第四章 相关分析:数据一致性测量

4.1 什么是相关?

4.2 积差相关(皮尔逊相关)

4.3 等级相关

4.4 质与量的相关

4.5 点二列相关和二列相关

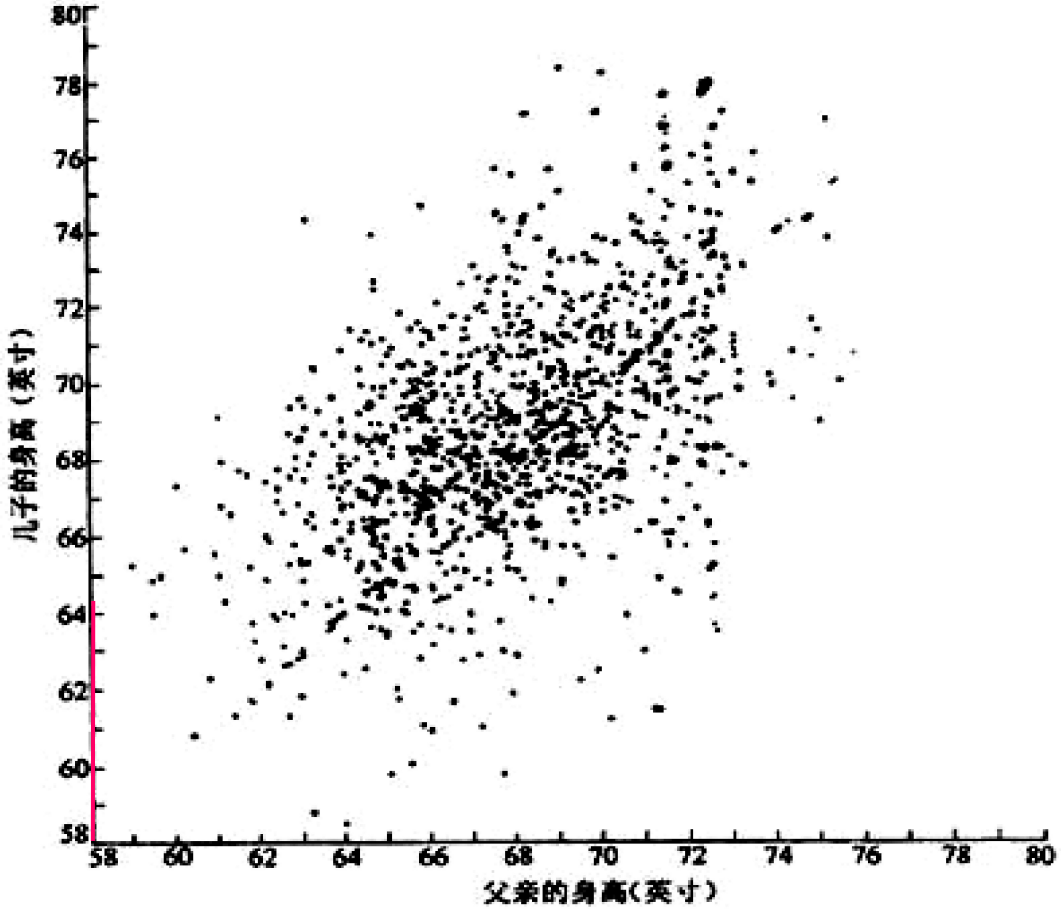
比较下面两种变量间的依存关系

函数关系
（确定性关系）

1. 购物费用=单价×数量

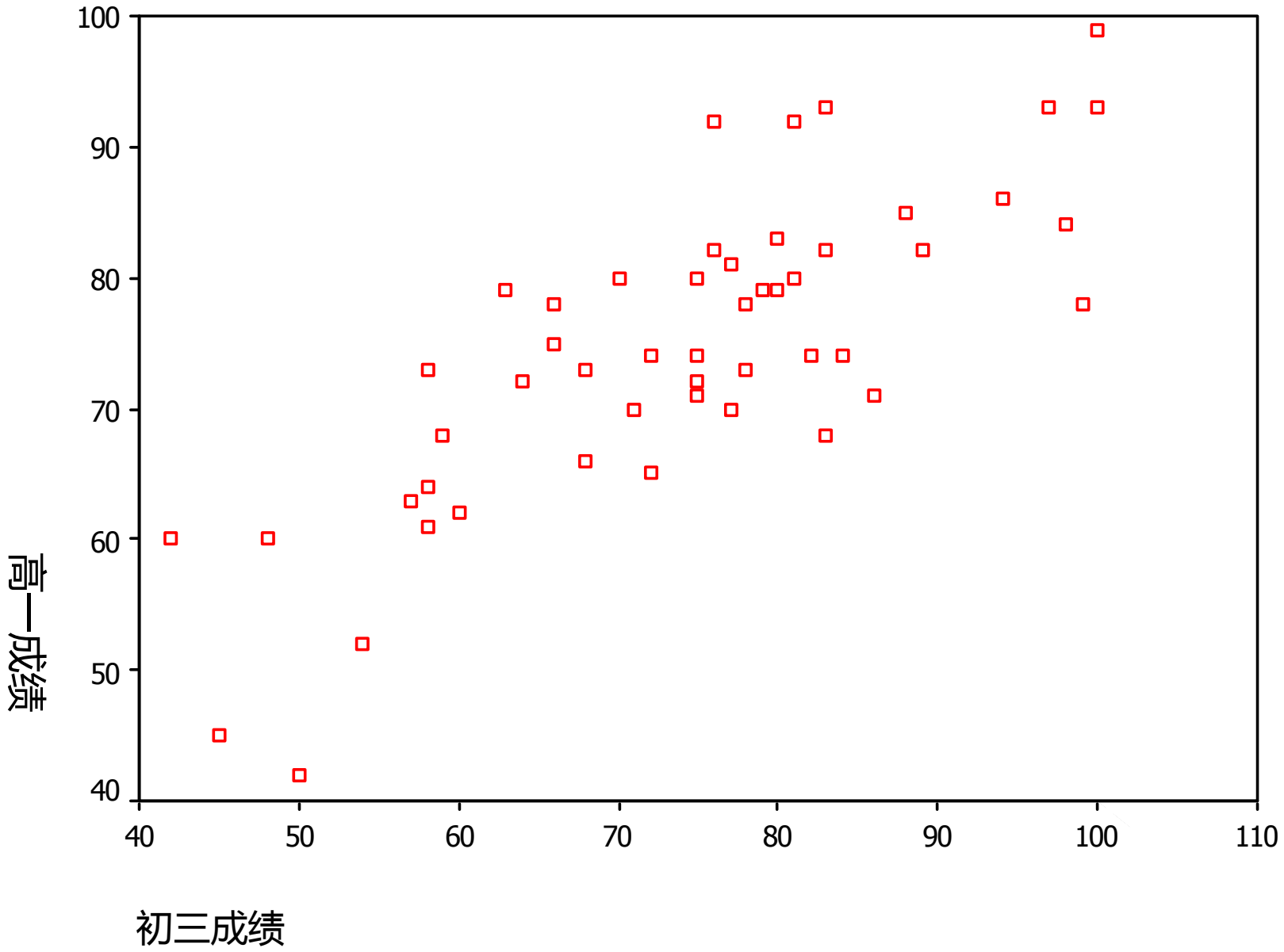
相关关系
（非确定性关系）

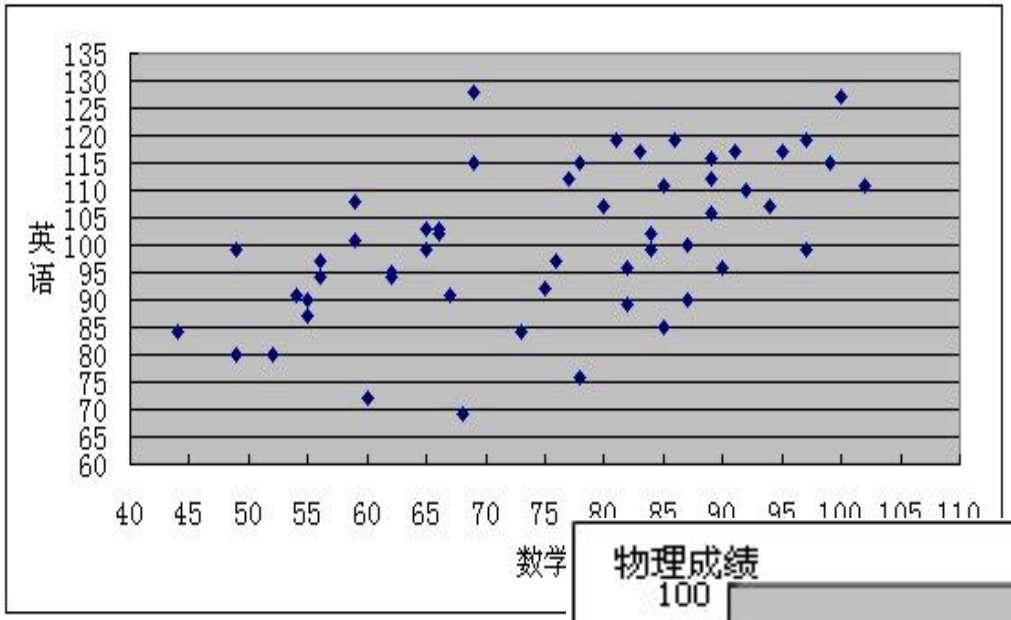
2. 个体的身高与其体重。

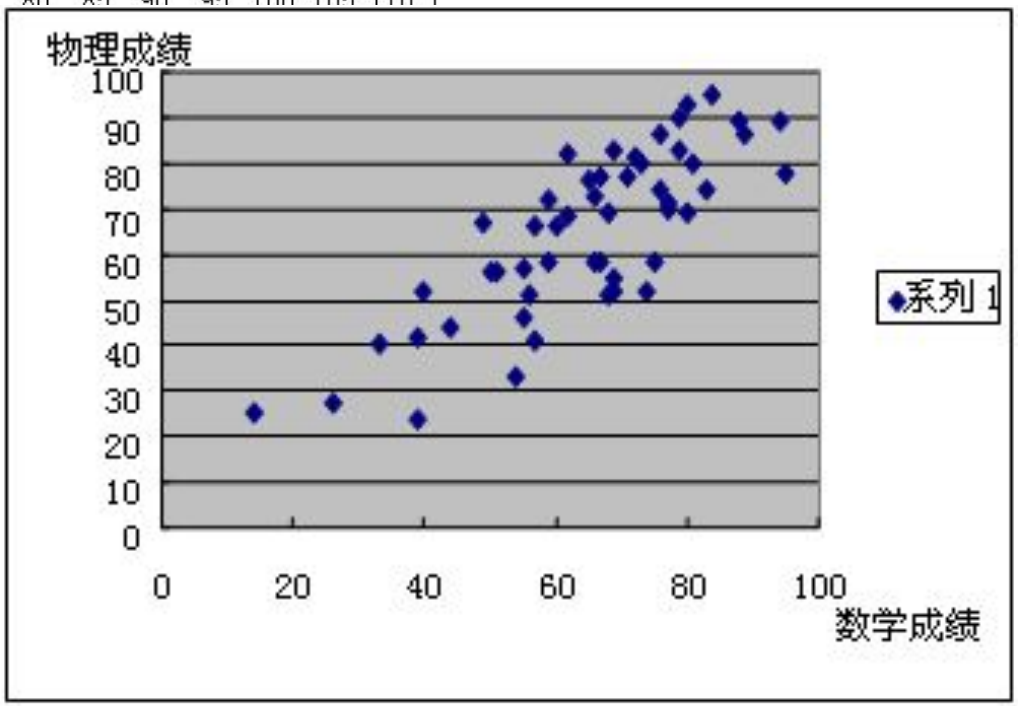


为了研究父亲与成年儿子身高之间的关系，卡尔.皮尔逊测量了1078对父子的身高。用水平轴X上的数代表父亲身高，垂直轴Y上的数代表儿子的身高，1078个点所形成的图形是一个散点图。它的形状象一块橄榄状的云，中间的点密集，边沿的点稀少，其主要部分是一个椭圆。

