·讲座·

医学科研中合理运用统计学的方略

——定性资料统计分析方法合理选用(1)

胡良平[1],单彬[1],刘惠刚[2],李子建[1]

(1. 北京军事医学科学院生物医学统计咨询中心,北京 100850; 2. 首都医科大学继续教育学院,北京 100036)

[中图分类号] R195.1 [文献标识码] C [文章编号] 1008-8830(2005)01-0095-03

1 引言

人们在从事医学科研工作中,经常需要处理定性资料。很多人习惯于用 χ^2 检验处理一切定性资料,这显然是不妥的,因为每一种统计分析方法都有其前提条件和适用场合。盲目套用,不仅达不到科学、严谨之目的,反而降低了说服力和可靠性。如何才能做到合理地选用统计分析方法处理定性资料呢? 首先,要弄清什么是定性资料,其次,要知道合理选用统计分析方法处理定性资料的要领,第三,才是学会具体的处理定性资料的统计分析方法。

2 何为定性资料

定性资料是指对每个研究对象的某些方面的特征和性质进行表达或描述所得的资料,其具体的取值要么是名义的,如血型(A、B、O、AB)、职业(工人、农民、军人、学生)、性别(男、女),等等;要么是有序的或等级的,如疗效(治愈、显效、好转、无效、死亡)、抗体滴度(+、++、+++、++++),等等。这些定性资料有些属于"原因"、有些属于"结果",若将每个受试者的定性变量的具体取值一一列出,则不便看出资料之间内在的联系,故人们常以表格的形式对资料进行整理或归纳,这种表格被称为"列联表"。列联表的类型很多,针对不同的类型应采用不同的统计分析方法。下面有一个实际的例子,原文作者就很肯定地认为其资料是定性资料,并在其论著的统计分析部分明确交代采用 χ^2 检验对资料进行了处理。试分析产生这种错误的根源是什么? 如何避免这种错误的再次发生?

例 1:原文题目:小儿皮肤血管瘤雌、孕激素受体的研究。原作者意在探讨雌激素受体(ER)、孕激素受体(PR)在血管瘤发生、发展中的意义。采用免疫组化方法对毛细血管瘤、混合型血管瘤、海绵状血管瘤、淋巴管瘤及正常皮肤组织的 ER、PR 受体进行检测。全部标本经 10% 福尔马林固定,常规石蜡包埋。每例选一典型蜡块,4 ~ 6 μm 切片,进行免疫组化染色,高倍镜下每例肿瘤区内计数 500 个细胞,计数 ER、PR 阳性细胞百分率,见表 1。统计方法用 χ^2 检验。

表 1 血管瘤、淋巴管瘤中 ER、PR 阳性率检测结果($\bar{x} \pm s$)

类别	例数	ER(%)	PR(%)
毛细血管瘤	45	74.18 ± 11.77	77.92 ± 10.54
混合型血管瘤	44	64.55 ± 12.34	68.12 ± 15.38
海绵状血管瘤	18	23.00 ± 7.89	25.12 ± 9.66
淋巴管瘤	23	26.93 ± 15.62	30.00 ± 18.87
正常皮肤	6	9.83 ± 6.69	11.00 ± 4.56

解说:正确判别统计资料的性质是合理选择统计分析方法的重要前提。根据统计指标的性质,统计资料一般分为定量资料和定性资料两大类,所谓定量资料,是指从每个观察单位(针对此资料,其观察单位是病例标本)上测得的指标是用具体的数值表示的。其又细分为计量资料(一般带度量衡单位和小数点)和计数资料〔一般只带度量衡单位而不带小数点,如脉搏次数(次/min)〕。所谓定性资料,是指从每个观察单位上测得的指标仅反映某一方面的性质,并不能用具体的数值表示。其又细分为名义资料和有序资料。对于本资料来说,测量细胞的结果是"阳性"或"阴性",且一般认为带有"率"的

[收稿日期]2004-04-05
[作者简介]胡良平(1955-),男,教授。主攻方向:实验设计与统计学。

资料就是定性资料,似应判为定性资料,然而问题的关键在于,原作者的观察单位并不是细胞本身,而是每一个病例标本,原作者关心的是四种疾病病例标本和一组正常人标本的 ER、PR 阳性细胞率之均值是否相同,从每一个病例标本中得到的是 ER 和 PR 阳性细胞率,是一具体的数值,因而应属于定量资料。如果仅从资料的表面现象(有"率")进行判断,而不考虑每一个数值的实际含义,没有从资料的本质上进行判断,很容易判断错误。

下面让我们再看一个资料,分析一下它是定性资料还是定量资料。

例 2:原文题目:美喘清与博利康尼治疗支气管哮喘各 40 例临床疗效与副作用比较。原作者选择 80 例哮喘病人随机分为美喘清组与博利康尼组各 40 例,记录各组病人发生疗效的时间,见表 2。所得结果用 χ^2 检验进行处理,认为美喘清较博利康尼发生疗效的时间早,且差异有显著性($P<0.05$)。

表 2　美喘清与博利康尼疗效发生时间比较　(h)

