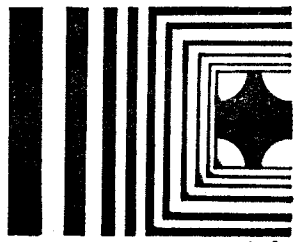
医用多因素分析及计算机程序

YIYONGDUOYINSUFENXIJIJISUANJICHENGXU

黄正南编著

湖南科学技术出版社

医用多因素分析及计算机程序



黄正南编著

湖南科学技术出版社

内 容 简 介

本书介绍多因素分析在医学领域中的应用，阐明各种多因素分析的基本概念，结合医学科研实例说明方法和步骤，并编制了BASIC计算程序。本书可作为广大医科学生、医师和医学科研工作者进一步学习医用统计方法的教材和参考书。

医用多因素分析及计算机程序

黄正南 编著

责任编辑：张碧金 朱 杰

*

湖南科学技术出版社出版发行

（长沙市展览馆路14号）

湖南省新华印刷二厂印刷

*

1980年9月第1版第1次印刷

1986年1月第2版第2次印刷

开本：850×1168毫米 1/32 印张：9.75 插页：平装1 精装5 字数：258,000

印数：（平装）1——7,900 （精装）1——500

统一书号：14204·25 定价：（平装）2.30元 （精装）3.70元

前　言

这本书的原版名为《医用多因素分析》，1980年初由湖南科学技术出版社出版。当时，计算机在医学领域的应用还刚开始，广大医务人员和医学生对计算机存在神秘感，对最新的统计方法也不很了解。鉴此，笔者本着让读者易学易用的原则来写的。五年过去了，现在计算机在医学领域已有普及之势，根据新的情况，决定对原书作较大改动、再版。

原书只有七讲，介绍了多因素的回归分析、判别分析、聚类分析和正交试验。目前扩充为十二讲，增加了主成分分析、因子分析、典型相关分析和 Logistic 回归分析等。各讲均有一定的独立性，读者可以结合自己的实际需要而选读。每一讲的内容分成两部分：第一部分介绍各种多因素分析的基本概念；第二部分通过实例说明分析的方法和步骤，为了条理化，并用方框图作了总结。读者在未用到某讲内容时，也可只先看基本概念，需用时再看方法和步骤。本书后部分还用BASIC语言编制了所介绍的多因素分析的计算程序，便于直接上机运算。

本书的编写得到了我院的领导、老师和同学的大力支持，于仁和同志协助编写了部分程序，卫生统计界前辈对我给予了具体指导和热情鼓励，在此深表谢意。我恳切希望继续得到前辈的指教和读者的帮助。

编　者　1985年3月于湖南医学院

目　录

第一讲　概　　论

近几十年来，随着电子计算机的发展，数理统计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，表现在以下两个方面：首先是表现在空间（因素或变量空间）广度上，这是横向的发展。传统的卫生统计学一般只研究一个或两个因素（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，其中很多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，其分布参数是不变的。现在可研究一个时间过程的随机现象，叫随机过程（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也有广泛的用途。上述两方面相对说来，由于多因素分析比随机过程涉及的数学知识要浅些，所以，在医学科研中，多因素分析比随机过程的发展也要快一些，运用也广一些。当然，空间广度的多因素分析和时间深度的随机过程两者是既有区别又紧密联系的。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，是需要掌握多因素分析和随机过程两方面知识的。

一、什么叫做多因素分析？

（一）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研究单位或个体，它是据研究目的而定的。有狭义的自然界的真正个体，如一个人，一头动物，一只眼睛等；也有广义的人定的个体，如一家人，一班学生，

一毫升水等。因素是描述研究单位各种特征的。描述研究单位特征的因素很多，把要研究的因素叫研究因素，它也是据研究目的而定的。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。

严格说来，总体是所有要研究的个体的研究因素的集合。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，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。然后由样本推断总体。

