时间序列分析

翟祥

北京林业大学

zhaixbh@126.com

教材介绍

- 教材
  - 《基于python的时间序列分析》，白晓东 编著，清华大学出版社.
- 计算平台
  - Python.
- 数据
  - 同学可根据自己的情况进行调整.

参考书

- 何书元. 《应用时间序列分析》. 北京大学出版社, 2003年.
- 吴喜之, 刘苗. 《应用时间序列分析》. 机械工业出版社, 2014年.
- 王燕. 《应用时间序列》. 中国人民大学出版社, 2023年.
- 王黎明, 王连, 杨楠. 《应用时间序列分析》. 复旦大学出版社, 2009年.
- Tsay R S. 《An Introduction to Analysis of Financial Data with R》. Wiley, 2013.

第一章 引言及基础知识

zhaixbh@126.com

本章结构

1. 引言
2. 基本概念
3. 时间序列建模的基本步骤
4. 数据预处理

时间序列的定义

- 时间序列的产生
  - 时间序列分析在人类早期的生产实践和科学研究中发挥了重要作用.
  - 比如： 7000 年前, 古埃及人为了发展农业, 把尼罗河涨落的情况逐天记录下来, 并进行了长期的观察. 他们发现, 在天狼星第一次和太阳同时升起后的两百天左右尼罗河开始泛滥, 洪水大约持续七八十天, 此后土地肥沃、适于农业种植. 由于掌握了尼罗河泛滥的规律, 古埃及的农业迅速发展, 从而创造了古埃及灿烂的史前文明.
  - 再如：德国天文学家、药剂师 S. H. Schwabe 从 1826 年至 1843年, 在每一个晴天, 认真审视太阳表面, 并且记录下每一个黑点, 对这些记录仔细研究后, 最终发现了太阳黑子活动有 11 年左右的周期性规律. 这一发现被视为天文学上最重要的发现之一.

时间序列的定义

- 时间序列的产生
  - 许多经济现象的发展都具有随时间演变的特征，如：宏观经济运行中的国内生产总值、消费支出、货币供应量等；又如：微观经济运行中的企业产品价格、销售量、销售额、利润等量;再如：金融市场中的股价指数、股票价格、成交量等变量的变化．将这些变量依时间先后记录下来并加以研究，揭示其中隐含的经济规律，预测未来经济行为，已经成为经济研究的重要手段．
  - 像上面这样按照时间的顺序把随机事件变化发展的过程记录下来就构成了一个时间序列,对时间序列进行观察、研究，找寻它变化发展的规律，预测它将来的走势就是时间序列分析．

时间序列的定义

- 时间序列的定义
  - 在统计研究中，一般将按时间顺序排列的一组随机变量

$$X_1, X_2, \ldots, X_t, \ldots \tag{1.1}$$

称为一个时间序列 (time series)，简记为 $\{X_t, t \in T\}$ 或 $\{X_t\}$.

  - 用 $x_1, x_2, \ldots, x_n$ (1.2)

或 $\{x_t, t = 1, 2 \ldots, n\}$

表示该随机序列的 n 个有序观测值，称为序列长度为 n 的观察值序列，有时也称观察值序列 (1.2)为时间序列 (1.1)的一个实现. 在上下文不引起歧义的情况下，有时一个时间序列也记为 $\{x_t\}$.

时间序列的定义

- 例 1.1
  - 把我国 1953—2016 年国内生产总值 (GDP) 按照时间顺序记录下来, 就构成了一个序列长度为 64 的国内生产总值观察值序列. 将数据按时间顺序逐一罗列或绘表罗列, 一般不易观察. 为此通常绘制时序图来观察趋势, 所谓**时序图**是指横轴表示时间, 纵轴表示时间序列的观察值而绘制的图.
  - 借助 R 软件强大的绘图功能可以绘制出许多漂亮的统计图. 如:用下列 R 语句,可绘出上述序列的时序图(见图1.1).

时间序列的定义

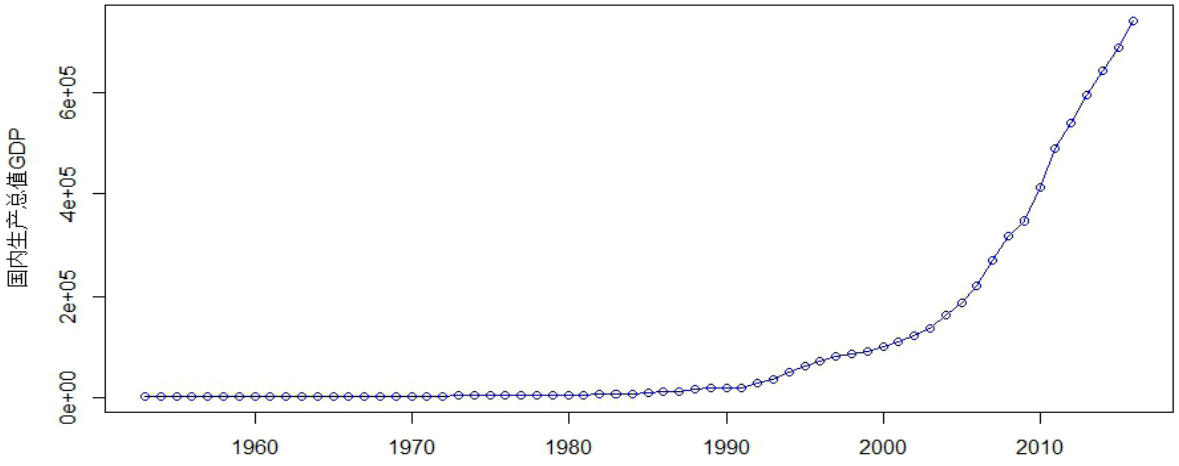


图1.1 中国1953-2016年国内生产总值年度时序图

- 从图 1.1 中可以看出, 我国 GDP 从 1992 年开始大幅度增长, 1998 年左右增长速度出现瓶颈, 而 2004 年之后, 除了 2009 年有小幅增速外, 几乎呈现直线型高速增长趋势. 为了更好地预测这种趋势, 我们关心的是相邻年度 GDP 的关联情况. 为此, 我们可以绘制我国当年 GDP 与上一年 GDP 的散点图. 接上面程序, 我们用下列 R 语句生成图 1.2.

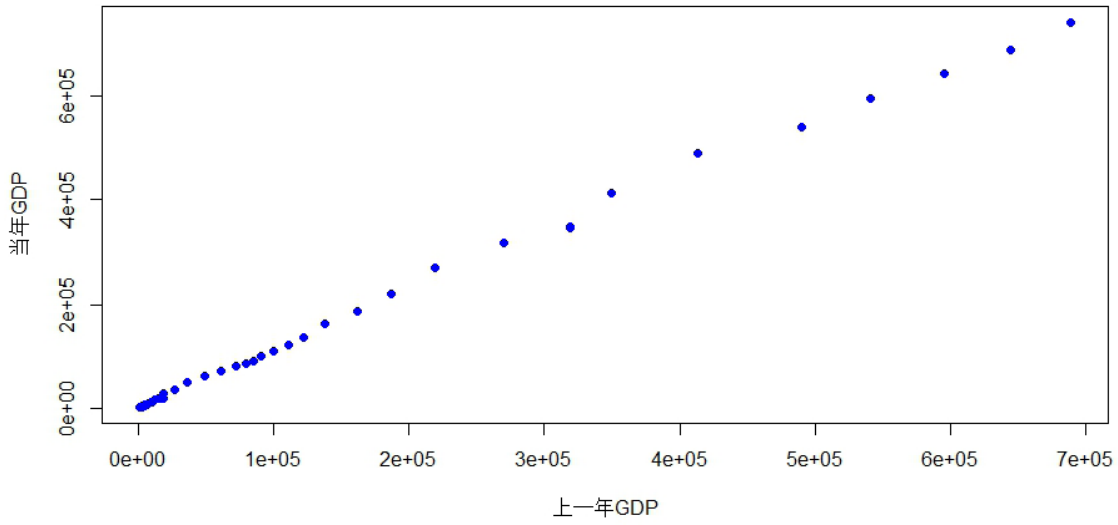
时间序列的定义

- > y <- GDP[-1]

  > x <- GDP[-64]

  > plot(x,y,xlab="上一年GDP",ylab="当年GDP", pch=16,col=1)

图1.2 中国当年GDP与上一年GDP的散点图



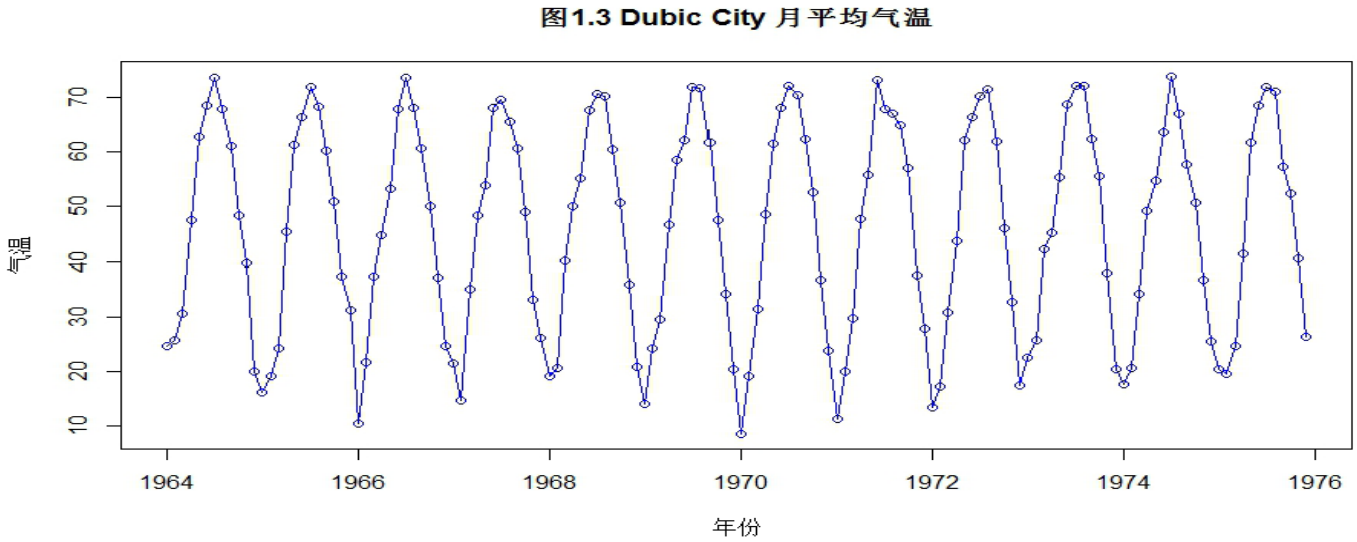
从图 1.2 看出相邻年度GDP 的关联呈线性.

