Time Series Analysis

Theory: Time Series Analysis

Probability distribution

Correlation and Autocorrelation

Spectrum and spectral analysis

Autoregressive-Moving Average (ARMA) modeling

Spectral analysis -- smoothed periodogram method

Detrending, Filtering and Smoothing

Laboratory exercises:

.....

Applied Time Series Analysis Course. David M. Meko, University of Arizona. Laboratory of Tree-Ring Research, , dmeko@LTRR.arizona.edu

Romà Tauler (IDAEA, CSIC, Barcelona)

Probability distribution

Probability density function or probability function

Probability function (also called *probability density function*, ***pdf***)
The probability function of the random variable X , denoted by $f(x)$ is the function that gives the probability of X taking the value x , for any real number x :

$$f(x) = P(X=x)$$

The most commonly used theoretical distribution is the normal distribution. Its probability density function (pdf) is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{(x-\mu}{2\sigma^2}\right]$$

where μ and σ are the population mean and standard deviation of X .
The standard normal distribution is the normal distribution with μ equal to 0 and σ equal to 1

Probability distribution

Cumulative distribution function (cdf)or distribution function

The ***distribution function*** of a random variable X is the function that gives the probability of X being less than or equal to a real number x :

$$F(x) = p(X \leq x) = \sum_{\mu<x} f(u)$$

For a theoretical distribution, the cdf can be computed as the integral of the probability density function.

The *empirical distribution function* $S(x)$ is a function of x, which equals the decimal fraction of the observations that are less than or equal to x_c for $-\infty < x_c < \infty$

Probability distribution

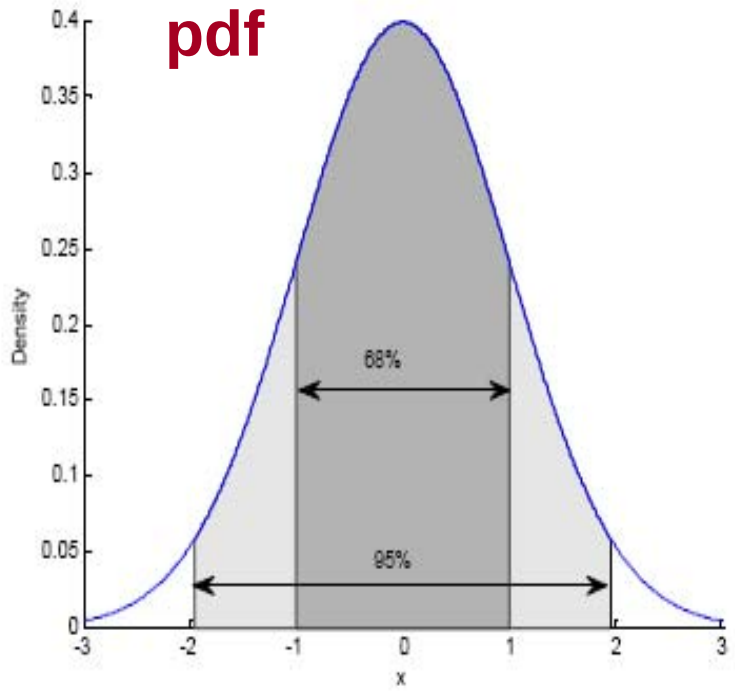


Figure2.2. Probability density function of standard normal distribution. Sixty-eight percent of the population is within ±1.0 of zero. Ninety-five percent is within ±1.96 of zero.

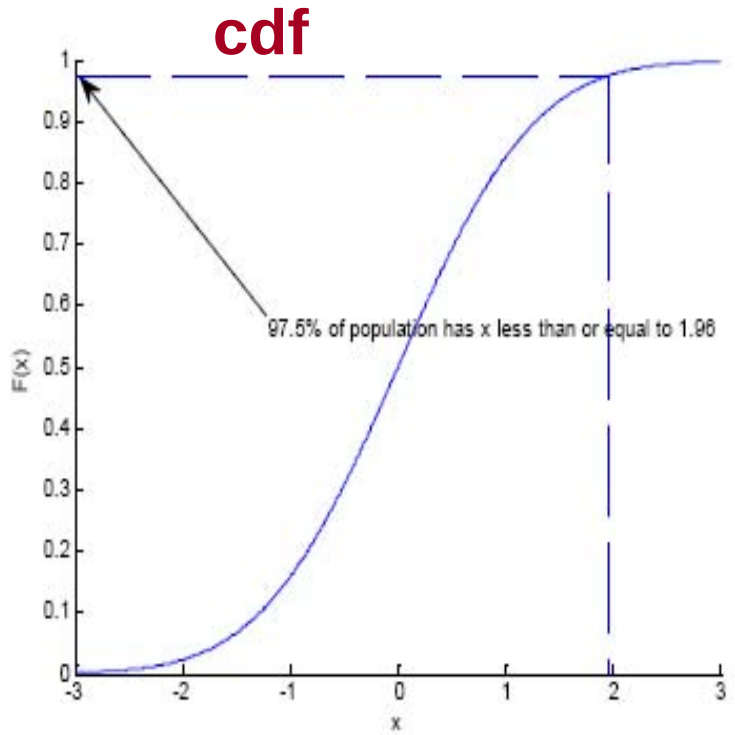
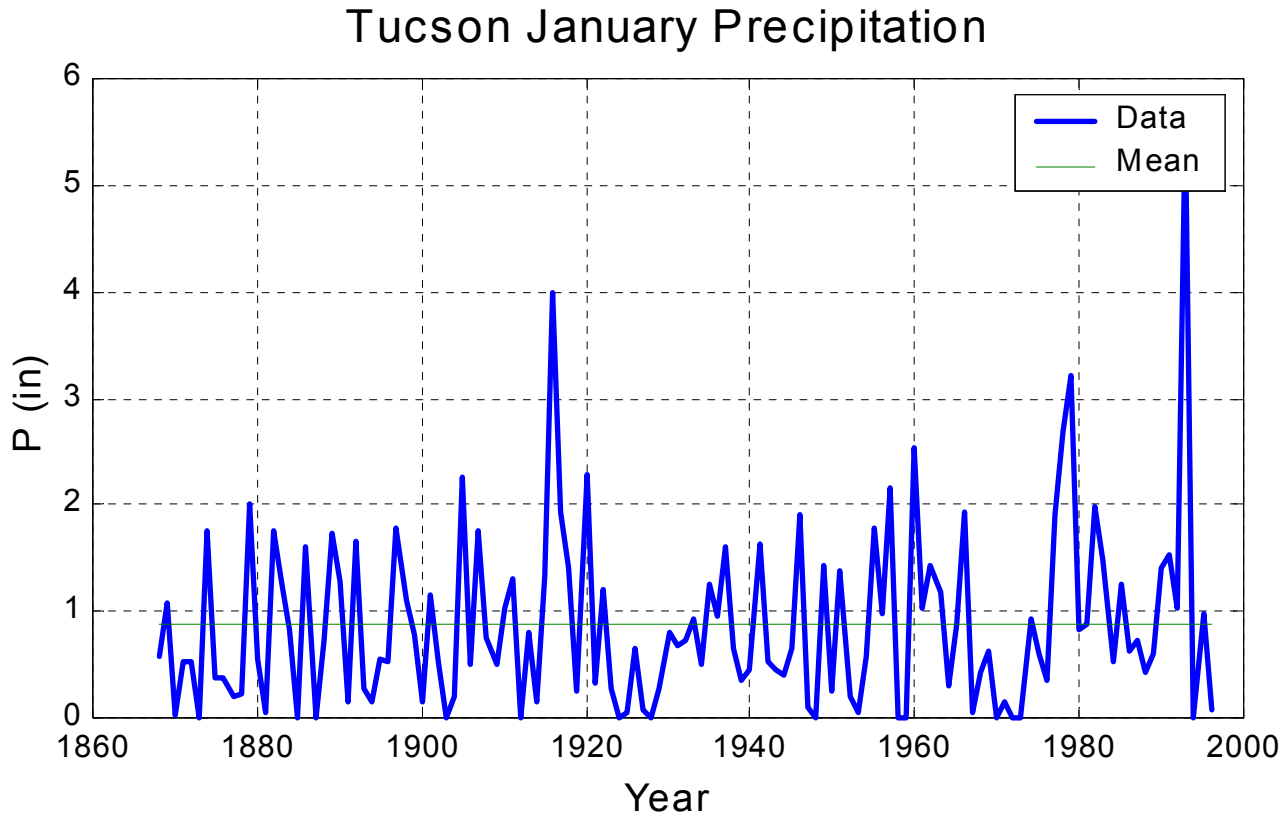


Figure2.3. Cumulative distribution function (cdf) of standard normal distribution.

Probability distribution
Time Series



A time series is a set of observations ordered in time, observed at a discrete set of evenly spaced time intervals: x_t at times t =1,2,..., N , where N is the length of the time series.

Probability distribution
Time Series

- A time series is broadly defined as any series of measurements taken at different times. Some distinctive properties of time series include
  - 1) continuous vs discrete,
  - 2) univariate vs multivariate,
  - 3) evenly sampled vs unevenly sampled,
  - 4) periodic vs aperiodic,
  - 5) stationary vs nonstationary,
  - 6) short vs long.

- These properties, as well as the sampling interval and temporal overlap of multiple series, must be considered in selecting a dataset for analysis

Probability distribution
Time Series

Probability distribution

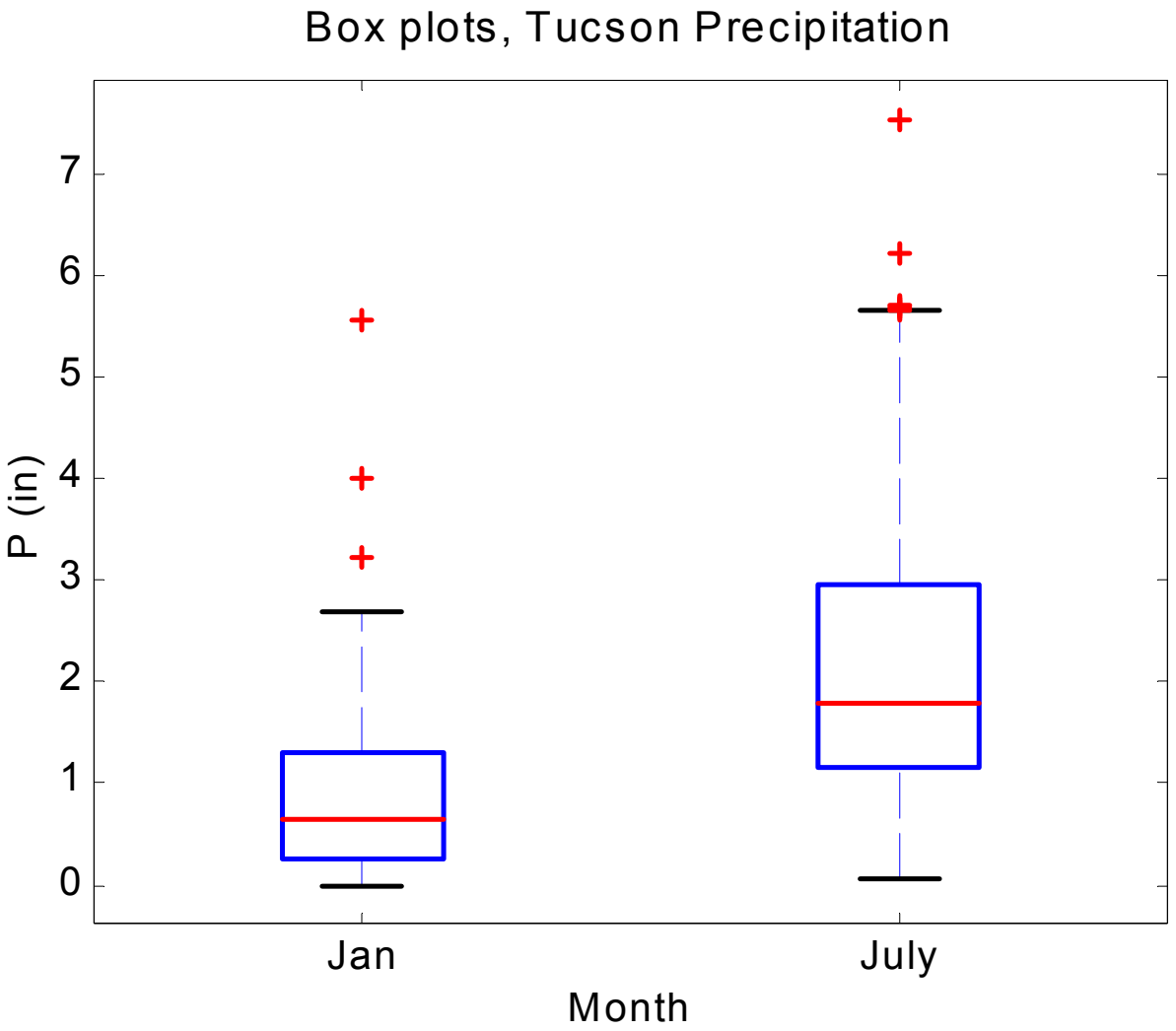
•The probability distribution of a time series describes the probability that an observation falls into a specified range of values.

•An empirical probability distribution for a time series can be arrived at simply by sorting and ranking the values of the series.

•Quantiles and percentiles are useful statistics that can be taken directly from the empirical probability distribution.

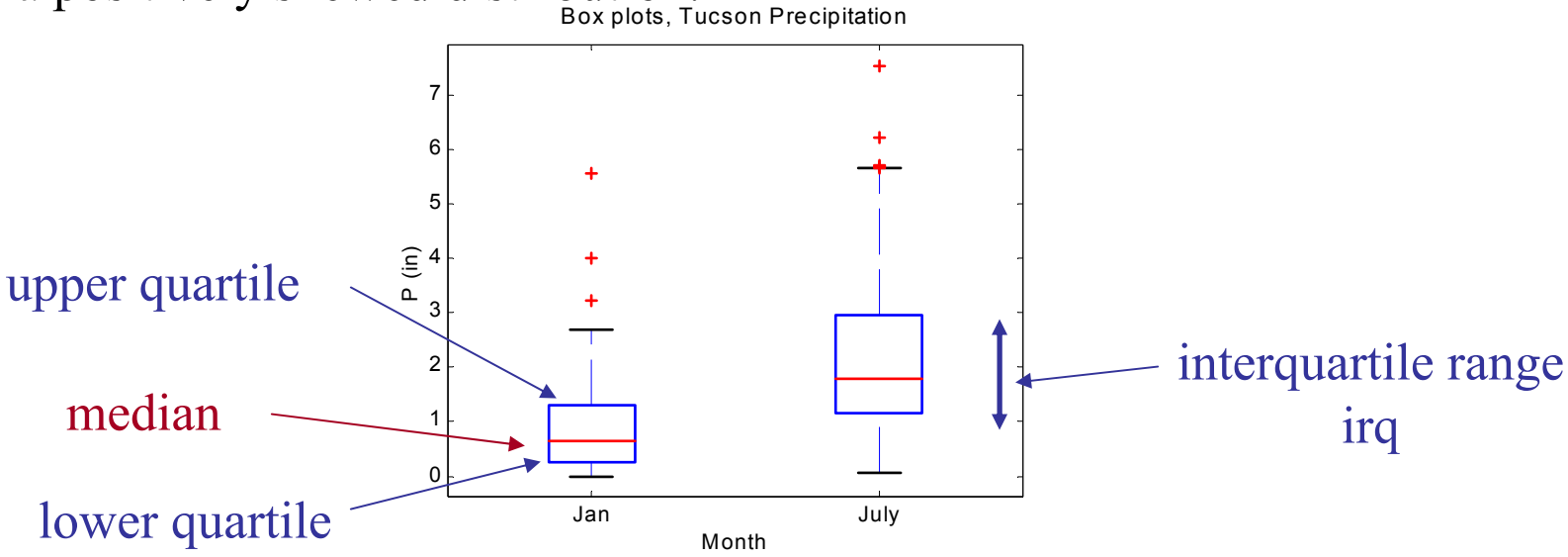
•Many parametric statistical tests assume the time series is a sample from a population with a particular population probability distribution. Often the population is assumed to be normal.

Probability distributions: Box plots



Probability distributions: Box plots

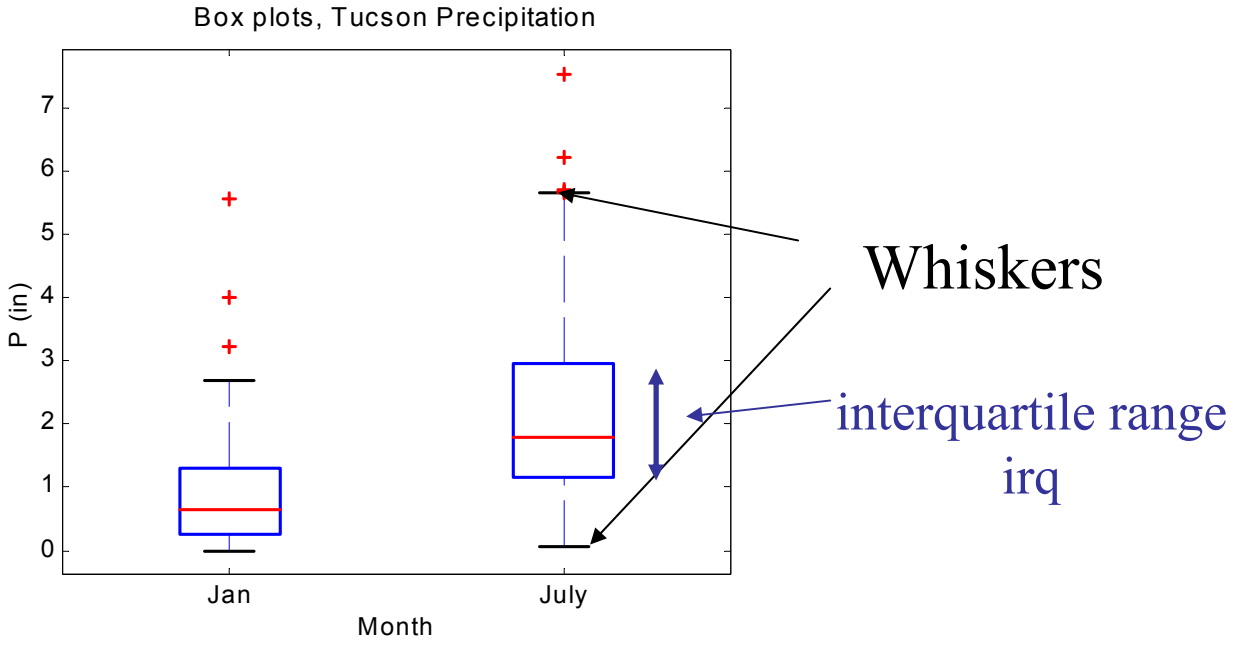
A **box plot** summarizes the information on the data distribution primarily in terms of the median, the upper quartile, and lower quartile. The "box" by definition extends from the upper to lower quartile. Within the box is a dot or line marking the median. The width of the box, or the distance between the upper and lower quartiles, is equal to the interquartile range, and is a measure of spread. The median is a measure of location, and the relative distances of the median from the upper and lower quartiles is a measure of symmetry "in the middle" of the distribution. is defined by the upper and lower quartiles. A line or dot in the box marks the median. For example, the median is approximately in the middle of the box for a symmetric distribution, and is positioned toward the lower part of the box for a positively skewed distribution.



Probability distributions: Box plots

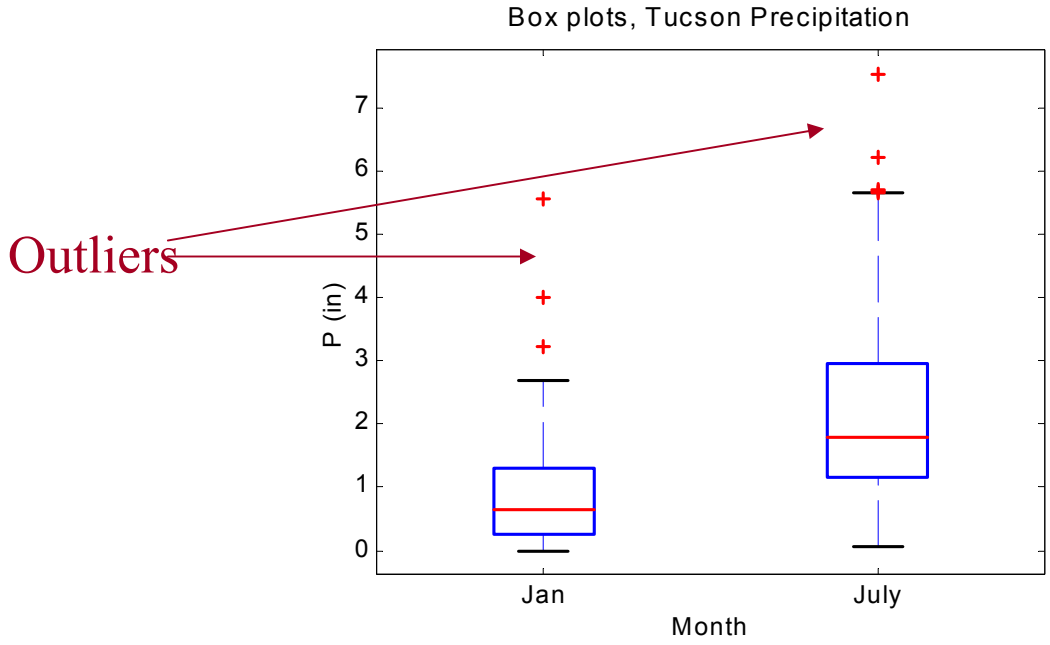
"Whiskers" are drawn outside the box at what are called the the "adjacent values." The upper adjacent value is the largest observation that does not exceed the upper quartile plus 1.5 iqr , where iqr is the interquartile range. The lower adjacent value is the smallest observation than is not less than the lower quartile minus1.5 iqr . If no data fall outside this 1.5 iqr buffer around the box, the whiskers mark the data extremes. The whiskers also give information about symmetry in the tails of the distribution.

