时间序列分析（初级）

时间序列预处理

陈垚翰

安徽大学 大数据与统计学院

安徽大學
Anhui University

本章结构

- 基本统计学概念
- 平稳性
- 随机性

基本统计学概念

概率分布

- 时间序列概率分布族

$$\left\{F_{t_1,t_2,\cdots,t_m}\left(x_1,x_2,\cdots,x_m\right)\right\}$$

$$\forall m \in (1,2,\cdots,m), \forall t_1,t_2,\cdots,t_m \in [0,T]$$

特征统计量

- 均值

$$\mu_t = \mathbb{E}(X_t) = \int_{-\infty}^{\infty} x dF_t(x)$$

- 方差

$$\mathrm{Var}(X_t) = \mathbb{E}\left[(X_t - \mu_t)^2\right] = \int_{-\infty}^{\infty} (x - \mu_t)^2 \, dF_t(x)$$

- 自协方差

$$\gamma(t,s) = \mathrm{Cov}\left(X_t, X_s\right) = \mathbb{E}\left[\left(X_t - \mu_t\right)\left(X_s - \mu_s\right)\right]$$

- 自相关系数

$$\rho(t,s) = \frac{\gamma(t,s)}{\sqrt{\mathrm{Var}\left(X_t\right) \cdot \mathrm{Var}\left(X_s\right)}}$$

平稳性

平稳时间序列的统计学定义

- 满足如下条件的序列称为 **严平稳** 序列
  - $\forall$ 正整数 $m, \forall t_1, t_2, \cdots, t_m \in T, \forall$ 正整数 τ, 有
    $$F_{t_1, t_2 \cdots t_m}(x_1, x_2, \cdots, x_m) = F_{t_{1+\tau}, t_{2+\tau} \cdots t_{m+\tau}}(x_1, x_2, \cdots, x_m)$$

平稳时间序列的统计学定义

- 满足如下条件的序列称为 **严平稳** 序列
  - $\forall$ 正整数 $m, \forall t_1, t_2, \cdots, t_m \in T, \forall$ 正整数 τ, 有
    $F_{t_1, t_2 \cdots t_m}(x_1, x_2, \cdots, x_m) = F_{t_{1+\tau}, t_{2+\tau} \cdots t_{m+\tau}}(x_1, x_2, \cdots, x_m)$
- 满足如下条件的序列称为 **宽平稳** 序列

(1) $\mathbb{E}\left(X_t^2\right) < \infty, \forall t \in T$
(2) $\mathbb{E}(X_t) = \mu, \mu$ 为常数, $\forall t \in [0, T]$
(3) $\gamma(t, s) = \gamma(k, k + s - t), \forall t, s, k$ 且 $k + s - t \in [0, T]$

严平稳与宽平稳的关系

- 严平稳 $\nRightarrow$ 宽平稳：服从 Cauchy 分布的严平稳序列不是宽平稳序列
- 宽平稳 $\nRightarrow$ 严平稳
- 当序列服从正态分布时，宽平稳 $\Rightarrow$ 严平稳

$$f_{t_1,t_2,\cdots,t_n}\left(\widetilde{X}_n\right)=(2\pi)^{-\frac{n}{2}}\left|\mathbf{\Gamma}_n\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left(\widetilde{X}_n-\widetilde{\mu}_n\right)'\mathbf{\Gamma}_n^{-1}\left(\widetilde{X}_n-\widetilde{\mu}_n\right)\right]$$

其中，

$$
\begin{aligned}
\widetilde{X}_n &= (X_1, X_2, \cdots, X_n)^\top \\
\widetilde{\mu}_n &= (\mathbb{E}(X_1), \mathbb{E}(X_2), \cdots, \mathbb{E}(X_n))^\top \\
\mathbf{\Gamma}_n &= \begin{pmatrix}
\gamma(t_1, t_1) & \gamma(t_1, t_2) & \cdots & \gamma(t_1, t_n) \\
\gamma(t_2, t_1) & \gamma(t_2, t_2) & \cdots & \gamma(t_2, t_n) \\
\vdots & \vdots & & \vdots \\
\gamma(t_n, t_1) & \gamma(t_n, t_2) & \cdots & \gamma(t_n, t_n)
\end{pmatrix}
\end{aligned}
$$

