
**CLODOALDO TEODOSIO SANTANA DA SILVA**

# UTILIZAÇÃO DE MODELOS DE REGRESSÃO PARA DADOS CIRCULARES COM OU SEM A PRESENÇA DE CENSURA

**LAVRAS - MG**

**2023**

**CLODOALDO TEODOSIO SANTANA DA SILVA**

# UTILIZAÇÃO DE MODELOS DE REGRESSÃO PARA DADOS CIRCULARES COM OU SEM A PRESENÇA DE CENSURA

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária para a obtenção do título de Doutor.

Profa. DSc. Carla Regina Guimarães Brighenti
Orientadora

**LAVRAS - MG**
**2023**

**Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).**

Silva, Clodoaldo Teodosio Santana da.
Utilização de modelos de regressão para dados circulares com ou sem a presença de censura / Clodoaldo Teodosio Santana da Silva. - 2023.
156 p.

Orientador(a): Carla Regina Guimarães Brighenti.

Tese (doutorado) - Universidade Federal de Lavras, 2023.
Bibliografia.

1. Análise de Sobrevivência. 2. Covariável circular. 3. Acidentes de trânsito. I. Brighenti, Carla Regina Guimarães. II. Título.

**CLODOALDO TEODOSIO SANTANA DA SILVA**

# UTILIZAÇÃO DE MODELOS DE REGRESSÃO PARA DADOS CIRCULARES COM OU SEM A PRESENÇA DE CENSURA
# USE OF REGRESSION MODELS FOR CIRCULAR DATA WITH OR WITHOUT THE PRESENCE OF CENSORING

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária para a obtenção do título de Doutor.

APROVADA em 24 de fevereiro de 2023.

| | |
|---|---|
| Prof. Dsc. Lucas Monteiro Chaves | UFLA |
| Profa. DSc. Luciane Teixeira Passos Giarola | UFSJ |
| Prof. DSc. Michel Cândido de Souza | UFVJM |
| Prof. DSc. Rinaldo Artes | INSPER |

Profa. DSc. Carla Regina Guimarães Brighenti
Orientadora

**LAVRAS - MG**
**2023**

*À minha querida esposa Maria*

## AGRADECIMENTOS

Em primeiro lugar agradeço a Ele, meu Deus, que me deu forças para contiunar nesta jornada.

À minha esposa Maria pelo incentivo e compreensão durante estes anos em que estive tão envolvido com os estudos, meus sinceros muito obrigado.

Aos colegas da pós graduação em Estatística e Experimentação Agropecuária pelo apoio, incentivo e ajuda que vocês me deram, gostaria de citar nominalmente todos, mas tenho o receio de esquecer algum nome. Enfim, a todos vocês muito obrigado.

Agradeço a Universidade Federal dos Vales do Jequitinhonha e Mucuri-UFVJM pelo afastamento durante o doutorado, o que tornou possível que eu me dedicasse à pesquisa. Aproveito para demonstrar minha gratidão pelo incentivo dado pelos colegas do Departamento de Ciências Exatas da FACSAE-UFVJM.

Também sou grato aos professores do Programa de Pós graduação em Estatística e Experimentação Agropecuária pelos ensinamentos, vocês são fontes de inspiração e motivação para continuar buscando conhecimento.

Meu agradecimento ao Major Campos, Comandante da $6^a$ Cia PM Rv, do estado de Minas Gerais que nos disponibilizou os dados relativos aos acidentes nas rodovias, os quais foram utilizados no quarto artigo incluído nesta tese.

Por fim, sou grato a minha orientadora, a professora Carla Regina Guimarães Brighenti por ter aceito o desafio de me orientar. Muito obrigado pela sua paciência e profissionalismo ao conduzir este trabalho

# RESUMO

Dados circulares estão presentes em diversos fenômenos, na orientação do voo dos pássaros, direção dos ventos, além disso algumas medidas podem ser transformadas em angulares, como os horários, meses e dias da semana. Através da estatística circular é possível fornecer resultados mais adequados para tais situações. Os modelos de regressão adaptados para a estatística circular podem ser classificados como linear-circular, circular-linear e circular-circular, de acordo com a característica das variáveis explicativas ou resposta estarem situadas na reta real ou em uma circunferência de um círculo unitário. Neste trabalho, foi feito, inicialmente, uma revisão dos principais aspectos da estatística circular e seus modelos de regressão, realizando-se a utilização de planilhas provenientes de dados de coordenadas geográficas e meteorológicos. Por fim, fez-se uma proposta de utilização de um modelo de regressão cuja resposta é uma medida situada na reta real (linear) e a variável explicativa é uma medida angular (circular) no contexto da análise de sobrevivência, com o objetivo de se estudar o tempo para o início de atendimento às vítimas de acidentes em rodovias no estado de Minas Gerais. O Modelo log-normal foi utilizado para modelar a variável resposta, sendo a covariável circular com censura. O ajuste mostrou-se satisfatório, evidenciando que a estatística para dados circulares associada a análise de sobrevivência é uma excelente ferramenta para este tipo de estudo.

**Palavras-chave:** Análise de sobrevivência. Covariável circular. Coordenadas geográficas. Acidentes de trânsito.

# ABSTRACT

Circular data are present in several phenomena, in the orientation of birds' flight, wind direction, in addition some measures can be transformed into angular ones, such as times, months and days of the week. Through circular statistics it is possible to provide more adequate results for such situations. Regression models adapted for circular statistics can be classified as linear-circular, circular-linear and circular-circular, according to whether the explanatory or response variables are located on the real line or on a circumference of a unit circle. In this research, initially, a review of the main aspects of circular statistics and their regression models was carried out, using spreadsheets from geographic and meteorological coordinate data. Finally, a proposal was made for the use of a regression model whose response is a measure located on the real line (linear) and the explanatory variable is an angular measure (circular) in the context of survival analysis, with the aim of to study the time for the beginning of assistance to victims of accidents on highways in the state of Minas Gerais. The log-normal model was used to model the response variable, with the circular covariate being censored. The fit was satisfactory, showing that statistics for circular data associated with survival analysis is an excellent tool for this type of study.



**Keywords:** Survival analysis. Circular covariate. Geographic coordinates. Traffic accidents.

## LISTA DE FIGURAS

**SUMÁRIO**

# PRIMEIRA PARTE

# 1 INTRODUÇÃO

Dados circulares são aqueles medidos no círculo e seus valores dados em graus ou radianos. Os dados circulares são medidos na circunferência e os dados lineares na reta real. Eles são utilizados em diversas áreas como: zootecnia, medicina, astronomia, física, biologia, psicologia, entre outras.

A presença de periodicidade pode ser observada na direção dos voos dos pássaros, no comportamento de animais em resposta a alguns estímulos, na direção do ventos e das correntes marítimas e podem ser medidos em graus ou radianos. Dados referentes a dias do ano, meses, horários, também podem ser facilmente convertidos em medidas circulares por meio de transformação apropriada.

A análise dos dados circulares ou direcionais não é feita da mesma maneira em que são analisados os dados na reta real. Exemplificando, se os ângulos 5º, 0º, 355º, fossem tratados como lineares, poderia se chegar a conclusão que a média é 120º. Mas, neste caso, avaliando a distribuição dos pontos no círculo, nota-se que a média circular é 0º. Afim de que se obtenha um resultado mais confiável, dados que envolvem ângulos ou medidas direcionais deveriam ser analisados por meio da estatística circular.

Os dados circulares podem ser representados de maneiras diferentes, e podem estar relacionados com direção, neste caso a orientação utilizada será geográfica, e o zero estará no ponto em que representa o pólo Norte; como também referir-se a horários do dia, neste caso, cada hora do dia estará associada a uma medida angular que pode ser medida em graus ou radianos. Ou, se é diretamente um ângulo, pode ser situada no ciclo trigonométrico, utilizando o sentido horário ou anti-horário.

Assim como nos dados lineares, nos angulares têm-se as medidas de posição, dispersão e modelos próprios de distribuição de probabilidade, bem como adptações para modelos de regressão.

Medidas circulares podem ser analisadas utilizando técnicas estatísticas como por exemplo modelos de regressão. Porém os modelos de regressão são classificados de acordo com as variáveis envolvidas, sendo que, se ela é uma medida situada na reta real (será chamada linear) ou na circunferência (ou circular). Assim, os modelos de regressão serão chamados de linear-circular (a variável

resposta é linear e a explicativa circular), circular-linear (a resposta é circular e a explicativa linear) ou circular-circular (as variáveis resposta e explicativa são medidas circulares).

Os objetivos deste trabalho são fazer uma revisão de literatura sobre a área da estatística para dados circulares e sua utilização, dando ênfase para as aplicações na análise de sobrevivência e propor a utilização do modelo de regressão linear - circular na análise de sobrevivência. Neste modelo as variáveis explicativas são circulares e a resposta é a duração (ou tempo) para que os atendimentos às vítimas seja iniciado, portanto, um número real, ou seja, pertence ao conjunto dos números reais $\mathbb{R}$.

O trabalho foi organizado em duas partes capítulos. Na primeira há uma revisão bibliográfica sobre estatística circular: Algumas definições e modelos de probabilidade para o caso em que a variável aleatória é circular, como também algumas propriedades, e a análise de sobrevivência. Em seguida os modelos de regressão tanto para dados circulares e na análise de sobrevivência são estudados. Na segunda parte são apresentados quatro artigos desenvolvidos de acordo com a teoria abordada no capítulo 1. O primeiro artigo apresenta a estatística circular descritiva de um conjunto de dados envolvendo coordenadas geográficas, a partir da longitude e latitude. Foram obtidos os ângulos entre cidades para verificar a distribuição das cidades de origem dos alunos do Profmat (Mestrado profissional em matemática) e o local de estudo, a cidade de São João del Rei, já que foi avaliado o curso ministrado no campus Santo Antonio da Universidade Federal de São João del Rei (UFSJ).

Além da estatistica descritiva, a correlação e regressão circular foram abordadas no segundo artigo utilizando o banco de dados meteorológicos obtido no site do INMET (Instituto nacional de meteorologia) referente aos anos de 2020 e 2021.

No terceiro artigo a ênfase foi trabalhar o mesmo banco de dados do segundo artigo, utilizando no entanto, a teoria de análise de sobrevivência, através de uma extensão do modelo de regressão semi- paramétrico de Cox, denominado modelo AG. E finalmente no artigo quatro fez-se uma proposta de abordagem de utilização de modelo de regressão linear-circular na análise de sobrevivência. A proposta foi avaliar o tempo até o início do atendimento de ocorrências de acidentes de trânsito, sendo esta a variável resposta linear que foi relacionada ao horário do acidente estudado, como uma medida angular. Os registros incompletos ou com tempo superior a 12 horas

foram tratados como sendo censurados. O banco de dados refere-se ao atendimento de ocorrência de acidentes de trânsito, em que a variável de estudo foi o tempo até o atendimento da ocorrência.

# 2 REFERENCIAL TEÓRICO

Neste capítulo serão considerados alguns aspectos da estatística para dados circulares, incluindo a estatística circular descritiva, distribuição de probabilidades e modelos de regressão. Além disso, será feita uma revisão dos conceitos básicos da análise de sobrevivência.

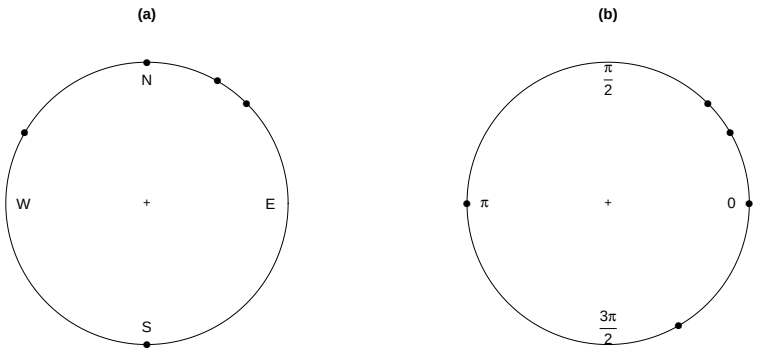
## 2.1 Estatística com dados circulares

Dados circulares são apropriados para a análise de informações cuja medida é angular. Estes dados estão presentes em informações relacionadas como por exemplo direção dos ventos e vôos dos passáros, ou medidas que são resultantes de dados lineares, provenientes da reta real, mas que pode ser convertidos em circulares, como horários, dias e meses do ano. Os dados podem ser medidos em graus ou em radianos.

Cada observação circular pode ser representada geometricamente como um ponto em círculo de raio um (círculo unitário). A depender dos dados, cada ângulo será medido no sentido horário ou anti-horário.

Na figura ( 2.1), os dados $0^\circ$, $30^\circ$, $45^\circ$, $180^\circ$ e $300^\circ$, estão representados no gráfico circular, usando o norte magnético como a origem e o $0^\circ$ no eixo das abscissas no sentido anti-horário.

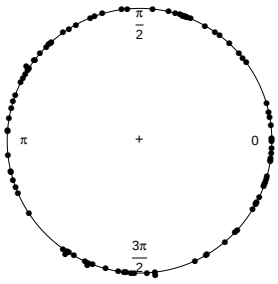
Figura 2.1 – (a) Pontos cardeais e (b) Sentido anti-horário




Fonte: O autor (2022)

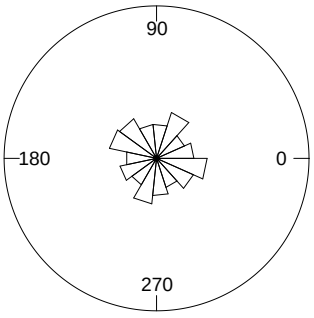
Assim como nas observações lineares, existem gráficos e diagramas para representar os dados circulares, entre eles o gráfico circular e o diagramas de rosas, conforme as figuras ( 2.2) e ( 2.3). Fazendo uma comparação com os dados lineares, o gráfico circular que é equivalente ao gráfico de pontos representados nos eixos coordenados e o diagrama de rosas que equivale ao histograma.

Figura 2.2 – Gráfico circular




Fonte: O autor (2022)

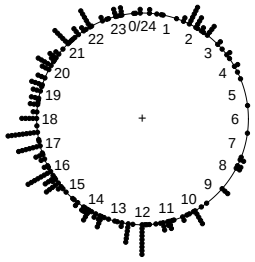
Figura 2.3 – Diagrama de rosas




Fonte: O autor (2022)

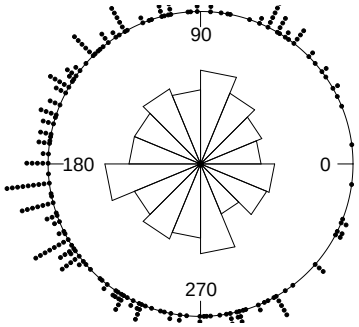
As figuras ( 2.4) e ( 2.5) mostram o gráfico circular e o diagrama de rosas de dados referente ao horário de chegada de 254 pacientes em uma unidade de terapia intensiva, num período de 12 meses (FISHER, 1993).

Figura 2.4 – Horário de chegada de 254 pacientes em uma unidade de terapia intensiva




Fonte: O autor (2022)

Figura 2.5 – Diagrama de rosas para o horário de chegada de 254 pacientes em uma unidade de terapia intensiva



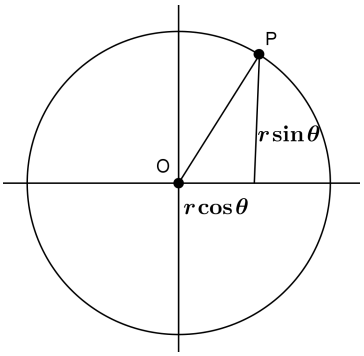

Fonte: O autor (2022)

Com isso em mente, pode-se afirmar que dado um ponto $P = (x, y)$ no plano cartesiano, este ponto pode também ser representado na forma trigonométrica, no sentido anti-horário, como

$x = r\cos\theta$ e $y = r\sin\theta$ , ou seja, $P$ também pode ser representado por $P = (r\cos\theta, r\sin\theta)$, figura ( 2.6).

No caso em que o circulo é unitário, ou seja, $r = 1$, o ponto $P$ será representado por $P = (\cos\theta, \sin\theta)$.

Figura 2.6 – Círculo




Fonte: O autor (2022)

Assim como se dá com as observações na reta real em que há medidas tendência central (média, mediana e moda) e medidas de dispersão, também existem estas medidas para estatística circular, porém os conceitos não são os mesmos utilizados para os dados que estão localizados na reta real ($\mathbb{R}$).

**Medidas de posição**

**Definição 1**- Suponha que sejam dados vetores unitários $\vec{v_1}, ..., \vec{v_n}$ com os correspondentes ângulos $\theta_1, ..., \theta_n$, respectivamente. A *média direcional* $\bar{\theta}$ de $\theta_1, ..., \theta_n$ é a direção da resultante de $\sum_{i=1}^{n} \vec{v_i}$, e ela tem a mesma direção que o centro de massa, $\bar{v}$, de $\vec{v_1}, ..., \vec{v_n}$ (MARDIA; JUPP, 2000).

Como cada $\vec{v_i}$, com $i = 1, ..., n$, pode ser representado por $(\cos\theta_i, \sin\theta_i)$, e o centro de massa $(\bar{C}, \bar{S})$, sendo

$$\begin{cases} \bar{C} = \dfrac{1}{n}\sum_{i=1}^{n}\cos\theta_i \\ \\ \bar{S} = \dfrac{1}{n}\sum_{i=1}^{n}\sin\theta_i \end{cases} \tag{2.1}$$

Assim a média direcional $\bar{\theta}$ é a solução das equações:

$$\begin{cases} \bar{C} = \bar{R}\cos\bar{\theta} \\ \\ \bar{S} = \bar{R}\sin\bar{\theta} \end{cases} \tag{2.2}$$

$R$ será chamado de comprimento resultante e $\bar{R}$ de comprimento médio resultante, o qual é dado por

$$\bar{R} = \sqrt{\bar{C}^2 + \bar{S}^2} \tag{2.3}$$

Por sua vez $R^2 = C^2 + S^2$ e $\bar{R} = \dfrac{R}{n}$.

Observe que $0 \leq \bar{R} \leq 1$. De fato, considerando $n$ vetores unitários $\vec{v_1}, ..., \vec{v_n}$.

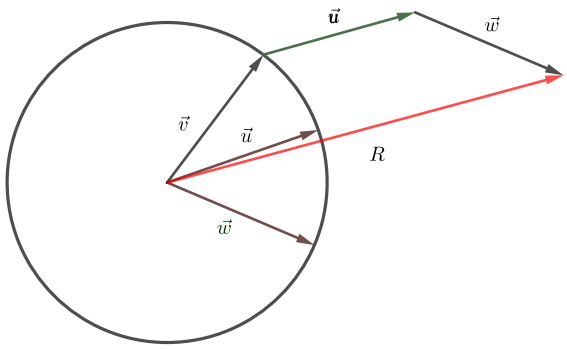
Tem - se que

$$0 \leq R = ||\vec{v_1} + ... + \vec{v_n}|| \leq ||\vec{v_1}|| + ... + ||\vec{v_n}|| = 1 + ... + 1 = n.1 = n$$

Logo, $0 \leq R \leq n$ e $0 \leq \dfrac{R}{n} = \bar{R} \leq 1$.

A figura ( 2.7) exemplifica como o comprimento resultante $R$ pode ser obtido para o caso em que se tem três vetores $\vec{u}$, $\vec{v}$, $\vec{w}$.

Figura 2.7 – Comprimento resultante $R$ para três vetores




Fonte: O autor (2022)

Considerando a média direcional como a solução da equação ( 2.2) tem-se :

$$\bar{\theta} = \begin{cases} \arctan\dfrac{\bar{S}}{\bar{C}}, & S > 0, C > 0 \\ \arctan\dfrac{\bar{S}}{\bar{C}} + \pi, & C < 0 \\ \arctan\dfrac{\bar{S}}{\bar{C}} + 2\pi, & S < 0, C > 0 \end{cases} \tag{2.4}$$

Como $\bar{S} = \dfrac{S}{n}$ e $\bar{C} = \dfrac{C}{n}$, a média direcional pode também ser dada por:

$$\bar{\theta} = \begin{cases} \arctan\dfrac{S}{C}, & S > 0, C > 0 \\ \arctan\dfrac{S}{C} + \pi, & C < 0 \\ \arctan\dfrac{S}{C} + 2\pi, & S < 0, C > 0 \end{cases} \tag{2.5}$$

A média direcional não está definida quando $S = C = 0$.

**Definição 2**- A *mediana direcional* dos ângulos $\theta_1, ..., \theta_n$ é qualquer ângulo $\phi$ tal que:

a) Metade dos pontos situam-se no arco $[\phi, \phi + \pi)$;

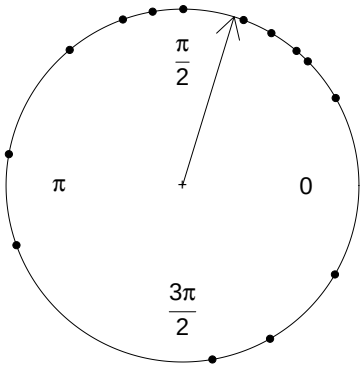
b) A maioria dos pontos estão mais próximos de $\phi$ do que de $\phi + \pi$.

Quando o tamanho da amostra, $n$, é ímpar a mediana é um único ângulo, e quando $n$ é par a mediana será a média aritmética entre os dois ângulos centrais e adjacentes. Exemplificando: Considere o conjunto de dados

$$30^\circ, 45^\circ, 50^\circ, 60^\circ, 70^\circ, 90^\circ, 100^\circ, 110^\circ, 130^\circ, 170^\circ, 200^\circ, 280^\circ, 300^\circ, 330^\circ$$

A média circular é $1,27$ radianos, que corresponde a $73^\circ$, figura ( 2.8), e o comprimento resultante é $\bar{R} = 0,37$.

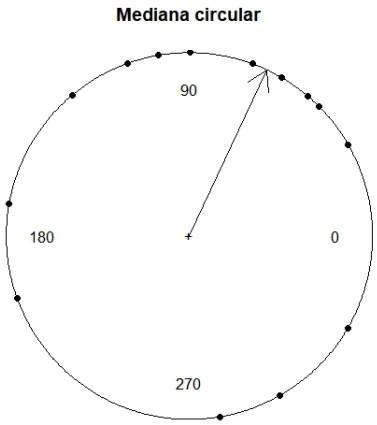
Figura 2.8 – Média circular




Fonte: O autor (2022)

Em virtude da quantidade par de dados a mediana direcional (circular) é a média aritmética entre os ângulos de $60^\circ$ e $70^\circ$, ou $1,047$ radianos e $1,220$ radianos, respectivamente. Assim a mediana é $65^\circ$ ($1,134$ radianos). Note que o valor da mediana é diferente do encontrado para o caso em que os dados são medidas lineares (não circulares). A representação gráfica da mediana está na figura ( 2.9).

Figura 2.9 – Mediana direcional




Fonte: O autor (2022)

De acordo com Mardia (1972) a mediana direcional $\phi$ também pode ser definida como a solução da equação:

$$\int_{\phi}^{\phi+\pi} f(\theta)d\theta = \frac{1}{2} \tag{2.6}$$

em que $f(\theta)$ é a função densidade de probabilidade de $\theta$.

Assim como nos dados lineares a mediana é o segundo quartil $Q_2$, logo pode ser definido os quartis $Q_1$ e $Q_3$, respectivamente, o primeiro e terceiro quartis, como as soluções das equações ( 2.7) e ( 2.8).

$$\int_{\phi-Q_1}^{\phi} f(\theta)d\theta = \frac{1}{4} \tag{2.7}$$

$$\int_{\phi}^{\phi+Q_3} f(\theta)d\theta = \frac{1}{4} \tag{2.8}$$

**Medidas de dispersão**

Se $\bar{R}$ tem um valor próximo de 1, os dados direcionais estão bem agrupados. Caso $\bar{R}$ esteja próximo de zero, ocorre uma forte dispersão dos dados. Notando assim, que $\bar{R}$ é uma medida de

concentração dos dados (MARDIA, 1972). A partir deste fato será dado o conceito da variância circular.

**Definição 3**- A **variância circular** é dada por

$$V = 1 - \bar{R} \tag{2.9}$$

com $0 \leq V \leq 1$.

Quanto mais próximo de 0 for o valor da variância circular mais concentrados estarão os dados.

O desvio padrão circular de uma amostra é definido pela expressão

$$v = \{-2\log(1-V)\}^{\frac{1}{2}} \tag{2.10}$$

Detalhes sobre a equação ( 2.10) pode ser obtida em Mardia (1972) capítulo 3 (§ 3.4.8d). Também pode ser utilizada uma aproximação para o desvio padrão circular,

$$S \approx (2V)^{\frac{1}{2}} \tag{2.11}$$

(JAMMALAMADAKA; SENGUPTA, 2001)

De fato, usando e expansão de Taylor para $\cos\theta$, com apenas dois termos, tem-se

$$\cos\theta \approx 1 - \frac{\theta^2}{2} \tag{2.12}$$

Então,

$$2\cos\theta \approx 2 - \theta^2$$

$$\theta^2 \approx 2(1-\cos\theta)$$

$$\sum_{i=1}^{n} \theta_i^2 \approx 2\sum_{i=1}^{n}(1-\cos\theta_i)$$

Com isso

$$\sum_{i=1}^{n}(\theta_i - \bar{\theta})^2 \approx 2\sum_{i=1}^{n}(1-\cos(\theta_i - \bar{\theta})) \tag{2.13}$$

Antes de desenvolver a equação ( 2.13) há a necessidade do Teorema 2.1.1.

**Teorema 2.1.1** *Se $\bar{\theta}$ é a direção circular média dos vetores $\theta_1, ..., \theta_n$, então*

$$\sum_{i=1}^{n} \sin(\theta_i - \bar{\theta}) = 0$$

$$\sum_{i=1}^{n} \cos(\theta_i - \bar{\theta}) = R$$

[**Demonstração**]

Usando relações trigonométricas e a equação ( 2.2)

$$\sum_{i=1}^{n} \sin(\theta_i - \bar{\theta}) = \cos\bar{\theta} \sum_{i=1}^{n} \sin\theta_i - \sin\bar{\theta} \sum_{i=1}^{n} \cos\theta_i$$

$$= S.\frac{C}{R} - C.\frac{S}{R} = 0$$

e

$$\sum_{i=1}^{n} \cos(\theta_i - \bar{\theta}) = \sum_{i=1}^{n} (\cos\theta_i \bar{\cos}\,\theta + \sin\theta_i \sin\bar{\theta})$$

$$= \cos\bar{\theta} \sum_{i=1}^{n} \cos\theta_i + \sin\bar{\theta} \sum_{i=1}^{n} \sin\theta_i$$

$$= C\cos\bar{\theta} + S\sin\bar{\theta} = \frac{C^2}{R} + \frac{S^2}{R} = R$$

Retornando à equação ( 2.13) tem-se:

$$\sum_{i=1}^{n} (\theta_i - \bar{\theta})^2 \approx 2 \sum_{i=1}^{n} (1 - \cos(\theta_i - \bar{\theta}))$$

$$= 2n - 2 \sum_{i=1}^{n} \cos(\theta_i - \bar{\theta})$$

$$= 2(n - R)$$

$$= 2(n - n\bar{R})$$

$$2n(1-\bar{R}) \approx \sum_{i=1}^{n}(\theta_i - \bar{\theta})^2$$

$$2(1-\bar{R}) \approx \frac{\sum_{i=1}^{n}(\theta_i - \bar{\theta})^2}{n}$$

$$S^2 \approx 2(1-\bar{R})$$

$$S^2 \approx 2V$$

Portanto

$$S \approx \sqrt{2V} = \sqrt{2(1-\bar{R})} \tag{2.14}$$

Seja $P_i$ um ponto no círculo unitário correspondente ao ângulo $\theta_i$, e $\alpha$ uma direção fixa, a distância circular $D$ entre $P$ e $P_i$ será a menor entre os dois ângulos formados por $\bar{OP}$ e $\bar{OP_i}$ então ela pode ser representada por:

$$D = min\{\alpha - \theta, 2\pi - (\alpha - \theta)\}$$

então $D = \alpha - \theta = \pi - (\pi - (\alpha - \theta))$ ou $D = 2\pi - (\alpha - \theta) = \pi + (\pi - (\alpha - \theta))$.

Portanto,

$$D = min\{\alpha - \theta, 2\pi - (\alpha - \theta)\} = \pi - |\pi - |\alpha - \theta|| \tag{2.15}$$

Jammalamadaka (2001) também define a distância circular, $\Delta$, entre dois ângulos centrais $\alpha$ e $\theta$ como:

$$\Delta = 1 - \cos(\alpha - \theta) \tag{2.16}$$

**Momentos Trigonométricos**

Características das variáveis aletórias podem ser analisadas por meio das potências do valor esperado (MAGALHÃES, 2006).

Já foi visto que $\bar{C} = \frac{1}{n}\sum_{i=1}^{n}\cos\theta_i$, $\bar{S} = \frac{1}{n}\sum_{i=1}^{n}\sin\theta_i$, com isso

$$\bar{C} + i\bar{S} = \bar{R}(\cos\bar{\theta} + i\sin\bar{\theta}) = \bar{R}e^{i\bar{\theta}}$$

Assim chamaremos de $m_1$, o momento trigonométrico de ordem 1, que será dado por

$$m_1 = \bar{R}e^{i\bar{\theta}}$$

Generalizando tem-se que o momento trigonométrico de ordem $p$, como $p = 1, 2, ...$ e $m_p = \bar{C}_p + i\bar{S}_p$, onde

$$\bar{C}_p = \frac{1}{n}\sum_{i=1}^{n} \cos p\theta_i$$

e

$$\bar{S}_p = \frac{1}{n}\sum_{i=1}^{n} \sin p\theta_i$$

Com isso o modelo de ordem $p$ será dado por

$$m_p = \bar{R}_p e^{i\bar{\theta}_i}$$

O p-ésimo momento trigonométrico centrado é dado por $m_p = C_p + iS_p$, onde

$$C_p = \frac{1}{n}\sum_{i=1}^{n} \cos p(\theta_i - \bar{\theta})$$

e

$$S_p = \frac{1}{n}\sum_{i=1}^{n} \sin p(\theta_i - \bar{\theta})$$

Como $\sum_{i=1}^{n} \sin(\theta_i - \bar{\theta}) = 0$, tem-se que $m_1 = \bar{R}$ pois

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} \cos(\theta_i - \bar{\theta}) = \frac{R}{n} = \bar{R} \tag{2.17}$$

## 2.2 Distribuições circulares de probabilidade

Modelos de probabilidade são importantes para a análise dos dados, pois é possível ajustar mais adequadamente os parâmetros aos dados.

Se $f(\theta)$ é uma funcão densidade de probabilidade de uma variável aleatória circular $\Theta$, então esta função satisfaz as seguintes propriedades:

a) $f(\theta) \geq 0$;

b) $f(\theta + 2\pi) = f(\theta)$;

c) $\int_0^{2\pi} f(\theta) d\theta = 1$.

O item b) mostra que a função densidade de probabilidade é periódica. Existem várias distribuições circulares usadas para a análise dos dados circulares. Serão apresentados resultados como o estimador de máxima verossimilhança e log-verossimilhança para alguns destes modelos.

O estimador de máxima verossimilhança é muito utilizado na estatística com a finalidade de se obter entre as estimativas do parâmetro desconhecido, "aquele que maximiza a probabilidade de obter a amostra particular observada, ou seja, o valor que torna aquela amostra mais provável"(BUSSAB; MORRETIN, 2010).

A ideia do método consiste em obter uma função de densidade de todas as observações, levando em consideração que os dados amostrais são independentes, então

$$f(x_1, ..., x_n|y) = f(x_1|y)....f(x_n|y)$$

onde $f(x)$ é a função densidade de probabilidade. Considerando agora que $x_1, ..., x_n$ são fixos e que $y$ pode variar, a função de verossimilhança será dada por:

$$L(y|x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i|y) \tag{2.18}$$

Em muitas situações pode ser usado o $\log L(y|x_1, ..., x_n)$, que é chamado de log-verossimilhança e é representado por

$$l(y|x_1, ..., x_n)$$

Assim a log-verossimilhança será dada por

$$l(y|x_1, ..., x_n) = \sum_{i=1}^{n} \log f(x_i|y)$$

O estimador $\hat{y}$ que maximiza a função de verossimilhança será dado pela solução da equação

$$\frac{\partial l(y|x_1,...x_n)}{\partial y} = 0$$

A seguir são exibidas algumas distribuições circulares, como a distribuição Uniforme, Cardióide, Wrapped e von Mises, bem como algumas propriedades das distribuições von Mises, Wrapped e Cardióide.

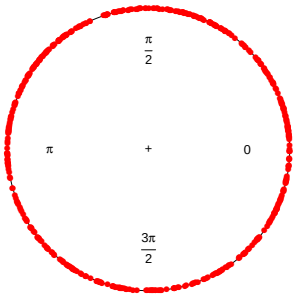
**Distribuição circular uniforme**

Se os dados estão distribuídos uniformemente sobre a circunferência, têm-se a distribuição circular uniforme, que é representada por $U_c$. A função densidade de probabilidade é dada por

$$f(\theta) = \frac{1}{2\pi}, 0 \leq \theta < 2\pi$$

Foi obtido, por simulação, uma amostra com 500 dados seguindo a distribuição circular uniforme, figura (2.10). Note que os dados estão distribuídos uniformemente ao redor do círculo.

Figura 2.10 – Distribuição circular uniforme




Fonte: O autor (2022)

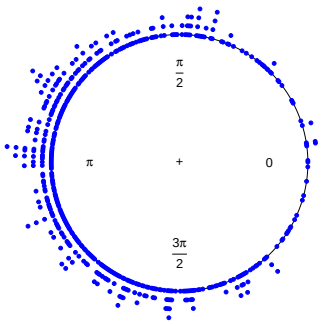
**Distribuição Cardióide**

A Distribuição Cardióide é representada por $C(\mu,\rho)$ , e sua função densidade de probabilidade é dada por

$$f(\theta) = \frac{1}{2\pi}(1+2\rho\cos(\theta-\mu)) \tag{2.19}$$

sendo $0 \leq \theta < 2\pi$, $0 \leq \rho \leq \frac{1}{2}$. A distribuição é simétrica e unimodal, o parâmetro $\rho$ é o comprimento médio resultante (ou parâmetro de concentração) e $\mu$ é a média direcional. Observe que se $\rho = 0$ tem-se a distribuição uniforme. A figura ( 2.11) mostra o gráfico circular de 500 dados, obtidos por simulação, que seguem a distribuição cardióide. Foi considerada a média circular $\mu = \pi$ rad e $\rho = 0,3$.

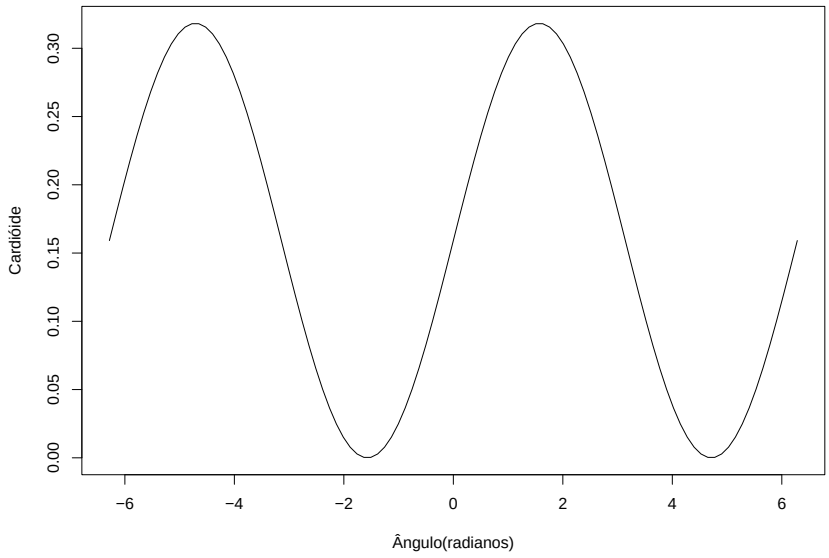
Figura 2.11 – Distribuição cardióide




Fonte: O autor (2022)

A figura ( 2.12) mostra o gráfico da funcão densidade da Cardióide para o caso em que $\mu = \frac{\pi}{2} rad$ e $\rho = \frac{1}{2}$.

Figura 2.12 – Gráfico da função densidade da distribuição cardióide




Fonte: O autor (2022)

A função de verossimilhança será dada por

$$L(\mu,\rho|\theta_1,...,\theta_n) = (\frac{1}{2\pi})^n \prod_{i=1}^{n}\{1+2\rho\cos(\theta_i-\mu)\}$$

A sua função de log-verossimilhança será:

$$l(\mu,\rho|\theta_1,...,\theta_n) = -n\log 2\pi + \sum_{i=1}^{n}\log\{1+2\rho\cos(\theta_i-\mu)\}$$

Os estimadores $\hat{\mu}$ e $\hat{\rho}$ são as soluções, respectivamente, de $\frac{\partial l}{\partial \mu} = 0$ e $\frac{\partial l}{\partial \rho} = 0$.

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{n}\frac{2\rho\sin(\theta_i-\mu)}{1+2\rho\cos(\theta_i-\mu)} \tag{2.20}$$

$$\frac{\partial l}{\partial \rho} = \sum_{i=1}^{n}\frac{2\cos(\theta_i-\mu)}{1+2\rho\cos(\theta_i-\mu)} \tag{2.21}$$

**Distribuição Wrapped (Arqueada)**

Dada uma distribuição na reta real $\mathbb{R}$, ela pode ser arqueada ao redor da circunferência do circulo unitário (MARDIA, 1972). Então, se $X$ é uma variável aleatória definida na reta real, a correspondente variável aleatória circular $X_W$ pode ser obtida definindo:

$$X_W \equiv X mod 2\pi$$

logo dado $0 \leq \theta \leq 2\pi$

$$X_W(\theta) = \{X(\theta + 2k\pi), k \in \mathbb{Z}\}$$

pois se $x_w \equiv x mod 2\pi$ então $x_w = x + 2k\pi, k \in \mathbb{Z}$

Se $X$ tem função de distribuição $F$, então a função de distribuição de $X_W$ é representada por $F_W$.

e

$$F_W(\theta) = \sum_{k=-\infty}^{\infty} \{F(\theta + 2k\pi) - F(2k\pi)\}, k \in \mathbb{Z}, 0 \leq \theta \leq 2\pi \tag{2.22}$$

Se $f(x)$ é a função densidade de probabilidade de uma variável aleatória $X$, então para a variável aleatória circular $X_W$ a função densidade de probabilidade $f_W$ será definida por (RAO; GIRIJA, 2020; MARDIA, 1972):

$$f_W(\theta) = \sum_{k=-\infty}^{+\infty} f(\theta + 2k\pi) \tag{2.23}$$

com $k \in \mathbb{Z}$ e $0 \leq \theta \leq 2\pi$

A distribuição Wrapped $X_w$ satisfaz as seguintes propriedades (MARDIA; JUPP, 2000):

*Propriedade 2.1*

a) $(X + Y)_w = X_w + Y_w$;

b) Se $\phi$ é a função característica de $X$ então a função característica de $X_w$ é dada por $\phi(p)$ com $p = 0, \pm 1, \pm 2...$;

c) Se $\phi$ é integrável então $X$ tem função densidade e

$$f_w(\theta) = \sum_{k=-\infty}^{\infty} f(\theta + 2k\pi) = \frac{1}{2\pi}\left[1 + 2\sum_{p=1}^{+\infty}(\alpha_p \cos p\theta + \beta_p \sin p\theta)\right] \tag{2.24}$$

e $\phi(p) = \alpha_p + i\beta_p$.

Existem vários tipos de distribuições Wrapped (arqueadas) como : Wrapped Normal, Wrapped Poisson, Wrapped Cauchy e outras (MARDIA; JUPP, 2000). Neste trabalho será abordado apenas a Wrapped Cauchy.

