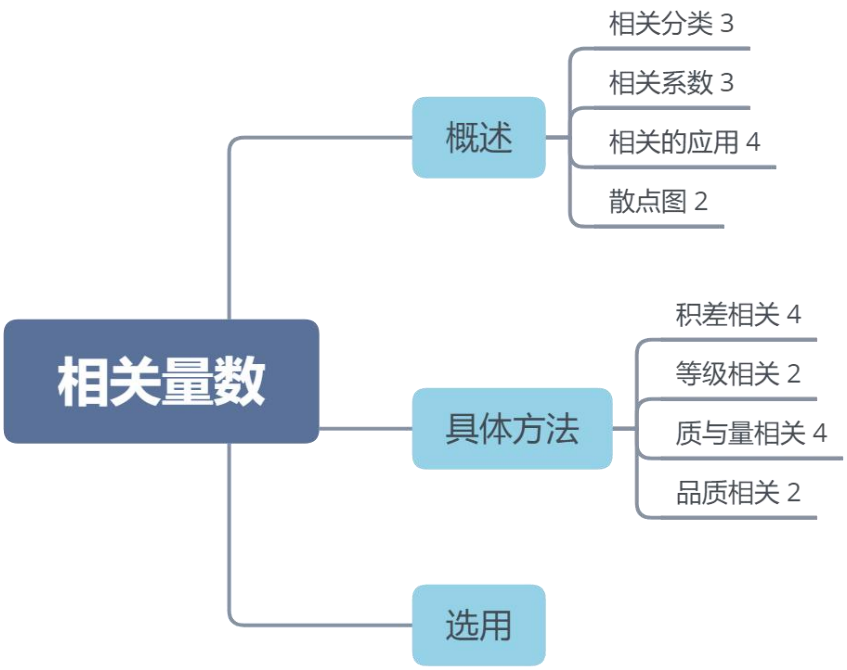
第五章　相关量数



一、概述

（一）相关分类

1.正相关

一列变量变动时，另一变量同时发生或大或小的相同方向的变动。

2.负相关

一列变量变动时，另一变量同时发生或大或小的相反方向的变动。

3.零相关

一列变量变动时， 另一变量做无规律的变动。

（二）相关系数

相关关系强度的指标。作为样本统计量时用 r 表示，作为总体参数时一般用ρ表示。 它是和平均数、标准差一样应用广泛的统计量。

1.特点

（1）取值范围是[-1，1]，这里讲的相关是线性相关。当然，即使是线性相关为 0 时，仍可能存在曲线相关。

（2）正负号表示相关的方向，正值表正相关，负值表负相关。其中，r =+1，完全正相关；r =-1,完全负相关；r=0，零相关，即不存在线性相关。

（3）相关系数的取值的大小表示相关的强弱程度，正负号仅意味着方向。相关系数的绝对值接近 1，相关程度越密切；绝对值接近 0，关系越不密切。

2.解释

（1）存在相关关系不一定存在因果关系。原因：①无法找到两个变量相互影响的方向；②无法排除由于存在第三个变量同时引起了这两个变量的变化，即伪相关。

（2）相关系数 r 是一个比值，不能进行加减乘除运算。

（3）r 值受到样本量的影响。

（4）在纯理论研究中，即使是很小的相关，若在统计上有显著性，也能说明心理规律。

3.应用

0≤|r|<0.2 时，可能没有相关；0.2≤|r|<0.4 时，弱相关；0.4≤|r|<0.6,中等程度相关；0.6≤|r|<0.8 时，强相关；0.8≤|r|<1 时，非常强的相关；|r|=1 时，完全相关。

（三）相关的应用

1.预测：已知两个变量以某种系统方式相关联，就可以用其中一个变量对另一个变量做精确的预测。

2.效度：可以用相关来计算效度。

3.信度：相关可以测验信度，信度高时，两个测量的相关是高度正相关。

4.理论验证：理论的预测可以由两个变量间的相关来检验。

（四）散点图

散点图是用圆点多少和分布疏密来表示两个变量的相关程度的统计图。

1.若所有散点呈直线，直线呈左低右高为完全正相关，形状右低左高为完全负相关。

2.若所有散点呈椭圆状，则说明两变量之间为线性相关，形状左低右高为正相关，形状右低左高为负相关；若所有散点呈圆形，就为零相关或弱相关。

二、具体方法

（一）积差相关

积差相关也就是 Pearson 相关，又称为积矩相关，是计算相关的最常用和最基本的方法。

1.适用范围

（1）数据要成对出现，即若干个体中每个个体都有两种不同的观测值，并且每对数据与其他对子相互独立。（一个人的数学和语文成绩）

（2）两列变量各自总体的分布都是正态的，至少接近正态。

（3）两个相关的变量是连续变量，即两列数据都是测量数据。

（4）两列变量之间的关系应是线性的。

2.离差积和（SP）

（1）定义式

$$SP = (X - M_X)\ (Y - M_Y)$$

（2）计算式

$$SP = XY - \frac{\sum X \sum Y}{n}$$

3.积差相关的计算

（1）基本式

$$r=\frac{\sum\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)}{Ns_Xs_Y}=\frac{\sum xy}{\sqrt{\sum x^2\times\sum y^2}},\quad x=X-\overline{X},\quad y=Y-\overline{Y}$$

（2）使用原始数据

$$r=\frac{\sum XY-\frac{\sum X\sum Y}{N}}{\sqrt{\sum X^2-\frac{\left(\sum X\right)^2}{N}}\sqrt{\sum Y^2-\frac{\left(\sum Y\right)^2}{N}}}$$