50名同学初三和高一成绩的散点图



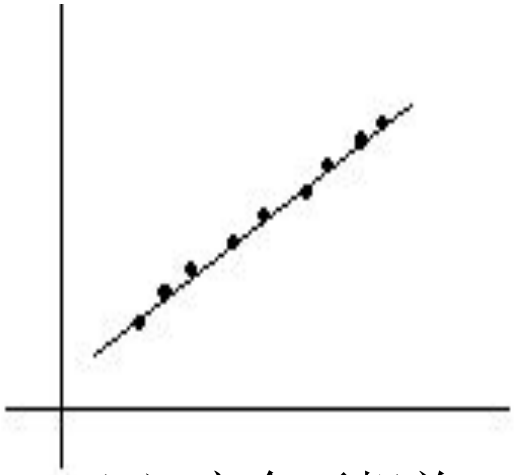

英语
135
130
125
120
115
110
105
100
95
90
85
80
75
70
65
60
40 45 50 55 60 65 70 75 80 85 90 95 100 105 110
数学


物理成绩
100
90
80
70
60
50
40
30
20
10
0
0 20 40 60 80 100
数学成绩
系列 1

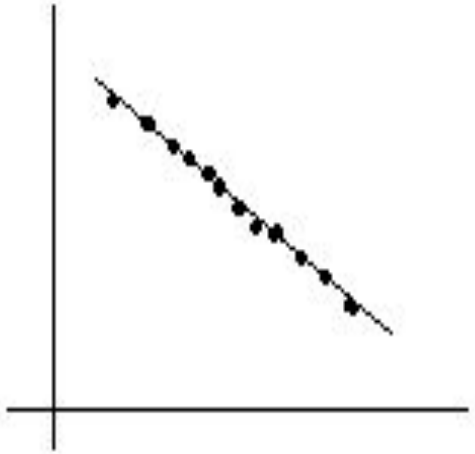
第 4 题

相关关系的其他例子

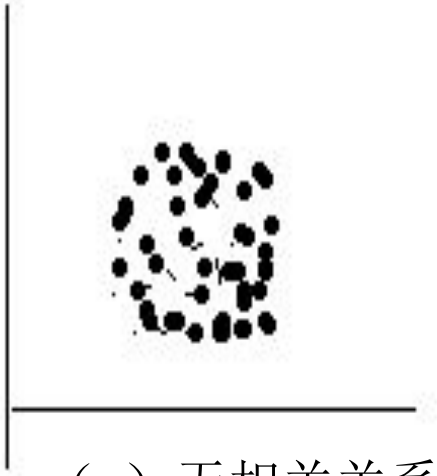
- 商品销售额(y)与广告费支出(x)之间的关系
- 教育发展(y)与经济增长(x)之间的关系
- 收入水平(y)与受教育程度之间的关系(x)
- 情绪智力与社会适应
- 完美主义倾向与抑郁情绪
- 学习动机与学习成绩