**Distribuição Wrapped Cauchy**

Esta distribuição é simétrica, unimodal e é obtida arqueando a função densidade de probabilidade Cauchy, definida sobre a reta real ($\mathbb{R}$), dada por

$$f(x) = \frac{1}{\pi} \quad \frac{a}{a^2 + (x - \mu^2)}, \quad -\infty < x < +\infty, \quad a > 0$$

A partir daí é possível escrever a função densidade de probabilidade da distribuição Wrapped Cauchy dada por:

$$g(\theta) = \frac{1}{2\pi} \quad \frac{1 - \rho^2}{1 + \rho^2 - 2\rho\cos(\theta - \mu)}, \quad 0 \leq \theta \leq 2\pi, \quad 0 \leq \ \rho \leq 1 \tag{2.25}$$

De fato, considerando que a função característica da distribuição Cauchy é dada por $\phi(t) = e^{-a|t| + it\mu}$ com $t \in \mathbb{R}$, a função característica da Wrapped Cauchy, de acordo com o item $b$) da propriedade (2.1), (MARDIA, 1972; ALLAHHAM, 2015) será dada por:

$$\phi(p) = e^{-a|p| + ip\mu}$$

Com isso, tem-se que

$$\phi(p) = e^{-a|p|}.e^{ip\mu} = e^{-a|p|}.(cosp\mu + i\sin p\mu)$$

Então

$$\phi(p) = \alpha_p + i\beta_p$$

sendo

$$\alpha_p = e^{-a|p|}\cos p\mu$$

$$\beta_p = e^{-a|p|}\sin p\mu$$

Do item $c)$ da propriedade (2.1) obtem-se

$$g(\theta) = \sum_{k=-\infty}^{+\infty} f(\theta + 2k\pi) = \frac{1}{2\pi}\left[1 + 2\sum_{p=1}^{\infty}(\alpha_p \cos p\theta + i\sin p\theta)\right]$$

Então

$$g(\theta) = \frac{1}{2\pi}\left[1 + 2\sum_{p=1}^{\infty}\left[e^{-a|p|}(\cos p\theta\cos p\mu + \sin p\theta\sin p\mu)\right]\right]$$

$$g(\theta) = \frac{1}{2\pi}[1 + 2\sum_{p=1}^{\infty}(e^{-a|p|}\cos p(\theta - \mu))] \tag{2.26}$$

Fazendo $\rho = e^{-a}$ e substituindo em ( 2.26):

$$g(\theta) = \frac{1}{2\pi}(1 + 2\sum_{p=1}^{\infty}(\rho^{|p|}\cos p(\theta - \mu)) \tag{2.27}$$

Para simplificar a notação considere

$$y = \rho e^{-i(\theta-\mu)}$$

então

$$y^p = \rho^p e^{-ip(\theta-\mu)}$$

$$y^p = \rho^p \cos p(\theta - \mu) - i\rho^p \sin p(\theta - \mu) \tag{2.28}$$

como

$$|y| = |\rho e^{-i(\theta-\mu)}| = |\rho||e^{-i(\theta-\mu)}| = |\rho|.1 = |e^{-a}|$$

e $a > 0$, então $0 < |y| = |\rho| < 1$.

Com isso pode se notar que a série geométrica $\sum_{p=1}^{\infty} y^p$ é convergente, pois $|y| < 1$

$$\sum_{p=1}^{\infty} y^p = \frac{y}{1-y} = \frac{\rho e^{-i(\theta-\mu)}}{1-\rho e^{-i(\theta-\mu)}}$$

$$= \sum_{p=1}^{\infty} \frac{\rho^{-i(\theta-\mu)}}{1-\rho^{-i(\theta-\mu)}} \frac{1-\rho e^{i(\theta-\mu)}}{1-\rho e^{i(\theta-\mu)}}$$

$$= \frac{\rho e^{-i(\theta-\mu)} - \rho^2}{1-\rho e^{i(\theta-\mu)} - \rho e^{-i(\theta-\mu)} + \rho^2}$$

$$= \frac{\rho\cos(\theta-\mu) - \rho^2}{1-2\rho\cos(-\mu)+\rho^2} - \quad \frac{\sin(\theta-\mu)}{1-2\rho\cos(\theta-\mu)+\rho^2} i$$

Considerando apenas a parte real da série obtém-se

$$\frac{\rho\cos(\theta-\mu) - \rho^2}{1-2\rho cos(\theta-\mu)+\rho^2} \tag{2.29}$$

Das equações ( 2.27), ( 2.28) e ( 2.29) chega-se a conclusão que

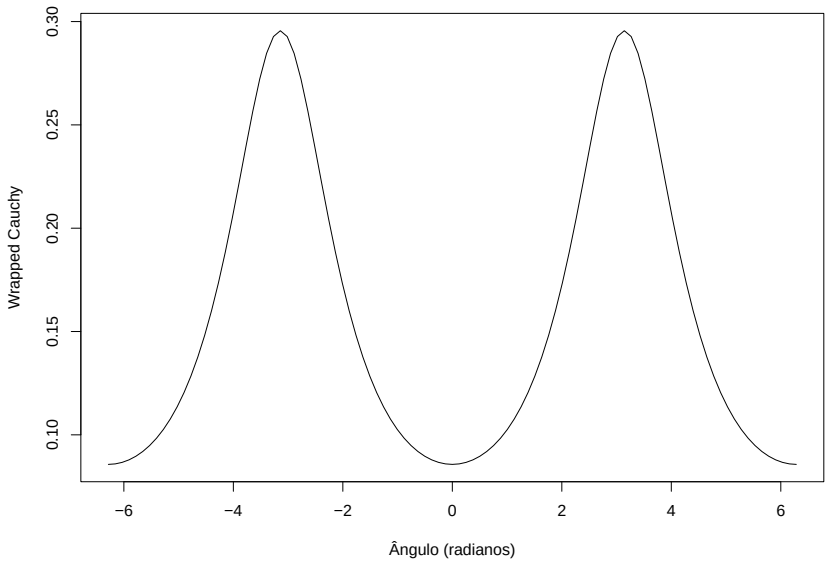
$$g(\theta) = \frac{1}{2\pi}[1 + 2\frac{\rho\cos(\theta-\mu) - \rho^2}{1-2\rho cos(\theta-\mu)+\rho^2}]$$

Portanto

$$g(\theta) = \frac{1}{2\pi} \quad \frac{1-\rho^2}{1+\rho^2-2\rho\cos(\theta-\mu)} \tag{2.30}$$

Que é a função densidade de probabilidade da distribuição Wrapped Cauchy. Os parâmetros deste modelo são $\mu$ e $\rho$. Na figura ( 2.14) temos um exemplo de dados , simulados, com a distribuição Wrapped Cauchy quando $\mu = \pi rad$ e $\rho = 0,3$.

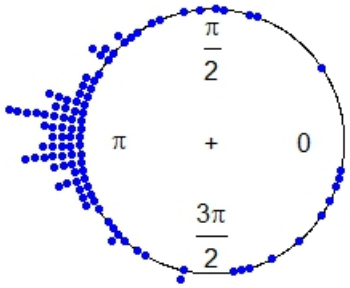
Figura 2.13 – Gráfico da função densidade de probabilidade da Wrapped Cauchy




Fonte: O autor (2022)

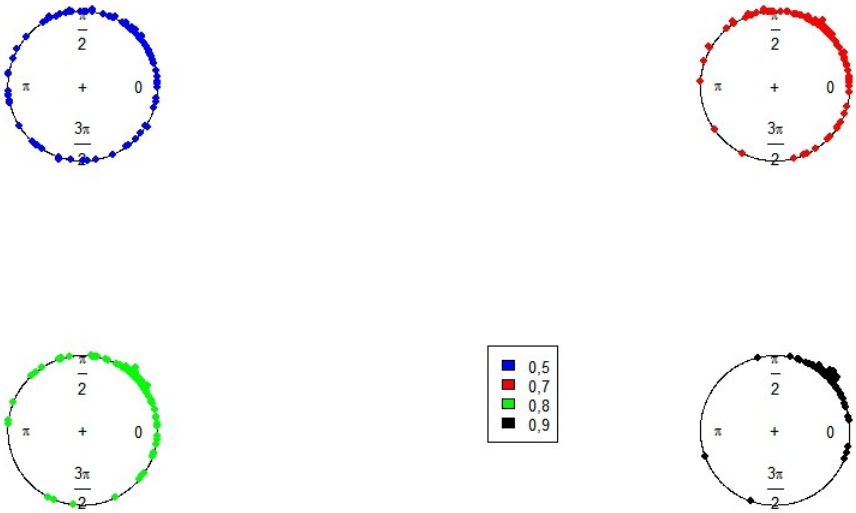
Nesta distrubuição quando o valor parâmetro $\rho$ se aproxima de 0 (zero), a distribuição converge para a distribuição uniforme e quando O valor de $\rho$ se aproxima de 1 (um), os pontos da distribuição se concentram próximos do valor da média circular $\mu$. Isto está ilustrado na figura ( 2.15), onde os dados foram simulados, e mantido fixo o valor da média circular $\mu = \frac{\pi}{4} rad$, e variou-se os valores do parâmetro $\rho$, sendo utilizado os seguintes valores $\rho = 0,5; 0,7; 0,8; 0,9$.

Figura 2.14 – Distribuição Wrapped Cauchy




Fonte: O autor (2022)

Figura 2.15 – Gráfico circular da Wrapped Cauchy, com $\mu = \frac{\pi}{4} rad$ e $\rho = 0,5;0,7;0,8;0,9$




Fonte: O autor (2022)

**Inferência da Distribuição Wrapped Cauchy**

Considerando que

$$g(\theta;\mu,\rho) = \frac{1}{2\pi}.\frac{1-\rho^2}{1+\rho^2-2\rho\cos(\theta-\mu)}$$

com $\leq \mu \leq 2\pi$ e $0 \leq \rho \leq 1$ tem-se que

$$g(\theta;\mu,\rho) = \frac{1}{2\pi} \quad \frac{1-\rho^2}{1+\rho^2-2\rho(\cos\theta\cos\mu+\sin\theta\sin\mu)}$$

tem-se que

$$g(\theta;\mu,\rho) = \frac{1-\rho^2}{2\pi\left(1-\frac{2\rho\cos\theta\cos\mu}{1+\rho^2}-\frac{2\rho\sin\theta\sin\mu}{1+\rho^2}\right)(1+\rho^2)}$$

fazendo $\mu_1 = \frac{2\rho\cos\mu}{1+\rho^2}$ e $\mu_2 = \frac{2\rho\sin\mu}{1+\rho^2}$, será obtida uma nova parametrização para a função $f(\theta;\mu,\rho)$ (ALLAHHAM, 2015) :

$$g(\theta;\mu_1,\mu_2) = \frac{1}{2\pi}\frac{1-\rho^2}{1+\rho^2} \quad \frac{1}{(1-\mu_1\cos\theta-\mu_2\sin\theta)} \tag{2.31}$$

Observe que

$$\frac{1-\rho^2}{1+\rho^2} = 1-\frac{2\rho^2}{1+\rho^2} \tag{2.32}$$

mas $0 \leq \rho \leq 1$ e

$$0 \leq \rho^2 \leq 1 \tag{2.33}$$

De 2.33 obtem-se

$$0 \leq 2\rho^2 \leq 1+\rho^2$$

daí

$$0 \leq \frac{2\rho^2}{1+\rho^2} \leq 1$$

com isso nota-se que

$$1-\frac{2\rho^2}{1+\rho^2}\geq 0$$

e a equação ( 2.32) pode ser escrita da seguinte maneira:

$$\frac{1-\rho^2}{1+\rho^2}=\sqrt{\left(1-\frac{2\rho^2}{1+\rho^2}\right)^2}$$

Logo

$$\frac{1-\rho^2}{1+\rho^2}=\sqrt{1-\frac{4\rho^2}{1+\rho^2}+\frac{4\rho^4}{(1+\rho^2)^2}}$$

e

$$\frac{1-\rho^2}{1+\rho^2}=\sqrt{1-\frac{4\rho^2(1+\rho^2)+4\rho^4}{(1+\rho^2)^2}}$$

Portanto

$$\frac{1-\rho^2}{1+\rho^2}=\sqrt{1-\frac{4\rho^2}{(1+\rho^2)^2}} \tag{2.34}$$

Mas tem-se que

$$\mu_1^2+\mu_2^2=\frac{4\rho^2\cos^{\mu}}{(1+\rho^2)^2}+\frac{4\rho^2\sin^2\mu}{(1+\rho^2)^2}$$

então

$$\mu_1^2+\mu_2^2=\frac{4\rho^2}{(1+\rho^2)^2} \tag{2.35}$$

Das equações ( 2.34) e ( 2.35), pode ser obtida a expressão

$$\frac{1-\rho^2}{1+\rho^2}=\sqrt{1-\mu_1^2-\mu_2^2}$$

Então:

$$g(\theta;\mu_1,\mu_2)=\frac{1}{2\pi.c.(1-\mu_1\cos\theta-\mu_2\sin\theta)} \tag{2.36}$$

em que

$$c=\frac{1}{\sqrt{1-\mu_1^2-\mu_2^2}}$$

Para simplificar os cálculos na obtenção da função de verossimilhança será realizada mais uma parametrização (JAMMALAMADAKA; SENGUPTA, 2001):

$$\eta_1 = c\mu_1 \quad \text{e} \quad \eta_2 = c\mu_2$$

daí $\eta_1^2 + \eta_2^2 = c^2(\mu_1^2 + \mu_2^2)$ como

$$c^2 = 1 + c^2(\mu_1^2 + \mu_2^2) = 1 + \eta_1^2 + \eta_2^2$$

A função densidade de probabilidade será escrita da seguinte maneira:

$$g(\theta;\eta_1,\eta_2) = \frac{1}{2\pi.(c - \eta_1\cos\theta - \eta_2\sin\theta)} \tag{2.37}$$

e a função de verossimilhança será dada por

$$L = L(\eta_1,\eta_2,\theta_i) = \prod_{i=1}^{n} \frac{1}{2\pi(c - \eta_1\cos\theta_i - \eta_2\sin\theta_i)}$$

$$L = \left(\frac{1}{2\pi}\right)^n \prod_{i=1}^{n} \frac{1}{(c - \eta_1\cos\theta_i - \eta_2\sin\theta_i)}$$

$$l = \log L = -nlog(2\pi) - \sum_{i=1}^{n} \log(c - \eta_1\cos\theta_i - \eta_2\sin\theta_i)$$

encontrando as derivadas $\frac{\partial l}{\partial \eta_1}$ e $\frac{\partial l}{\partial \eta_2}$ tem-se :

$$\frac{\partial l}{\partial \eta_1} = -\sum_{i=1}^{n} \frac{c^{-1}\eta_1 - \cos\theta_i}{c - \eta_1\cos\theta_i - \eta_2\sin\theta_i} \tag{2.38}$$

e

$$\frac{\partial l}{\partial \eta_2} = -\sum_{i=1}^{n} \frac{c^{-1}\eta_2 - \sin\theta_i}{c - \eta_1\cos\theta_i - \eta_2\sin\theta_i} \tag{2.39}$$

igualando as equações ( 2.38) e ( 2.39) a zero e lembrando que $\eta_1 = c.\mu_1$ e $\eta_2 = c.\mu_2$ :

$$\frac{\partial l}{\partial \eta_1} = \frac{1}{c}\sum_{i=1}^{n} \frac{1}{1 - \mu_1\cos\theta_i - \mu_2\sin\theta_i} \quad (\cos\theta_i - \mu_1)$$

$$\frac{\partial l}{\partial \eta_1} = \frac{1}{c}\sum_{i=1}^{n} w_i(\cos\theta_i - \mu_1) = 0 \tag{2.40}$$

analogamente,

$$\frac{\partial l}{\partial \eta_2} = \frac{1}{c}\sum_{i=1}^{n} w_i(\sin\theta_i - \mu_2) = 0 \tag{2.41}$$

sendo

$$w_i = \frac{1}{1 - \mu_1\cos\theta_i - \mu_2\sin\theta_i}$$

A partir de ( 2.40) e ( 2.41) obtem-se

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} w_i\cos\theta_i}{\sum_{i=1}^{n} w_i} \tag{2.42}$$

$$\hat{\mu_2} = \frac{\sum_{i=1}^{n} w_i\sin\theta_i}{\sum_{i=1}^{n} w_i} \tag{2.43}$$

**Distribuição von Mises** $vM(\mu,\kappa)$

A função densidade de probabilidade da distribuição von Mises é dada por:

$$g(\theta,\mu,\kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa\cos(\theta-\mu)} \tag{2.44}$$

sendo $I_0$ a função de Bessel modificada, de primeiro tipo e ordem 0, que pode ser definida por

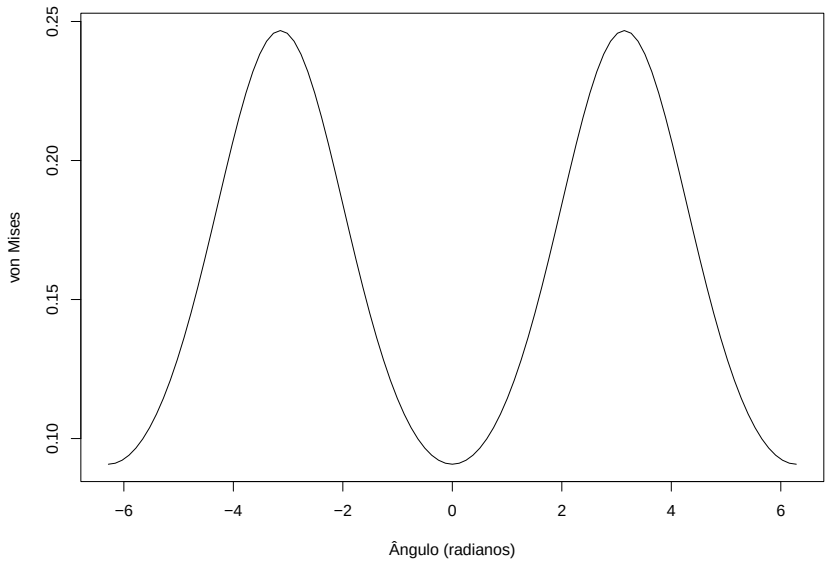
$$I_0(\kappa) = \frac{1}{2\pi}\int_0^{2\pi} e^{\kappa\cos\theta}\, d\theta \tag{2.45}$$

A distribuição von Mises será denotada por $vM(\mu,\kappa)$, o parâmetro $\mu$ é a média circular e $\kappa$ é o parâmetro de concentração. A distribuição é unimodal e simétrica para $\theta = \mu$. Observe que quando $\kappa = 0$, tem-se a distribuição uniforme. Do ponto de vista da inferência estatística esta distribuição pode ser considerada muito útil para a análise de dados circulares (MARDIA; JUPP, 2000).

A distribuição von Mises, $vM(\mu, \kappa)$ é a mais comum para amostras circulares unimodais, e em muitos aspectos a distribuição está para dados circulares assim como a distribuição normal está para os dados na reta. Esta distribuição no círculo é análoga a distribuição normal na reta real (FISHER, 1993).

Na figura ( 2.16) está representado o gráfico da função densida da distribuição von Mises, para o caso em que $\mu = \pi rad$ e $\kappa = 0,5$.

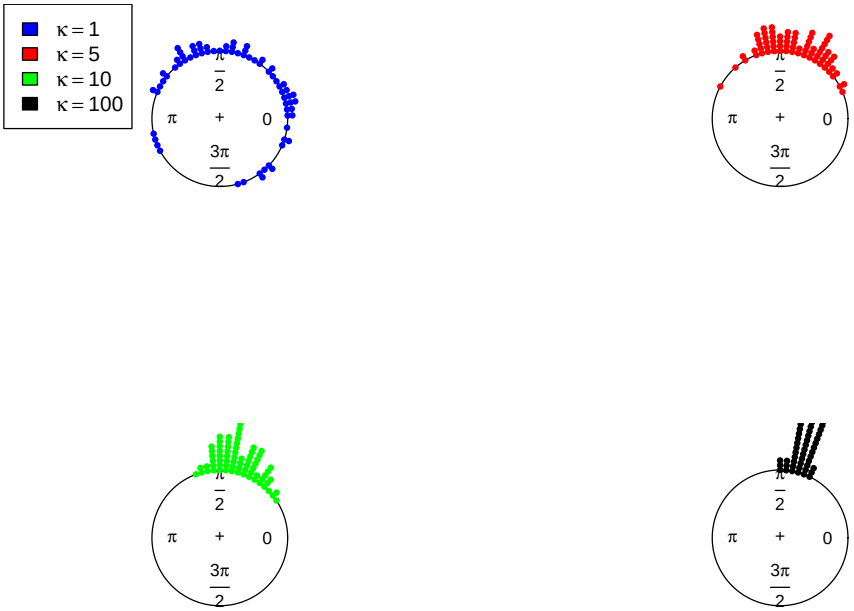
Figura 2.16 – Gráfico da distribuição de probabilidade da von Mises




Fonte: O autor (2022)

Para exemplificar a função do parâmetro $\kappa$ da distribuição, a figura ( 2.17) mostra a representação gráfica da simulação de dados da distribuição von Mises. Mantendo constante a média direcional $\mu = \pi rad$, e utilizando os valores $1;5;10;100$ para o parâmetro de concentração $\kappa$. Percebe-se que quanto maior o valor de $\kappa$, menor é a dispersão dos dados.

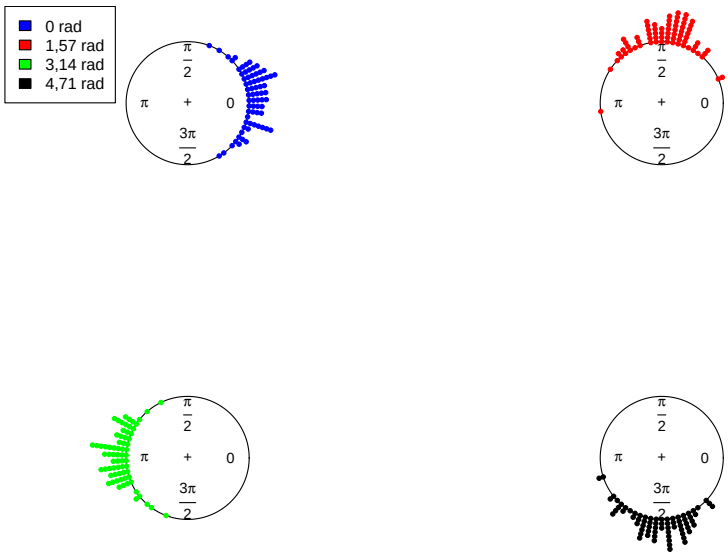
Figura 2.17 – Distribuição von Mises para $\kappa = 1, 5, 10, 100$




Fonte: O autor (2022)

Na figura ( 2.18), na qual os dados foram simulados, foi mantido constante o valor do parâmetros $\kappa = 5$ e variou-se a média direcional. Foram usados para a média os ângulos $0°, 90°, 180°, 270°$. No gráfico os valores estão em radianos. Note que o dados se concentram em torno da média direcional.

Figura 2.18 – Distribuição von Mises para $\mu = 0$ rad; $1,57$ rad; $3,14$ rad; $4,71$ rad




Fonte: O autor (2022)

**Inferência para a distribuição von Mises**

Algumas informações para a distribuição von Mises serão consideradas nesta secção.

**Máxima Verossimilhança**

Considere as medidas angulares $\theta_1, \theta_2, ..., \theta_n$ provenientes da distribuição von Mises $vM(\mu, \kappa)$. A função de verossimilhança será dada por:

$$L(\mu, \kappa, \theta_1, , ..., \theta_n) = \prod_{i=1}^{n} \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos(\theta_i - \mu)) \tag{2.46}$$

A função de log verosimilhança será:

$$l(\mu, \kappa, \theta_i, , ..., \theta_n) = \log\left(\prod_{i=1}^{n} \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos(\theta_i - \mu))\right) \tag{2.47}$$

$$= \log \left[ \left( \frac{1}{2\pi I_0(\kappa)} \right)^n \exp \sum_{i=1}^{n} (\kappa \cos(\theta_i - \mu)) \right]$$

$$l(\mu, \kappa.\theta_i, ..., \theta_n) = -n \log 2\pi - n \log I_0(\kappa) + \kappa \sum_{i=1}^{n} \cos(\theta_i - \mu) \tag{2.48}$$

derivando a equação ( 2.48) em relação a $\mu$ e $\kappa$ temos:

$$\frac{\partial l}{\partial \mu} = -\kappa \sum_{i=1}^{n} \sin(\theta_i - \mu)$$

$$\frac{\partial l}{\partial \kappa} = -n \frac{I_0'(\kappa)}{I_0(\kappa)} + \sum_{i=1}^{n} \cos(\theta_i - \mu)$$

Encontrando os valores máximos de $l$, ou seja, fazendo $\frac{\partial l}{\partial \mu} = 0$ e $\frac{\partial l}{\partial \kappa} = 0$, logo

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{n} \sin(\alpha_i - \mu) = 0$$

$$= \sum_{i=1}^{n} \sin\theta_i \cos\mu - \cos\theta_i \sin\mu = 0$$

$$= nS\cos\mu - nC\sin\mu = 0$$

$$\hat{\mu} = \arctan \frac{S}{C} = \bar{\theta}_0 \tag{2.49}$$

e

$$\frac{\partial l}{\partial \kappa} = -n \frac{I_0'(\kappa)}{I_0(\kappa)} + \sum_{i=1}^{n} \cos(\theta_i - \mu) \tag{2.50}$$

assim

$$\frac{dI_0(\kappa)}{d\kappa} = I_1(\kappa) \tag{2.51}$$

como

$$\sum_{i=1}^{n} \cos(\theta_i - \mu) = R$$

Temos que

$$-n\frac{I_1(\kappa)}{I_0(\kappa)}+R=0$$

logo

$$\frac{I_1(\kappa)}{I_0(\kappa)}=\frac{R}{n} \tag{2.52}$$

O estimador $\hat{\kappa}$ é a solução da equação $A(\hat{\kappa})=\bar{R}$, onde $A(\kappa)=\dfrac{I_0'(\kappa)}{I_0(\kappa)}$.

Os valores $\hat{\kappa}$ são obtidos por métodos numéricos. Mardia (2000) e Fisher (1993) sugerem a seguinte aproximação para o parâmetro $\kappa$ :

$$\hat{\kappa}=\begin{cases} 2\bar{R}+\bar{R}^3+\dfrac{5\bar{R}^5}{6}, & \bar{R}<0,53 \\ -0,4+1,39\bar{R}+\dfrac{0,43}{1-\bar{R}}, & 0,53\leq\bar{R}<0,85 \\ \dfrac{1}{\bar{R}^3-4\bar{R}^2+3\bar{R}}, & R\geq 0,85 \end{cases} \tag{2.53}$$

Quando o tamanho da amostra e os valores de $R$ são pequenos o estimador pode ser fortemente viesado. Para $n\leq 15$ o viés pode ser corrigido usando a aproximação (FISHER, 1993):

$$\hat{\kappa}=\begin{cases} max(\hat{\kappa}-\dfrac{1}{n\hat{\kappa}},0), \hat{\kappa}<2 \\ \dfrac{(n-1)^3\hat{\kappa}}{n^3+n}, \hat{\kappa}\geq 2 \end{cases} \tag{2.54}$$

**Matriz Informação de Fisher**

A matriz informação de Fisher representada por $I$ é dada por

$$I=E\left[-\begin{bmatrix} \frac{\partial^2 l}{\partial\mu^2} & \frac{\partial^2 l}{\partial\mu\partial\kappa} \\ \frac{\partial^2 l}{\partial\kappa\partial\mu} & \frac{\partial^2 l}{\partial\kappa^2} \end{bmatrix}\right]$$

Como

$$\frac{\partial^2 l}{\partial\mu^2}=-\kappa\sum_{i=1}^{n}\cos(\theta_i-\mu)=-\kappa R$$

$$\frac{\partial^2 l}{\partial \kappa^2} = -nA'(\kappa)$$

Em que

$$A(\kappa) = \frac{I_0'(\kappa)}{I_0(\kappa)} \quad ,A'(\kappa) = \frac{I_0''(\kappa)}{I_0(\kappa)} - A(\kappa)^2 \quad \text{e} I_0''(\kappa) = I_0(\kappa) - \frac{I_0'(\kappa)}{\kappa}$$

( ver (JAMMALAMADAKA; SENGUPTA, 2001)). Então

$$A'(\kappa) = 1 - \frac{A(\kappa)}{\kappa} - A(\kappa)^2$$

e

$$\frac{\partial^2 l}{\partial \mu \partial \kappa} = \frac{\partial^2 l}{\partial \kappa \partial \mu} = \sum_{i=1}^{n} \sin(\theta_i - \mu) = 0$$

Teremos que a matriz informação de Fisher será dada por

$$I = \begin{bmatrix} \frac{A(\hat{\kappa})}{\hat{\kappa}} & 0 \\ 0 & 1 - \frac{A(\hat{\kappa})}{\hat{\kappa}} - A(\hat{\kappa})^2 \end{bmatrix}$$

**Boxplot circular**

Ferramentas para visualização e análise dos dados são bastantes úteis. Uma delas é o box-plot. No caso de dados lineares é representado graficamente, por um retângulo, em que estão representados a mediana, o primeiro e o terceiro quartil. A medida do lado deste retângulo será chamada de $d_q$, que é o intervalo interquartil. Segmentos de reta partindo do primeiro quartil ($Q_1$) e do terceiro quartil ($Q_3$) são traçados e, com isso, é possível estabelecer os limites inferior($L_i$) e superior ($L_s$). Dados por

$$L_i = Q_1 - \nu(Q_3 - Q_1)$$

$$L_s = Q_3 + \nu(Q_3 + Q_1)$$

O valor geralmente utilizado para o fator $\nu$ quando a distribuição dos dados é a Normal é $1,5$, pois para esta distribuição isto resulta que 99% dos dados estejam contidos entre os dois

limites (NUZZO, 2016). As observações cuja medidas são menores que o limite inferior ($L_i$) e maiores que o limite superior ($L_s$) são chamados de outliers.

Apesar desta representação ser bastante útil, foi necessário desenvolver um boxplot específico para dados circulares (BUTTARAZZI; PANDOLFO; PORZIO, 2018; ABUZAID; MOHAMED; HUSSIN, 2012). Isto se tornou necessário, uma vez que, quando se trata de medidas circulares, o ponto inicial que pode ser o zero ou o polo norte, a depender dos dados. Bem como o sentido deles, horário ou anti-horário, podem afetar na visualização dos mesmos. No caso de dados lineares isto não acontece.

Com relação ás medidas circulares ou direcionais a constante $\nu$ pode assumir outros valores. Abuzaid, Mohamed e Hussin (2012), utilizando a distribuição von Mises e por simulação, verificaram que podem ser usados os seguintes valores para a constante:

$$\begin{cases} 1,5; \text{se} \quad 2 \leq \kappa \leq 3 \\ 2,5; \text{se} \quad \kappa > 3 \end{cases}$$

Para $\kappa < 2$ o valor da constante não está definido.

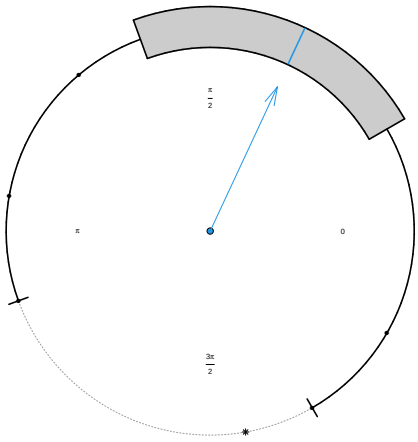
Buttarazzi, Pandolfo e Porzio (2018) também utilizando a distribuição von Mises, verificaram que o fator de multiplicação, $\nu$, pode variar de 0,493, quando $\kappa = 0$, atingindo um valor máximo de 2,168. Quando o valor de $\kappa$ é muito grande indicando pouca dispersão dos dados, o fator de multiplicação será 1,5. Isto faz sentido, pois para um valor do parâmetro $\kappa$ suficientemente grande a distribuição von Mises se aproxima da distribuição normal (JAMMALAMADAKA; SENGUPTA, 2001).

Nesta tese para construção do boxplot circular foi usado o pacote bpDir, que utiliza a proposta de Buttarazzi, Pandolfo e Porzio (2018).

A figura ( 2.19) mostra um exemplo do uso do boxplot circular com os mesmos dados da figura ( 2.9).

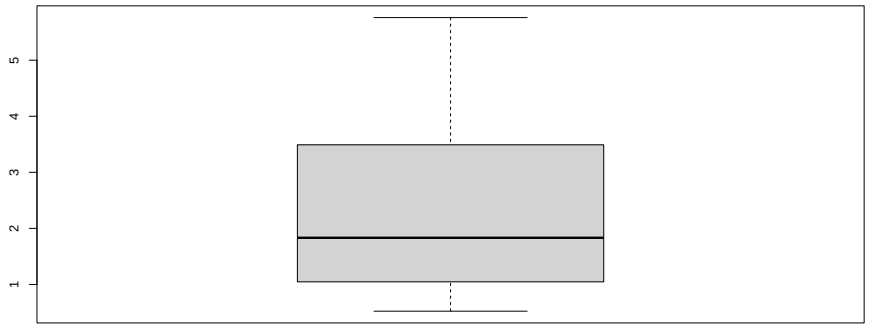
Observe que há uma vantagem na utilização do boxplot circular. No caso do uso do boxplot para dados lineares com medidas circulares poderia ocorrer interpretação errada. Por exemplo, no boxplot da figura ( 2.20), com os mesmos dados utilizados na construção do boxplot circular, não consta nenhum outlier. Porém no boxplot circular consta uma medida discrepante.

Figura 2.19 – Boxplot circular



Fonte: O autor (2022)

Figura 2.20 – Boxplot linear com dados circulares




Fonte: O autor (2022)

## 2.3 Correlação e Regressão para dados circulares

Será visto inicialmente a correlação entre as variáveis angulares e lineares; os modelos de regressão para dados circulares e a análise de resíduos para se verificar a qualidade do ajuste.

**Correlação**

A correlação pode ser utilizada quando se deseja avaliar a relação entre relação entre duas variáveis aleatórias, avaliar a dependência entre uma variável circular e uma variável linear ou entre duas variáveis circulares. No caso em que as variáveis são lineares o modelo de correlação de Pearson é muito utilizado. Quando se trata de variáveis circulares serão anlisadas duas situações: correlação linear-circular e a circular-circular.

**Linear-circular**

Uma medida de correlação entre uma variável linear $x$ que pode ser, por exemplo, radiação diária ou umidade relativa do ar e uma variável angular $\alpha$ a qual pode ser a direção dos ventos e meses do ano é dada por (MARDIA; SUTTON, 1978):

$$R^2_{x,\alpha} = \frac{r^2_{xc} + r^2_{xs} - 2r_{xc}r_{xs}r_{cs}}{1 - r^2_{cs}} \tag{2.55}$$

em que

$$r_{xc} = corr(x, \cos\alpha)$$

$$r_{xs} = corr(x, \sin\alpha)$$

$$r_{cs} = corr(\cos\alpha, \sin\alpha)$$

Quando o valor de $R^2_{x,\alpha}$ está próximo de 0 não há relação entre as variáveis e quando o valor está próximo de 1 há evidência de uma forte relação entre as variáveis. De acordo com Jammalamadaka e Sengupta (2001) quando $X$ e $\alpha$ são independentes e $X$ tem distribuição normal, então

$$\frac{(n-3)R^2}{1-R^2} \sim F_{2,n-3}$$

**Correlação circular-circular**

Considerando duas variáveis circulares $\theta$ e $\phi$, uma medida para o coeficiente de correlação circular- circular é (FISHER; LEE, 1992):

$$\rho_T = \frac{\sum_{i,j=1}^{n} \sin(\theta_i - \theta_j)\sin(\phi_i - \phi_j)}{\sqrt{\sum_{i,j=1}^{n} \sin^2(\theta_i - \theta_j)\sum_{i,j=1}^{n} \sin^2(\phi_i - \phi_j)}} \tag{2.56}$$

Para fins computacionais pode ser utilizada a seguinte alternativa:

$$\rho_T = \frac{4(AB - CD)}{\sqrt{(n^2 - E^2 - F^2)(n^2 - G^2 - H^2)}} \tag{2.57}$$

sendo

$$A = \sum_{i,j=1}^{n} \cos\theta_i \cos\phi_j$$

$$B = \sum_{i,j=1}^{n} \sin\theta_i \sin\phi_j$$

$$C = \sum_{i,j=1}^{n} \cos\theta_i \sin\phi_j$$

$$D = \sum_{i,j=1}^{n} \sin\theta_i \cos\phi_j$$

$$E = \sum_{i=1}^{n} \cos 2\theta_i$$

$$F = \sum_{i=1}^{n} \sin 2\theta_i$$

$$G = \sum_{j=1}^{n} \cos 2\phi_j$$

$$H = \sum_{j=1}^{n} \cos 2\phi_j$$

## 2.4 Modelos de Regressão

A análise de regressão é utilizada para investigar a relação existente entre uma variável, que é chamada de variável resposta, e as variáveis explicativas. A relação entre estas variáveis se dá por meio de um modelo. No caso em que as variáveis têm como suporte a reta real, os modelos de regressão podem ser lineares ou não lineares.

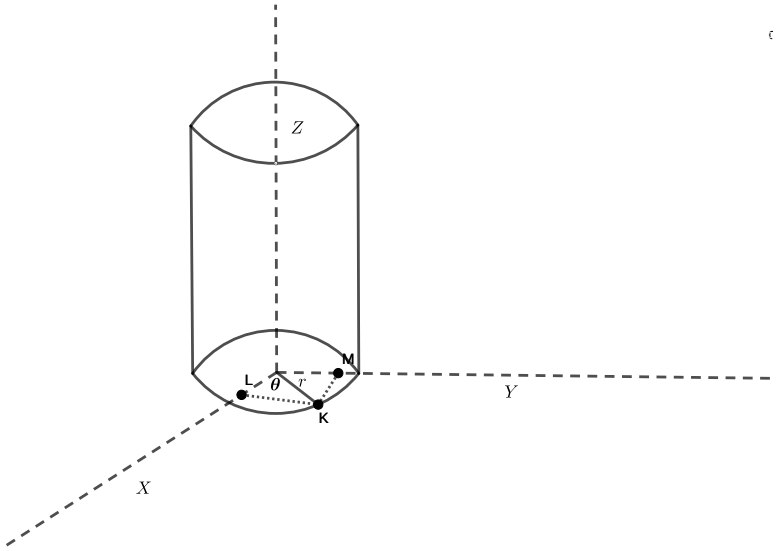
Quando se trata de variáveis circulares ou medidas que podem ser tranformadas em ângulos, os modelos de regressão podem ser classificados em três tipos: Circular-circular, circular-linear ou linear-circular (JAMMALAMADAKA; SENGUPTA, 2001). Esta classificação se dá devido ao suporte dos dados, se estão na reta real ou na circunferência unitária.

a) Linear - Circular : Quando a variável explicativa é circular e a variável resposta é linear.
b) Circular- Linear : Quando a variável explicativa é linear e a variável resposta é circular.
c) Circular- Circular : Quando tanto a variável resposta como explicativa são circulares.

### Modelo de Regressão linear - circular

Ocorrem situações em que a variável explicativa é uma medida circular $\theta$ e a variável resposta $x$ é linear, neste caso $(x, \theta)$ assume valores num cilindro. Assim, foi proposto um modelo baseado numa distribuição cilíndrica (MARDIA; SUTTON, 1978; ANDERSON-COOK; NOBLE, 2001) em que a altura da curva no cilindro é a parte linear do modelo (ver figura 2.21). Este modelo proposto inicialmente por Mardia e Sutton (1978), e considera uma distribuição conjunta, na qual a variável circular $\Theta$ tem distribuição von Mises - $vM(\mu_0, \kappa)$, e a componente linear $x$ segue a distribuição Normal condicionada ao valor de $\theta$.

Figura 2.21 – Regressão cilíndrica




Fonte: O autor (2022)

Assim, a distribuição conjunta é dada por

$$f(x,\theta) = \frac{1}{2\pi I_0(\kappa)} \exp\{\kappa \cos(\theta - \mu_0)\} \frac{1}{\sigma_c^2\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_c)^2}{2\sigma_c^2}\right) \tag{2.58}$$

Sendo $\mu_0$ e $\kappa$, respectivamente, a média circular e o parâmetro de concentração da distribuição von Mises, $\mu_c$ e $\sigma_c^2$ são a média e a variância da distribuição Normal, respctivamente. Os parâmetros $\rho_1$ e $\rho_2$ descrevem a associação entre a variável linear e a circular, e $x$, $\mu_c \in \mathbb{R}$; $\theta$, $\mu_0 \in [0, 2\pi)$, $\kappa > 0$, além disso,

$$\mu_c = \mu + \frac{\sigma}{\sqrt{\kappa}}\{\rho_1(cos\theta - \cos\mu_0) + \rho_2(\sin\theta - \sin\mu_0)\} \tag{2.59}$$

$\sigma_c^2 = \sigma^2(1-\rho^2)$ e $\rho^2 = \rho_1^2 + \rho_2^2, \quad 0 \leq \rho \leq 1.$

De ( 2.59) tem-se que

$$\mu_c = \mu - \frac{\sigma}{\sqrt{\kappa}}(\rho_1 \cos\mu_0 + \rho_2 \sin\mu_0) + \frac{\sigma}{\sqrt{\kappa}}\rho_1 \cos\theta + \frac{\sigma}{\sqrt{\kappa}}\rho_2 \sin\theta$$

Considerando

$$\beta_0 = \mu - \frac{\sigma}{\sqrt{\kappa}}(\rho_1 \cos\mu_0 + \rho_2 \sin\mu_0)$$

$$\beta_1 = \frac{\sigma}{\sqrt{\kappa}}\rho_1$$

$$\beta_2 = \frac{\sigma}{\sqrt{\kappa}}\rho_2$$

a esperança condicional $E(x|\theta)$ é dada por:

$$E(x|\theta) = \beta_0 + \beta_1 \cos\theta + \beta_2 \sin\theta \tag{2.60}$$

Então o modelo linear-circular será:

$$y_i = \beta_0 + \beta_1 \cos\theta_i + \beta_2 \sin\theta_i + \varepsilon_i \tag{2.61}$$

Em virtude da distribuição conjunta utilizada o erro $\varepsilon_i$ pode ser considerado Normal, com média zero e variância constante.

