Case Study

LIDE Young Researchers Fellowship Program

Laboratorio de Investigación para el Desarrollo del Ecuador

1 Introduction

The case study project is the final deliverable for the *Introduction to Statistics* module. The project is designed to give you an opportunity to apply the statistical concepts you have learned in the course in a real-world setting.

You will work with a dataset and create a research questions in groups of up to 6 students. You will analyze the data, using what you've learned in the Introduction to Statistics course (and the R companion module, if you choose to), and present findings in a report and an academic-style presentation.

2 General guidelines

1. **Groups**: You will work in groups of up to 6 students. You can choose your own group members. If you do not have a group until **October 13th**, I will assign you to a group randomly among others who do not have a group.
    - I will provide a Google Sheets link where you will write down your group members.
2. **Deliverables**: You will submit a GitHub repository, which includes all analysis, a small paper (report), and a slides presentation. See below for more details.
3. **Data**: You will use the Registro de Entradas y Salidas, as provided by the Ecuadorian statistics agency, INEC. See below for more details.
4. **Research question**: You will need to come up with a research question that you will answer using the selected dataset. See below for more details.
5. **Methods**: You will need to use the statistical methods you have learned in the course to answer your research question. See below for more details.

6. **Software**: You can use any software you like to analyze the data. We recommend using R, as you have learned in the companion module, or Stata, as it was taught in the previous module. See below for more details.
7. **Class presentation**: You will present your findings in a 10-minute academic-style presentation. See below for more details. Attendance is mandatory to all presentations, as I will consider this as part of your grade for participation.
8. **Meetings**: You will need to meet with me at least once to discuss your project. You can schedule a meeting with me through Google Calendar (I will provide a link).

3 Dataset overview: Registro de Entradas y Salidas Internacionales (International Arrivals and Departures Registry) 2023

The Registro de Entradas y Salidas Internacionales (International Arrivals and Departures Registry), ESI for short, is a dataset provided by the Ecuadorian statistics agency, INEC. The dataset contains information on the number of international arrivals and departures in Ecuador in 2023. I have uploaded the dataset to the module repository, and you can access it here.

The dataset contains, among others, microdata on individual international arrivals and departures. The dataset contains information on demographics, reason of travel, arrival site, etc.

The dataset is in Stata format, but you can read it in R using the `haven` package. The dataset only includes data for 2023, but you may download prior years from the INEC website if your research question requires it.

4 Research question or topic

You need to come up with a specific research question that you will answer using the dataset. The research question should be specific, and should be answerable using the data provided. Example research questions could be:

- Do international departures increase during the summer months?
- Are there differences in the number of international arrivals between different regions of Ecuador? E.g. comparing coast to highlands.
- Are there differences in the number of international arrivals of different nationalities? E.g. comparing arrivals from the US to arrivals from Europe.
- Are tourism arrivals fundamentally different from business arrivals in their quantity?

You may pick one of these research questions, or come up with your own. Know that if you pick one of these I will expect a more detailed and rigorous nalysis, as these are more general questions and you didn't have to come up with them.

You may not only use the data provided, but use other data sources to answer your research question if it requires it. For instance, you may correlate with businessses data, or with weather data, or with other tourism data. You may benefit a lot from joining with other ESI datasets for other years, or with other datasets from INEC.

I encourage groups that may want to go a different way and not use this dataset. If you have a different dataset and research question in mind, please let me know and we can discuss it (especially if it relates to a group member's research paper or field of study). However, I do expect that at least one group will use this dataset.

5 Methods and analysis approaches

You will need to use the statistical methods you have learned in the course to answer your research question. All papers will have a structure where descriptive analyses will be needed, and then you will need to use inferential statistics to answer your research question.

You may use any statistical method you like, as long as it is appropriate for the research question. Below, a list (not exhaustive) of methods you may use:

- Single-sample z-test or t-test: To compare a sample mean to a hypothesized population mean.
- Two-sample z-test or t-test: To compare two sample means, assuming equal or unequal variances.
- Paired t-test: To compare two related samples.
- Proportion test (single-sample or two-sample): To compare a sample proportion to a hypothesized population proportion, or to a sample proportions to each other.
- Experimental hypothesis testing: To compare means or proportions between groups in an experimental setting. Two-way or multi-way ANOVA. You should be able to justify why are you using this method (e.g. you have a treatment and control group).
- Simulation: If you have a complex research question, you may use simulation to answer it, based on observed patterns in the ESI dataset.

