# Especialização em *Data Science* e Estatística Aplicada

## Módulo IV - Análise de Séries Temporais

**Prof. Dr. Eder Angelo Milani**

**Goiânia, 2025**

IME INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
FEN FACULDADE DE ENFERMAGEM
UFG UNIVERSIDADE FEDERAL DE GOIÁS

# Conteúdo Programático

- Principais conceitos de séries temporais.
- Funções de autocovariância e de autocorrelação.
- Definição e estimação de tendência e sazonalidade.
- Métodos de suavização.
- Modelagem Box-Jenkins: modelos AR, MA, ARMA, ARIMA e SARIMA.
- Análise de séries temporais interrompidas.

# Conteúdo - Aula 1

1. Principais conceitos de séries temporais
   - Considerações gerais
   - Algumas séries temporais
   - Por que não usar regressão linear?
   - Estacionariedade
   - Valores ausentes

2. Função de autocovariância e autocorrelação

# Introdução

**Análise de séries temporais**

"O termo análise de séries temporais é a tentativa de extrair um resumo significativo e informações estatísticas de pontos de dados organizados em ordem cronológica. É feita a fim de diagnosticar comportamentos passados e predizer comportamentos futuros." — Aileen Nielsen, Análise Prática de Séries Temporais (2021), p. 1

# Introdução

Uma série temporal é qualquer conjunto de observações ordenadas no tempo. São exemplos de séries temporais:

(i) número de casos mensal de tuberculose no Estado de Goiás;
(ii) número de casos mensal de dengue no Estado de Goiás;
(iii) número de casos de queimada no bioma cerrado;
(iv) número de nascidos vivos em Goiás;
(v) índice de preços ao consumidor amplo (IPCA);
(vi) registro de marés no porto de Santos.

# Introdução

Nos exemplos (i) - (v) temos séries temporais *discretas*, enquanto (vi) é um exemplo de uma séries *contínua*.

Muitas vezes, uma série temporal discreta é obtida através de amostragem de uma série temporal contínua em intervalos de tempos iguais, $\Delta t$. Assim, para analisar a série (vi) será necessário amostrá-lá (em intervalos de tempo de uma hora, por exemplo), convertendo a série contínua, observada no intervalo $[0, T]$, digamos, em uma série discreta com $N$ pontos, onde $N = T/\Delta t$.

# Introdução

Para motivar a discussão, considere o exemplo a seguir. Suponha que queiramos medir a temperatura do ar, de dado local, durante 24 horas. Vamos designar por $Z(t)$ a temperatura no instante $t$ (dado em horas, por exemplo).

Designando-se por $Z(15)$ o valor da temperatura no instante $t = 15$, teremos um número real, por exemplo, $Z(15) = 30$.

Na modelagem, para cada $t$ fixo, teremos os valores de uma variável aleatória $Z(t)$, que terá certa distribuição de probabilidade.

# Introdução

Na realidade, o que chamamos de série temporal é uma parte de uma trajetória, dentre muitas que poderiam ter sido observadas. Em algumas situações, quando temos dados experimentais, é possível observar algumas trajetórias do processo sob consideração, mas na maioria dos casos, quando não é possível fazer experimentações, temos uma só trajetória para análise.

Um exemplo de experimentação: suponha que estamos estudando como a pressão arterial de um paciente responde à administração de um determinado medicamento. Podemos repetir o experimento com diferentes pacientes, sob condições controladas e semelhantes (mesma dose, horário, dieta etc.), e medir a pressão arterial a cada 10 minutos por, digamos, 2 horas. Assim, obtemos várias trajetórias da série temporal, uma para cada paciente.

Temos referido o parâmetro $t$ como sendo o tempo, mas a série $Z(t)$ poderá ser função de algum outro parâmetro físico, como espaço ou volume.

# Introdução

De modo bastante geral, uma série temporal poderá ser um vetor $\boldsymbol{Z(t)}$, de ordem $r \times 1$, onde, por usa vez, $\boldsymbol{t}$ é um vetor $p \times 1$. Por exemplo, considere a série

$$\boldsymbol{Z(t)} = [Z_1(\boldsymbol{t}), Z_2(\boldsymbol{t}), Z_3(\boldsymbol{t})]',$$

sendo que as três componentes denotam, respectivamente, a altura, a temperatura e a pressão de um ponto do oceano e $\boldsymbol{t} =$ (tempo, latitude, longitude). Dizemos que a série é multivariada ($r = 3$) e multidimensional ($p = 3$).

