第二章 时间序列的预处理

本章结构

01 平稳序列的定义

02 平稳性检验

03 纯随机性检验

2.1、平稳性的定义

- 概率分布的意义
  - 随机变量族的统计特性完全由它们的联合分布函数或联合密度函数决定
- 时间序列概率分布族的定义

$$\{F_{t_1,t_2,\cdots,t_m}(x_1,x_2,\cdots,x_m)\ ,\ \forall 整数 m \geq 1,\ \forall t_1,t_2,\cdots,t_m \in T\}$$

- 实际应用的局限性

特征统计量

- 均值函数 $\mu_t = EX_t = \int_{-\infty}^{\infty} x dF_t(x)$
- 方差函数 $DX_t = E(X_t - \mu_t)^2 = \int_{-\infty}^{\infty} (x - \mu_t)^2 dF_t(x)$
- 自协方差函数 $\gamma(t,s) = E(X_t - \mu_t)(X_s - \mu_s)$
- 自相关系数（ACF） $\rho(t,s) = \frac{\gamma(t,s)}{\sqrt{DX_t \cdot DX_s}}$

平稳时间序列的定义

- 严平稳
  - 严平稳是一种条件比较苛刻的平稳性定义，它认为只有当序列所有的统计性质都不会随着时间的推移而发生变化时，该序列才能被认为平稳。

- 严平稳序列的定义：

$$\forall \text{正整数} m,\ \forall t_1, t_2, \cdots, t_m \in T,\ \forall \text{正整数} \tau, \text{有}$$

$$F_{t_1, t_2 \cdots t_m}(x_1, x_2, \cdots, x_m) = F_{t_{1+\tau}, t_{2+\tau} \cdots t_{m+\tau}}(x_1, x_2, \cdots, x_m)$$

▪ 宽平稳

– 宽平稳是使用序列的特征统计量来定义的一种平稳性。它认为序列的统计性质主要由它的低阶矩决定，所以只要保证序列低阶矩平稳（二阶），就能保证序列的主要性质近似稳定。

宽平稳序列的定义：

1) $EX_t^2 < \infty, \forall t \in T$

2) $EX_t = \mu, \mu$为常数，$\forall t \in T$

3) $\gamma(t,s) = \gamma(k, k+s-t)$，$\forall t, s, k$且$k+s-t \in T$

严平稳与宽平稳的关系

- 一般关系
  - 严平稳条件比宽平稳条件苛刻，通常情况下，严平稳（低阶矩存在）能推出宽平稳成立，而宽平稳序列不能反推严平稳成立
- 特例
  - 序列严平稳，若其不存在低阶矩时，序列不是宽平稳的。

例如：服从柯西分布的严平稳序列不是宽平稳序列。

  - 当序列服从多元正态分布时，宽平稳等价于严平稳。

平稳时间序列的统计性质

- 常数均值
- 自协方差函数和自相关函数只依赖于时间的平移长度而与时间的起止点无关
  - 延迟 k 自协方差函数

$$\gamma(k)=\gamma(t,t+k),\forall k\text{为整数}$$

  - 延迟 k 自相关系数

$$\rho_k=\frac{\gamma(k)}{\gamma(0)}$$

自相关系数的性质

- 规范性　$\rho_0 = 1$ ，且 $|\rho_k| \le 1$　，$\forall k$
- 对称性　$\rho_k = \rho_{-k}$
- 非负定性　Γ_m为非负定阵，$\forall$正整数m

$$\Gamma_m = \begin{pmatrix} \rho_0 & \rho_1 & \cdots & \rho_{m-1} \\ \rho_1 & \rho_0 & \cdots & \rho_{m-2} \\ \vdots & \vdots & \cdots & \vdots \\ \rho_{m-1} & \rho_{m-2} & \cdots & \rho_0 \end{pmatrix}$$

