Time Series Data Analysis and Forecasting

MSDS

Module 2

Topic Covered

- Measures of central tendency
- Measure of dispersion
- Types of time series data
- Numerical description
- Autocorrelation functions
- partial autocorrelation functions
- Differencing method
- Log transformation

Measures of central tendency

- Measures of central tendency are summary statistics represent the center point or typical value of a dataset.
- These measures are mean, median, and mode
- These statistics indicate where most values in a distribution fall and are also referred to as central location of a distribution.
- Choosing the best measure of central tendency depends on the type of data you have.

Measures of central tendency

Mean

- The mean is arithmetic average, to calculate mean add up all of values and divide by the number of observations in your dataset.

$$\frac{x_1 + x_2 + \cdots + x_n}{n}$$

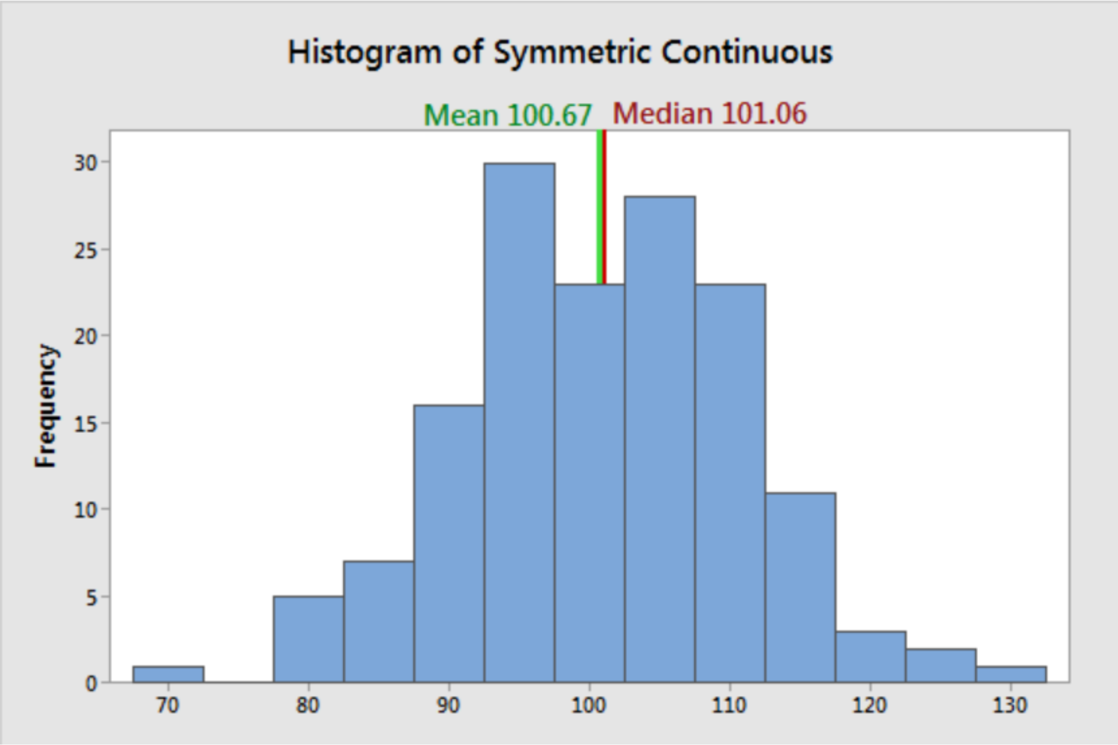
Median

- The median is the middle value. It is the value that splits the dataset in half, making it a natural measure of central tendency.
- To find the median, order your data from smallest to largest, and then find the data point that has an equal number of values above it and below it

Mode

- The mode is the value that occurs the most frequently in your data set, making it a different type of measure of central tendency than the mean or median.
- To find the mode, sort the values in your dataset by numeric values or by categories. Then identify the value that occurs most often.

