
**UNIVERSIDADE ESTADUAL DA PARAÍBA**
**CAMPUS I**
**CENTRO DE CIÊNCIAS E TECNOLOGIA**
**DEPARTAMENTO DE ESTATÍSTICA**
**CURSO DE BACHARELADO EM ESTATÍSTICA**

**KAILCA LÍBIA SANTOS MARTINS**

# INFERÊNCIA EM CADEIAS DE MARKOV

**CAMPINA GRANDE - PB**
**2024**

KAILCA LÍBIA SANTOS MARTINS

# INFERÊNCIA EM CADEIAS DE MARKOV

Trabalho de Conclusão de Curso (Artigo) apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Estatística.

**Orientador:** Prof. Dra. Divanilda Maia Esteves

**CAMPINA GRANDE - PB**
**2024**

M386i Martins, Kailca Líbia Santos.
Inferência em Cadeias de Markov [manuscrito] / Kailca Líbia Santos Martins. - 2024.
20 f.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2024.

"Orientação : Prof. Dra. Divanilda Maia Esteves, Departamento de Estatística - CCT".

1. Processos estocásticos. 2. Processos markovianos. 3. Cadeias de ordem superior. I. Título

21. ed. CDD 519.233

Elaborada por Lêda Cristina Diniz Andrade - CRB - 15/1032 BC

KAILCA LÍBIA SANTOS MARTINS

INFERÊNCIA EM CADEIAS DE MARKOV

Trabalho de Conclusão de Curso (Artigo) apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Estatística.

Trabalho aprovado em 21 de novembro de 2024.

**BANCA EXAMINADORA**

DMEsteves

Prof. Dra. Divanilda Maia Esteves (Orientadora)
Universidade Estadual da Paraíba (UEPB)

Mª das Vitórias A. Seratim

Prof. Me. Maria das Vitórias Alexandre Serafim
Universidade Estadual da Paraíba (UEPB)

Prof. Dr. Sílvio Fernando Alves Xavier Júnior
Universidade Estadual da Paraíba (UEPB)

*Dedico este trabalho aos meus pais, que sempre me apoiaram e acreditaram no meu potencial.*

“Porque sou eu que conheço os planos que tenho para vocês, diz o Senhor,
planos de fazê-los prosperar e não causar dano, planos de dar a vocês esperaça e um futuro.”
(Jeremias 29:11)

**SUMÁRIO**

# INFERÊNCIA EM CADEIAS DE MARKOV

Kailca Líbia Santos Martins*
Divanilda Maia Esteves†

## RESUMO

Um processo estocástico é uma sequência de variáveis aleatórias usada para modelar eventos que sucedem de maneira aleatória. Já a propriedade Markoviana refere-se à ação de perda de memória nesse processo, em que situações futuras serão fundamentadas apenas em $k$ estados mais recentes, independentemente do que tenha acontecido antes. Esse trabalho tem a finalidade de inferir informações por meio do modelo de Markov. Para isso, foram utilizados dois bancos de dados que fazem alusão à cotação da bolsa de valores durante os anos de 2021 e 2023, obtidos no site *Infomoney*[1]. Para a análise dos dados foi utilizado o software R. Além de testes estatísticos, critérios de informação foram empregados para avaliar a viabilidade da aplicação de cadeias de Markov aos dados considerados. A análise realizada buscou compreender qual modelo de cadeias de Markov os dados se adequam de forma mais eficiente, sendo de primeira, segunda ou terceira ordem. Por meio dos cálculos das probabilidades e testes de hipóteses, ficou comprovado que o modelo de melhor ajuste é o de primeira ordem para os dados de 2021 e um modelo independente para o período de 2023. Diante disso, pode-se ressaltar que o estudo sobre cadeias de Markov segue sendo de grande importância, pois promove uma abordagem matemática eficiente em inúmeras áreas do conhecimento, a exemplo de modelagem de jogos de azar, tratamento analítico de filas, preços de ativos, sistemas de comunicação, sequências de DNA, previsão do tempo, entre outras.

**Palavras-chaves**: Processos estocásticos; processos markovianos; cadeias de ordem superior.

## ABSTRACT

A stochastic process is a sequence of random variables used to model events that occur randomly. The Markovian property refers to the memory loss action in this process, in which future situations will be based only on the $k$ most recent states, regardless of what happened before. This work aims to infer information through the Markov model. For this, two databases that refer to stock market prices during the years 2021 and 2023, obtained from the website *Infomoney*, were used. The R software was used to analyze the data. In addition to statistical tests, information criteria were used to assess the feasibility of applying Markov chains to the data considered. The analysis performed sought to understand which Markov chain model the data fits most efficiently, being first, second or third order. Through probability calculations and hypothesis testing, it was proven that the best-fitting model is the first-order model for the 2021 data and an independent model for the 2023 period. In view of this, it can be highlighted that the study of Markov chains continues to be of great importance, as it promotes an efficient mathematical approach in numerous areas of knowledge, such as modeling gambling, analytical queue treatment, asset prices, communication systems, DNA sequences, weather forecasting, among others.

