

第五章 相关系数

第一节 相关概述

第二节 积差相关系数

第三节 其他相关系数

第一节 相关概述

一、相关的含义

客观现象之间的数量联系存在着函数关系和相关关系。当一个或几个变量取定值时，另一个变量有确定的值与之对应，称为函数关系，可用 $Y=f(X)$ 表示。

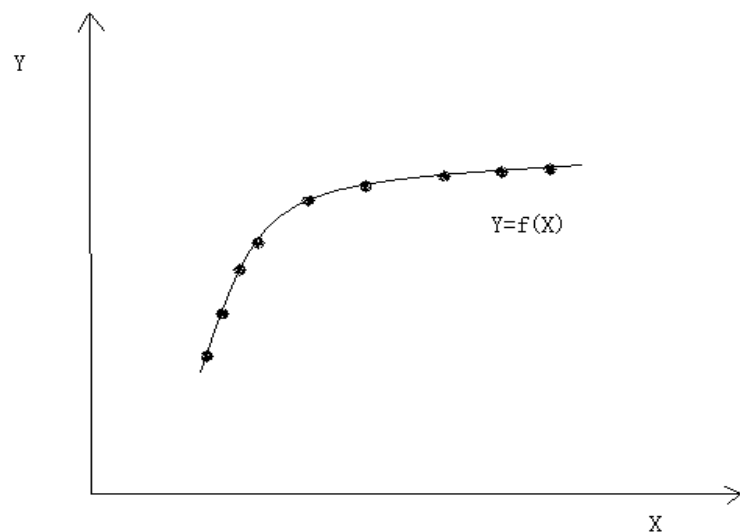


图5-0(a) 函数关系

所谓**相关**就是指事物或现象之间的相互关系。事物之间在数量上的变化关系有的是属于因果关系（一种现象是另一种现象的原因，另一种现象是这种现象的结果），有的却不能直接作出因果关系的解释。当一个或几个相互联系的变量取一定数值时，与之相对应的另一个变量的值虽然不确定，但它仍然按某种规律在一定范围内变化，变量间的这种关系，被称为**相关关系**，如图5-0（b）。

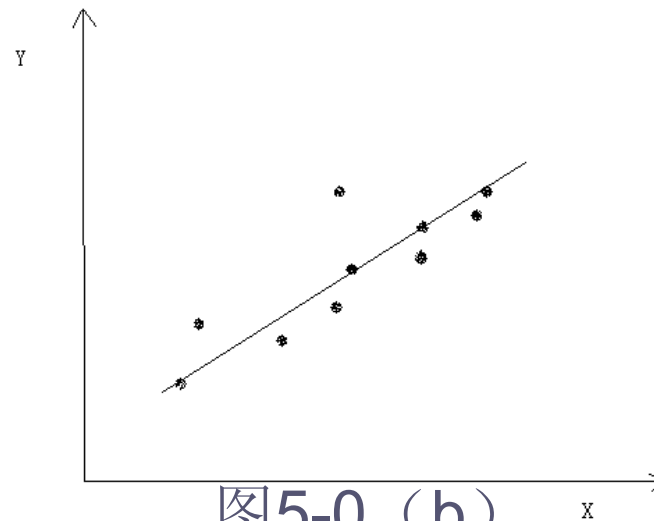


图5-0（b）

二、相关的种类

（一）从变化方向上划分

1、正相关。两个变量中，一个变量增大，另一个变量对应值也随之增大；或一个变量值减小，另一个变量对应值也随之减小，两列变量变化方向相同。如学生的学习成绩与智商之间的关系；教师工作积极性与学校民主管理程度之间的相关，学校办学经费与教学设施之间的相关等。

2、负相关：两个变量中，一个变量增大，另一个变量对应值也随之减少；或一个变量值减小，另一个变量对应值也随之增大，两列变量变化方向相反。如学生学习能力水平与其解题时间的关系；运动员赛跑与所用时间之间的相关；学生学习能力与识记所用时间之间的相关等。

3、零相关。两变量值的变化方向无规律。如学生的身高与学生成绩的变化关系。

（二）从变量的个数上划分

1、简相关。两个变量之间的相关关系。如在一定年龄阶段，儿童身高与年龄的关系。本课所研究的都是简相关。

2、复相关。一个变量与两个或两个以上变量间的相关关系。如教师教学的成效与教师思维能力、教学方法、学生的学习准备情况之间的关系。

（三）从变量相互关系的程度上划分

1、高度相关。即两个变量相互联系非常密切。如大学生的学习成绩和智商的关系。当两个变量变化关系达到一一对应的密切程度时，数量变化就是确定性关系了，则称为完全相关。

2、低度相关。即两变量存在相互联系，但其关系并不密切。

三、相关散布图

它是表示两种事物之间的相关性及联系的模式。以直角坐标的横轴表示 x 列变量，纵轴表示 y 列变量，在相关的两变量对应值的垂直相交处画点，构成相关散布图。如图5-1。