多因素分析的一个首要问题是：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都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了多元正态分布。甚至可把这个要求更加放宽，明知有个别单变量（如两分变量和不连续变量，见下文）不符合单变量正态分布，因此总体则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。这一方面是据数学中心极限定理，不管总体服从什么分布，当样本含量很大时，样本统计量近似服从正态分布；另方面是对建立的多因素分析模式进行是否符合实际检验，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：研究的因数个数m愈多，需要的样本含量n就愈大。有人提出一个粗糙的准则：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。（本书的一些例题， 没有达到这个样本含量要求，有些是为了节省篇幅和简化运算。）

（二）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三类。

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。如人的身高(厘米)、体重(公斤)和血压(毫米汞柱)等；又如家庭的人口数(人/家)和水中某种毒物含量(毫克/升)等。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，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两类；血型有O型、A型、B型和AB型四类等。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。其特点是等级是主观划分的，没有大小，但有秩序，必需自低（或弱）到高(或强)，或自高到低排列，故等级因素又可叫半定量因素；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、中度痛和重度痛四级；反映可分为－和＋两级等。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。

可把定性因素和等级因素统称属性变量。多因素分析又可叫多变量(一般变量和属性变量)分析或简称多元分析。可把计数资料和等级资料统称计数资料。

在多因素分析中，和单因素、双因素分析一样，计量资料有计量资料的统计分析方法，计数资料有计数资料的统计分析方法，不过计量资料的多因素分析方法远多于计数资料的而已。

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；如果用计量资料的多因素分析方法，要把定性因

素和/或等级因素转换成定量因素。

定量因素转换成等级因素比较容易，只要根据专业知识，规定划分变量值间隔，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60毫米汞柱定为低血压，将60毫米汞柱到90毫米汞柱定为正常，高于90毫米汞柱定为高血压，则舒张压就分成低血压、正常和高血压三级的等级指标了。同样可把身高这个定量指标转换成矮、中和高三级的等级指标。

定性因素和等级因素转换成定量因素比较难办，这个问题叫做指标的数量化，迄今尚未得到很好的解决。本书用现在的一般解决办法：定性指标转换成取值（0,1）的两分变量x。若定性指标只有两类，则转换成一个取值（0,1）的两分变量x。如性别用x表示，男为$x=0$和女为$x=1$（也可女为$x=0$和男为$x=1$）。若定性指标有m类，则转换成$m-1$个取值（0,1）的两分变量x_1，x_2，…，x_{m-1}。如血型用x_1、x_2和x_3表示：O型为$x_1=1$、$x_2=0$和$x_3=0$；A型为$x_1=0$、$x_2=1$和$x_3=0$；B型为$x_1=0$、$x_2=0$和$x_3=1$；AB型为$x_1=0$、$x_2=0$和$x_3=0$。等级指标则按等级自低到高转换成一个取值0，1，2，…（或1，2，3，…等）的不连续变量x。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四级，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$x=3$。等级指标的如此转换是太主观了，现在国内外很多人正在从事等级指标的数量化的研究，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数量化模式。

（三）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，主要优点有下述两个。

1.取得原始资料容易：单因素、双因素分析一次只能研究一个或二个因素，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，也可把其纳入研究因素内，最后分析不考虑就行了。这样原始资料就容易取得。而且要在科研开始前就定出对结果的一个或二个研究因素，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一个或二个研究因素，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：如果某个结果受多个因素影响，单因素(加结果因素就可叫双因素)分析一个一个地研究其因素对结果的作用(影响)，如果设计得不好，往往会被因素间的交互作用所蒙蔽。特殊情况时会把有作用判断为无作用或无作用判断为有作用。而多因素分析同时研究多个因素对结果的作用，既可研究各因素的单独作用，又可研究因素间的交互作用，分析就全面了。

二、什么是多因素分析的主要内容和主要任务？

(一) 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。有些多因素分析方法（研究因素间的互依性的方法，见下文）可分为就因素论述的方法和就样品论述的方法两类。在实际科研中，主要用到论述因素的方法，所以本书的多因素分析主要是就因素论述方法，把很多就样品论述的方法省略了。

多因素分析方法就目的而言可分成下述两大类。

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，研究原因因素对结果因素的作用或结果因素对原因因素的依赖。本书的回归分析、判别分析、Logistic 回归分析和正交试验属此范畴。

2.研究因素间的互依性：这种情况认为各因素是平等的，研究各因素间的彼此关系或彼此影响。本书的聚类分析、主成分分析、因子分析和典型相关分析属此范畴。如前所述，研究因素间互依性的方法也可用作研究样品间的互依性。

