An Overview of Statistical Methods Used in PLoS Publications

Ye Zhang

10/11/2017

Introduction

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, the statistical methods presented in the Journal of American Medical Informatics Association and the International Journal of Medical Informatics have been studied [1], and so as the original publications in South African Medical Journal [2].

The Public Library of Science (PLoS) is a nonprofit open access science, technology and medicine publisher, innovator and advocacy organization with a library of open access journals and other scientific literature under an open content license [3]. It has 7 journals, PLoS one, PLoS Biology, PLoS Medicine, PLoS Comutational Biology, PLoS Genetic, PLoS Neglected Tropical Diseases and PLos Pathogens. In the project, our objective is to describe the statistical analysis methods reported in all PLoS published articles. We exam the most commonly used statistical methods as well as their usage distribution among different fields. We also describe the use of statistical methods trends over the last 10-15 years.

Methods

Preliminary Exploration

Before searching and downloading data about usage of statistical analysis methods in PLoS publications and information related to these publications, it is important to establish a decent pool of key words, such as "Hypothesis testing", "t-test", "linear regression", "machine learning", et al., which are the most commonly used statistical analysis methods and should be mentioned frequently in these articles. In order to establish the pool for the key words, first a list of full articles with the word "statistics" in "abstract" is searched using R package `rplos`, which contains functions that can be used for PLoS article searching and information download. By indicating "statistic" in the "abstract" part, we can achieve result `outide_id` containing all the DOIs of all the full articles that we are interested in and then download the abstracts of these articles. Here I randomly downloaded abstracts of 500 full articles with the word "statistics" in their abstracts. After tidying up this preliminary data using R package `tidyr` and `tidytext` and removing all the numbers, I unnest the tokens using `word`, `bigram` (two words combination) and `trigram` (three words combination) respectively and calculated the frequency of these `word`, `bigram` and `trigram`. After going through these three data frames ordered with frequenc I achieved a summary of thestatistical methods most frequently mentioned in the 500 abstracts, named `dic` for subsequent data collection.

Data Collection

The dataset for this project should include information of all the PLoS full articles that used the statistical methods within the key words pool `dic` created through preliminary exploration. With R package `rplos`, I downloaded "title", "DOI", "PLoS journal" and "Date of publication" of all the published full articles with the key word in their "Materials and Methods" section.

Data Analysis

After extracting information of articles with key words from the website, I can start to explore the usage of these statistical methods and their distribution among the different areas. With the dataset established, it's possible to figure out the most commonly utilized analyses techniques, and coorelation between these techniques and the fields (the PLoS journals) and publication years. Take the key word "t-test" as an example, I can figure out how many times the "t test" is mentioned over the years as well as in articles among 7 different fields.