时间序列的定义

- 例 1.2
  - 将美国爱荷华州杜比克 (Dubic) 市 144 个月的平均气温 (单位: °F) 按时间顺序记录下来, 就得到长度为 144 的观察值序列.
  - 用下列 R 语句生成图 1.3.

```
> t <- scan("E:/DATA/CHAP1/data1.2.txt")

> t <- ts(t, start=c(1964,1),frequency=12)

> plot(t,type="o",xlab="年份",ylab="气温",col=4)
```

时间序列的定义



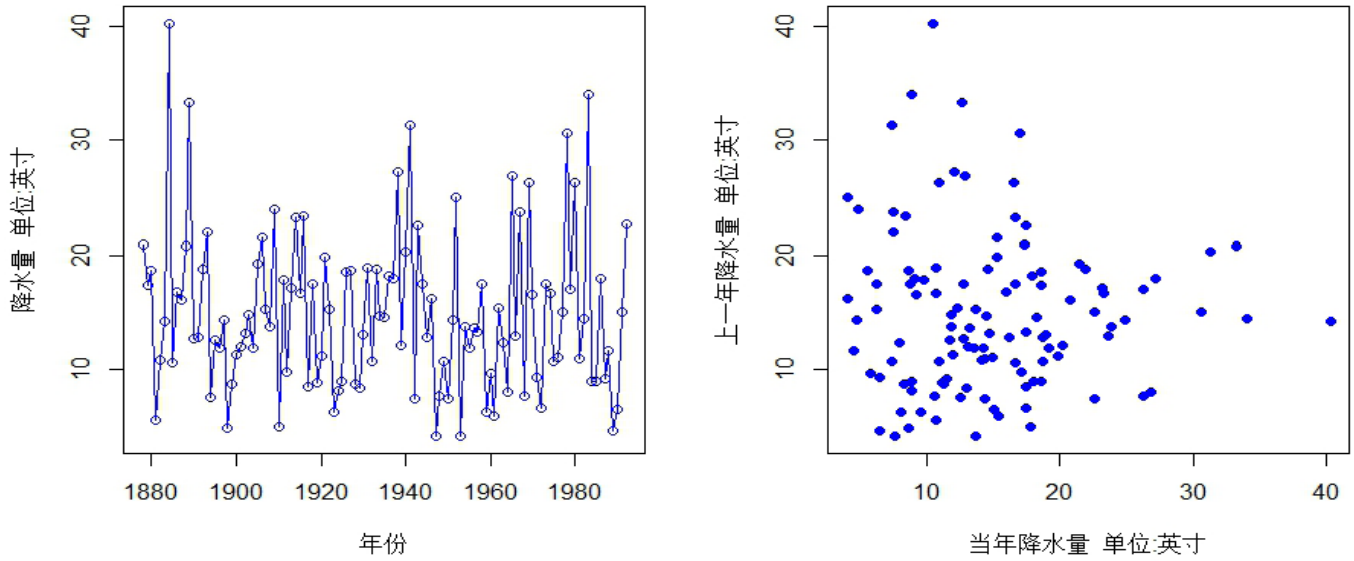
- 从图 1.3 可以看出, 这些观察值显示了很强的季节性趋势. 后面的章节中将会通过构造季节指数的方式, 对这类数据建模.

时间序列的定义

- 例 1.3
  - 将美国加利福尼亚州洛杉矶地区 115 年来的年降水量记录下来, 构成一个序列长度为 115 的观察值序列.
  - 用下列 R 语句生成时序图 1.4.

```
> t <- scan("E:/DATA/CHAP1/data1.2.txt")

> t <- ts(t, start=c(1964,1),frequency=12)

> plot(t,type="o",xlab="年份",ylab="气温",col=4)
```

时间序列的定义



- 从图 1.4 中可以看出，该地区降水量没有明显的趋势性. 接下来观察相邻年份的相关关系. 从图 1.5 看出，相邻各点没有明显的相关关系. 像这种既无明显的趋势，也无明显的相关关系的数据，从统计建模和预测角度看没有研究的意义.

时间序列的定义

- 从上述例子可以看出，时间序列中观察值的取值随着时间的变化而不同，反映了相关指标在不同时间进行观察所得到的结果. 这些观察值可以是一个时期内的数据，也可能是一个时间点上的数据，通常存在前后时间上的相依性. 从整体上看，时间序列往往呈现某种趋势性或出现季节性变化的现象，这种相依性就是系统的动态规律性，也是进行时间序列分析的基础.

- 总之，我们进行时间序列研究的目的是想揭示随机时序 $\{X_t\}$ 的性质，而要实现这个目标就要分析它的观察值序列 $\{x_t\}$ 的性质，由观察值序列的性质来建立恰当的模型，从而推断随机时序 $\{X_t\}$ 的性质.

时间序列的分类

- 时间序列的分类
  - 1. 一元时间序列与多元时间序列.

    每个时间点只观察一个变量的时间序列称为一元时间序列.

    如果每个时间点同时观察多个变量的时间序列则称为多元时间序列.

    多元时间序列不仅描述了各个变量的变化情况，而且还蕴含了各变量间的相互依存关系. 例如，考察某国或某地区经济运行情况，就需要同时观察某国国内或某地区内生产总值、消费支出、投资额、货币供应量等一系列指标，既要分析每个指标的动态变化情况还要分析各个指标之间的动态影响关系.

时间序列的分类

- 时间序列的分类
  - 2. 连续时间序列与离散时间序列

时间序列是按照时间顺序记录的一系列观测值, 这种观测值可能是按连续的时间记录的,也可能是按离散的时间点来记录的. 相应地, 通常把这两类序列分别称为连续时间序列 和离散时间序列.

例 1.1~ 例 1.3 都是离散时间序列, 而利用脑电图记录仪记录大脑活动情况则可视为连续时间序列. 对于连续时间序列, 可通过等间隔抽取样本使之转化为离散时间序列加以研究. 一般地, 如果时间间隔足够小, 那么我们可以认为这种过程几乎不会损失原序列的信息.

时间序列的分类

- 时间序列的分类
  - 3. 平稳时间序列与非平稳时间序列

按时间序列的统计性质, 可将时间序列分为平稳时间序列和非平稳时间序列. 关于时间序列的平稳性与非平稳性在之后的学习中详细讨论.

  - 此外, 还可以按照模型的表示形式分为线性时间序列和非线性时间序列, 等等.

时间序列分析的方法回顾

- 时间序列分析的方法回顾
  - 1. 描述性时间序列分析

早期的时间序列分析是通过直观的数据比较或绘图观测，寻找序列中蕴含的发展规律，这种分析方法就称为描述性时间序列分析. 该方法不采用复杂的模型和分析方法，仅仅是按照时间顺序收集数据，描述和呈现序列的波动，常常能使人们发现意想不到的规律，具有操作简单、直观有效的特点. 人们在进行时间序列分析时，往往首先进行描述性分析.

  - 2. 统计时间序列分析

随着研究领域的不断拓展，单纯的描述性时间序列分析方法越来越显示出局限性. 在许多问题中，随机变量的发展会表现出很强的随机性，想通过对序列的简单观察和描述总结出随机变量发展变化的规律，

时间序列分析的方法回顾

- 时间序列分析的方法回顾
  - 并准确预测出它们将来的走势通常非常困难. 为了准确地估计随机序列发展变化的规律, 从 20 世纪 20 年代开始, 学术界利用数理统计学原理分析时间序列内在的相关关系, 由此开辟了一门应用统计学科 —— 时间序列分析.
  - 从时间序列分析方法的发展历史来看, 其大致可分为两类: 频域 (frequency domain) 分析方法和时域 (time domain) 分析方法.
  - 频域分析方法也称为“频谱分析”或“谱分析”方法. 早期的频域分析方法假设任何一种无趋势的时间序列都可以分解成若干不同频率的周期波动, 借助 Fourier 分析从频率的角度揭示时间序列的规律. 20 世纪 60 年代, Burg 在分析地震信号时提出最大熵谱估计理论. 该理论克服了传统谱分析所固有的分别率不高和频率泄露

时间序列分析的方法回顾

- 时间序列分析的方法回顾
  - 等缺点，使得谱分析进入一个新的阶段，称为现代谱分析. 谱分析方法是一种非常有用的纵向数据分析方法，目前已广泛应用于物理学、天文学、海洋学、气候学、电力和通信工程等领域. 谱分析方法的最大的缺点是，需要较强的数学基础才能熟练使用，而且分析结果较为抽象，难以解释.

  - 时域分析方法的基本思想是事件的发展通常都具有一定的惯性，这种惯性用统计学语言来描述就是序列值之间存在一定的相关关系，而且这种相关关系具有某种统计规律性. 我们分析的重点就是从序列自相关的角度揭示时间序列的某种统计规律. 相对于谱分析方法，它具有理论基础扎实、操作步骤规范、分析结果易于解释等优点. 目前已经广泛应用于自然科学和社会科学的各个领域，成为时间序列分析的主流方法之一.

本章结构

1. 引言
2. 基本概念
3. 时间序列建模的基本步骤
4. 数据预处理

时间序列与随机过程

- 时间序列与随机过程
  - 我们知道，随机变量是分析随机现象的重要工具，对于简单的随机现象，用一个随机变量就可以了，如某时段内共享单车的使用量，某时刻候车的人数，等等. 而对于复杂的随机现象，用一个随机变量描述就不够了，需要用若干个随机变量来描述. 一般地，将一族随机变量放在一起就构成一个随机过程. 具体地，有下面的定义：我们将概率空间（Ω,F,P）上的一族随机变量 $\{X_t, t \in T\}$ 称为一个随机过程 (stochastic process)，其中 t 是参数，它属于某个集合 T，通常称 T 为参数集 (parameter set).

  - 参数集 T 可以是离散集合，也可以是连续集. 若 T 为一连续集，则$\{X_t\}$为一连续型随机过程. 若 T 为离散集，则称$\{X_t\}$为一离散型随机过程. 当参数集为某时间集合时，则相应的随机过程就为时间序列. 可见，时间序列仅仅是随机过程的特殊情况，因此随机过程的许多概念和性质同样适用于时间序列.

概率分布族及其特征

- 时间序列的有限维分布族
  - 设 $\{X_t, t \in T\}$ 为一个时间序列, 对于任意一个 $t \in T$, X_t 是一个随机变量, 它的分布函数 $F_{X_t}(x)$ 可以通过 $F_{X_t}(x) = \mathrm{P}(X_t \le x)$ 得到, 这一分布函数称为**时间序列的一维分布**.
  - 对于 $t_1, t_2 \in T$, 有两个随机变量 X_{t_1} X_{t_2} 与之对应, X_{t_1} X_{t_2} 的联合分布函数

$$F_{X_{t_1}, X_{t_2}}(x_1, x_2) = \mathrm{P}(X_{t_1} \le x_1, X_{t_2} \le x_2)$$

称为**时间序列的二维联合分布**.

概率分布族及其特征

- 时间序列的有限维分布族
  - 一般地，任取正整数 n ，$t_1,t_2,...,t_n \in T$ 则 n 维向量的联合分布

$$F_{X_{t_1},X_{t_2},...,X_{t_n}}(x_1,x_2,...,x_n)=\mathrm{P}(X_{t_1}\le x_1,X_{t_2}\le x_2,...,X_{t_n}\le x_n)$$

这些有限维分布函数的全体

$$\left\{F_{X_{t_1},X_{t_2},...,X_{t_n}}(x_1,x_2,...,x_n),\forall n\in Z^+,\forall t_1,t_2,...,t_n\in T\right\}$$

被称为**时间序列** $\{X_t,t\in T\}$ **的有限维分布族**.

