Thurs, 2-07-19

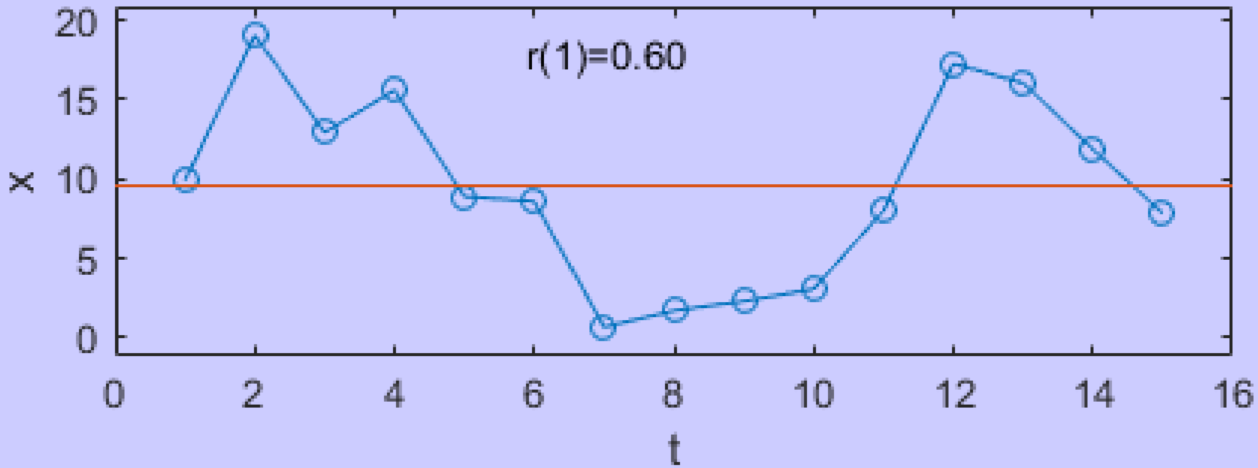
3. Autocorrelation (cont.)

1. Effective sample size
2. Confidence interval for acf
3. Stationarity
4. Sample runs of a3

– Assignment a3: due Tues, Feb 12

Effective sample size

$x_t, \quad t = 1, \ldots, N$ Positively autocorrelated time series



- *N* observations, but *N*'<*N* independent pieces of information
- *N*' is "effective" sample size
- *N*'<*N* → increased uncertainty in statistics computed from x_t

… effective sample size

- Assume x_t is generated by a process with autocorrelation at lag 1 only
- Theory (see notes) gives:

$$N' = \left(\frac{1-r_1}{1+r_1}\right)N$$ is defined as the effective sample size

Say,

$$N = 100, \quad r_1 = 0.5$$

$$N' = \frac{1-0.5}{1+0.5}N = \frac{0.5}{1.5}N = \frac{N}{3} = 33$$

… effective sample size

- The relevance of the effective sample size is that statistics computed on the time series have increased uncertainty.
- This is because those statistics are based on fewer pieces of independent information than implied by the sample size.
- An example is the variance of the sample mean…

Variance of time series

$$\operatorname{var}\left[\bar{x}\right] = \frac{s^2}{N}$$

Sample size, before adjustment

Lag-1 autocorrelation

$$\operatorname{var}\left[\bar{x}\right] = \frac{s^2}{N'} = \frac{s^2}{N}\underbrace{\left(\frac{1+r_1}{1-r_1}\right)}$$

"Variance inflation factor", or time between independent events

Confidence interval for acf:
“population” autocorrelation

Sample autocovariance and autocorrelation

c_k and r_k are viewed as samples from populations

⇓

γ_k and ρ_k -- corresponding population functions

autocovariance *autocorrelation*

- Assume the time series was generated by random variables with zero autocorrelation
- Then the sample autocorrelation (r_k) is normally distributed with a specified expected value (“E”) and variance (“var”)
- Consider the lag-1 sample autocorrelation, r_1 …

Confidence interval for acf …

Distribution of $r_{1;}$ assuming x_t is "iid", or independent and identically distributed, and N is large :

