How similar are applied disciplines in their usage of statistical techniques?

A study of research articles published in the Public Library of Science

Bonnie Smith

October 24, 2017

1. Introduction

As the developers and implementers of statistical methodology, statisticians clearly have a vested interest in the adoption of these methods by applied scientific researchers. As a measure of progress in this direction, it is important to know which statistical techniques are widely used, and by whom. The answers have the potential to suggest directions in which new methodology would be helpful, as well as to shed light on possible shortcomings of communication or promotion on the part of the statistics community. If certain disciplines tend to use more (or more advanced) statistical methods, this information would also be of interest to statisticians seeking positive collaborative experiences. Biostatistics as a community has especially strong ties to biology, medicine, and genetics. Thus for these fields in particular, we would like to know which techniques are being used by which disciplines, and also to be cognisant of any trends over time that exist.

In this paper, we investigate the following questions: Among subfields of biology and medicine, what differences exist in terms of which statistical techniques are widely used? Can we see a change in journals' usage of statistical techniques over the course of several years? We address these question as they apply specifically to research articles from the Public Library of Science[1]. The Public Library of Science (PLoS) is a collection of open-access journals featuring scientific research from a number of disciplines connected to biology and medicine. Because the full text of every article appearing in a PLoS journal is freely available online, the PLoS collection provides an accessible dataset for study.

We find that a number of differences across disciplines do exist. In particular, Computational Biology is least similar to the the other disciplines studied, with Medicine also standing out in a number of respects. There is not a dramatic change in usage of statistical techniques over the course of the last 8 years, though here again differences by field do exist. Biology shows the greatest overall increase in statistical usage, while Neglected Tropical Diseases has shown the least change in its patterns of usage.

2. Data

2.1 Data source. All of our data were collected from the Public Library of Science website. There are six discipline-specific journals in the PLoS umbrella: PLoS Biology, PLoS Medicine, PLoS Computational Biology, PLoS Genetics, PLoS Pathogens, and PLoS Neglected Tropical Diseases. There is also an omnibus journal, PLoS ONE. However, we have chosen not to consider PLoS ONE for the purpose of this analysis, as our main goal will be to make comparisons of usage of statistical techniques across specific disciplines. The Public Library of Science journals were all founded relatively recently: the initial publication date of the six journals ranges from October 2003 (PLos Biology) to October 2007 (PLoS Neglected Tropical Diseases). Thus in examining change over time, we are constrained to a somewhat narrow timespan.

Throughout, we restrict our attention to full-length research articles only, as opposed to articles appearing in the 'viewpoints', 'correspondence', or other section of a journal. The total number of full-length research articles that have appeared in each of our six journals of interest to date is shown in Table 1 below.

Note that throughout, we are working with the full population of PLoS research articles. Thus we give descriptive statistics only, as there is no opportunity for inference.

2.2 Missingness. Some research articles do not contain a Materials and Methods section, and thus are missing information as to what statistical techniques, if any, were used. We do not remove these cases, but rather impute a value of zero for each technique; that is, we assume that the missingness is actually due to the fact that no statistical method was used.

The overall percent of missingness for each journal is shown in Table 1 below. Note that PLoS Computational Biology has a much higher rate of missingness than any other journal. Thus our findings for PLoS Computational Biology have the potential to be especially sensitive to the correctness of our missingness assumption.

Table 1: **Counts of research articles, and percent missingness.** For each journal, the total number of research articles, and the total number of research articles that contain a Materials and Methods section are shown, as well as the percent of that journal's research articles that are missing a Materials and Methods section.

