回归分析在农业科学中的应用

丁　希　泉

吉林省农业科学院 情报资料室

一九七八年

目　　录

第三章　农业科学中的曲线回归分析

第四章　农业科学中的多元回归分析

第一章　农业科学中的相关分析

在农业科学中，有很多事情彼此之间有密切关系。比如，作物生态受环境因素影响很大。有些环境因素的改变，能引起农作物生育产量的变化。大家都知道，土壤水分的多少与作物生育产量间的关系十分密切。土壤水分少（或降雨少），会使作物生育不良，造成严重减产。适宜的土壤水分，会使作物生育状况良好，产量提高。其它因素如温度、降水、日照等对作物生育产量都有极大的影响。因为这些因素是作物必不可少的生活条件。当然，不同作物品种、不同地区，各因素与作物生态间的关系也是不同的，有的与这个因素关系密切，有的与那个因素关系密切。怎样才能表示出它们之间的关系是否密切？它们之间的关系密切到什么样的程度呢？

为了便于讨论，我们将生态因子（如株高、干物重、产量等）、环境因素（如温度、水分、日照等）等等均用数学上常用的名称："变量"来代表。相关系数就是表示两个连续变量之间关系密切程度的最常用的量数。它的好处是明了又便于计算。它的缺点是仅能表示相互关系的大小，不能表示是什么样的关系，这就要用另一个表示变量之间制约关系的方法——回归分析来解决（将在后面介绍）。相关分析与回归分析有着密切的联系。应用时，一般先用相关分析确定变量之间关系明显后，再用回归分析来说明它们是什么样的关系，即求出一个变量变化时制约着另一个变量变化的方程式。当然，根据需要或者试验数据不适于相关分析，也可以直接用回归分析。但应注意到，绝不能把实质上没有关系的两件事物硬放在一起来分析，那样做会得出错误甚至荒谬的结论。

§1、相关分析原理

假设有两个连续变量X与Y，它们均为正态分布，并且X与Y之间有一定的关系，那么，二元正态分布密度函数为：

$$P(X\cdot Y)=\frac{1}{2\pi\,\sigma x\sigma y\sqrt{1-\rho^2}}\,e^{XP}\left\{-\frac{1}{2(1-\rho^2)}\times\left(\frac{(x-\bar{U}x)^2}{\sigma x^2}-2\rho\cdot\frac{(x-\bar{U}x)(y-\bar{U}y)}{\sigma x\sigma y}+\frac{(y-\bar{U}y)^2}{\sigma y^2}\right)\right\}$$

式中$\bar{U}x$——X的总体平均值；

$\bar{U}y$——Y的总体平均值；

σx——X的总体标准差；

σy——Y的总体标准差；

ρ——X与Y的总体相关系数；

e^{XP}——指数函数符号，即$e^{XP}\ Z=e^{Z}$。当Z的表达式较繁时，用$e^{XP}\ Z$的写法比较方便。

该函数的图形为一钟形曲面（如图1）。

此为试读，需要完整PDF请访问：www.ertongbook.com

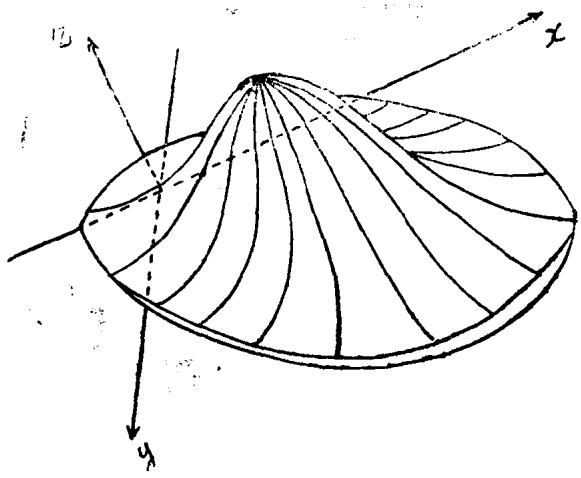

图1 二元正态分布曲面

$$\because \sigma_{xy} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\bar{U}x)(y-\bar{U}y)\, P(x\cdot y)\cdot dxdy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\bar{U}x)(y-\bar{U}y)\cdot \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\right.$$

$$\left.\cdot\left(\frac{(x-\bar{U}x)^2}{\sigma x^2} - 2\rho\frac{(x-\bar{U}x)(y-\bar{U}y)}{\sigma x\,\sigma y} + \frac{(y-\bar{U}y)^2}{\sigma y^2}\right)\right\}\cdot dxdy$$

$$= \sigma x\,\sigma y\,\rho$$

$$\therefore \rho = \frac{\sigma xy}{\sigma x\cdot\sigma y} \qquad (2)$$

公式（2）表示x与y两变量之间的总体相关系数或称理论相关系数。我们根据样本资料求的相关系数，以r表示，它是总体相关系数的估计值。实际上，我们只能计算出r值，并以此来判断x与y之间的相关性。