Sample r_1 is normally distributed, with

$$\mathrm{E}\left(r_1\right) = -1/N$$

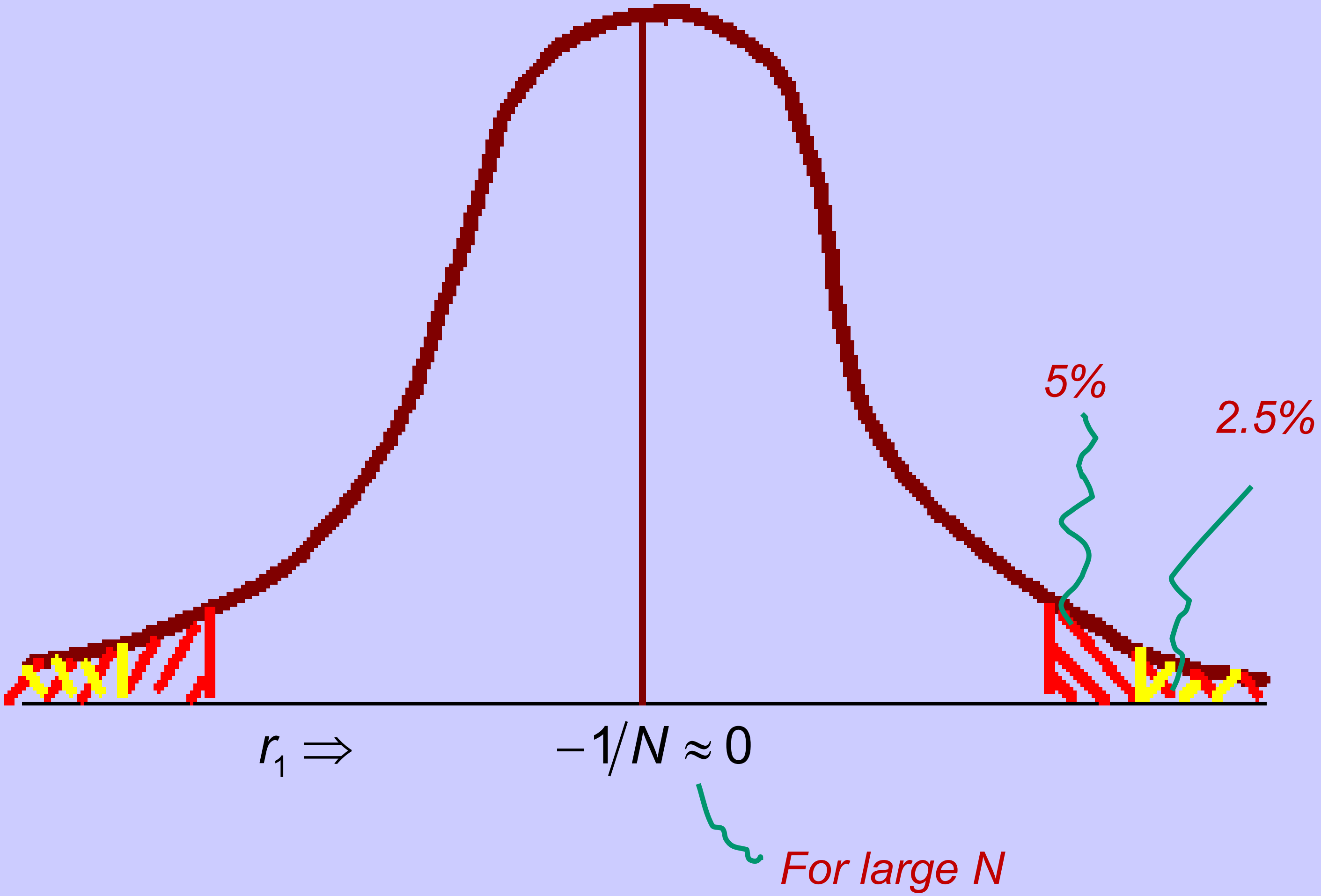
$$\mathrm{var}\left(r_1\right) = 1/N$$

i.e., asymptotically, r ~ $\mathcal{N}\left(-1/N, 1/N\right)$

as N→ ∞

Confidence interval for acf …

Distribution of sample lag-1 autocorrelations



Confidence interval for acf …

