杨永岐

农业气象中的统计方法

气象出版社

农业气象中的统计方法

杨永歧

气象出版社

内 容 简 介

本书介绍了目前农业气象中常用的一些数理统计方法的具体应用，内容包括相关分析、直线回归、曲线回归、多元线性回归、多项式回归、逐步回归和积分回归等问题。叙述通俗，说理清楚，并列举大量实例说明其计算方法和步骤，尽量做到条理化、表格化。本书可供农业气象工作者学习使用，也可供农学、生物学等专业的科技人员参考。

农业气象中的统计方法

杨永歧

*

气象出版社出版

（北京西郊白石桥路46号）

北京丰台岳各庄印刷厂印刷　新华书店北京发行所发行

*

开本787×1092 1/32　印张11.75　字数261千字

1983书5月第1版　1983年5月第1次印刷

印数：1—7,000　统一书号：13194·0105

定价：1.00元

目　　录

前　言

农业气象学是介于农学和气象学之间的边缘学科，它的研究范围非常广泛，研究手段和研究方法也是多样的。广大气象工作者在农业气象科学试验和农业气候资源调查中，得到一些十分庞杂的数据、资料。为了深入地研究农业生物的生长、发育、产量和品质与气象条件的关系，定量地估计各种气象要素及其波动对农业的效应；揭示农业生物对气象条件的基本要求和农业气象灾害等，以便为充分合理地利用农业气候资源，扬长避短，趋利避害，实现高产优质管理以及发展多种经营提供科学依据。为此，需要将这些大量的数据、资料进行数学处理，以期揭露事物的本质和内在规律。数理统计方法就可以帮助我们分析资料，处理数据，揭示农业生物和气象条件之间的内在关系。实践证明，只要我们掌握的数据、资料充分可靠，选用的数学模型合理，它就会帮助我们发现一般的分析方法所不易发现的规律，因而把试验或考察数据、资料本身所包含的信息和某些固有规律提炼出来。但应当指出的是，数理统计方法运用于农业气象学科时，必须结合本学科的理论基础和实践经验灵活掌握，切不可生搬硬套某些公式或例子（把本来毫不相干的数据、资料进行统计分析），从而得出一些荒谬的结论。

随着国民经济的发展，现代大农业对环境条件中的光、热、水、气等气象因子提出了愈来愈科学的要求，使得农业气象条件分析逐步从定性向定量发展，从简单的数值关系逐步向复杂的数学模型发展。因此，数理统计方法在农业气象中的应用也愈来愈广泛；尤其是计算技术的进步，为这种应用开辟了更加广阔的前景，其内容更加丰富，方法日臻完善。在这种情况下，农业气象工作者迫切希望有一本通俗易懂、简明实用的农业气象统计方法的书，以适应当前实际工作的需要。近年来，东北地区陆续举办了以应用为主的农业气象训练班，我们讲授一些农业气象中的数理统计方法。本书就是在原有讲稿的基础上经过整理、加工、修改而形成的。

我们的目的是向广大的实际工作者，尽可能多地介绍一些实用方法。因此，书中尽量避免一些复杂的数学推导，而直接引用其最终公式，并以具体例子为线索，简要地说明各种统计方法的直观原理，详细地介绍具体的计算步骤和分析方法。为了便于比较，某些方法的例题资料是连续使用的。凡是具有中等数学基础知识的读者都可以看懂，只要掌握了这些方法的基本原理，就会在具体的运算过程中举一反三，有所创见，有所前进。鉴于目前农业气候服务工作的需要，书末附有农业气候资料整理方法，以备参考。

本书初稿曾得到沈阳农学院、辽宁农业科学院、吉林农业科学院和气象界的老师和同行们的大力支持和帮助，并提出了许多宝贵意见；书中还引用一些单位和同志的工作成果资料，在此一并表示感谢。由于笔者的专业知识和技术水平所限，书中难免出现缺点和错误，恳请读者提出批评指正。

作者

1981年12月于沈阳

第一章 相关分析

在自然界中，各种事物，各种现象之间是相互联系、相互影响的。在农业气象中，许多因素之间也存在着密切程度不同的关系。有时一个因素同另一个因素之间有关系，有时一个因素同时与其它几个因素之间有关系；有时它们之间的关系比较直观、简单，有时它们之间的关系是那么复杂、那么令人捉摸不定等等。为估计它们之间关系的密切程度，则需要运用相关分析的方法。

