·医学综合研究·

医学论文中常见统计学错误例析

田云鹏，陈丽

浙江医学杂志社，浙江 杭州 310003

摘要：列举医学论文中常见的统计学问题，包括 t 检验的常见误用和 χ^2 检验的常见误用。t 检验的常见误用有误用 t 检验代替配对 t 检验、误用 t 检验代替单因素方差分析、误用 t 检验代替重复测量数据的方差分析，χ^2 检验的常见误用有误用 χ^2 检验代替 Fisher 确切概率法、误用 χ^2 检验代替秩和检验、误用 χ^2 检验代替配对 χ^2 检验。通过对医学论文中常见的统计学问题进行案例分析发现，统计学方法误用的主要原因有忽略统计学方法的应用前提，未能根据实际情况和统计分析的目的来正解选用统计学方法，未能充分理解研究资料是否满足参数检验的条件，不能正确判断计量资料所对应的实验设计类型。为了提高统计学应用的准确性，首先，应聘请统计学审稿专家对稿件进行质量把关；其次，编辑人员应加强自身统计学专业素养，提高统计学专业知识水平，在编辑稿件过程中加强对稿件中统计学处理的思考和把握，遇到有疑问的地方要及时翻阅统计学资料或请教相关统计学专家以求正确解决；最后，作者在进行科研设计时，要多与统计学专家交流合作，从源头上避免设计上的漏洞和偏差。从作者、编辑、审稿专家等三方提高统计学的应用能力，不仅能保证科研的科学性、可靠性和有效性，同时能减少对读者的错误信息的引导。

关键词：医学论文；统计学；误用

中图分类号： R195.1 **文献标识码**： A **文章编号**：1674-4152(2017)10-1791-04

DOI：10.16766/j.cnki.issn.1674-4152.2017.10.044

An analysis of common statistical errors in medical papers *TIAN Yun-peng, CHEN Li. Magazine House of "Zhejiang Medicine", Hangzhou, Zhejiang* 310003, *China*

Abstract: The common statistical mistakes seen in medical researches, including common misuse of t test and χ^2 test, are enumerated in this paper. The common misuse of t test includes t test instead of paired t test, t test instead of one-way ANOVA, and t test instead of a repeated measures analysis of variance. χ^2 test is incorrectly used instead of Fisher exact probability method, rank test, or paired χ^2 test. The case analysis on the common statistical problems in medical papers shows that the main reasons for misuse of statistical methods are the application premise error of a special statistical method, incorrectly choice of statistical methods divorced from the actual situation and the purpose of statistical analysis, not fully understanding whether the research data to meet the conditions of parameter testing, and incorrectly determine the measurement data corresponding to the experimental design type. To ensure the accuracy of statistical applications in a scientific journal, the editorial department should invite statistical review experts to control the quality of the manuscript; the editors should strengthen their own statistical professional quality, improve the level of statistical professional knowledge, deeply think and grasp the statistical analysis of the manuscripts in the process of editing manuscripts, timely look over the statistical data and consult the relevant statistical expert where you have a question. Finally, the authors should have more exchanges with the statistical experts in the research design, from the source to avoid the design of the loopholes and deviations. It is suggested to improve the application ability of statistics from the authors, editors and peer review experts to ensure the scientificalness, reliability and validity of scientific research, and reduce the induction of the error messages to readers.

Key words: Medical papers; Statistics; Misuse

医学期刊是发表医学科研成果相关论文的重要平台，优质的稿件是医学期刊质量的保证。判断稿件质量的好坏，主要从专业角度和统计学分析两方面来评价。统计学为医学研究提供了数据分析的手段和方法，是医学研究不可分割的部分。迟殿元等[1]认为统计结果的正确与否与医学论文的质量密切相关。医学研究中统计学的误用将直接影响研究的结果可信度。郝丽洁等[2]对某医学期刊 413 篇稿件进行统计学使用情况分析，结果发现 255 篇在数据处理中存在统计学误用的情况，统计学误用发生率为 61.74%。本文中，笔者主要运用案例解析的方式对稿件编辑过程中碰到的常见的统计学误用进行分析，以期对读者和编辑同行有所帮助。