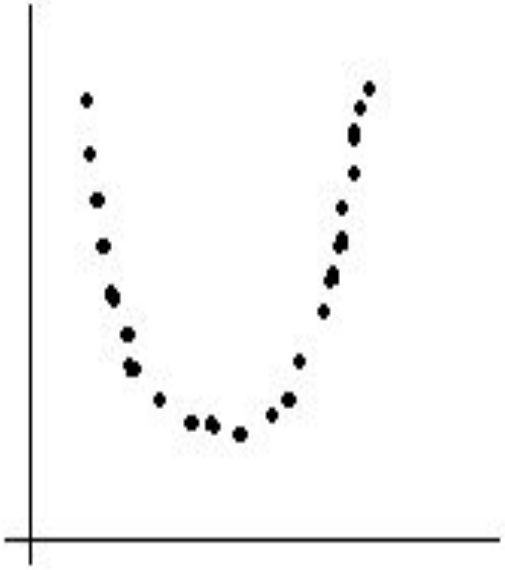
（a）完全正相关



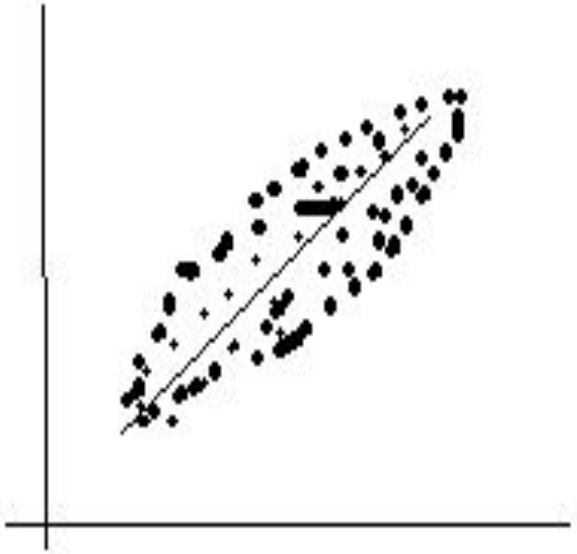
（b）完全负相关



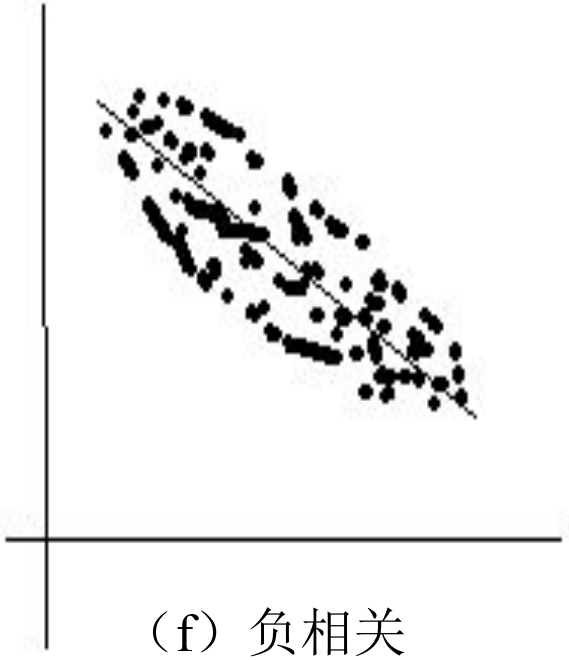
（c）无相关关系



（d）非线性关系



（e）正相关



（f）负相关

4.1 导致变量间相关的原因

（1）因果关系：一种现象是另一种现象的因，而另一种现象则是果。如：

下雨地上会湿。

两类现象在发展变化的方向与大小方面存在一定关系，但有时不能确定两者中哪个是因，哪个是果，或者说可能互为因果。

（2）共变关系：两事物本身之间没有直接的关系，但它们都受第三种现象的影响而发生变化。如：

婴儿身高和树苗高度的关系（均受时间t的影响）；

拥有金表和长寿间的关系（均受富足程度的影响）。

相关（统计学意义）：指具有相关关系的不同现象之间的关系程度。相关分三种情况：

（1）正相关：两列变量变动方向相同。同时增大，同时减少。

如身高与体重的关系。

（2）负相关：两列变量变动方向相反。一个增大，另一个却减少；反之亦然。如年龄越大，走路速度越慢。

（3）零相关：两列变量之间无关系。如学习成绩与身高的关系。

2、相关系数

相关系数是两列变量相关程度的数字表现形式。样本相关系数用r表示；总体相关系数用ρ表示。 ρ，r∈[-1.00，1.00]

（1）完全相关：-1.00或1.00，说明两个变量之间为确定关系；

（2）不完全相关：|ρ，r |<1;

