非定量数据分析及其应用

周光亚 夏立显 编著

科学出版社

非定量数据分析及其应用

周光亚　夏立显　编著

科　学　出　版　社

1993

（京）新登字 092 号

内 容 简 介

由于非定量数据的广泛存在，至今已发展了种种不同的适用于这种数据的分析方法，并已应用在地质、气象、林业、医学、企业管理、产品设计等方面.

本书主要介绍非定量数据的分析方法，其中包括林知己夫建立的数量化方法和克拉斯卡尔等人发展的多维标度法. 本书着重介绍方法的实际意义、计算公式和基本计算步骤，而不追求严格的数学推导.

本书可供大专院校数理统计专业师生、从事实际工作的科技人员阅读.

非定量数据分析及其应用

周光亚 夏立显 编著

责任编辑 毕 颖

[illegible] 出版

[illegible] 19 号

邮政编码：100707

昌平马池口印刷厂印刷

新华书店北京发行所发行 各地新华书店经售

*

1993年8月第 一 版 开本：787×1092 1/32

1993年8月第一次印刷 印张：9 1/2

印数：1— 500 字数：207 000

ISBN 7-03-003247-0/O·587

定价：8.10 元

前　言

由于非定量数据的广泛存在，至今已经发展了种种不同的适用于这种数据的分析方法，并已应用于许多领域，获得了较好的效果．董文泉和我们合写的《数量化理论及其应用》（吉林人民出版社，1979）一书，曾系统地介绍了日本林知己夫教授的数量化理论Ⅰ，Ⅱ，Ⅲ，Ⅳ，其中数量化理论Ⅰ，Ⅱ，Ⅲ分别对应于回归分析、判别分析和对应因子分析．这些方法的特点是非定量数据的定量转化及其分析同时进行，应用起来比较直接．而数量化理论Ⅳ则与前三种不同，它是根据多个被研究对象两两间相比较所得到的种种（非）相似性度量，合理地将各对象表示为欧氏空间中的点，以实现对各对象的定量表示，为进一步分析各对象间的种种关系提供定量依据．除此以外，还有许多种这样处理问题的方法，它们统称为多维标度法或多维尺度构成法(Multidimensional Scaling)，简称为MDS．林知己夫后期的工作以及克拉斯卡尔（Kruskal），加特曼（Gattman）等许多人的工作均涉及到了这种方法．由于这种方法仅依赖较少的信息便可使用，所以它有着广泛的应用．例如在地质、社会和心理现象这些难以取得准确的定量数据的领域中，尤其具有其特殊的功效．

在1983年第四届全国多元分析会议中，我们曾和其他作者一起提交了一个文集——多维标度法及其在地质学中的应用，以期将此方法介绍给有兴趣的同行．此后，在长春地质学

院数学地质研究室的几届研究生中多次讲授了此课，并结合毕业论文作了若干工作，取得了好的效果。中国科学院应用数学研究所方开泰研究员一直关心和支持我们在这方面的工作，在他的热情推动下，我们又收集了一些有关文献，尽可能吸取各家之长，以完成此书。我们对某些结果作了改进，并把我们的同事、研究生和我们自己的部分成果也吸收到书中。

为使非数学专业出身的科技工作者能够阅读本书，并将书中介绍的方法用于自己的工作中，我们在讲法上力求通俗易懂。在本书的正文中，我们着重介绍方法的实际意义、计算公式和基本计算步骤，而不追求严格的数学推导。至于对数学严格性证明有兴趣的读者，可参看书后所附的参考文献。

对于书中的一些不妥甚至错误之处，敬请读者批评指正。如果这本书能起到"抛砖引玉"的作用，我们将会感到非常欣慰。

本书是在长春地质学院和吉林大学许多同志的热情支持下完成的。景毅教授始终给予了热情鼓励和关怀，王世称教授对这些方法的地质应用给了具体和有益的指导，董文泉教授对本书的编写给了不少具体的帮助。此外，赵文、赵振全、王锦功、姜诗章、范继璋、成秋明、林娜、安龙哲、刘玉芦、许雅明等同志都在不同方面提供了帮助。杨毅恒同志对书中所涉及的全部方法都编制了计算机程序，并用于计算许多实际问题。本书中大量采用了他的计算结果。