平稳时间序列的统计性质

- 常数均值

$$\mathbb{E}(X_t) = \mu, \quad \forall t \in [0, T]$$

- 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

$$\gamma(t, s) = \gamma(k, k + s - t), \forall t, s, k \text{ 且 } k + s - t \in [0, T]$$

所以

$$\gamma(s - t) \doteq \gamma(t, s), \forall t, s \in [0, T]$$

- 延迟 k 自协方差函数

$$\gamma(k) = \gamma(t, t + k) \quad \text{且} \quad \gamma(0) = \gamma(t, t) = \mathrm{Var}(X_t)$$

- 延迟 k 自相关函数

$$\rho_k = \frac{\gamma(t, t+k)}{\sqrt{\mathrm{Var}\,(X_t) \cdot \mathrm{Var}\,(X_{t+k})}} = \frac{\gamma(k)}{\gamma(0)}$$

自相关系数的性质

- 规范性

$$\rho_0 = 1 \text{ 且 } |\rho_k| \leqslant 1, \forall k$$

- 对称性

$$\rho_k = \rho_{-k}$$

- 非负定性

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

是非负定矩阵

平稳时间序列的意义

- 时间序列结构的特殊性
  - 可列多个随机变量，而每个变量只有一个样本观察值

平稳时间序列的意义

- 时间序列结构的特殊性
  - 可列多个随机变量，而每个变量只有一个样本观察值
- 平稳性的意义
  - 减少了参数空间的维度，增加了样本容量
  - 简化了时序分析的难度

