第3章　多元相关分析

——各个类型的变量均适用的场合

3.1 多元相关分析概述

整个多元统计分析的目标都是描述随机变量之间的关系。从建模的逻辑说，对于自变量，人们总希望每个自变量都能独当一面，各司其职，而多个自变量之间相关程度较低，最好不相关甚至独立；对于因变量也是如此。至于因变量与自变量则希望彼此之间相关程度较高，最好完全相关，所以相关分析乃是建模与多元统计分析的基础。

相关分析依涉及的随机变量数目可以分为三大类型：一对一的，一对多的，多对多的。在以往的统计学相关课程里，已经明确介绍的是一对一的类型，虽未明确指明但有提及某些一对多类型的具体例子，因此本章的内容主要讨论多对多的类型，但是鉴于一对多可以视为多对多的特例，其关键处理手法相似，故亦一并讨论。而根据随机变量的类型，三种类型又可各自细分为几种具体情形，总的交叉组合如表 3.1 所示有 16 个之多。

表 3.1 不同数据类型变量交叉组合

	一个分类变量	一个数值变量	多个分类变量	多个数值变量
一分类	一对一	一对一	一对多	一对多
一数值	一对一	一对一	一对多	一对多
多分类	多对一	多对一	多对多	多对多
多数值	多对一	多对一	多对多	多对多

然而，考虑到相关分析对双方变量是完全对称的，一个分类变量对多个数值变量与多个数值变量对一个分类数值变量被看做是同样的情形，于是经简化后总的组合情形为 10 种，分别是：

1. 一对一类型的 3 种具体情形：

(1) 一个分类变量对一个分类变量；

(2) 一个分类变量对一个数值变量；

(3) 一个数值变量对一个数值变量。

2. 一对多类型的 4 种具体情形：

(1) 一个分类变量对多个分类变量；

(2) 一个分类变量对多个数值变量；

(3) 一个数值变量对多个分类变量；

(4) 一个数值变量对多个数值变量。

3. 多对多类型的 3 种具体情形：

(1) 多个分类变量对多个分类变量；

(2) 多个分类变量对多个数值变量；

(3) 多个数值变量对多个数值变量。

3.2　一对一的类型

作为准备，首先讨论一下一对一的类型。如上所述，一对一类型有三种具体情形：一个分类变量对一个分类变量，一个分类变量对一个数值变量，一个数值变量对一个数值变量。由于分类变量与数值变量不同，不具可加性，所以三种情形的相关分析无法依据相同的原理。

3.2.1　一个分类变量对一个分类变量的情形

基本原理是从根据“随机事件相关”概念出发，经由“随机变量值相关”概念转换，最终构造一个统计量，衡量“随机变量相关”程度。

考虑随机事件 A 与 B，其相关程度的大小可以理解为不独立程度的高低。当随机事件 A 与 B 相互独立时，根据事件独立的定义，应有

$$P(B \mid A) = \frac{P(AB)}{P(A)} = P(B), \qquad P(A \mid B) = \frac{P(AB)}{P(B)} = P(A)$$

两式皆等同于

$$P(AB) = P(A)P(B)$$

当随机事件 A 与 B 不相互独立时，上面各式则不成立，且各式左端与右端

$$P(B \mid A) \text{ 与 } P(B) \text{ 因而 } P(AB) \text{ 与 } P(A)P(B)$$

$$P(A \mid B) \text{ 与 } P(A) \text{ 因而 } P(AB) \text{ 与 } P(A)P(B)$$

差异越大，则距离独立状态越远，相关程度就越高。故可以构造统计量

$$r(AB) = \left| \frac{P(A)P(B) - P(AB)}{P(A)P(B)} \right| = \left| 1 - \frac{P(AB)}{P(A)P(B)} \right|$$

来表示随机事件 A 与 B 的相关程度。该统计量的物理意义是 $P(AB)$ 以 $P(A)P(B)$ 为参照的相对误差，其实质则是两个分类变量的实际状态到独立状态的相对距离。

