统计分析专栏

单一响应变量统计分析在烟草学研究中应用的若干问题

杨锦忠，宋希云

青岛农业大学 / 山东省旱作农业技术重点实验室，青岛 266109

摘　要：正确选择和应用统计分析方法是研究工作成败的重要因素之一。本文在简述统计分析一般流程、解答生物统计学应用常见问题的基础上，重点介绍了如何根据响应变量和解释变量的性质及其组合，正确选择和应用统计方法的原则与要点，强调了统计诊断的重要性。此外，还介绍了随机模拟抽样、重复测量、稳健回归、分位数回归、广义线性模型、Bootstrap 技术、元分析、测量误差模型等最新统计方法的应用场合。

关键词：响应变量；解释变量；统计方法选择；统计分析流程；统计诊断；统计应用指南

doi：10.3969/j.issn.1004-5708.2014.04.020

中图分类号：O29　　**文献标志码**：A　　**文章编号**：1004-5708（2014）04-0108-07

Guides to statistical techniques for single response variables in tobacco science research

YANG Jinzhong, SONG Xiyun

Shandong Provincial Key Laboratory of Dry Farming Techniques, Qingdao Agricultural University, Qingdao 266109, China

Abstract: Responsible selection of statistical techniques is one of the key factors in tobacco science research. Principles and requirements for selection and use of statistical techniques were introduced according to combinations and features of both response and explanatory variables, with focus on significance of statistical diagnosis option. Such new statistical methods as random simulation of missing values, repeated measures, robust regression, quantile regression, generalized linear models, bootstrap techniques, meta-analysis and measurement error models were also discussed as to when and where they can be applied by tobacco research professionals.

Keywords: response variable; explanatory variable; statistical method selection; statistical analysis procedure; statistical diagnosis; statistical guide

计算技术及计算手段今非昔比，极大地推动了统计分析及其应用进步，新的方法和应用领域不断出现。包括许多开源和免费软件在内的统计分析程序日益普及，科技人员足不出户就可以完成任意复杂的分析计算。可是，科技论文的统计错误也着实令人担扰，即使世界顶级刊物也不例外。据报道[1]，在《科学》和《自然》等世界前五名杂志上发表的 513 篇神经科学论文中，有 79 篇使用了不正确的统计方法，约占受检论文总数的 15%。我国科学研究中应用统计分析的深度和广度正在不断增加，学术论文中也同样存在诸如分析方法选择不当、分析选项失误、分析结果解释失真、结果应用有误等统计错误。为进一步提升我国烟草学术论文的统计分析水平，本文拟就单一响应变量的统计分析应用的常见问题，特别是如何选择正确的分析方法，提出针对性的建议。

基金项目：泰山学者岗位（20090510）；山东省旱地作物水分高效利用创新团队（20121025）

作者简介：杨锦忠（1963—），教授，从事数字农业研究，Email：jzyang@qau.edu.cn

通讯作者：宋希云（1961—），教授，从事作物遗传育种研究，Tel: 0532-86080009, Email: songxy@qau.edu.cn

收稿日期：2013-10-12

1　统计分析方法选择概述

1.1　关联试验研究与统计分析的两种变量

从试验研究角度看，任何试验数据都是由变量及其取值组成的。变量有两大类：响应变量（response variable）和解释变量（explanatory variable）相对应。以烟草学为例，响应变量表示研究者最关注的烟草性状，如产量或者品质等，打算通过研究剖析其变化规律或者进行预测。解释变量则表示能够影响响应变量的其它因素，如烟草的基因型、生产措施、土壤因素、

环境因素等。由此可知，在因果关系研究中，结果是响应变量，原因是解释变量；在遗传关系研究中，一般来说亲代是解释变量，子代是响应变量；在方差分析中，被分析的变量，如产量是响应变量，各种具体的变异来源，如处理、试验因素等是解释变量；在回归分析中，依变量是响应变量，自变量是解释变量。值得指出的是，一个变量属于响应变量还是解释变量，并不是固定不变的，而是因分析目标而异。例如，品种比较试验中，烟草基因型是解释变量（产量是响应变量），但是在品种识别研究中，烟草基因型却是响应变量（烟草各种性状是解释变量）。