**Keywords**: Stochastic processes; markovian processes; higher-order chains.

* Aluno do curso de Estatística, Depto de Estatística, UEPB, Campina Grande, PB, kailca.martins@aluno.uepb.edu.br
† Prof. de Estatística, Depto de Estatística, UEPB, Campina Grande, PB, diana.maia@servidor.uepb.edu.br
1 <https://www.infomoney.com.br/cotacoes/b3/etf/etf-qbtc11/historico/>

# 1 INTRODUÇÃO

Um processo estocástico é uma sequência de variáveis aleatórias $\{X(t), t \in T\}$ usada para modelagem de um experimento que se desenvolve aleatoriamente ao longo do tempo ou espaço. Neste caso, $X(t)$ representa o estado do processo no tempo $t$. O conjunto $T$ que indexa as variáveis $X(t)$ pode ser discreto ou contínuo, o que reflete o modo como o processo é observado, sendo em intervalos sistemáticos de tempo ($T$ discreto) ou a qualquer momento ($T$ contínuo). Além disso, as variáveis consideradas podem ser discretas ou contínuas (BARBOSA, 2009).

As cadeias de Markov são um caso particular dos processos estocásticos que seguem a propriedade Markoviana, ou seja, em que as situações futuras são influenciadas apenas pelo estado mais recente que se tem conhecimento. Na verdade, essas cadeias assim descritas são as cadeias de ordem 1. Uma definição mais ampla considera que uma cadeia de Markov de ordem $k$ é um processo estocástico, cujas variáveis $X(t)$ são discretas e cujo estado presente do processo é influenciado pelas $k$ observações mais recentes (ROSS, 1996).

As cadeias de ordem maior do que 1 não são tão frequentemente usadas quanto as de ordem 1 e isso deve ser explicado em parte porque à medida que a ordem aumenta, aumenta também a quantidade de parâmetros do modelo, o que requer amostras maiores para estimar tais parâmetros (BHAT; MILLER, 2002). Neste trabalho, foram estudadas as cadeias de Markov de ordem 1 e de ordens superiores, do ponto de vista prático, dando ênfase na parte da inferência estatística. Além das definições e principais características dessas cadeias, foram vistos conteúdos como estimação dos parâmetros e testes de hipótese.

Outra característica presente nesse estudo é a escolha de uma cadeia de Markov com apenas dois estados, tornado a análise dos resultados bem mais simplificada. Com eles, a matriz de transição possuirá apenas quatro elementos para um modelo de primeira ordem, o que facilita a estimação dos parâmetros a partir dos dados observados. Além disso, a visualização e a interpretação das probabilidades de transição são mais intuitivas, permitindo um diagnóstico mais objetivo da cadeia.

A modelagem foi aplicada a dois conjuntos de dados referentes às cotações diárias da bolsa de valores (Ibovespa) no período de 2021 e 2023. O objetivo é investigar se o comportamento das cotações pode ser ajustado a um modelo Markoviano e, caso isso suceda, determinar a ordem mais adequada para representar esses dados.

Segundo o site *Infomoney*[2], o Ibovespa é o principal índice da bolsa brasileira, criado em 1968. O índice é calculado com base no desempenho de ações das maiores empresas de capital aberto, como Vale, Petrobras, Itaú, Bradesco, Banco do Brasil, Eletrobras, entre outros. Portanto, é de grande valia associar um indicador econômico nacional a um modelo matemático pertinente e inferir informações sobre essa relação.

[2] <https://www.infomoney.com.br/cotacoes/b3/etf/etf-qbtc11/historico/>

# 2 FUNDAMENTAÇÃO TEÓRICA

## 2.1 Processos Estocásticos

Um processo estocástico pode ser definido como uma família de variáveis aleatórias que descreve o desenvolvimento de um processo físico ao longo do tempo e espaço, podendo ser classificado como discreto ou contínuo. Assim, seja $X(t) : t \in T$, em que, para cada $t \in T$, $X(t)$ é uma variável aleatória que, por exemplo, pode representar o número de acidentes automobilísticos em uma determinada cidade no mês $t$. Além disso, tem-se um conjunto chamado de espaço de estados do processo, que denota todos os valores prováveis que $X(t)$ pode assumir. O conjunto de índices dessas variáveis $X(t)$ é representado por $T$, e o parâmetro $t$, que varia dentro desse conjunto $T$, geralmente é simbolizado pelo tempo (ROSS, 2007).

Como citado anteriormente, dentro de um processo estocástico há um conjunto chamado espaço de estados, geralmente caracterizado por $S$. Quando o espaço de estados $S$ é um conjunto enumerável de valores inteiros, temos um processo estocástico discreto. No entanto, se esse espaço de estados é formado por valores em um intervalo contínuo, o processo é chamado de processo estocástico contínuo. De maneira semelhante, temos as mesmas nomeações para o conjunto de índices $T$. Se enumerável, o processo é classificado como de tempo discreto; nesse caso, $T$ geralmente um subconjunto dos números naturais. Quando o conjunto de índices $T$ é um intervalo contínuo, o processo é classificado como de tempo contínuo (SANTOS, 2017).

Outro ponto importante é entender a estrutura de dependência entre as variáveis. Um caso bem conhecido é o da independência, que é bastante considerado em análises estatísticas, mas que nem sempre será válido. O caso que será considerado aqui é o da dependência de Markov, que tem como princípio a propriedade markoviana de "perda de memória" (SANTOS, 2017). Neste trabalho, os resultados obtidos permitem observar os dois pontos mencionados acima, embora o foco esteja no modelo de Markov.

## 2.2 Cadeias de Markov

Cadeias de Markov são processos estocásticos que seguem a propriedade Markoviana. Diante disso, pode-se afirmar que um processo estocástico nada mais é do que o estudo de eventos que sucedem de maneira aleatória. Já a propriedade Markoviana refere-se à ação de "perda de memória" nesse processo, em que situações futuras serão fundamentadas apenas em seu estado mais recente, independentemente do que tenha acontecido antes (ROSS, 2007).