Reproducibility

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

Results and Discussion

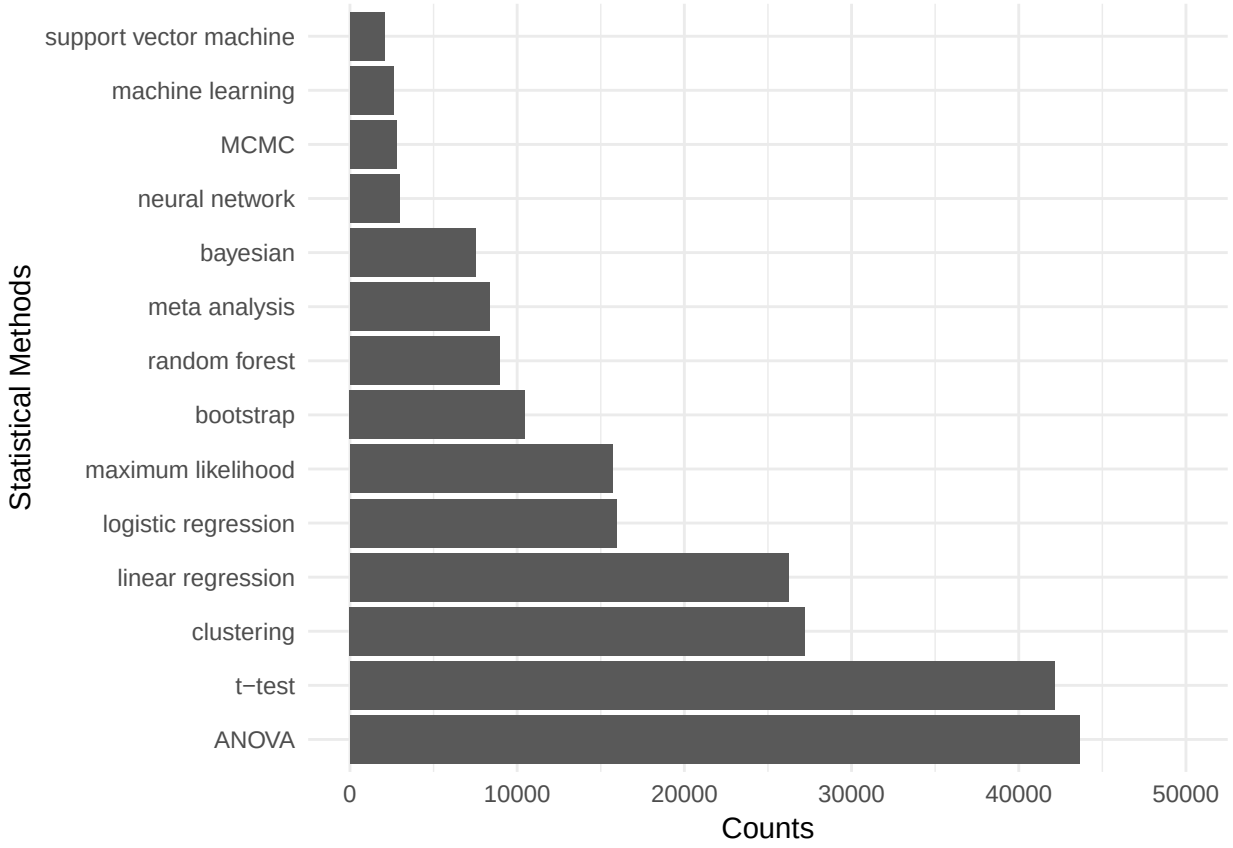

Figure 1 A barplot of the number of PLoS publications using each statistical analysis methods. ANOVA and t-test are top two popular statistical analyses methods among PLoS publications. Clustering and linear regression methods are the second tier with half of the counts. SVM and machine learning are used in the smallest number of PLoS publications.

The dateset for analysis contains information for statistical analysis methods used in PLoS publication. By counting the number of articles that mentioned the statistical analysis methods in their "materials and

methods" section, the popularity comparison of the statistical methods are summarized and presented as **Figure 1**. As shown 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 top two popular statistical analyses methods among PLoS publications, each being used in over 40,000 publications. Clustering and linear regression methods are the second tier with half of the counts. Though not as popular as the top two, logistic regression, maximum likelihood and bootstrap methods still have been used in over 10,000 articles. SVM and machine learning are the least popular methods that have been used in PloS publications.

The usage distrbution of statistical analysis methods in different fields are explored by calcuating the counts of each method in each PLOS journal, as shown in **Figure 2**. The top two popular methods ANOVA and t-test have similar ditribution and have been applied in Pathogen and Genetics most frequently. In the field of Computational biology and Neglected tropical diseases, clustering method is most commonly used. Besides, linear regression and maximum likelihood methods have also widely used in the field of Computational biology, genetics and Neglected tropical diseases.

All the statistical methods have similar trends in the past 10~15 years. Their usage keeps increasing since 2005, which is the year PLoS started, until reaches the top in 2013. And its' quite obvious that the increasing accelarates after 2010, while it starts to drop down slowly after reaching the top. However, one possible reason for the declining is that the PLoS publication shrinks since 2013 [4], which can affect the trend presented in **Figure 3**.