Na forma matricial $Y = \mathbf{X}\beta^T$, ou seja

$$\mathrm{X}\beta^T = \begin{pmatrix} 1 & \cos\theta_1 & \sin\theta_1 \\ 1 & \cos\theta_2 & \sin\theta_2 \\ \vdots & \vdots & \vdots \\ 1 & \cos\theta_n & \sin\theta_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

em que $\beta$ é o vetor dos parâmetros que serão estimados.

**Modelos de Regressão circular-linear**

Fisher e Lee (1992) propuseram um modelo de regressão para a variável resposta circular e as variáveis exploratórias lineares. Então suponha que $(\theta_1, x_1), ..., (\theta_n, x_n)$ em que $\theta_i$ são observa-

ções circulares independentes, seguindo a distribuição de von Mises, e $x_i$ são variáveis exploratórias, sendo $i = 1, \cdots, n$.

Podem ser considerados três modelos (FISHER; LEE, 1992) :

a) Modelo para a média direcional $\mu$ em relação ao vetor de covariáveis $\mathbf{X} = (x_1, ..., x_k)^T$;
b) Modelo de dispersão em relação ao vetor de covariáveis $\mathbf{X} = (x_1, ..., x_k)^T$;
c) Modelo misto, para a média direcional e a dispersão simultâneamente, em relação ao vetor de covariáveis $\mathbf{X} = (x_1, ..., x_n)^T$.

**Modelo de regressão circular-linear para a média**

Suponha que $(\theta_1, x_1), \cdots, (\theta_n, x_n)$ são $n$ observações, nas quais $\theta_i$ seguem a distribuição von Mises e $x_i$ são variáveis explicativas, com $i = 1, \cdots, n$. Suponha que os parâmetros de concentração são todos iguais, ou seja, $\kappa_1 = \cdots = \kappa_n = \kappa$ e a média direcional $\mu_i$ está relacionada com o vetor de variáveis $X_i$ por meio da equação

$$\mu_i = \mu + g(\beta^T \mathbf{X}_i)$$

O vetor $\beta = (\beta_1, \cdots, \beta_k)$ é formado pelos parâmetros do modelos que serão estimados e a função $g$ é chamada de função de ligação, em que $g : \mathbb{R} \to (-\pi, \pi)$, duas vezes diferenciável, monótona e tal que $g(0) = 0$ (FISHER, 1993), com isso pode ser utilizada a função proposta por Fisher e Lee (1992) dada por:

$$g(u) = 2\tan^{-1}(u) \tag{2.62}$$

Com isso tem-se que a funçãode verossimilhança $L$ é

$$L(\theta_i, \mu_i, \kappa) = \prod_{i=1}^{n} \frac{1}{2\pi I_0(\kappa)} \exp\left[\kappa_i \cos(\theta_i - \mu_i)\right]$$

$$= \frac{1}{2\pi I_0(\kappa)} \exp\left[\kappa_i \cos(\theta_i - \mu + g(\beta^T X_i)\right]$$

A função log verossimilhança será dada por:

$$l(\theta_1,...\theta_n,\mu_i,\kappa) = -n\log I_0(\kappa) + \kappa\sum_{i=1}^{n}\cos(\theta_i - \mu - g(\beta^T x_i)) \tag{2.63}$$

Derivando a função ( 2.63) em relação a $\mu$ e $\theta$ e igualando as derivadas a zero é possível encontrar os estimadores $\hat{\mu}$, $\hat{\kappa}$ e $\hat{\beta}$.

Mas para que isso seja possível serão utilizadas as relações a seguir, indicadas por Fisher e Lee (1992)

$$u_i = \sin(\theta_i - \mu - g(\beta^T x_i))$$

$$u = (u_1....,u_n)^T$$

$$X = (x_1,...,x_n)^T$$

$$G = \text{diag}(g'(\beta^T x_1),...,g'(\beta^T x_n))$$

$$S = \sum \frac{\sin(\theta_i - g(\beta^T x_i))}{n}$$

$$S = \sum \frac{\cos(\theta_i - g(\beta^T x_i))}{n}$$

$$R^2 = S^2 + C^2$$

Encontrando $\dfrac{\partial f}{\partial \mu}$ temos

$$\frac{\partial f}{\partial \mu} = \kappa\sum_{i=1}^{n}\sin(\theta_i - \mu - g(\beta^T x_i))$$

$$= \kappa\sum_{i=1}^{n}\sin(\theta_i - g(\beta^T x_i) - \mu)$$

$$= \kappa\sum_{i=1}^{n}\left(\sin(\theta_i - g(\beta^T x_i)\cos\mu) - \sin\mu(\cos(\theta_i - g(\beta^T x_i))\right)$$

$$= \kappa(nS\cos\mu - nC\sin\mu)$$

Para se obter o estimador de máxima verossimilhança calcula-se

$$\frac{\partial f}{\partial \mu} = 0$$

assim

$$n\kappa(S\cos\hat{\mu} - C\sin\hat{\mu}) = 0$$

$$S\cos\hat{\mu} = C\sin\hat{\mu}$$

$$S^2\cos^2\hat{\mu} = C^2\sin^2\hat{\mu}$$

$$S^2\cos^2\hat{\mu} = C^2(1-\cos^2\hat{\mu})$$

$$(S^2 + C^2)\cos^2\hat{\mu} = C^2$$

logo

$$R^2\cos^2\hat{\mu} = C^2$$

encontrando a raiz quadrada de ambos os membros da igualdade tem-se

$$R|\cos\hat{\mu}| = |C|$$

Serão analisadas as seguintes situações:

a) Se $0 < \hat{\mu} < \frac{\pi}{2}$ ou $\frac{3\pi}{2} < \hat{\mu} < 2\pi$, tem-se que $\cos\hat{\mu} > 0$ e $C > 0$. Assim,

$$R\cos\hat{\mu} = C$$

b) se $\dfrac{\pi}{2} < \hat{\mu} < \pi$ ou $\pi < \hat{\mu} < \dfrac{3\pi}{2}$, daí $\cos\hat{\mu} < 0$ e $|C| = -C$, pois $C < 0$. Com isso

$$|\cos\hat{\mu}| = -\cos\hat{\mu}$$

e

$$R\cos\hat{\mu} = C$$

Analogamente tem-se que $R\sin\hat{\mu} = S$, daí o estimador $\hat{\mu}$ tem que satisfazer as seguintes equações (LEAL, 2006)

$$R\cos\hat{\mu} = C \tag{2.64}$$

$$R\sin\hat{\mu} = S \tag{2.65}$$

Obtendo $\frac{\partial f}{\partial \beta}$

$$\frac{\partial f}{\partial \beta} = \kappa \sum_{i=1}^{n} \sin(\theta_i - \mu - g(\beta^T x_i)).g'(\beta^T x_i)x_i$$

$$= \kappa \sum_{i=1}^{n} x_i g'(\beta^T x_i)\sin(\theta_i - \mu - g(\beta^T x_i))$$

Para encontrar as estimativas dos estimadores da log-verossimilhança será resolvida a equação $\frac{\partial f}{\partial \beta} = 0$ logo

$$\sum_{i=1}^{n} x_i g'(\beta^T x_i)\sin(\theta_i - \mu - g(\beta^T x_i)) = 0$$

$$\sum_{i=1}^{n} x_i g'(\hat{\beta}^T x_i)u_i = 0$$

$$x_1 g'(\hat{\beta} x_1)u_1 + ... + x_n g'(\hat{\beta}_n)u_n = 0$$

$$X^T G u = 0$$

Falta determinar $\frac{\partial f}{\partial \kappa} ento \frac{\partial f}{\partial \kappa} = -n\frac{I_0'(\kappa)}{I_0(\kappa)} + \sum_{i=0}^{n}\cos(\theta_i - \mu - g(\beta^T x_i))$
como

$$A_1(\kappa) = \frac{I_0'(\kappa)}{I_0(\kappa)}$$

$$\frac{\partial f}{\partial \kappa} = -nA_1(\kappa) + \sum_{i=0}^{n}\cos(\theta_i - g(\beta^T x_i) - \mu)$$

$$= -nA_1(\kappa) + \sum_{i=1}^{n}\cos(\theta_i - g(\beta^T x_i)\cos\mu + \sin(\theta_i - g(\beta^T x_i)\sin\mu)$$

assim,

$$-nA_1(\hat{\kappa}) + nC\cos\hat{\mu} + nS\sin\hat{\mu} = 0$$

$$nA_1(\hat{\kappa}) = nC\cos\hat{\mu} + nS\sin\hat{\mu}$$

das equações ( 2.64) e ( 2.65) conclui-se que

$$A_1(\hat{\kappa}) = \frac{C^2}{R} + \frac{S^2}{R} = R$$

logo

$$A_1(\hat{\kappa}) = R$$

Para encontrar as estimativas $\hat{\mu}$, $\hat{\beta}$ e $\hat{\kappa}$ é necessário resolver as equações

$$\begin{cases} X^T G u = 0 \\ R\sin\hat{\mu} = S \\ R\cos\hat{\mu} = C \\ A_1(\hat{\kappa}) = R \end{cases}$$

**Modelo de regressão circular-linear para a dispersão**

Considere as $n$ observações $(\theta_1, x_1), ..., (\theta_n, x_n)$ , em que $\theta_i$, com $i = 1, ..., n$ são as variáveis respostas seguindo a distribuição de von Mises. Suponha que $\mu_1, ..., \mu_n = \mu$, e os parâmetros de concentração $\kappa_1, ..., \kappa_n$ estão relacionados com $x_i$ pela função de ligação $h$ dada por

$$\kappa_i = h(\gamma^T x)$$

onde

$$\gamma^T \mathbf{X} = \gamma_0 + \gamma_1 x_1 + ... + \gamma_k x_k$$

A função de log verossimilhança

$$l = l(\mu, \gamma; y) = -\sum_{i=1}^{n} \log I_0(\kappa_i) + \sum_{i=1}^{n} \kappa_i \cos(\theta_i - \mu)$$

$$= \sum_{i=1}^{n} \log(I_0(\gamma^T x_i)) + \sum_{i=1}^{n} n\cos(\theta_i - \mu)$$

Com a finalidade de obter os estimadores de máxima verossimilhança serão encontrados, $\frac{\partial l}{\partial \mu}, \frac{\partial l}{\partial \gamma_i}, \frac{\partial l}{\partial \alpha}$

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{n} \kappa_i \sin(\theta_i - \mu)$$

$$= \sum_{i=1}^{n} \kappa_i \left(\sin\theta_i \cos\mu - \sin\mu \cos\theta_i\right)$$

$$= \cos\mu \sum_{i=1}^{n} \kappa_i \sin\theta_i - \sin\mu \sum_{i=1}^{n} \kappa_i \cos\theta_i$$

fazendo

$$S = \sum_{i=1}^{n} \kappa_i \sin(\theta_i)$$

$$C = \sum_{i=1}^{n} \kappa_i \cos(\theta_i)$$

logo

$$\frac{\partial l}{\partial \mu} = S\cos\mu - C\sin\mu$$

assim

$$S\cos\hat{\mu} - C\sin\hat{\mu} = 0$$

$$S^2\cos^2\hat{\mu} - C^2\sin^2\hat{\mu} = 0$$

$$S^2\cos^2\hat{\mu} - C^2(1 - \cos^2\hat{\mu} = 0$$

$$S^2 + C^2 = C^2\cos^2\hat{\mu} = 0$$

$$R^2 = C^2\cos^2\hat{\mu}$$

$$R = |C||\cos\hat{\mu}|$$

daí, $R = C\cos\hat{\mu}$ e analogamente é obtido $R = S\sin\hat{\mu}$.

Calculando agora $\frac{\partial l}{\partial \gamma_i}$:

$$\frac{\partial l}{\partial \gamma_i} = -\sum_{i=1}^{n} \frac{I_0'(\kappa_i)}{I_0(\kappa_i)} h'(\gamma_i^T x) \frac{\partial \kappa_i}{\partial \gamma_i} + \sum_{i=1}^{n} h'(\gamma_i^T x) \frac{\partial \kappa_i}{\partial \gamma_i} \cos(\theta_i - \mu)$$

como $I_1(\kappa_i) = I_0'(\kappa_i)$ e $A_1(\kappa_i) = \dfrac{I_1(\kappa_i)}{I_0(\kappa_i)}$ tem-se que

$$\frac{\partial l}{\partial \gamma_i} = -\sum_{i=1}^{n} A_1(\kappa_i) h'(\gamma_i^T x_i) x_{ij} + h'(\gamma^T x_i) x_{ij} \cos(\theta_i - \mu)$$

$$= -\sum_{i=1}^{n} h'(\gamma_i^T x_i)(A_1(\kappa_i) - \cos(\theta_i - \mu)) x_{ij}$$

Então os estimadores $\hat{\gamma}_i$ com $j = 0, 1, ..., k$ têm que satisfazer a condição abaixo:

$$\sum_{i=1}^{n} \left[ h'(\hat{\gamma}_i^T x_i) \left( -A_1(\hat{\kappa}_i) + \cos(\hat{\theta}_i - \hat{\mu}) \right) \right] x_{ij} = 0 \tag{2.66}$$

Os estimadores $\hat{\mu}$, $\hat{\kappa}$ e $\hat{\gamma}_i$ não são obtidos algebricamente e sim por métodos numéricos.

**Modelos de regressão circular-linear misto**

Nas subsecções anteriores foi modelado a média ou a dispersão separadamente, mas há situações em que ocorrem conjuntos de dados nos quais a média e a dispersão dependem das covariáveis. Estes modelos são chamados de mistos e são ajustados para a média e dispersão, simultâneamente. A função de log-verosimilhança será dada por

$$l = l(\mu, \beta, \gamma, \theta) = -\sum_{i=1}^{n} \log I_0(\kappa_i) + \sum_{i=1}^{n} \kappa_i \cos(\theta_i - \mu - g(\beta^T x_i))$$

na qual

$$\kappa_i = h(\gamma^T x_i)$$

Os estimadores podem ser obtidos combinando os métodos usados nas duas secções anteriores.

As estimativas dos estimadores $\hat{\mu}$, $\hat{\kappa}$ e $\hat{\gamma}_i$ não são obtidos algebricamente e sim por métodos numéricos.

**Modelo de regressão circular-circular**

No caso em que tanto a variável explicativa como a variável resposta são circulares, como por exemplo a direção dos ventos e os meses do ano, o modelo a ser usado é o de regressão circular-circular. Serão analisadas três modelos diferentes para o caso em que as variáveis resposta e explicativa são circulares.

1- Considerando $(\alpha, \beta)$, um par de variáveis aleatórias direcionais. Os ângulos $\alpha$ e $\beta$ podem ser representados como vetores unitários em função dos seus respectivos valores de seno e cosseno. Com isso, dado $f(\alpha, \beta)$ a distribuição conjunta, com $0 < \alpha, \beta \leq 2\pi$. Para predizer $\beta$ para um valor dado de $\alpha$, será considerado a regressão do vetor $e^{i\beta}$ dado $\alpha$ (JAMMALAMADAKA; SENGUPTA, 2001), dada por:

$$E(e^{i\beta}|\alpha) = g(\alpha) = g_1(\alpha) + ig_2(\alpha) \tag{2.67}$$

e

$$\begin{cases} E(\cos\beta|\alpha) = g_1(\alpha) \\ E(\sin\beta|\alpha) = g_2(\alpha) \end{cases} \tag{2.68}$$

A partir da equação ( 2.68) pode ser obtido $\beta$, dado por

$$\mu(\alpha) = \hat{\beta} = \arctan\frac{g_2(\alpha)}{g_1(\alpha)} \tag{2.69}$$

em que $\mu(\alpha)$ é a média direcional de $\beta$ dado $\alpha$.

As funções $g_1(\alpha)$ e $_2(\alpha)$. Como $g_1(\alpha)$ e $g_2(\alpha)$ são funções periódicas, com período $2\pi$, elas podem ser expressas utilizando as suas expansões da série de Fourier.

Dada uma função $f(x)$, a série

$$\frac{a_0}{2} + \sum_{n=1}^{n} (a_n \cos nx + b_n \sin nx) \tag{2.70}$$

onde $a_n$ e $b_n$ são os coeficientes de Fourier de $f$, denomina-se *série de Fourier* de $f$. Então as funções $g_1$ e $g_2$ serão aproximadas por polinômios trigonométricos com graus adequados,

$m$, de modo que:

$$\begin{cases} g_1(\alpha) \approx \sum_{k=0}^{m}(A_k \cos k\alpha + B_k \sin k\alpha) \\ g_2(\alpha) \approx \sum_{k=0}^{m}(C_k \cos k\alpha + D_k \sin k\alpha) \end{cases} \tag{2.71}$$

da equação ( 2.71) pode se afirmar que

$$\begin{cases} \cos\beta = \sum_{k=0}^{m}(A_k \cos k\alpha + B_k \sin k\alpha) + \varepsilon_1 \\ \sin\beta = \sum_{k=0}^{m}(C_k \cos k\alpha + D_k \sin k\alpha) + \varepsilon_2 \end{cases} \tag{2.72}$$

em que o vetor $\varepsilon$=$(\varepsilon_1, \varepsilon_2)$ é o vetor de erros, considerado aleatório com média zero e matriz de covariância dada por $\Sigma$.

Os parâmetros do modelo a serem estimados são $A_k, B_k, C_k, D_k$ para $k = 0, 1, ..., m$. Observe que quando $k = 0$ será considerado $C_k = D_k = 0$.

2- Considere agora as variáveis circulares $X$ e $Y$, com médias circulares $\alpha$ e $\beta$, respectivamente, e $\omega$ o parâmetro de forma com $-1 \leq \omega \leq 1$, sendo que a variável resposta $Y$ tem distribuição von Mises (DOWNS; MARDIA, 2002). O modelo considerado é dado por:

$$\tan\frac{1}{2}(Y-\beta) = \omega \tan\frac{1}{2}(X-\alpha) \tag{2.73}$$

então

$$Y = \beta + 2\arctan(\omega \tan\frac{1}{2}(X-\alpha)) \tag{2.74}$$

3- Outro modelo de regressão circular-circular pode ser obtido por meio da Transformação de Möbius, na proposta inicial o erro angular é dado pela distribuição Wrapped Cauchy (KATO; SHIMIZU; SHIEH, 2018).

Considere $x$ e $y$ variáveis angulares, $\beta_0$ e $\beta_1$ números complexos tais que, $|\beta_0| = 1$. A Transformação de Möbius é dada por

$$y = \beta_0 \frac{x + \beta_1}{1 + \bar{\beta}_1 x}\varepsilon \tag{2.75}$$

sendo $\varepsilon$ o vetor de erros.

**Transformação de Möbius**

Considere $y$ a variável resposta e $x$ a variável explicativa, ambas medidas angulares tomadas num círculo unitário. O modelo de regressão é dado por

$$y = \beta_0 \frac{x+\beta_1}{1+\bar{\beta}_1 x} \tag{2.76}$$

Em que as varíaveis $y$ e $x$ são angulares, $\beta_0$ e $\beta_1$ são números complexos, sendo $|\beta_0| = 1$ e $|\beta_1| \neq 1$.

A Transformação de Möbius é uma função bijetora, que leva o círculo de raio 1 no círculo de raio 1. Para verificar isso basta mostrar que $|y| = 1$ (PAULA, 2018). Como $|\beta_0| = 1$, será considerado $\beta_0 = 1$, além disso, $x = e^{i\alpha}$, $\beta_1 = re^{i\nu}$. Então

$$|y| = \frac{|x+\beta_1|}{|1+\bar{\beta}_1 x|} = \frac{|e^{i\alpha} + re^{i\nu}|}{|1+re^{i(\alpha-\nu)}|}$$

$$= \frac{|\cos\alpha + r\cos\nu + i(\sin\alpha + r\sin\nu)|}{|1+r\cos(\alpha-\nu) + ir\sin(\alpha-\nu)|}$$

$$= \frac{\sqrt{(cos\alpha + r\cos\nu)^2 + (\sin\alpha + r\sin\nu)^2}}{\sqrt{[1+r\cos(\alpha-\nu)]^2 + [r\sin(\alpha-\nu)]^2}}$$

$$= \frac{\sqrt{1+2r\cos(\alpha-\nu)+r^2}}{\sqrt{1+2r\cos(\alpha-\nu)+r^2}}$$

$$|y| = 1$$

Escrevendo $Y = e^{i\theta}$, $X = e^{i\alpha}$, $\beta_0 = e^{i\mu}$ e $\beta_1 = re^{i\nu}$, então ( 2.76) pode ser escrita da seguinte maneira:

$$e^{i\theta} = e^{i\mu} \frac{e^{i\alpha} + re^{i\nu}}{re^{i(\alpha-\nu)} + 1} \tag{2.77}$$

Na expressão ( 2.76) o parâmetro $\beta_0$ é chamado de parâmetro de rotação e $\beta_1$ é o parâmetro de atração. A concentração dos pontos está na direção de $\frac{\beta_1}{|\beta_1|}$, e cresce nesta mesma na direção à medida que $\beta_1$ cresce.

O modelo de regressão será dado por

$$Y = \beta_0 \frac{X + \beta_1}{1 + \bar{\beta}_1 X} \varepsilon \tag{2.78}$$

Onde as variáveis $Y$ e $X$ são angulares, com $|x| = 1$, $\beta_0$ e $\beta_1$ números complexos, tais que $|\beta_0| = 1$ , $|\beta_1| \neq 1$, e $\varepsilon$ é o vetor de erros.

**Análise de Resíduos**

A análise de resíduos é uma ferramenta importante para se verificar a qualidade do ajuste, bem como detectar a presença de observações extremas ou outliers, pois a presença deles pode afetar a qualidade do ajuste.

Para os modelos em que a resposta é circular (circular-circular e circular-linear) a maneira de analisar os resíduos segue uma metodologia diferente (SOUZA; PAULA, 2002; LEAL, 2006).

Uma medida útil é a função desvio (deviance), muito utilizada no contexto dos modelos lineares generalizados.

Souza e Paula (2002) propuseram, considerando o modelo de regressão circular-linear, no qual a variável resposta foi modelada pela von Mises, a deviance angular.

A função desvio é dada por $D(y, \hat{\mu}) = \sum_{i=1}^{n} d_i^2$ sendo

$$d_i = \pm\sqrt{2}(l_i(y_i, \tilde{\mu}_i, \kappa) - l_i(y_i, \hat{\mu}_i, \kappa))^{\frac{1}{2}} \tag{2.79}$$

Em que $l_i(y_i,)$ é a contribuição de $y_i$ para a log verossimilhança total, $\tilde{\mu}_i$ é o estimador de máxima verossimilhança de $\mu_i$ baseado apenas em $y_i$ e $\hat{\mu}$ é o estimador de máxima verossimilhança baseado na amostra completa (SOUZA, 1999)

Como a variável resposta segue a distribuição von Mises $vM(\mu, \kappa)$ a log verossimilhança é dada por

$$l(y_i, \mu_i, \kappa) = -n\log(I_0(\kappa)) + \kappa \sum_{i=1}^{n} \cos(y_i - \mu_i)$$

Com isso

$$l_i(y_i,\mu_i,\kappa) = -\log(I_0(\kappa)) + \kappa\cos(y_i-\mu_i)$$

e

$$l_i(y_i,\tilde{\mu}_i,\kappa) = -\log(I_0(\kappa)) + \kappa\cos(y_i-\tilde{\mu}_i)$$

$$l_i(y_i,\hat{\mu}_i,\kappa) = -\log(I_0(\kappa)) + \kappa\cos(y_i-\hat{\mu}_i)$$

utilizando a equação (2.79)

$$d_i = \pm\sqrt{2}\,(-\log(I_0(\kappa)) + \kappa\cos(y_i-\tilde{\mu}_i) + \log(I_0(\kappa)) - \kappa\cos(y_i-\hat{\mu}_i))^{\frac{1}{2}}$$

$$d_i = \pm\sqrt{2}\,(\kappa\cos(y_i-\tilde{\mu}_i) - \kappa\cos(y_i-\hat{\mu}_i))^{\frac{1}{2}}$$

como $\cos(y_i-\tilde{\mu}_i) = 1$ pois $\tilde{\mu}_i$ é o estimador da máxima verossimilhança baseado apenas em $y_i$.

$$d_i = \pm\sqrt{2}\,[\kappa(1-\cos(y_i-\hat{\mu}_i))]^{\frac{1}{2}} \tag{2.80}$$

Usando a relação trigonométrica $\cos(2\theta) = \cos^2(\theta) - \sin^2(\theta)$, tem-se que

$$\sin^2(\theta) = \frac{1-cos(2\theta)}{2}$$

então $\sin(\frac{\theta}{2}) = \pm\left(\frac{1-\cos(\theta)}{2}\right)^{\frac{1}{2}}$

$$\sqrt{2}\sin(\frac{\theta}{2}) = \pm(1-\cos(\theta))^{\frac{1}{2}}$$

A equação ( 2.80) pode ser escrita como:

$$d_i = \pm 2\sqrt{\kappa}\sin\left[\frac{y_i-\hat{\mu}}{2}\right] \tag{2.81}$$

Souza e Paula (2002) propuseram uma correção para a função desvio, equação ( 2.81), dada por $d_i^*$:

$$d_i^* = \pm 2\sqrt{\kappa}\frac{\sin(\frac{1}{2}(y_i - \hat{\mu}_i))}{(1-\hat{h}_{ii}^*)^{\frac{1}{2}}} \tag{2.82}$$

sendo que $\hat{h}_{ii}^*$ é o elemento da diagonal principal da matriz $\mathbf{H}^* = \mathbf{GX}(\mathbf{X^T G^2 X})^{-1}\mathbf{X^T G}$, com $G = diag(g'(\beta \mathbf{x_i}))^{\mathbf{T}}$, avaliada na estimativa de máxima verossimilhança. Eles também consideraram o resíduo $r_i$ dado por:

$$r_i = \pm\left(\frac{2}{\pi}\right)^{\frac{1}{2}} \frac{\sin\left(\frac{1}{2}(y_i - \hat{\mu})\right)}{I_0(\hat{\kappa})e^{-\hat{\kappa}}} \tag{2.83}$$

cuja distribuição de probabilidade é equivalente à distribuição normal padrão $N(0,1)$.

Considere agora outra maneira de se obter os resíduos para dados circulares. A distância circular para a medida observada $\theta_i$ e ajustada $\hat{\theta}_j$, é dada por

$$d_j = \pi - |\pi - |\theta_j - \hat{\theta}_j||$$

O Erro Circular Médio (MCEs) (RAMBLI et al., 2016) dado por:

$$MCEs = \frac{1}{n}\sum_{j=1}^{n} \sin(\frac{d_j}{2}) \tag{2.84}$$

sendo $n$ o tamanho da amostra e MCEs $\in [0,1]$.

Rambli et al.(2016) considerando o modelo de regressão circular-circular proposto por Downs e Mardia (2002,), e dada pela equação ( 2.74), para a variável resposta, propuseram uma medida chamada DMCEs, que é a máxima diferença absoluta entre os dados completos e reduzidos, ou seja:

$$DMCEs = max_j\{|MCEs - MCEs_{(-j)}|\} \tag{2.85}$$

Onde $MCEs$ e $MCEs_{(-j)}$ são relativos aos dados completos e aos dados sem a j-ésima observação, respectivamente. Se o valor de $DCMEs$ for superior a um determinado valor de corte a j-ésima observação é um outlier.

Com relação ao modelo linear-circular, como a variável resposta pode ser considerada como tendo a distribuição normal, os resíduos poderiam ser os mesmos utilizados nos modelos de regres-

são linear. No entanto, alguns autores recentemente, tem proposto outras medidas para avaliar a qualidade do ajuste (SADIKON et al., 2018).

Com os dados modelados pelo modelo linear-circular Sadikon et al. (2018) propuseram uma medida baseada no conceito dos K- vizinhos mais próximos. Como o resíduo é dado pela diferença entre o valor real e o valor ajustado, então

$$e_i = y_i - \hat{y}_j \tag{2.86}$$

Supondo que a distância euclidiana $d$ entre os resíduos $e_i$ e $e_j$ pode-se definir por

$$d(e_i, e_j) = |e_i - e_j|, \quad i, j = 1, ..., n \quad i \neq j \tag{2.87}$$

ordenando as distâncias $d(x_i, x_1), ..., d(x_i, x_n)$, obtem-se $d_{(1)}(x_i, x_1), ..., d_{(k)}(x_i, x_k)$. A primeira distância mais próxima é a menor distância.

Uma outra medida para os resíduos é baseada no resíduo quantílico (ANDRADE; PEREIRA; ARTES, 2022), o qual foi proposto para o modelo circular - linear, e foi utilizado para situações em que a variável resposta segue outras distibuições além da von Mises.

Considerando $\theta$ a variável circular e $\tau$ o vetor de parâmetros, o resíduo é dado por

$$r^*_{qi} = \Phi^{-1}\{\hat{F}^*(\theta_i; \hat{\tau}_i)\} \tag{2.88}$$

No qual $F^*(\theta_0; \tau) = P(Md - \pi \leq \theta \leq \theta_0)$ é a funcão de distribuição, $Md$ é mediana direcional e $\hat{F}^*$ é o estimador de $F^*$ avaliado na mediana direcional $\hat{Md}_i$.

## 2.5 Análise de Sobrevivência

A análise de sobrevivência pode ser entendida como um conjunto de procedimentos estatísticos que são úteis para realizar a análise de dados nos quais o variável de interesse é o tempo até que um determinado evento ocorra (KLEINBAUM; KLEIN, 2012). O tempo pode ser dado em anos, meses, dias, horas e é contado a partir do início do acompanhamento de um indíviduo

até que o evento que está sendo analisado ocorra. O evento pode ser : morte, doenças, início dos atendimentos de acidentes, por exemplo.

Considerando T uma variável aleatória, não negativa, que descreve o tempo t até que um determinado evento ocorra. O tempo que vai do início da observação até a ocorrência do evento é chamado de tempo de falha.

Alguns dados podem estar com as informações parciais ou incompletas, quando isto acontece dizemos que estes dados são censurados. Estas informações, mesmo censuradas, podem ser relevantes para a pesquisa, e não levar em conta estes dados pode conduzir a conclusões equivocadas.

De acordo com Colosimo e Giolo (2006), a censura pode ser classificada como censura do tipo I, esta ocorre quando o estudo termina após o período que havia sido estabelecido previamente. A censura do tipo II é aquela em que o estudo termina depois que o evento que está sendo estudado ocorreu em um número pré-estabelecido de indivíduos. Além dessa, há um terceiro tipo de censura quando o indivíduo é retirado do estudo antes que o tempo de falha tenha ocorrido.

Colosimo e Giolo (2006) também afirmam que há a censura intervalar. Neste caso o tempo exato de falha não é conhecido, mas sabemos que ele ocorreu, em algum momento, dentro do intervalo $(L,U]$, onde $L \leq t \leq U$, onde $t$ é o tempo de falha.

Os dados relativos a falha e a censura serão representados pelo par $(t_i, \delta_i)$, onde $\delta_i$ é a função indicadora de falha ou censura, seja

$$\delta_i = \begin{cases} 0, \text{ se o dado é censurado} \\ 1, \text{ se o dado não é censurado} \end{cases} \tag{2.89}$$

Outro conceito importante é a função de sobrevivência, representada por $S(t)$. A função de sobrevivência é definida como sendo a probabilidade de um indivíduo sobreviver mais que um tempo especificado $t$, ou seja, é a probabilidade de que a variável aleatória $T$ seja maior que um tempo especificado $t$, isto é, $S(t) = P(T \geq t)$. Então

$$S(t) = P(T \geq t) = 1 - F(t) \tag{2.90}$$

onde $F(t)$ é a função densidade acumulada de T com função densidade de probabilidade $f(t)$.

Note que

$$f(t) = -\frac{dS(t)}{dt} \tag{2.91}$$

Vale mencionar também o conceito de função de risco (ou taxa de falha), denotada por $\lambda(t)$, esta função é dada por

$$\lambda(t) = \lim_{\Delta \to 0} \frac{P(t \leq T < t + \Delta t) | T \geq t}{\Delta t} \tag{2.92}$$

A função de risco $\lambda(t)$ dá a maneira em que a taxa instantânea de falha é modificada com o tempo.

A função de risco acumulada é dada por

$$\Lambda(u) = \int_0^t \lambda(u)\, du \tag{2.93}$$

Algumas relacões podem ser estabelecidas entre as funções consideradas anteriormente, podemos definir a função risco como sendo $\lambda(t) = \frac{f(t)}{S(t)}$, portanto

$$\lambda(t) = -\frac{d(\log(S(t))}{dt} \tag{2.94}$$

$$\Lambda(t) = -log(S(t)) \tag{2.95}$$

e

$$S(t) = exp\{-\int_0^t \lambda(u)\, du\} \tag{2.96}$$

Considerando $\delta_i$ como a função indicadora de falha ou censura e $S(t)$ a função de sobrevivência, a função de verossimilhança será representada por:

$$L(\theta) = \prod_{i \in O} f(t_i) \prod_{i \in D} S(t_i) \prod_{i \in E} [1 - S(t_i)] \tag{2.97}$$

em que $O$ é o conjunto de observaçãoes que sofreram o evento, $E$ o conjunto de observações censuradas à esquerda e $D$ as observações censuras à direita (CARVALHO et al., 2011).

Quando ocorrer censura à direita a função de verossimilhança será dada por

$$L(\theta) = \prod_{i=1}^{n} [f(t_i, \theta)]^{\delta_i} [S(t_i, \theta)]^{1-\delta_i} \tag{2.98}$$

e quando ocorrer censura à esquerda a verossimilhança será

$$L(\theta) = \prod_{i=1}^{n} f(t_i, \theta)^{\delta_i} \prod_{i=1}^{n} (1 - S(t_i, \theta))^{1-\delta_i} \tag{2.99}$$

**Estimador de Kaplan-Meier**

Considere os $k$ tempos de falhas ordenados e distintos, $t_1 < t_2 < ... < t_k$, e $d_j$ o número de falhas em $t_j$, com $j = 1, ..., n$ e $n_j$ é o número de indivíduos que estão sob risco em $t_j$ (COLOSIMO; GIOLO, 2006). O estimador de Kaplan-Meier é dado por:

$$\hat{S}(t) = \prod_{j:t_j<t} \left( \frac{n_j - d_j}{n_j} \right) = \prod \left( 1 - \frac{d_j}{n_j} \right) \tag{2.100}$$

**Alguns modelos probabilísticos**

Serão mencionados nesta secção alguns modelos de distribuição mais frequentemente usados na análise de sobrevivência.

**Distribuição Exponencial**

É um dos modelos mais simples para descrever o tempo de falha. Neste modelo a função taxa de falha é constate, isto é para $t > 0$ tem-se que

$$\lambda(t) = \alpha \tag{2.101}$$

Como a taxa de falha é mantida constante ao longo do tempo, a distribuição exponencial tem uma propriedade conhecida como de falta de memória (LAWLESS, 2003).

A função densidade de probabilidade para a variável aleatória $T$ é dada por

$$f(t) = \alpha \exp(-\alpha t) \tag{2.102}$$

e a função de sobrevivência é

$$S(t) = \exp\{-\alpha t\} \tag{2.103}$$

**Distribuição de Weibull**

A sua função densidade de probabilidade da distribuição Weibull é

$$f(t) = \alpha\gamma(\alpha t)^{\gamma-1} \exp[-(\alpha t)^{\gamma}] \tag{2.104}$$

a função de Sobrevivência é

$$S(t) = \exp[-(\alpha t)^{\gamma}] \tag{2.105}$$

e a função taxa de falha (função de risco) é

$$\lambda(t) = \alpha\gamma(\alpha t)^{\gamma-1} \tag{2.106}$$

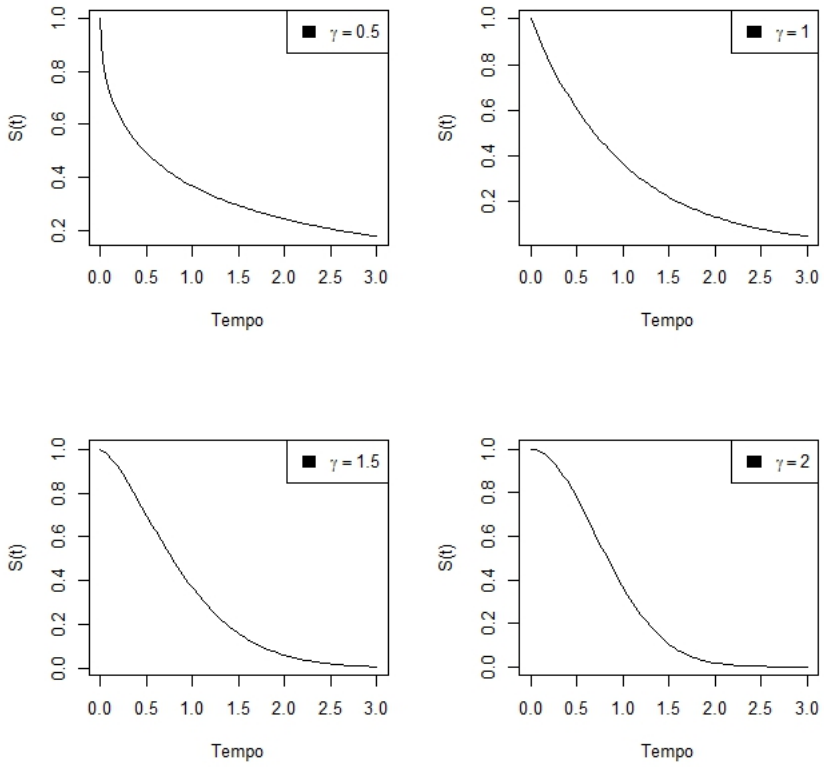
Com relação à função de risco (taxa de falha)(COLOSIMO; GIOLO, 2006) tem-se que:

a) Quando $\gamma > 1$, a taxa de falha ($\lambda(t)$) é crescente;
b) Quando $\gamma < 1$, a taxa de falha é decrescente;
c) Quando $\gamma = 1$, a taxa de falha é constante.

$\gamma$ é o parâmetro de forma e $\alpha$ o parâmetro de escala. Observe que se $\gamma = 1$ tem-se a distribuição exponencial, ou seja a distribuição exponencial é um caso particular da distribuição Weibull.

A figura ( 2.22) apresenta o gráfico da função de sobrevivência da distribuição Weibull, considerando o parâmetro de escala $\alpha = 1$ e diversos valores para o parâmetro de forma $\gamma = 0,5; 1; 1,5; 2$

Figura 2.22 – Gráficos da função de sobrevivência da distribuição Weibull para $\alpha = 1$ e $\gamma = 0,5; 1; 1,5; 2$




Fonte: O autor (2022)

**Distribuição Log- Normal**

A função densidade de probabilidade de uma variável aleatória $T$é dada por

$$f(t) = \frac{1}{\sigma t\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right), \quad t > 0$$

$\mu$ é a média do logarítmo do tempo de falha e $\sigma$ é o desvio padrão.

A função de sobrevivência é dada por:

$$S(t) = \Phi\left(\frac{-\log(t)+\mu}{\sigma}\right) \tag{2.107}$$

Em que $\Phi$ é a função distribuição acumulada de uma normal padrão.

