Usage of Most Common Statistical Methods Varies by Field and Trends Downward in PLoS Publications

Ye Zhang

Biomedical Engineering Department, Johns Hopkins University, Baltimore, MD 21218

Abstract

Statistical analysis has become an essential component of scientific research. To examine the usage of statistical analysis methods in research fields of different disciplines, a study exploring the most frequently used statistical analyses in the Public Library of Science (*PLoS*) was conducted. The use of 14 most frequently reported statistical methods was examined among all the full articles published on *PLoS* journals since 2005. In this study, the number of articles referring to each method was computed and results indicate that ANOVA and t-test are the most frequently reported methods with over 40,000 articles. Also, the usage distribution of these methods in six research fields, including biology, computational biology, genetics, medicine, neglected tropical diseases and pathogen, have been studied and suggested different fields have their preferred statistical methods. In addition, all these methods have been found to share a similar application trend, which reaches its peak in 2013 but starts to decline afterwards. These trends were discussed in light of the shrinking of journal outputs. These results highlight the application of statistical methods in scientific research and provide insight into statistical analyses training for people conducting scientific research.

Introduction

In this Information Age with overwhelwing volume of data, statistical analyses have been widely applied in scientific research and different statistical analysis methods have been preferred in research fields of different disciplines. For example, Analysis of Variance (ANOVA), has been popular in medical and agricultural research, because it can help compare the effetiveness of different treatments or any other factors [1]. In the past decades, researchers have been tracking the commonly used statistical analysis methods and their usage distribution in different fields. Several studies have explored the prevalence of different statistical analysis methods in specific journals, generally with the goal of identifying the statistical knowledge needed in specific areas. For example, surveys about statistical techniques used in the *American Educational Research Journal (AERJ)*, *Review of Educational Research (RER)* and *Educational Researcher (ER)* between 1978 and 1997 have been published and reported that the most frequently used statistical methods in these three journals were descriptive statistics, ANOVA/ANCOVA, correlation/regression, qualitative techniques, bivariate correlation, and multivariate [2]. The statistical methods presented in two high-impact factor biomedical informatics journals, the *Journal of American Medical Informatics Association (JAMIA)* and *International Journal of Medical Informatics (IJMI)* from 2002 to 2007 have also been examined and found that descriptive statistics was most often used in both journals, and chi-square and t-test were used much more frequently in *JAMIA* [3].

The Public Library of Science (*PLoS*) is a nonprofit open access science, technology and medicine publisher with a library of open access journals and scientific literature under an open content license [4]. It consists of 7 journals focusing on research fields of different disciplines, including *PLoS ONE*, *PLoS Biology*, *PLoS Medicine*, *PLoS Computational Biology*, *PLoS Genetic*, *PLoS Neglected Tropical Diseases* and *PLos Pathogens*. In this project, my objective is to explore the statistical analysis methods reported in *PLoS* articles. More specifically, this project is designed to examine the most commonly used statistical methods presented in *PLoS* journals, as well as the usage distribution of these methods among different research fields. The trends of these frequently used methods over the last 10-15 years are also explored and dicussed in this project.

Methods

Preliminary Exploration

Before I started to search and download data for analysis, the first step of our study was to define reasonable boundaries for my search. My study about the usage of statistical analysis methods was conducted among full articles published in the 7 journals on *PLoS* websites. Other types of short, non-research articles like "essays" or "perspectives" were excluded from my search because they were missing the "abstract"" and "materials and methods" sections. Given the objective of this project, it is important to establish a decent pool of "key words"" associated with statistical methods, such as "hypothesis testing", "t-test", "linear regression", "machine learning", et al. To establish this pool of key words, a list of full articles with the word "statistics" in their "abstract" was targeted using R package `rplos` [5], which contains functions that can be used for accessing articles from the *PLoS* using their API. With `rplos` and other R packages like `fulltext` [6] and `XML` [7], I was able to download abstracts of 500 full articles randomly from *PLoS*, which had the word "statistics" in their abstracts (see final_code Part II). After tidying up the downloaded 500 abstracts using R packages `dplyr` [8], `tidyr` [9], `tidytext` [10] and `stringr` [11], I summarized all the single word, "bigram" (knowns as two words combination) and "trigram" (known as three words combination) respectively and calculated their appearance frequency in these 500 abstracts. The top 20 most frequently mentioned "word", "bigram" and "trigram" were shown in **Supplementary Table 1**, **2** and **3**. Based on these three summarys, I achieved a table of 14 statistical methods that are most frequently mentioned in these 500 randomly picked abstracts, as shown in **Table 1** as below.