1.2　统计分析的一般流程

生物统计学教科书一般都侧重于介绍统计原理与方法，很少涉及科研活动的分析实践。如何避免失误，最大限度地挖掘试验数据中包含的信息，则是分析实践必须考虑的问题。这一问题的答案就是借鉴学术界的成功经验，认真执行统计分析的一般流程，养成良好的分析习惯。完整的分析流程包括核实数据、选择合理的分析方法并付诸行动、进行统计诊断、分析结果解释与报告等环节。

首先，要仔细审核录入计算机的数据正确性。

其次，审查数据中是否存在异常值、离群点。结合烟草学专业知识运用统计方法对数据进行整理，同时对是否存在异常值、离群点进行判断和甄别。常用的统计方法主要包括描述统计量和统计图，描述统计量主要有最大 / 小值、标准差和变异系数等；统计图主要有：次数分布直方图、任意两个连续型变量的相关散点图、连续性变量对离散型变量的盒形图和柱形图。

如发现异常值或者离群点后，结合专业知识可以直接删除，即当作缺失值。一般不建议用平均值替换缺失值，因为违背了随机性要求。许多统计软件碰到有缺失值的变量时，会将整个记录删除然后进行后续分析，这样必然发生信息损失，并且有可能形成不平衡数据。目前解决缺失值问题的较好办法是，利用计算机进行随机模拟抽样，然后再逐一分析模拟样本，最后合并全部模拟样本的分析结果，结合专业知识得出结论，SAS 软件有这一功能模块。如果是回归分析问题，则使用稳健回归 (Robust regression) 方法（如广义极大似然、岭回归、主成分回归等）取代最小二乘法，可以有效减轻离群点的干扰，获得较好结果。另外，通过单位变换或者线性变换，使参与计算的所有变量保持在相同的数量级，可以有效降低计算误差。

第三，选择恰当的方法分析数据（详见下文）。选定某一种方法后，确定适当的分析选项，包括图形输出和统计诊断选项，以对各种假定进行检验，要充分利用这一功能，起码作者认为非常重要的科学发现和技术发明一定要过统计诊断关。对于复杂的统计方法，最好请教经验丰富的领域专家和统计专家。

第四，查看计算机软件运行日志和分析结果。切记不要忽略运行日志。若日志报告了错误，提出了警告，则要仔细检查数据和分析程序，找出问题所在并加以解决，然后重新分析，有时甚至需要重新选择分析方法。

查看分析结果时，先结合图形输出看统计诊断结果，然后再看参数估计和显著性检验结果。只有诊断显示基本正常时，才能使用参数估计和显著性检验结果。否则，一切分析结果都是无效的，甚至是错误的。例如，如果响应变量对解释变量的散点图呈现明显的大喇叭形状，说明误差方差与自变量相关，此时不能使用普通最小二乘回归，而应该使用加权最小二乘法，或者使用分位数回归 (Quantile regression)[2]。普通最小二乘回归只能描述自变量对于依变量均值变化的影响，而分位数回归能更精确地描述自变量对于依变量的各种分位数以及条件分布形状的影响。现在已有分位数回归软件，如免费的 Quantreg[3]。

最后，对分析结果进行专业解释，并阐述产生结果的原因。专业解释力求通俗易懂。解释时要注意区分因果关系和非因果关系。分析结论是应用统计分析工具的产物，仅提供了处理效应的估计及其显著性，在某种意义上仅是一种“概率证明”。若要论证结论的可靠性，最好能够找到产生结果的原因，作为证据。概率论证和专业论证相互补充，结论才更具有说服力。

1.3　选择统计方法的一般考虑

统计分析方法是为了满足人们解决实践问题的需要而逐步发展起来的，所以，在选择统计分析方法时，首先应当回答以下几个问题：试验目的和分析目标是什么？响应变量是什么？有哪些解释变量？变量的类型和数目？试验设计的类型？试验数据是否满足统计方法的基本假定？试验数据的性质在很大程度上决定了采用什么样分析方法。在计算机统计软件普及之前，分析步骤繁琐、计算复杂、工作量大是方法选择的限制因素之一，现在计算手段问题已经不复存在，最大限制因素是科技人员的统计学知识和统计咨询意愿。

所谓单一响应变量的统计分析乃指研究者在分析时只关注一个响应变量的变化，欲明确一个或者几个解释变量是如何引起响应变量变化的。在明确分析单个响应变量之后，根据两种变量的组合不同，可以采用的统计分析方法随之改变。表 1 给出了它们之间的对应关系。