组别	例数					
	0.5	1	2	24	48	72
美喘清	8	9	8	8	4	3
博利康尼	2	4	6	8	10	10

解说:严格地说,每个哮喘患者都能提供一个药物发生疗效的时间,因而此资料从本质上讲应为定量资料,表 2 只是为了表达的方便列出不同时间点上的频数分布,并不代表此资料中的结果变量就为定性资料。原作者采用一般 χ^2 检验对资料进行处理,χ^2 检验所能回答的问题与原作者的分析目的不一致。此时得出的结论只能是美喘清组和博利康尼组在不同起效时间的构成上存在的差别是否具有统计学意义,并不能得出两组起效时间之间的差别具有统计学意义。最好将此资料中的发生疗效的时间还原为原始值,然后按成组设计定量资料进行统计学分析。如果每组发生疗效的时间符合正态分布且两组发生疗效的时间满足方差齐性,则可以进行成组设计资料的 t 检验,如果不满足 t 检验的前提条件,则可采用非参数统计,如成组设计两样本比较的秩和检验。如果每位患者药物起效时间不像表 2 所表示的那样精确,只是一个时间段(如:0 ~ ≤0.5, >0.5 ~ ≤1, ……, >48 ~ ≤72,通常的表达方式为:0 ~ ,0.5 ~ ,1.0 ~ ,2.0 ~ ,24.0 ~ ,48.0 ~ 72.0),此时的表 2 资料叫做"结果变量为有序变量的单向有序列联表资料",可以选用的统计分析方法有:秩和检验,Ridit 分析等。

3　定性资料统计分析方法合理选择要领

我们知道,若观测结果为定性资料(如治愈、未愈),一般不以个体为计量单位,而是以处理组为计量单位,换句话说,资料以分组且用表格的形式呈现出来,这种表通称为"列联表"。列联表有二维列联表和高维列联表之分。具体地说,就是列联表中涉及几个定性变量就称为几维列联表。一般来说,一个列联表中只有一个结果变量,其他都是原因变量,但也有少数列联表中的变量不包含结果变量,此时只能考察全部定性变量全部水平组合下的频数分布情况。如何才能合理选用统计分析方法处理列联表资料呢?其要领在于以下两点:第一,根据列联表中定性变量的性质对列联表资料进行命名,正确地说出其名称后,与其对应的统计分析方法是十分有限的几种,可选择的范围就大大缩小了;第二,根据定性资料所具备的前提条件和分析目的进一步缩小选择统计分析方法的范围,此时可供选用的方法一般只有 1、2 种,最多时可能会有 3 到 4 种。

对于二维列联表而言,一般可分为以下 4 类:第 1 类:双向无序的二维列联表。当表中小于 5 的理论频数的个数小于总格子数的 1/5 时,需要选用 Fisher 的精确检验,否则,可以选用 χ^2 检验。若是 2×2 表(或称四格表),应根据实验设计类型,选择相应的统计分析方法:若属横断面研究设计,当缺乏统计软件进行 Fisher 精确计算时,还可考虑用校正的 χ^2 检验;若属队列研究设计或病例-对照研究设计,先将其视为横断面研究设计资料处理,当得到 $P<0.05$时,在求出相对危险度 RR(队列研究设计时用)或比数比 OR(病例-对照研究设计时用)后,用 χ^2_{MH}计算公式检验 RR(或 OR)是否等于 1。第 2 类:结果变量为有序变量的二维列联表。此时,所选用的统计分析方法必须与结果变量的有序性有联系,可供选用的统计分析方法有"秩和检验"、"Ridit 分析"与"有序变量的 Logistic 回归分析"。第 3 类:双向有序且属性不同的二维列联表。此时,根据实际情况可能提出四个不同的分析目的,将对应四套分析方法:目的一:仅关心结果变量的有序性,可将其视为前面介绍的"第 2 类列联表",选择相应的统计分析方法;目的二:希望研究两有序变量之间的相关性的高低,此时宜选用定性资料的相关分析,如 Spearman 秩相关分析或典型相关分析;目的三:希望研究两有序变量之间是否存在线性变化趋势,可

以选用线性趋势检验;目的四:希望研究各实验分组中的频数分布是否相同,此时可选用一般 χ^2 检验或 Fisher 精确检验(注意:此时的结论不应该是"行变量与列变量之间呈正相关或负相关关系")。第 4 类:双向有序且属性相同的二维列联表。这种表一般都是考察用两种类似的检测方法检测同一批样品,看检测结果是否一致,故需要做一致性检验或称 Kappa 检验。若是 2×2 表时,通常称为配对设计的四格表,此时,常检验不一致部分相差是否具有统计学意义,用 McNemar χ^2 检验;当然,也可做一致性检验。但这两种检验的目的和对检验结果的解释是不同的。

对于高维列联表而言,一般可分为以下 3 类:第 1 类:结果变量为二值变量的高维列联表。可以选用多元 Logistic 回归分析或对数线性模型分析,某些特殊情况下还可选用加权 χ^2 检验。第 2 类:结果变量为多值有序变量的高维列联表。可以选用有序变量的多元 Logistic 回归分析。第 3 类:结果变量为多值名义变量的高维列联表。可以选用对数线性模型分析或扩展的多元 Logistic 回归分析。

4 结尾

本文仅从概念和方法选择角度讲述了定性资料统计分析方法合理选用方面的知识,属于"纸上谈兵"。从下一讲开始,我们将针对各种各样的实际资料,讲授如何叫出他们的名称,如何合理选择统计分析方法,如何判断资料是否具备相应统计分析方法所要求的前提条件,如何具体实现统计分析。只要读者按本系列讲座坚持学习下去,就有可能成为统计学的行家,对提高医学科研工作的科学性与严谨性会大有裨益。

[参 考 文 献]

[1] 胡良平主编. 现代统计学与 SAS 应用. 北京:军事医学科学出版社,2002:188-189.
[2] 胡良平,李子建主编. 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,2003:189-217,298-320.

(本文编辑:钟乐)