相关分析的误用表现与解决方案

魏　瑜，马开平
（南京农业大学 工学院，南京 210031）

摘　要：每种统计方法都有其特定的应用场合，如果误用，不仅得不出正确的结论，甚至会出现误导，给科学研究和生产实践带来负面的影响。文章以统计中经常采用的相关分析为例，研究了相关分析的常见的误用情况，并指出采用主题串讲并给出索引和反例教学法来解决相关分析的误用问题。

关键词：相关分析；应用场合；知识索引；反例教学法
中图分类号：C829.29　　**文献标识码**：A　　**文章编号**：1002-6487(2015)02-0086-03

大部分统计学相关教材[1~3]，包括概率论与数理统计教材[4]，在介绍相关分析时，只介绍了Pearson积矩法线性相关系数，而且对Pearson相关系数的应用场合强调得不够，只有少部分统计学教材给出了相对完整的介绍。而很多初学者，甚至一些运用统计作为工具的非统计专业的研究者，在进行相关分析时，也只计算并检验Pearson相关系数，并据此做出是否相关的推断，甚至做出因果推断，这是对相关分析的误用，这种误用会导致错误的分析结论，从而导致错误的决策。

1　相关分析的种类

变量间的关系可分为函数关系和相关关系，其中相关关系是一种非确定性的关系，这种关系在日常生活和科学研究中非常常见。例如，生物学研究的父亲身高(Y)与子女身高(X)之间的关系；农学研究的粮食亩产量(Y)与施肥量(X_1)、降雨量(X_2)、温度(X_3)之间的关系；体育科学研究的中年人三个生理指标：体重(Y_1)、腰围(Y_2)、脉搏(Y_3)和三个训练指标拉单杠次数(X_1)、仰卧起坐次数(X_2)、跳高(X_3)间的关系，这些变量间显然不是独立的，而又没有确切到可由其中的一个或几个变量去精确地决定另一个或几个变量的程度，这就是相关关系。研究相关关系的分析方法即相关分析，即研究现象之间是否存在某种依存关系，并对具体有依存关系的现象探讨其相关方向以及相关程度的分析。

从不同的角度来看，变量间的相关关系有着不同的分类，可以按变量的计量尺度、变量的个数、变量间的依存形式等角度对相关分析进行分类，这些分类间存在着交叉，每类相关关系的分析方法各有所不同。在实际应用时，要根据具体情况选择不同的相关分析方法。

1.1　按分析变量的计量尺度来分

变量有定类、定序、定距和定比四个不同层次的计量尺度。不同尺度的变量其相关分析的度量指标不同。

表1　　分尺度的相关系数计算方法

变量类型	相关系数
两个定类变量	Lambda，tau-y，Φ，列联，Cramer V
两个定序变量	Gamma，dy，Kendall's τ，Spearman's ρ
两个定距变量	简单线性回归，Pearson
定类变量与定距变量	相关比率与非线性相关
定类变量与定序变量	Lambda，tau-y
定序变量与定距变量	相关比率

姚宝玺，李育安，曹维芳[5]总结了两变量相关关系的度量见表1。

(1) 定类尺度：也称类别尺度或名义尺度，是将调查对象分类，标以各种名称，并确定其类别的方法。类与类之间是平等的。最常见的例子就是性别，分男、女。

(2) 定序尺度：也称等级尺度或顺序尺度，是按照某种逻辑顺序将调查对象排列出高低或大小，确定其等级及次序的一种尺度。如学历，有小学、中学、大学等之分。一般讲来，它们之间是有一个先后顺序的：读完小学、再进中学、其后上大学。

(3) 定距尺度：也称等距尺度或区间尺度，是一种不仅能将变量(社会现象)区分类别和等级，而且可以确定变量之间的数量差别和间隔距离的方法。如智商，甲有智商为130，乙的是65，相差65，但不能说甲比乙聪明一倍。