**Distribuição Gama e Gama Generalizada**

A função de densidade da distribuição gama possui dois parâmetros: o parâmetro de forma $k$ e o de escala $\alpha$ e é dada por:

$$f(t) = \frac{t^{k-1}}{\Gamma(k)\alpha^{\kappa}} \exp\left(-\frac{t}{\alpha}\right) \tag{2.108}$$

Em que $\Gamma(k)$ é a função gama dada por

$$\Gamma(k) = \int_0^{\infty} \exp(-x)x^{k-1}dx$$

A função de sobrevivência é dada por

$$S(t) = \int_t^{\infty} \frac{u^{k-1}}{\Gamma(k)\alpha^{k}} \exp\left(-\frac{u}{\alpha}\right) du \tag{2.109}$$

Uma extensão da distribuição Gama é a distribuição Gama Generalizada que é caracterizada por três parâmetros positivos $\gamma$, $k$ e $\alpha$. A função densidade de probabilidade é dada por:

$$f(t) = \frac{\gamma}{\Gamma(k)\alpha^{\gamma k}} t^{k\gamma-1} \exp\left(-\left(\frac{t}{\alpha}\right)^{\gamma}\right) \tag{2.110}$$

Com isso tem-se que

a) Se $\gamma = k = 1$ então T $\sim$ Exp($\alpha$);
b) Se $k = 1$ então T$\sim$ Weibull($\gamma, \alpha$);
c) Se $\gamma = 1$ então T$\sim$ Gama($k, \alpha$).

Para estimar os parâmetros dos modelos mencionados é usado o método da máxima verossimilhança, porém nem sempre tem-se uma solução algébrica. Sendo necessária a utilização de métodos numéricos para obtenção das soluções.

Vale destacar que existem outros modelos distribuições para modelar o tempo de falha, como : log-logístico, Rayleigh, Weibull exponencial (MUDHOLKAR; SRIVASTAVA, 1993), Weibull inversa. E com relação a dados circulares algumas distribuições de probabilidade foram pro-

postas (BAKOUCH; CHESNEAU; LEAO, 2018; SOUZA et al., 2019; KUMAR; SINGH; SINGH, 2015).

**Modelos de regressão na Análise de sobrevivência**

No contexto da análise de sobrevivência a variável resposta, o tempo até a ocorrência de um evento, segue uma distribuição de probabilidade, que é uma função contínua, não negativa e assimétrica. Considerando, por exemplo, o modelo $T = \exp\{\beta_0 + \beta_1 x\}\varepsilon$ é considerado como não linear, pois as derivadas parciais em relação aos parâmetros do modelo, dependem de $\beta_0$ e $\beta_1$ de fato

$$\frac{\partial T}{\partial \beta_0} = \exp\{\beta_0 + \beta_1 x\}\varepsilon$$

$$\frac{\partial T}{\partial \beta_1} = \exp\{\beta_0 + \beta_1 x\}.x\varepsilon$$

No entanto, o modelo pode ser linearizado, da seguinte maneira

$$y = \log T = \beta_0 + \beta_1 x + \nu \tag{2.111}$$

onde $\nu = \log \varepsilon$, Note que neste modelo a distribuição dos erros não é necessariamente normal.

O modelo ( 2.111) pode ser generalizado, incluindo um parâmetro de forma $\sigma$. O tempo T é dado por:

$$\log T = \mu + \sigma\nu \tag{2.112}$$

em que $\mu$ é o parâmetro de locação e $\sigma$ o parâmetro de escala. $\sigma$ é uma variável aleatória que segue alguma distribuição de probabilidade que possa representar $y = \log T$ (CARVALHO et al., 2011; COLOSIMO; GIOLO, 2006).

Vale mencionar que este modelo é log-linear para $T$ e um modelo linear para $Y$, e que o vetor de covariáveis tem um efeito multiplicativo em $T$, uma vez que $T = \exp\{\beta_0 + \beta_1 x\}\exp\{\sigma\nu\}$. Os modelos paramétricos para o tempo de sobrevivência que podem ser linearizados, usando o

logaritmo natural são chamados de modelo de tempo de vida acelerado, pois as covariáveis têm a função de acelerar ou desacelerar o tempo de vida (HOSMER; LEMESHOW; MAY, 2008).

Sendo $\mathbf{X}^T = (1, x_1, ..., x_p)$ o vetor de covariáveis e $\beta = (\beta_0, \beta_1, ..., \beta_n)^T$ o vetor de parâmetros que serão estimados.

Alguns modelos serão considerados.

**Modelo de regressão exponencial**

A função de sobrevivência do modelo de regressão exponencial é dado por:

$$S(t|x) = \exp\left(-\left(\frac{t}{\exp\{x^T\beta\}}\right)\right) \tag{2.113}$$

e a função de risco será

$$\lambda(t|x) = \exp\{x^T\beta\} \tag{2.114}$$

**Modelo de regressão Weibull**

Considerando o tempo de sobrevivência $T$, e $\alpha = \exp\{x^T\beta\}$, onde $x$ é o vetor de covariáveis e $\beta$ é o vetor de parâmetros, a função de sobrevivência é dada por:

$$S(t|x) = \exp\left(-\left(\frac{t}{\exp\{x^T\beta\}}\right)^{\frac{1}{\sigma}}\right) \tag{2.115}$$

Considerando $Y = \log T$, a função de sobrevivência será:

$$S(y|x) = \exp\left(-\exp\left(\frac{y - x^T\beta}{\sigma}\right)\right) \tag{2.116}$$

**Modelo de regressão log-normal**

Considerando $Y = \log T$, as funções de densidade de probabilidade da log normal e de sobrevivência são dadas, respectivamente, por

$$f(y|x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y - x^T\beta}{\sigma}\right)^2\right) \tag{2.117}$$

$$S(y|x) = 1 - \Phi\left(\frac{y - x^T\beta}{\sigma}\right) \tag{2.118}$$

**Critério para seleção do modelo**

Alguns critérios para a seleção do modelo que serão utilizados para o ajuste dos dados são propostos, entre estes o critério de informação de Akaike (AIC) e o critério de informação Bayesiano (BIC).

O critério de informação de Akaike (AIC) é dado por

$$AIC = -2\log(L) + 2p$$

e o critério de informação Bayesiano (BIC) é dado por

$$BIC = -2\log(L) + p\log(n)$$

em que $L$ é a função de verossimilhança estimada nos valores dos parâmetros do modelo, $p$ é o número de parâmetros e $n$ é o tamanho da amostra. Será indicado como melhor modelo o que apresentar menor valor de AIC e BIC.

**Modelo de Cox**

O modelo de Cox é utilizado para modelar a função de risco, considerando que o efeito das covariáveis na função é dado por

$$\lambda(t|x) = \lambda_0(t)\exp(\beta_1 x_1 + ... + \beta_n x_n) = \lambda_0(t)\exp(\mathbf{x}^T\beta) \tag{2.119}$$

$\lambda_0(t)$ é chamado de risco basal ou função de taxa de falha de base, pois $\lambda(t|\mathbf{x}) = \lambda_0(t)$ quando $\mathbf{x = 0}$.

O modelo é composto pelo produto de um componente não-paramétrico $\lambda_0(t)$ e um paramétrico $\exp(\beta_1 x_1 + ... + \beta_n x_n)$, sendo por isso considerado um modelo semi- paramétrico.

O modelo de Cox também é denominado modelo de riscos proporcionais, pois a razão das taxas de falhas de dois indivíduos diferentes $i$ e $j$, com covariáveis $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ e $\mathbf{x}_j = (x_{j1}, ..., x_{jp})$ é constante, ou seja a razão das funções taxa de falha entre dois indivíduos $i$ e $j$ não depende do tempo $t$ (COLOSIMO; GIOLO, 2006). De fato,

$$\frac{\lambda(t|\mathbf{x}_i)}{\lambda(t|\mathbf{x}_j)} = \frac{\lambda_0(t)\exp(\mathbf{x}_i^T\beta)}{\lambda_0(t)\exp(\mathbf{x}_j^T\beta)} = \frac{\exp(\mathbf{x}_i^T\beta)}{\exp(\mathbf{x}_j^T\beta)}$$

No modelo de Cox pressupõe-se que as taxas de falhas são proporcionais, e por ser semi-paramétrico é bastante flexível. Este modelo é bastante utilizado na medicina.

**Análise de Resíduos na Análise de Sobrevivência**

Nos modelos de regressão linear os resíduos podem ser definidos como a diferença entre os dados reais e o valor ajustado, porém devido a presença da censura, os resíduos na análise de sobrevivência não são vistos desta maneira (HOSMER; LEMESHOW; MAY, 2008).

**Resíduos de Schoenfeld**

Os resíduos de Schoenfeld são utilizados para se investigar se os riscos são proporcionais, o resíduo é dado por:

$$r_{ik} = \delta_i(x_{ik} - a_{ik}) \tag{2.120}$$

sendo $\delta_i$ a função indicadora de falha ou censura e $a_{ik}$ é dada da seguinte maneira:

$$a_{ik} = \frac{\sum_{j \in R(t_i)} x_{jk} \exp(x_j \hat{\beta})}{\sum_{j \in R(t_i)} \exp(x_j \hat{\beta})} \tag{2.121}$$

$R(t_i)$ é o conjunto dos indivíduos que estão sob risco no tempo $t_i$ e $x_{ik}$ é o valor da k-ésima covariável do j-ésimo indivíduo.

O resíduo padronizado de Schoenfeld no tempo $t_i$ é dado por:

$$r_i^* = [\hat{V}(r_i)]^{-1} r_i \tag{2.122}$$

sendo $V(r_i)$ a matriz de covariância estimada do vetor de resíduos de Schoenfeld $r_i = (r_{i1}, ..., r_{ik})$.

O gráfico dos resíduos padronizados de Schoenfeld contra os tempos de sobrevivência permite verificar se há indicação de não proporcionalidade, se for satisfeita a suposição de proporcionalidade não deve existir alguma tendência no gráfico de $r*_{ik}$ versus tempo de sobrevivência, sendo portanto uma reta horizontal (CARVALHO et al., 2011).

**Resíduos de Cox-Snell**

Os resíduos de Cox-Snell auxiliam na avaliação do ajuste global do modelo e eles são dados por:

$$\hat{e}_i = \hat{\Lambda}(t_i|x_i) \tag{2.123}$$

em que $\Lambda(\cdot)$ é a função de taxa de falha acumulada. Os resíduos de Cox - Snell, de acordo com o modelo de regressão considerado, podem ser dados por:

a) Para o modelo exponencial

$$\hat{e}_i = t_i \exp[-X_i'\hat{\beta}]$$

b) Para o modelo Weibull

$$\hat{e}_i[t_i \exp\{-X_i'\hat{\beta}\}^{\gamma}$$

c) Para o modelo Log-normal

$$\hat{e}_i = -\log\left[1 - \Phi\left(\frac{\log(t_i) - X_i'(\hat{\beta})}{\hat{\sigma}}\right)\right]$$

Os resíduos de Cox-Snell $\hat{e}_i$ vêm de uma população homogênea, e se o modelo for adequado para representar os dados, os resíduos devem seguir uma distribuição exponencial padrão (COLOSIMO; GIOLO, 2006)

**Resíduos Martingale**

Os resíduos martingale são baseados em processos de contagem indivídual e são dados por:

$$M_i = \delta_i - \hat{e}_i \tag{2.124}$$

sendo $\delta_i$ a função indicadora de falha ou censura e $\hat{e}_i$ é o resíduo de Cox-Snell. Vale mencionar que os resíduos $M_i$ não são simetricamente distribuídos em torno do 0 (zero) e quando o tempo de sobrevivêcia é censurado o resíduo é negativo (CARVALHO et al., 2011).

**Resíduos Deviance**

Os resíduos deviance são definidos por

$$r_{D_i} = \text{sinal}(\hat{M}_i)\sqrt{-2(\tilde{l}_i - l_i)} \tag{2.125}$$

em que o sinal de $(\hat{M}_i)$ é o sinal do resíduo martingale para o i-ésimo indivíduo, sendo $\tilde{l}_i$ e $l_i$ os valores do logaritmo da função de verossimilhança para a observação $i$ do modelo dado e do modelo

saturado, respectivamente. O resíduo deviance deve estar distribuído simetricamente em torno da reta $y = 0$ (CARRASQUINHA; VERÍSSIMO; VINGA, 2018).

**Modelos para eventos múltiplos**

Ao analisar alguns dados pode ser verificado a ocorrência de um mesmo evento, para um indivíduo, várias vezes durante o período de acompanhamento (HOSMER; LEMESHOW; MAY, 2008) ou diferentes tipos de eventos decorrentes de um mesmo fator de risco em estudo podem ocorrer (CARVALHO et al., 2011).

Se o uso de um medicamento causa efeitos colaterais num mesmo indivíduo, em um determinado período de estudo, isto pode ser considerado eventos múltiplos, ou ainda, podem ocorrer as mesmas reações adversas em vários períodos de tempo. Estas reações podem ser independentes, ou certa reação ocorre apenas após outra reação. O que caracteriza também eventos múltiplos.

Há vários modelos para a análise dos eventos múltiplos, como o modelo AG (Andersen-Gill), proposto por Andersen e Gill (ANDERSEN; GILL, 1982) e o que pode ser considerado como uma extensão do modelo de Cox é dado por:

$$\lambda_i(t) = \lambda_0(t)\exp[\mathbf{x}_i^T(t)\beta] \tag{2.126}$$

Este modelo é semelhante ao modelo de Cox, equação ( 2.119), a diferença está no fato de que no modelo de Cox o indivíduo não está mais em risco, quando o evento ocorre e no modelo AG o índivíduo continua no grupo de risco, mesmo com a ocorrência do evento, pois o evento pode ocorrer mais de uma vez (COLOSIMO; GIOLO, 2006). O risco basal $\lambda_0(t)$ não varia e os eventos são considerados independentes ao longo do tempo (CARVALHO et al., 2011).

No modelo PWP (Prentice-Williams-Peterson) há uma estrutura de dependência entre os tempos dos indivíduos, isto significa que o indivíduo só vai estar em risco para o evento $m$ após a ocorrência do evento $m-1$ (CARVALHO et al., 2011), a função taxa de falha é dada por:

$$\lambda_{im} = \lambda_{0m}(t)\exp\{\mathbf{x}_i^T\beta_m\} \tag{2.127}$$

em que $\lambda_{ij}$ é o risco do indivíduo $i$ sofrer o evento $j$

No caso do modelo WLW (Wei-Lin-Weissfel) é possível obter uma taxa de falha para cada indivíduo e ele permanece sob risco para o $j$-ésimo evento até a ocorrência deste evento (COLOSIMO; GIOLO, 2006). A formulação da função taxa de falha para o j-ésimo evento do i-ésimo índivíduo é idêntica ao modelo PWP, equação (2.127).

## 2.6 Distribuições circulares com censura

Nesta secção serão analisados os procedimentos para se obter os estimadores das distribuições Cardióide, von Mises e Wrapped Cauchy, com a presença da censura. O método usado para obter os estimadores será o da máxima verossimilhança. Alguns autores fazendo uso do conceito de censura, dados incompletos ou faltantes, discutiram sobre este aspecto na estatística circular (DEVARAAJ; GIRIJA; RAO, 2014; JAMMALAMADAKA; MANGALAM, 2009).

### Distribuição Cardióide

Considerando as variáveis circulares $\theta_1, ..., \theta_n$ e os parâmetros $\mu$ e $\rho$ da distribuição Cardióide, utilizando método da máxima verossimilhança, estes parâmetros serão estimados para o caso de censura à direita. Lembrando que $\delta_i$ é a função indicadora de falha ou censura

$$\delta_i = \begin{cases} 0, \text{se o dado é censurado} \\ 1, \text{se o dado é não censurado} \end{cases}$$

A função densidade de probabilidade da distribuição cardióide é dada por:

$$f(\theta, \mu, \rho) = \frac{1}{2\pi}(1 + 2\rho\cos(\theta - \mu)) \tag{2.128}$$

e a função de distribuição é

$$F(\theta) = \frac{1}{2\pi}\left[\theta + 2\rho\sin(\theta - \mu) + 2\rho\sin\mu\right] \tag{2.129}$$

Para o caso de censura à direita a função de verossimilhança $L = L(\theta_i; \mu, \rho)$ é dada por

$$\begin{aligned} L &= \prod_{i=1}^{n} \left( \frac{1}{2\pi}(1 + 2\rho\cos(\theta_i - \mu)) \right)^{\delta_i} \\ &\cdot \left( 1 - \frac{1}{2\pi}(\theta + 2\rho\sin(\theta - \mu) + 2\rho\sin\mu) \right)^{1-\delta_i} \end{aligned} \tag{2.130}$$

Então

$$\begin{aligned} L &= \prod_{i=1}^{n} \frac{1}{2\pi} \left( (1 + 2\rho\cos(\theta_i - \mu))^{\delta_i} \right. \\ &\left. (2\pi - (\theta + 2\rho\sin(\theta - \mu) + 2\rho\sin\mu))^{1-\delta_i} \right. \end{aligned} \tag{2.131}$$

Encontrando o log da verossimilhança, $l = l(\theta_i; \mu, \rho) = \log L(\theta_i; \mu, \rho)$ tem-se

$$l = \log\frac{1}{2\pi} + \sum_{i=1}^{n} \delta_i \log(1 + 2\rho\cos(\theta_i - \mu) + \sum_{i=1}^{n} (1 - \delta_i) \log\left(2\pi - (\theta_i + 2\rho\sin(\theta_i - \mu) + 2\rho\sin\mu)\right)$$

Em seguida são obtidas as derivadas parciais $\frac{\partial l}{\partial \mu}$ e $\frac{\partial l}{\partial \rho}$

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= \sum_{i=1}^{n} \delta_i \left( \frac{2\rho\sin(\theta_i - \mu)}{1 + 2\rho\cos(\theta_i - \mu)} \right) \\ &+ \sum_{i=1}^{n} (1 - \delta_i) \left( \frac{2\rho(\cos(\theta_i - \mu) - \cos\mu)}{2\pi - (\theta_i + 2\rho\sin(\theta_i - \mu) + 2\rho\sin\mu)} \right) \end{aligned} \tag{2.132}$$

$$\begin{aligned} \frac{\partial l}{\partial \rho} &= \sum_{i=1}^{n} \delta_i \left( \frac{2\cos(\theta_i - \mu)}{1 + 2\rho\cos(\theta_i - \mu)} \right) \\ &- \sum_{i=1}^{n} (1 - \delta_i) \left( \frac{2(\sin(\theta_i - \mu) + \sin\mu)}{2\pi - (\theta_i + 2\rho\sin(\theta_i - \mu) + 2\rho\sin\mu)} \right) \end{aligned} \tag{2.133}$$

Igualando as equações ( 2.132) e ( 2.133) a zero é possível encontrar as estimativas dos estimadores $\hat{\mu}$ e $\hat{\rho}$.

**Distribuição Wrapped Cauchy**

Considere as variáveis circulares $\theta_1, ..., \theta_n$, os parâmetros $\mu$ e $\rho$ da distribuição Wrapped Cauchy, também podem ser estimados utilizando método da máxima verossimilhança para o caso de censura à direita. Sendo $\delta_i$ a função indicadora de censura, a função de verossimilhança $L = L(\mu, \rho; \theta_i)$ é dado por :

$$L = \prod_{i=1}^{n} \left[ \left( \frac{1}{2\pi} \frac{1-\rho^2}{1+\rho^2-2\rho\cos(\theta_i-\mu)} \right)^{\delta_i} \left( 1 - \frac{1}{2\pi} \arccos\left( \frac{(1+\rho^2)\cos(\theta_i-\mu)-2\rho}{1+\rho^2-2\rho\cos(\theta_i-\mu)} \right) \right)^{1-\delta_i} \right. \quad (2.134)$$

$$L = \prod_{i=1}^{n} \left( \frac{1}{2\pi} \frac{1-\rho^2}{1+\rho^2-2\rho\cos(\theta_i-\mu)} \right)^{\delta_i} \left( \frac{1}{2\pi} \right)^{1-\delta_i} \left( 2\pi - \arccos\left( \frac{(1+\rho^2)\cos(\theta_i-\mu)-2\rho}{1+\rho^2-2\rho\cos(\theta_i-\mu)} \right) \right)^{1-\delta_i} \quad (2.135)$$

Com isso

$$L = \prod_{i=1}^{n} \frac{1}{2\pi} \left( \frac{1-\rho^2}{1+\rho^2-2\rho\cos(\theta_i-\mu)} \right)^{\delta_i} \cdot \left( 2\pi - \arccos\left( \frac{(1+\rho^2)\cos(\theta_i-\mu)-2\rho}{1+\rho^2-2\rho\cos(\theta_i-\mu)} \right) \right)^{1-\delta_i} \quad (2.136)$$

Obtendo agora a função log verossimilhança

$$
\begin{aligned}
l &= \log\frac{1}{2\pi} + \sum_{i=1}^{n}\delta_i \log\left(\frac{1-\rho^2}{1+\rho^2-2\rho\cos(\theta_i-\mu)}\right) \\
&\quad + \sum_{i=1}^{n}(1-\delta_i)\log\left(\frac{(1+\rho^2)\cos(\theta_i-\mu)-2\rho}{1+\rho^2-2\rho\cos(\theta_i-\mu)}\right) \qquad (2.137)
\end{aligned}
$$

Encontrando as derivadas parcias $\frac{\partial l}{\partial \mu}$ e $\frac{\partial l}{\partial \mu}$ da função ( 2.137) e igualando a zero, é possível encontrar as estimativas dos estimadores $\hat{\mu}$ e $\hat{\rho}$.

**Distribuição von Mises**

Considere os ângulos $\theta_1, \theta_2, ..., \theta_n$ e assumindo que eles seguem a distribuição von Mises com média $\mu$ e parâmetro de concentração $\kappa$, que é representada por $vM(\mu, \kappa)$. Suponha que algumas observações são censuradas por intervalos do tipo $(l, r)$, onde $l$ e $r$ são arcos na circunferência (JAMMALAMADAKA; MANGALAM, 2009).

A função de máxima verossimilhança será dada por :

$$
\begin{aligned}
L(\mu, \kappa, \theta_i, \theta_2, ..., \theta_n) &= \frac{1}{(2\pi I_0(\kappa))^n}\exp\left[\kappa\sum_{i=1}^{n}\delta_i\cos(\theta_i-\mu)\right] \\
&\quad \prod_{i=1}^{n}\left[\int_{l_i}^{r_i}\exp[\kappa\cos(t-\mu)]dt\right]^{1-\delta_i} \qquad (2.138)
\end{aligned}
$$

sendo que $\delta_i$ é a função indicadora de censura.

A função log verosssimilhança será dada por

$$
\begin{aligned}
l(\mu, \kappa, \theta_1, ..., \theta_n) &= -n(\log 2\pi + \log I_0(\kappa)) + \kappa\sum_{i=1}^{n}\delta_i\cos(\theta_i-\mu) \\
&\quad + \sum_{i=1}^{n}(1-\delta_i)\log\left[\int_{l_i}^{r_i}\exp[\kappa\cos(t-\mu)]dt\right] \qquad (2.139)
\end{aligned}
$$

O objetivo é encontrar os estimadores $\hat{\mu}$ e $\hat{\kappa}$, que maximizam a função de log verossimilhança. Algumas relações serão utilizadas (JAMMALAMADAKA; MANGALAM, 2009), para facilitar a obtenção dos estimadores:

$$A(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)}$$

$$B_{0i}(\mu,\kappa) = \int_{l_i}^{r_i} \exp[\kappa\cos(t-\mu)]\,dt$$

$$B_{1i}(\mu,\kappa) = \int_{l_i}^{r_i} \sin(t-\mu)\exp[\kappa\cos(t-\mu)]\,dt$$

$$B_{2i}(\mu,\kappa) = \int_{l_i}^{r_i} \cos(t-\mu)\exp[\kappa\cos(t-\mu)]\,dt$$

$$B_{3i}(\mu,\kappa) = \int_{l_i}^{r_i} \sin^2(t-\mu)\exp[\kappa\cos(t-\mu)]\,dt$$

$$B_{4i}(\mu,\kappa) = \int_{l_i}^{r_i} \cos^2(t-mu)\exp[\kappa\cos(t-\mu)]\,dt$$

$$B_{5i}(\mu,\kappa) = \int_{l_i}^{r_i} \sin(t-\mu)\cos(t-\mu)\exp[\kappa\cos(t-\mu)]\,dt$$

A partir destas relações obtem-se

$$\frac{\partial B_{0i}}{\partial \mu} = \kappa B_{1i}$$

$$\frac{\partial B_{0i}}{\partial \kappa} = B_{2i}$$

$$\frac{\partial B_{1i}}{\partial \mu} = \kappa B_{3i} - B_{2i}$$

$$\frac{\partial B_{1i}}{\partial \kappa} = B_{5i}$$

$$\frac{\partial B_{2i}}{\partial \kappa} = B_{4i}$$

Para determinar os estimadores são calculadas $\frac{\partial l}{\partial \mu}$ e $\frac{\partial l}{\partial \kappa}$. Logo

$$\frac{\partial l}{\partial \mu} = \kappa \sum_{i=1}^{n} \left[ \delta_i \sin(\alpha_i - \mu) + (1 - \delta_i) \frac{B_{1i}(\mu, \kappa)}{B_{0i}(\mu, \kappa)} \right] \tag{2.140}$$

e

$$\frac{\partial l}{\partial \kappa} = -nA(\kappa) + \sum_{i=1}^{n} \left[ \delta_i \cos(\alpha_i - \mu) + (1 - \delta_i) \frac{B_{2i}(\mu, \kappa)}{B_{0i}(\mu, \kappa)} \right] \tag{2.141}$$

Ao igualar a zero as equações ( 2.140) e ( 2.141) os estimadores $\hat{\mu}$ e $\hat{\kappa}$ são obtidos.

Vale ressaltar que as estimativas da máxima verossimilhança das três distribuições analisadas não são obtidas por métodos algébricos.

Nos modelos analisados nesta secção, foi considerada a censura do ponto de vista conceitual, dados incompletos ou faltantes, sem se considerar o tempo até a ocorrência do evento.

# 3 CONSIDERAÇÕES GERAIS

A partir da teoria estudada foram produzidos quatro artigos abordando os temas de estatística circular e análise de sobrevivência. São eles:

a) Estatística circular aplicada aos dados de localização dos municípios de origem dos alunos do PROFMAT-UFSJ-Campus Santo Antônio;
b) Regressão linear-circular para modelagem de dados meteorológicos na cidade de São João del Rei;
c) Modelos com eventos múltiplos para avaliação dos picos de radiação solar na cidade de São João del Rei- MG;
d) Modelo de Regressão com covariável circular: teoria e aplicação em dados sobre atendimento de ocorrência de acidentes de trânsito.

A princípio são artigos independentes, trabalhados inclusive com bancos de dados diferentes, que exigiram manipulação de planilhas de grandes dimensões com características próprias. A proposta foi apresentar a teoria abordada nos três primeiros artigos, utilizando diferentes aplicações e posteriormente no quarto artigo reunir as teorias em uma única proposta de adaptação de modelo de regressão para incluir dados circulares com censura.

No primeiro artigo, a utilização de coordenadas geográficas permitiu verificar a ampla abrangência do uso da estatística circular e trabalhar conceitos básicos. No segundo, a ideia da regressão circular foi abordada com um conjunto de dados tradicional na área, que são os dados meteorológicos. No entanto, percebe-se que grande parte dos artigos com tais dados trabalha apenas com a estatística circular descritiva e, recentemente, a inclusão com modelos de regressão tem sido realizada e sua teoria divulgada.

Já no terceiro artigo, a mudança está na forma de trabalhar os dados pois, utilizou-se o mesmo banco de dados meteorológicos mas agora com a teoria estatística da análise de sobrevivência, que usualmente trabalha com dados da área de saúde. Neste caso, a finalidade foi também apresentar uma maneira diferente para explorar esses dados considerando ainda uma adaptação do tradicional modelo de Cox, mas incluindo a teoria para os eventos repetidos. Apesar deste artigo não discutir dados circulares, ele foi utilizado para investigar a presença de dados censurados em um conjunto de dados comum na estatística circular, que são dados meteorológios.

Finalmente, ao obter os dados de ocorrências de acidentes de trânsito registrada pela Polícia Militar de Minas Gerais, foi possível analisar os dados tanto no contexto da estatística circular, quanto na abordagem da análise de sobrevivência, possibilitando propor um modelo não usual na literatura, que permite a associação entre variável resposta linear e covariável circular com censura. Seria possível ainda, neste conjunto de dados, explorar também eventos repetidos em diferentes regiões geográficas o que permitiria abordar todas as teorias anteriores em um único banco de dados.

# REFERÊNCIAS


ABUZAID, A. H.; MOHAMED, I. B.; HUSSIN, A. G. Boxplot for circular variables. **Computational Statistics**, v. 27, n. 3, p. 381–392, 2012.

ALLAHHAM, N. **On the simple angular regression model with Wrapped Cauchy error**. Tese (Doutorado) — Al Azhar University-Gaza, 2015.

ANDERSEN, P. K.; GILL, R. D. Cox's regression model for counting processes: a large sample study. **The annals of statistics**, p. 1100–1120, 1982.

ANDERSON-COOK, C. M.; NOBLE, R. B. An alternate model for cylindrical data. **Nonlinear Analysis**, v. 47, p. 2011–2022, 2001.

ANDRADE, A. C.; PEREIRA, G. H.; ARTES, R. The circular quantile residual. **Computational Statistics & Data Analysis**, p. 107612, 2022.

BAKOUCH, H. S.; CHESNEAU, C.; LEAO, J. A new lifetime model with a periodic hazard rate and an application. **Journal of Statistical Computation and Simulation**, v. 88, n. 11, p. 2048–2065, 2018.

BUSSAB, W. O.; MORRETIN, P. A. **Estatística básica**. 6. ed. São Paulo: Saraiva, 2010.

BUTTARAZZI, D.; PANDOLFO, G.; PORZIO, G. C. A boxplot for circular data. **Biometrics**, v. 74, n. 4, p. 1492–1501, 2018.

CARRASQUINHA, E.; VERÍSSIMO, A.; VINGA, S. Consensus outlier detection in survival analysis using the rank product test. **bioRxiv**, p. 1–27, 2018.

CARVALHO, M. S. et al. **Análise de sobrevivência: teoria e aplicações em saúde**. 2. ed. Rio de Janeiro: Editora FIOCRUZ, 2011.

COLOSIMO, E.; GIOLO, S. **Análise de sobrevivência aplicada**. São Paulo: Blucher, 2006.

DEVARAAJ, V. J.; GIRIJA, S.; RAO, A. D. Estimation of parameters in cardioid distribution from censored samples. **International Journal of Innovative Research in Science & Engineering**, v. 2, n. 11, p. 774–781, 2014.

DOWNS, T.; MARDIA, K. Circular regression. **Biometrika**, v. 89, n. 3, p. 683–697, 2002.

FISHER, N. **Statistical analysis of circular data**. New York: Cambridge University Press, 1993.

FISHER, N.; LEE, A. J. Regression models for an angular response. **Biometrics**, v. 48, n. 11, p. 665–677, 1992.

HOSMER, D.; LEMESHOW, S.; MAY, S. **Applied survival analysis: regression modeling of time-to-event data**. 2. ed. New Jersey: John Wiley & Sons, 2008.

JAMMALAMADAKA, S.; SENGUPTA, A. **Topics in circular statistics**. Singapore: World Scientific, 2001.

JAMMALAMADAKA, S. R.; MANGALAM, V. A general censoring scheme for circular data. **Statistical Methodology**, v. 9, p. 280–289, 2009.

KATO, S.; SHIMIZU, K.; SHIEH, G. A circular-circular regression model. **Statistica Sinica**, p. 633–645, 2018.

KLEINBAUM, D.; KLEIN, M. **Survival analysis - A self- learning text**. 3. ed. New York: Springer, 2012.

KUMAR, D.; SINGH, U.; SINGH, S. K. A new distribution using sine function-its application to bladder cancer patients data. **Journal of Statistics Applications & Probability**, v. 4, n. 3, p. 417–427, 2015.

LAWLESS, J. **Statistical models an methods for lifetime data**. 2. ed. New York: Jonh Willey and Sons, 2003.

LEAL, G.-M. G. **Análise de resíduos em modelos de regressão von Mises.** Dissertação (Mestrado em Matemática) — Universidade Federal de Campina Grande, 2006.

MAGALHÃES, M. N. **Probabilidade e variáveis aleatórias**. São Paulo: Edusp, 2006.

MARDIA, K.; JUPP, P. E. **Directional statistics**. London: Jonh Willey and Sons, 2000.

MARDIA, K.; SUTTON, T. W. A model for cylindrical variables with applications. **Royal Statistical Society**, v. 40, n. 2, p. 229–233, 1978.

MARDIA, K. V. **Statistics of directional data**. London: Academic Press, 1972.

MUDHOLKAR, G. S.; SRIVASTAVA, D. K. Exponentiated weibull family for analyzing bathtub failure-rate data. **IEEE transactions on reliability**, v. 42, n. 2, p. 299–302, 1993.

NUZZO, R. L. The box plots alternative for visualizing quantitative data. **American academy of physical medicine and rehabilitation**, v. 8, n. 3, p. 268–272, 2016.

PAULA, F. V. **Extended circular distributions: mathematical properties, inference and regression model**. Tese (Doutorado em Estatística) — Universidade Federal de Pernambuco, 2018.

RAMBLI, A. et al. Procedure for detecting outliers in a circular regression model. **PLoS One**, v. 11, n. 4, 2016.

RAO, A. D.; GIRIJA, S. **Angular statistics**. Boca Raton: Chapman and Hall/CRC, 2020.

SADIKON, N. H. et al. A new discordancy test on a regression for cylindrical data. **Sains Malaysiana**, v. 47, n. 6, p. 1319–1326, 2018.

SOUZA, F. A. M. **Influência local e análise de resíduos em modelos de regressão von Mises**. Tese (Doutorado em Estatística) — Universidade de São Paulo, 1999.

SOUZA, F. A. M.; PAULA, G. A. Deviance residuals for an angular response. **Australian & New Zealand Journal of Statistics**, v. 44, n. 3, p. 345–356, 2002.

SOUZA, L. et al. On the sin-g class of distributions: theory, model and application. **Journal of Mathematical Modeling**, University of Guilan, v. 7, n. 3, p. 357–379, 2019.

# SEGUNDA PARTE

# ARTIGO 1

## Estatística circular aplicada aos dados de localização dos municípios de origem dos alunos do PROFMAT - UFSJ - Campus Santo Antonio

# Estatística circular aplicada aos dados de localização dos municípios de origem dos alunos do PROFMAT-UFSJ - Campus Santo Antonio

**Clodoaldo Teodosio Santana da Silva** teoelania@gmail.com
Universidade Federal dos Vales do Jequitinhonha e Mucuri, Teófilo Otoni, MG, Brazil
Universidade Federal de Lavras, Lavras, MG, Brasil

**Carla Regina Guimarães Brighenti** carlabrighenti@ufsj.edu.br
Universidade Federal de São João del Rei, São João del Rei, MG, Brazil

**Luiz Fernando Silva Resende** luiz.resende@estudante.ufla.br
Universidade Federal de Lavras, Lavras, MG, Brazil

**Resumo**

O PROFMAT - Programa de Mestrado Profissional em Matemática em Rede Nacional- é um programa criado com o objetivo de contribuir com a qualificação dos professores da educação básica. O curso é ofertado em Rede Nacional sob coordenação da SBM (Sociedade Brasileira de Matemática) e o IMPA(Instituto de Matemática Pura e Aplicada) em parceira com 77 instituições de ensino superior em todo o Brasil, entre elas a Universidade Federal de São João del Rei- MG (UFSJ). Os alunos matriculados no curso são oriundos, em sua maioria, de municípios próximos à São João del Rei. Assim, estabelecer a abrangência, ou mesmo a distribuição de localização de origem dos alunos atendidos pelo programa no entorno da instituição de ensino superior parceira da SBM, pode ser um fator de interesse para gestão de coordenadores e, posterior divulgação entre os professores de matemática. Desta forma, ter acesso as coordenadas geográficas dos municípios de origem dos alunos fornece um conjunto de dados de interesse, sendo que tais dados podem ser transformados em dados angulares em relação à cidade sede do curso. Para realizar análises com dados angulares se faz necessário o uso da estatística circular, isto é, a análise considerando dados que não devem ser organizados simplesmente na reta real e sim estabelecendo uma disposição circular. Diante do exposto, o objetivo deste trabalho foi utilizar a estatística circular para analisar a distribuição de localização de origem dos alunos matriculados no PROFMAT na UFSJ, Campus Santo Antonio.

**Palavras-chave**

Média circular, Variância circular, Comprimento médio.

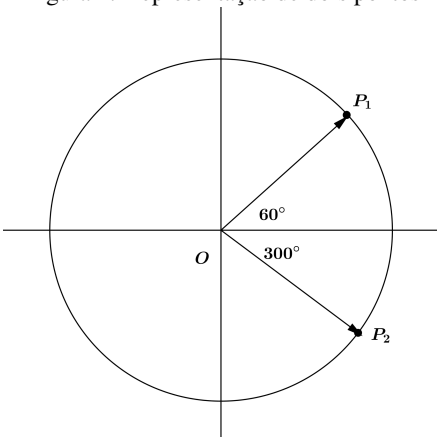
 

## 1 Introdução

Quando se pretende estudar fenômenos que são provenientes de medidas que envolvem ângulos como, direção dos ventos, ângulos de voos de pássaros, a melhor ferramenta é a estatística circular. No entanto muitas técnicas estatísticas são mais adequadas para dados localizados na reta real, como altura, idade e peso. Estes dados são considerados como lineares.

Contudo os dados circulares não devem ser analisados da mesma maneira como se tratam os lineares. O motivo desta afirmação está no fato de que se medidas angulares fossem tratadas como lineares seriam obtidos resultados equivocados. Por exemplo, a média aritmética dos ângulos $60^\circ$ e $300^\circ$ é $180^\circ$, no entanto observando a Figura 1, nota-se que o valor mais apropriado para esta média é $0^\circ$. Nota-se então que há vantagem de utilizar a estatística circular para que se obtenha a interpretação correta dos dados. Neste contexto os dados analisados são medidas angulares, ou que podem ser

Figura 1: Representação de dois pontos




**Fonte: Os autores**

transformadas em ângulos, como é o caso dos meses do ano e coordenadas geográficas. Já existem vários estudos utilizando dados circulares, pode ser mencionado o usa da estatística circular para analisar o efeito da direção dos ventos e a concentração de ozônio [3] e um outro estudo sobre a dispersão dos alunos matriculados na Universidade Federal do Recôncavo da Bahia- UFRB [7].

O objetivo desta pesquisa foi analisar, via estatística circular, a localização das cidades de origem dos alunos do programa de mestrado, PROFMAT, da UFSJ - Campus Santo Antônio, a partir das coordenadas geográficas. A relevância deste estudo está no fato que é possível verificar a distribuição dos alunos nas cidades situadas na região de atuação da UFSJ.

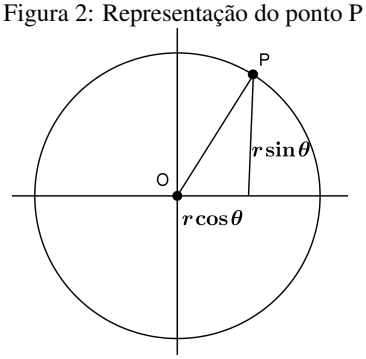
Isto tornará possível verificar em que localidades é mais importante divulgar o curso, tendo em vista a necessidade de qualificação dos professores da educação básica.

Para realização da pesquisa foram analisadas as 26 cidades mineiras de origem dos 56 alunos matriculados no PROFMAT no período de 2017- 2021.

## 2 Estatística Circular Descritiva

Um ponto $P$ do plano cartesiano pode ser representado por $(x, y)$, este ponto também pode ser visto como um ponto numa circunferência de centro na origem $(0, 0)$ e raio $r$, assim pode ser estabelecida uma relação : $x = r\cos\theta$ e $y = r\sin\theta$, então $P = (r\cos\theta, r\sin\theta)$ (Figura 2).

Figura 2: Representação do ponto P




**Fonte: Os autores**

No caso em que o raio é unitário, isto é, $r = 1$, tem-se $x = \cos\theta$ e $y = \sin\theta$, um ponto $P$ num círculo unitário pode ser representado por $P = (\cos\theta, \sin\theta)$.

A partir deste fato será dado o conceito da **Média direcional**, a qual tem a mesma direção que o centro de massa do circulo unitário ( ver [6]).