§ 1 相关分析的意义

农业生物的生长、发育、产量和品质的变化，无不受环境条件所制约。环境条件中经常发生变化的光、热、水、二氧化碳等气象因素，能不同程度地引起农业生物的生育状况和经济性状的改变。例如太阳辐射量的多少决定着光合产物的积累量，日照时数的变化能引起开花的迟早，积温的数量在一定程度上决定着作物产量的多寡，二氧化碳浓度决定着植物的同化量等等。这些因素，都是农业生物生长发育和产量形成所不可缺少的生活因子。诚然，不同的农业生物、品种以及不同的处理方法，各因素与农业生物之间的关系也不尽相同。不仅如此，生活因子的年际变化，气象要素的强度和时空分布也都在颇大程度上左右着农业生物的生命过程。于是，就提出这样的问题：如农林作物的生育速度、开花、结实、产量、品质等，牲畜的生长、繁殖，鱼儿的产卵，蚕儿的发育、放养、结茧等等，究竟与哪些时段、哪些气象因子有关系？相关的程度如何？哪些因子的影响是主要的？哪些因子的影响是次要的？为此，我们借助调查研究、系统观测等手段取得足够的数据、资料，并将农业生物方面（如发育速度、干物质重量、千粒重、产量等）和环境条件中（如太阳辐射、日照时数、温度、降水量、二氧化碳浓度等）的资料都用“变量”表示，然后分析变量之间的关系。相关分析就是其中的一种方法，也是数理统计中最普通最常用的一种方法，在回归分析中占有重要地位。

§ 2 相关与函数

科学研究的实践表明，变量之间的关系可以分为两种类型：相关关系和函数关系。若变量之间存在着完全确定的关系，如在一定电阻 R 的电路中，电流强度 I 与加在该电路两端的电压 V 存在着完全确定的关系，即欧姆定律：

$$I=\frac{V}{R}$$

由上式可知，每给定一个电压值，就有一个电流强度值与之对应，上式完全确定了电压与电流强度的关系。又如，在匀速运动中，距离 S 与所需时间 t 有如下完全确定的关系：

$$S=vt$$

即是说，当速度一定时，距离与时间有一一对应的关系，这种完全确定性的关系，称之

为函数关系。

然而，在实际工作中，在农业气象的许多问题中，变量之间的关系并非那么简单。例如，粳稻产量与生育期总积温这两个变量就不存在着完全确定性的关系，生育期总积温相同，在不同的年份，即使是同一块田里，粳稻的产量也往往不同。又如，高粱出苗到三叶期的日数，与该时段的平均气温这两个变量之间也不存在着完全确定性的函数关系，即该时段的平均气温相同，不同的年份其出苗到三叶期所经历的日数也不尽相同。具体地说，粳稻产量虽然有随着积温的增加而增加的趋势，但有的年份产量增加的多，有的年份增加的少，个别年份产量甚至会减少；高粱出苗到三叶期所经历的日数虽然随着该时段的平均气温增加而缩短，但有的年份缩短的快，有的年份缩短的慢，甚至有个别年份没有缩短。出现这类情况的原因是相当复杂的，因为影响一个变量的因子是多样的，影响因子之间又相互制约，加上一些偶然性因素的作用，使变量之间的关系形成了不确定性。比如，高粱出苗至三叶期所经历的日数，除受该时段平均气温影响外，还受温度分布状况的影响、水分供应状况的影响等。总之，在诸影响因素之中，有些因素属于人们一时尚没认识或尚未被发现的，有些虽然认识，但一时尚无法测量或控制的。再加上测量的误差、农业技术的改进、品种的更新、农业灾害以及某些试验条件的改变等等，这就使得在农业气象研究中，变量之间的关系变得特别复杂，加剧了这种关系的不确定性。

既然农业气象研究中，这种不确定性的关系如此复杂，那末是否就无规律可循呢？当然不是。偶然性寓于必然性之中，只要经过艰苦细致的调查研究，多年的系统观测，取得充分可靠的资料，选择适宜的数学模型，就会揭示某些变量之间存在某些客观规律性。虽然变量之间的关系属于不确定性的，但总是有规律可循的。一般说来，生育期总积温愈高，粳稻产量也高；高粱出苗至三叶期平均温度愈高，则该发育期的持续日数就愈短。我们称变量之间的这种关系为相关关系，即变量之间的关系密切，但又不能由一个或几个变量的数值精确地求出另一个变量的值。由于这种关系具有统计上的意义，故也称统计相关，简称相关。

应当指出，相关关系和函数关系虽然是两种不同类型的变量关系，但它们之间并无严格的界限。由于测量误差等原因，确定性的函数关系在实际中往往通过相关关系表示出来；当对变量之间的内部规律已充分了解时，相关关系在一定条件下，从一定的统计意义上看，便可以转化为确定性的函数关系。古今无数的科学活动的实践表明，许多定律都是在大量实验数据的基础上，经过从感性到理性的提高，才总结出来的，即由统计上的相关，转化为完全确定的函数关系。

