工商管理硕士(MBA)系列教材

现代应用统计分析方法

王明华　主编

中国统计出版社

工商管理硕士(MBA)系列教材

现代应用统计分析方法

主　编　　王明华

中国统计出版社

(京)新登字 041 号

图书在版编目(CIP)数据

现代应用统计分析方法/王明华主编．
－北京：中国统计出版社，1999.12
工商管理硕士(MBA)系列教材
ISBN 7－5037－3104－4/C.1695

I.现...
II.王...
III.统计分析－研究生－教材
IV.C813

中国版本图书馆 CIP 数据核字(1999)第 40187 号

中国统计出版社出版
(北京三里河月坛南街 75 号　100826)
新华书店经销
北京科技印刷厂印刷
*
850×1168 毫米　32 开本　10.125 印张　25 万字
1999 年 12 月第 1 版　1999 年 12 月北京第 1 次印刷
印数：1—5000 册
ISBN 7－5037－3104－4/C.1695
定价：14.40 元

序

中南财经大学是我国1978年首批恢复招收硕士研究生的普通高等院校之一，1993年经国务院学位办正式批准开始招收工商管理硕士(MBA)研究生。为了适应工商管理硕士教育的需要，我们组织了部分专业基础扎实、教学经验丰富、重视理论联系实际、熟悉国外工商管理教育的教师，在比较短的时间里，编写了这套工商管理硕士(MBA)系列教材。这套教材共24本。除供工商管理硕士研究生(MBA)使用之外，还可作为经济管理类各专业研究生和本科生的选修教材，同时也可作为从事经济管理理论研究和实际工作的干部自学参考书。

工商管理硕士教育在我国尚处在试点阶段，我们组织编写工商管理硕士系列教材也只是一种初步尝试，由于经验不足，肯定存在某些缺陷，甚至错误。我们将继续努力开拓，希望在不久的将来，能奉献给广大读者一套体系完备，内容更适用，方法更科学的工商管理硕士系列教材，望广大读者不吝赐教。

中南财经大学工商管理教材编审委员会

1994年7月

前　言

本书主要是为工商管理硕士(MBA)研究生编写的教材。根据工商管理硕士(MBA)研究生的培养目标,本书系统地介绍了社会经济研究中最常用的一些统计方法,重点放在各种统计方法的实际应用上。

本书的取材主要考虑到:一　连续性。注意与本科生数理统计课程的教学大纲内容相衔接,使修过本科数理统计课程的读者能顺利地过渡到本书内容的学习。二　系统性。本书各部分内容力求既相对独立、完整,又相互联系,组成一个有机的整体。既有经典的统计方法,又融入一些新方法、新成果。三　实用性。本书所介绍的统计方法是处理社会经济问题的最常见的统计方法。在介绍有关方法的同时,特别注意到应用这些方法处理实际问题时应该考虑的问题,包括这些问题的识别及解决办法。

本书力求通俗易懂,简洁明了,方法程序化。在介绍各种有关统计方法与思想的同时,辅以典型实际例子,以便于读者掌握应用。本书的重点是统计方法的运用,对统计方法应用中的问题的识别及处理作了较系统、较详细的介绍。以期读者在学习本书内容后,能够将其直接应用于经济管理的实践和研究中去。在编写方法上考虑到数学推证往往是学习经济专业的读者在阅读中的主要困难,因而尽量避免繁杂的数学推导,使读者易读易懂,增强本书的可读性。

本书得到了华中理工大学博士生导师黄志远教授的指导。他对本书内容进行了全面的、详细的审查。在此表示衷心的感谢。

在本书与广大读者见面之际,我还要特别感谢中南财经大学副校长王寿安教授、科研处处长帅重庆教授。在本书的编写过程

中，自始至终得到了他们的指导和帮助。他们为本书的出版做了大量的组织工作。

由于本人水平有限，加之时间仓促，书中难免不当之处。衷心希望广大读者指正。

编　者

一九九九年四月于武昌

目　　录

第一章　方差分析与多重比较

方差分析是由 R.A.Fisher 在本世纪 20 年代首先提出的，并成功地应用于实际工作中。方差分析是利用对样本观测值的分析，来研究实际问题中各个不同因素的变化对问题结果的影响。它是一种非常重要的数理统计方法。