Table 1: Key words of statistical analysis methods

	methods		methods
1.	logistic regression	8.	linear regression
2.	meta analysis	9.	machine learning
3.	bootstrap	10.	maximum likelihood
4.	ANOVA	11.	neural network
5.	clustering	12.	random forest
6.	bayesian	13.	support vector machine
7.	t-test	14.	MCMC

Data Collection

Considering the goal of this project, the dataset should include information of all the *PLoS* full articles that used statistical methods associated with the "key words" in **Table 1** via preliminary exploration. Considering that methods used in one publication are usually described in "materials and methods" section, I used R package `rplos` to look for exact matches to "key words" listed in **Table 1** in "materials and methods" of all the *PLoS* full articles. For example, I searched the text term "support vector machine" but not its abbrevation "SVM" (see final_code Part IV). For each article with the appearance of key word, I downloaded the article title, DOI, the *PLoS* journal it belongs to and the date of publication. In this way, I achieved my data set consisting of 14 large lists. Each list corresponded to one statistical method and included all the information of articles that mentioned this method in their "materials and methods" section (see final_code Part IV).

Data Analysis

After extracting information of articles I'm interested in, the frequencies of these 14 statistical methods used in *PLoS* articles and their percentages in each journal were computed. With the dataset, I was capable of figure out the most frequently used analyses techniques, and coorelation between these techniques and the research fields (the PLoS journals). To explore the trends of application of these analyses, the frequencies of each method in each year were computed and plotted since 2005. All the data collection and analysis occurred in RStudio (version 1.0153) utilizing the R language version 3.42 [12]. Only counts and frequencies of the statistical methods reported in the articles were used for the data analysis. All the tables and figures were also produced in RStudio using R packages `plyr` [13], `knitr` [14], `kableExtra` [15], `ggplot2` [16] and `scales` [17] (see final_code Part V, VI and VII).

Reproducibility

Everything presented in this report and supplementary information are reproduced in the R markdown file `final_code.Rmd`. In order to save time for knitting, the preliminary exploration data (downloaded 500 abstracts) and dataset downloaded from websites for analysis were save as `abs500.RData` and `data.RData` respectively and uploaded under the directory `"data/"` on GitHub. To reproduce the exact same results in this report, the uploaded data files must be used because the publications on *PLoS* website increases over time.

Results and Discussion

This project described the most commonly used statistical analysis methods in research articles published on *PLoS*. A data set with 14 large lists was created, in which each list corresponded to one method and included titles, DOIs, journals and publication dates of articles utilizing this method. With this data set, the frequencies of fourteen leading statistical methods and their usage distribution among seven *PLoS* journals were computed, as well as the trends of application of these methods since 2005.