	Biology	Medicine	Comp. Bio.	Genetics	Pathogens	N.T.D.
Total # of articles	2367	1381	4747	6205	5280	4780
# articles with a methods section	2344	1370	4377	6165	5266	4746
% articles with no methods section	.97	.8	7.79	.64	.27	.71

3. Methods

3.1 Determining keywords of interest. We first construct a list comprising those statistical techniques that will be included in our study. In order to form this list, we take a sample of 72 articles from our restricted dataset (consisting of only those articles that have a Materials and Methods section), with 12 articles randomly selected from each of the 6 journals. We read through the Materials and Methods sections of each of these 72 articles, and make note of any statistical methods mentioned. Our findings are cataloged in the Supplementary Materials, Section 1. We then distill this catalog into a list of 20 statistical techniques of interest.

In our analysis we equate usage of a statistical technique with mention of the technique in the text of the Materials and Methods section of the article. Given that we have chosen this approach, we should look for means of assessing the false discovery rate—how common is it for authors to mention a statistical technique that they are *not* using? In reading through the Materials and Methods sections in our 72-article sample, we observed 121 instances of a statistical technique being mentioned. Of these, only 3 were instances where the statistical technique named was *not* being used in the analysis (rather, the authors were justifying their choice not to use that particular technique). Based on this sample, we estimate the false discovery rate at only around 2.5%.

3.2 Differences across journals For each keyword-journal pair, we compute the number and proportion of that journal's research articles that contain the keyword. There were 5 techniques—Benjamini-Hochberg correction, k-means clustering, generalized linear models, linear mixed models, and ROC curve—that were used in less than 5% of articles for each journal, which were subsequently omitted from further analysis. For each of the remaining 15 keywords, we compute the standard deviation in proportions of each journal's articles that used that keyword. This allows us to see which techniques correspond to a greater variety in amount of usage by discipline. As a summary measure of how different a given pair of journals are in their overall usage of techniques, we also compute the square of the Euclidean distance between the vector of proportions for the two journals.

3.3 Difference over time In order to address our question as to changes over time, we focus on articles from selected dates, comparing a group of older articles versus a group of more recent articles. Because the youngest PLoS journal began publication in 2007, we take as our earlier group of articles those whose year of publication was 2009 or 2010, and articles published in 2016 or 2017 as our more recent group. For each

journal, we select the 7-8 most popular statistical techniques from our list of keywords, and compute the proportion of the journal's articles from 2009-10 that used the technique, and the proportion of the journal's articles from 2016-17 that used the technique.

4. Results

4.1 Comparative usage of statistical techniques by journal. Comparative usage by journal for each of the most widely used statistical techniques that we studied is shown in Figure 1. The table showing proportions for all 20 techniques, as well as the raw counts, may be found in the Supplementary Materials, Section 2.