- 非唯一性

一个平稳时间序列一定唯一决定了它的自相关函数，但一个自相关函数未必唯一对应着一个平稳时间序列。

平稳性的意义

- 在平稳序列场合，序列的均值等于常数，这意味着原本含有可列多个随机变量的均值序列变成了只含有一个变量的常数序列。

$$\{\mu_t, t \in T\} \quad \Rightarrow \quad \{\mu, t \in T\}$$

- 原本每个随机变量的均值（方差，自相关系数）只能依靠唯一的一个样本观察值去估计，现在由于平稳性，每一个统计量都将拥有大量的样本观察值。

- 这极大地减少了随机变量的个数，并增加了待估变量的样本容量。极大地简化了时序分析的难度，同时也提高了对特征统计量的估计精度

检验方法

方法一：图检验（本章介绍）

- 时序图检验
- 自相关图检验

方法二：构造检验统计量进行假设检验（第四章介绍）

- 单位根检验

平稳性的检验（图检验方法）

1、时序图检验（目测法）

– 根据平稳时间序列均值、方差为常数的性质，平稳序列的时序图应该显示出该序列始终在一个常数值附近随机波动，而且波动的范围有界、无明显趋势及周期特征。

回顾宽平稳的定义：

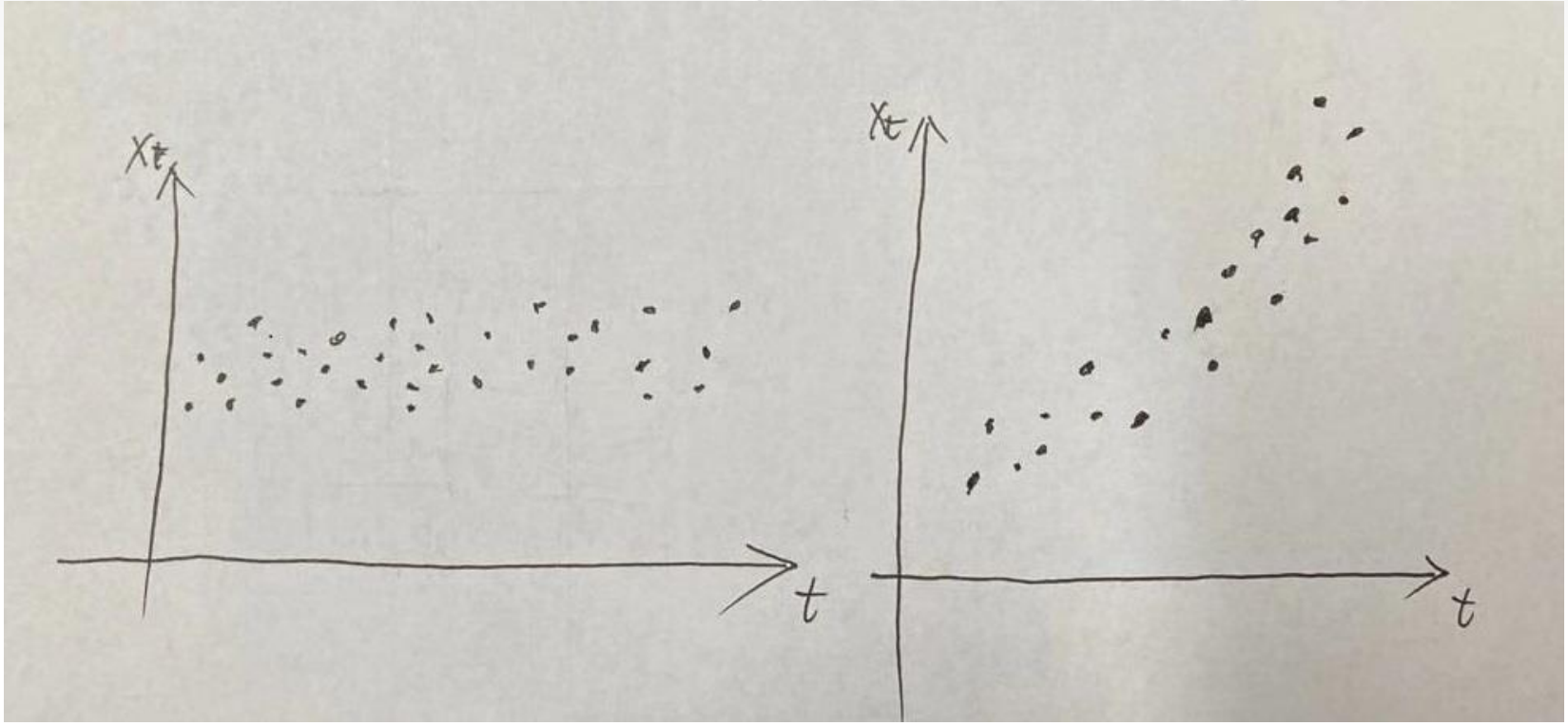
1) $EX_t^2 < \infty, \forall t \in T$

2) $EX_t = \mu, \mu$为常数，$\forall t \in T$

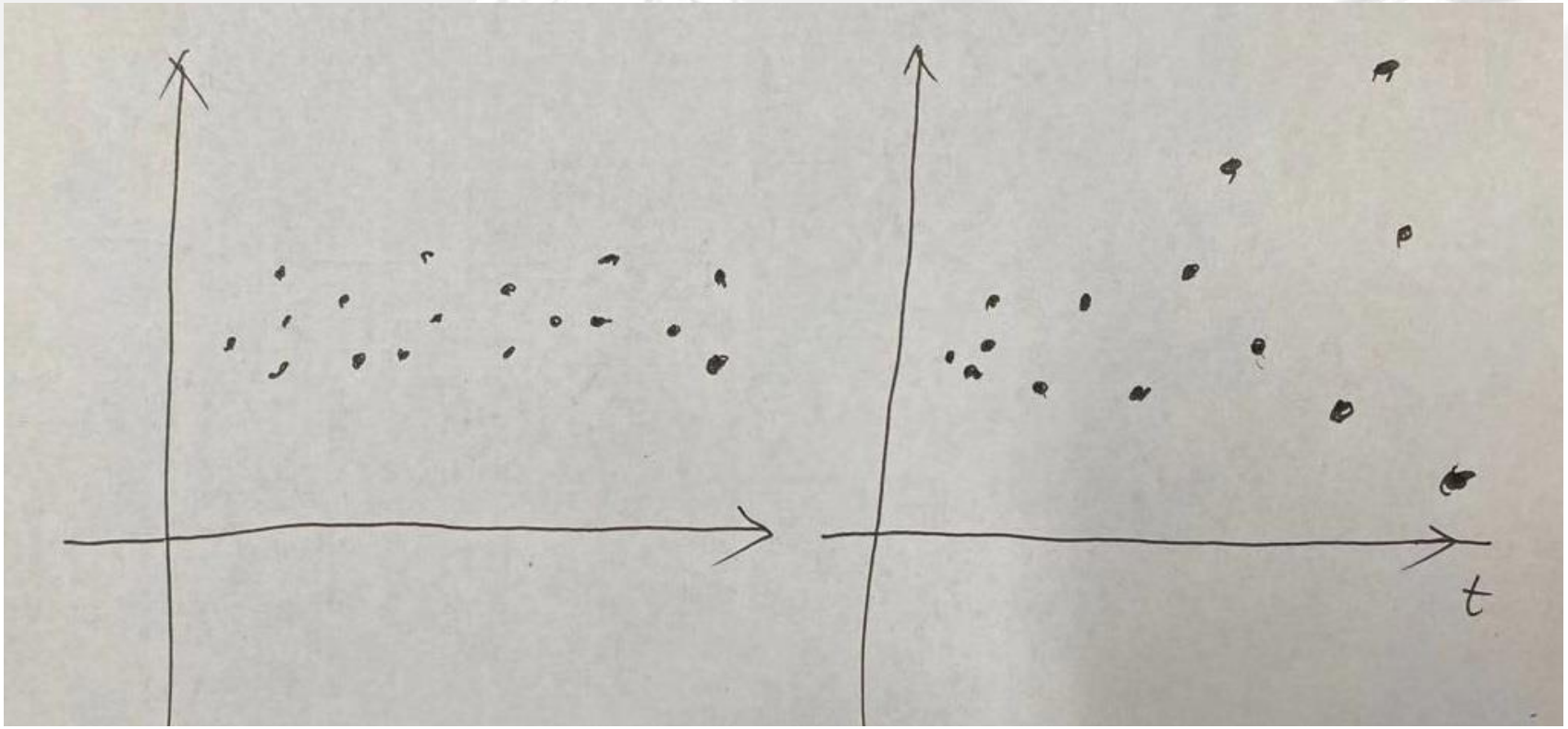
3) $\gamma(t,s) = \gamma(k, k+s-t)$，$\forall t, s, k$且$k+s-t \in T$

判断一个序列是不是平稳序列有三个评判标准：

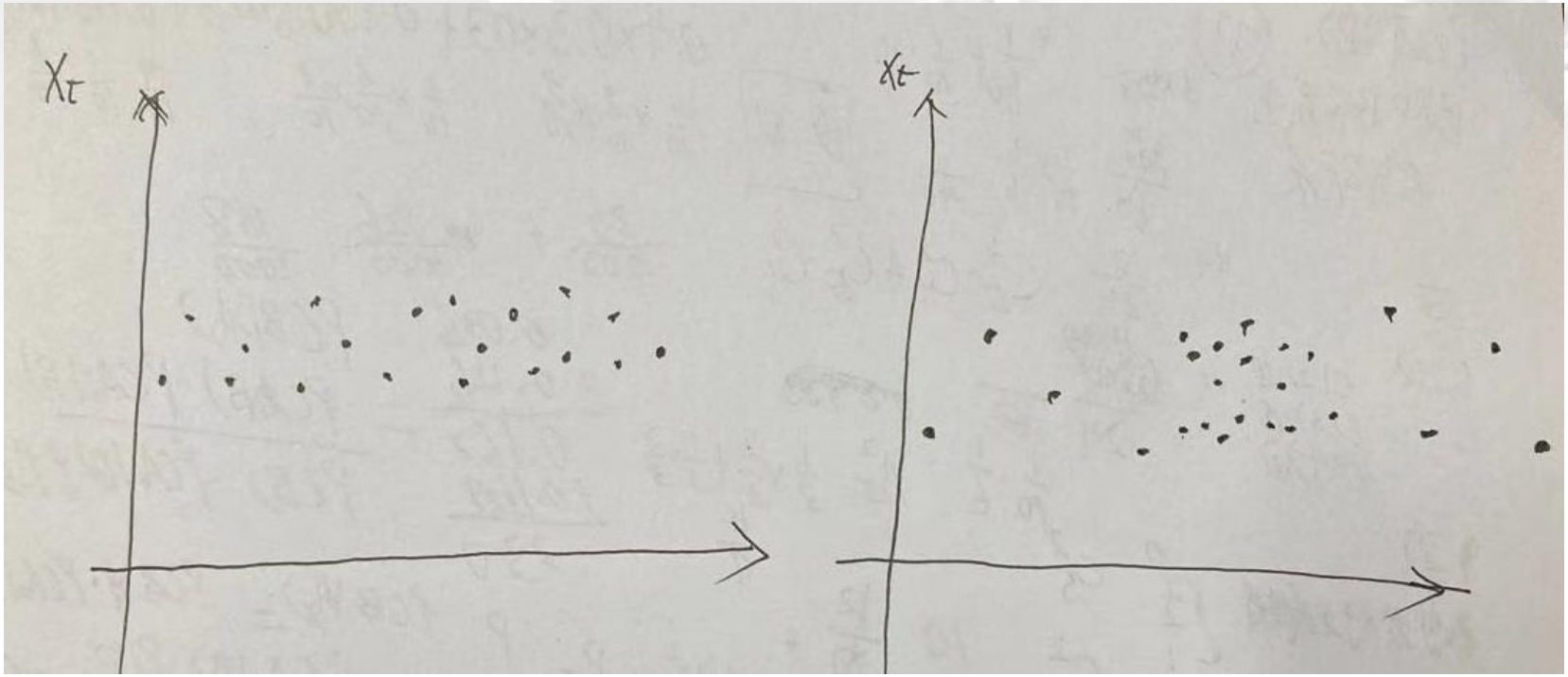
1、均值，是与时间t 无关的常数。下图（左）满足平稳序列的条件，下图（右）很明显具有时间依赖。