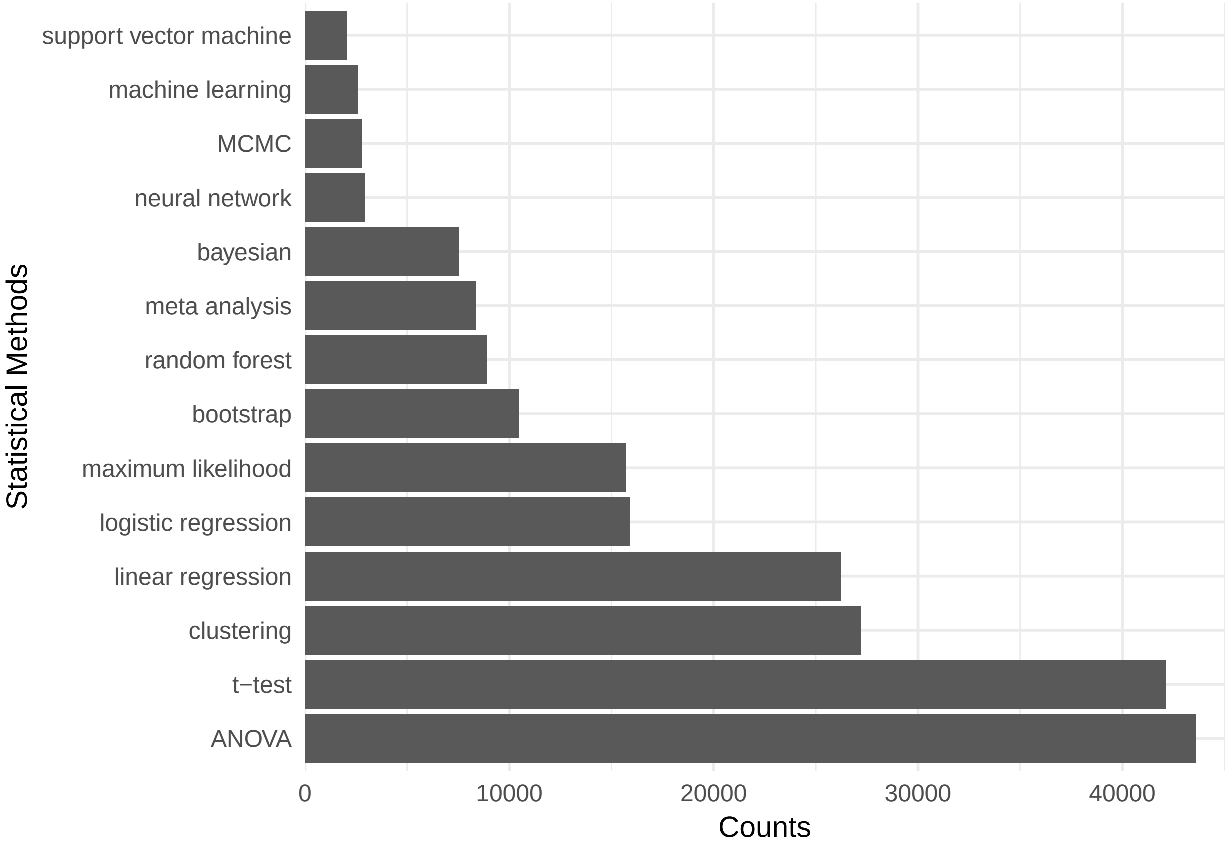

Figure 1. A barplot of the number of *PLoS* articles reporting each statistical analysis methods. ANOVA and t-test are the top two popular statistical analyses methods each having been used in over 40,000 articles. Clustering and linear regression methods are the second tier with 27202 and 26222 articles respectively. Though less frequently mentioned, logistic regression, maximum likelihood and bootstrap methods still have

been reported in over 10,000 articles. SVM and machine learning are reported in the smallest number of articles.

First, the popularity comparison of fourteen statistical analysis methods were computed by counting the number of articles that reported the methods in their "materials and methods" section, as presented in **Figure 1**. All the methods have been used in over 2000 articles, indicating that they are all pretty commonly used methods, which matches the preliminary exploration. Among these methods, ANOVA and t-test are the top two popular statistical analyses, each having been used in over 40,000 articles. Clustering and linear regression methods are the second tier with around 3/5 of the counts of ANOVA. Though less frequently reported than the top two, logistic regression, maximum likelihood and bootstrap methods still have been reported in over 10,000 articles. Other analyses like random forest, meta analysis and bayesian methods have been mentioned in 8916, 8350 and 7514 articles respectively. Neural network, MCMC, SVM and machine learning are the least frequently reported methods in *PLoS* publications, being referenced in less than 3000 articles.

