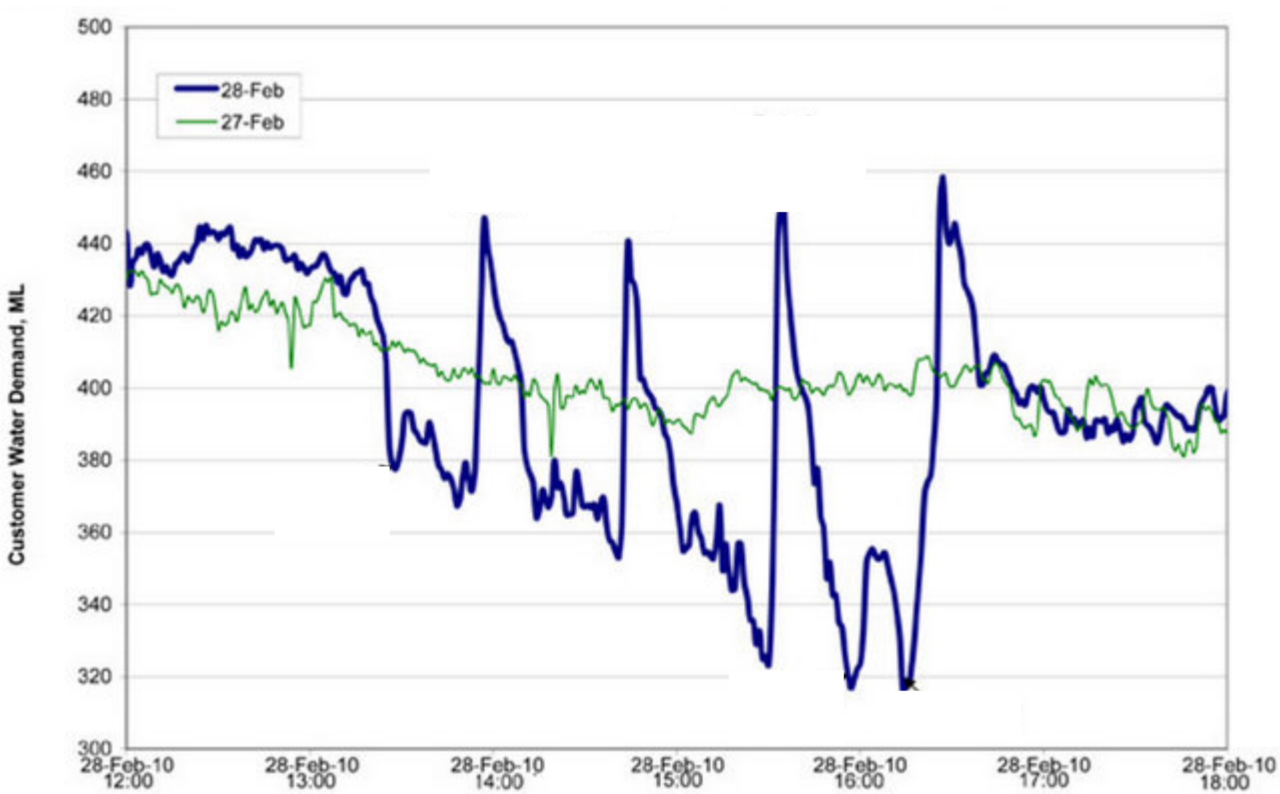
# Covariance, stationarity & some useful operators

Mark Scheuerell

*FISH 507 – Applied Time Series Analysis*

5 January 2017

# Example of a time series



# Topics for today

- Expectation, mean & variance
- Covariance & correlation
- Stationarity
- Autocovariance & autocorrelation
- Correlograms
- White noise
- Random walks
- Backshift & difference operators

# Expectation, mean & variance

- The *expectation* (E) of a variable is its mean value in the population
- $E(x) \equiv$ mean of $x = \mu$
- $E([x - \mu]^2) \equiv$ mean of squared deviations about $\mu$

  $\equiv$ variance $= \sigma^2$
- Can estimate $\sigma^2$ from sample as

$$\mathrm{Var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

# Covariance

- If we have 2 variables (*x*, *y*) we can generalize variance

  $$\sigma_x^2 = \mathrm{E}\left[(x-\mu_x)(x-\mu_x)\right]$$

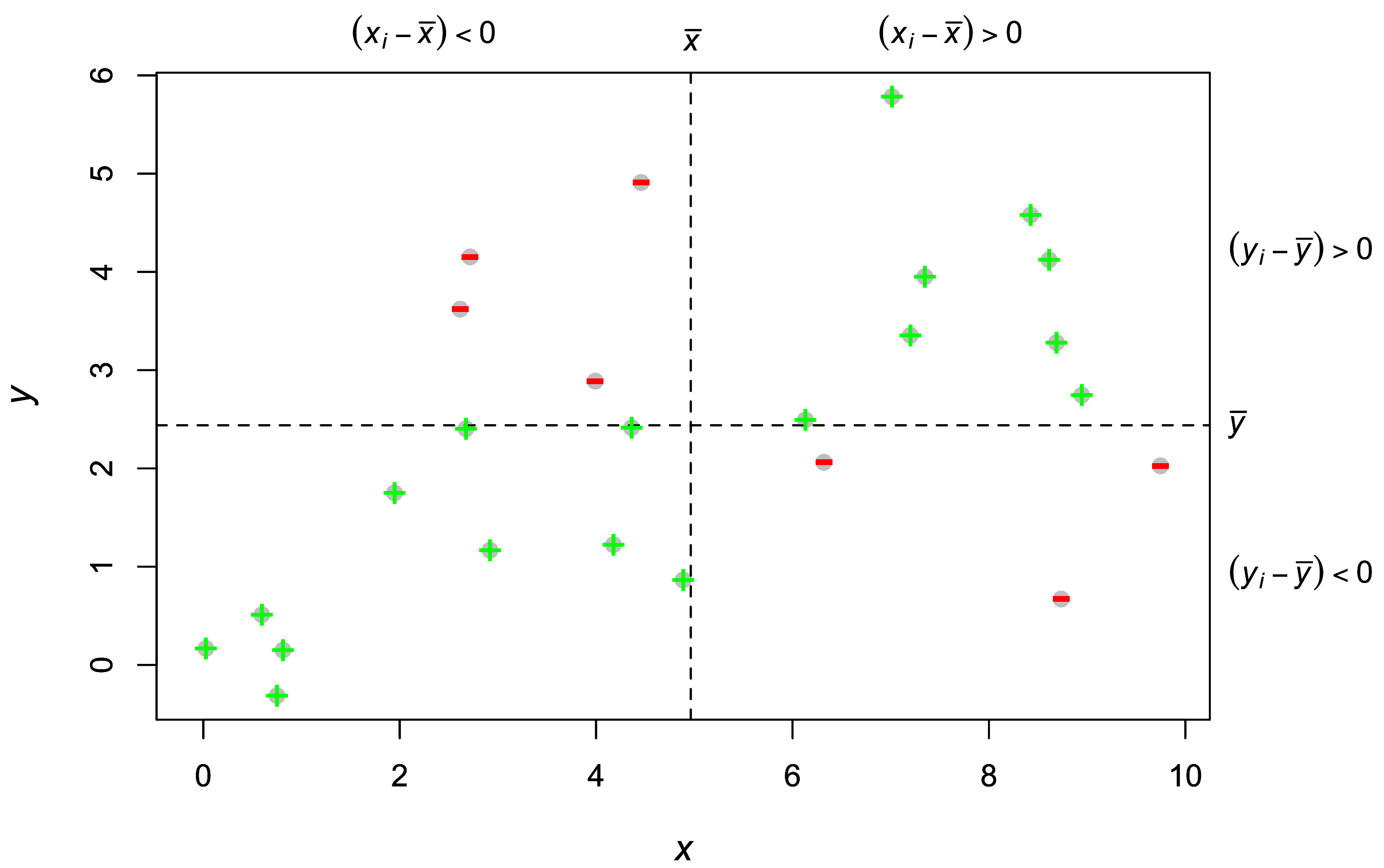
  to *covariance*

  $$\gamma(x,y) = \mathrm{E}\left[(x-\mu_x)(y-\mu_y)\right]$$

- Can estimate γ from sample as

  $$\mathrm{Cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$$

# Graphical example of covariance



# Correlation

- *Correlation* is a dimensionless measure of the linear association between 2 variables *x* & *y*

- It is simply the covariance standardized by the standard deviations

$$\rho(x,y)=\frac{\mathrm{E}\left[(x-\mu_x)(y-\mu_y)\right]}{\sigma_x\sigma_y}=\frac{\gamma(x,y)}{\sigma_x\sigma_y}\in\left[-1,1\right]$$