2、方差，是与时间t 无关的常数。这个特性叫做方差齐性。



3、协方差，只与时期间隔k有关，与时间t 无关的常数。



思考：若序列有明显的周期性，是否平稳？

例2-1

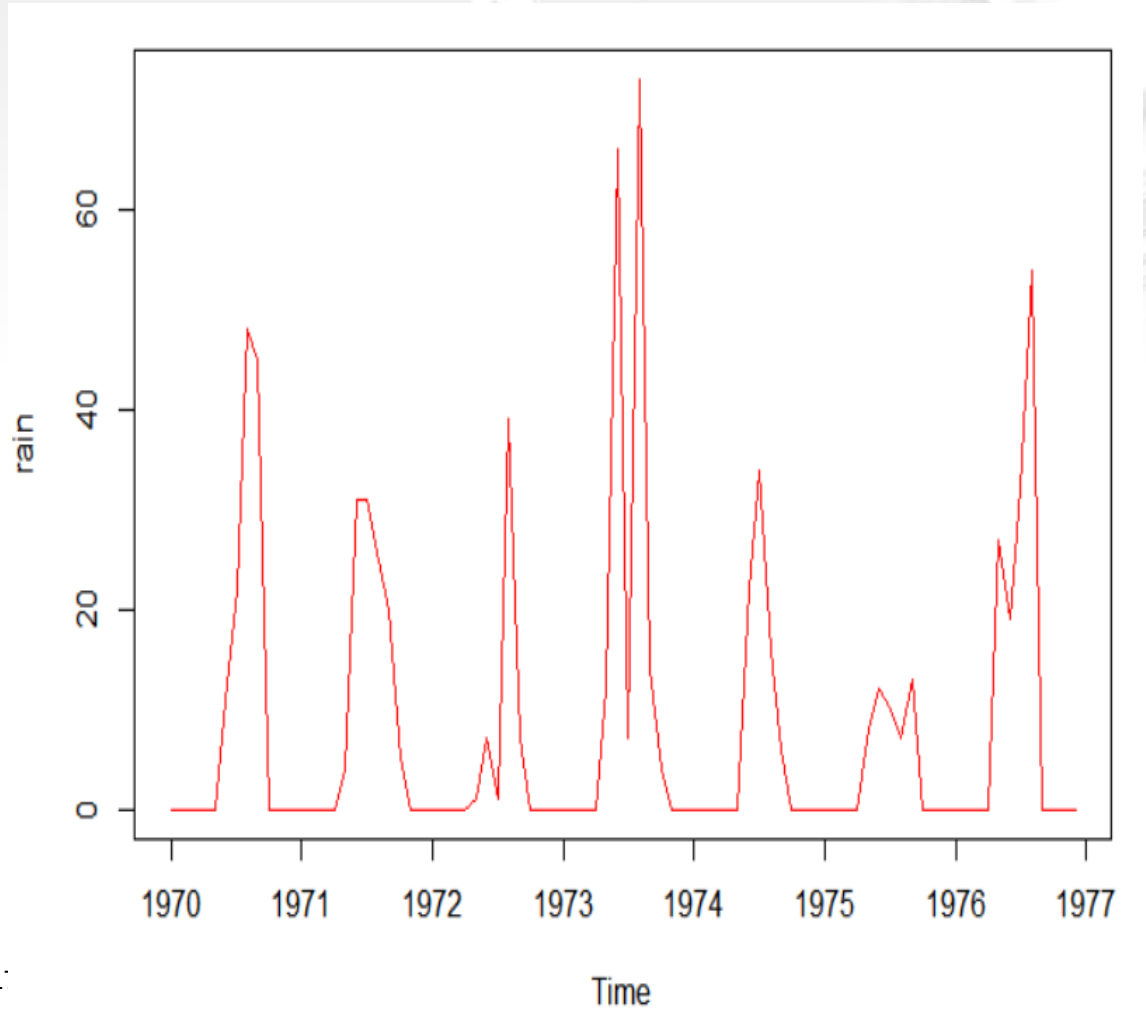
- 绘制1978-2012年我国第三产业占国内生产总值的比例序列的时序图，根据时序图判断该序列的平稳性。

该序列时序图清晰显示：

序列有明显的递增趋势特征，所以是非平稳序列。

例2-2

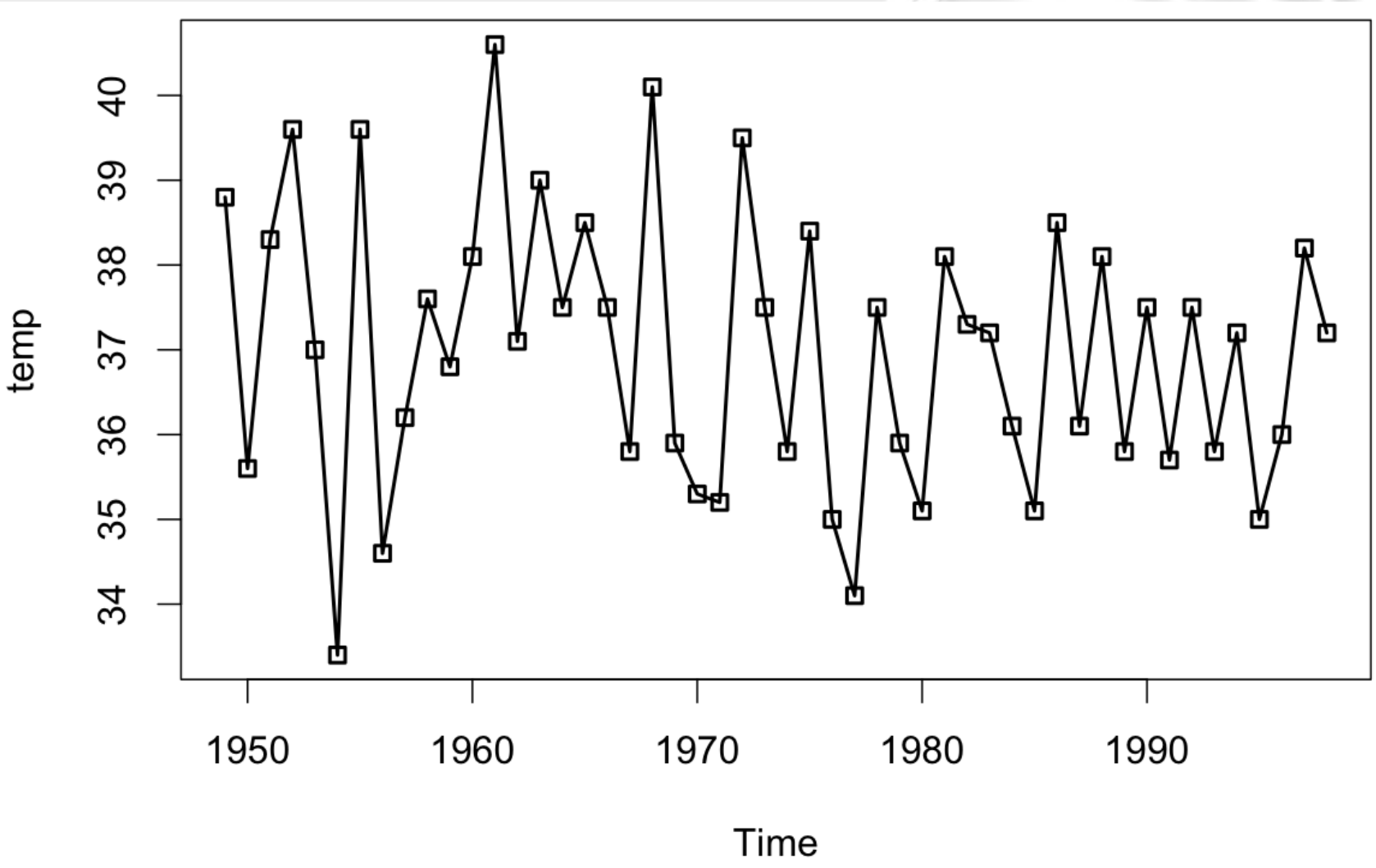
- 绘制 1970—1976 年加拿大 Coppermine地区月度降雨量序列的时序图，根据时序图判断该序列平稳性