表 1 解释变量和响应变量不同组合对应的统计方法选择

Tab. 1 Selection of statistical methods based on response and explanatory variables

解释变量	响应变量	
	离散型	连续型
离散型	卡方检验等非参数方法	*t*- 检验，*F*- 检验，方差分析
连续型	Logistic 回归，判别分析，典型变量分析	回归分析，相关分析
离散型和连续型	Logistic 回归，广义线性模型	协方差分析，广义线性模型

例如，株高、产量、施肥量等取值可以连续变化的变量即所谓连续型变量，而取值数目有限的变量是离散型变量，又进一步划分为二值变量（如发病、不发病）、等级变量（如灾害等级、烟叶等级）、名义变量（如烟草品种名称）。

根据统计理论，烟草研究的分析目标有：（1）明确响应变量的各种性质，诸如描述统计量，即平均数、变异数、峰度和偏倚度、以及概率分布类型等；（2）明确统计量的差异，如比较烟叶丰产性（平均数）或者稳产性（变异数）；（3）解析响应变量发生变化的原因并进行定量估算和甄别；（4）建立响应变量与解释变量之间的数量关系式并进行预测或者控制；如此等等。若响应变量是新定义的烟草性状，则描述统计就成为认识新性状的首选分析方法。

值得指出，统计方法的选择应该早在试验研究设计阶段就进行，否则，等试验结束之后再考虑，就可能出现测定指标不全、数据量偏少、数据取值范围偏颇、条件控制失当等问题，严重时导致统计分析失败，无法实现研究的预期目标。

2 统计分析应用的若干热点问题

2.1 参数的点估计和区间估计

点估计是统计描述的重要内容，区间估计则是统计推断的重要内容。尽管区间估计的计算复杂性和工作量都大得多，但是，当今计算机统计软件非常普及，相对于漫长的试验数据收集过程而言，统计分析的计算时间极其短暂，因此，推荐尽可能使用区间估计。

经典统计学的区间估计是基于先验概率分布，如正态分布建立的，由于研究对象的概率分布常常是未知的，其应用受到很大限制。现代统计学发明了许多区间估计的新方法，例如基于重抽样技术的 Bootstrap[4-5]，Jackknife 等。这些方法原则上能够解决包括区间估计在内的各种统计问题，如参数估计、显著性检验等。

2.2 t- 检验、F- 检验与非参数检验

同样是用于两个总体平均数的比较，t- 检验的基本假定多（试验指标服从正态分布，试验误差是独立的、随机的），非参数检验少（常常只要求试验指标的分布是对称的）。所以，非参数检验的普适性大得多。不过，非参数检验是相当保守的方法，当试验数据符合 t- 检验的基本假定时，使用非参数检验将增加犯第二类错误的风险。

同样是用于多个总体平均数的比较，F- 检验的基本假定多（试验指标服从正态分布，试验误差是独立的、随机的，全部处理的试验误差是同质的），非参数检验少（常常只要求试验指标的分布是对称的）。所以，非参数检验的普适性大得多。不过，非参数检验是相当保守的假设检验方法，当试验数据符合 F- 检验的基本假定时，使用非参数检验将增加犯第二类错误的风险。

2.3 假设检验

假设检验，又称显著性检验，是经典统计推断的重要内容。假设检验的结果分“显著”和“不显著”，极易引起读者误解，甚至作者误释。即使结论是“显著”，也只是对总体特征的定性推断，不象区间估计那样作定量推断。鉴于上述原因，假设检验已经在越来越多的场合受到质疑。

另外，烟草学领域的数据采集能力已经大大超越过去，样本容量和同源样品的试验测试项目（即试验指标）都大大增加，传统的假设检验方法面临着新考验。对来自同源样品的多个试验指标分别进行显著性检验，由于这些指标之间相关性会导致实际的显著水平会低于其名义值，必须对此进行矫正，例如在农学方面的应用 [6]。

2.4 “显著”和“不显著”

显著性检验（即假设检验）的结论，不外乎两种：“显著”或者“不显著”。它们究竟是什么意思呢？我国在上个世纪 30 年代从国外引进了生物统计

学，“显著”一词对应英文的“Significant”一词。“Significant”是一个多义词，意思是“有意义的；重要的；有效的；非偶然的”。根据显著性检验的统计学原理，“Significant”最贴切的译文是“非偶然的”，“Not significant”自然就成了“偶然的”。由于历史的原因，“显著”和“不显著”的译法一直沿袭下来了。

