

· 讲座 ·

医学科研中如何用好应用统计学的方法

陈青山 孟晶 杨剑 周亚敏 刘晓玲

统计学是一门透过同质事物的变异性、揭示内在事物规律性和实质性的科学,确切地讲,是一门关于客观数据分析的科学,研究数据的收集、整理和分析,包括理论和应用两个方面^[1]。医学应用统计学,侧重于实际应用,是在传承和借鉴传统医学统计学“理论-原理-公式-应用”模式基础上,创造性地以“目的-数据库-变量类型-变量间关系”模式为指导的统计学^[2]。它遵循简单实用的原则,力避复杂的数学原理和公式推导,以实际问题为导向,以建立统计数据库、分清变量类型为基础,以分析变量与变量间关系为核心阐述统计学分析方法,对于广大医学科研工作者,具有内容简单、思维明确、操作可行、方法实用的特点。因此,学好用好医学应用统计学需要掌握如下一些基本方法。

一、明确研究目的和研究设计

研究目的是研究设计的目标和方向,科学研究的基本要素及其基本原则是科研设计的基础和指南。完整的科研设计包括专业设计和统计设计两部分:专业设计是指课题的实际意义和研究价值,入选对象的诊断标准、纳入标准及排除标准等,决定研究课题的先进性和实用性;统计设计包括选择研究类型与设计方案,确定研究总体、样本量、观察指标、随机化分组或抽样方法,以及数据的质量控制和统计分析方法等,影响课题的可信度和科学价值。因此,正确的统计学分析一定要建立在明确的研究目的和研究设计的基础之上,那些事先没有研究目的和研究设计,事后找来一堆数据进行统计分析都是不可取的。在医学论文的撰、编、审、读过程中经常遇到的问题是研究的题目与课题设计、论文内容不符,包括文章的方法解决不了论文的目的、文章的结果说明不了论文的题目、文章的讨论偏离了论文的主题;还有是目的不明确、设计不合理。如题目过小,论文不够字数,而一些无关紧要的变量指标或结果被分析被讨论;又如题目过大,论文的全部内容不足以说明研究的目的,使论文的论点难以立足^[3-4]。所以,合理明确的论文题目或目的以及研究设计方案是撰、编、审、读者应当关注的首要问题。此外,样本含量是否满足,抽样是否随机,偏倚是否控制等,也是不可忽视的问题。

二、建好分析用的数据库

数据库即存放数据的“仓库”,是指将不同研究对象不同观测指标的观察结果逐一有序记录的二维表格形式。二维表中除第一行属于观察指标外,其余每一行代表一个观察对象的所有观察指标值(即数据);每一列代表某项观察指标所有观察对象的观察值。严格的数据库数据可以直接应用相关软件进行统计分析。由于不同软件对文字存在可识别性问题,一般在统计分析时要求数据库的数据值全部用阿拉伯数字表示,必要时可在适当位置附加批注。对于论文作者来讲,统计分析需要借助于

统计分析软件计算,而统计分析软件都要有完整、符合要求的数据或数据库,所以建好分析数据库是统计分析的需要。此外,建好分析数据库还可以理清分析思路。在试验或调查研究中获取的数据有时多而零散,如果不能进行科学的整理汇总,就会显得杂乱无章,理不清头绪,抓不住要点,甚至无所适从,最后可能束之高阁、弃之不用,造成数据的极大浪费。相反,建好数据库,可以使观察对象的研究指标一目了然,使研究思路清晰明确。因此,建好数据库是正确统计分析的前提和基础,甚至决定了论文分析结果的成败。对于编、审、读者来讲,一般由于篇幅的限制,往往得不到数据库数据,而只有作者在数据库数据基础上经统计描述计算后给出的诸如各指标均数 $\bar{x}$ 、标准差 s 或中位数 M 、百分位数 P_x 的“二手”数据,或将研究对象的某一指标按其数值大小或特征属性分组,清点各组观察单位出现的个数或频数的频数表数据等。无论是否能够得到数据库数据,作者在统计分析过程中一定依据数据库数据进行计算,得出结果。如果对“二手”数据或频数表数据的结果等存在疑惑,编辑、审稿专家或读者有权要求作者提供数据库数据以检查其完整性、准确性和真实性,确保研究数据的质量。假若在投稿须知中对数据库数据作出必要的要求,无疑对于保证刊物的发表质量有着积极的意义。

