

• 流行病学与统计学方法 •

统计学方法的合理选择和常见误区

史芸萍 刘军廷 么鸿雁 于石成 王琦琦 胡跃华

【摘要】 统计学是帮助人们透过“偶然性”掌握事物发展规律的重要工具,但现状却是很多学者不重视统计学甚至不具备基本的统计技能。本文以“统计资料类型”为着手点,对统计学方法的选择进行了系统介绍,并汇总了文献中常见的统计方法误区,期望引起广大学者对统计学方法的重视。

【关键词】 统计学; 人口统计学; 生物统计学

Proper selection and common pitfalls of statistical methods SHI Yun-ping*, LIU Jun-ting, YAO Hong-yan, YU Shi-cheng, WANG Qi-qi, HU Yue-hua. * Department of Statistics and Information, Beijing Center for Diseases Control and Prevention, Beijing 100013, China

Corresponding author: HU Yue-hua, Email: huyueer@163.com

【Abstract】 Statistics is an important tool that helping people master the law of things through the accidental, however, currently many scholars do not care statistics enough or even have no basic statistical skills. This article systemly introduces the selection of statistical method based on types of statistical data, as well as summarizes common statistical methods pitfalls in literature, in order to make scholars pay more attention to statistics.

【Key words】 Statistics; Demography; Biometry

统计学的应用贯穿于整个科学研究过程中,包括研究设计、资料收集整理、分析方法和软件选择,以及最终的结果解释。英国著名统计学家高尔顿曾说:“统计学具有处理复杂问题的超凡能力,当科学探索者在前进的过程中荆棘载途时,惟有统计学可以帮助他们打开一条通道。”可是,统计学也很容易被误用或滥用。*Nature* 杂志于 2014 年发表的评论文章 *Number crouch* 中提到:很多研究者缺乏统计学基本知识,普遍存在统计学方法误用现象,在一些基础科学领域这种情况十分严重^[1]。关于文献中的统计学问题,早在二十世纪 30 年代就已展开讨论^[2],来自不同领域的研究者们发现,很多科学文献中存在统计学错误,甚至包括那些发表在影响因子很高的杂志的文章^[3-4]。二十世纪 80 年代,国外的医学文献审稿人发现,统计学方法使用不正确的文章大约占 50%^[5]。我国学者也指出:“我国科技期

刊,尤其是医学期刊中,存在大量的误用和滥用统计学方法的现象^[6]。笔者就统计学方法的合理选择进行汇总分析,并指出统计学方法选择的常见误区,旨在帮助读者避免常见错误,正确使用统计学方法。

一、统计资料分类

合理选择统计学分析方法的首要前提是正确识别统计资料类型。资料类型的现代划分法将资料先分成定量资料和定性资料两大类。其中,定量资料是测量每个观察单位某项指标值大小所得到的资料;定性资料是观测每个观察单位某项指标质的状况所得的资料。

定量资料又划分为计量资料和计数资料。计量资料是指指标的取值可以带度量衡单位,甚至可以带表示测量精度的小数位的定量资料,例如测量正常成年男女的身高(cm)、体质量(kg)、血红蛋白(g/L)、体温(℃)等所得的资料。计数资料是指指标的取值可以带度量衡单位,但不可以带小数,只能取整数,通常为正整数的定量资料;例如,测得正常成年男子脉搏数(次/min)、引体向上次数(次/min)、现有子女数(个)等所得的资料。

定性资料又划分为名义资料和有序资料。名义资料是指指标的不同状况之间在本质上没有数量大小或先后顺序之分的定性资料,可分为二项分类名义资料和多项分类名义资料。二项分类名义资料举

doi:10.3969/j.issn.1000-6621.2016.05.004

基金项目:中国疾病预防控制中心青年科研基金课题(2016A201)

作者单位:100013 北京市疾病预防控制中心信息统计中心(史芸萍);首都儿科研究所流行病学研究室(刘军廷);中国疾病预防控制中心流行病学办公室(么鸿雁、于石成、王琦琦、胡跃华)

通信作者:胡跃华,Email: huyueer@163.com

例:检查某小学学生大便中的蛔虫卵,以每名学生对观察单位,结果分为蛔虫卵阳性或阴性所得的资料;观察某药治疗某病患者的治疗结果,以每例患者为观察单位,结果分为治愈和未治愈所得的资料。多项分类名义资料举例:某学校全体师生按血型 A、B、O、AB 来记录每个人的情况所得的资料;某乡镇全体居民按人群分类(散居儿童、幼托儿童、学生、教师、农民、工人、干部职员、家务及待业等)记录每个人的情况所得资料。有序资料是指标的不同状况之间在本质上有数量大小或先后顺序之分的定性资料,例如:测定某化工厂全体职员血清反应,结果可分为“一、士、十、++”四级所得到的资料;观察某药治疗某病患者的治疗效果,分为治愈、显效、好转、无效、死亡所得到的资料。