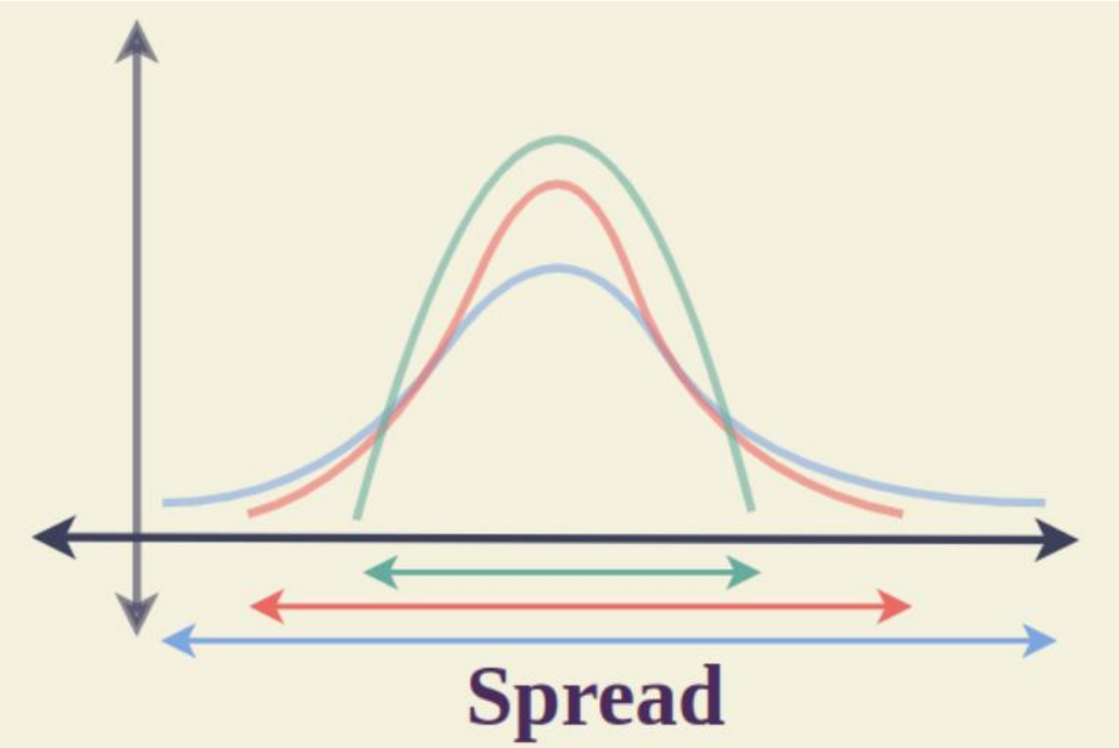
Mean Vs Median



Given data- ***10, 20, 60, 40, 25, 35*** *where* ***n = 6***
Arithmetic mean= = (10 + 20 + 60 + 40 + 25 + 35)/6= 190/6 = 31.66

Measure of dispersion

- It is used to represent the **scattering of data.**
- It show the various aspects of the data spread across various parameters
- It helps to understand if the data points are close together or far apart.



Different Measures to find dispersion

- Range
- Variance
- Standard Deviation
- Mean Deviation
- Quartile Deviation

Variance

- It measures variability of given data from the mean.
- Variance is equal to square of standard deviation.

$$Variance(\sigma^2) = \frac{(x-\bar{x})^2}{n}$$

Where x is observation data given,

$\bar{x}$ is the mean of the data

n number of observation

Example- Find the variance for the data 1, 2, 5, 4, 8, 4, here n=6

Arithmetic mean($\bar{x}$)= *(1 + 2 + 5 + 4 + 8 + 4)/6 = 24/6=4*

Variance= $(\sigma^2) = \frac{(x-\bar{x})^2}{n}$ = $[(1-4)^2 + (2-4)^2 + (5-4)^2 + (4-4)^2 + (8-4)^2 + (4-4)^2]/6 = (9 + 4 + 1 + 0 + 16 + 0)/6$

$= 30/6 = 5$

Standard Deviation

- It measures amount of variation/dispersion of a set of values.
- Dispersion tells how much data is spread out.
- A lower standard deviation indicates that **data is close to center**.
- higher value of standard deviation represents that **data spread is more**.

$$Standard\ Deviation = \sqrt{Variance(\sigma^2)}$$

Mean Deviation

- mean deviation of the data set as the value which tells us how far each data is from the centre point of the data set.
- It is the average of the deviation.