由于随机事件可用随机变量值来表示，例如用 $X=1$ 和 $X=0$ 分别表示抛掷硬币出现正面和反面的结果，所以可用 $P_i = P(X=i)$ 表示事件 $X=i$ 发生的概率，$P_j = P(Y=j)$ 表示事件 $Y=j$ 发生的概率，$P_{ij} = P(X=i, Y=j)$ 来表示 $X=i$ 和 $Y=j$ 两个事件同时发生的概率。类似地，可用 $r(i,j) = \left| \frac{P_{ij} - P_i P_j}{P_i P_j} \right|$ 来表示 $r(AB)$。

这样，我们就已经定义了随机事件的相关程度和随机变量值的相关程度的衡量指标，那么，能否在此基础上仿此定义随机变量的相关程度的衡量指标呢？

与随机变量值的相关程度不同的是，随机变量的相关程度必须考虑所有随机变量可能取值的情况而非单一随机变量值的情况。为此，要将 $r(i,j)=\left|\frac{P_{ij}-P_iP_j}{P_iP_j}\right|$ 对所有 $X=i$ 和 $Y=j$ 的可能组合求平均，且为了避免使用绝对值函数，现考虑以

$$r(i,j)^2=\left|\frac{P_{ij}-P_iP_j}{P_iP_j}\right|^2=\left(\frac{P_{ij}-P_iP_j}{P_iP_j}\right)^2$$

代替之，这样构造的用来表征两个随机变量间相关程度的指标记为 χ^2。

$$\chi^2=\sum_{i,j}\left(\frac{P_{ij}-P_iP_j}{P_iP_j}\right)^2P_iP_j$$

可理解为两个分类型随机变量的一般相关状态相对于完全独立状态的相对误差之平方的均值，其物理意义为两个分类型随机变量的相关状态到完全独立状态的相对距离长短，χ^2 值越大，离独立状态越远，其相关程度就越高。

鉴于 χ^2 统计量是一个和式，与其中的加项项数有关，此外由于

$$P_i=\sum_jP_{ij}\ 与\ P_j=\sum_i=P_{ij}$$

两个关系式的存在，显然还有一个自由度的问题，所以出于消除加项项数和自由度的影响以及使随机事件及随机变量值的相关系数尽量可比的考虑，围绕 χ^2 统计量，人们构造了一系列不同统计量衡量两个分类型随机变量的相关程度，常用的是以下 3 种。

Pearson 列联系数　$\varphi=\sqrt{\dfrac{\chi^2}{n+\chi^2}}$

Cramer 关联系数　$\nu_1=\sqrt{\dfrac{\chi^2}{n\max\{c-1,r-1\}}},\nu_2=\sqrt{\dfrac{\chi^2}{n\min\{c-1,r-1\}}}$

Cramer 修正关联系数　$v=\sqrt{\dfrac{\chi^2}{n(c-1)(r-1)}}$

其中，c 为列变量的类数，r 为行变量的类数。

由于 χ^2 里含有 $r(i,j)^2$ 的因子，因此上述各式都要开平方，具有将这些系数与 $r(AB)$ 和 $r(i,j)$ 比较的内蕴。易见这些系数（统以 ρ 表示）均满足：

（1）对称性；

（2）$\rho\leqslant1$；

（3）$\rho=\pm1$ 时，代表 X 和 Y 可以相互线性表出；

（4）$\rho=0$ 时，代表 X 和 Y 相互独立。

其实，凡是满足上述 4 个条件的统计量都可视为两个分类变量的相关系数。

例 3.1（吸烟与肺癌的关系）　吸烟的恶果之所以会引起人们严重的关注，最早是由于对肺癌患者吸烟情况的观察。1972 年英国医生泰勒歌德博士说，他所看到的肺癌患者几乎都是吸烟的。随着很多医生关于肺癌患者吸烟情况报道资料的不断积累，人们越来越感到有必要对吸烟恶果问题进行科学研究。

在某个关于这个问题的对照统计实验中，选择 63 个肺癌病例，以及 43 个与肺癌患者年龄、性别和其他属性相类似的健康人。后者构成对照组。然后分别调查肺癌患者和对照组中的人的吸烟情况。调查结果见表 3.2。