理论上，时间序列的所有统计性质都可通过有限维分布族推导出来，但是在实际应用中，要想得到一个时间序列的有限维分布族几乎是不可能的，而且有限维分布族在使用中通常涉及非常复杂的数学运算，因而一般情况下，我们很少直接使用有限维分布族进行时间序列分析.

概率分布族及其特征

- 数字特征
  - 1. 均值函数

对时间序列$\{X_t, t\in T\}$ 来说, 任意时刻的序列值 X_t 都是一个随机变量. 假设它的分布函数为$F_{X_t}(x)$, 那么当 $\mu_t = EX_t = \int_{-\infty}^{+\infty} x dF_{X_t}(x) < \infty$ 对于所有的 $t\in T$ 成立时,我们称 μ_t 为时间序列$\{X_t, t\in T\}$ 的**均值函数** (mean function). 它反映的是时间序列$\{X_t, t\in T\}$ 在各个时刻的平均取值水平, 通常也可记为 EX_t.

  - 2. 方差函数

对于所有的 $t\in T$, $\int_{-\infty}^{+\infty} x^2 dF_{X_t}(x) < \infty$ 成立时, 我们称

$$\sigma_t^2 = Var(X_t) = \int_{-\infty}^{+\infty} (x-\mu_t)^2 dF_{X_t}(x)$$

为时间序列$\{X_t, t\in T\}$ 的 的方差函数 (variance function). 它反映了序列值围绕其均值做随机波动时的平均波动程度.

概率分布族及其特征

- 数字特征
  - 3. 自协方差函数

类似于随机变量间的协方差, 在时间序列分析中, 我们可以定义自协方差函数 (autocovariance function) 的概念. 对于时间序列 $\{X_t, t \in T\}$ 来说, 任取 $s,t \in T$,称

$$\gamma(s,t) = E[(X_t - \mu_t)(X_s - \mu_s)]$$

为序列 $\{X_t, t \in T\}$ 的自协方差函数.

  - 4. 自相关函数

同样地, 类似于随机变量间的相关系数, 我们可以定义时间序列的自相关函数 (autocorrelation function, ACF). 我们称

概率分布族及其特征

- 数字特征

$$\rho(s,t)=Cor(X_t,X_s)=\frac{\gamma(s,t)}{\sqrt{Var(X_t)}\sqrt{Var(X_s)}}$$

为序列$\{X_t,t\in T\}$的自相关函数. 时间序列的自协方差函数和自相关函数反映了不同时刻的两随机变量的相关程度

- 5. 偏自相关函数

自相关函数虽然反映了时间序列$\{X_t,t\in T\}$在两个不同时刻X_t和X_s的相依程度, 但是这种相关包含了X_s通过X_s和X_t之间的其他变量传递到对$X_t(s<t)$的影响, 也就是说自相关函数实际上掺杂了其他变量的影响. 为了剔除中间变量的影响, 可引入偏自相关函数 (partial autocorrelation function, PACF) 的概念. 偏自相关函数的定义为

概率分布族及其特征

- 数字特征

$$\beta(s,t)=Cor(X_t,X_s\,|\,X_{s+1},\ldots,X_{t-1})=\frac{Cov(X_t,X_s\,|\,X_{s+1},\ldots,X_{t-1})}{\sqrt{Var(X_t)}\sqrt{Var(X_s)}},0<s<t.$$

  - 一般来讲，一个时间序列的上述数字特征与时间有关，因而可看成关于时间的函数. 不同类型时间序列的数字特征会随时间变化呈现不同的变化规律，如有些时间序列的均值函数或方差函数不随时间的变化而变化，有些时间序列的自相关函数或偏自相关函数会出现随时间推移而逐渐变小的规律，等等. 在之后的章节中，我们将详细讨论不同类型时间序列在数字特征中表现出的差异.

平稳时间序列的定义

- 1. 严平稳时间序列

严平稳是一种条件较为严格的平稳性定义，它要求序列的所有有限维分布不随时间的推移而发生变化，从而序列的全部统计性质也不会随着时间的推移而发生变化. 具体地，定义如下：

设 $\{X_t, t \in T\}$ 为一时间序列. 若对于任意正整数 n，$t_1, t_2, \ldots, t_n \in T$ 以及任意正数 h 都有

$$F_{X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_n+h}}(x_1, x_2, \ldots, x_n) = F_{X_{t_1}, X_{t_2}, \ldots, X_{t_n}}(x_1, x_2, \ldots, x_n)$$

则称时间序列 为严平稳时间序列 (strictly stationary time series).

- 严平稳时间序列的定义所要求的条件过分严格. 实际中，要想知道时间序列的有限维分布族是极其困难的事情，所幸时间序列的主要统计性质是由它的低阶矩决定的，因此可以把严平稳的条件放宽，仅要求其数字特征不随时间发生变化，这样就得到了宽平稳的概念.

平稳时间序列的定义

- 2. 宽平稳时间序列

一般地, 如果一个时间序列 $\{X_t, t \in T\}$ 满足如下三个条件:

(1)对于任意 $t \in T$，有 $EX_t = \mu$；(2)对于任意 $t \in T$，有 $EX_t^2 < \infty$；

(3) 对于任意的 $s,t,k \in T$, $k+t-s \in T$ 有

$$\gamma(s,t) = \gamma(k, k+t-s), 0 < s < t.$$

则称 $\{X_t, t \in T\}$ 为宽平稳时间序列 (weakly stationary time series). 宽平稳也称为弱平稳或二阶矩平稳.

宽平稳的条件显然比严平稳的条件宽泛得多, 更具有操作性, 它只要求二阶矩具有平稳性,二阶以上的矩没有做任何要求. 一般情况下, 宽平稳不一定是严平稳; 严平稳也不一定是宽平稳.

平稳时间序列的定义

- 2. 宽平稳时间序列

如: 服从柯西分布的严平稳序列就不是宽平稳序列, 因为它不存在一、二阶矩, 所以无法验证它二阶矩平稳. 不过, 存在二阶矩的严平稳序列一定是宽平稳的. 宽平稳一般推不出严平稳, 但当序列服从多元正态分布时, 由宽平稳可以推出严平稳.

例 1.4 如果一个时间序列$\{X_t, t \in T\}$，满足: 任取正整数n，$t_1, t_2, ..., t_n \in T$

相应的 n 维随机变量$(X_{t_1}, X_{t_2}, ..., X_{t_n})$服从 n 维正态分布,那么我们称其为正态时间序列.从正态随机序列的密度函数可以看出, 它的 n 维分布仅由均值向量和协方差阵决定, 因此对于正态随机序列而言, 宽平稳一定严平稳.

需要强调的是, 在实际应用中, 如果不做说明, 我们所说的平稳指的就是宽平稳.

平稳时间序列的一些性质

- 1.延迟 k 自协方差函数

设 $\{X_t, t \in T\}$ 为一平稳时间序列，则称

$$\gamma(k) = \gamma(t, t+k), \qquad \forall t, t+k \in T$$

为该时间序列的延迟 k 自协方差函数.

- 根据平稳时间序列的定义可知，平稳序列具有常数方差：

$$\mathrm{Var}(X_t) = \gamma(t,t) = \gamma(0), \qquad \forall t \in T$$

- 2.延迟 k 自相关函数

称 $$\rho(k) = \frac{\gamma(t, t+k)}{\sqrt{\mathrm{Var}(X_t)}\sqrt{\mathrm{Var}(X_{t+k})}} = \frac{\gamma(k)}{\gamma(0)}$$ 为延迟 k 自相关函数

平稳时间序列的一些性质

- 容易验证延迟 k 自相关函数具有如下三个性质：

① 规范性： $\rho(0)=1$ 且 $|\rho(k)|\leq 1, \forall k$

② 对称性： $\rho(k)=\rho(-k)$

③ 非负定性：

根据协方差阵的非负定性，可得对于任意正整数 m，相关阵为非负定矩阵.

虽然一个平稳时间序列唯一决定了它的自相关函数，但是一个自相关函数未必唯一对应一个平稳时间序列，因而延迟 k 自相关函数 ρ(k) 对应模型并不唯一. 这个性质给我们根据样本自相关函数来确定模型增加了难度. 在后面的章节将进一步说明这个问题

平稳性假设的意义

数理统计学是利用样本信息来推测总体信息，时间序列分析作为数理统计学的一个分支也不例外。根据统计学常识，要分析一个 n 维随机向量 $\mathbf{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$，需要如下数据(见表 1.1)：

表 1.1　数据表

样本 \ 随机变量	X_1	$\cdots$	X_n
1	x_{11}	$\cdots$	x_{n1}
2	x_{12}	$\cdots$	x_{n2}
$\vdots$	$\vdots$		$\vdots$
m	x_{1m}	$\cdots$	x_{nm}

我们希望维数 n 越小越好，而对于每个变量希望样本容量 m 越大越好，这是因为维数越小分析过程越简单，样本容量越大，分析结果越可靠.

平稳性假设的意义

但是对于时间序列而言，它在任意时刻 t 的序列值都是一个随机变量，而且由于时间的不可重复性，该变量在任意一个时刻只能获得唯一的样本观察值，其数据结构如下 (见表 1.2).

表 1.2　数据表

随机变量 / 样本	X_1	$\cdots$	X_t	$\cdots$
1	x_1	$\cdots$	x_t	$\cdots$

由于某时刻对应的随机变量的样本容量太小，用该数据直接分析此刻的随机变量基本不会得到可用的结果，因此必须借用一些辅助信息，才能得到些有用的结果. 序列平稳性假设是解决该问题的有效途径之一.

平稳性假设的意义

比如：如果一个时间序列是平稳的，那么其均值函数是常数函数，也即$\{\mu_t, t \in T\}$变成了常数序列$\{\mu\}$。这样，本来每个随机变量X_t的均值μ_t只能凭借唯一的样本观察值x_t来估计,即$\hat{\mu}_t = x_t$．现在由于

$$\mu_t \equiv \mu$$

于是每个样本观测值x_t都变成了μ的样本观察值

$$\mu = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

于是，不但提高了对均值函数的估计精度，而且大大降低了时序分析的难度。

平稳性假设的意义

再如：基于平稳性可计算出延迟 k 自协方差函数的估计值：

$$\hat{\gamma}(k)=\frac{1}{n-k}\sum_{t=1}^{n-k}(x_t-\bar{x})(x_{t+k}-\bar{x})$$

和总体方差的估计值：

$$\hat{\gamma}(0)=\frac{1}{n}\sum_{t=1}^{n}(x_t-\bar{x})^2$$

进而可得，延迟 k 自相关函数的估计值：

$$\hat{\rho}(k)=\frac{\hat{\gamma}(k)}{\hat{\gamma}(0)} \qquad \forall 0<k<n$$

当延迟阶数 k 远远小于样本容量时，有

$$\hat{\rho}(k)=\frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}\approx\frac{\sum_{t=1}^{n-k}(x_t-\bar{x})(x_{t+k}-\bar{x})}{\sum_{t=1}^{n}(x_t-\bar{x})^2} \qquad \forall 0<k<n$$

本章结构

1. 引言
2. 基本概念
3. 时间序列建模的基本步骤
4. 数据预处理

时间序列建模的基本步骤

- 从实际数据出发，对时间序列建模一般可遵循四个步骤，即模型识别、模型估计、模型检验和模型应用. 通常上述四个步骤需要经过多次反复，才能达到比较满意的效果.
  - 1.模型识别：

所谓模型识别就是根据时间序列的统计特征选择适当的拟合模型. 模型识别主要包含如下内容：

（1）依照所研究的问题科学地收集数据.

