Análise e Previsão de Séries Temporais

Aula – 1: Introdução às séries temporais

Eraylson Galdino

egs@cin.ufpe.br

Agenda

- Séries Temporais:
  - Definições
  - Exemplos
- Modelos simples com média zero:
  - Ruído I.I.D
  - Processo Binário
  - Random Walk
- Modelos com tendência e sazonalidade:
  - Tendência
  - Sazonalidade
- Modelos Estacionários:
  - Estacionariedade
  - Função de Autocorrelação
  - Correlograma

Séries Temporais: Definição

George Box & Gwilym Jenkis

*Uma série temporal é um conjunto de **observações ordenadas**, x_t, cada uma observada em um instante de tempo.*

Peter J Brockwell

*Uma série temporal é um conjunto de observações, x_t, cada uma **registrada num tempo específico t.***

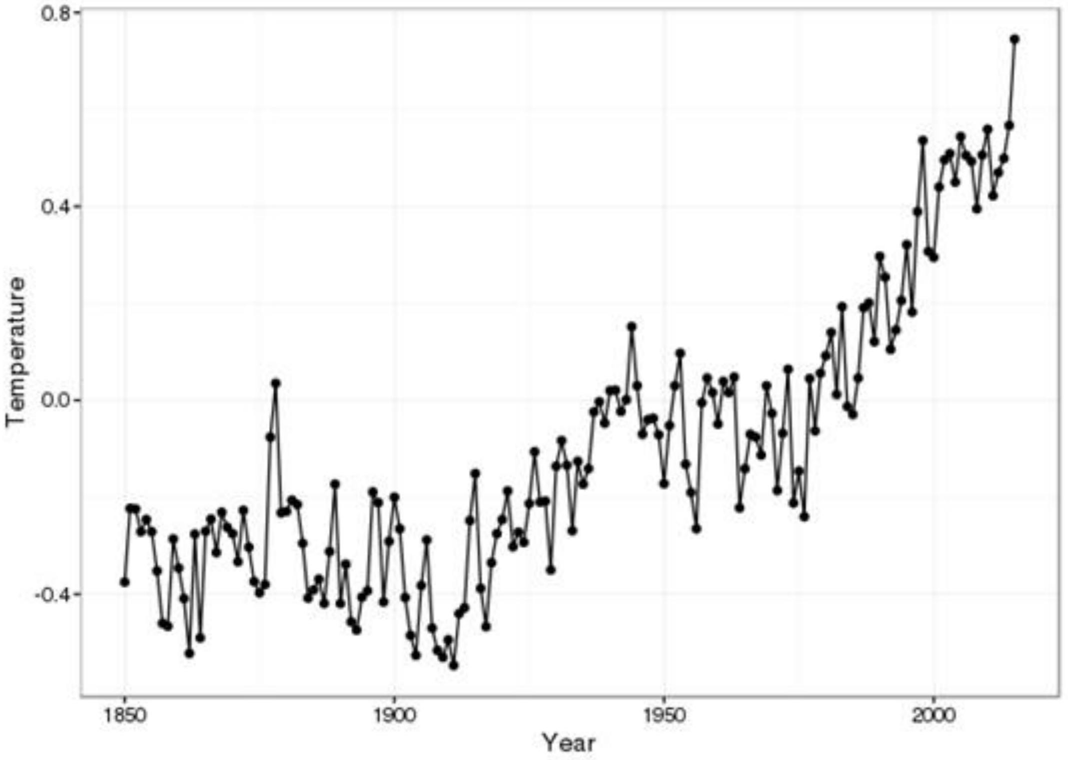
Eraylson G. Silva

*Uma série temporal é um conjunto de observações ordenadas, x_t, cada uma registrada num tempo específico **t**, resultante de um comportamento que pode ser representado por um **modelo matemático.***

Séries Temporais: Definição

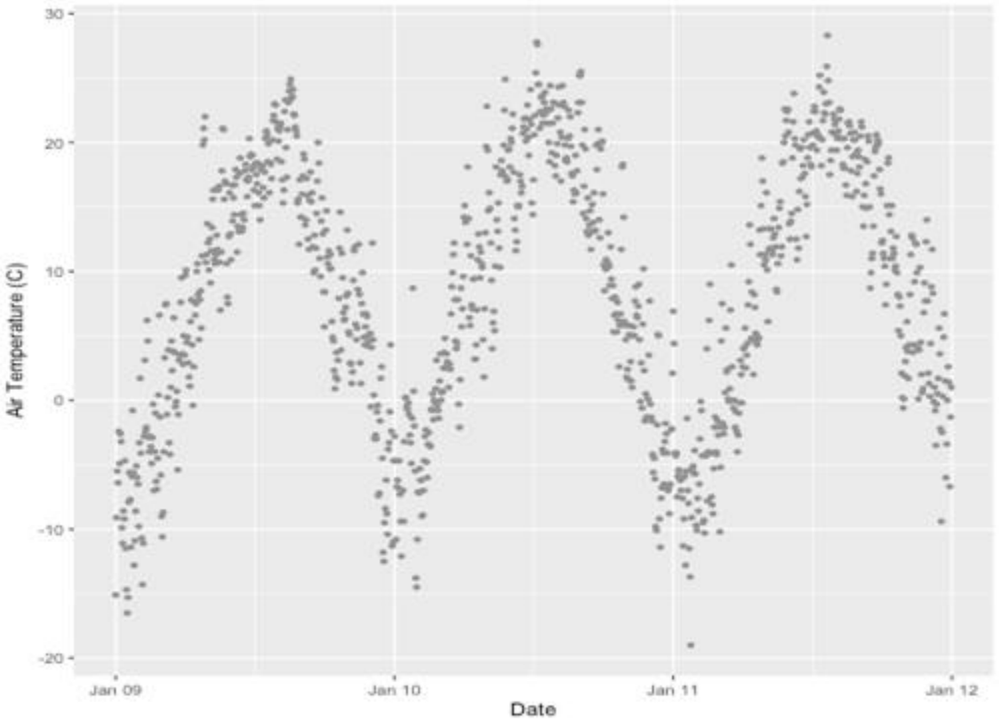
- **Séries temporais discretas**: são séries em que o **intervalo de observações** (**t**) pertence a um **conjunto discreto**. Ou seja, as observações são feitas em **intervalos de tempo fixos**;
- **Séries temporais contínuas:** são séries em que as **observações são obtidas continuamente** através de algum **intervalo no tempo**, por exemplo, quando T = [0, 1];

Séries Temporais: Definição



Exemplo de uma série temporal **discreta**

Séries Temporais: Definição



Exemplo de uma série temporal **contínua**

Análise de Séries Temporais

- Objetivos:
  - Compreender as **características** de tal **fenômeno** temporal;
  - **Selecionar** e **estimar** um **modelo estocástico** que possivelmente tenha **gerado** o conjunto de dados;
- Aplicações:
  - Entendimento de um comportamento temporal;
  - Classificar um comportamento temporal;
  - Detecção de anomalias;
  - Previsão;

Análise de Séries Temporais

- Campos de Aplicação:
  - Mercado Financeiro;
  - Meteorologia;
  - Medicina;
  - Controle de Qualidade e Processos;
  - Governo;
  - Epidemiologia;
  - Transporte;

Análise de Séries Temporais

- No geral, Análise de Séries Temporais consiste em:
  - Analisar um conjunto de dados (ao longo do tempo);
  - **Selecionar** **e** **estimar** um **modelo** matemático que possivelmente tenha **gerado** o conjunto de dados;

Modelos de Séries Temporais

Todos os modelos estão errados,
mas alguns são úteis.