相关系数值的范围为$-1 \le \rho \le 1$。当$\rho = 0$时，说明变量x与y之间没有相关性，即x与y是相互独立的正态分布。当$\rho = 1$或$\rho = -1$时，表明两者有完全的正相关或负相关，它说明可以根据每一个x值求出与之相对应的y值，即回归式$y = a + bx$所表示的关系，称作y依x的回归式，b则称为y依x的回归系数。反之，也可根据y的任何一个值求出与之相对应的x值，即回归式$x = a' + b'y$所表示的关系，b'则称为x依y的回归系数。当$-1 < \rho < 1$时，表明两者有一定的关系，且$|\rho|$值越大，两者的相关关系越大。上述情况如图2。

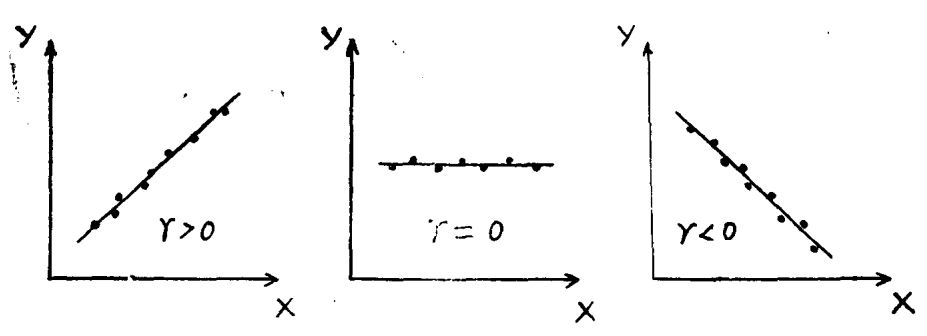

图2 相关系数几种情况

§2、直线相关分析

当两个变量x_i与y_i之间呈简单的直线关系时，这是最简单的相关分析，其相关系数称为单相关系数。

一、单相关系数的计算方法

1、一般的计算公式：根据相关系数的定义，由公式（2）得出计算相关系数r_{xy}的公式为：

$$r_{xy}=\frac{S_{xy}}{S_x\cdot S_y}=\frac{\Sigma xy}{\sqrt{\Sigma x^2\cdot\Sigma y^2}}=\frac{\Sigma(x_i-\bar{x})\Sigma(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2\cdot\Sigma(y_i-\bar{y})^2}} \qquad (3)$$

式中$x_i=x_i-\bar{x}$，$y_i=y_i-\bar{y}$；x_i、y_i均为原始数据。$\bar{x}$为x_i的平均值；$\bar{y}$为y_i的平均值。

2、直接计算法：

$$\because\quad \Sigma xy=\Sigma(x_i-\bar{x})(y_i-\bar{y})=\Sigma x_iy_i-n\bar{x}\bar{y}$$
$$\Sigma x^2=\Sigma(x_i-\bar{x})^2=\Sigma x_i^2-n\bar{x}^2$$
$$\Sigma y^2=\Sigma(y_i-\bar{y})^2=\Sigma y_i^2-n\bar{y}^2$$

代入（3）式：

$$r=\frac{\Sigma x_iy_i-n\bar{x}\bar{y}}{\sqrt{(\Sigma x_i^2-n\bar{x}^2)(\Sigma y_i^2-n\bar{y}^2)}} \qquad (4)$$

又因为 $\bar{x}=\frac{\Sigma x_i}{n}$，$\bar{y}=\frac{\Sigma y_i}{n}$

所以（4）式又变为

$$r=\frac{\Sigma x_iy_i-n\left(\frac{\Sigma x_i}{n}\right)\cdot\left(\frac{\Sigma y_i}{n}\right)}{\sqrt{\left[\Sigma x_i^2-n\cdot\left(\frac{\Sigma x_i}{n}\right)^2\right]\cdot\left[\Sigma y_i^2-n\cdot\left(\frac{\Sigma y_i}{n}\right)^2\right]}}$$
$$=\frac{n\Sigma x_iy_i-\Sigma x_i\cdot\Sigma y_i}{\sqrt{[n\Sigma x_i^2-(\Sigma x_i)^2]\cdot[n\Sigma y_i^2-(\Sigma y_i)^2]}} \qquad (5)$$

一般地，当样本个数不太大时（$n\leq30$），大多采用此公式进行计算，并且计算误差较小。

例1　根据高粱产量与8月平均气温的资料（表1）计算相关系数。

先在表内完成各栏的计算。再根据表中第（3）（4）（5）栏内的数字，由公式（3）得：

$$r=\frac{1601.8}{\sqrt{7.77\times626049}}=\frac{1601.8}{2.787\times791.232}=\frac{1601.8}{2205.2}=0.7264$$