表 3.2　吸烟与肺癌对照统计实验数据

	吸烟		不吸烟		合计	
	频数	频率/%	频数	频率/%	频数	频率/%
肺癌患者	60	56.6	3	2.8	63	59.4
对照组	32	30.2	11	10.4	43	40.6
合计	92	86.8	14	13.2	106	100.0

此例中，i 为是否患肺癌，j 为是否吸烟，可依据 $\chi^2=\sum_{i,j}\left(\frac{P_{ij}-P_iP_j}{P_iP_j}\right)^2P_iP_j$ 算出 χ^2 统计量，其自由度为 1，p 值为 0.00188，所以我们有充分的理由相信，是否吸烟与是否患肺癌是相关的。吸烟者相较于不吸烟者，更有可能患肺癌。

例 3.2（人脑左右半球与良性、恶性肿瘤的关系）　某医疗组织曾经做过一项调查，希望了解肿瘤长在人脑的左半球还是右半球与其是否为恶性肿瘤，是否有相关关系。该组织收集了 16 个患者的发病情况，具体见表 3.3。

表 3.3　肿瘤性质与其位置的数据

	良性肿瘤		恶性肿瘤		合计	
	频数	频率/%	频数	频率/%	频数	频率/%
左半球	9	56.3	3	18.8	12	75.0
右半球	1	6.3	3	18.8	4	25.0
合计	10	62.5	6	37.5	16	100.0

同样，按照上文中的公式 $\chi^2=\sum_{i,j}\left(\frac{P_{ij}-P_iP_j}{P_iP_j}\right)^2P_iP_j$ 进行计算，可求得该统计量的 p 值为 0.0736，在 $\alpha=0.05$ 的显著性水平下，我们并没有充分的理由断定，肿瘤所在人脑的位置与其是否为恶性肿瘤明显相关。

例 3.3(苏联党员与受教育年限的关系) 1957 年,苏联政府对其人口中党派成员的分布进行了一次调查,并且与其受教育程度、性别、年龄等因素分别进行了双向分类,希望探索党派成员与以上各个因素之间是否存在相关关系。其中,是否为共产党员与受教育年限的调查数据如表 3.4 所示。

表 3.4 苏联党员与受教育年限关系调查数据

	小于 4 年		4～7 年		8～10 年		大于 10 年		合计	
	频数	频率/%	频数	频率/%	频数	频率/%	频数	频率/%	频数	频率/%
是党员	0	0.0	2290478	1.6	4580957	3.2	1367696	0.9	8239131	5.7
不是党员	48398440	33.3	38685090	26.6	45243862	31.1	4764765	3.3	137092157	94.3
合计	48398440	33.3	40975568	28.2	49824819	34.3	6132461	4.2	145331288	100.0

依据公式 $\chi^2 = \sum_{i,j}\left(\frac{P_{ij} - P_iP_j}{P_iP_j}\right)^2 P_iP_j$ 可算出其 χ^2 统计量,值得注意的是,在该案例中,受教育年限包括小于 4 年,4 ～ 7 年,8 ～ 10 年与大于 10 年四个水平,所以 χ^2 统计量的自由度是 $(r-1)(c-1)=3$,p 值约等于 0,故我们有充分的理由相信,是否为苏共党员与其受教育年限存在相关关系。一个人的受教育程度越高,他(她)就有更大的可能性是苏共党员。这也从另一个侧面反映了苏共是一个受教育程度较高的政党,具有精英党派的特征。

3.2.2 一个分类变量对一个数值变量的情形

基本原理是假如该分类变量与另一个数值变量相关,那么对应不同分类变量值的数值变量的值不仅应不同,而且应是系统性的不同,其表现为对应不同分类变量值的数值变量的均值应有显著不同。这与分类变量的相关有着内在的联系,例如是否患肺病与是否吸烟这两个分类变量被医学实验证明是高度相关的,而是否患肺病会影响寿命这一数值变量,因此吸烟者寿命短,不吸烟者寿命长,尽管会有一些反例,但从统计上说,这是大概率事件,可以看做是系统性的规律。于是,分类变量不同取值所对应的不同组的数值变量均值之间就应有明显差异。不同组的数值变量均值之间的差异称为组间差,同组内部的数值变量取值与其均值之间的差异称为组内差。于是构造一个比例统计量

