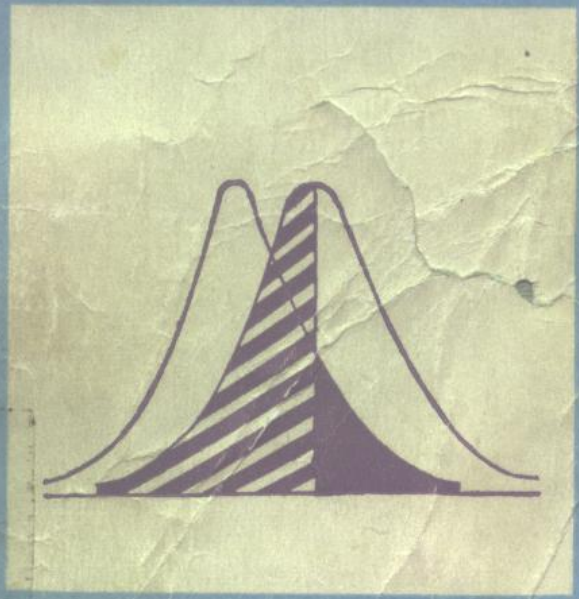
卫生统计应用丛书

医用多元分析

史秉璋　杨　琦　编著



人民卫生出版社

卫生统计应用丛书

医用多元分析

史秉璋　杨　琦　编著

田凤调　审

人民卫生出版社

医用多元分析

史秉璋　杨　琦　编著

人民卫生出版社出版

登记证号：（京）081号

（北京市崇文区天坛西里10号）

北京市卫顺排版厂印刷

新华书店北京发行所发行

787×1092毫米32开本　8印张　189千字

1990年5月第1版　1990年5月第1版第1次印刷

印数：00,001—10,000

ISBN 7-117-01277-3/R·1278　定价：2.65元

〔科技新书目214—230〕

卫生统计应用丛书编写说明

为了提高我国卫生统计知识水平，促进卫生统计工作的发展，更好地适应我国四化建设的需要，经过较长时间的同行酝酿，并经与人民卫生出版社协商，决定编写出版这套丛书。

本套丛书以介绍卫生统计基础知识、基本方法为主，注意实用性、科学性。既照顾到读者实际接受的可能性，又要求反映出时代的特点，介绍新的内容。

主要读者对象是：卫生统计专业工作者；医务人员和卫生防疫人员；医学界有关专业的科研与教学工作者；也可作为医学院校学生与研究生的参考书。

卫生统计应用丛书的选题包括下列几个方面：医学统计方法，居民健康统计，卫生资源统计，卫生业务统计，计算机应用技术，卫生统计工作改革等。每册一般为10～15万字，分批出版。

卫生统计应用丛书编委会

1988年12月

前　言

多元统计分析简称多元分析或多变量分析，医学中也常称之为多因素分析，它是研究客观事物中多种因素间相互依赖的统计规律性的一个数理统计学分支。因为客观世界中的任何事物实际上都是多种因素影响着它的形成、变化和发展，而且各种因素间又存在着广泛而又错综复杂的联系，所以多元分析有着广泛的应用意义。医学现象尤其复杂多变，例如疾病的发生，病情的变化就往往受到多种因素的支配，各种病因之间也常存在着千丝万缕的内在联系和相互制约，多元分析在研究这些问题时有着广泛的用武之地。

由于多元分析的计算比较复杂，运算工作量很大，常规计算工具很难胜任。因此，直到近三十年来，在电子计算机的使用日益普遍后，才使它的理论和应用得到了迅速的发展。

多元分析的应用面很广，本书对目前在医学中已得到一定应用的若干方法，如Hotelling T^2检验、多元回归及相关分析、逐步回归分析、判别分析、聚类分析、主成分分析、典型相关分析等作一概述，对比较成熟、应用也比较广泛的，如多元回归、多元相关、逐步回归、判别分析等，讨论稍详细些；对应用还不十分普遍的，则只作简略叙述。

考虑到电子计算机的应用已日益广泛，不少多元分析方法的计算程序已陆续问世，有的甚至已做成软件包，应用十分方便，一般不再需要每个读者自己动手编写这类程序，故本书中不专门列出用算法语言编写的计算程序，需要时可参

阅有关参考书。同时我们还考虑到有些读者只能依靠计算器来进行必要计算，故对每一方法的计算步骤都作了深入浅出的叙述，并附有详细的计算实例。需要说明的是：本书例题的计算均由电子计算机进行连续计算而完成，在计算过程中保留了较多的数字位数，而在例题的中间计算过程中，一般只列出4～6位有效数字，故当用计算器逐步运算时，所得结果与书上数字可能略有出入。