<u>Two</u>-tailed test of significance of r_1

$\text{H0}: \; \rho_1 = 0$

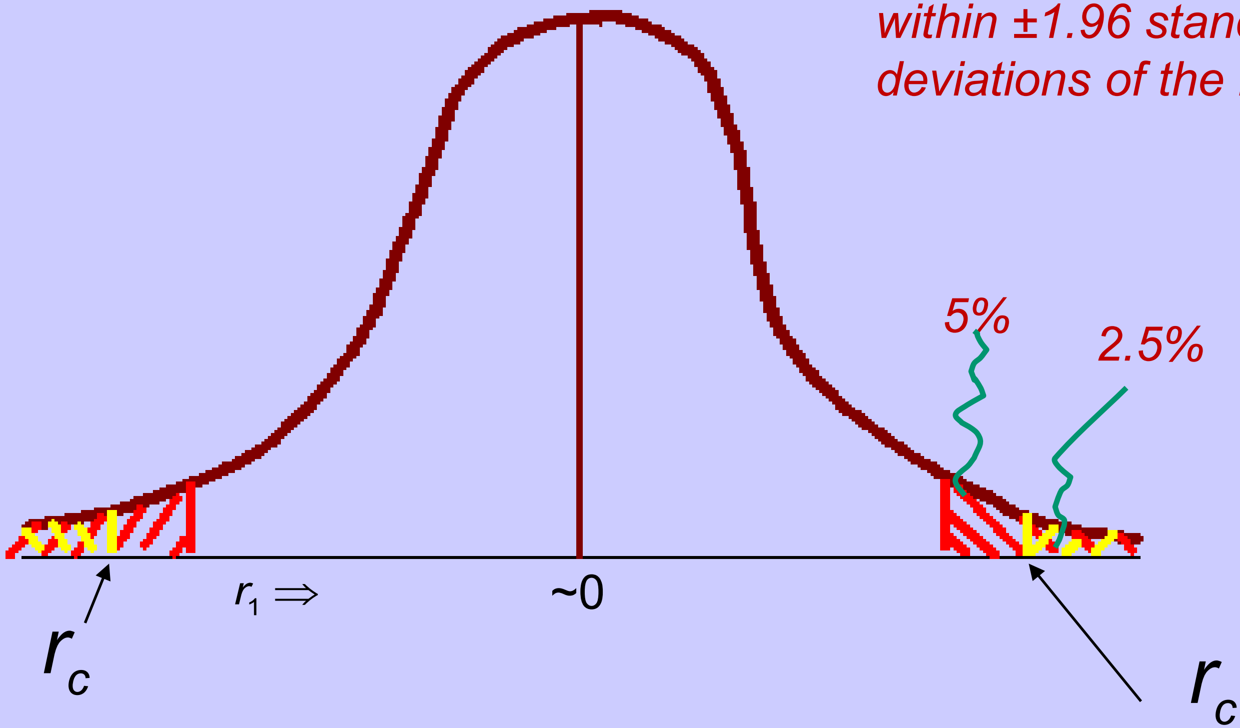
$\text{H1}: \; \rho_1 \neq 0$

$|r_1| > \dfrac{1.96}{\sqrt{N}} \rightarrow$ reject H0 at $\alpha = 0.05$

Recall that

$$\text{std}(r_1) = \sqrt{\text{var}(r_1)} = 1/\sqrt{N}$$

and that 95% of the normal distribution falls within ±1.96 standard deviations of the mean



