中国地质大学
CHINA UNIVERSITY OF GEOSCIENCES
武汉 · WUHAN

多元统计方法:原理及应用

信息与计算科学系 ***

主要内容

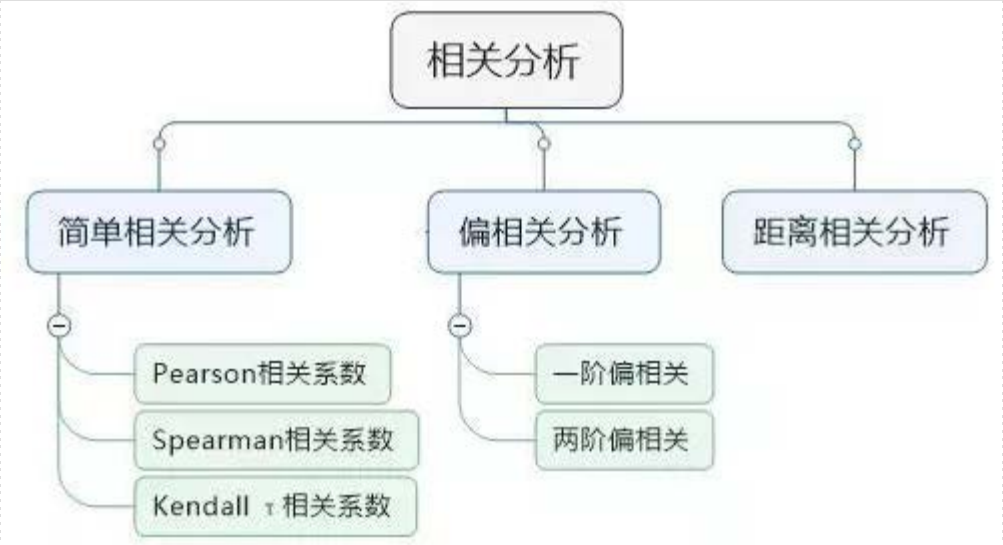
1. 相关分析
2. 回归分析
3. 聚类分析
4. 判别分析
5. 主成分分析
6. 因子分析
7. 对应分析
8. 典型相关分析

