医学论文中成组 t 检验 P 值错误及其原因分析

■相丹凤[1)]　高　永[2)]　周英智[3)]

收稿日期:2018-07-30
修回日期:2018-10-10

1)《医学综述》杂志社,北京市通州区北苑通典铭居 F 座 806 室　101100
2)滨州医学院《中国医院统计》编辑部,山东省烟台市莱山区观海路 346 号　264003
3)山东大学科技期刊社,山东省济南市山大南路 27 号　250100

摘　要　**【目的】**调查医学论文中成组 t 检验的 P 值错误,分析错误原因,并提出相应措施。**【方法】**选取 236 种医学期刊,每种期刊选取 1 项成组 t 检验,核验其 P 值,应用 χ^2 检验、Mantel-Haenszel 法、二项式 logistic 回归分析 P 值错误。**【结果】**236 项成组 t 检验中,50 项存在 P 值错误。单因素分析结果显示,核心期刊与非核心期刊相比,P 值错误发生率差异具有统计学意义($\chi^2=4.871$,$P=0.027$);给出具体 P 值组与未给出具体 P 值组相比,P 值错误发生率差异具有统计学意义($\chi^2=15.440$,$P<0.0001$)。将是否给出具体 P 值作为混杂因素,比较核心期刊与非核心期刊 P 值错误发生率,差异无统计学意义($\chi^2=2.703$,$P=0.100$)。多因素分析结果显示,是否方差齐(OR 值为 0.470,95%CI 为 0.230~0.961)、是否给出具体 P 值(OR 值为 5.459,95%CI 为2.311~12.895)具有统计学意义。**【结论】**医学论文成组 t 检验 P 值错误较多。为及时发现 P 值错误,期刊编辑应当重视对统计学方法应用条件的审查,要求作者给出统计描述以及统计推断的具体结果,能够利用简单易学的统计学软件核实 P 值。

关键词　医学论文;成组 t 检验;P 值错误;原因分析;方差齐性
DOI:10.11946/cjstp.201807300679

P 值是医学论文重要的统计学内容,是最终准确推断结论的重要依据,如果 P 值出现错误,就会严重影响对结果的正确判断,甚至得出与事实相反的结论。美国统计学会理事会于 2016 年发表了 P 值声明[1],引起了学者们对 P 值更高度的重视。t 检验是最常用的统计方法之一[2-3],编辑同仁对医学论文中 t 检验的应用情况进行了大量调查,发现其中存在很多问题,例如对不符合正态分布或方差不齐的 2 组资料进行成组 t 检验[4-5],使用 t 检验处理多组间或重复测量设计的资料等[2,6-7]。高永等[8]研制了基于 Excel 的统计分析系统,对于完全随机设计的 2 个样本均数的比较,可以输入论文中样本例数、均数、标准差,系统根据方差齐性检验结果自动选择 t 或 t' 检验,并且给出具体的 t(或 t')值和 P 值。笔者在工作中利用该系统检验来稿的数据时,发现 t 检验中 P 值错误较多。查阅相关文献,发现有关 t 检验 P 值判断错误的系统调查报道较少。所以,本研究以成组 t 检验(又称独立样本 t 检验)为例,利用该系统调查公开发表的医学论文中出现 P 值错误的情况,并分析导致 P 值错误的原因,以期引起编辑同仁对统计学结果错误的高度重视,并采取有力措施来提高医学论文的统计学质量。

1　资料与方法

1.1　资料来源

对中国知网数据库收录的 296 种综合医药卫生期刊(http://navi.cnki.net/KNavi/Journal.html#)按下列条件进行检索:全文出现“成组 t 检验”“独立样本 t 检验”或“t 检验”,发表时间为 2014—2018 年。按时间顺序排序,每种期刊尽可能选用最新发表的 1 篇论文。每篇论文选择 1 项成组 t 检验,录入 2 组例数、均数、标准差及 t 值、P 值。若原文中没有明确说明成组 t 检验或独立样本 t 检验,只说明使用 t 检验,笔者将对其进行核实,确保所用方法为成组 t 检验,排除配对 t 检验、秩和检验等。

1.2　方差齐性判断及论文中 P 值的核验

具体方法见文献[8]。因为本研究所纳入的成

作者简介:相丹凤(ORCID:0000-0002-6766-6881),学士,编辑,E-mail:1244820516@qq.com;周英智,博士,编审。
通信作者:高　永(ORCID:0000-0002-3564-4880),学士,编审,E-mail:bygaoy@126.com。