Como outro exemplo, considere $\boldsymbol{Z(t)}$ como sendo o número de acidentes ocorridos em rodovias do Estado de São Paulo, por mês. Aqui, $r = 1$ e $p = 2$, com $\boldsymbol{t} =$ (mês, rodovia).

# Introdução

Obtida a série temporal $Z(t_1), Z(t_2), \ldots, Z(t_n)$, observada nos instantes $t_1, t_2, \ldots, t_n$, podemos estar interessados em:

(a) descrever o comportamento da série, verificando a existência de tendências, variações sazonais e ciclos;

(b) fazer previsões de valores futuros da série, estas podem ser a curto prazo, como para série de vendas, produção ou estoque, ou a longo prazo, como para séries populacionais, de produtividade etc.;

(c) procurar periodicidade relevantes nos dados.

# Número de casos mensal de tuberculose

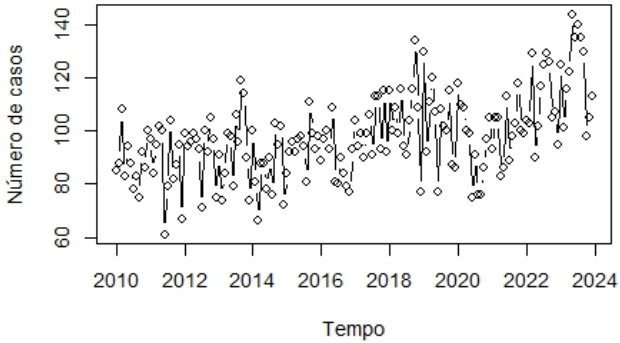


Figura 1: Número de casos mensal de tuberculose no Estado de Goiás, entre janeiro de 2010 a dezembro de 2023.

# Número de casos de dengue no Estado de Goiás

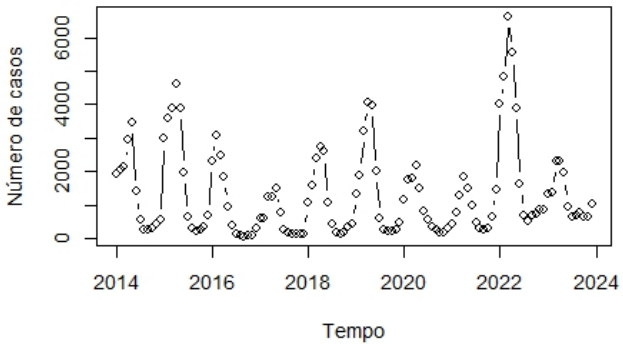


Figura 2: Número de casos mensal de dengue no Estado de Goiás (UF da residência é Goiás) com exame sorológico (IgM) positivo, de janeiro de 2014 a dezembro de 2023.

# Número de casos de queimadas no bioma cerrado

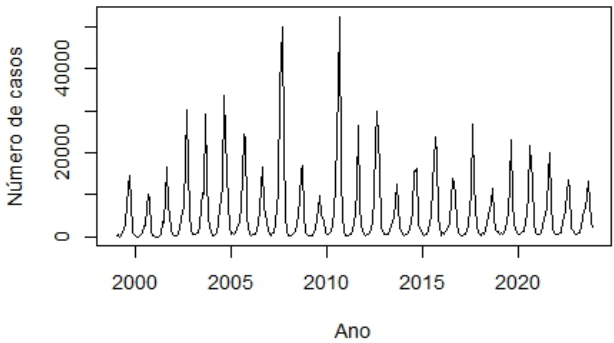


Figura 3: Número mensal de casos de queimadas no bioma cerrado no período de janeiro de 1999 até dezembro de 2023.