§3 直线相关

假设自变量为 X，因变量为 Y，当它们之间的关系呈简单的直线关系时，即 X 的增加（或减少），引起 Y 的增加（或减少），或者 X 的增加（或减少），引起 Y 的减少（或增加），我们称这两个变量的关系为直线相关，这是一种最普通、最简单的相关。

一、相关系数的定义

描述两个变量线性关系密切程度的数量化指标，就叫做单相关系数，或样本单相关

系数，简称相关系数。由此可见，相关系数是两个变量之间联系性强度的一种度量，通常用 r 表示：

$$r=\frac{l_{XY}}{\sqrt{l_{XX}\cdot l_{YY}}}=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\cdot\sum(Y-\bar{Y})^2}}$$

$$=\frac{\sum XY-\frac{1}{n}(\sum X)(\sum Y)}{\sqrt{\left[\sum X^2-\frac{1}{n}(\sum X)^2\right]\cdot\left[\sum Y^2-\frac{1}{n}(\sum Y)^2\right]}} \qquad (1.1)$$

其中 $l_{XX}=\sum(X-\bar{X})^2$ 为 X 的离差平方和，$l_{YY}=\sum(Y-\bar{Y})^2$ 为 Y 的离差平方和，$l_{XY}=\sum(X-\bar{X})(Y-\bar{Y})$ 为 X，Y 的离差乘积之和。可见，这样的相关系数是立足于样本资料计算的，称为样本相关系数，它不受度量单位的影响，是一个无量纲的数量，其符号决定于 l_{XY} 的符号，与回归系数 b（见第二章）的符号一致。

二、相关系数的意义

不难证明，对于 X，Y 的任何数值，(1.1)式中分子的绝对值永远不会大于分母的值。因此相关系数的取值范围：

$$0\leqslant|r|\leqslant1 \qquad (1.2)$$

作不同相关系数的散点图(见图 1.1)。由散点的分布状况，可以比较直观的理解相关系数的意义。

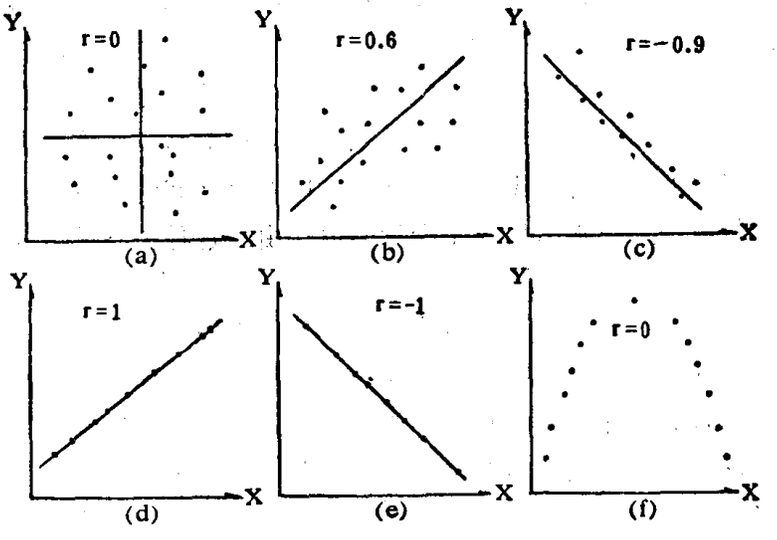

图 1.1 相关系数散点示意图

(1) $r=0$，在图 1.1 a 中，点子散布在通过$(\bar{X},\ \bar{Y})$并平行于 X 轴的直线上下，也散布在通过$(\bar{X},\ \bar{Y})$并平行于 Y 轴的直线左右，即样本点子相等地分布在这两条直线所分割的四个象限中，因此，$l_{XY}=0$，则 $b=0$。这就说明 Y 的变化与 X 无关系。此时，X 与 Y 毫无线性关系，在通常情况下，散点的分布是无规则的。

(2) $0<|r|<1$，这是绝大多数的情形，X 与 Y 之间存在着一定的线性关系。这时散点有集中在不是轴的一条直线附近的趋势。当 $r>0$ 时，$l_{XY}>0$，$b>0$，散点有 Y

随 X 的增加而增加的趋势，此时称 X 与 Y 呈正相关（图 1.1 b）；当 $r<0$ 时，$l_{xy}<0$，$b<0$，散点有 Y 随 X 的增加而减少的趋势，此时称 X 与 Y 呈负相关（图 1.1 c）。当 r 的绝对值比较小时，散点离直线较分散，而当 r 的绝对值较大时，散点就比较靠近直线。