2.5　处理间差异显著时的处理效应

只要将试验误差控制得非常小，或者重复次数足够多，处理间差异即使非常微小，也可能获得“差异显著”的结论。但是，从专业实践角度看，这种微小的处理效应却不见得有应用价值。例如，处理间的烟草产量仅有 0.25% 差异，统计分析表明差异显著，但是，从应用角度看这种差异是微不足道的。只有处理间差异大而且显著的处理效应才能够认为是重要的。

烟草学领域的数据采集能力已经大大超越过去，大样本数据越来越多。在这种背景下，更应该明确处理效应的显著性确切含义，严格区分处理效应的统计学显著性和烟草学重要性。

2.6　解释与报告分析结果

分析结果的适用范围既取决于试验设计，又取决于分析时对效应类型的假定。根据固定效应模型得到的结果，稳妥的解释类似于这样：在与试验相同或者相似条件下，供试处理具有某种效果。由一个或少数地点试验结果推演到整个地区，由一个或少数烟草品种试验结果推演到烟草作物，都是根据专业知识进行的类比，不属于统计分析范畴，无论作者还是读者都必须充分认识到这一点。根据随机效应模型得出的结论适用范围大于固定效应模型。要防止按固定效应模型分析，却按随机模型下结论的做法。即使按随机效应模型分析，试验结论的可靠性还主要取决于参试处理的代表性。

服从正态分布的变量一般报告算术平均数和标准差，其它分布的变量则报告中位数和百分位距，大样本数据还要报告峰度和偏度。面向纯专业人员的报告，还应列出参数的区间估计，即对于被研究总体的平均数、平均数差异、方差、变异系数、相关系数、回归系数、回归预测值都要给出区间估计。有关统计分析结果报告的更多建议参见文献 [7]。

2.7　综合分析不同研究者的试验结果

当回顾前人研究进展时，不应该满足于只使用文字概括和归纳它们的异同，而应当进行元分析（Meta-analysis）。元分析是文献综述的一种量化方法，对同一问题的多项研究结果作系统性评价和总结，借助各种统计分析技术获得一般规律性认识，已在生命科学中得到广泛应用 [8-9]。简单做法是对处理效应按单变量描述统计进行汇总，复杂做法是进行处理效应的差异显著性检验 [10]。

2.8　统计分析的结论

可以把统计分析当作一个黑箱系统，输入是试验数据，输出是试验结论，统计方法是系统过程。结论是否有效、可靠，取决于输入和系统过程的质量。因此，评价统计分析结论时，要考虑以下几个方面问题。首先，试验数据是否有效。由于试验处理违背随机原则、试验实施发生差错、数据采集发生差错、数据抄录差错、甚至伪造数据等，造成试验数据本身存在错误，就会导致无效，甚至错误的试验结论。只有正确可靠的试验数据才有可能得出有效的结论。其次，统计方法的使用是否得当。任何一种统计方法都是一个数学定理，定理成立的前提是它的所有条件都得到满足。如果试验数据不满足统计分析的基本假定，生搬硬套统计方法，统计分析的显著性和置信度就不再拥有字面上的意义，此时，分析结论变得无效了。只有当试验数据满足统计分析的基本假定时，分析结果才是有效的。第三，是否对分析结果进行合理解释。分析结果常常以数字、公式或者图表的形式出现，由于对统计方法的原理不甚了解，一知半解，作者常常会错误地解释分析结果，或者解释不准确，引起读者误解。只有对分析结果结合专业知识进行正确解释，才是有效的试验结论。第四，研究性质。根据试验数据来源可以把研究分为两类：观察性研究和实验性研究。前者是非随机化的研究，在自然状态下对研究对象的特征进行观查、记录，后者是在人为控制条件下，遵循试验设计之随机、重复和局部控制三原则实施试验。一般而言，与观察性研究相比，实验性研究的统计分析结论更有说服力。观察性研究的最大风险来自漏掉对响应变量有重要作用的因素，以及误把伴随因素当作原因。最后，能否从生物学角度阐明分析结果的合理性。分析结果是试验数据的高度概括与抽象，这种归纳若能够从生物学角度（包括遗传学、生理学、生物化学、生物物理等分支学科）论述其合理性，则试验结论就拥有了统计学和生物学的双重证据，说服力大大增加。这方面例子可以参阅文献 [11]。