- Can estimate $\gamma$ from sample as

$$\mathrm{Cor}(x,y)=\frac{\mathrm{Cov}(x,y)}{\mathrm{sd}(x)\mathrm{sd}(y)}$$

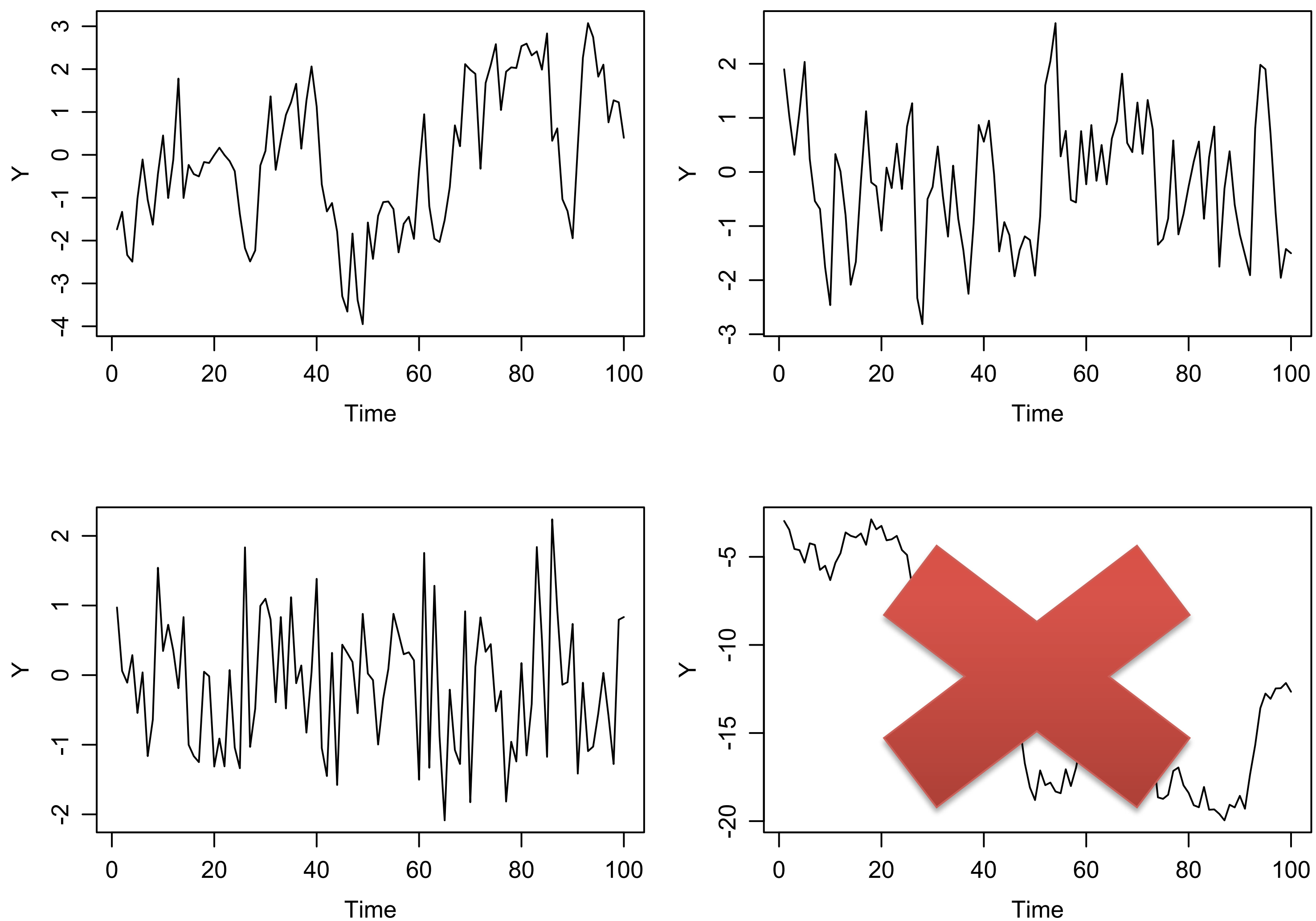
# The ensemble & stationarity

- Consider again the mean function for a time series:
  $\mu(t) = E(x_t)$
- The expectation is taken across an *ensemble* (population) of all possible time series
- With only 1 sample, however, we must estimate the mean at each time point by the observation
- If $E(x_t)$ is constant across time, we say the time series is *stationary* in the mean

# Stationarity of time series

- *Stationarity* is a convenient assumption that allows us to describe the statistical properties of a time series.
- In general, a time series is said to be stationary if there is
  1) no systematic change in the mean or variance,
  2) no systematic trend, and
  3) no periodic variations or seasonality

# Which of these are stationary?



# Autocovariance function (ACVF)

- For stationary ts, we can define the *autocovariance function* (ACVF) as a function of the time lag (k)

$$\gamma_k = \mathrm{E}\left[(x_t - \mu_x)(x_{t+k} - \mu_x)\right]$$

- Very "smooth" series have large ACVF for large k; "choppy" series have ACVF near 0 for small k

- Can estimate $\gamma_k$ from sample as

$$c_k = \frac{1}{\boxed{n}} \sum_{t=1}^{\boxed{n-k}} (x_t - \bar{x})(x_{t+k} - \bar{x})$$

# Autocorrelation function (ACF)

- The *autocorrelation function* (ACF) is simply the ACVF normalized by the variance

$$\rho_k = \frac{\gamma_k}{\sigma^2} = \frac{\gamma_k}{\gamma_0}$$

- ACF measures the correlation of a time series against a time-shifted version of itself (& hence the term "auto")

- Can estimate $\gamma_k$ from sample as

$$r_k = \frac{c_k}{c_0}$$

# Properties of the ACF

The ACF has several important properties, including

1) $-1 \leq r_k \leq 1$,
2) $r_k = r_{-k}$ (ie, it's an "even function"),
3) $r_k$ of periodic function is itself periodic
4) $r_k$ for sum of 2 indep vars is sum of $r_k$ for each

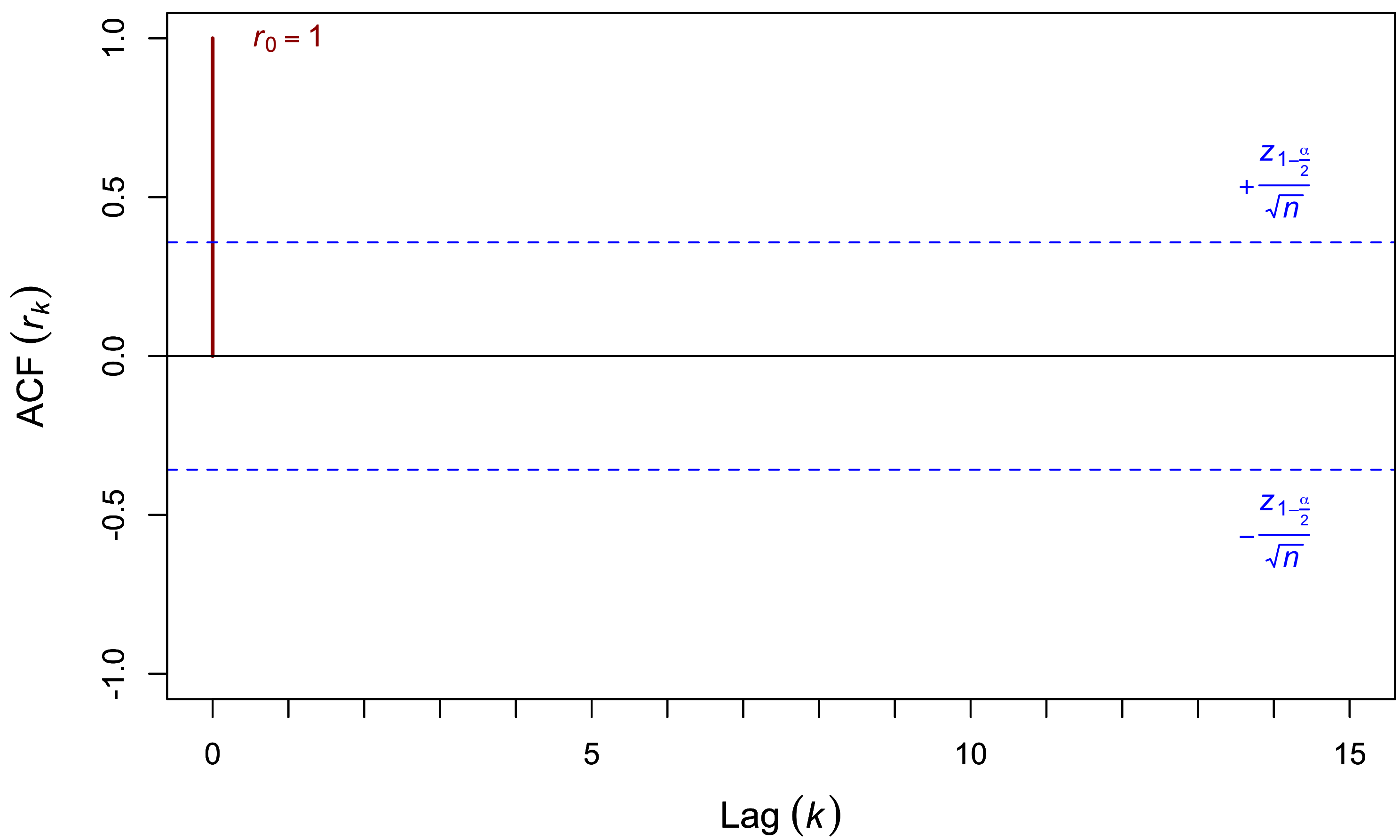
# The correlogram

- The common graphical output for the ACF is called the *correlogram*, and it has the following features:

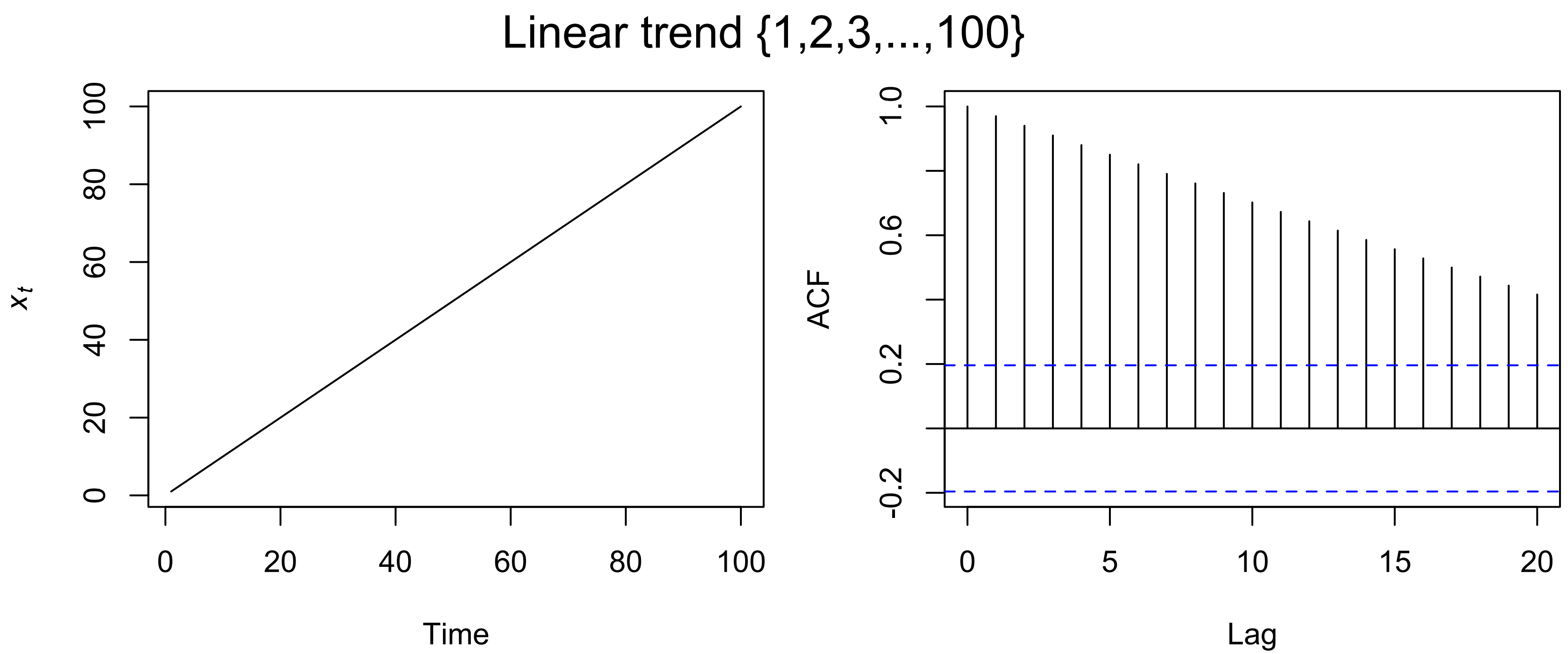
1) x-axis indicates lag (0 to $k$);
2) y-axis is autocorrelation $r_k$ (-1 to 1);
3) lag-0 correlation ($r_0$) is always 1 (it's a ref point);
4) If $\rho_k = 0$, then sampling distribution of $r_k$ is approx. normal, with var = 1/n;
5) Thus, a 95% conf interval is given by