1 相关分析概述 correlation analysis

一. 概念与分类

相关分析是研究两个或两个以上随机变量间的相关关系的统计分析方法。例如，人的身高和体重之间；空气中的相对湿度与降雨量之间的相关关系都是相关分析研究的问题。

相关分析，常用的方法类别有：简单相关分析、偏相关分析、复相关分析、距离相关分析等。其中前两种方法比较常见。

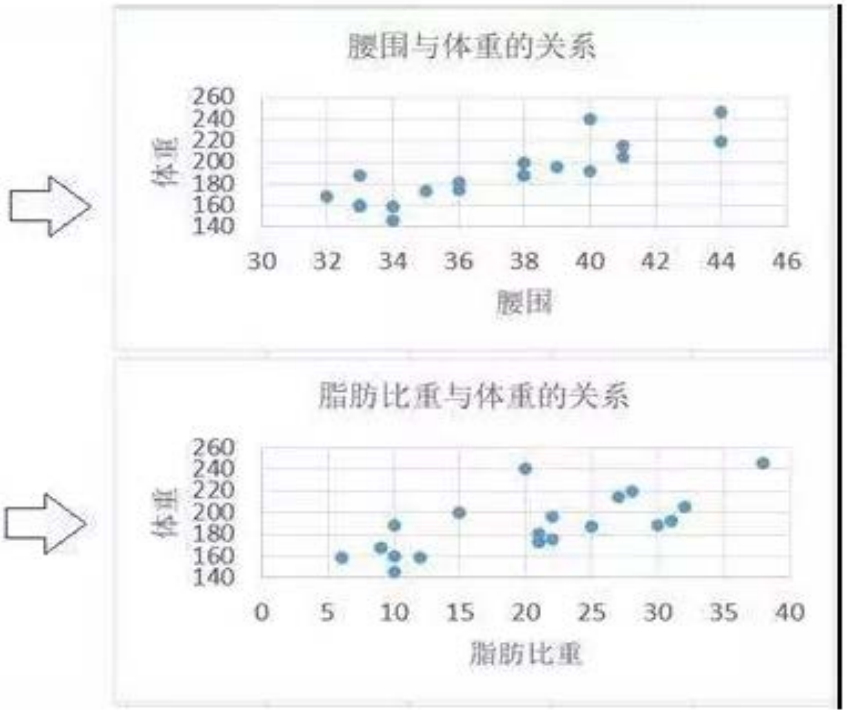


1 相关分析概述 correlation analysis

1. 简单相关分析

简单相关分析研究两个随机变量间的相关关系。可以用散点图直观反映，同时可用相关系数体现两者之间的相关程度。

腰围	体重	脂肪比重
32	1	6
36	181	21
38	200	15
33	159	6
39	196	22
40	192	31
41	205	32
35	173	21
38	187	25
38	188	30
33	188	10
40	240	20
36	175	22
32	168	9
44	246	38
33	160	10
41	215	27
34	159	12
34	146	10
44	219	28



散点图显示，随着腰围的增加，体重也在增加。说明腰围和体重是存在相关关系的，而且应该是正相关。同样，脂肪比重与体重也是正相关的。

1 相关分析概述 correlation analysis

1. 简单相关关系

可视化的优点是：直观，其缺点是：无法准确度量。比如腰围和脂肪比重，对体重的影响程度到底有多大？或者说，这两个因素中哪个因素对体重的影响会更大？散点图无法给出答案。

相关系数（Correlation Coefficient），是专门用来衡量两个变量之间的线性相关程度的指标，常用的相关系数有皮尔逊（Pearson）相关系数、Spearman相关系数和肯德尔（kendall）相关系数。需要的别注意的是：这三种相关系数适用于不同情形的数据类型，对数据分布有不同要求！

1 相关分析概述 correlation analysis

1. 简单相关关系

Pearson相关适用于正态分布定距数据；Spearman等级相关系数适用于不明分布定距数据，只要求两个变量是成对的等级评定数据；Kendall相关系数适用于多列不明分布定序数据，又称为和谐系数或者一致性系数，例如让K个评委对N个对象进行排序性评价，常用于信度分析。如果采用不合适系数计算变量之间的相关性，将不能客观准确评价实际情况。

三种相关系数的计算公式

$$R=\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2\sum(Y_i-\bar{Y})^2}}$$

$$R=1-\frac{6\sum d_i^2}{N(N^2-1)}$$

$$r=\frac{N_c-N_d}{n(n-1)/2}=\frac{2(N_c-N_d)}{n(n-1)}$$

1 相关分析概述 correlation analysis

Spearman等级相关系数

从等级的角度研究变量之间的关联程度。

1、*Spearmam* 等级相关系数 r_s

设样本为n对配对样本,即:$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$.

将变量x与y的观察值按一定次序（从高到低）排列，依次给以等级值 $1,2,\cdots,n$ ， 即: $(x_i,y_i)\rightarrow(x_i',y_i')$.

该系数是以变量没有相同等级为前提的。若观察值相等，则它们的等级值取它们所对应的等级值的平均值。

例子:

交卷名次: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

考试成绩: 90, 74, 74, 60, 68, 86, 92, 60, 78, 74, 78, 64

1 相关分析概述 correlation analysis

相关分析基本步骤

简单相关分析的基本步骤如下：

- 基本步骤：
  - 1、画散点图，观察两个变量是否有规律变化。
  - 2、根据变量类型或正态性检验，选择合适的相关系数公式。
  - 3、计算相关系数r，评估相关程度。
  - 4、进行显著性检验，如果P<α（一般取0.05），表示存在显著相关性。
  - 5、给出结论。