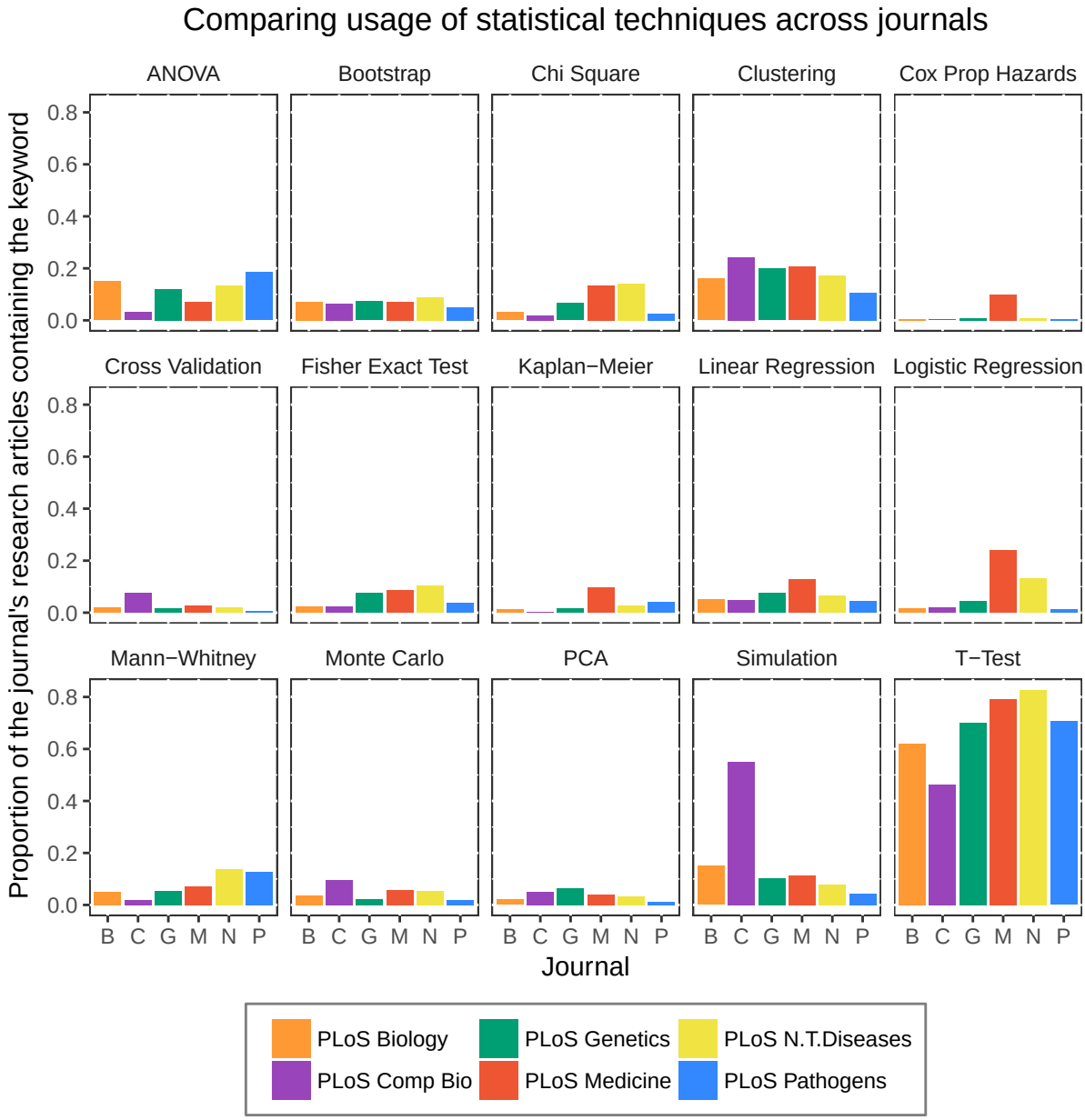

Figure 1: Popularity of certain statistical techniques varies considerably by journal. Each bar shows the proportion of all research articles appearing in the given journal that contain the given keyword.

Note especially how Computational Biology stands out from the other journals in its usage of clustering,

cross validation, simulation, and t-tests. Medicine stands out in its usage of Cox proportional hazards, Kaplan-Meier, and logistic regression. Techniques where we see a lot of variability in usage across journals include simulation, t-test, and logistic regression. Standard deviations for each of the techniques are shown in Table 2.

Table 2: **Standard deviation in journal proportions for each keyword.** For each keyword, we compute the standard deviation in the 6 values of proportionate usage by journal.

ANOVA	Bootstrap	Chi square	Clustering	Cox prop.	Cross val.	Fisher exact	Kaplan-Meier
0.055	0.012	0.055	0.047	0.038	0.024	0.035	0.034

Linear regr.	Logistic regr.	Mann-Whitney	Monte Carlo	PCA	Simulation	T test
0.031	0.091	0.047	0.028	0.019	0.188	0.131

The difference between each pair of journals in terms of their usage of the 20 statisitcial techniques we have examined is summarized in Table 3 below.