(3) $|r|=1$，这是在实际中少有的情形，X 与 Y 呈完全的线性关系。此时，所有散点都集中在与轴不平行的一根直线上。当 $r=1$ 时，称为完全的正相关(图 1.1 d)；当 $r=-1$ 时，称为完全的负相关（图 1.1 e）。实际上，此时 X 与 Y 之间存在着完全确定性的函数关系。

综上讨论可知，相关系数 r 确实可以度量两个变量 X 与 Y 之间线性关系密切的程度。当 r 的绝对值越接近 0 时，X 与 Y 之间的线性相关程度越小；反之，r 的绝对值越接近 1 时，X 与 Y 之间的线性相关也就越密切，相关的程度就愈高。这里要指出的是，相关系数只是表示两个变量 X 与 Y 之间的线性关系的密切程度，它只是两个变量之间的联合关系的资料。因此，相关与描述性的技术相联。当 r 很小时，甚至接近0的情形下，就可以认为 X 与 Y 之间没有线性关系，这时有两种情况：一种是表示两个变量之间没有线性关系（图 1.1 a）；另一种是表示两个变量之间有非线性关系（图 1.1 f）。

三、相关系数的计算

这种相关系数因具有较大的实用价值，故介绍它的计算方法。在实际工作中，发现社会产量资料的变化与年际间有关，如产量有随年份的增长而增加的趋势，为揭示它们之间关系的密切程度，需要计算产量与年序的相关系数。

[例 1.1] 利用某县水稻产量资料(Y)，试计算它与年份序列(X)的相关系数。

具体计算步骤如下：

(1) 原终资料列表（见表 1.1）；

表 1.1 水稻产量逐年变化的原始记录 (n = 18)

年份	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970
时间序列 X	1	2	3	4	5	6	7	8	9	10
产量(斤/亩) Y	342	354	584	533	484	559	661	683	485	634
年份	1971	1972	1973	1974	1975	1976	1977	1978	总计	
时间序列 X	11	12	13	14	15	16	17	18	171	
产量(斤/亩) Y	778	701	862	922	919	714	856	936	12007	

将年份按时间顺序排列为 1，2，…，18，其对应的产量资料照抄，以便计算产量与各年代的相关系数。

(2) 列表分别计算两个变量的平方以及它们的交叉乘积（见表 1.2）；

(3) 列表计算各变量的离差平方和 l_{xx}，l_{yy} 及它们的离差乘积之和 l_{xy}，并按(1.1)式计算水稻产量与年份的相关系数 r（见表 1.3）。

这种计算步骤既方便记忆，又便于检查。在使用袖珍计算器时，步骤 (2) 与 (3) 中

表 1.2　水稻产量与年份的相关系数的计算(Ⅰ)

编　　号	X	Y	X^2	Y^2	XY
1	1	342	1	116964	342
2	2	354	4	125316	708
3	3	584	9	341056	1752
4	4	533	16	284089	2132
5	5	484	25	234256	2420
6	6	559	36	312481	3354
7	7	661	49	436921	4627
8	8	683	64	466489	5464
9	9	485	81	235225	4365
10	10	634	100	401956	6340
11	11	778	121	605284	8558
12	12	701	144	491401	8412
13	13	862	169	743044	11206
14	14	922	196	850084	12908
15	15	919	225	844561	13785
16	16	714	256	509796	11424
17	17	856	289	732736	14552
18	18	936	324	876096	16848
Σ	171	12007	2109	8607755	129197

表 1.3　水稻产量与年份的相关系数的计算(Ⅱ)

$\Sigma X = 171$	$\Sigma Y = 12007$	$n = 18$
$\bar{X} = 9.5$	$\bar{Y} = 667.06$	
$\Sigma X^2 = 2109$	$\Sigma Y^2 = 8607755$	$\Sigma XY = 129197$
$(\Sigma X)^2/n = 1624.5$	$(\Sigma Y)^2/n = 8009336.1$	$(\Sigma X)(\Sigma Y)/n = 114066.5$
$l_{XX} = 484.5$	$l_{YY} = 598418.9$	$l_{XY} = 15130.5$

$$r = \frac{l_{XY}}{\sqrt{l_{XX} \cdot l_{YY}}} = \frac{15130.5}{\sqrt{(484.5)(598418.9)}} = 0.8886$$

的数据可以一次得到，但要注意变量个数正确、无误的输入，这是计算成败的关键。

四、符号相关

当样本容量很大时，或从诸多因素中挑取有关因素时，为避免繁琐的相关系数的计算，往往在初级普查相关因子时，采用符号相关的分析方法，也称为图解相关，其精度比直接利用(1.1)式计算要差些。具体计算步骤（见图 1.2)：

