·医学统计·

相关系数的正确理解和表达*

易 东,陈 军,刘 岭,张彦琦,陈品一,伍亚舟,赵增炜
(第三军医大学卫生统计学教研室,重庆 400038)

doi:10.3969/j.issn.1671-8348.2011.34.038 文献标识码:B 文章编号:1671-8348(2011)34-3518-02

两个随机变量的相关系数在医学领域中有着广泛的应用,相关分析也是统计方法研究的核心。但是此方法在实际应用中容易出错,其主要原因是:作者对该方法的本质未彻底理解;同时该方法在实际应用中具有一些不确定性。因此,本文结合多年的教学情况和数据处理经验,简要介绍如何正确理解该方法,以及在实际应用中应该注意的问题。

1 相关系数的定义

1.1 Pearson 相关 双变量正态分布资料的直线相关性一般用 Pearson 相关系数 r 来进行描述。Pearson 相关系数又称积差相关系数,它是说明具有直线关系的两个变量间相关关系的密切程度与相关方向的指标。

$$r=\frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2\Sigma(Y-\overline{Y})^2}}=\frac{l_{XY}}{\sqrt{l_{XX}l_{YY}}} \quad (1)$$

相关系数 r 没有单位,其值为 $-1\leqslant r\leqslant 1$。

1.2 Spearman 等级相关 在实际工作中,常遇到有些资料并不呈正态分布,或总体分布类型未知,或等级资料的情形。对于这些资料,分析变量间的关系就不宜用 Pearson 相关系数来描述变量间的相关关系,而常选用 Spearman 等级相关进行统计推断。Spearman 等级相关又称为秩相关,是将原始数据资料经过秩转换后计算等级相关系数 r_s,说明两个变量间直线相关关系的密切程度和相关方向。这类方法对原变量分布不作要求,属于非参数统计方法。

$$r_s=1-\frac{6\Sigma d^2}{n(n^2-1)} \quad (2)$$

上式中:d 为每对观察值 X、Y 的秩次之差,n 为观察值的对子数。样本等级相关系数 r_s 界于 -1 与 1 之间,r_s 值为正表示正相关,r_s 值为负表示负相关,r_s 值为 0 表示零相关。

1.3 双向有序分组变量的线性趋势相关 该相关分析方法主要适用于双向有序且属性不同的行×列表中两有序变量间是否存在线性相关关系,或是否存在线性变化趋势。首先,计算行×列表的 χ^2 值,然后将总的 χ^2 值分解成两回归分量。

$$\chi^2_{总}=\chi^2_{线性回归}+\chi^2_{偏线性回归} \quad (3)$$

若两分量均有统计学意义,说明两分类变量存在相关关系,但关系不是简单的直线关系;若线性回归分量有统计学意义,偏线性回归分量无统计学意义时,说明两分类变量不仅存在相关关系,而且是线性关系。

$$\chi^2_{线性回归}=\frac{b^2}{S_b^2} \quad (4)$$

$\upsilon_{线性回归}=1$,$\upsilon_{偏线性回归}=\upsilon_{总}-\upsilon_{线性回归}$,$b=\frac{l_{XY}}{l_{XX}}$,$S_b^2=\frac{l_{XY}}{n\cdot l_{XX}}$。

1.4 偏相关分析 相关分析是分析两个变量间线性关系的程度。但是,实际中常常因为第 3 个变量的作用,使相关系数不能真正反映两个变量间的线性程度。例如身高、体质量与肺活量的关系。如果使用 Pearson 相关计算其相关系数,可以得出肺活量与身高和体质量均存在较强的线性关系。但实际上,如果对体质量相同的人分析身高和肺活量,是否身高值越大,肺活量也越大?结论是否定的。而正是因为身高与体质量有着线性关系,体质量与肺活量也存在线性关系,因此,得出身高与肺活量也存在线性关系的错误结论。偏相关关系的任务就是在研究两个变量之间的线性关系时控制可能产生影响的变量。

2 相关系数的理解

2.1 相关系数的一般理解 一般认为相关系数是衡量观测数据之间相关程度的一个指标,在众多统计学基本理论书籍中,“相关关系”可能被谨慎地定义为“变量间具有密切关联而又不能用函数关系精确表达的关系”,也可能被宽泛地定义为“客观现象之间存在的互相依存关系”。因而,相关系数就是“两个变量 X 和 Y 之间线性关系强弱的一种描述性测度”或“反映相关关系密切程度的重要指标”。概言其要,相关系数是对事物之间相互影响之密切程度的度量。