Critical or threshold r_1

Confidence interval for acf …

<u>One</u>-tailed test of significance of r_1

$H0:\ \rho_1 = 0$

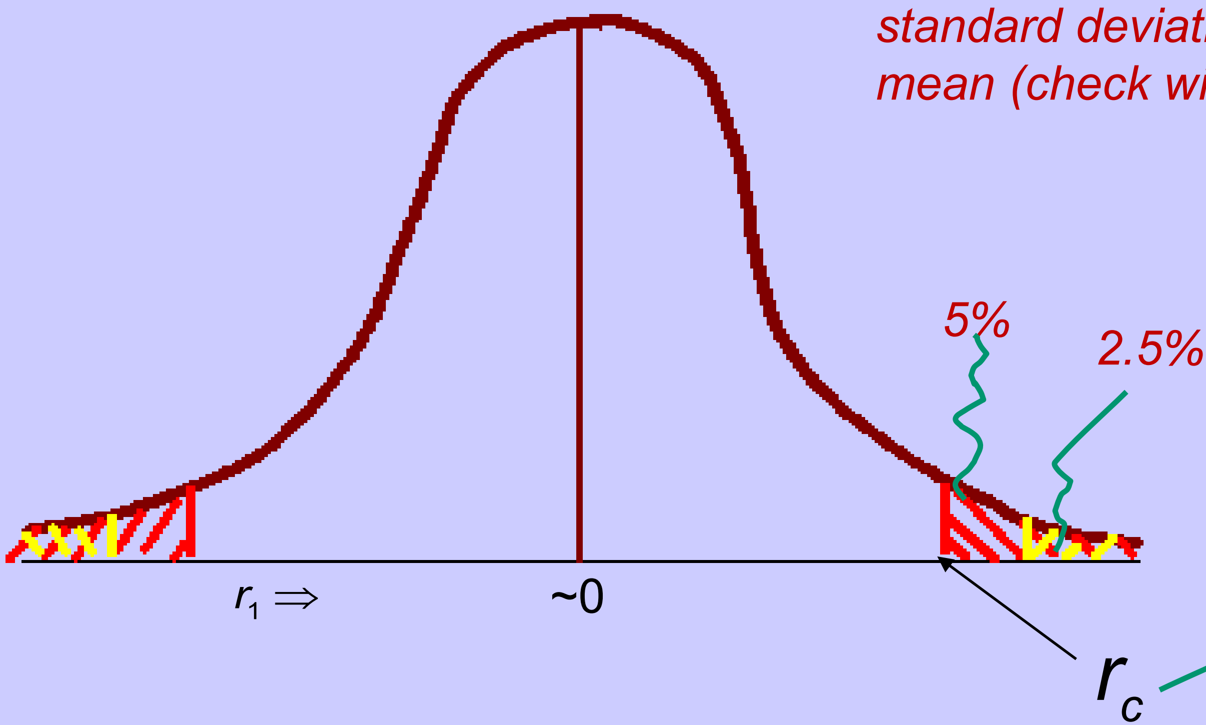
$H1:\ \rho_1 > 0$

$|r_1| > \frac{1.645}{\sqrt{N}} \rightarrow$ reject H0 at $\alpha = 0.05$

Recall that

$\mathrm{std}(r_1) = \sqrt{\mathrm{var}(r_1)} = 1/\sqrt{N}$

and that 95% of the normal distribution falls below +1.645 standard deviations above the mean (check with disttool)



A sample r_1 significant at α=0.05 by a 1-tailed test may not be significant by a 2-tailed test

Confidence interval for acf …

- While, strictly, the confidence interval on *r*(*k*) would widen with increasing *k,* it is often assumed that *N* is large. Then, if N>>k, horizontal confidence intervals can be used to assess significance of the acf at different lags
- As an approximation, the 95% confidence interval for a two-tailed test can be plotted at