Um processo $\{X_n : n \geq 0\}$ assumindo valores em um conjunto $S$ é uma cadeia de Markov de ordem $k$ se

$$\mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_1 = x_1, X_0 = x_0) = \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_{n-k} = x_{n-k}). \tag{2.2.1}$$

Se, em particular

$$\mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_{n-k} = x_{n-k}) = \mathbb{P}(X_k = x_k | X_{k-1} = x_{k-1}, \ldots, X_0 = x_0), \tag{2.2.2}$$

para qualquer $n \geq 0$, diz-se que a cadeia é homogênea no tempo e, neste caso, frequentemente as probabilidades acima são representadas em forma matricial.

Quando $k = 1$, temos uma cadeia de Markov de ordem 1, a qual é frequentemente chamada apenas de Cadeia de Markov.

Para uma cadeia de Markov $\{X_n, n \geq 0\}$ com espaço de estados $S$, a função de transição da cadeia é dada por

$$P_n(x, y) = P(X_n = y | X_{n-1} = x), \tag{2.2.3}$$

onde $x, y \in S$, e é tal que para quaisquer $x, y \in S$

$$P_n(x,y) \geq 0 \quad \text{e} \quad \sum_{y \in S} P_n(x,y) = 1.$$

Quando as probabilidades acima não dependem de $n$, então a cadeia é dita ser homogênea no tempo. Neste caso, as probabilidades de transição, $P(x,y)$, podem ser representadas matricialmente em uma matriz $P$ chamada de matriz de transição a um passo da cadeia, ou apenas, matriz de transição da cadeia. Para uma cadeia de ordem 1, por exemplo, definindo-se

$$P(x,y) = \mathbb{P}(X_{n+1} = y | X_n = x) = \mathbb{P}(X_1 = y | X_0 = x), \tag{2.2.4}$$

então pode-se representar tais probabilidades na forma

$$P = \begin{bmatrix} P(0,0) & P(0,1) & P(0,2) & \cdots \\ P(1,0) & P(1,1) & P(1,2) & \cdots \\ P(2,0) & P(2,1) & P(2,2) & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}.$$

Tal matriz, por construção tem todos os termos não negativos e a soma dos elementos de uma linha qualquer deve ser igual a 1.

Para ordens maiores, tal representação também pode ser usada, ainda que a medida que a ordem da cadeia aumenta, aumenta também o tamanho dessa matriz. Por exemplo, em uma cadeia com dois estados, a matriz de transição será $2 \times 2$ para uma cadeia de ordem 1, $4 \times 4$ para uma cadeia de ordem 2 e $8 \times 8$ para uma cadeia de ordem 3.

Segundo Bhat e Miller (2002), a análise de uma cadeia de ordem superior com grandes valores é bastante difícil. Porém, para ordens "pequenas" como 2 ou 3, quando espaço de estados é pequeno, cadeias de Markov de ordem superior podem ser analisadas de modo similar às cadeias de Markov de primeira ordem. Considere uma cadeia de Markov $\{X_n, n \geq 0\}$ de segunda ordem com estados $\{0,1\}$. Neste caso, as probabilidades de transição podem ser representadas matricialmente como

$$\begin{bmatrix} P(00,00) & P(00,01) & P(00,10) & P(00,11) \\ P(01,00) & P(01,01) & P(01,10) & P(01,11) \\ P(10,00) & P(10,01) & P(10,10) & P(10,11) \\ P(11,00) & P(11,01) & P(11,10) & P(11,11) \end{bmatrix}$$

sendo que as probabilidades $P(xy,zw) = 0$ se $y \neq z$, pois

$$P(xy,zw) = P(X_{n+1} = w, X_n = z | X_n = y, X_{n-1} = x). \tag{2.2.5}$$

A matriz acima é conveniente para operações matriciais por ser quadrada, mas para fins de interpretação, uma forma mais adequada é

$$\begin{bmatrix} P(00,0) & P(00,1) \\ P(01,0) & P(01,1) \\ P(10,0) & P(10,1) \\ P(11,0) & P(11,1) \end{bmatrix}$$

com

$$P(xy,z) = P(X_{n+1} = z | X_n = y, X_{n-1} = x). \tag{2.2.6}$$

## 2.3 Inferência estatística para cadeias de Markov

Nesta etapa do trabalho, será descrito aspectos da inferência estatística, a qual engloba a estimativa de elementos de probabilidades e a realização de testes de hipóteses, particularmente no contexto das cadeias de Markov. Segundo Bhat e Miller (2002), a área inferencial aplicada a processos estocásticos tem experimentado um crescimento expressivo nas últimas décadas. E que por mais que os aspectos probabilísticos desses processos seja mais enfatizado, os testes paramétricos e as estimativas desempenham um papel importante e não devem ser desprezadas. Diante desse fato, será abordado na sequência a inferência em tempos discretos para as cadeias de Markov.