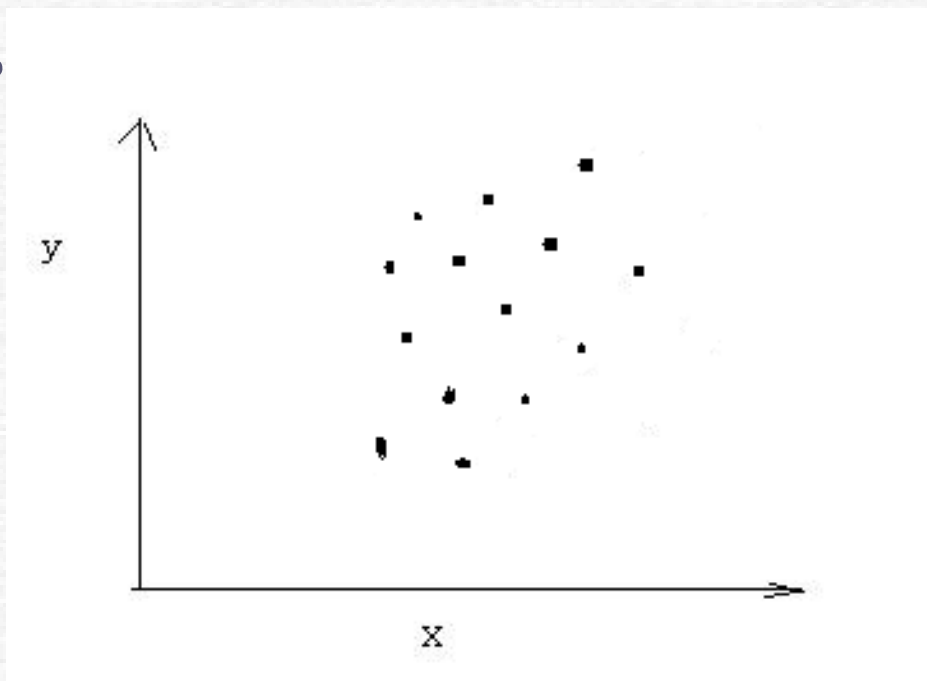


图5-1 散布图

相关散布图的用途：

1、判断相关是否直线式。

当两变量之间呈曲线趋势，其相关散布图呈弯月状，说明两变量之间是非线性关系，如图5-2(a)。

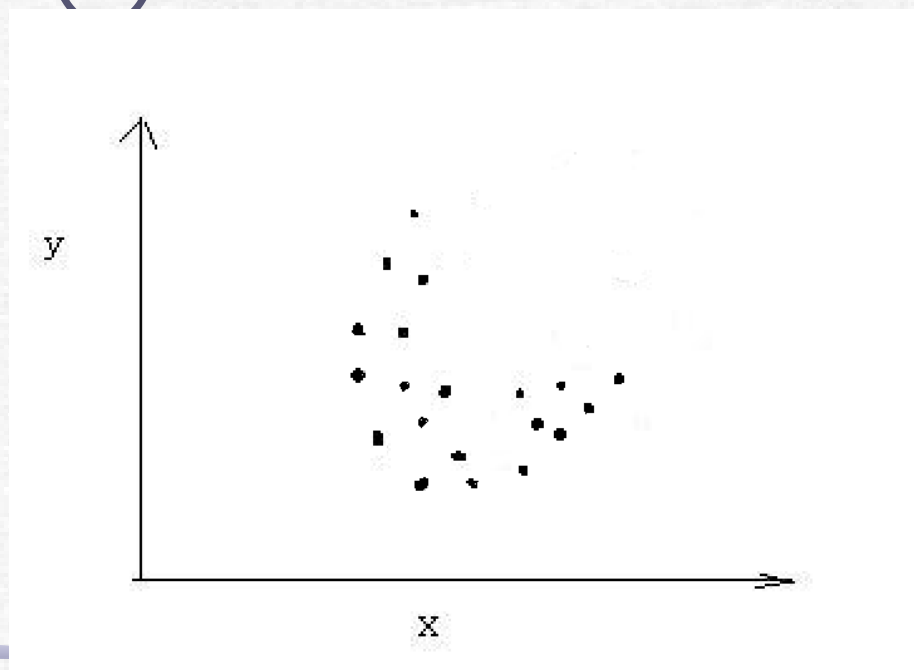


图5-2(a) 曲线相关

当两变量间呈线性趋势，其相关散布图是椭圆形，说明两变量之间是线性关系，称为直线相关，如图5-2(b)。

