Exploring Time Series Data Using StatsModels

Janani Ravi

CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Specialized models for time-series data, x-axis denotes time

Autoregressive models have y variables which depend on previous y values

White noise error terms

Moving average models over white noise

ARMA models combine autoregression and moving averages

Stationarity and Time Series Data

Time Series

A time series is a sequence of data taken at successive and usually equally spaced points in time.

Time Series

A time series is a sequence of data taken at successive and usually equally spaced points in time.

Time Series

A time series is a sequence of data taken at successive and usually equally spaced points in time

Types of Time Series Models

AR(p)

Autoregressive models

MA(q)

Moving average models

ARMA(p,q)

Combination of other two

Time series models are especially vulnerable to problems of **non-stationarity**

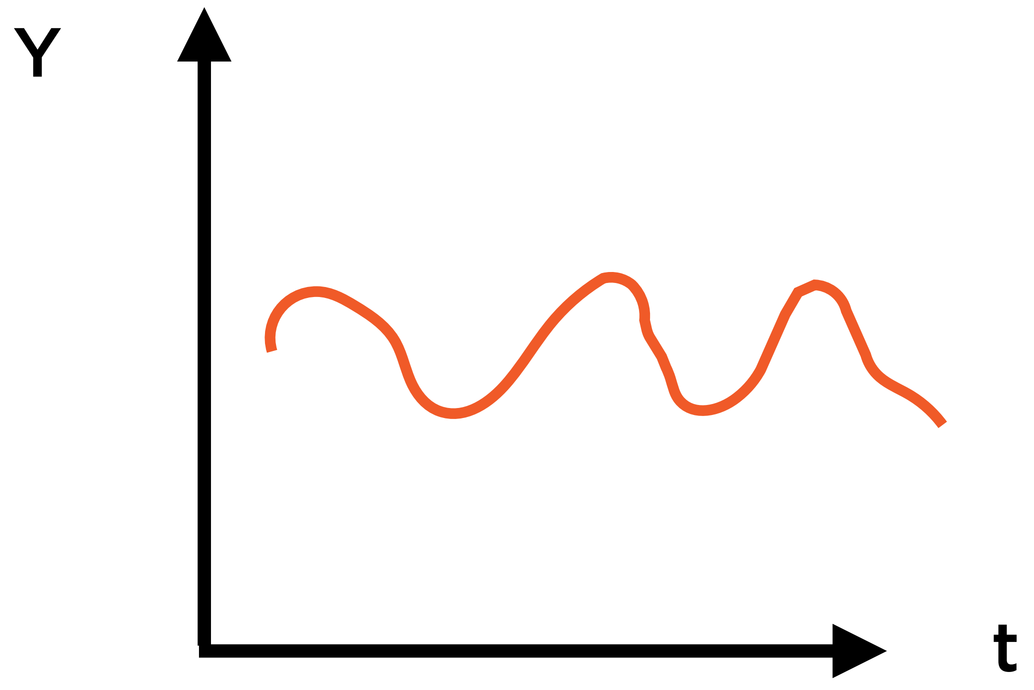
Non-stationary Data

Mean changes over time

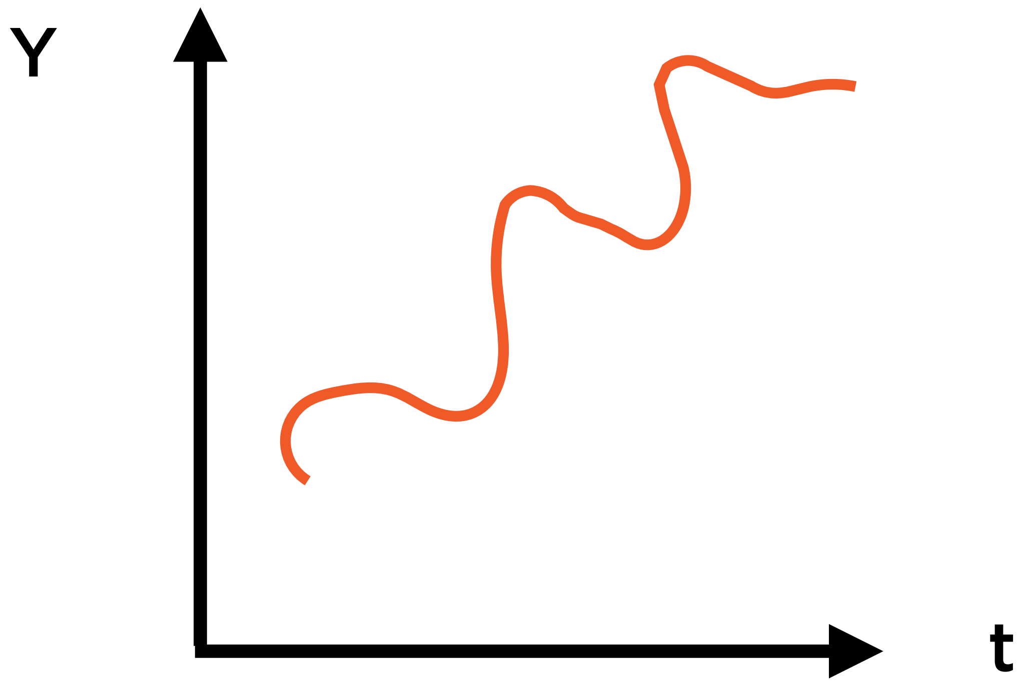
Variance changes over time

Autocorrelation changes over time

Varying Mean



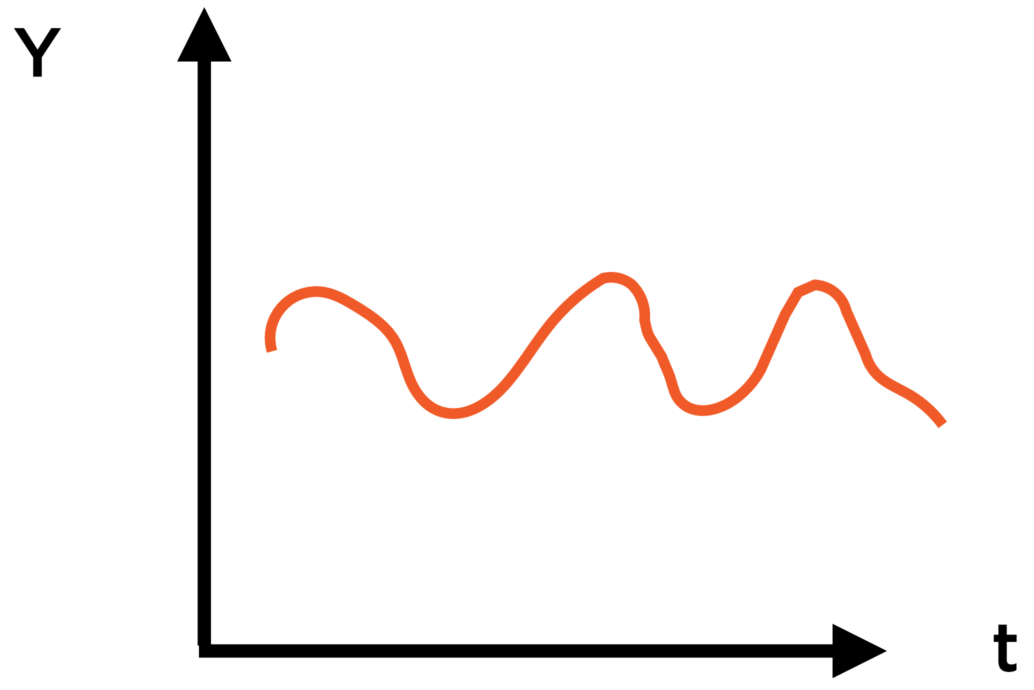
Stationary



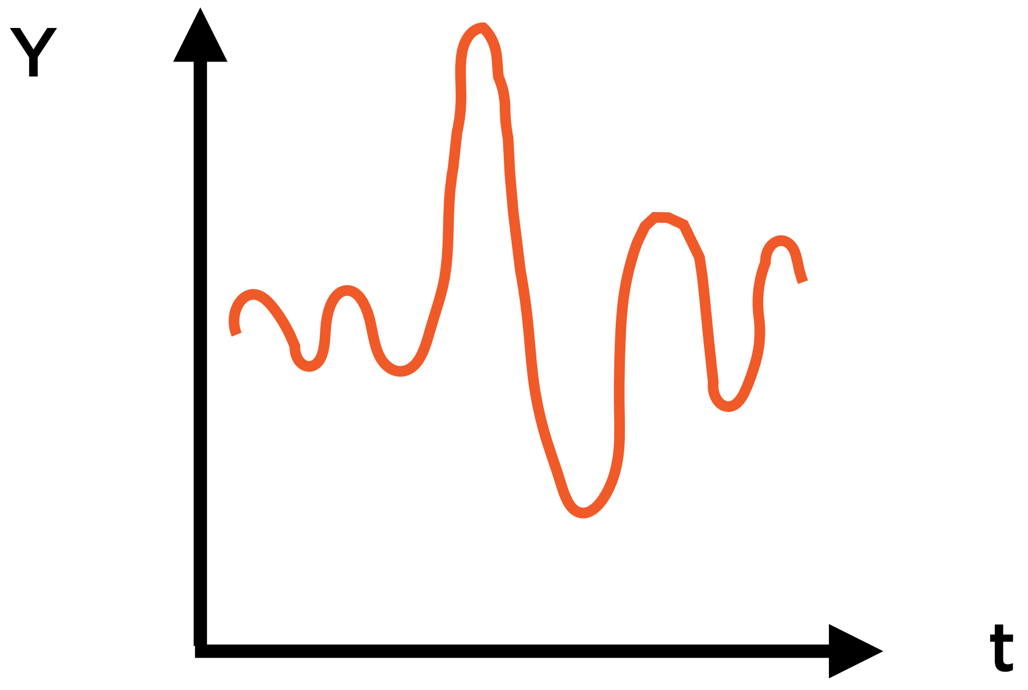
Non-stationary