$$\Lambda = \frac{\text{组间差}}{\text{组内差} + \text{组间差}}$$

假如一个分类变量与一个数值变量相关,该统计量应大于 0,假如完全不相关,则该统计量应等于 0,这与方差分析的思想实质上是一致的。

$$\Lambda=\frac{\text{组间差}}{\text{组内差}+\text{组间差}}=\frac{1}{\frac{\text{组内差}}{\text{组间差}}+1}\propto\frac{\text{组间差}}{\text{组内差}}\propto\frac{\text{平均组间差}}{\text{平均组内差}}=F$$

方差分析认为，分类变量不同取值所对应的不同组的数值变量均值之间有明显差异的标准，是平均组间差要超过平均组内差一定程度，具体说在置信水平 $1-\alpha$ 下，应有

$$F\geqslant F_{\alpha}(L-1,n-L)$$

其中 n 仍然是指样本量，L 是指组数，与分层抽样中的层数含义相当。

但方差分析并未使用上述名正言顺的组间差定义，其组间差采用的定义是各组均值与总均值的差异，具体说是 $\sum_{h=1}^{L}n_h(\bar{y}_h-\bar{y})^2$ 而非 $\sum_{h=1}^{L}\sum_{k=1}^{L}n_kn_h(\bar{y}_h-\bar{y}_k)^2$。那么这两者之间存在怎样的逻辑与数量关系呢？

定理　$\sum_{h=1}^{L}\sum_{k=1}^{L}n_kn_h(\bar{y}_h-\bar{y}_k)^2=2n\sum_{h=1}^{L}n_h(\bar{y}_h-\bar{y})^2$

证明　$\bar{y}_h-\bar{y}=\bar{y}_h-\sum_{h=1}^{L}\frac{n_k}{n}\bar{y}_k=\frac{1}{n}\sum_{k=1}^{L}n_k(\bar{y}_h-\bar{y}_k)$

$$\sum_{h=1}^{L}n_h(\bar{y}_h-\bar{y})=\sum_{h=1}^{L}n_h\bar{y}_h-\sum_{h=1}^{L}n_h\bar{y}=\sum_{h=1}^{L}n_h\bar{y}_h-n\bar{y}=0$$

$$\begin{aligned}\sum_{h=1}^{L}n_h(\bar{y}_h-\bar{y})^2&=\sum_{h=1}^{L}n_h(\bar{y}_h^2+\bar{y}^2-2\bar{y}_h\bar{y})\\&=\sum_{h=1}^{L}n_h(\bar{y}_h^2-\bar{y}^2)=\sum_{h=1}^{L}n_h(\bar{y}_h-\bar{y})(\bar{y}_h+\bar{y})\\&=\sum_{h=1}^{L}n_h(\bar{y}_h-\bar{y})\bar{y}_h+\sum_{h=1}^{L}n_h(\bar{y}_h-\bar{y})\bar{y}=\sum_{h=1}^{L}n_h\bar{y}_h(\bar{y}_h-\bar{y})\\&=\sum_{h=1}^{L}n_h\bar{y}_h\left[\frac{1}{n}\sum_{k=1}^{L}n_k(\bar{y}_h-\bar{y}_k)\right]\\&=\frac{1}{n}\sum_{h=1}^{L}\sum_{k=1}^{L}n_kn_h\bar{y}_h(\bar{y}_h-\bar{y}_k)=\frac{1}{n}\sum_{h=1}^{L}\sum_{k=1}^{L}n_kn_h\bar{y}_k(\bar{y}_k-\bar{y}_h)\\&=\frac{1}{2}\left[\frac{1}{n}\sum_{h=1}^{L}\sum_{k=1}^{L}n_kn_h\bar{y}_h(\bar{y}_h-\bar{y}_k)+\frac{1}{n}\sum_{h=1}^{L}\sum_{k=1}^{L}n_kn_h\bar{y}_k(\bar{y}_k-\bar{y}_h)\right]\\&=\frac{1}{2n}\sum_{h=1}^{L}\sum_{k=1}^{L}n_kn_h(\bar{y}_h-\bar{y}_k)^2\end{aligned}$$

