

第 1 章 引言及基础知识

学习目标与要求

1. 了解时间序列分析的发展简史.
2. 理解时间序列的基本概念和主要特征.
3. 理解时间序列分析的基本步骤.
4. 掌握 R 语言的基本操作.
5. 学会时间序列数据预处理的方法.

1.1 引言

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. 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 (1789—1875) 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.

另外, 许多经济现象的发展都具有随时间演变的特征. 例如: 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.

1.1.1 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

$$X_1, X_2, \dots, X_t, \dots \quad (1.1)$$

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

$$x_1, x_2, \dots, x_n \quad (1.2)$$

或

$$\{x_t, t = 1, 2, \dots, n\}$$

表示该随机序列的 n 个**有序观察 (测) 值**, 称为序列长度为 n 的**观察 (测) 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 $\{x_t\}$.

下面介绍一些时间序列的例子.

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. 图 1.1 为国内生产总值年度时间序列的时序图. 该图是用下列 R 语句生成的 (全书中假设所涉及的数据文件存放在 E 盘的 DATA 子目录下, 读者可根据自己的情况进行调整).

```
> x <- read.table("E:/DATA/CHAP1/data1.1.csv", sep=",", header=T)
> GDP <- ts(x$GDP, start=1953)
> plot(GDP, type="o", xlab="年份", ylab="国内生产总值(GDP)", col=1)
```

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. 从图 1.2 看出相邻年度 GDP 的关联呈线性.

```
> y <- GDP[-1]
> x <- GDP[-64]
> plot(x,y,xlab="上一年GDP", ylab="当年GDP", pch=16, col=1)
```

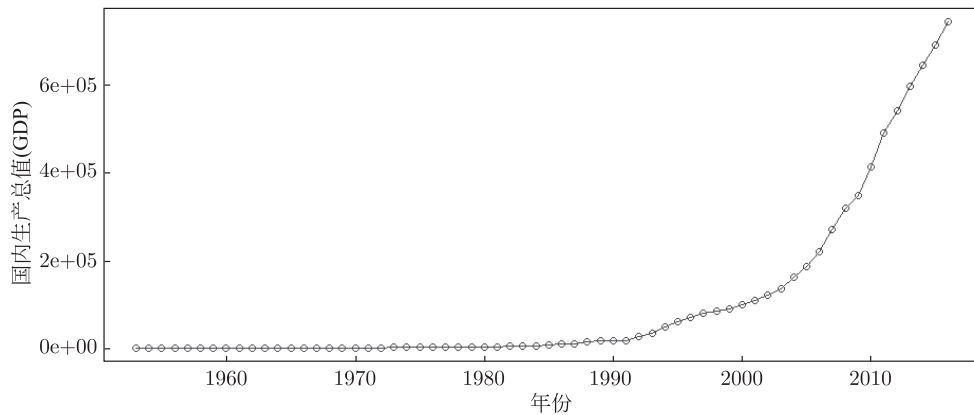


图 1.1 中国 1953 年至 2016 年国内生产总值年度时序图

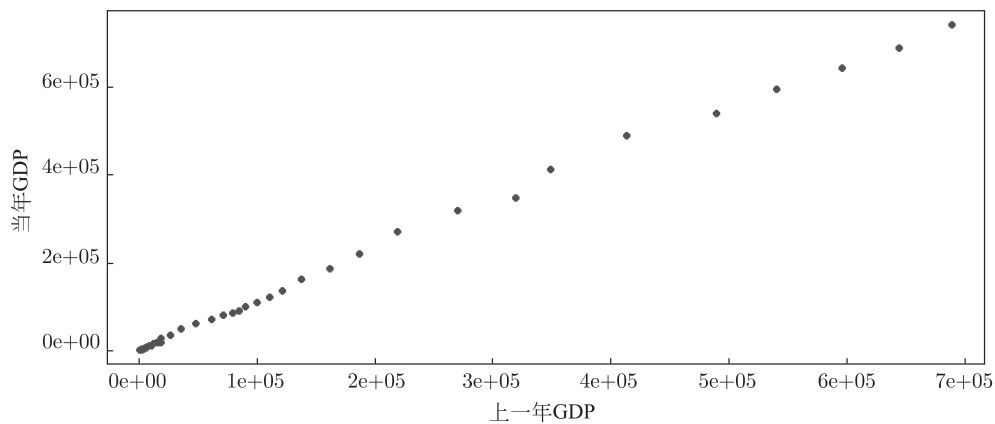


图 1.2 中国当年 GDP 与上一年 GDP 的散点图

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: $^{\circ}\text{F}$) 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. 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.

```
> 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)
```

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 语句生成其时序图 (见图 1.4).