（二）多因素分析的主要任务

几乎所有的多因素分析方法的一个主要任务都是要求简化研究问题的复杂性。这既可减少计算工作量，又可抓住主要矛盾，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。如一般回归分析和 Logistic 回归分析中只把对因变量作用显著的自变量引入回归方程；判别分析中只把对判别结果有作用的判别指标选入判别表或判别函数；聚类分析中通过指标聚类选出少量几个典型指标来代表原来众多指标；正交试验选出使结果最好的因素的最优搭配等。

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)，这样就可减少许多相关系数的参数。主成分分析、因子分析和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，感兴趣的是各变量的均数、方差和变量间的相关系数。则原来参数个数为：均数m个，方差m个，相关系数$\frac{1}{2}m(m-1)$个，共计有 $m+m+\frac{1}{2}m(m-1)=\frac{1}{2}m(m+3)$个参数。如果减少一个变量，则减少的参数个数为$\frac{1}{2}m(m+3)-\frac{1}{2}(m-1)(m+2)=m+1$；如果把m个相关的原变量转换成 m 个独立的新变量，则减少的参数个数为$\frac{1}{2}m(m-1)$。

三、多因素分析包括哪些主要方法？

现在数理统计发展的几种多因素分析的主要方法，差不多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，先作一个扼要介绍，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，选择阅读本书各讲内容。

（一）回归分析

作出以自变量（原因因素）估计因变量（结果因素）的回归方程。自变量和因变量都是定量指标，定性指标和/或等级指标要转换成定量指标。

回归分析可分为一般回归分析和逐步回归分析。一般回归分析是一次建立回归方程，在建立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会进入回归方程，因此预先定自变量的工作很重要。不能遗漏对因变量有作用的自变量，不能增加对因变量无作用的变量，这在实际科研中比较难办。逐步回归分析是逐步建立回归方程，在建立回归方程的过程中只挑选对因变量有作用的自变量进入回归方程，对因变量无作用的变量不能进入回归方程或进入回归方程后被剔除，因此可把一些怀疑对因变量有作用的变量来让逐步回归挑选，这就给实际科研带来很大方便。

（二）判别分析

作出以判别指标（原因因素）判别事物属性（结果因素）分类或分级的判别指数表或判别函数。事物属性一定是定性指标或等级指标，判别指标可是定性指标、等级指标或定量指标。

判别分析就事物属性分类而言可分为两种：若事物属性分成两类，叫两类判别；若事物属性分成多类，则叫多类判别。

判别分析方法就判别指标性质而言可分为两种：若判别指标是定性指标和/或等级指标用计数资料的判别分析方法，这时定量指标要转换成等级指标；判别指标是定量指标用计量资料的判别分析方法，这时定性指标和等级指标要转换成定量指标。计数资

料的判别分析方法比较简单，得出的结果是判别指数表；计量资料的判别分析方法得出的结果是判别函数。

判别分析和回归分析一样，也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建立判别函数，在建立判别函数的过程中不能挑选判别指标，预先凭理论或经验定的判别指标都会进入判别函数，因此预先定判别指标的工作很重要。逐步判别分析是逐步建立判别函数，在建立判别函数的过程中只挑选对判别事物属性有作用的判别指标进入判别函数，无判别作用的指标不能进入判别函数或进入判别函数后被剔除，因此可把一些怀疑有判别作用的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概率型判别方法和非概率型判别方法。概率型判别方法计算事物属性各类发生的概率，哪类发生的概率大就判别属哪一类；非概率型判别方法只适用于计量资料的两类判别，用临界值判断，算出判别函数值大于临界值判别为一类，小于临界值判别为另一类。