For example, if the distance from the top of the box to the upper whisker exceeds the distance from the bottom of the box to the lower whisker, the distribution is positively skewed in the tails. Skewness in the tails may be different from skewness in the middle of the distribution. For example, a distribution can be positively skewed in the middle and negatively skewed in the tails.



Probability distributions: Box plots

Any points lying outside the 1.5 iqr around the box are marked by individual symbols as “outliers”. These points are outliers in comparison to what is expected from a normal distribution with the same mean and variance as the data sample. For a standard normal distribution, the median and mean are both zero, and: q at 0.25 = −0.67449, q at 0.75 =0.67449, iqr = q 0.75 − q 0.25 =1.349, where q 0.25and q. 075are the first and third quartiles, and iqr is the interquartile range. We see that the whiskers for a standard normal distribution are at data values: Upper whisker = 2.698 , Lower whisker = -2.698

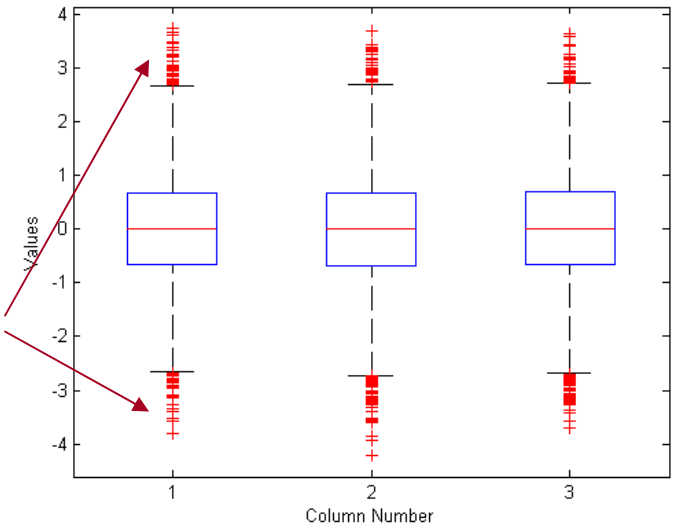


Probability distributions: Box plots

From the cdf of the standard normal distribution, we see that the probability of a lower value than x=−2.698 is 0.00035.
This result shows that for a normal distribution, roughly 0.35 percent of the data is expected to fall below the lower whisker. By symmetry, 0.35 percent of the data are expected above the upper whisker. These data values are classified as outliers.
Exactly how many outliers might be expected in a sample of normally distributed data depends on the sample size. For example, with a sample size of 100, we expect no outliers, as 0.35 percent of 100 is much less than 1. With a sample size of 10,000, however, we would expect 35 positive outliers and 35 negative outliers for a normal distribution.

For a normal distribution
>> varnorm=randn(10000,3);
>> boxplot(varnorm)

0.35% of 10000
are approx. 35 outliers
at each whisker side

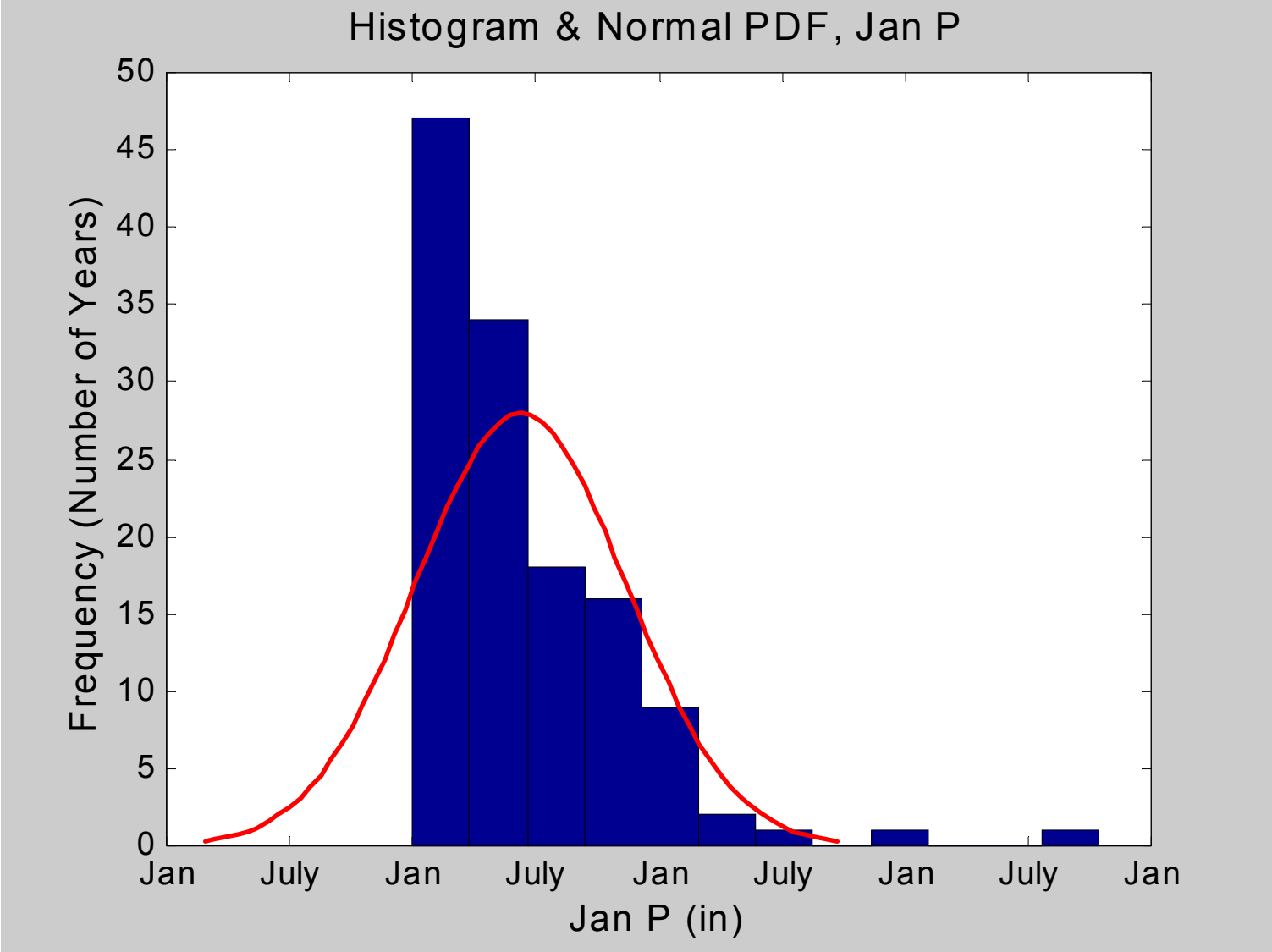


Probability distributions: Box plots

It is therefore not surprising to find some outliers in box plots of very large data sample, and the existence of a few outliers in samples much larger than 100 does not necessarily indicate lack of normality. “Notched” boxplots plotted side by side can give some indication of the significance of differences in medians of two sample.

Given a sample of data with N observations and interquartile range iqr. How wide should the notch in the box plot be for
a) 95 percent confidence interval about the median, a
b) visual assessment of whether two medians are statistically different at the 95 percent level?

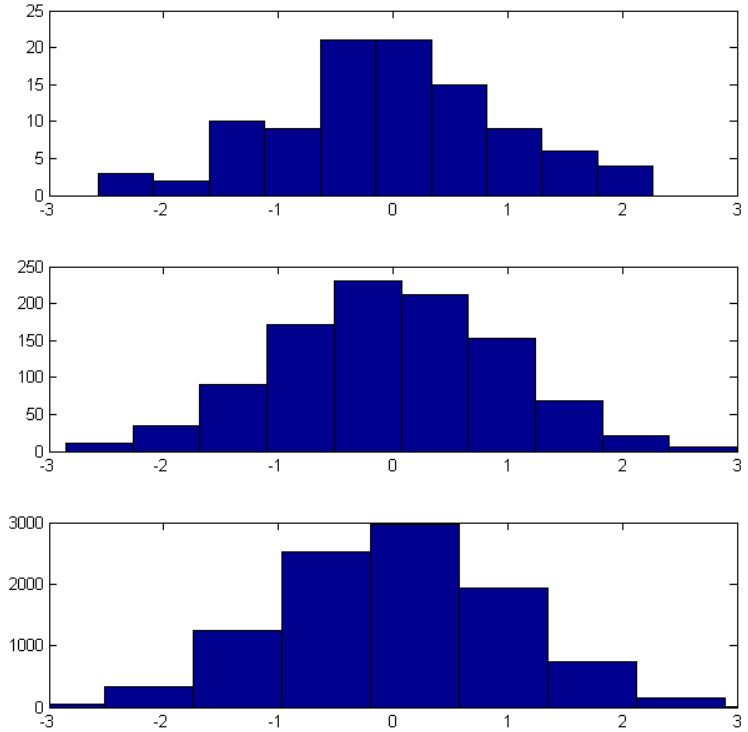
Probability distributions: Histogram plots



Probability distributions: Histogram plots

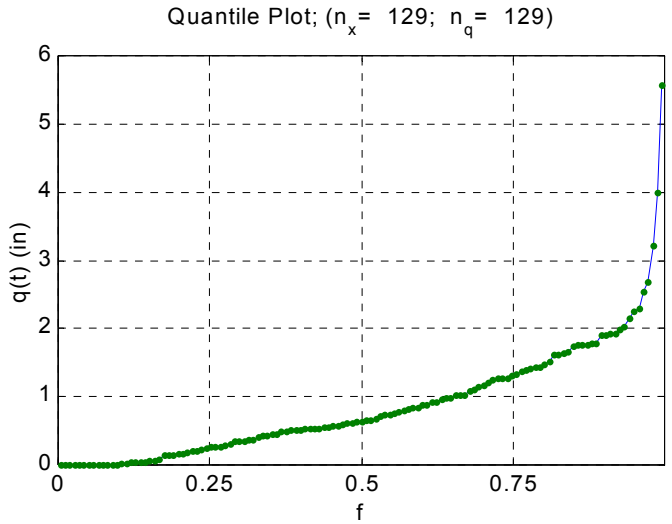
For a normal distribution

```
>> varnorm(:,1)=randn(100,1);
>> varnorm2=randn(1000,1);
>> varnorm3=randn(10000,1);
>> subplot(3,1,1),hist(varnorm1);
>> subplot(3,1,2),hist(varnorm2);
>> subplot(3,1,3),hist(varnorm3);
```



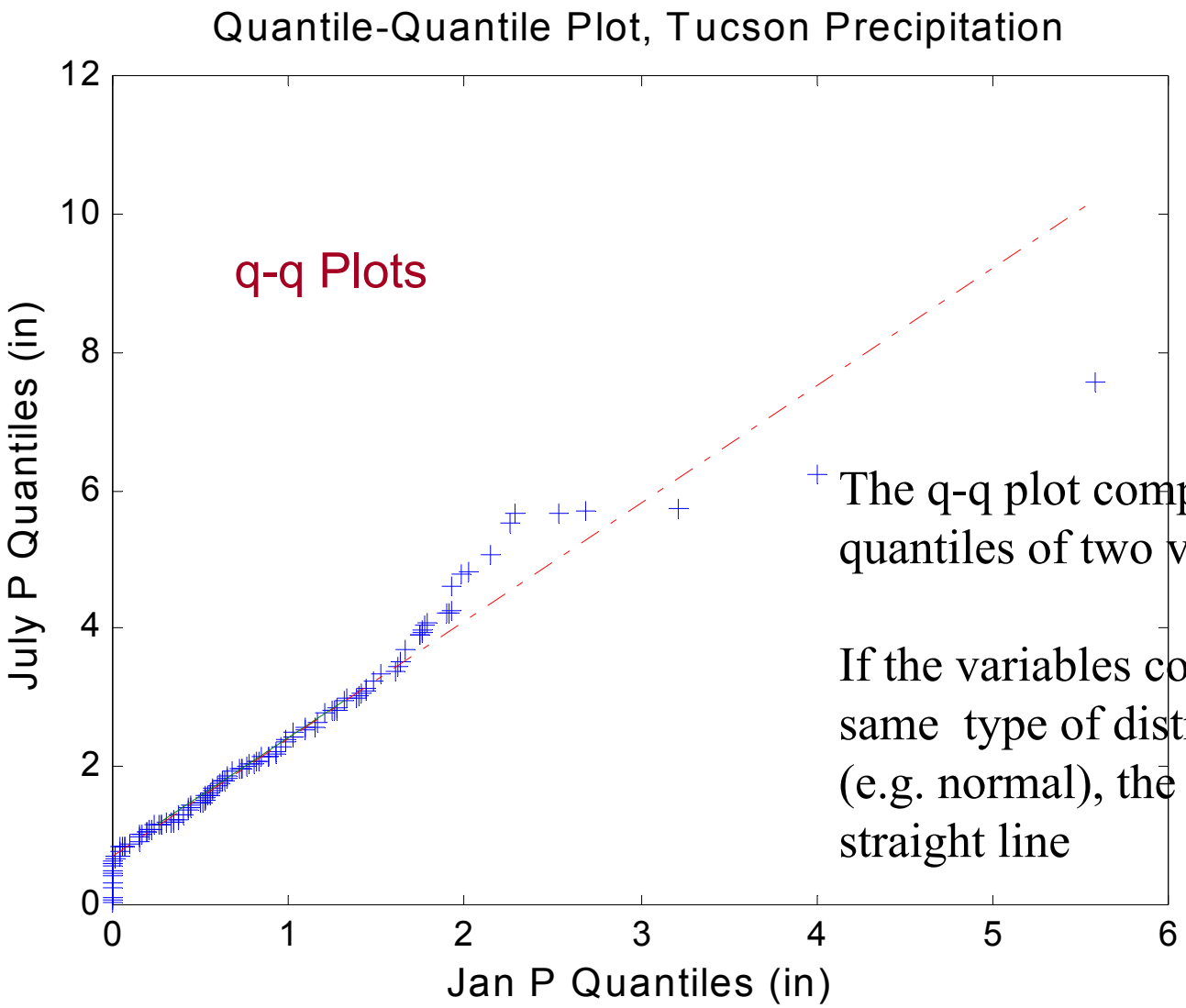
Probability distribution
Time Series

Quantile Plots. The f quantile *is the data value* below which approximately a decimal fraction f of the data is found. That data value is denoted **q(f)**. Each data point can be assigned an f-value. Let a time series x of length n be sorted from smallest to largest values, such that the sorted values have rank i =1,2,..., n The f-value for each observation is computed as $\mathbf{f_i = (i-0.5)/n.}$



Quantile plot Probability, location, spread, range, outliers

Probability distributions: q-q plots



The q-q plot compares the quantiles of two variables.

If the variables come from the same type of distribution (e.g. normal), the q-q plot is a straight line

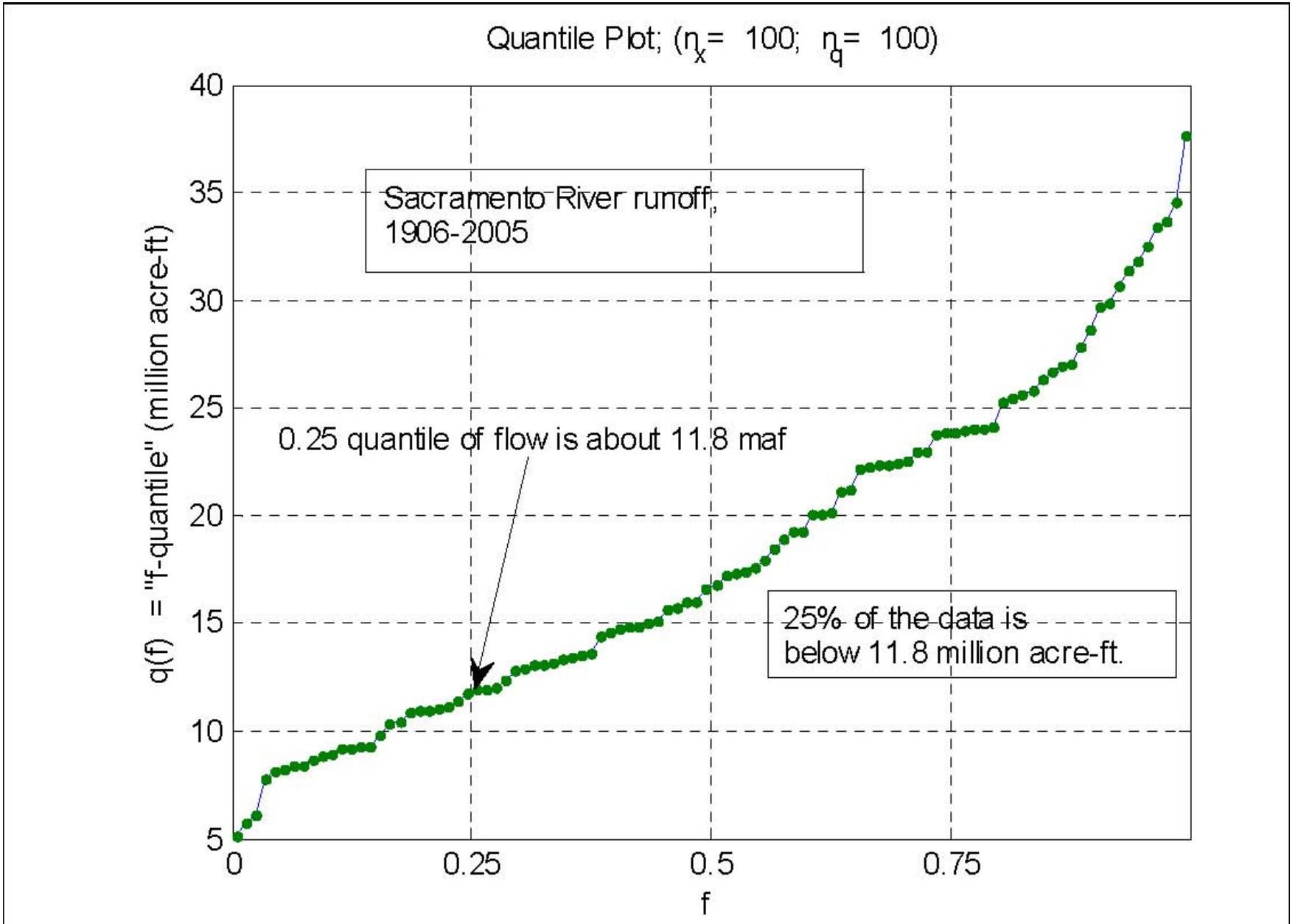
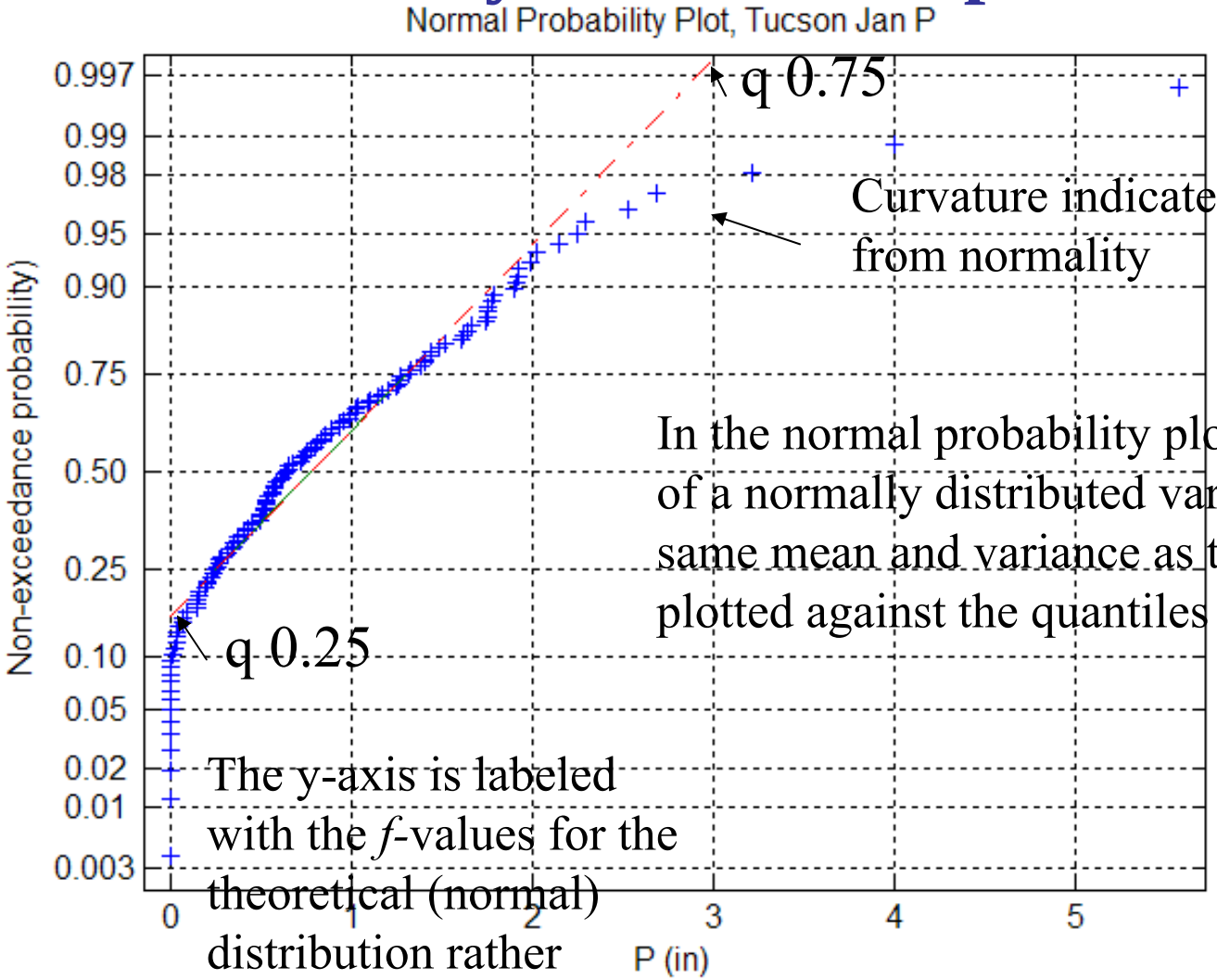


Figure 2.1. Quantile plot of natural flow of Sacramento River. Data are water-year totals for 1906-2005 in million acre-ft (maf).