§1　问题的提出

在实际中，许多管理问题都与相互制约，相互依存的不同因素联系着。人们需要知道这些因素对问题结果的影响是否存在着显著的差异。例如在对若干所院校的学生进行统一的标准考试中，人们感兴趣的问题之一是不同院校的学生智力商数是否存在着显著的差异；在经营管理中，人们希望了解商品的不同位置摆放是否对商品的销售量有明显的差异；在产品生产中，为了节省原材料，选用若干个具有代表性的操作工人分别在同等条件下，用三种不同的操作法进行操作，人们关心的是工人的技术及操作法对节省原材料是否有显著的影响等等。诸如此类的问题还可举出很多。这些问题都可归纳为有关多个总体均值的比较问题。用数理统计的形式将其表述出来就是检验假设。

$$H_0:\mu_1=\mu_2=\cdots=\mu_R \qquad (1-1-1)$$

其中 μ_i 为第 i 个总体的均值

检验有关总体均值的假设我们曾经学过，那是对两个总体的均值进行比较。如果对于假设(1－1－1)我们采用这种检验方法，两个两个的总体均值进行比较或检验，这将会明显地存在着如下的缺陷：第一，两个总体均值进行比较，每次检验所利用的数据仅

是被比较的两个总体的样本资料信息，而不是利用的所有被研究的总体的样本资料所提供的信息，有些信息资料没有得到充分地利用。而且即使在足够一致的总体群中，几乎经常会发现至少存在一对总体均值在假设检验中被认为存在着差异。这种现象往往是由被比较总体的样本所引起。第二，若被比较的总体数目较大，计算量将会很大。n 个总体的均值进行两两比较，将要作 $\frac{n(n-1)}{2}$ 次的假设检验，例如 10 个总体均值进行两两比较，需要进行 45 次假设检验，这意味着要计算 45 个相关的统计量和 45 个相关的临界值。如果总体数目再增多一些，计算量将是非常可观。可见，涉及检验多个总体均值的假设采用两两进行的比较是不适宜的。这就提出了一个问题，如何充分地利用样本数据资料所提供的信息，作出合理的推断，检验假设(1－1－1)。

方差分析方法用于解决多总体均值比较就能克服上述的问题。它是将全部的总体均值同时进行比较，以推断它们之间是否存在着显著的差异，其计算量也较少。再者方差分析方法是将所有样本资料集中一起来使用，这样也增加了数据资料的稳定性。

方差分析研究某因素对试验或观测的结果是否存在显著的影响，是通过研究该因素的变化对试验或观测结果是否有显著影响来推断的。

例 1.1.1　有三条生产线生产同一种型号的产品，对每一条生产线各观测其 5 天的日产量，得数据如下表(表 1－1－1)

表 1－1－1

产量 / 设备 \ 日期	1	2	3	4	5
生产线Ⅰ	57	41	41	49	48
生产线Ⅱ	64	65	54	72	57
生产线Ⅲ	48	45	56	48	51

问不同的生产线的日产量是否有显著的差异?

从表上数据资料可看到,每一行为同一生产线不同日期的抽样结果,每一列为不同生产线的日产量结果。生产线Ⅰ、Ⅱ、Ⅲ日产量平均值分别为47.2、62.4、49.6。显然存在着差别。产生这个差别明显有两个可能原因:一是由抽样所致;一是由不同的生产线所致。那么这种差别是由哪个原因所引起的呢?这正是我们想要了解的。如果主要由前者所致,则不能认为三条生产线的日产量有明显的差异;如果主要由后者所致,则可以认为三条生产线的日产量有明显差异。方差分析解决这些问题的具体做法就是研究哪种原因是主要的。它是将全部的样本数据的总离差平方和分解为两部分,一部分是由抽样引起,一部分是由不同的生产线所引起。分析哪一部分占主要的,从而在数量上得出统计推断。