统计分析的结论是否符合客观，是否有用，除上述因素外，还取决于试验数据的信息量是否充足。信息量不足常常会得出违背客观的结论。统计方法在一定程度上能够排除偶然现象的干扰，获得对事物本质的认识。这个“程度”的大小，取决于试验数据的信息量。信息量越大，试验结论就越能够揭示事物的本

质。当比较两个总体的平均数（不妨假定二者不相等）时，如果试验数据的信息量足够大（此处指试验误差小和重复次数多），则获得“显著”结论的可能性就高，否则，信息量不足，就可能获得“不显著”结论。又如，在区间估计中，区间长度决定了估计的精度，置信度则反映估计的把握大小。在一定的置信度下，如 90%，信息量不足（此处指试验误差大和重复次数少）造成区间太宽，失去实用价值。

3 统计方法的选择

一般统计咨询都习惯于基于统计理论提出统计方法的适用情形，本文尝试根据试验研究任务提出统计方法建议。

3.1 单个处理（样本）的统计分析

单样本的统计分析工具最为丰富 [5]，根据研究目标，可以选择的分析内容包括：总体分布类型（如正态、二项、负二项、泊松、指数、韦布尔等）检验，均值、方差、偏度、峰度、变异系数等参数检验与区间估计，样本容量估计，数据的可视化展示（如直方图、茎叶图、盒形图），如此等等。

3.2 两个处理比较试验的统计分析

两个处理无配对数据进行 t- 检验，应先做 F- 检验以判断方差同质性。若同质，则用等方差 t- 检验，否则用异方差的 t- 检验。t- 检验要求响应变量符合正态分布，不符合正态性假定的数据，要使用适合于离散型响应变量的非参数统计。

对于 2×2 和 2×R 列联表，建议使用 Fisher 氏精确检验，不使用卡方检验。卡方检验只是精确检验的近似，精度随期望次数而减小，期望次数小于 5 时效果很差。

两处理数据不仅要进行差异显著性检验，有条件时还要进行区间估计，后者更容易理解、更全面。针对两个处理的均值比较的情形，表 2 罗列了不同试验数据性质和响应变量性质对应的统计方法，供大家参考。

表 2 响应变量性质影响两个处理均值比较方法的选择

Tab. 2 Effect of response variables on the selection of comparison methods for 2 treatments of average

统计量	试验单元控制	连续型响应变量	离散型响应变量		
			二值变量	等级变量	名义变量
平均数	无配对	服从或者可变换为正态分布，用 t- 检验，否则同等级变量。	2×2 列联表的 Fisher 氏精确检验，卡方检验	2×R 列联表的 Fisher 氏精确检验，卡方检验	2×R 列联表的 Fisher 氏精确检验，卡方检验
平均数	配对	服从或者可变换为正态分布，单样本 t- 检验，否则同等级变量。	McNemar 检验	Wilcoxon 符号秩检验	——

上述方法适用于两个处理间均值比较。若比较它们的方差，则在响应变量服从或者可以变换为正态分布时，用 *F*- 检验。否则，使用 Levene 检验。

3.3 多个处理比较试验的统计分析

当处理水平是连续型变量的不同值时，应当进行回归分析和相关分析。若响应变量也是连续型变量，则先绘制散点图观察变化趋势，再根据专业理论或者散点图趋势选择直线回归还是曲线回归，以及曲线的类型。经典的线性回归假定自变量没有误差，这在某些场合下并不符合事实，例如，利用土壤速效氮含量预测作物产量的线性回归问题，土壤氮浓度就有误差。此时，应该使用回归的测量误差模型或者 EIV 模型 [12]。特别地，若响应变量是非正态的，则不宜使用常见的直线相关分析，而应当作秩相关分析等非参数检验。若响应变量是离散型变量，可以进行判别分析，或者进行 logistic 回归分析。对于非正态数据，还可以考虑采用广义线性模型（见下节详述）。

