MAT 222 — Fall 2019
Project Guidelines

Overview

The objective of this project is to put the material you have learned in MAT 221 and MAT 222 into practice by examining *real-world data*. You will investigate questions of your choosing based on data collected from Chicago public schools. The data will be analyzed using Minitab (or a standard statistical package of your choice, e.g. R, SAS, SPSS, or STATA). You will then summarize the questions, data, analyses, and conclusions in a written report. The project will follow the following schedule:

Project Deadlines:

Initial Draft Due: Thursday November 21
Final Report Due: Thursday December 5 (Last Day of Class)

Data Analysis

During this stage, you analyze your data by using all the statistics you have learned in MAT 221 and MAT 222. Typically, these analyses will consist of estimates of a parameter, i.e. a confidence interval and its interpretation, or a hypothesis test of some parameter/difference along with its interpretation. Be sure to explain your methods clearly so that a lay person or a person with very little statistical knowledge could follow what is going on in the paper. Furthermore, be sure to check the assumptions required for any statistical computation you perform. Beyond this, analyze other possible sources of error in the project. You may – and should – use visual representations of data wherever appropriate to explain data, analyses, or check assumptions.

Following the analysis, you need to write a report consisting of the following: an introduction introducing the study, dataset, and the questions you will analyze, a discussion of how the data was collected and a summary of the variables (along with accompanying charts and tables), statistical analyses with explanation and interpretation of the results, along with an error discussion, a conclusion summarizing the results, and a bibliography of cited sources. You must cite any and all sources of the dataset, including the original study and where you obtained your data, and you must cite the course textbook. *Plagiarism of any kind is unacceptable and will result in a 0 for the project as well as academic consequences.* Be sure to include any statistical package outputs. Only data required for the statistical results or your discussion should be included. When appearing 'in-line', they should be centered, have proper size, and have proper spacing. Similarly, graphs and plots should be large enough to be readable but not so large that they simply 'kill space.' Mathematical symbols should be typeset in equation style in text and not using the normal font that other text appears. Equations should be numbered and the equation referenced when discussed. Citations should also appear in-text.

A rough draft of the project will be due on Thursday November 21st. The project should be as complete as possible as this will be your last chance for detailed feedback on what you have written, in addition to the fact that you have very little time to edit the paper based on these comments

before the final due date. Do not worry if your paper receives a lot of feedback! This is to help you ensure that your final project will be as good as possible, ensuring the best possible grade on the project. Moreover, this is most likely your first time writing an extensive mathematical paper, so there will naturally be a lot to be desired in a draft! The more you turn in for this draft, the more feedback you can receive, leaving you time to make the changes and be done with the project before the due date. Be aware, the Writing Center can help you with your phrasing, paper layout issues, and citations. Make use of all the free writing help the university offers you!

Final Report

The final report is due Thursday December 5th. This is the last day of class. You must turn in a paper copy of the paper by the end of this class. Late submissions and email submissions will not be accepted. You will also submit the paper via a portion on Blackboard for checking using the plagiarism detection and prevention system Turnitin. You should submit your papers to Turnitin, via an upload portion on Blackboard, to check that all sources you use have been properly acknowledged and cited. Turnitin compares submitted documents against documents on the Internet and against student papers submitted to Turnitin at Syracuse University and at other colleges and universities. Your knowledge of the subject matter of this course and your writing level and style are taken into account in interpreting the originality report. Keep in mind that all papers you submit for this class will become part of the Turnitin.com reference database solely for the purpose of detecting plagiarism of such papers.

When turning in this paper electronically on Blackboard, it must be submitted as a PDF. Many word processors have the option of saving or exporting the document into PDF form. You should do this to submit the paper. You can upload the paper and continue to change the uploaded file until the due date. But review the final submission file thoroughly before uploading. It is your responsibility to check the file, both for content and to be sure that it opens properly. Be especially careful if you have been using/saving multiple versions of the paper with the same name that you have uploaded the correct file. Do not close your browser until the uploading is complete. Corrupted files will not be graded, so once the file is uploaded, be especially sure to check that it has uploaded correctly.

Dataset/Analysis Requirements:

(i) Requirements for the dataset are it must have:

  (a) *At least* three quantitative variables;

  (b) *At least* two categorical variables; and

  (c) *At least* 30 observations.

(ii) Analyses must include:

  (a) Graphical summaries of *all* important variables analyzed, including side-by side (when appropriate) boxplots of each quantitative variable analyzed, and when appropriate visualizations of categorical variables,

  (b) Numerical summaries of *all* quantitative variables analyzed, including a 5-number summary, mean, and standard deviation reported.

(c) A 1-sample t-test/confidence interval. The hypothesis test may be constructed arbitrarily (within reason). An analysis of Type I error, Type II error, and power for this test must be discussed.

(d) At least one 2-sample t-procedure or 2-sample p-procedure.

(e) Multiple linear regression model;

(f) *at least one* other analysis methods of your choice, e.g. Chi-Square test of independence, one-way ANOVA, two-way ANOVA, etc. Ideally, you should have around five total statistical analyses.

(g) An error analysis for the statistical analyses.

Report Requirements:

(i) The report must be understandable to a non-statistician/lay-person or a person with very little statistical knowledge.

(ii) The report must have proper introduction and conclusion sections.

(iii) Datasets or any other information obtained from external sources must be properly referenced. You can see the Writing Center for help with this. [You must at least reference the source(s) of your dataset and the course textbook.]

(iv) The report should follow the standard conventions in scientific writing. Moreover, the report should follow general good writing standards, i.e. proper grammar and spelling.