Assim, considere agora os pontos $P_1, P_2, ..., P_n$ em uma circunferência unitária, e seus respectivos ângulos $\theta_i, ..., \theta_n$ em relação ao eixo $x$. O centro de massa será dado por $M = (\bar{C}, \bar{S})$ onde

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n}\cos\theta_i \tag{1}$$

$$\bar{S} = \frac{1}{n}\sum_{i=1}^{n}\sin\theta_i \tag{2}$$

Seja

$$\bar{R}^2 = \bar{C}^2 + \bar{S}^2 \tag{3}$$

Note que das equações (1) e (2) tem-se que $C = \sum_{i=1}^{n} \cos\theta_i$ e $S = \sum_{i=1}^{n} \sin\theta_i$

A **média direcional**, $\bar{\theta}$, é a solução das equações :

$$\begin{cases} \cos\bar{\theta} = \dfrac{\bar{C}}{\bar{R}} \\ \\ \sin\bar{\theta} = \dfrac{\bar{S}}{\bar{R}} \end{cases} \tag{4}$$

e satisfaz as equações (5)

$$\bar{\theta} = \begin{cases} \arctan\dfrac{\bar{S}}{\bar{C}}, S > 0, C > 0 \\ \\ \arctan\dfrac{\bar{S}}{\bar{C}} + \pi, C < 0 \\ \\ \arctan\dfrac{\bar{S}}{\bar{C}} + 2\pi, S < 0, C > 0 \end{cases} \tag{5}$$

A medida $\bar{R}$ é chamada de **comprimento médio resultante**, observe que $\bar{R} = \dfrac{R}{n}$, com $0 \leq \bar{R} \leq 1$, e quanto mais próximo de 1 está o valor de $\bar{R}$ menos dispersos estarão os dados. A partir deste fato, pode-se obter uma medida para avaliar a dispersão dos dados. Se $\bar{R}$ tem um valor próximo de 1, então os dados direcionais estão bem agrupados, caso $\bar{R}$ esteja próximo de zero há uma dispersão dos dados. Notando assim, que $\bar{R}$ é uma medida de concentração dos dados( [5],[2]).

**A variância circular** $V$ é dada por

$$V = 1 - \bar{R} \tag{6}$$

onde $0 \leq V \leq 1$.

Observe que quanto mais próximo de 0(zero) for o valor na variância circular mais concentrados estão os dados e quanto mais próximo de 1 há uma dispersão dos dados.

O **desvio padrão circular** é dado por $\nu = [-2\ln(1-V)]^{\frac{1}{2}}$ ([6])

Pode também ser utilizado uma aproximação para o desvio padrão, $\nu$, que é dada por

$$\nu \approx (2V)^{\frac{1}{2}} \tag{7}$$

Considerando os valores de $\theta$ tais que $0 \leq \theta \leq 1$ rad, e usando e expansão de

Taylor (com apenas dois termos) para $\cos\theta$ (ver [4]) tem-se:

$$\cos\theta \approx 1 - \frac{\theta^2}{2} \tag{8}$$

De (8)

$$2\cos\theta \approx 2 - \theta^2$$

$$\theta^2 \approx 2(1 - \cos\theta)$$

$$\sum_{i=1}^{n} \theta_i^2 \approx 2\sum_{i=1}^{n}(1 - \cos\theta_i)$$

Com isso

$$\sum_{i=1}^{n}(\theta_i - \bar{\theta})^2 \approx 2\sum_{i=1}^{n}(1 - \cos(\theta_i - \bar{\theta})) \tag{9}$$

Antes de desenvolver a igualdade (9), será mostrado que

$$\sum_{i=1}^{n} \sin(\theta_i - \bar{\theta}) = 0 \tag{10}$$

$$\sum_{i=1}^{n} \cos(\theta_i - \bar{\theta}) = R \tag{11}$$

Usando relações trigonométricas e as equações (4) tem-se:

$$\sum_{i=1}^{n} \sin(\theta_i - \bar{\theta}) = \cos\bar{\theta}\sum_{i=1}^{n}\sin\theta_i - \sin\bar{\theta}\sum_{i=1}^{n}\cos\theta_i$$

$$= S\frac{\bar{C}}{\bar{R}} - C\frac{\bar{S}}{\bar{R}}$$

$$= S\frac{C}{R} - C\frac{S}{R} = 0$$

e

$$\sum_{i=1}^{n} \cos(\theta_i - \bar{\theta}) = \sum_{i=1}^{n}(\cos\theta_i\cos\bar{\theta} + \sin\theta_i\sin\bar{\theta})$$

$$= \cos\bar{\theta}\sum_{i=1}^{n}\cos\theta_i + \sin\bar{\theta}\sum_{i=1}^{n}\sin\theta_i$$

$$= C\cos\bar{\theta} + S\sin\bar{\theta} = \frac{C^2}{R} + \frac{S^2}{R} = R$$

Retornando à equação (9) :

$$\sum_{i=1}^{n}(\theta_i-\bar{\theta})^2 \approx 2\sum_{i=1}^{n}(1-\cos(\theta_i-\bar{\theta}))$$

$$\sum_{i=1}^{n}(\theta_i-\bar{\theta})^2 \approx 2n-2\sum_{i=1}^{n}\cos(\theta_i-\bar{\theta}) =$$

$$= 2(n-R) = 2(n-n\bar{R})$$

$$= 2n(1-\bar{R})$$

Então

$$\sum_{i=1}^{n}(\theta_i-\bar{\theta})^2 \approx 2n(1-\bar{R})$$

$$\frac{\sum_{i=1}^{n}(\theta_i-\bar{\theta})^2}{n} \approx 2(1-\bar{R})$$

$$\nu^2 \approx 2(1-\bar{R})$$

Portanto

$$\nu \approx \sqrt{2(1-\bar{R})} = \sqrt{2V} \tag{12}$$

De acordo com [5] para valores da variância circular $V$ próximos de zero, pode ser utilizada a aproximação dada em (12) para o desvio padrão circular $\nu$.

## 3 Metodologia

Os dados foram obtidos através das informações dos estudantes do PROFMAT registradas no Sistema Integrado de Gestão das Atividades Acadêmicas da Universidade Federal de São João Del Rei, localizada no município de São João Del Rei – MG. Para este trabalho, fez-se uso dos dados provenientes das cidades de origem dos estudantes matriculados no curso ofertado no campus Santo Antônio. Foi analisado o período de $2017-2021$, no qual há $56$ alunos matriculados oriundos de $26$ cidades mineiras.

Utilizou-se métodos estatísticos circulares para averiguar as frequências direcionais dos dados fornecidos. A partir das coordenadas geográficas de cada município fornecida pelo Instituto Brasileiro de Geografia Estatística (IBGE), foram obtidas as distâncias e direções para determinação das medidas angulares, tendo a cidade sede da UFSJ como município de referência.

Para o cálculo das distâncias, de acordo com [1] se faz necessário obter a distância entre dois pontos geográficos $P_1$ e $P_2$ se calculando então a diferença de latitude (DLA) e diferença de longitude (DLO). Para se obter os valores da DLA e DLO é necessário calcular a diferença entre as latitudes e longitudes entre as coordenadas de $P_1$ e $P_2$. Essas diferenças são aplicáveis para pequenas distancias, segundo [1] para ângulos

maiores que 7 graus deve-se levar em consideração a curvatura da terra. Calculada as diferenças, o próximo passo é transformar os valores angulares em distância, que generalizamos como D, porém para o cálculo das diferença de latitude e diferença de longitude chamaremos respectivamente de DLA e DLO.

$$D = [G(60) + M + \frac{S}{60}](1852) \tag{13}$$

A medida D está quantificada em graus(G), minutos(M) e segundos(S), o valor $1852$ é referente a transformação em Milhas Náuticas. Por meio da expressão (13) as diferenças (D) foram convertidas em distância, considerando que uma milha náutica (1 NM), equivale 1 minuto da circunferência terrestre. Assim tem-se que uma milha equivale a $1852$ metros ou $1,852Km$.

Pelas propriedades de um triangulo retângulo, sabe-se que a tangente é a divisão do cateto oposto pelo adjacente, seja $P_1$ a origem em um plano cartesiano, ou seja, o ponto $(0, 0)$, podemos deduzir que o ângulo $\alpha$ do seguimento $P_1$ e $P_2$ em um círculo trigonométrico, será:

$$\tan(\alpha) = \frac{DLA}{DLO} \tag{14}$$

Sabendo-se o valor da tangente, o valor de $\alpha$ será dado por:

$$\arctan|(\alpha)| = \tan(\alpha^{-1}) \tag{15}$$

O valor de $\alpha$ é dado em modulo pois se faz necessário encontrar o quadrante ao qual $\alpha$ pertence, e o quadrante será dado de acordo com os sinais do DLA e DLO.

A cidade de São João del Rei – MG foi considerada como $0^\circ$ (zero grau). Para realização das analises utilizou-se o pacote circular e CircStats do software R [8].

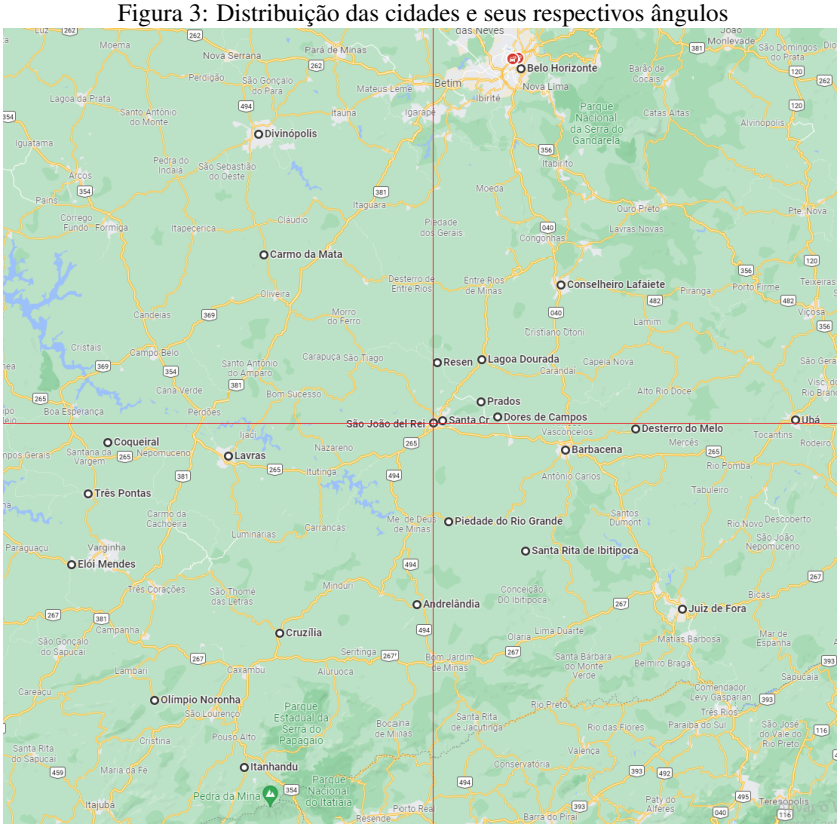
## 4 Resultados e Discussões

A Tabela 1 mostra os ângulos encontrados e a distância, em relação à cidade de referência, para cada uma das 26 cidades de origem dos alunos matriculados no PROFMAT.

A distribuição dos alunos, nas diversas cidades, está representada nas Figuras 3 e 4, e percebe-se que alunos de várias regiões do entorno de São João del Rei frequentam este programa de pós graduação.

Considerando que os dados estão distribuídos numa circunferência de raio 1, a

Figura 3: Distribuição das cidades e seus respectivos ângulos




**Source: Os autores**

média circular obtida foi $\bar{\theta} = 0,398$ rad ou $\bar{\theta} = 22^{\circ}$, o comprimento médio resultante foi de $\bar{R} = 0,29$ e a variância circular $V = 0,71$. Note que os alunos estão distribuídos em toda a região do entorno de São João del Rei. Além disso, os alunos estão mais concentrados nas cidade situadas ao norte da cidade sede do curso.

Em seguida foi encontrada a média circular e a variâncias dos dados

## 5 Conclusão

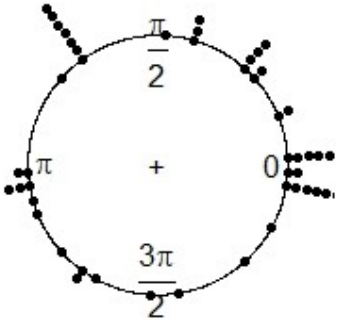
Os dados dos alunos do programa de pós graduação PROFMAT indicam que a procura de alunos pelo curso é proveniente das diversas regiões nas quais a UFSJ está inserida. Estes dados podem ser um indicativo da necessidade de formação de professores em nível mestrado profissional em Matemática nestas regiões, e ainda sugerir a necessidade de adequações na universidade e na cidade de São João de Rei, no sentido da melhoria na infraestrutura para a recepção destes alunos.

Não há uma concentração dos alunos matriculados no curso em apenas uma região

| Cidade | Núm. de alunos | Lat. | Longit. | Direção ($^\circ$) | Distância(Km) |
|---|---|---|---|---|---|
| São João del Rei | 7 | $21^\circ 8' 9"$ | $44^\circ 15' 43"$ | 0 | 0,00 |
| Santa Cruz de Minas | 1 | $21^\circ 7' 12"$ | $44^\circ 13' 22"$ | 22 | 4,69 |
| Ubá | 2 | $21^\circ 7' 12"$ | $42^\circ 56' 34"$ | 1 | 146,60 |
| Prados | 1 | $21^\circ 3' 28"$ | $44^\circ 4' 48"$ | 23 | 22,00 |
| Resende Costa | 1 | $20^\circ 55' 19"$ | $44^\circ 14' 16"$ | 84 | 23,54 |
| Lagoa Dourada | 3 | $20^\circ 54' 50"$ | $44^\circ 4' 48"$ | 50 | 32,05 |
| Dores de Campos | 1 | $21^\circ 6' 32"$ | $44^\circ 1' 22"$ | 6 | 26,70 |
| Barbacena | 9 | $21^\circ 13' 33"$ | $43^\circ 46' 26"$ | 350 | 55,14 |
| Desterro de Melo | 2 | $21^\circ 8' 49"$ | $43^\circ 31' 4"$ | 359 | 82,7 |
| Ouro Branco | 1 | $20^\circ 31' 15"$ | $43^\circ 41' 31"$ | 47 | 93,17 |
| Divinópolis | 7 | $20^\circ 8' 33, 6"$ | $44^\circ 53' 25, 2"$ | 127 | 130,59 |
| Conselheiro Lafaiete | 2 | $20^\circ 39' 36"$ | $43^\circ 47' 9"$ | 45 | 74,80 |
| Belo Horizonte | 3 | $19^\circ 55' 0"$ | $43^\circ 56' 0"$ | 75 | 140,30 |
| Carmo da Mata | 1 | $20^\circ 33' 28"$ | $44^\circ 52' 15"$ | 44 | 93,30 |
| Coqueiral | 2 | $21^\circ 11' 20"$ | $45^\circ 26' 27"$ | 183 | 131,13 |
| Lavras | 2 | $21^\circ 14' 42"$ | $45^\circ 0' 0"$ | 188 | 82,90 |
| Três Pontas | 1 | $21^\circ 22" 13"$ | $45^\circ 30' 44"$ | 191 | 141,30 |
| Varginha | 1 | $21^\circ 233' 3"$ | $45^\circ 25' 48"$ | 199 | 137,70 |
| Elói Mendes | 1 | $21^\circ 36' 36"$ | $45^\circ 33' 54"$ | 200 | 154,00 |
| Olímpio Noronha | 1 | $22^\circ 4' 4"$ | $45^\circ 15' 50"$ | 223 | 152,00 |
| Cruzília | 2 | $21^\circ 50' 20"$ | $44^\circ 48' 28"$ | 232 | 99,00 |
| Itanhandu | 1 | $22^\circ 17' 45"$ | $45^\circ 44' 56"$ | 240 | 149,00 |
| Andrelândia | 1 | $21^\circ 44' 24"$ | $44^\circ 18' 32"$ | 266 | 67,30 |
| Piedade do Rio Grande | 1 | $21^\circ 28' 8"$ | $44^\circ 11' 45"$ | 281 | 37,7, |
| Santa Rita de Ibitipoca | 1 | $21^\circ 33' 46"$ | $43^\circ 54' 54"$ | 309 | 61,00 |
| Juiz de Fora | 1 | $21^\circ 41' 20"$ | $43^\circ 20' 40"$ | 329 | 119,00 |

Tabela 1: Cidades de origem dos alunos do PROFMAT entre 2017 e 2021 suas respectivas coordenadas geográficas com distâncias em relação à São João del Rei - MG

Figura 4: Gráfico circular com a distribuição dos alunos




**Fonte: Os autores**

específica. Cerca de $30\%$ dos alunos são oriundos da região ao norte de São João del Rei. Com isso, há a indicativo de uma necessidade maior divulgação do curso nas cidades do entorno, o que poderia resultar numa demanda maior de professores da educação básica em se matricular no programa. Esta pesquisa pode ser estendida para outros cursos da instituição, tanto de graduação como de pós-graduação.

## 6 Agradecimentos

Os autores agradecem o apoio dado pela Universidade Federal dos Vales do Jequitinhonha e Mucuri, Universidade Federal de São João del Rei, e ao Departamento de Estatística da Universidade Federal de Lavras.

## Referências

[1] Falconi, Carlos Eduardo. Calculando distâncias e direções utilizando coordenadas geográficas. < *https://www.pilotopolicial.com.br/calculando-distancias-e-direcoes-utilizando-coordenadas-geograficas/*>, 2009.

[2] Fisher, Nicholas. *Statistical anlysis of circular data*. Cambridge University Press, New York, 1993.

[3] Jammalamadaka, Sreenivasa Rao and Lund, Ulrich J. The effect of wind direction on ozone levels: a case study. *Enviromment ecology statistics*, (13):287–298, 2006.

[4] Jammalamadaka, Sreenivasa Rao and SenGupta, A. *Topics in circular statistics*. World Scientific, New Jersey, 2001.

[5] Kantil, Mardia and Jupp, Peter. *Directional statistics*. Jonh Wiley & Sons, London, 2000.

[6] Mardia, Kantil. *Statistics of directional data*. Academic Press, London, 1972.

[7] Menezes, Eliane Barbosa. Um estudo direcional dos estudantes que chegam à UFRB: Uma aplicação da estatística circular. [trabalho de conclusão de curso -TCC]. *UFRB - Universidade Federal do Recôncavo da Bahia*, 2018.

[8] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. ISBN 3-900051-07-0.

## ARTIGO 2

**Regressão linear-circular para modelagem de dados meteorológicos na cidade de São João del Rei**

# Regressão linear-circular para modelagem de dados meteorológicos na cidade de São João del Rei


## Resumo

Modelos de regressão são utilizados para analisar o comportamento de uma variável em relação às outras. Quando os dados são provenientes da reta real podem ser usados, por exemplo, modelos de regressão linear. No entanto, quando os dados são resultantes de medidas angulares, que podem surgir a partir de observações relacionadas à direção de voos de pássaros, dados meteorológicos, meses do ano, assim, a melhor alternativa é utilizar os modelos de regressão para dados circulares. Neste trabalho, os dados são provenientes do INMET e referentes à estação meteorológica situada na cidade de São João del Rei-MG, no período de 01/01/2020 a 31/12/2021. O modelo linear-circular foi selecionado, sendo a variável resposta a radiação solar, a variável explicativa, circular, os meses do ano e a covariável umidade relativa do ar foi inserida. Para o ajuste foi utilizado o software R. Os parâmetros foram significativos ao nível de significância de 1%. Para validar o modelo foi utilizada a análise dos resíduos, e indicou ajuste satisfatório. Através da estatística circular foi possível verificar os períodos de picos de radiação, e em que meses do ano ocorreram altas radiações.



**Palavras chave**: Ventos, Radiação, Correlação, Direção.


## Introdução

Uma das maneiras de se analisar o comportamento de uma variável em relação a outra é utilizando a regressão linear. Com isso tem-se a variável resposta ($Y_i$) e a variável preditora ($x_i$), podendo ocorrer a inclusão de covariáveis. Uma suposição da regressão linear simples ou múltipla é que a variável resposta tem distribuição normal (CASELLA; BERGER, 2018; UYANIK; GÜLER, 2007). Ocorrem situações em que as variáveis seguem outras distribuições, como binomial, Poisson e Weibull. Nesse caso, podem ser utilizados os modelos lineares generalizados. Assim, a variável resposta $Y_i$ é considerada como pertencente a uma família de distribuições chamada de família exponencial. A variável $x_i$ está relacionada com a média de $Y_i$ por meio de uma função de ligação (LOBATO; VEIGA, 2020; MARQUES; VILLELA, 2020). Tanto na regressão linear como nos modelos lineares generalizados os dados são considerados como provenientes da reta real. No entanto, ocorrem situações em que os dados são resultantes de medidas circulares ou angulares, estes podem ser vistos como localizados num círculo de raio unitário. Eles podem surgir a partir de observações relacionadas à direção de voos de pássaros, dados meteorológicos, ritmos circadianos, horários, dias ou meses do ano (THORUP; RABOL; ERNI, 2007; JAMMALAMADAKA; LUND, 2006; BRIGHENTI et al, 2014). Quando se trata de medidas circulares, a melhor forma de analisá-los é utilizar a estatística para dados circulares (FISHER, 1993), pois o uso da estatística usual para dados lineares pode levar a erros de interpretação (MARDIA, 1972). Jammalamadaka (2001) e Mardia & Jupp (2000) citam modelos de regressão específicos para dados circulares. Estes modelos podem ser denominados linear-circular ou circular – linear, quando a variável resposta é uma medida linear (situada na reta real) e a variável explicativa é circular ou vice-versa, ou circular-circular, quando as variáveis explicativas e resposta são circulares.

Com relação à distribuição para a variável aleatória circular, existem diversos modelos de probabilidade, como uniforme, cardióide, wrapped e von Mises. Neste artigo, os

dados circulares serão considerados como provenientes da distribuição von Mises. Vale mencionar que Fisher (1993) e Mardia & Jupp (2000) consideram que essa distribuição, em muitos aspectos, tem propriedades semelhantes à distribuição normal. Autores como, Jammalamadaka & Lund (2006) e SenGupta & Ugwowo (2006), utilizaram modelos de regressão para dados circulares para analisarem fenômenos meteorológicos, nos quais as variáveis circulares foram as direções dos ventos, dias e meses do ano.

Este artigo está organizado em seis seções a segunda e terceira fornecem detalhes sobre a estatística para dados circulares, os modelos de distribuição de probabilidade bem como os modelos de regressão próprios desta área da estatística. As duas seções seguintes contêm uma aplicação a dados reais, de uma estação meteorológica situada na cidade de São João del Rei - MG, uma análise dos resultados obtidos e a escolha de um modelo de regressão para ajuste dos dados. E, por fim, na última parte é feita a conclusão da pesquisa realizada.

Foram analisados dados angulares como direção dos ventos, dias e meses dos anos e outros dados meteorológicos, avaliando a correlação entre eles. Neste contexto o objetivo principal deste trabalho foi ajustar modelos de regressão envolvendo variáveis circulares. Deve-se destacar a relação existente entre a radiação e outras variáveis meteorológicas como umidade relativa do ar e temperatura (OLIVEIRA; CAVAZZANA; SOUZA, 2019; DALMAGO el al., 2006). Assim, no contexto da estatística circular pretende-se ajustar a variável resposta, a radiação, em função das covariáveis: temperatura, velocidade dos ventos, umidade relativa do ar, pressão e meses do ano.

## Estatística para dados circulares

Um ponto $P$ do plano pode ser representado por $(x, y)$ no caso de coordenadas retangulares ou pode ser representado por $(r, \theta)$ em coordenadas polares. Pode ser estabelecida uma relação entre as coordenadas retangulares e polares da seguinte maneira:

$x = r\ cos\theta$ e $y = r \sin\theta$

Considerando uma circunferência de raio unitário, isto é, $r = 1$, $x = \cos\theta$ e $y = \sin\theta$

Portanto, um ponto $P$ no círculo unitário pode ser representado por $P = (\cos\theta, \sin\theta)$.

Considere agora uma amostra aleatória dada pelos vetores unitários $\vec{v_1}, ..., \vec{v_n}$ e seus respectivos ângulos $\theta_1, ...\theta_n$. A média direcional $\bar{\theta}$ de $\theta_1, ..., \theta_n$ é a direção resultante de $\vec{v_1} + ... + \vec{v_n}$ dos vetores $\vec{v_1}, ..., \vec{v_n}$. Ela tem a mesma direção do centro de massa $\vec{v}$ de $\vec{v_1}, ..., \vec{v_n}$ (MARDIA; JUPP, 2000).

Como o vetor $\vec{v}$ é unitário, então $\vec{v} = (C, S) = (\cos\theta, \sin\theta)$, daí o centro de massa será representado por:

$$(\bar{C}, \bar{S}) = (\frac{1}{n}\sum_{i=1}^{n}\cos\theta_i, \frac{1}{n}\sum_{i=1}^{n}\sin\theta_i) \tag{1}$$

Portanto $\bar{\theta}$, a média circular, é a solução das equações:

$$\begin{cases} \bar{C} = \bar{R}\cos\bar{\theta} \\ \bar{S} = \bar{R}\sin\bar{\theta} \end{cases} \tag{2}$$

onde

$$\bar{R}^2 = \bar{C}^2 + \bar{S}^2$$

$\bar{R} = \dfrac{R}{n}$, e esta medida é chamada de comprimento médio resultante e está associada à concentração dos dados. Note que $0 \leq \bar{R} < 1$. Para valores de $\bar{R}$ próximos 1(um) tem-se

que a concentração dos dados e valores de $\bar{R}$ próximos de 0 (zero) há uma forte dispersão dos dados (MARDIA, 1972).

A variância circular $V$ é dada por $V = 1 - \bar{R}$, e o desvio padrão $\nu$ é dado por

$$\nu = \sqrt{-2\log(1-V))} \tag{3}$$

## Modelo de Probabilidade

Para a análise de dados provenientes de medidas angulares ou direcionais são utilizados modelos de probabilidade próprios. Se $f(\theta)$ é uma função de densidade de probabilidade de uma variável aleatória circular $\Theta$, então esta função satisfaz as seguintes propriedades:

i) $f(\theta) \geq 0$

ii) $f(\theta + 2\pi) = f(\theta)$

iii) $\int_0^{2\pi} f(\theta)d\theta = 1$

Um dos modelos de probabilidade muito utilizado para dados circulares é o modelo von Mises, o qual será denotado por $vM(\mu, \kappa)$, o parâmetro $\mu$ é a média direcional e $\kappa$ é o parâmetro de concentração (JAMMALAMADAKA; SENGUPTA, 2001).

O modelo é dado por:

$$g(\theta, \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa\cos(\theta-\mu)} \tag{4}$$

em que $I_0$ é a função de Bessel modificada, de primeiro tipo e ordem 0, que pode ser definida por

$$I_0(\kappa) = \frac{1}{2\pi}\int_0^{2\pi} e^{\kappa\cos\theta}\, d\theta \tag{5}$$

## Correlação e regressão para dados circulares

Em muitas situações práticas, quando se deseja verificar se há alguma relação ou dependência entre duas variáveis aleatórias, usa-se uma medida para verificar se esta dependência ocorre, a correlação, e ela pode ser:

1. Linear- linear quando as variáveis são lineares, ou são provenientes de dados localizados na reta real;

2. Linear- circular, quando uma variável é linear e a outra é circular;

3. Circular- circular no caso em que as duas variáveis são circulares.

Para a correlação linear-linear pode ser utilizado o coeficiente de correlação (CASELLA; BERGER, 2018) dado por:

$$\rho_p = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \tag{6}$$

Uma medida de correlação entre uma variável linear $x$ e uma variável angular $\alpha$, correlação linear-circular (MARDIA; SUTTON, 1978), é dada por:

$$R^2_{x,\alpha} = \frac{r^2_{xc} + r^2_{xs} - 2r_{xc}r_{xs}r_{cs}}{1 - r^2_{cs}} \tag{7}$$

onde

$$r_{xc} = corr(x, \cos\alpha)$$
$$r_{xs} = corr(x, \sin\alpha)$$
$$r_{cs} = corr(\cos\alpha, \sin\alpha)$$

Quando o valor de $R^2_{x,\alpha}$ está próximo de 0 não há relação entre as variáveis e quando o valor está próximo de 1 há evidência de uma forte relação entre as variáveis.

Considere agora uma amostra dada por $(\theta_1, \phi_1), ..., (\theta_n, \phi_n)$ variáveis circulares e suas respectivas médias direcionais, $\bar{\theta}$ e $\bar{\phi}$ Jammalamadaka & Senguta (2001) sugerem a seguinte medida:

$$r_{c,n} = \frac{\sum_{i=1}^{n} \sin(\theta_i - \bar{\theta}) \sin(\phi_i - \bar{\phi})}{\sqrt{\sum_{i=1}^{n} \sin^2(\theta_i - \bar{\theta}) \sin^2(\phi_i - \bar{\phi})}} \tag{8}$$

Se as variáveis $\theta$ e $\phi$ são independentes o valor de $r_{c,n}$ será bem próximo de zero, quanto mais próximo o valor de $r$ for de 1 maior será a associação entre as variáveis.

Quando as variáveis têm natureza linear pode ser verificada a forma da associação entre elas considerando vários tipos de ajustes, tais como a regressão linear propriamente dita ou mesmo a regressão não-linear, pois estes casos envolvem a definição do tipo de regressão considerando os parâmetros. No entanto, nestes casos as estatísticas são provenientes de dados trabalhados na reta real, ou dados que podem ser plotados no $\mathbb{R}^n$. Já a regressão que envolve variáveis circulares, devido à natureza periódica dos dados, há modelos próprios os quais podem ser circular-linear, linear-circular e circular-circular.

Neste artigo foi considerado o modelo de regressão linear-circular, quando a variável resposta é uma medida linear, quando a variável explicativa é circular e a variável resposta é linear. Com isso, considerando a variável resposta $Y$ e a variável explicativa circular $\theta$, o modelo de regressão sugerido por Mardia & Jupp (2000) é dado por:

$$Y_i = \beta_0 + \beta_1 \cos\theta_i + \beta_2 \sin\theta_i + \epsilon_i \tag{9}$$

com $i = 1, ..., n$ e $\beta_0, \beta_1, \beta_2$ os parâmetros a serem estimados.

## Metodologia

Os dados analisados nesta pesquisa foram obtidos do INMET e refere-se à cidade de São João del Rei no estado de Minas Gerais. Nesta cidade há uma estação meteorológica automatizada localizada no Campus Tancredo Neves - CTAN, da Universidade Federal de São João del Rei -UFSJ. Os dados são referentes aos anos de 2020 e 2021, para que fosse possível verificar o comportamento periódico, ambos no período de 1 de janeiro a 31 de dezembro. E incluem variáveis como temperatura máxima, umidade relativa do ar, velocidade dos ventos, radiação e direção dos ventos. Os dados são diários e relativos aos horários das 0h as 23h. As unidades de medidas para cada variável estão listadas na Tabela ( 1).

Inicialmente foram obtidas as médias mensais para cada variável, salientando que para os dados circulares é calculada a média circular.

Foram encontradas as correlações lineares, linear- circular e circular-circular e em seguida os dados foram ajustados ao modelo linear- circular para as médias mensais e meses do ano, vale destacar que as medidas para meses do ano foram transformadas em circulares.

Os dados foram ajustados para o modelo de regressão linear-circular, onde as variáveis circulares foram a direção dos ventos e meses do ano. A análise inicial se deu com o modelo apenas com a variável circular, e depois covariáveis foram incluídas.

Para validar a escolha do modelo foi realizada a análise dos resíduos, o gráfico com os resíduos e QQ plot. Para análise e ajuste dados foram utilizados os pacotes circular e CircStats do software R (R Development Core Team, 2021).

## Resultados e discussões

As médias para a pressão, radiação, temperatura máxima diária, umidade relativa do ar, velocidade e direção dos ventos para os anos de 2020 e 2021 estão mostradas na figura ( 1).

Com base nos gráficos apresentadas na figura ( 1), observa-se que para os anos de 2020 e 2021 eles apresentam comportamento bem semelhantes, exceto nos gráficos da pressão atmosférica e direção dos ventos. Com relação à pressão média mensal, vê-se que no ano de 2020 houve uma queda da pressão atmosférica entre os meses de abril e setembro e no gráfico da direção dos ventos há diferença considerável desta variável circular entre os meses de abril e junho no ano de 2021.

Para obtenção da correlação entre as diversas variáveis da planilha de dados meteorológicos verificou-se a ausência de registros no período das 23h às 8h para a variável pressão, e para outras variáveis faltavam os dados em alguns dias nos horários das 9h, 10h e 23h dessa maneira considerou-se apenas os dados do período das 11h às 21h.

Para a análise de correlação e regressão a variável direção dos ventos foi tratada como circular e as demais lineares, desta forma obteve-se valores para os coeficientes de correlação de correlação linear - linear e linear- circular para os anos de 2020 e 2021, conforme mostrado nas tabelas ( 2) e ( 3).

Na correlação linear-linear o maior valor obtido foi entre a temperatura e a radiação, tanto para 2020 como 2021, seguida dos valores de temperatura e pressão, no entanto houve uma correlação em 2020 e negativa em 2021, indicando realmente uma alteração devido a pressão conforme observado na figura ( 1), fato semelhante aconteceu com a pressão e radiação.

O ajuste do modelo de regressão foi realizado para os casos linear-circular e circular-circular, sendo considerado neste caso os meses do ano transformado em medidas angulares.

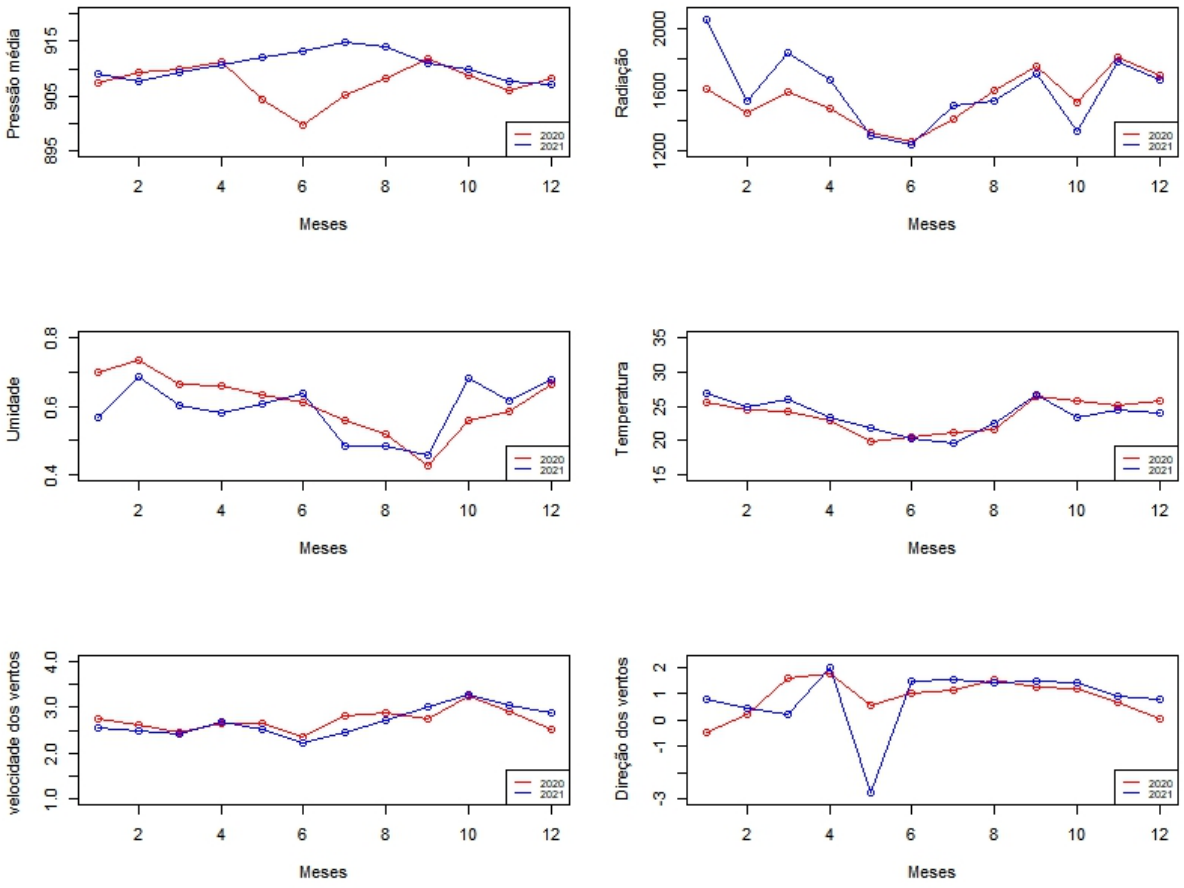
Inicialmente foi estudado o modelo 1, que contém apenas a covariável circular e em seguida foram inseridas as outras covariáveis lineares, considerando a variável resposta radiação. Para o modelo circular-circular, considerando a direção dos ventos e os meses do ano, como variáveis circulares, não foi obtida significância. No caso dos modelos de regressão linear-circular, foi obtida significância com a variável circular meses do ano, sendo a variável direção dos ventos desconsiderada. A tabela ( 4) mostra as estimativas dos parâmetros dos modelos analisados. O Modelo 1 é referente ao modelo considerando as covariáveis meses dos anos e os Modelos 2 até 11 considera além da covariável circular as demais covariáveis lineares, sendo os 4 primeiros a inclusão de apenas uma delas e os demais modelos duas.

Tabela 1: Unidades de medidas

| **Variável** | **Unidade de medida** |
|---|---|
| Pressão | mB |
| Radiação | $KJ/m^2$ |
| Direção dos ventos | Radianos |
| Velocidade dos ventos | m/s |
| Umidade relativa do ar | % |
| Temperatura máxima | ºC |

Fonte: Os autores (2022)

Figura 1: Média das variáveis: umidade, temperatura, pressão, radiação, velocidade e direção dos ventos em função dos meses do ano de 2020 e 2021




Fonte: Os autores (2022)

Tabela 2: Correlação linear-linear e linear-circular relativa ao ano de 2020

| | Radiação | Pressão | Temperatura | Umidade | Velocidade dos Ventos |
|---|---|---|---|---|---|
| Radiação | | 0,559 | 0,767 | -0,294 | 0,312 |
| Pressão | | | 0,630 | -0,105 | 0,264 |
| Temperatura | | | | -0,025 | 0,304 |
| Umidade | | | | | -0,434 |
| Direção dos ventos | 0,038 | 0,267 | 0,126 | 0,391 | 0,029 |

Fonte: Os autores (2022)

Tabela 3: Correlação linear-linear e linear-circular relativa ao ano de 2021

| | Radiação | Pressão | Temperatura | Umidade | Velocidade dos Ventos |
|---|---|---|---|---|---|
| Radiação | | -0,498 | 0,783 | -0,228 | 0,087 |
| Pressão | | | -0,694 | -0,643 | -0,366 |
| Temperatura | | | | 0,023 | 0,310 |
| Umidade | | | | | 0,059 |
| Direção dos ventos | 0,330 | 0,397 | 0,347 | 0,191 | 0,104 |

Fonte: Os autores (2022)

Tabela 4: Estimativas dos parâmetros dos modelos

| *Modelos* | Intercepto | Cos $\theta$ | Sen$\theta$ | Umid. | Temp | Velocidade | Pressão |
|---|---|---|---|---|---|---|---|
| 1 | 1568,78* | 178,33* | -0,81 | | | | |
| 2 | 3130,87* | 278,46* | 187,70* | -2603,42* | | | |
| 3 | 64,22 | 20,68 | -7,26 | | 63,57* | | |
| 4 | 2058,19* | 206,86* | -41,62 | | | -180,95 | |
| 5 | -7724,66 | 184,89* | 4,38 | | | | 10,22 |
| 6 | 3760,83* | 336,48* | 213,25* | -2930,69* | -18,32 | | |
| 7 | 7909,80 | 279,01* | 192,37* | -2704,36* | | | -5,19 |
| 8 | 3186,78* | 281,66* | 181,04* | -2587,14* | | -24,29 | |
| 9 | -2517,64 | 27,08 | -5,6 | | 61,74* | | 2,89 |
| 10 | 537,22 | 48,79 | -45,69 | | 63,09* | -170,67 | |
| 11 | -6305,43 | 209,78* | -32,75 | | | 9,15 | -162,22 |
| 9,15 | | | | | | | |

Obs.: Valores seguidos de asteriscos correspondem ao valor $p < 0,05$

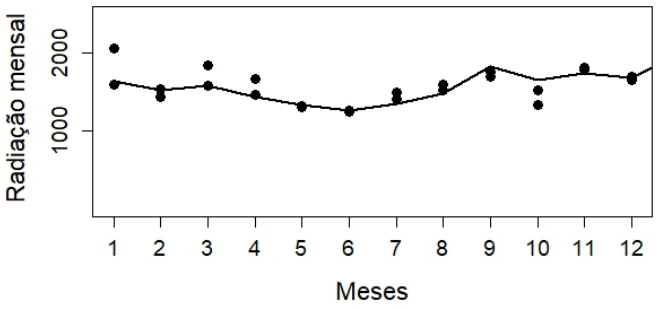
Fonte : Os autores (2022)

O modelo escolhido foi o 2, já que todos os parâmetros são significativos, e o $R^2$ ajustado apresentou o valor 0,864, isto significa que 86,4% dos dados são explicados pelo modelo. Dessa forma obtém-se a seguinte equação de ajuste do modelo linear-circular:

$$\widehat{\text{Radiação}} = 3130,87 + 278,46\cos(\theta) + 187,7\sin(\theta) - 2603,42\text{Umidade} \qquad (10)$$

sendo que $\theta$ corresponde aos meses do ano em radianos. Os valores preditos pelo modelo estão apresentados na figura ( 2).

