法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：小象
  - 新浪微博：ChinaHadoop

第二章　时间序列的预处理

主讲教师　　周仕君

本章结构

- 平稳性检验
- 纯随机性检验

2.1平稳性检验

- 特征统计量
- 平稳时间序列的定义
- 平稳时间序列的统计性质
- 平稳时间序列的意义
- 平稳性的检验

概率分布

- 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

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

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

- 实际应用的局限性

特征统计量

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

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。
- 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。

平稳时间序列的统计定义

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

  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
  - 不存在低阶矩的严平稳序列不满足宽平稳条件，例如服从柯西分布的严平稳序列就不是宽平稳序列
  - 当序列服从多元正态分布时，宽平稳可以推出严平稳

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
- 对称性
- 非负定性
- 非唯一性

平稳时间序列的意义

- 时间序列数据结构的特殊性
  - 可列多个随机变量，而每个变量只有一个样本观察值
- 平稳性的重大意义
  - 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

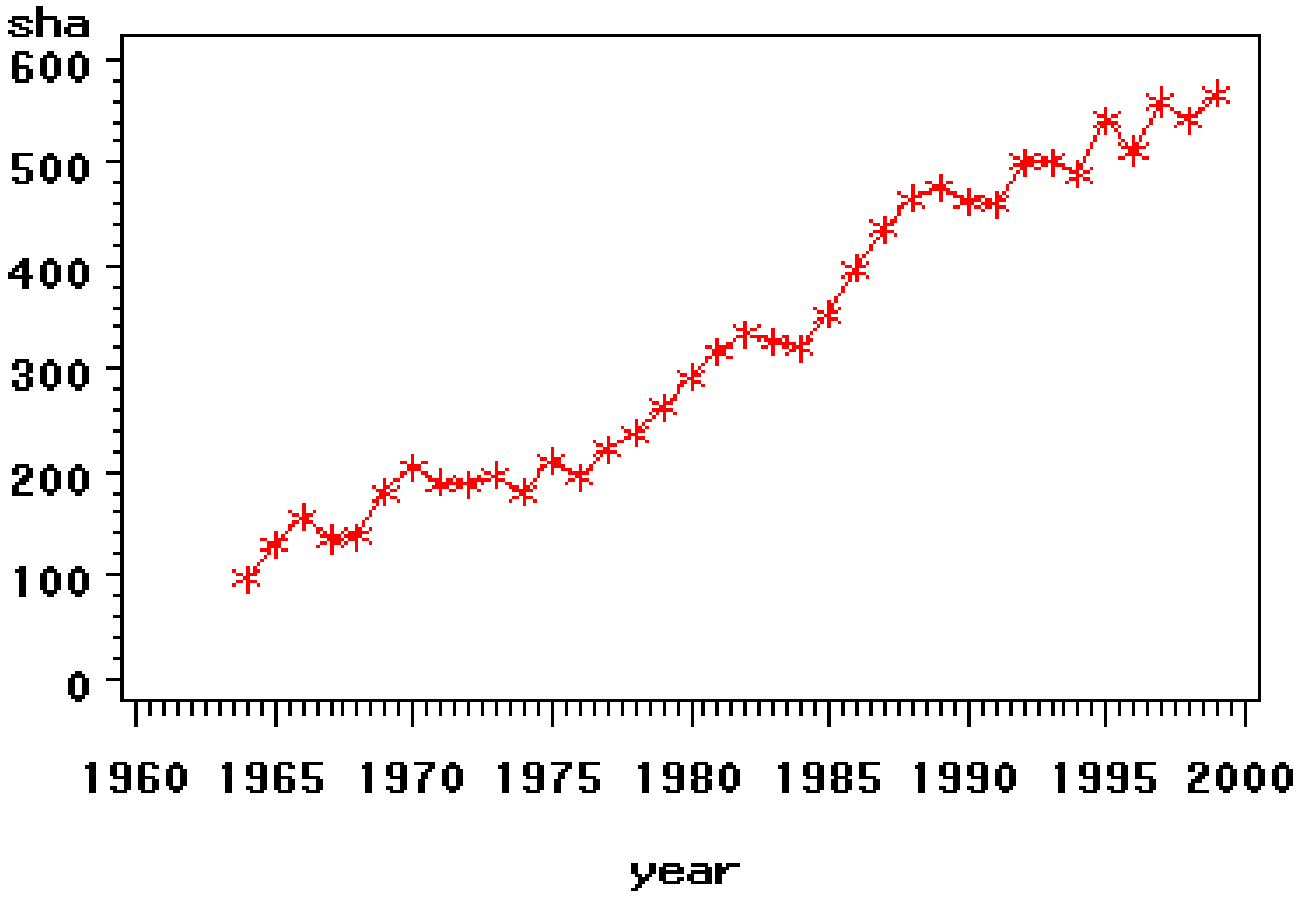
平稳性的检验（图检验方法）

- 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

例题

- 例2.1
  - 检验1964年——1999年中国纱年产量序列的平稳性
- 例2.2
  - 检验1962年1月——1975年12月平均每头奶牛月产奶量序列的平稳性
- 例2.3
  - 检验1949年——1998年北京市每年最高气温序列的平稳性