Probability distributions

Normality distribution plots



Curvature indicates departure from normality

In the normal probability plot, the quantiles of a normally distributed variable with the same mean and variance as the data are plotted against the quantiles of the data.

The y-axis is labeled with the *f*-values for the theoretical (normal) distribution rather than with the quantiles of the normal variate.

Probability distribution
Distribution tests

Lilliefors test

The Lilliefors test evaluates the hypothesis that the sample has a normal distribution with unspecified mean and variance against the alternative hypothesis that the sample does not have a normal distribution. The main difference from the well-known Kolmogorov-Smirnov test (K-S test) is in the assumption about the mean and standard deviation of the normal distribution.

K-S test

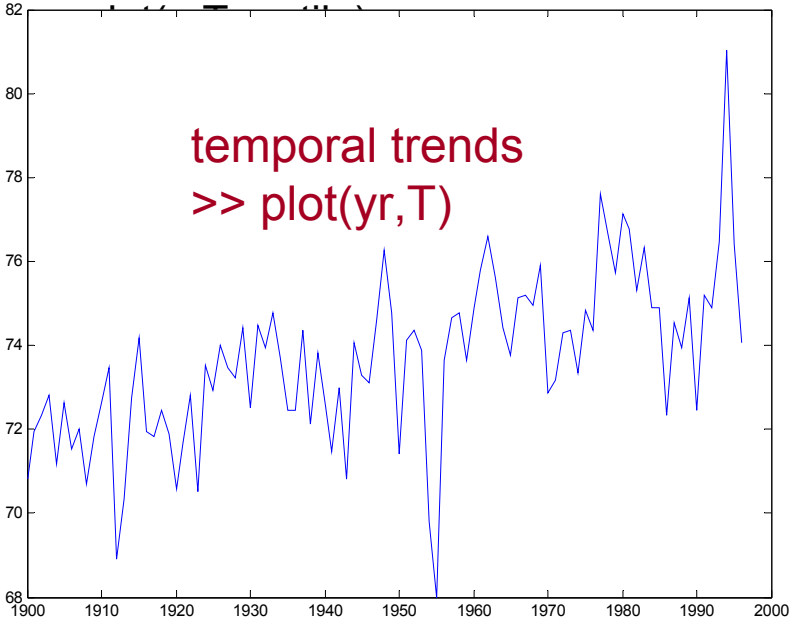
The K-S test assumes the mean and standard deviation of the population normal distribution are known; Lilliefors test does not make this assumption. In the analysis of empirical data, moreoften than not the mean and variance of the population normal distribution are unknown, and must be estimated from the data. Hence Lilliefors test is generally more relevant than the K-S test.

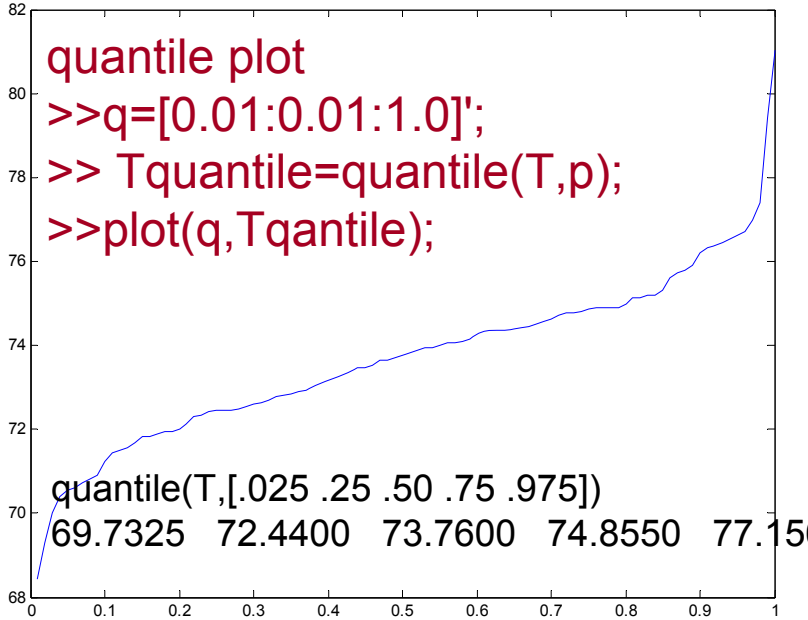
Probability distribution
Example of application

in MATLAB
load Tucson
whos

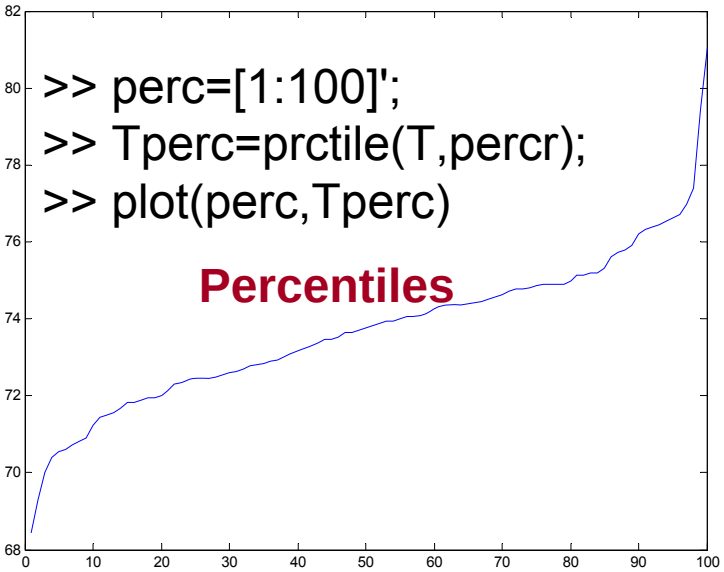
Name	Size	Bytes	Class	Attributes
T	97x1	776	double	
vlist	2x40	160	char	
yr	97x1	776	double	

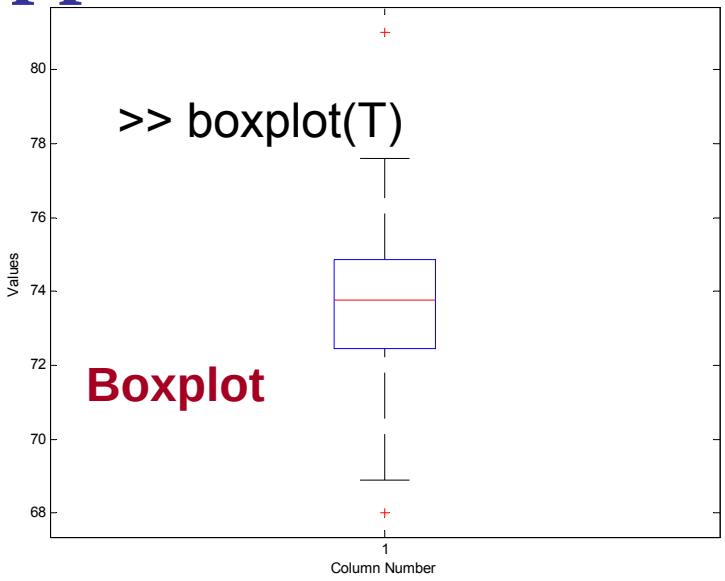
```
>> plot(yr,T)
>>q=[0.01:0.01:1.0]';
>> Tquantile=quantile(T,p);   & returns quantiles of the values
```





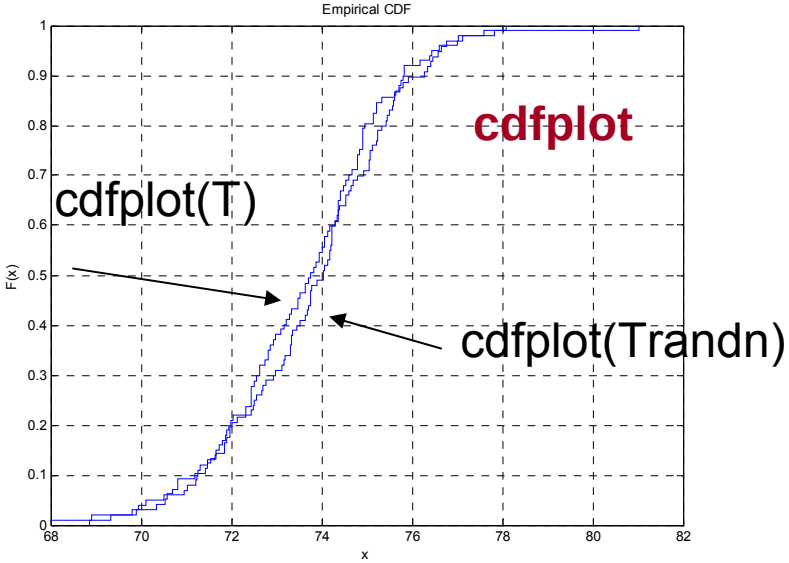
Probability distribution
Example of application





Comparison of distributions
cdfplot(T) plots an empirical cumulative distribution function (CDF) of the observations in the data sample vector T

xrandn=randn(100,1);
Trandn=Tm+xrandn*Ts.

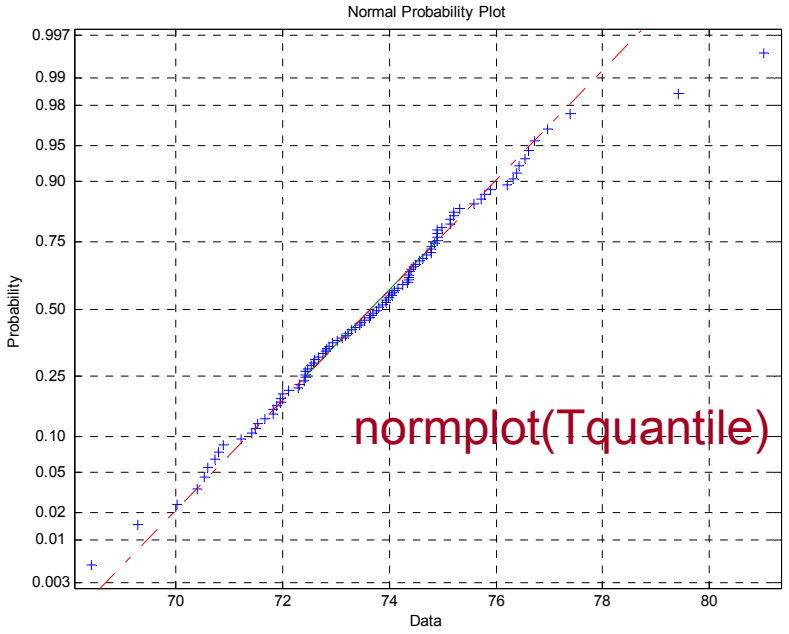


Probability distribution
Example of application

Normal Probability Plot

The plot has the sample data displayed with the plot symbol '+'. Superimposed on the plot is a line joining the first and third quartiles of each column of X (a robust linear fit of the sample order statistics.)

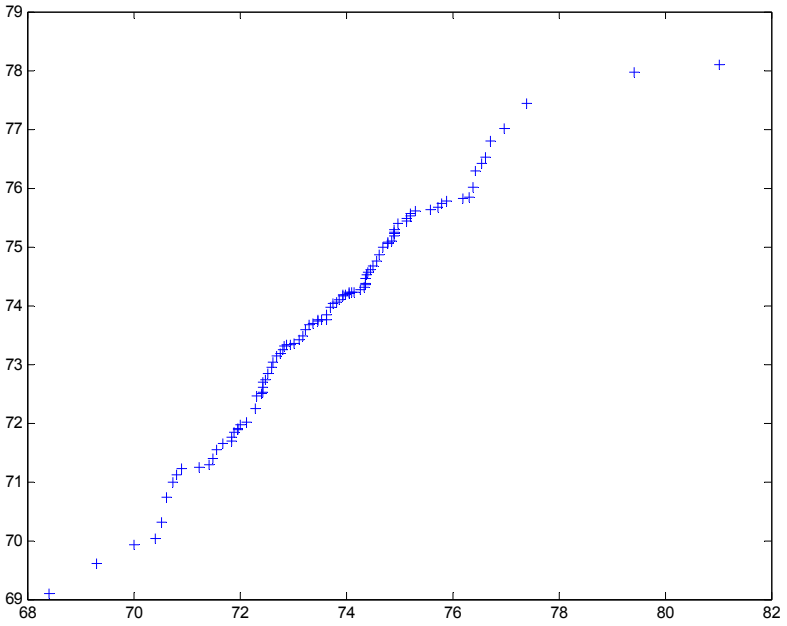
This line is extrapolated out to the ends of the sample to help evaluate the linearity of the data.



Probability distribution
Example of application

Comparison of T values with random vales with the same mean and std

```
xrandn=randn(100,1);
Trandn=Tm+xrandn*Ts
Tquantilerandn=quantile(Trandn,q)
plot(Tquantilerandn,Tquantile)
plot(Tquantile,Tquantilerandn,'+')
```



Normality tests

1)Test Lillietest

[H,alfa] = LILLIETEST(T)
H =0; alfa = 0.4810

2) Test Kolmogorof

Tcdf=normcdf(T,Tm,Ts)
[H,P] = KSTEST(T,[T,Tcdf])
H = 0; P = 0.8431

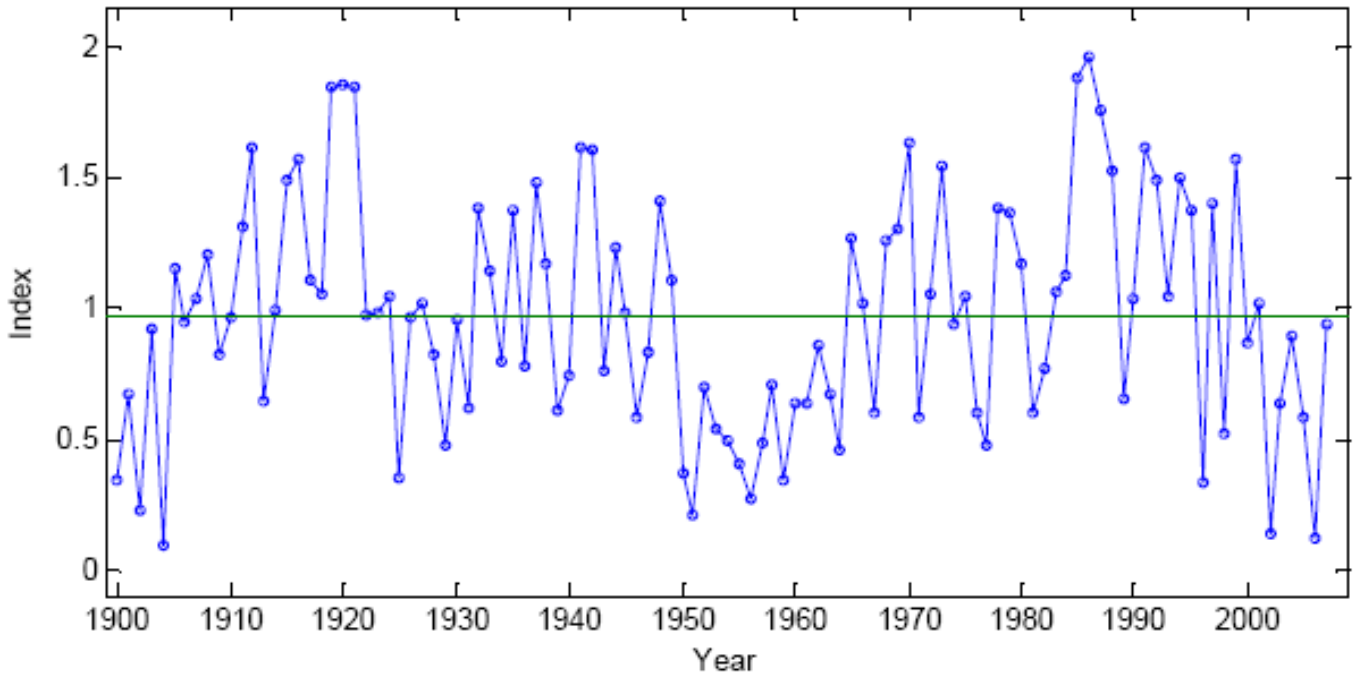


Figure 2.1. Time series plot of MEAF tree-ring index for period 1901-2007. Each observation is a dimensionless index of tree-ring width for a year. Horizontal line is the 1901-2007 mean. Higher values indicates wide rings and lower values narrow rings. This time series has a uniform time step of one year.

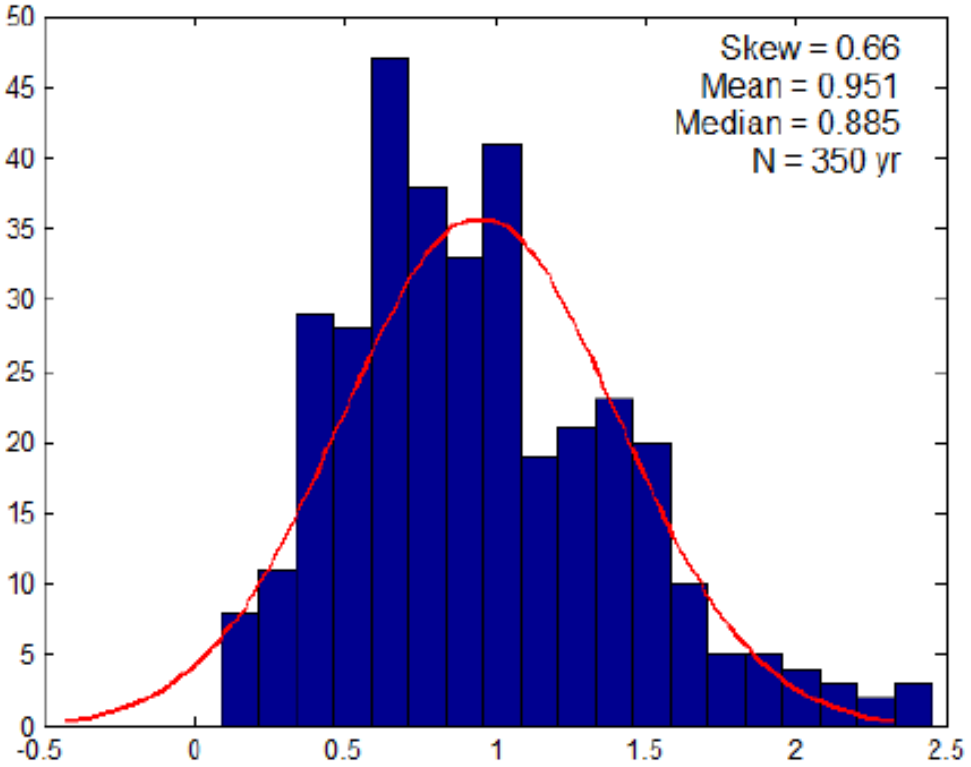


Figure 2.4. Histogram of MEAF tree-ring index. Data covers 350 years 1658-2007. Smooth red line is theoretical pdf of normal distribution with same mean and standard deviation as the tree-ring index.

