

利用统计方法与规律发现论文数据造假

刘 清 海

中山大学学报编辑部, 510080, 广州

摘 要 期刊编辑在防范学术不端过程中起着重要作用, 其一即发现问题。期刊编辑可以在 EXCEL 中验证统计学结果, 以发现统计数据造假; 可以利用统计检验量分布值与 P 值对应规律以及区间值与 P 值对应规律, 发现数据造假。这些方法为防范学术不端体系增加了利器, 期待有更多学者发掘相关规律与分享成果。如欲发现论文的数据造假, 期刊编辑首先须提高数据审查意识, 其次应提升统计学素养与能力。发现数据造假, 还必须要求来稿提供统计检验量值与精确 P 值。在论文评议流程中, 建议增设数据与统计学审查流程, 以加强对数据的把关力度。

关键词 学术不端; 论文造假; 数据造假; 统计学; 数据审查

Using statistical methods and rules to discover data fabrication in journal papers//LIU Qinghai

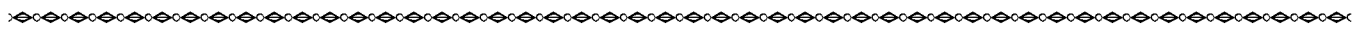
Abstract Journal editors play an important role in preventing academic misconduct. To detect statistical fabrication, journal editors can verify statistical results in EXCEL, and use the statistical law of the tested distribution value and corresponding P value or the law of interval value and corresponding P value. These findings are helpful to prevent academic misconduct, and it is expected that more scholars join the group to explore the relevant laws and share the findings. To find out the data falsification, journal editors must firstly improve their awareness of data review, and secondly, enhance their statistical literacy and ability. In addition, the tested distribution value and accurate P value are needed in papers. We suggested that data and statistical review process be inserted in paper review process so as to strengthen the data check.

Keywords academic misconduct; paper fabrication; data fabrication; statistics; data check

Author's address Editorial Department of Journal of SUN Yat-sen University, 510080, Guangzhou, China

DOI:10.16811/j.cnki.1001-4314.2018.06.019

学术不端事件在全世界范围内存在已久^[1-2], 从 2017 年 4 月《Tumor Biology》集中批量撤稿 107 篇以来尤其得到国内各界关注。研究其成因、社会背景、防范措施文献已有不少, 多数学者认同防范学术不端不是一个部门或少数人可以完成的, 必须是由多个部门通力合作, 甚至有学者提出, 应构建全球性的反学术不端合作网络^[3-6]。虽然学术不端行为涉及了论文发表、职称申报、成果署名、成果认定与鉴定、基金申请、学术评价、学术荣誉等多个方面, 但各个方面或多或少都与论文相关。因此, 笔者^[5]曾提出, 期刊编辑在防范学术不端流程当中起着不可忽视的作用, 编辑除了在编辑流程中利用自动查重软件等简单措施查出重复率高的稿件、一稿多投等不端行为之外, 还在诚信教育、发现与控制、认定与惩罚、建档与利用等方面也可发挥重要作用。本文希望从期刊编辑发现问题的角度入手, 为防范学术不端体系增加资料与方法。由于医学论文的统计学问题较多^[7-9], 以前较关注统计学报告项目、统计学方法、统计图表等, 但对统计结果与数据的符合程度还关注较少, 因此, 国内外已有不少期刊开始增设统计学审查的把关, 如《Science》和《Nature》



30(1): 19

[11] PETRO C. Undergraduates' use of Google vs. library resources: a four-year cohort study[J]. *College & Research Libraries*, 2016, 77(5): 614

[12] AKEROYD J. Discovery systems; are they now the library? [J]. *Learned Publishing*, 2017, 30(1): 88

[13] ZHU J. Should publishers work with library discovery technologies and what can they do? [J]. *Learned Publishing*, 2017, 30(1): 71

[14] 董文鸳. 科研人员对 Google Scholar 的采纳行为及启示 [J]. *图书与情报*, 2014(4): 11

[15] KILLORAN J B. How to use search engine optimization techniques to increase website visibility? [J]. *IEEE Transactions on Professional Communication*, 2013, 56(1): 50

[16] Google algorithm's top ranking factors[EB/OL]. (2018-01-21)[2016-03-01]. <http://www accuracast.com/seo-weekly/ranking-factors.php>

