

## **Project 2**

### **STAT 207 - Data Science Exploration**

Due: Tuesday, March 26 by 11:59 pm on GitHub

#### **Main Goal of Analysis**

The main goal of this project is to use a dataset to make inferences about a population, quantifying uncertainty in the process. The secondary goal is to contrast the inferences with summaries for a sample.

You are required to perform two main analytical tasks:

1. Estimating a parameter with a confidence interval
2. Defining and testing theories about an unknown parameter.

Additional descriptions for these tasks can be found later in this document. If you think that there are additional questions or analyses that would add additional insights to your overall research goal, you're more than welcome to pursue these in addition to what is required in this document.

#### **Project Format**

Your project will be submitted as a written report to GitHub. This report is worth 40 points. Additional details for the report can be found below.

#### **Group Structure**

You can work in groups of up to 3 people, or you can work individually.

- If you work with a group of 3, you must do at least 25% of the work in order to get full credit.
- If you work with a group of 2, you must do at least 33% of the work in order to get full credit.

#### **Dataset Options**

You can choose your own dataset, or you can choose from one of the three supplied datasets below. There are several places you can go to find interesting datasets, but here are some places to start:

- <https://www.kaggle.com/datasets>
- <https://corgis-edu.github.io/corgis/csv>
- <https://archive.ics.uci.edu/ml/datasets.php>
- <https://github.com/fivethirtyeight/data>
- For sports data, you may choose to explore (nflfastr.com for NFL, billpetti.github.io/baseballr for MLB, cfbfastr.sportsdataverse.org/index.html for CFB, and sportsdataverse.org for more sports data).

**Choosing your own data:**

If you choose your own data, it must meet the following specifications.

1. It must have at least three variables total (five for later projects)
  - a. Variables that have uninformative information don't count and won't be useful. Examples of uninformative variables include those that provide the observational unit (a row name or row id) or are a linear combination of other variables in the datasets. If you aren't sure, come ask!
2. It must have at least two categorical variable
  - a. For at least one of these categorical variables, you will use only two levels; that is, only two of the possible options that it contains. You can filter your data at the time to accommodate this need
3. It must have at least one quantitative variable
4. It must have at least 50 rows

You may continue using the same dataset from Project 1, or you could change to a different dataset.

**Provided dataset:**

1. Video Games Data (video\_games.csv)
  - a. This dataset has information about the sales and playtime of over a thousand video games released between 2004 and 2010. The playtime information was collected from crowd-sourced data on "How Long to Beat"
  - b. This was originally collected and curated by Dr. Joe Cox
  - c. This data was originally downloaded on 2/8/2024 from here:  
<https://researchportal.port.ac.uk/en/datasets/video-games-dataset>
  - d. Read more about this data here:  
<https://researchportal.port.ac.uk/en/publications/what-makes-a-blockbuster-video-game-an-empirical-analysis-of-us-s>

## Project Report Specifications

Deadline: Tuesday, March 26 by 11:59 pm on GitHub

Format:

- Jupyter notebook
- This should be a clean data analysis report that you could submit to an employer or client (not a homework assignment). At the very least, your report should have a title, headings for each section, and be written in paragraphs and with complete sentences.
- You can use and modify the attached project\_02\_template.ipynb file as a template for this report if you'd like. You can add and delete as many cells as you'd like in this file.

### 1. Introduction [6 points]

Goal: In your introduction, you should orient the reader to what they are about to read. This will help to prepare the reader, so that they can figure out how to connect the different components that they will read.

While working towards this goal, you should complete/address the following:

- a. Title: Give your research report a title
- b. Dataset introduction: You should briefly introduce the reader to the context of your dataset and the available data.
- c. Populations and Samples: Does your data represent a population or a sample? What is the corresponding population of interest? If your data represents a population, you should take a random sample of your data for this project.
- d. Motivation: Describe the motivation for why you or someone else would want to explore your dataset (or a dataset of this type). Describe one person or the type of person who would find these insights useful. You can give background research (with citations) if this would help back up your motivation.
- e. Research Questions: You will answer two **sets** of research questions. In your introduction, you should briefly describe your research questions and why you (or someone else) would be interested in answering these research questions. How could the answers to these research questions be used?

### 2. Confidence Interval Analytical Task [18 points]

Goal: For your confidence interval analytical task, you should pick one of your variables of interest (either categorical or quantitative). You will explore the behavior of this variable *in the dataset* (descriptive analytics) and *in the population* (inference).

While working towards this goal, you should complete/address the following:

- a. State your research questions: Your research question should have two parts. For example, you might ask "What are important features of my variable in this dataset? What are a range of reasonable values for at least one important feature of this variable for the underlying population?"
  - i. Remember, descriptive analytics only involves describing relationships in the dataset that you have, so your first research question should be *just* about the data. Then, inferential statistics involves answering research questions about populations given a random sample from that population.

Your second research question should reference the population from which your data were collected.