于是即有

$$\sum_{h=1}^{L}\sum_{k=1}^{L}n_kn_h(\bar{y}_h-\bar{y}_k)^2=2n\sum_{h=1}^{L}n_h(\bar{y}_h-\bar{y})^2=2\sum_{k=1}^{L}n_k\sum_{h=1}^{L}n_h(\bar{y}_h-\bar{y})^2$$

例 3.4（行业服务质量）　为了对几个行业的服务质量进行评价，消费者协会

在零售业、旅游业、航空公司与家电制造业这四个行业中分别抽取了不同的企业作为样本。最近一年中消费者对着 23 家企业投诉的次数如表 3.5 所示。

表 3.5　不同行业投诉次数数据

观测值	行　业			
	零售业	旅游业	航空公司	家电制造业
1	57	68	31	44
2	66	39	49	51
3	49	29	21	65
4	40	45	34	77
5	34	56	40	58
6	53	51		
7	44			

消费者协会希望了解消费者对服务质量的投诉与行业之间是否存在相关关系，换句话说，就是希望了解不同行业间，消费者对服务质量的投诉量是否有明显差异。对于一个分类变量与一个数值型变量之间的相关关系，我们利用方差分析的思想，分别计算其组内差与组间差。其中，组间差为 1456.6，组内差为 2708，分别除以相对应的自由度后，得到平均组间差与平均组内差。用平均组间差除以平均组内差，就可得到一个服从分布 $F(k-1,n-k)$ 的 F 统计量。经检验，该统计量的 p 值为 0.0388，所以在 $\alpha=0.05$ 的显著性水平下，我们有充分的理由相信，消费者对服务质量的投诉与行业之间是存在明显的相关关系的。

例 3.5（判别机动割草机拥有者与非拥有者）　考虑某城市中的两群人，拥有机动割草机的人群与不拥有机动割草机的人群。在一次促销活动中，某机动割草机的制造商为预测最佳销售前景，想把该市家庭分为可能购买割草机的家庭和不太可能购买的家庭，所依据的变量是草坪面积的大小。12 个现在的拥有者与 12 个非拥有者两个随机样本所产生的数据如表 3.6 所示。

表 3.6　是否拥有机动割草机者的草坪面积数据　　单位：$100\times m^2$

拥有者	18.4	16.8	21.6	20.8	23.6	19.2	17.6	22.4	20.0	20.8	22.0	20.0
非拥有者	19.6	20.8	17.2	20.4	17.6	17.6	16.0	18.4	16.4	18.8	14.0	14.8

从数据中我们可以看到，割草机拥有者倾向于拥有较大的草坪。与上例不同的是，本例感兴趣的点在于利用草坪大小与是否拥有机动割草机的相关关系，对潜在用户群进行判别。同样，由于只有机动割草机拥有者与机动割草机非拥有者两

个水平，故可利用 t 检验对两组调查对象的草坪大小是否有显著差异进行检验。检验得到的 p 值为 0.005，故我们有充分的理由相信草坪大小与是否拥有机动割草机是存在相关关系的。

3.2.3　一个数值变量与另一个数值变量的情形

基本原理是通过两个随机变量变化趋势的一致性程度来反映线性相关关系。考虑极端情况，如果 X 与 Y 完全线性相关，自然可以互为线性表出 $Y=\beta_0+\beta X$。

对于一个值，是 $Y_i=\beta_0+\beta X_i$，对于均值，则是 $\bar{Y}=\beta_0+\beta\bar{X}$，离差为 $Y_i-\bar{Y}=\beta(X_i-\bar{X})$ 或 $(Y_i-\bar{Y})=\beta(X_i-\bar{X})$。离差两端同乘 $(X_i-\bar{X})$，有

$$(Y_i-\bar{Y})(X_i-\bar{X})=\beta(X_i-\bar{X})^2$$

离差和则为 $\sum_{i=1}^{n}(Y_i-\bar{Y})(X_i-\bar{X})=\beta\sum_{i=1}^{n}(X_i-\bar{X})^2$。