（3）不相关：当相关系数在0附近时，说明两个变量之间毫无关系。

正相关时，相关系数为正，取值在0~1之间；负相关时，相关系数为负，取值在-1~0之间。

4.2 积差相关

也称积矩相关、皮尔逊相关。

1、适用条件

（1）两个变量总体服从正态分布；

（2）两列变量之间的关系应是线性的；

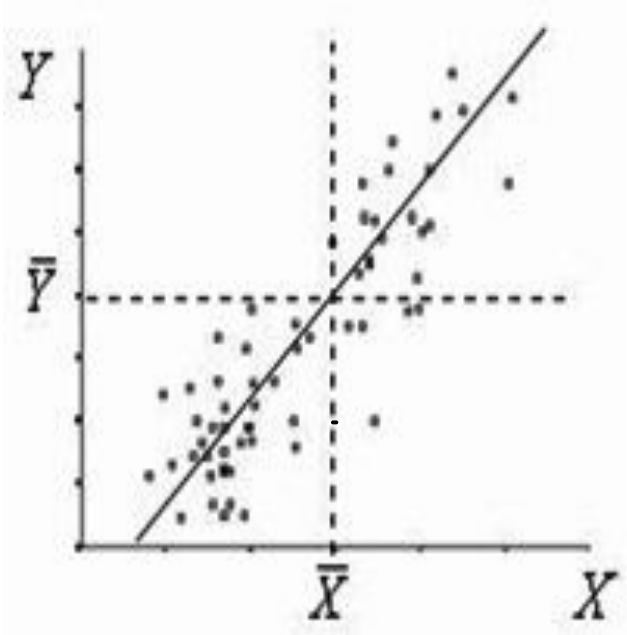
（3）样本容量不低于30。

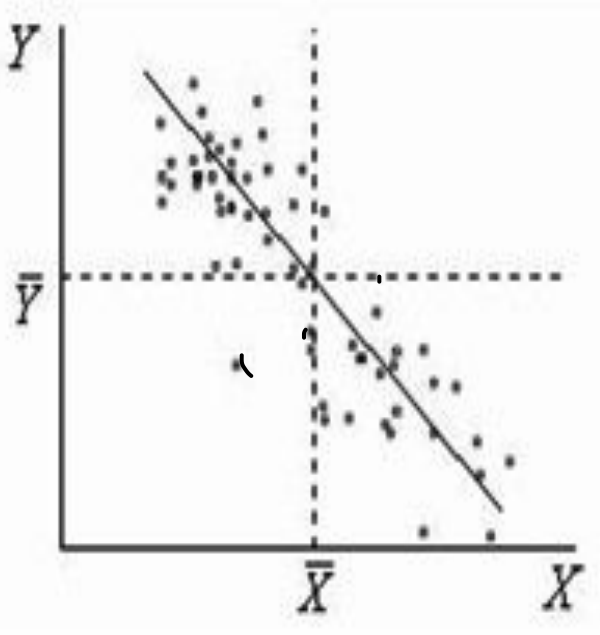
2、计算积差相关系数的公式

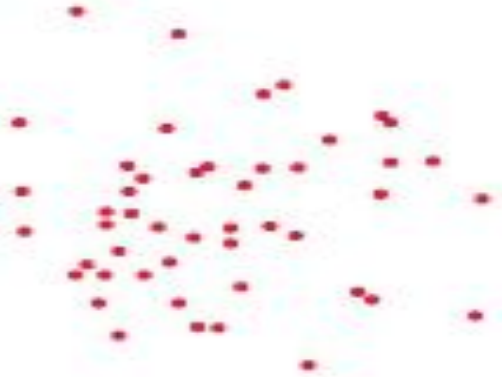
$$r = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n S_X S_Y} = \frac{S_{XY}^2}{S_X S_Y} = \frac{1}{n}\sum Z_X Z_Y$$

$$= \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) \Big/ n}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \Big/ n}\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \Big/ n}}$$

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


Y
$\bar{Y}$
$\bar{X}$
X


Y
$\bar{Y}$
$\bar{X}$
X



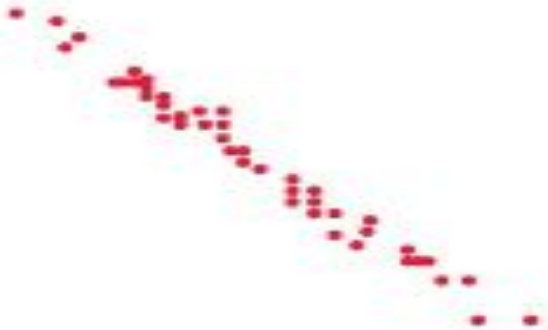
Correlation $r = 0$

Correlation $r = -0.3$

Correlation $r = 0.5$

Correlation $r = -0.7$

Correlation $r = 0.9$



Correlation $r = -0.99$

例4.1 某研究考察数学成绩和化学成绩的相关性。X表示数学测验分数，Y表示化学测验分数。求两门功课成绩的关联性。

X	Y	x	y	x^2	y^2	xy
74	82	-1.6	-1.7	2.56	2.89	2.72
71	75	-4.6	-8.7	22.16	75.69	40.02
80	81	4.4	-2.7	19.36	7.26	-11.88
85	89	9.4	5.3	88.36	28.09	49.82
76	82	0.4	-1.7	0.16	2.89	-0.68
77	89	1.4	5.3	1.96	28.09	7.42
77	88	1.4	4.3	1.96	18.49	6.02
68	84	-7.6	.3	57.76	0.09	-2.28
74	80	-1.6	-3.7	2.56	13.69	5.92
74	87	-1.6	3.3	2.56	10.89	-5.28
756	837	0	0	198.40	188.07	91.80

Σ

解：$\overline{X}=75.6, \overline{Y}=83.7, x=X-\overline{X}, y=Y-\overline{Y},$

$$\sum\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)=91.8, \sum\left(X-\overline{X}\right)^2=198.40,$$

$$\sum\left(Y-\overline{Y}\right)^2=188.07,$$

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

$$=\frac{91.80}{\sqrt{198.40}\times\sqrt{188.07}}=0.48$$

给定显著性水平α＝0.05，自由度df=n-2=10-2=8, $r_{0.05}$=0.632, $r< r_{0.05}$, 所以，没有足够的证据表明该测验有一定的效度。

表4.1 皮尔逊相关系数r的临界值

$n-2$	$\alpha = .05$	$\alpha = .01$
2	.950	.990
3	.878	.959
4	.811	.917
5	.754	.874
6	.707	.834
7	.666	.793
8	.632	.765
9	.602	.735
10	.576	.708
11	.553	.684
12	.532	.661
13	.514	.641
14	.497	.623
15	.482	.606
16	.468	.590
17	.456	.575
18	.444	.561
19	.433	.549
20	.423	.537
25	.381	.487
30	.349	.449
35	.325	.418
40	.304	.393
45	.288	.372
50	.273	.354
60	.250	.325
70	.232	.302
80	.217	.283
90	.205	.267
100	.195	.254

例4.2 计算身高与体重的相关系数

被试	身高（X）	体重（Y）	X^2	Y^2	XY
1	170	50	28900	2500	8500
2	173	45	29929	2025	7785
3	160	47	25600	2209	7520
4	155	44	24025	1936	6820
5	173	50	29929	2500	8650
6	188	53	35344	2809	9964
7	178	50	31684	2500	8900
8	183	49	33489	2401	8967
9	180	52	32400	2704	9360
10	165	45	27225	2025	7425
N=10	ΣX=1725	ΣY=485	ΣX^2= 298525	ΣY^2= 23609	ΣXY= 83891

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

$$= \frac{83891 - \frac{1725 \times 485}{10}}{\sqrt{298525 - \frac{1725^2}{10}}\sqrt{23609 - \frac{485^2}{10}}}$$

$$= \frac{228.5}{\sqrt{962.5}\sqrt{86.5}} = 0.792$$

给定显著性水平α=0.01，自由度df=n-2=10-2=8, $r_{0.05}$=0.765, r> $r_{0.01}$, 所以，有足够的证据表明身高和体重之间有线性关系。

课堂练习

一个公司的销售经理收集到关于该公司销售员的工龄与其年销售额的数据如下：

销售员	工龄（年）	年销售额（千元）
1	1	80
2	3	97
3	4	92
4	4	102
5	6	103
6	8	111
7	10	119
8	10	123
9	11	117
10	13	136