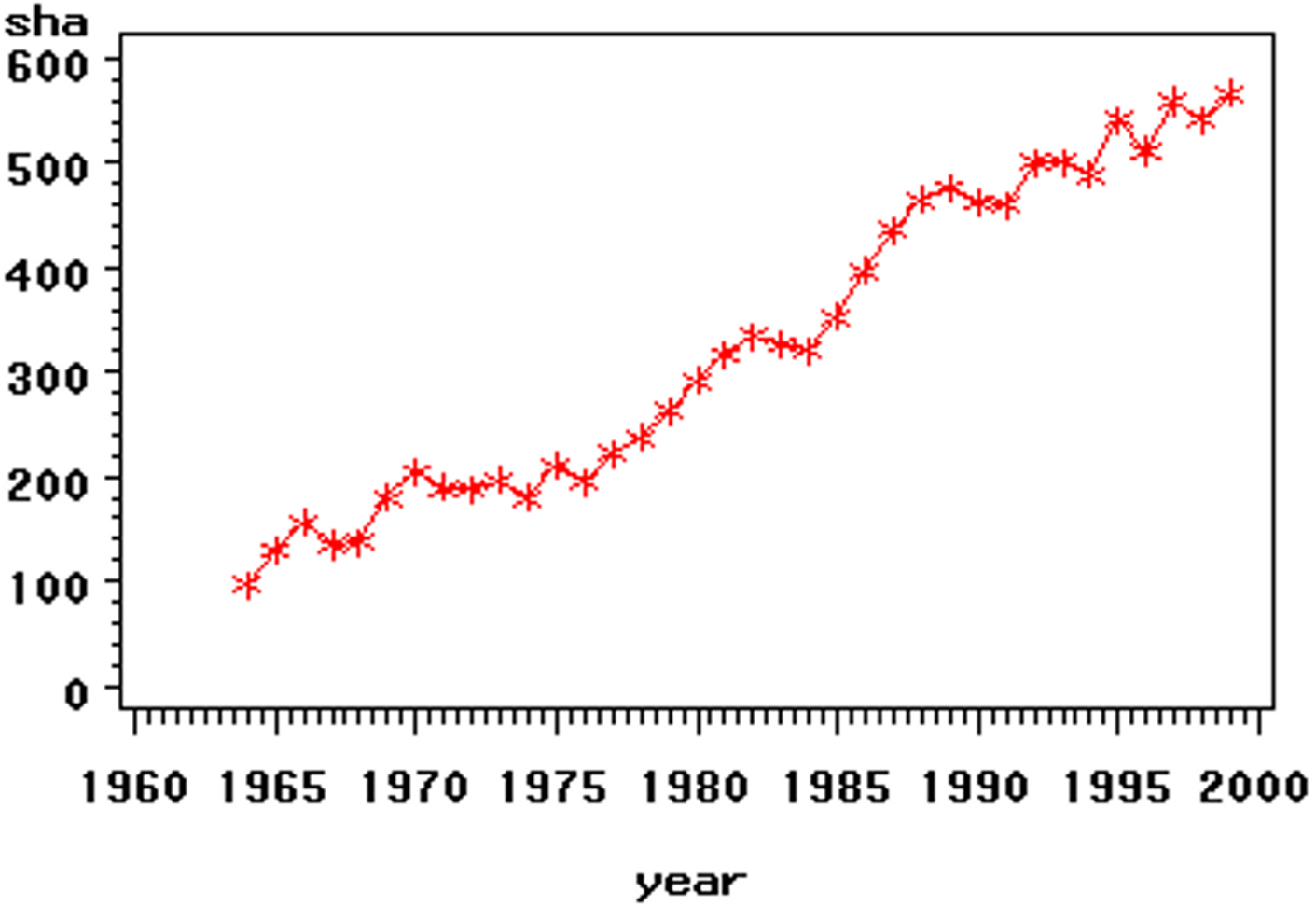
平稳性的检验（图检验法）

- 时序图检验
  - 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

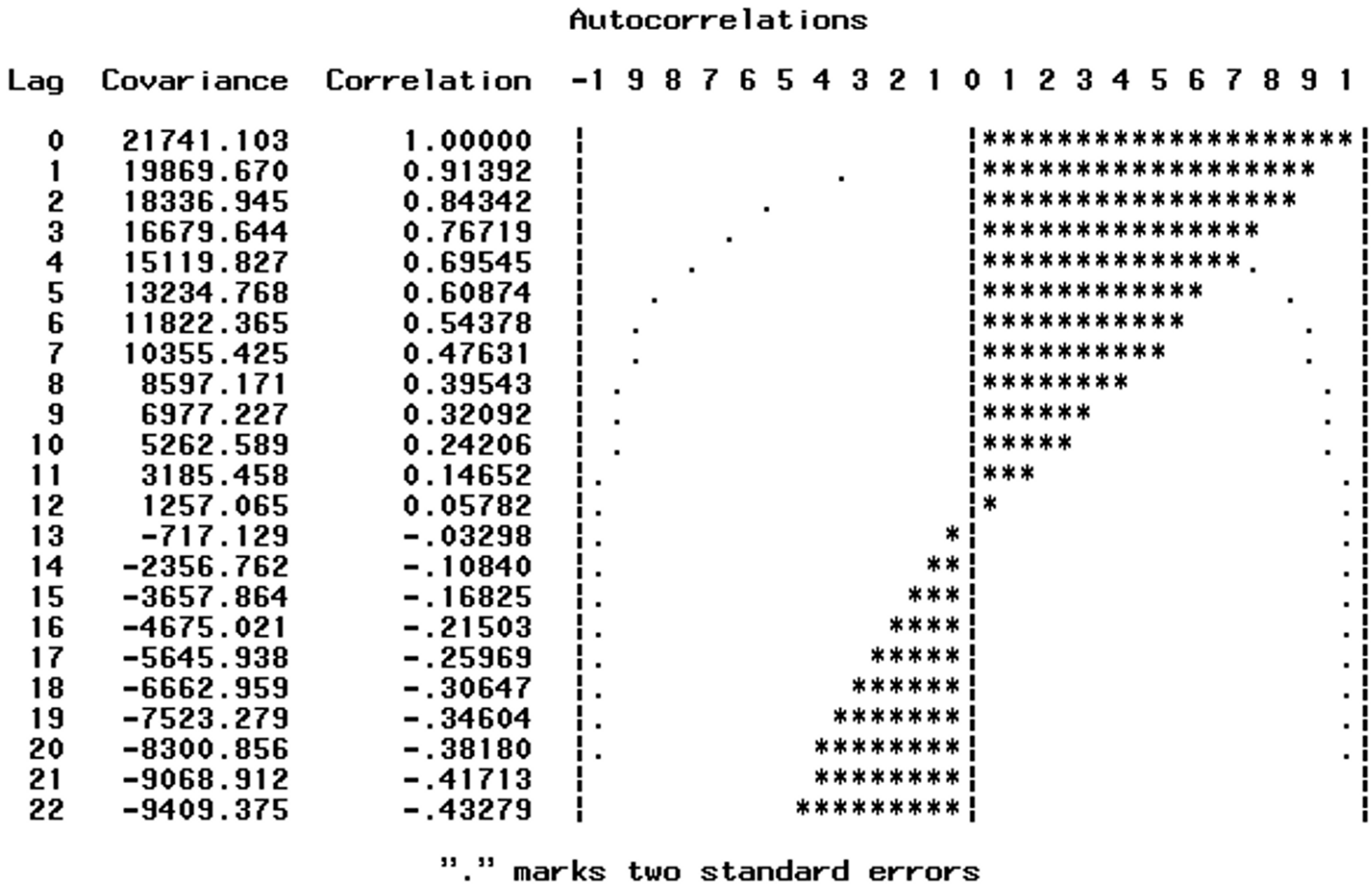
平稳性的检验（图检验法）

- 时序图检验
  - 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
- 自相关图检验
  - 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