2.2 相关系数刻画表象上的数量关系,是事物之关系紧密程度的估计值 无论是 Pearson 相关系数还是 Spearman 相关系数都只是一个具体的数值,没有单位,只是刻画两个变量间表象上是否存在数量相关关系,相关系数的任务就是对相关关系给予定量的描述。

从认识无限性的角度讲,两个变量间的强弱相关关系是不可知的,要探讨它,只能以表象上的明暗程度作为实质上强弱程度的“估计值”,把探究强弱程度的希望寄托于对明暗程度的探索上。相关系数就是描述现象间联系之明暗程度的典型方法。因此,相关系数是事物相互关系密切程度的一个“投射”,逻辑上并不能认为它反映了事物本质联系的密切程度。

2.3 产生相关关系的客观起因 从事物本质属性角度看,相关关系的产生可能有以下几种情况:(1)受到干扰的因果关系。如汽车的行驶里程与耗油量,二者成正比,但受道路、风速、驾驶特点等因素影响,使这一很明显的因果关系产生了波动,从而体现出多值依赖的因果关系,表象上具有一定的非决定论色彩。(2)同一原因的诸多结果之间的关系,即一干多枝。如人的体质量与肺活量,它们都基本取决于身高,呈正相关关系,但实质上二者之间不存在因果关系。在统计学中,如果观测到的两个变量间的关系可以通过引入第 3 个变量来解释,这种关系就被称为“伪关系”或偏相关关系。(3)不存在因果关系而局部同律,比如“小孩身高”与“小树高度”呈正相关关系,这两个变

* 基金项目:国家自然科学基金面上资助项目(81172773)。

量在性质上基本相互独立，在一段时间内出现相同的走向，只能形成局部的描述性解释，而无法找出因果关系。这种情况还包括“偶然地巧合”。绝对地看，观察毕竟是有限的，由观察所归纳的“模式”只能实事求是地说明过去，把这一模式延伸到未来就存在着或然性。

2.4 对待因果关系统计规律的态度　统计规律本质上属于归纳概括，运用统计规律不能对事物的变化过程做出真实的因果论证，仅能提供具有或然性的统计描述，统计规律和运动形式规律是两种不同的规律，统计规律正是由于包含偶然因素而与因果关系相区别。

3　相关系数的错误解析

3.1 相关系数的应用条件和表达　在统计学上，根据资料类型的不同，可以分为计量资料、计数资料和等级资料，不同类型资料有不同的相关系数应用条件。若两计量资料均为正态分布资料或为双向无序行×列表资料，则可以用 Pearson 相关系数描述两变量间的相关关系；若两计量资料不全为正态分布资料或资料分布类型未知又或为单向有序行×列表资料，则可以用 Spearman 等级相关系数来描述两变量间的相关关系；若为双向有序且属性不同行×列表类型资料，则可以用线性趋势分析描述两变量间的相关关系或趋势。相关系数的正确表达为：(1)必须有专业知识为依据；(2)必须绘制散点图，并正确分析散点图；(3)计算关键的统计量(如 r、a、b)，并进行假设检验；(4)结合专业和统计学知识判断所作的统计分析是否有实用价值。

3.2 脱离专业背景知识，盲目进行相关分析　(1)实例：某人于其子出生当天在门前植小树一棵，以示纪念。后每隔一段时间，测量小树高度及其子身高，发现二者存在直线相关关系，经相关分析检测二者之间的相关系数($r=0.974\ 7$，$P<0.05$)。结论是小树身高与其子的身高存在很强的正相关关系[1]。(2)错误解析：该资料中，“小孩身高与小树高度存在直线相关关系”毫无专业依据，这样的计算结果纯属“数字游戏”，没有实际意义。事实上，小孩身高与时间存在相关关系，而小树高度与时间也存在相关关系。因此，小孩身高与小树高度存在的直线相关关系只是一种表象，是他们与时间存在某种相关关系的一种伴随关系。