因本书的读者对象为具有一般医学统计基础的医学院校师生，卫生防疫及医学科研工作者，因此编写时力求叙述通俗，注重实用，避免繁琐的数学推导，以使广大医学工作者能有所借鉴。

著　者

目　录

第一章 Hotelling T^2 检验

§1.1 Student t 检验的回顾

读者熟知，单变量计量资料的假设检验，一般用 Student t 检验法，通常有两类问题：

1．检验一样本是否来自某已知总体 设有某正态总体 $N(\mu_0,\sigma^2)$，现手头有一个大小为 n 的样本，其样本均数和标准差分别为 $\bar{X}$ 及 s，$\bar{X}$ 是总体均数 μ 的估计值，问此样本是否来自均数为 μ_0 的总体。为此，先作假设 H_0：$\mu=\mu_0$，然后求统计量

$$t=\sqrt{n}\frac{|\bar{X}-\mu_0|}{s} \qquad (1.1)$$

它遵从自由度为 $n-1$ 的 t 分布。如果规定检验水准为 α，则：

(1) 如 $t\geqslant t_\alpha$，则在 α 水准上拒绝 H_0，接受其对立假设 H_1，(即 $\mu\neq\mu_0$)，认为此样本均数与总体均数有显著差别，不大可能来自此总体；

(2) 如 $t<t_\alpha$，则在 α 水准上不拒绝 H_0，认为样本均数与总体均数差别不显著，有可能来自此总体。

上述 t_α 是在 α 水准下由 t 分布决定的值，可由 t 值表查得。

由于在 $\nu_1=1$ 时，

$$F=t^2 \qquad (1.2)$$

故 t 检验也可查 F 值表以判断其显著性。此时，$\nu_1=1$，

$\nu_2=n-1$。

2．检验两样本是否来自同一总体，即两样本所代表的总体均数有否显著差别　设两样本来自两个具有公共方差的总体 $N(\mu_1,\sigma^2)$ 及 $N(\mu_2,\sigma^2)$，显然，此时的检验假设 H_0 为 $\mu_1=\mu_2$。设两样本容量分别为 n_1 及 n_2，先分别算出两样本的均数 $\bar{X}_1$，$\bar{X}_2$ 及标准差 s_1，s_2，再求 t 值

$$t=\sqrt{\frac{n_1n_2}{n_1+n_2}}\cdot\frac{|\bar{X}_1-\bar{X}_2|}{s} \qquad (1.3)$$

式中

$$s=\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} \qquad (1.4)$$

此 t 值遵从自由度为 n_1+n_2-2 的 t 分布。同样，当 $t\geqslant t_\alpha$ 时，则在 α 水准上拒绝 H_0，而接受其对立假设 H_1（即 $\mu_1\neq\mu_2$），认为两样本均数有显著差别，不大可能来自同一总体；当 $t<t_\alpha$ 时，则在 α 水准上不接受 H_0，认为两样本均数差别不显著，有可能来自同一总体。

此时，t 值和 F 值有下述关系：在 $\nu_1=1$，$\nu_2=n_1+n_2-2$ 时

$$F=t^2 \qquad (1.5)$$

故也可查 F 值表以判断其显著性。

§1.2　Hotelling T^2 检验的基本概念

在很多医学问题中，当作以上假设检验时（如检验两样本是否来自同一总体）所依据的指标可能不止一个。例如，当比较两组风湿性和类风湿性关节炎患者的病情程度时，就不能只用一个指标，如果采用血沉（X_1），抗“O”（X_2），白细胞计数（X_3）三个指标，则资料就呈如下格式：

	编号	血沉(X_1)	抗"O"(X_2)	白细胞计数(X_3)
A组	1			
	2	…	…	…
	3			
	⋮	⋮	⋮	⋮
	n_1	…	…	…
B组	1			
	2	…	…	…
	3			
	⋮	⋮	⋮	⋮
	n_2	…	…	…

这三项指标都是测得值越高病情越重。如果采用通常的 t 检验法对每个指标作检验，则只有出现下列情况之一时，才能作出明确判断：

(1) 两组间 X_1，X_2，X_3 均有显著差别，且大小趋势一致（例如A组各指标均数都大于B组对应指标的均数）。

(2) 两组间各指标均无显著差别。

反之，如果出现下列情况之一，就难以得出明确结论：

(1) 两组的各指标间虽有显著差别，但趋势不一（如A组的$\overline{X}_1$，$\overline{X}_2$ 大于B组的$\overline{X}_1$，$\overline{X}_2$，但A组的$\overline{X}_3$却小于B组的$\overline{X}_3$）。

