如何在大数据时代中写好统计学论文

摘要

本文通过对文献数据库中 4 本杂志在 2018 年中刊登的所有论文数据进行分析，在对关键字分析中得到论文研究方法和思想的特点；在对引用文章的数量上可以看到各位作者如何更好的能建立论文的权威与可信度；此外，我们还对四个不同的杂志分开观察，来得到部分作者的论文合作关系图；最终得到大数据时代下统计学研究一些技巧。

一、背景介绍

1.1 大数据时代

不论你愿不愿意承认，大数据时代已经来临了。大数据潮流引领的技术变革正在悄无声息地改变着各行各业。虽说“大数据”是近些年才火热起来的词汇，但可以说“大数据”其实一直存在，只是由于技术的局限性使得人们在很长的一段时间里没有办法能够使用全量数据。

但是随着技术的发展与革新，现在人们可以使用大数据技术来处理海量的数据了，这使得很多之前只能停留在理论研究层面的算法和思想现在能够付诸行动，比如现在很火爆的深度学习。与此同时，大数据技术这一新兴的工具也让人们拥有了一种新的思维模式，即大数据思维。大数据思维注重全量样本数据而不是局部数据，注重相关性而不是因果关系。通过分析和挖掘数据将其转化为知识，再由知识提炼成智慧以获取洞察。大数据思维在很多行业都有用武之地，比如在银行行业，基于大数据的风险控制体系就是一个很好的例子。而数据往往和统计有着密切的联系。

1.2 统计学研究

统计学是通过搜索、整理、分析、描述数据等手段，以达到推断所测对象的本质，甚至预测对象未来的一门综合性科学。统计学用到了大量的数学及其它学科的专业知识，其应用范围几乎覆盖了社会科学和自然科学的各个领域。

统计对大数据的生命力和应用价值都有着至关重要的作用。统计学对科研工作者有多重要呢？只要是做科研，就需要进行科学研究的探索和撰写科研论文，在这其中，都离不了从实验设计、收集和分析数据的步骤，这些都必须要用到统计学方法。

然而，大家一想起统计学，脑海里免不了出现那些晦涩难懂的统计学原理和计算公式，还有各种试验数据的统计方法的选择，不免头痛。接下来，我们将对统计学学术论文的进行分析得到它们的特点，带你了解统计学研究的常用方法。

二、 数据来源与说明

数据来自在文献数据库网站“web of science”爬取论文数据。分别为《ANNALS OF STATISTICS》、《JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION》、《JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES B-STATISTICAL METHODOLOGY》、《BIOMETRIKA》4 本杂志在 2018 年中刊登的所有论文数据，总共有 393 篇文献，它们的作品数量分别为：127、144、47、75。由扇形图可以看出文献更多的来源于《ANNALS OF STATISTICS》、《JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION》两本杂志；