Time series that trends over time is non-stationary - mean is changing over time

Varying Variance



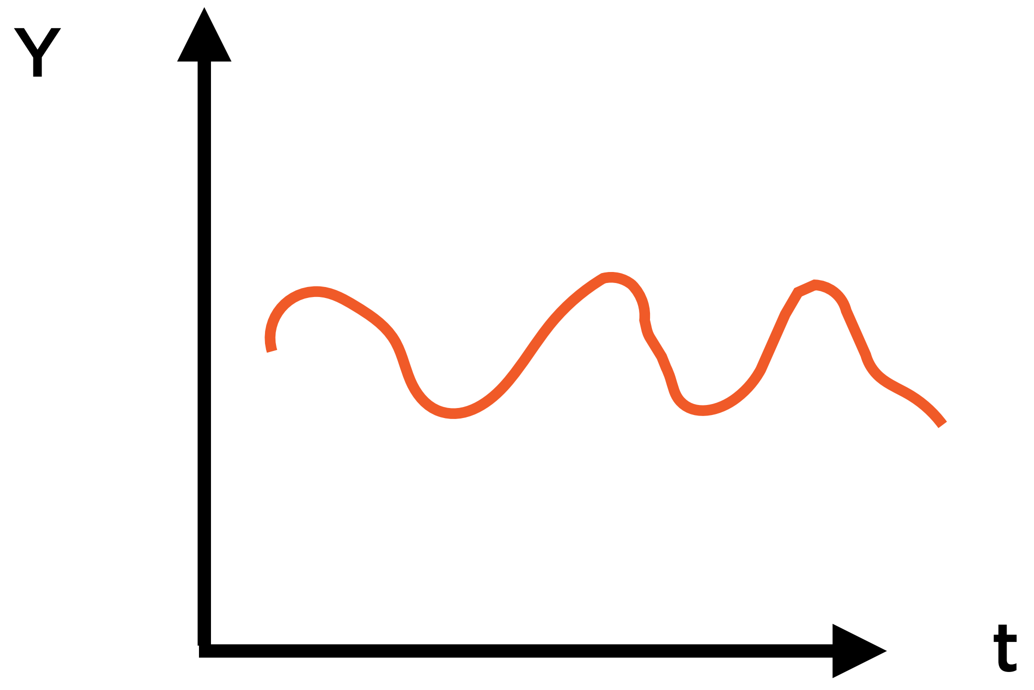
Stationary



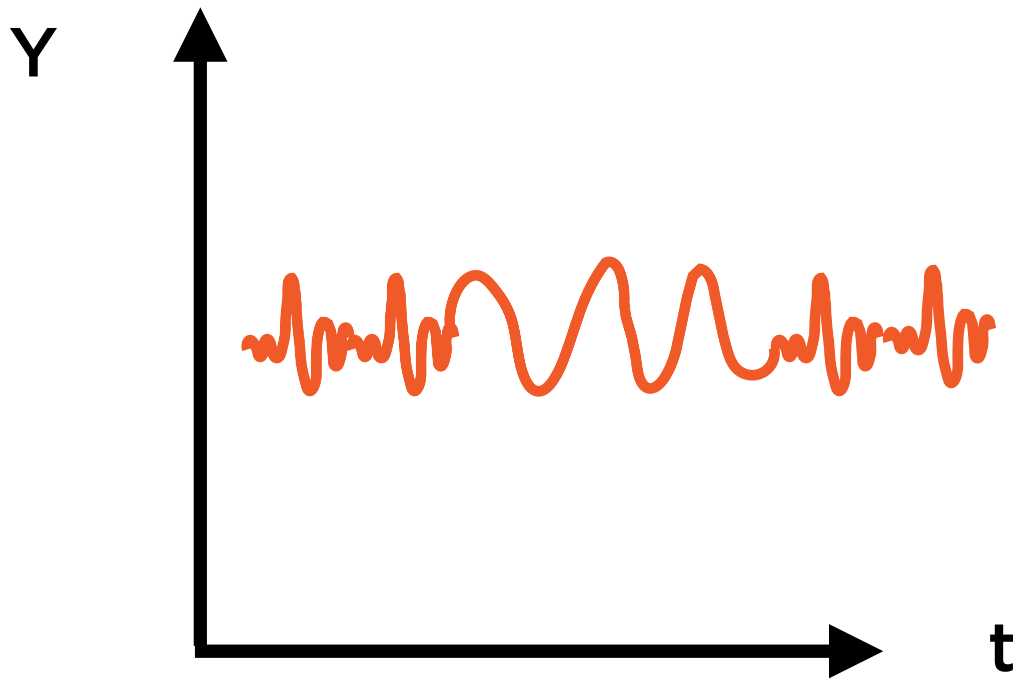
Non-stationary

Time series that has periods of higher volatility is non-stationary - variance is changing over time

Varying Autocorrelation

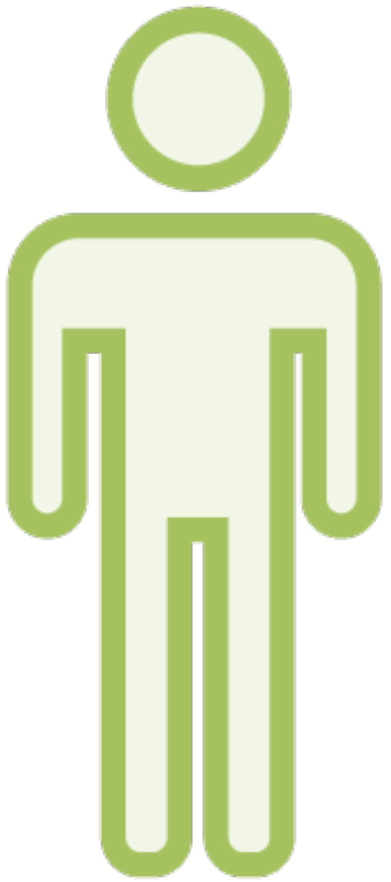


Stationary



Non-stationary

The spread in the data becomes further away and then closer

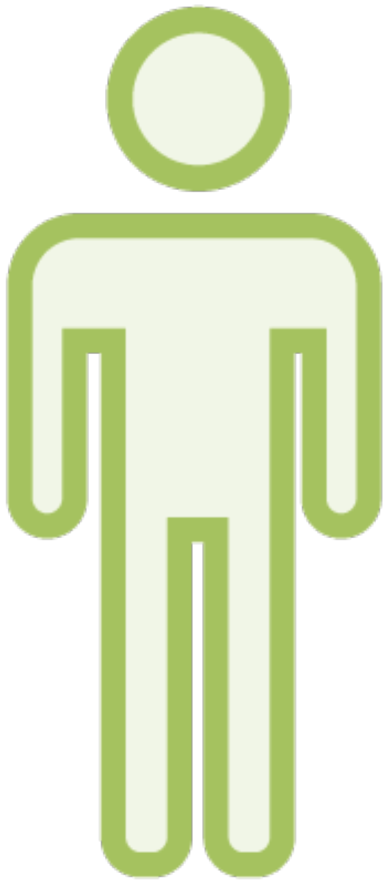


Stationary Data

Mean of time series does not change over time

Variance of time series does not change over time (homoscedasticity)

Autocorrelation does not change over time

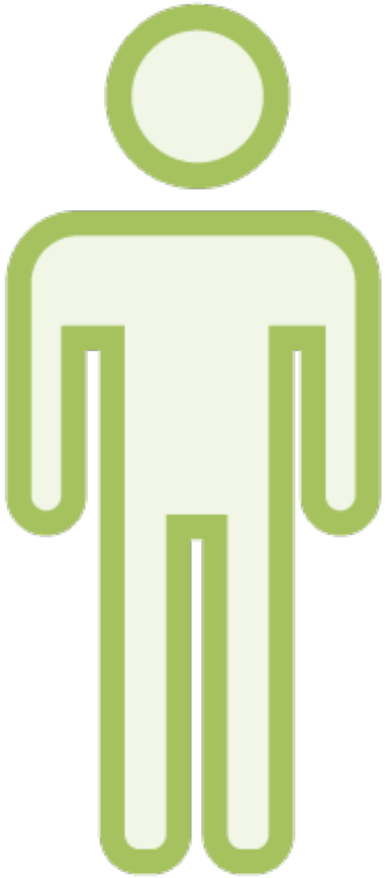


Stationary Data

Applying regression to non-stationary data yields poor model

Inflated R^2

Problems associated with heteroscedasticity



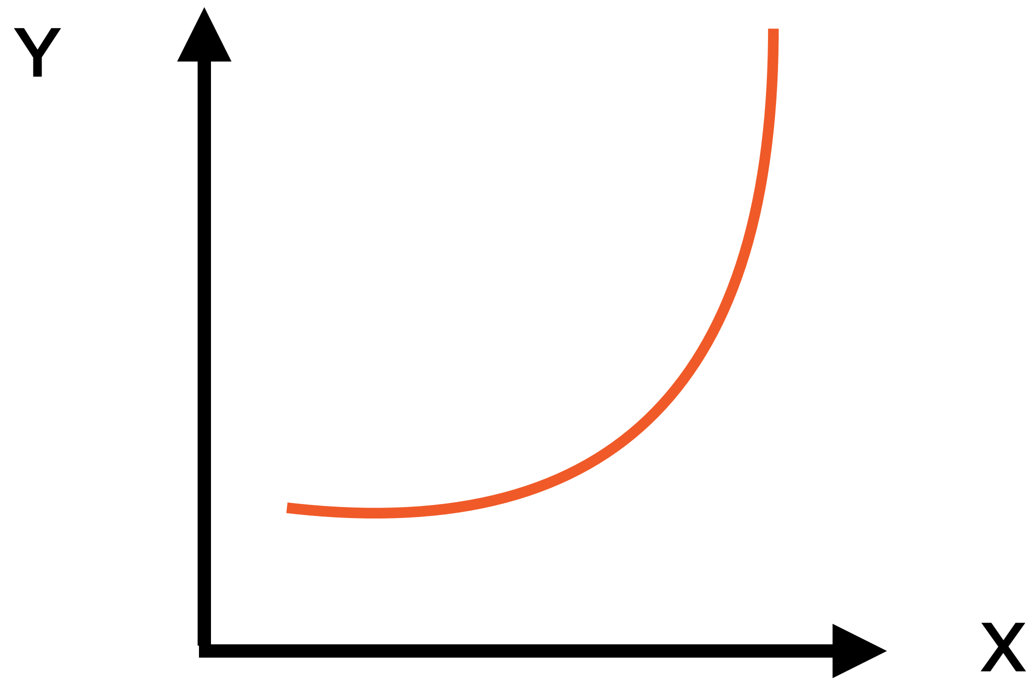
Stationary Data

Statistical tests exist to test for non-stationarity

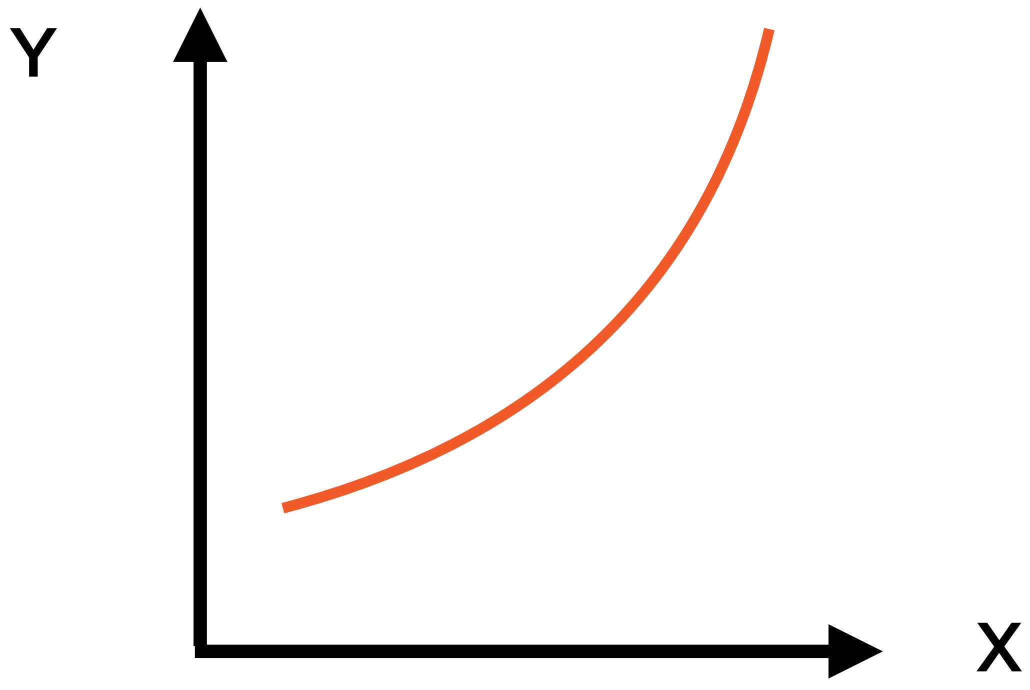
In practice, simple forms of non-stationarity can be found from plotting data

More complex forms require statistical tests

Beware of Non-stationary Data



Exponential



Polynomial

Smoothly trending data will lead to poor quality regression and time series models