通信作者：田云鹏，E-mail：523406098@qq.com

1 t 检验的常见误用

t 检验的使用条件是随机样本、来自正态分布的总体、符合方差齐性。在日常编辑工作中，经常会碰到作者一看到定量资料就直接套用 t 检验。常见的 t 检验误用类型有：①误用 t 检验代替配对 t 检验；②误用 t 检验代替单因素方差分析；③误用 t 检验代替重复测量数据的方差分析。

1.1 误用 t 检验代替配对 t 检验

例 1，某研究对 2

组患者治疗前后简式 Fugl-Meyer 下肢运动量表(FMA)、临床痉挛指数(CSI)、临床神经功能缺损程度评分(NDS)和 Barthel 指数进行比较,具体数据见表1。

表1　2组患者治疗前后 FMA、CSI、NDS 和 Barthel 指数比较

组别	例数	时间	FMA 评分	CSI 评分	NDS 评分	Barthel 指数
治疗组	37	治疗前	15.17±4.35	12.26±0.76	23.14±4.54	46.02±4.46
		治疗后	33.33±4.67	4.34±1.68	7.16±3.68	68.25±3.92
对照组	37	治疗前	15.22±5.14	12.08±0.84	22.95±4.33	48.14±3.94
		治疗后	28.21±5.53	7.04±0.88	11.54±5.12	55.23±4.32

原稿中本例采用两独立样本 t 检验对组内治疗前后患者 FMA、CSI、NDS 和 Barthel 指数进行比较。但观察表1发现,组内治疗前后比较为自身对照研究,应采用配对 t 检验。而作者错误地采用两独立样本 t 检验对组内治疗前后进行比较,无形中扩大了样本例数,容易得出假阳性结果,也会增大混杂因素,降低精准度和结果的可信度。故本例应使用配对 t 检验。

1.2　误用 t 检验代替单因素方差分析　例2,某研究作者将维生素 D 营养状况分为充足、不足、缺乏和严重缺乏4组,并统计分析不同维生素 D 营养状况下骨密度情况,见表2(该表为作者投稿时原表)。

表2　不同维生素 D 营养状况时骨密度情况

维生素 D 营养状况分组	例数	骨密度
充足组	137	61.34±2.484
不足组	236	54.20±1.751[a]
缺乏组	196	50.24±1.996[a]
严重缺乏组	8	46.50±11.090

注:维生素 D 充足组与其他各营养组的骨密度比较采用 t 检验,[a]$P<0.05$。

作者原稿中统计方法是采用两独立样本 t 检验,t 检验只能用于两个独立样本均数之间的比较差异是否有统计学意义,而本例研究的是4个相互独立样本均数之间的比较差异是否有统计学意义,所以应采用单因素方差分析。若单因素方差分析结果差异有统计学意义,再进行4个样本均数之间的两两比较,本例两两比较方法可采用 Dunnett-t 检验。另外原稿直接进行 t 检验也会增加Ⅰ类错误的概率。经重新修改后,得出不同维生素 D 营养状况组间骨密度比较差异有统计学意义($F=3.671$, $P<0.05$)。经两两比较后显示,维生素 D 充足组骨密度分别高于不足组和缺乏组(均 $P<0.05$)。

1.3　误用 t 检验代替重复测量数据的方差分析　例3,某研究中作者对健康干预前后患者菌斑指数(PLI)、牙龈指数(GI)、探诊深度(PD)、附着丧失(AL)等牙周检查指数进行比较,见表3(该表为作者投稿时原表)。

表3存在以下问题:①作者采用 t 检验对健康干预后不同时间点的牙周检查指数进行比较;②表格中 a 和 b 标注未说明是哪些数据之间进行比较;③P 值含义理解错误。在编辑工作中,经常会看到有些作者用“无显著差异”“显著差异”和“非常显著差异”来形容 $P>0.05$、$P<0.05$ 和 $P<0.01$。事实上 P 值只能描述被比较对象之间差异是否有统计学意义,而不能描述被比较对象之间差异有多大。