根据表中第（6）（7）（8）栏内的数字，由式（4）得：

表1　　××县高粱产量与8月平均气温的相关系数计算

	y	x	(1) y_- $(y-\bar{y})$	(2) x_- $(x-\bar{x})$	(3) x^2	(4) y^2	(5) xy	(6) xy	(7) x^2	(8) y^2
	1253	22.8	56	0.6	0.36	3136	33.6	28568.4	519.84	1570009
	1583	23.5	386	1.3	1.69	148996	501.8	37200.5	552.25	2505889
	1005	22.4	-192	0.2	0.04	63864	-38.4	22512.0	501.76	1010025
	1268	21.9	71	-0.3	0.09	5041	-21.3	27769.2	479.61	1607824
	1230	21.2	33	-1.0	1.00	1089	-33.0	26076.0	449.44	1512900
	1095	22.4	-102	0.2	0.04	10404	-20.4	24528.0	501.76	1199025
	1523	23.1	326	0.9	0.81	106276	293.4	35181.3	533.61	2319529
	915	22.0	-282	-0.2	0.04	79524	56.4	20130.0	484.00	837225
	825	20.6	-272	-1.6	2.56	138384	595.2	16995.0	424.36	680625
	1058	21.3	-139	-0.9	0.81	19321	125.1	22535.4	453.69	1119364
	1275	22.7	78	0.5	0.25	6084	39.0	28942.5	515.29	1625625
	1088	22.0	-109	-0.2	0.04	11881	21.8	23936.0	484.00	1183744
	1440	22.4	243	0.2	0.04	59049	48.6	32256.0	501.76	2073600
合计	15558	288.3			7.77	626049	1601.8	346630.3	6401.37	19245384
平均	1197	22.2								

$$r=\frac{13\times346630.3-288.3\times15558.0}{\sqrt{[13\times6401.37-(288.3)^2]\,[13\times19245384-(15558)^2]}}$$

$$=\frac{20822.5}{\sqrt{100.92\times8138628.0}}=\frac{20822.5}{28659.2}=0.7266$$

两个公式所得结果是相同的。但从表中的数值计算看出，当数值较大时，用公式（3）计算较方便，可以避免很大数值的计算。当数值较小时，用公式（4）计算则比较方便。

3、回归分析中计算相关系数

从回归分析（参见第二章 §4）可以知道，如果y依x变量的回归式为$y=a_1+b_1x$，那么，回归系数b_1为：

$$b_1=\frac{\Sigma xy}{\Sigma x^2}=\frac{s_y}{s_x}\cdot r_{xy} \qquad (6)$$

同样地，x依y的回归式$x=a_2+b_2y$，其回归系数b_2为：

$$b_2=\frac{\Sigma xy}{\Sigma y^2}=\frac{s_x}{s_y}\cdot r_{xy} \qquad (7)$$

因而　$b_1\cdot b_2=(\frac{s_y}{s_x}\cdot r_{xy})\cdot(\frac{s_x}{s_y}\cdot r_{xy})=r^2_{xy}$

$$\therefore \quad r = \pm\sqrt{b_1 \cdot b_2} \tag{8}$$

另外，在回归分析中利用最小二乘法求出直线式$\hat{y} = a + bx$，那么，距回归线的总误差Q（见公式87）为：

$$Q = \Sigma d_i^2 = \Sigma (y_i - \hat{y}_i)^2 = \Sigma (y_i - a - bx)^2$$

$$\because \quad \bar{y} = a + b\bar{x} \qquad \therefore \quad a = \bar{y} - b\bar{x}$$

代入Q的表达式，得：

$$Q = \Sigma (y - \bar{y} + b\bar{x} - bx)^2 = \Sigma [(y - \bar{y}) - b(x - \bar{x})]^2$$

$$= \Sigma (y - \bar{y})^2 - b^2 \Sigma (x - \bar{x})^2$$

$$\therefore \quad b^2 \Sigma (x - \bar{x})^2 = \Sigma (y - \bar{y})^2 - Q \tag{9}$$

于是，由公式（3）（6）（9）得：

$$r^2 = \frac{[\Sigma (x - \bar{x})(y - \bar{y})]^2}{\Sigma (x - \bar{x})^2 \cdot \Sigma (y - \bar{y})^2} = \frac{b^2 \cdot \Sigma (x - \bar{x})^2}{\Sigma (y - \bar{y})^2}$$

$$= \frac{\Sigma (y - \bar{y})^2 - Q}{\Sigma (y - \bar{y})^2} = 1 - \frac{\Sigma (y - \hat{y})^2}{\Sigma (y - \bar{y})^2} \tag{10}$$