Other methods may be used (such as regression, logistic regression, etc.), but you will need to justify why you are using them and do them correctly. If you are unsure about the method you are using, please review the course material and ask me if further clarification is needed.

6 Deliverables

6.1 GitHub repository

You will need to create a GitHub repository for your project. The repository should include all the files needed to run the analysis. This includes the data, the scripts, the report, and the presentation. The repository should be public, so that I can access it.

Because you will be working in groups, you will need to decide on a workflow for the repository. I recommend being careful with the way that you use the repository, as it can be easy to overwrite each other's work or run into merge conflicts. For instance, you can assign different parts of the analysis to different group members, and have them work on separate scripts. You can then merge the scripts together at the end in a master file. A more advanced approach would be to use branches for each group member, and then merge the branches together at the end (however, this is not necessary for this project).

A tentative structure for the repository could be:

```
project/
  data/
     raw_data.csv
     cleaned_data.csv
  scripts/
     data_cleaning.R
     analysis.R
  report/
     report.Rmd
     report.pdf
  presentation/
     slides.Rmd
     slides.pdf
  README.md
```

6.2 Small paper

In the project's small paper, you will need to present your research question, your method, your results, and your conclusions. The paper should be written in an academic style, and should include the following sections:

- **Introduction**: Present your research question and why it is important.
- **Data**: Describe the data you are using, and how you are using it.

- **Methods**: Describe the statistical methods you are using to answer your research question.
- **Results**: Present your results, including any tables or figures that you have created.
- **Discussion**: Discuss your results, and what they mean for your research question.
- **Conclusion**: Summarize your findings and what they mean for your research question. (typically this is part of the discussion).

While a literature review is not required, you should acknowledge any sources that you use in your analysis. This includes data sources, code snippets, and any supporting literature. You should provide a bibliography in the last section of the paper, as per a standard bibliography style (e.g. APA, Chicago, etc.).

The paper may be written in any software you desire, but it must be submitted in PDF format to your GitHub repository. Submissions in RMarkdown, Quarto, or LaTeX are encouraged, and may be eligible for extra credit. Ideally, the paper is written in English, please let me know if you require an exception and I will evaluate it.

Further, if you desire to do a literature review, it should review the state-of-the-art in the research question you are asking, and should be written in a way that is understandable to the general public. Papers and findings that relate to yours should be cited and discussed in this review. A literature review will be considered for a good amount of extra credit.

There is no minimum or maximum length for the paper, but it should be long enough to cover all the required sections. I would expect a paper to be at least 5 pages long, but this is just a guideline.

6.3 Slides

You will need to present your findings in a 10-minute academic-style presentation to the class. This requires the use of slides in an academic format and styling. The slides should include the following sections:

- **Introduction**: Present your research question and why it is important.
- **Data**: Describe the data you are using, and how you are using it.
- **Methods**: Describe the statistical methods you are using to answer your research question.
- **Results**: Present your results, including any tables or figures that you have created.
- **Conclusion**: Summarize your findings and what they mean for your research question.

The format is up to you but should be academic in nature (e.g. no flashy transitions, no animations, etc. [1]). The slides should be submitted in PDF format to your GitHub repository. Submissions in RMarkdown, Quarto, or LaTeX are encouraged, and may be eligible for extra credit.

The presentation itself need not be in English, so these slides can be in Spanish if you prefer. However, I would prefer the slides to be in English (eligible for extra credit).

7 Presentation

You will use the submitted slides to present your findings to the class. The presentation will be 10 minutes long, and will be followed by a 5-minute question and answer session. The presentation should be academic in nature, and should present your research question, your method, your results, and your conclusions. I will go all academic on you, so be prepared for tough questions (the best wau to prepare is to know your analysis inside and out, and having it done adequately, hence, the importance of the meetings).