表3　健康干预前后患者牙周检查指数比较

检查时间	PLI	GI	PD(mm)	AL(mm)
入组时	2.36±0.58	2.35±0.54	3.90±1.20	3.17±1.60
健康干预4周	2.14±0.49	2.28±0.50	3.50±1.12	2.92±1.47
健康干预8周	1.96±0.42	2.01±0.39	2.95±0.86[a]	2.84±1.02
健康干预3个月	1.45±0.31	1.52±0.34[a]	2.65±0.68	2.56±0.98
健康干预6个月	1.35±0.30[a]	1.24±0.22[b]	2.40±0.45[b]	2.29±0.62[a]
t 值[a]	2.58	2.12	2.42	2.25
P 值	<0.05	<0.05	<0.05	<0.05
t 值[b]		2.84	3.83	
P 值		<0.01	<0.01	

注:[a]表示出现显著性差异,[b]表示出现非常显著差异。

观察该组数据可发现,每例患者在不同时间点均进行了重复测量,因此应采用重复测量数据的方差分析。首先对重复测量数据进行球对称检验,当球对称检验 $P>0.05$ 时,说明数据满足 Huynh-Feldt 条件,可以直接使用 F 值;当球对称检验 $P<0.05$ 时,应采用 Huynh-Feldt 法校正方差分析结果,然后使用 Bonferroni 法进行组内各时间点两两比较。该例重复测量数据经球对称检验后,均 $P<0.05$,所以需校正方差分析结果。另外两两比较结果显示除健康干预4周 GI、PD 和 AL 与干预前比较,差异均无统计学意义(均 $P>0.05$),其他指标各时间点两两比较差异均有统计学意义(均 $P<0.05$)。修正后的统计结果见表4。

表4　健康干预前后患者牙周检查指数比较

检查时间	PLI	GI	PD(mm)	AL(mm)
干预前	2.34±0.35	2.35±0.31	3.83±0.70	3.20±0.95
健康干预4周后	2.13±0.27	2.30±0.28[a]	3.49±0.66[a]	2.95±0.80[a]
健康干预8周后	1.93±0.27	1.95±0.21	2.93±0.51	2.79±0.59
健康干预3个月	1.46±0.18	1.51±0.20	2.65±0.36	2.66±0.54
健康干预6个月	1.25±0.16	1.24±0.14	2.39±0.26	2.28±0.33
F 值	267.21	376.96	112.40	23.19
P 值	<0.05	<0.05	<0.05	<0.05

注:[a]表示单个指标健康干预4周后与干预前比较,差异均无统计学意义($P>0.05$);其他时间点各指标与干预前比较,差异均有统计学意义($P<0.05$)。

2　χ^2 检验的常见误用

在日常编辑工作中,笔者经常可以看到作者将 χ^2 检验作为计数资料分析的“万能钥匙”。有些作者甚至不考虑 χ^2 检验的适用条件,只要一看到是计数资料,就一律盲目套用 χ^2 检验。常见的 χ^2 检验误用类型有:①误用 χ^2 检验代替 Fisher 确切概率法;②误用 χ^2 检验代替秩和检验;③误用 χ^2 检验代替配对 χ^2 检验。

2.1　误用 χ^2 检验代替 Fisher 确切概率法　例4,某研

究中作者对 2 组裸鼠标本中 Aurora-A 蛋白表达情况进行分析，见表 5（该表为作者投稿时原表）。

表 5 2 组裸鼠标本中 Aurora-A 蛋白表达情况

组别	只数	Aurora-A 蛋白表达阳性	χ^2 值	P 值
PANC-1 组	19	11	5.023	<0.05
PANC-1/R2 组	18	15		

当 $n<40$ 或理论频数（T）<1 时，需使用四格表资料的 Fisher 确切概率法。本例中 $n=37$，不能直接使用四格表卡方检验，应使用 Fisher 确切概率法，直接得出 $P=0.151$，说明 2 组 Aurora-A 蛋白表达阳性率比较差异无统计学意义，与原先作者的统计结果完全相反。