$$\pm \frac{z_{1-\alpha/2}}{\sqrt{n}}$$

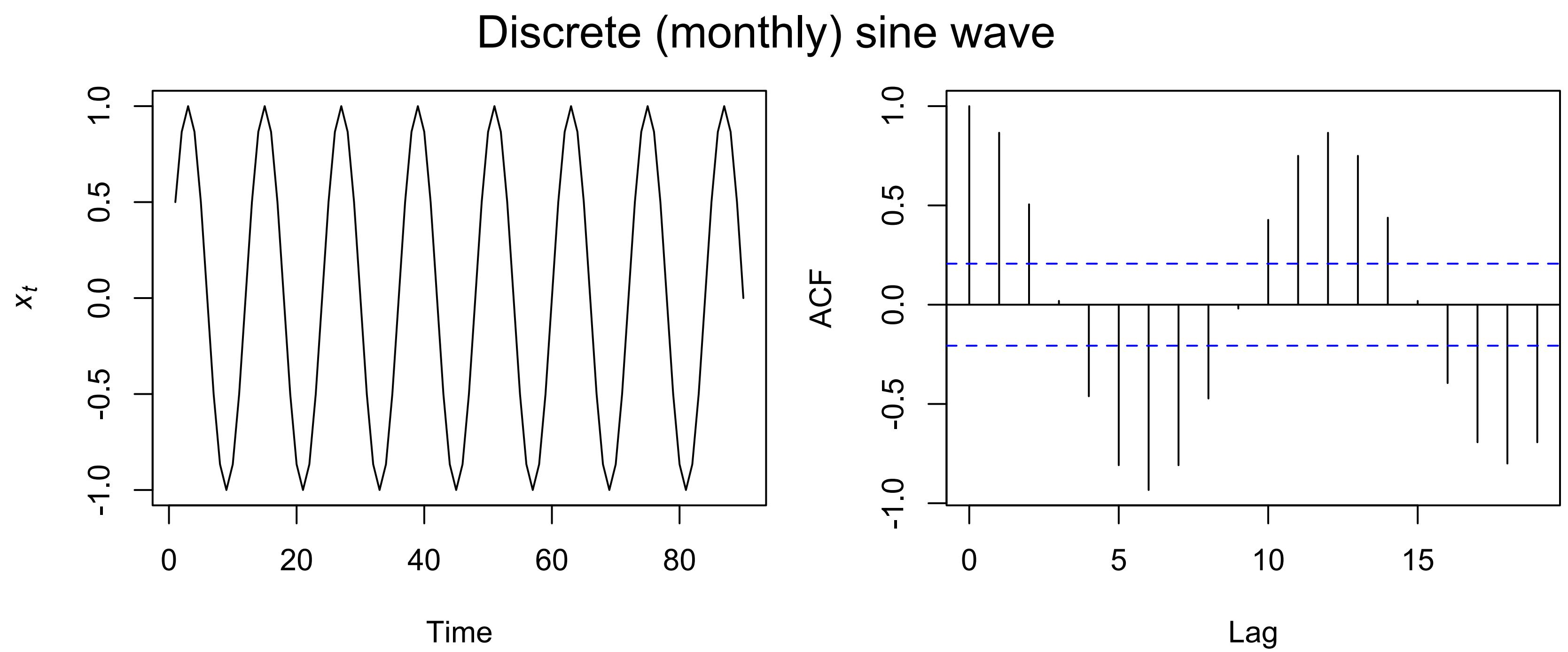
# The correlogram



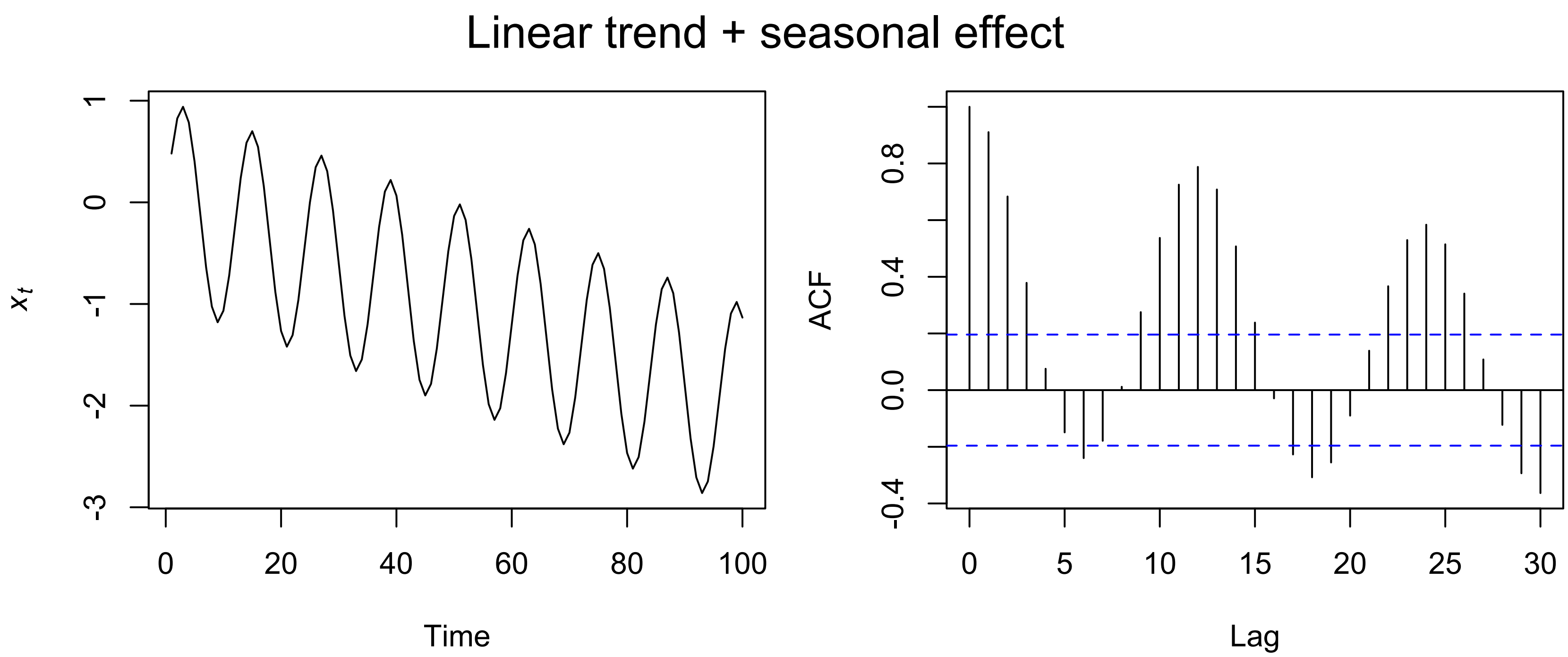
# Correlogram for deterministic trend



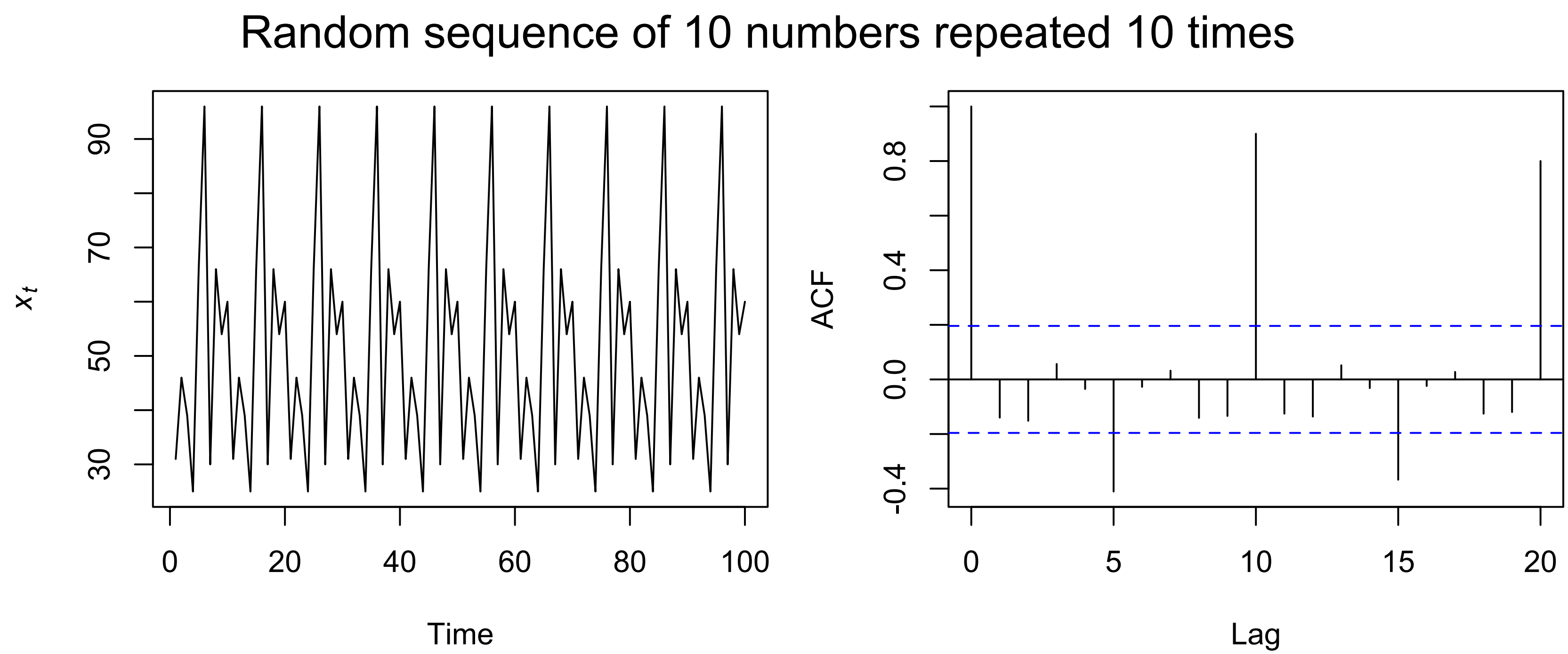
# Correlogram for sine wave



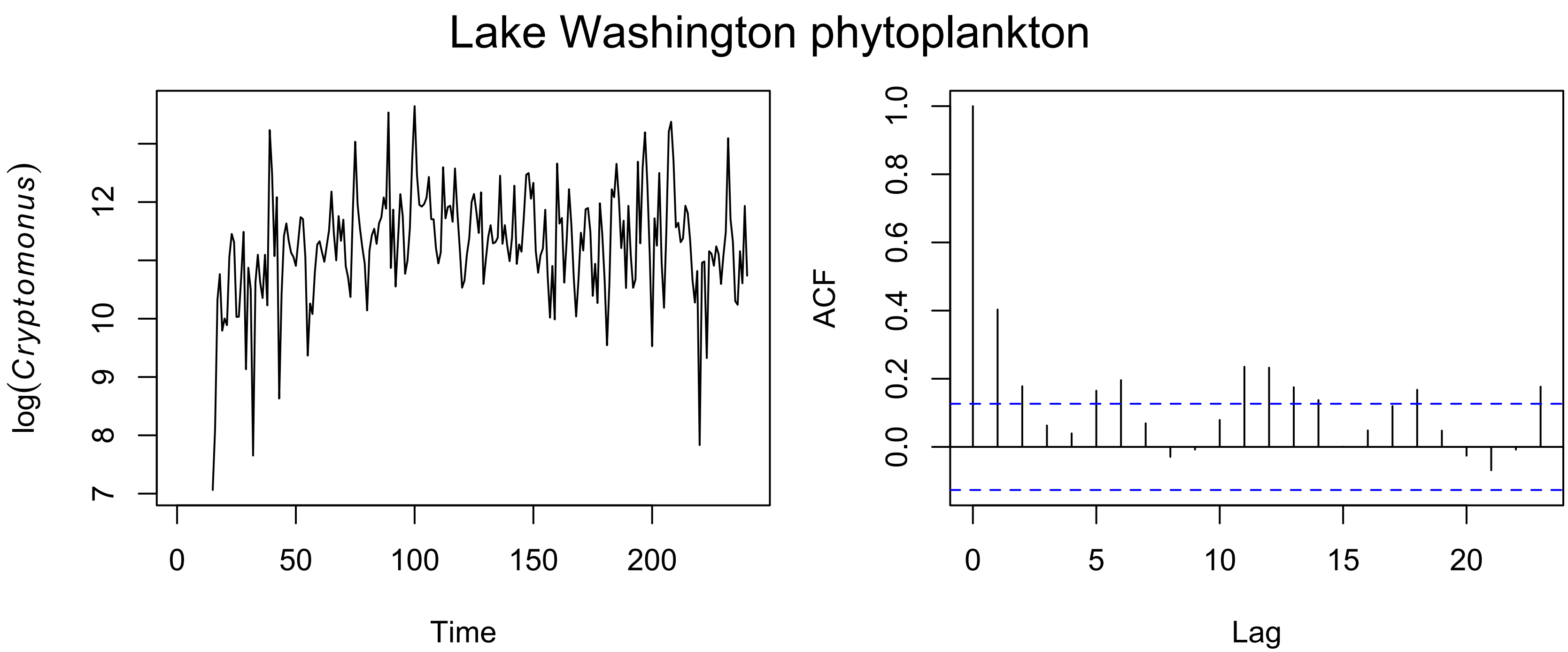
# Correlogram for trend + season



# Correlogram for random sequence



# Correlogram for real data



# Partial autocorrelation function

- The partial *autocorrelation function* (PACF) measures the linear correlation of a series $x_t$ and $x_{t+k}$ with the linear dependence of $\{x_{t-1}, x_{t-2}, \ldots, x_{t-(k-1)}\}$ removed

