生物医学期刊中常见的 *P* 值使用错误

陈 章 颖
中山大学附属第一医院期刊中心,510080,广州

摘 要 在生物医学期刊的编辑工作中,时常可以发现稿件中 *P* 值的误用。文章分析稿件中 *P* 值的意义、报告、描述和解释的误用情况及其对文章科学性的影响,提出相应的解决方法,并指出置信区间对 *P* 值局限性的补充作用。认为生物医学期刊编辑应当增强统计学观念,在处理稿件时严把统计学质量关,以达到提高期刊学术质量的目的。

关键词 生物医学期刊;统计学错误;*P* 值

Common misuses of the *P* value in the editing work of biomedical journals//CHEN Zhangying

Abstract In the editing work of biomedical journals, statistical errors can be found frequently, of which the most common is the misuse of the *P* value. This paper analyzes the misuse of the *P* value from its significance, report, description and explanation as well as its effect on science of the papers. We also propose the appropriate solutions, and point out that the confidence interval can play a complementary role to the limitations of *P* value. Editors of biomedical journals should develop a vigilant mind on statistics, guarantee the quality of statistics while dealing with the manuscripts, which could increase the submission of high quality papers and improve the academic quality of the journal.

Keywords biomedical journal; statistical error; *P* value

Author's address Editorial Department of the First Affiliated Hospital of Sun Yat-sen University, 510080, Guangzhou, China

2004 年,西班牙统计学家 Garcia-Berthou 等查阅了 2001 年《自然》和 2002 年《BMC 医学研究方法学》刊登的论文,结果发现其中分别有 38% 和 25% 的文章出现至少 1 处统计学错误[1]。笔者在生物医学科研论文的编辑工作中,经常看到作者虽然使用统计学方法对研究资料进行分析,但由于对统计学的理解和表述不恰当或错误,导致得出与实际结果不一致的结论,影响论文的科学性。而在统计学运用的错误中,最常见的是在显著性检验中对 *P* 值的误用。

1 *P* 值的意义

P 值是指在零假设(或称为无效假设、原假设或检验假设)成立的前提下,出现目前样本数据对应的统计值(如 *t* 值、*F* 值、χ^2 值等)乃至比它更极端数值的概率。统计学上根据显著性检验原理推算出来的 *P* 值的实际意义是表示研究拒绝特定零假设时可能犯Ⅰ类错误(即假阳性)概率的理论值。$P<0.05$ 说明样本之间的差异来源于抽样误差的可能性小于 5%(小概率事件),根据统计学原理,在一次抽样过程中小概率事件是不可能发生的;因此,可以认为差异是由于样本数据对应的总体之间存在差别所致,而不是由于抽样误差导致的[2]。

部分作者不理解显著性检验的基本思想以及 *P* 值的意义,没有对所观察的样本数据进行统计学计算,而是单纯根据其差别大小直接得出 $P<0.05$ 或 $P>0.05$ 的推断,导致文章的结论出现错误。而编辑在处理稿件时,有时会有先入为主的观念,认为作者已经对数据进行统计分析,仅根据数值的大小简单地审阅稿

率也相应增多;但其设计和编排格式各异,存在很多问题。笔者在分析已检索的科技期刊中正交试验表使用现状的基础上,针对存在的问题,提出了自己的观点,并从编辑和出版的角度给出了正交试验表的组合应用模式,即因素水平表、正交试验设计方案与结果表、极差分析表、方差分析表,这 4 个表格须同时存在。此组合模式科学与否,敬请同人评析。