曹鸿兴副研究员仔细审阅了本书原稿，提出了许多宝贵的意见。

数量化理论的创始人，日本著名统计学家林知己夫教授和田中豊教授对本书的写作一直非常关注。他们在与作者多年的交往中提出了许多建设性的意见，并提供了大量的有关

资料。

作者对上面提到的诸位表示衷心的感谢！

吉 林 大 学　周光亚
长春地质学院　夏立显

目　录

第一章　相似性和非相似性度量

作为各种MDS方法的前提和基础，首先是建立研究对象之间的某种相似性或非相似性度量。这种度量指标选用的是否适当，直接关系到MDS方法应用效果的好坏。关于相似性和非相似性度量各种定义方法的讨论虽然并非MDS方法本身的内容，但它在研究MDS方法的理论和应用当中，却是不可缺少的一步。我们将在本章专门介绍这方面的内容。

作为与定量变量（包括比例尺度和间隔尺度的变量）有关的样品之间的非相似性度量，可以采用明科夫斯基(Minkowski)距离，它是普通欧氏距离的推广。有关的相似性度量，有人们熟知的相似性系数和相关系数等。我们这里不再涉及上述内容，仅介绍与定性变量（包括名义尺度和有序尺度变量）有关的一些相似性度量指标的定义方法。

定性变量之间的相似性度量通常称为"关联系数"，以便与相似性系数和相关系数有所区别。关联性一词来自英文association，它的含意很不确切，人们对它的理解有宽有窄。我们认为这类指标主要有以下两个特点：（1）它不单刻画变量间线性关系的强弱，而着重反映其它的非线性关系的强弱；（2）它主要用于反映定性变量之间的相似性。

为了研究两个定性变量之间的关联性，常常取一个容量为n的样本。这些样本观测值的记录方法有许多种。例如，在林知己夫的数量化理论中，采用项目(item)、类目(category）反映表对观测结果作最详细的记载[8]，但我们这里

介绍的大部分方法都是先把观测结果整理成列联表的形式，然后再由此出发讨论变量之间或样品之间的关联性度量。

本章第一节讨论名义尺度变量间的关联性。这里随机变量X，Y的不同状态x_i，y_i通常用不同的数字来表示，但它们只是作为不同的符号代表不同的状态，这些状态之间没有大小或先后顺序的关系。把各种状态交换顺序相当于把列联表中各行或各列交换顺序，对于这种换序，这一节所讨论的关联性度量都不发生变化。

第三节讨论有序尺度变量间的关联性。有序尺度变量的各种状态有一定的顺序关系。若用不同的数值表示各种状态的话，数值的大小仅表示状态间的顺序，而不表示状态间差异的大小。把各种状态交换顺序时，也就是把列联表中的各行或各列交换位置时。这一节里所介绍的关联性度量一般都要发生变化。

对于二态变量，它当然既可以是名义尺度变量，也可以是有序尺度变量，有时甚至可以理解为定量变量。因此，对于各种变量所引进的关联性度量，也都适用于二态变量。但是，对于二态变量还有些独特的关联性度量指标，对此将在第二节进行讨论。