二、描述性统计方法

(一)常用方法

对于定量资料,常用的描述指标分为两类,一类是描述数据分布集中趋势的指标,另一类是描述数据分布的离散程度的指标。其中,描述集中趋势的指标主要包括算术均数、几何均数、中位数;描述离散趋势的统计指标主要包括极差、四分位数间距、方差、标准差、变异系数。

对于定性资料,常用的描述指标有相对数,包括比和率。相对数可以分为结构相对数和强度相对数,比反映部分与整体或某一部分与另一部分之间的关系,属结构相对数;率反映某种事物或现象发生的强度,属强度相对数。常用的相对数指标包括率、构成比、相对比、动态数列的定基比和环比。

(二)定量资料常见误区

1. 使用均数描述偏态资料的集中趋势:算术均数适用于描述对称分布资料,特别是满足正态分布资料的集中趋势;几何均数适用于原始数据不对称、但对数转换后呈对称分布的资料。而中位数则适用于任何分布的资料,包括对称的、偏态的、开口的。需要根据资料分布情况来选择合适的描述指标,不能一出现对资料进行描述就采用均数加减标准差的形式。例如,描述正常人血铅含量时,因血铅含量呈偏态分布,不应再用均值来描述。《中国防痨杂志》2015 年 11 月第 37 卷第 11 期的一篇文章^[7]中,在描述“取样时病程”和“取样时抗结核治疗时间”也是采用均数加减标准差的形式,但文献中数据显示,标准差与均值接近甚至标准差大于均值,数据资料明显不服从正态分布,应采用中位数(四分位间距)描述更为恰当。

2. 混淆标准差和标准误:标准差用来描述数据分布的离散程度,表示观察值之间波动的大小;而标准误用来描述样本均数间的离散程度,可以表示均数抽样误差的大小。前者可与均数一起计算正常值范围,后者可计算置信区间,文献中有混用现象。

(三)定性资料常见误区

1. 构成比与率混淆滥用:构成比只能说明事物各组成部分的比重或分布,率是用来描述某种现象发生的频率或强度。两者概念和计数方法都不同,所得结论也不同。文献中容易发生用构成比代替率来表达强度,并得出错误结论。还要注意,不能用构成比的动态分析代替率的动态分析。例如文献^[7]中,描述研究对象的一般情况时,制作的表 1 将构成比以“率(%)”来描述,且表中数据与文中描述不符。

2. 计算相对数的分母过小:如果观察单位过少,那么偶然性就大,则其可靠性就差。例如,观察某药物疗效时,5 例患者,4 例有效,1 例无效,不能说该药物有效率为 80%。一般当观察例数较少时,如观察例数少于 30 例时,采用绝对数表示为好。通常观察单位足够多时,计算出的相对数比较稳定,能够正确反映实际情况。

3. 相对数比较时忽略了可比性:影响率或构成比变化的因素很多,除了要比较的研究因素外,其余的影响因素应尽可能相同或相近。一旦其他影响因素不同或者差异很大而直接进行分析和比较时,往往会得到错误的结果和结论。但实际应用中,很多学者只关注研究因素,而把其余的影响因素,如观察时间、地区、民族、性别、年龄等因素忽略了。例如,分析某工厂不同工种工人的疾病发病规律时,要考虑年龄、工龄、工种、病程、环境条件等因素;比较两地人口某病发病及死亡情况时,应考虑两地人口构成。《中国防痨杂志》2015 年第 37 卷第 9 期中的一篇文章^[8]对正常组和异常组利福平血药浓度进行了比较,但治疗时间的长短对血药浓度有直接影响,文中就未考虑或未对治疗和服药时间因素进行交代。

三、推断性统计学方法

(一)常用方法

对于定量资料,常用到的基本的统计学分析方法有 *t* 检验、方差分析、相关与回归等。要想做到合理选用定量资料统计分析方法,关键在于做好以下两点:第一,正确判断定量资料所应用的实验设计类型;第二,检查资料是否满足“正态分布、独立性和方差齐性”3 个前提条件。

对于定性资料,当原因变量全是定性变量时,常

用到的基本的统计学分析方法包括卡方检验、秩和检验、Spearman 秩相关检验、Kappa 检验等;当原因变量既有定性变量,又有定量变量时,需要根据实验目的来选择是将资料离散化还是设置哑变量采用 logistic 回归分析方法;当原因变量全为定量变量时,可直接采用 logistic 回归分析方法。