# Número de nascidos vivos em Goiás

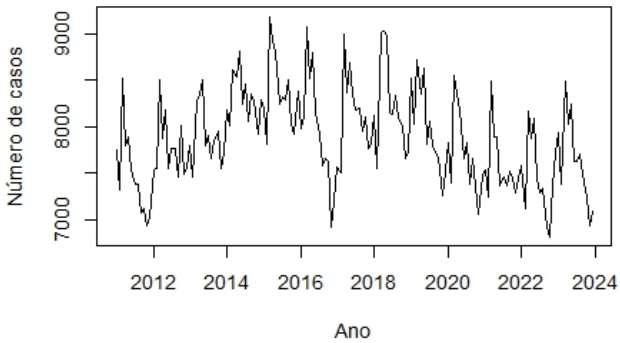


Figura 4: Número mensal de nascidos vivos em Goiás no período de janeiro de 2011 até dezembro de 2023

# IPCA

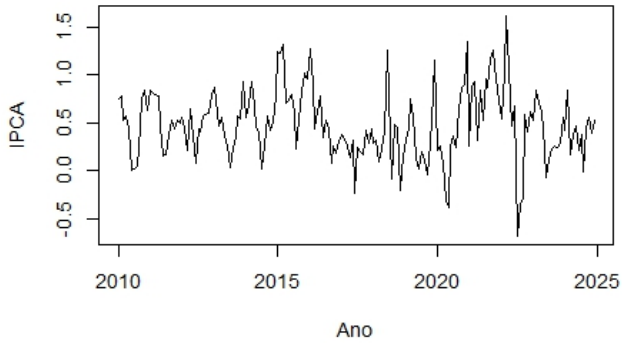


Figura 5: IPCA (índice nacional de preços ao consumidor amplo) do período de julho de 1994 até dezembro de 2024

# Introdução

**Aplicação no** *software* **R.**

# Por que não usar regressão linear?

Por que a regressão linear tradicional não deve ser aplicada diretamente a séries temporais?

A regressão linear tradicional assume que os resíduos:

- têm média zero;
- têm variância constante (homocedasticidade);
- são não correlacionados entre si (ausência de autocorrelação);
- são independentemente e identicamente distribuídos (i.i.d.), em muitos casos com distribuição normal.

# Por que não usar regressão linear?

O problema com séries temporais é que os valores ordenados no tempo apresentam, frequentemente, dependência temporal, ou seja

- $Y_t$ depende de $Y_{t-1}, Y_{t-2}, \ldots$,
- os resíduos $e_t$, em geral, apresentam autocorrelação.

Ignorar a autocorrelação nos resíduos pode levar a problemas como: inferência incorreta (teste de hipóteses e intervalos de confiança), previsões ineficientes e resíduos com padrões não aleatórios.

# Estacionariedade

Uma das suposições mais frequentes que se faz a respeito de uma série temporal é a de que ela é estacionária, ou seja, ela se desenvolve no tempo aleatoriamente ao redor de uma média constante, refletindo alguma forma de equilíbrio estável.

Todavia, a maior parte das séries que encontramos na prática apresentam alguma forma de não-estacionariedade. Assim, as séries econômicas e financeiras apresentam, em geral,+- tendências, sendo o caso mais simples aquele que a série flutua ao redor de uma reta, com inclinação positiva ou negativa (tendência linear). Podemos ter, também, uma forma de não-estacionariedade explosiva, como o crescimento de uma colônia de bactérias.

# Estacionariedade

**Definição:** Um processo estocástico $Z = \{Z(t), t \in T\}$ diz-se fracamente estacionário ou estacionário de segunda ordem se, e somente se,

(i) $E(Z(t)) = \mu(t) = \mu$, constante, para todo $t \in T$;

(ii) $E(Z^2(t)) < \infty$, para todo $t \in T$;

(iii) $\gamma(t_1, t_2) = Cov(Z(t_1), Z(t_2))$ é uma função de $|t_1 - t_2|$.

A partir de agora estaremos interessados somente nesta classe de processos, que denominaremos simplesmente de processos estacionários.

# Estacionariedade

Pausa para uma revisão ...

Seja $X$ uma variável aleatória contínua, a esperança ou média da variável aleatória $X$ é definida por

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx,$$

sendo que $f(x)$ é a função densidade de probabilidade da variável aleatória $X$. No caso onde a variável aleatória $X$ é discreta, a esperança é dada por

$$E(X) = \sum_{x} x P(X = x),$$

sendo que $P(X = x)$ é a probabilidade da variável aleatória $X$ assumir o valor $x$.

# Estacionariedade

Pausa para uma revisão ...