(4) 定比尺度：也称比例尺度或等比尺度，除了有上述三种尺度的全部性质之外，还可测量不同变量(社会现象)之间的比例或比率关系。

高层次变量可以通过相应的转换转化为低层次变量，适合于低层次变量的分析方法也适合于高层次的变量，但适合于高层次变量的分析方法并不适合于低层次的变量。

因为在社会学研究中，经常会遇到定类和定序数据，李沛良[6]在《社会研究的统计应用》一书中详细介绍了数据的不同尺度及相应的分析方法。

1.2　按需分析变量的多少来分

按需分析变量的多少可分为简单相关和多元相关，其中多元相关又包括复相关、偏相关和典型相关。

基金项目：国家自然科学基金青年科学基金项目(71101072)；南京农业大学工学院教育教改项目
作者简介：魏　瑜(1974-)，女，安徽巢湖人，硕士，讲师，研究方向：数据统计分析与管理。

简单相关分析即分析两个变量间的相关关系，根据不同的变量类型、依存关系采用不同的相关系数。

复相关分析通常采用复相关系数来反映一个变量y与其它多个变量之间线性相关程度。应用时多用多元线性回归方程的可决系数的根号直接求得，其值介于0到1之间，无正负之分。

偏相关分析通常采用偏相关系数来度量在对其他变量的影响进行控制的条件下，多个变量中某两个变量之间的线性相关程度。

当然对于非线性关系，也有相应的复相关和偏相关分析，只是比较复杂。

典型相关分析是研究两组变量之间相关关系的多元分析方法。它借用主成分分析降维的思想，分别对两组变量提取主成分，且使从两组变量提取的主成分之间的相关程度达到最大，而从同一组内部提取的各主成分之间互不相关，用从两组之间分别提取的主成分的相关性来描述两组变量整体的线性相关关系。

黄良文[7]在《统计学原理》一书中有关于简单相关、复相关和偏向关的详细介绍；而典型相关则需参考多元统计分析方面的书[8]。

变量间还可能有更复杂的结构关系，有些变量可能通过中介变量与另外的变量发生关系，有些变量可能起着调节的作用。这些变量间的复杂结构关系应该用结构方程模型来分析，这些知识则需要参考结构方程模型方面的书[9]。

1.3 按变量间按依存形式来分

根据变量间按依存形式可分为线性相关和非线性相关。

线性相关根据具体情况可以选择Pearson积矩相关系数（两变量）、复相关系数、偏相关系数和典型相关分析。

非线性相关分成两类，一类可以用某种非线性模型去模拟，则可以用相关指数来度量其相关程度。相关指数即判断变量之间是否显著存在某种类型的非线性相关关系的尺度，是对非线性回归模型进行拟合时所得到的可决系数。另一类是不可以用某种非线性模型去模拟变量间的关系，但也许只表现为秩相关，即等级相关，此时，即使两个变量都是数值型变量（定距、定比变量），也可以用Spearman秩相关系数和Kendall秩相关系数来度量它们的等级相关程度。具体内容可参见黄良文编写的《统计学原理》[7]。

2 相关分析的常见误用表现

2.1 不区分数据类型而都采用数值相关

虽然关于定性、定类数据的分析方法在李沛良等编写的社会研究方法方面的书[6]中的介绍比较详细，但一般的《应用统计学》或《统计学原理》书[8]中，尤其是管理科学与工程类的统计学教材[9]中关于定性、定类数据的分析的内容很少涉及，但管理学科中尤其是人力资源管理中也会碰到定性、定类数据的分析。而这些学生由于知识结构的缺陷，容易不管何种数据类型而都采用数值相关分析，这样得出的结论是不准确的，甚至是错误的。

2.2 将相关关系和因果关系视为等同

相关关系只表示变量间具有某种共变关系，并不表示变量间的因果。变量间具有相关关系是变量间具有因果关系的必要但不充分条件，即

相关关系 ⇐ 因果关系

边玉芳[10]在《警惕心理学研究中的统计误用》也指出仅仅根据一个相关系数，无法确定事物之间的因果关系，单一的相关证据并不能做出有效的因果陈述。相关分析一般只用于分析两个变量间的关联程度，要说明蕴含在相关背后的、对这种相关加以解释的本质则要借助于理论。或进一步对一些变量进行控制后作深入的研究，也可以对相关研究进行改进，如作交叉-滞后-组相关程序的研究（一种追踪研究，可以得出因素间的交叉-滞后相关），经过多重检验来提高相关研究的解释力。