该序列时序图清晰显示：

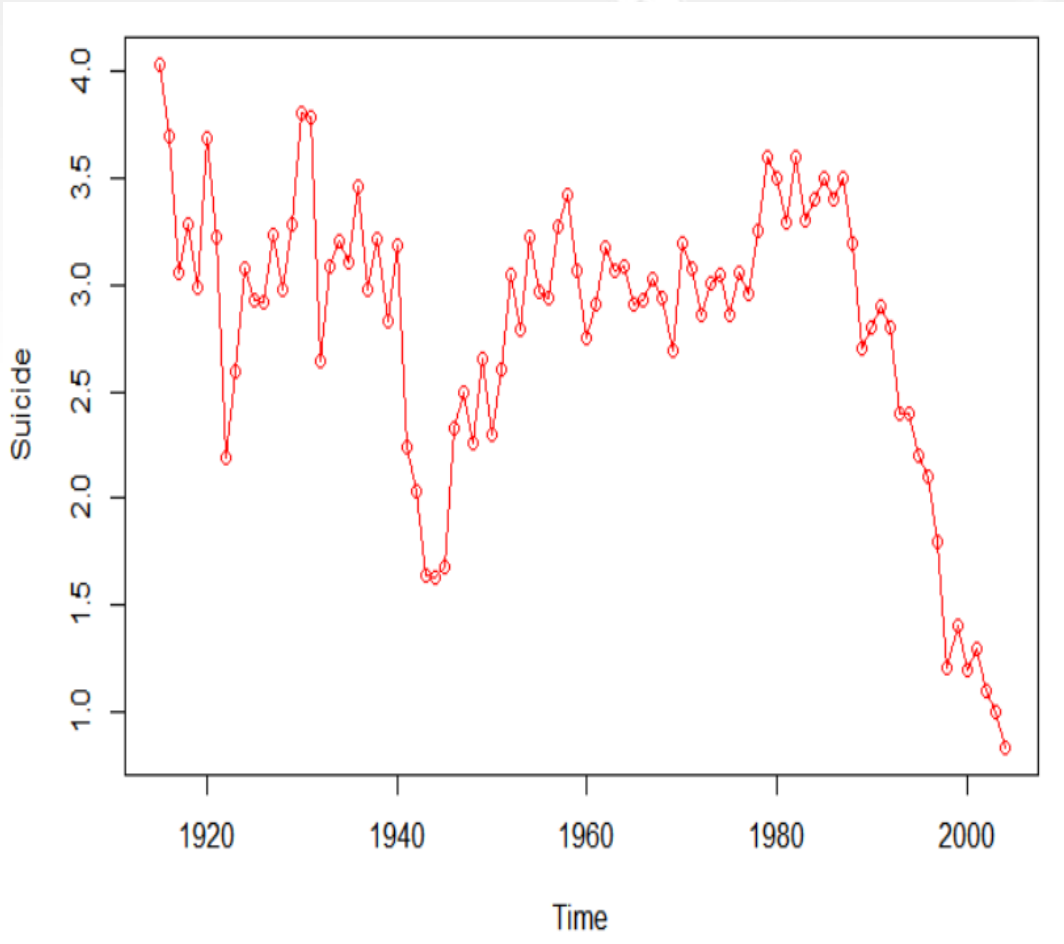
序列有明显的周期特征，所以是非平稳序列。

北京市每年最高气温时序图



例2-3

- 绘制1915-2004年澳大利亚自杀率序列（每10万人自杀人口数）的时序图，根据时序图判断该序列的平稳性。



该序列时序图显示：

- 从1915年开始澳大利亚每年的自杀率长期围绕在10万分之3附近波动，而且波动范围长期在10万分之2至10万分之4之间，这呈现出平稳序列的特征。但是看序列的最后20年的波动，自杀率又是一路递减，这是有趋势吗？如果是趋势，这就是非平稳特征。
- 这时，通过时序图判断序列的平稳性有点困难。
- 这时，可以借助序列自相关图的性质进一步辅助识别。

2、自相关图检验法。

由宽平稳定义，可如下构造序列的k阶自协方差函数的估计：

$$\hat{\gamma}(k)=\frac{\sum_{t=1}^{n-k}(X_t-\bar{X})(X_{t+k}-\bar{X})}{n-k},\qquad 0<k<n$$

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

k阶自相关系数函数的估计：

$$\hat{\rho}_k=\frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}\approx\frac{\sum_{t=1}^{n-k}(X_t-\bar{X})(X_{t+k}-\bar{X})}{\sum_{t=1}^{n}(X_t-\bar{X})^2},\qquad 0<k\ll n$$

平稳序列通常具有短期相关性。该性质用自相关系数来描述就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向零。

R语言中运用acf函数来绘制序列的自相关图:

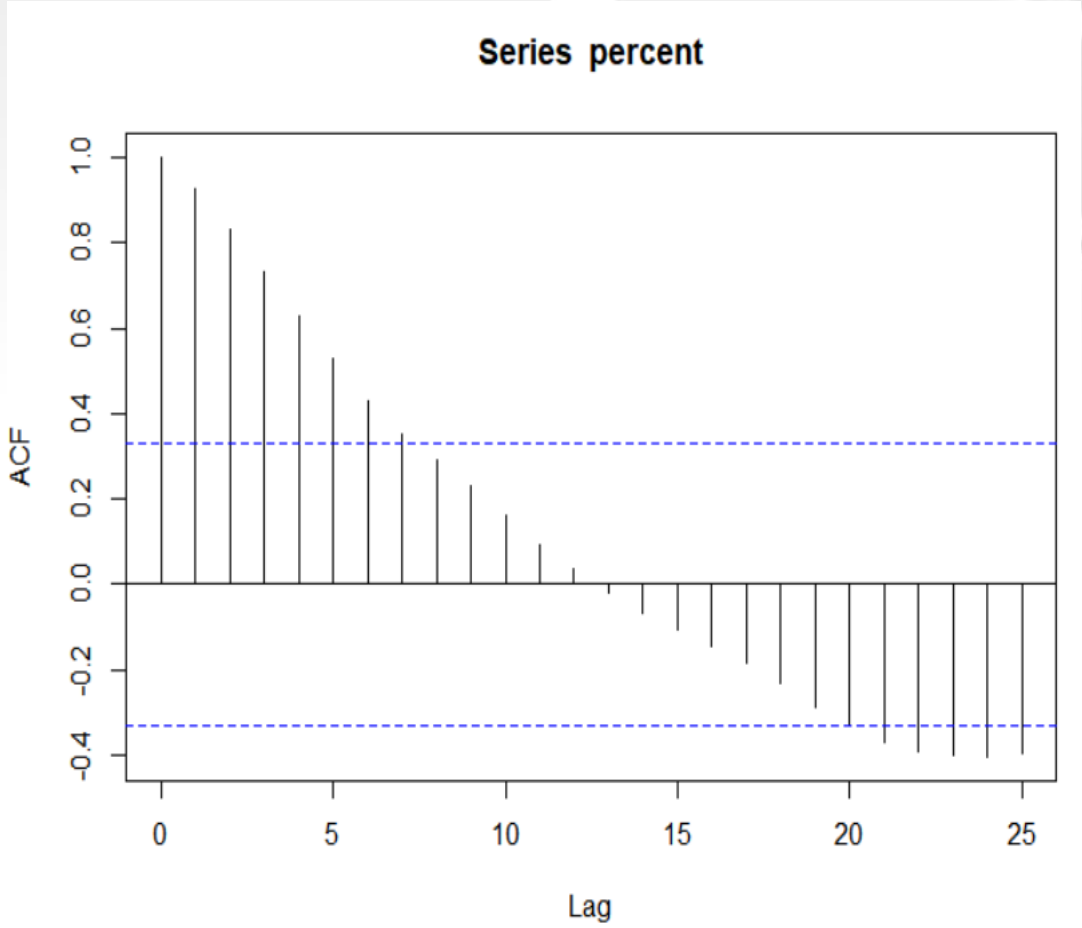
acf(x, lag=)