(1) 将两个变量 X，Y 的所有数据点在平面坐标图中；

(2) 然后作平行于 X 轴的直线 a 将所有散点上下平分，作平行于 Y 轴的直线 b 将所有散点左右平分；

(3) 将 a，b 两条直线所划分的四个象限分别命名：一、三象限分别为 n_1、n_2；二、四象限分别为 n_3、n_4，然后分别数出每个象限内的散点数目（注意：正好被直线 a，b 所切割的点子不计）；

(4) 根据散点群体的方向，令 n_1、n_2 为正，n_3、n_4 为负，则 $n_+ = n_1 + n_2$、$n_- = n_3 + n_4$；

(5) 根据公式计算相关系数：

$$r = \sin\left(\frac{n_+}{n_+ + n_-} - \frac{1}{2}\right)\pi \tag{1.3}$$

[例 1.2] 利用例 1.1 的资料，说明符号相关的计算方法。

参照本节所讲各步骤作相关散点图 1.2：

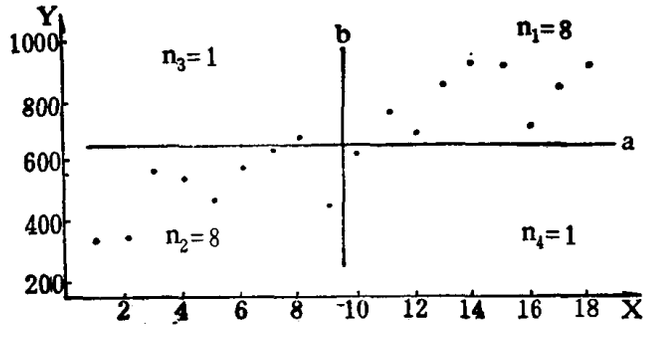

图 1.2 相关系数的图解

由上图可知：$n_1 = n_2 = 8$，$n_3 = n_4 = 1$，$n_+ = n_1 + n_2 = 8 + 8 = 16$，$n_- = n_3 + n_4 = 1 + 1 = 2$。根据 (1.3) 式计算符号相关系数：

$$r = \sin\left(\frac{16}{16+2} - \frac{1}{2}\right)\pi = \sin\left(\frac{7}{18}\pi\right) = \sin 70^\circ = 0.9397.$$

由上述分析和计算，我们不难发现，n_+ 与 n_- 的数值相差越大，散点越趋于直线，相关越密切，计算所得的相关系数也越大；反之，n_+ 与 n_- 的数值相差越小，图中散点越离散，两个变量的相关越不密切，计算所得的相关系数也越小；若 n_+ 与 n_- 的数值相等，即图 1.2 中被直线 a，b 所分成的四个象限的点子数目相等时，则相关为零，这两个变量之间毫无线性相关。这一事实再一次说明了相关系数的意义。

五、秩次相关

在实际工作中，我们常常遇到这样的情况：所收集的数据、资料，无法判明其母体[1] 的内在分布，这就需要用到秩次相关的方法，也称为等级相关。将收集的资料按其特征分成若干等级，计算两个变量的相关系数公式为

1) 统计分析的对象的全体称为母体或总体。

$$r_s = 1 - \frac{6\sum D^2}{n(K^2-1)} \tag{1.4}$$

其中，D 为两个变量的等级差；K 为所分的等级数；n 为样本个数，即 D 的个数。

若将收集的资料直接编成次序，则计算两个变量的相关系数公式为

$$r_s = 1 - \frac{6\sum D^2}{n(n-1)} \tag{1.5}$$

现在用两个例子来说明秩次相关系数的计算方法及有关技术处理。

[例 1.3] 在作物农业气象条件的调查研究中发现，某公社稻谷空壳瘪粒状况与花期温度有关，即该时期温度低，稻谷空瘪率也高，试计算它们的相关系数。

首先将调查的历年稻谷空瘪状况分为最低、低、正常、较高四个等级；然后将对应的各年花期温度分为最高、高、正常、较低四个等级，并将等级排成秩次 1，2，3，4（见表 1.4）。

表 1.4 稻谷空瘪(Y)与花期温度(X)的等级相关

年 代	X	Y	D	D^2	年 代	X	Y	D	D^2
1961	4	3	−1	1	1971	3	2	−1	1
1962	3	4	1	1	1972	4	4	0	0
1963	2	1	−1	1	1973	2	3	1	1
1964	1	1	0	0	1974	2	1	−1	1
1965	3	4	1	1	1975	1	1	0	0
1966	4	4	0	0	1976	4	4	0	0
1967	2	2	0	0	1977	3	3	0	0
1968	1	2	1	1	1978	1	1	0	0
1969	4	4	0	0	Σ				9
1970	2	3	1	1					

