气象统计预报

王得民　施　能　编

气　象　出　版　社

前　言

我院自1975年起开设气象统计预报课，已达十五年之久。在1979年9月由王得民、孙照渤、施能、程极益等四同志编成本课程试用教材，从1977年级本科生开始使用。经过五届教学实践后，发现原教材中所存在的一些缺点和错误。于1985年对教材进行了修改。在那次修改中，由施能同志补充编写本书第七章，天气预报经济效果评定及最优利用章节，其余部分由王得民同志负责修改。全部修改稿由吴洪宝同志审校。由于水平所限，以及计算机普及应用，气象统计方法、理论的发展，因此本书肯定还存在一些缺点和问题，欢迎读者提出批评指正的意见。

本书的出版得到气象出版社的大力支持，书中的插图由汪仲梅同志协助绘制，在这里向他们致以衷心的感谢。

编者于南京气象学院

1990年10月

目　录

第一章　概率统计预报模式

第二章　相　关　与　回　归

第三章　判　别　分　析

第四章　聚　类　分　析

第五章　时间序列分析在气象预报中的应用

第六章　气象要素场的正交展开

第七章　天气预报质量检查与成绩评定

第一章　概率统计预报模式

第一节　概率统计预报模式的产生

地球上气象过程的发展历史保留在气象资料之中。成千上万的气象台站，不停顿地进行长年累月的观测，积累了多种多样的气象资料。这些资料的数量是极其丰富的，即使是存贮量很大的现代计算机，也只能存贮其中的一部分。对得到的这些历史资料，进行统计的逻辑的分析，可以建立不同时效的气象预报方法。这样建立起来的预报方法，称为统计预报方法。显然，预报的纯统计方法是不存在的。实际在建立气象预报方法时，除了对气象资料作统计分析之外，常常同时进行一些具有物理内容的计算，以便从庞大的历史资料中，找出天气过程演变的规律性，有效地提取对于预报具有信息意义的那部分。由于不同的预报问题，有不同的物理学和统计学的联系，这样就使预报问题变得更加复杂化，很难提出统一的预报方案。

为了对根据历史资料的统计作天气预报问题有一般的了解，在这里先作如下的说明。

若在一份预报中包含有几个预报项目，每个预报项目用一个变量表示。设要预报的L个量为y_1、y_2、…、y_L，它的全部可写为L维的(行)向量。若

$$Y_t = (y_1\ y_2 \cdots\cdots y_L)_t = (y_{t1}\ y_{t2} \cdots\cdots y_{tL})$$

附标t表示时间，y_{t1}、y_{t2}、…、y_{tL}表示时间t的y_1、y_2、…、y_L的观测值。Y_t称为预报对象向量，用如上的行向量表示。对Y_t的N次观测值，依次列于表1.1内的后部。

表1.1

时间	预报中利用的量（时间$t' \leqslant t-\tau$观测的预报因子的值）	时间t观测的预报量（预报对象的值）
1	$(x_{11}\ x_{12} \cdots\cdots x_{1m})$	$(y_{11}\ y_{12} \cdots\cdots y_{1L})$
2	$(x_{21}\ x_{22} \cdots\cdots x_{2m})$	$(y_{21}\ y_{22} \cdots\cdots y_{2L})$
⋮	……	……
t	$(x_{t1}\ x_{t2} \cdots\cdots x_{tm})$	$(y_{t1}\ y_{t2} \cdots\cdots y_{tL})$
⋮	……	……
N	$(x_{N1}\ x_{N2} \cdots\cdots x_{Nm})$	$(y_{N1}\ y_{N2} \cdots\cdots y_{NL})$

设Y是Y_t的N次观测值组成的矩阵，即

$$Y=\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_t \\ \vdots \\ Y_N \end{pmatrix}=\begin{pmatrix} y_{11} & y_{12} & \cdots\cdots & y_{1L} \\ y_{21} & y_{22} & \cdots\cdots & y_{2L} \\ & & & \\ y_{t1} & y_{t2} & \cdots\cdots & y_{tL} \\ & & & \\ y_{N1} & y_{N2} & \cdots\cdots & y_{NL} \end{pmatrix}$$