[17] SU Ao-Jan, HU Y C, KUZMANOVIC A. How to improve your Google ranking: myths and reality[C]//2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. [S.l.]: IEEE, 2010: 49

[18] 刘艳民, 魏清华. 图书馆发现系统功能及相关度排序评价研究[J]. *情报科学*, 2015, 33(3): 90

[19] 武丽娜, 左阳, 窦天芳. PRIMO 发现系统应用的可用性设计调研及评测[J]. *图书馆杂志*, 2017, 36(5): 47

[20] 郑雯译. 国外发现系统分面过滤功能比较: EDS、Summon、Primo[J]. *大学图书情报学刊*, 2016, 34(1): 26

(2018-05-29 收稿; 2018-10-30 修回)

等^[10]。笔者所在编辑部也增加了一个“统计学审查”的任务,在审查过程中发现,有些存在统计学问题的论文,很可能也涉及研究数据的造假,而如果把捏造统计学结果也算作数据造假的一种,那么这种可能性就变成了肯定。下面就本人对医学论文统计学审查中总结的2种发现统计结果造假的方法跟大家分享,期望更多的期刊发现规律,利用规律,助推学术净化。

1 验证统计结果,发现统计结果造假

笔者在统计学审查时,遇见一篇文章,其原文的表2如本文的表1所示(略去2个指标并略去计量单位),论文作者按本刊要求提供了统计检验量的 t 值,但是并未提供精确的 P 值,因此,笔者对其统计结果产生怀疑。为了方便验证简单的统计学结果,笔者用EXCEL做了一个模板,用以检验无原始数据但有样本量、均数及标准差的 t 检验结果。EXCEL中,可设置 $t = \text{ABS}((a_1 - a_2) / \text{SQRT}(S^2 * (1/n_1 + 1/n_2)))$ 。其中: a_1 、 a_2 为2组均数(以数据所在格子代入,下面的标准差、样本量等同理以格子代入); n_1 、 n_2 为2组样本量; S^2 为合并方差, $S^2 = (n_1 - 1) * b_1^2 + (n_2 - 1) * b_2^2$,其中 b_1 、 b_2 为2组标准差。 $P = \text{TDIST}(t, n_1 + n_2 - 2, 2)$,括号内数据分别代表 t 值、自由度、双侧检验。由于原文表2未提供样本量,笔者在其表1中找到了分组依据与结果,各为8例,以这些数据代入检验得出,2组年龄比较, $t = 0.3475$, $P = 0.7334$;WBC数比较, $t = 1.1418$, $P = 0.2727$;……。5个检测指标,无一统计量正确!遂要求作者提供原始数据及重新统计,但作者以各种理由搪塞,最后提供的数据与原稿出入较大,再以其提供的数据进行 t 检验核校,还是存在问题,最后认定该稿至少统计学结果造假,还涉嫌原始数据造假,即使通过了主编定稿,笔者还是建议编辑与主编将其退稿。

表1 2组比较 t 检验结果论文原数据(省略计量单位)

组别	年龄	WBC数	采集量	CD34+细胞数	MNC数
A组	32 ± 12	21.6 ± 8.7	189 ± 113	0.88 ± 0.45	3.0 ± 1.2
B组	34 ± 11	26.8 ± 9.5	202 ± 108	2.94 ± 0.98	5.3 ± 2.4
t 值	-0.241	-0.351	-1.006	-2.442	-3.891
P 值	>0.05	<0.05	>0.05	<0.05	<0.01

除了 t 检验外,卡方检验也容易在EXCEL中验证其统计检验结果。笔者亦在EXCEL中做了模板,以方便统计学审查时使用。不过,有些比较复杂的统计学方法,不易在EXCEL中实验,可能必须找作者提供原始资料在SPSS等专门的统计软件中验证。