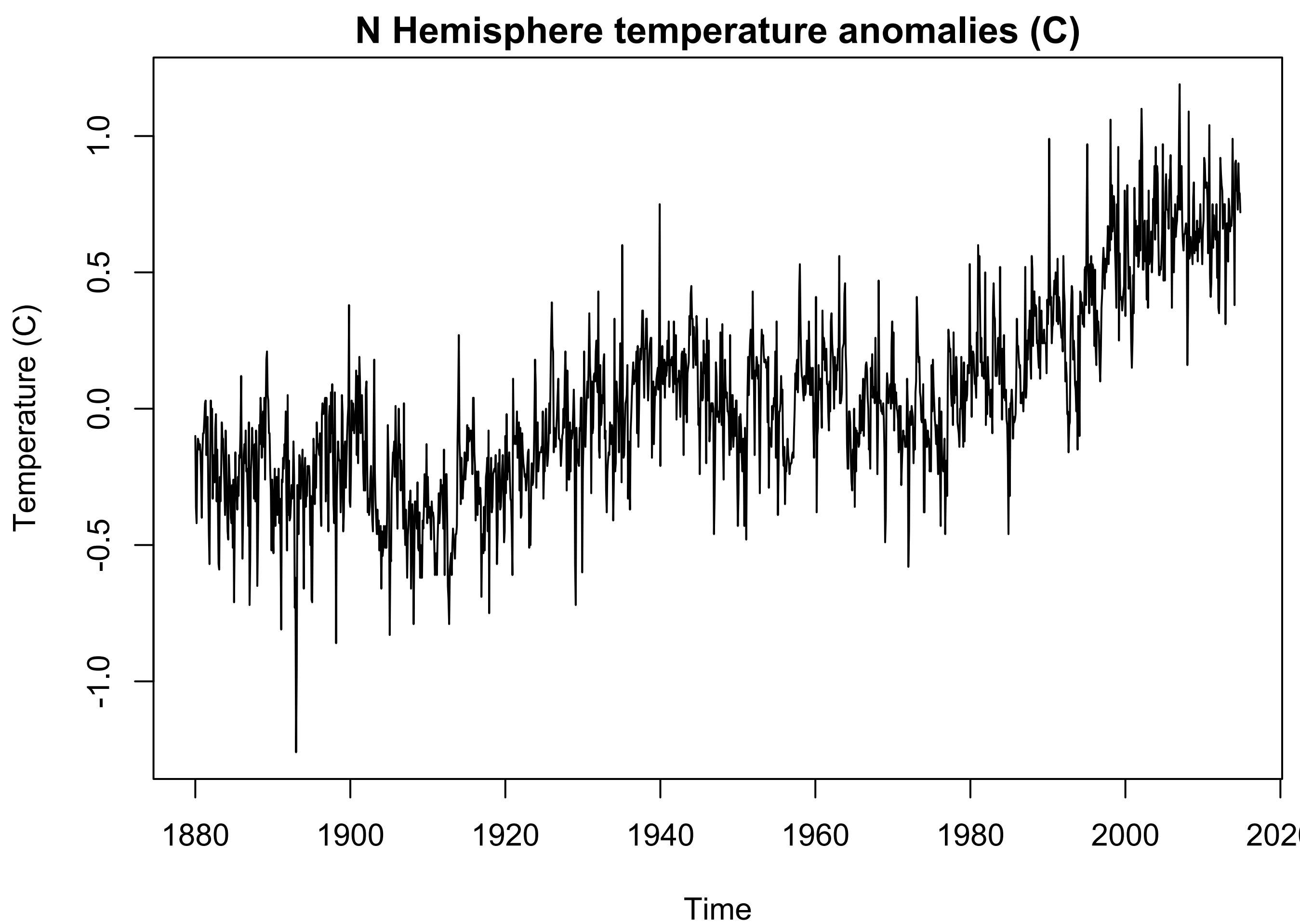
- It is defined as

$$\phi_{kk} = \begin{cases} \mathrm{Cor}\left(x_1, x_0\right) = \rho(1) & \text{if } k = 1 \\ \mathrm{Cor}\left(x_k - x_k^{k-1}, x_0 - x_0^{k-1}\right) & \text{if } k \geq 2 \end{cases} \qquad -1 \leq \phi_{kk} \leq 1$$

$$x_k^{k-1} = \beta_1 x_{k-1} + \beta_2 x_{k-2} + \cdots + \beta_{k-1} x_1$$

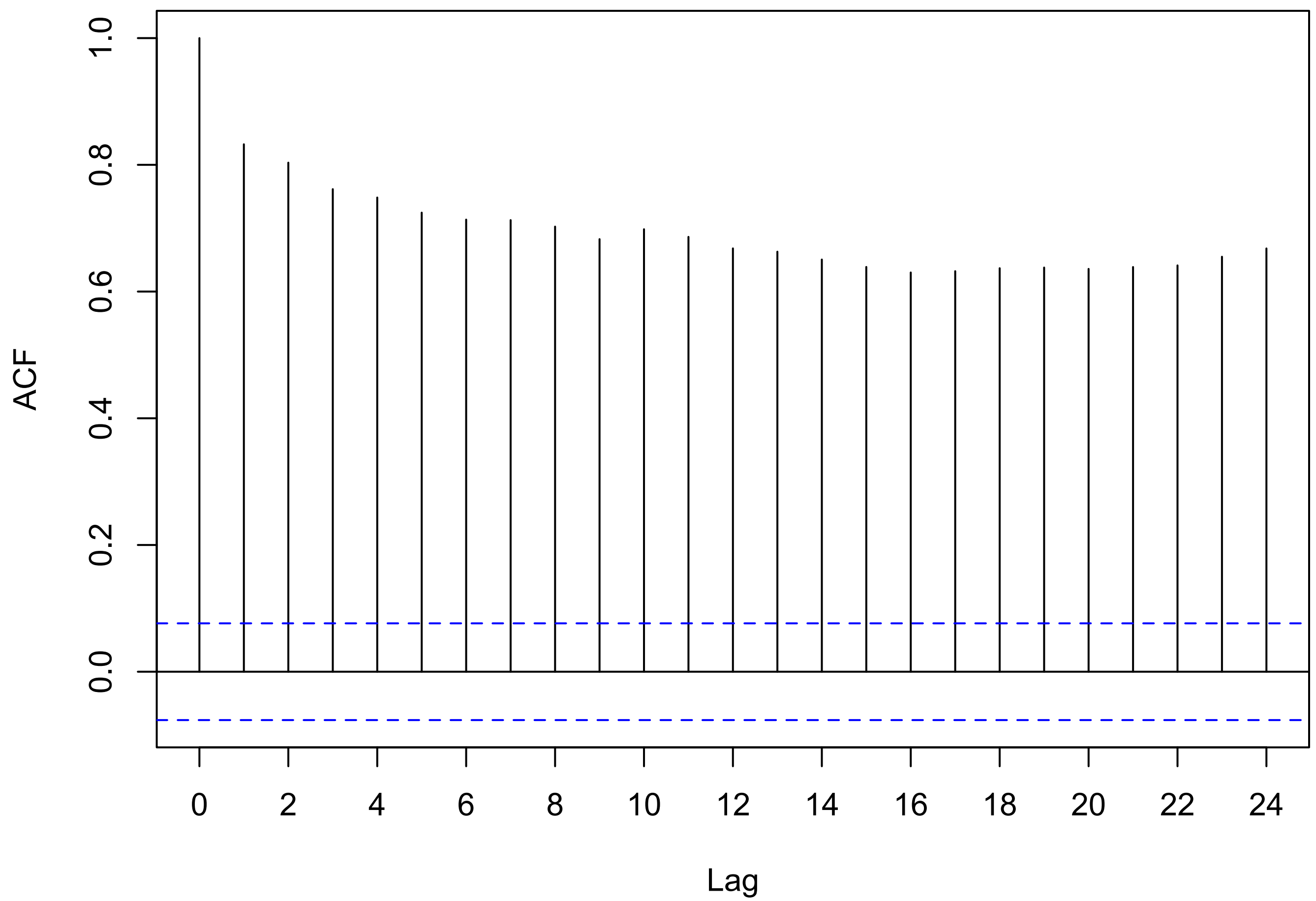
$$x_0^{k-1} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1}$$