预报时，选择时间 t 的前期 t′观测的m个量作预报因子，这时要求

$$t' \leqslant t-\tau$$

显然，这是企图在时间t之前t′做出预报，τ为预报时效。若将选出的m个预报因子写成m维的预报因子的向量形式，这些因子在时间 t 的观测值组成的预报因子(行)向量为：

$$X_t=(x_1\ x_2\ \cdots\cdots\ x_m)_t=(x_{t1}\ x_{t2}\ \cdots\cdots\ x_{tm})$$

在N次观测内，由预报因子向量组成的矩阵为：

$$X=\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_t \\ \vdots \\ X_N \end{pmatrix}=\begin{pmatrix} x_{11} & x_{12} & \cdots\cdots & x_{1m} \\ x_{21} & x_{22} & \cdots\cdots & x_{2m} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ x_{t1} & x_{t2} & \cdots\cdots & x_{tm} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ x_{N1} & x_{N2} & \cdots\cdots & x_{Nm} \end{pmatrix}$$

应当注意，预报对象向量Y_t和预报因子向量X_t的附标t取相同值时，后者是比前者在较早时间得到的。

如果将上表的全部，用以下几种简要的符号来表达：

$$X \longrightarrow Y$$

$$X_t \longrightarrow Y_t$$

$$(x_1\ x_2\ \cdots\cdots\ x_m)_t \longrightarrow (y_1\ y_2\ \cdots\cdots\ y_L)_t$$

中间的箭头的意思是“相应”或“引起”。

建立预报方法的问题在于：利用历史资料选出对Y_t具有预报信息的那些因子去组成X_t，然后再去求得一定的变换或算法，将X_t变为Y_t，亦即需求出：

$$Y_t=J(X_t) \qquad t=1,\ 2,\ \cdots,\ N \qquad (1.1)$$

这里的 J 称为决策规则。实践中指出：应用历史资料去找到这种理想的变换是极其困难的。还需指出的是，即使用N次资料成功地找到这种变换，而对于没有包含在上述资料内的X_{N+1}，也不能用这种变换去求得相应的Y_{N+1}，亦即

$$Y_{N+1} \neq J(X_{N+1})$$

因此，对预报方法要放宽要求。如果要求找到这样的变换 J，使得：

$$\widehat{Y}_t = J(X_t) \qquad (1.2)$$

其中 $\widehat{Y}_t$ 是预报对象 Y_t 的预报发布值组成的向量，显然要求 $\widehat{Y}_t$ 与 Y_t 之间差异达到最小。也就是说，要求计算的 $\widehat{Y}_t$ 与实测的 Y_t 达到最小差异或最大相似，实质上就是要求这种变换是最优的。

(1.2)式说明存在预报误差 δ_t，因而可换写成：

$$Y_t = J(X_t) + \delta_t \qquad (1.2)'$$

(1.2)或(1.2)′式和(1.1)式一样，是一种预报模式。它们都是用右方的 X_t 经过变换J去预报 Y_t。如果变换J是从基本的物理定律数学地引入，那么这种预报模式称做理论模式。在气象预报中，这种类型的预报模式就是热力学—流体动力学模式。在这种模式中，预报对象和预报因子都是非随机的物理量，它们之间存在确定性关系，预报由这些确定性规律做出。如果变换 J 是由统计预报因子和预报对象的观测资料之间的关系，用一定的函数模式而得出，则称这种预报模式是经验模式。在这种模式中，预报因子和预报对象都作为随机变量来讨论，因而概率论和数理统计方法是建立这种模式的主要工具，常又称做概率统计模式，简称为统计模式，这时预报是应用统计规律性作出。

再说明一点，统计模式(1.2)是用来归纳出预报对象和预报因子之间统计联系的规律。因为预报对象和预报因子是属于不同时间的，则可以认为：预报的统计方法的依据，是大气过程和天气现象之间存在着非同期的统计联系。