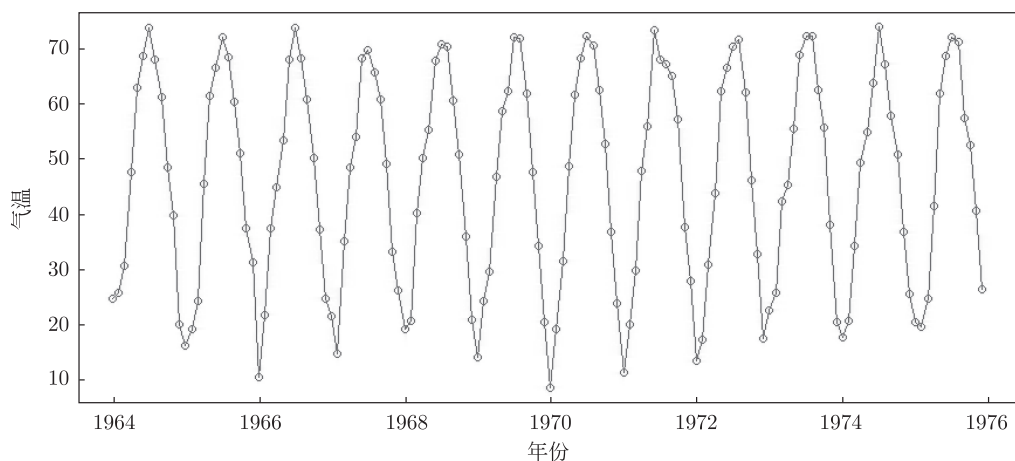


图 1.3 美国爱荷华州杜比克市月平均气温图

```

> x <- scan("E:/DATA/CHAP1/data1.3.txt")
> x <- ts(x, start=1878)
> plot(x,type="o",xlab="年份",ylab="降水量 单位:英寸",col=4)

```

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. 由下列 R 语句生成图 1.5. 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.

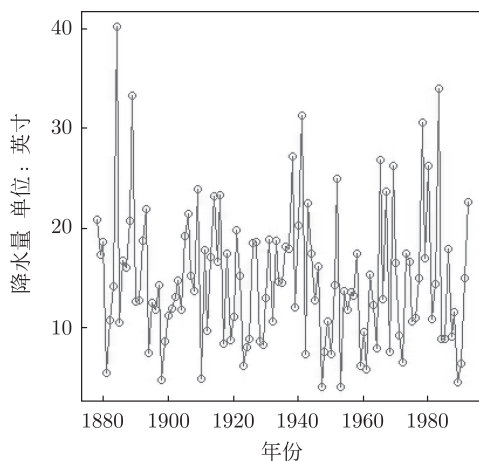


图 1.4 洛杉矶年降水量的时序图

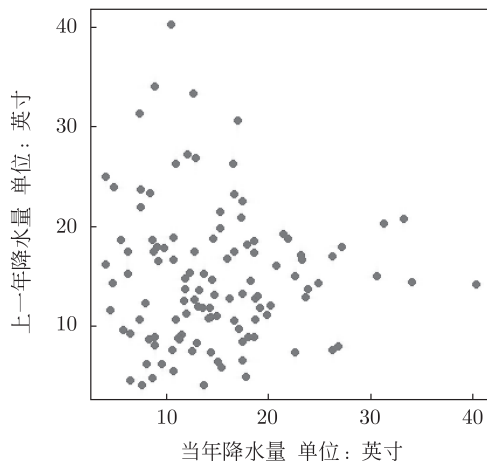


图 1.5 洛杉矶当年与上一年降水量散点图

```
> a <- x[-1]
> b <- x[-115]
> plot(a,b,xlab="当年降水量 单位:英寸",ylab="上一年降水量 单位:英寸",pch=+19,col=4)
```

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.

1.1.2 时间序列的分类

在现实中存在不同类别的时间序列. 根据所研究问题的不同, 可以对时间序列做如下不同的分类.

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.

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**. 例如, 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.

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.

1.1.3 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,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辨率不高和频率泄露等缺点,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.

时域分析方法的起源可以追溯到20世纪20年代英国统计学家 G. U. Yule 在分析和预测市场变化规律等问题中提出的自回归模型 (AR 模型).同时,英国科学家 G. T. Walker 爵士在研究气象问题时得到著名的 Yule-Walker 方程.这些开创性工作奠定了时域分析方法的基础.20世纪60年代之后,随着计算机技术和数据处理技术的迅速发展,时间序列分析的理论和应用得到迅猛发展.1970年,统计学家 G. E. P. Box 和 G. M. Jenkins 在梳理、发展已有研究成果的基础上,合作出版了 *Time Series Analysis: Forecasting and Control* 一书.该书系统地阐述了对求和自回归移动平均模型 (ARIMA 模型) 的识别、估计、检验和预测的原理和方法.这些方法

已经成为经典的时域分析方法.