（三）聚类分析

1.指标聚类法：在存在众多指标的情况下，可以将相近指标聚成类，每类找一个典型指标，从而用少量几个典型指标来代替原来的众多指标，称指标聚类法。

2.样品聚类法：在样品个数甚多的情况下，可以把相近样品聚成类，然后作比较研究，称样品聚类法。

（四）主成分分析

由多个原指标找出少量几个新的综合指标的方法。综合指标综合了原来所有指标的方差信息，依综合信息由多到少把新指标依次叫做第一主成分，第二主成分，……等。

（五）因子分析

寻找决定多个原指标又不能直接测得的少量几个公共因子（新指标）的方法。公共因子反映了原来所有指标之间的相关变化。因子分析和主成分分析不同，主成分分析是以综合指标来反映原指标；因子分析是以公共因子来解释原指标。

（六）典型相关分析

寻找代表两组指标的典型变量和相应的典型相关系数的方法。典型指标之间的相关提取了原来所有两组指标间的相关信息。依提取信息由多到少，即典型相关系数由大到小依次把新的每对典型变量叫做第一对典型变量，第二对典型变量，……等。

（七）Logistic回归分析

作出以危险因素（原因因素）估计某病在某段时间内发病概率（结果因素）的Logistic回归方程。Logistic回归方程和一般回归方程不同，因为发病概率的变化范围只能从0到1。

Logistic回归分析方法采取逐步引入危险因素法，只挑选对发病概率有作用的危险因素进入回归方程，因此可把一些怀疑对发病有作用的因素来让Logistic回归挑选。

Logistic回归分为成组Logistic回归模型和配对Logistic回归模型。成组Logistic回归用分层来控制干扰因素齐同，估计的参数多；配对Logistic回归用配对来控制干扰因素齐同，估计的参数少，较成组模型要准确一些。本书只介绍配对Logistic回归分析。

（八）正交试验

寻找使试验结果（结果因素）最好的各因素（原因因素）各水平的最优搭配的方法。试验结果是定量指标，各因素可是定性指标、等级指标或定量指标。定性指标的类别和等级指标的等级是水平，如果是定量指标则用不连续的几个取值作水平。关于试验结果是定性指标或等级指标的正交试验，国内有人研究，并已取得一定成绩。

四、多因素分析在医学科研中有哪些主要用途？

多因素分析在医学科研中的应用愈来愈广泛，笔者的知识和经验有限，见闻不够，以下所列很不全面，而且比较机械，仅供读者参考。

1.推算不易测得的指标：回归分析。

2.指标对时间等的变化曲线拟合：回归分析。

3.绘制疾病的患病、发病和死亡等趋势地图，寻找疾病的高

发区和低发区：回归分析。

4.寻找疾病病因：回归分析、判别分析和Logistic回归分析。

5.寻找疾病发病概率和危险因素的数量关系： Logistic 回归分析。

6.疗效估计、生存期预报：回归分析和判别分析。

7.流行病预报：回归分析和判别分析。

8.疾病的计量诊断：判别分析。

9.寻找多个指标的典型指标：指标聚类分析。

10.寻找样品的分类方法：样品聚类分析。

11.寻找反映多个指标的综合指标：主成分分析。

12.寻找决定多个指标的公共因子：因子分析。

13.寻找两组指标间的相关关系：典型相关分析。

14.寻找最优药物（试剂）配制、最佳细菌繁殖条件和最优治疗方案等：正交试验。

五、如何进行医用多因素分析的科研？

医用多因素分析作为统计方法是为医学科研服务的，因此从科研设计、具体实施到结果考核都要和专业知识紧密相结合。下面仅就其作为统计方法而言，简述其步骤。

1.科研设计：明确专业目的，选择多因素分析方法。

2.收集资料：通过现场调查、临床试验或动物实验等收集资料，资料要完整、准确和及时。然后根据选用的多因素分析方法通过指标转换把资料规范化，使其全部成为定量指标，或定性指标和等级指标。

3.数据处理：编制程序，上电子计算机计算。这一步如果自己作有困难，可利用别人编制的现成程序，或和计算机工作人员合作完成此步。但至少自己要有一套核对程序本身和程序运行是否对，精度是否达到要求的检验题目。能正确和顺利的设计这些检验题目，是本书各讲的方法和步骤部分的最低要求。

4.结果考核：检查所用的多因素分析数学模式和所得结果是否与专业理论和经验相符合，如果不符合，则不能公布结果，并

要追查不符合的原因。

5.结果应用：在实际应用中继续检验所用的多因素分析数学模式，如必要则进行适当修正。