Convert Series to Returns

$$y'_{12} = \log y_2 - \log y_1$$

$$x'_{12} = \log x_2 - \log x_1$$

Log Differences

$$y'_{12} = (y_2 - y_1)/y_1$$

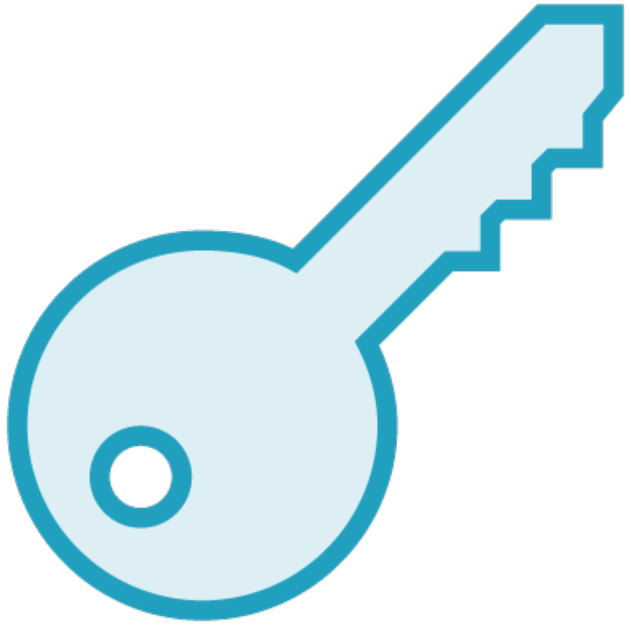
$$x'_{12} = (x_2 - x_1)/x_1$$

Returns

Take first differences of smooth data converting either to log differences or returns

Autoregressive and Moving Average Models

X Causes Y



Cause

Independent variable

Effect

Dependent variable

Linear Regression

$$y = A + Bx$$

$$
\begin{array}{ccc}
y_1 & = & A + Bx_1 \\
y_2 & = & A + Bx_2 \\
y_3 & = & A + Bx_3 \\
\ldots & & \ldots \\
y_n & = & A + Bx_n
\end{array}
$$

Linear Regression

$$y = A + Bx$$

$$
\begin{aligned}
y_1 &= A + Bx_1 + e_1 \\
y_2 &= A + Bx_2 + e_2 \\
y_3 &= A + Bx_3 + e_3 \\
&\cdots \\
y_n &= A + Bx_n + e_n
\end{aligned}
$$

Y_{t-1} Causes Y_t

Cause

Rain yesterday

Effect

Rain today as well

Autoregression

$$y_t = A + By_{t-1}$$

$$\begin{aligned}
y_1 &= A + By_0 \\
y_2 &= A + By_1 \\
y_3 &= A + By_2 \\
&\ldots \\
y_n &= A + By_{n-1}
\end{aligned}$$

Autoregression

$$y_t = A + By_{t-1}$$

$$\begin{aligned}
y_1 &= A + By_0 + e_1 \\
y_2 &= A + By_1 + e_2 \\
y_3 &= A + By_2 + e_3 \\
&\dots \\
y_n &= A + By_{n-1} + e_n
\end{aligned}$$

$$Y_t = C + \sum_{i=1}^{p} \phi_i \boxed{X_t} + \epsilon_t$$

General Form of Linear Model

The error terms ϵ_t in OLS are assumed to be zero-mean, constant-variance and normally distributed

$$Y_t = C + \sum_{i=1}^{p} \phi_i \boxed{Y_{t-i}} + \epsilon_t$$

General Form of Autoregressive Model

Notice that Y is now on both sides of the equation - hence the name

$$\boxed{Y_t} = C + \sum_{i=1}^{p} \phi_i \boxed{Y_{t-i}} + \epsilon_t$$

General Form of Autoregressive Model

Notice that Y is now on both sides of the equation - hence the name

$$Y_t = C + \sum_{i=1}^{\boxed{p}} \phi_i Y_{t-i} + \epsilon_t$$

AR(p) Model

Because last p values of Y influence current value of Y

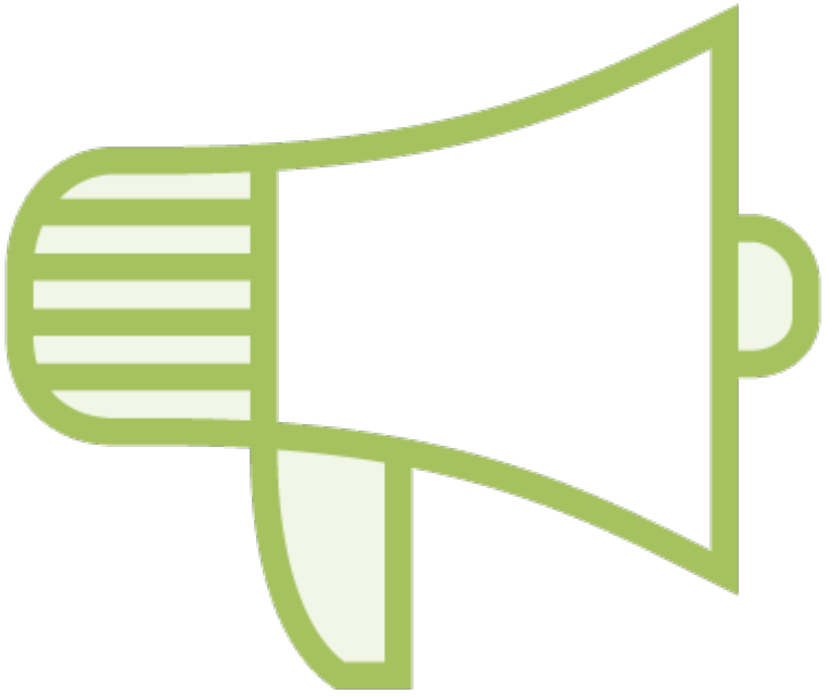
Autoregressive Models

Future values of Y depend on **past** values of Y and on **current** value of white noise

$$Y_t = C + \sum_{i=1}^{p} \phi_i Y_{t-i} + \boxed{\epsilon_t}$$

White Noise Error Terms

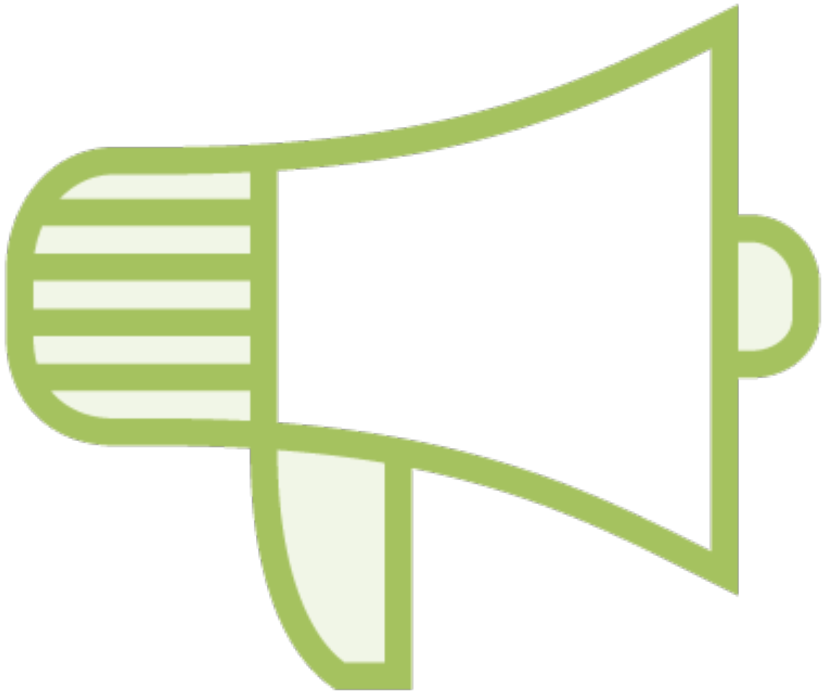
The error terms ϵ_t in AR(p) are still assumed to be zero-mean, constant-variance and normally distributed



White Noise Error Terms

Error terms form a white noise process

- Mean zero
- Constant variance
- Normally distributed
- Independent and Identically Distributed (IID)



White Noise Error Terms

AR models define $Y_t \sim Y_{t-1}\ Y_{t-2} \ldots Y_{t-p}$

Have a single error term ϵ_t

Can also define $Y_t \sim \epsilon_{t-1}\ \epsilon_{t-2} \ldots \epsilon_{t-q}$

Such models are called MA models

MA(q) ~ Moving Average of last q values of ϵ

$$Y_t = C + \sum_{i=1}^{\boxed{p}} \boxed{\phi_i Y_{t-i}} + \epsilon_t$$

AR(p) Model

Because last p values of Y influence current value of Y

$$Y_t = \mu + \boxed{\epsilon_t + \theta_1 \epsilon_{t-1} + \ldots.. + \theta_q \epsilon_{t-q}}$$

MA(q) Model

Value of Y depends on last q values of the white noise process

Moving Average Models

Future values of Y depend on **past** values of white noise alone

$$Y_t = \boxed{\mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots\ldots + \theta_q \epsilon_{t-q}} + \boxed{\phi_1 Y_{t-1} + \ldots\ldots + \phi_p Y_{t-p}}$$

ARMA(p,q) Model

Combine AR(p) and MA(q)

$$Y_t = \boxed{\mu} + \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots.. + \theta_q \epsilon_{t-q} + \phi_1 Y_{t-1} + \ldots.. + \phi_p Y_{t-p}$$

ARMA(p,q) Model

μ is the mean (constant)

$$Y_t = \mu + \boxed{\epsilon_t} + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q} + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p}$$

ARMA(p,q) Model

ϵ_t is the current period error; zero-mean, constant-variance, normally distributed

$$Y_t = \mu + \epsilon_t + \boxed{\theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q}} + \phi_1 Y_{t-1} + ... + \phi_p Y_{t-p}$$

ARMA(p,q) Model

MA(q) component of the ARMA(p,q)

$$Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots\ldots + \theta_q \epsilon_{t-q} + \boxed{\phi_1 Y_{t-1} + \ldots\ldots + \phi_p Y_{t-p}}$$

ARMA(p,q) Model

AR(p) component of the ARMA(p,q)

ARMA Models

Future values of Y depend on **past** values of Y and on **current** and **past** values of white noise

Specifying AR, MA and ARMA Models

Finding p,q in AR(p) and MA(q)

Correlation

Autocorrelation

ACF Plot

Find q for
MA(q) Model

Partial Correlation

Partial Autocorrelation

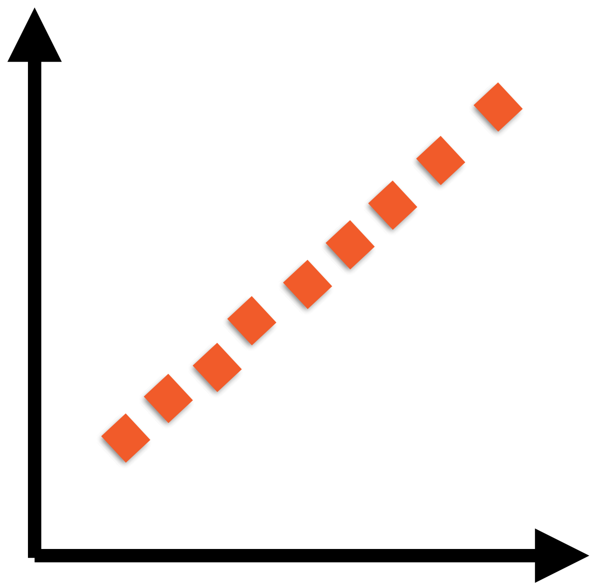
PACF Plot

Find p for
AR(p) Model

Correlation

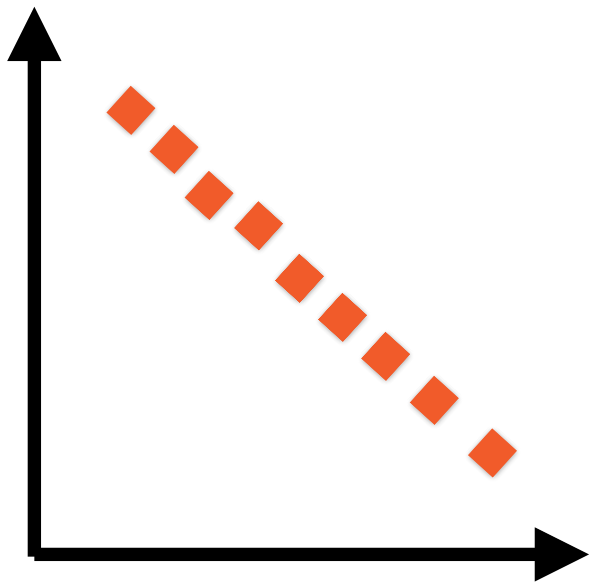
The measure of the relationship between two items or variables