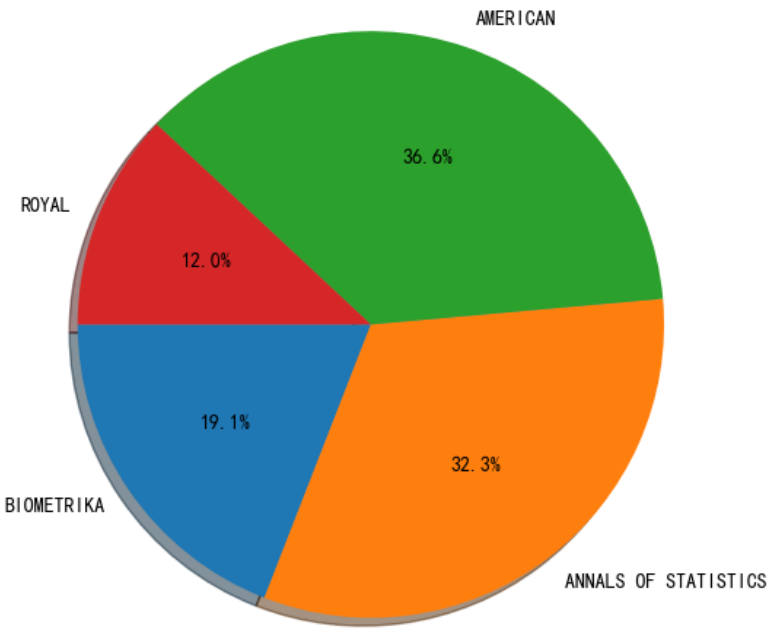

图 1 数据来源

三、 数据分析

3.1 关键字分析

关键词是从文章的题名、摘要、正文中抽出的，并能表达全文内容主题，具有是在意义的单词或术语。而作为统计学学术论文，我们能够通过关键字看到文章的研究方法和大致研究内容。经过统计，在 393 篇论文中，除去 9 篇没有收集到关键字的论文，总共有 1641 个关键字，其中出现次数最多的 5 个关键为 Causal inference（因果推断）、Variable selection（变量选择）、False discovery rate（错误

发现率)、Lasso（Lasso 算法)、Functional data（函数型数据分析)。观察关键字词云能够发现模型、回归分析等显著的字眼。

图 2 关键字词云

再对各个出版社进行分开统计分析,我们可以看到 lasso 算法、Causal inference（因果推断)、False discovery rate（错误发现率）在不止一本杂志上排行上榜，而 Variable selection（变量选择）更是在美国统计中出现高达 12 次。

表 1 不同出版社关键字统计

出版社	文章总数	未收集到关键字的文章数	关键字总数	出现次数前三的关键字
ANNALS OF STATISTICS	127	0	634	sparsity:7 robustness:5 bootstrap:4
ROYAL	47	0	226	Causal inference:5 Quantile regression:4 False discovery rate:3
AMERICAN	144	6	603	Variable selection:12 False discovery rate:5 LASSO:4
BIOMETRIKA	75	3	309	Causal inference:5 Lasso:4 Crossvalidation:3

3.2 论文引用文献数分析

许多作者通常会有一个疑问——论文要引用多少文献才是适当的？其实这没有一个标准答案，只要论文需要，就都得放进去。关于引用要放多少并没有太多明确的规定，但是过多或过少的引用数量，都隐射着一个信息，作者学术能力和其论文的不足。原因如下：如果引用数量不足，尤其是大众比较熟悉的主题，读者可能会认为作者没有做好足够的文献研究。如果作者没有下功夫做好研究，读者会质疑论文是否值得花时间阅读。读者会想:论文有什么创新点和价值？作者是不是对学术研究不太在行？此外，读者也可能认为你有抄袭的可能，因为当引用文献过少时很可能是因为作者没有将参考的相关文献列入引用文献中。如果引用数量过多，读者会想你是否有彻底地分析与研究主题相关的文献。如果是文献回顾类型的文章，那引用数量则另当别论。论文需专注于研究主题的调查与发现，千万不要将自己的心血结晶埋没在一堆文献与其他研究的讨论之中。让读者看到你发现了什么，以及这个新的发现是否偏离或符合现在学术圈对于你论文主题的理解。

所以说，适当的文献数量能也能更好的能建立论文的权威与可信度，经过统计该数据集所有论文所引用的文献数，结果如图 3 所示

	citation
count	393.000000
mean	37.839695
std	16.314894
min	1.000000
25%	28.000000
50%	36.000000
75%	45.000000
max	113.000000

图 3 所有论文引用文献信息统计

由图可以看出，在 393 篇论文中，平均每篇论文引用文献数约 38 篇，最高引用文献数量高达 113 篇，最低引用文献数也仅有 1 篇，但是根据四分位数的结果来看，大部分论文引用文献数目为 36 篇左右。

3.3 作者合作网络图

有些学者喜欢小步快跑，喜欢研究一些相对容易的小问题，一路沿着一个大方向发着，积累到一个阶段后，偶尔也能发 1-2 篇高质量的。有些学者则喜欢憋大招，像数学家张益唐一样，语不惊人誓不休，博士毕业直到快 60 岁就没发什么小论文，他不喜欢研

究小问题，甚至对小问题有些鄙视，认为这些问题迟早会一般的人解决的，而大问题需要天才来解决。但无论何种性格的学者，一年发表 3-4 篇自己满意的论文，应该是顶了天了，至于发 8-10 篇，其质量可想而知，因为一个人的精力是有限的。