例2.1时序图


sha
600
500
400
300
200
100
0
1960 1965 1970 1975 1980 1985 1990 1995 2000
year

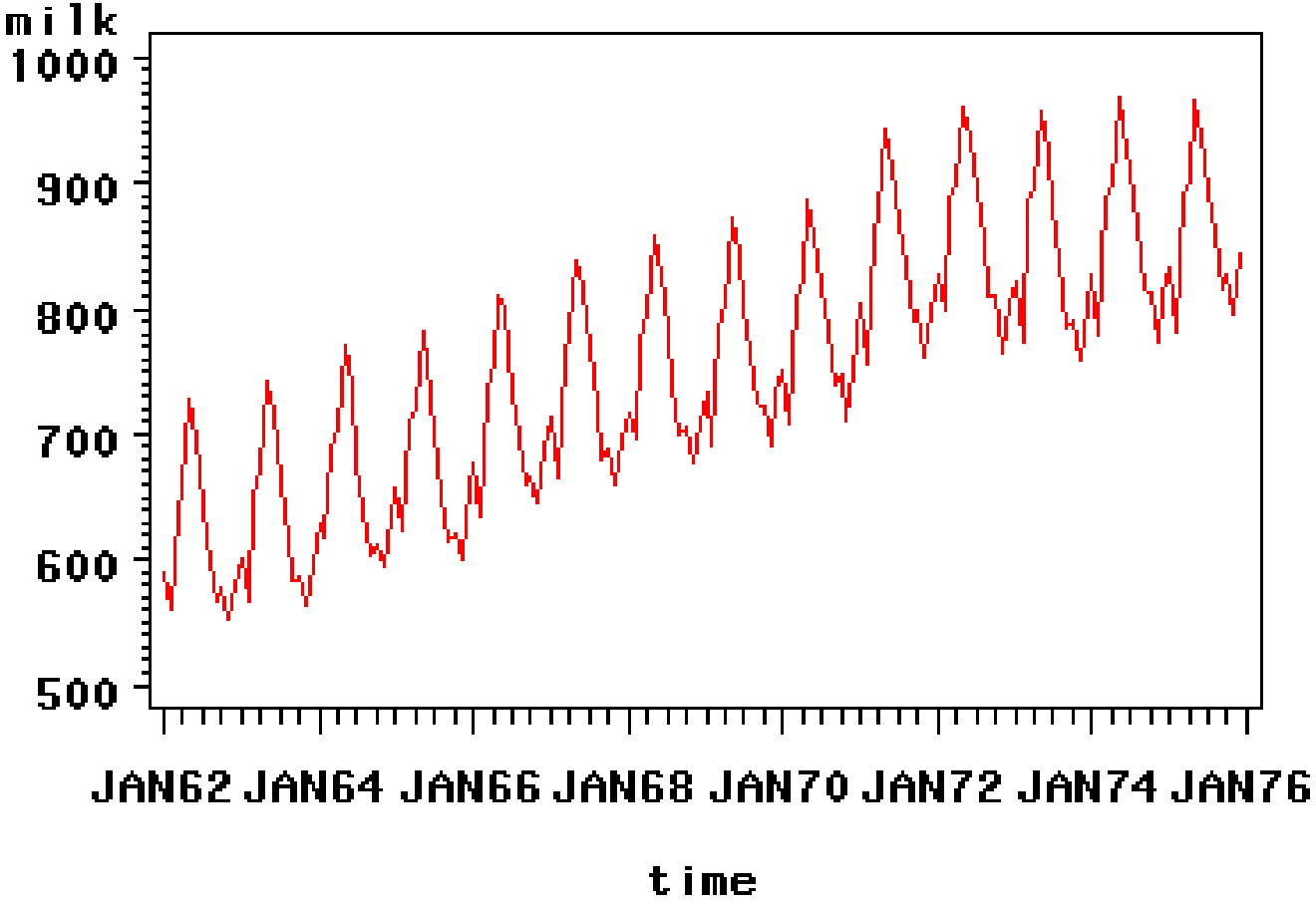
例2.1自相关图

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

例2.2时序图