求工龄与年销售额之间的相关系数。

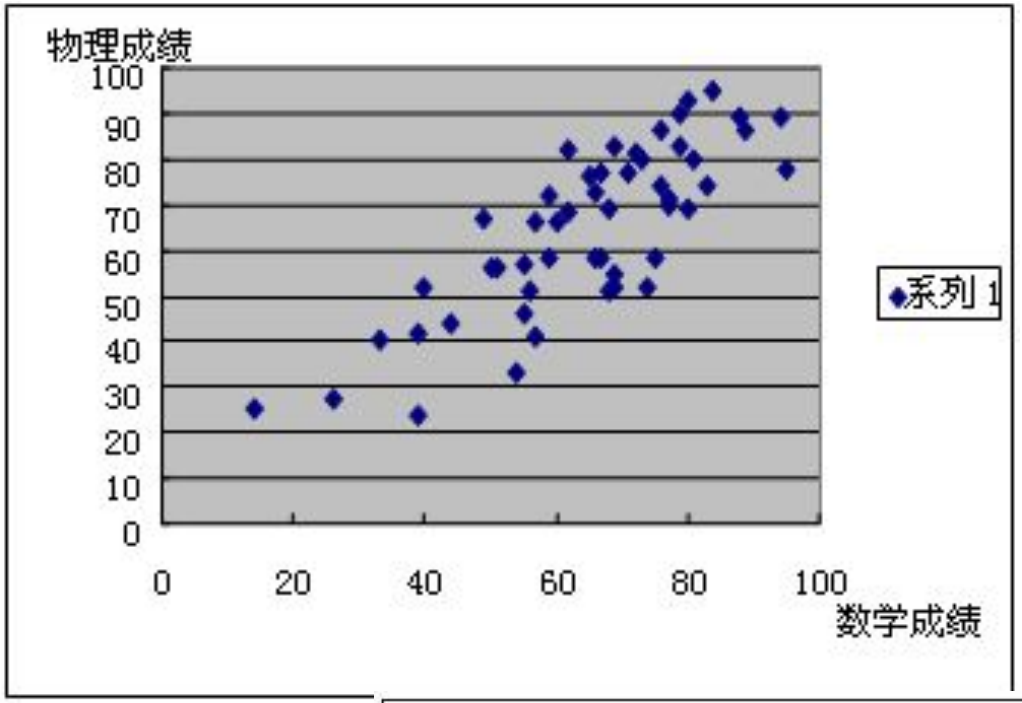
计算相关系数时应注意的问题：

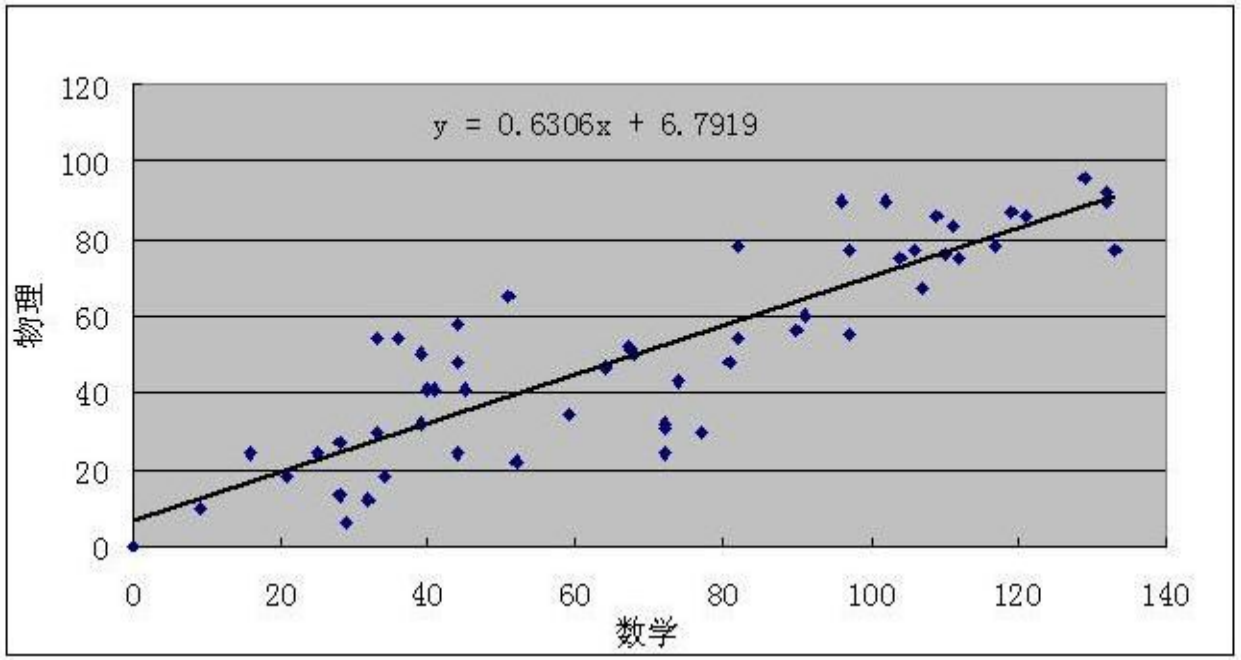
（1）相关系数受样本容量n的影响，以n>=30为宜。如果n很小，样本相关系数 *r* 可能不可靠。

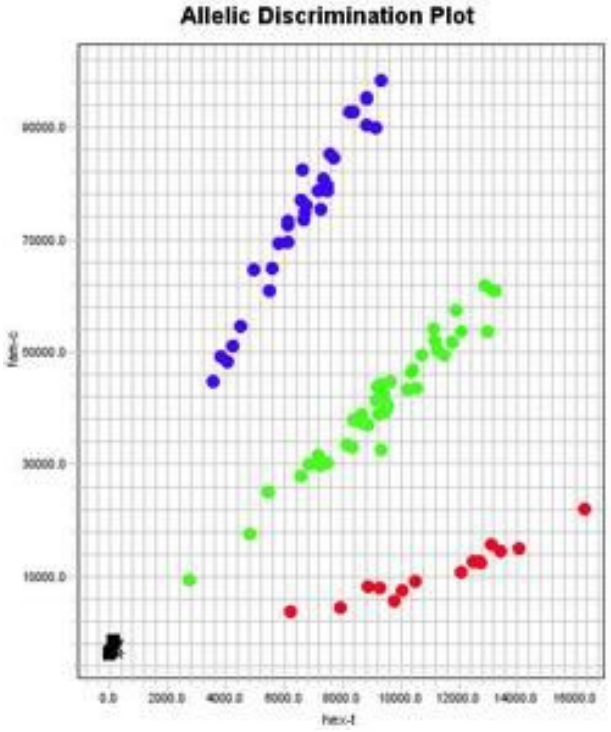
（2）相关系数不是等距量表值，更不是等比量表。不能说r=0.5是r=0.25的两倍。