组 t 检验均未提及单侧检验,所以 P 值采用双侧检验结果。如果测算的 t(或 t')值和 P 值与论文中差别较大,则怀疑论文中的数据错误。考虑到“四舍五入”的情况,利用均数、标准差计算的结果与利用原始数据计算的结果可能存在一定偏差,但是差别不应太大。进一步利用该系统测算 t(或 t')值和 P 值因均数、标准差的“四舍五入”造成的波动范围,如果论文结果在这个范围外,则确定原文结果错误。例如某研究设观察组、对照组各 60 例,观察组麻醉时间为(66.5±2.7) min,对照组为(68.4±3.1) min,2 组比较 $t=0.721$,$P>0.05$[9]。利用软件对上述结果进行核验,2 组资料方差齐,$t=3.580$,$P=0.0005$(双侧),与论文结果差别较大。根据 2 个样本均数比较 $t(t')$ 检验计算公式,在样本数量不变的情况下,2 个均数差值越大、标准差越小,则 t(或 t')越大、P 值越小,反之亦然。考虑到“四舍五入”因素,均数66.5 的精确值在 66.45 与 66.55 之间,同理可以给出其他均数、标准差的精确值所在范围。据此,可以计算出因均数、标准差的“四舍五入”造成的 P 值波动范围为 0.0002 ~ 0.0011,由此可以确定原文 $P>0.05$ 错误。

1.3　纳入分析的因素

(1)是否为核心期刊[10]。有学者比较核心期刊与非核心期刊文献数量增长速度[11]、篇尾空白处理[12]等方面的差别,受此启发,笔者尝试比较核心期刊与非核心期刊成组 t 检验 P 值错误的比例。(2)样本量大小。成组 t 检验要求资料符合正态分布以及方差齐。样本量较大时,对非正态分布、方差不齐的 2 组资料比较采用成组 t 检验,可能对结果影响不大,但对于小样本资料结果影响较大。(3)方差齐性。2 组定量资料比较时,如果方差不齐,不能采用成组 t 检验,而应当采用 t' 检验或秩和检验,如果误用成组 t 检验,则会影响 P 值。(4)t 值及具体 P 值。医学论文应当给出确切的统计量和 P 值,医学论文中缺少具体统计量及 P 值的问题已经引起了编辑同仁的重视[13-14]。本研究尝试分析是否给出 t 值及具体 P 值与 P 值错误是否有关。

1.4　统计学处理

采用 SPSS 22.0 软件进行数据处理,采用相对数对 P 值错误进行表述,应用 χ^2 检验对 2 组间差异进行单因素分析,利用Mantel-Haenszel 法进行分层分析,采用二项式 logistic 回归进行多因素分析,检验水准 $\alpha=0.05$(双侧)。

2　结果

2.1　基本情况

296 种期刊中,除去停刊、近年未被收录、未检索到合适论文等 60 种期刊,纳入统计期刊共 236 种,每种期刊选择 1 项成组 t 检验。其中 2014 年 1 项,2015 年 4 项,2016 年 14 项,2017 年 83 项,2018 年 134 项;50 项存在 P 值错误,占比 21.19%。

2.2　P 值错误单因素分析

2.2.1　是否为核心期刊

236 项成组 t 检验中,109 项来自核心期刊,占比 46.19%,127 项来自非核心期刊,占比 53.81%。核心期刊 P 值错误共 30 项,占比 27.52%,非核心期刊 P 值错误共 20 项,占比 15.75%,2 组差异具有统计学意义。

2.2.2　样本量大小

根据文献[15]的方法,将 2 组中至少 1 组样本量≤60 定义为小样本资料。236 项成组 t 检验中,43 项为大样本资料,占比 18.22%,193 项为小样本资料,占比 81.78%。大样本组 P 值错误 9 项,占比 20.93%,小样本组 P 值错误 41 项,占比 21.24%,2 组差异无统计学意义。

2.2.3　方差齐性

236 项成组 t 检验中,方差齐 171 项,占比 72.46%,方差不齐 65 项,占比 27.54%。方差齐组的 P 值错误 31 项,占比 18.13%,方差不齐组的 P 值错误 19 项,占比 29.23%,2 组差异无统计学意义。

2.2.4　是否给出 t 值

236 项成组 t 检验中,给出 t 值 152 项,占比 64.41%,未给出 t 值 84 项,占比 35.59%。给出 t 值组 P 值错误 34 项,占比 22.37%,未给出 t 值组 P 值错误 16 项,占比 19.05%,2 组差异无统计学意义。