（2）根据时间序列的数据作出相关图，求出相关函数进行分析. 相关图能够显示出序列变化的趋势性和周期性等特征，这些特征不但隐含着序列的平稳性的一些特点，而且能够发现跳点和拐点. 而这些跳点和拐点也是模型识别的重要参考因素.

时间序列建模的基本步骤

（3）判别时间序列是平稳的还是非平稳的. 一般来讲, 判别时间序列的平稳性有两种方法,一种是图检验法; 另一种是构造统计量进行假设检验的方法. 图检验法是根据时序图和自相关图显示的特征做出平稳性判别的方法. 它的优点是操作简便、运用广泛; 它的缺点是判别结论带有很强的主观色彩, 因此最好能够用统计检验方法加以辅助判别. 目前最常用的平稳性统计检验方法是单位根检验.

（4）判别时间序列是否是纯随机序列. 当对一个时间序列进行了平稳性判别之后, 序列被分成了平稳序列和非平稳序列两类. 对于非平稳序列通常要通过进一步的检验、变换或处理, 才能够确定适当的拟合模型.

时间序列建模的基本步骤

对于平稳序列来讲，我们需要检验其是否是纯随机的，因为只有那些序列值之间具有密切相关关系的序列，才值得我们花时间去挖掘历史数据中的有效信息，用来预测序列未来的发展. 如果序列值彼此之间没有任何相关性，那就意味着该序列是一个没有记忆的序列，过去的行为对将来的发展没有丝毫影响，这种序列称为纯随机序列. 从统计分析的角度而言,纯随机序列没有任何分析的价值.

（5）综合考虑时间序列的统计特征辨识合适的模型类型，初步确定模型结构.

至于常见的时间序列模型有哪些，它们分别具有哪些统计特征，以及如何根据样本信息估计数字特征、识别拟合模型等，将在后续章节详细研究.

时间序列建模的基本步骤

- 2.模型估计：

依照样本信息进行模型识别之后，我们得到了所分析的时间序列大概服从什么样的模型类型和模型结构，模型的最终形式还需要估计模型的参数之后才能够确定. 模型的参数决定了不同时刻随机变量之间的相依关系，也即反映了随机变量随时间变化的记忆性大小和记忆期的长短. 当参数确定了，变量的动态关系也就确定了.

在数理统计中，估计时间序列模型参数的常用方法有：矩估计、极大似然估计和最小二乘估计. 矩估计 是用样本矩代替相应的总体矩，并通过求解相应的方程而得到参数估计的方法；极大似然估计是使得样本出现概率最大，也就是使得似然函数达到最大而得到参数估计的方法；最小二乘估计是使得模型拟合的残差平方和达到最小，从而求得参数估计的方法. 这三种方法都有各自的优点和不足.

时间序列建模的基本步骤

- 3.模型检验：

在模型识别时，为了简化问题我们会提出一些假设，这些假设往往因人而异，带有主观因素,因此必须对模型本身进行检验. 同时，由于参数估计方法本身也有许多缺点，而且有些参数贡献不大，甚至可以忽略，所以对所估出的参数也必须进行检验.

由于上述两个原因，所以时间序列模型的检验有两类，一类是模型的显著性检验；另一类是模型参数的显著性检验. 这两类检验统称为模型的诊断性检验.

时间序列建模的基本步骤

- 3.模型检验：

模型的显著性检验主要是检验模型的有效性. 一个模型是否有效主要看它提取的相关信息是否充分, 一个好的拟合模型应该确保提取出了观察值序列中几乎所有的样本相关信息, 换言之, 拟合残差项中将不再蕴含任何相关信息, 即残差序列应该为白噪声序列 (其概念见后续章节). 反之, 如果残差序列为非白噪声序列, 那就意味着残差序列中还残留着相关信息未被提取,这就说明拟合模型不够有效, 需重新选择模型进行拟合.

时间序列建模的基本步骤

- 3.模型检验：

模型参数的显著性检验主要是检验模型中每一个参数是否显著异于零. 目的是要找出贡献不大的参数并将其剔除, 使得模型更为精简和准确. 一般地, 如果模型中包含了不显著的参数,不但使得模型参数冗余, 影响自由度, 而且也会影响其他参数的估计精度.

在实际应用中, 如果模型的诊断性检验没有通过, 则需要重新识别、估计和检验, 直到得到一个满意的拟合模型.

如果一个模型通过了检验, 说明在一定的置信水平下, 该模型能够有效地拟合观察值序列的波动, 但这种有效模型有时并不是唯一的. 面对多个显著有效的模型, 到底选择哪个来统计推断更好呢？为了解决这个问题, 一般需要引进一些信息准则来进行模型优化. 具体地, 在后继章节中, 我们结合具体模型来详细论述.

时间序列建模的基本步骤

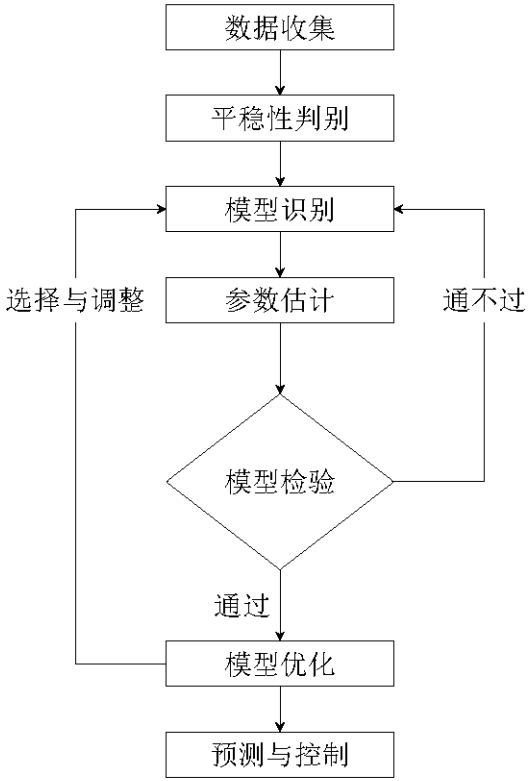
- 4.模型应用:

时间序列模型的应用主要包括变量动态结构分析、预测和控制.

- 动态结构分析是指用已经估计出参数的模型, 对变量的动态变化情况进行考察. 例如, 对于自回归模型 (AR 模型), 可以考察它的记忆特征和记忆衰减情况; 对于滑动平均模型 (MA 模型), 可以考察外部冲击对变量的影响情况和对外部冲击的记忆期限. 动态结构分析对于认识经济金融变量的运行规律具有重要作用.
- 预测是时间序列建模的最重要的目的, 是指用已经估计出参数的模型, 对变量未来变化进行预报.
- 控制是指根据时间序列模型调整输入变量使得系统发展过程保持在目标值上. 当运用时间序列模型进行预测、发现预测值会偏离目标值时, 便可进行必要的控制, 调整当前值使之朝预定目标靠近.

时间序列建模的基本步骤

- 总之，时间序列建模过程包括模型识别、模型估计、模型检验和优化，并可能反复多次才能达到比较满意的效果，最终投入使用. 时间序列分析完整的流程可用图 1.6 表示.



本章结构

1. 引言
2. 基本概念
3. 时间序列建模的基本步骤
4. 数据预处理

时序图与自相关图的绘制

- 时间序列数据建模之初，我们应该对数据有个初步的认识，如：大致观察一下数据的趋势性、季节性；序列值相近时期的相关性；初步判断一下序列的平稳性；判断序列是否有研究的必要,即检验一下序列是否是白噪声. 这些数据建模前的分析都称为数据的预处理.
- 时序图与自相关图的绘制
  - 进行时间序列分析的第一步，通常是利用序列值画出时序图和自相关图进行观察.
  - 时序图就是一张二维平面图，一般横坐标表示时间，纵坐标表示序列取值. 通过观察时序图，我们能够获得序列值的趋势和走向.
  - 自相关图是平面上的悬垂线图，横坐标表示延迟时期数，纵坐标表示自相关系数，悬垂线表示自相关系数的大小.通过自相关图我们能够大致获得不同时刻序列值之间的相关关系.

时序图与自相关图的绘制

- R 语言拥有操作简便、功能强大的绘图软件包, 可以绘制出各种各样的精美统计图. 借助于该功能我们可绘制出所需要的序列时序图和自相关图. 下面, 我们通过时序图和自相关图的绘制, 顺便介绍 R 的简单绘图命令.
- R 中最常用的绘图命令是 plot( ). plot( ) 含有丰富的参数, 用它们可绘制出美观的时序图.plot( ) 函数的命令格式如下:
- > plot(x, y, type, main, sub, xlab, ylab, xlim, ylim, pch, lty, lwd, col…)
- 该函数常用参数的说明:

时序图与自相关图的绘制

- > plot(x, y, type, main, sub, xlab, ylab, xlim, ylim, pch, lty, lwd, col…)
- 该函数常用参数的说明:
- - x, y: 各绘图点横坐标, 纵坐标构成的向量.
- - type: 指定绘图的类型. 取“p”为点图；取“l”为线图；取“b”为点连线；取“o”为线穿过点；取“h”为悬垂线；取“s”为阶梯线.
- - main: 指定主标题. - sub: 指定副标题.
- - xlab: 指定 x 轴的标签；ylab: 指定 y 轴的标签.
- - xlim: 指定横轴的上下限, 取值为上下限构成的向量；ylim: 指定纵轴的上下限, 取值为上下限构成的向量.

时序图与自相关图的绘制

- > plot(x, y, type, main, sub, xlab, ylab, xlim, ylim, pch, lty, lwd, col…)
- 该函数常用参数的说明:
- - pch: 指定观察点的符号, 可取从 1 ~ 25 的整数.
- - lty: 指定连线类型, 可取从 1 ~ 6 的整数.
- - lwd: 指定连线的宽度, 取整数.
- - col: 指定颜色, 可取正整数, 或指定颜色参数.

时序图与自相关图的绘制

- **例 1.10** 现给出 2000 年 1 月至 2012 年 10 月新西兰人出国旅游目的地数据. 我们下面一起来认识这组时间序列数据, 并用 R 绘制时序图来分析.