(v) Do not discuss the analyses as "I did a regression. . . ", "I then plotted the box-plot . . . ", etc. Use instead phrases such as "a regression was conducted. . . ", "the generated box plot shows. . . "

(vi) For each statistical analysis, the Minitab output (graphical or session window ouput) must be included and must be clearly explained within the body of the report. However, take note of formatting issues discussed below.

(vii) Use font size 11 with a 'standard' font. There should be 1 in margins, except perhaps at the top/bottom of the page where 1.5 in is fine.

(viii) Graphical or numerical outputs taken from Minitab must be formatted cleanly (e.g., resize the graphs such that an entire page is not a graph, indent or box numerical outputs). Graphs/numerical results should come with appropriate labels and captions. You should include only portions relevant to your analyses and/or discussion. Table labels precede the tables and are labeled without punctuation, whereas figures are labeled after the figure and are labeled with punctuation.

The following documents will help you get started on understanding what goes into creating good scientific/mathematical exposition. See also the sample project provided by the instructor. Note that this sample *does not* meet the project guidelines. The sample is to demonstrate good writing for this project, clean typesetting and formatting, and a good statistical analysis of the dataset.

Guide to Scientific Writing: https://ugr.ue.ucsc.edu/sites/default/files/jyi_guide_to_scientific_writing.pdf (especially pages 11–16 & 17–38 for additional specifics)

Guide to Manuscript Writing: http://www.ugresearchsci.ucla.edu/docs/guide%20to%20science%20writing.doc

Should you opt to choose your own project area, you alone will be responsible for finding an appropriate dataset. While this option is not recommended, it is possible to have a successful project. Take your time and explore many different datasets from many different sources. You should consider the following criterion when selecting your project:

- Interest: The dataset should interest you! You will spend a lot of time working with this data. If the data is centered around a topic you do not care about, the process will be painful. This should be fun (or as fun as it can be). The goal is to approximate what it is like to be on the statistical calculation end of a real world study. You should get a sense of satisfaction finally using all you have learned examining real hypotheses using a real dataset. If you do not get a sense of excitement at the possibilities while looking at the dataset, it is probably not the one for you.
- Questions: The dataset should contain enough variables and data for you to ask interesting questions. Even if the dataset looks interesting, if there is not 'enough' in the dataset to ask lots of interesting questions, then the project and report probably will not turn out the way you had hoped.
- Understandability: The world is complicated, so it comes as no surprise that many datasets are complicated. The variables or data inside a dataset can be hard to understand. Even an interesting dataset can prove to be a nightmare if you deeply struggle trying to figure out what the variables are or what the numbers/categories represent. Pick a dataset that you can 'get your head around'. However, even 'simple' datasets can seem daunting at first. So if a dataset seems interesting, spend a bit of time reading about the data/study and staring at the data. Give the dataset a fair chance! You can also ask for help! But if, after spending a bit of time and effort, you still do not understand the dataset or cannot think of lots of questions based on the data, the dataset is probably not for you.

These are not the only considerations! Be sure your dataset meets the standards and guidelines for the project! If you hand in a project based on a dataset that does not follow these standards, *you will receive a 0 for the project!*

Once you choose a dataset, you will need to formulate some questions you will investigate using this dataset. Specifically, determine what population parameters you want to estimate or what hypotheses you want to test. You should think of around ten, *but at least five*, questions you will investigate. At this stage, *before performing any analyses*, you will need to determine the significance level you will use for your inferences and hypotheses tests. Remember, low α helps reduce chances of a Type I error but can increase your chance of a Type II error. Think carefully about the α-level you will choose.

You will then summarize the your chosen dataset, your proposed questions, and your proposed statistical analyses, among other things, to your instructor for approval. Appropriate feedback on

the project will be given, and the project will either be approved or sent back for edits, or even possibly rejected outright. *Do not wait to ask questions and be sure your dataset will work for this project!*

Suggested Data Sources:

- UC Irvine Machine Learning Repository: http://archive.ics.uci.edu/ml/index.php
  - A few examples of the datasets which can be found there are: Housing, Liver Disorders, Annealing, Automobiles, Echocardiograms, Labor Relations, Statlog, Concrete Compressive Strength, Wine Quality, Forest Fires, Breast Cancer Wisconsin, Pima Indians Diabetes, Energy Efficiency, Glass Identification, Breast Tissue, Leaves
- StatLib–Datasets Archive: http://lib.stat.cmu.edu/datasets/
- StatSci.org Data Sets: http://www.statsci.org/datasets.html
- United Nations Statistics Database: https://unstats.un.org/unsd/databases.htm
- Economic and Social Development: https://unstats.un.org/unsd/demographic/products/socind/default.htm
- National Statistics Office: https://unstats.un.org/home/nso_sites/
- United Nations Statistics Division: https://unstats.un.org/home/international_agencies/
- University of Florida Datasets: http://users.stat.ufl.edu/~winner/datasets.html
- Wolfram Data Repository: https://datarepository.wolframcloud.com/
- IRS Tax Statistics: https://www.irs.gov/statistics
- Eurostat: https://ec.europa.eu/eurostat/data/database
- Vanderbilt University: http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets
- CDC: https://www.cdc.gov/DataStatistics/
- National Heart, Lung, and Blood Institute: https://biolincc.nhlbi.nih.gov/studies/
- Microsoft R ApplicatioN Network: https://mran.revolutionanalytics.com/documents/data
- Dryad: https://datadryad.org/discover?query=&submit=Search#advanced

You can also find larger lists at this website https://github.com/awesomedata/awesome-public-datasets and at this website https://www.dataquest.io/blog/free-datasets-for-projects/. If your dataset is not from one of the sites listed above, both the website and the dataset must be approved by the instructor. This is to ensure both the dataset and the website are reliable, and that they will be appropriate for the project. Moreover, datasets from the above list will not necessarily be approved. But choosing a dataset from the above list will increase the odds.