Positive and Negative Correlation



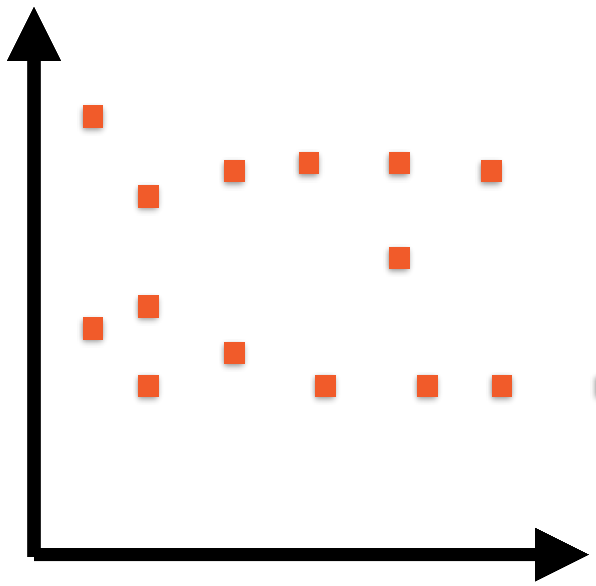
Correlation = +1

As X increases, Y increases linearly



Correlation = -1

As X increases, Y decreases linearly



Correlation = 0

Changes in X independent* of changes in Y

self

Autocorrelation

Autocorrelation

Measures the relationship between a variable’s current value and past value

Partial Autocorrelation

Conceptually similar to autocorrelation; based on partial correlation of a series with lagged versions of itself