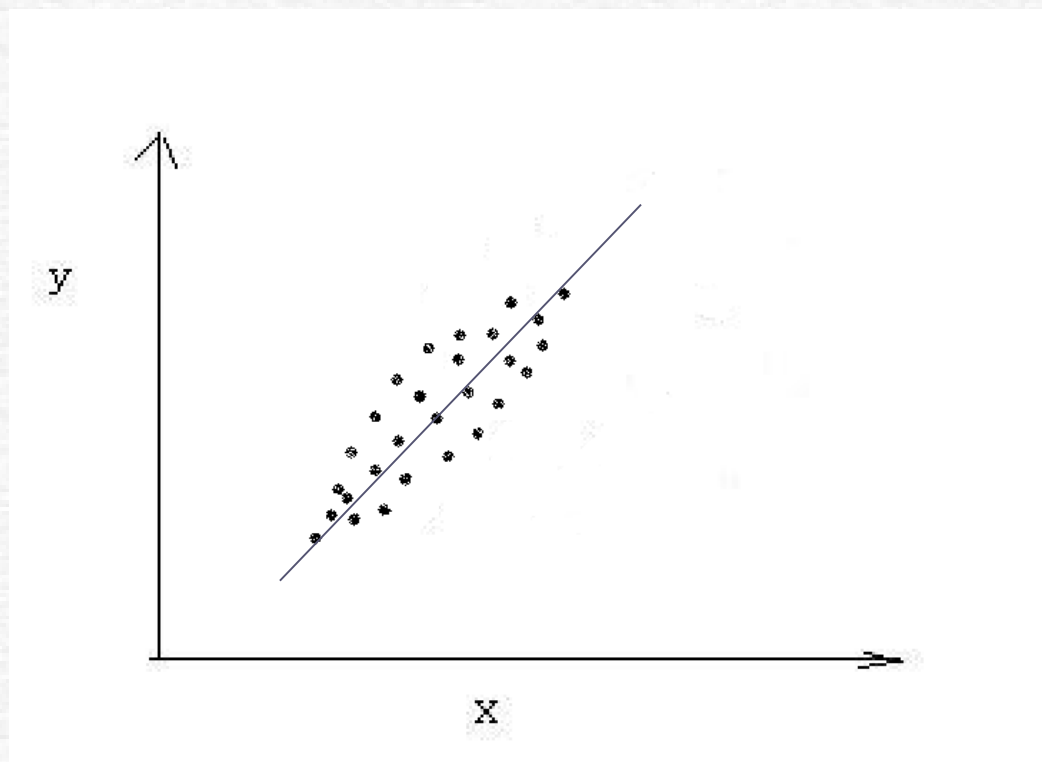


图5-2(b) 直线相关

2、判断相关密切程度高低

相关散布图的形状和疏密，反映着相关程度的高低。如图5-3（a），散布图的椭圆形状较狭长，称为高度相关。

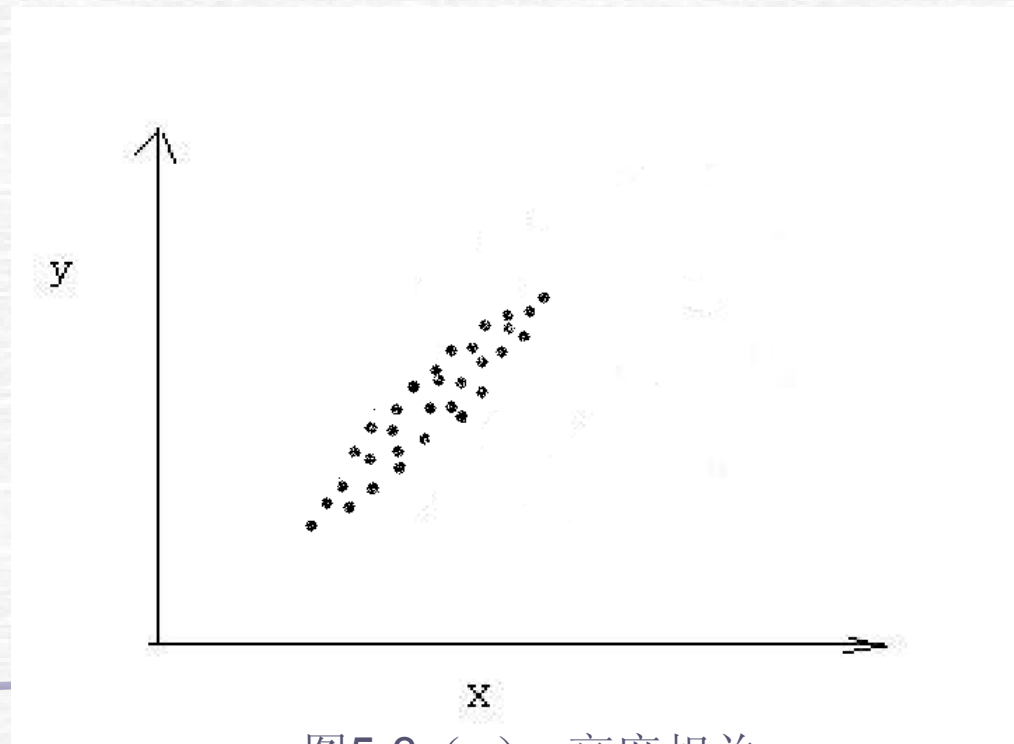


图5-3（a） 高度相关