5 参考文献

[1] 郝拉娣,于化东. 正交试验设计表的使用分析[J]. 编辑学报,2005,17(5):334-335
[2] 刘瑞江,张业旺,闻崇炜,等. 正交试验设计和分析方法研究[J]. 实验技术与管理,2010,27(9):52-55
[3] 程红,李莉. 学术期刊正交试验类稿件的审读方法[J]. 编辑学报,2012,24(5):450-452
[4] 陈浩元. 科技书刊标准化 18 讲[M]. 北京:北京师范大学出版社,1998:135
[5] 王秀丽. 科技论文三线表中的常见问题分析[J]. 编辑学报,2006,18(4):267-268
[6] 田军,王新英,刘文革. 科技论文表格常见错误评析[J]. 编辑学报,2005,17(6):421-422
[7] 郭青,李小萍,梁秋野. 医学期刊表格设计编排原则及常见问题[J]. 编辑学报,2011,23(4):335-336
[8] 《正交试验设计法》编写组. 正交试验设计法[M]. 上海:上海科学技术出版社,1979:13
[9] 郝拉娣,张娴,刘琳. 科技论文中正交试验结果分析方法的使用[J]. 编辑学报,2007,19(5):340-341

(2013-08-22 收稿;2013-11-20 修回)

件,未能纠正错误;因此,在工作中需要对作者报告的统计值进行核算,避免错误的统计学推论发表。此外,由于 P 值是对样本数据对应的统计值概率的估计,在文章中应该注明所采用的统计学方法,并列出统计值,而不能只单纯写出 P 值的数值或范围。

2 P 值的报告

过去由于受到条件的限制,需要在计算出统计量后手工查询相应的统计界值表来估计 P 值的范围,因此在论文中仅需要报告 $P<0.05$ 或 $P>0.05$ 等;但是,随着计算机和统计分析软件的发展和普及,现在已经可以便捷地计算出精确的 P 值。

报告精确的 P 值,在实战中很有意义。在统计分析中可能会遇到接近临界值如 $P=0.051$ 或 $P=0.049$ 的情况,两者的统计学意义差别并不是很大,往往需要研究者根据研究的具体情况和专业知识进行分析来推断结果是否有统计学意义,而不能一概而论[3]。

报告精确的 P 值可以促使作者或者其他研究者在后续研究中根据犯Ⅰ类错误的概率、检验效能和专业知识设计研究,确定合理的样本量,以获得更为明确的推断。如果仅报告 P 值的范围,单纯根据固定的界值进行判断,则会丢失这部分信息,影响结果的科学性。此外,P 值是循证医学最重要的证据之一,某些荟萃分析方法如 Fisher 法需要根据精确的 P 值来对同类研究结果进行综合评价和分析[4]。因此,编辑在处理稿件过程中应该尽量要求作者报告出精确的 P 值。

应注意的是,在使用 SPSS 等统计软件进行计算时,如果得出的 P 值过小,就会出现计算机输出 $P=0.000\,0$ 的情况。有些作者直接将 $P=0.000\,0$ 写入稿件,这是不妥当的;因为这里的 P 值实际上并不等于0,只是受到软件显示的限制数值修约后才输出0.000 0的结果。在实际报告中,对此情况应该写成 $P<0.001$ 或 $P<0.000\,1$。一般认为,科研论文中报告到 $P<0.000\,1$ 已经足够,低于 $0.000\,1$ 的 P 值不需要报告出具体数值[1]。

3 P 值的描述

在编辑工作中,经常会看到作者在显著性检验得到 $P>0.05$、$P<0.05$ 或 $P<0.01$ 时,对结果的描述分别为“差别不显著”“差别显著”“差别非常显著”,这也是错误的。

如前文所述,P 值是当零假设为真时出现样本观察结果的概率,根据小概率事件的原理,如果 P 值小于所确定的检验水准 α(通常取 0.05 或 0.01),拒绝零假设,则认为事件发生的可能性非常低,P 值越小,拒绝零假设的理由就越充分。

P 值不但和样本的实际差别有关,还取决于样本量大小、抽样误差等;因此,不能单纯从 P 值的大小判断样本实际差别的大小,总体参数间差别的大小只能根据专业知识来判断。因此,对于 $P>0.05$、$P<0.05$ 或 $P<0.01$ 的统计结果,应该使用“差别无统计学意义”“差别有统计学意义”“差别有高度统计学意义”来描述。