说明：x:变量名。

lag：延迟阶数。

例2.1续

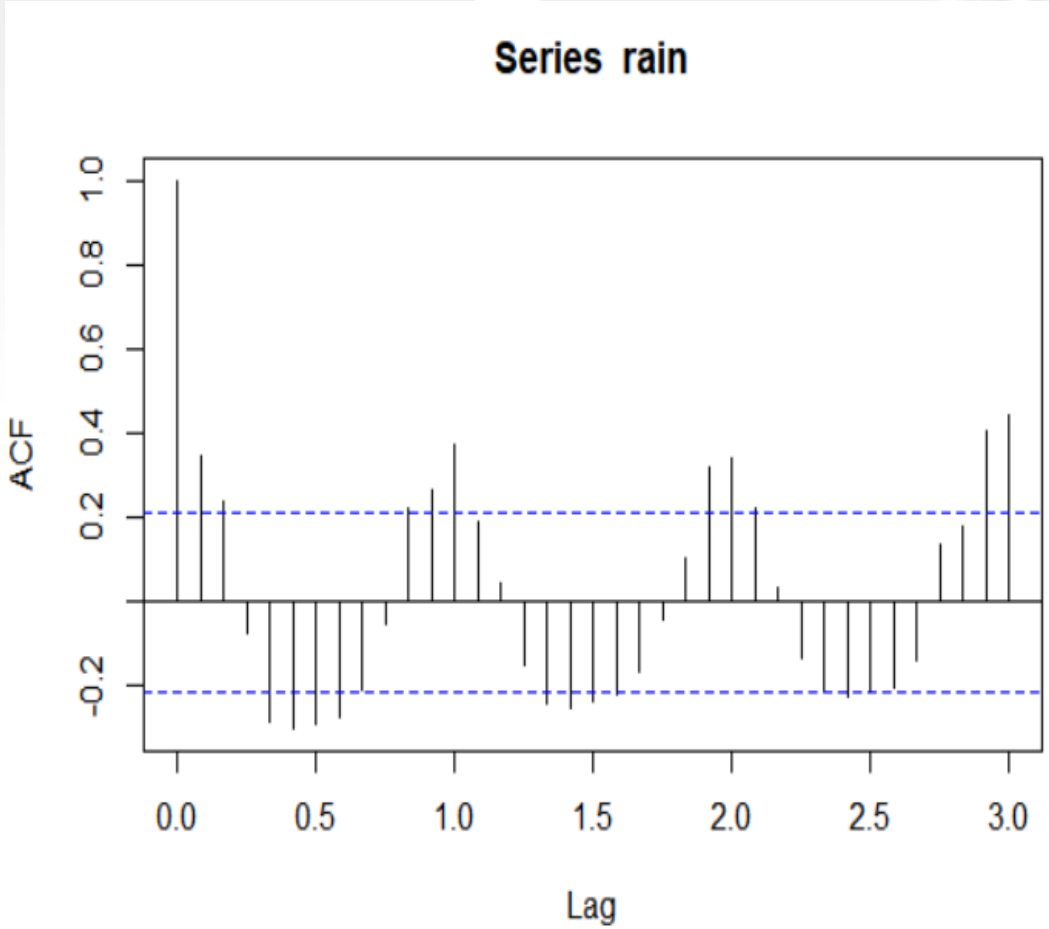
- 绘制1978-2012年我国第三产业占国内生产总值的比例序列的自相关图



- 该序列自相关图呈现出明显的三角对称性，这是有趋势的非平稳序列常见的自相关图特征.
- 根据该序列自相关图我们可以认为该序列非平稳，且可能具有长期趋势这和该序列时序图呈现的单调递增性是一致的.

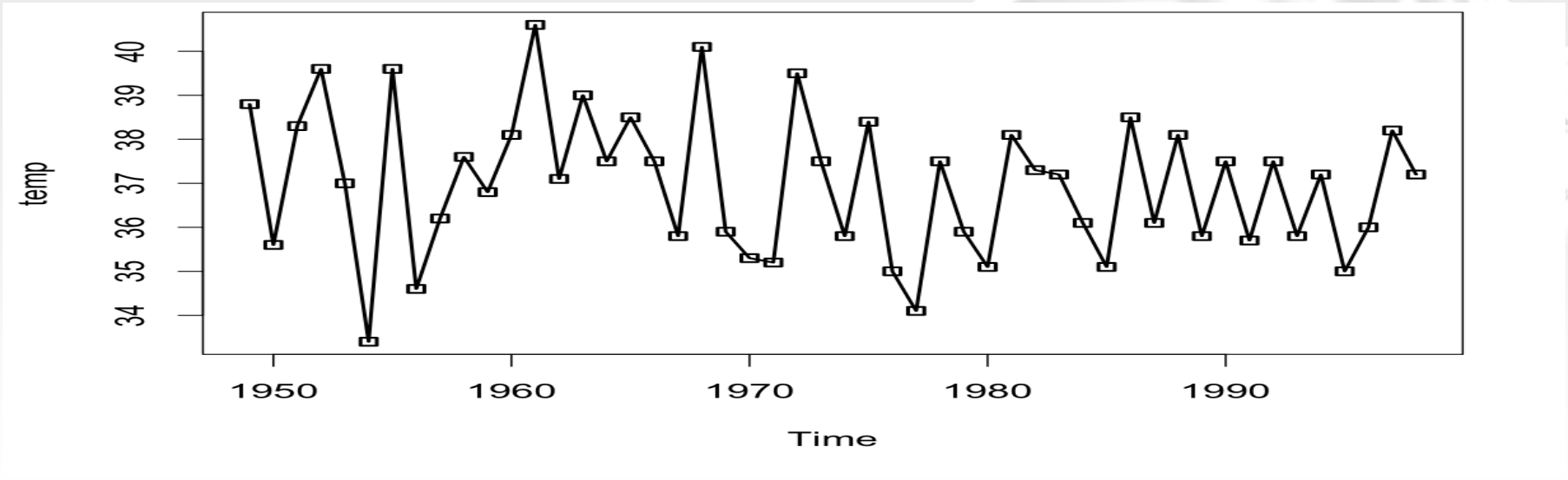
例2.2续

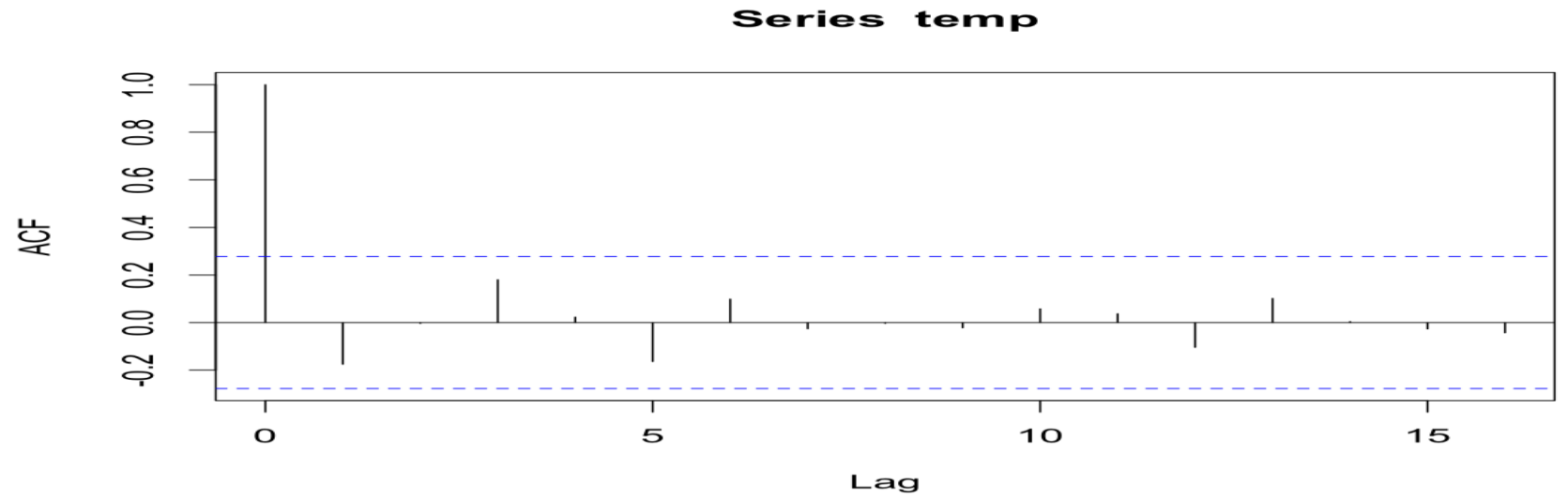
- 绘制 1970—1976 年加拿大 Coppermine地区月度降雨量序列的自相关图



- 该序列自相关图呈现明显的三角函数(正弦或余弦)波动规律.这是具有周期性变化的非平稳序列的一种典型的自相关图特征，而且这种周期性几乎不衰减，直到第3个周期(延迟了36阶)，自相关系数依然落人两倍标准差之外.
- 根据自相关图的长期相关性和余弦变化特征，我们可以认为该序列非平稳且具有稳定的周期变化规律.这和该序列时序图(图2-2)呈现的季节性特征是一致的.