    总结分析结论，并从业务层面给出业务判断，以及业务策略。

1 相关分析概述 correlation analysis

相关分析是否一定靠谱？统计也会撒谎？

例一：想长寿吗？来吸烟吧！（……）

这个例子是一个基于合理数据的严肃研究。英国某健康研究机构随机抽取出了1314名志愿者，其中582名吸烟者，732名不吸烟者。20年后，跟踪调查显示，吸烟者的死亡率24%，而不吸烟者死亡率为31%，并给出了这样一份统计报告（部分）：

	Survival Status		
Smoker	Dead	Alive	Total
Yes	139 (24%)	443 (76%)	582
No	230 (31%)	502 (69%)	732
Total	**369**	**945**	**1,314**

1 相关分析概述 correlation analysis

相关分析是否一定靠谱？统计也会撒谎？

例一：想长寿吗？来吸烟吧！（……）

如果只看表中数据，人们会感到诧异：难道吸烟能增寿？数据没有造假，一定出了什么问题！人们很快发现分析中漏掉了年龄因素。实际上，随机抽取的1314个人中，吸烟者中65岁以上只占8.4%，而不吸烟者中65岁以上者占比26.4%.发现了这一情况，该如何处理？分年龄组！然后看看每个年龄组的吸烟者与不吸烟者的死亡率之间的关系。

	Age Group			
Smoker	18–34	35–54	55–64	65+
Yes	2.8%	17.2%	44.3%	85.7%
No	2.7%	9.5%	33.1%	85.5%

1 相关分析概述 correlation analysis

相关分析是否一定靠谱？统计也会撒谎？

例一：想长寿吗？来吸烟吧！（……）

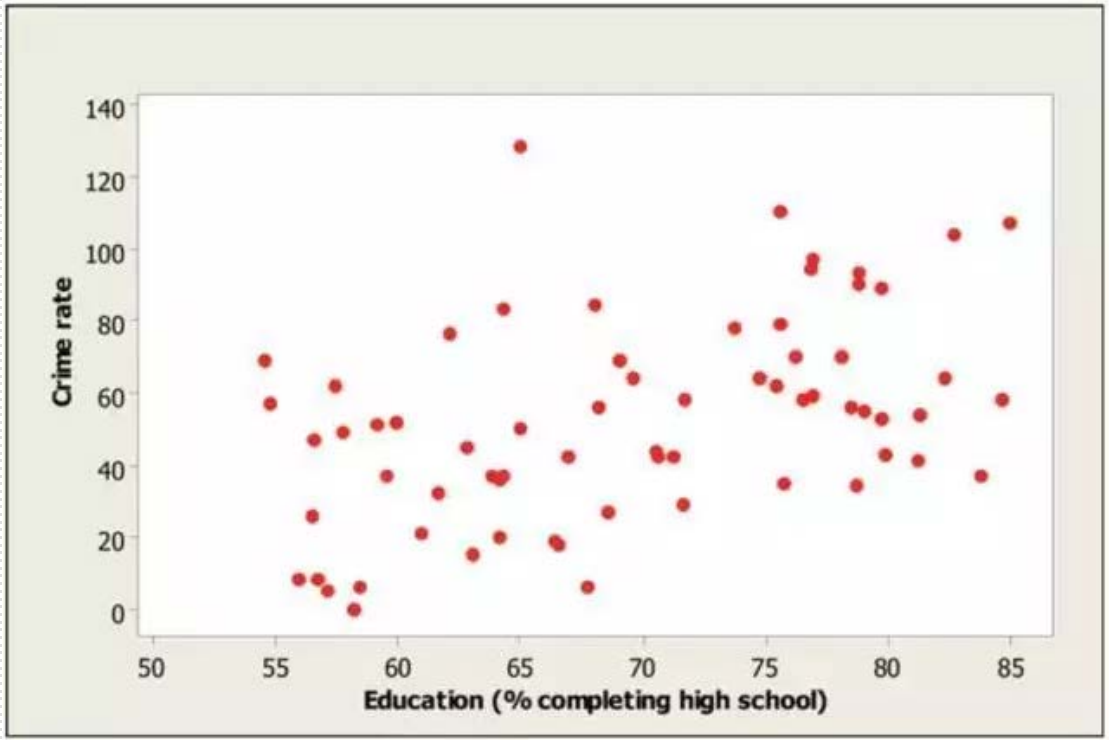
这个例子告诉我们“年龄”这种会对结果产生重大影响但是却没有被考虑的变量，会对本问题的相关分析带来很大干扰。我们把它叫做“潜在变量”（**lurking variable**）。它有时候真可谓是“杀人于无形之中”，稍不注意就可能会要了整个统计分析报告的命。这种结果直接被潜在变量给反转的现象，我们在统计学里面称之为辛普森悖论（**Simpson’s Paradox**）。

