阮敬 博士

首都经济贸易大学研究生院 副院长
首都经济贸易大学统计学院 教授

© ruanjing@msn.com

相关分析

- 现实世界任何事物之间都存在或多或少的必然联系， 数据之间也不例外。 在现实生活中，人们最常见的便是数据之间的函数关系。 在数据间的函数关系下， 一个 （些） 数据发生变动，与之对应的另一个（些）数据会严格按照函数关系发生相应的变动，这种变动情况可以根据函数的具体形式进行精确度量。 但实际上， 数据之间的变动情况还会受到其他没有考虑到或者根本无法考虑的因素的影响，使得数据变动状况很少真正能够用函数的形式来具体描述，数据之间的关系往往体现为相互依存的非函数关系。 而这种关系人们可以根据数据本身的特征和自身经验进行大概的判定。

- 本章将要介绍的相关分析便是分析两个变量或两组变量之间的相互依存关系的一种典型方法。

两变量之间的相关分析

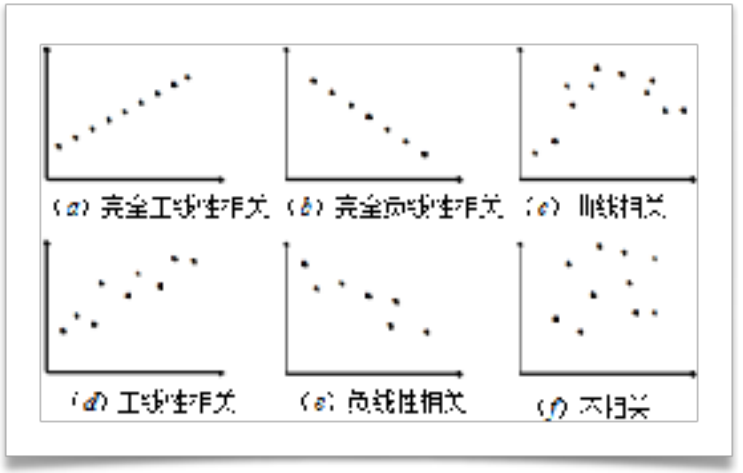
- 相关分析（Correlation Analysis）主要分析两个变量之间的相互依存关系，在介绍相关分析之前，应当先区分变量或数据之间的两种主要关系。
  - 函数关系： 当一个或几个变量取一定的值时， 另一个变量有确定值与之具体严格相对应，则称这种关系为函数关系；
  - 相关关系： 变量之间的影响不能够用具体的函数来度量， 但变量之间的关系确实存在数量上不是严格对应的相互依存关系，称之为相关关系。
- 函数关系是确定性的，往往把发生变动的变量称之为“自变量”，受自变量变动影响而发生变动的变量称之为“因变量”。如牛顿第二定律：F=ma，m 代表质量，a 代表加速度，当 m 不变的时候，a 增加一倍成为 2a，则代表力的 F 变量随之发生变动，也会增加一倍，变为 2F；再如北京市出租车 15 公里内的单价在 5 点到 23 点之间是 2 元/公里，起步价是 3公里 10 元， 某人在早上 10 点钟打车走了 8 公里的路程， 可以根据函数关系精确计算出其应付的出租车价格为 10+（8-3）*2=20 元。

两变量之间的相关分析

- 相关关系是不确定的，主要考察变量之间的相互影响，这种影响不存在方向性，即变量A 与变量 B 相关和变量 B 与变量 A 相关是一致的。相关关系主要体现为变量之间的相互依存关系，如身高和体重之间的关系便是相关关系的一种体现。通常情况下，一个人身高比较高，其体重也会相应比较重，但不能说身高增加 1 厘米，体重就会增加 2 公斤，因为还有例外，即有些身高比较高但比较瘦的人，其体重反而不如身高比较低的人的体重重。身高和体重这两个变量之间虽然不能用函数关系来描述，但是从总体上来说， 这两个变量是存在一定的关系的， 这种关系便是相互依存的关系。 此外， 相关分析不具有传递性， 即 A 和 C 相关，B 和 C 相关，则 A 和 B 不一定相关。
- 相关分析根据其分析方法和处理对象不同， 可以分为简单相关分析、 偏相关分析和非参数相关分析等，本节将对这些分析过程进行详细介绍；此外，相关分析根据相关关系表现形式的不同， 又可以分为线性相关分析和非线性相关分析， 本节主要介绍线性相关的内容和分析过程。