$$\text{Mean Deviation} = \frac{\Sigma_1^n |x_i - \mu|}{n}$$

where

$$\mu = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

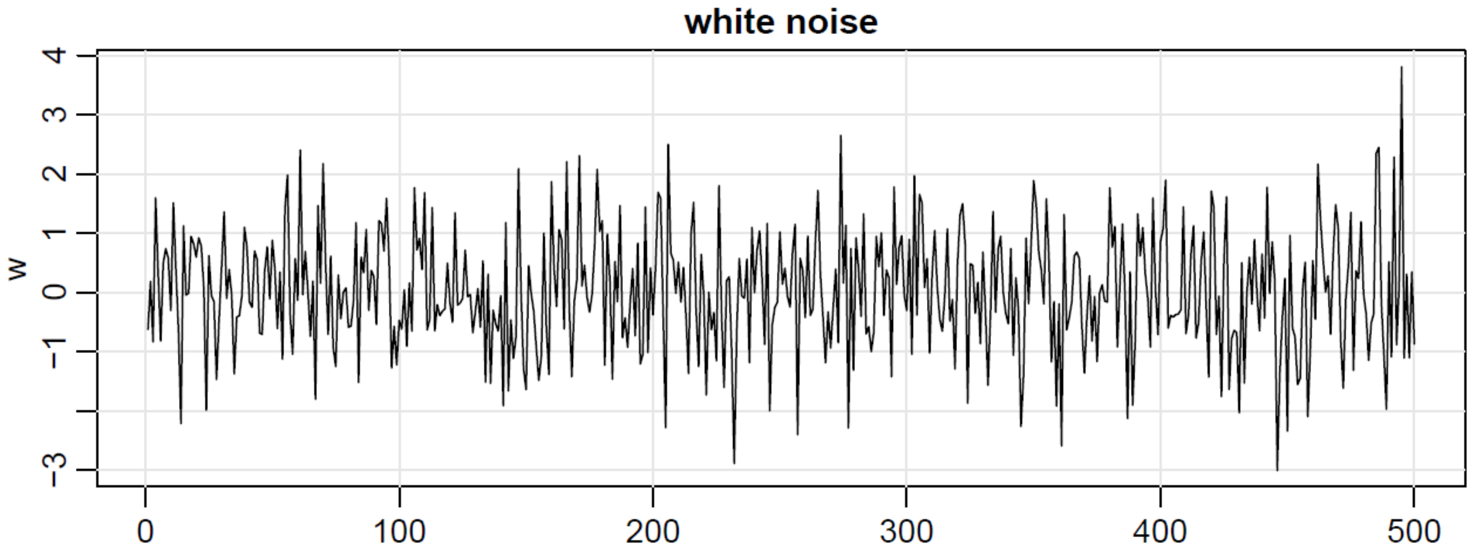
Example- Find the mean deviation of the data set, {4, 5, 6, 7, 8} about the mean of the data set. Then we first find the mean of the data set,m the central tendency.

- Mean = (4 + 5 + 6 + 7 + 8)/5 = 6

Mean Deviation = (2+1+0+1+2)/5
⇒ Mean Deviation = 1.2

White Noise in Time series

- Time series generated from uncorrelated variables is used as a model for noise in engineering applications where it is called white noise
- Noise is **independent and identically distributed (iid)** random variables with mean 0 and variance σ_w^2
- Represented by $\boldsymbol{w_t = iid(0, \sigma_w^2)}$