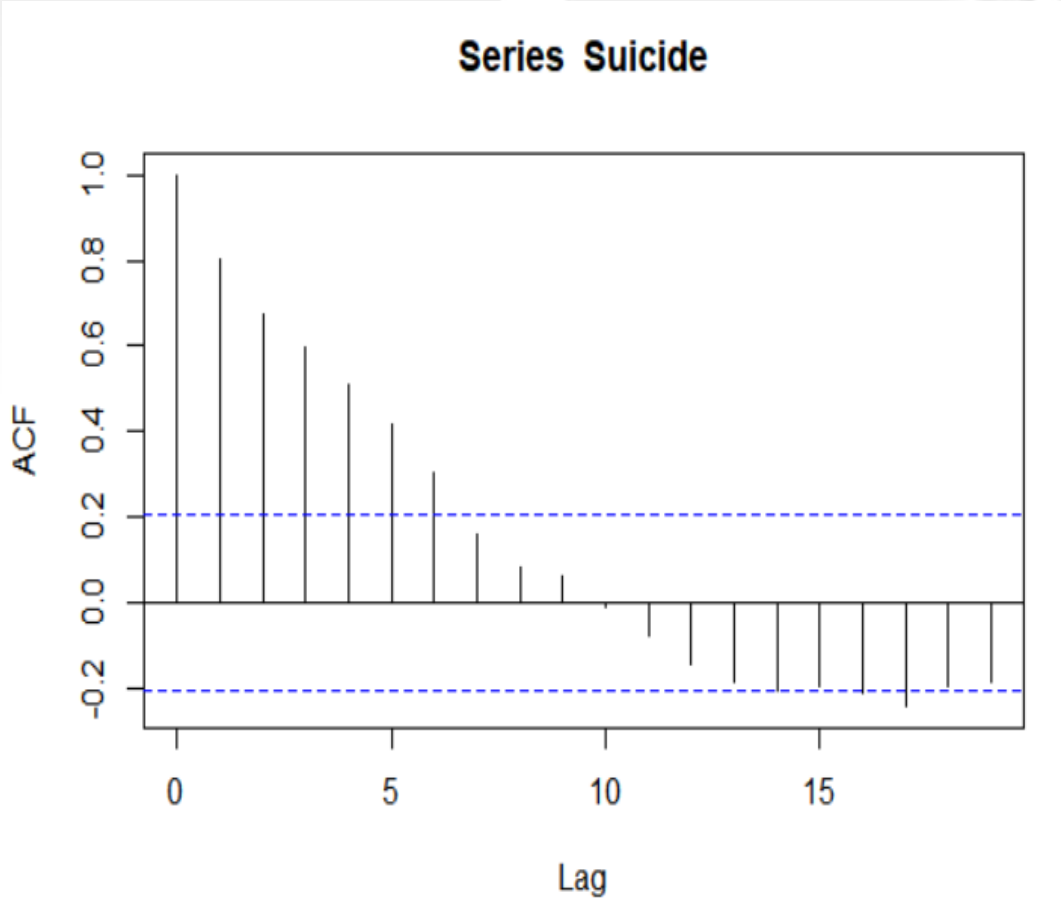
1949-1998年北京市每年最高气温序列。





例2.3续

- 绘制1915-2004年澳大利亚自杀率序列（每10万人自杀人口数）的自相关图



- 该序列自相关系数延迟15阶之后依然显著非零，这说明该序列自相关系数具有长期相关性，而且自相关图呈现出明显的倒三角特征，这是具有单调趋势的非平稳序列的典型特征.
- 根据自相关图特征，我们可以认为该序列非平稳，且具有长期趋势.
- 在该序列时序图难以判别平稳性的情况下，自相关图可以帮助我们进一步识别序列的平稳性.