Table 4: Pairwise 'distance' in keyword usage between journals

	Bio	Med	CB	Gen	Path	NTD	Row Sums
Bio	0.000	0.125	0.215	0.020	0.031	0.090	0.481
Med	0.125	0.000	0.397	0.073	0.126	0.042	0.763
CB	0.215	0.397	0.000	0.285	0.385	0.427	1.709
Gen	0.020	0.073	0.285	0.000	0.031	0.042	0.451
Path	0.031	0.126	0.385	0.031	0.000	0.061	0.634
NTD	0.090	0.042	0.427	0.042	0.061	0.000	0.662

Note that Computational Biology and Neglected Tropical Diseases are the most dissimilar pair of journals in their usage. Computational Biology and Medicine are also quite dissimilar, as are Computational Biology and Pathogens. Biology, Pathogens, and Genetics are all fairly similar. The column on the far right gives a measure of how different each journal is from the others overall, and shows that Computational Biology is the most distinct.

4.2 Changes across time.

For each journal, we compare how prevalant certain key statistical techniques were in 2009-2010 (early on in the life of the journal), versus in 2016-17. Notice we are focused here only on the techniques that were most widely used in that particular journal. Results are displayed in Figure 2 below. Observe that Biology saw a sizable increase in the proportion of articles that used a technique, for several techniques. Medicine, on the other hand, saw some of its techniques become less widely used over time, and small increases in usage for others. Computational Biology and Genetics show similar patterns as Medicine.