三、分清楚指标(或变量)的性质和类型

指标,即观察指标,是由研究目的确定的观察对象的内在属性特征或其相关的影响因素。例如,需要研究本体感觉训练对脑卒中偏瘫患者运动功能(本体感觉、平衡功能)的影响^[5],那么本体感觉、平衡功能反映了脑卒中偏瘫患者运动功能的特征,分别称为研究的本体感觉指标、平衡功能指标,影响本体感觉和平衡功能的有关因素,比如年龄、性别、病种、病程等,称为研究的年龄指标、性别指标、病种指标和病程指标。变量即观察变量,也称变化的量,实际上就是观察指标,一般特指用于数学、统计或软件计算的分析指标。例如,脑卒中偏瘫患者运动功能的本体感觉、平衡功能指标,在统计计算时,分别称为本体感觉变量和平衡功能变量。

按变量是否影响其它变量或是否受到其它变量的影响有影响变量和结果变量之分。影响变量,也称自变量,是指自身变化并影响结果变量变化的量;结果变量,又称因变量,是指随影响变量变化而变化的量,看作是影响变量变化的结果。如果分析康复训练对冠心病患者有氧运动功能的影响,那么康复训练可看作是影响变量,有氧运动功能则为结果变量^[6];如果分析不同性别之间冠心病患者有氧运动功能是否存在统计学差异,那么性别是影响变量,有氧运动功能是结果变量。分清楚变量的性质,即什么是结果变量、什么是影响变量,是选择统计分析方法的第一步。一般而言,那些相对固有的、不易改变的特征(如性别、籍贯等)或易于被人控制的处理因素(如实验分组、疫苗接种与否等)作为影响变量或影响因素;而那些容易变化、较难确定的观察效应或结局(如疗效、患病与

否等)作为结果变量,看成是最后观察的结果。但影响变量和结果变量的划分是相对的,视研究目的和具体情况而定,有时甚至不加区分。

从数据库、数据分析的角度来看,变量是指那些能反映数据库数据的内在数量关系,可用于统计计算包括软件计算的指标。一般而言,不同的研究目的决定了不同的数据库,实际上决定了组成数据库的不同变量。变量的类型分为数值变量和分类变量。数值变量,又称定量变量,是指能用定量方法测定的、具有数值大小、高低或多少的指标,变量值一般有度量衡单位,可以带小数点,如身高、体重、血压等;分类变量,又称定性变量,是指能用定性的方法确定的、观察单位某项属性或特征分类的指标。根据分类变量的分类项数和各项数间有无等级程度差异分为二项分类变量、多项无序分类变量、多项有序分类变量,如表 1。

表 1 分类变量的不同类别与举例

类别	项数	等级次序	举例
二项分类变量	二项	无或有	性别(男、女)、考核(合格、不合格)
多项无序分类变量	多项	无	血型(A、B、AB、O)
多项有序分类变量	多项	有	营养状况(优、良、中、差)

从应用统计学选择统计分析方法的角度考虑,变量可考虑分为数值变量、多项有序分类变量、多项无序分类变量、二项分类变量四种。此外,不同类别变量可遵循下列顺序转化:数值变量→多项有序分类变量→多项无序分类变量→二项分类变量,称为降级转化,但这种转化过程会不断丧失蕴藏的数据信息,导致统计分析过程中假阴性结果的不断增加。至于逆向转化即升级转化,尽管理论上认同,但实际应用中不建议采用。

很多研究表明,掌握好统计分析的应用条件,正确选择统计分析方法是学习并应用统计学的一个突出难点^[7]。对于医学论文作者而言,分清楚数据库中变量的性质(影响变量与结果变量)、类型(数值变量、多项有序分类变量、多项无序分类变量、二项分类变量)以及它们之间的降级转化关系(数值变量→多项有序分类变量→多项无序分类变量→二项分类变量)是学好用好应用统计分析的基础,可以有效避免张冠李戴、缺乏原则地选错统计分析方法;对于文章的编审和读者来说,这是判断作者正确选择统计学分析方法与否的一个简单

有效的途径。

四、正确选用统计学方法

应用统计学(严格而言是指统计学的假设检验)可以简单地看作是一门关于结果变量与影响变量之间关系分析的科学。由于结果变量(因变量)、影响变量(自变量)各有 4 种类型,所以相互组合有 16 种情形,相对应的有 16 种首选的统计分析方法(表 2 中第一个或用☆表示的方法),如,二项分类变量与二项分类变量关系的分析选用两个率比较的 χ^2 检验(四格表 χ^2 检验),二项分类变量与多项无序分类变量关系的分析选用多个率比较的 χ^2 检验,多项无序分类变量与二项分类变量关系的分析选用两个构成比比较的 χ^2 检验,多项无序分类变量与多项无序分类变量关系的分析选用多个构成比比较的 χ^2 检验;数值变量与二项分类变量关系的分析选用 t -检验,数值变量与多项无序分类变量关系的分析选用完全随机设计的 F -检验,数值变量与数值变量关系的分析选用 Pearson 直线相关回归分析^[8-9],等等。如果首选统计方法的条件不适合,一般通过降级转化选择“低”一级或“低”二级、三级的统计方法或其它统计方法。如, t -检验是数值变量与二项分类变量关系分析时首选的统计方法,如果该方法的条件不适合,此时将 t -检验中数值变量“降级”当作多项有序分类变量看待,故可次选 Wilcoxon 秩和检验,如果再“降级”,依次低选两构成比比较的 χ^2 检验,甚至四格表 χ^2 检验。又如,如果 Pearson 直线相关回归分析的条件不符合,可根据情况将其中的一个或两个数值变量“降一级”,选择 Spearman 等级相关,如果再“降级”,相应可以选择秩和检验、Logistic 回归或者 t -检验、 χ^2 检验,等等。其它仿此,详见表 2。它涵盖了基本统计分析的绝大部分,是应用统计分析的核心内容。