由上表可知：$n=18$，$\Sigma D^2=9$，且 $K=4$，代入(1.4)式得

$$r_s = 1 - \frac{6\times 9}{18(4^2-1)} = 0.8$$

即稻谷空壳瘪粒状况与花期温度的相关关系为 0.8。

[例 1.4] 某地有一批水稻分期插秧及其相应的齐穗期资料(见表 1.5)，试计算这两个变量的秩次相关系数。

首先将插秧日期按早到晚次序排列，并编成秩次；然后将对应的齐穗期随之排上，也按早到晚编成秩次；计算秩次差 D 及其平方和 ΣD^2（见表 1.5）：

由上表可知：$n=18$，$\Sigma D^2=57$，代入(1.5)式得

$$r_s = 1 - \frac{6\times 57}{18(18^2-1)} = 0.9412$$

值得注意的是，在应用秩次相关编成秩次时，如遇两个变量的数值相等时，则取其

表 1.5　水稻插秧日与齐穗日的秩次相关

编号	插秧日	秩次 X	齐穗日	秩次 Y	D	D^2	编号	插秧日	秩次 X	齐穗日	秩次 Y	D	D^2
1	10/5	1	28/7	1	0	0	10	28/5	10	8/8	10	0	0
2	12/5	2	30/7	2.5	0.5	0.25	11	30/5	11	6/8	7	−4	16
3	14/5	3	30/7	2.5	−0.5	0.25	12	2/6	12	16/8	15	3	9
4	16/5	4	1/8	4	0	0	13	4/6	13	17/8	16	3	9
5	18/5	5	5/8	6	1	1	14	6/6	14	13/8	13	−1	1
6	20/5	6	3/8	5	−1	1	15	8/6	15	12/8	12	−3	9
7	22/5	7	7/8	8.5	1.5	2.25	16	10/6	16	15/8	14	−2	4
8	24/5	8	7/8	8.5	0.5	0.25	17	12/6	17	18/8	17	0	0
9	26/5	9	10/8	11	2	4	18	14/6	18	19/8	18	0	0

秩次的平均数填入其中，如表 1.5 中，齐穗期有两个都是七月三十日，则取其秩次 2 与 3 的算术平均值 2.5 填入其中。

单相关系数与符号相关、秩次相关系数进行比较，其精度以单相关系数为最高，符号相关系数为最低，尤其是在样本容量不大的情况下更是如此。因此，在数字信息比较系统的情况下，还是计算单相关系数为好，特别是在袖珍计算器已经普及的今天，计算单相关系数已是轻而易举的事了。但在没有系统数据，而变量又是不连续的情况下，秩次相关就充分显示其优越性了。

六、相关系数的显著性检验

前述几节所讨论的相关系数都是以样本资料为基础进行的，故称其为样本相关系数。它不是真正总体的相关系数，而只是总体相关系数的一个近似值，它的精确程度取决于样本的观测方法、样本容量的多少、样本资料的准确性以及其他偶然因素等。那末究竟怎样判断样本相关系数能否代表总体相关系数呢？即如何估计样本相关系数对总体相关系数的近似程度以及两个样本相关系数的比较等，为此需要对相关系数进行显著性检验。只有当样本相关系数达到某一水准显著时，才认为样本相关系数能够代表总体相关系数。下面介绍几种相关系数显著性检验的方法。

1. *或然差法*

当样本容量很大时，取自同一母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：

$$\sigma_r = \frac{1-r^2}{\sqrt{n-1}} \doteq \frac{1-r^2}{\sqrt{n}} \tag{1.6}$$

根据相关系数 r 的正态分布曲线，我们很容易求出由于偶然机会而得到的相关系数 $r \geqslant r_0$ 的概率。在实际应用中，一般采用的概率标准为 $\alpha = 0.05$ 与 $\alpha = 0.01$，此时的 t 值分别为 $t \geqslant 1.96$ 与 $t \geqslant 2.60$。因而只有 $r \geqslant 1.96t$ 与 $r \geqslant 2.60t$ 时，才表示相关系数分

别达到 0.05 与 0.01 的显著水平。

但在实际业务中，往往并不采用这样的具体计算，而用相关系数 r 大于或等于或然误差的四倍来表示相关显著。或然误差公式为

$$P\cdot E\cdot r=\pm 0.6745\,\sigma_r=\pm 0.6745\frac{1-r^2}{\sqrt{n}} \tag{1.7}$$

