高 等 学 校 教 材

时间序列分析
简 明 教 程

张树京　齐立心　编著

清 华 大 学 出 版 社
北方交通大学出版社
·北京·

内 容 简 介

本书根据电子信息类本科教学计划，着重介绍时间序列分析的基本模型和算法，适用作信息工程、通信工程、控制工程等相关专业的本科教材或教学参考用书。全书共分 6 章，其内容安排如下：第 1 章动态数据预处理，第 2 章时间序列模型，第 3 章模型参数估计，第 4 章模型定阶方法，第 5 章时间序列建模，第 6 章时间序列应用。各章均有小结和习题，有助于教学。

图书在版编目 (CIP) 数据

时间序列分析简明教程/张树京，齐立心编著. —北京：北方交通大学出版社，2003.9
ISBN 7-81082-136-9

Ⅰ. 时… Ⅱ. 张… Ⅲ. 时间序列分析－高等学校－教材 Ⅳ. O211.61

中国版本图书馆 CIP 数据核定 (2003) 第 039552 号

责任编辑：韩 乐
印 刷 者：北京东光印刷厂
出版发行：北方交通大学出版社 邮编：100044 电话：010-51686045，62237564
清 华 大 学 出 版 社 邮编：100084
经 销：各地新华书店
开 本：787×960 印张：11.5 字数：258 千字
版 次：2003 年 9 月第 1 版 2003 年 9 月第 1 次印刷
印 数：2000 册 定价：23.00 元

前　言

时间序列(Time Series)分析是概率统计学科的一个分支,它是运用概率统计的理论和方法来分析随机数据序列(或称动态数据序列),并对其建立数学模型,进行参数估计,对模型定阶,以及进一步应用于预报、预测、自适应控制、最佳滤波等诸多方面。

时间序列分析方法与随机过程理论有所区别,前者是先对实测数据建立数学模型,并在此基础上进一步分析随机数据的统计特性;后者是在对实测数据统计所得的先验概率知识基础上来分析其统计特性。由于人们所能获得的实测数据总是有限的,而理论上的先验概率要求在无限多的样本数据基础上统计才能获得,因此实际上我们能够获得的先验概率只能是在一定置信度条件下的近似,亦即尽量接近真实的概率(密度)分布,这是随机过程理论和方法在实际应用时的困难。时间序列分析方法可以克服这个困难,它是在有限个样本数据总量的情况下建立起相当精确的数学模型,从而获得具有一定精度(用模型误差方差来表示)的统计特性,与真实结果非常接近,因此在实际应用时比较方便,可操作性较好。总之,随机过程分析方法在理论上严谨求实,但操作性较差;而时间序列分析方法在使用时方便实用,但是,要想建立精度相当高的时序模型不仅要求模型参数最佳地估计,而且模型阶数也要合适,因此建模过程也是相当复杂的。这两种对随机数据序列的分析方法都有各自的研究和应用领域,应视不同的分析对象和要求而定。

时间序列分析方法最早起源于1927年,数学家耶尔(Yule)提出建立自回归(AR)模型来预测市场变化的规律。接着,在1931年,另一位数学家瓦尔格(Walker)在AR模型的启发下,建立了滑动平均(MA)模型和自回归、滑动平均(ARMA)混合模型,初步奠定了时间序列分析方法的基础,当时主要应用在经济分析和市场预测领域。20世纪60年代,时间序列分析理论和方法迈入了一个新的阶段,伯格(Burg)在分析地震信号时最早提出最大熵谱(MES)估计理论,后来有人证明AR模型的功率谱估计与最大熵谱估计是等效的,并称之为现代谱估计。它克服了用传统的傅里叶功率谱分析(又称经典谱分析)所带来的分辨率不高和频率漏泄严重等固有的缺点,从而使时间序列分析方法不仅在时间域内得到应用,而且扩展到频率域内,得到更加广泛的应用,特别是在各种工程领域内应用功率谱的概念更加方便和普遍。到20世纪70年代以后,随着信号处理技术的发展,时间序列分析方法不仅在理论上更趋完善,尤其是在参数估计算法、定阶方法及建模过程等方面都得到了许多改进,进一步地迈向实用化,各种时间序列分析软件也不断涌现,逐渐成为分析随机数据序列不可缺少的有效工具之一。