当处理水平可以看作是离散型变量的不同取值时，若响应变量是连续型变量，进行方差分析；若响应变量是离散型变量，则进行卡平方检验。

在选择处理均值的多重比较方法时，应当避免使用国内常用的 Duncan 检验，因为它只控制比较水平的第一类错误率，却不能控制整体水平的第一类错误率，容易出现假显著，国际上有的学术杂志甚至禁止论文使用该方法。对于正态分布的响应变量，单一自由度比较法适用于早在试验设计时就列入计划的正交

对比，Dunnett 法适用于检验对照与每个处理的差异，Hsu 法适用于检验比较每个处理与剩余处理中最好者的差异，Tukey 法适用于检验任意两个处理之间的差异。上述次序也是各种方法检验功率由大到小的顺序。非正态分布的连续型响应变量可以使用 Kruskal-Wallis 检验或者 Dunn 检验进行多重比较 [5]。

比较多个处理的方差是否相等，即所谓的方差齐性检验，在烟草中有重要应用，因为均匀性（方差越小均匀性越好）是烟草工业的重要质量指标，而且许多统计方法诸如方差分析、回归分析、主成分分析、判别分析等都要求满足方差齐性假定。若响应变量符合正态分布，则使用 Bartlett 检验方差齐性，否则，使用 Levene 检验。国内常用的 Bartlett 检验对非正态性非常敏感，效果不理想（实际的第一类错误概率大于名义显著性水平）。

3.4　多因素试验的统计分析

对于多因素多水平试验结果（全部试验因素、区组，以及附加的或者隐含的观测变量都是解释变量）的统计分析，国内常见教科书介绍的各种试验设计的方差分析，实际上只适用于响应变量符合正态分布的情形。若有连续型解释变量，则多元回归分析、协方差分析、通用线性模型分析更有效，能够提供更丰富信息。表 3 概括了此类试验数据针对响应变量和解释变量组合可供选择的主要统计方法。

表 3　多因素试验的响应变量和解释变量组合对统计方法选择的影响

Tab. 3　Effect of response and explanatory variables in multi-factor experiments on the selection of statistical methods

解释变量	响应变量		
	正态	连续非正态	离散型
离散型	方差分析	Friedman 中位数检验	对数 - 线性回归，Logistic 回归
连续型	多元线性或曲线回归	广义线性模型	Logistic 回归，广义线性模型
离散型和连续型	协方差分析，通用线性模型	广义线性模型	Logistic 回归，广义线性模型

计算机性能的快速提升为复杂试验数据分析技术的广泛应用提供了必要条件，反过来促进了统计分析技术的进步，近年来发展迅速的广义线性模型 (Generalized linear model) 就是一例。目前广义线性模型分析可以应用于符合下列指数族分布的响应变量：正态、指数、珈马、逆高斯、泊松、二项式、多项式。在此情形下，经典的多元回归分析、协方差分析、通用线性模型只是广义线性模型的特例 [13-14]。

烟草学研究经常需要在不同时间对同一小区或者植株进行重复测量。例如，烟叶采收期长达 6 ～ 12 周，需要多次采收。这就是所谓的重复测量现象（Repeated measures），国内农学类统计教科书称之为时间裂区。重复测量可以看作为一个隐含的试验因素，属于解释变量。对于含有重复测量的数据，建议不再按时间裂区进行分析，而使用专门的分析方法，因为后者提供了非常丰富的分析选项，如复合对称、一阶自回归、非均质自回归、非确定结构和收敛的 Toeplitz 等 [15]，从而能够获得更客观的结果。

3.5　多个试验的联合分析

相同的试验方案，在不同地点和年份或者季节实施，得到大量的数据。对这些数据进行联合分析，除增加误差自由度，提高分析精度外，还可以剖析地点或者年份效应，以及它们与试验因素的互作，加深对响应变量变化规律的认识，扩大试验结论的适用范围。在联合分析之前，先要解决方差齐性问题。

首先，进行单个试验的分析，获得每个试验的误差方差，然后，进行方差齐性检验（亦称同质性检验）。若符合齐性假定，则进行联合分析，否则，通过适当数据转换，直至符合要求后再进行联合分析，或者选择能够分析异质方差数据的方法。

若要进行方差分析，还必须先明确地点和年份的效应类型。生物学试验的年份一般都是随机效应，地点或为固定效应（当结论应用范围局限于参试地点时），或为随机效应（当结论推广至更多地点时，参试地点只是全部地点集合的一个子集）。