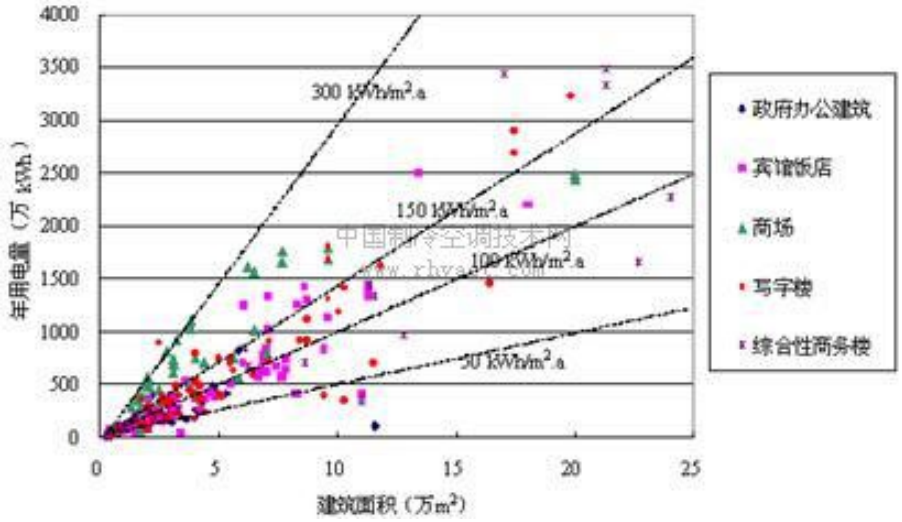
（3）存在相关关系不一定存在因果关系。

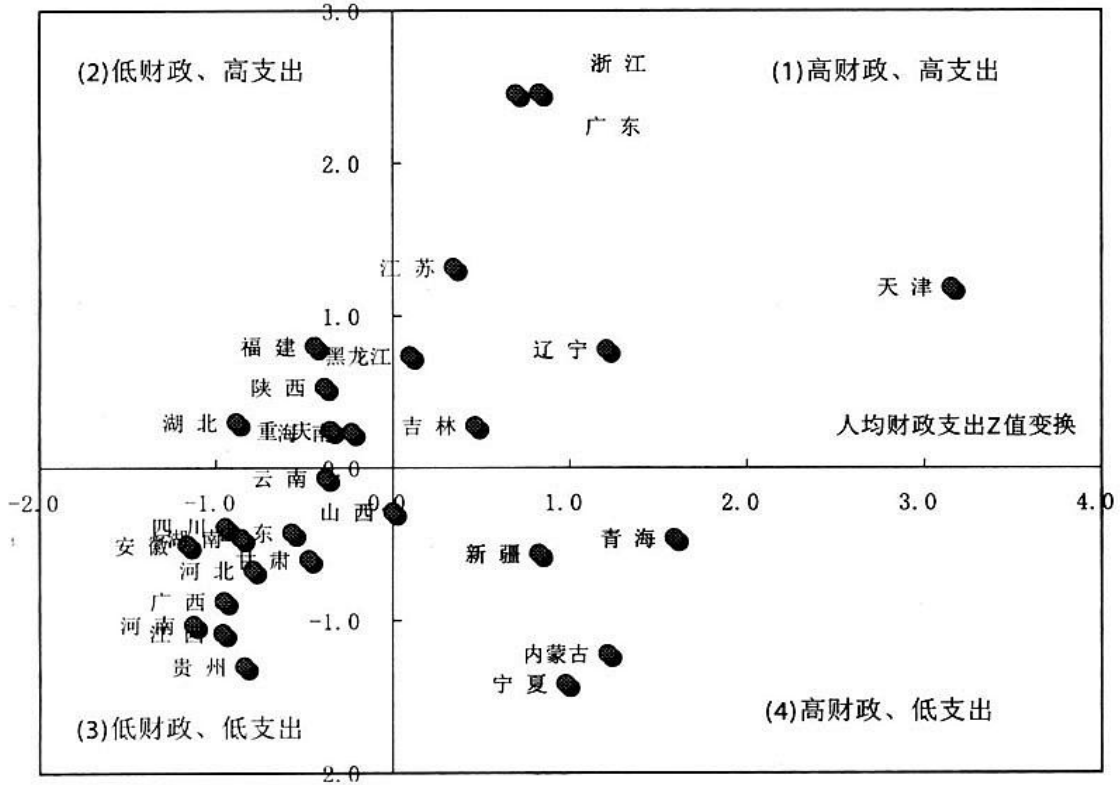
（4）没有线性相关，不一定没有关系，可能是非线性的。

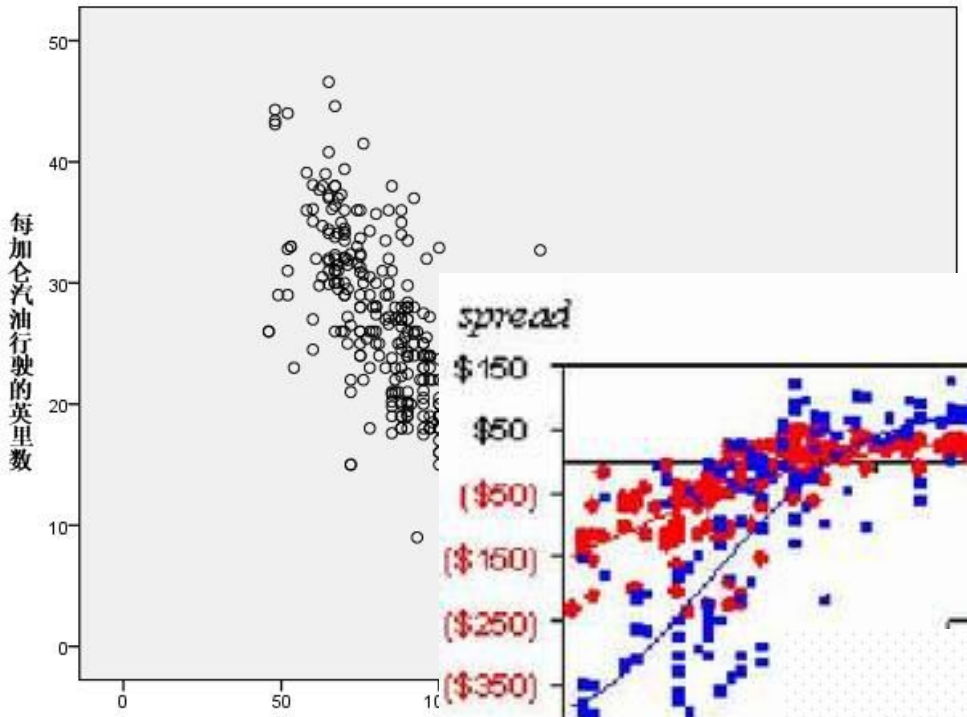

物理成绩
100
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
数学成绩
系列 1


y = 0.6306x + 6.7919
120
100
80
60
40
20
0
物理
0
20
40
60
80
100
120
140
数学


Allelic Discrimination Plot


年用电量（万 kWh）
建筑面积（万m²）
300 kWh/m².a
150 kWh/m².a
100 kWh/m².a
50 kWh/m².a
政府办公建筑
宾馆饭店
商场
写字楼
综合性商务楼


(2)低财政、高支出
(1)高财政、高支出
(3)低财政、低支出
(4)高财政、低支出
人均财政支出Z值变换
浙 江
广 东
江 苏
天 津
福 建
黑龙江
辽 宁
陕 西
湖 北
吉 林
云 南
山 西
青 海
新 疆
内蒙古
宁 夏
贵 州
广 西
河 北