或者写成

$$r^2 = 1 - \frac{S_d^2}{S_y^2}, \qquad \text{即} \quad r = \pm\sqrt{1 - \frac{S_d^2}{S_y^2}} \tag{11}$$

式中 S_d—y距回归线的变异程度，即距回归的标准差；S_y—y对于平均数$\bar{y}$的变异程度，即为样本的标准差。

从上述的公式也可以清楚的看出：相关分析与回归分析是密切联系的。

4、利用相关表计算相关系数（相关表法）

当样本内试验数据的个数n很大时，用上面的公式计算相关系数很麻烦，工作量很大。在这种情况下，我们利用相关表计算相关系数就比较方便，可以减少很多工作量。

表2、　　相关表（X为0—50cm层土壤湿度 y为0—100cm层土壤湿度）

y的组距 \ X的组距	S_x	t_x	18.1～21.0	21.1～24.0	24.1～27.0	27.1～30.0	30.1～33.0	33.1～36.0	36.1～39.0	39.1～42.0
			19.5	22.5	25.5	28.5	31.5	34.5	37.5	40.5
S_y	t_y	f_{xy}	−2	−1	0	1	2	3	4	5
18.1－21.0	19.5	−1	2	1						
21.1－24.0	22.5	0		41	15					
24.1－27.0	25.5	1			28	8	1			
27.1－30.0	28.5	2				11	2	1		
30.1－33.0	31.5	3						2	1	
33.1－36.0	34.5	4								1

相关表（见表2）的编制过程如下：将资料中X值由最小值到最大值，按一定的间距分为m组，表2中的组距为3.0， m＝8；Y值也由最小值到最大值按一定的间距分为L组，表2中的组距为3.0，L＝6。然后按着每对X、Y值找出所在组交叉处的格子登记上“—”，最后数出每格内“—”数，即成为表2的形式。例如，数据X＝19，Y＝20，则应在X的18.1—21.0组与Y为18.1—21.0组相交处(即表内第一个格子内)记上“—”依此类推。

如果以S_{xi}表示横坐标各组的组中距，S_{yj}表示纵坐标各组的组中值，f_{ij}表示横坐标属于第i组，纵坐标属于第j组的数据个数，那么，

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

$$= \frac{\sum_{i=1}^{m}\sum_{j=1}^{L} f_{ij}(S_{xi}-\bar{X})(S_{yj}-\bar{y})}{\sqrt{\sum_{i=1}^{m} f_i (X_i-\bar{X})^2 \cdot \sum_{j=1}^{L} f_j (y_j-\bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{m}\sum_{j=1}^{L} f_{ij} S_{xi} S_{yj} - n\bar{X}\bar{y}}{\sqrt{\left(\sum_{i=1}^{m} f_i S_{xi}^2 - n\bar{X}^2\right)\cdot\left(\sum_{j=1}^{L} f_j S_{yj}^2 - n\bar{y}^2\right)}} \qquad (12)$$

式中 $f_i = \sum_{j=1}^{L} f_{ij}$，$f_j = \sum_{i=1}^{m} f_{ij}$，

$$\bar{X} = \frac{\sum_{i=1}^{m} f_i S_{xi}}{n}, \qquad \bar{y} = \frac{\sum_{j=1}^{L} f_j S_{yj}}{n}。$$

上式也可以写成：

$$r = \frac{\sum_{i=1}^{m}\sum_{j=1}^{L} f_{ij} S_{xi} S_{yj} - n\cdot\frac{\sum_{i=1}^{m} f_i S_{xi}}{n}\cdot\frac{\sum_{j=1}^{L} f_j S_{yj}}{n}}{\sqrt{\left(\sum_{i=1}^{m} f_i S_{xi}^2 - n\cdot\left(\frac{\sum_{i=1}^{m} f_i S_{xi}}{n}\right)^2\right)\left(\sum_{j=1}^{L} f_j S_{yj} - n\left(\frac{\sum_{j=1}^{L} f_j S_{yj}}{n}\right)^2\right)}}$$

$$= \frac{n\sum_{i=1}^{m}\sum_{j=1}^{L} f_{ij} S_{xi} S_{yj} - \left(\sum_{i=1}^{m} f_i S_{xi}\right)\left(\sum_{j=1}^{L} f_j S_{yj}\right)}{\sqrt{\left[n\sum_{i=1}^{m} f_i S_{xi}^2 - \left(\sum_{i=1}^{m} f_i S_{xi}\right)^2\right]\left[n\sum_{j=1}^{L} f_j S_{yj}^2 - \left(\sum_{j=1}^{L} f_j S_{yj}\right)^2\right]}} \qquad (13)$$