(2) 两组间有些指标有显著差别（趋势一致或不一致），有些却无显著差别。

在这些情况下，常规的单指标 Student t 检验法就无能为力了。为此，引进多变量的 Hotelling T^2 检验，它在很多方面都与 t 检验有类似之处。事实上，我们也可把 Student

t 检验看作是 Hotelling T^2 检验中变量数为 1 时的特例。它通常也用于处理下述两类问题。

§1.3 检验一样本是否来自均向量为 μ_0的 p 元正态总体$N(\mu_0, \Sigma)$

其中的 μ_0 是形如 $\begin{pmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_p\end{pmatrix}$ 的列向量，Σ为 p 个变量间的协方差阵 $\begin{pmatrix}V_{11}V_{12}\cdots V_{1p}\\ V_{21}V_{22}\cdots V_{2p}\\ \cdots\cdots\\ V_{p1}V_{p2}\cdots V_{pp}\end{pmatrix}$。

设手头有一个大小为n的样本，各观测了 p 个变量的值，其数据格式为

观察单位	变量				
	1	2	3	…	p
1		…	…	…	…
2		…	…	…	…
3		…	…	…	…
4		…	…	…	…
⋮		⋮	⋮	⋮	⋮
n		…	…	…	…
均值$\bar{X}$	$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	…	$\bar{X}_p$

则在算出样本的均向量$\bar{X}$(它所估计的总体均向量可记为μ_0)及协方差阵V后，即可作假设H_0：$\mu=\mu_0$，并计算统计量

$$T^2=n\,[\bar{X}-\mu_0]'V^{-1}[\bar{X}-\mu_0] \qquad (1.6)$$

式中 $[\bar{X}-\mu_0]'$ 是列向量 $[\bar{X}-\mu_0]$ 的转置，V^{-1} 是 V 的逆矩阵，此 T^2 遵从自由度为 p 的 χ^2 分布，并且有：

(1) 如 $T^2\geqslant\chi^2_{\alpha(p)}$，则在 α 水准上拒绝 H_0，接受 H_1，即认为此样本不大可能来自总体 $N(\mu_0,\Sigma)$；

(2) 如 $T^2<\chi^2_{\alpha(p)}$，则在 α 水准上不拒绝 H_0，即认为此样本有可能来自总体 $N(\mu_0,\Sigma)$。

同样，T^2 与 F 间存在一定关系，故也可由 T^2 再算出 F

$$F=\frac{n-p+1}{np}T^2 \qquad (1.7)$$

然后据 $\nu_1=p$，$\nu_2=n-p+1$ 的 F 分布作推断。

例 1.1　用胸腺素治疗 15 例病毒性心肌炎细胞免疫功能低下症，得结果如下：

表1.1　胸腺素治疗前后免疫球蛋白测得值

例号	IgG(X_1)			IgA(X_2)			IgM(X_3)		
	治前	治后	差数	治前	治后	差数	治前	治后	差数
1	1810	1654	−156	246	196	−50	292	243	−49
2	1744	1568	−176	213	208	−5	286	272	−14
3	1806	1743	−63	226	214	−12	297	276	−21
4	1712	1584	−128	238	168	−70	265	274	9
5	1642	1649	7	227	242	15	307	289	−18
6	1685	1543	−142	260	198	−62	246	265	19
7	1728	1624	−104	138	212	74	312	288	−24
8	1695	1500	−195	196	207	11	266	262	−4
9	1760	1340	−420	233	179	−54	243	259	16
10	1690	1454	−236	256	196	−60	334	296	−38
11	1667	1453	−214	297	209	−88	285	263	−22
12	1703	1564	−139	212	223	11	296	274	−22
13	1715	1644	−71	228	237	9	249	260	11
14	1699	1543	−156	236	205	−31	266	262	−4
15	1733	1684	−49	202	197	−5	308	288	−20