2.3 用简单相关分析复杂现象

如果研究者只用简单相关去分析一个复杂的现象，那么结果可能犹如“盲人摸象”，看到的只是一个现象的局部，而得出错误的结论。

为了形象地说明，借用黄良文[7]书中的一个简单的反例，收集某地各年的A产品的消费量，居民人均收入和价格资料，见表2。

表2　　某城市有关A产品的需求统计

年次	1	2	3	4	5	6	7	8	9	10
销售量X_1（百件）	10	10	15	13	14	20	18	24	19	23
单价X_2（百元）	2	3	2	5	4	3	4	3	5	4
居民人均收入X_3（10元）	5	7	8	9	9	10	10	12	13	15

画出销售量X_1与单价X_2散点图见图1，可以看出两者之间有着很弱的正线性相关关系。

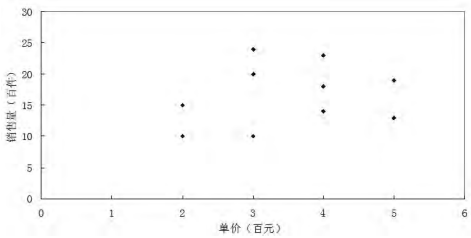

图1 某城市有关A产品销售量与单价散点图

计算简单线性Pearson相关系数r_{12}=0.227，Sig.(2-tailed)=0.529；这似乎与经济学的需求曲线相悖。而计算控制居民人均收入X_3不变的销售量X_1与单价X_2的偏相关系数$r_{12.3}$=-0.680，Sig.(2-tailed)=0.044，在0.05的显著性水平上通过偏相关系数不为零的统计检验，且偏相关系数为负值，说明如果居民人均收入不变，该商品的销售量会随着单价的上升而呈显著的负线性相关关系。如果只分析销售量X_1与单价X_2的简单相关则会得出错误的结论。

2.4 用线性相关分析非线性相关问题

同样举两个简单的反例。第一个是用简单线性Pearson积矩相关系数度量一个曲线相关的反例，设两变量数据见表3（为使得例题简单，效果明显，数据是假设的），计算其简单线性Pearson积矩相关系数r=0.0138，几乎无线性相关关系。画出X1与X2的散点图见图2，可见X1与X2

呈抛物线曲线相关，用抛物线去模拟得出其相关指数R^2=0.9970，相关度很高。

表3　　曲线相关例题数据

X1	9	5	2	1	2	5	9.2
X2	0	1	2	3	4	5	6

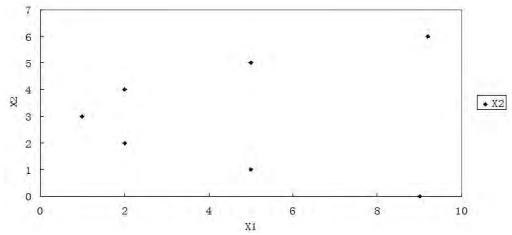

图2 曲线相关例题X1与X2的散点图

第二个是用简单线性Pearson积矩相关系数度量一个等级相关的反例，例题同样引自黄良文[7]一书。统计了几个学生的复习时间和考试成绩见表4。计算得Pearson相关系数r= 0.587, Sig.(2-tailed)=0.075。结论好像考试成绩与复习时间呈弱的线性相关。

表4　　复习时间与考试成绩数据表

复习时间X1(天)	3	4	1	2	5	8	10	9	11	13
考试成绩X2	86	87	4	85	93	91	95	94	95	96

但画出两者的散点图见图3。仔细观察和分析可知，虽然考试成绩和复习时间的线性相关关系很弱，也不呈现某种曲线相关，但是其单调性却相当明显，计算Spearman秩相关系数r_d= 0.985，Sig.(2-tailed)=0.000，即考试成绩和复习时间呈高度等级相关关系。