2.2.5　是否给出具体 P 值

总体分为给出和未给出具体 P 值,前者包括 P 值为 0.00、0.000、0.0000 和其他具体值。实际 P 值并不等于 0,当 P 值太小时,统计软件会四舍五入为 $P=0.0000$,在论文中应描述为 $P<0.001$ 或 $P<0.0001$[16-18]。因此,将二者也归为给出具体 P 值。其他为未给出具体 P 值,包括 $P<0.01$、$P<0.05$、$P>0.05$、$P>0.1$。236 种期刊中,给出具体 P 值 126 项,占比 53.39%,未给出具体 P 值 110 项,占比 46.61%。给出具体 P 值组中 P 值错误 39 项,占比 30.95%,未给出具体 P 值组中 P 值错误 11 项,占比

10.00%,2 组差异具有统计学意义。

以上单因素分析结果见表 1。

表 1 236 项成组 *t* 检验 *P* 值错误单因素分析

因素	类别	*P* 值正确		*P* 值错误		χ^2	*P*
		数量 /项	占比 /%	数量 /项	占比 /%		
核心期刊	是	79	72.48	30	27.52	4.871	0.027
	否	107	84.25	20	15.75		
样本数量	大	34	79.07	9	20.93	0.002	0.964
	小	152	78.76	41	21.24		
方差齐	是	140	81.87	31	18.13	3.477	0.062
	否	46	70.77	19	29.23		
给出 *t* 值	是	118	77.63	34	22.37	0.357	0.550
	否	68	80.95	16	19.05		
给出具体 *P* 值	是	87	69.05	39	30.95	15.440	<0.0001
	否	99	90.00	11	10.00		

2.3 *P* 值错误分层分析

将是否给出具体 *P* 值作为混杂因素,采用 Mantel-Haenszel 分层分析法比较核心期刊与非核心期刊 *P* 值错误发生率,结果表明差异无统计学意义($\chi^2=2.703$,$P=0.100$)。

2.4 *P* 值错误多因素分析

各变量的赋值情况见表 2。将上述因素均纳入模型,得到 236 项成组 *t* 检验 *P* 值错误二项式 logistic 回归分析结果(表 3)。可以看出,是否方差齐(*OR* 值为 0.470,95%*CI* 为 0.230~0.961)、是否给出具体 *P* 值(*OR* 值为 5.459,95% *CI* 为 2.311~12.895)具有统计学意义。

表 2 各变量的赋值情况

因素类型	因素	赋值	
		是	否
因变量	*P* 值是否错误	1	0
自变量	是否核心期刊	1	0
	是否为大样本	1	0
	是否方差齐	1	0
	是否给出 *t* 值	1	0
	是否给出具体 *P* 值	1	0

表 3 236 项成组 *t* 检验 *P* 值错误二项式 logistic 回归分析

因素	*b*	S_b	Wald χ^2	*P*	*OR*	95%*CI*
是否核心期刊	0.659	0.343	3.695	0.055	1.932	0.987~3.782
样本量大小	−0.162	0.445	0.132	0.716	0.851	0.356~2.034
是否方差齐	−0.754	0.364	4.282	0.039	0.470	0.230~0.961
是否给出 *t* 值	−0.595	0.426	1.955	0.162	0.551	0.239~1.270
是否给出具体 *P* 值	1.697	0.439	14.978	<0.0001	5.459	2.311~12.895

3 原因分析与建议

本研究发现,医学论文成组 *t* 检验中 *P* 值错误发生率高达 21.19%,严重影响了论文的学术质量,需要引起高度重视。其可能原因主要包括:统计分析软件操作失误;写作过程中 *P* 值笔误;写作过程中样本例数、均数、标准差数据笔误,造成核验 *P* 值本身错误而误判;统计方法不当,例如符合正态分布但方差不齐时没用 *t*′检验;手工计算错误;排版错误;数据造假等。为避免成组 *t* 检验 *P* 值错误,提高医学论文的统计学质量,提出以下建议。

(1) 重视成组 *t* 检验的应用条件。进行成组 *t* 检验,特别是样本量较小时,用于 2 组比较的资料必须符合正态分布。笔者在收集研究资料的过程中发现,资料不符合正态分布的情况较为常见。例如某研究采用成组 *t* 检验比较胆管癌和胆总管结石患者血清 CA199 水平,2 组数值分别为(413.09±355.35) U/mL和(183.48±322.24) U/mL,标准差接近甚至超过均数,初步可以判断为非正态分布[19]。应当首先对数据进行正态分布检验,若为非正态分布,改为中位数及四分位间距描述,采用 Wilcoxon 秩和检验进行 2 组比较[20-21]。