这个公式的物理意义是，在前述一组相关系数中，任选一个，其误差介于 $-0.6745\,\sigma_r$ 与 $+0.6745\,\sigma_r$ 之间的或然误差率为 50%。由此可见，误差大于 $4(P\cdot E\cdot r)$ 的或然率是可以忽略不计的，据此可检验相关系数是否达到显著水平。

若相关系数是由秩次相关计算的，其或然误差公式为

$$P\cdot E\cdot r=\pm 0.7063\frac{1-r^2}{\sqrt{n}} \tag{1.8}$$

或然误差法适用于样本数目足够大时，即 $n>500$，至少应当 $n>100$ 时应用效果才好，否则会产生较大的误差，因为样本数目较少时，r 的分布不是正态分布，应用此法不太恰当，需要采用其它方法去检验相关系数的显著性。

2. *t 检验法*

当样本容量较少时，相关系数 r 的分布往往偏倚，即呈一偏态分布，因此不能用上述方法进行相关系数的显著性检验。根据费希尔 (Fisher) 指出的，假设总体相关系数 $\rho=0$，可以采用 t 检验方法去检验相关系数的显著性。则 t 的统计量为

$$t=r\sqrt{(n-2)/(1-r^2)} \tag{1.9}$$

它遵循自由度 df (或 f) $=n-2$ 的 t 分布。于是用样本相关系数计算统计量，并给定 α，根据 t 分布表 (附表 1)，查得自由度 $df=n-2$ 时 t 的机率值 t_α。若 $|t|\geqslant t_\alpha$，则表示相关显著；若 $|t|<t_\alpha$，则相关不显著。

[例 1.5] 根据例 1.1 的资料计算所得的单相关系数 $r=0.8886$，试检验其相关系数是否显著。

由例 1.1 可知，样本容量 $n=18$，则由 (1.9) 式得

$$t=0.8886\sqrt{(18-2)/(1-0.8886^2)}=7.7491$$

给定 $\alpha=0.01$，查附表 1，得到自由度 $df=16$ 时，其 $t_{0.01}=2.921$。由于 $|t|>t_{0.01}$，故该相关系数 $r=0.8886$ 在 $\alpha=0.01$ 的水平上显著。这说明从正态总体随机抽得的全部样品中，有 1% 样品的 t 的绝对值大于 2.921，即有 99% 这类样品的 t 值处于 ± 2.921 之间。结论是水稻产量与年序的相关系数是高度显著的。

秩次相关也可以采用 t 检验法进行显著性检验。则 t 的统计量为

$$t=r_s\sqrt{(n-2)/(1-r^2)} \tag{1.10}$$

如根据例 1.4 的资料，所得的秩次相关系数 $r_s=0.9412$，根据(1.10)式计算得

$$t=0.9412\sqrt{(18-2)/(1-0.9412^2)}=11.1434$$

给定 $\alpha=0.01$，查附表 1，得到 $df=16$ 时，$t_{0.01}=2.921$。由于 $|t|>t_\alpha$，故该相关系数 $r_s=0.9412$ 在 $\alpha=0.01$ 水平上显著，即水稻齐穗期与插秧期呈显著正相关。

3. *相关系数检验表*

如前所述，相关系数经 t 检验后，再查 t 分布表去判断相关显著与否，其过程还是

显得繁琐。为减轻这冗繁的计算查表，有人根据前述原理编制了直接利用相关系数进行显著性测定的“相关系数检验表”（见附表 2）。一般地说，由于抽样误差的影响，使相关系数 r 达到显著的值与抽样的个数 n 有关。在附表 2 中，给出了不同的 $(n-2)$ 在两种显著性水平上（$\alpha=0.05$ 与 $\alpha=0.01$）达到显著的相关系数最小值。例如，在例 1.1 中的 $n=18$，即抽取了 18 个样本，则自由度为 $df=(n-2)=16$，相关系数 $r=0.8886$。查附表 2 得到 $\alpha=0.05$ 时的相关系数 $r_{0.05}=0.468$。由于 $r>r_{0.05}$，故可以认为相关系数 $r=0.8886$ 在 $\alpha=0.05$ 的水平上显著；若 $r>r_{0.01}(=0.59)$ 时，则相关系数 $r=0.8886$ 在 $\alpha=0.01$ 的水平上显著。显然，α 愈小，其相关系数的显著程度愈高；α 愈大，显著程度愈低。可见，对相关系数的显著性检验，这两种方法所得的结论是一致的。