1 相关分析概述 correlation analysis

相关分析是否一定靠谱？统计也会撒谎？

例二：高学历者更容易犯罪？

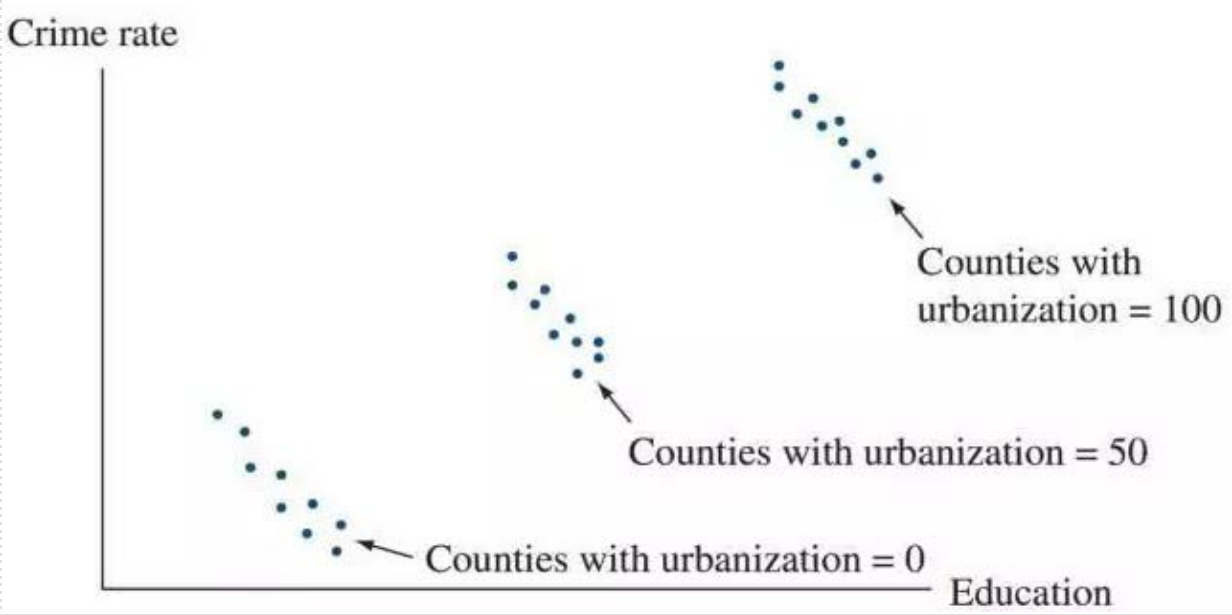
曾经有人调查美国各地居民的学历与犯罪率数据，数据的散点图如下：



散点图是不是体现了这样的倾向：受教育程度越高的地方，犯罪率也越高？计算出二者的相关系数为0. 47.