Autocorrelation

More likely

Today

Tomorrow

Autocorrelation

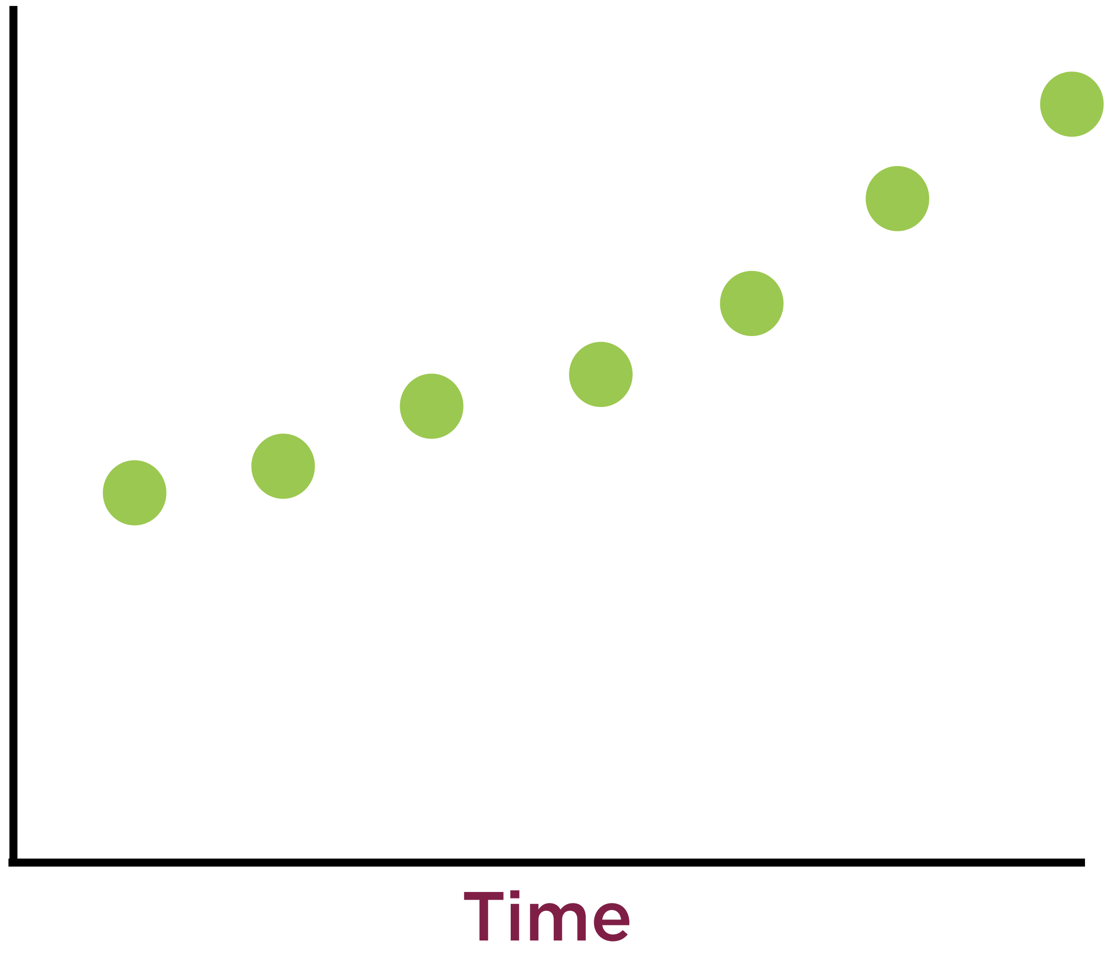
Less likely

Today

Tomorrow

Autocorrelation

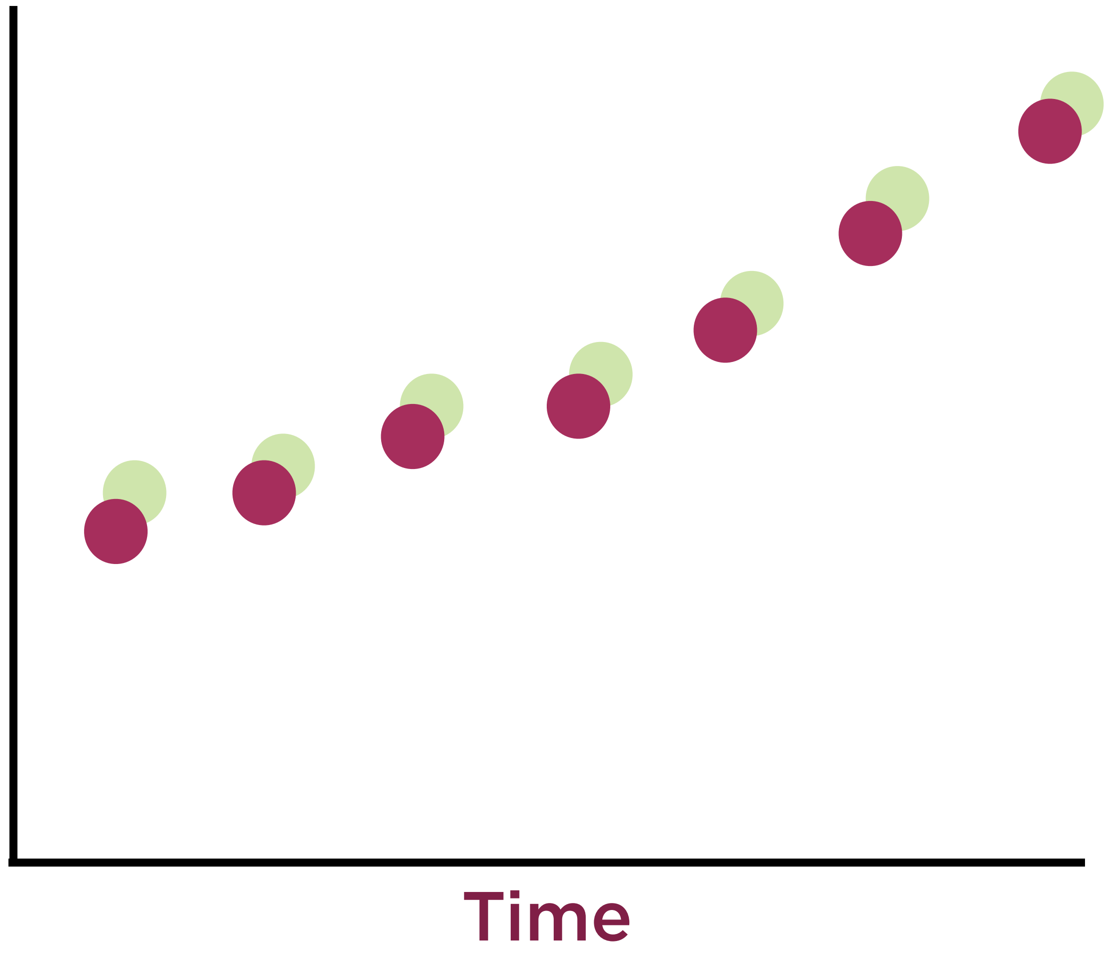
Same time series is used twice



Original form

Autocorrelation

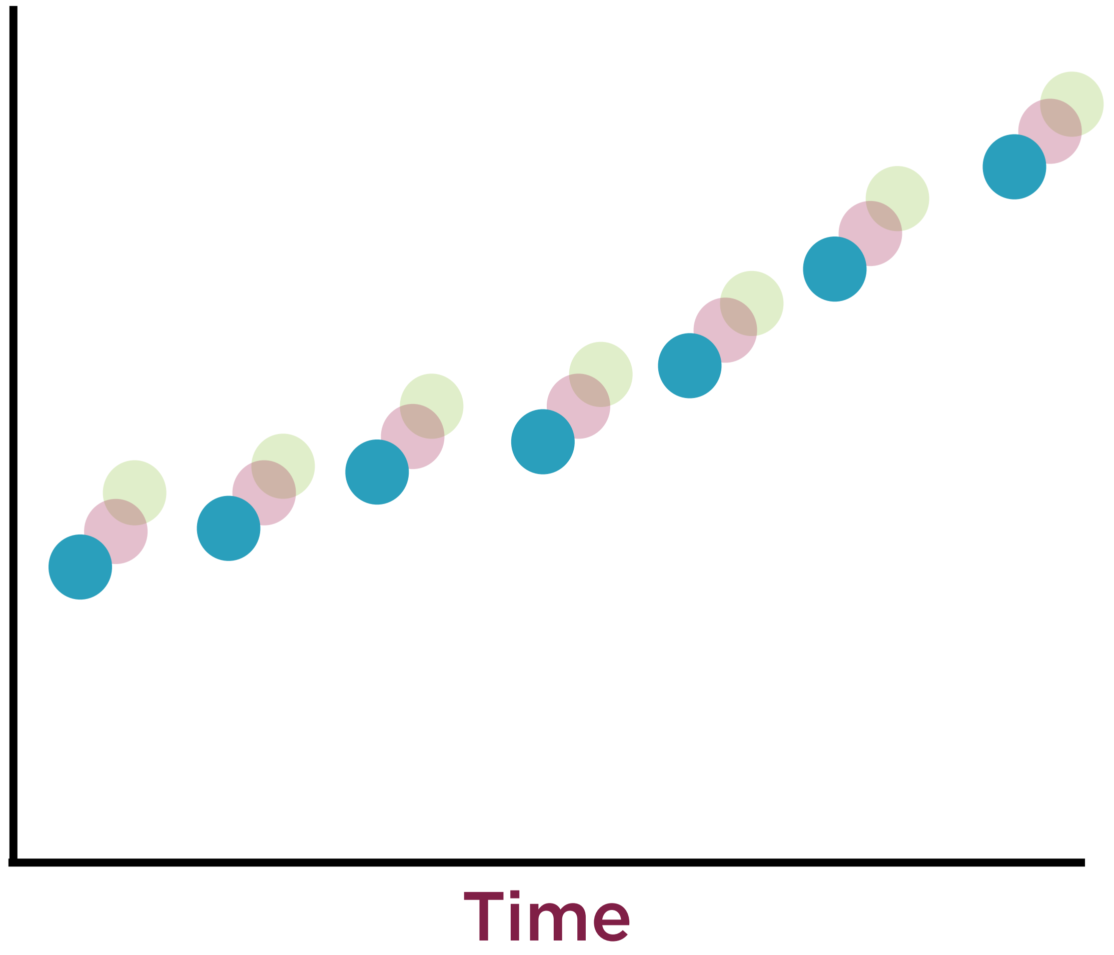
Same time series is used twice



Lagged over one or more time periods

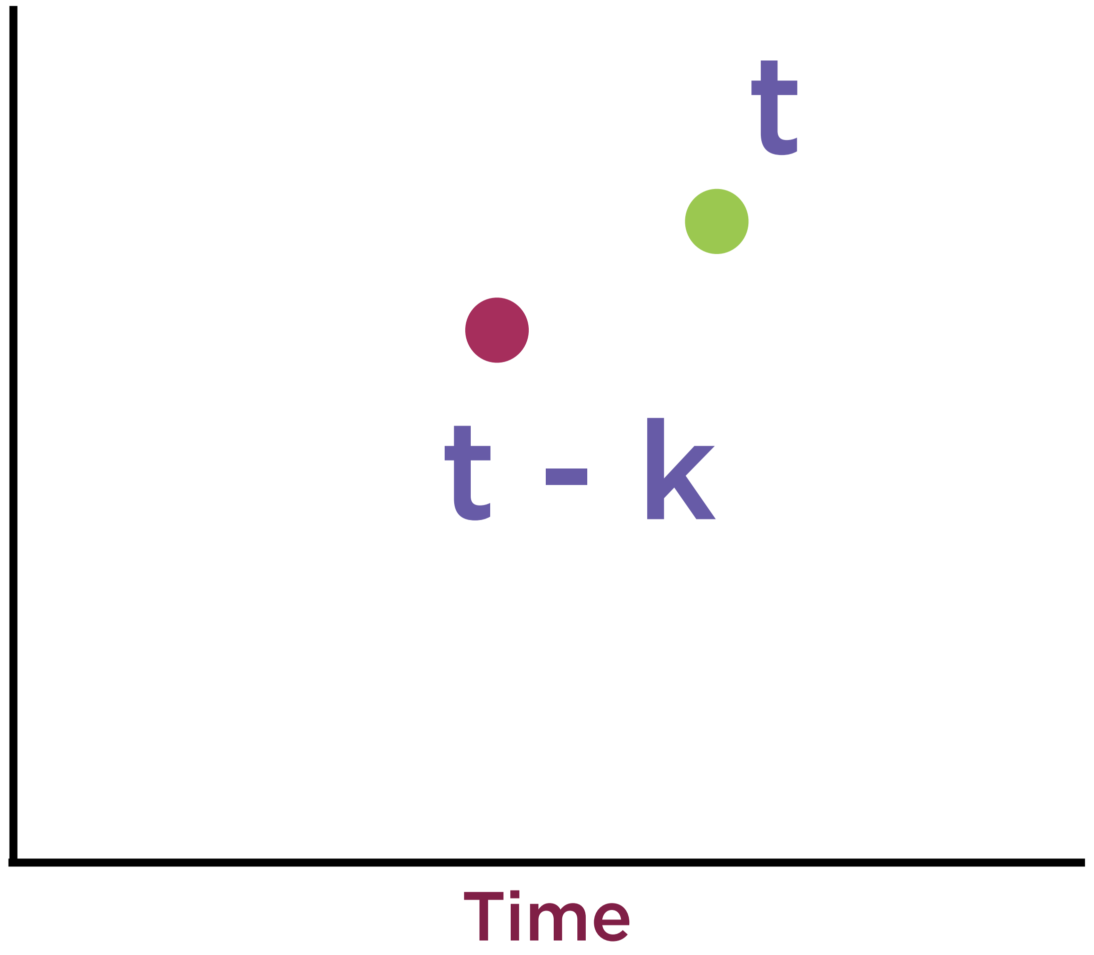
Autocorrelation

Same time series is used twice



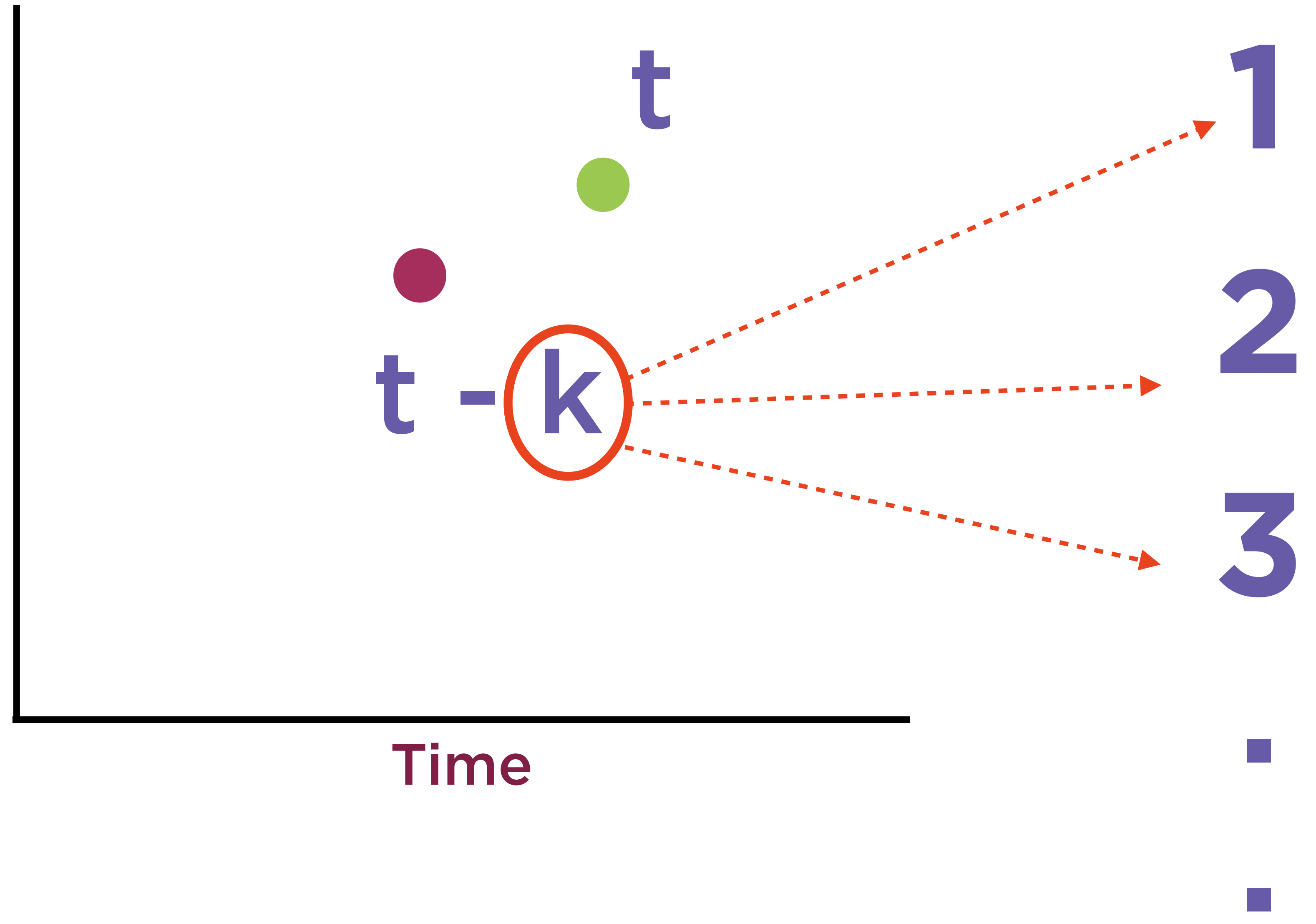
Lagged over one or more time periods

