

M MULTIPLE
FACTORS ANALYSIS
IN MEDICAL RESEARCH

医用多因素分析

(第三版)

■黄正南编著 ■湖南科学技术出版社

医用多因素分析

**MULTIPLE FACTORS ANALYSIS IN
MEDICAL RESEARCH**

(第三版)

黄正南编著

湖南科学技术出版社

内 容 提 要

《医用多因素分析》自1980年初版至今已修订扩充2次，初版时内容较少，共7章，主要介绍多元线性回归分析、判别分析、聚类分析和多因素的正交试验；1986年第二版更名为《医用多因素分析及计算机程序》，列12章，增加了主成分分析、因子分析、典型相关分析和配对资料的logistic回归分析，并附有计算机程序。本次为第三版，恢复原书名，但内容已大大扩充，增至17章，增加了岭回归分析和通路分析、有序样品聚类分析、成组资料的logistic分析和Cox回归分析，其内容基本上是重撰的，力争反映国内外医用多因素分析研究方面的诸多新进展，也融入作者从事医用多因素分析研究的成果，考虑到现在医学界各医疗、防疫、教学和科研等单位一般都有各种多因素分析方法的统计软件，故不再附计算机程序。

本书适合从事医学科研工作的医师、医学院校老师和医学科研人员等使用，也可作为医科学生和研究生进一步学习医学统计方法的教材。

医用多因素分析

(第三版)

黄正南 编著

责任编辑：王一方

*

湖南科学技术出版社出版发行

(长沙市展览馆路3号)

湖南省新华书店经销

湖南省新华印刷三厂印刷

厂址：长沙市韶山路158号

邮码：410004

(印装质量问题请直接与本厂联系)

*

1995年5月第3版第3次印刷

开本：850×1168毫米 1/32 印张：13.625 插页：4 字数：360,000

印数：19,100-23,300

ISBN 7-5357-1080-8

R·229 定价：14.50元

地科171-16

湘新登字004号

本书的第一版名为《医用多因素分析》，1980年出版，共7章，介绍了多元线性回归分析、判别分析、聚类分析和多因素的正交试验；第二版名为《医用多因素分析及计算机程序》，1986年出版，共12章，增加了主成分分析、因子分析、典型相关分析和配对资料的logistic回归分析，并为所介绍的各种多因素分析方法编制了计算机程序。现在第三版又恢复原书名，即《医用多因素分析》，其内容有如下说明：

1. 第三版扩充成17章，增加了岭回归分析和通路分析、有序样品聚类分析、成组资料的logistic回归分析和Cox回归分析。

2. 由于现在医学界各医疗、防疫、教学和科研等单位，一般都已有从国内外获得的各种多因素分析方法的统计软件，所以第三版不附所介绍的各种多因素分析方法的计算机程序。

3. 第三版对第二版的原12章内容都作了很大的改动和扩充，基本上是重写的，其依据是我15年来从事医用多因素分析的教学、科研和科研咨询的体会和经验，以及我在美国洛杉矶加州大学公共卫生学院生物统计系进修1年多变量统计方法的收获和心得。

2006/03

4. 第三版和第二版一样, 各章内容均有一定的独立性, 读者可结合自己开展医学科研的实际需要而选读。每章内容一般分成两部分: 第一部分介绍各种多因素分析方法的基本概念; 第二部分通过实例阐述其方法和步骤。读者在未用到某章内容时, 也可只先看其基本概念, 需要时再看其方法和步骤。基本概念部分要仔细阅读; 方法和步骤部分可据读者需要而定, 如果是用相应统计软件上计算机运算, 只需仔细阅读实例的初始数据的规范化、运算的最终结果和结果的分析及应用, 对具体的运算过程可不阅读或浏览一遍即可。

5. 第三版增加了两部分内容: 一是书末的“附 4: 各章公式的证明”, 用以论证一些多因素分析方法的原理及公式来源; 二是绝大多数章都有的每章的方法和步骤的最后的“几点说明”, 作为该章基本内容的补充和扩充, 特别是说明相应统计软件可能会输出的其他运算结果, 很多“说明”打了*号。公式证明和打*号内容供用本书作医学研究生和医学统计方法学习班教材的老师作教学参考用, 也为已学过多因素分析方法的读者深入钻研参考, 初学者不要阅读。