1 相关分析概述 correlation analysis

有了例一的经验，我们肯定能确定，这里也存在潜在变量！具体开展研究时，需要依据背景理论确定几个候选的潜在变量。这里，学者确定了一个很可能是重要的潜在变量：城市化水平！于是他们把调查的地区分为三个等级城市化的分类样本，得到如下的散点图



1 相关分析概述 correlation analysis

通过分类之后的数据计算，我们可以发现：在类似的城市化程度上，教育水平是跟犯罪率是负相关的！而随着城市化程度越来越高，其实犯罪率和教育水平都会升高，这一点是跟我们的常识相符的。美国的大城市犯罪率就是要高一些，而大城市的居民通常更有可能接受到高中以上的教育。所以之前我们看到的教育水平和犯罪率的正相关性其实并没有太大的参考价值。

本部分内容主要参考了微信公众号“量化研究方法”中相关分析资料。

1 相关分析概述 correlation analysis

偏相关

研究在多变量的情况下，当控制其他变量影响后，两个变量间的程度, 又称净相关或部分相关。

复相关

研究一个变量与另一组变量之间的相关程度。例如, 职业声望同时受到一系列因素（收入、文化、权力……）的影响，那么这一系列因素的总和与职业声望之间的关系，就是复相关。

距离分析

用于测度变量之间相似性，可以采用距离或者相似系数等统计量来衡量变量或者观测量之间的相近（相似）程度。

2 回归分析概述 Regression analysis

一. 基本原理

回归分析是研究两个或两个以上变量间的影响关系的重要统计分析方法。具体而言，它能研究变量间影响模式、影响方向和影响程度。当存在多个解释变量时，有时还要考虑解释变量之间是否存在交互作用。

二、基本假定（多元线性回归分析）

解释变量之间不相关（否则意味着多重共线性）；

回归模型形式是线性的（包含可以变换成线性的非线性情形）

随机误差项有多条假定，简单说：独立同分布，服从均值为零，方差为σ平方的正态分布。

实际应用中需要检验数据是否满足这些假定！（回归诊断）

2 回归分析概述 Regression analysis

三. 基本步骤

1. 分析实际问题，确定候选解释变量和被解释变量；
2. 设定模型形式（利用背景理论和统计数据）；
3. 收集数据，必要时做数据预处理；
4. 估计参数（OLS、RRS、PLS、Lasso Regression）；
5. 假设检验（方程和系数的显著性检验）；
6. 回归诊断（模型形式、共线性问题、异方差问题、正态性问题、独立性问题，是否有异常数据等）；
7. 模型改进；
8. 应用（解释现象、预测及生产控制）。

2 回归分析概述 Regression analysis

四. 应该注意的几个问题

应用回归分析方法研究实际问题时，要注意以下几个问题

1. 模型形式与变量选择既要有理论依据，又符合统计学原则；
2. 回归建模过程要完整，每一环节尽量精致化处理；
3. 要对计算结果做出明确的解释，并且要判断是否与背景理论或者常理相一致，如果出现了不一致的情况，不要轻易下结论，先要检查建模过程是否存在问题；
4. 要有数据质量及样本意识，即注意检查数据的可靠性、是否存在缺失或错误数据，样本数据是否具有代表性，与研究目的是否相符合，样本量是否足够大；
5. 注意相关方法的适用性（参数估计及假设检验）；
6. 不要混淆拟合与预测。

2 回归分析概述 Regression analysis

五. 关于检验问题

回归模型检验包括实际意义检验和统计检验两个方面。实际意义检验：主要涉及参数估计值的符号和取值范围，如果模型结果与现实理论以及人们的实践经验不符，说明回归模型不能很好地解释该现象。尤其是对社会经济现象进行回归分析时，常会遇到实际意义检验不能通过的情况，主要是因为社会经济现象的统计数据无法象自然科学中的数据那样通过有效控制的实验得到，这样得到的数据可能不满足线性回归分析的要求.