2.2 误用 χ^2 检验代替秩和检验 误用 χ^2 检验代替秩和检验的情况时常发生在有序分类变量资料的比较中，对于单向有序分类变量，如果选择 χ^2 检验只能比较各组数据的差别，而无法比较各组强度之间的差别。有序分类变量可选用两个独立样本比较的 Wilcoxon 秩和检验或多个样本比较的 Kruskal-Wallis H 检验等。

例 5，某研究中作者对 2 组糖尿病眼病患者术前 1 d护理满意度进行比较，见表 6（该表为作者投稿时原表）。

表 6 2 组患者术前 1 d 护理满意度比较[例(%)]

组别	例数	护理满意度		
		满意	一般	不满意
干预组	51	37(72.5)	14(27.5)	0(0.0)
对照组	51	15(29.4)	31(60.8)	5(9.8)

注：2 组患者护理满意度总体比较，$P<0.01$。

该例作者将护理满意度分为满意、一般和不满意 3 个等级，是结果变量为有序变量的等级资料。原稿中作者采用 χ^2 检验对 2 组患者护理总体满意度进行比较，笔者认为宜采用两个独立样本比较的 Wilcoxon 秩和检验。经统计，得出结果为 $Z=4.51$，$P<0.01$。虽然两种方法都是差异有统计学意义的，但 χ^2 检验分析得出的结果是 2 组不同满意度等级之间的频数差异是否有统计学意义，而秩和检验得出的是 2 组不同满意度等级之间的差异是否有统计学意义。

例 6，表 7 为某研究作者对农村不同月龄婴儿喂养情况进行比较的原表。

表 7 农村不同月龄婴儿喂养情况

月龄	纯母乳喂养		部分母乳喂养		配方奶喂养		合计
	人数	构成比(%)	人数	构成比(%)	人数	构成比(%)	
1~2	38	32.76	63	54.31	15	12.93	116
3~4	43	45.26	35	36.84	17	17.90	95
5~6	27	36.00	35	46.67	13	17.33	75
合计	108	37.76	133	46.50	45	15.74	286
χ^2 值	3.609		6.407		1.167		
P 值	0.165		0.041		0.558		

本例中作者将婴儿喂养分为纯母乳喂养、部分母乳喂养和配方奶喂养 3 种喂养状况。就母乳喂养来说，3 种喂养状况呈有序等级梯次，属于多组单向有序定性资料（即单向有序 R×C 表），故表中分别就不同喂养状况用 χ^2 检验作不同月龄组间比较欠妥，笔者认为应采用 Kruskal-Wallis H 检验。结果得出农村不同月龄组婴儿不同喂养方式之间差异无统计学意义（$H=2.532$，$P>0.05$）。

2.3 误用 χ^2 检验代替配对 χ^2 检验 例 7，某研究采用甲乙两种方法测定 70 例恶性肿瘤患者体内 CK20 基因表达阳性率，比较两种方法测定的阳性率是否有差异，见表 8（该表为作者投稿时原表）。

表 8 两种方法测定结果比较

甲法	乙法		合计	χ^2 值	P 值
	+	−			
+	20	29	49	0.045	0.831
−	8	13	21		
合计	28	42	70		

本例原稿中作者采用一般四格表 χ^2 检验对两种方法测定的阳性率进行比较，结果得出差异无统计学意义。而正确的统计方法是采用配对四格表 χ^2 检验，另外要注意的是本例中 $b+c<40$，应采用校正配对 χ^2 检验，结果得出 $\chi^2=10.81$，$P<0.05$，差异有统计学意义，说明两种方法测定的阳性率有差别，显然甲法要由于乙法。

3 讨 论

科研设计是开展科技研究的第一步，关系着研究的成败。科研设计由专业设计和统计设计两部分组成，严谨的专业设计加上正确的统计学方法应用能够保证论文的科学性和可靠性[3]。一般来说，专业设计是保证科研结果的有用性与先进性的基础，而统计设计则是对科研结果的技术支持[4-5]。科研结果往往通过科研资料来解读，但是直接去解读科研资料一般是无法获取想要的结果或规律性的东西。因此，通常的做法是对科研资料进行合理的统计学处理，对统计结果进行解读，才能获得开展科技研究的真正目的[6]。