多因素分析方法是近代先进的统计分析方法, 为促进我国医学科研的现代化, 应在医学科研中推广使用。本书的读者是已从事或将从事医学科研工作的医师、医学院校老师、医学科研人员、医科学生和研究生等, 为了使本书内容及时更新以跟上现代科学技术发展的步伐; 为了书中所介绍的各种多因素分析方法对读者来说能通俗和实用, 使之学得会并用得上, 初版后又 2 次修订扩充, 此版修订扩充历时 3 年, 始成定稿。知我鉴我, 尽在书中。恳切希望广大读者批评指正。

本书的编著参阅了国内外大量有关多变量统计分析的著作和论文, 对这些著作和论文的作者, 谨在此深表谢意。

黄正南

1994 年 10 月

于湖南医科大学

CONTENTS

1	概论	[1]
2	多元线性回归分析	[18]
3	逐步回归分析	[42]
4	岭回归分析和通路分析	[64]
5	计量资料的两类判别分析	[84]
6	计量资料的多类判别分析	[110]
7	计数资料的判别分析	[139]
8	指标聚类分析	[156]
9	样品聚类分析	[167]
10	有序样品聚类分析	[182]
11	主成分分析	[193]
12	因子分析	[219]

13	典型相关分析	[245]
14	配对资料的 logistic 回归分析	[270]
15	成组资料的 logistic 回归分析	[299]
16	Cox 回归分析	[321]
17	多因素的正交试验	[344]
附 1	F 界值表	[362]
附 2	χ^2 界值表	[372]
附 3	部分常用正交表	[373]
附 4	各章公式的证明	[383]
附 5	英汉名词对照	[421]
	参考文献	[426]

1 概 论

近几十年来，随着计算机的高速发展，数理统计学的理论和方法有了很大的突破。卫生统计学（或医学统计学）随之也向前发展了。这种发展，表现在两个方面：首先是在空间（因素或变量空间）广度上的发展，这是横向的发展。传统的卫生统计学一般只研究 1 个或 2 个因素（医学上称指标），如同时研究多个因素，计算工作量繁复，有些是手工计算所不可能完成的。由于计算机运算速度快，不怕繁复，可同时研究几个甚至几十个因素，从而发展了很多多因素统计分析方法，这些统计分析方法在医学科研中有广泛的用途。其次是在时间深度上的发展，这是纵向的发展。传统的卫生统计学一般只研究一个固定时刻或时期的随机现象，其分布参数是固定不变的。现在可研究一个时间过程的随机现象，叫做随机过程（或时间序列），其分布参数随时间而变化。发展了很多随机过程模式，其中很多模式在医学科研中有广泛的用途。上述两个方面，在医学科研中，由于需要不同，还由于多因素分析比随机过程所涉及的数学知识要浅些，因此多因素分析比随机过程的运用要广泛得多、普及得多。当然，空间广度的多因素分析和时间深度的随机过程是既有区别又有联系的。本书专门论述多因素分析，有时也会涉及时间的效应。关于随机过程可参考其他专著。但作为一个全面的卫生统计学家，以推广和普及医学科研中的近代先进统计方法为己任，是需要掌握多因素

分析和随机过程两方面知识的。

1.1 多因素分析的基本概念

1.1.1 研究单位和研究因素，总体和样本

一个科学研究的最小研究单元叫做研究单位或个体，是据研究目的而定的。有狭义的自然界的真正个体，如 1 个人、1 头动物、1 只眼睛等；也有广义的人定的个体，如 1 家人、1 班学生、1 升水等。因素是描述研究单位的各种特征的。描述研究单位特征的因素很多，把要研究的因素叫做研究因素，也是根据研究目的而定的。有狭义的研究单位本身具有的因素；也有广义的研究单位所在环境具有的因素。如人是研究单位，身高、体重和血压等是本身具有的因素；而所在环境中的人群、空气、水和土壤中的各种疾病传染源、毒物等就是环境因素。统计学的特点是研究群体的数量特征，因此统计学要研究的个体因素必须是可直接或间接观测因素取值的因素。