1964年——1999年中国纱年产量（时序图）



1964年——1999年中国纱年产量（自相关图）

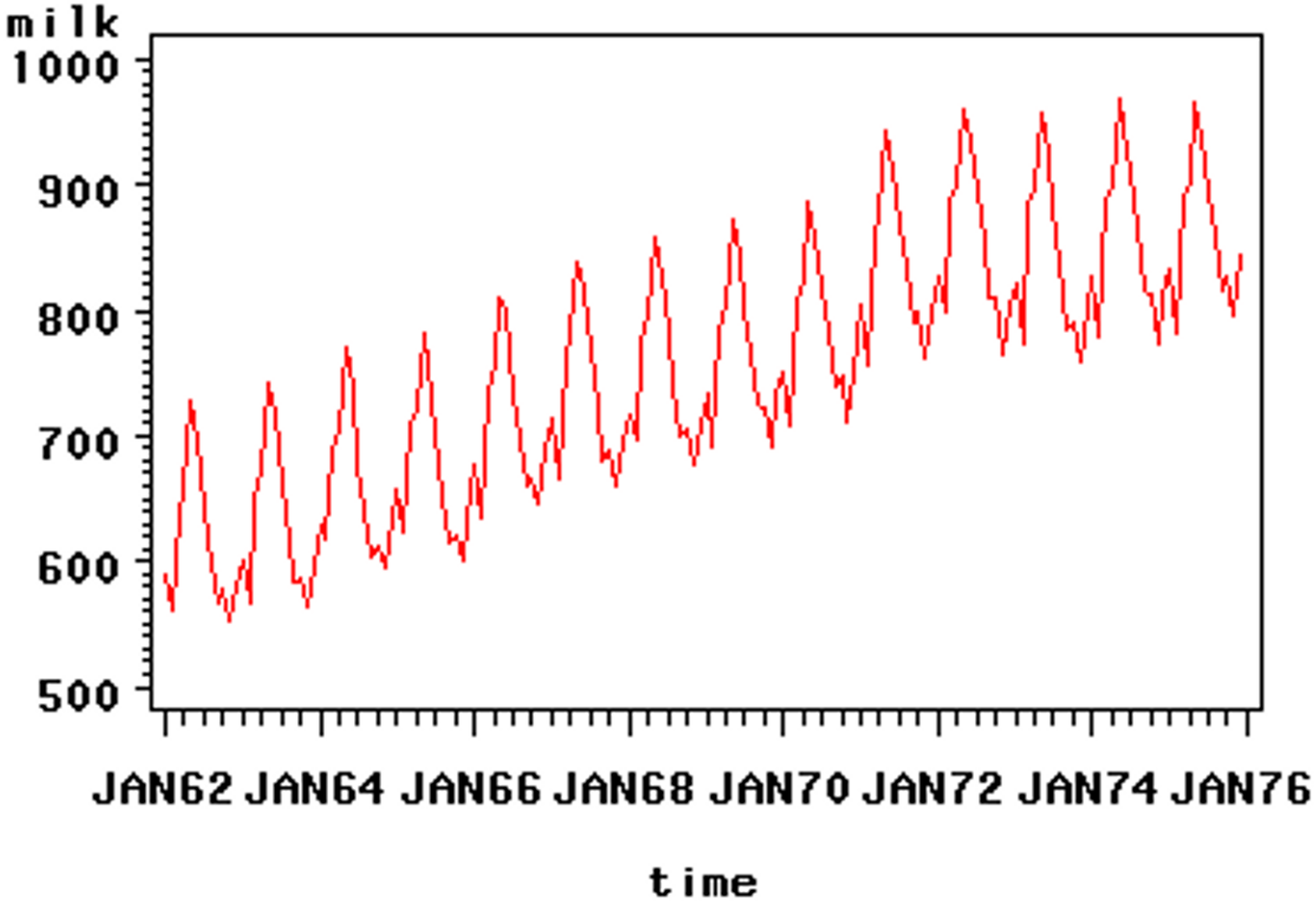


Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
0	21741.103	1.00000	|**********|
1	19869.670	0.91392	|********** .
2	18336.945	0.84342	|********** .
3	16679.644	0.76719	|********** .
4	15119.827	0.69545	|********** .
5	13234.768	0.60874	|********** .
6	11822.365	0.54378	|********** .
7	10355.425	0.47631	|********** .
8	8597.171	0.39543	|******** .
9	6977.227	0.32092	|****** .
10	5262.589	0.24206	|***** .
11	3185.458	0.14652	|*** .
12	1257.065	0.05782	|* .
13	-717.129	-.03298	. * | .
14	-2356.762	-.10840	. ** | .
15	-3657.864	-.16825	. *** | .
16	-4675.021	-.21503	. **** | .
17	-5645.938	-.25969	. ***** | .
18	-6662.959	-.30647	. ****** | .
19	-7523.279	-.34604	. ******* | .
20	-8300.856	-.38180	. ******** | .
21	-9068.912	-.41713	******** |
22	-9409.375	-.43279	********* |