简单相关分析

- 简单相关分析主要分析两个变量之间相互依存的关系， 人们可以通过主观观测和客观测度指标来衡量。
- 主观观测变量之间的相关关系，主要是通过两个变量之间散点图的手段来进行分析的。而客观测度主要是通过统计分析的方法， 计算相关系数， 利用相关系数数值的符号和大小来判定相关关系的方向和强弱。
- **1. 用图形描述相关关系**
  - 利用散点图可以描绘出两个变量的相互影响状况。 选定两个要分析的变量， 把其中任意一个变量指定为二维坐标轴的横轴， 另一个变量指定为纵轴之后， 就可以根据两个变量的每一对数值在二维坐标轴上描点， 所有描出来的点在一起形成了散点图。 根据散点图的不同表现情形，主要有以下几种类型，如图所示。



简单相关分析

- 上图中的（a）和（b）表示了两个变量之间的函数关系，而且这种关系是线性的，可以用一条直线方程来描述两个变量之间一一对应的严格关系。其中（a）表示随着一个变量的增加（减少），另一个变量对应的也增加（减少），这种同增同减的情况被称之为“正相关”；而（b）所描绘的是一个变量的增加（减少），另一个变量减少（增加），这种反向变动的情况被称之为“负相关”。
- 而（c）中描绘了变量之间的曲线相关关系，变量之间的变动关系随着曲线的形式发生，但是这种变动关系同样不能用严格的数学函数表示。
- （d）和（e）分别描述的是正线性相关和负线性相关关系。在这两个图中，只能够看到两个变量变动状况的趋势是直线的，与（a）和（b）相比，二者之间的变动不能够用直线方程严格对应。在（f）中，基本上看不出两个变量之间有相互依存的关系。
- 根据散点图来描述相关关系比较简单和直观，在 SAS 系统中可以使用第 7.2.6 小节中介绍过的图形绘制方法来绘制散点图。但是如果要对相关关系进行进一步分析和下结论，仅用图形来描述就显得主观性比较强。因此，还可以使用相关关系的测度指标——相关系数来衡量变量之间的相互依存关系。

简单相关分析

- 2．用相关系数测度相关关系
- 相关系数是描述线性相关程度和方向的统计量，根据样本收集的数据计算的相关系数通常用字母 r 表示（r 也可称之为样本相关系数）。r 的正负符号表示相关关系的方向，r 的绝对值大小表示相关关系的强弱程度。设有两个变量分别是 x 和 y，根据样本数据计算相关系数的方法主要采用 Pearson 提出的方法，即 Pearson 相关系数：

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

- 如两个变量之间的正向关系可用线性函数表示，则相关系数 r=+1，表示完全正线性相关；如果两个变量之间的负向关系可用线性函数表示，则相关系数 r=-1，表示完全负线性相关。相关系数 r 的取值范围为[-1，+1]，具体有以下几种情况：
  - r=+1，表示完全正线性相关；
  - r=-1，表示完全负线性相关；
  - r<0，表示负线性相关；
  - r>0，表示正线性相关；
  - r=0，表示不存在线性关系。

当计算出来的 r =0 时，只是表示线性关系不存在，但是变量之间有可能存在其他形式的相关关系（如曲线相关）。

简单相关分析

- 此外，|r|的大小可以根据经验，表示不同程度的线性相关关系：
  - |r|<0.3，表示低度线性相关；
  - 0.3≤|r|<0.5，表示中低度线性相关；
  - 0.5≤|r|<0.8，表示中度线性相关；
  - 0.8≤|r|<1.0，表示高度线性相关。
- 上述这种对相关程度的大致判断只是从状态上描述了变量之间的相关关系， 但是相关系数 r 是根据样本数据计算出来的一个统计量，从样本数据分析出来的相关关系，是否能够对总体数据下结论呢？这需要对相关系数的显著性进行检验。