### *2.3.1 Estimativa de uma matriz de probabilidade de transição*

Considere uma amostra de tamanho $n$ de uma cadeia de Markov finita com $s$ estados, $S = \{1,2,\ldots,s\}$. Seja $n_{ij}$ o número de transições de $i$ para $j$ $(i,j = 1,2,\ldots,s)$, sendo $\sum_{j=1}^{S} n_{ij} = n_i$. Estes valores podem ser representados como

| | 1 | 2 | 3 | ... | $S$ | *Soma* |
|---|---|---|---|---|---|---|
| 1 | $n_{11}$ | $n_{12}$ | $n_{13}$ | ... | $n_{1s}$ | $n_{1.}$ |
| 2 | $n_{21}$ | $n_{22}$ | $n_{23}$ | ... | $n_{2s}$ | $n_{2.}$ |
| 3 | $n_{31}$ | $n_{32}$ | $n_{33}$ | ... | $n_{3s}$ | $n_{3.}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $s$ | $n_{s1}$ | $n_{s2}$ | $n_{s3}$ | ... | $n_{ss}$ | $n_{s.}$ |
| | $n_{.1}$ | $n_{.2}$ | $n_{.3}$ | ... | $n_{.s}$ | $n$ |

As estimativas de máxima verossimilhança das probabilidades $P(i,j)$, as quais serão denotadas por $\hat{P}(i,j)$ $(i,j = 1,2,\ldots,S)$, são (BHAT; MILLER, 2002)

$$\hat{P}(i,j) = \frac{n_{ij}}{n_{i.}} \qquad i,j = 1,2,\ldots,s. \tag{2.3.1}$$

Ou seja, as estimativas dos elementos de uma matriz de probabilidade de transição $(\hat{P})$ são obtidas como a razão entre o número de transições do estado $i$ para o estado $j$ e o número de vezes em que a cadeia esteve no estado $i$ seguido de qualquer estado na realização da cadeia.

De modo similar obtém-se as probabilidades de transição estimadas para uma cadeia de ordem maior. Por exemplo, para uma cadeia de ordem 2,

$$\hat{P}(ij,k) = \frac{n_{ijk}}{n_{ij.}} \qquad i,j,k = 1,2,\ldots,s. \tag{2.3.2}$$

e, para ordem 3,

$$\hat{P}(ijk,p) = \frac{n_{ijkp}}{n_{ijk.}} \qquad i,j,k,p = 1,2,\ldots,s. \tag{2.3.3}$$

### *2.3.2 Teste de Independência versus Dependência de Primeira Ordem*

Suponha que desejamos realizar um teste para a hipótese nula de que as observações coletadas são independentes, contra a alternativa de que o processo observado é uma cadeia de Markov de primeira ordem. Em outras palavras, a hipótese nula será

$$H_0 : P = P^0,$$

em que $P^0$ tem linhas idênticas sob a conjectura de independência. Mais especificamente, seja $P$ composto de $s$ linhas idênticas $\pi = (\pi_1, \pi_2, \ldots, \pi_s)$. De acordo com (BHAT; MILLER, 2002), quando essas probabilidades não são conhecidas as estimativas de máxima verossimilhança podem ser determinadas como segue. Seja $n_j = \sum_{i=1}^{s} n_{ij}$. Então, a função de log-verossimilhança pode ser escrita como:

$$L(\pi) = \ln B + \sum_{j=1}^{s} n_{.j} \ln \pi_j. \tag{2.3.4}$$

Com o auxílio da equação 2.3.4 as estimativas de máxima verossimilhança serão:

$$\hat{\pi_j} = \frac{n_j}{n}, \quad j = 1, \ldots, s. \tag{2.3.5}$$

A estatística $\chi^2$ para testar independência versus dependência de primeira ordem toma a forma:

$$\sum_{i=1}^{s} \sum_{j=1}^{s} \frac{(n_{ij} - n_i n_{.j}/n)^2}{n_i n_{.j}/n} \tag{2.3.6}$$

com graus de liberdade $(s-1)^2$.

Assim, essa estatística de teste é utilizada para averiguar a hipótese nula de que os dados coletados seguem um modelo independente ou a hipótese alternativa de que as observação se adequam a um modelo de ordem 1.

### *2.3.3 Teste para Ordem de uma Cadeia de Markov*

O teste de independência pode ser generalizado para permitir um teste para ordem, possivelmente maior que 1, de uma cadeia de Markov. Como retratado anteriormente, as cadeias de Markov de ordem superior a 1 podem ser reduzidas a uma cadeia de Markov de primeira ordem expandindo o espaço de estados para indicar a dependência de ordem superior. A seguir, temos este procedimento usando uma cadeia de Markov de segunda ordem com probabilidades de transição estacionárias. Seja (BHAT; MILLER, 2002)

$$P_{ijk} = P(X_n = k | X_{n-1} = j, X_{n-2} = i) \tag{2.3.7}$$

e $n_{ijk}$ o correspondente número de transições. Além disso, se $n_{ij.} = \sum_k n_{ijk}$, a estimativa de $P_{ijk}$ é obtida como:

$$\hat{P}_{ijk} = \frac{n_{ijk}}{n_{ij.}}. \tag{2.3.8}$$

A hipótese nula de que a cadeia de Markov é de primeira ordem contra a alternativa de que é de ordem 2 pode ser dada como:

$$H_0 : P_{ijk} = P_{jk}, \quad i, j, k = 1, 2, \ldots, s.$$

A estatística $\chi^2$ para testar a hipótese nula será:

$$\lambda = 2 \cdot \sum_{i=1}^{s} \sum_{j=1}^{s} \sum_{k=1}^{s} n_{ijk} \ln \left( \frac{n_{ijk}}{n_{ij.}} \cdot \frac{n_i}{n_{ij}} \right). \tag{2.3.9}$$

com distribuição $\chi^2$ e $s(s-1)^2$ graus de liberdade. Entretanto, quando for apicada a uma cadeia de Markov de ordem $r$ a correspondente estatística $\chi^2$ terá $s^{(r-1)}(s-1)^2$ graus de liberdade.