Autocorrelation



Lagged over one or more time periods

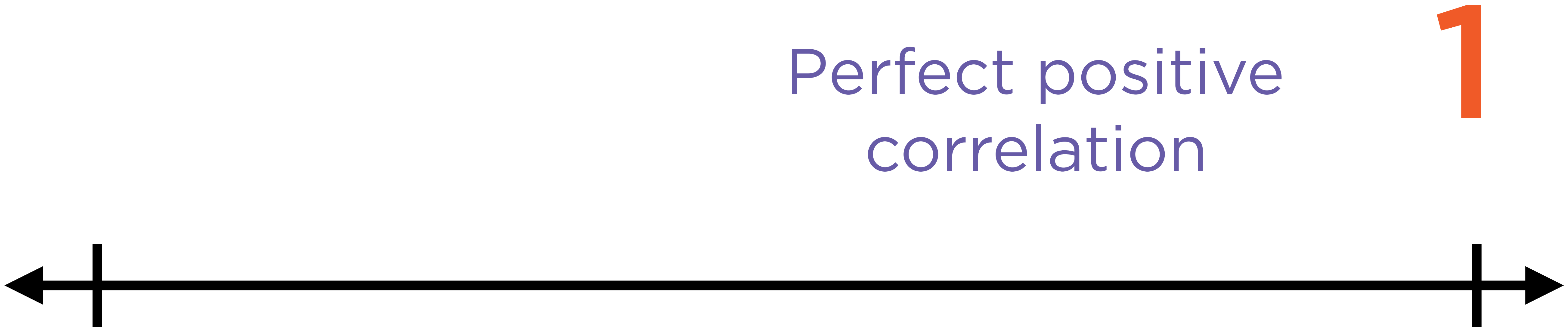
Autocorrelation



Lagged over one or more time periods

Autocorrelation

Autocorrelation

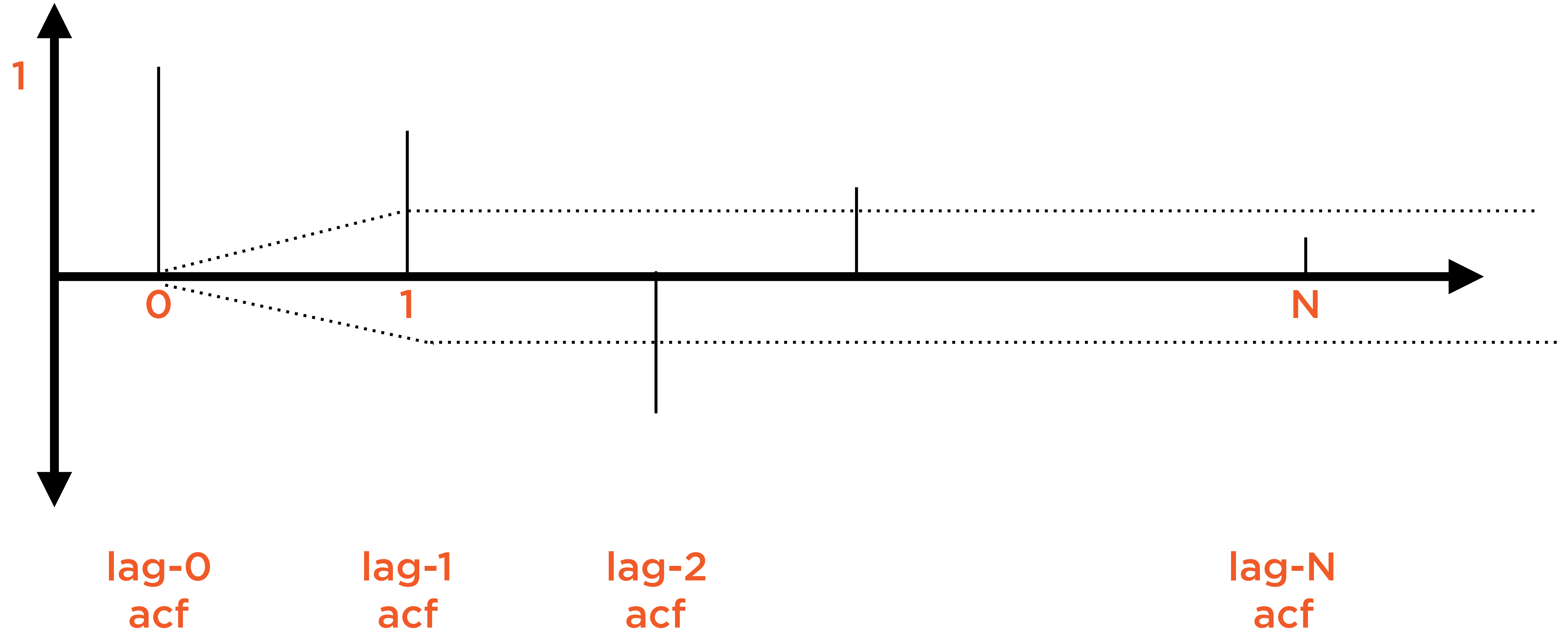


Autocorrelation

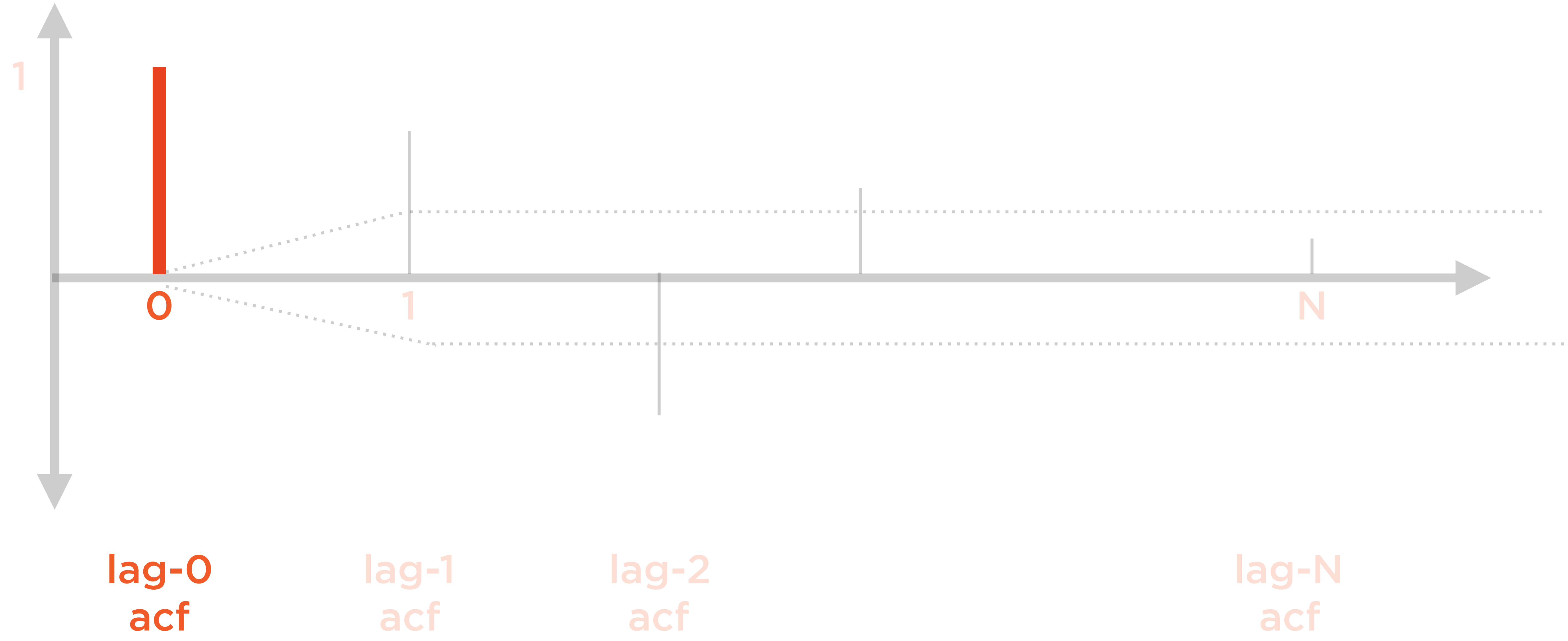
-1

Perfect negative correlation

ACF Plot

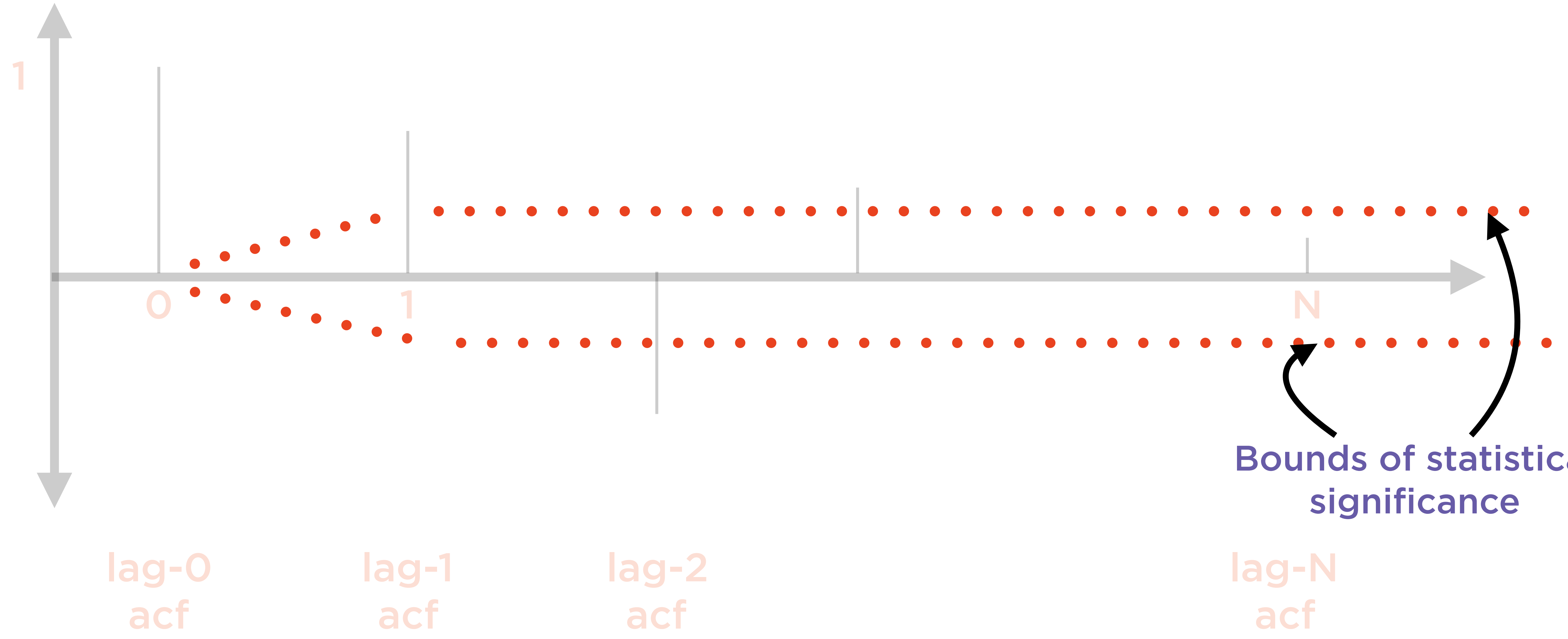


ACF Plot

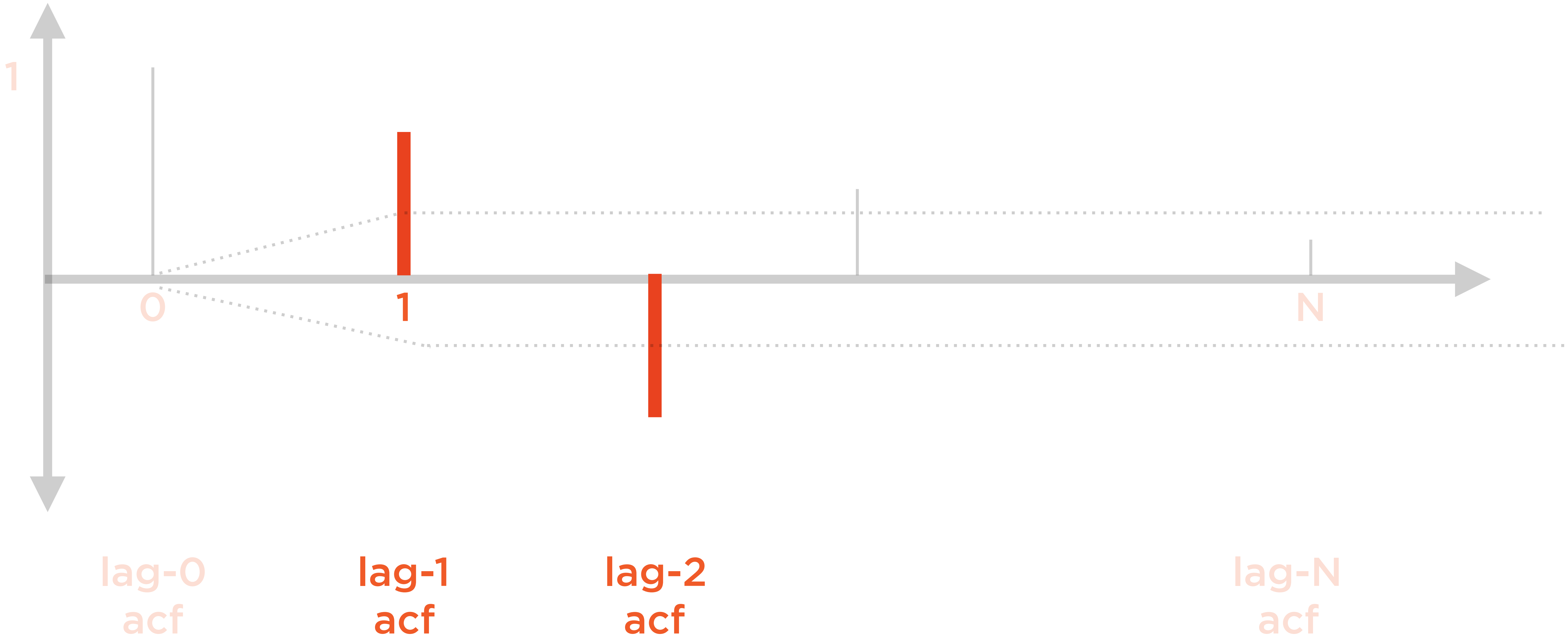


Lag-0 acf always = 1

ACF Plot

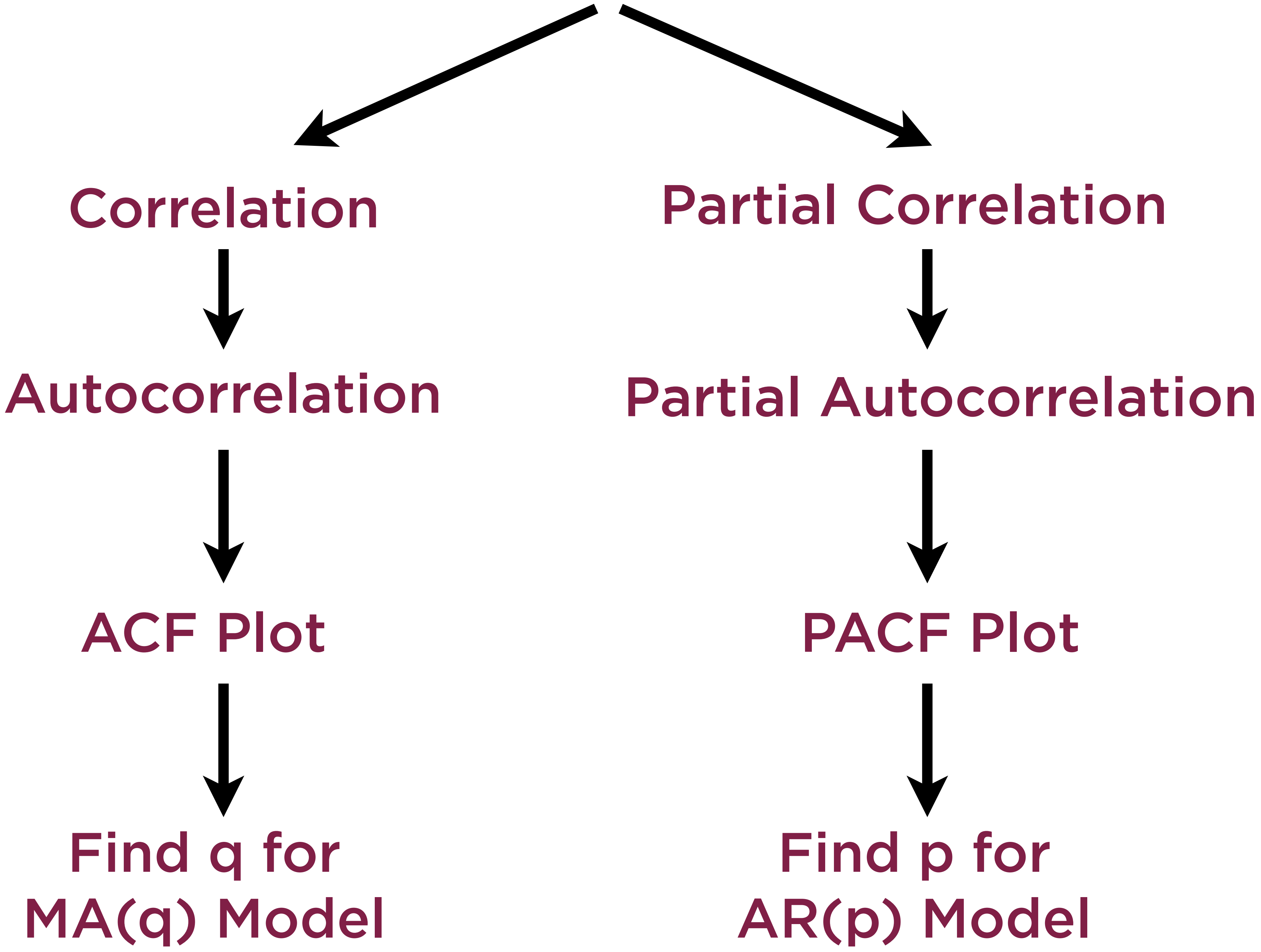

1
0