经过统计得到参与发表论文数最多的几位作者如下表所示：

表 2 发表论文数最多的几位作者

Cai,T.Tony@University of Pennsylvania	7
Fan,Jianqing@Fudan University@Princeton University	7
Liu,Han@Princeton University	6

这三位作者的发表论文数量均达到了 6 篇以上，可见其实力非同一般。
之后对四本杂志分开分析，找到他们发表论文数最多的两位作者，对他们和其他作者的合作关系进行绘制合作网络关系图

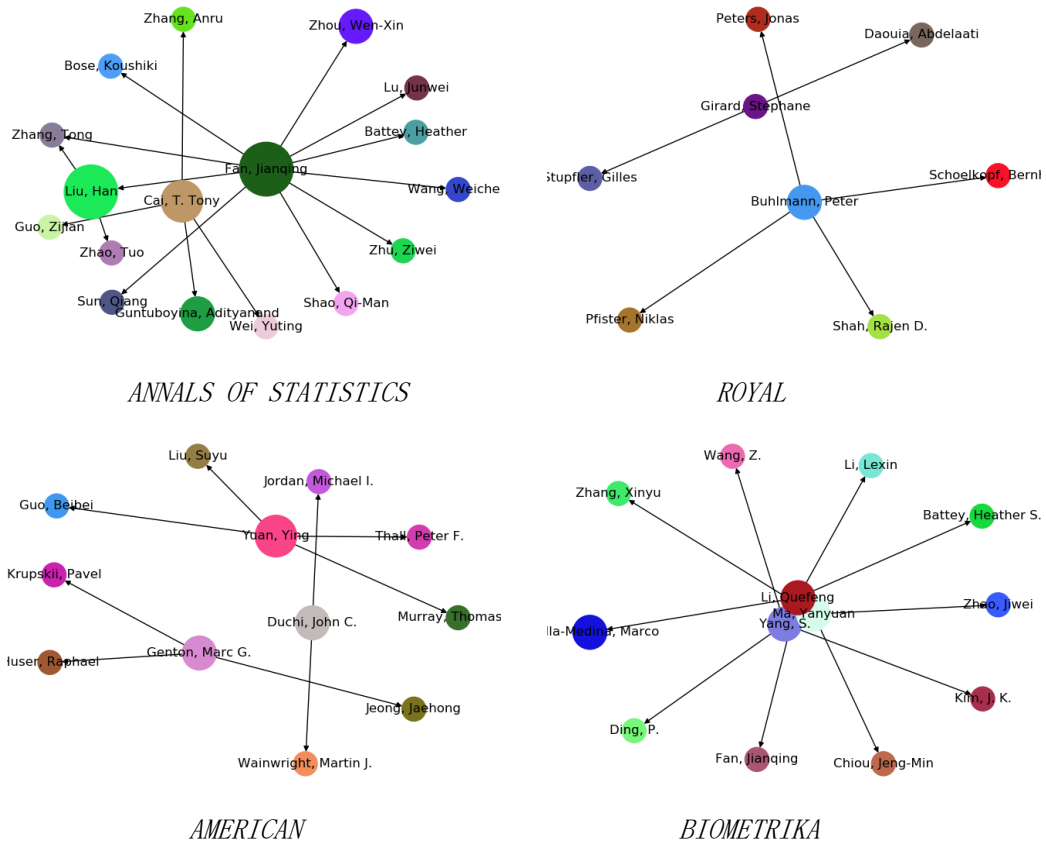

图 4 各出版社发表文章最多的两名作者合作关系图

图中每个点代表一位作者，点越大，作者参与发表的论文数越多，两点之间连线表示这两位作者合作过，由图可以看到，这些作者的论文“朋友圈”也是很广泛的

四、 结论

要想在统计学方面对于大数据时代的发展潮流更好的适应,就必须有更好的数据思维能力，写出更优秀的统计学论文，通过以上对四大统计杂志的论文进行分析，给出以下建议：

（1） 熟练的掌握 Causal inference（因果推断）、Variable selection（变量选择）、False discovery rate（错误发现率）、Lasso（Lasso 算法）、Functional data（函数型数据分析）等统计研究方法，更好的建立模型；

（2） 引用文献数量控制在 36 左右；

（3） 多结识一些学习上的朋友，因为有朋友我们的学习生活才不会太枯燥，朋友可以在我们在学习上帮助我们。