随着时间序列分析方法的日趋成熟,其应用领域越来越广泛,主要集中在预报预测领域,例如气象预报、市场预测、地震预报、人口预测、汛情预报、产量预测,等等。另一个应用领域是精密测控,例如精密仪器测量、精密机械制造、航空航天轨道跟踪和监控,以及遥控遥测、精细化工控制等。再一个应用领域是安全检测和质量控制。在工程施工和维修中经常会出现异常

险情,采用仪表监测和时间序列分析方法可以随时发现问题,及早排除故障,以保证生产安全和质量要求。以上仅仅列举了某些应用领域,实际上还有许多应用,不胜枚举。

根据电子信息工程类专业本科教学计划,许多高等学校已将时间序列分析列为一门必修课程,本书就是针对这个要求而编写的,适用于信息工程、通信工程、控制工程等相关专业的本科教材或教学参考用书。由于教学时数有限,本书偏重于时间序列分析的基础内容。

本书本着由浅入深、由易到难的原则,在先修课程(如概率论、统计学、随机过程等)的基础上,着重介绍时间序列分析的基本模型和基本计算方法,并且各章配有适当的例题和习题,以便帮助读者进一步理解。各章都有小结,有助于复习和教学。

本书受北方交通大学出版基金资助。

作　者

2003 年 9 月

目　录

第 1 章　动态数据预处理

在日常生活和科学实验中经常会遇到大量的实测数据，它们从表面上看来是杂乱无章的，但实际上具有一定的统计规律，因此人们总是想从这些有序的数据中找出它们的统计特性，并称它们为统计数据。最典型的例子就是在通信信道中存在的白噪声，它在示波器上显示为一堆茅草波形，但经过统计分析后发现它的统计特性符合高斯型（正态）概率分布，因此又称为高斯型噪声。

统计数据具有动态随机变化的属性，它们不断地更新，因此有时又称为动态随机数据。动态数据的统计特性可以用概率（密度）分布来描述，但是属于动态数据的随机过程往往具有很复杂的多维概率分布特性，实际上难以分析和应用。

时间序列分析是另外一种描述动态数据统计特性的理论和方法，它不同于用多维概率分布来描述动态数据的随机过程理论，其突出优点是方便和实用。有人将随机过程称为大样本理论，因为多维概率分布是要建立在无限多样本数据的统计基础上。时间序列分析则可以从有限的样本数据中拟合成具有一定精度的时间序列模型，因此它又可称为小样本理论。

在建立时间序列模型之前，必须先对动态数据进行必要的预处理，以便剔除那些不符合统计规律的异常样本，并对这些样本数据的基本统计特性进行检验，以确保建立时间序列模型的可靠性和置信度，并满足一定的精度要求。

本章将要研究动态数据预处理内容，主要包括平稳性检验、正态性检验、周期性检验和独立性检验。此外，还要对某些确知规律性的数据进行趋势性检验，以便保持被统计数据的纯随机性质。

1.1　平稳性检验

时间序列的平稳性是我们建模的重要前提。一般说来，某个实测过程如果它的系统参数和运行时周围的条件不改变，即可视为平稳的。不过这仅是一种定性判据，只从这样的物理背景来下结论，有时是不太可靠的，还需要依据一些统计方法进一步检验。

在检验时间序列的平稳性时，必须考虑以下两个内容：一是序列的均值（$\bar{x}_i$）和方差（σ_i^2）是否为常数，二是序列的自相关函数（r_i）是否仅与时间间隔有关，而与此间隔端点位置无关。下面介绍平稳性检验的两种常用方法。

1. 平稳性的参数检验法

设样本序列 $x_1, x_2, \cdots, x_N$ 足够长，即 N 相当大。把样本序列分成 k 个子序列，即取

$N=kM$，M 是一个较大的正整数，k 也是一个正整数。分段后的样本序列为

$$\left.\begin{array}{c} x_{11},x_{12},\cdots,x_{1M} \\ x_{21},x_{22},\cdots,x_{2M} \\ \vdots \\ x_{k1},x_{k2},\cdots,x_{kM} \end{array}\right\}=\{x_{ij}\} \tag{1-1}$$