Modelos Simples de Séries Temporais

- Selecionar um **modelo probabilístico** adequado para os dados é uma das partes mais **importantes** da **análise** de séries temporais.
- Geralmente, é **suposto** que cada **observação** **x** seja um valor **resultante** de uma determinada **variável aleatória** **X**;
- Um **modelo** de série temporal para um determinado **dado observado** $\boldsymbol{x_t}$ é a **especificação** da **composição de distribuições** da sequência da variável aleatória $\boldsymbol{X_t}$ em que $\boldsymbol{x_t}$ é uma observação;

Modelos Simples de Séries Temporais

- Modelos com média zero:
  - Ruído i.i.d. (independente e identicamente distribuído);
  - Processo binário;
  - *Random Walk* (Passeio Aleatório);
- Modelos com tendência e sazonalidade

Modelos Simples com média zero: Ruído i.i.d

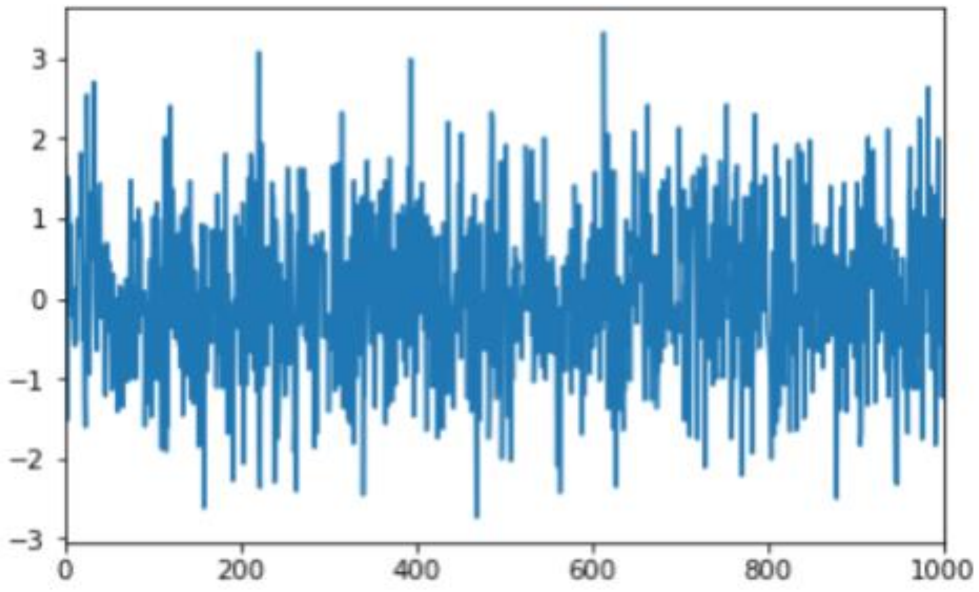
- Modelo mais simples;
- Sem os componentes de **tendência** e **sazonalidade**;
- As **observações** são resultados de **variáveis aleatórias** i.i.d. com média zero;
- Uma sequência de variáveis aleatórias $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, ..., são ruídos i.i.d;

$$P[X_1 \leq x_1, \ldots, X_n \leq x_n] = P[X_1 \leq x_1] \cdots P[X_n \leq x_n] = F(x_1) \cdots F(x_n)$$

- F(.) é uma função de distribuição cumulativa;
- **Não existe dependência** entre as observações;
- Através do valor $\boldsymbol{x_n}$ **não é possível prever** o valor de $\boldsymbol{x_{n+h}}$;
- Apesar de **não ser interessante para predição**, é um modelo **importante** para **construção** e **entendimento** de modelos mais complexos;

Modelos Simples com média zero: Ruído i.i.d

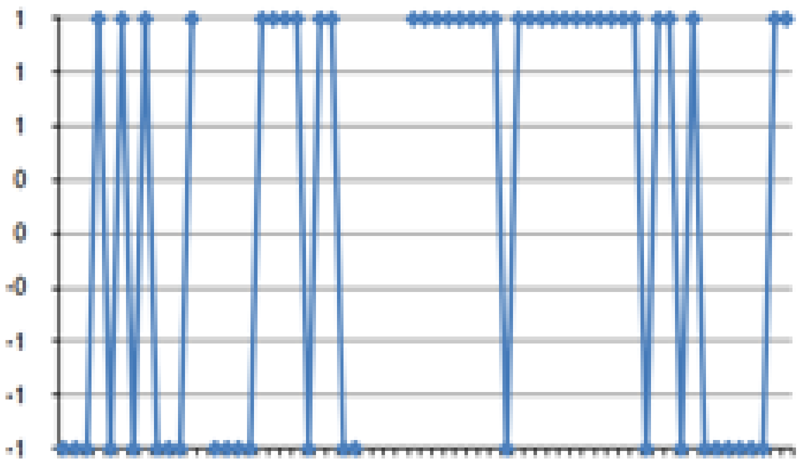
- Um ruído i.i.d com distribuição normal com média zero e variância σ^2 é também chamado de **ruído branco** gaussiano;