问如何评价总的治疗效果？

解：如果分别对 IgG(X_1)，IgA(X_2)，IgM(X_3) 进行配对 t 检验，则可得下列结果：

对 IgG，$\Sigma X_1=-2242$，$\bar{X}_1=-149.4667$，$s_1=99.5008$，$s_{\bar{X}_1}=25.691$，

$$t=\frac{|\bar{X}_1|}{s_{\bar{X}_1}}=\frac{|-149.4667|}{25.691}=5.8179$$

查表，$t_{0.001(14)}=4.14$

现　$t>t_{0.001(14)}$，$\therefore P<0.001$。

对 IgA，$\Sigma X_2=-320$，$\bar{X}_2=-21.1333$，$s_2=43.0811$，$s_{\bar{X}_2}=11.1235$，

$$t=\frac{|\bar{X}_2|}{s_{\bar{X}_2}}=\frac{|-21.13331|}{11.1235}=1.8999$$

查表，$t_{0.05(14)}=2.145$。

现　$t<t_{0.05(14)}$，$\therefore P>0.05$。

对 IgM，$\Sigma X_3=-181$，$\bar{X}_3=-12.0667$，$s_3=19.6704$，$s_{\bar{X}_3}=5.0789$，

$$t=\frac{|\bar{X}_3|}{s_{\bar{X}_3}}=\frac{|-12.0667|}{5.0789}=2.3759$$

$\because t>t_{0.05(14)}$，$\therefore P<0.05$。

以上三个指标中，有二个（即X_1与X_3）的差数均数与零有显著差别，意味着治疗后 IgG 与 IgM 都有显著下降，但 X_2 的差数均数却与零没有显著差别，也即治疗后 IgA 的下降不明显。这样，总的来说，胸腺素究竟有否改善免疫功能的作用就很难得出确切评价。

例1.2对例1.1资料，应用式(1.6)进行Hotelling T^2检验。

解：

$$(\bar{X}-\mu_0)=\begin{pmatrix}-149.4667-0\\-21.1333-0\\-12.0667-0\end{pmatrix}=\begin{pmatrix}-149.4667\\-21.1333\\-12.0667\end{pmatrix}$$

计算各变量的离均差平方和，分别记为 l_{11}，l_{22}，l_{33}，以及两变量之间的离均差积和，分别记为 l_{12}，l_{13}，l_{23}，得

$$l_{11}=138605.73,\quad l_{22}=25983.73,\quad l_{33}=5416.93$$

$$l_{12}=28668.07,\quad l_{13}=5485.47,\quad l_{23}=1518.13$$

分别除以14，即得协方差阵V的各元素

$$V=\begin{pmatrix}9900.4095 & 2047.7190 & -391.8190\\2047.7190 & 1855.9810 & -108.4381\\-391.8190 & -108.4381 & 386.9238\end{pmatrix}$$

求出V阵的逆阵，记为V^{-1}，得

$$V^{-1}=\begin{pmatrix}0.00013429 & -0.00014255 & 0.000096038\\-0.00014255 & 0.00069909 & 0.000051568\\0.000096038 & 0.000051568 & 0.002696200\end{pmatrix}$$

由式（1.6）

$$T^2=n(\bar{X}-\mu_0)'V^{-1}(\bar{X}-\mu_0)$$

$$=15(-149.4667\ -21.1333\ -12.0667)$$

$$\begin{pmatrix}0.00013429 & -0.00014255 & 0.000096038\\-0.00014255 & 0.00069909 & 0.000051568\\0.000096038 & 0.000051568 & 0.002696200\end{pmatrix}\begin{pmatrix}-149.4667\\-21.1333\\-12.0667\end{pmatrix}$$

$$=47.6555$$

现变量数 $p=3$，$\chi^2_{0.001(3)}=16.266$

$\because T^2>\chi^2_{0.001(3)}$，$\therefore P<0.001$。

故得结论为：如把 IgG，IgA，IgM 三个指标综合起来分析，则可认为用胸腺素治疗前后免疫功能有显著变化。

§1.4 检验两样本是否来自同一多元正态总体

检验两样本是否来自同一多元正态总体，即检验两样本所代表的两总体均向量间有无显著差别。

设有容量分别为 n_1，n_2 的两个样本，它们来自两个具有公共协方差 Σ 的 p 维正态总体 $N(\mu_A, \Sigma)$ 和 $N(\mu_B, \Sigma)$，并作假设 H_0：$\mu_A = \mu_B$。

数据格式为

编号		变量			
	1	2	3	…	p
样本A	1	…	…	…	…
	2	…	…	…	…
	3	…	…	…	…
	⋮	⋮	⋮	⋮	⋮
	n_A	…	…	…	…
样本B	1	…	…	…	…
	2	…	…	…	…
	3	…	…	…	…
	⋮	⋮	⋮	⋮	⋮
	n_B	…	…	…	…

分别算出两样本的均向量 $\bar{X}_A$，$\bar{X}_B$ 及合并协方差矩阵 V，则统计量