3.3 相关分析结果只有相关系数结果而无相关系数的假设检验结果　(1)实例：某研究者要分析咪唑安定在 10～90 μg/mL 范围内浓度与色谱图的峰面积之间的关系。结果为经相关分析及线性回归可得：回归方程为 $y=144\ 849x-811\ 86$(x 为浓度，y 为峰面积)，相关系数 $r=0.999\ 9$。结论为二者的线性关系良好[2-3]。(2)错误解析：该研究对咪唑安定在 10～90 μg/mL 范围内浓度与色谱图的峰面积之间的关系进行了相关分析所得相关系数 $r=0.999\ 9$ 值非常接近 1，那么是否就可以直接下二者之间具有很强的正相关关系呢？这样是不可以的。因为不同样本空间大小对应一个临界相关系数值，若统计值高于它，就代表相关关系显著，否则为不显著。例如：若有 30 组数据，临界相关系数为 0.361，0.6 的相关系数就代表相关关系显著；若只有 3 组数据，临界相关关系则为 0.997，0.99 的相关系数就代表相关关系不显著。因此，统计相关系数时必须与临界相关系数对比之后即进行相关系数的假设检验以后才有统计学意义。

3.4 对相关系数假设检验结果认识不清　(1)实例：某研究者要研究 EGFR、HER-2、TopoⅡα 在胃癌组织中的表达与临床病例特征之间的关系。结果是 EGFR 与 HER-2 之间($r=0.317$，$P=0.001$)，EGFR 与 TopoⅡα 之间($r=0.257$，$P=0.007$)，HER-2 与 TopoⅡα 之间($r=0.270$，$P=0.005$)的表达均呈正相关[4]。(2)错误解析：该研究所得 EGFR、HER-2、TopoⅡα 之间的相关系数 r 最大也只有 0.317，属于非常弱的直线相关关系。如果对其进行直线回归，决定系数 R^2 只有 0.1 左右，也就是说在总的离均差平方和中因变量的总体变异能被自变量解释的部分仅占 10％左右，决定系数太小，也就是说此时进行回归分析及直线回归意义不大[5]。那么，当 $P<0.05$，相关系数 r 大到何种程度才有意义呢？目前，仍没有一种确定的方法。作者认为是否有意义，主要与样本量和实际问题有关。例如在心理学调查问题中，往往是很大的样本，这时若 r 较小，也可以认为两个变量之间有某种正或负的总体相关趋势。但是，如果样本量较小时，其相关性的意义不大。

3.5 误认为“无直线相关”就是“无相关”　(1)实例：某研究者研究肿瘤坏死因子-α 在溃疡性结肠炎中的作用。其结果中有这样的叙述：结果显示 UC 结肠黏膜肿瘤坏死因子-α mRNA 的表达与疾病活动性评分无相关性($r=-0.10$，$P=0.55$)[6]。(2)错误解析：进行相关性的讨论，得先根据医学常识判断是否有进行相关分析的意义，再画出反映两个变量同时变化的散点图，当散点图显示的结果表明值得进行直线相关分析时，才能进行适当的统计计算和假设检验，该研究未给出也未提及散点图，而仅根据相关系数 r 值作出无相关性的判断是不妥当的，忽略了所研究的两个定量指标之间呈曲线相关的可能性[7]。

参考文献：

[1] 高辉，胡良平，李长平，等. 如何正确进行直线相关与回归分析[J]. 中西医结合学报，2008，6(12)：1311-1314.

[2] 胡良平，刘惠刚. 相关与回归分析错误辨析(3)[J]. 函授继续医学教育，2006，11(1)：25-32.

[3] 田睿，甘勇军，徐颖. 咪唑安定微乳滴鼻液的制备及质量控制[J]. 第三军医大学学报，2011，33(1)：82-85.

[4] 梁秀菊，高明，周云. 胃癌组织 EGFR、HER-2 与 TopoⅡ的表达及临床意义[J]. 第三军医大学学报，2010，32(22)：2443-2445.

[5] 刘显初，郑利平，黄胜贤，等. 冠心病血清总胆红素与氧化修饰低密度脂蛋白的关系研究[J]. 中华检验医学杂志，2003，26(8)：479-480.

[6] 严瑾，欧阳钦，刘卫平，等. 肿瘤坏死因子-α 在溃疡性结肠炎中的表达及其作用探讨[J]. 胃肠病学，2005，10(5)：269-272.

[7] 李建志，孙建东，郭秀芳，等. 山东省寿光市沿海地区低氟水分布调查[J]. 地方病通报，2001，16(1)：21-22.

(收稿日期：2011-06-06　修回日期：2011-07-18)