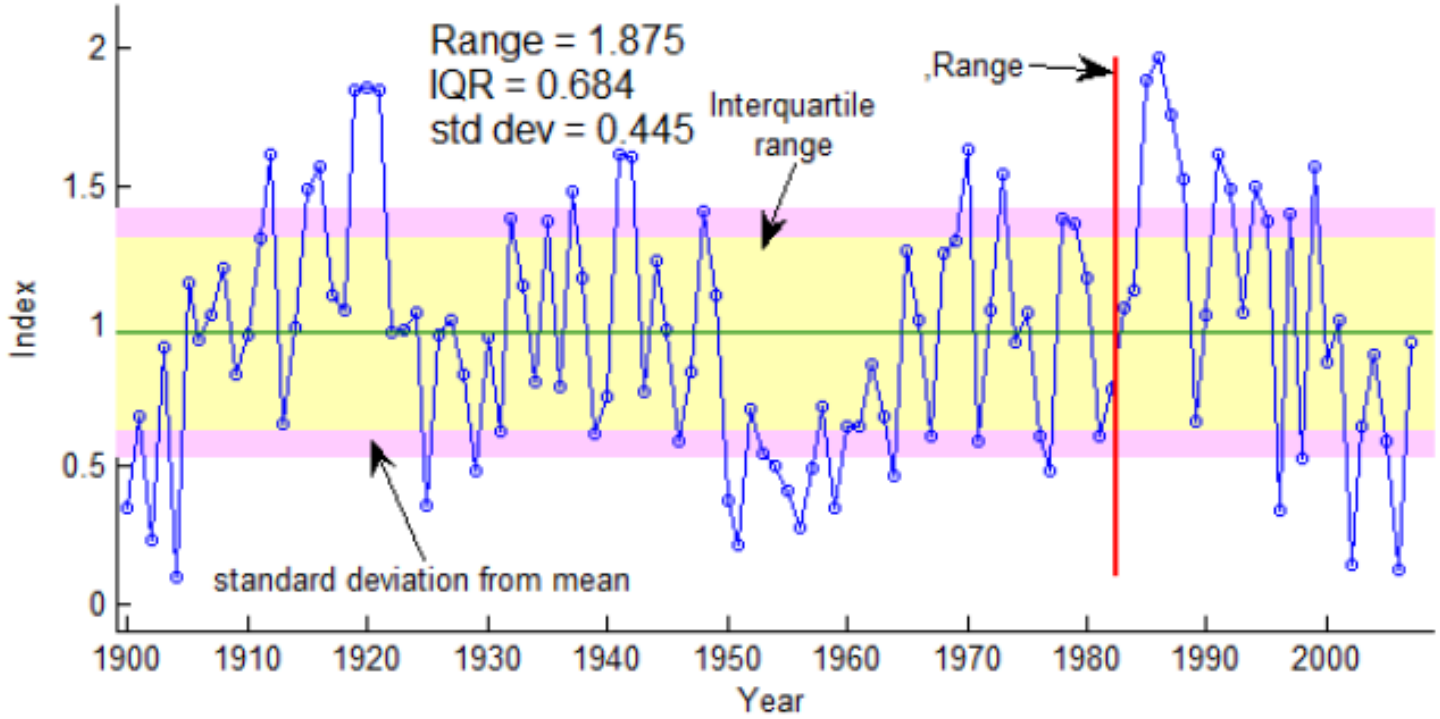


Figure 2.5. Measures of spread illustrated in time series plot of MEAF tree-ring index, 1901-2007. Plotted series same as in Figure 2.1. Range extends from highest to lowest value. Interquartile range (iqr) covers the middle 50% of observations (all except the highest 25% and lowest 25%). Standard deviaton is computed from squared departures from mean; in this plot the purple band delineates observations within ±1 standard deviation of the mean. For a normal distribution, 68% of the observations would fall within ±1 standard deviation of the mean. It is therefore expected that for normally distributed data the iqr would be narrower than the band marking ±1 standard deviations from the mean.

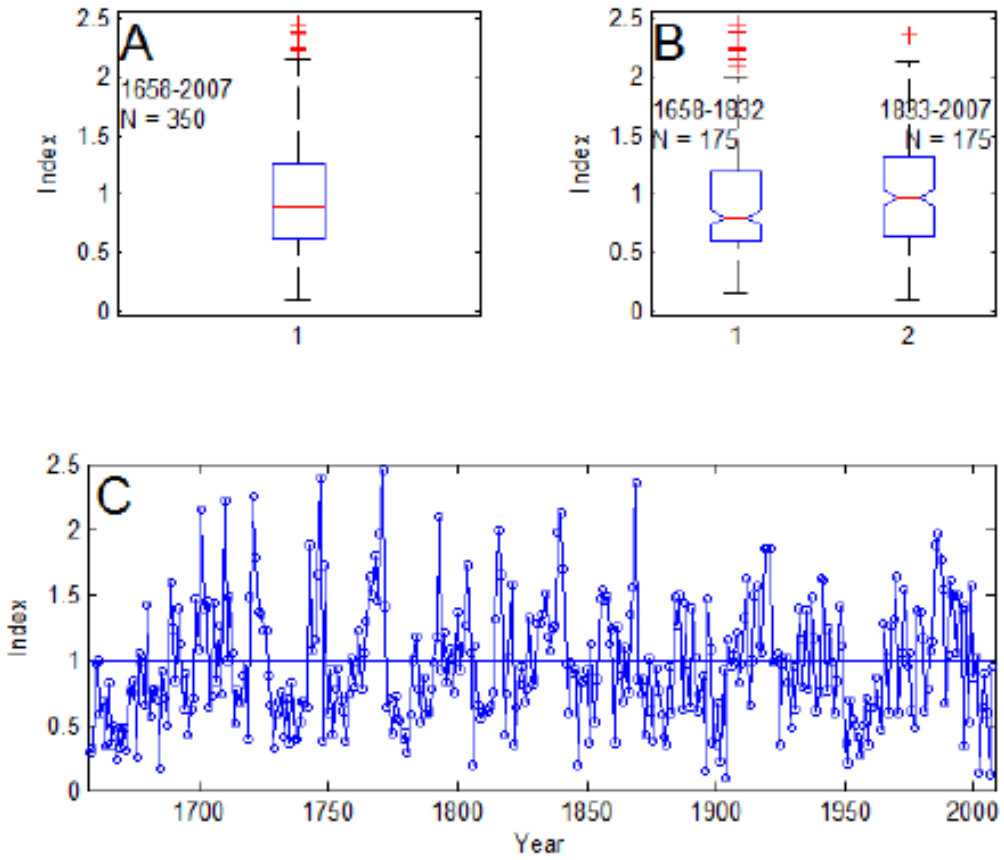


Figure 2.7. Boxplots and time series plot of tree-ring index for site MEAF. (A) Boxplot for entire series. (B) Notched boxplots for first and second halves of series. (C) Time series plot of entire series. Boxplot for entire series shows mild positive skew in middle of distribution (median toward lower half of box) and positive skew in tails. Boxplots for halves emphasize the relative skewness of first half. That notches do not overlap indicates sub-period medians different (α=0.05). Greater skew of first half obvious from time series plot, but difference in medians not noticeable. In fact, median is 0.79 for first half and 0.96 for second half.

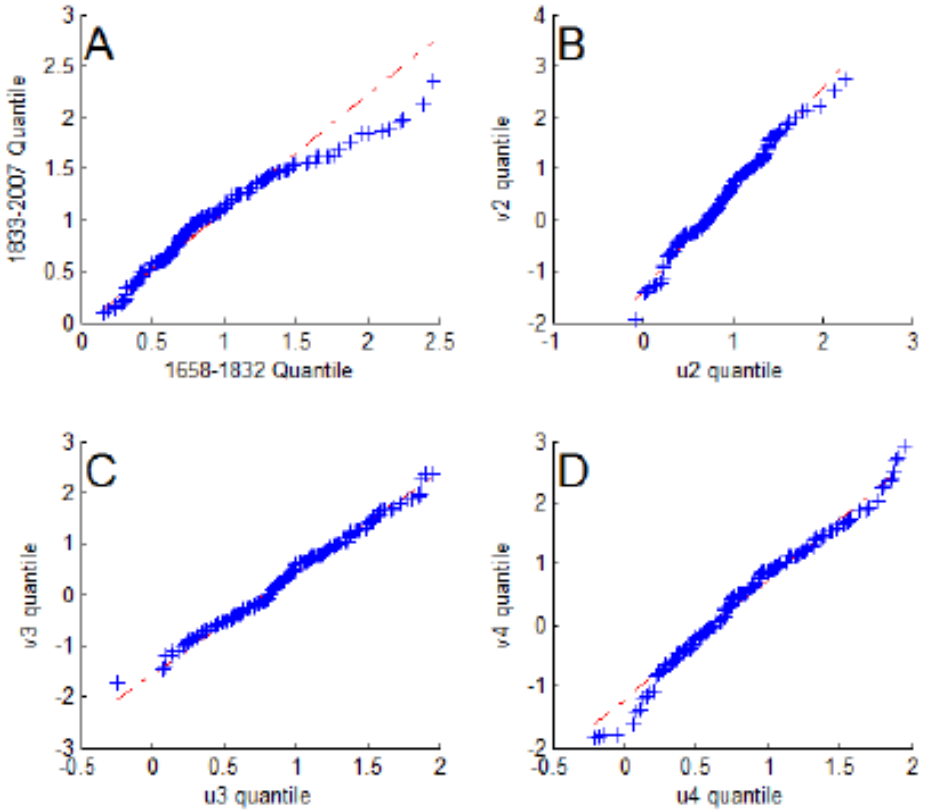


Figure 2.8. Examples of quantile-quantile plots. (A) First and last halves (175 years each) of MEAF tree-ring index (time series plot in Figure 2.7). (B-D) pairs of 175-year samples from random normal distribution. The random series for the ordinate (v2, v3, v4) were drawn from a random normal distribution with mean and standard deviation the same as that of the 1658-1832 tree-ring index. The random series for the abscissa (u2,u2, u4) were drawn from a random normal distribution with mean ½ that of the 1658-1832 tree-ring index and standard deviation twice that of the 1658-1832 tree-ring index. Differences in mean and variance (or standard deviation) do not affect the q-q plot: if two series come from the same form of distribution (e.g., normal), the q-q plot should fall along a straight line regardless of differences in the mean and variance.

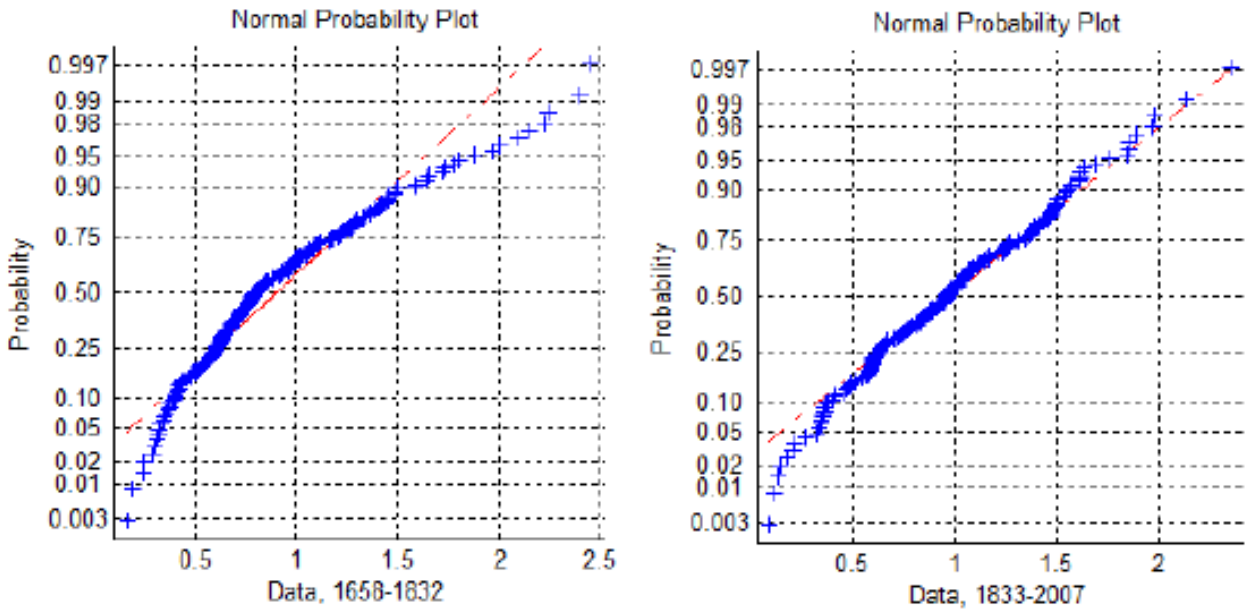


Figure 2.9. Normal probability plots for first and last halves of MEAF tree-ring index. If normally distributed, the sample should plot along the straight dashed line. At lower end of distribution (low-growth), both series are "pulled in" toward the center relative to a normally distributed series with the same mean and variance as the sample. The normal distribution would therefore have a longer tail on the left side than the sample. At the higher end of the distribution (high growth), the data for 1658-1832 is stretched out relative to the normal distribution, reflecting the high values and skew note previous (see Figure 2.7). The 1833-2007 segment, in contrasts, conforms well to a normal distribution over the high range.

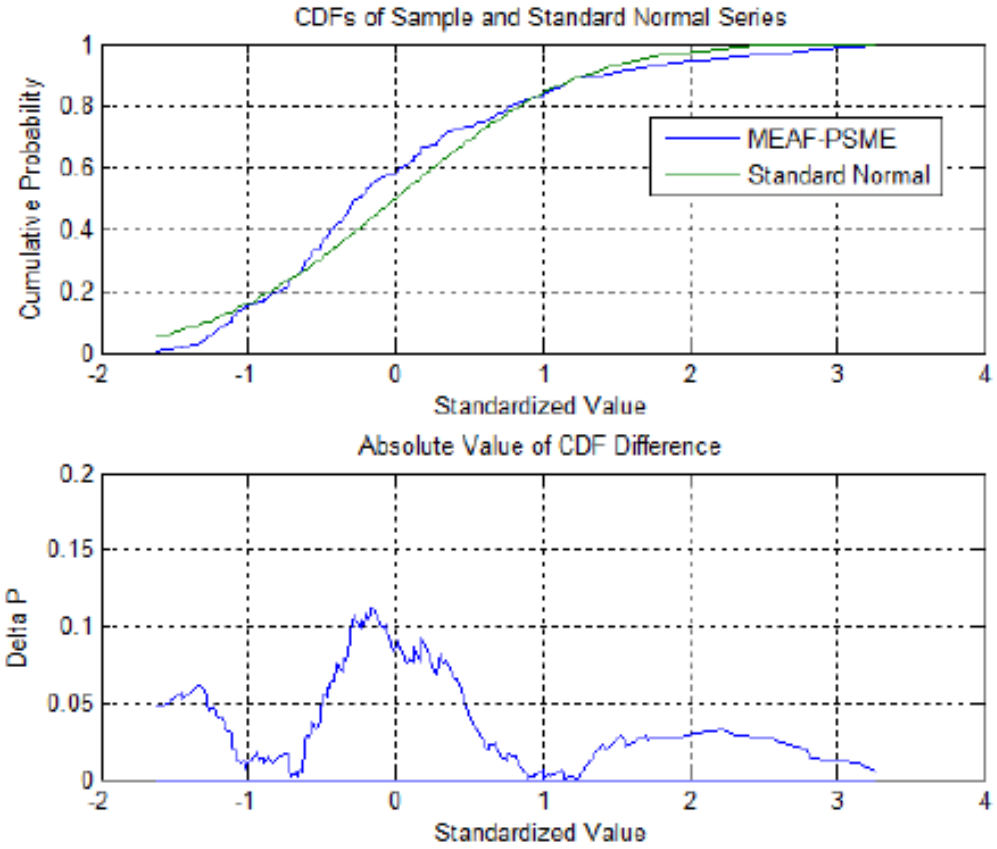


Figure 1.10. Graphical summary of Lilliefors statistic applied to test for normality of a tree-ring index. Test series is the 1658-1832 portion of MEAF tree-ring series. (Top) Empirical cdf's of a standard normal series and of the Z-scores of the tree-ring series. (Bottom) Difference between the two cdf's as a function of the standardized value of the tree-ring series. This difference is labeled "Delta P". The maximum absolute value of Delta P is defined as the Lilliefors statistic. For these data, the Lilliefors statistic is large enough to be significant (α<.05) such that the hypothesis of normality is rejected.

Time Series Analysis

Theory: Time Series Analysis

Probability distribution

Correlation and Autocorrelation

Spectrum and spectral analysis

Autoregressive-Moving Average (ARMA) modeling

Spectral analysis -- smoothed periodogram method

Detrending, Filtering and Smoothing

Laboratory exercises:

,…

Applied Time Series Analysis Course. David M. Meko, University of Arizona. Laboratory of Tree-Ring Research,

Email: dmeko@LTRR.arizona.edu

Romà Tauler (IDAEA, CSIC, Barcelona)

Correlation

Univariate correlation between two variables:

•Scattersplots are useful for checking whether the relationship is linear.

•The Pearson product-moment correlation coefficient is probably the single most widely used statistic for summarizing the relationship between two variables.

•Pearson correlation coefficient measures strength of linear relationship.

•The statistical significance of a correlation coefficient depends on the sample size, defined as the number of independent observations.

•If time series are autocorrelated, an "effective" sample size, lower than the actual sample size, should be used when evaluating significance.

Correlation

Correlation coefficient, mathematical definition

$$r == \frac{\mathrm{cov}(x,y)}{s_x s_y} = \frac{1}{(N-1)} \sum_{t=1}^{N} z_{t,x} z_{t,y},$$

Scaling

"Z-score" expression

$$z_{t,x} = \frac{(x_t - \bar{x})}{s_x}, s_x = \sqrt{\frac{\sum_{t=1}^{N} (x_t - \bar{x})^2}{N-1}}$$

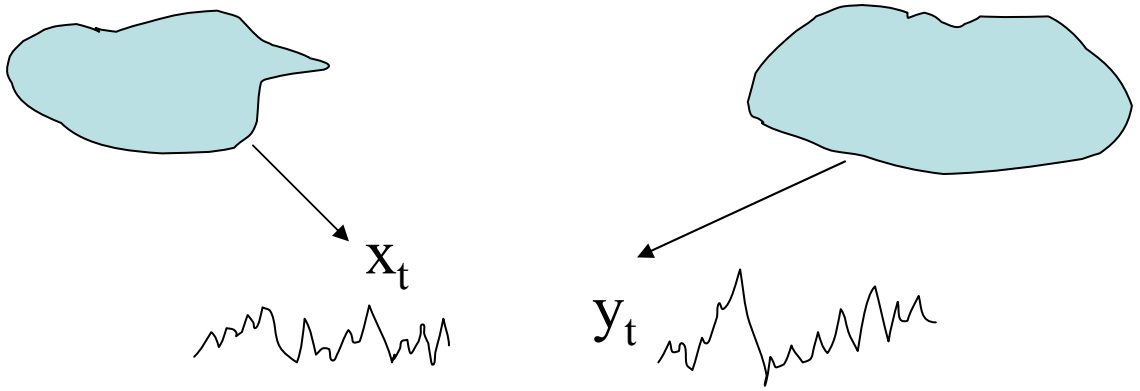
$$z_{t,y} = \frac{(y_t - \bar{y})}{s_y}, s_y = \sqrt{\frac{\sum_{t=1}^{N} (y_t - \bar{y})^2}{N-1}}$$

Departures

$$\mathrm{cov}(x,y) = \frac{1}{N-1} \sum_{t=1}^{N} (x_t - \bar{x})(y_t - \bar{y})$$

Correlation

Statistical significance: Testing H0 that ρ=0



Assume:

- Populations normally distributed
- Populations uncorrelated
- Pairs of observations drawn at random
- Sample size “large”

Correlation

Testing H0 that ρ=0

If assumptions true, sample correlation coefficient is normally distributed with

Mean = 0, Standard deviation = $1/(N-2)^{1/2}$

This information yields theoretical confidence bands for the correlation coefficient *r*.