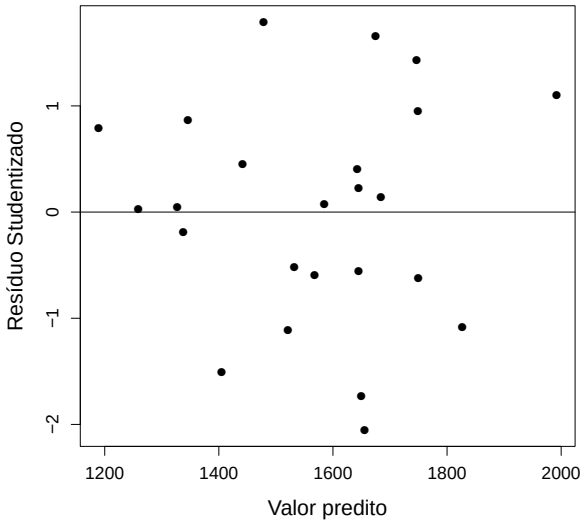
Figura 2: Gráfico dos valores preditos do modelo




Fonte: Os autores (2022)

Para validar a escolha do modelo 2 será realizada a análise dos resíduos. Uma maneira pode ser realizada utilizando os valores ajustados e resíduos Studentizado (PEWSEY; NEUHÄUSER; RUXTON, 2013).

Figura 3: Gráfico dos resíduos

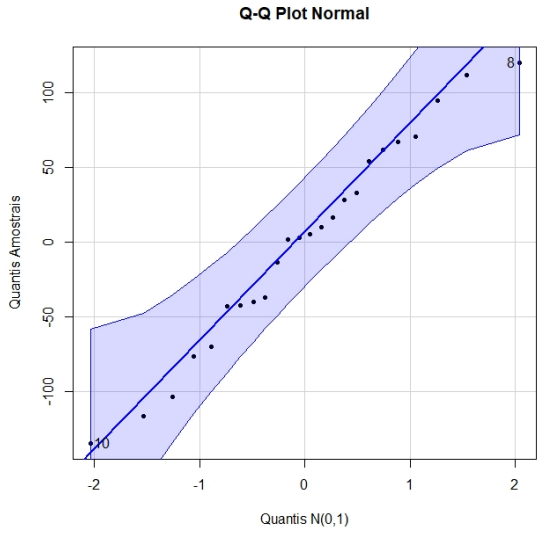



Fonte: Os autores (2022)

Os erros são considerados independentes logo, eles não devem estar correlacionados, eles também são considerados identicamente distribuídos e com variância constante e média zero. Na figura ( 3), observa-se que os erros estão distribuídos aleatoriamente em torno do zero, indicando que não há nenhum padrão, isso evidencia que a variância é constante.

O gráfico do QQ plot, mostrado na figura ( 4), mostra também a qualidade do ajuste, pois nota-se que os erros estão normalmente distribuídos.

Figura 4: Gráfico QQ plot




Fonte: Os autores (2022)

A normalidade é comprovada pelo teste de Shapiro - Wilk, o valor da estatística do teste foi $W = 0,97905$ e o valor p $= 0,8777$, então a hipótese nula, de que os erros estão normalmente distribuídos não é rejeitada.

Com relação a homocedasticidade, foi comprovada pelo teste de Barlett, o valor p obtido para a estatística do teste igual a $0,4534$ assim, a hipótese nula não é rejeitada, e a variância do erro é considerada constante.

## Conclusões

Para conjuntos de dados que envolvem medidas angulares ou circulares a estatística circular é uma excelente ferramenta. Foi possível ajustar os dados para o modelo linear - circular, devido à relação existente entre meses do ano e radiação, foi incluída como covariável a umidade relativa do ar. A inclusão desta covariável melhorou o ajuste. Nesse caso, verificou-se que no período estudado, os anos de 2020 e 2021, há um pico de radiação no mês de setembro e valores de altas radiações ocorreram a partir do mês de agosto e permaneceram alta até o mês de janeiro. Tais informações podem ser úteis para analisar se numa região como essa, pode ser viável investigar a utilização de fontes de energia alternativas que dependem da incidência da energia solar.

# Referências


BRIGHENTI, C. R. G. et al. Distribuição von Mises na avaliação de dados apícolas, **Archivos de zootecnia**, v.63, n.243, p.461-471, 2014.

BURIOL, G. A. *et al.* Estimativa da radiação solar global a partir dos dados de insolação para Santa Maria – RS, **Ciência Rural**, v. 42, n.9, p. 1563-1567, 2012.

CAMELO, H. N. *et al.*, Predição de velocidade do vento em municípios do Nordeste brasileiro através de regressão linear e não linear para fins de geração eólica, **Revista Brasileira de Geografia Física**, v. 9, n. 3, p. 927-939, 2016.

CASELLA, G.; BERGER, R. **Inferência estatística**, Cengage Learning, São Paulo, 2018.

DALMAGO, G. A. *et al.* Evapotranspiração máxima da cultura de pimentão em estufa plástica em função da radiação solar, da temperatura, da umidade relativa e do déficit de saturação do ar. **Ciência Rural**, v. 36, p. 785–792, 2006.

FISHER, N.I. **Statistical analysis of circular data**, Cambridge University Press, New York, 1993.

FISHER, N.I.; LEE, A. J., Regression models for an angular response, **Biometrics**, JSTOR, 665-677, 1992.

FREITAS, J.R. *et al.*, Análise em séries temporais da radiação solar na Cidade do Recife/PE, **Research, Society and Development**, v. 9, n. 9, 2020.

HOWELL, J. R.; MENGUC, M. P.; SIEGEL, R. Thermal Radiation Heat Transfer. **CRC Press**, New York, 2016.

JAMMALAMADAKA, S. R.; SENGUPTA, A. **Topics in circular statistics**, World Scientific, New Jersey, 2001.

JAMMALAMADAKA, S. R.; LUND, U. J., The effect of wind direction on ozone levels: a case study, **Enviromment Ecology Statistics**, . 13, p. 287-298, 2006.

LIRA, M.A.T. ; DA SILVA, E.M. ; ALVES, J.M.B., Estimativa dos recursos eólicos no litoral cearense usando a teoria da regressão linear,**Revista Brasileira de Meteorologia**, v.26, n.3, p. 349 -366, 2011.

LOBATO JUNIOR, D.; VEIGA, R. D. Análise de Diagnóstico em modelos de regressão normal e logístico, **Revista brasileira de biometria**, v. 38, n.4, p. 449- 482, 2020.

MARDIA, K. V. **Statistics of directional data**. London: Academic Press, 1972.

MARDIA, K. V.; JUPP, P. E, **Directional statistics**, Jonh Wiley & Sons, London, 2000.

MARDIA, K.; SUTTON, T. W. A model for cylindrical variables with applications. **Royal Statistical Society**, v. 40, n. 2, p. 229–233, 1978.

MARQUES, C.; VILLELA, R. A desigualdade de renda e a pandemia de covid-19 nas regiões metropolitanas do Brasil, **Revista Brasileira de Estatística**, v.78, n. 244, p. 32-53.

MATTIUZZI, H. V.; MARCHIORO, E. O Comportamento dos ventos em Vitória(ES): A gestão e interpretação dos dados climatológicos. **Revista Geonorte**, v. 2, n. 4, p. 983-993, 2012.

OLIVEIRA, S. S.; CAVAZZANA, G. H.; SOUZA, A. Estimativa da radiação solar global em função da temperatura do ar e isolinhas para o estado de Mato Grosso do Sul, Brasil. **Revista Brasileira de Gestão Ambiental e Sustentabilidade**, v. 6, n. 12, p. 93–108, 2019.

PEWSEY, A.; NEUHÄUSER, M.; RUXTON, G. D. **Circular Statistics in R**, Oxford University Press, New York, 2013.

R Development Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2021. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org>.

SENGUPTA, A; UGWUOWO, F. I. Asymmetric circular-linear multivariate regression models with applications to environmental data, **Enviroment Statistical ecology**, v.13, p.299-309 2006.

SILVA, R. *et al.*, Estudo da variabilidade da radiação solar no Nordeste do Brasil, **Revista Brasileira de Engenharia Agrícola e Ambiental**, v. 14, n. 5, p. 501- 509, 2010.

THORUP, K.; RABOL, J; ERNI, B. Estimating variation among individuals in migration Direction, **Journal of Avian Biology**, v.38, n.2, p.182-189, 2007.

UYANIK, G. K.; GÜLER, N. A study on multiple linear regression analysis. **Procedia – social and behavioral sciences**, v.106, p. 234-240, 2007.

# ARTIGO 3

## Modelos com eventos múltiplos para avaliação dos picos de radiação solar na cidade de São João del Rei- MG

*Disciplinarum Scientia*. Série: Naturais e Tecnológicas, Santa Maria, v. 23, n. 1, p. 1-11, 2022.
Recebido em: 05.11.2021. Aprovado em: 10.03.2022.

ISSN 2176-462X DOI: 10.37779/nt.v23i1.4093

# MODELOS COM EVENTOS MÚLTIPLOS PARA AVALIAÇÃO DOS PICOS DE RADIAÇÃO SOLAR NA CIDADE DE SÃO JOÃO DEL REI - MG[1]

## *MODELS WITH MULTIPLE EVENTS TO EVALUATE THE PEAKS OF SOLAR RADIATION IN THE CITY OF SÃO JOÃO DEL REI - MG*

**Clodoaldo Teodosio Santana da Silva[2] e Carla Regina Guimarães Brighenti[3]**

**RESUMO**

Devido à necessidade de utilização de fontes renováveis de energia, visando não apenas a utilização de usinas hidrelétricas como matriz energética, fontes alternativas têm sido discutidas. Uma das alternativas é o uso de energia solar. O estado de Minas Gerais tem um potencial bastante alto para fazer uso da energia solar, pois a incidência da radiação solar no estado é, em média, igual a 2314 KJ /m$^2$ por hora. Ao longo do dia, a radiação solar pode exceder valores absolutos que permitem maior geração de energia e que podem compensar a baixa produção em outros horários. O objetivo do trabalho foi analisar o horário de ocorrência de picos de radiação utilizando o modelo de regressão não paramétrico de Cox. Os dados analisados são referentes ao período de 01/01/2020 a 31/12/2020, da cidade de São João del Rei no estado de Minas Gerais. A cidade foi escolhida para este estudo pois há uma estação meteorológica automática localizada na Universidade Federal de São João del Rei - UFSJ. Foram inseridas as covariáveis umidade relativa do ar e temperatura instantânea e, devido a ocorrência de horários múltiplos com radiação acima do valor médio, em um mesmo dia, foi considerada uma extensão do modelo de Cox, chamada de modelo AG, utilizado para modelar eventos múltiplos. Inicialmente realizou-se uma análise descritiva dos dados, seguida da estimação dos parâmetros e análise de resíduos do modelo ajustado. Por meio do modelo obtido, foi possível verificar a probabilidade de ocorrência de picos de radiação ao longo do dia.

**Palavras-chave**: Modelo AG, Modelo de Cox, Energia solar.

***ABSTRACT***

*Due to the need to use renewable energy sources, aiming not only at the use of hydroelectric plants as an energy matrix, alternative sources have been discussed. One of the alternatives is the use of solar energy. The state of Minas Gerais has a very high potential to make use of solar energy, as the incidence of solar radiation in the state is, on average, equal to 2314 KJ/m$^2$ per hour. Throughout the day, solar radiation can exceed absolute values that allow greater energy generation and that can compensate for low production at other times. The objective of this work was to analyze the time of occurrence of radiation peaks using the Cox non-parametric regression model. The analyzed data refer to the period 01/01/2020 to 12/31/2020, in the city of São João del Rei in the state of Minas Gerais. The city was chosen for this study because there is an automatic meteorological station located at the Federal University of São João del Rei - UFSJ. The covariates relative humidity and instantaneous temperature were inserted and, due to the occurrence of multiple times with radiation above the mean value, in the same day, an extension of the Cox model, called the AG model, used to model events was considered. multiples. Initially, a descriptive analysis of the data was carried out,*

1 Parte da Tese de Doutorado do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária (PPGEE).

2 Doutorando - PPGEE/Universidade Federal de Lavras - UFLA e professor da Universidade Federal dos Vales de Jequitinhonha e Mucuri - UFVJM. E-mail: teoelania@gmail.com

3 Orientadora - PPGEE/Universidade Federal de Lavras - UFLA e professora da Universidade Federal de São João del Rei - UFSJ. E-mail: carlabrighenti@ufsj.edu.br

*followed by the estimation of the parameters and analysis of the residues of the adjusted model. Through the model obtained, it was possible to verify the probability of occurrence of radiation peaks throughout the day.*

***Keywords:*** *AG model, Cox's model, solar energy.*

## INTRODUÇÃO

A matriz energética do Brasil ainda está associada às hidroelétricas, porém alguns fatores, como períodos de estiagem, problemas ambientais, custo tem tornado necessário a busca por alternativas para a matriz energética. Uma fonte alternativa para a produção de energia elétrica para o país é o melhor aproveitamento do recurso solar (PEREIRA *et al*., 2017).

A localização geográfica do Brasil favorece o uso e aproveitamento da energia solar, e esta fonte de energia não gera prejuízos para o meio ambiente. A energia solar pode ser usada como fonte de energia térmica ou ser convertida em energia elétrica, com a utilização de células fotovoltaicas. Segundo Wanderley e Campos (2013) o uso da energia fotovoltaica é uma fonte de energia ilimitada, disponível em várias partes do mundo, os materiais podem ser reciclados e há ainda um benefício econômico, que é a geração de empregos.

Vale destacar que "as fontes alternativas solar e eólica são as de maior potencial para a utilização e geração distribuída que se caracteriza pelo uso de geradores próximo ao local de consumo, sem a necessidade de linhas de transmissão e conectadas diretamente às redes de distribuição de baixa tensão da concessionária de energia" (MILANO, 2018).

Sendo o Brasil um país com grande potencial para utilização de energia solar, há várias pesquisas sobre o assunto para avaliar o potencial nas diferentes regiões. Minas Gerais é um estado que tem aproveitado bastante essa fonte renovável de energia. Segundo Maiomi e Cardoso (2020) o estado tem grandes oportunidades para o uso de células fotovoltaicas ou outras tecnologias relacionadas com a energia solar. Isso é confirmado pela Associação Brasileira de Energia Solar (ABSOLAR, 2021). Dados desta associação mostram que Minas tem liderança na geração distribuída de energia solar, detendo 19,9% de todo o parque nacional, duas cidades mineiras estão entre as maiores geradoras de energia solar no Brasil, Uberlândia (2º) e Belo Horizonte (10º).

Para maximizar o aproveitamento da radiação solar, é necessário, além de ajustar a posição do painel solar de acordo com a latitude local, avaliar o período do ano em que se requer mais energia (GUIA TÉCNICO, 2014). O aproveitamento racional da energia solar no sentido de produzir instalações bem dimensionadas e economicamente viáveis só é possível a partir de informações solarimétricas consistentes da região em questão (PRADO *et al*., 2017).

Um dado indispensável para um projeto de sistema fotovoltaico são os índices de incidência de radiação solar na localidade onde o sistema será implantado (TIBA *et al*., 2000). De acordo com o levantamento desenvolvido pelo Atlas solarimétrico do Brasil os valores de incidência de

radiação para Minas Gerais revelam médias anuais em torno de 4,5 a 6,5 KWh / $m^2$ por dia para toda a extensão do estado.

Vários trabalhos destacam os maiores picos de radiação solar incidente nos diferentes horários ao longo do dia. As variações da radiação solar e da temperatura ambiente sobre um módulo fotovoltaico afetam diretamente a operação das células fotovoltaicas (MIDEKSA; KALLBEKKEN, 2010). O aumento da temperatura reduz a potência gerada, principalmente, por diminuir a tensão elétrica na célula mesmo que a corrente elétrica aumente de modo inexpressivo. Além disso, as elevadas temperaturas podem degradar significativamente as células fotovoltaicas e, portanto, reduzir a vida útil do módulo. Segundo Cantor (2017), existe uma relação direta entre aumento da temperatura dos painéis fotovoltaicos com a redução da corrente respectivamente. Consequentemente, quanto maior for o tempo de exposição solar, maior será a temperatura dos painéis e menor será sua eficiência de geração.

Além da temperatura, outros fatores de perdas são significativos para a análise do potencial solar brasileiro, como é o caso da umidade. Segundo (FRANCISCO, 2019), a umidade possui uma relação inversa com a energia gerada. Isso pode ser explicado pelo fato de que uma maior umidade do ar pode indicar formações de nuvens e precipitação, causando um bloqueio parcial do sol e, por consequência, uma redução na geração de energia.

A intensidade da radiação solar muda a cada instante em função da rotação da terra e sua translação ao redor do sol. Ao nível do mar, ao meio-dia com céu limpo (sem nenhuma nuvem) a intensidade da radiação solar atinge um valor próximo de 1000 W/$m^2$. Este valor está relacionado ao chamado Hora de Sol Pico (HSP) que é a insolação diária (ou mensal, ou anual, dependendo da medida de tempo utilizada) que é recebida por uma determinada superfície, levando-se em consideração aspectos, como a localização específica, ângulo de inclinação e orientação, levando-se em consideração aspectos, como a localização específica, ângulo de inclinação e orientação. A HSP é obtida dividindo-se a irradiação do local (kWh/$m^2$) por 1000 W/$m^2$ (ou 3600 KJ/$m^2$)..

Assim, o pico da hora solar indica a quantidade de horas em que uma irradiação solar padronizada de 1000 W/ $m^2$ é recebida naquele local e, se as demais condições padronizadas também fossem cumpridas exatamente, traria o número de horas por dia em que um painel fotovoltaico forneceria sua potência de pico. Os dados referentes à tabela HSP são obtidos nos chamados mapas solarimétrico e, no Brasil existem boas opções disponíveis para consulta.

Os horários de Sol pico podem alterar em função de diversos fatores tais como a temperatura e a umidade (WANDERLEY; CAMPOS, 2013). Desta forma, avaliar os horários do dia em que há maior incidência de radiação solar, principalmente associada a valores de temperatura e umidade se torna uma ferramenta importante para planejamento de instalação de painéis fotovoltaicos.

Como não só os valores máximos de radiação solar, mas os valores que excedem determinado padrão, como por exemplo, a média global da região, pode-se dizer que ocorrem eventos múltiplos de interesse. Além disso, há dias em que este valor padrão, considerado como evento de interesse, não é excedido

e tal fato deve ser também considerado no modelo a ser ajustado, o que caracteriza uma incompletude de um evento em determinados dias. Assim, para modelagem de horário de eventos múltiplos, associado ao estudo de covariáveis tais como a umidade e temperatura, pode-se utilizar uma extensão do modelo semi-paramétrico de Cox, denominado modelo AG (Andersen - Gill) (ANDERSEN; GILL, 1982).

Diante deste cenário promissor para o uso desta fonte alternativa de energia, este trabalho propõe o uso de modelos de eventos múltiplos para analisar a ocorrência da radiação acima de um valor pré-estabelecido, incluindo covariáveis como a temperatura e umidade relativa do ar, a partir dos dados provenientes da cidade de São João del Rei- Minas Gerais.

## METODOLOGIA

São João del Rei é uma cidade de Estado do Minas Gerais que se estende por 1 464,3 km$^2$ e contava com 90.082 habitantes no último censo ocorrido em 2010. Está situada na latitude 21$^o$ 8' 11'' Sul e na longitude 44$^o$ 15' 43'' Oeste, com altitude média de 904 metros. O município foi escolhido como modelo para esta análise por contar com uma estação meteorológica auxiliar automática, instalada no Campus Tancredo Neves - CTan/UFSJ-Universidade Federal de São João del Rei, em convênio com o Instituto Nacional de Meteorologia - INMET / Ministério da Agricultura, Pecuária e Abastecimento - MAPA.

Com o objetivo de encontrar os valores da radiação solar em São João del Rei foram utilizados os dados do programa *Sundata* disponibilizado pelo Centro de Referência para Energia Solar e Eólica Sérgio Brito - CRESESB/ Centro de Pesquisas de Energia Elétrica -CEPEL (CRESESB, 2021). O *Sundata* é uma plataforma hospedada na web composta por modelos climatológicos, capazes de informar dados estimados de energia solar (Wh/m$^2$) a partir de consultas por coordenadas geográficas. O programa fornece os dados de irradiação solar para no mínimo 3 localidades disponíveis próximas do ponto de interesse. São fornecidos os valores de irradiação solar, em KWh/m$^2$ por dia no plano horizontal, correspondentes às diárias médias mensais para os 12 meses do ano.

Os dados analisados são referentes aos valores diários no período de 01/01/2020 a 31/12/2020, obtidos no site do INMET, nos quais constam a radiação em cada horário do dia e além da radiação, foram consideradas a temperatura e a umidade relativa do ar.

Para a análise dos dados foi usada uma técnica estatística denominada análise de sobrevivência. Nesta análise a variável estudada é o tempo até a ocorrência de um evento de interesse. Outro aspecto relevante é a presença de dados censurados, que podem ser compreendidos como dados faltantes, incompletos ou perdidos. No caso deste trabalho a variável de interesse é o horário até a ocorrência da radiação acima de 2500 KJ/m$^2$.

Escolheu-se esta irradiância devido ao fato de que o valor mínimo da média de radiação solar no Estado de Minas Gerais é igual a 4,5 KW/m$^2$ por hora, para cada dia, o que corresponde

a uma média horária de 0,643 KW por hora. O que equivale a um valor médio de 2314 KJ/m$^2$ registrado a cada hora. Nos dias em que não houve radiação superior 2500 KJ/m$^2$ considerou-se a ocorrência de uma censura.

Além disso, a ocorrência de vários horários, em um mesmo dia, em que o limite de radiação estabelecido é atingido, origina os chamados eventos múltiplos, que devem ser tratados com modelos estatísticos apropriados.

Com o objetivo de ajustar os dados e escolher o modelo para fazer o ajuste, foi utilizado o modelo semi-paramétrico de Cox. O modelo de regressão de Cox permite a análise de dados cuja resposta é o tempo até a ocorrência do evento de interesse (COLOSIMO; GIOLO, 2006). Como o objetivo é verificar o tempo até a ocorrência da radiação acima de 2500 KJ/m$^2$ o modelo de Cox é adequado.

O modelo dado pela expressão

$$\lambda(t|x) = \lambda_0(t)exp\{\boldsymbol{x}^T\beta\} \quad (1)$$

Possui um componente não paramétrico $\lambda_0\ (t)$ denominado taxa de falha basal, isto se dá porque quando x = 0 tem-se que

$$\lambda(t|x) = \lambda_0(t)$$

e um componente paramétrico dado por $exp\{x^T\beta\}$, onde $\beta$ é o vetor de parâmetros que se deseja estimar.

Quando há eventos múltiplos, os quais são entendidos como aqueles que ocorrem mais de uma vez, podem ser utilizadas extensões do modelo de Cox.

Uma destas extensões é o modelo AG (Andersen - Gill) (ANDERSEN; GILL, 1982), neste modelo os eventos são ordenados e independentes, isto é, o momento de ocorrência de cada evento é "independente do tempo decorrido anteriormente ou do número de eventos ocorridos até então" (CARVALHO *et al*., 2011). No modelo AG O risco basal não varia entre os eventos, que neste trabalho correspondem ao número de horários em que a irradiância solar foi igual ou superior a 2500 KJ/m$^2$.

O modelo AG é dado pela expressão

$$\lambda(t|x) = \lambda_0(t)exp\{\boldsymbol{x}^T\beta_i\} \quad (2)$$

Para verificar a qualidade do ajuste será feita a análise de resíduos. Segundo Carvalho *et al* (2011) os resíduos escore são bastante úteis para se avaliar a influência de cada observação no modelo que foi ajustado. Uma maneira bem simples é calcular para cada observação *i*

$$\Delta\beta = \hat{\beta} - \hat{\beta}_{(-i)} \quad (3)$$

Que é a diferença entre o vetor de parâmetros estimados pelo modelo e o mesmo vetor sem o dia *i*. Se esta diferença for zero a *i*- ésima observação tem pouca importância no modelo. Assim obtém-se

$$\Delta\beta = I^{-1}\,D\,I \quad (4)$$

onde I é a matriz identidade e D é "aproximadamente igual a matriz de resíduos escore escalonada pela variância dos parâmetros *β"* (CARVALHO *et al*., 2011).

Para estimar os parâmetros do modelo selecionado foi utilizado o pacote *Survival* do software R (R DEVELOPMENT CORE TEAM, 2009).

## RESULTADOS E DISCUSSÕES

A radiação média anual para a cidade de São João del Rei corresponde a 4,83KWh /m$^2$ ao dia, sendo que o valor máximo atingido foi de 4.067 KJ/m$^2$ no dia 25 do mês de janeiro de 2020 às 16 horas. Entre 01 de janeiro de 2020 e 31 de dezembro de 2020 observou-se que foram acumulados 1.741 MWh / m$^2$. Na figura 1 tem-se o gráfico da radiação média mensal.

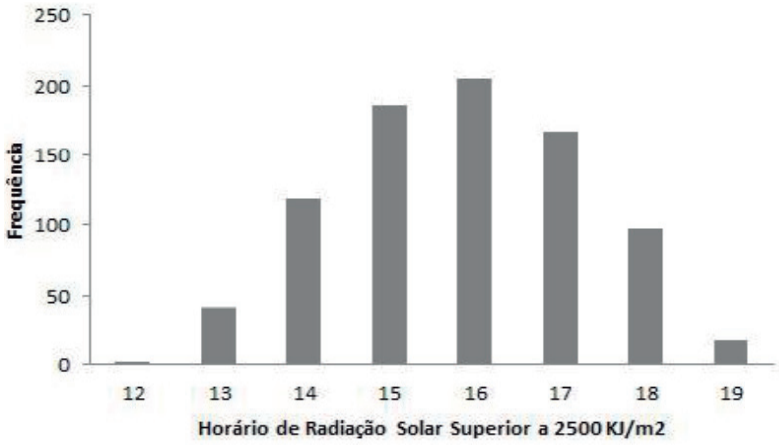
**Figura 1** - Radiação diária no plano horizontal em São João del Rei - MG.


Fonte: Os autores.

Os horários do evento de radiação superior a 2500 KJ / m$^2$ ocorreram entre 12 e 19h, sendo o horário mais frequente o das 16h. No total foram obtidos 833 horários de radiação superior a 2500 KJ/ m$^2$, sendo que a média de tal evento às 15 h e 48 min, com radiação média igual a 2971 ± 340,36 KJ /m$^2$. Em 98 dias do ano não foi atingido o pico estabelecido.

Foi elaborado o gráfico dos dados horários referentes aos dias quando a radiação solar incidente foi superior a 2.500 KJ / m$^2$ (Figura 2).

**Figura 2** - Gráfico em colunas dos horários de Radiação superior a 2500 KJ/m$^2$ em São João del Rei - MG.




Fonte: Os autores

Para realizar a análise considerando os eventos múltiplos, inicialmente foi encontrado o estimador não paramétrico de Kaplan-Meier. Este estimador foi proposto por Kaplan e Meier (1958) é utilizado para estimar a função de sobrevivência S(t), e é dado pela função:

$$\hat{S}(t) = \prod_{j:\, t_j < t} \left( \frac{n_j - d_j}{n_j} \right) \quad (5)$$

onde $t_j$ com j=1, ..., k são os horários de ocorrência dos picos de radiação estabelecidos; $d_j$ é o número de ocorrências em $t_j$ e $n_j$ é o número de horários que podem ocorrer os picos em $t_j$.

Com os dados obtidos para esta pesquisa no site do INMET foi elaborada a Tabela 1 com os valores do estimador de Kaplan-Meier (EKM) para cada horário em que ocorreu picos da radiação solar.

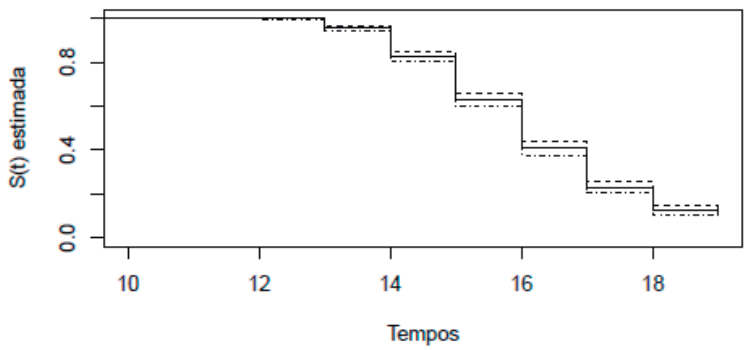
**Tabela 1** - Estimador de Kaplan- Meier

| Tempo | Sob risco | Ocorrência | Kaplan - Meier |
|---|---|---|---|
| 12 | 931 | 1 | 0,999 |
| 13 | 930 | 41 | 0,955 |
| 14 | 889 | 119 | 0,827 |
| 15 | 770 | 186 | 0,627 |
| 16 | 584 | 205 | 0,407 |
| 17 | 379 | 167 | 0,228 |
| 18 | 212 | 97 | 0,124 |
| 19 | 114 | 17 | 0,105 |

Fonte: os autores

Por meio do estimador de Kaplan-Meier observa-se (Tabela 1) que até as 12h ocorreu apenas um evento, ou seja, apenas uma vez neste horário a radiação ultrapassou o valor de 2500 KJ/m$^2$. No horário das 13h há a ocorrência de 41 casos, às 14h vê-se 119 casos, e às 19h houve 17 ocorrências. Sendo que em 98 dias não ocorreu radiação superior a 2500 KJ/m$^2$, os eventos destes dias foram considerados como censurados. A representação gráfica do estimador de Kaplan - Meier em função dos horários de ocorrência de radiação superior a 2500 KJ/m$^2$ está representado na figura 3. Considera-se neste caso a possibilidade de ocorrência de 100% dos eventos a partir das 12 h e ajusta-se por meio da técnica o percentual já ocorrido a cada intervalo. Assim, estima-se que 10,5% dos eventos foram censurados.

**Figura 3** - Estimador de Kaplan - Meier.




Fonte: Os autores

A seguir foram acrescentadas as covariáveis $X_1$ (Umidade relativa do ar) e $X_2$ (Temperatura instantânea) no modelo de Cox adaptado para eventos múltiplos AG. Analisando-se 4 possibilidades de ajuste do modelo:

1. Modelo 1: Inclusão da covariável umidade relativa do ar - $X_1$.
2. Modelo 2: Inclusão da covariável temperatura instantânea - $X_2$.
3. Modelo 3: Inclusão simultânea das covariáveis umidade e temperatura.
4. Modelo 4: Além das covariáveis umidade e temperatura, foi incluída a interação entre estas duas covariáveis ($X_1$. $X_2$).

Para seleção do modelo foi utilizado o teste da razão de verossimilhança sendo e o modelo 4 selecionado (Tabela 2).

**Tabela 2** - Resumo dos modelos

| Modelos | Covariáveis | Estimativas | Log verossimilhança |
|---|---|---|---|
| 1 | $X_1$ | $\hat{\beta}$= - 0,0499 | - 4032,14 |
| 2 | $X_2$ | $\hat{\beta}$= 0,19748 | - 4045,11 |
| 3 | $X_1$ e $X_2$ | $\widehat{\beta_1}$ = - 0,03165 e $\widehat{\beta_2}$ = 0,10512 | - 3995,57 |
| 4 | $X_1$, $X_2$ e $X_1$.$X_2$ | $\widehat{\beta_1}$= -0,27206, $\widehat{\beta_2}$= -0,2477 e $\widehat{\beta_3}$= 0,00907 | - 3730,92 |

Fonte: Os autores

Para validar o uso do modelo AG, é necessário que os tempos sejam independentes e isto ocorre quando o valor da variância robusta não é maior que duas vezes o valor da variância estimada (CARVALHO *et al*., 2011). De acordo com os resultados obtidos para a variância robusta (Tabela 3) tem-se que o valor obtido é próximo da variância estimada, isto indica que o modelo AG é adequado.

**Tabela 3** - Seleção do modelo 4 - com a inclusão das covariáveis umidade, temperatura e a interação entre umidade e temperatura.

| | ***β*** | **Variância** | **Variância robusta** |
|---|---|---|---|
| Umidade | - 0,27206 | 0,02149 | 0,02507 |
| Temperatura | - 0,2477 | 0,03414 | 0,03963 |
| Interação | 0,00907 | 0,00079 | 0,00086 |

Fonte: Os autores

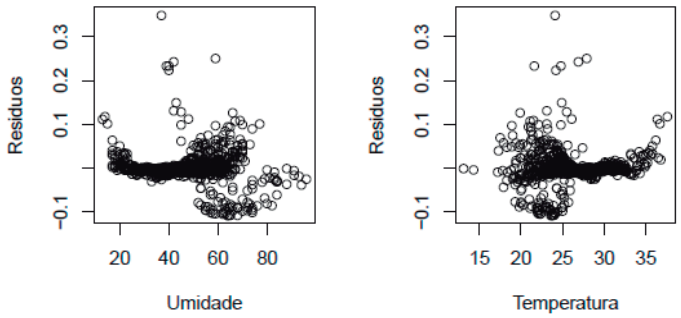
Assim de acordo com os parâmetros ajustados, o modelo escolhido é dado pela seguinte expressão:

$$\lambda_i = \lambda_0(t) = e^{[-0{,}27206x_1(t) - 0{,}2477x_2(t) + 0{,}00907x_1(t)\,x_2(t)]} \quad (6)$$

Em que $x_1(t)$ é a umidade relativa do ar, $x_2(t)$ a temperatura instantânea e $x_1(t).x_2(t)$ é a interação entre a temperatura e a umidade.

No gráfico dos resíduos (figura 4) são analisados os valores que influenciam a estimativa dos parâmetros no modelo selecionado.

**Figura 4** - Resíduos escore




Fonte: Os autores

A partir do modelo encontrado serão analisados, como exemplo de aplicação, alguns cenários, levando em consideração as variações da temperatura e da umidade relativa do ar. No primeiro cenário será mantida constante a umidade relativa do ar, no segundo será fixada a temperatura e por fim será considerada a variação tanto da umidade como da temperatura.

1) Mantendo-se a umidade relativa do ar em 65% e a temperatura varia de 30º C para 31º C, para um mesmo horário em que a radiação ultrapassa 2500 KJ/m$^2$ então:

$$\lambda(t \mid x_1, x_2) = \lambda_0(t \mid x_1, x_2) = \lambda_0(t \mid x_1, x_2)\, e^{-0{,}2706.65 - 0{,}2477.30 + 0{,}0097.65.30}$$
$$\lambda(t \mid x_1, x_2) = \lambda_0(t \mid x_1, x_2) = \lambda_0(t \mid x_1, x_2)\, e^{-0{,}2706.65 - 0{,}2477.31 + 0{,}0097.65.31}$$

$$\frac{\lambda(t \mid x_1\,(65),\ x_2\,(30))}{\lambda(t \mid x_1\,(65)\ \ x_2\,(31))} = e^{0{,}3828} = 1{,}47$$

Isto significa que o risco da radiação ultrapassar o valor de 2500 KJ / m$^2$ na cidade de São João del Rei é 1,47 vezes maior quando a temperatura varia de 30º C para 31º C, mantendo a umidade relativa do ar.

2) Mantendo-se a temperatura 30º C e a umidade relativa do ar variando de 65 % para 68 %, então,

$$\lambda(t \mid x_1, x_2) = \lambda_0(t \mid x_1, x_2) = \lambda_0(t \mid x_1, x_2)\, e^{-0{,}2706.65 - 0{,}2477.30 + 0{,}0097.65.30}$$
$$\lambda(t \mid x_1, x_2) = \lambda_0(t \mid x_1, x_2) = \lambda_0(t \mid x_1, x_2)\, e^{-0{,}2706.68 - 0{,}2477.30 + 0{,}0097.68.30}$$

$$\frac{\lambda(t \mid x_1\,(65),\ x_2\,(30))}{\lambda(t \mid x_1\,(68),\ x_2\,(30))} = e^{0{,}06128} = 1{,}06$$

O risco da radiação ultrapassar o valor de 2500 KJ / m$^2$ é 1,06 vezes maior.

3) No terceiro cenário tanto a temperatura como a umidade variam de 30º C e 65 % para 31º C e 68 % logo,

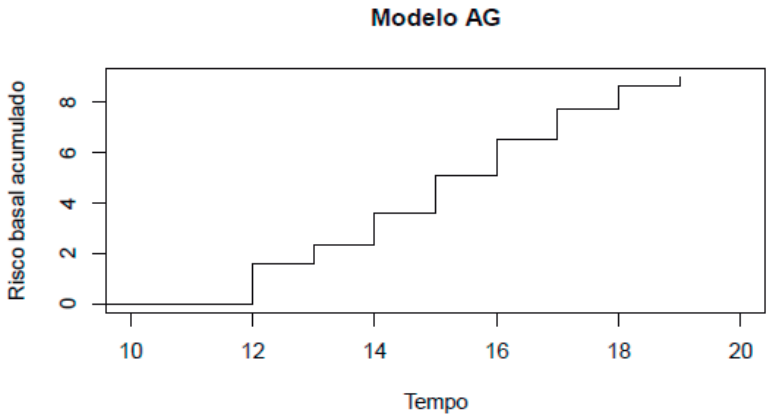
$$\lambda(t \mid x_1, x_2) = \lambda_0(t \mid x_1, x_2) = \lambda_0(t \mid x_1, x_2)\, e^{-0{,}2706.65 - 0{,}2477.30 + 0{,}0097.65.30}$$
$$\lambda(t \mid x_1, x_2) = \lambda_0(t \mid x_1, x_2) = \lambda_0(t \mid x_1, x_2)\, e^{-0{,}2706.68 - 0{,}2477.31 + 0{,}0097.68.31}$$

$$\frac{\lambda(t \mid x_1\,(65),\ x_2\,(30))}{\lambda(t \mid x_1\,(68),\ x_2\,(31))} = e^{0{,}473} = 1{,}6$$

O risco aumenta cerca de 60 %.

Na figura 5 tem-se a representação gráfica do risco basal acumulado para o modelo AG, observe que o risco aumenta à medida que um novo episódio ocorre.

**Figura 5** - Risco basal acumulado do Modelo AG




Fonte: Os autores

## CONCLUSÕES

O uso do modelo de eventos múltiplos AG pode ser uma alternativa para análise dos dados envolvendo radiação e o horário de ocorrência de picos de radiação. Por meio desta técnica verificou-se o risco de ocorrência dos picos de radiação em função da temperatura e umidade ao longo do dia, o que permite que se tenha uma ideia do melhor horário para que se obtenha um resultado mais eficiente no que diz respeito à utilização da luz solar como fonte de energia alternativa, viabilizando uma economia a longo prazo.

## REFERÊNCIAS

ABSOLAR. **Minas Gerais, a locomotiva da energia solar no Brasil**. Disponível em: https://bit.ly/37BhUWh. Acesso em: 20 de abril de 2021.

ANDERSEN, P. K., GILL, R. D. Cox's regression model for counting processes: a large sample study. **The annals of statistics**, p. 1100-1120, 1982.

CANTOR, G. A. R. **Influência dos fatores climáticos no desempenho de módulos fotovoltaicos em regiões de clima tropical**. Dissertação de Mestrado, Universidade Federal da Paraíba, 2017.

CARVALHO, M. *et al.* **Análise de sobrevivência: Teoria e aplicações em saúde**, 2. ed. Fiocruz, 2011.

COLOSIMO, E., GIOLO, S. **Análise de sobrevivência aplicada**, 1. ed. Blücher, 2006.

CRESESB (2021). **Centro de referência para as energias solar e eólica Sérgio de S. Brito**. Disponível em: https://bit.ly/3KS3Zti. Acesso em: 15 abr. 2021.