解: 首先, 我们读取数据的前两行来认识所给数据的结构, 然后再决定用 read.table( ) 中的哪些参数. 具体命令及运行结果如下:

```
>readLines("E:/DATA/CHAP1/NZTravellersDestination.csv", n=2)

[1]"DATE,Australia,CookIslands,Fiji,Samoa,China,India,Thailand, UnitedKingdom,UnitedStates,Other“

[2]"2000M01,23203,1071,1936,2403,2285,2270,862,5418,4109,30660"
```

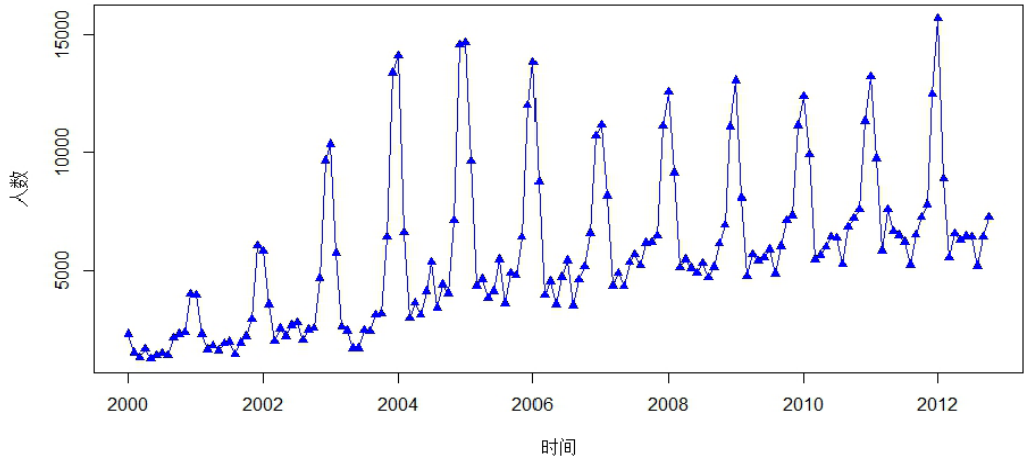
时序图与自相关图的绘制

- 其次, 我们研究新西兰人 2001 年 1 月至 2012 年 10 月之间月均到中国旅游人数的变化. 具体命令如下, 运行结果见图 1.7.

```
> x <-
read.csv("E:/DATA/CHAP1/NZTravellersDestination.csv",
+header =T)

> x1 <- ts(x$China,start=c(2000,1),frequency = 12)

> plot(x1,type="o",xlab="时间",ylab="人数
+",pch=17,col="blue")
```

时序图与自相关图的绘制

- **图 1.7**: 2000 年 1 月至 2012 年 10 月新西兰人月均来中国旅游时序图



- 从时序图 1.7, 我们可以看到从 2000 年 1 月到 2012 年新西兰人月均来中国旅游人数有增长趋势但是增幅和增速不大. 同时新西兰人月均来中国旅游人数有明显的季节性, 每年的圣诞节前后来华旅游人数最多.

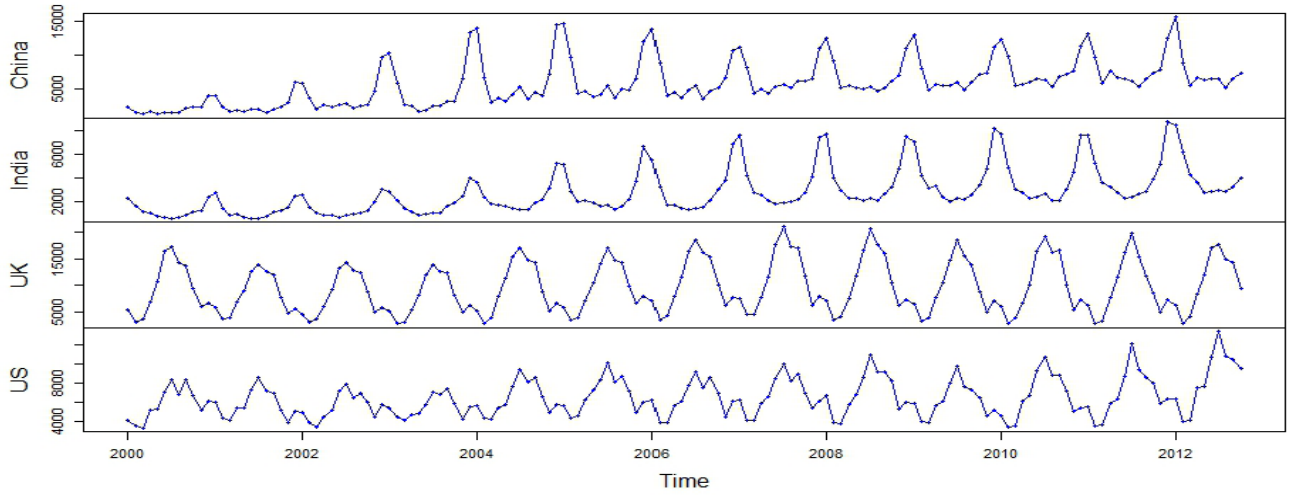
时序图与自相关图的绘制

- 我们也可以在同一个窗口，绘制不同变量的时序图，来比较这些变量之间变化. 下面比较新西兰人从 2000 年 1 月到 2012 年 10 月月均到中国、印度、英国和美国四国旅游人数变化情况.
- 具体命令如下，运行结果见图 1.8.

```
> China <-x$China; India <-x$India
> UK <-x$UnitedKingdom; US <-x$UnitedStates
> y <- data.frame(China,India,UK,US)
> Tourism <- ts(y,start=c(2000,1),frequency = 12)
> par(mfrow=c(4,1))
> plot(Tourism,type="o",pch=16,col="blue",main=" ")
```

时序图与自相关图的绘制

- **图 1.8:** 2000 年 1 月至 2012 年 10 月新西兰人月均出国旅游时序图



- 从图 1.8 可以看出, 2000 年 1 月至 2012 年 10 月新西兰人月均到中国和印度的人数都有增.长趋势, 来中国旅游的人数增长更快些. 去英国和美国的人数比较稳定. 另外, 这四组数据都有明显的季节性. 在每年 12 月左右去中国、印度旅游的人数最多; 而在夏季新西兰人更乐意去英国和美国旅游.

时序图与自相关图的绘制

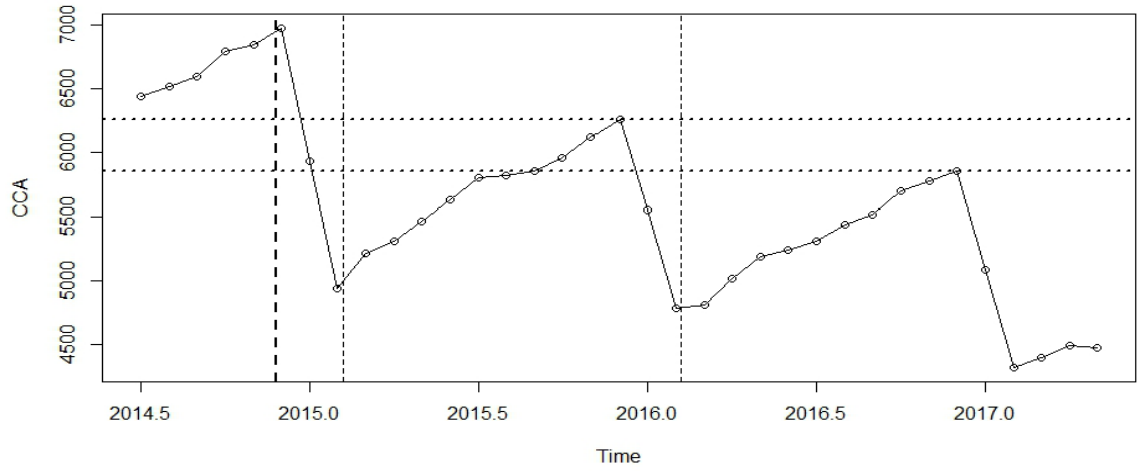
- 在绘图时, 为了突出比较效果, 可以使用 abline( ) 为图形添加参照线. 参照线可以是水平线,也可以是垂线, 还可以是线性回归线. 下面通过例子来说明 abline( ) 的使用.
- **例 1.11 :** 绘制 2014 年 7 月至 2017 年 5 月北京市商品住宅施工面积累计值 的时序图, 并添加辅助线来比较. 具体命令如下, 运行结果见图 1.9.
- 解:
- > w <- read.csv("E:/DATA/CHAP1/Beijing commodity +housing.csv", header=T)

  > z <- w$CCA; y <- na.spline(z)

  > CCA <- ts(y,start=c(2014,7),frequency = 12)

时序图与自相关图的绘制

- > plot(CCA,type="o",col=1)
- > abline(v=2014.9,lty=2, col=1,lwd=2)
- > abline(v=c(2015.1,2016.1),lty=2,col=1)
- > abline(h=c(6261.21,5857.61), lty=3, col=1,lwd=2)
- **图 1.9 :**2014 年 7 月至 2017 年 5 月北京市商品住宅施工面积累计值时序图



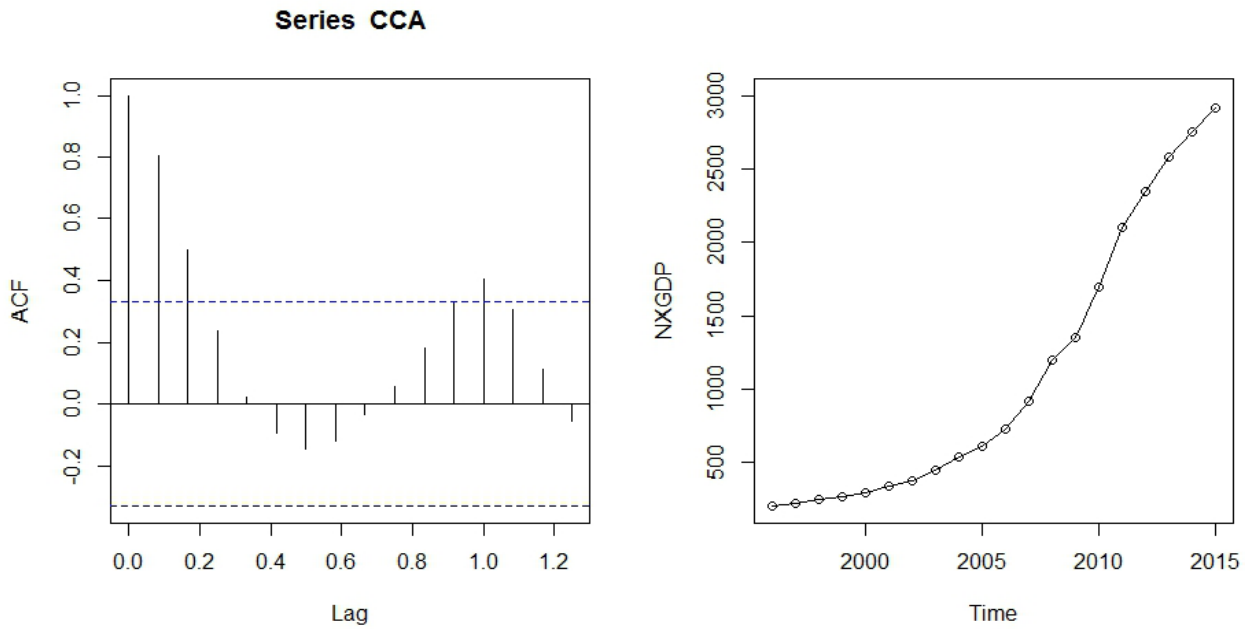
时序图与自相关图的绘制

- 在 R 中绘制自相关图使用函数 acf( ), 这个函数的命令格式为:
- acf(x, lag)
- 函数 acf( ) 的参数说明:
- -x: 是时间序列数据构成的向量.
- -lag: 是延迟的阶数. 若用户不特殊指定的话, 系统会根据序列长度自动指定延迟阶数.
- **例 1.12 :** 接例 1.11, 绘制 2014 年 7 月至 2017 年 5 月北京市商品住宅施工面积累计值的自相关图, 命令如下, 运行结果如图 1.10 所示.
- > acf(CCA)　　#绘制自相关图

时序图与自相关图的绘制

- 图 1.10 中的虚线为自相关函数的 2 倍标准差位置. 一般地, 如果悬垂线夹在两条虚线之间,那么可认为此时自相关函数非常接近于零.