## 2.4 Critérios de Informação

Se uma boa estimativa para a log-verossimilhança puder ser obtida através dos dados observados, esta estimativa poderá ser utilizada como um critério para comparar modelos. Assim, um modo de comparar $n$ modelos é simplesmente comparar os pesos da função maximizada $L(\hat{\theta}_i)$. Entretanto, tal método não fornece uma verdadeira comparação, haja vista que, em não conhecendo o verdadeiro modelo $g(x)$, primeiramente o método da máxima verossimilhança estima os parâmetros de cada modelo $g_i(x), i = 1,2,\ldots,n$, e posteriormente são utilizados os mesmos dados para estimar $E_g\left[\log f(x|\hat{\theta})\right]$, isto introduz um viés em $L(\hat{\theta}_i)$, sendo que, a magnitude deste viés varia de acordo com a dimensão do vetor de parâmetros (SABINO, 2019).

Um critério de informação é utilizado para verificar se um modelo estatístico aplicado a um determinado conjunto de dados é ajustado com aptidão. Os critérios comparam as funções de log-verossimilhança dos modelos, mas com uma "penalização" pela quantidade de parâmetros e como uma forma de correção do viés implementado na função (EMILIANO, 2009). Isso é feito para favorecer a escolha de um modelo mais parcimonioso.

Existem diversos critérios de informação; entretanto, foram considerados neste trabalho apenas dois: AIC e BIC. Em ambos os casos, a medida considerada no critério será calculada para cada modelo proposto, sendo considerado mais adequado aquele que tiver a menor medida.

### *2.4.1 Critérios de Informação de Akaike*

O valor do critério de informação de Akaike (AIC) (AKAIKE, 1974) é dado por:

$$\text{AIC} = 2 \cdot L(\hat{\theta}) + 2 \cdot p, \tag{2.4.1}$$

em que $p$ é o total de parâmetros do modelo e $L(\hat{\theta})$ é o máximo da função de verossimilhança, ou seja, a função de verossimilhança considerando que o estimador de máxima verossimilhança $\hat{\theta}$ é o parâmetro do modelo.

### *2.4.2 Critérios de Informação de Bayesiano*

O valor do critério de informação Bayesiano (BIC) (SCHWARZ, 1978) é dado por:

$$\text{BIC} = 2 \cdot L(\hat{\theta}) + p \cdot \ln(n), \tag{2.4.2}$$

sendo $L(\hat{\theta})$ e $p$ as mesmas quantidades consideradas para o AIC, e $n$ é o tamanho da amostra.

## 3 RESULTADOS E DISCUSSÃO

Os bancos de dados obtidos foram retirados do site *Infomoney* e fazem alusão à cotação da bolsa de valores (Ibovespa) durante o ano de 2021 e 2023, entre os dias $04/01/2021$ a $30/12/2021$ e $02/01/2023$ a $28/12/2023$. Eles são compostos por 247 observações (apenas dias úteis) e 7 variáveis, que são: data, abertura, fechamento, valor máximo, valor mínimo, volume e variação, sendo esta última fornecida em valor decimal. Para a análise dos dados, utilizou-se a primeira observação da variável variação, que fornece um valor aproximado de -0,14 e -3.06, respectivamente.

Os dados foram inicialmente importados e visualizados no R (R Core Team, 2020), seguidos pela criação de uma estrutura condicional para codificar a variável "variação" em 0 ou 1, utilizando a função *ifelse*. Após a binarização dos valores, foi aplicado o comando *table* para realizar a contagem dos eventos sucedidos. Em seguida, ocorreu a estimação dos parâmetros, com a criação dos estados e contagem dessas ocorrências. Para isso, se fez necessário utilizar os pacotes `dplyr` e `magrittr`, como também as funções *sapply*, *lag* e *lead*. Essas funções (*lag* e *lead*) foram aplicadas para realizar comparações entre os valores observados e seus respectivos valores anteriores e posteriores, permitindo a análise do comportamento da cadeia.

Considerou-se uma sequência de variáveis aleatórias $\{X_n, n \geq 0\}$, com

$$X_n = \begin{cases} 0, & \text{se a cotação da bolsa de valores caiu no } n\text{-ésimo dia} \\ 1, & \text{caso contrário.} \end{cases}$$

Para a estimação dos parâmetros, fez-se uma contagem da quantidade de zeros e uns na amostra, bem como a quantidade de vezes que aparece cada um dos pares $00, 01, 10, 11$. Como é necessário atribuir um ponto de referência para o desenvolvimento da análise, foi utilizado o que equivale ao dia 04/01/2021, portanto, esse valor não será atribuído na amostra. Logo, ela terá tamanho 246. É possível ver na Tabela 1 a quantidade de vezes que aparece cada um dos pares $00, 01, 10, 11$, bem como a contagem do número de zeros e uns.