若提出衡量预报误差的尺度，亦即损失为

$$L = L(\widehat{Y}, Y)$$

或相反的预报有效尺度如收益等。据(1.2)式可知：

$$L = L[J(X), Y]$$

根据历史资料能估计损失的平均值或数学期望

$$\overline{L} = E\{L[J(X), Y]\}$$

$\overline{L}$ 取决于决策规则 J，称为风险函数。统计研究的任务，是要使风险函数达到最小而去求得决策规则 J。选择决策规则具体说来有许多种不同的途径，但却常导致类似的结果。

模式(1.2)表示了预报对象对预报因子的依赖关系。在概率统计模式中，提出研究预报对象取决于预报因子的概率分布函数 $F(Y|X)$ 的问题。这时，研究的对象是概率或概率的预报。如果写成象(1.2)那样的形式：

$$F(Y|X) = J(X) \qquad (1.3)$$

(1.3)式讨论的是概率的统计预报方法。假设这种关系对时间来说是稳定的。如果

$F(Y|X)=F(Y)$，亦即在用到初始信息范围内，Y在统计上不决定于X，这时不能用X去预报Y。而作为概率预报，可以应用气候的分布函数$F(Y)$。

近代，由于数值的动力学预报的进步，在统计预报模式中的预报因子，引入动力学模式中采用的变量。这样构成的动力与统计相结合的预报模式，和经典的统计模式容易区别开来。在这种预报模式中，有一类称做完全预报（Perfect Prog.）方法，简称为PPM方法，预报时利用求出预报对象和同期动力学预报模式中选用变量之间的统计关系。这时认为动力学模式对这些选用变量的预报是完全正确的，在这种基础上按统计关系去做出预报对象的预报。另一类称做模式输出统计（Model output staticstics）方法，简称为MOS方法，预报时利用求出预报对象和预期时效内动力学模式中输出变量之间的统计关系，亦即将动力学模式的输出作为统计模式的输入，按统计模式作出预报对象的预报。上述两类方法在导出与使用变换J都和经典统计模式不一样，故列出表1.2以示差别。

表1.2

方　　法	导　出　变　换	预　报　中　应　用
经典统计方法	$\hat{Y}_t = J_1(X_0)$	$\hat{Y}_t = J_1(X_0)$
完全预报方法	$\hat{Y}_t = J_2(X_t)$	$\hat{Y}_t = J_2(\hat{X}_t)$
模式输出统计方法	$\hat{Y}_t = J_3(\hat{X}_t)$	$\hat{Y}_t = J_3(\hat{X}_t)$

注：N的附标0表示初始时刻，故有$t>0$。

第二节　预报对象和预报因子

预报模式（1.2）的两侧，是预报对象和预报因子。比较起来，表述预报对象的未来状态比预报因子的初始状态要简单一些，我们先分析一下预报对象的特征。在预报对象中列入日常预报项目的那些气象要素或天气现象，有些能用数量表征，有些则不能用数量表征。前一类例如气温是几度，雨量有几毫米。后一类例如有无雷暴，降水性质是连续性、间歇性还是阵性。但是，不论是否能用数量表征，常可以将每个项目分为有限的几类，做这个项目的分类预报。一般说，根据预报使用部门的要求，大部分预报项目只需作出定性的分类预报就可以了。如在短期预报中，雨量只需报出小雨、中雨、大雨、暴雨等级；风只需报出风向属那一方位，风力几级。在长期预报中，分类可更少些，只需报出雨量、雨日是否正常，偏多或偏少，特多或特少；温度是否正常，偏冷或偏暖，特冷或特暖等情况。分成3到5级就已经是足够了。因此，从这一方面说来，预报就有了简化的可能。在基层气象台站的日常工作中，常常采用一些简便而实用的预报方法，大多是从做分类预报入手的。当然，在实际的预报问题中，也有连续变量的预报，描述它的

空间分布即场的预报，后者即通常所说的区域预报，进一步研究这些较为复杂的问题也是很有必要的。

预报对象的分类，能用一种或几种气象要素来进行。分类时要考虑实际的需要，利用资料的容量和计算的可能，以及相应量的准确性和自然的变异性。例如，对一种要素 y_i 选择比 k_i' 要更多的类别是没有意义的，这时

$$k_i' = \frac{y_{i\max} - y_{i\min}}{\delta y_i} + 1$$

其中 $y_{i\max}$ 和 $y_{i\min}$ 相应为 y_i 的最大值和最小值，δy_i 是它的测量的准确性。

最简单的分类预报就是分成两类，称为二分类预报或正反预报。例如有无霜冻，有无雷暴，有无低温阴雨天气等等。二分类预报是分类预报的基础，因为有限的分类预报总可以划分成几个二分类问题来处理。