为了计算方便，可以引入参变量后再进行计算。设参变数 $t_{x_i}=\frac{S_{x_i}-A}{C}$，其中C表示X值的组距，A表示任选的某一组的组中值，一般选频数f_i最大的那一组。参变量 $t_{y_j}=\frac{S_{y_j}-B}{d}$，其中d表示y值的组距，B为任选的某一组的组中值。将 t_{x_i} 与 t_{y_j} 代入（13）式，得：

$$r=\frac{n\sum_{i=1}^{m}\sum_{j=1}^{L}f_{ij}t_{xi}t_{yj}-\left(\sum_{i=1}^{m}f_{xi}t_{xi}\right)\left(\sum_{j=1}^{L}f_{yj}t_{yj}\right)}{\sqrt{\left[n\sum_{i=1}^{m}f_{xi}t_{xi}^{2}-\left(\sum_{i=1}^{m}f_{xi}t_{xi}\right)^{2}\right]\left[n\sum_{j=1}^{L}f_{yj}t_{yj}^{2}-\left(\sum_{j=1}^{L}f_{yj}t_{yj}\right)^{2}\right]}} \qquad (14)$$

例2　根据表3的资料，计算公主岭黑土地0—100cm层的土壤湿度与0—50cm层的土壤湿度的相关系数。

我们采用公式（14）形式来计算相关系数，其计算过程如表3。表内各栏的计算说明如下：

纵行　第8栏：$-4=-2\times2$；$-42=(-1)\times1+(-1)\times41$；……。

第9栏：$8=(-2)^2\times2$；$42=(-1)^2\times1+(-1)^2\times41$；……。

第10栏：$-2=(-1)\times2$；$-1=(-1)\times1+0\times41$；……。

第11栏：$4=(-2)\times(-2)$；$1=(-1)\times(-1)$；……。

横行　第（10）栏：$-3=(-1)\times2+(-1)\times1$；……$37=1\times28+1\times8+1\times1$……。

第（11）栏：$3=(-1)^2\times2+(-1)^2\times1$；…$37=1^2\times28+1^2\times8+1^2\times1$……。

第（12）栏：$-5=(-2)\times2+(-1)\times1$；…

$10=0\times28+1\times8+2\times1$；……。

第（13）栏：$5=(-1)\times(-5)$；…$10=1\times10$……。

将表内数值代入公式（14），得：

$$r=\frac{124\times101-(-3)\times75}{\sqrt{[124\times149-(-3)^2][124\times139-(75)^2]}}$$

$$=\frac{12749}{\sqrt{18467\times11611}}=\frac{12749}{135.9\times107.75}$$

$$=\frac{12749.0}{14642.15}=0.8707$$

此外，在观测资料数目较大时，或者从许多因素中挑出有关因素时，还有一种求相关系数的近似方法。首先，根据数据点出两变量X与y的相关图，图3表示大豆产量与开花—成熟期降水量的关系。然后作一条水平线，将点上下平分；再作一条垂直线，将点左右平分。这两条线将平面分成四块，分别数每一块的点数（在平分线上的点不算），得出相关方向上两块的点数（即点数多的两块）$n_1=n_2=14$ 其余两块$n_3=n_4=4$。

此为试读，需要完整PDF请访问：www.ertongbook.com

表 3　　相关系数的计算

编号	编号 / t_{yj} / t_{xi} / $f_{xi\,yj}$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
	t_{xi}	-2	-1	0	1	2	3	4	5	Σf_{yj}	$t_{yj}f_{yj}$	$t_{yj}^2 f_{yj}$	$\Sigma t_{xi} f_{yj}$	$t_{yj}\Sigma t_{xi} f_{yj}$
1	-1	2	1							3	-3	3	-5	5
2	0		41	15						56	0	0	-41	0
3	1			28	8	1				37	37	37	10	10
4	2				11	2	1			14	28	56	18	36
5	3						2	1		3	9	27	10	30
6	4								1	1	4	16	5	20
7	Σf_{xi}	2	42	43	19	3	3	1	1	124	75	139	—	101
8	$\Sigma t_{xi} f_{xi}$	-4	-42	0	19	6	9	4	5	-3				
9	$\Sigma t_{xi}^2 f_{xi}$	8	42	0	19	12	27	16	25	149				
10	$\Sigma t_{yj} f_{xi}$	-2	-1	28	30	5	8	3	4	—				
11	$t_{xi}\Sigma t_{yj} f_{xi}$	4	1	0	30	10	24	12	20	101				

设 $n_+ = n_1 + n_2 = 28$，$n_- = n_3 + n_4 = 8$ 则由相关系数近似计算公式：

$$\dot{r} = \sin\left(\frac{n_+}{n_+ + n_-} - \frac{1}{2}\right)\pi \tag{15}$$

得：

$$\dot{r} = \sin\left(\frac{28}{28+8} - \frac{1}{2}\right)\pi = \sin\frac{5}{18}\pi = \sin 50^\circ = 0.7660$$