# Revisiting the temperature ts

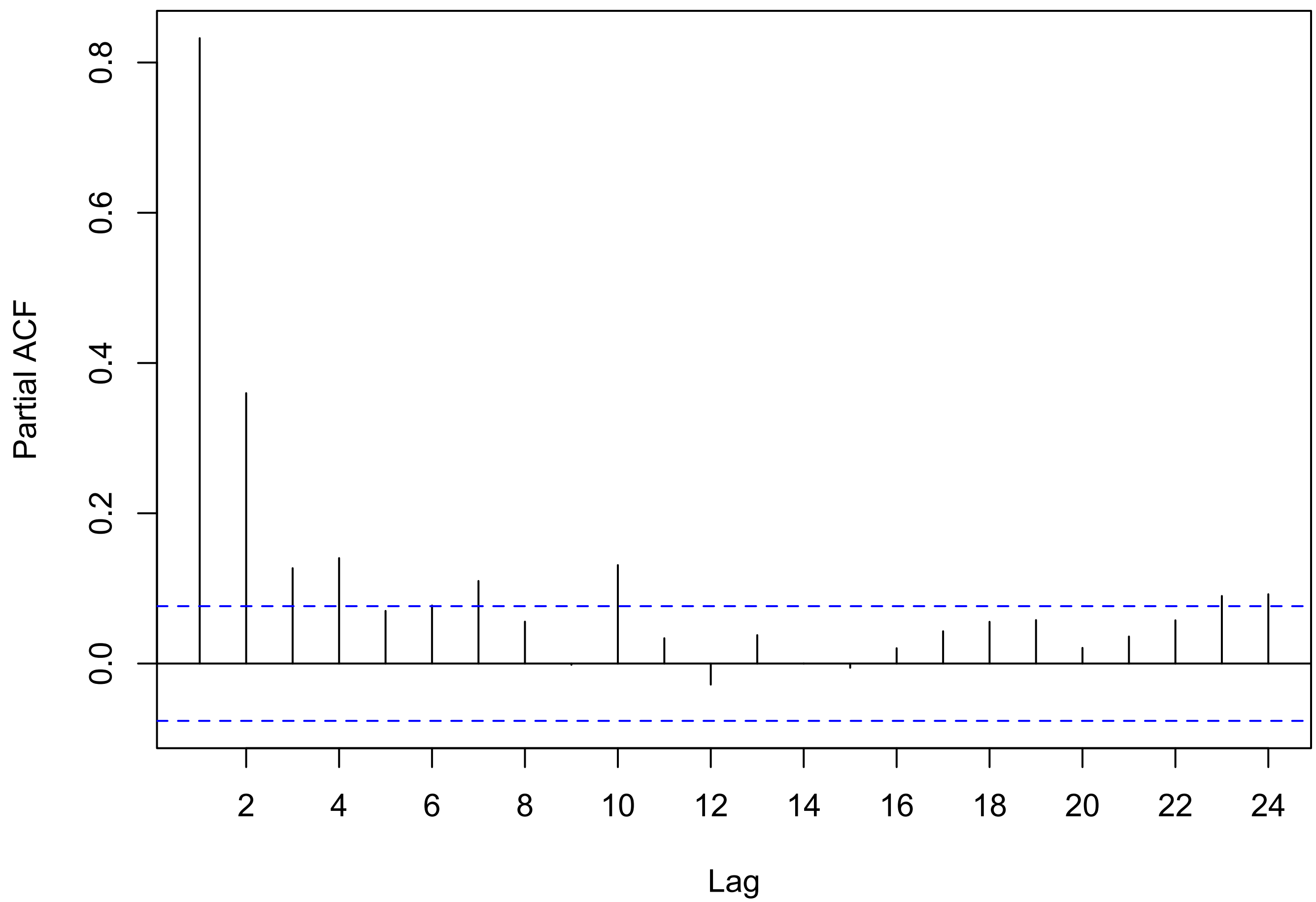



Data from http://www.ncdc.noaa.gov/

# ACF of temperature ts



# PACF of temperature ts



# Cross-covariance function (CCVF)

- Often we are interested in looking for relationships between 2 different time series
- We can extend the idea of autocovariance to examine the covariance between 2 different ts
- Define the cross-covariance function (CCVF) for x & y

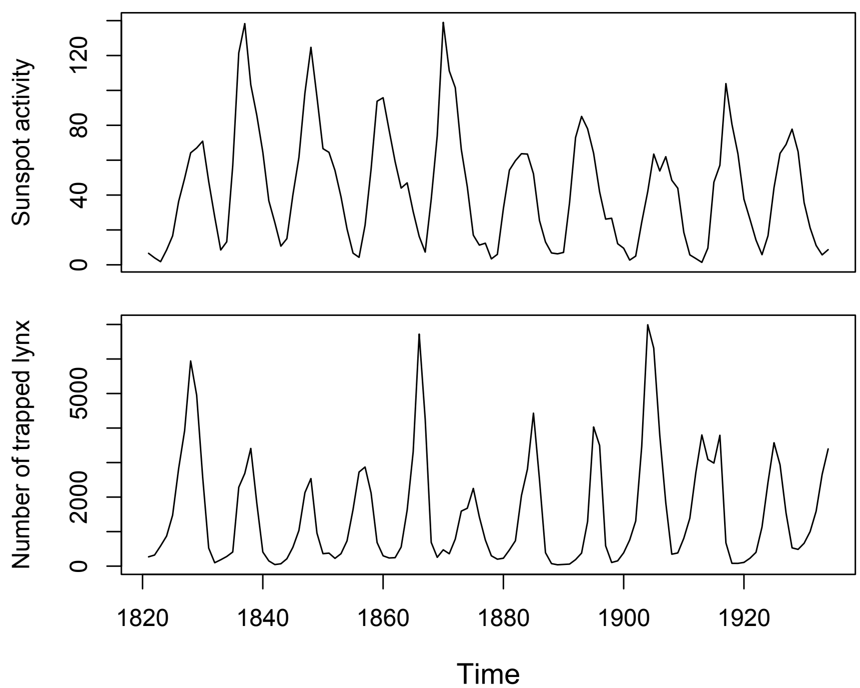
$$g_k^{xy} = \frac{1}{n}\sum_{t=1}^{n-k}(y_t - \bar{y})(x_{t+k} - \bar{x})$$

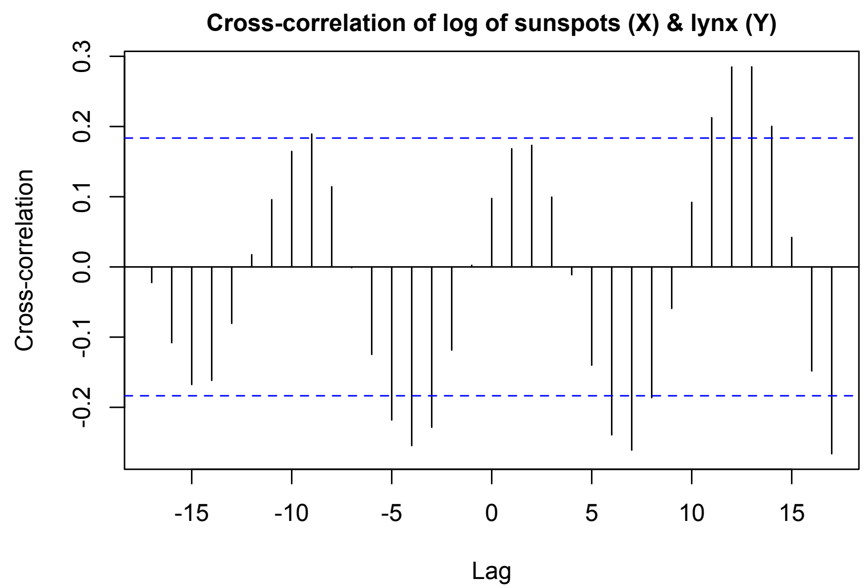
# Cross-correlation function (CCF)

- The *cross-correlation function* (CCF) is the CCVF normalized by standard deviations of *x* & *y*

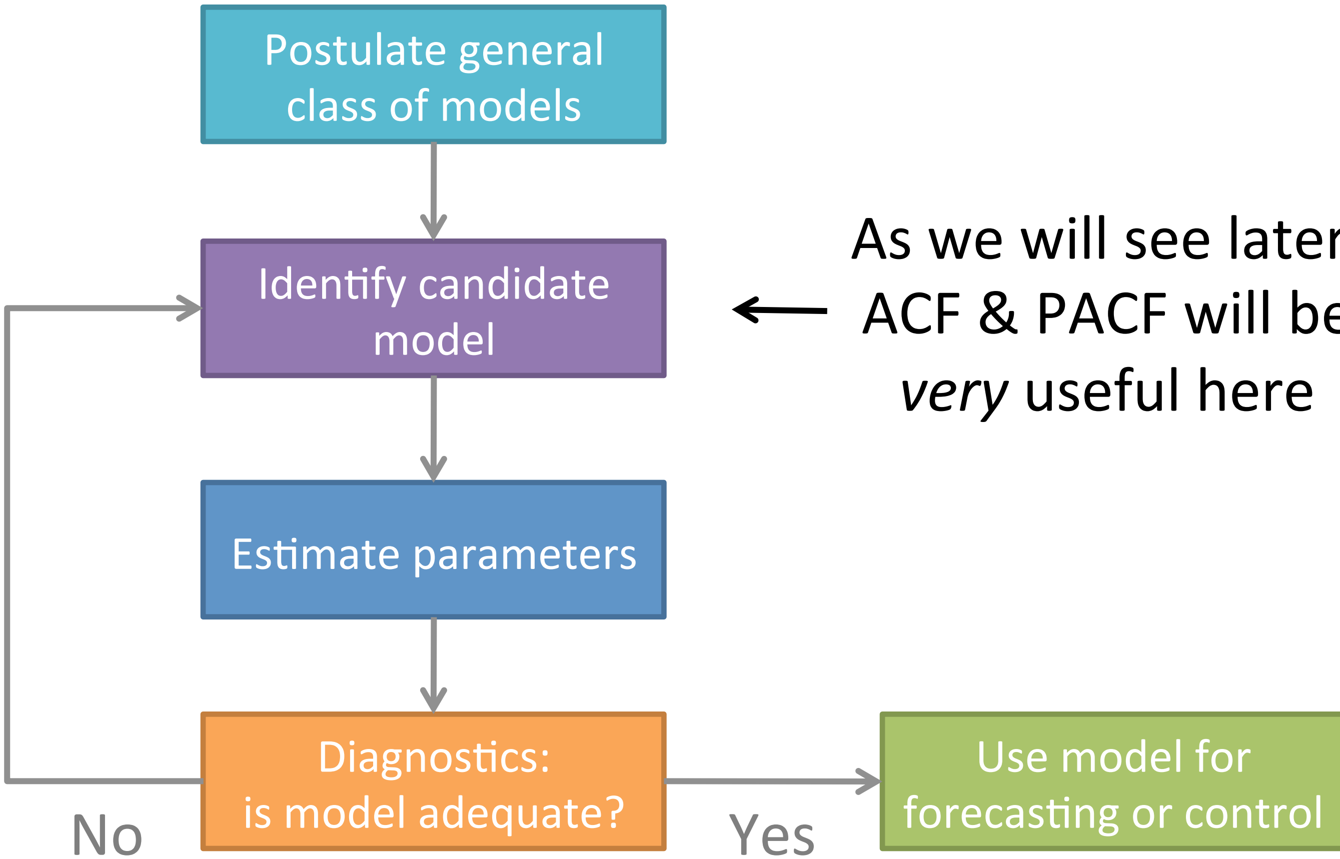
$$r_k^{xy} = \frac{g_k^{xy}}{\sqrt{\mathrm{SD}_x \mathrm{SD}_y}}$$

# CCF for sunspots and lynx



# Iterative approach to model building




As we will see later, ACF & PACF will be *very* useful here

# White noise (WN)

A time series $\{w_t : t = 1,2,3,\ldots,n\}$ is *discrete white noise* if the variables $w_1, w_2, w_3, \ldots, w_n$ are

1) *independent*, and
2) *identically distributed* with a mean of zero

*Note*: At this point we are making **no** assumptions about the distributional form of $\{w_t\}$!