若将 y_i 分成 k_i'' 类（$k_i'' \leqslant k_i'$），则对 Y 中可能形成的天气状态总数为

$$k_L = \prod_{i=1}^{L} k_i''$$

实际上，其中有些天气状态是不可能观测得到的。例如，不可能有如下的结合："晴天"和"有强的降水"，"雾"和"低于50%的空气湿度"等等。我们从历史资料内统计出各种天气状态出现的气候频率，除了气候频率等于零的那些天气状态，得出了实际能观测到的 $k(k \leqslant k_L)$ 种天气状态。以下将表1.1内的历史资料转换成表1.3内的形式。

表1.3

时间	预报中利用的量（时间 $t' \leqslant \tau$ 的预报因子向量）	时间 t 的天气状态 $(\Phi_i)_t$，$i \ni (1, \overline{k})$
1	$(x_1\ x_2 \cdots\cdots x_m)_1$	$(\Phi_i)_1$
2	$(x_1\ x_2 \cdots\cdots x_m)_2$	$(\Phi_i)_2$
⋮	……	…
t	$(x_1\ x_2 \cdots\cdots x_m)_t$	$(\Phi_i)_t$
⋮	……	…
N	$(x_1\ x_2 \cdots\cdots x_m)_N$	$(\Phi_i)_N$

上表可用以下的符号表达：

$$X_t \longrightarrow (\Phi_i)_t \qquad t=1,\ 2,\ \cdots,\ N;\ i \in (1,k)$$

i 表示数集 $(1,\overline{k})$ 中的第 i 个。这时建立预报方法的问题变为：利用历史资料，选出对

$(\Phi_t)_t$具有预报信息的那些因子去组成X_t，然后再去求得一定的变换或算法，将X_t变为$(\Phi_t)_t$，亦即求得：

$$(\hat{\Phi}_t)_t = J(X_t) \tag{1.4}$$

这里的变换J要求使得时间t的预报信息$(\hat{\Phi}_t)_t$和相应的观测状态$(\Phi_t)_t$之间关系是最优的。例如，使不成功次数达到最少等，具体做法，不详细讨论。

以下分析一下预报因子的特征。预报因子的挑选是统计预报的关键。可供挑选的预报因子，不仅包括历史资料各种气象要素的观测值，而且包括各种各样的导出量。例如表示空间分布的梯度、涡度、垂直风切变、垂直运动速度、层结稳定度；表示大范围水平分布的一些指数的如环流指数；由气象要素水平分布和流场配置关系决定的平流输送量。又如表示气象要素随时间的变化量，在一定时间内的平均量、累积量等。预报因子中，也可包括非大气因子，例如太阳活动指数、海水温度、下垫面状况等。可以看出，可供挑选的预报因子是非常多的。对于某项预报，挑选那些因子来做预报，需要尽可能地了解预报因子和预报对象之间的关系、相制约的机制和过程。革命导师马克思应用唯物辩证法分析了资产阶级社会最简单、最基本的、最常见的、最平常的、碰到亿万次关系—商品交换，暴露了商品社会的一切矛盾。我们要学会用唯物辩证法去观察和分析天天碰到的天气和天气的转变过程，找出制约天气发生和转变的主要矛盾。如果从统计方法的角度来看，应该有一个标准，用来去衡量预报对象和预报因子之间关系的紧密程度，从而剔除掉与预报对象关系不紧密的因子，筛选出关系比较密切的因子。由于可供挑选的因子非常多，找因子时一方面要依赖于计算工具提高挑选效率，另一方面要求前述的衡量标准尽量要简单且易于计算，这样才可以从大量因子中筛选出关系好的因子来。在实际工作中，常常在计算前对预报因子加以处理，变成特征资料，研究它与预报量之间的关系。当我们找出一些与预报量关系密切的因子后，这时就应该考虑用什么方法去综合这些因子的作用，去作出预报量的预报。