Studies about the usage distribution of statistical methods in research fields of different disciplines were conducted by calcuating the percentage of each research field occupied using the same method, as shown in **Figure 2**. At the beginning of data analysis, the number of articles reporting each statistical method in each research fields was counted, as presented in **Supplementary Table 4**. It's worth mentioning that articles published in *PLoS ONE* was excluded during the analysis for usage distribution here because *PLoS ONE* is a general journal covering a wide variety of topics without a specific field of discipline, unlike the other six journals focusing on a specific research area. In addition, there are far more publications in *PLoS ONE* than the other six journals, which could affect the analysis of distrbution of methods among the other six journals. From **Figure 2** we can find that the top two popular methods ANOVA and t-test have similar ditribution and have been reported in "Pathogen" most frequently (34% and 33% respectively). The t-test and meta analysis have also been frequently reported in "Genetics" (34% and 33% respectively). It's interesting to see that some methods are obviously preferred in some specific research field, such as logistic regression with its 52% references in the field of neglected tropical diseases. Similarly, "Computational Biology" is the field where neural network (77%), SVM (51%) and machine learning (50%) are most frequently reported. There's no article published on *PLoS Medicine* using neural network method. Meta analysis and logistic regression are probably the most commonly used methods in the field of medicine (15% and 17% respectively). In addition, Bayesian analysis are often reported in the field of Computational Biology (33%) and Neglected Tropical Diseases (30%). The results have suggested that research fields of different disciplines prefer different statistical methods, which can be explained by the different types of experiment data collected and goals to achieve in these fields. For example, the identification of disease biomarkers plays a significant role in the field of genetics and pathogen research. It asks for accurate distinguishing of differentially expressed genes in two or more groups, in which t-test is undoubtly the most popular approach [19]. While in the field of medicine, meta-analysis can satisfy the need to integrate results of multiple clinical trials performed by different research groups at differet locations, and eventually derive a more precise conclusion, which is perfect for evidence-based medicine [20].

All the statistical methods have similar trends in the past 10~15 years, as presented in **Figure 3**. The number of articles reporting these methods keeps increasing since 2005, which is the year most *PLoS* journals started, until the curve reaches its top in 2013. And we can observe that the increasing rate is relatively slow at the beginning and it accelarates after 2010, while the trend curve starts to drop down slowly after reaching the top. One possible reason for the decreasing is that the number of research articles published in *PLoS* shrinks since 2013, which can cause the decrease in the counts of statistical methods reported. It has been reported that articles published on *PLoS ONE* journal, which covers over 90% of total output of *PLoS*, has been declining since 2013 [20]. There might be multiple factors that cause the reduction in *PLoS ONE* publication since 2013, including the falling of its impact factor from 4.079 in 2012 to 2.606 in 2013 [21], and the rising of its publication fees [22, 23]. Since this study uses the number of articles reporting each statistical method to plot trends, the reduction in *PLoS* production definitely affect the trending curves.