如果散布图的椭圆形状比较粗，称为低度相关。如图5-3（b）。

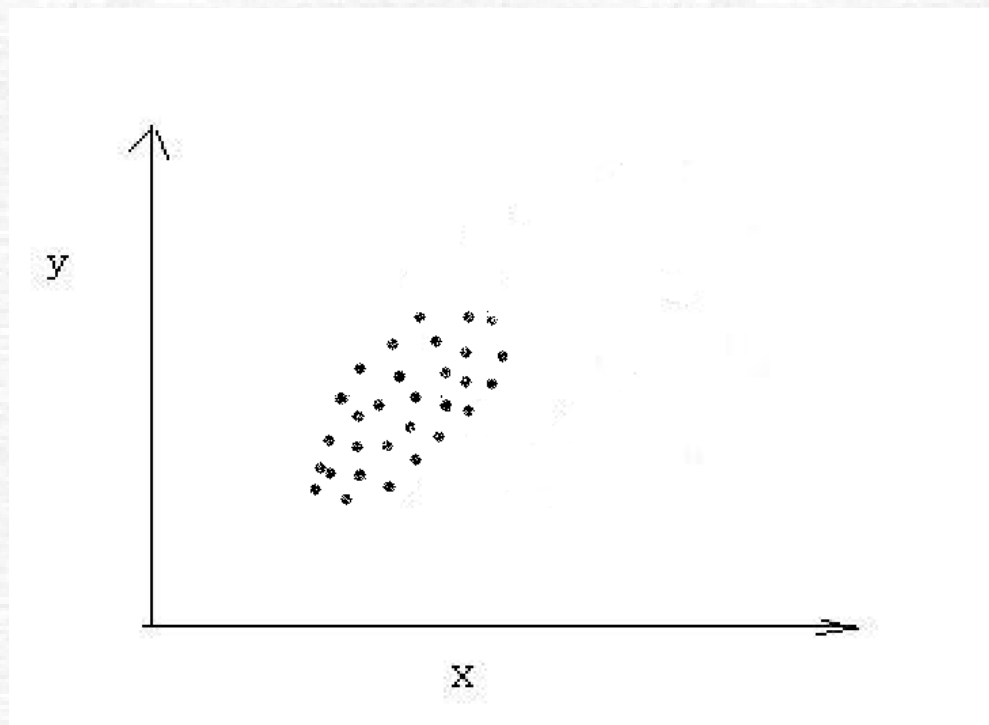


图5-3b 低度相关

3、判断相关变化方向

正相关：散布点主要位于一、三象限。如图5-4(a), 即一个变量增加（或减少），另一个变量也增加（或减少）。

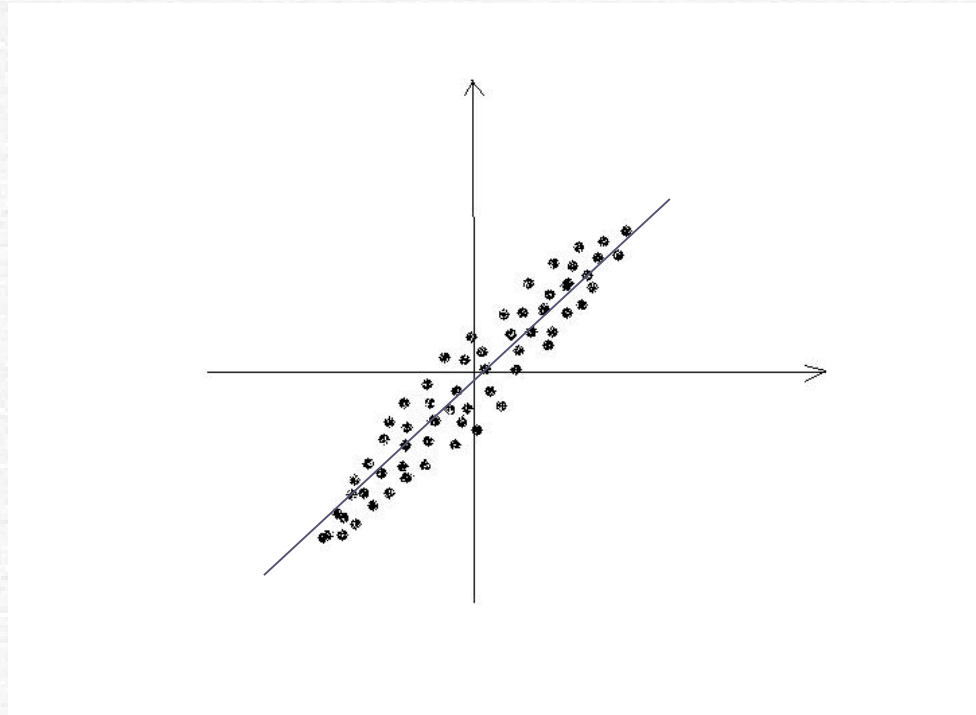


图5-4 (a) 正相关

