Course Project: Guidelines for Proposal

MATH 345-545 – Fall 2019

For most research groups, the course project will consist in the statistical analysis of one or several datasets. However, if your group is interested in working on a substantial problem in probability theory or in developing a software package for a given statistical task, feel free to describe such activity in your project proposal. Also, if you are interested in undertaking a long-term research project (2+ semesters), please let me know and I can help you organize the MATH 345-545 project as the first part of a longer project.

What to write in the project proposal?

1. **Names of the group members.**
2. **Research subject:** provide some background information and give a sense of why this topic is interesting/relevant.
3. **Data description:**
    - Source (URL for data found on the web),
    - Collection date (at least approximately),
    - Collection method (survey, controlled experiment, observational), sampling method if applicable (telephone, internet, in person; random sample, convenience sample, stratified sample…)
    - Number of statistical units in data set (= sample size)
    - Study variables: name and type (numerical + discrete/continuous; categorical, ordinal…)
    - Discuss missing data is necessary (proportion, causes, potential impact on data analyses…).
4. **Research questions:** At least 4 questions, at most 7-8.
5. **Data:** upload the data you will analyze on Blackboard. If the data are too large, simply provide a web link to the data.

Where to find data for your project?

- You are welcome to use data from your work or from academic research you may be engaged in.
- If some faculty member on campus works in an research area you are interested in, you can (politely!) ask them if they could provide you with data and maybe point you to a couple basic questions that you could investigate in your course project.
- There are millions of datasets on the internet. For example, try a google search “astronomy data” and see what comes up. Examples of websites are given below.

Choosing a data set (or several)

- You will be working for 5 weeks or so on the course project, so you should **select a data set you are really interested in!**
- Your research work and analyses will have to be personal and original. For this reason, avoid data that have already been widely analyzed (unless you have an original research question that has not been addressed before). In particular, *avoid textbook data*.
- Use recent data (less than 20 years old!).
- As much as possible, *avoid time series data* (i.e. repeated measurements over time like the daily temperature in a location over a year). The statistical methods learned in this course mostly pertain to independent measurements and may not be adequate for data that display temporal dependence.

Defining research questions

- Your research questions should be genuinely interesting to you and they should be meaningful to the topic you are studying. If you are not exactly sure of how to address these questions from a statistical perspective, I can help.
- In terms of statistical methods, the course covers descriptive statistics (numerical + graphical) and probability modeling, mostly. You will also learn about quantifying the association between variables through notions of covariance and correlation. Typically, your project will mostly revolve about descriptive statistics but also include 1-2 questions where you study (and try to model) the probability distribution of a variable.
- If your group has research questions that go beyond the scope of this course, I can briefly explain how to use other statistical methods (for example linear regression or chi square tests for categorical data) to handle these questions.
- The goal of the project proposal is to make sure that the data and research questions selected by your group are meaningful and suitable for statistical analysis. That is, that you will be able to apply (several of) the statistical methods learned in class to address your group's research questions.
- As your course project progresses, you will be free to add, modify, or delete research questions according to the project's needs.

Examples of data sources

Among 1,000's of possibilities, in no particular order:

- World/U.S.: economy, energy, environment, demographics, employment, education, social inequalities, science…
  http://data.gov
  http://data.worldbank.org
  http://census.gov
  http://www.pewresearch.org
  http://www.datacenter.org/research-tools/web-resources
- U.S. Elections
  http://www.electoral-vote.com
- Data Challenge Expo 2019: New York City Housing and Vacancy Survey
  https://www.census.gov/programs-surveys/nychvs.html
- City of Boston
  https://data.cityofboston.gov
- UMass Boston
  https://www.umb.edu/oirap/common_data_set
- Machine learning & Data Science
  http://archive.ics.uci.edu/ml/
  https://www.kaggle.com/datasets
  http://rs.io/100-interesting-data-sets-for-statistics/
  https://www.springboard.com/blog/free-public-data-sets-data-science-project/
- Sports
  www.nba.com
  http://www.baseball-reference.com
- Wikipedia usage
  https://dumps.wikimedia.org/other/analytics/
- U.S. Movies
  http://www.the-numbers.com
  http://www.imdb.com
- Crime in the U.S.
  http://www.ucrdatatool.gov

An example of project proposal

Group members

Joe W, Samir X, Rita Y, and Wenli Z.

Topic

Rio 2016 Olympic Games.

The Rio 2016 Games were very exciting (and even nail-biting!) to watch on more than one count. For example, they were the first Olympic Games ever to be held in South America and only the third to be held in a developing country. (The 1968 games took place in Mexico City and the 1988 games in Seoul, South Korea.) In addition, the lead-up to the Rio Games was plagued by controversies, including the instability of Brazil's federal government; the country's economic crisis; health and safety concerns surrounding the Zika virus and significant pollution in the Guanabara Bay; and a doping scandal involving Russia, which affected the participation of its athletes in the Games.

In this project we will analyze the performances of the participating countries in the sports represented at the Rio 2016 Games; we will also investigate the potential adverse effects of the Zika virus during these Games. In addition, we will touch on broader aspects of the history of the Olympic games such as countries' performances over time and organization costs.

Data sources

https://www.rio2016.com
https://www.rio2016.com/en/medal-count-country
https://www.rio2016.com/en/medal-count-athletes
https://www.rio2016.com/en/medal-count-sports
https://en.wikipedia.org/wiki/2012_Summer_Olympics
https://arxiv.org/pdf/1607.04484v1.pdf

Data description

There are three data sources. The main source is the official website of the Rio 2016 Olympic Games, rio2016.com. The second source, Wikipedia, provides results from the 2012 summer Olympics and will be used for comparison. For these two sources, the data can be viewed as censuses: the entire event was observed (all results of all competitions in all sports) without error. The third source is a 2016 Oxford study about the costs of the Olympic games since 1960.

For the first and second data sources, the main study variable is the medal count. This is a numerical, discrete variable. The associated statistical units are countries (n=207). For the Oxford study, the variable of interest is the cost of the Olympic games in 2015 dollars (numerical, continuous). Here, the statistical units are all Olympic games since 1960 (n=30, 15 in summer and 15 in winter). Some data are missing and, more importantly, there is substantial uncertainty in the reported costs.

Research questions

1. Which country won the most medals in total? The most gold medals? How were medals distributed across countries? How many new world records were set? How many athletes were infected by the Zika virus? *(Descriptive statistics)*
2. Can the number of medals by country be modeled with a commonly used probability distribution? Same question for the organization cost. *(Modeling)*
3. Given the uncertainty in reporting organization costs, we want to build confidence intervals for quantities such as the average, the median, and the variance of the cost. *(Estimation and inference)*
4. Looking at numbers of gold medals (or just medals), can the 2016 performance of a country be predicted by its 2012 performance? (*Regression and correlation analysis*)
5. A more general question (requires additional data): is it true that during Olympic games, the organizing country has an advantage, that is, wins more medals than usual?

Data

[Attach Excel files on Blackboard.]