现在统计学学者普遍认为,既然 P 值已经可以算出精确的数值,就没有必要把 0.05 或 0.01 看作是特殊的界值,即使 $P<0.01$ 也只需描述“差别有统计学意义”即可[3]。

在早期的统计报告中,对于 $P>0.05$、$P<0.05$ 或 $P<0.01$ 也可以采用“差别无显著性意义”“差别有显著性意义”“差别有非常显著性意义”来表示。虽然这种表达方式本身并没有错误,但是经常会被缩写或误解为实际差别是否显著,从而背离 P 值的本意;因此,现在更提倡使用差别是否有统计学意义的说法来描述 P 值。

4 P 值的解释

即使显著性检验的计算和描述过程都无误,但对 P 值的解释不当也会影响文章的科学性。对 $P>0.05$ 描述为“差别无统计学意义”,其统计学含义是尚不能拒绝零假设,即当前的数据不能提供足够的证据来证明2组数据之间的差异有统计学意义;因此,从统计学的角度来说,$P>0.05$ 可得到“差别无统计学意义”的结论,但从研究的角度来说,$P>0.05$ 不能得到“2组数据没有差别”的结论[3]。例如在比较2组药物的疗效时,如显著性检验结果为 $P>0.05$,可能是由于样本量少、测量误差比较大而出现假阴性的结果,我们不能轻易下结论认为“两组药物的疗效相同”,只能说尚不足以认为2组药物之间疗效的差异有统计学意义,即“尚不能认为2组药物的疗效不同”。

1978年,Freiman 等对在《新英格兰医学杂志》等权威杂志的71篇阴性结果论文进行分析,发现其中有67篇可能是由于样本量不足等原因导致检验效能不足而造成的假阴性[6];因此,对阴性结果作出解释要谨慎,$P>0.05$ 不能作为样本基本相同的判断依据,这一点在编辑的审稿中经常被忽视。在实际研究特别是随机对照试验中,如果出现阴性结果,建议应报告研究的检验效能,为读者和后续研究者提供更多的信息。

在生物医学研究中,很多人把 P 值当作是判断科学性的“金标准”,在得出与专业理论知识相悖的阳性结果时,他们也宁愿选择相信 P 值,结果导致结论与实际规律不符甚至互相矛盾[7]。例如对某药物的降血糖效果进行评价,经过大量人群试验后发现治疗后空腹血糖和餐后2 h血糖与治疗前相比差别有统计学

意义,但我们不能轻易对该药物的疗效作出有效的结论。根据临床评价标准,空腹血糖和餐后 2 h 血糖降低≥10%才是有意义的,如果低于 10%,即使在统计学上有意义,在临床实践上也没有实际意义。

统计学上是否有意义和专业上是否有意义是 2 个不同的概念,科研者和编辑都应该正确认识显著性检验和 P 值的概念和意义,作结论时要充分结合统计结果和专业知识,坚持"专业为主,统计为辅"的观点来对待科研论文。

5　P 值和置信区间

显著性检验的 P 值在结果表达上有一定的局限性,只能说明 2 组或多组间总体参数的差别是否有统计学意义,但不能说明差别的程度有多大或者是否有实际意义。特别是当样本量较小时,样本间的差异完全可能是由随机误差引起的,而 P 值并不能显示这种随机误差。在研究中可能会由于随机误差的影响,导致出现单纯从 P 值得到样本之间存在差异的错误结论。

在循证医学研究中,经常使用置信区间(或称为可信区间)对样本的总体进行推断[8]。置信区间是按一定的概率去估计总体参数所在的范围,它按预先给定的概率(常取 95%或 99%)来确定未知参数值的可能范围,这个范围就被称为所估计参数值的置信区间。它可以用于估计总体参数,也可以用于显著性检验,95%的置信区间和 $\alpha=0.05$ 的显著性检验是等价的[9]。如果置信区间很大,说明可能存在很大的随机误差,结论具有很大的不确定性,并警示研究者应改进研究条件如扩大样本量等重新进行研究。采用置信区间来解释"差别无统计学意义"的结果也有重要的意义[10]。