负相关：若散布点主要位于二、四象限，如图5-4（b），即一个变量增加（或减少），另一个变量也减少（或增加）。

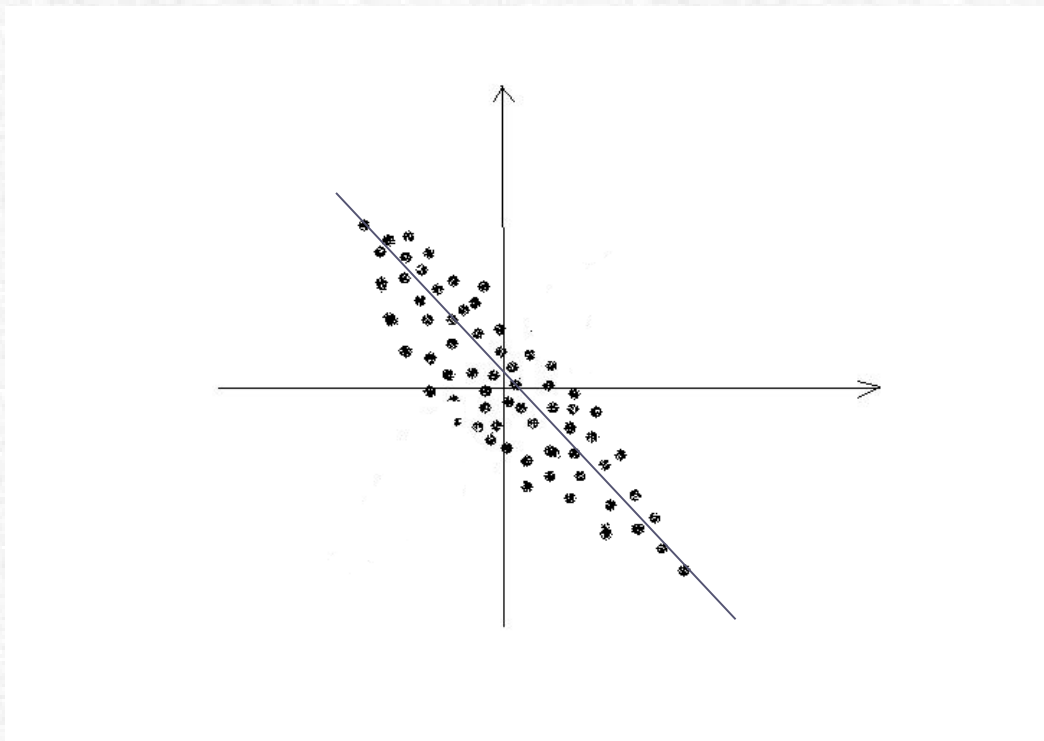


图5-4（b） 负相关

零相关：散布点的变化无一定规律。如图5-4（c）。

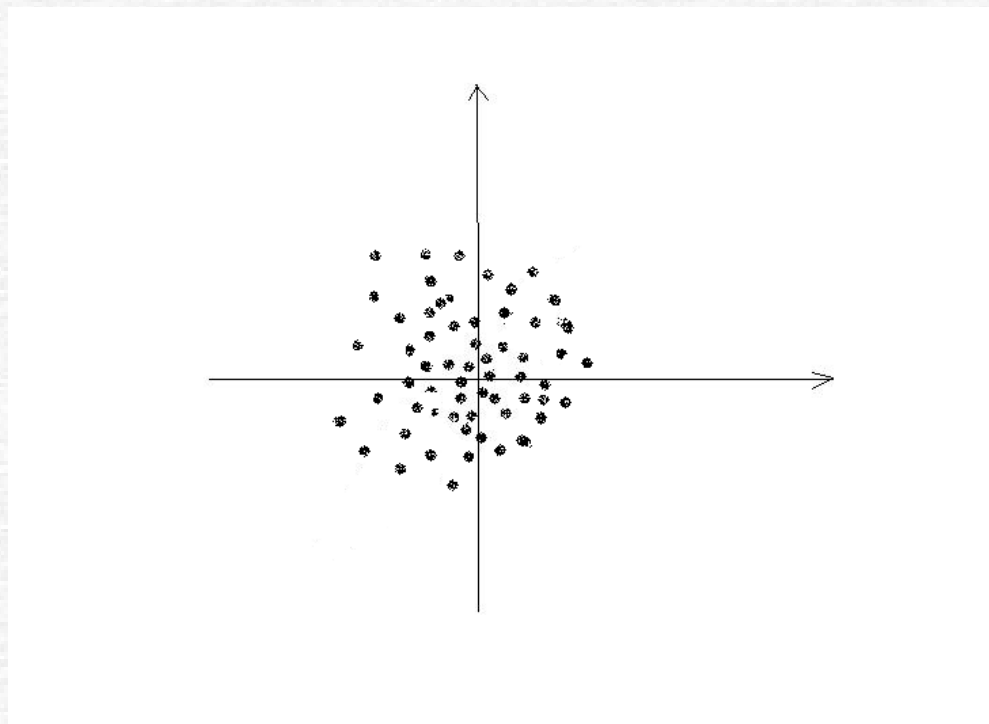
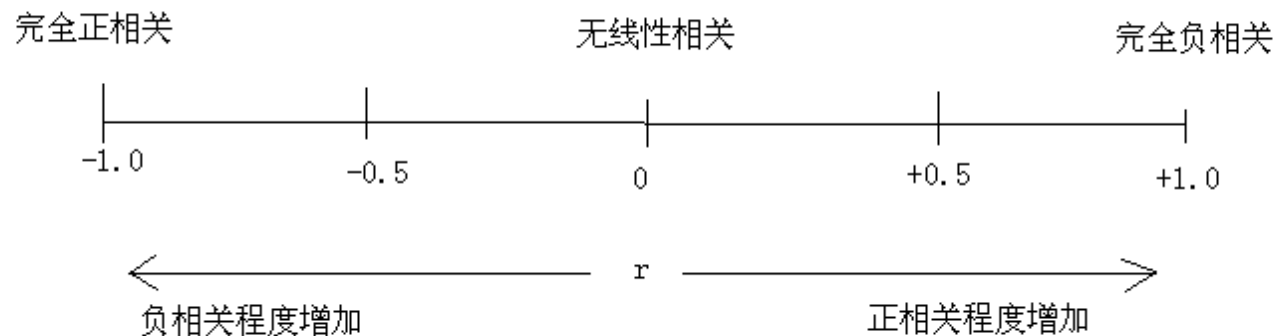