Modelos Simples com média zero: Processo binário

- As observações só podem assumir dois valores possíveis;

$$P[X_t = 1] = p, \quad P[X_t = -1] = 1 - p,$$



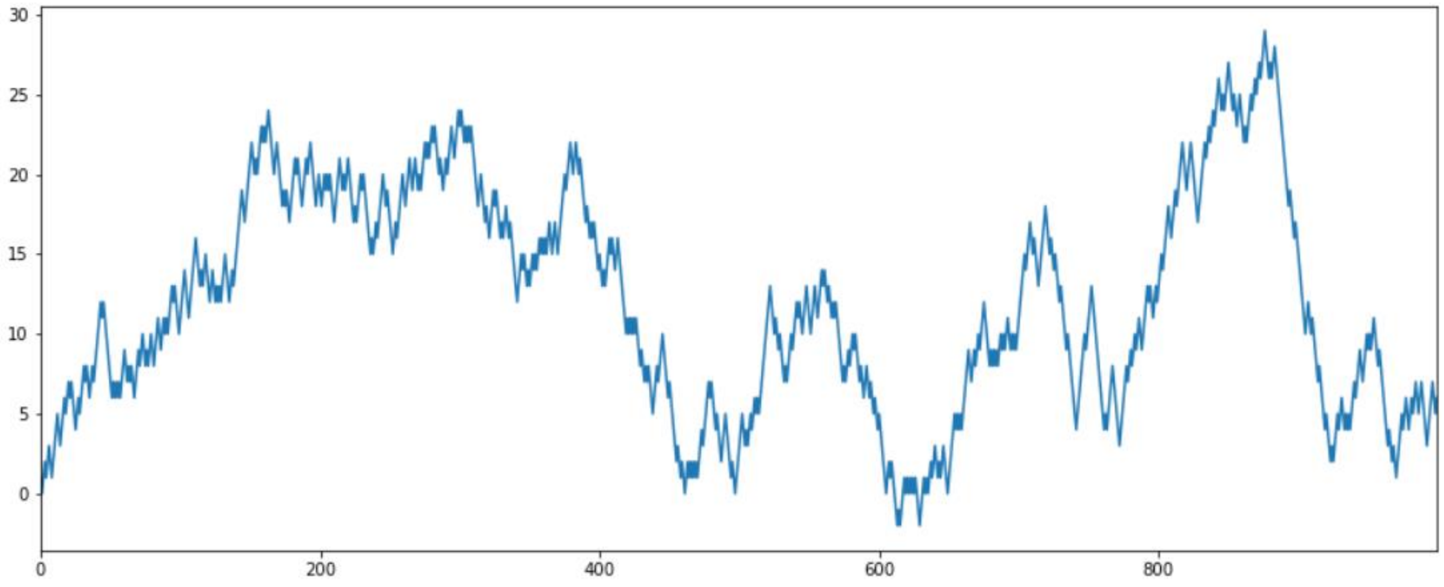
Modelos Simples com média zero: *Random Walk*

- Ideia remete a "**caminhada do bêbado**". Tomada de vários **passos consecutivos**, cada qual em uma **direção aleatória**;
- É **obtido** através da **soma cumulativa** de variáveis i.i.d aleatórias:

$$S_t = X_1 + X_2 + \cdots + X_t, \quad \text{for } t = 1, 2, \ldots,$$

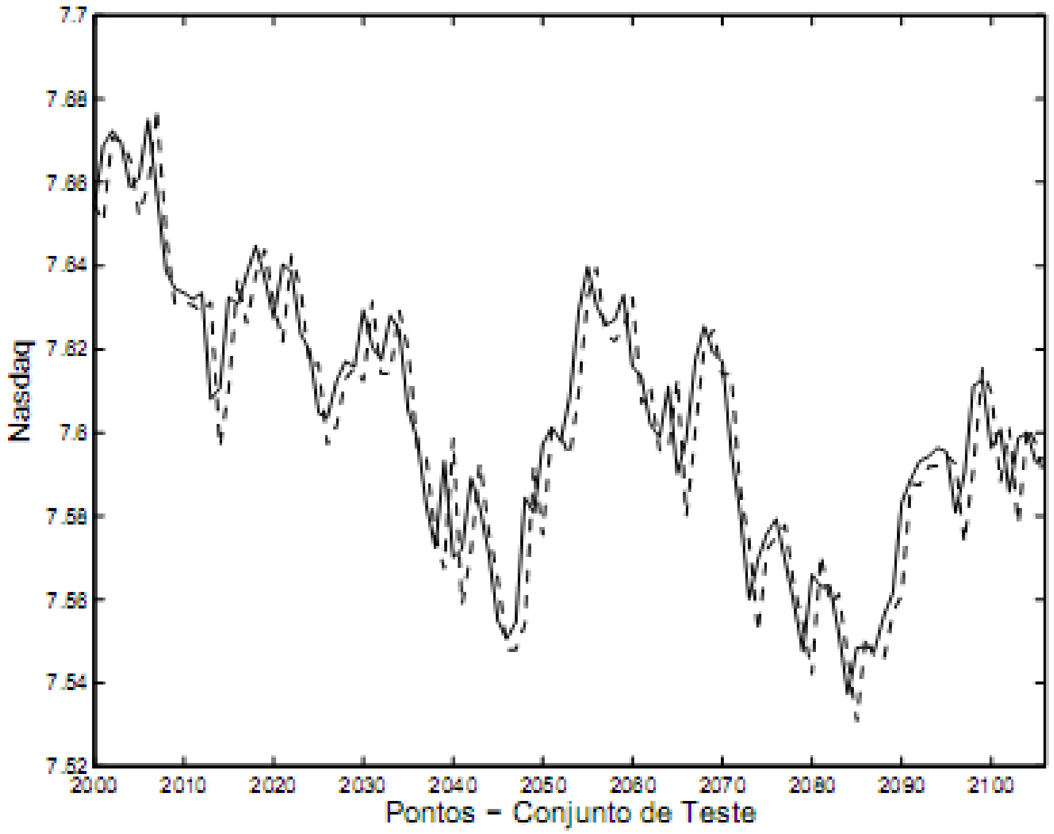
- é um ruído i.i.d;
- Se $\boldsymbol{X_t}$ é resultado de um processo binário, então é chamado de *Random Walk* simétrico simples;
  - Caso de *Random Walk* aplicado em jogo de baseball