第二章　相关与回归

广大的劳动人民在长期与自然作斗争的过程中，积累了丰富的预报经验。这些经验，往往反映气象要素或天气现象相互之间存在着非同时的相关联系。气象台站的工作人员，需将这种相关联系数量地表示出来，同时还研究它的可靠程度，以便在实际预报工作中应用时加以考虑。相关和回归的分析，便是研究这些预报关系的一种基本的重要方法。

第一节　相关和回归的概念

一、　相关与回归

在数学中，反映两个变量之间的函数关系可表示为：

$$y = f(x)$$

其中x是自变量，y是因变量。当x取值确定了，y的取值便确定了。如果因变量y由多个自变量x_1、x_2、…、x_m所确定，可表示为：

$$y = f(x_1, x_2, \cdots, x_m)$$

其右方为多元函数。总之，这两者都是反映自变量和因变量之间有确定的函数关系。

类似地，如果去研究随机变量之间的关系，並且企图用某个或某些随机变量的取值去推测另一个随机变量的可能取值。在这些问题中，其性质和变量之间的函数关系有着根本的区别。例如讨论随机变量ξ和η，可能ξ的取值对η的取值有影响，也可能没有影响。如果是有影响，认为ξ和η之间存在着相关联系，那末如何估计这种联系的紧密程度？能不能找到它们之间联系的数学形式，用ξ的取值去估计η呢？这些问题，都属于本章讨论范围之内。

（一）　回归线和回归面

设(ξ, η)是连续型的二维随机变量。如果用$f(\xi)$去估计η，在最小二乘方的意义下是最好的。若用$E\eta$表示随机变量η的数学期望，用$E(\eta|\xi=x)$表示它的条件数学期望。这时，需具有条件：

$$E[\eta - f(\xi)]^2 = \min \tag{2.1}$$

由方差的性质可知，当$f(x)=E(\eta|\xi=x)$时，上式成立。于是称

$$y=f(x)=E(\eta|\xi=x) \tag{2.2}$$

为η关于ξ的回归线。推广到研究随机变量η与m个随机变量ξ_1、ξ_2、…、ξ_m之间的关系，也是根据条件

$$E[\eta - f(\xi_1, \xi_2, \cdots, \xi_m)]^2 = min \tag{2.3}$$

去求，将m维空间中的曲面

$$y = f(x_1, x_2, \cdots, x_m) = E(\eta | \xi_1 = x_1, \xi_2 = x_2, \cdots, \xi_m = x_m) \tag{2.4}$$

称为η关于ξ_1、ξ_2、…、ξ_m的回归面。

（二） 线性回归和线性相关

如果表示回归线和回归面的函数是线性函数，即

$$y = E(\eta | \xi = x) = \beta_0 + \beta x \tag{2.5}$$

$$y = E(\eta | \xi_1 = x_1, \xi_2 = x_1, \cdots, \xi_m = x_m)$$

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \tag{2.6}$$

则分别称之为回归直线和回归超平面。

常常在度量随机变量ξ和η之间的线性相关程度时，采用(线性)相关系数$\rho_{\xi\eta}$，它可表示为：

$$\rho_{\xi\eta} = \frac{E(\xi - E\xi)(\eta - E\eta)}{\sqrt{D\xi \cdot D\eta}} \tag{2.7}$$

式中Eξ、Eη表示ξ、η的数学期望，Dξ、Dη表示ξ、η的方差。如果ξ与η线性相关，则有$|\rho_{\xi\eta}| = 1$；如果ξ与η线性相互独立，则$\rho_{\xi\eta} = 0$。