现在我们来介绍方差分析中的基本概念与进行方差分析的前提条件。

一、基本概念

因素、因素水平

所谓因素是指对试验或观测结果发生影响和作用的分组变量。而根据因素或分组变量的变化分成的等级或组别称为因素的水平。例如我们在前面的介绍的例子中,生产设备是因素,生产线Ⅰ、生产线Ⅱ、生产线Ⅲ是因素水平。

固定因素、随机因素

随机因素是随机变量,该因素的因素水平是来自于由许多因素水平所构成的总体的随机样本。例如在考虑原材料节省问题中,为分析工人的技术因素的影响,我们从所有工人中抽选一部分工人作为代表,这时技术因素便是随机因素。固定因素是指它的因素水平不再是随机样本而是确定的。在方差分析中,根据因素是固定的还是随机的,可分为固定效应的方差分析和随机效应的方差分析。

单因素方差分析、双因素方差分析

在进行方差分析时，如果只研究一个因素对试验结果的影响，而其它条件不变，称为单因素方差分析，如果研究的是两个因素对试验结果的影响，其它条件不变，称为双因素方差分析。

二、方差分析的前提条件

方差分析的假设前提条件是

1. 样本是相互独立的随机样本；

2. 各样本皆来自于正态总体 $N(\mu_i、\sigma_i^2)$

其中 $\mu_i、\sigma_i^2$ 为未知参数；

3. 总体方差具有齐性，即 $\sigma_1^2=\sigma_2^2=\cdots=\sigma_k^2=\sigma^2$

在下面的讨论中，如果没有特殊声明，我们总是假定条件是被满足的。

§2 单因素方差分析

一、固定效应的单因素方差分析

设因素 A 有 R 个水平(或总体) A_i，它服从正态分布 $N(\mu_i, \sigma^2)$ $i=1,2,\cdots k$，现从总体 $A_1, A_2, \cdots A_R$ 中分别抽取容量为 $n_1, n_2 \cdots n_k$ 的样本。记

X_{ij} 为总体 A_i 中第 j 次抽取的样本观测值，$j=1,2\cdots n_i$。将其数据资料归于表 1－2－1。

表 1－2－1

总体	样本容量	样本观测值			
A_1	n_1	x_{11}	x_{12}	$\cdots$	x_{1n1}
A_2	n_2	x_{21}	x_{22}	$\cdots$	x_{2n2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
A_k	n_k	x_{k1}	x_{k2}	$\cdots$	x_{knk}

则固定效应的单因素方差分析模型可表为：

$$x_{ij} = \mu_i + \varepsilon_{ij} \qquad (1-2-1)$$

$$i = 1, 2\cdots\cdots k, \quad j = 1, 2\cdots\cdots n_k$$

其中 μ_i 为总体 A_i 的均值，ε_{ij} 是随机误差项，它服从正态分布 $N(0, \sigma^2)$

若我们又记

$$n = \sum_{i=1}^{k} n_j$$

$$\mu = \frac{1}{n}\sum_{i=1}^{k} n_i\mu_i$$

$$\mu_i = \mu + a_i$$

则模型(1－2－1)又可表为

$$x_{ij} = \mu + a_i + \varepsilon_{ij} \qquad (1-2-2)$$

其中 $\sum_{i=1}^{k} n_i a_i = 0$, $\quad \varepsilon_{ij} \sim N(0, \sigma^2)$

固定效应的单因素方差分析的任务就是要检验假设

$$H_0: \mu_1 = \mu_2 = \cdots\cdots = \mu_{kj} = \mu \qquad (1-2-3)$$

H_1：至少存在 $i \neq j$ 有 $\mu_i \neq \mu_j$ $\qquad (1-2-3')$

或者

$$H_0: a_1 = a_2 = \cdots\cdots = a_k = 0 \qquad (1-2-4)$$

H_1：至少存在一个 i 有 $a_i \neq 0$ $\qquad (1-2-4')$

为检验假设(1－2－3)或(1－2－4)我们又记

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \qquad i = 1, 2\cdots\cdots k$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} \qquad x_{ij} = \frac{1}{n}\sum_{i=1}^{k} n_i\bar{x}_i$$