Modelos Simples com média zero: *Random Walk*



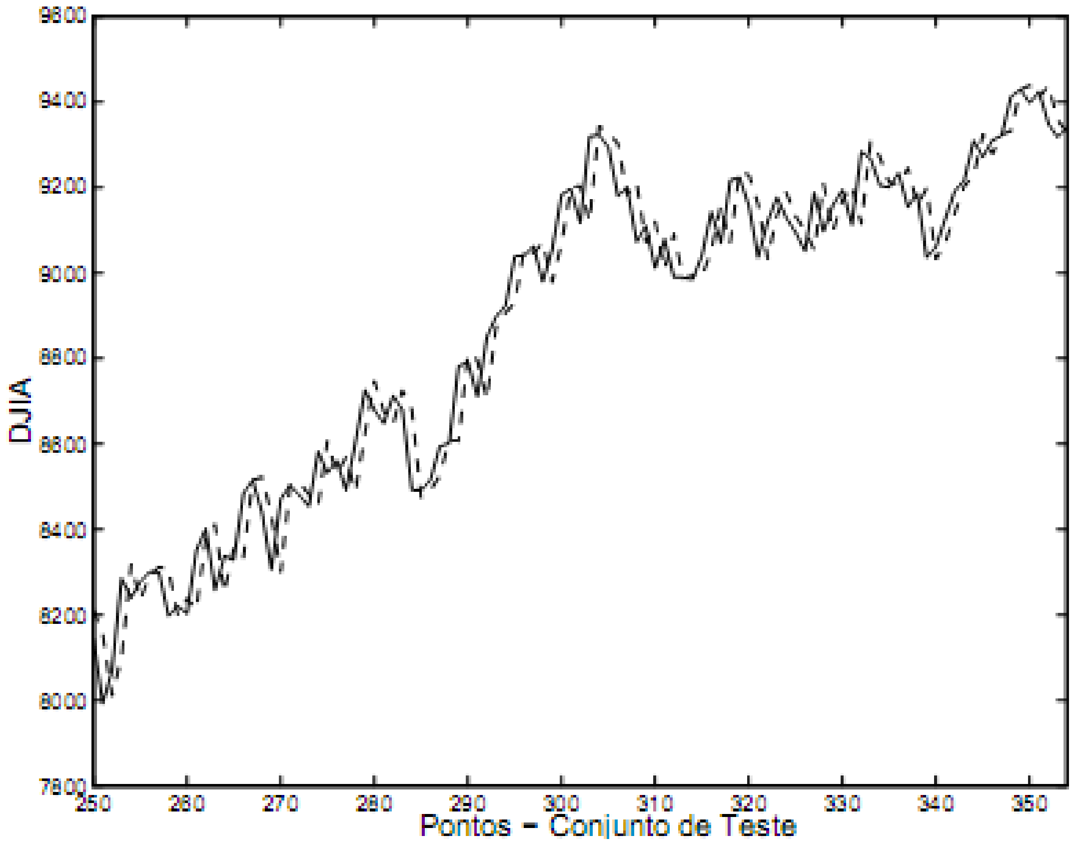
Modelos Simples com média zero: *Random Walk*

- Problema na previsão de um *Random Walk: Série Nasdaq*



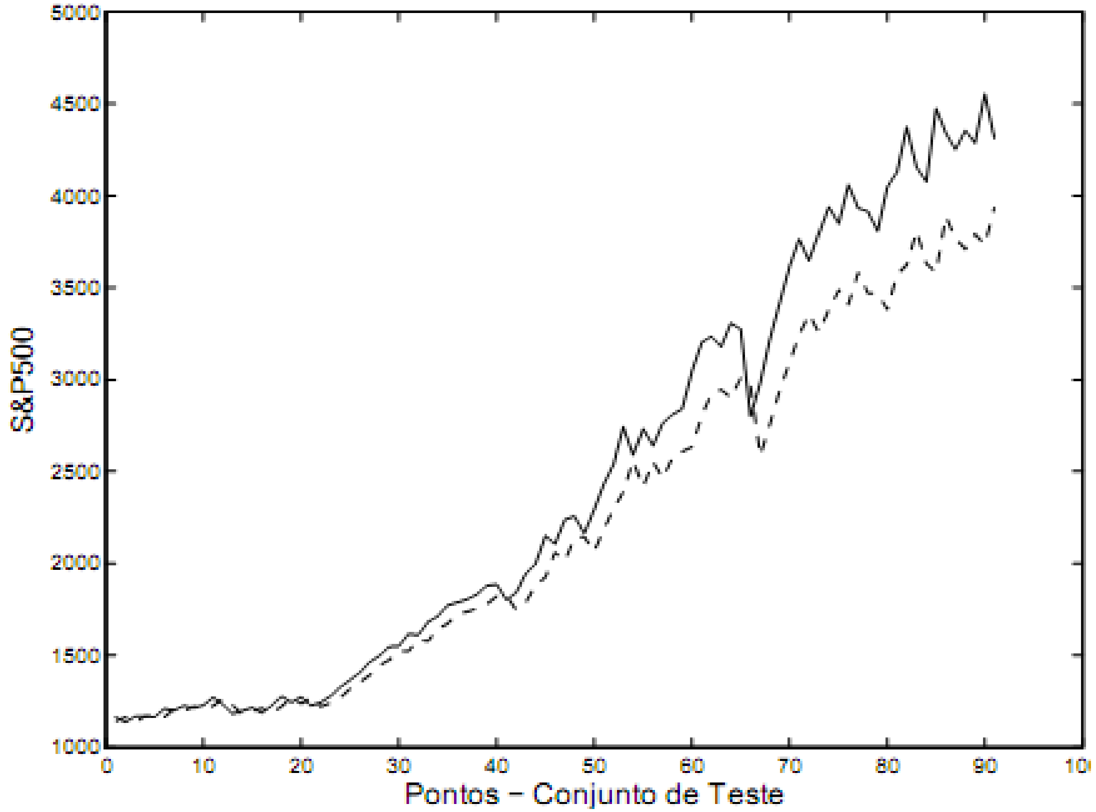
Modelos Simples com média zero: *Random Walk*

- Problema na previsão de um *Random Walk: Série DJIA*



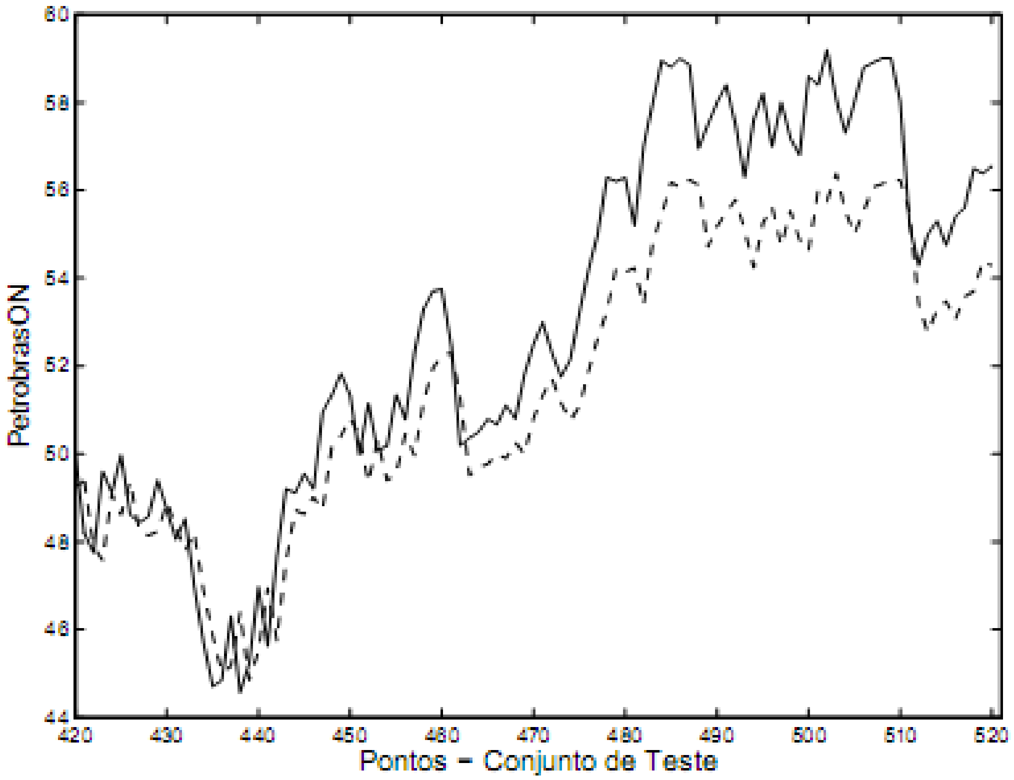
Modelos Simples com média zero: *Random Walk*