Tabela 1 – Contagem do número de vezes que aparece cada par de estados na amostra de 2021

| | 0 | 1 | $n_i$ |
|---|---|---|---|
| 0 | 48 | 71 | 119 |
| 1 | 70 | 57 | 127 |
| | | | 246 |

Fonte: Elaborado pelos autores, 2024

Deste modo, considerando que a amostra seja uma cadeia de ordem 1, a estimativa da matriz de transição será

$$\hat{P} = \begin{bmatrix} 0,4 & 0,6 \\ 0,55 & 0,45 \end{bmatrix}$$

Na sequência, foram estimados também os parâmetros para uma cadeia de ordem 2, obtendo

$$\hat{P} = \begin{bmatrix} 0,42 & 0,58 \\ 0,57 & 0,43 \\ 0,40 & 0,60 \\ 0,53 & 0,47 \end{bmatrix}.$$

Tais valores foram calculados com base nos dados da Tabela 2, que apresenta a quantidade de vezes que cada um dos trios $000, 001, 010, 011, 100, 101, 110$ e $111$ aparece na amostra, assim como a quantidade de vezes que os pares aparecem seguidos de um estado qualquer.

Tabela 2 – Contagem do número de vezes que aparece cada trio de estados na amostra de 2021

| | 0 | 1 | $n_{ij}$ |
|---|---|---|---|
| 00 | 20 | 28 | 48 |
| 01 | 40 | 30 | 70 |
| 10 | 28 | 42 | 70 |
| 11 | 30 | 27 | 57 |

Fonte: Elaborado pelos autores, 2024

Na Tabela 3, expõe-se o número de vezes que cada uma das quadras $0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001, 1010, 1011, 1100, 1101, 1110$ e $1111$ aparecem, assim como a quantidade de vezes que os trios aparecem seguidos de um estado qualquer.

Tabela 3 – Contagem do número de vezes que aparece cada quadra de estados na amostra de 2021.

| | 0 | 1 | $n_{ijk}$ |
|---|---|---|---|
| 000 | 7 | 13 | 20 |
| 001 | 18 | 10 | 28 |
| 010 | 13 | 26 | 39 |
| 011 | 15 | 15 | 30 |
| 100 | 13 | 15 | 28 |
| 101 | 22 | 20 | 42 |
| 110 | 14 | 16 | 30 |
| 111 | 15 | 12 | 27 |

Fonte: Elaborado pelos autores, 2024

E daí, obtém-se a matriz de transição considerando uma cadeia de ordem 3:

$$\hat{P} = \begin{bmatrix} 0,35 & 0,65 \\ 0,64 & 0,36 \\ 0,33 & 0,67 \\ 0,50 & 0,50 \\ 0,47 & 0,53 \\ 0,52 & 0,48 \\ 0,47 & 0,53 \\ 0,56 & 0,44 \end{bmatrix}.$$

Em seguida, foi efetuado um teste para verificar se a amostra é independente ou segue uma cadeia de Markov de ordem 1. Esse teste é denominado Teste de Independência versus Dependência de Primeira Ordem e apresenta as seguintes hipóteses:

$H_0$ : O processo observado é independente
$H_1$ : O processo observado é uma cadeia de Markov de ordem 1

Neste caso, a estatística de teste segue distribuição $\chi^2$ e apresenta a seguinte fórmula:

$$\chi^2_{\text{Calculado}} = \sum_{i=1}^{S}\sum_{j=1}^{S} \frac{(n_{ij} - n_i n_j/n)^2}{n_i n_j/n} \sim \chi^2_{(s-1)^2}$$

Para a realização dos cálculos, foram usados os dados da Tabela 1:

$$\chi^2_{\text{Calculado}} = \frac{(48 - \frac{119\times118}{246})^2}{\frac{119\times118}{246}} + \frac{(71 - \frac{119\times128}{246})^2}{\frac{119\times128}{246}} + \frac{(70 - \frac{127\times118}{246})^2}{\frac{127\times118}{246}} + \frac{(57 - \frac{127\times128}{246})^2}{\frac{127\times128}{246}} = 5,38$$

Diante disso, a estatística de teste é $\chi^2 = 5,38$. Como a cadeia tem apenas dois estados, sob a hipótese $H_0$ (independência), a estatística de teste tem distribuição $\chi^2_{(1)}$ e p-valor aproximado de 0.020, indicando que a amostra segue uma cadeia de Markov de primeira ordem, uma vez que o p-valor < 0.05. Dessa forma, como a decisão do teste foi favorável à cadeia de ordem 1, um novo teste será efetuado para decidir entre ordem 1 e ordem 2.

Foi executado o Teste para Ordem de uma Cadeia de Markov, que apresenta as hipóteses:

$$H_0 : \text{O processo observado é uma cadeia de Markov de ordem 1}$$
$$H_1 : \text{O processo observado é uma cadeia de Markov de ordem 2}$$

e estatística de teste igual a

$$\lambda = 2 \cdot \sum_{i=1}^{s} \sum_{j=1}^{s} \sum_{k=1}^{s} n_{ijk} \ln \left( \frac{n_{ijk}}{n_{ij.}} \cdot \frac{n_i}{n_{ij}} \right) \sim \chi^2_{s^{(r-1)}(s-1)^2},$$

em que:

- $n_{ijk}$ = número de transições observadas para a cadeia de ordem 2,
- $n_{ij.} = \sum_k n_{ijk}$ é a soma das transições observadas no estado $ij$,
- $n_i$ = número de vezes que o estado $i$ foi visitado na cadeia de ordem 1,
- $n_{ij}$ = número de vezes que o estado $ij$ foi visitado na cadeia de ordem 2.