当然,应用统计分析除了单一变量分析、两变量间关系的分析以外,其它诸如一个自变量和多个因变量、多个自变量和多个因变量之间关系的分析当属多变量关系分析的内容。由于分类变量与数值变量各不相同,不同个数不同变量的组合方式多种多样,所以相应的统计方法也有很多种,主要有:1 个数值变量与多个数值变量之间的关系,如多元相关回归分析^[10];1 个分类变量与多个数值变量之间的关系,如多因素方差分析^[11]、重复设计方差分析^[12];1 个数值变量与混合多个变量之间的关系,如协方差分析^[13]、COX 模型^[14];1 个分类变量与混合多变量之间的关系,如 Logistic 回归分析;多个数值变量与多个数值变量之间的关系,如典则相关等^[15-16]。

表 2 两变量关系分析的统计方法

影响变量(自变量)	结果变量(因变量)			
	二项分类变量	多项无序分类变量	多项有序分类变量	数值变量
二项分类变量	四格表卡方检验(两样本率分析)	2×C 卡方检验(两样本构成比分析)	①Wilcoxon 秩和检验☆ ②R×C 卡方检验	① t 检验☆ ②Wilcoxon 秩和检验
多项无序分类变量	R×2 卡方检验(多样本率分析)	R×C 卡方检验(多样本构成比分析)	①Kruskal-Wallis H 秩和检验☆ ②R×C 卡方检验	①完全随机设计的方差分析☆ ②Kruskal-Wallis H 秩和检验
多项有序分类变量	①二项 Logistic 回归☆ ②R×2 卡方检验	①无序多分类 Logistic 回归☆ ②R×C 卡方检验	①Spearman 等级相关☆ ②有序多分类 Logistic 回归 ③R×C 关联性分析 ④Kruskal-Wallis H 秩和检验	①Spearman 等级相关☆ ②完全随机设计的方差分析 ③Kruskal-Wallis H 秩和检验
数值变量	①二项 Logistic 回归☆ ②R×2 卡方检验	①无序多分类 Logistic 回归☆ ②R×C 卡方检验	①Spearman 等级相关☆ ②有序多分类 Logistic 回归☆ ③R×C 关联性分析	①直线相关回归☆ ②Spearman 等级相关

因此,医学论文的作者,在分清楚数据库中分析变量的性质、类型和降级转化关系时,可以应用表 2 迅速确定首选的统计方法以及备选或次选的统计方法,应用相关软件或计算工具快速实现统计分析。医学论文的编审和读者也可应用该表 2,准确判断作者是否正确选择了统计学分析方法,甚至分析错误选择统计方法的原因与后果。如,欲评价某种药物的降舒张压效果,试验组用该降压药、对照组不用药,假如测量的舒张压值符合 t 检验的条件(正态分布、方差齐等),很明显该数据首选 t 检验的统计方法^[17]。如果降级转化可以选择 Wilcoxon 秩和检验,如果降压效果转化为有效和无效两种情况,甚至可以选择四格表 χ^2 检验。但需要注意的是,如此降级选择统计方法,可能出现假阴性或漏诊错误,即把差异有统计学意义的结果(有降压效果)当作差异无统计学意义的结果(无降压效果)看待,从而低估药物的作用,在论文讨论中至少要加以必要的说明,否则统计方法不能视为正确有效,可作为退修或退稿处理。

五、熟悉常用的统计分析软件

统计分析软件是统计分析的必备工具,常用的统计分析软件有:统计分析系统 SAS、社会学统计程序包 SPSS、微软公司电子表格系统 Microsoft Office Excel 等。

SAS(statistics analysis system)是统计分析系统的英文简称,最早由北卡罗来纳大学的两位生物统计学研究生编制,1976 年由 SAS 软件研究所正式推出。SAS 完全针对专业用户进行设计,以编程为主。其最大特点是分析模块调用,功能强大,深浅皆宜,简短编程即可同时对多个数据文件进行分析。但对一般用户而言,人机界面不太友好,最初编写使用程序时可能会存在各种难度^[18]。

