Project Outline

Sophie Berube

8/31/2017

Option 4: Perform an analysis of the statistical analyses in all published PLoS papers. What are the most common techniques? How to they vary by field? Are there any trends over the last 10-15 years?

Ideal dataset: the ideal dataset in this case would be all published PLoS papers since the publication of the first journal, PLOS Biology in October of 2003. Since currently PLOS publishes 7 journals, an ideal data set would include all papers published in each of these journals since the first issue of each journal.

If papers have been retracted those shouldn't be incorporated in the dataset. For each paper, the title, date of publication, PLoS journal (which would indicate the field of study the article relates to) and authors should be noted as well as the actual text of the paper and perhaps if possible the figures in the papers. This will involve loading text files from the web into R, and possibly figures.

Narrowing the Dataset: Since this project will aim to look at statistical analyses performed, we are probably interested in the Materials and Methods section of each article, which usually outlines what kind of analyses were performed for the paper and more specifically if there is a subsection of Materials and Methods with "Statistics" or a related word in the title this would be a subsection of interest.

How to observe trends from the dataset: Since we are looking for trends in types of data analysis it would probably be useful to look for certain key phrases like "linear regression model", "ANOVA", "log linear model". Part of the analysis will involve identifying possible types of statistical analyses that could be carried out by researchers. In observing these trends, it might be useful to also analyze plots or graphs in the results section of the paper.

Once key words are extracted from these articles, it might be possible to see what covariates might be related to certain types of analysis, for instance it would be possible to monitor how common ANOVA is in certain years, if the frequency of ANOVA use changes over time, or if ANOVA is more commonly found in Biology articles than infectious disease articles.