that, on average, each year only has 12 full moons. Thus, we will only have about 36 full moon days to compare against non-full moon days. This relatively small sample size may impact our results.

Question of Interest

- The question of interest should drive the central story of your entire project.
- You should use data from a sample to infer something about a property (parameter) of the population.

Bad Questions and Good Questions

Bad Questions of interest:

- The question is descriptive rather than statistical, e.g.:

❖Which exoplanet is the most similar size to Earth?

❖Which states had the highest COVID rates?

- The question is too broad or ambitious to answer with statistical inference, e.g.:

❖Who was the best player in the league last season?

❖How well can we predict the number of crimes in a given day in a NYC borough?

Bad Questions and Good Questions

Good questions of interest:

- Questions are about specific population parameters and are therefore appropriate for this project, e.g.:

  - Is there a difference in the average age of smokers and non-smokers among Wisconsin adults? (Difference in means)
  - Is there a relationship between in-state tuition costs and number of undergraduate students enrolled among American universities? (Linear regression)
  - Is there a difference in the proportion of car accidents which are fatal between night (9pm to 6am) and day?

Project Proposal **(DUE: W 11/20, 11:59pm)**

- The names of the students in the group.
- A brief description of how the group will communicate and work together.
- The questions of interest that motivate the planned study.
- Links to where one can find and download the dataset(s) you are going to use.
- A description of the data. (Data cleaning methods)
- Include a description of a graph which will show how the data informs the primary question of interest.
- Identify which type of inference method you are going to use.

Turn in: submit a knitted .html file and the corresponding R Markdown file

Project Draft **(DUE: W 12/4, 11:59pm)**

- Introduction
  - Provide background and motivation
  - Question(s) of interest.
  - Thesis statement.
- Background
  - Data description
  - Background information needed to interpret questions
  - Describe any unusual factors in the data
  - Data cleaning part
  - AT LEAST ONE relevant graph

Project Draft **(DUE: W 12/4, 11:59pm)**

- Statistical Analysis
  - Include all of your technical vocabulary.
  - Including a definition of your parameter(s) of interest.
  - Declaration of what type of inference you are using.
  - Checking of assumptions, stating of hypotheses, reporting of at least one confidence interval OR p-value.
- Discussion
  - Interpret your confidence interval(s)/p-value(s).
  - Discuss any shortcomings of the analysis.
  - Discuss any additional questions that came up during your analysis

Turn in: dataset(s), R Markdown file and a knitted HTML file which contains your *full* report.

Example of Analysis

Initial Impressions

As seen in the figure below, in our sample set, the mean number of crimes on full moon days is approximately equal to that on non-full moon days.

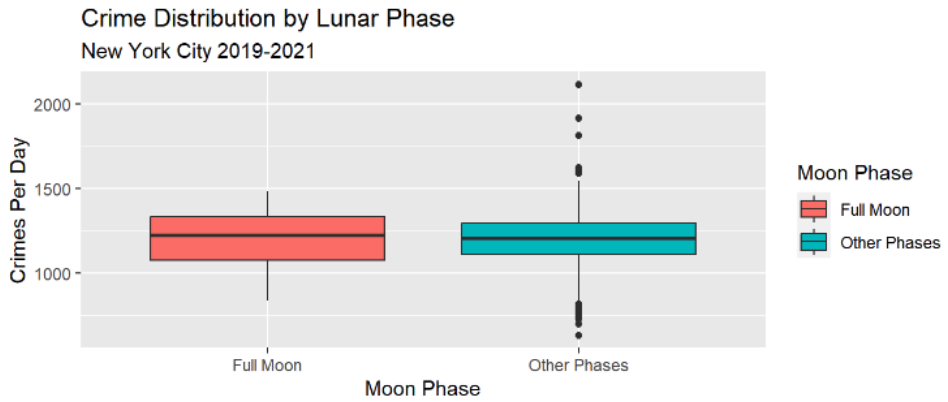

Figure 1. Distribution of crimes in our sample set by lunar phase.

While this plot may hint at the crime rates not being affected by the phase of the moon, we need to conduct further analysis to say anything conclusively.

Statistical Model

Let X_{1i} be the number of crimes that occurred on the ith full moon day in our sample set and let X_{2j} be the number of crimes that occurred on the jth non-full moon day. These random variables are modeled by

$$X_{1i} \sim \mathrm{F}(\mu_1, \sigma_1) \quad i = 1, \dots, n_1$$
$$X_{2j} \sim \mathrm{F}(\mu_2, \sigma_2) \quad j = 1, \dots, n_2$$

for some unknown distribution F, where μ_1 and σ_1 are the respective mean and standard deviation of X_{1i}, μ_2 and σ_2 are the same of X_{2j}, and n_1 and n_2 are the respective number of full moon and non-full moon days in our dataset.

Hypotheses

With these models, we will then conduct hypothesis testing with

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 > \mu_2$$

using a two-sample t-test for a difference in means.

Interpretation

> We do not have statistically significant evidence to conclude that the mean crime rates differ from full moon days to non-full moon days ($p = 0.4407754$, one-sided t-test).

Peer Review **(DUE: SATURDAY 12/7, 11:59pm)**

- Each individual student will be asked to read and comment on the draft report HTML of another randomly assigned group.
- Give at least one specific thing they did well and one specific thing they could improve in each section.
- Good Example:

"Choosing to graph the number of gold medals with gold points, silver medals with silver points, and bronze medals with bronze points is a really nice touch! However, the outlier in the top right is really stretching your graph out - maybe consider a log transformation?"

- Bad Example:

"Graphs are really good. "

Project Presentation **(During last discussion section, M 12/9)**

- A real-life terms description of what kind of information is in your dataset.
- A real-life terms description of what parameter you are interested in, and what kind of test you used to assess your hypotheses.
- Visualizations or summary statistics that aid your central point
- A discussion part.

Note: Maximum of 8 minutes for the presentation.

Not everyone has to speak during the presentation.

Everyone is expected to contribute to the presentation one way or another.

Final Project Report **(DUE: W 12/11, 11:59pm)**

- Instructions like project draft.
- Account for your official grader's feedback on your first draft in the final report.
- Expected to read the feedback from your peer reviews and consider making changes based on that feedback. (not required to implement every piece)

Note:

- The goal of the .html full report is to concisely, persuasively, and thoroughly describe your analysis and results.
- Although we have no word or page limit for reports, please note that longer reports are not necessarily better.
- You may NOT use inference methods from outside the course.
- The expectation is that each member of the group will share in the work of completing the project.


Thank you for listening!