SPSS(statistical package for the social science)是社会学统计程序包的英文简称,20 世纪 60 年代末由美国斯坦福大学的 3 位研究生研制,1975 年由芝加哥 SPSS 总部推出。SPSS 系统的最大特点是菜单操作,方法齐全,绘制图形、表格较为方便,输出结果比较直观。但其统计分析功能略显逊色,特别是难以同时分析处理多个数据文件^[19]。

Microsoft Office Excel 是美国微软公司开发的电子表格系统,是目前应用最为广泛的办公室表格处理软件之一。Excel 作为 Office 软件的一员被众多用户所熟知,具有数据处理、函数运算、数据库、图表制作等功能,进行统计分析时具有易得、快速、直观、简单、运算可视等优点,是建立数据库,并进行常用统计分析的好工具^[20]。

其中,SAS、SPSS 是国际通用的统计分析计算软件。即便如此,不同软件仍各有利弊、互有长短,用户可根据需求和使用习惯,选择一种或几种软件进行数据分析。特别一提的是 Microsoft Office Excel,由于其独特的优势^[20],统计计算功能也逐渐得到开发应用,如 Excel 统计分析程序^[21]等,必将得到广大科技论文作者、编审和读者认可与使用。

参 考 文 献

- [1] 孙振球,徐勇勇. 医学统计学[M]. 北京:人民卫生出版社,2010:1-6.
- [2] 袁卫,刘畅,张云. 我国统计教材建设的历史回顾与现实思考[J]. 统计研究,2004(6):55-60.
- [3] 肖丽娟. 从编辑角度谈医学论文写作中的统计学应用问题[J]. 苏州大学学报,2006,26(5):90-92.
- [4] Rumsey DJ. The challenge of developing statistical literacy, reasoning and thinking [M]. Berlin: Springer Netherlands,2004:3-15.
- [5] 谢凌峰,黄晓琳,黄杰,等. 本体感觉训练对脑卒中偏瘫患者运动功能及日常生活活动能力的影响[J]. 中华物理医学与康复杂志,2012,34(8):592-595.
- [6] 郑华,原晓晶. 康复训练对冠心病患者有氧运动功能的影响[J]. 中华物理医学与康复杂志,2012,34(4):293-295.
- [7] 伍亚舟,易东,张彦琦,等. 探讨医学院校医学统计学教育教学的困境与改革[J]. 医学教育,2013,42(33):4096-4097.
- [8] 胡良平,李子健. 医学统计学基础与典型错误辨析[M]. 北京:军事医学科学出版社,2003:4-5.
- [9] 肖丽娟,孙茂民. 医学论文中统计学处理常见问题及应对措施[J]. 编辑学报,2010,22(6):500-502.
- [10] 黄霞,胡玉民,陈星,等. 脑瘫儿童母亲生活质量及其影响因素分析[J]. 中华物理医学与康复杂志,2013,35(1):13-17.
- [11] 郑光新,赵晓鸥,常智跃. 等速运动系统测试膝关节本体感觉功能的信度研究[J]. 中华物理医学与康复杂志,2013,35(8):609-612.
- [12] 高春华,徐乐义,黄杰,等. MOTomed 智能运动训练系统对脑卒中偏瘫患者平衡及下肢运动功能的影响[J]. 中华物理理论与实践,2013,19(8):609-612.
- [13] 何建明. 蒙特利尔认知评估表与简易精神状态检查评定老年人轻度认知功能障碍的比较[J]. 中华老年医学杂志,2010,29(6):725-728.
- [14] 邵子力,梁惠宏,曹良启,等. Cox 模型多因素分析影响复发性肝癌再切除的预后因素[J]. 国际外科学杂志,2011,38(7):510-513.
- [15] 王保进. 多变量分析-统计软件与数据分析[M]. 北京:北京大学出版社,2007:55.
- [16] Hardle W, Simar L. Applied multivariate statistical analysis[M]. Berlin: Springer Berlin Heidelberg,2003:3-18.
- [17] 胡良平,高辉. 如何正确运用 t 检验[J]. 中西医结合学报,2008,6(2):209-212.
- [18] 洪楠,侯军. 国外统计软件的特色与风格简介[J]. 现代医学仪器与应用,2008,20(1):65-67.
- [19] 李国春. SPSS 统计软件简介[J]. 循证医学,2003,3(4):249-252.
- [20] 唐朝霞,赖伏虎,王舒,等. Excel 在卫生统计工作中的综合应用[J]. 中国医院统计,2006,13(3):287-288.
- [21] 陈青山,顾大勇. Excel 统计分析[M]. 广州:暨南大学出版社,2012:5-10.

(修回日期:2014-05-30)

(本文编辑:汪 玲)