总体是所有要研究的个体（指据研究目的所定同质个体）的研究因素值的集合。为简化起见，可把总体看作所有要研究的个体的集合。卫生统计学的方法和任务是：从总体随机抽取一些个体，测量各研究因素值，每个被抽到的个体叫做样品，由样品构成样本，样本中的样品数叫样本含量，计算描述样本的统计量（即样本统计指标）；然后由样本信息（个体研究因素值和样本统计量）推断描述总体特征的参数（即总体统计指标）。

多因素分析的第一个重要问题是：对于每种多因素分析方法要求总体符合什么分布？很多多因素分析模型在理论上要求总体符合多元正态分布。但在实际科研中，这个要求是很难满足的，也很难用统计方法检验这个要求是否达到了。现在实际可行的办法是：如果就每个单个的随机变量而言，在理论上或经验上知道它们都服从单变量正态分布，那么就认为由多个这样的随机变量构成的总体服从多元正态分布。甚至可把这个要求更加放宽，明

知有个别变量（如两分变量和不连续变量，见下文）不符合单变量正态分布，因此总体不符合多元正态分布，也用要求多元正态分布总体的多因素分析方法。这一方面是据数理统计学的中心极限定理，不管原变量总体服从什么分布，当样本含量很大时，样本统计量近似服从正态分布；另一方面是对建立的多因素分析模式进行是否符合实际的检验，如果符合实际，达到了要求，就可接受该模式。

多因素分析的第二个重要问题是：样本含量 n 和研究因素个数 m 之间应该有什么样的比例关系？这个问题迄今尚得不到很好的回答。原则上是：研究因素的个数 m 越多，需要的样本含量 n 就越大。有人提出一个大致的准则：样本含量 n 至少是研究因素个数 m 的 5~10 倍，但这是很主观的。（本书的例题，有些没有达到这个样本含量要求，是为了简化运算和节省篇幅。）

1.1.2 因素和资料的分类及转换

因素（或指标）和相应的统计资料按其性质可分成下述 3 类。

1. 定量因素（定量指标）和计量资料：定量因素是对研究单位的定量特征的描述，有大小和单位，叫变量（数值变量），其取值叫变量值。如人的身高（m）、体重（kg）和血压（kPa）等；又如家庭人口数（人/家）、水中某种毒物含量（mg/L）等。全部由变量构成的统计资料叫做计量资料，是一群单变量值、双变量值或多变量值。

2. 定性因素（定性指标）和计数资料：定性因素是对研究单位的定性特征的描述，有类别。其特点是类别是客观存在的，各类无秩序，可以任意排列；类和类之间界线清楚，不会错判。如人的性别有男和女 2 类；血型有 O 型、A 型、B 型和 AB 型 4 类等。全部由定性因素构成的统计资料叫做计数资料，是指每个因素的每类有多少个个体。

3. 等级因素（等级指标）和等级资料：等级因素是对研究

单位的等级特征的描述，分等级。其特点是级别是主观划分的，各级没有大小，但有秩序，必需自低（或弱）到高（或强）或自高（或强）到低（或弱）排列，故等级因素又可叫做半定量因素；级和级之间界线模糊，可能错判。如腹痛可分为不痛（正常）、轻度痛、中度痛和重度痛 4 级；某种反应可分为-和+ 2 级等。全部由等级因素构成的统计资料叫做等级资料，是指每个因素的每级有多少个个体。

用 3 类因素描述研究单位的特征，就区别个体的因素特异性而言，定量因素最好。凡能采用定量因素来描述研究单位的研究特征时，应尽量采用定量因素，其获得的统计信息会最多。还要注意因素客观性。定性因素分类是客观的，观测得出的定量因素值也是客观的。但用等级因素一定要确定分级的客观标准，减少评判级别的主观性；如果定量因素值是评分得出的，就更要仔细确定计分的客观标准，尽量减少评定分数的主观性。

可把定性因素称为名义变量，其取值是各类；等级因素称为等级变量，其取值是各等级。也可把定性因素和等级因素统称属性变量或分类变量（等级看作有序分类）。多因素分析在数理统计学中叫做多变量（数值变量和属性变量）分析或简称多元分析。多因素分析常把计数资料和等级资料统称计数资料。

基本上各种多因素分析方法都要求所有的因素或所有的原因因素是定量因素，用计量资料的多因素分析方法；只有个别多因素分析方法所有的因素或所有的原因因素也可是定性和（或）等级因素，用计数资料的多因素分析方法。