4　结束语

烟草学研究内容包罗万象，相应的数据分析需求也非常多，国内常见农学类生物统计学书籍中介绍的统计方法在满足这些需求方面已经发挥了很大作用，但是，其中绝大多数方法因正态性假定而受到很大应用限制，所幸现代统计学和计算机软件的极速发展为

我们突破这种限制提供了非常丰富的选择。本文从响应变量和解释变量角度入手，依据研究处理和试验任务的复杂程度不同，分别介绍了如何选择各种适宜的统计分析方法，特别是各种新技术的应用场合，强调了统计诊断对于正确分析试验数据的重要性。此外，还分析了多年来从事生物统计学教学和咨询中发现的共性问题，希望生物统计学这一研究工具在烟草学研究中能够发挥更大作用。本文仅仅展示了适用于烟草学研究统计方法的冰山一角，侧重于单个响应变量数据的分析，关于多个响应变量数据的分析问题将另文专门介绍。

参考文献

[1] Nieuwenhuis S, Forstmann B U, Wagenmakers E J. Erroneous analyses of interactions in neuroscience: a problem of significance[J]. Nature Neuroscience, 2011, 14:1105-1107.

[2] Koenker R, Bassett J G. Regression quantiles [J]. Econometric, 1978, 46: 33-50.

[3] Koenker R, Portnoy S, Tian P, et al. Quantreg: Quantile Regression [M]. R package version 5. New York: Cambridge University Press, 2013.

[4] Efron B. Bootstrap Methods: Another Look at the Jackknife [J]. The Annals of Statistics, 1979, 7 (1): 1-26.

[5] 茆诗松 . 统计手册 [M]. 北京：科学出版社，2006.

[6] 赵春明，韩仲志，杨锦忠 , 等 . 玉米果穗 DUS 性状测试的图像处理应用研究 [J]. 中国农业科学 , 2009, 42(11): 4100-4105.

[7] 杨锦忠，宋希云 . 烟草学术论文的统计学表达与展示 [J]. 中国烟草学报 , 2013,19 (4):114-118.

[8] 杨锦忠 , 陈明利 , 张洪生 . 中国 1950s 到 2000s 玉米产量 - 密度关系的 Meta 分析 [J]. 中国农业科学 , 2013, 46 (17): 3562-3570.

[9] 杨锦忠 , 张洪生 , 杜金哲 . 玉米产量 - 密度关系年代演化趋势的 Meta 分析 [J]. 作物学报 , 2013, 39 (3):515-519.

[10] Fleiss J L. The statistical basis of meta-analysis [J]. Statistical Methods in Medical Research, 1993, 2: 121-145.

[11] 杨锦忠 , 张洪生 , 赵延明 , 等 . 玉米穗粒重与果穗三维几何特征关系的定量研究 [J]. 中国农业科学 , 2010,43(21): 4367-4374.

[12] Fuller W A. Measurement Error Models [M]. New York: John Wiley & Sons, 1987.

[13] Hardin J, Hilbe J. Generalized Linear Models and Extensions [M]. College Station: Stata Press, 2007.

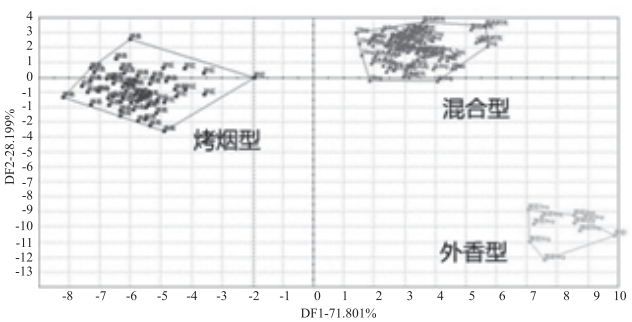
[14] 费宇 . 线性和广义线性混合模型及其统计诊断 [M]. 北京：科学出版社 , 2013.

[15] 卢纹岱 . SPSS 统计分析 [M]. 4 ed. 北京 : 电子工业出版社，2010.

更正启事

本刊 2014 年第 20 卷第 3 期（上期）第 5 页“田书霞，杨振民，胡林，等《卷烟主流烟气粒相物的电子鼻分析》”一文“图 6 中南海（特高）卷烟风格判别”做如下更正：

错误图：



更正为：