2 利用统计规律,发现统计结果造假

2.1 统计检验量分布值与 P 值对应规律

除了验证统计学结果之外,有时不需要借助数据的计算,也可发现一些统计学造假。仍以表1为例,我们不看某一指标,而是综览所有指标,可以发现,从左至右, t 值的绝对值是逐渐增加的,按照 t 分布的规律,由于是同样的2组比较,样本量一样,因此, P 值应当逐渐减小,虽然表中并未提供精确 P 值,但“WBC数”与“采集量”的 P 值范围已经违反了规律,因此,可以确定,统计结果存在问题。一般地,其他变量固定时,随着统计检验量绝对值的增加, P 值相应变小。利用这种统计检验量的分布值与 P 值关系规律的思路,我们不但可以发现 t 检验与卡方检验这些简单的统计检验中的问题,还可以发现回归分析等稍微较复杂的统计学方法的检验结果中存在的问题。表2为笔者统计学审查时遇到的多元线性回归结果,在没有原始数据的情况下无法经简单的数据统计验证,但是可以看出,前3行的数据随着 t 值的增大, P 值并没有相应减小,而后3行的数据则符合该规律;因此,判断此表统计学结果存在问题,结合前后数据,可以认为,“性别”及“职业”因素的检验结果可能有误。之后,我们提醒作者注意核对统计学结果,但作者修回后并未修改这些错误,这更增加了论文统计学造假的嫌疑,最后稿件被认为存在造假嫌疑而作为退稿处理,作者也未有任何的申诉,验证了我们的推断。

表2 多元线性回归结果分析论文原数据

影响因素	回归系数	标准误	标准回归系数	t 值	P 值
性别	0.753	0.162	0.256	4.651	0.001
职业	0.827	0.162	0.279	5.106	0.004
文化程度	1.455	0.237	0.346	6.149	<0.001
家庭月收入	0.983	0.196	0.277	5.011	0.003
介入治疗	1.035	0.167	0.343	6.212	<0.001
药物副作用	0.982	0.170	0.316	5.776	0.004
抑郁	0.896	0.180	0.283	4.974	0.006

2.2 区间值与 P 值对应情况 除了统计检验量的分布值与 P 值有对应关系之外,其他还有一些值也与 P 值相关,如置信区间与 P 值。在分析某疾病风险因素的时候,除了要知道某变量是否为某病的风险因素之外,还应了解这个风险有多大,统计学上常用优势比(odds ratio, OR)或相对危险度(relative risk, RR)等来表示,且还要提供效应大小的区间范围,如95%的置信区间(95% confidence interval, 95% CI)。通常情况下,如效应值远小于1,认为该变量为保护性因素,如远大于1则认为其为伤害性因素,如约等于1则认为其对疾病没有影响。由于抽样的影响,效应值并非一个固定值,可以通过样本计算其95% CI,如果效应的95% CI包含了1,则统计检验按检验水平0.05接受

零假设,暂不能认为该因素为风险因素;如95% CI的范围皆在1的同侧(即两端值皆大于1或皆小于1),则认为该因素为风险因素。

表3为笔者统计学审查时遇到的论文数据,可以看出,“出血量”的95% CI右端值刚好为1,其 P 值也刚好为0.05,这是符合统计学原理的,而“血流稳定性”的OR较大,其95% CI皆大于1,理论上,其 P 值应远小于0.05,但论文提供的 P 值为0.901,可以判断该表格统计结果有误。同理可以推断,“伴有DIC”这个变量的结果符合统计学规律,而“血管造影发现出血”则不符合统计学规律。后来文章按上例处理,得到同样的结果,验证了我们的推断。

表3 风险因素的logistic回归论文原数据

变量	OR	OR 95% CI	P 值
出血量	0.997	0.994 ~ 1.000	0.05
血流稳定性	8.817	5.210 ~ 9.661	0.901
伴有DIC	0.107	0.014 ~ 0.835	0.033
血管造影发现出血	2.369	1.625 ~ 3.723	0.665