多因素分析的第三个重要问题是：如果在研究因素中，既有定量因素，又有定性和（或）等级因素，怎么处理？回答是要进行因素的转换。如果用计数资料的多因素分析方法，要把定量因素转换成等级因素，这是个别情况；如果用计量资料的多因素分析方法，要把定性和（或）等级因素转换成定量因素，这是一般情况。

定量因素转换成等级因素比较容易，只要据专业知识，规定

划分变量值区间的标准,就可将其转换成相应的等级。如血压的舒张压本是定量指标,若按规定把舒张压低于 8kPa (60 mmHg) 定为低血压, 8kPa (60mmHg) 到 12kPa (90 mmHg) 定为正常, 高于 12kPa (90mmHg) 定为高血压, 则舒张压就为分成低血压、正常和高血压 3 级的等级指标了。同样可把身高这个定量指标转换成矮、中和高 3 级的等级指标。

定性因素和等级因素转换成定量因素比较难办, 叫做指标的数量化。定性指标的数量化是引入指示变量、即取值 (0, 1) 的两分变量 x (这种转换成的 (0, 1) 变量称为伪变量或哑变量)。若定性指标只有 2 类, 则用 1 个 (0, 1) 变量 x 。如性别用 x 表示: 男为 $x=0$, 女为 $x=1$ 。若定性指标有 k ($k \geq 2$) 类, 则用 $k-1$ 个 (0, 1) 变量 $x_1, x_2, \dots, x_{k-1}$ 。如血型用 x_1, x_2 和 x_3 表示: O 型为 $x_1=0, x_2=0$ 和 $x_3=0$, A 型为 $x_1=1, x_2=0$ 和 $x_3=0$, B 型为 $x_1=0, x_2=1$ 和 $x_3=0$, AB 型为 $x_1=0, x_2=0$ 和 $x_3=1$ 。等级指标的数量化问题迄今尚未得到很好的解决, 现在一般是按等级自低到高用 1 个取值 0, 1, 2, ... (或 1, 2, 3, ...等) 的不连续变量 x (这种转换成的不连续变量也称为伪变量) 表示。如咳嗽这个等级指标分不咳嗽 (正常)、轻度咳嗽、中度咳嗽和重度咳嗽 4 级, 用 x 表示: 不咳嗽为 $x=0$, 轻度咳嗽为 $x=1$, 中度咳嗽为 $x=2$, 重度咳嗽为 $x=3$ 。等级指标的如此数量化是太主观了。现在国内外很多学者正在从事等级指标的数量化研究, 如探求描述人的健康状况、生活质量等的指标及数量化研究。读者在以后的实际科研工作中, 如有可能, 可选用当时自己认为较合理的等级指标的数量化模式。另外, 等级指标若要分别得出每个等级的作用时, 仍定用指示变量 (0, 1) 变量作指示变量。符合因果因素因个变量因个变量值分带简单可变量分为连续变量和来连续或间断) 变量。数值变量一般是连续变量 (如身高、体重、年龄等) 但也可是不连续变量 (如家庭人口数); 名义变量转换成的取值 (0, 1) 变量和等级变量转换成的取值 0, 1, 2, ... 变量是不连

续变量。

1.1.3 多因素分析和单因素、双因素分析比较的优点

多因素分析和单因素、双因素分析比较，主要优点有下述 2 个。

1. 取得原始资料容易：单因素、双因素分析一次只能研究 1 个或 2 个因素，要用严格的科研设计（调查设计或实验设计）来保证研究因素外其他影响结果的因素、即干扰因素的齐同，因此取得原始资料困难。而多因素分析可同时研究几个甚至几十个因素，一方面可把研究的因素增多，不必被迫把一些要研究因素作干扰因素处理；另方面有些干扰因素难于控制齐同，也可把它们纳入研究因素之内，最后分析时不考虑就行了。而且要在科研开始前就定出对结果的 1 个研究因素（加结果因素就为双因素），有时本身就是困难的事。如高血压的致病因素很多，研究高血压很难预先只定出 1 个研究因素，而多因素分析可把所有怀疑的高血压致病因素作研究因素处理。