统计检验：是利用统计学的检验理论检验回归模型的可靠性，具体又分为拟合优度检验、模型的显著性检验（F 检验）和模型参数的显著性检验（t 检验）。

3 聚类分析概述 Cluster analysis

一. 基本概念

聚类分析指将物理或抽象对象的集合分组为由类似的对象组成的多个类的分析过程。分为R型聚类和Q型聚类。

聚类分析的基于相似性作分类。聚类源于很多领域，包括数学，计算机科学，统计学，生物学和经济学。在不同的应用领域，很多聚类技术都得到了发展，这些技术方法被用作描述数据，衡量不同数据源间的相似性，以及把数据源分类到不同的簇中。

按照机器学习的观点，聚类分析属于无指导（或者无监督的）的机器学习类型。它是一种探索性的统计方法，即学习过程不依赖于事先标记类别的训练样本。

3 聚类分析概述 Cluster analysis

二. 基本方法

传统意义上常用的聚类方法有系统聚类法、K-均值法、有序样品最优分割法和模糊聚类法等。

从数据挖掘角度看，聚类方法可分为①基于代表的聚类，主要是K-均值及其拓展方法，还有以EM算法为基础的期望最大聚类方法；②层次式聚类方法，包括聚合型与分化型两种。聚合型自下而上，分化型自上而下；③基于密度的聚类，这种方法可以使用于非凸簇，第①类方法适用于凸簇；④谱聚类和图聚类。

3 聚类分析概述 Cluster analysis

三. 聚类结果的验证

对于聚类分析结果进行验证和评估主要包含三个方面的任务①聚类评价，用于评估聚类的质量和优度；②聚类稳定性，主要考虑聚类结果对于算法、参数的敏感性；③聚类的趋向性，用于评估应用聚类分析方法是否合适，即数据本身是否有固有的分组结构。

根据以上任务，分别提出了一些有效性度量和统计量，又分为三个类别：

外部验证（采用与数据无关的标准）、内部验证（从数据自身导出标准）和相对验证（比较不同的聚类）。

3 聚类分析概述 Cluster analysis

四. 聚类分析中注意的问题

聚类分析中需要考虑一下几个重要问题：

① 属性或者特征变量选择。尽量选择具有重要分类指示意义的变量，变量之间相关性不要太强；

② 分类数目。一般原则是：在将不同对象分开的前提下，尽量分为较少的类别数目。可以参考专业知识以及统计方法确定类数。

③ 聚类统计量。R型聚类一般采用相似或者相关系数，Q型聚类采用距离度量，必要时先进行数据预处理（如标准化）。

④ 验证分类结果的有效性。将样本随机分为两组，一组用于聚类实验，另一组用来验证。

4 判别分析概述 Discriminant analysis

一. 基本概念

判别分析的作用也是对研究对象进行分类（classify），但它与聚类不同，是一种有指导（有监督）的学习方法，需要带标签的训练样本数据进行学习。

传统意义上的多元统计分析中，常用的判别方法有距离判别法、费歇判别法、贝叶斯判别法和逐步判别法，以及逻辑斯蒂回归方法。现代意义上的分类方法众多，分别是基于概率的分类方法、决策树分类法、费歇判别法的拓展方法—核判别法、支持向量机和神经网络方法等等。