由于模型中的参数 μ_i 未知，我们用 μ_i 的最优无偏估计 $\bar{x}_j$ 去估计它，并考虑总离差平方和

$$S_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}[(x_{ij}-\bar{x}_i)+(\bar{x}_i-\bar{x})]^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(\bar{x}_i-\bar{x}) + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_i-\bar{x})^2$$

因为 $$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(\bar{x}_i-\bar{x}) = \sum_{i=1}^{k}(\bar{x}_i-\bar{x})\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)$$

$$= \sum_{i=1}^{k}(\bar{x}_i-\bar{x})(\sum_{j=1}^{n_i}x_{ij}-n_i\bar{x}_i)$$

$$= 0$$

所以 $$S_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_i-\bar{x})^2$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2 + \sum_{i=1}^{k}n_i(\bar{x}_i-\bar{x})^2$$

记 $$S_E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2$$

$$S_A = \sum_{i=1}^{k}n_i(\bar{x}_i-\bar{x})^2$$

则 $$S_T = S_E + S_A \qquad (1-2-5)$$

显然，S_E 是样本观测值与其平均值的离差平方和。它反映同一总体内样本观测值的波动状况，其大小说明抽样的误差大小。S_A 是不同总体的样本均值与总均值的离差平方和。反映出不同总体之间的差异。

由公式(1－2－5)不难看到，如果 S_E 占的比重过大，则 S_A 占的比重减少，这说明不同总体均值间的差异小；如果 S_E 占的比重小，则 S_A 占的比重大，这说明不同总体均值间的差异存在。因此，我们可通过分析 S_E 与 S_A 占的比重来判断假设(1－2－3)是否成立。事实上，

$$E(S_E) = E[\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2]$$

$$= \sum_{i=1}^{k}E\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2$$

$$= \sum_{j=1}^{k}(n_i-1)\sigma^2 = (n-k)\sigma^2$$

$$E(S_A)=E[\sum_{i=1}^{k}n_i(\bar{x}_i-\bar{x})^2]$$

$$=\sum_{i=1}^{k}n_iE(\bar{x}_i-\bar{x})^2$$

$$=\sum_{i=1}^{k}n_i[D(\bar{x}_i-\bar{x})-E^2(\bar{x}_i-\bar{x})]$$

$$=(k-1)\sigma^2+\sum_{i=1}^{k}n_i(\mu_i-\mu)^2$$

记 $$MS_A=\frac{S_A}{k-1}$$

$$MS_E=\frac{S_E}{n-k}$$

则 MS_E 是 σ^2 的无偏估计，与假设(1－2－3)成立与否无关，而 MS_E 只有在假设(1－2－3)成立时，才是 σ^2 的无偏估计，否则 MS_E 的期望有偏大于 σ^2 的趋势。由此可见

$$F=\frac{MS_A}{MS_E}$$

在假设(1－2－3)不成立时，有偏大的趋势；而假设(1－2－3)成立时，它应在 1 附近波动。因此，我们可利用比值 F 来检验假设(1－2－3)。

可以证明，在假设(1－2－3)成立时，$F=\frac{MS_A}{MS_E}$ 是服从参数为 $k-1, n-k$ 的 F 分布。于是检验假设(1－2－3)的程序应是：

对于给定的显著性水平 α，当 $F>F_\alpha(k-1, n-k)$ 时，就拒绝假设(1－2－3)，当 $F<F_\alpha(k-1, n-k)$ 时，就接受假设(1－2－3)。

将检验假设(1－2－3)的程序可归于如下所示的方差分析表

方差分析表

方差来源	离差平方和	自由度	均方	F 值
因素 A	S_A	$k-1$	MS_A	MS_A/MS_E
误　差	S_E	$n-k$	MS_E	
总　和	S_T	$n-1$		

