优化试验中常见的统计学错误

刘 宁[1],梁媛媛[2],程 新[3]*　(1. 山东胜利股份有限公司,山东济南 250101;2. 江西农业大学经济与贸易学院,江西南昌 330045;3. 江西农业大学生物科学与工程学院,江西南昌 330045)

摘要　以科技文献中常见的优化试验为例,介绍了一些常见的统计学错误,并提出了相应的解决对策。
关键词　科技论文;优化试验;统计学;错误
中图分类号　S131　　**文献标识码**　A　　**文章编号**　0517 - 6611(2009)22 - 10333 - 02

Common Errors of Statistical Analysis in Optimization Experimentation

LIU Ning et al　(Shandong Shengli CO.,LTD, Jinan, Shangdong 250101)

Abstract　With common optimization experiment in scifentific literature as example, the common errors in technological article were introduced, some countermeasures for these errors were provided.
Key words　Technological article; Optimization experimentation; Statistic; Error

优化试验是科学研究中常见的试验形式[1],目前国内所发表的科技论文中,涉及到优化试验的占很大比重。存在统计学错误一直是严重影响科技论文质量的重要因素之一。目前,大量科技文献中都存在统计学应用错误的情况。以医学科技论文为例,20 世纪 90 年代的统计结果表明,医学论文中统计方法的应用错误率高达 60% 以上。2002 年统计表明,随着统计学知识的普及,错误率有所下降,但仍在 30% 左右[2]。为此,部分杂志还曾经专门约稿邀请统计学专家对杂志的统计学错误进行综合点评[3]。笔者以优化试验为例,对科技论文中常见的统计学错误予以分析,并提出一些针对性的处理意见,以期为广大科技论文撰写者及提高论文撰写质量提供一定的帮助。

1　单因素试验中的常见统计学错误

单因素试验是一种较简单的试验设计,这种试验设计方法应用灵活,结果处理简便,不需要研究人员对试验设计有深入的理论知识,因此是很多优化试验所采用的基础方法。尽管该方法较简单,但在不了解统计学原理的情况下,仍会出现大量统计学错误。常见的错误有以下几种:

1.1　未遵循随机化原则　随机化是试验设计的基本原则之一,是在试验中排除非试验因素干扰的重要手段。随机化原则包括既要随机抽样,又要随机分组。常用的随机化方法有随机数字表法、抽签法等[1]。而目前很多科技论文或不标明是否采用随机化方法,或者是不管实际是否采用了随机抽样或分组方法,也未表明采用何种随机方法[4]。

1.2　未标明试验类型　常用的试验设计方法有完全随机法和随机单位组(随机区组)法。在试验对象个体较均一的情况下,可采用完全随机法,但当试验对象个体间存在一定差异的情况下,根据局部控制原则,可将整个试验环境或试验单位分成若干个小环境或小组,在小环境或小组内使非处理因素尽量一致,这种试验方法可较好地降低试验误差[1]。完全随机设计和随机单位组设计不仅在试验设计方法上有区别,而且在统计方法选择上也有区别。因此,根据试验对象特性,合理选择设计方法较有必要。而目前在很多科技论文中,作者一般不提供试验分组方法,即使表明"随机单位组"方法,也不提供具体分组原则和方法。

1.3　试验结果缺少统计推断　试验结果一般都针对样本结果,若要对整体试验做出相应结论,则必须通过统计推断方法进行。常用的单因素试验统计方法有 t-检验法、方差分析法、χ^2 检验法等。而目前很多科技论文缺少该环节,而采用样本统计量直接替代总体参数,最终得出结论。这种处理方法无法区分试验效应与误差效应,因此得出的结论缺乏说服力。

1.4　对显著性检验 P 值存在错误解释　尽管统计学显著性在科学研究中应用广泛,但其含义常被误解,将 P 值错误地认为是判断各处理间差异大小的证据。P 值在假设检验中只是一个概率,它所反应的是支持无效假设可能性的大小。P 值很小,则推翻无效假设。P 值越小说明越有理由推翻或拒绝无效假设,接受备择假设,并不能得出 P 值越小两者相差越大的结论[1,5]。同时,存在统计学意义上的显著性,并不代表在实践中同样具有应用价值[1,5]。

1.5　试验结果的表达方式有错误　试验结果的描述性分析一般分为集中程度和离散程度的描述。目前常采用"算术平均数 + 标准差"或"算术平均数 + 置信半径"等表示方法。而科技论文通常只给出算术平均数,而未在文中列出表示离散程度的统计量,这样使读者无法从中判断试验误差。同时,算术平均数并不适合于所有情况,当数据分布为非正态分布时,采用算术平均数并不能很好地描述数据的集中程度,因此,可采用几何平均数、中位数等统计量。

