# Statistical techniques used for work published in the Public Library of Science

Class project for Advanced Data Science I

*Bonnie Smith*

*October 11, 2017*

**Introduction**

The Public Library of Science (PLoS) is a collection of open-access journals representing a variety of sub-fields with connections to biology. The discipline-specific journals in the collection are PLoS Biology, PLoS Medicine, PLoS Computational Biology, PLoS Genetics, PLoS Pathogens, and PLoS Neglected Tropical Diseases. The inception of these six journals ranges from October 2003 (PLos Biology) to October 2007 (PLoS Neglected Tropical Diseases). Note that there is also a catch-all journal, PLoS ONE, which we shall ignore for the purpose of this analysis.

In this project, we examine the statistical techniques used in research articles published in the six discipline-specific PLoS journals. We identify common techniques, and whether they vary by journal and year of publication date. Specifically, we address two questions of interest:

1. Out of a list of statistical techniques used in PLoS articles (which we formulate as discussed below), identify the 5 most widely used in each discipline in 2016, and the 5 most widely used in each disipline in 2010.
2. Group a subset of statistical techniques into 4 groups by theme, and examine the popularity of each group in each year from 2010 to 2016. Identify disciplines where the change over time is greatest and least.

**Methods**

Using the `rplos` package, which allows us to interface with the PLoS API directly in R, we search the library of PLoS papers for certain keywords, comprising our statistical techniques of interest. Identification of this list of keywords is discussed below. We restrict our search to full-length 'research articles' only (as opposed to 'viewpoints', 'correspondence', etc.); the following table gives the number of full-length research articles that have appeared in each journal to date.

Table 1: Total number of research articles that have appeared in each PLoS journal

| PLos Bio | PLoS Medicine | PLoS Comp. Bio. | PLoS Genetics | PLoS Pathogens | PLoS Trop. Diseases |
|---|---|---|---|---|---|
| 2366 | 1375 | 4731 | 6196 | 5271 | 4761 |

We further restrict our search to the "Materials and Methods" section of the articles, and stratify the results both by journal and by year of publication.

In order to form our list of keywords to search, we take a random sample of 100 PLoS articles, which is representative of all PLoS research articles in terms of journal and time period. We read in the 'Materials and Methods' sections of the articles in our sample, check each by hand, and add any statistical methods mentioned to our list of keywords. This list has been refined slightly and roughly organized into categories below.

**List of Keywords:**

- **ANOVA**. **Chi-square**. **Fisher's exact test**. **T test**.
- **Kaplan-Meier**. **Cox** proportional hazard/Cox regression.
- **Wilcoxon** rank-sum test/Mann-Whitney U-test. **Nonparametric**.
- **Cross validation**. **Machine learning**. **Lasso** regression. Lowess. Kernel density. Penalized spline regression.
- K-means **clustering**/ hierarchical clustering. Singular value decomposition/**principal components**.
- **Linear regression**. **Logistic regression**. Generalized linear models /Poisson regression. Likelihood ratio test. Mixed-effect model/**random-effect model**/linear mixed model. **Generalized estimating equations**.
- **Simulations**. **Bootstrap**.
- **Markov chain Monte Carlo**. Multiple imputation. Hidden Markov model. **Empirical Bayes**. **Bayesian**.

In order to examine trends across journal and accross years, for each keyword, for each of 4 journals, we compute the proportion of the journal's research articles from 2009 and 2010 which contain that keyword. We then do the same for 2016-2017. Results are displayed below.

Table 2: Proportion of a journal's research articles conta
keyword, in 2009-2010 vs. 2016-2017

| | Biology, 2009-10 | Biology, 2016-17 | Comp. Bio., 2009-2010 | Comp. Bio., 2016-17 | Ge |
|---|---|---|---|---|---|
| ANOVA | 0.169 | 0.269 | 0.027 | 0.046 | |
| t test | 0.614 | 0.756 | 0.482 | 0.531 | |
| chi square | 0.029 | 0.063 | 0.017 | 0.018 | |
| Fisher exact test | 0.029 | 0.026 | 0.027 | 0.017 | |
| Wilcoxon | 0.042 | 0.085 | 0.027 | 0.028 | |
| linear regression | 0.050 | 0.089 | 0.043 | 0.062 | |
| logistic regression | 0.016 | 0.030 | 0.010 | 0.029 | |
| Kaplan-Meier | 0.024 | 0.015 | 0.000 | 0.005 | |
| Cox | 0.008 | 0.022 | 0.000 | 0.011 | |
| generalized estimating equation | 0.000 | 0.004 | 0.000 | 0.000 | |
| random effect model | 0.003 | 0.000 | 0.004 | 0.002 | |
| cross validation | 0.019 | 0.055 | 0.074 | 0.100 | |
| machine learning | 0.003 | 0.004 | 0.032 | 0.043 | |
| lasso | 0.008 | 0.004 | 0.003 | 0.030 | |
| clustering | 0.151 | 0.262 | 0.261 | 0.238 | |
| principal components | 0.029 | 0.018 | 0.052 | 0.057 | |
| simulation | 0.143 | 0.210 | 0.553 | 0.633 | |
| bootstrap | 0.058 | 0.092 | 0.065 | 0.072 | |
| Markov chain Monte Carlo | 0.005 | 0.018 | 0.014 | 0.030 | |
| empirical Bayes | 0.003 | 0.004 | 0.004 | 0.006 | |
| Bayesian | 0.048 | 0.063 | 0.082 | 0.109 | |
| nonparametric | 0.032 | 0.063 | 0.017 | 0.014 | |
| (any word) | 1.000 | 1.000 | 1.000 | 1.000 | |