(二) 定量资料常见误区

1. 直接进行分析未判断方法的前提条件:文献中会出现拿到定量资料后不进行判断甚至在不满足前提条件的情况下,仍旧直接进行统计学分析。但不论是 t 检验、方差分析还是相关与回归,都有着方法自身的前提条件。如 t 检验要求资料正态、独立和方差齐性;直线相关分析要求双变量满足双变量正态分布等。当资料不满足条件时,可寻找其他分析方法。例如, t 检验如果资料不满足正态性条件,可以对数据进行变量变换后再进行正态性检验,然后进行分析;或者,可以采用非参数检验方法。

2. 选择统计学分析方法时未正确判断实验设计类型:多因素实验设计本身就复杂,再加上收集资料表达形式多样化(未必一定是标准的样子),一旦缺乏一定的统计能力,就会被资料的表象所迷惑,不能正确判断实验设计类型,例如,将两因素析因设计看成单因素四水平设计或成组设计。因此,成组、配伍组、配对、交叉等基本实验设计方法一定要分清楚。

3. t 检验的误用:不考察定量资料所对应的实验设计类型,不检查定量资料是否满足“正态、独立和方差齐性”的条件,只要看到资料为定量资料,就使用 t 检验进行分析。成组设计、配对设计、单因素多水平设计的 t 检验互相混用,甚至对多组均数的比较也采用 t 检验进行两两比较。

4. 方差分析的误用:医学文献中经常涉及到重复测量资料,因重复测量数据间存在一定的自相关性和随机误差的多层次性,直接使用普通的方差分析方法来分析重复测量资料是不妥的。正确判断资料实验设计类型,如果为重复测量资料,应该采用重复测量的方差分析方法进行分析。

(三) 定性资料常见误区

1. 对于四格表,不区分普通卡方与确切概率法:很多人在看到普通四格表定性资料时,随手就进行普通卡方检验。其实应首先判断资料的情况。例如,对于普通的四格表,当样本量 ≥ 40 并且所有理论频数 > 5 时,可以用普通的卡方检验;当样本量 ≥ 40 而有理论频数在 1~5 之间时,应使用校正的卡

方检验;当样本量 < 40 或者有理论频数 < 1 时,则应该使用确切概率法,不能在应该使用确切概率法时还在用普通的卡方检验。而在分析 $R \times C$ 表资料时,除了上述方法外,一旦分析的数据出现了理论频数 < 1 ,或者理论频数 ≥ 1 且 < 5 的格子数超过格子总数的 $1/5$,还可以通过增加样本量,使理论频数增大;或者根据专业知识,删去理论频数太小的行或列,或者看能否将理论频数太小的行或列与性质相近的邻行或邻列进行合并来解决。

2. 对于复杂的 $R \times C$ 列联表,不判断资料是否有序:不同情况的列联表,需选取不同的分析方法。对于复杂的 $R \times C$ 列联表,要在错综复杂的实际问题中,看清事物的本质面目。例如,单向有序列联表可采用“秩和检验或 Ridit 分析,或有序资料的 logistic 回归模型”进行处理;两有序变量之间的相关关系可采用 Spearman 秩相关分析或典型相关分析进行处理;两有序变量之间的线性变化趋势可采用线性趋势检验进行分析。对于高维列联表资料,根据分析目的,则可选用加权卡方检验、对数线性模型、一般的 logistic 回归模型或有序资料的 logistic 回归模型进行分析。

综上所述,只有在科学地进行实验设计,合理地进行资料收集和整理,准确地判断实验类型的前提下,才有可能选取正确的统计学分析方法。在避开常见统计学错误的同时,还需注意统计图表的表达、统计结果的表述等,也需要遵循统计学要求。

参 考 文 献

- [1] Number crunch. Nature, 2014, 506(13):131-132.
- [2] Mainland D. Chance and the blood count. Can Med Assoc J, 1934, 30(2):225-227.
- [3] Gore SM, Jones G, Thompson SG. The Lancet's statistical review process: areas for improvement by authors. Lancet, 1992, 340(8811):100-102.
- [4] Schor S, Karten I. Statistical evaluation of medical journal manuscripts. JAMA, 1966, 195(13):1123-1128.
- [5] Glantz SA. Biostatistics: how to detect, correct and prevent errors in the medical literature. Circulation, 1980, 61(1):1-7.
- [6] 胡良平,李子建. 医学统计学基础与典型错误辨析. 北京:军事医学科学出版社,2003:2.
- [7] 梁瑞霞,谷蕴婷,董伟杰,等. 两种分子检测技术快速诊断骨关节结核及其耐药性的研究. 中国防痨杂志, 2015, 37(11):1126-1129.
- [8] 邓国防,孙丽珍,詹森林,等. 有机阴离子转运多肽 *1B1* 基因多态性和利福平血药浓度对肝毒性的影响. 中国防痨杂志, 2015, 37(9):933-937.

(收稿日期:2016-04-17)

(本文编辑:李敬文)