本章第四节讨论样品之间的关联性度量，其中许多指标是由定量变量间相似性度量指标转化而来的。

第一节　名义尺度变量间的关联性

一、关于列联表的一些定义和记号

设定性变量X所有可能的状态为x_1，x_2，$\cdots$，x_r，定性变量Y所有可能的状态为y_1，y_2，$\cdots$，y_c。二维随机变量（X，

Y）的联合分布为

$$P(X=x_i,\ Y=y_j)\triangleq\pi_{ij},\ i=1,\ 2,\cdots,r;\ j=1,2,\cdots,c.$$

由此可以得到X和Y各自的边缘分布为

$$P(X=x_i)=\sum_{j=1}^{c}\pi_{ij}\triangleq\pi_{i.},\ i=1,2,\cdots,r,$$

$$P(Y=y_j)=\sum_{i=1}^{r}\pi_{ij}\triangleq\pi_{.j},\ j=1,\ 2,\cdots,c,$$

条件分布为

$$P(Y=y_j|X=x_i)=\pi_{ij}/\pi_{i.}\triangleq\pi_{(j)i},$$

$$P(X=x_i|Y=y_j)=\pi_{ij}/\pi_{.j}\triangleq\pi_{i(j)},$$

$$i=1,\ 2,\cdots,r;\ j=1,\ 2,\cdots,c.$$

这些概率可归纳在表1.1中。

表 1.1　(X, Y)的概率分布表

X \ Y	y_1	y_2	$\cdots$	y_c	Σ
x_1	π_{11}	π_{12}	$\cdots$	π_{1c}	$\pi_{1.}$
x_2	π_{21}	π_{22}	$\cdots$	π_{2c}	$\pi_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_r	π_{r1}	π_{r2}	$\cdots$	π_{rc}	$\pi_{r.}$
Σ	$\pi_{.1}$	$\pi_{.2}$	$\cdots$	$\pi_{.c}$	1

根据X和Y的各个状态，可以对n个样品按两种不同的方式进行分类，两种分法的不同组合可以把n个样品交叉分成rc类。用n_{ij}表示使$X=x_i$，$Y=y_j$的样品数，它们可以列成表1.2的形式，这个表就叫做列联表。

表 1.2 $r\times c$列联表

X \ Y	y_1	y_2	$\cdots$	y_c	Σ
x_1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r\cdot}$
Σ	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot c}$	n

表中最后一列给出n个样品按变量X分类的结果，它们是对应行诸元素之和

$$n_{i\cdot}=\sum_{j=1}^{c}n_{ij},\quad i=1,\ 2,\cdots,r.$$

最后一行给出n个样品按变量Y分类的结果，它们是对应列诸元素之和

$$n_{\cdot j}=\sum_{i=1}^{r}n_{ij},\quad j=1,\ 2,\cdots,c.$$

若把n个样品看作是n次独立随机试验的结果，则n_{ij}就是试验结果落入表中各个网格的频数。而落入各个网格的频率

$$P_{ij}\triangleq n_{ij}/n,\quad i=1,\ 2,\cdots,r;\quad j=1,\ 2,\cdots,c.$$

可以作为相应概率π_{ij}的估计值。

表1.2含有r个行和c个列，叫做$r\times c$列联表。只有一个状态的"变量"实际上是常量，没有讨论价值。具有两个状态的变量，如$r=2$时的X，因其在实际中遇到的较多，而且对

于它们还有些独特的分析方法，特别把它们叫做二态变量，并在后面专列一节对它们进行讨论。以后说到多态变量时，虽然也包括二态情形在内，但其所强调的是多于两个状态的情形。

二、与χ^2统计量有关的关联性度量[20]

我们知道，随机变量X和Y独立的充要条件是

$$\pi_{ij}=\pi_{i.}\pi_{.j},\ i=1,2,\cdots,r;\ j=1,2,\cdots,c.$$

为了衡量这两个变量之间的关联程度，皮尔森（Pearson）引进了均方列联的概念，其表达式如下：

$$\varphi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(\pi_{ij}-\pi_{i.}\pi_{.j})^2}{\pi_{i.}\pi_{.j}}$$

$$=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\pi_{ij}^2}{\pi_{i.}\pi_{.j}}-1. \qquad (1.1)$$

可以证明均方列联有以下几条性质：

（1）不论（X,Y）的概率分布如何，恒有

$$0\leqslant\varphi^2\leqslant q-1,$$

其中$q=\min(r,\ c)$.