- Problema na previsão de um *Random Walk: Série S&P 500*



Modelos Simples com média zero: *Random Walk*

- Problema na previsão de um *Random Walk: Série Petrobras*



Modelos Simples com média zero: *Random Walk*

- Problema na previsão de um *Random Walk: Série S&P 500*

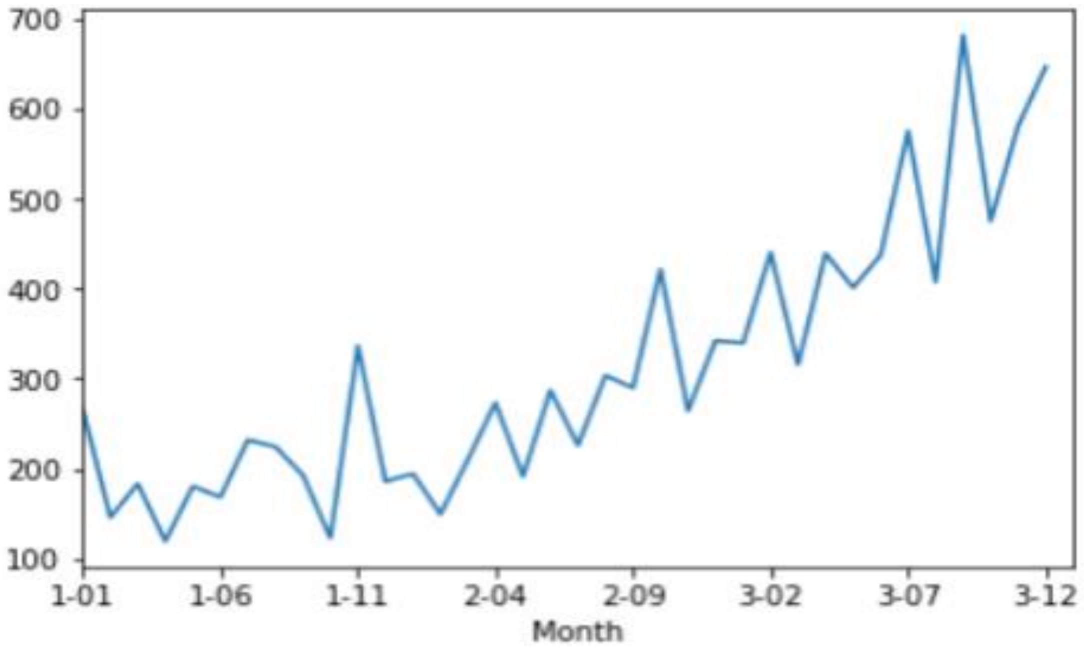
Real	Previsão
0,685264	0,69427
0,690659	0,688232
0,681649	0,69365
0,685226	0,684601
0,691443	0,688193
0,700727	0,694438
0,707789	0,703763

Modelos com tendência e sazonalidade

- Algumas séries é perceptível que **não podem ser modeladas** por **modelos simples** de média zero;
- São **geradas** com **componentes** de **tendência** e **sazonalidade**;
- **Tendência**: mudança sistemática na série temporal que não aparenta ser periódico;
- **Sazonalidade**: comportamento que se repete durante um período de tempo;

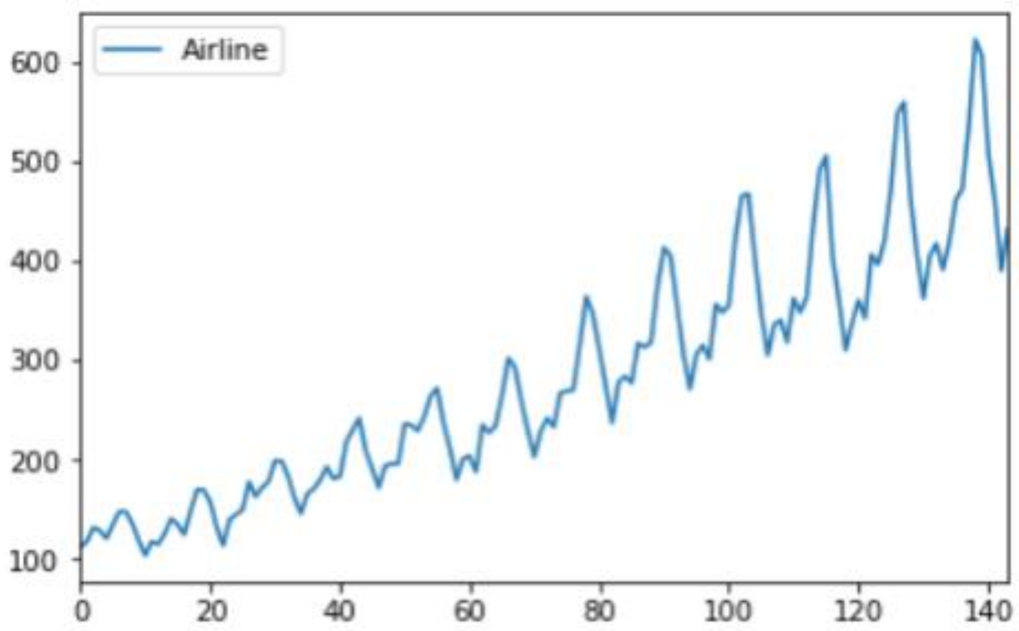
Modelos com tendência e sazonalidade

- Série com tendência: Vendas de Shampoo



Modelos com tendência e sazonalidade

- Série com sazonalidade: Vendas de Passagens aéreas



Modelos com tendência e sazonalidade

- Modelo com tendência:

$$X_t = m_t + Y_t.$$

- $\boldsymbol{m_t}$ é o **componente de tendência** e pode ser **estimado** através do método de **mínimo quadrado**:

$$m_t = a_0 + a_1 t + a_2 t^2.$$

- Valores de **a** são ajustados para minimizar a função:

$$\sum_{t=1}^{n} (x_t - m_t)^2$$

Modelos com tendência e sazonalidade

- Modelo com Sazonalidade:

$$X_t = s_t + Y_t$$

- s_t é o **componente de sazonalidade**, resultado de uma **função periódica** que pode ser obtida através da **soma de ondas senoidais**:

$$s_t = \alpha_0 + \sum_{j=1}^{k}(a_j cos(\lambda_j t) + b_j sin(\lambda_j t))$$

- **a** e **b** são parâmetros desconhecidos e lambda λ são as frequências

Uma Abordagem Geral para Modelagem de Séries Temporais

- A abordagem consiste em **plotar a série** e analisar alguns aspectos:
  - Tendência
  - Sazonalidade
  - Alterações acentuadas no comportamento
  - Observações discrepantes com os dados
- **Remover tendências** e componentes **sazonais** para obter resíduos estacionários;
- **Escolher** um modelo para ajustar aos resíduos. Utilizando várias estatísticas amostrais como função de autocorrelação
- Alcançar **a previsão original** da série através da **previsão dos resíduos** junto com os valores estimados da **tendência** e **sazonalidade**;

Processo Estacionário

- É um processo que se **mantém em equilíbrio estatístico** com propriedades probabilísticas que **não se alteram no tempo**;
  - $\boldsymbol{X_t}$ uma série temporal com E($\boldsymbol{x_t}^2$) < ∞
  - A Média pode ser representada por: $\mu_x(t) = E(X_t)$
  - A covariância pode ser representada por: $\gamma_x(r,s) = Cov(X_r, X_s) = E[(X_r - \mu_x(r))(X_s - \mu_x(s))]$
- O processo é considerado **fracamente estacionário** se:
  - As características de $\boldsymbol{x_t}$ são iguais para todo $\boldsymbol{t}$;
  - Média e variância constantes;
  - $\mu_x(t)$ é independente de $\boldsymbol{t}$ e $\gamma_x(t+h,t)$ é independente de $\boldsymbol{t}$ para cada $\boldsymbol{h}$;

Modelos Estacionários: Função de Autocorrelação

- Seja X_t uma série estacionária, a Função de Autocovariância do lag **h** é:

$$\gamma_x = Cov(X_{t+h}, X_t)$$

- A Função de Autocorrelação de X_t do lag **h** é definida como:

$$\rho_x(h) \equiv \frac{\gamma_x(h)}{\gamma_x(0)} = Cor(X_{t+h}, X_t)$$

- A autocorrelação é a correlação entre uma série e ela mesma defasada;

Analisando Estacionariedade: Exemplos

- Ruído i.i.d.:
  - Se $\boldsymbol{X_t}$ é um ruído i.i.d e E($\boldsymbol{X_t}^2$) = σ² < ∞ e E($\boldsymbol{X_t}$) = 0 para todo $\boldsymbol{t}$, então a primeira condição para ser um modelo estacionário é satisfeita.
  - Assumindo a independência dos dados temos:

$$\gamma_X(t+h, t) = \begin{cases} \sigma^2, & \text{if } h = 0, \\ 0, & \text{if } h \neq 0, \end{cases}$$

Não dependente de $\boldsymbol{t}$

  - {$\boldsymbol{X_t}$}~IID(0, σ²) é um modelo estacionário.

Analisando Estacionariedade: Exemplos

- Ruído Branco:
  - Se X_t é uma sequência de **variáveis aleatórias descorrelacionadas**, onde X_t ~WN(0, σ^2) então X_t é **estacionário com características semelhantes ao ruído i.i.d** (covariância);
  - Todo processo IID(0, σ^2) é um WN(0, σ^2), porém o inverso não é verdadeiro;
- Random Walk:
  - Se S_t é um random walk definida por:
    - E(S_t) = 0, E(S_t^2) = tσ^2 < ∞ para todo t, e para h ≥ 0;
    - Será estacionária se $\gamma_s(t+h,t)$ não depender de **t**. Pois caso contrario à medida que **t** aumenta, a variância cresce indefinidamente, violando uma das condições de estacionariedade;

Função de Autocorrelação Amostral

- Utilizada para analisar o **grau de dependência** temporal na série;
- **Auxilia na seleção** de possíveis modelos de séries temporais estacionárias;
- Definição:
  - Seja $\boldsymbol{x_t}, \ldots, \boldsymbol{x_n}$ observações de uma série temporal. A média amostral é definida como:

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$$

  - A função de autocovariância amostral é definida como:

$$\hat{\gamma}(h) = n^{-1}\sum_{t=1}^{n-|h|}(x_{t+|h|} - \bar{x})(x_t - \bar{x}), \quad -n < h < n$$

  - A função de autocorrelação amostral:

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \quad -n < h < n.$$

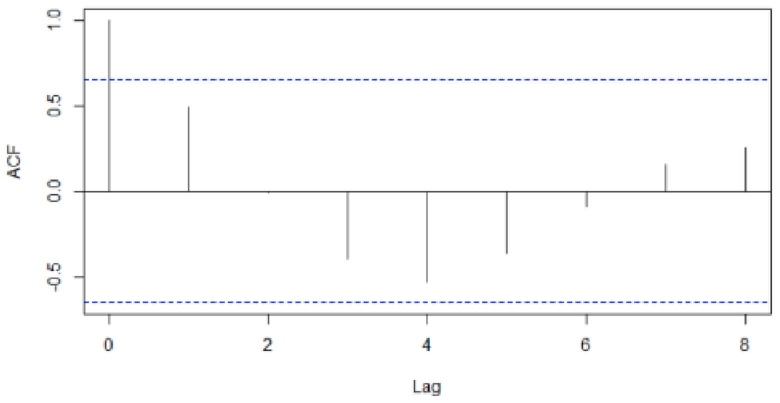
Função de Autocorrelação Parcial

- As autocorrelações para intervalos sucessivos são formalmente dependentes;
- A F.A.C parcial é utilizada para **obter uma informação** sobre autocorrelação na série **sem esta influência em cascata**;
- A autocorrelação parcial de atraso k corresponde a autocorrelação entre $\boldsymbol{x_t}$ e $\boldsymbol{x_{t-k}}$ que não é explicada pelos atrasos de 1 a k;