Substituindo os valores da Tabela 3 na estatística de teste tem-se $\chi^2_{cal} \approx 0,14$. Como a cadeia tem apenas dois estados, a estatística de teste tem distribuição $\chi^2_{(2)}$ e valor 0,14 ao nível de 5% de significância, indicando que a amostra é uma cadeia de Markov de primeira ordem. Desse modo, como a decisão do teste foi favorável novamente à cadeia de ordem 1, não será necessário efetuar um novo teste.

Na sequência, apresentam-se na Tabela 4, os valores dos critérios de informação AIC e BIC para averiguar, de outra perspectiva, qual modelo é mais adequado aos dados.

Tabela 4 – Valores de AIC e BIC supondo que os dados de 2021 são independentes, seguem uma cadeia de Markov de ordem 1 (CM 1), cadeia de Markov de ordem 2 (CM 2), cadeia de Markov de ordem 3 (CM 3)

| | Ind. | CM 1 | CM 2 | CM 3 |
|---|---|---|---|---|
| AIC | 346 | 343,2 | 350 | 362,5 |
| BIC | 353 | 357,2 | 378 | 418,7 |

Fonte: Elaborado pelos autores, 2024

Com isso, percebe-se que o modelo mais adequado é o de primeira ordem, visto que, a partir dos testes realizados, as hipóteses aceitas foram favoráveis à cadeia de ordem 1. Além disso, vimos que ele apresenta um menor valor para o critério de informação AIC se comparado ao modelo independente e menores valores de AIC e BIC em relação aos modelos de ordem 2 e 3. Portanto, será melhor ajustado aos dados.

O passo a passo anterior também foi executado para os dados de 2023. Como feito anteriormente, realizou-se a contagem de zeros, uns, pares, trios e quadras que aparecem na amostra. De modo similar, foi atribuído um ponto de referência para a análise, que equivale ao dia 02/01/2023. Logo, esse valor não fará parte da amostra, que contará apenas com 246 observações. Posteriormente, foram montadas as matrizes de transição para a primeira, segunda e terceira ordem. A quantidade de vezes que aparecem os estados, pares, trios e quadras pode ser vista na Tabela 5, Tabela 6 e Tabela 7.

Tabela 5 – Contagem do número de vezes que aparece cada par de estados na amostra de 2023

| | 0 | 1 | $n_i$ |
|---|---|---|---|
| 0 | 66 | 56 | 122 |
| 1 | 56 | 68 | 124 |
| | | | 246 |

Fonte: Elaborado pelos autores, 2024

Tabela 6 – Contagem do número de vezes que aparece cada trio de estados na amostra de 2023

| | 0 | 1 | $n_{ij}$ |
|---|---|---|---|
| 00 | 35 | 31 | 66 |
| 01 | 24 | 32 | 56 |
| 10 | 30 | 25 | 55 |
| 11 | 32 | 36 | 68 |

Fonte: Elaborado pelos autores, 2024

Tabela 7 – Contagem do número de vezes que aparece cada quadra de estados na amostra de 2023

| | 0 | 1 | $n_{ijk}$ |
|---|---|---|---|
| 000 | 20 | 15 | 35 |
| 001 | 14 | 16 | 30 |
| 010 | 14 | 10 | 24 |
| 011 | 18 | 14 | 32 |
| 100 | 15 | 15 | 30 |
| 101 | 10 | 15 | 25 |
| 110 | 17 | 15 | 32 |
| 111 | 14 | 22 | 36 |

Fonte: Elaborado pelos autores, 2024

Em seguida, foram realizadas as estimativas das probabilidades e a organização nas matrizes de transição. A matriz de transição estimada para a cadeia de ordem 1 é

$$\hat{P} = \begin{bmatrix} 0,54 & 0,46 \\ 0,45 & 0,55 \end{bmatrix},$$

para a cadeia de ordem 2 é

$$\hat{P} = \begin{bmatrix} 0,53 & 0,47 \\ 0,43 & 0,57 \\ 0,55 & 0,45 \\ 0,47 & 0,53 \end{bmatrix},$$

e para ordem 3 é

$$\hat{P} = \begin{bmatrix} 0,57 & 0,43 \\ 0,47 & 0,53 \\ 0,58 & 0,42 \\ 0,56 & 0,44 \\ 0,50 & 0,50 \\ 0,40 & 0,60 \\ 0,53 & 0,47 \\ 0,39 & 0,61 \end{bmatrix}.$$

Posteriormente, foi implementado o Teste de Independência versus Dependência de Primeira Ordem para verificar se a amostra é independente ou segue uma cadeia de Markov de ordem 1. Utilizando a estatística de teste para encontrar o $\chi^2_{\text{Cal}}$, obtém-se

$$\chi^2_{\text{Calculado}} = \frac{(66-\frac{122\times122}{246})^2}{\frac{122\times122}{246}} + \frac{(56-\frac{122\times124}{246})^2}{\frac{122\times124}{246}} + \frac{(56-\frac{124\times122}{246})^2}{\frac{124\times122}{246}} + \frac{(68-\frac{124\times124}{246})^2}{\frac{124\times124}{246}} = 1,97$$

Como a cadeia tem apenas dois estados, sob a hipótese $H_0$ (independência), a estatística de teste tem distribuição $\chi^2_{(1)}$, e é igual a 1,97. Com isso, o p-valor será aproximadamente 0.1604, indicando que a amostra é independente.

A próxima etapa será analizar os resultados dos critérios de informação e os valores calculados estão na Tabela 8.