这里 $x_{ij}=x_{(i-1)M+j}$，$i=1,2,\cdots,k$；$j=1,2,\cdots,M$。

对于 k 个子序列，可以分别计算它们的样本均值，样本方差和样本自协方差函数。它们的定义如下（详见第 3 章）：

$$\left.\begin{array}{l} \bar{x}_i=\dfrac{1}{M}\displaystyle\sum_{j=1}^{M}x_{ij} \\ \sigma_i^2=\dfrac{1}{M}\displaystyle\sum_{j=1}^{M}(x_{ij}-\bar{x}_i)^2 \\ r_i(\tau)=\dfrac{1}{M}\displaystyle\sum_{j=1}^{M-1}(x_{ij}-\bar{x}_i)(x_{i,j+\tau}-\bar{x}_i)/\sigma_i^2 \end{array}\right\} \tag{1-2}$$

$$i=1,2,\cdots,k;\tau=1,2,\cdots,m,m\ll M$$

由平稳性的假定，以上各统计量对不同的子序列 i 不应有显著的差异，否则就应否定 $\{x_t\}$ 是平稳序列的假定。

设 $\{x_t\}$ 具有理论上的均值 μ，方差 σ^2 和自相关函数 ρ_τ，这时样本统计量 $\bar{x}_i$，σ_i^2 及 $r_i(\tau)$ 的方差可由随机变量四阶矩的算式得到

$$\left.\begin{array}{l} \text{① 样本均值的方差 } \sigma_1^2=D(\bar{x}_i)=\dfrac{1}{M^2}E\left(\displaystyle\sum_{j=1}^{M}\sum_{l=1}^{M}(x_{ij}-\mu)(x_{il}-\mu)\right)= \\ \qquad\dfrac{\sigma^2}{M^2}\displaystyle\sum_{j=1}^{M}\sum_{l=1}^{M}\rho_{j-l}=\dfrac{\sigma^2}{M^2}\left[1+2\sum_{j=1}^{M}\left(1-\dfrac{j}{M}\right)\rho_j\right] \\ \text{② 样本方差的方差 } \sigma_2^2=D(\sigma_i^2)=\dfrac{2\sigma^2}{M^2}\left(1+2\displaystyle\sum_{j=1}^{M}\left(1-\dfrac{j}{M}\right)\rho_j^2\right) \\ \text{③ 样本自相关的方差}\sigma_3^2(\tau)=D(r_i(\tau))\approx\dfrac{1}{M-\tau}\left[1+\rho_\tau^2+2\displaystyle\sum_{j=1}^{M-\tau}\left(1-\dfrac{j}{M-\tau}\right)\times\right. \\ \qquad\left.(\rho_j^2+\rho_{j+\tau}\rho_{j-\tau})\right] \end{array}\right\} \tag{1-3}$$

采用统计检验方法，取显著水平 $\alpha=0.05$ 和 2σ 原则，此时置信度为 $1-\alpha=0.95$，当

$$\left.\begin{aligned}&|\bar{x}_i-\bar{x}_j|>1.96\sqrt{2\sigma_1^2}\\&|\sigma_i^2-\sigma_j^2|>1.96\sqrt{2\sigma_2^2}\\&|r_i(\tau)-r_j(\tau)|>1.96\sqrt{2\sigma_3^2(\tau)}\end{aligned}\right\}\tag{1-4}$$
$$(i\neq j,i,j=1,2,\cdots,k;\tau=1,2,\cdots,m)$$

成立时，可拒绝 $\{x_t\}$ 为平稳序列的假设，即该序列不具有平稳性。但一般并不知道 $\{x_t\}$ 的理论方差与自相关函数，因此无法直接得出 σ_1^2,σ_2^2 和 $\sigma_3^2(\tau)$，仅能以它们的样本估计值代之。因此，这个方法还不够理想，一般还要结合物理背景判断在过程运行中周围条件及有关参数是否维持不变来确定是否平稳的。