如果根据公式（5）计算，则得相关系数为0.8245。虽然两个结果的数值不一样，但差别不大。如果用相关系数的显著性测定方法来检验，两者之间差异不显著（见例7）。它们均达到0.1%的显著水准，说明大豆产量与开花—成熟期降水量之间的相关关系极为明显。

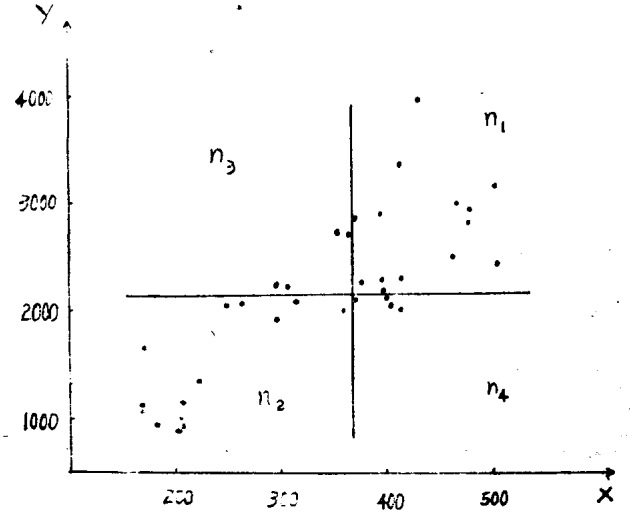

图3　　相关系数 r 的近似计算

二、等级相关时的相关系数计算方法

在实际工作中，常常遇到调查数据或试验结果不是具体的数值，而只是表示它们等级的号码。例如，我们将作物生长状况分为好、较好、一般、较差、差五等；将产量分为高产、一般、低产；将降水量分为极多、多雨、一般、少雨、极少五等，或者分为旱，一般，涝三等。并将它们用数字1、2、3、4、5来表示。在这种情况下它们之间的相关称为等级相关，也可以计算出相关系数。但是，由于等级不可能分得很多，一般也就是3～5等，而等级少时所计算出的相关系数受到偶然性影响较大，所以结果并不十分可靠，只能反映出一般的情况，可作参考之用。

计算等级相关，相关系数的公式为史皮曼（SPearman）公式：

$$\rho = 1 - \frac{6\sum D^2}{n(n^2-1)} \tag{16}$$

式中，n为项数；$D = Y - X$，即相关的两变量的等级差值。

求得ρ后，再由下式求得相关系数 r：

$$r = 2\,\mathrm{Sin}\left(\frac{\pi\rho}{6}\right) \tag{17}$$

皮尔逊（Pearson）根据公式（17），制作出ρ与 r的对照表（附表1），求得ρ后，直接查表得出 r。

例3××县农民谚语说“秋后雨水多，来夏淹山坡”，为了检验这一谚语，我们将降雨量分为少，偏少，偏多，多四等，其界限分别为平均降雨量的70%以下，70—100%，101—130%，130%以上，并记为1，2、3、4等，如表4，X表示秋雨（10—11月）等级，y表示夏季（6—9月）降雨等级。

表 4　　秋雨与夏季降雨的等级相关

年 份	X	y	D	D^2	年 份	X	y	D	D^2
1955	3	3	0	0	1963	3	3	0	0
56	4	3	1	1	64	1	3	2	4
57	2	1	1	1	65	1	2	1	1
58	2	3	1	1	66	2	2	0	0
59	4	4	0	0	67	3	1	2	4
60	2	3	1	1	68	1	2	1	1
61	2	2	0	0	69	1	1	0	0
62	3	3	0	0	70	2	1	1	1

∵　$n=16$，$\Sigma D^2=15$

∴　$$\rho = 1 - \frac{6\times 15}{16\times(16^2-1)} = 1 - 0.0221 = 0.9779$$

查附表1，得 $r=0.9818$。

如果用相关系数的显著性测定方法来检验（见下节），这是高度显著的，说明当年秋雨多少与来年夏季雨水的多少有极为密切的关系，即农民谚语：“秋后雨水多，来夏淹山坡”是正确的。

三、相关系数显著性测定方法

从前面的各种计算方法表明，我们所进行的相关系数计算都是以样本资料为基础的，它不是真正的总体相关系数，只是总体相关系数的一个近似值。它的近似程度随取样方法、样本单元数的多少及其样本资料的准确程度等的不同而异。那么，我们究竟怎样判断样本相关系数对总体相关系数的近似程度，即样本的相关系数能否代表总体的相关情况？还有计算出的样本相关系数与另一个标准相关系数比较；二个样本相关系数之间的比较等，诸如此类问题都是相关系数显著性测定要解决的问题。