例如某种药物和常规组相比,其效应量的 95%置信区间为(-5%,50%)。由于零假设(即效应量等于 0)在这个区间内,因此显著性检验作出不能拒绝零假设的结论($P>0.05$),即不能认为 2 种药物的疗效有区别;但是,我们通过置信区间可以看到这个区间包括了 50%的效应量,也就是说没有排除这种药物能比对照组增加 50%有效率的可能,这一点应该值得研究者重视。如果出现这种结果,也需要谨慎作出结论。

旨在提高随机对照临床试验报告质量的 CONSORT 声明,从 1995 年的第 1 版到 2010 年的最新版都建议在结局和评估的报告中应该描述效应估计值及其精确度(如 95%置信区间)[11-12],但目前国内的生物医学期刊对置信区间的要求并不严格;因此,在编辑工作中应重视置信区间的作用,鼓励作者报告置信区间,既可以增加文章的信息量,也可以和国外的学术期刊接轨,提高期刊的可信度。

6　结束语

生物医学论文应具有科学性、创新性和实用性,其中科学性是关键。统计学作为一种保证研究论文科学性的手段,现在已经得到人们的普遍重视。国内的生物医学期刊都很重视来稿的统计学问题,并提出相应的统计学要求,对 P 值的解释和表达也有规定;但从实际效果上看,生物医学论文统计学应用的缺陷率仍然相当高,仅在对 P 值的使用上就存在诸多问题,从而影响文章的可信度:因此,编辑应当建立良好的统计学观念,在处理稿件时严把统计学质量关,并针对来稿中的统计学错误及时与作者沟通,杜绝有统计学缺陷的文章发表;有条件时还可以举办统计学定期定向培训,强化作者的统计学意识,以增加优质的稿源,从而提高期刊的学术质量,达到双赢的效果。

7　参考文献

[1] García-Berthou E, Alcaraz C. Incongruence between test statistics and P values in medical papers [J]. BMC Med Res Methodol, 2004, 28(4): 13

[2] 林爱华,宇传华. 假设检验 [M] //方积乾. 生物医学研究的统计方法. 北京: 高等教育出版社, 2007: 80-93

[3] 徐勇勇. 医学研究报告中的统计学新概念 [J]. 中华医学杂志, 1995, 75(3): 178-182

[4] 徐勇勇,赵清波. 第十讲 如何在论文中正确表达和解释统计结果 [J]. 中华预防医学杂志, 2002, 36(4): 284-286

[5] 张弓,肖景榕. 正确理解生物统计学的 P 值 [J]. 现代肿瘤医学, 2006, 14(1): 102

[6] Freiman J A, Chalmers T C, Smith H Jr, et al. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials [J]. N Engl J Med, 1978, 299(13): 690-694

[7] 张海涛,王滨有. 浅谈流行病学中 P 值滥用的原因及其解决对策 [J]. 中国卫生统计, 2011, 28(1): 95-97

[8] 康德英,王家良,洪旗,等. 循证医学中统计结果的准确表达: P 值与可信区间 [J]. 华西医学, 2000, 15(4): 402-403

[9] 刘关键,洪旗. 可信区间的用途和意义 [J]. 中国循证医学, 2001, 1(4): 235-238

[10] 段乃华,王元佳. 显著性检验与可信区间 [J]. 上海精神医学, 2010, 23(1): 62-63

[11] 杜亮,陈耀龙,陈敏,等. 从 CONSORT 到 GPP: 医学研究报告规范简介 [J]. 编辑学报, 2008, 20(4): 367-370

[12] 汪谋岳. 2010 年新版 CONSORT 声明简介 [J]. 中国科技期刊研究, 2011, 22(2): 309-310

(2013-09-24 收稿; 2013-12-24 修回)