统计学处理在论文中具有重要的地位，能否正确应用统计学方法直接关系到论文研究结果的差异有无统计学意义[7]。由于统计方法的错误使用或使用不当造成的论文科学性不强、学术质量不高的现象，已引起广大编辑的重视[1]。从笔者日常编辑的稿件来看，高质量的稿件往往专业设计较缜密，且统计学方法使用正确；而学术质量不高，得不出可靠结论的稿件，其科研设计往往较粗糙，且统计学方法误用较多，甚至有些论文根本就不使用统计学方法。

目前医学期刊中统计学方面出现的错误相当普遍，一方面与作者统计学知识欠缺有关，另一方面也与编辑和审稿专家有关。由于大部分的医学期刊编辑不是统计学专业毕业的，所以在初审和稿件编辑过程中较难发现稿件中存在的统计学问题。另外，审稿专家

在审稿过程中重点审查的是专业领域的学术问题，通常对统计学相关问题只是粗略带过，最终造成这类稿件进入终审环节，使结果不可靠的文章得以发表，大大降低了期刊的学术水平和学术质量[8]。

提高医学期刊中统计学应用的质量是一项任重而道远的工作，它涉及到作者、审稿专家及编辑等各个环节，需要大家的共同努力[9-10]。首先，对于运用到统计学方法的稿件，可以请统计学审稿专家对稿件进行质量把关。其次，编辑人员也不能过分依赖统计学审稿专家，因为统计学专家一般都身兼数职，很难把全部的精力投入到稿件统计学问题的详细审阅中[11-12]。编辑人员应加强自身统计学专业素养，提高统计学专业知识水平，在编辑稿件过程中加强对稿件中统计学处理的思考和把握，遇到有疑问的地方要及时翻阅统计学资料或请教相关统计学专家以求正确解决。最后，对于某些稿件来说，由于在最初的实验设计阶段就存在一些问题，致使后期的统计学处理无法满足研究者的分析需求。这就需要从源头抓起，加强作者对统计学知识的掌握和了解，提高作者的统计学处理能力[13]。作者在进行科研设计时应明确科研目的，全面考虑研究的各方面因素，在分析数据时，对于不同性质资料、不同类型研究设计应选用合适的统计学方法。当研究中所涉及的数据分析较复杂时，建议与统计学专家交流合作，这样可以避免设计上的漏洞和偏差，而编辑部和一些医学机构则可以定期举办关于稿件统计学处理的专题培训班，还可以在期刊上开设相关栏目，对读者进行系统的统计学继续教育，提供参考资料[14]。

总之，不论是作者还是编辑都应该加强自身的统计学意识，从根本上解决稿件中存在的统计学问题，杜绝有问题的稿件发表在期刊上。作者和编辑都应该掌握一定的统计学知识，这样在进行科研设计及研究的过程中就会减少错误的产生，从而提高研究结果的准确性，保证了论文的严谨性、科学性、可靠性和有效性，同时也减少了对读者的错误信息的引导[15]。

参考文献

[1] 迟殿元. 科研设计与统计学在医学论文中的问题与对策[J]. 齐齐哈尔医学院学报，2012，33(11)：1548-1549.

[2] 郝丽洁，贾崇奇. 某医学期刊的统计学应用情况分析[J]. 中外医学研究，2016，14(13)：153-154.

[3] 彭芳，董燕萍，金建华，等. 医学期刊编辑应重视论文中统计学审查[C]. 中国科技期刊新挑战：第九届中国科技期刊发展论坛论文集，2013：391-394.

[4] 方积乾. 卫生统计学[M]. 7 版. 北京：人民卫生出版社，2012：266.

[5] 颜虹. 医学统计学[M]. 2 版. 北京：人民卫生出版社，2014：173.