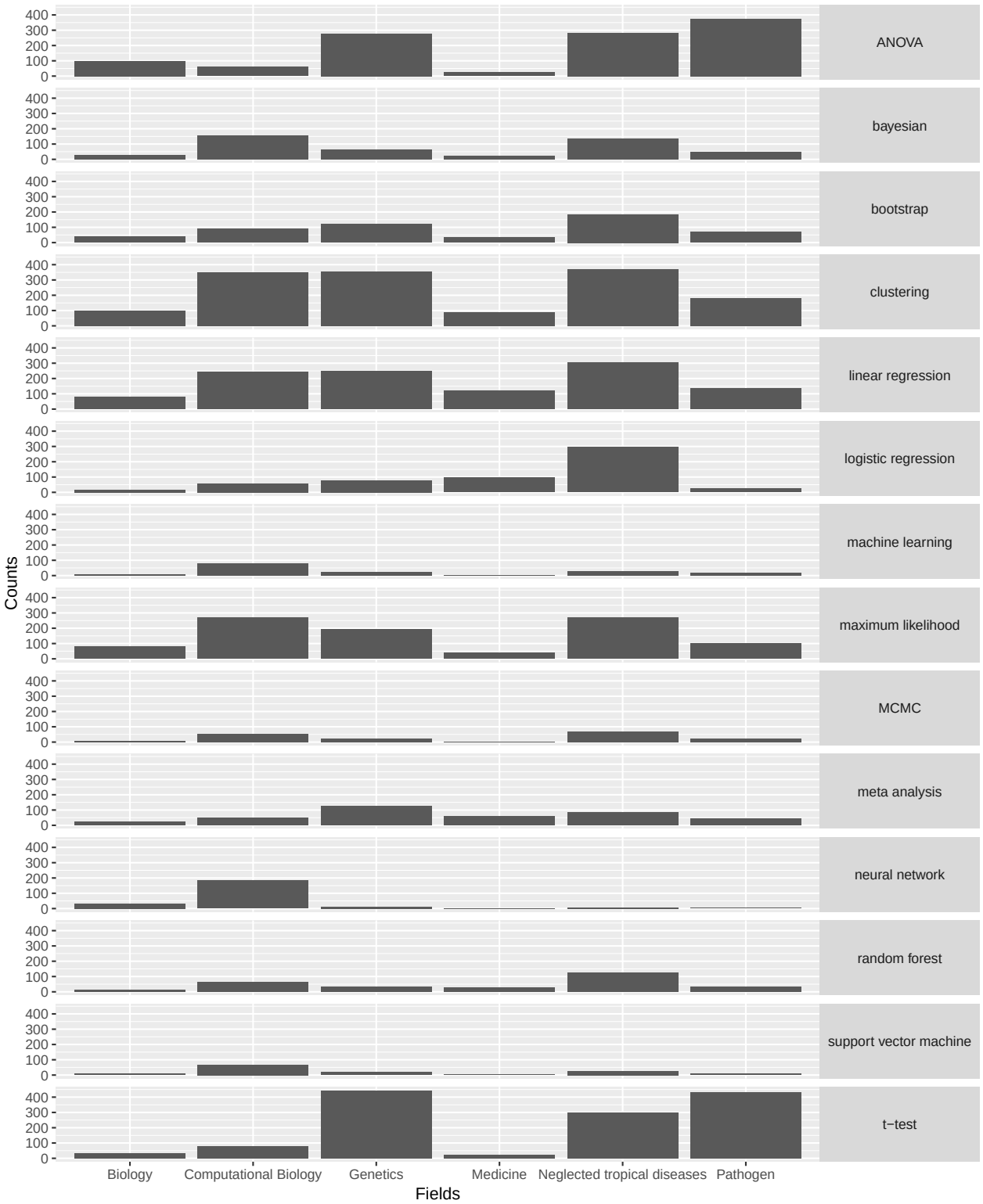

Figure 2 A distribution of statistical analysis methods in each field. ANOVA and t-test are most popular in Pathogen and Genetics. clustering method is most commonly used in the field of Computational biology and Neglected tropical diseases. Linear regression and maximum likelihood methods have also been applied in Computational biology, Genetics and Neglected tropical diseases.