分类工作中，分类性能度量与分类器（classifier）的效果评估方法也很重要。分类器是属性变量的组合（函数）。

4 判别分析概述 Discriminant analysis

二. 方法要求的条件或适用性

距离判别法与费歇判别法对数据分布没有要求，贝叶斯判别法一般要求数据服从（多元）正态分布。

基于概率的分类方法一般要求属性变量相互独立，但现实中这个假定往往不能满足。朴素贝叶斯分类方法往往能取得良好效果。

决策树分类法可以对数值型数据以及类别型数据进行分类，但是数据必须是离散型而不是连续型的。

支持向量机是一种二分类方法，根据数据数据特点可采用线性支持向量机和非线性支持向量机。

4 判别分析概述 Discriminant analysis

三. 分类的评估

分类性能度量：误判率指标或正确率指标、ROC分析（Receiver Operation Characteristic）。

分类器评估：

① K折交叉验证：将数据集等分为K个子集，将其中每个子集作为测试集，其余子集作为训练集。

② 自助抽样（Bootstrap Resampling）：每次从数据集中抽出K个样本规模相同的子样；

③ 置信区间：对某个选定的性能指标的均值和方差进行估计，确定其置信区间；

④ 配对t检验：对不同分类器做显著性差异检验。

4 判别分析概述 Discriminant analysis

四. 影响判别（分类）效果的重要因素

（1）训练样本的数量和质量：传统判别分析应用时样本量一般不大，每组训练样本最好多余20个。数据可靠。大数据时代，可能遇到海量数据或者高维数据，则要考虑数据化简、其他数据预处理（如清洗数据、补缺）工作。另外，还要注意不平衡数据问题。

（2）分类方法：无论是传统还是现代分类方法，都有其优劣之处，尽量选择适合实际问题、数据特点的分类方法，或者采用不同方法进行对比分析。

（3）属性变量选择：尽量选择那些具有分类能力的特征变量，删除不重要的特征变量。

（4）组间差异问题：需要存在显著的组间统计差异。

5 主成分分析概述 Principal Components Analysis

一. 基本原理和应用

主成分分析是一种数据化简方法，它通过将原始变量作适当线性组合（数据挖掘有非线性组合，即核主成分分析）构建综合变量来代替原始变量进行统计分析。面临高维数据时，主成分是很常用的数据化简方法。

主成分分析的应用：实际应用包括综合评价、主成分回归、主成分聚类等。一般不说主成分模型，说主成分方法，主要原因是它的作用是构建新变量做统计分析。

主成分降维的主要目标有三个：减少变量个数；构建不相关综合变量（正交化）；提供一个框架用于解释结果。

5 主成分分析概述 Principal Components Analysis

二. 主成分应用时的几个注意问题

1. 综合评价的应用：在国内，很多领域的研究者应用主成分方法进行综合评价，但关于主成分的这种应用也备受争议，争议的主要问题有：①应用条件问题，几乎所有的多元统计分析教科书都没有谈到这种方法需要具备怎样的条件，这很容易导致众多应用工作不能令人信服。一般要考虑几个条件：变量之间应该有一定相关性！样本量与变量个数之比不能太小；求出的主成分能否合理解释其现实意义。

②排序性评价问题，有学者指出，当主成分存在无序变量时，主成分会带来错误的综合评价结果。实际应用中

5 主成分分析概述 Principal Components Analysis

二. 主成分应用是考虑的几个问题

综合评价应用（续）往往只有第一主成分的符号一致性比较好，所以有人认为只用第一主成分做综合评价，但这样一来，又被质疑贡献率不够，这个问题仍然没有得到圆满解决，不过有些做法可供参考：对于负向（第一）主成分，可以取负号让其变成正向综合指标，而其他主成分会存在较多不同符号的变量，则没有统一方法处理，只能相机行事。

③实际意义的解释：如果不能解释清楚主成分的实际意义，怎么能给出令人信服的评价工作，这个问题往往被应用工作者忽视了。

5 主成分分析概述 Principal Components Analysis

二. 主成分应用是考虑的几个问题

综合评价应用（续）④主成分个数的选取：一般性的选取原则有累计贡献率≥85%；或者特征值≥1的个数等，但不太过于机械，完全可能视具体情况主观决定主成分个数。

主成分的其他应用（主成分回归、主成分聚类）讨论的不多，似乎没有太多争议。

建议阅读文献：

林海明，对主成分分析法运用中十个问题的解析，统计与决策，2007年8月；

林海明，杜子芳，主成分分析综合评价应该注意的问题，统计研究，2013年8月。

6 因子分析概述 Factor Analysis

一. 因子分析的基本原理

因子分析也是一种数据化简方法，它是主成分分析的推广。在早期，这种方法在社科领域（心理学、教育学、经济学、社会学等）应用比较多，如今其他领域也广泛应用这种方法。

在社科领域，有许多“潜在变量”比如“态度”“爱好”“智力”“认识”“能力”对事物的结果具有重要影响，但它们都是不可观测的量。因子分析的目的就是找出这些对事物有决定性影响的“潜变量”，并用结构性的模型表达出来。