A collection of 500 such random variables, σ_w^2=1

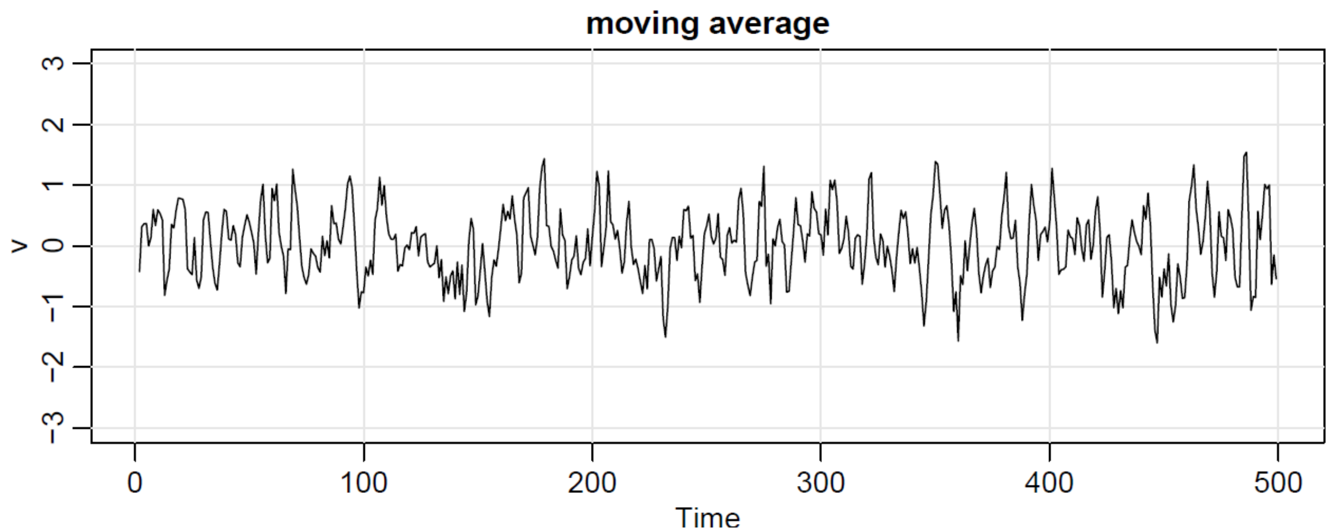
Moving averages

One can replace white noise series $\boldsymbol{w_t}$ by moving average that smooths the series.

Consider replacing w_t in previous slide example by average of its current value and its immediate neighbors in past and future.

Given by equation---

$$v_t = \frac{1}{3}(w_{t-1} + w_t + w_{t+1})$$

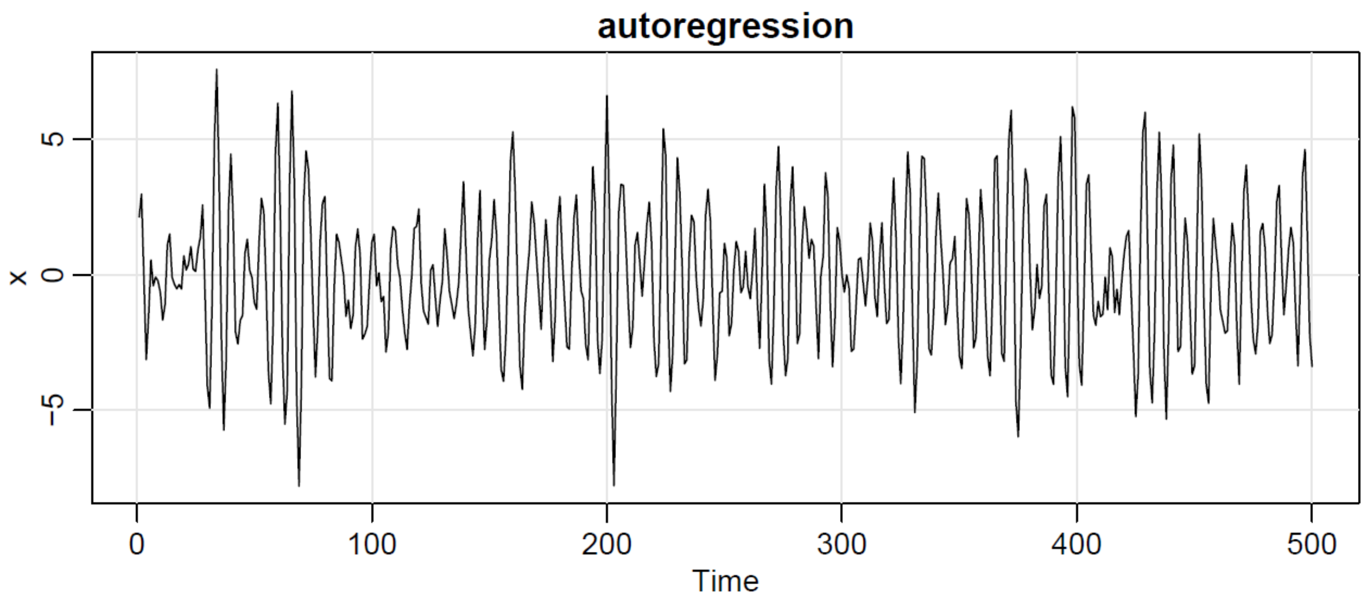


Autoregressions

- By calculating output using second order equation.

$$x_t = x_{t-1} - 0.9x_{t-2} + w_t$$

- Do it successively for t=1,2,…,500
- This equation represents a regression or prediction of the current value x_t of a time series as a function of past two values.