- b. Dataset cleaning: Be sure to clean the data for your variable of interest. You should filter your data to be sure it only contains relevant data for your population of interest. Be sure to describe these steps, along with any limitations.
- c. Descriptive analytics:
  - i. Numerical summaries: Calculate at least one appropriate numerical summary for your variable of interest.
  - ii. Visualization: Generate an appropriate visualization to observe your variable of interest.
  - iii. Interpret the results of these two descriptive analytic tasks.
  - iv. What is an appropriate measure to describe a relevant feature of your variable of interest?
- d. Create a confidence interval:
  - i. Select an appropriate confidence level
  - ii. Simulate a sampling distribution for your statistic of interest
  - iii. Use this simulated sampling distribution to estimate your confidence interval
- e. Interpret your confidence interval & confidence level.
  - i. Provide a formal interpretation for both of these values.
- f. Finally, discuss how your approach and conclusions address your research question.

### 3. Hypothesis Testing Analytical Task [18 points]

Goal: For your hypothesis testing analytical task, you should pick one categorical variable and a variable as your primary variable of interest (either categorical or quantitative, but be sure to pick the opposite type of the variable from your confidence interval question). You will explore the relationship between these variables *in the dataset* (descriptive analytics) and *in the population* (inference).

While working towards this goal, you should complete/address the following:

- a. State your research question: For instance, you could ask “What is the relationship between x and y in this dataset? Is there an association between x and y in my population?”
- b. Dataset cleaning: Be sure to clean the data for your variable of interest. You should also filter your data to be sure it only contains relevant data for your population(s) of interest. Be sure to describe these steps, along with any limitations.
- c. Population definition: Do you have one population of interest with two variables recorded for it? Or do you have two (or more) populations of interest with samples for each?
- d. Descriptive analytics:
  - i. Numerical summaries: Calculate at least one set of appropriate numerical summaries to observe the relationship between your variables of interest.

- ii. Visualization: Generate an appropriate visualization to observe the relationship between your variables of interest.
  - iii. Interpret the results of these two descriptive analytic tasks.
  - iv. What is an appropriate measure to describe a relevant feature of your variable of interest for each of your groups?
- e. Perform a hypothesis test:
  - i. For the hypothesis test, we will use only two levels for one of our categorical variables. This allows us to use our previously discussed approaches. Note that this may result in a simplification/adjustment of your original data.
  - ii. State your hypotheses, including a definition and description of your parameter of interest.
  - iii. Select a significance level.
  - iv. Check the conditions for this test
  - v. Simulate a sampling distribution for your statistic of interest
  - vi. Calculate a p-value for this test based on the sampling distribution.
  - vii. Evaluate your results to make a decision and state a conclusion
- f. Interpret your significance level & p-value:
  - i. Provide a formal interpretation for both of these values.
- g. Finally, discuss how your approach and conclusion address your research question

#### 4. Conclusion [6 points]

- a. Summarization: Summarize your confidence interval and hypothesis test tasks in the conclusion. Provide at least a paragraph. (This will likely be a restatement of what you have already included in your report).
- b. Limitations: What limitations did you face in your analysis, results, or interpretations? What challenges did you face in your data analysis? What contextual information is important before you make strong claims from these results? How might these affect how the person you described in the introductory motivation paragraph uses these results?
- c. Future work: If you (or someone else) were to conduct future work based on these analyses, what kind of research questions or analyses might that entail?

The remaining 2 points of this project will be graded on writing quality, clarity, conciseness, and professional and neat formatting of the report.

#### Intended Audience/Reader of your Project

The intended audience of your report/presentations should be someone who has the same level statistical/python knowledge as you and your STAT 207 classmates. **Theoretically, you should be able to send your report to one of your classmates (who is not on your team), and they should be able to understand everything that you did and the claims that you are making.**

## Grading

In addition to being graded for correctness and completion (as noted), this project will be graded on a qualitative basis. Qualitatively, we will be looking for the following things:

- **Clarity about Analyses, Algorithms, and Data Choices**
  - Someone who has taken a STAT207-level class should be able to read through your report and easily be able to do the following:
    - Replicate what you did in your analyses.
    - Know why you made the choices that you did in your analyses.
- **Clarity about Motivation (i.e. the “so what?”) of your analyses**
  - Beginning of the Report
    - Someone who is **about to** read the body of your report should be able to clearly answer the questions:
      - Why should I (or someone else) care about the report that I am about to read?
      - What research questions do they intend to answer?
      - How do these research questions relate to their motivation?
    - Therefore, in the introduction of your report you should make this clear.
  - Middle of the Report:
    - While **in the middle of** your report, your audience should be able to clearly answer the question:
      - How do each of these analyses/algorithms/data choices that they’re making/using tie back into the overarching motivation of this whole analysis?
    - Therefore, each new analysis/model/algorithm/data choice that you make, you should explain this and make it clear to your audience.
  - End of the Report:
    - Someone who has **just finished** reading your report should be able to clearly answer the questions:
      - Why should I (or someone else) care about the analysis that I just read?
      - Did their analyses and conclusions answer the research questions that they stated at the beginning of the report? If so, how? What were the answers to these research questions?
      - How would the results/answers to these research questions be useful to someone?
    - Therefore, in the conclusion of your report you should make this clear.