注意斜率 β 是固定的，由于 $\sum_{i=1}^{n}(X_i-\bar{X})^2\geqslant 0$，所以 $\sum_{i=1}^{n}(Y_i-\bar{Y})(X_i-\bar{X})$ 的符号与斜率 β 一致。为了使两个随机变量的量纲同一，并消除 n 的大小影响，定义

$$r=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma_x}\right)\left(\frac{Y_i-\bar{Y}}{\sigma_y}\right)$$

易见 $r=\beta\frac{\sigma_x}{\sigma_y}$ 即 $r\propto\beta$。

因此 r 可以很好地反映两个随机变量变化趋势的方向与紧密程度，这就是著名的 Pearson 相关系数。其正值表示两个随机变量同向变化，负值表示两个随机变量逆向变化。绝对值则表示了两者变化趋势的紧密程度。

由于在普通情形（即一个变量不能为另一变量线性表出）下，$(Y_i-\bar{Y})(X_i-\bar{X})$ 对于不同的 i 其符号未必一致，所以 $\sum_{i=1}^{n}(Y_i-\bar{Y})(X_i-\bar{X})$ 可以看做各项求和正负相抵后的整体结果，因此 r 是一个可以推广到普通情形的反映线性相关关系的指标。

3.3　多对多类型

在一对一场合，针对不同类型的随机变量对，人们分别从到独立状态的距离、组间差所占比例和变化紧密性与一致性程度定义了 3 种明显不同的相关系数。这些相关系数的共同特点是(1)均无量纲，绝对值在 0 与 1 之间；(2)绝对值大小反映随机变量相关关系的密切程度；(3)计算具有唯一性。

在多对多的情形，一组分类型随机变量对另一组分类型随机变量之间的相关系数，线性运算仍是不允许的，一种自然的也是别无选择的方案是通过分别将两组的组内所有变量交叉组合为一个超级分类变量（该超级分类变量的水平数等于相应组内所有变量水平数的乘积，如一组有 3 个变量，其水平数分别为 4，2，5，则 3 个变量所交叉合并成的超级分类变量其水平数为 $4\times2\times5=40$ 个），从而将多对多问题化为一对一问题，用两个超级分类变量的相关系数来反映两组分类变量之间的线性相关关系。同样的手法也可用于一对多的情形，不赘。

但这样的做法并不能移植到确定两组数值型随机变量之间的相关系数的过程中，对于两组数值型随机变量之间的相关系数的定义，必须另觅途径。一个解决问题的思路是借鉴一个因变量对多个自变量的回归分析，将若干自变量的一个线性组合看做一个变量，定义该变量组合与因变量的 Pearson 相关系数为复相关系数或决定系数（此即数值变量的一对多情形，故以后也不再讨论一对多的情形）。这一线性组合是通过最小二乘法获得的，其理论上与因变量的线性相关系数是所有线性组合中是最大[①]的。既然对一组随机变量可以如此，对两组有何不可？事实上，以两组随机变量的拥有最大相关系数的线性组合对的相关系数代表两组随机变量的相关系数，就可以在**确保唯一性**的前提下实现对看似复杂无解的数值型多对多相关关系的刻画。这种以“代表”方式来反映整体相关关系的方法称为典型相关分析。典型一词，在中文里具有代表的意思。例如所谓典型调查就是以总体里的代表性个体为调查客体的调查。

例 3.6（美国密尔沃基房屋售价） 某机构曾经在美国威斯康星州密尔沃基的一个住宅区做过一项关于房屋售价的调查。该组织走访了住宅区内 20 个家庭，调查了这 20 个家庭的房屋售价、总居住面积与评估价值这三项指标。具体数据如表 3.7 所示。

表 3.7　美国密尔沃基房屋售价数据

总居住面积/10^3 ft^2	估价/千美元	售价/千美元
15.31	57.3	74.8
15.20	63.8	74.0
16.25	65.4	72.9
14.33	57.0	70.0
14.57	63.8	74.9

① 谋求因变量与自变量的相关系数最大的理由显而易见，相关系数越大自变量对因变量的解释能力越强。