2. 平稳性的非参数检验法

平稳性的非参数检验法又称为游程检验法（或轮次检验法）。该方法只涉及一组实测数据，而不需要假设数据的分布规律，因此本方法具有很好的实用性。

在保持随机序列原有顺序的情况下，游程定义为具有相同符号的序列，这种符号可把观测值分成两个互相排斥的类。例如观测序列的值是 x_i $(i=1,2,\cdots,N)$，其均值为 $\bar{x}$，用符号“+”表示 $x_i\geqslant\bar{x}$，而“−”表示 $x_i<\bar{x}$。按符号“+”和“−”的出现顺序将原序列写成如下形式，例如

$$\underbrace{+\ +\ +}_{1}\ \underbrace{-}_{2}\ \underbrace{+}_{3}\ \underbrace{-}_{4}\ \underbrace{+}_{5}\ \underbrace{-\ -\ -\ -}_{6}\ \underbrace{+}_{7}$$

“+”号和“−”号共 14 个，分 7 个游程。每个游程的长短在这里并不重要。游程太多或太少都被认为是存在非平稳性趋势。游程检验所判断的原假设为：“样本数据出现的顺序没有明显的趋势，就是平稳的”。我们采用的样本统计量有

$$N_1=\text{一种符号出现的总数}$$
$$N_2=\text{另一种符号出现的总数}$$
$$\gamma=\text{游程的总数}$$

其中 γ 作为检验统计量。对于显著水平 $\alpha=0.05$ 的双边检验，由表 1-1 给出概率分布左右两侧为 $\alpha/2=0.025$ 时的上限 γ_U 和下限 γ_L。如果 γ 在界限以内，则接受原假设；否则拒绝原假设。

例如有 $N=22$ 的观测序列，其值超过均值者记为“+”；反之记为“−”，得符号序列如下

$$+\ +\ -\ -\ -\ +\ -\ -\ +\ +\ +\ +\ +\ -\ -\ +\ -\ -\ +\ +\ -\ +$$

我们按 $\alpha=0.05$ 检验顺序的随机性。因 $N_1=12$ (+)，$N_2=10$ (−)，故可取 $r=11$ (双边)。查表 1-1 得原假设的接受域为 $7\leqslant\gamma\leqslant17$，故原序列没有明显的潜在趋势，可以认为是平稳的。

表 1-1 是游程检验分布表，表中 N_1 和 N_2 分别表示符号 + 和 − 的数目，γ_L 和 γ_U 给出显著水平 $\alpha=0.05$ 时游程总数的下限和上限。由于游程检验属于双边检验，故应将显著水平划分成两边 $\alpha/2=0.025$，只要实际的游程总数 γ (双边) 在表 1-1 中给出的 γ_L 和 γ_U 界限内，则平稳性的假设就可以接受。此时 $P(\gamma\leqslant\gamma_L)+P(\gamma\geqslant\gamma_U)=0.05$