4. *Z 转换法*

当样本容量不大时，假设总体相关系数 $\rho\neq 0$，或者 $\rho_1-\rho_2=0$ 以及比较两个样本相关系数的差异时，用上述的方法显然是不适宜的。这时，相关系数 r 的分布特别偏斜，对于这样的偏态分布，费氏给出了从 r 到 Z 的转换关系，使这个 Z 值近乎正态分布，我们称通过 Z 值去检验相关系数的方法为 Z 转换法。这个变换的 Z 值的平均数为

$$\mu=0.5\ln\frac{1+\rho}{1-\rho} \tag{1.11}$$

标准差为

$$\sigma_z=\frac{1}{\sqrt{n-3}} \tag{1.12}$$

可见这个标准差实际上和在样品所抽自的总体中的相关值无关，只与样本容量有关。Z 对 r 的转换关系：

$$Z=0.5\ln\frac{1+r}{1-r} \tag{1.13}$$

在比较一个样本相关系数 r 与一个总体相关系数 ρ（假设 $\rho\neq 0$）的差异时，其判别式为

$$t=\frac{Z-r}{\sigma_z} \tag{1.14}$$

其自由度 $df=\infty$，$t_{0.01}=2.576$。

若比较两个样本相关系数 r_1 与 r_2 或者两个总体相关系数 ρ_1 与 ρ_2 的差异时，仿照（1.11）式、（1.13）式，得 Z 转换公式：

$$\begin{cases} Z_1=0.5\ln\dfrac{1+r_1}{1-r_1} \\ Z_2=0.5\ln\dfrac{1+r_2}{1-r_2} \end{cases} \tag{1.15}$$

$$\begin{cases} \mu_1 = 0.5 \ln \dfrac{1+\rho_1}{1-\rho_1} \\ \mu_2 = 0.5 \ln \dfrac{1+\rho_2}{1-\rho_2} \end{cases} \tag{1.16}$$

显然，$(Z_1 - Z_2)$的分布近乎正态分布，其平均数为$(\mu_1 - \mu_2)$，标准差为

$$\sigma_{z_1 - z_2} = \sqrt{\sigma_{z_1}^2 + \sigma_{z_2}^2} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}} \tag{1.17}$$

于是得到比较两个样本相关系数差异显著与否的公式：

$$t = \frac{Z_1 - Z_2}{\sigma_{z_1 - z_2}} = \frac{Z_1 - Z_2}{\sqrt{\dfrac{1}{n_1 - 3} + \dfrac{1}{n_2 - 3}}} \tag{1.18}$$

给定α，查附表1，得t_α值。若$|t| \geqslant t_\alpha$，则两个样本相关系数差异显著；反之，$|t| < t_\alpha$，则差异不显著，即样本相关系数之间的差异可能是由于随机误差造成的，并非是真正的差异。

在实际工作中，常常利用 Z 值转换公式制成的 Z 值表（附表 3）中的 Z 值去计算 t 值的大小。现举例说明计算方法。

[例 1.6] 由例1.1的计算结果可知：水稻产量与年序的单相关系数 $r_1 = 0.8886$，$n_1 = 18$；由例1.2可知：利用符号相关公式计算所得上述两变量的相关系数$r_2 = 0.9397$，$n_2 = 18$。试检验这两个相关系数的差异。

根据两个样本相关系数查附表3，当 $r_1 = 0.8886$ 时，得 $Z_1 = 1.42$；当 $r_2 = 0.9397$ 时，得 $Z_2 = 1.74$。代入(1.17)式得

$$\sigma_{z_1 - z_2} = \sqrt{\frac{1}{18 - 3} + \frac{1}{18 - 3}} = \sqrt{\frac{2}{15}}$$

再代入(1.18)式得

$$t = \frac{1.42 - 1.74}{\sqrt{\dfrac{2}{15}}} = -0.864$$

查附表 1，由于 Z 近乎正态分布，与样本含量无关，故 $df = \infty$ 时，$t_{0.01} = 2.576$。因为 $|t| < t_{0.01}$，所以这两个相关系数 r_1 与 r_2 的差异不显著，即直接计算的单相关系数与符号相关系数无显著差异，其仅有的小差异是由于随机误差造成的，而不是真正的差异。

5. 符号检验表

符号检验表（附表 4）仅适用于符号相关的情形。在符号相关作图（详见本节四）后，如无需知道两个变量的符号相关系数的大小时，我们可以利用作图时所给定的符号值直接查符号检验表，以判定两个变量间的相关程度。因为在普查因子时，我们并不太关心相关系数的大小，而特别关心两个变量的相关是否显著，因此这种检验是十分简捷的。如本节四那样，作图后得到的符号值：$n_+ = 16$，$n_- = 2$，于是 $N = n_+ + n_- = 16 + 2 = 18$。