Seja $X$ e $Y$ duas variáveis aleatórias, com valor esperado dado respectivamente por $E(X)$ e $E(Y)$, a covariância entre $X$ e $Y$ é dada por

$$Cov(X, Y) = E(XY) - E(X)E(Y).$$

A covariância mede a tendência linear conjunta das duas variáveis aleatórias.

# Estacionariedade

Uma série pode ser estacionária durante um período muito longo, mas pode ser estacionária apenas em períodos muito curtos, mudando de nível e/ou inclinação.

A classe de modelos ARIMA será capaz de descrever de maneira satisfatória séries estacionárias e séries não-estacionárias, mas que não apresentem comportamento explosivo. Este tipo de não-estacionariedade é chamado homogêneo.

# Estacionariedade

A série pode ser estacionária, flutuando ao redor de um nível, por certo tempo, depois mudar de nível e flutuar ao redor de um novo nível e assim por diante, ou então mudar de inclinação, ou ambas as coisas. A Figura 6 ilustra esta forma de não-estacionariedade.

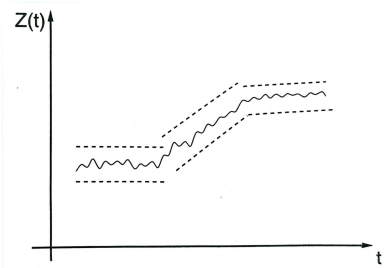


Figura 6: Série não-estacionária quanto ao nível e inclinação - Figura retirada de Morettin e Toloi (2006)

# Estacionariedade

Como a maioria dos procedimentos de análise estatística de séries temporais supõe que estas sejam estacionárias, é necessário transformar os dados originais, se estes não formam uma série estacionária.

A transformação mais comum consiste em tomar diferenças sucessivas da série original, até se obter uma série estacionária. A primeira diferença de $Z(t)$ é definida por

$$\Delta Z(t) = Z(t) - Z(t-1),$$

a segunda diferença é

$$\Delta^2 Z(t) = \Delta[\Delta Z(t)] = \Delta[Z(t) - Z(t-1)] = Z(t) - 2Z(t-1) + Z(t-2).$$

De modo geral, a $n$-ésima diferença de $Z(t)$ é $\Delta^n Z(t) = \Delta[\Delta^{n-1} Z(t)]$.

# Estacionariedade

Considere a série temporal dada a seguir.

| $t$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $Z(t)$ | 10 | 9 | 11 | 8 | 10 |

Calcule a primeira diferença de $Z(t)$ - $\Delta Z(t) = Z(t) - Z(t-1)$.

# Estacionariedade

Considere a série temporal dada a seguir.

| $t$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $Z(t)$ | 10 | 9 | 11 | 8 | 10 |

Calcule a primeira diferença de $Z(t)$ - $\Delta Z(t) = Z(t) - Z(t-1)$.

$$
\begin{aligned}
\Delta Z(2) &= Z(2) - Z(1) = 9 - 10 = -1 \\
\Delta Z(3) &= Z(3) - Z(2) = 11 - 9 = 2 \\
\Delta Z(4) &= Z(4) - Z(3) = 8 - 11 = -3 \\
\Delta Z(5) &= Z(5) - Z(4) = 10 - 8 = 2
\end{aligned}
$$

# Estacionariedade

Considere a série temporal dada a seguir.

| $t$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $Z(t)$ | 10 | 9 | 11 | 8 | 10 |

Calcule a segunda diferença de $Z(t)$ - $\Delta^2 Z(t) = Z(t) - 2Z(t-1) + Z(t-2)$.

# Estacionariedade

Considere a série temporal dada a seguir.

| $t$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $Z(t)$ | 10 | 9 | 11 | 8 | 10 |

Calcule a segunda diferença de $Z(t)$ - $\Delta^2 Z(t) = Z(t) - 2Z(t-1) + Z(t-2)$.

$$
\begin{aligned}
\Delta^2 Z(3) &= Z(3) - 2Z(2) + Z(1) = 11 - 2 \times 9 + 10 = 3 \\
\Delta^2 Z(4) &= Z(4) - 2Z(3) + Z(2) = 8 - 2 \times 11 + 9 = -5 \\
\Delta^2 Z(5) &= Z(5) - 2Z(4) + Z(3) = 10 - 2 \times 8 + 11 = 5
\end{aligned}
$$

# Estacionariedade

Em situações normais, será suficiente aplicar uma ou duas diferenças para que a série se torne estacionária. Na Figura 7 é apresentada a série do número de casos de nascidos vivos em Goiás, acompanhada de sua primeira diferença.