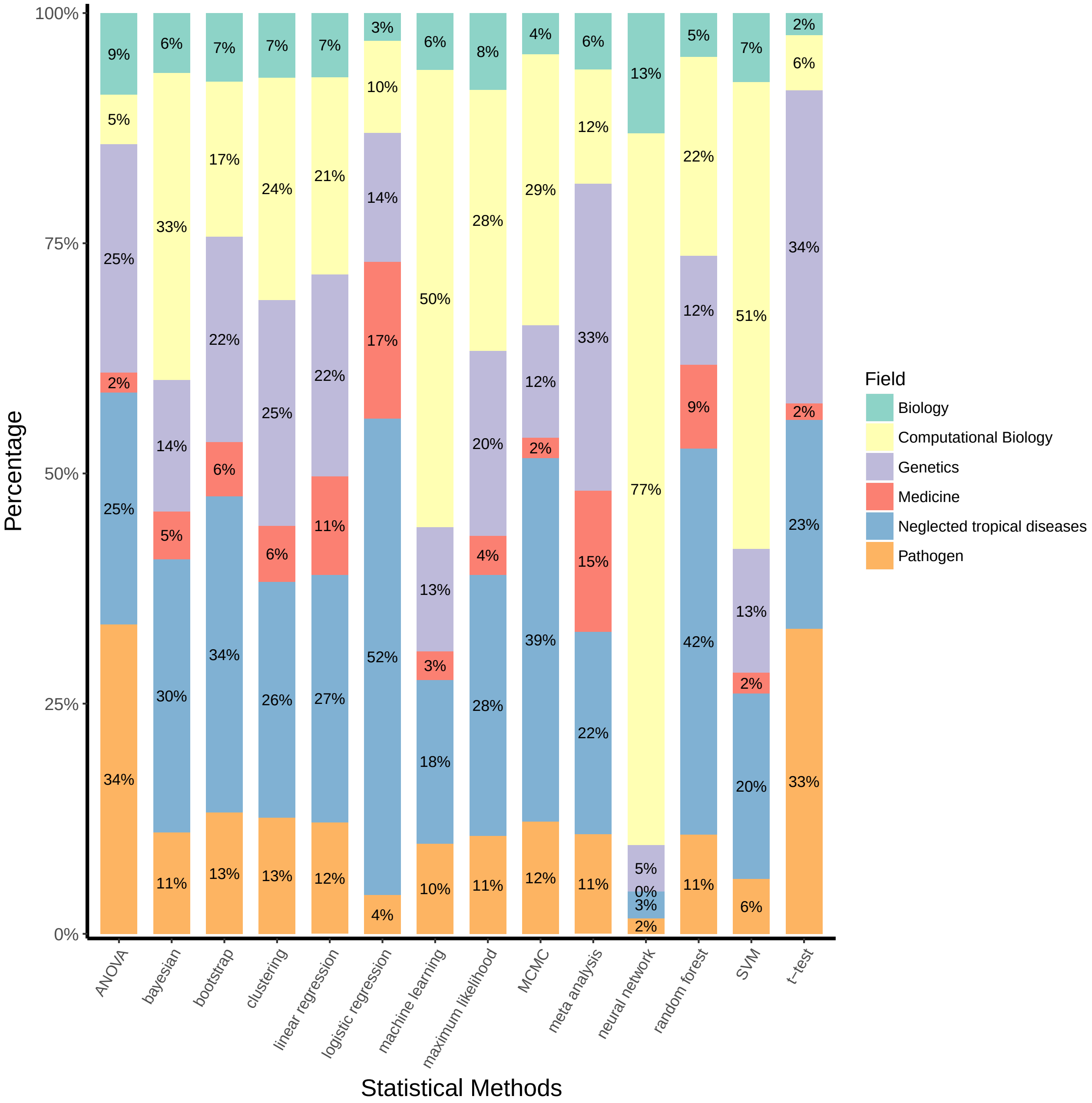

Figure 2. A distribution of statistical analysis methods in each field. ANOVA and t-test are most popular in pathogen research. T-test has also been frequently reported in the field of genetics, as meta analysis in medicine research. Logistic regression is obviously preferred in the field of neglected tropical diseases (52%). Computational Biology is the field where neural network (77%), SVM (51%) and machine learning (50%) are most frequently reported. There's no article in medicine field using neural network method.