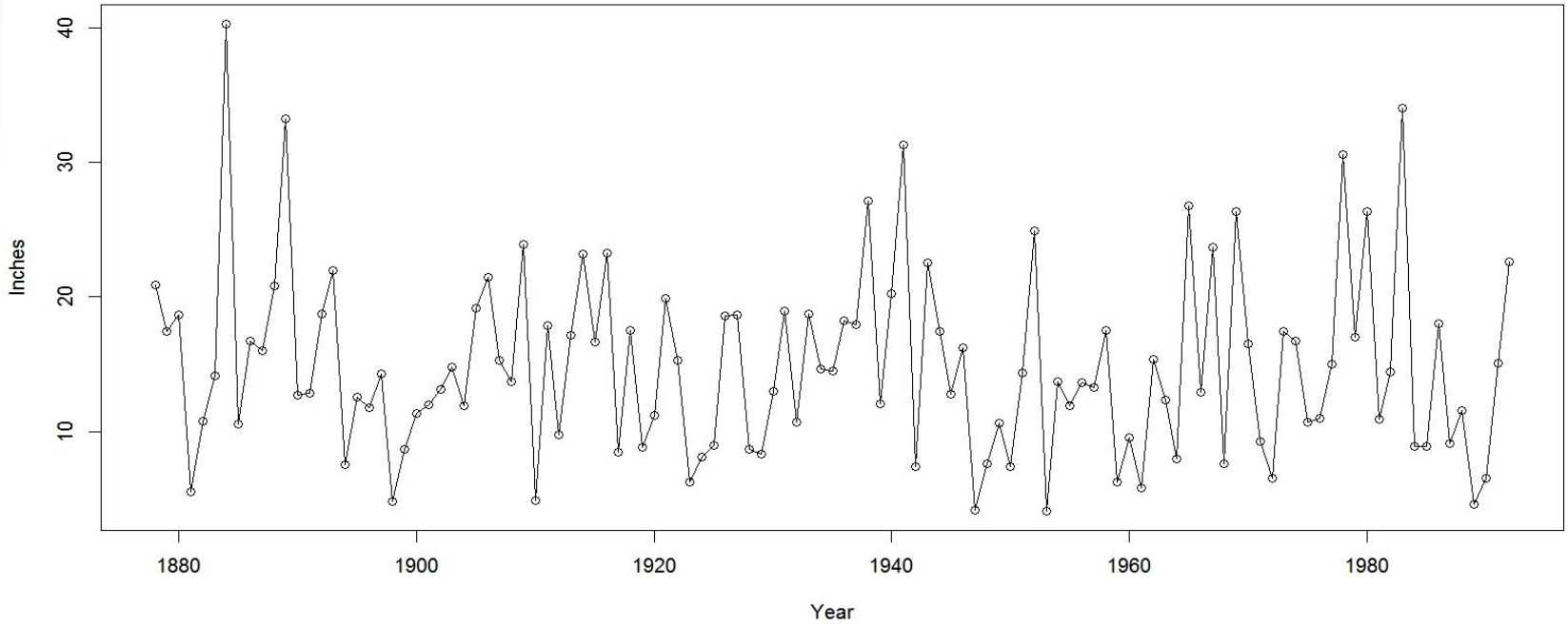
本章结构

1. 平稳性检验

2. 纯随机性检验

2.2 纯随机性检验

回顾：洛杉矶100多年的降水量时间序列图，这个序列没有什么相关性，也就没有什么分析的价值。



纯随机序列的定义

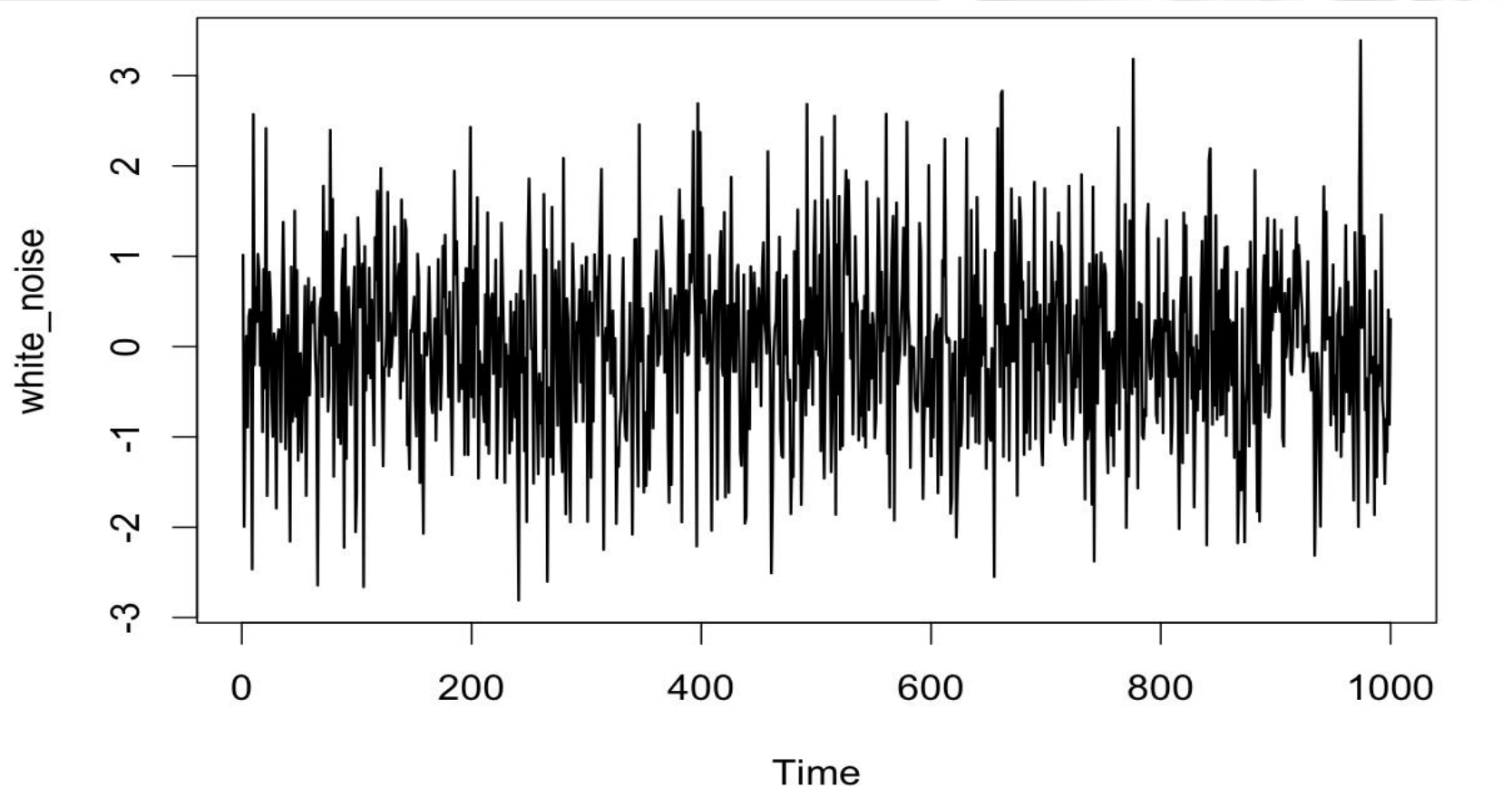
纯随机序列也称为白噪声序列，它满足如下两条性质

$$(1) EX_t = \mu, \forall t \in T$$

$$(2) \gamma(t,s) = \begin{cases} \sigma^2, t = s \\ 0, t \neq s \end{cases}, \forall t, s \in T$$

思考：白噪声序列是否是平稳序列？

例2.4：标准正态白噪声序列时序图



正态随机数命令：rnorm(n=, mean=, sd=)

白噪声序列的性质

- 纯随机性和方差齐性
  - 各序列值之间没有任何相关关系，方差恒等

$$\gamma(k) = 0, \quad \forall k \neq 0; \qquad \gamma(0) = \sigma^2$$

- 不再含有可提取的信息
  - 线性回归中，我们要求误差项为白噪声，残差分析就是分析其是否为白噪声。

$$Y_t = \beta X_t + \varepsilon_t$$

纯随机性检验

直观上，两个序列相隔越远，相关性就越小。所以，一般我们只检验前m阶的相关系数是否为零。

- 原假设：延迟期数小于或等于 m 期的序列值之间相互独立

$$H_0: \rho_1 = \rho_2 = \cdots = \rho_m = 0, \forall m \geq 1$$

- 备择假设：延迟期数小于或等于 m 期的序列值之间有相关性

$$H_1: \text{至少存在某个} \rho_k \neq 0, \forall m \geq 1, \ k \leq m$$

检验量

- Q统计量：

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

- LB统计量：

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

提示：抽样分布的推导，Barlett证明了：

$$\hat{\rho}_k \dot{\sim} N(0, \frac{1}{n}) \quad , \forall k \neq 0$$

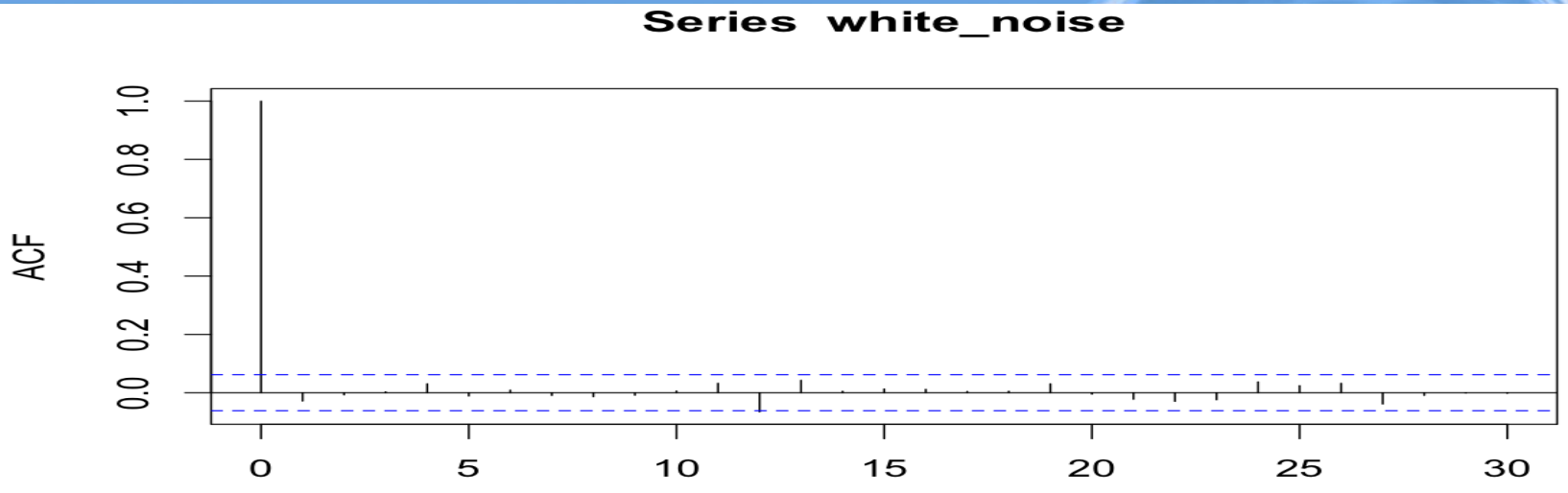
判别原则

- 拒绝域
  - 当检验统计量大于 $\chi^2_{1-\alpha}(m)$ 分位点，或该统计量的 P 值小于 α 时，则可以以 $1-\alpha$ 的置信水平拒绝原假设，认为该序列为非白噪声序列。否则，接受原假设。

R 语言运用Box.test函数进行随机性检验。

```
Box.test(x, type= , lag=)
式中:
-x: 变量名;
-type: 检验统计量类型.
(1) type=''Box-Pierce'', 输出白噪声检验的 Q 统计量. 该统计量为系统默认输出结果.
(2) type=''Ljung-Box'', 输出白噪声检验的 LB 统计量.
-lag: 延迟阶数. lag=n 表示输出滞后 n 阶的白噪声检验统计量. 忽略该选项时, 默认输出滞后 1 阶的检验统计量结果.
```