为了简化计算，可将所有的样本数据同加(或减)一个常数，不影响 S_A, S_E 的数值。或同乘(或除)一个常数，不影响比值 F 的数值。具体计算可通过表 1－2－2 来实现。

表 1－2－2

总体	样本容量	样本观测值				$\sum$	$(\sum)^2$	$(\sum)^2/n_i$	$\sum^2$
A_1	n_1	x_{11}	x_{12}	$\cdots$	x_{1n_1}	$\sum x_{1j}$	$(\sum x_{1j})^2$	$(\sum x_{1j})^2/n_1$	$\sum x_{1j}^2$
A_2	n_2	x_{21}	x_{22}	$\cdots$	x_{2n_2}	$\sum x_{2j}$	$(\sum x_{2j})^2$	$(\sum x_{2j})^2/n_2$	$\sum x_{2j}^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_k	n_k	x_{k1}	x_{k2}	$\cdots$	x_{kn_k}	$\sum x_{kj}$	$(\sum x_{kj})^2$	$(\sum x_{kj})^2/n_k$	$\sum x_{kj}^2$
$\sum$	$\sum n_i$					$\sum\sum x_{ij}$		$\sum(\sum x_{ij})^2/n_i$	$\sum\sum x_{ij}^2$

记 $$Q=\sum(\sum x_{ij})^2/n_i,$$

$$P=\frac{1}{n}(\sum\sum x_{ij})^2$$

$$R=\sum\sum x_{ij}^2$$

则 $$S_A=\sum_{i=1}^{k}n_i(\bar{x}_i-\bar{x})^2$$

$$=\sum_{i=1}^{k}(\sum_{j=1}^{n_i}x_{ij})^2/n_i-\frac{1}{n}(\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij})^2$$

$$=Q-P$$

$$S_E=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}^2-\sum_{i=1}^{k}(\sum_{j=1}^{n_i}x_{ij})^2/n_i$$

$$=R-Q$$

$$S_T=S_E+S_A$$

例 1.2.1　假设例 1.1.1 中的问题满足方差分析的所有条件。对给定的显著性水平 $\alpha=0.05$，作方差分析。

解　将各样本数据同减去 41，并按表 1－2－2 计算得表 1－2－3

表 1-2-3

产量/日期 / 设备	1	2	3	4	5	n_i	Σ	$(\Sigma)^2$	$(\Sigma)^2/n_i$	Σ^2
生产线 I	16	0	0	8	7	5	31	961	192.2	369
生产线 II	23	24	13	31	16	5	107	11449	2289.8	2491
生产线 III	7	4	15	7	10	5	43	1849	369.8	439
					Σ	15	181		2851.8	3299

$$Q = 2851.8 \qquad P = \frac{1}{15}(181)^2 = \frac{32761}{15} = 2184.1$$

$$R = 3299$$

$$S_A = Q - P = 2851.8 - 2184.1 = 667.7$$

$$S_E = R - Q = 3299 - 2851.8 = 447.2$$

$$S_T = S_A + S_E = 1114.9$$

于是得方差分析表(表 1-2-4)

表 1-2-4

方差来源	离差平方和	自由度	均方	F 值
因　素	667.7	2	333.85	8.96
误　差	447.2	12	37.27	
总　和	1114.9	14		

检验假设 $H_0: \mu_1 = \mu_2 = \mu_3$; $H_1: \mu_i$ 中至少有一对不相等

对于给定的显著性水平 $\alpha = 0.05$, 查 F 分布临界值表得临界值 $F_{0.05}(2, 12) = 3.89$, 显然, F 值大于 3.89, 故可以认为不同的生产线的日产量有显著的差异。

二、随机效应的单因素方差分析

在许多实际工作中, 由于各种原因, 使我们无法对研究的因素的所有水平进行考察, 只能从其中选择一部分因素水平进行试验, 来推断某因素的不同水平对试验结果的影响是否存在着显著的差