1.6　未设置重复观测值　很多试验由于科研人员的疏忽或试验条件的限制,并未设置重复观测值,这给试验结果的处理和统计带来诸多问题。设置重复的主要作用是估计与降低试验误差。若同一处理仅实施于一个试验单位,则只能得到一个观测值,但无从分辨出差异,因而无法估计试验误差。只有当同一处理实施在 2 个或 2 个以上的试验单位中,获得 2 个或 2 个以上观测值时,才能估算出试验误差。

2　正交试验中常见的统计学错误

正交试验是一种常见的试验设计方法,它具有设计简单、应用范围广、统计方便等特性,因此在优化领域得到了广泛的应用[1]。正交试验是一种不完全试验设计法,除了前文

作者简介　刘宁(1977 -),女,山东鄄城人,硕士,工程师,从事生物制药领域研究工作。* 通讯作者。
收稿日期　2009-04-13

所提到的一些错误外,在进行正交试验设计和统计时仍存在一些注意事项。在正交试验中也存在一些常见的统计学错误,主要表现在以下几方面:

2.1 以直观分析替代统计推断 直观分析是指利用试验所得到的数据,在不进行统计推断的前提下,进行一些简单、直观的分析方法。这些数据均为利用样本所得到的信息,可获得观测对象的一些特征,是其他分析的基础。但仅依靠直观分析,并不能区分出处理效应和误差效应,因此也无法得出确切的结论。因此,在对正交试验结果进行直观分析的基础上,应采用方法分析、多重比较、回归分析等方法,对结果进行进一步分析,但目前很多科技论文中,仅给出直观分析结果,并得出结论,如有些医学论文仅标明"经统计学处理",便得出结论,而不注明具体的统计学方法。统计方法交待不清或不予交待,使读者无法判断论文结论的正确性。因此,在论文中应说明具体的统计方法,如有特殊情况还应说明是否采用校正,这样才能使论文具有说服力。

2.2 缺少重复观测数据 根据正交试验的特点,在保留至少一个空列的情况下,正交试验的数据可以在无重复观测值的情况下进行方法分析[1]。因此,很多科研工作者利用该"特点"不设置重复观测值,设置重复观测值是估计试验误差必不可少的前提。而在未设置重复观测值的正交试验中,其误差是由"空列"来估计的,而根据空列估计的误差并不是真正的试验误差,因此,这种方法所得到的结果缺少一定说服力。

2.3 缺少试验设计方法 根据试验设计方法的不同,正交试验的结果分析存在一定差异。如完全随机法和随机单位组(区组)设计法的试验结果不同。而目前所发表的大部分涉及到正交试验的科技论文中,均未提到其试验设计采用的方法。一些论文虽说明了采用随机单位组设计法,但未注明具体的手段,如抽签法、随即数字表法等,这种表述方式缺少说服力。

2.4 对正交试验结果的错误解释 目前,大量正交试验在判断对结果影响最显著的因子时,均通过直观分析的极差大小[6-7]或通过 P 值进行决断。P 值大小不能用来判断因素对试验结果的影响程度;而极差分析同样不能作为判断因子影响程度大小的依据。首先,极差分析属于直观分析,无统计推断作基础,无法区分处理效应和误差效应;其次,极差分析结果只能说明当某因素改变一定范围时,对试验结果造成的影响,这种结果不能进行相互比较。如,A 因素从 10 g/L 变化到 30 g/L,极差为 2.0,B 因素从 3 g/L 变化到 5 g/L,极差为 1.5,但并不代表 A 因素的影响强度比 B 因素大,由于两者的变化不在同一水平,尤其是当比较不同单位的因素之间影响强度时,更无法得出结论[8]。

3 回归分析中的常见错误

回归分析是一种常见的分析手段。这种分析方法将相关变量区分为自变量和因变量,然后通过建立回归方程,总结出自变量和因变量的关系,从而达到分析的目的。在优化试验中,均匀设计、响应面分析等试验结果都需要通过回归分析完成。回归分析,尤其是多元回归分析,是一种较为复杂的分析手段,其中需要注意的问题较多。

3.1 以 r 值或 R^2 值替代统计推断 相关系数 r 或决定系数 R^2 常被用来表示回归方程的显著性,这是一种错误的方法[1]。回归方程是否显著,需要通过 F 检验或 t 检验来进行。一个 r 或 R^2 较大的方程,虽然常常是显著的,但两者并不存在必然的因果关系。因此,对于建立的方程,作者应给出显著性检验的结果。