（三） 样本相关系数的计算

在实际问题中，往往是根据样本观测资料，去计算样本相关系数r_{xy}。如果随机变量ξ及η的容量为n的样本资料，其测值依次为x_1、x_2、…、x_n及y_1、y_2、…、y_n，则r_{xy}可这样计算：

$$r_{xy} = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2 \cdot \frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y})^2}}$$

$$= \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{n}(x_t - \bar{x})^2 \cdot \sum_{t=1}^{n}(y_t - \bar{y})^2}}$$

$$= \frac{\sum_{t=1}^{n} x_t y_t - n\bar{x}\,\bar{y}}{\sqrt{\left(\sum_{t=1}^{n} x_t^2 - n\bar{x}^2\right)\left(\sum_{t=1}^{n} y_t^2 - n\bar{y}^2\right)}}$$

$$= \frac{\sum_{t=1}^{n} x_t y_t - \frac{1}{n}\sum_{t=1}^{n} x_t \sum_{t=1}^{n} y_t}{\sqrt{\left[\sum_{t=1}^{n} x_t^2 - \frac{1}{n}\left(\sum_{t=1}^{n} x_t\right)^2\right]\left[\sum_{t=1}^{n} y_t^2 - \frac{1}{n}\left(\sum_{t=1}^{n} y_t\right)^2\right]}} \qquad (2.8)$$

上式中$\bar{x}$、$\bar{y}$表示样本平均数，即$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$，$\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$。

以下举例说明样本相关系数的计算。

例1　设有样本资料如表2.1内所示。表中的$\Delta y_t = y_t - \bar{y}$，$\Delta x_t = x_t - \bar{x}$。据(2.8)式的第二式可得样本相关系数

$$r_{xy} = \frac{24}{\sqrt{64 \cdot 10}} = \frac{24}{25.3} = 0.9486$$

表2.1

t	1	2	3	4	5	6	7	合　计	平　均
y	3	5	4	8	0	-2	3	21	3
x	4	4	5	6	3	2	4	28	4
Δy	0	2	1	5	-3	-5	0		
Δx	0	0	1	2	-1	-2	0		
$(\Delta y)^2$	0	4	1	25	9	25	0	64	
$(\Delta x)^2$	0	0	1	4	1	4	0	10	
$\Delta y \cdot \Delta x$	0	0	1	10	3	10	0	24	

例2　设y表示南京初霜日期距9月30日的天数，x表示南京2月份降水量。在1951—1971年的21年内，其测值如表2.2内所示：

表2.2

t	1951	52	53	54	55	56	57	58	59	60	61
y	55	65	49	49	39	36	30	23	30	19	37
x	94	107	82	77	65	6	51	29	90	26	18
t	1962	63	64	65	66	67	68	69	70	71	合计
y	36	57	43	57	28	44	30	38	47	47	859
x	46	18	78	57	27	55	4	45	84	40	1099

据表2.2的资料，求样本相关系数。

因$\bar{x}$、$\bar{y}$不是整数，采用(2.8)式的第4式比较准确。可求得：

$$\sum_{t=1}^{n} x_t y_t = 48843, \quad \frac{1}{n}\sum_{t=1}^{n} x_t \sum_{t=1}^{n} y_t = \frac{1}{21} \times 859 \times 1099 = 44954,$$

$$\sum_{t=1}^{n} x_t^2 = 75965, \quad \frac{1}{n}\left(\sum_{t=1}^{n} x_t\right)^2 = \frac{1}{21} \times (1099)^2 = 57514,$$

$$\sum_{t=1}^{n} y_t^2 = 38053, \quad \frac{1}{n}\left(\sum_{t=1}^{n} y_t\right)^2 = \frac{1}{21} \times (859)^2 = 35137,$$

$$r_{xy} = \frac{3889}{\sqrt{18451 \times 2916}} = \frac{3889}{7335} = 0.5302$$

从计算中可见，随着样本容量的增加，各测值数值大或要求准确度高，用手工计算样本相关系数是相当麻烦的，因此需用简化方法估计两个变量之间的相关程度。在这些简化的方法中，秩相关系数是工作中常用的一种。

（四） 秩相关系数

秩相关系数也叫顺序相关系数。它是两种变量的测值用它们在样本中的大小顺序(即秩)代替而计算出的相关系数。这时，容量为n的样本资料，无论对x_t或y_t，各测值秩的和都等于：

$$\sum_{t=1}^{n} x_t = \sum_{t=1}^{n} y_t = 1 + 2 + \cdots + n = \frac{1}{2} n(n+1)$$