简单相关分析

- **3. 相关系数的显著性检验**
- 相关系数的显著性检验主要就是根据样本数据计算的样本相关系数 r，利用 t 统计量，根据 r 服从自由度为（n-2）的 t 分布的假定，对总体相关系数（通常用 ρ 表示）是否等于 0进行假设检验。如果在一定的显著性水平下，拒绝 ρ=0 的原假设，则表示样本相关系数 r是显著的。因此，该问题又可以归结为一个假设检验的问题，其原假设和备择假设是：

$$H_0: \ \rho=0 \ , \ H_1: \ \rho \neq 0$$

- 该假设检验问题的分析过程和得到结论的方法， 与第 8 章假设检验的分析过程一致。 相关系数显著性的检验也可适用于本章后面介绍的其他相关分析方法。

简单相关分析

- 某杂志为了评价市场上所销售汽车最高时速与汽车自身相应指标的影响， 收集了各大厂商生产的各种系列和型号的中级汽车的最高时速、车身自重、轮胎尺寸、发动机马力等指标数据，如图所示。试对这些指标进行相关分析。

Dataset of SASUSER.CAR_CORR

Obs	Brand_Model 品牌和车型	Weight 车身自重	Circle 轮胎尺寸	Max_Speed 最高时速（英里）	Horsepower 马力
1	Acura Legend V6	3265	42	163	160
2	Audi 100	2935	39	141	130
3	BMW 535i	3640	39	209	208
4	Buick Century	2880	41	151	110
5	Buick Riviera V6	3465	41	231	165
			……		
27	Saab 9000S	3065	40	121	130
28	Sterling 827 V6	3295	42	163	160
29	Toyota Cressida	3480	36	100	190
30	Volvo 740 GL	3140	37	141	114

简单相关分析

- 在本例中， 采用相关分析的方法可以分析各个变量之间的相互依存关系， 对于主要考察最高时速这个变量与其他因素之间的关系， 也可以利用相关分析来判定那个或哪些因素与最高时速有密切的相关关系。
- 简单相关分析可使用 CORR 过程来实现，CORR 过程的主要语法如下：

```
PROC CORR <选项> ;
 BY 变量列表;
 FREQ 变量;
 ID 变量列表;
 PARTIAL 变量列表; /*指定偏相关分析的控制变量*/
 VAR 变量列表;
 WEIGHT 变量;
 WITH 变量列表;
```

- CORR 过程的语法较为简单。 如果要分析指定变量列表中两两变量之间的相关关系， 只需在 VAR 语句中把这些变量列出来即可；也可以使用 WITH 语句来分析指定变量与 VAR变量列表中所列示变量的相关关系。 该过程主要通过过程选项来控制进行分析的内容。 本书将在后续章节的程序注释中对这些控制选项进行介绍。

简单相关分析

- 本例使用 CORR 过程的具体程序如下：

```
proc corr data=sasuser.car_corr pearson;    /*关键字 PEARSON 指定计算 Pearson 相关系数*/
  var weight circle max_speed horsepower;
/*指定相关分析的变量列表*/
run;
```

- 程序运行之后，除了能够得到如图(上)所示的各变量统计量信息之外，还可得到如图（下）所示的相关分析表。

The CORR Procedure

4 Variables: Weight Circle Max_Speed Horsepower

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Weight	30	3305	318.58113	99145	2765	4000	车身重量
Circle	30	40.50000	1.88917	1215	36.00000	43.00000	轮胎尺寸
Max_Speed	30	105.23333	45.61061	3157	121.00000	302.00000	最高时速（英里）
Horsepower	30	152.90000	43.53702	4587	90.00000	278.00000	马力