成组 *t* 检验的另一个应用条件是方差齐,如果符合正态分布但方差不齐应该取 *t*′检验的 *P* 值。但本组资料方差不齐的比例高达 27.54%,均未提及采用 *t*′检验,由此推测,许多方差不齐的 2 组比较很可能采用的是成组 *t* 检验的 *P* 值,导致 *P* 值不精确甚至错误。二项式 logistic 回归分析结果也显示,方差齐减少了 *P* 值错误的可能性。如果统计学方法选择错误,统计学处理结果的正确性将无从谈起,因此编辑审核稿件时,一定要首先审核所用的统计学方法是否正确。

(2) 要求论文作者给出观察指标的描述分析。例如比较 2 组正态分布的资料时,要求作者给出样本量、均数、标准差等指标,这是论文写作的基本要求,也便于利用这些数据核实统计推断结果。如果觉得结果可疑,可以请作者提供原始数据,通过统计学软件进行核查。

(3) 要求论文作者给出统计量和具体 P 值。医学论文要给出确切的统计量和 P 值,包括中华医学会系列杂志在内的许多医学期刊都对此做出了明确要求[22-23]。但本研究发现,236 项成组 t 检验中,未给出 t 值和具体 P 值的比例分别高达35.59%和46.61%。不给出 t 值和具体 P 值,不利于判断 2 组比较统计学差异的具体程度。另外,本研究结果显示,给出具体 P 值的论文中 P 值错误发生率较高,主要因为本研究所用的判别方法更容易发现具体 P 值的错误。例如文献[24]比较痛经女性组与正常女性组经期 SCL-90 各因素的均值,其中"强迫"一项的 P 值为 0.003,笔者利用文中数据测算的结果是 P 值为 0.0003,波动范围为 0.0003~0.0004,因此判断原文 P 值错误。如果原文给出的不是具体值,而是 $P<0.01$ 或 $P<0.05$,则不会判为错误。

4　结语

统计学处理是医学论文的重要内容,统计学结果错误将严重影响论文的学术质量。利用基于 Excel 的统计分析系统核验了中国知网收录的综合医药卫生期刊中成组 t 检验的 P 值,发现 P 值错误较多,必须引起高度重视。期刊编辑应当重视对统计学方法应用条件的审查;要求作者给出统计描述以及统计推断的具体结果,必要时请作者提供原始数据,通过统计学软件进行核查,严防统计数据造假等学术不端现象;可以利用简单易学的统计学软件核实 P 值;做好校对工作,及时发现排版导致的错误。由于本研究只调查了综合医药卫生期刊,结果可能与国内医学期刊的整体情况有一定出入,有待扩大范围做进一步的深入研究。

参考文献

[1] Wasserstein R L. ASA 关于统计意义和 p-值的声明[J]. 方积乾,译. 中国卫生统计,2016,33(3):549-552.

[2] 马莉,高贵现.《山东医药》杂志常见统计学错误及分析[J]. 编辑学报,2016,28(1):32-34.

[3] 陈姗姗,孙琴. 学术不端论文的几大隐性特征及其辅助鉴别方法——以医学科研论文为例[J]. 湖北师范大学学报(自然科学版),2018,38(3):72-76.

[4] 陈文娟,汤雷,马莉. 医学期刊常见 t 检验应用错误及案例分析[J]. 编辑学报,2016,28(3):237-239.

[5] 陈景景,谭晓蕾,徐晓静. 护理期刊来稿常见统计学问题及其对策分析[J]. 科技与出版,2017(2):87-91.

[6] 刘静. 统计学方法的正确使用问题(一)[J]. 心肺血管病杂志,2017,36(1):76.

[7] 田云鹏,陈丽. 医学论文中常见统计学错误例析[J]. 中华全科医学,2017,15(10):1791-1794.

[8] 高永,张中文,石德文,等. 基于 Excel 的统计分析系统在期刊编辑部审稿中的应用[J]. 编辑学报,2013,25(5): 478-481.

[9] 潘洪秀,朱晓文,季方茹,等. 右美托咪定联合芬太尼对高血压脑出血患者术后脑糖氧代谢及 Toll 样受体表达的影响[J]. 中国老年保健医学,2018,16(1):23-27.

[10] 中国科学技术信息研究所. 2017 年版中国科技期刊引证报告(核心版)[M]. 北京:科学技术文献出版社,2017.

[11] 夏成锋,林江娇. 2007—2013 年中国科技期刊主要计量指标统计分析[J]. 中国科技期刊研究,2016,27(8):900-903.