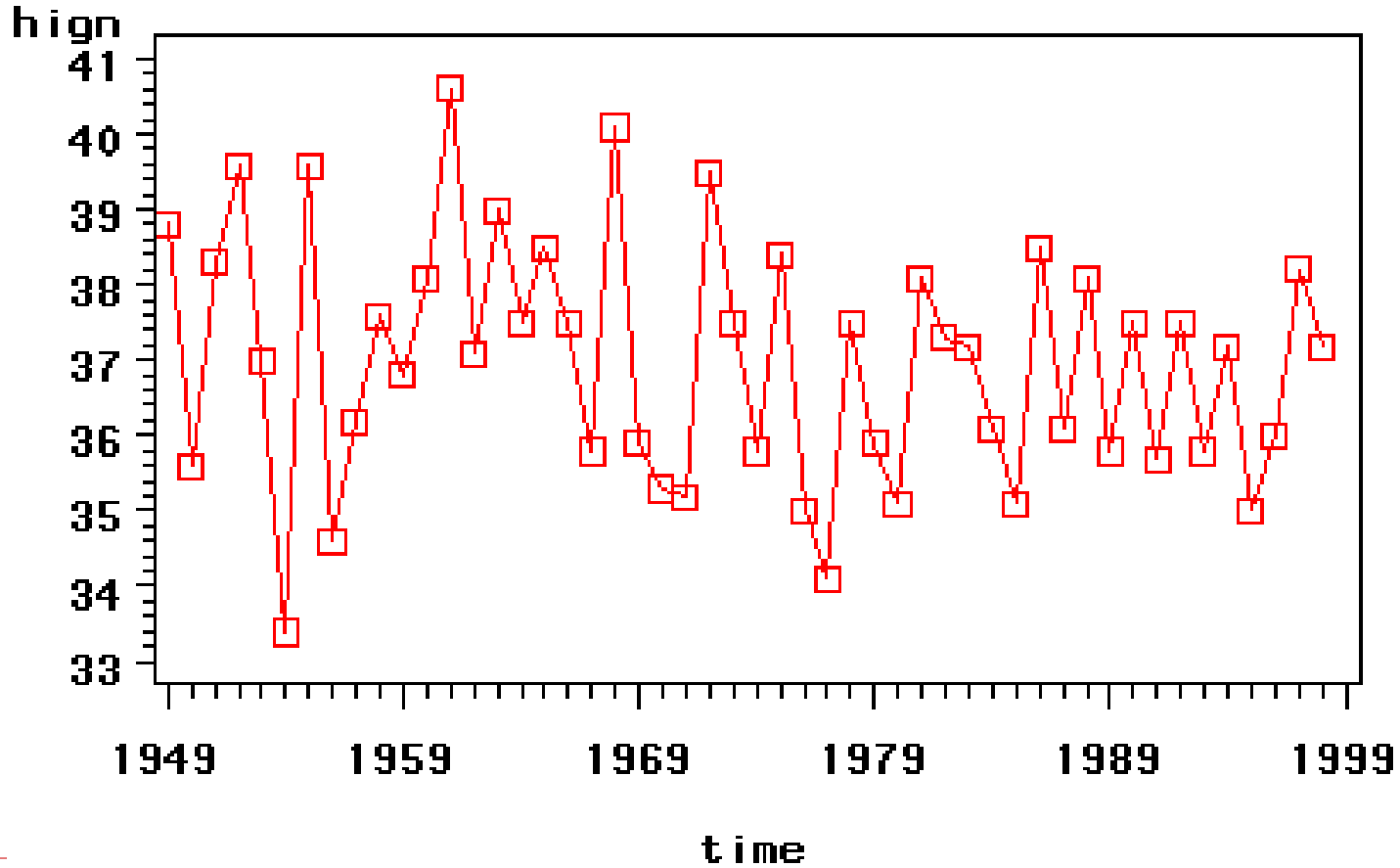

milk
1000
900
800
700
600
500
JAN62 JAN64 JAN66 JAN68 JAN70 JAN72 JAN74 JAN76
time

例2.2 自相关图

Lag	Covariance	Correlation	Autocorrelations
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