For example, $w_t$ might be distributed as

- DiscreteUniform({-2,-1,0,1,2})
- Normal(0,1)

# White noise (WN)

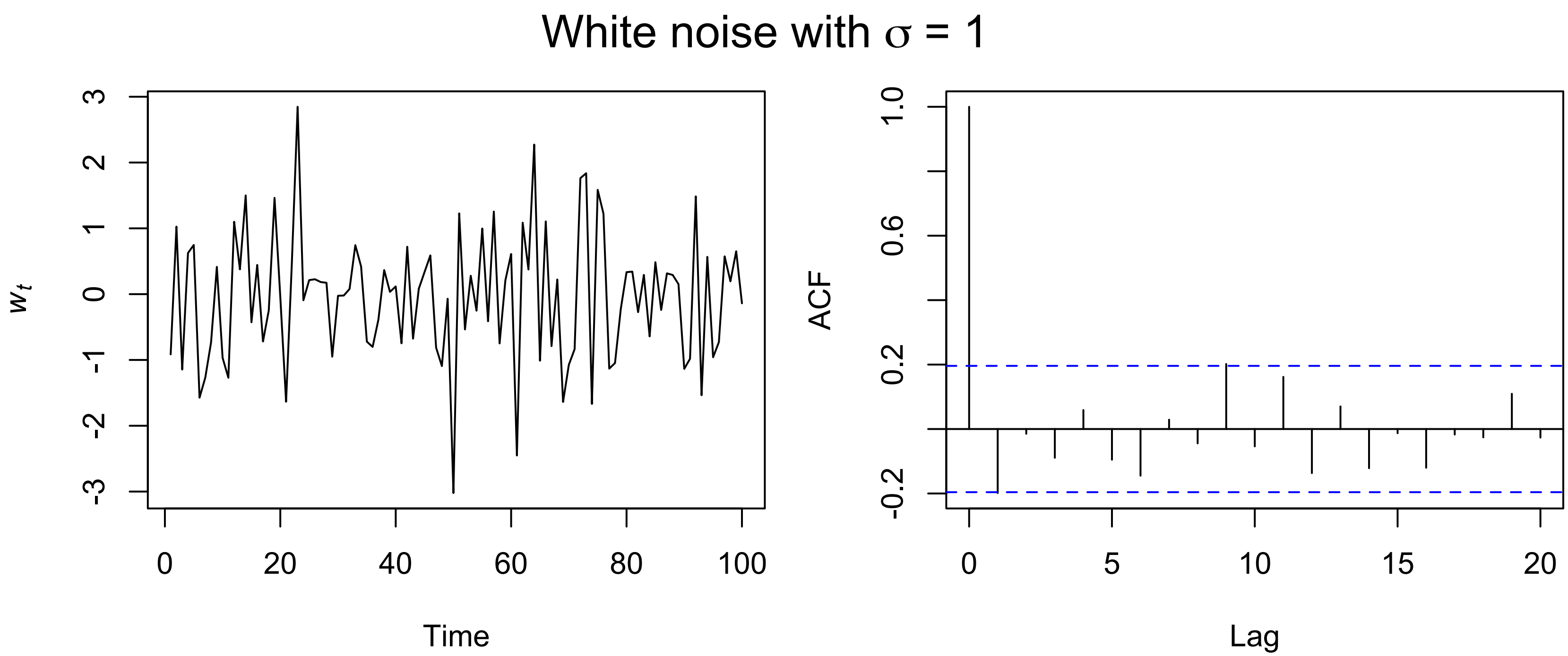
A time series $\{w_t : t = 1,2,3,\ldots,n\}$ is *discrete white noise* if the variables $w_1, w_2, w_3, \ldots, w_n$ are

1) *independent*, and
2) *identically distributed* with a mean of zero

Gaussian WN has the following 2nd-order properties:

$$\mu_w = 0 \qquad \gamma_k = \begin{cases} \sigma^2 & \text{if } k = 0 \\ 0 & \text{if } k \neq 0 \end{cases} \qquad \rho_k = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{if } k \neq 0 \end{cases}$$

# White noise


White noise with σ = 1
$w_t$
Time
0
20
40
60
80
100
3
2
1
0
-1
-2
-3
ACF
Lag
0
5
10
15
20
1.0
0.6
0.2
-0.2

# Random walk (RW)

A time series $\{x_t : t = 1,2,3,\ldots,n\}$ is a *random walk* if

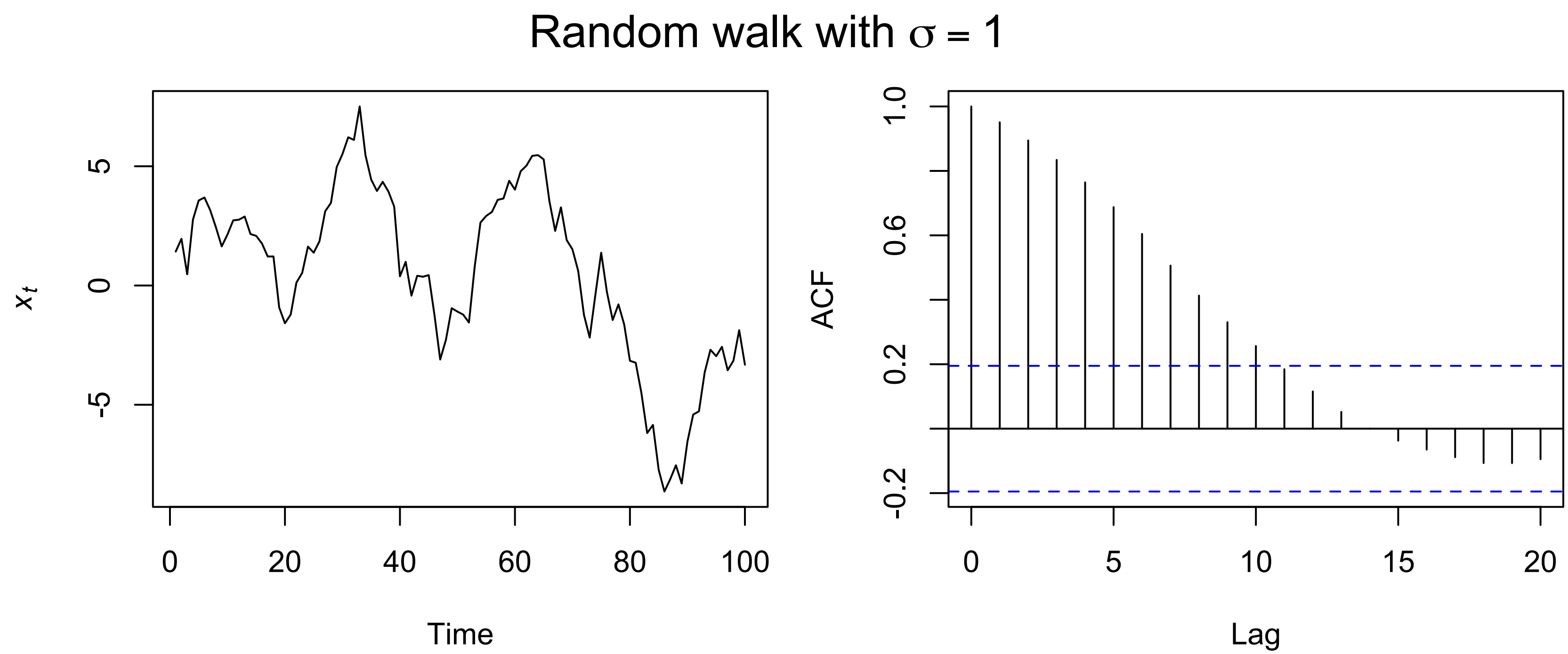
1) $x_t = x_{t-1} + w_t$, and
2) $w_t$ is white noise

RW has the following 2nd-order properties:

$$\mu_w = 0 \qquad \gamma_k(t) = t\sigma^2 \qquad \rho_k(t) = \frac{t\sigma^2}{\sqrt{t\sigma^2 (t+k)\sigma^2}} = \frac{1}{\sqrt{1 + k/t}}$$

*Note*: Random walks are NOT stationary!