例2.4：标准正态白噪声序列纯随机性检验



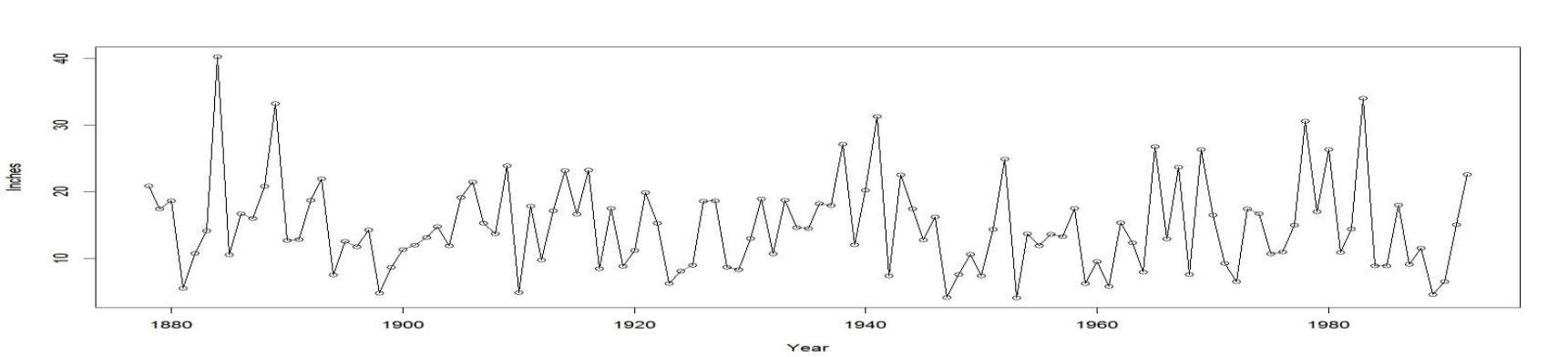

```
> Box.test(white_noise,lag=6)

        Box-Pierce test

data:  white_noise
X-squared=5.0327,  df=6,  p-value=0.5396

> Box.test(white_noise,lag=12)

        Box-Pierce test

data:  white_noise
X-squared=9.811,  df=12,  p-value=0.6325
```

检验结果

延迟	Q_{LB} 统计量检验	
	Q_{LB} 统计量值	P值
延迟6期	5.03	0.5396
延迟12期	9.81	0.6325

由于 P 值显著大于显著性水平 $\alpha = 0.05$，所以该序列不能拒绝纯随机的原假设。

回顾：洛杉矶降水量时间序列图。

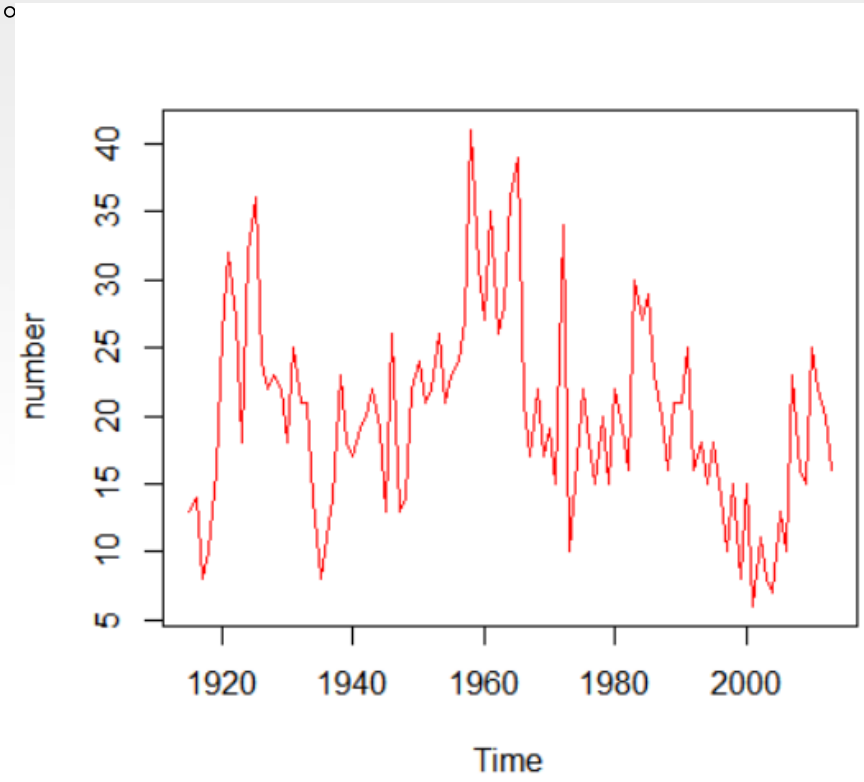


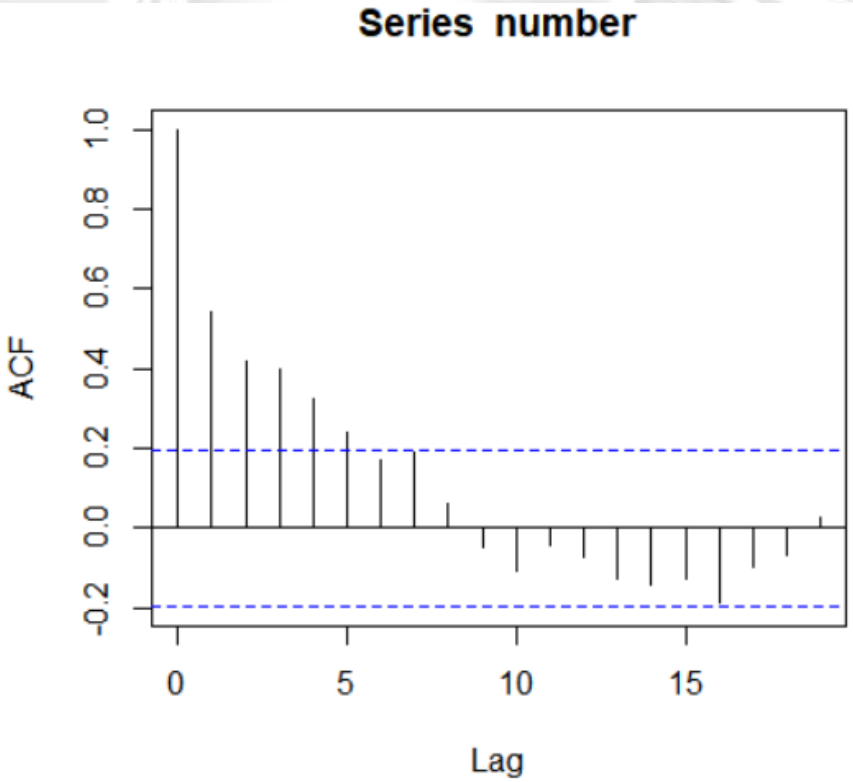
延迟阶数	LB统计量检验	
	LB检验量的值	P值
6	5.56	0.4744
12	13.55	0.3304

由于 P 值显著大于显著性水平 $\alpha = 0.05$ ，所以该序列不能拒绝纯随机的原假设。

例2-5

- 对1900—1998年全球7级以上地震发生次数序列进行平稳性和纯随机性检验





- 时序图显示该序列没有明显的趋势和周期.
- 自相关图显示，除了延迟1一5阶的自相关系数在两倍标准差之外，其他自相关系数均在两倍标准差之内.我们可以认为该序列具有短期相关性.
- 因此，我们可以判断该序列为平稳序列.

例2-5

- 对1900—1998年全球7级以上地震发生次数序列进行纯随机性检验。

```
        Box-Ljung test

data:  number
X-squared = 84.734, df = 6, p-value = 3.331e-16
```

  - 检验结果显示，延迟6阶的LB统计量的P值显著小于显著性水平0.05，所以拒绝原假设，认为该序列为非白噪声序列。

- 本利通过图检验和纯随机性检验，我们可以认为全球每年发生7.0+级地震次数序列是平稳非白噪声序列。在统计时序分析领域，平稳非白噪声序列被认为是值得分析且最容易分析的一种序列。下一章我们将详细介绍对平稳非白噪声序列的建模及预测方法。