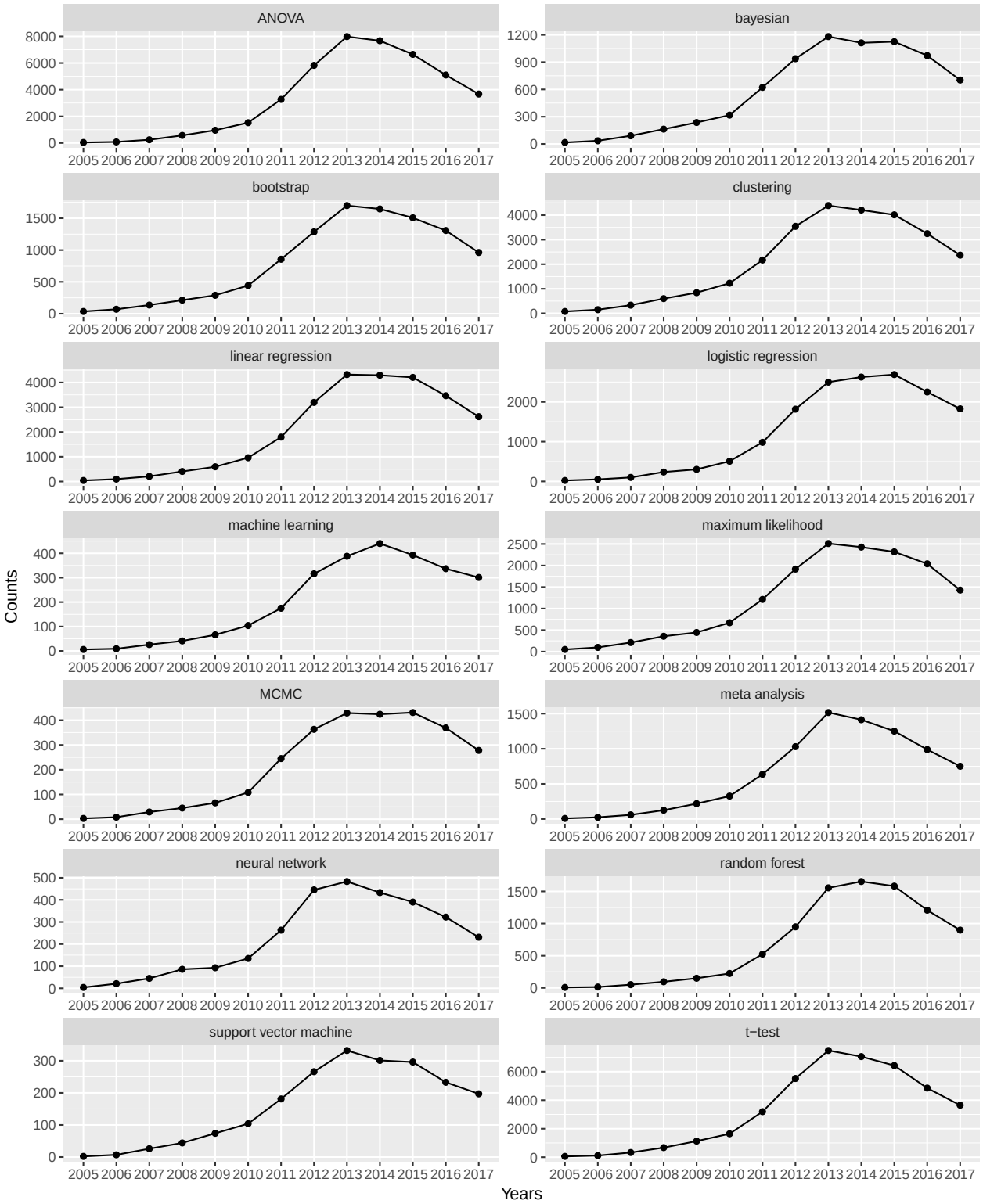

Figure 3 An illustration of trends of statistical analysis methods usage between 2005 and 2017. Their counts keep increasing since 2005 and reach the top in 2013. And the increasing accelarates during 2010 to 2013, while it starts to drop down slowly after reaching the top.

Conclusion

This report explores the most commonly used statistical analysis methods in PLoS publications Our analysis suggests that the t-test and ANOVA are the most frequently referenced methods, and they have been prefered to be applied the field of Genetics and Pathogen. The publication trend has also been studied and shows all the statistical methods have been increasely referenced in PLoS publications, but the publication number start to decline since 2013, which could be correlated with the publication shrinking starting around the same time.

Reference

1. Scotch M, Duggal M, Brandt C, Lin Z, Shiffman R: Use of statistical analysis in the biomedical informatics literature. JAMIA 2010, 17(1):3–5.
2. Becker PJ, Viljoen E, Wolmarans L, IJsselmuiden CB: An assessment of the statistical procedures used in original papers published in the SAMJ during 1992. South African medical journal 1995, 85(9):881–884.
3. https://www.plos.org.
4. https://scholarlykitchen.sspnet.org/2016/01/06/plos-one-shrinks-by-11-percent/.