- Correlação x Causalidade:
  - Correlação: Uma variável **A** tem um comportamento semelhante à **B**;
  - Causalidade: Uma variável **B** influencia o comportamento da variável **B;**
  - http://www.tylervigen.com/spurious-correlations

Correlograma

- Forma gráfica de analisar a autocorrelação;
- O correlograma traça as autocorrelações em **diversas defasagens**;
- Através da análise utilizando o correlograma é possível **entender se a série é aleatória** ou possui alguma **tendência** ou **sazonalidade**;
- Frequentemente utilizado para **analisar os resíduos** de um modelo;



Correlograma

- O Gráfico é **composto por valores** exibidos em pontos/barra que representam o **coeficiente de correlação** amostral **r**. Esse **coeficiente** é uma medida de direção e **grau** entre **duas variáveis** quantitativas que se **associam linearmente**;
- Representando as variáveis por (x,y), o coeficiente é calculado através da seguinte fórmula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \text{ onde } \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ e } \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

- O coeficiente **r** para o par de variáveis (x,y) é o quociente entre a **covariância amostral** das variáveis x e y e o **produto dos respectivos desvios padrões**:

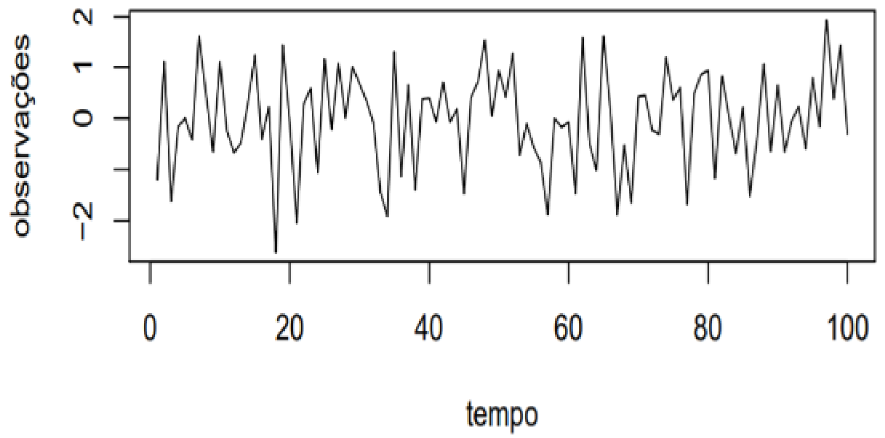
$$r = \frac{Cov(x,y)}{s_x s_y}$$

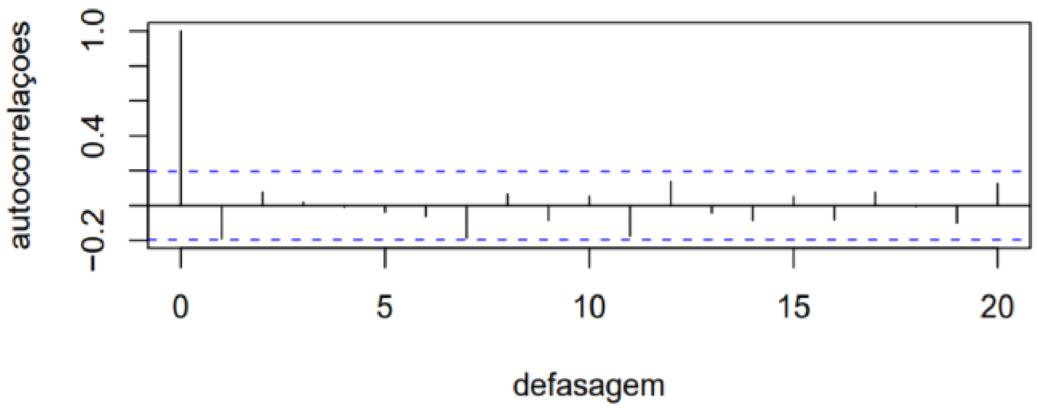
Correlograma

- Para inferir se em um dado lag existe ou não correlação é utilizado intervalos de confiança;
- Para um intervalo de confiança de 95% os limites são $\pm 1.96/\sqrt{n}$
- Os lag com coeficiente **r** fora do intervalo de confiança são considerados significantes.
- O coeficiente com valor 1 representa correlação máxima positiva;
- O coeficiente com valor -1 representa correlação máxima negativa;
- O coeficiente com valor 0 representa que não existe correlação;

Correlograma: Exemplo

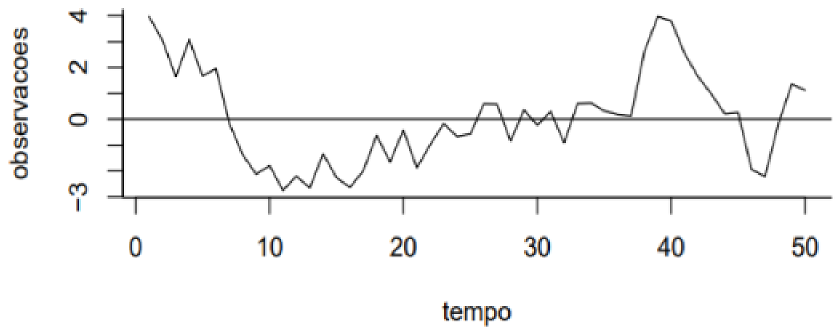
- Série aleatória:
  - Através do correlograma é possível analisar se uma série é aleatória ou não;
  - Em uma série completamente aleatória os lags são não correlacionados, ou seja, espera-se que o coeficiente de autocorrelação amostral rk seja próximo à zero, k = 1,2,....