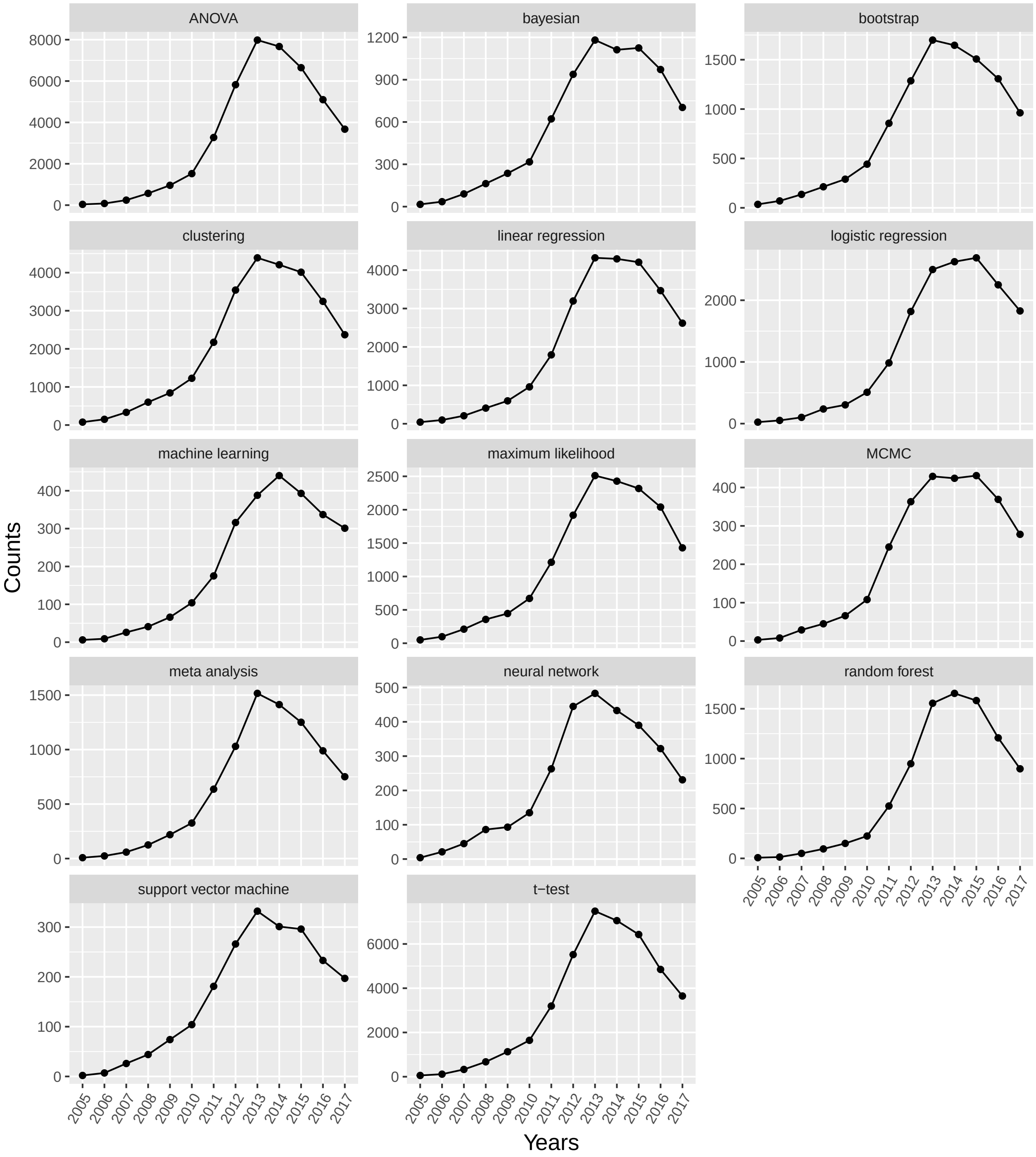

Figure 3. An illustration of trends of statistical analysis methods application between 2005 and 2017. Their usage counts keep increasing since 2005 and reach the top in 2013. The increasing accelarates during 2010 to 2013, while it starts to drop down slowly after reaching the top.

In summary, this project studies the commonly used statistical methods and their usage distribution as well as application trend in *PLoS* journals. The data collection for this study highly relies on the series of "Key words" resultant from preliminary exploration. It is possible that the use of some statistical methods were not detected through my preliminary search, so that may affect the completeness of the set of statistical methods. For the data collection, since I searched for methods using an exact match to a specific text term, some articles might not be detected because they described the method using different phasing. Thus alternative keywords for the same statistical methods and proximity word searching should be considered for the improvement of this study. In addition, the R packages I'm using can only detect the appearance of certain method in the article, however, they cannot tell how many times the method was actually referenced in the article, which could be intereting to look at for future discussion. After all, this study highlights the significance of statistical analyses in scientific researches and provide insight into statistical analysis training for future scientists.

