# Class Project: Analysis of PLOS Analyses

*Bonnie Smith*

*September 11, 2017*

**Question:**

In this project, we perform an analysis of the statistical techniques used in published Public Library of Science (PLoS) papers. We identify the most common techniques and whether they vary by field, as well as trends over time.

**Background about PLoS:**

The fields represented in PLOS journals are Biology (since October 2003), Computational Biology (since June 2005), Neglected Tropical Disease (since October 2007), Medicine (since October 2004), Genetics (since July 2005), and Pathogens (since September 2005).

Statistical methods used are discussed in a "Methods", "Materials and Methods", or "Materials and Models" section. (Note: there are other names as well—what is the best way to find them all?)

**Data Collection:**

We will begin by forming a list of common statistical methods used by authors publishing in these journals: Using the `tm` package, we will text mine the appropriate sections of a sample of papers from each of the 6 journals in order to help form this list. (More details about how to do this.)

We will then use the `rplos` package to scrape the PLOS website. For each paper, we will record the journal/field, the month/year of the issue it appeared in, and an indicator for each statistical technique from our list as to whether it was used in this paper.

As an example, if we are searching only for papers that use ANOVA, we have the following code:

```
library(rplos)

Anova_journals=searchplos(q="materials_and_methods:ANOVA",
        fl=c("title","journal","publication_date"),limit=8)
Anova_journals
```

```
## $meta
##   numFound start maxScore
## 1    43352     0       NA
##
## $data
##    journal     publication_date
## 1 PLoS ONE 2014-02-27T00:00:00Z
## 2 PLoS ONE 2012-06-22T00:00:00Z
## 3 PLOS ONE 2015-11-06T00:00:00Z
## 4     none 2012-08-14T00:00:00Z
## 5 PLOS ONE 2015-03-16T00:00:00Z
## 6 PLoS ONE 2014-06-03T00:00:00Z
## 7 PLoS ONE 2008-04-02T00:00:00Z
## 8 PLoS ONE 2014-02-28T00:00:00Z
##
## 1                                                                  Occlusal Enamel Complexity in Middl
## 2                                                               Recovery of Barotrauma Injuries in Chin
## 3 Phenotypic Buffering in a Monogenean: Canalization and Developmental Stability in Shape and Size o
## 4                           Acute Peripheral but Not Central Administration of Olanzapine Induces Hyp
```

```
## 5                                                                                                  La
## 6                                  Variation in Seed Germination of 134 Common Species on the Eastern
## 7                                                                  Appetite Enhancement and Weight
## 8                                                                                Complexity
```

As a check that our list of candidate methods was relatively complete, for any paper that matched zero of the methods, we will inspect the paper by hand, augmenting our list as needed based on the results.

**Analysis:**

Once we have the cleaned data, for each statistical technique on our list we will model the proportion of PLOS papers to use that technique, adjusting for time and field.

**Conclusions:**

Our analysis shows that there is variation by field. ...