# Random walk (RW)



# The backward shift operator (B)

- Define the *backward shift operator* by

$$\mathbf{B}x_t = x_{t-1}$$

- Or, more generally as

$$\mathbf{B}^k x_t = x_{t-k}$$

- So, RW model can be expressed as

$$x_t = \mathbf{B}x_t + w_t$$

$$(1-\mathbf{B})x_t = w_t$$

$$x_t = (1-\mathbf{B})^{-1} w_t$$

# The difference operator ( $\nabla$ )

- Define the first *difference operator* as

$$\nabla x_t = x_t - x_{t-1}$$

- So, first differencing a RW model yields WN

$$\nabla\left(x_t = x_{t-1} + w_t\right)$$

$$x_t - x_{t-1} = x_{t-1} - x_{t-1} + w_t$$

$$x_t - x_{t-1} = w_t$$

# The difference operator ( $\nabla$ )

- Differences of order *d* are then defined by

$$\nabla^d = (1 - B)^d$$

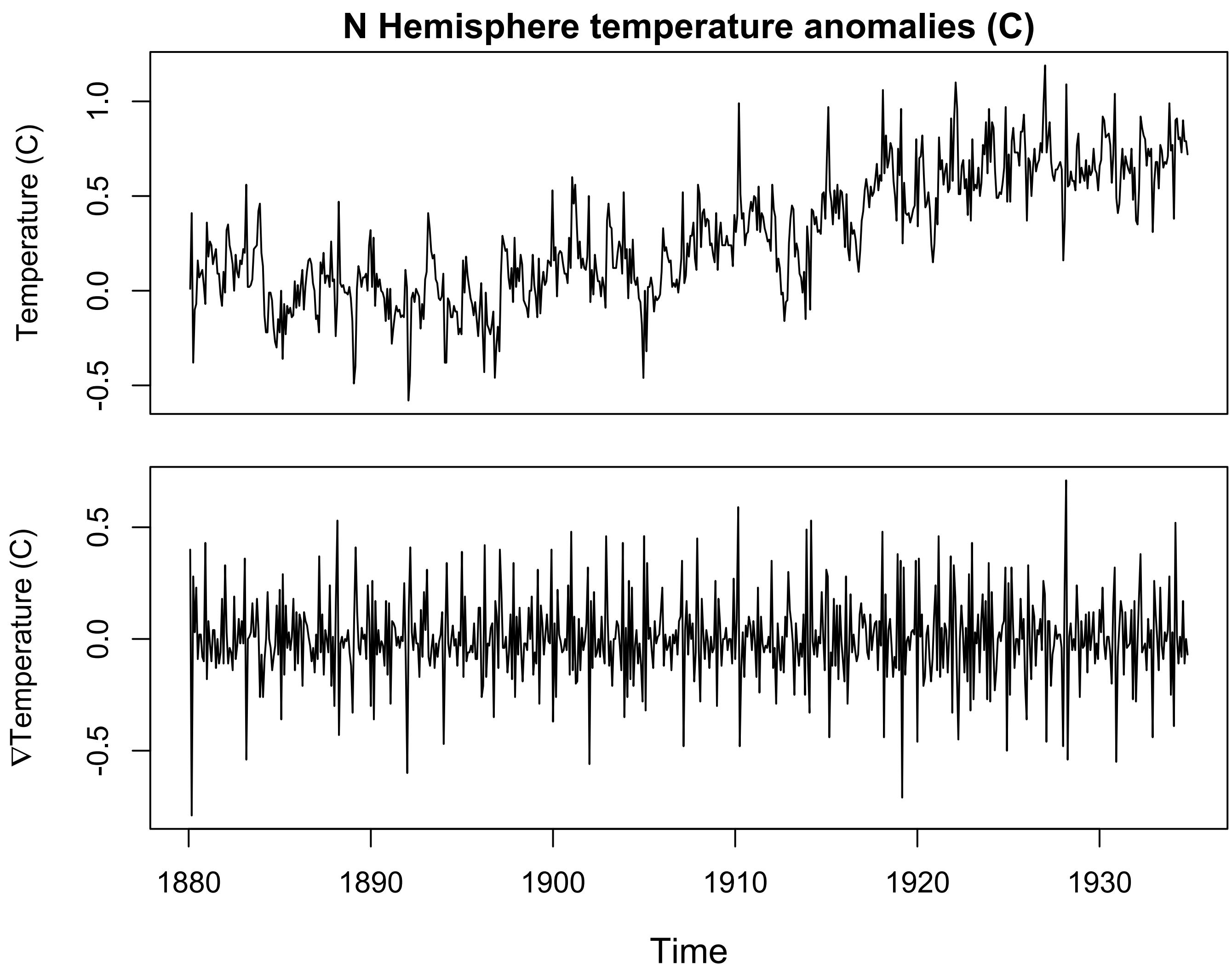
- For example, twice differencing a ts

$$\nabla^2 x_t = (1 - B)^2 x_t$$

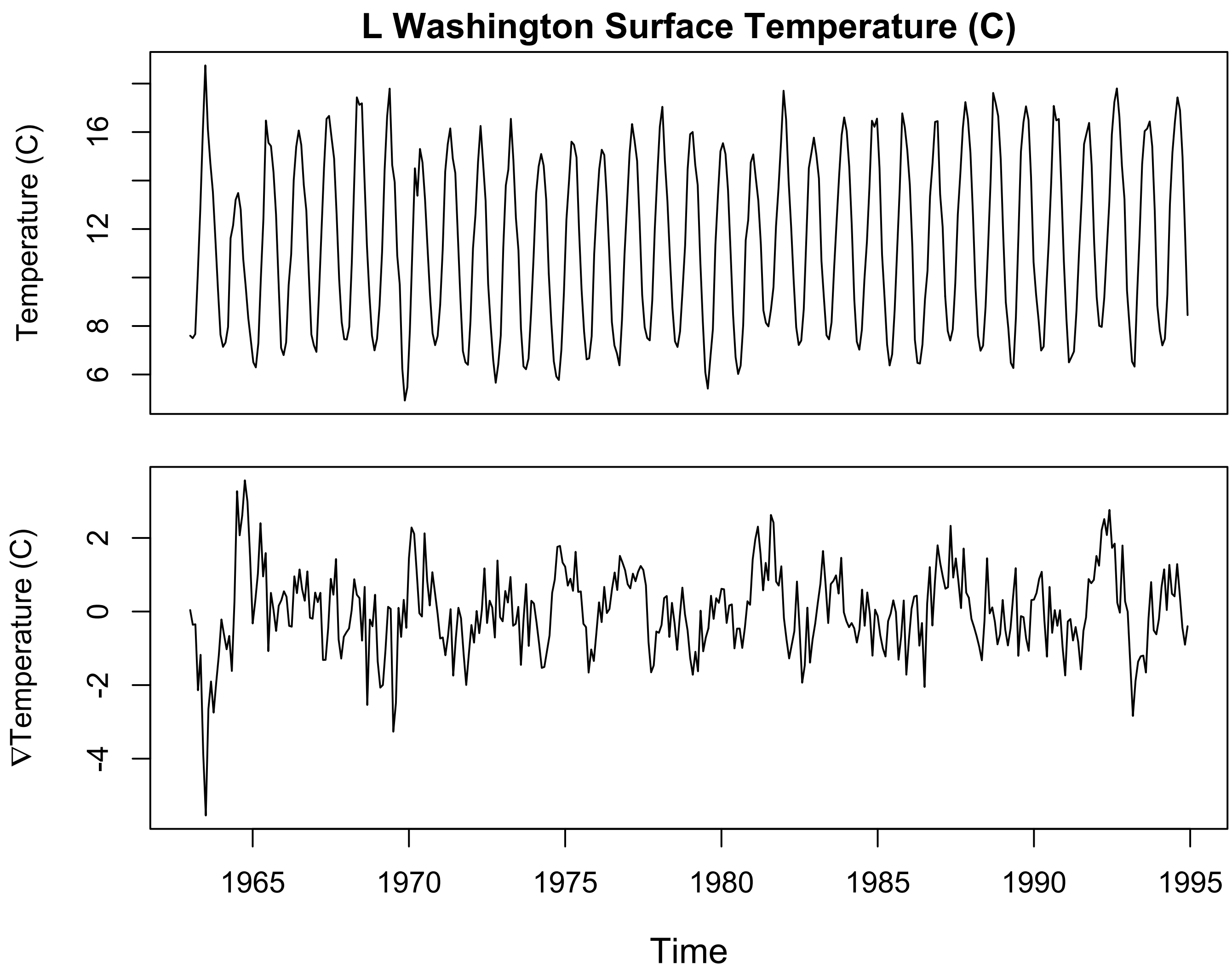
# Difference to remove trend/season

- Differencing is a very simple means for removing a trend or seasonal effect
- The 1$^{st}$-difference removes a linear trend, a 2$^{nd}$-difference would remove a quadratic trend, etc.
- For seasonal data, using a 1$^{st}$-difference with *lag = period* removes both trend & seasonal effects
- Pro: no parameters to estimate
- Con: no estimate of stationary process

# First-difference to remove trend



# First-difference* to remove season




*At lag = 12

# Topics for today

- Expectation, mean & variance
- Covariance & correlation
- Stationarity
- Autocovariance & autocorrelation
- Correlograms
- White noise
- Random walks
- Backshift & difference operators