（2）$\varphi^2=0$的充要条件是X与Y独立.

（3）$\varphi^2=q-1$的充要条件是X与Y中必有一个以概率1为另一个的函数.

下面只给出（1）和（3）的证明.

（1）的证明：$\varphi^2\geqslant 0$是显然的。由

$$\pi_{ij}\leqslant\pi_{i.},\ \pi_{ij}\leqslant\pi_{.j},\ i=1,\ 2,\cdots,r;\ j=1,\ 2,\cdots,c$$

得到

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\pi_{ij}^2}{\pi_{i.}\pi_{.j}}=\sum_{i=1}^{r}\frac{1}{\pi_{i.}}\sum_{j=1}^{c}\frac{\pi_{ij}^2}{\pi_{.j}}\leqslant\sum_{i=1}^{r}\frac{1}{\pi_{i.}}\sum_{j=1}^{c}\pi_{ij}$$

$$=\sum_{i=1}^{r}\frac{1}{\pi_{i.}}\pi_{i.}=r.$$

类似地还可以得到

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\pi_{ij}^2}{\pi_{i.}\pi_{.j}}=\sum_{j=1}^{c}\frac{1}{\pi_{.j}}\sum_{i=1}^{r}\frac{\pi_{ij}^2}{\pi_{i.}}\leqslant c.$$

从而推出

$$\varphi^2\leqslant q-1.$$

（3）的证明：设Y以概率1为X的函数，则对每个i都对应一个$j(i)$使得

$$P(Y=y_{j(i)}|X=x_i)=1,$$

即有

$$\pi_{ij(i)}/\pi_{i.}=1,$$

于是有

$$\pi_{ij}=\begin{cases}\pi_{i.}, & j=j(i)\\ 0, & j\neq j(i).\end{cases}$$

由此可见，在概率分布表1.1中，每行仅有一个非零元素，整个表内只能有r个非零元素。这时必有$r\geqslant c$，否则将保证不了所有的$\pi_{.j}>0$，$j=1,2,\cdots,c$（注意如果某个$\pi_{.j}=0$，则表明j状态是变量y实际上不能取得的，在研究问题的开始就应该把这个状态删除）。因此

$$\sum_{i}\sum_{j}\frac{\pi_{ij}^2}{\pi_{i.}\pi_{.j}}=\sum_{i}\sum_{j}\frac{\pi_{ij}^2}{\pi_{i.}\pi_{.j}}=\sum_{i}\sum_{j}\frac{\pi_{i.}\pi_{ij}}{\pi_{i.}\pi_{.j}}$$

$$(\pi_{ij}>0)$$

$$=\sum_{j}\frac{1}{\pi_{.j}}\sum_{i}\pi_{ij}=\sum_{j}1=c=q.$$

由此得到

$$\varphi^2=q-1.$$

同理可证，在X为Y的函数时也有$\varphi^2=q-1$.

反之，如Y并非以概率1为X的函数，则必有(i_0, j_0)使

$$0<\frac{\pi_{i_0j_0}}{\pi_{i_0\cdot}}=P(Y=y_{j_0}|X=X_{i_0})<1,$$

即有

$$0<\pi_{i_0j_0}<\pi_{i_0\cdot},$$

从而

$$\sum_i\sum_j\frac{\pi_{ij}^2}{\pi_{i\cdot}\pi_{\cdot j}}<\sum_i\sum_j\frac{\pi_{i\cdot}\pi_{ij}}{\pi_{i\cdot}\pi_{\cdot j}}=\sum_i\sum_j\frac{\pi_{ij}}{\pi_{\cdot j}}=c.$$