例2.3时序图


hign
41
40
39
38
37
36
35
34
33
1949
1959
1969
1979
1989
1999
time

例2.3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

2.2 纯随机性检验

- 纯随机序列的定义
- 纯随机性的性质
- 纯随机性检验

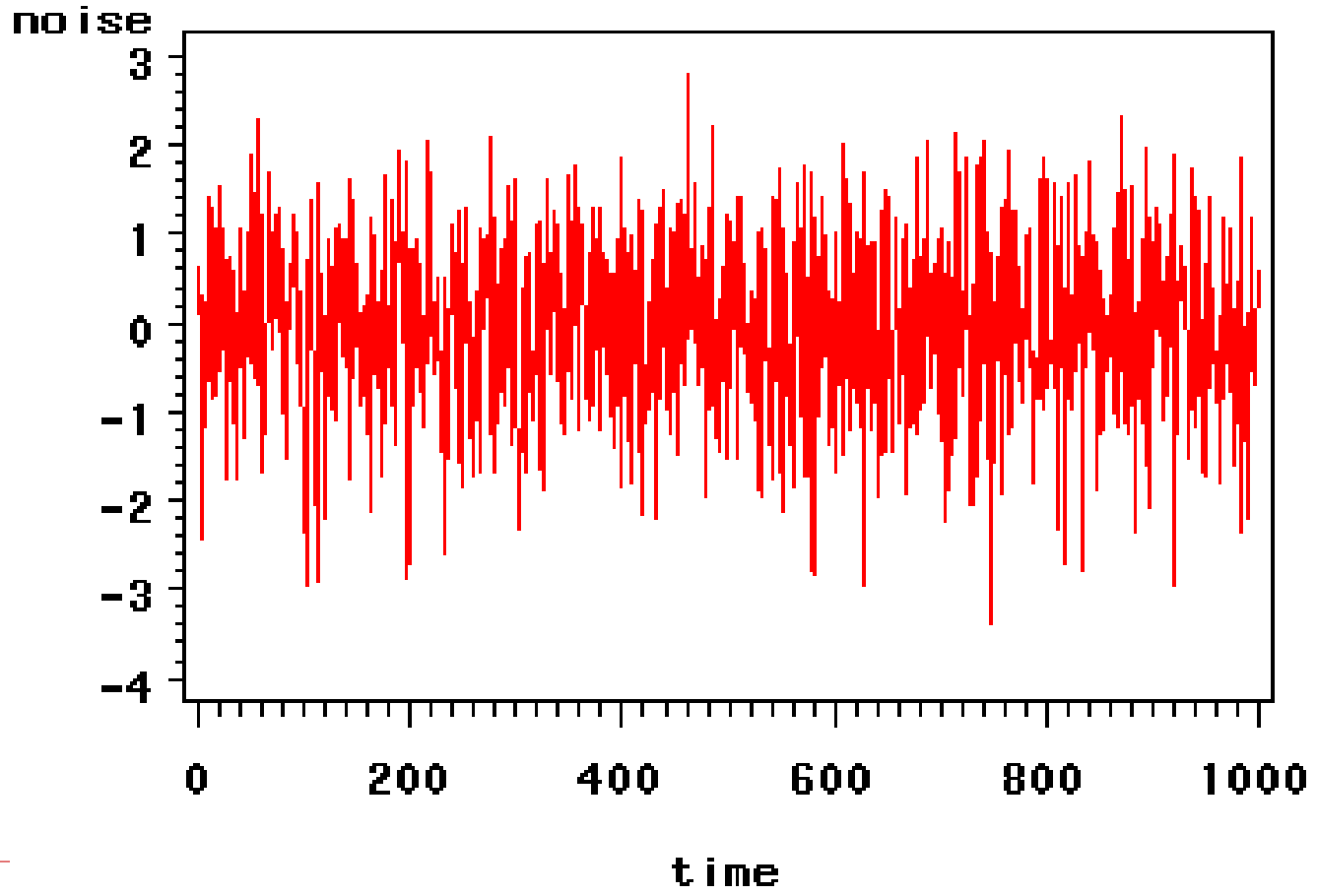
纯随机序列的定义

- 纯随机序列也称为白噪声序列，它满足如下两条性质

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

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

标准正态白噪声序列时序图



白噪声序列的性质

- 纯随机性 $\gamma(k)=0,\ \forall k\neq 0$
  - 各序列值之间没有任何相关关系，即为“没有记忆”的序列
- 方差齐性 $DX_t=\gamma(0)=\sigma^2$
  - 根据马尔可夫定理，只有方差齐性假定成立时，用最小二乘法得到的未知参数估计值才是准确的、有效的

纯随机性检验

- 检验原理
- 假设条件
- 检验统计量
- 判别原则

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

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

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

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

- 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

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

判别原则

- 拒绝原假设
  - 当检验统计量大于 $1-\alpha$ 分位点，或该统计量的P值小于α 时，则可以以 $\chi^2_{1-\alpha}(m)$的置信水平拒绝原假设，认为该序列为非白噪声序列
- 接受原假设
  - 当检验统计量小于 $1-\alpha$ 分位点，或该统计量的P值大于α 时，则认为在 $\chi^2_{1-\alpha}(m)$的置信水平下无法拒绝原假设，即不能显著拒绝序列为纯随机序列的假定

例2.4：标准正态白噪声序列纯随机性检验

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	1.005528	1.00000	|**********|
1	-0.0010647	-.00106	.
2	-0.036747	-.03655	*.
3	-0.0062567	-.00622	.
4	0.011938	0.01187	.
5	-0.025139	-.02500	*.
6	-0.014428	-.01435	.
7	0.0088565	0.00881	.
8	-0.010179	-.01012	.
9	-0.026893	-.02674	*.
10	-0.024882	-.02475	.
11	-0.014021	-.01394	.
12	0.035527	0.03533	.*