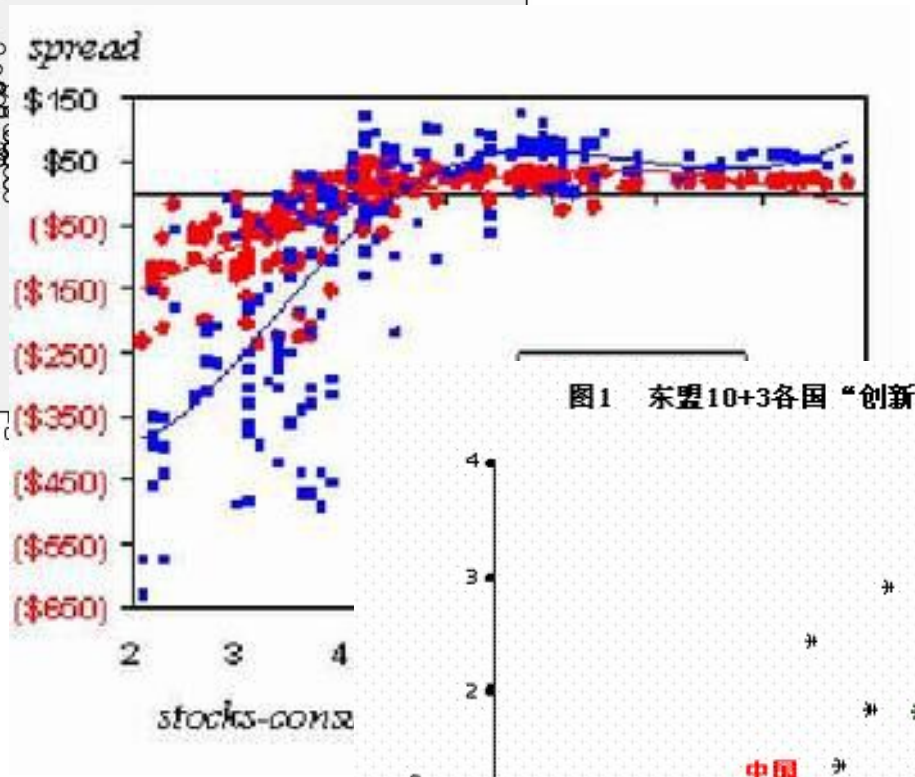
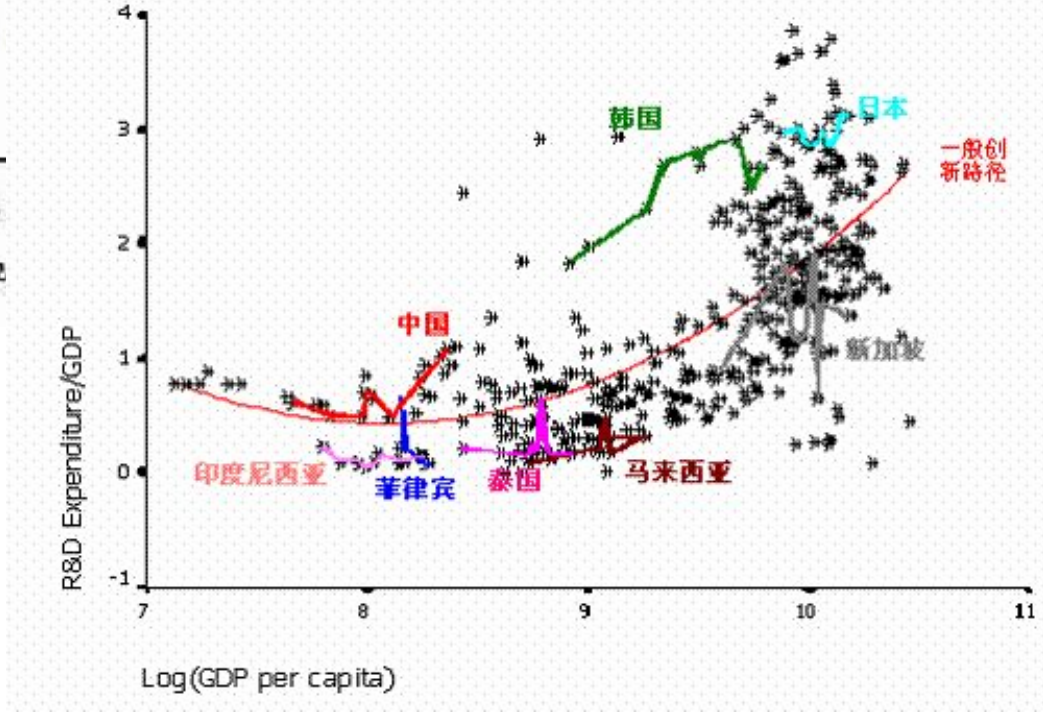


图1　东盟10+3各国“创新路径”



4.3 等级相关

顺序量表的数据，或非正态的等距等比数据，或单调变化但偏离线性的关系，可以计算等级相关。

例如：年龄与流体智力的关系（成年前）

优点：对总体没有特别要求，是非参数的相关方法，适用面广。

缺点：与积差相关相比，精度稍差。

1、斯皮尔曼等级相关

（1）适用资料：顺序量表或数值型变量按其大小排列赋以等级；两个变量均为等级变量的呈线性相关的资料。

（2）计算公式：

$$r_R = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}$$

D＝R_X-R_Y--各对偶等级之差， n为等级数目，

R_X---X变量的等级，R_Y---Y变量的等级

（3）例4.3 一家咨询公司想调查某种类型产品的质量与其市场份额的关系。通过调查，获得该行业12家公司产品的质量等级。

原始数据都是等级数据，可以直接代入公式。

质量等级与市场份额等级表

公司	质量等级（X）	市场份额（Y）	D_i	D_i^2
A	4	3	1	1
B	6	7	-1	1
C	9	5	4	16
D	7	6	1	1
E	1	2	-1	1
F	3	4	-1	1
G	11	12	-1	1
H	5	9	-4	16
I	8	8	0	0
J	12	10	2	4
K	10	11	-1	1
L	2	1	1	1

$$\sum_{i=1}^{n} D_i^2 = 44 \qquad r_R = 1 - \frac{6 \times 44}{12^3 - 12} \approx 0.85$$

公司的质量形象与其市场份额等级成正相关。

例4.4 研究学校内儿童问题行为与母亲耐心程度的关系。用X表示儿童的问题程度分数，Y表示母亲的不耐心程度分数。原始分数不是等级数据，要先化为等级数据，再代入公式。

家庭	儿童得分（X）	母亲得分（Y）	R_X	R_Y	$D=R_X-R_Y$	D^2
1	72	79	8	6	2	4
2	40	62	3	3	0	0
3	52	53	6	2	4	16
4	87	89	9	9	0	0
5	39	81	2	7	-5	25
6	95	90	10	10	0	0
7	12	10	1	1	0	0
8	64	82	7	8	-1	1
9	49	78	5	5	0	0
10	46	70	4	4	0	0
n=10			$\Sigma R_X=55$,	$\Sigma R_Y=55$	$\Sigma D=0$	$\Sigma D^2=46$

解：

$$r_R = 1 - \frac{6\sum D^2}{n\left(n^2-1\right)} = 1 - \frac{6\times 46}{10\times\left(10^2-1\right)} = 0.72$$

在观测变量没有相同等级时能够保证$\Sigma R_X = \Sigma R_Y$，$\Sigma R_X^2 = \Sigma R_Y^2$。如果观测变量出现相同等级时$\Sigma R_X = \Sigma R_Y$，但$\Sigma R_X^2 \neq \Sigma R_Y^2$。 ΣR^2 随相同等级数目的增多而有规律地减少。

此时，可再用积差相关公式计算；或者采用校正公式。

此时的斯皮尔曼等级相关公式为：

$$r_R = \frac{\sum x^2 + \sum y^2 - \sum D^2}{2\sqrt{\sum x^2} \times \sqrt{\sum y^2}},$$