1
N
Bounds of statistical significance
lag-0 acf
lag-1 acf
lag-2 acf
lag-N acf

ACF Plot



Lag-1 and lag-2 series also correlated

Finding p,q in AR(p) and MA(q)

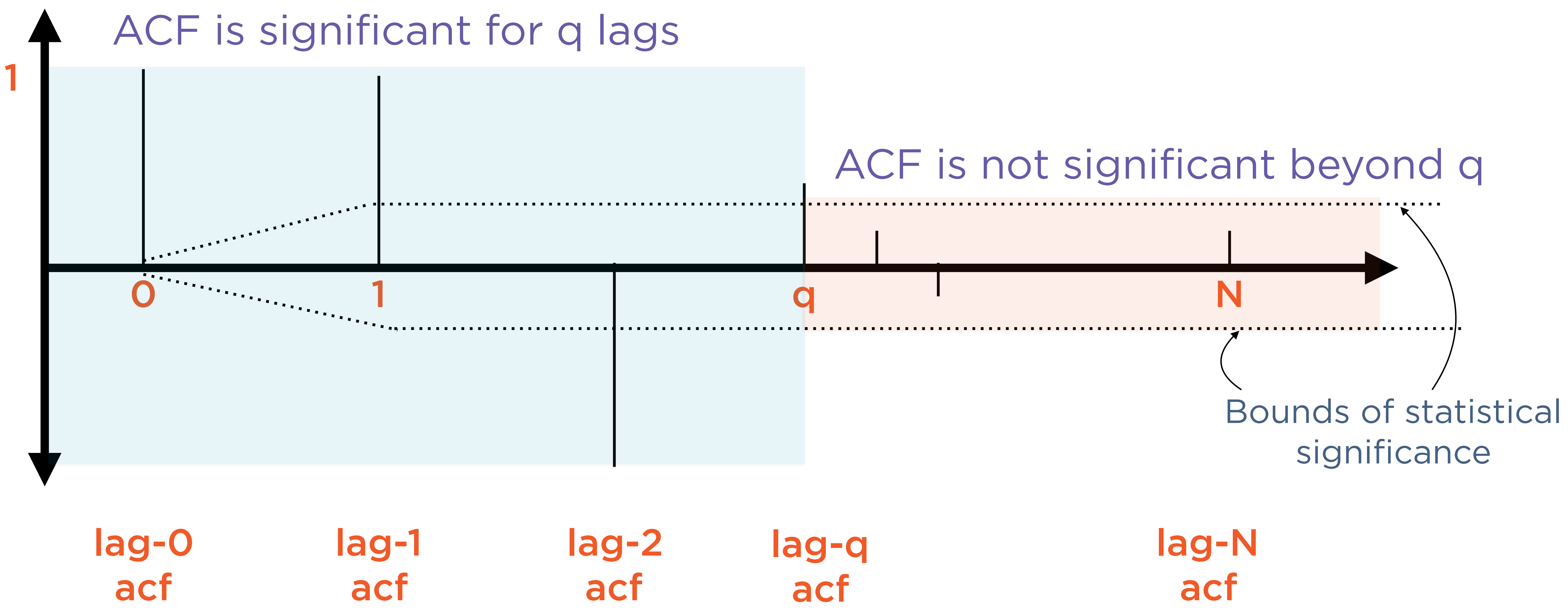


ACF and PACF Plots

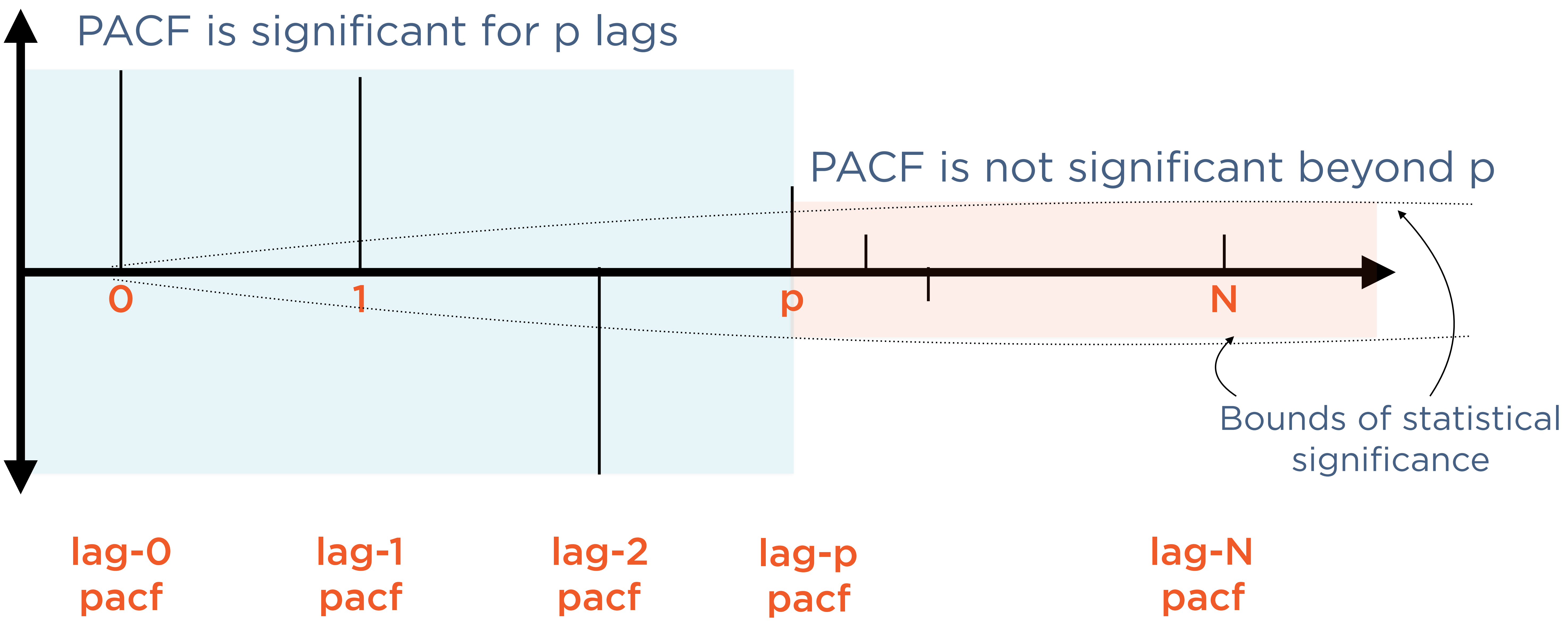
The ACF plot of an MA(q) process cuts off after q lags

The PACF plot of an AR(p) process cuts off after p lags

ACF Plot and MA(q)



PACF Plot and AR(p)



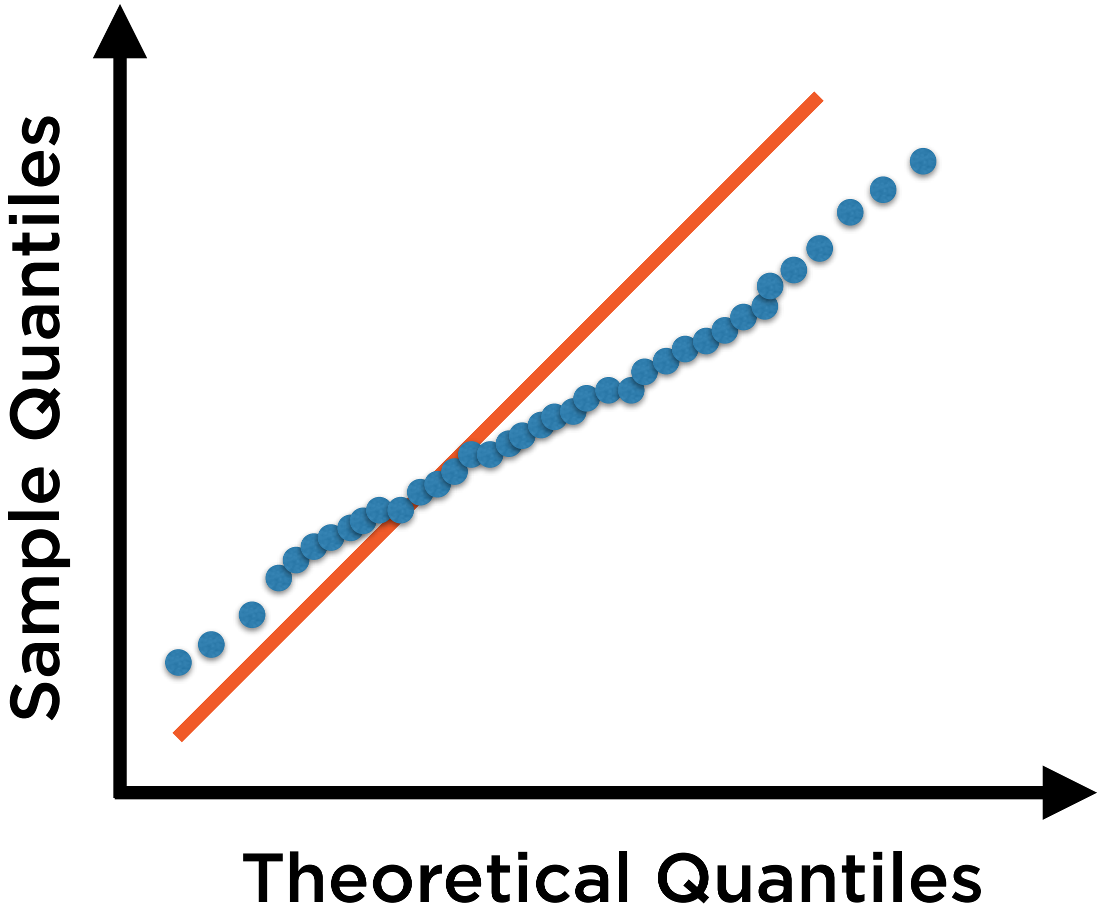
Principle of Parsimony: Keep p and q as small as possible