FRANCISCO, A. C. C. *et al.* Influência de parâmetros meteorológicos na geração de energia em painéis fotovoltaicos: um caso de estudo do Smart Campus Facens, SP, Brasil. **Revista Brasileira de Gestão Urbana**, 11.2019

GUIA TÉCNICO. **Energia fotovoltaica:** manual sobre tecnologias, projeto e instalação. 2014. Disponível em: https://bit.ly/3N58WRi. Acesso em: 06 de maio de 2021.

KAPLAN, E. L., MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American statistical association**, 53(282), 457-481,1958

MAIOMI, F. P., CARDOSO, R. B. Análises de viabilidades econômicas para alternativas de utilização da energia solar em residência do estado de Minas Gerais, Brasil. **Research, Society and Development**, 9(8), 1-39, 2020.

MIDEKSA, T. K., KALLBEKKEN, S. The impact of climate change on the electricity market: A review. **Energy Policy**, 38(7), 3579-3585, 2010.

MILANO, J. Proposta de utilização de sistema híbrido eólico/solar de energia em estabelecimentos comerciais na Ilha do Mel- PR. **Revista Ciência e Natura**, 40 e 66, 2018.

PEREIRA, E. B. *et al.* **Atlas brasileiro de energia solar**. INPE, 2017

PRADO, R. *et al.* **Levantamento do estado da arte: Energia solar**. Projeto tecnologias para construção habitacional mais sustentável. São Paulo Projeto FINEP, 2386(04), 2017.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Disponível em: http://www.R-project.org, 2009.

TIBA, C. *et al.* Atlas solarimétrico do Brasil: banco de dados terrestre. Recife: **Ed. Universitária da UFPE**, 2020.

WANDERLEY, A., CAMPOS, A. Perspectivas de inserção da energia solar fotovoltaica na geração de energia elétrica no Rio Grande do Norte. **Holos**, 3(29), 3-14, 2013.

ARTIGO 4

**Modelo de Regressão com covariável circular: teoria e aplicação em dados sobre atendimento de ocorrência de acidentes de trânsito**

# Modelo de Regressão com covariável circular: teoria e aplicação em dados sobre atendimento de ocorrência de acidentes de trânsito


## Resumo

Acidentes de trânsito têm sido causa de preocupação devido aos custos com prejuízos tanto sociais quanto econômicos associados a eles. Alguns fatores devem ser levados em consideração quanto à ocorrência destes acidentes, entres eles os horários, dias da semana, meses do ano e as localidades nas quais estes acidentes ocorrem, pois os cenários do trânsito podem variar. A modelagem do tempo para o atendimento aos envolvidos nos acidentes, analisando uma possível periodicidade nos horários, pode contribuir para melhorar o planejamento dos órgãos de trânsito. Nesta pesquisa foi analisado o tempo até o início dos atendimentos de acidentes de trânsito, considerando três regiões do estado de Minas Gerais durante os anos 2020 e 2021 e sua relação com o horário de ocorrência. Assim sendo, faz-se necessário utilizar uma estatística própria para dados circulares. Além disso, devido a falhas ou ausências de registro na planilha de dados ocasionadas pelo atraso prolongado no atendimento de algumas ocorrências ou problemas técnicos, pode-se tratar esses dados como censura. Portanto, propôs-se neste trabalho o uso da técnica de análise de sobrevivência, utilizando a covariável circular, ou seja, um modelo de regressão linear-circular foi utilizado no contexto da análise de sobrevivência para estimar o tempo de início do atendimento às vítimas de acidentes de trânsito. A qualidade do ajuste foi verificada por meio da análise de resíduos e o modelo foi considerado satisfatório.



**Palavras-chave**: Análise de sobrevivência, Log normal, vítimas.


## Introdução

Acidentes automobilísticos têm sido causa de preocupação, pois além das perdas de vidas, causam problemas sociais e econômicos, que vão desde afastamentos do trabalho, invalidez e outros relacionados com a vítima e seus parentes (MARÍN; QUEIROZ, 2000; RESENDE et al., 2012; DINIZ *et al.*, 2008).

Deslandes (1999) verificou, ao analisar a distribuição dos atendimentos por hora de entrada em hospitais, no estado de São Paulo, que nos horários das 8 às 20 horas há uma demanda maior, exceto nos finais de semana, quando o cenário se inverte, o período com maior número de atendimentos é das 20 às 8 horas. Foi observado, também, um aumento dos casos a partir da sexta-feira e uma diminuição aos domingos.

Em acidentes que envolvem vítimas, quanto mais rápido for o atendimento menores poderão ser os riscos, pois a demora do atendimento pode levar a consequências como o agravamento das condições de saúde dos envolvidos no acidente e até a morte (MALESVIO; SOUZA, 2002). Logo, quanto mais rápido for o socorro melhores podem ser os resultados, inclusive preservar a vida da vítima.

Portanto, é relevante avaliar o tempo para o início dos atendimentos às vítimas, pois quanto mais rápido ele acontecer poderá implicar na redução na morte ou em outros problemas para a vítima (LADEIRA; BARRETO, 2008). Além disso, também é importante investigar o horário em que o acidente ocorre, já que eles podem ser mais comuns em determinados horários e períodos do dia, dependendo da região (DIAS et al., 2018; ASCARI et al., 2013).

A qualidade do registro no local das ocorrências é muito importante, pois essas informações serão úteis para identificar os dias e horários com maior demanda e que podem requerer maior atenção (DESLANDES, 2018).

Para se avaliar o tempo para o início dos atendimentos às vítimas foi utilizado uma técnica estatística chamada análise de sobrevivência, cujo objetivo é estimar o tempo até a ocorrência de um evento (COLOSIMO; GIOLO, 2006). No caso desta pesquisa, o tempo até o início do atendimento a um acidente de trânsito.

Uma proposta deste trabalho é utilizar a variável explicativa, horário dos acidentes, que será considerada como uma medida angular e, por isso, utilizar a estatística para dados circulares (FISHER, 1993; MARDIA, 1972). Ou seja, o modelo de regressão próprio para dados circulares, no qual a variável explicativa é circular (o horário do acidente) e a variável resposta é linear (o tempo até o início do atendimento). Assim, o modelo linear - circular proposto por Mardia e Sutton (1978) foi utilizado no contexto da análise de sobrevivência.

Para a realização desta pesquisa fez-se uso de um banco de dados obtido junto à Polícia Militar de Minas Gerais (PMMG) que contém registros diários dos acidentes, com os horários e tipo de acidentes e o horário de início dos atendimentos por parte da polícia, em três regiões atendidas por três batalhões da polícia militar localizados nas cidades de Lavras, São João del Rei e Teófilo Otoni.

Inicialmente algumas informações da estatística circular foram consideradas, inclusive foi verificado se os dados de cada região, por ano, se ajustavam à distribuição von Mises. O tempo decorrido desde o horário do acidente até o início aos atendimentos das vítimas foi avaliado.

Em algumas situações o tempo para se iniciar o atendimento foi superior a 12 h ou 720 minutos, e esta informação foi considerada como censurada, o que significa que a informação foi tida como perdida ou faltante, já que para a análise desta pesquisa o tempo máximo para início dos atendimentos foi de 12h (CARVALHO et al., 2011).

Este artigo está organizado da seguinte maneira, na primeira seção é feita uma revisão de alguns conceitos da estatística circular, na segunda uma abordagem sobre a análise de sobrevivência, incluindo alguns modelos de distribuição de probabilidade. Nas duas seguintes foram discutidas a aplicação relativa aos horários dos acidentes em algumas rodovias de municípios mineiros e a análise do tempo para se iniciar os atendimento às pessoas envolvidas nos acidentes, e por fim as conclusões.

## Estatística circular

Dados circulares são apropriados para a análise de informações cuja medida é angular. Esses dados estão presentes em informações relacionadas como, por exemplo, direção dos ventos e vôos dos pássaros, ou medidas, que podem ser convertidos em angulares, como horários, dias e meses do ano. Os dados podem ser medidos em graus ou em radianos.

Cada observação circular pode ser representada geometricamente como um ponto em circunferência de um círculo de raio $r$. Dado um ponto $P = (x, y)$ no plano cartesiano, este ponto pode também ser representado na forma trigonométrica, no sentido anti-horário, como $x = r\cos\theta$ e $y = r\sin\theta$, ou seja, $P$ também pode ser representado por $P = (r\cos\theta, r\sin\theta)$. Se o círculo é unitário, ou seja, $r = 1$, o ponto $P$ será representado por $P = (\cos\theta, \sin\theta)$ (FISHER, 1993).

Assim como se dá com as observações na reta real, em que há medidas tendência central como média, mediana e moda, medidas de dispersão, também existem essas medidas para

estatística circular, porém os conceitos não são os mesmos utilizados para os dados que estão localizados na reta real ($\mathbb{R}$).

Considerando $(\bar{C}, \bar{S})$, no qual

$$\begin{cases} \bar{C} = \dfrac{1}{n}\sum_{i=1}^{n}\cos\theta_i \\ \\ \bar{S} = \dfrac{1}{n}\sum_{i=1}^{n}\sin\theta_i \end{cases} \tag{1}$$

Então a média direcional $\bar{\theta}$ é a solução das equações:

$$\begin{cases} \bar{C} = \bar{R}\cos\bar{\theta} \\ \\ \bar{S} = \bar{R}\sin\bar{\theta} \end{cases} \tag{2}$$

$R$ será chamado de comprimento resultante e $\bar{R}$ de comprimento médio resultante, o qual é dado por

$$\bar{R} = \sqrt{\bar{C}^2 + \bar{S}^2} \tag{3}$$

Por sua vez $R^2 = C^2 + S^2$ e $\bar{R} = \dfrac{R}{n}$.

Observe que $0 \leq \bar{R} \leq 1$.

A média direcional é a solução da equação 2 e :

$$\bar{\theta} = \begin{cases} \arctan\dfrac{S}{C}, & S > 0, C > 0 \\ \\ \arctan\dfrac{S}{C} + \pi, & C < 0 \\ \\ \arctan\dfrac{S}{C} + 2\pi, & S < 0, C > 0 \end{cases} \tag{4}$$

A média direcional não está definida quando $S = C = 0$.

A mediana direcional dos ângulos $\theta_1, ..., \theta_n$ é qualquer ângulo $\phi$ tal que:

a) Metade dos pontos situa-se no arco $[\phi, \phi + \pi)$;

b) maioria dos pontos está mais próxima de $\phi$ do que de $\phi + \pi$.

Quando o tamanho da amostra, $n$, é ímpar, a mediana é um único ângulo, e quando $n$ é par a mediana será a média aritmética entre os dois ângulos centrais e adjacentes.

De acordo com Mardia (1972), a mediana direcional $\phi$ também pode ser definida como a solução da equação:

$$\int_{\phi}^{\phi+\pi} f(\theta)d\theta = \frac{1}{2} \tag{5}$$

na qual $f(\theta)$ é a função densidade de probabilidade de $\theta$.

Assim como nos dados lineares, a mediana é o segundo quartil $Q_2$, e em função dele podem ser determinados os quartis $Q_1$ e $Q_3$, respectivamente, o primeiro e terceiro quartis. Com esses valores, é possível ainda construir um boxplot circular (BUTTARAZZI; PANDOLFO; PORZIO, 2018; ABUZAID; MOHAMED; HUSSIN, 2021).

Se $\bar{R}$ tem um valor próximo de 1 os dados direcionais estão bem agrupados, caso $\bar{R}$ esteja próximo de zero há uma dispersão dos dados. Notando, assim, que $\bar{R}$ é uma medida de concentração dos dados (MARDIA, 1972). A partir deste fato será dado o conceito da variância circular.

A variância circular é dada por

$$V = 1 - \bar{R} \tag{6}$$

onde $0 \leq V \leq 1$.

Quanto mais próximo de 0 for o valor da variância circular mais concentrados estarão os dados.

O desvio padrão circular de uma amostra é definido pela expressão

$$\nu = \{-2\log(1-V)\}^{\frac{1}{2}} \tag{7}$$

### Modelo de probabilidade

Para a análise de dados provenientes de medidas angulares ou direcionais são utilizados modelos de probabilidade próprios. Se $f(\theta)$ é uma função de densidade de probabilidade de uma variável aleatória circular $\Theta$, então esta função satisfaz as seguintes propriedades:

a) $f(\theta) \geq 0$;

b) $f(\theta + 2\pi) = f(\theta)$;

c) $\displaystyle\int_0^{2\pi} f(\theta)d\theta = 1.$

Um dos modelos de probabilidade muito utilizado para dados circulares é o modelo von Mises, o qual será denotado por $vM(\mu, \kappa)$, o parâmetro $\mu$ é a média direcional e $\kappa$ é o parâmetro de concentração (JAMMALAMADAKA; SENGUPTA, 2001).

O modelo é dado por:

$$g(\theta, \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa\cos(\theta-\mu)} \tag{8}$$

onde $I_0$ é a função de Bessel modificada, de primeiro tipo e ordem 0, que pode ser definida por

$$I_0(\kappa) = \frac{1}{2\pi}\int_0^{2\pi} e^{\kappa\cos\theta}\, d\theta \tag{9}$$

## Análise de Sobrevivência

A análise de sobrevivência pode ser entendida como um conjunto de procedimentos estatísticos que são úteis para realizar a análise de dados nos quais a variável de interesse é o tempo até que um determinado evento ocorra (KLEINBAUM; KLEIN, 2012). O tempo pode ser dado em minutos e é contado a partir do início de um evento, o qual pode ser o início para o atendimento às vítimas de um acidente. Quando o evento ocorre, diz-se que houve a falha, no estudo em questão o evento ocorre quando o atendimento às vítimas é iniciado.

Os dados relativos à falha e à censura serão representados pelo par $(t_i, \delta_i)$, onde $\delta_i$ é a função indicadora de falha ou censura, assumindo que $\delta_i = 1$ se o dado indica que ocorreu a falha (não é censurado) e $\delta_i = 0$ se o dado é censurado.

A função de sobrevivência, representada por $S(t)$, é definida como sendo a probabilidade de um indivíduo sobreviver mais que um tempo especificado $t$, ou seja, é a probabilidade de que a variável aleatória $T$ seja maior que um tempo especificado $t$, isto é, $S(t) = P(T \geq t)$.

Então

$$S(t) = P(T \geq t) = 1 - F(t) \tag{10}$$

sendo $F(t)$ a função densidade acumulada de T com função densidade de probabilidade $f(t)$.

Vale mencionar também o conceito de função de risco (ou taxa de falha), denotada por $\lambda(t)$, esta função é dada por

$$\lambda(t) = \lim_{\Delta \to 0} \frac{P(t \leq T < t + \Delta t) | T \geq t}{\Delta t} \tag{11}$$

A função de risco $\lambda(t)$ dá a maneira em que a taxa instantânea de falha é modificada com o tempo.

A função de risco acumulada é dada por

$$\Lambda(u) = \int_0^t \lambda(u)\, du \tag{12}$$

A função de verossimilhança será representada por:

$$L(\theta) = \prod_{i=1}^{n} [f(t_i, \beta)]^{\delta_i} [S(t_i, \theta)]^{1-\delta_i} \tag{13}$$

sendo $\delta_i$ a função indicadora de falha ou censura, $\beta$ é o vetor de parâmetros e $S(t)$ é a função de sobrevivência.

### Estimador de Kaplan- Meier

O estimador não paramétrico de Kaplan-Meier é utilizado para estimar a função de sobrevivência, e é muito utilizado para verificar se o modelo paramétrico se ajusta bem aos dados.

Considere os $k$ tempos de falhas ordenados e distintos, $t_1 < t_2 < ... < t_k$, e $d_j$ o número de falhas em $t_j$, com $j = 1, ..., n$ e $n_j$ é o número de indivíduos que estão sob risco em $t_j$ (COLOSIMO; GIOLO, 2006). O estimador de Kaplan - Meier é dado por:

$$\hat{S}(t) = \prod_{j:t_j<t} \left( \frac{n_j - d_j}{n_j} \right) = \prod \left( 1 - \frac{d_j}{n_j} \right) \tag{14}$$

## Alguns modelos probabilísticos

Existem três abordagens na análise de sobrevivência. A não paramétrica (quando é utilizado o estimador de Kaplan-Meier), o semiparamétrico - quando se utiliza o modelo de Cox, e o paramétrico- neste caso, se faz uso de alguns modelos para estimar a variável aleatório o tempo, como Weibull, Gama e Log normal.

A sua função densidade de probabilidade da distribuição Weibull é

$$f(t) = \alpha\gamma(\alpha t)^{\gamma-1} \exp\left[-(\alpha t)^{\gamma}\right] \tag{15}$$

a função de Sobrevivência é

$$S(t) = \exp\left[-(\alpha t)^{\gamma}\right] \tag{16}$$

e a função taxa de falha (função de risco) é

$$\lambda(t) = \alpha\gamma(\alpha t)^{\gamma-1} \tag{17}$$

A função densidade de probabilidade de uma variável aleatória $T$é dada por

$$f(t) = \frac{1}{\sigma t\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right), \quad t > 0$$

$\mu$ é a média do logarítmo do tempo de falha e $\sigma$ é o desvio padrão.

A função de sobrevivência da distribuição log-normal é dada por:

$$S(t) = \Phi\left(\frac{-\log(t)+\mu}{\sigma}\right) \tag{18}$$

em que $\Phi$ é a função distribuição acumulada de uma normal padrão.

Outra função de densidade utilizada é distribuição gama. Ela possui dois parâmetros: o parâmetro de forma $k$ e o de escala $\alpha$ e é dada por:

$$f(t) = \frac{t^{k-1}}{\Gamma(k)\alpha^{\kappa}} \exp\left(-\frac{t}{\alpha}\right) \tag{19}$$

em que $\Gamma(k)$ é a função gama dada por

$$\Gamma(k) = \int_0^{\infty} \exp(-x)x^{k-1}dx$$

A função de sobrevivência é dada por

$$S(t) = \int_t^{\infty} \frac{u^{k-1}}{\Gamma(k)\alpha^{k}} \exp\left(-\frac{u}{\alpha}\right) du \tag{20}$$

# Modelos de regressão na Análise de sobrevivência

No contexto da análise de sobrevivência, a variável resposta, o tempo até a ocorrência de um evento, segue uma distribuição de probabilidade, que é uma função contínua, não negativa e assimétrica. Considerando, por exemplo, o modelo $T = \exp\{\beta_0 + \beta_1 x\}\epsilon$.

Este modelo pode ser linearizado da seguinte maneira

$$y = \log T = \beta_0 + \beta_1 x + \nu \tag{21}$$

sendo $\nu = \log\epsilon$, Note que neste modelo a distribuição dos erros não é necessariamente normal.

O modelo ( 21) pode ser generalizado, incluindo um parâmetro de forma $\sigma$. O tempo T é dado por:

$$\log T = \mu + \sigma\nu \tag{22}$$

Tem-se que $\mu$ é o parâmetro de locação e $\sigma$ o parâmetro de escala. $\sigma$ é uma variável aleatória que segue alguma distribuição de probabilidade que possa representar $y = \log T$ (CARVALHO et al., 2011; COLOSIMO; GIOLO, 2006).

Considerando $\mathbf{x}^T = (1, x_1, ..., x_p)$ o vetor de covariáveis e $\beta = (\beta_0, \beta_1, ..., \beta_n)^T$ o vetor de parâmetros que serão estimados. Se $T$ é o tempo de sobrevivência, e $\alpha = \exp\{x^T\beta\}$, onde $x$ é o vetor de covariáveis e $\beta$ é o vetor de parâmetros, a função de sobrevivência para a distribuição Weibull é dada por:

$$S(t|x) = \exp\left(-\left(\frac{t}{\exp\{x^T\beta\}}\right)^{\frac{1}{\sigma}}\right) \tag{23}$$

Se $Y = \log T$, a função de sobrevivência será:

$$S(y|x) = \exp\left(-\exp\left(\frac{y - x^T\beta}{\sigma}\right)\right) \tag{24}$$

Considere agora que $Y = \log T$, as funções de densidade de probabilidade e de sobrevivência da distribuição log normal são dadas, respectivamente, por

$$f(y|x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y - x^T\beta}{\sigma}\right)^2\right) \tag{25}$$

$$S(y|x) = 1 - \Phi\left(\frac{y - x^T\beta}{\sigma}\right) \tag{26}$$

Com a finalidade de escolher o modelo que melhor se ajusta aos dados, podem ser utilizados alguns critérios como, o critério de informação de Akaike (AIC) e o critério de informação Bayesiano (BIC).

O critério de informação de Akaike (AIC) é dado por

$$AIC = -2\log(L) + 2p$$

e o critério de informação Bayesiano (BIC) é dado por

$$BIC = -2\log(L) + p\log(n)$$

em que $L$ é a função de verossimilhança estimada nos valores dos parâmetros do modelo, $p$ é o número de parâmetros e $n$ é o tamanho da amostra. Será indicado como melhor modelo o que apresentar menor valor de AIC e BIC.

A análise de resíduos é importante para se verificar a qualidade do ajuste. Os resíduos de Cox-Snell auxiliam na avaliação do ajuste global do modelo e eles são dados por:

$$\hat{e}_i = \hat{\Lambda}(t_i|x_i) \tag{27}$$

sendo $\Lambda()$ a função de taxa de falha acumulada. Exemplificando, para o modelo de regressão log-normal o resíduo de Cox-Snell será:

$$\hat{e}_i = -\log\left[1 - \Phi\left(\frac{\log(t_i) - X_i'(\hat{\beta})}{\hat{\sigma}}\right)\right] \tag{28}$$

Os resíduos de Cox-Snell $\hat{e}_i$ vêm de uma população homogênea, e se o modelo for adequado para representar os dados, os resíduos devem seguir uma distribuição exponencial padrão (COLOSIMO; GIOLO, 2006)

Os resíduos martingale, são baseados em processos de contagem individual e são dados por:

$$M_i = \delta_i - \hat{e}_i \tag{29}$$

em que $\delta_i$ é a função indicadora de censura e $\hat{e}_i$ é o resíduo de Cox-Snell, vale mencionar que os resíduos $M_i$ não são simetricamente distribuídos em torno do 0 (zero) (CARVALHO et al., 2011).

Os resíduos deviance são definidos por

$$r_{D_i} = \text{sinal}(\hat{M}_i)\sqrt{-2(\tilde{l}_i - l_i)} \tag{30}$$

no qual o sinal de $(\hat{M}_i)$ é o sinal do resíduo martingale para o i-ésimo indivíduo e $\tilde{l}_i$ e $l_i$ são os valores do logaritmo da função de verossimilhança para a observação $i$ do modelo dado e saturado, respectivamente. O resíduo deviance está distribuído simetricamente em torno da reta $y = 0$ (CARRASQUINHA; VERÍSSIMO; VINGA, 2018).

## Metodologia

Nesta pesquisa foi analisado o tempo até o início dos atendimentos de acidentes de trânsito considerando três localidades (Lavras, São João del Rei e Teófilo Otoni) do estado de Minas Gerais durante os anos de 2020 e 2021. O que está sendo considerado como tempo para início do atendimento é o tempo decorrido entre o horário do acidente e horário em que a viatura policial chega ao local do acidente. O banco de dados foi obtido junto a Polícia Militar de Minas Gerais (PMMG) e contém registros de acidentes de trânsito, com o horário, data , localização, e o tipo de acidente.

As localidades foram separadas pela proximidade das cidades de Lavras, São João del Rei e Teófilo Otoni, que abrigam um batalhão de polícia atendendo as cidades listadas a seguir:

> **Lavras** - Boa Esperança, Bom Sucesso, Campo Belo, Cana Verde, Candeias, Carmo da Cachoeira, Ibituruna, Ijaci, Itumirim, Itutinga, Ingaí, Lavras, Luminárias, Perdões, Ribeirão Vermelho, Santo Antonio do Amparo, Varginha.
>
> **São João del Rei** - Andrelândia, Barroso, Bom Jardim, Conceição da Barra de Minas, Coronel Xavier Chaves, Dores de Campos, Lagoa Dourada, Madre Deus, Nazareno, Piedade do Rio Grande, Prados, Resende Costa, Ritápolis, Santa Cruz de Minas, São João del Rei, São Tiago, São Vicente, Tiradentes.
>
> **Teófilo Otoni** - Ataleia, Carlos Chagas, Frei Gaspar, Ladainha, Malacacheta, Novo Oriente de Minas, Ouro Verde, Poté, Teófilo Otoni.

Registrou-se 1160 horários no período estudado, os anos 2020 e 2021, distribuídos da seguinte maneira:

1. - Região de Lavras : 469 horários.
2. - Região de São João del Rei: 273 horários.
3. - Região de Teófilo Otoni: 418 horários.

Cada horário de acidente foi transformado numa medida angular, utilizando equivalência que cada 1 hora corresponde a 15°. Para a variável circular, horário do acidente, algumas medidas da estatística circular descritiva como média direcional, mediana, variância e desvio padrão foram obtidas. Isto foi feito para o período de 2020 e 2021, em conjunto.

Na análise dos dados observou-se que em algumas informações pode ter ocorrido erros nos registros, pois o horário do início do atendimento era um horário anterior ao da ocorrência do fato. Em outras situações havia dados em que o horário do acidente era o mesmo do horário da chamada, ou diferiam de poucos segundos. Assim, para não comprometer a análise estas situações foram excluídas. Dessa forma, os 670 dados restantes estavam distribuídos da seguinte maneira: 300 para Lavras e região, 217 para São João del Rei e região e 153 para Teófilo Otoni e região.

Além disso, quando o tempo para iniciar um atendimento for superior a 12 h ou 720 minutos, a informação foi considerada censurada, pois nestes casos houve falhas de comunicação ao relatar o atendimento do acidente. Assim, foram encontrados 36 dados censurados, sendo 15 para a região de Lavras, 10 para São João del Rei e 11 em Teófilo Otoni.

O passo seguinte foi testar os modelos de distribuição Weibull, Log normal e Gama para variável resposta, tempo até o início do atendimento e adaptar um modelo linear-circular, no qual a variável explicativa circular é o horário do acidente. O modelo é baseado numa distribuição no cilindro (MARDIA; SUTTON, 1978; ANDERSON-COOK; NOBLE, 2001), em que a altura da curva no cilindro é a parte linear do modelo, figura ( 1). O modelo, proposto inicialmente por Mardia e Sutton (1978), considera uma distribuição conjunta, na qual a variável circular $\Theta$ tem distribuição von Mises ($vM(\mu_0, \kappa)$), e a componente linear $x$ segue a distribuição Normal condicionada ao valor de $\theta$.

Assim, a distribuição conjunta é dada por

$$f(x|\theta) = \frac{1}{2\pi I_0(\kappa)} \exp\{\kappa \cos(\theta - \mu_0)\} \frac{1}{\sigma_c^2\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_c)^2}{2\sigma_c^2}\right) \tag{31}$$

em que $x, \mu_c \in \mathbb{R}$, $\theta, \mu_0 \in [0, 2\pi)$, $\kappa > 0$ e

$$\mu_c = \mu + \frac{\sigma}{\sqrt{\kappa}}\{\rho_1(cos\theta - \cos\mu_0) + \rho_2(\sin\theta - \sin\mu_0)\} \tag{32}$$

com $\sigma_c^2 = \sigma^2(1-\rho^2)$ e $\rho^2 = \rho_1^2 + \rho_2^2, 0 \leq \rho \leq 1$

Como uma medida angular pode ser vista como um ponto no círculo de uma circunferência unitária, pode ser escrito como $(\cos\theta, \sin\theta)$. Cada horário de um acidente será visto como um ponto numa circunferência e esta medida será associada a um ângulo, medido em radianos, e o tempo $t$ será a variável resposta, $0 \leq t \in \mathbb{R}$.

Em relação ao modelo proposto, a variável resposta tempo é sempre positiva, $t > 0$, mas o modelo pode ser linearizado e $-\infty < y = \log(t) < \infty$. Os outros parâmetros permanecem inalterados assim, essa alteração não viola as pressuposições do modelo original.

O modelo estudado por Mardia e Sutton (1978), equação (31), será adaptado para a análise de sobrevivência. Na ideia original se trata de uma função de densidade conjunta dada pela distribuição normal e a distribuição von Mises. Nesta proposta, será a distribuição conjunta entre a distribuição log normal e a von Mises, não violando nenhum pressuposto do modelo proposto originalmente.

O modelo de regressão é dado por

$$T = \exp\{\beta_0 + \beta_1 \cos(\theta_i) + \beta_2 \sin(\theta_i) + \nu_i\}$$

O qual pode ser linearizado

$$Y = \log T = \beta_0 + \beta_1 \cos\theta_i + \beta_2 \sin\theta_i + \epsilon_i \tag{33}$$

Assim, $y_i(t)$ segue a distribuição log-normal, e $\theta_1, ..., \theta_n$ são as variáveis exploratórias angulares, $\epsilon_i$ é o erro, que é considerado como tendo média 0 e variância constante; $\beta_0, \beta_1, \beta_2$ são os parâmetros a serem estimados.

Na forma matricial o modelo será dado por

$$\mathbf{Y}_i = \mathbf{X}\beta^T + \epsilon_i \tag{34}$$

A estimação dos parâmetros do modelo é realizada por meio do método da máxima verossimilhança, da seguinte maneira:

$$L = \prod_{1=1}^{n} \left[ \frac{1}{\sqrt{2\pi}\sigma} \exp\left( -\frac{1}{2} \left( \frac{y - \mathbf{X}\beta^T}{\sigma} \right)^2 \right) \right]^{\delta_i} \left[ 1 - \Phi\left( \frac{y - \mathbf{X}\beta^T}{\sigma} \right) \right]^{1-\delta_i} \tag{35}$$

Encontrando $l = \log L$, que é o log verossimilhança,

$$l = \sum_{i=1}^{n} \left[ \delta_i \left( \log\left( \frac{1}{\sqrt{2\pi}\sigma} \right) - \frac{1}{2} \left( \frac{y - \mathbf{X}\beta^T)}{\sigma} \right)^2 \right) + (1 - \delta_i) \log\left( 1 - \Phi\left( \frac{y - \mathbf{X}\beta^T}{\sigma} \right) \right) \right] \tag{36}$$

Onde $\delta_i$ é a função indicadora de falha ou censura e $\mathbf{X}\beta^T$ é dada por:

$$\begin{pmatrix} 1 & \cos\theta_1 & \sin\theta_1 \\ 1 & \cos\theta_2 & \sin\theta_2 \\ \vdots & \vdots & \vdots \\ 1 & \cos\theta_n & \sin\theta_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

sendo $\beta$ a matriz dos parâmetros que serão estimados. As estimativas dos parâmetros foram obtidas por meio do pacote survival do software R (R Development Core Team, 2021), sendo utilizado o método de Newton-Raphson.

Para seleção do modelo foram utilizados os critérios AIC e BIC. Covariáveis dicotômicas foram incluídas (acidente com ou sem vítimas, o acidente ocorreu no sábado / domingo ou segunda - sexta).

O teste log-rank foi utilizado para verificar se havia diferença ou não entre os tempos de início dos atendimentos, quando há vítimas no acidente e quando não há. Também se verificou tal fato quando o acidente ocorre nos dias de sábado ou domingo, ou nos demais dias da semana, sem levar em consideração se neste dia houve ou não feriados.

Após a seleção do modelo, foi realizada a análise de resíduos com a finalidade de avaliar a qualidade do ajuste. Uma maneira de avaliar a qualidade do modelo é por método gráfico. Para utilizar este método realiza-se a linearização da função de sobrevivência e faz uma comparação como o estimador de Kaplan-Meier.

Outro método gráfico é comparar o modelo proposto com o estimador de Kaplan-Meier, neste caso, o modelo se ajusta bem aos dados se os pontos do gráfico estiverem próximos da reta $y = x$, onde $y$ é a função de sobrevivência do modelo que está sendo avaliado e $x$ a função do estimador de Kaplan- Meier.

Para que o modelo de regressão esteja bem ajustado, os resíduos de Cox-Snell devem seguir a distribuição exponencial padrão, os modelos são considerados adequados quando curvas de sobrevivência se aproximam das respectivas curvas do estimador de Kaplan-Meier. Outros resíduos que podem ser aplicados são os resíduos deviance e martingale.

## Resultados

Foram construídos gráficos dos horários dos 1160 acidentes ocorridos durante os anos de 2020 e 2021, das três localidades em conjunto. Um gráfico linear (no qual no eixo X está representado o número de acidentes e no eixo Y o horário dos acidentes) e um gráfico circular, onde cada ponto no círculo unitário é o horário de um acidente. A figura ( 2) mostra estes gráficos.

Nota-se que apresentação do gráfico circular permite uma melhor visualização dos horários de ocorrência dos acidentes.

Com os horários dos acidentes convertidos em radianos foi calculada, para efeito de comparação, a média aritmética. O valor encontrado foi $3,65$ radianos, o qual corresponde ao horário de $13h56min$. No entanto, o valor da média direcional foi de $4,00$ radianos, a qual equivale ao horário das $15h16min$, note que há uma diferença de mais de uma hora. De fato, considerar dados circulares como lineares pode conduzir a resultados equivocados. A seguir serão analisados separadamente os anos de 2020 e 2021. As médias direcionais foram, respectivamente $4,07$ radianos e $3,96$ radianos, sendo os horários 15h32min, para o ano de 2020 e 15h 07 min para o ano de 2021. Na figura ( 3) estão representados os horários dos acidentes para cada ano.

Em seguida foram analisados os dados das três regiões próximas às cidades mineiras de Lavras, São João del Rei e Teófilo Otoni. Foram registrados 418 horários de acidentes para a região de Teófilo Otoni, 469 para a região de Lavras e 273 para São João del Rei.

Como os valores do $\bar{R}$, o comprimento médio resultante, não está muito próximo de 1, a variância $V$ dos dados é alta, pois $V = 1 - \bar{R}$, outro fator a ser observar é que média circular das regiões de Lavras e São João del Rei são bem próximas, indicando que os acidentes nestas regiões ocorrem, em média, em horários bem próximos.

Na tabela ( 1) estão listados o total de ocorrências, a média circular, o respectivo horário e comprimento médio.

Os gráficos dos horários dos acidentes por localidade e ano estão na figura ( 4).

Considerando que os dados seguem a distribuição von Mises foi construído o boxplot circular, os gráficos estão nas figuras (5) e ( 6), nos boxplots há a presença de outliers, estes são referentes aos horários em que não é comum a ocorrência de acidentes, ou pouco provável que eles ocorram.

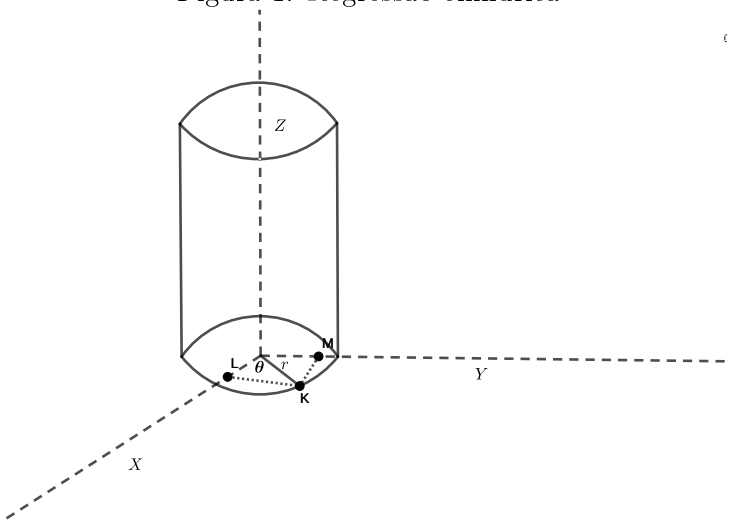
Na tabela ( 2) estão algumas medidas da estatística descritiva dos dados dos acidentes nas estradas e a estimativa do parâmetro $\kappa$ da distribuição von Mises.

Através da mediana direcional, percebe-se que a metade dos acidentes ocorrem até às 15h em Lavras e até às 16h na região de São João del Rei e Teófilo Otoni. Com relação ao parâmetro $\kappa$ da distribuição von Mises, como $\kappa < 1$ para as regiões de São Del Rei e Teófilo Otoni no ano de 2020 e Teófilo Otoni em 2021, nota-se claramente que não há uma concentração dos dados em determinados horários específicos.

Após entender como estão distribuídos os horários dos acidentes, foi avaliado o tempo de início dos atendimento às vítimas, isto é, o tempo em que a viatura policial foi acionada para que o atendimento às vítimas fosse iniciado.

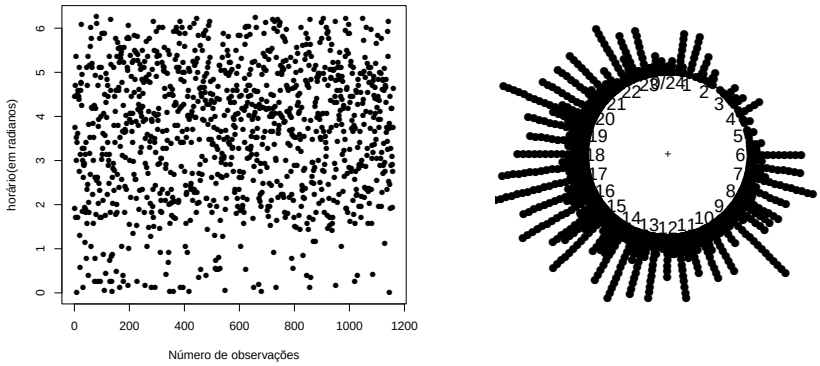
O primeiro passo foi encontrar o estimador não paramétrico de Kaplan-Meier para

Figura 1: Regressão cilíndrica




Fonte: Os autores (2022)

Figura 2: Horário das ocorrências dos acidentes



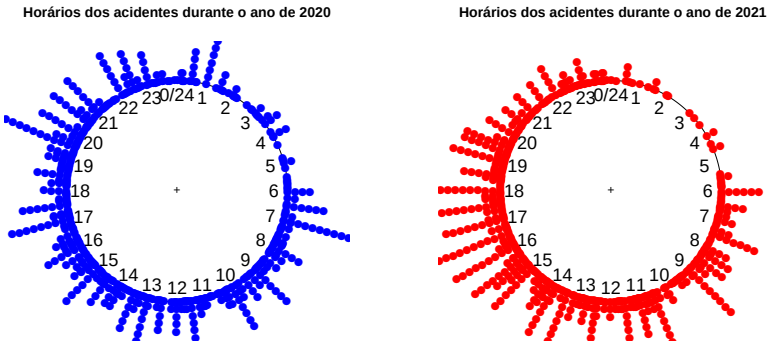

Fonte: Os autores (2022)

Tabela 1: Média Direcional, média dos horários dos acidentes e $\bar{R}$ para cada região

| Região | Número de dados | Média Direcional | Horário | $\bar{R}$ |
|---|---|---|---|---|
| Teófilo Otoni | 418 | 4,17 | 15h56min | 0,33 |
| Lavras | 469 | 3,91 | 14h56min | 0,29 |
| São João del Rei | 273 | 3,87 | 14h44min | 0,27 |

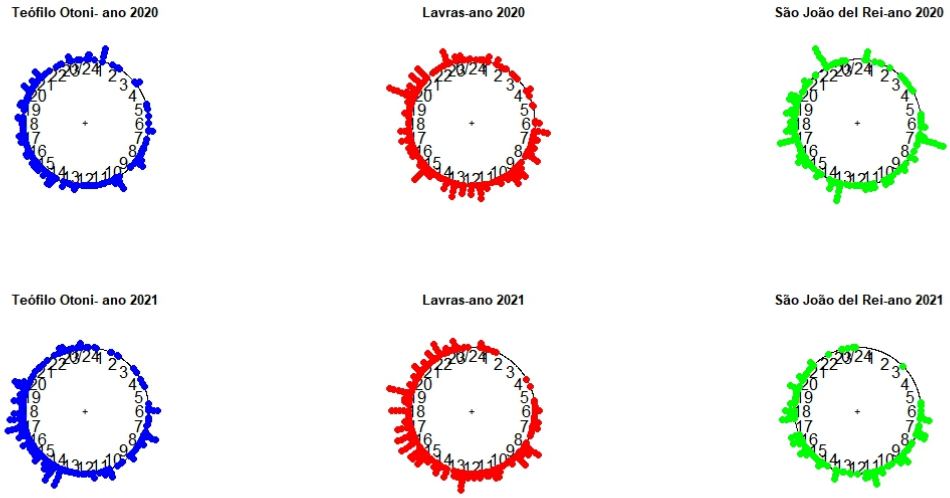
Fonte: Os autores (2022)

Figura 3: Horário das ocorrências dos acidentes para os anos 2020 e 2021



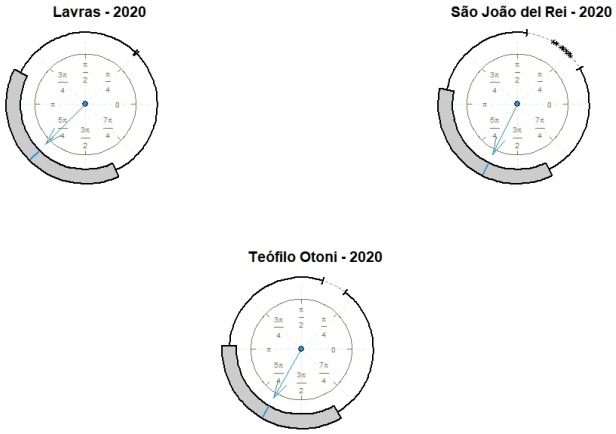

Fonte: Os autores (2022)

Figura 4: Horário das ocorrências dos acidentes por regiões e ano



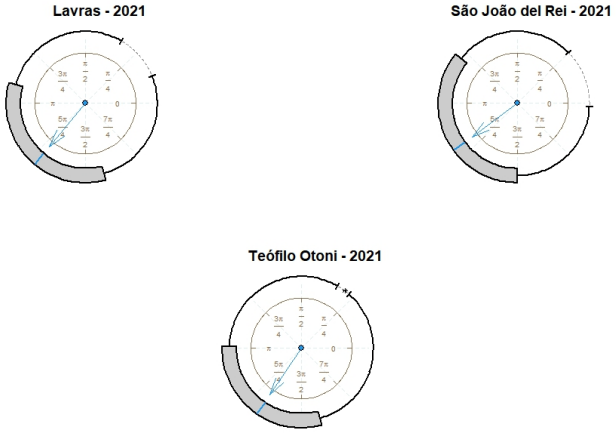

Fonte: Os autores (2022)

Figura 5: Gráfico boxplot das Localidades de Lavras, São João del Rei e Teófilo Otoni referente ao ano de 2020




Fonte: Os autores (2022)

Figura 6: Gráfico boxplot da localidades de Lavras, São João del Rei e Teófilo Otoni para o ano de 2021




Fonte: Os autores (2022)

cada localidade. As curvas estão representadas na figura ( 7) e nota-se que, com relação ao tempo para início dos atendimentos aos envolvidos em acidentes no estado de Minas Gerais durante os anos de 2020 e 2021, conjuntamente, este tempo é menor em Lavras e São João del Rei, e maior na localidade de Teófilo Otoni.