- **图 1.10：**商品住宅施工面积累计值自相关图



数据平稳性的图检验

- **1. 时序图检验**

根据平稳性的定义, 平稳时间序列的均值和方差均为常数, 因此平稳时间序列的时序图应该围绕一条水平线上下波动, 而且波动的范围有界. 如果序列时序图显示出了明显的趋势性或周期性, 那么它通常不是平稳的时间序列. 根据这个性质, 许多时间序列通过时序图就可看出它的非平稳性.

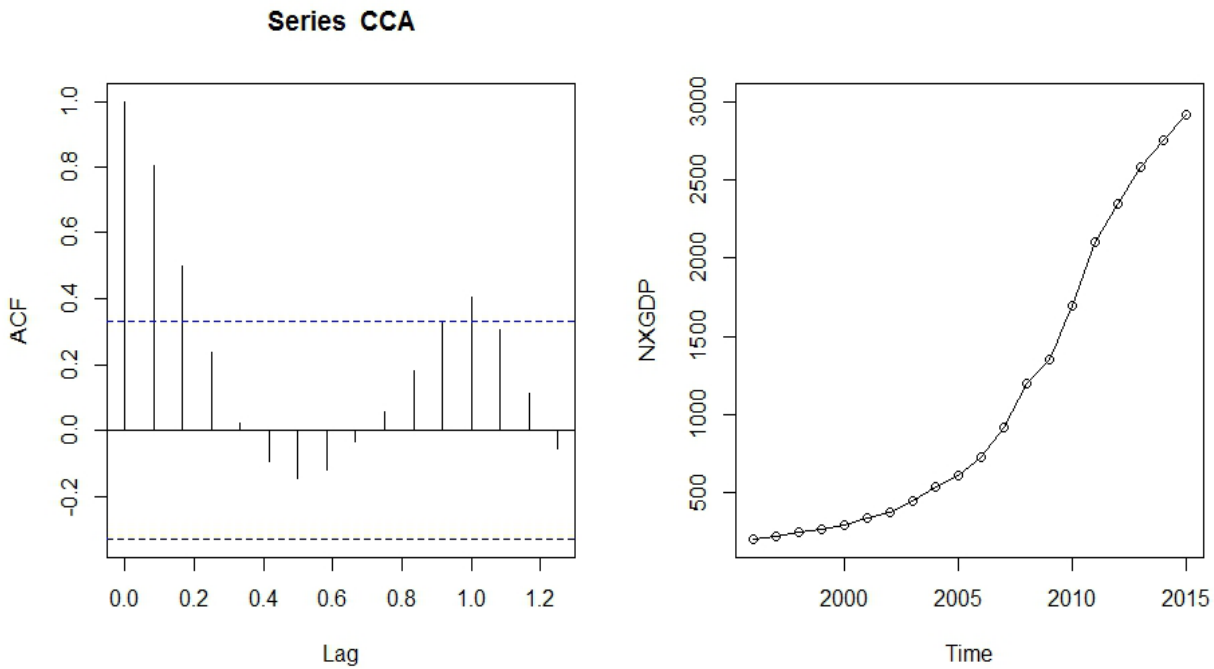
例 1.13：绘制 1996 年至 2015 年宁夏回族自治区地区生产总值 (单位: 亿元) 的时序图. 命令如下, 运行结果见图 1.11.

```
> a <-
read.table(file="E:/DATA/CHAP1/SMGDP.csv",sep=",",header
=T)
```

数据平稳性的图检验

```
> NXGDP <- ts(a$NX,start=1996)

> plot(NXGDP,type="o",ylim=c(200,3000))
```

图 1.11：宁夏回族自治区地区生产总值时序图

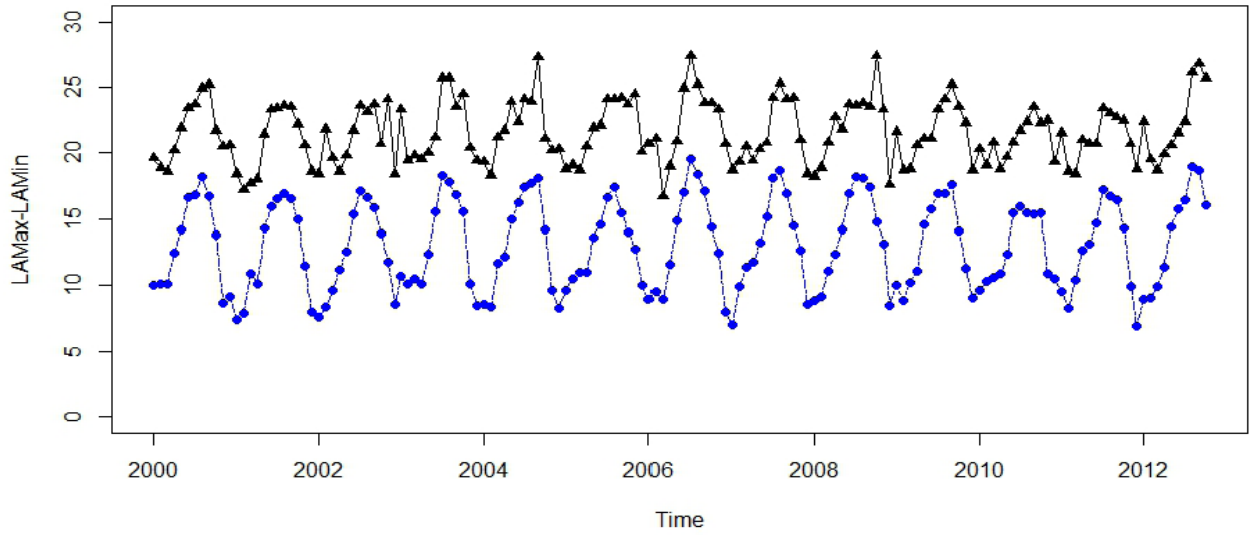


数据平稳性的图检验

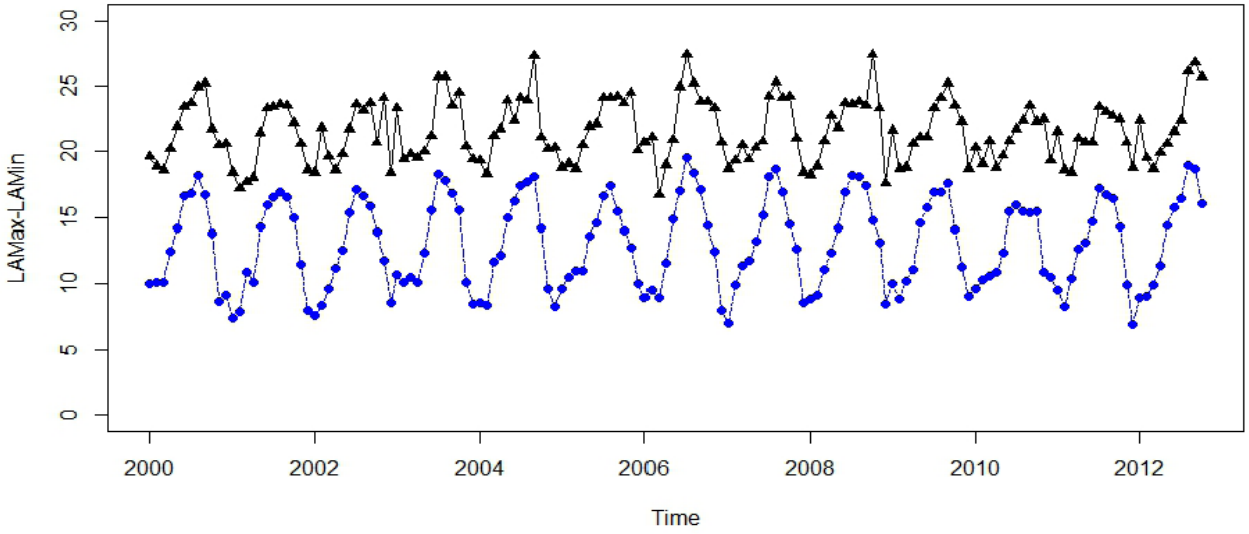
- 例 1.15 绘制 2000 年 1 月至 2012 年 10 月美国洛杉矶月平均最高到最低气温时序图. 具体命令如下, 运行结果见图 1.12.
- > Temp <- read.csv("E:/DATA/CHAP1/TempUSA.csv",header=T)
- > LAMax <- Temp$LosAngelesMax;LAMin <- Temp$LosAngelesMin
- > LAMax <- ts(LAMax,start=c(2000,1),frequency = 12)
- > LAMin <- ts(LAMin,start=c(2000,1),frequency = 12)
- >plot(LAMax,type="o",pch=17,lty=1,ylim=c(0,30),ylab="LAMax-LAMin")
- > points(LAMin,type="o",pch=16,lty=6,col="blue")

数据平稳性的图检验

- 如果连续地调用函数 plot( ), 那么每次都会创建一个新图来取代前一次调用产生的图. 为了解决该问题, 我们可以在调用 plot( ) 之后, 通过调用 points( ) 来叠加图形. 函数 points( ) 的参数与函数 plot( ) 的参数相同.



数据平稳性的图检验



从图 1.12 可以看出，洛杉矶月平均最高气温和月平均最低气温分别围绕在 22.5 ℃ 和 13 ℃附近随机波动，没有明显的趋势，却有很强的周期，因此还不能断定为平稳序列. 不过，我们可以通过自相关图来进一步识别.

数据平稳性的图检验

- **2. 自相关图检验**

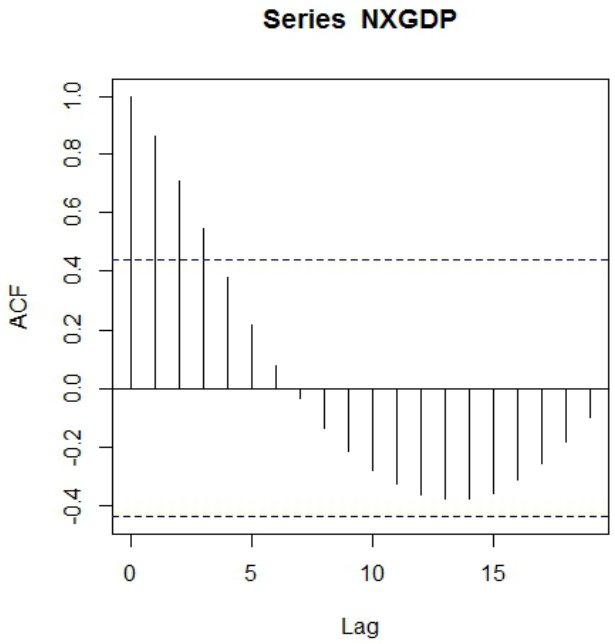
平稳时间序列的一个显著特点是序列值之间具有短期相关性, 这一点我们将在后继的章节中给予证明. 短期相关性突出的表征是, 随着延迟期数的增加, 平稳序列的自相关系数会很快地衰减为零. 而非平稳序列的自相关系数衰减为零的速度通常比较慢. 利用自相关系数的上述特点, 我们可以进一步识别序列的平稳性.