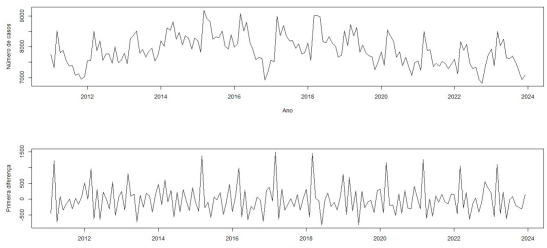


Figura 7: A série temporal do número de casos de nascidos vivos em Goiás, de janeiro de 2011 a dezembro de 2023, acompanhada de sua primeira diferença

# Estacionariedade

**Aplicação no** *software* **R.**

# Procedimento de previsão

Suponha que temos observações de uma série temporal até o instante $t$ e queiramos prever o valor da série no instante $t + h$, como na Figura 8.

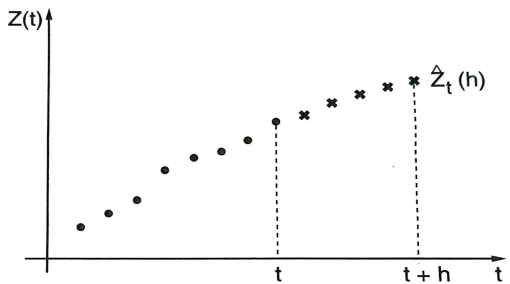


Figura 8: Observações de uma série temporal com previsões de origem $t$ e horizonte $h$ - Figura retirada de Morettin e Toloi (2006)

# Procedimento de previsão

Diremos que $\hat{Z}_t(h)$ é a previsão de $Z(t+h)$, de origem $t$ e horizonte $h$.

Exemplo: Suponha que a série temporal $Z(t)$ tenha sido observada até o instante de tempo $t = 120$, e queremos as previsões para os instantes de tempo $t = 121, 122$ e $123$. Utilizando da notação, temos:

$$\begin{aligned} \hat{Z}_{120}(1) &- \text{é a previsão para } Z(121) \\ \hat{Z}_{120}(2) &- \text{é a previsão para } Z(122) \\ \hat{Z}_{120}(3) &- \text{é a previsão para } Z(123) \end{aligned}$$

# Procedimento de previsão

Exemplo: Considere agora que um novo valor da série foi observado, ou seja, $Z(121)$, como fica a notação para as previsões para os instantes de tempo $t = 122$ e $123$, considerando a nova observação?

# Procedimento de previsão

Exemplo: Considere agora que um novo valor da série foi observado, ou seja, $Z(121)$, como fica a notação para as previsões para os instantes de tempo $t = 122$ e $123$, considerando a nova observação?

A notação agora é dada por

$$\hat{Z}_{121}(1) \quad - \quad \text{é a previsão para } Z(122)$$
$$\hat{Z}_{121}(2) \quad - \quad \text{é a previsão para } Z(123)$$

**Muita atenção à origem e horizonte de previsão!**

# Procedimento de previsão

A palavra previsão sugere que se quer ver uma coisa antes que ela exista. Alguns autores preferem a palavra predição, para indicar algo que deverá existir no futuro. Ainda outros utilizam o termo projeção. Nestas notas, iremos usar consistentemente a palavra previsão, com o sentido indicado acima.

# Como quantificar o erro de previsão

O erro absoluto médio (EAM), o erro quadrado médio (EQM) e o erro médio absoluto percentual (MAPE), são dados, respectivamente, por

$$\begin{aligned}
\text{EAM} &= \frac{1}{k}\sum_{h=1}^{k}\left|Z(t+h)-\hat{Z}_t(h)\right|, \\
\text{EQM} &= \frac{1}{k}\sum_{h=1}^{k}\left(Z(t+h)-\hat{Z}_t(h)\right)^2, \\
\text{MAPE} &= 100\left(\frac{1}{k}\sum_{h=1}^{k}\left|\frac{Z(t+h)-\hat{Z}_t(h)}{Z(t+h)}\right|\right)\%,
\end{aligned}$$

sendo que $k$ é a quantidade de previsões realizadas.