[12] 袁丽霞,陈雯,黄明睿. 动物科技类期刊补白的统计与分析[J]. 编辑学报,2014,26(6): 550-552.

[13] 冉明会. 医学期刊编辑应重视摘要中统计学著录问题的审编[J]. 编辑学报,2014,26(3):238-240.

[14] 接雅俐,陈汐敏,顾璟,等. 医学论著统计学报告水平评价量表的制作与初步应用[J]. 中国科技期刊研究,2012,23(6): 982-987.

[15] 孙振球,徐勇勇. 医学统计学[M]. 4 版. 北京:人民卫生出版社,2014:37.

[16] 吴艳妮,周春兰,江霞,等. 国内护理学统计源期刊论文中报告精确 P 值常见错误:$P=0.000$[J]. 编辑学报,2016,28(2): 133-134.

[17] 冯国双,罗凤基. 医学案例统计分析与 SAS 应用[M]. 北京:北京大学医学出版社,2011:55.

[18] 姜春霞. 论医学期刊编辑的统计学审核[J]. 中国科技期刊研究,2014,25(6):782-784.

[19] 陈攀,冯留顺,潘洁,等. CA199 和总胆红素水平在胆管疾病应用中的临床分析[J]. 医药论坛杂志,2018,39(2):62-64.

[20] 张军锋,董海原. 医学论文审稿中常见的统计学错误:定量资料统计方法的误用分析[J]. 中国药物与临床,2017,17(10): 1558-1560.

[21] 周英智,靳光华. 利用文中数据识别统计学错误[J]. 编辑学报,2016,28(1):29-31.

[22] 《中华神经创伤外科电子杂志》编辑部. 中华医学会系列杂志对来稿中统计学处理的有关要求[J]. 中华神经创伤外科电子杂志,2018,4(1):28.

[23] 《郑州大学学报(医学版)》投稿指南[EB/OL]. [2018-05-03]. http://jms.zzu.edu.cn/Corp/fixed.aspx? id=40.

[24] 冯思仪,程兰. 痛经的精神心理因素探讨[J]. 暨南大学学报(自然科学与医学版),2017,38(5):437-442.

作者贡献声明:

相丹风:收集分析数据,撰写论文;

高　永:提供基于 Excel 的统计分析系统,指导资料分

析,修订论文;

周英智:确定研究方向,修订论文。

P-value errors in two-sample *t*-test of medical papers and reason analysis

XIANG Danfeng[1)], **GAO Yong**[2)], **ZHOU Yingzhi**[3)]

1) Publishing House of *Medical Recapitulate*, F806 Tongdianmingju, Beiyuan, Tongzhou District, Beijing 101100, China
2) Editorial Office of *Chinese Journal of Hospital Statistics*, Binzhou Medical University, 346 Guanhai Road, Laishan District, Yantai 264003, China
3) Shandong University Scientific Journals Press, 27 Shanda Nanlu, Jinan 250100, China

Abstract: [**Purposes**] This paper investigates *P*-value errors in two-sample *t*-test of medical papers, analyzes the causes of errors, and puts forward the corresponding measures. [**Methods**] We reviewed 236 two-sample *t*-test cases from 236 medical journals. The *P*-value was verified and the *P*-value errors were analyzed by χ^2 test, Mantel-Haenszel method and binomial logistic regression. [**Findings**] Among the 236 cases, 50 cases have *P*-value errors. The univariate analysis shows that there is a significant difference between core journals and non-core journals in incidence of *P*-value errors ($\chi^2 = 4.871$, $P = 0.027$), and there is a significant difference between cases with concrete *P*-value and cases without concrete *P*-value in incidence of *P*-value errors ($\chi^2 = 15.440$, $P < 0.0001$). When we regard whether the concrete *P*-value is described as the confounding factor, no significant difference in the incidence of *P*-value errors is found between core journals and non-core journals ($\chi^2 = 2.703$, $P = 0.100$). The multivariate analysis shows that whether the variance is equal ($OR = 0.470$, 95% CI: 0.230-0.961) and whether the concrete *P*-value is described ($OR = 5.459$, 95% CI: 2.311-12.895) are significant variables. [**Conclusions**] There is a high incidence of *P*-value errors in two-sample *t*-test of medical papers. In order to discover the *P*-value errors in time, editors should pay attention to the examination of the application conditions of statistical methods, ask the authors to give statistical description and statistical inference result, and verify *P*-value by simple statistical software.

Keywords: Medical paper; Two-sample *t*-test; *P*-value error; Reason analysis; Homogeneity of variance

(本文责编:梁永霞)