Conclusion

This report explores the most frequently used statistical analysis methods in *PLoS* publications. My analysis suggests that ANOVA and t-test are the most frequently reported methods with over 40,000 articles refering to. The usage distribution of these methods in six research fields have also been studied and results present that fields of different disciplines have their preferred statistical methods. In addition, the application trend of these frequently used methods over the last 10-15 years are explored and shows that all the statistical methods have been increasely reported in *PLoS* publications, but the number of articles reaches its peak in 2013 and starts to decrease afterwards, which could be correlated with the output drops at *PLoS* which started around the same time.

Reference

1. Hsu, T.(2005) Research methods and data analysis procedures used by educational researchers. International Journal of Research & Method in Education, 28(2), 109-133.
2. Elmore, P. B. & Woehlke, P. L.(1998) Twenty years of research methods employed in the American Educational Research Journal, Educational Researcher, and Review of Educational Research, paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA, April.
3. Scotch M, Duggal M, Brandt C, Lin Z & Shiffman R.(2010) Use of statistical analysis in the biomedical informatics literature. JAMIA, 17(1):3–5.
4. https://www.plos.org.
5. Scott Chamberlain, Carl Boettiger and Karthik Ram (NA). rplos: Interface to the Search API for 'PLoS' Journals. R package version 0.7.8.9110. https://github.com/ropensci/rplos.
6. Scott Chamberlain (2016). fulltext: Full Text of 'Scholarly' Articles Across Many Data Sources. R package version 0.1.8. https://CRAN.R-project.org/package=fulltext.
7. Duncan Temple Lang and the CRAN Team (2017). XML: Tools for Parsing and Generating XML Within R and S-Plus. R package version 3.98-1.9. https://CRAN.R-project.org/package=XML.
8. Hadley Wickham, Romain Francois, Lionel Henry and Kirill Müller (2017). dplyr: A Grammar of Data Manipulation. R package version 0.7.4. https://CRAN.R-project.org/package=dplyr.
9. Hadley Wickham and Lionel Henry (2017). tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. R package version 0.7.1. https://CRAN.R-project.org/package=tidyr.
10. Silge J and Robinson D (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *JOSS*, *1*(3). doi: 10.21105/joss.00037 (URL: http://doi.org/10.21105/joss.00037), <URL: http://dx.doi.org/10.21105/joss.00037>.
11. Hadley Wickham (2017). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.2.0. https://CRAN.R-project.org/package=stringr.
12. R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
13. Yihui Xie (2017). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.17. Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963. Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

14. Hao Zhu (NA). kableExtra: Construct Complex Table with ‘kable’ and Pipe Syntax. http://haozhu233.github.io/kableExtra/, https://github.com/haozhu233/kableExtra.
15. Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.
16. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.
17. Hadley Wickham (2017). scales: Scale Functions for Visualization. R package version 0.5.0. https://CRAN.R-project.org/package=scales.
18. Hadley Wickham and Winston Chang (2017). devtools: Tools to Make Developing R Packages Easier. R package version 1.13.3. https://CRAN.R-project.org/package=devtools.
19. Jeanmougin M, de Reynies A, Marisa L, Paccard C, Nuel G, et al. (2010) Should We Abandon the t-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies. PLoS One 5: e12336.
20. Haidich A. B.(2010) Meta-analysis in medical research. Hippokratia 14, 29–37.
21. http://www.sciencemag.org/news/2014/06/output-drops-worlds-largest-open-access-journal.
22. https://scholarlykitchen.sspnet.org/2016/02/02/as-plos-one-shrinks-2015-impact-factor-expected-to-rise/.
23. http://blogs.plos.org/plos/2015/09/plos-publication-costs-update/.
24. https://scholarlykitchen.sspnet.org/2016/01/06/plos-one-shrinks-by-11-percent/.