3 讨论与建议

3.1 进一步发掘和利用统计规律,为防范学术不端提供利器 虽然政府、学术界和民间都非常关心学术不端问题,也在不断地研究和下发文件,但事实上是,学术不端问题改善并不明显。据廖杉等^[11]报道,论文代发现象与日俱增,被代理期刊越来越多,期刊由非核心期刊向部分核心期刊扩散,代发论文的队伍趋向于专业化,产业链的利润持续走高。由于代发论文为枪手所写,而枪手为了多出成果,显然不会有时间去做实验或调查研究,所提供数据必然是捏造或抄袭篡改的,这样的数据在统计学上经常是经不起推敲的。鉴于此,期刊有必要加强对来稿中数据的审查。然而,关于如何对数据进行审查的文献较少,一般编辑常无从下手。笔者在统计学审查的过程中发现,掌握一些基本的统计学规律,利用一些简单的统计学结果复核工具,就能发现来稿论文的不少数据问题——本文表1~表3的数据不论其原始数据是否真实,统计学结果肯定是有问题的。本文提供的几种利用统计学规律发现数据造假的方法,包括简单统计检验方法的EXCEL模板数据验证法、利用统计量分布值与等值关系规律法、利用效应区间值与等值关系规律法等,有助于医学期刊编辑在编辑实践中发现一些简单和明显的数据造假的论文,也有助于有统计学基础的编辑开拓思想,进一步发现更多规律,供同行利用,为防范学术不端提供更多利器。对于一些比较复杂的或隐蔽的数据造假,期待同行发掘更多的规律、更好或更简捷的方法,并利用这些

方法来发现和防范学术不端论文。

3.2 提升编辑统计学审查意识,提高发现数据造假的能力与水平 不论期刊编辑是否有相应的统计学知识,编辑对于数据审查的意识都是至关重要的。吴艳妮等^[12]提出,对于稿件修改前后存在数据变更、调查样本量较大、统计检验量值与 P 值不符,以及结果部分数据与前文不符等的稿件须提高警惕,最好要求作者提供原始数据统计表格和统计结果图供编辑分析识别。这就是审查意识,只有意识到位,才有可能发现问题,才有可能最终解决问题。虽然说对于以上所提的内容,期刊编辑可以不必再深入学习统计学知识就基本可以掌握,但是提升统计学素养,无疑对于更好地理解以上内容、更主动地审查论文数据、发现更多的审查数据的方法与规律是有益处的。国内医学期刊编辑中已有不少发表过重视医学论文统计学问题的文章^[7-9,13],不过,相对期刊编辑群体来说,掌握统计学知识的编辑比例还是偏小。国内已发表医学论文仍然存在较多统计学问题也从反面说明,我们科技期刊,尤其是医学期刊编辑掌握统计学知识的比例太小了。为此,笔者建议:1)科技期刊编辑行业的相关会议多开展统计学相关知识的培训和讲座;2)期刊编辑自觉地研修统计学知识,甚至攻读统计学专业的更高学位;3)把统计学家请入编辑部进行统计学知识培训;4)期刊编辑经常与统计学家进行交流,提升统计学意识与技能。

3.3 改革论文评审流程,增设数据与统计学审查关

在统计数据的审查过程当中,有一个很重要的条件,那就是论文提供了统计检验量值,如 t 值、卡方值等以及相应精确 P 值。如果没有这个条件,审查的难度就大了许多。我们的前期研究^[9]发现,国内外对于论文是否需要提供统计检验量值差别很大,国内专家多数主张必须提供,国外的文章则较少提供,即使最近的CONSORT声明^[15]也只是提到了要提供效应大小及其区间值,而未提及统计检验量值。我们猜测,原因可能在于国内论文统计学结果造假过多,因此,必须提供统计检验量以供统计学家审查核验^[9]。这样,首先就需要期刊在稿约中声明来稿必须提供统计检验量值,或者提供国内统计学报告项目自查清单^[15]供作者自查。

然而,仅靠作者提供统计检验量值和自查报告项目显然是不够的。国外一直对医学论文的统计学问题比较关注,因此于20多年前就开始研究提高医学论文统计学报告质量的方法,也研制出了CONSORT清单等。然而,数据本身的问题不能由报告项目解决,因此,2014年《Science》和《Nature》都相继推出了专门的数据审查流程,而《Science》的数据编辑委员会由7名

科学家组成,专门负责论文的数据审查工作,其委员之一的帕玛嘉尼教授认为,数据编辑必将是期刊未来发展趋势^[10]。澳大利亚的 Vaux 教授^[16]对科技期刊(细胞与分子生物学)的数据错误用“触目惊心”(alarming)来形容,认为其对审稿专家提出了更高的要求,必须具备数据分析能力。为此,应当提倡和建议有条件的期刊编辑部在评审论文时增加数据审查或统计学审查的流程,为节约计,也可以在主编定稿之后对拟录用稿件进行数据和统计学审查,而不必与同行评议同步进行。数据与统计学审查人员最好是数理或统计学专家,如果编辑部有统计学专业人员,则统计学编辑也可胜任。