可见，利用线性相关分析工具去分析非线性相关问题也会得出错误的结论。

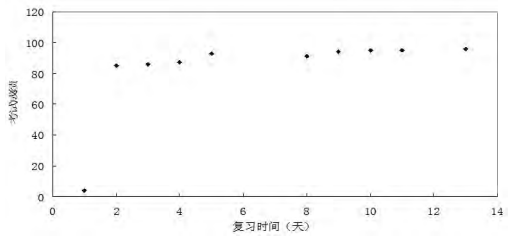

图3 复习时间与考试成绩散点图

3　解决方案

3.1　主题串讲并给出索引

综上所述，相关分析的各种方法分散在不同的教材和文章中，这使得学生尤其是非统计专业的学生或以统计为研究工具的其他专业研究者对相关分析的知识支离破碎，没能形成一个整体，从而容易导致误用。

也许是考虑到学时和学生的接受能力有限，大部分的教材只侧重介绍了相关分析的某一方面。当然教材编写者不可能、也没必要将所有的相关分析知识都囊括在一本统计学书中，教材编写者只需要给学生一个知识的全貌，并强调各种相关分析方法的应用场合，同时标出知识的出处和进一步学习的参考书目，给出知识的索引即可。这并不需要太多的篇幅和数学基础。这样读者就能够知道什么样的问题该用什么样的分析方法，即使不知道具体怎么做，也能够知道在哪里可以查到答案。虽然吴喜之[11]的《统计学：从数据到结论》一书由浅入深地把统计最基本和最有用的部分用相对简单的语言做了较全面的介绍，只可惜没有指出每个知识点进一步学习的相应参考书，即没给出索引，而且对于相关分析的知识介绍也不够全面，且分散在三章中，这不便于读者进一步地学习研究。所以在统计大师没有编出更合适的教材之前，需要授课教师来进行主题知识串讲，并给出知识的出处和进一步学习的参考文献，给写生一个知识的全貌和索引。

3.2　反例教学法

案例教学是一种实践教学活动，可实现统计理论与实践的有效结合，是培养学生应用能力的一种有效方法，而反例教学则可以给学生留下更深刻的印象。

马晓绒，姚红梅[12]认为可以借助反例引导学生辨析、纠正错误，掌握基本知识和基本概念。为了有效地帮助学生认清是非，巩固知识，提高纠错能力，恰当的选择反例，能使学生对知识得到进一步的扩大和深化，有助于增进学生对知识掌握的深度和广度，并能使学生养成严格推理、全面分析问题的习惯，提高防错能力。所以举几个容易用错的反例可以让学生充分地、直观地认识到相关分析方法误用带来的后果，使其避免发生类似的错误。

4　结论

考虑到有不少读者的数学基础比较薄弱，本文尽量避开数学表达式，用通俗易懂的语言给出了相关分析的全貌和索引，强调了各种分析方法的应用场合，研究了相关分析的常见误用之处。并指出可以采用主题串讲并给出索引和反例教学法来避免误用。本文所用的正是这两种方法，这两种方法也适用于其他统计知识的学习。

参考文献：

[1]黄良文.统计学 [M].北京:中国统计出版社,2009.

[2]贾俊平,何晓群,金勇进.统计学(第五版) [M].北京:中国人民大学出版社,2012.

[3]刘思峰,吴和成,菅利荣.应用统计学 [M].北京:高等教育出版社,2009.

[4]盛骤,谢式千,潘承毅.概率论与数理统计[M].北京:高等教育出版社,2010.

[5]姚宝玺,李育安,曹维芳.两变量相关关系的度量[J]. 统计与决策,2007,(1).

[6]李沛良.社会研究的统计应用[M].北京:社会科学文献出版社,2002.

[7]黄良文.统计学原理[M].北京:中国统计出版社,2000.

[8]约翰逊,威克恩. 实用多元统计分析[M].北京:清华大学出版社,2008.

[9]邱皓政,林碧芳.结构方程模型[M].北京:中国轻工业出版社,2009.

[10]边玉芳. 警惕心理学研究中的统计误用[J]. 心理科学进展,2002,(4).

[11]吴喜之.统计学:从数据到结论 [M].北京:中国统计出版社,2008.

[12]马晓绒,姚红梅. 反例方法的作用与施教时机[J]. 西安联合大学学报,2001,(4).

(责任编辑/亦　民)