$$0 \pm 1.96 / \sqrt{N}$$

where *N* is the sample size

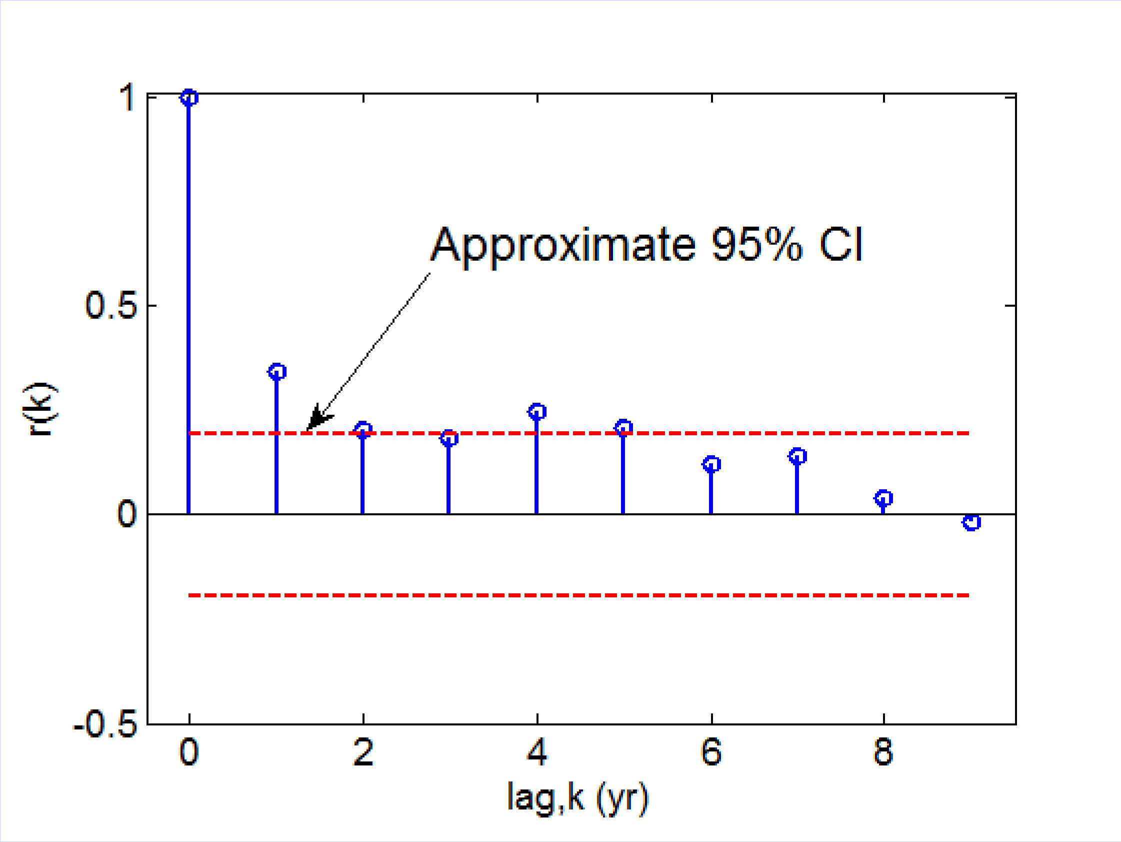
and for a one-tailed test

$$0 \pm 1.645 / \sqrt{N}$$

Confidence interval for acf …

Example for a tree-ring chronology with N=107

$$\frac{1.96}{\sqrt{N}} = \frac{1.96}{\sqrt{107}} \simeq 0.19$$



Confidence interval for acf …

As a further approximation, 1.96 is often rounded to 2, and the 95% interval is plotted as:

$$0 \pm 2\Big/\sqrt{N}$$

Confidence interval for acf …

- The horizontal confidence intervals just described are based on an assumption that the population has zero autocorrelation
- If the sample autocorrelations at lower lags are assumed to reflect “true” autocorrelation (of the process), the confidence interval for the sample acf widens at higher lags
- A modified CI can be computed based on assumption that lower-lag theoretical autocorrelations are non-zero

“Large-lag” standard error of acf

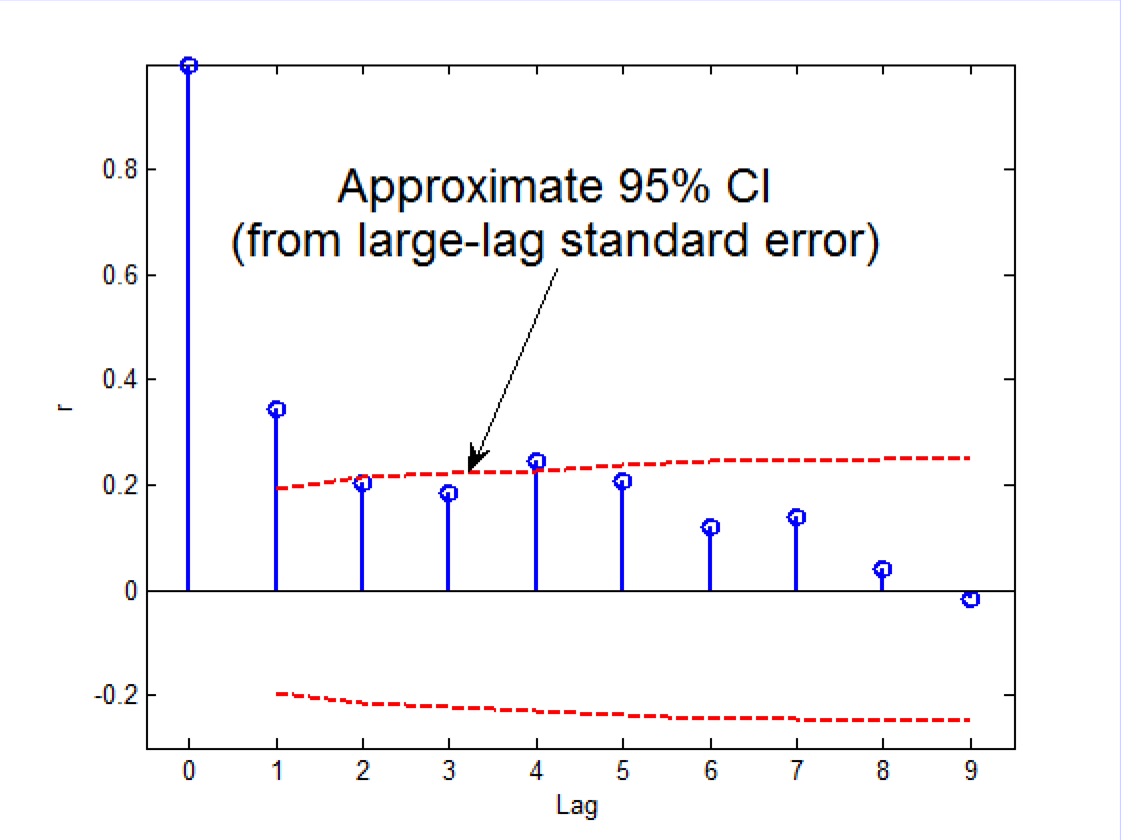
“Large-lag” standard error is defined as square root of

$$\mathrm{Var}(r_k) \simeq \frac{1}{N}\left(1 + 2\sum_{i=1}^{K} r_i^{\,2}\right), \quad K < k$$