3.2 自变量相关 在进行多元回归分析时,若自变量之间高度相关,某些回归参数的估计值则极不稳定,甚至出现有悖常理、难以解释的情况[9]。因此,在进行回归分析前,应通过相关检验方法,判断自变量之间的相关性。若自变量高度相关,则可以通过主成分回归等方法将原有变量处理后,再加以回归。目前发表的科技论文中,很多作者均未考虑自变量的相关性问题,因此很多回归方程与理论知识、现实经验等有所抵触。

3.3 自变量的选择问题 在回归分析中,方程的自变量并不是越多越好,过多的自变量会带来诸多误差,且容易造成变量的自相关。因此,通过合适的手段,将一些不显著的自变量从方程中"剔除"较有必要。目前常用的方法有前进法、后退法、逐步回归法等[9],而在发表的科技论文中,很多作者不注意这个问题,无论自变量显著与否,都将其引入方程中,最终造成一系列问题。而且很多作者为了追求高的 R^2 值,刻意增加方程的自变量数目,这都是不正确的处理方法。

3.4 回归分析方法不完整 很多科技工作者在完成回归方程的建立后,便误认为已完成工作。实际上,回归方程尤其是多元回归方程是否合适,需要通过大量后续工作完成。常见的方法包括回归诊断[10]、残差分析等。一个 R^2 较大或已通过显著性检验的方程,并不代表是符合要求的方程。一方面,方程应该满足统计分析的理论要求;另一方面,还应与实际情况、理论知识相符合。这就要求科技工作者在进行回归分析时,能较好地掌握统计学知识,同时多参阅他人的文献,综合分析后再得出结论。

4 优化试验中其他常见统计学错误

4.1 不注明统计学软件的名称、版本号及计算方法 目前,统计学软件已逐步开始替代计算器,成为科研工作的良好助手。但很多科研人员在发表文章时,并未说明所使用的统计学软件名称和版本号[3]。这种表述的方法与国际惯例相违背,也不便于读者核对结果。此外,对于很多软件来说,同一个名称下的算法,可能有很多种,如在 R 程序中[11],用来做非线性拟合的算法就包括 Gauss-Newton 迭代、Golub-Pereyra 迭代等。作者在撰写科技论文时,应给出具体算法,以便于读者核对。

4.2 直接用 P 值范围描述统计结果 有些医学论文中未注明统计方法和统计量。严格地说,撰写科技论文时,应注明精确的统计量值(如 t 值、F 值等)[5]。同时,目前提倡用精确的 P 值描述试验结果,而不应笼统地以 $P>0.05$ 或 $P<0.05$ 代替。

5 结论

统计学的思维方法和统计技术是科学研究的重要工具。统计学方法运用的质量是影响科研工作质量和科研论文水

(下转第 10360 页)

择参考其他试验方法,分别采用:甲醇-0.015 mol/L 醋酸钠(5∶95)[4],流动相:甲醇-0.1 磷酸(5∶95)[5]作为流动相,考虑到 V_C 本身化学结构的影响,最终选择甲醇-0.5%磷酸(8∶75)作为流动相,分离效果比较好。

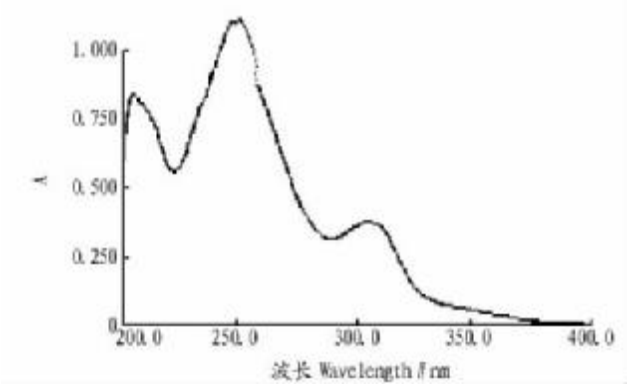

图 2 紫外扫描图谱

Fig. 2 UV scanning map

2.2 线性范围 按照上述色谱条件测定,以峰面积为纵坐标,进样量为横坐标作标准曲线,得回归方程为:$Y = 0.21 + 2\,467.32X$,$r = 0.998$。结果表明,进样量在 0.324 ~ 5.120 μg 范围峰面积与进样量呈良好线性关系。

2.3 精密度试验 对同一浓度对照品溶液,分别进样 5 次,每次 20 μl,测得 V_C 含量,计算相对标准偏差 *RSD* 为 0.2%,表明仪器精密度良好。

2.4 稳定性试验 取同一样品溶液分别于 0、5、10、15、20 h 时进样 20 μl,峰面积基本不变,*RSD* < 2.1%,说明样品在 20 h 内稳定性较好,可以作为试验方法的依据。

2.5 重复性试验 取同一批号的供试品,按供试品溶液制备项下操作,平行制备 5 份,按色谱条件测定百分含量,Vc 的平均含量为 10.81 μg/g,*RSD* 为 1.23%,说明该样品成分测定的重复性较好。