"." marks two standard errors

1962年1月——1975年12月平均每头奶牛月产奶量（时序图）



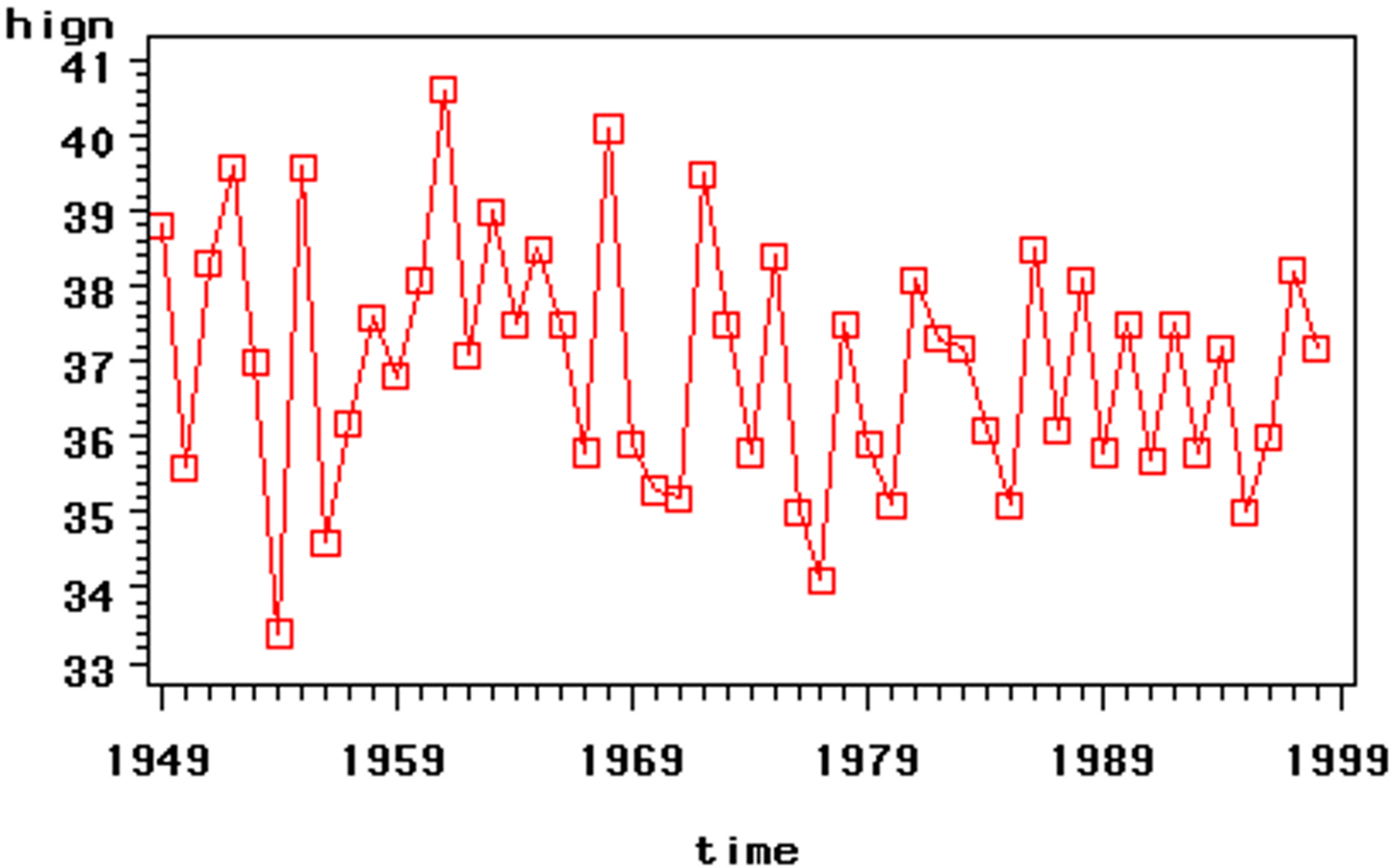
1962年1月——1975年12月平均每头奶牛月产奶量（自相关图）

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
0	10383.588	1.00000	|**********|
1	9257.734	0.89157	|*********. |
2	8080.289	0.77818	|******** |
3	6440.643	0.62027	|****** |
4	5053.314	0.48666	|**** |
5	4445.713	0.42815	|**** |
6	3904.890	0.37606	|**** |
7	4306.827	0.41477	|**** |
8	4716.761	0.45425	|**** |
9	5833.655	0.56181	|***** |
10	7128.946	0.68656	|****** |
11	7980.333	0.76855	|******** |
12	8773.234	0.84491	|*********. |
13	7735.639	0.74499	|******* |
14	6621.269	0.63767	|****** |
15	5084.621	0.48968	|**** |
16	3775.004	0.36355	|*** |
17	3176.849	0.30595	|*** |
18	2646.859	0.25491	|** |
19	2984.458	0.28742	|*** |
20	3328.659	0.32057	|*** |
21	4324.928	0.41652	|**** |
22	5489.933	0.52871	|***** |
23	6265.032	0.60336	|****** |
24	6986.088	0.67280	|****** |

"." marks two standard errors

1949年——1998年北京市每年最高气温（时序图）



1949年——1998年北京市每年最高气温（自相关图）

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
0	2.569604	1.00000	|**********|
1	-0.449960	-.17511	| . ***| . |
2	-0.0091078	-.00354	| . | . |
3	0.463204	0.18026	| . |**** . |
4	0.059232	0.02305	| . | . |
5	-0.421428	-.16400	| . ***| . |
6	0.253512	0.09866	| . |** . |
7	-0.067559	-.02629	| . * | . |
8	-0.0083274	-.00324	| . | . |
9	-0.057247	-.02228	| . | . |
10	0.148917	0.05795	| . |* . |
11	0.095461	0.03715	| . |* . |
12	-0.267799	-.10422	| . ** | . |
13	0.260969	0.10156	| . |** . |
14	0.011069	0.00431	| . | . |
15	-0.069243	-.02695	| . * | . |