The 0.975 probability point of the normal distribution is 1.96. Approximately 95% of the sample correlations should therefore fall within about ±1.96 standard deviations of zero. If the sample size is, say, *N* = 200 , the 95% confidence interval is

$$\frac{-1.96}{\sqrt{200-2}} \text{ to } \frac{+1.96}{\sqrt{200-2}} = -0.1393 \text{ to } +0.1393$$

A computed correlation coefficient greater in absolute magnitude than 0.1393 is judged "**significantly different than zero**" at the 0.05 alpha level, which corresponds to the 95% significance level. In this case, a two-tailed test with an alpha level of α = 0.05 , the null hypothesis of zero correlation is rejected.

Correlation

Testing H0 that ρ=0

The same critical threshold *r* would apply at the alpha level α = 0.05/ 2 = 0.025 for a one tailed test. In the one-tailed test the hypotheses might be:

H0: correlation coefficient is zero
H1: correlation coefficient is "greater than" zero

In the example above, a computed correlation of *r* = 0.15 would indicate rejection of the null hypothesis at the 0.05 level for the two-tailed test and rejection of the null hypothesis at the 0.025 level for the one-tailed test.

Whether to use a one-tailed or two-tailed test depends on the context of the problem.

If only positive correlations (or only negative) seem plausible from the physical relationship in question, the one-tailed test is appropriate.

Otherwise the two-tailed test should be used.

Correlation

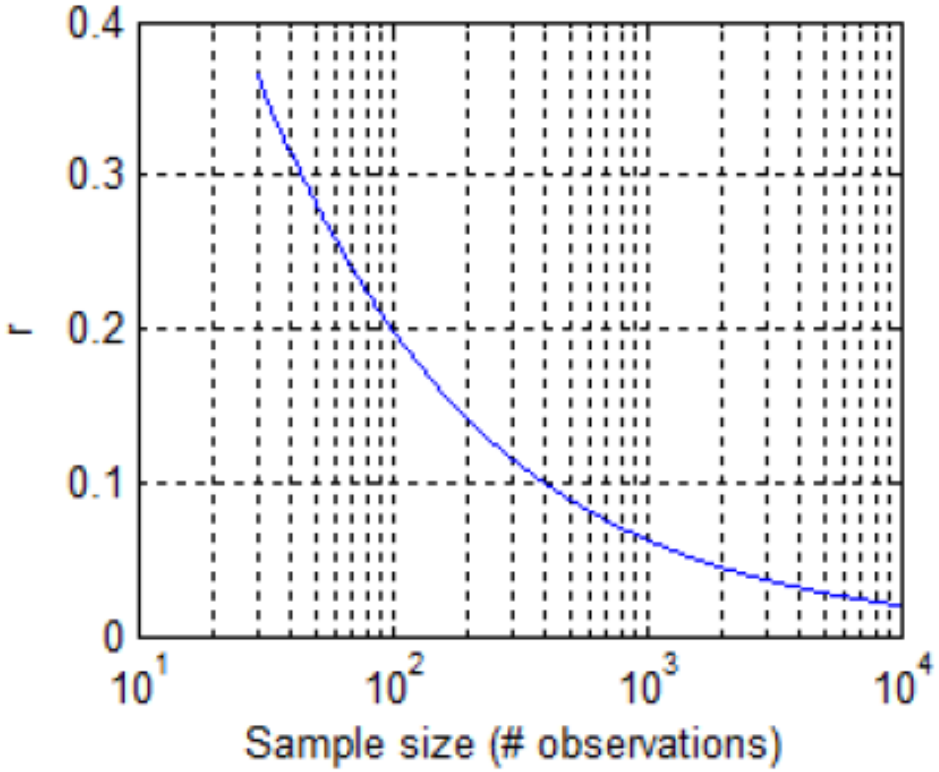


Figure 3.3. Critical level of correlation coefficient (95 percent significance) as a function of sample size. The critical level drops from r=0.20 for a sample size of 100 to r=0.02 for a sample size of 10,000.

Correlation

Testing H0 that ρ=0

What can be done if the scatterplot of *y* vs *x* is nonlinear?

• **Log-transform** $v = \log_{10}(y)$
compresses scale at high end of distribution; useful on *y* when scatterplot of *y* on *x* shows increasing scatter with increasing *y*

in hydrology, frequently used to transform discharge data to normality

• **Power transformation** $v=y^p$
most often used are square-root transform ($p = 0.5$) and squaring ($p = 2$);
square root transform has similar effect to log-transform
p is usually restricted to positive values
if *p* is negative, transformation $v= -y^p$ preferred

Autocorrelation

Autocorrelation refers to the correlation of a time series with its own past and future values. Autocorrelation is sometimes called "*serial correlation*", which refers to the correlation between members of a series of numbers arranged in time. Alternative terms are "*lagged correlation*", and "*persistence*"

Time series are frequently autocorrelated because of *inertia* or *carryover* processes in the physical system.

Autocorrelation complicates the application of statistical tests by reducing the effective sample size. Autocorrelation can also complicate the identification of significant covariance or correlation between time series (e.g., correlation of precipitation with a tree-ring series).

Autocorrelation

Three tools for assessing the autocorrelation of a time series are:

(1) the time series plot

(2) the lagged scatterplot,

(3) the autocorrelation function.

Autocorrelation: *Lagged scatterplot*

The simplest graphical summary of autocorrelation in a time series is the ***lagged scatterplot***, which is a scatterplot of the time series against itself offset in time by one to several years.

Let the time series of length N be , 1,..., x_i i= N.

The lagged scatterplot for lag k is a scatterplot of the last $N - k$ observations against the first $N - k$ observations.

For example, for lag-1, observations x_2 x_3, , ,x_N are plotted against observations x_1, x_2, ... $x_{N-1.}$

Autocorrelation: *Lagged scatterplot*

A **random scattering** of points in the lagged scatterplot indicates a lack of autocorrelation. Such a series is also sometimes called "random", meaning that the value at time *t* is independent of the value at other times.

Alignment from lower left to upper right in the lagged scatterplot indicates **positive autocorrelation**.

Alignment from upper left to lower right indicates **negative autocorrelation**.

An attribute of the lagged scatterplot is that it can display auto-correlation regardless of the form of the dependence on past values.

Autocorrelation: *Lagged scatterplot*

An assumption of linear dependence is not necessary.

An organized curvature in the pattern of dots might suggest **nonlinear dependence** between time separated values. Such nonlinear dependence might not be effectively summarized by other methods (e.g., the autocorrelation function, which is described later).

Another attribute is that the lagged scatterplot can show if the autocorrelation is driven by one or more **outliers** in the data.

This again would not be evident from the ***acf*** (autocorrelation function).

Autocorrelation: *Lagged scatterplot*

Lagged scatterplots are drawn for lags 1-8 years.

The straight line that appears on these plots is fit by least squares, and it is intended to aid in judging the preferred orientation of the pattern of points.

The correlation coefficient for the scatterplot summarizes the strength of the **linear relationship** between present and past values.

Autocorrelation: *Autocorrelation function (ACF)*

The correlation coefficient between x and y is given by:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum (x_i - \bar{x})^2\right]^{1/2} \left[\sum (y_i - \bar{y})^2\right]^{1/2}}$$

Autocorrelation: *Autocorrelation function (ACF)*

A similar idea can be applied to time series for which successive observations are correlated. Instead of two different time series, the correlation is computed between one time series and the same series ***lagged by one or more time units***. For the first-order autocorrelation, the lag is one time unit.

$$r_1 = \frac{\sum_{t=1}^{N-1}(x_t - \bar{x}_1)(x_{t+1} - \bar{x}_2)}{\left[\sum_{t=1}^{N-1}(x_t - \bar{x}_1)^2\right]^{1/2}\left[\sum_{t=1}^{N-1}(x_{t+1} - \bar{x}_2)^2\right]^{1/2}}$$

Where $\bar{x}_1$ is the mean of the first $N-1$ observations and $\bar{x}_2$ is the mean of the last $N-1$ observations. it is called the autocorrelation coefficient or serial correlation coefficient.

Autocorrelation: *Autocorrelation function (ACF)*

For N reasonably large, the difference between the sub-period means $\overline{x}_1$ and $\overline{x}_2$ can be ignored and r_1 can be approximated by

$$r_1 = \frac{\sum_{t=1}^{N-1}(x_t - \overline{x})(x_{t+1} - \overline{x})}{\sum_{t=1}^{N}(x_t - \overline{x})^2}$$

where $\overline{x} = \sum_{i=1}^{N} x_i / N$ is the overall mean. This equation can be generalized to give the correlation between observations separated by k years:

$$r_k = \frac{\sum_{i=1}^{N-k}(x_i - \overline{x})(x_{i+k} - \overline{x})}{\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

The quantity r_k is called the ***autocorrelation coefficient at lag k***. The plot of the autocorrelation function as a function of lag is also called the ***correlogram.***

Autocorrelation: *Autocorrelation function (ACF)*

Theoretical distribution of autocorrelation coefficient if population is not autocorrelated

Assuming

1. Series random (no autocorrelation)
2. Series identically and normally distributed
3. Weak stationarity

A 95% Confidence level (one sided) or band (two sided) is:

$$r_k(95\%) = \frac{-1 + 1.645\sqrt{N-k-1}}{N-k} \qquad \text{one sided}$$

$$r_k(95\%) = \frac{-1 \pm 1.96\sqrt{N-k-1}}{N-k} \qquad \text{two sided}$$

Autocorrelation: *Autocorrelation function (ACF)*

A common application is to test the first-order, or lag-1 auto-correlation (k=1). The 95% signif level *for one-tailed test* is:

$$r_{1,.95} = \frac{-1 + 1.645\sqrt{N-2}}{N-1}$$

$H_0 = r_1 \le 0$, null hypothesis

$H_1 = r_1 > 0$, alternative hypothesis

N	$r_{1,0.95}$
30	0.27
100	0.15
1000	0.05

Autocorrelation: *Autocorrelation function (ACF)*

An approximation:
95% Confidence interval on r(k)

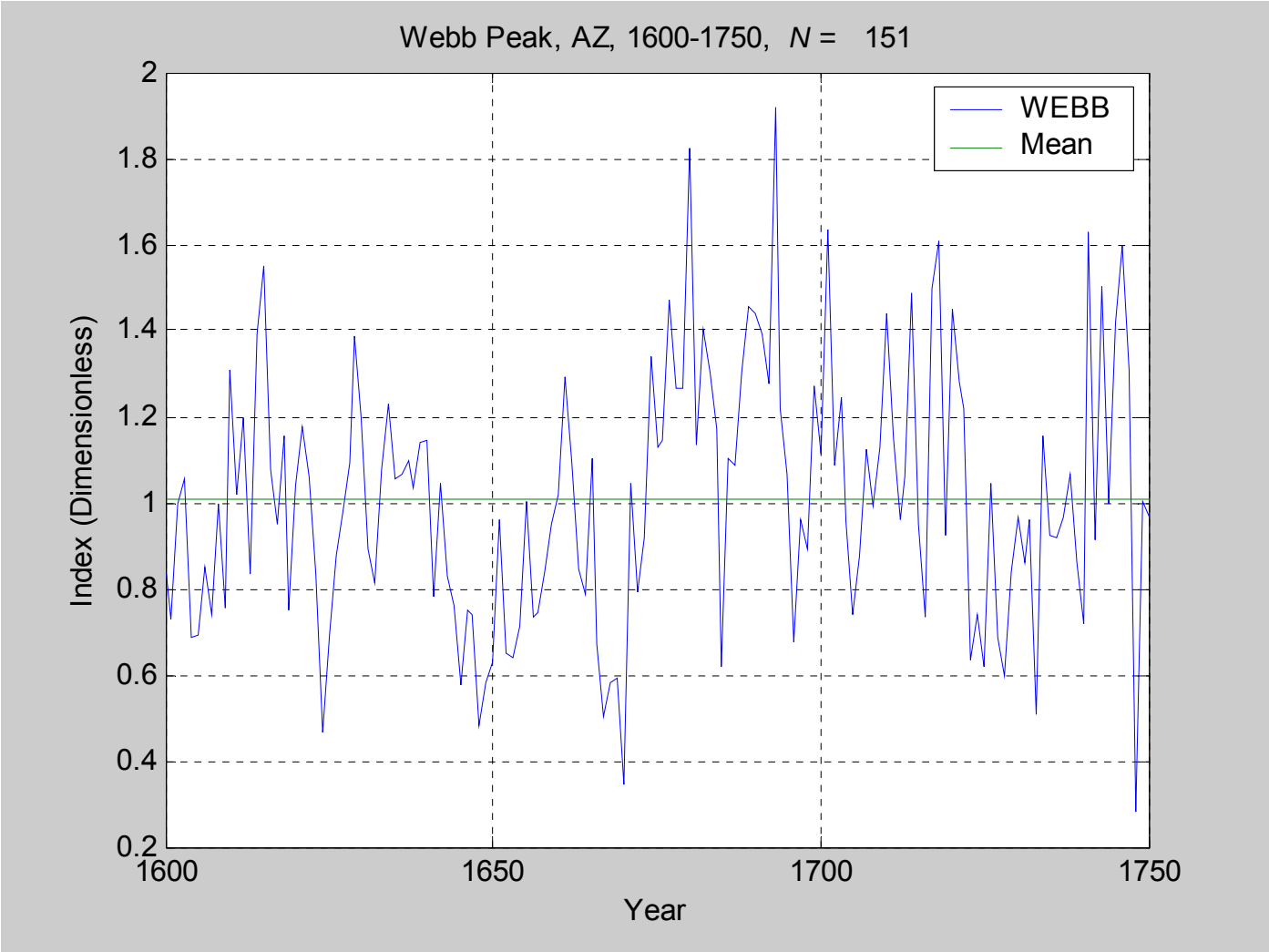
$$0 \pm 2/\sqrt{N}$$

- Appropriate for a *two-tailed test*
- Would give flat (horizontal lines) confidence band symmetric around zero

But often we find that the autocorrelaton at low lags is clearly non-zero. So the assumption of zero autocorrelation cannot be made.

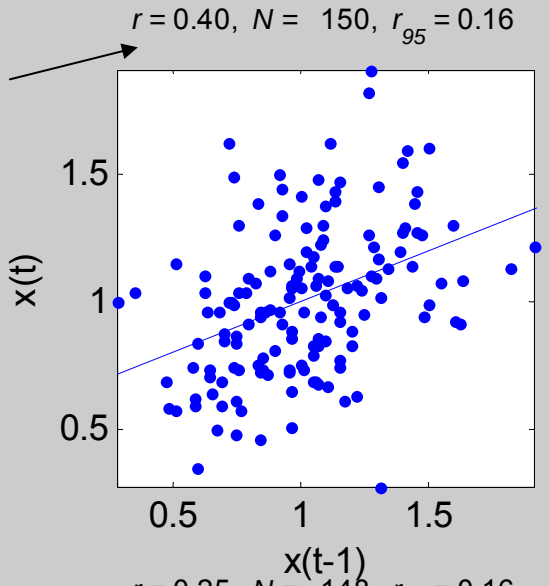
What is the confidence band for the whole acf if we cannot assume zero population autocorrelation?

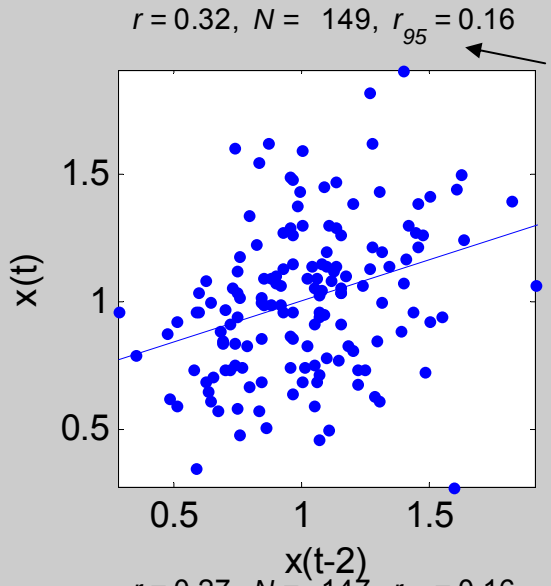
Autocorrelation: *Time series plots*
Example



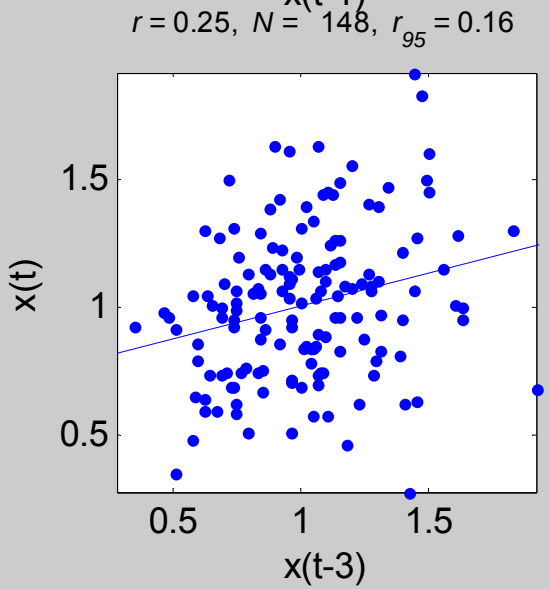
Autocorrelation: *Lagged scatterplot* Example

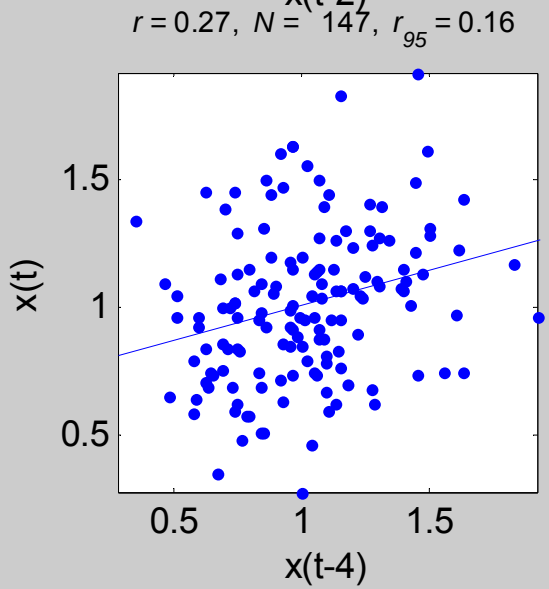
experimental correlation coefficient





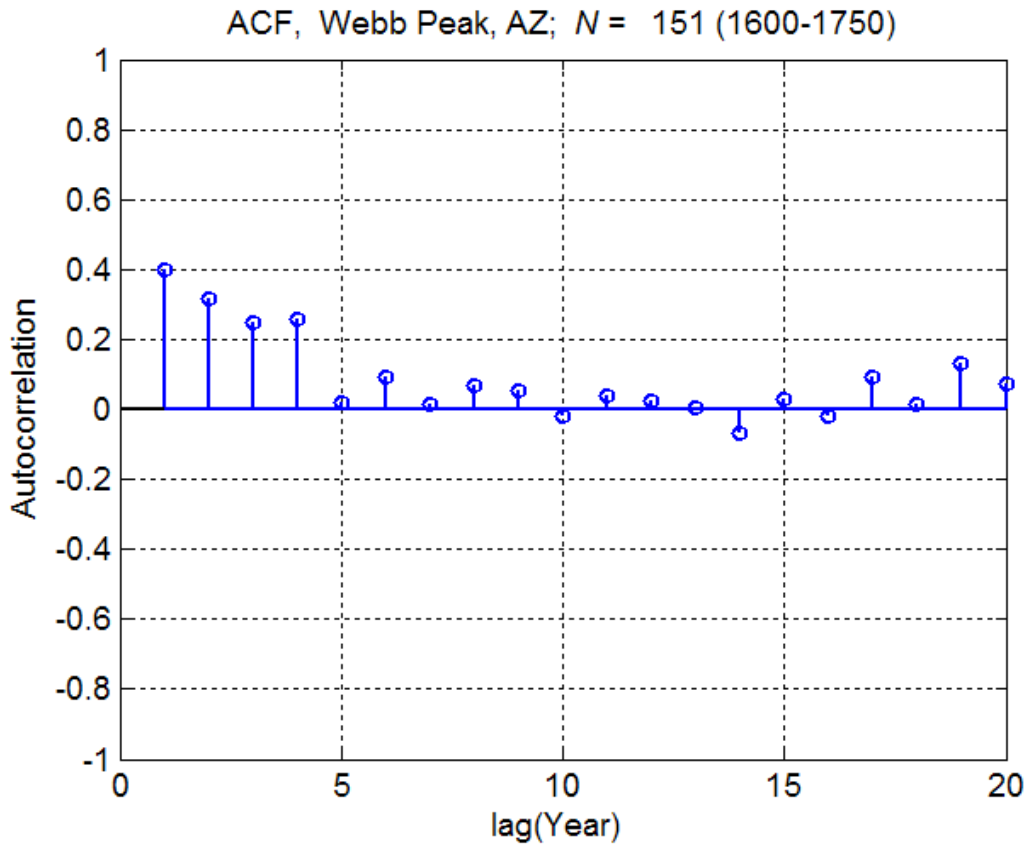
significant correlation at the 95% significance level





Autocorrelation: *Autocorrelation function (ACF)*

Example



Correlogram

Autocorrelation: *Autocorrelation function (ACF)*

Estimation of the Large-Lag Standard Error of acf

Previously defined confidence band based on assumption that true autocorrelation is zero.