# Lidando com valores ausentes

Dados ausentes são comuns em séries temporais. Por exemplo, na área de assistência médica, os dados ausentes podem ser ocasionados por:

(i) paciente que não tomou as medidas desejadas;

(ii) paciente que estava com a saúde em boas condições, assim não havia a necessidade de tomar uma medição específica;

(iii) dispositivo médico que apresentou um defeito técnico aleatório;

(iv) erro ocorrido na entrada de dados.

# Lidando com valores ausentes

Geralmente de maneira bem arriscada, os dados ausentes são ainda mais comuns na análise de séries temporais do que na análise transversal de dados, porque a informação da amostragem longitudinal é bastante pesada: séries temporais incompletas são bem comuns e, por causa disso, se desenvolveu métodos para lidar com a ausência nos dados registrados.

Os métodos mais comuns para lidar com dados ausentes em séries temporais são:

- Imputação - quando preenchemos os dados ausentes com base em observações do conjunto de dados;
- Interpolação - quando usamos pontos de dados vizinhos a fim de estimar o valor ausente.

# Lidando com valores ausentes

**Método de preenchimento** *forward fill*
Uma das formas mais simples de se preencher os valores ausentes é repetir o último valor conhecido para o valor ausente. Observe que não é necessário recorrer a cálculos matemáticos complexos.

Considere o caso em que o valor ausente é resultado de uma decisão médica devido ao fato que o profissional da saúde não achou necessário repetir um exame, porque esperava que a medição do paciente fosse normal. Em muitos casos, isso significa que poderíamos aplicar o *forward fill* aos valores ausentes com o último valor conhecido, já que esse era o pressuposto que motivou o profissional da saúde a não refazer a medição.

Obs.: também pode ser utilizado o *backward fill* - que é utilizar o primeiro valor após o dado ausente para a imputação.

# Lidando com valores ausentes

**Método de média móvel ou mediana móvel**
Podemos imputar dados com uma média móvel ou mediana móvel. A média móvel é semelhante a um *forward fill* com relação a usar valores passados a fim de "predizer" os valores futuros ausentes (a imputação pode ser uma forma de previsão). Mas com a média móvel, utiliza-se alguns valores.

Há diversas situações em que uma imputação de dados de média móvel melhor se adequa à tarefa em questão do que um *forward fill*. Por exemplo, se os dados forem ruidosos e você tiver razões para duvidar do valor de qualquer ponto de dados individual em relação a uma média geral.

Obs.: a média móvel não precisa necessariamente ser uma média aritmética, pois pode ser dado peso maior as observações mais recentes.

# Lidando com valores ausentes

**Interpolação**
A interpolação é um método para determinar os valores dos pontos de dados ausentes com base em restrições geométricas sobre como queremos que os dados gerais se comportem. Por exemplo, uma interpolação linear restringe os dados ausentes a um ajuste linear com pontos vizinhos conhecidos.

Há muitas situações em que uma interpolação linear ou spline é oportuna. Pense na temperatura média semanal, quando há uma tendência conhecida de aumento ou redução das temperaturas dependendo da época do ano.

Obs.: assim como acontece com a média móvel, a interpolação pode ser feita de modo que considere os dados passados e futuros ou só do passado.

# Lidando com valores ausentes

**Aplicação no** *software* **R.**

# Função de autocovariância e autocorrelação

**Definição:** A autocorrelação, também conhecida como correlação serial, é a correlação de um sinal com uma cópia atrasada de si mesmo. Informalmente, é a semelhança entre as observações em função do *lag* de tempo entre elas.

A autocorrelação dá ideia de como os pontos de dados em diferentes pontos de tempo estão linearmente relacionados entre si em função de sua diferença de tempo.

Embora seja simples calcular a função de autocorrelação (FAC, ou do inglês ACF) com um código customizado, geralmente é melhor usar uma função predefinida, como a função *acf* do R. Algumas das vantagens são: plotagem automática com rótulos úteis e um número razoável (geralmente, mas nem sempre) de *lags* para se calcular a FAC.