"." marks two standard errors

随机性

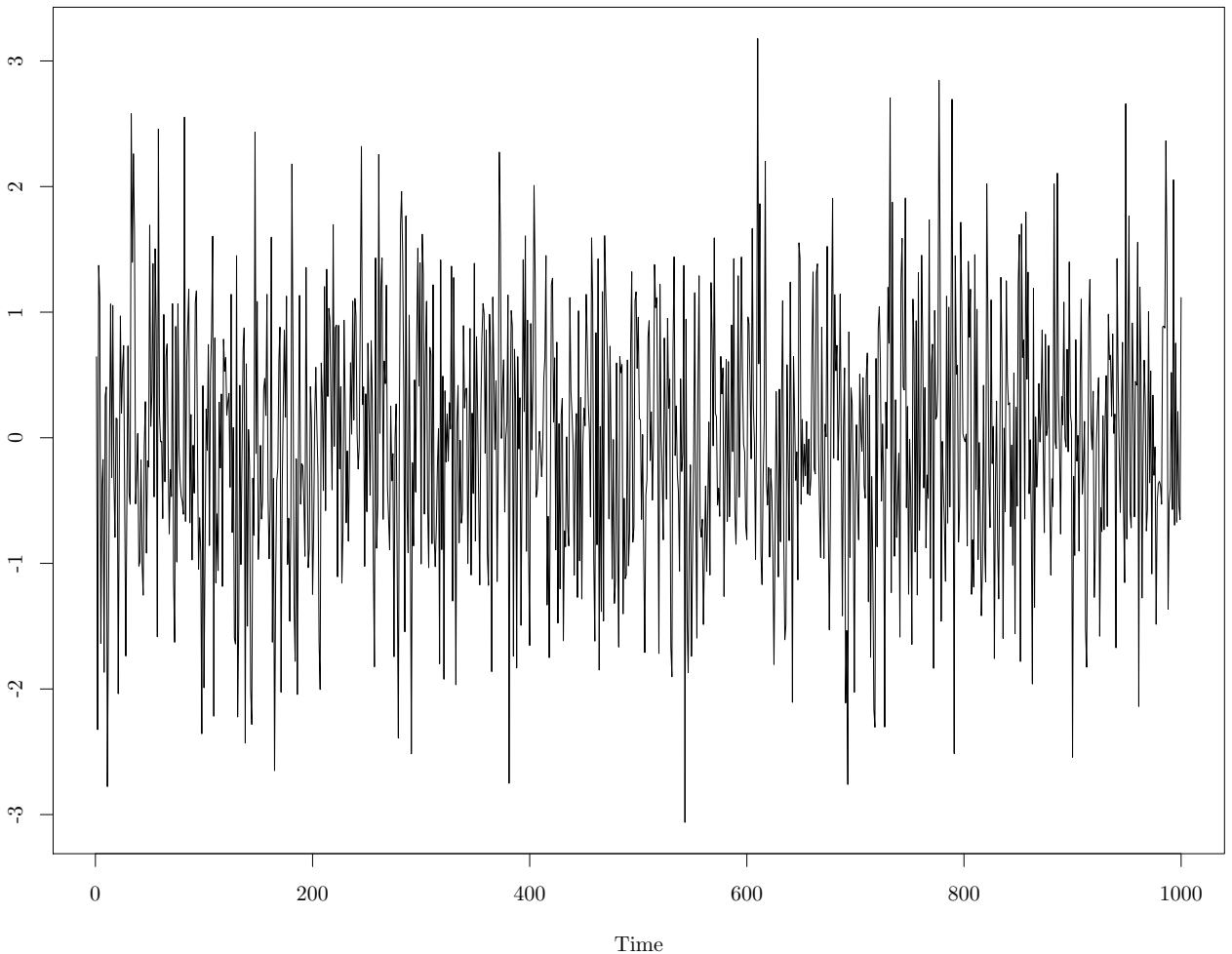
纯随机序列的定义

- 纯随机序列也称为白噪声序列，它满足如下两条性质

(1) $\mathbb{E}(X_t) = \mu, \forall t \in [0, T]$

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

标准正态白噪声序列时序图



Barlett 定理

- 如果一个时间序列是纯随机的，得到一个观察期数为 n 的观察序列，那么该序列的延迟非零期的样本自相关系数将近似服从均值为零，方差为序列观察期数倒数的正态分布

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

假设条件

- 原假设：延迟期数小于或等于 m 期的序列值之间相互独立

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

- 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

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

检验统计量

- Q 统计量

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

- LB 统计量

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