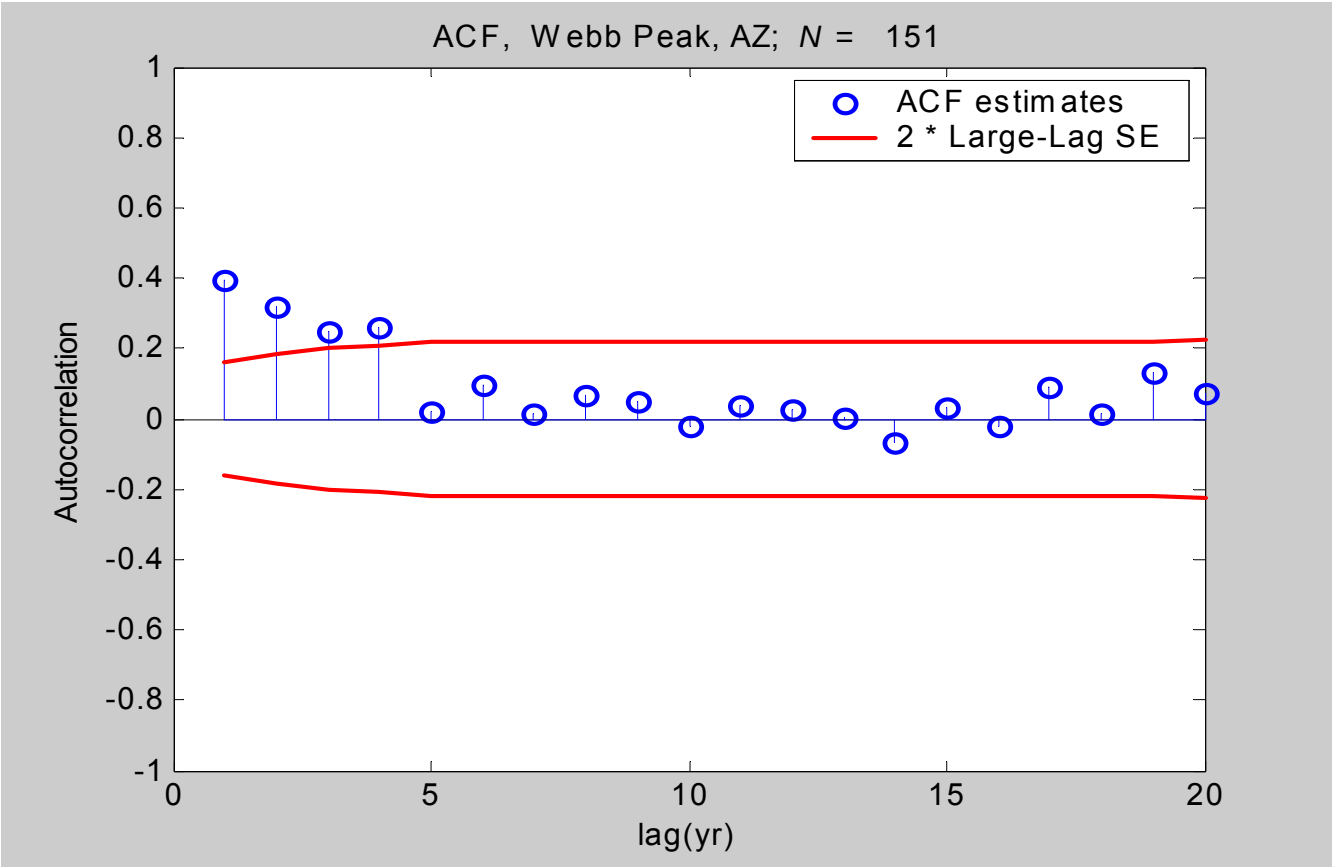
If not zero, band widens around r(k) at higher lags depending on the r(k) at lower lags.

"Large-lag" standard error is defined as square root of

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

Error bars on acf's can be estimated from the above equation

Autocorrelation: *Autocorrelation function (ACF)*



ACF and confidence band

Autocorrelation: *Autocorrelation function (ACF)*

Effective Sample Size

If a time series of length *N* is autocorrelated, the number of *independent observations* is fewer than *N*.
Essentially, the series is not random in time, and the information in each observation is not totally separate from the information in other observations. The reduction in number of independent observations has implications for hypothesis testing. Some standard statistical tests that depend on the assumption of random samples can still be applied to a time series despite the autocorrelation in the series. The way of circumventing the problem of autocorrelation is *to adjust the sample size for autocorrelation*. The number of independent samples after adjustment is fewer than the number of observations of the series.

Autocorrelation: *Autocorrelation function (ACF)*

Calculation of the "effective" sample size, or sample size adjusted for autocorrelation.

This equation is derived based on the assumption that the autocorrelation in the series represents ***first-order*** autocorrelation (dependence on lag-1 only). In other words, the governing process is ***first-order autoregressive***, or ***Markov***. Computation of the effective sample size requires only the sample size and first-order sample autocorrelation coefficient. The **"effective" sample size** is given by:

$$N' = N\frac{1-r_1}{1+r_1}$$

where N is the sample size, N' is the effective samples size, and 1 r is the first-order autocorrelation coefficient. For example, a series with a sample size of 100 years and a first order autocorrelation of 0.50 has an adjusted sample size of

$$N' = 100\frac{1-0.5}{1+0.5} = 100\frac{0.5}{1.5} = 33\,years$$

Autocorrelation: *Autocorrelation function (ACF)*

Calculation of the "effective" sample size, or sample size adjusted for autocorrelation

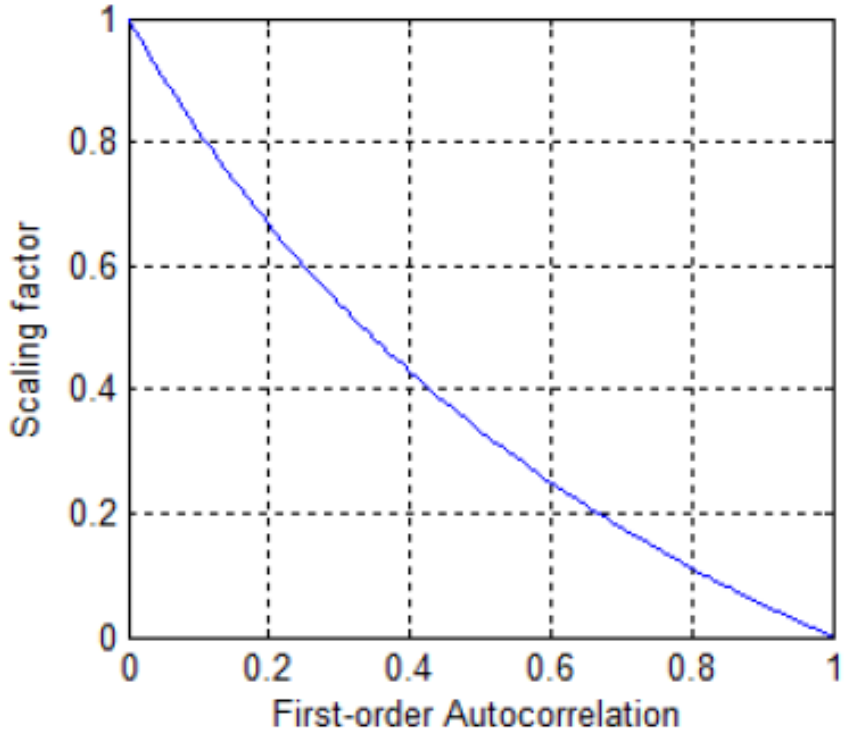


Figure 3.5. Scaling factor for computing effective sample size from original sample size for autocorrelated time series. For a given first-order

Autocorrelation: *Time series plots*
Example

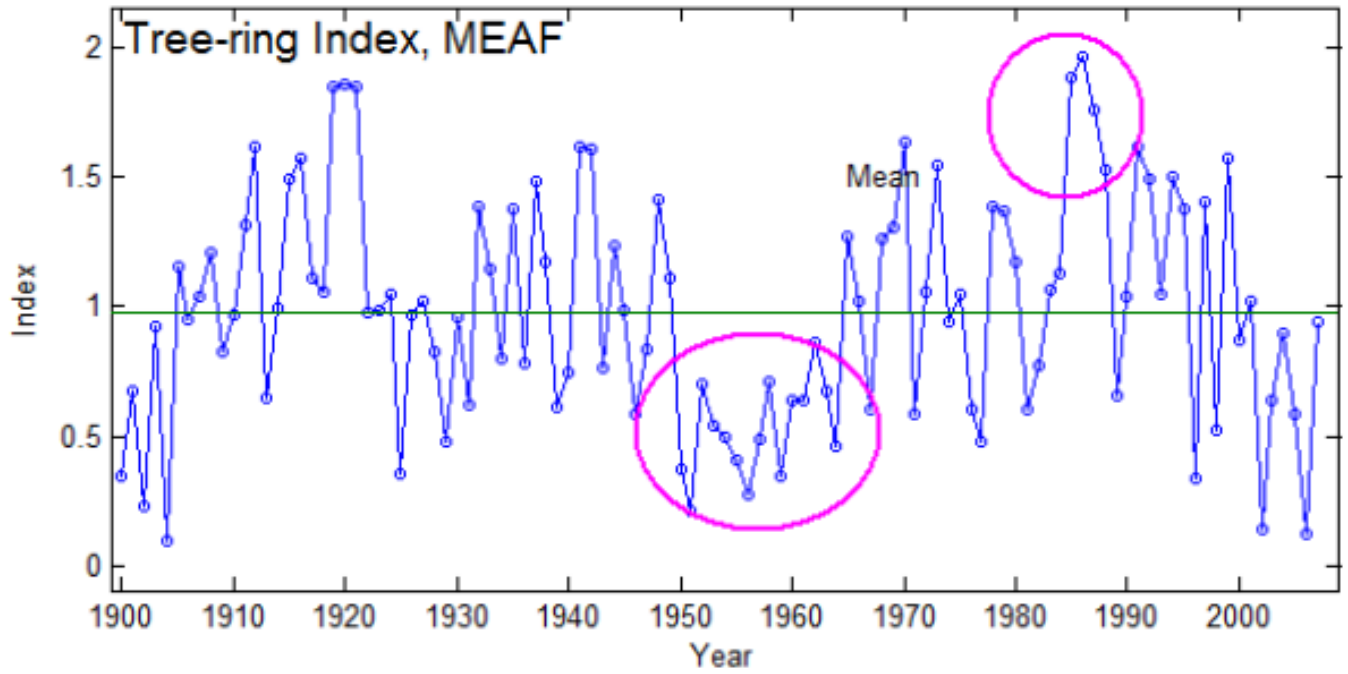


Figure 3.1. Time series plot illustrating signatures of persistence. Tendency for highs to follow highs or lows to follow lows (circled segments) characterize series with persistence, or positive autocorrelation.

Autocorrelation: *Lagged scatterplot* Example

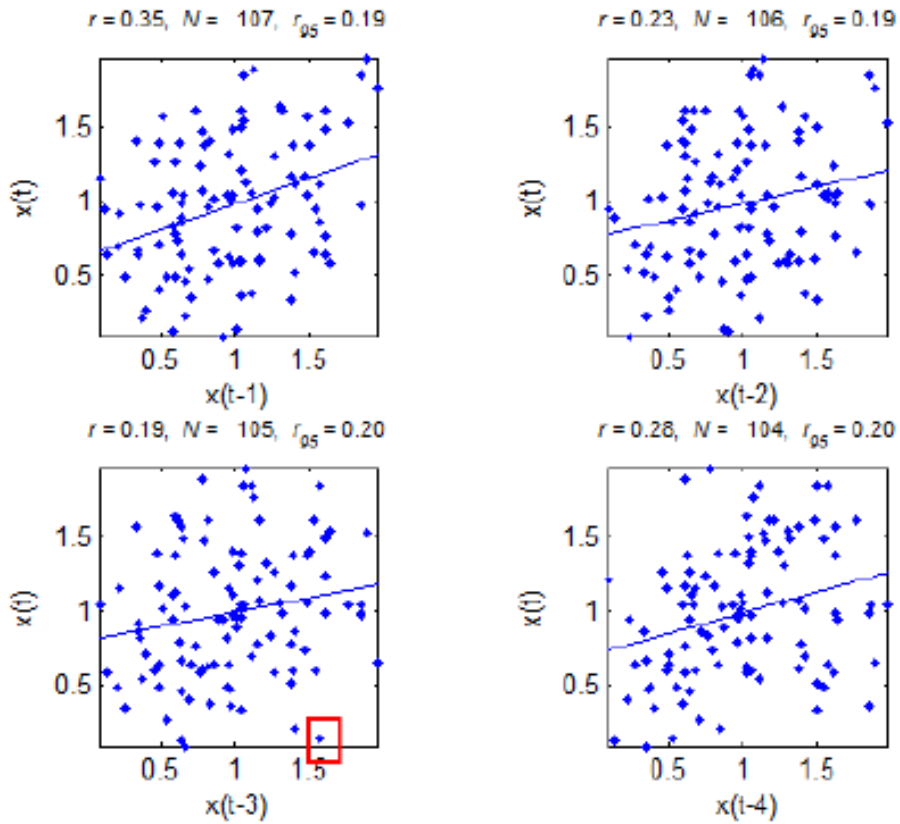


Figure 3.2. Lagged scatterplots of tree-ring series MEAF. These are scatterplots of the series in Figure 3.1 with itself offset by 1, 2, 3 and 4 years. Annotated above a plot is the correlation coefficient, the sample size, and the threshold level of correlation needed to reject the null hypothesis of zero population correlation with 95 percent significance (α=0.05). The threshold is exceeded at lags 1,2, and 4, but not at lag 3. At an offset of 3 years, the juxtaposition of high-growth 1999 with low-growth 2002 exerts high influence (point in red rectangle).

Autocorrelation: *Autocorrelation function (ACF)*
Example

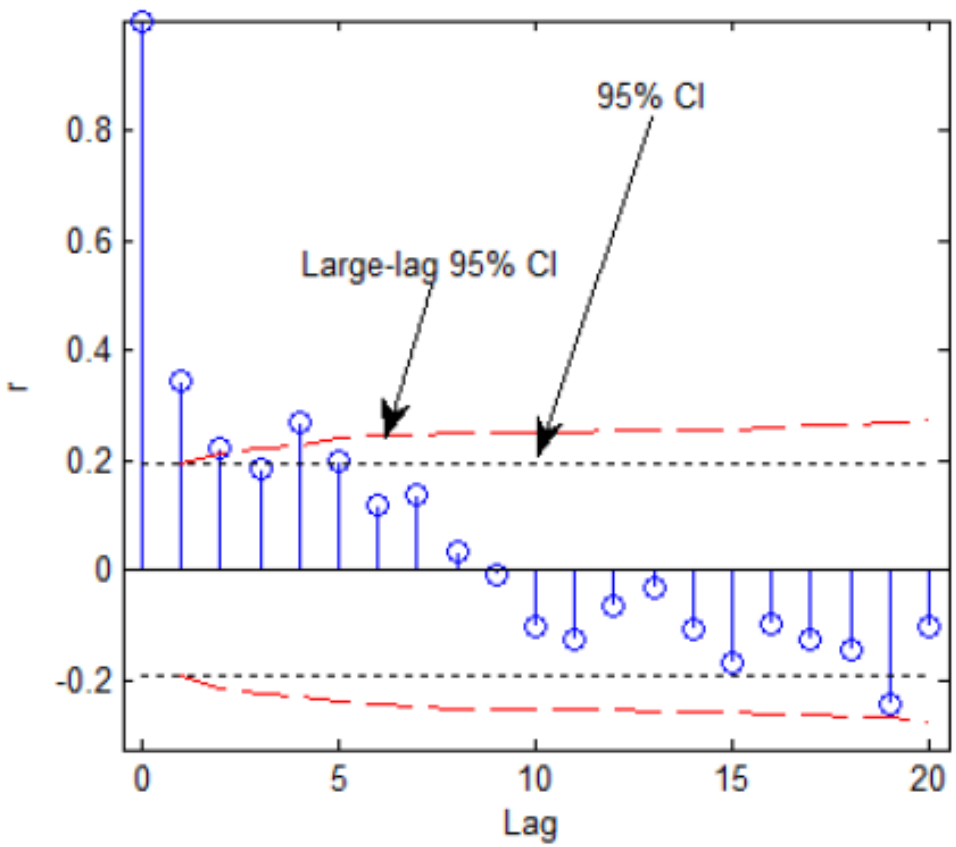


Figure 3.4. Sample autocorrelation with 95% confidence intervals for MEAF tree-ring index, 1900-2007. Dotted line is simple approximate confidence interval at 2/sqrt(N), where N is the sample size. Dashed line is large-lag standard error.

Time Series Analysis

Theory: Time Series Analysis

Probability distribution

Correlation and Autocorrelation

Spectrum and spectral analysis

Autoregressive-Moving Average (ARMA) modeling

Spectral analysis -- smoothed periodogram method

Detrending, Filtering and smoothing

Laboratory exercises:

,…

Applied Time Series Analysis Course. David M. Meko, University of Arizona. Laboratory of Tree-Ring Research, Email: dmeko@LTRR.arizona.edu

Romà Tauler (IDAEA, CSIC, Barcelona)

Spectrum

1. The frequency domain
2. Sinusoidal model of a time series
3. Harmonic analysis
4. Spectral analysis

Spectrum: the frequency domain

The spectrum of a time series is the ***distribution of variance of the series as a function of frequency***. The object of spectral analysis is to estimate and study the spectrum of the time data series.

The spectrum contains no new information beyond that in the autocovariance function (acvf), and in fact the spectrum can be computed mathematically by transformation of the acvf.

But the spectrum and acvf present the information on the variance of the time series from complementary viewpoints. ***The acf summarizes information in the time domain and the spectrum in the frequency domain***

Spectrum: the frequency domain

The spectrum of a time series is the variance of the series as a function of frequency

"The spectrum of a time series is analogous to an optical spectrum. An **optical** spectrum shows the contributions of different wavelengths or frequencies to the energy of a given light source.

The spectrum of a time series shows the contributions of oscillations with various frequencies to the variance of a time series." --**Panofsky (1958, p. 141)**

Panofsky, H.A., and Brier, G.W., 1958, Some applications of statistics to meteorology: The Pennsylvania State University Press, 224 p. *[Harmonic analysis; Climatology applications]*

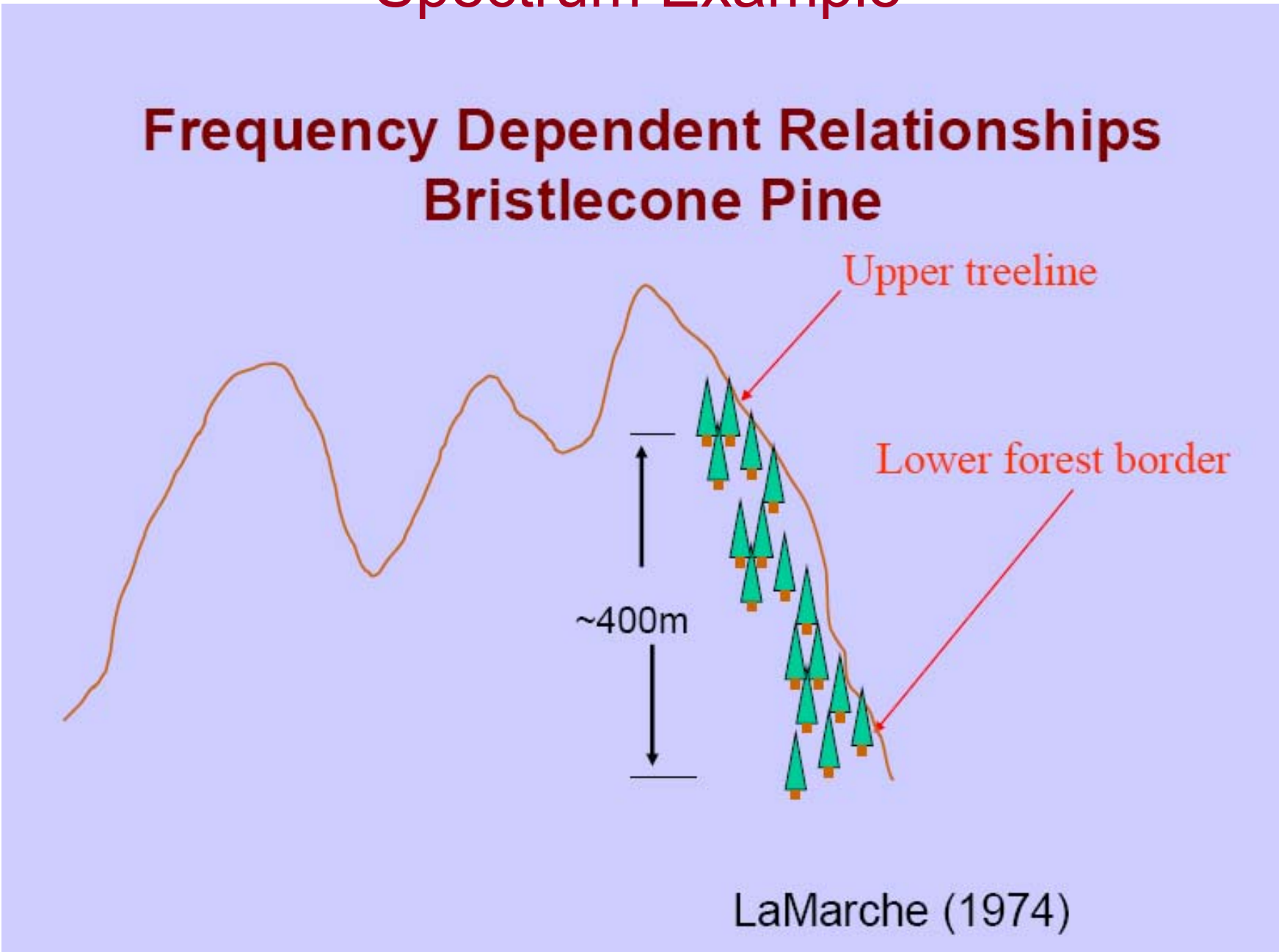
Spectrum: the frequency domain

Why study the spectrum?