# Função de autocovariância e autocorrelação

Seja $\{X(t), t \in \mathbb{Z}\}$ um processo estacionário real discreto, a função de autocovariância (FACV) é dada por

$$\gamma_\tau = Cov(X_t, X_{t+\tau}) = E(X_t X_{t+\tau}) - E(X_t)E(X_{t+\tau}).$$

A FAC do processo é definida por

$$\rho_\tau = \frac{\gamma_\tau}{\gamma_0},\ \tau \in \mathbb{Z},$$

note que $\rho_0 = 1$. Além disso, temos que $-1 \leq \rho_\tau \leq 1$.

# Função de autocovariância e autocorrelação

Dadas observações $X_1, \ldots, X_n$, a FAC $\rho_j$ é estimada por

$$r_j = \frac{c_j}{c_0},\ j = 0, 1, \ldots, N-1,$$

onde $c_j$ é a estimativa da função de autocovariância $\gamma_j$, sendo dada por

$$c_j = \frac{1}{N} \sum_{t=1}^{N-j} [(X_t - \overline{X})(X_{t+j} - \overline{X})],\ j = 0, 1, \ldots, N-1,$$

sendo $\overline{X} = \frac{1}{N} \sum_{t=1}^{N} X_t$ a média amostral. Aqui, colocamos $c_{-j} = c_j$ e $r_{-j} = r_j$.

# Função de autocovariância e autocorrelação

Em termos matemáticos, existem alguns fatores importantes sobre a FAC:

(i) a FAC de uma função periódica tem a mesma periodicidade do processo original.

(ii) Todas as séries temporais têm uma autocorrelação de 1 no *lag* 0.

(iii) A autocorrelação de uma amostra de ruído branco terá um valor de aproximadamente 0 em todos os *lags* diferentes de 0.

(iv) A FAC é simétrica em relação aos *lags* negativos e positivos, assim apenas os *lags* positivos precisam ser considerados explicitamente.

(v) Uma regra estatística para determinar uma estimativa FAC diferente de zero significativa é dada por uma "região crítica" com fronteira em $\pm 1,96 \times \sqrt{n}$. Essa regra se baseia em um tamanho de amostra grande o bastante e em uma variância finita para o processo.

# Função de autocovariância e autocorrelação

**Definição:** Dizemos que $\{\epsilon_t, t \in \mathbb{Z}\}$ é ruído branco discreto se as v.a. $\epsilon_t$ são não correlacionadas, isto é, $Cov(\epsilon_t, \epsilon_s) = 0,\ t \neq s$.

Tal processo será estacionário se $E(\epsilon_t) = \mu_\epsilon$ e $Var(\epsilon_t) = \sigma^2_\epsilon$, para todo $t$. Segue que a FACV de $\epsilon$ é dada por

$$\gamma_\tau = Cov(\epsilon_n, \epsilon_{n+\tau}) = \begin{cases} \sigma^2_\epsilon, & \text{se } \tau = 0 \\ 0, & \text{se } \tau \neq 0. \end{cases}$$

Obviamente, se as v.a. $\epsilon_t$ são independentes, elas também serão não-correlacionadas. Uma sequência de v.a. i.i.d., como definida acima, é chamada um processo puramente aleatório.

# Função de autocovariância e autocorrelação

Ilustramos na Figura 9 a FAC de um ruído branco. De agora em diante iremos supor que $\mu_\epsilon = 0$. Escrevemos, brevemente, $\epsilon_t \sim RB(0, \sigma^2_\epsilon)$.

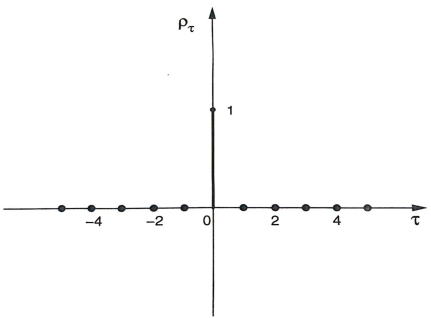


Figura 9: FAC de um ruído branco - Figura retirada de Morettin e Toloi (2006)

# Função de autocovariância e autocorrelação

**Aplicação no** *software* **R.**

# Especialização em *Data Science* e Estatística Aplicada

## Módulo IV - Análise de Séries Temporais

**Prof. Dr. Eder Angelo Milani**

**edermilani@ufg.br**

IME INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

FEN FACULDADE DE ENFERMAGEM

UFG UNIVERSIDADE FEDERAL DE GOIÁS