图5-4（c）零相关

四、相关系数

通过相关散布图的形状，我们大概可以判断变量之间相关程度的强弱、方向和性质，但并不能得知其相关的确切程度。为精确了解变量间的相关程度，还需作进一步统计分析，求出描述变量间相关程度与变化方向的量数，即相关系数。总体相关系数用 ρ （读“柔”）表示，样本相关系数用 r 表示。

相关系数 r 的取值范围是 $-1 \leq r \leq 1$ ，一般取小数点后两位。 r 的正负号表明两变量间变化的方向； $|r|$ 表明两变量间相关的程度， $r > 0$ 表示正相关， $r < 0$ 表示负相关， $r = 0$ 表示零相关。 $|r|$ 越接近于1，表明两变量相关程度越高，它们之间的关系越密切。



附加说明:

(1) 两变量间存在相关, 仅意味着变量间有关联, 并不一定是因果关系。

(2) 相关系数不是等距的测量单位。

r 是一个比值, 不是由相等单位度量而来, 不能进行加、减、乘、除运算。如

$r_1=0.25, r_2=0.5, r_3=0.75$, 不能认为 $r_1=r_3-r_2$

或 $r_2=2r_1$ 。

(3) 相关系数受变量取值区间大小及观测值个数的影响较大。

变量的取值区间越大，观测值个数越多，相关系数受抽样误差的影响越小，结果就越可靠，如果数据较少，本不相关的两列变量，计算的结果可能相关，如学生的身高与学习成绩。本书所举例题，数据较少，仅为说明计算方法时较方便。

(4) 相关系数在特定情况下使用才具有意义。

如高中生身高与体重的相关系数用在儿童身上就没有意义。

(5) 通过实际观测值计算的相关系数，须经过显著性检验确定其是否有意义。

表5-0 $|r|$ 的取值与相关程度

$ r $ 的取值范围	$ r $ 的意义
0.00-0.19	极低相关
0.20-0.39	低度相关
0.40-0.69	中度相关
0.70-0.89	高度相关
0.90-1.00	极高相关

*如何判断两个变量的相关性

- (1) 找出两个变量的正确相应数据。
- (2) 画出它们的散布图（散点图）。
- (3) 通过散布图判断它们的相关性。
- (4) 给出相关（ r ）的解答。
- (5) 对结果进行评价和检验。

进一步阅读资料:

- 1.程俊玲等.中小学教师工作压力状况及相关因素调查研究.教育理论与实践,2004(6)
- 2.陈小异.大学生自我容纳与人格特征的相关研究.统计教育,2004(4)

在线资源:

- 1.SPSS10.0 相关分析, [http://www.bioon.com/biology/class45/spss/200407/55214.html\(55175.html\)](http://www.bioon.com/biology/class45/spss/200407/55214.html(55175.html))
- 2.<http://www.fjmu.edu.cn/news/stat/jbtj2.htm#g4>

第二节 积差相关系数

一、概念及适用条件

(一) 概念

积差相关，又称积矩相关（或皮尔逊（英国）相关）。公式为

$$r = \frac{\sum xy}{N \bullet \sigma_x \bullet \sigma_y} \quad (5.1)$$

式中：x为X变量的离差，即 $x = X - \bar{X}$ ；y为Y变量的离差，即 $y = Y - \bar{Y}$ ； σ_x 为X的标准差； σ_y 为Y的标准差。

（二）适用条件

- 1、两变量均应由测量得到的连续变量。
- 2、两变量所来自的总体都应是正态分布，或接近正态的单峰对称分布。
- 3、变量必须是成对的数据。
- 4、两变量间为线性关系。

二、计算方法

(一) 基本公式计算法 步骤:

- (1) 计算 $\bar{X}$, $\bar{Y}$, σ_x , σ_y ;
- (2) 计算 xy ;
- (3) 计算 $\sum xy$;
- (4) 将上述数据代入公式 5.1, 求得 r .