"." marks two standard errors

样本自相关图

检验结果

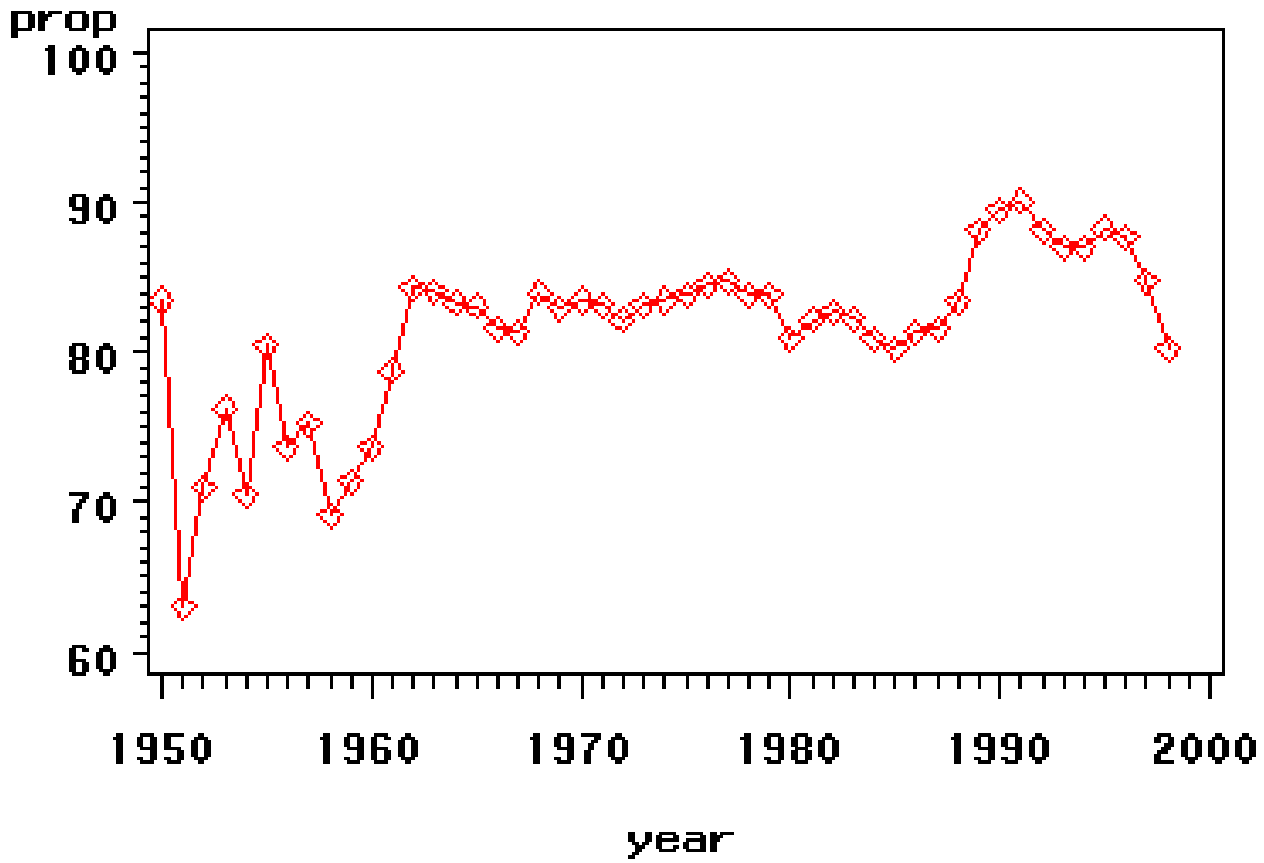
延迟	统计量检验	
	Q_{LB}统计量值	P值
延迟6期	2.36	0.8838
延迟12期	5.35	0.9454

由于P值显著大于显著性水平α，所以该序列不能拒绝纯随机的原假设。

例2.5

- 对1950年——1998年北京市城乡居民定期储蓄所占比例序列的平稳性与纯随机性进行检验

例2.5时序图



例2.5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	30.725523	1.00000	|********************|
1	21.583411	0.70246	|************** . |
2	18.293557	0.59539	|************ . |
3	14.684303	0.47792	|********** . |
4	10.080193	0.32807	|******* . |
5	10.931717	0.35579	|******* . |
6	9.318240	0.30327	|****** . |
7	8.944975	0.29113	|****** . |
8	4.927541	0.16037	|*** . |
9	1.842114	0.05995	|* . |
10	-1.151434	-.03747	| .* |
11	-2.369343	-.07711	| .** |
12	-1.130247	-.03679	| .* |

"." marks two standard errors

例2.5白噪声检验结果

延迟阶数	LB统计量检验	
	LB检验统计量的值	P值
6	75.46	<0.0001
12	82.57	<0.0001