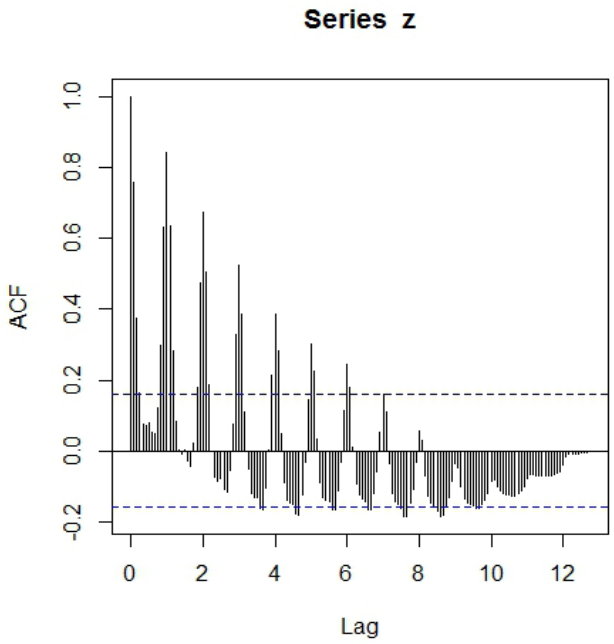
例 1.16：绘制 1996 年至 2015 年宁夏回族自治区地区生产总值的自相关图. 纵轴是自相关系数值, 大小用悬垂线表示; 横轴是延迟期数. 具体命令如下, 运行结果见图 1.13.

```
> a <-
read.table(file="E:/DATA/CHAP1/SMGDP.csv",sep=",",header
=T)
```

数据平稳性的图检验

- > NXGDP <- ts(a$NX,start=1996)
- > acf(NXGDP,lag=20)





数据平稳性的图检验

- 从图 1.13 中, 我们发现序列的自相关系数递减到零的速度比较缓慢, 而且在较长延迟期里自相关系数一直为正, 之后又一直为负. 在自相关图上显示出三角对称的关系, 这是具有单调趋势的非平稳序列的一种典型的自相关图形式. 这和该序列的时序图 (图 1.11) 显示的单调递增性是一致的.
- 例 1.17 绘制 2000 年 1 月至 2012 年 10 月新西兰人月均来华旅游人数的自相关图. 具体命令如下, 运行结果见图 1.14.
- > x <-read.csv("E:/DATA/CHAP1/NZTravellersDestinatio

  +n.csv",header= T)
- > China <- x$China; z <-ts(China,start=c(2000,1),frequen

  +cy= 12); acf(z,lag=154)

数据平稳性的图检验

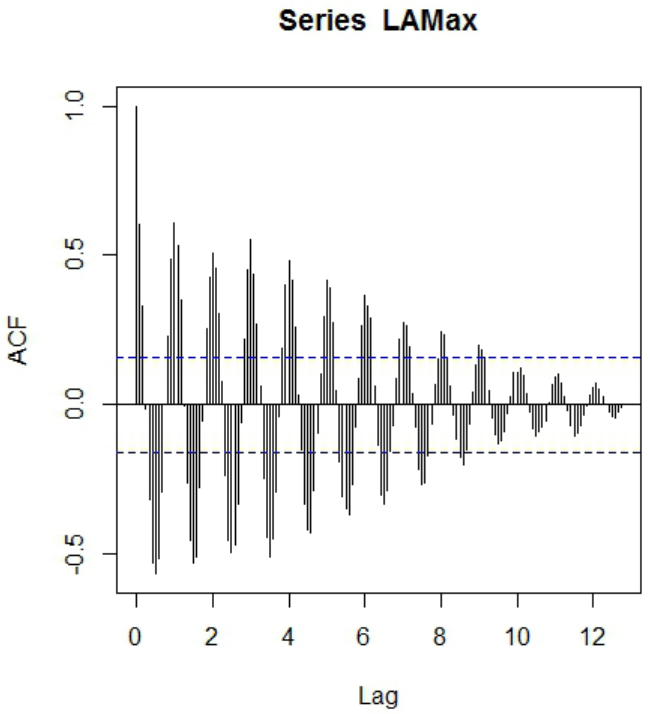
- 自相关图 1.14 显示序列自相关系数衰减到零的速度非常缓慢, 而且呈现明显的周期规律,这是具有周期变化规律和递增趋势的非平稳序列的典型特征. 自相关图显示出来的特征与时序图 1.7 显示的带长期递增趋势和周期的性质也高度吻合.
- **例 1.18:** 绘制 2000 年 1 月至 2012 年 10 月美国洛杉矶月平均最高气温的自相关图. 具体命令如下, 运行结果见图 1.15.

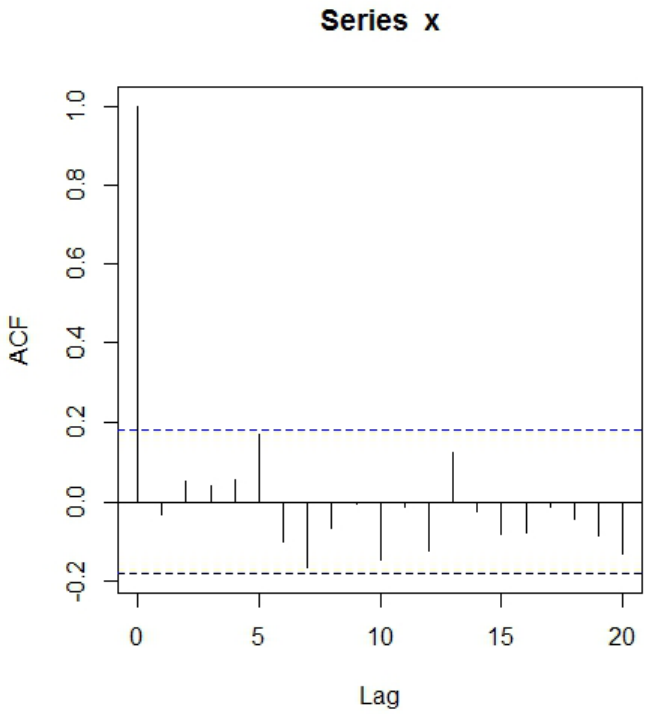
```
Temp <-read.csv("E:/DATA/CHAP1/TempUSA.csv",header=T)

> LAMax <- Temp$LosAngelesMax;LAMin <-
+Temp$LosAngelesMin
```

- > LAMax <- ts(LAMax,start=c(2000,1),frequency = 12)
- > acf(LAMax,lag=154)

数据平稳性的图检验

- 自相关图 1.15 显示序列自相关系数呈现长期周期衰减，而且自相关系数值正、负基本各半地居于横轴上下两侧，从而判断序列是非平稳序列.





数据平稳性的图检验

- **例 1.19**：绘制美国加利福尼亚州洛杉矶地区 115 年来的年降水量的自相关图. 具体命令如下, 运行结果见图 1.16.
- > x <- scan("E:/DATA/CHAP1/data1.3.txt")
- > x <- ts(x, start=1878)
- > acf(x)
- 从图 1.16 看出, 自相关系数 1 阶延迟之后, 立即衰减到零附近, 也不具有明显的周期特征,因此可以判断序列是平稳序列.

数据的纯随机性检验

- 当识别出一个序列是平稳时间序列之后，我们需要进一步分析该序列是否为纯随机序列，因为如果是纯随机序列的话，那么意味着序列值之间没有相关关系，该序列就成为所谓的无记忆序列,即过去的行为对将来的发展没有丝毫影响，这样的序列从统计分析的角度而言无任何研究的意义.

- 1. 纯随机序列的概念和性质

如果一个时间序列$\{X_t, t \in T\}$满足如下条件:

(1) 对于$\forall t \in T$有$EX_t = \mu$; (2) 对于$\forall t, s \in T$,有

$$\gamma(t,s) = \begin{cases} \sigma^2, t = s; \\ 0, t \neq s. \end{cases}$$

那么称$\{X_t, t \in T\}$ 为纯随机序列 (pure random sequences), 或称为白噪声序列 (whitenoise series), 简记为$X_t \sim WN(\mu, \sigma^2)$

数据的纯随机性检验

- 显然白噪声序列一定是平稳序列，而且是最简单的平稳序列. 需要注意的是，虽然白噪声序列简记为 $X_t \sim WN(\mu,\sigma^2)$，但是 X_t 不一定服从正态分布.

- 从白噪声序列的定义易得

$$\gamma(k)=0,\forall k\neq 0.$$

这说明白噪声序列的各项之间没有任何相关关系，序列在进行无序的纯随机波动，这是白噪声序列的本质特征. 在统计分析中，如果某个随机事件呈现出纯随机波动的特征，那么该随机事件就不含有任何值得提取的有用信息，从而分析应该终止.

数据的纯随机性检验

- 相反地，如果序列的某个延迟 k 自协方差函数不为零，即

$$\gamma(k) \neq 0, \exists k \neq 0.$$

那么说明该序列不是纯随机序列，其间隔 k 期的序列值之间存在一定程度的相互影响关系，也即具有相关信息. 我们分析的目的就是要把这种相关信息从观察值序列中提取出来. 如果观察值序列中蕴含的相关信息完全被提取出来，那么剩下的残差序列就应该呈现出纯随机序列的性质. 因此，纯随机性还是判别相关信息提取是否充分的一个判别标准.

- 2. 纯随机性的检验

纯随机性检验也称为白噪声检验，在学习它之前，我们先通过时序图和自相关图感知一下白噪声序列的序列值走向和相关程度的表现.