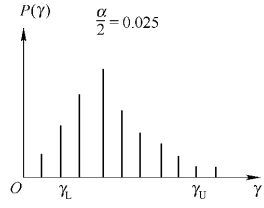

表 1-1 游程检验用 γ 分布表

N_2 / N_1		2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	γ_L											2	2	2	2
	γ_U														
3	γ_L					2	2	2	2	2	2	2	2	2	2
	γ_U														
4	γ_L				2	2	2	3	3	3	3	3	3	3	3
	γ_U				9	9									
5	γ_L			2	2	3	3	3	3	3	4	4	4	4	4
	γ_U			9	10	10	11	11							
6	γ_L		2	2	3	3	3	3	4	4	4	4	5	5	5
	γ_U			9	10	11	12	12	13	13	13	13			
7	γ_L		2	2	3	3	3	4	4	5	5	5	5	5	6
	γ_U				11	12	13	13	14	14	14	14	15	15	15
8	γ_L		2	3	3	3	4	4	5	5	5	6	6	6	6
	γ_U				11	12	13	14	14	15	15	16	16	16	16
9	γ_L		2	3	3	4	4	5	5	5	6	6	6	7	7
	γ_U					13	14	14	15	16	16	16	17	17	17
10	γ_L		2	3	3	4	5	5	5	6	6	7	7	7	7
	γ_U					13	14	15	16	16	17	17	18	18	18
11	γ_L		2	3	4	4	5	5	6	6	7	7	7	8	8
	γ_U					13	14	15	15	17	17	18	19	19	19
12	γ_L	2	2	3	4	4	5	6	6	7	7	7	8	8	8
	γ_U					13	14	16	16	17	18	19	19	20	20
13	γ_L	2	2	3	4	5	5	6	6	7	7	8	8	9	9
	γ_U						15	16	17	18	19	19	20	20	21
14	γ_L	2	2	3	4	5	5	6	7	7	8	8	9	9	9
	γ_U						15	16	17	18	19	20	20	21	22
15	γ_L	2	3	3	4	5	6	6	7	7	8	8	9	9	10
	γ_U						15	16	18	18	19	20	21	22	22

当 N_1 或 N_2 超过 15 时可以用正态分布来近似,即可利用正态分布表 (表 1-2) 来确定检验的接受域和否定域。此时用的统计量为

$$Z=\frac{\text{游程数}-\text{游程的期望数}}{\text{游程标准差}}=\frac{\gamma-\mu_\gamma}{\sigma_\gamma} \tag{1-5}$$

式中

$$\mu_\gamma = \frac{2N_1N_2}{N} + 1$$

$$\sigma_\gamma = \left[\frac{2N_1N_2(2N_1N_2 - N)}{N^2(N-1)}\right]^{1/2}$$

$$N = N_1 + N_2$$

对于 $\alpha = 0.05$ 的显著水平，如果 $|Z| \leqslant 1.96$（按 2σ 原则），则可接受原假设，否则就拒绝。

1.2　正态性检验

正态性是动态随机数据最重要的统计特性，目前常用的时间序列模型就是建立在具有正态概率分布特性的白噪声基础上的。

正态分布的概率密度函数（PDF）可记为

$$p(x) = (2\pi\sigma^2)^{-\frac{1}{2}}\exp[-(x-\mu)^2/(2\sigma^2)] \tag{1-6}$$

式（1-6）中 μ 和 σ^2 分别为样本总体的均值和方差。概率分布是概率密度函数的积分

$$P(x < X) = (2\pi\sigma^2)^{-\frac{1}{2}}\int_{-\infty}^{X}\exp(-(x-\mu)^2/2\sigma^2)\,\mathrm{d}x =$$

$$(2\pi)^{-\frac{1}{2}}\int_{-\infty}^{(X-\mu)/\sigma}\exp\left(-\frac{1}{2}x^2\right)\mathrm{d}x =$$

$$\Phi((X-\mu)/\sigma) \tag{1-7}$$

其中 Φ 称为“概率积分”，如表 1-2 所示。随机变量处于 α 和 β 之间的概率为

$$P(\alpha \leqslant x \leqslant \beta) = \Phi((\beta-\mu)/\sigma) - \Phi((\alpha-\mu)/\sigma) \tag{1-8}$$

“卡埃平方（χ^2）拟合优度检验”是一种检验动态数据正态性的有效方法，它是利用 χ^2 统计量作为观察到的 PDF 和理论密度函数之间偏差的量度，两者是否相同可通过分析 χ^2 的样本分布来检验。如果数据是正态的，则应落入第 j 组区间中的数据个数（称为组区间中的期望频数）为

$$F_0 = N\Phi\left(\frac{a-\mu}{\sigma}\right)$$

$$F_j = N\left[\Phi\left(\frac{d_j-\mu}{\sigma}\right) - \Phi\left(\frac{d_{j-1}-\mu}{\sigma}\right)\right] \tag{1-9}$$

$$F_{k+1} = N\left[1 - \Phi\left(\frac{b-\mu}{\sigma}\right)\right],\ k \text{ 是数据分组数}$$