1、或差法：

当样本单元数很大时，取自同一总体的各样本相关系数 r就接近于正态分布，其平均

数即是总体相关系数ρ。

相关系数 r的标准差为：

$$\sigma_r = \frac{1-r^2}{\sqrt{n-1}} \doteq \frac{1-r^2}{\sqrt{n}} \tag{18}$$

于是，我们可以得到相应于相关系数 r的正态分布曲线公式为：

$$P(r_0) = 1 - \sqrt{\frac{2}{\pi}} \int_0^{t_0} e^{-\frac{t_0^2}{2}} \cdot dt \tag{19}$$

式中 $t = \frac{r_0}{\sigma_r}$

从而可求出由于偶然机会而得到的相关系数 $r \geq r_0$的概率。在实际应用中，一般采用的概率标准（亦即显著水准）是0.05与0.01，此时的 t值分别为 $t \geq 1.96$（$P = 0.05$时）与 $t \geq 2.6$（$P = 0.01$时）。因而，只有 $r \geq 1.96\sigma_r$与 $r \geq 2.6\sigma_r$时才表示相关系数分别达到0.05与0.01的显著水准，即相关显著或很显著。

一般，并不采用上述那样具体的计算，只是用相关系数大于（或等于）或然差（P.E.r）的四倍就认为相关显著。这种相关程度仅达0.05的显著水准。其P.E.r公式为

$$P.E.r = \pm 0.6745 \frac{1-r^2}{\sqrt{n}} \tag{20}$$

并且习惯写成 $r \pm$ P.E.r的形式。

若相关系数是由等级相关时计算出来的，其或然差公式则为：

$$P.E.r = \pm 0.7063 \frac{1-r^2}{\sqrt{n}} \tag{21}$$

当然，这种方法只能在样本单元数很大时（$n > 100$，严格要求应是 $n \geq 500$）应用。对于 n较小时，r的分布就不是正态分布，因而再应用此法是不恰当的，需要采用另外的方法进行相关系数的显著性测定。

例4　由例2知：$n = 124$，$r = 0.8707$，试检验此相关是否显著。

根据公式（20）得：

$$P.E.r = \pm 0.6745 \frac{1-(0.8707)^2}{\sqrt{124}} = \pm 0.0146$$

因　$r > 4$ P.E.r $= 0.0584$

所以，此相关显著。　0—50厘米层土壤湿度与0—100厘米层土壤湿度之间关系很密切。

2、t 测验法：

当样本单元数较小时，相关系数 r的分布就不是正态分布，而是一个偏态分布（如图4）。因此，不能用 或差法来测定 r的显著性。费雪氏（Fisher）指出，假设总体相关系数$\rho = 0$，那么，新变量：

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{22}$$

是服从于自由度为n－2的 t分布，亦称学生氏分布，其公式为

$$P(t) = 1 - 2\int_{t_0}^{\infty} \frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{m\pi}\,\Gamma\left(\frac{m}{2}\right)} \left(\frac{t^2}{m} + 1\right)^{-\frac{m+1}{2}} \qquad (23)$$

式中m为自由度，$m = n - 2$。

从而可求出因偶然机会所得到的相关系数 $r \geq r_0$的机率。并且已有根据 t分布制成的表，即 t分布表。因而容易查出自由度为n－2时， t值的机率值 P（t）。若$P < P_0$（$P_0 = 0.05$或0.01），即为相关显著。反之，$P > P_0$，即为不显著。

但是，这样计算、查表仍还很麻烦。于是，费雪氏根据这个原理，编制了直接根据相关系数 r与自由度 n－2就可以查出相关显著程度的表，即相关系数显著性检验表（附表2）。