Pearson Correlation Coefficients, N = 30
Prob > |r| under H0: Rho=0

	Weight	Circle	Max_Speed	Horsepower
Weight 车身重量	1.00000	0.07549 0.6918	0.85459 <.0001	0.82559 <.0001
Circle 轮胎尺寸	0.07549 0.6918	1.00000	0.26369 0.1591	-0.02830 0.8820
Max_Speed 最高时速（英里）	0.85459 <.0001	0.26369 0.1591	1.00000	0.75015 <.0001
Horsepower 马力	0.82559 <.0001	-0.02830 0.8820	0.75015 <.0001	1.00000

简单相关分析

- 相关分析的结果首先给出的是如上图（上）所示各个变量的样本量、样本均值、样本标准差、和、最小值、最大值等样本统计量，变量的标签也会对应的与变量名字一同出现。图11-4 表示 Pearson 相关系数矩阵，同时在该表头下也列示了样本相关系数 r 显著性检验的原假设，即 ρ=0。
- 在相关系数矩阵中，行列交叉对应的数值即为行变量及其对应列变量之间的 Pearson 相关系数。每个相关系数的下面，都会给出用于检验其显著性的 P 值。
- 在上图（下）中可以看到，矩阵的主对角线均为 1，表示变量自己与自己是函数关系。对于最高时速 Max_Speed 变量，其与车身自重、轮胎尺寸、马力 3 个变量的相关系数分别为0.85459、0.26369、0.75015。其中最高时速和轮胎尺寸之间的相关系数 0.26369 在 α=0.05 的条件下不显著（即 P 值=0.1591>>α=0.05） 。因此，在不考虑其他因素的作用下，最高时速与车身自重存在显著的高度正线性相关，与发动机马力存在显著的中度正线性相关关系。

偏相关分析

- 简单相关分析有时不能够真实反映现象之间的关系。 如上述的例 11-1， 发动机作为汽车的心脏，可以说对汽车的各项指标均会产生一定的影响。因此，在研究其他指标与最高时速指标之间的相关关系时， 会不知不觉地在变量之间加入发动机相关指标， 对所研究的变量有影响，而这种影响由于相关关系的不可传递性，往往会得到错误的结论。
- 所以，在进行相关分析时往往要控制这种变量，剔除其对其他变量的影响之后，来研究变量之间的相关关系。 这种剔除其他变量影响之后再进行相关分析的方法称之为偏相关分析（Partial Correlation Analysis） 。
- 仍然以例 11-1 为例，在收集的 4 个指标中，依据常识，发动机马力这个变量对最高时速影响非常大， 在考虑最高时速与其他变量的相关关系时， 有可能包含了发动机马力因素的影响，因此，考虑剔除发动机马力变量影响的偏相关分析。

偏相关分析

- 使用 CORR 过程的 PARTIAL 语句指定控制变量便可进行偏相关分析，本例程序如下：

```
proc corr data=sasuser.car_corr pearson;
  var weight circle max_speed;
  partial horsepower;
run;
```

- 程序运行之后除了可得到类似上图（上）的统计量结果之外，还可得到如右图所示的偏相关系数的计算结果。
- 从偏相关分析结果中看到， 控制住发动机动力变量的影响之后， 最高时速与车身自重的相关系数有所降低，具体数值为 0.63053，处于中度线性相关的范围；而轮胎尺寸与最高时速的相关系数大幅提升，且该相关系数在显著性水平 α=0.05 条件下显著。这个分析结果，尤其是轮胎尺寸与最高时速之间的关系，较简单相关分析结果而言与实际状况更加接近。 汽车轮胎的尺寸增加，在一定程度上可以增加车身的抓地性能，从而提升速度；反过来，汽车速度不断增加，在其他条件不变的情况下，没有一定尺寸的轮胎，其最高速度也很难提升。

Pearson Partial Correlation Coefficients, N = 30
Prob > |r| under H0: Partial Rho=0

	Weight	Circle	Max_Speed
Weight 车身自重	1.00000	0.17525 0.3632	0.63053 0.0002
Circle 轮胎尺寸	0.17525 0.3632	1.00000	0.43105 0.0196
Max_Speed 最高时速（英里）	0.63053 0.0002	0.43105 0.0196	1.00000