式中 a 和 b 是两个端点值，$\Phi(\cdot)$ 是正态分布的累计积分，如表 1-2 内阴影面积。

式 (1-9) 中的 F_j 和观察到的频数 N_j 之间的偏差为 $(N_j - F_j)$，显然

$$\sum_{j=0}^{k+1} N_j = \sum_{j=0}^{k+1} F_j = N \tag{1-10}$$

故总的偏差必为 0。根据 Pearson 定理，样本的 χ^2 统计量如下：

$$\chi^2 = \sum_{j=0}^{k+1} (N_j - F_j)^2 / F_j \tag{1-11}$$

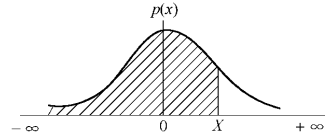

表 1-2 标准正态分布函数表 (假定 $\mu = 0, \sigma^2 = 1$)

X	0	1	2	3	4	5	6	7	8	9
-3.0	0.0013	0.0010	0.0007	0.0005	0.0003	0.0002	0.0002	0.0001	0.0001	0.0000
-2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0126	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0238	0.0233
-1.8	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0300	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0570	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985

续表

X	0	1	2	3	4	5	6	7	8	9
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8620
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8831
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9278	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9430	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9648	0.9653	0.9664	0.9671	0.9678	0.9686	0.9693	0.9700	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9762	0.9767

续表

X	0	1	2	3	4	5	6	7	8	9
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9863	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

假定这个样本 χ^2 统计量近似为卡埃平方分布，则可以将它和理论的卡埃平方分布（记为 $\chi^2_{n,\alpha}$）作比较，这时自由度 n 等于 $(k+2)$（如果把范围两端的组也算上的话）减去一些线性约束的数目，其中一个约束是当前 $(k+1)$ 个组区间的频数已知时，由于总频数为 N，最后一个组区间的频数也就知道了。另外两个约束是由于同理论正态概率密度函数拟合观察数据的频数直方图而引起的，这就是用样本均值和样本方差，而不是用真正的均值和方差来计算 $\{F_j\}$。因此，如果利用全部 $\{N_j\}$，则自由度

$$n=(k+2)-3=k-1 \tag{1-12}$$

实际 n 值可能比这还要小些，因为 $F<2$ 的一些组可能和其他组合并。

在 χ^2 的自由度正确确定之后，可以作如下假设检验：若假设 x 变量是正态分布的，在把观察数据分组列入 $k+2$ 个组区间后，利用样本均值和方差，计算 F_j 并求出 χ^2。样本 PDF 对正态分布的任何偏离都会使 χ^2 增大。如果

$$\chi^2 \leqslant \chi^2_{n,\alpha} \tag{1-13}$$

则接受（在 α 显著水平上）数据为正态分布的假设；反之，如果 χ^2 大于 $\chi^2_{n,\alpha}$，则在 α 显著水平上拒绝上述假设。常用的显著水平为 5%，10%和 20%，即相应于 95%，90%和 80%的置信水平，如表 1-3 所示。式（1-13）常称为检验判别公式。

表 1-3 是卡埃平方（χ^2）分布表，表中 k 是数据分组数，α 是给定的显著水平，表内给出在自由度为 n 和显著水平为 α 时的 $\chi^2_{n,\alpha}$ 值。只要实际计算出来的 χ^2 值小于 $\chi^2_{n,\alpha}$ 值，则正态性的假设是可以接受的。因为 χ^2 只有正数域，故这里是单边检验方法。