Types of Time Series Data

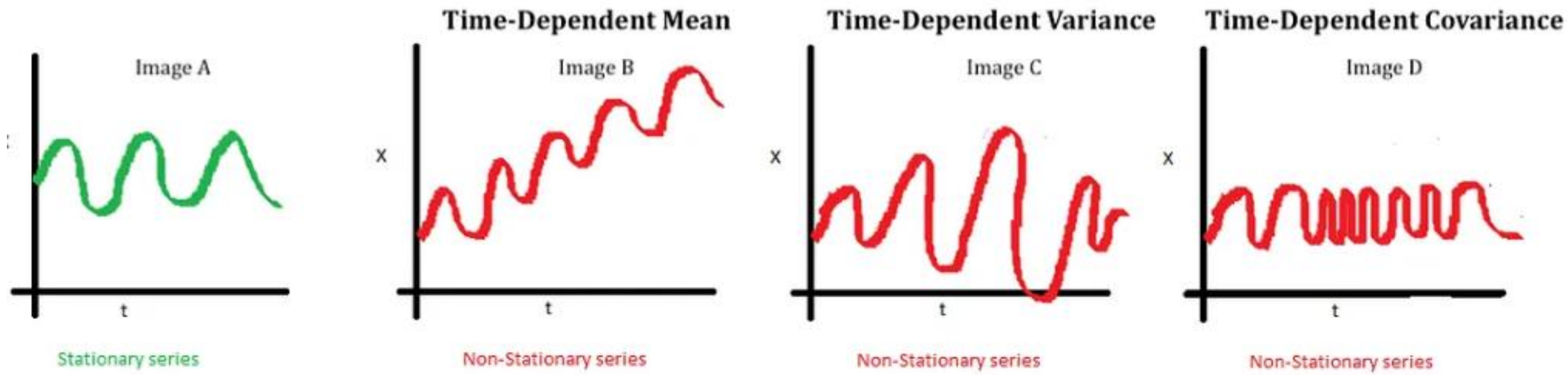
There are two major types

-Stationary

-Non-stationary

Stationary: A dataset should follow thumb rules without having Trend, Seasonality, Cyclical, and Irregularity components of time series.

- **mean** value of them should be completely constant in data.
- **variance** should be constant with respect to time-frame
- **Covariance** measures relationship between two variables.