数据的纯随机性检验

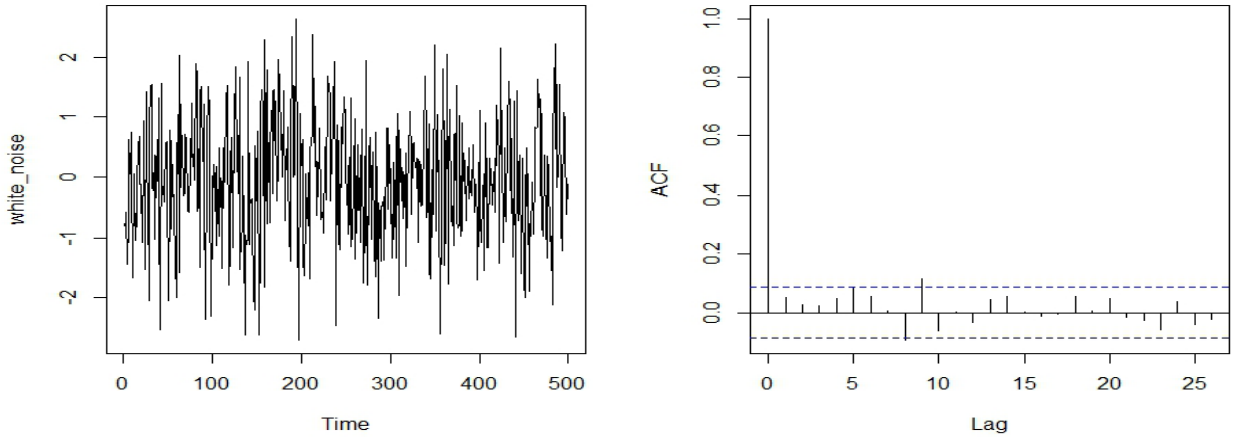
- **例 1.20**：随机产生 500 个服从标准正态分布的白噪声观察值序列，并绘制时序图和自相关图.
- 解: 利用 R 提供的随机数生成器，我们可以产生随机数. 其操作非常简单，一般都是在对应分布的名前加前缀 r，如正态分布随机数生成器是 rnorm( )，均匀分布随机数生成器是 runif( )，泊松
- 分布随机数生成器是 rpois( )，等等. 正态分布随机数生成器 rnorm( ) 的命令格式为:
- rnorm(n=, mean=, sd=)
- 函数 rnorm() 参数说明:
- -n: 将产生的随机数个数; -mean: 正态分布的均值，缺省默认值为 0; -sd: 标准差，缺省默认值为 1.

数据的纯随机性检验

本例的命令如下, 运行结果见图 1.17.

```
> white_noise <- rnorm(500)
> white_noise <- ts(white_noise)
> par(mfrow=c(2,1))
> plot(white_noise, main=" ")
> acf(white_noise,main=" ")
```

数据的纯随机性检验



标准正态白噪声序列的序列值围绕横轴波动，波动范围有界. 但波动既无趋势性，也无周期性，表现出明显的随机性. 图也显示白噪声序列的样本自相关系数并非都为零，但是这些自相关系数都非常小，在零值附近做小幅波动. 这提示我们应该考虑样本自相关系数的分布，构造统计量来检验序列的纯随机性.

数据的纯随机性检验

- 现在我们来学习白噪声检验. 首先我们介绍一个关于白噪声序列延迟非零期的样本自相关函数渐近分布的定理, 该定理由 Barlett 给出.

定理 1.1 如果$\{X_t, t \in T\}$是一个白噪声序列, 而$\{x_t, 1 \le t \le n\}$为该白噪声序列的一个观察期数为 n 的观察值序列, 那么该序列延迟非零期的样本自相关函数近似服从均值为零,且方差为序列观察期数倒数的正态分布, 即

$$\hat{\rho}(k) \sim N(0, 1/n), \forall k \neq 0.$$

- 下面借助于定理 1.1, 构造检验统计量来检验序列的纯随机性. 根据检验对象提出如下假设条件:

数据的纯随机性检验

- 原假设 $H_0: \rho(1)=\rho(2)=\cdots=\rho(m)=0, \forall m \geq 1;$
- 备择假设 H_1: 至少存在某个 $\rho(k) \neq 0, \forall m \geq 1, k \leq m.$

原假设 H_0 意味着延迟期数小于或等于 m 的序列值之间互不相关; 备择假设 H_1表明延迟期数小于或等于 m 的序列值之间存在某种相关性.

在样本容量 n 很大的情况下, Box 和 Pierce 构造了如下统计量:

$$Q_{BP}=n\sum_{k=1}^{m}\hat{\rho}^2(k)$$

其中, n 为序列观察期数; m 为指定延迟期数. 根据正态分布和卡方分布之间的关系, 易得 Q_{BP} 近似服从自由度为 m 的卡方分布, 即

$$Q_{BP}=n\sum_{k=1}^{m}\hat{\rho}^2(k)\sim\chi^2(m)$$

数据的纯随机性检验

- 当统计量 Q_{BP} 大于 $\chi^2_{1-\alpha}(m)$ 分位数，或它的 p 值小于 α，则以 $1-\alpha$ 的置信水平拒绝原假设，并有理由认为备择假设成立，即该序列为非白噪声序列；否则，接受原假设，认为该序列为白噪声序列.

- 在小样本情形，统计量 Q_{BP} 检验效果已不太精确. 为克服这一缺陷，Box 和 Ljung 将统计量 Q_{BP} 修正为统计量 Q_{LP} ：

$$Q_{LB} = n(n+2)\sum_{k=1}^{m}\frac{\hat{\rho}^2(k)}{n-k}$$

其中，n 为序列观察期数；m 为指定延迟期数. Box 和 Ljung 也证明了统计量 Q_{LB} 同样近似服从自由度为 m 的卡方分布.

- 统计量 Q_{BP} 和 Q_{LB} 统称为 Q 统计量. 在实际中，各种检验场合普遍采用的 Q 统计量通常指的是 Q_{LB} 统计量.

数据的纯随机性检验

- 在 R 中使用函数 Box.test() 进行白噪声检验. 该函数的命令格式为

Box.test(x, type=, lag=)

- 函数 Box.test() 的参数说明:

- -x: 变量名, 可以是数值向量, 也可以是一元时间序列名.

- -type: 检验统计量类型:

(1) type=“Box-Pierce”, 输出白噪声检验的 Q_{BP} 统计量. 该统计量是默认输出结果.(2) type=“Ljung-Box”, 输出白噪声检验的 Q_{LB} 统计量.

- -lag: 延迟阶数. lag=n 表示输出滞后 n 阶的白噪声统计量. 忽略该选项时, 默认输出滞后1 阶的检验统计量结果.

数据的纯随机性检验

- 在利用函数 Box.test( ) 进行白噪声检验时，我们一般取延迟阶数不会太大，这是因为平稳序列通常具有短期相关性. 如果序列值之间存在显著的相关关系，通常只存在于延迟时期比较短序列值之间. 如果一个平稳序列短期延迟的序列值之间都不存在显著的相关关系，通常长期延迟之间就更不会存在显著的相关关系了. 同时，如果一个序列显示了短期相关性，那么该序列就一定不是白噪声序列，我们就可以对序列值之间的相关性进行分析. 由于对平稳序列而言，自相关函数随着延迟期数的增长而逐渐趋于零，因此假若考虑的延迟期数太长，反而可能淹没了该序列的短期相关性. 这一点我们在之后的章节将会进一步阐释.

数据的纯随机性检验

- **例 1.21**：计算例 1.20 中白噪声序列分别延迟 6 期和 12 期的 Q_{BP} 统计量的值，并判断该序列的随机性 ($\alpha = 0.05$). 具体命令及运行结果如下:

- 解: > Box.test(white_noise,lag=6)

  Box-Pierce test

  data: white_noise

  X-squared = 2.0304, df = 6, p-value = 0.9169

  > Box.test(white_noise,lag=12)

  Box-Pierce test

  data: white_noise

  X-squared = 6.838, df = 12, p-value = 0.8681

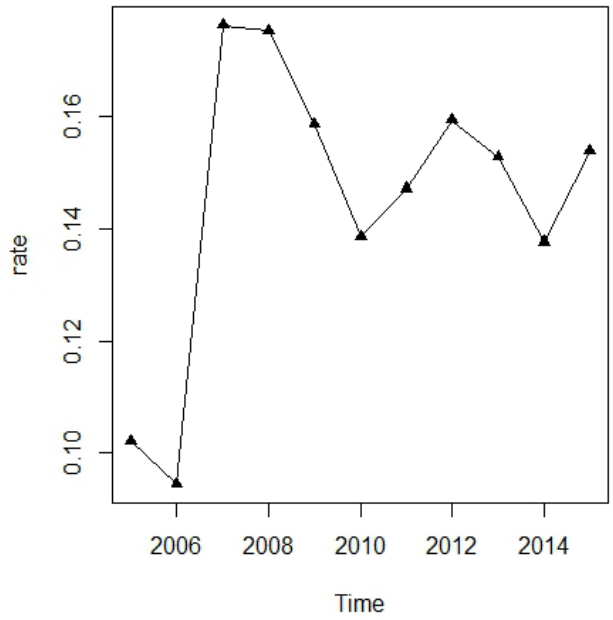
数据的纯随机性检验

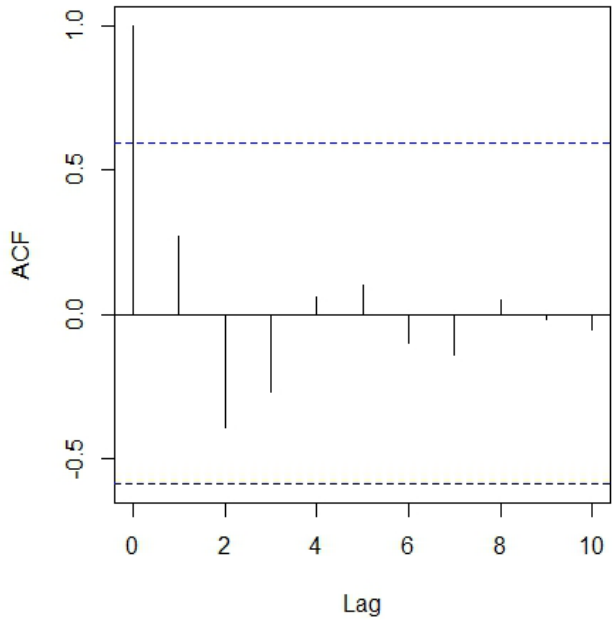
- **例 1.22**：对 2005 年至 2015 年苏格兰百岁老人男女之比序列的平稳性和纯随机性进行检验. 具体命令及运行结果如下:

解: > z <- read.csv("E:/DATA/CHAP1/Centenarians by sex.csv",header=T)

```
> z <- z$ratio

> ratio <- ts(z,start=2005)

> par(mfrow=c(1,2))

> plot(ratio,type="o",main=" ",ylab="rate",pch=17)

> acf(ratio,main=" ")
```

数据的纯随机性检验


rate
0.10
0.12
0.14
0.16
2006
2008
2010
2012
2014
Time


ACF
-0.5
0.0
0.5
1.0
0
2
4
6
8
10
Lag

数据的纯随机性检验

```
> for(i in 1:2) print(Box.test(ratio,lag=5*i))

Box-Pierce test

data: ratio

X-squared = 3.4811, df = 5, p-value = 0.6263

Box-Pierce test

data: ratio

X-squared = 3.8636, df = 10, p-value = 0.9533
```

练习

1. 什么是时间序列? 请列举生活中观察到的时间序列的例子, 并收集关于这些例子的观测

值. 根据所收集到的观测值序列, 绘制时序图.

2. 时间序列分析的方法有哪些? 分别简述时域方法和频域方法的发展轨迹和特点.

3. 何谓严平稳? 何谓宽平稳? 简述它们之间的区别和联系?

4. 简述平稳性假设的统计意义.

5. 简述时间序列建模全过程.

6. 何谓时间序列平稳性的图检验法? 请简述图检验思想.

7. 什么是白噪声序列? 简述白噪声检验方法及其操作过程