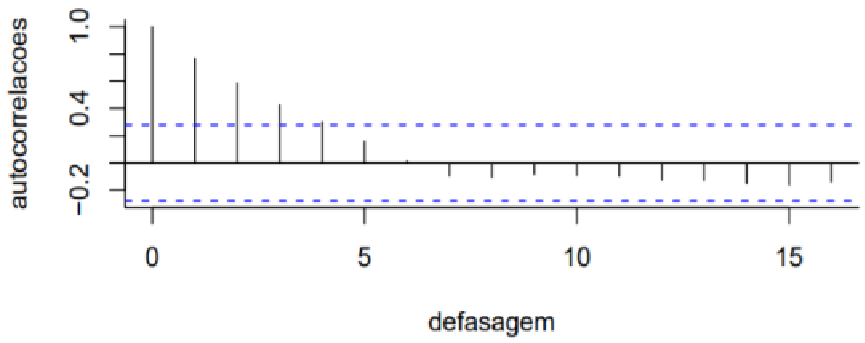




Correlograma: Exemplo

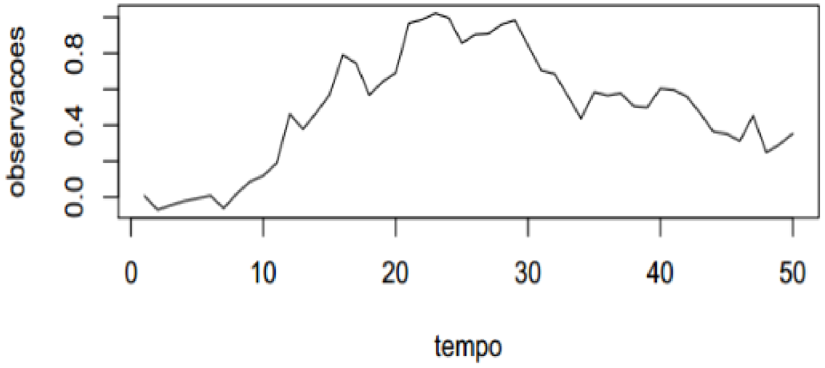
- Correlação de curto prazo:
  - É dita de curto prazo quando uma observação acima da média tende a ser seguida por uma ou mais observações acima da média. O mesmo ocorre para observações abaixo da média.
  - O correlograma desta série será composto por um valor relativamente grande para r1 seguido por valores que tendem a ficar sucessivamente menores;
  - A partir de uma certa defasagem k os valores de rk tendem a ser aproximadamente zero.

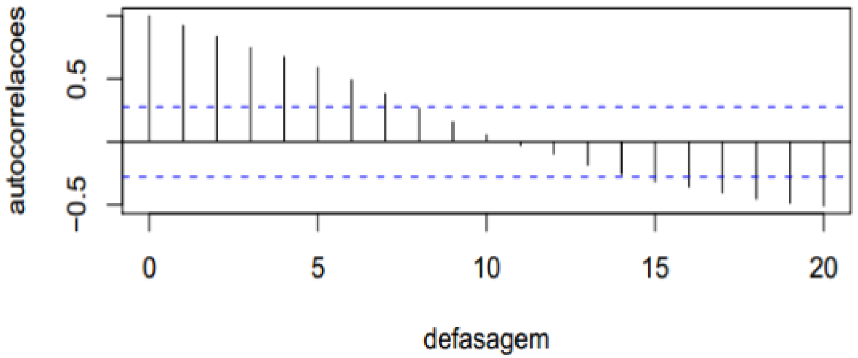




Correlograma: Exemplo

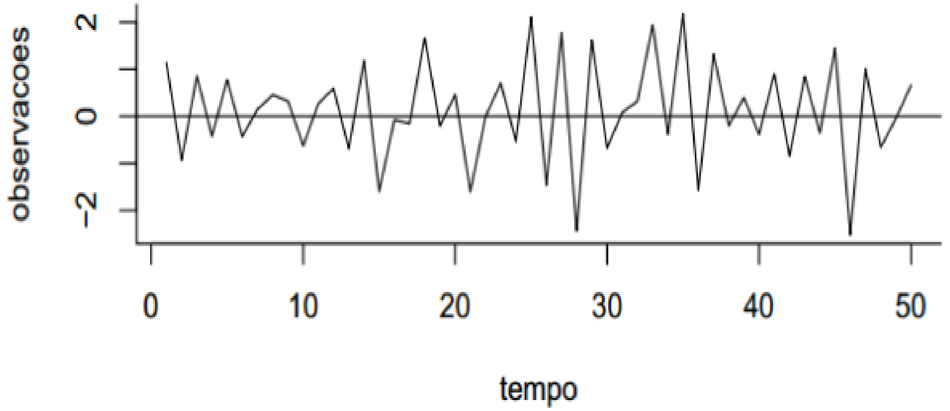
- Séries não estacionária:
  - Para uma série com tendência os valores do coeficiente r não decaem para zero a não ser em defasagens grandes. Isso ocorre pois uma observação de um lado da média tende a ser seguida por um grande número de observações da mesma média (mesmo lado) por conta da tendência;
  - Nesse caso, pouca ou nenhuma informação pode ser obtida do correlograma pois a tendência dominará outras características;

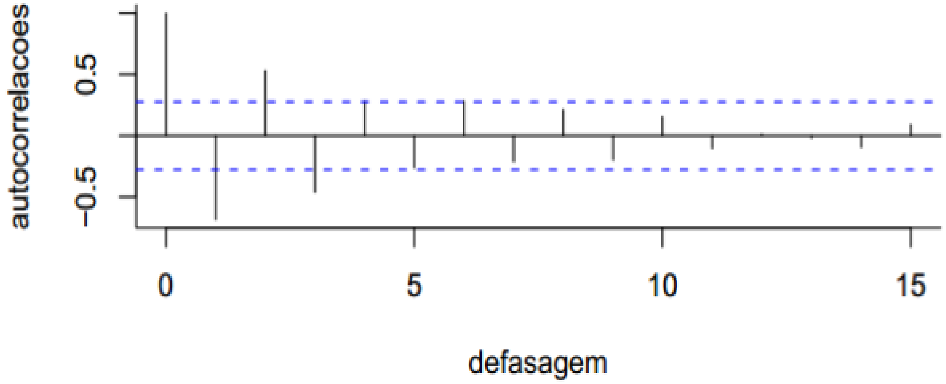




Correlograma: Exemplo

- Correlação negativa:
  - Quando os valores das observações tendem a se alternar acima e abaixo de uma média, o coeficiente de correlação também tende a se alternar;
  - O valor de r1 será negativo enquanto o valor de r2 será positivo já que as observações defasadas de 2 períodos (lags) tendem a estar do mesmo lado da média;





Resumo

- Conceitos:
  - Séries Temporais;
  - Modelos simples com média zero
    - Ruído I.I.D
    - Processo Binário
    - Random Walk
  - Modelos com tendência e sazonalidade
    - Tendência
    - Sazonalidade
  - Modelos Estacionários
    - Estacionariedade
    - Função de Autocorrelação
    - Correlograma

Extras

- Repositórios de séries temporais:
  - https://datamarket.com/data/list/?q=
  - https://research.cs.aalto.fi/aml/datasets.shtml
  - http://archive.ics.uci.edu/ml/index.php
  - https://www.kaggle.com/datasets
- Aula com código em Python:
  - https://github.com/EraylsonGaldino/timeseries/blob/master/TS%20-%20Aula%2001.ipynb

Referências

- BOX, G. E. P. and JENKINS, G. M. (2008). Time series analysis: forecasting and control, 4nd. ed., San Francisco: Holden-Day.
- Brockwell, Peter J. and Davis, Richard A. (2002). Introduction to Time Series and Forecasting, 2nd. ed., Springer-Verlag