非参数相关分析

- 简单相关分析和偏相关分析广泛应用于定量数据或连续型数据的研究。 对于某些数据尤其是定性数据的相关分析而言，如果用 Pearson 法计算相关系数，很难得到定性数据的协方差和标准差。因此，可以考虑其他的方法对这些数据尤其是顺序数据进行相关分析。
- 对于上述情况的相关分析往往从数据值的次序入手， 并借助非参数统计分析的思想。 次序在数列中代表了某个具体变量值的位置、 等级或秩， 因此这类相关分析通常称之为非参数相关分析、等级相关分析或秩相关分析，其计算的相关系数对应的称为非参数相关系数、等级相关系数或秩相关分析。
- 非参数相关系数计算方法较多，常见的主要有 Spearman、Kendall tau-b 和 Hoeffding's D相关系数等。

非参数相关分析

- 1. Spearman 相关系数
- 该相关系数主要测度顺序变量间的线性相关关系， 在计算过程中只考虑变量值的顺序而不考虑变量值的大小。
- 其计算过程为：首先把变量值转换成在样本所有变量值中的排列次序，再利用 Pearson方法求解转换后的两个变量对应的排列次序（即“秩”或等级）的相关系数。其具体计算公式为：
-

$$r=\frac{\sum(R_{x_i}-\overline{R}_x)(R_{y_i}-\overline{R}_y)}{\sqrt{\sum(R_{x_i}-\overline{R}_x)^2\cdot\sum(R_{y_i}-\overline{R}_y)^2}}$$

- 其中，R_{x_i}和R_{y_i}分别表示第 i 个 x 变量和 y 变量经过排序后的次序，$\overline{R_x}$和$\overline{R_y}$分别表示R_{x_i}和R_{y_i}的均值。

非参数相关分析

- 2. Kendall tau-b 系数：
- 该系数与 Spearman 相关系数作用类似，主要测度顺序变量间的线性相关关系，其计算过程中也是只考虑变量值的顺序而不考虑变量值的大小。在 Kendall tau-b 系数计算过程中，除对数据进行排序之外，还应当综合考虑该排序与变量值的具体情况，即：
  - 同序对：在两个变量上排列顺序相同的一对变量；
  - 异序对：在两个变量上排列顺序相反的一对变量。
- 上述对子的数目简称为对子数，设 P 为同序对子数，Q 为异序对子数，T_x为在 x 变量上是同序但在 y 变量上不是同序的对子数，T_y为在 y 变量上是同序在 x 变量上不是同序的对子数，则 Kendall tau-b 系数为：

$$\tau_b = \frac{P-Q}{\sqrt{(P+Q-T_x)(P+Q+T_y)}}$$

- τ_b的取值范围与简单相关系数相同，即$\tau_b \in [-1,+1]$

非参数相关分析

- 3. Hoeffding's D 系数：
- 该系数主要用于测度顺序变量或具有等级水平变量间的线性相关关系， 其具体计算公式如下：

$$D = 30 \times \frac{(n-2)(n-3)D_1 + D_2 - 2(n-2)D_3}{n(n-1)(n-2)(n-3)(n-4)}$$

其中：

$D_1 = \sum (Q_i - 1)(Q_i - 1)$；

$D_2 = \sum (R_i - 1)(R_i - 2)(S_i - 1)(S_i - 2)$；

$D_3 = \sum (R_i - 2)(S_i - 2)(Q_i - 1)$。

- R_i、S_i分别表示变量 x 、 y 的排列顺序；Q_i表示 1 加上变量 x 和 y 的值均小于这两个变量中的第 i 个值时的个数，也称之为双变量等级。
- 上述相关系数计算方法也可应用于定量数据中， 在相关分析中只要除去定量数据的数值意义即可。如例 11-1 所示的数据，同样可用于非参数相关系数的计算，只是计算结果所代表的相关含义会发生相应的变化。