QQ-plot

Q-Q plot (Quantile-Quantile plot)

Graph used to compare data to a standard distribution, usually to verify visually whether data is normally distributed

Q-Q Plot

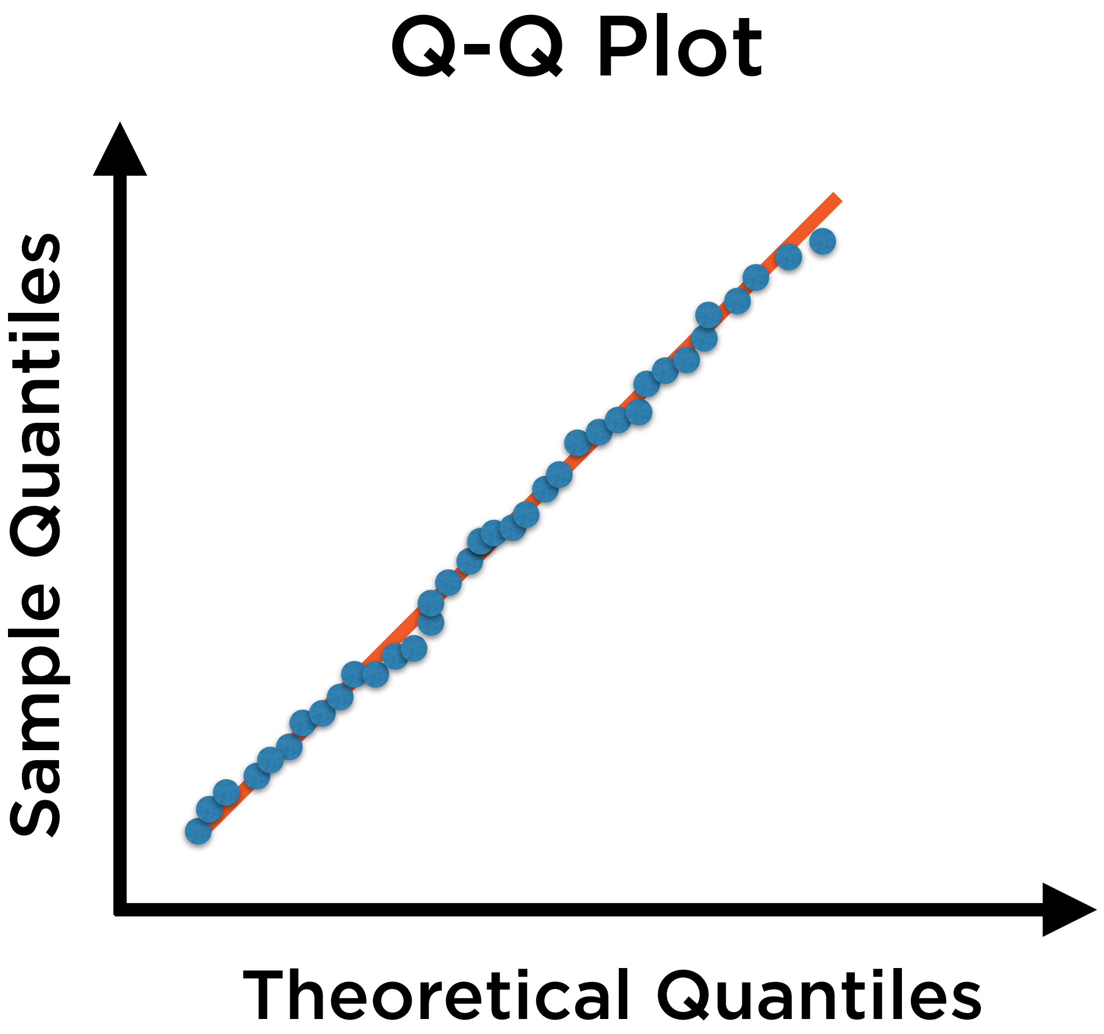


Plot is constructed by sorting data from low to high

- blue points

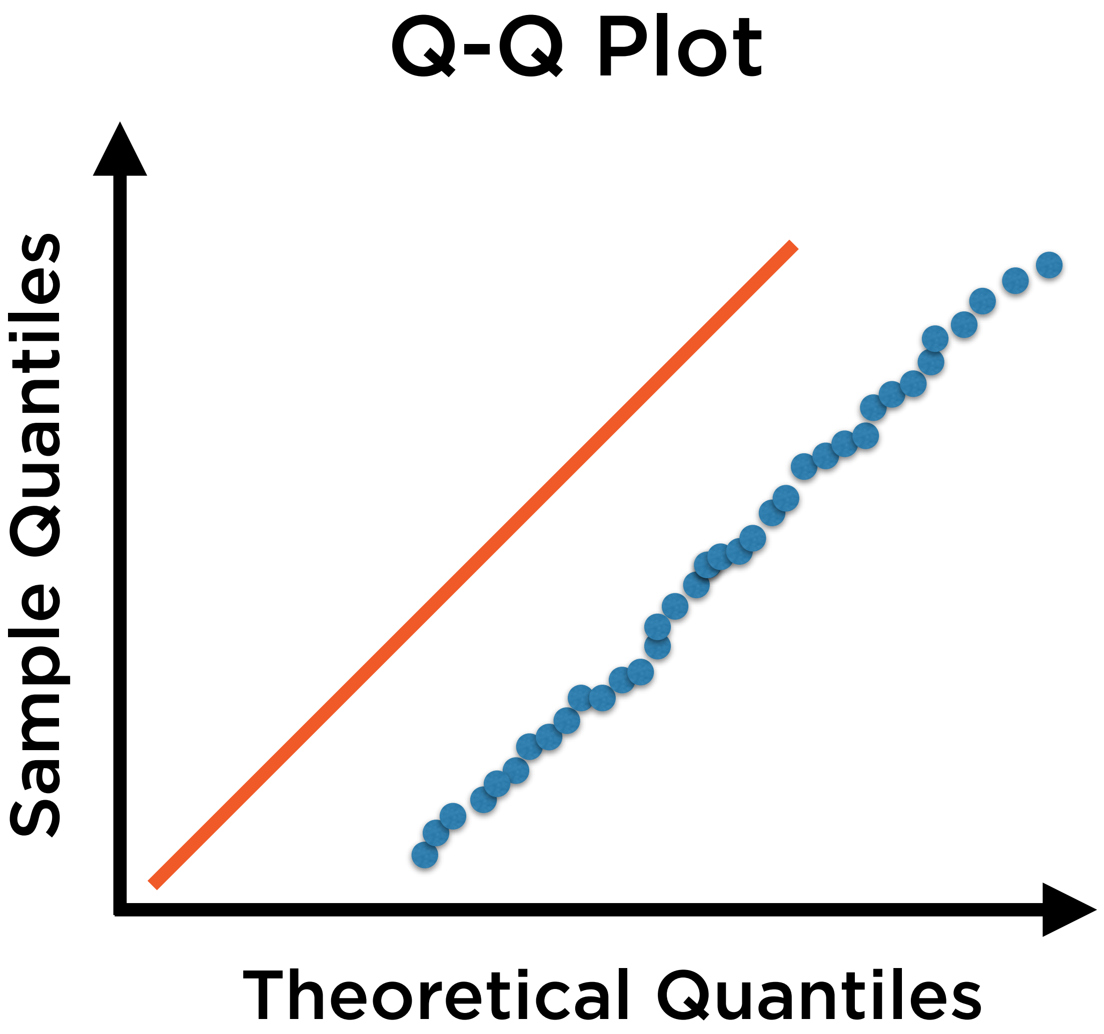
Then plotting against expected number of points in each quantile

- solid orange line
- Standard normal N(0,1)



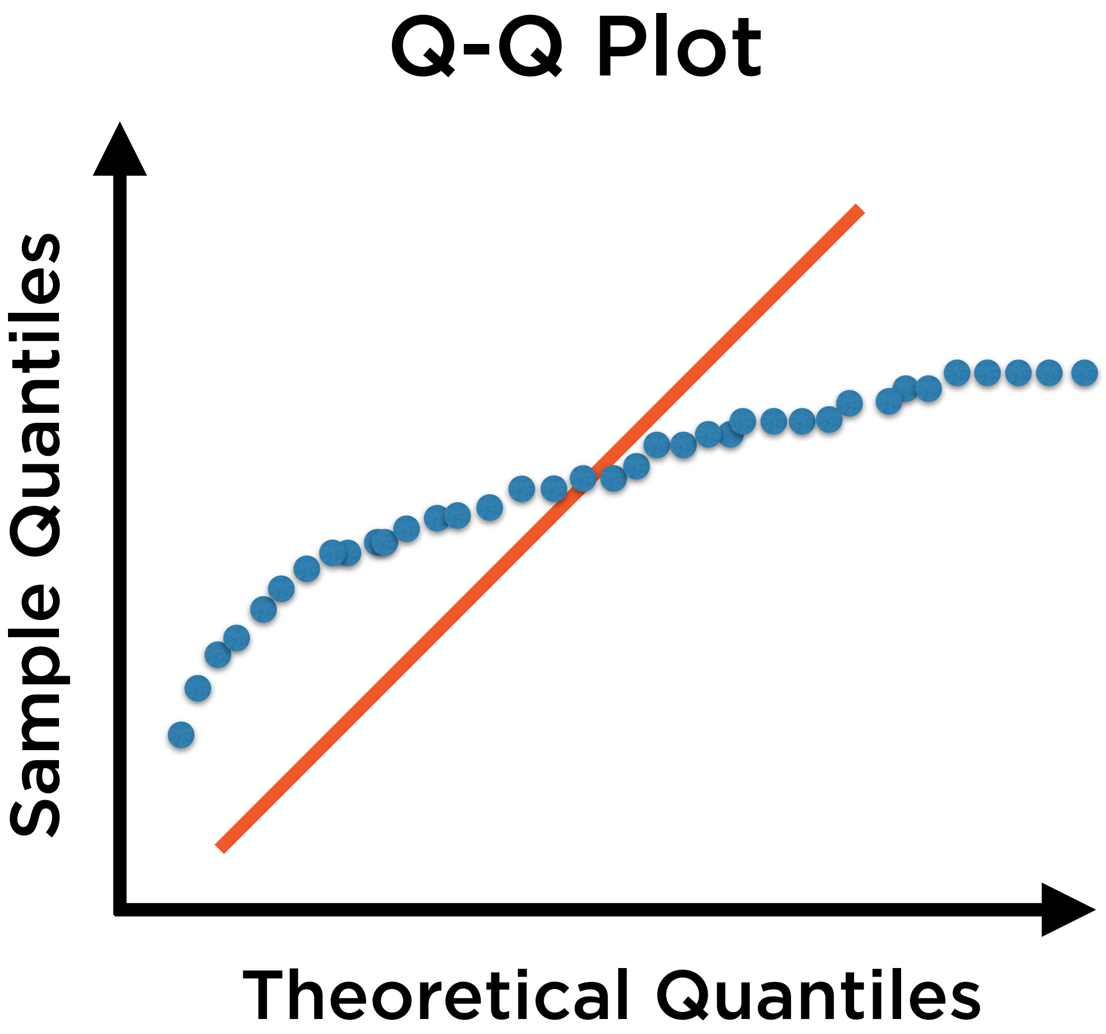
Perfectly normal data will lie entirely along line

Sample data and theoretically expected data agree perfectly



Offset from orange line indicates mean of data is not zero

Remember that orange line represents standard normal N(0,1)



Too many data points at low quantiles (small values)

Too few data points at high quantiles (large values)

Data is negatively skewed

Demo

Working with time series data

Summary

Specialized time series models

Autoregressive models

White noise error terms

Moving Average models

ARMA models combine AR and MA

Introduction to Machine Learning

Understanding Machine Learning with Python

Building Machine Learning Models in Python with scikit-learn

Python Packages for Data Science

Pandas Fundamentals

Introduction to Data Visualization in Python

Building Data Visualizations Using Plotly