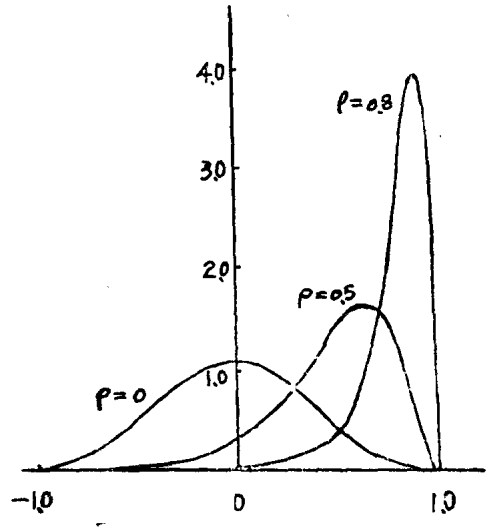

图4　正态相关下，样本相关系数r的分布曲线（n＝10）

例5　从例1知高粱产量与8月气温的相关系数 $r = 0.7266$，$n = 13$，试检验相关程度？

∵ 自由度d.f.＝n－2＝13－2＝11

查附表2，得：

5％显著水准时，$r_{0.05} = 0.5529$，　1％显著水准时，$r_{0.01} = 0.6835$，

因为，$r > r_{0.01}$。所以，相关很显著，即高粱产量与8月气温之间的关系很密切。

3、Z测定法

当样本单元数小，并且要比较的总体相关系数$\rho \neq 0$；或者是由两个样本分别得到相关系数 r_1与 r_2，要比较这两个相关系数是否有差异，如果仍用上面的方法是不适合的。这时，相关系数 r的分布特别偏斜，对于这样的偏态分布，费雪又引进一个新变量 Z 进行变换，使其Z值分布符合于正态分布（图5），这个变换步骤很复杂，其最后的变换式为：

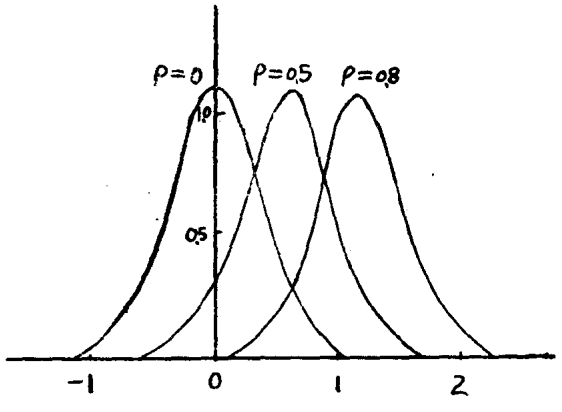

图5　正态相关下，$Z = \frac{1}{2}\ln\frac{1+r}{1-r}$的分布曲线（n＝10）

$$Z = \frac{1}{2}\ln\frac{1+r}{1-r} = 1.1513\log\frac{1+r}{1-r} \qquad (24)$$

及
$$\xi = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} = 1.1513\log\frac{1+\rho}{1-\rho} \tag{25}$$

式中 r与ρ 分别为样本与总体的相关系数。 于是Z — ξ则近似于正态分布， 其平均数为0，标准差(σ_z)为$1/\sqrt{n-3}$； Z近似于正态分布，其平均数为 ξ，标准差为$1/\sqrt{n-3}$。它的概率公式为：

$$P(Z) = 1 - \sqrt{\frac{2}{\pi}}\int_0^{z_0} e^{-\frac{Z^2}{2\sigma_z^2}} \cdot dZ \tag{26}$$

从而可以求出因偶然机会所得到的相关系数 $r \geq \rho$ 的机率P。 若$P < P_0$（ $P_0 = 0.05$或0.01），则为显著。也就是说明，在总体相关系数ρ 与样本相关数 r之间差异显著。

如果不是上面那样将一个样本的相关系数 r与一个总体相关系数ρ （$\rho \neq 0$）进行比较，而是比较二个样本相关系数 r_1与 r_2，或者是二个总体相关系数ρ_1与ρ_2进行比较，用Z测验法来测定如下：

根据公式（24）、（25）或者查附表3得：

$$\begin{cases} Z_1 = \frac{1}{2}\ln\frac{1+r}{1-r} \\ Z_2 = \frac{1}{2}\ln\frac{1+r}{1-r} \end{cases}$$

及
$$\begin{cases} \xi_1 = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} \\ \xi_2 = \frac{1}{2}\ln\frac{1+\rho}{1-\rho} \end{cases}$$

于是$Z_1 - Z_2$的分布近似于正态分布，其平均数为 $\xi_1 - \xi_2$，标准差（$\sigma_{z_1-z_2}$）为：

$$\sigma_{z_1-z_2} = \sqrt{\sigma_{z_1}^2 + \sigma_{z_2}^2} = \sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}} \tag{27}$$

因而
$$t = \frac{Z_1 - Z_2}{\sigma_{z_1-z_2}} = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}} \tag{28}$$

查附表4得出 $t_{0.05}$与 $t_{0.01}$值。若 $t > t_{0.05}$或 $t_{0.01}$，则二个相关系数差异显著或很显著，亦即它们的差异可能是真正的差异。反之，若 $t < t_{0.05}$，则二个相关系数差异不显著，即它们的差异可能是由于误差所造成的。

例6　今有大豆产量与开花期降水量的相关系数 $r_1 = 0.5805$，$n_1 = 17$；$r_2 = 0.5525$，$n_2 = 12$，它们分别是根据分期播种资料与历年产量资料求得的，试问两者结果是否一致？

解：查附表3

当$n_1 = 17$，　$r_1 = 0.5805$时，得$Z_1 = 0.66$

　　$n_2 = 12$，　$r_2 = 0.5525$时，得$Z_2 = 0.63$

于是由（28）式得：

此为试读，需要完整PDF请访问：www.ertongbook.com