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \varepsilon_i \quad , \quad (\ i = 1,2,\cdots,p\ )$$

6 因子分析概述 Factor Analysis

一. 因子分析的基本原理（续）

从因子模型。我们可以看出，因子分析就是利用少数几个潜在变量（公因子）去解释可以观测的显性变量，或者说，我们把可以观测的变量的影响因素分解为两个部分：一是对所有变量都有影响的公因子，另外一部分是对每个变量单独有影响的特殊因子，其作用类似于回归分析中的随机误差项。

因子模型有几个基本假定：观测变量以及公因子都是标准化的，均值为0，方差为1；公因子与特殊因子不相关

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \varepsilon_i \quad , \quad (\ i = 1,2,\cdots,p\ )$$

6 因子分析概述 Factor Analysis

二. 因子模型中的几个统计量的意义

因子载荷：a_{ij}

变量共同度：$h_i^2 = \sum_{j=1}^{q} a_{ij}^2$

公因子的方差贡献：$g_i^2 = \sum_{i=1}^{p} a_{ij}^2$

正交因子模型（公因子互不相关）与斜交因子模型（公因子之间存在一定相关性）

因子旋转：有时为了更好地解释公因子实际意义，需要将因子进行旋转，利用旋转后的模型进行分析，初始因子模型与旋转因子模型，哪个更合适？要看具体情况。

6 因子分析概述 Factor Analysis

三. 因子分析与主成分分析的异同

参见林海明等人相关论文。

四. 因子分析的应用

（1）解释现象，探求原因（潜变量）；

（2）开展评价性工作，这点与主成分方法类似；

（3）对样品、变量进行分类（依据因子载荷分布、因子得分）。

五. 如何判断因子模型的优劣？

要看几个统计量值：KMO，变量共同度等

7 对应分析概述 Correspondence Analysis

一. 对应分析的基本原理

现实问题的研究中，经常需要对定性变量之间的关系进行分析，例如，研究婚姻状况（已婚、离异、丧偶、分居、未婚）与幸福状况（非常幸福、比较幸福、不幸福）的关系，研究者往往不仅想研究婚姻状况与幸福状况的整体关系，还想研究每种具体状况与幸福感之间的关系，这就需要采用对应分析。

市场营销中，厂家或者销售部门需要了解顾客的需要及喜好，可以调研后做对应分析。

7 对应分析概述 Correspondence Analysis

一. 对应分析的基本思想

对应分析比前面介绍的多元统计方法产生要晚一些，在不同国家有多种名称，在我国也叫相应分析，R—Q型因子分析。

R型、Q型因子分析分别研究变量和样品的内在影响因素及其结构，但事实上，变量与样品是同一个事物的两个层面，单纯的因子分析只能研究某个层面，不能全面探讨问题，在同一个“平台”上同时分析变量和样品，可以对现象或事物进行更深入的研究。

7 对应分析概述 Correspondence Analysis

三. 对应分析的软件实现

SPSS可以做对应分析，但数据文件的格式有特定形式，操作过程比其他方法也有所不同，请查阅相关资料。

8 典型相关分析 Canonical Correlation Analysis

一． 基本思想

典型相关分析研究两组随机变量之间的关系。

二. 具体做法：借鉴主成分方法的思想，分别将两组变量构建综合变量，以两个综合变量之间的相关性作为两组变量相关性的度量 。

三. SPSS软件的实现：需要调用相关模块，利用一个语法文件。

四. 结果的解释：典型相关系数是最重要的结果，但典型相关变量的系数，以及冗余指数也有意义。