秩平均等于

$$\overline{x}=\overline{y}=\frac{1}{2}(n+1)$$

秩的平均和等于

$$\sum_{t=1}^{n}x_t^2=\sum_{t=1}^{n}y_t^2=1^2+2^2+\cdots+n^2=\frac{n}{6}(n+1)(2n+1)$$

且有

$$\sum_{t=1}^{n}x_t^2-n\overline{x}^2=\sum_{t=1}^{n}y_t^2-n\overline{y}^2=\frac{n}{12}(n^2-1)$$

应用计算 r_{xy} 的第 3 式，並考虑以上关系可得：

$$r_{xy}=\frac{\sum_{t=1}^{n}x_ty_t-n\overline{x}^2}{\sum_{t=1}^{n}x_t^2-n\overline{x}^2}=1-\frac{\sum_{t=1}^{n}x_t^2-\sum_{t=1}^{n}x_ty_t}{\sum_{t=1}^{n}x_t^2-n\overline{x}^2}$$

设 $d_t=x_t-y_t$，则有：

$$\sum_{t=1}^{n}d_t^2=\sum_{t=1}^{n}(x_t-y_t)^2=\sum_{t=1}^{n}x_t^2-2\sum_{t=1}^{n}x_ty_t+\sum_{t=1}^{n}y_t^2$$

$$=2\left(\sum_{t=1}^{n}x_t^2-\sum_{t=1}^{n}x_ty_t\right)$$

代入上式后可得：

$$r_{xy}=1-\frac{6\sum_{t=1}^{n}d_t^2}{n(n^2-1)} \tag{2.9}$$

对上面例 2 计算该样本的秩相关系数。各测值的秩由小到大计算，测值相等时用平均秩代替。将例 2 的各次测值用它的秩代替，可得表2.3。用表2.3内的资料求得：

$$\sum_{t=1}^{n}d_t^2=751,\quad r_{xy}=1-\frac{6\times751}{21\times(21^2-1)}=0.512$$

可见两次算得的样本相关系数，在数值上差别不大，故常用秩相关系数粗略地估计相关系数的大小。

表2.3

t	1951	52	53	54	55	56	57	58	59	60	61
y	18	21	16.5	16.5	11	7.5	5	2	5	1	9
x	20	21	17	15	14	2	11	7	19	5	3.5
t	1962	63	64	65	66	67	68	69	70	71	
y	7.5	19.5	12	19.5	3	13	5	10	14.5	14.5	
x	10	3.5	16	13	6	12	1	9	18	8	

二、 相关系数的统计检验

由样本资料算得相关系数的大小，是由取得的样本所决定的。而样本是从总体内随机抽取的，所以样本相关系数应是一个随机变量。特别对于较短的样本资料，从同一总体内抽取的样本算得的相关系数可能在数值上差得很大，怎样根据样本确定这两个随机变量总体之间是否确实存在相关呢？这就需要进行对相关系数的统计检验。

(一) 总体相关系数ρ=0的假设检验

进行相关系数的统计检验时，首先要求得样本相关系数 r 的概率分布函数，这是一个比较复杂的问题。

假设讨论相关的两个随机变量，其联合分布服从二元正态分布时，则其样本相关系数 r 的概率密度函数为：

$$f(r)=\frac{n-2}{\pi}(1-\rho^2)^{\frac{n-4}{2}}(1-r^2)^{\frac{n-4}{2}}\int_0^1\frac{x^{n-2}}{(1-\rho r x)^{n-1}\sqrt{1-x^2}}dx$$

其中ρ是总体相关系数，n是样本容量。

当 $\rho=0$，则有：

$$f(r)=\frac{n-2}{\pi}(1-r^2)^{\frac{n-4}{2}}\int_0^1 x^{n-2}(1-x^2)^{-\frac{1}{2}}dx$$

作变换 $z=x^2$，可得：

$$f(r)=\frac{n-2}{2\pi}(1-r^2)^{\frac{n-4}{2}}\int_0^1 Z^{\frac{n-1}{2}-1}(1-Z)^{\frac{1}{2}-1}dZ$$