在此基础上, 人们不断拓展研究方法. 20 世纪 80 年代以来, 统计学家逐步转向多变量场合、异方差场合和非线性场合的时间序列分析方法的研究, 并取得突破性的进展. 1982 年, R. F. Engle 在研究英国通货膨胀率的建模问题时, 提出了自回归条件异方差模型 (ARCH 模型). 而 Bollerslov 在 1985 年提出的广义自回归条件异方差模型 (GARCH 模型) 则进一步放宽了自回归条件异方差模型的约束条件. 之后, Nelson 等人又提出指数广义自回归条件异方差模型 (EGARCH 模型)、方差无穷广义自回归条件异方差模型 (IGARCH 模型) 和依均值广义自回归条件异方差模型 (GARCH-M 模型) 等限制条件更为宽松的异方差模型, 大大推广和补充了自回归条件异方差模型. 它们比传统的方差齐性模型更准确地刻画了金融市场风险的变化过程. R. F. Engle 因此获得 2003 年的诺贝尔经济学奖. 在多变量方面, C. Granger 于 1987 年提出协整理论, 极大地促进了多变量时间序列分析方法的发展. C. Granger 也于 2003 年获得诺贝尔经济学奖. 在非线性场合, 各种新的模型纷纷被提出. Granger 和 Anderson 在 1978 年提出双线性模型; 汤家豪于 1989 年提出门限自回归模型; Chen 和 Tasy 于 1993 年提出非线性可加模型, 等等. 非线性模型是个异常广阔的研究领域, 在该领域中, 模型构造、参数估计、参数检验等各方面都有大量的研究工作需要完成.

1.2 基本概念

在本节中, 我们介绍一些时间序列分析过程中的基本概念. 这些基本概念表明了本书中所研究的时间序列的主要统计性质.

1.2.1 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 $(\Omega, \mathcal{F}, P)$ 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.

1.2.2 概率分布族及其特征

由数理统计的知识可知, 分布函数能够完整地描述一个随机变量的统计性质. 同样, 要刻画时间序列的统计特征, 就要探讨一系列随机变量的统计分布.

设 $\{X_t, t \in T\}$ 为一个随机过程, 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 $F_{X_t}(x) = P(X_t \leq x)$ 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} 的联合分布函数为 $F_{X_{t_1}, X_{t_2}}(x_1, x_2) = P(X_{t_1} \leq x_1, X_{t_2} \leq x_2)$, 称为**时间序列的二维联合分布**.

一般地, 任取正整数 n 以及 $t_1, t_2, \dots, t_n \in T$, 则 n 维向量 $(X_{t_1}, X_{t_2}, \dots, X_{t_n})^T$ 的联合分布函数为

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

这些有限维分布函数的全体

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

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 $\{X_t, t \in T\}$ 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. 事实上, 在时间序列分析中, 更简单实用的方法是通过数字特征来研究其统计规律. 常用的关于时间序列的数字特征有如下几种.

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 $t \in T$ 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 $E(X_t)$.

2. 方差函数

当对于所有 $t \in T$

$$\int_{-\infty}^{+\infty} x^2 dF_{X_t}(x) < \infty$$

成立时, 我们称

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

为时间序列 $\{X_t, t \in T\}$ 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.

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 $\{X_t, t \in T\}$, 任取 $t, s \in T$, 称

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

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

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

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_t 和 X_s 之间的其他变量 $X_{s+1}, X_{s+2}, \dots, X_{t-1}$ 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

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

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.

1.2.3 平稳时间序列的定义

对时间序列进行统计推断时,通常要对其做出某些简化的假设,其中最重要的假设是平稳性. 根据限制条件的严格程度,时间序列的平稳性可分为严平稳和宽平稳两个层面.

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 , 任取 $t_1, t_2, \dots, t_n \in T$ 以及任意正数 h , 都有

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

则称时间序列 $\{X_t, t \in T\}$ 为**严平稳时间序列 (strictly stationary time series)**, 简称**严平稳序列**.

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 $\{X_t, t \in T\}$ 的有限维分布族是极其困难的事情, 而在此基础上判断一个时间序列是否属于严平稳则更难, 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.

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) 对于任意的 $t \in T$, 有 $EX_t = \mu$, μ 为常数;
- (2) 对于任意的 $t \in T$, 有 $EX_t^2 < \infty$;
- (3) 对于任意的 $s, t, k \in T$, 且 $k+t-s \in T$ 有

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

则称 $\{X_t, t \in T\}$ 为**宽平稳时间序列 (weakly stationary time series)**, 简称**宽平稳序列**. 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. 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.