Error bars on acf’s in class assignment A3 are based on the above equation

“Large-lag” standard error of acf

Example for a tree-ring chronology with *N*=107



Stationarity

- Many time series methods involving inference assume stationarity
- Stationarity refers to the **process** that generated the time series; you can infer stationarity of the process from characteristics of the observed time series
- A time series is sometimes described as stationary if it does not have any obvious sign of trend – this is **stationarity in the mean**
- **“Weak” stationarity** is stationarity in the mean and covariance. This means that the random variables generating the series have mean and covariance that do not depend on **absolute time**
- **“Strict” stationarity** means that the joint probability distribution of the random variables is not a function of absolute time. If the random variables are normally distributed and weakly stationary, they are also strictly stationary

Weak stationarity

Covariance of X_i with X_j

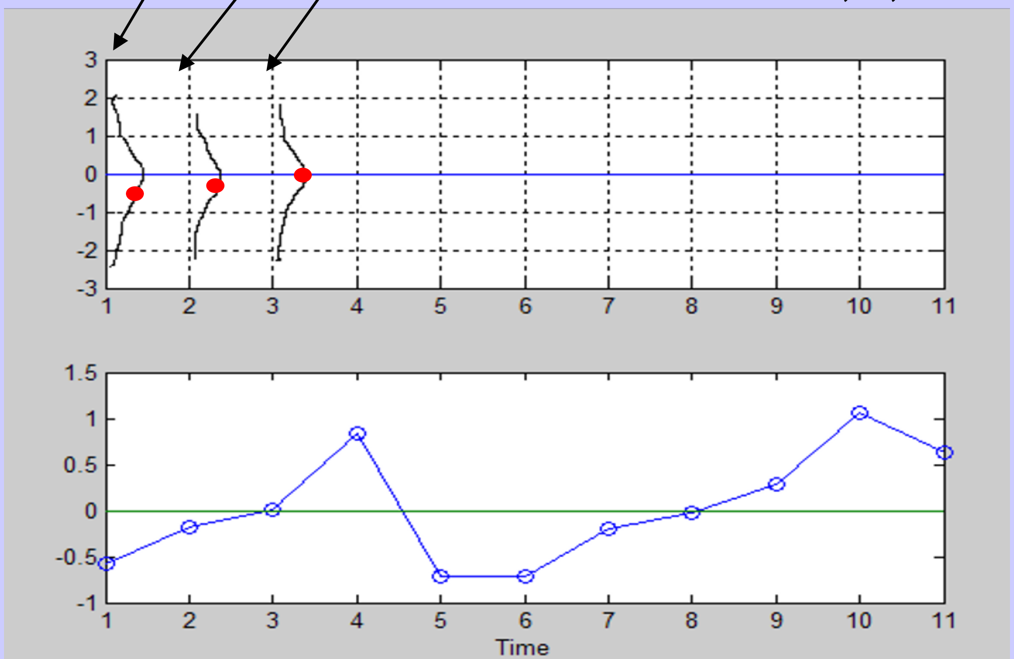
1) $\mu_1 = \mu_2 = \mu_3, \ldots$

2) $\gamma_{i,j}(k), k = 0,1,2,\ldots$ function of lag, k, only

X_1 X_2 X_3

Random variables associated with times 1, 2, 3

Stochastic process



Generating Process

Observed series (realization)

Short-memory process vs long-memory process

- Short: acf decays “quickly” to zero
- Long: acf does not decay quiickly

Both are stationary

Population autocorrelation

Short memory $\sum_{k=0}^{\infty} |\rho_k|$ converges

Long memory $\sum_{k=0}^{\infty} |\rho_k| \rightarrow \infty$

- In practice, VERY difficult to distinguish long-term memory from nonstationarity
- Chatfield (204, p 261) discusses in detail
- In hydrology, long memory processes have been invoked to explain the Nile River discharge record

Trial runs of geosa3…