4 结论

本文从期刊编辑发现论文数据造假的方法入手,分享了几种发现与验证论文数据造假的方法,为防范学术不端体系增加了利器,期待更多编辑同行与统计学专家发现相关规律与分享成果。如欲发现论文数据造假,期刊编辑首先必须提升数据审查意识,其次提升统计学素养与能力。对于期刊编辑部,欲发现数据造假,还必须要求来稿提供统计检验量值与精确 P 值。随着数据问题的增多,建议科技期刊增设数据与统计学审查流程,加强对数据的把关力度。

5 参考文献

- [1] 秦惠基. 强化文稿发表的编辑控制机制: 一件科学丑闻的教训[J]. 编辑学报, 1992, 4(3): 140
- [2] 杨杨. 强化科学论文造假控制机制的举措: 舍恩事件的反思[J]. 编辑学报, 2002, 15(6): 450

- [3] 余毅, 胡澜, 张凌之. 高校科技期刊学术不端防范体系研究[J]. 科技与出版, 2013(3): 110
- [4] 孔艳, 张铁明. 学术不端研究综述及建立遏制学术不端的“第三类法庭”[J]. 编辑学报, 2013, 25(5): 422
- [5] 刘清海. 科技期刊编辑在防范学术造假论文中的作用[J]. 编辑学报, 2014, 26(3): 258
- [6] 付中静. 全球合作撤销国际学术论文的原因及净化效果对比分析[J]. 情报杂志, 2017, 36(6): 61
- [7] 史红, 姜永茂, 游苏宁. 重视医学论文中有关统计学分析的描述[J]. 编辑学报, 2006, 18(3): 189
- [8] 邱芬, 曾令霞, 国荣. 统计学审查在医学论文审稿中的必要性[J]. 中国科技期刊研究, 2011, 22(4): 574
- [9] 刘清海, 方积乾. 国内外医学论文统计学报告质量的比较研究[J]. 中国科技期刊研究, 2008, 19(2): 236
- [10] 郜书楷. 《科学》杂志怎么应对数据造假[J]. 现代出版, 2016(4): 76
- [11] 廖杉, 龙云飞, 秦曰龙. 学术期刊论文代发产业链的调查分析[J]. 中国科技期刊研究, 2012, 23(4): 606
- [12] 吴艳妮, 周春兰. 科技期刊编辑对学术不端论文的识别: 以《护理学报》为例[J]. 编辑学报, 2015, 27(4): 361
- [13] 王玖, 徐天和, 祁爱琴, 等. 医学论文统计学误用及其防范对策[J]. 编辑学报, 2002, 14(6): 417
- [14] SCHULZ K F, ALTMAN D G, MOHER D, et al. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials[J]. BMJ, 2010, 340:322
- [15] 刘清海, 方积乾. 应重视医学临床试验论文统计学问题: 统计学报告项目自查清单的研制[J]. 中华医学杂志, 2007, 87(34): 2446
- [16] VAUX D L. Know when your numbers are significant[J]. Nature, 2012, 492(7428): 180
(2018-04-13 收稿;2018-05-02 修回)

使用“同人”还是“同仁”?

近 10 多年来,《编辑学报》中屡见“编辑同人”;作者原稿中写作“编辑同仁”,到刊出时均改为“编辑同人”。常有同人提出异议,认为“同人”错了,应该用“同仁”;还有同人在自校文稿时直接把“同人”改为“同仁”。那么,到底该用“同人”还是“同仁”呢?

以《现代汉语词典》(第 6 版)为参考依据,该版中词条“同人”下的释义为:“称在同一单位工作的人或同行业的人。也作同仁。”“同仁”词条下的释义为:“同‘同人’。”《辞海》也有相同的表述:“同人”为“志

趣相同或共事的人”,“亦作‘同人’”。这就很清楚,“同人”“同仁”同义,但“同人”是首选词,“同仁”为“也作”,是非推荐词。《编辑学报》统一使用了首选词“同人”,应该得到鼓励。当然,如果有报刊还在使用“同仁”,只要是统一的,也不宜说人用错了。不过在这里还是要强调,你在《编辑学报》上发表文章,请遵守它的规范使用“同人”。

(郝 远)