2. 可从整体分析结果：单个因素对结果因素的作用为单独作用。多个（包括 2 个）因素对结果因素的作用为联合作用，包括各因素对结果因素的单独作用和交互作用。因此，如果某结果因素受多个因素影响，研究多个因素对结果因素的作用，那么各因素对结果因素的作用就不再是单独作用，而是其在联合作用中的贡献。单因素（加结果因素就为双因素）分析一个一个地研究其因素对结果因素的作用，就只考虑了单独作用。这一方面会把没有单独作用但和其他因素在一起有联合作用的因素忽略掉；另方面有些因素虽有单独作用，但其单独作用可被其他 2 个或多个因素的联合作用所代替，而在联合作用中是多余的。而多因素分析同时研究多个因素对结果因素的联合作用，既考虑了各因素对结果因素的单独作用，又要考虑各因素间对结果因素的交互作用，分析就全面了。高良成）量变对变虽虽一量变直姓。量变（洵直如如并量变义谷；（爔口人致案成）量变对变不虽何出出直不虽量变... 5, 1, 0 直如如如并量变对变味量变 (1, 0)

1.2 多因素分析的主要内容和主要任务

1.2.1 多因素分析的主要内容

多因素分析是研究多个相依因素（变量）之间的关系以及具有这些因素的样品（个体）之间的关系，但主要是研究多个相依因素之间的关系。本书全面介绍了研究多个相依因素之间关系的多因素分析方法，包括多元线性回归分析、判别分析、指标聚类分析、主成分分析和因子分析、典型相关分析、logistic 回归分析、Cox 回归分析、多因素的正交试验。多因素分析方法的聚类分析、主成分分析和因子分析，可分为就因素论述的方法（称 R 型多因素分析）和就样品论述的方法（称 Q 型多因素分析）。聚类分析本书除介绍指标聚类分析（R 型聚类分析），也介绍样品聚类分析（Q 型聚类分析）。但主成分分析和因子分析，本书只介绍多指标的主成分分析和因子分析（R 型主成分分析和因子分析），由于多样品的主成分分析和因子分析（Q 型主成分分析和因子分析）在医学科研中用得少，就不介绍了。

研究多个相依因素之间关系的多因素分析方法，就研究目的而言可分为下述两大类。

1. 研究因素间的依赖性：这种情况有原因因素和结果因素之分，一般原因因素有多个（包括 2 个），结果因素只有 1 个（若有多个结果因素，则一般每次研究 1 个结果因素）。研究多个原因因素对结果因素的作用或结果因素对多个原因因素的依赖。本书的多元线性回归分析、判别分析、logistic 回归分析、Cox 回归分析、多因素的正交试验属此范畴。

2. 研究因素间的互依性：这种情况各因素是平等的。研究多个因素各因素间的彼此关系或彼此影响。本书的指标聚类分析、主成分分析和因子分析、典型相关分析属此范畴。

1.2.2 多因素分析的主要任务

研究多个相依因素之间关系的各种多因素分析方法，其主要

任务是要求简化研究问题的复杂性，从而抓住主要矛盾，使研究问题明朗化。这可从下述两方面入手。

1. 直接减少因素（变量）个数：直接减少因素个数是从原有的因素中选出一些典型的、有代表性的和能说明问题的因素，舍弃一些不典型的和无代表性的因素。这种因素舍弃甚至以不惜损失少量资料信息为代价。如多元线性回归分析的逐步回归分析、logistic 回归分析、Cox 回归分析中只把对应变量作用有显著性的自变量引入回归方程；判别分析的逐步判别分析只把对判别分类作用有显著性的判别指标引入判别函数；指标聚类分析中通过指标聚类挑选出少量几个典型指标来代表原来众多指标；多因素的正交试验中寻找出对试验指标作用有显著性的试验因素。

2. 通过变量变换减少参数个数：原来研究的多个变量往往是彼此相关的，有多个相关系数为参数，可通过变量变换把彼此相关的原变量（原指标）转换成彼此独立的新变量（新指标），这样就减少许多相关系数的参数。主成分分析和因子分析、典型相关分析都属此范畴。

直接减少变量个数和通过变量变换减少参数个数两者是紧密联系不可分的。直接减少变量个数一定减少了参数个数；变量变换在减少了相关系数的参数的同时，新变量个数往往会少于原变量个数。