2.6 回收率试验 准确称取已知含量的猕猴桃 5 份,各 0.5 g,分别准确加入对照品 1.0 mg,按样品溶液的制备项下操作,在上述液相条件下进行 HPLC 分析。每个样品测定 3 次,根据峰面积计算平均回收率,其结果为 99.87%,*RSD* 为 1.8%。结果表明,回收率符合测定要求。

2.7 样品的含量测定 取 3 批样品,按上述供试品溶液的制备方法制备供试品溶液,依上述色谱条件测定含量,分别进样 20 μl,记录色谱图。由表 1 可知,猕猴桃 V_C 平均含量为 10.81 μg/g。

表 1 样品中 V_C 的含量

Table 1 V_C Contents of samples

组别 Groups	样品含量//μg/g Samples contents	*RSD*//%
1	10.83	0.10
2	10.80	0.13
3	10.81	0.11

3 结论与讨论

3.1 色谱条件的选择 试验表明,流动相中磷酸的含量对 V_C 的影响较大,V_C 在中性和弱酸性条件下稳定,经试验确定试验中磷酸体积比。对 V_C 对照品溶液进行了紫外全波长扫描结果显示,V_C 在 258 nm 处有最大吸收,故检测波长定为 258 nm。

3.2 提取溶液的选择 V_C 具有亲水性,分别采用不同浓度的甲醇提取,其含量测定结果差异显著,提取效果及稳定性均好,其中 75% 甲醇提取率最高,因此,用 75% 甲醇为提取溶剂,并辅助超声提取。

3.3 方法的耐用性 固定流动相和检测波长,试验中通过改变柱温、流速、色谱柱部分色谱条件测定维生素 C 的含量,测定结果均相同,表明该方法耐用性好。该方法简便、准确,可作为猕猴桃的定量质控标准。

参考文献

[1] 马清温,万鹏,孙震晓,等. 山东药用植物[M]. 济南:山东科技出版社,1998:140.
[2] 余纲哲. 食品资源化学[M]. 汕头:汕头大学出版杜,1996:122-125.
[3] 凌育赵,刘经亮. 猕猴桃果酱中维生素 C 测定方法的比较研究[J]. 中国调味品,2009(2):101-102.
[4] 陈再浩,郑建明,王智. 高效液相色谱法测定维生素 C 片中 Vc 含量[J]. 分析仪器,2008(6):37-38.
[5] 李文玉,李虎将. HPLC 法测定维生素 C 制剂[J]. 中国药事,2008,22(9):810.

(上接第 10334 页)

平的重要因素。统计分析方法的正确运用已日益引起广大科学研究者的重视。目前仍有一些科研人员在试验设计和数据收集与整理、资料的分析与表达、结果的解释与陈述方面存在不少漏洞,从而了影响论文的质量。从目前国内发表的研究论文来看,这种情况并不少见。笔者以优化试验为例,对一些常见的统计学错误进行了归纳总结和整理,希望能够对广大科研工作者提供一些帮助。

参考文献

[1] 明道绪. 生物统计附实验设计[M]. 北京:中国农业出版社,2003.
[2] 李永红. 医学论文中常见的统计学错误与处理[J]. 海峡预防医学杂志,2007,13(1):94-95.
[3] 胡良平,高辉.《中西医结合学报》2006 年第 1 期论文中统计学应用错误辨析[J]. 中西医结合学报,2008,6(1):98-106.
[4] 吴青. 医学论文中常见的统计学错误分析[J]. 山东医学高等专科学校学报,2008,30(4):298-300.
[5] 潘发明,夏果,廖芳芳,等. 临床科研论文中常见的统计学错误分析(二)[J]. 安徽医药,2008,12(6):576-577.
[6] 杨盛,侯红萍. 高效降解纤维素混合菌的筛选及其产酶条件的研究[J]. 中国酿造,2008(21):20-23.
[7] 韩晶,李宝坤,李开雄. 嗜热 β-葡聚糖酶产生菌的筛选及其培养基优化研究[J]. 中国酿造,2008(21):33-36.
[8] 刘月华,施云芬,周兆梅. 凝集型酵母聚凝体解离的研究[J]. 中国酿造,2008(21):41-43.
[9] 方积乾. 医学统计学与电脑实验[M]. 上海:上海科学技术出版社,2001.
[10] 刘沛. 回归诊断是多元回归资料再开发的有力工具[J]. 中国卫生统计,1993,10(3):39-40.
[11] 程新,魏赛金,江莉,等. 统计软件 R 及其在《生物统计学》实验教学中的应用[J]. 统计教育,2008(4):29-31.