本表给出 $\chi_\alpha^2(k)$ 值使 $P[\chi^2 > \chi_\alpha^2(k)] = \alpha$ (阴影面积)。

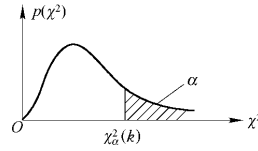

表 1-3　χ^2 分布表

n \ α	0.25	0.10	0.05	0.025	0.01	0.005
1	1.323	2.706	3.841	5.024	6.635	7.879
2	2.773	4.605	5.991	7.378	9.210	10.597
3	4.108	6.251	7.815	9.348	11.345	12.838
4	5.385	7.779	9.488	11.143	13.277	14.860
5	6.626	9.236	11.071	12.833	15.086	16.750
6	7.841	10.645	12.592	14.449	16.812	18.548
7	9.037	12.017	14.067	16.013	18.475	20.278
8	10.219	13.362	15.507	17.535	20.090	21.955
9	11.389	14.684	16.919	19.023	21.666	23.589
10	12.549	15.987	18.307	20.483	23.209	25.188
11	13.701	17.275	19.675	21.920	24.725	26.757
12	14.845	18.549	21.026	23.337	26.217	28.299
13	15.984	19.812	22.362	24.736	27.688	29.819
14	17.117	21.004	23.685	26.119	29.141	31.319
15	18.245	22.307	24.996	27.488	30.578	32.801
16	19.369	23.542	26.296	28.845	32.000	34.267
17	20.489	24.769	27.587	30.191	33.409	35.718
18	21.605	25.989	28.869	31.526	34.805	37.156
19	22.718	27.204	30.144	32.852	36.191	38.582
20	23.828	28.412	31.410	34.170	37.566	39.997
21	24.935	29.615	32.671	35.479	38.932	41.401
22	26.039	30.813	33.924	36.781	40.289	42.796
23	27.141	32.007	35.172	38.076	41.638	44.181
24	28.241	33.196	36.415	39.364	42.980	45.559
25	29.339	34.382	37.652	40.646	44.314	46.928
26	30.435	35.563	38.885	41.923	45.642	48.290
27	31.528	36.741	40.113	43.194	46.963	49.645
28	32.620	37.916	41.337	44.461	48.278	50.993
29	33.711	39.087	42.557	45.722	49.588	52.336
30	34.800	40.256	43.773	46.979	50.892	53.672

例如有 2 000 个样本数据，将它们划分成 $k=16$ 个分组，相应的自由度为 $n=k-1=15$。在给定显著水平为 $\alpha=0.05$ 的情况下，由表 1-3 可以查得 $\chi^2_{n,\alpha}=24.996$，因此如果实际计算所得的 χ^2 数值小于该 $\chi^2_{n,\alpha}$ 值，则所给出的一组样本数据是符合正态性假设的。

对于用 χ^2 拟合优度检验检查正态性的组区间数目，有人还给出了总体样本量和分组数目应满足的最优关系式

$$k\ (\text{分组区间数目})=1.87\ (N-1)^{2/5} \tag{1-14}$$

式 (1-14) 中假定数据是无关的。

另外，在应用 χ^2 检验时的一个准则是每个区间中的期望频数至少应为 2。由于范围两端的期望频数最少，因此上述要求可以用来确定 a 和 b。参数 a 应满足式 (1-7) 中的 $P\ (x<X)=2/N$，即

$$2\leqslant N\left[(2\pi)^{-\frac{1}{2}}\int_{-\infty}^{(a-\mu)/\sigma}\exp\left(-\frac{1}{2}x^2\right)\mathrm{d}x\right] \tag{1-15}$$

由式 (1-15) 可以求得 a，又由平均值 $\mu=(b-a)/2$ 得参数 b 为

$$b=2\mu+a \tag{1-16}$$

这时分组区间数目为

$$k=r\ (\text{最小区间数目})-2 \tag{1-17}$$

以上三种方法都可以用来确定分组区间数目。

表 1-4 样本总量为 N，$\alpha=0.05$ 时最小的分组区间数目 (r)

N	r	N	r
200	16	20 000	94
400	20	40 000	129
600	24	70 000	162
800	27	100 000	187
1 000	30	200 000	247
1 500	35	400 000	326
2 000	39	700 000	407
4 000	57	1 000 000	470
7 000	65	1 140 000	500
10 000	74		

因此，如果给出数据总量为 $N=2\ 000$，则由表 1-4 得最小的分组区间应为 $r=39$，相应的 $k=37$。最后在表 1-3 中可查出当 $n=36$ 时的 $\chi^2_{n,\alpha}=52.19$，此时给定的显著水平仍为 $\alpha=0.05$。