例如，原研究有 m 个变量，则原来的参数个数为：各变量的均数 m 个、方差 m 个，各变量间的相关系数 $m(m-1)/2$ 个，共计有 $m+m+m(m-1)/2=m(m+3)/2$ 个参数。如果减少 1 个变量，则减少的参数个数为 $m(m+3)/2-(m-1)(m+2)/2=m+1$ （1 个均数、1 个方差、 $m-1$ 个相关系数）；如果把 m 个相关的原变量转换成 m 个独立的新变量，则减少的参数（即相关系数）个数为 $m(m-1)/2$ 。

1.3 多因素分析的主要方法

本书介绍的现在数理统计学中发展的几种多因素分析的主要方法，都在医学科研中得到了很好的应用。而且可以说，多因素分析在医学领域内的应用是比较成功的。下面把本书叙述的几种多因素分析的主要方法，先作扼要介绍，使读者预先有一个总的认识，并可结合自己从事科研的实际需要，选择阅读本书的各章内容。

1.3.1 多元线性回归分析

作出以自变量（原因因素）估计应变量（结果因素）的多元线性回归方程。应变量和自变量都是定量指标。如果原因因素中有少数定性指标和（或）等级指标要转换成定量指标。

多元线性回归分析分为一般多元线性回归分析和逐步回归分析（或逐步多元线性回归分析）。一般多元线性回归分析是一次拟合多元线性回归方程，在拟合回归方程的过程中不能挑选自变量。也就是说，预先凭理论或经验定的自变量都引入回归方程，因此预先定自变量的工作很重要。不能遗漏对应变量作用有显著性的自变量，特别是不能增加对应变量作用无显著性的自变量，这在实际科研中难于作到。逐步回归分析是逐步拟合多元线性回归方程，在拟合回归方程的过程中挑选对应变量作用有显著性的自变量引入回归方程，对应变量作用无显著性的自变量不会引入回归方程或引入回归方程后会被剔除，因此可把一些对应变量作用可能有显著性的自变量来让逐步回归挑选，这就给实际科研带来很大方便。

岭回归分析和通路分析是作为多元线性回归分析内容的补充和发展。如果用一般多元线性回归分析解释各自变量对应变量的作用不满意时，若不宜用逐步回归分析，则可用岭回归分析。如果要把各自变量和应变量之间的数量关系用几何图形来形象化，则用通路分析。通路分析可描述原因变量（远因变量）、中间变

量（近因变量）和应变量之间的关系；还可描述多个应变量和多个自变量之间错综复杂的关系。

1.3.2 判别分析

作出以判别指标（原因因素）判别个体属性（结果因素）分类或分级的判别函数或判别指数表。个体属性是定性指标或等级指标；判别指标可全是定量指标，也可全是定性指标和（或）等级指标。

判别分析就个体属性而言可分为2种：若个体属性分成两类，叫做两类判别；若个体属性分成多类，叫做多类判别。判别分析就判别指标性质而言可分成2种：若判别指标是定量指标，叫做计量资料的判别分析，如有定性指标和（或）等级指标要转换成定量指标；若判别指标是定性指标和（或）等级指标，叫做计数资料的判别分析，如有定量指标要转换成等级指标。计量资料的判别分析得出的结果是判别函数；计数资料的判别分析得出的结果是判别指数表。

计量资料的判别分析就拟合判别函数而言，和回归分析拟合回归方程一样，可分为一般判别分析和逐步判别分析。一般判别分析是一次拟合判别函数，不能挑选判别指标，预先凭理论或经验定的判别指标都引入判别函数，因此预先定判别指标的工作很重要。逐步判别分析是逐步拟合判别函数，挑选对判别分类作用有显著性的判别指标引入判别函数，对判别分类作用无显著性的判别指标不会引入判别函数或引入判别函数后会被剔除，因此可把一些对判别分类作用可能有显著性的判别指标来让逐步判别挑选。

判别分析方法就判别原理而言可分为非概率型判别方法和概率型判别方法。非概率型判别方法只适用于计量资料的两类判别，用判别函数和两类分界值判别样品（个体）的类别，算出判别函数值大于或等于分界值属第一类，小于分界值属第二类；概率型判别方法用多个（包括2个）判别函数判别样品（个体）的类别，算出多个判别函数值，哪个判别函数值最大（即归属哪类