同理，当X并非以概率1为Y的函数时，又可推出

$$\sum_i\sum_j\frac{\pi_{ij}^2}{\pi_{i\cdot}\pi_{\cdot j}}<r.$$

总之，当X与Y不能以概率1有函数关系时，必有

$$\sum_i\sum_j\frac{\pi_{ij}^2}{\pi_{i\cdot}\pi_{\cdot j}}<q,$$

即

$$\varphi^2<q-1.$$

将φ^2除以其最大可能值$q-1$，便得到归一化的均方列联：

$$\theta^2\triangleq\frac{\varphi^2}{q-1}=\frac{1}{q-1}\left(\sum_i\sum_j\frac{\pi_{ij}^2}{\pi_{i\cdot}\pi_{\cdot j}}-1\right).\qquad(1.2)$$

在使用上，它比φ^2更为方便．从φ^2的性质立即可以推出θ^2的如下性质：

（1）$0\leqslant\theta^2\leqslant1$.

（2）$\theta^2=0$的充要条件是X与Y独立.

（3）$\theta^2=1$的充要条件是X与Y中必有一个以概率1为

另一个的函数.

上述性质表明，θ^2或φ^2可以作为X与Y间关联程度的度量指标. 它们的值越大，X与Y的关联程度越强，最强时可以达到两者间有某种的函数关系，最弱时两者独立.

将φ^2和θ^2表达式中的概率都用相应的样本估计值代替，得到统计量

$$\hat{\varphi}^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(\frac{n_{ij}}{n}-\frac{n_{i\cdot}n_{\cdot j}}{n^2}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n^2}}$$

$$=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}}-1, \tag{1.3}$$

$$\hat{\theta}^2=\frac{1}{q-1}\left(\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}}-1\right). \tag{1.4}$$

它们分别叫做样本均方列联和归一化的样本均方列联. $\hat{\varphi}^2$和$\hat{\theta}^2$分别与φ^2和θ^2有类似的性质，在实用上可以用它们来衡量两个变量之间的关联性.

我们知道，当X与Y独立时，统计量

$$\chi^2=n\hat{\varphi}^2=n\left(\sum_{i}\sum_{j}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}}-1\right) \tag{1.5}$$

渐近地服从自由度为$(r-1)(c-1)$的χ^2分布[3]. 因此可以用它来检验两个变量间的独立性.

此外，方开泰和潘恩沛[11]还介绍了另外几种与χ^2有关的类似的指标：

$$S_{xy}(1)=\sqrt{\frac{\chi^2}{\chi^2+n}}=\sqrt{\frac{\hat{\varphi}^2}{\hat{\varphi}^2+1}}, \tag{1.6}$$

$$S_{xy}(2)=\sqrt{\frac{\hat{\varphi}^2}{\max(r-1,c-1)}}, \tag{1.7}$$

$$S_{xy}(3)=\sqrt{\hat{\theta}^2}=\hat{\theta} \tag{1.8}$$

$$S_{xy}(4)=\sqrt{\frac{\hat{\varphi}^2}{\sqrt{(r-1)(c-1)}}}. \tag{1.9}$$

值得注意的是，当$r=c=2$时，即当X和Y都是二态变量时，$\hat{\varphi}^2$和$\hat{\theta}^2$有简单的表达式

$$\hat{\theta}^2=\hat{\varphi}^2=\frac{(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}. \tag{1.10}$$

后面我们将要指出，它还有另外的含意。

三、最优预测系数[17]

现在从另一个角度考虑变量间的关联性，所遵循的原则是把一个变量对另一个变量的预测能力作为关联性强弱的度量标准（见文献[17]）。

在我们对变量Y的取值状态进行预测时，需要按两种不同情况分别处理。

（1）当X的状态未知时，选择使Y的边缘概率最大的状态作为Y的预测状态，即取满足以下条件的状态y_m：

$$\pi_{.m}=\max(\pi_{.1},\ \pi_{.2},\cdots,\pi_{.c}).$$

这时出错概率为

$$p_1 \triangleq 1-\pi_{.m}.$$

（2）当X的状态已知为x_i时，我们只须考虑表1.1中的第i行，即考虑在条件$X=x_i$之下Y的条件概率，把条件概率最大的状态作为Y的预测状态，所取的状态y_{m_i}满足以下条