[6] 沈宁，刘一松，郭春雪，等. 在生理学研究中要正确运用统计学[J]. 中国应用生理学杂志，2016，32(2)：185-190.

[7] 闫娟，李国琪. 医学稿件中统计学方法误用研究的回顾和评价[J]. 今传媒，2013，21(2)：58-59.

[8] 廖薇薇，舒安琴. 医学期刊编辑如何发现统计学错误：实例分析[J]. 新闻研究导刊，2017，8(2)：218-220.

[9] 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.

[10] 尚永刚. 编辑应重视医学期刊中数据的审查与检验[J]. 编辑学报，2016，28(1)：26-28.

[11] 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.

[12] 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.

[13] 张维，邓强庭，冷怀明. 医学期刊中容易误用的统计学方法辨析[J]. 编辑学报，2013，25(5)：435-437.

[14] 刘霞，周晴，周英智. 医学论文中 2 组等级疗效比较问题分析与建议[J]. 编辑学报，2017，29(2)：125-127.

[15] 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.

（本文编辑：季群） 收稿日期：2017-04-05

（上接第 1693 页）

[4] 楼小航，刘继峰，吴纪龙，等. 重症多形红斑与中毒性表皮坏死松解症患者的感染分析[J]. 中华医院感染学杂志，2016，26(10)：2281-2283.

[5] Leenutaphong V，Sivayathorn A，Suthipinittharm P，et al. Stevens-Johnson syndrome and toxic epidermal necrolysis in Thailand[J]. Int J Dermatol，1993，32(6)：428-431.

[6] 龚春燕，申国庆，蔡果，等. 重症多形红斑型药疹 25 例和中毒性表皮坏死松解症 17 例回顾性分析[J]. 药学与临床研究，2014，22(6)：539-541.

[7] 平晓芳，卢桂玲. 重症多形红斑型药疹 52 例和中毒性表皮坏死松解症 31 例回顾性分析[J]. 中国皮肤性病学杂志，2013，27(2)：148-150.

[8] 姜涛. 老年性药源性糖尿病低血糖昏迷的临床疗效分析[J]. 中国社区医师，2015，31(9)：19，21.

[9] 许敏，王佑民. 药源性糖尿病[J]. 药品评价，2013，10(23)：14-17，22.

[10] 李久旭，梁潇，张楠，等. 27 例药源性低血糖临床特点分析及监护[J]. 中国药业，2016，25(3)：60-62.

[11] 陈霞，谭兵. 1 例卡马西平致中毒性表皮坏死松解症病例报道及文献分析[J]. 中国药房，2014，25(4)：360-362.

[12] 李蕾，邹先彪. 重症多形红斑型药疹的诊治进展[J]. 世界临床药物，2013，34(6)：321-324.

[13] 郑敬旭，叶小舟，吴雪梅. 沙利度胺致中毒性表皮坏死松解型药疹 1 例[J]. 中国医院药学杂志，2014，34(14)：1235-1236.

[14] 祝伦，蒋法兴. 纳米银抗菌敷料外用治疗重症药疹大面积表皮松解[J]. 安徽医学，2013，34(11)：1714-1717.

[15] 刘荣荣，朱红，何春涤. 糖皮质激素联合丙种球蛋白治疗重症药疹临床分析[J]. 中华实用诊断与治疗杂志，2012，26(7)：698-699.

[16] 张杰，李美洲，朱红卫. 重症药疹的护理[J]. 河北医药，2016，38(19)：3036-3038.

[17] 荆鲁华，王燕飞. 免疫球蛋白联合糖皮质激素、环磷酰胺治疗重症药疹的临床疗效观察[J]. 中国实用医药，2011，6(3)：139-140.

[18] 石家宴，付敏，杨义成. 优质护理理念在重型药疹患者并发症预防中的应用[J]. 浙江临床医学，2013，15(8)：1262-1263，1264.

[19] 范宝荣，陈慧，连石. 97 例药疹临床资料分析[J]. 北京医学，2015，37(1)：70-71.

（本文编辑：季群） 收稿日期：2016-11-15