Para fins de comparação, para Teófilo Otoni, foram registrados 153 dados e o tempo médio para início do atendimento foi de $172,14$ minutos; para a região de Lavras, foram registrados 300 dados e o tempo médio para início do atendimento foi de $90,05$ minutos e São João del Rei, foram 217 dados e tempo de $101,75$ minutos.

A seguir foi feita uma comparação do modelo utilizando, diretamente, a variável circular, o horário em que o acidente ocorreu, medido em radianos, e este modelo foi comparado com o modelo proposto, regressão linear-circular (MARDIA; SUTTON, 1978). Em em todos o casos a variável resposta, foi modelada pelos modelos Weibull, Log-normal e Gama e o critério de escolha foi o AIC. Na tabela ( 3) estão os valores do AIC para os modelos :

1. Modelo I : $T = \exp\{\beta_0 + \beta_1\theta_i\}$
2. Modelo II : $T = \exp\{\beta_0 + \beta_1\cos\theta_i + \beta_2\sin\theta_i\}$

Em que a covariável circular é ângulo correspondente aos horários dos acidentes.

Os valores do AIC para o modelo linear-circular foram menores, e a melhor distribuição foi a log normal.

Além da covariável, horário da ocorrência do acidente, uma variável circular, também foram incluídas duas variáveis dicotômicas:

1) Acidentes com ou sem vítimas. Sendo considerado 1 se houver vítimas e 0 (zero) caso não haja vítimas.

2) O acidente ocorreu no período de segunda a sexta-feira, ou no sábado ou domingo. Neste caso, o valor foi 1 se ocorreu no final de semana e 0 (zero) se ocorreu em outros dias da semana.

Também foram testados os modelos com de uma a inclusão da covariável. A variável é dicotômica : existência ou não de vítimas. A tabela ( 4) mostra os valores do AIC dos modelos em análise.

O modelo Log normal foi o modelo selecionado, o modelo linear-circular, o modelo completo foi dado por

$$y(t) = \beta_0 + \beta_1\cos(\theta) + \beta_1\sin(\theta) + \beta_3 V + \beta_4 S + \epsilon_i \tag{37}$$

Em que $\theta$ é a variável circular, $V$ é a covariável existência ou de vítimas, $S$ é a covariável período da semana em que o acidente ocorreu (sábado /domingo ou segunda-sexta) e $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ são os parâmetros que serão estimados e $\epsilon_i$, vetor de erros. Quatro modelos foram testados. Os resultados estão na tabela ( 5),sendo que os valores que estão com asteriscos indicam que o valor p é menor que $0,05$. O critério de escolha foi os menores valores para o AIC e o BIC.

O modelo selecionado possui a variável circular e foi incluída a covariável dicotômica, acidente com ou sem vítimas.

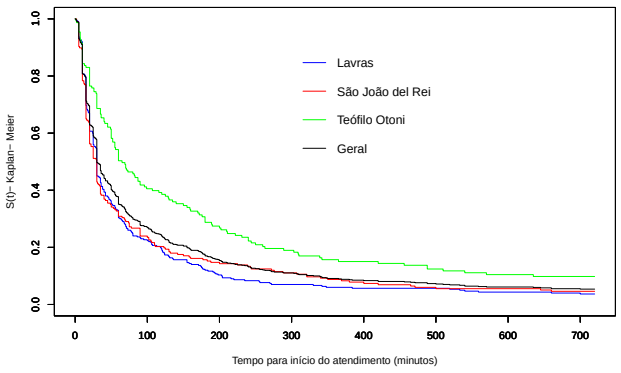
As curvas de Kaplan - Meier foram construídas para comparar se há diferenças entre o tempo de início dos atendimentos quando há ou não vítimas e se o acidente ocorre sábado / domingo ou segunda-sexta-feira, figuras ( 8) e ( 9). O teste log-rank foi aplicado para

Tabela 2: Resumo da estatística descritiva e parâmetro de concentração ($\kappa$) para os anos de 2020 e 2021

| Ano | Medidas descritivas | Lavras | São João del Rei | Teófilo Otoni |
|---|---|---|---|---|
| 2020 | Média direcional | 3,88(14h 49min) | 4,11(15h 4min) | 4,24(16h 12min) |
| | Mediana | 3,93(15h 01min) | 4,25(16h 14min) | 4,19(16h) |
| | $\bar{R}$ | 0,23 | 0,21 | 0,29 |
| | Variância | 0,77 | 0,79 | 0,71 |
| | Desvio padrão | 0,73 | 0,68 | 1,57 |
| | $\kappa$ | 0,48 | 0,42 | 0,61 |
| 2021 | Média direcional | 3,93(15h 01min) | 3,70(14h 08min) | 4,12(15h 44min) |
| | Mediana | 4,01(15h 19min) | 3,78(14h 27min) | 4,12(15h 44min) |
| | $\bar{R}$ | 0,34 | 0,36 | 0,36 |
| | Variância | 0,66 | 0,64 | 0,64 |
| | Desvio padrão | 0,73 | 0,95 | 0,95 |
| | $\kappa$ | 0,72 | 0,77 | 0,78 |

Fonte: Os autores (2022)

Figura 7: Curva de Kaplan-Meier para o tempo de início dos atendimentos às vítimas dos acidentes por localidade




Fonte: Os autores (2022)

Tabela 3: Valores do AIC para os Modelos Weibull, Log-normal e Gama

| | | AIC | | |
|---|---|---|---|---|
| | Modelos | Weibull | Log-normal | Gama |
| Modelo I | Geral | 7161,19 | 7027,23 | 7215,55 |
| | Teófilo Otoni | 1691,12 | 1666,83 | 1702,29 |
| | Lavras | 3125,44 | 3030,92 | 3160,87 |
| | São João del Rei | 2260,61 | 2192,51 | 2292,65 |
| Modelo II | Geral | 7081,13 | 6905,45 | 7154,90 |
| | Teófilo Otoni | 1690,19 | 1666,54 | 1700,56 |
| | Lavras | 3112,46 | 3030,66 | 3138,07 |
| | São João del Rei | 2252,41 | 2184,41 | 2283,87 |

Fonte: Os autores (2022)

Tabela 4: Valores do AIC para os Modelos Weibull, Log-normal e Gama, com a inclusão de uma covariável em cada modelo

| | | AIC | | |
|---|---|---|---|---|
| | Modelos | Weibull | Log-normal | Gama |
| Com ou sem vítimas | Geral | 7046,86 | 6888,31 | 7114,03 |
| | Teófilo Otoni | 1678,51 | 1657,71 | 1688,24 |
| | Lavras | 3103,42 | 3028,84 | 3126,11 |
| | São João del Rei | 2248,18 | 2184,63 | 2276,81 |
| Sábado / domingo ou Segunda - Sexta-feira | Geral | 7080,58 | 6906,61 | 7152,94 |
| | Teófilo Otoni | 1692,06 | 1668,39 | 1702,47 |
| | Lavras | 3113,29 | 3032,30 | 3138,60 |
| | São João del Rei | 2249,37 | 2184,91 | 2277,47 |

Fonte: Os autores (2022)

avaliar se há diferenças entre os tempos de início dos atendimentos, nas situações em que há ou não vítimas; o acidente ocorre aos sábados / domingos ou no período de segunda - sexta. O resultado está na tabela ( 6). O valor p da estatística do teste log-rank, mostra que não há diferença entre os tempos de início dos atendimentos quando os acidentes ocorrem nos finais de semana e no período de segunda a sexta-feira. Essa constatação reforça a não escolha da covariável do período da semana em que o acidente ocorreu.

Uma maneira de avaliar a qualidade do modelo é por métodos gráficos, na figura ( 10), está representado a gráfico das probabilidades de sobrevivência dos resíduos estimados pelo estimador de Kaplan-Meier e pela log normal padrão e suas respectivas curvas de sobrevivência. Quando o modelo se ajusta bem aos dados os pontos do gráfico estão bem próximos da reta $y = x$, em que $y$ a função de sobrevivência da log normal e $x$ a função do estimador de Kaplan- Meier. Também as curvas de sobrevivência da log normal padrão e Kaplan - Meier, avaliadas nos resíduos, estão bem próximas.

O resíduo deviance foi utilizado para avaliar o modelo, nota-se que os resíduos estão distribuídos aleatoriamente em torno da reta $y = 0$, evidenciando que os dados estão bem ajustados, figura ( 11).

Na figura ( 12) há a representação gráfica dos resíduos martingales. Nos gráficos é possível observar que alguns horários estão mal ajustados, isto pode ser notado pelos pontos cujos resíduos são mais negativos se comparados aos demais. Uma explicação para isto é que em algumas situações o início dos atendimentos foi superior a 720 minutos, um dado censurado. O tempo longo produziu uma estimativa maior (negativamente) para o resíduo. No entanto, de um modo geral, os resíduos indicam o bom ajuste do modelo proposto para os dados do tempo de início dos atendimentos.

Diante da verificação dos resíduos o modelo de regressão escolhido para o tempo foi o log normal, e utilizou modelo linear-circular (cujas variáveis explicativas são circulares e a resposta é linear), e foi incluída a variável dicotômica se houve ou não vítimas. Assim, o modelo para a função de sobrevivência é dada por:

$$S(t) = \Phi\left(\frac{-\log(tempo) + \beta_0 + \beta_1 \cos\theta_i + \beta_2 \sin\theta_i + \beta_3 x_3}{\sigma}\right) \tag{38}$$

sendo $\Phi()$ a distribuição acumulada da normal padrão e

$$x_3 = \begin{cases} 1, & \text{se houve vítimas} \\ 0, & \text{se não houve vítimas} \end{cases}$$

Os parâmetros do modelo selecionado estão na tabela ( 7), é possível observar que:

1. $\beta_1$ e $\beta_2$ são positivos, nos horários das 0 h às 6 h os atendimentos demoram mais.

2. Nos horários das 12 h às 18 h a probabilidade dos atendimentos serem mais rápidos é maior.

3. Para a localidade de Lavras, como $\beta_2$ é muito menor que $\beta_1$, quando os acidentes ocorrem entre 6 h e 12 h eles são atendidos mais rapidamente.

4. O coeficiente $\beta_3$ negativo indica que quando há vítimas os atendimentos são mais rápidos.

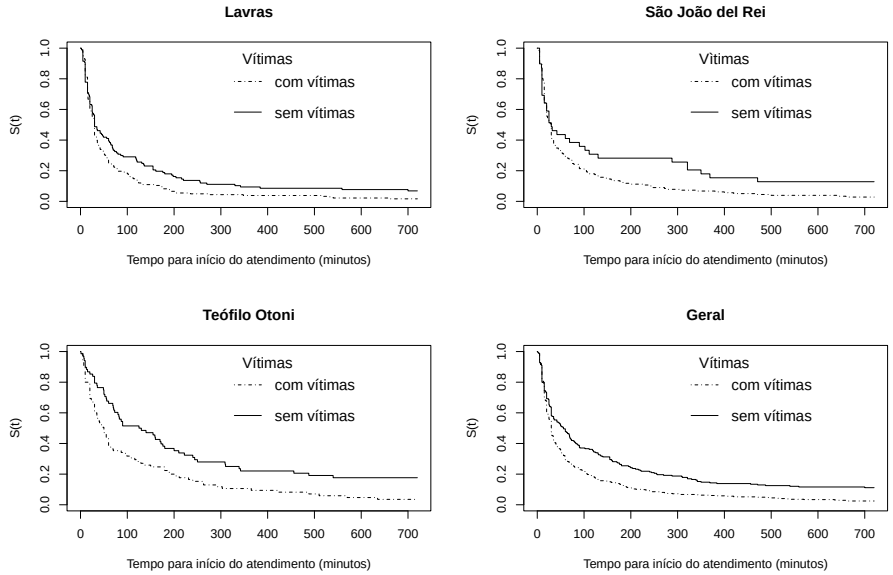
A razão de risco entre os tempos medianos foi estimada para o caso com ou sem vítimas:

Tabela 5: Valores dos parâmetros para o modelo Log-normal, com seus respectivos AIC e BIC

| Modelos | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\sigma$ | AIC | BIC |
|---|---|---|---|---|---|---|---|---|
| Geral | 4,23* | 0,29* | 0,23* | -0,51* | -0,08 | 1,42 | 6889,81 | 6916,85 |
| | 3,87* | 0,28* | 0,24* | | | 1,44 | 6905,45 | 6923,48 |
| | 4,21* | 0,28* | 0,23* | -0,51* | | 1,42 | 6888,31 | 6910,85 |
| | 3,91* | 0,29* | 0,24* | | -0,11 | 1,43 | 6906,61 | 6929,14 |
| Lavras | 3,87* | 0,20 | 0,04 | -0,30 | -0,08 | 1,30 | 3030,60 | 3052,86 |
| | 3,66* | 0,20 | 0,05 | | | 1,31 | 3030,66 | 3045,48 |
| | 3,84* | 0,20 | 0,04 | 0,05* | | 1,30 | 3028,84 | 3047,36 |
| | 3,69* | 0,20* | 0,05 | | -0,10 | 1,31 | 3032,30 | 3050,81 |
| S.J.del Rei | 4,14* | 0,39* | 0,41* | -0,35 | -0,25 | 1,39 | 2184,97 | 2205,25 |
| | 3,76* | 0,37* | 0,44* | | | 1,40 | 2184,41 | 2197,93 |
| | 4,03* | 0,37* | 0,41* | -0,33 | | 1,39 | 2184,63 | 2201,53 |
| | 3,85* | 0,39* | 0,43* | | -0,24 | 1,39 | 2184,91 | 2201,81 |
| Teófilo Otoni | 4,84* | 0,27 | 0,32 | -0,85* | 0,19 | 1,53 | 1659,23 | 1677,41 |
| | 4,43* | 0,25 | 0,33 | | | 1,58 | 1666,54 | 1678,66 |
| | 4,89* | 0,27 | 0,31 | -0,84* | | 1,53 | 1657,71 | 1672,87 |
| | 4,39* | 0,25 | 0,34 | | 0,11 | 1,58 | 1668,39 | 1683,54 |

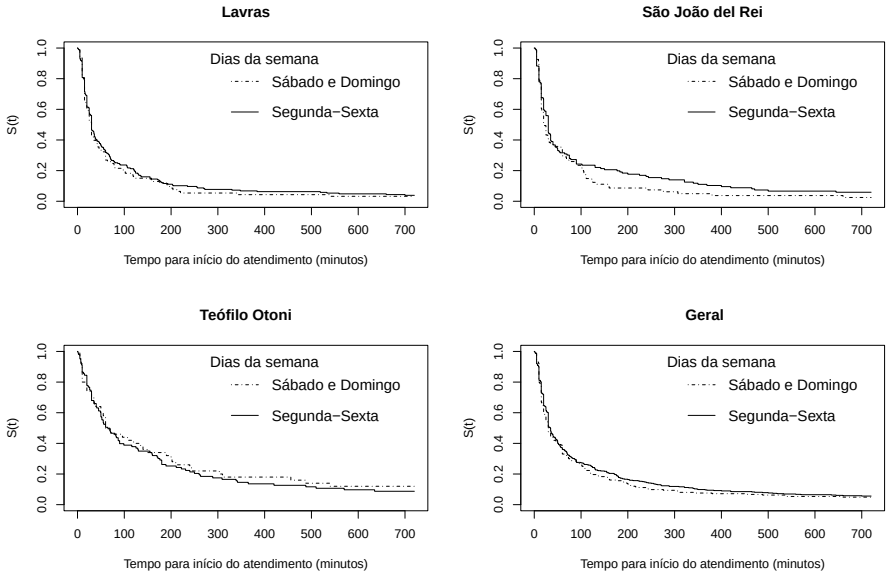
Fonte: Os autores (2022)

Figura 8: Curva de sobrevivência - Kaplan-Meier para a variável vítimas




Fonte: Os autores (2022)

Figura 9: Curva de sobrevivência - Kaplan-Meier para a variável dias da semama




Fonte: Os autores (2022)

Tabela 6: Resultado do teste log-rank

| | Com/sem vítimas | | Finais de semana/segunda-sexta | |
|---|---|---|---|---|
| Regiões | | | | |
| | Estatística do teste | valor p | Estatística do teste | valor p |
| Lavras | 6,38 | 0,01 | 0,46 | 0,50 |
| São João del Rei | 4,22 | 0,04 | 2,22 | 0,10 |
| Teófilo Otoni | 11,10 | $<$ 0,01 | 0,28 | 0,60 |
| Geral | 24,90 | $<$ 0,01 | 1,02 | 0,30 |

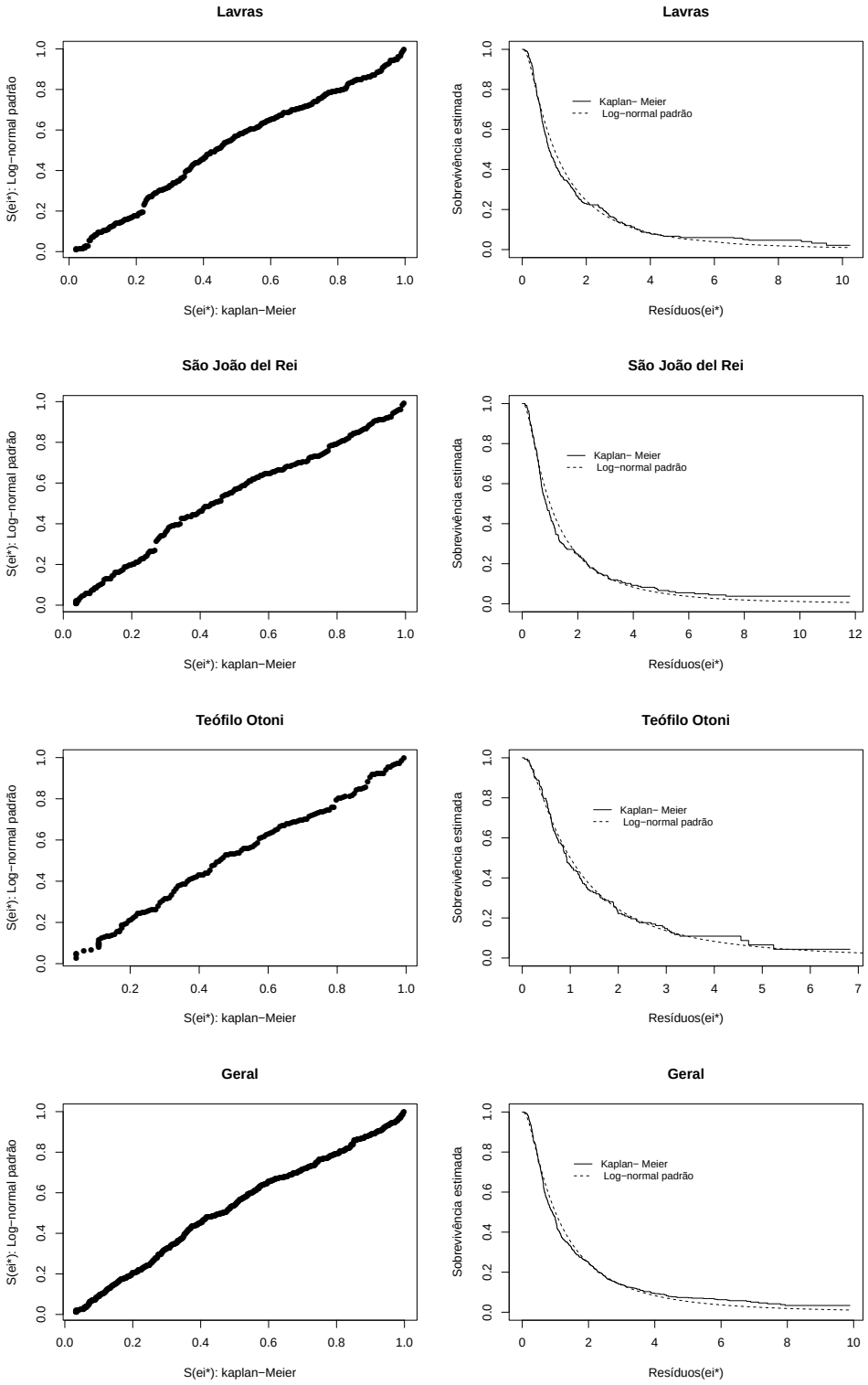
Fonte: Os autores (2022)

Tabela 7: Estimativas dos parâmetros e erro padrão (EP) para o modelo Log-normal

| Modelos | $\beta_0(EP)$ | $\beta_1(EP)$ | $\beta_2(EP)$ | $\beta_3(EP)$ | $\sigma(EP)$ |
|---|---|---|---|---|---|
| Geral | 4,21 (0,10) | 0,28 (0,09) | 0,23 (0,08) | -0,51 (0,12) | 1,42 (0,03) |
| Lavras | 3,84 (0,13) | 0,20 (0,12) | 0,04 (0,11) | -0,30 (0,15) | 1,30 (0,04) |
| S.J.del Rei | 4,03 (0,23) | 0,37 (0,16) | 0,41 (0,13) | -0,33 (0,25) | 1,39 (0,05) |
| Teófilo Otoni | 4,89 (0,19) | 0,27 (0,19) | 0,31 (0,18) | -0,84 (0,25) | 1,53 (0,06) |

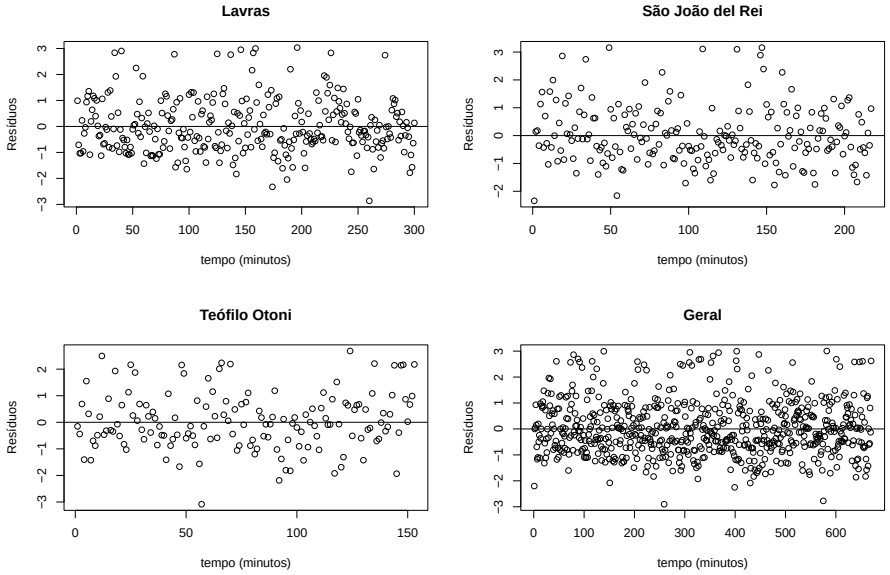
Fonte: Os autores (2022)

Figura 10: Sobrevivência dos resíduos estimados pela Log normal padrão e pelo método de Kaplan-Meier para as três localidades



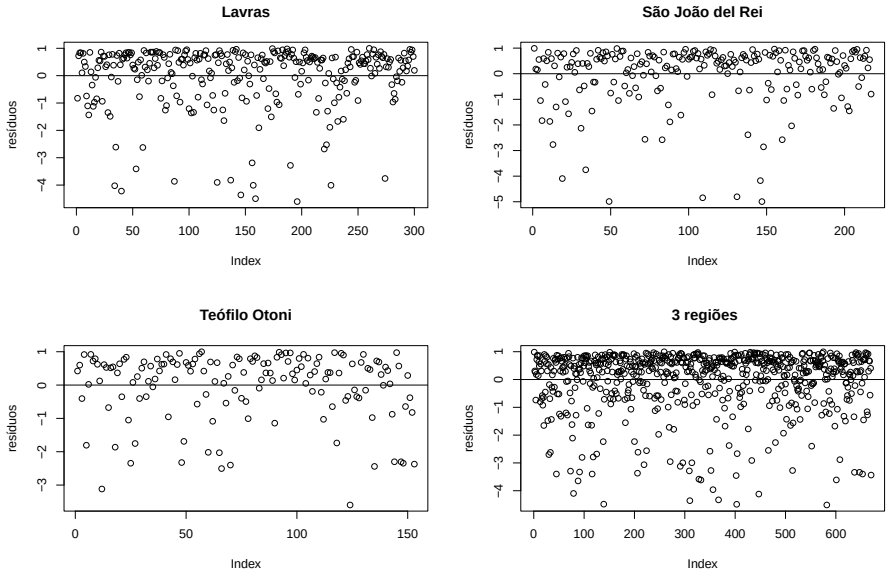

Fonte: Os autores (2022)

Figura 11: Resíduos deviance




Fonte: Os autores (2022)

Figura 12: Resíduos martingale




Fonte: Os autores (2022)

$$\frac{t_{0,5}(x_3=0,\hat{\beta})}{t_{0,5}(x_3=1,\hat{\beta})} = \frac{\exp[\sigma z_{0,5}]\exp[\beta_0+\beta_1 x_1+\beta_2 x_2]}{\exp[\sigma z_{0,5}]\exp[\beta_0+\beta_1 x_1+\beta_2\ x_2+\beta_3 x_3]} = \frac{1}{\exp(\beta_3)} \tag{39}$$

em que $\hat{\beta}$ é o vetor de parâmetros do modelo que foram estimados.

A razão de risco entre os tempos medianos para cada região está na tabela ( 8), comparando os resultados, percebe-se que em todas as localidades quando há vítimas, os atendimentos são mais rápidos. Por exemplo, para a localidade de Lavras, quando não há vítimas o tempo para início dos atendimentos é 67% maior, ou demora mais, em relação ao tempo para início dos atendimentos quando há vítimas.

Tabela 8: Razão de risco para tempos medianos por localidade

| Modelos | Razão de riscos |
|---|---|
| Geral | 1,67 |
| Lavras | 1,35 |
| S.J.del Rei | 1,39 |
| Teófilo Otoni | 2,33 |

Fonte: Os autores (2022)

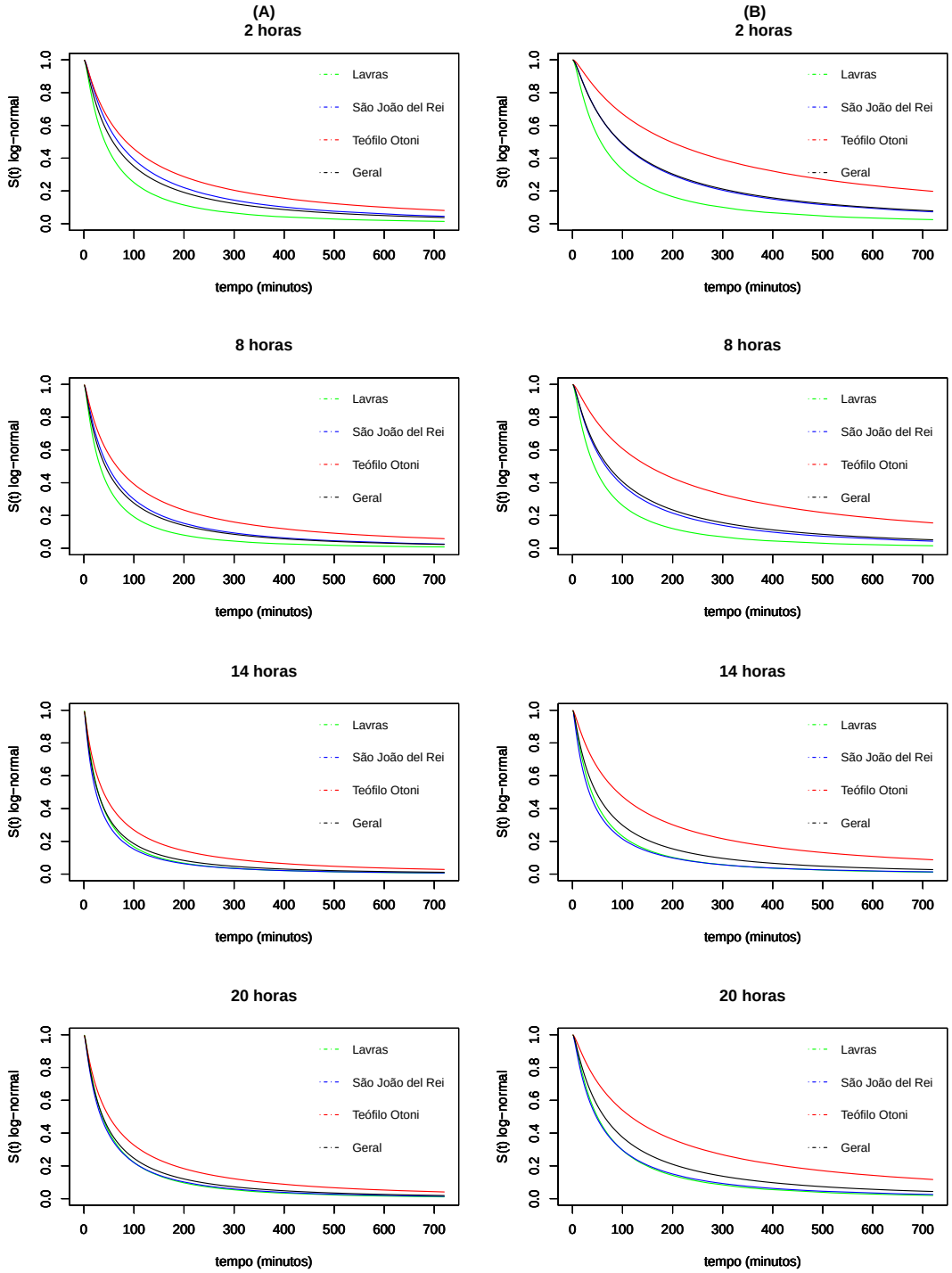
A seguir, para compreensão do modelo proposto, serão analisadas algumas situações quando os ângulos medem 30°, 120°, 210° e 300°, que são os equivalentes aos horários 2h, 8h, 14h e 20h, para cada uma das medidas angulares será verificada a situação em que os acidentes são com vítimas ou sem vítimas, observe que foram escolhidas medidas em cada um dos quadrantes. Isto será feito para os modelos de cada localidade e o geral, os gráficos estão na figura ( 13). Em todos os cenários a localidade de Teófilo Otoni tem um tempo para início dos atendimentos maior, e durante às 2 h e 8 h da manhã o tempo é menor para Lavras, à tarde e à noite (14 h e 20 h) Lavras e São João del Rei têm comportamentos semelhantes.

## Conclusão

As técnicas da estatística circular têm se mostrado bastante úteis ao se estudar dados em que há a presença de medidas angulares e periódicas. Com a utilização dessas informações, na análise de sobrevivência, foi possível estudar horários de acidentes (que são medidas que podem ser convertidas em angulares) e verificar o tempo para início dos atendimentos, das três localidades estudadas. Nota-se que a região de Teófilo Otoni tem um maior tempo para início dos atendimentos, os motivos pelos quais isto ocorre podem ser investigados, apesar disso, quando ocorrem vítimas nos acidentes, o tempo para início dos atendimentos é menor em todas as localidades e horários, o que é muito importante para a preservação das vidas.

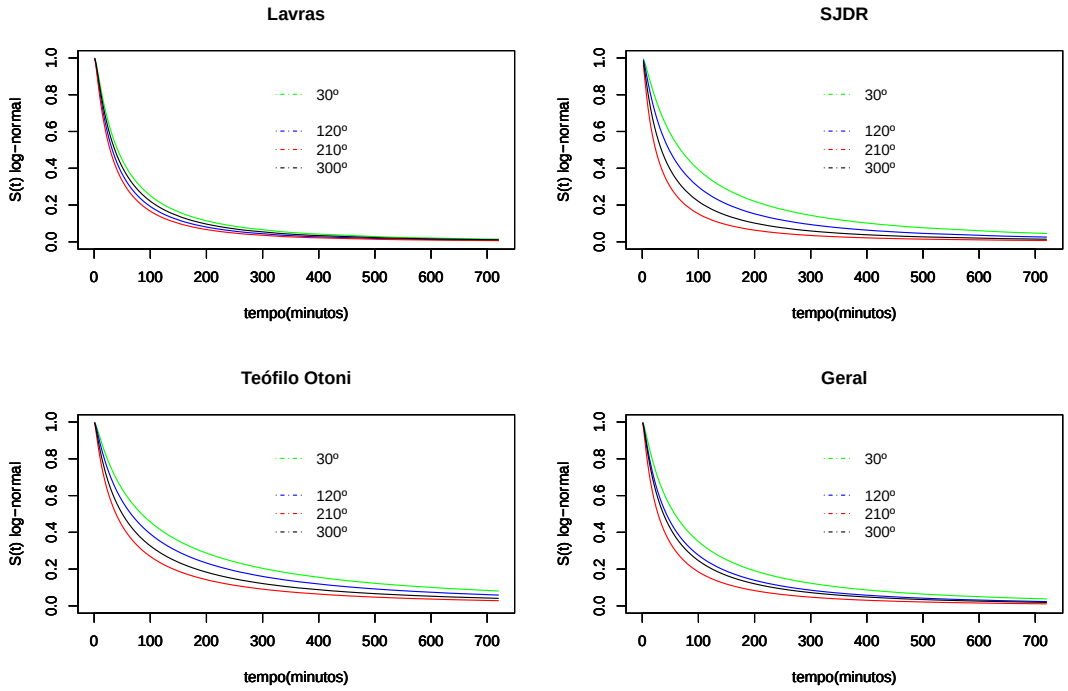
Notou-se também que não havia diferença significativa entre o tempo para se iniciar os atendimentos, caso os acidentes ocorressem nos finais de semana ou nos demais dias. Esta covariável não foi incluída no modelo, mas o parâmetro estimado para esta covariável foi negativa em todos os modelos, exceto para Teófilo Otoni, tal fato indica que o tempo para início do atendimento é um pouco menor quando os acidentes ocorrem aos sábados e domingos em relação aos demais dias da semana, no entanto essa diferença não foi significativa.

Figura 13: Curvas de sobrevivência para os horários das $2h$, $8h$, $14h$, $20h$ equivalendo aos ângulos 30°, 120°, 210°, 300°, respectivamente, para início dos atendimentos de acidentes: (A) com vítimas; (B) sem vítimas




Fonte: Os autores (2022)

Figura 14: Curvas de sobrevivência por localidades por horário $2h, 8h, 14h, 20h$, respectivamente, $30°, 120°, 210°, 300°$




Fonte: Os autores (2022)

# Referências


ABUZAID, A. H.; MOHAMED, I. B.; HUSSIN, A. G. Boxplot for circular variables. **Computational Statistics**, v. 27, n. 3, p. 381–392, 2012.

ANDERSON-COOK, C. M.; NOBLE, R. B. An alternate model for cylindrical data. **Nonlinear Analysis**, v. 47, p. 2011–2022, 2001.

ASCARI, R. A. et al. Perfil epidemiológico de vítimas de acidente de trânsito. **Revista de Enfermagem da UFSM**, v. 3, n. 1, p. 112–121, 2013.

BUTTARAZZI, D.; PANDOLFO, G.; PORZIO, G. C. A boxplot for circular data. **Biometrics**, v. 74, n. 4, p. 1492–1501, 2018.

CARRASQUINHA, E.; VERÍSSIMO, A.; VINGA, S. Consensus outlier detection in survival analysis using the rank product test. **bioRxiv**, p. 1–27, 2018.

CARVALHO, M. S. et al. **Análise de sobrevivência: teoria e aplicações em saúde**. 2. ed. Rio de Janeiro: Editora FIOCRUZ, 2011.

COLOSIMO, E.; GIOLO, S. **Análise de sobrevivência aplicada**. São Paulo: Blucher, 2006.

DESLANDES, S. F. O atendimento às vítimas de violência na emergência:"prevenção numa hora dessas?". **Ciência  Saúde Coletiva**, v. 4, p. 81–94, 1999.

DIAS, E. G. et al. Acidentes de trânsito com motocicleta atendidos pelo SAMU em uma cidade do Norte de Minas. **Saúde (Santa Maria)**, 2018.

DINIZ, A. M. A. et al. Acidentes de trânsito em Minas Gerais (2003): uma abordagem espacial. **Revista de Biologia e Ciências da Terra**, v. 8, n. 1, 2008.

FISHER, N. **Statistical analysis of circular data**. New York: Cambridge University Press, 1993.

JAMMALAMADAKA, S.; SENGUPTA, A. **Topics in circular statistics**. Singapore: World Scientific, 2001.

KLEINBAUM, D.; KLEIN, M. **Survival analysis - A self- learning text**. 3. ed. New York: Springer, 2012.

LADEIRA, R. M.; BARRETO, S. M. Fatores associados ao uso de serviço de atenção pré-hospitalar por vítimas de acidentes de trânsito. **Cadernos de Saúde Pública**, v. 24, p. 287–294, 2008.

MALVESTIO, M. A. A.; SOUSA, R. M. C. Suporte avançado à vida: atendimento a vítimas de acidentes de trânsito. **Revista de Saúde Pública**, v. 36, p. 584–589, 2002.

MARDIA, K.; SUTTON, T. W. A model for cylindrical variables with applications. **Royal Statistical Society**, v. 40, n. 2, p. 229–233, 1978.

MARDIA, K. V. **Statistics of directional data**. London: Academic Press, 1972.

MARÍN, L.; QUEIROZ, M. S. A atualidade dos acidentes de trânsito na era da velocidade: uma visão geral. **Cadernos de Saúde Pública**, v. 16, p. 7–21, 2000.

R Development Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2021. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org>.

RESENDE, V. D. et al. Ocorrências de acidentes de trânsito atendidas pelo serviço de atendimento móvel de urgência em belo horizonte.**Revista de Enfermagem do Centro-Oeste Mineiro**, 2012.

## CONSIDERAÇÕES FINAIS

O estudo da estatística circular impõe alguns desafios, se por um lado é uma área que pode ser bastante utilizada em diversos contextos, por outro lado há a necessidade de mais pesquisa e informações, o que consideramos positivo, já que, apesar de não ser uma área nova, mais trabalhos na perspectiva da estatística aplicada faz-se necessário.

Durante a realização deste trabalho este desafio tornou-se evidente, o que inlcui a necessidade de investigar mais funções e pacotes em software como R.

Pretende-se dar sequência a pesquisa, principalmente do último artigo incluindo os dados de localização geográfica dos acidentes, o que tornaria possível avaliar se há ou não a ocorrência de eventos recorrentes em alguns municípios.

Outro ponto a investigar são os resíduos na análise de sobrevivência com a inclusão das variáveis circulares, para avaliar com mais precisão a qualidade do ajuste.