其中：

$$\sum x^2 = \frac{n\left(n^2\text{-}1\right)}{12} - \sum \frac{t\left(t^2 - 1\right)}{12},$$

$$\sum y^2 = \frac{n\left(n^2\text{-}1\right)}{12} - \sum \frac{t\left(t^2 - 1\right)}{12}$$

2、肯德尔和谐系数

（1）适合于多个评价者，评价多个事物的等级变量资料。

（2）公式：

$$W=\frac{\sum_{i=1}^{n}R_i^2-\frac{\left(\sum_{i=1}^{n}R_i\right)^2}{n}}{\frac{1}{12}k^2\left(n^3-n\right)}$$

n为被评价事物的数目（＝等级数）

k为评价者的数目（＝等级变量的列数）。

（3）例4.3 有10人对七件广告作品进行等级评价，结果如下表，问这10人的评价是否具有一致性？

N=7	评价者 k=10											
作品号	1	2	3	4	5	6	7	8	9	10	R_i	R_i^2
1	3	5	2	3	4	4	3	2	4	3	33	1089
2	6	6	7	6	7	5	7	7	6	6	63	3969
3	5	4	5	7	6	6	4	4	5	4	50	2500
4	1	1	1	2	2	2	2	1	1	2	15	225
5	4	3	4	4	3	3	5	6	3	5	40	1600
6	2	2	3	1	1	1	1	3	2	1	17	289
7	7	7	6	5	5	7	6	5	7	7	62	3844

$$SS_R = \sum R_i^2 - \frac{\left(\sum R_i\right)^2}{n} = 13516 - \frac{280^2}{7} = 2316$$

$$W = \frac{2316}{\frac{1}{12} \times 10^2 \times \left(7^3 - 7\right)} = \frac{2316}{2800} = 0.827$$

所以，10人对7个作品的评价具有较高的一致性。排名为（从小到大）：4、6、1、5、3、7、2。

$$W=\frac{\sum_{i=1}^{n}R_i^2-\frac{\left(\sum_{i=1}^{n}R_i\right)^2}{n}}{\frac{1}{12}k^2\left(n^3-n\right)}$$

若k个评价者的评价完全一致，则等级和的最大方差$SS_R=k^2$（n^3-n）/12，W=1；

若完全没有相关，则各事物的等级之和相等，$SS_R=0$，W=0，W∈[0，1]。

和谐系数=实际离差平方和/最大可能离差平和。

4.4 质与量相关

质与量相关指一列变量为数值型（等距、等比）数据，另一列变量为类别变量，求两列变量的直线相关，称为质与量相关。包括：点二列相关、二列相关和多系列相关。

1、点二列相关

（1）适用资料

两列变量中一列为等距或等比的测量数据而且总体分布为正态，另一列变量为类别（名义）变量，分为两类。

点二列相关多用于编制是非测验题评价测验内部一致性等问题。每个题目（二分名义变量）与总分（数值）变量的相关，称为每个题目的区分度。相关高说明该题答对答错与总分的一致性高，即区分度高。

（2）计算公式

$$r_{pb}=\frac{\overline{X_p}-\overline{X_q}}{S_t}\sqrt{pq}$$

其中，$\overline{X_p}$ --与一个二分变量值对应的连续变量均值；

$\overline{X_q}$ --与另一二分变量值对应的连续变量均值；

p，q是二分变量两个值各自所占的比率，p+q=1；

S_t--连续变量的标准差；$r_{pb}\in[0，1]$

（3）例4.4 有一是非选择测验，共有50题，每题选对得2分，满分为100分。现有20人的总成绩及对第5题的选答情况，问第5题与总分的相关程度如何？

学生	总分	第 5 题选答情况
1	84	对
2	82	错
3	76	错
4	60	错
5	72	错
6	74	错
7	76	错
8	84	对
9	88	对
10	90	对
11	78	对
12	80	错
13	92	对
14	94	对
15	96	对
16	88	对
17	90	对
18	78	错
19	76	错
20	74	错

n=20，$\overline{X_t}$=81.6，St=8.66，答对人数10,答错人数10，p=答对学生的比率=10/20=0.5, q=1-p=0.5, $\overline{X_p}$=88.4, $\overline{X_q}$=74.8,

$$r_{pb}=\frac{\overline{X_p}-\overline{X_q}}{S_t}\sqrt{pq}=\frac{88.4-74.8}{8.66}\sqrt{0.5\times0.5}=0.785$$

第5题与总分相关较高,相关系数为0.785,即第5题的答对答错与总分有一致性。该题的区分度较高！

2、二列相关

（1）适用资料

适用于两列变量都为正态等距变量，但其中一列变量被人为地划分成两类。

二列相关与点二列相关的主要区别在于二分变量是否正态。

（2）计算公式

$$r_b = \frac{\overline{X_p} - \overline{X_q}}{S_t} \times \frac{pq}{y}$$

式中 S_t 是连续变量的标准差；$\overline{X_p}$ 为某一二分变量对应的连续变量的均值；X_q 为与另一二分变量对应的连续变量的均值；p为某一二分变量值所占的比率；y为p的正态曲线的高度，查正态表得到。

（3）例4.5下表为10名考生一次测验的卷面总分和一道回答题的得分，试求该问答题的区分度（该回答题满分为10分，因此得6分核分以上则认为该题通过）。

考生	A	B	C	D	E	F	G	H	I	J
卷面总分	75	57	73	65	67	56	63	61	65	67
回答题得分	7	6	7	4	7	4	4	4	7	6
n=10，S_y=6.12，p=6/10=0.6，q=4/10=0.4，Xp=67.33，Xq=61.25										

解：回答题得分被认为划分为通过和不通过两类，应求二列相关。查正态分布表：当p=0.60时y=0.3866，

$$r_b = \frac{\overline{X_P} - \overline{X_q}}{S_t} \times \frac{pq}{y} = \frac{67.33 - 61.25}{6.12} \times \frac{0.60 \times 0.40}{0.3866}$$

$$= 0.62$$

3、多系列相关

（1）适用资料

适用于两列变量都为正态等距变量，但其中一列变量被认为地划分成多项分类变量。

（2）计算公式

$$r_s = \frac{\sum\left(\left(y_l - y_h\right)\cdot \overline{X_i}\right)}{S_t \sum \frac{\left(y_l - y_h\right)^2}{p_i}}$$

其中，p_i为每系列的次数比率；y_l为每一名义变量下限的正态曲线高度，由p_i查正态分布表；y_h为每一名义变量上限的正态曲线高度，由p_i查正态分布表；$\overline{X}_i$是每一名义变量对应的连续变量的平均数；S_t为连续变量的标准差。

（3）例如已知若干名学生的学习能力测验分数与教师对该部分学生的评价，分A、B、C、D四等。学生的学习能力可视为正态分布。问能力测验与教师评价的一致性如何？

解决该问题需要用多系列相关。经过计算r_s=0.717，具体计算见P_{143}。此结果表明能力测验与教师的评价基本一致。

函数关系 变量间具有的严格确定的依存关系

相关关系 变量间确实存在、但数量上不严格对应、无法精确表达的关系

函数关系与相关关系之间并无严格的界限：有函数关系的变量间，由于有测量误差及各种随机因素的干扰，可表现为相关关系；对具有相关关系的变量有深刻了解之后，相关关系有可能转化为或借助函数关系来描述。

本 章 结 束