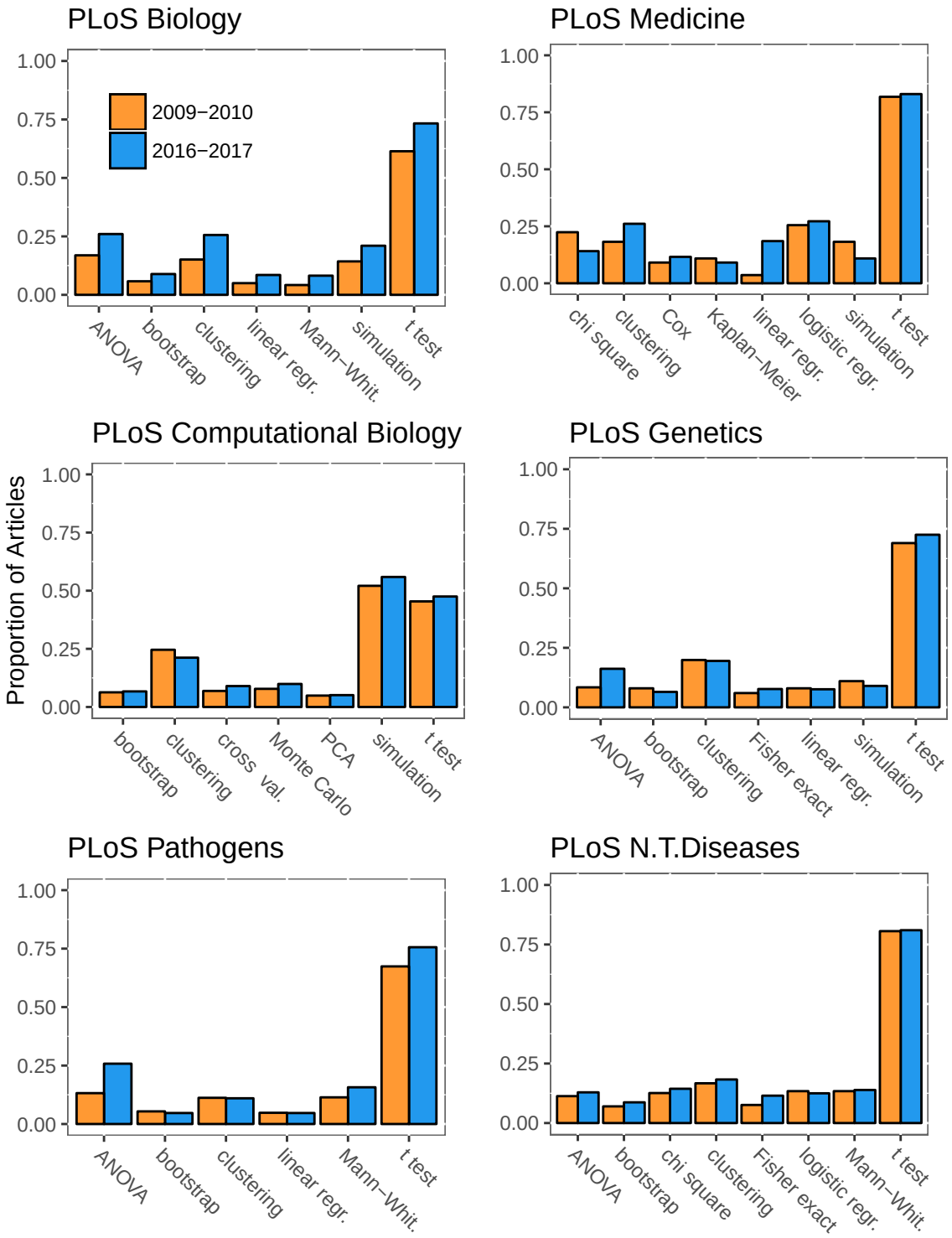

Figure 2: Prevalance of selected statistical techniques in PLoS journals, 2009-10 vs. 2016-17. Each bar shows the proportion of that journal's articles from the given time period that contain the indicated keyword.

In order to make overall comparisons across journal, we consider the change in proportions for each keyword, and also the absolute value of the change in proportions for each keyword, and sum over keywords (Table 5). Biology makes the greatest gains, has an average of roughly 1/2 more statistical techniques used per article in 2016-17 than in 2009-10. Neglected Tropical Diseases is the most consistent in its usage between 2009-10

and 2016-17, with very little increase or decrease in proportionate usage of any technique.

Table 5: **Changes in usage between 2009-10 to 2016-17, by journal.** Each measure is summed over the keywords chosen for that journal, as displayed in Figure 2.

	Biology	Medicine	Comp.Bio.	Genetics	Pathogens	N.T.Diseases
Sum for 2009-10	1.227	1.897	1.480	1.303	1.134	1.626
Sum for 2016-17	1.715	2.005	1.553	1.390	1.375	1.732
Sum of changes	0.488	0.108	0.073	0.087	0.241	0.106
Sum of |changes|	0.488	0.456	0.141	0.173	0.261	0.124

5. Conclusion

Our goal in this study was to examine keyword-journal-time relationships for only those keywords which were included in our initial list of techniques, which we formed by sampling and reading Methods sections. We saw that there were certain techniques, such as simulations and logistic regression, whose popularity varied widely across the 6 PLoS journals. Other techniques such as bootstrap had a much more uniform usage across the different disciplines. Computational Biology was the most dissimilar from other journals overall in its usage, and was especially dissimilar from subfields related to medicine. Some journals do show substantial gains in usage of many techniques from 2009-10 to 2016-17, while Computational Biology, Genetics, and Neglected Tropical Diseases were all fairly consistent in their usage between these two times.

By using a sample of Methods sections to form the list of statistical techniques to be investigated, the design of this study was such that only the more popular techniques tended to come under consideration. In future work we consider focusing on certain lesser-used techniques of particular interest. In addition to considering overall differences in usage of techniques by discipline, it would also to be useful to give some measure of the sophistication of the statistical methodology being employed in each of the disciplines.

References

1. Public Library of Science. https://www.plos.org
2. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
3. Scott Chamberlain, Carl Boettiger, and Karthik Ram (2016). rplos: Interface to PLOS Journals search API. R package version 0.6.4
4. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.
5. "Multiple graphs on one page (ggplot2)", Cookbook for R.

   http://www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)/

   Retrieved October 22, 2017, 6:42 p.m.