- Describe important timescales of variability
- Gain insight to underlying physical mechanisms (from biology, chemistry, geology, physics..) of the system
- Forecast

The spectrum is of interest because many natural phenomena have **variability that is frequency-dependent**, and understanding the frequency dependenc may yield information about the underlying physical mechanisms. Spectral analysis can help in this objective

Spectrum Example



Spectrum Example

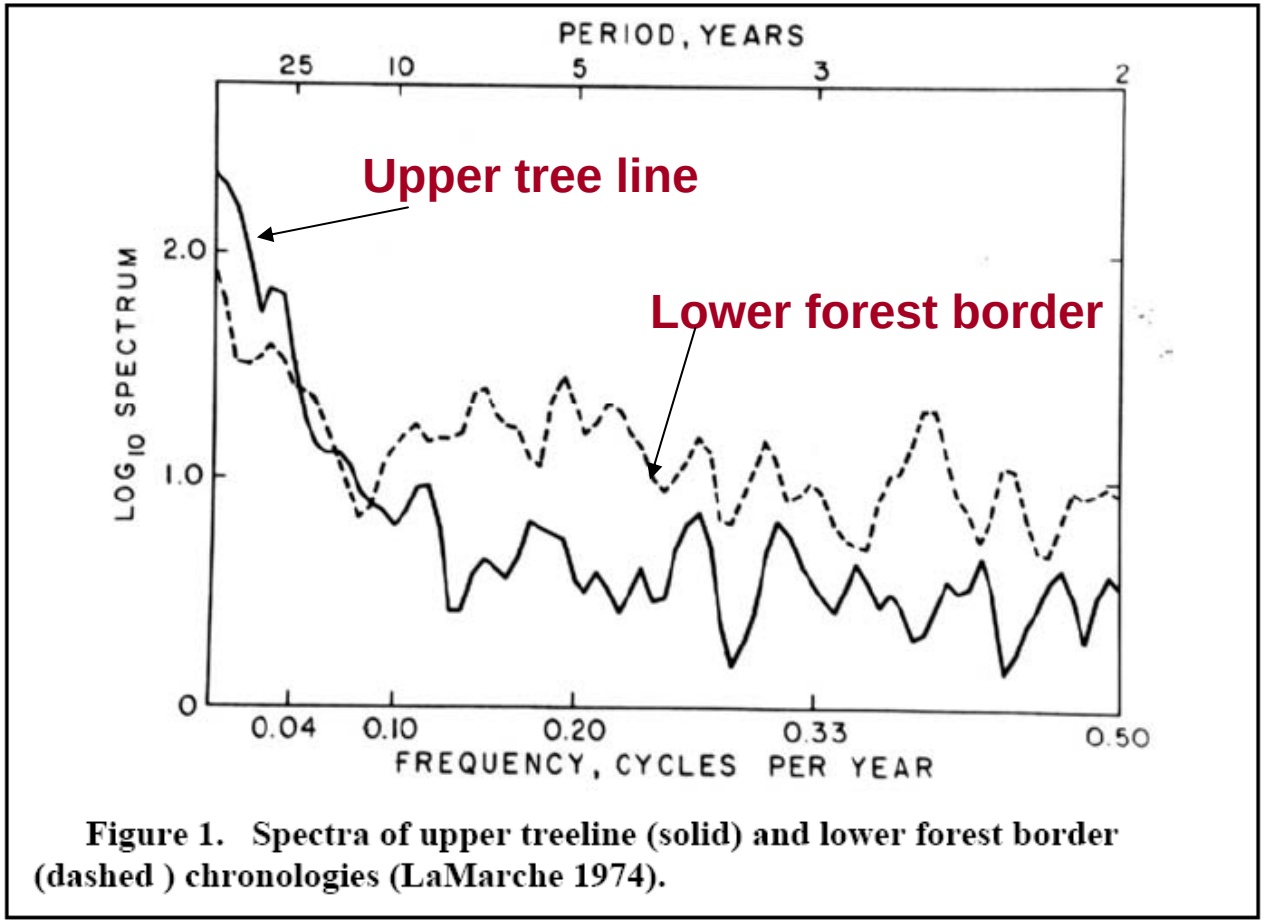


Figure 1. Spectra of upper treeline (solid) and lower forest border (dashed) chronologies (LaMarche 1974).

LaMarche, V.C., 1974, Frequency-dependent relationships between tree-ring series along an ecological gradient and some dendroclimatic implications, Tree-Ring Bulletin 34, 1-20.
LaMarche, V. C., and Fritts, H.C., 1972, Tree-rings and sunspot numbers: Tree-Ring Bulletin, v. 32, p. 19-33.

Spectrum: sinusoidal model for a time series

- Time Domain vs Frequency Domain
- Frequency domain terminology
- Sinusoidal model for a time series

Spectrum: sinusoidal model for a time series

In the time domain, variations are studied as a function of time.

For example, the time series plot of an annual tree-ring index displays variations in tree-growth from year to year, and the acf summarizes the persistence of a time series in terms of correlation between lagged values for different numbers of years of lag.

In the frequency domain, the variance of a time series is studied as a function of frequency or wavelength of the variation. The main building blocks of variation in the frequency domain are sinusoids, or sines and cosines.

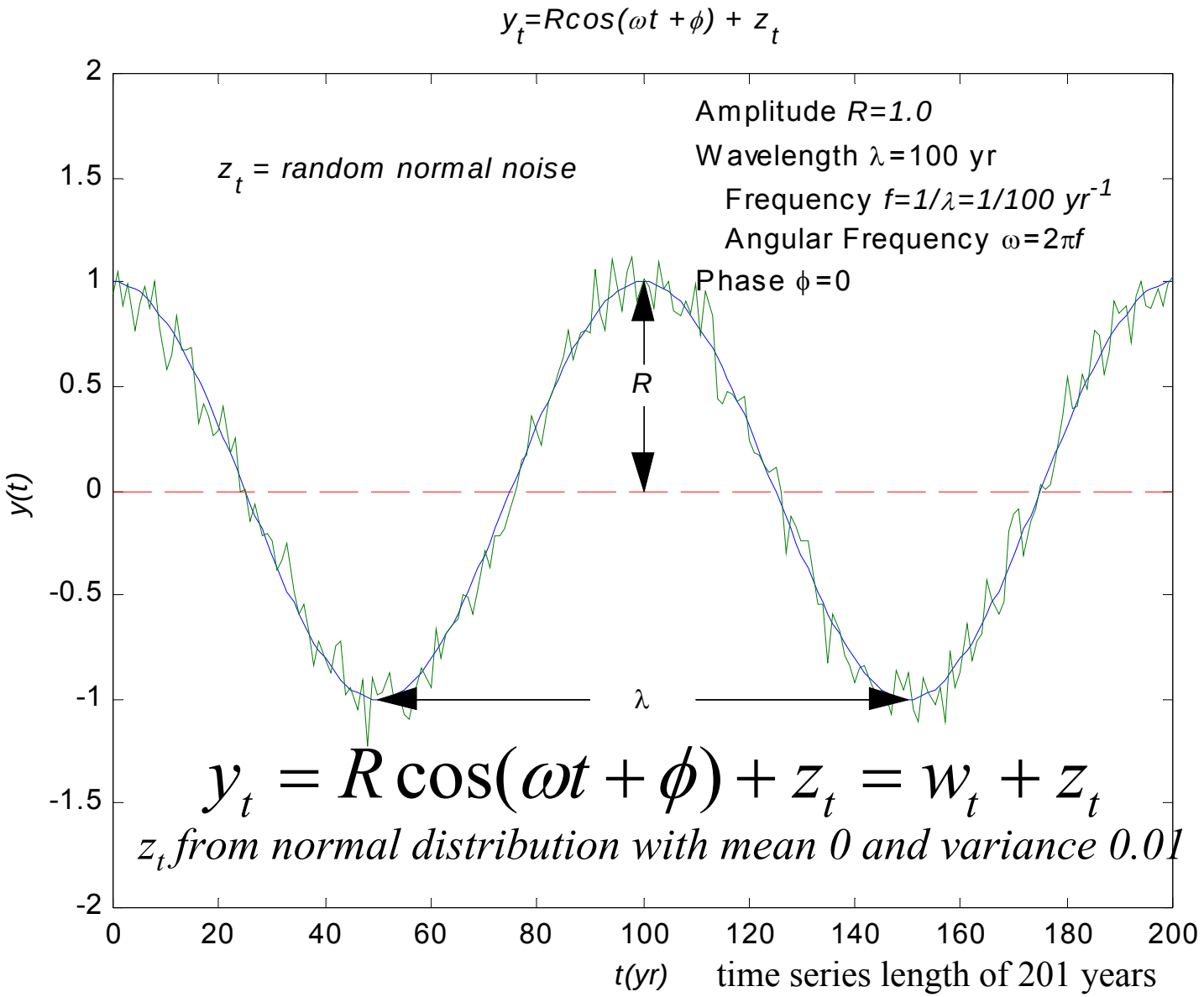
Spectrum: sinusoidal model for a time series

In discussing the frequency domain, it is helpful to start with definitions pertaining to waves. For simplicity, we will use a time increment of one year. Consider the simple example of an annual time series y_t generated by superimposing random normal noise on a cosine wave:

$$y_t = R\cos(\omega t + \phi) + z_t = w_t + z_t$$

where t is time (years, for this example), z_t is the random normal component in year t, w_t is the sinusoidal component; and R, ω and φ are the amplitude, angular frequency (radians per year), and phase of the sinusoidal component.

Spectrum: sinusoidal model for a time series



$$y_t = R\cos(\omega t + \phi) + z_t = w_t + z_t$$

z_t *from normal distribution with mean 0 and variance 0.01*

Spectrum: sinusoidal model for a time series

The ***peaks*** are the high points in the wave; the ***troughs*** are the low points. The wave varies around a mean of zero. The vertical distance from zero to the peak is called the ***amplitude***. The variance of the sinusoidal component is proportional to the square of the amplitude: $var\ (w_t) = R^2 / 2$. The ***phase*** φ describes the offset in time of the peaks or troughs from some fixed point in time.

From the relationship between variance and amplitude, the sinusoidal component in this example has a variance of 50 times that of the noise (0.5 is 50 times 0.01).

The angular frequency ω describes the number of radians of the wave in a unit of time, where 2π radians corresponds to a complete cycle of the wave (peak to peak). In practical applications, the frequency is often expressed by f, the number of cycles per time interval. The relationship between the two frequency measures is given by: $f = \omega / (2\pi)$. The wavelength, or period, of the cosine wave is the distance from peak to peak, and is the inverse of the frequency $\lambda = 1/f$.

Spectrum: sinusoidal model for a time series

A frequency of one cycle per year corresponds to an angular frequency of 2π radians per year and a wavelength of 1 year. The frequency of the cosine wave in previous Figure is $f = 1/100 = 0.01$ cycles per year and the angular frequency is $\omega = 2\pi f = 0.0628$ radians per year

A frequency of one cycle every two years corresponds to an angular frequency of π radians per year, or a wavelength of 2 years. In the analysis of annual time series, this frequency of ***$f = 0.5$ cycles / yr*** or ***$\omega = \pi$ radians / yr*** corresponds to what is called the ***Nyquist frequency***, which is the highest frequency for which information is given by the spectral analysis.

Another important frequency in spectral analysis is the ***fundamental*** frequency, also referred to as the first ***harmonic***. If the length of a time series is *N* years, the fundamental frequency is ***$1/N$*** . The corresponding fundamental period is *N* years, or the length of the time series. For example, the fundamental period of a time series of length 500 years is 500 years – a wave that undergoes a complete cycle over the full length of the time series.

Spectrum: sinusoidal model for a time series

$$X_t = \mu + \sum_{j=1}^{[N/2]} \left[A_j \cos\left(2\pi f_j t\right) + B_j \sin\left(2\pi f_j t\right) \right], \quad t=1,2,\ldots,N$$

where μ is a constant term, the notation $[N/2]$ refers to the greatest integer less than or equal to $N/2$, and the frequencies f_j

$$f_j \equiv j/N, \qquad 1 \le j \le [N/2]$$

where N is the sample size

The frequencies of the sinusoids are at intervals of $1/N$ and are called the Fourier frequencies, or standard frequencies

Fourier, or standard, frequencies

For example, for a 500-year treering series, the standard frequencies are at 1/ 500, 2 / 500,.. cycles per year. The highest standard frequency is $f=(N/2)/N = 1/2 = 0.5$, which corresponds to a wavelength of two years.

Spectrum: sinusoidal model for a time series

Variances at the standard frequencies

A_j and B_j are random variables with expected values of 0

$$E\{A_j\} = E\{B_j\} = 0$$

$$E\{A_j^2\} = E\{B_j^2\} = \sigma_j^2$$

variance at the standard frequencies

$$E\{A_j A_k\} = E\{B_j B_k\} = 0, \text{ for } j \neq k$$

$$E\{A_j B_k\} = 0, \text{ for all } j, k$$

$$E\{X_t\} = \mu$$

Variance at jth standard frequency is proportional to squared amplitude of sinusoidal component at that frequency.

Spectrum: sinusoidal model for a time series

Total variance for sinusoidal model

$$\sigma^2 = E\left\{\left(X_t - \mu\right)^2\right\} = \sum_{j=1}^{[N/2]} \sigma_j{}^2$$

Total variance of series is sum of variance contributions at the N/2 standard frequencies.

The variance of the series X_t is the sum of the sum of the variances associated with the sinusoidal components at the different standard frequencies.

Thus the variance of the series can be decomposed into components at the standard frequencies -- the variance can be expressed as a function of frequency.

Spectrum: sinusoidal model for a time series

Definition of the spectrum in terms of sinusoidal model

spectrum at frequency j → $S_j \equiv \sigma_j^2, \qquad 1 \le j \le [N/2]$

$$\sigma^2 = \sum_{j=1}^{N/2} S_j$$

The spectrum at standard frequency j is defined as the contributed variance at that frequency. The spectrum summed over all standard frequencies therefore equals the total variance of the series.

A plot of S_j against frequencies f_j shows the variance contributed by the sinusoidal terms at each of the standard frequencies.

The shape of the spectral values S_j **plotted against** f_j indicates which frequencies are most important to the variability of the time series

Spectrum: sinusoidal model for a time series

Relation of the spectrum with the autocorrelation function of X_j

spectrum at frequency j

$$\text{acf} \longrightarrow \rho_k = \frac{\sum_{j=1}^{N/2} \sigma_j^2 \cos(2\pi f_j k)}{\sum_{j=1}^{N/2} \sigma_j^2}$$

The acf is expressed as a cosine transform of the spectrum. Similarly the spectrum can be shown to be the Fourier transform of the acf. The spectrum and acf are therefore different characterizations of the same time series information.

The acf is a time-domain characterization and the spectrum is a frequency-domain characterization. From a practical point of view, the spectrum and acf are complementary to each other. Which is most useful depends on the data and the objective of analysis.

Spectrum: Harmonic Analysis

Harmonic Analysis
Periodogram Analysis
Fourier analysis

- Assume sinusoidal model applies exactly
- Compute sinusoidal components at standard frequencies
- Interpret components (e.g., importance as inferred from variance accounted for)

Spectrum: Harmonic Analysis

Harmonic Analysis
Periodogram Analysis
Fourier analysis

In ***harmonic analysis***, the frequencies j/N, $j = 1,...,N/2$ are referred to as the harmonics: $1/N$ is the first harmonic, $2/N$ the second harmonic, etc.
Any series can be decomposed mathematically into its $N/2$ harmonics.
The sinusoidal components at all the harmonics effectively describe all the variance in a series.
A plot of the variance associated with each harmonic as a function of frequency has been referred above as the "***spectrum***", for the hypothesized model.
Such a plot of variance (sometimes scaled in different ways) against frequency is also called the ***periodogram*** of the series, and the analysis is called ***periodogram analysis***
Spectral analysis, departs from ***periodogram analysis*** in an important way: in spectral analysis, *the time series is regarded as just one possible realization from a random process*, and the objective is to estimate the spectrum of that process using just the observed time series.

Spectrum: Spectral Analysis

- View the time series as short random sample from infinitely long series; a single realization of a process
- Acknowledge that random sampling fluctuations can produce spurious peaks in the computed periodogram of the short sample
- Using the sample, estimate the spectrum of this infinitely long series (population), explicitly accounting for sampling variability

Spectrum: Spectral Analysis

For any stationary stochastic process with a population autocovariance function acvf $\gamma(k)$, there exists a monotonically increasing function, $F(\omega)$, such that

$$\text{acvf} \longrightarrow \gamma(k) = \int_0^{\pi} \cos \omega k dF(\omega)$$

where $\gamma(k)$, is the spectral representation of the autocovariance function,
and $F(\omega)$ is called the *spectral distribution function*.

$F(\omega)$ has a direct physical interpretation, it is the contribution to the variance of the series which is accounted for by frequencies in the range $(0,\omega)$

Spectrum: Spectral Analysis

Spectral distribution function:

$F(\omega) =$ contribution to the variance of the series which is accounted for by frequencies in the range $(0,\omega)$

normalized spectral distribution function:

$$F^*(\omega) = F(\omega)/\sigma_X^2$$

which gives the *proportion* of variance accounted for by frequencies in the range $(0,\omega)$, and like a cdf, reaches a maximum of 1.0, since $F^*(\pi)=1$.

Spectrum: Spectral Analysis

Spectral density function or spectrum:

$$f(\omega) = \frac{\mathrm{d}F(\omega)}{d\omega} \equiv \text{(power) spectral density function}$$

The spectrum is the derivative of the spectral distribution function with respect to frequency. A point on the spectrum therefore represents the "variance per unit of frequency" at a specific frequency. If $\mathrm{d}\omega$ is an increment of frequency, the product $f(\omega)\mathrm{d}\omega$ is the contribution to the variance from the frequency range $(\omega,\omega+\mathrm{d}\omega)$.

In a graph of the spectum, therefore, the area under the curve bounded by two frequencies represents the variance in that frequency range, and the total area underneath the curve represents the variance of the series. A peak in the spectrum represents relatively high variance at frequencies in corresponding region of frequencies below the peak.

Spectrum: Spectral Analysis

Calculation of the spectrum of a time series

The acvf can be expressed as a cosine transform of the spectral density function, or spectrum. The inverse relationship is the ***Fourier transform*** **of the acfv:**

$$f(\varpi) = \frac{1}{\pi}\left[\gamma(0) + 2\sum_{k=1}^{\infty}\gamma(k)\cos\varpi k\right]$$

The ***normalized spectrum*** is accordingly defined as

$$f^*(\varpi) = f(\varpi)/\sigma_X^2$$

$$f^*(\varpi) = \frac{1}{\pi}\left[1 + 2\sum_{k=1}^{\infty}\rho(k)\cos\varpi k\right]$$

k is the lag

which therefore gives the normalized Fourier transform of the acf and an "obvious" estimator for the spectrum is the Fourier transform of the complete sample acvf.