例1 某学校为调查学生学习各科目之间的能力迁移问题，随机抽取10名学生的政治与语文成绩见表5-1，请计算其相关程度。

学生序号	X (政治)	Y (语文)	x	y	xy
1	74	82	-1.6	-1.7	2.72
2	71	75	-4.6	-8.7	40.02
3	80	81	4.4	-2.7	-11.88
4	85	89	9.4	5.3	49.82
5	76	82	0.4	-1.7	-0.68
6	77	89	1.4	5.3	7.42
7	77	88	1.4	4.3	6.02
8	68	84	-7.6	0.3	-2.28
9	74	80	-1.6	-3.7	5.92
10	74	87	-1.6	3.3	-5.28
$\bar{x}$ σ	75.6	83.7			
	4.454211	4.33705			91.8

解：依表5-1的资料，计算结果为

$$\begin{aligned} r &= \frac{\sum xy}{N \bullet \sigma_x \bullet \sigma_y} \\ &= \frac{91.80}{10 \times 4.454 \times 4.337} \\ &= 0.475 \end{aligned}$$

即 10名学生的政治与语文成绩的相关程度为0.475。

（二）原始数据计算法

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

课后练习：用原始数据计算法计算例5-1。

第三节 其他相关系数

一、等级相关系数

（一）斯皮尔曼等级相关

1、概念及适用条件

（1）概念

两变量是等级测量数据，且总体不一定呈正态分布，样本容量也不一定大于30，这样两变量的相关，称为等级相关（斯皮尔曼相关）

(2) 适用条件

①两变量的资料为等级测量数据，且具有线性关系。

②连续变量的测量数据，按其大小排成等级，亦可用等级相关计算。

③不要求总体呈正态分布。

2、计算方法

$$r_R = 1 - \frac{6 \sum D^2}{N(N^2 - 1)} \quad (5.4)$$

式中：D为两变量每对数据的等级之差；N表示样本容量。

计算步骤:

- (1) 计算两变量等级之差 D ;
- (2) 计算 D^2 ;
- (3) 计算 ΣD^2 ;
- (4) 代入公式 (5.4), 求得 r_R

例3 求10名学生的语文成绩与阅读能力成绩之间的等级相关系数。

表5-3 10名学生的语文成绩与阅读能力成绩相关计算表

序号	X（语文等级）	Y（阅读等级）	D	D ²
1	8	8	0	0
2	6	7	-1	1
3	5	4	1	1
4	3	2	1	1
5	2	1	1	1
6	4	5	-1	1
7	7	6	1	1
8	9	10	-1	1
9	1	3	-2	4
10	10	9	1	1
Σ				12

解：将有关数据代入公式（5.4）得

$$\begin{aligned} r_R &= 1 - \frac{6 \sum D^2}{N(N^2 - 1)} \\ &= 1 - \frac{6 \times 12}{10(100 - 1)} = 0.927 \end{aligned}$$

如果求相关的是连续变量，计算时先把两组数据分别按大小排成等级，最大值取为1等，其它类推。若出现相同的等级分数时，可用它们所占等级位置的平均数作为它们的等级。

例4 某校为了研究学生自学能力与学业成绩之间的关系，随机抽取10名学生的自学能力和学科成绩，见表5-4，求其相关系数。

表5-4 10名学生的自学能力和学科成绩相关计算表

序号	X（能力）	等级	Y（成绩）	等级	D	D ²
1	90	3.5	88	4	-1	0.25
2	85	7	80	6	1	1
3	70	10	80	6	4	16
4	85	7	79	8	-1	1
5	90	3.5	95	2.5	1	1
6	80	9	70	10	-1	1
7	85	7	75	9	-2	4
8	100	1	98	1	0	0
9	87	5	80	6	-1	1
10	92	2	92	2.5	-1	0.25
Σ						25.5

解：

$$\begin{aligned} r_R &= 1 - \frac{6 \sum D^2}{N(N^2 - 1)} \\ &= 1 - \frac{6 \times 25.5}{10(100 - 1)} = 0.85 \end{aligned}$$

即学生的自学能力与学习成绩的相关程度为0.85。

（二）肯德尔和谐系数

1、概念及适用条件

（1）概念

当多个变量值以等级顺序表示时，这几个变量之间的一致性程度，称为肯德尔和谐系数或肯德尔W系数。

（2）适用条件

适用于两列以上等级变量。如了解几个评定者对同一组学生成绩等级评定的一致性程度等。

2、计算方法

它以符号W表示，公式为

$$W = \frac{SS_R}{\frac{1}{12} K^2 (n^3 - n)}$$

式中： SS_R 为R的离差平方和，

$$SS_R = \sum R^2 - \left(\sum R \right)^2 / n;$$

K 为等级变量的列数或评价者数目；

n 为被评对象的数目。

计算步骤：略

例5 某评价小组7人依据已确定的4项内容对某教师打分，将分数转换为等级后的结果见表5-5，求这7人对该教师评价意见的一致性程度。

表5-5 7人评价某教师意见资料表

	一	二	三	四		
1	2.5	4	2.5	1		
2	3.5	3.5	1.5	1.5		
3	2.5	2.5	1	4		
4	4	2	2	2		
5	3	4	1.5	1.5		
6	1	2	3	4		
7	2	4	2	2		
R	18.5	22	13.5	16	$\Sigma =$	70
R^2	342.25	484	182.25	256	$\Sigma R^2 =$	1264.5

解：将上述数据代入公式（5.5）中得

$$\begin{aligned} W &= \frac{SS_R}{\frac{1}{12} K^2 (n^3 - n)} \\ &= \frac{1264.5 - \frac{(70)^2}{4}}{\frac{1}{12} \times 7^2 (4^3 - 4)} \\ &= 0.161 \end{aligned}$$

实际上，当出现相同等级时，应校正W系数，其校正公式为

$$W_c = \frac{SS_R}{\frac{1}{12} K^2 (n^3 - n) - K \sum C}$$

$$\text{式中} : \sum C = \sum \frac{(t^3 - t)}{12}$$

t 为相同等级数 .