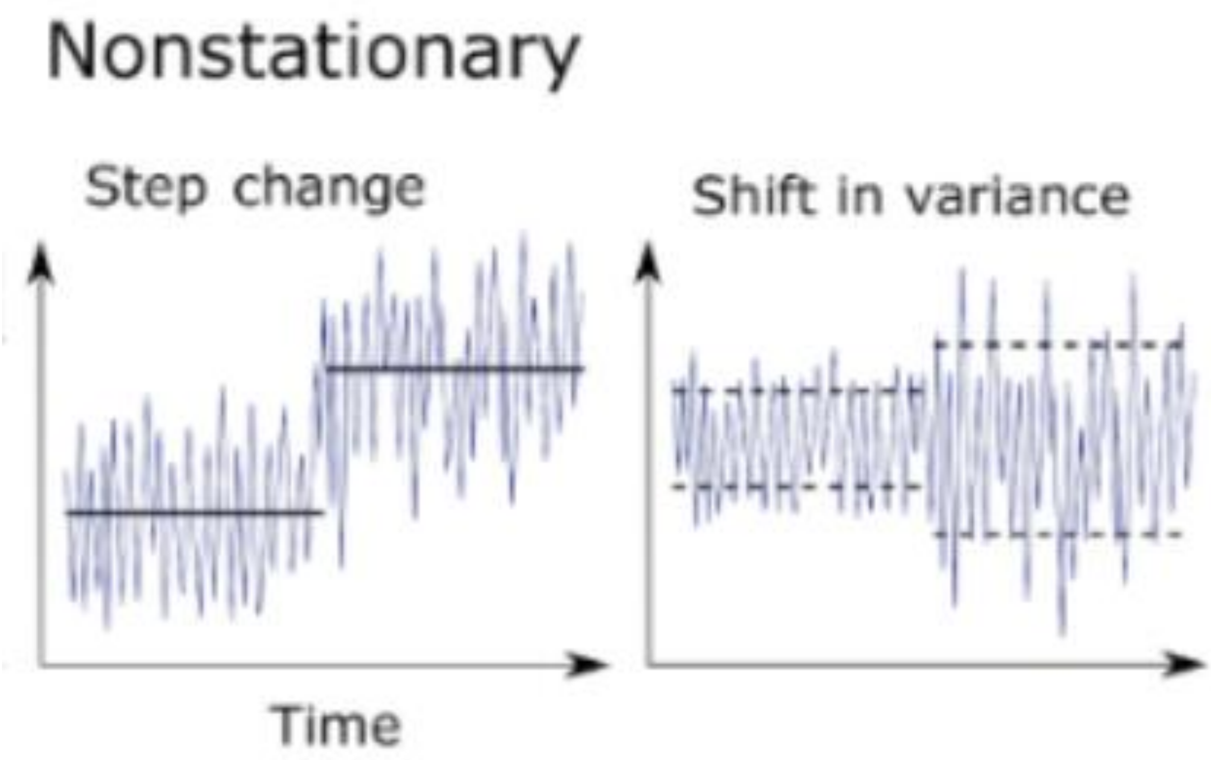
Stationary Time Series

A time series is said to be **strictly stationary** if its properties are not affected by a change in the time origin.

- If the joint probability distribution of the observations $y_t, y_{t+1}, \ldots, y_{t+n}$ is exactly the same as the joint
- probability distribution of the observations $y_{t+k}, y_{t+k+1}, \ldots, y_{t+k+n}$ then the time series is strictly stationary.
- When $n = 0$ the stationarity assumption means that the probability distribution of y_t is the same for all time periods
- Stationary implies a type of statistical **equilibrium** or **stability** in the data.

Non-stationary Time Series

- If either the **mean-variance** or **covariance** is changing with respect to time, the dataset is called non-stationary.
- A simple example of a non-stationary process is a random walk



Types of stationarity

When it comes to identifying if the data is stationary, it means identifying the fine-grained notions of stationarity in the data.

Types of stationarity observed in time series data include

- **Trend Stationary** – A time series that does not show a trend.
- **Seasonal Stationary** – A time series that does not show seasonal changes.
- **Strictly Stationary** – The joint distribution of observations is invariant to time shift.

Why checking stationarity is important?

- Non-stationary data can lead to **unreliable model** outputs and **inaccurate predictions**, just because the models aren't expecting it.
- Easier modeling and forecasting.
- Stationarity simplifies complexities within time series data, making it **easier to model** and forecast than non-stationary time series.
- When the statistical properties of a time series remain constant over time, it’s much easier to use historical data to develop accurate models of the time series and forecast future values of the series.
- By confirming stationarity, analysts can identify any potential issues in the data that might violate this essential assumption.

Testing for Stationarity

- When investigating a time series, one need to check stationary before applying various models.
- determining that time series is **constant** in mean and variance are constant and **not dependent on time**.
- Some methods to check stationarity are :-

  -by visualization

  -Autocorrelation Function (ACF)