10 minutes may feel like a short time, but it is enough to present your research question, your method, your results, and your conclusions (at least this is the way that conferences work in academia). I don't expect all group members to speak, but if one member does not speak, I will expect them to answer questions and to have contributed to the repository.

I am okay with the presentations being in Spanish, but I would prefer them to be in English (both slides and oral presentation). Presentations in English are eligible for extra credit.

The order of the presentations will be random, and I will let you know the order in advance in the Google Sheet.

8 Software

I strongly suggest using R or Stata for this project. These tools have been taught in the program, and I can commit to providing support for a project that uses these tools.

I can provide support for projects that use Python or Julia, but I cannot guarantee the same level of support as for R or Stata. Other software such as MATLAB, SPSS, SAS, Excel or Minitab are not recommended, as they are not suited for reproducible research. However, if you insist on using one of these, you will need to ensure that you use the right approaches to make a reproducible analysis. For instance, if you where to choose Excel, you would need to provide a VBA script that runs the analysis from start to finish (hence, I do not recommend using Excel).

[1] I am okay with some memes if they are tasteful.

8.1 Reproducibility

When I refer to reproducibility, I mean that someone else than you should be able to reproduce your analysis by following the instructions you provide in your repository. This means that you should provide a README file that explains how to run your analysis, and you should provide a script that runs the analysis from start to finish. Specific things that you should take into account are:

- File paths: You should use relative file paths, so that the analysis can be run from any computer.
- The GitHub repository should be public and include ALL the files needed to run the analysis.
- Plenty of commits with meaningful titles should be present for yourself to not get lost in the analysis and be able to return to a previous state if needed.
- The `README` or documentation files should explain how to run the analysis, and should include a master script that runs the analysis from start to finish.
- While I do not ask you to use Docker or other containerization tools which are the state-of-the-art for reproducibility, you should ensure that the analysis can be run on a fresh computer with minimal setup.
    - This means documenting session information and packages used in the documentation. Do research on how to do this in your software of choice.

9 Meetings

As we do not have enough time to have multiple submissions (e.g. drafts), I will expect you to meet with me at least once to discuss your project. You can schedule a meeting with me through Google Calendar here.

In the meeting, I would expect that you discuss your research question and your method, to evaluate if you are on the right track. If you choose to pick a pre-defined research question, I would still expect you to meet me to discuss your approach, specific method. Other meetings can be scheduled to talk about the progress of the project, code, or any other questions you might have.

10 Academic integrity and source acknowledgment

You should acknowledge all sources that you use in your analysis. This includes data sources, code snippets, and any other sources that you use. You should provide a bibliography in your report. This includes a citation for the dataset.

Any academic dishonesty will be severely penalized. This includes plagiarism, cheating, and any other form of dishonesty. This includes free-riding in group projects. I will be monitoring the repositories and the commits to ensure that everyone is contributing to the project. Any free-rider will receive a failing grade for the project, which means an automatic failing grade for the course.

If you are unsure about what constitutes academic dishonesty, please ask me.

11 Deadlines and grading

- **Group formation deadline**: **October 13th**. This must be done in the Google Sheet here.
- **Slides submission deadline**: **October 20th**. Submit by Google Classroom.
- **Report submission deadline**: **October 27th**. Submit by Google Classroom. Must incorporate feedback from the presentation.
- **Presentations**: Monday, **October 21st**, potentially Tuesday and Wednesday if we have too many groups. Attendance is mandatory.

The project will be graded on the following criteria:

- Quality of analysis: 30%. This includes the quality of the research question, the appropriateness of the method, and the quality of the results, considering the coding and the statistical analysis.
- Quality of the report: 30%. This includes the structure of the paper, and the quality of the figures and tables, and that the paper complies with everything that I set out in the previous sections.
- Quality of the presentation: 20%. This includes the quality of the slides, the quality of the oral presentation, and the ability to answer questions.
- Attended meeting: 10%. This includes whether the group met with me to discuss the project.
- Writing quality: 10%. This includes the quality of the writing in the report and the slides. While grammar and style is important, I will focus on the overall quality of the writing, i.e. whether the writing is clear and concise enough to convey the desired meaning. Further, this also includes the quality of the documentation in the repository. Also, the bibliography should be in a clear format and should be complete.