（3）使用标准分数

$$r=\frac{1}{N}\sum Z_XZ_Y$$

（4）其他

①差法公式

$$r=\frac{\sum x^2+\sum y^2-\sum(x-y)^2}{2\times\sqrt{\sum x^2}\times\sqrt{\sum y^2}}=\frac{\sum(x+y)^2-\sum x^2-\sum y^2}{2\times\sqrt{\sum x^2}\times\sqrt{\sum y^2}}$$

②合并

A. $r-Z$

B. $\overline{Z}=\frac{\sum(n_i-3)Z_i}{\sum(n_i-3)}$（$n_i$不同）或$\overline{Z}=\frac{\sum Z_i}{N}$（$n_i$相同）

C. $Z-r$

（二）等级相关

1.斯皮尔曼等级相关

是等级相关的一种，又称斯皮尔曼ρ系数，常用符号 rR 或 rS 表示。

（1）适用范围

①只有两列变量，变量为顺序型数据或称名数据；

②两列变量之间的关系应是线性的；

③当研究考察的变量的总体分布非正态时。

（2）计算

①等级差数法

$$r_R=1-\frac{6\sum D^2}{N^3-N},\quad D=R_X-R_Y,$$

R_X 和 R_Y 两变量各自等级序数。

②存在相同等级

$$r_R=\frac{\sum x^2+\sum y^2-\sum D^2}{2\times\sqrt{\sum x^2\times\sum y^2}},\quad \sum x^2=\frac{N^3-N}{12}-\sum C_X,\quad \sum C_X=\sum\frac{n^3-n}{12}$$

$$\sum y^2=\frac{N^3-N}{12}-\sum C_y\quad \sum C_y=\frac{n^3-n}{12}$$

2.肯德尔等级相关

包括适合两列等级变量的交错系数和相容系数，也包括适合多列变量的肯德尔 W 系数和 U 系数，这里主要介绍后一种:

（1）肯德尔 W 系数

又称肯德尔和谐系数，原始数据资料的获得一般采用等级评定法，即让 K 个被试对 N 件事物进行等级评定，或者让 1 个被试先后 K 次评价 N 件事物。

其原理是 W 为评价者评价的一致性除以最大变异的可能性。

（1）肯德尔 W 系数

$$W=\frac{s}{\frac{K^2}{12}(N^3-N)},\quad s=\sum R_i^2-\frac{(\sum R_i)^2}{N}$$

R_i：评价对象获得的 K 个等级之和，N：等级评定的对象的数目，K：等级评定者的数目。

（2）肯德尔 U 系数

$$U=\frac{8(\sum r_{ij}^2-K\sum r_{ij})}{N(N-1)K(K-1)}+1$$

r_{ij} 为对偶比较记录表中 $i>j$ 格中的择优分数。

等级相关的适用范围较积差相关的大，又对总体分布不做要求。但其精确度要差于积差相关，因此凡是符合积差相关的资料，都不用等级相关计算。

（三）质与量相关

当需要计算相关的两列变量一列是按性质划分的类别（即二分变量），而另一列是等距或等比数据时使用。

二分变量又分为真正二分变量和人为二分变量。其中，真正二分变量又称为离散型二分变量或二分称名变量。

1.点二列相关

适用于一列数据为真正二分变量，另一列数据为等距或等比数据。

$$r_{pb}=\frac{\overline{X}_p-\overline{X}_q}{s_t}\cdot\sqrt{pq}$$

$\overline{X}_p$ 是与二分称名变量的一个值对应的连续变量的平均数，$\overline{X}_q$ 是与二分称名变量的另一个值对应的连续变量的平均数，p 与 q 是二分称名变量两个值各自所占的比率（$p+q=1$），s_t 是连续变量的标准差。

2.二列相关

适用于两列变量都是正态等距或等比变量，但其中一列变量被人为地分成两类。即适用于一列数据为人为二分变量， 另一列数据为等距或等比数据。

$$r_b=\frac{\overline{X}_p-\overline{X}_q}{s_t}\cdot\frac{pq}{y}$$

y 为标准正态曲线中 p 值对应的高度，通过查正态分布表得到。

3.两者区别

主要区别是二分变量是否为正态分布，选用这两种方法的总原则是：

（1）若对数据分布形态是否为正态分布不明确，那么不论数据是真正二分变量还是人为二分变量，都使用点二列相关。

（2）若确认数据分布形态为正态分布，都应用二列相关。

4.多列相关

一列为等距或等比的测量数据，一列为被人为划分为多种类别的数据（如学习成绩划分为优、良、中、差）。

（四）品质相关

主要用于表示 R×C（行×列）表的两个变量之间的关联程度。品质相关处理的数据类型一般都是计数数据，而非测量性数据。

1.四分相关

适用于两个变量都是连续变量，且每一个变量的变化都被人为地分为两种类型，即两个变量都是人为二分变量。

2.φ系数

适用于两变量都是真正二分变量，但是除四分相关之外的四格表（计数）资料都可以使用此方法计算相关外，它还是表示两变量两项分类资料相关程度的最常用的一种相关系数。

$$r_{\Phi}=\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

其中 a、b、c、d 分别为四格表中左上、右上、左下、右下的数据。

3.联列表相关

又称均方相依系数、接触系数等，一般用 C 表示，由二因素的 R×C 列联表资料求得。

三、选用

先确定两列变量类型，在选择具体方法。

第一列变量		第二列变量				
		二分变量		等级数据	等距数据	等比数据
		人为二分变量	真正二分变量			
二分变量	人为二分变量	四分相关	φ系数		二列相关	二列相关
	真正二分变量	φ系数	φ系数		点二列相关	点二列相关
等级数据				等级相关	等级相关	等级相关
等距数据		二列相关	点二列相关	等级相关	积差相关	积差相关
等比数据		二列相关	点二列相关	等级相关	积差相关	积差相关