  -**Augmented Dickey-Fuller Test (ADF)**

Autocovariance Functions

- It is defined as the second moment product for all s and t,

$$\gamma_x(s,t) = cov(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)]$$

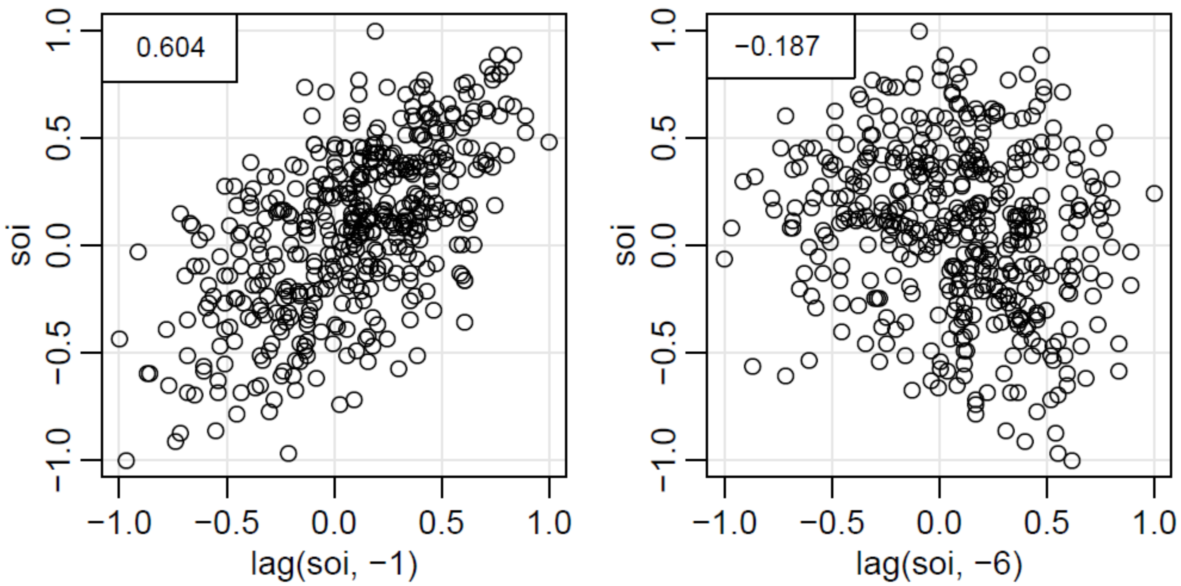
- It measures the linear dependence between two points on the same series observed at different times.
- Vary smooth series exhibit autocovariance functions that stay large even when t and s are far

Autocorrelation Functions(ACF)

ACF measures the linear predictability of series at time t, say x_t, Using only x_s.
It is defined as ,

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}.$$

$where\ \gamma(s,t) is\ autocovariance\ function$



Augmented Dickey-Fuller Test

it is based on two hypothesis:

1. The null hypothesis states that there exists a unit root in the time series and is non-stationary.
2. The alternative hypothesis states that there exists no unit root in the time series and is stationary or trend stationary.

$$y_t = c + \beta_t + \alpha Y_{t-1} + \phi \Delta Y_{t-1} + e_t$$

where,

yt= value in the time series at time t or lag of 1 time series

delta yt = first difference of the series at time (t-1)

Formula for ADF test is----

$$y_t = c + \beta_t + \alpha Y_{t-1} + \phi \Delta Y_{t-1} + \phi_2 \Delta Y_{t-2} .. + \phi_p \Delta Y_{t-p}$$

Non stationary Vs Stationary

stationary Time Series	Non-Stationary Time Series
Statistical properties of a stationary time series are **independent of the point in time** where it is observed.	Statistical properties of a non-stationary time series is **a function of time** where it is observed.
Mean, variance and other statistics of a stationary time series remains **constant**. Hence, the conclusions from the analysis of stationary series is reliable.	Mean, variance and other statistics of a non-stationary time series **changes with time**. Hence, the conclusions from the analysis of a non-stationary series might be misleading.
A stationary time series always **reverts to the long-term mean.**	A non-stationary time series **does not revert** to the long term mean.
A stationary time series **will not** have **trends, seasonality**, etc.	Presence of trends, seasonality makes a series non-stationary.

How to remove non-stationarity?

- One can fix a non-stationary time series by making it “stationary.”
- A non-stationary time series is like a toy car that doesn’t run in a straight line. Sometimes it goes fast and sometimes it goes slow, so it’s hard to predict what it will do next.

Common methods to convert non-stationary to stationary are:-

- **By differencing**
- **By seasonal differencing**
- **By log transformation**

Differencing

- It is a way to make a non-stationary time series stationary
- compute differences between consecutive observations. This is known as **differencing**.
- Differencing can help stabilise the mean of a time series by removing changes in the level of a time series, and therefore eliminating (or reducing) trend and seasonality.
- There can do it upto two-levels, first order difference and second order difference .

$$y'_t = y_t - y_{t-1}$$

Seasonal Differencing

- A seasonal difference is the difference between an observation and the previous observation from the same season.
- Where m=the number of seasons.
- These are called “lag-m” differences, because we substract observation after a lag of m period.