例5中第一个人评的有2个等级相同，第二个人评的有2个3.5和2个1.5等级...所以 $\sum C$ 为

$$\sum C = \frac{2^3 - 2}{12} + \frac{2^3 - 2}{12} + \frac{2^3 - 2}{12} + \frac{2^3 - 2}{12} + \frac{3^3 - 3}{12} + \frac{2^3 - 2}{12} + \frac{3^3 - 3}{12}$$

$$= 6.5$$

$$W_c = \frac{SS_R}{\frac{1}{12} K^2 (n^3 - n) - K \sum C}$$

$$= \frac{39.5}{\frac{1}{12} \times 7^2 (4^3 - 4) - 7 \times 6.5}$$

$$= 0.198$$

二、点二列相关

（一）概念及适用条件

1、概念

两列变量一列是正态连续变量，另一列是二分变量，描述这两个变量之间的相关，称为点二列相关。

2、适用条件

一列是正态连续变量，另一列是二分变量（如男与女，对与错等）。

（二）计算方法

点二列相关系数以表示 r_{pb} ，公式为

$$r_{pb} = \frac{\bar{X}_p - \bar{X}_q}{\sigma_x} \cdot \sqrt{p \cdot q}$$

式中： p 为二分变量中某一项所占比例； q 为二分变量中另一项所占比例， $p+q=1$ ； $\bar{X}_p$ 为二分变量中比例为 p 部分所对应的连续变量的平均数； $\bar{X}_q$ 为二分变量中比例为 q 部分所对应的连续变量的平均数。 σ_x 为连续变量的标准差。

例6 随机抽取某区初二数学期末考试卷15份，试计算第二题的得分与总分相一致的程度（即试题的区分度，它是衡量试题鉴别能力的指标值）。数据见表5-6。

表5-6 数据表

学生序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
第二题得分	1	0	1	0	0	0	1	1	0	1	1	1	1	1	1
总分	69	71	79	86	87	87	81	89	81	92	91	93	92	93	93

解：（1）求答对第二题的比率 p 和答错的比率 q ：

$$p=10/15=0.67$$

$$q=1-p=0.33$$

（2）求 $\bar{X}_p$ 和 $\bar{X}_q$ ，分别为答对和答错第二题学生成绩的平均数：

$$\bar{X}_p = 87.2(\text{分})$$

$$\bar{X}_q = 82.4(\text{分})$$

（3）求 σ_x ，所有考生的总分的标准差：

$$\sigma_x = 7.597(\text{分})$$

将上述数据代入公式（5.7），可得

$$r_{pb} = \frac{87.2 - 82.4}{7.597} \times \sqrt{0.67 \times 0.33}$$
$$= 0.297$$

即该试卷第二题的区分度为0.297。

三、 ϕ 相关

(一) 概念及适用条件

1、概念

当两变量均为二分变量时，描述这两个变量之间的相关，称为 ϕ 相关。

2、适用条件

两变量均为二分变量；或资料整理为 2×2 列联表一形式。

(二) 计算方法

ϕ 相关以符号 r_ϕ 表示，其计算公式为

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

式中：a、b、c、d分别表示四格表中的实际次数，如表5-7所示。

表5-7 2×2列联表

	Y1	Y2	Σ
X1	a	b	a+b
X2	c	d	c+d
Σ	a+c	b+d	a+b+c+d

例7 某区为研究性别与学习数学的关系，随机抽取100名学生，以数学成绩85分为线进行分类，求性别与数学成绩间的相关系数。

表5-8 100名学生成绩分布表

	85分以上	85分以下	Σ
男生	31	18	49
女生	29	22	51
Σ	60	40	100

$$r_{\phi} = \frac{31 \times 22 - 18 \times 29}{\sqrt{(31 + 18)(31 + 29)(18 + 22)(29 + 22)}} \\ = 0.065$$

即性别与数学成绩间的相关系数为0.065。

第四节 解释与应用相关系数时应注意的问题

略：[见第一节 四](#)

作业：

1、某小组10名学生物理期中与期末考试成绩如下，请用相关散点图分析其成绩动态变化情况并提出指导意见。

序 号	1	2	3	4	5	6	7	8	9	10
期 中	78	74	60	42	53	63	70	90	54	82
期 末	85	69	76	53	70	59	82	95	60	67

2、某校为研究高中模拟考试与高考之间的相关程度，随机抽取为**20**名学生模拟考试与高考的数学成绩如下，请计算其相关系数。

序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
模拟考	79	80	84	81	78	69	90	111	83	121	72	76	90	61	83	72	69	100	81	104
高考	85	69	76	53	70	59	82	95	60	67	70	79	84	62	80	76	72	100	85	100

3、4位教师对5名学生的论文水平按等级评定，结果如下表，求评定结果的一致性程度。

教师 论文		教师			
		1	1	1	1
1	1	1	2	3	2
2	2	3	3	2	1
3	3	4	5	4	5
4	4	5	4	5	4
5	5	2	1	1	3