Tabela 8 – Valores de AIC e BIC supondo que os dados de 2023 são independentes, seguem uma cadeia de Markov de ordem 1 (CM 1), cadeia de Markov de ordem 2 (CM 2), cadeia de Markov de ordem 3 (CM 3)

| | Ind. | CM 1 | CM 2 | CM 3 |
|---|---|---|---|---|
| AIC | 346.4 | 347.04 | 353.5 | 365.3 |
| BIC | 353.4 | 361.08 | 381.6 | 421.4 |

Fonte: Elaborado pelos autores, 2024

Verifica-se que, para o ano de 2023, o modelo mais adequado é o de independência, pois ele apresenta valores menores de AIC e BIC. Então, será melhor ajustado aos dados.

## 4 CONCLUSÃO

O presente trabalho teve como objetivo analisar o comportamento dos modelos Markovianos em relação a dois conjuntos de dados sobre as cotações da bolsa de valores durante os anos de 2021 e 2023. Neste projeto, foram exibidas probabilidades de transições organizadas em forma matricial, testes de hipóteses e critérios de informação.

Em relação ao ano de 2021, tanto os testes de hipóteses utilizados para verificar qual modelo apresenta melhor adequação aos dados analisados quanto o AIC indicaram que o modelo de ordem 1 era o mais adequado. Realizando a mesma análise para os dados referentes a 2023, pode-se relatar que os dados se adequam a um modelo independente, pois esse foi o resultado obtido tanto na realização do teste de hipótese quanto nos cálculos dos critérios de informação. Diante desse fato, vê-se que houve uma mudança no comportamento dos dados ao longo desses 2 anos. Isso pode ser explicado pelos acontecimentos sociais e políticos sucedidos no país nesse intervalo de tempo.

Posto isto, constata-se que inferir informações sobre cadeias de Markov tem uma importância significativa, uma vez que contribuem de forma efetiva em diversas áreas do conhecimento, como finanças, robótica, inteligência artificial e até ciências sociais. Para finalizar, é importante retratar sobre o conhecimento adquirido na execução deste trabalho, pois ele proporcionou um amadurecimento dos conceitos vistos na disciplina de processos estocásticos, bem como na elaboração do projeto de iniciação científica voltado a esse tema no período da graduação. Além disso, contribuiu para o interesse do estudo das teorias e cálculos matemáticas mais avançados, por meio do ingresso em um programa de mestrado em tempos futuros.

## REFERÊNCIAS

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, IEEE, v. 19, n. 6, p. 716–723, 1974. Citado na página 12.

BARBOSA, H. L. *Métodos Estatísticos em Cadeias de Markov*. Dissertação (Dissertação de Mestrado) — Universidade Federal do Rio Grande do Norte, Natal, RN, 2009. Programa de Pós-Graduação em Matemática Aplicada e Estatística. Citado na página 7.

BHAT, U. N.; MILLER, G. K. *Elements of Applied Stochastic Processes*. [S.l.]: Wiley Series in Probability and Statistics, 2002. Citado 4 vezes nas páginas 7, 9, 10 e 11.

EMILIANO, P. C. *Fundamentos e aplicações dos critérios de informação: Akaike e Bayesiano*. 2009. Trabalho Acadêmico. Disponível online. Citado na página 12.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2020. Disponível em: <http://www.R-project.org/>. Citado na página 13.

ROSS, S. M. *Stochastic Processes*. [S.l.]: John Wiley & Sons, 1996. Citado na página 7.

ROSS, S. M. *Introduction to Probability Models*. [S.l.]: Academic Press, 2007. Citado na página 8.

SABINO, A. L. C. *Uma aplicação de cadeias de Markov de alta ordem com dados de precipitação no Estado da Paraíba*. Campina Grande: [s.n.], 2019. Trabalho de Conclusão de Curso (Graduação em Estatística). Available at Universidade Estadual da Paraíba. Citado na página 12.

SANTOS, B. V. d. *Tópicos Avançados de Processos Markovianos e Aplicações*. 2017. Trabalho Acadêmico ou Livro. Citado na página 8.

SCHWARZ, G. Estimating the dimension of a model. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. ISSN 0090-5364. Disponível em: <http://www.jstor.org/stable/2958889>. Citado na página 12.

## AGRADECIMENTOS

Primeiramente, a Deus, pela oportunidade de ter chegado até aqui. A Ele também devo a força e a coragem de sair de uma cidade a 249 km de distância, deixando minha casa e família para buscar a realização de um sonho. Foi difícil, tive medo e muitas vezes pensei em desistir, mas Deus sempre me guiou e me deu impulso para continuar.

Aos meus pais, Aluísio e Lindalva, e meu irmão, Klaelson, por todo o apoio emocional e financeiro.

Aos meus tios, Lúcia e Araújo, e meus primos, Kaio e Kauê, por me acolherem em sua casa durante o meu primeiro ano.

A minha amiga de infância (16 anos de amizade), Erika, que mesmo distante sempre me incentivou nos dias difíceis.

A professora Dra. Divanilda Maia Esteves pelas orientações e correções na construção deste trabalho.

A banca examinadora, composta pela Prof. Me. Maria das Vitórias Alexandre Serafim e pelo Prof. Dr. Sílvio Fernando Alves Xavier Júnior, pela análise e contribuições feitas ao longo da defesa deste trabalho.

E, por fim, aos amigos que fiz durante o curso: Pedro, Vitória, Cauanny, Davi e Jéssica.