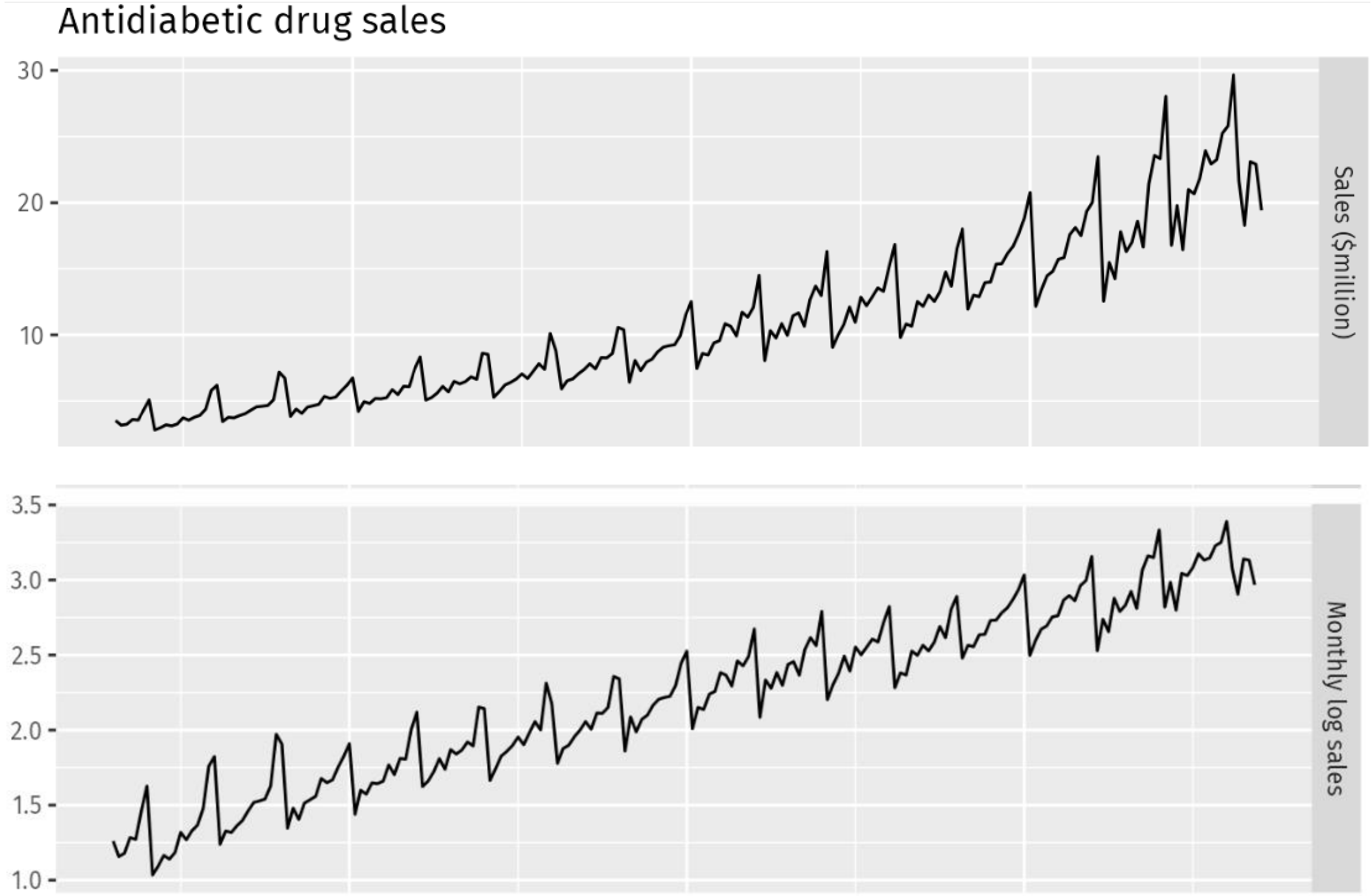
$$y'_t = y_t - y_{t-m}$$

Log transformation

- Log transformation can be used to stabilize the variance of a series with non-constant variance. This is done using log() fuction.
- One limitation of log transformation is that it can be applied only to positively valued time series.
- Taking a log shrinks the values towards 0.
- For values that are close to 1, the shrinking is less and for the values that are higher, the shrinking is more, thus reducing the variance

```
cbind("Sales ($million)" = a10,
      "Monthly log sales" = log(a10),
      "Annual change in log sales" = diff(log(a10),12)) %>%
  autoplot(facets=TRUE) +
    xlab("Year") + ylab("") +
    ggtitle("Antidiabetic drug sales")
```

Log transformation



Review Question

- What is time series data?
- Differentiate between seasonality and trends.
- Discuss two methods to covert non-stationary time series to stationary time-series.