MATH 23c FINAL PROJECT

Last modified: March 10, 2019

Project Summary - This is an open-ended project wherein students apply methods from the course to real-world data. The project is intended to be completed in groups of two, although working alone or in groups of three is permitted. Students are expected to find a dataset (either from work, research, or other outside interests), import that data into R, and use some of the methods learned in the course to explore the data and test the existence of relationships between the features.

The points below are broken into required and additional points. To achieve full credit, you must do all of the required points, and a subset of 10 of the additional points. To facilitate grading, leave an index/comment in your long script identifying where/how you think you earned your points within your analysis.

Required dataset standards

1. A dataframe.
2. At least two categorical or logical columns.
3. At least two numeric columns.
4. At least 20 rows, preferably more, but real-world data may be limited.

Required graphical displays (all graphs must be colored and nicely labeled)

1. A barplot.
2. A histogram.
3. A probability density graph overlaid on a histogram.
4. A contingency table.

Required analysis

1. A permutation test.
2. A p-value or other statistic based on a distribution function.
3. Analysis of a contingency table.
4. Comparison of analysis by classical methods (chi-square, CLT) and simulation methods.

Required submission uploads

1. A .csv file with the dataset
2. A long, well-commented script that loads the dataset, explores it, and does all the analysis.
3. A shorter .Rmd with compiled .pdf or .html file that presents highlights in ten minutes.
4. A one-page handout that explains the dataset and summarizes the analysis.

Additional points for creativity or complexity – You may attempt as many as you like, for a maximum possible of 10 points

1. A data set with lots of columns, allowing comparison of many different variables.
2. A data set that is so large that it can be used as a population from which samples are taken.
3. A one-page document that discusses ethical issues related to collection of the dataset.
4. A one-page document that discusses ethical issues raised by conclusions reached from analysis of the data.
5. A graphical display that is different from those in the textbook or in the class scripts.
6. Appropriate use of R functions for a probability distribution other than binomial, normal, or chi-square.
7. Appropriate use of integration to calculate a significant result.
8. A convincing demonstration of a relationship that might not have been statistically significant but that turns out to be so.

9. A convincing demonstration of a relationship that might have been statistically significant but that turns out not to be so.

10. Professional-looking software engineering (e.g defining and using your own functions).

11. Nicely labeled graphics using ggplot, with good use of color, line styles, etc., that tell a convincing story.

12. An example where permutation tests or other computational techniques clearly work better than classical methods.

13. Appropriate use of novel statistics (e.g. trimmed mean, maximum or minimum, skewness, ratios).

14. Use of linear regression.

15. Calculation and display of a logistic regression curve.

16. Appropriate use of covariance or correlation.

17. Use of theoretical knowledge of chi-square, gamma, or beta distributions.

18. Use of theoretical knowledge of sampling distributions.

19. A graphical display that is different from those in the class scripts.

20. Calculation of a confidence interval.

21. Appropriate use of quantiles to compare distributions.

22. Team consists of exactly two members (otherwise, 1 or 3 is a possibility).

23. A video of the short script is posted on YouTube and a link to it is left in your long script.

Subjective impression – if these folks were applying for a job that requires computerized statistical analysis, I would

1. Immediately disband the search committee and hire them. (4 points)

2. Add them to a short list of leading candidates. (3 points)

3. Regard them as well qualified, but not among the best candidates. (2 points)

4. View them as acceptable, but not be in a rush to hire. (1 point)