Spectrum: Spectral Analysis

Calculation of the spectrum of a time series using Blackman-Tukey method

The Blackman-Tukey applies the Fourier transform to **a truncated, smoothed acvf rather than to the entire acvf.**

The Blackman-Tukey estimation method consists of taking a Fourier transform of the truncated sample acvf using a weighting procedure. Because the precision of the acvf estimates decreases as lag k increases, it seems reasonable to give less weight to the values of the acvf at high lags. Such an estimator is given by:

$$f^*(\varpi) = \frac{1}{\pi}\left[\lambda_0 c_0 + 2\sum_{k=1}^{M} \lambda_k c_k \cos \varpi k\right]$$

where λ_k are the weights called the ***lag window***, and $M(< N)$ is called the *truncation point*. They are selected decreasing weight toward higher lags, such that the higher-lag acvf values are discounted.

Spectrum: Spectral Analysis

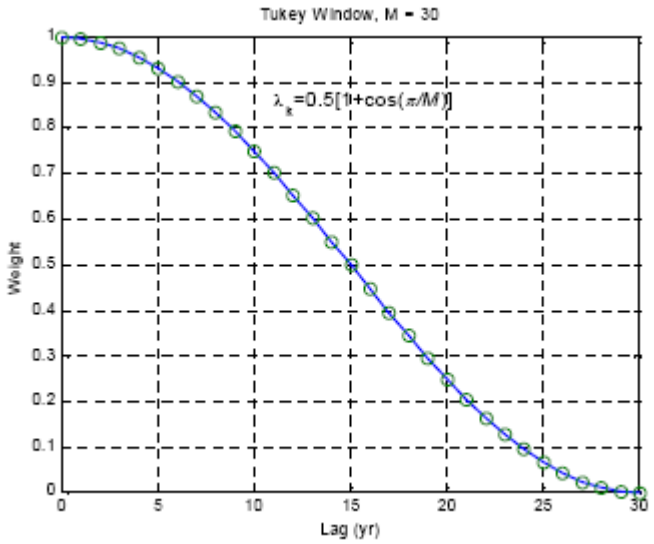
Calculation of the spectrum of a time series using Blackman-Tukey method

One popular form of lag window is the ***Tukey window.***

$$\lambda_k = 0.5\left(1+\cos\frac{\pi k}{M}\right), k = 0,1,..,M$$

where k is the lag, M is the width of the lag window –also called the *truncation point* -, and λ_k *is the weight at lag k.*

The window for a lag window of width 30 is shown in the figure. Tthe truncation point *M must be choosen.* This is generally done by trial and error, with a subjective evaluation of which window best displays the important spectral features. The choice of *M* affects the *bias, variance,* and *bandwidth* of the spectral estimations



Spectrum: Spectral Analysis

Calculation of the spectrum of a time series using Blackman-Tukey method

Smaller *M* means increased bias. *Bias* refers to the tendency of spectral estimates to be less extreme (both highs and lows) than the true spectrum. Increased bias is manifested in a "flattening out" of the estimated spectrum, such that peaks are not as high as they should be, and troughs not as low. This bias is acknowledged, but not explicitly expressed as a function of *M*.

Smaller M means smaller variance of spectral estimates (narrower confidence bands)

Smaller M means increased bandwidth (decreased resolution of frequency of features)

It has to be chosen subjectively so as to balance 'resolution' against 'variance'. The smaller the value of *M*, the smaller will be the variance of $\hat{f}(\omega)$ but the larger will be the bias

Spectrum: Spectral Analysis

Aliasing effect

Aliasing refers to the phenomenon in which spectral features on the frequency range {0, 0.5}can be produced by variability at frequencies higher than the Nyquist frequency (i.e., higher than f = 0.5 cycles per year).

Whether aliasing is a problem or not depends on the sampling interval and the frequencies of variability in the data, and is most easily illustrated for sampled rather than aggregated time series.

Aliasing produces false spectrum peaks

Spectrum: Spectral Analysis

Example of Aliasing effect

For example, imagine a time series of air temperature and a sampling interval of ***18 hours***. The Nyquist frequency corresponds to a wavelength of twice the sampling interval, or ***36 hours***. Air temperature has roughly a diurnal, or ***24-hour***, cycle – a cycle at a higher frequency than the Nyquist frequency.

If the first sample happens to coincide with the daily peak, the second sample will be ***6 hr*** before the peak on the second day, the third sample will be ***12 hr*** before the peak on the third day, the fourth sample will be ***18 hr*** before the peak on the fourth day, and the fifth sample will be ***24 hr*** before the peak on the fifth day. This fifth sample is again at the daily peak. If the sampling is continued, the series would tend to peak at observations ***1, 5, 9, 13, etc***. The spacing between peaks is 4 sample points, or ***4x18=72 hr***.

A spectrum of the series sampled in this way would have a spectral peak at wavelength ***72 hrs***. This is a false spectral peak, and is really the ***24-hr cycle aliased to 72 hours.***

Spectrum: Spectral Analysis

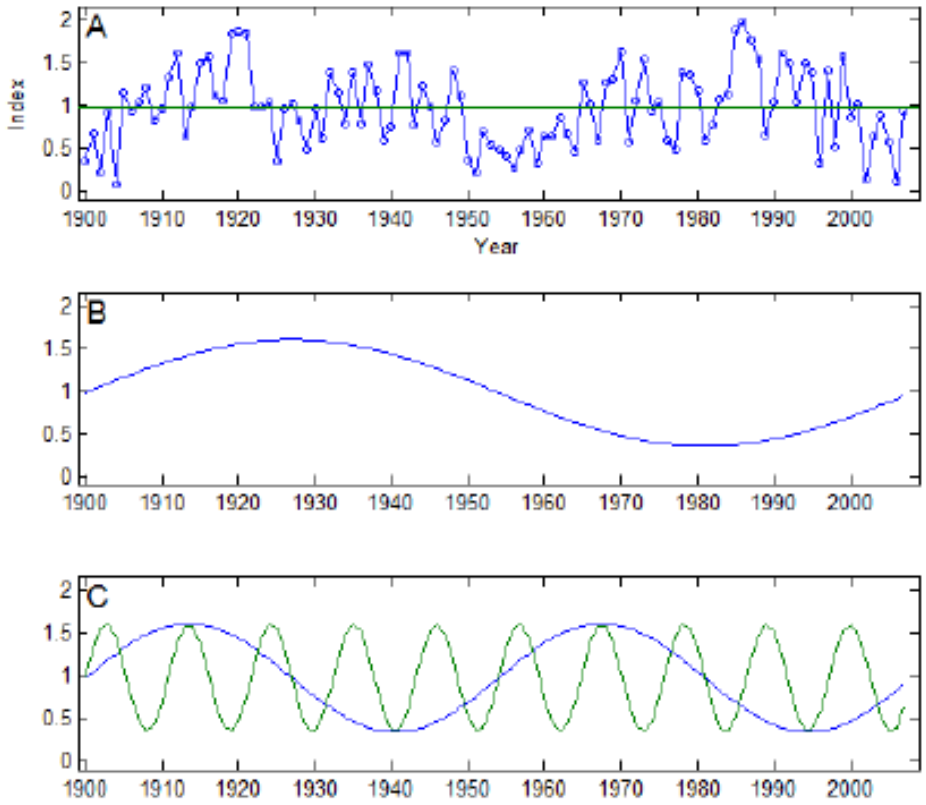


Figure 4.3. Illustration of fundamental and Fourier frequencies. (A) Tree-ring index MEAF, with length 108 years. (B) Sinusoidal time series at fundamental frequency (wavelength 108 yr) of the tree-ring series. (C) Sinusoidal time series at Fourier frequencies 2/N and 10/N (wavelengths 54 yr and 10.8 yr), where N=108 years. Series in B and C are scaled to same mean and variance as tree-ring series. As the tree-ring series is not periodic, its peaks and troughs are irregularly spaced, unlike those of the sinusoids. Theoretically, the variance of the tree-ring series in (A) could be decomposed into contributions from all Fourier frequencies.

Spectrum: Spectral Analysis

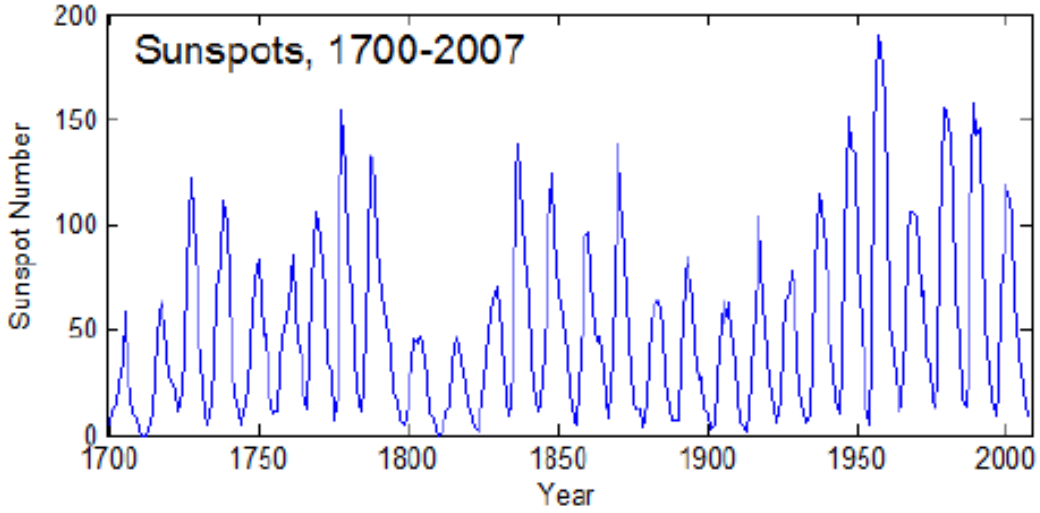


Figure 4.4. Time series of Wolf Sunspot Number. Source: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/SUNSPOT_NUMBERS/YEARLY.PLT

Spectrum: Spectral Analysis

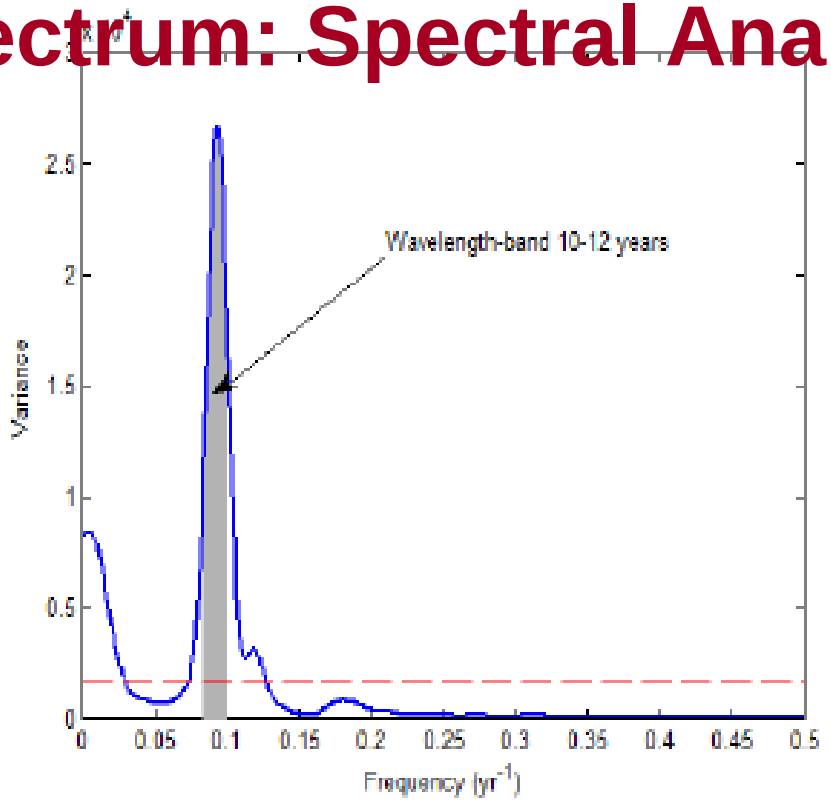


Figure 4.5. Spectrum of Wolf Sunspot Cycle, 1700-2007. Shading covers frequency range (1/12) yr^{-1} to (1/10) yr^{-1}, corresponding to wavelengths between 10 and 12 years. Ratio of shaded area to total area under curve is fraction of series variance contributed by wavelength-band 10-12 years. Horizontal dashed line is at the mean of the spectrum; its ordinate is the variance of the time series (here, variance≈0.1635E4). Area under dashed line equals area under solid curve, and is proportional to the series variance. The coefficient of proportionality is ½ given that the frequency axis extends from 0 to 0.5.

Spectrum: Spectral Analysis

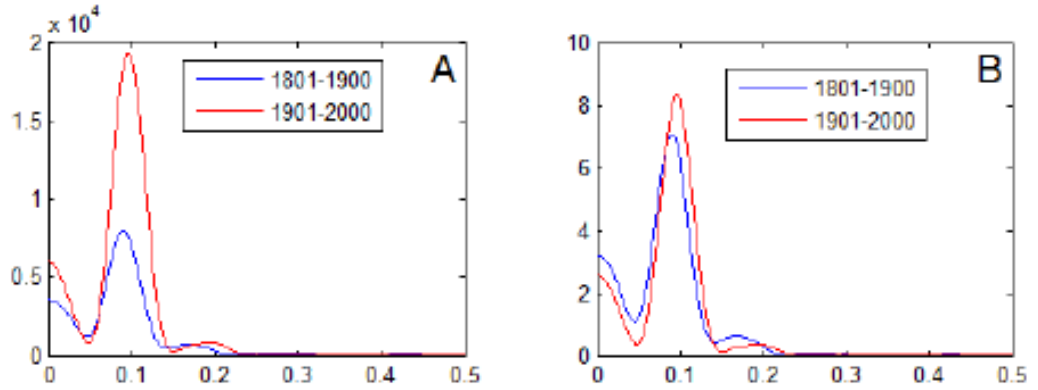


Figure 4.6. Comparison of spectra and normalized spectra. (A) Spectra for two different sub-periods of the Wolf Sunspot series. (B) Normalized spectra for the same two sub-periods. Series variance for 1801-1900 is only about half that for 1901-2000. From the time series plot (Figure 4.4), it can be seen that the variance for 1801-1900 is lower than for 1901-2000. The greater area under the spectrum for 1901-2000 reflects this difference (A). The differences in total variance have been removed in the normalized spectra (B), and the areas under the two spectra are equal.

Spectrum: Spectral Analysis

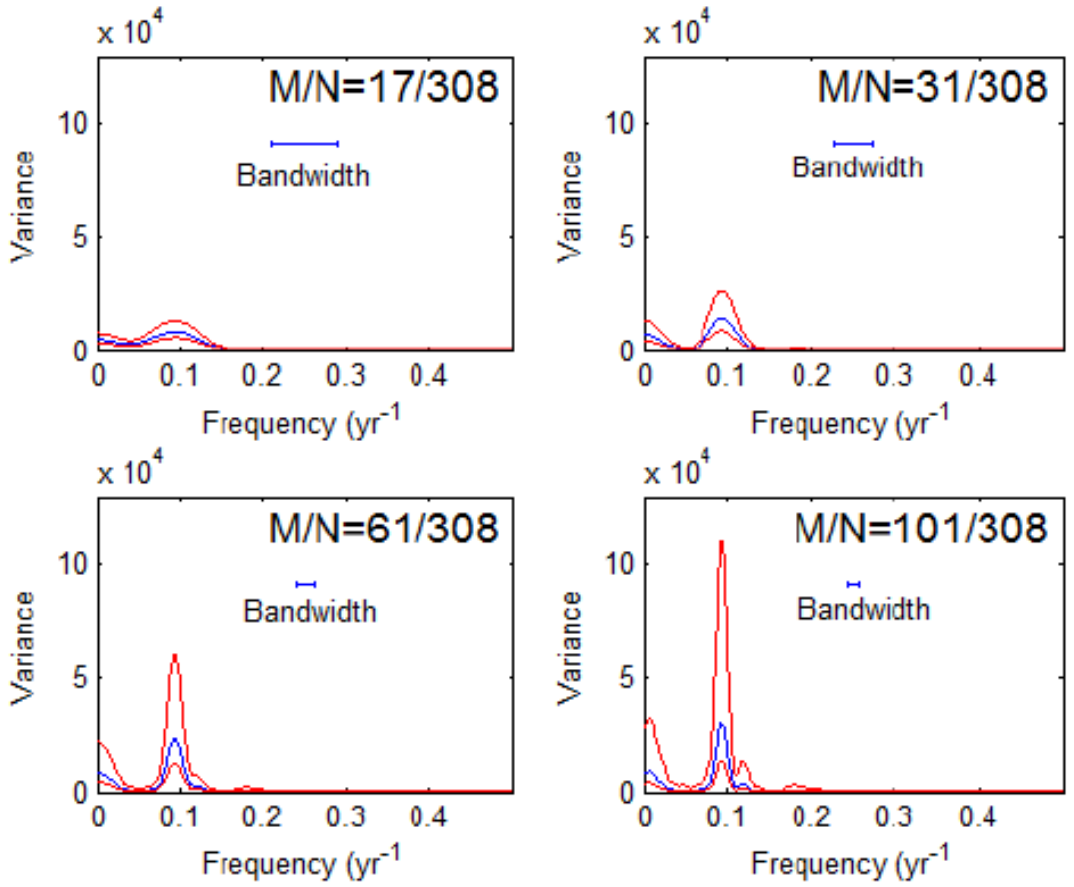


Figure 4.8. Illustration of "window closing" process in spectral estimation. Plots are spectra of 1700-2007 Wolf Sunspot Number using different setting of lag window M. With increasing M the following occurs: 1) bandwidth narrows, allowing greater frequency resolution of the main peak near 11 years, 2) spectral peak near 11 years becomes larger, and 3) confidence interval around spectrum widens.

Spectrum: Spectral Analysis

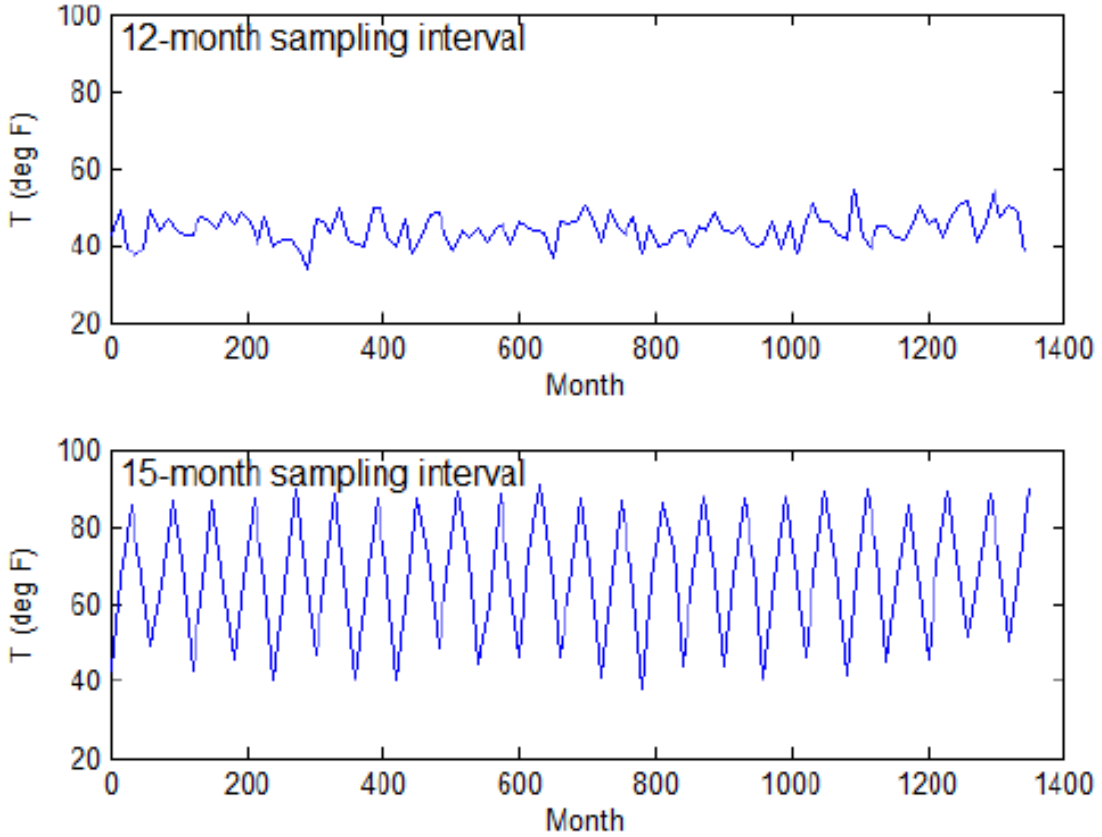


Figure 4.9. Ilustration of aliasing for a mean monthly maximum temperature time series. The original data (not plotted) is PRISM monthly mean maximum temperature, 1895-2007 (1356 months) for 106°W, 36°N. At top is the series sampled at a 12-month interval (months 1, 13, 25, …, or every January). At bottom is the series sample at a 15-month interval (months 1, 16, 31, ….). Bottom plot has 10 peaks per 600 months, or a 5-year cycle.